Vorlesung über Approximationstheorie

Jochen Werner

Wintersemester 1984/85

Inhaltsverzeichnis

Kapitel 1

Einführung, Beispiele, Übersicht

1.1 Einführung

Eine Approximationsaufgabe ist durch die folgenden Daten gegeben:

1. Einen (reellen) linearen normierten Raum X, dessen Norm mit $\|\cdot\|$ bezeichnet wird (der Raum, in dem sich alles "abspielt"),
2. ein Element $z \in X$ (das zu approximierende Element),
3. eine Menge $M \subset X$ (Menge der Elemente, mit denen approximiert wird)

und besteht in der Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \|x - z\|, \quad x \in M.$$

Eine (globale) Lösung x^* von (P), also ein $x^* \in M$ mit $\|x^* - z\| \leq \|x - z\|$ für alle $x \in M$, heißt (global) *beste Approximierende* an z in M. Entsprechend ist eine *lokal beste Approximierende* definiert als ein $x^* \in M$ mit der Eigenschaft, dass eine Umgebung U von x^* gibt mit $\|x^* - z\| \leq \|x - z\|$ für alle $x \in M \cap U$. Mit

$$d(z, M) := \inf_{x \in M} \|x - z\|$$

bezeichnet man den *Abstand* (oder auch den *Minimalabstand*) von $z \in X$ zu M. Wie bei den (allgemeineren) Optimierungsaufgaben stellen sich naheliegenderweise bei konkret gegebenen Approximationsaufgaben Fragen nach der Existenz sowie der Eindeutigkeit einer Lösung, nach notwendigen sowie hinreichenden Optimalitätsbedingungen, nach der numerischen Behandlung. Ist $\{M_k\} \subset X$ aufsteigend, also $M_k \subset M_{k+1}$, $k = 0, 1, \ldots$, und schreibt man zur Abkürzung $E_k(z) := d(z, M_k)$, so gilt $E_{k+1}(z) \leq E_k(z)$. Falls $\lim_{k\to\infty} E_k(z) = 0$ (eine Aussage vom Weierstraß-Typ), versucht man noch zu untersuchen, wie schnell $\{E_k(z)\}$ gegen Null konvergiert und wie diese Konvergenz von der "Glattheit" von z abhängt (Aussagen vom Jackson-Typ).

Man sieht, dass sich hier ein großes Programm auftut. Wir werden bei den meisten Fragen bzw. Problemen einen funktionalanalytischen Zugang verfolgen, d. h. die Probleme werden zunächst in einem möglichst allgemeinen Rahmen behandelt. Dies kann man

vergleichen mit dem aufstieg auf einen Aussichtsturm, von dem man sich einen besseren Überblick und eine Orientierung erhofft. Wenn man allerdings zu hoch steigt, so ist man in den Wolken und sieht gar nichts mehr. Außerdem wollen wir auch immer wieder von dem Turm herabsteigen, um uns die Dinge aus der Nähe anzusehen (Beispiele!). Für den Aufstieg benötigen wir Hilfsmittel und zwar solche, die die Funktionalanalysis bereithält. Wir werden uns diese Treppen zum Teil selber bauen, d. h. die benötigten Hilfsmittel selbst entwickeln, oder sie einfach benutzen oder noch einfacher: mit dem Fahrstuhl fahren.

Der Zusammenhang zwischen Approximations- und Optimierungsaufgaben ist offensichtlich. Approximationsaufgaben sind spezielle Optimierungsaufgaben, bei denen die Zielfunktion f im wesentlichen durch eine Norm gegeben ist, nämlich durch $f(x) := \|x - z\|$. Diese Zielfunktion hat den großen Vorteil, eine *konvexe* Funktion zu sein, aber auch den großen Nachteil, i. Allg. nicht im klassischen Sinne differenzierbar zu sein. Trotzdem wird die Vorgehensweise oft ähnlich wie in der Optimierung sein.

1.2 Beispiele von Approximationsaufgaben

Es folgt eine Sammlung von Beispielen für Approximationsaufgaben.

Beispiel 1.2.1 (Lineare Regression, Prinzip der kleinsten Quadrate) Im einfachsten Fall liegen zu Zeiten t_i (diese seien nicht alle gleich) Beobachtungen z_i, $i = 1, \ldots, m$, vor. Es wird ein linearer Zusammenhang zwischen den t_i und den z_i vermutetet, dass also $z_i \approx at_i + b$, $i = 1, \ldots, m$, mit Konstanten a und b und man sucht nach den "besten" Konstanten. Z. B. gibt t_i die Länge eines i-ten Säuglings bei der Geburt und z_i die zugehörige Schwangerschaftsdauer an. In Tabelle 1.1 haben wir spezielle Werte angegeben. Nach dem Prinzip der kleinsten Quadrate (auf dessen statistische

t [cm]	48	49	50	51	52
z [Tage]	277.1	279.3	281.4	283.2	284.8

Tabelle 1.1: Schwangerschaftsdauer und Länge bei der Geburt

Grundlagen wir nicht eingehen wollen), werden a, b als Lösung von

$$\text{(P)} \qquad \text{Minimiere} \quad f(a,b) := \sum_{i=1}^{m} (at_i - z_i)^2, \quad (a,b) \in \mathbb{R}^2,$$

bestimmt. Wieso ist dies eine Approximationsaufgabe? Hierzu setze man $X := \mathbb{R}^m$ und wähle als Norm die euklidische Norm $\|\cdot\|_2$, $z := (z_1, \ldots, z_m)^T$ und

$$M := \left\{ a \begin{pmatrix} t_1 \\ \vdots \\ t_m \end{pmatrix} + b \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} : a, b \in \mathbb{R} \right\}.$$

Man erkennt, dass die Aufgabe (P) äquivalent zu der Approximationsaufgabe mit den Daten X, z und M ist. Bekanntlich kann die eindeutige Lösung (a^*, b^*) von (P) aus der notwendigen und hinreichenden Optimalitätsbedingung

$$\nabla f(a^*, b^*) = \begin{pmatrix} \frac{\partial f}{\partial a}(a^*, b^*) \\ \frac{\partial f}{\partial b}(a^*, b^*) \end{pmatrix} = \begin{pmatrix} 2\sum_{i=1}^{m}(a^* t_i + b^* - z_i)t_i \\ 2\sum_{i=1}^{m}(a^* t_i + b^* - z_i) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

bestimmt werden. Mit den Mittelwerten

$$\overline{t} := \frac{1}{m}\sum_{i=1}^{m} t_i, \qquad \overline{z} := \frac{1}{m}\sum_{i=1}^{m} z_i$$

erhält man

$$a^* := \frac{\sum_{i=1}^{m} t_i z_i - m\overline{t}\,\overline{z}}{\sum_{i=1}^{m} t_i^2 - m\overline{t}^2} = \frac{\sum_{i=1}^{m}(t_i - \overline{t})(z_i - \overline{z})}{\sum_{i=1}^{m}(t_i - \overline{t})^2}$$

und

$$b^* := \overline{z} - a^*\overline{t}.$$

Die sogenannte *Regressionsgerade* ist dann durch $z(t) := a^*(z - \overline{t}) + \overline{z}$ gegeben. In Abbildung 1.1 geben wir zu den Daten aus Tabelle 1.1 die zugehörige Regressionsgerade an. □

Beispiel 1.2.2 (Nichtlineare diskrete Approximation im Mittel) Es wird angenommen, $z = (z_1, \ldots, z_m)^T \in \mathbb{R}^m$ sei ein bekannter Vektor, der z. B. durch m Beobachtungen gewonnen ist. Mit $n < m$ wird allgemeiner als in Beispiel 1.2.1 ein i. Allg. nichtlinearer funktionaler Zusammenhang

$$z_i \approx g_i(y_1, \ldots, y_n), \qquad i = 1, \ldots, m,$$

vermutet. Bei der *diskreten Approximation im Mittel* bzw. einem *nonlinear least square fit* bestimmt man die gesuchten Parameter als Lösung der Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(y) := \sum_{i=1}^{m}(g_i(y_1, \ldots, y_n) - z_i)^2, \quad y \in \mathbb{R}^{\cdot}$$

Im obigen Beispiel 1.2.1 ist z. B.

$$g_i(y_1, y_2) := y_1 t_i + y_2, \quad i = 1, \ldots, m.$$

Ein spezielles Beispiel zielt schon auf die später im kontinuierlichen Fall genauer zu untersuchende *Exponentialsummenapproximation*. Beim radioaktiven Zerfall eines Gemisches zweier radioaktiver Substanzen werden zu Zeiten $0 \le t_1 < t_2 < \cdots < t_m$ Intensitäten I_i, $i = 1, \ldots, m$, der emittierenden Strahlung gemessen. Gesucht sind Zerfallskonstanten b_1, b_2 für die beiden Stoffe, die im Mischungsverhältnis a_1/a_2 vorliegen

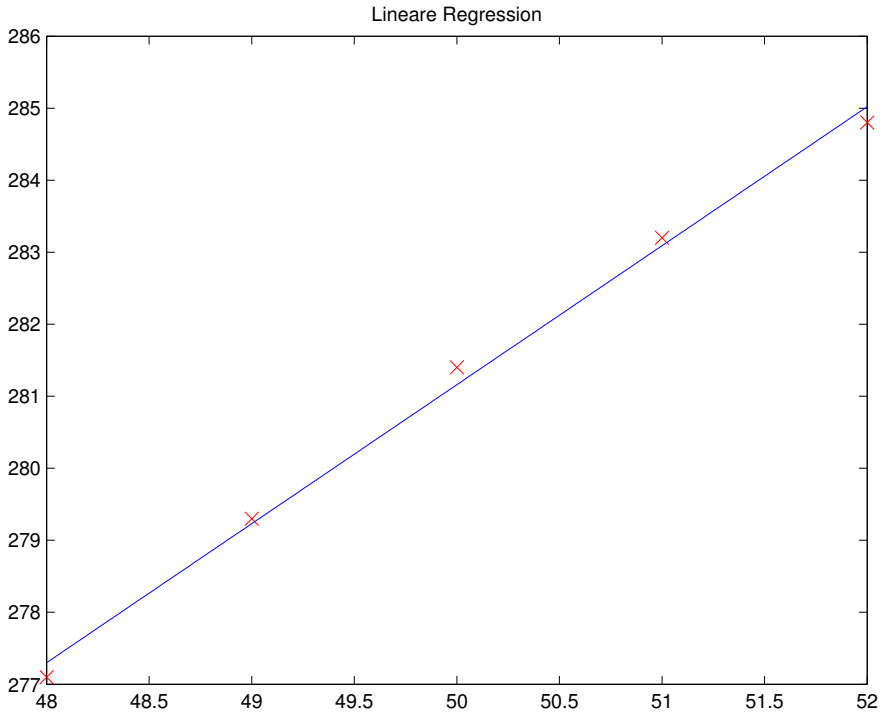

Abbildung 1.1: Die Regressionsgerade zu den Daten in Tabelle 1.1

mögen. Die Intensität ist dann proportional zu $a_1 e^{-b_1 t} + a_2 e^{-b_2 t}$ und man versucht, die Aufgabe

$$\text{(P)} \quad \text{Minimiere} \quad f(a_1, a_2, b_1, b_2) := \sum_{i=1}^{m} (a_1 e^{-b_1 t_i} + a_2 e^{-b_2 t_i} - I_i)^2, \quad (a_1, a_2, b_1, b_2) \in \mathbb{R}^4,$$

zu lösen. Allgemeiner hat man Ansätze der Form

$$g_i(a, b) := \sum_{j=1}^{p} a_j e^{-b_j t_i}$$

bei Abklingvorgängen oder

$$g_i(a, b, c) := \sum_{j=1}^{p} a_j \sin(b_j t_i - c_j)$$

bei periodischen Vorgängen. Denkbar ist aber auch ein rationaler Ansatz

$$g_i(a_0, \ldots, a_p, b_0, \ldots, b_q) := \frac{\sum_{j=0}^{p} a_j t_i^j}{\sum_{j=0}^{q} b_j t_i^j},$$

wobei nur solche Koeffizienten sinnvoll sind, für die der Nenner nicht verschwindet. Hier handelt es sich um eine diskrete Approximation durch rationale Funktionen mit vorgeschriebenem Zähler- und Nennergrad.

Typisch an vielen Approximationsaufgaben ist, dass die Menge M, mit der approximiert wird, in parametrisierter Form vorliegt, also etwa als

$$M := \left\{ \begin{pmatrix} g_1(y) \\ \vdots \\ g_m(y) \end{pmatrix} : y \in Y \right\}$$

mit einer Teilmenge Y des $\mathbb{R}^n$. Man ist nicht nur an der besten Approximierenden aus M interessiert, sondern vor allem an den Parametern, die diese erzeugen.

Auch bei der diskreten Approximation kann es sinnvoll sein, bezüglich einer anderen Norm als der euklidischen zu approximieren. Bei der *diskreten L_∞-Approximation* bzw. der *diskreten Tschebyscheff-Approximation* legt z. B. die Maximumnorm im $\mathbb{R}^m$ zu Grunde, während bei der *diskreten L_1-Approximation* die Betragssummennorm im $\mathbb{R}^m$ benutzt wird. Wir werden später sehen, dass trotz der scheinbaren Ähnlichkeit der Probleme völlig unterschiedliche Phänomene auftreten. □

Beispiel 1.2.3 (Lineare Tschebyscheff-Approximation) Sei $B \subset \mathbb{R}^N$ kompakt und $X := C(B)$ der lineare Raum der auf B stetigen reellwertigen Funktionen. Auf X wird die *Maximumnorm* (oder auch *Tschebyscheff-Norm*) definiert durch

$$\|x\|_\infty := \max_{t \in B} |x(t)|.$$

Ferner sei $M \subset C(B)$ ein endlichdimensionaler linearer Teilraum, etwa

$$M := \operatorname{span}\{v_1, \ldots, v_n\}$$

mit linear unabhängigen $v_1, \ldots, v_n$. Die Aufgabe, ein vorgegebenes $z \in X$ durch Elemente aus M bezüglich der Maximumnorm zu approximieren nennt man eine *lineare Tschebyscheffsche Approximationsaufgabe*. Für den Spezialfall $B := [a, b]$ und $M := \Pi_n$ (linearer Raum der Polynome vom Grad $\leq n$) hat P. L. Tschebyscheff diese Aufgabe behandelt. Hierbei hat man sich die zu approximierende funktion z als "kompliziert" etwa in dem sinne vorzustellen, dass sie nicht durch endlich viele elementare arithmetische Operationen berechnet werden kann, während die Elemente von M "einfach" berechenbar sind. Da bei der Tschebyscheff-Approximation die maximale Betrags-Abwerichung zur gegebenen Funktion z minimiert wird, wird sie z. B. zur Berechnung der elementaren Funktionen (Wurzel-Funktion, trigonometrische Funktionen usw.) benutzt. Wie dies für die Wurzelfunktion $z(t) := \sqrt{t}$ geschehen kann, wird ausführLich bei J. Werner (1992, S. 2ff.) beschrieben.

Beispiel 1.2.4 (Rationale Tschebyscheff-Approximation) Wie im letzten Beispiel sei $X := C(B)$ versehen mit der Maximumnorm $\|\cdot\|_\infty$. Diesmal sei allerdings die Menge M der Funktionen, mit denen approximiert wird, gegeben durch

$$M := \left\{ \frac{\sum_{j=1}^p \alpha_j u_j}{\sum_{k=1}^q \beta_k v_k} : \begin{array}{l} \alpha_j \in \mathbb{R}\ (j = 1, \ldots, p), \\ \beta_k \in \mathbb{R}\ (k = 1, \ldots, q), \end{array} \sum_{k=1}^q \beta_k v_k(t) \neq 0 \text{ für alle } t \in B \right\}.$$

Hierbei sind u_j, $j = 1, \ldots, p$, und v_k, $k = 1, \ldots, q$, vorgegebene "einfach" zu berechnende Funktionen aus $C(B)$. Ist z. B. $B := [0, 1]$ und $z(t) := e^t$, so ist die beste Tschebyscheff-Approximierende bezüglich Π_2, der Polynome vom Grad ≤ 2, gegeben durch

$$x^*(t) \approx 1.008757 + 0.854740\, t + 0.846029\, t^2$$

und Minimalabstand ist $\approx 8.75 \cdot 10^{-3}$. Approximiert man bezüglich $R_{1,1}$, der Menge der rationalen Funktionen mit Zähler- und Nennergrad ≤ 1, so ist die beste Approximierende gegeben durch

$$x^*(t) \approx \frac{0.995705 + 0.668203\, t}{1 - 0.388848\, t}$$

und der Minimalabstand ist $\approx 4.37 \cdot 10^{-3}$, also nur etwa halb so groß wie im linearen Fall (siehe G. Meinardus (1967, S. 167)). In beiden Fällen hat man drei Parameter zu speichern (der konstante Term im Nenner der rationalen Funktion sei auf 1 normiert). Im ersten Fall benötigt man zur Auswertung zwei Additionen und zwei Multiplikaten (man verwende das Horner-Schema!), im zweiten braucht man zwei Additionen, zwei Multiplikationen und noch eine Division. In Abbildung 1.2 stellen wir den Defekt bei linearer und bei rationaler Approximation der Exponentialfunktion dar. Eine ra-

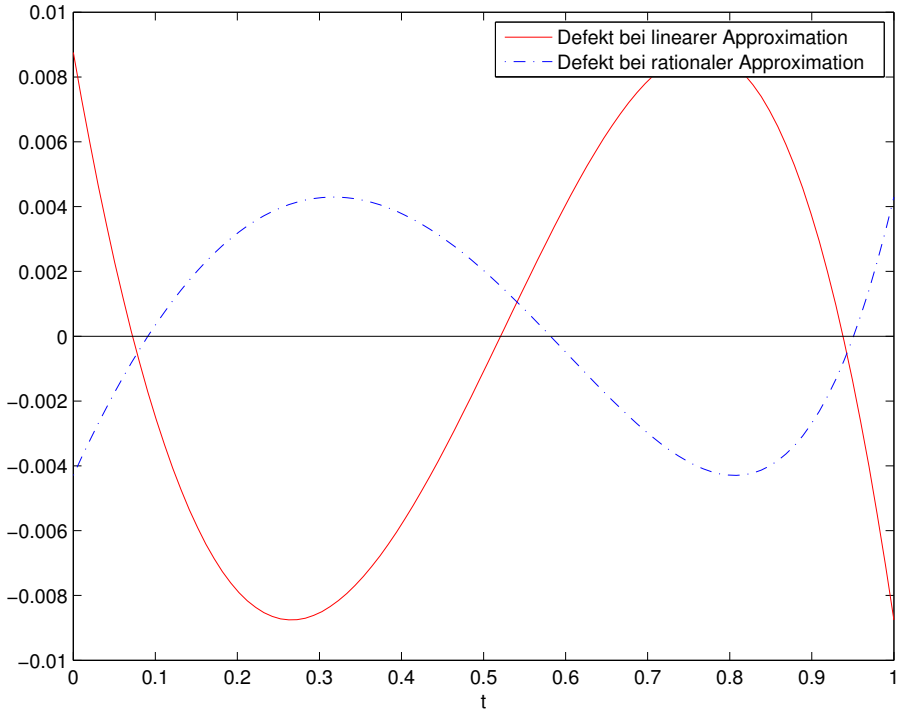

Abbildung 1.2: Defekt bei linearer bzw. rationaler Approximation von $z(t) := \exp(t)$

tionale Approximation ist u. a. dann sinnvoll, wenn die zu approximierende Funktion z außerhalb von B eine Singularität besitzt (die sozusagen in B "hereinstrahlt") und

die offenbar durch eine rationale Approximation besser als durch eine lineare erfasst werden kann. □

Beispiel 1.2.5 (Approximation mit Exponentialsummen) In Beispiel 1.2.2 sahen wir, dass es bei Abkling- oder Zerfallserscheinungen sinnvoll sein kann, eine gegebene Funktion $z \in C[\alpha, \beta]$ durch Elemente aus

$$E_r^0 := \left\{ \sum_{j=1}^{r} a_j e^{-b_j t} : a_j, b_j \in \mathbb{R}\,(j = 1, \ldots, r) \right\}$$

zu approximieren. Dies kann bezüglich verschiedener Normen geschehen, also etwa im Tschebyscheffschen Sinne bezüglich der Maximumnorm $\|\cdot\|_\infty$ oder auch im L_2-Sinne, also bezüglich der Norm

$$\|x\|_2 := \left(\int_\alpha^\beta x(t)^2 \, dt \right)^{1/2}.$$

Allerdings braucht nicht zu jedem $z \in C[\alpha, \beta]$ eine beste Approximierende aus E_r^0 zu existieren. Hierzu geben wir ein Beispiel an (siehe G. Meinardus (1967, S. 178)). Sei $z \in C[0, 1]$ definiert durch $z(t) := te^t$. Wir zeigen, dass es eine Folge $\{x_k\} \subset E_2^0$ gibt, die gleichmäßig auf $[0, 1]$ gegen z konvergiert, für die also $\lim_{k\to\infty} \|x_k - z\|_\infty = 0$. Da $z \notin E_2^0$ ist gezeigt, dass es keine beste Approximierende an z in E_2^0 gibt. Hierzu definiere man $x_k \in E_2^0$ durch

$$x_k(t) := -ke^t + ke^{(1+1/k)t}, \qquad k = 1, 2, \ldots.$$

Für $t \in [0, 1]$ ist

$$\begin{aligned} 0 &\leq x_k(t) - z(t) \\ &= e^t(ke^{t/k} - k - t) \\ &= e^t k \sum_{j=2}^{\infty} \frac{t^j}{k^j j!} \\ &\leq \frac{1}{k} e \sum_{j=2}^{\infty} \frac{1}{j!} \\ &= \frac{1}{k} e(e - 2). \end{aligned}$$

Hieraus folgt $\lim_{k\to\infty} \|x_k - z\|_\infty = 0$ und damit die Nichtexistenz einer besten Approximierenden an z in E_2^0. Daher ersetzt man E_r^0 durch "verallgemeinerte" Exponentialsummen

$$E_r := \left\{ \sum_{j=1}^{s} p_j(t) e^{b_j t} : p_j \in \Pi_{m_j} \text{ mit } \sum_{j=1}^{s} (m_j + 1) \leq r \right\}.$$

Hierin wird schließlich die Existenz einer besten Approximierenden nachgewiesen werden können.

Bei Anwendungen kann es sinnvoll Oder sogar notwendig sein, die Menge M, mit der approximiert wird, durch weitere Nebenbedingungen einzuschränken. Dies können Interpolationsbedingungen, Vorzeichenbedingungen an die Koeffizienten oder ähnliches sein. □

1.3 Übersicht

Eine gewisse Einteilung der Vorlesung ist schon durch die Angabe der Problemstellungen in Abschnitt 1.1 vorgegeben, wobei wir uns allerdings nicht genau an die angegebene Reihenfolge halten werden. Das didaktische Konzept besteht (nach einigem Schwanken) darin, einen funktionalanalytischen Zugang an den Anfang zu stellen und zumindest Existenz, Eindeutigkeit und notwendige Optimalitätsbedingungen in einem allgemeinen Rahmen zu behandeln, wobei natürlich möglichst viele konkrete Beispiele die theoretischen Ergebnisse anreichern sollen. Der Nachteil bei diesem Zugang besteht darin, dass wir schon am Anfang der Vorlesung einige funktionalanalytische Hilfsmittel ohne Beweis bringen müssen, da wir es nicht für sinnvoll halten, etwa den Trennungssatz für konvexe Mengen, den Satz von Hahn-Banach und ähnliches in dieser Vorlesung noch einmal zu beweisen. Ein "alternativer" Zugang wäre, eine Einteilung nach Problemklassen vorzunehmen, also etwa zunächst über lineare Tschebyscheff-Approximation, dann über L_2-Approximation, rationale Approximation usw. zu sprechen.

Das zu der Vorlesung im WS 1984/85 angefertigte handschriftliche Manuskript wurde von mir seit September 2014 zu einem LaTeXManuskript verarbeitet, wobei möglichst viel aus dem fast 30 Jahre alten Manuskript übernommen werden soll. Insbesondere werden wir uns auch in dieser Ausarbeitung einer verhältnismäßig alten Vorlesung auf die *univariate Approximation*, also die Approximation von Funktionen *einer* Veränderlichen beschränken.

Kapitel 2

Steilkurs: Lineare Funktionalanalysis

Wir können für diese Vorlesung leider kaum Vorkenntnisse aus der Funktionalanalysis voraussetzen und bringen daher in den nächsten Abschnitten die für die Approximationstheorie wichtigsten funktionalanalytischen Hilfsmittel, zum großen Teil natürlich ohne Beweis.

2.1 Lineare normierte Räume, Hilberträume

In dem Raum, in dem "sich alles abspielt" benötigt man, um Approximationsaufgaben formulieren und behandeln zu können, eine Abstandsbegriff. Metrische Räume sind uns zu allgemein, sie haben i. Allg. keine lineare Struktur, daher sind lineare normierte Räume und Prä-Hilbert- bzw. Hilberträume der angemessene Rahmen.

Definition 2.1.1 (a) Ein Paar $(X, \|\cdot\|)$ heißt ein (reeller) *linearer normierter Raum*, falls X ein linearer Raum (über $\mathbb{R}$) ist[1] und $\|\cdot\|\colon X \longrightarrow \mathbb{R}$ eine Abbildung (*Norm*) ist mit

1. $\|x\| \geq 0$ für alle $x \in X$ und $\|x\| = 0$ genau dann wenn $x = 0$ (Definitheit).
2. $\|\alpha x\| = |\alpha| \, \|x\|$ für alle $\alpha \in \mathbb{R}$, $x \in X$ (Homogenität).
3. $\|x + y\| \leq \|x\| + \|y\|$ für alle $x, y \in X$ (Dreiecksungleichung).

(b) Ein Paar $(X, (\cdot, \cdot))$ heißt ein (reeller) *Prä-Hilbertraum*, falls X ein linearer Raum (über $\mathbb{R}$) ist und $(\cdot, \cdot)\colon X \times X \longrightarrow \mathbb{R}$ eine Abbildung (*inneres Produkt*) ist mit

1. $(x, x) \geq 0$ für alle $x \in X$ und $(x, x) = 0$ genau dann wenn $x = 0$.
2. $(x, y) = (y, x)$ für alle $x, y \in X$.
3. $(\alpha x, y) = \alpha(x, y)$ für alle $\alpha \in \mathbb{R}$, $x, y \in X$.
4. $(x + y, z) = (x, z) + (y, z)$ für alle $x, y, z \in X$.

[1] Das Nullelement in X wird mit 0 bezeichnet. Eine Verwechslung mit der skalaren Null kann nur durch Böswilligkeit eintreten.

Bemerkung 2.1.2 Ist $(X, (\cdot,\cdot))$ ein Prä-Hilbertraum, so ist auf X durch

$$\|x\| := (x,x)^{1/2}$$

eine Norm definiert, $(X, (\cdot,\cdot))$ wird als in *kanonischer Weise* zu einem linearen normierten Raum. Die ersten beiden Eigenschaften einer Norm sind trivialerweise erfüllt. Um die Dreiecksungleichung einzusehen, beweist man zunächst die *Cauchy-Schwarzsche Ungleichung*:

- $|(x,y)| \leq \|x\|\,\|y\|$ für alle $x, y \in X$.

Denn: Seien $x, y \in X$ vorgegeben. O. B. d. A. ist $y \neq 0$. Dann ist

$$\begin{aligned} 0 \quad &\underbrace{\leq}_{1.} \quad \left(x - \frac{(x,y)}{(y,y)}y, x - \frac{(x,y)}{(y,y)}y\right) \\ &\underbrace{=}_{2.,3.,4.} \quad (x,x) - 2\frac{(x,y)^2}{(y,y)} + \frac{(x,y)^2}{(y,y)^2}(y,y) \\ &= \quad (x,x) - \frac{(x,y)^2}{(y,y)}, \end{aligned}$$

woraus die Behauptung folgt. Für beliebiges $x, y \in X$ ist daher

$$\begin{aligned} \|x+y\|^2 &= (x+y, x+y) \\ &= (x,x) + 2(x,y) + (y,y) \\ &= \|x\|^2 + 2(x,y) + \|y\|^2 \\ &\leq \|x\|^2 + 2\,\|x\|\,\|y\| + \|y\|^2 \\ &= (\|x\| + \|y\|)^2, \end{aligned}$$

woraus die Dreiecksungleichung folgt.

Ist $(X, \|\cdot\|)$ ein linearer normierter Raum, so gibt es genau dann ein inneres Produkt $(\cdot,\cdot)$ auf $X \times X$ mit $\|x\|^2 = (x,x)$ für alle $x \in X$, wenn die sogenannte *Parallelogrammgleichung*

$$\|x+y\|^2 + \|x-y\|^2 = 2(\|x\|^2 + \|y\|^2) \quad \text{für alle } x, y \in X$$

gilt. Diese veranschaulichen wir uns in Abbildung 2.1. Dass die Parallelogrammgleichung in einem Prähilbertraum gilt, weist man durch einfaches Nachrechnen nach. Gilt umgekehrt die Parallelogrammgleichung, so definiere man $(\cdot,\cdot)\colon X \times X \longrightarrow \mathbb{R}$ durch

$$(x,y) := \frac{1}{4}(\|x+y\|^2 - \|x-y\|^2)$$

und weise nach, dass hierdurch ein inneres Produkt auf X definiert ist. □

Beispiele: Die wichtigsten Normen im $\mathbb{R}^n$ sind die *Maximumnorm*

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

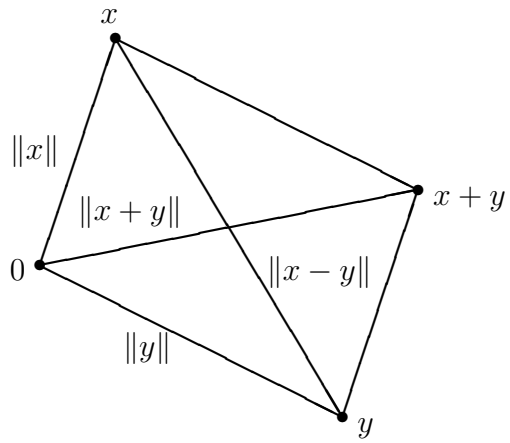

Abbildung 2.1: Die Parallelogrammgleichung

die *euklidische Norm*

$$\|x\|_2 := \left(\sum_{j=1}^{n} x_j^2\right)^{/2}$$

(diese wird durch das innere Produkt $(x, y) := x^T y$ erzeugt, mit diesem inneren Produkt ist der $\mathbb{R}^n$ also ein Prä-Hilbertraum), sowie die *Betragssummennorm* oder L_1*-Norm*

$$\|x\|_1 := \sum_{j=1}^{n} |x_j|.$$

Allgemeiner kann man für $1 \leq p < \infty$ die L_p*-Norm* durch

$$\|x\|_p := \left(\sum_{j=1}^{n} |x_j|^p\right)^{1/p}$$

definieren. Um einzusehen, dass dies in der Tat eine Norm ist, hat man lediglich noch die Dreiecksungleichung nachzuweisen. O. B. d. A. ist $p > 1$.

(a) Seien $a, b \geq 0$ und $p, q \in (1, \infty)$ mit $1/p + 1/q = 1$ gegeben. Dann ist

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q} \qquad \text{(Youngsche Ungleichung)}.$$

Denn: O. B. d. A. sind $a, b > 0$. Die Exponentialfunktion ist auf $\mathbb{R}$ strikt konvex, da ihre zweite Ableitung auf ganz $\mathbb{R}$ positiv ist. Wegen $1/p \in (0, 1)$ und $1/p+1/q = 1$ ist daher

$$\begin{aligned} ab &= e^{\ln(a)} e^{\ln(b)} \\ &= e^{(1/p)\ln(a^p)+(1/q)\ln(b^q)} \\ &\leq \frac{1}{p} e^{\ln(a^p)} + \frac{1}{q} e^{\ln(b^q)} \\ &= \frac{a^p}{p} + \frac{b^q}{q}. \end{aligned}$$

(b) Seien $x, y \in \mathbb{R}^n$ sowie $p, q \in (1, \infty)$ mit $1/p + 1/q = 1$ gegeben. Dann ist

$$\sum_{j=1}^{n} |x_j y_j| \le \|x\|_p \, \|y\|_q \qquad \text{(Höldersche Ungleichung)}.$$

Denn: O. B. d. A. ist $x, y \neq 0$. Man setze

$$a_j := \frac{|x_j|}{\|x\|_p}, \quad b_j := \frac{|y_j|}{\|y\|_q}, \qquad j = 1, \ldots, n.$$

Eine Anwendung von (a) und Summation liefert die Behauptung.

(c) Seien $x, y \in \mathbb{R}^n$ und $p \in (1, \infty)$. Dann ist

$$\|x + y\|_p \le \|x\|_p + \|y\|_p \qquad \text{(Minkowskische Ungleichung))}.$$

Denn: Sei $q := p/(p-1)$. Dann ist

$$\begin{aligned}
\|x+y\|_p^p &= \sum_{j=1}^{n} |x_j + y_j|^p \\
&= \sum_{j=1}^{n} |x_j + y_j| \, |x_j + y_j|^{p-1} \\
&\le \sum_{j=1}^{n} |x_j| \, |x_j + y_j|^{p-1} + \sum_{j=1}^{n} |y_j| \, |x_j + y_j|^{p-1} \\
&\le (\|x\|_p + \|y\|_p) \left(\sum_{j=1}^{n} |x_j + y_j|^{(p-1)q} \right)^{1/q} \\
&\qquad \text{(zweimalige Anwendung der Hölderschen Ungleichung)} \\
&= (\|x\|_p + \|y\|_p) \left(\sum_{j=1}^{n} |x_j + y_j|^p \right)^{1-1/p} \\
&= (\|x\|_p + \|y\|_p) \|x+y\|_p^{p(1-1/p)} \\
&= (\|x\|_p + \|y\|_p) \frac{\|x+y\|_p^p}{\|x+y\|_p},
\end{aligned}$$

woraus die Behauptung folgt.

$X := C[\alpha, \beta]$ sei der lineare Raum der auf dem kompakten Intervall $[\alpha, \beta] \subset \mathbb{R}$ reellwertigen stetigen Funktionen. Die für die Approximationstheorie und viele Anwendungen wichtigste Norm ist die *Maximum- oder Tschebyscheff-Norm*, definiert durch

$$\|x\|_\infty := \max_{t \in [\alpha, \beta]} |x(t)|.$$

Um kontinuierliche und diskrete Approximation gleichzeitig zu erfassen, kann es zweckmäßig sein, zu verallgemeinern und von einer kompakten Menge $B \subset \mathbb{R}^N$ auszugehen,

den linearen Raum $C(B)$ der auf B definierten reellwertigen stetigen Funktionen zu betrachten und diesen zu einem linearen normierten Raum zu machen, indem man als Norm

$$\|x\|_\infty := \max_{t\in B} |x(t)|$$

definiert. Hierbei schreiben wir mit gutem Gewissen max statt sup, da eine stetige Funktion, hier $|x(\cdot)|$, auf einer kompakten Menge, hier B, ihre Extrema, hier Maximum, annimmt. Auch auf $C[\alpha, \beta]$ kann man für $p \in [1, \infty)$ eine Lp-Norm definieren durch

$$\|x\|_p := \left(\int_\alpha^\beta |x(t)|^p \, dt \right)^{1/p}.$$

Hierbei sind vor allem die Fälle $p = 1, 2$ von Interesse.

Auf weitere wichtige Beispiele von linearen normierten Räumen, vor allem sogenannten Sobolev-Räumen, werden wir erst später bei Anwendungen eingehen. □

Im Weiteren benutzen wir die folgende Bezeichnung. Ist $(X, \|\cdot\|)$ ein linearer normierter Raum, $x \in X$ und $r > 0$, so sei

$$B[x; r] := \{y \in X : \|y - x\| \leq r\}$$

die "abgeschlossene" Kugel um x mit dem Radius r und

$$B(x; r) := \{y \in X : \|y - x\| < r\}$$

die "offene" Kugel um x mit dem Radius r. In Abbildung 2.2 veranschaulichen wir uns Einheitskugeln (Kugeln mit dem Mttelpunkt 0 und dem Radius 1) bezüglich verschiedener Normen im $\mathbb{R}^n$. In Abbildung 2.3 ist ein $x \in C[0, 2\pi]$ gestrichelt vorgegeben.

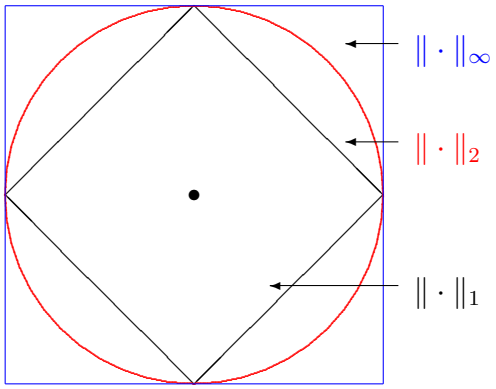

Abbildung 2.2: Einheitskugeln im $\mathbb{R}^2$ bezüglich $\|\cdot\|_1$, $\|\cdot\|_2$ und $\|\cdot\|_\infty$

Jedes $y \in C[0, 2\pi]$ mit $x(t) - 1 \leq y(t) \leq x(t) + 1$ für alle $t \in [0, 2\pi]$ bzw. aus dem angegebenen "Schlauch" liegt in der (abgeschlossenen) Kugel um x mit dem Radius 1.

Nun werden kurz verschiedene wichtige Begriffe der Funktionalanalysis angegeben. Stets ist $(X, \|\cdot\|)$ ein linearer normierter Raum.

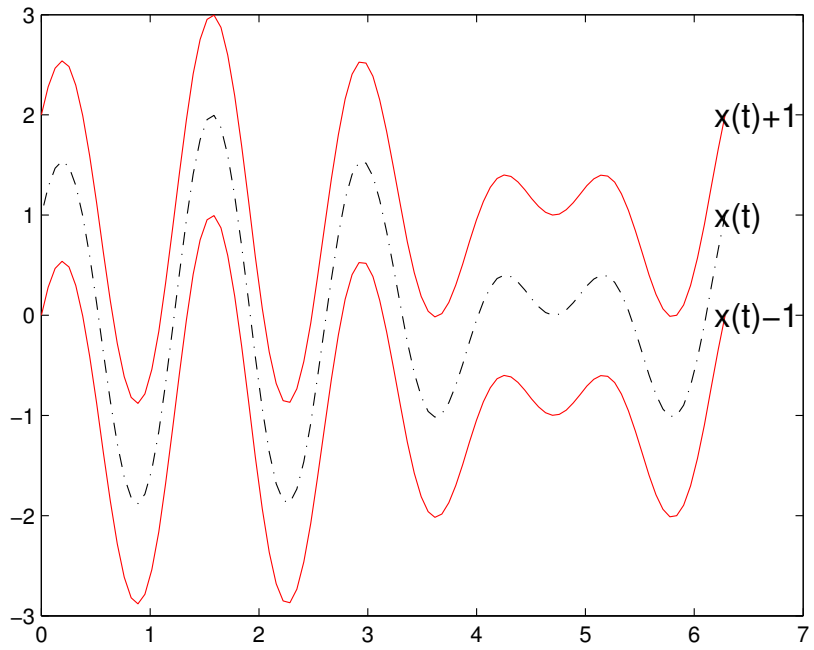

Abbildung 2.3: Die Kugel $B[x;1]$ bezüglich der Maximumnorm

1. Sei $A \subset X$, Dann heißt

$$\text{int}\,(A) := \{a \in A : \text{Es existiert ein } \epsilon > 0 \text{ mit } B[a;\epsilon] \subset A\}$$

das *Innere* von A. A heißt *offen*, wenn $A = \text{int}\,(A)$. Z. B. ist mit einem $x \in X$ und $r > 0$ die Kugel $B(x;r)$ offen (Beweis?)

2. Die Norm $\|\cdot\|$ definiert einen *Konvergenzbegriff* auf X. Eine Folge $\{x_k\} \subset X$ *konvergiert* gegen ein $x \in X$, was wir mit $x_k \to x$ oder $\lim_{k\to\infty} x_k = x$ bezeichnen, falls $\lim_{k\to\infty} \|x_k - x\| = 0$. Es gelten die üblichen Rechenregeln.

3. Sei $A \subset X$. Dann heißt

$$\text{cl}\,(A) := \{a \in X : \text{Es existiert } \{a_k\} \subset A \text{ mit } a_k \to a\}$$

der *Abschluss* von A. Offensichtlich ist

$$\text{cl}\,(A) = \{a \in X : B[a;\epsilon] \cap A \neq \emptyset \text{ für alle } \epsilon > 0\}.$$

A heißt *abgeschlossen*, wenn $A = \text{cl}\,(A)$. Z. B. ist mit einem $x \in X$ und $r > 0$ die Kugel $B[x;r]$ abgeschlossen. Die Menge A heißt *dicht* in X, falls $\text{cl}\,(A) = X$.

4. Seien $(X_1, \|\cdot\|_1)$ und $(X_2, \|\cdot\|_2)$ zwei lineare normierte Räume, $T\colon D \subset X_1 \longrightarrow X_2$ eine Abbildung mit dem Definitionsbereich $D \subset X_1$.

(a) T heißt ein *linearer Operator* (oder Abbildung, Transformation), falls $D \subset X_1$ ein linearer Teilraum ist und $T(\alpha x + \beta y) = \alpha T(x) + \beta T(y)$ für alle $\alpha, \beta \in \mathbb{R}$ und alle $x, y \in D$. Bei einem linearen Operator schreibt man häufig Tx statt $T(x)$.

(b) T heißt *stetig in* $\hat{x} \in D$,

$$\{x_k\} \subset D, \quad x_k \to \hat{x}, \Longrightarrow T(x_k) \to T(\hat{x})$$

bzw. es zu jedem $\epsilon > 0$ ein $\delta = \delta(\epsilon) > 0$ mit

$$x \in B[\hat{x}; \delta] \cap D \Longrightarrow T(x) \in B[T(\hat{x}); \epsilon]$$

gibt. T heißt *stetig auf* D, wenn T in jedem $\hat{x} \in D$ stetig ist.

(c) Mit $L(X_1, X_2)$ wird die Menge der linearen stetigen Operatoren von X_1 nach X_2 bezeichnet. Also ist $L(X, \mathbb{R})$ die Menge der linearen stetigen Abbildungen (auch *Funktionale* genannt) des linearen normierten Raumes X nach $\mathbb{R}$. Wir schreiben X^* statt $L(X, \mathbb{R})$ und nennen dies den *Dualraum* von X.

Bemerkungen: 1. Ist $X := C[\alpha, \beta]$ und $\|\cdot\| := \|\cdot\|_\infty$, so konvergiert eine Folge $\{x_k\} \subset X$ genau dann (der Maximumnorm nach) gegen ein $x \in X$, wenn die Folge $\{x_k\}$ *gleichmäßig auf* $[\alpha, \beta]$ gegen x konvergiert.

2. Für jeden endlich dimensionalen linearen Raum, speziell den $\mathbb{R}^n$, gilt: Konvergiert eine Folge $\{x_k\} \subset X$ bezüglich einer Norm gegen ein $x \in X$, so auch bezüglich jeder anderen Norm. Dies liegt daran, dass alle Normen auf einem endlichdimensionalen linearen Raum *äquivalent* sind, was heißen soll, dass zu je zwei Normen $\|\cdot\|_a$ und $\|\cdot\|_b$ auf X Konstanten $c, C > 0$ existieren mit $c\,\|x\|_a \le \|x\|_b \le C\,\|x\|_a$ für alle $x \in X$ existieren. Z. B. ist $\|x\|_\infty \le \|x\|_2 \le \sqrt{n}\,\|x\|_\infty$ für alle $x \in \mathbb{R}^n$.

3. Ist $L \subset X$ ein linearer Teilraum, so auch $\mathrm{cl}\,(L)$. in Kürze werden wir einsehen, dass endlichdimensionale lineare Teilräume eines linearen normierten Raumes abgeschlossen sind. Das ist für beliebige lineare Teilräume nicht richtig. Der lineare Teilraum Π der Polynome (in einer Variablen) liegt nach dem *Weierstraßschen Approximationssatz* (auf ihn kommen wir später zurück) dicht in $(C[\alpha, \beta], \cdot\|_\infty)$, es ist aber $\Pi \neq C[\alpha, \beta]$.

4. Die Normabbildung $\|\cdot\|\colon X \longrightarrow \mathbb{R}$ ist stetig, da $|\|x\| - \|y\|| \le \|x - y\|$ für alle $x, y \in X$.

5. Seien $(X_1, \|\cdot\|_1)$ und $(X_2, \|\cdot\|_2)$ lineare normierte und $T\colon X_1 \longrightarrow X_2$ ein linearer Operator. Dann ist $T \in L(X_1, X_2)$ genau dann, wenn

$$\|T\| := \sup_{x \neq 0} \frac{\|T(x)\|}{\|x\|} < +\infty.$$

Denn: Ist $T \in L(X_1, X_2)$, so existiert zu $\epsilon := 1$ ein $\delta = \delta(1) > 0$ mit

$$\|x\|_1 \le \delta \Longrightarrow \|T(x)\|_2 \le 1.$$

Für jedes $x \neq 0$ ist also $\|T(\delta\, x/\|x\|_1)\|_2 \le 1$ und daher $\sup_{x \neq 0} \|T(x)\|_2/\|x\|_1 \le 1/\delta$. Ist umgekehrt $\|T\| < +\infty$, so ist $\|T(x)\|_2 \le \|T\|\,\|x\|_1$ für alle $x \in X_1$. Daher ist T

in 0 stetig, wegen der Linearität von T auch auf ganz X_1. Die Menge $L(X_1, X_2)$ der stetigen, linearen Abbildungen von X_1 nach X_2 ist in kanonischer Weise ein linearer Raum. Dieser lineare Raum wird zu einem linearen normierten Raum, indem man $\|\cdot\|\colon L(X_1, X_2) \longrightarrow \mathbb{R}$ für $T \in L(X_1, X_2)$ durch

$$\|T\| := \sup_{x \neq 0} \frac{\|T(x)\|_2}{\|x\|_1}$$

definiert (Beweis?).

6. Jede lineare Abbildung zwischen endlichdimensionalen linearen normierten Räumen ist (automatisch) stetig. Zu jedem $l \in (\mathbb{R}^n)^*$ existiert genau ein $y \in \mathbb{R}^n$ mit $l(x) = y^T x$ für alle $x \in \mathbb{R}^n$. Definiert man umgekehrt für ein gegebenes $y \in \mathbb{R}^n$ durch $l(x) := y^T x$ eine Abbildung $l\colon \mathbb{R}^n \longrightarrow \mathbb{R}$, so ist $l \in (\mathbb{R}^n)^*$. Die Abbildung $I\colon (\mathbb{R}^n)^* \longrightarrow \mathbb{R}^n$ definiert durch $I(l) := y$ ist eine bijektive lineare Abbildung zwischen $(\mathbb{R}^n)^*$ und $\mathbb{R}^n$. Daher können $(\mathbb{R}^* n)^*$ und $\mathbb{R}^n$ identifiziert werden.

7. Sei $X := C[\alpha, \beta]$ und $\|\cdot\| := \|\cdot\|_\infty$. Man definiere $l\colon C[\alpha, \beta] \longrightarrow \mathbb{R}$ durch

$$l(x) := \int_\alpha^\beta x(t)\, dt.$$

Dann ist $l \in (C[\alpha, \beta])^*$, denn offenbar ist l linear und

$$|l(x)| \leq \int_\alpha^\beta |x(t)|\, dt \leq (\beta - \alpha)\, \|x\|_\infty$$

und daher

$$\|l\| = \sup_{x \neq 0} \frac{|l(x)|}{\|x\|_\infty} \leq (\beta - \alpha),$$

also l stetig. □

Definition 2.1.3 Sei $(X, \|\cdot\|)$ ein linearer normierter Raum.

1. Eine Folge $\{x_k\} \subset X$ heißt eine *Cauchy-Folge*, falls es zu jedem $\epsilon > 0$ ein $K(\epsilon) \in \mathbb{N}$ mit $\|x_k - x_l\| \leq \epsilon$ für alle $k, l \geq K(\epsilon)$ gibt.

2. Eine Teilmenge $A \subset X$ heißt *vollständig*, wenn zu jeder Cauchy-Folge $\{x_k\} \subset A$ ein $x \in A$ mit $x_k \to x$ existiert.

3. Ist ein linearer normierter Raum vollständig, so heißt er ein *Banachraum*.

4. Ist ein Prä-Hilbertraum $(X, (\cdot, \cdot))$ vollständig, so heißt er ein *Hilbertraum*.

Bemerkungen: 1. Jede konvergente Folge in einem linearen normierten Raum ist eine Cauchy-Folge. Daher ist jede vollständige Teilmenge eines linearen normierten Raumes abgeschlossen.

2. Jeder endlichdimensionale lineare Teilraum eines linearen normierten Raumes ist vollständig und daher auch abgeschlossen. Zur Begründung geben wir uns einen endlichdimensionalen linearen Teilraum $V = \operatorname{span}\{v_1, \ldots, v_n\} \subset X$ mit linear unabhängigen $v_1, \ldots, v_n$ vor. Sei $\{x_k\} \subset V$ eine Cauchy-Folge. Jedes $x_k \in V$ besitzt eine eindeutige Darstellung $x_k = \sum_{j=1}^n \alpha_{kj} v_j$. Wir zeigen, dass auch $\{a_k\} \subset \mathbb{R}^n$ mit

$a_k := (\alpha_{k1}, \ldots, \alpha_{kn})^T$ Cauchy-Folgen und daher wegen des Cauchyschen Konvergenzkriteriums konvergent sind. Hierzu definiere man

$$c := \min_{a=(\alpha_j)\in\mathbb{R}^n : \|a\|_2=1} \left\| \sum_{j=1}^{n} \alpha_j v_j \right\|,$$

wobei wir ausnutzen, dass das Minimum einer stetigen Funktion auf einer Kompakten Menge angenommen wird. Offenbar ist $c > 0$. Wegen $\|x_k - x_l\| \geq c \|a_k - a_l\|_2$ ist auch $\{a_k\}$ eine Cauchy-Folge und folglich konvergent gegen ein $a = (\alpha_j) \in \mathbb{R}^n$. Daher konvergiert die Folge $\{x_k\} \subset V$ gegen $x := \sum_{j=1}^{n} \alpha_j v_j$, V ist also vollständig. □

Beispiele: 1. Der $\mathbb{R}^n$ ist bezüglich jeder Norm vollständig.

2. Der lineare normierte Raum $(C[\alpha, \beta], \|\cdot\|_\infty)$ ist ein Banachraum. Denn sei $\{x_k\} \subset C[\alpha, \beta]$ eine Cauchy-Folge. Für ein beliebiges $t \in [\alpha, \beta]$ ist $|x_k(t) - x_l(t)| \leq \|x_k - x_l\|_\infty$. Daher ist $\{x_k(t)\} \subset \mathbb{R}$ für jedes $t \in [\alpha, \beta]$ eine Cauchy-Folge. Da $\mathbb{R}$ vollständig ist, ist $\{x_k(t)\}$ konvergent. Sei $x(t) := \lim_{k\to\infty} x_k(t)$. Wir zeigen, dass $\{x_k\}$ gleichmäsig auf $[\alpha, \beta]$ gegen x konvergiert. Hierzu sei $\epsilon > 0$ vorgegeben, $t \in [\alpha, \beta]$ beliebig. Da $\{x_k\} \subset C[\alpha, \beta]$ eine Cauchy-Folge ist, existiert ein $K(\epsilon) \in \mathbb{N}$ mit

$$|x_k(t) - x_l(t)| \leq \|x_k - x_l\| \leq \epsilon \quad \text{für alle } k, l \geq K(\epsilon).$$

Mit $l \to \infty$ folgt wegen $x_l(t) - x(t)$, dass

$$|x_k(t) - x(t)| \leq \epsilon \quad \text{für alle } k \geq K(\epsilon) \text{ und alle } t \in [\alpha, \beta].$$

Dies bedeutet, dass die Folge $\{x_k\}$ gleichmäßig gegen x konvergiert. Der gleichmäßige Limes einer Folge stetiger Funktionen ist bekanntlich stetig, der Beweis erfolgt mit einem $\epsilon/3$-Argument. Damit ist die Behauptung bewiesen.

3. Man betrachte den linearen normierten Raum $(C[\alpha, \beta], \|\cdot\|_1)$ mit der Norm

$$\|x\|_1 := \int_\alpha^\beta |x(t)|\, dt.$$

Dann ist $(C[\alpha, \beta], \|\cdot\|_1)$ nicht vollständig! O. B. d. A. ist $[\alpha, \beta] = [0, 1]$. Man betrachte die Folge $\{x_k\}$, die für $k \geq 2$ mit $a_k := \frac{1}{2} + \frac{1}{k}$ durch

$$x_k(t) := \begin{cases} 0, & 0 \leq t \leq \frac{1}{2}, \\ k(t - \frac{1}{2}), & \frac{1}{2} \leq t \leq a_k, \\ 1, & a_k \leq t \leq 1 \end{cases}$$

definiert ist, siehe Abbildung 2.4. Wir zeigen, dass die Folge $\{x_k\}$ eine Cauchy-Folge in $(C[0, 1], \|\cdot\|_1)$ ist. Hierzu sei $k \geq l$. Dann ist $a_k \leq a_l$ und $x_k(t) \geq x_l(t)$ für alle $t \in [0, 1]$

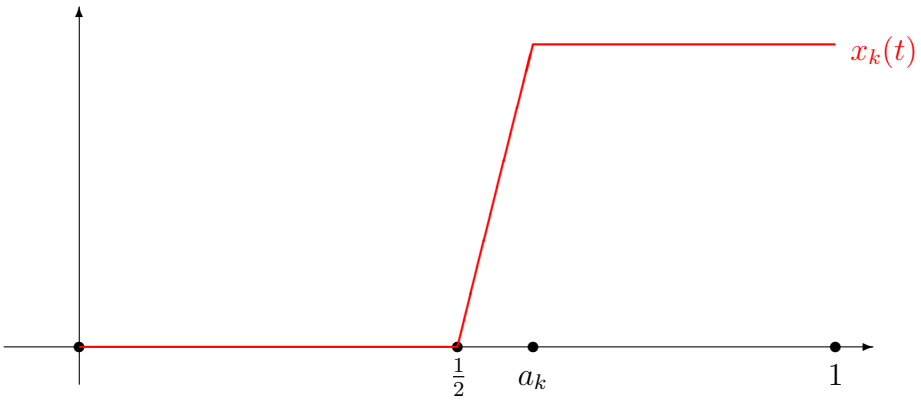

Abbildung 2.4: Eine Cauchy-Folge in $(C[0,1], \|\cdot\|_1)$

und daher

$$\begin{aligned}
\|x_k - x_l\|_1 &= \int_0^1 (x_k(t) - x_l(t))\,dt \\
&= \int_{\frac{1}{2}}^{a_k} (x_k(t) - x_l(t))\,dt + \int_{a_k}^{a_l} (x_k(t) - x_l(t))\,dt \\
&= \int_{\frac{1}{2}}^{a_k} (k-l)\left(t - \frac{1}{2}\right) dt + \int_{a_k}^{a_l} \left[1 - l\left(t - \frac{1}{2}\right)\right] dt \\
&= (k-l)\frac{1}{2}\left(t - \frac{1}{2}\right)^2 \Big|_{\frac{1}{2}}^{a_k} + (a_l - a_k) - l\frac{1}{2}\left(t - \frac{1}{2}\right)^2 \Big|_{a_k}^{a_l} \\
&= (k-l)\frac{1}{2k^2} + \frac{1}{l} - \frac{1}{k} - \frac{l}{2}\left(\frac{1}{l^2} - \frac{1}{k^2}\right) \\
&= \frac{1}{2}\left(\frac{1}{l} - \frac{1}{k}\right) \\
&\leq \frac{1}{2l}.
\end{aligned}$$

Hieraus liest man ab, dass $\{x_k\}$eine Cauchy-Folge in $(C[0,1], \|\cdot\|_1)$ ist. Angenommen, es existiert ein $x \in C[0,1]$ mit $\|x_k - x\|_1 \to 0$. Wegen

$$\begin{aligned}
\|x_k - x\|_1 &= \int_0^{\frac{1}{2}} |x(t)|\,dt + \int_{\frac{1}{2}} a_k |x_k(t) - x(t)|\,dt + \int_{a_k}^1 |1 - x(t)]\,dt \\
&\to \int_0^{\frac{1}{2}} |x(t)|\,dt + \int_{\frac{1}{2}}^1 |1 - x(t)|\,dt.
\end{aligned}$$

Daher ist

$$\int_0^{\frac{1}{2}} |x(t)|\,dt + \int_{\frac{1}{2}}^1 |1 - x(t)|\,dt = 0,$$

so dass x notwendigerweise die Gestalt

$$x(t) = \begin{cases} 0, & t \in [0, \frac{1}{2}), \\ 1, & t \in (\frac{1}{2}, 1] \end{cases}$$

besitzt, was ein Widerspruch zur Stetigkeit von x ist.
4. Entsprechend sind $(C[\alpha, \beta], \|\cdot\|_p)$, $1 \leq p \leq \infty$, mit

$$\|x\|_p := \left(\int_\alpha^\beta |x(t)|^p \, dt \right)^{1/p}$$

zwar lineare normierte Räume, aber keine Banachräume. Erst eine "Vervollständigung" von $C[\alpha, \beta]$ bezüglich $\|\cdot\|_p$ führt zu Banachräumen. Dies sind die Räume $L^p[\alpha, \beta]$ der auf $[\alpha, \beta]$ messbaren Funktionen x, für dir $|x|^p$ im Lebesguesche Sinne auf $[\alpha, \beta]$ integrierbar sind. Der Beweis hierfür kann noch nicht einmal angedeutet werden. Entsprechend ist $L^2[\alpha, \beta]$ mit dem inneren Produkt

$$(x, y) := \int_\alpha^\beta x(t)y(t) \, dt$$

ein Hilbertraum. □

2.2 Konvexe Mengen in linearen normierten Räumen

Konvexe Mengen (und Funktionen) spielen in der Optimierung und daher auch in der approximationstheorie eine sehr große Rolle. Die nötigen funktionalanalytischen Hilfsmittel werden in diesem Abschnitt bereitgestellt. Im folgenden sei $(X, \|\cdot\|)$ stets ein linearer normierter Raum.

Definition 2.2.1 Eine Menge $A \subset X$ heißt *konvex*, falls aus $x, y \in A$ und $\lambda \in [0, 1]$ folgt, dass $(1 - \lambda)x + \lambda y \in A$, wenn also mit zwei Punkten aus A auch die gesamte Verbindungsstrecke zu A gehört.

Beispiele: 1. Offene bzw. abgeschlossene Kugeln $B(x; r($ bzw. $B[x; r]$ sind konvex.
2. Sei $M \subset X$ konvex und $z \in X$. Gegeben sei die konvexe Approximationsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \|x - z\|, \quad x \in M.$$

Dann ist die sogenannte *metrische Projektion von z auf M*, nämlich die Menge

$$P_M(z) := \{x^* \in M : x^* \text{ ist beste Approximierende für } z \text{ in } M\}$$

konvex. □

Für das folgende Lemma geben wir nur einen kurzen Beweis an.

Lemma 2.2.2 *Sei $A \subset X$ konvex. Dann gilt:*

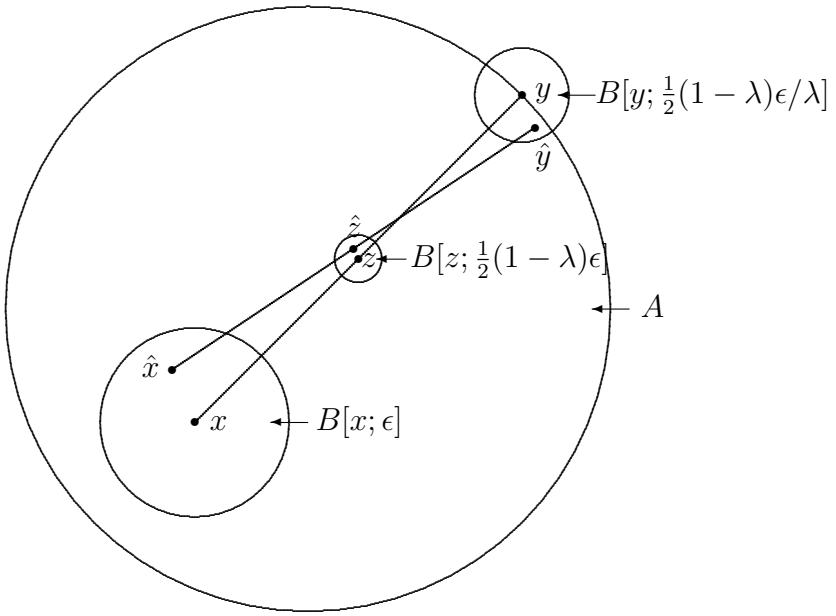

Abbildung 2.5: Veranschaulichung des Beweises von Lemma 2.2.2

1. *Sei* $x \in \operatorname{int}(A)$ *und* $y \in \operatorname{cl}(A)$. *Dann ist*

$$[x, y) := \{(1-\lambda)x + \lambda y : \lambda \in [0, 1)\} \subset \operatorname{int}(A).$$

2. $\operatorname{cl}(A)$ *ist konvex.*

3. $\operatorname{int}(A)$ *ist konvex.*

4. *Ist* $\operatorname{int}(A) \neq \emptyset$, *so ist* $\operatorname{cl}(\operatorname{int}(A)) = \operatorname{cl}(A)$.

Beweis: Wir zeigen nur den ersten Teil des Satzes. Die anderen Teile sind sehr einfach zu beweisen oder folgen hieraus. In Abbildung 2.5 veranschaulichen wir den Beweis. Wir geben uns ein $\lambda \in (0, 1)$ vor, setzen $z := (1-\lambda)x + \lambda y$ und zeigen $z \in \operatorname{int}(A)$. Wegen $x \in \operatorname{int}(A)$ existiert ein $\epsilon > 0$ mit $B[x; \epsilon] \subset A$. Wir haben zu zeigen, dass es um z eine ganz in A gelegene Kugel gibt. Genauer zeigen wir, dass $B[z; \frac{1}{2}(1-\lambda)\epsilon] \subset A$. Hierzu sei $\hat{z} \in B[z; \frac{1}{2}(1-\lambda)\epsilon]$ beliebig. Wegen $y \in \operatorname{cl}(A)$ existiert in jeder Kugel um y ein Element von A. Insbesondere gibt es ein $\hat{y} \in B[y; \frac{1}{2}(1-\lambda)\epsilon/\lambda] \cap A$. Nun definiere man

$$\hat{x} := \frac{1}{1-\lambda}\hat{z} - \frac{\lambda}{1-\lambda}\hat{y}.$$

Aus

$$z = (1-\lambda)x + \lambda y, \qquad \hat{z} = (1-\lambda)\hat{x} + \lambda\hat{y}$$

folgt

$$x - \hat{x} = \frac{1}{1-\lambda}(z - \hat{z}) - \frac{\lambda}{1-\lambda}(y - \hat{y})$$

und hieraus

$$\|x - \hat{x}\| \leq \frac{1}{1-\lambda}\|z_\lambda - \hat{z}\| + \frac{\lambda}{1-\lambda}\|y - \hat{y}\| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Also ist $\hat{x} \in B[x;\epsilon] \subset A$. Wegen der Konvexität von A ist $\hat{z} = (1-\lambda)\hat{x} + \lambda\hat{y} \in A$, womit der erste Teil des Satzes bewiesen ist. □

Der Durchschnitt konvexer Mengen ist konvex (Beweis?). Da ferner der ganze Raum X trivialerweise konvex ist, macht die folgende Definition einen Sinn.

Definition 2.2.3 Sei $A \subset X$. Die kleinste konvexe Menge, die A enthält, also der Durchschnitt aller konvexen Mengen, die A enthalten, heißt die *konvexe Hülle* von A und wird mit $\mathrm{co}\,(A)$ bezeichnet.

In Abbildung 2.6 links besteht A aus fünf Punkten in der Ebene, die zugehörige konvexe Hülle ist offenbar ein Viereck. Rechts in derselben Abbildung geben wir ebenfalls eine

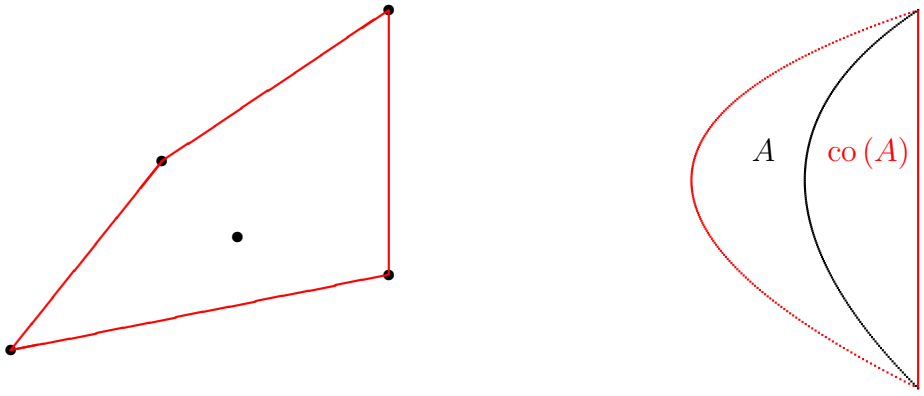

Abbildung 2.6: Eine Menge A und ihre konvexe Hülle $\mathrm{co}\,(A)$

Menge A und ihre zugehörige konvexe Hülle $\mathrm{co}\,(A)$ an.

Das folgende Lemma dient dazu, die Elemente von $\mathrm{co}\,(A)$ durch die von A beschreiben zu können.

Lemma 2.2.4 *Sei $A \subset X$. Dann ist*

$$\mathrm{co}\,(A) = \left\{ \sum_{i=1}^{m} \lambda_i a_i : \lambda_i \geq 0,\ a_i \in A\ (i = 1,\dots,m),\ \sum_{i=1}^{m} \lambda_i = 1,\ m \in \mathbb{N} \right\}.$$

Beweis: Die in der Behauptung rechtsstehende Menge werde mit K bezeichnet. Trivialerweise ist $A \subset K$. Weiter ist K konvex (Beweis?) und daher $\mathrm{co}\,(A) \subset K$. Für die umgekehrte Inklusionsbeziehung zeige man durch vollständige Induktion, dass für alle $m \in \mathbb{N}$ Elemente der Form $\sum_{i=1}^{m} \lambda_i a_i$ mit $\lambda_i \geq 0$, $a_i \in A$, $i = 1,\dots,m$, $\sum_{i=1}^{m} \lambda_i = 1$ zu $\mathrm{co}\,(A)$ gehören.Für $m = 1, 2$ ist dies trivial. Angenommen, es sei für Elemente mit $m \geq 2$ Summanden richtig. Für den Induktionsschritt von m nach $m+1$ setzen wir zur Abkürzung $\Lambda_m := \sum_{i=1}^{m} \lambda_i$. Wegen $\Lambda_m = 1 - \lambda_{m+1}$ ist dann unter der Benutzung der Induktionsvoraussetzung

$$\sum_{i=1}^{m+1} \lambda_i a_i = \Lambda_m \underbrace{\sum_{i=1}^{m} (\lambda_i/\Lambda_m) a_i}_{\in \mathrm{co}\,(A)} + \lambda_{m+1} a_{m+1} \in \mathrm{co}\,(A).$$

Damit ist das Lemma bewiesen. □

Sehr wichtig in der Optimierung und damit auch in der Approximationstheorie ist die Trennung konvexer Mengen durch Hyperebenen. Im $\mathbb{R}^n$ haben Hyperebenen die Form

$$H = \{x \in \mathbb{R}^n : y^T x = \gamma\}$$

mit $(y, \gamma) \in (\mathbb{R}^n \setminus \{0\}) \times \mathbb{R}$. Entsprechend definieren wir eine (abgeschlossene) Hyperebene in X durch

$$H = \{x \in X : l(x) = \gamma\}$$

mit $(l, \gamma) \in (X^* \setminus \{0\}) \times \mathbb{R}$. Diese definiert Halbräume

$$H^- := \{x \in X : l(x) \leq \gamma\}, \qquad H^+ := \{x \in X : l(x) \geq \gamma\}$$

und man sagt, dass Mengen $A, B \subset X$ durch die Hyperebene H getrennt wird, wenn $A \subset H^-$, $B \subset H^+$ (oder $A \subset H^+$, $B \subset H^-$). Nichtdisjunkte Mengen oder disjunkte, nichtkonvexe Mengen haben i. Allg. keine Chance, durch eine Hyperebene getrennt zu werden. Daher ist der folgende Satz schon fast das optimale Ergebnis. Dessen Beweis kann man z. B.bei J. Werner (1984, S. 71) finden.

Satz 2.2.5 (Eidelheit) *Seien $A, B \subset X$ nichtleer, konvex,* $\operatorname{int}(A) \neq \emptyset$ *und* $\operatorname{int}(A) \cap B = \emptyset$. *Dann können* $\operatorname{cl}(A)$ *und* $\operatorname{cl}(B)$ *durch eine (abgeschlossene) Hyperebene in X getrennt werden, d. h. es existiert $(l, \gamma) \in (X^* \setminus \{0\}) \times \mathbb{R}$ mit*

$$l(a) \leq \gamma \leq l(b) \quad \text{für alle } a \in \operatorname{cl}(A),\ b \in \operatorname{cl}(B)$$

und es gilt sogar

$$l(a) < \gamma \quad \text{für alle } a \in \operatorname{int}(A).$$

Für einen *endlichdimensionalen* linearen normierten Raum kann man auf die Voraussetzung $\operatorname{int}(A) \neq \emptyset$ verzichten. Einen Beweis findet man z. B. bei J. Werner (1984, S. 64). Einen elementaren Beweis findet man auch bei O. L. Mangasarian (1969, S. 49).

Satz 2.2.6 *Seien $A, B \subset \mathbb{R}^n$ nichtleer und konvex, $A \cap B = \emptyset$. Dann existiert ein Paar $(y, \gamma) \in (\mathbb{R}^n \setminus \{0\}) \times \mathbb{R}$ mit*

$$y^T a \leq \gamma \leq y^T b \quad \text{für alle } a \in A,\ b \in B.$$

Einen Beweis der letzten beiden Sätze können wir hier noch nicht einmal andeuten. Für die nächsten Folgerungen wollen wir aber wenigstens die Idee zum Beweis angeben, sie sollte als Übung vollständig ausgeführt werden.

Satz 2.2.7 (Strikter Trennungssatz) *Sei $B \subset X$ nichtleer, konvex und abgeschlossen. Sei $x \in X \setminus B$. Dann können $\{x\}$ und B durch eine abgeschlossene Hyperebene strikt getrennt werden, d. h. es existiert $(l, \gamma) \in (X^* \setminus \{0\}) \times \mathbb{R}$ mit*

$$l(x) < \gamma < l(b) \quad \text{für alle } b \in B.$$

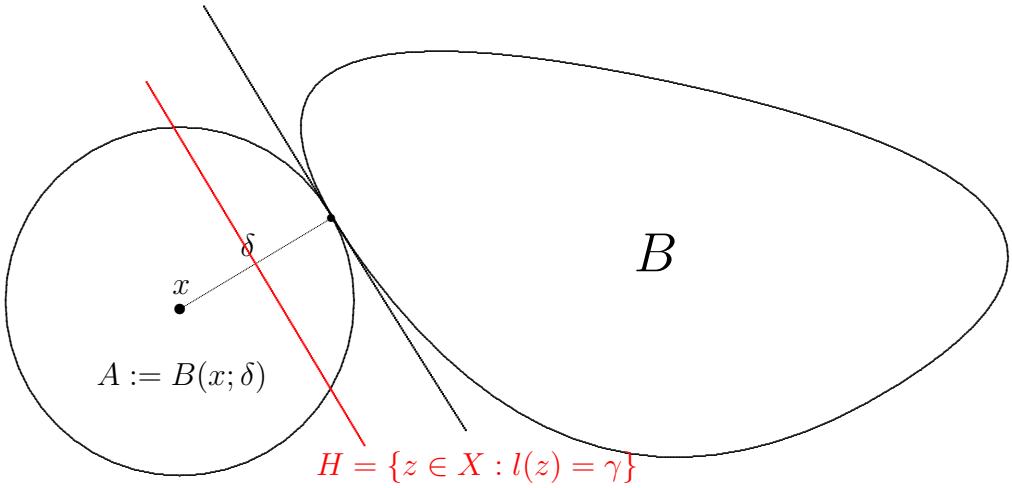

Abbildung 2.7: Strikter Trennungssatz: Veranschaulichung des Beweises

Beweis: In Abbildung 2.7 veranschaulichen wir uns den Beweis. Sei $\delta := \inf_{b\in B} \|x-b\|$ (der Abstand von x zu B). Da B abgeschlossen ist, ist $\delta > 0$ (Beweis?). Sei $A := B(x;\delta)$ die offene Kugel um x mit dem Radius δ. Dann ist A nichtleer, konvex, offen und $A \cap B = \emptyset$. Der Satz von Eidelheit ergibt die Existenz von $(l, \tilde{\gamma}) \in (X^* \setminus \{0\}) \times \mathbb{R}$ mit

$$l(y) < \tilde{\gamma} \leq l(b) \quad \text{für alle } y \in B(x;\delta),\ b \in B.$$

Mit $\gamma := \frac{1}{2}(l(x) + \tilde{\gamma})$ folgt die Behauptung. □

Als Anwendung des strikten Trennungssatzes beweisen wir (mit einer "kleinen" Lücke) das berühmte Farkas-Lemma.

Lemma 2.2.8 *Sei $B \in \mathbb{R}^{k\times n}$, $d \in \mathbb{R}^k$. Dann gilt genau eine der beiden folgenden Aussagen.*

1. *$Bx = d$, $x \geq 0$ besitzt eine Lösung $x \in \mathbb{R}^n$.*
2. *$B^T z \geq 0$, $d^T z < 0$ besitzt eine Lösung $z \in \mathbb{R}^k$.*

Beweis: Die Aussagen 1. und 2. können nicht beide wahr sein (Beweis?). Angenommen, 1. sei falsch. Dann ist

$$d \notin K := \{Bx : x \geq 0\}.$$

Offenbar ist K nichtleer und konvex. K ist aber auch abgeschlossen (die Lücke im Beweis besteht darin, dass wir dies *nicht* beweisen. Wegen des strikten Trennungssatzes existiert ein Paar $(z, \gamma) \in (\mathbb{R}^k \setminus \{0\}) \times \mathbb{R}$ mit $z^T d < \gamma < z^T Bx$ für alle $x \geq 0$. Hieraus folgt $B^T z \geq 0$, $d^T z < 0$, d. h. 2. ist wahr. □

Für einen Beweis ohne Lücke sei auf J. Werner (1984, S. 37ff.) verwiesen. Aus dem Trennungssatz sollen weitere Folgerungen gezogen werden. Hierzu definieren wir zunächst

Definition 2.2.9 Sei $A \subset X$. Eine (abgeschlossene) Hyperebene

$$H = \{x \in X : l(x) = \gamma\}$$

mit $(l, \gamma) \in (X^* \setminus \{0\}) \times \mathbb{R}$ heißt eine *Stützhyperebene* für A, falls

1. $A \cap H \neq \emptyset$,
2. A ganz in einem der Halbräume H^+ oder H^- liegt, wobei

$$H^+ := \{x \in X : l(x) \geq \gamma\}, \qquad H^- := \{x \in X : l(x) \leq \gamma\}.$$

Ein Punkt $a \in A$ heißt *Stützpunkt* von A, wenn es eine Stützhyperebene H für A mit $a \in A \cap H$ gibt.

In Abbildung 2.8 veranschaulichen wir uns die eingeführten Begriffe. Ein Punkt des

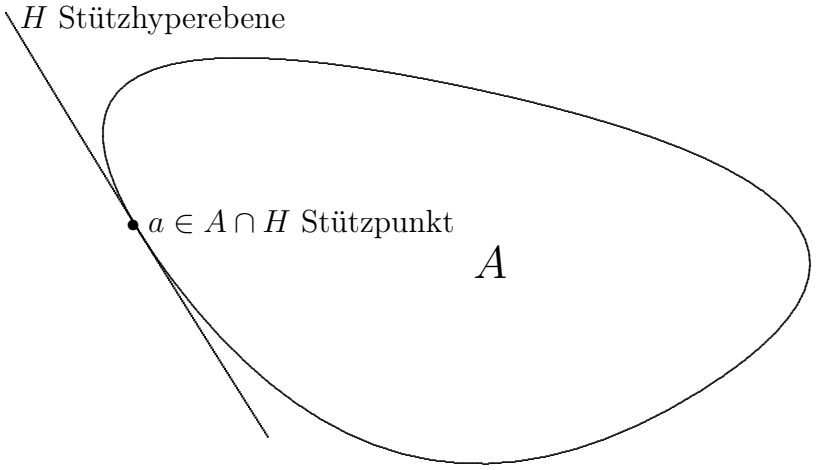

Abbildung 2.8: Stützhyperebene und Stützpunkt

Inneren $\operatorname{int}(A)$ einer Menge A kann kein Stützpunkt von A sein. Daher ist das folgende Ergebnis schon fast optimal.

Satz 2.2.10 *Sei $A \subset X$ nichtleer, konvex und abgeschlossen. Ferner sei $\operatorname{int}(A) \neq \emptyset$ oder X endlichdimensional. Dann ist jeder Randpunkt von A, d. h. jedes $a \in A \setminus \operatorname{int}(A)$ ein Stützpunkt von A.*

Beweis: Ist $\operatorname{int}(A) \neq \emptyset$ und $a \in A \setminus \operatorname{int}(A)$, so können $\{a\}$ und $\operatorname{int}(A)$ wegen Satz 2.2.5, dem Satz von Eidelheit, durch eine (abgeschlossene) Hyperebene getrennt werden und dies ist offenbar die gesuchte Stützhyperebene. Ist dagegen $\operatorname{int}(A) = \emptyset$ und X endlichdimensional, so liegt A selbst schon in einer Hyperebene (Beweis?). □

Fast zum Schluss des allgemeinen Teils dieses Abschnitts bringen wir noch Folgerungen aus den Trennungssätzen, die mit dem Namen Hahn-Banach verbunden sind.

Satz 2.2.11 (Fortsetzungssatz von Hahn-Banach) *Sei $L \subset X$ ein linearer Teilraum und $l_0 \in L^*$, also $l_0 \colon L \longrightarrow \mathbb{R}$ linear und*

$$\|l_0\|_L := \sup_{x \in L \setminus \{0\}} \frac{|l_0(x)|}{\|x\|} < +\infty.$$

Dann existiert eine Fortsetzung $l \in X^*$ von l (also $l(x) = l_0(x)$ für alle $x \in L$) mit $\|l\| = \|l_0\|_L$.

Beweis: O. B. d. A. ist $l_0 \neq 0$. Wir definieren

$$A := \{(x,t) \in X \times \mathbb{R} : \|l_0\|_L \, \|x\| < t\}, \qquad B := \{(y, l_0(y)) \in X \times \mathbb{R} : y \in L\}.$$

Dann sind A und B konvex (B sogar ein linearer Teilraum von $X \times \mathbb{R}$) und A offen in $X \times \mathbb{R}$ (wobei als Norm in $X \times \mathbb{R}$ z. B. $\|(x,t)\| := \max(\|x\|, |t|)$ genommen wird). Es ist $A \cap B = \emptyset$, denn andernfalls gibt es ein $x \in L$ mit $\|l_0\|_L \, \|x\| < l_0(x)$, ein Widerspruch. Wegen Satz 2.2.5, dem Satz von Eidelheit, können A und B durch eine (abgeschlossene) Hyperebene in $X \times \mathbb{R}$ getrennt werden. Daher (man benutze $(X \times \mathbb{R})^* = X^* \times \mathbb{R}$) existiert $((l,s),\gamma) \in (X^* \times \mathbb{R} \setminus \{(0,0)\}) \times \mathbb{R}$ mit

$$(*) \qquad \begin{cases} l(x) + s \cdot t < \gamma \leq l(y) + s \cdot l_0(y) \\ \quad \text{für alle } (x,t) \in A,\ y \in L. \end{cases}$$

Da L ein linearer Teilraum von X und $l(\cdot) + s \cdot l_0(\cdot)$ auf L nach unten beschränkt ist, ist $l(y) + s \cdot l_0(y) = 0$ für alle $y \in L$. Da $(0,1) \in A$, ist $s < \gamma \leq 0$ und daher o. B. d. A. $s = -1$. Also ist l eine Fortsetzung von l_0. Ferner ist $l(x) \leq \|l_0\|_L \, \|x\|$ für alle $x \in X$. Denn gäbe es ein $x \in X$ mit $l(x) > \|l_0\|_L \, \|x\|$, so wäre $(x, l(x)) \in A$ und daher wegen $(*)$ (unter Benutzung von $s = -1$) $l(x) - l(x) < \gamma \leq 0$, ein Widerspruch. Vertauschen von x mit $-x$ liefert $|l(x)| \leq \|l_0\| \, \|x\|$ für alle $x \in X$ bzw. $\|l\| \leq \|l_0\|_L$. Andererseits ist

$$\|l\| = \sup_{x \in \setminus \{0\}} \frac{|l(x)|}{\|x\|} \geq \sup_{x \in L \setminus \{0\}} \frac{|l(x)|}{\|x\|} = \sup_{x \in L \setminus \{0\}} \frac{|l_0(x)|}{\|x\|} = \|l_0\|_L.$$

Insgesamt ist l die gesuchte Fortsetzung von l_0 und der Satz von Hahn-Banach ist bewiesen. □

Eine einfache Folgerung ist

Korollar 2.2.12 *Sei $x \in X$. Dann gibt es ein $l \in X^*$ mit $l(x) = \|x\|$ und $\|l\| = 1$.*

Beweis: Zunächst sei $x \neq 0$ und $L := \operatorname{span}(x) = \{\alpha x : \alpha \in \mathbb{R}\}$. Sei $l_0 \in L^*$ durch $l_0(\alpha x) = \alpha \, \|x\|$ definiert. Dann ist $l_0(x) = \|x\|$ und

$$\|l_0\|_L = \sup_{\alpha \neq 0} \frac{|l_0(\alpha x)|}{\|\alpha x\|} = 1.$$

Eine Fortsetzung l von l_0 nach Hahn-Banach ist das gesuchte Element. Ist dagegen $x = 0$, so gibt es (zumindestens dann, wenn X nicht nur aus dem Nullelement 0 besteht) ein $l \in X^*$ mit $\|l\| = 1$ (und $l(x) = \|x\|$). □

Bemerkung: Ist $x \in X$ und $l \in X^*$ mit $\|l\| \leq 1$, so ist natürlich

$$l(x) \leq |l(x)| \leq \|l\| \, \|x\| \leq \|x\|.$$

Die Aussage von Korollar 2.2.12 kann daher auch als die Aussage

$$\|x\| = \max_{l \in X^*, \|l\| \leq 1} l(x) \qquad \text{für alle } x \in X$$

formuliert werden. □

Zum Schluss dieses Abschnitts über konvexe Mengen in linearen normierten Räumen geben wir noch zwei Sätze über konvexe Mengen im $\mathbb{R}^n$ an. Die erste Aussage werden wir nicht beweisen, obwohl der Beweis nicht schwierig ist, siehe z. B. J. Werner (1984, S. 43).

Satz 2.2.13 (Carathéodory) *Sei $A \subset \mathbb{R}^n$ und $x \in \operatorname{co}(A)$. Dann ist x eine Konvexkombination von höchstens $n+1$ Punkten aus A. Genauer lässt sich x in der Form $x = \sum_{i=1}^{m} \mu_i a_i$ darstellen, wobei $m \leq n+1$, $a_i \in A$, $\mu_i \geq 0$, $i = 1, \ldots, m$ und $\sum_{i=1}^{m} \mu_i = 1$.*

Eine Folgerung ist

Satz 2.2.14 *Ist $A \subset \mathbb{R}^n$ kompakt, so ist auch die konvexe Hülle $\operatorname{co}(A)$ von A kompakt.*

Beweis: Sei

$$S := \left\{ \lambda = (\lambda_i) \in \mathbb{R}^{n+1} : \lambda \geq 0,\ \sum_{i=1}^{n+1} \lambda_i = 1 \right\}.$$

Dann ist $S \subset \mathbb{R}^{n+1}$ kompakt. Man definiere die Abbildung

$$\psi \colon S \times \underbrace{A \times \cdots \times A}_{n+1 \text{ Faktoren}} \longrightarrow \operatorname{co}(A)$$

durch

$$\psi(\lambda, a_1, \ldots, a_{n+1}) := \sum_{i=1}^{n+1} \lambda_i a_i.$$

Wegen des Satzes von Carathéodory ist

$$\psi(S \times A \times \cdots \times A) = \operatorname{co}(A).$$

Da ψ stetig ist, ist $\operatorname{co}(A)$ als stetiges Bild einer kompakten Menge selbst kompakt. □

2.3 Kompaktheit in linearen normierten Räumen. Schwache Konvergenz. Reflexive Räume

Existenzaussagen bei Optimierungsaufgaben und daher auch in der Approximationstheorie beruhen i. Allg. auf Kompaktheitsaussagen (und der Tatsache, dass eine stetige reelwertige Funktion auf einer kompakten Menge ihre Extrema annimmt).

Im folgenden sei weiter $(X, \|\cdot\|)$ ein linearer normierter Raum. Eine Menge $A \subset X$ heißt *kompakt* (bzw. *folgenkompakt*), wenn es zu jeder Folge $\{a_k\} \subset A$ eine Teilfolge $\{a_{k_j}\} \subset \{a_k\}$ und ein $a \in A$ mit $\lim_{j\to\infty} a_{k_j} = a$. Die weiteren üblichen Definitionen von kompakt (endliche Überdeckungseigenschaft usw.) sind mit obiger Definition äquivalent.

Lax gesprochen gibt es in unendlichdimensionalen linearen normierten Räumen nicht allzu viele kompakte Mengen. Das zeigt der folgende

Satz 2.3.1 *Die Einheitskugel $B[0;1] := \{x \in X : \|x\| \leq 1\}$ in einem linearen normierten Raum $(X, \|\cdot\|)$ ist genau dann kompakt, wenn X endlichdimensional ist.*

Beweis: Ist X endlichdimensional, so ist die Einheitskugel $B[0;1]$ als beschränkte und abgeschlossene Menge in einem endlichdimensionalen Raum kompakt. Umgekehrt nehmen wir an, X sei unendlichdimensional und zeigen, dass $B[0;1]$ nicht kompakt ist, da eine Folge $\{x_k\} \subset B[0;1]$ mit $\|x_k - x_l\| \geq \frac{1}{2}$ für alle $k \neq l$ existiert (und aus dieser kann offensichtlich keine konvergente Teilfolge ausgewählt werden). Zur Konstruktion der Folge $\{x_k\}$ benutzen wir die folgende Aussage, die auch *Lemma von Riesz* genannt wird:

- *Sei $L \subset X$, $L \neq X$, ein echter linearer Teilraum des linearen normierten Raumes $(X, \|\cdot\|)$ und $\delta > 0$ eine reelle Zahl. Dann gibt es ein $x_\delta \in X$ mit $\|x_\delta\| = 1$ und*

$$d(x_\delta, L) := \inf_{x \in L} \|x - x_\delta\| \geq 1 - \delta.$$

Denn: Für $\delta \geq 1$ ist die Aussage trivial, so dass wir $\delta \in (0,1)$ annehmen können. Da L ein *echter* linearer Teilraum von X ist, existiert ein $y \in X \setminus L$. Da L abgeschlossen ist, ist $d(y, L) = \inf_{x \in L} \|x - y\| > 0$. Es existiert ein $z \in L$ mit

$$\|z - y\| \leq \frac{d(y, L)}{1 - \delta}.$$

Nun setze man

$$x_\delta := \frac{y - z}{\|y - z\|}.$$

Dann ist $\|x_\delta\| = 1$. Für beliebiges $x \in L$ ist ferner

$$\begin{aligned} \|x - x_\delta\| &= \left\| \left(x + \frac{z}{\|y - z\|} - \frac{1}{\|y - z\|} y \right\| \right. \\ &= \frac{1}{\|y - z\|} \| \underbrace{\|y - z\| \, x + z}_{\in L} - y \| \\ &\geq \frac{d(y, L)}{\|y - z\|} \\ &\geq 1 - \delta. \end{aligned}$$

Damit ist das Lemma von Riesz bewiesen[2].

[2]In $(X, \|\cdot\|) := (\|R^n, \|\cdot\|_2)$ ist obige Aussage auch noch für $\delta = 0$ richtig. Man wähle nämlich ein beliebiges von 0 verschiedenes Element senkrecht zu L und gewinne x_0 durch Normieren dieses Elementes auf die Länge 1. In unendlichdimensionalen Räumen ist dies i. Allg. nicht möglich. Sei z. B.

$$X := \{x \in C[0,1] : x(0) = 0\}, \quad \|\cdot\| := \|\cdot\|_\infty, \quad L := \left\{ x \in X : \int_0^1 x(t)\, dt. \right.$$

Offenbar ist L ein abgeschlossener echter linearer Teilraum von X. Angenommen, es existiert ein $x_0 \in X$ mit $\|x_0\|_\infty = 1$ und $\|x - x_0\|_\infty \geq 1$ für alle $x \in L$. Definiert man $x \in X$ durch

$$x(t) := x_0(t) - \frac{n+1}{n} \int_0^1 x_0(t)\, dt \cdot t^{1/n},$$

Dieses Lemma wenden wir nun zur Konstruktion eine Folge $\{x_k\}$ mit $\|x_k\| = 1$, $k = 1, \ldots$, aus der sich keine konvergente Folge auswählen lässt, wiederholt mit $\delta := \frac{1}{2}$ an. Sei $x_1 \in X$ mit $\|x_1\| = 1$ beliebig gewählt, setze $L_1 := \operatorname{span}\{x_1\}$. Eine Anwendung des Lemma von Riesz liefert die Existenz von $x_2 \in X$ mit $\|x_2\| = 1$ und $\|x - x_2\| \geq \frac{1}{2}$ für alle $x \in L_1$. Angenommen, $x_1, \ldots, x_k$ mit $\|x_j\| = 1$ und $\|x_i - x_j\| \geq \frac{1}{2}$ für $1 \leq i < j \leq k$ seien schon gefunden. Setze $L_k := \operatorname{span}\{x_1, \ldots, x_k\}$. Da X unendlichdimensional, ist L_k ein echter linearer Teilraum von X, als endIichdimensionaler linearer Teilraum ist L_k ferner abgeschlossen. Eine Anwendung des Lemmas von Riesz auf L_k liefert die Existenz von $x_{k+1} \in X$ mit $\|x_{k+1}\| = 1$ und $\|x - x_{k+1}\| \geq \frac{1}{2}$ für alle $x \in L_k$. Also ist die Existenz einer Folge $\{x_k\} \subset B[0;1]$ mit $\|x_k - x_j\| \geq \frac{1}{2}$ für alle $k \neq j$ und die Aussage des Satzes bewiesen. □

Aus einer beschränkten Folge in einem unendlichdimensionalen linearen normierten Raum lässt sich also i. Allg. keine konvergente Teilfolge auswählen. Bezüglich eines schwächeren Konvergenzbegriffs kann dies aber sehr wohl möglich sein.

Definition 2.3.2 Eine Folge $\{x_k\} \subset X$ heißt *schwach konvergent* gegen ein $x \in X$ (hierfür schreiben wir auch $w\text{-}\lim_{k\to\infty} x_k = x$ oder $x_k \rightharpoonup x$), falls $\lim_{k\to\infty} l(x_k) = l(x)$ für jedes $l \in X^*$.

Bemerkungen: 1. Es gelten die üblichen Regeln:

$$x_k \rightharpoonup x, \quad y_k \rightharpoonup y_k, \quad \alpha,\ \beta \in \mathbb{R} \Longrightarrow \alpha x_k + \beta y_k \rightharpoonup \alpha x + \beta y.$$

2. Der schwache Limes einer schwach konvergenten Folge ist eindeutig bestimmt. Denn sind x^1 und x^2 schwache Limiten einer Folge, so ist $l(x^1 - x^2) = 0$ für alle $l \in X^*$. Aus Korollar 2.2.12 folgt $x^1 - x^2 = 0$ bzw. $x^1 = x^2$.

3. Eine schwach konvergente Folge ist beschränkt, d. h. es gilt die Implikation

$$x_k \rightharpoonup x \Longrightarrow \{\|x_k\|\} \text{ ist beschränkt.}$$

Der übliche Beweis hierfür ist nicht trivial und beruht auf dem *Prinzip der gleichmäßigen Beschränktheit*, siehe z. B. F. Hirzebruch, W. Scharlau (1971, S. 61). Dieses Prinzip (wir werden es nicht beweisen) sagt aus:

- *Sei $(X, \|\cdot\|)$ ein Banachraum, sei $(Y, \|\cdot\|)$ ein linearer normierter Raum und $\{T_k\}_{k\in K} \subset L(X,Y)$ eine punktweise beschränkte Menge linearer, stetiger Abbildungen von X in Y, d. h. es ist*

$$\sup_{k\in K} \|T_k x\| < +\infty \quad \textit{für alle } x \in X.$$

---

so ist $x \in L$ und folglich

$$1 \leq \frac{n+1}{n}\left|\int_0^1 x_0(t)\,dt\right|, \qquad n = 1, 2, \ldots.$$

Mit $n \to \infty$ folgt also

$$1 \leq \left|\int_0^1 x_0(t)\,dt\right|.$$

Andererseits folgt aus $\|x_0\|_\infty = 1$, $x(0) = 0$ und der Stetigkeit von x_0, dass $|\int_0^1 x_0(t)\,dt| < 1$, ein Widerspruch.

Dann ist $\{T_k\} \subset L(X,Y)$ (gleichmäßig) beschränkt, d. h. es ist

$$\sup_{k\in K} \|T_k\| < +\infty.$$

Jetzt beachten wir, dass sich der lineare normierte Raum X kanonisch in den *Bidualraum* $X^{**} := (X^*)^*$ einbetten lässt. Man definiere die Abbildung $i\colon X \longrightarrow X^{**} := (X^*)^*$ von X in den Bidualraum X^{**} durch $i(x)(l) := l(x)$. Dann ist $i(x)$ für jedes $x \in X$ eine lineare und stetige Abbildung von X^* nach $\mathbb{R}$, also ein Element des Bidualraums X^{**}. Die Linearität von $i(x)$ bei gegebenem $x \in X$ ist für $\alpha_1, \alpha_2 \in \mathbb{R}$ und $l_1, l_2 \in X^*$ aus

$$i(x)(\alpha_1 l_1 + \alpha_2 l_2) = (\alpha_1 l_1 + \alpha_2 l_2)(x) = \alpha_1 l_1(x) + \alpha_2 l_2(x) = \alpha_1 i(x)(l_1) + \alpha_2 i(x)(l_2)$$

ersichtlich. Daher ist $i(X)$ ein linearer Raum. Für $x \in X$ und beliebiges $l \in X^*$ ist $|i(x)(l)| = |l(x)| \le \|l\|\,\|x\|$ und folglich $i(x)\colon X^* \longrightarrow \mathbb{R}$ linear und stetig, also $i(x) \in X^{**}$. Weiter liest man aus $|i(x)| \le \|x\|\,\|l\|$ auch $\|i(x)\| \le \|x\|$ ab. Wegen Korollar 2.2.12 existiert zu $x \in X$ ein $l \in X^*$ mit $\|l\| = 1$ und $l(x) = \|x\|$. Folglich ist $\|i(x)\| \ge \|x\|$ und insgesamt $\|i(x)\| = \|x\|$ für alle $x \in X$. Nun ist es einfach, die Aussage 3. mit Hilfe des Prinzips der gleichmäßigen Beschränktheit zu beweisen. Schwache Konvergenz $x_k \rightharpoonup x$ bedeutet $\lim_{k\to\infty} l(x_k) = l(x)$ bzw. $\lim_{k\to\infty} i(x_k)(l) = i(x)(l)$ für alle $l \in X^*$. Dies impliziert $\sup_{k\in\mathbb{N}} \|i(x_k)(l)\| < +\infty$ für alle $l \in X^*$. Es ist $\{i(x_k)\} \subset L(X^*, \mathbb{R}) = X^{**}$ und X^* ist ein Banachraum (Beweis?). Das Prinzip der gleichmäßigen Beschränktheit ist daher anwendbar und liefert

$$\sup_{k\in\mathbb{N}} \|i(x_k)\| = \sup_{k\in\mathbb{N}} \|x_k\| < +\infty,$$

und das ist die Behauptung.

4. Die starke Konvergenz $x_k \to x$ impliziert die schwache Konvergenz $x_k \rightharpoonup x$. Denn für ein beliebiges $l \in X^*$ ist

$$|l(x_k) - l(x)| = |l(x_k - x)| \le \|l\|\,\|x_k - x\|.$$

Ist X endlichdimensional, so gilt auch die Umkehrung (Beweis?). □

Im günstigsten Fall kommt bei späteren Anwendungen der Begriff der schwachen Konvergenz nur in Beweisen von Aussagen vor (und nicht in den Voraussetzungen oder der Behauptung). Es ist aber trotzdem wichtig den Dualraum der wichtigsten Räume, und dies sind vor allem Funktionenräume, zu kennen. Hierzu benötigt man Integrations- bzw. Maßtheorie und daher können wir die folgenden Ergebnisse im wesentlichen nur ohne Beweis angeben.

Beispiele: 1. Sei $X := \operatorname{span}\{x_1, \ldots, x_n\}$ ein endlichdimensionaler linearer Raum mit linear unabhängigen $\{x_1, \ldots, x_n\}$. Jedes $x = \sum_{j=1}^n \alpha_j x_j \in X$ kann mit dem Koordinatenvektor $\alpha = (\alpha_j) \in \mathbb{R}^n$ identifiziert werden. Der Dualraum X^* der linearen (und automatisch stetigen) linearen Funktionale kann mit dem $\mathbb{R}^n$ identifiziert werden. Denn die Abbildung $i\colon X^* \longrightarrow \mathbb{R}^n$, definiert durch $i(l) := (l(x_j)) \in \mathbb{R}^n$, bildet offensichtlich X^* linear und bijektiv auf den $\mathbb{R}^n$ ab.

2. Auf S. 19 hatten wir schon kurz die L^p-Räume $L^p[\alpha,\beta]$ angesprochen. Noch einmal: Für $1 \le p < \infty$ sei $L^p[\alpha,\beta]$ die Menge der auf $[\alpha,\beta]$ messbaren Funktionen $x(\cdot)$, für die $\int_\alpha^\beta |x(t)|^p\,dt$ im Lebesgueschen Sinne existiert, wobei zwei fast überall gleiche Funktionen, die sich also nur auf einer Nullmenge aus $[\alpha,\beta]$ unterscheiden, identifiziert werden. Mit $\|x\|_p := (\int_\alpha^\beta |x(t)|^p\,dt)^{1/p}$ ist $(L^p[\alpha,\beta], \|\cdot\|_p)$ ein Banachraum. Ferner gilt:

(a) Für $1 < p < \infty$ ist $(L^p[\alpha,\beta])^* = L^q[\alpha,\beta]$ mit $1/p + 1/q = 1$.

Genauer: Zu jedem $l \in (L^p[\alpha,\beta])^*$ gibt es genau ein $y \in L^q[\alpha,\beta]$ mit

$$l(x) = \int_\alpha^\beta y(t)x(t)\,dt \quad \text{für alle } x \in L^p[\alpha,\beta]$$

und es ist $\|l\| = \|y\|_q$. Bei vorgegebenem $y \in L^q[\alpha,\beta]$ ist umgekehrt durch die Vorschrift $l(x) := \int_\alpha^\beta y(t)x(t)\,dt$ ein Element $l \in (L^p[\alpha,\beta])^*$ mit $\|l| = \|y\|_q$ gegeben. Man beachte, dass insbesondere $(L^2[\alpha,\beta])^* = L^2[\alpha,\beta]$ ist.

(b) Es ist $(L^1[\alpha,\beta])^* = L^\infty[\alpha,\beta]$.

Genauer gilt: Zu jedem $l \in (L^1[\alpha,\beta])^*$ gibt es genau ein $y \in L^\infty[\alpha,\beta]$ mit

$$l(x) = \int_\alpha^\beta y(t)x(t)\,dt \quad \text{für alle } x \in L^1[\alpha,\beta]$$

und es ist $\|l\| = \|y\|_\infty$. Bei vorgegebenem $y \in L^\infty[\alpha,\beta]$ ist umgekehrt durch die Vorschrift $l(x) := \int_\alpha^\beta y(t)x(t)\,dt$ ein Element $l \in (L^1[\alpha,\beta])^*$ mit $\|l\| = \|y\|_\infty$ gegeben. Hierbei ist $L^\infty[\alpha,\beta]$ die Menge der auf $[\alpha,\beta]$ messbaren Funktionen $x(\cdot)$, die im wesentlichen beschränkt sind, für die also eine Konstante $c > 0$ existiert mit $|x(t)| \le c$ für fast alle $t \in [\alpha,\beta]$, wobei wiederum zwei fast überall gleiche Funktionen identifiziert werden. Mit der Norm

$$\|x\|_\infty := \inf\{c > 0 :\ |x(t)| \le c \text{ für fast alle } t \in [\alpha,\beta]$$

ist $(L^\infty[\alpha,\beta], \|\cdot\|)$ ein Banachraum.

Diese Aussagen können hier noch nicht einmal andeutungsweise bewiesen werden. Wir verweisen nur auf F. Hirzebruch, W. Scharlau (1971, S. 80).

3. Für $1 \le p < \infty$ sind die Folgenräume l^p definiert durch

$$l^p := \left\{ x = \{x_k\} : \sum_{k=1}^\infty |x_k|^p < \infty \right\},$$

während

$$l^\infty := \left\{ x = \{x_k\} : \sup_{k\in\mathbb{N}} |x_k| < \infty \right\}.$$

Mit den Normen

$$\|x\|_p := \begin{cases} \left(\sum_{k=1}^\infty |x_k|^p\right)^{1/p}, & 1 \le p < \infty, \\ \sup_{k\in\mathbb{N}} |x_k|, & p = \infty \end{cases}$$

sind $(l^p, \|\cdot\|_p)$ für $1 \le p \le \infty$ Banachräume. Entsprechend den Ergebnissen in 2. gilt hier:

(a) Für $1 < p < \infty$ ist $(l^p)^* = l^q$ mit $1/p + 1/q = 1$.

Genauer: Zu jedem $l \in (l^p)^*$ gibt es genau ein $y \in l^q$ mit $l(x) = \sum_{k=1}^{\infty} y_k x_k$ und $\|l\| = \|y\|_q$.

(b) Es ist $(l^1)^* = l^\infty$.

Genauer: Zu jedem $l \in (l^1)^*$ gibt es genau ein $y \in l^\infty$ mit $l(x) = \sum_{k=1}^{\infty} y_k x_k$ und $\|l\| = \|y\|_\infty$.

3. Die Berechnung des Dualraums von $(C[\alpha,\beta], \|\cdot\|_\infty)$ gehört in jede vernünftige Vorlesung über Funktionalanalysis. Wir können das Ergebnis hier nur skizzieren.

(a) Sei $\mathcal{D}$ die Menge der Zerlegungen $\Delta : \alpha = t_0 < t_1 < \cdots < t_k = \beta$, $k \in \mathbb{N}$. Eine Funktion $v\colon [\alpha,\beta] \longrightarrow \mathbb{R}$ heißt von *beschränkter Schwankung* auf $[\alpha,\beta]$, falls

$$TV(v) := \sup_{\Delta\in\mathcal{D}} \sum_{j=1}^{k} |v(t_j) - v(t_{j-1})| < \infty.$$

Hierbei heißt $TV(v)$ die *Totalvariation* von v. Die Menge der Funktionen von von beschränkter Schwankung (bounded variation) auf $[\alpha,\beta]$ wird mit $BV[\alpha,\beta]$ bezeichnet[3].

(b) In der Analysis wird bewiesen:

– Sei $x \in C[\alpha,\beta]$, $v \in BV[\alpha,\beta]$. Ist $\{\Delta_k\} \subset \mathcal{D}$ eine Folge von Unterteilungen $\Delta^{(k)} : \alpha = t_0^{(k)} < t_1^{(k)} < \cdots < t_k^{(k)} = \beta$ mit

$$|\Delta^{(k)}| := \max_{j=1,\ldots,k} (t_j^{(k)} - t_{j-1}^{(k)}) \to 0.$$

Ferner sei $s^{(k)} = (s_1(k), \ldots, s_k^{(k)})$ mit $t_{j-1}^{(k)} \le s_j^{(k)} \le t_j^{(k)}$, $j = 1,\ldots,k$ und

$$\Sigma(\Delta^{(k)}, s^{(k)}) := \sum_{j=1}^{k} x(s_j^{(k)})(v(t_j^{(k)}) - v(t_{j-1}^{(k)})).$$

[3]Ist $v\colon [\alpha,\beta] \longrightarrow \mathbb{R}$ monoton, etwa $v(s) \le v(t)$ für $\alpha \le s \le t \le \beta$, so ist $v \in BV[\alpha,\beta]$. Denn ist $\Delta : \alpha = t_0 < \cdots < t_k = \beta$ eine Zerlegung von $[\alpha,\beta]$, so ist

$$\sum_{j=1}^{k} |v(t_j) - v(t_{j-1})| = \sum_{j=1}^{k} (v(t_j) - v(t_{j-1})) = v(\beta) - v(\alpha),$$

also $TV(v) = v(\beta) - v(\alpha)$. Ebenso leicht sieht man ein, dass $C^1[\alpha,\beta] \subset BV[\alpha,\beta]$, wobei $C^1[\alpha,\beta]$ für die auf $[\alpha,\beta]$ stetig differenzierbaren Funktionen steht und es ist $TV(v) \le \|v'\|_\infty\,(\beta-\alpha)$ für $v \in C^1[\alpha,\beta]$. Ferner überlegt man sich leicht, dass Treppenfunktionen von beschränkter Schwankung sind.

Dann existiert das sogenannte *Riemann-Stieltjes-Integral*

$$\int_\alpha^\beta x(t)\,dv(t) := \lim_{k\to\infty} \Sigma(\Delta^{(k)}, s^{(k)})$$

und ist von der Wahl von $\{\Delta^{(k)}\}$ und $s^{(k)}\}$ unabhängig.

(c) Ist $v \in BV[\alpha,\beta]$ und definiert man $l\colon C[\alpha,\beta] \longrightarrow \mathbb{R}$ durch

$$l(x) := \int_\alpha^\beta x(t)\,dv(t),$$

so ist l natürlich linear. Weiter ist offenbar (Beweis?) $|l(x)| \le TV(v)\,\|x\|_\infty$, also ist $l \in (C[\alpha,\beta])^*$ und $\|l\| \le TV(v)$. Umgekehrt existiert zu jedem $l \in (C[\alpha,\beta])^*$ ein $v \in BV[\alpha,\beta]$ mit

- $l(x) = \int_\alpha^\beta x(t)\,dv(t)$ für alle $x \in C[\alpha,\beta]$,
- $\|l\| = TV(v)$.

Dies ist nicht trivial (benutzt wird der Fortsetzungssatz von Hahn-Banach) und soll hier nicht bewiesen werden. Man beachte, dass v nicht eindeutig durch l bestimmt ist. Daher können wir $(C[\alpha,\beta])^*$ und $BV[\alpha,\beta]$ *nicht* identifizieren.

□

Oben haben wir gesehen, dass $(L^2[\alpha,\beta])^* = L^2[\alpha,\beta]$ und $(l^2)^* = l^2$. Etwas entsprechendes ist allgemein für Hilberträume richtig.

Satz 2.3.3 (Rieszscher Darstellungssatz) *Sei $(X,(\cdot,\cdot))$ ein Hilbertraum. Dann ist die Abbildung $j\colon X \longrightarrow X^*$, definiert durch $j(y)(x) := (y,x)$, ein isometrischer Isomorphismus, d. h. $j\colon X \longrightarrow X^*$ ist linear und bijektiv und es ist $\|j(y)\| = \|y\|$ für alle $y \in X$.*

Beweis: Für jedes $y \in X$ ist die Abbildung $j(y)\colon X \longrightarrow \mathbb{R}$ (offensichtlich) linear und stetig, da wegen der Cauchy-Schwarzschen Ungleichung $|j(y)(x)| = |(y,x)| \le \|y\|\,\|x\|$ für alle $x \in X$. Also ist $j(y) \in X^*$ und $\|j(y)\| \le \|y\|$. Für $y \neq 0$ ist ferner einerseits $j(y)(y/\|y\|) = (y, y/\|y\|) = \|y\|$, andererseits $j(y)(y/\|y\|) \le \|j(y)\| \cdot 1 = \|j(y)\|$, insgesamt also $\|j(y)\| = \|y\|$ für alle $y \in X$. Dass $j\colon X \longrightarrow X^*$ linear ist, ist klar, ebenso, dass j injektiv ist (Beweis?). Zu zeigen bleibt also die Surjektivität von j, dass es also zu jedem $l \in X^*$ ein $y \in X$ mit $l(x) = (y,x)$ für alle $x \in X$ gibt.

Sei $N := \{x \in X : l(x) = 0\}$. Dann ist N ein abgeschlossener linearer Teilraum von X. O. B. d. A. ist $N \neq X$ (andernfalls ist $l = 0$ und es kann $y = 0$ gewählt werden). Man wähle $z_0 \in X \setminus N$ und definiere $\delta := \inf_{x\in N} \|x - z_0\|$. Da N abgeschlossen ist, ist $\delta > 0$. Wir zeigen, dass das Infimum bei der Definition von δ angenommen wird. Sei hierzu $\{x_k\} \subset N$ eine *Minimalfolge*, also $\|x_k - z_0\| \to \delta$. Es wird gezeigt, dass $\{x_k\} \subset N$ eine Cauchy-Folge ist und folglich gegen ein $x_0 \in N$ mit $\|x_0 - z_0\| = \delta$ konvergiert. Für $k,p \in \mathbb{N}$ ist

$$\delta \le \left\| \frac{x_k + x_p}{2} - z_0 \right\| \le \frac{1}{2} \underbrace{\|x_k - z_0\|}_{\to\delta} + \frac{1}{2} \underbrace{\|x_p - z_0\|}_{\to\delta},$$

woraus wir

$$\lim_{k,p\to\infty}\left\|\frac{x_k+x_p}{2}-z_0\right\|=0$$

ablesen. Wegen der Parallelogrammgleichung (siehe Seite 10) ist

$$\|x_k-x_p\|^2=2\underbrace{\|x_k-z_0\|^2}_{\to\delta^2}+2\underbrace{\|x_p-z_0\|^2}_{\to\delta^2}-4\underbrace{\left\|\frac{x_k+x_p}{2}-z_0\right\|^2}_{\to\delta^2}\to 0,$$

also $\{x_k\}$ eine Cauchy-Folge. Daher existiert $x_0\in N$ mit

$$\|x_0-z_0\|=\delta=\inf_{x\in N}\|x-z_0\|.$$

In Abbildung 2.9 veranschaulichen wir uns die Situation. Sei $y_0:=x_0-z_0$. Wir zeigen,

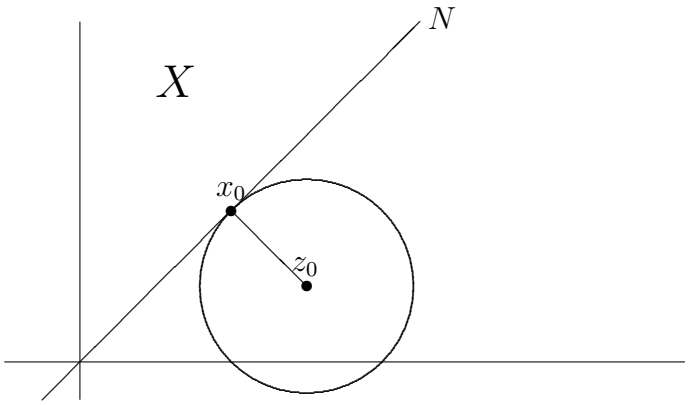

Abbildung 2.9: Beweis des Rieszschen Darstellungssatzes

dass $(y_0,x)=0$ für alle $x\in N$ (dass also y_0 senkrecht zu N ist). Für $x=0$ ist dies trivial, sei also $x\neq 0$. Dann ist

$$\begin{aligned}
\delta &\leq \left\|\underbrace{x_0-\frac{(y_0,x)}{\|x\|^2}\,x}_{\in N}-z_0\right\|^2\\
&= \left(y_0-\frac{(y_0,x)}{\|x\|^2}\,x,\,y_0-\frac{(y_0,x)}{\|x\|^2}\,x\right)\\
&= \|y_0\|^2-\frac{(y_0,x)^2}{\|x\|^2}+\frac{(y_0,x)^2}{\|x\|^2}\\
&= \delta-\frac{(y_0,x)^2}{\|x\|^2}
\end{aligned}$$

und folglich $(y_0,x)=0$ für alle $x\in N$. Nun können wir das gesuchte y angeben, und zwar setzen wir

$$y:=\frac{l(y_0)}{\|y_0\|^2}\,y_0.$$

Hierbei ist $l(y_0) \neq 0$ und folglich auch $l(y) \neq 0$, da $y_0 \notin N$. Um $l(x) = (y, x)$ für alle $x \in X$ zu zeigen, geben wir uns $x \in X$ beliebig vor. Dann ist

$$x = \underbrace{\left(x - \frac{l(x)}{l(y)}y\right)}_{\in N} + \frac{l(x)}{l(y)}y$$

und daher

$$(y, x) = \frac{l(x)}{l(y)}\|y\|^2 = l(x),$$

der Satz ist bewiesen. □

Bemerkungen: 1. Ohne es extra auszusprechen, haben wir einen ersten Existenzsatz für beste Approximierende bewiesen:

- *Ist M ein abgeschlossener linearer Teilraum in einem Hilbertraum $(X, (\cdot, \cdot))$, so existiert zu jedem $z \in X$ eine beste Approximierende in M.*

2. Aus dem Darstellungssatz von Riesz folgt:

- *Sei $(X, (\cdot, \cdot))$ ein Hilbertraum.*

  (a) *Es ist $x_k \rightharpoonup x$ genau dann, wenn $(y, x_k) \to (y, x)$ für alle $y \in X$.*

  (b) *Es ist $x_k \to x$ genau dann, wenn $x_k \rightharpoonup x$ und $\|x_k\| \to \|x\|$.*

□

Nun sei $(X, \|\cdot\|)$ wieder ein linearer normierter Raum. Aufbauend auf dem Begriff der schwachen Konvergenz definieren wir:

Definition 2.3.4 Eine Teilmenge $A \subset X$ heißt

(a) *schwach folgenabgeschlossen*, falls

$$\{x_k\} \subset A,\ x_k \rightharpoonup x \Longrightarrow x \in A,$$

(b) *schwach relativ folgenkompakt*, falls jede Folge $\{x_k\} \subset A$ eine schwach konvergente Teilfolge enthält,

(c) *schwach folgenkompakt*, falls jede Folge $\{x_k\} \subset A$ eine gegen ein $x \in A$ schwach konvergente Folge enthält.

Bemerkung: Offenbar ist eine schwach folgenabgeschlossene Menge abgeschlossen und eine (folgen)kompakte Menge auch schwach folgenkompakt. □

Eine hübsche Anwendung des strikten Trennungssatzes ist

Satz 2.3.5 *$A \subset X$ sei nichtleer, abgeschlossen und konvex. Dann ist A schwach folgenabgeschlossen.*

Beweis: Sei $\{x_k\} \subset A$ und $x_k \rightharpoonup x$. Angenommen, es ist $x \notin A$. Aus dem strikten Trennungssatz 2.2.7 folgt die Existenz von $(l, \gamma) \in (X^* \setminus \{0\}) \times \mathbb{R}$ mit $l(x) < \gamma < l(a)$ für alle $a \in A$. Insbesondere ist $l(x) < \gamma < l(x_k)$, wegen $x_k \rightharpoonup x$ ist $l(x) < \gamma \leq l(x)$, ein Widerspruch. □

Insbesondere ist die Einheitskugel $B[0;1]$ schwach folgenabgeschlossen. I. Allg. ist sie aber nicht schwach folgenkompakt: Denn sei z. B. $(X, \|\cdot\|) := (C[\alpha, \beta], \|\cdot\|_\infty)$. Definiert man

$$x_k(t) := \left(\frac{t-\alpha}{\beta-\alpha}\right)^k, \quad k \in \mathbb{N},$$

so ist $\{x_k\} \subset B[0;1]$. Angenommen, es existiert eine Teilfolge $\{x_{k_j}\} \subset \{x_k\}$ und ein $x \in C[\alpha, \beta]$ mit $x_{k_j} \rightharpoonup x$. Für $t \in [\alpha, \beta]$ definiere man $l_t \in X^*$ durch $l_t(z) := z(t)$. Dann gilt

$$x(t) = l_t(x) \leftarrow l_t(x_k) \rightarrow \begin{cases} 0, & t \in [\alpha, \beta), \\ 1, & t = \beta, \end{cases}$$

ein Widerspruch zur Stetigkeit von x. Wegen dieses Beispiels ist die abgeschlossene Einheitskugel also nicht in jedem linearen normierten Raum schwach folgenkompakt. Andererseits ist diese Aussage in einigen wichtigen Räumen richtig. Hierzu benötigen wir:

Definition 2.3.6 Ein linearer normierter Raum $(X, \|\cdot\|)$ heißt *reflexiv*, falls die sogenannte kanonische Abbildung $i\colon X \longrightarrow X^{**}$ von X in den *Bidualraum* $X^{**} := (X^*)^*$, definiert durch $i(x)(l) := l(x)$, surjektiv ist.

Bemerkung: 1. Sei $(X, \|\cdot\|)$ ein linearer normierter Raum und $i\colon X \longrightarrow X^{**}$ die durch $i(x)(l) := l(x)$ definierte kanonische Abbildung. Dass i linear und isometrisch ist (d. h. $\|i(x)\| = \|x\|$ für alle $x \in X$), hatten wir uns früher schon überlegt. Daher kann man X mit dem linearen Teilraum $i(X)$ von X^{**} identifizieren. Ist X reflexiv, so kann man X mit X^{**} identifizieren.

2. Da der Dualraum $(X^*, \|\cdot\|)$ eines linearen normierten Raumes vollständig ist (Beweis?), ist ein reflexiver linearer normierter Raum ein Banachraum. □

Beispiele: 1. Jeder endlichdimensionale lineare normierte Raum ist reflexiv.

2. Jeder Hilbertraum $(X, (\cdot,\cdot))$ ist reflexiv, da ja wegen des Rieszschen Darstellungssatzes schon $X = X^*$.

3. Für $1 < p < \infty$ ist $(L^p[\alpha, \beta])^{**} = (L^q[\alpha, \beta])^* = L^p[\alpha, \beta]$ mit $1/p + 1/q = 1$. Also sind $L^p[\alpha, \beta]$ und entsprechend l^p für $1 < p < \infty$ reflexiv. $L^1[\alpha, \beta]$ und $L^\infty[\alpha, \beta]$ sind dagegen nicht reflexiv, ebenso wenig wie $(C[\alpha, \beta], \|\cdot\|_\infty)$. □

Den folgenden Satz, gelegentlich benannt nach Eberlein-Šmulian, können wir hier nicht beweisen. Einen Beweis kann man z. B. bei F. Hirzebruch, W. Scharlau (1971, S. 68) finden.

Satz 2.3.7 *Die abgeschlossene Einheitskugel $B[0;1]$ ist in einem reflexiven Banachraum schwach folgenkompakt.*

Insbesondere ist jede beschränkte Menge, d. h. eine Menge, die in einer hinreichend großen Kugel enthalten ist, in einem reflexiven linearen normierten Raum schwach relativ folgenkompakt.

Zum Schluss wollen wir noch auf die schwach-$*$-Konvergenz in X^* eingehen. Auf dem Dualraum X^* eines linearen normierten Raumes kann man nämlich neben der starken und der schwachen Konvergenz einen weiteren Konvergenzbegriff erklären.

Definition 2.3.8 Sei $(X, \|\cdot\|)$ ein linearer normierter Raum. Eine Folge $\{l_k\} \subset X^*$ heißt *schwach-$*$-konvergent* gegen $l \in X^*$, wofür wir auch $w\text{-}* \lim_{k\to\infty} l_k = l$ oder $l_k \underset{*}{\rightharpoonup} l$ schreiben, falls

$$\lim_{k\to\infty} l_k(x) = l(x) \quad \text{für alle } x \in X.$$

Bemerkungen: 1. Natürlich gelten die üblichen Rechenregeln für Limiten. Ferner ist der schwach-$*$-Limes einer schwach-$*$-konvergenten Folge eindeutig bestimmt.

2. Seien $\{l_k\} \subset X^*$ und $l \in X^*$. Dann gelten die Implikationen

$$l_k \to l \Longrightarrow l_k \rightharpoonup l \Longrightarrow l_k \underset{*}{\rightharpoonup} l.$$

Denn die erste Implikation ist sowieso klar. Zur zweiten beachte man, dass man X mittels der kanonischen Abbildung $i\colon X \longrightarrow X^{**}$, $i(x)(l) := l(x)$, in X^{**} einbetten kann. Aus $l_k \rightharpoonup l$ folgt $i(x_k)(l_k) \to i(x)(l)$ für jedes $x \in X$ bzw. $l_k \underset{*}{\rightharpoonup} l$. Ist $(X, \|\cdot\|)$ reflexiv, so ist i surjektiv und jedes Element aus X^{**} ist durch $i(x)$ mit einem gewissen $x \in X$ gegeben. Hier gilt also auch die Umkehrung, dass nämlich $l_k \underset{*}{\rightharpoonup} l$ die schwache Konvergenz $l_k \rightharpoonup l$ impliziert. □

Unser letztes Ergebnis in diesem Abschnitt ist ein tiefer Satz, den wir nur in einem Spezialfall beweisen wollen, siehe z. B. F. Hirzebruch, W. Scharlau (1971, S. 64).

Satz 2.3.9 (Banach-Alaoglu) *Sei $(X, \|\cdot\|)$ ein linearer normierter Raum. Dann ist die abgeschlossene Einheitskugel $B^*[0;1] := \{l \in X^* : \|l\| \le 1\}$ schwach-$*$-folgenkompakt, d. h. aus jeder Folge $\{l_k\} \subset B^*[0;1]$ ist eine gegen ein $l \in B^*[0;1]$ schwach-$*$-konvergente Teilfolge $\{x_{k_j}\}$ auswählbar.*

Beweis: Wir setzen (unnötigerweise) voraus, dass X *separabel* ist, d. h. dass in X eine abzählbare, dichte[4] Teilmenge $\{x_k\}$ existiert. Sei $\{l_k\} \subset B^*[0;1]$ beliebig. Die Folge $\{l_k(x_1)\} \subset \mathbb{R}$ ist beschränkt, da $|l_k(x_1)| \le \|x_1\|$, und enthält daher eine konvergente Teilfolge $\{l_{k1}(x_1)\}$. Ebenso ist $\{l_{k1}(x_1)\}$ beschränkt, enthält also eine konvergente Teilfolge $\{l_{k2}(x_1)\}$. So fahre man fort und bilde die Diagonalfolge $\{l_{kk}\} \subset B^*[0;1]$. Nach Konstruktion existiert $\lim_{k\to\infty} l_{kk}(x_j)$ für alle $j \in \mathbb{N}$. Nun zeigen wir, dass $\lim_{k\to\infty} l_{kk}(x)$ sogar für alle $x \in X$ existiert. Sei hierzu $x \in X$ beliebig, $\epsilon > 0$ vorgegeben. Da $\{x_k\} \subset X$ dicht ist, gibt es ein $j \in \mathbb{N}$ mit $\|x - x_j\| \le \epsilon/3$. Da $\{l_{kk}(x_j)\}$ als konvergente Folge eine Cauchy-Folge ist, gibt es ein $K(\epsilon) \in \mathbb{N}$ mit

$$|l_{kk}(x_j) - l_{pp}(x_j)| \le \frac{\epsilon}{3} \quad \text{für } k, p \ge K(\epsilon).$$

[4] Zur Erinnerung: Eine Teilmenge $A \subset X$ heißt *dicht*, in X, wenn $\mathrm{cl}\,(A) = X$.

Für alle $k, p \geq K(\epsilon)$ ist daher

$$\begin{aligned} |l_{kk}(x) - l_{pp}(x)| &\leq |l_{kk}(x - x_j)| + |(l_{kk} - l_{pp}(x_j)| + |l_{pp}(x - x_j)| \\ &\leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} \\ &= \epsilon. \end{aligned}$$

Also ist $\{l_{kk}(x)\} \subset \mathbb{R}$ für jedes $x \in X$ eine Cauchy-Folge und als solche konvergent. Nun definiere man $l\colon X \longrightarrow \mathbb{R}$ durch

$$l(x) := \lim_{k\to\infty} l_{kk}(x).$$

Offenbar ist $l \in B^*[0;1]$ und $l_{kk} \underset{*}{\rightharpoonup} l$. □

Bemerkung: Mit obigem Beweis haben wir für separable Hilberträume auch Satz 2.3.7 bewiesen. □

Kapitel 3

Approximationstheorie in linearen normierten Räumen

In diesem Kapitel wollen wir noch nicht auf konkrete Approximationsaufgaben eingehen, sondern Approximationstheorie in einem funktionalanalytischen Rahmen betreiben. Wir wollen uns also z. B. fragen, unter welchen Voraussetzungen an den Raum $(X, \|\cdot\|)$ und die Teilmenge $M \subset X$ für jedes $z \in X$ mindestens (oder genau) eine beste Approximierende $x^* \in M$ existiert. Hierbei werden möglichst allgemeine Konzepte verfolgt, die erst später auf so interessante Spezialfälle wie rationale Approximation und Approximation durch Exponentialsummen angewandt werden, wobei die Konzepte i. Allg. geeignet modifiziert werden.

3.1 Existenz- und Eindeutigkeitsaussagen

Grundsätzlich sei in diesem Abschnitt $(X, \|\cdot\|)$ ein (reeller) linearer normierter Raum, $M \subset X$ und $z \in X$. Wir haben in der Einleitung schon definiert, was eine beste Approximierende an z in M ist. Wir wollen dies wiederholen und einige weitere grundlegende Definitionen anfügen.

Definition 3.1.1 (a) Ein $x^* \in M$ mit $\|x^* - z\| \leq \|x - z\|$ für alle $x \in M$ bzw.

$$d(z, M) := \inf_{x \in M} \|x - z\| = \|x^* - z\|$$

heißt *beste Approximierende* an z in M, $d(z, M)$ der (Minimal)*Abstand* von z zu M.

(b) Die Teilmenge $M \subset X$ heißt *Existenzmenge* (oder *proximinal*), wenn es zu jedem $z \in X$ (mindestens) eine beste Approximierende an z in M gibt.

(c) Die Teilmenge $M \subset X$ heißt *Tschebyscheff-Menge* oder kurz *T-Menge*, wenn es zu jedem $z \in X$ genau eine beste Approximierende an z in M gibt.

(d) Die Abbildung $P_M \colon X \longrightarrow 2^X$ (=Menge der Teilmengen von X), definiert durch

$$P_M(z) := \left\{ x^* \in M : \|x^* - z\| = \inf_{x \in M} \|x - z\| \right\}$$

heißt *metrische Projektion* von X auf M.

Die Menge M ist also proximinal, falls $P_M(z) \neq \emptyset$ für alle $z \in X$. Ferner ist M eine T-Menge, falls $P_M(z)$ für alle $z \in X$ einpunktig ist.

In diesem Abschnitt wollen wir untersuchen, unter welchen voraussetzungen an $(X, \|\cdot\|)$ und M die Menge M eine Existenzmenge oder sogar eine T-Menge ist. Um nicht gleich am anfang zu viele Begriffe einführen zu müssen, beginnen wir mit den einfachsten, und wahrscheinlich schon bekannten Ergebnissen. Diese werden später zum Teil wesentlich verallgemeinert. Wir sammeln zunächst einige triviale Ergebnisse, die so einfach sind, dass wir sie noch nicht einmal als Lemma oder Satz formulieren.

(a) Ist M kompakt, so ist M eine Existenzmenge.

Denn: Die Abbildung $x \mapsto \|x-z\|$ ist für jedes $z \in X$ stetig (Beweis?) und nimmt daher auf der kompakten Menge M ihr Minimum an.

(b) Ist $M \subset X$ eine Existenzmenge, so ist M abgeschlossen.

Denn: Sei $z \in \mathrm{cl}\,(M)$. Dann ist $\inf_{x\in M} \|x - z\| = 0$. Andererseits existiert nach Voraussetzung eine beste Approximierende $x^* \in M$ an z in M. Dann ist $\|x^*-z\| = 0$ und folglich $z \in M$.

Andererseits existiert die Abgeschlossenheit und Beschränktheit von M in unendlichdimensionalen linearen normierten Räumen *nicht*, dass M eine Existenzmenge ist. Hierzu geben wir ein Beispiel an.

Beispiel: Sei $(X, \|\cdot\|)$ ein unendlichdimensionaler linearer normierter Raum. Wie wir beim Beweis von Satz 2.3.1 (Folgerung aus dem Lemma von Riesz) gezeigt haben, existiert eine Folge $\{x_k\} \subset X$ mit $\|x_k\| = 1$ und $\|x_k - x_j\| \geq \frac{1}{2}$ für $k \neq j$. Sei

$$M := \left\{ \left(1 + \frac{1}{k}\right) x_k : k \in \mathbb{N} \right\}.$$

Dann ist M offenbar beschränkt (es ist $\|x\| \leq 2$ für jedes $x \in M$) und auch abgeschlossen (Beweis?). Es gibt aber keine beste Approximierende an 0 in M. □

(c) Ist $M \subset X$ abgeschlossen, so ist $P_M(z)$ für jedes $z \in X$ abgeschlossen.

Denn: Sei $\{x_k^*\} \subset P_M(z)$ eine Folge bester Approximierender an z in M und $x_k^* \to x^*$. Dann ist wegen der Stetigkeit der Norm

$$\|x^* - z\| \leftarrow \|x_k^* - z\| = \inf_{x\in M} \|x - z\|.$$

Also ist $\|x^* - z\| = \inf_{x\in M} \|x - z\|$. Wegen $\{x_k^*\} \subset M$ und der Abgeschlossenheit von M ist $x^* \in M$ und daher $x^* \in P_M(z)$.

(d) Ist $M \subset X$ konvex, so ist $P_M(z)$ für jedes $z \in X$ konvex.

Denn: Seien $x^*, y^* \in P_M(z)$ und $\lambda \in [0,1]$. Dann ist auch $(1-\lambda)x^* + \lambda y^* \in M$ und

$$\begin{aligned} \|(1-\lambda)x^* + \lambda y^* - z\| &= \|(1-\lambda)(x^* - z) + \lambda(y^* - z)\| \\ &\leq (1-\lambda)\|x^* - z\| + \lambda\|y^* - z\| \\ &= \inf_{x \in M} \|x - z\|. \end{aligned}$$

Also ist $(1-\lambda)x^* + \lambda y^* \in P_M(z)$ und $P_M(z)$ konvex.

Es folgen zwei einfache Existenzsätze.

Satz 3.1.2 *Ein endlichdimensionaler linearer Teilraum $M \subset X$ ist eine Existenzmenge.*

Beweis: Sei $z \in X$ beliebig. Dann ist $M_0 := M \cap B[0; 2\|z\|]$ kompakt (Beweis?), es existiert also eine beste Approximierende $x^* \in M_0$ an z in M_0. Ist $x \in M \setminus B[0; 2\|z\|]$, so ist

$$\|x - z\| \geq \|x\| - \|z\| > 2\|z\| - \|z\| = \|0 - z\| \geq \|x^* - z\|,$$

also ist x^* auch beste Approximierende an z in M, die Behauptung ist bewiesen. □

Bemerkungen: Durch die letzte Aussage ist die Existenz einer Lösung der meisten wichtigen *linearen* Approximationsaufgaben gesichert. Ist z. B. $X := C[\alpha, \beta]$ versehen mit der Maximumnorm $\|x\|_\infty := \max_{t \in [\alpha,\beta]} |x(t)|$, und $M := \Pi_n$ der $(n+1)$-dimensionale lineare Raum der Polynome vom Grad $\leq n$, so ist M eine Existenzmenge (sogar eine Tschebyscheff-Menge, wie wir später sehen werden).

Die Voraussetzung *endlichdimensional* in der letzten Aussage darf man nicht weglassen, wie das Beispiel $(X, \|\cdot\|) := (C[\alpha,\beta], \|\cdot\|_\infty)$ mit $M := \Pi$ der Menge aller Polynome zeigt. I. Allg. kann man auch nicht *endlichdimensional* durch *abgeschlossen* ersetzen. Hierzu geben wir ein Beispiel an (siehe D. Braess (1986, S. 25)). Sei

$$c_0 := \{x = \{x_k\} \subset \mathbb{R} : \lim_{k \to \infty} x_k = 0\}$$

der lineare Raum der reellen Nullfolgen. Auf c_0 definiere man eine Norm durch

$$\|x\|_\infty := \sup_{k \in \mathbb{N}} |x_k|.$$

Dann ist $(c_0, \|\cdot\|_\infty)$ ein lineare normierter Raum (sogar ein Banachraum). Nun definieren wir $l \colon c_0 \longrightarrow \mathbb{R}$ durch

$$l(x) := \sum_{k=1}^{\infty} 2^{-k} x_k$$

und anschließend

$$M := \{x \in c_0 : l(x) = 0\}.$$

Da l linear und wegen

$$|l(x)| \leq \left(\sum_{k=1}^{\infty} 2^{-k} \right) \|x\|_\infty = \|x\|_\infty, \quad x \in c_0,$$

auch stetig ist, ist M eine linearer, abgeschlossener Teilraum von c_0. Wir wollen uns überlegen, dass $P_M(z) = \emptyset$ für jedes $z \in c_0 \setminus M$. Hierzu gebe man sich ein beliebiges $z \in c_0 \setminus M$ vor und setze $\lambda := l(z)$. Wegen $z \notin M$ ist $\lambda \neq 0$. Zunächst zeigen wir, dass $d(z, M) \leq |\lambda|$. Hierzu definiere man

$$u^k := z - (1 - 2^{-k})^{-1}(\underbrace{\lambda, \ldots, \lambda}_{k}, 0, 0, \ldots).$$

Dann ist $u^k \in c_0$ und

$$l(u^k) = l(z) - (1 - 2^{-k})^{-1}\lambda \sum_{j=1}^{k} 2^{-j} = \lambda - (1 - 2^{-k})^{-1}\lambda(1 - 2^{-k}) = 0$$

und folglich $u^k \in M$. Daher ist

$$d(z, M) \leq \|u^k - z\|_\infty = (1 - 2^{-k})^{-1}|\lambda|.$$

Mit $k \to \infty$ folgt $d(z, M) \leq |\lambda|$. Wir machen einen Widerspruchsbeweis und nehmen an, dass eine beste Approximierende $x^* \in M$ an z in M existiert. Da auch $x^* - z$ eine Nullfolge ist, gibt es ein $K \in \mathbb{N}$ mit $|x_k^* - z_k| \leq \frac{1}{2}|\lambda|$ für alle $k \geq K$. Unter Berücksichtigung der schon bewiesenen Ungleichung $\|x^* - z\| = d(z, M) \leq |\lambda|$ erhalten wir

$$\begin{aligned}
|\lambda| &= |l(x^* - z)| \\
&= \left|\sum_{k=1}^{\infty} 2^{-k}(x_k^* - z_k)\right| \\
&= \left|\sum_{k=1}^{K-1} 2^{-k}(x_k^* - z_k) + \sum_{k=K}^{\infty} 2^{-k}(x_k^* - z_k)\right| \\
&\leq \Bigg(\underbrace{\sum_{k=1}^{K-1} 2^{-k}}_{=1-(1/2)^{K-1}}\Bigg)|\lambda| + \Bigg(\underbrace{\sum_{k=K}^{\infty} 2^{-k}}_{=(1/2)^{K-1}}\Bigg)\frac{1}{2}|\lambda| \\
&= \left[1 - \left(\frac{1}{2}\right)^K\right]|\lambda] \\
&< |\lambda|,
\end{aligned}$$

ein Widerspruch. Damit ist gezeigt, dass ein abgeschlossener linearer Teilraum eines linearen normierten Raumes i. Allg. keine Existenzmenge ist.

Unter den Voraussetzungen von Satz 3.1.2 kann man nicht erwarten, dass M sogar eine Tschebyscheff-Menge ist. Die Abbildung 3.1 soll dies verdeutlichen. Hier ist

$$(X, \|\cdot\|) := (\mathbb{R}^2, \|\cdot\|_\infty), \qquad M := \{(x_1, x_2) \in \mathbb{R}^2 : x_1 = 0\}$$

und $z := (z_1, 0)$ mit $z_1 > 0$. Hier ist

$$P_M(z) = \{(0, t) : |t| \leq \|z\|_\infty\}.$$

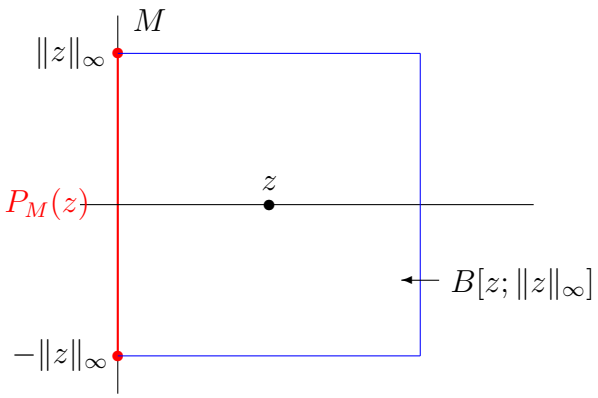

Abbildung 3.1: Ein endlichdimensionaler Teilraum ist i. Allg. keine T-Menge

Man ahnt, dass Eindeutigkeitsaussagen etwas mit der "Krümmung" von Kugeln in $(X, \|\cdot\|)$ zu tun haben könnten. □

Der folgende Satz ist der wohlbekannte Projektionssatz für abgeschlossene konvexe Mengen in einem Hilbertraum. Auch wenn dies die Zielsetzung dieses Abschnitts überschreitet, ist in ihm für einen speziellen Fall fast alles an Aussagen enthalten, was man sich in diesem Zusammenhang nur wünscht, nämlich:

1. Existenz und Eindeutigkeit einer Lösung,
2. Charakterisierung einer Lösung,
3. Stetige Abhängigkeit der Lösung von dem zu approximierenden Element.

Satz 3.1.3 (Projektionssatz) *Sei $(X, (\cdot,\cdot))$ ein Hilbertraum, $M \subset X$ eine nichtleere, abgeschlossene, konvexe Menge. Dann gilt:*

1. *M ist eine T-Menge.*

   Für jedes $z \in X$ gibt es also genau eine beste Approximierende $x^* = P_M(z)$ an z in M. Hierbei wird die metrische Projektion auf eine T-Menge sinnvollerweise als eine Abbildung von X nach M und nicht nach 2^M aufgefasst.

2. *Es ist $x^* = P_M(z)$ bzw. x^* die beste Approximierende an z in M genau dann, wenn*

   $$(z - x^*, x - x^*) \leq 0 \quad \text{für alle } x \in M.$$

3. *Für beliebige $y, z \in X$ ist*

   $$\|P_M(y) - P_M(z)\| \leq \|y - z\|,$$

   insbesondere ist die metrische Projektion $P_M \colon X \longrightarrow X$ stetig.

Beweis: 1. Sei $z \in X$. Wir zeigen zunächst die Existenz einer besten Approximierenden an z in M, anschließend die Eindeutigkeit.

(a) Sei $\{x_k\} \subset M$ eine Minimalfolge, d. h. es ist $\|x_k - z\| \to d(z, M)$. Dann ist $\{x_k\}$ eine Cauchy-Folge, denn wegen der Parallelogrammgleichung ist

$$\begin{aligned} \|x_k - x_l\|^2 &= 2\|x_k - z\|^2 + 2\|x_l - z\|^2 - 4\left\|\frac{x_k + x_l}{2} - z\right\|^2 \\ &\leq 2\|x_k - z\|^2 + 2\|x_l - z\|^2 - 4\,d(z, M)^2 \\ &\to 0. \end{aligned}$$

Wegen der Konvexität von M ist nämlich $\frac{1}{2}(x_k + x_l) \in M$ und $\|\frac{1}{2}(x_k + x_l) - z\| \geq d(z, M)$. Da X nach Voraussetzung vollständig ist, konvergiert die Folge $\{x_k\}$ gegen ein $x^* \in X$, wegen der Abgeschlossenheit von M ist $x^* \in M$ und wegen der Stetigkeit der Norm ist $\|x^* - z\| = d(z, M)$, also x^* eine beste Approximierende an z in M.

(b) Seien $x_1^*, x_2^* \in M$ zwei beste Approximierende an z in M. Wir definieren die Folge $\{x_k\}$ durch

$$x_k := \begin{cases} x_1^*, & k \text{ gerade}, \\ x_2^*, & k \text{ ungerade}. \end{cases}$$

Wegen $\|x_k - z\| = d(z, M)$, $k \in \mathbb{N}$, ist $\{x_k\}$ eine Minimalfolge, wegen (a) eine Cauchy-Folge und damit konvergent. Hieraus folgt $x_1^* = x_2^*$.

2. Der Beweis für die Charakterisierung der besten Approximierenden zerfällt in zwei Teile.

(a) (Notwendige Optimalitätsbedingung).

Sei $x^* \in M$ die beste Approximierende an z in M und $x \in M$ beliebig. Wegen der Konvexität von M ist $x^* + t(x - x^*) \in M$ für alle $t \in (0, 1]$. Für $t \in (0, 1]$ ist daher

$$\begin{aligned} 0 &\leq \frac{1}{t}[\|x^* + t(x - x^*) - z\|^2 - \|x^* - z\|^2] \\ &= 2\,(x^* - z, x - x^*) + t\,\|x - x^*\|^2. \end{aligned}$$

Mit $t \to 0+$ ist also

$$(z - x^*, x - x^*) \leq 0 \quad \text{für alle } x \in M.$$

(b) (Hinreichende Optimalitätsbedingung).

Sei $x^* \in M$ und

$$(z - x^*, x - x^*) \leq 0 \quad \text{für alle } x \in M.$$

Für beliebiges $x \in M$ ist dann

$$\begin{aligned} \|x - z\|^2 &= \|(x - x^*) + (x^* - z)\|^2 \\ &= \|x - x^*\|^2 + 2\underbrace{(x^* - z, x - x^*)}_{\geq 0} + \|x^* - z\|^2 \\ &\geq \|x - x^*\|^2 + \|x^* - z\|^2 \\ &\geq \|x^* - z\|^2, \end{aligned}$$

also x^* eine beste Approximierende an z in M.

3. Seien $y, z \in X$, $P_M(y)$ bzw. $P_M(z)$ seien die zugehörigen besten Approximierenden. Wegen 2. (a) gelten die Ungleichungen

$$(y - P_M(y), P_M(z) - P_M(y)) \leq 0, \qquad (z - P_M(z), P_M(y) - P_M(z)) \leq 0.$$

Eine Addition dieser beiden Ungleichungen ergibt

$$(P_M(y) - P_M(z) - (y - z), P_M(y) - P_M(z)) \leq 0$$

bzw.

$$\begin{aligned} \|P_M(y) - P_M(z)\|^2 &\leq (y - z, P_M(y) - P_M(z)) \\ &\leq \|y - z\| \, \|P_M(y) - P_M(z)\| \end{aligned}$$

und hieraus folgt

$$\|P_M(y) - P_M(z)\| \leq \|y - z\|,$$

die Behauptung. □

Bemerkung: Teile des letzten Satzes werden wir im folgenden noch wesentlich verallgemeinern und uns vor allem von der Voraussetzung lösen, dass X ein Hilbertraum ist. Die Charakterisierung 2. einer besten Approximierenden lässt sich sehr schön geometrisch veranschaulichen, siehe Abbildung 3.2. Die Bedingung $(z - P_M(z), x - P_M(z)) \leq 0$

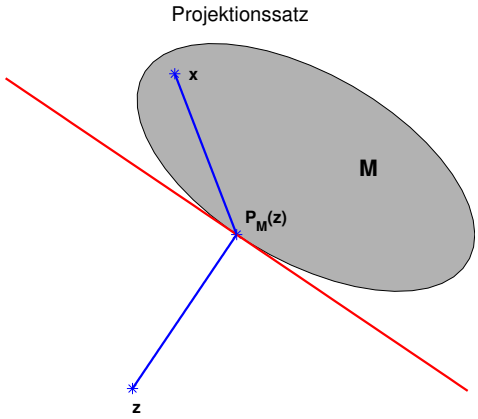

Abbildung 3.2: Charakterisierung einer besten Approximierenden

sagt aus, dass der cos des Winkels zwischen $z - P_M(z)$ und $x - P_M(z)$ nichtnegativ ist bzw. der Winkel zwischen $\pi/2$ und $3\pi/2$ liegt. Für affin lineares M (bzw. einen verschobenen linearen Teilraum) bedeutet 2., dass $P_M(z)$ durch $(z - P_M(z), x - P_M(z)) = 0$ für alle $x \in M$ charakterisiert ist. Ist dagegen M ein (abgeschlossener) linearer Teilraum, so ist $P_M(z)$ dadurch charakterisiert, dass $z - P_M(z)$ auf M senkrecht steht bzw. $(z - P_M(z), x) = 0$ für alle $x \in M$ gilt. In Abbildung 3.3 veranschaulichen wir beide Fälle. Die Konsequenzen dieser Beziehung sind in Spezialfällen sicher bekannt, es sei nur das Stichwort "Normalgleichungen" genannt. □

Beispiel: Eine chemische Mischung A werde während eines festen Zeitintervalls $[0, T]$ einer Flüssigkeit in einem Tank zugeführt. Ein gewisser kritischer Wert x der hierdurch

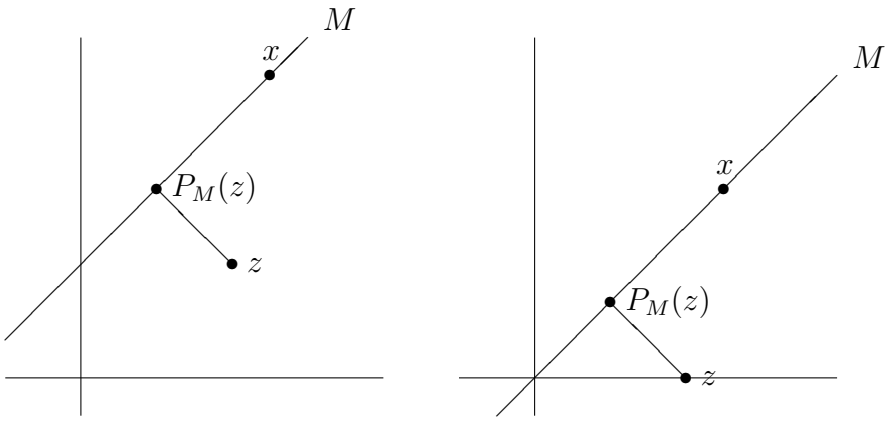

Abbildung 3.3: Die Projektion auf einen (affin)linearen Teilraum

entstehenden Mischung werde bestimmt durch die Stärke u einer Komponente von A, die zeitlich geändert werden kann. Die zeitliche Änderung von x sei linear in x und u, genauer gelte

$$(*) \qquad \dot{x} = \alpha x + \beta u$$

mit $\alpha, \beta > 0$. ZurAnfangszeit 0 sei der kritische Wert bekannt:

$$(**) \qquad x(0) = x_0.$$

Als Nebenbedingung könnte noch sinnvoll sei, dass im ganzen Zeitintervall der kritische Wert ein gewisses Plateau p nicht unterschreitet:

$$x(t) \geq p \quad \text{für alle } t \in [0, T].$$

Ferner stehe u nur bis zu einer gewissen Maximalstärke zur Verfügung, es sei also $0 \leq u(t) \leq \gamma$ für $t \in [0, T]$. Schließlich seien die Kosten proportional zu $\int_0^T u(t)^2\, dt$ und diese seien unter den angegebenen Nebenbedingungen zu minimieren. Die Lösung von $(*)$, $(**)$ ist gegeben durch

$$x(t) = e^{\alpha t} x_0 + \int_0^t e^{\alpha(t-s)} \beta u(s)\, ds.$$

Als Ausgangsraum nehme man den Hilbertraum $L^2[0, T]$ mit dem üblichen inneren Produkt und der zugehörigen Norm. Dann hat man also die Aufgabe

$$\text{(P)} \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad \|u\|_2 = (\int_0^T u(t)^2\, dt)^{1/2} \quad \text{auf} \\ M := \left\{ u \in L^2[0,T] : \begin{array}{l} x_0 + \beta \int_0^t e^{-s} u(s)\, ds \geq e^{-\alpha t} p \quad \text{für } t \in [0,T] \\ 0 \leq u(t) \leq \gamma \quad \text{fast überall auf } [0,T]. \end{array} \right\} \end{array} \right.$$

Die Menge M ist offensichtlich konvex, aber auch abgeschlossen (Beweis?, nicht ganz einfach). Ist also $M \neq \emptyset$, so besitzt das angegebene Problem (P) eine eindeutige Lösung. □

Der folgende Satz ist eine Verallgemeinerung der Existenzaussage in Satz 3.1.3, dem Projektionssatz.

Satz 3.1.4 *Sei $(X, \|\cdot\|)$ ein reflexiver Banachraum. Eine nichtleere, abgeschlossene, konvexe Menge $M \subset X$ ist eine Existenzmenge.*

Beweis: Sei $z \in X$ und $\{x_k\} \subset M$ eine Minimalfolge, also $\|x_k - z\| \to d(z, M)$. Dann ist $\{x_k\}$ beschränkt, wegen Satz 2.3.7 kann aus $\{x_k\}$ eine schwach konvergente Teilfolge ausgewählt werden: $x_{k_j} \to x^*$. Als abgeschlossene, konvexe Menge ist M nach Satz 2.3.5 schwach folgenabgeschlossen. Also ist $x^* \in M$. Zu zeigen bleibt $\|x^* - z\| = d(z, M)$. Auch hierzu wenden wir auf geschickte Weise Satz 2.3.5 an. Bei festem $\epsilon > 0$ ist $x_{k_j} \in B[z; d(z, M) + \epsilon]$ für fast alle j. Da $x_{k_j} \rightharpoonup x^*$ und $B[z; d(z, M) + \epsilon]$ als abgeschlossene, konvexe Menge schwach abgeschlossen ist, ist $x^* \in B[z; d(z, M) + \epsilon]$. Dies gilt für jedes $\epsilon > 0$, also ist $\|x^* - z\| \leq d(z, M)$ bzw. $\|x^* - z\| = d(z, M)$. □

Wenn man sich die Existenzbeweise zu Satz 3.1.3, dem Projektionssatz, und Satz 3.1.4 noch einmal ansieht, erkennt man eine Gemeinsamkeit: man geht bei festem $z \in X$ aus von einer Minimalfolge $\{x_k\} \subset M$, also $\|x_k - z\| \to d(z, M)$, und zeigt, dass diese eine konvergente Teilfolge enthält (im Beweis von Satz 3.1.3 konnte sogar die Konvergenz der *ganzen* Folge bewiesen werden) bzw. eine schwach konvergente Teilfolge enthält. Hieraus folgt jeweils, dass der entsprechende Limes beste Approximierende an z in M ist. Daher liegt die folgende Definition nahe.

Definition 3.1.5 Eine Menge $M \subset X$ heißt *approximativ kompakt* bzw. *approximativ schwach kompakt*, wenn es zu jedem $z \in X$ und jeder Minimalfolge $\{x_k\} \subset M$, also $\|x_k - z\| \to d(z, M)$, eine gegen ein $x^* \in M$ konvergente bzw. schwach konvergente Teilfolge $\{x_{k_j}\} \subset \{x_k\}$ gibt.

Offenbar gilt dann, wir wollen dies gar nicht als Satz formulieren, weil es offensichtlich ist: Eine approximativ kompakte bzw. approximativ schwach kompakte Menge $M \subset X$ ist eine Existenzmenge.

Bevor wir Beispiele und Gegenbeispiele zu approximativer Kompaktheit geben, zeigen wir im folgenden Satz einen interessanten Zusammenhang zwischen Existenz, Eindeutigkeit und der Stetigkeit der metrischen Projektion.

Satz 3.1.6 *Die metrische Projektion P_M auf eine approximativ kompakte T-Menge $M \subset X$ ist stetig.*

Beweis: Sei $\{z_k\} \subset X$ und $z_k \to z$. Zu zeigen ist $P_M(z_k) \to P_M(z)$. Nun ist

$$\begin{aligned} d(z, M) &\leq \|P_M(z_k) - z\| \\ &\leq \|P_M(z_k) - z_k\| + \|z_k - z\| \\ &\leq \|P_M(z) - z_k\| + \|z_k - z\| \\ &\leq \|P_M(z) - z\| + 2\,\|z_k - z\| \\ &= d(z, M) + 2\,\|z_k - z\|. \end{aligned}$$

Daher ist $\{P_M(z_k)\}$ eine Minimalfolge. Da M approximativ kompakt ist, kann aus $\{P_M(x_k)\}$ eine konvergente Teilfolge ausgewählt werden, deren Limes wegen der Eindeutigkeit der besten Approximierenden notwendig $P_M(z)$ ist. Wir überlegen uns, dass dann auch die gesamte Folge $\{P_M(z_k)\}$ gegen $P_M(z)$ konvergiert. Denn wäre dies nicht

der Fall, so gäbe es ein $\epsilon > 0$ und eine Teilfolge $\{P_M(z_{k_j})\}$ mit $\|P_M(z_{k_j}) - P_M(z)\| \geq \epsilon$ für alle j. Auch $\{P_M(z_{k_j})\}$ ist eine Minimalfolge, auch aus ihr ist eine konvergente Teilfolge auswählbar, die notwendig gegen $P_M(z)$ konvergiert, ein Widerspruch. □

Beim Beweis des ersten Teiles des Projektionssatzes 3.1.3 haben wir gezeigt, dass eine abgeschlossene, konvexe Menge in einem Hilbertraum eine approximativ kompakte T-Menge ist. Dieses Ergebnis kann man etwas allgemeiner fassen. Hierzu benötigen wir

Definition 3.1.7 Ein linearer normierter Raum $(X, \|\cdot\|)$ heißt

(a) *strikt konvex*, falls

$$\|x\| = \|y\| = 1,\ x \neq y \Longrightarrow \|\tfrac{1}{2}(x+y)\| < 1,$$

(b) *uniform konvex* (oder auch *gleichmäßig konvex*), falls es zu jedem $\epsilon > 0$ ein $\delta = \delta(\epsilon)$ mit

$$\|x\| = \|y\| = 1,\ \|x-y\| \geq \epsilon \Longrightarrow \|\tfrac{1}{2}(x+y)\| < 1-\delta$$

gibt.

Bemerkungen und Beispiele: 1. Offenbar gilt: Eine konvexe Existenzmenge in einem strikt konvexen Raum ist eine T-Menge.

2. Jeder uniform konvexe Raum ist offenbar strikt konvex. Ein *endlichdimensionaler* strikt konvexer Raum ist auch uniform konvex. Zur Begründung definiere man zu vorgegebenen $\epsilon > 0$ die Menge

$$S_\epsilon := \{(x,y) \in X \times X : \|x\| = \|y\| = 1,\ \|x-y\| \geq \epsilon\}.$$

Als abgeschlossene, beschränkte Menge des endlichdimensionalen Raumes $X \times X$ ist S_ϵ kompakt. Die Abbildung $\phi\colon S_\epsilon \longrightarrow \mathbb{R}$, definiert durch $\phi(x,y) := \|\frac{1}{2}(x+y)\|$, ist stetig und nimmt auf S_ϵ ihr Maximum an. Es existiert also $(x^*, y^*) \in S_\epsilon$ mit

$$\|\tfrac{1}{2}(x+y)\| \leq \|\tfrac{1}{2}(x^*+y^*)\| \quad \text{für alle } (x,y) \in S_\epsilon.$$

Wegen der strikten Konvexität ist $q := \|\frac{1}{2}(x^*+y^*)\| < 1$. Mit $\delta := 1-q$ folgt die Behauptung.

3. Wichtige Beispiele uniform konvexer Räume sind prä-Hilberträume. Zur Begründung sei $\epsilon > 0$ vorgegeben, ferner sei $\|x\| = \|y\| = 1$ und $\|x-y\| \geq \epsilon$. Wegen der Parallelogrammgleichung ist

$$\|\tfrac{1}{2}(x+y)\|^2 = 1 - \tfrac{1}{4}\|x-y\|^2 \leq 1 - \tfrac{1}{4}\epsilon^2$$

und daher

$$\|\tfrac{1}{2}(x+y)\| \leq (1-\tfrac{1}{4}\epsilon^2)^{1/2} = 1-\delta$$

mit

$$\delta := 1 - (1-\tfrac{1}{4}\epsilon^2)^{1/2}.$$

4. Weitere interessante Beispiele uniform konvexer Räume sind die Räume l^p, L^p für $1 < p < \infty$. Zum Nachweis unterscheidet man die Fälle $p \geq 2$ und $1 < p < 2$.

(a) Sei $p \geq 2$.

Zunächst zeigen wir: Für alle $x \in [0,1]$ ist

$$(*) \qquad \left(\frac{1+x}{2}\right)^p + \left(\frac{1-x}{2}\right)^p - \frac{1}{2}(1+x^p) \leq 0.$$

Für den Beweis folgen wir E. HEWITT, K. STROMBERG (1965, p. 223). Für $x = 0$ ist $(*)$ richtig, da

$$\left(\frac{1}{2}\right)^p + \left(\frac{1}{2}\right)^p - \frac{1}{2} = \frac{1}{2}(2^{-p+2} - 1) \leq 0$$

wegen $p \geq 2$. Für $x \in (0,1]$ definieren wir

$$\begin{aligned} \Phi(x) &:= \frac{2^p}{x^p}\left[\left(\frac{1+x}{2}\right)^p + \left(\frac{1-x}{2}\right)^p - \frac{1}{2}(1+x^p)\right] \\ &= \left(\frac{1}{x}+1\right)^p + \left(\frac{1}{x}-1\right)^p - 2^{p-1}\left(\frac{1}{x^p}+1\right). \end{aligned}$$

Zu zeigen ist $\Phi(x) \leq 0$ für $x \in (0,1]$. Offenbar ist $\Phi(1) = 0$. Wir sind daher fertig, wenn wir $\Phi'(x) \geq 0$ für $x \in (0,1)$ zeigen können. Nun ist

$$\Phi'(x) = -\frac{p}{x^{p+1}}\Big[\underbrace{(1+x)^{p-1} + (1-x)^{p-1} - 2^{p-1}}_{=:\,\Psi(x)}\Big].$$

Wir zeigen nun, dass $\Psi(x) \leq 0$ für $x \in (0,1]$ bzw. $\Phi'(x) \geq 0$ für $x \in (0,1)$ und $\Phi(x) \leq 0$ für $x \in (0,1]$. Es ist $\Psi(1) = 0$ und

$$\Psi'(x) = (p-1)[(1+x)^{p-2} - (1-x)^{p-2}] \geq 0 \quad \text{für } x \in (0,1]$$

wegen $p-2 \geq 0$, damit ist die Ungleichung $(*)$ bewiesen. Aus $(*)$ kann man leicht die *Clarksonsche Ungleichung* für $p \geq 2$ herleiten, nämlich

$$(**) \qquad \left\|\frac{x+y}{2}\right\|_p^p + \left\|\frac{x-y}{2}\right\|_p^p \leq \frac{1}{2}\|x\|_p^p + \frac{1}{2}\|y\|_p^p$$

für $x, y \in L^p[\alpha,\beta]$ (bzw. l^p). Zum Nachweis von $(**)$ zeigen wir, dass für $p \geq 2$ und alle $x, y \in \mathbb{R}$ die Ungleichung

$$(***) \qquad \left|\frac{x+y}{2}\right|^p + \left|\frac{x-y}{2}\right|^p \leq \frac{1}{2}|x|^p + \frac{1}{2}|y|^p$$

gilt. Hier kann man offenbar annehmen, dass $0 < |x| \leq |y|$. Ersetzt man x durch $|x|/|y| \in (0,1]$ in $(*)$, so erhält man nach Multiplikation mit $|y|^p$, dass

$$\begin{aligned} \frac{1}{2}|x|^p + \frac{1}{2}|y|^p &\geq \left(\frac{|x|+|y|}{2}\right)^p + \left(\frac{|y|-|x|}{2}\right)^p \\ &\geq \left|\frac{x+y}{2}\right|^p + \left|\frac{x-y}{2}\right|^p, \end{aligned}$$

womit auch $(***)$ bewiesen ist. Ersetzt man in $(***)$ nun x durch $x(t)$, y durch $y(t)$ bzw. x durch x_j, y durch y_j und integriert über $[\alpha, \beta]$ bzw. summiert, so erhält man die Clarksonsche Ungleichung in $L^p[\alpha, \beta]$ bzw. in l^p. Aus dieser erhält man leicht, dass die L^p- bzw. l^p-Räume uniform konvex sind für $2 \leq p < \infty$. Sei hierzu ein $\epsilon > 0$ und x, y mit $\|x\|_p = \|y\|_p = 1$ und $\|x - y\|_p \geq \epsilon$ vorgegeben. Wegen der Clarksonschen Ungleichung $(**)$ ist

$$\left\| \frac{1}{2}(x+y) \right\|_p^p \leq 1 - \left(\frac{\epsilon}{2}\right)^p$$

und folglich $\|\frac{1}{2}(x+y)\|_p \leq 1-\delta(\epsilon)$ mit $\delta(\epsilon) := 1-(1-(\epsilon/2)^p)^{1/p}$. Die Clarksonsche Ungleichung ersetzt also die für Prä-Hilberträume geltende Parallelogrammgleichung.

Ähnlich wie in (a) kann man für $1 < p < 2$ vorgehen. Allerdings ist der Beweis der ersten Ungleichung, die wir gleich angeben werden, merkwürdigerweise sehr viel komplizierter als im Fall $p \geq 2$.

(b) Sei $1 < p \leq 2$.

Mit $1/p + 1/q = 1$ gilt für alle $x \in [0, 1]$, dass

$$(*) \qquad (1+x)^q + (1-x)^q \leq 2(1+x^p)^{1/(p-1)}.$$

Den Beweis für diese Ungleichung wollen wir nicht führen. Wir verweisen lediglich auf E. Hewitt, K. Stromberg (1965, p. 225). Hieraus erhält man dann, ähnlich wie in (a), leicht die Clarksonsche Ungleichung für $1 < p \leq 2$, nämlich

$$\left\| \frac{x+y}{2} \right\|_p^q + \left\| \frac{x-y}{2} \right\|_p^q \leq \left[\frac{1}{2}\|x\|_p^p + \frac{1}{2}\|y\|_p^p \right]^{1/(p-1)}$$

für $x, y \in L^p[\alpha, \beta]$ (bzw. l^p). Hieraus wiederum erhält man wieder leicht, dass die L^p- bzw. l^p-Räume auch für $1 < p < 2$ uniform konvex sind.

5. Natürlich gibt es auch wichtige Räume, die nicht strikt konvex, geschweige denn uniform konvex sind.

(a) $(\mathbb{R}^n, \|\cdot\|_\infty)$ ist für $n \geq 2$ nicht strikt konvex (Beweis?).

(b) $(C[\alpha, \beta], \|\cdot\|_\infty)$ ist nicht strikt konvex (Beweis?).

6. Ein interessanter Satz von Milman sagt aus, dass ein uniform konvexer Banachraum reflexiv ist, siehe z. B. F. Hirzebruch, W. Scharlau (1971, S. 78). □

Als letzten Existenz- und Eindeutigkeitssatz formulieren wir einen Satz, der die entsprechenden Aussagen des Projektionssatzes für konvexe, abgeschlossene Mengen in einem Hilbertraum verallgemeinert.

Satz 3.1.8 *Eine abgeschlossene, konvexe Teilmenge M eines uniform konvexen Banachraumes $(X, \|\cdot\|)$ ist eine approximativ kompakte T-Menge.*

Beweis: Im ersten Teil des Beweises zeigen wir, dass M eine approximativ kompakte Existenzmenge ist. Sei $z \in X$ beliebig. O. B. A. ist $z \notin M$. Wegen der Abgeschlossenheit von M ist daher der Abstand von z zu M positiv: $d := d(z, M) > 0$. Sei $\{x_k\} \subset M$ eine Minimalfolge, also $\|x_k - z\| \to d$. Wir wollen mit Hilfe der uniformen Konvexität zeigen, dass $\{x_k\}$ eine konvergente Folge ist, wegen der Stetigkeit der Norm und der Abgeschlossenheit von M also gegen ein Element $x^* \in M$ mit $\|x^* - z\| = d$ konvergiert. Damit wird gezeigt sein, dass M eine approximativ kompakte Existenzmenge ist. Die Eindeutigkeit einer besten Approximierenden im zweiten Teil des Beweises zu zeigen wird dann einfach sein.

Nun zeigen wir die Konvergenz der Minimalfolge $\{x_k\}$. Sei

$$z_k := \frac{1}{d}(x_k - z), \qquad y_k := \frac{x_k - z}{\|x_k - z\|}$$

und daher $z_k - y_k \to 0$. Sei $\epsilon > 0$ vorgegeben und $\delta = \delta(\epsilon)$ wie bei der Definition der uniformen Konvexität bestimmt, als:

$$\|x\| = \|y\| = 1,\ \|x - y\| \geq \epsilon \Longrightarrow \|\tfrac{1}{2}(x + y)\| < 1 - \delta.$$

Es ist $\|y_k\| = \|y_p\| = 1$ und

$$\|\tfrac{1}{2}(z_k + z_p)\| = \frac{1}{d}\,\|\underbrace{\tfrac{1}{2}(x_k + x_p)}_{\in M} - z\| \geq 1$$

und daher

$$\begin{aligned} \|\tfrac{1}{2}(y_k + y_p)\| &= \|\tfrac{1}{2}(z_k + z_p) + \tfrac{1}{2}(y_k - z_k) + \tfrac{1}{2}(y_p - z_p)\| \\ &\geq 1 - \tfrac{1}{2}\|y_k - z_k\| - \tfrac{1}{2}\|y_p - z_p\|. \end{aligned}$$

Wegen $y_k - z_k \to 0$ existiert $K(\epsilon) \in \mathbb{N}$ mit $\|y_k - z_k\| < \delta(\epsilon)$ für alle $k \geq K(\epsilon)$. Für $k, p \geq K(\epsilon)$ ist daher $\|\frac{1}{2}(y_k + y_p)\| > 1 - \delta(\epsilon)$ und daher $\|y_k - y_p\| < \epsilon$. Also ist $\{y_k\} \subset X$ eine Cauchy-Folge. Da X ein Banachraum ist, ist $\{y_k\}$ gegen ein $y \in X$ konvergent und wegen $\|y_k\| = 1$ ist $\|y\| = 1$. Ferner ist

$$x_k = \|x_k - z\|\, y_k + z \to d\, y + z.$$

Insgesamt ist im ersten Teil des Beweises gezeigt, dass M eine approximativ kompakte Existenzmenge ist.

Im zweiten Teil des Beweises zeigen wir, dass eine beste Approximierende an z in M eindeutig ist. Seien $x_1^*, x_2^* \in M$ zwei beste Approximerende an z in M, also

$$\|x_1^* - z\| = \|x_2^* - z\| = d$$

bzw.

$$\|(x_1^* - z)/d\| = \|(x_2^* - z)/d\| = 1.$$

Ferner ist

$$\|\tfrac{1}{2}[(x_1^* - z)/d + (x_2^* - z)/d]\| = \|[\underbrace{\frac{1}{2}(x_1^* + x_2^*)}_{\in M} - z]/d\| \geq 1.$$

Aus der strikten Konvexität folgt $(x_1^*-z)/d = (x_2^*-z)/d$ und damit $x_1^* = x_2^*$. Insgesamt ist alles bewiesen. □

Bemerkung: Wegen Satz 3.1.6 und Satz 3.1.8 ist die metrische Projektion auf eine abgeschlossene, konvexe Menge in einem uniform konvexen Banachraum stetig. □

Die Begriffe *approximativ kompakt* und *approximativ schwach kompakt* ergeben sich im Zusammenhang mit Existenzmengen auf ganz natürliche Weise und man könnte hoffen, möglichst alle Existenzaussagen hierauf zurückzuführen. Das ist leider nicht so, wie das folgende Beispiel von F. DEUTSCH (1980) zeigt.

Beispiel: Sei $(X, \|\cdot\|) := (C[0,1], \|\cdot\|_\infty$ und

$$M := \left\{ x(t) = \frac{1}{1+at} : a \geq 0 \right\} \cup \{0\}.$$

Wir wollen die folgenden Aussagen nachweisen:

1. M ist eine Existenzmenge.

   Denn: Sei $z \in C[0,1]$, o. B. d. A. ist $z \neq 0$. Sei $\{x_k\} \subset M$ mit $x_k(t) = 1/(1+a_k t)$ eine Minimalfolge, also $\|x_k - z\|_\infty \to d := d(z, M)$. Wir unterscheiden zwei Fälle.

   Im ersten Fall ist $\{a_k\} \subset \mathbb{R}$ beschränkt. Dann existiert eine Teilfolge $\{a_{k_j}\} \subset \{a_k\}$ und ein $a^* \in [0, \infty)$ mit $a_{k_j} \to a^*$. Wir definieren $x^* \in M$ durch

   $$x^*(t) := \frac{1}{1+a^*t}.$$

   Für $t \in [0,1]$ ist dann

   $$|x_{k_j}(t) - x^*(t)| = \frac{|a_{k_j} - a^*|\, t}{(1+a_{k_j}t)(1+a^*t)} \leq |a_{k_j} - a^*|.$$

   Also ist $\|x_{k_j} - x^*\|_\infty \leq |a_{k_j} - a^*|$, folglich $x_{k_j} \to x^* \in M$ und $\|x^* - z\| = d(z, M)$ und damit x^* eine beste Approximierende an z in M.

   Im zweiten Fall ist $\{a_k\} \subset \mathbb{R}$ nicht beschränkt. Dann existiert eine Teilfolge $\{a_{k_j}\} \subset \{a_k\}$ mit $a_{k_j} \to \infty$. Offenbar ist

   $$\lim_{j\to\infty} x_{k_j}(t) = \begin{cases} 1, & t = 0, \\ 0, & t \in (0,1]. \end{cases}$$

   Wir wollen zeigen, dass $x^* = 0$ beste Approximierende an z in M ist, dass also $\|z\|_\infty = d$. Angenommen, es wäre $d < \|z\|_\infty$. Nun definiere man die positive Zahl

   $$\delta := \frac{1}{4}(\|z\|_\infty - d).$$

   Mit einem $t_0 \in [0,1]$ mit $|z(t_0)| = \|z\|_\infty$ ist dann

   $$|z(t_0)| > \|z\|_\infty - \delta = d + 3\delta.$$

Da z auf $[0,1]$ stetig ist, gibt es ein $t_1 \in (0,1]$ mit $|z(t_1)| > d + 2\delta$. Wegen $x_{k_j}(t_1) \to 0$ ist $|x_{k_j}(t_1) - z(t_1)| > d + 2\delta$ für alle hinreichend großen j und daher auch

$$\|x_{k_j} - z\|_\infty \geq |x_{k_j}(t_1) - z(t_1)| > d + \delta$$

für alle hinreichend großen j. Da $\{x_{k_j}\}$ eine Minimalfolge ist, folgt $d \geq d + \delta$, ein Widerspruch zu $\delta > 0$.

2. M ist nicht approximativ schwach kompakt.

   Denn: Sei $z(t) := \frac{1}{2}$. Für beliebiges $x \in M$ ist

$$\|x - z\|_\infty \geq |x(0) - z(0)| = \frac{1}{2},$$

   da $x(0) = 0$ oder $x(0) = 1$. Folglich ist $d := d(z; M) \geq \frac{1}{2}$. Hieraus erkennt man, dass $x^* = 0$ und $x^*(t) = 1 = 1/(1 + 0 \cdot t)$ beste Approximierende an z in M sind und $d = \frac{1}{2}$ ist. Man definiere $\{x_k\} \subset M$ durch

$$x_k(t) := \frac{1}{1 + k\,t}.$$

   Dann ist

$$|x_k(t) - z(t)| = \frac{1}{2} \cdot \frac{|1 - k\,t|}{1 + k\,t}.$$

   In Abbildung 3.4 veranschaulichen wir uns f_k für $k = 1, 3, 5$. Man zeigt leicht,

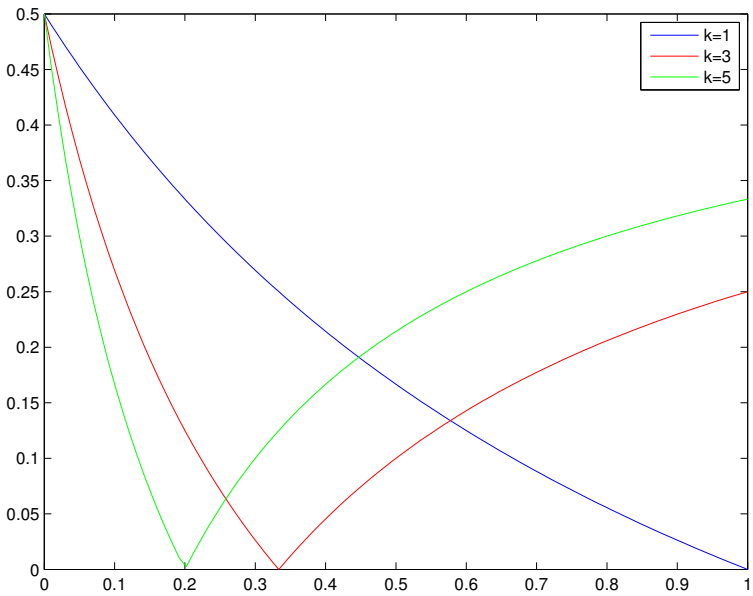

Abbildung 3.4: Der Defekt $|x_k(z) - z(t)|$ für $k = 1, 3, 5$

   dass $\|x_k - z\|_\infty = \frac{1}{2}$, insbesondere ist $\{x_k\}$ eine Minimalfolge. Angenommen, es

existiert eine Teilfolge $\{x_{k_j}\} \subset \{x_k\}$ und ein $x^* \in M$ mit $x_{k_j} \rightharpoonup x^*$. Die schwache Konvergenz auf $C[0,1]$ impliziert die punktweise Konvergenz (Beweis?). Daher würde

$$x^*(t) = \begin{cases} 1, & t = 0, \\ 0, & t \in (0,1] \end{cases}$$

folgen, ein Widerspruch zu $x^* \in M$.

3. M ist nicht approximativ kompakt.

   Dies folgt sofort aus 2.

□

Das letzte Beispiel war insofern ein negatives Beispiel, als es zeigte, dass man *nicht* jede Existenzaussage auf den Nachweis der approximativen Kompaktheit zurückführen kann. Um diesen Abschnitt aber noch mit einem positiven Resultat (siehe F. DEUTSCH (1980, p. 146)) abzuschließen, zeigen wir, dass die Menge der rationalen Funktionen in $L^p[\alpha,\beta]$ für $1 < p < \infty$ approximativ kompakt, also eine Existenzmenge ist.

Satz 3.1.9 *Sei $I := [\alpha,\beta]$ ein kompaktes Intervall, m,n nichtnegative ganze Zahlen und*

$$R_{m,n} := \left\{ \frac{p}{q} : p \in \Pi_m,\ q \in \mathbb{R}^n,\ q(t) > 0 \text{ auf } I \right\}.$$

Dann ist $R_{m,n}$ für $1 \le p < \infty$ in $L^p(I)$ approximativ kompakt[1].

Beweis: Sei $z \in L^p(I)$ vorgegeben und $\{u_k\} \subset R_{m,n}$ eine zugehörige Minimalfolge, also $\|u_k - z\|_p \to d := d(z, R_{m,n})$. Als Minimalfolge ist $\{u_k\}$ beschränkt, es existiert also ein $c > 0$ mit $\|u_k\|_p \le c$ für alle k. Es ist $u_k = p_k/q_k$ mit $p_k \in \Pi_m$, $q_k \in \Pi_n$ und $q_k(t) > 0$ für $t \in I$. Bei dieser Darstellung können wir o. B. d. A. annehmen, dass

$$\|q_k\|_\infty = \max_{t \in I} q_k(t) = 1.$$

Dann ist aber

$$\|p_k\|_p = \left(\int_I |p_k(t)|^p \, dt \right)^{1/p} \le \left(\int_I \left| \frac{p_k(t)}{q_k(t)} \right|^p dt \right)^{1/p} = \|u_k\|_p \le c.$$

Also ist $\{p_k\}$ eine beschränkte Folge in dem endlichdimensionalen linearen normierten Raum $(\Pi_m, \|\cdot\|_p)$. Da in diesem Raum alle Normen äquivalent sind, ist $\{p_k\}$ auch in $(\Pi_m, \|\cdot\|_\infty)$ eine beschränkte Folge. Ebenso ist $\{q_k\}$ eine beschränkte Folge in $(\Pi_n, \|\cdot\|_\infty)$. Indem man notfalls zu Teilfolgen übergeht kann man o. B. d. A. annehmen, dass $p_k \to p^* \in \Pi_m$ und $q_k \to q^* \in \Pi_n$ gleichmäßig auf I. Es ist $\|q^*\|_\infty = 1$, insbesondere ist q^* nicht das Nullpolynom. Die Schwierigkeiten, mit denen wir uns jetzt noch herumschlagen müssen, rühren daher, dass wir natürlich nicht a priori sichern können, dass $q^*(t) > 0$ für alle $t \in I$. Selbstverständlich ist $q^*(t) \ge 0$ für alle $t \in I$. Ferner kann

[1] In der Formulierung dieses Satzes kommt p in zweierlei Bedeutung vor. Das sollte aber nicht zu Verwirrungen führen!

q^* als (nichttriviales) Polynom vom Grad $\leq n$ höchstens n Nullstellen in I besitzen. Es genügt daher zu zeigen: Besitzt q^* eine Nullstelle $t_0 \in I$ mit der Vielfachheit ν, ist also

$$q^*(t) = (t - t_0)^\nu q(t) \quad \text{mit } q \in \Pi_{n-\nu},\ q(t_0) \neq 0,$$

so hat p^* in t_0 eine Nullstelle der Vielfachheit $\geq \nu$. Angenommen, dies sei nicht richtig. Dann ist

$$p^*(t) = (t - t_0)^\mu p(t) \quad \text{mit } p \in \Pi_{m-\mu},\ p(t_0) \neq 0$$

mit einer nichtnegativen ganzen Zahl μ mit $\mu < \nu$. Mit $\lambda := \nu - \mu \geq 1$ gibt es dann eine Umgebung U_0 von t_0 in I und ein $\delta > 0$ derart, dass

$$\frac{p^*(t)}{q^*(t)} = \frac{p(t)}{(t - t_0)^\lambda q(t)}, \qquad \left|\frac{p(t)}{q(t)}\right| \geq \delta \quad \text{für } t \in U_0.$$

Für jedes $t \in I$, welches keine Nullstelle von q^* ist, gilt

$$\left|\frac{p_k(t)}{q_k(t)}\right|^p \to \left|\frac{p^*(t)}{q^*(t)}\right|^p,$$

ferner ist

$$\int_I \left|\frac{p_k(t)}{q_k(t)}\right|^p dt \leq c^p.$$

Aus dem Lemma von Fatou (man informiere sich über dieses!) folgt

$$\int_I \left|\frac{p^*(t)}{q^*(t)}\right| dt = \int_I \liminf_{k\to\infty} \left|\frac{p_k(t)}{q_k(t)}\right| dt \leq \liminf_{k\to\infty} \int_I \left|\frac{p_k(t)}{q_k(t)}\right| dt \leq c^p.$$

Insbesondere ist

$$c^p \geq \int_I \left|\frac{p^*(t)}{q^*(t)}\right| dt \geq \int_{U_0} \left|\frac{p^*(t)}{q^*(t)}\right| dt \geq \delta^p \int_{U_0} \frac{dt}{|t - t_0|^{\lambda p}} = +\infty$$

wegen $\lambda p \geq 1$, ein Widerspruch. Man kann also im Ausdruck p^*/q^* durch die Nullstellen von q^* kürzen und erhält, indem man notfalls noch Zähler und Nenner mit -1 multipliziert, dass $p^*/q^* \in R_{m,n}$. Es bleibt zu zeigen, dass

$$u_k = \frac{p_k}{q_k} \to \frac{p^*}{q^*} =: u^* \quad \text{in } L^p(I).$$

Hierzu überlegen wir uns zunächst, dass $\|u_k - z\|_p \to \|u^* - z\|_p$, was insbesondere impliziert, dass u^* beste Approximierende an z in $R_{m,n}$ und damit $R_{m,n}$ eine Existenzmenge in $L^p(I)$ ist. Denn wegen des Lemmas von Fatou ist

$$\begin{aligned}
\|u^* - z\|_p^p &= \int_I |u^*(t) - z(t)|^p \, dt \\
&= \int_I \liminf_{k\to\infty} |u_k(t) - z(t)|^p \, dz \\
&\leq \liminf_{k\to\infty} \int_I |u_k(t) - z(t)|^p \, dt \\
&= d(z, R_{m,n})^p \\
&\leq \|u^* - z\|_p^p,
\end{aligned}$$

folglich ist

$$\lim_{k\to\infty} \|u_k - z\|_p = d(z, M) = \|u^* - z\|_p.$$

Aus $\|u_k - z\|_p \to \|u^* - z\|_p$ und $u_k(t) \to u^*(t)$ fast überall auf I folgt $\|u_k - u^*\|_p \to 0$, siehe E. HEWITT, K. STROMBERG (1965, p. 209). Damit ist der Satz bewiesen. □

Zusammenfassend kann man zu diesem Abschnitt sagen, dass wir verschiedene Konzepte zum Nachweis der Existenz bester Approximierender angegeben haben. Wir haben die wichtigen Begriffe *Existenzmenge* und *Tschebyscheff (T)-Menge* kennengelernt. Die Existenz bester Approximierender ist auch für einige konkrete Aufgaben bewiesen worden, z. B. Approximation bezüglich endlichdimensionaler linearer Teilräume (z. B. Π_n) eines linearen normierten Raumes (z. B. $(C[\alpha,\beta], \|\cdot\|_\infty)$ oder $(L^p[\alpha,\beta], \|\cdot\|_p)$) und rationaler Funktionen in $L^p[\alpha,\beta]$, $1 \le p < \infty$. Für weitere nichtlineare Aufgaben, etwa T-Approximation bezüglich rationaler Funktionen und T- und L^p-Approximation bezüglich verallgemeinerter Exponentialsummen werden wir später Existenzbeweise nachholen. Außerdem haben wir einfache Eindeutigkeitsaussagen erhalten, die mit dem Begriff *strikt konvex* im Zusammenhang stehen.

3.2 Charakterisierung bester Approximierender

In diesem Abschnitt sei $(X, \|\cdot\|)$ ein (reeller) linearer normierter Raum, $M \subset X$ und $z \in X$. Die zu diesen Daten gehörende Approximationsaufgabe heißt *konvex*, falls M konvex und entsprechend *linear*, falls M ein linearer Teilraum von X ist. Dass die Verhältnisse bei konvexen Approximationsaufgaben angenehmer als im allgemeinen Fall sind, zeigt schon das folgende einfache

Lemma 3.2.1 *Eine lokale beste Approximierende einer konvexen Approximationsaufgabe ist sogar eine (globale) beste Approximierende.*

Beweis: Sei $x^* \in M$ eine lokale beste Approximierende an z in M, wobei $M \subset X$ konvex ist. Dann existiert eine Umgebung U von x^* mit $\|x^* - z\| \le \|u - z\|$ für alle $u \in M \cap U$. Sei nun $x \in M$ beliebig. Dann ist $u_z := (1-t)x^* + tx \in M \cap U$ für alle $t \in [0, t^*]$ mit hinreichend kleinem $t^* \in (0, 1]$. Insbesondere ist also

$$\begin{aligned} \|x^* - z\| &\le \|u_{t^*} - z\| \\ &= \|(1-t^*)(x^* - z) + t^*(x - z)\| \\ &\le (1-t^*)|\,\|x^* - z\| + t^*\,\|x - z\|, \end{aligned}$$

woraus $\|x^* - z\| \le \|x - z\|$ folgt. □

Bemerkung: Wir wollen uns überlegen, dass für die verallgemeinerte rationale Tschebyscheff-Approximation eine Aussage gilt, die der von Lemma 3.2.1 vollständig entspricht, obwohl es sich hierbei i. Allg. um keine konvexe Approximationsaufgabe handelt.

Sei $B \subset \mathbb{R}^N$ kompakt, $(C(B), \|\cdot\|_\infty)$ der Banachraum der auf B definierten reellwertigen stetigen Funktionen versehen mit der Maximumnorm $\|\cdot\|_\infty$, definiert durch

$$\|x\|_\infty := \max_{t\in B} |x(t)|.$$

Seien $P, Q \subset C(B)$ lineare Teilräume und

$$Q_+ := \{q \in Q : q(t) > 0 \text{ für alle } t \in B\} \neq \emptyset.$$

Schließlich sei

$$R := \left\{ \frac{p}{q} : p \in P,\ q \in Q_+ \right\} \subset C(B)$$

die zugehörige Menge der verallgemeinerten rationalen Funktionen. Wir wollen zeigen:

- *Ist $p^*/q^* \in R$ eine lokal beste Approximierende an ein $z \in C(B)$ in R, so ist p^*/q^* auch eine global beste Approximierende an z in R.*

Denn: Nach Voraussetzung existiert eine Umgebung U von p^*/q^* mit

$$\|p^*/q^* - z\|_\infty \leq \|u - z\|_\infty \quad \text{für alle } u \in R \cap U.$$

Sei $p/q \in R$ beliebig. Für ein hinreichend kleines $\lambda_0 \in (0, 1]$ ist

$$u_0 := \frac{(1-\lambda_0)p^* + \lambda_0 p}{(1-\lambda_0)q^* + \lambda_0 q} \in R \cap U.$$

Daher ist

$$\begin{aligned}
\left\| \frac{p^*}{q^*} - z \right\|_\infty &\leq \left\| \frac{(1-\lambda_0)p^* + \lambda_0 p}{(1-\lambda_0)q^* + \lambda_0 q} - z \right\|_\infty \\
&= \left| \frac{(1-\lambda_0)p^*(t_0) + \lambda_0 p(t_0)}{(1-\lambda_0)q^*(t_0) + \lambda_0 q(t_0)} - z(t_0) \right| \\
&= \frac{|(1-\lambda_0)(p^*(t_0) - q^*(t_0)z(t_0)) + \lambda_0(p(t_0) - q(t_0)z(t_0))|}{(1-\lambda_0)q^*(t_0) + \lambda_0 q(t_0)}
\end{aligned}$$

mit einem $t_0 \in B$ nach Definition der Maximumnorm. Hieraus folgt

$$\begin{aligned}
((1-\lambda_0)q^*(t_0) + \lambda_0 q(t_0)) \left\| \frac{p^*}{q^*} - z \right\|_\infty &\leq (1-\lambda_0)q^*(t_0) \left| \frac{p^*(t_0)}{q^*(t_0)} - z(t_0) \right| \\
&\quad + \lambda_0 q(t_0) \left| \frac{p(t_0)}{q(t_0)} - z(t_0) \right| \\
&\leq (1-\lambda_0)q^*(t_0) \left\| \frac{p^*}{q^*} - z \right\|_\infty \\
&\quad + \lambda_0 q(t_0) \left\| \frac{p}{q} - z \right\|_\infty
\end{aligned}$$

und dann

$$\left\| \frac{p^*}{q^*} - z \right\|_\infty \leq \left\| \frac{p}{q} - z \right\|_\infty,$$

womit die Behauptung bewiesen ist. □

Nun kommen wir zu notwendigen Optimalitätsbedingungen bei konvexen Approximationsaufgaben. Gegeben sei also die Aufgabe

(P) $$\text{Minimiere} \quad f(x) := \|x - z\| \quad \text{auf } M,$$

wobei $M \subset X$ konvex ist. Sei $x^* \in M$ eine Lösung von (P) bzw. eine beste Approximierende an z in M (lokal oder global, das spielt wegen Lemma 3.2.1 keine Rolle). Für beliebiges $x \in M$ ist

$$\|x^* - z\| \leq \| \underbrace{x^* + t(x - x^*)}_{\in M} - z\| \quad \text{für } t \in [0,1]$$

und daher

$$0 \leq \frac{1}{t}[\|x^* + t(x - x^*) - z\| - \|x^* - z\|] \quad \text{für } t \in (0,1].$$

Wenn man in dieser Ungleichung rechts den Grenzübergang $t \to 0+$ machen könnte, hätten wir eine erste notwendige Optimalitätsbedingung. Wir werden zeigen, dass man diesen Grenzübergang ganz allgemein für *konvexe* Funktionen machen kann. Es folgt daher zunächst ein Exkurs über konvexe Funktionen, einseitige Richtungsableitung und das Subdifferential konvexer Funktionen.

Definition 3.2.2 Eine Abbildung $f\colon X \longrightarrow \mathbb{R}$ heißt *konvex* (auf X), falls

$$x, y \in X,\ t \in [0,1] \Longrightarrow f((1-t)x + ty) \leq (1-t)f(x) + tf(y).$$

Bei unseren Anwendungen wird i. Allg. $f(x) = \|x - z\|$ sein. Wegen

$$\begin{aligned} f((1-t)x + ty) &= \|(1-t)(c-z) + t(y-z)\| \\ &\leq (1-t)\,\|x-z\| + t\,\|y-z\| \\ &= (1-t)f(x) + tf(y) \end{aligned}$$

ist f konvex.Anschaulich sehen konvexe Funktionen bekanntlich wie in Abbildung 3.5 aus.

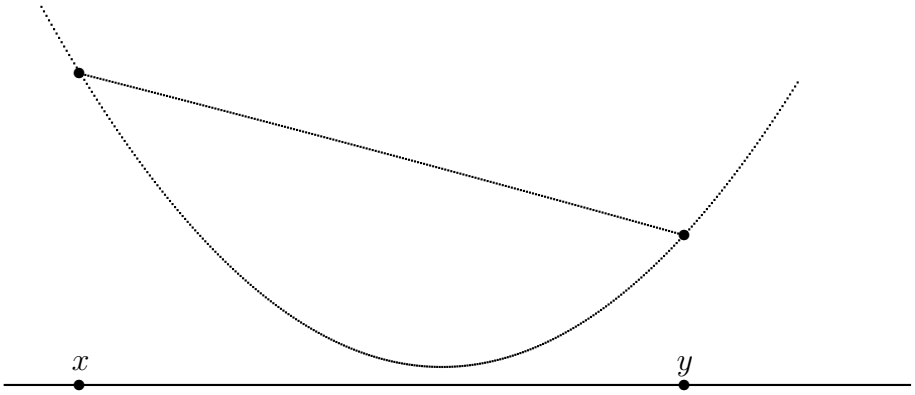

Abbildung 3.5: Eine konvexe Funktion

Sehr wichtig ist nun der folgende Satz, in dem die Existenz der einseitigen Richtungsableitung (auch Gateaux-Variation genannt) für konvexe Funktionen gezeigt wird.

Satz 3.2.3 *Sei $f\colon X \longrightarrow \mathbb{R}$ konvex. Für jedes $x \in X$, $p \in X$ existiert dann*

$$f'(x;p) := \lim_{t \to 0+} \frac{f(x+tp) - f(x)}{t},$$

die sogenannte einseitige Richtungsableitung von f in x in Richtung p. Die Abbildung $f'(x;\cdot)\colon X \longrightarrow \mathbb{R}$, die Gateaux-Variation bzw. G-Variation von f in x, besitzt die folgenden Eigenschaften:

1. *Es ist*
$$f(x) - f(x-p) \leq f'(x;p) \leq f(x+p) - f(x) \quad \text{für alle } p \in X.$$
2. *Die Abbildung $f'(x;\cdot)\colon X \longrightarrow \mathbb{R}$ ist positiv homogen, d. h. es ist*
$$f'(x;\alpha p) = \alpha f'(x;p) \quad \text{für alle } \alpha \geq 0,\ p \in X.$$
3. *Die Abbildung $f'(x;\cdot)\colon X \longrightarrow \mathbb{R}$ ist subadditiv, d. h. es ist*
$$f'(x;p+q) \leq f'(x;p) + f'(x;q) \quad \text{für alle } p, q \in X.$$
4. *Die Abbildung $f'(x;\cdot)\colon X \longrightarrow \mathbb{R}$ ist konvex.*
5. *Es ist $-f'(x;-p) \leq f'(x;p)$ für alle $p \in X$.*

Beweis: Für feste $x, p \in X$ definieren wir $\phi\colon (0,\infty) \longrightarrow \mathbb{R}$ durch
$$\phi(t) := \frac{f(x+tp) - f(x)}{t}$$
und zeigen

(a) Es ist $f(x) - f(x-p) \leq \phi(t)$ für alle $t > 0$.
Denn: Für $t > 0$ ist
$$\begin{aligned} \frac{1}{1+t} f(x) + \frac{t}{1+t} f(x) &= f(x) \\ &= f\left(\frac{1}{1+t}(x+tp) + \frac{t}{1+t}(x-p)\right) \\ &\leq \frac{1}{1+t} f(x+tp) + \frac{t}{1+t} f(x-p) \end{aligned}$$
und daher
$$\frac{t}{1+t}[f(x) - f(x-p)] \leq \frac{1}{1+t}[f(x+tp) - f(x)],$$
woraus (a) folgt.

(b) ϕ ist nicht fallend auf $(0,\infty)$ und $\phi(s) \leq f(x+p) - f(x)$ für alle $s \in (0,1]$.
Denn: Für $0 < s \leq t$ ist
$$\begin{aligned} f(x+sp) - f(x) &= f\left(\frac{s}{t}(x+tp) + \frac{t-s}{t} x\right) - f(x) \\ &\leq \frac{s}{t} f(x+tp) + \frac{t-s}{t} f(x) - f(x) \\ &= \frac{s}{t}[f(x+tp) - f(x)] \end{aligned}$$
und damit $\phi(s) \leq \phi(t)$. Für $s \in (0,1]$ ist insbesondere
$$\phi(s) \leq \phi(1) = f(x+p) - f(x).$$

Aus (a) und (b) folgt die Existenz der Richtungsableitung von f in Richtung p sowie die Eigenschaft 1. Die Eigenschaft 2. gilt offensichtlich. Zum Nachweis von 3. beachte man, dass

$$\begin{aligned} f(x+t(p+q)) &= f\left(\frac{1}{2}(x+2tp)+\frac{1}{2}(x+2tq)\right) \\ &\leq \frac{1}{2}f(x+2tp)+\frac{1}{2}f(x+2tq). \end{aligned}$$

Für $t>0$ ist daher

$$\frac{f(x+t(p+q))-f(x)}{t} \leq \frac{f(x+2tp)-f(x)}{2t}+\frac{f(x+2tq)-f(x)}{2t}.$$

Mit $t\to 0+$ folgt die Subadditivität von $f'(x;\cdot)$. Die Eigenschaft 4., also die Konvexität von $f'(x;\cdot)$ ist eine unmittelbare Folgerung aus 2. und 3. Schließlich ist

$$0=f'(x;0)\leq f'(x;p)+f'(x;-p),$$

womit auch 5. bewiesen ist. □

Nun können wir schon eine erste, ganz einfache notwndige und hinreichende Optimalitätsbedingung für die konvexe Optimierungsaufgabe (P) angeben.

Satz 3.2.4 *Gegeben sei die konvexe Approximationsaufgabe*

(P) $$\textit{Minimiere} \quad f(x):=\|x-z\|, \quad x\in M.$$

Dann ist $x^*\in M$ genau dann eine beste Approximierende an z in M, wenn

$(*)$ $$0\leq f'(x^*;x-x^*) \quad \textit{für alle } x\in M.$$

Beweis: Zunächst nehmen wir an, $x^*\in M$ sei eine beste Approximierende an z in M. Aus

$$0\leq \frac{1}{t}[f(x^*+t(x-x^*))-f(x^*)] \quad \text{für } t\in(0,1]$$

folgt $(*)$ mit $t\to 0+$. Umgekehrt gelte $(*)$ und es sei $x\in M$. Dann ist

$$0\leq f'(x^*;x-x^*)\leq f(x^*+(x-x^*))-f(x^*),$$

folglich $f(x^*)\leq f(x)$. Daher ist $x^*\in M$ eine beste Approximierende an z in M. □

Nun will man die Bedingung $(*)$ im konkreten Fall natürlich "ausschlachten". Dazu muss man die Gateaux-Variation der Norm kennen. Diese wollen wir in einigen Beispielen ausrechnen.

Beispiele: Es sei stets $f(x):=\|x-z\|$. Offenbar ist

$$f'(z;p)=\lim_{t\to 0+}\frac{\|z+tp-z\|-\|z-z\|}{t}=\|p\|,$$

wir brauchen $f'(x;p)$ also nur für $x\neq z$ auszurechnen. Im folgenden sei also $x\neq z$.

1. Sei $(X, (\cdot,\cdot))$ ein Prä-Hilbertraum und $\|\cdot\|$ die durch $\|x\| := (x,x)^{1/2}$ definierte zugehörige Norm. Für $x \neq z$ ist dann

$$\begin{aligned} f'(x;p) &= \lim_{t\to 0+} \frac{1}{t}\{[(x-z,x-z)+2t(x-z,p)+t^2(p,p)]^{1/2} - (x-z,x-z)^{1/2}\} \\ &= \frac{(x-z,p)}{\|x-z\|}, \end{aligned}$$

wobei wir die Regel von de L'Hospital angewandt haben.

2. Sei $X := \mathbb{R}^n$. Die G-Variation der euklidischen Norm haben wir schon in 1. ausgerechnet.

(a) Sei $\|\cdot\| := \|\cdot\|_1$, $\|x\|_1 := \sum_{j=1}^n |x_j|$.

Sei $J_0 := \{j \in \{1,\dots,n\} : x_j = z_j\}$. Bei vorgegebenem $p \in \mathbb{R}^n$ und für $j \notin J_0$ ist $\operatorname{sign}(x_j + th_j - z_j) = \operatorname{sign}(x_j - z_j)$ für alle hinreichend kleinen $t > 0$ und daher

$$\begin{aligned} f'(x;p) &= \sum_{j=1}^n \lim_{t\to 0+} \frac{1}{t}(|x_j + tp_j - z_j| - |x_j - z_j|) \\ &= \sum_{j\in J_0} \lim_{t\to 0+} \frac{1}{t}(|x_j + tp_j - z_j| - |x_j - z_j|) \\ &\qquad + \sum_{j\notin J_0} \lim_{t\to 0+} \frac{1}{t}(|x_j + tp_j - z_j| - |x_j - z_j|) \\ &= \sum_{j\in J_0} |p_j| + \sum_{j\notin J_0} \operatorname{sign}(x_j - z_j)p_j. \end{aligned}$$

(b) Sei $\|\cdot\| := \|\cdot\|_\infty$, $\|x\|_\infty := \max_{j=1,\dots,n}[x_j|$.

Sei $B(x-z) := \{j \in \{1,\dots,n\} : |x_j - z_j| = \|x-z\|_\infty\}$. Für $x \neq z$ ist dann

$$f'(x;p) = \max_{j\in B(x-z)} \operatorname{sign}(x_j - z_j)p_j.$$

Dies wollen wir hier nicht beweisen, da es sich in Kürze als Spezialfall erweist.

(c) Sei $\|\cdot\| := \|\cdot\|_p$, $\|x\|_p := (\sum_{j=1}^n |x_j|^p)^{1/2}$ mit $1 < p < \infty$.

Für $x \neq z$ ist

$$f'(x;p) = \frac{\sum_{j=1}^n \operatorname{sign}(x_j - z_j)|x_j - z_j|^{p-1}p_j}{(\sum_{j=1}^n |x_j - z_j|^p)^{1/q}}$$

mit $1/p + 1/q = 1$, wie man nach leichter Rechnung erhält.

3. Sei $(X, \|\cdot\|) := (C(B), \|\cdot\|_\infty)$, wobei $B \subset \mathbb{R}^n$ eine kompakte Menge ist (eigentlich genügt: B ist ein kompakter metrischer Raum) und $\|x\|_\infty := \max_{t\in B} |x(t)|$. Wir zeigen: Für $x \neq z$ ist

$$f'(x;p) = \max_{t\in B(x-z)} \operatorname{sign}(x(t) - z(t))p(t)$$

mit

$$B(x-z) := \{t \in B : |x(t) - z(t)| = \|x - z\|_\infty\}.$$

Zum Beweis müssen wir zeigen, dass

$$\lim_{s \to 0+} \frac{\|x + tp - z\|_\infty - \|x - z\|_\infty}{s} = \max_{t \in B(x-z)} \operatorname{sign}(x(t) - z(t))p(t),$$

wobei uns die Existenz des links stehenden Limes schon bekannt ist.

(a) Sei $s_k \to 0+$. Nach Definition der Maximumnorm existiert eine Folge $\{t_k\} \subset B$ mit

$$\|x + s_k p - z\|_\infty = |x(t_k) + s_k p(t_k) - z(t_k)|, \qquad k = 1, 2, \ldots.$$

Da B kompakt ist, besitzt $\{t_k\}$ eine konvergente Teilfolge. O. B. d. A. ist $\{t_k\}$ selbst schon konvergent: $t_k \to t \in B$. Offenbar ist $t \in B(x-z)$ und wegen $x \neq z$ ist $x(t) - z(t) \neq 0$. Für alle hinreichend großen k ist daher

$$\operatorname{sign}(x(t_k) + s_k p(t_k) - z(t_k)) = \operatorname{sign}(x(t_k) - z(t_k)) = \operatorname{sign}(x(t) - z(t)).$$

Für alle hinreichend großen k ist folglich

$$\begin{aligned} f'(x;p) &= \frac{\|x + s_k p - z\|_\infty - \|x - z\|_\infty}{s_k} \\ &= \frac{\operatorname{sign}(x(t) - z(t))(x(t_k) + s_k p(t_k) - z(t_k)) - \|x - z\|_\infty}{s_k} \\ &\leq \operatorname{sign}(x(t) - z(t))p(t_k). \end{aligned}$$

Mit $k \to \infty$ folgt

$$f'(x;p) \leq \operatorname{sign}(x(t) - z(t))p(t) \leq \max_{t \in B(x-z)} \operatorname{sign}(x(t) - z(t))p(t).$$

(b) Sei $t \in B(x-z)$ und $s_k \to 0+$. Für alle hinreichend großen k ist dann

$$\operatorname{sign}(x(t) + s_k p(t) - z(t)) = \operatorname{sign}(x(t) - z(t))$$

und daher

$$\begin{aligned} \frac{\|x + s_k p - z\|_\infty - \|x - z\|_\infty}{s_k} &\geq \frac{|x(t) + s_k p(t) - z(t)| - |x(t) - z(t)|}{s_k} \\ &= \operatorname{sign}(x(t) - z(t))p(t). \end{aligned}$$

Mit $k \to \infty$ folgt $f'(x;p) \geq \operatorname{sign}(x(t) - z(t))p(t)$ und damit auch

$$f'(x;p) \geq \max_{t \in B(x-z)} \operatorname{sign}(x(t) - z(t))p(t).$$

Aus (a) und (b) folgt die Behauptung. □

Als Folgerung aus dem letzten Beispiel und dem in Satz 3.2.4 gewonnenen Ergebnis erhalten wir

Satz 3.2.5 (Kolmogoroff-Kriterium) *Sei* $(X, \|\cdot\|) := (C(B), \|\cdot\|_\infty)$, $M \subset X$ *konvex und* $z \in X \setminus M$. *Dann ist* $x^* \in M$ *genau dann eine beste Approximierende an* z *in* M, *wenn*

$$\max_{t \in B(x^*-z)} (x^*(t) - z(t))(x(t) - x^*(t)) \geq 0 \quad \textit{für alle } x \in M$$

bzw.

$$\min_{t \in B(x^*-z)} (z(t) - x^*(t))(x(t) - x^*(t)) \leq 0 \quad \textit{für alle } x \in M.$$

Ist $M \subset X$ *sogar ein linearer Teilraum, so ist* $x^* \in M$ *genau dann eine beste Approximierende an* z *in* M, *wenn*

$$\max_{t \in B(x^*-z)} (x^*(t) - z(t))x(t) \geq 0 \quad \textit{für alle } x \in M$$

bzw.

$$\min_{t \in B(x^*-z)} (z(t) - x^*(t))x(t) \leq 0 \quad \textit{für alle } x \in M.$$

Hierbei ist stets $B(x^* - z) := \{t \in B : |x^*(t) - z(t)| = \|x^* - z\|_\infty\}$.

Beweis: Aus Satz 3.2.4 und dem obigen Beispiel 3. erhält man als Kriterium für die Optimalität von $x^* \in M$, dass

$$0 \leq \max_{t \in B(x^*-z)} \operatorname{sign}(x^*(t) - z(t))(x(t) - x^*(t)) \quad \text{für alle } x \in M.$$

Eine Multiplikation dieser Beziehung mit $\|x^* - z\|_\infty = |x^*(t) - z(t)|$, $t \in B(x^* - z)$, liefert unter Berücksichtigung von

$$\operatorname{sign}(x^*(t) - z(t))|x^*(t) - z(t)| = x^*(t) - z(t)$$

die erste Behauptung. Der Rest ergibt sich in trivialer Weise. □

Auf die Anwendung des Kolmogoroff-Kriteriums in der T-Approximation kommen wir später zu sprechen.

Ausführlich sind wir auf den Begriff der einseitigen Richtungsableitung bzw. der G-Variation eingegangen. Ein weiterer wichtiger und nützlich Begriff wird nun definiert.

Definition 3.2.6 Sei $f\colon X \longrightarrow \mathbb{R}$ konvex und $x \in X$. Dann heißt

$$\partial f(x) := \{l \in X^* : l(y - x) \leq f(y) - f(x) \text{ für alle } y \in X\}$$

das *Subdifferential* von f in x, ein Element $l \in \partial f(x)$ heißt *Subgradient* von f in x.

Bemerkungen: 1. Wir setzen hier stets voraus, dass der Definitionsbereich von f der ganze Raum X ist, außerdem kommen als Elemente des Subdifferentials nur *stetige* lineare Funktionale in Frage. Hierin unterscheidet sich obige Definition von anderen.

2. Um die anschauliche Bedeutung des Subdifferentials zu erläutern, bilde man den sogenannten *Epigraphen*

$$\operatorname{epi}(f) := \{(y, t) \in X \times \mathbb{R} : f(y) \leq t\}$$

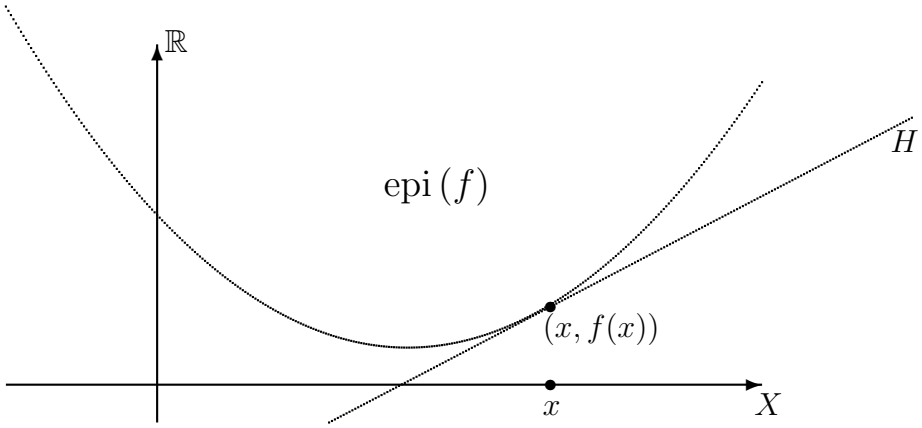

Abbildung 3.6: Der Epigraph einer konvexen Funktion

von f, siehe Abbildung 3.6. Dann gilt offenbar: Es ist $l \in \partial f(x)$ genau dann, wenn

$$H := \{(y,t) \in X \times \mathbb{R} : l(y) - t = l(x) - f(x)\}$$

eine Stützhyperebene für epi(f) in $(x, f(x))$ ist, d. h. es ist $(x, f(x)) \in \text{epi}(f) \cap H$ und $\text{epi}(f) \subset H^-$.

3. Das Subdifferential $\partial f(x)$ einer konvexen Funktion f in einem Punkt $x \in X$ hat offenbar die folgenden Eigenschaften:

(a) $\partial f(x) \subset X^*$ ist konvex und abgeschlossen.

(b) $\partial f(x)$ ist schwach-$*$-folgenabgeschlossen, d. h. ist $\{l_k\} \subset \partial f(x)$ und $l_k \overset{*}{\rightharpoonup} l$ (bzw. $l_k(y) \to l(y)$ für alle $y \in X$), so ist $l \in \partial f(x)$.

(c) Ist f in x stetig, so ist $\partial f(x)$ beschränkt, d. h. in einer Kugel um 0 mit einem hinreichend großen Radius enthalten.

Denn: Da f in x stetig ist, gibt es ein $\delta > 0$ mit $|f(y) - f(x)| \leq 1$ für alle $y \in B[x;\delta]$. Sei $l \in \partial f(x)$ und $p \in B[0;\delta]$. Dann ist

$$l(\pm p) = l(x \pm p - x) \leq f(\underbrace{x \pm p}_{\in B[x;\delta]}) - f(x) \leq 1$$

und folglich $|l(p)| \leq 1/\delta$ für alle $p \in B[0;\delta]$ bzw. $\|l\| \leq 1/\delta$. Also ist $\partial f(x) \subset B[0;1/\delta]$ und damit $\partial f(x)$ beschränkt.

(d) Wegen (b), (c) und des Satzes von Banach-Alaoglu (Satz 2.3.9) ist das Subdifferential $\partial f(x)$ einer konvexen Funktion f, die in x stetig ist, schwach-$*$-folgenkompakt.

□

Wir fassen zusammen und ergänzen.

Satz 3.2.7 *Sei $f\colon X \longrightarrow \mathbb{R}$ konvex und stetig. Dann ist das Subdifferential $\partial f(x)$ von f in x eine nichtleere, konvexe, schwach-$*$-folgenkompakte Teilmenge von X^*.*

Beweis: Zu zeigen bleibt lediglich noch, dass $\partial f(x) \neq \emptyset$. Wir betrachten den Epigraphen

$$\operatorname{epi}(f) := \{(y,t) \in X \times \mathbb{R} : f(y) \leq t\}$$

von f. Wegen der Stetigkeit von f ist $\operatorname{epi}(f)$ abgeschlossen. Ferner ist $\operatorname{int}(\operatorname{epi}(f)) \neq \emptyset$, da z. B. $(y, f(y)+1) \in \operatorname{int}(\operatorname{epi}(f))$ für alle $y \in X$ (Beweis?). Ferner ist $(x, f(x))$ ein Randpunkt von $\operatorname{epi}(f)$. Wegen Satz 2.2.10 existiert durch $(x, f(x))$ eine Stützhyperebene an $\operatorname{epi}(f)$, es existiert also ein (nichttriviales) Paar $(l_0, t_0) \in X^* \times \mathbb{R} \setminus \{(0,0)\}$ mit

$$l_0(y) - t_0 \cdot t \leq l_0(x) - t_0 \cdot f(x) \quad \text{für alle } (y,t) \in \operatorname{epi}(f).$$

Dann ist notwendig $t_0 > 0$ (Beweis?) und daher ist $l := (1/t_0) l_0 \in \partial f(x)$. □

Bemerkung: Zum Nachweis für $\partial f(x) \neq \emptyset$ genügt es, die Stetigkeit von f in x vorauszusetzen (siehe z. B. J. Werner (1984, S. 83)). Ist ferner $X = \mathbb{R}^n$, so folgt aus der Konvexität von f die Stetigkeit von f (siehe z. B. J. Werner (1984, S. 83)). □

Beispiel: Die konvexe Funktion $f\colon X \longrightarrow \mathbb{R}$ sei definiert durch $f(x) := \|x - z\|$ mit $z \in X$. Was ist $\partial f(x)$? Wir wollen uns überlegen, dass

$$\partial f(x) = \begin{cases} \{l \in X^* : \|l\| = 1,\ l(x-z) = \|x-z\|\}, & x \neq z, \\ B[0;1], & x = z. \end{cases}$$

Zum Beweis nehmen wir zunächst $x \neq z$ an.

(a) Sei $l \in X^*$, $\|l\| = 1$ und $l(x-z) = \|x-z\|$. Für ein beliebiges $y \in X$ ist dann

$$\begin{aligned} l(y-x) &= l(y-z) - l(x-z) \\ &\leq \|y-z\| - \|x-z\| \\ &= f(y) - f(x), \end{aligned}$$

also $l \in \partial f(x)$.

(b) Sei $l \in \partial f(x)$. Dann ist

$$(*) \qquad l(y-x) \leq f(y) - f(x) = \|y-z\| - \|x-z\| \quad \text{für alle } y \in X.$$

Setzt man $y = x \pm p$ in $(*)$, so erhält man

$$\pm l(p) \leq \|x \pm p - z\| - \|x - z\| \leq \|p\|,$$

folglich $|l(p)| \leq \|p\|$ bzw. $\|l\| \leq 1$. Setzt man andererseits $y = z$ in $(*)$, so erhält man $l(z-x) \leq -\|x-z\|$ bzw.

$$\|x-z\| \leq l(x-z) \leq \|l\|\,\|x-z\|\| \leq \|x-z\|.$$

Also ist $l(x-z) = \|x-z\|$ und $\|l\| = 1$ wegen $x \neq z$.

Offenbar ist

$$\partial f(z) = \{l \in X^* : l(y-z) \leq \|y-z\| \text{ für alle } y \in X\} = B[0;1].$$

Damit haben wir obige Aussage verifiziert. □

Der folgende Satz stellt eine Verbindung zwischen der Richtungsableitung und dem Subdifferential einer konvexen Funktion her. Hierdurch kann in einem darauffolgenden Satz neben Satz 3.2.4 eine weitere Charakterisierung einer besten Approximierenden einer konvexen Approximationsaufgabe angegeben werden.

Satz 3.2.8 *Sei $f\colon X \longrightarrow \mathbb{R}$ konvex und stetig. Für jedes $x \in X$ ist dann*

$$f'(x;p) = \max_{l \in \partial f(x)} l(p) \quad \textit{für alle } p \in X.$$

Beweis: Sei $l \in \partial f(x)$ und $p \in X$. Für $t > 0$ ist

$$l(p) = \frac{1}{t} l(x+tp-x) \leq \frac{1}{t}[f(x+tp) - f(x)].$$

Mit $t \to 0+$ folgt $l(p) \leq f'(x;p)$, folglich ist

$$\sup_{l \in \partial f(x)} l(p) \leq f'(x;p).$$

Zu zeigen bleibt daher die Existenz eines $l \in \partial f(x)$ mit $l(p) \geq f'(x;p)$. Hierzu definiere man die beiden Mengen

$$\begin{aligned} A &:= \{(y,s) \in X \times \mathbb{R} : f(y) < s\}, \\ B &:= \{(x+tp, f(x) + tf'(x;p)) \in X \times \mathbb{R} : t \geq 0\}. \end{aligned}$$

Dann sind A und B nichtleer, konvex, $\operatorname{int}(A) \neq \emptyset$ und $A \cap B = \emptyset$, denn andernfalls gäbe es ein $t \geq 0$ mit

$$f(x+tp) < f(x) + tf'(x;p) = f(x) + f'(x;tp) \leq f(x) + [f(x+tp) - f(x)] = f(x+tp),$$

ein Widerspruch. Hierbei haben wir die ersten beiden Aussagen von Satz 3.2.3 über Eigenschaften der Gateaux-Variation ausgenutzt. Aus Satz 2.2.5, dem Satz von Eidelheit, folgt, dass sich A und B durch eine (abgeschlossene) Hyperebene in $X \times \mathbb{R}$ trennen lassen. Es existiert also ein Paar $(l_0, \lambda_0) \in X^* \times \mathbb{R} \setminus \{(0,0)\}$ mit

$$l_0(y) - \lambda_0 \cdot s \leq l_0(x+tp) - \lambda_0 \cdot [f(x) + tf'(x;p)] \quad \text{für alle } (y,s) \in A,\ t \geq 0.$$

Notwendigerweise ist $\lambda_0 > 0$ (Beweis). Mit $l := (1/\lambda_0) l_0$ ist

$$l(y) - f(y) \leq l(x+tp) - [f(x) + tf'(x;p)] \quad \text{für alle } y \in X,\ t \geq 0.$$

Setzt man hier $t = 0$, so erhält man $l \in \partial f(x)$. Setzt man dagegen $y = x$ und $t = 1$, so folgt $f'(x;p) \leq l(p)$. Insgesamt ist der Satz bewiesen. □

Nun geben wir die schon angekündigte weitere Charakterisierung einer besten Approximierenden einer konvexen Approximationssaufgabe an. Diese wird bei D. Braess (1986, S. 8) ein *verallgemeinertes Kolmogoroff-Kriterium* genannt.

Satz 3.2.9 *Gegeben sei die konvexe Optimierungsaufgabe*

(P) $$\text{Minimiere} \quad f(x) := \|x - z\|, \quad x \in M.$$

Dann ist $x^* \in M$ genau dann eine beste Approximierende an z in M, wenn es zu jedem $x \in M$ ein $l \in X^*$ mit

$(*)$ $$\|l\| = 1, \quad l(x^* - z) = \|x^* - z\|, \quad 0 \le l(x - x^*)$$

gibt.

Beweis: Sei $x^* \in M$ eine beste Approximierende an z in M, ferner sei $x \in M$ beliebig. Aus den Sätzen 3.2.4 und 3.2.8 erhalten wir

$$0 \le f'(x^*; x - x^*) = \max_{l \in \partial f(x^*)} l(x - x^*).$$

Für $x^* \neq z$ ist

$$\partial f(x^*) = \{l \in X^* : \|l\| = 1,\ l(x^* - z) = \|x^* - z\|\}$$

(siehe obiges Beispiel), so dass für $x^* \neq z$ zu jedem $x \in M$ die Existenz eines $l \in X^*$ mit $(*)$ gesichert ist. Ist dagegen $x^* = z$, so ist $\partial f(x^*) = B[0;1]$ und man erhält die Behauptung leicht aus dem Korollar 2.2.12 zum Satz von Hahn-Banach. Umgekehrt sei $x^* \in M$ und $x \in M$. Zu diesem $x \in M$ gebe es ein $l \in X^*$ mit $(*)$. Dann ist

$$\|x^* - z\| = l(x^* - z) = \underbrace{l(x^* - x)}_{\le 0} + l(x - z) \le \|l\|\,\|x - z\| \le \|x - z\|,$$

also $x^* \in M$ eine beste Approximierende an z in M. □

Zum Schluss dieses Abschnitts geben wir einen weiteren Charakterisierungssatz an, siehe z. B. D. Braess (1986, S. 6).

Satz 3.2.10 *Gegeben sei die konvexe Approximationsaufgabe*

(P) $$\text{Minimiere} \quad f(x) := \|x - z\|, \quad x \in M.$$

Es sei $z \in X \setminus M$. Dann ist $x^* \in M$ genau dann eine beste Approximierende an z in M, wenn es ein $l \in X^*$ mit

$$\|l\| = 1, \quad l(x^* - z) = \|x^* - z\|, \quad 0 \le l(x - x^*) \quad \textit{für alle } x \in M$$

gibt.

Beweis: Sei $x^* \in M$ eine beste Approximierende an z in M. Wegen $z \notin M$ ist

$$0 < \delta := \|x^* - z\| = \inf_{x \in M} \|x - z\|.$$

Wegen der Optimalität von x^* ist ferner $B(z; \delta) \cap M = \emptyset$. Der Trennungssatz 2.2.5 von Eidelheit impliziert die Existenz von $l \in X^* \setminus \{0\}$, $\gamma \in \mathbb{R}$ mit

$(*)$ $$l(z \pm \delta p) \le \gamma \le l(x) \quad \text{für alle } p \in B[0;1],\ x \in M.$$

Wegen $l \neq 0$ ist o. B. d. A. $\|l\| = 1$. Wir wollen zeigen, dass wir mit diesem l das gesuchte Element gefunden haben. Setzt man $x = x^*$ in $(*)$, so erhält man

$$\pm\delta l(p) \leq l(x^* - z) \leq \|x^* - z\| = \delta \quad \text{für alle } p \in B[0;1].$$

daher ist

$$1 = \|l\| = \sup_{p \in B[0;1]} |l(p)| \leq \frac{l(x^* - z)}{\delta} = \frac{l(x^* - z)}{\|x^* - z\|} \leq 1,$$

also ist $l(x^* - z) = \|x^* - z\|$. Wegen $(*)$ ist schließlich

$$l(x^*) = l\left(z + \|x^* - z\| \frac{x^* - z}{\|x^* - z\|}\right) \leq l(x) \quad \text{für alle } x \in M$$

bzw.

$$0 \leq l(x - x^*) \quad \text{für alle } x \in M.$$

damit genügt $l \in X^*$ den geforderten Bedingungen. Sei umgekehrt $x^* \in M$ und ein $l \in X^*$ ein Element mit

$$\|l\| = 1, \quad l(x^* - z) = \|x^* - z\|, \quad 0 \leq l(x - x^*) \quad \text{für alle } x \in M.$$

Für ein beliebiges $x \in M$ ist dann

$$\|x^* - z\| = l(x^* - z) = \underbrace{l(x^* - x)}_{\leq 0} + l(x - z) \leq \|x - z\|$$

und folglich $x^* \in M$ eine beste Approximierende an z in M. □

Beispiel: Sei $(X, \|\cdot\|) := (C[-1,1], \|\cdot\|_1)$, $M := \Pi_1$ (Menge der Polynome vom Grad ≤ 1) und $z \in C[-1,1] \setminus \Pi_1$ eine konvexe Funktion. Was ist die beste Approximierende an z in Π_1 (bezüglich der 1-Norm)? Wir werden zeigen: Ist $x^* \in \Pi_1$ diejenige lineare

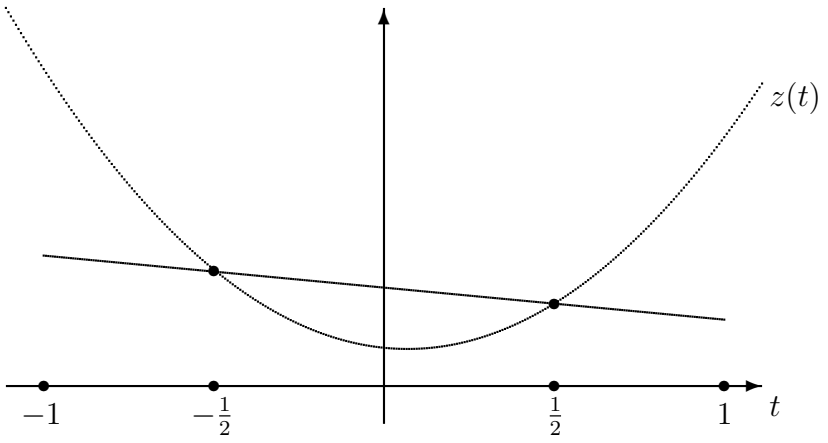

Abbildung 3.7: Beste Approximierende bezüglich 1-Norm an konvexes z in Π_1

Funktion, die z in den Punkten $-\frac{1}{2}$ und $\frac{1}{2}$ interpoliert, so ist x^* beste Approximierende an z in Π_1, siehe Abbildung 3.7. Beim Nachweis beachten wir zunächst, dass

$$x^*(z) - z(t) \begin{cases} \leq 0, & t \in [-1, -\frac{1}{2}], \\ \geq 0, & t \in [-\frac{1}{2}, \frac{1}{2}], \\ \leq 0, & t \in [\frac{1}{2}, 1] \end{cases}$$

wegen der Konvexität von z. Nun definieren wir $l\colon C[-1,1] \longrightarrow \mathbb{R}$ durch

$$l(x) := -\int_{-1}^{-\frac{1}{2}} x(t)\,dt + \int_{-\frac{1}{2}}^{\frac{1}{2}} x(t)\,dt - \int_{\frac{1}{2}}^{1} x(t)\,dt = \int_{-1}^{1} w(t)x(t)\,dt$$

mit

$$w(t) := \begin{cases} -1, & t \in [-1, -\frac{1}{2}], \\ 1, & t \in [-\frac{1}{2}, \frac{1}{2}], \\ -1, & t \in [\frac{1}{2}, 1] \end{cases}$$

Offenbar ist l linear, wegen

$$|l(x)| \leq \int_{-1}^{1} |w(t)|\,|x(t)|\,dt \leq \int_{-1}^{1} |x(t)|\,dt = \|x\|_1$$

ist l auch stetig und $\|l\| \leq 1$. Wegen

$$l(x^* - z) = \int_{-1}^{1} w(t)\,(x^*(t) - z(t))\,dt = \int_{-1}^{1} |x^*(t) - z(t)|\,dt = \|x^* - z\|_1$$

und $x^* \neq z$ (wegen $z \notin \Pi_1$) ist $\|l\| = 1$. Um Satz 3.2.10 anwenden zu können, müssen wir noch nachweisen, dass $l(x - x^*) \geq 0$ für alle $x \in \Pi_1$. Um $l(x)$ für $x \in \Pi_1$ auszurechnen, setzen wir $x_0(t) := 1$ und $x_1(t) := t$ ein. Es ist

$$l(x_0) = -\frac{1}{2} + 1 - \frac{1}{2} = 0$$

und

$$l(x_1) = -\frac{1}{2}\left(\frac{1}{4} - 1\right) + \frac{1}{2}\left(\frac{1}{4} - \frac{1}{4}\right) - \frac{1}{2}\left(1 - \frac{1}{4}\right) = 0.$$

Also ist $l(x) = 0$ für alle $x \in \Pi_1$ und damit auch $0 = l(x - x^*)$ für alle $x \in \Pi_1$. Satz 3.2.10 liefert die Optimalität von x^*. □

3.3 Eigenschaften von T-Mengen

In diesem Abschnitt untersuchen wir die Frage, welche Eigenschaften T-Mengen[2] *notwendig* haben müssen. So soll z. B. gezeigt werden, dass unter bestimmten zusätzlichen Bedingungen T-Mengen in *glatten Räumen* (Definition folgt gleich) konvex sein müssen. Untersuchungen dieser Art stehen im Zusammenhang mit der berühmten Frage von V. Klee: Ist eine T-Menge in einem Hilbertraum notwendig konvex? Diese Problemstellung ist sicherlich hauptsächlich von theoretischem Interesse, weil sie aber die Approximationstheorie in linearen normierten Räumen stark beeinflusst hat, gehen wir hierauf in diesem Abschnitt ein.

[2]Zur Erinnerung: Die Teilmenge M eines linearen normierten Raumes X heißt *Tschebyscheff-Menge* oder kurz *T-Menge*, wenn es zu jedem $z \in X$ genau eine beste Approximierende an z in M gibt.

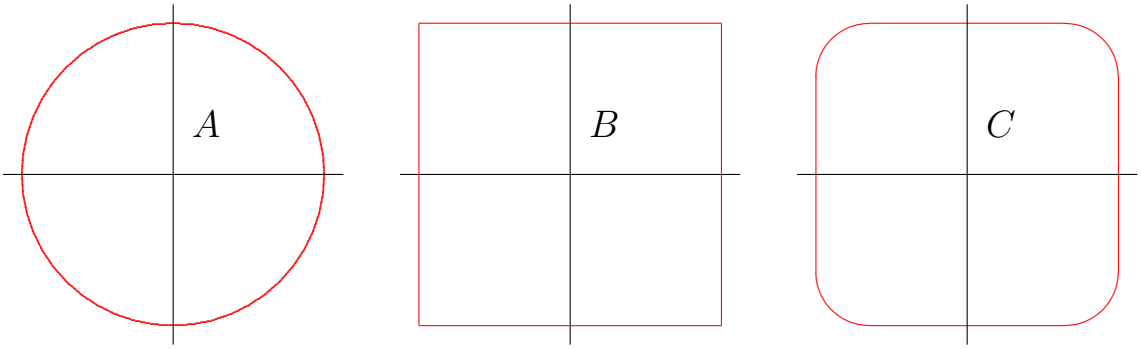

Abbildung 3.8: Einheitskugeln im $\mathbb{R}^2$ bezüglich geeigneter Norm

Definition 3.3.1 Ein linearer normierter Raum $(X, \|\cdot\|)$ heißt *glatt*, wenn jeder Randpunkt der Einheitskugel $B[0;1]$ in X Stützpunkt genau einer Stützhyperebene[3] ist.

Beispiel: In Abbildung 3.8 geben wir drei (abgeschlossene) Mengen A, B und C im $\mathbb{R}^2$ an, die jeweils (abgeschlossene) Einheitskugel einer geeigneten Norm auf $\mathbb{R}^2$ sind. Für die Mengen A und B ist das klar, die zugrunde liegende Norm ist die euklidische Norm bzw. die Maximumnorm. Natürlich ist $(\mathbb{R}^2, \|\cdot\|_2)$ sowohl strikt konvex[4] als auch glatt, während $(\mathbb{R}^2, \|\cdot\|_\infty)$ weder strikt konvex noch glatt ist. Bezüglich welcher Norm ist aber C die abgeschlossene Einheitskugel? Hierzu definieren wir die Abbildung (sogenanntes *Minkowski-Funktional*) $p_C\colon \mathbb{R}^2 \longrightarrow \mathbb{R}$ durch

$$p_C(x) := \inf\{\lambda > 0 : x \in \lambda C\}.$$

Dass p_C wohldefiniert ist bzw. $\{\lambda > 0 : x \in \lambda C\} \neq \emptyset$ für alle $x \in \mathbb{R}^2$ gilt, liegt daran, dass 0 ein innerer Punkt von C ist. Wir wollen uns davon überzeugen, dass $p_C(\cdot)$ eine Norm ist und

$$C = \{x \in \mathbb{R}^2 : p_C(x) \leq 1\}$$

gilt. Natürlich ist $p_C(x) \geq 0$ für alle $x \in \mathbb{R}^2$ und $p_C(x) = 0$ genau dann wenn $x = 0$. Zum Nachweis der Homogenität beachte man, dass C *symmetrisch* ist, also $x \in C$ genau dann, wenn $-x \in C$. Für $\alpha \neq 0$ und $\lambda > 0$ ist daher offenbar $\alpha x \in \lambda C$ genau dann, wenn $x \in (\lambda/|\alpha|)C$ und folglich

$$\begin{aligned} p_C(\alpha x) &= \inf\{\lambda > 0 : \alpha x \in \lambda C\} \\ &= \inf\{\lambda > 0 : x \in (\lambda/|\alpha|)C\} \\ &= |\alpha| \inf\{\lambda/|\alpha| > 0 : x \in (\lambda/|\alpha|)C\} \\ &= |\alpha| \inf\{\mu > 0 : x \in \mu C\} \\ &= |\alpha| p_C(x). \end{aligned}$$

[3] In Definition 2.2.9 sind die Begriffe *Stützhyperebene* und *Stützpunkt* eingeführt worden.

[4] Zur Erinnerung, siehe Definition 3.1.7: Ein linearer normierter Raum $(X, \|\cdot\|)$ heißt *strikt konvex*, falls

$$\|x\| = \|y\| = 1,\ x \neq y \Longrightarrow \|\tfrac{1}{2}(x+y)\| < 1.$$

Beim Nachweis der Dreiecksungleichung für $p_C(\cdot)$ ist die Konvexität von C entscheidend. Seien $x, y \in \mathbb{R}^2$ beliebig und s, t reelle Zahlen mit $s > p_C(x)$, $t > p_C(y)$. Wir überlegen uns, dass dann $x \in sC$, $y \in tC$. Denn: Nach Definition von $p_C(x)$ existiert ein $\lambda \in [0, s)$ mit $x = \lambda c \in \lambda C$ mit $c \in C$. Daher ist

$$x = s\Big[\underbrace{\frac{\lambda}{s}c + \left(1 - \frac{\lambda}{s}\right) \cdot 0}_{\in C}\Big] \in sC,$$

wobei wir die Konvexität von C und $0 \in C$ ausgenutzt haben. Entsprechend ist $y \in tC$. Daher ist

$$x + y \in sC + tC = (s+t)\left(\frac{s}{s+t}C + \frac{t}{s+t}C\right) \subset (s+t)C,$$

wobei wir wieder die Konvexität von C benutzt haben. Daher ist $p_C(x+y) \leq s + t$. Da dies für *alle* s, t mit $s > p_C(x)$, $t > p_C(y)$ gilt, ist die Dreiecksungleichung

$$p_C(x + y) \leq p_C(x) + p_C(y)$$

bewiesen. Also ist $p_C(\cdot)$ eine Norm auf $\mathbb{R}^2$. Aus $x = 1 \cdot x \in C$ folgt $p_C(x) \leq 1$, d. h. es ist

$$C \subset \{x \in \mathbb{R}^2 : p_C(x) \leq 1\}.$$

Ist $p_C(x) < 1$, so ist (siehe obige Überlegung) $x \in C$. Wir haben also bewiesen, dass

$$\{x \in \mathbb{R}^2 : p_C(x) < 1\} \subset C \subset \{x \in \mathbb{R}^2 : p_C(x) \leq 1\}.$$

Wegen der Abgeschlossenheit von C folgt $C = \{x \in \mathbb{R}^2 : p_C(x) \leq 1\}$. Damit haben wir eine Norm gefunden bezüglich der C gerade die abgeschlossene Einheitskugel ist. Offenbar ist $(\mathbb{R}^2, p_C(\cdot))$ zwar glatt, aber nicht strikt konvex[5]. □

Als kleine Fingerübung beweisen wir die folgenden Aussagen (siehe z. B. V. Barbu, T. Precupanu (2012, S. 35)).

Satz 3.3.2 *Sei $(X, \|\cdot\|)$ ein linearer normierter Raum. Dann gilt:*

1. *Ist $(X^*, \|\cdot\|)$ strikt konvex, so ist $(X, \|\cdot\|)$ glatt.*
2. *Ist $(X^*, \|\cdot\|)$ glatt, so ist $(X, \|\cdot\|)$ strikt konvex.*
3. *Ist $(X, \|\cdot\|)$ reflexiv und glatt, so ist $(X^*, \|\cdot\|)$ strikt konvex.*
4. *Ist $(X, \|\cdot\|)$ reflexiv und strikt konvex, so ist $(X^*, \|\cdot\|)$ glatt.*

[5]Man gebe eine abgeschlossene, konvexe Menge $D \subset \mathbb{R}^2$ an, die 0 im Innern enthält und für die $(\mathbb{R}^2, p_D(\cdot))$ zwar strikt konvex, aber nicht glatt ist.

Beweis: Zum Beweis der ersten Aussage des Satzes sei $x \in X$ mit $\|x\| = 1$ ein Randpunkt der Einheitskugel $B[0;1]$ in X. Wir haben zu zeigen, dass x Stützpunkt genau einer Stützhyperebene für $B[0;1]$ ist. Wegen $\{x\} \cap B(0;1) = \emptyset$ und dem Trennungssatz 2.2.5 von Eidelheit (bzw. dem Satz 2.2.10) ist x Stützpunkt von $B[0;1]$, es existiert also eine Stützhyperebene für $B[0;1]$ mit Stützpunkt x. Mit einem $l \in X^* \setminus \{0\}$, o. B. d. A. $\|l\| = 1$, ist

$$H := \{y \in X : l(y) = l(x)\}$$

genau dann eine Stützhyperebene für $B[0;1]$ mit Stützpunkt x, wenn $l(y) \leq l(x)$ für alle $y \in B[0;1]$. Dies wiederum ist gleichwertig mit

$$1 = \|l\| = \sup_{y \in B[0;1]} |l(y)| \leq l(x) \leq \|l\| \, \|x\| = 1.$$

Zu zeigen ist also, dass genau ein $l \in X^*$ mit $\|l\| = 1$ und $l(x) = 1$ existiert. Die Existenz ist gerade bewiesen worden, die Eindeutigkeit folgt aus der strikten Konvexität von $(X^*, \|\cdot\|)$. Denn angenommen, l_1 und l_2 sind zwei Elemente aus X^* mit $\|l_1\| = \|l_2\| = 1$ sowie $l_1(x) = l_2(x) = 1$. Ist $l_1 \neq l_2$, so ist $\|\frac{1}{2}(l_1 + l_2)\| < 1$ wegen der strikten Konvexität von X^*. Aus $1 = \frac{1}{2}(l_1 + l_2)(x)$ folgt

$$1 \leq \|\tfrac{1}{2}(l_1 + l_2)\| \, \|x\| = \|\tfrac{1}{2}(l_1 + l_2)\|,$$

ein Widerspruch.

Nun sei $(X^*, \|\cdot\|)$ glatt, weiter seien $x_1, x_2 \in X$ mit $\|x_1\| = \|x_2\| = 1$ und $x_1 \neq x_2$ gegeben. Wir haben zu zeigen, dass $\|\frac{1}{2}(x_1 + x_2)\| < 1$. Angenommen, das sei nicht der Fall, es sei also $\|\frac{1}{2}(x_1 + x_2)\| = 1$. Wegen des Korollars 2.2.12 zum Satz von Hahn-Banach existiert ein $l_0 \in X^*$ mit

$$l_0(\tfrac{1}{2}(x_1 + x_2)) = \|\tfrac{1}{2}(x_1 + x_2)\| = 1.$$

Hieraus folgt $l_0(x_1) = l_0(x_2) = 1$ (da $l_0(x_1) \leq \|l_0\| \, \|x_1\| = 1$ und entsprechend $l_0(x_2) \leq 1$). Durch

$$H_1^* := \{l \in X^* : l(x_1) = 1\}, \qquad H_2^* := \{l \in X^* : l(x_2) = 1\}$$

sind dann zwei verschiedene Stützhyperebenen der Einheitskugel in X^* mit dem Stützpunkt l_0 gegeben, was ein Widerspruch zur Glattheit von $(X^*, \|\cdot\|)$ ist.

Die beiden restlichen Aussagen folgen aus den beiden ersten, indem man berücksichtigt, dass für reflexives $(X, \|\cdot\|)$ dieser Raum mit $((X^*)^*, \|\cdot\|)$ identifiziert werden kann. □

Grundlage der folgenden Ausführungen, vor allem aber auch Motivation für die Definition einer *Sonne*, ist eine Beobachtung, die als ein einfaches Lemma formuliert wird.

Lemma 3.3.3 *Ist $M \subset X$ und $x^* \in M$ eine beste Approximierende an $z \in X \setminus M$ in M, so ist x^* auch beste Approximierende an $z_\lambda := x^* + \lambda(z - x^*)$ in M für jedes $\lambda \in [0,1]$, also jeden Punkt der Verbindungsstrecke von x^* und z.*

Beweis: Sei $x \in M$ beliebig, $\lambda \in [0,1]$. Dann ist

$$\begin{aligned} \|x - z_\lambda\| &\geq \|x - z\| - \|z - z_\lambda\| \\ &\geq \|x^* - z\| - \|z - z_\lambda\| \\ &= \|x^* - z\| - \|(1-\lambda)(z - x^*)\| \\ &= \lambda \|x^* - z\| \\ &= \|x^* - z_\lambda\|, \end{aligned}$$

und dies ist die Behauptung. □

Definition 3.3.4 Eine Existenzmenge $M \subset X$ heißt eine *Sonne*, wenn es zu jedem $z \notin M$ eine beste Approximierende $x^* \in M$ an z in M gibt, die für jedes $\lambda > 0$ auch beste Approximierende an $z_\lambda := x^* + \lambda(z - x^*)$ in M ist, also jeden Punkt der von x^* ausgehenden Geraden durch z. Eine solche beste Approximierende heißt *Sonnenpunkt* für z. Gilt dies sogar für *jede* beste Approximierende an z, so heißt M eine *strikte Sonne*.

In Abbildung 3.9 veranschaulichen wir uns die Definition einer Sonne.

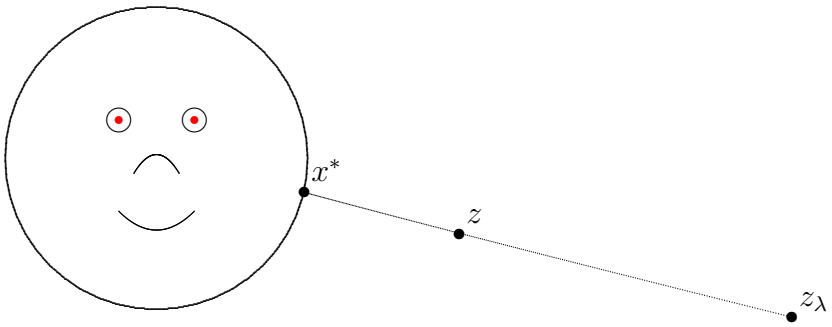

Abbildung 3.9: Veranschaulichung der Definition einer Sonne

Bemerkungen: 1. Es gilt:

- *Eine konvexe Existenzmenge ist eine strikte Sonne.*

Denn: Sei $z \notin M$ und $x^* \in P_M(z)$ eine beste Approximierende an z in M und $z_\lambda := x^* + \lambda(z - x^*)$. Ist $x \in M$, so ist $\|x - z_\lambda\| \geq \|x^* - z_\lambda\|$ für $\lambda \in [0,1]$ schon in Lemma 3.3.3 bewiesen worden. Für $\lambda > 1$ ist

$$\|x - z_\lambda\| = \lambda \left\| \underbrace{\frac{1}{\lambda} x + \frac{\lambda - 1}{\lambda} x^*}_{\in M} - z \right\| \geq \lambda \|x^* - z\| = \|x^* - z_\lambda\|.$$

Daher ist x^* für jedes $\lambda > 0$ eine beste Approximierende an z_λ in M und daher eine strikte Sonne.

2. Ist M eine T-Menge, so ist M natürlich genau dann eine Sonne, wenn M eine strikte Sonne ist. I. Allg. ist dies nicht richtig, denn es gibt Sonnen, die keine strikten Sonnen sind. Man überlege sich hierzu ein Beispiel.

3. Eher beiläufig (weil es so einfach ist) haben wir im Anschluss an die Definition 3.1.7 des Begriffes "strikt konvex" erwähnt, dass eine konvexe Existenzmenge in einem strikt konvexen Raum eine T-Menge ist. Es gilt sogar:

- *Eine Sonne in einem strikt konvexen Raum ist eine T-Menge.*

Denn: Sei $z \notin M$. Da M nach Voraussetzung eine Sonne ist, gibt es einen Sonnenpunkt $x_1^* \in P_M(z)$, also eine beste Approximierende an z, die auch noch beste Approximierende an jeden Punkt z_λ des von x_1^* ausgehenden Strahls durch z ist, insbesondere an

$$z_2 = x_1^* + 2(z - x_1^*) = 2z - x_1^*.$$

Sei $x_2^* \neq x_1^*$ eine weitere beste Approximierende an z in M. Wegen der strikten Konvexität ist $\|\frac{1}{2}(x_1^* + x_2^*) - z\| < \delta$ mit $\delta := \|x_1^* - z\| = \|x_2^* - z\|$. Dann ist aber

$$\begin{aligned} \|x_2^* - (2z - x_1^*)\| &= 2\,\|\tfrac{1}{2}(x_1^* + x_2^*) - z\| \\ &< 2\delta \\ &= 2\,\|x_1^* - z\| \\ &= \|x_1^* - (2z - x_1^*)\|, \end{aligned}$$

ein Widerspruch dazu, dass x_1^* beste Approximierende an $2z - x_1^*$ ist.

4. Es gilt (siehe D. Braess (1986, S. 32)):

- *Ist $M \subset X$ eine strikte Sonne und $z \notin M$, so ist eine lokal beste Approximierende $x^* \in M$ an z in M sogar eine (global) beste Approximierende an $z$in M.*

Denn: Da $x^* \in M$ eine lokal beste Approximierende an z in M ist, existiert ein $r > 0$ mit $x^* \in P_{M \cap B[x^*;r]}(z)$. Sei $\lambda := r/(2\,\|x^* - z\|)$, indem man r notfalls kleiner wählt kann man o. B. d. A. annehmen, dass $\lambda \in (0, 1]$. Anschließend definieren wir $z_\lambda := x^* + \lambda(z - x^*)$. Wir wollen uns überlegen, dass $x^* \in P_M(z_\lambda)$. Da z auf dem von x^* ausgehenden Strahl durch z_λ liegt und M eine strikte Sonne ist, ist dann $x^* \in P_M(z)$, die Behauptung also bewiesen.

(a) Sei $x \in M \cap B[x^*; r]$.

Dann ist

$$\begin{aligned} \|x - z_\lambda &= \|x - x^* - \lambda(z - x^*)\| \\ &= \|x - z - (1 - \lambda)(x^* - z)\| \\ &\geq \|x - z\| - (1 - \lambda)\,\|x^* - z\| \\ &= \underbrace{\|x - z\| - \|x^* - z\|}_{\geq 0} + \lambda\,\|x^* - z\| \\ &\geq \lambda\,\|x^* - z\| \\ &= \|x^* - z_\lambda\|. \end{aligned}$$

(b) Sei $x \in M \setminus B[x^*; r]$.

Dann ist

$$\begin{aligned} \|x - z_\lambda\| &\geq \|x - x^*\| - \lambda \|x^* - z\| \\ &\geq r - \frac{r}{2} \\ &= \frac{r}{2} \\ &= \lambda \|x^* - z\| \\ &= \|x^* - z_\lambda\|. \end{aligned}$$

Damit ist obige Aussage bewiesen. □

Es gibt viele interessante Sätze über T-Mengen und Sonnen, insbesondere darüber, unter welchen Bedingungen diese konvex sind. Wir können und wollen nur eine kleine Auswahl angeben. Ziel ist ein Satz der folgenden Art:

- *Unter "gewissen Voraussetzungen" ist eine T-Menge konvex.*

Hierzu wird gezeigt:

1. Unter "gewissen Voraussetzungen" ist eine T-Menge eine Sonne.
2. Charakterisierung von Sonnen.
3. Unter "gewissen Voraussetzungen" ist eine Sonne konvex.

Der tiefste Satz wird der erste sein. Hierzu benötigen wir eine Definition.

Definition 3.3.5 Eine Menge $M \subset X$ heißt *beschränkt kompakt*, wenn der Durchschnitt von M mit jeder abgeschlossenen Kugel kompakt ist.

Offenbar gilt (siehe Definition 3.1.5):

$$\begin{aligned} M \text{ kompakt} &\Longrightarrow M \text{ beschränkt kompakt}, \\ &\Longrightarrow M \text{ approximativ kompakt}, \\ &\Longrightarrow M \text{ approximativ schwach kompakt}, \\ &\Longrightarrow M \text{ Existenzmenge}, \\ &\Longrightarrow M \text{ abgeschlossen}. \end{aligned}$$

Z. B. sind abgeschlossene Mengen in einem endlichdimensionalen Raum und endlichdimensionale Teilräume eines linearen normierten Raumes beschränkt kompakt.

Dass der folgende Satz nicht ganz trivial ist, sieht man daran, dass zu seinem Beweis das folgende klassische Ergebnis benutzt wird, das wir jetzt ohne Beweis angeben.

Schauderscher Fixpunktsatz: *Sei $(X, \|\cdot\|)$ ein Banachraum, $A \subset X$ nichtleer, abgeschlossen und konvex und $\Psi\colon A \longrightarrow X$ stetig. Ferner gelte*

1. *$\Psi(A) \subset A$, d. h. Ψ bildet A in sich ab,*

2. $\mathrm{cl}\,(\Psi(A))$ *ist kompakt (bzw.* $\Psi(A)$ *relativ kompakt.*

Dann besitzt A *mindestens einen Fixpunkt in* A*, es existiert also ein* $x \in A$ *mit* $\Psi(x) = x$*.*

Ist $X = \mathbb{R}^n$ und $A = B[0;1]$ die abgeschlossene Einheitskugel, so erhält man aus dem Schauderschen den Brouwerschen Fixpunktsatz, dass also eine stetige Abbildung der Einheitskugel in sich mindestens einen Fixpunkt besitzt. Andererseits wird der Brouwersche Fixpunktsatz zum Beweis des Schauderschen Fixpunktsatzes benötigt.

Nun kommt der angekündigte erste Satz (siehe D. Braess (1986, S. 40), dort findet man auch genauere Literaturhinweise).

Satz 3.3.6 *Jede beschränkt kompakte T-Menge* M *in einem Banachraum* $(X, \|\cdot\|)$ *ist eine (strikte) Sonne.*

Beweis: Eine beschränkt kompakte Menge ist approximativ kompakt (denn jede Minimalfolge ist beschränkt). Die metrische Projektion $P_M\colon X \longrightarrow M \subset X$ auf die approximativ kompakte T-Menge M ist nach Satz 3.1.6 stetig. Wir machen einen Widerspruchsbeweis und nehmen an, M sei keine Sonne. Dann gibt es ein $z \in X$ mit der Eigenschaft, dass es auf dem von $x^* := P_M(z)$ ausgehenden Strahl durch z einen Punkt gibt, für den x^* nicht die beste Approximierende ist. Da dieser Punkt wegen Lemma 3.3.3 nicht zu der Strecke zwischen x^* und z liegen kann, existiert ein $\lambda_0 > 1$ mit $x^* \neq P_M(z_{\lambda_0})$, wobei z_λ für $\lambda > 0$ durch $z_\lambda := x^* + \lambda(z - x^*)$ definiert ist. Wegen Lemma 3.3.3 ist $x^* = P_M(z_\lambda)$ für $\lambda \in (0,1]$. Daher ist die folgende Definition sinnvoll:

$$\lambda_1 := \sup\{\lambda > 0 : x^* = P_M(z_\lambda)\}.$$

Wiederum wegen Lemma 3.3.3 ist $x^* = P_M(z_\lambda)$ für alle $\lambda \in (0, \lambda_1)$ und daher $\lambda_1 \leq \lambda_0$. Wegen der Stetigkeit der metrischen Projektion ist

$$x^* = \lim_{\lambda \nearrow \lambda_1} P_M(z_\lambda) = P_M(z_{\lambda_1}).$$

Daher ist $1 \leq \lambda_1 < \lambda_0$. Zur Abkürzung setzen wir $z^1 := z_{\lambda_1}$. Dann ist x^* beste Approximierende an z^1, aber kein $z^1_\lambda := x^* + \lambda(z^1 - x^*)$ mit $\lambda > 1$ besitzt x^* als beste Approximierende. Da $z^1 \notin M$ und M als beschränkt kompakte Menge insbesondere abgeschlossen ist, haben z^1 und M einen positiven Abstand voneinander und $r > 0$ kann so klein gewählt werden, dass $M \cap B[z^1; 2r] = \emptyset$. Zur Abkürzung setzen wir $B := B[z^1; r]$ und definieren $\Phi\colon M \longrightarrow B$ durch

$$\Phi(x) := z^1 + r \cdot \frac{z^1 - x}{\|z^1 - x\|}$$

sowie $\Psi\colon B \longrightarrow B$ durch $\Psi := \Phi \circ P_M$. In Abbildung 3.10 machen wir uns die Konstruktion der Abbildung Φ klar. Wir wollen die Voraussetzungen des Schauderschen Fixpunktsatzes nachprüfen. Als Komposition stetiger Abbildungen ist Ψ eine stetige Abbildung. Die Menge B ist nichtleer, abgeschlossen und konvex, ferner ist $\psi(B) \subset B$.

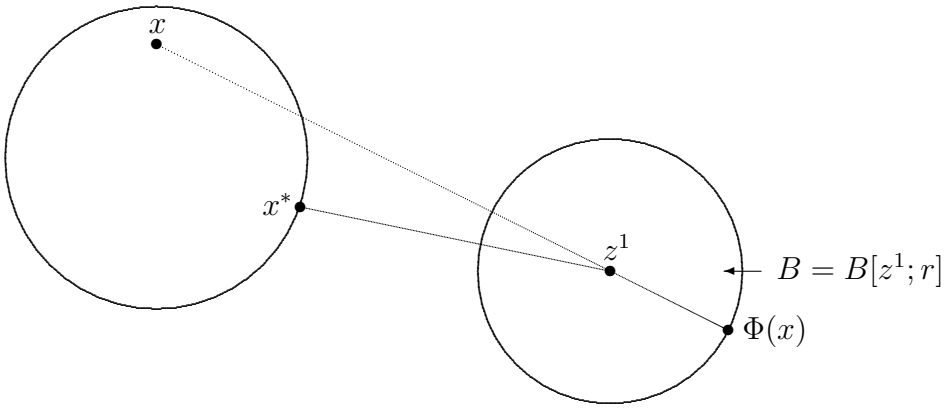

Abbildung 3.10: Die Konstruktion von Φ

Zu zeigen bleibt, dass $\mathrm{cl}\,(\Psi(B))$ kompakt ist. Hierzu zeigen wir zunächst, dass $P_M(B)$ beschränkt ist. Denn ist $y \in B$, so ist

$$\begin{aligned}
\|P_M(y)\| &\leq \|y\| + \|P_M(y) - y\| \\
&\leq \|y\| + \|x^* - y\| \\
&\leq \|y - z^1\| + \|z^1\| + \|x^* - z^1\| + \|z^1 - y\| \\
&\leq 2r + \|z^1\| + \|x^* - z^1\| \\
&=: c.
\end{aligned}$$

Also ist $P_M(B) \subset M \cap B[0;c]$. Da M beschränkt kompakt ist, ist $M \cap B[0;c]$ und damit auch $\Phi(M \cap B[0;c])$ kompakt. Wegen $\Psi(B) \subset \Phi(M \cap B[0;c])$ ist $\mathrm{cl}\,(\Psi(B))$ kompakt. Aus dem Schauderschen Fixpunktsatz erhält man die Existenz von $h \in B$ mit $\Psi(h) = \Phi(P_M(h)) = h$. Nach Konstruktion ist Φ eine Abbildung mit der Eigenschaft, dass z^1 für jedes $x \in M$ auf der Verbindungsstrecke zwischen x und $\Phi(x)$ liegt, siehe Abbildung 3.10. Insbesondere liegt z^1 auf der Verbindungsstrecke von $P_M(h)$ zu h. Also ist

$$\begin{aligned}
\|P_M(h) - h\| &= \|P_M(h) - z^1\| + \|z^1 - h\| \\
&\geq \|x^* - z^1\| + \|z^1 - h\| \\
&\qquad \text{(da } x^* \text{ beste Approximierende an } z^1\text{)} \\
&\geq \|x^* - h\|.
\end{aligned}$$

Aus der Eindeutigkeit der besten Approximierenden folgt $x^* = P_M(h)$. Damit ist $\Phi(x^*) = h$ bzw.

$$\begin{aligned}
h &= z^1 + \frac{r}{\|z^1 - x^*\|}(z^1 - x^*) \\
&= x^* + \underbrace{\left(1 + \frac{r}{\|z^1 - x^*\|}\right)}_{=:\lambda > 1}(z^1 - x^*) \\
&= z^1_\lambda.
\end{aligned}$$

Also ist x^* auch für z^1_λ mit einem $\lambda > 1$ eine beste Approximierende, was einen Widerspruch zur Definition von z^1 bedeutet. Der Satz ist damit schließlich bewiesen. □

Als einfaches Korollar zum vorigen Satz notieren wir:

Korollar 3.3.7 *Eine T-Menge M in einem endlichdimensionalen linearen normierten Raum ist eine Sonne.*

Beweis: Eine T-Menge ist als Existenzmenge abgeschlossen. Abgeschlossene Mengen in einem linearen normierten Raum sind beschränkt kompakt. Da außerdem endlichdimensionale Räume vollständig sind, liefert Satz 3.3.6 die Behauptung. □

Auf unserem Weg zu einer Aussage der Form

- *Unter "gewissen Voraussetzungen" ist eine T-Menge konvex*

geben wir nun eine notwendige und hinreichende Bedingung dafür an, dass $M \subset X$ eine strikte Sonne ist. Der Gültigkeitsbereich des verallgemeinerten Kolmogoroff-Kriteriums (Satz 3.2.9) wird damit von konvexen Mengen auf strikte Sonnen ausgedehnt.

Satz 3.3.8 *Sei $(X, \|\cdot\|)$ ein linearer normierter Raum und $M \subset X$ eine Existenzmenge. Dann sind die folgenden beiden Bedingungen äquivalent:*

(a) M ist eine strikte Sonne.

(b) Für jedes $z \notin M$ ist $x^* \in P_M(z)$ (bzw. x^* eine beste Approximierende an z in M) genau dann, wenn es zu jedem $x \in M$ ein $l \in X^*$ mit

$$(*) \qquad \|l\| = 1, \quad l(x^* - z) = \|x^* - z\|, \quad 0 \le l(x - x^*)$$

gibt.

Beweis: Sei M eine strikte Sonne, die Bedingung (a) also erfüllt. Sei $z \notin M$ und $x^* \in P_M(z)$. Nach Definition einer strikten Sonne ist $x^* \in P_M(x^* + \lambda(z - x^*))$ für jedes $\lambda > 0$. Für jedes $x \in M$ und alle $\lambda > 0$ ist also

$$\begin{aligned} \lambda \|x^* - z\| &= \|x^* - (x^* + \lambda(z - x^*))\| \\ &\le \|x - (x^* + \lambda(z - x^*))\| \\ &= \lambda \left\| x^* + \frac{1}{\lambda}(x - x^*) - z \right\|. \end{aligned}$$

Mit $f(x) := \|x - z\|$ und $t > 0$ ist daher (setze $t = 1/\lambda$)

$$\frac{1}{t}[f(x^* + t(x - x^*)) - f(x^*)]$$

und folglich

$$0 \le f'(x^*; x - x^*) = \max_{l \in \partial f(x^*)} l(x - x^*),$$

wobei wir Satz 3.2.8 benutzt haben. Wegen $z \notin M$ ist $x^* \neq z$ und folglich (siehe das Beispiel auf S. 65)

$$\partial f(x^*) = \{l \in X^* : \|l\| = 1,\ l(x^* - z) = \|x^* - z\|\}.$$

Folglich existiert zu jedem $x \in M$ ein $l \in X^*$ mit $(*)$. Existiert umgekehrt zu jedem $x \in M$ ein $l \in X^*$ mit $(*)$, so ist $x^* \in P_M(x^*)$, wie wir am Schluss des Beweises von Satz 3.2.9 gezeigt haben (hier wird die Konvexität von M nicht benutzt). Damit ist gezeigt, dass aus (a) die Bedingung (b) folgt.

Nun sei die Bedingung (b) erfüllt. Sei $z \notin M$ und $x^* \in P_M(z)$. Sei weiter $\lambda > 0$ und $z_\lambda := x^* + \lambda(z - x^*)$. Zu zeigen ist $x^* \in P_M(z_\lambda)$. Hierzu sei $x \in M$ und $l \in X^*$ ein stetiges lineares Funktional, das $(*)$ genügt. Dann ist

$$\begin{aligned} \|x^* - z_\lambda\| &= \lambda \|x^* - z\| \\ &= \lambda\, l(x^* - z) \\ &= \underbrace{l(x^* - x)}_{\leq 0} + l(x - (\underbrace{x^* + \lambda(z - x^*)}_{= z_\lambda})) \\ &\leq l(x - z_\lambda) \\ &\leq \|l\|\, \|x - z_\lambda\| \\ &= \|x - z_\lambda\|, \end{aligned}$$

also $x^* \in P_M(z_\lambda)$. Damit ist gezeigt, dass M eine strikte Sonne ist bzw. die Bedingung (a) erfüllt ist. □

Mit Hilfe des folgenden Satzes erhalten wir ein erstes Ergebnis der gewünschten Art (siehe D. Braess (1986, S. 34)).

Satz 3.3.9 *Die folgenden beiden Aussagen sind in einem linearen normierten Raum $(X, \|\cdot\|)$ äquivalent:*

(a) $(X, \|\cdot\|)$ ist glatt.

(b) Jede Sonne in X ist konvex.

Beweis: Um (a)$\Longrightarrow$(b) zu zeigen, nehmen wir an, M sei eine Sonne in dem glatten Raum $(X, \|\cdot\|)$. Seien $x_1, x_2 \in M$ und $\alpha \in (0,1)$. Angenommen, $z := (1-\alpha)x_1 + \alpha x_2 \notin M$. Da M eine Sonne ist, gibt es ein $x^* \in P_M(z)$ mit $x^* \in P_M(z_\lambda)$ für alle $\lambda > 0$, wobei wieder $z_\lambda := x^* + \lambda(z - x^*)$ gesetzt wurde. Wie im ersten Teil des Satzes 3.3.8 gezeigt wurde, folgt hieraus mit $f(x) := \|x - z\|$, dass

$$0 \leq f'(x^*; x - x^*) = \max_{l \in \partial f(x^*)} l(x - x^*) \quad \text{für alle } x \in M.$$

Das Subdifferential $\partial f(x^*)$ von f in x^* ist wegen $x^* \neq z$ wieder durch

$$\partial f(x^*) = \{l \in X^* : \|l\| = 1,\ l(x^* - z) = \|x^* - z\|\}$$

gegeben. Wendet man dies mit $x = x_1, x_2$ an, so erhält man die Existenz von $l_1, l_2 \in X^*$ mit

$$\|l_i\| = 1, \quad l_i(x^* - z) = \|x^* - z\|, \quad 0 \leq l_i(x_i - x^*) \quad (i = 1, 2).$$

Da $z \notin M$ ist $(x^* - z)/\|x^* - z\|$ ein Punkt der Einheitskugel $B[0,1]$, der auf den Hyperebenen $H_i := \{y \in X : l_i(y) = 1\}$, $i = 1, 2$, liegt, welche wegen $\|l_i\| = 1$, $i = 1, 2$,

Stützhyperebenen für $B[0;1]$ sind. Da $(X, \|\cdot\|)$ glatt ist, ist $l_1 = l_2$. Mit $l := l_1 = l_2$ ist daher

$$\begin{aligned} \|x^* - z\| &= l(x^* - z) \\ &= l(x^* - (1-\alpha)x_1 - \alpha x_2) \\ &= -(1-\alpha)\underbrace{l(x_1 - z)}_{\geq 0} - \alpha \underbrace{l(x_2 - z)}_{\geq 0} \\ &\leq 0, \end{aligned}$$

ein Widerspruch zu der Annahme $z \notin M$. Damit ist (a)$\Longrightarrow$(b) nachgewiesen.

Zum Nachweis von (b)$\Longrightarrow$(a) zeigen wir: Wenn $(X, \|\cdot\|)$ nicht glatt ist, so existiert in X eine nichtkonvexe Sonne. Da X nicht glatt ist, existiert ein Randpunkt x der Einheitskugel mit zwei verschiedenen Stützhyperebenen H_1, H_2 mit x als Stützpunkt, also $l_1, l_2 \in X^*$ mit $l_1(x) = l_2(x) = 1$, $\|l_1\| = \|l_2\| = 1$ und $l_1 \neq l_2$. Nun definiere man

$$M_1 := \{y \in X : l_1(y) = 0,\ l_2(y) \geq 0\}, \quad M_2 := \{y \in X : l_2(y) = 0,\ l_1(y) \geq 0\}$$

und

$$M := M_1 \cup M_2.$$

Die Konstruktion der Menge M wollen wir uns anhand von Abbildung 3.11 deutlich machen. Hierbei sei $(X, \|\cdot\|) = (\mathbb{R}^2, \|\cdot\|_\infty)$. Wir haben die Einheitskugel und den

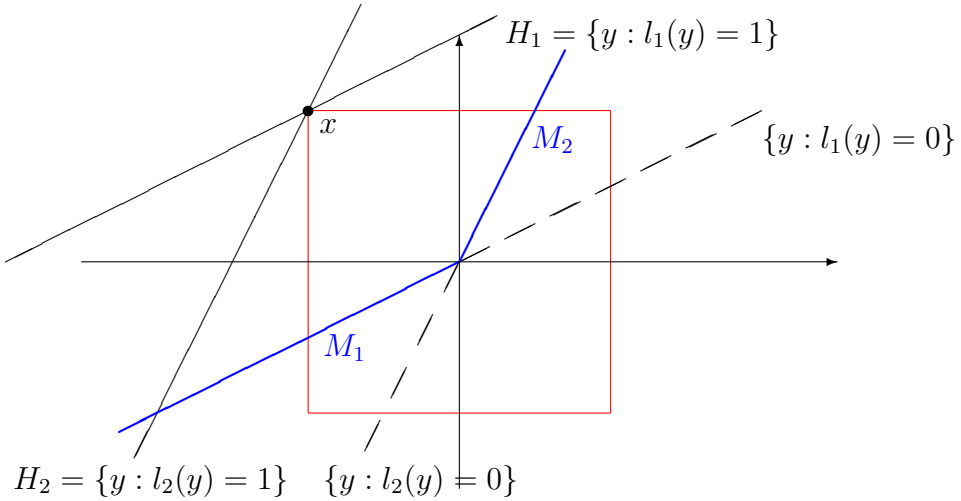

Abbildung 3.11: Die Konstruktion einer nichtkonvexen Sonne

Randpunkt x mit zwei verschiedenen Stützhyperebenen eingetragen. Wir werden jetzt zeigen, dass M eine nichtkonvexe Sonne ist. Um nachzuweisen, dass M nichtkonvex ist, nehmen wir an, es seien $y_1 \in (M_1 \setminus M_2) \subset M$ bzw. $l_1(y_1) = 0$, $l_2(y_1) > 0$ und $y_2 \in (M_2 \setminus M_1) \subset M$ bzw. $l_2(y_2) = 0$, $l_1(y_2) > 0$ und zeigen, dass ein Punkt auf der Verbindungsstrecke zwischen y_1 und y_2 nicht zu M gehört. Mit $\lambda \in (0,1)$ sei $y := (1-\lambda)y_1 + \lambda y_2$. Dann ist

$$l_1(y) = \lambda l_1(y_2) > 0, \qquad l_2(y) = (1-\lambda) l_2(y_1) > 0,$$

also $y \notin M$, womit gezeigt ist, dass M nichtkonvex ist. Nun zeigen wir, dass M eine Sonne ist. Sei $z \in X \setminus M$. Wir werden ein $x^* \in P_M(z)$ mit $x^* \in P_M(x^* + t(z - x^*))$ für alle $t > 0$ angeben, womit bewiesen sein wird, dass M eine Sonne ist. O. B. d. A. ist $l_1(z) \leq l_2(z)$ (andernfalls vertausche man l_1 und l_2). Sei $x^* := z - l_1(z) \cdot x$. Dann ist

$$l_1(x^*) = l_1(z) - l_1(z)\underbrace{l_1(x)}_{=1} = 0, \qquad l_2(x^*) = l_2(z) - l_1(z)\underbrace{l_2(x)}_{=1} \geq 0,$$

also $x^* \in M_1 \subset M$. Ferner ist $d(z, M) \leq \|x^* - z\| = |l_1(z)|$. Um $x^* \in P_M(z)$ zu beweisen, unterscheiden wir zwei Fälle. Im ersten Fall ist $l_1(z) \geq 0$. Ist $y \in M_1$, so ist

$$\|y - z\| \geq -l_1(y - z) = l_1(z) = \|x^* - z\|,$$

ist dagegen $y \in M_2$, so ist

$$\|y - z\| \geq -l_2(y - z) = l_2(z) \geq l_1(z) = \|x^* - z\|.$$

Im zweiten Fall ist $l_1(z) < 0$. Für alle $y \in M$ ist $l_1(y) \geq 0$ und daher

$$\|y - z\| \geq l_1(y - z) \geq -l_1(z) = |l_1(z)| = \|x^* - z\|.$$

Also ist $x^* \in P_M(z)$. Zu zeigen bleibt, dass $x^* \in P_M(x^* + t(z - x^*))$ für alle $t > 1$ (für $t \in [0, 1]$ ist dies wegen Lemma 3.3.3 sowieso klar). Zur Abkürzung sei

$$z_t := x^* + t(z - x^*) = z - (1 - t)l_1(z) \cdot x.$$

Dann ist $l_1(z_t) = tl_1(z)$ und

$$l_2(z_t) - l_1(z_t) = l_2(z) - (1 - t)l_1(z) - tl_1(z) = l_2(z) - l_1(z)$$

und $x^* \in P_M(z_t)$ kann ganz genau so bewiesen werden wie oben $x^* \in P_M(z)$. Damit ist alles bewiesen. □

Fügt man Satz 3.3.6 und Satz 3.3.9 zusammen, so erhält man (siehe D. Braess (1986, S. 41))

Satz 3.3.10 *Eine beschränkt kompakte T-Menge in einem glatten Banachraum ist konvex.*

Berücksichtigt man ferner, dass ein strikt konvexer endlichdimensionaler linearer normierter Raum ein uniform konvexer Banachraum (siehe eine Bemerkung im Anschluss an Definition 3.1.7), berücksichtigt man außerdem Satz 3.1.8, so erhält man leicht die folgende Aussage.

Satz 3.3.11 *In einem glatten, strikt konvexen endlichdimensionalen linearen normierten Raum $(X, \|\cdot\|)$ sind folgende Aussagen gleichwertig:*

(a) $M \subset X$ ist eine T-Menge.

(b) $M \subset X$ ist abgeschlossen und konvex.

Beweis: Es gelte (a), $M \subset X$ sei also ein T-Menge. Als Existenzmenge ist M abgeschlossen. Nach Korollar 3.3.7 ist die T-Menge M in dem endlichdimensionalen linearen normierten Raum X eine Sonne. Da weiter $(X, \|\cdot\|)$ als glatt vorausgesetzt wurde, ist nach Satz 3.3.9 die Sonne M konvex und damit (b) erfüllt. Umgekehrt sei (b) erfüllt, also $M \subset X$ abgeschlossen und konvex. Nach Satz 3.1.8 ist M eine (approximativ kompakte) T-Menge, also (a) erfüllt. □

Ohne Beweis geben wir an (siehe D. Braess (1986, S. 45)):

- *Eine approximativ kompakte T-Menge in einem uniform konvexen linearen normierten Raum ist eine Sonne.*

Da Hilberträume uniform konvex und glatt sind, ferner Sonnen in einem glatten Raum notwendig konvex sind, gilt:

- *Eine approximativ kompakte T-Menge in einem Hilbertraum ist konvex.*

3.4 Untere Schranken für den Minimalabstand, Dualität bei konvexen Approximationsaufgaben

Wiederum sei generell $(X, \|\cdot\|)$ ein linearer normierter Raum, $M \subset X$ eine Teilmenge und $z \in X$. Wir betrachten die (primale) Approximationsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \|x - z\|, \quad x \in M.$$

Der *Abstand* bzw. *Minimalabstand* von z zu M ist bekanntlich definiert durch

$$d(z, M) := \inf_{x \in M} \|x - z\|.$$

Obere Schranken für $d(z, M)$ zu gewinnen ist trivial, da $d(z, M) \leq \|x - z\|$ für alle $x \in M$. Sehr viel schwieriger ist es dagegen, *untere* Schranken zu erhalten. Weshalb ist dies ein wichtiges Problem? Durch untere Schranken und trivialerweise erhältliche obere Schranken für $d(z, M)$ sind Einschließungen des Minimalabstandes möglich. Wird eine obere Schranke durch eine Näherungslösung $\hat{x} \in M$ gefunden und ist eine bekannte untere Schranke nur "unwesentlich" kleiner als $\|\hat{x} - z\|$, so wird man in der Praxis eventuell mit $\hat{x}$ zufrieden sein.

In der Optimierung erhält man untere Schranken für den *Wert* einer Optimierungsaufgabe (dieser entspricht dem Minimalabstand bei approximationsaufgaben) durch den schwachen Dualitätssatz bzw. durch den Wert der dualen Zielfunktion in einem dual zulässigen Punkt. Wir wollen nun zeigen, dass man bei Approximationsaufgaben ganz ähnlich vorgehen kann und diese Ergebnisse dann auf T-Approximationsaufgaben anwenden.

Zur Herleitung der dualen Approximationsaufgabe definieren wir die Menge

$$\Lambda := \{(x - y, \|x - z\| + r) \in X \times \mathbb{R} : x \in X,\ y \in M,\ r \geq 0\}$$

und beachten:

1. Es ist $(0, \beta) \in M$ genau dann, wenn ein $x \in M$ mit $\|x - z\| \leq \beta$ existiert.

   Das Approximationsproblem, z durch Elemente aus M zu approximieren ist also äquivalent damit, β unter der Nebenbedingung $(0, \beta) \in \Lambda$ zu minimieren.

2. Ist M konvex, so ist auch Λ konvex.

   Denn: Seien $P_i := (x_i - y_i, \|x_i - z\| + r_i) \in \Lambda$, $i = 0, 1$ und $\lambda \in [0, 1]$. Zur Abkürzung setze man

   $$x_\lambda := (1-\lambda)x_0 + \lambda x_1, \quad y_\lambda := (1-\lambda)y_0 + \lambda y_1, \quad r_\lambda := (1-\lambda)r_0 + \lambda r_1.$$

   Dann ist

   $$\begin{aligned} & (1-\lambda)P_1 + \lambda P_2 \\ = \; & (1-\lambda)(x_1 - y_1, \|x_1 - z\| + r_1) \\ & \quad + \lambda(x_2 - y_2, \|x_2 - z\| + r_2) \\ = \; & (x_\lambda - \underbrace{y_\lambda}_{\in M}, (1-\lambda)\|x_1 - z\| + \lambda\|x_2 - z\| + \underbrace{r_\lambda}_{\geq 0}) \\ = \; & (x_\lambda - y_\lambda, \|x_\lambda - z\| + \underbrace{r_\lambda + (1-\lambda)\|x_1 - z\| + \lambda\|x_2 - z\| - \|x_\lambda - z\|}_{\geq 0}) \\ \in \; & \Lambda. \end{aligned}$$

Genau wie in der Optimierung stellt man zu der (primalen) Approximationsaufgabe

(P) $$\text{Minimiere} \quad f(x) := \|x - z\|, \quad x \in M.$$

das duale Problem auf:

(D) Unter allen abgeschlossenen Hyperebenen in $X \times \mathbb{R}$, die nicht parallel zu $\mathbb{R}$ sind und Λ im nichtnegativen Halbraum enthalten, ist diejenige zu bestimmen, deren Schnitt mit der $\mathbb{R}$-Achse maximal ist.

Diese bisher rein verbale Definition des dualen Problems muss nun mathematisch genau gefasst werden.

Die nichtvertikalen abgeschlossenen Hyperebenen in $X \times \mathbb{R}$ können dargestellt werden durch

$$H(l, \alpha) := \{(x, q) \in X \times \mathbb{R} : l(x) + q = \alpha\}.$$

Hierbei ist α der Schnitt von $H(l, \alpha)$ mit der $\mathbb{R}$-Achse in $X \times \mathbb{R}$. In Abbildung 3.14 verdeutlichen wir uns die Situation. Ferner ist $\Lambda \subset H^+(l, \alpha)$ genau dann, wenn

$$\alpha \leq l(x - y) + \|x - z\| + r \quad \text{für alle } x \in X,\ y \in M \text{ und } r \geq 0.$$

Dies wiederum ist genau dann der Fall, wenn

(∗) $$\alpha \leq l(x - y) + \|x - z\| \quad \text{für alle } x \in X,\ y \in M.$$

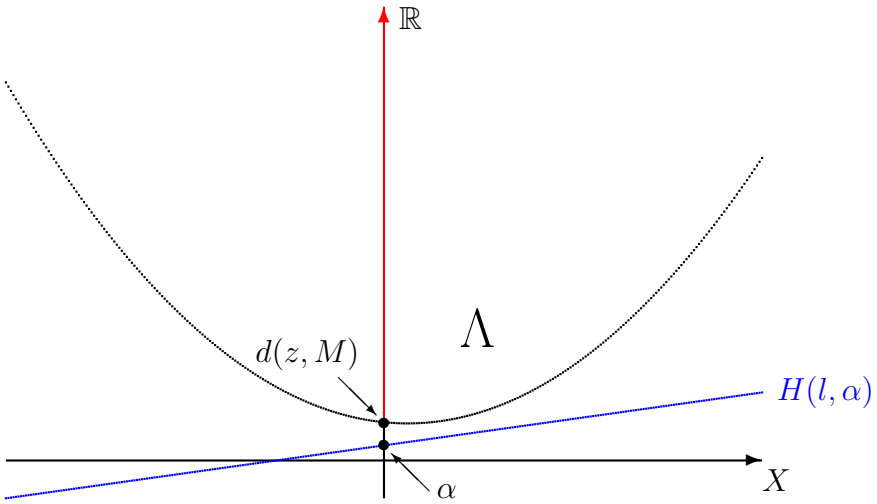

Abbildung 3.12: Primale und duale Approximationsaufgabe

Man weist leicht nach, dass $(*)$ genau dann gilt, wenn

$$\sup_{y\in M} l(y) < +\infty, \qquad \|l\| \le 1, \qquad \alpha \le l(z) - \sup_{y\in M} l(y).$$

Als duale Approximationsaufgabe erhält man daher

$$\text{(D)} \qquad \begin{cases} \text{Maximiere} \quad \phi(l) := l(z) - \sup\limits_{y\in M} l(y) \quad \text{auf} \\ N := \left\{ l \in X^* : \|l\| \le 1,\ \sup\limits_{y\in M} l(y) < +\infty \right\} \end{cases}$$

Bemerkung: Ist $M \subset X$ ein *linearer* Teilraum, so ist $\sup_{y\in M} l(y) < +\infty$ genau dann, wenn $l(y) = 0$ für alle $y \in M$ bzw. wenn l ein Element von

$$M^\perp := \{ l \in X^* : l(y) = 0 \ \text{ für alle } y \in M \}$$

ist. Das duale Problem ist in diesem Fall also durch

$$\text{(D)} \qquad \text{Maximiere} \quad \phi(l) := l(z) \quad \text{auf} \quad N := \{ l \in X^* : \|l\| \le 1,\ l \in M^\perp \}$$

gegeben. □

Aufgrund der Herleitung der dualen Aufgabe ist die folgende Aussage evident.

Satz 3.4.1 (Schwacher Dualitätssatz) *Gegeben sei die Approximationsaufgabe*

$$\text{(P)} \qquad \textit{Minimiere} \quad f(x) := \|x - z\|, \quad x \in M$$

und die hierzu duale Aufgabe

$$\text{(D)} \qquad \begin{cases} \textit{Maximiere} \quad \phi(l) := l(z) - \sup\limits_{y\in M} l(y) \quad \textit{auf} \\ N := \left\{ l \in X^* : \|l\| \le 1,\ \sup\limits_{y\in M} l(y) < +\infty \right\} \end{cases}$$

Dann gilt:

1. *Ist $x \in M$ und $l \in N$, so ist $\phi(l) \leq f(x)$. Insbesondere ist $\phi(l) \leq d(z, M)$ für jedes $l \in N$.*

2. *Sind $x \in M$, $l \in N$ und $\phi(l) = f(x)$, so ist x beste Approximierende an z in M und l Lösung von (D).*

Beweis: Seien $x \in M$ und $l \in N$. Dann ist

$$\phi(l) = l(z) - \sup_{y \in M} l(y) \leq l(z - x) \leq \|x - z\| = f(x).$$

Die übrigen Behauptungen folgen hieraus sofort. □

Man beachte, dass in Satz 3.4.1 keinerlei Konvexitätsvoraussetzungen gemacht wurden. Anschaulich ist aber klar, dass man ohne Konvexität von M (bzw. von Λ) mit einer *Dualitätslücke* bzw.

$$\sup_{l \in N} \phi(l) < \inf_{x \in M} f(x)$$

rechnen muss. Hierauf werden wir erst später eingehen. Jetzt wollen wir als Anwendung einige bekannte Aussagen über untere Schranken des Minimalabstands auf das obige allgemeine Prinzip zurückführen.

Zunächst betrachten wir eine *lineare Tschebyscheffsche Approximationsaufgabe bezüglich eines Haarschen Teilraumes*. Der zugrunde gelegte Raum ist im folgenden stets $(X, \|\cdot\|) = (C[\alpha, \beta], \|\cdot\|_\infty)$.

Definition 3.4.2 Ein n-dimensionaler linearer Teilraum $M \subset C[\alpha, \beta]$ heißt ein *n-dimensionaler Haarscher Teilraum* (oder auch *Haarsches System* oder *genügt der Haarschen Bedingung*), falls jedes $x \in M \setminus \{0\}$ höchstens $n - 1$ Nullstellen besitzt.

Das bekannteste und wichtigste Beispiel eines n-dimensionalen Haarschen Teilraumes ist natürlich $M = \Pi_{n-1}$, die Menge der Polynome vom Grad $\leq n - 1$.

Für den folgenden Satz werden wir zwei Beweise angeben. Zunächst einen direkten und später einen, der den Zusammenhang mit dem schwachen Dualitätssatz deutlich macht.

Satz 3.4.3 (de La Vallée Poussin) *Sei $M \subset C[\alpha, \beta]$ ein n-dimensionaler Haarscher Teilraum, $z \in C[\alpha, \beta]$. Ferner sei $x \in M$ und $\alpha \leq t_1 < \cdots < t_{n+1} \leq \beta$ mit*

$$[x(t_j) - z(t_j)][x(t_{j+1}) - z(t_{j+1})] < 0, \qquad j = 1, \ldots, n,$$

d. h. der Defekt bzw. genauer das Vorzeichen des Defekts $x - z$ alterniert in den Punkten t_j. Dann ist

$$\min_{j=1,\ldots,n+1} |x(t_j) - z(t_j)| \leq d(z, M) \leq \|x - z\|_\infty.$$

Beweis: Zu zeigen ist natürlich nur die linke Ungleichung. Angenommen, es existiert ein $\hat{x} \in M$ mit

$$\|\hat{x} - z\|_\infty < \min_{j=1,\ldots,n+1} |x(t_j) - z(t_j)|.$$

Insbesondere ist dann $|\hat{x}(t_j) - z(t_j)| < |x(t_j) - z(t_j)|$, $j = 1, \ldots, n + 1$. Wir werden uns gleich überlegen, dass $\hat{x} - x \in M \setminus \{0\}$ in den t_i alternierendes Vorzeichen hat,

also mindestens n Nullstellen besitzt, was einen Widerspruch dazu ergibt, dass M ein n-dimensionaler Haarscher Teilraum ist. Zur Begründung machen wir eine Fallunterscheidung. Ist $x(t_j) - z(t_j) > 0$, so ist $|\hat{x}(t_j) - z(t_j)| < x(t_j) - z(t_j)$, also

$$\hat{x}(t_j) - z(t_j) = \hat{x}(t_j) - z(t_j) - (x(t_j) - z(t_j)) \leq |\hat{x}(t_j) - z(t_j)| - (x(t_j) - z(t_j)) < 0.$$

Ist dagegen $x(t_j) - z(t_j) < 0$, so ist $|\hat{x}(t_j) - z(t_j)| < -(x(t_j) - z(t_j))$, also

$$\hat{x}(t_j) - x(t_j) = \hat{x}(t_j) - z(t_j) - (x(t_j) - z(t_j)) > \hat{x}(t_j) - z(t_j) + |\hat{x}(t_j) - z(t_j)| \geq 0.$$

Da $x - z$ nach Voraussetzung in den t_i dem Vorzeichen nach alterniert, trifft dies auch auf $\hat{x} - x$ zu und der Satz ist bewiesen. □

Dieser einfache und natürliche Beweis ist natürlich nicht zu schlagen. Trotzdem geben wir noch einen Beweis mit Hilfe des schwachen Dualitätssatzes an. Der schwache Dualitätssatz liefert das gewünschte Ergebnis, wenn wir ein

$$l \in N := \{l \in C[\alpha, \beta]^* : \|l\| \leq 1,\ l(y) = 0 \ \text{ für alle } y \in M\}$$

mit

$$\min_{j=1,\ldots,n+1} |x(t_j) - z(t_j)| \leq l(z)$$

finden können. Nun liegt dieser Konstruktion ein allgemeines Prinzip zugrunde, das wir später beim Alternantensatz und auch beim Remez-Algorithmus wieder antreffen werden und das wir daher als Satz formulieren. Anschließend kommen wir auf den Beweis des Satzes von de La Vallée Poussin zurück.

Satz 3.4.4 *Sei $M := \operatorname{span}\{x_1, \ldots, x_n\} \subset C[\alpha, \beta]$ ein n-dimensionaler Haarscher Teilraum von $C[\alpha, \beta]$ und $\alpha \leq t_1 < \cdots < t_{n+1} \leq \beta$. Dann gilt:*

1. *Das Gleichungssystem*

$$(*) \qquad \sum_{j=1}^{n+1} x_i(t_j) q_j = 0 \quad (i = 1, \ldots, n), \quad \sum_{j=1}^{n+1} |q_j| = 1$$

   besitzt eine bis auf einen Faktor ± 1 eindeutige Lösung $q = (q_1, \ldots, q_{n+1})^T$ und es ist $q_j q_{j+1} < 0$, $j = 1, \ldots, n$, die q_j alternieren also im Vorzeichen.

2. *Definiert man $l \in C[\alpha, \beta]^*$ durch*

$$l(y) := \sigma \cdot \sum_{j=1}^{n+1} q_j y(t_j),$$

   wobei $\sigma \in \{+1, -1\}$ und $q = (q_1, \ldots, q_{n+1})^T$ eine Lösung von $(*)$ ist, so ist $\|l\| \leq 1$ und $l(y) = 0$ für alle $y \in M$.

Beweis: Die $n \times n$-Matrix $(x_i(t_j))_{1\le i,j\le n}$ ist nichtsingulär, denn andernfalls könnten die Zeilen nichttrivial zu 0 kombiniert werden, d. h. es es existierten $c_1, \dots, c_n$, nicht alle gleich 0, mit

$$0 = \sum_{i=1}^{n} c_i x_i(t_j) = \left(\sum_{i=1}^{n} c_i x_i\right)(t_j), \quad j = 1, \dots, n.$$

Mit $x := \sum_{i=1}^{n} c_i x_i \in M \setminus \{0\}$ hätte man ein nichttriviales Element des n-dimensionalen Haarschen Teilraums M mit den n Nullstellen t_j, $j = 1, \dots, n$, ein Widerspruch. Das Gleichungssystem

$$\sum_{j=1}^{n+1} x_i(t_j) q_j = 0 \quad (i = 1, \dots, n)$$

ist äquivalent zu

$$\begin{pmatrix} q_1 \\ \vdots \\ q_n \end{pmatrix} = -q_{n+1}(x_i(t_j))^{-1} \begin{pmatrix} x_1(t_{n+1}) \\ \vdots \\ x_n(t_{n+1}) \end{pmatrix} = q_{n+1} \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix},$$

wobei

$$\begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix} := -(x_i(t_j))^{-1} \begin{pmatrix} x_1(t_{n+1}) \\ \vdots \\ x_n(t_{n+1}) \end{pmatrix}.$$

Durch die Zusatzbedingung

$$1 = \sum_{j=1}^{n+1} |q_j| = |q_{n+1}| \left(1 + \sum_{j=1}^{n} |p_j|\right)$$

ist

$$q_{n+1} = \pm \frac{1}{1 + \sum_{j=1}^{n} |p_j|}$$

bis auf einen Faktor ± 1 festgelegt, was dann auch für $q = (q_1, \dots, q_{n+1})^T$ gilt. Zu zeigen bleibt im ersten Teil des Beweises, dass $q_j q_{j+1} < 0$, $j = 1, \dots, n$. Hierzu beachten wir, dass $\sum_{j=1}^{n+1} y(t_j) q_j = 0$ für alle $y \in M = \operatorname{span}\{x_1, \dots, x_n\}$, da $\sum_{j=1}^{n+1} x_i(t_j) q_j = 0$, $i = 1, \dots, n$. Nun sei $k \in \{1, \dots, n\}$ fest. Ein $y_k \in M$ ist durch die n Interpolationsbedingungen

$$y_k(t_j) = \begin{cases} 0, & j \in \{1, \dots, n+1\} \setminus \{k, k+1\}, \\ 1, & j = k \end{cases}$$

eindeutig festgelegt, da M ein n-dimensionaler linearer Teilraum von $C[\alpha, \beta]$ ist. Dann ist

$$0 = \sum_{j=1}^{n+1} y_k(t_j) q_j = q_k + y_k(t_{k+1}) q_{k+1}.$$

Nun ist aber $y_k(t_{k+1}) > 0$, denn andernfalls hätte y_k wegen $y_k(t_k) = 1$ außer den $n-1$ Nullstellen in t_j, $j \in \{1, \ldots, n+1\} \setminus \{k, k+1\}$, noch eine Nullstelle in $(t_k, t_{k+1}]$. Also ist $q_k = -y_k(t_{k+1})q_{k+1}$ mit $y_k(t_{k+1}) > 0$, daher $q_k q_{k+1} < 0$. Damit ist der erste Teil des Satzes bewiesen. Der zweite Teil des Satzes ist nach dem schon bewiesenen fast trivial. Für $y \in C[\alpha, \beta]$ ist z. B.

$$|l(y)| = \left|\sum_{j=1}^{n+1} q_j y(t_j)\right| \leq \sum_{j=1}^{n+1} |q_j|\,|y(t_j)| \leq \underbrace{\left(\sum_{j=1}^{n+1} |q_j|\right)}_{=1} \|y\|$$

und daher $\|l\| \leq 1$. Damit ist das Satz bewiesen. □

Bemerkung: Im ersten Teil des letzten Satzes wurde gezeigt, dass das Gleichungssystem

$$\sum_{j=1}^{n+1} x_i(t_j)q_j = 0 \quad (i = 1, \ldots, n), \quad \sum_{j=1}^{n+1} |q_j| = 1$$

eine bis auf einen Faktor ± 1 eindeutige Lösung $q = (q_1, \ldots, q_{n+1})^T$ besitzt und $q_j q_{j+1} < 0$, $j = 1, \ldots, n$, gilt. Der Faktor sei so gewählt, dass $\operatorname{sign}(q_1) = \operatorname{sign}(x(t_1) - z(t_1))$. Dann ist $q_j = \lambda_j \operatorname{sign}(x(t_j) - z(t_j))$ mit $\lambda_j > 0$, $j = 1, \ldots, n+1$, und $\sum_{j=1}^{n+1} \lambda_j = 1$. Anschließend definiere man $l \in C[\alpha, \beta]^*$ durch

$$l(y) := -\sum_{j=1}^{n+1} q_j y(t_j).$$

Dann ist l dual zulässig (siehe zweiter Teil des obigen Satzes), wegen des schwachen Dualitätssatzes ist also

$$\begin{aligned}
d(z, M) &\geq l(z) \\
&= -\sum_{j=1}^{n+1} q_j z(t_j) \\
&= \sum_{j=1}^{n+1} q_j [x(t_j) - z(t_j)] \\
&= \sum_{j=1}^{n+1} \lambda_j \operatorname{sign}(x(t_j) - z(t_j))[x(t_j) - z(t_j)] \\
&= \sum_{j=1}^{n+1} \lambda_j |x(t_j) - z(t_j)| \\
&\geq \underbrace{\left(\sum_{j=1}^{n+1} \lambda_j\right)}_{=1} \min_{j=1,\ldots,n+1} [x(t_j) - z(t_j)| \\
&= \min_{j=1,\ldots,n+1} [x(t_j) - z(t_j)|.
\end{aligned}$$

Damit haben wir einen zweiten Beweis zum Satz von de La Vallée Poussin erhalten. Die zusätzliche Arbeit, die wir eben in einen zweiten Beweis des Satzes von de La Vallée Poussin hineingesteckt haben, ist nicht umsonst. Z. B. haben wir eine *konstruktive* Methode gefunden, um den Abstand $d(z, M)$ von $z \in C[\alpha, \beta]$ zu dem Haarschen Unterraum M nach unten abzuschätzen. Durch Variation der *Referenz* $\{t_1, \ldots, t_{n+1}\}$ kann man hoffen, diese Abschätzung zu verbessern. □

Jetzt wollen wir noch untere Schranken für den Minimalabstand bei der *rationalen T-Approximation* in $C[\alpha, \beta]$ angeben bzw. herleiten. Wie gerade eben sei der zugrunde liegende Raum wieder durch $(X, \|\cdot\|) = (C[\alpha, \beta], \|\cdot\|_\infty)$ gegeben. Mit Π bezeichnen wir die Menge aller (reellen) Polynome in einer Variablen. Für $p \in \Pi$ bezeichne ∂p den *Grad* von p (wir vereinbaren $\partial 0 = -\infty$). Sei Π_n die Menge der Polynome vom Grad $\leq n$. Weiter sei

$$R_{m,n} := \left\{ \frac{p}{q} : p \in \Pi_m,\ q \in \Pi_n,\ \frac{p}{q} \text{ irreduzibel, } q(t) > 0 \text{ für alle } t \in [\alpha, \beta] \right\}.$$

Ist $r = p/q \in R_{m,n}$, so heißt $d(r) := \min(m - \partial p, n - \partial q)$ der *Defekt* von r. Die rationale Funktion r heißt *ausgeartet*, falls $d(r) > 0$, andernfalls *nichtausgeartet* oder *normal*. Die irreduzible Darstellung der 0 sei $0/1$.

Beispiel: Sei $[\alpha, \beta] = [0, 1]$ und $r(t) := 1/(1 + t)$. Als Element von $R_{0,1}$ ist r normal. Als Element von $R_{2,3}$ ist $d(r) = 2$ und r ausgeartet. □

Wir werden später die Existenz und Eindeutigkeit bester T-Approximierender in $R_{m,n}$ beweisen und Charakterisierungen angeben. Jetzt interessiert uns lediglich ein dem Satz von de La Vallée Poussin entsprechender Satz für rationale T-Approximation.

Satz 3.4.5 *Sei $z \in C[\alpha, \beta]$, $r = p/q \in R_{m,n}$ und $N := m + n + 1 - d(r)$. Wenn es Punkte $\alpha \leq t_1 < \cdots < t_N < t_{N+1} \leq \beta$ mit*

$$[r(t_j) - z(t_j)][r(t_{j+1}) - z(t_{j+1})] < 0, \quad j = 1, \ldots, N,$$

gibt, so ist

$$\min_{j=1,\ldots,N+1} |r(t_j) - z(t_j)| \leq d(z, R_{m,n}) \leq \|r - z\|_\infty.$$

Beweis: Angenommen, es gibt ein $\hat{r} \in R_{m,n}$ mit

$$\|\hat{r} - z\|_\infty < \min_{j=1,\ldots,N+1} |r(t_j) - z(t_j)|.$$

Wörtlich wie beim Beweis von Satz 3.4.3, dem Satz von de La Vallée Poussin, folgt dann, dass $\hat{r}(\cdot) - r(\cdot)$ in den t_i, $i = 1, \ldots, N + 1$, alternierendes Vorzeichen hat, also mindestens $N = m + n + 1 - d(r)$ Nullstellen in $[\alpha, \beta]$ besitzt. Wir wollen zeigen, dass dies mindestens eine Nullstelle zu viel ist! Sei $\hat{r} = \hat{p}/\hat{q}$. dann ist

$$\hat{r} - r = \frac{\hat{p}}{\hat{q}} - \frac{p}{q} = \frac{\hat{p}q - p\hat{q}}{\hat{q}q}.$$

Nun ist aber

$$\partial(\hat{p}q - p\hat{q}) \leq \max(\partial(\hat{p}q), \partial(p\hat{q})) \leq \max(m + \partial q, n + \partial p) = m + n - d(r),$$

wobei die letzte Gleichung sich durch genaueres Hinsehen ergibt. Daher besitzt $\hat{r} - r$ höchstens $m + n - d(r)$ Nullstellen und wir haben einen Widerspruch erhalten. □

Beispiel: Sei $[\alpha, \beta] = [0, 1]$, $z(t) = e^{-t}$, $m = n = 1$ und

$$r(t) := \frac{1.002 - 0.4t}{1 + 0.64t}.$$

Dann ist $d(r) = 0$, wir benötigen also vier Punkte t_i, $i = 1, \ldots, 4$, in denen der Defekt $r - z$ im Vorzeichen alterniert. In Abbildung 3.13 ist der Defekt dargestellt. Wegen

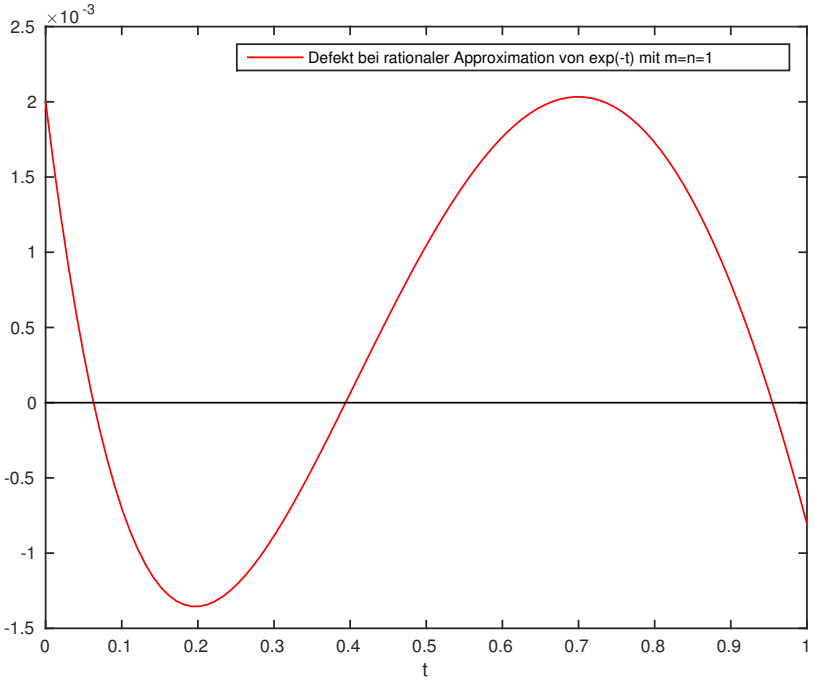

Abbildung 3.13: Der Defekt $r(t) - \exp(-t)$

i	t_i	$r(t_i) - z(t_i)$
1	0	0.0020
2	0.2	−0.0014
3	0.7	0.0020
4	1	−0.0008

ist $0.002 \leq d(z, R_{1,1})$. □

Wenn man versucht, Satz 3.4.5 auf den schwachen Dualitätssatz zurückzuführen, so hat man Schwierigkeiten. Diese liegen vor allem daran, dass das oben angegebene duale Problem für rationale Approximationsaufgaben nicht adäquat ist. Das "richtige" duale Problem ist bei L. Collatz, W. Krabs (1973.S. 125 ff.) angegeben. Hierauf wollen wir aber nicht näher eingehen.

Wenn man über untere Schranken für dem Minimalabstand spricht, muss man auch etwas über die von L. Collatz eingeführten H-Mengen bei der T-Approximation sagen. Bevor wir dies tun, wollen wir noch einmal auf die *geometrische Interpretation* des dualen Problems eingehen.
Das primale Problem ist

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \|x - z\|, \quad x \in M,$$

die hierzu duale Aufgabe ist

$$\text{(D)} \qquad \begin{cases} \text{Maximiere} \quad \phi(l) := l(z) - \sup\limits_{y \in M} l(y) \quad \text{auf} \\ N := \left\{ l \in X^* : \|l\| \leq 1,\ \sup\limits_{y \in M} l(y) < +\infty \right\}. \end{cases}$$

Sei $l \in N \setminus \{0\}$ und $\gamma := \sup_{y \in M} l(y)$. Die Hyperebene

$$H(l, \gamma) := \{x \in X : l(x) = \gamma\}$$

enthält M im nichtpositiven Halbraum $H^-(l, \gamma)$ und "stützt" M. Wir nehmen ferner an, dass z im M gegenüberliegenden Halbraum liegt, dass also $\phi(l) = l(z) - \gamma > 0$. Wir werden uns gleich überlegen, dass $\phi(l) = d(z, H(l, \gamma))$. In Abbildung 3.14 verdeutlichen wir uns die Situation. Das duale Problem kann man dann folgendermaßen formulieren:

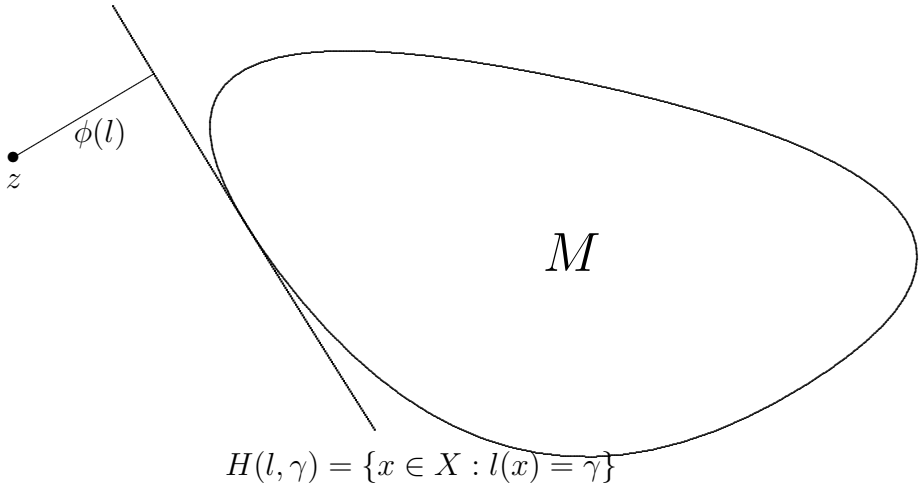

Abbildung 3.14: Veranschaulichung des dualen Problems

- Unter allen (abgeschlossenen) Hyperebenen, die M stützen und z im gegenüberliegenden Halbraum enthalten, ist diejenige zu finden, deren Abstand zu M maximal ist.

Nun überlegen wir uns, dass $\phi(l)$ gleich dem Abstand $d(z, H(l, \gamma))$ von z zur Hyperebene $H(l, \gamma)$ ist, falls $\|l\| = 1$. Denn ist $x \in H(l, \gamma)$ bzw. $l(x) = \gamma$, so ist

$$\|x - z\| \geq l(z) - l(x) = l(z) - \gamma = \phi(l),$$

also ist

$$d(z, H(l,\gamma)) = \inf_{x \in H(l,\gamma)} \|x - z\| \geq \phi(l).$$

Zum Nachweis der umgekehrten Ungleichung geben wir uns ein $\epsilon \in (0,1)$ vor. Wegen

$$1 = \|l\| = \sup_{x:\|x\|=1} |l(x)|$$

existiert ein $\hat{x} \in X$ mit $\|\hat{x}\| = 1$ und $|l(\hat{x})| \geq 1-\epsilon$. Da man notfalls $\hat{x}$ durch $-\hat{x}$ ersetzen kann, können wir $l(\hat{x}) \geq 1 - \epsilon$ annehmen. Nun definiere man

$$x := z - \frac{\phi(l)}{l(\hat{x})}\hat{x}.$$

Dann ist $l(x) = \gamma$ bzw. $x \in H(l,\gamma)$ und daher

$$d(z, H(l,\gamma)) \leq \|x - z\| = \frac{\phi(l)}{l(\hat{x})} \leq \frac{\phi(l)}{1-\epsilon}.$$

Mit $\epsilon \to 0+$ folgt $d(z, H(l,\gamma)) \leq \phi(l)$, insgesamt also die Behauptung.
Ohne die Voraussetzung, dass z im M gegenüberliegenden Halbraum liegt und für beliebiges $l \in N \setminus \{0\}$ ist offenbar

$$d(z, H(l,\gamma)) = \frac{|\phi(l)|}{\|l\|}.$$

Hieraus kann man ein *allgemeines Prinzip zur Gewinnung unterer Schranken* für den Minimalabstand herleiten:

Lemma 3.4.6 *Sei $(X, \|\cdot\|)$ ein linearer normierter Raum, $M \subset X$ und $z \in X$. Ferner sei I eine Indexmenge mit der folgenden Eigenschaft:*

Zu jedem $i \in I$ gibt es eine Hyperebene

$$H_i = H(l_i, \gamma_i) := \{x \in X : l_i(x) = \gamma_i\}$$

mit

$$z \notin H_i^- := \{x \in X : l_i(x) \leq \gamma_i\}$$

und $M \subset \bigcup_{i\in I} H_i^-$.

Dann ist $\inf_{i\in I} d(z, H_i) \leq d(z, M)$.

Beweis: Wegen $M \subset \bigcup_{i\in I} H_i^-$ existiert zu jedem $x \in M$ ein $i \in I$ mit $x \in H_i^-$, also $l_i(x) \leq \gamma_i$. Dann ist

$$\|l_i\| \, \|x - z\| \geq l_i(z) - l_i(x) \geq l_i(z) - \gamma_i$$

und daher

$$\|x - z\| \geq d(z, H_i) \geq \inf_{i\in I} d(z, H_i).$$

Hieraus folgt $\inf_{i\in I} d(z, H_i) \leq d(z, M)$, die Behauptung. □

Bemerkung: Der schwache Dualitätssatz 3.4.1 ergibt sich als Spezialfall, wenn die Indexmenge I in Lemm 3.4.6 *einpunktig* ist. □

Die folgende Definition findet man bei L. Collatz, W. Krabs (1973, S. 101).

Definition 3.4.7 Sei $(X, \|\cdot\|) := (C(B), \|\cdot\|_\infty)$ mit kompaktem $B \subset \mathbb{R}^N$, ferner sei $M \subset C(B)$. Eine Teilmenge $D \subset B$ heißt eine *H-Menge* (bezüglich M), wenn D die Vereinigung zweier nichtleerer Mengen D_1, D_2 ist derart, dass kein Paar $x, \hat{x} \in M$ existiert mit

$$x(t) - \hat{x}(t) \begin{cases} < 0, & t \in D_1, \\ > 0, & t \in D_2. \end{cases}$$

Durch Kombination mit Lemma 3.4.6, dem allgemeinen Prinzip zur Gewinnung unterer Schranken für den Minimalabstand, erhält man

Satz 3.4.8 *Sei $(X, \|\cdot\|) := (C(B), \|\cdot\|_\infty)$ mit kompaktem $B \subset \mathbb{R}^N$, ferner sei $M \subset C(B)$ und $z \in C(B)$. Die Menge $D = D_1 \cup D_2 \subset B$ sei eine H-Menge und $\hat{x} \in M$ ein Element mit*

$$\hat{x}(t) - z(t) \begin{cases} > 0, & t \in D_1, \\ < 0, & t \in D_2. \end{cases}$$

Dann ist

$$\inf_{t \in D} |\hat{x}(t) - z(t)| \leq d(z, M).$$

Beweis: Wir wollen Lemma 3.4.6 anwenden, setzen hierzu $I := D$ und definieren für $t \in D$ die stetigen linearen Funktionale $l_t \in C(B)^*$ durch

$$l_t(x) := \epsilon(t)x(t)$$

mit

$$\epsilon(t) := \begin{cases} -1, & t \in D_1, \\ t \in D_2 \end{cases}$$

und $\gamma_t \in \mathbb{R}$ durch $\gamma_t := \epsilon(t)\hat{x}(t)$. Für die zugehörigen (abgeschlossenen) Hyperebenen

$$H_t := \{x \in C(B) : l_t(x) = \gamma_t\} = \{x \in C(B) : \epsilon(t)x(t) = \epsilon(t)\hat{x}(t)\}$$

müssen die Voraussetzungen von Lemma 3.4.6 nachgewiesen werden. Für $t \in D$ ist zunächst $z \notin Z_t^-$, also $l_t(z) > \gamma_t$ bzw. $\epsilon(t)z(t) > \epsilon(t)\hat{x}(t)$, da $\epsilon(t)(\hat{x}(t) - z(t)) < 0$ für $t \in D$. Zum Nachweis der zweiten Voraussetzung $M \subset \bigcup : t \in DH_t^-$ geben wir uns ein $x \in M$ beliebig vor. Da D eine H-Menge ist, gibt es zum Paar $(x, \hat{x})$ ein $t \in D_1$ mit $x(t) - \hat{x}(t) \geq 0$ oder ein $t \in D_2$ mit $x(t) - \hat{x}(t) \leq 0$. In jedem Fall gibt es ein $t \in D$ mit $\epsilon(t)(x(t) - \hat{x}(t)) \leq 0$ bzw. $l_t(x) \leq \gamma_t$. Also ist $x \in H_t^-$ und $M \subset \bigcup_{t \in D} H_t^-$. Eine Anwendung von Lemma 3.4.6 liefert

$$\inf_{t \in D} d(z, H_t) \leq d(z, M) = \inf_{x \in M} \|x - z\|_\infty.$$

Offenbar ist $\|l_t\| = 1$ und daher, wie wir oben bemerkt haben,

$$d(z, H_t) = \frac{l_t(z) - \gamma_t}{\|l_t\|} = \epsilon(t)(z(t) - \hat{x}(t)) = |\hat{x}(t) - z(t)|.$$

Damit ist der Satz schließlich bewiesen. □

Weiter wollen wir auf das Konzept der H-Mengen nicht eingehen, sondern verweisen nur auf L. Collatz, W. Krabs (1973, S. 100 ff.).

Zum Schluss dieses Abschnitts wollen wir noch den *starken Dualitätssatz* für konvexe Optimierungsaufgaben formulieren und beweisen.

Satz 3.4.9 (Starker Dualtätssatz) *Sei $(X, \|\cdot\|)$ ein linearer normierter Raum, $M \subset X$ sei nichtleer, konvex und $z \in X$. Gegeben sei die konvexe Approximationsaufgabe*

$$(*) \qquad \textit{Minimiere} \quad f(x) := \|x - z\| \quad \textit{auf} \quad M$$

und das hierzu duale Problem

$$\text{(D)} \qquad \begin{cases} \textit{Maximiere} \quad \phi(l) := l(z) - \sup\limits_{y \in M} l(y) \quad \textit{auf} \\ N := \left\{ l \in X^* : \|l\| \leq 1,\ \sup\limits_{y \in M} l(y) < +\infty \right\}. \end{cases}$$

Dann ist (D) lösbar und es tritt keine Dualitätslücke auf, d. h. es existiert ein $l^* \in N$ mit

$$d(z, M) = \phi(l^*) = \max_{l \in N} \phi(l).$$

Besitzt (P) eine Lösung $x^* \in M$, so ist

$$-l(x^* - z) = \|x^* - z\|.$$

Beweis: Der einfachste und vielleicht auch natürlichste Beweis besteht darin, auf die ursprüngliche geometrische Interpretation des dualen Problems zurückzukehren und die Menge

$$\Lambda := \{(x - y, \|x - z\| + r) \in X \times \mathbb{R} : x \in X,\ y \in M,\ r \geq 0\}$$

zu betrachten. Mit M ist auch Λ konvex. Zunächst überlegen wir uns:

- Es ist

$$\{(x - y, \|x - z\| + r) \in X \times \mathbb{R} : x \in X,\ y \in M,\ r > 0\} \subset \operatorname{int}(\Lambda),$$

  insbesondere besitzt Λ ein nichtleeres Inneres.

Denn: Sind $\hat{x} \in X$, $\hat{y} \in M$ und $\hat{r} > 0$, so ist $(\hat{x} - \hat{y}, \|\hat{x} - z\| + \hat{r}) \in \operatorname{int}(\Lambda)$, da noch eine ganze Kugel um diesen Punkt zu Λ gehört. Hierzu zeigen wir, dass

$$(\hat{x} - \hat{y}, \|\hat{x} - z\| + \hat{r}) + B[0; \tfrac{1}{2}\hat{r}] \times [-\tfrac{1}{2}\hat{r}, \tfrac{1}{2}\hat{r}] \subset \Lambda.$$

Denn für $(h, q) \in B[0; \frac{1}{2}\hat{r}] \times [-\frac{1}{2}\hat{r}, \frac{1}{2}\hat{r}]$ ist

$$\begin{aligned} (\hat{x} - \hat{y}, \|\hat{x} - z\| + \hat{r}) + (h, q) &= (\hat{x} - \hat{y} + h, \|\hat{x} - z\| + \hat{r} + q) \\ &= (\hat{x} + h - \hat{y}, \|\hat{x} + h - z\| + \hat{q}) \end{aligned}$$

mit

$$\begin{aligned}\hat{q} &:= \|\hat{x}-z\| - \|\hat{x}+h-z\| + \hat{r} + q \\ &\geq \|\hat{x}-z\| - \|\hat{x}-z\| - \|h\| + \hat{r} + q \\ &\geq -\tfrac{1}{2}\hat{r} + \hat{r} - \tfrac{1}{2}\hat{r} \\ &= 0.\end{aligned}$$

Damit ist die angegebene Behauptung bewiesen. Weiter gilt offensichtlich:

- Es ist $(0, d(z,M)) \notin \operatorname{int}(\Lambda)$.

Aus dem Satz von Eidelheit (Satz 2.2.5) folgt, dass $\{(0, d(z,M))\}$ und Λ durch eine abgeschlossene Hyperebene in $X \times \mathbb{R}$ getrennt werden können. Es existiert also ein Paar $((l^*, \lambda^*), \gamma) \in (X^* \times \mathbb{R} \setminus \{(0,0)\}) \times \mathbb{R}$ mit

$$(*) \qquad \lambda^* d(z,M) \leq \gamma \leq l^*(x-y) + \lambda^*(\|x-z\| + r) \quad \text{für alle } x \in X,\ y \in M,\ r \geq 0,$$

und

$$(**) \qquad \lambda^* d(z,M) \leq \gamma < l^*(x-y) + \lambda^*(\|x-z\| + r) \quad \text{für alle } x \in X,\ y \in M,\ r > 0.$$

Wählt man $y \in M$ beliebig und setzt $x = y$, so sagt $(**)$ aus, dass

$$\lambda^* d(z,M) \leq \gamma < \lambda^*(\|x-z\| + r) \quad \text{für alle } r > 0,$$

woraus offenbar $\lambda^* > 0$ folgt. O. B. d. A. ist $\lambda^* = 1$, so dass $(*)$ mit $r = 0$ aussagt, dass

$$d(z,M) \leq l^*(x-y) + \|x-z\| \quad \text{für alle } x \in X,\ y \in M.$$

Für festes $x \in X$ ist also

$$\sup_{y \in M} l^*(y) \leq \|x-z\| + l^*(x) - d(z,M) < +\infty$$

und daher

$$d(z,M) \leq \|x-z\| + l^*(x) - \sup_{y \in M} l^*(y) \quad \text{für alle } x \in X.$$

Insbesondere ist

$$\inf_{x \in X}[\|x-z\| + l^*(x)] > -\infty.$$

Hieraus folgt $\|l^*\| \leq 1$ (Beweis?). Ferner ist

$$\inf_{x \in X}[\|x-z\| + l^*(x)] = l^*(z) + \inf_{x \in X}[\|x-z\| + l^*(x-z)] = l^*(z).$$

Insgesamt ist $l^* \in N$ dual zulässig und

$$d(z,M) \leq \phi(l^*) = l^*(z) - \sup_{y \in M} l^*(y) \leq d(z,M),$$

wobei die zweite Ungleichung aus dem schwachen Dualitätssatz folgt. Also ist $l^* \in N$ eine Lösung von (D) und $d(z,M) = \phi(l^*)$. Damit ist der erste Teil des Satzes bewiesen.

Im zweiten Teil des Satzes nehmen wir an, $x^* \in M$ sei eine Lösung von (P), also eine beste Approximierende an z in M. Dann ist

$$\|x^* - z\| = d(z, M) = \phi(l^*) = l^*(z) - \sup_{y \in M} \leq l^*(z - x^*) \leq \|z - x^*\|,$$

also ist $l^*(z - x^*) = \|z - x^*\|$ bzw. $-l^*(x^* - z) = \|x^* - z\|$. Damit ist der starke Dualitätssatz für konvexe Optimierungsaufgaben bewiesen. □

Bemerkung: Besitzt (P) eine Lösung x^*, so existiert wegen Satz 3.4.9 ein $l^* \in N$ mit $\|x^* - z\| = \phi(l^*)$. Hieraus folgt

1. $\|l^*\| \leq 1$,
2. $l^*(z - x^*) = \|z - x^*\|$,
3. $l^*(x^* - x) \geq 0$ für alle $x \in M$.

Man vergleiche diese Aussage mit der von Satz 3.2.10. □

Kapitel 4

Lineare Tschebyscheff-Approximation

4.1 Kolmogoroff-Kriterium, Alternantensatz, Eindeutigkeit und starke Eindeutigkeit

Wir betrachten Approximationsaufgaben, bei denen $(X, \|\cdot\|) = (C(B), \|\cdot\|_\infty)$ mit kompaktem $B \subset \mathbb{R}^N$, $\|x\|_\infty := \max_{t\in B} |x(t)|$ und $M \subset C(B)$ ein n-dimensionaler linearer Teilraum ist, etwa

$$M = \operatorname{span}\{x_1, \ldots, x_n\}.$$

Schließlich sei das zu Approximierende Element $z \in C(B) \setminus M$ vorgegeben. Zunächst geben wir für diesen Fall noch einmal das Kolmogoroff-Kriterium (siehe Satz 3.2.5) an.

Satz 4.1.1 (Kolmogoroff-Kriterium) *Ein $x^* \in M$ ist genau dann eine beste Approximierende an z in M, wenn*

$$\max_{t\in B(x^*-z)} \operatorname{sign}(x^*(t) - z(t))\, x(t) \geq 0 \quad \text{für alle } x \in M,$$

wobei $B(x^* - z) := \{t \in B : |x^*(t) - z(t)| = \|x^* - z\|_\infty\}$.

Dieses Kolmogoroff-Kriterium steht in einem engen Zusammenhang mit dem nächsten Charakterisierungssatz.

Satz 4.1.2 *Ein $x^* \in M$ ist genau dann eine beste Approximierende an z in M, wenn $0 \in \operatorname{co}(K)$ mit*

$$K := \{\operatorname{sign}(x^*(t) - z(t)) \begin{pmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{pmatrix} : t \in B(x^* - z)\},$$

wobei wieder $B(x^* - z) := \{t \in B : |x^*(t) - z(t)| = \|x^* - z\|_\infty\}$.

Beweis: Als stetiges Bild der kompakten Menge $B(x^* - z)$ ist offenbar $K \subset \mathbb{R}^n$ und wegen Satz 2.2.14 auch $\operatorname{co}(K)$ kompakt.

Im ersten Teil des Satzes nehmen wir an, $x^* \in M$ sei eine beste Approximierende an z in M. Angenommen, es sei $0 \notin \operatorname{co}(K)$. Wegen des strikten Trennungssatzes 2.2.7 lassen sich $\{0\}$ und $\operatorname{co}(K)$ strikt durch eine Hyperebene im $\mathbb{R}^n$ trennen, es existiert also ein $y = (y_j) \in \mathbb{R}^n$ mit

$$y^T k < y^T 0 = 0 \quad \text{für alle } k \in \operatorname{co}(K).$$

Insbesondere ist also

$$\operatorname{sign}(x^*(t) - z(t)) \sum_{j=1}^{n} y_j x_j(t) < 0 \quad \text{für alle } t \in B(x^* - z).$$

Da $x := \sum_{j=1}^{n} y_j x_j \in M$, ist dies ein Widerspruch zum Kolmogoroff-Kriterium.
Im zweiten Teil des Satzes nehmen wir an, es sei $0 \in \operatorname{co}(K)$. Wegen Satz 2.6 existieren $m \in \mathbb{N}$, $\lambda_i \geq 0$, $i = 1, \ldots, m$, mit $\sum_{i=1}^{m} \lambda_i = 1$ und $t_i \in B(x^* - z)$, $i = 1, \ldots, m$, mit

$$0 = \sum_{i=1}^{m} \lambda_i \operatorname{sign}(x^*(t_i) - z(t_i)) \begin{pmatrix} x_1(t_i) \\ \vdots \\ x_n(t_i) \end{pmatrix}.$$

Mit einem beliebigen $x = \sum_{j=1}^{n} y_j x_j$ ist dann

$$\begin{aligned} 0 &= \sum_{i=1}^{m} \lambda_i \operatorname{sign}(x^* * (t_i) - z(t_i)) x(t_i) \\ &\leq \max_{i=1,\ldots,m} \operatorname{sign}(x^*(t_i) - z(t_i)) x(t_i) \\ &\leq \max_{t \in B(x^* - z)} \operatorname{sign}(x^*(t) - z(t)) x(t). \end{aligned}$$

Aus dem Kolmogoroff-Kriterium (dem hinreichenden Teil) folgt, dass $x^* \in M$ eine beste Approximierende an z in M ist. □

Für Eindeutigkeitsaussagen und Alternanteneigenschaften bei linearer T-Approximation werden spezielle Eigenschaften des linearen Teilraums $M \subset C(B)$ benötigt. Die folgende Definition begegnete uns schon in 3.4.2.

Definition 4.1.3 Ein n-dimensionaler linearer Teilraum $M = \operatorname{span}\{x_1, \ldots, x_n\} \subset C(B)$ heißt ein *Haarscher Teilraum* der Dimension n, wenn eine der drei folgenden äquivalenten Bedingungen erfüllt ist.

1. Jedes $x \in M \setminus \{0\}$ besitzt höchstens $n - 1$ Nullstellen in B.
2. Sind n paarweise verschiedene Punkte $t_i \in B$ sowie $y_1, \ldots, y_n \in \mathbb{R}$ gegeben, so existiert ein $x \in M$, welches der Interpolationsbedingung $x(t_i) = y_i$, $i = 1, \ldots, n$, genügt.
3. Sind $t_i \in B$, $i = 1, \ldots, n$, paarweise verschieden, so ist die $n \times n$-Matrix

   $$(x_i(t_j))_{1 \leq i,j \leq n} \in \mathbb{R}^{n \times n}$$

   nichtsingulär.

Bevor wir Beispiele von Haarschen Teilräumen angeben, wollen wir den Zusammenhang mit der Eindeutigkeit bei linearer T-Approximation deutlich machen. Der folgende Satz wird auch *Haars Eindeutigkeitssatz* genannt, siehe z. B. E. W. Cheney (1966, S. 81).

Satz 4.1.4 *Sei $M \subset C(B)$ ein n-dimensionaler Teilraum*[1]*. Dann ist M genau dann eine T-Menge, wenn M ein Haarscher Teilraum ist.*

Beweis: Im ersten Teil des Beweises nehmen wir an, M sei eine T-Menge. Angenommen, M sei kein Haarscher Teilraum. Um einen Widerspruch zu erreichen wollen wir eine Funktion $z \in C(B)$ konstruieren, die mehr als eine beste Approximierende in M besitzt. Wegen unserer zum Widerspruch zu führenden Annahme existiert ein $x_0 \in M \setminus \{0\}$ mit n paarweise verschiedenen Nullstellen $t_1, \ldots, t_n \in B$. O. B. d. A. ist $\|x_0\|_\infty = 1$. Ist $M = \operatorname{span}\{x_1, \ldots, x_n\}$, so besitzt also das lineare Gleichungssystem

$$\sum_{j=1}^{n} \alpha_j x_j(t_i) = 0, \quad i = 1, \ldots, n,$$

eine nichttriviale Lösung $\alpha = (\alpha_j)$. Dann besitzt auch das homogene lineare Gleichungssystem

$$\sum_{i=1}^{n} \beta_i x_j(t_i) = 0, \quad j = 1, \ldots, n,$$

eine nichttriviale Lösung $\beta = (\beta_i)$. Wir können annehmen, dass $\sum_{i=1}^{n} |\beta_i| = 1$. Nun definieren wir $z_0 \colon \{t_1, \ldots, t_n\} \longrightarrow \mathbb{R}$ durch

$$z_0(t_i) := \begin{cases} \operatorname{sign}(\beta_i), & \beta_i \neq 0, \\ 0, & \beta_i = 0, \end{cases} \qquad i = 1, \ldots, n.$$

Man kann z_0 zu einer auf B stetigen Funktion $z_0 \in C(B)$ mit $\|z_0\|_\infty = 1$ fortsetzen. Denn als kompakte Menge ist B *normal*, d. h. eine auf einer abgeschlossenen Teilmenge $A \subset B$ definierte stetige Funktion $z_0 \colon A \longrightarrow \mathbb{R}$ kann so stetig auf B fortgesetzt werden, dass $\max_{t \in B} |z_0(t)| = \max_{t \in A} |z_0(t)|$ (Fortsetzungssatz von Tietze). Sei $z := z_0(|x_0| - 1)$. Wir werden zeigen, dass

$$\{\alpha x_0 : |\alpha| \leq 1\} \subset P_M(z),$$

also insbesondere mehr als eine beste Approximierende an z in M existiert und daher M keine T-Menge ist, ein Widerspruch zur Annahme. Hierzu wenden wir die hinreichende Optimalitätsbedingung in Satz 3.2.10 an, indem wir nachweisen, dass das durch $l(y) := \sum_{i=1}^{n} \beta_i y(t_i)$ definierte Element $l \in C(B)^*$ den Bedingungen

$$\|l\| = 1, \qquad l(\alpha x_0 - z) = \|\alpha x_0 - z\|, \qquad l(y) = 0 \quad \text{für alle } y \in M$$

genügt, wobei $\alpha \in \mathbb{R}$ mit $|\alpha| \leq 1$ beliebig ist. Für beliebiges $y \in C(B)$ ist

$$|l(y)| \leq \sum_{i=1}^{n} |\beta_i| \, |y(t_i)| \leq \Bigg(\underbrace{\sum_{i=1}^{n} |\beta_i|}_{=1} \Bigg) \|y\|_\infty$$

[1] Wegen Satz 3.1.2 ist M eine Existenzmenge.

und daher $\|l\| \leq 1$. Wegen $l(z_0) = 1 = \|z_0\|$ ist $\|l\| = 1$. Zum Nachweis der zweiten Eigenschaft beachten wir, dass

$$l(\alpha x_0 - z) = -l(z) = \sum_{i=1}^{n} \underbrace{\beta_i z_0(t_i)}_{|\beta_i|}(1 - \underbrace{|x_0(t_i)|}_{=0}) = 1 \leq \|\alpha x_0 - z\|.$$

Für $t \in B$ und $|\alpha| \leq 1$ ist andererseits

$$|\alpha x_0(t) - z(t)| \leq |\alpha|\,|x_0(t)| + |z(t)| \leq \underbrace{|\alpha|}_{\leq 1}\,|x_0(t)| + \underbrace{|z_0(t)|}_{\leq 1}(1 - \underbrace{|x_0(t)|}_{\leq 1}) \leq 1,$$

insgesant also

$$l(\alpha x_0 - z) = 1 = \|\alpha x_0 - z\|$$

für alle $\alpha \in \mathbb{R}$ mit $|\alpha| \leq 1$. Wegen

$$\sum_{i=1}^{n} \beta_i x_j(t_i) = 0, \quad j = 1, \ldots, n,$$

ist $l(y) = 0$ für alle $y \in M$. Damit haben wir den gewünschten Widerspruch erhalten.
Nun sei $M \subset C(B)$ ein Haarscher Teilraum. Es wird die Eindeutigkeit einer besten Approximierenden in M an jedes $z \in C(B)$ bewiesen. O. B. d. A. ist $z \in C(B) \setminus M$. Wir zeigen zunächst:

- Ist $x^* \in P_M(z)$, so enthält

  $$B(x^* - z) := \{t \in B : |x^*(t) - z(t)| = \|x^* - z\|_\infty\}$$

  mindestens $n + 1$ Punkte.

Denn andernfalls existiert ein $x \in M$ mit $x(t_i) = z(t_i) - x^*(t_i)$ für alle $t_i \in B(x^* - z)$. Dann ist aber

$$\max_{t \in B(x^* - z)} \operatorname{sign}(x^*(t) - z(t))x(t) = -\|x^* - z\|_\infty < 0,$$

ein Widerspruch zum Kolmogoroff-Kriterium. Nun kommen wir zum Eindeutigkeitsbeweis und nehmen an, es seien $x_1^*, x_2^* \in P_M(z)$. Wegen der Konvexität von $P_M(z)$ ist auch $x^* := \frac{1}{2}(x_1^* + x_2^*) \in P_M(z)$. Wegen der gerade eben bewiesenen Aussage gibt es paarweise verschiedene $t_1, \ldots, t_{n+1} \in B(x^* - z)$. Sei $\delta := d(z, M)$ der Minimalabstand von z zu M bzw. der Abstand von x^*, x_1^* bzw. x_2^* zu z. Für $i = 1, \ldots, n + 1$ ist dann

$$|x^*(t_i) - z(t_i)| = |\tfrac{1}{2}[(x_1^*(t_i) - z(t_i)) + (x_2^*(t_i) - z(t_i))]| = \delta$$

sowie

$$|x_1^*(t_i) - z(t_i)| \leq \delta, \qquad |x_2^*(t_i) - z(t_i)| \leq \delta.$$

Da $(\mathbb{R}, |\cdot|)$ strikt konvex ist, ist

$$x_1^*(t_i) - z(t_i) = x_2^*(t_i) - z(t_i), \qquad i = 1, \ldots, n + 1.$$

Daher besitzt $x_1^* - x_2^* \in M$ mindestens $n+1$ Nullstellen. Also ist $x_1^* = x_2^*$ und das ist die Eindeutigkeit einer besten Approximierenden. □

Beispiele: 1. Sei $B = I$ ein reelles Intervall und Π_n die Menge der Polynome vom Grad $\leq n$. Dann ist Π_n ein $(n+1)$-dimensionaler Haarscher Teilraum von $C(I)$. Denn jedes $x \in \Pi_n \setminus \{0\}$ besitzt höchstens n Nullstellen in I.

2. Sei

$$\mathcal{T}_n := \left\{ x(\phi) = \sum_{k=0}^{n} a_k \cos k\phi + \sum_{k=1}^{n} b_k \sin k\phi : a_k, b_k \in \mathbb{R} \right\}$$

der Raum der trigonometrischen Polynome vom Grad $\leq n$. Sei $B := [0, \beta]$ mit $\beta < 2\pi$. Dann ist $\mathcal{T}_n$ ein $(2n+1)$-dimensionaler Haarscher Teilraum von $C(B)$. Denn: Mit $i = \sqrt{-1}$ ist

$$\begin{aligned} \sum_{k=0}^{n} (a_k \cos k\phi + b_k \sin k\phi) &= \sum_{k=0}^{n} \left[\frac{a_k}{2}(e^{ik\phi} + e^{-ik\phi}) + \frac{b_k}{2i}(e^{ik\phi} - e^{-ik\phi}) \right] \\ &= \sum_{k=0}^{n} \left[\frac{1}{2}(a_k - ib_k)e^{ik\phi} + \frac{1}{2}(a_k + ib_k)e^{-ik\phi} \right] \\ &= e^{-in\phi} \sum_{k=0}^{2n} c_k z^k \end{aligned}$$

mit $z = e^{i\phi}$. Hieraus liest man ab, dass ein nicht identisch verschwindendes Element von $\mathcal{T}_n$ höchstens $2n$ Nullstellen in $[0, 2\pi)$ besitzt. Etwas natürlicher ist es, $\mathcal{T}_n$ als Teilraum von $C_{2\pi}$, dem linearen Raum der 2π-periodischen stetigen Funktionen in $C[0, 2\pi]$ zu betrachten, wobei 0 und 2π zu identifizieren sind.

3. Seien $q_1, \ldots, q_n$ nichtnegative ganze Zahlen und $\lambda_1, \ldots, \lambda_n$ paarweise verschiedene reelle Zahlen. Dann ist

$$E_n(q, \lambda) := \left\{ \sum_{k=1}^{n} p_k(t) e^{\lambda_k t} : p_k \in \Pi_{q_k},\ k = 1, \ldots, n \right\}$$

auf jedem (reellen) Intervall I ein $N := (\sum_{k=1}^{n} q_k) + n$-dimensionaler Haarscher Teilraum von $C(I)$.

Denn: Man hat zu zeigen, dass jedes nichttriviale Element aus $E_n(q, \lambda)$ höchstens $N-1$ reelle Nullstellen besitzt. Dies geschieht durch vollständige Induktion nach n. Für den *Induktionsanfang* ist $n = 1$. Offenbar besitzt $p_1(t)e^{\lambda_1 t}$ mit $p_1 \in \Pi_{q_1}$ höchstens $q_1 = q_1 + 1 - 1$ reelle Nullstellen. Die *Induktionsannahme* besteht darin, dass $E_{n-1}(q, \lambda)$ für beliebige nichtnegative ganze Zahlen $q_1, \ldots, q_{n-1}$ und paarweise verschiedenene $\lambda_1, \ldots, \lambda_{n-1}$ ein Haarscher Teilraum der Dimension $\sum_{k=1}^{n-1} q_k + n - 1$ ist. Im *Induktionsschluss* sei $x(t) := \sum_{k=1}^{n} p_k(t) e^{\lambda_k t}$ ein nichttriviales Element aus $E_n(q, \lambda)$. Sei

$$y(t) := x(t)e^{-\lambda_n t} = p_n(t) + \sum_{k=1}^{n-1} p_k(t) e^{(\lambda_k - \lambda_n)t}.$$

Differenziert man y genau $(q_n + 1)$-mal, so verschwindet der erste Summand. Daher besitzt $y^{(q_n+1)}$ nach Induktionsannahme höchstens $\sum_{k=1}^{n-1} q_k + n - 2 + 1$ Nullstellen.

Jetzt wenden wir sukzessive den Satz von Rolle an. Nach diesem liegt zwischen je zwei Nullstellen von $y^{(q_n)}$ mindestens eine Nullstelle von $y^{(q_n+1)}$. Daher besitzt $y^{(q_n)}$ höchstens $\sum_{k=1}^{n-1} q_k + n - 2 + 1$ Nullstellen, $y^{(q_n-1)}$ höchstens $\sum_{k=1}^{n-1} q_k + n - 2 + 2$ Nullstellen, bis man schließlich erhält, dass $y = y^{(q_n-q_n)}$ und damit auch x höchstens

$$\sum_{k=1}^{n-1} q_k + n - 2 + q_n + 1 = \sum_{k=1}^{n} q_k + n - 1$$

Nullstellen besitzt. □

Bemerkung: Sei $B := [-1,1] \times [-1,1] \subset \mathbb{R}^2$. Die Polynome $\sum_{k=0}^{m} \sum_{l=0}^{n} a_{kl} s^l t^l$ in zwei Variablen bilden für $m, n \geq 1$ keinen Haarschen Teilraum von $C(B)$, da z. B. das Polynom $s-t$ längs einer Geraden, also sogar in unendlich vielen Punkten verschwindet. Der Satz von Mairhuber-Curtis sagt aus, dass dies "kein Wunder" ist. Genauer gilt (siehe D. Braess (1986, S. 12)):

- Ist $B \subset \mathbb{R}^N$ kompakt und enthält $C(B)$ einen n-dimensionalen Haarschen Teilraum mit $n \geq 2$, so ist B homöomorph zu einer abgeschlossenen Teilmenge der Kreisperipherie.

Daher beschränken wir uns im folgenden auf kompakte Intervalle $B = [\alpha, \beta]$. □

Unser nächstes Ziel ist es, den folgendes Satz zu beweisen.

Satz 4.1.5 (Alternantensatz) *Sei $M = \operatorname{span}\{x_1, \ldots, x_n\}$ ein n-dimensionaler Haarscher Teilraum von $C[\alpha, \beta]$. Dann ist $x^* \in M$ genau dann eine beste T-Approximierende an $z \in C[\alpha, \beta]$, wenn es $n+1$ Punkte $t_j \in [\alpha, \beta]$, $j = 1, \ldots, n$, gibt mit*

(a) $\alpha \leq t_1 < t_2 < \cdots < t_{n+1} \leq \beta$,

(b) $|x^*(t_j) - z(t_j)| = \|x^* - z\|_\infty$, $j = 1, \ldots, n+1$,

(c) $x^*(t_j) - z(t_j) = (-1)^{j+1}(x^*(t_1) - z(t_1))$, $j = 1, \ldots, n+1$.

Beweis: Sei $x^* \in M$ eine beste T-Approximierende an z in M. O. B. d. A. ist $z \notin M$. Wegen Satz 4.1.2 ist

$$0 \in \operatorname{co}\left(\{\operatorname{sign}(x^*(t) - z(t)) \begin{pmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{pmatrix} : t \in B(x^* - z)\}\right)$$

mit

$$B(x^* - z) := \{t \in [\alpha, \beta] : |x^*(t) - z(t)| = \|x^* - z\|_\infty\}.$$

Wegen des Satzes von Carathéodory (Satz 2.2.13) existieren eine natürliche Zahl $m \leq n+1$ und $\lambda_j > 0$, $j = 1, \ldots, m$, mit $\sum_{j=1}^{m} \lambda_j = 1$ sowie $t_j \in B(x^* - z)$, $j = 1, \ldots, m$, mit

$$0 = \sum_{j=1}^{m} \lambda_j \operatorname{sign}(x^*(t_j) - z(t_j)) x_i(t_j), \qquad i = 1, \ldots, n.$$

Die t_j können als der Größe nach geordnet angenommen werden, so dass also $\alpha \le t_1 < t_2 < \cdots < t_m \le \beta$. Nach Satz 3.4.4 besitzt das Gleichungssystem

$$\sum_{j=1}^{n+1} x_i(t_j) q_j = 0, \quad i = 1, \ldots, n, \qquad \sum_{j=1}^{n+1} |q_j| = 1$$

eine bis auf einen Faktor ± 1 eindeutige Lösung $q = (q_1, \ldots, q_{n+1})^T$ und es gilt $q_j q_{j+1} < 0$, $j = 1, \ldots, n$, die q_j alternieren also im Vorzeichen. Hieraus folgt $m = n + 1$ und

$$\operatorname{sign}(x^*(t_j) - z(t_j)) = -\operatorname{sign}(x^*(t_{j+1}) - z(t_{j+1})), \quad j = 1, \ldots, n.$$

Insgesamt genügen die $n + 1$ Punkte $t_j \in [\alpha, \beta]$ den Bedingungen (a), (b) und (c).

Gibt es $n + 1$ Punkte $t_j \in [\alpha, \beta]$, die den Bedingungen (a), (b) und (c) genügen, so folgt aus Satz 3.4.3, dem Satz von de La Vallée Poussin, dass x^* eine beste T-Approximierende an z in M ist. Insgesamt ist der Alternantensatz bewiesen. □

Erste konkrete Anwendungen des Alternantensatzes werden wir im nächsten Abschnitt kennenlernen. Es sei hier schon darauf hingewiesen, dass der Alternantensatz Grundlage des wichtigsten numerischen Verfahrens der linearen T-Approximation ist, nämlich des Remez-Verfahrens. Hier wollen wir nur ein Beispiel angeben.

Beispiel: Sei $[\alpha, \beta] := [\frac{1}{2}, 1]$ und $M := \operatorname{span}\{1/\sqrt{t}, \sqrt{t}\}$, $z(t) := 1$. Die entsprechende T-Approximationsaufgabe besteht dann darin, die Quadratwurzel $\sqrt{t}$ in $[\frac{1}{2}, 1]$ so durch eine lineare Funktion zu approximieren, dass der maximale relative Fehler minimal wird. M ist ein 2-dimensionaler Haarscher Teilraum von $C[\frac{1}{2}, 1]$, man benötigt also drei Alternantenpunkte $\frac{1}{2} \le t_1 < t_2 < t_3 \le 1$. Für die beste Approximierende machen wir den Ansatz

$$x^*(t) = a\frac{1}{\sqrt{t}} + b\sqrt{t}.$$

Nimmt man $t_1 = \frac{1}{2}$, $t_3 = 1$ an, so ist notwendig $x^*(t_1) - z(t_1) = x^*(t_3) - z(t_3)$ und daher

$$a(\sqrt{2} - 1) - b\left(1 - \frac{1}{\sqrt{2}}\right) = 0$$

bzw.

$$\frac{a}{b} = \frac{1}{\sqrt{2}}.$$

Aus

$$\frac{d}{dt}(x^*(t) - z(t))_{t=t_2} = 0,$$

der notwendigen Bedingung für ein Extremum in t_2 im Innern von $[\frac{1}{2}, 1]$, erhält man

$$-\frac{1}{2} a t_2^{-3/2} + \frac{1}{2} b t_2^{-1/2} = 0$$

und hieraus

$$t_2 = \frac{a}{b} = \frac{1}{\sqrt{2}}.$$

Aus

$$a + b - 1 = x^*(t_3) - z(t_3) = -(x^*(t_2) - z(t_2)) = -at_2^{-1/2} - bt_2^{1/2} + 1 = -2\sqrt{ab} + 1$$

erhält man $\sqrt{a} + \sqrt{b} = \sqrt{2}$, zusammen mit $b = \sqrt{2}a$ folgt

$$a = \frac{2}{(1 + 2^{1/4})^2}, \qquad b = \frac{2^{3/2}}{(1 + 2^{1/4})^2}.$$

Durch $x^*(t) = a + bt$ ist dann diejenige lineare Funktion gegeben, die den maximalen *relativen* Fehler bei der Approximation von $\sqrt{t}$ auf $[\frac{1}{2}, 1]$ minimiert. In Abbildung 4.1 geben wir den relativen Fehler $d_1(t) := (a + bt - \sqrt{t})/\sqrt{t}$ und den absoluten Fehler

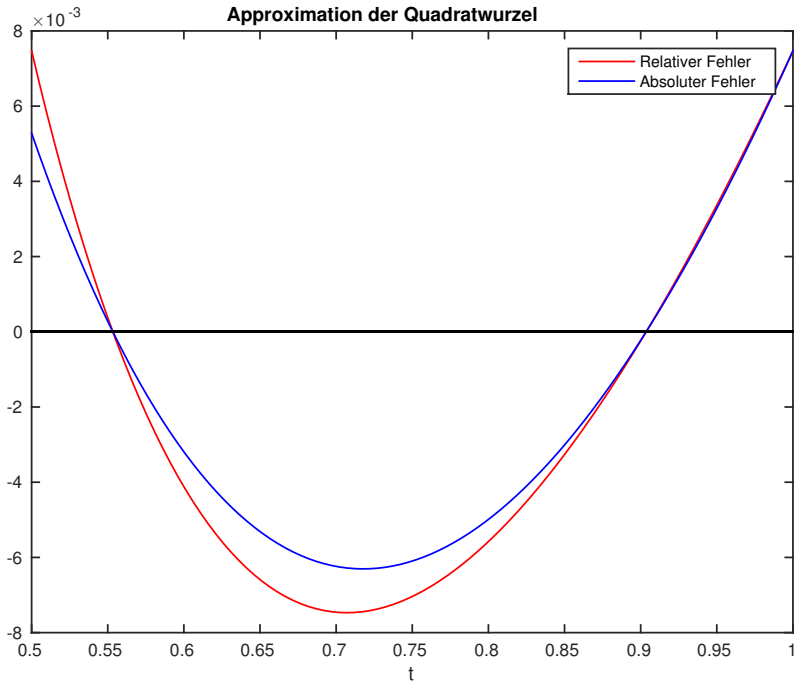

Abbildung 4.1: Relativer und absoluter Fehler

$d_2(t) := a+bt-\sqrt{t}$ an. Will man den maximalen *absoluten* Fehler bei der Approximation der Quadratwurzel durch eine lineare Funktion auf $[\frac{1}{2}, 1]$ minimieren, so erhält man als beste Approximierende $x^*(t) = a + bt$ mit

$$a = \frac{3}{8}\left(\frac{3}{2}\sqrt{2} - 1\right), \qquad b = 2 - \sqrt{2}.$$

Siehe auch Abschnitt 41 in *Merkwürdige Mathematik*. □

Zum Schluss dieses Abschnitts gehen wir noch auf die *starke Eindeutigkeit* und die *Stetigkeit der metrischen Projektion* ein.

Wir definieren für ein allgemeines Approximationsproblem in einem linearen normierten Raum:

Definition 4.1.6 Sei $(X, \|\cdot\|)$ ein linearer normierter Raum und $M \subset X$ eine nichtleere Menge. Ein $x^* \in M$ heißt *stark* (oder *streng*) *eindeutige* beste Approximierende an $z \in X$, falls eine positive Konstante $c > 0$ existiert mit

$$\|x - z\| \geq \|x^* - z\| + c\,\|x - x^*\| \quad \text{für alle } x \in M.$$

Offenbar ist eine stark eindeutige beste Approximierende auch eine eindeutige beste Approximierende.

Beispiel: Sei $(X, (\cdot,\cdot))$ ein Hilbertraum und $M \subset X$ ein abgeschlossener linearer Teilraum, ferner $z \in X$. Dann existiert genau eine beste Approximierende x^* an z in M und diese ist charakterisiert durch $x^* - z \perp M$. Für beliebiges $x \in M$ ist dann wegen des Satzes von Pythagoras

$$\begin{aligned} \|x - z\|^2 &= \|(x - x^*) + (x^* - z)\|^2 \\ &= \|x^* - z\|^2 + \|x - x^*\|^2. \end{aligned}$$

Hieraus erkennt man sehr leicht, dass *keine* starke Eindeutigkeit vorliegt. Dies würde nämlich bedeuten, dass der Abstand von x zu z linear mit dem Abstand von x zu x^* wächst, während er hier nur von zweiter Ordnung wächst. □

Den folgenden Satz findet man z. B.auch bei E.W.Cheney (1966, S. 80).

Satz 4.1.7 *Sei* $(X, \|\cdot\|) = (C(B), \|\cdot\|_\infty)$, $B \subset \mathbb{R}^N$ *kompakt und*

$$M = \operatorname{span}\{x_1, \ldots, x_n\} \subset C(B)$$

ein n*-dimensionaler Haarscher Teilraum. Dann ist die beste Approximation* $x^* \in M$ *an ein* $z \in C(B)$ *stark eindeutig.*

Beweis: O. B. d. A. ist $z \notin M$ und daher $\delta := d(z, M) > 0$. Wegen Satz 4.1.2 ist

$$0 \in \operatorname{co}\left(\{\operatorname{sign}(x^*(t) - z(t)) \begin{pmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{pmatrix} : t \in B(x^* - z)\}\right)$$

mit

$$B(x^* - z) := \{t \in B : |x^*(t) - z(t)| = \|x^* - z\|_\infty\}.$$

Wegen des Satzes von Carathéodory (Satz 2.2.13) existieren $m \leq n + 1$, $\lambda_j > 0$, $j = 1, \ldots, m$, mit $\sum_{j=1}^m \lambda_j = 1$ und $t_j \in B(x^* - z)$, $j = 1, \ldots, m$, mit

$$\sum_{j=1}^m \lambda_j \operatorname{sign}(x^*(t_j) - z(t_j)) x_i(t_j) = 0, \qquad i = 1, \ldots, n.$$

Da M ein Haarscher Teilraum ist, ist $m = n + 1$. Denn wäre $m < n + 1$, so wähle man paarweise verschiedene $t_{m+1}, \ldots, t_n \in B$, die auch noch von $t_1, \ldots, t_m$ verschieden sind. Dann hat das lineare Gleichungssystem $\sum_{j=1}^n y_j x_i(t_j) = 0$, $i = 1, \ldots, n$, eine

nichttriviale Lösung. Also ist $(x_i(t_j))_{1\le i,j\le n}$ singulär und daher existiert ein $x \in M\setminus\{0\}$ mit den n Nullstellen $t_1,\dots,t_n \in B$, ein Widerspruch. Daher ist

$$\sum_{j=1}^{n+1} \lambda_j \operatorname{sign}(x^*(t_j)-z(t_j))x_i(t_j) = 0, \qquad i=1,\dots,n,$$

und folglich

$$\sum_{j=1}^{n+1} \lambda_j \operatorname{sign}(x^*(t_j)-z(t_j))x(t_j) = 0, \qquad \text{für alle } x \in M.$$

Für $x \in M \setminus \{0\}$ ist dann

$$\psi(x) := \max_{j=1,\dots,n+1} \operatorname{sign}(x^*(t_j)-z(t_j))x(t_j) > 0.$$

Denn wäre $\operatorname{sign}(x^*(t_j)-z(t_j))x(t_j) \le 0$, $j=1,\dots,n+1$, für ein $x \in M$, so wäre $x(t_j)=0$, $j=1,\dots,n$, und damit $x=0$. Definiere

$$c := \min_{y\in M, \|y\|_\infty=1} \psi(y).$$

Dann ist $c>0$, da die stetige Funktion ψ auf der kompakten Menge $M\cap\{y : \|y\|_\infty = 1\}$ positiv ist. Nun wollen wir zeigen, dass die eben definierte Konstante c als Konstante bei der starken Eindeutigkeit verwandt werden kann. Sei $x \in M$, o. B. d. A. $x \neq x^*$. Sei $y := (x-x^*)/\|x-x^*\|_\infty$. Wegen

$$c \le \psi(y) = \max_{j=1,\dots,n+1} \operatorname{sign}(x^*(t_j)-z(t_j))y(t_j)$$

gibt es ein $j \in \{1,\dots,n+1\}$ mit

$$c \le \operatorname{sign}(x^*(t_j)-z(t_j))y(t_j).$$

Dann ist aber

$$\begin{aligned}
\|x-z\|_\infty &\ge \operatorname{sign}(x^*(t_j)-z(t_j))(x(t_j)-z(t_j)) \\
&= \operatorname{sign}(x^*(t_j)-z(t_j))[(x(t_j)-x^*(t_j))+(x^*(t_j)-z(t_j))] \\
&= \|x^*-z\|_\infty + \operatorname{sign}(x^*(t_j)-z(t_j))(x(t_j)-x^*(t_j)) \\
&= \|x^*-z\|_\infty + \operatorname{sign}(x^*(t_j)-z(t_j))y(t_j)\,\|x-x^*\|_\infty \\
&\ge \|x^*-z\|_\infty + c\,\|x-x^*\|_\infty
\end{aligned}$$

und dies ist die Behauptung. □

Zum Schluss dieses Abschnitts gehen wir noch kurz auf die Stetigkeit der metrischen Projektion auf Haarsche Teilräume ein. Wir wissen bisher: Haarsche Teilräume von $(C(B),\|\cdot\|_\infty)$ sind T-Mengen. Als endlichdimensionale Räume sind sie beschränkt kompakt und erst recht approximativ kompakt. Die metrische Projektion auf eine approximativ kompakte T-Menge ist stetig (Satz 3.1.6). Mit Hilfe der starken Eindeutigkeit können wir mehr zeigen, nämlich die Lipschitz-Stetigkeit der metrischen Projektion in jedem Punkt. Genauer gilt:

Satz 4.1.8 *Sei $(X, \|\cdot\|) = (C(B), \|\cdot\|_\infty)$, $B \subset \mathbb{R}^N$ kompakt und $M \subset C(B)$ ein n-dimensionaler Haarscher Teilraum. Zu jedem $z_0 \in C(B)$ existiert dann eine Konstante $c_0 > 0$ derart, dass*

$$\|P_M(z) - P_M(z_0)\|_\infty \leq c_0 \|z - z_0\|_\infty \quad \text{für alle } z \in C(B),$$

wobei $P_M \colon C(B) \longrightarrow M$ die metrische Projektion auf M ist.

Beweis: Nach Satz 4.1.7 ist $P_M(z_0)$ stark eindeutige beste Approximierende an z_0, es existiert also eine Konstante $c > 0$ mit

$$\|x - z_0\|_\infty \geq \|P_M(z_0) - z_0\|_\infty + c\,\|x - P_M(z_0)\|_\infty \quad \text{für alle } x \in M.$$

Setzt man hier $x = P_M(z)$, so erhält man

$$\begin{aligned}
\|P_M(z) - P_M(z_0)\|_\infty &\leq \frac{1}{c}[\|P_M(z) - z_0\|_\infty - \|P_M(z_0) - z_0\|_\infty] \\
&\leq \frac{1}{c}[\|P_M(z) - z\|_\infty + \|z - z_0\|_\infty - \|P_M(z_0) - z_0\|_\infty] \\
&= \frac{1}{c}[\underbrace{\|P_M(z) - z\|_\infty - \|P_M(z_0) - z\|_\infty}_{\leq 0} + \|z - z_0\|_\infty \\
&\qquad + \underbrace{\|P_M(z_0) - z\|_\infty - \|P_M(z_0) - z_0\|_\infty}_{\leq \|z - z_0\|_\infty}] \\
&\leq \frac{2}{c}\,\|z - z_0\|_\infty,
\end{aligned}$$

die Behauptung ist also mit $c_0 := 2/c$ bewiesen. □

Ganz zum Schluss dieses Abschnitts über (allgemeine) lineare T-Approximation formulieren und beweisen wir noch das *Invarianzprinzip* von Meinardus.

Satz 4.1.9 *Sei $(X, \|\cdot\|) = (C(B), \|\cdot\|_\infty)$, $B \subset \mathbb{R}^N$ kompakt und $M \subset C(B)$ ein endlichdimensionaler linearer Teilraum, ferner sei $z \in C(B)$. Sei $T \colon B \longrightarrow B$ stetig und $A \colon C(B) \longrightarrow C(B)$ linear und stetig mit $\|A\| \leq 1$. Ferner gelte*

1. *$Az \circ T = z$, also $Az(Tt) = z(t)$ für alle $t \in B$ (Symmetrieeigenschaft der zu approximierenden Funktion z),*
2. *$Ax \circ T \in M$ für alle $x \in M$.*

Dann existiert ein $x^* \in P_M(z)$ mit $Ax^* \circ T = x^*$.

Bevor wir in den Beweis einsteigen, wollen wir andeuten, in welchen Fällen der Satz anwendbar ist. Sei z. B. $B := [-1, 1]$. Die Abbildung $T \colon B \longrightarrow B$ sei definiert durch $T(t) := -t$. Die Abbildung $A \colon C(B) \longrightarrow C(B)$ sei definiert durch $Ax(t) := -x(t)$. Dann ist offenbar A linear, stetig und $\|A\| = 1$. Die Bedingung 1. besagt dann, dass $-z(-t) = z(t)$, dass also z ungerade ist. Die Bedingung 2. besagt: Ist $x \in M$ und y definiert durch $y(t) := -x(-t)$, so ist auch $y \in M$. Die Aussage des Invarianzprinzips

besteht darin, dass auch eine beste *ungerade* Approximierende existiert. Ist etwa $M = \operatorname{span}\{1, 1-t^2\}$, $z(t) = t^3$, so ist die beste Approximierende zwar nicht eindeutig, aber es existiert eine *ungerade* beste Approximierende an z in M. Denn definiert man den Defekt

$$d_{(a,b)}(t) := a + b(1-t^2) - t^3,$$

so ist $d_{(a,b)}(1) = a - 1$ und $d_{(a,b)}(-1) = a + 1$ und folglich

$$\|d_{(a,b)}\|_\infty \geq \max(|a-1|, |a+1|) \geq 1 \quad \text{für alle } (a,b) \in \mathbb{R}^2.$$

Also ist $d(z, M) \geq 1$. Wegen $\|d_{(0,0)}\|_\infty = 1$ ist $x_0^* = 0$ eine (ungerade) beste Approximierende. Weiter ist auch $\|d_{(0,b)}\|_\infty = 1$ für jedes $b \in [-1, 1]$ und daher $x_b^*(t) = b(1-t^2) - t^3$ für jedes $b \in [-1, 1]$ eine beste Approximierende an z in M. In Abbildung 4.2 geben

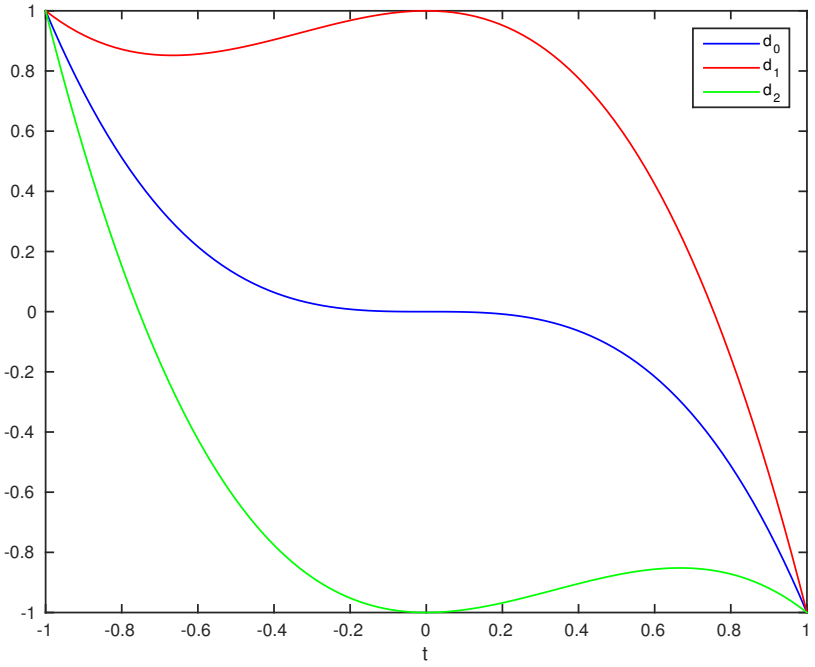

Abbildung 4.2: Optimaler Defekt für $b = 0$, $b = 1$ und $b = -1$

wir den Defekt $d_0 := d_{(0,0)}$, $d_1 := d_{(0,1)}$ und $d_2 = d_{(0,-1)}$ über $[-1, 1]$ an.

Beweis von Satz 4.1.9: Die Menge $P_M(z)$ der besten Approximierenden an z in M ist konvex, abgeschlossen und beschränkt, als Teilmenge des endlichdimensionalen linearen normierten Raumes $(M, \|\cdot\|_\infty)$ also kompakt. Wir definieren die Abbildung $S\colon P_M(z) \longrightarrow M$ durch $S(x) := Ax \circ T$ und zeigen

1. S bildet $P_M(z)$ in sich ab, d. h. es gilt $S(P_M(z)) \subset P_M(z)$,
2. S ist stetig.

Aus dem Brouwerschen Fixpunktsatz folgt die Existenz eines Fixpunktes von S in $P_M(z)$ und dies ist genau die Behauptung. Zu zeigen bleiben also die Aussagen 1. und 2.

1. Für $x \in P_M(z)$ ist

$$\begin{aligned} d(z,M) &= \|x - z\|_\infty \\ &\geq \|x \circ T - z \circ T\|_\infty \\ &\geq \|Ax \circ T - Az \circ T\|_\infty \\ &\qquad \text{Linearität von } A \text{ und } \|A\| \leq 1 \\ &= \|\underbrace{Ax \circ T}_{=Sx} - z\|_\infty \end{aligned}$$

   und daher ist $Sx \in P_M(z)$.

2. Wir zeigen, dass S sogar lipschitzstetig auf $P_M(z)$ und damit insbesondere stetig ist. Seien hierzu $x, y \in P_M(z)$. Dann ist

$$\begin{aligned} \|Sx - Sy\|_\infty &= \|Ax \circ T - Ay \circ T\|_\infty \\ &\leq \|x \circ T - y \circ T\|_\infty \\ &\leq \|x - y\|_\infty, \end{aligned}$$

   und damit ist auch 2. und der ganze Satz bewiesen.

□

Zusammenfassung: Ziel dieses Abschnitts war es, die klassischen Charakterisierungssätze und Eindeutigkeitsaussagen bei linearer T-Approximation zu präsentieren. Es wurde noch nicht auf die Approximation durch spezielle Teilräume eingegangen. so werden in den Anwendungen neben Polynomen und trigonometrischen Polynomen auch lineare Räume aus Splines betrachtet. Hierzu sei

$$\alpha = t_0 < t_1 < \cdots < t_k < t_{k+1} = \beta$$

eine feste Zerlegung von $[\alpha, \beta]$. Dann nennt man Elemente von

$$S_{n,k} := \{x \in C^{n-1}[\alpha, \beta] : x|_[t_j, t_{j+1}] \in \Pi_n,\ j = 0, \ldots, k\}$$

Splines vom Grad n. Die Dimension von $S_{n,k}$ ist $n + k + 1$. Der lineare Teilraum $S_{n,k}$ ist *kein* Haarscher Teilraum von $C[\alpha, \beta]$. □

4.2 Spezielle T-Approximationsaufgaben. T-Polynome

In diesem Abschnitt sollen die Ergebnisse des letzten Abschnitts zur Lösung spezieller Aufgaben angewandt werden. Es sei hier schon betont, dass wir auf den wichtigen Themenkreis, asymptotische Aussagen für $E_n(z) := d(z, \Pi_n)$ zu machen, erst später ausführlich eingehen werden. Es wird dann zwischen sätzen verschiedenen Typs unterschieden. Bei den *Jackson-Sätzen* wird aus der Glattheit von z auf das asymptotische Verhalten von $\{E_n(z)\}_{n\in\mathbb{N}}$ geschlossen. Bei den *Bernstein-Sätzen* wird umgekehrt aus dem asymptotischen Verhalten von $\{E_n(z)\}_{n\in\mathbb{N}}$ auf die Glattheit von z geschlossen.

Wir geben zunächst die Definition der Tschebyscheff (T)-Polynome an und zeigen dann, dass diese (im wesentlichen) Lösungen einer speziellen T-Approximationsaufgabe sind. Definition und einige Eigenschaften der T-Polynome T_n geben wir im nächsten Lemma an.

Lemma 4.2.1 *Für nichtnegatives ganzes n sei $T_n\colon [-1,1] \longrightarrow \mathbb{R}$ definiert durch*

$$T_n(t) := \cos(n \arccos t).$$

Dann gilt:

1. *$\{T_n(t)\}$ genügt der Rekursionsformel*

$$T_0(t) = 1, \quad T_1(t) = t, \quad T_{n+1}(t) + T_{n-1}(t) = 2tT_n(t).$$

2. *$T_n \in \Pi_n$ und $T_n(t) = 2^{n-1}t^n + p(t)$ mit $p \in \Pi_{n-2}$, $n = 1, 2, \ldots$.*

3. *$T_n(-t) = (-1)^n T_n(t)$ (Symmetrieeigenschaft).*

4. *Es ist*

$$\frac{2}{\pi}\int_{-1}^{1} \frac{T_m(t)T_n(t)}{\sqrt{1-t^2}}\,dt = \begin{cases} 2, & m = n = 0, \\ 1, & m = n \neq 0, \\ 0, & m \neq n. \end{cases}$$

5. *Es ist*

$$T_n(t) = \frac{1}{2}[(t + \sqrt{t^2-1})^n + (t - \sqrt{t^2-1})^n].$$

Beweis: Offenbar ist $T_0(t) = 1$, $T_1(t) = t$. Aus dem Additionstheorem für den Cosinus

$$\cos(a \pm b) = \cos a \cos b \mp \sin a \sin b$$

folgt durch Addition

$$\cos(a+b) + \cos(a-b) = 2\cos a \cos b.$$

Setzt man hier $a = n\phi$, $b = \phi$, so ist

$$\cos(n+1)\phi + \cos(n-1)\phi = 2\cos n\phi \cos\phi.$$

Mit $\phi = \arccos t$ folgt

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

also die behauptete Rekursionsformel. Die zweite Behauptung, dass nämlich $T_n(t) = 2^{n-1}t^n + p(t)$ mit $p \in \Pi_{n-2}$ für $n \in \mathbb{N}$ ist, beweisen wir durch vollständige Induktion nach n. Für $n = 1$ und $n = 2$ ist die Behauptung wegen $T_1(t) = t$ und $T_2(t) = 2t^2 - 1$ richtig. Angenommen, die Aussage ist für natürliche Zahlen $\leq n$ richtig. Dann ist

$$T_{n+1}(t) = 2tT_n(t) - T_{n-1}(t) = 2t(2^{n-1}t^n + p(t)) - (2^{n-2}t^{n-1} + q(t))$$

mit $p \in \Pi_{n-2}$ und $q \in \Pi_{n-3}$ und folglich $T_{n+1}(t) = 2^n t^{n+1} + r(t)$ mit $r \in \Pi_{n-1}$. Damit ist auch die zweite Aussage bewiesen. Die dritte Aussage erhält man ebenfalls leicht durch vollständige Induktion aus der Rekursionsformel. Mit Hilfe der Substitution $t = \cos\phi$, die $dt = -\sin\phi\, d\phi = -\sqrt{1-t^2}\, dt$ nach sich zieht, erhalten wir

$$\begin{aligned}\frac{2}{\pi}\int_{-1}^{1} \frac{T_m(t)T_n(t)}{\sqrt{1-t^2}}\, dt &= -\frac{2}{\pi}\int_{\pi}^{0} \cos m\phi \cos n\phi\, d\phi \\ &= \frac{1}{\pi}\int_{-\pi}^{\pi} \cos m\phi \cos n\phi\, d\phi \\ &= \frac{1}{\pi}\int_{-\pi}^{\pi} \frac{1}{2}[\cos(m-n)\phi + \cos(m+n)\phi]\, d\phi.\end{aligned}$$

Wegen

$$\int_{-\pi}^{\pi} \cos k\phi\, d\phi = 0, \quad k \in \mathbb{Z} \setminus \{0\}$$

folgt die vierte Behauptung. Zum Beweis der fünften Behauptung beachten wir, dass

$$\cos n\phi \pm i\, \sin n\phi = e^{\pm in\phi} = (e^{\pm\phi})^n = (\cos\phi \pm i\, \sin\phi)^n$$

und daher

$$\cos n\phi = \frac{1}{2}[(\cos\phi + i\, \sin\phi)^n + (\cos\phi - i\, \sin\phi)^n].$$

Mit $t = \cos\phi$, $i\, \sin\phi = \sqrt{t^2-1}$ folgt die letzte Behauptung. □

Die Polynome $T_0, T_1, \ldots$ nennt man *Tschebyscheff (T)-Polynome* (erster Art). Offenbar besitzt T_n die n Nullstellen

$$t_j := \cos\frac{2(n-j)+1}{2n}\pi, \qquad j = 1, \ldots, n,$$

in $(-1, 1)$ und es ist

$$-1 < t_1 < t_2 < \cdots < t_n < 1.$$

In Abbildung 4.3 haben wir die ersten T-Polynome über dem Intervall $[0, 1]$ (Symmetrieeigenschaft!) aufgetragen.

Nun kommen wir zu dem angekündigten Zusammenhang zwischen T-Approximation und T-Polynomen.

Satz 4.2.2 *Sei $n \in \mathbb{N}$ und $p^* \in \Pi_{n-1}$ die beste T-Approximierende an $z(t) := t^n$ in Π_{n-1} über dem Intervall $B := [-1, 1]$. Dann ist*

$$t^n - p^*(t) = 2^{-n+1} T_n(t)$$

und daher

$$E_{n-1}(t^n) = d(z, \Pi_{n-1}) = 2^{-n+1}.$$

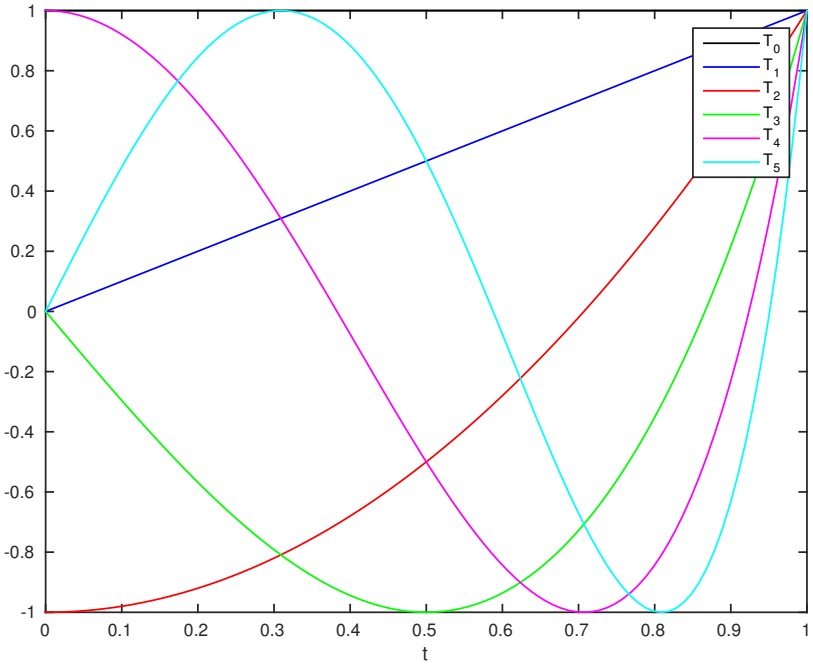

Abbildung 4.3: Die ersten Tschebyscheff-Polynome auf $[0, 1]$

Beweis: Wir wenden die hinreichende Bedingung im Alternantensatz 4.1.5 an und zeigen, dass $p^* \in \Pi_{n-1}$, definiert durch

$$p^*(t) := t^n - 2^{-n+1} T_n(t),$$

(die eindeutige) beste T-Approximierende an $z(t) := t^n$ in Π_{n-1} ist. Da $T_n \in \Pi_n$ den höchsten Koeffizienten 2^{n-1} hat, ist $p^* \in \Pi_{n-1}$. Offensichtlich ist $\|p^* - z\|_\infty = 2^{-n+1}$ und die Punkte

$$t_j = \cos \frac{n-j}{n} \pi, \qquad j = 0, \ldots, n,$$

bilden eine Alternante der Länge $n + 1$, da

(a) $-1 = t_0 < t_1 < \cdots < t_n = 1$,

(b) $|p^*(t_j) - z(t_j)| = 2^{-n+1} = \|p^* - z\|_\infty$, $j = 0, \ldots, n$,

(c) $p^*(t_j) - z(t_j) = -2^{-n+1} T_n(t_j) = (-1)^{n+j+1} 2^{-n+1}$, $j = 0, \ldots, n$.

Aus dem Alternantensatz folgt die Behauptung. □

Bemerkung: Gelegentlich wird der letzte Satz 4.2.2 etwas anders gewendet. Stellt man nämlich die Aufgabe, dasjenige Polynom n-ten Grades mit höchstem Koeffizienten gleich 1 zu bestimmen, dessen Maximumnorm über dem Intervall $[-1, 1]$ möglichst klein ist, so ist diese Aufgabe äquivalent dazu, $z(t) := t^n$ durch Elemente aus Π_{n-1} über dem Intervall $[-1, 1]$ nach Tschebyscheff zu approximieren. Aus Satz 4.2.2 erhalten wir die Antwort. Denn $2^{-n+1} T_n$ ist dasjenige Polynom n-ten Grades mit höchstem Koeffizienten

1, das minimale Maximumnorm über $[-1,1]$ besitzt bzw. die Nullfunktion am besten approximiert. □

Man kann und sollte sich natürlich fragen, wozu die T-Polynome gut sind, weshalb z. B. die Extremaleigenschaft der T-Polynome wie sie in Satz 4.2.2 oder der anschließenden Bemerkung deutlich wird, nützlich ist. Hierzu geben wir in Form von Bemerkungen verschiedene Antworten.

Bemerkungen: 1. Aus der Numerischen Mathematik (siehe z. B. J. WERNER (1992, S. 140)) ist die folgende Aussage über den Fehler bei der Polynominterpolation bekannt:

- *Sei $x \in C^n[\alpha,\beta]$ und $L_{n-1}(x) \in \Pi_{n-1}$ das durch paarweise verschiedene Knoten $t_j \in [\alpha,\beta]$, $j = 1,\dots,n$, festgelegte Interpolationspolynom. Dann gibt es zu jedem $t \in [\alpha,\beta]$ ein $\xi(t) \in (\alpha,\beta)$ mit*

$$x(t) - L_{n-1}(x)(t) = \frac{x^{(n)}(\xi(t))}{n!} \prod_{j=1}^{n} (t - t_j).$$

Mit

$$w(t) := \prod_{j=1}^{n} (t - t_j)$$

ist also

$$\|x - L_n(x)\|_\infty \le \frac{\|x^{(n)}\|_\infty}{n!} \|w\|_\infty.$$

Diese Fehlerabschätzung wird optimal, wenn die Knoten t_j so gelegt werden, dass $\|w\|_\infty$ minimal ist. Anders gewendet: Die t_j sind die Nullstellen desjenigen Polynoms n-ten Grades mit höchstem Koeffizienten 1, dessen Minimalabweichung von der Nullfunktion minimal ist. Für $[\alpha,\beta] = [-1,1]$ kennen wir die Antwort. Denn $2^{-n+1}T_n$ ist das gesuchte Polynom, dessen Nullstellen

$$t_j = \cos \frac{2(n-j)+1}{2n}\pi, \qquad j = 1,\dots,n,$$

die gesuchten Knoten sind. Ist $[\alpha,\beta] \neq [-1,1]$, so transformiere man $[\alpha,\beta]$ mittels

$$s = \frac{1}{\beta - \alpha}[2t - (\alpha + \beta)]$$

auf $[-1,1]$. Aus

$$\cos \frac{2(n-j)+1}{2n}\pi = \frac{1}{\beta - \alpha}[2t_j - (\alpha + \beta)], \quad j = 1,\dots,n,$$

erhält man die gesuchten Knoten

$$t_j = \frac{\alpha + \beta}{2} + \frac{\beta - \alpha}{2} \cos \frac{2(n-j)+1}{2n}\pi, \quad j = 1,\dots,n.$$

In dem bekannten Runge-Beispiel (siehe z. B. J. WERNER (1992, S. 144) oder M. J. D. POWELL (1981, S. 37 ff.)) ist $x(t) := 1/(1+t^2)$ und $[\alpha,\beta] = [-5,5]$. In Abbildung 4.4

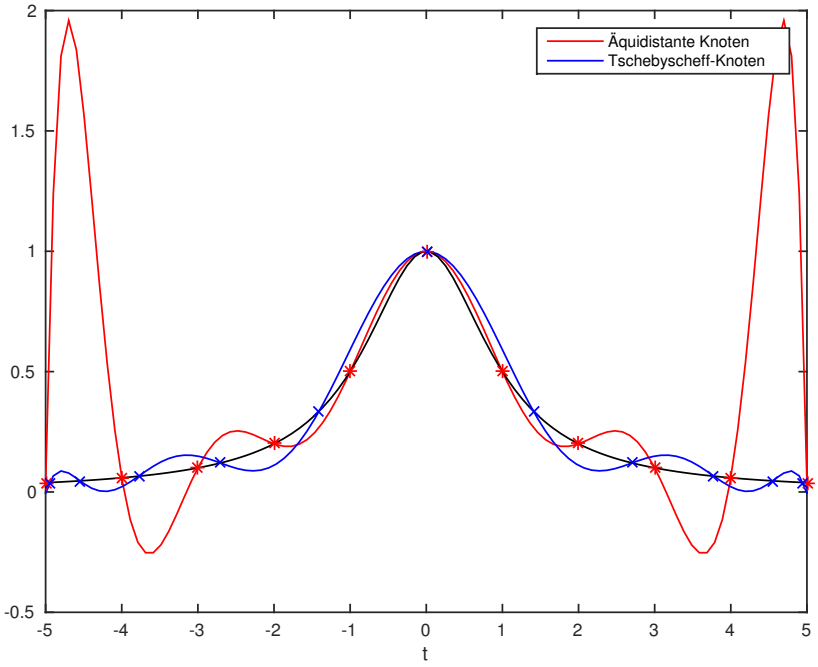

Abbildung 4.4: Das Runge-Beispiel mit äquidistanten Knoten und Tschebyscheff-Knoten

geben wir zu 11 vorgegebenen äquidistanten Knoten bzw. Tschebyscheff-Knoten das zugehörige Interpolationspolynom 10-ten Grades an.

2. Im nächsten Abschnitt gehen wir auf den Remez-Algorithmus zur numerischen Berechnung der besten T-Approximierenden an eine vorgegebene Funktion $z \in C[\alpha, \beta]$ in einem n- oder $n+1$-dimensionalen Haarschen Teilraum M an. Sei etwa $M = \Pi_n$. Wegen des Alternantensatzes ist eine beste Approximierende $x^* \in \Pi_n$ charakterisiert durch die Existenz einer zugehörigen Alternante $\{t_0, \ldots, t_{n+1}\}$ mit

(a) $\alpha \leq t_0 < \cdots < t_{n+1} \leq \beta$,

(b) $|x^*(t_j) - z(t_j)| = \|x^* - z\|_\infty$, $j = 0, \ldots, n+1$,

(c) $\operatorname{sign}(x^*(t_j) - z(t_j)) = \epsilon(-1)^j$, $j = 0, \ldots, n+1$, mit $\epsilon \in \{-1, 1\}$.

Das Remez-Verfahren startet mit einer *Ausgangs-Referenz*

$$T_0 : \alpha \leq t_0^0 < \cdots < t_{n+1}^0 \leq \beta.$$

Eine Verwechslung der Start-Referenz T_0 mit dem T-Polynom T_0 sollte vermieden werden! Anschließend wird ein *diskretes* Problem, nämlich

$$(\mathrm{P}_0) \qquad \text{Minimiere} \quad f_0(x) := \|x - z\|_{\infty, T_0} = \max_{j=0,\ldots,n+1} |x(t_j^0) - z(t_j^0)| \quad \text{auf} \quad \Pi_n,$$

gelöst. Dieses Problem besitzt wegen Satz 4.1.4 eine eindeutige Lösung x_0, da Π_n ein $(n+1)$-dimensionaler Haarscher Teilraum von $C(T_0)$ ist. Anschließend wird getestet,

ob x_0 schon beste Approximierende auf $[\alpha, \beta]$ ist. Ist dies nicht der Fall, so wird die Ausgangs-Referenz T_0 zu T_1 verändert und das (P_0) entsprechende Problem (P_1) gelöst. Auf die Einzelheiten wollen wir hier nicht eingehen, sondern uns nur die folgende naheliegende Frage stellen: Wie sollte T_0 in $[\alpha, \beta]$ platziert werden? Günstig wäre es offenbar, wenn die durch Lösung des diskreten Problems (P_0) gewonnene Näherung x_0 "möglichst nahe" bei der Lösung x^* liegt. Durch den nächsten Satz wird begründet, weshalb es günstig ist, als Ausgangs-Referenz die Extremstellen des T-Polynoms T_{n+1} zu nehmen, jedenfalls dann, wenn $[\alpha, \beta] = [-1, 1]$ (andernfalls muss man auf dieses Intervall transformieren). □

Satz 4.2.3 *Sei*

$$t_j := \cos \frac{n+1-j}{n+1}\pi, \qquad j = 0, \ldots, n+1.$$

Ist $z \in \Pi_{n+1}$ *und* $x_0 \in \Pi_n$ *die Lösung der Aufgabe*

$$(\mathrm{P}_0) \qquad \textit{Minimiere} \quad f_0(x) := \max_{j=0,\ldots,n+1} |x(t_j) - z(t_j)| \quad \textit{auf} \quad \Pi_n,$$

so ist $x_0 = x^*$ *auch beste Approximierende an* z *in* Π_n *auf dem Intervall* $[-1, 1]$.

Beweis: Zunächst zeigen wir:

- Es existiert $\rho \in \mathbb{R}$ mit

$$x_0(t_j) - z(t_j) = (-1)^j \rho, \qquad j = 0, \ldots, n+1.$$

Denn[2]: Wir wenden Satz 4.1.2 mit $B := \{t_0, \ldots, t_{n+1}\}$ und $M := \Pi_n$ an. Sei etwa $\{p_0, \ldots, p_n\}$ eine Basis von M und daher $M = \operatorname{span}\{p_0, \ldots, p_n\}$. Hiernach ist

$$0 \in \operatorname{co}\left(\{\operatorname{sign}(x_0(t_j) - z(t_j)) \begin{pmatrix} p_0(t_j) \\ \vdots \\ p_n(t_j) \end{pmatrix} : t_j \in B(x_0 - z)\}\right)$$

mit

$$B(x_0 - z) := \{t_j : |x_0(t_j) - z(t_j)| = \max_{i=0,\ldots,n+1} |x_0(t_i) - z(t_i)|\}.$$

Der Satz von Carathéodory (Satz 2.2.13) und die Tatsache, dass Π_n ein $(n+1)$-dimensionaler Haarscher Teilraum ist, liefern $B = B(x_0 - z)$ und die Darstellung

$$0 = \sum_{j=0}^{n+1} \underbrace{\lambda_j \operatorname{sign}(x_0(t_j) - z(t_j))}_{=:q_j} p_i(t_j), \qquad i = 0, \ldots, n,$$

mit

$$\lambda_j > 0, \quad j = 0, \ldots, n+1, \qquad \sum_{j=0}^{n+1} \lambda_j = 1.$$

[2]Eigentlich beweisen wir jetzt unnötigerweise den Alternantensatz noch einmal für diskrete Approximationsaufgaben.

Nach Satz 3.4.4 besitzt das Gleichungssystem

$$\sum_{j=0}^{n+1} q_j p_i(t_j) = 0, \quad i = 0, \ldots, n, \qquad \sum_{j=0}^{n+1} |q_j| = 1$$

eine bis auf einen gemeinsamen Faktor $\epsilon \in \{-1, 1\}$ eindeutige Lösung $(q_0, \ldots, q_{n+1})^T$, deren Komponenten im Vorzeichen alternieren. Daher ist

$$\operatorname{sign}(x_0(t_j) - z(t_j)) = \epsilon(-1)^j, \quad j = 0, \ldots, n+1,$$

und folglich

$$x_0(t_j) - z(t_j) = (-1)^j \rho \quad \text{mit} \quad \rho := \epsilon \max_{i=0,\ldots,n+1} |x_0(t_i) - z(t_i)|.$$

Bisher ging weder $z \in \Pi_{n+1}$ ein, noch dass die t_j Extremstellen von T_{n+1} sind. Wegen

$$x_0 - z \in \Pi_{n+1}, \qquad x_0(t_j) - z(t_j) = (-1)^j \rho, \quad j = 0, \ldots, n+1,$$

sowie

$$T_{n+1} \in \Pi_{n+1}, \qquad T_{n+1}(t_j) = (-1)^j (-1)^{n+1}, \quad j = 0, \ldots, n+1,$$

ist $x_0 - z = (-1)^{n+1} \rho T_{n+1}$ und damit $|\rho| = \|x_0 - z\|_\infty$. Die vhinreichende Richtung aus dem Alternantensatz 4.1.5 liefert, dass $x_0 = x^*$ beste Approximierende an z in Π_n (bezüglich des Intervalls $[-1, 1]$ ist). □

Bemerkung: Den allgemeinen Fall $[\alpha, \beta] \neq [-1, 1]$ führt man wiederum durch eine Transformation auf den speziellen zurück und erhält, dass man

$$t_j := \frac{1}{2}(a+b) + \frac{1}{2}(b-a) \cos \frac{n+1-j}{n+1} \pi, \qquad j = 0, \ldots, n+1,$$

als Start-Referenz wählen sollte. □

4.3 Die Numerische Behandlung linearer T-Approximationsaufgaben

In diesem Abschnitt sei eine Approximationsaufgabe gegeben, bei der $(X, \|\cdot\|) = (C[\alpha, \beta], \|\cdot\|_\infty)$ der zugrunde liegende lineare normierte Raum ist, mit Elementen eines $(n+1)$-dimensionalen Haarschen Teilraums $M = \operatorname{span}\{x_0, \ldots, x_n\}$ von $C[\alpha, \beta]$ approximiert wird und $z \in C[\alpha, \beta]$ die zu approximierende Funktion ist.

Ein $(n+2)$-Tupel $T = (t_0, \ldots, t_{n+1}) \in \mathbb{R}^{n+2}$ mit[3] $\alpha \leq t_0 < t_1 < \cdots < t_{n+1} \leq \beta$ nennen wir eine *Referenz* (der Länge $n+2$). Die Menge aller Referenzen bezeichnen wir mit $\mathcal{T}$. Entscheidend für die numerische Behandlung ist die folgende Aussage, in der wir lediglich frühere Ergebnisse zusammenfassen.

[3] Gelegentlich werden wir das $(n+2)$-Tupel T auch als *Menge* auffassen und z. B. $t \in T$ schreiben. Das sollte zu keinen Verwirrungen führen.

Lemma 4.3.1 *(a) Sei $T = (t_0, \ldots, t_{n+1}) \in \mathcal{T}$ eine Referenz. Dann besitzt die diskrete Approximationsaufgabe*

$$(\mathrm{P}_T) \qquad \textit{Minimiere} \quad f_T(x) := \|x - z\|_{T,\infty} = \max_{t \in T} |x(t) - z(t)| \quad \textit{auf} \quad M$$

genau eine Lösung $x_T^* \in M$ und diese ist charakterisiert durch

1. $|x_T^*(t_j) - z(t_j)| = \|x_T^* - z\|_{T,\infty}$, $j = 0, \ldots, n+1$,
2. $\operatorname{sign}(x_T^*(t_j) - z(t_j)) = -\operatorname{sign}(x_T^*(t_{j+1}) - z(t_{j+1}))$, $j = 0, \ldots, n$.

Es ist

$$\begin{aligned} \|x_T^* - z\|_{T,\infty} &\leq d(z, M) \quad \textit{(de La Vallée Poussin)} \\ &\leq \|x_T^* - z\|_\infty. \end{aligned}$$

Ist daher $\|x_T^* - z\|_{T,\infty} = \|x_T^* - z\|_\infty$, so ist x_T^* die beste Approximierende an z in M.

(b) Ist $x^* \in M$ die beste Approximierende an z in M, so gibt es eine Referenz $T = (t_0, \ldots, t_{n+1}) \in \mathcal{T}$ mit

1. $|x_T^*(t_j) - z(t_j)| = \|x_T^* - z\|_\infty$, $j = 0, \ldots, n+1$,
2. $\operatorname{sign}(x_T^*(t_j) - z(t_j)) = -\operatorname{sign}(x_T^*(t_{j+1}) - z(t_{j+1}))$, $j = 0, \ldots, n$,

also mit $x^* = x_T^*$ und $\|x^* - z\|_\infty = \|x_T^* - z\|_{T,\infty}$.

Nun liegt es nahe, wie man im Prinzip vorgeht. Einzelne Bausteine im folgenden Algorithmus werden wir später noch näher beschreiben müssen. Wir geben jetzt das *allgemeine Remez-Verfahren* an.

- Wähle eine Start-Referenz $T_0 \in \mathcal{T}$.
- Für $k = 0, 1, \ldots$:
  - Setze $T := T_k$ und berechne die Lösung x_T^* von

    $$(\mathrm{P}_T) \qquad \text{Minimiere} \quad f_T(x) := \max_{t \in T} |x(t) - z(t)| \quad \text{auf} \quad M.$$

    Setze $\rho_T := f_T(x_T^*)$.
  - Berechne $\|x_T^* - z\|_\infty = \max_{t \in [\alpha, \beta]} |x_T^*(t) - z(t)|$.
  - Falls $\rho_T = \|x_T^* - z\|_\infty$, dann:
    - x_T^* ist die beste Approximierende an z in M, STOP.
  - Andernfalls:
    - Bestimme eine neue Referenz $T_{k+1} = (t_0^{(k+1)}, \ldots, t_{n+1}^{(k+1)}) \in \mathcal{T}$ mit
      1. $|x_T(t) - z(t)| \geq \rho_T$ für $t \in T_{k+1}$,
      2. Es existiert $s \in T_{k+1}$ mit $|x_T^*(s) - z(s)| = \|x_T^* - z\|_\infty$,

3. $\operatorname{sign}(x_T^*(t_j^{(k+1)}) - z(t_j^{(k+1)})) = \epsilon(-1)^j,\ j = 0,\dots,n+1$ mit $\epsilon \in \{-1,1\}$.

Später werden wir auf die folgenden Fragen näher eingehen müssen.

(i) Wie berechnet man bei vorgegebenem $T \in \mathcal{T}$ die Lösung x_T^* von (P_T)?

(ii) Wie berechnet man ein Maximum einer gegebenem Funktion (hier: $|x_T^* - z|$) auf einem Intervall?

(iii) Welche Auswahlregel $T_k \longrightarrow T_{k+1}$ genügt den Bedingungen 1.–3.?

Zunächst gehen wir nur auf (i) ein und geben im folgenden Lemma u. a. an, wie die Koeffizienten $a_T = (a_{i,T})$ der Lösung $x_T^* = \sum_{i=0}^n a_{i,T} x_i$ von (P_T) berechnet werden können. Danach werden wir schon einen Konvergenzsatz für das oben angegebene allgemeine Remez-Verfahren (ohne Spezifikation der Auswahlregel) beweisen. Schließlich werden wir noch ausführlich auf die Frage (iii) eingehen.

Lemma 4.3.2 *Sei $T = \{t_0,\dots,t_{n+1}\} \in \mathcal{T}$ und $x_T^* = \sum_{i=0}^n a_{i,T} x_i \in M$ die Lösung von*

$$(\mathrm{P}_T) \qquad \textit{Minimiere} \quad \|x - z\|_{T,\infty} = \max_{t\in T} |x(t) - z(t)| \quad \textit{auf} \quad M$$

sowie $\rho_T := \|x_T^* - z\|_{T,\infty}$. Dann gilt:

(a) Das Gleichungssystem

$$(*) \qquad \sum_{i=0}^n a_i x_i(t_j) + (-1)^j \rho = z(t_j), \quad j = 0,\dots,n+1,$$

besitzt eine eindeutige Lösung (a,ρ) und es ist $a = a_T$ und $|\rho| = \rho_T$.

(b) Sei $\sigma_T := \operatorname{sign}(x_T^*(t_{n+1}) - z(t_{n+1}))$ und $\lambda_T := (\lambda_{0,T},\dots,\lambda_{n+1,T})^T$ die eindeutige Lösung von

$$(**) \qquad \sum_{j=0}^{n+1} \lambda_j (-1)^{n+1-j} x_i(t_j) = 0, \quad i = 0,\dots,n, \qquad \sum_{j=0}^{n+1} \lambda_j = 1.$$

Dann ist $\lambda_{j,T} > 0$, $j = 0,\dots,n+1$, und

$$\rho_T = \sigma_T \sum_{j=0}^{n+1} \lambda_{j,T} (-1)^{n-j} z(t_j).$$

Beweis: Der erste Teil des Satzes folgt aus Lemma 4.3.1. Denn hiernach besitzt (P_T) eine (eindeutige) Lösung $x_T^* = \sum_{i=0}^n a_{i,T} x_i$. Mit $\rho_T := \|x_T^* - z\|_{T,\infty}$ genügt diese Lösung den Bedingungen

$$|x_T^*(t_j) - z(t_j)| = \rho_T, \quad j = 0,\dots,n+1,$$

und

$$\operatorname{sign}(x_T^*(t_j) - z(t_j)) = -\operatorname{sign}(x_T^*(t_{j+1}) - z(t_{j+1})), \quad j = 0,\dots,n.$$

Dann ist aber

$$\begin{aligned} x_T^*(t_j) - z(t_j) &= \operatorname{sign}(x_T^*(t_j) - z(t_j))\rho_T \\ &= (-1)^j \operatorname{sign}(x_T^*(t_0) - z(t_0))\rho_T, \quad j = 0, \ldots, n+1. \end{aligned}$$

Mit

$$\rho := -\operatorname{sign}(x_T^*(t_0) - z(t_0))\rho_T$$

ist also

$$x_T^*(t_j) + (-1)^j \rho = \sum_{i=0}^{n} a_{i,T} x_i + (-1)^j \rho = z(t_j), \quad j = 0, \ldots, n+1.$$

Damit ist gezeigt, dass $(*)$ eine Lösung besitzt. Da dies für beliebige rechte Seiten gilt ist das lineare Gleichungssystem $(*)$ mit $n+2$ Gleichungen und ebenso vielen Unbekannten auch *eindeutig* lösbar.
Das Gleichungssystem

$$\sum_{j=0}^{n+1} q_j x_i(t_j) = 0, \quad i = 0, \ldots, n, \qquad \sum_{j=0}^{n+1} |q_j| = 1$$

besitzt nach Satz 3.4.4 genau eine Lösung $(q_0, \ldots, q_{n+1})^T$ mit $q_{n+1} > 0$. Ferner alternieren die q_j im Vorzeichen. Daher ist auch $(**)$ eindeutig lösbar und für die Lösung $(\lambda_{0,T}, \ldots, \lambda_{n+1,T})^T$ gilt $q_j = \lambda_{j,T}(-1)^{n+1-j}$ mit $\lambda_{j,T} > 0$, $j = 0, \ldots, n+1$. Ferner ist

$$\begin{aligned} \rho_T &= \sum_{j=0}^{n+1} \lambda_{j,T} \rho_T \\ &= \sum_{j=0}^{n+1} \lambda_{j,T} |x_T^*(t_j) - z(t_j)| \\ &= \sum_{j=0}^{n+1} \lambda_{j,T} \operatorname{sign}(x_T^*(t_j) - z(t_j))(x_T^*(t_j) - z(t_j)) \\ &= \sigma_T \sum_{j=0}^{n+1} \lambda_{j,T}(-1)^{n+1-j}(x_T^*(t_j) - z(t_j)) \\ &= \sigma_T \sum_{j=0}^{n+1} \lambda_{j,T}(-1)^{n-j} z(t_j), \end{aligned}$$

da

$$\begin{aligned} \sum_{j=0}^{n+1} \lambda_{j,T}(-1)^{n-j} x_T^*(t_j) &= -\sum_{j=0}^{n+1} q_j \left(\sum_{i=0}^{n} a_{i,T} x_i(t_j) \right) \\ &= -\sum_{i=0}^{n} a_{i,T} \Bigg(\underbrace{\sum_{j=0}^{n+1} q_j x_i(t_j)}_{=0} \Bigg) \\ &= 0. \end{aligned}$$

Damit ist das Lemma bewiesen. □

Beispiel: Sei $[\alpha, \beta] := [0, 1]$, $M := \Pi_2$, $z(t) := t^3$. Sei ferner

$$T := (0, 0.3, 0.8, 1\} =: \{t_0, t_1, t_2, t_3).$$

Zu bestimmen sei die Lösung $x_T^* \in M$ der diskreten Optimierungsaufgabe

$$(\mathrm{P}_T) \qquad \text{Minimiere} \quad f_T(x) := \|x - z\|_{T,\infty} = \max_{t \in T} |x(t) - z(t)| \quad \text{auf} \quad M.$$

Durch Lösen des linearen Gleichungssystem $(*)$ in Lemma 4.3.2 (wir haben die Monome 1, t und t^2 als Basis von Π_2 gewählt, was i. Allg. nicht empfehlenswert ist) erhalten wir

$$x_T^*(t) = 0.03 - 0.56t + 1.50t^2, \qquad \rho_T := \|x_T^* - z\|_{T,\infty} = 0.03.$$

In Abbildung 4.5 geben wir den Defekt $x_T^* - z$ auf dem Intervall $[0, 1]$ an. Man erkennt,

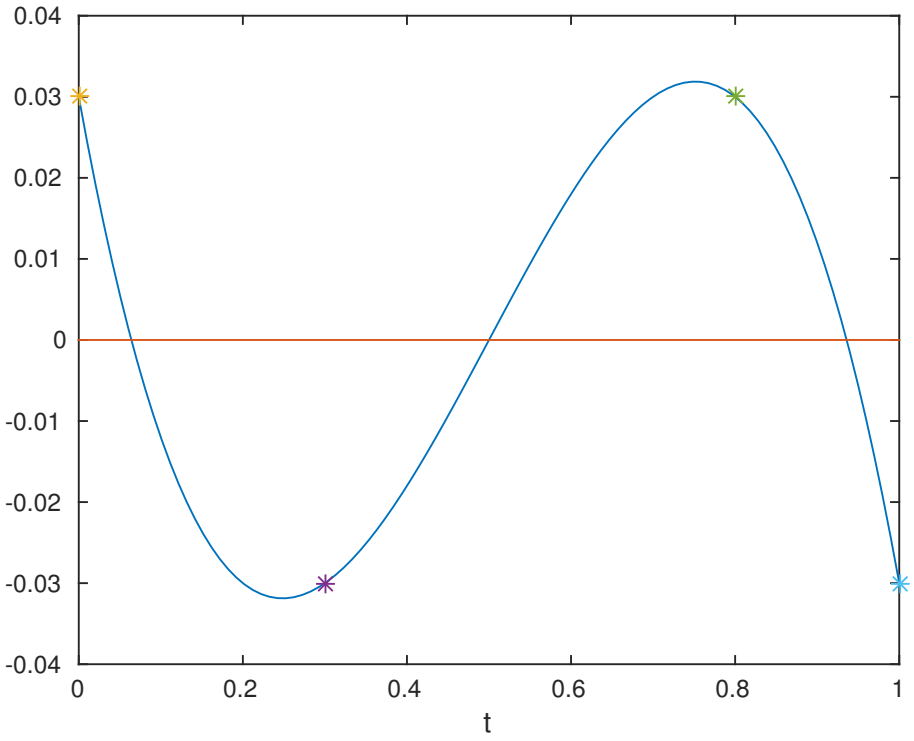

Abbildung 4.5: Defekt $x_T^* - z$ mit $T = (0, 0.3, 0.8, 1)$

dass $\|x_T^* - z\|_\infty$ nur unwesentlich größer als $\|x_T^* - z\|_{T,\infty}$ ist und damit schon eine gute Approximation für die beste Approximierende an z in Π_2 sein dürfte. Wegen Satz 4.2.3 und der anschließenden Bemerkung wissen wir, wie wir diese berechnen können. Und zwar haben wir das diskrete Problem bezüglich der Referenz $T = (0, 0.25, 0.75, 1)$ zu lösen. Diesmal erhalten wir

$$x_T^*(t) = 0.03125 - 0.5625t + 1.50t^2, \qquad \rho_T := \|x_T^* - z\|_\infty = 0.03125.$$

In Abbildung 4.6 geben wir wieder den Defekt $x_T^* - z$ an. □

Für den Konvergenzsatz benötigen wir das folgende Lemma.

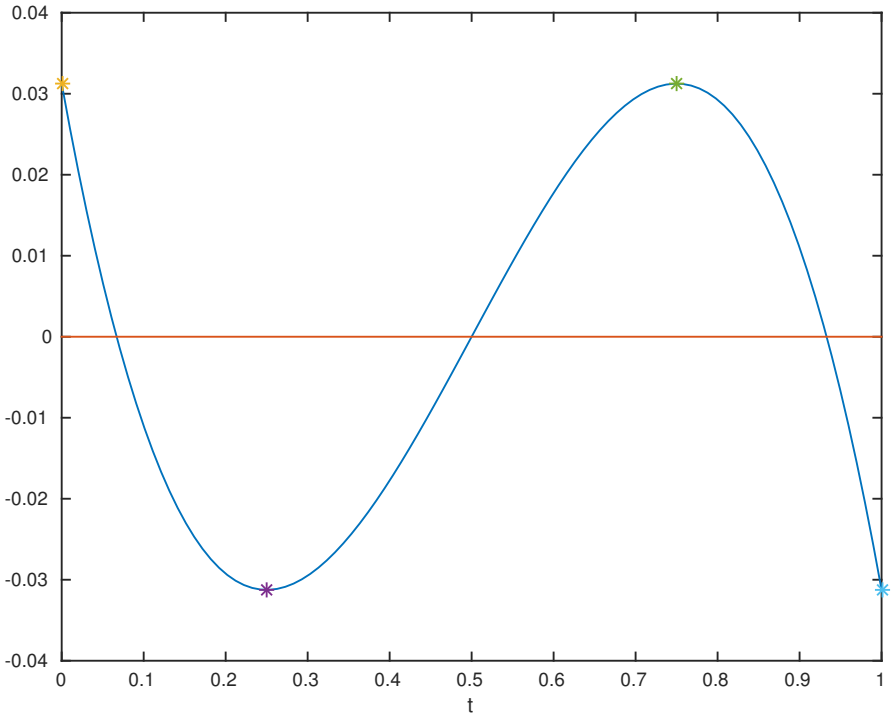

Abbildung 4.6: Defekt $x_T^* - z$ mit $T = (0, 0.25, 0.75, 1)$

Lemma 4.3.3 *Für $T \in \mathcal{T}$ seien a_T, ρ_T und λ_T eie in Lemma 4.3.2 definiert. Dann sind die Abbildungen $a\colon \mathcal{T} \longrightarrow \mathbb{R}^{n+1}$, $\rho\colon \mathcal{T} \longrightarrow \mathbb{R}$ und $\lambda\colon \mathcal{T} \longrightarrow \mathbb{R}^{n+2}$ mit $a(T) := a_T$ usw. stetig. Ist ferner $d > 0$, so ist*

$$K_d := \{T \in \mathcal{T} : \rho_T \geq d\}$$

kompakt.

Beweis: Die Stetigkeit von a, ρ und λ dürfte klar sein, da z. B. (a_T, ρ_T) im wesentlichen Lösung eines linearen Gleichungssystems ist, dessen Koeffizientenmatrix und rechte Seite stetig von T abhängen[4], siehe die Teile (a) und (b) von Lemma 4.3.2. Da K_d eine Teilmenge der beschränkten Menge $\mathcal{T} \subset \mathbb{R}^{n+2}$ ist, bleibt zu zeigen, dass K_d abgeschlossen ist. Sei hierzu $\{T_k\} \subset K_d$ eine Folge mit $T_k \to T \in \mathbb{R}^{n+2}$. Dann ist natürlich $T \in \mathrm{cl}\,(\mathcal{T})$. Ist $T \in \mathcal{T}$, so folgt aus $\rho_{T_k} \geq d$ und der Stetigkeit von ρ, dass $\rho_T \geq d$ und damit $T \in K_d$. Daher bringen wir jetzt die Annahme $T \in \mathrm{cl}\,(\mathcal{T}) \setminus \mathcal{T}$ zum Widerspruch. Offenbar besteht T wegen $T \in \mathrm{cl}\,(\mathcal{T}) \setminus \mathcal{T}$ aus höchstens $n+1$ paarweise verschiedenen Komponenten. Da M ein $(n+1)$-dimensionaler linearer Teilraum von $C[\alpha, \beta]$ ist und T höchstens $n+1$ paarweise verschiedene Komponenten besitzt, existiert ein $x \in M$ mit $x(t) = z(t)$ für $t \in T$. Nun ist

$$d \leq \rho_{T_k} = \|x_{T_k}^* - z\|_{T_k,\infty} \leq \|x - z\|_{T_k,\infty}$$

[4] ρ_T ist eigentlich der *Absolutbetrag* der letzten Komponente des entsprechenden Gleichungssystems.

und

$$\lim_{k\to\infty} \|x - z\|_{T_k,\infty} = \lim_{k\to\infty} \max_{t\in T_k} |x(t) - z(t)| = \max_{t\in T} |x(t) - z(t)| = 0,$$

womit wir den gewünschten Widerspruch erreicht haben. □

Nun folgt der Konvergenzsatz.

Satz 4.3.4 *Sei $M \subset C]\alpha,\beta]$ ein $(n+1)$-dimensionaler Haarscher Teilraum und $z \in C[\alpha,\beta]\setminus M$. Dann bricht das oben angegebene allgemeine Remez-Verfahren zur Lösung der T-Approximationsaufgabe entweder nach endlich vielen schritten mit der Lösung x^* ab oder es liefert Folgen $\{T_k\} \subset \mathcal{T}$, $\{x_k\} \subset M$ und $\{\rho_k\} \subset \mathbb{R}_+$, wobei $x_k := x^*_{T_k}$, $\rho_k := \rho_{T_k}$, mit folgenden Eigenschaften:*

(a) Es existiert $q \in (0,1)$ mit

$$d(z,M) - \rho_{k+1} \le q[d(z,M) - \rho_k], \qquad k = 0,1,\ldots,$$

d. h. die Folge $\{\rho_k\}$ konvergiert wenigstens linear gegen den Minimalabstand $d(z,M)$.

(b) Die Folge $\{x_k\} \subset M$ konvergiert gleichmäßig gegen die beste Approximierende x^* an z in M.

Beweis: Das Verfahren breche nicht vorzeitig ab, liefere also Folgen $\{T_k\}$ von Referenzen, $\{x_k\}$ von diskreten besten Approximierenden und $\{\rho_k\}$ von diskreten Minimalabständen.

Da das Verfahren nicht vorzeitig abbricht, ist

$$\rho_k = \|x_k - z\|_{T_k,\infty} < \|x_k - z\|_\infty, \qquad k = 0,1,\ldots.$$

Wegen Teil (a) von Lemma 4.3.1 ist daher

$$\rho_k = \|x_k - z\|_{T_k,\infty} < d(z,M) < \|x_k - z\|_\infty, \qquad k = 0,1,\ldots.$$

Unter Beachtung des Teiles (b) von Lemma 4.3.2 sei[5] $\lambda^{(k)} := \lambda_{T_k}$ und $\sigma_k := \sigma_{T_k}$. Wegen Lemma 4.3.2 (b) mit $T = T_{k+1}$ ist

$$\begin{aligned}
\rho_{k+1} &= \sigma_{k+1} \sum_{j=0}^{n+1} \lambda_j^{(k+1)} (-1)^{n-j} z(t_j^{(k+1)}) \\
&= \sigma_{k+1} \sum_{j=0}^{n+1} \lambda_j^{(k+1)} (-1)^{n+1-j} [x_k(t_j^{(k+1)}) - z(t_j^{(k+1)})] \\
&= \underbrace{\epsilon\sigma_{k+1}}_{=1} \sum_{j=0}^{n+1} \lambda_j^{(k+1)} |x_k(t_j^{(k+1)}) - z(t_j^{(k+1)})| \\
&= \sum_{j=0}^{n+1} \lambda_j^{(k+1)} |x_k(t_j^{(k+1)}) - z(t_j^{(k+1)})|,
\end{aligned}$$

[5] Hier schreiben wir den Iterationsindex nach oben, da wir auch auf die Komponenten zugreifen werden.

denn wegen Austauschregel 3. ist

$$\operatorname{sign}\left(x_k(t_j^{(k+1)}) - z(t_j^{(k+1)})\right) = \epsilon(-1)^{n+1-j}$$

mit $\epsilon \in \{-1, 1\}$. Wegen $\sum_{j=0}^{n+1} \lambda_j^{(k+1)} = 1$ ist

$$\rho_{k+1} = \rho_k + \sum_{j=0}^{n+1} \lambda_j^{(k+1)} [\underbrace{|x_k(t_j^{(k+1)}) - z(t_j^{(k+1)})| - \rho_k}_{\geq 0 \text{ wegen Austauschregel 1.}}].$$

Nach Austauschregel 2. für die Referenz T_{k+1} existiert $j_k \in \{0, \ldots, n+1\}$ mit

$$|x_k(t_{j_k}^{(k+1)}) - z(t_{j_k}^{(k+1)})| = \|x_k - z\|_\infty.$$

Daher ist

$$\begin{aligned} \rho_{k+1} &\geq \rho_k + \lambda_{j_k}^{(k+1)} [\|x_k - z\|_\infty - \rho_k] \\ &\geq \rho_k + \lambda_{j_k}^{(k+1)} \underbrace{[d(z, M) - \rho_k]}_{>0} \\ &> \rho_k. \end{aligned}$$

Als monoton wachsende, nach oben durch $d(z, M)$ beschränkte Folge ist $\{\rho_k\}$ konvergent und folglich $\lim_{k\to\infty}(\rho_{k+1} - \rho_k) = 0$. *Angenommen* es existiert eine Konstante $c > 0$ mit $\lambda_j^{(k)} \geq c$ für alle $j \in \{0, \ldots, n+1\}$ und $k = 1, 2, \ldots$. Dann wäre

$$\rho_{k+1} \geq \rho_k + c[d(z, M) - \rho_k]$$

und folglich

$$d(z, M) - \rho_{k+1} \leq \underbrace{(1 - c)}_{=:q}[d(z, M) - \rho_k]$$

und der erste Teil des Konvergenzsatzes wäre bewiesen. Nun ist $0 \leq \rho_0 < \rho_1 < \cdots$ und daher (Lemma 4.3.3) ist

$$K_{\rho_1} := \{T \in \mathcal{T} : \rho_T \geq \rho_1\}$$

kompakt. Für alle $k \geq 1$ ist $T_k \in K_{\rho_1}$. Da $\lambda \colon \mathcal{T} \longrightarrow \mathbb{R}^{n+2}$ stetig, jede Komponente von $\lambda(T)$, $T \in \mathcal{T}$ positiv und K_{ρ_1} kompakt ist, existiert ein $c > 0$ mit $\lambda_j^{(k)} \geq c$ für $j \in \{0, \ldots, n+1\}$ und $k = 1, 2, \ldots$. Damit ist die lineare Konvergenz von $\{\rho_k\}$ gegen $d(z, M)$ bewiesen.

Die Abbildung $a \colon \mathcal{T} \longrightarrow \mathbb{R}^{n+1}$ ist stetig auf der kompakten Menge K_{ρ_1}, folglich ist $\{a(T_k)\}$ und damit auch die Folge $\{x_k\} = \{\sum_{i=0}^{n} a_i^{(k)} x_i\}$ beschränkt. Daher ist aus $\{x_k\} \subset M$ eine konvergente Teilfolge auswählbar. Wir überlegen uns, dass *jede* konvergente Teilfolge $\{x_{k_j}\} \subset \{x_k\}$ notwendig die eindeutige beste Approximierende $x^* \in M$ an z in M als Limes besitzt. Denn angenommen, $\tilde{x} \in M$ (als endlichdimensionaler linearer Teilraum ist M abgeschlossen) sei Limes der Folge $\{x_{k_j}\}$.Wegen $\rho_k \to d(z, M)$ gilt insbesondere $\rho_{k_j} \to d(z, M)$. Nun gilt aber

$$(*) \qquad \lim_{k\to\infty} (\|x_k - z\|_\infty - \rho_k) = 0,$$

woraus dann $\|\tilde{x} - z\|_\infty - d(z, M) = 0$, wegen der Eindeutigkeit der besten Approximierenden also $x_{k_j} \to x^*$ folgt. Zum Nachweis von $(*)$ beachten wir, dass mit der obigen Konstanten $c > 0$ gilt:

$$\begin{aligned} c(\|x_k - z\|_\infty - \rho_k) &\leq \lambda_{j_k}^{(k+1)}(\|x_k - z\|_\infty - \rho_k) \\ &\leq \rho_{k+1} - \rho_k \\ &\to 0. \end{aligned}$$

Damit ist gezeigt, dass jede konvergente Teilfolge von $\{x_k\}$ denselben Limes, nämlich x^* besitzt. Hieraus folgt aber sehr einfach die Konvergenz der *gesamten* Folge $\{x_k\}$ gegen x^*, womit dann auch Teil (b) des Konvergenzsatzes bewiesen ist. □

Mit dem letzten Satz wurde unter verhältnismäßig allgemeinen Austauschregeln für die Referenzen ein globaler Konvergenzsatz für das Remez-Verfahren zur Lösung linearer T-Approximationsaufgaben bezüglich Haarscher Teilräume bewiesen.

Jetzt wollen wir auf Austauschregeln eingehen, die die Bedinungen 1.–3. des allgemeinen Remez-Verfahrens erfüllen. Zunächst geben wir die Regeln noch einmal an. Sei $T_k = (t_0^{(k)}, \ldots, t_{n+1}^{(k)}) \in \mathcal{T}$. Ferner sei $x_k := x_{T_k}^*$ die Lösung von

$$(\mathrm{P_k}) \qquad \text{Minimiere} \quad \|x - z\|_{T_k,\infty} = \max_{t \in T_k} |x(t) - z(t)| \quad \text{auf} \quad M$$

und $\rho_k := \|x_k - z\|_{T_k,\infty}$. Zu bestimmen ist eine neue Referenz $T_{k+1} = (t_0^{(k+1)}, \ldots, t_{n+1}^{(k+1)})$ mit

1. $|x_k(t_j^{(k+1)}) - z(t_j^{(k+1)})| \geq \rho_k$, $j = 0, \ldots, n+1$.
2. Es existiert $j_k \in \{0, \ldots, n+1\}$ mit $|x_k(t_{j_k}^{(l+1)}) - z(t_{j_k}^{(k+1)})| = \|x_k - z\|_\infty$.
3. $\operatorname{sign}(x_k(t_j^{(k+1)}) - z(t_j^{(k+1)})| = -\operatorname{sign}(x_k(t_{j+1}^{(k+1)}) - z(t_{j+1}^{(k+1)}))$, $j = 0, \ldots, n$. D. h. der Defekt $x_k - z$ alterniert nicht nur in den Punkten der alten Referenz T_k sondern auch in denen der neuen Referenz T_{k+1}.

Zunächst betrachten wir den 1. Remez-Algorithmus mit *Ein-Punkt-Austausch*. Hierbei wird ein beliebiger Punkt $s^* \in [\alpha, \beta]$ mit $|x_k(s^*) - z(s^*)| = \|x_k - z\|_\infty$ zu T_k hinzugenommen und dafür ein gewisser anderer Punkt aus T_k herausgeworfen. Welcher Punkt dies ist, hängt von der Lage von s^* ab. muss natürlich gewährleistet sein, dass der Defekt $x_k - z$ auch in den Punkten der neuen Referenz im Vorzeichen alterniert. Wir definieren $\sigma^* := \operatorname{sign}(x_k(s^*) - z(s^*))$ und unterscheiden drei Fälle, nämlich ob s^* links von $t_0^{(k)}$, zwischen $t_0^{(k)}$ und $t_{n+1}^{(k+1)}$ oder rechts von $t_{n+1}^{(k)}$ liegt.

1. Sei $s^* \in [\alpha, t_0^{(k)})$.

   Wir setzen

$$T_{k+1} := \begin{cases} (s^*, t_1^{(k)}, \ldots, t_{n+1}^{(k)}), & \sigma^* = \operatorname{sign}(x_k(t_0^{(k)}) - z(t_0^{(k)}), \\ (s^*, t_0^{(k)}, \ldots t_n^{(k)}), & \sigma^* \neq \operatorname{sign}(x_k(t_0^{(k)}) - z(t_0^{(k)}). \end{cases}$$

2. Sei $s^* \in (t_j^{(k)}, t_{j+1}^{(k)})$.

   Wir setzen

$$T_{k+1} := \begin{cases} (t_0^{(k)}, \ldots, t_{j-1}^{(k)}, s^*, t_{j+1}^{(k)}, \ldots, t_{n+1}^{(k)}), & \sigma^* = \operatorname{sign}(x_k(t_j^{(k)}) - z(t_j^{(k)})), \\ (t_0^{(k)}, \ldots, t_j^{(k)}, s^*, t_{j+2}^{(k)}, \ldots, t_{n+1}^{(k)}), & \sigma^* \neq \operatorname{sign}(x_k(t_j^{(k)}) - z(t_j^{(k)})). \end{cases}$$

3. Sei $s^* \in (t_{n+1}^{(k)}, \beta]$.

   Wir setzen

$$T_{k+1} := \begin{cases} (t_0^{(k)}, t_1^{(k)}, \ldots, t_n^{(k)}, s^*), & \sigma^* = \operatorname{sign}(x_k(t_{n+1}^{(k)}) - z(t_{n+1}^{(k)})), \\ (t_1^{(k)}, t_2^{(k)}, \ldots, t_{n+1}^{(k)}, s^*), & \sigma^* \neq \operatorname{sign}(x_k(t_{n+1}^{(k)}) - z(t_{n+1}^{(k)})). \end{cases}$$

Bemerkungen: 1. Man sollte möglichst davon Gebrauch machen, dass sich bei dem Gleichungssystem

$$\sum_{i=0}^{n} a_i x_i(t_j) + (-1)^j \rho = z(t_j), \qquad j = 0, \ldots, n+1,$$

von Schritt zu Schritt nur eine Zeile ändert.

2. Approximiert man bezüglich $M = \Pi_n$ auf dem Intervall $[\alpha, \beta]$, so sollte man die Ausgangs-Referenz $T_0 = (t_0^{(0)}, \ldots, t_{n+1}^{(0)})$ mit

$$t_j^{(0)} := \frac{1}{2}(\alpha + \beta) + \frac{1}{2}(\beta - \alpha) \cos \frac{n+1-j}{n+1}\pi, \qquad j = 0, \ldots, n+1,$$

wählen , also die Extremstellen des $(n+1)$-te T-Polynoms. Die Begründung hierfür liefert Satz 4.2.3, denn dieser impliziert, dass das Remez-Verfahren für $z \in \Pi_{n+1}$ mit dieser Ausgangs-Referenz in einem Schritt fertig ist.

3. Es ist zweckmäßig, als Basis von Π_n die T-Polynome T_i, $i = 0, \ldots, n$, zu nehmen. Die Auswertung von

$$p_n(t) = \frac{a_0}{2} + \sum_{i=1} a_i T_i(t)$$

bei gegebenen $a_0, \ldots, a_n$ und t sollte nach dem folgenden "Horner-ähnlichen" Verfahren erfolgen.

- Setze $b_{n+1} := 0$, $b_n := a_n$.
- Für $i = n-1, n-2, \ldots, 0$:
  - Berechne $b_i := 2tb_{i+1} - b_{i+2} + a_i$.
- Es ist $p_n(t) = (b_0 - b_2)/2$.

Zur Begründung beachte man, dass

$$\begin{aligned} p_n(t) &= \frac{a_0}{2} + \sum_{i=0}^{n-1}(b_i - 2tb_{i+1} + b_{i+2})T_i(t) + b_n T_n(t) \\ &= \frac{a_0}{2} + \sum_{i=2}^{n} b_i \underbrace{(T_i(t) - 2tT_{i-1}(t) + T_{i-2}(t))}_{=0} + b_1 T_1(t) - b_0 T_0(t) \\ &= \frac{1}{2}(b_0 - 2tb_1 + b_2) + b_1 t - b_2 \\ &= \frac{1}{2}(b_0 - b_2), \end{aligned}$$

wobei wir die Rekursionsformel in Lemma 4.2.1 benutzt haben. □

Beispiel: Sei $[\alpha, \beta] = [0, \pi/2]$, $M := \Pi_3$ und $z(t) := \sin(t)$. Wir nehmen als Ausgangsreferenz

$$T_0 = (t_0^{(0)}, t_1^{(0)}, t_2^{(0)}, t_3^{(0)}, t_4^{(0)}) := (0, 0.2300, 0.7854, 1.3408, 1.5708)$$

die Extremstellen des T-Polynoms T_4. Der Defekt $x_0 - z$ ist in Abbildung 4.7 dargestellt. Es ist $\rho_0 = 0.0013586698$. In Abbildung 4.7 erkennt man, dass man schon eine ziemlich

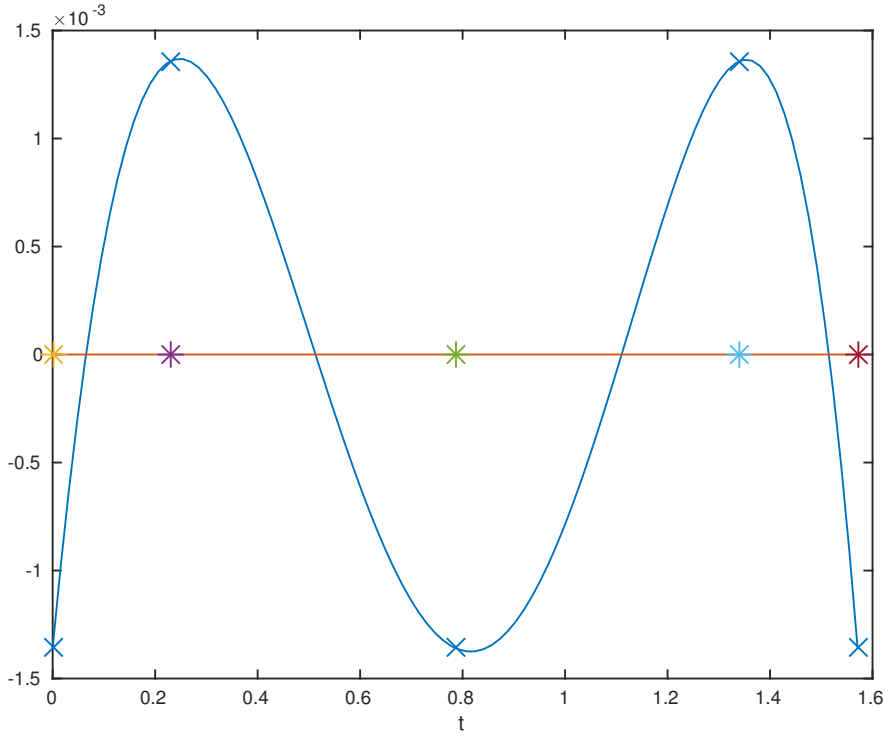

Abbildung 4.7: Erster Schritt des Remez-Verfahrens. Der Defekt $x_0 - z$.

gute Näherung erhalten hat, da ρ_0 nur unwesentlich kleiner als $\|x_0 - z\|_\infty$ ist. Nun

müssen wir ein $s^* \in [\alpha, \beta]$ mit $|x_0(s^*) - z(s^*)| = \|x_0 - z\|_\infty$ bestimmen. Wir suchen in der Nähe der inneren Referentpunkte von T_0 und sehen uns den Defekt $x_0 - z$ dort wie mit einer Lupe an. Dies geschieht in Abbildung 4.8. Man stellt fest, dass man

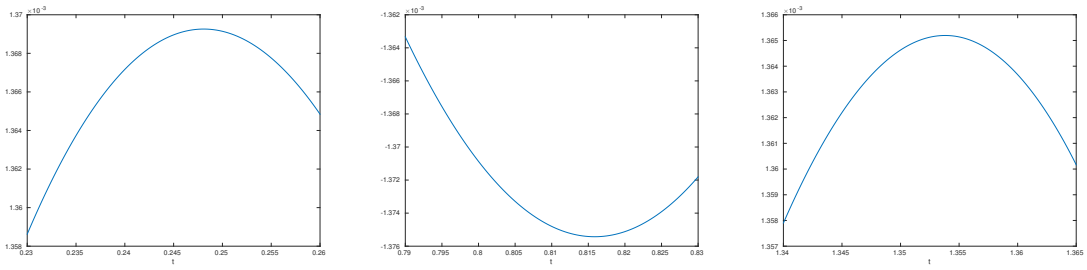

Abbildung 4.8: Der Defekt in der Nähe innerer Referenzpunkte von T_0

$s^* = 0.8160$ wählen kann, so dass man als neue Referenz

$$T_1 = (t_0^{(1)}, t_1^{(1)}, t_2^{(1)}, t_3^{(1)}, t_4^{(1)}) := (0, 0.2300, 0.8160, 1.3408, 1.5708)$$

erhält. Den Defekt $x_1 - z$ geben wir in Abbildung 4.9 an. Es ist $\rho_1 = 0.0013628656$ und

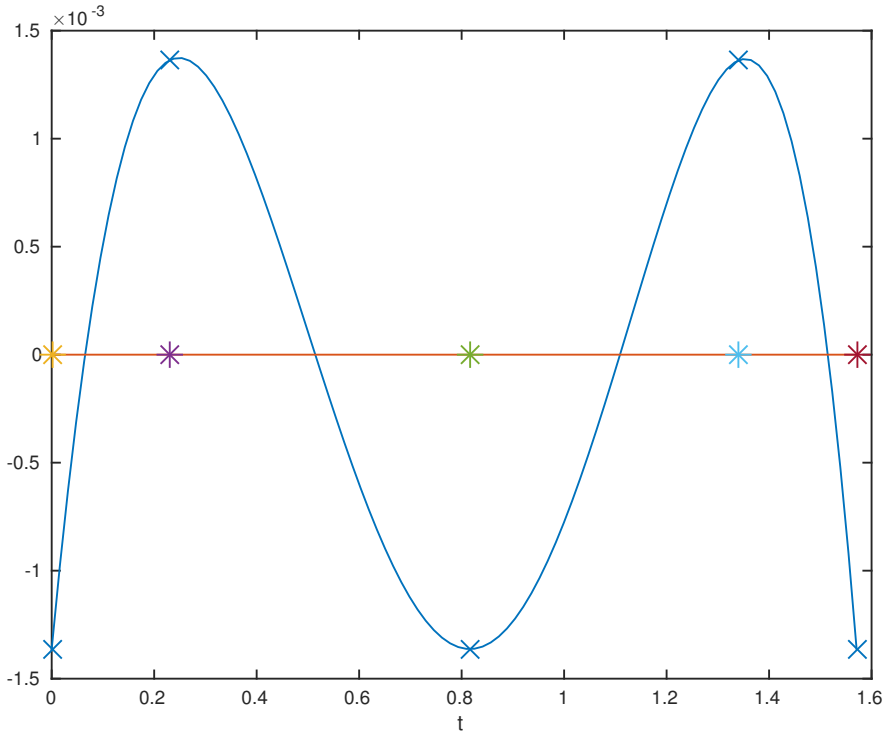

Abbildung 4.9: Der Defekt $x_1 - z$

$$x_2(t) = -0.0014 + 1.0253t - 0.0707t^2 - 0.1125t^3$$

eine Näherung an die beste Approximierende an z in Π_3 auf dem Intervall $[0, \pi/2]$. $\square$

Beim *Simultan-Austausch* bzw. dem 2. Remez-Verfahren wählt man die neue Referenz $T_{k+1} = (t_0^{(k+1)}, \ldots, t_{n+1}^{(k+1)})$ so, dass bei $t_j^{(k+1)}$ ein lokales Extremum von $x_k - z$ bzw. ein lokales Maximum von $|x_k - z|$ liegt. Man muss sich nur überlegen, dass dies so möglich ist, dass die Forderungen 1.–3. an die neue Referenz T_{k+1} erfüllt sind. Hierauf wollen wir aber nicht näher eingehen.

4.4 Diskrete lineare T-Approximation

In diesem Abschnitt wollen wir kurz gesondert auf die diskrete T-Approximation eingehen. Zwar sind fast alle Ergebnisse schon in Abschnitt 4.1 enthalten, aber es lohnt sich hoffentlich trotzdem, diesen Fall noch einmal gesondert zu betrachten. Als Literatur sei G. A. Watson (1980, S. 25 ff.) empfohlen.

Sei $B = \{t_1, \ldots, t_k\} \subset \mathbb{R}^N$. Eine Approximationsaufgabe in $C(B)$ ist eigentlich eine Aufgabe im $\mathbb{R}^k$, da $z \in C(B)$ durch den Vektor $z = (z(t_1), \ldots, z(t_k))^T$ gegeben ist. Ist $M = \operatorname{span}\{x_1, \ldots, x_n\} \subset C(B)$ ein n-dimensionaler linearer Teilraum, so können $x_1, \ldots, x_n$ als Vektoren des $\mathbb{R}^k$ aufgefasst werden und daher ist notwendig $n \leq k$. Das diskrete T-Approximationsproblem besteht darin, bei gegebenen $z \in \mathbb{R}^k$, $X = (\, x_1 \ \cdots \ x_n \,) \in \mathbb{R}^{k \times n}$ mit $\operatorname{Rang}(X) = n$ (und damit $n \leq k$) die Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(a) := \|Xa - z\|_\infty, \quad a \in \mathbb{R}^n,$$

zu lösen. Ist z. B. $N = 1$, $M = \Pi_{n-1} = \operatorname{span}\{1, t, \ldots, t^{n-1}\}$, so ist

$$X = \begin{pmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_k & t_k^2 & \cdots & t_k^{n-1} \end{pmatrix}.$$

Die *Existenz* mindestens einer Lösung von (P) ist gesichert. Um etwas über die *Charakterisierung* einer Lösung a^* von (P) auszusagen, definieren wir für gegebenes $a \in \mathbb{R}^n$ die Indexmenge

$$I(a) := \{i \in \{1, \ldots, k\} : |(Xa)_i - z_i| = \|Xa - z\|_\infty\}.$$

Das *Kolmogoroff-Kriterium* (siehe Satz 4.1.1) besagt:

- *Ein $a^* \in \mathbb{R}^n$ ist genau dann eine Lösung von (P), wenn*

$$\max_{i \in I(a^*)} \operatorname{sign}\left((Xa^*)_i - z_i\right) (Xa)_i \geq 0 \quad \textit{für alle } a \in \mathbb{R}^n.$$

Auch die Charakterisierungsaussage in Satz 4.1.2 kann übertragen werden:

- *Ein $a^* \in \mathbb{R}^n$ ist genau dann eine Lösung von (P), wenn*

$$(*) \qquad 0 \in \operatorname{co}\left(\{\operatorname{sign}\left((Xa^*)_i - z_i\right) \begin{pmatrix} x_{i1} \\ \vdots \\ x_{in} \end{pmatrix} : i \in I(a^*)\}\right),$$

wobei $x_j = (x_{1j}, \ldots, x_{nj})^T$ die j-te Spalte von X ist, $j = 1, \ldots, n$.

Diese Aussage kann geringfügig umformuliert werden. Denn wegen des Satzes von Carathéodory (Satz 2.2.13) gilt $(*)$ genau dann, wenn eine Indexmenge $I^* \subset I(a^*)$ mit $|I^*| \leq n+1$ sowie $\lambda_i > 0$, $i \in I^*$, mit

$$0 = \sum_{i \in I^*} \lambda_i \operatorname{sign}\left((Xa^*)_i - z_i\right) x_{ij}, \quad j = 1, \ldots, n, \qquad 1 = \sum_{i \in I^*} \lambda_i.$$

Definiert man $\lambda = (\lambda_i) \in \mathbb{R}^k$, indem man $\lambda_i = 0$ für $i \notin I^*$ setzt, definiert man ferner $\mu = (\mu_i) \in \mathbb{R}^k$ durch

$$\mu_i := \lambda_i \operatorname{sign}\left((Xa^*)_i - z_i\right), \quad i = 1, \ldots, k,$$

so erhält man (siehe G. A. Watson (1980, S. 27)):

- Ein $a^* \in \mathbb{R}^n$ ist genau dann eine Lösung von (P), wenn eine Indexmenge $I^* \subset I(a^*)$ mit $|I^*| \leq n+1$ und ein Vektor $\mu \in \mathbb{R}^k \setminus \{0\}$ mit $\mu_i = 0$, $i \notin I^*$, $X^T\mu = 0$ und $\mu_i \operatorname{sign}\left((Xa^*)_i - z_i\right) \geq 0$, $i \in I^*$, existiert.

Die Übertragung des Begriffes "n-dimensionaler Haarscher Teilraum von $C(B)$" ist fast offensichtlich:

- Die Matrix $X \in \mathbb{R}^{k \times n}$ genügt der Haarschen Bedingung, falls jede $n \times n$-Untermatrix von X nichtsingulär ist.

Offenbar genügt $X \in \mathbb{R}^{k \times n}$ der Haarschen Bedingung genau dann, wenn jedes nichttriviale Element von $M := \{Xa : a \in \mathbb{R}^n\}$ höchstens $n-1$ verschwindende Komponenten besitzt bzw. M ein n-dimensionaler Haarscher Teilraum des $\mathbb{R}^k$ ist. Es gilt dann (siehe Satz 4.1.4 bzw. G. A. Watson (1980, S. 32)):

- *Genügt X der Haarschen Bedingung, so ist eine Lösung a^* von (P) eindeutig.*

Denn: Wir können $k \geq n+1$ annehmen, da X für $k = n$ nichtsingulär und damit $a^* = X^{-1}z$ die eindeutige Lösung von (P) ist. Sei $a^* \in \mathbb{R}^n$ eine Lösung von (P). Wie wir gerade eben gesehen haben, existiert eine Indexmenge $I^* \subset I(a^*)$ mit $|I^*| \leq n+1$ und ein Vektor $\mu \in \mathbb{R}^k \setminus \{0\}$ mit

$$\mu_i = 0, \quad i \notin I^*, \quad i \notin I^*, \qquad X^T\mu = 0, \qquad \mu_i \operatorname{sign}\left((Xa^*)_i - z_i\right) > 0, \quad i \in I^*.$$

Da X der Haarschen Bedingung genügt, ist $|I^*| = n+1$. Denn andernfalls könnte I^* (notfalls) zu einer n-elementigen Indexmenge aus $\{1, \ldots, k\}$ ergänzt werden. Die aus diesen Zeilen von X gebildete $n \times n$-Matrix wäre wegen der Haarschen Bedingung nichtsingulär und insbesondere die zu I^* gehörenden Zeilen von X linear unabhängig, ein Widerspruch dazu, dass der Nullvektor eine nichttriviale Linearkombination gerade dieser Zeilen ist. Nun sei $a^{**} \in \mathbb{R}^n$ ebenfalls eine Lösung von (P). Für $i \in I^*$ ist

$$\begin{aligned} \mu_i \left((Xa^*)_i - z_i\right) &= \mu_i \operatorname{sign}\left((Xa^*)_i - z_i\right) \left|(Xa^*)_i - z_i\right| \\ &= \underbrace{\mu_i \operatorname{sign}\left((Xa^*)_i - z_i\right)}_{>0} \|Xa^* - z\|_\infty, \quad i \in I^*. \end{aligned}$$

Daher ist

$$\begin{aligned}
\|Xa^* - z\|_\infty \sum_{i\in I^*} |\mu_i| &= \left|\sum_{i\in I^*} \mu_i \left((Xa^*)_i - z_i\right)\right| \\
&= \left|\sum_{i\in I^*} \mu_i z_i\right| \\
&= \left|\sum_{i\in I^*} \mu_i((Xa^{**})_i - z_i)\right| \\
&\leq \sum_{i\in I^*} |\mu_i|\,|(Xa^{**})_i - z_i| \\
&\leq \|Xa^{**} - z\|_\infty \sum_{i\in I^*} |\mu_i| \\
&= \|Xa^* - z\|_\infty \sum_{i\in I^*} |\mu_i|.
\end{aligned}$$

Also gilt in dieser Gleichheits-Ungleichungskette durchgehend das Gleichheitszeichen und folglich $(Xa^*) - z_i = (Xa^{**})_i - z_i$, $i \in I^*$, bzw. $(X(a^* - a^{**}))_i = 0$, $i \in I^*$. Da X der Haarschen Bedingung genügt, folgt $a^* = a^{**}$ und das ist der Beweis für die behauptete Eindeutigkeitsaussage.

Weiter gilt:

- *Genügt X der Haarschen Bedingung, so ist die Lösung a^* von (P) stark eindeutig, d. h. es gibt eine Konstante $\gamma > 0$ mit*

$$\|Xa - z\|_\infty \geq \|Xa^* - z\|_\infty + \gamma\,\|a - a^*\| \quad \textit{für alle } a \in \mathbb{R}^n.$$

Denn: Wegen Satz 4.1.7 gibt es eine Konstante $c > 0$ mit n

$$\|Xa - z\|_\infty \geq \|Xa^* - z\|_\infty + c\,\|Xa - Xa^*\|_\infty \quad \text{für alle } a \in \mathbb{R}^n.$$

Da Rang $(X) = n$ ist $\delta := \min_{\|b\|_\infty = 1} \|Xb\|_\infty > 0$ und folglich

$$\|Xa - Xa^*\|_\infty \geq \|a - a^*\|_\infty \quad \text{für alle } a \in \mathbb{R}^n.$$

Mit $\gamma := c\delta$ folgt die Behauptung.

Bemerkung: Aus der Eindeutigkeit einer Lösung von (P) folgt ihre starke Eindeutigkeit, die Haarsche Bedingung muss hierbei *nicht* erfüllt sein (siehe G. A. Watson (1980, S. 34)). Denn sei $a^* \in \mathbb{R}^n$ eine (eindeutige) Lösung von (P). Offenbar[6] können wir o. B. d. A. annehmen, dass $\|Xa^* - z\| > 0$. Man definiere

$$\gamma := \min_{\|c\|_\infty = 1} \max_{i\in I(a^*)} \operatorname{sign}\left((Xa^*)_i - z_i\right)(Xc)_i.$$

[6] Aus der Eindeutigkeit einer Lösung von (P) folgt die lineare Unabhängigkeit der Spalten von X bzw. Rang $(X) = n$. Wie oben ist $\delta := \min_{\|b\|_\infty = 1} \|Xb\|_\infty > 0$ und folglich

$$\|Xa - z\|_\infty = \|X(a - a^*)\|_\infty \geq \underbrace{\|Xa^* - z\|_\infty}_{=0} + \delta\,\|a - a^*\|_\infty \quad \text{für alle } a \in \mathbb{R}^n$$

und das ist die starke Eindeutigkeit von a^*.

Wir wollen uns überlegen, dass $\gamma > 0$. Angenommen, es wäre $\gamma \leq 0$. Dann existiert ein $c^* \in \mathbb{R}^n \setminus \{0\}$ mit

$$\operatorname{sign}\left((Xa^*)_i - z_i\right)(Xc^*)_i \leq 0, \quad i \in I(a^*).$$

Wir wollen uns überlegen, dass $\|X(a^* + tc^*) - z\|_\infty \leq \|Xa^* - z\|_\infty$ für alle hinreichend kleinen $t > 0$ ist, was dann ein Widerspruch dazu ist, dass a^* *eindeutige* Lösung von (P) ist. Zu zeigen ist, dass für alle $i = 1, \ldots, k$ gilt:

$$|(X(a^* + tc^*))_i - z_i| \leq \|Xa^* - z\|_\infty \quad \text{für alle hinreichend kleinen } t > 0.$$

Dies ist für $i \in \{1, \ldots, k\} \setminus I(a^*)$ trivialerweise richtig. Sei daher $i \in I(a^*)$. Für alle hinreichend kleinen $t > 0$ ist dann

$$\begin{aligned}
|(X(a^* + tc^*))_i - z_i| &= |(Xa^*)_i - z_i + t(Xc^*)_i| \\
&= |\operatorname{sign}\left((Xa^*)_i - z_i\right)\|Xa^* - z\|_\infty + t(Xc^*)_i| \\
&= |\underbrace{\|Xa^* - z\|_\infty}_{>0} + t\operatorname{sign}\left((Xa^*)_i - z_i\right)(Xc^*)_i| \\
&= \|Xa^* - z\|_\infty + t\underbrace{\operatorname{sign}\left((Xa^*)_i - z_i\right)(Xc^*)_i}_{\leq 0} \\
&\leq \|Xa^* - z\|_\infty.
\end{aligned}$$

Durch einen Widerspruchsbeweis haben wir damit $\gamma > 0$ nachgewiesen. Nun können wir die starke Eindeutigkeit der eindeutigen Lösung a^* nachweisen. Sei $a \in \mathbb{R}^n$ beliebig. Wir zeigen, dass

$$\|Xa - z\|_\infty \geq \|Xa^* - z\|_\infty + \gamma\|a - a^*\|_\infty.$$

O. b. d. A. ist $a \neq a^*$. Man definiere

$$c := \frac{a - a^*}{\|a - a^*\|_\infty}.$$

Nach Definition von γ ist

$$\gamma \leq \max_{i \in I(a^*)} \operatorname{sign}\left((Xa^*)_i - z_i\right)(Xc)_i,$$

so dass ein $i \in I(a^*)$ mit $\gamma \leq \operatorname{sign}\left((Xa^*)_i - z_i\right)(Xc)_i$ existiert. Dann ist aber

$$\begin{aligned}
\|Xa - z\|_\infty &\geq \operatorname{sign}\left((Xa^*)_i - z_i\right)\left((Xa)_i - z_i\right) \\
&= \operatorname{sign}\left((Xa^*)_i - z_i\right)\left((Xa^*)_i - z_i\right) + \operatorname{sign}\left((Xa^*)_i - z_i\right)(X(a - a^*))_i \\
&= |(Xa^*)_i - z_i| + \operatorname{sign}\left((Xa^*)_i - z_i\right)(Xc)_i\|a - a^*\|_\infty \\
&= \|Xa^* - z\|_\infty + \operatorname{sign}\left((Xa^*)_i - z_i\right)(Xc)_i\|a - a^*\|_\infty \\
&\geq \|Xa^* - z\|_\infty + \gamma\|a - a^*\|_\infty
\end{aligned}$$

und das ist die starke Eindeutigkeit von a^*. □

Einige wenige Bemerkungen wollen wir noch zur *numerischen Behandlung* des diskreten linearen T-Approximationsproblems (P) machen. Weitergehende Aussagen werden

bei G. A. WATSON (1980, S. 34 ff.). Wir wollen nur einige wenige Bemerkungen zum Zusammenhang von (P) mit linearen Optimierungsaufgaben machen.

Die diskrete T-Approximaionsaufgabe (P) ist äquivalent zu

$$\text{Minimiere} \quad \delta \quad \text{auf} \quad M := \{(a, \delta) \in \mathbb{R}^n \times \mathbb{R} : \|Xa - z\|_\infty \leq \delta\}.$$

Diese Aufgabe wiederum kann als eine lineare Optimierungsaufgabe geschrieben werden:

$$\begin{cases} \text{Minimiere} \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix}^T \begin{pmatrix} a \\ \delta \end{pmatrix} \quad \text{unter der Nebenbedingung} \\ \qquad\qquad \begin{pmatrix} X & -e \\ -X & -e \end{pmatrix} \begin{pmatrix} a \\ \delta \end{pmatrix} \leq \begin{pmatrix} z \\ -z \end{pmatrix}, \end{cases}$$

wobei $e := (1, \ldots, 1)^T \in \mathbb{R}^k$.

Wenn man die Möglichkeit dazu hat und ein konkretes diskretes T-Approximationsproblem lösen will, so kann man die Optimization Toolbox von MATLAB benutzen. Dies wollen wir anhand zweier Beispiele demonstrieren.

Beispiele: 1. Zu lösen sei ein diskretes lineares T-Approximationsproblem mit (siehe G. A. WATSON (1980, S. 36))

$$X := \begin{pmatrix} 1 & 2 \\ 2 & 3 \\ 2 & 4 \\ 1 & 0 \end{pmatrix}, \qquad z := \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

Das folgende MATLAB-Programm

```
X=[1 2;2 3;2 4;1 0];z=ones(4,1);e=ones(4,1);
A=[X -e;-X -e];b=[z;-z];c=[zeros(2,1);1];
x=linprog(c,A,b);a=x(1:2);delta=a(3);
```

liefert das verhältnismäßig schlechte Ergebnis

$$a = \begin{pmatrix} 0.666666666502920 \\ 0.000000001707576 \end{pmatrix}, \qquad \delta = 0.333333331748690.$$

Dies ist darauf zurückzuführen, dass `linprog` per default eine Innere-Punkt-Methode (was immer das ist, darauf wollen wir nicht eingehen) benutzt, eine Methode, welche für sehr hochdimensionale Probleme (was obige Aufgabe nicht ist) anderen Methoden vorzuziehen ist. Zwingt man `linprog`, das Simplex- bzw. das duale Simplexverfahren zu benutzen, so sind die Ergebnisse sehr viel besser. So ergibt

```
X=[1 2;2 3;2 4;1 0];z=ones(4,1);e=ones(4,1);
A=[X -e;-X -e];b=[z;-z];c=[zeros(2,1);1];
options=optimoptions(@linprog,'Algorithm','dual-simplex');
x=linprog(c,A,b,[],[],[],[],[],options); a=x(1:2);delta=x(3);
```

das Ergebnis

$$a = \begin{pmatrix} 0.666666666666667 \\ 0 \end{pmatrix}, \qquad \delta = 0.333333333333333.$$

Statt `'dual-simplex'` hätte man auch `'simplex'` eingeben können, wobei MATLAB darüber informiert, dass diese Funktion künftig wegfallen könnte.

2. Die Funktion $f(t) := \cos((\pi/2)t)$ soll im Intervall $[0,1]$ durch ein Polynom vierten Grades im Tschebyscheffschen Sinne approximiert werden bezüglich der elf äquidistanten Abszissen $t_i := (i-1)/10$, $i = 1, \ldots, 11$. Gesucht ist also eine Lösung der Aufgabe

$$\text{Minimiere} \quad f(a) := \max_{i=1,\ldots,11} \left| \sum_{j=1}^{5} a_j t_i^{j-1} - \cos((\pi/2)t_i) \right|, \quad a \in \mathbb{R}^5.$$

Mit $X := (t_i^{j-1}) \in \mathbb{R}^{11\times 5}$ und $z := (\cos((\pi/2)t_i)) \in \mathbb{R}^{11}$ ordnet sich dieses Problem der obigen Aufgabenstellung unter. Wir berechnen die Lösung und plotten den Defekt zwischen dem gewonnenen Polynom vierten Grades und der zu approximierenden Funktion f. Hierbei benutzen wir wieder MATLAB sowie die Funktion `linprog`. Als Ergebnis von

```
i=(1:11)';t=(i-1)/10;z=cos((pi*t)/2);e=ones(11,1);X=e;
for j=1:4
   e=e.*t;
   X=[X e];
end;
e=ones(11,1);
A=[X -e;-X -e];b=[z;-z];c=[zeros(5,1);1];
options=optimoptions(@linprog,'Algorithm','dual-simplex');
x=linprog(c,A,b,[],[],[],[],[],options);
a=x(1:5);delta=x(6);a=a(5:-1:1);
s=linspace(0,1);defekt=polyval(a,s)-cos((pi*s)/2);
plot(s,defekt,'r');hold on;
disdef=polyval(a,t)-z;plot(t,disdef,'b*','MarkerSize',12);
plot(t,0*t);
hold off
```

erhalten wir (Koeffizienten in aufsteigender Reihenfolge! Eine Umkehrung der Koeffizienten im obigen Programm haben wir nur durchgeführt, um `polyval` anwenden zu können)

$$a = \begin{pmatrix} 0.999896084731383 \\ 0.004960107066475 \\ -1.271044520660073 \\ 0.093307119901566 \\ 0.172985124229266 \end{pmatrix}, \qquad \delta = 1.039152686165926e-04.$$

In Abbildung 4.10 haben wir den Defekt gezeichnet. □

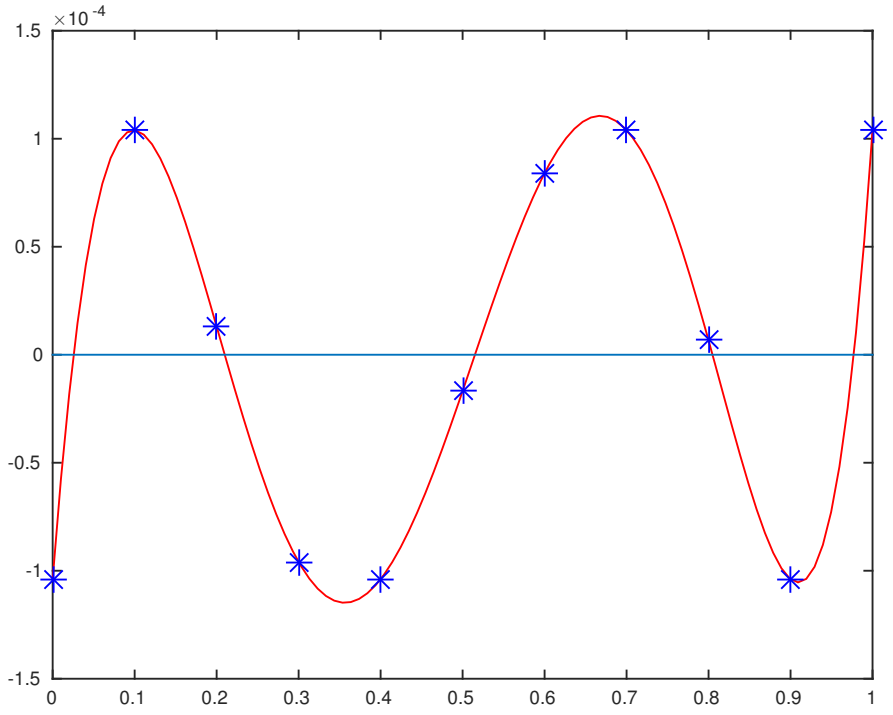

Abbildung 4.10: Defekt bei diskreter T-Approximation

Die obige zu der diskreten T-Approximationsaufgabe äquivalente lineare Optimierungsaufgabe hat nicht die Simplex-Normalform. Durch Übergang zur dualen linearen Optimierungsaufgabe kann dies weitgehend erreicht werden. Weiter wollen wir auf die Anwendung von Methoden der linearen Optimierung auf die Aufgabe (P) nicht eingehen.

Kapitel 5

Rationale T-Approximation

5.1 Existenz, Eindeutigkeit und Charakterisierung einer besten Approximierenden

In diesem Kapitel soll die Approximation stetiger Funktionen durch rationale Funktionen bezüglich der Maximumnorm untersucht werdeb. Rationale Funktionen bieten sich an, da ihre Funktionswerte alleine durch die Grundrechenarten berechnet werden können und da man zu Recht hofft, dass eine Approximation mittels rationaler Funktionen gegenüber der mit Polynomen für gewisse Funktionenklassen wesentliche Genauigkeitsverbesserungen mit sich bringt.

Zunächst formulieren wir ein Ergebnis, das wir in Form einer Bemerkung im Anschluss an Lemma 3.2.1 schon bewiesen haben.

Lemma 5.1.1 *Sei $B \subset \mathbb{R}^N$ kompakt, $(C(B), \|\cdot\|_\infty)$ der Banachraum der auf B definierten reellwertigen stetigen Funktionen versehen mit der Maximumnorm $\|\cdot\|_\infty$, definiert durch*

$$\|x\|_\infty := \max_{t\in B} |x(t)|.$$

Seien $P, Q \subset C(B)$ lineare Teilräume und

$$Q_+ := \{q \in Q : q(t) > 0 \text{ für alle } t \in B\} \neq \emptyset.$$

Schließlich sei

$$R := \left\{ \frac{p}{q} : p \in P,\ q \in Q_+ \right\} \subset C(B)$$

die zugehörige Menge der verallgemeinerten rationalen Funktionen. Ist dann $p^*/q^* \in R$ eine lokal beste Approximierende an ein $z \in C(B)$ in R, so ist p^*/q^* auch eine global beste Approximierende an z in R.

Bemerkungen: 1. Für welche endlichdimensionalen linearen Teilräume $Q \subset C(B)$ ist die Menge Q_+, definiert durch

$$Q_+ := \{q \in Q : q(t) > 0 \ \text{ für alle } t \in B\},$$

nichtleer? Es gilt:

- *Ist $Q \subset C[\alpha,\beta]$ ein n-dimensionaler Haarscher Teilraum, so ist $Q_+ \neq \emptyset$.*

Denn: Sei $z := 1$ die Funktion, die auf $[\alpha,\beta]$ konstant gleich 1 ist und $x^* \in Q$ die (eindeutige) beste Approximierende an z in Q. Es ist $x^* \neq 0$, denn $x^* - z$ muss ja im Vorzeichen alternieren. Für alle $t \in [\alpha,\beta]$ ist also

$$|x^*(t) - \underbrace{z(t)}_{=1}| \leq \|x^* - z\|_\infty < \|0 - 1\|_\infty = 1,$$

also $-1 < x^*(t) - 1 < 1$ und damit $0 < x^*(t)$ für alle $t \in [\alpha,\beta]$, also $x^* \in Q_+$.

2. Die Aussage von Lemma 5.1.1 ist nicht mehr verwunderlich, wenn man sich überlegt, dass die Menge R der verallgemeinerten rationalen Funktionen in $(C(B), \|\cdot\|_\infty)$ eine strikte Sonne bilden (siehe Bemerkung 4. im Anschluss an Definition 3.3.4), sofern sie eine Existenzmenge ist. Dies ist der Inhalt des nächsten Satzes. □

Satz 5.1.2 *Sei $(X, \|\cdot\|) = (C(B), \|\cdot\|_\infty)$ mit kompakter Menge $B \subset \mathbb{R}^N$. Seien $P, Q \subset C(B)$ lineare*[1] *Teilräume und*

$$Q_+ := \{q \in Q : \ q(t) > 0 \text{ für alle } t \in B \neq \emptyset.$$

Ferner sei

$$R := \left\{ \frac{p}{q} : p \in P, \ q \in Q_+ \right\}.$$

Dann gilt: Ist $r^* \in R$ eine beste Approximierende an $z \in C(B) \setminus R$, so ist r^* auch beste Approximierende an $z_\lambda := r^* + \lambda(z - r^*)$ für jedes $\lambda > 0$, d. h. R ist eine strikte Sonne (sofern R eine Existenzmenge ist).

Beweis: Sei $z \in C(B) \setminus R$ und $r^* \in P_M(z)$.

- Es gilt[2] das Kolmogoroff-Kriterium, d. h. mit $f(r) := \|r - z\|_\infty$ ist

$$f'(r^*; r - r^*) \geq 0 \quad \text{für alle } r \in R.$$

Denn: Sei $r^* = p^*/q^*$ und $r = p/q \in R$. Für $s \in [0,1]$ definiere man

$$r_s := \frac{(1-s)p^* + sp}{(1-s)q^* + sq} \in R.$$

Dann ist

$$r_s - r^* = \frac{sq}{(1-s)q^* + sq}(r - r^*)$$

und daher

$$r_s - r^* - s\frac{q}{q^*}(r - r^*) = \frac{s^2 q(q^* - q)}{q^*((1-s)q^* + sq)} = o(s).$$

[1] Hier würde sogar die Konvexität genügen!

[2] Man beachte, dass Satz 3.2.4 *nicht* anwendbar ist, da R nicht konvex ist.

Hierbei benutzen wir das Landau-Symbol $o(\cdot)$, d. h. wir schreiben $g(s) = o(s)$, wenn $\lim_{s\to 0+} g(s)/s = 0$. Da $r^* \in P_R(z)$ ist $\|r^* - z\|_\infty \le \|r_s - z\|_\infty$ für alle $s \in [0,1]$ und daher

$$\begin{aligned} \|r^* - z\| &\le \|r_s - z\| \\ &\le \left\| r_s - r^* - s\frac{q}{q^*}(r - r^*) \right\| + \left\| r^* + s\frac{q}{q^*}(r - r^*) - z \right\|. \end{aligned}$$

Für $s \in (0,1]$ ist daher

$$\underbrace{-\frac{1}{s}\left\| r_s - r^* - s\frac{q}{q^*}(r - r^*) \right\|}_{\to 0} \le \frac{1}{s}\left[\left\| r^* + s\frac{q}{q^*}(r - r^*) - z \right\| - \|r^* - z\| \right].$$

Mit $s \to 0+$ folgt

$$\begin{aligned} 0 &\le f'\left(r^*; \frac{q}{q^*}(r - r^*) \right) \\ &= \max_{t \in B(r^* - z)} \operatorname{sign}\left(r^*(t) - z(t)\right) \frac{q(t)}{q^*(t)} (r(t) - r^*(t)) \end{aligned}$$

mit

$$B(r^* - z) := \{t \in B : |r^*(t) - z(t)| = \|r^* - z\|_\infty\}$$

(siehe Beispiel 3. im Anschluss an Satz 3.2.4). Wegen $q(t)/q^*(t) > 0$ für alle $t \in B$ ist dann auch

$$0 \le f'(r^*; r - r^*) = \max_{t \in B(r^* - z)} \operatorname{sign}\left(r^*(t) - z(t)\right)(r(t) - r^*(t)).$$

Damit ist obige Zwischenbehauptung bewiesen.
Nach Satz 3.2.8 ist

$$f'(r^*; r - r^*) = \max_{l \in \partial f(r^*)} l(r - r^*),$$

ferner ist (siehe Beispiel im Anschluss an Satz 3.2.7)

$$\partial f(r^*) = \{l \in C(B)^* : \|l\| = 1,\ l(r^* - z) = \|r^* - z\|\}.$$

Wegen der oben bewiesenen Hilfsaussage existiert zu jedem $r \in R$ ein $l \in \partial f(r^*)$ mit $0 \le l(r - r^*)$. Hieraus folgt aber unmittelbar die Behauptung (siehe Satz 3.3.8): Sei $\lambda > 0$, $z_\lambda := r^* + \lambda(z - r^*)$ und $r \in R$ beliebig. Hierzu existiert $l \in \partial f(r^*)$ mit $0 \le l(r - r^*)$ und daher ist

$$\begin{aligned} \|r^* - z_\lambda\|_\infty &= \lambda \|r^* - z\|_\infty \\ &= l(\lambda(r^* - z)) \\ &\le l(r - r^* + \lambda(r^* - z)) \\ &= l(r - z_\lambda) \\ &\le \|l\| \, \|r - z_\lambda\|_\infty \\ &= \|r - z_\lambda\|_\infty, \end{aligned}$$

also ist $r^* \in P_R(z_\lambda)$ für alle $\lambda > 0$ und genau das sollte bewiesen werden. □

Nun gehen wir auf die eigentliche rationale Approximation ein und betrachten die Approximation in $(C[\alpha,\beta], \|\cdot\|_\infty)$ mittels Elementen aus

$$R_{m,n}[\alpha,\beta] := \left\{ \frac{p}{q} : p \in \Pi_m,\ q \in \Pi_n,\ \ q(t) > 0 \text{ für alle } t \in [\alpha,\beta] \right\}.$$

Die Darstellung $r = p/q$ von Elementen aus $R_{m,n}[\alpha,\beta]$ kann als irreduzibel angenommen werden. Die irreduzible Darstellung der Null sei durch $0/1$ festgelegt. Wie üblich bezeichnen wir mit ∂p den Grad eines Polynoms p, es sei $\partial 0 = -\infty$.

Zunächst beweisen wir den folgenden Existenzsatz (siehe z. B. E. W. Cheney (1966, S. 154), D. Braess (1986, S. 109)).

Satz 5.1.3 *$R_{m,n}[\alpha,\beta]$ ist eine Existenzmenge in $(C[\alpha,\beta], \|\cdot\|_\infty)$.*

Beweis: Sei $z \in C[\alpha,\beta]$ die zu approximierende Funktion und $\{r_j\} = \{p_j/q_j\} \subset R_{m,n}[\alpha,\beta]$ eine Minimalfolge, also $\|r_j - z\|_\infty \to d(z, R_{m,n}[\alpha,\beta])$. Die Polynome p_j, q_j werden als irreduzibel angenommen und o. B. d. A. ist $\|q_j\|_\infty = 1$. Daher besitzt $\{q_j\} \subset \Pi_n$ eine konvergente Teilfolge und da $\{r_j\}$ als Minimalfolge beschränkt ist, ist $|p_j(t)| \le c\,|q_j(t)| \le c$ für alle $t \in [\alpha,\beta]$ mit einer Konstanten $c > 0$. Also ist auch $\{p_j\}$ beschränkt und besitzt daher ebenfalls eine konvergente Teilfolge. O. B. d. A. ist also $p_j \to p \in \Pi_m$, $q_j \to q \in \Pi_n$ gleichmäßig auf $[\alpha,\beta]$. Aus $|p_j(t)| \le c\,|q_j(t)|$ auf $[\alpha,\beta]$ folgt $|p(t)| \le c\,|q(t)|$ für alle $t \in [\alpha,\beta]$, so dass jede Nullstelle von q in $[\alpha,\beta]$ auch eine Nullstelle von p ist. Sei $r^* = p^*/q^*$ die irreduzible Darstellung von p/q. Offenbar ist $r^* \in R_{m,n}[\alpha,\beta]$ und

$$\lim_{j\to\infty} \frac{p_j(t)}{q_j(t)} = r^*(t)$$

für alle $t \in [\alpha,\beta]$ bis auf, möglicherweise die höchstens n Nullstellen von q in $[\alpha,\beta]$. Bis auf endlich viele $t \in [\alpha,\beta]$ ist also

$$\begin{aligned} |r^*(t) - z(t)| &= \lim_{j\to\infty} \left| \frac{p_j(t)}{q_j(t)} - z(t) \right| \\ &\le \lim_{j\to\infty} \|r_j - z\|_\infty \\ &= d(z, R_{m,n}[\alpha,\beta]). \end{aligned}$$

Da $r^* - z \in C[\alpha,\beta]$ ist $|r^*(t) - z(t)| \le d(z, R_{m,n}[\alpha,\beta])$ für *alle* $t \in [\alpha,\beta]$, also

$$\|r^* - z\| \le d(z, R_{m,n}[\alpha,\beta]).$$

Daher ist $r^* \in R_{m,n}[\alpha,\beta]$ eine beste Approximierende an z in $R_{m,n}[\alpha,\beta]$. □

Bemerkung: Allgemeine rationale Approximationsaufgaben können nicht auf die gleiche Art behandelt werden, da ganz entscheidend einging, dass man durch gemeinsame lineare Faktoren in Zähler und Nenner kürzen kann. Sei etwa $P := \operatorname{span}\{t^2\}$, $Q := \operatorname{span}\{1, t\}$, $[\alpha,\beta] := [0,1]$ und $z(t) := t$. Dann existiert keine beste Approximierende an z in

$$R := \left\{ \frac{at^2}{b+ct} : a,b,c \in \mathbb{R},\ \ b + ct > 0 \text{ für alle } [t \in [0,1] \right\},$$

denn $d(z,R)=0$ (setze $a=b=c$, $b\to 0+$), aber $z\notin R$. Also ist R nicht abgeschlossen und das ist ja eine notwendige Bedingung dafür, dass R eine Existenzmenge ist. □

Die Existenzfrage bezüglich der gewöhnlichen rationalen T-Approximation ist also positiv beantwortet. Bei der *Charakterisierung* bester Approximationen werden wir möglichst Lange den allgemeinen Fall betrachten. Seien $P,Q\subset C(B)$ also $(m+1)$- bzw. $(n+1)$-dimensionale lineare Teilräume und

$$R:=\left\{\frac{p}{q}:p\in P,\ q\in Q\ \text{ mit } q(t)>0 \text{ für alle } t\in B\right\}.$$

Im Beweis von Satz 5.1.2 haben wir zu Beginn nachgewiesen: Ist $r^*\in P_R(z)$, so ist

$$0\le f'(r^*;r-r^*)=\max_{t\in B(r^*-z)}\operatorname{sign}(r^*(t)-z(t))(r(t)-r^*(t))\quad\text{für alle } r\in R.$$

Da die Umkehrung trivialerweise richtig ist, gilt insgesamt das folgende *Kolmogoroff-Kriterium für die rationale T-Approximation*:

Satz 5.1.4 *Sei $(X,\|\cdot\|)=(C(B),\|\cdot\|_\infty)$ mit kompaktem $B\subset\mathbb{R}^N$ und $R\subset C(B)$ wie oben die Menge verallgemeinerter rationaler Funktionen, ferner sei $z\in C(B)\setminus R$. Dann ist $r^*\in R$ genau dann eine beste Approximierende an z in R, wenn*

$$0\le\max_{t\in B(r^*-z)}\operatorname{sign}(r^*(t)-z(t))(r(t)-r^*(t))\quad\textit{für alle } r\in R,$$

wobei

$$B(r^*-z):=\{t\in B:|r^*(t)-z(t)|=\|r^*-z\|_\infty\}.$$

Bemerkung: Das Kolmogoroff-Kriterium kann auf diverse äquivalente Weisen formuliert werden, z. B.

$$\begin{aligned} r^*\in P_R(z)\quad&\Longleftrightarrow\quad 0\le\max_{t\in B(r^*-z)}\operatorname{sign}(r^*(t)-z(t))(r(t)-r(t))\quad\text{für alle } r\in R,\\ &\Longleftrightarrow\quad 0\ge\min_{t\in B(r^*-z)}(z(t)-r^*(t))(r(t)-r^*(t))\quad\text{für alle } r\in R.\end{aligned}$$

□

Eine weitere äquivalente Version wollen wir als Satz formulieren.

Satz 5.1.5 *Unter den Voraussetzungen von Satz 5.1.4 ist $r^*\in R$ genau dann eine beste Approximierende an z in R, wenn*

$$0\le\max_{t\in B(r^*-z)}\operatorname{sign}(r^*(t)-z(t))x(t)\quad\textit{für alle } x\in P+r^*Q,$$

wobei

$$P+r^*Q:=\{p+r^*q:p\in P,\ q\in Q\}.$$

Beweis: Sei $r^*=p^*/q^*$ eine beste Approximierende an z in R, ferner seien $p\in P$, $q\in Q$ beliebig und $x:=p+r^*q$. Für alle hinreichend kleinen $s>0$ ist

$$r_s:=\frac{(1-s)p^*+sp}{(1-s)q^*-sq}\in R.$$

Dann ist

$$r_s - r^* = \frac{s}{(1-s)q^* - sq}(p + r^*q)$$

und

$$r_s - r^* - \frac{s}{q^*}(p + r^*q) = \frac{s^2(q^* + q)}{((1-s)q^* - sq)q^*}(p + r^*q) = o(s).$$

Wie beim Beweis von Satz 5.1.2 erhält man nun mit $f(x) := \|x - z\|_\infty$ aus $r^* \in P_R(z)$, dass

$$\begin{aligned} 0 &\leq f'\left(r^*; \frac{1}{q^*}(p + r^*q)\right) \\ &= f'\left(r^*; \frac{1}{q^*}x\right) \\ &= \max_{t \in B(r^* - z)} \operatorname{sign}(r^*(t) - z(t)) \frac{1}{q^*(t)} x(t), \end{aligned}$$

wegen $q^*(t) > 0$ für alle $t \in B$ ist daher auch

$$0 \leq \max_{t \in B(r^* - z)} \operatorname{sign}(r^*(t) - z(t))x(t) \quad \text{für alle } x \in P + r^*Q.$$

Umgekehrt nehmen wir an, es sei $r^* = p^*/q^* \in R$ und

$$0 \leq \max_{t \in B(r^* - z)} \operatorname{sign}(r^*(t) - z(t))x(t) \quad \text{für alle } x \in P + r^*Q.$$

Sei $r = p/q \in R$ beliebig. Dann ist (setze $x := p + r^*(-q)$)

$$0 \leq \max_{t \in B(r^* - z)} \operatorname{sign}(r^*(t) - z(t))(p(t) - r^*(t)q(t))$$

und daher auch

$$0 \leq \max_{t \in B(r^* - z)} \operatorname{sign}(r^*(t) - z(t))(r(t) - r^*(t)).$$

Wegen Satz 5.1.4 ist r^* eine beste Approximierende an z in R. □

Entscheidend für eine weitere Charakterisierung einer besten Approximierenden r^* ist der lineare Raum $P + r^*Q$. Ist $\dim(P) = m + 1$, $\dim(Q) = n + 1$, so ist

$$\dim(P + r^*Q) \leq m + n + 1.$$

Denn: Ist etwa

$$P = \operatorname{span}\{p_0, \ldots, p_m\}, \qquad Q = \operatorname{span}\{q_0, \ldots, q_n\},$$

so ist

$$P + r^*Q = \operatorname{span}\{p_0, \ldots, p_m, r^*q_0, \ldots, r^*q_n\}.$$

Aber $\{p_0, \ldots, p_m, r^*q_0, \ldots, q_n\}$ sind linear abhängig, da aus

$$r^* = \frac{\sum_{i=0}^m a_i^* p_i}{\sum_{i=0}^n b_i^* q_i}$$

folgt, dass

$$\sum_{i=0}^{m} a_i^* p_i - \sum_{i=0}^{n} b_i r^* q_i = 0.$$

Wir sind jetzt in einer Position, dass wir wie in der linearen T-Approximation vorgehen können.

Satz 5.1.6 *Gegeben sei ein Approximationsproblem mit* $(X, \|\cdot\|) = (C(B), \|\cdot\|_\infty)$ *mit kompaktem* $B \subset \mathbb{R}^N$ *gegeben, bei welchem mit Elementen aus*

$$R := \left\{ \frac{p}{q} : p \in P,\ q \in Q,\ \ q(t) > 0 \text{ für alle } t \in B \right\}$$

approximiert wird. P *bzw.* Q *seien hierbei* $(m+1)$*- bzw.* $(n+1($*-dimensionale lineare Teilräume von* $C(B)$ *und* $z \in C(B) \setminus R$ *das zu approximierende Element. Sei* $r^* \in R$*,* $k := \dim(P + r^*Q)$ *und* $P + r^*Q = \operatorname{span}\{x_1, \ldots, x_k\}$*. Dann gilt:*

1. *Es ist* r^* *genau dann eine beste Approximierende an* z *in* R*, wenn*

$$(*) \qquad 0 \in \operatorname{co}\left(\{\operatorname{sign}(r^*(t) - z(t)) \begin{pmatrix} x_1(t) \\ \vdots \\ x_k(t) \end{pmatrix} : t \in B(r^* - t)\}\right),$$

   wobei

$$B(r^* - z) := \{t \in B : |r^*(t) - z(t)| = \|r^* - z\|_\infty\}.$$

   Wegen des Satzes von Carathéodory (Satz 2.2.13) ist $(*)$ *gleichwertig mit der Existenz von* $l \in \{1, \ldots, k+1\}$*,* $\lambda_j > 0$*,* $t_j \in B(r^* - z)$*,* $j = 1, \ldots, l$*, mit*

$$0 = \sum_{j=1}^{l} \lambda_j \operatorname{sign}(r^*(t_j) - z(t_j)) x(t_j) \quad \text{für alle } x \in P + r^*Q.$$

2. *Ist* $P + r^*Q$ *ein* k*-dimensionaler Haarscher Teilraum von* $C(B)$*, so ist in 1. notwendig* $l = k + 1$*. In diesem Fall ist also* r^* *genau dann eine beste Approximierende an* z *in* R*, wenn* $\lambda_j > 0$*,* $t_j \in B(r^* - z)$*,* $j = 1, \ldots, k+1$*, existieren mit*

$$0 = \sum_{j=1}^{k+1} \lambda_j \operatorname{sign}(r^*(t_j) - z(t_j)) x(t_j) \quad \text{für alle } x \in P + r^*Q.$$

3. *Ist* $P + r^*Q$ *ein* k*-dimensionaler Haarscher Teilraum von* $C(B)$ *und* $r^* \in R$ *eine beste Approximierende an* z *in* R*, so ist* r^* *die einzige beste Approximierende an* z *in* R*.*

Beweis: 1. Sei $r^* \in R$ eine beste Approximierende an z in R. Wäre

$$0 \notin \operatorname{co}\left(\{\operatorname{sign}(r^*(t) - z(t)) \begin{pmatrix} x_1(t) \\ \vdots \\ x_k(t) \end{pmatrix} : t \in B(r^* - t)\}\right),$$

so würde wie im Beweis von Satz 4.1.2 die Existenz eines $y \in P + r^*Q$ mit

$$\max_{t \in B(r^*-z)} \operatorname{sign}(r^*(t) - z(t))y(t) < 0$$

folgen, ein Widerspruch zu Satz 5.1.5. Gilt umgekehrt $(*)$, so ist

$$0 \leq \max_{t \in B(r^*-z)} \operatorname{sign}(r^*(t) - z(t))x(t) \quad \text{für alle } x \in P + r^*Q$$

bzw. $r^* \in P_R(z)$, denn andernfalls existiert $y = (y_1, \ldots, y_k)^T \in \mathbb{R}^k$ mit

$$\operatorname{sign}(r^*(t) - z(t))\, y^T \begin{pmatrix} x_1(t) \\ \vdots \\ x_k(t) \end{pmatrix} < 0 \quad \text{für alle } t \in B(r^* - z),$$

ein Widerspruch zu $(*)$. Der Rest von 1. ist evident.

2. Sei $P + r^*Q$ ein k-dimensionaler Haarscher Raum. Angenommen, in 1. sei $l \leq k$. Man ergänze (notfalls) $\{t_1, \ldots, t_l\}$ durch $t_{l+1}, \ldots, t_k \in B$ so, dass die t_j, $j = 1, \ldots, k$, paarweise verschieden sind. Da $P + r^*Q$ ein Haarscher Teilraum von $C(B)$ ist, ist die Matrix $(x_i(t_j))_{i,j=1,\ldots,k+1}$ nichtsingulär. Dies ist ein Widerspruch zu

$$\sum_{j=1}^{l} \underbrace{\lambda_j \operatorname{sign}(r^*(t_j) - z(t_j))}_{\neq 0} x_i(t_j) = 0, \quad i = 1, \ldots, k.$$

Damit ist auch 2. bewiesen.

3. Sei $r^* \in R$ eine beste Approximierende an z in R und $P + r^*Q$ ein k-dimensionaler Haarscher Teilraum von $C(B)$. Sei auch $r_1^* = p_1^*/q_1^* \in P_R(z)$. Wegen 2. existieren $\lambda_j > 0$, $t_j \in B(r^* - z)$, $j = 1, \ldots, k+1$, mit

$$0 = \sum_{j=1}^{k+1} \lambda_j \operatorname{sign}(r^*(t_j) - z(t_j))x(t_j) \quad \text{für alle } x \in P + r^*Q,$$

insbesondere ist

$$(*) \qquad 0 = \sum_{j=1}^{k+1} \lambda_j \operatorname{sign}(r^*(t_j) - z(t_j))(-p_1^*(t_j) + r^*(t_j)q_1^*(t_j)).$$

Für $t \in B(r^* - z)$ ist aber

$$\begin{aligned} \operatorname{sign}(r^*(t) - z(t))(r_1^*(t) - z(t)) &\leq \|r_1^* - z\|_\infty \\ &= \|r^* - z\|_\infty \\ &= |r^*(t) - z(t)| \\ &= \operatorname{sign}(r^*(t) - z(t))(r^*(t) - z(t)). \end{aligned}$$

Für $t \in B(r^* - z)$ ist daher

$$\begin{aligned} 0 &\leq \operatorname{sign}(r^*(t) - z(t))(r^*(t) - r_1^*(t)) \\ &= \operatorname{sign}(r^*(t) - z(t)) \underbrace{\frac{1}{q_1^*(t)}}_{>0}(-p_1^*(t) + r^*(t)q_1^*(t)) \end{aligned}$$

und daher

$$0 \leq \operatorname{sign}(r^*(t) - z(t))(-p_1^*(t) + r^*(t)q_1^*(t)) \quad \text{für alle } t \in B(r^* - z).$$

Aus $(*)$ folgt

$$-p_1^*(t_j) + r^*(t_j)q_1^*(t_j) = 0, \quad j = 1, \ldots, k+1.$$

Da $P + r^*Q$ nach Voraussetzung ein Haarscher Teilraum ist, ist $-p_1^* + r^*q_1^* = 0$, also $r_1^* = r^*$, und damit ist die Eindeutigkeit einer besten Approximierenden bewiesen. □

Nun kehren wir zurück zu der uns vor allem interessierenden speziellen rationalen Approximation auf einem Intervall $B = [\alpha, \beta]$. Sei also $P = \Pi_m$, $Q = \Pi_n$ und

$$R_{m,n} := \left\{ \frac{p}{q} : p \in P,\ q \in Q,\ \ q(t) > 0 \text{ für alle } t \in [\alpha, \beta] \right\}.$$

Die Elemente $r = p/q \in R_{m,n}[\alpha, \beta]$ werden als irreduzibel angenommen. Wir erinnern an eine Definition, die schon Abschnitt 3.4 vorkam.

Definition 5.1.7 Ist $r = p/q \in R_{m,n}[\alpha, \beta]$, so heißt $d(r) := \min(m - \partial p, m - \partial q)$ der *Defekt von* r. Wir nennen r *ausgeartet*, falls $d(r) > 0$, andernfalls *normal* oder *nichtausgeartet*.

Die Menge $R_{m,n}[\alpha, \beta]$ ist nicht nur eine Existenzmenge (Satz 5.1.3), sondern sogar eine T-Menge, falls $\Pi_m + r^*\Pi_n$ für jedes $r^* \in R$ ein Haarscher Teilraum von $C[\alpha, \beta]$ ist. Genau dies wird im nächsten Satz bewiesen (siehe z. B. E. W. Cheney (1966, S. 162)).

Satz 5.1.8 *$r^* \in R_{m,n}[\alpha, \beta]$ habe die irreduzible Darstellung $r^* = p^*/q^*$. Dann ist $\Pi_m + r^*\Pi_n$ ein Haarscher Teilraum der Dimension $k := m + n - d(r^*) + 1$.*

Beweis: Ist $r^* = 0 = 0/1$, so ist $d(r^*) = \min(m + \infty, n - 0) = n$ und Π_m ist in der Tat ein Haarscher Teilraum der Dimension $m + 1 = m + n - d(r^*) + 1$. Sei daher o. B. d. A. $r^* \neq 0$. Der Beweis zerfällt in zwei Teile. Im ersten zeigen wir:

- Es ist $\Pi_m \cap r^*\Pi_n = p^*\Pi_{d(r^*)}$ und $\dim(\Pi_m + r^*\Pi_n) = m + n - d(r^*) + 1$.

Denn: Wegen der Dimensionsformel für die Summe zweier endlichdimensionaler Teilräume eines Vektorraums gilt

$$\begin{aligned} \dim(\Pi_m + r^*\Pi_n) &= \dim(\Pi_m) + \dim(r^*\Pi_n) - \dim(\Pi_m \cap r^*\Pi_n) \\ &= m + 1 + n + 1 - \dim(\Pi_m \cap r^*\Pi_n). \end{aligned}$$

Wir zeigen, dass $\Pi_m \cap r^*\Pi_n = p^*\Pi_{d(r^*)}$, daher gilt $\dim(\Pi_m \cap r^*\Pi_n) = d(r^*)+1$ und damit die Zwischenbehauptung. Sei $r^*q \in \Pi_m \cap r^*\Pi_n$. Da p^*/q^* eine irreduzible Darstellung von r^* ist, wird q durch q^* geteilt, es ist also $q = q'q^*$. Folglich ist $n \geq \partial q = \partial q' + \partial q^*$ und daher $\partial q' \leq n - \partial q^*$. Andererseits ist $r^*q = p^*q' \in \Pi_m$, also

$$\partial q' \leq \min(m - \partial p^*, n - \partial q^*) = d(r^*).$$

Also ist $r^*q = p^*q' \in p^*\Pi_{d(r^*)}$. Damit ist $\Pi_m \cap r^*\Pi_n \subset p^*\Pi_{d(r^*)}$ nachgewiesen. Sei umgekehrt $q' \in \Pi_{d(r^*)}$. Wir zeigen, dass $p^*q' \in \Pi_m \cap r^*\Pi_n$. Wegen

$$\partial(p^*q') \leq \partial p^* + \partial q' \leq \partial p^* + d(r^*) \leq m$$

ist $p^*q' \in \Pi_m$. Andererseits ist $p^*q' = (p^*/q^*)q'q^* = r^*(q'q) \in r^*\Pi_n$, da

$$\partial(q'q) \leq \partial q^* + d(r^*) \leq n.$$

Insgesamt ist die Zwischenbehauptung $\dim(\Pi_m + r^*\Pi_n) = m + n - d(r^*) + 1$ bewiesen. Nun kommen wir zum Schluss des Beweises. Sei $x = p + r^*q \in \Pi_m + r^*\Pi_n$ nichttrivial. Angenommen, x besitzt $k = m + n - d(r^*) + 1$ (paarweise) verschiedene Nullstellen in $[\alpha, \beta]$. Das gilt dann auch für $xq^* = pq^* + p^*q$ und dies ist ein Polynom vom Grad

$$\partial(pq^* + p^*q) \leq \max(m + \partial q^*, n + \partial p^*) = m + n - d(r^*),$$

welches höchstens $m+n-d(r^*)$ Nullstellen besitzen kann. Damit ist der Satz vollständig bewiesen. □

Als Folgerung aus Satz 5.1.3 (Existenzsatz für rationale T-Approximation), Teil 3. von Satz 5.1.6 und Satz 5.1.8 halten wir fest:

Satz 5.1.9 *$R_{m,n}[\alpha, \beta]$ ist eine T-Menge in $(C[\alpha, \beta], \|\cdot\|_\infty)$.*

Bemerkung: Für die allgemeine rationale Approximation in $(C(B), \|\cdot\|_\infty)$ gilt i. Allg. *nicht*, dass $P + r^*Q$ ein Haascher Teilraum von $C(B)$ ist, wenn $P, Q \subset C(B)$ Haarsche Teilräume sind und $r^* \in R$. hierzu geben wir ein Beispiel (siehe E. W. Cheney (1966, S. 169)) an. Sei $B := [0, 3]$, $P := \operatorname{span}\{1, t^2\}$, $r^*(t) := (1 + t^2)/(1 + t)$, ferner sei $x(t) := 6 + t^2 - 6r^*(t)$. Dann ist $x \in P + r^*Q$ nichttrivial mit drei Nullstellen in $[0, 3]$, nämlich $0, 2, 3$. Andererseits ist $P + r^*Q$ ein 3-dimensionaler Teilraum von $C[0, 3]$. □

Es folgt der *Alternantensatz für die rationale T-Approximation*.

Satz 5.1.10 *Sei $z \in C[\alpha, \beta] \setminus R_{m,n}[\alpha, \beta]$. Dann ist $r^* \in R_{m,n}[\alpha, \beta]$ mit der irreduziblen Darstellung $r^* = p^*/q^*$ genau dann die beste T-Approximierende an z in $R_{m,n}[\alpha, \beta]$, wenn $r^* - z$ eine $(m + n - d(r^*) + 2)$-punktige Alternante besitzt, wenn es also genau $m + n - d(r^*) + 2$ Punkte t_j gibt mit*

(a) $\alpha \leq t_0 < \cdots < t_{m+n-d(r^*)+1} \leq \beta$,

(b) $|r^*(t_j) - z(t_j)| = \|r^* - z\|_\infty$, $j = 0, \ldots, m + n - d(r^*) + 1$,

(c) $r^*(t_j) - z(t_j) = (-1)^j(r^*(t_0) - z(t_0))$, $j = 0, \ldots, m + n - d(r^*) + 1$.

Beweis: Der Beweis besteht lediglich in einem Zusammenfügen früherer Ergebnisse. Wegen der Teile 1. und 2. von Satz 5.1.6 ist $r^* \in R_{m,n}[\alpha,\beta]$ genau dann eine beste Approximierende an z in $R_{m,n}[\alpha,\beta]$, wenn $\lambda_j > 0$, $t_j \in B(r^*-z)$, $j = 0,\dots,m+n-d(r^*)+1$, existieren mit

$$\sum_{j=0}^{m+n-d(r^*)+1} \lambda_j \operatorname{sign}(r^*(t_j)-z(t_j))x(t_j) = 0 \quad \text{für alle } x \in \Pi_m + r^*\Pi_n.$$

Ist $r^* \in R_{m,n}[\alpha,\beta]$ beste Approximierende an z in $R_{m,n}[\alpha,\beta]$, so sind (a) und (b) erfüllt, da die t_j als der Größe nach angeordnet angenommen werden können. Die Eigenschaft (c) folgt aus dem ersten Teil von Lemma 3.4.4. Existieren zu $r^* \in R_{m,n}[\alpha,\beta]$ umgekehrt $m+n-d(r^*)+2$ Punkte t_j mit den Eigenschaften (a)–(c), so ist r^* wegen Satz 3.4.5, dem Satz von de La Vallée Poussin, beste Approximierende an z in $R_{m,n}[\alpha,\beta]$. □

Bemerkung: Man hat also bei der rationalen T-Approximation eine weitgehende Analogie zur linearen T-Approximation. Der einzige, allerdings unangenehme Unterschied besteht darin, dass man die Zahl der Alternantenpunkte nicht a priori kennt, da man mit Ausartung der besten Approximierenden rechnen muss. □

Wir haben bisher die Existenz, Eindeutigkeit und eine Charakterisierung bester rationaler T-approximierender bewiesen. Es fehlt noch die starke Eindeutigkeit. Diese ist allerdings nur für *normale* beste Approximierende richtig. Auch hier zeigt sich also ein Unterschied zur linearen T-Approximation.

Satz 5.1.11 *Sei $z \in C[\alpha,\beta]$ und $r^* \in R_{m,n}[\alpha,\beta]$ die beste T-Approximierende an z in $R_{m,n}[\alpha,\beta]$. Ist r^* normal bzw. $d(r^*)=0$, so ist r^* stark eindeutig, d. h. es existiert eine Konstante $c>0$ mit*

$$\|r-z\|_\infty \geq \|r^*-z\|_\infty + c\,\|r^*-r\|_\infty \quad \textit{für alle } r \in R_{m,n}[\alpha,\beta].$$

Beweis: O. B. d. A. ist $z \notin R_{m,n}[\alpha,\beta]$, denn andernfalls ist $r^* = z$ und es kann $c = 1$ gewählt werden. Wegen der Normalität von r^* ist $\Pi_m + r^*\Pi_n$ mit $l := m+n+1$ ein l-dimensionaler Haarscher Teilraum (siehe Satz 5.1.8). Wegen des zweiten Teiles von Satz 5.1.6 existieren $\lambda_j > 0$, $t_j \in B(r^*-z)$, $j = 0,\dots,l$, mit

$$\sum_{j=0}^{l} \lambda_j \operatorname{sign}(r^*(t_j)-z(t_j))x(t_j) = 0 \quad \text{für alle } x \in \Pi_m + r^*\Pi_n.$$

Für $x \in (\Pi_m + r^*\Pi_n)\setminus\{0\}$ ist also

$$\psi(x) := \max_{j=0,\dots,l} \operatorname{sign}(r^*(t_j)-z(t_j))x(t_j) > 0.$$

Angenommen, die Behauptung sei falsch. Dann existiert eine Folge $\{r_k\} \subset R_{m,n}[\alpha,\beta]\setminus\{r^*\}$ mit

$$c_k := \frac{\|r_k-z\|_\infty - \|r^*-z\|_\infty}{\|r_k-r^*\|_\infty} \to 0.$$

Die Folge $\{r_k\}$ ist beschränkt, denn

$$\begin{aligned}\|r_k - r^*\|_\infty - \|r^* - z\|_\infty &\leq \|r_k - z\|_\infty \\ &= \|r^* - z\|_\infty + c_k \|r_k - r^*\|_\infty \\ &\leq \|r^* - z\|_\infty + \frac{1}{2}\|r_k - r^*\|_\infty\end{aligned}$$

für alle hinreichend großen k. Hieraus folgt $\|r_k - r^*\|_\infty \leq 4\|r^* - z\|_\infty$ für alle hinreichend großen k und das ist die Beschränktheit von $\{r_k\}$. Sei $r_k = p_k/q_k$, $r^* = p^*/q^*$. O. B. d. A. ist $\|q_k\|_\infty = \|q^*\|_\infty = 1$. Da $\{r_k\}$ beschränkt ist, können wir o. B. d. A. annehmen (notfalls gehe man zu Teilfolgen über), dass $p_k \to p$, $q_k \to q$.

Nun zeigen wir, dass $\psi(p - r^*q) \leq 0$ und folglich $p - r^*q = 0$ gilt. Denn für $j = 0, \ldots, l$ ist

$$\begin{aligned}c_k \|r_k - r^*\|_\infty &= \|r_k - z\|_\infty - \|r^* - z\|_\infty \\ &\geq \operatorname{sign}(r^*(t_j) - z(t_j))(r_k(t_j) - z(t_j)) \\ &\qquad - \underbrace{\operatorname{sign}(r^*(t_j) - z(t_j))(r^*(t_j) - z(t_j))}_{=\|r^* - z\|_\infty} \\ &= \operatorname{sign}(r^*(t_j) - z(t_j))(r_k(t_j) - r^*(t_j)) \\ &= \frac{1}{q_k(t_j)} \operatorname{sign}(r^*(t_j) - z(t_j))(p_k(t_j) - r^*(t_j) q_k(t_j)).\end{aligned}$$

Hieraus erhalten wir

$$q_k(t_j) c_k \|r_k - r^*\|_\infty \geq \operatorname{sign}(r^*(t_j) - z(t_j))(p_k(t_j) - r^*(t_j) q_k(t_j)).$$

Mit $k \to \infty$ folgt

$$0 \geq \operatorname{sign}(r^*(t_j) - z(t_j))(p(t_j) - r^*(t_j) q(t_j)), \quad j = 0, \ldots, l.$$

Also ist

$$\psi(p - r^*q) = \max_{j=0,\ldots,l} \operatorname{sign}(r^*(t_j) - z(t_j))(p(t_j) - r^*(t_j) q(t_j)) \leq 0$$

und damit $p = r^*q$.

Im nächsten Teil des Beweises zeigen wir, dass $p = p^*$ und $q = q^*$. Da $r^* \in R_{m,n}[\alpha, \beta]$ normal ist, ist $d(r^*) = 0$. Folglich ist (erster Teil von Satz 5.1.8)

$$p = r^*q \in \Pi_m \cap r^*\Pi_n = p^*\Pi_{d(r^*)} = p^*\Pi_0.$$

Daher existiert eine Konstante c_0 mit $p = p^*c_0$. Dann ist $p^*q = pq^* = p^*c_0q^*$. Jetzt machen wir eine Fallunterscheidung. Ist $p^* \neq 0$, so ist $q = c_0q^*$. Wegen $q^* > 0$ und $q \geq 0$ $\|q^*\|_\infty = \|q\|_\infty = 1$ ist $c_0 = 1$, also $p = p^*$, $q = q^*$. Ist dagegen $p^* = 0$, so ist $r^* = 0 = 0/1$ und folglich $0 = d(r^*) = n - 0$, damit $n = 0$. Also sind q und q^* Konstanten, $q^* = 1 = q$ und damit wieder $p = p^*$, $q = q^*$.

Im letzten Teil des Beweises zeigen wir, dass dass die Folge $\{c_k\}$ durch eine positive Konstante nach unten beschränkt ist, was einen Widerspruch zu der Annahme $c_k \to 0$

liefert. Wegen $q_k \to q^* > 0$ kann die Existenz eines $\epsilon > 0$ mit $q_k(t) \geq \epsilon$ für alle $t \in [\alpha, \beta]$ und alle $k \in \mathbb{N}$ angenommen werden. Man definiere

$$\delta := \min_{x \in \Pi_m + r^* \Pi_n, \|x\|_\infty = 1} \psi(x).$$

Wegen des ersten Teils des Beweises ist $\delta > 0$. Sei nun $k \in \mathbb{N}$ fest. Wegen

$$\psi\left(\frac{p_k - r^* q_k}{\|p_k - r^* q_k\|_\infty}\right) \geq \delta$$

existiert ein $j \in \{0, \ldots, l\}$ mit

$$\operatorname{sign}(r^*(t_j) - z(t_j))(p_k(t_j) - r^*(t_j) q_k(t_j)) \geq \delta \|p_k - r^* q_k\|_\infty.$$

Aus (siehe oben im Beweis)

$$\begin{aligned} c_k \|r_k - r^*\|_\infty &\geq \frac{1}{q_k(t_j)} \operatorname{sign}(r^*(t_j) - z(t_j))(p_k(t_j) - r^*(t_j) q_k(t_j)) \\ &\geq \frac{\delta}{q_k(t_j)} \|p_k - r^* q_k\|_\infty \\ &\geq \delta \|p_k - r^* q_k\|_\infty \\ &\qquad \text{(wegen } 0 < q_k(t_j) \leq 1\text{)} \\ &\geq \delta\epsilon \|r_k - r^*\|_\infty \\ &\qquad \text{(wegen } \epsilon \leq q_k(t) \text{ für alle } t \in [\alpha, \beta]\text{)} \end{aligned}$$

folgt $c_k \geq \delta\epsilon$ für alle k, ein Widerspruch zu $c_k \to 0$. Damit ist die starke Eindeutigkeit einer normalen besten Approximierenden bewiesen. □

Aus der starken Eindeutigkeit nicht ausgearteter bester Approximierender erhält man nun ähnlich wie bei der linearen T-Approximation (siehe Satz 4.1.8) die lokale Lipschitzstetigkeit der metrischen Projektion.

Satz 5.1.12 *Sei $z_0 \in C[\alpha, \beta]$. Die zugehörige beste T-Approximierende r_0^* in $R_{m,n}[\alpha, \beta]$ sei nicht ausgeartet bzw. normal. Dann existiert eine Konstante $c_0 > 0$ mit*

$$\|P_{R_{m,n}[\alpha,\beta]}(z) - P_{R_{m,n}[\alpha,\beta]}(z_0)\|_\infty \leq c_0 \|z - z_0\|_\infty \quad \textit{für alle } z \in C[\alpha, \beta],$$

wobei $P_{R_{m,n}[\alpha,\beta]} \colon C[\alpha, \beta] \longrightarrow R_{m,n}[\alpha, \beta]$ die metrische Projektion auf $R_{m,n}[\alpha, \beta]$ ist.

Beweis: Der Beweis ist völlig analog dem von Satz 4.1.8. Wir wiederholen ihn trotzdem. Nach Satz 5.1.11 ist $r_0^* = P_{R_{m,n}[\alpha,\beta]}(z_0)$ stark eindeutige beste Approximierende an z_0, es existiert also eine Konstante $c > 0$ mit

$$\|r - z_0\|_\infty \geq \|P_{R_{m,n}[\alpha,\beta]}(z_0) - z_0\|_\infty + c \|P_{R_{m,n}[\alpha,\beta]}(z_0) - r\|_\infty$$

für alle $r \in R_{m,n}[\alpha,\beta]$. Sei $z \in C[\alpha,\beta]$ beliebig, setze $r := P_{R_{m,n}[\alpha,\beta]}(z)$. Dann ist

$$\begin{aligned}
& \|P_{R_{m,n}[\alpha,\beta]}(z) - P_{R_{m,n}[\alpha,\beta]}(z_0)\|_\infty \\
\leq\ & \frac{1}{c}[\|P_{R_{m,n}[\alpha,\beta]}(z) - z_0\|_\infty - \|P_{R_{m,n}[\alpha,\beta]}(z_0) - z_0\|_\infty] \\
\leq\ & \frac{1}{c}[\|P_{R_{m,n}[\alpha,\beta]}(z) - z\|_\infty + \|z - z_0\|_\infty - \|P_{R_{m,n}[\alpha,\beta]}(z_0) - z_0\|_\infty] \\
=\ & \frac{1}{c}[\underbrace{\|P_{R_{m,n}[\alpha,\beta]}(z) - z\|_\infty - \|P_{R_{m,n}[\alpha,\beta]}(z_0) - z\|_\infty}_{\leq 0} + \|z - z_0\|_\infty \\
& \qquad + \underbrace{\|P_{R_{m,n}[\alpha,\beta]}(z_0) - z\|_\infty - \|P_{R_{m,n}[\alpha,\beta]}(z_0) - z_0\|_\infty}_{\leq \|z - z_0\|_\infty}] \\
\leq\ & \frac{2}{c}\|z - z_0\|_\infty,
\end{aligned}$$

die Behauptung ist also mit $c_0 := 2/c$ bewiesen. □

Bemerkung: Elemente von

$$N_{m,n}[\alpha,\beta] := \{z \in C[\alpha,\beta] : d(P_{R_{m,n}[\alpha,\beta]}(z) = 0\}$$

heißen *normale* Funktionen bezüglich $R_{m,n}[\alpha,\beta]$. Also heißt $z \in C[\alpha, /beta]$ normal, wenn die zugehörige beste Approximierende an z in $R_{m,n}[\alpha,\beta]$ nicht ausgeartet bzw. normal ist. Es stellt sich nun natürlich die Frage, wie "wahrscheinlich" es ist, dass ein vorgegebenes $z \in C[\alpha,\beta]$ normal ist, bzw. wie "groß" die Menge $N_{m,n}[\alpha,\beta]$ ist. Von E. W. Cheney, H. L. Loeb (1964) ist gezeigt worden, dass $N_{m,n}[\alpha,\beta]$ eine offene und dichte Menge in $(C[\alpha,\beta], \|\cdot\|_\infty)$ ist. □

Als Ergänzung zu Satz 5.1.12 kann gezeigt werden, dass die metrische Projektion in einem $z_0 \in C[\alpha,\beta] \setminus (N_{m,n}[\alpha,\beta] \cup R_{m,n}[\alpha,\beta])$ nicht stetig ist (siehe z. B. D. Braess (1986, S. 115)). Dies wollen wir nicht beweisen, sondern nur an einem Beispiel demonstrieren.

Beispiel: In $R_{0,1}[0,1]$ definiere man

$$r_0^*(t) := 0, \qquad r_\lambda^*(t) = \frac{\lambda}{\lambda + t} \quad (\lambda > 0).$$

Für $\lambda \geq 0$ sei z_λ die stückweise lineare Funktion mit Knoten in $0, \frac{1}{2}, 1$, für die

$$z_\lambda(0) = \tfrac{1}{2}, \qquad z_\lambda(\tfrac{1}{2}) = r_\lambda^*(\tfrac{1}{2}) + \tfrac{1}{2}, \qquad z_\lambda(1) = r_\lambda^*(1) - \tfrac{1}{2}.$$

Dann ist $r_\lambda^* \in P_{R_{0,1}[0,1]}(z_\lambda)$ für alle $\lambda \geq 0$. Dies ist für $\lambda = 0$ offensichtlich, siehe Abbildung 5.1. Denn wegen des Alternantensatzes ist $r_0^* = 0/1 \in R_{0,1}[0,1]$, wenn es zu $r_0^* - z_0$ eine $0 + 1 - 1 + 2 = 2$-elementige Alternante gibt. Durch $\{0,1\}$ ist eine solche gegeben. Denn offenbar ist

$$r_0^*(0) - z_0(0) = -(r_0^*(1) - z_0(1)) = -\tfrac{1}{2},$$

ferner ist offensichtlich $\|r_0^* - z_0\|_\infty = \frac{1}{2}$. Insbesondere ist z_0 nicht normal, da die zugehörige beste Approximierende ausgeartet ist. Um nachzuweisen, dass $r_\lambda^* = P_{R_{0,1}[0,1]}(z_\lambda)$

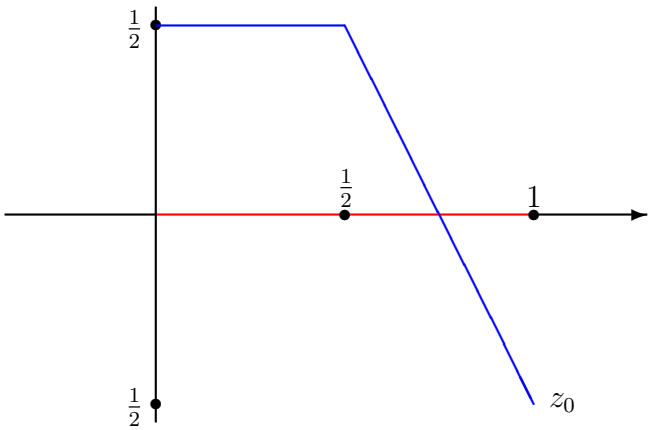

Abbildung 5.1: Es ist $r_0^* = P_{R_{0,1}[0,1]}(z_0)$

für $\lambda > 0$ hat man die Existenz einer $0+1-0+2 = 3$-elementigen Alternante zu $r_\lambda^* - z_\lambda$ nachzuweisen. Eine solche Alternante ist durch $\{0, \frac{1}{2}, 1\}$ gegeben. Denn offenbar ist

$$r_\lambda^*(0) - z_\lambda(0) = -(r_\lambda^*(\tfrac{1}{2}) - z_\lambda(\tfrac{1}{2}) = r_\lambda^*(1) - z_\lambda(1) = \tfrac{1}{2},$$

so dass nur noch $\|r_\lambda^* - z_\lambda\|_\infty = \frac{1}{2}$ zu zeigen ist. Wir verweisen lediglich auf Abbildung 5.2, in der wir z_λ, r_λ^* und $r_\lambda^* - z_\lambda$ für $\lambda = \frac{1}{2}$ eingetragen haben. □

Zum Schluss dieses Abschnitts über die rationale T-Approximation in $C[\alpha, \beta]$ wollen wir einige wenige Bemerkungen zur *diskreten* rationalen T-Approximation machen. Während man bei der kontinuierlichen rationalen T-Approximation eine fast vollständige Analogie zur kontinuierlichen linearen T-Approximation vorfindet ("fast", da man mit ausgearteten besten Approximierenden rechnen muss), so ist dies bei der diskreten rationalen T-Approximation nicht mehr der Fall. Denn schon die Existenz einer besten Approximierenden ist im diskreten rationalen Fall i. Allg. nicht gesichert. Hierzu geben wir gleich ein Beispiel an, ein weiteres Beispiel findet man bei G. A. Watson (1980, S. 193 ff.).

Beispiel: Sei $B := \{0, 1\}$ und $z \in C(B)$ gegeben durch $z(0) = 1$, $z(1) = 0$. Dann gilt:

(a) Es ist

$$\inf_{r \in R_{0,1}[0,1]} \max_{t \in B} |r(t) - z(t)| = 0.$$

Denn: Für $r \in R_{0,1}[0, 1]$ mache man den Ansatz $r(t) = a/(1 + bt)$ mit $1 + b > 0$. Dann ist

$$\max_{t \in B} |r(t) - z(t)| = \max\left(|a - 1|, \frac{|a|}{1 + b}\right).$$

Mit $a = 1$ und $b \to \infty$ erhalten wir die Behauptung.

(b) Es existiert kein $r \in R_{0,1}[0, 1]$ mit $\max_{t \in B} |r(t) - z(t)|$ bzw. $r(t) = z(t)$ für $t \in B$.

Denn: Ein $r \in R_{0,1}[0, 1]$ hat die Form $r(t) = a/(1+bt)$. Aus $r(1) = z(1) = 0$ folgt $a = 0$. Dann ist aber $0 = r(0) \neq z(0) = 1$.

□

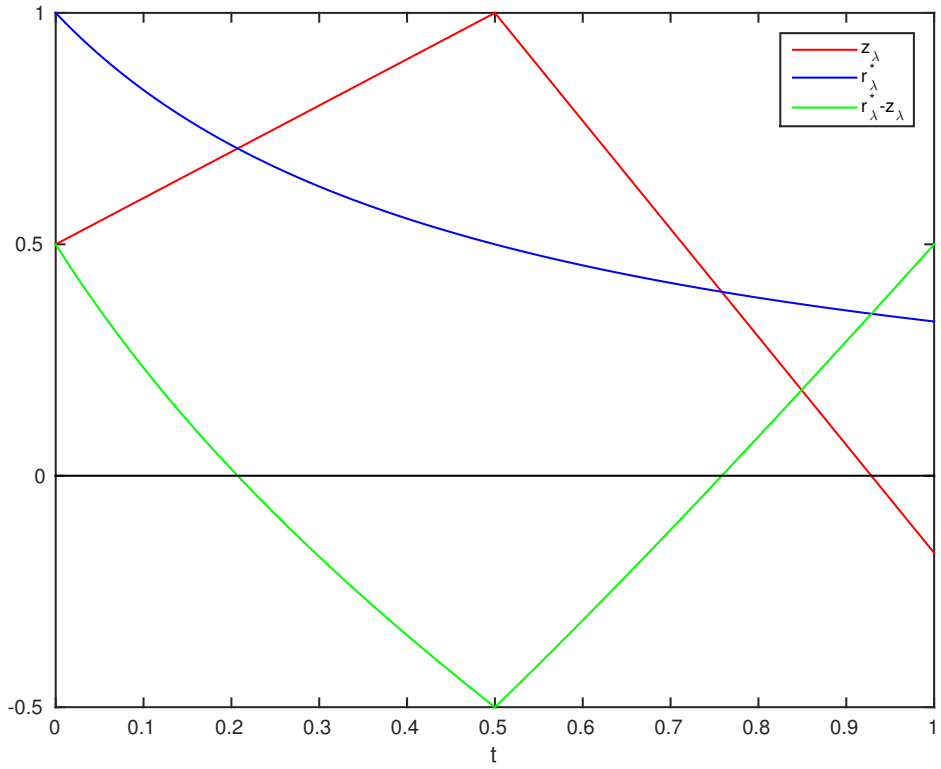

Abbildung 5.2: z_λ, r^*_λ und $r^*_\lambda - z_\lambda$ für $\lambda = \frac{1}{2}$

5.2 Bemerkungen zur numerischen Behandlung rationaler T-Approximation

Wir wollen uns in diesem kurzen abschnitt nit Bemerkungen zum Differential Correction Algorithmus und zum Remez-Verfahren begnügen.

Bei der Schilderung des *Differential Correction Algorithmus* können wir von einer *verallgemeinerten* rationalen T-Approximationsaufgabe ausgehen. Seien also $P, Q \subset C(B)$ endlichdimensionale lineare Teilräume und

$$R := \left\{ \frac{p}{q} : p \in P,\ q \in Q,\ \ q(t) > 0 \text{ für alle } t \in B \right\}.$$

Ferner sei $z \in C(B) \setminus R$ vorgegeben. Wir geben das Verfahren (siehe E. W. Cheney (1966, S. 171)) im folgenden Satz an und machen außerdem eine Konvergenzaussage.

Satz 5.2.1 *Man betrachte den folgenden Algorithmus zur Lösung der verallgemeinerten rationalen T-Approximationsaufgabe.*

- *Wähle einen Startwert* $r_0 = p_0/q_0 \in R$.
- *Für* $k = 0, 1, \ldots$:

– *Berechne* $\Delta_k := \|r_k - z_k\|_\infty$.

– *Berechne eine Lösung* $(p_{k+1}, q_{k+1}) \in P \times Q$ *der Aufgabe*

$$(P_k) \qquad \begin{cases} \text{Minimiere} \quad \delta_k(p,q) := \max\limits_{t\in B}(|p(t) - z(t)q(t)| - \Delta_k q(t)) \\ \text{auf} \quad \{(p,q) \in P \times Q : \|q\|_\infty = 1\}. \end{cases}$$

– *Falls* $\min_{t\in B} q_{k+1}(t) \le 0$, *dann: STOP, es ist* $r_k \in P_R(z)$.

– *Falls* $\delta_k(p_{k+1}, q_{k+1}) = 0$, *dann: STOP, es ist* $r_k \in P_R(z)$.

– *Setze* $r_{k+1} := p_{k+1}/q_{k+1}$.

Dann gilt (siehe z. B. E. W. Cheney *(1966, S. 171)):*

1. *Das Verfahren ist durchführbar.*
2. *Bricht das Verfahren nicht vorzeitig mit einer Lösung ab, so liefert es eine Folge* $\{r_k\} \subset R$ *mit der Eigenschaft, dass* $\Delta_k = \|r_k - z\|_\infty \searrow d(z, R)$.
3. *Ist* $P_R(z) \neq \emptyset$, *existiert also eine beste Approximierende an* z *in* R, *so ist die Konvergenz von* $\{\Delta_k\}$ *gegen* $d(z, R)$ *mindestens linear, es existiert also eine Konstante* $\vartheta \in (0, 1)$ *mit*

$$\Delta_{k+1} - d(z,R) \le \vartheta(\Delta_k - d(z,R)), \qquad k = 0, 1, \ldots.$$

Beweis: 1. Um die Durchführbarkeit zu zeigen, müssen wir nachweisen, dass (P_k) lösbar ist und die im Algorithmus gemachten Aussagen zum Abbruch des Verfahrens zu Recht bestehen.

(a) Die Aufgabe (P_k) ist lösbar.

Denn: Wähle $(\hat{p}, \hat{q}) \in P \times Q$ mit $\hat{q}\|_\infty = 1$ beliebig und definiere die Niveaumenge

$$\hat{W}_k := \{(p,q) \in P \times Q : \|q\|_\infty = 1,\ \delta_k(p,q) \le \delta_k(\hat{p}, \hat{q})\}.$$

Dann ist $\hat{W}_k$ offensichtlich abgeschlossen, aber auch beschränkt. Ist nämlich $(p, q) \in \hat{W}_k$, so ist

$$|p(t) - z(t)q(t)| - \Delta_k q(t) \le \delta_k(\hat{p}, \hat{q}) \quad \text{für alle } t \in B$$

und daher

$$\begin{aligned} |p(t)| - \|z\|_\infty &\le |p(t)| - |z(t)|\,|q(t)| \\ &\le |p(t) - z(t)q(t)| \\ &\le \Delta_k q(t) + \delta_k(\hat{p}, \hat{q}) \\ &\le \Delta_k + \delta_k(\hat{p}, \hat{q}), \end{aligned}$$

also

$$\|p\|_\infty \le \|z\|_\infty + \Delta_k + \delta_k(\hat{p}, \hat{q}), \qquad \|q\|_\infty = 1.$$

Damit ist $\hat{W}_k$ abgeschlossen und beschränkt in dem endlichdimensionalen Raum $P \times Q$, also kompakt. Die stetige Funktion $\delta_k(\cdot, \cdot)$ nimmt auf $\hat{W}_k$ ihr Minimum an, womit (a) bewiesen ist.

(b) Ist $\min_{t\in B} q_{k+1}(t) \le 0$, so ist $r_k \in P_R(z)$.

Denn: Es wird hierbei natürlich angenommen, dass $r_k = p_k/q_k \in R$, dass also $q_k(t) > 0$ für alle $t \in B$. Angenommen, die Behauptung sei nicht richtig, es würde also ein $r = p/q \in R$ mit $\|q\|_\infty = 1$ und

$$|r(t) - z(t)| \le \|r - z\|_\infty < \|r_k - z\|_\infty = \Delta_k \quad \text{für alle } t \in B$$

existieren. Dann ist

$$\begin{aligned}
\delta_k(p_{k+1}, q_{k+1}) &\le \delta_k(p, q) \\
&= \max_{t\in B}(|p(t) - z(t)q(t)| - \Delta_k q(t)) \\
&= \max_{t\in B}[(\underbrace{|r(t) - z(t)| - \Delta_k}_{<0})\underbrace{q(t)}_{>0}] \\
&< 0.
\end{aligned}$$

Nimmt q_{k+1} ein Minimum auf B in $t_0 \in B$ an, ist also

$$q_{k+1}(t_0) = \min_{t\in B} q_{k+1}(t) \le 0,$$

so ist

$$\begin{aligned}
\delta_k(p_{k+1}, q_{k+1}) &= \max_{t\in B}(|p_{k+1}(t) - z(t)q_{k+1}(t)| - \Delta_k q_{k+1}(t)) \\
&\ge |p_{k+1}(t_0) - z(t_0)q_{k+1}(t_0)| - \Delta_k \underbrace{q_{k+1}(t_0)}_{\le 0} \\
&\ge 0.
\end{aligned}$$

Wir haben einen Widerspruch erhalten, womit (b) bewiesen ist.

(c) Es ist $\delta_k(p_{k+1}, q_{k+1}) \le 0$. Ist $\delta_k(p_{k+1}, q_{k+1}) = 0$, so ist $r_k \in P_R(z)$.

Denn: Wir brauchen nur wie beim Beweis von (b) vorzugehen. Denn es ist

$$\begin{aligned}
\delta_k(p_{k+1}, q_{k+1}) &\le \delta_k(p_k, q_k) \\
&= \max_{t\in B}[(\underbrace{|r_k(t) - z(t)| - \Delta_k}_{\le 0})\underbrace{q_k(t)}_{>0}] \\
&\le 0.
\end{aligned}$$

Ist $r_k \notin P_R(z)$, so ist $\delta_k(p_{k+1}, q_{k+1}) < 0$, wie wir im ersten Teil des Beweises von (b) nachgewiesen haben. Damit ist auch (c) bewiesen.

2. Das Verfahren breche nicht vorzeitig ab, liefere also eine Folge $\{r_k\} = \{p_k/q_k\} \subset R$ mit $\delta_k(p_{k+1}, q_{k+1}) < 0$, $k = 0, 1, \ldots$. Dann ist

$$\begin{aligned}
0 &> \delta_k(p_{k+1}, q_{k+1}) \\
&= \max_{t\in B}[(|r_{k+1}(t) - z(t)| - \Delta_k)\underbrace{q_{k+1}(t)}_{\in(0,1]}] \\
&\ge \max_{t\in B}(|r_{k+1}(t) - z(t)| - \Delta_k) \\
&= \Delta_{k+1} - \Delta_k.
\end{aligned}$$

Also ist $\{\Delta_k\}$ eine monoton fallende, nach unten durch $d(z,R)$ beschränkte Folge, daher konvergent. Es sei etwa $\Delta_k \to L \geq d(z,R)$. Wäre $L > d(z,R)$, so existiert ein $r = p/q \in R$ mit $\|r - z\|_\infty < L$, wobei o. B. d. A. $\|q\|_\infty = 1$. Dann ist

$$|r(t) - z(t)| \leq \|r - z\|_\infty < L \leq \Delta_k \quad \text{für alle } t \in B.$$

Mit $\alpha := \min_{t\in B} q(t)$ ist daher

$$\begin{aligned} \delta_k(p_{k+1}, q_{k+1}) &\leq \delta_k(p,q) \\ &= \max_{t\in B}[(|r(t) - z(t)| - \Delta_k)q(t)] \\ &\leq \alpha \max_{t\in B}(|r(t) - z(t)| - \Delta_k) \\ &= \alpha(\|r - z\|_\infty - \Delta_k) \end{aligned}$$

und folglich

$$\begin{aligned} \Delta_{k+1} &\leq \delta_k(p_{k+1}, q_{k+1}) + \Delta_k \\ &\leq \alpha\,(\|r - z\|_\infty - \Delta_k) + \Delta_k. \end{aligned}$$

Mit $k \to \infty$ ist

$$L \leq \underbrace{\alpha}_{>0}\,(\underbrace{\|r - z\|_\infty - \Delta_k}_{>0}) + L,$$

ein Widerspruch. Damit ist $\Delta_k \searrow d(z,R)$ nachgewiesen.
3. Sei $r^* = p^*/q^* \in P_R(z)$, o. B. d. A. ist wieder $\|q^*\|_\infty = 1$. Wie gerade eben erhalten wir

$$\delta_k(p_{k+1}, q_{k+1}) \leq \delta_k(p^*, q^*) \leq \alpha^*\,(d(z,R) - \Delta_k)$$

mit $\alpha^* := \min_{t\in B} q^*(t) \in (0,1]$. Hieraus folgt

$$\Delta_{k+1} - \Delta_k \leq \delta_k(p_{k+1}, q_{k+1}) \leq \alpha^*\,(d(z,R) - \Delta_k)$$

und damit

$$\begin{aligned} \Delta_{k+1} - d(z,R) &= \Delta_{k+1} - \Delta_k + \Delta_k - d(z,R) \\ &\leq \alpha^*\,(d(z,R) - \Delta_k) + \Delta_k - d(z,R) \\ &= \vartheta\,(\Delta_k - d(z,R)) \end{aligned}$$

mit $\vartheta := 1 - \alpha^* \in (0,1)$. Damit ist die lineare Konvergenz von $\{\Delta_k\}$ gegen $d(z,R)$ bewiesen. □

Bemerkungen: 1. Der Name "Differential Correction"-Algorithmus sollte erklärt bzw. die Vorgehensweise des Verfahrens motiviert werden. Es wird sich herausstellen, dass das Differential Correction Verfahren, wie viele gute Verfahren, mit dem Newton-Verfahren verwandt ist. Die Aufgabe, $r = p/q \in P_R(z)$ zu bestimmen, schreibe man als Optimierungsaufgabe:

$$\text{Minimiere} \quad f(\Delta, p, q) := \Delta \quad \text{auf}$$

$$M := \left\{ (\Delta, P, Q) \in \mathbb{R} \times P \times Q : \begin{array}{c} -\Delta q(t) \leq p(t) - z(t)q(t) \leq \Delta q(t)\ \forall\, t \in B, \\ \|q(t) > 0\ \forall\, t \in B,\ \|q\|_\infty = 1. \end{array} \right\}$$

Dies ist eine *nichtlineare* Optimierungsaufgabe (der Term $\Delta \cdot q$ ist nichtlinear). Es liegt also nahe, die "Philosophie" des Newton-Verfahrens zu übernehmen und das Problem in einer aktuellen Näherung (Δ_k, p_k, q_k) zu linearisieren. Linearisiert man die Restriktion

$$|p(t) - z(t)q(t)| \leq \Delta q(t)$$

in (Δ_k, p_k, q_k), so erhält man

$$|p(t) - z(t)q(t)| \leq \Delta q(t) = \Delta_k q_k(t) + (\Delta - \Delta_k)q_k(t) + (q(t) - q_k(t))\Delta_k + \cdots .$$

Umordnen ergibt

$$|p(t) - z(t)q(t)| - \Delta_k q(t) \leq (\Delta - \Delta_k)q_k(t) + \cdots .$$

Daher liegt es nahe, $(p_{k+1}, q_{k+1}) \in P \times Q$ als Lösung der Aufgabe

$$(P_k) \qquad \begin{cases} \text{Minimiere} \quad \delta_k(p,q) := \max\limits_{t \in B}(|p(t) - z(t)q(t)| - \Delta_k q(t)) \\ \text{auf} \quad \{(p,q) \in P \times Q : \|q\|_\infty = 1\} \end{cases}$$

zu bestimmen. Unter der Voraussetzung, dass $q_k(t) > 0$ für alle $t \in B$ ist, bietet es sich außerdem an, $(p_{k+1}, q_{k+1}) \in P \times Q$ als Lösung von

$$(Q_k) \qquad \begin{cases} \text{Minimiere} \quad \eta_k(p,q) := \max\limits_{t \in B}(|p(t) - z(t)q(t)| - \Delta_k q(t))/q_k(t) \\ \text{auf} \quad \{(p,q) \in P \times Q : \|q\|_\infty = 1\} \end{cases}$$

zu berechnen. Dies ist das "originale" differential correction Verfahren von Cheney-Loeb (siehe E. W. Cheney, H. L. Loeb (1961)). Konvergenzaussagen zu diesem Verfahren findet man bei E. W. Cheney, M. J. D. Powell (1987).

2. Unklar ist natürlich, wie das Hilfsproblem im obigen Algorithmus gelöst werden kann. Ist B diskret, so führt das Hilfsproblem auf eine lineare Optimierungsaufgabe, andernfalls auf ein sogenanntes semiinfinites Programm. Hierauf wollen wir aber nicht mehr eingehen. □

Sei nun $B := [\alpha, \beta]$, $P := \Pi_m$, $Q := \Pi_n$ und daher $R = R_{m,n}[\alpha, \beta]$. Wir nehmen an, $r^* \in P_{R_{m,n}[\alpha,\beta]}(z)$ sei normal und daher wegen Satz 5.1.11 stark eindeutig, d. h. es existiert eine Konstante $c > 0$ mit

$$\|r - z\|_\infty \geq \|r^* - z\|_\infty + c\,\|r^* - r\|_\infty \quad \text{für alle } r \in R_{m,n}[\alpha, \beta].$$

Wir nehmen an, das Verfahren (genauer: modifiziertes differential correction Verfahren) aus Satz 5.2.1 breche nicht vorzeitig ab und liefere daher eine Folge $\{r_k\} \subset R_{m,n}[\alpha, \beta]$. Mit $\Delta_k := \|r_k - z\|_\infty$ und einer Konstanten $\vartheta \in [0, 1)$ ist

$$\Delta_{k+1} - d(z, R_{m,n}[\alpha, \beta]) \leq \vartheta(\Delta_k - d(z, R_{m,n}[\alpha, \beta]))$$

wegen Satz 5.2.1 3. Dies ergibt

$$\Delta_k - d(z, R_{m,n}[\alpha, \beta]) \leq \vartheta^k(\Delta_0 - d(z, R_{m,n}[\alpha, \beta])).$$

Wegen der starken Eindeutigkeit von r^* ist daher

$$\begin{aligned}\|r_k - r^*\|_\infty &\leq \frac{1}{c}(\|r_k - z\|_\infty - \|r^* - z\|_\infty) \\ &= \frac{1}{c}(\Delta_k - d(z, R_{m,n}[\alpha,\beta])) \\ &\leq \frac{1}{c}(\Delta_0 - d(z, R_{m,n}[\alpha,\beta]))\vartheta^k.\end{aligned}$$

Wir haben also den folgenden Satz (siehe auch E. W. CHENEY (1966, S. 172)) bewiesen:

Satz 5.2.2 *Sei $r^* \in P_{R_{m,n}[\alpha,\beta]}(z)$ normal. Bricht das Verfahren aus Satz 5.2.1 nicht vorzeitig mit der besten Approximierenden r^* an z in $R_{m,n}[\alpha,\beta]$ ab, so liefert es eine Folge $\{r_k\} \subset R_{m,n}[\alpha,\beta]$ mit $\|r_k - z\|_\infty \leq A\vartheta^k$, $k = 0, 1, \ldots$, wobei $A > 0$ und $\vartheta \in [0, 1)$ Konstanten sind.*

Einige Bemerkungen wollen wir nun noch zum *Remez-Verfahren* für rationale T-Approximationsaufgaben machen. Der Alternantensatz der rationalen T-Approximation (siehe Satz 5.1.10) suggeriert ein Verfahren, das dem Remez-Verfahren der linearen T-Approximation analog ist. Es besteht allerdings die Schwierigkeit, dass die Länge der Alternante nicht a priori bekannt ist. Wir werden uns das Leben hier aber verhältnimäßig einfach machen und *annehmen*, dass $r^* = P_{R_{m,n}[\alpha,\beta]}(z)$ normal ist. Dann wissen wir (Satz 5.1.10):

- Ein $r^* \in R_{m,n}[\alpha,\beta]$ ist genau dann die beste Approximierende an z in $R_{m,n}[\alpha,\beta]$, wenn $t_i \in [\alpha,\beta]$, $i = 0, \ldots, N := m + n + 1$, existieren mit

  (a) $\alpha \leq t_0 < \cdots < t_N \leq \beta$,

  (b) $|r^*(t_i) - z(t_i)| = \|r^* - z\|_\infty$, $i = 0, \ldots, N$,

  (c) $r^*(t_{i+1}) - z(t_{i+1}) = -(r^*(t_i) - z(t_i))$, $i = 0, \ldots, N-1$.

Der zweite Remez-Algorithmus (Simultan-Austausch) zur Lösung rationaler T-Approximationsaufgaben sieht dann im Prinzip folgendermaßen aus:

- Setze $N := m + n + 1$. Wähle Startreferenz $T_0 = (t_0^{(0)}, \ldots, t_N^{(0)})$ mit

$$\alpha \leq t_0^{(0)} < t_1^{(0)} < \cdots < t_N^{(0)} \leq \beta.$$

- Für $k = 0, 1, \ldots$:

  - Bestimme $\rho_k \in \mathbb{R}$ und $r_k \in R_{m,n}[\alpha,\beta]$ mit

$$r_k(t_i^{(k)}) + (-1)^i\rho_k = z(t_i^{(k)}), \qquad i = 0, \ldots, N.$$

  - Falls $\rho_k = \|r_k - z\|_\infty$, dann: STOP. Es ist $r_k = P_{R_{m,n}[\alpha,\beta]}(z)$.

– Bestimme eine neue Referenz $T_{k+1} = (t_0^{(k+1)}, \ldots, t_N^{(k+1)})$ mit

$$\alpha \le t_0^{(k+1)} < t_1^{(k+1)} < \cdots < t_N^{(k+1)} \le \beta$$

und

1. Bei $t_i^{(k+1)}$ ist ein lokales Extremum von $r_k - z$.
2. Es existiert ein $t_{i_k}^{(k+1)}$ mit $|r_k(t_{i_k}^{(k+1)}) - z(t_{i_k}^{(k+1)})| = \|r_k - z\|_\infty$.
3. $r_k - z$ alterniert im Vorzeichen in den $t_i^{(k+1)}$.

Während die Bestimmung einer neuen Referenz T_{k+1} im Prinzip keine Schwierigkeiten macht und genau wie im linearen Fall durchgeführt werden kann, ist die Bestimmung von ρ_k und $r_k = p_k/q_k \in R_{m,n}[\alpha, \beta]$ mit

$$r_k(t_i^{(k)}) + (-1)^i \rho_k = z(t_i^{(k)}), \qquad i = 0, \ldots, N,$$

sehr viel schwieriger, denn dies führt nicht mehr wie im linearen Fall auf ein lineares Gleichungssystem, sondern, wie wir sehen werden, auf eine Eigenwertaufgabe. Dies wollen wir näher untersuchen, lassen dabei natürlich den Iterationsindex k weg und gehen daher von einer Referenz $T = (t_0, \ldots, t_N)$ aus, wobei nach wie vor $N := m+n+1$ gesetzt ist. Wir folgen im wesentlichen der Darstellung bei M. J. D. Powell (1981, S. 113 ff.). Mit dem Ansatz

$$p(t) = \sum_{j=0}^{m} a_j t^j, \qquad q(t) = \sum_{j=0}^{n} b_j t^j$$

sind $(a, b, \rho) \in \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}$ so zu bestimmen, dass

$$(*) \qquad \begin{cases} \displaystyle\sum_{j=0}^{m} a_j t_i^j + (-1)^i \rho \sum_{j=0}^{n} b_j t_i^j = \left(\sum_{j=0}^{n} b_j t_i^j \right) z(t_i), \qquad i = 0, \ldots, N, \\ \displaystyle\sum_{j=0}^{n} b_j t^j > 0 \quad \text{für alle } t \in [\alpha, \beta]. \end{cases}$$

Nun zeigen wir:

- Es ist

$$\sum_{i=0}^{N} t_i^k \prod_{\substack{j=0 \\ j \ne i}}^{N} \frac{1}{t_j - t_i} = 0, \qquad k = 0, 1, \ldots, N-1.$$

Denn: Sei $f\colon [\alpha, \beta] \longrightarrow \mathbb{R}$. Das Lagrangesche Interpolationspolynom $p \in \Pi_N$, das f an den Stellen $t_0, \ldots, t_N$ interpoliert, hat die Darstellung

$$p(t) = \sum_{i=0}^{N} f(t_i) l_i(t) \qquad \text{mit} \qquad l_i(t) := \prod_{\substack{j=0 \\ j \ne i}}^{N} \frac{t - t_j}{t_i - t_j}, \quad i = 0, \ldots, N.$$

Für $f(t) := t^k$, $k = 0, \dots, N$, stimmen f und das zugehörige Interpolationspolynom überein, es ist also

$$t^k = \sum_{i=0}^{N} t_i^k \prod_{\substack{j=0 \\ j\neq i}}^{N} \frac{t - t_j}{t_i - t_j}, \qquad k = 0, \dots, N.$$

Für $k = 0, \dots, N-1$ vergleiche man in dieser Gleichung nun links und rechts die Koeffizienten von t^N. Man erhält

$$0 = \sum_{i=0}^{N} t_i^k \prod_{\substack{j=0 \\ j\neq i}}^{N} \frac{1}{t_i - t_j},$$

eine Multiplikation mit $(-1)^N$ liefert die Behauptung.
Nun multipliziere man die Gleichung

$$(z(t_i) - (-1)^i \rho) \sum_{j=0}^{n} b_j t_i^j = \sum_{j=0}^{m} a_j t_i^j$$

mit $t_i^k \prod_{\substack{s=0 \\ s\neq i}}^{N} 1/(t_s - t_i)$, $i = 0, \dots, N$, und summiere anschließend über i von 0 bis N. Dann erhält man

$$\begin{aligned} \sum_{i=0}^{N} (z(t_i) - (-1)^i \rho) \Big(\sum_{j=0}^{n} b_j t_i^{j+k} \Big) \prod_{\substack{s=0 \\ s\neq i}}^{N} \frac{1}{t_s - t_i} &= \sum_{i=0}^{N} \sum_{j=0}^{m} a_j t_i^{j+k} \prod_{\substack{s=0 \\ s\neq i}}^{N} \frac{1}{t_s - t_i} \\ &= \sum_{j=0}^{m} a_j \underbrace{\sum_{i=0}^{N} t_i^{j+k} \prod_{\substack{s=0 \\ s\neq i}}^{N} \frac{1}{t_s - t_i}}_{=0,\ k=0,\dots,n}. \end{aligned}$$

Also ist

$$\sum_{i=0}^{N} (z(t_i) - (-1)^i \rho) \Big(\sum_{j=0}^{n} b_j t_i^{j+k} \Big) \prod_{\substack{s=0 \\ s\neq i}}^{N} \frac{1}{t_s - t_i} = 0, \qquad k = 0, \dots, n,$$

bzw.

$$\sum_{j=0}^{n} \underbrace{\Big(\sum_{i=0}^{N} z(t_i) t_i^{j+k} \prod_{\substack{s=0 \\ s\neq i}}^{N} \frac{1}{t_s - t_i} \Big)}_{=:A_{kj}} b_j = \rho \sum_{j=0}^{n} \underbrace{\Big(\sum_{i=0}^{N} (-1)^i t_i^{j+k} \prod_{\substack{s=0 \\ s\neq i}}^{N} \frac{1}{t_s - t_i} \Big)}_{=:B_{kj}} b_j, \quad k = 0, \dots, n.$$

Führt man dann die $(n+1) \times (n+1)$-Matrizen $A := (A_{kj})$, $B := (B_{kj})$ ein, so hat man also die Aufgabe $(b, \rho) \in \mathbb{R}^{n+1} \times \mathbb{R}$ zu bestimmen mit

$$(**) \qquad Ab = \rho B b, \quad \sum_{j=0}^{n} b_j t^j > 0 \qquad \text{für alle } t \in [\alpha, \beta].$$

Offensichtlich sind A und B symmetrisch. Weiter ist B positiv definit. Denn sei $c = (c_0, \ldots, c_n)^T \in \mathbb{R}^{n+1}$. Dann ist

$$\begin{aligned}
c^T B c &= \sum_{k=0}^{n}\sum_{j=0}^{n} c_k c_j B_{kj} \\
&= \sum_{k=0}^{n}\sum_{j=0}^{n} c_k c_j \sum_{i=0}^{N}(-1)^i t_i^{j+k} \prod_{\substack{s=0\\ s\neq i}}^{N} \frac{1}{t_s - t_i} \\
&= \sum_{i=0}^{N}\sum_{k=0}^{n}\sum j = 0^n (c_k t_i^k)(c_j t_i^j)(-1)^i \prod_{\substack{s=0\\ s\neq i}}^{N} \frac{1}{t_s - t_i} \\
&= \sum_{i=0}^{N}\left(\sum_{j=0}^{n} c_j t_i^j\right)^2 (-1)^i \prod_{\substack{s=0\\ s\neq i}}^{N} \frac{1}{t_s - t_i} \\
&= \sum_{i=0}^{N}\left(\sum_{j=0}^{n} c_j t_i^j\right)^2 \prod_{\substack{s=0\\ s\neq i}}^{N} \frac{1}{|t_s - t_i|} \\
&> 0 \quad \text{für } c \neq 0,
\end{aligned}$$

also ist B positiv definit. Die verallgemeinerte Eigenwertaufgabe $Ab = \rho B b$ ist also äquivalent zu der Eigenwertaufgabe $B^{-1/2}AB^{-1/2}b = \rho b$. Da die Matrix $B^{-1/2}AB^{-1/2}$ symmetrisch ist, sind die $n+1$ Eigenwerte ρ reell. Zu jedem ρ gibt es einen zugehörigen Eigenvektor b, hiermit kann man $q(t) = \sum_{j=0}^{n} b_j t^j$ bilden. Angenommen, man hat einen Eigenwert ρ und einen zugehörigen Eigenvektor b so bestimmt, dass $\sum_{j=0}^{n} b_j t^j > 0$ für alle $t \in]\alpha, \beta]$. Dann kann man den Vektor $a = (a_0, \ldots, a_n)^T$ und damit $p(t) = \sum_{j=0}^{m} a_j t^j$ aus dem linearen Gleichungssystem

$$\sum_{j=0}^{m} a_j t_i^j = (z(t_i) - (-1)^i \rho)\left(\sum_{j=0}^{n} b_j t_i^j\right), \qquad i = 0, \ldots, m,$$

bestimmen. Also ist p dajenige Polynom aus Π_m, das durch die Interpolationsbedingung, an der Stelle t_i den Wert $(z(t_i) - (-1)^i \rho) q(t_i)$ zu haben, festgelegt ist. Nun stellt sich die Frage: Gibt es einen oder mehrere Eigenwerte ρ von $B^{-1/2}AB^{-1/2}$ derart, dass für einen zugehörigen Eigenvektor b gilt, dass $\sum_{j=0}^{n} b_j t^j > 0$ für alle $t \in [\alpha, \beta]$? Nun gilt:

- Sind $r, r^* \in R_{m,n}[\alpha, \beta]$ mit $r = p/q$, $r^* = p^*/q^*$, und ist

$$\begin{cases} r(t_i) - (-1)^i \rho = z(t_i), \qquad r^*(t_i) - (-1)^i \rho^* = z(t_i), \\ \qquad\qquad (i = 0, \ldots, m+n+1), \end{cases}$$

  so ist $r = r^*$ und $\rho = \rho^*$.

Denn: Es ist

$$r(t_i) - r^*(t_i) = (-1)^i(\rho - \rho^*), \qquad i = 0, \ldots, m+n+1.$$

Daher besitzt $r - r^*$ mindestens $m+n+1$ Nullstellen. Andererseits ist

$$r - r^* = \frac{p}{q} - \frac{p^*}{q^*} = \frac{pq^* - qp^*}{qq^*},$$

wobei $pq^* - qp^* \in \Pi_{m+n}$ das Nullpolynom ist oder höchstens $m+n$ Nullstellen in $[\alpha, \beta]$ besitzt. Folglich ist $r = r^*$ und dann auch $\rho = \rho^*$. Daher gibt es *höchstens* ein Paar $(r, \rho) \in R_{m,n}[\alpha, \beta] \times \mathbb{R}$ mit $r(t_i) - (-1)^i\rho = z(t_i)$, $i = 0, \ldots, m+n+1$. Leider kann man nicht zeigen, dass es *genau* eine Lösung liegt, was daran liegt, dass diskrete rationale T-Approximationsaufgaben nicht lösbar zu sein brauchen.

Beispiel: Sei $[\alpha, \beta] := [-1, 1]$, $z(t) := t$, $m := 0$, $n := 1$ und $(t_0, t_1, t_2) := (-1, 0, 1)$. Dann ist

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}.$$

Zu bestimmen sind also die Eigenwerte von

$$\begin{pmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 \end{pmatrix}$$

und dies sind $\rho_1 = 1/\sqrt{2}$, $\rho_2 = -1/\sqrt{2}$ mit zugehörigen Eigenvektoren $b_1 = (1, 1)^T$, $b_2 = (1, -1)^T$. Aber $1 + t$ und $1 - t$ haben jeweils eine Nullstelle in $[-1, 1]$. □

Kapitel 6

T-Approximation mit Exponentialsummen

6.1 Der Existenzsatz für die Exponentialapproximation

Die rationalen Funktionen und die Exponentialsummen gehören zu den am meisten in der nichtlinearen Approximationstheorie untersuchten konkreten Funktionenfamilien.
Bei der Untersuchung von Abklingvorgängen versucht man, eine gegebene Funktion z durch reine Exponentialsummen bzw. Elemente aus

$$E_n^0 := \left\{ u : u(t) = \sum_{i=1}^{n} a_i e^{\lambda_i t},\ a_i,\ \lambda_i \in \mathbb{R}\ (i = 1, \ldots, n) \right\}$$

zu approximieren, in der T-Approximation natürlich bezüglich der Maximumnorm $\|\cdot\|_\infty$ auf einem Intervall $[\alpha, \beta]$. Nun stellt man aber leicht fest, dass E_n^0 *keine* Existenzmenge in $(C[\alpha, \beta], \|\cdot\|_\infty)$ ist.
Beispiel: Wir geben ein $z \in C[0,1] \setminus E_2^0$ mit $d(z, E_2^0) = 0$ an (siehe G. Meinardus (1967, S. 178)). Dies zeigt, dass E_2^0 keine Existenzmenge ist.
Sei $z(t) := te^t$. Definiere $u_k \in E_2^0$ durch

$$u_k(t) := -ke^t + ke^{1+1/k)t}, \qquad k = 1, 2, \ldots.$$

Für $t \in [0,1]$ ist

$$\begin{aligned} 0 &\leq u_k(t) - z(t) \\ &= e^t(ke^{t/k} - k - t) \\ &= e^t k \sum_{j=2}^{\infty} \frac{t^j}{k^j j!} \\ &\leq \frac{e}{k} \sum_{j=2}^{\infty} \frac{1}{j!} \\ &= \frac{e(e-2)}{k}. \end{aligned}$$

Also ist $\lim_{k\to\infty} \|u_k - z\|_\infty = 0$ und daher $d(z, E_2^0) = 0$. Da $z \notin E_2^0$ ist E_2^0 keine Existenzmenge. □

Es wird die Menge der erweiterten (extended) Exponentialsummen oder nur Exponentialsummen betrachtet:

$$E_n := \left\{ u : u(t) = \sum_{i=1}^{l} p_i(t) e^{\lambda_i t},\ p_i \in \Pi,\ \lambda_i \in \mathbb{R}\ (i = 1, \ldots, l),\ \sum_{i=1}^{l} (1 + \partial p_i) \le n \right\}.$$

Für $u \in E_n$ mit $u(t) = \sum_{i=1}^{l} p_i(t) e^{\lambda_i t}$ kann angenommen werden, dass $\lambda_1 < \cdots < \lambda_l$. Dann heißt $l = l(u)$ die *Länge* von u, $k = k(u) = \sum_{i=1}^{l}(1 + \partial p_i)$ die *Ordnung* von u. Unser Ziel in diesem Abschnitt wird es sein nachzuweisen, dass E_n eine Existenzmenge in $(C[\alpha, \beta], \|\cdot\|_\infty)$ ist, eine Aussage, die von H. Werner (1969) als erstem bewiesen und dessen Beweis von E. Schmidt (1970) vereinfacht wurde. Trotzdem wird der Beweis dieses Existenzsatzes immer noch ziemlich aufwendig sein und daher ist es nützlich, eine Art Fahrplan durch den Beweis aufzustellen.

1. Sei $z \in C[\alpha, \beta]$ vorgegeben. Eine Minimalfolge $\{u_k\} \subset E_n$ (d. h. $\|u_k - z\|_\infty \to d(z, E_n)$) ist natürlich beschränkt, d. h. es existiert eine Konstante $K > 0$ mit $\|u_k\|_\infty \le K$ für alle $k \in \mathbb{N}$. Hierbei ist $\|\cdot\|_\infty$ stets die Maximumnorm bezüglich des Grundintervalls $[\alpha, \beta]$.

2. Sei $\{u_k\} \subset E_{n,K} := \{u \in E_n : \|u\|_\infty \le K\}$. Dann existiert eine Teilfolge $\{u_{k_i}\} \subset \{u_k\}$ und ein $u^* \in E_n$ mit der Eigenschaft, dass $\{u_{k_i}\}$ auf jedem Teilintervall[1] $[a, b]$ mit $\alpha < a < b < \beta$ gleichmäßig gegen u^* konvergiert. Insbesondere konvergiert $\{u_{k_i}\}$ punktweise auf (α, β) gegen u^*.

3. Man wende 2. auf die Minimalfolge $\{u_k\}$ aus 1. an. Hiernach existiert insbesondere eine Teilfolge $\{u_{k_i}\} \subset \{u_k\}$, die punktweise auf (α, β) gegen ein $u^* \in E_n$ konvergiert. Bei festem $t \in (\alpha, \beta)$ ist also

$$\begin{aligned} |u^*(t) - z(t)| &= \lim_{i\to\infty} |u_{k_i}(t) - z(t)| \\ &\le \lim_{i\to\infty} \|u_{k_i} - z\|_\infty \\ &= d(z, E_n). \end{aligned}$$

   Da $u^* - z \in C[\alpha, \beta]$, ist $|u^*(t) - z(t)| \le d(z, E_n)$ für *alle* $t \in [\alpha, \beta]$, also $\|u^* - z\| \le d(z, E_n)$ und folglich $\|u^* - z\|_\infty = d(z, E_n)$. Daher ist u^* eine beste Approximierende an z in E_n.

Die Arbeit beim Existenzbeweis steckt also im Nachweis von 2. Hierzu sind einige Vorbereitungen nötig, die zum Teil auch für sich von Interesse sind und mit der Exponentialsummenapproximation zunächst gar nichts zu tun haben.

[1] Man kann nicht erwarten, dass eine auf dem *ganzen* Intervall $[\alpha, \beta]$ gleichmäßig konvergente Teilfolge auswählbar ist. Sei z. B. $[\alpha, \beta] = [0, 1]$ und $u_k(t) = e^{-k(1-t)}$. Dann gilt

$$u_k(t) \to \begin{cases} 1, & t = 1, \\ 0, & t \in [0, 1). \end{cases}$$

Satz 6.1.1 (Vergleichssatz der Interpolation) *Seien $y, z \in C^{n+1}[\alpha, \beta]$ mit*

$$|z^{(n+1)}(t)| \leq y^{(n+1)}(t) \quad \textit{für alle } t \in [\alpha, \beta].$$

Zu vorgegebenen $n+1$ Punkten $t_0 < t_1 < \cdots < t_n$ in $[\alpha, \beta]$ seien $L_n(z)$ bzw. $L_n(y)$ aus Π_n die zugehörigen Interpolationspolynome. Dann ist

$$|z(t) - L_n(z)(t)| \leq |y(t) - L_n(y)(t)| \quad \textit{für alle } t \in [\alpha, \beta].$$

Beweis: Wir nehmen zunächst an, dass $y^{(n+1)}(t) > 0$ für alle $t \in [\alpha, \beta]$. Angenommen, es gibt einen Punkt $\hat{t} \in [\alpha, \beta]$ mit$|z(\hat{t}) - L_n(z)(\hat{t})| > /y(\hat{t}) - L_n(y)(\hat{t})|$. Definiere

$$\lambda := \frac{y(\hat{t}) - L_n(y)(\hat{t})}{z(\hat{t}) - L_n(z)(\hat{t})}.$$

Dann ist $|\lambda| < 1$. Die Hilfsfunktion

$$h := y - L_n(y) - \lambda(z - L_n(z))$$

hat die $n+2$ Nullstellen $t_0, \ldots, t_n$ und $\hat{t}$ (notwendig von den t_i verschieden). Eine wiederholte Anwendung des Satzes von Rolle zeigt, dass $h^{(n+1)}$ eine Nullstelle besitzt. Andererseits ist

$$h^{(n+1)}(t) = y^{(n+1)}(t) - \lambda z^{(n+1)}(t) \geq (1 - |\lambda|) y^{(n+1)}(t) > 0 \quad \text{für alle } t \in [\alpha, \beta],$$

ein Widerspruch. Ohne die Voraussetzung $y^{(n+1)} > 0$ erhält man die Behauptung durch ein Störungsargument. Definiere $y_\epsilon(t) := y(t) + \epsilon t^{n+1}$ mit $\epsilon > 0$. dann ist $y_\epsilon^{(n+1)} = y^{(n+1)} + \epsilon(n+1)! > 0$ und

$$|z^{(n+1)}(t)| \leq y^{(n+1)}(t) < y_\epsilon^{(n+1)}(t) \quad \text{für alle } t \in [\alpha, \beta].$$

Nach dem ersten Teil ist

$$|z(t) - L_n(z)(t)| \leq |y_\epsilon(t) - L_n(y_\epsilon)(t)| \quad \text{für alle } t \in [\alpha, \beta].$$

Da $L_n(y_\epsilon)$ stetig von ϵ abhängt, folgt die Behauptung mit $\epsilon \to 0+$. □

Eine leichte Folgerung ist der folgende Satz.

Satz 6.1.2 (Vergleichssatz der T-Approximation) *Seien $y, z \in C^{n+1}[\alpha, \beta]$ mit*

$$|z^{(n+1)}(t)| \leq y^{(n+1)}(t) \quad \textit{für alle } t \in [\alpha, \beta].$$

Dann ist $d(z, \Pi_n) \leq d(y, \Pi_n)$.

Beweis: Da zu $P_{\Pi_n}(y)$ eine Alternante der Länge $n+2$ existiert, besitzt $P_{\Pi_n}(y) - y$ mindestens $n+1$ Nullstellen in $[\alpha, \beta]$. Daher ist $P_{\Pi_n}(y)$ Interpolationspolynom bezüglich $n+1$ dieser Nullstellen. Sei $L_n(z)$ das zu z und diesen Nullstellen gehörende Interpolationspolynom. Nach dem Vergleichssatz der Interpolation 6.1.1 ist

$$|z(t) - L_n(z)(t)| \leq |y(t) - P_{\Pi_n}(y)(t)| \quad \text{für alle } t \in [\alpha, \beta].$$

Dann ist aber

$$d(z, \Pi_n) \leq \|z - L_n(z)\|_\infty \leq \|y - P_{\Pi_n}(y)\|_\infty = d(y, \Pi_n),$$

womit die Behauptung bewiesen ist. □

Eine wichtige Folgerung ist

Satz 6.1.3 (Bernstein) *Sei $z \in C^{n+1}[-1,1]$. Dann gibt es ein $\eta \in [-1,1]$ mit*

$$d(z,\Pi_n) = \frac{1}{2^n(n+1)!}|z^{(n+1)}(\eta)|.$$

Beweis: Sei

$$y(t) := c\frac{t^{n+1}}{(n+1)!}$$

mit noch unbestimmter Konstante c. Dann ist

$$d(y,\Pi_n) = \frac{|c|}{2^n(n+1)!},$$

wie sehr leicht aus Satz 4.2.2 folgt. Nun definiere man

$$a := \min_{t\in[-1,1]} z^{(n+1)}(t), \qquad b := \max_{t\in[-1,1]} z^{(n+1)}(t).$$

Wir machen eine Fallunterscheidung.

1. $a < 0 < b$.

   Sei $c := \max(-a,b)$. Dann ist

   $$|z^{(n+1)}(t)| = \max(-z^{(n+1)}(t), z^{(n+1)}(t)) \le \max(-a,b) = c = y^{(n+1)}(t).$$

   Aus Satz 6.1.2 folgt

   $$0 = \frac{0}{2^n(n+1)!} \le d(z,\Pi_n) \le d(y,\Pi_n) = \frac{c}{2^n(n+1)!}.$$

   Da $z^{(n+1)}$ wegen des Zwischenwertsatzes alle Werte zwischen 0 und c annimmt, folgt die Behauptung.

2. $0 \le a \le b$.

   Indem man zunächst $c := b$ setzt, erhält man

   $$d(z,\Pi_n) \le \frac{b}{2^n(n+1)!}.$$

   Aus Satz 6.1.2 folgt aber auch

   $$\frac{a}{2^n(n+1)!} \le d(z,\Pi_n).$$

   Eine Anwendung des Zwischenwertsatzes liefert wieder die Behauptung.

3. $a \le b \le 0$.

   Man ersetze z durch $-z$ und wende 2. an.

Damit ist der Satz bewiesen. □

Die für uns wichtige Folgerung aus diesen Vergleichssätzen ist der folgende Satz, siehe auch D. Braess (1986, S. 171).

Satz 6.1.4 *Sei I ein abgeschlossenes reelles Intervall der Länge $d > 0$ und $y \in C^m(I)$, $m \geq 1$. Dann gibt es ein $\xi \in I$ mit*

$$|y^{(m)}(\xi)| \leq \frac{2^{2m-1}m!}{d^m}\|y\|_\infty.$$

Hier ist $\|\cdot\|_\infty$ die Maximumnorm bezüglich des Intervalls I, also $\|y\|_\infty := \max_{t\in I}|y(t)|$.

Beweis: Sei $I = [\alpha, \alpha + d]$. Man mache die Variablentransformation

$$s = \frac{2}{d}(t - \alpha) - 1$$

und setze

$$z(s) := y\left(\frac{d}{2}(s+1) + \alpha\right).$$

Dann ist $z \in C^m[-1, 1]$ und

$$z^{(m)}(s) = \left(\frac{d}{2}\right)^m y^{(m)}\left(\frac{d}{2}(s+1) + \alpha\right).$$

Da das Nullpolynom zu Π_{m-1} gehört, ist $d(z, \Pi_{m-1}) \leq \|z\|_\infty = \|y\|_\infty$. Hierbei hätten wir eigentlich genauer $\|z\|_{\infty,[-1,1]}$ statt $\|z\|_\infty$ und $\|y\|_{\infty,I}$ statt $\|y\|_\infty$ schreiben müssen. Wendet man Satz 6.1.3 mit $n := m - 1$ an, so erhält man die Existenz von $\eta \in [-1, 1]$ mit

$$d(z, \Pi_{m-1}) = \frac{1}{2^{m-1}m!}|z^{(m)}(\eta)| = \frac{d^m}{2^{2m-1}m!}\left|y^{(m)}\Big(\underbrace{\frac{d}{2}(\eta+1)+\alpha}_{=:\xi}\Big)\right|.$$

Berücksichtigt man $d(z, \Pi_{m-1}) \leq \|y\|_\infty$, so erhält man die Behauptung. □

Die letzten vier Sätze haben mit Exponentialsummen nichts zu tun. Dies wird nun langsam anders. Zunächst geben wir ein Ergebnis an, das wir im Zusammenhang mit Beispielen zu Haarschen Teilräumen schon einmal kennengelernt haben.

Satz 6.1.5 *Ist $u \in E_n$, also*

$$u(t) = \sum_{i=1}^{l} p_i(t)e^{\lambda_i t}$$

mit $k(u) := \sum_{1=1}^{l}(1 + \partial p_i) \leq n$, so besitzt u höchstens $k(u) - 1$ reelle Nullstellen oder verschwindet identisch.

Beweis: Im Anschluss an Satz 4.1.4 hatten wir in Beispiel 3. gezeigt, dass bei vorgegebenen nichtnegativen ganzen Zahlen $q_1, \ldots, q_l$ und paarweise verschiedenen reellen Zahlen $\lambda_1, \ldots, \lambda_l$ der lineare Raum

$$E_l(q, \lambda) := \left\{u : u(t) = \sum_{i=1}^{l} p_i(t)e^{\lambda_i t},\ p_i \in \Pi_{q_i}\ (i = 1, \ldots, l)\right\}$$

ein $\sum_{i=1}^{l}(1+q_i)$-dimensionaler Haarscher Teilraum von $C(I)$ ist, wobei $I \subset \mathbb{R}$ ein kompaktes Intervall ist. Hieraus folgt die Behauptung. □

Der technisch schwierigste Beitrag zum Existenzbeweis für die T-Approximation mit Exponentialsummen ist der Beweis des folgenden Satzes, siehe D. Braess (1986, S. 171 ff.). Der Deutlichkeit halber gebrauchen wir jetzt die Bezeichnung $\|\cdot\|_{\infty,[a,b]}$ für die Maximumnorm auf dem Intervall $[a,b]$.

Satz 6.1.6 *Sei $n \in \mathbb{N}$. Dann existiert eine nur von n abhängende Konstante $c > 0$ derart, dass*

$$\|u'\|_{\infty,[\alpha+d,\beta-d]} \leq \frac{c}{d}\|u\|_{\infty,[\alpha,\beta]}$$

für alle $u \in E_n$ und alle $d \in (0,(\beta-\alpha)/2]$, $\alpha < \beta$.

Beweis: Wir überlegen uns, dass es genügt, folgendes zu zeigen:

- Es existiert eine nur von n abhängende Konstante $c > 0$ mit

  $$\left|u'\left(\frac{a+b}{2}\right)\right| \leq \frac{2c}{b-a}\|u\|_{\infty,[a,b]} \tag{1}$$

  für alle $u \in E_n$ und alle $a < b$.

Denn angenommen, (1) sei bewiesen. Seien $\alpha < \beta$, $d \in (0,(\beta-\alpha)/2]$ beliebig, ferner sei $t \in [\alpha+d, \beta-d]$ und $u \in E_n$ beliebig. Wir machen eine Fallunterscheidung.

- Es ist $\frac{1}{2}(\alpha+\beta) \leq t$.

  In Abbildung 6.1 verdeutlichen wir die Situation. Es ist $t = \frac{1}{2}[(2t-\beta)+\beta]$.

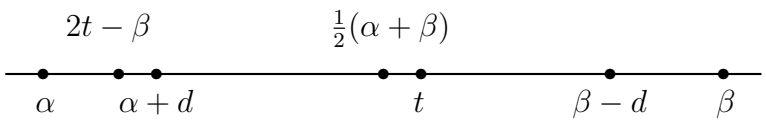

Abbildung 6.1: Der Fall $\frac{1}{2}(\alpha+\beta) \leq t$

  Wendet man daher (1) mit $a := 2t-\beta$, $b := \beta$ an, so erhält man

  $$|u'(t)| \leq \frac{c}{\beta-t}\|u\|_{\infty,[2t-\beta,\beta]} \leq \frac{c}{d}\|u\|_{\infty,[\alpha,\beta]} \quad \text{für alle } t \in [\alpha+d,\beta-d]$$

  und damit die Behauptung.

- Es ist $t < \frac{1}{2}(\alpha+\beta)$.

  Man vergleiche Abbildung 6.2.Diesmal ist $t = \frac{1}{2}[(2t-\alpha)+\alpha]$. Nun wende man (1) mit $a := \alpha$ und $b := 2t-\alpha$ an. Man erhält

  $$|u'(t)| \leq \frac{c}{t-\alpha}\|u\|_{\infty,[\alpha,2t-\alpha]} \leq \frac{c}{d}\|u\|_{\infty,[\alpha,\beta]} \quad \text{für alle } t \in [\alpha+d,\beta-d]$$

  und hieraus die Behauptung.

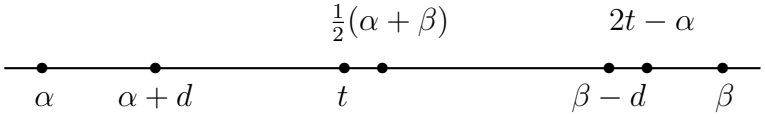

Abbildung 6.2: Der Fall $t < \frac{1}{2}(\alpha+\beta)$

Nun kommt ein weiterer Reduktionsschritt. Es genügt nämlich zu zeigen:

- Es existiert eine nur von n abhängende Konstante $c > 0$ mit

$$|u'(0)| \le c\,\|u\|_{\infty,[-1,1]} \quad \text{für alle } u \in E_n. \tag{2}$$

Denn angenommen, die Existenz einer Konstanten $c > 0$ mit (2) sei bewiesen. Seien $a < b$ und $v \in E_n$ beliebig, etwa $v(t) = \sum_{i=1}^{l} p_i(t)e^{\lambda_i t}$ mit paarweise verschiedenen λ_i und Polynomen p_i mit $\sum_{i=1}^{l}(1+\partial p_i) \le n$. Man definiere die Funktion u durch

$$u(t) := v\left(\frac{b-a}{2}\cdot t + \frac{a+b}{2}\right).$$

Offensichtlich ist auch $u \in E_n$ und wegen (2) ist

$$\frac{(b-a)}{2}\left|v'\left(\frac{a+b}{2}\right)\right| = |u'(0)| \le c\,\|u\|_{\infty,[-1,1]} = c\,\|v\|_{\infty,[a,b]},$$

woraus (1) folgt.
Wegen der Homogenität in der Ungleichung (2) genügt es schließlich zu zeigen:

- Es existiert eine nur von n abhängende Konstante $c > 0$ mit

$$|u'(0)| \le c \quad \text{für alle } u \in E_n \text{ mit } \|u\|_{\infty,[-1,1]} = 1. \tag{3}$$

Nun kommt ein letzter Reduktionsschritt. Da ja $\Pi_{n-1} \subset E_n$ liegt es nahe, die Polynome gesondert zu behandeln. Zunächst zeigen wir die Gültigkeit von (3) sogar für alle $u \in \Pi_n$, anschließend für alle $u \in E_n \setminus \Pi_n$.

- Es existiert eine nur von n abhängende Konstante $c_0 = c_0(n)$ mit

$$|u'(0)| \le c_0 \quad \text{für alle } u \in \Pi_n \text{ mit } \|u\|_{\infty,[-1,1]} = 1.$$

Denn: Dies folgt sofort aus der Markoffschen Ungleichung

$$\|u'\|_{\infty,[-1,1]} \le n^2\,\|u\|_{\infty,[-1,1]} \quad \text{für alle } u \in \Pi_n$$

(siehe z. B. E. W. Cheney (1966, S. 91)). Ein direkter Beweis ist naheliegenderweise der folgende: Die Abbildung $l\colon (\Pi_n, \|\cdot\|_{\infty,[-1,1]}) \longrightarrow (,|\cdot|)$, definiert durch $l(u) := u'(0)$, ist eine lineare Abbildung zwischen endlichdimensionale linearen normierten Räumen, daher auch stetig, und folglich

$$|l(u)| = |u'(0)| \le \|l\|\,\|u\|_{\infty,[-1,1]} \quad \text{für alle } u \in \Pi_n.$$

Nun zeigen wir schließlich:

- Es existiert eine nur von n abhängende Konstante $c = c(n) > 0$ mit

$$|u'(0)| \le c \quad \text{für alle } u \in E_n \setminus \Pi_n \text{ mit } \|u\|_{\infty,[-1,1]} = 1.$$

Denn: Wir definieren die Intervalle

$$I_j := \left[-\frac{j}{n}, -\frac{(j-1)}{n}\right], \qquad I'_j := \left[\frac{j-1}{n}, \frac{j}{n}\right], \qquad j = 1, \ldots, n,$$

der Länge $d := 1/n$. Man definiere

$$r_j := 2^{2j-1} j! n^j, \qquad j = 1, \ldots, n,$$

und anschließend

$$C_j := \sum_{i=j}^{n} r_i, \qquad j = 1, \ldots, n.$$

Wir wollen zeigen, dass obige Behauptung • mit $c = C_1$ gilt. Angenommen, dies wäre nicht der Fall. Dann gibt es ein $u \in E_n \setminus \Pi_n$ mit $\|u\|_{\infty,[-1,1]} = 1$ und $|u'(0)| > C_1$. O. B. d. A. können wir $u'(0) > C_1$ annehmen (notfalls ersetze man u durch $-u$). Wir werden zeigen, dass $u^{(n+1)}$ mindestens n Nullstellen in $[-1, 1]$ besitzt. Zusammen mit $u^{(n+1)} \in E_n$ und $u^{(n+1)} \neq 0$ (da $u \notin \Pi_n$) ergibt dies einen Widerspruch zu Satz 6.1.5, denn nach diesem hat $u \in E_n \setminus \{0\}$ höchstens $k(u) - 1 \le n - 1$ Nullstellen.
Wegen Satz 6.1.4, angewandt auf I_1 bzw. I'_1, gibt es Punkte $\xi_1 \in I_1$, $\xi'_1 \in I'_1$, mit

$$\left\{ \begin{array}{l} |u'(\xi_1)| \\ |u'(\xi'_1)| \end{array} \right\} \le \frac{2^{2\cdot 1-1} \cdot 1!}{(1/n)^1} = 2n = r_1 \le C_1 < u'(0).$$

Daher ist $\xi_1 < 0 < \xi'_1$. Aus dem Mittelwertsatz folgt die Existenz von $\zeta_1 \in (\xi_1, 0)$, $\zeta'_1 \in (0, \xi'_1)$ mit

$$u'(0) - u'(\xi_1) = -u''(\zeta_1)\xi_1 > C_1 - r_1 = C_2,$$
$$u'(\xi'_1) - u'(0) = u''(\zeta'_1)\xi'_1 < r_1 - C_1 = -C_2$$

bzw.

$$u''(\zeta_1) > \left(-\frac{1}{\xi_1}\right) C_2 > C_2, \qquad u''(\zeta'_1) < -\left(\frac{1}{\xi'_1}\right) C_2 < -C_2.$$

Also besitzt u'' eine Nullstelle in (ζ_1, ζ'_1), wobei

$$-\frac{1}{n} < \zeta_1 < 0 < \zeta'_1 < \frac{1}{n}.$$

Wir zeigen nun:

- Für $j = 2, \ldots, n+1$ gibt es Punkte ζ_{j-1}, ζ'_{j-1} mit

  (a) $-(j-1)/n < \zeta_{j-1} < 0 < \zeta'_{j-1} < (j-1)/n$,

  (b) $u^{(j)}(\zeta_{j-1}) > C_j$, $(-1)^{j-1} u^{(j)}(\zeta'_{j-1}) > C_j$,

(c) $u^{(j)}$ besitzt mindestens $j-1$ Nullstellen in $(\zeta_{j-1}, \zeta'_{j-1})$.

Denn: Für $j = 2$ sind die Aussagen (a)–(c) richtig, wie wir gerade gesehen haben. Angenommen, die Aussagen (a)–(c) seien für $j \in \{2, \ldots, n\}$ richtig. Wegen Satz 6.1.4, angewandt auf I_j bzw. I'_j statt I, u statt y und j statt m, existieren $\xi_j \in I_j$, $\xi'_j \in I'_j$ mit

$$|u^{(j)}(\xi_j)| \le r_j, \qquad |u^{(j)}(\xi'_j)| \le r_j.$$

Hierbei ist $\xi_j \le -(j-1)/n < \zeta_{j-1}$ und entsprechend $\zeta'_{j-1} < (j-1)/n \le \xi'_j$. Wegen des Mittelwertsatzes existieren $\zeta_j \in (\xi_j, \zeta_{j-1})$ und $\zeta'_j \in (\zeta'_{j-1}, \xi'_j)$ mit

$$u^{(j)}(\zeta_{j-1}) - u^{(j)}(\xi_j) = u^{(j+1)}(\zeta_j)(\underbrace{\zeta_{j-1} - \xi_j}_{\in(0,1)}) > C_j - r_j = C_{j+1}$$

und damit $u^{(j+1)}(\zeta_j) > C_{j+1}$ sowie

$$(-1)^{j-1}(u^{(j)}(\zeta'_{j-1}) - u^{(j)}(\xi'_j)) = (-1)^j u^{(j+1)}(\zeta'_j)(\underbrace{\xi'_j - \zeta'_{j-1}}_{\in(0,1)}) > C_j - r_j = C_{j+1}$$

und damit $(-1)^j u^{(j+1)}(\zeta'_j) > C_{j+1}$. Damit sind (a) und (b) für $j+1$ statt j nachgewiesen. Bleibt also der Nachweis von (c) für $j+1$ statt j, um den Induktionsbeweis abzuschließen.

Nach Induktionsannahme besitzt $u^{(j)}$ mindestens $j-1$ Nullstellen in $(\zeta_{j-1}, \zeta'_{j-1})$. Sei t_1 die kleinste und t_{j-1} die größte dieser Nullstellen. Wegen des Satzes von Rolle besitzt $u^{(j+1)}$ mindestens $j-2$ Nullstellen in (t_1, t_{j-1}). Ferner ist $u^{(j)}(t) > 0$ für $t \in (\zeta_{j-1}, t_1)$, $u^{(j)}(t_1) = 0$ und folglich $u^{(j+1)}(t_1) \le 0$. Zusammen mit $u^{(j+1)}(\zeta_j) > C_{j+1} > 0$ folgt aus dem Zwischenwertsatz, dass $u^{(j+1)}$ auch noch eine Nullstelle in $(\zeta_j, t_1]$ besitzt. Entsprechend ist $(-1)^{j-1}u^{(j)}(t) > 0$ für $t \in (t_{j-1}, \zeta'_{j-1})$, $u^{(j)}(t_{j-1}) = 0$. zusammen mit $(-1)^j u^{(j+1)}(\zeta'_j) > C_{j+1} > 0$ folgt auch noch die Existenz einer Nullstelle von $u^{(j+1)}$ in $[t_{j-1}, \zeta'_j)$. Insgesamt hat $u^{(j+1)}$ also mindestens j Nullstellen in (ζ_j, ζ'_j). Der Induktionsbeweis ist abgeschlossen. Daher besitzt $u^{(n+1)} \in E_n \setminus \{0\}$ mindestens n Nullstellen in $(-1, 1)$, ein Widerspruch zu Satz 6.1.5. □

Jetzt ist das Ziel nicht mehr fern.Vor der Formulierung und dem Beweis des letzten entscheidenden Hilfssatzes (dies ist genau die Aussage 2. in dem ganz am Anfang angebenen "Fahrplan" zum Beweis des Existenzsatzes für die T-Approximation mit Exponentialsummen) müssen wir noch an ein funktionalanalytisches Hilfsmittel erinnern, das wir ohne Beweis (einen Beweis findet man in jedem Funktionalanalysis-Lehrbuch) zitieren.

Satz von Arzela-Ascoli *Sei $\{x_k\} \subset C[\alpha, \beta]$. Es gelte:*

(a) Die Folge $\{x_k\}$ ist beschränkt in $(C[\alpha, \beta], \|\cdot\|_\infty)$, d. h. es existiert eine Konstante $C > 0$ mit $\|x_k\|_\infty \le C$ für alle $k \in \mathbb{N}$.

(b) Die Folge $\{x_k\}$ ist gleichgradig stetig, d. h. es existiert zu jedem $\epsilon > 0$ ein $\delta = \delta(\epsilon) > 0$ mit

$$s, t \in [\alpha, \beta],\ |s - t| \le \delta \Longrightarrow |x_k(s) - x_k(t)| \le \epsilon \quad \text{für alle } k \in \mathbb{N}.$$

Dann kann aus $\{x_k\}$ eine auf $[\alpha,\beta]$ gleichmäßig konvergente Teilfolge ausgewählt werden. D. h. es existiert eine Teilfolge $\{x_{k_i}\} \subset \{x_k\}$ und ein $x \in C[\alpha,\beta]$ mit

$$\lim_{i\to\infty} \|x_{k_i} - x\|_\infty = 0.$$

Der folgende Satz stammt von E. SCHMIDT (1970) und ist das entscheidende Hilfsmittel zum Beweis des Existenzsatzes.

Satz 6.1.7 *Sei $\{u_k\} \subset E_n$ eine in $(C[\alpha,\beta], \|\cdot\|_{\infty,[\alpha,\beta]})$ beschränkte Folge. Dann existiert eine Teilfolge $\{u_{k_j}\} \subset \{u_k\}$ und ein $u^* \in E_n$ mit der Eigenschaft, dass $\{u_{k_j}\}$ auf jedem Teilintervall $[a,b]$ mit $\alpha < a < b < \beta$ gleichmäßig gegen u^* konvergiert.*

Beweis: 1. Man wähle $2n+2$ Punkte a_i, b_i, $i = 1,\dots,n+1$, mit

$$\alpha < a_1 < \cdots < a_{n+1} < b_{n+1} < \cdots < b_1 < \beta.$$

Eine wiederholte Anwendung von Satz 6.1.6 liefert, dass die Folge $\{u_k^{(i)}\}$ in $C[a_i,b_i]$, $i = 1,\dots,n+1$, beschränkt ist, insbesondere also in $C[a_{n+1},b_{n+1}]$. Es existiert also eine Konstante $K > 0$ mit

$$\|u_k^{(i)}\|_{\infty,[a_{n+1},b_{n+1}]} \le K \quad \text{für alle } k \in \mathbb{N} \text{ und } i = 0,1,\dots,n+1.$$

Für $i = 0,1,\dots,n$ gilt dann: Ist $s,t \in [a_{n+1},b_{n+1}]$, so ist

$$|u_k^{(i)}(s) - u_k^{(i)}(t)| \le \|u_k^{(i+1)}\|_{\infty,[a_{n+1},b_{n+1}]}\,|s-t| \le K\,|s-t|.$$

Daher ist $\{u_k^{(i)}\}$, $i = 0,\dots,n$, in $C[a_{n+1},b_{n+1}]$ beschränkt und gleichgradig stetig. Eine wiederholte Anwendung des Satzes von Arzela-Ascoli liefert die Existenz einer Teilfolge $\{u_{k_j}\} \subset \{u_k\}$ sowie von $u_i^* \in C[a_{n+1},b_{n+1}]$, $i = 0,\dots,n$, mit

$$\lim_{j\to\infty} \|u_{k_j}^{(i)} - u_i^*\|_{\infty,[a_{n+1},b_{n+1}]} = 0, \qquad i = 0,\dots,n.$$

Wir werden zeigen, dass $\{u_{k_j}\}$ die gesuchte Teilfolge ist. Um Schreibarbeit zu sparen, nehmen wir im folgenden allerdings an, dass kein Übergang zu einer Teilfolge nötig ist, dass also schon

$$\lim_{k\to\infty} \|u_k^{(i)} - u_i^*\|_{\infty,[a_{n+1},b_{n+1}]} = 0, \qquad i = 0,\dots,n.$$

Sei $u^* := u_0^*$. Dann ist offenbar $u_i^* = (u^*)^{(i)}$, $i = 1,\dots,n$.

2. Nun zeigen wir, dass $u^* \in E_n$. Als Exponentialsumme hat $u_k \in E_n$ die Form

$$u_k(t) = \sum_{i=1}^{l_k} p_{i,k}(t) e^{\lambda_{i,k} t} \quad \text{mit} \quad \sum_{i=1}^{l_k} (1 + \partial p_{i,k}) \le n.$$

Für das Weitere muss der Zusammenhang zwischen Exponentialsummen und Lösungen von Differentialgleichungen mit konstanten Koeffizienten geklärt werden. Einem $u \in E_n$

der Ordnung k ordnen wir einen Differentialoperator L der Ordnung k zu, und zwar auf die folgende Weise. Hat u die Darstellung

$$u(t) = \sum_{i=1}^{l} p_i(t)e^{\lambda_i t} \quad \text{mit} \quad k := \sum_{i=1}^{l}(1 + \partial p_i) \leq n,$$

so sei der Differentialoperator L definiert durch

$$L := \prod_{i=1}^{l}\left(\frac{d}{dt} - \lambda_i\right)^{1+\partial p_i}.$$

Ist z. B. $n := 2$ und $u(t) := te^t$, so ist $u \in E_2$ und die Ordnung von u ist $k = 2$. Der zugehörige Differentialoperator ist

$$L = \left(\frac{d}{dt} - 1\right)^2$$

bzw.

$$Ly = \left(\frac{d}{dt} - 1\right)\left(\frac{dy}{dt} - y\right) = \frac{d^2y}{dt^2} - 2\frac{dy}{dt} + y.$$

Einsetzen von u ergibt

$$Lu(t) = \left(\frac{d}{dt} - 1\right)(e^t + te^t - te^t) = \left(\frac{d}{dt} - 1\right)e^t = 0.$$

Nun wollen wir uns überlegen, dass die letzte Gleichung kein Zufall ist, dass also eine Exponentialsumme $u \in E_n$ der Ordnung k einer linearen, homogenen Differentialgleichung k-ter Ordnung mit konstanten Koeffizienten genügt, also

$$u^{(k)} + a_{k-1}u^{(k-1)} + \cdots + a_1u' + a_0u = 0,$$

deren zugehöriges charakteristisches Polynom

$$\chi(\lambda) := \lambda^k + a_{k-1}\lambda^{k-1} + \cdots + a_1\lambda + a_0$$

gerade die λ_i als Wurzeln der Vielfachheit $1+\partial p_i$, $i = 1, \ldots, l$, besitzt. Dies liegt daran, dass

$$\begin{aligned}
\left(\frac{d}{dt} - \lambda_i\right)^{1+\partial p_i}(p_i(t)e^{\lambda_i t}) &= \left(\frac{d}{dt} - \lambda_i\right)^{\partial p_i}(p_i'(t) + \lambda_i p_i(t) - \lambda_i p_i(t))e^{\lambda_i t} \\
&= \left(\frac{d}{dt} - \lambda_i\right)^{1+\partial p_i-1}(p_i'(t)e^{\lambda_i t}) \\
&= \left(\frac{d}{dt} - \lambda_i\right)^{1+\partial p_i-2}(p_i''(t)e^{\lambda_i t}) \\
&\vdots \\
&= \underbrace{p_i^{(1+\partial p_i)}(t)}_{=0}e^{\lambda_i t} \\
&= 0.
\end{aligned}$$

Die Frequenzen λ_i in der Darstellung von $u \in E_n$ sind stets als paarweise verschieden und der Größe nach geordnet angenommen:

$$\lambda_1 < \lambda_2 < \cdots < \lambda_l.$$

Schreibt man jedes λ_i entsprechend seiner Vielfachheit $1 + \partial p_i$ als Wurzel des charakteristischen Polynoms

$$\chi(\lambda) = \prod_{i=1}^{l} (\lambda - \lambda_i)^{1+\partial p_i}$$

auf, so erhält man

$$(\lambda_1 =)\mu_1 \leq \mu_2 \leq \cdots \leq \mu_k (= \lambda_l).$$

Dann genügt

$$u(t) = \sum_{i=1}^{l} p_i(t) e^{\lambda_i t}$$

der Differentialgleichung k-ter Ordnung

$$\prod_{i=1}^{k} \left(\frac{d}{dt} - \mu_i \right) u(t) = 0.$$

Jetzt kehren wir zum eigentlichen Beweis zurück. Die Ordnung

$$k(u_k) = \sum_{i=1}^{l_k} (1 + \partial p_{i,k}) \leq n$$

kann als konstant angenommen werden (notfalls gehe man zu einer Teilfolge über), etwa maximal gleich n (andernfalls ersetze man n durch $n-1$ usw.). Also gilt

$$(*) \qquad \prod_{i=1}^{n} \left(\frac{d}{dt} - \mu_{i,k} \right) u_k = 0.$$

Die Folgen $\{\mu_{i,k}\}_{k \in \mathbb{N}}$ sind beschränkt oder unbeschränkt. Indem man zu Teilfolgen übergeht und notfalls umnummeriert, kann man annehmen, dass

$$\lim_{k \to \infty} \mu_{i,k} = \mu_i, \qquad i = 1, \ldots, q, \quad (\{\mu_{i,k}\} \text{ beschränkt})$$

und

$$\lim_{k \to \infty} \frac{1}{\mu_{i,k}} = 0, \qquad i = q+1, \ldots, n, \quad (\{\mu_{i,k}\} \text{ unbeschränkt}).$$

Dividiert man die Differentialgleichung $(*)$ durch $\prod_{i=q+1}^{n} (-\mu_{i,k})$, so erhält man

$$0 = \prod_{i=1}^{q} \left(\frac{d}{dt} - \mu_{i,k} \right) \prod_{i=q+1}^{n} \left(1 - \frac{1}{\mu_{i,k}} \frac{d}{dt} \right) u_k.$$

Mit $k \to \infty$ folgt wegen der gleichmäßigen Konvergenz von $\{u_k^{(i)}\}$ gegen $(u^*)^{(i)}$ auf $[a_{n+1}, b_{n+1}]$,dass

$$0 = \prod_{i=1}^{q} \left(\frac{d}{dt} - \mu_i \right) u^*, \quad \text{falls } q \geq 1$$

bzw. $u^* = 0$, falls $q = 0$. Oben haben wir uns überlegt, dass ein $u \in E_n$ der Ordnung k einer linearen Differentialgleichung mit konstanten Koeffizienten k-ter Ordnung genügt, deren charakteristisches Polynom nur reelle Nullstellen besitzt. Wie aus der Theorie linearer Differentialgleichungen bekannt ist, gilt aber auch die Umkehrung. Damit ist nachgewiesen, dass $u^* \in E_q \subset E_n$.

Wir halten fest, was wir bisher bewiesen haben (wir sind nämlich leider noch nicht fertig):

- *Sei $\{u_k\} \subset E_n$ beschränkt in $(C[\alpha, \beta], \|\cdot\|_{\infty,[\alpha,\beta]})$. Seien a_{n+1}, b_{n+1} mit $\alpha < a_{n+1} < b_{n+1} < \beta$ gegeben. Dann existiert eine Teilfolge $\{u_{k_j}\} \subset \{u_k\}$ und ein $u^* \in E_n$ derart, dass $\{u_{k_j}\}$ gleichmäßig auf $[a_{n+1}, b_{n+1}]$ gegen u^* konvergiert.*

Zu zeigen bleibt, dass $\{u_{k_j}\}$ nicht nur auf $[a_{n+1}, b_{n+1}]$ sondern auf *jedem* Teilintervall $[a, b]$ mit $\alpha < a < b < \beta$ gegen u^* konvergiert. Hierbei können wir natürlich annehmen, dass $[a_{n+1}, b_{n+1}]$ eine *echte* Teilmenge von $[a, b]$ ist. Angenommen, die Behauptung sei nicht richtig. Dann existiert eine Teilfolge von $\{u_{k_j}\}$, wiederum mit $\{u_{k_j}\}$ bezeichnet, und ein $\epsilon > 0$ mit $\|u_{k_j} - u^*\|_{\infty,[a,b]} \geq \epsilon$ für alle j. Durch Wiederholung des obigen Beweises mit $[\tilde{a}_{n+1}, \tilde{b}_{n+1}] := [a, b]$ erhält man die Existenz einer Teilfolge von $\{u_{k_j}\}$, die gleichmäßig auf $[a, b]$ gegen ein $u^{**} \in E_n$ konvergiert und es ist $\|u^{**} - u^*\|_{\infty,[a,b]} \geq \epsilon$. Da aber notwendig $u^* = u^{**}$ auf $[a_{n+1}, b_{n+1}]$, ist auch $u^* = u^{**}$ auf $[a, b]$. Man hat einen Widerspruch zu $\|u^{**} - u^*\|_{\infty,[a,b]} \geq \epsilon$ erhalten und der Satz ist schließlich bewiesen. □

Nach den Vorbemerkungen vor Beginn aller Hilfssätze ist damit bewiesen:

Satz 6.1.8 *Die Menge E_n der Exponentialsummen der Ordnung $\leq n$ ist eine Existenzmenge in $(C[\alpha, \beta], \|\cdot\|_\infty)$.*

Durch ein Beispiel am Anfang dieses Abschnitts haben wir gezeigt, dass die Menge E_n^0 der reinen Exponentialsummen der Ordnung $\leq n$ keine Existenzmenge in $(C[\alpha, \beta], \|\cdot\|_\infty)$ ist. Erstaunlicherweise gilt aber:

Satz 6.1.9 *Die Menge*

$$E_n^+ := \left\{ \sum_{i=1}^{n} a_i e^{\lambda_i t} : \lambda_i \in \mathbb{R},\ a_i \geq 0\ (i = 1, \ldots, n) \right\}$$

der reinen nichtnegativen Exponentialsummen der Ordnung $\leq n$ ist eine Existenzmenge in $(C[\alpha, \beta], \|\cdot\|_\infty)$.

Beweis: Es genügt offenbar zu zeigen:

- Ist $\{u_k\} \subset E_n^+$ beschränkt in $(C[\alpha, \beta], \|\cdot\|_\infty)$, so existiert eine Teilfolge $\{u_{k_j}\} \subset \{u_k\}$ und ein $u^* \in E_n^+$ derart, dass $\{u_{k_j}\}$ auf jedem Intervall $[a, b] \subset (a, b)$ gleichmäßig gegen u^* konvergiert.

Denn: Sei

$$u_k(t) = \sum_{i=1}^{n} a_{i,k} e^{\lambda_{i,k} t} \qquad \text{mit } a_{i,k} \geq 0,\ i = 1, \ldots, n,\ k \in \mathbb{N}.$$

Aus der Beschränktheit von $\{u_k\}$ folgt die von $\{v_{i,k}\}_{k\in\mathbb{N}}$, $i = 1, \ldots, n$, wobei $v_{i,k}(t) := a_{i,k} e^{\lambda_{i,k} t}$. O. B. d. A. ist also $n = 1$ und $u_k(t) = a_k e^{\lambda_k t}$, $a_k \geq 0$. Nach Satz 6.1.7 existiert eine Folge $\{u_{k_j}\} \subset \{u_k\}$ und ein $u^* \in E_1$ derart, dass $\{u_{k_j}\}$ auf jedem Teilintervall $[a, b] \subset (\alpha, \beta)$ gleichmäßig gegen u^* konvergiert. Ist $\{\lambda_{k_j}\}$ unbeschränkt, so ist $u^* = 0$, im anderen Fall ist $u^*(t) = ae^{\lambda t}$ und notwendig $a \geq 0$. Insgesamt ist also $u^* \in E_n^+$, der Satz ist bewiesen. □

6.2 Charakterisierung und Eindeutigkeit bester T-Approximierender in E_n^0, E_n^+ und E_n

Wie im letzten Abschnitt sei

$$E_n^0 := \left\{ u : u(t) = \sum_{i=1}^{n} a_i e^{\lambda_i t},\ a_i,\ \lambda_i \in \mathbb{R}\ (i = 1, \ldots, n) \right\}$$

die Menge der reinen Exponentialsummen der Ordnung $\leq n$,

$$E_n^+ := \left\{ \sum_{i=1}^{n} a_i e^{\lambda_i t} : \lambda_i \in \mathbb{R},\ a_i \geq 0\ (i = 1, \ldots, n) \right\}$$

der reinen nichtnegativen Exponentialsummen der Ordnung $\leq n$,

$$E_n := \left\{ u : u(t) = \sum_{i=1}^{l} p_i(t) e^{\lambda_i t},\ p_i \in \Pi,\ \lambda_i \in \mathbb{R}\ (i = 1, \ldots, l),\ \sum_{i=1}^{l} (1 + \partial p_i) \leq n \right\}$$

die Menge der Exponentialsummen der Ordnung $\leq n$. Vorgegeben sei stets ein $z \in C[\alpha, \beta]$, o. B. d. A. ist $z \notin E_n^0$ (bzw. E_n^+, E_n). Ist $u \in E_n^0$ (bzw. E_n^+, E_n), so heißen $N + 1$ Punkte

$$\alpha \leq t_0 < t_1 < \cdots < t_N \leq \beta$$

eine *Alternante der Länge* $N + 1$ zu $u - z$, falls

(1) $|u(t_i) - z(t_i)| = \|u - z\|_\infty$, $i = 0, \ldots, N$,

(2) $\operatorname{sign}(u(t_i) - z(t_i)) = -\operatorname{sign}(u(t_{i+1}) - z(t_{i+1}))$, $i = 0, \ldots, N - 1$.

Uns interessieren in diesem Abschnitt notwendige und hinreichende Bedingungen dafür, dass ein $u^* \in E_n^0$ (bzw. E_n^+, E_n) beste Approximierende an z bezüglich E_n^0 (bzw. E_n^+, E_n) ist.

Wir erinnern an die Definition der Länge $l(u)$ und der Ordnung $k(u)$ von $u \in E_n$. Ist $u = 0$, so ist $l(u) = k(u) = 0$. Ist dagegen $u(t) = \sum_{i=1}^{l} p_i(t) e^{\lambda_i t}$ mit $p_i \in \Pi \setminus \{0\}$, $i = 1, \ldots, l$, und $\lambda_1 < \cdots < \lambda_l$, so ist $l(u) := l$ die *Länge* von u und $k(u) := \sum_{i=1}^{l} (1 + \partial p_i)$

die *Ordnung* von u. Offenbar ist $k(u) \geq l(u)$ und $k(u) = l(u)$ für $u \in E_n^0$. Unser Ziel wird es zunächst sein, eine Charakterisierung einer besten Approximierenden (also notwendige und hinreichende Bedingungen) in E_n^0 mittels Alternanteneigenschaften nachzuweisen, sowie die Eindeutigkeit bester Approximierender in E_n^0 zu beweisen (man beachte jedoch, dass E_n^0 keine Existenzmenge in $(C[\alpha,\beta], \|\cdot\|_\infty)$ ist). Hierzu werden wir jeweils einen direkten Beweis angeben und danach die Einordnung in eine wesentlich allgemeinere Theorie von Meinardus-Schwedt und varisolventer Familien von Rice andeuten. Anschließend werden wir die gleichen Aufgabenstellungen in E_n betrachten, wobei schon darauf hingewiesen sei, dass wir dort keinen Eindeutigkeits- und keinen Charakterisierungssatz haben, sondern nur notwendige und hinreichende Bedingungen mit Lücke.

Wir beginnen mit dem angekündigten Charakterisierungssatz bezüglich E_n^0.

Satz 6.2.1 *Ein $u^* \in E_n^0$ mit $u^*(t) = \sum_{i=1}^{k(u^*)} a_i^* e^{\lambda_i^* t}$ (hier ist $a_i^* \neq 0$, $i = 1,\dots,k(u^*)$, $\lambda_1^* < \dots < \lambda_{k(u^*)}^*$) ist genau dann beste Approximierende an $z \in C[\alpha,\beta]$, wenn es eine Alternante der Länge $n + k(u^*) + 1$ zu $u^* - z$ gibt.*

Beweis: 1. Wir zeigen zunächst durch ein Argument vom de La Vallée Poussin-Typ, dass die angegebene Bedingung hinreichend ist. Sei also

$$(\alpha \leq)\ t_0 < \dots \leq t_{n+k(u^*)}\ (\leq \beta)$$

eine Alternante zu $u^* - z$. Angenommen, es gibt ein $u \in E_n^0$ mit $\|u - z\|_\infty < \|u^* - z\|_\infty$. Für $i = 0,\dots,n + k(u^*)$ ist insbesondere

$$\begin{aligned} 0 &< |u^*(t_i) - z(t_i)| - |u(t_i) - z(t_i)| \\ &\leq \operatorname{sign}(u^*(t_i) - z(t_i))(u^*(t_i) - z(t_i)) - \operatorname{sign}(u^*(t_i) - z(t_i))(u(t_i) - z(t_i)) \\ &= \operatorname{sign}(u^*(t_i) - z(t_i))(u^*(t_i) - u(t_i)) \end{aligned}$$

und daher

$$\operatorname{sign}(u^*(t_i) - z(t_i)) = \operatorname{sign}(u^*(t_i) - u(t_i)), \qquad i = 0,\dots,n + k(u^*).$$

Daher alterniert das Vorzeichen von $u^* - u$ in den $n + k(u^*) + 1$ Alternantenpunkten, $u^* - u$ besitzt also mindestens $n + k(u^*)$ Nullstellen in $[\alpha,\beta]$. Andererseits hat $u^* - u$ die Form

$$u^*(t) - u(t) = \sum_{i=1}^{k(u^*)+n} \tilde{a}_i e^{\tilde{\lambda}_i t},$$

folglich ist $u^* - u \in E_{k(u^*)+n}^0$. Da $u^* - u \neq 0$, besitzt $u^* - u$ nach Satz 6.1.5 höchstens $n + k(u^*) - 1$ Nullstellen, ein Widerspruch, und der erste Teil des Satzes ist bewiesen.

2. Notwendige Optimalitätsbedingungen sind immer schwieriger zu beweisen als hinreichende Bedingungen. Aber wir wissen ja zum Glück ziemlich genau, wie wir vorzugehen haben. Der reinen Exponentialsumme $u \in E_n^0$ mit $u(t) = \sum_{i=1}^n a_i e^{\lambda_i t}$ ordnen wir den Parametervektor

$$x = (a_1,\dots,a_n,\lambda_1,\dots,\lambda_n)^T \in \mathbb{R}^{2n}$$

zu, wobei etwa $a_1, \ldots, a_k \neq 0$, $a_i = 0$, $i = k+1, \ldots, n$, und $\lambda_1 < \cdots < \lambda_k$, $\lambda_{k+1}, \ldots, \lambda_n$ beliebig, aber sämtliche λ_i paarweise verschieden. Hierbei ist $k = k(u)$ die Ordnung von u. Hiermit definieren wir

$$F(x,t) := \sum_{i=1}^{n} a_i e^{\lambda_i t}.$$

Sei $u^*(t) = F(x^*, t)$ eine (globale) beste T-Approximierende an z in E_n^0, also

$$\|F(x^*, \cdot) - z\|_\infty \leq \|F(x, \cdot) - z\|_\infty \quad \text{für alle } x \in \mathbb{R}^{2n}.$$

Sei

$$x^* = (a_1^*, \ldots, a_n^*, \lambda_1^*, \ldots, \lambda_n^*)^T, \qquad p = (b_1, \ldots, b_n, \mu_1, \ldots, \mu_n)^T.$$

Dann ist

$$F(x^* + sp, t) - F(x^*, t) = s \sum_{i=1}^{n} (b_i + a_i^* \mu_i t) e^{\lambda_i^* t} + o(s).$$

Wegen (wir verändern jetzt die Schreibweise ein wenig, indem wir z. B. $\|u(t)\|_\infty$ statt der korrekteren Schreibweise $\|u(\cdot)\|_\infty$ oder $\|u\|_\infty$ benutzen)

$$\begin{aligned} \|F(x^*,t) - z(t)\|_\infty &\leq \|F(x^* + sp, t) - z(t)\|_\infty \\ &\leq \|F(x^* + sp) - F(x^*,t) - s\textstyle\sum_{i=1}^n (b_i + a_i^*\mu_i t) e^{\lambda_i^* t}\|_\infty \\ &\qquad + \|F(x^*,t) + s\textstyle\sum_{i=1}^n (b_i + a_i^*\mu_i t) e^{\lambda_i^* t} - z(t)\|_\infty \end{aligned}$$

ist für alle $s > 0$ (bei einer lokal besten Approximierenden: für alle *hinreichend kleinen* $s > 0$)

$$\underbrace{-\frac{1}{s}\|f(x^* + sp, t) - F(x^*,t) - s\textstyle\sum_{i=1}^n (b_i + a_i^*\mu_i t) e^{\lambda_i^* t}\|_\infty}_{=o(s)}$$

$$\leq \frac{1}{s}[\|F(x^*,t) + s\textstyle\sum_{i=1}^n (b_i + a_i^*\mu_i t) e^{\lambda_i^* t} - z(t)\|_\infty - \|F(x^*,t) - z(t)\|_\infty].$$

Mit $u^*(t) = F(x^*, t)$ und $s \to 0+$ folgt (siehe Beispiel 3 im Anschluss an Satz 3.2.4)

$$0 \leq \max_{t \in B(u^* - z)} \operatorname{sign}(u^*(t) - z(t)) \sum_{i=1}^{n} (b_i + a_i^* \mu_i t) e^{\lambda_i^* t},$$

wobei wieder

$$B(u^* - z) := \{t \in [\alpha, \beta] : |u^*(t) - z(t)| = \|u^* - z\|_\infty\}.$$

Definiert man also

$$W(x^*) := \left\{ \sum_{i=1}^{n} q_i(t) e^{\lambda_i^* t}, \quad \begin{array}{ll} q_i \in \Pi_1 & (i = 1, \ldots, k(u^*)), \\ q_i \in \Pi_0 & (i = k(u^*) + 1, \ldots, n) \end{array} \right\},$$

so ist

$$0 \leq \max_{t \in B(u^* - z)} \operatorname{sign}(u^*(t) - z(t)) h(t) \quad \text{für alle } h \in W(x^*).$$

In Beispiel 3. auf Seite 101 im Anschluss an Satz 4.1.4 hatten wir nachgewiesen:

- *Seien $q_1, \ldots, q_n$ nichtnegative ganze Zahlen und $\lambda_1, \ldots, \lambda_n$ paarweise verschiedene reelle Zahlen. Dann ist*

$$E_n(q, \lambda) := \left\{ \sum_{k=1}^{n} p_k(t) e^{\lambda_k t} : p_k \in \Pi_{q_k},\ k = 1, \ldots, n \right\}$$

auf jedem (reellen) Intervall I ein $N := (\sum_{k=1}^{n} q_k) + n$-dimensionaler Haarscher Teilraum von $C(I)$.

Wegen dieses Ergebnisses ist $W(x^*)$ ein $(n + k(u^*))$-dimensionaler Haarscher Teilraum von $C[\alpha, \beta]$. Hieraus folgt in der üblichen Weise die Existenz von Punkten t_i, $i = 0, \ldots, n + k(u^*)$, mit

1. $\alpha \le t_0 < t_1 < \cdots < t_{n+k(u^*)} \le \beta$,
2. $|u^*(t_i) - z(t_i)| = \|u^* - z\|_\infty$ bzw. $t_i \in B(u^* - z)$, $i = 0, \ldots, n + k(u^*)$,
3. Es existieren $\lambda_i > 0$, $i = 0, \ldots, n + k(u^*)$, mit

$$0 = \sum_{i=0}^{n+k(u^*)} \lambda_i \operatorname{sign}(u^*(t_i) - z(t_i)) h(t_i) \quad \text{für alle } h \in W(x^*).$$

Berücksichtigt man nun noch den ersten Teil von Lemma 3.4.4, so erhält man, dass die t_i, $i = 0, \ldots, n + k(u^*)$, eine Alternante für $u^* - z$ bilden. Der Satz ist bewiesen. □

Bemerkungen: 1. Im letzten Satz haben wir eigentlich sogar bewiesen: Ist $u^* \in E_n^0$ eine lokal beste Approximierende an z in E_n^0, so existiert eine Alternante der Länge $n + k(u^*) + 1$ für $u^* - z$ und hieraus folgt (hinreichende Bedingung), dass u^* sogar eine global beste Approximierende an z in E_n^0 ist. Insbesondere ist eine lokal beste Approximierende in E_n^0 auch eine global beste Approximierende.

2. Den ersten Teil des Beweises können wir zu einem Satz vom de La Vallée Poussin-Typ ausbauen:

- Sei $u \in E_n^0$ und $k = k(u)$ die Ordnung von u. Gibt es dann $n + k + 1$ Punkte t_i mit

  (i) $\alpha \le t_0 < t_1 < \cdots < t_{n+k} \le \beta$,

  (ii) Mit $\sigma \in \{-1, 1\}$ ist $\operatorname{sign}(u(t_i) - z(t_i)) = \sigma(-1)^i$, $i = 0, \ldots, n + k$,

  so ist

$$\min_{i=0,\ldots,n+k} |u(t_i) - z(t_i)| \le d(z, E_n^0).$$

3. Im Laufe des durchsichtigen Beweises zum letzten Satz schimmerte ein allgemeines Prinzip durch. Hierauf wollen wir aber erst später eingehen. □

Zunächst folgt der Eindeutigkeitssatz für E_n^0.

Satz 6.2.2 *Zu jedem $z \in C[\alpha, \beta]$ existiert höchstens eine beste T-Approximierende in E_n^0.*

Beweis: Seien u und v zwei beste Approximierende an z in E_n^0. Wir benutzen Bezeichnungen wie im zweiten Teil von Satz 6.2.1. Sei also

$$u(t) = F(x,t) = \sum_{i=1}^{n} a_i e^{\lambda_i t}, \qquad v(t) = F(y,t) = \sum_{i=1}^{n} b_i e^{\mu_i t}.$$

O. B. d. A. ist $k(u) \le k(v)$, die Ordnung von u also nicht größer als die von v. Nach Satz 6.2.1 gibt es eine Alternante der Länge $n + k(u) + 1$ zu $u - z$, also Punkte t_i, $i = 0, \ldots, n + k(u)$, mit:

1. $\alpha \le t_0 < \cdots < t_{n+k(u)} \le \beta$,
2. $|F(x,t_i) - z(t_i)| = d(z, E_n^0)$, $i = 0, \ldots, n + k(u)$,
3. $\operatorname{sign}(F(x,t_i) - z(t_i)) = \sigma(-1)^i$, $i = 0, \ldots, n + k(u)$, wobei $\sigma \in \{-1, 1\}$.

Für $i = 0, \ldots, n + k(u)$ ist dann

$$\sigma(-1)^i[F(y,t_i) - z(t_i)] \le |F(x,t_i) - z(t_i)| = \sigma(-1)^i[F(x,t_i) - z(t_i)]$$

und folglich

$$(*) \qquad 0 \le \sigma(-1)^i[F(x,t_i) - F(y,t_i)], \qquad i = 0, \ldots, n + k(u).$$

Ist $F(x,t_i) - F(y,t_i) = 0$, $i = 0, \ldots, n+k(u)$, so ist $F(x,t) = F(y,t)$ für alle t bzw. $u = v$. Denn $F(x,\cdot) - F(y,\cdot) \in E^0_{n+k(u)}$ (siehe erster Teil des Beweises von Satz 6.2.1). Daher wird nun angenommen, dass $F(x,t_j) - F(y,t_j) \ne 0$ für ein $j \in \{0, \ldots, n+k(u)\}$. Wegen $(*)$ ist $\operatorname{sign}(F(x,t_j) - F(y,t_j)) = \sigma(-1)^j$. Wir setzen $p = (c_1, \ldots, c_n, \nu_1, \ldots, \nu_n)^T$. Dann ist

$$\begin{aligned} F(x,t) - F(y - sp, t) &= F(x,t) - \sum_{i=1}^{n} (b_i - sc_i) e^{(\mu_i - s\nu_i)t} \\ &= F(x,t) - \sum_{i=1}^{n} b_i e^{\mu_i t} + s \underbrace{\sum_{i=1}^{n} (c_i + b_i \nu_i t) e^{\mu_i t}}_{=:P(t)} + o(s) \\ &= F(x,t) - F(y,t) + sP(t) + o(s). \end{aligned}$$

Hierbei ist

$$P \in W(y) := \left\{ \sum_{i=1}^{n} q_i(t) e^{\mu_i t} : \begin{array}{l} q_i \in \Pi_1 \ (i = 1, \ldots, k(v)), \\ q_i \in \Pi_0 \ (i = k(v)+1, \ldots, n) \end{array} \right\}.$$

Da $W(y)$ ein Haarscher Teilraum von $C[\alpha, \beta]$ der Dimension $n + k(v) \ge n + k(u)$ ist, existiert ein $P \in W(y)$ und damit ein $p \in \mathbb{R}^{2n}$ mit $P(t_i) = \sigma(-1)^i$ für alle $i \in \{0, \ldots, n + k(u)\} \setminus \{j\}$. Dann ist

$$\begin{aligned} \sigma(-1)^i[F(x,t_i) - F(y - sp, t_i)] &= \underbrace{\sigma(-1)^i[F(x(t_i) - F(y,t_i)]}_{\ge 0} \\ &\quad + s\sigma(-1)^i P(t_i) + o(s). \end{aligned}$$

Für $i \in \{0, \dots, n + k(u)\} \setminus \{j\}$ ist daher

$$\sigma(-1)^i[F(x,t_i) - F(y - sp, t_i)] \geq s \underbrace{\sigma(-1)^i P(t_i)}_{=1} + o(s) = s + o(s) > 0$$

für alle hinreichend kleinen $s > 0$. Für $i = j$ ist dagegen

$$\begin{aligned} \sigma(-1)^i[F(x,t_i) - F(y - sp, t_i)] &= \underbrace{\sigma(-1)^j[F(x,t_j) - F(y,t_j)]}_{>0} + sP(t_j) + o(s) \\ &> 0 \quad \text{für alle hinreichend kleinen } s > 0. \end{aligned}$$

Also besitzt $F(x,t) - F(y - sp, t) \in E_{n+k(u)}^0$ für alle hinreichend kleinen $s > 0$ mindestens $n + k(u)$ Nullstellen, verschwindet also identisch. Aus Stetigkeitsgründen folgt mit $s \to 0+$, dass $F(x,t) = F(y,t)$ für alle t bzw. $u = v$. Damit ist die Eindeutigkeit bewiesen. □

Bemerkung: Oben behaupteten wir, dass bei den Beweisen der letzten beiden Sätze ein allgemeineres Prinzip durchschimmert. Wir wollen hier den Zusammenhang mit der Theorie von Meinardus-Schwedt (siehe G. MEINARDUS (1967, S. 141 ff.)) beleuchten. Hierzu sei $X \subset \mathbb{R}^m$ offen (z. B. $X = \mathbb{R}^{2n}$) und $M := \{F(x, \cdot) : x \in X\}$, wobei $F \in C(X \times [\alpha, \beta])$ (und damit $F(x, \cdot) \in C[\alpha, \beta]$ für alle $x \in X$). Uns interessieren Charakterisierungs- und Eindeutigkeitsaussagen für die Approximationsaufgabe, $z \in C[\alpha, \beta]$ im T-Sinne durch Elemente aus M zu approximieren. Oben, bei der Approximation mit reinen Exponentialsummen, ist z. B.

$$x = (a_1, \dots, a_n, \lambda_1, \dots, \lambda_n)^T, \qquad F(x,t) = \sum_{i=1}^{n} a_i e^{\lambda_i t}.$$

Entscheidend haben wir benutzt, dass wir F nach den Parametern x differenzieren können. Daher setzen wir jetzt voraus:

(a) Die partiellen Ableitungen $\partial F / \partial x_j$, $j = 1, \dots, m$, existieren und sind auf $X \times [\alpha, \beta]$ stetig. Mit

$$F_x(x,t) = \left(\frac{\partial F}{\partial x_1}(x,t), \dots, \frac{\partial F}{\partial x_m}(x,t) \right)^T$$

bezeichnen wir den Gradienten von F bezüglich x. Im obigen Fall der reinen Exponentialsummen ist

$$F_x(x,t) = (e^{\lambda_1 t}, \dots, e^{\lambda_n t}, t a_1 e^{\lambda_1 t}, \dots, t a_n e^{\lambda_n t})^T.$$

Für $x \in X$ sei

$$W(x) := \{F_x(x, \cdot)^T p : p \in \mathbb{R}^m\}.$$

Dies ist offenbar ein endlichdimensionaler linearer Teilraum von $C[\alpha, \beta]$.

(b) Die *lokale Haarsche Bedingung* sei erfüllt, d. h. für jedes $x \in X$ sei $W(x)$ ein Haarscher Teilraum von $C[\alpha, |beta]$, etwa der Dimension $d(x)$.D. h. jedes nichttriviale Element von $W(x)$ besitzt höchstens $d(x) - 1$ Nullstellen in $C[\alpha, \beta]$.

(c) Die *Nullstellenbedingung* sei erfüllt: Für $x, y \in X$ verschwindet $F(x,\cdot) - F(y,\cdot)$ identisch oder besitzt höchstens $d(x) - 1$ Nullstellen in $[\alpha, \beta]$.

Man überlege sich, dass die letzten beiden Sätze unter diesen Voraussetzungen verallgemeinert werden können (siehe G. MEINARDUS (1967, S. 142–147)). □

Nun studieren wir die T-Approximation bezüglich der Menge E_n^+ der nichtnegativen reinen Exponentialsummen. Eines unserer Ziele wird es sein, nachzuweisen, dass E_n^+ eine T-Menge in $(C[\alpha,\beta], \|\cdot\|_\infty)$ ist. Der nächste Satz wird ein wichtiges Hilfsmittel sein (siehe z. B. D. BRAESS (1986, S. 170)).

Satz 6.2.3 (Descartes-Regel) *Sei $u \in E_n^0 \setminus \{0\}$, $u(t) = \sum_{i=1}^n a_i e^{\lambda_i t}$ mit $\lambda_1 < \cdots < \lambda_n$, $a_i \neq 0$, $i = 1, \ldots, n$ (andernfalls wäre $u \in E_{n-1}^0$). Sei N die Anzahl der reellen Nullstellen von u und W die Anzahl der Vorzeichenwechsel in $(a_1, a_2, \ldots, a_n)$. Dann ist $N \leq W$. Ist $N = W$ und t_N die größte Nullstelle von u, so ist $\operatorname{sign} u(t) = \operatorname{sign} a_n$ für alle $t > t_N$.*

Beweis: Der Beweis wird durch Induktion nach W, der Anzahl der Vorzeichenwechsel in $(a_1, \ldots, a_n)$, erbracht. Ist $W = 0$, so sind alle a_i, $i = 1, \ldots, n$, positiv oder negativ, so dass u keine reelle Nullstelle besitzt und es ist $N = 0$. Der Induktionsanfang ist damit gelegt. Sei nun $W > 0$, die Aussage sei für $W - 1$ richtig. Dann gibt es ein $j < n$ mit $a_j a_{j+1} < 0$. Man wähle $\lambda \in (\lambda_j, \lambda_{j+1}$. Dann ist

$$v(t) := \frac{d}{dt}(e^{-\lambda t} u(t)) = \sum_{i=1}^n b_i e^{(\lambda_i - \lambda)t} \quad \text{mit} \quad b_i := a_i(\lambda_i - \lambda).$$

Für $i = 1, \ldots, j$ ist $\lambda_i \leq \lambda_j < \lambda$ und daher $\operatorname{sign} b_i = -\operatorname{sign} a_i$. In $(b_1, \ldots, b_j)$ treten also so viele Vorzeichenwechsel wie in $(a_1, \ldots, a_j)$ auf. Entsprechend ist $\operatorname{sign} b_i = \operatorname{sign} a_i$ für $i = j+1, \ldots, n$, in $(b_{j+1}, \ldots, b_n)$ treten also so viele Vorzeichenwechsel wie in $(a_{j+1}, \ldots, a_n)$ auf. Ferner ist $\operatorname{sign} b_j = -\operatorname{sign} a_j = \operatorname{sign} a_{j+1} = \operatorname{sign} b_{j+1}$. In $(b_1, \ldots, b_n)$ treten also $W - 1$ Vorzeichenwechsel auf. Nach Induktionsannahme besitzt v also höchstens $W - 1$ Nullstellen und daher $e^{-\lambda t} u$ bzw. u höchstens W Nullstellen. Damit ist der Induktionsschluss erfolgt. Rechts von t_N ist u von einem Vorzeichen. Wegen $\lim_{t\to\infty} e^{-\lambda_n t} u(t) = a_n$ ist $\operatorname{sign} u(t) = \operatorname{sign} a_n$ für alle $t > t_N$. Der Satz ist bewiesen. □

Bevor wir den Alternantensatz für die T-Approximation bezüglich E_n^+ formulieren und beweisen, führen wir noch eine Bezeichnung ein. Ist $u \in E_n^0$, also $u(t) = \sum_{i=1}^n a_i e^{\lambda_i t}$ mit $\lambda_1 < \cdots < \lambda_n$, so bezeichne $k^+(u)$ die Anzahl der positiven a_i, $k^-(u)$ die Anzahl negativer a_i. Offenbar ist

$$k^+(u) + k^-(u) = k(u),$$

wobei $k(u)$ die Ordnung von u ist.

Satz 6.2.4 *Sei $z \in C[\alpha,\beta]$ und $u^* \in E_n^+$. Dann ist u^* genau dann eine beste Approximierende an z in E_n^+, wenn*

(i) $u^* - z$ eine Alternante der Länge $2n + 1$ besitzt

oder

(ii) $u^* - z$ *eine Alternante* $\alpha \le t_0 < \cdots < t_{2k(u^*)} \le \beta$ *der Länge* $2k(u^*)+1$ *mit* $(u^* - z)(t_{2k(u^*)}) > 0$ *besitzt.*

Beweis: O. B. d. A. ist natürlich $z \notin E_n^+$. Wir zeigen zunächst, dass die Bedingungen (i) oder (ii) hinreichend dafür sind, dass u^* eine beste Approximierende an z in E_n^+ ist. Hierzu machen wir eine Fallunterscheidung und nehmen im *ersten Fall* $k(u^*) = n$ an. Wegen (i) (oder (ii)) besitzt $u^* - z$ dann eine Alternante der Länge $n + k(u^*) + 1 = 2n+1$. Aus dem Alternantensatz 6.2.1 folgt, dass u^* beste Approximierende an z in E_n^0 und damit (wegen $u^* \in E_n^+$) erst recht in E_n^+ ist. Im *zweiten Fall* ist $k(u^*) < n$. Wir können annehmen, dass (ii) erfüllt ist. Der (hinreichende Teil in) Satz 6.2.1 liefert, dass u^* beste Approximierende an z in $E_{k(u^*)}^0$ und damit wegen $u^* \in E_{k(u^*)}^+$ auch in $E_{k(u^*)}^+$ ist. Zu zeigen bleibt, dass u^* auch beste Approximierende an z in E_n^+ ist. Angenommen, dies sei nicht der Fall, es existiere also ein $u \in E_n^+$ mit $\|u - z\|_\infty < \|u^* - z\|_\infty$. Dann ist

$$\begin{aligned}(-1)^i[u(t_i) - z(t_i)] &\le \|u - z\|_\infty \\ &< \|u^* - z\|_\infty \\ &= (-1)^i[u^*(t_i) - z(t_i)], \quad i = 0, \ldots, 2k(u^*).\end{aligned}$$

Hieraus folgt

$$(-1)^i[u^*(t_i) - u(t_i)] > 0, \qquad i = 0, \ldots, 2k(u^*).$$

Folglich besitzt $u - u^*$ mindestens $2k(u^*)$ Nullstellen in $(t_0, t_{2k(u^*)})$, d. h. es ist $2k(u^*) \le N(u - u^*)$, und es ist $(u - u^*)(t_{2k(u^*)}) < 0$. Da $u - u^* = \sum_{i=1}^m b_i e^{\mu_i t}$ mit $\mu_1 < \cdots < \mu_m$, $m \le n + 2k(u^*)$, eine von der Nullfunktion verschiedene reine Exponentialsumme ist, folgt aus der Descartes-Regel in Satz 6.2.3 auch noch $N(u - u^*) \le W(u - u^*)$, wobei $W(u - u^*)$ die Anzahl der Vorzeichenwechsel in $(b_1, \ldots, b_m)$ ist. Insgesamt ist also

$$2k(u^*) \le N(u - u^*) \le W(u - u^*).$$

Wir unterscheiden zwei Fälle. Im ersten Fall ist $2k(u^*) = W(u - u^*)$. Dann ist also $2k(u^*) = N(u - u^*) = W(u - u^*)$. Wegen (der letzten Aussage in) Satz 6.2.3 und $(u - u^*)(t_{2k(u^*)}) < 0$ ist $b_m < 0$. Dann ist

$$k(u^*) + 1 \le k^-(u - u^*) \le \underbrace{k^-(u)}_{=0} + k^+(u^*) = k(u^*),$$

ein Widerspruch. Im zweiten Fall ist $2k(u^*) < W(u - u^*)$, also $2k(u^*) + 1 \le W(u - u^*)$. Dann ist wieder $k(u^*) + 1 \le k^-(u - u^*)$, was wie gerade eben zu einem Widerspruch führt. Damit ist gezeigt, dass die Bedingungen (i) oder (ii) hinreichend dafür sind, dass $u^* \in E_n^+$ eine beste Approximierende an z in E_n^+ ist.

Sei $u^* \in E_n^+$ eine beste Approximierende an z in E_n^+ ist. Wir machen eine Fallunterscheidung. Im *ersten Fall* ist $k(u^*) = n$. Dann sind alle Koeffizienten a_i^* von u^* positiv. Eine kleine Störung in den Parametern liefert nach wie vor Elemente von E_n^+. Genau wie beim Beweis von Satz 6.2.1 folgt die Existenz einer Alternante der Länge $n + k(u^*) + 1 = 2n + 1$ für $u^* - z$. In diesem Fall ist also die Bedingung (i) erfüllt. Im *zweiten Fall* ist $k(u^*) < n$. Zur Abkürzung sei $k := k(u^*($. Dann ist $u^* \in P_{E_n^+}(z) \cap E_k^+$ und damit $u^* \in P_{E_k^+}(z)$. Wegen des ersten Falls existiert zu $u^* - z$ eine Alternante der

Länge $2k+1$. Wir zeigen, dass es sogar eine Alternante der Länge $2k+2$ für u^*-z gibt. Indem man die ersten $2k+1$ oder die letzten $2k+1$ Punkte dieser Alternante nimmt, erhält man eine, für die $(u^*-z)(t_{2k})>0$ ist, für die also Bedingung (ii) erfüllt ist. Angenommen, es gibt keine Alternante der Länge $2k+2$ für u^*-z. Wegen Satz 6.2.1 ist u^* dann *keine* beste Approximierende an z in E_{k+1}^0, es gibt also ein $u\in E_{k+1}^0$ mit $\|u-z\|_\infty<\|u^*-z\|_\infty$. Sei $\alpha\le t_0<\cdots<t_{2k}\le\beta$ eine Alternante zu u^*-z mit $(u^*-z)(t_{2k})<0$ (wäre $(u^*-z)(t_{2k})>0$, so wären wir schon fertig). Wegen (siehe den zweiten F all im ersten Teil des obigen Beweises)

$$\begin{aligned}(-1)^{i+1}[u(t_i)-z(t_i)] &\le \|u-z\|_\infty\\ &< \|u^*-z\|_\infty\\ &= (-1)^{i+1}[u^*(t_i)-z(t_i)],\quad i=0,\dots,2k,\end{aligned}$$

ist

$$(-1)^i[u(t_i)-u^*(t_i)]>0,\qquad i=0,\dots,2k.$$

Daher besitzt $u-u^*$ mindestens zwei Nullstellen in (t_0,t_{2k}) und es ist $(u-u^*)(t_{2k})>0$. Bezeichnet man wie oben mit $N(u-u^*)$ die Anzahl der Nullstellen von $u-u^*$ und mit $W(u-u^*)$ die Anzahl der Vorzeichenwechsel in $(b_1,\dots,b_m)$, wobei $u(t)-u^*(t)=\sum_{i=1}^m b_i e^{\mu_i t}$ mit $\mu_1<\cdots<\mu_m$, $m\le 2k+1$, so liefert die Descartesche Regel zunächst

$$2k\le N(u-u^*)\le W(u-u^*).$$

Auch hier machen wir eine Fallunterscheidung. Im ersten Fall ist $2k=W(u-u^*)$. Wegen Satz 6.2.3 ist $b_m>0$. Daher ist

$$k+1\le k^+(u-u^*)\le k^+(u).$$

Die Anzahl positiver Koeffizienten von $u\in E_{k+1}^0$ ist also mindestens gleich $k+1$. Da die Ordnung von u aber gleich $k+1$ ist $u\in E_{k+1}^+\subset E_n^+$, ein Widerspruch dazu, dass u^* beste Approximierende an z in E_n^+ ist. Im zweiten Fall ist $2k<W(u-u^*)$. Wiederum folgt

$$k+1\le k^+(u-u^*)\le k^+(u).$$

Wie gerade eben ergibt sich ein Widerspruch und der Satz ist bewiesen. □

Mit Hilfe des letzten Charakterisierungssatzes und der Eindeutigkeit bester T-Approximierender in E_n^0 erhält man leicht:

Satz 6.2.5 *E_n^+ ist eine T-Menge in $(C[\alpha,\beta],\|\cdot\|)$.*

Beweis: Wegen Satz 6.1.9 ist E_n^+ eine Existenzmenge in $(C[\alpha,\beta],\|\cdot\|_\infty)$. Zu zeigen bleibt also die Eindeutigkeit bester T-Approximierender in E_n^+. Seien $u_1,u_2\in E_n^+$ zwei beste Approximierende an z in E_n^+. O. B. d. A. ist $k(u_1)\ge k(u_2)$. Wegen Satz 6.2.4 existiert zu u_1-z eine Alternante der Länge $2k(u_1)+1$. Wegen Satz 6.2.1 ist u_1 die nach Satz 6.2.2 eindeutige beste T-Approximierende an z in $E_{k(u_1)}^0$. Wegen $k(u_2)\le k(u_1)$ ist auch $u_2\in E_{k(u_1)}^0$. Da $u_1,u_2\in E_n^+$ beides beste T-Approximierende an z in E_n^+ sind, ist $\|u_1-z\|_\infty=\|u_2-z\|_\infty$. Also sind $u_1,u_2\in E_{k(u_1)}^0$ beide beste T-Approximierende an

z in $E_{k(u_1)}^0$. Wegen der Eindeutigkeit bester T-Approximierender in $E_{k(u_1)}^0$ ist $u_1 = u_2$, womit der Satz bewiesen ist. □

Nun untersuchen wir noch die T-Approximation bezüglich der (erweiterten oder verallgemeinerten) exponentialsummen. Durch ein Beispiel (siehe D. Braess (1986, S. 195)) machen wir uns zunächst klar, dass wir keine Eindeutigkeit erwarten können. Bei der Argumentation wird allerdings der gleich folgende Alternantensatz schon benutzt.

Beispiel: Sei $z \in C[-1,1]$, $z(t) = z(-t)$ für $t \in [-1,1]$ (bzw. z symmetrisch), $z(t) > 0$ für $t \in [-1,1]$ und z für $t > 0$ monoton fallend. Z. B. sei $z(t) = 1/(1+t^2)$. Angenommen, $u^* \in E_2$ sei die einzige beste Approximierende an z in E_2. Dann ist u^* notwendig selbst gerade, hat also die Form $u^*(t) = a\cosh \lambda t$ mit $a > 0$. Die Funktion $u^* - z$ ist monoton wachsend in $[0,1]$, besitzt dort also höchstens eine Nullstelle. Daher besitzt $u^* - z$ in $[-1,1]$ höchstens zwei Nullstellen, kann also nur eine Alternante von höchstens der Länge 3 besitzen. Notwendig für eine beste Approximierende ist aber, wie wir gleich sehen werden, die Existenz einer Alternante der Länge $n + l(u^* + 1 \geq 2 + 1 + 1 = 4$. Daher gibt es mindestens zwei beste Approximierende an z in E_2. □

Es folgt nun der angekündigte Alternantensatz für Exponentialsummen. In etwas allgemeinere Form, Bemerkungen hierzu machen wir später, findet man diese Aussage bei D. Braess (1986, S. 196).

Satz 6.2.6 *Sei $u^* \in E_n$, $l(u^*)$ die Länge und $k(u^*)$ die Ordnung von u^*. Ferner sei $z \in C[\alpha,\beta]$. Dann gilt:*

(a) Besitzt $u^* - z$ eine Alternante der Länge $n + k(u^*) + 1$, so ist u^* eine beste T-Approximierende an z in E_n.

(b) Ist u^* eine lokal beste T-Approximierende an z in E_n, so besitzt $u^* - z$ eine Alternante der Länge $n + l(u^*) + 1$.

Beweis: Im wesentlichen muss der Beweis von Satz 6.2.1 nur kopiert werden. Wir überlassen dies als eine Übungsaufgabe. Man beachte hierbei, dass die Nullstellenbedingung und die lokale Haarsche Bedingung erfüllt sind. □

Wie wir in einem Beispiel zeigten, ist E_n keine T-Menge in $(C[\alpha,\beta], \|\cdot\|_\infty)$, da die Eindeutigkeit einer besten Approximierenden i. Allg. nicht gesichert ist. Es gilt aber (siehe D. Braess (1986, S. 196)):

Satz 6.2.7 *Sei $z \in C[\alpha,\beta]$ und $u^* \in E_n^0$ beste Approximierende an z in E_n^0. Dann ist u^* die einzige beste T-Approximierende an z in E_n.*

Beweis: O. B. d. A. ist $z \notin E_n^0$. Es ist $k(u^*) = l(u^*)$, da $u^* \in E_n^0$. Nach Satz 6.2.1 existiert zu $u^* - z$ eine Alternante der Länge $n + k(u^*) + 1$. Wegen Satz 6.2.6 (a) ist u^* eine beste Approximierende an z in E_n. Sei $u \in E_n$ eine weitere beste T-Approximierende an z in E_n, also $\|u - z\|_\infty = \|u^* - z\|_\infty$. Sind $t_0, \ldots, t_{n+k(u^*)}$ die Alternantenpunkte von

$u^* - z$, so ist also

$$\begin{aligned} \operatorname{sign}(u^*(t_i) - z(t_i))[u(t_i) - z(t_i)] &\leq |u(t_i) - z(t_i)| \\ &\leq \|u - z\|_\infty \\ &= \|u^* - z\|_\infty \\ &= |u^*(t_i) - z(t_i)| \\ &= \operatorname{sign}(u^*(t_i) - z(t_i))[u^*(t_i) - z(t_i)] \end{aligned}$$

für $i = 0, \ldots, n + k(u^*)$ und folglich

$$0 \leq \operatorname{sign}(u^*(t_i) - z(t_i))[u^*(t_i) - u(t_i)], \qquad i = 0, \ldots, n + k(u^*).$$

Daher besitzt $u^* - u \in E_{n+k(u^*)}$ mindestens $n + k(u^*)$ Nullstellen. Wegen Satz 6.1.5 folgt $u^* = u$. □

Bemerkung: Ein großer Teil der Aussagen dieses Abschnitts kann auf sogenannte γ-Polynome verallgemeinert werden. Hierbei sei $\Lambda \subset \mathbb{R}$ und $\gamma \in C(\Lambda \times [\alpha, \beta])$. Dann heißt $u(t) = \sum_{i=1}^{k} a_i \gamma(\lambda_i, t)$ mit $a_i \in \mathbb{R}$, $\lambda_i \in \Lambda$, $i = 1, \ldots, k$, ein *γ-Polynom der Ordnung* k, wenn sich u nicht als eine entsprechende summe mit $k - 1$ Summanden darstellen lässt. Die reinen Exponentialsummen erhält man, indem man $\gamma(\lambda, t) := e^{\lambda t}$ und $\Lambda := \mathbb{R}$ setzt. Man kann untersuchen, unter welchen Voraussetzungen eine Descartes-Regel für γ-Polynome (siehe Definition 1.1 bei D. Braess (1986, S. 182)) gilt. Reine γ-Polynomr, für die das der Fall ist, nennt man dann eine Descartes-Familie, sie wird z. B. mit G_n^0 bezeichnet. Weiter kann man G_n^+ und G_n als Verallgemeinerung von E_n^+ und E_n definieren, worauf wir aber nicht mehr eingehen wollen, sondern nur auf D. Braess (1986, Chapter VII) verweisen. □

6.3 Varisolvenz

Zum Schluss dieses Kapitels wollen wir noch, allerdings ziemlich kurz, auf den von J. R. Rice eingeführten Begriff der *Varisolvenz* eingehen (siehe J. R. Rice (1969) und auch D. Braess (1986, S. 66 ff.)).

Definition 6.3.1 Sei $M \subset C[\alpha, \beta]$ nichtleer.

1. M heißt *lokal solvent* in $u_0 \in M$ vom Grad $m = m(u_0) \geq 1$, falls es zu jedem $\epsilon > 0$ und beliebigen m paarweise verschiedenen Punkten $t_i \in [\alpha, \beta]$, $i = 1, \ldots, m$, ein $\delta = \delta(u_0, \epsilon, t_1, \ldots, t_m) > 0$ gibt mit: Zu jedem $y = (y_i) \in \mathbb{R}^m$ mit $|u_0(t_i) - y_i| \leq \delta$, $i = 1, \ldots, m$, existiert ein $u \in M$ mit $\|u - u_0\|_\infty \leq \epsilon$ und $u(t_i) = y_i$, $i = 1, \ldots, m$.

2. M besitzt die *Eigenschaft Z* vom Grad $n = n(u_0)$ in $u_0 \in M$, falls $u - u_0$ für jedes $u \in M$ höchstens $n - 1$ Nullstellen besitzt oder identisch verschwindet.

3. M heißt eine *varisolvente Familie*, falls M in jedem Punkt $u \in M$ sowohl lokal solvent ist als auch die Eigenschaft Z besitzt und die Grade übereinstimmen: $m(u) = n(u)$.

4. M besitzt die *Dichte-Eigenschaft*, falls es zu gegebenen $u_0 \in M$ und $\epsilon > 0$ zwei Elemente $u_1, u_2 \in M$ mit $\|u_i - u_0\|_\infty \leq \epsilon$, $i = 1, 2$, und $u_1(t) < u_0(t) < u_2(t)$ für alle $t \in [\alpha, \beta]$ gibt.

Beispiele: 1. Sei $M \subset C[\alpha, \beta]$ ein n-dimensionaler Haarscher Teilraum.

(a) Seien paarweise verschiedene $t_1, \ldots, t_n$ aus $[\alpha, \beta]$ vorgegeben. Zu beliebig vorgegebenem $y = (y_i) \in \mathbb{R}^n$ existiert genau ein $u \in M$ mit $u(t_i) = y_i$, $i = 1, \ldots, n$. Die Abbildung $y \mapsto u$ ist stetig. Daher ist M in jedem $u_0 \in M$ lokal solvent vom konstanten Grad n.

(b) Offensichtlich besitzt M die Nullstelleneigenschaft Z vom Grad n in jedem $u_0 \in M$.

(c) Aus (a) und (b) folgt, dass M varisolvent vom Grad n ist.

(d) Die Dichte-Eigenschaft ist erfüllt, da es in jedem Haarschen Teilraum von $C[\alpha, \beta]$ ein positives Element gibt (siehe die Bemerkung im Anschluss an Lemma 5.1.1).

2. Sei $M := E_1 = \{u(t) = ae^{\lambda t} : a, \lambda \in \mathbb{R}\}$ (siehe D. Braess (1986, S. 66)). Wir überlegen uns, dass E_1 varisolvent vom Grad

$$m(u) = \begin{cases} 2, & u \neq 0, \\ 1, & u = 0 \end{cases}$$

ist. Denn:

(i) Wir zeigen, dass M lokal solvent in jedem $u_0 \in E_1$ ist. Zunächst sei $u_0 \neq 0$ und $t_1 \neq t_2$. Ferner seien y_1, y_2 von Null verschiedene reelle Zahlen vom gleichen Vorzeichen, etwa dem von u_0. Das Interpolationsproblem

$$ae^{\lambda t_i} = y_i, \quad i = 1, 2,$$

kann explizit gelöst werden:

$$\lambda = \frac{1}{t_2 - t_1} \ln \frac{y_2}{y_1}, \qquad a = y_1 e^{-\lambda t_1}.$$

Weiter sind die Parameter a und λ stetige Funktionen von y_1, y_2. Daher ist ist E_1 lokal solvent vom Grad 2 in $u_0 \neq 0$. Nun sei $u_0 = 0$. Dann hat $u - u_0$ keine Nullstelle, wenn $\in E_1$, $u \neq u_0$. Da E_1 den eindimensionalen linearen Raum der konstanten Funktionen enthält, ist E_1 lokal solvent vom Grad 1 in $u_0 = 0$. Man beachte, dass eine globale Interpolationseigenschaft unangemessen ist. Denn für $y_1 = -y_2 \neq 0$ existiert kein $u \in E_1$ mit $u(t_i) = y_i$, $i = 1, 2$.

(ii) Für $u_0 \neq 0$ und $u \in E_1$ ist $u - u_0 \in E_2$, besitzt also höchstens eine Nullstelle oder verschwindet identisch. Für $u_0 = 0$ besitzt $u - u_0 \in E_1$ keine Nullstelle oder verschwindet identisch. Also besitzt E_1 die Eigenschaft Z vom Grad $n(u) = m(u)$, insgesamt ist E_1 varisolvent.

Offensichtlich ist auch die Dichte-Eigenschaft erfüllt. □

Gegenüber der Vorgehensweise, die wir oben verfolgt haben (lokale Haarsche Bedingung usw.) ist der wesentliche Unterschied der, dass bei varisolventen Familien keine Differenzierbarkeitsvoraussetzungen eingehen. Da wir aber trotzdem zwischen einfachen und doppelten bzw. mehrfachen Nullstellen unterscheiden müssen, definieren wir (siehe D. Braess (1986, S. 68)):

Definition 6.3.2 Sei $t_0 \in (\alpha, \beta)$ eine Nullstelle von $u \in C[\alpha, \beta]$. Dann heißt t_0 eine *mehrfache Nullstelle* von u, falls es eine Umgebung U von t_0 in $[\alpha, \beta]$ gibt mit $f(t) \geq 0$ oder $f(t) \leq 0$ für alle $t \in U$ und $f(t) \neq 0$ für t aus dem Rand von U. Mehrfachen Nullstellen wird die Vielfachheit 2 zugeordnet. Alle anderen Nullstellen haben die Vielfachheit 1.

Das folgende Lemma dient dazu nachzuweisen, dass die Nullstelleneigenschaft Z richtig bleibt, wenn man die Nullstellen entsprechend ihrer Vielfachheit (im obigen Sinne) zählt.

Lemma 6.3.3 *Sei $M \subset C[\alpha, \beta]$ varisolvent. Zu $u, u_1 \in M$ gebe es $m + 1 = m(u) + 1$ Punkte*

$$(\alpha \leq) t_0 < t_1 < \cdots < t_m (\leq \beta)$$

mit

$$\sigma(-1)^i [u_1(t_i) - u(t_i)] \geq 0, \qquad i = 0, \ldots, m,$$

wobei $\sigma \in \{-1, 1\}$. Dann ist $u = u_1$.

Beweis: O. B. d. A. ist $m(u) \leq m(u_1)$. Ist $u_1(t_i) - u(t_i) = 0$, $i = 0, \ldots, m$, so hat $u_1 - u$ mindestens $m(u) + 1$ Nullstellen, verschwindet also wegen der Nullstelleneigenschaft Z identisch. Für ein $j \in \{0, \ldots, m\}$ sei daher $\sigma(-1)^j [u_1(t_j) - u(t_j)] > 0$. Da M lokal solvent vom Grad $m(u_1)$ in u_1 ist, gibt es zu $\epsilon := |u_1(t_j) - u(t_j)| > 0$ ein $\delta > 0$ mit: Zu jedem $y = (y_i)_{i \in \{0,\ldots,m\} \setminus \{j\}} \in \mathbb{R}^m$ mit $|u_1(t_i) - y_i| \leq \delta$, $i = 0, \ldots, m$, $i \neq j$, existiert ein $u_2 \in M$ mit $\|u_1 - u_2\| < \epsilon$ und $u_2(t_i) = y_i$, $i = 0, \ldots, m$, $i \neq j$. Insbesondere setze man

$$y_i := u_1(t_i) + (-1)^i \sigma \delta, \qquad i \in \{0, \ldots, m\} \setminus \{j\}.$$

Daher existieren ein $\delta > 0$ und ein $u_2 \in M$ mit

(i) $u_2(t_i) = u_1(t_i) + (-1)^i \sigma \delta$, $i \in \{0, \ldots, m\} \setminus \{j\}$,

(ii) $\|u_1 - u_2\|_\infty < |u_1(t_j) - u(t_j)|$.

Hieraus wollen wir schließen, dass $u_2 - u$ mindestens $m(u)$ Nullstellen besitzt, was wegen $u_2 \neq u$ (siehe (ii)) ein Widerspruch ist. Nun ist

$$(-1)^i \sigma [u_2(t_i) - u(t_i)] = \underbrace{(-1)^i \sigma [u_1(t_i) - u(t_i)]}_{\geq 0} + \underbrace{\delta}_{>0} > 0$$

für $i \in \{0, \ldots, m\} \setminus \{j\}$ und

$$\begin{aligned} \sigma(-1)^j[u_1(t_j) - u_2(t_j)] &\leq |u_1(t_j) - u_2(t_j)| \\ &\leq \|u_1 - u_2\|_\infty \\ &< |u_1(t_j) - u(t_j)| \\ &= \sigma(-1)^j[u_1(t_j) - u(t_j)], \end{aligned}$$

folglich

$$\sigma(-1)^j[u_2(t_j) - u(t_j)] > 0.$$

Insgesamt ist also

$$\sigma(-1)^i[u_2(t_i) - u(t_i)] > 0, \qquad i = 0, \ldots, m,$$

woraus die Existenz von mindestens $m = m(u)$ Nullstellen von $u_2 - u \neq 0$ folgt, ein Widerspruch. □

Lemma 6.3.4 *Sei $M \subset C[\alpha, \beta]$ varisolvent und $u, u_1 \in M$, $u \neq u_1$. Dann besitzt $u - u_1$ höchstens $m(u) - 1$ Nullstellen, wobei mehrfache Nullstellen doppelt gezählt sind.*

Beweis: Im Widerspruch zur Behauptung werde angenommen, $u - u_1$ habe $m = m(u)$ Nullstellen, wobei jede Nullstelle entsprechend ihrer Vielfachheit (1 oder 2) zu zählen ist. Seien $s_1 < \cdots < s_k$ die Nullstellen von $u - u_1$. Man füge nun noch gewisse Punkte hinzu, so dass man insgesamt $m+1$ Punkte $t_0 < \cdots < t_m$ mit $\sigma(-1)^i[u(t_i) - u_1(t_i)] \geq 0$, $i = 0, \ldots, m$, erhält. Dies mache man auf die folgende Weise: Zunächst gehören alle Nullstellen $s_1, \ldots, s_k$ zu den gesuchten $t_0, \ldots, t_m$. Für jede mehrfache Nullstelle wird ein weiterer Punkt hinzugefügt. Ist s_i eine mehrfache Nullstelle, so nehme man einen Punkt aus (s_i, s_{i+1}) bzw. (s_k, β) hinzu, ferner einen Punkt links von der ersten mehrfachen Nullstelle. Die erweiterte Menge enthält $m + 1$ Punkte, diese seien (nach Umordnung) $t_0 < \cdots < t_m$, es ist $\sigma(-1)^i[u(t_i) - u_1(t_i)] \geq 0$, $i = 0, \ldots, m$, mit $\sigma \in \{-1, 1\}$. Wegen Lemma 6.3.3 ist $u = u_1$, was aber gerade ausgeschlossen war. Damit ist das Lemma bewiesen. □

Die hinreichende Bedingung in einem Alternantensatz wird i. Allg. durch ein Argument vom de La Vallée Poussin-Typ nachgewiesen. Den Beweis des folgenden Satzes (siehe z. B. D. Braess (1986, S. 69)), den man eine nichtlineare Version des Satzes von de La Vallée Possin bezeichnen könnte, können wir weglassen, da wir den entsprechenden Beweisschluss schon wiederholt gemacht haben.

Satz 6.3.5 *$M \subset C[\alpha, \beta]$ habe die Eigenschaft Z vom Grad m in $u_0 \in M$, ferner sei $z \in C[\alpha, \beta]$. Es gebe $m + 1$ paarweise verschiedene Punkte $t_0 < t_1 < \cdots < t_m$ in $]\alpha, \beta]$ mit*

$$\operatorname{sign}(u_0(t_i) - z(t_i)) = \sigma(-1)^i, \qquad i = 0, \ldots, m,$$

wobei $\sigma \in \{-1, 1\}$. Dann ist

$$\min_{i=0,\ldots,m} |u_0(t_i) - z(t_i)| \leq d(z, M).$$

Schwieriger ist der Beweis des folgenden Alternantensatzes, für den wir auf D. Braess (1986, S. 69) verweisen.

Satz 6.3.6 *Sei $M \subset C[\alpha, \beta]$ varisolvent, $z \in C[\alpha, \beta]$ und $u^* \in M$. Es sei $u^* - z$ nicht konstant*[2]*. Dann ist u^* genau dann eine beste Approximierende an z in M, wenn $u^* - z$ eine Alternante der Länge $m(u) + 1$ besitzt.*

Als letzter Satz in diesem Zusammenhang wird der Eindeutigkeitssatz für varisolvente Familien angegeben.

Satz 6.3.7 *sei $M \subset C[\alpha, \beta]$ varisolvent, die Dichte-Eigenschaft sei erfüllt. Dann gibt es zu jedem $z \in C[\alpha, \beta]$ höchstens eine beste Approximierende in M.*

Beweis: Sei $u^* \in P_M(z)$. Wegen des Alternantensatzes gibt es zu $u^* - z$ eine Alternante $t_0 < t_1 < \cdots < t_{m(u^*)}$. Ist auch $u_1 \in P_M(z)$, so folgt in bewährter Weise

$$\sigma(-1)^i[u^*(t_i) - u_1(t_i)] \geq 0, \qquad i = 0, \ldots, m(u^*),$$

mit $\sigma \in \{-1, 1\}$. Wegen Lemma 6.3.3 folgt die Eindeutigkeit. □

Bemerkung: Der Vorteil des zuletzt geschilderten Ansatzes ist es, dass nicht über eine Parametrisierung der Menge M, mit der approximiert wird, vorgegangen wird, sondern nur im Funktionenraum argumentiert wird. Will man für konkrete Fälle etwa die lokale Solvenz von M nachweisen, so wird man i. Allg. doch auf die lokale Haarsche Bedingung zurückgreifen müssen. wegen des Satzes über implizite Funktionen impliziert diese nämlich lokale Solvenz. Insbesondere ist E_n^0 varisolvent, E_n aber nicht. □

[2] Auf die Voraussetzung, dass $u^* - z$ nicht konstant ist, kann verzichtet werden, wenn M der Dichteeigenschaft genügt.

Kapitel 7

Verschiedenes

7.1 Der Satz von Stone-Weierstraß

Der berühmte Weierstraßsche Approximationssatz sagt aus, dass sich jede auf einem kompakten Intervall stetige Funktion auf diesem Intervall gleichmäßig durch eine Folge von Polynomen approximieren lässt. Eine weitreichende Verallgemeinerung stammt von M. H. Stone (1937, 1948). Dieser Satz von Stone-Weierstraß ist für mich einer der erstaunlichsten Sätze der Mathematik. Ich hörte als Student das erste Mal von diesem Satz in einer Vorlesung von Hel Braun über Topologie an der Universität Hamburg. In dieser speziellen Vorlesung wurde Hel Braun krankheitshalber von Emil Artin vertreten!

Satz 7.1.1 (Stone-Weierstraß) *Sei $B \subset \mathbb{R}^n$ kompakt und $C(B)$ die Menge der auf B definierten, stetigen und reellwertigen Funktionen. Auf $C(B)$ definieren wir die Maximumnorm $\|\cdot\|_\infty$ durch $\|x\|_\infty := \max_{t\in B} |x(t)|$. Sei $\mathcal{A} \subset C(B)$ eine Teilalgebra, d. h. ein linearer Teilraum mit der Eigenschaft, dass mit $x, y \in \mathcal{A}$ auch[1] $x \cdot y \in \mathcal{A}$. Es gelte:*

1. *Es ist $1 \in \mathcal{A}$, wobei $1(t) := 1$ für alle $t \in B$, d. h. $\mathcal{A}$ enthält die konstanten Funktionen.*
2. *Zu $s, t \in B$ mit $s \neq t$ existiert ein $x \in \mathcal{A}$ mit $x(s) \neq x(t)$, d. h. $\mathcal{A}$ trennt Punkte von B.*

Dann ist $\mathcal{A}$ dicht in $C(B)$, d. h. zu jedem $x \in C(B)$ existiert eine Folge $\{x_k\} \subset \mathcal{A}$, die auf B gleichmäßig gegen x konvergiert, für die also $\lim_{k\to\infty} \|x - x_k\|_\infty = 0$.

Beweis: Mit $\mathrm{cl}(\mathcal{A})$ bezeichnen wir den Abschluss von $\mathcal{A}$ in $C(B)$ bezüglich gleichmäßiger Konvergenz auf B. D. h. es sei

$$\mathrm{cl}(\mathcal{A}) := \{x \in C(B) : \text{ Es existiert } \{x_k\} \subset \mathcal{A} \text{ mit } \lim_{k\to\infty} \|x - x_k\|_\infty = 0.\}$$

Wir haben zu zeigen, dass $\mathrm{cl}(\mathcal{A}) = C(B)$. Der Beweis erfolgt in mehreren Schritten.

(a) $\mathrm{cl}(\mathcal{A})$ ist eine Teilalgebra von $C(B)$, d. h. mit $x, y \in \mathrm{cl}(\mathcal{A})$ sowie $\alpha \in \mathbb{R}$, sind $x + y$, $x \cdot y$ sowie $\alpha x \in \mathrm{cl}(\mathcal{A})$.

[1] Für $x, y \in C(B)$ ist natürlich $x \cdot y \in C(B)$ durch $(x \cdot y)(t) := x(t)y(t)$ definiert.

Der Beweis hierfür ist offensichtlich.

(b) Sei $a > 0$ gegeben. Dann gibt es eine Folge $\{p_k\}$ von Polynomen mit

$$\lim_{k\to\infty} \max_{t\in[-a,a]} | \, |t| - p_k(t)| = 0,$$

welche also gleichmäßig auf $[-a, a]$ gegen die Betragsfunktion konvergiert.

Denn: O. B. d. A. ist $a = 1$. Definiere die Folge $\{p_k\}$ von Polynomen durch $p_0(t) := 0$ sowie $p_{k+1}(t) := p_k(t) + \frac{1}{2}(t^2 - p_k(t)^2)$, $k = 0, 1, \ldots$. In Abbildung 7.1 haben wir die Betragsfunktion sowie p_1, p_2 und p_3 dargestellt. Durch vollständige Induktion nach k

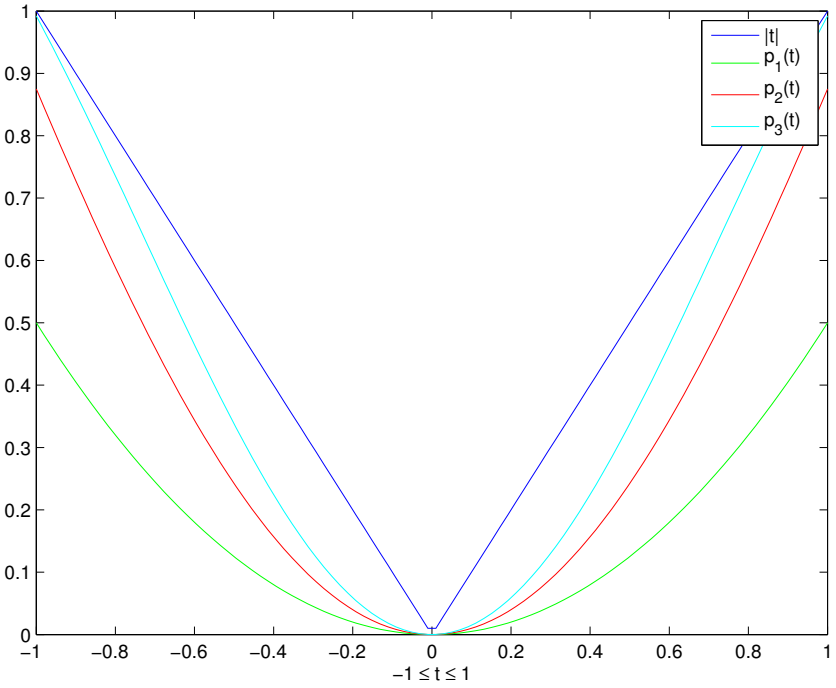

Abbildung 7.1: Approximation der Betragsfunktion durch Polynome

zeigt man, dass $p_k(t) \le |t|$ und $0 \le p_k(t) \le p_{k+1}(t)$ für alle $t \in [-1, 1]$. Denn für $k = 0$ sind diese beiden Aussagen offenbar richtig. Sei dies auch für k der Fall. Wegen

$$|t| - p_{k+1}(t) = \underbrace{(|t| - p_k(t))}_{\ge 0}\underbrace{(1 - \tfrac{1}{2}(|t| + p_k(t)))}_{\ge 0} \ge 0$$

ist die erste Aussage auch für $k + 1$ richtig, was für die zweite Aussage wegen der Rekursionsformel evident ist. Also konvergiert $\{p_k\}$ auf $[-1, 1]$ punktweise und ist monoton nicht fallend. Der Limes ist die Betragsfunktion, insbesondere ist diese stetig. Aus dem Satz von Dini[2] folgt die auf $[-1, 1]$ gleichmäßige Konvergenz von $\{p_k\}$ gegen die Betragsfunktion.

[2]Der Satz von Dini sagt aus: Eine auf einem kompakten Intervall I monoton nicht fallende oder monoton nicht wachsende, punktweise konvergente Funktionenfolge mit stetiger Grenzfunktion ist auf I sogar gleichmäßig konvergent.

(c) Ist $x \in \mathrm{cl}(\mathcal{A})$, so ist auch $|x| \in \mathrm{cl}(\mathcal{A})$, wobei $|x|(t) := |x(t)|$ für alle $t \in B$.

Denn: Wegen (b) existiert eine Folge $\{p_k\}$ von Polynomen mit

$$\lim_{k\to\infty} \max_{|\xi|\le\|x\|_\infty} ||\xi| - p_k(\xi)| = 0.$$

Da $\mathrm{cl}(\mathcal{A})$ nach (a) eine Teilalgebra von $C(B)$ ist und die konstanten Funktionen enthält, ist $p_k \circ x \in \mathrm{cl}(\mathcal{A})$ (wobei natürlich $(p_k \circ x)(t) := p_k(x(t))$). Wegen

$$\begin{aligned} \| \, |x| - p_k \circ x\|_\infty &= \max_{t\in B} | \, |x(t)| - p_k(x(t))| \\ &\le \max_{|\xi|\le\|x\|_\infty} | \, |\xi| - p_k(\xi)| \\ &\to 0 \quad \text{mit } k \to \infty \end{aligned}$$

ist $|x| \in \mathrm{cl}(\mathcal{A})$.

(d) Mit $x, y \in \mathrm{cl}(\mathcal{A})$ sind auch $\max(x, y), \min(x, y) \in \mathrm{cl}(\mathcal{A})$, wobei

$$\max(x, y)(t) := \max(x(t), y(t)), \quad \min(x, y)(t) := \min(x(t), y(t)).$$

Denn: Dies folgt sofort aus

$$\max(x, y) = \frac{1}{2}(x + y + |x - y|), \quad \min(x, y) = \frac{1}{2}(x + y - |x - y|)$$

sowie (a) und (c).

(e) Zu $s, t \in B$ mit $s \neq t$ und $\alpha, \beta \in \mathbb{R}$ existiert ein $x_{st} \in \mathcal{A}$ mit $x_{st}(s) = \alpha$ und $x_{st}(t) = \beta$.

Denn: Da $\mathcal{A}$ nach Voraussetzung Punkte trennt, gibt es ein $x \in \mathcal{A}$ mit $x(s) \neq x(t)$. Dann setze man

$$x_{st} := \frac{\alpha - \beta}{x(s) - x(t)} x + \frac{\beta x(s) - \alpha x(t)}{x(s) - x(t)} 1.$$

Nach diesen Vorbereitungen kommt jetzt der eigentliche Beweis des Satzes. Dieser erfolgt in den nächsten beiden Schritten (f) und (g).

(f) Seien $x \in C(B)$ und $\epsilon > 0$ beliebig. Dann existiert zu jedem $t \in B$ ein $x_t \in \mathrm{cl}\,(\mathcal{A})$ mit $x_t(t) = x(t)$ und $x_t(b) < x(b) + \epsilon/2$ für alle $b \in B$.

Denn: Zu jedem Paar $(s, t) \in B \times B$ existiert wegen (e) für $s \neq t$ ein $x_{st} \in \mathcal{A}$ mit $x_{st}(s) = x(s)$, $x_{st}(t) = x(t)$. Für $s \in B$ setze man $x_{ss} := x(s)1$. Nun definiere man

$$O_{st} := \{b \in B : x_{st}(b) < x(b) + \epsilon/2\}.$$

Dann ist O_{st} offen und $s \in O_{st}$, daher hat man durch $B = \bigcup_{s\in B} O_{st}$ für beliebiges $t \in B$ eine offene Überdeckung von B. Da B kompakt ist, existiert eine endliche Teilüberdeckung

$$B = \bigcup_{i=1}^{n} O_{s_i t}.$$

Hierbei sind n und die s_i i. Allg. von t abhängig. Man setze

$$x_t := \min(x_{s_1 t}, \ldots, x_{s_n t}).$$

Wegen (d) ist $x_t \in \mathrm{cl}\,(\mathcal{A})$. Wegen $x_{s_i t}(t) = x(t)$, $i = 1, \ldots, n$, ist auch $x_t(t) = x(t)$. Ein beliebiges $b \in B$ ist in $O_{s_i t}$ mit einem gewissen $i \in \{1, \ldots, n\}$ enthalten, so dass $x_t(b) \leq x_{s_i t}(b) < x(b) + \epsilon/2$. Damit ist der Beweisschritt (f) beendet.

(g) Es existiert ein $y \in \mathrm{cl}(\mathcal{A})$ mit $x(b) - \epsilon/2 < y(b) < x(b) + \epsilon/2$ für alle $b \in B$ bzw. mit $\|x - y\| < \epsilon/2$.

Denn: Man definiere

$$O_t := \{b \in B : x_t(b) > x(b) - \epsilon/2\}.$$

Für jedes $t \in B$ ist O_t offen und $t \in O_t$ wegen $x_t(t) = x(t)$ (siehe (f)). Dann ist auch $B = \bigcup_{t \in B} O_t$ eine offene Überdeckung von B. Wiederum kann eine endliche Teilüberdeckung $B = \bigcup_{j=1}^{m} O_{t_j}$ mit $\{t_1, \ldots, t_m\} \subset B$ ausgewählt werden. Nun setze man

$$y := \max(x_{t_1}, \ldots, x_{t_m}).$$

Bei gegebenem $b \in B$ ist einerseits $y(b) = x_{t_j}(b)$ für ein gewisses $j \in \{1, \ldots, m\}$ und daher $y(b) < x(b) + \epsilon/2$. Andererseits ist $b \in O_{t_k}$ mit einem gewissen $k \in \{1, \ldots, m\}$. Daher ist $y(b) \geq x_{t_k}(z) > x(b) - \epsilon/2$. Auch der Beweisschritt (g) ist beendet und damit der Satz von Stone-Weierstraß bewiesen. □

Folgerung Sei $B \subset \mathbb{R}^N$ kompakt und $\mathcal{P}$ die Menge der reellwertigen Polynome in N Variablen. Dann ist $\mathcal{P}$ eine Teilalgebra von $C(B)$, die die konstanten Funktionen enthält und Punkte von B trennt. Wegen des Satzes von Stone-Weierstraß ist $\mathrm{cl}(\mathcal{P}) = C(B)$, jede auf B stetige, reellwertige Funktion kann also auf B gleichmäßig durch Polynome approximiert werden. Dies gilt insbesondere für $B := [a, b] \subset \mathbb{R}$, womit auch der klassische Weierstraßsche Approximationssatz bewiesen ist.

Nun soll eine dem Satz von Stone entsprechende Aussage für $C_{\mathbb{C}}(B)$, die Menge der auf der kompakten Menge $B \subset \mathbb{R}^N$ stetigen *komplexwertigen* Funktionen bewiesen werden. Die direkte Übertragung vom Reellen ins Komplexe ist falsch. Denn sei $B := \{z \in \mathbb{C} : |z| \leq 1\}$ und $\Pi \subset C_{\mathbb{C}}(B)$ die Menge der (komplexen) Polynome. Ein Element von $\mathrm{cl}\,(\Pi)$ ist dann analytisch auf einer offenen Obermenge von B und daher $\mathrm{cl}\,(\Pi) \subset C_{\mathbb{C}}(B)$, $\mathrm{cl}\,(\Pi) \neq C_{\mathbb{C}}(B)$, obwohl Π die konstanten Funktionen enthält und Punkte trennt. Aber es gilt:

Satz 7.1.2 *Sei $\mathcal{A} \subset C_{\mathbb{C}}(B)$ eine Teilalgebra, d. h. ein linearer Teilraum (über $\mathbb{C}$) mit $x \cdot y \in \mathcal{A}$, falls $x, y \in \mathcal{A}$. Es gelte:*

1. *Es ist $1 \in \mathcal{A}$, wobei $1(t) := 1$ für alle $t \in B$. Also enthält $\mathcal{A}$ die konstanten Funktionen.*
2. *Zu $s, t \in B$ mit $s \neq t$ existiert ein $x \in \mathcal{A}$ mit $x(s) \neq x(t)$. Also trennt $\mathcal{A}$ Punkte von B.*
3. *Ist $x \in \mathcal{A}$, so ist $\overline{x} \in \mathcal{A}$. Hierbei ist $\overline{x}(t) := \overline{x(t)}$, wobei der Querstrich natürlich den Übergang zum konjugiert komplexen bedeutet.*

Dann ist $\mathrm{cl}\,(\mathcal{A}) = C_{\mathbb{C}}(B)$.

Beweis: Sei $\mathcal{A}_{\mathbb{R}}$ die Unteralgebra der reellwertigen Funktionen aus $\mathcal{A}$. Dann enthält $\mathcal{A}_{\mathbb{R}}$ die reellwertigen konstanten Funktionen und ist punktetrennend. Denn sind $s, t \in B$ mit $s \neq t$, so existiert nach Voraussetzung ein $x \in \mathcal{A}$ mit $x(s) \neq x(t)$. Dann sind der Realteil $\Re(x) = \frac{1}{2}(x + \overline{x})$ und der Imaginärteil $\Im(x) = \frac{1}{2i}(x - \overline{x})$ Elemente aus $\mathcal{A}_{\mathbb{R}}$ und es ist $\Re(x(s)) \neq \Re(x(t))$ oder $\Im(x(s)) \neq \Im(x(t))$. Wegen des (reellen) Satzes von Stone ist also $\mathrm{cl}\,(\mathcal{A}_{\mathbb{R}}) = C_{\mathbb{R}}(B)$, wobei $C_{\mathbb{R}}(B)$ natürlich die Menge der auf B stetigen reellwertigen Funktionen bedeutet. Wegen $\mathcal{A} = \mathcal{A}_{\mathbb{R}} + i\mathcal{A}_{\mathbb{R}}$ und $C_{\mathbb{C}}(B) = C_{\mathbb{R}}(B) + iC_{\mathbb{R}}(B)$ ist dann

$$\mathrm{cl}\,(\mathcal{A}) = \mathrm{cl}\,(\mathcal{A}_{\mathbb{R}}) + i\mathrm{cl}\,(\mathcal{A}_{\mathbb{R}}) = C_{\mathbb{C}}(B) + iC_{\mathbb{R}}(B) = C_{\mathbb{C}}(B),$$

was zu beweisen war. □

Anwendung Als Anwendung des letzten Satzes wollen wir den zweiten Weierstraßschen Approximationssatz beweisen, dass nämlich die Menge der reellwertigen trigonometrischen Polynome

$$\mathcal{T} := \left\{ u : u(t) = \frac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos kt + b_k \sin kt),\ a_k, b_k \in \mathbb{R},\ (k = 0, \ldots, n),\ n \in \mathbb{N} \right\}$$

dicht (bezüglich gleichmäßiger Konvergenz) im Raum $C_{\mathbb{R}}^{2\pi}$ der 2π-periodischen, reellwertigen stetigen Funktionen ist. Sei $B := \{z \in \mathbb{C} : |z| = 1\}$ und $C_{\mathbb{C}}^{2\pi}$ der lineare Raum der 2π-periodischen, stetigen, komplexwertigen Funktionen, normiert durch die Maximumnorm auf $[0, 2\pi]$. Definiert man $\phi\colon C_{\mathbb{C}}(B) \longrightarrow C_{\mathbb{C}}^{2\pi}$ durch $\phi(x)(t) := x(e^{it})$, so ist ϕ linear und bijektiv, ferner $\|\phi(x)\|_\infty = \|x\|_\infty$. Also ist ϕ auch isometrisch. Als lineare normierte Räume können daher $(C_{\mathbb{C}}(B), \|\cdot\|_\infty)$ und $(C_{\mathbb{C}}^{2\pi}, \|\cdot\|_\infty)$ identifiziert werden. Nun definiere man

$$\mathcal{T}_{\mathbb{C}} := \left\{ u : u(t) = \sum_{k=-n}^{n} c_k e^{ikt},\ c_k \in \mathbb{C}\ (k = -n, \ldots, n),\ n \in \mathbb{N} \right\}$$

und setze $\mathcal{A} := \phi^{-1}(\mathcal{T}_{\mathbb{C}})$. Dann genügt $\mathcal{A} \subset C_{\mathbb{C}}(B)$ den Voraussetzungen von Satz 7.1.2, also ist $\mathrm{cl}\,(\mathcal{A}) = C_{\mathbb{C}}(B)$ und dann auch $\mathrm{cl}\,(\mathcal{T}_{\mathbb{C}}) = C_{\mathbb{C}}^{2\pi}$. Da aber leicht einzusehen ist, dass $\mathcal{T}_{\mathbb{C}} = \mathcal{T} + i\mathcal{T}$ und $C_{\mathbb{C}}^{2\pi} = C_{\mathbb{R}}^{2\pi} + iC_{\mathbb{R}}^{2\pi}$, folgt der zweite Weierstraßsche Approximationssatz.

7.2 Der Satz von Korovkin

In diesem kurzen Abschnitt stellen wir einen weiteren schönen Beweis des Weierstraßschen Approximationssatzes vor. Dieser benutzt den jetzt folgenden Satz von Korovkin.

Satz 7.2.1 (Korovkin) *Sei* $\{K_n\}$ *eine Folge linearer, stetiger Operatoren des linearen normierten Raumes* $(C[\alpha, \beta], \|\cdot\|_\infty)$ *in sich, die außerdem noch monoton*[3] *sind. Sei*

[3] In $C[\alpha, \beta]$ sei eine Halbordnung $\leq$ dadurch eingeführt, dass man $x \leq y$ für $x, y \in C[\alpha, \beta]$ schreibt, falls $x(t) \leq y(t)$ für alle $t \in [\alpha, \beta]$ gilt. Ein Operator $K\colon C[\alpha, \beta] \longrightarrow C[\alpha, \beta]$ heißt dann *monoton*, falls aus $x \leq y$ folgt, dass $K(x) \leq K(y)$.

$x_k \in C[\alpha,\beta]$ durch $x_k(t) := t^k$ definiert. Es gelte

$$\lim_{n\to\infty} \|K_n(x_k) - x_k\|_\infty = 0, \qquad k = 0,1,2.$$

Dann gilt

$$\lim_{n\to\infty} \|K_n(x) - x\|_\infty = 0 \quad \textit{für alle } x \in C[\alpha,\beta].$$

Beweis: Seien $x \in C[\alpha,\beta]$ und $\epsilon > 0$ vorgegeben. Da x auf dem kompakten Intervall $[\alpha,\beta]$ gleichmäßig stetig ist, existiert ein $\delta = \delta(\epsilon) > 0$ mit

$$s,t \in [\alpha,\beta],\ |s-t| \le \delta \Longrightarrow |x(s) - x(t)| \le \frac{\epsilon}{2}.$$

Weiter gilt

$$s,t \in [\alpha,\beta],\ |s-t| > \delta \Longrightarrow |x(s) - x(t)| \le 2\,\|x\|_\infty \le 2\,\|x\|_\infty \left(\frac{s-t}{\delta}\right)^2$$

und folglich

$$s,t \in [\alpha,\beta] \Longrightarrow |x(s) - x(t)| \le \frac{\epsilon}{2} + 2\,\|x\|_\infty \left(\frac{s-t}{\delta}\right)^2.$$

Für alle $s,t \in [\alpha,\beta]$ ist daher

$$p_s(t) := x(s) - \frac{\epsilon}{2} - 2\,\|x\|_\infty \left(\frac{s-t}{\delta}\right)^2 \le x(t) \le x(s) + \frac{\epsilon}{2} + 2\,\|x\|_\infty \left(\frac{s-t}{\delta}\right)^2 =: q_s(t).$$

Da K_n monoton ist, ist $K_n(p_s) \le K_n(x) \le K_n(q_s)$ für alle $s \in [\alpha,\beta]$. Weiter ist

$$\begin{aligned} q_s(t) &= x(s) + \frac{\epsilon}{2} + 2\,\|x\|_\infty \left(\frac{s-t}{\delta}\right)^2 \\ &= \underbrace{\left(x(t) + \frac{\epsilon}{2} + \frac{2\,\|x\|_\infty s^2}{\delta^2}\right)}_{=:a(s)} \underbrace{x_0(t)}_{=1} \underbrace{-\frac{4\,\|x\|_\infty\, s}{\delta^2}}_{=:b(s)} \underbrace{x_1(t)}_{=t} + \underbrace{\frac{2\,\|x\|_\infty}{\delta^2}}_{=:c} \underbrace{x_2(t)}_{=t^2} \\ &= a(s)x_0(t) + b(s)x_1(t) + cx_2(t). \end{aligned}$$

Da K_n linear ist und $\lim_{n\to\infty} \|K_n(x_k) - x_k\|_\infty = 0$, $k = 0,1,2$, vorausgesetzt ist, erhalten wir für alle hinreichend großen n:

$$\begin{aligned} \|K_n(q_s) - q_s\|_\infty &= \|a(s)[K_n(x_0) - x_0]\|_\infty + b(s)[K_n(x_1) - x_1 + c[K_n(x_2) - x_2]\|_\infty \\ &\le a\,\|K_n(x_0) - x_0\|_\infty + b\,\|K_n(x_1) - x_1\|_\infty + c\,\|K_n(x_2) - x_2\|_\infty \\ &\le \frac{\epsilon}{2} \end{aligned}$$

für alle $s \in [\alpha,\beta]$. Hierbei haben wir

$$a := \max_{s\in[\alpha,\beta]} |a(s)|, \qquad b := \max_{s\in[\alpha,\beta]} |b(s)|$$

gesetzt. Entsprechend ist auch

$$\|K_n(p_s) - p_s\|_\infty \leq \frac{\epsilon}{2} \quad \text{für alle } s \in [\alpha, \beta]$$

und alle hinreichend großen n. Daher existiert ein $N_0 \in \mathbb{N}$ mit

$$-\frac{\epsilon}{2} \leq K_n(q_s)(t) - q_s(t) \leq \frac{\epsilon}{2}, \qquad -\frac{\epsilon}{2} \leq K_n(p_s)(t) - p_s(t) \leq \frac{\epsilon}{2}$$

und damit auch

$$p_s(t) - \frac{\epsilon}{2} \leq K_n(p_s)(t) \leq K_n(x)(t) \leq K_n(q_s)(t) \leq q_s(t) + \frac{\epsilon}{2}$$

für alle $n \geq N_0$ und alle $s, t \in [\alpha, \beta]$. Insbesondere ist

$$p_s(t) - \frac{\epsilon}{2} \leq K_n(x)(t) \leq q_s(t) + \frac{\epsilon}{2}$$

für alle $n \geq N_0$ und alle $s, t \in [\alpha, \beta]$. Mit $s = t \in [\alpha, \beta]$ ist also

$$p_t(t) - \frac{\epsilon}{2} = x(t) - \epsilon \leq K_n(x)(t) \leq x(t) + \epsilon = q_t(t) + \frac{\epsilon}{2}$$

für alle $n \geq N_0$ und alle $t \in [\alpha, \beta]$. Damit ist gezeigt, dass

$$\|K_n(x) - x\|_\infty \leq \epsilon \quad \text{für alle } n \geq N_0.$$

Der Satz von Korovkin ist bewiesen. □

Bemerkung: Den ersten Weierstraßschen Approximationssatz erhält man aus dem Satz von Korovkin, indem als K_n die *Bernstein-Operatoren* $B_n \colon C[0,1] \longrightarrow \Pi_n$ nimmt, welche durch

$$B_n(x)(t) := \sum_{i=0}^{n} \binom{n}{i} x\left(\frac{i}{n}\right) t^i (1-t)^{n-i}$$

definiert sind. Hierbei wird angenommen, dass $[\alpha, \beta] = [0, 1]$, was man durch eine einfache Variablentransformastion erreichen kann. Offensichtlich sind die Bernstein-Operatoren linear und monoton. Man definiere $x_k \in C[0,1]$, $k = 0, 1, 2$, durch $x_k(t) := t^k$. Dann ist

$$B_n(x_0)(t) = \sum_{i=0}^{n} \binom{n}{i} t^i (1-t)^{n-i} = (t + (1-t))^n = 1 = x_0(t).$$

Man kann leicht zeigen, dass

$$B_n(x_1)(t) = \sum_{i=0}^{n} \binom{n}{i} \left(\frac{i}{n}\right) t^i (1-t)^{n-i} = t = x_1(t)$$

und

$$B_n(x_2)(t) = \sum_{i=0}^{n} \binom{n}{i} \left(\frac{i}{n}\right)^2 t^i (1-t)^{n-i} = t^2 + \frac{t(1-t)}{n} = x_2(t) + \frac{t(1-t)}{n}.$$

Die Voraussetzungen des Satzes von Korovkin sind also erfüllt und man erhält die Aussage des Weierstraßschen Approximationssatzes. □

7.3 Die Müntzschen Sätze

In diesem wiederum kurzen Abschnitt wollen wir auf die klassischen Müntzschen Sätze hinweisen. Eine ausführliche Darstellung findet man bei E. W. CHENEY (1966, S. 193 ff.).

Das abstrakte Problem, das durch die Müntzschen Sätze in einem Spezialfall gelöst wird, ist das folgende:

- Sei $(X, \|\cdot\|)$ ein linearer normierter Raum und $S \subset X$ eine Teilmenge. Mit $\operatorname{span}(S)$ werde der von S erzeugte lineare Teilraum bezeichnet, also der kleinste S enthaltende lineare Teilraum von X. Unter welchen Voraussetzungen an S ist dann $\operatorname{cl}(\operatorname{span}(S)) = X$?

Beispiel: Sei $(X, \|\cdot\|) = (C[\alpha, \beta], \|\cdot\|_\infty)$ uns $S = \{1, t, t^2, \ldots\}$. Dann ist $\operatorname{span}(S) = \Pi$ und $\operatorname{cl}(\operatorname{span}(S)) = X$ wegen des Weierstraßschen ASpproximationssatzes. □

Teilmengen $S \subset X$ mit $\operatorname{cl}(\operatorname{span}(S)) = X$ heißen *fundamental* in X. Da man leicht zeigen kann, dass

$$\operatorname{span}(S) = \left\{ \sum_{i=1}^{n} \alpha_i s_i : \alpha_i \in \mathbb{R},\ s_i \in S\ (i = 1, \ldots, n),\ n \in \mathbb{N} \right\}$$

die Menge der endlichen Linearkombinationen aus Elementen von S ist, ist also eine Menge S genau dann fundamental, wenn man jedes Element von X durch eine Folge aus endlichen Linearkombinationen von Elementen aus S beliebig genau approximieren kann.

In den (beiden) Müntzschen Sätzen ist

(a) $(X, \|\cdot\|) = (C[0,1], \|\cdot\|_2)$, $S = \{t^{p_1}, t^{p_2}, \cdots\}$ mit $0 \le p_1 < p_2 < \cdots$

bzw.

(b) $(X, \|\cdot\|) = (C[0,1], \|\cdot\|_\infty)$, $S = \{1, t^{p_1}, t^{p_2}, \ldots\}$ mit $1 \le p_1 < p_2 < \cdots$.

Verblüffend an den Müntzschen Sätzen ist, dass eine Verbindung zwischen zwei scheinbar unkorellierten Tatsachen aufgezeigt wird, nämlich

$$S = \{1, t, t^2, \ldots\} \text{ ist fundamental in } (C[0,1], \|\cdot\|_\infty)$$

und

$$\sum_{j=1}^{\infty} \frac{1}{j} = +\infty.$$

Die beiden Müntzschen Sätze sagen nämlich aus, dass in den oben angegebenen Fällen (a) und (b) die jeweilige Menge S genau dann fundamental in $(C[0,1], \|\cdot\|_2)$ bzw. $(C[0,1], \|\cdot\|_\infty)$ ist, wenn

$$\sum_{j=2}^{\infty} \frac{1}{p_j} = +\infty.$$

Wir formulieren und beweisen (mit kleineren Lücken) nun den ersten Müntzschen Satz.

Satz 7.3.1 *$S = \{t^{p_1}, t^{p_2}, \ldots\}$ mit $0 \le p_1 < p_2 < \cdots$ ist genau dann fundamental in $(C[0,1], \|\cdot\|_2)$, wenn $\sum_{j=2}^{\infty} 1/p_j = +\infty$ (wir summieren erst ab $j = 2$, da $p_1 = 0$ möglich ist).*

Beweis: Man definiere $M_n := \operatorname{span}\{t^{p_1}, \ldots, t^{p_n}\}$, einen endlichdimensionalen Teilraum in dem Prä-Hilbertraum $(C[0,1], (\cdot,\cdot))$, wobei das innere Produkt natürlich durch

$$(x, y) = \int_0^1 x(t)y(t)\,dt$$

gegeben ist. Die zugehörige Norm ist

$$\|x\|_2 = (x,x)^{1/2} = \left(\int_0^1 x^2(t)\,dt\right)^{1/2}.$$

Ist $z \in C[0,1]$, so ist der Abstand von z zu M_n (bezüglich $\|\cdot\|_2$) definiert als

$$d(z, M_n) := \inf_{x \in M_n} \|x - z\|_2$$

(hier hätten wir auch min statt inf schreiben können, da ein Approximationsproblem bezüglich eines endlichdimensionalen linearen Teilraums stets lösbar ist) und es gilt offenbar:

- S ist genau dann fundamental in $(C[0,1], \|\cdot\|_2)$, wenn $\lim_{n\to\infty} d(z, M_n) = 0$ für alle $z \in C[0,1]$.

Dies gilt natürlich viel allgemeiner. Wichtig ist nun der erste Schritt im Beweis:

(a) Es ist

$$\lim_{n\to\infty} d(z, M_n) = 0 \quad \text{für alle } z \in C[0,1]$$

genau dann, wenn

$$\lim_{n\to\infty} d(z, M_n) = 0 \quad \text{für alle } z \in C[0,1] \text{ mit } z(t) = t^m,\ m = 0, 1, 2, \ldots.$$

Denn: Zu zeigen ist natürlich nur die Richtung $\Longleftarrow$. Diese folgt aber sofort aus dem Weierstraßschen Approximationssatz. Nach diesem ist Π dicht in $(C[0,1], \|\cdot\|_\infty)$ und daher trivialerweise auch in $(C[0,1], \|\cdot\|_2)$.

Die Hauptarbeit beim Beweis steckt im Nachweis der folgenden Aussage:

(b) Mit $z(t) := t^m$, $m = 0, 1, \ldots$, ist

$$d^2(z, M_n) = \frac{1}{2m+1} \prod_{j=1}^{n} \left(\frac{m - p_j}{m + p_j + 1}\right)^2.$$

Denn: Zur Abkürzung sei $x_j(t) := t^{p_j}$. O. B. d. Ä. ist $m \notin \{p_1, \ldots, p_n\}$. Wir wissen, dass genau eine beste Approximierende an z in M_n existiert (beachte: $(C[0,1], \|\cdot\|_2)$ ist strikt konvex und $M_n \subset C[0,1]$ ein endlichdimensionaler linearer Teilraum, sei dies etwa $x^* = \sum_{j=1}^n \alpha_j x_j$. Diese beste Approximierende ist durch $x^* - z \perp M_n$ charakterisiert. Insbesondere ist

$$(x^* - z, x_i) = 0, \qquad i = 1, \ldots, n.$$

Dies ergibt die *Normalgleichungen*

$$\sum_{j=1}^n (x_i, x_j)\alpha_j = (z, x_i), \qquad i = 1, \ldots, n.$$

Mit $d := d(z, M_n)$ ist ferner

$$d^2 = \|x^* - z\|_2^2 = (x^* - z, x^* - z) = (z - x^*, z)$$

bzw.

$$(x^*, z) + d^2 = \sum_{j=1}^n (x_j, z)\alpha_j + d^2 = (z, z).$$

Insgesamt hat man damit für die $n+1$ Unbekannten $\alpha_1, \ldots, \alpha_n, d^2$ das lineare Gleichungssystem

$$\begin{aligned} \sum_{j=1}^n (x_i, x_j)\alpha_j + 0 \cdot d^2 &= (z, x_i), \quad i = 1, \ldots, n, \\ \sum_{j=1}^n (x_j, z)\alpha_j + 1 \cdot d^2 &= (z, z). \end{aligned}$$

Aus der Cramerschen Regel erhält man

$$d^2 = \frac{\det \begin{pmatrix} (x_1,x_1) & \cdots & (x_1,x_n) & (x_1,z) \\ \vdots & \ddots & \vdots & \vdots \\ (x_n,x_1) & \cdots & (x_n,x_n) & (x_n,z) \\ (z,x_1) & \cdots & (z,x_n) & (z,z) \end{pmatrix}}{\det \begin{pmatrix} (x_1,x_1) & \cdots & (x_1,x_n) \\ \vdots & \ddots & \vdots \\ (x_n,x_1) & \cdots & (x_n,x_n) \end{pmatrix}}.$$

Berücksichtigt man

$$(x_i, x_j) = \int_0^1 x_i(t)x_j(t)\,dt = \int_0^1 t^{p_i+p_j}\,dt = \frac{1}{p_i + p_j + 1}$$

und

$$(x_i, z) = \int_0^1 x_i(t)z(t)\,dt = \int_0^1 t^{p_i+m}\,dt = \frac{1}{p_i + m + 1},$$

so erhält man

$$d^2 = \frac{\det\begin{pmatrix} \frac{1}{2p_1+1} & \cdots & \frac{1}{p_1+p_n+1} & \frac{1}{p_1+m+1} \\ \vdots & \ddots & \vdots & \vdots \\ \frac{1}{p_n+p_1+1} & \cdots & \frac{1}{2p_n+1} & \frac{1}{p_n+m+1} \\ \frac{1}{p_1+m+1} & \cdots & \frac{1}{p_n+m+1} & \frac{1}{2m+1} \end{pmatrix}}{\det\begin{pmatrix} \frac{1}{2p_1+1} & \cdots & \frac{1}{p_1+p_n+1} \\ \vdots & \ddots & \vdots \\ \frac{1}{p_n+p_1+1} & \cdots & \frac{1}{2p_n+1} \end{pmatrix}}.$$

Nun benutzt man eine Identität, die auf Cauchy zurückgeht:

$$\det\begin{pmatrix} \frac{1}{a_1+b_1} & \cdots & \frac{1}{a_1+b_n} \\ \vdots & \ddots & \vdots \\ \frac{1}{a_n+b_1} & \cdots & \frac{1}{a_n+b_n} \end{pmatrix} = \frac{\prod\limits_{1\le i<j\le n}(a_j-a_i)(b_j-b_i)}{\prod\limits_{1\le i,j\le n}(a_i+b_j)}.$$

Diese Identität wollen wir nicht beweisen, sondern nur auf E. W. Cheney (1966, S. 195) hinweisen. Setzt man $a_i = b_i = p_i + \frac{1}{2}$, $i = 1, \ldots, n$, so erhält man

$$\det\begin{pmatrix} \frac{1}{2p_1+1} & \cdots & \frac{1}{p_1+p_n+1} \\ \vdots & \ddots & \vdots \\ \frac{1}{p_n+p_1+1} & \cdots & \frac{1}{2p_n+1} \end{pmatrix} = \frac{\prod\limits_{1\le i<j\le n}(p_j-p_i)^2}{\prod\limits_{1\le i,j\le n}(p_i+p_j+1)}.$$

Entsprechend ist

$$\begin{aligned} &\det\begin{pmatrix} \frac{1}{2p_1+1} & \cdots & \frac{1}{p_1+p_n+1} & \frac{1}{p_1+m+1} \\ \vdots & \ddots & \vdots & \vdots \\ \frac{1}{p_1+p_n+1} & \cdots & \frac{1}{2p_n+1} & \frac{1}{p_n+m+1} \\ \frac{1}{p_1+m+1} & \cdots & \frac{1}{p_n+m+1} & \frac{1}{2m+1} \end{pmatrix} \\ &= \frac{\prod\limits_{1\le i<j\le n}(p_j-p_i)^2 \prod\limits_{j=1}^{n}(m-p_j)^2}{\prod\limits_{1\le i,j\le n}(p_i+p_j+1)\prod\limits_{j=1}^{n}(m+p_j+1)^2(2m+1)}. \end{aligned}$$

Damit erhalten wir, wie in (b) behauptet, die Beziehung

$$d^2(z, M_n) = \frac{1}{2m+1} \prod_{j=1}^{n} \frac{(m-p_j)^2}{(m+p_j+1)^2}.$$

(c) Sei $0 < a_j \neq 1$ und $a_j \to 0$. Dann gilt

$$\prod_{j=1}^{\infty}(1-a_j) = 0 \Longleftrightarrow \sum_{j=1}^{\infty} a_j = +\infty.$$

Denn (siehe E. W. CHENEY (1966, S. 196)): Man wähle m so groß, dass $a_j < \frac{1}{2}$ für alle $j \geq m$. Für alle solche j ist dann

$$\begin{aligned} \left|\frac{\ln(1-a_j)}{a_j} + 1\right| &= \left|\frac{-a_j - a_j^2/2 - a_j^3/3 - \cdots}{a_j} + 1\right| \\ &= \left|\frac{a_j}{2} + \frac{a_j^2}{3} + \frac{a_j^3}{4} + \cdots\right| \\ &< \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^3 + \cdots \\ &= \frac{1}{2}. \end{aligned}$$

Für $j \geq m$ ist daher

$$-\frac{3}{2} < \frac{\ln(1-a_j)}{a_j} < -\frac{1}{2}.$$

Hieraus folgt, dass die beiden Reihen $\sum_{j=1}^{\infty} \ln(1-a_j)$ und $\sum_{j=1}^{\infty} a_j$ entweder beide konvergieren oder beide divergieren. Daher gilt $\sum_{j=1}^{\infty} a_j = +\infty$ genau dann wenn $\sum_{j=1}^{\infty} \ln(1-a_j) = \ln \prod_{j=1}^{\infty}(1-a_j) = -\infty$ bzw. $\prod_{j=1}^{\infty}(1-a_j) = 0$.

(d) Mit $z(t) := t^m$, $m = 0, 1, \ldots$, gilt $\lim_{n\to\infty} d(z, M_n) = 0$ genau dann, wenn $\sum_{j=2}^{\infty}(1/p_j) = \infty$. Wegen (a) ist dann der Satz bewiesen.

Denn: Wegen

$$d(z, M_n) = \frac{1}{\sqrt{2m+1}} \prod_{j=1}^{n} \left|\frac{m-p_j}{m+p_j+1}\right|$$

gilt $\lim_{n\to\infty} d(z, M_n) = 0$ genau dann, wenn

$$(*) \qquad \lim_{n\to\infty} \prod_{j=1}^{n} \left|\frac{m-p_j}{m+p_j+1}\right| = \prod_{j=1}^{\infty} \left|\frac{m-p_j}{m+p_j+1}\right| = 0.$$

Um dies nachzuweisen, machen wir eine Fallunterscheidung (siehe A. SCHÖNHAGE (1971, S. 50 ff.)). Im ersten Fall ist die Folge $\{p_j\}$ beschränkt, etwa $p_j \leq \sigma$ für alle j. Dann ist

$$\sum_{j=2}^{\infty} \frac{1}{p_j} = \infty,$$

und auch $(*)$ ist erfüllt. Denn es gilt

$$\frac{|m-p_j|}{m+p_j+1} \le \frac{m+p_j}{m+p_j+1} \le \frac{m+\sigma}{m+\sigma+1}$$

und damit

$$\prod_{j=1}^{n}\left|\frac{m-p_j}{m+p_j+1}\right| \le \left(\frac{m+\sigma}{m+\sigma+1}\right)^n \to 0 \quad \text{mit } n\to\infty.$$

Im zweiten Fall kommen alle $m = 0, 1, 2, \ldots$ unter den p_j vor, d. h. es ist $\{0, 1, \ldots\} \subset \{p_1, p_2, \ldots\}$. Dann ist einerseits $d(z, M_n) = 0$ für alle hinreichend großen n und damit $(*)$ erfüllt, andererseits $\sum_{j=2}^{\infty}(1/p_j)$ divergent, da diese Reihe die harmonische Reihe als Teilreihe enthält. Im dritten Fall ist $\{p_j\}$ nicht beschränkt bzw. $p_j \to \infty$ (da die p_j aufsteigend angeordnet sind) und es existiert ein m mit $m \notin \{p_1, p_2, \ldots\}$. Da $\lim_{n\to\infty} d(z, M_n) = 0$ äquivalent zu $(*)$ ist und $p_j \ge m+1$ für fast alle (d. h. bis auf endlich viele) j gilt, ist $(*)$ äquivalent zu

$$0 = \prod_{j:p_j\ge m+1}\frac{|m-p_j|}{m+p_j+1} = \prod_{j:p_j\ge m+1}\frac{p_j-m}{m+p_j+1} = \prod_{j:p_j\ge m+1}\Bigg(1-\underbrace{\frac{2m+1}{m+p_j+1}}_{\in(0,1)}\Bigg).$$

Wegen (c) ist dies wiederum äquivalent zu

$$(**) \qquad \sum_{j:p_j\ge m+1}\frac{1}{m+p_j+1} = \infty.$$

Für alle j mit $p_j \ge m+1$ ist

$$\frac{1}{p_j} > \frac{1}{m+p_j+1} \ge \frac{1}{2p_j}$$

und daher ist $(**)$ äquivalent mit

$$\sum_{j:p_j\ge m+1}\frac{1}{p_j} = \infty \quad \text{bzw.} \quad \sum_{j=2}\frac{1}{p_j} = \infty.$$

Damit ist der erste Müntzsche Satz bewiesen. □

Für einen Beweis des zweiten Müntzschen satzes verweisen wir auf die zahlreiche literatur, etwa E. W. Cheney (1966, S. 197 ff.) und A. Schönhage (1971, S. 49 ff.).

7.4 Die Jackson-Sätze

Durch die Jackson-Sätze wird der T-Minimalabstand $E_n(z) = d(z, \Pi_n)$ in Abhängigkeit von der Glattheit der zu approximierenden Funktion z und n abgeschätzt. Hierbei betrachtet man das Intervall $[-1, 1]$, so dass

$$E_n(z) = \inf_{p\in\Pi_n}\|z-p\|_{\infty,[-1,1]} = \inf_{c_0,\ldots,c_n}\max_{t\in[-1,1]}\left|z(t)-\sum_{j=0}^{n}c_jt^j\right|.$$

Man beginnt mit entsprechenden Abschätzungen in $(C_{2\pi}, \|\cdot\|)$ und $\mathcal{T}_n$ statt Π_n, wobei $C_{2\pi}$ die Menge der 2π-periodischen reellwertigen Funktionen bedeutet und

$$\mathcal{T}_n := \left\{ u : u(t) = \frac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos kt + b_k \sin kt) \right\}$$

die Menge der trigonometrischen Polynome vom Grad $\leq n$ ist. Für $z \in C_{2\pi}$ definieren wir wie oben

$$E_n(z) := \inf_{u \in \mathcal{T}_n} \|z - u\|_\infty.$$

Erstes Ziel ist der Beweis des ersten Jackson-Satzes (siehe z. B. E. W. Cheney (1966, S. 142)).

Satz 7.4.1 (Jackson I) *Für jedes $z \in C^1_{2\pi}$ ist*

$$E_n(z) \leq \frac{\pi}{2(n+1)} \|z'\|_\infty,$$

wobei die Konstante $\pi/(2(n+1))$ bestmöglich ist.

Zum Beweis werden einige Hilfssätze benötigt, die wir zunächst beweisen (siehe z. B. E. W. Cheney (1966, S. 139 ff.)).

Lemma 7.4.2 *Für $k \in \{0, \ldots, n\}$ ist*

$$\int_0^\pi \sin kt \cdot \operatorname{sign}(\sin nt)\, dt = 0.$$

Beweis: Da der Integrand eine gerade Funktion ist, genügt es

$$\int_{-\pi}^{\pi} \sin kt \cdot \operatorname{sign}(\sin nt)\, dt = 0$$

zu zeigen. Da $\sin kt$ eine Linearkombination von e^{ikt} und e^{-ikt} ist, braucht man nur

$$I := \int_{-\pi}^{\pi} e^{imt} \cdot \operatorname{sign}(\sin nt)\, dt = 0$$

für $|m| < n$ nachzuweisen. Macht man die Variablentransformation $t = s + \pi/n$, so erhält man

$$\begin{aligned} I &= \int_{-\pi-\pi/n}^{\pi-\pi/n} e^{im(s+\pi/n)} \cdot \operatorname{sign}(\underbrace{\sin(ns+\pi)}_{=-\sin ns})\, ds \\ &= -e^{im\pi/n} \int_{-\pi-\pi/n}^{\pi-\pi/n} e^{ims} \cdot \operatorname{sign}(\sin ns)\, ds \\ &= -e^{im\pi/n} I, \end{aligned}$$

da der Integrand 2π-periodisch ist und das Integrationsintervall daher durch $[-\pi, \pi]$ ersetzt werden kann. Also ist

$$(1 + e^{im\pi/n}) I = 0.$$

Wegen $|m| < n$ ist $1 + e^{im\pi/n} \neq 0$ und daher $I = 0$. Das Lemma ist bewiesen. □

Lemma 7.4.3 *Es ist*

$$\min_{\alpha_1,\ldots,\alpha_{n-1}\in\mathbb{R}} \int_0^\pi \left| t - \sum_{k=1}^{n-1} \alpha_k \sin kt \right| dt = \frac{\pi^2}{2n}.$$

Beweis: Im ersten Teil des Beweises zeigen wir, dass

$$\inf_{\alpha_1,\ldots,\alpha_{n-1}\in\mathbb{R}} \int_0^\pi \left| t - \sum_{k=1}^{n-1} \alpha_k \sin kt \right| dt \geq \frac{\pi^2}{2n}.$$

Hierzu seien $\alpha_1, \ldots, \alpha_{n-1} \in \mathbb{R}$ beliebig vorgegeben. Dann ist

$$\begin{aligned}
\int_0^\pi \left| t - \sum_{k=1}^{n-1} \alpha_k \sin kt \right| dt &\geq \left| \int_0^\pi \left(t - \sum_{k=1}^{n-1} \alpha_k \sin kt \right) \cdot \operatorname{sign}(\sin nt)\, dt \right| \\
&= \left| \int_0^\pi t \cdot \operatorname{sign}(\sin nt)\, dt \right| \\
&\qquad \text{(Anwendung von Lemma 7.4.2)} \\
&= \left| \sum_{k=0}^{n-1} \int_{k\pi/n}^{(k+1)\pi/n} t \cdot \operatorname{sign}(\sin nt)\, dt \right| \\
&= \left| \sum_{k=0}^{n-1} (-1)^k \int_{k\pi/n}^{(k+1)\pi/n} t\, dt \right| \\
&= \left| \sum_{k=0}^{n-1} (-1)^k \frac{1}{2} \left[\left(\frac{(k+1)\pi}{n} \right)^2 - \left(\frac{k\pi}{n} \right)^2 \right] \right| \\
&= \frac{\pi^2}{2n^2} \left| \sum_{k=0}^{n-1} (-1)^k (2k+1) \right| \\
&= \frac{\pi^2}{2n}.
\end{aligned}$$

Für den letzten Schritt benutze man, dass

$$\sum_{k=0}^{n-1} (-1(^k(2k+1) = (-1)^{n-1} n,$$

was man leicht durch vollständige Induktion beweist. Damit ist nachgewiesen, dass

$$\inf_{\alpha_1,\ldots,\alpha_{n-1}\in\mathbb{R}} \int_0^\pi \left| t - \sum_{k=1}^{n-1} \alpha_k \sin kt \right| dt \geq \frac{\pi^2}{2n}.$$

Nun muss gezeigt werden, dass die untere Schranke $\pi^2/(2n)$ für eine spezielle Wahl von $\alpha_1, \ldots, \alpha_{n-1}$ angenommen wird. Hierzu sieht man sich im ersten Teil des Beweises an,

wo bei der Berechnung der unteren Schranke Ungleichungen eingingen, wo man also eventuell etwas verschenkt hat. Man erkennt, dass

$$\Phi(t) := t - \sum_{k=1}^{n-1} \alpha_k \sin kt$$

sein Vorzeichen in $(0,\pi)$ genau dort wechseln muss, wo es $\operatorname{sign}(\sin nt)$ tut, also in

$$t_i := \frac{i\pi}{n}, \qquad i = 1, \ldots, n-1.$$

Daher sollten $\alpha_1, \ldots, \alpha_{n-1}$ so bestimmt werden, dass

$$\sum_{k=1}^{n-1} \alpha_k \sin kt_i = t_i, \qquad i = 1, \ldots, n-1.$$

Dass dies in eindeutiger Weise möglich ist, ist eine Konsequenz der Tatsache, dass $\{\sin t, \ldots, \sin(n-1)t\}$ ein $(n-1)$-dimensionales Haarsches System auf $(0,\pi)$ ist[4]. Sind $\alpha_1, \ldots, \alpha_{n-1}$ auf diese Weise bestimmt, so besitzt Φ auf $(0,\pi)$ genau in den Punkten t_i, $i = 1, \ldots, n-1$, eine Nullstelle. Dass die so definierte Funktion Φ in den Punkten t_i, $i = 1, \ldots, n-1$, das Vorzeichen wechselt, überlegen wir uns indem wir das Gegenteil annehmen. Dann verschwindet Φ' in jedem der Intervalle (t_i, t_{i+1}), $i = 1, \ldots, n-2$, einmal in $(0, t_1)$ und in mindestens einem der Punkte t_i, also in mindestens n Punkten in $(0,\pi)$. Die Ableitung Φ' besitzt andererseits die Form $\Phi'(t) = 1 + \sum_{k=1}^{n-1} \beta_k \cos kt$ und hat daher höchstens $n-1$ Nullstellen in $(0,\pi)$. Damit ist das Lemma bewiesen. □

Bei festem $n \in \mathbb{N}$ und festen $A_1, \ldots, A_n$, an denen später "gedreht" wird, definiere man den Operator

$$L \colon C_{2\pi} \longrightarrow C_{2\pi}$$

durch

$$L(x)(t) := \frac{a_0}{2} + \sum_{k=1}^{n} A_k (a_k \cos kt + b_k \sin kt).$$

Hierbei seien a_k, b_k Fourier-Koeffizienten zu x, d. h.

$$a_k := \frac{1}{\pi} \int_{-/pi}^{\pi} x(s) \cos ks \, ds, \qquad b_k := \frac{1}{\pi} \int_{-\pi}^{\pi} x(s) \sin ks \, ds.$$

Dann gilt

Lemma 7.4.4 *Ist $x \in C^1_{2\pi}$, so ist*

$$(L(x) - x)(t) = \frac{1}{\pi} \int_{-\pi}^{\pi} \Phi(s) x'(t + \pi - s) \, ds$$

mit

$$\Phi(s) := \frac{1}{2} s + \sum_{k=1}^{n} \frac{(-1)^k}{k} A_k \sin ks.$$

[4]Denn angenommen, $\sum_{k=1}^{n-1} \alpha_k \sin kt$ besitze $n-1$ Nullstellen t_j in $(0,\pi)$, also $2(n-1)+1$ Nullstellen in $(-\pi,\pi)$. Andererseits ist $\mathcal{T}_{n-1}$ ein $2(n-1)+1$-dimensionaler Haarscher Teilraum auf $[0,2\pi)$ bzw. $[-\pi,\pi)$ (siehe Seite 101). Daher ist $\alpha_1 = \cdots = \alpha_{n-1} = 0$ und folglich $\{\sin t, \ldots, \sin(n-1)t\}$ ein $(n-1)$-dimensionales Haarsches System auf $(0,\pi)$.

Beweis: Es ist

$$\begin{aligned}
\frac{1}{\pi}\int_{-\pi}^{\pi}\Phi(s)x'(t+\pi-s)\,ds &= -\frac{1}{\pi}\Phi(s)x(t+\pi-s)\Big|_{s=-\pi}^{s=+\pi} \\
&\quad + \frac{1}{\pi}\int_{-\pi}^{\pi}\Phi'(s)x(t+\pi-s)\,ds \\
&\quad \text{(partielle Integration)} \\
&= -\frac{1}{\pi}\left[\frac{1}{2}\pi x(t)+\frac{1}{2}\pi x(t+2\pi)\right] \\
&\quad + \frac{1}{\pi}\int_{-\pi}^{\pi}\Phi'(s)x(t+\pi-s)\,ds \\
&\quad \text{(wegen } \Phi(\pm\pi)=\pm\tfrac{1}{2}\pi\text{)} \\
&= -x(t) \\
&\quad + \frac{1}{\pi}\int_{-\pi}^{\pi}\left[\frac{1}{2}+\sum_{k=1}^{n}(-1)^k A_k\cos ks\right]x(t+\pi-s)\,ds \\
&\quad \text{(wegen } x\in C_{2\pi} \text{ und der Definition von } \Phi\text{)} \\
&= -x(t) \\
&\quad + \frac{1}{\pi}\int_{-\pi}^{\pi}\left[\frac{1}{2}+\sum_{k=1}^{n}A_k\cos k(t-\tau)\right]x(\tau)\,d\tau \\
&\quad \text{(Mache Variablentransformation } \tau=t+\pi-s\text{)} \\
&= \frac{1}{\pi}\int_{-\pi}^{\pi}\left[\frac{1}{2}+\sum_{k=1}^{n}A_k(\cos kt\cos k\tau+\sin kt\sin k\tau)\right]x(\tau)\,d\tau \\
&\quad - x(t) \\
&\quad \text{(Additionstheorem für den Cosinus)} \\
&= (L(x)-x)(t) \\
&\quad \text{(Definition von } L \text{ und der Fourier-Koeffizienten).}
\end{aligned}$$

Damit ist das Lemma bewiesen. □

Nun können wir Satz 7.4.1 beweisen.

Beweis von Satz 7.4.1: Nach Lemma 7.4.4 ist, egal wie $A_1,\ldots,A_n$ gewählt werden:

$$\begin{aligned}
E_n(z) &= \inf_{u\in\mathcal{T}_n}\|z-u\|_\infty \\
&\le \|z-L(z)\|_\infty \\
&\le \|z'\|_\infty\frac{1}{\pi}\int_{-\pi}^{\pi}\left|\frac{t}{2}+\sum_{k=1}^{n}\frac{(-1)^k}{k}A_k\sin kt\right|dt \\
&= \|z'\|_\infty\frac{1}{\pi}\int_{0}^{\pi}\left|t+\sum_{k=1}^{n}\frac{2(-1)^k}{k}A_k\sin kt\right|dt.
\end{aligned}$$

Wegen Lemma 7.4.3 können $A_1,\ldots,A_n$ so gewählt werden, dass

$$E_n(z)\le\frac{\pi}{2(n+1)}\|z'\|_\infty.$$

Für den Beweis, dass die Konstante bestmöglich ist, verweisen wir auf E. W. CHENEY (1966, S. 142). □

Die Jackson-Sätze verlaufen nun sozusagen in zwei Richtungen: Einerseits betrachtet man weiter die Approximationsgüte bezüglich trigonometrischer Polynome und verändert die Glattheitsvoraussetzungen an die zu approximierende Funktion z, andererseits geht man (durch eine Variablentransformation) zu Aussagen zur Approximationsgüte bezüglich algebraischer Polynome über.

Das folgende Korollar findet man z. B. bei E. W. CHENEY (1966, S. 145).

Korollar 7.4.5 *Ist $z \in C^k_{2\pi}$ und $n > k$, so ist* [5]

$$E_n(z) \le \left(\frac{\pi}{2n+2}\right)^k \|z^{(k)}\|_\infty.$$

Beweis: Sei

$$e_n(z) := \inf_{u \in \mathcal{T}_n^0} \|z - u\|_\infty,$$

wobei

$$\mathcal{T}_n^0 := \left\{ u : u(t) = \sum_{k=1}^{n} (a_k \cos kt + b_k \sin kt),\ a_k, b_k \in \mathbb{R} \right\}.$$

Wir werden zeigen, dass die folgende Ungleichungskette gilt:

$$\begin{aligned} E_n(z) &\le \frac{\pi}{2n+2} e_n(z') \\ &\le \left(\frac{\pi}{2n+2}\right)^2 e_n(z'') \\ &\vdots \\ &\le \left(\frac{\pi}{2n+2}\right)^{k-1} e_n(z^{(k-1)}) \\ &\le \left(\frac{\pi}{2n+2}\right)^k \|z^{(k)}\|_\infty, \end{aligned}$$

womit das Korollar bewiesen sein wird.

Im ersten Schritt zeigen wir

(a) Es ist

$$E_n(z) \le \frac{\pi}{2n+2} e_n(z').$$

[5] Die Konstante kann man sogar verkleinern zu $\frac{1}{2}\pi(n+1)^{-k}$, was aber nicht gezeigt werden soll.

Denn: Sei $u^0 \in P_{\mathcal{T}_n^0}(z')$. Man definiere $U^0 \in \mathcal{T}_n$ durch $U^0(t) := \int_0^t u^0(s)\,ds$. Wegen Satz 7.4.1 ist y

$$\begin{aligned} E_n(z) &= E_n(z - U_0) \\ &\leq \frac{\pi}{2n+2}\|(z-U^0)'\|_\infty \\ &= \frac{\pi}{2n+2}\|z' - u^0\|_\infty \\ &= \frac{\pi}{2n+2}e_n(z'). \end{aligned}$$

Im nächsten Schritt zeigen wir:

(b) Es ist

$$e_n(z^{(\nu)}) \leq \frac{\pi}{2n+2}e_n(z^{(\nu+1)}), \qquad \nu = 1, \ldots, k-2.$$

Denn: Wir erinnern an den Operator $L\colon C_{2\pi} \longrightarrow C_{2\pi}$, der vor Lemma 7.4.4 in Abhängigkeit von Konstanten $A_1, \ldots, A_n$ definiert ist. Offenbar ist $L(z^{(\nu)}) \in \mathcal{T}_n^0$, $\nu = 1, \ldots, k$, da der konstante Term in der Fourier-Entwicklung von $z^{(\nu)}$ wegen

$$\frac{2}{\pi}\int_{-\pi}^{\pi} z^{(\nu)}(t)\,dt = \frac{2}{\pi}[z^{(\nu-1)}(\pi) - z^{(\nu-1)}(-\pi)] = 0, \qquad \nu = 1, \ldots, k.$$

verschwindet. Nun sei $u^\nu \in P_{\mathcal{T}_n^0}(z^{(\nu+1)})$ und $U^\nu(t) := \int_0^t u^\nu(s)\,ds$. Dann ist

$$\begin{aligned} e_n(z^{(\nu)}) &\leq \|z^{(\nu)} - (\underbrace{U^\nu + L(z^{(\nu)} - U^\nu)}_{\in \mathcal{T}_n^0})\|_\infty \\ &= \|(z^{(\nu)} - U^\nu) - L(z^{(\nu)} - U^\nu)\|_\infty \\ &\leq \frac{\pi}{2n+2}\|(z^{(\nu)} - U^\nu)'\|_\infty \\ &= \frac{\pi}{2n+2}\|z^{(\nu+1)} - u^\nu\|_\infty \\ &= \frac{\pi}{2n+2}e_n(z^{(\nu+1)}). \end{aligned}$$

Im letzten Schritt zeigen wir

(c) Es ist

$$e_n(z^{(k-1)}) \leq \frac{\pi}{2n+2}\|z^{(k)}\|_\infty.$$

Denn: Die Behauptung folgt aus

$$e_n(z^{(k-1)}) \leq \|z^{(k-1)} - \underbrace{L(z^{(k-1)})}_{\in \mathcal{T}_n^0}\|_\infty \leq \frac{\pi}{2n+2}\|z^{(k)}\|_\infty.$$

Damit ist der Satz bewiesen. □

Bevor wir auf Abschätzungen für $E_n(z) = d(z, \Pi_n)$ auf $[-1, 1]$ eingehen, folgen noch der zweite und der dritte Jackson-Satz.

Satz 7.4.6 (Jackson II) *Sei $z \in C_{2\pi}$ mit*

$$|z(s) - z(t)| \leq \lambda\,|s - t| \quad \textit{für alle } s, t \in \mathbb{R}.$$

Dann ist

$$E_n(z) \leq \frac{\pi\lambda}{2n+2}.$$

Gegenüber Jackson I wird also lediglich $\|z'\|_\infty$ durch die Lipschitzkonstante λ ersetzt.

Beweis: Sei $\delta > 0$ fest und

$$\Phi_\delta(t) := \frac{1}{2\delta}\int_{t-\delta}^{t+\delta} z(s)\,ds.$$

Dann ist

$$|\Phi_\delta'(t)| = \frac{1}{2\delta}|z(t+\delta) - z(t-\delta)| \leq \lambda.$$

Also ist $\Phi_\delta \in C_{2\pi}^1$ und $\|\Phi_\delta'\|_\infty \leq \lambda$. Aus Jackson I folgt

$$E_n(\Phi_\delta) \leq \frac{\pi\lambda}{2n+2}.$$

Ferner ist

$$\begin{aligned}
|\Phi_\delta(t) - z(t)| &\leq \frac{1}{2\delta}\int_{t-\delta}^{t+\delta} |z(s) - z(t)|\,ds \\
&\leq \frac{\lambda}{2\delta}\int_{t-\delta}^{t+\delta} |s - t|\,ds \\
&= \frac{\lambda}{2\delta}\left(\int_{t-\delta}^{t}(t-s)\,ds + \int_{t}^{t+\delta}(s-t)\,ds\right) \\
&= \frac{\lambda}{2}\delta.
\end{aligned}$$

Ist nun $p \in P_{\mathcal{T}_n}(\Phi_\delta)$, so ist

$$\begin{aligned}
E_n(z) &\leq \|p - z\|_\infty \\
&\leq \underbrace{\|p - \Phi_\delta\|_\infty}_{=E_n(\Phi_\delta)} + \underbrace{\|\Phi_\delta - z\|_\infty}_{\leq(\lambda/2)\delta} \\
&\leq \frac{\pi\lambda}{2n+2} + \frac{\lambda}{2}\delta.
\end{aligned}$$

Mit $\delta \to 0+$ folgt die Behauptung. □

Schließlich folgt noch

Satz 7.4.7 (Jackson III) *Für alle $z \in C_{2\pi}$ ist*

$$E_n(z) \leq \frac{3}{2}\omega\left(\frac{\pi}{n+1}\right).$$

Hierbei ist[6] *der sogenannte Stetigkeitsmodul* ω *von* z *definiert durch*

$$\omega(\delta) := \sup_{|s-t|\le\delta} |z(s) - z(t)|.$$

Beweis: Wie beim Beweis von Satz 7.4.6, dem Satz Jackson II, definiere man

$$\Phi_\delta(t) := \frac{1}{2\delta}\int_{t-\delta}^{t+\delta} z(s)\,ds$$

bei vorgegebenem $\delta > 0$. Dann ist

$$|\Phi'_\delta(t)| = \frac{1}{2\delta}|z(t+\delta) - z(t-\delta)| \le \frac{\omega(2\delta)}{2\delta}.$$

Ferner ist

$$\begin{aligned} |\Phi_\delta(t) - z(t)| &\le \frac{1}{2/delta}\int_{t-\delta}^{t+\delta} |z(s) - z(t)|\,ds \\ &= \frac{1}{2\delta}\left(\int_{t-\delta}^{t} |z(s) - z(t)|\,ds + \int_{t}^{t+\delta} |z(s) - z(t)|\,ds\right) \\ &\le \omega(\delta). \end{aligned}$$

Dann erhält man, genau wie im Beweis von Satz 7.4.6,

$$E_n(z) \le \frac{\pi}{2n+2}\cdot\frac{\omega(2\delta)}{2\delta} + \omega(\delta) \le \omega(2\delta)\left[1 + \frac{\pi}{4\delta(n+1)}\right].$$

Setzt man $\delta := 1/(2n+2)$, so erhält man die Behauptung. □

Nun sei der Ausgangsraum $(C[-1,1], \|\cdot\|_\infty)$ und

$$E_n(z) = d(z, \Pi_n) = \inf_{p\in\Pi_n} \|p - z\|_{\infty,[-1,1]}.$$

In dem folgenden Satz (siehe z. B. E. W. Cheney (1966, S. 147 ff.)) werden die den bisherigen Jackson-Sätzen entsprechenden Aussagen zusammengefasst.

Satz 7.4.8 (Jackson IV) *Sei* $z \in C[-1,1]$. *Dann gilt:*

(a) Es ist

$$E_n(z) \le \frac{3}{2}\omega\left(\frac{\pi}{n+1}\right).$$

(b) Ist $|z(s) - z(t)| \le \lambda\,|s-t|$ *sür alle* $s, t \in [-1,1]$, *so ist*

$$E_n(z) \le \frac{\pi\lambda}{2n+2}.$$

[6] Es kann gezeigt werden, dass in der Abschätzung $\frac{3}{2}$ durch 1 ersetzt werden kann und dass die abschätzung dann optimal ist.

(c) Ist $z \in C^{(k)}[-1,1]$ und $n > k$, so ist

$$E_n(z) \leq \left(\frac{\pi}{2}\right)^k \frac{\|z^{(k)}\|_\infty}{(n+1)n\cdots(n-k+2)}.$$

Beweis: □Definiere $y(t) := z(\cos t)$. dann ist y eine gerade, 2π-periodische stetige Funktion.Man wende mit geringen Zusatzüberlegungen die bisherigen Sätze an. Genaueres findet man bei E. W. Cheney (1966, S. 147 ff.).

Literaturverzeichnis

[1] Barbu, V. and T. Precupanu (2012) *Convexity and Optimization in Banach Spaces. Fourth Edition*. Springer-Verlag, Berlin-Heidelberg-New York-London-Paris-Tokyo.

[2] Braess, D. (1967) Über die Approximation mit Exponentialsummen. Computing 2, 209–321.

[3] Braess, D. (1986) *Nonlinear Approximation Theory*. Springer-Verlag, Berlin-Heidelberg-New York-London-Paris-Tokyo.

[4] Cheney, E. W. (1966) *Introduction to Approximation Theory.* McGraw-Hill Book Company, New York-St. Louis-San Francisco-Toronto-London-Sydney.

[5] Cheney, E. W. and H. L. Loeb (1961) Two New algorithms for Rational Approximation. Numer. Math. 3, 72–75.

[6] Cheney, E. W. and H. L. Loeb (1964) Generalized Rational Approximation. J. SIAM Numer. Anal. Ser. B Vol. 1, 11–25.

[7] Cheney, E. W. and M. J. D. Powell (1987) The Differential Correction Algorithm for generalized Rational Functions. Constr. Approx. 3, 249–256.

[8] Collatz, L. und W. Krabs (1973) *Approximationstheorie. Tschebyscheffsche Approximation mit Anwendungen.* B. G. Teubner, Stuttgart.

[9] Deutsch, F. (1980) Existence of Best Approximations. Journal of Approximation Theory 28, 132–154.

[10] Hewitt, E. and K. Stromberg (1965) *Real and Abstract Analysis*. Springer-Verlag, Berlin-Heidelberg-New York.

[11] Hirzebruch, F. und W. Scharlau (1971) *Einführung in die Funktionalanalysis*. Bibliographisches Institut, Mannheim-Wien-Zürich.

[12] Mangasarian, O. L. (1969) *Nonlinear Programming*. McGraw-Hill Book Company, New York.

[13] Meinardus, G. (1967) *Approximation of Functions: Theory and Numerical Methods*. Springer-Verlag, Berlin-Heidelberg-New York.

[14] Powell, M. J. D. (1981) *Approximation theory and methods*. Cambridge University Press, Cambridge-NewYork-New Rochelle-Melbourne-Sydney.

[15] Rice, J. R. (1969) *The Approximation of Functions. II. Nonlinear and Multivariate Theory*. Addison-Wesley, Reading, Mass.

[16] Schmidt, E. (1970) Zur Kompaktheit bei Exponentialsummen. Journal of Approximation Theory 3, 445–454.

[17] Schönhage, A. (1971) *Approximationstheorie*. Walter de Gruyter & Co, Berlin-New York.

[18] Watson, G. A. (1980) *Approximation Theory and Numerical Methods*. John Wiley & Sons, Chichester-New York-Brisbane-Toronto.

[19] Werner, H. (1969) Der Existenzsatz für das Tschebyscheffsche Approximationsproblem mit Exponentialsummen. In: *Funktionalanalytische Methoden der numerischen Mathematik* (L. Collatz und H. Unger, eds). ISNM 12, 133–143. Birkhäuser, Basel.

[20] Werner, J. (1984) *Optimization.Theory and Applications*. Friedr. Vieweg & Sohn, Braunschweig-Wiesbaden.

[21] Werner, J. (1992) *Numerische Mathematik 1*. Friedr. Vieweg & Sohn Verlagsgesellschaft, Braunschweig-Wiesbaden.