Mehr merkwürdige Mathematik

Jochen Werner
jochen.christa@t-online.de

Inhaltsverzeichnis

1 Einleitung

Dies ist eine Fortsetzung meiner Sammlung *Merkwürdige Mathematik*. Ich werde mich gelegentlich, vor allem am Anfang, auf für mich etwas dünneres Eis vorwagen, da ich z. B. mehr Respekt *vor* Algebra und Zahlentheorie als Kenntnisse *von* diesen mathematischen Teildisziplinen habe. Es wird aber auch auf einige Aussagen über Optimierungsaufgaben in linearen normierten Räumen und Anwendungen z. B. auf optimale Steuerungsprobleme eingegangen, die man etwa bei J. Werner (1984, 1985) finden

kann. Aufgenommen in diese Sammlung habe ich aber auch Fragen der nichtlinearen Funktionalanalysis, z. B. die analytische Einführung des Brouwerschen Abbildungsgrades und hierauf aufbauend Definition und Eigenschaften des Leray-Schauderschen Abbildungsgrades. Außerdem gab es für mich hier eine Gelegenheit, Ergebnisse aus meiner Dissertation darzustellen (und diese für mich noch einmal nachzuvollziehen).

2 Der euklidische Algorithmus

Wir wollen den euklidischen Algorithmus zur Berechnung des größten gemeinsamen Teilers nichtnegativer Zahlen sowie von Polynomen mit Koeffizienten in einem Körper schildern.

2.1 Nichtnegative ganze Zahlen

Mit dem euklidischen Algorithmus kann der größte gemeinsame Teiler $\mathrm{ggT}(a, b)$ zweier nichtnegativer ganzer Zahlen a und b berechnet werden[1]. Der euklidische Algorithmus gilt als der älteste bekannte nicht-triviale Algorithmus. In Buch VII (Proposition 1 und 2) der Elemente des Euklid wurde der Algorithmus für ganze Zahlen formuliert. Genauer lautet die Proposition 1 in Buch VII (siehe `http://www.opera-platonis.de/euklid/eb7/eb702.htm`):

- Wird von zwei ungleichen Zahlen ausgehend, immer wieder die kleinere von der größeren Zahl subtrahiert, und bleibt schließlich der Rest Eins, dann sind sie teilerfremd.

In Proposition 2 wird die Aufgabe formuliert:

- Zu zwei Zahlen, die nicht teilerfremd sind, den größten gemeinsamen Teiler (zu) finden.

Grundlage des euklidischen Algorithmus ist die *Division mit Rest*:

- Zu ganzen Zahlen $a \geq 0$ und $b > 0$ existieren eindeutig ganze Zahlen $q \geq 0$ und r mit $0 \leq r < b$ derart, dass $a = qb + r$.

In MATLAB ist die Division mit Rest z. B. durch `q=floor(a/b); r=a-q*b` realisierbar. Es ist leicht einzusehen, dass dann

$$\mathrm{ggT}(a, b) = \begin{cases} a, & \text{falls } b = 0, \\ \mathrm{ggT}(b, r), & \text{sonst.} \end{cases}$$

Dies ist Grundlage des *euklidischen Algorithmus*, den wir im folgenden Satz schildern.

[1] Hierbei heißt $d \in \mathbb{N}$ *größter gemeinsamer Teiler* von a und b, wenn d gemeinsamer Teiler von a und b ist und jeder andere gemeinsame Teiler von a und b auch d teilt. Die Existenz und Eindeutigkeit des größten gemeinsamen Teilers ist leicht einzusehen. Es wird vereinbart, dass $\mathrm{ggT}\,(0, 0) = 0$.

Satz 2.1 *Seien $a, b \in \mathbb{N}$ gegeben. Setze $r_0 := a$ und $r_1 := b$. Man berechne sukzessive nichtnegative ganze Zahlen $r_2, \ldots, r_{n-1}, r_n$ und $q_1, \ldots, q_n$ mit*

$$\begin{array}{rcll} r_0 & = & q_1 r_1 + r_2, & 0 < r_2 < r_1, \\ r_1 & = & q_2 r_2 + r_3, & 0 < r_3 < r_2, \\ & \vdots & & \vdots \\ r_{n-2} & = & q_{n-1} r_{n-1} + r_n, & 0 < r_n < r_{n-1}, \\ r_{n-1} & = & q_n r_n + 0. & \end{array}$$

Dann gilt:

(a) r_n teilt a und b.

(b) Es existieren $s, t \in \mathbb{Z}$ mit $sa + tb = r_n$.

(c) Es ist $r_n = \mathrm{ggT}\,(a, b)$.

Beweis: Wegen $r_{n-1} = q_n r_n$ gilt $r_n \mid r_{n-1}$. Sei $k \leq n$ und $r_n \mid r_{k-1}, \ldots, r_{n-1}, r_n$. Wegen $r_{k-2} = q_{k-1} r_{k-1} + r_k$ gilt dann auch $r_n \mid r_{k-2}$. Damit erhält man induktiv, dass r_n auch $r_1 = g$ und $r_0 = f$ teilt und (a) ist nachgewiesen. Zum Beweis von (b) zeigen wir für $k = 0, 1, \ldots, n$ die Existenz von $s_k, t_k \in \mathbb{Z}$ mit $s_k a + t_k b = r_k$. Dies ist für $k = 0$ und $k = 1$ trivialerweise richtig mit $(s_0, t_0) := (1, 0)$ und $(s_1, t_1) := (0, 1)$. Wir nehmen an, die Aussage sei für $k - 2$ und $k - 1$ richtig. Dann ist

$$\begin{array}{rcl} r_k & = & r_{k-2} - q_{k-1} r_{k-1} \\ & = & (s_{k-2} a + t_{k-2} b) - q_{k-1}(s_{k-1} a + t_{k-1} b) \\ & = & (\underbrace{s_{k-2} - q_{k-1} s_{k-1}}_{=:s_k}) a + (\underbrace{t_{k-2} - q_{k-1} t_{k-1}}_{=:t_k}) b \\ & = & s_k a + t_k b, \end{array}$$

damit ist die Aussage auch für k richtig, der Induktionsschritt vollzogen und auch (b) mit $s := s_n$, $t := t_n$ bewiesen. Aus (b) erhalten wir, dass jeder Teiler von a und b auch r_n teilt. Daher ist r_n der ggT von a und b und (c) ist bewiesen. □

Bemerkung 2.2 Die Aussage in (b) und (c), dass sich der ggT von zwei Zahlen $a, b \in \mathbb{N}$ als ganzzahlige Linearkombination von a und b darstellen lässt, wird gelegentlich *Lemma von Bézout* genannt. Beim Beweis von (b) haben wir den sogenannten *erweiterten euklidischen Algorithmus* geschildert. Eine einfache MATLAB-Implementation ist durch

```
function [g,s,t]=extEuclid(a,b);
s=1;t=0;
u=0;v=1;
while b>0
   q=floor(a/b);
   r=a-q*b;
   a=b;b=r;
```

```
   tmp=u;u=s-q*u;s=tmp;
   tmp=v;v=t-q*v;t=tmp;
end;
g=a;
```

gegeben. In MATLAB selbst gibt es die Funktion `gcd`. Durch `g=gcd(a,b)` wird der ggT g von a und b berechnet, durch `[g,s,t]=gcd(a,b)` werden neben dem ggT g auch noch die sogenannten *Bézout-Koeffizenten* $s, t \in \mathbb{Z}$ mit $g = sa + tb$ ausgegeben. □

2.2 Polynome mit Koeffizienten aus einem Körper

Der euklidische Algorithmus kann verallgemeinert werden zur Berechnung eines größten gemeinsamen Teilers zweier Polynome über einem Körper. Hierzu müssen einige grundlegende Begriffe eingeführt werden.

Definition 2.3 *Ein* Körper *ist eine Menge* $\mathbb{K}$ *versehen mit zwei binären Verknüpfungen (Addition bzw. Multiplikation)*

$$+ : \mathbb{K} \times \mathbb{K} \longrightarrow \mathbb{K}, \qquad \cdot : \mathbb{K} \times \mathbb{K} \longrightarrow \mathbb{K},$$

die den folgenden Bedingungen bzw. Körperaxiomen *genügen:*

1. $(\mathbb{K}, +)$ *ist eine abelsche (kommutative) Gruppe, d. h. die folgenden additiven Eigenschaften sind erfüllt:*
   (a) $a + (b + c) = (a + b) + c$ *für alle* $a, b, c \in \mathbb{K}$ *(Assoziativgesetz).*
   (b) $a + b = b + a$ *für alle* $a, b \in \mathbb{K}$ *(Kommutativgesetz).*
   (c) *Es gibt ein Element* $0 \in \mathbb{K}$ *mit* $0 + a = a = a + 0$ *für alle* $a \in \mathbb{K}$ *(Existenz des neutralen Elements).*
   (d) *Zu jedem* $a \in \mathbb{K}$ *existiert das* additive Inverse $-a \in \mathbb{K}$ *mit* $(-a) + a = 0$.
2. *Multiplikative Eigenschaften:*
   (a) $a \cdot (b \cdot c) = (a \cdot b) \cdot c$ *für alle* $a, b, c \in \mathbb{K}$ *(Assoziativgesetz).*
   (b) $a \cdot b = b \cdot a$ *für alle* $a, b \in \mathbb{K}$ *(Kommutativgesetz).*
   (c) *Es gibt ein Element* $1 \in \mathbb{K} \setminus \{0\}$ *mit* $1 \cdot a = a = a \cdot 1$ *für alle* $a \in \mathbb{K}$ *(Existenz des neutralen Elements).*
   (d) *Zu jedem* $a \in \mathbb{K} \setminus \{0\}$ *existiert das* multiplikative Inverse $a^{-1} \in \mathbb{K}$ *mit* $a^{-1} \cdot a = 1 = a \cdot a^{-1}$.

   Mit $\mathbb{K}^* := \mathbb{K} \setminus \{0\}$ *ist also* $(\mathbb{K}^*, \cdot)$ *eine abelche (kommutative) Gruppe.*
3. *Zusammenspiel zwischen additiver und multiplikativer Verknüpfung:*
   (a) $a \cdot (b + c) = a \cdot b + a \cdot c$ *für alle* $a, b, c \in \mathbb{K}$ *(Links-Distributivgesetz).*
   (b) $(a + b) \cdot c = a \cdot c + b \cdot c$ *für alle* $a, b, c \in \mathbb{K}$ *(Rechts-Distributivgesetz).*

Mit $\mathbb{K}[x]$ bezeichnen wir die Menge der Polynome mit Koeffizienten aus $\mathbb{K}$ und der Unbestimmten x. Es sei also

$$\mathbb{K}[x] := \Big\{ \sum_{i=0}^{d} c_i \cdot x^i : c_0, \ldots, c_d \in \mathbb{K},\, d = 0, 1, \ldots \Big\}.$$

Ist $p = \sum_{i=0}^{d} c_i \cdot x^i \in \mathbb{K}[x]$ und $c_d \neq 0$, so heißt $\deg(p) := d$ der *Grad* von p. Vereinbarungsgemäß ist der Grad des Nullpolynoms gleich -1. Natürlich ist der Körper $\mathbb{K}$ in $\mathbb{K}[x]$ enthalten, da zu $\mathbb{K}[x]$ auch die konstanten Polynome gehören. In $\mathbb{K}[x]$ ist in offensichtlicher Weise eine Addition $+$ und eine Multiplikation $\cdot$ erklärt.

Zunächst formulieren und beweisen wir den sogenannten *Divisionssatz*, dessen entsprechende Aussage für nichtnegative ganze Zahlen offensichtlich ist.

Satz 2.4 (Divisionssatz) *Sei $\mathbb{K}$ ein Körper und $f, g \in \mathbb{K}[x]$ mit $g \neq 0$. Dann gibt es $q, r \in \mathbb{K}[x]$ mit $\deg(r) < \deg(g)$ mit $f = q \cdot g + r$. Erfüllen $q_1, r_1 \in \mathbb{K}[x]$ dieselben Bedingungen, ist also $\deg(r_1) < \deg(g)$ und $f = q_1 \cdot g + r_1$, so ist $q = q_1$ und $r = r_1$.*

Beweis: Wir wenden vollständige Induktion nach dem Grad von f an, wobei $g \in \mathbb{K}[x]$ mit $\deg(g) \geq 0$ (da $g \neq 0$ vorausgesetzt wurde) fixiert ist. Ist $\deg(f) < \deg(g)$, so ist die Existenzaussage mit $(q, r) = (0, f)$ richtig. Der Induktionsanfang ist damit gelegt. Wir können also annehmen, dass $\deg(f) \geq \deg(g)$. Sei

$$f = \sum_{i=0}^{m} a_i \cdot x^i, \qquad g = \sum_{i=0}^{n} b_i \cdot x^i,$$

wobei $m := \deg(f) \geq \deg(g) =: n$. Nun definiere man $f' \in \mathbb{K}[x]$ durch

$$f' := f - \frac{a_m}{b_n} \cdot x^{m-n} \cdot g,$$

wobei a_m/b_n natürlich für $a_m \cdot b_n^{-1}$ steht. Dann ist $\deg(f') < \deg(f)$. Die Induktionsannahme liefert ein Paar $(q_0, r) \in \mathbb{K}[x] \times \mathbb{K}[x]$ mit $\deg(r) < \deg(g)$ und $f' = q_0 \cdot g + r$. Dann ist

$$f = (\underbrace{q_0 + (a_m/b_n) \cdot x^{m-n}}_{=:q}) \cdot g + r,$$

womit die *Existenz* eines Paares (q, r) mit den gewünschten Eigenschaften bewiesen ist. Zum Beweis der *Eindeutigkeit* nehmen wir an, (q, r) und (q', r') seien zwei Paare aus $\mathbb{K}[x] \times \mathbb{K}[x]$ mit

$$f = q \cdot g + r = q' \cdot g + r'$$

und $\deg(r) < \deg(g)$ und $\deg(r') < \deg(g)$. Dann ist

$$0 = (q - q') \cdot g + (r - r')$$

und daher $(q' - r') \cdot g = r - r'$. Die linke Seite ist das Nullpolynom oder ein Polynom mit einem Grad, der mindestens so groß wie $\deg(g)$ ist, während auf der rechten Seite $\deg(r - r') < \deg(g)$ gilt. Daher sind beide Seiten gleich 0, folglich $r = r'$, $q = q'$. $\square$

Seien $f, g \in \mathbb{K}[x]$ gegeben. Wir sagen, dass g das Polynom f *teilt* (oder g ein *Teiler* von f ist, ausgedrückt durch $g \mid f$), wenn $f = q \cdot g$ mit $q \in \mathbb{K}[x]$. Ein Polynom $p \in \mathbb{K}[x]$ heißt ein *größter gemeinsamer Teiler* (ggT) von f und g, wenn p sowohl f als auch g teilt, und für jedes $q \in \mathbb{K}[x]$, das ebenfalls f und g teilt, gilt $\deg(q) \leq \deg(p)$. Man beachte, dass der ggT zweier Polynome *nicht* eindeutig bestimmt ist, da ein nichttriviales Vielfaches eines ggT ebenfalls ein ggT ist.

Durch wiederholte Anwendung des Divisionssatzes kann man einen größten gemeinsamen Teiler zu zwei Polynomen $f, g \in \mathbb{K}[x]$ bestimmen. Dies ist der *euklidische Algorithmus*: Am Anfang setzt man $r_0 := f$, $r_1 := g$ und berechnet sukzessive auf Grund des Divisionssatzes $r_2, \ldots, r_{n-1}, r_n \in \mathbb{K}[x]$ und $q_1, \ldots, q_n \in \mathbb{K}[x]$ mit

$$\begin{array}{rcll} r_0 & = & q_1 \cdot r_1 + r_2, & \deg(r_2) < \deg(r_1), \\ r_1 & = & q_2 \cdot r_2 + r_3, & \deg(r_3) < \deg(r_2), \\ & \vdots & & \vdots \\ r_{n-2} & = & q_{n-1} \cdot r_{n-1} + r_n, & \deg(r_n) < \deg(r_{n-1}), \\ r_{n-1} & = & q_n \cdot r_n + 0. & \end{array}$$

Dann ist r_n ein ggT von f und g. Genauer gilt der Reihe nach:

(a) r_n teilt f und g.

(b) Es existieren $s, t \in \mathbb{K}[x]$ mit $s \cdot f + t \cdot g = r_n$.

(c) r_n ist ein ggT von f und g.

(d) Ein ggT von f und g ist bis auf Multiplikation mit einem Element aus $\mathbb{K}^* := \mathbb{K} \setminus \{0\}$ eindeutig bestimmt.

(e) Ist $d \in \mathbb{K}[x]$ ein ggT von f und g, so existieren $s, t \in \mathbb{K}[x]$ mit $d = s \cdot f + t \cdot g$.

Denn: Die Aussagen (a)–(c) können genau wie beim Beweis von Satz 2.1 in Unterabschnitt 2 bewiesen werden. Ist d ein weiterer ggT von f und g, so ist d ein Teiler von r_n und es gilt $\deg(d) = \deg(r_n)$. Hieraus folgt auch (d). Die Aussage (e) folgt unmittelbar aus (c) und (d). Dies entspricht der Aussage des Lemmas von Bézout für nichtnegative ganze Zahlen.

3 Der Satz von Wedderburn

Der (kleine) Satz von Wedderburn sagt aus, dass ein endlicher Schiefkörper sogar ein Körper ist, also die Multiplikation notwendigerweise kommutativ ist. Hierbei ist ein *Schiefkörper* eine Menge $\mathbb{K}$, die mit zwei binären Verknüpfungen $+$ und $\cdot$ versehen ist, für die alle Körperaxiome erfüllt sind mit eventueller Ausnahme der multiplikativen Kommutativität. In einem Schiefkörper (und erst recht in einem Körper) sind die (additiv bzw. multiplikativ) neutralen Elemente sowie die (additiv bzw. multiplikativ) inversen Elemente offenbar eindeutig bestimmt.

Der folgende überraschende und schöne Satz stammt von J. H. M. WEDDERBURN (1905). Wir geben den brillanten Beweis von E. WITT[2](1931) an, siehe auch A. WEIL (1974, S. 2) und M. AIGNER, G. M. ZIEGLER (2002). Die Arbeit von Witt kann man hier nachlesen.

Satz 3.1 (Wedderburn) *Ein endlicher, d. h. aus endlich vielen Elementen bestehender, Schiefkörper $\mathbb{K}$ ist kommutativ, also ein Körper.*

Beweis: Für ein beliebiges $s \in \mathbb{K}$ sei

$$C_s := \{x \in \mathbb{K} : x \cdot s = s \cdot x\}$$

der *Zentralisator von s* bzw. die Menge aller Elemente aus $\mathbb{K}$, die mit s kommutieren. Mit $Z(\mathbb{K})$ bezeichnen wir die Menge aller Elemente von $\mathbb{K}$, welche mit allen Elementen von $\mathbb{K}$ vertauschbar sind und nennen diese Menge das *Zentrum von K*. Dann ist also

$$Z(\mathbb{K}) = \bigcap_{s \in \mathbb{K}} C_s.$$

Offenbar ist $Z(\mathbb{K})$ ein *endlicher Körper*. Denn $Z(\mathbb{K})$ ist in dem endlichen Schiefkörper $\mathbb{K}$ enthalten, alle Elemente von $Z(\mathbb{K})$ sind miteinander vertauschbar, 0 und 1 liegen in $Z(\mathbb{K})$ und schließlich folgt aus $x \in Z(\mathbb{K})$ bzw. $x \in Z(\mathbb{K}) \setminus \{0\}$, dass auch $-x \in Z(\mathbb{K})$ bzw. $x^{-1} \in Z(\mathbb{K})$. Mit $q := |Z(\mathbb{K})|$ bezeichnen wir die Anzahl[3] der Elemente von $Z(\mathbb{K})$.

Nun fassen wir $\mathbb{K}$ bzw. für $s \in \mathbb{K}$ den Teilschiefkörper $C_s \subset \mathbb{K}$ als Vektorraum über dem Körper $Z(\mathbb{K})$ auf[4], ihre Dimension sei n bzw. n_s. Dann ist $|\mathbb{K}| = q^n$ und entsprechend $|C_s| = q^{n_s}$, denn z. B. lässt sich jedes Element von $\mathbb{K}$ in eindeutiger Weise als Linearkombination der n Basiselemente mit den q Koeffizienten aus $Z(\mathbb{K})$ darstellen. Unser Ziel wird darin bestehen nachzuweisen, dass $n = 1$, daher der Körper $Z(\mathbb{K}) \subset \mathbb{K}$ und der Schiefkörper $\mathbb{K}$ dieselbe Anzahl q von Elementen besitzen und folglich übereinstimmen. Dies geschieht dadurch, dass wir ab jetzt $n \geq 2$ annehmen und diese Annahme zum Widerspruch führen.

Wir definieren $\mathbb{K}^* := \mathbb{K} \setminus \{0\}$ und auf $\mathbb{K}^*$ die Relation $\sim$ dadurch, dass $u \sim v$ für $u, v \in \mathbb{K}^*$ genau dann gilt, wenn ein $x \in \mathbb{K}^*$ mit $v = x \cdot u \cdot x^{-1}$ existiert. Offenbar ist $\sim$

[2]Über Ernst Witt hörte ich vor vielen Jahren die folgende Geschichte, die, wenn sie nicht wahr wenigstens gut erfunden ist. Nach einem Vortrag in Hannover wurde Witt darauf hingewiesen, dass er einen braunen und einen schwarzen Schuh anhatte. Nach kurzer Überlegung war die Reaktion: "Das Merkwürdige ist, dass ich zu Hause noch so ein Paar habe."

[3]Diese Anzahl ist notwendigerweise eine Primzahlpotenz, siehe Unterabschnitt 4.1. Dies ist aber für die Argumentation hier nicht wichtig.

[4]Ist $x \in C_s$ für ein $s \in \mathbb{K}$ und $\alpha \in Z(\mathbb{K})$, so ist

$$(\alpha \cdot x) \cdot s = \alpha \cdot (x \cdot s) = (x \cdot s) \cdot \alpha = (s \cdot x) \cdot \alpha = s \cdot (x \cdot \alpha) = s \cdot (\alpha \cdot x),$$

also auch $\alpha \cdot x \in C_s$. Sind andererseits $x, y \in C_s$ für ein $s \in \mathbb{K}$, so ist

$$(x + y) \cdot s = x \cdot s + y \cdot s = s \cdot x + s \cdot y = s \cdot (x + y),$$

also auch $x + y \in C_s$.

eine Äquivalenzrelation[5]. Mit

$$\Gamma_s := \{x \cdot s \cdot x^{-1} : x \in \mathbb{K}^*\}$$

bezeichnen wir die Menge aller Elemente aus $\mathbb{K}^*$, die äquivalent zu $s \in \mathbb{K}^*$ sind und nennen diese Menge die zu s gehörende *Äquivalenzklasse*. Die Menge $\mathbb{K}^*$ ist disjunkte Vereinigung ihrer Äquivalenzklassen. Man beachte, dass $|\Gamma_s| = 1$ für ein $s \in \mathbb{K}^*$ genau dann, wenn $s \in Z(\mathbb{K}) \setminus \{0\}$. Bei vorgegebenem $s \in \mathbb{K}^*$ definieren wir $C_s^* := C_s \setminus \{0\}$ und setzen $k := |\Gamma_s|$, $l := |C_s^*| = q^{n_s} - 1$. Sei

$$\Gamma_s = \{x_1 \cdot s \cdot x_1^{-1}, \ldots, x_k \cdot s \cdot x_k^{-1}\}, \qquad C_s^* = \{y_1, \ldots, y_l\}.$$

Zu einem beliebigen $x \in \mathbb{K}^*$ existiert ein $i \in \{1, \ldots, k\}$ mit $x \cdot s \cdot x^{-1} = x_i \cdot s \cdot x_i^{-1}$. Hieraus erhalten wir $(x_i^{-1} \cdot x) \cdot s = s \cdot (x_i^{-1} \cdot x)$ bzw. $x_i^{-1} \cdot x \in C_s^*$. Daher existiert ein $j \in \{1, \ldots, l\}$ mit $x_i^{-1} \cdot x = y_j$ bzw. $x = x_i \cdot y_j$ und folglich ist

$$\mathbb{K}^* = \{x_i \cdot y_j : i = 1, \ldots, k,\ j = 1, \ldots, l\}.$$

Wir wollen uns überlegen, dass $\mathbb{K}^*$ genau kl Elemente besitzt bzw. die Elemente $x_i \cdot y_j$ mit $i = 1, \ldots, k$, $j = 1, \ldots, l$, paarweise verschieden sind. Angenommen, es ist $x_i \cdot y_j = x_{i'} \cdot y_{j'}$ für Paare $(i, j), (i', j') \in \{1, \ldots, k\} \times \{1, \ldots, l\}$. Dann ist $x_{i'} \cdot x_i^{-1} = y_{j'} \cdot y_j^{-1} \in C_s^*$. Folglich ist $x_i \cdot s \cdot x_i^{-1} = x_{i'} \cdot s \cdot x_{i'}^{-1}$, daher $i = i'$ und danach $j = j'$. Daher ist bewiesen, dass $|\mathbb{K}^*| = kl = |\Gamma_s|\,|C_s^*|$ und insbesondere

$$\frac{|\mathbb{K}^*|}{|C_s^*|} = \frac{q^n - 1}{q^{n_s} - 1} = |\Gamma_s|$$

für alle $s \in \mathbb{K}^*$.

Oben erinnerten wir daran, dass $\mathbb{K}^*$ die disjunkte Vereinigung von Äquivalenzklassen ist. Genau $q - 1$ dieser Äquivalenzklassen sind einpunktig und bestehen aus den von 0 verschiedenen Elementen des Zentrums. Sind *alle* Äquivalenzklassen einpunktig, so stimmen $\mathbb{K}^*$ und $Z(\mathbb{K}) \setminus \{0\}$ bzw. $\mathbb{K}$ und $Z(\mathbb{K})$ überein und die Aussage des Satzes ist richtig. Daher nehmen wir im folgenden an, dass es Äquivalenzklassen mit mehr als einem Element gibt. Diese Äquivalenzklassen seien mit $\Gamma_{s_1}, \ldots, \Gamma_{s_p}$ bezeichnet. Da $\mathbb{K}^*$ die disjunkte Zerlegung

$$\mathbb{K}^* = (Z(\mathbb{K}) \setminus \{0\}) \cup \bigcup_{i=1}^{p} \Gamma_{s_i}$$

besitzt, ist

$$(*) \qquad q^n - 1 = |\mathbb{K}^*| = |Z(\mathbb{K}) \setminus \{0\}| + \sum_{i=1}^{p} |\Gamma_{s_i}| = q - 1 + \sum_{i=1}^{p} \frac{q^n - 1}{q^{n_{s_i}} - 1}.$$

Zur Abkürzung setzen wir $n_i := n_{s_i}$ und notieren, dass

$$2 \le |\Gamma_{s_i}| = \frac{q^n - 1}{q^{n_i} - 1} \in \mathbb{N}, \qquad i = 1, \ldots, p.$$

[5] D. h. es gilt $u \sim u$ (Reflexivität), aus $u \sim v$ folgt $v \sim u$ (Symmetrie), aus $u \sim v$ und $v \sim w$ folgt $u \sim w$ (Transitivität) für alle $u, v, w \in \mathbb{K}^*$.

Hieraus wollen wir schließen, dass $n_i \mid n$, $i = 1, \ldots, p$ bzw. n durch n_i geteilt wird. Sei $i \in \{1, \ldots, p\}$ fest. Wir wissen, dass $(q^{n_i} - 1) \mid (q^n - 1)$ und bezeichnen mit r den Rest bei der Division von n durch n_i. D. h. es sei $n = an_i + r$ mit $a \in \mathbb{N}$ und $0 \leq r < n_i$. Wegen

$$(q^n - 1) - (q^{n_i} - 1) = (q^{an_i+r} - 1) - (q^{n_i} - 1) = q^{n_i}(q^{(a-1)n_i+r} - 1)$$

folgt aus $(q^{n_i} - 1) \mid (q^n - 1)$, dass $(q^{n_i} - 1) \mid (q^{(a-1)n_i+r} - 1)$, wobei wir noch benutzt haben, dass q^{n_i} und $(q^{n_i} - 1)$ relativ prim[6] sind. In dieser Weise kann man fortfahren und erhält, dass $(q^{n_i} - 1) \mid (q^r - 1)$ mit $0 \leq r < n_i$. Hieraus folgt $r = 0$ bzw. $n_i \mid n$, $i = 1, \ldots, p$.

Ist $n \in \mathbb{N}$, so heißt $\zeta \in \mathbb{C}$ eine *n-te Einheitswurzel*, wenn $\zeta^n = 1$ bzw. ζ ein Element von

$$C_n := \{\zeta \in \mathbb{C} : \zeta^n = 1\}$$

ist. Die n-ten Einheitswurzeln sind Nullstellen von $x^n - 1$, daher gibt es höchstens n verschiedene n-te Einheitswurzeln. Einheitswurzeln haben den Betrag 1, weiter ist das Produkt von n-ten Einheitswurzeln wieder eine n-te Einheitswurzel. Eine spezielle n-te Einheitswurzel ist

$$\zeta_n := e^{2\pi i/n}.$$

Da ζ_n^k, $k = 0, \ldots, n-1$, paarweise verschiedene n-te Einheitswurzeln sind, ist

$$C_n = \{1, \zeta_n, \zeta_n^2, \ldots, \zeta_n^{n-1}\}.$$

In Abbildung 1 veranschaulichen wir die n-ten Einheitswurzeln für $n = 1, \ldots, 10$. Das n-te *Kreisteilungspolynom* (engl.: cyclotomic polynomial) ist definiert durch

$$\Phi_n(x) := \prod_{\substack{k=1 \\ \mathrm{ggT}(k,n)=1}}^{n} (x - \zeta_n^k), \qquad n \in \mathbb{N}.$$

Offenbar ist $\Phi_n(\cdot)$ ein Polynom mit dem höchsten Koeffizienten 1 vom Grade $\phi(n)$, der Anzahl der zu n teilerfremden Zahlen kleiner oder gleich n bzw. der Anzahl primitiver n-ter Einheitswurzeln[7]. In Abbildung 1 haben wir für $n = 1, \ldots, 10$ die primitiven n-ten Einheitswurzeln durch • gekennzeichnet, die übrigen n-ten Einheitswurzeln durch •. Nun zeigen wir:

[6] Zwei natürliche Zahlen k und n heißen *relativ prim* oder *teilerfremd*, wenn es keine natürliche Zahl außer der Eins gibt, die beide Zahlen teilt bzw. der größte gemeinsame Teiler $\mathrm{ggT}(k, n)$ von k und n gleich 1 ist. Insbesondere sind zwei natürliche Zahlen, deren Differenz 1 ist, relativ prim.

[7] Eine n-te Einheitswurzel $\zeta \in C_n$ heißt *primitiv*, wenn alle n-ten Einheitswurzeln Potenzen von ζ sind. Dann gilt:

- Die n-te Einheitswurzel ζ_n ist primitiv.
- Für $k \in \mathbb{N}$ ist die n-te Einheitswurzel ζ_n^k genau dann primitiv, wenn $\mathrm{ggT}(k, n) = 1$, also k und n relativ prim sind.

Denn: Die erste Aussage ist trivialerweise richtig, sodass nur die zweite Aussage zu zeigen bleibt. Sind k und n relativ prim bzw. $\mathrm{ggT}(k, n) = 1$, so existieren wegen des Lemmas von Bézout (siehe Abschnitt 2) $s, t \in \mathbb{Z}$ mit $sk + tn = 1$. Dann ist

$$\zeta_n^{sk} = \zeta_n^{1-tn} = \zeta_n.$$

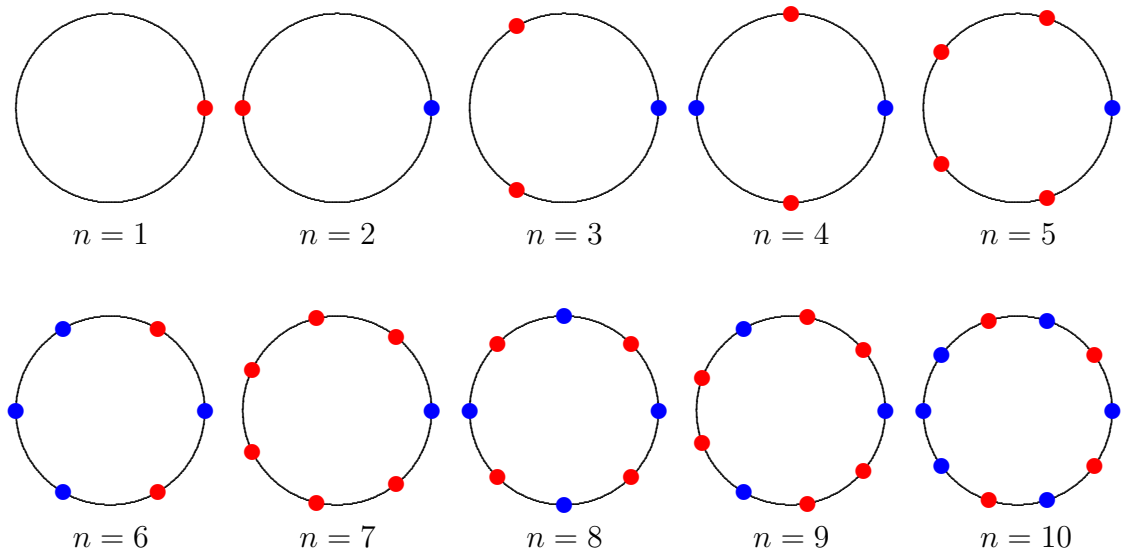

Abbildung 1: Die n-ten Einheitswurzeln für $n = 1, \ldots, 10$

- Für alle $n \in \mathbb{N}$ ist

  $$(**) \qquad x^n - 1 = \prod_{\substack{d=1 \\ d|n}}^{n} \Phi_d(x).$$

  Durch Vergleich der Grade der Polynome auf beiden Seiten von $(**)$ erhält man als einfache Folgerung die Beziehung

  $$n = \sum_{\substack{d=1 \\ d|n}}^{n} \phi(d).$$

Denn:

$$\begin{aligned} x^n - 1 &= \prod_{k=1}^{n} (x - \zeta_n^k) \\ &= \prod_{\substack{d=1 \\ d|n}}^{n} \prod_{\substack{k=1 \\ \text{ggT(k,n)=d}}}^{n} (x - \zeta_n^k) \end{aligned}$$

Da ζ_n eine primitive Einheitswurzel ist, gilt dies auch für ζ_n^k. Denn ist η eine beliebige Einheitswurzel, so gibt es ein $l \in \mathbb{N}$ mit $\eta = \zeta_n^l = (\zeta_n^k)^{sl}$. Umgekehrt nehmen wir jetzt an, ζ_n^k sei eine primitive n-te Einheitswurzel. Es gibt ein $s \in \mathbb{N}$ derart, dass $(\zeta_n^k)^s = \zeta_n^{sk} = \zeta_n$. Hieraus folgt $\zeta_n^{1-sk} = 1$ und hieraus die Existenz eines $t \in \mathbb{Z}$ mit $sk + tn = 1$. Hieraus liest man ab, dass k und n teilerfremd bzw. relativ prim sind. Denn ist $d \in \mathbb{N}$ ein gemeinsamer Teiler von k und n, so teilt d auch $sk + tn = 1$, woraus $d = 1$ folgt.

$$
\begin{aligned}
&= \prod_{\substack{d=1 \\ d|n}}^{n} \prod_{\substack{k=1 \\ \mathrm{ggT(k,n/d)=1}}}^{n/d} (x - \zeta_n^{kd}) \\
&= \prod_{\substack{d=1 \\ d|n}}^{n} \prod_{\substack{k=1 \\ \mathrm{ggT(k,n/d)=1}}}^{n/d} (x - \zeta_{n/d}^{k}) \\
&= \prod_{\substack{d=1 \\ d|n}}^{n} \Phi_{n/d}(x) \\
&= \prod_{\substack{d=1 \\ d|n}}^{n} \Phi_d(x).
\end{aligned}
$$

Als Nächstes zeigen wir:

- Für jedes $n \in \mathbb{N}$ ist $\Phi_n(\cdot)$ ein Polynom mit ganzzahligen Koeffizienten. Ferner ist der konstante Koeffizient von $\Phi_n(\cdot)$ entweder 1 oder -1. Insbesondere folgt aus $(**)$ wegen $n \mid n$, dass $\Phi_n(x) \mid (x^n - 1)$ bzw. ein Polynom $p(\cdot)$ mit ganzzahligen Koeffizienten und $x^n - 1 = p(x)\Phi_n(x)$ existiert.

Denn: Wir beweisen die Behauptung durch vollständige Induktion nach n. Wegen $\Phi_1(x) = x - 1$ ist die Behauptung für $n = 1$ richtig. Nun sei $n \geq 2$ und es wird angenommen, $\Phi_m(\cdot)$ sei für $m < n$ ein Polynom mit ganzzahligen Koeffizienten und einem konstanten Koeffizienten, der entweder 1 oder -1 ist. Aus $(**)$ und der Induktionsannahme erhalten wir, dass

$$x^n - 1 = p(x)\Phi_n(x)$$

mit einem Polynom $p(\cdot)$ mit ganzzahligen Koeffizienten, dessen konstanter Koeffizient entweder 1 oder -1 ist. Außerdem wissen wir, dass der höchste Koeffizient von $p(\cdot)$ (und von $\Phi_n(\cdot)$) gleich 1 ist. Mit einem gewissen $l \in \mathbb{N}$ ist also

$$p(x) = \sum_{j=0}^{l} p_j x^j, \qquad \Phi_n(x) = \sum_{k=0}^{n-l} a_k x^k$$

mit $p_0, \ldots, p_l \in \mathbb{Z}$ mit $p_0 \in \{1, -1\}$ und $p_l = 1$. Durch Vergleich der konstanten Koeffizienten erhalten wir $-1 = p_0 a_0$ und damit $a_0 \in \{1, -1\}$. Angenommen, $a_0, a_1, \ldots, a_{k-1} \in \mathbb{Z}$. Ein Vergleich der Koeffizienten von x^k, $1 \leq k < n$, in $x^n - 1 = p(x)\Phi_n(x)$ liefert

$$0 = \sum_{j=0}^{k} p_l a_{k-j} = \underbrace{\sum_{j=1}^{k} p_j a_{k-j}}_{\in \mathbb{Z}} + p_0 a_k.$$

Wegen $p_0 \in \{1, -1\}$ folgt $a_k \in \mathbb{Z}$. Insgesamt ist gezeigt, dass $\Phi_n(\cdot)$ ganzzahlige Koeffizienten besitzt.

- Für $1 \leq d < n$ und $d \mid n$ gilt

$$(***) \qquad \Phi_n(x) \mid \frac{x^n - 1}{x^d - 1}.$$

Denn: Wegen $(**)$ ist

$$\begin{aligned}
x^n - 1 &= \prod_{\substack{m=1 \\ m|n}}^{n} \Phi_m(x) \\
&= \Phi_n(x) \prod_{\substack{m=1 \\ m|n}}^{n-1} \Phi_m(x) \\
&= \Phi_n(x) \prod_{\substack{m=1 \\ m|d}}^{d} \Phi_m(x) \prod_{\substack{m=1 \\ m \nmid d,\, m|n}}^{n-1} \Phi_m(x) \\
&= \Phi_n(x)(x^d - 1) \underbrace{\prod_{\substack{m=1 \\ m \nmid d,\, m|n}}^{n-1} \Phi_m(x)}_{=:f(x)} \\
&= \Phi_n(x)(x^d - 1) f(x).
\end{aligned}$$

Hierbei ist $f(\cdot)$ ein Polynom mit ganzzahligen Koeffizienten, da wir gezeigt haben, dass dies für jedes $\Phi_m(\cdot)$ der Fall ist.
Nun kommt das furiose Finale des Beweises. Aus $(**)$ und $(***)$ sowie der sogenannten *Klassenformel* $(*)$ folgt $\Phi_n(q) \mid (q-1)$. Ist andererseits $\zeta = a + ib \neq 1$ eine n-te Einheitswurzel (eine solche existiert, da wir $n \geq 2$ annehmen) und $q \geq 2$, so ist notwendigerweise $a < 1$ und

$$\begin{aligned}
|q - \zeta|^2 &= |q - (a + ib)|^2 \\
&= (q - a)^2 + b^2 \\
&= q^2 - 2aq + \underbrace{a^2 + b^2}_{=|\zeta|^2 = 1} \\
&= q^2 - 2aq + 1 \\
&= (q-1)^2 + \underbrace{2(1-a)q}_{>0} \\
&> (q-1)^2
\end{aligned}$$

und daher auch $|q - \zeta| > q - 1$ und $|\Phi_n(q)| > q - 1$, was $\Phi_n(q) \mid (q-1)$ widerspricht. Damit ist der Satz bewiesen. □

4 Endliche Körper

In diesem Abschnitt wollen wir den Beweis eines Ergebnisses nachholen, über welches im Abschnitt über Lateinische Quadrate in der Sammlung *Merkwürdige Mathematik*

gesagt wurde, dass Kenntnisse darüber zur Allgemeinbildung gehören. Insbesondere wollen wir hier mit Satz 4.17 die dort (ohne Beweis) gemachte Aussage

- Ist $q = p^m$ eine Primzahlpotenz, so gibt es (bis auf Isomorphie genau) einen Körper $\mathbb{F}_q$ mit q Elementen

vollständig beweisen. Hierbei kommt es uns darauf an, möglichst wenige Hilfsmittel aus der Algebra zu benutzen. Als Literatur seien z. B. L. Childs (1979), H. Kurzweil (2008) genannt. Wir werden *nicht* auf die zahlreichen Anwendungen endlicher Körper eingehen.

4.1 Endliche Körper haben eine Primzahlpotenz als Ordnung

Zunächst beweisen wir eine Umkehrung der gerade angegebenen Aussage.

Satz 4.1 *Sei $\mathbb{K}$ ein endlicher Körper mit q Elementen. Dann ist $q = p^m$ mit einer Primzahl p und einem $m \in \mathbb{N}$.*

Beweis: Für $k \in \mathbb{N}$ schreiben wir $k \cdot 1$ für das Element aus $\mathbb{K}$, das durch k-faches Summieren der 1, dem bezüglich der Multiplikation neutralen Elements, entsteht. Wir definieren die *Charakteristik* $\mathrm{char}(\mathbb{K})$ des Körpers $\mathbb{K}$ (mit den additiven bzw. multiplikativen neutralen Elementen 0 bzw. 1, die nicht mit den entsprechenden ganzen Zahlen verwechselt werden sollten) durch

$$\mathrm{char}(\mathbb{K}) := \min\{k \in \mathbb{N} : \underbrace{1 + \cdots + 1}_{=k\cdot 1} = 0\}.$$

Zunächst zeigen wir:

- Die Charakteristik $p := \mathrm{char}(\mathbb{K})$ eines endlichen Körpers $\mathbb{K}$ ist eine Primzahl.

Denn: Angenommen, p ist keine Primzahl. Dann ist $p = kl$ mit $k, l \in \mathbb{N}$ und $k, l < p$. Wegen $0 = p \cdot 1 = (k \cdot 1) \cdot (l \cdot 1)$ ist $k \cdot 1 = 0$ oder $l \cdot 1 = 0$. Dies ist ein Widerspruch zur Minimalität von p.

- Sei $p := \mathrm{char}(\mathbb{K})$ die Charakteristik des endlichen Körpers $\mathbb{K}$. Dann ist

$$\mathbb{F}_p := \{i \cdot 1 : i = 0, \ldots, p-1\}$$

  ein in $\mathbb{K}$ enthaltener Körper, und zwar der kleinste in $\mathbb{K}$ enthaltene Körper, der sogenannte *Primkörper*.

Denn: Dass $\mathbb{F}_p \subset \mathbb{K}$ ein Teilkörper von $\mathbb{K}$ ist (bzw. $\mathbb{K}/\mathbb{F}_p$ eine *Körpererweiterung*), erkennt man daran, das $0 = 0 \cdot 1$ und $1 = 1 \cdot 1$ in $\mathbb{F}_p$ liegen, ferner mit $a, b \in \mathbb{F}_p$ auch $a + b \in \mathbb{F}_p$, $a \cdot b \in \mathbb{F}_p$ und $-a \in \mathbb{F}_p$ wegen $p \cdot 1 = 0$. Hier wurde noch nicht benutzt, dass p eine Primzahl ist. Zu zeigen bleibt, dass Inverse von Elementen aus $\mathbb{F}_p \setminus \{0\}$ existieren und selbst in $\mathbb{F}_p \setminus \{0\}$ liegen. Sei $i \cdot 1 \in \mathbb{F}_p \setminus \{0\}$, also $i \in \{1, \ldots, p-1\}$. Zu zeigen ist die Existenz eines $j \in \{1, \ldots, p-1\}$ mit $ij \equiv 1 \bmod p$. Da p eine Primzahl

ist und $i \in \{1, \ldots, p-1\}$ ist $\mathrm{ggT}(i,p) = 1$. Das Lemma von Bézout (siehe Schluss von Abschnitt 2) liefert die Existenz von $s, t \in \mathbb{Z}$ mit $si + tp = 1$. Hieraus folgt

$$1 = 1 \cdot 1 = (si + tp) \cdot 1 = (si) \cdot 1 + t \underbrace{p \cdot 1}_{=0} = (si) \cdot 1.$$

Da s kein ganzzahliges Vielfaches von p ist (andernfalls wäre $(si) \cdot 1 = 0$), ist $j := s \bmod p \in \{1, \ldots, p-1\}$. Offenbar ist dann $(i \cdot 1) \cdot (j \cdot 1) = (ij) \cdot 1 = 1$ und daher $j \cdot 1 = (i \cdot 1)^{-1}$. Damit ist nachgewiesen, dass $\mathbb{F}_p$ ein Körper ist. Da $\mathbb{F}_p \subset \mathbb{K}$ in jedem Teilkörper von $\mathbb{K}$ enthalten sein muss, ist $\mathbb{F}_p$ Primkörper von $\mathbb{K}$.

Nun kommen wir zum Schluss des Beweises. $\mathbb{K}$ ist Oberkörper von $\mathbb{F}_p$ und damit ein (endlichdimensionaler) $\mathbb{F}_p$-Vektorraum. Ist $\{a_1, \ldots, a_m\} \subset \mathbb{K}$ eine Basis dieses Vektorraumes[8], so lässt sich jedes Element $a \in \mathbb{K}$ in eindeutiger Weise in der Form

$$a = \sum_{i=1}^{m} \alpha_i \cdot a_i$$

mit $\alpha_i \in \mathbb{F}_p$, $i = 1, \ldots, m$, darstellen. Da $\mathbb{F}_p$ genau p Elemente enthält, liest man hieraus ab, dass $\mathbb{K}$ genau p^m Elemente enthält. Der Satz ist bewiesen. □

Bemerkung 4.2 Bei vorgegebener Primzahl p ist $\mathbb{Z}/p\mathbb{Z}$ definiert als die Menge aller *Restklassen* $[i]$, $i = 0, \ldots, p-1$, derjenigen ganzen Zahlen, die bei einer Division durch p den Rest i ergeben. Also ist

$$\mathbb{Z}/p\mathbb{Z} := \{[i] : i = 0, \ldots, p-1\}$$

mit

$$[i] := \{kp + i : k \in \mathbb{Z}\}, \qquad i = 0, \ldots, p-1.$$

Auf $\mathbb{Z}/p\mathbb{Z}$ können binäre Verknüpfungen

$$+ : \mathbb{Z}/p\mathbb{Z} \times \mathbb{Z}/p\mathbb{Z} \longrightarrow \mathbb{Z}/p\mathbb{Z}, \qquad \cdot : \mathbb{Z}/p\mathbb{Z} \times \mathbb{Z}/p\mathbb{Z} \longrightarrow \mathbb{Z}/p\mathbb{Z}$$

durch

$$[i] + [j] := [(i + j) \bmod p], \qquad [i] \cdot [j] := [(ij) \bmod p]$$

erklärt werden. Z. B. sind in $\mathbb{Z}/7\mathbb{Z}$ die Addition und die Multiplikation durch die folgenden Verknüpfungstabellen, siehe Tabelle 1, gegeben. Hierbei lassen wir die eckigen Klammern bei den Restklassen fort. Mit dem Nullelement [0] (Menge der durch p teilbaren ganzen Zahlen) als additivem neutralen Element und dem Einselement [1] (Menge $1 + p\mathbb{Z}$ der ganzen Zahlen, die bei Division durch p den Rest 1 ergeben) als multiplikativem neutralen Element ist $\mathbb{Z}/p\mathbb{Z}$ ein aus p Elementen bestehender Körper der Charakteristik p, der sogenannte *Restklassenkörper*. Das multiplikative Inverse einer Restklasse $[i] \in \mathbb{Z}/p\mathbb{Z}$ berechnet man mit Hilfe des erweiterten euklidischen Algorithmus. Mit diesem erhält man $s, t \in \mathbb{Z}$ mit $si + tp = 1$ und weiß, dass mit $j := s \bmod p$ die Beziehung $[j] = [i]^{-1}$ gilt. Ferner ist jeder aus p Elementen bestehender Körper

[8] Die Dimension des $\mathbb{F}_p$-Vektorraums $\mathbb{K}$ bezeichnet man auch mit $[\mathbb{K} : \mathbb{F}_p]$ und nennt dies den *Grad der Körpererweiterung* $\mathbb{K}/\mathbb{F}_p$.

+	0	1	2	3	4	5	6
0	0	1	2	3	4	5	6
1	1	2	3	4	5	6	0
2	2	3	4	5	6	0	1
3	3	4	5	6	0	1	2
4	4	5	6	0	1	2	3
5	5	6	0	1	2	3	4
6	6	0	1	2	3	4	5

·	0	1	2	3	4	5	6
0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6
2	0	2	4	6	1	3	5
3	0	3	6	2	5	1	4
4	0	4	1	5	2	6	3
5	0	5	3	1	6	4	2
6	0	6	5	4	3	2	1

Tabelle 1: Verknüpfungstabellen für Addition und Multiplikation in $\mathbb{Z}/7\mathbb{Z}$

$\mathbb{F}_p$ (wobei p eine Primzahl ist) *isomorph* zu dem Restklassenkörper $\mathbb{Z}/p\mathbb{Z}$. Denn die durch $\sigma(i \cdot 1) := [i]$, $i = 0, \ldots, p-1$, definierte Abbildung $\sigma\colon \mathbb{F}_p \longrightarrow \mathbb{Z}/p\mathbb{Z}$ ist bijektiv, ferner ist $\sigma(a+b) = \sigma(a) + \sigma(b)$ und $\sigma(a \cdot b) = \sigma(a) \cdot \sigma(b)$ für alle $a, b \in \mathbb{F}_p$, also σ ein Isomorphismus. Daher schreibt man auch häufig $\mathbb{F}_p$ statt $\mathbb{Z}/p\mathbb{Z}$. □

Bemerkung 4.3 Wir halten noch einmal fest, was wir im obigen Satz bewiesen haben.

1. Die Charakteristik eines endlichen Körpers $\mathbb{K}$ ist eine Primzahl.

2. Ist p die Charakteristik des endlichen Körpers $\mathbb{K}$, so ist der kleinste in $\mathbb{K}$ enthaltene Körper $\mathbb{F}_p$, also der Primkörper von $\mathbb{K}$, isomorph zum Restklassenkörper $\mathbb{Z}/p\mathbb{Z}$.

3. Ist $\mathbb{K}$ ein endlicher Körper der Charakteristik p, so ist die Ordnung von $\mathbb{K}$, also die Anzahl der Elemente von $\mathbb{K}$, gleich $q = p^m$, wobei $m := [\mathbb{K} : \mathbb{F}_p]$ die Dimension des $\mathbb{F}_p$-Vektorraums $\mathbb{K}$ bzw. der Grad der Körpererweiterung $\mathbb{K}//\mathbb{F}_p$ ist.

□

4.2 Die multiplikative Gruppe eines endlichen Körpers ist zyklisch

Eine (multiplikative) abelsche Gruppe G heißt *zyklisch*, wenn sie von einem einzelnen Element $g \in G$ erzeugt wird, also jedes Element von G eine Potenz von g ist bzw.

$$G = \{g^k : k \in \mathbb{Z}\}$$

gilt. Für ein Element $a \in G$ heißt

$$\operatorname{ord}(a) := \begin{cases} \infty, & \text{falls } a^k \neq 1 \text{ für alle } k \in \mathbb{N}, \\ \min\{k \in \mathbb{N} : a^k = 1\}, & \text{sonst} \end{cases}$$

die *Ordnung* des Gruppenelements $a \in G$. Wir beginnen mit einem klassischen Resultat.

Satz 4.4 (Lagrange) *Sei G eine endliche (multiplikative) Gruppe und $U \subset G$ eine Untergruppe. Dann ist die Anzahl $|U|$ der Elemente von U ein Teiler der Anzahl $|G|$ der Elemente von G. Speziell ist für jedes $a \in G$ die Ordnung* $\operatorname{ord}(a)$ *von a ein Teiler von $|G|$.*

Beweis: Auf G definiere man die binäre Relation $\sim$ durch

$$x \sim y :\Longleftrightarrow y \cdot x^{-1} \in U.$$

Aus den Gruppenaxiomen folgt, dass $\sim$ eine Äquivalenzrelation ist. Die Äquivalenzklasse eines Elementes $x \in G$ ist $U \cdot x := \{a \cdot x : a \in U\}$. Offensichtlich ist $|U \cdot x| = |U|$ für alle $x \in G$. Da andererseits G disjunkte Vereinigung von Äquivalenzklassen ist, ist $|G|$ gleich dem Produkt aus $|U|$ und der Anzahl der Äquivalenzklassen, insbesondere ist $|U|$ ein Teiler von $|G|$. Ist $k := \operatorname{ord}(a)$ die Ordnung eines Elementes $a \in G$, so ist $U := \{a, a^2, \ldots, a^k\}$ eine zyklische Untergruppe von G und daher $|U| = \operatorname{ord}(a)$ ein Teiler von $|G|$. □

Unser nächstes Ziel ist es, den folgenden Satz zu beweisen. Für diesen gibt es viele Beweise, siehe z. B. hier oder hier. Einen Beweis findet man auch bei A. Weil (1974, S. 2) und J.-P. Serre (1973, S. 4), dessen Beweis wir im wesentlichen folgen.

Satz 4.5 *Sei $\mathbb{K}$ ein Körper, $\mathbb{K}^* := \mathbb{K} \setminus \{0\}$ und $(\mathbb{K}^*, \cdot)$ die multiplikative Gruppe zu $\mathbb{K}$. Dann ist jede endliche Untergruppe G von $\mathbb{K}^*$ zyklisch. Ist speziell $\mathbb{K}$ ein endlicher Körper mit $q := |\mathbb{K}|$ Elementen, so gilt:*

1. *Die multiplikative Gruppe $(\mathbb{K}^*, \cdot)$ ist zyklisch.*
2. *Es ist $x^q - x = 0$ für alle $x \in \mathbb{K}$.*

Beweis: Zunächst zeigen wir:

- Ein Polynom $p \in \mathbb{K}[x] \setminus \{0\}$ vom Grad d besitzt höchstens d Nullstellen in $\mathbb{K}$.

Denn: Wir beweisen die Aussage durch vollständige Induktion nach dem Grad d. Die Aussage ist für $d = 0, 1$ richtig. Sei also $d \geq 2$ und die Aussage für kleinere Grade bewiesen. Ist $a \in \mathbb{K}$ eine Nullstelle von p, so ist $p = (x - a) \cdot q$ mit $\deg(q) = d - 1$. Denn wegen des Divisionssatzes 2.4(siehe Unterabschnitt 2.2) ist $p = (x - a) \cdot q + r$, wobei r konstant ist, woraus sich durch Einsetzen von a notwendigerweise $r = 0$ ergibt. Nach Induktionsvoraussetzung hat q maximal $q - 1$ Nullstellen. Für $b \in \mathbb{K}$ ist $p(b) = (b - a) \cdot q(b)$. Daher ist b eine Nullstelle von p genau dann, wenn $b = a$ oder b eine Nullstelle von q ist. Daher hat p maximal d Nullstellen.

Nun kommen wir zum eigentlichen Beweis des Satzes. Sei G eine endliche Untergruppe von $\mathbb{K}^*$ und $n := |G|$ die Anzahl der Elemente von G. Für $d = 1, \ldots, n$ bezeichnen wir mit a_d die Anzahl der Elemente aus G mit der Ordnung d. Dann ist $a_d = 0$, wenn d kein Teiler von n ist. Ist dagegen $a_d > 0$, so existiert ein Element $g \in G$ mit der Ordnung d. Die Untergruppe $U := \{g, g^2, \ldots, g^d\}$ ist eine zyklische Gruppe der Ordnung d. Jedes Element von U ist eine Lösung der Gleichung $x^d - 1 = 0$. Da diese Gleichung höchstens d Lösungen besitzen kann, sind die Elemente von U, also $g, g^2, \ldots, g^d$, *alle* Lösungen von $x^d - 1 = 0$. Daher ist *jedes* Element aus G der Ordnung

d in U enthalten. Ein Element $g^k \in U$, $k = 1, \ldots, d$, hat aber genau dann die Ordnung d, wenn k und d teilerfremd sind bzw. $\operatorname{ggT}(k, d) = 1$ gilt (siehe die entsprechende Aussage über primitive Einheitswurzeln). Also ist a_d, die Anzahl der Elemente in G mit der Ordnung d, gleich der Anzahl $\phi(d)$ der Zahlen k zwischen 1 und d die zu d teilerfremd sind. Auf Seite 11 haben wir im Beweis des Satzes von Wedderburn die Beziehung

$$\sum_{\substack{d=1 \\ d|n}}^{n} \phi(d) = n$$

nachgewiesen. Andererseits ist auch

$$\sum_{\substack{d=1 \\ d|n}}^{n} a_d = \sum_{d=1}^{n} a_d = n,$$

da jedes Element eine Ordnung besitzt. Folglich ist

$$\sum_{\substack{d=1 \\ d|n}}^{n} (\underbrace{\phi(d) - a_d}_{\geq 0}) = 0$$

und daher $a_d = \phi(d)$ für alle $d \in \{1, \ldots, n\}$ mit $d \mid n$. Insbesondere ist $a_n = \phi(n) > 0$. Daher enthält die Gruppe G mit n Elementen mindestens ein Element der Ordnung n (genauer sogar $\phi(n)$ Elemente der Ordnung n) und ist folglich zyklisch. Die Behauptung ist bewiesen.

Ist speziell $\mathbb{K}$ ein endlicher Körper, so ist $\mathbb{K}^*$ bezüglich der Multiplikation eine endliche Untergruppe und daher wegen der gerade eben bewiesenen Aussage zyklisch. Die Gleichung $x^q - x = 0$ ist für $x = 0$ trivialerweise richtig. Wir überlegen uns, dass $x^{q-1} = 1$ für alle $x \in \mathbb{K}^*$. Nun ist $q - 1$ die Ordnung der multiplikativen Gruppe $\mathbb{K}^*$, wegen des Satzes von Lagrange ist die Ordnung jedes Elementes aus $\mathbb{K}^*$ ein Teiler von $q - 1$. Zu jedem $x \in \mathbb{K}^*$ gibt es daher ein $k \in \{1, \ldots, q-1\}$ mit $k \operatorname{ord}(x) = q - 1$. Folglich ist

$$x^{q-1} = x^{k \operatorname{ord}(x)} = (x^{\operatorname{ord}(x)})^k = 1^k = 1$$

und damit natürlich auch $x^q - x = 0$. □

Beispiel 4.6 Wir betrachten den Restklassenkörper $\mathbb{K} := \mathbb{Z}/23\mathbb{Z}$. Wegen des vorigen Satzes wissen wir insbesondere, dass die multiplikative Gruppe $\mathbb{K}^* := \mathbb{K} \setminus \{0\}$ eine zyklische Gruppe mit $n := 22$ Elementen ist. Welche Elemente erzeugen diese Gruppe und wie viele gibt es davon? Als Ordnungen der Elemente von $\mathbb{K}^*$ kommen nur die Teiler von 22 in Frage, also 1, 2, 11 und 22. Die Anzahl der Elemente in $\mathbb{K}^*$ der Ordnung 22 ist (siehe den Beweis des letzten Satzes) gleich $\phi(22) = 10$, der Anzahl der zu 22 teilerfremden Zahlen zwischen 1 und 22 (dies sind nämlich $1, 3, 5, 7, 9, 13, 15, 17, 19, 21$). Um ein erzeugendes Element zu finden, experimentieren wir. Die Restklasse [1] hat natürlich die Ordnung 1, kommt also als erzeugendes Element nicht in Frage. Was ist die Ordnung der Restklasse [2]? Es ist $[2]^5 = [2^5] = [9]$ und daher

$$[2]^{11} = [9]^2 \cdot [2] = [12] \cdot [2] = [1],$$

daher ist ord ([2]) = 11. Jetzt untersuchen wir die Restklasse [3]. Es ist $[3]^3 = [4]$ und daher

$$[3]^{11} = [4]^3 \cdot [3]^2 = [18] \cdot [9] = [1],$$

daher ist auch [3] kein die multiplikative Gruppe K^* erzeugendes Element. Der nächste Kandidat [4] ist ebenfalls nicht primitiv, da $[4]^{11} = [2]^{22} = [1]$. Versuchen wir es jetzt mit [5]. Es ist $[5]^2 = [2]$ und daher

$$[5]^{11} = [2]^5 \cdot [5] = [9] \cdot [5] = [22] \neq [1].$$

Daher hat [5] die Ordnung 22 und ist ein erzeugendes Element von $\mathbb{K}^*$. Alle erzeugenden Elemente sind durch $[5]^k$ mit $k \in \{1, 3, 5, 7, 9, 13, 15, 17, 19, 21\}$ gegeben. □

4.3 Kongruenzklassen modulo eines Polynoms

Grundlegend für das Weitere ist der Übergang von einem Körper $\mathbb{K}$ zunächst zu der Menge $\mathbb{K}[x]$ der Polynome mit Koeffizienten aus $\mathbb{K}$ und dann zur Menge $\mathbb{K}[x]/N(x)$, der Menge der *Kongruenzklassen modulo N* bezüglich eines vorgegebenen Polynoms $N \in \mathbb{K}[x]$. Zwei Polynome $f, g \in \mathbb{K}[x]$ heißen *kongruent modulo N*, in Zeichen $f \equiv g \bmod N$, wenn $N \mid (f - g)$ bzw. $f - g = N \cdot p$ mit einem $p \in \mathbb{K}[x]$ ist. Hierdurch ist eine Äquivalenzrelation erklärt. Die zu einem $f \in \mathbb{K}[x]$ gehörende Äquivalenzklasse ist

$$[f]_N := \{g \in \mathbb{K}[x] : f \equiv g \bmod N\} = \{f + N \cdot p : p \in \mathbb{K}[x]\}.$$

Mit $\mathbb{K}[x]/(N(x))$ bezeichnen wir die Menge der Äquivalenz- bzw. Kongruenzklassen von Polynomen aus $\mathbb{K}[x]$ modulo N, d. h. es ist

$$\mathbb{K}[x]/(N(x)) := \{[f]_N : f \in \mathbb{K}[x]\}.$$

In $\mathbb{K}[x]/(N(x))$ können eine Addition + und eine Multiplikation · durch

$$[f]_N + [g]_N := [f + g]_N, \qquad [f]_N \cdot [g]_N := [f \cdot g]_N$$

erklärt werden. Ist speziell $g \in \mathbb{K}[x]$ durch $g(x) := x$ definiert und $f \in K[x]$ gegeben durch

$$f(x) = \sum_{i=0}^{d} c_i \cdot x^i = \sum_{i=0}^{d} c_i \cdot g(x)^i,$$

so ist

$$[f]_N = \Big[\sum_{i=0}^{d} c_i \cdot g^i\Big]_N = \sum_{i=0}^{d} [c_i \cdot g^i]_N = \sum_{i=0}^{d} [c_i]_N \cdot [g]_N^i.$$

Definition 4.7 Ein Polynom $N \in \mathbb{K}[x]$ heißt *irreduzibel* in $\mathbb{K}[x]$, wenn es sich nicht als Produkt von zwei Polynomen aus $\mathbb{K}[x]$ mit einem Grad ≥ 1 darstellen lässt.

Ist $p \in \mathbb{N}$ eine Primzahl, so ist $\mathbb{Z}/p\mathbb{Z}$ ein Körper. Entsprechend ist $\mathbb{K}[x]/(N(x))$ ein Körper, wenn $N \in \mathbb{K}[x]$ irreduzibel ist:

Satz 4.8 *Sei $\mathbb{K}$ ein Körper und $N \in \mathbb{K}[x]$ irreduzibel. Dann ist $\mathbb{K}[x]/(N(x))$ mit der oben erklärten Addition und Multiplikation sowie den additiv bzw. multiplikativ neutralen Elementen $0 = [0]_N$ bzw. $1 = [1]_N$ ein Körper.*

Beweis: Offenbar genügt es zu zeigen, dass jedes vom Nullelement verschiedenes Element aus $\mathbb{K}[x]/(N(x))$ ein (multiplikativ) inverses Element besitzt. Sei also $g \in \mathbb{K}[x]$ mit $[g]_N \neq [0]_N$ gegeben. Insbesondere ist N kein Teiler von g. Ist dann $d \in \mathbb{K}[x]$ ein ggT von N und g, so ist $d \in \mathbb{K}^* := \mathbb{K} \setminus \{0\}$, also d ein von Null verschiedenes konstantes Polynom, ferner existieren Polynome $s, t \in \mathbb{K}[x]$ mit

$$d = s \cdot g + t \cdot N,$$

siehe die Aussage (e) am Schluss von Unterabschnitt 2.2. Nach Division der letzten Gleichung durch d bzw. Multiplikation mit d^{-1} erhalten wir

$$1 = (\underbrace{d^{-1}s \cdot s}_{=:s_1}) \cdot g + (\underbrace{d^{-1} \cdot t}_{=:t_1}) \cdot N = s_1 \cdot g + t_1 \cdot N.$$

Ein Übergang zu Restklassen ergibt

$$[1]_N = [s_1 \cdot g + t_1 \cdot N]_N = [s_1 \cdot g]_N + \underbrace{[t_1 \cdot N]_N}_{=[0]_N} = [s_1]_N \cdot [g]_N.$$

Mit $h := s_1 \in \mathbb{K}_0[x]$ ist $[h]_N = [g]_N^{-1}$ das inverse Element zu $[g]_N \neq 0$. Daher ist $\mathbb{K}[x]/(N(x))$ ein Körper. □

Wenn wir wüssten, dass es zu jedem $m \in \mathbb{N}$ und jeder Primzahl p ein irreduzibles Polynom $N \in \mathbb{F}_p[x]$ vom Grad m gibt, so wäre durch den folgenden Satz die Existenz endlicher Körper der Ordnung $q = p^m$ bewiesen.

Satz 4.9 *Sei p eine Primzahl und $N \in \mathbb{F}_p[x]$ ein irreduzibles Polynom mit Koeffizienten aus $\mathbb{F}_p$ vom Grad $m \in \mathbb{N}$. Dann ist $\mathbb{F}_p[x]/(N(x))$ ein Körper mit $q := p^m$ Elementen.*

Beweis: Dass $\mathbb{F}_p[x]/(N(x))$ ein Körper ist, haben wir gerade eben bewiesen. Wegen $[N]_N = [0]_N$ (daher ist x^m kongruent zu einem Polynom vom Grad $\leq m-1$ mit Koeffizienten aus $\mathbb{F}_p$) sind Elemente von $\mathbb{F}_p[x]/(N(x))$ Kongruenzklassen, die von Polynomen mit Koeffizienten aus $\mathbb{F}_p$ vom Grad $\leq m-1$ erzeugt werden. Genauer existiert eine bijektive Abbildung T zwischen $\mathbb{F}_p[x]/(N(x))$ und

$$P_{m-1}(\mathbb{F}_p) := \{a_0 + a_1 x + \cdots + a_{m-1}x^{m-1} : a_0, a_1, \ldots, a_{m-1} \in \mathbb{F}_p\},$$

definiert durch

$$T([f]_N) := r,$$

wobei $r \in P_{m-1}(\mathbb{F}_p)$ der Rest der Division von f durch N ist. Daher ist r kongruent f modulo N ist, also ist $r \in [f]_N$. Da die m Koeffizienten $a_0, \ldots, a_{m-1}$ jeweils genau p Werte annehmen können, ist $|P_{m-1}(\mathbb{F}_p)| = p^m$ und die Behauptung des Satzes ist bewiesen. □

Der folgende Satz wird gelegentlich nach *Kronecker* benannt.

Satz 4.10 (Kronecker) *Sei $\mathbb{K}$ ein Körper.*

1. *Ist $N \in \mathbb{K}[x]$ irreduzibel, so ist $\mathbb{L} := \mathbb{K}[x]/(N(x))$ ein Körper, der einen zu $\mathbb{K}$ isomorphen Körper enthält und in dem N die Nullstelle $\alpha := [x]_N \in \mathbb{L}$ besitzt.*

2. *Ist $f \in \mathbb{K}[x]$, so gibt es eine Körpererweiterung $\mathbb{L}/\mathbb{K}$, in dem f eine Nullstelle besitzt.*

Beweis: Wie wir in Satz 4.8 bewiesen haben, ist $\mathbb{L}$ ein Körper. Wir definieren $\phi\colon \mathbb{K} \longrightarrow \mathbb{L}$ durch $\phi(a) := [a]_N$. Aus $\phi(a) = \phi(b)$ bzw. $[a]_N = [b]_N$ folgt $a = b$. Folglich ist $\phi(\mathbb{K})$ ein zu $\mathbb{K}$ isomorpher Teilkörper von $\mathbb{L}$. Sei $g \in \mathbb{K}[x]$ gegeben durch $g(x) := x$, ferner sei $\alpha := [g]_N$. Hat $N \in \mathbb{K}[x]$ die Darstellung

$$N(x) = \sum_{i=0}^{d} c_i \cdot x^i,$$

so ist

$$[0]_N = [N]_N = \Big[\sum_{i=0}^{d} c_i \cdot g^i\Big] = \sum_{i=0}^{d} [c_i]_N \cdot \alpha^i = \sum_{i=0}^{d} \phi(c_i) \cdot \alpha^i.$$

Daher hat N als Element von $\mathbb{L}[x]$ die Nullstelle α. Die Aussage des zweiten Teiles des Satzes erhält man, in dem man den ersten Teil des Satzes auf einen irreduziblen Faktor N von f anwendet. □

Bemerkung 4.11 Ist $\mathbb{K}$ ein Körper und $N \in \mathbb{K}[x]$ irreduzibel, so ist $\mathbb{L} := \mathbb{K}[x]/(N(x))$ ebenfalls ein Körper. Hat N wie im Beweis des Satzes von Kronecker den Grad d und ist $f \in \mathbb{K}[x]$ ein Polynom vom Grad r mit Koeffizienten $a_0, \ldots, a_r$, so hat ein typisches Element $[f]_N \in \mathbb{L}$ die Darstellung

$$[f]_N = \Big[\sum_{i=0}^{r} a_i \cdot x^i\Big]_N = \sum_{i=0}^{r} [a_i]_N \cdot \alpha^i = \sum_{i=0}^{r} \phi(a_i) \cdot \alpha^i.$$

Daher können Elemente von $\mathbb{L}$ als Polynome mit Koeffizienten aus $\phi(\mathbb{K})$, ausgewertet an der Nullstelle $\alpha \in \mathbb{L}$ von N, aufgefasst werden. Hierbei ist $\phi\colon \mathbb{K} \longrightarrow \mathbb{L}$ die durch $\phi(a) := [a]_N$ definierte Abbildung im Beweis des letzten Satzes. □

Nun können wir zeigen:

Satz 4.12 *Sei $\mathbb{K}$ ein Körper und $f \in \mathbb{K}[x]$ ein Polynom mit $d := \deg(f) \geq 1$ und höchstem Koeffizienten a_d. Dann gibt es einen $\mathbb{K}$ (bzw. einen hierzu isomorphen Körper) enthaltenden Körper $\mathbb{L}$ und $\alpha_1, \ldots, \alpha_d \in \mathbb{L}$ derart, dass f als Element von $\mathbb{L}[x]$ die Darstellung*

$$f = a_d \cdot \prod_{i=1}^{d} (x - \alpha_i),$$

also f über $\mathbb{L}$ in Linearfaktoren zerfällt. Also existiert zu $f \in \mathbb{K}[x]$ ein Zerfällungskörper, also eine Körpererweiterung bzw. ein Oberkörper von $\mathbb{K}$, über dem f in Linearfaktoren zerfällt und der bezüglich dieser Eigenschaft minimal ist.

Beweis: Da wir f durch f/a_d ersetzen können, können wir o. B. d. A. annehmen, dass der höchste Koeffizient von f gleich 1 ist. Wir beweisen die Behauptung durch vollständige Induktion nach dem Grad d von $f \in \mathbb{K}[x]$. Ist $d = 1$, so ist f linear und die Behauptung mit $\mathbb{L} := \mathbb{K}$ trivialerweise richtig. Sei nun $\deg(f) = d$. Wir nehmen an, die Behauptung sei für Polynome $g \in \mathbb{K}[x]$ mit $\deg(g) = d-1$ richtig. Wir stellen $f \in \mathbb{K}[x]$ als Produkt irreduzibler Polynome aus $\mathbb{K}[x]$ dar: $f = N_1 \cdot \cdots \cdot N_s$. Dass eine solche Faktorisierung möglich ist, zeigt man leicht durch vollständige Induktion nach dem Grad des gegebenen Polynoms. Ist $\deg(N_i) = 1$, $i = 1, \ldots, s$, so ist die Behauptung mit $\mathbb{L} := \mathbb{K}$ richtig. Andernfalls sei nach eventueller Umnummerierung $\deg(N_1) > 1$. Sei $\mathbb{L}' := \mathbb{K}[x]/(N_1(x))$. Dann ist $\mathbb{L}'$ ein $\mathbb{K}$ (bzw. ein isomorphes Bild von $\mathbb{K}$) enthaltender Körper, welcher wegen des vorigen Satzes eine Nullstelle α von N_1 enthält. In $\mathbb{L}'[x]$ ist also $N_1(x) = (x-\alpha)M_1(x)$ mit $M_1 \in \mathbb{L}'[x]$. Daher besitzt f in $\mathbb{L}'[x]$ die Faktorisierung

$$f(x) = (x-\alpha)M_1(x) \cdot N_2(x) \cdot \cdots \cdot N_s(x).$$

Definiert man $g \in \mathbb{L}'$ durch

$$g := M_1 \cdot N_2 \cdot \cdots \cdot N_s,$$

so ist $\deg(g) = d-1$. Wendet man die Induktionsannahme auf den Körper $\mathbb{L}'$ und $g \in \mathbb{L}'[x]$ an, so erhält man die Existenz eine $\mathbb{L}'$ (bzw. eines isomorphen Bildes) enthaltenden Körper $\mathbb{L}$ derart, dass g in $\mathbb{L}[x]$ ein Produkt linearer Faktoren ist. Das gilt dann aber auch für f wegen $f(x) = (x-\alpha) \cdot g(x)$. Der Satz ist damit bewiesen. □

4.4 Das Minimalpolynom

Ist $\mathbb{K}$ ein Körper, so nennen wir einen Körper $\mathbb{L} \supset \mathbb{K}$ einen Oberkörper von $\mathbb{K}$ bzw. $\mathbb{L}/\mathbb{K}$ eine Körpererweiterung.

Definition 4.13 Sei $\mathbb{L}/\mathbb{K}$ eine Körpererweiterung.

1. Ein $\alpha \in \mathbb{L}$ heißt *algebraisch über* $\mathbb{K}$, wenn ein nichttriviales $f \in \mathbb{K}[x]$ mit $f(\alpha) = 0$ existiert.

2. Eine Körpererweiterung $\mathbb{L}/\mathbb{K}$ heißt *algebraisch über* $\mathbb{K}$, wenn jedes Element in $\mathbb{L}$ algebraisch über $\mathbb{K}$ ist.

3. Sei $\alpha \in \mathbb{L}$ algebraisch über $\mathbb{K}$. Das *Minimalpolynom* für α ist unter allen Polynomen aus $\mathbb{K}[x]$ mit höchstem Koeffizienten 1 und α als Nullstelle dasjenige, welches minimalen Grad besitzt[9]. Weiter heißt α *algebraisch vom Grade* n, wenn n der Grad des Minimalpolynoms zu α ist.

[9]Wir müssen uns überlegen, dass das Minimalpolynom wohldefiniert ist, dass es also unter allen Polynomen aus $\mathbb{K}[x]$ mit höchstem Koeffizienten 1 und einer Nullstelle in α genau eines mit minimalem Grad gibt. Da $\alpha \in \mathbb{L}$ algebraisch über $\mathbb{K}$ ist, gibt es ein Polynom aus $\mathbb{K}[x]$ mit dem höchsten Koeffizienten 1, welches α als Nullstelle besitzt. Sind g_1 und g_2 zwei verschiedene Polynome aus $\mathbb{K}[x]$ mit $\alpha \in \mathbb{L}$ als Nullstelle, höchtem Koedffizienten 1 und dem gleichen minimalen Grad r, so ist $g_1 - g_2$ ein von Null verschiedenes Polynom aus $\mathbb{K}[x]$ mit einem Grad $< r$, welches α als Nullstelle besitzt. Normiert man $g_1 - g_2$ so, dass der höchste Koeffizient 1 ist, so hat man ein Polynom aus $\mathbb{K}[x]$ mit einem kleineren Grad als r, welches den höchsten Koeffizienten 1 und α als Nullstelle besitzt. Damit hat man einen Widerspruch erhalten und gezeigt, dass wir zu Recht von *dem* Minimalpolynom zu $\alpha \in \mathbb{L}$ sprechen können.

Die wichtigsten Eigenschaften des Minimalpolynoms sind im folgenden Satz zusammengefasst.

Satz 4.14 *Sei $\mathbb{L}/\mathbb{K}$ eine Körpererweiterung, $\alpha \in \mathbb{L}$ algebraisch über $\mathbb{K}$ und $g \in \mathbb{K}[x]$ das Minimalpolynom für α. Dann gilt:*

1. *g ist in $\mathbb{K}[x]$ irreduzibel.*
2. *Sei $f \in \mathbb{K}[x]$. Dann ist $f(\alpha) = 0$ genau dann, wenn f durch g geteilt wird bzw. ein $p \in \mathbb{K}[x]$ mit $f = g \cdot p$ existiert.*
3. *Mit $\mathbb{K}(\alpha)$ wird der Durchschnitt aller Teilkörper von $\mathbb{L}$ bezeichnet, die sowohl $\mathbb{K}$ als auch α enthalten. Dann gilt:*
   (a) *Die durch $\phi([f]_g) := f(\alpha)$ (wohl)definierte Abbildung*
   $$\phi\colon \mathbb{K}[x]/(g(x)) \longrightarrow S := \{f(\alpha) : f \in \mathbb{K}[x]\}$$
   ist ein Isomorphismus. Ferner ist $S = \mathbb{K}(\alpha)$.
   (b) *Sei $n := \deg(g)$. Dann ist $[\mathbb{K}(\alpha) : \mathbb{K}] = n$ und $\{1, \alpha, \ldots, \alpha^{n-1}\}$ eine Basis des $\mathbb{K}$-Vektorraums $\mathbb{K}(\alpha)$.*
   (c) *Es ist $[\mathbb{L} : \mathbb{K}] = [\mathbb{L} : \mathbb{K}(\alpha)]\,[\mathbb{K}(\alpha) : \mathbb{K}]$.*

Beweis: Zum Nachweis, dass das Minimalpolynom g für α irreduzibel ist, machen wir einen Widerspruchsbeweis und nehmen an, es sei $g = g_1 \cdot g_2$ mit $g_i \in \mathbb{K}[x]$ und $\deg(g_i) \geq 1$, $i = 1, 2$. Wegen $\deg(g) = \deg(g_1) + \deg(g_2)$ haben sowohl g_1 als auch g_2 kleineren Grad als g. Wegen $0 = g(\alpha) = g_1(\alpha) \cdot g_2(\alpha)$ hat o. B. d. A. g_1 die Nullstelle α. Nach Normierung von g_1 auf höchsten Koeffizienten 1 hat man ein Polynom mit kleinerem Grad als das Minimalpolynom gefunden, welches in α eine Nullstelle besitzt. Dies ist ein Widerspruch zur Definition des Minimalpolynoms und die Irreduzibilität des Minimalpolynoms für α ist bewiesen.

Sei $f \in \mathbb{K}[x]$ und $f(\alpha) = 0$. Wegen des Divisionssatzes 2.4 existieren $p, r \in \mathbb{K}[x]$ mit $f = g \cdot p + r$ und $\deg(r) < \deg(g)$. Da $f(\alpha) = 0$ und $g(\alpha) = 0$ ist $r(\alpha) = 0$. Nach Definition des Minimalpolynoms ist dies nur möglich, wenn r das Nullpolynom ist bzw. f durch g geteilt wird. Wird umgekehrt f durch g geteilt, existiert also ein $p \in \mathbb{K}[x]$ mit $f = g \cdot p$, so folgt aus $g(\alpha) = 0$ auch $f(\alpha) = 0$.

Wir erinnern daran, dass

$$\mathbb{K}[x]/(g(x)) := \{[f]_g : f \in \mathbb{K}[x]\},$$

wobei $[f]_g$ für $f \in \mathbb{K}[x]$ durch

$$[f]_g := \{f + g \cdot p : p \in \mathbb{K}[x]\}$$

definiert ist. Jetzt zeigen wir, dass $S := \{f(\alpha) \in \mathbb{L} : f \in \mathbb{K}[x]\}$ isomorph zu $\mathbb{K}[x]/(g(x))$ ist. Hierzu definieren wir $\phi\colon \mathbb{K}[x]/(g(x)) \longrightarrow S$ durch $\phi([f]_g) := f(\alpha)$ und zeigen, dass ϕ ein (wohldefinierter) Isomorphismus zwischen $\mathbb{K}[x]/(g(x))$ und S ist. Die Abbildung

ϕ ist wohldefiniert, da $(f + g \cdot p)(\alpha) = f(\alpha)$ für jedes $p \in \mathbb{K}[x]$ wegen $g(\alpha) = 0$. Weiter ist die Abbildung ϕ auch injektiv. Denn sind $f, h \in \mathbb{K}[x]$ und $\phi([f]_g) = \phi([h]_g)$, so ist $f(\alpha) = h(\alpha)$ bzw. $(f - h)(\alpha) = 0$. Wegen des gerade eben bewiesenen Teil des Satzes existiert ein $p \in \mathbb{K}[x]$ mit $f - h = g \cdot p$. Also ist $h \in [f]_g$ und daher $[f]_g = [h]_g$, womit die Injektivität von ϕ bewiesen ist. Dass ϕ surjektiv ist, ist offensichtlich. Weiter ist $\phi([1]_g) = 1_{\mathbb{K}}$ sowie

$$\phi([f]_g + [h]_g) = \phi([f + h]_g) = (f + h)(\alpha) = f(\alpha) + h(\alpha) = \phi([f]_g) + \phi([h]_g)$$

und

$$\phi([f]_g \cdot [h]_g) = \phi([f \cdot h]_g) = (f \cdot h)(\alpha) = f(\alpha) \cdot h(\alpha) = \phi([f]_g) \cdot \phi([h]_g).$$

Also [10] sind S und $\mathbb{K}[x]/(g(x))$ isomorph. Nach Satz 4.8 ist $\mathbb{K}[x]/(g(x))$ und dann auch S ein Körper. Wegen $\mathbb{K} \subset S \subset \mathbb{K}(\alpha)$ ist $S = \mathbb{K}(\alpha)$ der kleinste $\mathbb{K}$ und α enthaltende Körper. Damit ist bewiesen, dass $\mathbb{K}(\alpha)$ und $\mathbb{K}[x]/(g(x))$ isomorphe Körper sind und $S = \mathbb{K}(\alpha)$ gilt.

Sei $n := \deg(g)$. Gerade eben haben wir bewiesen, dass $\{f(\alpha) : f \in \mathbb{K}[x]\} = \mathbb{K}(\alpha)$. Daher existiert zu jedem $\beta \in \mathbb{K}(\alpha)$ ein $f \in \mathbb{K}[x]$ mit $\beta = f(\alpha)$. Wegen des Divisionssatzes 2.4 existieren $p, r \in \mathbb{K}[x]$ mit $f = g \cdot p + r$ und $\deg(r) < \deg(g) = n$. Wegen $g(\alpha) = 0$ ist

$$\beta = f(\alpha) = r(\alpha)$$

und daher ist jedes Element β aus $\mathbb{K}(\alpha)$ eine Linearkombination von $1, \alpha, \ldots, \alpha^{n-1}$ mit Koeffizienten aus $\mathbb{K}$. Dies bedeutet, dass $\mathbb{K}(\alpha)$ als $\mathbb{K}$-Vektorraum höchstens die Dimension n besitzt. Wir zeigen jetzt, dass $1, \alpha, \ldots, \alpha^{n-1}$ linear unabhängige Elemente des $\mathbb{K}$-Vektorraums $\mathbb{K}(\alpha)$ sind, womit dann $[\mathbb{K}(\alpha) : \mathbb{K}] = n$ nachgewiesen ist. Hierzu nehmen wir an, mit $a_0, \ldots, a_{n-1} \in \mathbb{K}$ sei

$$a_0 + a_1 \cdot \alpha + \cdots + a_{n-1} \cdot \alpha^{n-1} = 0.$$

Dann hat

$$h := a_0 + a_1 \cdot x + \cdots + a_{n-1} \cdot x^{n-1} \in \mathbb{K}[x]$$

eine Nullstelle in α. Wären nicht alle Koeffizienten $a_0, \ldots, a_{n-1}$ gleich Null, so könnte h so normiert werden, dass der höchste Koeffizient gleich 1 ist. Man hätte ein Polynom aus $\mathbb{K}[x]$ mit kleinerem Grad als das Minimalpolynom, welches ebenfalls in α eine Nullstelle besitzt. Dies ist ein Widerspruch zur Definition des Minimalpolynoms und die Behauptung ist bewiesen.

Sei $k := [\mathbb{L} : \mathbb{K}(\alpha)]$, $l := [\mathbb{K}(\alpha) : \mathbb{K}]$ und

$$\{\alpha_1, \ldots, \alpha_k\} \subset \mathbb{L} \qquad \text{bzw.} \qquad \{\beta_1, \ldots, \beta_l\} \subset \mathbb{K}(\alpha)$$

eine Basis des $\mathbb{K}(\alpha)$-Vektorraums $\mathbb{L}$ bzw. des $\mathbb{K}$-Vektorraums $\mathbb{K}(\alpha)$. Wir zeigen, dass

$$\{\alpha_i \cdot \beta_j\}_{\substack{i=1,\ldots,k \\ j=1,\ldots,l}} \subset \mathbb{K}(\alpha)$$

[10] Nur angemerkt sei, dass dies auch mit Hilfe des sogenannten ersten Isomorphie-Satzes hätte bewiesen werden können. Wir haben einen direkten Beweis vorgezogen.

eine Basis des $\mathbb{K}$-Vektorraums $\mathbb{L}$ ist, womit die Behauptung bewiesen sein wird. Hierzu zeigen wir zunächst, dass sich jedes Element aus $\mathbb{L}$ als eine $\mathbb{K}$-Linearkombination der $\alpha_i \cdot \beta_j$ dargestellt werden kann und anschließend, dass die $\alpha_i \cdot \beta_j$, $i = 1, \ldots, k$, $j = 1, \ldots, l$ linear unabhängig sind. Sei also $a \in \mathbb{L}$ beliebig. Da $\{\alpha_1, \ldots, \alpha_k\}$ eine Basis des $\mathbb{K}(\alpha)$-Vektorraums $\mathbb{L}$ ist, kann a eindeutig in der Form

$$a = \sum_{i=1}^{k} \gamma_i \cdot \alpha_i$$

mit $\{\gamma_1, \ldots, \gamma_k\} \subset \mathbb{K}(\alpha)$ dargestellt werden. Jedes $\gamma_i \in \mathbb{K}(\alpha)$, $i = 1, \ldots, k$, lässt sich eindeutig als Linearkombination von $\beta_j \in \mathbb{K}(\alpha)$, $j = 1, \ldots, l$, mit Koeffizienten r_{ij} aus $\mathbb{K}$ darstellen:

$$\gamma_i = \sum_{j=1}^{l} r_{ij} \cdot \beta_j.$$

Daher ist

$$a = \sum_{i=1}^{k} \gamma_i \alpha_i = \sum_{i=1}^{k} \Big(\sum_{j=1}^{l} r_{ij} \cdot \beta_j \Big) \cdot \alpha_i = \sum_{i=1}^{k} \sum_{j=1}^{l} r_{ij} \cdot \alpha_i \cdot \beta_j.$$

Hieraus liest man auch ab (betrachte den Fall $a = 0$), dass die $\alpha_i \cdot \beta_j$, $(i, j) \in \{1, \ldots, k\} \times \{1, \ldots, l\}$ linear unabhängig sind und die behauptete Aussage ist bewiesen. □

4.5 Existenz und Eindeutigkeit eines Körpers mit p^m Elementen

Wir wissen schon, dass jeder endliche Körper eine Primzahlpotenz als Ordnung besitzt (siehe Satz 4.1). Das Ziel in diesem Unterabschnitt besteht darin nachzuweisen, dass es zu jeder Primzahlpotenz $q = p^m$ im wesentlichen (d. h. bis auf Isomorphie) genau einen Körper $\mathbb{F}_q$ der Ordnung q gibt. Dies ist für $m = 1$ bzw. $q = p$ klar, da der Restklassenkörper $\mathbb{F}_p := \mathbb{Z}/p\mathbb{Z}$ im wesentlichen, d. h. bis auf Isomorphie, der einzige Körper mit p Elementen ist.

Beispiel 4.15 Wir wollen einen Körper mit $q = 2^2 = 4$ Elementen konstruieren. Wir suchen einen $\mathbb{F}_2$ enthaltenden Körper $\mathbb{K}$, dessen vier Elemente wegen der Aussage 2. im Satz in Unterabschnitt 4.2 Nullstellen von

$$f(x) := x^4 - x = x \cdot (x - 1) \cdot (x^2 + x + 1)$$

sind. Wir definieren $N(x) := x^2 + x + 1$. In $\mathbb{F}_2$ ist $-1 = 1$ und daher $x^2 \equiv x + 1 \bmod N$. Elemente von $\mathbb{F}_2[x]/(N(x))$ sind daher Äquivalenzklassen, die von Polynome bis zu einem Grad 1 mit Koeffizienten aus $\mathbb{F}_2$ erzeugt werden. Folglich hat $\mathbb{F}_2[x]/(N(x))$ genau vier Elemente, nämlich $[0]_N$, $[1]_N$, $[x]_N$ und $[x + 1]_N$. Die Additions- bzw. Multiplikationstabellen in $\mathbb{F}_2[x]/(N(x))$ sind durch

$+$	0	1	x	$x+1$
0	0	1	x	$x+1$
1	1	0	$x+1$	x
x	x	$x+1$	0	1
$x+1$	$x+1$	x	1	0

$\cdot$	0	1	x	$x+1$
0	0	0	0	0
1	0	1	x	$x+1$
x	0	x	$x+1$	1
$x+1$	0	$x+1$	1	x

gegeben, wobei wir die eckigen Klammern und den Index $_N$ fortgelassen haben. Offenbar ist

$$x^2 + x + 1 = (x + 1) + (x + 1) = 0, \qquad (x + 1)^2 + (x + 1) + 1 = x + (x + 1) + 1 = 0.$$

Die vier Elemente von $\mathbb{F}_2[x]/(N(x))$ genügen also der Gleichung $x^4 - x = 0$. Da $N \in \mathbb{F}_2[x]$ irreduzibel ist, ist $\mathbb{F}_2[x]/(N(x))$ mit der so erklärten Addition und Multiplikation ein Körper ist, was hier aber natürlich direkt nachgeprüft werden kann. Damit ist ein Körper mit vier Elementen gefunden. □

Beispiel 4.16 Jetzt wollen wir noch einen Schritt weitergehen und einen Körper mit $q = 2^3 = 8$ Elementen konstruieren, wobei wir weitgehend der Vorgehensweise im letzten Beispiel folgen werden. Wir gehen wieder aus von dem Restklassenkörper $\mathbb{F}_2$. In $\mathbb{F}_2[x]$ betrachten wir das Polynom

$$f(x) := x^8 - x = x(x - 1)(x^6 + x^5 + x^4 + x^3 + x^2 + x + 1).$$

Der dritte Faktor zerfällt in $\mathbb{F}_2[x]$ in zwei Faktoren, es ist nämlich

$$x^6 + x^5 + x^4 + x^3 + x^2 + x + 1 = (x^3 + x^2 + 1)(x^3 + x + 1),$$

wobei wir $2 = 0$ in $\mathbb{F}_2$ ausgenutzt haben. Beide Faktoren sind *irreduzibel* in $\mathbb{F}_2[x]$, d. h. sie können nicht als Produkt von zwei nichtkonstanten Polynomen aus $\mathbb{F}_2[x]$ dargestellt werden. Wir setzen $N(x) := x^3 + x^2 + 1$ (ebenso hätten wir auch den anderen Faktor wählen können) und gehen wie im vorigen Beispiel vor. Wegen $x^3 \equiv x^2 + 1 \bmod N$ besteht der Körper $\mathbb{F}_2[x]/(N(x))$ aus Kongruenzklassen, die von Polynomen bis zu einem Grad 2 mit Koeffizienten aus $\mathbb{F}_2$ erzeugt werden. Daher hat $\mathbb{F}_2[x]/(N(x))$ genau acht Elemente, nämlich

$$[0]_N, \quad [1]_N, \quad [x]_N, \quad [x + 1]_N, \quad [x^2]_N, \quad [x^2 + 1]_N, \quad [x^2 + x]_N, \quad [x^2 + x + 1]_N.$$

Wenn wir wieder die eckigen Klammern und den Index $_N$ fortlassen, erhalten wir die Additionstabelle

$+$	0	1	x	$x+1$	x^2	x^2+1	x^2+x	x^2+x+1
0	0	1	x	$x+1$	x^2	x^2+1	x^2+x	x^2+x+1
1	1	0	$x+1$	x	x^2+1	x^2	x^2+x+1	x^2+x
x	x	$x+1$	0	1	x^2+x	x^2+x+1	x^2	x^2+1
$x+1$	$x+1$	x	1	0	x^2+x+1	x^2+x	x^2+1	x^2
x^2	x^2	x^2+1	x^2+x	x^2+x+1	0	1	x	$x+1$
x^2+1	x^2+1	x^2	x^2+x+1	x^2+x	1	0	$x+1$	x
x^2+x	x^2+x	x^2+x+1	x^2	x^2+1	x	$x+1$	0	1
x^2+x+1	x^2+x+1	x^2+x	x^2+1	x^2	$x+1$	x	1	0

sowie die Multiplikationstabelle

$\cdot$	0	1	x	$x+1$	x^2	x^2+1	x^2+x	x^2+x+1
0	0	0	0	0	0	0	0	0
1	0	1	x	$x+1$	x^2	x^2+1	x^2+x	x^2+x+1
x	0	x	x^2	x^2+x	x^2+1	x^2+x+1	1	$x+1$
$x+1$	0	$x+1$	x^2+x	x^2+1	1	x	x^2+x+1	x^2
x^2	0	x^2	x^2+1	1	x^2+x+1	$x+1$	x	x^2+x
x^2+1	0	x^2+1	x^2+x+1	x	$x+1$	x^2+x	x^2	1
x^2+x	0	x^2+x	1	x^2+x+1	x	x^2	$x+1$	x^2+1
x^2+x+1	0	x^2+x+1	$x+1$	x^2	x^2+x	1	x^2+1	x

Damit haben wir einen Körper mit acht Elementen gefunden. □

Satz 4.17 *Zu jeder Primzahlpotenz $q = p^m$ gibt es einen Körper $\mathbb{F}_q$ der Ordnung $q = p^m$. Dieser ist bis auf Isomorphie eindeutig bestimmt.*

Beweis: Gegeben sei eine Primzahl p, ein $m \in \mathbb{N}$, ferner sei $q := p^m$. Zunächst zeigen wir die *Existenz* eines Körpers $\mathbb{F}_q$ der Ordnung q. Für $m = 1$ bzw. $q = p$ ist dies klar, da der Restklassenkörper $\mathbb{F}_p := \mathbb{Z}/p\mathbb{Z}$ die Ordnung p besitzt. Auch für $m > 1$ gehen wir von $\mathbb{K} := \mathbb{F}_p$ aus, einem Körper der Charakteristik p, und definieren $f \in \mathbb{K}[x]$ durch $f(x) := x^q - x$. Im vorigen Unterabschnitt haben wir bewiesen, dass es einen $\mathbb{K}$ (bzw. ein isomorphes Bild von $\mathbb{K}$) enthaltenden Körper $\tilde{\mathbb{L}}$ gibt, über dem f in Linearfaktoren zerfällt, also $\alpha_1, \ldots, \alpha_q \in \tilde{\mathbb{L}}$ mit

$$f(x) = x^q - x = \prod_{i=1}^{q} (x - \alpha_i)$$

existieren. Nun definieren wir

$$\mathbb{L} := \{x \in \tilde{\mathbb{L}} : f(x) = 0\} = \{x \in \tilde{\mathbb{L}} : x^q = x\} = \{\alpha_1, \ldots, \alpha_q\}.$$

Wir zeigen jetzt, dass $|\mathbb{L}| = q$ bzw. $\alpha_1, \ldots, \alpha_q$ paarweise verschieden sind. Hierzu definieren wir die formale Ableitung D: $\tilde{\mathbb{L}}[x] \longrightarrow \tilde{\mathbb{L}}[x]$ durch

$$D\Big(\sum_{i=0}^{d} c_i \cdot x^i\Big) := \sum_{i=0}^{d-1} (i+1) c_{i+1} \cdot x^i.$$

Wegen $f(x) = x^q - x$ ist dann

$$D(f)(x) = qx^{q-1} - 1 = \underbrace{q \cdot 1}_{=0} \cdot x^{q-1} - 1 = -1 \neq 0,$$

da $\tilde{\mathbb{L}}$ die Charakteristik p besitzt. Ist andererseits $\alpha \in \{\alpha_1, \ldots, \alpha_q\}$ eine mehrfache Nullstelle von f, so lässt sich f darstellen als $f(x) = (x-\alpha)^2 \cdot g(x)$ mit $g \in \tilde{\mathbb{L}}[x]$. Dann ist aber

$$D(f)(x) = 2(x - \alpha) \cdot g(x) + (x - \alpha)^2 \cdot D(g)(x),$$

und folglich $D(f)(\alpha) = 0$. Dies ist eine Widerspruch, alle Nullstellen von f in $\tilde{\mathbb{L}}$ sind einfach und folglich $|\mathbb{L}| = q$. Die Existenz eines Körpers mit $q = p^m$ Elementen ist bewiesen, wenn wir uns davon überzeugt haben, dass mit $\tilde{\mathbb{L}}$ auch $\mathbb{L}$ ein Körper ist. Die additiv bzw. multiplikativ neutralen Elemente 0 und 1 aus $\tilde{\mathbb{L}}$ liegen in $\mathbb{L}$. Weiter bleibt zu zeigen, dass mit $x, y \in \mathbb{L}$ auch $x + y$, $x \cdot y$ sowie $-x$ und x^{-1} (falls $x \neq 0$) zu $\mathbb{L}$ gehören. Zunächst zeigen wir die Abgeschlossenheit von $\mathbb{L}$ bezüglich der Addition, also

$$x, y \in \mathbb{L} \Longrightarrow x + y \in \mathbb{L}$$

bzw.

$$x, y \in \tilde{\mathbb{L}}, \quad x^{p^m} = x,\ y^{p^m} = y \Longrightarrow (x + y)^{p^m} = x + y.$$

Diese Aussage zeigen wir durch vollständige Induktion nach m. Der Induktionsanfang liegt bei $m = 1$ bzw. der Aussage

$$x, y \in \tilde{\mathbb{L}}, \quad x^p = x,\ y^p = y \Longrightarrow (x + y)^p = x + y.$$

Nach der binomischen Formel ist

$$(x+y)^p = x^p + px^{p-1}y + \binom{p}{2}x^{p-2}\cdot y^2 + \cdots + \binom{p}{p-1}x\cdot y^{p-1} + y^p.$$

Die Binomialkoeffizienten

$$\binom{p}{k} = \frac{p(p-1)\cdot\cdots\cdot(p-k+1)}{1\cdot 2\cdot\cdots\cdot k}, \qquad k = 1,\ldots,p-1,$$

sind durch p teilbar. Da mit $\mathbb{K}$ auch $\tilde{\mathbb{L}}$ die Charakteristik p besitzt, ist $\binom{p}{k}\cdot 1 = 0$ und somit $(x+y)^p = x^p + y^p$. Hieraus folgt aber, dass der Induktionsanfang gelegt ist. Ist die Aussage für m richtig, so ist sie auch für $m+1$ richtig. Denn

$$(x+y)^{p^{m+1}} = (x+y)^{p^m\cdot p} = ((x+y)^{p^m})^p = (x+y)^p = x+y.$$

Damit ist die Abgeschlossenheit von $\mathbb{L}$ bezüglich der Addition nachgewiesen. Die Abgeschlossenheit von $\mathbb{L}$ bezüglich der Multiplikation bzw. die Gültigkeit der Aussage

$$x, y \in \tilde{\mathbb{L}}, \quad x^q = x,\ y^q = y \Longrightarrow (x\cdot y)^q = x\cdot y$$

ist offenbar wegen $(x\cdot y)^q = x^q\cdot y^q$ richtig. Mit x gehört auch das additiv inverse Element $-x$ zu $\mathbb{L}$, d. h. es gilt die Implikation

$$x \in \tilde{\mathbb{L}}, \quad x^q = x \Longrightarrow (-x)^q = -x.$$

Denn es ist

$$(-x)^q = (-1)^q x = (-1)^{q+1}(-x).$$

Zu zeigen bleibt also: Ist $q = p^m$, so ist $(-1)^{q+1} = 1$. Für $p = 2$ und damit gerades q ist dies richtig, da in diesem Falle $-1 = 1$ bzw. $x = -x$ gilt. Für $p > 2$ ist p ungerade, damit auch $q = p^m$ für alle $m \in \mathbb{N}$ ungerade und folglich $(-1)^{q+1} = 1$. Damit ist gezeigt, dass $\mathbb{L}$ ein Körper mit $q = p^m$ Elementen ist.

Jetzt kommen wir zum Beweis der (im wesentlichen) *Eindeutigkeit* eines Körpers mit $q = p^m$ Elementen. Beim Beweis orientieren wir uns an dieser Quelle, siehe auch E. Artin (1942, Theorem 10). Sei $\mathbb{L}$ der im ersten Teil des Beweises konstruierte Körper mit $q = p^m$ Elementen. Dieser enthält den Primkörper $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ (bzw. ein isomorphes Bild). Die Elemente von $\mathbb{L}$ sind die q paarweise verschiedenen Nullstellen von $f \in \mathbb{F}_p[x]$ mit $f(x) := x^q - x$ in einem Körper $\tilde{\mathbb{L}}$, über dem f in Linearfaktoren zerfällt. Daher ist $\mathbb{L} \supset \mathbb{F}_p$ ein *Zerfällungskörper* von $f \in \mathbb{F}_p[x]$, d. h. f zerfällt über $\mathbb{L}$ in Linearfaktoren und $\mathbb{L}$ ist minimal bezüglich dieser Eigenschaft. Sei nun $\mathbb{F}$ ein weiterer Körper mit $q = p^m$ Elementen, von dem wir wieder annehmen können, dass er $\mathbb{F}_p$ enthält. Mit $\mathbb{F}^* := \mathbb{F}\setminus\{0\}$ ist die multiplikative Gruppe $(\mathbb{F}^*,\cdot)$ zyklisch und daher wegen des Satzes von Lagrange $x^q - x = 0$ für alle $x \in \mathbb{F}$, wie wir in Unterabschnitt 4.2 bewiesen haben. Daher hat f genau q paarweise verschiedene Nullstellen in $\mathbb{F}$ und somit zerfällt $f \in \mathbb{F}_p[x]$ in $\mathbb{F}$ in Linearfaktoren. Da die q Nullstellen von f in $\mathbb{F}$ paarweise verschieden sind, ist $\mathbb{F}$ ebenfalls ein Zerfällungskörper von $f \in \mathbb{F}_p[x]$. Es bleibt zu zeigen, dass $\mathbb{F}$ und $\mathbb{L}$ isomorph sind. Dies geschieht durch einen Beweis der Aussage, dass ein Zerfällungskörper für ein Polynom im wesentlichen eindeutig bestimmt ist:

- Sei $\mathbb{K}$ ein Körper und $f \in \mathbb{K}[x]$ ein Polynom mit $\deg(f) \geq 1$. Sind dann $\mathbb{L}_1$ und $\mathbb{L}_2$ zwei Zerfällungskörper für f, so sind $\mathbb{L}_1$ und $\mathbb{L}_2$ isomorph. Genauer existiert ein Isomorphismus $F\colon \mathbb{L}_1 \longrightarrow \mathbb{L}_2$ mit $F(k) = k$ für alle $k \in \mathbb{K}$.

Da $\mathbb{L}_2$ ein Zerfällungskörper für $f \in \mathbb{K}[x]$ ist, ist $\mathbb{L}_2 \supset \mathbb{K}$ eine Körpererweiterung von $\mathbb{K}$. Sei $i\colon \mathbb{K} \longrightarrow \mathbb{L}_2$ die natürliche Inklusionsabbildung und $\underline{i}\colon \mathbb{K}[x] \longrightarrow i(\mathbb{K})[x] \subset \mathbb{L}_2[x]$ definiert durch

$$\underline{i}(a_0 + a_1 \cdot x + \cdots + a_d \cdot x^d) := i(a_0) + i(a_1) \cdot x + \cdots + i(a_d) \cdot x^d.$$

Wir werden zeigen, dass wir $i\colon \mathbb{K} \longrightarrow \mathbb{L}_2$ zu einem Ringhomomorphismus[11] $F\colon \mathbb{L}_1 \longrightarrow \mathbb{L}_2$ erweitern können. Angenommen, wir hätten dies schon bewiesen. Wir zeigen, dass $F\colon \mathbb{L}_1 \longrightarrow \mathbb{L}_2$ sogar ein Isomorphismus ist, also auch noch injektiv und surjektiv ist. Ein Ringhomomorphismus zwischen Körpern ist injektiv[12]. Jetzt überlegen wir uns, dass $F\colon \mathbb{L}_1 \longrightarrow \mathbb{L}_2$ auch surjektiv ist. Mit $\mathbb{L}_1$ ist offenbar auch $F(\mathbb{L}_1) \subset \mathbb{L}_2$ ein Körper. Da $f \in \mathbb{K}[x]$ über $\mathbb{L}_1$ in Linearfaktoren zerfällt, existieren mit $d := \deg(f)$ Nullstellen $\alpha_1, \ldots, \alpha_d \in \mathbb{L}_1$ von f in $\mathbb{L}_1$. Dann besitzt $\underline{i}(f) = f$ in $\mathbb{L}_2$ die Nullstellen $F(\alpha_1), \ldots, F(\alpha_d)$ aus dem Körper $F(\mathbb{L}_1)$, d. h. f zerfällt in $F(\mathbb{L}_1) \subset \mathbb{L}_2$ in Linearfaktoren. Da $\mathbb{L}_2$ ein Zerfällungskörper für f ist, ist $F(\mathbb{L}_1) = \mathbb{L}_2$, also $F\colon \mathbb{L}_1 \longrightarrow \mathbb{L}_2$ auch surjektiv und insgesamt ein Isomorphismus. Da F eine Erweiterung von i, der natürlichen Inklusionsabbildung von K nach $\mathbb{L}_2$ ist, ist $F(k) = k$ für alle $k \in \mathbb{K}$. Es bleibt, die oben offen gelassene Lücke zu schließen, dass man nämlich die natürliche Inklusionsabbildung $i\colon \mathbb{K} \longrightarrow \mathbb{L}_2$ zu einem Ringhomomorphismus $F\colon \mathbb{L}_1 \longrightarrow \mathbb{L}_2$ erweitern kann. Dies geschieht dadurch, dass wir die folgende etwas allgemeinere Aussage beweisen:

- Sei $\mathbb{K}$ ein Körper, $f \in \mathbb{K}[x]$ und $\mathbb{L}_1$ ein Zerfällungskörper für f. Ist dann $\mathbb{L}_2$ ein Oberkörper von $\mathbb{K}$ bzw. $\mathbb{L}_2/\mathbb{K}$ eine Körpererweiterung und $i\colon \mathbb{K} \longrightarrow \mathbb{L}_2$ ein

[11] Eine Abbildung $F\colon \mathbb{L}_1 \longrightarrow \mathbb{L}_2$ heißt ein *Ringhomomorphismus*, wenn $F(1_{\mathbb{L}_1}) = 1_{\mathbb{L}_2}$, $F(a+b) = F(a) + F(b)$ und $F(a \cdot b) = F(a) \cdot F(b)$ für alle $a, b \in \mathbb{L}_1$.

[12] Denn: Sei $F\colon \mathbb{L}_1 \longrightarrow \mathbb{L}_2$ ein Ringhomomorphismus zwischen den Körpern $\mathbb{L}_1, \mathbb{L}_2$. Seien $a, b \in \mathbb{L}_1$ mit $F(a) = F(b)$ vorgegeben. Dann ist

$$F(a-b) = F(a + (-b)) = F(a) + F(-b) = F(a) - F(b) = 0_{\mathbb{L}_2}.$$

Hierbei haben wir benutzt, dass aus $0_{\mathbb{L}_1} = 0_{\mathbb{L}_1} + 0_{\mathbb{L}_1}$ folgt, dass $F(0_{\mathbb{L}_1}) = F(0_{\mathbb{L}_1}) + F(0_{\mathbb{L}_1})$ und damit $F(0_{\mathbb{L}_1}) = 0_{\mathbb{L}_2}$. Folglich ist

$$0_{\mathbb{L}_2} = F(0_{\mathbb{L}_1}) = F(b + (-b)) = F(b) + F(-b)$$

für alle $b \in \mathbb{L}_1$. Daher ist $F(-b) = -F(b)$ für alle $b \in \mathbb{L}_1$. Zum Nachweis der Injektivität von F ist $a = b$ bzw. $a - b = 0$ zu zeigen. Angenommen, dies sei nicht der Fall, es sei also $a - b \neq 0_{\mathbb{L}_1}$. Mit $x := (a-b)^{-1}$ ist

$$1_{\mathbb{L}_2} = F(1_{\mathbb{L}_1}) = F(x \cdot (a-b)) = F(x) \cdot F(a-b) = F(x) \cdot 0_{\mathbb{L}_2} = 0_{\mathbb{L}_2},$$

wobei die letzte Gleichung aus

$$F(x) \cdot 0_{\mathbb{L}_2} = F(x) \cdot (0_{\mathbb{L}_2} + 0_{\mathbb{L}_2}) = F(x) \cdot 0_{\mathbb{L}_2} + F(x) \cdot 0_{\mathbb{L}_2}$$

folgt. Damit haben wir einen Widerspruch zu $0_{\mathbb{L}_2} \neq 1_{\mathbb{L}_2}$ erhalten und es ist nachgewiesen, dass ein Ringhomomorphismus zwischen Körpern injektiv ist.

Ringhomomorphismus, so kann i genau dann zu einem Ringhomomorphismus $F\colon \mathbb{L}_1 \longrightarrow \mathbb{L}_2$ erweitert werden, wenn $\underline{i}(f) \in i(\mathbb{K})[x]$ über $\mathbb{L}_2$ in Linearfaktoren zerfällt[13]. Hierbei ist $\underline{i}\colon \mathbb{K}[x] \longrightarrow i(\mathbb{K})[x] \subset \mathbb{L}_2[x]$ wie oben durch

$$\underline{i}(a_0 + a_1 \cdot x + \cdots + a_d \cdot x^d) := i(a_0) + i(a_1) \cdot x + \cdots + i(a_d) \cdot x^d$$

definiert. Offenbar ist $\underline{i}\colon \mathbb{K}[x] \longrightarrow i(\mathbb{K})[x]$ ein Ringhomomorphismus.

Der Beweis besteht aus zwei Teilen. Im ersten Teil zeigen wir die einfache Richtung ($\Longrightarrow$). Da $\mathbb{L}_1$ ein Zerfällungskörper von $f \in \mathbb{K}[x]$ ist, besitzt f in $\mathbb{L}_1$ (nicht notwendig verschiedene) Nullstellen $\alpha_k \in \mathbb{L}_1$, $k = 1, \ldots, d$, wobei $d := \deg(f)$. Das Polynom $f \in \mathbb{K}[x]$ hat eine Darstellung

$$f = a_0 + a_1 \cdot x + \cdots + a_d \cdot x^d$$

mit Koeffizienten $a_0, \ldots, a_d \in \mathbb{K}$. Andererseits hat f in $\mathbb{L}_1$ die Darstellung

$$f = a_d \cdot \prod_{k=1}^{d} (x - \alpha_k).$$

Daher können die Koeffizienten $a_0, \ldots, a_{d-1} \in \mathbb{K}$ in Abhängigkeit vom führenden Koeffizienten a_d und den Nullstellen $\alpha_1, \ldots, \alpha_d \in \mathbb{L}_1$ dargestellt werden. Genauer ist

$$a_{d-k} = (-1)^k a_d \cdot \sum_{1 \le i_1 < i_2 < \cdots < i_k \le d} \alpha_{i_1} \cdot \alpha_{i_2} \cdot \cdots \cdot \alpha_{i_k}, \qquad k = 1, \ldots, d.$$

Z. B. ist

$$\begin{aligned}
a_{d-1} &= -a_d \cdot (\alpha_1 + \alpha_2 + \cdots + \alpha_d), \\
a_{d-2} &= a_d \cdot ((\alpha_1 \cdot \alpha_2 + \cdots + \alpha_1 \cdot \alpha_d) + (\alpha_2 \cdot \alpha_3 + \cdots + \alpha_2 \cdot \alpha_d) + \cdots + \alpha_{d-1} \cdot \alpha_d), \\
&\vdots \\
a_0 &= (-1)^d a_d \cdot (\alpha_1 \cdot \alpha_2 \cdot \cdots \cdot \alpha_d).
\end{aligned}$$

Da $F\colon \mathbb{L}_1 \longrightarrow \mathbb{L}_2$ eine Erweiterung des Ringhomomorphismus $i\colon \mathbb{K} \longrightarrow \mathbb{L}_2$ von $\mathbb{K}$ auf den Oberkörper $\mathbb{L}_1$ und ebenfalls ein Ringhomomorphismus ist, ist

$$i(a_{d-k}) = F(a_{d-k}) = (-1)^k i(a_d) \cdot \sum_{1 \le i_1 < i_2 < \cdots < i_k \le d} F(\alpha_{i_1}) \cdot F(\alpha_{i_2}) \cdot \cdots \cdot F(\alpha_{i_k})$$

für $k = 1, \ldots, d$. Hierbei haben wir ausgenutzt, dass $F(a_d) = i(a_d)$, da $a_d \in \mathbb{K}$ und F eine Erweiterung von i ist. Daher ist

$$\begin{aligned}
\underline{i}(f) &= \underline{i}(a_0 + a_1 \cdot x + \cdots + a_{d-1} \cdot x^{d-1} + a_d \cdot x^d) \\
&= i(a_0) + i(a_1) \cdot x + \cdots + i(a_{d-1}) \cdot x^{d-1} + i(a_d) \cdot x^d \\
&= F(a_0) + F(a_1) \cdot x + \cdots + F(a_{d-1}) \cdot x^{d-1} + i(a_d) \cdot x^d \\
&= i(a_d) \cdot \prod_{k=1}^{d} (x - F(\alpha_k)).
\end{aligned}$$

[13]Man beachte, dass mit $\mathbb{K}$ auch $i(\mathbb{K})$ ein (zu $\mathbb{K}$ isomorpher) Körper ist. Denn als Ringhomomorphismus zwischen Körpern ist i injektiv.

Daher zerfällt $\underline{i}(f) \in i(\mathbb{K})[x][x]$ über $\mathbb{L}_2$ in Linearfaktoren. Genau dies war für die Richtung ($\Longrightarrow$) zu zeigen. Jetzt kommen wir zum zweiten Teil und zeigen die schwierigere Richtung ($\Longleftarrow$). Wir nehmen also an, dass $\underline{i}(f) \in i(\mathbb{K})[x]$ über $\mathbb{L}_2$ in Linearfaktoren zerfällt und beweisen die Behauptung, dass der Ringhomomorphismus $i\colon \mathbb{K} \longrightarrow \mathbb{L}_2$ zu einem Ringhomomorphismus $F\colon \mathbb{L}_1 \longrightarrow \mathbb{L}_2$ erweitert werden kann, durch Induktion nach $[\mathbb{L}_1 : \mathbb{K}]$, wobei wir daran erinnern, dass mit $[\mathbb{L}_1 : \mathbb{K}]$ der Grad der Körpererweiterung $\mathbb{L}_1/\mathbb{K}$ bzw. die Dimension des $\mathbb{K}$-Vektorraumes $\mathbb{L}_1$ bezeichnet wird. Ist $[\mathbb{L}_1 : \mathbb{K}] = 1$, so existiert ein $x \in \mathbb{L}_1 \setminus \{0\}$ derart, dass $\mathbb{L}_1 = \{k \cdot x : k \in \mathbb{K}\}$. Es ist leicht sich zu überlegen, dass notwendigerweise $x \in \mathbb{K}$ und damit $\mathbb{L}_1 = \mathbb{K}$ gilt. Denn zu $x \in \mathbb{L}_1$ existiert ein $k_x \in \mathbb{K}$ mit $k_x \neq 0$ und $k_x \cdot x = 1$. Also ist $x = k_x^{-1} \in \mathbb{K}$, damit $\mathbb{L}_1 = \mathbb{K}$ und es ist nichts zu zeigen. Daher nehmen wir jetzt an, es sei $n := [\mathbb{L}_1 : \mathbb{K}] > 1$ und die Aussage sei für kleinere Grade als n bewiesen. Sei $g \in \mathbb{K}[x]$ ein nichtlinearer, irreduzibler Faktor von $f \in \mathbb{K}[x]$, sei etwa $f = g \cdot p$ mit $p \in \mathbb{K}[x]$. Da $\mathbb{L}_1$ ein Zerfällungskörper für $f \in \mathbb{K}[x]$ ist, besitzt g eine Nullstelle $\alpha \in \mathbb{L}_1$. Denn sei etwa

$$f(x) = \kappa \cdot \prod_{k=1}^{d} (x - \alpha_k)$$

eine Darstellung von f als Produkt linearer Faktoren mit $\alpha_k \in \mathbb{L}_1$, $k = 1, \ldots, d$. Jetzt fassen wir g als ein Polynom aus $\mathbb{L}_1[x]$ auf. Es gibt wegen des zweiten Teiles des Satzes von Kronecker (Satz 4.10) eine Körpererweiterung $\tilde{\mathbb{L}}_1/\mathbb{L}_1$, in welcher g eine Nullstelle α besitzt. Dann ist

$$f(\alpha) = \kappa \cdot \prod_{k=1}^{d} (\alpha - \alpha_k) = \underbrace{g(\alpha)}_{=0} \cdot p(\alpha) = 0.$$

Daher muss α mit einem der α_k übereinstimmen, also in $\mathbb{L}_1$ liegen. Bis auf ein von 0 verschiedenes Vielfaches stimmt g mit dem Minimalpolynom für α überein. Dies folgt offenbar aus dem zweiten Teil von Satz 4.14. Aus Teil 3 (a) desselben Satzes folgt die Existenz eines Isomorphismus

$$S\colon \mathbb{K}[x]/(g(x)) \longrightarrow \mathbb{K}(\alpha),$$

wobei $\mathbb{K}(\alpha)$ der kleinste Körper in $\mathbb{L}_1$ ist, der $\mathbb{K}$ und α enthält. Mit $g \in \mathbb{K}[x]$ ist auch $\underline{i}(g) \in i(\mathbb{K})[x]$ irreduzibel. Folglich ist die durch

$$T([h]_g) := [\underline{i}(h)]_{\underline{i}(g)}, \qquad [h]_g \in \mathbb{K}[x]/(g(x))$$

definierte Abbildung

$$T\colon \mathbb{K}[x]/(g(x)) \longrightarrow i(\mathbb{K})[x]/(\underline{i}(g)(x))$$

eine Abbildung zwischen Körpern. Offenbar ist T sogar ein Ringhomomorphis zwischen Körpern, und damit injektiv. Da T aber auch surjektiv ist, ist T ein Isomorphismus. Nach Voraussetzung zerfällt $\underline{i}(f) \in i(\mathbb{K})[x]$ über $\mathbb{L}_2$ in Linearfaktoren, es ist also

$$\underline{i}(f)(x) = i(\kappa) \cdot \prod_{k=1}^{d} (x - \beta_k)$$

mit $\beta_k \in \mathbb{L}_2$, $k = 1, \ldots, d$. Da $\underline{i}(g) \in i(\mathbb{K})[x]$ ein irreduzibler Faktor von $\underline{i}(f) \in i(\mathbb{K})[x]$, schließt man wie gerade eben, dass $\underline{i}(g)$ eine Nullstelle $\beta \in \{\beta_1, \ldots, \beta_d\}$ besitzt und durch

$$U([\underline{i}(h)]_{\underline{i}(g)}) := \underline{i}(h)(\beta)$$

ein Isomorphismus

$$U\colon i(\mathbb{K})[x]/(\underline{i}(g)(x)) \longrightarrow i(\mathbb{K})(\beta)$$

definiert ist. Hierbei ist $i(\mathbb{K})(\beta)$ der kleinste Körper in $\mathbb{L}_2$, der $i(\mathbb{K})$ und $\beta \in \mathbb{L}_2$ enthält. Schließlich bezeichne

$$V\colon i(\mathbb{K})(\beta) \longrightarrow \mathbb{L}_2$$

die natürliche Inklusionsabbildung. Damit haben wir einen Ringhomomorphismus

$$W := VUTS^{-1}\colon \mathbb{K}(\alpha) \xrightarrow{S^{-1}} \mathbb{K}[x]/(g(x)) \xrightarrow{T} i(\mathbb{K})[x]/(\underline{i}(g)(x)) \xrightarrow{U} i(\mathbb{K})(\beta) \xrightarrow{V} \mathbb{L}_2.$$

Wir wollen uns überlegen, dass $W(k) = i(k)$ für alle $k \in \mathbb{K}$ gilt, also W eine Erweiterung des Ringhomomorphismus $i\colon \mathbb{K} \longrightarrow \mathbb{L}_2$ von $\mathbb{K}$ auf $\mathbb{K}(\alpha)$ ist. Sei also $k \in \mathbb{K}$ gegeben. Dann haben wir

$$k \xrightarrow{S^{-1}} \{k + g \cdot p : p \in \mathbb{K}[x]\} \xrightarrow{T} \{i(k) + \underline{i}(g) \cdot \underline{i}(p) : p \in \mathbb{K}[x]\} \xrightarrow{U} i(k),$$

da $\underline{i}(g)(\beta) = 0$. Also ist W in der Tat eine Erweiterung von i von $\mathbb{K}$ auf $\mathbb{K}(\alpha)$. Wegen Teil 3 (c) von Satz 4.14 ist

$$[\mathbb{L}_1 : \mathbb{K}] = [\mathbb{L}_1 : \mathbb{K}(\alpha)] \underbrace{[\mathbb{K}(\alpha) : \mathbb{K}]}_{>1}.$$

Hierbei ist $[\mathbb{K}(\alpha) : \mathbb{K}]$ nach Teil 3 (b) von Satz 4.14 gleich dem Grad des Minimalpolynoms für α bzw. dem Grad des nichtlinearen, irreduziblen Faktors g von f und dieser ist größer als 1. Folglich ist

$$[\mathbb{L}_1 : \mathbb{K}(\alpha)] = \frac{[\mathbb{L}_1 : \mathbb{K}]}{[\mathbb{K}(\alpha) : \mathbb{K}]} < [\mathbb{L}_1 : \mathbb{K}].$$

Jetzt können wir die Induktionsannahme auf den Fall anwenden, dass $\mathbb{K}(\alpha)$ (statt $\mathbb{K}$) der Grundkörper ist. Daher kann W und damit auch i zu einem Ringhomomorphismus von $\mathbb{L}_1$ nach $\mathbb{L}_2$ erweitert werden. Damit ist auch der schwierigere Teil ($\Longleftarrow$) der Hilfsbehauptung und schließlich auch der gesamte Satz bewiesen. □

5 Trennung konvexer Mengen in linearen normierten Räumen

In diesem und dem nächsten Abschnitt setzen wir elementare Kenntnisse der linearen Funktionalanalysis voraus. So sei etwa bekannt, was ein linearer normierter Raum (die Norm wird stets mit $\|\ \|$ bezeichnet, wobei wir Normen in unterschiedlichen Räumen nicht unterschiedlich bezeichnen, das entsprechende gilt für das Nullelement 0) und ein

Banachraum ist, ferner was man unter linearen, stetigen bzw. beschränkten Abbildungen zwischen linearen normierten Räumen und ihrer Norm (die ebenfalls wieder mit $\| \ \|$ bezeichnet wird) versteht.

Auf die Trennung konvexer Mengen im $\mathbb{R}^n$ durch Hyperebenen sind wir in Abschnitt 49 der *Merkwürdigen Mathematik* (J. Werner (2013)) eingegangen. Ziel in diesem Abschnitt wird es sein, Trennungssätze für konvexe Mengen in linearen normierten Räumen zu formulieren und zu beweisen.

5.1 Hyperebenen in einem linearen normierten Raum

Zunächst muss der Begriff *Hyperebene* geklärt werden. Ein *affiner Teilraum* A eines linearen Raumes E ist eine Menge, die mit zwei Punkten auch die gesamte Gerade durch diese Punkte enthält, für die also die Implikation

$$x, y \in A, \ \lambda \in \mathbb{R} \Longrightarrow (1 - \lambda)x + \lambda y \in A$$

gilt.

Definition 5.1 Sei E ein (reeller) linearer Raum. Ein affiner Teilraum $H \subset E$ heißt *Hyperebene* in E, falls H ein maximaler, echter affiner Teilraum von E ist, d. h. wenn

1. $H \subset E$ ist ein affiner Teilraum und es ist $H \neq E$.
2. Ist M ein affiner Teilraum von E mit $H \subset M$, so ist $M = E$ oder $M = H$.

In einem linearen normierten Raum ist der Abschluss eines affinen Teilraums ebenfalls ein affiner Teilraum. Eine Hyperebene H in einem linearen normierten Raum E ist daher entweder abgeschlossen, also cl $(H) = H$, oder dicht in E, also cl $(H) = E$.

Ist E ein linearer normierter Raum, so bezeichnen wir mit $E^* := L(E, \mathbb{R})$ den /em Dualraum von E, d. h. die Menge der linearen, stetigen Abbildungen von E nach $\mathbb{R}$. E^* ist in kanonischer Weise ein linearer Raum, der für $l \in E^*$ durch $\|l\| := \sup_{x \neq 0} |l(x)|/\|x\|$ zu einem linearen normierten Raum wird.

Nun folgt der Nachweis dafür, dass eine abgeschlossene Hyperebene in einem linearen normierten Raum E mit Hilfe eines von Null verschiedenen Elementes des Dualraums E^* dargestellt werden kann. Natürlich ist dadurch noch nicht die Existenz einer abgeschlossenen Hyperebene bzw. eines nichttrivialen Elements in E^* bewiesen!

Satz 5.2 *Sei E ein linearer normierter Raum. Dann ist $H \subset E$ genau dann eine abgeschlossene Hyperebene in E, wenn ein Paar $(l, \gamma) \in (E^* \setminus \{0\}) \times \mathbb{R}$ existiert mit $H = \{x \in E : l(x) = \gamma\}$.*

Beweis: Im ersten Teil des Beweises nehmen wir an, es sei $(l, \gamma) \in (E^* \setminus \{0\}) \times \mathbb{R}$ und $H := \{x \in E : l(x) = \gamma\}$. Zu zeigen ist, dass H eine abgeschlossene Hyperebene in E ist. Offensichtlich ist H ein affiner Teilraum von E. Es ist $H \neq E$, also H ein echter affiner Teilraum von E. Denn andernfalls wäre notwendig $\gamma = 0$ und $l(x) = 0$ für alle $x \in E$ und damit l das Nullelement in E^*, was ausgeschlossen ist. H ist auch abgeschlossen, da $l \colon E \longrightarrow \mathbb{R}$ stetig ist. Zu zeigen bleibt, dass H ein *maximaler*, echter

affiner Teilraum von E ist. Sei hierzu $z \in E \setminus H$ und $M \subset E$ ein affiner Teilraum, der H und z enthält. Wir zeigen, dass $M = E$ und damit H maximal ist. Sei $x_0 \in H$ fest vorgegeben und $x \in E$ beliebig. Dann ist

$$x = \underbrace{\frac{l(x-x_0)}{l(z-x_0)}}_{=:\alpha(x)}(z-x_0) + \underbrace{\Big(x - \frac{l(x-x_0}{l(z-x_0)}(z-x_0)\Big)}_{=:y(x)} = \alpha(x)(z-x_0) + y(x)$$

mit $\alpha(x) \in \mathbb{R}$ und x_0, $y(x) \in H \subset M$ sowie $z \in M$. Folglich ist

$$x = (1-\alpha(x))y(x) + \alpha(x)(z - x_0 + y(x)) \in M,$$

da

$$z - x_0 + y(x) = \frac{1}{2}(\underbrace{2z - x_0}_{\in M}) + \frac{1}{2}(\underbrace{2y(x) - z_0}_{\in M}) \in M.$$

Damit ist der erste Teil des Beweises abgeschlossen.

Im zweiten Teil des Beweises nehmen wir an, H sei eine abgeschlossene Hyperebene in E mit $0 \in H$. Am Schluss geben wir an, wie man den allgemeinen Fall hierauf zurückführt. Zunächst zeigen wir, dass eine nichttriviale lineare Abbildung $l\colon E \longrightarrow \mathbb{R}$ mit

$$H = \{x \in E : l(x) = 0\}$$

existiert, anschließend wird die Stetigkeit von l bewiesen. Man wähle ein $z \notin H$ beliebig. Ein beliebiges $x \in E$ besitzt eine Darstellung $x = \alpha(x)z + y(x)$ mit $\alpha(x) \in \mathbb{R}$ und $y(x) \in H$, wie wir im ersten Teil des Beweises gesehen haben. Diese Darstellung ist offenbar eindeutig. Nun definiere man $l\colon E \longrightarrow \mathbb{R}$ durch $l(x) := \alpha(x)$. Offenbar ist l eine nichttriviale lineare Abbildung von E nach $\mathbb{R}$ und $H = \{x \in E : l(x) = 0\}$. Zu zeigen bleibt die Stetigkeit bzw. Beschränktheit von l. Da H nach Voraussetzung abgeschlossen und $z \notin H$ ist, besitzt z einen positiven Abstand von H, d. h. es ist

$$d := \inf_{y \in H} \|z - y\| > 0.$$

Dann ist

$$\sup_{x \neq 0} \frac{|l(x)|}{\|x\|} = \sup_{x \neq 0} \frac{|\alpha(x)|}{\|\alpha(x)z + y(x)\|} = \sup_{x \neq 0} \frac{1}{\|z + y(x)/\alpha(x)\|} \leq \frac{1}{d},$$

womit die Stetigkeit von l bewiesen ist. Zum Schluss befreien wir uns von der Voraussetzung $0 \in H$. Man wähle hierzu ein beliebiges $x_0 \in H$ und setze $V := H - x_0$. Dann ist V eine abgeschlossene Hyperebene in E mit $0 \in V$. Also existiert nach dem gerade eben bewiesenen Ergebnis ein $l \in E^* \setminus \{0\}$ mit $V = \{y \in E : l(y) = 0\}$. Dann ist aber

$$H = \{x \in E : l(x) = l(x_0)\}.$$

Denn für $x = x_0 + y \in H$ mit $y \in V$ ist $l(x) = l(x_0)$ und daher

$$H \subset \{x \in E : l(x) = l(x_0)\}.$$

Nach dem schon bewiesenen ersten Teil dieses Satzes steht hier rechts ein echter affiner Teilraum von E. Wegen der Maximalitätseigenschaft von Hyperebenen gilt sogar Gleichheit und der Satz ist bewiesen. □

Wie im endlichdimensionalen Fall erzeugt eine abgeschlossene Hyperebene

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

einen nichtnegativen (abgeschlossenen) Halbraum H^+ sowie einen nichtpositiven (abgeschlossenen) Halbraum H^- mittels

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

Offenbar können die aus dem endlichdimensionalen Fall (siehe Abschnitt 49 aus *Merkwürdige Mathematik* (J. Werner (2013)) bekannten Trennungsbegriffe in naheliegenderweise auf lineare normierte Räume übertragen werden.

5.2 Inneres und Abschluss konvexer Mengen

In einem Satz formulieren wir einfache Aussagen über konvexe Mengen in einem linearen normierten Raum.

Satz 5.3 *Sei E ein linearer normierter Raum und $A \subset E$ eine nichtleere, konvexe Teilmenge. Dann gilt:*

1. *Ist $x \in \text{int}\,(A)$ und $y \in \text{cl}\,(A)$, so ist*

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

2. $\text{int}\,(A)$ *und* $\text{cl}\,(A)$ *sind konvex.*

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

Beweis: Im ersten Teil des Beweises geben wir uns ein $\lambda \in (0, 1)$ vor, setzen $z := (1 - \lambda)x + \lambda y$ und zeigen $z \in \text{int}\,(A)$. Wegen $x \in \text{int}\,(A)$ existiert ein $\epsilon > 0$ mit $B[x; \epsilon] \subset A$, wobei $B[x; \epsilon]$ die abgeschlossene Kugel um x mit dem Radius $\epsilon > 0$ bedeutet. Wir haben zu zeigen, dass es um z eine ganz in A gelegene Kugel gibt. Genauer zeigen wir, dass $B[z; (1 - \lambda)\epsilon/2] \subset A$. Hierzu sei $\hat{z} \in B[z; (1 - \lambda)\epsilon/2]$ beliebig. Wegen $y \in \text{cl}\,(A)$ existiert in jeder Kugel um y ein Element von A. Insbesondere gibt es ein $\hat{y} \in B[y; (1 - \lambda)\epsilon/(2\lambda)] \cap A$. 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

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

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.

Aus dem ersten Teil des Satzes folgt aus der Konvexität von A auch die von int (A). Zum Nachweis der Konvexität von cl (A) geben wir uns $x, y \in \text{cl}\,(A)$ sowie $\lambda \in [0,1]$ vor und setzen $z := (1-\lambda)x + \lambda y$. Zu zeigen ist, dass $A \cap B[z;\epsilon] \neq \emptyset$ für alle $\epsilon > 0$. Sei also ein $\epsilon > 0$ vorgegeben. Zu $x, y \in \text{cl}\,(A)$ existieren

$$x_\epsilon \in A \cap B[x;\epsilon], \qquad y_\epsilon \in A \cap B[y;\epsilon].$$

Wegen der Konvexität von A ist $z_\epsilon := (1-\lambda)x_\epsilon + \lambda y_\epsilon \in A$. Ferner ist

$$\|z_\epsilon - z\| = \|(1-\lambda)(x_\epsilon - x) + \lambda(y_\epsilon - y)\| \leq (1-\lambda)\epsilon + \lambda\epsilon = \epsilon,$$

insgesamt also $z_\epsilon \in A \cap B]z;\epsilon]$. Damit ist auch der zweite Teil des Satzes bewiesen.

Im dritten Teil des Satzes wird int $(A) \neq \emptyset$ bzw. die Existenz eines $x \in \text{int}\,(A)$ vorausgesetzt. Wegen int $(A) \subset A$ gilt trivialerweise die Inklusion cl (int (A)) $\subset$ cl (A). Wegen des ersten Teil des Satzes ist $[x, y) \subset \text{int}\,(A)$ für beliebiges $y \in \text{cl}\,(A)$. Hieraus folgt die umgekehrte Inklusionsbeziehung, womit der gesamte Satz bewiesen ist. □

5.3 Das Lemma von Stone

Ein Trennungssatz für konvexe Mengen ist insbesondere eine Aussage über die Existenz maximaler, echter affiner Teilräume. Es ist daher nicht überraschend, dass das Zorn'sche Lemma eine wichtige Rolle spielt.

Zorn'sches Lemma *Sei C eine halbgeordnete Menge, die induktiv geordnet ist. Dann besitzt $\mathcal{C}$ ein maximales Element.*

Hierbei bedeutet:

1. $\mathcal{C}$ ist *halbgeordnet*:

   In $\mathcal{C}$ ist eine $\leq$-Relation mit folgenden Eigenschaften erklärt:

   (a) Es ist $x \leq x$ für alle $x \in \mathcal{C}$.

   (b) Sind $x, y \in \mathcal{C}$, $x \leq y$ und $y \leq x$, so ist $x = y$.

   (c) Sind $x, y, z \in \mathcal{C}$, $x \leq y$ und $y \leq z$, so ist $x \leq z$.

2. $\mathcal{C}$ ist *induktiv geordnet* (bezüglich der Halbordnung $\leq$):

   Ist $\mathcal{F} \subset \mathcal{C}$ *total geordnet*, d. h. für $f, g \in \mathcal{F}$ ist $f \leq g$ oder $g \leq f$, so existiert ein $x \in \mathcal{C}$ mit $f \leq x$ für alle $f \in \mathcal{F}$.

3. $c \in \mathcal{C}$ ist ein *maximales Element* (bezüglich der Halbordnung $\leq$):

   Ist $y \in \mathcal{C}$ und $c \leq y$, so ist $y = c$.

Das Zorn'sche Lemma geht beim Beweis des folgenden schönen und für den Beweis von Trennungssätzen entscheidenden Lemmas von Stone ein, siehe z. B. R. B. HOLMES (1975, S. 7).

Lemma 5.4 (Stone) *Sei E ein linearer Raum und $A, B \subset E$ nichtleer, konvex und $A \cap B = \emptyset$. Dann existieren konvexe Mengen $C, D \subset E$ mit $A \subset C$, $B \subset D$ sowie $C \cap D = \emptyset$ und $C \cup D = E$.*

Beweis: Sei

$$\mathcal{C} := \{K \subset E : K \text{ konvex}, A \subset K, K \cap B = \emptyset\}.$$

In $\mathcal{C}$ führe man durch die Inklusionsbeziehung eine Halbordnung $\leq$ ein. Für $K_1, K_2 \in \mathcal{C}$ sei also definitionsgemäß $K_1 \leq K_2$ wenn $K_1 \subset K_2$. Dann ist $\mathcal{C}$ induktiv geordnet! Denn ist $\mathcal{F} \subset \mathcal{C}$ totalgeordnet, so setze man $K := \bigcup_{F \in \mathcal{F}} F$. Wir haben zu zeigen, dass $K \in \mathcal{C}$ und $F \leq K$ bzw. $F \subset K$ für alle $F \in \mathcal{F}$. Zunächst ist K konvex, da $\mathcal{F} \subset \mathcal{C}$ totalgeordnet ist[14]. Wegen $A \subset F$ und $F \cap B = \emptyset$ für alle $F \in \mathcal{F}$ ist auch $A \subset K$, $K \cap B = \emptyset$ und damit $K \in \mathcal{C}$, ferner $F \subset K$ bzw. $F \leq K$ für alle $F \in \mathcal{F}$. Das Zorn'sche Lemma liefert die Existenz eines in $\mathcal{C}$ maximalen Elementes $C \in \mathcal{C}$. Definiert man entsprechend

$$\mathcal{D} := \{K \subset E : K \text{ konvex}, B \subset K, C \cap K = \emptyset\},$$

so erhält man mit denselben Argumenten ein maximales Element $D \in \mathcal{D}$. Wir wollen zeigen, dass C und D die gesuchten Mengen sind. Als Elemente von $\mathcal{C}$ bzw. $\mathcal{D}$ sind C und D konvex, ferner Ist $A \subset C$, $B \subset D$ und schließlich $C \cap D = \emptyset$, da $D \in \mathcal{D}$. Zu zeigen bleibt daher, dass $C \cup D = E$. Den Beweis hierfür führen wir durch Widerspruch und nehmen an, es gäbe ein $x \in E$ mit $x \notin C \cup D$. Wegen der Maximalität von C ist co $(C \cup \{x\}) \cap B \neq \emptyset$. Hierbei bedeutet co $(C \cup \{x\})$ die *konvexe Hülle* von $C \cup \{x\}$, also der Durchschnitt aller konvexen Mengen, die C und x enthalten. Also gibt es ein $d_0 \in B$ mit $d_0 \in$ co $(C \cup \{x\})$ und dieser Punkt lässt sich (Beweis?) als Konvexkombination von x und einer gewissen endlichen Anzahl m von Elementen aus C darstellen, d. h. es existieren $\lambda_0, \lambda_1, \ldots, \lambda_m \geq 0$ mit $\sum_{i=0}^m \lambda_i = 1$ und $c_1, \ldots, c_m \in C$ mit

$$d_0 = \lambda_0 x + \sum_{i=1}^m \lambda_i c_i.$$

Natürlich ist $\lambda_0 \in [0, 1]$. Es ist aber sogar $\lambda_0 \in (0, 1)$. Denn wäre $\lambda_0 = 0$, so wäre $d_0 \in C \cap B$, ein Widerspruch zu $\emptyset = C \cap D \supset C \cap B$. Wäre $\lambda_0 = 1$, so wäre $d_0 = x \in B \subset D$, ein Widerspruch zu $x \notin D$. Wegen der Konvexität von C ist

$$c_0 := \sum_{i=1}^m \frac{\lambda_i}{1 - \lambda_0} c_i \in C.$$

[14]Denn sind $x, y \in K$, so existieren nach Definition von K Mengen $F_x, F_y \in \mathcal{F}$ mit $x \in F_x$, $y \in F_y$. Da $\mathcal{F}$ totalgeordnet ist, ist $F_x \subset F_y$ oder $F_y \subset F_x$. Ist etwa ersteres der Fall, so ist wegen der Konvexität von F_y auch $(1 - \lambda)x + \lambda y \in F_y \subset K$ für jedes $\lambda \in [0, 1]$, womit die Konvexität von K bewiesen ist.

Insgesamt existieren also $d_0 \in B \subset D$, $c_0 \in C$ und $\lambda_0 \in (0,1)$ mit

$$d_0 = (1-\lambda_0)c_0 + \lambda_0 x.$$

Ebenso folgt aus der Maximalität von D, dass co $(D \cup \{x\}) \cap C \neq \emptyset$. Hieraus wiederum folgt die Existenz von $c_1 \in C$, $d_1 \in D$, $\lambda_1 \in (0,1)$ mit

$$c_1 = (1-\lambda_1)d_1 + \lambda_1 x.$$

In Abbildung 2 verdeutlichen wir uns die Situation. *Anschaulich* ist klar, dass die

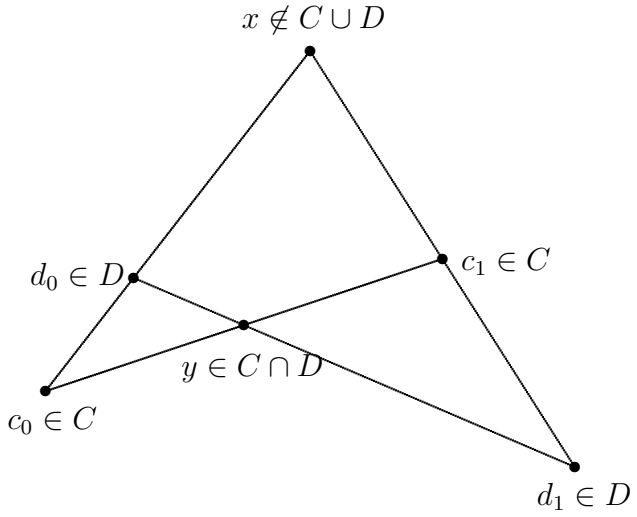

Abbildung 2: Beweis des Lemmas von Stone

beiden Geradenstücke $[c_0, c_1] \subset C$ und $[d_0, d_1] \subset D$ einen Schnittpunkt y besitzen. Dieser muss sowohl in C als auch in D liegen, was einen Widerspruch zu $C \cap D = \emptyset$ bedeutet. Der *analytische* Nachweis ist nicht schwierig. Nach einfacher Rechnung erhält man als Schnittpunkt

$$y = (1-\lambda)c_0 + \lambda c_1 = (1-\mu)d_0 + \mu d_1 \in C \cap D,$$

wobei

$$\lambda := \frac{\lambda_0}{\lambda_0 + \lambda_1(1-\lambda_0)}, \qquad \mu := \frac{\lambda_0(1-\lambda_1}{\lambda_1 + \lambda_0(1-\lambda_1)}.$$

Damit ist das Lemma von Stone bewiesen. □

5.4 Trennungssätze

In diesem Unterabschnitt formulieren und beweisen wir die wichtigsten Trennungssätze für konvexe Mengen in linearen normierten Räumen. Der folgende Trennungssatz stammt von M. Eidelheit (1936), siehe z. B. auch D. G. Luenberger (1969, S. 133).

Satz 5.5 (Eidelheit) *Sei E ein linearer normierter Raum und seien $A, B \subset E$ nichtleere, konvexe Teilmengen mit* int $(A) \neq \emptyset$ *sowie* int $(A) \cap B = \emptyset$. *Dann existiert ein Paar $(l, \gamma) \in (E^* \setminus \{0\}) \times \mathbb{R}$ bzw. eine abgeschlossene Hyperebene*

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

mit (abgeschlossenen) Halbräumen

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

derart, dass

1. $l(a) \leq \gamma \leq l(b)$ *für alle* $a \in A$, $b \in B$ *bzw.* $A \subset H^-$, $B \subset H^+$,
2. $l(a) < \gamma$ *für alle* $a \in$ int (A) *bzw.* int $(A) \subset$ int (H^-).

Beweis: Trotz der bisherigen Vorarbeiten ist der Beweis des Trennungssatzes von Eidelheit nicht ganz einfach. Zunächst wenden wir Lemma 5.4, das Lemma von Stone an. Hiernach existieren konvexe Mengen $C, D \subset E$ mit

$$C \cap D = \emptyset, \qquad C \cup D = E, \qquad \text{int } (A) \subset C, \qquad B \subset D.$$

Wegen int $(A) \neq \emptyset$ ist auch int $(C) \neq \emptyset$. Ferner ist

$$\text{cl } (D) = E \setminus \text{int } (C), \qquad \text{cl } (C) = E \setminus \text{int } (D).$$

Denn

$$\begin{aligned} \text{cl } (D) &= \{x \in E : B[x; \epsilon] \cap D \neq \emptyset \text{ für alle } \epsilon > 0\} \\ &= \{x \in E : B[x; \epsilon] \not\subset C \text{ für alle } \epsilon > 0\} \\ &= E \setminus \text{int } (C). \end{aligned}$$

Entsprechend ist cl $(C) = E \setminus$ int (D), wobei wir allerdings bisher nicht wissen, ob auch int $(D) \neq \emptyset$. Nun setze man

$$H := \text{cl } (C) \cap \text{cl } (D).$$

Im Rest des Beweises zeigen wir, dass H die gesuchte abgeschlossene Hyperebene ist. Klar ist, dass H als Durchschnitt von zwei abgeschlossenen, konvexen Mengen selbst abgeschlossen und konvex ist. Wegen int $(C) \neq \emptyset$ und cl $(D) = E \setminus$ int (C) ist H ferner eine *echte* Teilmenge von E. Die weiteren Beweisschritte sind gegeben durch

(a) H ist affiner Teilraum von E,

(b) H ist eine Hyperebene,

(c) H ist die gesuchte abgeschlossene Hyperebene, cl (C) der A enthaltende nichtpositive Halbraum und cl (D) der B enthaltende nichtnegative Halbraum.

Zum Beweis von (a) geben wir uns $x, y \in H$ und $\lambda \in \mathbb{R}$ vor, setzen $z := (1-\lambda)x + \lambda y$ und zeigen $z \in H$. Angenommen, dies sei nicht der Fall. Da H konvex ist, wäre notwendig $\lambda \notin [0,1]$, etwa $\lambda > 1$ (andernfalls vertausche man x und y). Da $z \notin H$, ist $z \in \text{int}\,(C)$ oder $z \in \text{int}\,(D)$. Aus dem ersten Teil von Satz 5.3 erhalten wir wegen $x \in \text{cl}\,(C) \cap \text{cl}\,(D)$, dass

$$z \in \text{int}\,(C) \Longrightarrow y = \frac{1}{\lambda} z + \frac{\lambda - 1}{\lambda} x \in \text{int}\,(C)$$

bzw.

$$z \in \text{int}\,(D) \Longrightarrow y = \frac{1}{\lambda} z + \frac{\lambda - 1}{\lambda} x \in \text{int}\,(D).$$

Dies ist jeweils ein Widerspruch zu

$$y \in H = \text{cl}\,(C) \cap \text{cl}\,(D) = (E \setminus \text{int}\,(D)) \cap (E \setminus \text{int}\,(C)).$$

Also ist H ein (echter, abgeschlossener) affiner Teilraum von E.

Um (b) zu beweisen, haben wir zu zeigen, dass H ein *maximaler* (echter) affiner Teilraum von E ist. Sei $z \notin H$ beliebig. Wir wählen $x_0 \in H$ beliebig. Wegen $z = 2x_0 - (2x_0 - z)$ ist z ein Punkt auf der Geraden durch x_0 und $2x_0 - z$ und folglich $2x_0 - z \notin H$ bzw. $2x_0 - z \in \text{int}\,(C) \cup \text{int}\,(D)$. Wegen $z \notin H$ ist $z \in \text{int}\,(C) \cup \text{int}\,(D)$. Ist $z \in \text{int}\,(C)$, so ist $2x_0 - z \in \text{int}\,(D)$. Denn wäre $2x_0 - z \in \text{int}\,(C)$, so würde auch der Mittelpunkt x_0 der Strecke von z nach $2x_0 - z$ zu $\text{int}\,(C)$ gehören, im Widerspruch zu $x_0 \in H$. Insbesondere folgt, dass auch $\text{int}\,(D) \neq \emptyset$. Entsprechend folgt aus $z \in \text{int}\,(D)$, dass $2x_0 - z \in \text{int}\,(C)$. O. B. d. A. betrachten wir nur den ersten Fall, es sei also $z \in \text{int}\,(C)$ und $2x_0 - z \in \text{int}\,(D)$ bzw. $2x_0 - z \notin \text{cl}\,(C)$, siehe Abbildung 3. Nun kommen wir zum Beweis dafür, dass H ein maximaler affiner Teilraum von E

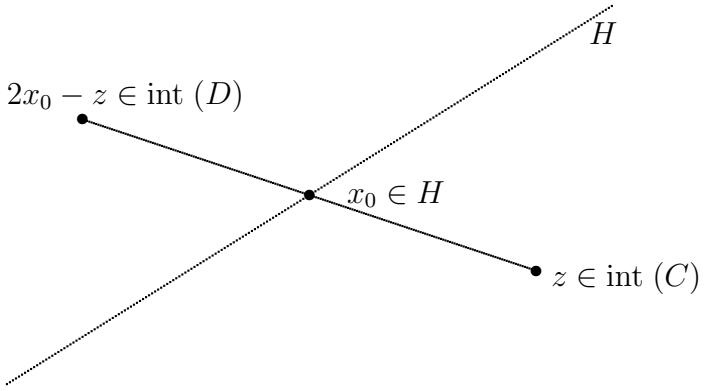

Abbildung 3: Erster Schritt für den Beweis, dass H Hyperebene ist

ist. Hierzu zeigen wir, dass ein affiner Teilraum V von E, der H und $z \notin H$ enthält, notwendig gleich E ist bzw. ein beliebiges $y \in E$ in V liegt. Da E die disjunkte Vereinigung der Mengen C und D ist, unterscheiden wir zwei Fälle. Im ersten Fall ist $y \in C$. Wir setzen

$$\lambda_1 := \min\{\lambda \in [0,1] : (1-\lambda)(2x_0 - z) + \lambda y \in \text{cl}\,(C)\},$$

wobei wir zu Recht min statt inf schreiben können, da die Menge aller λ aus $[0,1]$ mit $(1-\lambda)(2x_0 - z) + \lambda y \in \text{cl}\,(C)$ offensichtlich nichtleer und kompakt ist. Es ist $\lambda_1 > 0$, da $2x_0 - z \notin \text{cl}\,(C)$. Also ist

$$x_1 := (1-\lambda_1)(2x_0 - z) + \lambda_1 y \in \text{cl}\,(C).$$

Für $\lambda \in [0, \lambda_1)$ ist

$$(1-\lambda)(2x_0 - z) + \lambda y \notin \text{cl}\,(C) \quad \text{bzw.} \quad (1-\lambda)(2x_0 - z) + \lambda y \in \text{int}\,(D).$$

Folglich ist

$$x_1 \in \text{cl}\,(C) \cap \text{cl}\,(\text{int}\,(D)) = \text{cl}\,(C) \cap \text{cl}\,(D) = H.$$

Hierbei haben wir den dritten Teil von Satz 5.3 benutzt. Folglich ist

$$y = \frac{1-\lambda_1}{\lambda_1} z - 2\frac{1-\lambda_1}{\lambda_1} x_0 + \frac{1}{\lambda_1} x_1$$

eine affine Linearkombination (Koeffizienten addieren sich zu 1) der drei in V enthaltenden Punkte z, x_0 und x_1, also ist $y \in V$. In Abbildung 4 veranschaulichen wir uns diesen Beweisschritt. In dieser Abbildung haben wir auch noch den Punkt

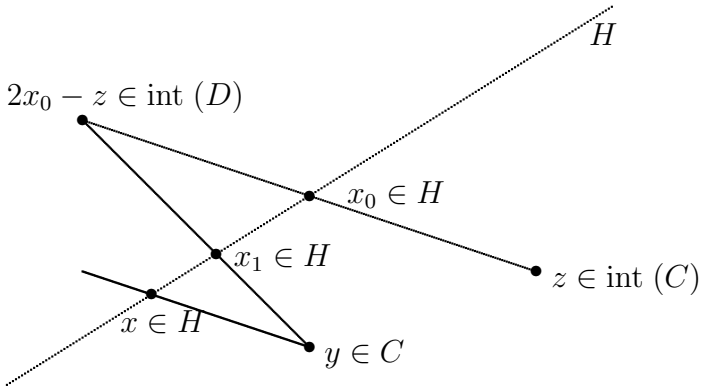

Abbildung 4: Zweiter Schritt ($y \in C$) für den Beweis, dass H Hyperebene ist

$$x := \frac{1}{\lambda_1} x_1 + \frac{\lambda_1 - 1}{\lambda_1} x_0 \in H$$

eingetragen. Mit $\alpha := (1-\lambda_1)/\lambda_1$ ist dann $y = \alpha(z - x_0) + x$. Der zweite Fall $y \notin C$ bzw. $y \in D$ kann entsprechend behandelt werden. Also ist H eine Hyperebene.

Im letzten Teil des Beweises zeigen wir, dass H die gesuchte abgeschlossene, trennende Hyperebene ist. Da H eine abgeschlossene Hyperebene ist, existiert nach Satz 5.2 ein Paar $(l, \gamma) \in (E^* \setminus \{0\}) \times \mathbb{R}$ mit

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

Sei $a_0 \in \text{int}\,(A)$ beliebig. Wegen $\text{int}\,(A) \subset C$ ist $a_0 \in \text{int}\,(C)$ und damit $a_0 \notin H$. O. B. A. ist $l(a_0) < \gamma$ (andernfalls multipliziere man l und γ mit -1). Wir zeigen zunächst, dass A in dem von H erzeugten nichtpositiven Halbraum liegt, also

$$A \subset H^- := \{x \in E : l(x) \leq \gamma\}$$

gilt. Hierzu sei $a \in A$ beliebig. Wäre $l(a) > \gamma$, so wäre

$$x := \Big(1 - \frac{\gamma - l(a_0)}{l(a) - l(a_0)}\Big)a_0 + \frac{\gamma - l(a_0)}{l(a) - l(a_0)}a$$

wegen $l(x) = \gamma$ einerseits ein Punkt aus H, andererseits wäre $x \in (a_0, a) \in \text{int}\,(A) \subset \text{int}\,(C)$, was ein Widerspruch zu $H \cap \text{int}\,(C) = \emptyset$ ist. Damit ist $A \subset H^-$ bewiesen. Jetzt folgt der Beweis für

$$B \subset H^+ := \{x \in E : l(x) \geq \gamma\}.$$

Sei $b \in B$ beliebig. Man definiere

$$\lambda_0 := \max\{\lambda \in [0,1] : (1-\lambda)b + \lambda a_0 \in \text{cl}\,(D)\}$$

und anschließend

$$x_0 := (1 - \lambda_0)b + \lambda a_0.$$

Die Menge aller λ aus $[0,1]$ mit $(1-\lambda)b + \lambda a_0 \in \text{cl}\,(D)$ ist kompakt, sodass wir bei der Definition von λ_0 zu Recht max statt sup geschrieben haben. Daher ist auch $x_0 \in \text{cl}\,(D)$. Wegen $a_0 \in \text{int}\,(C)$ bzw. $a_0 \notin \text{cl}\,(D)$ ist $\lambda_0 \in [0,1)$. Für $\lambda \in (\lambda_0, 1]$ ist $(1-\lambda)b + \lambda a_0 \notin \text{cl}\,(D)$ bzw. $(1-\lambda)b + \lambda a_0 \in \text{int}\,(C)$. Mit $\lambda \searrow \lambda_0$ folgt $x_0 \in \text{cl}\,(C)$. Insgesamt ist $x_0 \in \text{cl}\,(C) \cap \text{cl}\,(D) = H$. Folglich ist

$$\gamma = l(x_0) = (1 - \lambda_0)l(b) + \lambda_0 l(a_0)$$

und daher

$$l(b) - \gamma = \frac{\lambda_0}{1 - \lambda_0}[\gamma - l(a_0)] \geq 0$$

bzw. $B \subset H^+$. Zu zeigen bleibt schließlich noch, dass $l(a) < \gamma$ für alle $a \in \text{int}\,(A)$ bzw. $\text{int}\,(A) \subset \text{int}\,(H^-)$. Angenommen, es existiert ein $a_0 \in \text{int}\,(A)$ mit $l(a_0) = \gamma$. Wegen $a_0 \in \text{int}\,(A)$ existiert ein $\epsilon > 0$ mit $B[a_0; \epsilon] \subset A \subset H^-$. Für beliebiges $x \neq 0$ ist dann

$$a_0 \pm \epsilon \frac{x}{\|x\|} \in H^-$$

und daher

$$\gamma \geq l\Big(a_0 \pm \epsilon \frac{x}{\|x\|}\Big) = \gamma \pm \frac{\epsilon}{\|x\|} l(x),$$

also $l(x) = 0$ für alle $x \in E$. Dies ist ein Widerspruch zu $l \in E^* \setminus \{0\}$.

Damit ist der Trennungssatz von Eidelheit vollständig bewiesen. □

Der nun formulierte *starke Trennungssatz* stellt sich als eine einfache Folgerung aus dem Trennungssatz von Eidelheit heraus.

Satz 5.6 (Starker Trennungssatz) *Sei E ein linearer normierter Raum, $A \subset E$ nichtleer, konvex und abgeschlossen, ferner z ein Punkt in E außerhalb von A, also $z \notin A$. Dann können $\{z\}$ und A stark durch eine abgeschlossene Hyperebene in E getrennt werden, d. h. es existiert ein $l \in E^* \setminus \{0\}$ mit*

$$l(z) < \inf_{a \in A} l(a).$$

Beweis: Da A abgeschlossen ist und z kein Element von A ist, hat z zu A einen positiven Abstand d:

$$d := \inf_{a \in A} \|z - a\| > 0.$$

Nun definieren wir

$$A_0 := B\Big[z; \frac{d}{2}\Big], \qquad B_0 := A + B\Big[0; \frac{d}{2}\Big].$$

Dann sind A_0, $B_0 \subset E$ nichtleer und konvex, int $(A_0) \neq \emptyset$ und int $(A_0) \cap B_0 = \emptyset$. Aus dem Trennungssatz von Eidelheit folgt die Existenz eines Paares $(l, \gamma) \in (E^* \setminus \{0\}) \times \mathbb{R}$ bzw. einer abgeschlossenen Hyperebene $H := \{x \in E : l(x) = \gamma\}$ mit

$$B\Big[z; \frac{d}{2}\Big] \subset H^-, \qquad A + B\Big[0; \frac{d}{2}\Big] \subset H^+.$$

Für beliebige $a \in A$ ist daher

$$l(z) + \frac{d}{2}\|l\| \leq \gamma \leq l(a) - \frac{d}{2}\|l\|$$

und folglich

$$l(z) < l(z) + d\,\|l\| \leq \inf_{a \in A} l(a).$$

Damit ist der starke Trennungssatz bewiesen. □

5.5 Der Satz von Hahn-Banach

Es gibt viele Formulierungen des Satzes von Hahn-Bach. Wir wollen hier nur Formulierungen betrachten, bei denen der zugrunde liegende Raum ein linearer *normierter* Raum ist. Die folgende Aussage wird auch *geometrische Form des Satzes von Hahn-Banach* genannt, siehe z. B. D. G. Luenberger (1969, S. 133).

Satz 5.7 *Sei E ein linearer normierter Raum, $A \subset E$ konvex mit* int $(A) \neq \emptyset$. *Ist dann $V \subset E$ ein affiner Teilraum von E mit* int $(A) \cap V = \emptyset$, *so existiert eine abgeschlossene Hyperebene H in E mit $V \subset H$ und* int $(A) \cap H = \emptyset$ *bzw. ein Paar $(l, \gamma) \in (E^* \setminus \{0\}) \times \mathbb{R}$ mit $l(v) = \gamma$ und $l(a) < \gamma$ für alle $a \in$* int (A).

Beweis: Wegen des Satzes von Eidelheit existiert ein Paar $(l, \gamma_0) \in (E^* \setminus \{0\}) \times \mathbb{R}$ und hiermit die abgeschlossene Hyperebene $H_0 := \{x \in E : l(x) = \gamma_0\}$ mit int $(A) \subset$ int (H_0^-) bzw. $l(a) < \gamma_0$ für alle $a \in$ int (A) und $V \subset H_0^+$ bzw. $l(v) \geq \gamma_0$ für alle $v \in V$. Wir zeigen, dass l auf V konstant ist, also ein $\gamma \geq \gamma_0$ mit $l(v) = \gamma$ für alle $v \in V$ existiert. Dann ist offenbar (l, γ) das gesuchte Paar bzw. $H := \{x \in E : l(x) = \gamma\}$ die

gesuchte Hyperebene. Sei $w \in V$ beliebig, setze $\gamma := l(w)$. Da V ein affiner Teilraum von E ist, ist $(1-\lambda)w + \lambda v \in V$ für alle $\lambda \in \mathbb{R}$ und daher

$$\gamma_0 \leq l((1-\lambda)w + \lambda v) = \gamma + \lambda(l(v) - l(w))$$

für alle $\lambda \in \mathbb{R}$, was mit $\lambda \to +\infty$ für $l(v) < l(w)$ bzw. $\lambda \to -\infty$ für $l(v) > l(w)$ zum Widerspruch führt. Damit ist die geometrische Form des Satzes von Hahn-Banach bewiesen. □

Es folgt ein *Fortsetzungssatz von Hahn-Banach* für lineare normierte Räume.

Satz 5.8 *Sei E ein linearer normierter Raum und $L_0 \subset E$ ein linearer Teilraum. Die Abbildung $l_0 \colon L_0 \longrightarrow \mathbb{R}$ sei linear und stetig auf L_0, also*

$$\|l_0\| := \sup_{x \in L_0 \setminus \{0\}} \frac{|l(x)|}{\|x\|} < \infty.$$

Dann existiert ein $l \in E^*$ mit $l(x) = l_0(x)$ für alle $x \in L_0$ und $\|l\| = \|l_0\|$. Also kann $l_0 \in L_0^*$ zu einem $l \in E^*$ fortgesetzt werden, wobei die Norm erhalten bleibt.

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

$$A := \{(x,t) \in E \times \mathbb{R} : \|l_0\| \, \|x\| \leq t\}, \qquad B := \{(y, l_0(y)) \in E \times \mathbb{R} : y \in L_0\}.$$

Dann sind A und B nichtleer und konvex, B sogar ein linearer Teilraum von $E \times \mathbb{R}$. Der Beweis verläuft in drei Schritten. Im ersten Schritt zeigen wir, dass

$$\text{int } (A) = A_0 := \{(x,t) \in E \times \mathbb{R} : \|l_0\| \, \|x\| < t\} \neq \emptyset.$$

Ist $(x_0, t_0) \in \text{int } (A)$, so existiert ein $\epsilon > 0$ mit

$$(x_0, t_0) + B[0;\epsilon] \times [-\epsilon, \epsilon] \subset A,$$

insbesondere ist $(x_0, t_0 - \epsilon) \in A$ und daher $\|l_0\| \, \|x_0\| \leq t_0 - \epsilon < t_0$ bzw. $(x_0, t_0) \in A_0$. Damit ist int $(A) \subset A_0$ nachgewiesen. Zum Nachweis der umgekehrten Inklusionsbeziehung nehmen wir an, es sei $(x_0, t_0) \in A_0$. Wir definieren

$$\epsilon := \frac{t_0 - \|l_0\| \, \|x_0\|}{2} \min\Big(1, \frac{1}{\|l_0\|}\Big)$$

und zeigen, dass $(x_0, t_0) + B[0;\epsilon] \times [-\epsilon, \epsilon] \subset A$ bzw. $(x_0, t_0) \in \text{int } (A)$ und $A_0 \subset \text{int } (A)$. Mit $(x, s) \in B[0;\epsilon] \times [-\epsilon, \epsilon]$ folgt

$$\begin{aligned}
\|l_0\| \, \|x_0 + x\| &\leq \|l_0\| \, \|x_0\| + \|l_0\| \, \|x\| \\
&\leq \|l_0\| \, \|x_0\| + \frac{t_0 - \|l_0\| \, \|x_0\|}{2} \\
&= t_0 - \frac{t_0 - \|l_0\| \, ||x_0\|}{2} \\
&\leq t_0 - \epsilon \\
&\leq t_0 + s
\end{aligned}$$

und damit ist $(x_0, t_0) + (x, s) = (x_0 + x, t_0 + s) \in A$ und $A_0 \subset \text{int}\,(A)$. Da offensichtlich $A_0 \neq \emptyset$ ist der erste Beweisschritt abgeschlossen. Im zweiten Beweisschritt zeigen wir, dass $\text{int}\,(A) \cap B = \emptyset$. Denn gäbe es ein $(x, t) \in \text{int}\,(A) \cap B$, so wäre $x \in L$ und wegen des ersten Beweisteils $\|l_0\|\,\|x\| < t = l_0(x)$, was ein Widerspruch zur Definition von $\|l_0\|$ ist. Nun wenden wir im letzten Beweisteil den Trennungssatz von Eidelheit an. Wegen $(E \times \mathbb{R})^* = E^* \times \mathbb{R}$ existiert hiernach $((l, r), \gamma) \in ((E^* \times \mathbb{R}) \setminus \{(0, 0)\}) \times \mathbb{R}$ mit

$$l(x) + rt \leq \gamma \leq l(y) + rl_0(y) \qquad \text{für alle } (x, t) \in A,\ y \in L_0$$

bzw. eine abgeschlossene Hyperebene in $E \times \mathbb{R}$, die A und B trennt. Wegen $(0, 0) \in A$ und $0 \in L_0$ ist notwendig $\gamma = 0.$, d. h. die A und B trennende abgeschlossene Hyperebene geht durch bzw. enthält den Nullpunkt. Da L_0 ein linearer Teilraum ist, also insbesondere mit einem Punkt y auch $-y$ enthält, ist sogar

$$(*) \qquad l(x) + rt \leq 0 = l(y) + rl_0(y) \qquad \text{für alle } (x, t) \in A,\ y \in L_0.$$

Da $(0, 1) \in A$ ist $r \leq 0$. Wäre $r = 0$, so wäre $l(x) = 0$ für alle $x \in E$ bzw. $l = 0$, ein Widerspruch zu $(l, r) \neq (0, 0)$. Da man notfalls in $(*)$ durch $-r$ dividieren bzw. l durch $-l/r$ ersetzen kann, kann man o. B. d. A. $r = -1$ annehmen, sodass also

$$l(x) - t \leq 0 = l(y) - l_0(y) \qquad \text{für alle } (x, t) \in A,\ y \in L_0.$$

Hieraus liest man ab, dass l eine Fortsetzung von l_0 auf ganz E ist. Da $(x, \|l_0\|\,\|x\|) \in A$ für alle $x \in E$, ist ferner $l(x) - \|l_0\|\,\|x\| \leq 0$ und dann auch $|l(x)| \leq \|l_0\|\,\|x\|$ für alle $x \in E$. Hieraus folgt $\|l\| \leq \|l_0\|$. Da aber andererseits l eine Fortsetzung von l_0 ist, gilt $\|l\| \geq \|l_0\|$. Insgesamt ist also $\|l\| = \|l_0\|$ Damit ist der Fortsetzungssatz von Hahn-Banach in linearen normierten Räumen bewiesen. □

Im nächsten Satz geben wir weitere Folgerungen aus den Trennungssätzen an, die ebenfalls mit den Namen Hahn-Banach verbunden sind, diese findet man z. B. *hier*.

Satz 5.9 (Hahn-Banach) *Sei E ein linearer normierter Raum.*

1. *Sei $S^*[0; 1] := \{l \in E^* : \|l\| = 1\}$ der Rand der (abgeschlossenen) Einheitskugel in E^*. Dann ist*
$$\|x\| = \max_{l \in S^*[0;1]} l(x) \qquad \text{für alle } x \in E.$$
2. *Zu jedem $x \in E$ gibt es ein $l \in E^*$ mit $\|l\| = 1$ und $l(x) = \|x\|$.*
3. *Sei $U \subset E$ ein linearer Teilraum und $x \notin \text{cl}\,(U)$. Dann gibt es ein $l \in E^*$ mit $\|l\| = 1$ und $l(u) = 0$ für alle $u \in U$.*

Beweis: Beim Beweis der ersten Aussage können wir o. B. d. A. $x \neq 0$ annehmen. Dann besitzt

$$A := B[0; \|x\|] = \|x\|\, B[0; 1]$$

ein nichtleeres Inneres, welches x nicht enthält bzw. mit $\{x\}$ einen leeren Durchschnitt besitzt. Aus dem Trennungssatz 5.5 von Eidelheit folgt die Existenz einer A und x trennenden abgeschlossenen Hyperebene bzw. von $l^* \in E^* \setminus \{0\}$ mit

$$l^*(\|x\|\, z) \leq l^*(x) \qquad \text{für alle } z \in B[0; 1],$$

wobei wir (notfalls nach Division mit $\|l^*\|$) o. B. d. A. $\|l^*\| = 1$ bzw. $l^* \in S^*[0;1]$ annehmen können. Also ist

$$\|x\| = \sup_{z \in B[0;1]} l^*(\|x\|\, z) \leq l^*(x) \leq \sup_{l \in S^*[0;1]} l(x) \leq \|x\|,$$

womit die erste Behauptung bewiesen ist. Die zweite Aussage ist offenbar eine direkte Folgerung der ersten. Zum Beweis der dritten Aussage sei δ der Abstand von x zum linearen Teilraum U, es sei also

$$\delta := \inf_{u \in U} \|u - x\|.$$

Wegen $x \notin \mathrm{cl}\,(U)$ ist $\delta > 0$. Offensichtlich ist $\mathrm{int}\,(B[x;\delta]) \cap U = \emptyset$. Wegen Satz 5.5, dem Trennungssatz von Eidelheit, existiert ein $l \in E^* \setminus \{0\}$ mit

$$l(x) \leq l(u) \qquad \text{für alle } u \in U.$$

Wieder können wir o. B. d. A. annehmen, dass $\|l\| = 1$. Da U ein linearer Teilraum ist, folgt offenbar $l(u) = 0$ für alle $u \in U$. Damit ist auch die dritte Aussage des Satzes bewiesen. □

6 Der Satz von Lyusternik

Durch den Satz von Lyusternik aus dem Jahre 1934 wird unter bestimmten Voraussetzungen eine nichttriviale Teilmenge des Tangentialkegels an eine Menge M in einem Punkt $x^* \in M$ angegeben, siehe L. A. Ljusternik, W. I. Sobolew (1968, S. 342 ff.). In der Darstellung halten wir uns an J. Werner (1984, 1988), siehe auch J. Jahn (1994, S. 98 ff.) und J. Werner (2013, Abschnitt 50).

Definition 6.1 Sei X ein linearer normierter Raum, $M \subset X$ eine Teilmenge und $x^* \in M$. Dann heißt

$$T(M;x^*) := \left\{ p \in X : \begin{array}{l} \text{Es existieren Folgen } \{t_k\} \subset \mathbb{R}_+,\ \{r_k\} \subset X \text{ mit} \\ \text{i) } x^* + t_k p + r_k \in M \text{ für } k = 0, 1, \ldots, \\ \text{ii) } t_k \longrightarrow 0,\ r_k/t_k \longrightarrow 0 \end{array} \right\}$$

der *Tangentialkegel an M in x^**. Ein Element $p \in T(M;x^*)$ heißt *Tangentialrichtung an M in x^**.

Ein $p \in X$ ist also eine Tangentialrichtung an M in x^*, wenn eine Nullfolge $\{t_k\} \subset \mathbb{R}_+$ existiert mit der Eigenschaft, dass $x^* + t_k p$ für $k = 0, 1, \ldots$ nach einer "kleinen" Korrektur durch ein Element $r_k \in X$ in M liegt, also $x^* + t_k p + r_k \in M$ gilt. "Klein" bedeutet hierbei, dass $\{r_k/t_k\}$ eine Nullfolge ist, also gegen das Nullelement in X konvergiert. In dem folgenden Satz fassen wir einige einfache Eigenschaften des Tangentialkegels zusammen.

Satz 6.2 *Sei X ein linearer normierter Raum, $M \subset X$ und $x^* \in M$. Mit $T(M;x^*)$ werde der Tangentialkegel an M in x^* bezeichnet. Dann gilt:*

1. *Ist $x^* \in \text{int}\,(M)$, also x^* ein innerer Punkt von M, so ist $T(M;x^*) = X$. Jede Richtung ist also eine Tangentialrichtung an M in einem inneren Punkt von M.*

2. *$T(M;x^*)$ ist abgeschlossen.*

3. *Ist M konvex, so ist $T(M;x^*) = \text{cl}\,(\{\lambda(x-x^*) : \lambda \geq 0,\, x \in M\})$.*

Beweis: Ist $x^* \in \text{int}\,(M)$, so existiert zu jedem $p \in X$ ein $t^* > 0$ mit $x^* + tp \in M$ für alle $t \in [0, t^*]$. Hieraus folgt insbesondere $p \in T(M;x^*)$ und damit $T(M;x^*) = X$.
Um die Abgeschlossenheit von $T(M;x^*)$ zu zeigen, geben wir uns eine Folge $\{p^{(j)}\}_{j\in\mathbb{N}} \subset T(M;x^*)$ mit $\lim_{j\to\infty} p^{(j)} = p$ vor und zeigen, dass $p \in T(M;x^*)$. Nach Definition des Tangentialkegels existieren zu jedem $j \in \mathbb{N}$ Folgen $\{t_k^{(j)}\}_{k\in\mathbb{N}}$ und $\{r_k^{(j)}\}_{k\in\mathbb{N}} \subset X$ mit $x^* + t_k^{(j)} p^{(j)} + r_k^{(j)} \in M$ für alle k sowie $\lim_{k\to\infty} t_k^{(j)} = 0$ und $\lim_{k\to\infty} r_k^{(j)}/t_k^{(j)} = 0$. Zu jedem $j \in \mathbb{N}$ existiert ein $k(j) \in \mathbb{N}$ mit $0 < t_k^{(j)} \leq 1/j$ und $\|r_k^{(j)}\|/t_k^{(j)} \leq 1/j$ für alle $k \geq k(j)$. Nun definiere man dir Folgen $\{t_j\}_{j\in\mathbb{N}} \subset \mathbb{R}_+$ und $\{r_j\}_{j\in\mathbb{N}} \subset X$ durch

$$t_j := t_{k(j)}^{(j)}, \qquad r_j := r_{k(j)}^{(j)} + t_{k(j)}^{(j)}(p^{(j)} - p).$$

Dann ist $x^* + t_j p + r_j = t_{k(j)}^{(j)} p^{(j)} + r_{k(j)}^{(j)} \in M$ für alle $j \in \mathbb{N}$. Weiter ist

$$\lim_{j\to\infty} t_j = 0, \qquad \lim_{j\to\infty} \frac{r_j}{t_j} = 0$$

wegen

$$0 < t_j = t_{(k(j)}^{(j)} \leq \frac{1}{j}, \qquad \frac{\|r_j\|}{t_j} \leq \frac{\|r_{k(j)}^{(j)}\|}{t_{k(j)}^{(j)}} + \|p^{(j)} - p\| \leq \frac{1}{j} + \|p^{(j)} - p\|.$$

Insgesamt ist damit $p \in T(M;x^*)$, also die Abgeschlossenheit des Tangentialkegels $T(M;x^*)$ bewiesen.
Wir zeigen zunächst, dass

$$\{\lambda(x-x^*) : \lambda \geq 0,\, x \in M\} \subset T(M;x^*).$$

Hierzu geben wir uns $p = \lambda(x - x^*)$ mit $\lambda \geq 0$ und $x \in M$ vor. Dann ist

$$x^* + tp = x^* + \lambda t(x - x^*) \in M$$

für alle $t > 0$ mit $\lambda t \in [0, 1]$, also insbesondere alle hinreichend kleinen $t > 0$. Hieraus folgt $p \in T(M;x^*)$. Wegen der schon bewiesenen Abgeschlossenheit des Tangentialkegels ist

$$\text{cl}\,(\{\lambda(x-x^*) : \lambda \geq 0,\, x \in M\}) \subset T(M;x^*).$$

Umgekehrt sei nun $p \in T(M;x^*)$ eine Tangentialrichtung an M in x^*. Dann existieren Folgen $\{t_k\} \subset \mathbb{R}_+$ und $\{r_k\} \subset X$ mit $t_k \to 0$, $r_k/t_k \to 0$ und $x_k := x^* + t_k p + r_k \in M$. Dann ist

$$p_k := \frac{1}{t_k}(x_k - x^*) = p + \frac{r_k}{t_k} \in \{\lambda(x-x^*) : \lambda \geq 0,\, x \in M\}$$

und folglich

$$p = \lim_{k\to\infty} p_k \in \text{cl}\,(\{\lambda(x-x^*) : \lambda \geq 0,\, x \in M\}).$$

Damit ist auch die dritte Behauptung des Satzes bewiesen. □

Bemerkung 6.3 Um zu erläutern, weshalb der Tangentialkegel zur Gewinnung notwendiger Optimalitätsbedingungen von Bedeutung ist, betrachten wir die Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M$$

mit der *Zielfunktion* $f\colon X \longrightarrow \mathbb{R}$, einem linearen normierten Raum X und einer Teilmenge $M \subset X$. Sei nun $x^* \in M$ eine *lokale Lösung* von (P), es existiere also eine Kugel $B[x^*;\epsilon] := \{x \in X : \|x - x^*\| \leq \epsilon\}$ um x^* mit einem Radius $\epsilon > 0$ derart, dass $f(x^*) \leq f(x)$ für alle $x \in B[x^*;\epsilon] \cap M$. Ist $p \in T(M;x^*)$, also p eine Tangentialrichtung an M in x^*, so existieren nach Definition des Tangentialkegels eine Nullfolge $\{t_k\}$ und eine Folge $\{r_k\} \subset X$ derart, dass $\lim_{k\to\infty} r_k/t_k = 0$ und $x^* + t_k p + r_k \in M$. Für alle hinreichend großen k ist $x^* + t_k p + r_k \in B[x^*;\epsilon]$, da

$$\|t_k p + r_k\| = \underbrace{t_k}_{\to 0} \, \|p + \underbrace{r_k/t_k}_{\to 0}\| \to 0.$$

Also ist $x^* + t_k p + r_k \in B[x^*;\epsilon] \cap M$ und, da x^* eine lokale Lösung von (P) ist,

$$\frac{f(x^* + t_k p + r_k) - f(x^*)}{t_k} \geq 0$$

für alle hinreichend großen k. Unter einer geeigneten "Glattheitsbedingung" an f wird der Grenzwert

$$f'(x^*;p) := \lim_{k\to\infty} \frac{f(x^* + t_k p + r_k) - f(x^*)}{t_k}$$

für jedes $p \in X$ existieren und unabhängig von den Folgen $\{t_k\} \subset \mathbb{R}_+$ und $\{r_k\} \subset X$ sein. Als eine notwendige Bedingung dafür, dass $x^* \in M$ eine lokale Lösung von (P) ist, erhalten wir in diesem Falle also

$$f'(x^*;p) \geq 0 \qquad \text{für alle } p \in T(M;x^*).$$

Um diese Bedingung bei einer speziellen Menge M "auszuschlachten", ist es nützlich, eine möglichst große Teilmenge des Tangentialkegels $T(M;x^*)$ angeben zu können. Genau dies gelingt mit Hilfe des Satzes von Lyusternik.

6.1 Gâteaux- und Hadamard-Variation, Fréchet-Differential

Die wichtigsten Differenzierbarkeitsbegriffe für Abbildungen zwischen linearen normierten Räumen sind in der folgenden Definition zusammengefasst.

Definition 6.4 Seien X und Y lineare normierte Räume, $F\colon X \longrightarrow Y$ eine Abbildung und $x^* \in X$.

1. Eine (nicht notwendig lineare) Abbildung $F'(x^*;\cdot)\colon X \longrightarrow Y$ heißt

(a) *Gâteaux-Variation von F in x^**, wenn[15]

$$F'(x^*;p) = \lim_{t\to 0+} \frac{F(x^*+tp)-F(x^*)}{t}$$

für alle $p \in X$.

(b) *Hadamard-Variation von F in x^**, falls für alle $p \in X$ gilt:
Sind $\{t_k\} \subset \mathbb{R}_+$, $\{r_k\} \subset X$ Folgen mit $t_k \to 0$, $r_k/t_k \to 0$, so ist

$$F'(x^*;p) = \lim_{k\to\infty} \frac{F(x^*+t_kp+r_k)-F(x^*)}{t_k}.$$

2. Eine lineare[16], stetige Abbildung $F'(x^*)\colon X \longrightarrow Y$ heißt *Gâteaux-Differential von F in x^** , falls

$$F'(x^*)p = \lim_{t\to 0} \frac{F(x^*+tp)-F(x^*)}{t}$$

für alle $p \in X$. Existiert das Gâteaux-Differential von F in x^*, so heißt F in x^* *Gâteaux-differenzierbar*.

3. Eine lineare, stetige Abbildung $F'(x^*)\colon X \longrightarrow Y$ heißt *Fréchet-Differential von F in x^**, falls

$$\lim_{p\to 0} \frac{F(x^*+p)-F(x^*)-F'(x^*)p}{\|p\|} = 0.$$

Existiert das Fréchet-Differential von F in x^*, so heißt F in x^* *Fréchet-differenzierbar*.

4. F heißt in x^* *stetig Fréchet-differenzierbar*, wenn es eine Kugel B um x^* gibt mit:

(a) F ist in jedem Punkt $x \in B$ Fréchet-differenzierbar.

(b) Zu jedem $\epsilon > 0$ existiert ein $\delta > 0$ mit

$$x \in B, \quad \|x-x^*\| \le \delta \Longrightarrow \|F'(x)-F'(x^*)\| \le \epsilon,$$

d. h. die Abbildung $x \mapsto F'(x)$ von $B \subset X$ in den linearen normierten Raum $L(X,Y)$ der linearen stetigen Abbildungen von X nach Y ist stetig.

Bemerkung 6.5 Das Fréchet-Differential ist durch eine Eigenschaft charakterisiert. Man sollte sich also überlegen, dass eine Abbildung F zwischen linearen normierten Räumen X und Y nicht zwei verschiedene Fréchet-Differentiale besitzen kann. Denn sind $F_1'(x^*)$ und $F_2'(x^*)$ jeweils ein Fréchet-Differential von F in $x^* \in X$, so existiert zu einem vorgegebenen $\epsilon > 0$ ein $\delta > 0$ mit

$$\|p\| \le \delta \Longrightarrow \|F(x^*+p)-F(x^*)-F_i'(x^*)p\| \le \frac{\epsilon}{2}\|p\|, \quad i = 1,2.$$

[15]Man beachte, dass für eine Gâteaux-Variation nur die Existenz eines einseitigen Limes vorausgesetzt wird, ganz im Gegensatz zum Gâteaux-Differential.

[16]Bei einer linearen Abbildung T zwischen linearen (normierten) Räumen schreiben wir häufig (aber nicht immer) Tx statt $T(x)$. Daher ist $F'(x^*)p$ nur eine andere Schreibweise für $F'(x^*)(p)$.

Benutzt man die Linearität des Fréchet-Differentials und die Dreiecksungleichung, so erhält man

$$\|[F_1'(x^*) - F_2'(x^*)]p\| \leq \epsilon \, \|p\|,$$

zunächst für alle $p \in X$ mit $\|p\| \leq \delta$ und dann aus Homogenitätsgründen für alle $p \in X$. Dann ist aber auch

$$\|F_1'(x^*) - F_2'(x^*)\| = \sup_{p \neq 0} \frac{\|[F_1'(x^*) - F_2'(x^*)]p\|}{\|p\|} \leq \epsilon,$$

und mit $\epsilon \to 0+$ folgt $F_1'(x^*) = F_2'(x^*)$, also die behauptete Eindeutigkeit des Fréchet-Differentials. □

Klar ist, dass eine Hadamard-Variation oder ein Fréchet-Differential auch eine Gâteaux-Variation ist. Weiter ist klar, dass ein Fréchet-Differential auch ein Gâteaux-Differential ist. Durch den folgenden Satz werden nicht ganz offensichtliche Verbindungen zwischen den angegebenen Differenzierbarkeitsbegriffen spezifiziert.

Satz 6.6 *Seien X, Y lineare normierte Räume, $F\colon X \longrightarrow Y$ eine Abbildung, $x^* \in X$.*

1. *Ist $F'(x^*)$ Fréchet-Differential von F in x^*, so ist $F'(x^*)\colon X \longrightarrow Y$ auch Hadamard-Variation von F in x^*.*

2. *Ist $F'(x^*;\cdot)\colon X \longrightarrow Y$ Gâteaux-Variation von F in x^*, ist ferner F auf einer Kugel B um x^* lipschitzstetig, existiert also eine Konstante $C > 0$ mit*

$$\|F(x_1) - F(x_2)\| \leq C \, \|x_1 - x_2 \quad \text{für alle } x_1, x_2 \in B,$$

so ist $F'(x^*;\cdot)$ auch eine Hadamard-Variation von F in x^*.

Beweis: Sei $F'(x^*)$ Fréchet-Differential von F in x^*. Seien $p \in X$ und Folgen $\{t_k\} \subset \mathbb{R}_+$, $\{r_k\} \subset X$ mit $t_k \to 0$ und $r_k/t_k \to 0$ gegeben. Wegen $r_k/t_k \to 0$ existiert eine Konstante $C > 0$ mit $\|p + r_k/t_k\| \leq C$ für alle k. Sei $\epsilon > 0$ vorgegeben. Da $F'(x^*)$ Fréchet-Differential von F in x^* ist, existiert ein $\delta > 0$ derart, dass

$$\|F(x^* + q) - F(x^*) - F'(x^*)q\| \leq \frac{\epsilon}{2C} \, \|q\|$$

für alle $q \in X$ mit $\|q\| \leq \delta$. Wegen $t_k p + r_k \to 0$ bzw. $r_k/t_k \to 0$ ist $\|t_k p + r_k\| \leq \delta$ bzw. $\|F'(x^*)\| \, \|r_k\|/t_k \leq \epsilon/2$ für alle hinreichend großen k. Für diese k ist

$$\begin{aligned}
& \|F(x^* + t_k p + r_k) - F(x^*) - t_k F'(x^*)p\| \\
\leq \; & \|F(x^* + t_k p + r_k) - F(x^*) - F'(x^*)(t_k p + r_k)\| + \|F'(x^*) r_k\| \\
\leq \; & \frac{\epsilon}{2C} \underbrace{\|t_k p + r_k\|}_{\leq t_k C} + \underbrace{\|F'(x^*)\| \, \|r_k\|}_{\leq t_k \epsilon/2} \\
\leq \; & \epsilon t_k.
\end{aligned}$$

Für alle hinreichend großen k ist daher

$$\left\| \frac{F(x^* + t_k p + r_k) - F(x^*)}{t_k} - F'(x^*)p \right\| \leq \epsilon,$$

womit

$$F'(x^*)p = \lim_{k\to\infty} \frac{F(x^* + t_k p + r_k) - F(x^*)}{t_k}$$

und damit die erste Behauptung des Satzes bewiesen ist.

Sei jetzt $F'(x^*;\cdot)\colon X \longrightarrow Y$ Gâteaux-Variation von F in x^* und F auf einer Kugel B um x^* lipschitzstetig mit einer Lipschitzkonstanten $C > 0$. Um nachzuweisen, dass $F'(x^*;\cdot)$ auch Hadamard-Variation ist, geben wir uns $p \in X$ sowie Folgen $\{t_k\} \subset \mathbb{R}_+$, $\{r_k\} \subset X$ mit $t_k \to 0+$ und $r_k/t_k \to 0$ vor. Für alle hinreichend großen k ist $x^* + t_k p + r_k$, $x^* + t_k p \in B$ und daher folgt wegen

$$\begin{aligned}
& \left\| \frac{F(x^* + t_k p + r_k) - F(x^*)}{t_k} - F'(x^*;p) \right\| \\
\leq\ & \underbrace{\left\| \frac{F(x^* + t_k p) - F(x^*)}{t_k} - F'(x^*;p) \right\|}_{\to 0} + \underbrace{\frac{1}{t_k} \left\| F(x^* + t_k p + r_k) - F(x^* + t_k p) \right\|}_{\leq C\,\|r_k\|/t_k} \\
\to\ & 0
\end{aligned}$$

die Behauptung. □

Beispiel 6.7 Sei $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ eine Abbildung, die in $x^* \in \mathbb{R}^n$ stetig partiell differenzierbar ist, d. h. es existiert eine Kugel B um x^*, auf der die partiellen Ableitungen $\partial F_i/\partial x_j$ für $i = 1, \ldots, m$ und $j = 1, \ldots, n$ existieren und in x^* stetig sind. Hierbei sei

$$F(x) = (F_1(x_1, \ldots, x_n), \ldots, F_m(x_1, \ldots, x_n))^T.$$

Dann kann leicht nachgewiesen werden, dass F in x^* Fréchet-differenzierbar ist und das Fréchet-Differential $F'(x^*) \in L(\mathbb{R}^n, \mathbb{R}^m) = R^{m\times n}$ durch

$$F'(x^*) = \Big(\frac{\partial F_i}{\partial x_j}(x)\Big)_{\substack{i=1,\ldots,m \\ j=1,\ldots,n}}$$

gegeben ist, d. h. das Fréchet-Differential von F in x^* kann mit der *Funktionalmatrix* von F in x^* identifiziert werden. □

Beispiel 6.8 Sei X ein linearer normierter Raum und $F\colon X \longrightarrow \mathbb{R}$ definiert durch $F(x) := \|x\|$. Wir wollen uns überlegen, dass F in jedem x^* eine konvexe Gâteaux-Variation $F'(x^*;\cdot)\colon X \longrightarrow \mathbb{R}$ besitzt. Diese ist sogar eine Hadamard-Variation.

Für den ersten Teil wird lediglich ausgenutzt, dass $F\colon X \longrightarrow \mathbb{R}$ eine konvexe Funktion ist, also die Implikation

$$x, y \in X, \quad \lambda \in [0,1] \Longrightarrow F((1-\lambda)x + \lambda y) \leq (1-\lambda)F(x) + \lambda F(y)$$

gilt. Seien $x^* \in X$ und $p \in X$ beliebig vorgegeben. Wir definieren $\phi\colon (0,1] \longrightarrow \mathbb{R}$ durch

$$\phi(t) := \frac{F(x^* + tp) - F(x^*)}{t}.$$

Zum Nachweis der Existenz von $\lim_{t\to 0+}\phi(t)$ zeigen wir, dass ϕ auf $(0,1]$ nach unten beschränkt und monoton nicht fallend ist. Für $t\in(0,1]$ ist

$$F(x^*)=F\Big(\frac{1}{1+t}(x^*+tp)+\frac{t}{1+t}(x^*-p)\Big)\le\frac{1}{1+t}F(x^*+tp)+\frac{t}{1+t}F(x^*-p)$$

wegen der Konvexität von F. Daher ist

$$F(x^*)-F(x^*-p)\le\phi(t),\quad t\in(0,1],$$

also ϕ auf $(0,1]$ nach unten beschränkt. Zum Nachweis der Monotonie von ϕ sei $0<s\le t\le 1$. Dann ist

$$F(x^*+sp)-F(x^*)=F\Big(\frac{s}{t}(x^*+tp)+\frac{t-s}{t}x^*\Big)-F(x^*)\le\frac{s}{t}[F(x^*+tp)-F(x^*)]$$

wegen der Konvexität von F. daher ist

$$\phi(s)\le\phi(t)\le\phi(1)=F(x^*+p)-F(x^*),\qquad 0<s\le t\le 1.$$

Insgesamt ist die Existenz der Gâteaux-Variation $F'(x^*;\cdot)\colon X\longrightarrow\mathbb{R}$ nachgewiesen. Die Konvexität von $F'(x^*;\cdot)$ erkennt man leicht, indem man $F'(x^*;\alpha p)=\alpha F'(x^*;p)$ für alle $\alpha\ge 0$, $p\in X$ und $F'(x^*;p+q)\le F'(x^*;p)+F'(x^*;q)$ für alle $p,q\in X$ nachweist. Offensichtlich haben wir hier nur die Konvexität von F auf X ausgenutzt.

Dass die Gâteaux-Variation $F'(x^*;\cdot)$ von F in x^* sogar eine Hadamard-Variation ist, folgt aus Satz 6.6, da F wegen

$$|F(x)-F(y)|=|\|x\|-\|y\||\le\|x-y\|,\quad x,\,y\in X,$$

auf ganz X lipschitzstetig ist.

Sei speziell $B\subset\mathbb{R}^N$ kompakt, $X:=C(B)$ der lineare Raum der auf B stetigen reellwertigen Funktionen und

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

Mit dieser (Maximum-)Norm ist $C(B)$ ein linearer normierter Raum. Wir wissen daher, dass $F(x):=\|x\|_\infty$ in jedem $x^*\in X$ eine Gâteaux-Variation $F'(x^*;\cdot)$ besitzt und dass diese sogar eine Hadamard-Variation ist. Sie ist gegeben durch

$$(*)\qquad F'(x^*;p)=\begin{cases}\max\limits_{t\in B(x^*)}(\operatorname{sign}x^*(t))\,p(t), & x^*\ne 0,\\ \|p\|_\infty, & x^*=0,\end{cases}$$

wobei

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

Offenbar ist $B(x^*)$, die Menge derjenigen Punkte in B, in denen $|x^*(\cdot)|$ sein Maximum annimmt, als abgeschlossene Teilmenge einer kompakten Menge selbst kompakt. Beim Nachweis von $(*)$ können wir $x^*\ne 0$ annehmen. Der Beweis von $(*)$ zerfällt in zwei Teile

und benutzt die Tatsache, dass die Gâteaux-Variation $F'(x^*;\cdot)$ von F in x^* existiert. Sei zunächst $t \in B(x^*)$. Dann ist

$$\begin{aligned}\frac{F(x^*+sp)-F(x^*)}{s} &= \frac{\|x^*+sp\|_\infty - \|x^*\|}{s}\\ &\geq \frac{|x^*(t)+sp(t)|-|x^*(t)|}{s}\\ &\geq (\operatorname{sign} x^*(t))\, p(t)\end{aligned}$$

für alle hinreichend kleinen $s > 0$. Mit $s \to 0+$ folgt $F'(x^*;p) \geq (\operatorname{sign} x^*(t))\, p(t)$ und, da $t \in B(x^*)$ beliebig ist,

$$F'(x^*;p) \geq \max_{t\in B(x^*)} (\operatorname{sign} x^*(t))\, p(t).$$

Im zweiten Teil des Beweises von $(*)$ zeigen wir die andere Ungleichung, also $F'(x^*;p) \leq \max_{t\in B(x^*)}(\operatorname{sign} x^*(t))\, p(t)$. Hierzu sei $\{s_k\} \subset \mathbb{R}_+$ eine beliebige Nullfolge. Nach Definition der Maximum-Norm $\|\cdot\|_\infty$ existiert eine Folge $\{t_k\} \subset B$ mit

$$(**) \qquad \|x^*+s_kp\|_\infty = |x^*(t_k)+s_kp(t_k)|, \quad k = 1,2,\ldots.$$

Da B kompakt ist, besitzt $\{t_k\} \subset B$ eine gegen ein $t \in B$ konvergente Teilfolge. Da man notfalls zu dieser Teilfolge und der entsprechenden Teilfolge von $\{s_k\}$ übergehen kann, ist o. B. d. A. $\{t_k\}$ selbst konvergent gegen ein $t \in B$. Indem man in $(**)$ den Grenzübergang $k \to \infty$ macht, erhält man $\|x^*\|_\infty = |x^*(t)|$ bzw. $t \in B(x^*)$. Für alle hinreichend großen k ist

$$\operatorname{sign}(x^*(t_k)+s_kp(t_k)) = \operatorname{sign} x^*(t_k) = \operatorname{sign} x^*(t).$$

Daher ist

$$\begin{aligned}\frac{F(x^*+s_kp)-F(x^*)}{s_k} &= \frac{\|x^*+s_kp\|_\infty - \|x^*\|_\infty}{s_k}\\ &\leq \frac{|x^*(t_k)+s_kp(t_k)|-|x^*(t_k)|}{s_k}\\ &= (\operatorname{sign} x^*(t))\, p(t_k)\end{aligned}$$

für alle hinreichend großen k. Mit $k \to \infty$ folgt

$$F'(x^*;p) \leq (\operatorname{sign} x^*(t))\, pt(t) \leq \max_{t\in B(x^*)} (\operatorname{sign} x^*(t))\, p(t).$$

Damit ist $(*)$ nachgewiesen. □

Jetzt formulieren wir noch eine Kettenregel für die Hadamard-Variation bzw. das Fréchet-Differential.

Satz 6.9 *Seien X, Y, Z lineare normierte Räume, $x^* \in X$ und $G\colon X \longrightarrow Y$ sowie $H\colon Y \longrightarrow Z$ Abbildungen. Ferner sei $F := H \circ G\colon X \longrightarrow Z$ die zusammengesetzte Abbildung.*

1. *Besitzen die Abbildungen G bzw. H in x^* bzw. $G(x^*)$ die Hadamard-Variation $G'(x^*;\cdot)$ bzw. $H'(G(x^*),\cdot)$, so besitzt F die durch*

$$F'(x^*;\cdot) = H'(G(x^*); G'(x^*;\cdot))$$

gegebene Hadamard-Variation.

2. *Ist G bzw. H in x^* bzw. $G(x^*)$ Fréchet-differenzierbar mit Fréchet-Differential $G'(x^*)$ bzw. $H'(G(x^*)$, so ist auch F in x^* Fréchet-differenzierbar und besitzt das Fréchet-Differential*

$$F'(x^*) = H'(G(x^*))G'(x^*).$$

Beweis: Zum Beweis des ersten Teils des Satzes seien $p \in X$ sowie Folgen $\{t_k\} \subset \mathbb{R}_+$, $\{r_k\} \subset X$ mit $t_k \to 0$ und $r_k/t_k \to 0$ vorgegeben. Dann ist

$$\begin{aligned} & \frac{F(x^* + t_k p + r_k) - F(x^*)}{t_k} - H'(G(x^*); G'(x^*; p)) \\ = \; & \frac{H(G(x^*) + t_k G'(x^*; p) + q_k) - H(G(x^*))}{t_k} - H'(G(x^*); G'(x^*; p)) \end{aligned}$$

mit

$$q_k := G(x^* + t_k p + r_k) - G(x^*) - t_k G'(x^*; p).$$

Da $G'(x^*;\cdot)$ Hadamard-Variation von G in x^* ist, ist $q_k/t_k \to 0$. Da ferner $H'(G(x^*);\cdot)$ Hadamard-Variation von H in $G(x^*)$ ist, folgt die erste Behauptung.

Zum Beweis des zweiten Teiles des Satzes bemerken wir, dass $H'(G(x^*))G'(x^*)\colon X \longrightarrow Z$ linear und stetig ist. Daher bleibt zu zeigen, dass

$$(*) \qquad \lim_{p \to 0} \frac{H(G(x^* + p)) - H(G(x^*)) - H'(G(x^*))G'(x^*)p}{\|p\|} = 0.$$

Mit

$$\psi_G(p) := \frac{G(x^* + p) - G(x^*) - G'(x^*)p}{\|p\|}$$

und

$$\psi_H(q) := \frac{H(G(x^*) + q) - H(G(x^*)) - H'(G(x^*))q}{\|q\|}$$

gilt

$$\lim_{p \to 0} \psi_G(p) = 0, \qquad \lim_{q \to 0} \psi_H(q) = 0,$$

da G in x^* und H in $G(x^*)$ Fréchet-differenzierbar sind. Setzt man $q := G(x^*+p) - G(x^*)$ in die Definition von $\psi_H(q)$ ein, so erhält man

$$\begin{aligned} & H(G(x^* + p)) - H(G(x^*)) - H'(G(x^*))G'(x^*)p \\ = \; & H'(G(x^*))[G(x^* + p) - G(x^*) - G'(x^*)p \\ & \quad + \|G(x^* + p) - G(x^*)\| \, \psi_H(G(x^* p) - G(x^*))] \\ = \; & H'(G(x^*))[\|p\| \, \psi_G(p) + \|\|p\| \, \psi_G(p) + G'(x^*)p\| \, \psi_H(G(x^* + p) - G(x^*))]. \end{aligned}$$

Folglich ist

$$\begin{aligned} & \frac{\|H(G(x^*+p)) - H(G(x^*)) - H'(G(x^*))G'(x^*)p\|}{\|p\|} \\ = \; & \frac{\|H'(G(x^*))[\|p\|\,\psi_G(p) + \|\|p\|\,\psi_G(p) + G'(x^*)p\|\,\psi_H(G(x^*+p) - G(x^*))]\|}{\|p\|} \\ \leq \; & \|H'(G(x^*))\|\,[\|\psi_G(p)\| + (\|\psi_G(p)\| + \|G'(x^*)\|)\|\psi_H(G(x^*+p) - G(x^*)\|]. \end{aligned}$$

Wegen

$$\lim_{p\to 0} \psi_G(p) = 0, \qquad \lim_{p\to 0} \|\psi_H(G(x^*+p) - G(x^*))\| = 0$$

folgt $(*)$ und damit die Behauptung. Hierbei haben wir am Schluss noch ausgenutzt, dass das Fréchet-Differential $G'(x^*)$ stetig bzw. beschränkt ist und damit $\|G'(x^*)p\| \leq \|G'(x^*)\|\,\|p\|$ gilt und außerdem $\lim_{p\to 0}(G(x^*+p)-G(x^*)) = 0$ gilt. Dies wiederum folgt aus der Fréchet-Differenzierbarkeit von G in x^*, denn diese impliziert die Stetigkeit von G in x^*. Damit ist der Satz bewiesen. □

Beispiel 6.10 Die Abbildungen $f_i\colon \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, \ldots, m$, seien in $x^* \in \mathbb{R}^n$ stetig partiell differenzierbar. Hiermit sei die Abbildung $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ durch

$$F(x) := \max_{i=1,\ldots,m} |f_i(x)|$$

definiert. Dann besitzt F in x^* eine Hadamard-Variation, welche durch

$$F'(x^*;p) = \begin{cases} \max\limits_{i\in I(x^*)} (\operatorname{sign} f_i(x^*))\nabla f_i(x^*)^T p, & F(x^*) > 0, \\ \max\limits_{i=1,\ldots,m} |\nabla f_i(x^*)^T p|, & F(x^*) = 0, \end{cases}$$

gegeben ist. Dies folgt aus der Kettenregel für die Hadamard-Variation, wenn man berücksichtigt, dass $F = H \circ G$, wobei $G\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ durch

$$G(x) := (f_1(x), \ldots, f_m(x))^T$$

und $H\colon \mathbb{R}^m \longrightarrow \mathbb{R}$ durch

$$H(y) := \max_{i=1,\ldots,m} |y_i|$$

definiert sind. □

Letztes Ergebnis in diesem Unterabschnitt wird der folgende *Mittelwertsatz* sein.

Satz 6.11 *Seien X und Y lineare normierte Räume, $x, p \in X$ und die Abbildung $F\colon X \longrightarrow Y$ in jedem Punkt von $[x, x+p] := \{x+tp : t \in [0,1]\}$ Gâteaux-differenzierbar. Dann gilt:*

1. *Es ist*

$$\|F(x+p) - F(x)\| \leq \sup_{t\in(0,1)} \|F'(x+tp)\|\,\|p\|.$$

2. *Ist $T \in L(X,Y)$ bzw. $T\colon X \longrightarrow Y$ linear und stetig, so ist*

$$\|F(x+p) - F(x) - Tp\| \leq \sup_{t\in(0,1)} \|F'(x+tp) - T\|\,\|p\|.$$

Insbesondere ist

$$\|F(x+p) - F(x) - F'(x)p\| \leq \sup_{t\in(0,1)} \|F'(x+tp) - F'(x)\|\,\|p\|.$$

Beweis: Der zweite Teil von Satz 5.9 sagt aus:

- Sei E ein linearer normierter Raum und $x \in E$. Dann existiert ein $l \in E^* := L(E,\mathbb{R})$ bzw. eine lineare, stetige Abbildung $l\colon E \longrightarrow \mathbb{R}$ mit $\|l\| = 1$ und $l(x) = \|x\|$.

Daher gibt es eine lineare, stetige Abbildung $l\colon Y \longrightarrow \mathbb{R}$ mit $\|l\| = 1$ und

$$l(F(x+p) - F(x)) = \|F(x+p) - F(x)\|.$$

Nun definiere man die Abbildung $\phi\colon [0,1] \longrightarrow \mathbb{R}$ durch $\phi(t) := l(F(x+tp))$. Dann ist ϕ auf $[0,1]$ differenzierbar mit der Ableitung $\phi'(t) = l(F'(x+tp)p)$. Der Mittelwertsatz der Differentialrechnung liefert die Existenz eines $t_0 \in (0,1)$ mit $\phi(1) - \phi(0) = \phi'(t_0)$. Daher ist

$$\begin{aligned}
\|F(x+p) - F(x)\| &= l(F(x+p) - F(x)) \\
&= \phi(1) - \phi(0) \\
&= \phi'(t_0) \\
&= l(F'(x+t_0p)p) \\
&\leq \underbrace{\|l\|}_{=1}\, \|F'(x+t_0p)p\| \\
&\leq \|F'(x+t_0p)\|\,\|p\| \\
&\leq \sup_{t\in(0,1)} \|F'(x+tp)\|\,\|p\|,
\end{aligned}$$

und damit ist der erste Teil des Satzes bewiesen. Zum Beweis des zweiten Teiles wendet man den ersten Teil des Satzes auf $F - T$ an. □

6.2 Der Satz von Baire

Einer der klassischen Sätze der Funktionalanalysis ist der Satz von Baire, der für unsere Zwecke folgendermaßen formuliert wird:

Satz 6.12 (Baire) *Ist X ein Banachraum und $X = \bigcup_{k=1}^{\infty} A_k$ mit abgeschlossenen $A_k \subset X$, $k = 1,2,\ldots$, so ist* $\operatorname{int}(A_j) \neq \emptyset$ *für mindestens ein $j \in \mathbb{N}$, das Innere also mindestens einer der Mengen A_k nichtleer.*

Beweis: Mit $B[x;\epsilon]$ bezeichnen wir die abgeschlossene und mit $B(x;\epsilon)$ die offene Kugel um $x \in X$ mit dem Radius $\epsilon > 0$. Angenommen es sei int $(A_k) = \emptyset$, $k = 1, 2, \ldots$. Den gewünschten Widerspruch erhalten wir in vier Schritten.

Im ersten Schritt zeigen wir, dass $(X \setminus A_k) \cap B(x;\epsilon) \neq \emptyset$ für alle $x \in X$, $\epsilon > 0$ und $k \in \mathbb{N}$. Denn: O. B. d. A. ist $x \in A_k$ (andernfalls ist die Aussage trivial). Dann ist $B(x;\epsilon) \not\subset A_k$ (andernfalls ist $x \in$ int (A_k) im Widerspruch zur Annahme, dass int $(A_k) = \emptyset$). Dies impliziert aber, wie behauptet, $(X \setminus A_k) \cap B(x;\epsilon) \neq \emptyset$.

Im zweiten Schritt konstruieren wir nach der folgenden Vorschrift eine Folge $\{x_k\} \subset X$ und eine Folge $\{B_k\}$ abgeschlossener Kugeln $B_k = B[x_k;\epsilon_k]$:

- Wähle $x_0 \in X$ beliebig, setze $\epsilon_0 := 1$.
- Für $k = 0, 1, \ldots$:
  - Wähle $x_{k+1} \in (X \setminus A_{k+1}) \cap B(x_k;\epsilon_k)$.
    Dies ist wegen des ersten Beweisschrittes möglich.
  - Wähle $\epsilon_{k+1} \in (0, 1/2^{k+1}]$ mit $B[x_{k+1};\epsilon_{k+1}] \subset (X \setminus A_{k+1}) \cap B(x_k;\epsilon_k)$.
    Dies ist möglich, da $(X \setminus A_{k+1}) \cap B(x_k;\epsilon_k)$ offen ist und x_{k+1} enthält.
  - Setze $B_{k+1} := B[x_{k+1};\epsilon_{k+1}]$.

Nach Konstruktion ist offenbar $B_k \subset X \setminus A_k$ und $B_{k+1} \subset B_k$, $k = 1, 2, \ldots$.

Im dritten Teil des Beweises überlegen wir uns, dass $\{x_k\}$ eine Cauchyfolge (im Banachraum X) ist und ihr daher existierender Limes x in $\bigcap_{k=1}^{\infty} B_k$ liegt. Denn zunächst ist

$$x_{k+1} \in B_{k+1} \subset B_k = B[x_k;\epsilon_k] \subset B[x_k; 1/2^k], \quad k = 1, 2, \ldots,$$

und daher

$$\|x_{k+1} - x_k\| \leq \frac{1}{2^k}, \quad k = 1, 2, \ldots.$$

Hieraus folgt leicht, dass $\{x_k\} \subset X$ eine Cauchyfolge ist. Der Limes x der Folge $\{x_k\}$ liegt in allen B_k und damit in $\bigcap_{k=1}^{\infty} B_k$. Um dies einzusehen, sei $k \in \mathbb{N}$ beliebig vorgegeben. Da die Folge der Kugeln $\{B_l\}_{l \in \mathbb{N}}$ ineinander geschachtelt ist, ist $B_l \subset B_k$ und damit $x_l \in B_k$ für $l = k, k+1, \ldots$. Mit $l \to \infty$ folgt wegen der Konvergenz von $\{x_l\}$ gegen x und der Abgeschlossenheit von B_k, dass $x \in B_k$.

Jetzt kommt das Finale. Für den Limes x der im zweiten Teil des Beweises konstruierten Folge $\{x_k\}$ gilt

$$x \in \bigcap_{k=1}^{\infty} B_k \subset \bigcap_{k=1}^{\infty} (X \setminus A_k) = X \setminus \bigcup_{k=1}^{\infty} A_k = \emptyset,$$

ein Widerspruch. Die Annahme, die A_k hätten für alle $k \in \mathbb{N}$ ein leeres Inneres, ist damit widerlegt. Der Satz von Baire ist bewiesen. □

6.3 Ein Open Mapping Theorem

Das klassische Open Mapping Theorem der Funktionalanalysis sagt aus:

Satz 6.13 (Open Mapping Theorem) *Ist eine lineare stetige Abbildung T zwischen Banachräumen X und Y surjektiv, so ist T eine offene Abbildung, d. h. das Bild $T(O)$ einer offenen Menge $O \subset X$ ist offen.*

Das Open Mapping Theorem ist eine Folgerung der Aussage:

Satz 6.14 *Ist eine lineare stetige Abbildung T zwischen Banachräumen X und Y surjektiv, so existiert ein $\rho > 0$ mit $B[0;\rho] \subset T(B[0;1])$.*

Denn: Sei $O \subset X$ offen und $y \in T(O)$. Dann existiert ein $x \in O$ mit $y = Tx$. Da O offen ist, existiert ein $\epsilon > 0$ mit $B[x;\epsilon] \subset O$. Existiert ein $\rho > 0$ mit $B[0;\rho] \subset T(B[0;1])$, so ist

$$B[y;\epsilon\rho] = y + \epsilon\, B[0;\rho] = Tx + \epsilon\, B[0;\rho] \subset Tx + \epsilon\, T(B[0;1]) = T(B[x;\epsilon]) \subset T(O).$$

Also gibt es um jedes $y \in T(O)$ eine in $T(O)$ gelegene Kugel, d. h. $T(O)$ ist offen bzw. T eine offene Abbildung.

Wir werden in diesem Unterabschnitt eine auf J. Zowe, S. Kurcyusz (1979) zurückgehende Verallgemeinerung von Satz 6.14 und damit des klassischen Open Mapping Theorems 6.13 formulieren und beweisen. Vorher wollen wir aber zwei Folgerungen aus Satz 6.14 angeben. Die erste ist ein Satz über die Beschränktheit der Inversen einer linearen, stetigen und bijektiven Abbildung zwischen Banachräumen.

Satz 6.15 *Seien X und Y Banachräume, $A\colon X \longrightarrow Y$ sei linear, stetig und bijektiv. Dann existiert die inverse Abbildung $A^{-1}\colon Y \longrightarrow X$ und ist linear und stetig.*

Beweis: Natürlich existiert $A^{-1}\colon Y \longrightarrow X$ und ist linear. Zu zeigen bleibt die Stetigkeit bzw. Beschränktheit von A^{-1}, also dass

$$\|A^{-1}\| := \sup_{y \neq 0} \frac{\|A^{-1}y\|}{\|y\|} < \infty.$$

Wegen Satz 6.14 existiert ein $\rho > 0$ mit $B[0;\rho] \subset A(B[0;1])$. Sei $y \in Y \setminus \{0\}$ beliebig. Dann ist $(\rho/\|y\|)y \in B[0;\rho]$, folglich $A^{-1}((\rho/\|y\|)y) \in B[0;1]$ bzw. $\|A^{-1}y\|/\|y\| \leq 1/\rho$. Damit ist die Beschränktheit bzw. Stetigkeit von A^{-1} mit $\|A^{-1}\| \leq 1/\rho$ bewiesen. □

Die zweite Anwendung von Satz 6.14 ist eine Aussage, die der des sogenannten Hoffman-Lemmas (siehe A. J. Hoffman (1952) und z. B. O. Güler (2012, S. 299)) ähnelt. Dieses sagt aus:

Satz 6.16 (Hoffman) *Gegeben sei ein nichtleerer Polyeder $P := \{z \in \mathbb{R}^n : Az \leq b\}$, wobei $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$. Dann existiert eine allein von A abhängende positive Konstante $c = c(A)$ derart, dass*

$$\operatorname{dist}(x,P) := \inf_{z\in P} \|x - z\| \leq c\,\|(Ax-b)_+\| \qquad \text{für alle } x \in \mathbb{R}^n.$$

Für $y = (y_i) \in \mathbb{R}^m$ ist hierbei $y_+ \in \mathbb{R}^m$ definiert durch $y_+ := (\max(y_i, 0))$.

Beweis: Wegen der bekannten Äquivalenz der Normen im $\mathbb{R}^m$ bzw. $\mathbb{R}^n$ können wir annehmen, dass $\|\cdot\|$ die euklidische Norm im $\mathbb{R}^m$ bzw. $\mathbb{R}^n$ ist. Wir wollen einen Beweis bringen, der den Satz von Kuhn-Tucker (siehe z. B. Abschnitt 53 bei J. WERNER (2013)) benutzt. Für eine Indexmenge $I \subset \{1, \ldots, m\}$ seien $A_I \in \mathbb{R}^{|I| \times n}$ und $b_I \in \mathbb{R}^{|I|}$ in naheliegender Weise definiert, indem die zu I gehörenden Zeilen von A bzw. Komponenten von b aneinandergefügt werden. Wir beweisen zunächst die folgende Hilfsaussage:

- Sei $I \subset \{1, \ldots, m\}$. Hiermit definiere man

$$N_I := \{z \in \mathbb{R}^n : A_I z \geq 0_I\}, \qquad N_I^+ := \{y \in \mathbb{R}^n : y^T z \geq 0 \text{ für alle } z \in N_I\}.$$

  Dann existiert eine Konstante $\delta_I > 0$ mit

$$\delta_I \|y\| \leq \|(A_I y)_+\| \qquad \text{für alle } y \in N_I^+.$$

Um dies einzusehen, können wir zunächst annehmen, dass $N_I^+ \neq \{0\}$ (bzw. N_I echte Teilmenge des $\mathbb{R}^n$), da andernfalls die Aussage trivial ist. Man definiere

$$\delta_I := \min_{y \in N_I^+, \, \|y\|=1} \|(A_I y)_+\|.$$

Es ist $\delta_I > 0$, denn andernfalls existiert ein $y \neq 0$ mit $-y \in N_I$ und $y \in N_I^+$, was $-\|y\|^2 \geq 0$ implizieren und damit den Widerspruch $y = 0$ ergeben würde. Die angegebene Konstante δ_I tut offenbar das verlangte.

Bei gegebenem $x \in \mathbb{R}^n$ betrachte man die quadratische Optimierungsaufgabe

$$\text{Minimiere} \quad \tfrac{1}{2}\|z - x\|^2, \quad z \in P.$$

Die eindeutige Lösung $z(x) \in P$ ist die Projektion von x auf P und nach Kuhn-Tucker charakterisiert durch die Existenz von $u(x) \in \mathbb{R}^m$ mit

$$u(x) \geq 0, \qquad z(x) - z + A^T u(x) = 0, \qquad u(x)^T (Az(x) - b) = 0.$$

Mit $I(x) \subset \{1, \ldots, m\}$ werde die Indexmenge der in $z(x)$ aktiven (also voll ausgeschöpften bzw. als Gleichung erfüllten) Ungleichungsrestriktionen bezeichnet. Dann ist also

$$u_{I(x)} \geq 0, \qquad z(x) - x + A_{I(x)}^T u_{I(x)} = 0.$$

Um die obige Hilfsaussage benutzen zu können, überlegen wir uns, dass $x - z(x) \in N_{I(x)}^+$. Denn für ein beliebiges $z \in N_{I(x)}$ (also $A_{I(x)} z \geq 0$) ist

$$z^T (x - z(x)) = z^T A_{I(x)}^T u_{I(x)} = \underbrace{(A_{I(x)} z)}_{\geq 0}^T \underbrace{u_{I(x)}}_{\geq 0} \geq 0.$$

Mit obiger Hilfsaussage ist daher

$$\begin{aligned} \|(Ax - b)_+\| &\geq \|(A_{I(x)} z - b_{I(x)})_+\| \\ &= \|(A_{I(x)}(x - z(x)))_+\| \\ &\geq \delta_{I(x)} \|x - z(x)\| \\ &= \delta_{I(x)} \operatorname{dist}(x, P) \\ &\geq \delta \operatorname{dist}(x, P), \end{aligned}$$

wobei

$$\delta := \min_{I \subset \{1,\dots,m\}} \delta_I.$$

Mit $c := 1/\delta$ ist das Hoffman-Lemma bewiesen. □

Wir formulieren und beweisen jetzt eine Aussage, die der des Hoffman-Lemmas ähnelt:

Satz 6.17 *Seien X und Y Banachräume, $T\colon X \longrightarrow Y$ sei linear, stetig und surjektiv. Mit $y \in Y$ sei $M := \{z \in X : Tz = y\}$. Dann existiert eine alleine von T abhängende positive Konstante $c = c(T)$ mit*

$$d(x, M) := \inf_{z \in M} \|x - z\| \le c\,\|Tx - y\| \qquad \textit{für alle } x \in X.$$

Beweis: Wegen Satz 6.14 existiert ein $\rho > 0$ mit $B[0;\rho] \subset T(B[0;1])$. Zu einem beliebigen $x \in X$ (o. B. d. A. ist $x \notin M$ und daher $Tx - y \neq 0$) existiert daher ein $u \in B[0;1]$ mit

$$\rho \frac{Tx - y}{\|Tx - y\|} = Tu.$$

Folglich ist

$$z := x - \frac{\|Tx - y\|}{\rho} u \in M$$

und damit

$$\operatorname{dist}(x, M) \le \|x - z\| = \frac{1}{\rho}\,\|Tx - y\|\,\|u\| \le \frac{1}{\rho}\,\|Tx - y\|.$$

Mit $c := 1/\rho$ ist die Behauptung bewiesen. □

Nun kommen wir zu der angekündigten Verallgemeinerung von Satz 6.14 bzw. des Open Mapping Theorems 6.13, siehe J. Zowe, S. Kurcyusz (1979, Theorem 2.1). Vorher müssen wir aber noch einige einfache Bezeichnungen bzw. Definitionen voranschicken:

- Ist E ein reeller linearer Raum, A und B Teilmengen von E sowie $\alpha, \beta \in \mathbb{R}$, so sei

  $$\alpha A + \beta B := \{\alpha a + \beta b : a \in A,\, b \in B\}.$$

  Vereinfachungen dieser Schreibweise in Spezialfällen sind naheliegend. So schreibt man z. B. natürlich $A + \beta B$ statt $1A + \beta B$. Ist ferner $A = \{a\}$ einpunktig, so schreibt man $a + B$ statt $\{a\} + B$.

- Ist E ein linearer Raum und $K \subset E$, so heißt K ein *Kegel*, falls $\lambda x \in K$ für alle $x \in K$ und $\lambda \ge 0$.

- Ist E ein linearer Raum und $S \subset E$, so bezeichnet cone $(S) \subset E$ die *konvexe Kegelhülle* von S, d. h. den kleinsten konvexen Kegel, der S enthält bzw. den Durchschnitt aller konvexen Kegel, die S enthalten. Ist $S \subset E$ konvex, so ist

  $$\text{cone } (S) = \{\lambda x : x \in S,\, \lambda \ge 0\}.$$

  Denn einerseits muss die rechts stehende Menge nach Definition in cone (S) enthalten sein, andererseits enthält sie S und ist ein konvexer (Beweis?) Kegel.

Satz 6.18 (Zowe-Kurcyusz) *Seien X und Y Banachräume und $C \subset X$ sowie $K \subset Y$ nichtleer, konvex und abgeschlossen. Es seien $x^* \in C$, $y^* \in K$ und $T\colon X \longrightarrow Y$ linear und stetig. Ferner gelte*

$$\text{cone}\,(T(C - x^*) + (K - y^*) = Y.$$

Dann existiert ein $\rho > 0$ mit

$$B[0;\rho] \subset T((C - x^*) \cap B[0;1]) + (K - y^*) \cap B[0;1].$$

Beweis: Wir folgen sehr eng dem Beweis von J. Zowe, S. Kurcyusz (1979). Zur Abkürzung setzen wir

$$(C - x^*)_1 := (C - x^*) \cap B[0;1], \qquad (K - y^*)_1 := (K - y^*) \cap B[0;1].$$

Anschließend definieren wir

$$A_k := k\,[T((C - x^*)_1) + (K - y^*)_1], \qquad k = 1, 2, \ldots.$$

Der Beweis erfolgt in drei Schritten. Im ersten zeigen wir, dass $Y = \bigcup_{k=1}^{\infty} A_k$. Hierzu geben wir uns ein beliebiges $y \in Y$ und zeigen die Existenz eines $k \in \mathbb{N}$ mit $y \in A_k$. Wegen der Konvexität von $S := T(C - x^*) + (K - y^*)$ und $Y = \text{cone}\,(S)$ existieren $\lambda \geq 0$, $x \in C$ und $z \in K$ mit

$$y = \lambda[T(x - x^*) + (z - y^*)].$$

Nun wähle man $k \in \mathbb{N}$ so groß, dass $k \geq \lambda \max(1, \|x - x^*\|, \|z - y^*\|)$ und setze

$$\hat{x} := x^* + \frac{\lambda}{k}(x - x^*), \qquad \hat{z} := y^* + \frac{\lambda}{k}(z - y^*).$$

Wegen der Konvexität von C bzw. K ist $\hat{x} \in C$ bzw. $\hat{z} \in K$, ferner ist

$$y = k[T(\hat{x} - x^*) + (\hat{z} - y^*)] \in kA_1 = A_k.$$

Damit ist $Y = \bigcup_{k=1}^{\infty} A_k$ nachgewiesen.

Im zweiten Schritt bestimmen wir durch eine Anwendung des Satzes von Baire das gesuchte $\rho > 0$, von dem wir im dritten Schritt zeigen, dass es das Verlangte tut. Erst recht ist $Y = \bigcup_{k=1}^{\infty} \text{cl}(A_k)$. Nach Voraussetzung ist Y ein Banachraum. Aus Satz 6.12, dem Satz von Baire, folgt die Existenz eines $j \in \mathbb{N}$ mit $\text{int}\,(\text{cl}(A_j)) \neq \emptyset$. Sei $a \in \text{int}\,(\text{cl}\,(A_j))$ beliebig. Wegen $Y = \bigcup_{k=1}^{\infty} A_k$ gibt es ein $k \in \mathbb{N}$ mit $-a \in A_k$ bzw. $-(1/k)a \in A_1$ und $-(j/k)a \in A_j \subset \text{cl}(A_j)$. Mit A_j ist auch der Abschluss $\text{cl}(A_j)$ konvex. Der Nullpunkt von Y liegt "zwischen" $a \in \text{int}\,(\text{cl}\,(A_j))$ und $-(j/k)a \in \text{cl}(A_j)$ und daher[17] ist $0 \in \text{int}\,(\text{cl}\,(A_j))$ und auch $0 \in \text{int}\,(\text{cl}\,(A_1))$. Folglich existiert ein $\rho > 0$ mit $B[0;2\rho] \subset \text{cl}(A_1)$.

[17] Hierbei benutzen wir die folgende, leicht zu beweisende Aussage, siehe auch die erste Aussage in Satz 5.3:

- Sei E ein linearer normierter Raum und $K \subset E$ eine nichtleere, konvexe Teilmenge. Ist dann $a \in \text{int}\,(K)$ und $b \in K$, so ist

$$[a, b) := \{(1 - \lambda)a + \lambda b : \lambda \in [0, 1)\} \subset \text{int}\,(K).$$

Wir zeigen im dritten Schritt, dass dieses ρ das Verlangte tut. Es ist

$$B[0;\rho] = \frac{1}{2}B[0;2\rho] \subset \frac{1}{2}\mathrm{cl}(A_1) \subset \frac{1}{2}A_1 + B\Big[0;\frac{1}{2}\rho\Big],$$

folglich ist

$$(*) \quad B\Big[0;\Big(\frac{1}{2}\Big)^i\rho\Big] = \Big(\frac{1}{2}\Big)^i B[0;\rho] \subset \Big(\frac{1}{2}\Big)^{i+1}A_1 + B\Big[0;\Big(\frac{1}{2}\Big)^{i+1}\rho\Big], \qquad i = 0,1,2,\ldots.$$

Sei $y \in B[0;\rho]$ beliebig vorgegeben. Eine Anwendung von $(*)$ mit $i = 0$ ergibt für y eine Darstellung

$$y = \frac{1}{2}(Tu_1 + v_1) + r_1$$

mit

$$u_1 \in (C - x^*)_1, \qquad v_1 \in (K - y^*)_1, \qquad r_1 \in B\Big[0;\Big(\frac{1}{2}\Big)^1\rho\Big].$$

Eine Anwendung von $(*)$ mit $i = 1$ ergibt für r_1 die Darstellung

$$r_1 = \Big(\frac{1}{2}\Big)^2(Tu_2 + v_2) + r_2$$

mit

$$u_2 \in (C - x^*)_1, \qquad v_2 \in (K - y^*)_1, \qquad r_2 \in B\Big[0;\Big(\frac{1}{2}\Big)^2\rho\Big].$$

Fährt man so fort, so erhält man Folgen

$$\{u_k\} \subset (C - x^*)_1, \qquad \{v_k\} \subset (K - y^*)_1, \qquad \{r_k\} \subset Y$$

mit

$$y = T\underbrace{\Big(\sum_{i=1}^{k}\Big(\frac{1}{2}\Big)^i u_i\Big)}_{=:p_k} + \underbrace{\sum_{i=1}^{k}\Big(\frac{1}{2}\Big)^i v_i}_{=:q_k} + r_k, \qquad r_k \in B\Big[0;\Big(\frac{1}{2}\Big)^k\rho\Big] \qquad (k \in \mathbb{N}).$$

Als Durchschnitt konvexer, abgeschlossener Mengen, die das Nullelement von X bzw. Y enthalten, sind $(C-x^*)_1$ bzw. $(K-y^*)_1$ konvex und abgeschlossen, ferner enthalten sie das Nullelement von X bzw. Y. Insbesondere ist daher

$$p_k = \sum_{i=1}^{k}\Big(\frac{1}{2}\Big)^i u_i + \Big(1 - \sum_{i=1}^{k}\Big(\frac{1}{2}\Big)^i\Big)\cdot 0 \in (C - x^*)_1,$$

also $\{p_k\} \subset (C - x^*)_1$. Entsprechend ist $\{q_k\} \subset (K - y^*)_1$. Offensichtlich sind

$$\{p_k\} \subset (C - x^*)_1 \subset X, \qquad \{q_k\} \subset (K - y^*)_1 \subset Y$$

Cauchyfolgen. $(C - x^*)_1$ bzw. $(K - y^*)_1$ sind abgeschlossene Teilmengen der Banachräume X bzw. Y. Daher konvergieren $\{p_k\}$ bzw. $\{q_k\}$ gegen Elemente $p \in (C-x^*)_1$ bzw. $q \in (K - y^*)_1$. Da $\{r_k\}$ eine Nullfolge ist, erhält man aus

$$y = Tp_k + q_k + r_k$$

und der Stetigkeit von T für das vorgegebene $y \in B[0;\rho]$ die Darstellung

$$y = Tp + q \in T((C - x^*)_1) + (K - y^*)_1.$$

Das war zu zeigen. □

Bemerkung 6.19 Ist im Satz von Zowe-Kurcyusz speziell $C = X$ und $K = \{0\}$, so erhält man die Aussage von Satz 6.14, aus dem das klassische Open Mapping Theorem folgt. □

6.4 Der Satz von Lyusternik

Jetzt haben wir alle Hilfsmittel beisammen, um den Satz von Lyusternik zu beweisen.

Satz 6.20 (Lyusternik) *Seien X und Y Banachräume, $C \subset X$, $K \subset Y$ nichtleer, abgeschlossen und konvex. Sei $g\colon X \longrightarrow Y$ eine Abbildung, die im Punkte*

$$x^* \in M := \{x \in X : x \in C,\, g(x) \in -K\}$$

stetig Fréchet-differenzierbar (mit dem Fréchet-Differential $g'(x^*)$) ist. Mit $T(M;x^*)$ wird der Tangentialkegel an M in x^* bezeichnet. Ferner gelte die sogenannte Constraint Qualification

$$\text{(CQ)} \qquad \text{cone}\,(g(x^*) + g'(x^*)(C - x^*) + K) = Y.$$

Dann ist

$$\begin{aligned} L_0(M;x^*) &:= \{p \in X : p \in C - x^*,\, g'(x^*)p \in -(K + g(x^*))\} \\ &\subset L(M;x^*) \\ &:= \{p \in X : p \in \text{cone}\,(C - x^*),\, g'(x^*)p \in -\text{cone}\,(K + g(x^*))\} \\ &\subset \left\{ p \in X : \begin{array}{c} \textit{Es existieren } \hat{t} > 0 \textit{ und } r\colon [0,\hat{t}] \longrightarrow X \textit{ mit} \\ x^* + tp + r(t) \in M \textit{ für alle } t \in [0,\hat{t}],\, \lim_{t\to 0+} r(t)/t = 0 \end{array} \right\} \\ &\subset T(M;x^*), \end{aligned}$$

wobei $T(M;x^*)$ den Tangentialkegel an M in x^* bezeichnet.

Bemerkung: Bevor wir in den Beweis des Satzes von Lyusternik einsteigen, wollen wir einen Spezialfall betrachten. Sei nämlich $C := X$ und $K := \{0\}$ und damit $M = \{x \in X : g(x) = 0\}$. Die Bedingung (CQ) besagt dann, dass $g'(x^*) \in L(X,Y)$ surjektiv ist. Der Satz von Lyusternik sagt in diesem Falle aus, dass $\{p \in X : g'(x^*)p = 0\} \subset T(M;x^*)$. Dies ist von Lyusternik 1934 bewiesen worden, siehe auch J. Jahn (1994, S. 98 ff.). Ist z. B. $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}$, so ist $g'(x^*)$ genau dann surjektiv, wenn der Gradient $\nabla g(x^*)$ von g in x^* vom Nullvektor verschieden ist. Nach dem Satz von Lyusternik ist in diesem Falle jeder Vektor, der senkrecht auf $\nabla g(x^*)$ steht, eine Tangentialrichtung an die Hyperfläche M in x^*. □

Beweis des Satzes von Lyusternik: Wir setzen $y^* := -g(x^*)$ (dann ist $y^* \in K$) und erinnern an die Bezeichnungen

$$(C - x^*)_1 := (C - x^*) \cap B[0;1], \qquad (K - y^*)_1 := (K - y^*) \cap B[0;1]$$

die beim Beweis von Satz 6.18, dem Satz von Zowe-Kurcyusz, eingeführt wurden. Der nicht ganz einfache Beweis zerfällt in zwei Teile. Im ersten Teil, und dieser wird der anstrengendere sein, zeigen wir:

- Zu vorgegebenem $p \in X$ existieren Zahlen $t^* > 0$, $c_0 > 0$ und Abbildungen

  $$r\colon [0,t^*] \longrightarrow X, \qquad z\colon [0,t^*] \longrightarrow Y$$

  derart, dass für alle $t \in [0,t^*]$ gilt:

  (a) $\left\{ \begin{array}{c} r(t) \\ z(t) \end{array} \right\} \in c_0 \, \|g(x^* + tp) - g(x^*) - tg'(x^*)p\| \left\{ \begin{array}{c} (C - x^*)_1 \\ (K - y^*)_1 \end{array} \right),$

  (b) $g(x^*) + tg'(x^*)p = g(x^* + tp + r(t)) + z(t).$

Zum Beweis dieser Aussage können wir o. B. A. $p \neq 0$ annehmen, da man andernfalls $r = 0$, $z = 0$ setzen kann. Wegen Satz 6.18, dem Satz von Zowe-Kurcyusz, existiert ein $\rho > 0$ mit

$$B[0;\rho] \subset g'(x^*)((C - x^*)_1) + (K - y^*)_1.$$

Da g in x^* stetig Fréchet-differenzierbar ist, gibt es ein $\delta > 0$ mit

$$x \in B[x^*;\delta] \Longrightarrow \|g'(x) - g'(x^*)\| \leq \frac{\rho}{2}.$$

Wegen des Mittelwertsatzes 6.11 gilt

$$(*) \qquad x, x' \in B[x^*;\delta] \Longrightarrow \|g(x) - g(x') - g'(x^*)(x - x')\| \leq \frac{\rho}{2} \|x - x'\|.$$

Nun definiere man $t^* := \delta/(2\,\|p\|)$, wähle ein $t \in [0,t^*]$ fest und konstruiere Folgen $\{r_k\} \subset \text{cone}\,(C - x^*)$ und $\{z_k\} \subset \text{cone}\,(K - y^*)$ nach einer Vorschrift, die jetzt angegeben wird:

- Setze $r_0 := 0$, $z_0 := 0$.
- Für $k = 0, 1, \ldots$:
  - Berechne $y_k := g(x^*) + tg'(x^*)p - g(x^* + tp + r_k) - z_k$.
  - Falls $y_k = 0$, dann setze $r(t) := r_k$, $z(t) := z_k$, STOP.
  - Bestimme $\left\{ \begin{array}{c} u_k \\ v_k \end{array} \right\} \in \dfrac{\|y_k\|}{\rho} \left\{ \begin{array}{c} (C - x^*)_1 \\ (K - y^*)_1 \end{array} \right\}$ mit $y_k = g'(x^*)u_k + v_k$.
  - Setze $r_{k+1} := r_k + u_k$, $z_{k+1} := z_k + v_k$.

Diese Konstruktion ist nach Definition von ρ durchführbar, da

$$\frac{\rho}{\|y_k\|} y_k \in B[0;\rho] \subset g'(x^*)((C-x^*)_1) + (K-y^*)_1.$$

Wir zeigen, dass $\{r_k\}$ und $\{z_k\}$ Cauchyfolgen in den Banachräumen X bzw. Y sind und ihre daher existierenden Limiten $r = r(t)$ bzw. $z = z(t)$ mit $c_0 := 2/\rho$ den behaupteten Bedingungen (a) und (b) genügen.
Zur Abkürzung setzen wir

$$d(t) := \|g(x^* + tp) - g(x^*) - tg'(x^*)p\|, \qquad q := \frac{1}{2}.$$

Wegen $\|tp\| \le t^* \|p\| = \delta/2$ folgt aus $(*)$, dass

$$d(t) = \|g(x^* + tp) - g(x^*) - tg'(x^*)p\| \le \frac{\rho}{2} t \|p\| \le \frac{\rho}{4}\delta.$$

Durch vollständige Induktion nach k zeigen wir: Solange die Konstruktion nicht vorzeitig abbricht, ist

1. $\left\{ \begin{array}{c} r_k \\ z_k \end{array} \right\} \in \frac{1}{\rho}\Big(\frac{1-q^k}{1-q}\Big)\, d(t) \left\{ \begin{array}{c} (C-x^*)_1 \\ (K-y^*)_1 \end{array} \right\}$,
2. $x^* + tp + r_k \in B[x^*;\delta]$,
3. $\|y_k\| \le q^k d(t)$,
4. $\left\{ \begin{array}{c} u_k \\ v_k \end{array} \right\} \in \frac{q^k}{\rho}\, d(t) \left\{ \begin{array}{c} (C-x^*)_1 \\ (K-y^*)_1 \end{array} \right\}$.

Für den Induktionsbeweis benötigen wir zwei Aussagen über konvexe Mengen.

- Sei E ein linearer Raum und $A \subset E$ konvex. Dann gilt:

  (i) Sind $\lambda, \mu \ge 0$, so ist $\lambda A + \mu A \subset (\lambda+\mu)A$.
  Denn: Wir können annehmen, dass $\lambda + \mu > 0$. Mit $a_1, a_2 \in A$ ist

  $$\lambda a_1 + \mu a_2 = (\lambda+\mu)\Big(\underbrace{\frac{\lambda}{\lambda+\mu}a_1 + \frac{\mu}{\lambda+\mu}a_2}_{\in A}\Big) \in (\lambda+\mu)A.$$

  (ii) Ist $0 \in A$ und $0 \le \lambda \le \mu$, so ist $\lambda A \subset \mu A$.
  Denn: Wir können annehmen, dass $\mu > 0$. Mit $a \in A$ ist

  $$\lambda a = \mu\Big[\underbrace{\frac{\lambda}{\mu}a + \Big(1-\frac{\lambda}{\mu}\Big)0}_{\in A}\Big] \in \mu A.$$

Nun kommen wir zum Induktionsbeweis. Für $k = 0$ sind die Aussagen 1.–4. richtig, wobei wir $\|y_0\| = d(t)$ berücksichtigen. Angenommen, die Aussagen 1.–4. seien auch für k richtig.

1. Es ist

$$\begin{aligned} r_{k+1} &= r_k + u_k \\ &\in \frac{1}{\rho}\Big(\frac{1-q^k}{1-q}\Big) d(t)(C-x^*)_1 + \frac{q^k}{\rho} d(t)(C-x^*)_1 \\ &\qquad \text{(wegen der Induktionsannahmen 1. und 4.)} \\ &\subset \frac{1}{\rho}\Big(\frac{1-q^{k+1}}{1-q}\Big) d(t)(C-x^*)_1 \\ &\qquad \text{(wegen obiger Aussage (i)).} \end{aligned}$$

   Entsprechend zeigt man

$$z_{k+1} \in \frac{1}{\rho}\Big(\frac{1-q^{k+1}}{1-q}\Big) d(t)(K-y^*)_1.$$

   Damit ist 1. auch für $k+1$ richtig.

2. Wir hatten $q := 1/2$ gesetzt und $d(t) \le \rho\delta/4$ nachgewiesen. Aus 1. (für $k+1$) erhalten wir daher

$$\|r_{k+1}\| \le \frac{1}{\rho}\Big(\frac{1-q^{k+1}}{1-q}\Big) d(t) \le \frac{1}{\rho} 2 \frac{\rho}{4}\delta = \frac{/delta}{2}.$$

   Hieraus und aus $t\|p\| \le \delta/2$ folgt $x^* + tp + r_{k+1} \in B[x^*;\delta]$. Damit gilt 2. auch für $k+1$.

3. Nach Konstruktion der Folgen $\{r_k\}$, $\{z_k\}$ ist

$$\begin{aligned} \|y_{k+1}\| &= \|g(x^*) + tg'(x^*)p - g(x^* + tp + r_{k+1}) - z_{k+1}\| \\ &= \|g(x^*) + tg'(x^*)p - g(x^* + tp + r_{k+1}) - z_k - v_k\| \\ &= \|g(x^* + tp + r_{k+1}) - g(x^* + tp + r_k) - g'(x^*)u_k\| \\ &\qquad \text{(wegen } v_k = y_k - g'(x^*)u_k \text{ und der Definition von } y_k) \\ &\le \frac{\rho}{2}\|u_k\| \\ &\qquad \text{(wegen } (*) \text{ und } x^* + tp + r_k, x^* + tp + r_{k+1} \in B[x^*;\delta]) \\ &\le \frac{\rho}{2}\frac{q^k}{\rho} d(t) \\ &\qquad \text{(wegen der Induktionsannahme 4.)} \\ &= q^{k+1} d(t) \\ &\qquad \text{(da } q = 1/2). \end{aligned}$$

   Damit ist 3. für $k+1$ nachgewiesen.

4. Nach Konstruktion ist

$$u_{k+1} \in \frac{\|y_{k+1}\|}{\rho}(C-x^*)_1 \subset \frac{q^{k+1}}{\rho} d(t)(C-x^*)_1,$$

wobei 3. (für $k+1$) und obige Aussage (ii) über konvexe Mengen benutzt wurden. Entsprechend folgt

$$v_{k+1} \in \frac{q^{k+1}}{\rho} d(t)(K - y^*)_1.$$

Damit ist 4. auch für $k+1$ richtig.

Insgesamt ist der Induktionsbeweis abgeschlossen. Aus 1.–4. erhalten wir, dass $\{r_k\}$ und $\{z_k\}$ Cauchyfolgen mit

$$\{r_k\} \subset \frac{1}{\rho}\Big(\frac{1}{1-q}\Big) d(t)(C - x^*)_1 = c_0 d(t)(C - x^*)_1,$$
$$\{z_k\} \subset \frac{1}{\rho}\Big(\frac{1}{1-q}\Big) d(t)(K - y^*)_1 = c_0 d(t)(K - y^*)_1$$

sind, wobei wir an die Definition $c_0 := 2/\rho$ erinnern. Daher konvergieren $\{r_k\}$ bzw. $\{z_k\}$ gegen Elemente $r = r(t)$ bzw. $z = z(t)$. Da $c_0 d(t)(C - x^*)_1$ bzw. $c_0 d(t)(K - y^*)_1$ abgeschlossen sind, ist

$$r = r(t) \in c_0 d(t)(C - x^*)_1, \qquad z = z(t) \in c_0 d(t)(K - y^*)_1.$$

Da $\{u_k\}$ und $\{v_k\}$ gegen das Nullelement von X bzw. Y konvergieren und nach Konstruktion

$$y_k = g(x^*) + tg'(x^*)p - g(x^* + tp + r_k) - z_k = g'(x^*)u_k + v_k$$

gilt, folgt

$$g(x^*) + tg'(x^*)p = g(x^* + tp + r(t)) + z(t),$$

wobei die aus $(*)$ folgende Stetigkeit von g in $x^* + tp + r(t) \in B[x^*; \delta]$ benutzt wurde. Damit ist der erste Teil des Beweises abgeschlossen.

Nun kommen wir zum zweiten Teil des Beweises. Von den im Satz von Lyusternik behaupteten Inklusionsbeziehungen ist nur eine nichttrivial, nämlich die Aussage

$$\begin{aligned} L(M; x^*) &:= \{p \in X : p \in \text{cone}\,(C - x^*),\ g'(x^*)p \in -\text{cone}\,(K + g(x^*))\} \\ &\subset \left\{ p \in X : \begin{array}{c} \text{Es existieren } \hat{t} > 0 \text{ und } r\colon [0, \hat{t}] \longrightarrow X \text{ mit} \\ x^* + tp + r(t) \in M \text{ für alle } t \in [0, \hat{t}],\ \lim_{t \to 0+} r(t)/t = 0 \end{array} \right\}. \end{aligned}$$

Sei daher ein $p \in L(M; x^*)$ vorgegeben. Wir erinnern daran, dass wir $y^* := -g(x^*)$ zu Beginn des Beweises gesetzt haben. Im ersten Teil des Beweises hatten wir gezeigt, dass zu p positive Zahlen t^*, c_0 und Abbildungen

$$r\colon [0, t^*] \longrightarrow X, \qquad z\colon [0, t^*] \longrightarrow Y$$

existieren derart, dass mit

$$d(t) := \|g(x^* + tp) - g(x^*) - tg'(x^*)p\|$$

für alle $t \in [0, t^*]$ gilt:

(a) $\left\{ \begin{array}{c} r(t) \\ z(t) \end{array} \right\} \in c_0 d(t) \left\{ \begin{array}{c} (C - x^*)_1 \\ (K - y^*)_1 \end{array} \right\}$,

(b) $g(x^*) + tg'(x^*)p = g(x^* + tp + r(t)) + z(t)$.

Aus (a) folgt

$$\frac{\|r(t)\|}{t} \leq c_0 \frac{d(t)}{t}, \quad t \in [0, t^*],$$

und hieraus

$$\lim_{t \to 0+} \frac{r(t)}{t} = 0.$$

Zu zeigen bleibt daher, dass $x^* + tp + r(t) \in M$ für alle $t \in [0, \hat{t}]$ mit hinreichend kleinem $\hat{t} \in (0, t^*]$. Da wir in $M := \{x \in X : x \in C, g(x) \in -K\}$ explizite und implizite Restriktionen zusammengefasst haben, zerfällt der Beweis hierfür in zwei Teile.

(α) Wegen $p \in \text{cone}\,(C - x^*)$ ist $p = \lambda(c - x^*)$ mit $\lambda \geq 0$ und $c \in C$. Wegen (a) lässt sich $r(t)$ als $r(t) = c_0 d(t)(c(t) - x^*)$ mit $c(t) \in C$ darstellen. Dann ist

$$x^* + tp + r(t) = (1 - \lambda t - c_0 d(t))x^* + \lambda t c + c_0 d(t) c(t) \in C$$

als Konvexkombination von drei Elementen der konvexen Menge C, falls nur $1 - \lambda t - c_0 d(t) \geq 0$, was für alle hinreichend kleinen $t > 0$ wegen $\lim_{t \to 0+} d(t) = 0$ richtig ist.

(β) Wegen $g'(x^*)p \in -\text{cone}\,(K + g(x^*))$ ist $g'(x^*)p = -\mu(k + g(x^*))$ mit $\mu \geq 0$, $k \in K$. Wegen (a) unter Berücksichtigung von $y^* = -g(x^*)$ lässt sich $z(t)$ in der Form $z(t) = c_0 d(t)(k(t) + g(x^*))$ mit $k(t) \in K$ darstellen. Dies ergibt zusammen mit (b), ganz ähnlich wie in (α), dass

$$g(x^* + tp + r(t)) = -[(1 - \mu t - c_0 d(t))(-g(x^*)) + \mu t k + c_0 d(t) k(t)] \in -K$$

für alle hinreichend kleinen $t > 0$.

Damit ist die Existenz eines $\hat{t} \in (0, t^*]$ mit $x^* + tp + r(t) \in M$ für alle $t \in [0, \hat{t}]$ nachgewiesen. Der Satz von Lyusternik ist bewiesen. □

Im folgenden Satz wird der Satz von Lyusternik auf den endlichdimensionalen Fall spezialisiert.

Satz 6.21 *Die Abbildung $g: \mathbb{R}^n \longrightarrow \mathbb{R}^m$ mit $g(x) = (g_1(x), \ldots, g_m(x))^T$ sei auf einer Kugel um*

$$x^* \in M := \left\{ x \in \mathbb{R}^n : \begin{array}{ll} g_i(x) = 0, & i = 1, \ldots, m_0, \\ g_i(x) \leq 0, & i = m_0 + 1, \ldots, m \end{array} \right\}$$

stetig partiell differenzierbar, wobei $m_0 \in \mathbb{Z}$ und $0 \leq m_0 \leq m$. Mit

$$I^* := \{i \in \{m_0 + 1, \ldots, m\} : g_i(x^*) = 0\}$$

werden die in x^* aktiven Ungleichungsrestriktionen bezeichnet. Ferner gelte

(a) $\{\nabla g_1(x^*), \ldots, \nabla g_{m_0}(x^*)\}$ *sind linear unabhängig.*

(b) Es existiert ein $\hat{p} \in \mathbb{R}^n$ *mit*
$$\begin{array}{ll} \nabla g_i(x^*)^T \hat{p} = 0, & i = 1, \ldots, m_0, \\ \nabla g_i(x^*)^T \hat{p} < 0, & i \in I^*. \end{array}$$

Dann ist
$$\left\{ p \in \mathbb{R}^n : \begin{array}{ll} \nabla g_i(x^*)^T p = 0, & i = 1, \ldots, m_0, \\ \nabla g_i(x^*)^T p \leq 0, & i \in I^* \end{array} \right\} \subset T(M; x^*),$$
wobei $T(M; x^*)$ *den Tangentialkegel an* M *in* x^* *bezeichnet.*

Beweis: Mit $X := \mathbb{R}^n$, $Y := \mathbb{R}^m$, $C := \mathbb{R}^n$ und
$$K := \left\{ k = (k_i) \in \mathbb{R}^m : \begin{array}{ll} k_i = 0, & i = 1, \ldots, m_0, \\ k_i \geq 0, & i = m_0 + 1, \ldots, m \end{array} \right\}$$
wenden wir den Satz von Lyusternik an. Nach Beispiel 6.7 ist g in x^* stetig Fréchet-differenzierbar mit der Funktionalmatrix
$$g'(x^*) = \begin{pmatrix} \nabla g_1(x^*)^T \\ \vdots \\ \nabla g_m(x^*)^T \end{pmatrix}$$
als Fréchet-Differential. Mit diesen Bezeichnungen ist offenbar
$$\begin{aligned} L_0(M; x^*) &:= \{p \in \mathbb{R}^n : p \in C - x^*,\ g'(x^*)p \in -(K + g(x^*))\} \\ &= \left\{ p \in \mathbb{R}^n : \begin{array}{ll} \nabla g_i(x^*)^T p = 0, & i = 1, \ldots, m_0, \\ \nabla g_i(x^*)^T p \leq 0, & i \in I^* \end{array} \right\}. \end{aligned}$$

Es bleibt, die Gültigkeit der Constraint Qualification (CQ) im Satz von Lyusternik nachzuweisen. Mit dem oben definierten K lautet diese hier:

(CQ) $$\text{cone}\,(g(x^*) + g'(x^*)(\mathbb{R}^n) + K) = \mathbb{R}^m.$$

Zu zeigen ist also, dass es zu jedem $y \in \mathbb{R}^m$ (offenbar ist o. B. d. A. $y \neq 0$) ein nichtnegatives λ sowie Elemente $p \in \mathbb{R}^n$ und $k \in K$ mit
$$y = \lambda[g(x^*) + g'(x^*)p + k]$$
gibt. Komponentenweise bedeutet dies, dass ein $\lambda \geq 0$ und ein $p \in \mathbb{R}^n$ mit
$$y_i = \lambda \nabla g_i(x^*)^T p, \qquad i = 1, \ldots, m_0,$$
sowie
$$y_i \geq \lambda[g_i(x^*) + \nabla g_i(x^*)^T p], \qquad i = m_0 + 1, \ldots, m,$$
zu finden sind. Da $\{\nabla g_1(x^*), \ldots, \nabla g_{m_0}(x^*)\}$ linear unabhängig sind, existiert ein $q \in \mathbb{R}^n$ mit
$$y_i = \nabla g_i(x^*)^T q, \qquad i = 1, \ldots, m_0.$$

Mit noch unbekannten positiven λ und t machen wir für den gesuchten Vektor p den Ansatz $p = q/\lambda + t\hat{p}$. Für beliebige positive λ und t ist dann

$$y_i = \lambda \nabla g_i(x^*)^T p, \qquad i = 1, \dots, m_0.$$

Bei den Ungleichungsrestriktionen unterscheiden wir danach, ob diese in x^* aktiv sind oder nicht. Für $i \in I^*$, also eine in x^* aktive Ungleichungsrestriktion, ist

$$\lambda[\underbrace{g_i(x^*)}_{=0} + \nabla g_i(x^*)^T p] = \nabla g_i(x^*)^T q + \lambda t \underbrace{\nabla g_i(x^*)^T \hat{p}}_{<0}.$$

Werden also die positiven λ, t so gewählt, dass

$$\lambda t \geq \max_{i \in I^*} \Big(\frac{y_i - \nabla g_i(x^*)^T q}{\nabla g_i(x^*)^T \hat{p}} \Big), \tag{1}$$

so ist

$$y_i \geq \lambda [g_i(x^*) + \nabla g_i(x^*)^T p], \qquad i \in I^*.$$

Für $i \in \{m_0 + 1, \dots, m\} \setminus I^*$ (falls es solche i nicht gibt, so sind wir schon fertig) ist

$$\lambda[g_i(x^*) + \nabla g_i(x^*)^T p] = \lambda[\underbrace{g_i(x^*)}_{<0} + t \nabla g_i(x^*)^T \hat{p}] + \nabla g_i(x^*)^T q.$$

Die Idee bei der Wahl von λ und t besteht nun darin, zunächst $t > 0$ hinreichend klein und anschließend $\lambda > 0$ hinreichend groß zu wählen. Genauer sei

$$d := \max_{i \in \{m_0+1, \dots, m\} \setminus I^*} g_i(x^*).$$

Dann ist $d < 0$. Jetzt wähle man $t > 0$ so klein, dass

$$t \max_{i \in \{m_0+1, \dots, m\} \setminus I^*} \nabla g_i(x^*)^T \hat{p} \leq -\frac{d}{2} \tag{2}$$

und damit

$$g_i(x^*) + t \nabla g_i(x^*)^T \hat{p} \leq d - \frac{d}{2} = \frac{d}{2}, \qquad i \in \{m_0 + 1, \dots, m\} \setminus I^*.$$

Mit dem gewählten $t > 0$ bestimme man nun $\lambda > 0$ so groß, dass einerseits (1) gilt und andererseits

$$\lambda \frac{d}{2} \leq \min_{i \in \{m_0+1, \dots, m\} \setminus I^*} (y_i - \nabla g_i(x^*)^T q). \tag{3}$$

Damit ist die Constraint Qualification (CQ) nachgewiesen und die Aussage des Satzes folgt aus dem Satz von Lyusternik. □

In einem Korollar zum Satz von Lyusternik zeigen wir nun noch, dass die Aussage des Satzes von Lyusternik auch dann gilt, wenn der Ausgangsraum X lediglich ein linearer normierter Raum (und *kein* Banachraum) ist, dafür aber in der Restriktionenmenge keine expliziten Restriktionen (d. h. es ist $C = X$) und nur endlich viele Gleichungen (d. h. Y ist endlichdimensional und $K = \{0\}$) vorkommen.

Korollar 6.22 *Seien X und Y lineare normierte Räume, wobei Y endlichdimensional sei. Die Abbildung $g\colon X \longrightarrow Y$ sei in*

$$x^* \in M := \{x \in X : g(x) = 0\}$$

stetig Fréchet-differenzierbar (mit dem Fréchet-Differential $g'(x^*) \in L(X,Y)$ in x^*). Es gelte

$$\text{(CQ)} \qquad g'(x^*)(X) = Y,$$

d. h. $g'(x^*)$ sei surjektiv. Dann ist

$$\begin{aligned} L_0(M;x^*) &:= \{p \in X : g'(x^*)p = 0\} \\ &\subset \left\{ p \in X : \begin{array}{c} \text{Es existieren } t^* > 0 \text{ und } r\colon [0,t^*] \longrightarrow X \text{ mit} \\ x^* + tp + r(t) = 0 \text{ für alle } t \in [0,t^*],\ \lim_{t\to 0+} r(t)/t = 0 \end{array} \right\} \\ &\subset T(M;x^*), \end{aligned}$$

wobei $T(M;x^*)$ den Tangentialkegel an M in x^* bezeichnet.

Beweis: Sei $m := \dim(Y)$ und $Y = \text{span}\,\{y_1,\ldots,y_m\}$. Wegen der Surjektivität von $g'(x^*)$ existieren $\{x_1,\ldots,x_m\} \subset X$ mit $g'(x^*)x_i = y_i$, $i = 1,\ldots,m$. Definiert man

$$X_m := \text{span}\,\{x_1,\ldots,x_m\},$$

so ist $g'(x^*)(X_m) = Y$. Als endlichdimensionale lineare normierte Räume sind X_m und Y Banachräume. Satz 6.18, der Satz von Zowe-Kurcyusz, liefert die Existenz eines $\rho > 0$ mit $B[0;\rho] \subset g'(x^*)(X_m \cap B[0;1])$. Der erste Teil des Beweises von Satz 6.20, des Satzes von Lyusternik, zeigt, dass es zu beliebigem $p \in X$ positive Zahlen t^*, c_0 und eine Abbildung $r\colon [0,t^*] \longrightarrow X_m$ gibt derart, dass für alle $t \in [0,t^*]$ gilt:

(a) $\|r(t)\| \le c_0 \,\|g(x^* + tp) - g(x^*) - tg'(x^*)p\|$,

(b) $g(x^*) + tg'(x^*)p = g(x^* + tp + r(t))$.

Hieraus folgt die Behauptung. □

7 Der Satz von Kuhn-Tucker bei Optimierungsaufgaben in linearen normierten Räumen

In diesem Abschnitt betrachten wir die Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in X : x \in C,\ g(x) \in -K\}.$$

Hier heißt f die *Zielfunktion*, die Bedingungen $x \in C$ bzw. $g(x) \in -K$ nennen wir eine *explizite* bzw. eine *implizite* Restriktion. I. Allg. setzen wir voraus:

(V1) X ist ein (reeller) Banachraum, $f\colon X \longrightarrow \mathbb{R}$.

(V2) $C \subset X$ ist nichtleer, abgeschlossen und konvex. Die Bedingung $x \in C$ nennen wir eine *explizite Restriktion*.

(V3) Y ist ein (reeller) Banachraum, $g\colon X \longrightarrow Y$.

(V4) $K \subset Y$ ist ein nichtleerer, abgeschlossener und konvexer Kegel. Die Bedingung $g(x) \in -K$ nennen wir eine *implizite Restriktion*. Ist $\operatorname{int}(K) \neq \emptyset$, so sprechen wir von einer *Restriktion vom Ungleichungstyp*.

(V5) $x^* \in M$ ist eine lokale Lösung von (P), d. h. es existiert ein $\epsilon > 0$ mit $f(x^*) \leq f(x)$ für alle $x \in M \cap B[x^*; \epsilon]$. Ferner besitzt f in x^* eine konvexe Hadamard-Variation $f'(x^*; \cdot)\colon X \longrightarrow \mathbb{R}$ und g ist in x^* stetig Fréchet-differenzierbar.

Außerdem gelte die schon im Satz 6.20 von Lyusternik auftretende Constraint Qualification

$$\text{(CQ)} \qquad \operatorname{cone}\,(g(x^*) + g'(x^*)(C - x^*) + K) = Y.$$

Unser Ziel ist es, notwendige Optimalitätsbedingungen erster Ordnung für die lokale Lösung x^* von (P) herzuleiten. Diese gewinnt man auf einem Weg, den wir nun beschreiben wollen. Am Schluss der Argumentation werden wir das erhaltene Ergebnis in einem Satz zusammenfassen. Wir folgen ziemlich genau der Darstellung bei J. Werner (1988, Kurseinheit 5), siehe aber auch J.Zowe, S. Kurcyusz (1979, Theorem 3.1).

1. Aus (V5) folgt (siehe Bemerkung 6.3), dass

   $$0 \leq f'(x^*; p) \quad \text{für alle } p \in T(M; x^*),$$

   wobei (siehe Definition 6.1) $T(M; x^*)$ den Tangentialkegel an M in x^* bedeutet.

2. Der Satz von Lyusternik (siehe Satz 6.20) liefert, dass

   $$L_0(M; x^*) := \{p \in X : p \in C - x^*,\ g(x^*) + g'(x^*)p \in -K\} \subset T(M; x^*).$$

   Hier wird benutzt, dass X und Y Banachräume ((V1) und (V3)) und C bzw. K nicht nur konvex, sondern auch abgeschlossen ((V2) und (V4)) sind. Die Kegeleigenschaft von K wird noch nicht gebraucht.

3. Man definiere die Menge

   $$\Lambda_+ := \{(f'(x^*; p) + r, g(x^*) + g'(x^*)p + z) \in \mathbb{R} \times Y : p \in C - x^*,\ r > 0,\ z \in K\}.$$

   Dann ist $(0, 0) \notin \Lambda_+$, denn andernfalls existieren $p \in C - x^*$, $r > 0$ und $z \in K$ mit

   $$(0, 0) = (f'(x^*; p) + r, g(x^*) + g'(x^*)p + z).$$

   Dies bedeutet aber, dass $p \in L_0(M; x^*)$ und $f'(x^*; p) < 0$, ein Widerspruch zu 1. und 2. Die Menge $\Lambda_+ \subset \mathbb{R} \times Y$ ist konvex. Um dies einzusehen, geben wir uns zwei Punkte

   $$P_i := (f'(x^*; p_i) + r_i, g(x^*) + g'(x^*)p_i + z_i) \in \Lambda_+, \quad i = 0, 1,$$

aus Λ_+ und ein $\lambda \in [0,1]$ vor. Hierbei ist natürlich $p_i \in C - x^*$, $r_i > 0$, $z_i \in K$, $i = 0, 1$. Wir setzen

$$p_\lambda := (1-\lambda)p_0 + \lambda p_1, \qquad z_\lambda := (1-\lambda)z_0 + \lambda z_1, \qquad r_\lambda := (1-\lambda)r_0 + \lambda r_1.$$

Wegen der Konvexität von C bzw. K ist $p_\lambda \in C - x^*$ und $z_\lambda \in K$, ferner ist $r_\lambda > 0$. Da $f'(x^*;\cdot)\colon X \longrightarrow \mathbb{R}$ konvex ist ((V5)), ist

$$q_\lambda := (1-\lambda)f'(x^*;p_0) + \lambda f'(x^*;p_1) - f'(x^*;p_\lambda) \geq 0.$$

Dann ist

$$\begin{aligned}(1-\lambda)P_0 + \lambda P_1 &= ((1-\lambda)f'(x^*;p_0) + \lambda f'(x^*;p_1) + r_\lambda, g(x^*) + g'(x^*)p_\lambda + z_\lambda) \\ &= (f'(x^*;p_\lambda) + \underbrace{q_\lambda + r_\lambda}_{>0}, g(x^*) + g'(x^*)p_\lambda + z_\lambda) \\ &\in \Lambda_+.\end{aligned}$$

Also gehört mit P_0 und P_1 auch die gesamte Verbindungsstrecke $[P_0, P_1]$ zu Λ_+, d. h. Λ_+ ist konvex.

4. *Angenommen*, der Punkt $(0,0) \in \mathbb{R} \times Y$ und die konvexe Menge $\Lambda_+ \subset \mathbb{R} \times Y$ lassen sich durch eine abgeschlossene Hyperebene in $\mathbb{R} \times Y$ trennen. Als Folgerung dieser Annahme erhalten wir:

   - Es existiert ein Paar $(q^*, l^*) \in \mathbb{R} \times Y^* \setminus \{(0,0)\}$ mit

   (a) $$\begin{cases} 0 \leq q^*(f'(x^*;p) + r) + l^*(g(x^*) + g'(x^*)p + z) \\ \qquad \text{für alle } p \in C - x^*,\ r > 0,\ z \in K. \end{cases}$$

   Setzen wir in (a) speziell $p = 0$ und $z = 0$, so erhalten wir $0 \leq q^* r + l^*(g(x^*))$ bzw. $-l^*(g(x^*)) \leq q^* r$ für alle $r > 0$. Hieraus folgt $q^* \geq 0$ und mit $r \to 0+$ auch $-l^*(g(x^*)) \leq 0$ bzw. $l^*(g(x^*)) \geq 0$. Indem man $p = 0$ und $r = 0$ (dies ist nach einem Grenzübergang möglich) in (a) setzt, so erhält man $0 \leq l^*(g(x^*) + z)$ für alle $z \in K$. Hieraus folgt, dass $0 \leq l^*(z)$ für alle $z \in K$. Denn gäbe es ein $z_0 \in K$ mit $l^*(z_0) < 0$, so ist wegen $\lambda z_0 \in K$ für alle $\lambda \geq 0$ (hier geht ein, dass K wegen (V4) ein Kegel ist) auch

   $$0 \leq l^*(g(x)) + \lambda \underbrace{l^*(z_0)}_{<0} \qquad \text{für alle } \lambda \geq 0,$$

   was mit $\lambda \to +\infty$ einen Widerspruch ergibt. Da $-g(x^*) \in K$ und $0 \leq l^*(g(x^*))$, ist $l^*(g(x^*)) = 0$. Definieren wir also den zu $K \subset Y$ *dualen Kegel* K^+ durch

   $$K^+ := \{l \in Y^* : l(z) \geq 0 \text{ für alle } z \in K\},$$

   so haben wir bisher als Folge der *Annahme*, $(0,0)$ lasse sich von Λ_+ durch eine abgeschlossene Hyperebene trennen, unter den Voraussetzungen (V1)–(5) nachgewiesen:

- Es existiert ein Paar $(q^*, l^*) \in \mathbb{R} \times Y^* \setminus \{(0,0)\}$ mit $q^* \geq 0$, $l^* \in K^+$, $l^*(g(x^*)) = 0$ und

  (b) $\qquad 0 \leq q^* f'(x^*; p) + l^*(g'(x^*)p) \quad$ für alle $p \in C - x^*$.

Wir wollen uns überlegen, dass notwendigerweise $q^* > 0$ ist. Denn wäre $q^* = 0$, so erhielte man aus (b) wegen $l^*(g(x^*)) = 0$, dass $0 \leq l^*(g(x^*) + g'(x^*)p)$ für alle $p \in C - x^*$. Wir werden zeigen, dass dann $0 \leq l^*(y)$ für alle $y \in Y$ bzw. (ersetze y durch $-y$) sogar $l^*(y) = 0$ für alle $y \in Y$ bzw. $l^* = 0$, womit wir einen Widerspruch zu $(q^*, l^*) \neq (0,0)$ erhalten haben. Wir geben uns ein beliebiges $y \in Y$ vor. Wegen der Constraint Qualification (CQ) lässt sich y darstellen als

$$y = \lambda[g(x^*) + g'(x^*)p + z] \quad \text{mit } \lambda \geq 0,\ p \in C - x^* \text{ und } z \in K.$$

Hierbei haben wir ausgenutzt, dass $g(x^*) + g'(x^*)(C - x^*) + K$ konvex ist. Dann ist

$$l^*(y) = l^*(\lambda(g(x^*) + g'(x^*)p + z)) = \underbrace{\lambda}_{\geq 0}[\underbrace{l^*(g(x^*) + g'(x^*)p)}_{\geq 0} + \underbrace{l^*(z)}_{\geq 0}] \geq 0.$$

Hierbei haben wir $(**)$ und $l^* \in K^+$ ausgenutzt. Also ist $q^* > 0$ und daher o. B. d. A. $q^* = 1$, da wir notfalls l^* durch l^*/q^* ersetzen können. Unter den Voraussetzungen (V1)–(V5) und der *Annahme*, dass sich $(0,0) \in \mathbb{R} \times Y$ und Λ_+ durch eine abgeschlossene Hyperebene trennen lassen, haben wir damit die folgende Aussage erhalten:

- Es existiert $l^* \in Y^*$ mit $l^* \in K^+$, $l^*(g(x^*)) = 0$ und

  (c) $\qquad 0 \leq f'(x^*; x - x^*) + l^*(g'(x^*; x - x^*)) \quad$ für alle $x \in C$.

5. Jetzt stellt sich die Frage, unter welchen Voraussetzungen sich der Punkt $(0,0) \in \mathbb{R} \times Y$ von der konvexen Menge

$$\Lambda_+ := \{(f'(x^*; p) + r, g(x^*) + g'(x^*)p + z) \in \mathbb{R} \times Y : p \in C - x^*,\ r > 0,\ z \in K\}$$

durch eine abgeschlossene Hyperebene trennen lässt. Wegen $(0,0) \notin \Lambda_+$ bzw. $\{(0,0)\} \cap \Lambda_+ = \emptyset$ ist dies der Fall, wenn Y endlichdimensional ist. Denn zwei disjunkte konvexe Mengen im $\mathbb{R}^n$ lassen sich durch eine (notwendig abgeschlossene) Hyperebene trennen, siehe z. B. J. Werner (2013, Abschnitt 49). Auch in linearen normierten Räumen ist die Aussage wegen des Satzes 5.5 von Eidelheit richtig, wenn nur int $(\Lambda_+) \neq \emptyset$. Gesucht sind daher hinreichende Bedingungen dafür, dass int $(\Lambda_+) \neq \emptyset$. Wir geben zwei solcher Bedingungen an.

- Ist int $(K) \neq \emptyset$, so ist int $(\Lambda_+) \neq \emptyset$.

Denn: Ein Punkt $P_0 := (f'(x^*; p_0) + r_0, g(x^*) + g'(x^*)p_0 + z_0)$ mit $r_0 > 0$, $p_0 \in C - x^*$ und $z_0 \in$ int (K) ist ganz offensichtlich ein Punkt aus int (Λ_+).

- Die Hadamard-Variation $f'(x^*;\cdot)\colon X \longrightarrow \mathbb{R}$ sei in 0 stetig, was insbesondere der Fall ist, wenn f in x^* Fréchet-differenzierbar ist oder X endlichdimensional ist. Dann ist $(q_0, 0) \in \text{int}\,(\Lambda_+)$ für alle hinreichend großen q_0, insbesondere ist $\text{int}\,(\Lambda_+) \neq \emptyset$.

Denn: Aus dem Satz 6.18 von Zowe-Kurcyusz (setze $T := g'(x^*)$, $y^* := -g(x^*)$) folgt die Existenz eines $\rho > 0$ mit

$$(*) \qquad B[0;\rho] \subset g(x^*) + g'(x^*)((C - x^*) \cap B[0;1]) + K.$$

Nun nutzen wir aus, dass $f'(x^*;\cdot)$ nach Voraussetzung in 0 stetig ist. Insbesondere existiert daher zu $\epsilon := 1$ ein $\delta > 0$ mit

$$\|p\| \leq \delta \Longrightarrow f'(x^*;p) = f'(x^*;p) - f'(x^*;0) \leq 1.$$

Da die Hadamard-Variation $f'(x^*;\cdot)$ positiv homogen ist, d. h. für alle $\alpha \geq 0$ und $p \in X$ ist $f'(x^*;\alpha p) = \alpha f'(x^*;p)$, gilt die Implikation

$$\|p\| \leq 1 \Longrightarrow f'(x^*;p) \leq f_0 := \frac{1}{\delta}.$$

Wir zeigen, dass $(q_0, 0) \in \text{int}\,(\Lambda_+)$ für alle $q_0 > f_0$), indem wir

$$(q_0, 0) + \Big[-\frac{1}{2}(q_0 - f_0), \frac{1}{2}(q_0 - f_0)\Big] \times B[0;\rho] \subset \Lambda_+$$

nachweisen. Hierzu gebe man sich

$$(q, y) \in \Big[-\frac{1}{2}(q_0 - f_0), \frac{1}{2}(q_0 - f_0)\Big] \times B[0;\rho]$$

beliebig vor. Wegen $(*)$ lässt sich $y \in B[0;\rho]$ darstellen als

$$y = g(x^*) + g'(x^*)p + z \quad \text{mit} \quad p \in (C - x^*) \cap B[0;1],\ z \in K.$$

Dann ist

$$\begin{aligned}(q_0, 0) + (q, y) &= (q_0 + q, y) \\ &= (f'(x^*;p) + \underbrace{q_0 + q - f'(x^*;p)}_{=:r}, g(x^*) + g'(x^*)p + z) \\ &\in \Lambda_+,\end{aligned}$$

da

$$r := q_0 + q - f'(x^*;p) \geq q_0 - \frac{q_0 - f_0}{2} - f_0 = \frac{q_0 - f_0}{2} > 0.$$

Damit haben wir zwei hinreichende Bedingungen für $\text{int}\,(\Lambda_+) \neq \emptyset$ gewonnen.

Bemerkung: Bemerkenswert ist, dass die Constraint Qualification (CQ) in der obigen Argumentation an *zwei* Stellen eine Rolle spielt. *Einmal* wird (CQ) für die Anwendung des Satzes von Lyusternik benötigt. *Andererseits* spielt (CQ) auch für den Nachweis,

dass eine $(0,0)$ und Λ_+ trennende abgeschlossene Hyperebene in $\mathbb{R}\times Y$ nicht parallel zu $\mathbb{R}$ ist, eine wichtige Rolle. Weiter weisen wir darauf hin, dass die Voraussetzung, X und Y seien Banachräume, und nicht nur lineare normierte Räume, nur bei der Anwendung des Satzes von Lyusternik benutzt wird. □

Nun fassen wir das aus der obigen Argumentation resultierende Ergebnis in einem Satz zusammen. Für endlichdimensionale Optimierungsaufgaben ist ein entsprechendes Ergebnis von Kuhn-Tucker gewonnen worden, siehe z. B. J. Werner (2013, Abschnitt 53). Daher nennen wir auch den folgenden Satz einen Satz von Kuhn-Tucker.

Satz 7.1 (Kuhn-Tucker) *Gegeben sei die Optimierungsaufgabe*

$$\text{(P)} \qquad \textit{Minimiere} \quad f(x) \quad \textit{auf} \quad M := \{x \in X : x \in C,\ g(x) \in -K\}.$$

Die Voraussetzungen (V1)–(V5) und die Constraint Qualification

$$\text{(CQ)} \qquad \text{cone}\,(g(x^*) + g'(x^*)(C - x^*) + K) = Y$$

seien erfüllt. Ferner gelte eine der folgenden drei Bedingungen:

(a) $Y = \mathbb{R}^m$,

(b) $f'(x^*;\cdot)\colon X \longrightarrow \mathbb{R}$ *ist in* 0 *stetig.*

(c) int $(K) \neq \emptyset$.

Dann existiert $l^* \in Y^*$ *(sogenannter Lagrange-Multiplikator) mit*

1. $l^* \in K^+ := \{l \in Y^* : l(z) \geq 0$ *für alle* $z \in K\}$, $l^*(g(x^*)) = 0$.
2. $f'(x^*; x - x^*) + l^*(g'(x^*)(x - x^*)) \geq 0$ *für alle* $x \in C$.

Jetzt wollen wir noch einige Modifikationen und Spezialfälle zum Satz von Kuhn-Tucker angeben.

Bemerkungen: Betrachtet man neben der Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M := \{x \in X : x \in C,\ g(x) \in -K\}.$$

das scheinbar allgemeinere Problem

$$\text{(P}_0\text{)} \qquad \text{Minimiere} \quad f(x) \quad \text{auf} \quad M_0 := M \cap X_0,$$

wobei $X_0 \subset X$ *offen* ist, so stellt man fest, dass eine lokale Lösung von (P$_0$) auch eine lokale Lösung von (P) ist, sodass auch für (P$_0$) die Aussage des Satzes 7.1 von Kuhn-Tucker gültig bleibt.

Die Bedingung (b) in Satz 7.1 ist z. B. erfüllt, wenn die Zielfunktion f in x^* sogar Fréchet-differenzierbar ist oder wenn $X = \mathbb{R}^n$. Denn eine auf dem $\mathbb{R}^n$ konvexe, reellwertige Funktion ist auf dem ganzen $\mathbb{R}^n$ und damit auch im Nullpunkt stetig (siehe z. B. J. Werner (1984, S. 83)).

Ist $C = X$, $K = \{0\}$ und $Y = \mathbb{R}^m$ in der Optimierungsaufgabe (P), hat man also keine expliziten Restriktionen und nur endlich viele Gleichungen als implizite Restriktionen, so braucht X lediglich ein linearer normierter Raum (und kein Banachraum) zu sein, siehe Korollar 6.22.

Ist in Satz 7.1 speziell $C = X$ und die Hadamard-Variation $f'(x^*;\cdot)$ linear, was z. B. der Fall ist, wenn f in x^* Fréchet-differenzierbar ist, so lautet die Aussage des Satzes von Kuhn-Tucker (wir schreiben jetzt $f'(x^*)\cdot$ statt $f'(x^*;\cdot)$): Es existiert $l^* \in Y^*$ mit

1. $l^* \in K^+$, $l^*(g(x^*)) = 0$,
2. $f'(x^*) + l^* \circ g'(x^*) = 0$,

und dies nennt man (gewöhnlich allerdings nur im endlichdimensionalen Fall: $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$) die *Lagrangesche Multiplikatorenregel*.

Ist int $(K) \neq \emptyset$ in der Optimierungsaufgabe (P), sind also die impliziten Restriktionen vom Ungleichungstyp, so kommt man beim Beweis von Satz 7.1 mit einer Zusatzbedingung *ohne* den Satz von Lyusternik und daher ohne die Vollständigkeit von X und Y aus, ferner können die Differenzierbarkeitsvoraussetzungen abgeschwächt werden. Genauer gilt:

- *Gegeben sei die Optimierungsaufgabe*

  (P) $\qquad$ *Minimiere* $\quad f(x) \quad$ *auf* $\quad M := \{x \in X : x \in C,\ g(x) \in -K\}$.

  Hierbei seien X, Y lineare normierte Räume, $f\colon X \longrightarrow \mathbb{R}$, $g\colon X \longrightarrow Y$, $C \subset X$ nichtleer und konvex, $K \subset Y$ ein nichtleerer konvexer Kegel mit int $(K) \neq \emptyset$. *Ferner sei $x^* \in M$ eine lokale Lösung von (P), f besitze in x^* eine konvexe Hadamard-Variation $f'(x^*;\cdot)$ und g sei in x^* Gâteaux-differenzierbar mit dem Gâteaux-Differential $g'(x^*)$. Schließlich sei*

  $$L_+(M;x^*) := \{p \in X : p \in C - x^*,\ g(x^*) + g'(x^*)p \in -\text{int } (K)\} \neq \emptyset.$$

  Dann existiert $l^* \in Y^*$ mit

  1. $l^* \in K^+ := \{l \in Y^* : l(z) \geq 0$ *für alle* $z \in K\}$, $\quad l^*(g(x^*)) = 0$.
  2. $f'(x^*; x - x^*) + l^*(g'(x^*)(x - x^*)) \geq 0$ *für alle* $x \in C$.

Denn: Im zweiten Schritt haben wir den Satz von Lyusternik benutzt und wollen jetzt auch ohne diesen die Inklusion

$$L_0(M;x^*) := \{p \in X : p \in C - x^*,\ g(x^*) + g'(x^*)p \in -K\} \subset T(M;x^*)$$

nachweisen. Hierzu spielt der *Kegel der in x^* zulässigen Richtungen*

$$F(M;x^*) := \{p \in X : \text{Es existiert } t_0 > 0 \text{ mit } x^* + tp \in M \text{ für alle } t \in [0,t_0]\}$$

eine wichtige Rolle. Offensichtlich ist $F(M;x^*) \subset T(M;x^*)$. Wegen der Abgeschlossenheit des Tangentialkegels (zweiter Teil von Satz 6.2) ist cl $(F(M;x^*)) \subset T(M;x^*)$.

Daher genügt es, $L_0(M;x^*) \subset \text{cl}\,(F(M;x^*))$ nachzuweisen. Sei hierzu $p \in L_0(M;x^*)$, also $p = c - x^*$ mit $c \in C$ und $g(x^*) + g'(x^*)p \in -K$. Nun wähle man ein beliebiges $p_+ \in L_+(M;x^*)$, es ist also $p_+ = c_+ - x^*$ mit $c_+ \in C$, $g(x^*) + g'(x^*)p \in -\text{int}\,(K)$. Nun definieren wir $p(t) := (1-t)p + tp_+$ und zeigen, dass $p(t) \in L_+(M;x^*) \subset F(M;x^*)$ für alle $t \in (0,1]$. Denn $p(t) \in C - x^*$ für $t \in [0,1]$ ist wegen der Konvexität von C trivial, ferner ist

$$g(x^*) + g'(x^*)p(t) = (1-t)[\underbrace{g(x^*) + g'(x^*)p}_{\in -K}] + t[\underbrace{g(x^*) + g'(x^*)p_+}_{-\text{int}\,(K)}] \in -\text{int}\,(K)$$

für $t \in (0,1]$, also $p(t) \in L_+(M;x^*)$ für $t \in (0,1]$. Die Inklusion $L_+(M;x^*) \subset F(M;x^*)$ ist einfach zu beweisen[18], daher ist $p = \lim_{t\to 0+} p(t) \in \text{cl}\,(F(M;x^*))$. Folglich haben wir $L_0(M;x^*) \subset T(M;x^*)$ (ohne den Satz von Lyusternik) bewiesen. Wegen $\text{int}\,(K) \neq \emptyset$ lassen sich die oben definierte Menge Λ_+ und der Punkt $(0,0)$ durch eine abgeschlossene Hyperebene in $\mathbb{R} \times Y$ trennen. Diese Hyperebene ist nicht parallel zu $\mathbb{R}$. Beim Beweis hierfür haben wir oben die Constraint Qualification (CQ) benutzt. So können wir auch hier argumentieren, denn die Bedingung $L_+(M;x^*) \neq \emptyset$ impliziert die Gültigkeit von (CQ). Denn ist $p_+ \in L_+(M;x^*)$, so existiert ein $\epsilon > 0$ mit

$$g(x^*) + g'(x^*)p_+ + B[0;\epsilon] \subset -K.$$

Ist $y \in Y \setminus \{0\}$ beliebig, so ist

$$g(x^*) + g'(x^*)p_+ - \frac{\epsilon}{\|y\|}y = -z \in -K$$

mit $z \in K$. Daher ist

$$y = \frac{\|y\|}{\epsilon}[g(x^*) + g'(x^*)p_+ + z] \in \text{cone}\,(g(x^*) + g'(x^*)(C - x^*) + K),$$

womit die Gültigkeit von (CQ) nachgewiesen ist. Insgesamt ist die obige Aussage bewiesen.

Zum Schluss dieser Bemerkungen zum Satz von Kuhn-Tucker wollen wir auf den endlichdimensionalen Spezialfall eingehen. Hierzu formulieren und beweisen wir die folgende Aussage (siehe z. B. auch J. Werner (2013, Satz 3 in Abschnitt 53)):

[18] Sei $p \in L_+(M;x^*)$. Dann ist $p = c - x^*$ mit $c \in C$ und

$$g(x^*) + \lim_{t\to 0+} \frac{g(x^* + tp) - g(x^*)}{t} = g(x^*) + g'(x^*)p \in -\text{int}\,(K).$$

Daher kann $t_0 \in (0,1]$ so klein gewählt werden, dass

$$g(x^*) + \frac{g(x^* + tp) - g(x^*)}{t} \in -K, \quad t \in (0,t_0].$$

Für $t \in (0,t_0]$ ist wegen der Konvexität von C bzw. K dann $x^* + tp = x^* + t(c - x^*) \in C$ und

$$g(x^* + tp) = (1-t)g(x^*) + t\left(g(x^*) + \frac{g(x^* + tp) - g(x^*)}{t}\right) \in -K,$$

woraus wir $p \in F(M;x^*)$ erhalten.

- *Gegeben sei die Optimierungsaufgabe*

  (P) *Minimiere* $f(x)$ *auf* $M := \left\{ x \in \mathbb{R}^n : \begin{array}{l} g_i(x) = 0,\ i = 1, \ldots, m_0, \\ g_i(x) \leq 0,\ i = m_0 + 1, \ldots, m \end{array} \right\}$.

  Die Zielfunktion $f: \mathbb{R}^n \longrightarrow \mathbb{R}$ *sei in der lokalen Lösung* $x^* \in M$ *stetig partiell differenzierbar, die Abbildung* $g: \mathbb{R}^n \longrightarrow \mathbb{R}^m$ *mit* $g(x) = (g_1(x), \ldots, g_m(x))^T$ *sei auf einer Kugel um* x^* *stetig partiell differenzierbar. Mit*

  $$I^* := \{i \in \{m_0 + 1, \ldots, m\} : g_i(x^*) = 0\}$$

  werden die Indices der in x^* *aktiven Ungleichungsrestriktionen bezeichnet. Ferner gelte*

  (a) $\nabla g_1(x^*), \ldots, \nabla g_{m_0}(x^*)$ *sind linear unabhängig,*

  (b) Es existiert ein $\hat{p} \in \mathbb{R}^n$ *mit* $\begin{array}{ll} \nabla g_i(x^*)^T \hat{p} = 0, & i = 1, \ldots, m_0, \\ \nabla g_i(x^*)^T \hat{p} < 0, & i \in I^*. \end{array}$

  (Die Bedingungen (a) und (b) nennt man die Arrow-Hurwicz-Uzawa oder Mangasarian-Fromowitz Constraint Qualification.) Dann existiert ein $y^* = (y_i^*) \in \mathbb{R}^m$ *mit*

  1. $y_i^* \geq 0$ *und* $y_i^* g_i(x^*) = 0$ *für* $i = m_0 + 1, \ldots, m$.
  2. $\nabla f(x^*) + \sum_{i=1}^{m} y_i^* \nabla g_i(x^*) = 0$.

Denn: Wir wenden Satz 7.1 an mit $X := \mathbb{R}^n$, $Y := \mathbb{R}^m$, $C := \mathbb{R}^n$ und

$$K := \left\{ y = (y_i) \in \mathbb{R}^m : \begin{array}{ll} y_i = 0, & i = 1, \ldots, m_0 \\ y_i \geq 0, & i = m_0 + 1, \ldots, m \end{array} \right\}.$$

Die Glattheitsvoraussetzungen an f und g in Satz 7.1 sind erfüllt und es ist

$$f'(x^*; p) = f'(x^*)p = \nabla f(x^*)^T p, \qquad g'(x^*)p = \begin{pmatrix} \nabla g_1(x^*)^T p \\ \vdots \\ \nabla g_m(x^*)^T p \end{pmatrix}.$$

Der zu K duale Kegel K^+ ist offenbar gegeben durch

$$\begin{array}{rcl} K^+ & := & \{y \in \mathbb{R}^m : y^T z \geq 0 \ \text{ für alle } z \in K\} \\ & = & \{y = (y_i) \in \mathbb{R}^m : y_i \geq 0,\ i = m_0 + 1, \ldots, m\}. \end{array}$$

Zum Beweis der Aussage mit Hilfe von Satz 7.1 bleibt offenbar nachzuweisen, dass aus der Gültigkeit von (a) und (b) die Gültigkeit der Constraint Qualification (CQ) folgt, die in unserem Fall gegeben ist durch

(CQ) $$\text{cone } (g(x^*) + g'(x^*)(\mathbb{R}^n) + K) = \mathbb{R}^m.$$

Genau dies ist aber beim Beweis von Satz 6.21 geschehen. Damit ist die obige Aussage bewiesen. □

Jetzt stellen wir uns noch die Frage, welche notwendigen Optimalitätsbedingungen auch *ohne* die Constraint Qualification (CQ) bewiesen werden können. Als Antwort erhalten wir:

Satz 7.2 (F. John) *Gegeben sei die Optimierungsaufgabe*

$$\text{(P)} \qquad \textit{Minimiere} \quad f(x) \quad \textit{auf} \quad M := \{x \in X : x \in C,\ g(x) \in -K\}.$$

Die Voraussetzungen (V1)–(V5) seien erfüllt. Ferner gelte eine der folgenden drei Bedingungen:

(a) $Y = \mathbb{R}^m$,

(b) $f'(x^*;\cdot)\colon X \longrightarrow \mathbb{R}$ *ist in* 0 *stetig.*

(c) int $(K) \neq \emptyset$.

Schließlich sei cone $(g(x^*) + g'(x^*)(C - x^*) + K)$ *abgeschlossen oder* $Y = \mathbb{R}^m$. *Dann existiert* $(l_0^*, l^*) \in \mathbb{R} \times Y^* \setminus \{(0,0)\}$ *mit*

1. $l_0^* \geq 0, \quad l^* \in K^+ := \{l \in Y^* : l(z) \geq 0$ *für alle* $z \in K\}, \quad l^*(g(x^*)) = 0.$
2. $l_0^* f'(x^*; x - x^*) + l^*(g'(x^*)(x - x^*)) \geq 0$ *für alle* $x \in C$.

Beweis: Ist cone $(g(x^*) + g'(x^*)(C - x^*) + K) = Y$, so folgt die Aussage des Satzes mit $l_0^* := 1$ aus Satz 7.1. Ist dagegen cone $(g(x^*) + g'(x^*)(C - x^*) + K)$ ein echter abgeschlossener Teilkegel von Y oder $Y = \mathbb{R}^m$, so kann ein Punkt in Y, der nicht zu cone $(g(x^*) + g'(x^*)(C - x^*) + K)$ gehört, wegen des starken Trennungssatzes 5.6 in linearen normierten Räumen bzw. des Trennungssatzes im $\mathbb{R}^m$ von diesem Kegel getrennt werden. Hieraus folgt offenbar die Existenz eines $l^* \in Y^* \setminus \{0\}$ mit

$$0 \leq l^*(g(x^*) + g'(x^*)(x - x^*) + z) \quad \text{für alle } x \in C,\ z \in K.$$

Man weist leicht nach, dass 1. und 2. mit $l_0^* := 0$ erfüllt sind. Damit ist der Satz bewiesen. □

8 Optimale Steuerungsprobleme

Es ist das Ziel dieses Abschnitts, einen kleinen Einblick in die Theorie optimaler Steuerungsprobleme zu geben. Wir werden uns mit ziemlich einfachen Aussagen und Beispielen begnügen und folgen hierbei fast wörtlich der Darstellung bei J. Werner (1988), siehe auch A. Kirsch, W. Warth, J. Werner (1978). Durch zwei Beispiele wollen wir zunächst in die Problemstellung bei optimalen Steuerungsproblemen einführen.

8.1 Problemstellung

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 entstehenden Mischung (z. B. der pH-Wert) werde bestimmt durch die Stärke u einer Komponente von A, die zeitlich geändert bzw. gesteuert werden kann. Die zeitliche Änderung von x sei linear in x und u:

$$\dot{x} = ax + bu$$

mit Konstanten $a, b \in \mathbb{R}$. Hierbei wird mit $\dot{x}$ die Ableitung von x nach der Zeit t bezeichnet. Ferner sei zur Anfangszeit 0 der kritische Wert x_0 bekannt:

$$x(0) = x_0.$$

Es ist erstrebenswert, dass x sich im Mittel von einem idealen Wert, etwa 0, möglichst wenig unterscheidet. Die Kosten für die Stärke u seien proportional zu u^2. Man erhält dann etwa die folgende Optimierungsaufgabe:

$$\text{(P)} \quad \left\{ \begin{array}{c} \text{Minimiere} \quad F(x,u) := \dfrac{1}{2}\displaystyle\int_0^T [qx(t)^2 + u(t)^2]\,dt \quad \text{auf} \\ M := \left\{ (x,u) \in C^1[0,T] \times C[0,T] : \begin{array}{ll} \dot{x}(t) = ax(t) + bu(t) & \text{auf } [0,T], \\ x(0) = x_0 & \end{array} \right\}. \end{array} \right.$$

Hierbei ist $q > 0$ vorgegeben. □

Beispiel: Ein Zug der Masse m bewege sich mit vernachlässigbarer Reibung auf einer horizontalen, geraden Strecke. Zur Zeit t befinde er sich im Punkte $x(t)$. Die steuerbare äußere Kraft, die auf den Zug wirke, werde mit $u(t)$ bezeichnet, sodass

$$m\ddot{x} = u(t).$$

Zur Anfangszeit 0 befinde sich der Zug im Punkte x_0 und habe dort die Geschwindigkeit y_0. Die Aufgabe besteht darin, den Zug in minimaler Zeit in einem Zielpunkt, etwa dem Nullpunkt, zu stoppen. Hierbei ist die aufzuwendende äußere Kraft aus technischen Gründen beschränkt, etwa sei

$$|u(t)| \leq 1.$$

Es ist anschaulich klar, dass man auch unstetige Steuerungen u zulassen muss, damit das Problem eine Lösung besitzt, da man mit Sprüngen einer Lösung u^* zu gewissen Schaltzeiten rechnen muss. Man vermutet (zu Recht) sogar, dass eine optimale Steuerung vom sogenannten "bang-bang Typ" ist, also nur die extremalen Werte $+1$ (volle Kraft) und -1 (Vollbremsung) annimmt. Man kommt also auf ganz natürliche Weise zu der Frage, in welchem Funktionenraum eine optimale Steuerung zu suchen ist. Einerseits bieten sich Funktionenräume stückweise stetiger Funktionen an, die aber (wenn man die Sprungstellen nicht kennt) mit der Maximumnorm keinen Banachraum bilden, was für den Beweis notwendiger Optimalitätsbedingungen von großer Wichtigkeit

sein kann. Daher nimmt man i. Allg. als Raum der Steuerungen einen L^∞-Raum. Hier wollen wir uns noch nicht festlegen und formulieren die eben geschilderte Aufgabe als

$$\text{(P)} \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad F(x,u,T) := T \quad \text{auf} \\ M := \left\{ (x,u,T) : \begin{array}{l} m\ddot{x}(t) = u(t),\ |u(t)| \le 1 \ \text{ auf } [0,T] \\ x(0) = x_0,\ \dot{x}(0) = y_0,\ x(T) = 0,\ \dot{x}(0) = 0 \end{array} \right\}. \end{array} \right.$$

Man spricht hier von einem *zeitoptimalen Steuerungsproblem*, da die Aufgabe darin besteht, die Zeit, die man benötigt, um vom vorgegebenen Anfangszustand in den gewünschten Endzustand zu gelangen, zu minimieren. Im Gegensatz zum vorigen Beispiel ist hier also das Zeitintervall, auf dem die Prozessgleichung, hier $m\ddot{x} = u$, betrachtet wird, nicht schon vorher bekannt. Weitere Unterschiede zum vorigen Beispiel bestehen darin, dass hier ein Endzustand vorgegeben ist und zulässige Steuerungen einer Restriktion genügen müssen. Gemeinsam ist beiden Beispielen, dass die Prozessgleichung linear in der Steuerung u und der den Zustand des Systems beschreibenden Variablen x ist. □

Allgemein ist ein optimales Steuerungsproblem durch die folgenden Daten gegeben:

(a) Prozess,
(b) Anfangszustand des Systems,
(c) Endzustand (Ziel) des Systems,
(d) Steuerbereich,
(e) Zielfunktion (Kostenfunktional).

} Diese Daten bestimmen die Menge der zulässigen Lösungen

Wir wollen der eher vagen Erklärung eines optimalen Steuerungsproblems gleich ein mathematisches Modell gegenüberstellen, wobei allerdings schon gesagt werden soll, dass wir nur auf eine spezielle Klasse optimaler Steuerungsprobleme näher eingehen wollen. So betrachten wir z. B. nur die optimale Steuerung von Prozessen, die sich durch Systeme von gewöhnlichen Differentialgleichungen beschreiben lassen. Natürlich ist auch die Steuerung durch partielle Differentialgleichungen denkbar. Auf dieses nach wie vor aktuelle Forschungsgebiet gehen wir aber nicht ein.

(a) Es sei ein *Prozess* gegeben, der durch die Wahl von Steuerparametern bzw. einer Steuerfunktion beeinflusst werden kann. Man stelle sich z. B. die Bewegung eines Raumfahrzeuges vor, welche durch das Zünden von Steuerraketen verändert werden kann.

Der Prozess werde durch ein System von n gewöhnlichen Differentialgleichungen erster Ordnung auf dem Zeitintervall $[t_0, t_1]$ beschrieben, etwa durch $\dot{x} = f(x,u,t)$. Dabei wird die m-Vektorfunktion $u = (u_i)$ *Steuerfunktion* genannt. Die Variable $x = (x_j)$ beschreibt den Zustand des Systems und heißt *Trajektorie*.

(b) Das System, welches durch den Prozess verändert bzw. durch Wahl der Steuerung beeinflusst wird, ist zu einer vorgegebenen Anfangszeit in einem gewissen *Anfangszustand*. Dieser ist i. Allg. fest vorgegeben, er kann aber auch nur "innerhalb gewisser Grenzen" vorgeschrieben sein.

Der Zustand zur Zeit t des durch das Differentialgleichungssystem $\dot{x} = f(x, u, t)$ beschriebenen Prozesses ist durch $x(t)$ gegeben. Der Anfangszustand des Systems zur Anfangszeit t_0 ist daher fest vorgegeben, wenn eine Anfangsbedingung $x(t_0) = x_0$ mit $x_0 \in \mathbb{R}^n$ gegeben ist. Denkbar ist aber auch eine Bedingung der Form $x(t_0) \in Q_0$ mit gegebener Menge $Q_0 \subset \mathbb{R}^n$.

(c) Das System soll durch den gegebenen Prozess von einem (zulässigen) Anfangszustand in einen gewissen *Endzustand* überführt werden. Dieser ist entweder fest oder nur "innerhalb gewisser Grenzen" vorgegeben. Hier sind zwei verschiedene Aufgabenstellungen denkbar: Der Endzustand ist zu einer vorgegebenen *festen Endzeit* zu erreichen oder die Endzeit ist *frei*. Ist letzteres der Fall, so kann der zu erreichende Endzustand bzw. das Ziel auch von der Zeit abhängen.

Der Endzustand ist fest vorgegeben, falls $x(t_1) = x_1$ mit festem $x_1 \in \mathbb{R}^n$ gefordert wird. Der Endzustand ist nur "ungefähr" gegeben, wenn das System zur Endzeit t_1 einer Bedingung $x(t_1) \in Q_1$ mit $Q_1 \subset \mathbb{R}^n$ genügen muss (auch $Q_1 = \mathbb{R}^n$ ist möglich: keine Endbedingung). Dabei ist, je nach Aufgabenstellung, t_1 fest oder frei. Im letzteren Fall kann Q_1 auch von t_1 abhängen.

(d) Aus technischen Gründen sind i. Allg. nicht beliebige Steuerungen möglich sondern nur solche, deren Werte gewissen Beschränkungen genügen bzw. einem vorgegebenen *Steuerbereich* angehören. Wenn die Steuerung z. B. eine auf das System wirkende äußere Kraft ist, so ist diese gewöhnlich beschränkt.

I. Allg. ist eine Menge $\Omega \subset \mathbb{R}^m$, der *Steuerbereich*, gegeben. Diese Menge wird oft als konvex und abgeschlossen oder sogar kompakt vorausgesetzt. Zulässig ist eine Steuerung u nur dann, wenn $u(t) \in \Omega$ auf dem gesamten Zeitintervall $[t_0, t_1]$ ist, wobei dies häufig nur für fast alle Zeiten gefordert wird.

Tripel, die aus Zustand, Steuerung und Endzeit bestehen, heißen *zulässig* für das gegebene optimale Steuerungsproblem, wenn das System durch den Prozess und die gewählte Steuerung vom Anfangszustand in der Endzeit in den gewünschten Endzustand überführt wird und die Werte der Steuerung zum Steuerbereich gehören. Existiert ein zulässiges Tripel, so heißt das System *steuerbar*.

Ein Tripel (x, u, t_1) aus Trajektorie, Steuerung und Endzeit heißt *zulässig*, falls

1. (x, u, t_1) genügt in einem geeignet zu präzisierenden Sinne dem Differentialgleichungssystem
$$\dot{x} = f(x, u, t)$$
auf $[t_0, t_1]$.
2. $x(t_0) \in Q_0$, $x(t_1) \in Q_1$.
3. $u(t) \in \Omega$ auf $[t_0, t_1]$.

Die Menge der zulässigen Tripel wird mit M bezeichnet.

Ist die Endzeit t_1 fixiert, so bestehen die zulässigen Lösungen für das gegebene Steuerungsproblem natürlich nur aus *Paaren* (x, u) mit den Eigenschaften 1.–3.

(e) Auf der Menge der zulässigen Tripel sei ein gegebenes Kostenfunktional zu minimieren. Ein zulässiges Tripel heißt dann *optimal*, wenn es kein zulässiges Tripel mit kleineren Kosten gibt.

In dem gewählten Modell hat das Kostenfunktional gewöhnlich die folgende Form:
$$F(x, u, t_1) = g^0(x(t_0), x(t_1)) + \int_{t_0}^{t_1} f^0(x(t), u(t), t)\, dt$$
mit
$$g^0 \colon \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R},$$
$$f^0 \colon \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \longrightarrow \mathbb{R}.$$

Ein Tripel $(x^*, u^*, t_1^*) \in M$ heißt *optimal*, falls $F(x^*, u^*, t_j) \leq F(x, u, t_1)$ für alle $(x, u, t_1) \in M$.

Es dürfte nun klar sein, wie sich die obigen Beispiele dem angegebenen allgemeinen Steuerungsproblem unterordnen. Wichtig für uns ist, dass es sich bei Problemen der optimalen Steuerung um verhältnismäßig komplizierte Optimierungsaufgaben handelt, bei denen eine Lösung in einem Funktionenraum gesucht wird und i. Allg. unendlich viele Nebenbedingungen auftreten, etwa die Prozessgleichung.

Im folgenden betrachten wir ein optimales Steuerungsproblem auf einem *festen* Zeitintervall $[t_0, t_1]$, welches durch die folgenden Daten gegeben ist:

(a) Die Prozessgleichung ist durch ein in der Trajektorie lineares Differentialgleichungssystem der Form
$$\dot{x} = A(t)x + r(u, t)$$

gegeben. Hierbei seien

$$A\colon [t_0,t_1] \longrightarrow \mathbb{R}^{n\times n}, \qquad r\colon \mathbb{R}^m \times [t_0,t_1] \longrightarrow \mathbb{R}^n$$

stetig, wofür wir auch $A \in C([t_0,t_1],\mathbb{R}^{n\times n})$ bzw. $r \in C(\mathbb{R}^m\times[t_0,t_1],\mathbb{R}^n)$ schreiben. Wir betrachten also lediglich einen speziellen Fall, nämlich in der Trajektorie *lineare* Prozessgleichungen.

(b) Der Anfangszustand des Systems zur Anfangszeit t_0 sei fest vorgegeben:

$$x(t_0) = x_0$$

mit $x_0 \in \mathbb{R}^n$.

(c) Zur Endzeit t_1 soll sich die Trajektorie auf einer Fläche befinden, die durch $G(x) = 0$ beschrieben sei. Als Nebenbedingung hat man also

$$G(x(t_1)) = 0,$$

wobei $G\colon \mathbb{R}^n \longrightarrow \mathbb{R}^k$ mit $k \leq n$. Ist z. B. $G(x) := x - x_1$ mit $x_1 \in \mathbb{R}^n$, so ist der Endzustand vorgeschrieben. Ist $G = 0$, so ist der Endzustand frei.

(d) Der Steuerbereich $\Omega \subset \mathbb{R}^m$ sei nichtleer, konvex und abgeschlossen. Zugelassen ist hier natürlich der Fall, dass $\Omega = \mathbb{R}^m$, die Steuerungen also keinen Beschränkungen genügen müssen.

(e) Das Kostenfunktional sei gegeben durch

$$F(x,u) := g^0(x(t_1)) + \int_{t_0}^{t_1} f^0(x(t),u(t),t)\,dt,$$

wobei $g^0\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ und $f^0\colon \mathbb{R}^n \times \mathbb{R}^m \times [t_0,t_1] \longrightarrow \mathbb{R}$. Der erste Summand bewertet den erreichten Endzustand, der zweite die gewählte Steuerung und die resultierende Trajektorie auf dem Zeitintervall $[t_0,t_1]$.

Bezeichnungen: Sei $n \in \mathbb{N}$. Wir schreiben im folgenden $C_n[t_0,t_1]$ statt $C([t_0,t_1],\mathbb{R}^n)$ für die Menge der auf dem Intervall $[t_0,t_1]$ definierten und stetigen n-Vektorfunktionen. Entsprechend sind die Bezeichnungen $C_n^1[t_0,t_1]$, $C_{n\times n}[t_0,t_1]$ oder auch $C_{n\times n}^1[t_0,t_1]$ zu verstehen. Mit $C_n^{0,s}[t_0,t_1]$ bezeichnen wir die Menge der auf dem Intervall $[t_0,t_1]$ definierten, stückweise stetigen (mit höchstens endlich vielen Sprüngen, die nur im Inneren des Intervalls $[t_0,t_1]$ stattfinden dürfen) n-Vektorfunktionen, entsprechend sei $C_n^{1,s}[t_0,t_1]$ die Menge der auf dem Intervall $[t_0,t_1]$ definierten, stückweIse stetig differenzierbaren n-Vektorfunktionen. Ferner werde mit $L_m^\infty[t_0,t_1]$ die Menge der auf dem Intervall $[t_0,t_1]$ definierten, (Lebesgue) messbaren und im wesentlich beschränkten n-Vektorfunktionen bezeichnet. Hierbei heißt eine Funktion $u\colon [t_0,t_1] \longrightarrow \mathbb{R}^m$ *im wesentlichen beschränkt*, falls eine Konstante $c > 0$ mit $\|u(t)\| \leq c$ für fast alle $t \in [t_0,t_1]$ existiert. Definiert man auf $L_m^\infty[t_0,t_1]$ die *Maximumnorm* $\|\cdot\|_\infty$ durch

$$\|u\|_\infty := \inf\{c \in \mathbb{R} : \|u(t)\| \leq c \text{ für fast alle } t \in [t_0,t_1]\}$$

(hierbei ist $\|\cdot\|$ rechts vom Gleichheitszeichen eine Norm auf dem $\mathbb{R}^m$), so ist $L_m^\infty[t_0,t_1]$ ein Banachraum. □

In der Trajektorie x lineare Prozessgleichungen sind bei der Untersuchung optimaler Steuerungsprobleme angenehm, da man die Abhängigkeit zwischen einer Steuerung u und der resultierenden Trajektorie x bei Vorgabe einer Anfangsbedingung für die Trajektorie mit Hilfe eines *Fundamentalsystems* des homogenen Differentialgleichungssystems geschlossen angeben kann. An diesen aus der Analysis bzw. der Theorie gewöhnlicher Differentialgleichungen bekannten Sachverhalt erinnern wir im folgenden Lemma, das wir ohne Beweis angeben.

Lemma 8.1 *Seien $A \in C_{n\times n}[t_0,t_1]$, $r \in L_n^\infty[t_0,t_1]$ und $x_0 \in \mathbb{R}^n$ gegeben. dann gilt:*

1. *Es existiert genau ein $\Phi \in C_{n\times n}^1[t_0,t_1]$, das sogenannte normierte Fundamentalsystem zu $\dot{x} = A(t)x$, mit*

$$\dot{\Phi}(t) = A(t)\Phi(t) \qquad \text{für alle } t \in [t_0,t_1], \qquad \Phi(t_0) = I.$$

   Für jedes $t \in [t_0,t_1]$ ist $\Phi(t)$ nichtsingulär.

2. *Definiert man $x \in C_n[t_0,t_1]$ durch*

$$x(t) := \Phi(t)x_0 + \int_{t_0}^{t} \Phi(s)^{-1} r(s)\, ds,$$

   so ist x für fast alle $t \in [t_0,t_1]$ differenzierbar und es gilt

$$\dot{x}(t) = A(t)x(t) + r(t) \quad \text{für fast alle } t \in [t_0,t_1], \qquad x(t_0) = x_0.$$

   Ist $r \in C_n^{0,s}[t_0,t_1]$, so ist $x \in C_n^{1,s}[t_0,t_1]$.

Der gleich zu definierende *Steuerungsoperator* ordnet einer Steuerung die resultierende Trajektorie (bei vorgegebenem Anfangszustand) zu.

Definition 8.2 Sei $A \in C_{n\times n}[t_0,t_1]$ und Φ das zu $\dot{x} = A(t)x$ gehörende, nach Lemma 8.1 eindeutig existierende normierte Fundamentalsystem. Ist $r\colon \mathbb{R}^m \times [t_0,t_1] \longrightarrow \mathbb{R}^n$ stetig und $x_0 \in \mathbb{R}^n$, so nennt man die Abbildung

$$S\colon L_m^\infty[t_0,t_1] \longrightarrow C_n[t_0,t_1],$$

die durch

$$S(u)(t) := \Phi(t)x_0 + \int_{t_0}^{t} \Phi(s)^{-1} r(u(s),s)\, ds$$

definiert ist, den zu

$$(*) \qquad \dot{x} = A(t)x + r(u(t),t), \qquad x(t_0) = x_0$$

gehörenden *Steuerungsoperator*. Wegen Lemma 8.1 ist $x := S(u)$ bei gegebenem $u \in L_m^\infty[t_0,t_1]$ die (eindeutige) Lösung von $(*)$.

8.2 Das lokale Pontryaginsche Maximumprinzip für optimale Steuerungsprobleme auf einem festen Zeitintervall

Für das durch die obigen Daten (a)–(e) gegebene optimale Steuerungsproblem formulieren und beweisen wir nun das sogenannte lokale Pontryaginsche Maximumprinzip.

Satz 8.3 (Lokales Pontryaginsches Maximumprinzip) *Gegeben sei das optimale Steuerungsproblem auf dem festen Zeitintervall $[t_0, t_1]$, welches durch die Prozessgleichung $\dot{x} = A(t)x + r(u,t)$, den Anfangszustand $x(t_0) = x_0$, den Endzustand $G(x(t_1)) = 0$, den Steuerbereich $\Omega \subset \mathbb{R}^m$ und die Zielfunktion*

$$F(x,u) := g^0(x(t_1)) + \int_{t_0}^{t_1} f^0(x(t), u(t), t)\, dt$$

definiert ist. Hierbei sei $A \in C_{n\times n}[t_0, t_1]$, ferner seien

$$\begin{array}{rcl} r\colon \mathbb{R}^m \times [t_0,t_1] & \longrightarrow & \mathbb{R}^n, \\ (u,t) & \longrightarrow & r(u,t) \end{array} \qquad \begin{array}{rcl} G\colon \mathbb{R}^n & \longrightarrow & \mathbb{R}^k \\ x & \longrightarrow & G(x) \end{array}$$

mit $k \le n$ sowie

$$\begin{array}{rcl} g^0\colon \mathbb{R}^n & \longrightarrow & \mathbb{R}, \\ x & \longrightarrow & g^0(x) \end{array} \qquad \begin{array}{rcl} f^0\colon \mathbb{R}^n \times \mathbb{R}^m \times [t_0,t_1] & \longrightarrow & \mathbb{R} \\ (x,u,t) & \longrightarrow & f^0(x,u,t) \end{array}$$

stetig und nach x und u stetig partiell differenzierbar. Der Steuerbereich $\Omega \subset \mathbb{R}^m$ sei nichtleer, abgeschlossen und konvex. Das durch die angegebenen Daten gegebene optimale Steuerungsproblem ist äquivalent zu der Optimierungsaufgabe

$$\text{(P)} \quad \left\{ \begin{array}{l} \text{Minimiere} \quad F(u) := g^0(S(u)(t_1)) + \displaystyle\int_{t_0}^{t_1} f^0(S(u)(t), u(t), t)\, dt \quad \text{auf} \\ M := \{u \in L_m^\infty[t_0,t_1] : u(t) \in \Omega \text{ für fast alle } t \in [t_0,t_1],\ G(S(u)(t_1)) = 0\}, \end{array} \right.$$

wobei $S\colon L_m^\infty[t_0,t_1] \longrightarrow C_n[t_0,t_1]$ der zu

$$\dot{x} = A(t)x + r(u(t),t), \qquad x(t_0) = x_0$$

gehörende Steuerungsoperator (siehe Definition 8.2) ist. Sei $u^* \in M$ eine lokale Lösung von (P) und $x^* := S(u^*)$ die zugehörige Trajektorie. Zur Vereinfachung nehmen wir an, es sei sogar $u^* \in C_m^{0,s}[t_0,t_1]$ und damit $x^* \in C_n^{1,s}[t_0,t_1]$. Dann existiert ein Paar $(y_0^*, y^*) \in \mathbb{R} \times \mathbb{R}^k \setminus \{(0,0)\}$ mit $y_0^* \ge 0$ und der Eigenschaft: Ist $\eta^* \in C_n^{1,s}[t_0,t_1]$ die Lösung von

$$\begin{array}{rcll} -\dot{\eta} & = & A(t)^T\eta - y_0^*\nabla_x f^0(x^*(t), u^*(t), t) & \textit{(Adjungierte Gleichung)} \\ -\eta(t_1) & = & y_0^*\nabla g^0(x^*(t_1)) + G_x(x^*(t_1))^T y^* & \textit{(Transversalitätsbedimgung)}, \end{array}$$

so ist

$$(u - u^*(t))^T[r_u(u^*(t),t)^T\eta^*(t) - y_0^*\nabla_u f^0(x^*(t), u^*(t), t)] \le 0$$

für alle $u \in \Omega$, $t \in [t_0,t_1]$. Ist weiter Rang $(G_x(x^*(t_1))) = k$, *der Rang der Funktionalmatrix von G in $x^*(t_1)$ also maximal, so ist $(y_0^*, \eta^*) \ne (0,0)$.*

Beweis: Wir werden Satz 7.2 von F. John anwenden. Hierzu schreiben wir die Optimierungsaufgabe (P) in der Form

$$\text{(P)} \qquad \text{Minimiere} \quad F(u) \quad \text{auf} \quad M := \{u \in U : u \in C,\ g(u) = 0\}.$$

Hierbei ist $U := L_m^\infty[t_0, t_1]$ versehen mit der Maximumnorm ein Banachraum. Die Menge der expliziten Restriktionen ist

$$C := \{u \in L_m^\infty[t_0, t_1] : u(t) \in \Omega \ \text{ für fast alle } t \in [t_0, t_1]\}.$$

Die Restriktionsabbildung $g\colon L_m^\infty[t_0, t_1] \longrightarrow \mathbb{R}^k$ ist gegeben durch

$$g(u) := G(S(u)(t_1))$$

mit dem Steuerungsoperator S. Da nur Gleichungen auftreten, ist $K := \{0\}$. Die Zielfunktion ist

$$F(u) := g^0(S(u)(t_1)) + \int_{t_0}^{t_1} f^0(S(u)(t), u(t), t)\, dt.$$

Jetzt müssen wir die Voraussetzungen des Satzes von F. John nachweisen. Da Ω nichtleer und konvex ist, gilt das entsprechende auch für C. Wir müssen uns noch überlegen, dass mit Ω auch C abgeschlossen ist. Sei hierzu $\{u_j\} \subset C$ eine Folge, die gegen ein Element $u \in L_m^\infty[t_0, t_1]$ konvergiert, für die also

$$\lim_{j\to\infty} \|u_j - u\|_\infty = 0.$$

Nach Definition der Maximumnorm in $L_m^\infty[t_0, t_1]$ und wegen $u_j(t) \in \Omega$ für fast alle $t \in [t_0, t_1]$, existiert für jedes $j \in \mathbb{N}$ eine Menge $E_j \subset [t_0, t_1]$ vom Maße Null mit

$$\|u_j(t) - u(t)\| \le \|u_j - u\|_\infty, \quad u_j(t) \in \Omega \quad \text{für alle } t \in [t_0, t_1] \setminus E_j.$$

Setzt man $E := \bigcup_{j=1}^\infty E_j$, so ist auch E eine Menge vom Maß Null und

$$\lim_{j\to\infty} u_j(t) = u(t), \quad u_j(t) \in \Omega \quad \text{für alle } t \in [t_0, t_1] \setminus E.$$

Wegen der vorausgesetzten Abgeschlossenheit von Ω folgt hieraus $u(t) \in \Omega$ für alle $t \in [t_0, t_1]\setminus E$ bzw. $u \in C$, also die Abgeschlossenheit von C. Die Fréchet-Differenzierbarkeit von F bzw. g in u^* wollen wir nicht zeigen, sondern verweisen auf J. Werner (1988) und geben nur die jeweiligen Fréchet-Differentiale an. Der durch

$$S(u)(t) := \Phi(t)x_0 + \Phi(t) \int_{t_0}^t \Phi(s)^{-1} r(u(s), s)\, ds$$

definierte Steuerungsoperator (hierbei ist Φ das durch $\Phi(t_0) = I$ normierte Fundamentalsystem zu $\dot{x} = A(t)x$) besitzt als Abbildung von $L_m^\infty[t_0, t_1]$ in $C_n[t_0, t_1]$ in u^* das durch

$$S'(u^*)v(t) := \Phi(t) \int_{t_0}^t \Phi(s)^{-1} r_u(s)\, ds$$

gegebene Fréchet-Differential $S'(u^*)$, wobei zur Abkürzung

$$r_u(t) := r_u(u^*(t), t)$$

gesetzt ist. Damit ist dann

$$\begin{aligned} F'(u^*)v &= \nabla g^0(t_1)^T S'(u^*)v(t_1) \\ &\qquad + \int_{t_0}^{t_1} [\nabla_x f^0(t) S'(u^*)v(t) + \nabla_u f^0(t)^T v(t)]\, dt, \\ g'(u^*)v &= G_x(t_1) S'(u^*)v(t_1), \end{aligned}$$

wobei wir

$$\nabla_x f^0(t) := \nabla_x f^0(x^*(t), u^*(t), t), \qquad \nabla_u f^0(t) := \nabla_u f^0(x^*(t), u^*(t), t)$$

und entsprechend

$$\nabla g^0(t_1) := \nabla g^0(x^*(t_1)), \qquad G_x(t_1) := G_x(x^*(t_1))$$

mit $x^* := S(u^*)$ gesetzt haben. Da wir zur Vereinfachung angenommen haben, dass u^* sogar stückweise stetig ist, gilt

$$\nabla_x f^0(\cdot) \in C_n^{0,s}[t_0, t_1], \qquad \nabla_u f^0(\cdot) \in C_m^{0,s}[t_0, t_1], \qquad r_u(\cdot) \in C_{n\times m}^{0,s}[t_0, t_1],$$

wobei die Sprungstellen dieser Funktionen mit denen von u^* übereinstimmen. Eine Anwendung des Satzes 7.2 von F. John liefert die Existenz von $(y_0^*, y^*) \in \mathbb{R}\times\mathbb{R}^k \backslash \{(0,0)\}$ mit[19]

$$(*) \qquad 0 \le y_0^* F'(u^*)(u - u^*) + (y^*)^T g'(u^*)(u - u^*) \quad \text{für alle } u \in C.$$

Nun sei $\eta^* \in C_n^{1,s}[t_0, t_1]$ die Lösung von

$$-\dot\eta = A(t)^T \eta - y_0^* \nabla_x f^0(t), \qquad -\eta(t_1) = y_0^* \nabla g^0(t_1) + G_x(t_1)^T y^*.$$

Für ein beliebiges $u \in C \cap C_m^{0,s}[t_0, t_1]$ ist daher

$$\begin{aligned} 0 &\le y_0^* F'(u^*)(u - u^*) + (y^*)^T g'(u^*)(u - u^*) \\ &= [y_0^* \nabla g^0(t_1) + (y^*)^T G_x(t_1)]^T S'(u^*)(u - u^*)(t_1) \\ &\qquad + y_0^* \int_{t_0}^{t_1} [\nabla_x f^0)t)^T S'(u^*)(u - u^*)(t) + \nabla_u f^0(t)^T (u(t) - u^*(t))]\, dt \\ &\qquad \text{(Einsetzen der Fréchet-Differentiale } F'(u^*) \text{ und } g'(u^*)) \\ &= [y_0^* \nabla g^0(t_1) + (y^*)^T G_x(t_1)]^T \Phi(t_1) \int_{t_0}^{t_1} \Phi(t)^{-1} r_u(t)(u(t) - u^*(t))\, dt \\ &\qquad + y_0^* \int_{t_0}^{t_1} \nabla_x f^0(t)^T \Phi(t) \int_{t_0}^{t} \Phi(s)^{-1} r_u(s)(u(s) - u^*(s))\, ds\, dt \end{aligned}$$

[19] Beachte: $(\mathbb{R}^k)^*$ kann mit $\mathbb{R}^k$ identifiziert werden, da $\sigma\colon \mathbb{R}^k \longrightarrow (\mathbb{R}^k)^*$ mit $\sigma(y) := l$ mit $l(x) := y^T x$ für alle $x \in \mathbb{R}^k$ ein Isomorphismus ist.

$$
\begin{aligned}
&\quad + y_0^* \int_{t_0}^{t_1} \nabla_u f^0(t)^T (u(t) - u^*(t))\, dt \\
&\qquad \text{(Einsetzen von } S'(u^*)) \\
&= -\eta^*(t_1)^T \Phi(t_1) \int_{t_0}^{t_1} \Phi(t)^{-1} r_u(t)(u(t) - u^*(t))\, dt \\
&\quad + \int_{t_0}^{t_1} [\eta^*(t)^T A(t) + \dot\eta^*(t)^T] \Phi(t) \int_{t_0}^{t} \Phi(s)^{-1} r_u(s)(u(s) - u^*(s))\, ds\, dt \\
&\quad + y_0^* \int_{t_0}^{t_1} \nabla_u f^0(t)^T (u(t) - u^*(t))\, dt \\
&\qquad \text{(Berücksichtigung der Definition von } \eta^*) \\
&= -\eta^*(t_1)^T \Phi(t_1) \int_{t_0}^{t_1} \Phi(t)^{-1} r_u(t)(u(t) - u^*(t))\, dt \\
&\quad + \int_{t_0}^{t_1} [\eta^*(t)^T \dot\Phi(t) + \dot\eta^*(t)^T \Phi(t)] \int_{t_0}^{t} \Phi(s)^{-1} r_u(s)(u(s) - u^*(s))\, ds\, dt \\
&\quad + y_0^* \int_{t_0}^{t_1} \nabla_u f^0(t)^T (u(t) - u^*(t))\, dt \\
&\qquad \text{(wegen } \dot\Phi(t) = A(t)\Phi(t)) \\
&= -\eta^*(t_1)^T \Phi(t_1) \int_{t_0}^{t_1} \Phi(t)^{-1} r_u(t)(u(t) - u^*(t))\, dt \\
&\quad + \int_{t_0}^{t_1} \frac{d}{dt} [\eta^*(t)^T \Phi(t)] \int_{t_0}^{t} \Phi(s)^{-1} r_u(s)(u(s) - u^*(s))\, ds\, dt \\
&\quad + y_0^* \int_{t_0}^{t_1} \nabla_u f^0(t)^T (u(t) - u^*(t))\, dt \\
&= -\eta^*(t_1)^T \Phi(t_1) \int_{t_0}^{t_1} \Phi(t)^{-1} r_u(t)(u(t) - u^*(t))\, dt \\
&\quad + \eta^*(t_1)^T \Phi(t_1) \int_{t_0}^{t_1} \Phi(t)^{-1} r_u(t)(u(t) - u^*(t))\, dt \\
&\quad - \int_{t_0}^{t_1} \eta^*(t)^T r_u(t)(u(t) - u^*(t))\, dt \\
&\quad + y_0^* \int_{t_0}^{t_1} \nabla_u f^0(t)^T (u(t) - u^*(t))\, dt \\
&\qquad \text{(Partielle Integration)} \\
&= \int_{t_0}^{t_1} (u(t) - u^*(t))^T [-r_u(t)^T \eta^*(t) + y_0^* \nabla_u f^0(t)]\, dt.
\end{aligned}
$$

Also ist

$$
(**) \qquad \int_{t_0}^{t_1} (u(s) - u^*(s))^T p(s)\, ds \le 0 \quad \text{für alle } u \in C \cap C_m^{0,s}[t_0, t_1],
$$

wobei wir zur Abkürzung $p \in C_m^{0,s}[t_0, t_1]$ durch

$$p(s) := r_u(s)^T \eta(s) - y_0^* \nabla_u f^0(s)$$

definiert haben. Dies ist sozusagen das lokale Pontryaginsche Maximumprinzip in integrierter Form. Um das behauptete eigentliche lokale Pontryaginsche Maximumprinzip zu beweisen, gebe man sich $t \in [t_0, t_1]$ und $u \in \Omega$ beliebig vor. Da u^* höchstens endlich viele Sprungstellen im Innern des Intervalls $[t_0, t_1]$ besitzt, liegen für alle hinreichend kleinen $\epsilon > 0$ in $(t, t+\epsilon)$ keine Sprungstellen von u^* und damit auch keine Sprungstellen von p, falls $t \in [t_0, t_1)$. Für diese ϵ definiere man $u_\epsilon \in C \cap C_m^{0,s}[t_0, t_1]$ durch

$$u_\epsilon(s) := \begin{cases} u, & s \in (t, t+\epsilon), \\ u^*(s), & s \notin (t, t+\epsilon). \end{cases}$$

Einsetzen in $(**)$ und Multiplikation mit $1/\epsilon$ ergibt mit $\epsilon \to 0+$, dass

$$\begin{aligned} 0 &\geq \frac{1}{\epsilon} \int_{t_0}^{t_1} (u_\epsilon(s) - u^*(s))^T p(s)\, ds \\ &= \frac{1}{\epsilon} \int_t^{t+\epsilon} (u - u^*(s))^T p(s)\, ds \\ &\to (u - u^*(t+0))^T p(t+0). \end{aligned}$$

Entsprechend liegen für $t \in (t_0, t_1]$ und alle hinreichend kleinen $\epsilon > 0$ in $(t-\epsilon, t)$ keine Sprungstellen von $u*$ und damit keine Sprungstellen von p. Definiert man für diese ϵ, ähnlich wie oben, $u_\epsilon \in C \cap C_m^{0,s}[t_0, t_1]$ durch

$$u_\epsilon(s) := \begin{cases} u, & s \in (t-\epsilon, t), \\ u^*(s), & s \notin (t-\epsilon, t), \end{cases}$$

so erhält man mit $\epsilon \to 0+$, dass

$$\begin{aligned} 0 &\geq \frac{1}{\epsilon} \int_{t_0}^{t_1} (u_\epsilon(s) - u^*(s))^T p(s)\, ds \\ &= \frac{1}{\epsilon} \int_{t-\epsilon}^{t} (u - u^*(s))^T p(s)\, ds \\ &\to (u - u^*(t-0))^T p(t-0). \end{aligned}$$

Damit ist das lokale Pontryaginsche Maximumprinzip schließlich bewiesen. Ist schließlich Rang $(G_x(x^*(t_1))) = k$, sind also die $k \leq n$ Zeilen von $G_x(x^*(t_1))$ bzw. die k Spalten von $G_x(x^*(t_1))^T$ linear unabhängig, so würde aus $(y_0^*, \eta^*) = (0, 0)$ wegen der Endbedingung bei der Definition von η^* insbesondere auch $G_x(x^*(t_1))^T y^* = 0$ und damit $y^* = 0$ folgen, ein Widerspruch zu $(y_0^*, y^*) \neq (0, 0)$. Insgesamt ist der Satz damit bewiesen. □

Beispiel 8.4 Wir betrachten ein optimales Steuerungsproblem, das durch die folgenden Daten gegeben ist:

(a) Die Prozessgleichung ist $\dot{x} = x + u$, diese betrachten wir auf dem festen Intervall $[0, 1]$.

(b) Die Anfangsbedingung ist $x(0) = x_0$ mit gegebenem $x_0 \in \mathbb{R}$.

(c) Die Endbedingung ist $x(1) = 0$.

(d) Der Steuerbereich sei $\Omega := \mathbb{R}$.

(e) Die Zielfunktion ist

$$F(u) := \frac{1}{4} \int_0^1 u(t)^4 \, dt.$$

Wir wenden Satz 8.3 an. Hiernach existiert zu einer (lokal) optimalen Steuerung $u^* \in C^{0,s}[0, 1]$ ein Paar $(y_0^*, y^*) \in \mathbb{R} \times \mathbb{R} \setminus \{(0, 0)\}$ mit $y_0^* \geq 0$ und der Eigenschaft: Ist η^* die Lösung von

$$-\dot{\eta} = \eta, \qquad -\eta(1) = y^*,$$

so ist

$$(u - u^*(t))[\eta^*(t) - y_0^* u^*(t)^3] \leq 0 \quad \text{für alle } u \in \mathbb{R},\ t \in [0, 1]$$

und damit

$$\eta^*(t) = y_0^* u^*(t)^3 \quad \text{für alle } t \in [0, 1].$$

Wäre $y_0^* = 0$, so wäre $\eta^* = 0$ und dann auch $y^* = 0$, ein Widerspruch zu $(y_0^*, y^*) \neq (0, 0)$. Daher ist o. B. d. A. $y_0^* = 1$, also

$$\eta^*(t) = \eta_0^* e^{-t} = u^*(t)^3$$

bzw.

$$u^*(t) = (\eta_0^* e^{-t})^{1/3}.$$

Um die optimale Steuerung u^* zu erhalten, braucht also nur noch der Anfangszustand η_0^* der adjungierten Trajektorie berechnet zu werden. Einsetzen in die Prozessgleichung zeigt, dass man die optimale Trajektorie x^* als Lösung von

$$\dot{x} = x + (\eta_0^* e^{-t})^{1/3}, \qquad x(0) = x_0, \quad x(1) = 0$$

gewinnt. Berücksichtigt man zunächst nur die Anfangsbedingung $x(0) = x_0$ und benutzt man, dass $\Phi(t) := e^t$ ein "Fundamentalsystem" zu $\dot{x} = x$ ist, so erhält man

$$\begin{aligned} x^*(t) &= e^t x_0 + e^t \int_0^t e^{-s} (\eta_0^* e^{-s})^{1/3} \, ds \\ &= e^t x_0 + (\eta_0^*)^{1/3} e^t \int_0^t e^{-4s/3} \, ds \\ &= e^t x_0 - \frac{3}{4} (\eta_0^*)^{1/3} (e^{-t/3} - e^t). \end{aligned}$$

Die Endbedingung liefert

$$(\eta_0^*)^{1/3} = -\frac{4}{3}(1 - e^{-4/3})^{-1} x_0,$$

sodass die optimale Steuerung u^* durch

$$u^*(t) = -\frac{4}{3}(1 - e^{-4/3})^{-1} x_0 e^{-t/3}$$

gegeben ist. □

8.2.1 Spezialfall: Keine Endbedingung

Wir haben Satz 8.3 bewiesen, indem wir auf die Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad F(u) \quad \text{auf} \quad M := \{u \in U : u \in C,\ g(u) = 0\}$$

den Satz 7.2 von F. John anwandten. Hierbei ist

$$U := L_m^\infty[t_0, t_1], \qquad C := \{u \in U : u(t) \in \Omega \ \text{ für fast alle } t \in [t_0, t_1]\},$$

weiter sind die Abbildungen $F\colon U \longrightarrow \mathbb{R}$ und $g\colon U \longrightarrow \mathbb{R}^k$ durch

$$F(u) := g^0(S(u)(t_1)) + \int_{t_0}^{t_1} f^0(S(u)(t), u(t), t)\, dt, \qquad g(u) := G(S(u)(t_1))$$

mit dem Steuerungsoperator S gegeben, wobei $x := S(u)$ die Lösung der Anfangswertaufgabe

$$\dot{x} = A(t)x + r(u(t), t), \quad x(t_0) = x_0$$

ist. Ist die Constraint Qualification

$$\text{cone}\,(g'(u^*)(C - x^*)) = \mathbb{R}^k$$

erfüllt, so kann der Satz 7.1 von Kuhn-Tucker angewandt werden und o. B. d. A. $y_0^* = 1$ angenommen werden. Ist *keine* Endbedingung vorgeschrieben, handelt es sich bei (P) also um die Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad F(u) \quad \text{auf} \quad M := \{u \in U : u \in C\},$$

so erhält man als notwendige Bedingung für die lokale Optimalität von

$$u^* \in C_m^{0,s}[t_0, t_1] \cap M$$

(ganz ohne den Satz von Kuhn-Tucker!), dass

$$0 \le F'(u^*)(u - u^*) \quad \text{für alle } u \in C.$$

Eine Argumentation wie im Beweis von Satz 8.3, dem lokalen Pontryaginschen Maximumprinzip, liefert (setze $y_0^* := 1$ und $y^* := 0$ in $(*)$ auf Seite 89):

- Ist $\eta^* \in C_n^{1,s}[t_0, t_1]$ die Lösung von

  $$-\dot{\eta} = A(t)^T \eta - \nabla_x f^0(x^*(t), u^*(t), t), \qquad -\eta(t_1) = \nabla g^0(x^*(t_1)),$$

  wobei $x^* := S(u^*)$ die zu u^* gehörende Trajektorie ist, so ist

  $$(u - u^*(t))^T [r_u(u^*(t), t)^T \eta^*(t) - \nabla_u f^0(x^*(t), u^*(t), t)] \le 0$$

  für alle $u \in \Omega$, $t \in [t_0, t_1]$.

Beispiel 8.5 Wir betrachten ein optimales Steuerungsproblem mit den folgenden Daten:

(a) Die Prozessgleichung ist $\dot{x} = ax + bu$ mit $b \neq 0$, diese betrachten wir auf dem festen Zeitintervall $[0, T]$ mit gegebenem $T > 0$.

(b) Die Anfangsbedingung ist $x(0) = x_0$.

(c) Es ist keine Endbedingung gegeben.

(d) Der Steuerbereich sei $\Omega := \mathbb{R}$, an zulässige Steuerungen werden also keine Bedingungen gestellt.

(e) Die Zielfunktion ist

$$F(x, u) := \frac{1}{2} \int_0^T [qx(t)^2 + u(t)^2]\, dt$$

mit $q > 0$.

Die adjungierte Gleichung sowie die Transversalitätsbedingung lauten

$$(*) \qquad -\dot{\eta} = a\eta - qx^*(t), \qquad -\eta(T) = 0,$$

sei η^* die Lösung von $(*)$. Hierbei ist x^* die zu der (lokal) optimalen Steuerung u^* gehörende Trajektorie. Das lokale Pontryaginsche Maximumprinzip liefert die Aussage, dass

$$(u - u^*(t))[b\eta^*(t) - u^*(t)] \leq 0 \quad \text{für alle } u \in \mathbb{R},\ t \in [0, T].$$

Hieraus aber folgt $u^*(t) = b\eta^*(t)$. Daher gewinnt man (x^*, η^*) als Lösung von

$$(**) \qquad \begin{pmatrix} \dot{x} \\ \dot{\eta} \end{pmatrix} = \underbrace{\begin{pmatrix} a & b^2 \\ q & -a \end{pmatrix}}_{=:C} \begin{pmatrix} x \\ \eta \end{pmatrix}, \qquad \begin{pmatrix} x(0) \\ \eta(T) \end{pmatrix} = \begin{pmatrix} x_0 \\ 0 \end{pmatrix}.$$

Hat man hieraus (x^*, η^*) bestimmt, so gewinnt man die optimale Steuerung u^* aus $u^* = b\eta^*$. Aus der Theorie linearer, gewöhnlicher Differentialgleichungssysteme weiß man, wie man die Lösung von $(**)$ berechnen kann. Zunächst bestimme man das Fundamentalsystem $\Phi(t) := e^{Ct}$ (siehe Lemma 8.1), dann macht man für die Lösung von $(**)$ den Ansatz

$$\begin{pmatrix} x^*(t) \\ \eta^*(t) \end{pmatrix} = e^{Ct} \begin{pmatrix} \xi_1^* \\ \xi_2^* \end{pmatrix}$$

und berechnet die noch unbekannten Konstanten ξ_1^*, ξ_2^* aus der Anfangs-Endbedingung $x^*(0) = x_0$, $\eta^*(T) = 0$. Die Eigenwerte von C sind $\pm\lambda$ mit $\lambda := \sqrt{a^2 + qb^2}$. Wegen

$$C = \begin{pmatrix} 1 & 1 \\ \frac{q}{a+\lambda} & \frac{q}{a-\lambda} \end{pmatrix} \begin{pmatrix} \lambda & 0 \\ 0 & -\lambda \end{pmatrix} \begin{pmatrix} 1 & 1 \\ \frac{q}{a+\lambda} & \frac{q}{a-\lambda} \end{pmatrix}^{-1}$$

ist

$$\begin{aligned}
e^{Ct} &= \begin{pmatrix} 1 & 1 \\ \frac{q}{a+\lambda} & \frac{q}{a-\lambda} \end{pmatrix} \begin{pmatrix} e^{\lambda t} & 0 \\ 0 & e^{-\lambda t} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ \frac{q}{a+\lambda} & \frac{q}{a-\lambda} \end{pmatrix}^{-1} \\
&= \frac{a^2-\lambda^2}{2q\lambda} \begin{pmatrix} 1 & 1 \\ \frac{q}{a+\lambda} & \frac{q}{a-\lambda} \end{pmatrix} \begin{pmatrix} e^{\lambda t} & 0 \\ 0 & e^{-\lambda t} \end{pmatrix} \begin{pmatrix} \frac{q}{a-\lambda} & -1 \\ -\frac{q}{a+\lambda} & 1 \end{pmatrix} \\
&= \frac{a^2-\lambda^2}{2q\lambda} \begin{pmatrix} \frac{q}{a-\lambda}e^{\lambda t} - \frac{q}{a+\lambda}e^{-\lambda t} & -(e^{\lambda t} - e^{-\lambda t} \\ \frac{q^2}{a^2-\lambda^2}(e^{\lambda t} - e^{-\lambda t}) & -\frac{q}{a+\lambda}e^{\lambda t} + \frac{q}{a-\lambda}e^{-\lambda t} \end{pmatrix} \\
&= \frac{1}{2\lambda} \begin{pmatrix} (a+\lambda)e^{\lambda t} - (a-\lambda)e^{-\lambda t} & b^2(e^{\lambda t} - e^{-\lambda t}) \\ q(e^{\lambda t} - e^{-\lambda t}) & -(a-\lambda)e^{\lambda t} + (a+\lambda)e^{-\lambda t} \end{pmatrix} \\
&= \frac{1}{\lambda} \begin{pmatrix} a\sinh\lambda t + \lambda\cosh\lambda t & b^2\sinh\lambda t \\ q\sinh\lambda t & -a\sinh\lambda t + \lambda\cosh\lambda t \end{pmatrix}.
\end{aligned}$$

Aus $x^*(0) = x_0$, $\eta^*(T) = 0$ ergibt sich, dass ξ_1^*, ξ_2^* aus

$$\xi_1^* := x_0, \qquad \xi_2^* := \frac{q\sinh\lambda T}{a\sinh\lambda T - \lambda\cosh\lambda T} x_0$$

zu berechnen sind. Damit sind die optimale Trajektorie x^* und die optimale adjungierte Trajektorie η^* bestimmt:

$$x^*(t) := \frac{ax_0 + b^2\xi_2^*}{\lambda}\sinh\lambda t + x_0\cosh\lambda t, \qquad \eta^*(t) := \frac{qx_0 - a\xi_2^*}{\lambda}\sinh\lambda t + \xi_2^*\cosh\lambda t.$$

Die optimale Steuerung u^* ist durch

$$u^*(t) := b\eta^*(t)$$

gegeben. □

Beispiel 8.6 Ein optimales Steuerungsproblem sei durch die folgenden Daten gegeben:

(a) Die Prozessgleichung ist $\dot{x} = u$, diese betrachten wir auf dem festen Zeitintervall $[0, 1]$.

(b) Die Anfangsbedingung ist $x(0) = x_0$ mit vorgegebenem $x_0 \in \mathbb{R}$.

(c) Es ist keine Endbedingung gegeben.

(d) Der Steuerbereich sei $\Omega := [-1, 1]$.

(e) Die Zielfunktion ist

$$F(x, u) := \frac{1}{2}\int_0^1 [x(t)^2 + u(t)^2]\, dt.$$

Diesmal lauten die adjungierte Gleichung sowie die Transversalitätsbedingung

$$(*) \qquad -\dot{\eta} = -x^*(t), \qquad -\eta(1) = 0,$$

sei η^* die Lösung von $(*)$. Offenbar ist

$$\eta^*(t) = -\int_t^1 x^*(s)\, ds.$$

Hierbei ist x^* die zu der (lokal) optimalen Steuerung u^* gehörende Trajektorie. Das lokale Pontryaginsche Maximumprinzip liefert die Aussage, dass

$$(u - u^*(t))[\eta^*(t) - u^*(t)] \leq 0 \quad \text{für alle } u \in [-1,1],\ t \in [0,1].$$

Hieraus schließen wir:

(1) Ist $|\eta^*(t)| \leq 1$, so ist $u^*(t) = \eta^*(t)$.
Denn: Man setze $u := \eta^*(t)$ und erhält $(\eta^*(t) - u^*(t))^2 \leq 0$ bzw. $u^*(t) = \eta^*(t)$.

(2) Ist $\eta^*(t) \geq 1$, so ist $u^*(t) = 1$.
Denn: Man setze $u := 1$ und erhält

$$(1 - u^*(t))^2 \leq \underbrace{(1 - u^*(t))}_{\geq 0}\underbrace{[\eta^*(t) - u^*(t)]}_{\geq 0} \leq 0$$

bzw. $u^*(t) = 1$.

(3) Ist $\eta^*(t) \leq -1$, so ist $u^*(t) = -1$.
Denn: Man setze $u := -1$ und erhält

$$(-1 - u^*(t))^2 \leq \underbrace{(-1 - u^*(t))}_{\leq 0}\underbrace{[\eta^*(t)}_{\leq -1} - u^*(t)] \leq 0$$

bzw. $u^*(t) = -1$.

Es ist $\eta^*(1) = 0$. Wir definieren $\hat{t} \in [0,1)$ dadurch, dass wir $\hat{t} := 0$ setzen, falls $|\eta^*(t)| < 1$ für alle $t \in (0,1]$, andernfalls sei $\hat{t} \in (0,1)$ die erste Nullstelle von $|\eta^*(t)| - 1$ "links" von 1, es sei also $\eta^*(\hat{t})| = 1$ und $|\eta^*(t)| < 1$ für alle $t \in (\hat{t}, 1]$. Wir unterscheiden jetzt drei Fälle.

(α) Es ist $\hat{t} = 0$, also $|\eta^*(t)| < 1$ für alle $t \in (0,1]$.
Dann ist $\dot{x}^*(t) = u^*(t)$, $\dot{\eta}^*(t) = \dot{u}^*(t) = x^*(t)$ für $t \in (0,1]$, also

$$\ddot{x}^*(t) - x^*(t) = 0 \quad \text{für } t \in (0,1], \quad x^*(0) = x_0, \quad \dot{x}^*(1) = 0.$$

Hieraus folgt

$$x^*(t) = \frac{x_0}{\cosh 1} \cosh(1 - t)$$

und anschließend

$$\eta^*(t) = -\int_t^1 x^*(s)\,ds = -\frac{x_0}{\cosh 1}\sinh(1-t).$$

Man beachte, dass $|\eta^*(t)| < 1$ für alle $t \in (0,1]$ genau dann, wenn

$$|x_0| \le \frac{\cosh 1}{\sinh 1} = \coth 1 \approx 1.313035285499331.$$

(β) Es ist $\hat{t} \in (0,1)$ und $\eta^*(\hat{t}) = 1$.

Dann ist

$$\ddot{x}^*(t) - x^*(t) = 0 \quad \text{für } t \in (\hat{t},1], \quad \eta^*(\hat{t}) = -\int_{\hat{t}}^1 x^*(s)\,ds, \quad \dot{x}^*(1) = 0.$$

Aus der ersten und der letzten Bedingung erhält man, dass $x^*(t) = c_1\cosh(1-t)$. Hierbei ist die noch unbekannte Konstante c_1 aus

$$1 = -c_1\int_{\hat{t}}^1 \cosh(1-s)\,ds = c_1\sinh(1-s)\,\Big|_{s=\hat{t}}^{s=1} = -c_1\sinh(1-\hat{t})$$

zu bestimmen. Also ist

$$x^*(t) = -\frac{\cosh(1-t)}{\sinh(1-\hat{t})}, \quad u^*(t) = \dot{x}^*(t) = \frac{\sinh(1-t)}{\sinh(1-\hat{t})}, \quad t \in [\hat{t},1].$$

Nun müssen wir noch x^* und u^* auf $[0,\hat{t}]$ berechnen. Wir wollen uns überlegen, dass $\eta^*(t) \ge 1$ für alle $t \in [0,\hat{t}]$. Angenommen, dies wäre schon bewiesen. Wegen Fall (2) oben hätten wir $u^*(t) = 1$ und damit $x^*(t) = x^*(\hat{t}) + (t-\hat{t})$ für alle $t \in [0,\hat{t}]$, wir hätten also u^* und x^* auf dem ganzen Intervall $[0,1]$ bestimmt. Es ist $\eta^*(\hat{t}) = 1$ und $\dot{\eta}^*(\hat{t}) = x^*(\hat{t}) < 0$. Sei $\tilde{t} \in [0,\hat{t})$ minimal mit $\eta^*(t) \ge 1$ für alle $t \in [\tilde{t},\hat{t}]$. Dann ist $u^*(t) = 1$ für $t \in [\tilde{t},\hat{t}]$, also $\dot{x}^*(t) = 1$ und folglich $x^*(t) = x^*(\hat{t}) + (t-\hat{t})$ für $t \in [\tilde{t},\hat{t}]$. Dann ist

$$\begin{aligned}
\eta^*(\tilde{t}) &= -\int_{\tilde{t}}^1 x^*(s)\,ds \\
&= \underbrace{-\int_{\hat{t}}^1 x^*(s)\,ds}_{=1} - \int_{\tilde{t}}^{\hat{t}} x^*(s)\,ds \\
&= 1 - \int_{\tilde{t}}^{\hat{t}} [x^*(\hat{t}) + (s-\hat{t})]\,ds \\
&= 1 - x^*(\hat{t})(\hat{t}-\tilde{t}) - \frac{1}{2}(s-\hat{t})^2\,\Big|_{s=\tilde{t}}^{s=\hat{t}} \\
&= 1 \underbrace{-x^*(\hat{t})}_{>0}\underbrace{(\hat{t}-\tilde{t})}_{>0} + \underbrace{\frac{1}{2}(\hat{t}-\tilde{t})^2}_{>0} \\
&> 1.
\end{aligned}$$

Aus der Definition von $\tilde{t}$ folgt $\tilde{t} = 0$. Damit sind x^* und u^* auf $[0,1]$ berechnet, wir erhalten

$$x^*(t) = \begin{cases} -\dfrac{\cosh(1-t)}{\sinh(1-\hat{t})}, & t \in [\hat{t},1], \\ x^*(\hat{t}) + (t-\hat{t}), & t \in [0,\hat{t}), \end{cases} \quad u^*(t) = \begin{cases} \dfrac{\sinh(1-t)}{\sinh(1-\hat{t})}, & t \in [\hat{t},1], \\ 1, & t \in [0,\hat{t}). \end{cases}$$

Hierbei wird $\hat{t} \in (0,1)$ aus der Anfangsbedingung $x^*(0) = x_0$ bzw.

$$-\underbrace{\frac{\cosh(1-\hat{t})}{\sinh(1-\hat{t})}}_{\coth(1-\hat{t})} - \hat{t} = x_0$$

bestimmt. Diese Gleichung hat eine eindeutige Lösung $\hat{t} \in (0,1)$ falls

$$x_0 \in (-\infty, -\coth 1).$$

Dies sieht man leicht ein, wenn man $f(t) := -\coth(1-t) - t$ definiert und beachtet, dass $f(0) = -\coth 1$, $\lim_{t\to 1} f(t) = -\infty$ und f auf $(0,1)$ wegen $f'(t) = -1/\sinh^2(1-t) - 1$ monoton fallend ist.

(γ) Es ist $\hat{t} \in (0,1)$ und $\eta^*(\hat{t}) = -1$.

Es kann ganz ähnlich wie unter (β) geschlossen werden. Als Ergebnis erhält man

$$x^*(t) = \begin{cases} \dfrac{\cosh(1-t)}{\sinh(1-\hat{t})}, & t \in [\hat{t},1], \\ x^*(\hat{t}) + (\hat{t}-t), & t \in [0,\hat{t}), \end{cases} \quad u^*(t) = \begin{cases} -\dfrac{\sinh(1-t)}{\sinh(1-\hat{t})}, & t \in [\hat{t},1], \\ -1, & t \in [0,\hat{t}), \end{cases}$$

wobei $\hat{t} \in (0,1)$ aus der Anfangsbedingung $x^*(0) = x_0$ bzw.

$$\coth(1-\hat{t}) + \hat{t} = x_0$$

zu bestimmen ist. Diese Gleichung besitzt für

$$x_0 \in (\coth 1, +\infty)$$

genau eine Lösung.

Damit ist das oben angegebene optimale Steuerungsproblem gelöst. □

8.3 Ein zeitoptimales Steuerungsproblem mit linearem Prozess

Wir betrachten ein optimales Steuerungsproblem, bei dem die Endzeit t_1 *nicht* fest vorgegeben ist, sondern so zu bestimmen ist, dass ein gewisses Ziel in minimaler Zeit zu erreichen ist. Zur Vereinfachung betrachten wir nur lineare Prozesse. Das optimale Steuerungsproblem sei also durch die folgenden Daten gegeben:

(a) Die Prozessgleichung ist

$$\dot{x} = A(t)x + B(t)u,$$

diese betrachten wir auf einem Intervall $[t_0, T]$ mit der Anfangszeit t_0 und hinreichend großem $T > t_0$. Wir setzen voraus, dass $A \in C_{n\times n}[t_0, T]$, $B \in C_{n\times m}[t_0, T]$.

(b) Die Anfangsbedingung ist $x(t_0) = x_0$ mit vorgegebenem $x_0 \in \mathbb{R}^n$.

(c) Das Ziel ist eine von $t \in [t_0, T]$ *stetig* abhängende kompakte Menge $G(t) \subset \mathbb{R}^n$, was im Anschluss erklärt wird.

(d) Der Steuerbereich $\Omega \subset \mathbb{R}^m$ ist konvex und kompakt.

Der Steuerungsoperator S ordnet einer Steuerung $u \in L_m^\infty[t_0, T]$ die Lösung x der Prozessgleichung zu, welche der gegebenen Anfangsbedingung genügt. Es ist also

$$x(t) = S(u)(t) = \Phi(t)x_0 + \Phi(t) \int_{t_0}^{t} \Phi(s)^{-1} B(s)u(s)\, ds,$$

wobei Φ das normierte Fundamentalsystem zu $\dot{x} = A(t)x$ ist, siehe Definition 8.2. Die Menge der zulässigen Lösungen des durch die oben angegebenen Daten definierten Steuerungsproblems ist

$$M := \left\{ (u, t_1) \in L_m^\infty[t_0, T] \times \mathbb{R} : \begin{array}{c} t_1 \in (t_0, T),\ u(t) \in \Omega \ \text{ für fast alle } t \in [t_0, t_1], \\ S(u)(t_1) \in G(t_1) \end{array} \right\}.$$

(e) Die Zielfunktion ist

$$F(u, t_1) := t_1 - t_0 = \int_{t_0}^{t_1} 1\, dt.$$

Als zu diesen Daten gehörende Optimierungsaufgabe erhalten wir also

$$\text{(P)} \qquad \text{Minimiere} \quad F(u, t_1), \quad (u, t_1) \in M.$$

Mit $K(t)$ bezeichnen wir die in der Zeit $t \in [t_0, T]$ durch eine zulässige Steuerung *erreichbare* Menge. Hierbei nennen wir eine Steuerung *zulässig*, wenn ihre Werte wenigstens fast überall im Steuerbereich Ω liegen. Also ist

$$K(t) := \begin{cases} \{S(u)(t) : u \in L_m^\infty[t_0, t] : u(s) \in \Omega \ \text{ für fast alle } s \in [t_0, t]\}, & t \in (t_0, T], \\ \{x_0\}, & t = t_0. \end{cases}$$

Ist $K(t) \cap G(t) \neq \emptyset$ für ein $t \in [t_0, T]$, so kann der Prozess vom Anfangszustand x_0 durch eine zulässige Steuerung in der Zeit t in das Ziel $G(t)$ gesteuert werden. O. B. d. A. können wir $x_0 \notin G(t_0)$ annehmen. Das zeitoptimale Steuerungsproblem (P) ist daher äquivalent dazu, das kleinste $t_1 \in (t_0, T]$ mit $K(t_1) \cap G(t_1) \neq \emptyset$ zu bestimmen.

8.3.1 Der Hausdorff-Abstand kompakter Teilmengen des $\mathbb{R}^n$

Bevor wir gleich auf die Existenz einer Lösung des oben definierten zeitoptimalen Steuerungsproblems eingehen, müssen wir erklären, was wir unter einer stetigen Abbildung $t \mapsto G(t)$ verstehen. Entscheidend hierfür ist die folgende Definition.

Definition 8.7 Sei X die Menge der nichtleeren, kompakten Teilmengen des $\mathbb{R}^n$. Für $P, Q \in X$ definiere man

$$\rho(P,Q) := \max\Big(\max_{x\in P} d(x,Q), \max_{y\in Q} d(y,P)\Big),$$

wobei z. B. $d(x,Q) := \min_{y\in Q} \|x-y\|$ mit einer festen (z. B. der euklidischen) Norm $\|\cdot\|$ auf dem $\mathbb{R}^n$ den *Abstand* des Punktes x zur Menge Q bezeichnet[20]. Dann heißt $\rho(P,Q)$ der *Hausdorff-Abstand* der Mengen P und Q.

Satz 8.8 *Mit den Bezeichnungen von Definition 8.7 ist (X,ρ) ein metrischer Raum. D. h. $\rho\colon X \times X \longrightarrow \mathbb{R}$ ist eine Abbildung mit*

1. *Für $P, Q \in X$ ist $\rho(P,Q) = 0$ genau dann, wenn $P = Q$.*
2. *Es ist $\rho(P,Q) = \rho(Q,P)$ für alle $P, Q \in X$.*
3. *Es ist $\rho(P,R) \le \rho(P,Q) + \rho(Q,R)$ für alle $P, Q, R \in X$.*

Beweis: Ist $\rho(P,Q) = 0$, so ist $d(x,Q) = d(y,P) = 0$ für alle $x \in P$, $y \in Q$. Wegen der Abgeschlossenheit von P und Q bedeutet dies aber, dass $x \in Q$ für alle $x \in P$ und $y \in P$ für alle $y \in Q$, also $P \subset Q \subset P$ bzw. $P = Q$. Weiter ist offensichtlich

[20] Man beachte: Bei festem $x \in \mathbb{R}^n$ ist die Abbildung $y \mapsto \|x-y\|$ stetig, sodass bei der Definition von $d(x,Q)$ zu Recht min statt inf geschrieben wird, da eine stetige reellwertige Funktion auf einer kompakten Menge ihre Extrema annimmt. Weiter ist auch die Abbildung $d(\cdot,Q)\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ ist stetig. Hierzu seien $x, y \in \mathbb{R}^n$ vorgegeben. Dann existieren $q_x, q_y \in Q$ mit

$$d(x,Q) = \|x-q_x\| \le \|x-q\| \quad \text{für alle } q \in Q$$

bzw.

$$d(y,Q) = \|y-q_y\| \le \|y-q\| \quad \text{für alle } q \in Q.$$

Folglich ist

$$\begin{aligned} d(x,Q) - d(y,Q) &= \|x-q_x\| - \|y-q_y\| \\ &\le \|x-q_y\| - \|y-q_y\| \\ &\le \|(x-q_y)-(y-q_y)\| \\ &= \|x-y\|. \end{aligned}$$

Durch Vertauschen von x und y erhält man insgesamt

$$|d(x,Q) - d(y,Q)| \le \|x-y\| \quad \text{für alle } x, y \in \mathbb{R}^n,$$

womit die Stetigkeit von $d(\cdot,Q)$ auf $\mathbb{R}^n$ bewiesen ist. Da P als kompakt vorausgesetzt ist, nimmt $d(\cdot,Q)$ das Maximum auf P an.

$\rho(P,P) = 0$ für alle $P \in X$. Die Symmetrieeigenschaft 2. ist trivialerweise erfüllt. Zum Nachweis von 3. geben wir uns $P, Q, R \in X$ vor. Für $z \in R$ ist

$$\begin{aligned} d(z,P) &= \min_{x\in P} \|z-x\| \\ &\le \min_{x\in P} (\|z-y\| + \|y-x\|) \\ &\le \|z-y\| + \min_{x\in P} \|y-x\| \\ &= \|z-y\| + d(y,P) \\ &\le \|z-y\| + \max_{y\in Q} d(y,P) \quad \text{für alle } y \in Q. \end{aligned}$$

Dann ist aber auch

$$d(z,P) \le \min_{y\in Q} \|z-y\| + \max_{y\in Q} d(y,P) = d(z,Q) + \max_{y\in Q} d(y,P)$$

und folglich

$$\max_{z\in R} d(z,P) \le \max_{z\in R} d(z,Q) + \max_{y\in Q} d(y,P).$$

Vertauscht man in dieser Ungleichung x und z sowie P und R, so erhält man

$$\max_{x\in P} d(x,R) \le \max_{x\in P} d(x,Q) + \max_{y\in Q} d(y,R).$$

Dann ist aber

$$\begin{aligned} \rho(P,R) &= \max\Big(\max_{x\in P} d(x,R), \max_{z\in R} d(z,P)\Big) \\ &\le \max\Big(\max_{x\in P} d(x,Q) + \max_{y\in Q} d(y,R), \max_{z\in R} d(z,Q) + \max_{y\in Q} d(y,P)\Big) \\ &\le \max\Big(\max_{x\in P} d(x,Q), \max_{y\in Q} d(y,P)\Big) + \max\Big(\max_{y\in Q} d(y,R), \max_{z\in R} d(z,Q)\Big) \\ &= \rho(P,Q) + \rho(Q,R), \end{aligned}$$

insgesamt ist der Satz bewiesen. □

Die folgende Definition ist nun naheliegend.

Definition 8.9 Sei (X,ρ) der metrische Raum der kompakten Teilmengen des $\mathbb{R}^n$ versehen mit dem Hausdorff-Abstand. Eine Abbildung $G\colon [t_0,T] \longrightarrow X$ heißt *stetig in* $\hat{t} \in [t_0,T]$, falls es zu jedem $\epsilon > 0$ ein $\delta(\epsilon) > 0$ mit

$$t \in [t_0,T], \quad |t-\hat{t}] \le \delta(\epsilon) \Longrightarrow \rho(G(t),G(\hat{t})) \le \epsilon$$

gibt. Weiter heißt G *stetig auf* $[t_0,T]$, falls G in jedem $\hat{t} \in [t_0,T]$ stetig ist.

8.3.2 Die Existenz einer Lösung

Wichtigstes Hilfsmittel für den Beweis des Existenzsatzes bei zeitoptimalen Steuerungsproblemen mit einem linearen Prozess ist das folgende Ergebnis (siehe z. B. Theorem 1 bei E. B. Lee, L. Markus (1967, S. 69) oder auch Satz 9.1 bei H. Bauer, K. Neumann (1969, S. 96) sowie Lemma 7.6 und Lemma 7.8 bei J. Jahn (1994)).

Satz 8.10 *Sei $\Omega \subset \mathbb{R}^m$ konvex und kompakt. Dann ist die in der Zeit $t \in [t_0, T]$ erreichbare Menge*

$$K(t) := \begin{cases} \{S(u)(t) : u \in L_m^\infty[t_0,t] : u(s) \in \Omega \ \text{ für fast alle } s \in [t_0,t]\}, & t \in (t_0, T], \\ \{x_0\}, & t = t_0 \end{cases}$$

eine konvexe und kompakte Teilmenge des $\mathbb{R}^n$, ferner ist K stetig auf $[t_0, T]$. Hierbei ordnet der Steuerungsoperator S einer Steuerung $u \in L_m^\infty[t_0, T]$ die Lösung x der Prozessgleichung zu, welche der gegebenen Anfangsbedingung $x(t_0) = x_0$ genügt, siehe Definition 8.2.

Beweis: Sei $t \in (t_0, T]$ gegeben. Die Konvexität von $K(t)$ folgt sehr einfach aus der vorausgesetzten Konvexität[21] des Steuerbereichs Ω. Denn ist $\lambda \in [0, 1]$ und sind $u_1, u_2 \in L_m^\infty[t_0, t]$ zwei Steuerungen mit $u_1(s), u_2(s) \in \Omega$ für fast alle $s \in [t_0, t]$ und damit $S(u_1)(t), S(u_2)(t)$ zwei Punkte aus der erreichbaren Menge $K(t)$, so definiere man

$$u_\lambda(s) := (1 - \lambda)u_1(s) + \lambda u_2(s).$$

Es ist $u_\lambda(s) \in \Omega$ für fast alle $s \in [t_0, t]$ wegen der vorausgesetzten Konvexität von Ω und

$$\begin{aligned} (1-\lambda)S(u_1)(t) + \lambda S(u_2)(t) &= (1-\lambda)\Big(\Phi(t)x_0 + \Phi(t)\int_{t_0}^t \Phi(s)^{-1}B(s)u_1(s)\,ds\Big) \\ &\quad + \lambda\Big(\Phi(t)x_0 + \Phi(t)\int_{t_0}^t \Phi(s)^{-1}B(s)u_2(s)\,ds\Big) \\ &= \Phi(t)x_0 + \Phi(t)\int_{t_0}^t \Phi(s)^{-1}B(s)u_\lambda(s)\,ds \\ &= S(u_\lambda)(t) \\ &\in K(t). \end{aligned}$$

Damit ist die Konvexität von $K(t)$ bewiesen. Da $\Omega \subset \mathbb{R}^m$ als kompakte Menge auch beschränkt ist, ist auch die erreichbare Menge $K(t)$ beschränkt. Wir zeigen, dass $K(t) \subset \mathbb{R}^n$ auch abgeschlossen und daher insgesamt kompakt ist. Sei hierzu $x_k(t) := \{S(u_k)(t)\} \subset K(t)$ mit

$$\{u_k\} \subset U(t) := \{u \in L_m^\infty[t_0,t] : u(s) \in \Omega \ \text{ für fast alle } s \in [t_0,t]\}$$

eine Folge in $K(t)$ mit $\lim_{k\to\infty} x_k(t) = x(t)$. Wir zeigen $x(t) \in K(t)$, d. h. die Existenz eines $u^* \in U(t)$ mit $x(t) = S(u^*)(t)$. Die Menge $U(t)$ ist schwach kompakt, siehe E. B. Lee, L. Markus (1967, S. 157), d. h. zu der Folge $\{u_k\} \subset U(t)$ gibt es eine Teilfolge $\{u_{k_i}\}$, welche schwach gegen ein $u^* \in U(t)$ konvergiert. Insbesondere ist

$$\lim_{i\to\infty} \int_{t_0}^t \Phi(s)^{-1}B(s)u_{k_i}(s)\,ds = \int_{t_0}^t \Phi(s)^{-1}B(s)u^*(s)\,ds$$

[21] Erstaunlicherweise ist $K(t)$ auch konvex, wenn Ω nicht notwendig konvex ist, siehe Theorem 1A bei E. B. Lee, L. Marcus (1967, S. 164 ff.).

und damit auch

$$x(t) = \lim_{i\to\infty} x_{k_i}(t) = \lim_{i\to\infty} S(u_{k_i})(t) = S(u^*)(t),$$

womit die Abgeschlossenheit und folglich auch die Kompaktheit von $K(t)$ bewiesen ist. Nun zeigen wir, dass die Abbildung $K\colon [t_0,T] \longrightarrow X$ stetig ist, wobei (X,ρ) der metrische Raum X der kompakten Teilmengen des $\mathbb{R}^n$ versehen mit dem Hausdorff-Abstand ρ ist. Wir überlegen uns, dass $S(u)(\cdot)$ bei einem festen $u \in U(T)$ auf $[t_0,T]$ lipschitzstetig ist, also ein $c > 0$ mit

$$(*) \qquad \|S(u)(t_2) - S(u)(t_1)\| \le c\,|t_2 - t_1| \quad \text{für alle } t_1, t_2 \in [t_0,T]$$

existiert. Denn seien $t_1, t_2 \in [t_0,T]$ fest vorgegeben. Dann ist

$$\begin{aligned} S(u)(t_2) - S(u)(t_1) &= [\Phi(t_2) - \Phi(t_1)]x_0 + \Phi(t_2)\int_{t_0}^{t_2} \Phi(s)^{-1}B(s)u(s)\,ds \\ &\qquad - \Phi(t_1)\int_{t_0}^{t_1} \Phi(s)^{-1}B(s)u(s)\,ds \\ &= \Phi(t_2)\int_{t_1}^{t_2} \Phi(s)^{-1}B(s)u(s)\,ds \\ &\qquad + [\phi(t_2) - \Phi(t_1)]\Big[x_0 + \int_{t_0}^{t_1} \Phi(s)^{-1}B(s)\,u(s)\,ds\Big]. \end{aligned}$$

Wegen der Beschränktheit von Ω, der Beschränktheit der stetigen Matrizen $\Phi(\cdot)$, $\Phi(\cdot)^{-1}$ und $B(\cdot)$ sowie der aus

$$\Phi(t_2) - \Phi(t_1) = \int_{t_1}^{t_2} A(s)\Phi(s)\,ds$$

folgenden Lipschitzstetigkeit von $\Phi(\cdot)$ auf $[t_0,T]$ folgt die Existenz einer (von u sowie t_1, t_2 unabhängigen) Konstanten $c > 0$ mit $(*)$. Bei vorgegebenem $\epsilon > 0$ ist also

$$\|S(u)(t_2) - S(u)(t_1)\| \le \epsilon \quad \text{für alle } t_1, t_2 \in [t_0,T] \text{ mit } |t_2 - t_1| \le \delta(\epsilon) := \epsilon/c.$$

Wir wollen uns überlegen, dass

$$\rho(K(t_1), K(t_2)) \le \epsilon \quad \text{für alle } t_1, t_2 \in [t_0,T] \text{ mit } |t_1 - t_2| \le \delta(\epsilon),$$

womit die Stetigkeit von $K\colon [t_0,T] \longrightarrow X$ bewiesen sein wird. Wir geben uns $t_1, t_2 \in [t_0,T]$ mit $|t_1 - t_2| \le \delta(\epsilon)$ vor und zeigen, dass $d(x_1, K(t_2)) \le \epsilon$ für alle $x_1 \in K(t_1)$. Da wir analog $d(x_2, K(t_1)) \le \epsilon$ für alle $x_2 \in K(t_2)$ zeigen können, werden wir

$$\rho(K(t_1), K(t_2)) = \max\Big(\max_{x_1\in K(t_1)} d(x_1, K(t_2)), \max_{x_2\in K(t_2)} d(x_2, K(t_1))\Big) \le \epsilon$$

bewiesen haben. Sei also $x_1 \in K(t_1)$ gegeben. Dann existiert ein $u \in U(t_1)$ mit $x_1 = S(u)(t_1)$. Wir setzen u auf das ganze Intervall $[t_0,T]$ zu einem $u \in U(T)$ fort. Dann ist

$$d(x_1, K(t_2)) = \min_{x_2\in K(t_2)} \|S(u)(t_1) - x_2\| \le \|S(u)(t_1) - S(u)(t_2)\| \le \epsilon.$$

Insgesamt ist auch die Stetigkeit von K bewiesen. □

Der Existenzsatz für zeitoptimale Steuerungsprobleme ist nun eine verhältnismäßig einfache Folgerung aus Satz 8.10, siehe z. B. auch E. B. Lee, L. Markus (1967, S. 127), H. Bauer, K. Neumann (1969, S. 129) und Theorem 7.7 bei J. Jahn (1994, S. 202).

Satz 8.11 *Gegeben sei das zeitoptimale Steuerungsproblem mit dem linearen Prozess $\dot{x} = A(t)x + B(t)u$ mit auf dem Zeitintervall $[t_0, T]$ stetigen $n \times n$- bzw. $n \times m$-Matrizenfunktionen $A(\cdot)$ bzw. $B(\cdot)$, dem Anfangszustand $x_0 \in \mathbb{R}^n$ zur Anfangszeit t_0, dem konvexen und kompakten Steuerbereich $\Omega \subset \mathbb{R}^m$ und dem auf $[t_0, T]$ konvexen, kompakten und stetigen Ziel $G(\cdot)$. Mit $K(t) \subset \mathbb{R}^n$ sei die in der Zeit t durch eine zulässige Steuerung erreichbare Menge bezeichnet, es sei also*

$$K(t) := \begin{cases} \{S(u)(t) : u \in L_m^\infty[t_0, t] : u(s) \in \Omega \ \text{ für fast alle } s \in [t_0, t]\}, & t \in (t_0, T], \\ \{x_0\}, & t = t_0. \end{cases}$$

Es sei

$$M := \{t \in [t_0, T] : K(t) \cap G(t) \neq \emptyset\} \neq \emptyset,$$

es existiere also eine Zeit $t \in [t_0, T]$, in der der Prozess vom Anfangszustand x_0 durch eine zulässige Steuerung in das Ziel gesteuert werden kann. Dann ist

$$t_1^* := \inf\{t \in [t_0, T] : t \in M\} \in M,$$

das zeitoptimale Steuerungsproblem besitzt also eine Lösung.

Beweis: Wir haben $K(t_1^*) \cap G(t_1^*) \neq \emptyset$ zu zeigen. O. B. d. A. ist $t_1^* < T$, da andernfalls T die minimale Zeit ist, in der der Prozess vom Anfangszustand in das Ziel gesteuert werden kann. Angenommen, es sei $K(t_1^*) \cap G(t_1^*) = \emptyset$. Da $K(t_1^*)$ nach Satz 8.10 kompakt und $G(t_1^*)$ abgeschlossen (und sogar kompakt) ist, haben $K(t_1^*)$ und $G(t_1^*)$ einen positiven Abstand voneinander, d. h. es ist

$$0 < d := d(K(t_1^*), G(t_1^*)) = \inf\{\|k^* - g^*\| : k^* \in K(t_1^*),\, g^* \in G(t_1^*)\}.$$

Wegen der Stetigkeit von $K(\cdot)$ und $G(\cdot)$ in t_1^* existiert ein $\delta > 0$ mit $[t_1^*, t_1^* + \delta] \subset [t_0, T]$ derart, dass

$$t \in [t_1^*, t_1^* + \delta] \Longrightarrow \rho(K(t), K(t_1^*)) < \frac{d}{2},\ \rho(G(t), G(t_1^*)) < \frac{d}{2}.$$

Dann ist $K(t) \cap G(t) = \emptyset$ für alle $t \in [t_1^*, t_1^* + \delta]$, was ein Widerspruch zur Definition von t_1^* bedeutet! Denn seien $t \in [t_1^*, t_1^* + \delta]$ und $(k, g) \in K(t) \times G(t)$ beliebig vorgegeben. Wegen

$$d(k, K(t_1^*)) \leq \rho(K(t), K(t_1^*)) < \frac{d}{2}, \qquad d(g, G(t_1^*)) \leq \rho(G(t), G(t_1^*)) < \frac{d}{2}$$

existiert ein Paar $(k_1^*, g_1^*) \in K(t_1^*) \times G(t_1^*)$ mit

$$\|k - k_1^*\| < \frac{d}{2}, \qquad \|g - g_1^*\| < \frac{d}{2}.$$

Wegen $d = d(K(t_1^*), G(t_1^*))$ ist ferner $d \leq \|k_1^* - g_1^*\|$. Insgesamt ist dann

$$d \leq \|k_1^* - g_1^*\| \leq \|k_1^* - k\| + \|k - g\| + \|g - g_1^*\| < d + \|k - g\|,$$

also $0 < \|k - g\|$ bzw. $k \neq g$ und $K(t) \cap G(t) = \emptyset$. Damit ist der Satz bewiesen. □

8.3.3 Das Maximumprinzip

Unser Ziel ist es, das folgende *Maximumprinzip* für zeitoptimale Steuerungsprobleme mit linearem Prozess zu beweisen, siehe z. B. E. B. Lee, L. Markus (1967, S. 129), H. Bauer, K. Neumann (1969, S. 106).

Satz 8.12 *Gegeben sei das zeitoptimale Steuerungsproblem mit dem linearen Prozess* $\dot{x} = A(t)x + B(t)u$, *den auf dem Zeitintervall* $[t_0, T]$ *stetigen* $n \times n$- *bzw.* $n \times m$-*Matrizenfunktionen* $A(\cdot)$ *bzw.* $B(\cdot)$, *dem Anfangszustand* $x_0 \in \mathbb{R}^n$ *zur Anfangszeit* t_0, *dem konvexen und kompakten Steuerbereich* $\Omega \subset \mathbb{R}^m$ *und dem auf* $[t_0, T]$ *konvexen, kompakten und stetigen Ziel* $G(\cdot)$. *Sei* (u^*, t_1^*) *eine Lösung des angegebenen zeitoptimalen Steuerungsproblems. Dann existiert eine nichttriviale Lösung* η *von* $-\dot{\eta} = A(t)^T\eta$ *mit*

$$\eta(t)^T B(t)u^*(t) = \max_{u\in\Omega} \eta(t)^T B(t)u \quad \textit{für fast alle } t \in [t_0, t_1^*].$$

Ist sogar $u^* \in C_m^{0,s}[t_0, t_1^*]$, *so ist*

$$\eta(t)^T B(t)u^*(t) = \max_{u\in\Omega} \eta(t)^T B(t)u \quad \textit{für alle } t \in [t_0, t_1^*].$$

Beweis: Mit $K(t) \subset \mathbb{R}^n$ bezeichnen wir wieder die in der Zeit t durch eine zulässige Steuerung erreichbare Menge. Der Beweis erfolgt in drei Schritten. Sei $x^* := S(u^*)$ die zur optimalen Steuerung u^* gehörende optimale Trajektorie. Im ersten Schritt zeigen wir:

- Der Punkt $x^*(t_1^*)$ ist ein *Randpunkt* der abgeschlossenen Menge $K(t_1^*)$, es ist also $x^*(t_1^*) \in K(t_1^*) \setminus \text{int}\,(K(t_1^*))$.

Denn angenommen, es sei $x^*(t_1^*) \in \text{int}\,(K(t_1^*))$. Dann existiert ein $\epsilon > 0$ derart, dass die offene Kugel $B(x^*(t_1^*), \epsilon)$ um $x^*(t_1^*)$ mit dem Radius ϵ noch ganz in $K(t_1^*)$ enthalten ist. Wegen der Stetigkeit von $K(\cdot)$ in t_1^* existiert zu $\epsilon > 0$ ein $\delta = \delta(\epsilon) > 0$ mit

$$\rho(K(t), K(t_1^*)) < \frac{\epsilon}{4} \quad \text{für alle } t \in (t_1^* - \delta, t_1^*].$$

Dann ist die offene Kugel $B(x^*(t_1^*), \epsilon/2)$ um $x^*(t_1^*)$ mit dem Radius $\epsilon/2$ für alle $t \in (t_1^* - \delta, t_1^*]$ in $K(t)$ enthalten. Denn angenommen, zu einem $t \in (t_1^* - \delta, t_1^*]$ existiert ein $x \in B(x^*(t_1^*), \epsilon/2)$ mit $x \notin K(t)$. Da $K(t) \subset \mathbb{R}^n$ nichtleer, abgeschlossen und konvex ist, lassen sich $\{x\}$ und $K(t)$ stark durch eine (abgeschlossene) Hyperebene im $\mathbb{R}^n$ trennen, siehe Satz 5.6 oder auch Abschnitt 49 (Trennungssätze für konvexe Mengen im $\mathbb{R}^n$) bei J. Werner (2013). Es existiert also ein $c \in \mathbb{R}^n \setminus \{0\}$ mit

$$c^T x < \gamma := \inf_{k\in K(t)} c^T k.$$

Die Hyperebene

$$H := \{z \in \mathbb{R}^n : c^T z = \gamma\}$$

enthält also $K(t)$ im zugehörigen abgeschlossenen Halbraum $H^+ := \{z \in \mathbb{R}^n : c^T z \geq \gamma\}$ und $\{x\}$ im gegenüberliegenden offenen Halbraum. Die Idee besteht darin, die Existenz

eines Punktes $y \in B(x^*(t_1^*), \epsilon)$ nachzuweisen, der zu $K(t)$ einen Abstand $d(y, K(t)) \geq \epsilon/4$ besitzt. Dann ist y wegen $B(x^*(t_1^*), \epsilon) \subset K(t_1^*)$ ein Element von $K(t_1^*)$, folglich

$$\frac{\epsilon}{4} \leq d(y, K(t)) \leq \rho(K(t_1^*), K(t)) = \rho(K(t), K(t_1^*)) \quad \text{mit einem } t \in (t_1^* - \delta, t_1^*],$$

ein Widerspruch zur Definition von δ. Sei im folgenden $\|\cdot\|$ die euklidische Norm im $\mathbb{R}^n$. Wir definieren

$$y := x - \frac{\epsilon}{4\,\|c\|} c.$$

Dann ist

$$\|y - x^*(t_1^*)\| \leq \underbrace{\|y - x\|}_{=\epsilon/4} + \underbrace{\|x - x^*(t_1^*)\|}_{<\epsilon/2} < \epsilon,$$

also $y \in B(x^*(t_1^*), \epsilon) \subset K(t_1^*)$. Für ein beliebiges $k \in K(t)$ ist weiter (hier benutzen wir, dass $\|\cdot\|$ die euklidische Norm ist!)

$$\begin{aligned} \|y - k\|^2 &= \|y - x + x - k\|^2 \\ &= \|y - x\|^2 + 2(y - x)^T (x - k) + \|x - k\|^2 \\ &= \left(\frac{\epsilon}{4}\right)^2 + \frac{\epsilon}{2\,\|c\|} \underbrace{c^T (k - x)}_{>0} + \underbrace{\|x - k\|^2}_{>0} \\ &> \left(\frac{\epsilon}{4}\right)^2 \end{aligned}$$

und daher

$$d(y, K(t)) = \min_{k \in K} \|y - k\| \geq \frac{\epsilon}{4}.$$

Insgesamt ist damit gezeigt, dass $B(x^*(t_1^*), \epsilon/2) \subset K(t)$ für alle $t \in (t_1^* - \delta, t_1^*]$. Wegen der Stetigkeit von $G(\cdot)$ in t_1^* existiert zu $\epsilon > 0$ ein $\tilde{\delta} > 0$ mit

$$\rho(G(t_1^*), G(t)) < \frac{\epsilon}{2} \quad \text{für alle } t \in (t_1^* - \tilde{\delta}, t_1^*).$$

Wegen $x^*(t_1^*) \in G(t_1^*)$ ist auch

$$d(x^*(t_1^*), G(t)) \leq \max_{y \in G(t_1^*)} d(y, G(t)) \leq \rho(G(t_1^*), G(t)) < \frac{\epsilon}{2} \quad \text{für alle } t \in (t_1^* - \tilde{\delta}, t_1^*).$$

Daher existiert zu jedem $t \in (t_1^* - \tilde{\delta}, t_1^*)$ ein $x(t) \in G(t)$ mit $\|x^*(t_1^*) - x(t)\| < \epsilon/2$. Setzt man $\delta^* := \min(\delta, \tilde{\delta})$, so ist daher

$$\emptyset \neq B(x^*(t_1^*), \epsilon/2) \cap G(t) \subset K(t) \cap G(t) \quad \text{für alle } t \in (t_1^* - \delta^*, t_1^*),$$

ein Widerspruch zur Optimalität von t_1^*.

Im zweiten Schritt zeigen wir:

- Sei $K \subset \mathbb{R}^n$ nichtleer, konvex und abgeschlossen und $x \in K \setminus \text{int}\,(K)$ ein Randpunkt von K. Dann existiert ein $\eta_1 \in \mathbb{R}^n \setminus \{0\}$ mit $\eta_1^T k \leq \eta_1^T x$ für alle $k \in K$. Mit anderen Worten existiert eine $x \in K$ enthaltende *Stützhyperebene*

$$H := \{z \in \mathbb{R}^n : \eta_1^T z = \eta_1^T x\}$$

mit

$$K \subset H^- := \{z \in \mathbb{R}^n : \eta_1^T z \leq \eta_1^T x\},$$

die also K in einem zugehörigen Halbraum enthält.

Zum Beweis machen wir eine Fallunterscheidung. Ist int $(K) \neq \emptyset$, so lassen sich $\{x\}$ und die nichtleere konvexe Menge int (K) durch eine Hyperebene H trennen. Es existiert also ein $\eta_1 \in \mathbb{R}^n \setminus \{0\}$ mit

$$\eta_1^T k \leq \eta_1^T x \quad \text{für alle } k \in K$$

bzw. eine den Punkt x enthaltende Hyperebene

$$H := \{z \in \mathbb{R}^n : \eta_1^T z = \eta_1^T x\},$$

welche int (K) in einem Halbraum enthält, für die also z. B.

$$\text{int } (K) \subset H^- := \{z \in \mathbb{R}^n : \eta_1^T z \leq \eta_1^T x\},$$

woraus auch

$$K = \text{cl } (K) = \text{cl } (\text{int } (K)) \subset \text{cl } (H^-) = H^-$$

folgt, wobei wir Aussagen aus Satz 5.3 benutzt haben. Für int $(K) \neq \emptyset$ ist also obige Aussage bewiesen. Ist dagegen int $(K) = \emptyset$, so liegt K selber schon in einer Hyperebene und die Aussage ist trivialerweise richtig! Hierzu zeigen wir, dass span (K), die Menge aller (endlichen) Linearkombinationen von Elementen aus K, ein *echter* linearer Teilraum des $\mathbb{R}^n$ ist. Sei m die Maximalzahl linear unabhängiger Punkte in K, etwa seien $\{e_1, \ldots, e_m\} \subset K$ linear unabhängig. Wegen int $(K) = \emptyset$ ist $m \leq n-1$. Sei $V :=$ span $\{e_1, \ldots, e_m\}$. Dann ist $K \subset V$, denn andernfalls existierte ein $e_{m+1} \in K \setminus V$. Dieses Element e_{m+1} ist notwendig von $e_1, \ldots, e_m$ linear unabhängig, ein Widerspruch zur Maximalität von m. Also ist span $(K) \subset V$ und V ein echter linearer Teilraum des $\mathbb{R}^n$. Insbesondere ist K in einer Hyperebene des $\mathbb{R}^n$ enthalten.

Fasst man die ersten beiden Schritte zusammen, so erhalten wir:

- Es existiert ein $\eta_1 \in \mathbb{R}^n \setminus \{0\}$ mit $\eta_1^T k \leq \eta_1^T x^*(t_1^*)$ für alle $k \in K(t_1^*)$.

Im letzten Schritt zeigen wir:

- Sei η die (notwendig nichttriviale) Lösung von

  $$(*) \qquad -\dot{\eta} = A(t)^T \eta, \qquad \eta(t_1^*) = \eta_1.$$

  Dann ist

  $$\eta(t)^T B(t) u^*(t) = \max_{u \in \Omega} \eta(t)^T B(t) u \quad \text{für fast alle } t \in [t_0, t_1^*].$$

Mit $\Phi(\cdot)$ bezeichnen wir das durch $\Phi(t_0) = I$ normierte Fundamentalsystem zu $\dot{x} = A(t)x$. Man rechnet leicht nach, dass

$$\eta(t) = \Phi(t)^{-T}\Phi(t_1^*)^T\eta_1.$$

Denn η und die rechts stehende Funktion genügen $(*)$. Sei

$$U(t_1^*) := \{u \in L_m^\infty[t_0, t_1^*] : u(t) \in \Omega \text{ für fast alle } t \in [t_0, t_1^*]\}.$$

Mit $u \in U(t_1^*)$ und der zugehörigen Trajektorie $x := S(u)$ ist dann

$$\begin{aligned}
\eta_1^T x(t_1^*) &= \eta_1^T\Big[\Phi(t_1^*)x_0 + \Phi(t_1^*)\int_{t_0}^{t_1^*}\Phi(t)^{-1}B(t)u(t)\,dt\Big] \\
&= \eta_1^T\Phi(t_1^*)x_0 + \int_{t_0}^{t_1^*}\eta(t)^TB(t)u(t)\,ds \\
&\le \eta_1^T x^*(t_1^*) \\
&\qquad \text{(wegen } x(t_1^*) \in K(t_1^*) \subset H^-\text{)} \\
&= \eta_1^T\Phi(t_1^*)x_0 + \int_{t_0}^{t_1^*}\eta(t)^TB(t)u^*(t)\,dt.
\end{aligned}$$

Also ist

$$\int_{t_0}^{t_1^*}\eta(t)^TB(t)u(t)\,dt \le \int_{t_0}^{t_1^*}\eta(t)^TB(t)u^*(t)\,dt \quad \text{für alle } u \in U(t_1^*).$$

Nun sei $\hat{t} \in [t_0, t_1^*)$ ein Lebesgue-Punkt[22]zu $f(t) := \eta(t)^TB(t)u^*(t)$. Sei $u \in \Omega$ beliebig. Für hinreichend kleine $\epsilon > 0$ definiere man $u_\epsilon \in U(t_1^*)$ durch

$$u_\epsilon(t) := \begin{cases} u, & t \in [\hat{t}, \hat{t}+\epsilon], \\ u^*(t), & t \in [t_0, t_1^*] \setminus [\hat{t}, \hat{t}+\epsilon]. \end{cases}$$

Dann ist

$$\frac{1}{\epsilon}\int_{\hat{t}}^{\hat{t}+\epsilon}\eta(t)^TB(t)u\,dt \le \frac{1}{\epsilon}\int_{\hat{t}}^{\hat{t}+\epsilon}\eta(t)^TB(t)u^*(t)\,dt.$$

[22]Ist $f \in L^\infty[t_0, t_1]$, so heißt $\hat{t} \in [t_0, t_1)$ ein *Lebesgue-Punkt von* f, falls

$$\lim_{\epsilon\to 0+}\frac{1}{\epsilon}\int_{\hat{t}}^{\hat{t}+\epsilon}f(t)\,dt = f(\hat{t}).$$

Fast jeder Punkt aus $[t_0, t_1]$ ist ein Lebesgue-Punkt von f. Denn definiert man F durch $F(t) := \int_{t_0}^t f(s)\,ds$, so ist F absolut stetig auf $[t_0, t_1]$, daher fast überall auf $[t_0, t_1]$ differenzierbar und

$$f(\hat{t}) = \lim_{\epsilon\to 0+}[F(\hat{t}+\epsilon) - F(\hat{t})] = \lim_{\epsilon\to 0+}\frac{1}{\epsilon}\int_{\hat{t}}^{\hat{t}+\epsilon}f(t)\,dt$$

für fast alle $\hat{t} \in [t_0, t_1]$. Ist f stückweise stetig auf $[t_0, t_1]$, besitzt f also nur endlich viele Unstetigkeitsstellen in $[t_0, t_1]$, in denen f aber noch von rechts stetig ist, so ist jeder Punkt in $[t_0, t_1)$ ein Lebesgue-Punkt von f.

Mit $\epsilon \to 0+$ folgt $\eta(\hat{t})^T B(\hat{t})u \le \eta(\hat{t})^T B(\hat{t})u^*(\hat{t})$. Damit gilt

$$\eta(t)^T B(t)u^*(t) = \max_{u\in\Omega} \eta(t)^T B(t)u \quad \text{für fast alle } t \in [t_0, t_1^*].$$

Da für eine auf $[t_0, t_1^*]$ stückweise stetige Funktion jeder Punkt aus $[t_0, t_1^*)$ ein Lebesgue-Punkt ist, ist auch der Zusatz zum Maximumprinzip und damit der ganze Satz bewiesen. □

Beispiel 8.13 Wir betrachten ein zeitoptimales Steuerungsproblem mit den folgenden Daten. Der Prozess sei gegeben durch $\ddot{x} = u$, der Anfangszustand durch $x(0) = x_0$, $\dot{x}(0) = y_0$ mit gegebenen $(x_0, y_0) \in \mathbb{R} \times \mathbb{R}$. Wir schreiben diese Daten als System:

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}}_{=A(t)} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \underbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{=B(t)} u, \qquad \begin{pmatrix} x_1(0) \\ x_2(0) \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}.$$

Das (zeitunabhängige) Ziel sei der Ursprung, also $G := \{(0,0)\}$. Der Steuerbereich sei $\Omega := [-1, 1]$. Wegen des obigen Maximumprinzips existiert zu einer Lösung (u^*, T^*) des zugehörigen zeitoptimalen Steuerungsproblems eine nichttriviale Lösung $\eta(t) = (\eta_1(t), \eta_2(t))^T$ der adjungierten Gleichung $-\dot{\eta} = A(t)^T\eta$ bzw.

$$-\begin{pmatrix} \dot{\eta}_1 \\ \dot{\eta}_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}$$

mit

$$\eta_2(t)u^*(t) = \eta(t)^T B(t)u^*(t) = \max_{u\in[-1,1]} \eta(t)^T B(t)u = \max_{u\in[-1,1]} \eta_2(t)u$$

für fast alle $t \in [0, T^*]$ bzw. für alle $t \in [0, T^*]$, falls u^* sogar stückweise stetig ist. Offenbar existieren Konstanten c_1 und c_2 mit $|c_1|+|c_2| > 0$, d. h. c_1 und c_2 verschwinden nicht beide, mit

$$\eta_1(t) = c_1, \qquad \eta_2(t) = -c_1 t + c_2.$$

Aus $\eta_2(t)u^*(t) = \max_{u\in[-1,1]} \eta_2(t)u$ erhalten wir $u^* = \operatorname{sign}(\eta_2)$, d. h. u^* ist eine sogenannte *bang-bang-Steuerung*, nimmt also nur Werte aus dem Rand des Steuerbereichs an. Weiter lesen wir hieraus ab, dass u^* höchstens einen Sprung bzw. Unstetigkeitsstelle hat, da η_2 höchstens eine Nullstelle besitzt. Angenommen, der "switch" geschehe zur Zeit t^*. Dann sind für die optimale Trajektorie zwei Fälle möglich, je nachdem ob die optimale Steuerung zunächst gleich -1 und dann $+1$ oder erst gleich $+1$ und dann -1 ist. Wir betrachten zunächst den ersten Fall. Hier ist $(x_1^*, x_2^*)^T$ Lösung von

$$\begin{array}{llll} \dot{x}_1 = x_2 \\ \dot{x}_2 = -1 \end{array} \quad \text{auf } [0, t^*], \quad \begin{array}{l} x_1(0) = x_0, \\ x_2(0) = y_0, \end{array} \qquad \begin{array}{l} \dot{x}_1 = x_2 \\ \dot{x}_2 = 1 \end{array} \quad \text{auf } [t^*, T^*], \quad \begin{array}{l} x_1(T^*) = 0, \\ x_2(T^*) = 0. \end{array}$$

Jetzt unterscheiden wir wieder zwei Fälle, nämlich ob $t \in [0, t^*]$ oder $t \in [t^*, T^*]$. Auf $[0, t^*]$ ist x_1^* die Lösung von

$$\ddot{x} = -1, \qquad x(0) = x_0, \quad \dot{x}(0) = y_0,$$

woraus wir

$$x_1^*(t) = -\frac{1}{2}t^2 + x_0 + y_0 t, \qquad t \in [0, t^*],$$

erhalten. Auf $[t^*, T^*]$ ist x_1^* dagegen Lösung von

$$\ddot{x} = 1, \qquad x(T^*) = 0, \quad \dot{x}(T^*) = 0$$

und damit

$$x_1^*(t) = \frac{1}{2}t^2 + \frac{1}{2}(T^*)^2 - T^* t, \qquad t \in [t^*, T^*].$$

Die noch unbekannten t^*, T^* müssen aus der Stetigkeit von x_1^* und $\dot{x}_1^*$ in t^* bestimmt werden. Wegen

$$x_1^*(t^* - 0) = -\frac{1}{2}(t^*)^2 + x_0 + y_0 t^*, \qquad x_1^*(t^* + 0) = \frac{1}{2}(t^*)^2 + \frac{1}{2}(T^*)^2 - T^* t^*$$

ist die Stetigkeit von x_1^* in t^* gleichbedeutend mit

$$(*) \qquad (t^*)^2 + \frac{1}{2}(T^*)^2 - T^* t^* - x_0 - y_0 t^* = 0.$$

Ebenso ist die Stetigkeit von $\dot{x}_1^*$ in t^* wegen

$$\dot{x}_1^*(t^* - 0) = -t^* + y_0, \qquad \dot{x}_1^*(t^* + 0) = t^* - T^*$$

gleichbedeutend mit

$$t^* = \frac{T^* + y_0}{2}.$$

Man kann also die optimale Switch-Zeit t^* durch die optimale Endzeit T^* ausdrücken. Wegen $t^* \le T^*$ ist notwendigerweise $y_0 \le T^*$. Setzt man $t^* = \frac{1}{2}(T^* + y_0)$ in $(*)$ ein, so erhält man

$$\Big(\frac{T^* + y_0}{2}\Big)^2 + \frac{1}{2}(T^*)^2 - T^*\Big(\frac{T^* + y_0}{2}\Big) - x_0 - y_0\Big(\frac{T^* + y_0}{2}\Big) = 0$$

bzw. für T^* die quadratische Gleichung

$$(T^*)^2 - 2y_0 T^* - 4x_0 - y_0^2 = 0.$$

Ist $y_0 > 0$ und $2x_0 + y_0^2 = 0$, so ist $T^* = y_0 > 0$ und $t^* = T^*$, es ist als *kein* Umschalten nötig und die optimale Steuerung ist konstant gleich -1. Sei daher $2x_0 + y_0^2 > 0$. Wegen $y_0 \le T^*$ erhält man als einzige sinnvolle Lösung der quadratischen Gleichung für T^* den Wert

$$T^* = y_0 + \sqrt{4x_0 + 2y_0^2}.$$

Durch einfaches Nachrechnen kann man zeigen: Ist ($x_0 > 0$ und $y_0 > -\sqrt{2x_0}$) oder ($x_0 \le 0$ und $y_0 > \sqrt{-2x_0}$), so ist

$$T^* = y_0 + \sqrt{4x_0 + 2y_0^2} > 0, \qquad t^* = \frac{1}{2}(T^* + y_0) \in (0, T^*).$$

Im zweiten Fall ist die optimale Steuerung zunächst gleich $+1$, dann gleich -1. Hier ist die zugehörige optimale Trajektorie $(x_1^*, x_2^*)^T$ Lösung von

$$\begin{array}{llll} \dot{x}_1 = x_2 & & x_1(0) = x_0, & \dot{x}_1 = x_2 \\ \dot{x}_2 = 1 & \text{auf } [0, t^*], & x_2(0) = y_0, & \dot{x}_2 = -1 \end{array} \quad \text{auf } [t^*, T^*], \quad \begin{array}{l} x_1(T^*) = 0, \\ x_2(T^*) = 0. \end{array}$$

Man kann nun analog dem ersten Fall argumentieren. Die Stetigkeit von x_1^* in t^* ist gleichbedeutend mit

$$(**) \qquad (t^*)^2 + \frac{1}{2}(T^*)^2 - T^* t^* + x_0 + y_0 t^* = 0.$$

Ebenso ist die Stetigkeit von $\dot{x}_1^*$ in t^* äquivalent zu

$$t^* = \frac{T^* - y_0}{2}.$$

Wieder kann die optimale Switch-Zeit t^* durch die optimale Endzeit T^* ausgedrückt werden. Wegen $t^* \leq T^*$ ist $-y_0 \leq T^*$. Setzt man $t^* = \frac{1}{2}(T^* - y_0)$ in $(**)$ ein, so erhält man für T^* die quadratische Gleichung

$$(T^*)^2 + 2y_0 T^* + 4x_0 - y_0^2 = 0.$$

Ist $-y_0 > 0$ und $-2x_0 + y_0^2 = 0$, so ist $T^* = -y_0 > 0$ und $t^* = T^*$, es ist also kein Umschalten nötig und die optimale Steuerung ist konstant gleich $+1$. Sei daher jetzt $-2x_0 + y_0^2 > 0$. Wegen $-y_0 \leq T^*$ erhält man als einzige sinnvolle Lösung der quadratischen Gleichung für T^* den Wert

$$T^* = -y_0 + \sqrt{-4x_0 + 2y_0^2}.$$

Durch einfaches Nachrechnen kann man zeigen: Ist ($x_0 > 0$ und $y_0 < -\sqrt{2x_0}$) oder ($x_0 \leq 0$ und $y_0 < \sqrt{-2x_0}$), so ist

$$T^* = -y_0 + \sqrt{-4x_0 + 2y_0^2} > 0, \qquad t^* = \frac{1}{2}(T^* - y_0) \in (0, T^*).$$

Nun definieren wir $g\colon \mathbb{R} \longrightarrow \mathbb{R}$ durch

$$g(x_0) := \begin{cases} \sqrt{-2x_0}, & x_0 \leq 0, \\ -\sqrt{2x_0}, & x_0 > 0 \end{cases}$$

und anschließend

$$G_+ := \{(x_0, y_0) \in \mathbb{R} \times \mathbb{R} : y_0 > g(x_0)\}, \qquad G_- := \{(x_0, y_0) \in \mathbb{R} \times \mathbb{R} : y_0 < g(x_0)\}.$$

In Abbildung 5 veranschaulichen wir uns die Mengen G_+ und G_-. Für $(x_0, y_0) \in G_+$ ist ($x_0 \leq 0$ und $y_0 > \sqrt{-2x_0}$) oder ($x_0 > 0$ und $y_0 > -\sqrt{2x_0}$), folglich ist die optimale Steuerung u^* zunächst gleich -1, dann gleich $+1$ mit den angegebenen Umschalt- bzw.

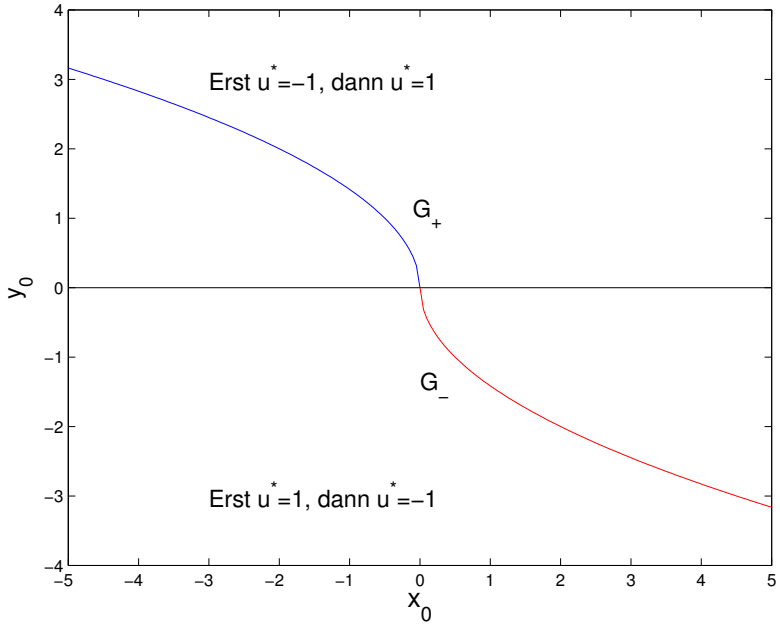

Abbildung 5: Lösung eines zeitoptimalen Steuerungsproblems

Endzeiten t^* bzw. T^*. Entsprechendes gilt für G_-. Der (gemeinsame) Rand von G_+ und G_- ist

$$\Gamma := \{(x_0, y_0) \in \mathbb{R} \times \mathbb{R} : y_0 = g(x_0)\}.$$

Es ist

$$\Gamma = \Gamma_- \cup \{(0,0)\} \cup \Gamma_+$$

mit

$$\Gamma_- := \{(x_0, y_0) \in \Gamma : x_0 < 0\} = \{(x_0, y_0) \in \mathbb{R} \times \mathbb{R} : x_0 < 0,\ 2x_0 + y_0^2 = 0\}$$

und

$$\Gamma_+ := \{(x_0, y_0) \in \Gamma : x_0 > 0\} = \{(x_0, y_0) \in \mathbb{R} \times \mathbb{R} : x_0 > 0,\ -2x_0 + y_0^2 = 0\}.$$

In Abbildung 5 haben wir Γ_- blau und Γ_+ rot gezeichnet. Für $(x_0, y_0) \in \Gamma_-$ ist $T^* = y_0 > 0$ und die optimale Steuerung u^* ist auf $[0, T^*]$ konstant gleich -1. Ist dagegen $(x_0, y_0) \in \Gamma_+$, so ist $T^* = -y_0 > 0$ und die optimale Steuerung u^* ist auf $[0, T^*]$ konstant gleich $+1$. Ist schließlich $(x_0, y_0) = (0, 0)$, so stimmen Anfangszustand und Ziel überein und die optimale Endzeit ist $T^* = 0$, das zeitoptimale Steuerungsproblem ist trivial. Damit ist das gegebene zeitoptimale Steuerungsproblem gelöst. □

8.3.4 Die Eindeutigkeit einer Lösung

Jetzt untersuchen wir die *Eindeutigkeit* einer zeitoptimalen Steuerung. Nach wie vor gehen wir aus von einem linearen Prozess $\dot{x} = A(t)x + B(t)u$ mit den auf dem Intervall

$[t_0, T]$ stetigen $n \times n$- bzw. $n \times m$-Matrizenfunktionen $A(\cdot)$ bzw. $B(\cdot)$, dem Anfangszustand $x_0 \in \mathbb{R}^n$ zur Anfangszeit t_0, dem konvexen und kompakten Steuerbereich $\Omega \subset \mathbb{R}^m$ und dem auf $[t_0, T]$ konvexen, kompakten und stetigen Ziel $G(\cdot)$. Mit $K(t)$ bezeichnen wir wieder die in der t durch eine zulässige Steuerung erreichbare Menge, es sei also

$$K(t) := \begin{cases} \{S(u)(t) : u \in L_m^\infty[t_0, t] : u(s) \in \Omega \text{ für fast alle } s \in [t_0, t]\}, & t \in (t_0, T], \\ \{x_0\}, & t = t_0. \end{cases}$$

Hierbei ordnet der Steuerungsoperator S einer Steuerung $u \in L_m^\infty[t_0, T]$ die Lösung x der Prozessgleichung zu, welche der gegebenen Anfangsbedingung $x(t_0) = x_0$ genügt, siehe Definition 8.2. Sei (u^*, t_1^*) eine Lösung des zugehörigen zeitoptimalen Steuerungsproblems. Klar ist, dass die optimale Endzeit

$$t_1^* := \{t \geq t_0 : K(t) \cap G(t) \neq \emptyset\}$$

eindeutig bestimmt ist. I. allg. ist die optimale Steuerung u^* aber *nicht* eindeutig bestimmt, wie das folgende Beispiel zeigt.

Beispiel 8.14 Man betrachte ein zeitoptimales Steuerungsproblem mit dem Prozess

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix},$$

der Anfangsbedingung

$$\begin{pmatrix} x_1(0) \\ x_2(0) \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \end{pmatrix},$$

dem Ziel $G := \{(0, 0)\}$ und dem Steuerbereich

$$\Omega := \left\{ u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \in \mathbb{R}^2 : |u_1| \leq 1,\ |u_2| \leq 1 \right\}.$$

Sei (u^*, T^*) eine Lösung des zeitoptimalen Steuerungsproblems und $x^* = (x_1^*, x_2^*)^T$ die zugehörige Trajektorie. Dann ist

$$0 = x_1^*(T^*) = -1 + \int_0^{T^*} u_1^*(s)\, ds \leq -1 + T^*.$$

Also ist $1 \leq T^*$, die Minimalzeit ist also mindestens 1. Andererseits kann der Anfangspunkt $(-1, 0)$ in der Zeit 1 für jedes $r \in [0, \frac{1}{2}]$ durch die Steuerung u in das Ziel $(0, 0)$ gesteuert werden, deren erste Komponente durch

$$u_1(t) := 1, \quad t \in [0, 1]$$

und deren zweite Komponente durch

$$u_2(t) := \begin{cases} 1, & t \in [0, \frac{1}{2} - r], \\ 0, & t \in (\frac{1}{2} - r, \frac{1}{2} + r), \\ -1, & t \in [\frac{1}{2} + r, 1] \end{cases}$$

gegeben ist. □

Die folgende Definition (siehe z. B. E. B.Lee, L. Markus (1967, S. 76)) wird sich als hilfreich erweisen.

Definition 8.15 Ein Steuerungsproblem mit dem linearen Prozess $\dot{x} = A(t)x + B(t)u$, dem Anfangszustand x_0 zur Zeit t_0 und dem Steuerbereich $\Omega \subset \mathbb{R}^m$ heißt *normal zur Zeit* $t_1 \geq t_0$, wenn zwei auf $[t_0, t_1]$ zulässige Steuerungen u_1 und u_2, die den Anfangszustand x_0 in denselben Punkt P_1 des Randes $\partial K(t_1)$ von $K(t_1)$ steuern, auf $[t_0, t_1]$ fast überall gleich sind. Ein Steuerungsproblem heißt *normal*, wenn es normal zu jeder Zeit $t_1 \in [t_0, T]$ ist.

Sei (u^*, t_1^*) eine Lösung des zeitoptimalen Steuerungsproblems und x^* die zugehörige Trajektorie. Beim Beweis von Satz 8.12 haben wir ganz am Anfang nachgewiesen:

- Der Punkt $x^*(t_1^*)$ ist ein *Randpunkt* der abgeschlossenen Menge $K(t_1^*)$, es ist also $x^*(t_1^*) \in \partial K(t_1^*) = K(t_1^*) \setminus \text{int}\,(K(t_1^*))$.

Ist also das Ziel einpunktig, so folgt aus der Normalität des Steuerungsproblems offenbar die Eindeutigkeit. Ferner hatten wir im Beweis von Satz 8.12 nachgewiesen, dass eine nichttriviale Lösung η von $-\dot{\eta} = A(t)^T\eta$ existiert mit:

1. Es existiert ein $\eta_1 \in \mathbb{R}^n \setminus \{0\}$ mit $\eta_1^T k \leq \eta_1^T x^*(t_1^*)$ für alle $k \in K(t_1^*)$.

2. Sei η die (notwendig nichttriviale) Lösung von

$$-\dot{\eta} = A(t)^T\eta, \qquad \eta(t_1^*) = \eta_1.$$

   Dann ist

$$\eta(t)^T B(t)u^*(t) = \max_{u\in\Omega} \eta(t)^T B(t)u \quad \text{für fast alle } t \in [t_0, t_1^*].$$

Sei Φ ein normiertes Fundamentalsystem zu $\dot{x} = A(t)x$. Es stellt sich die Frage, unter welchen Bedingungen bei vorgegebenem $\eta_1 \neq 0$ und dadurch bestimmtem $\eta(t) = \Phi(t)^{-T}\Phi(t_1^*)^T\eta_1$ eine (zulässige) Steuerung u^* eindeutig durch

$$\eta(t)^T B(t)u^*(t) = \max_{u\in\Omega} \eta(t)^T B(t)u$$

festgelegt ist.

Definition 8.16 Sei $K \subset \mathbb{R}^n$ eine abgeschlossene, konvexe Menge, die mehr als einen Punkt enthält.

1. Eine Hyperebene $H := \{x \in \mathbb{R}^n : \eta_1^T x = \gamma\}$ mit $\eta_1 \in \mathbb{R}^n \setminus \{0\}$ und $\gamma \in \mathbb{R}$ heißt eine *Stützhyperebene an* K, wenn $K \cap H \neq \emptyset$ und $K \subset H^- := \{x \in \mathbb{R}^n : \eta_1^T x \leq \gamma\}$.

2. Die Menge K heißt *strikt konvex*, wenn jede Stützhyperebene an K die Menge K in genau einem Punkt trifft.

Offenbar gilt: Ist $K \subset \mathbb{R}^n$ abgeschlossen, so existiert zu jedem $P \in \partial K$ eine Stützhyperebene H an K mit $P \in K \subset H$, wie wir uns beim Beweis von Satz 8.12 schon klar gemacht hatten. Eine strikt konvexe Menge K besitzt ein nichtleeres Inneres (denn wäre das Innere leer, so läge K in einer Hyperebene) und jeder Randpunkt von K ist eine *Ecke*, lässt sich also nicht als echte Konvexkombination von zwei verschiedenen Punkten aus K darstellen.

Im folgenden Satz, siehe auch E. B. Lee, L. Markus (1967, S. 76 ff.), wird eine Charakterisierung für die Normalität eines Steuerungsproblems angegeben.

Satz 8.17 *Gegeben sei ein Steuerungsproblem mit dem linearen Prozess $\dot{x} = A(t)x + B(t)u$, dem Anfangszustand $x_0 \in \mathbb{R}^n$ zur Anfangszeit t_0 und dem kompakten Steuerbereich $\Omega \subset \mathbb{R}^m$, der mehr als einen Punkt enthält (andernfalls ist eine Eindeutigkeitsaussage trivial). Dann gilt:*

1. *Ist das Steuerungsproblem zur Zeit t_1 normal, so ist $K(t_1)$ strikt konvex.*
2. *Das Steuerungsproblem ist genau dann zur Zeit t_1 normal, wenn die folgende Eindeutigkeitsaussage gilt: Ist η eine nichttriviale Lösung von $-\dot{\eta} = A(t)^T\eta$ und*

$$u_1, u_2 \in U(t_1) := \{u \in L_m^\infty[t_0, t_1] : u(t) \in \Omega \text{ fast überall auf } [t_0, t_1]\}$$

mit

$$\eta(t)^T B(t)u_1(t) = \eta(t)^T B(t)u_2(t) = \max_{u\in\Omega} \eta(t)^T B(t)u$$

fast überall auf $[t_0, t_1]$, so ist $u_1(t) = u_2(t)$ fast überall auf $[t_0, t_1]$.

Beweis: Der Beweis des ersten Teiles des Satz benutzt (ohne Beweis) ein tiefliegendes maßtheoretisches Ergebnis von Lyapunov, siehe E. B. Lee, L. Markus (1967, S. 163). Der Prozess sei zur Zeit t_1 normal. Angenommen, $K(t_1)$ sei nicht strikt konvex. Dann existiert eine Hyperebene H mit $K(t_1) \subset H^-$, die also $K(t_1)$ in einem abgeschlossenen Halbraum enthält, und zwei verschiedene Punkte $P_1, P_2 \in K(t_1) \cap H$. Dann sind $P_1, P_2 \in \partial K(t_1)$. Denn wäre z. B. $P_1 \in \text{int}\,(K(t_1))$, so existierte ein $\epsilon > 0$ mit

$$P_1 + B[0; \epsilon] \subset K(t_1) \subset H^-.$$

Da $P_1 \in H$ erhält man, dass die Kugel $B[0; \epsilon]$ in einem Halbraum einer Hyperebene durch den Nullpunkt liegt, ein Widerspruch. Aber auch das ganze Segment

$$[P_1, P_2] := \{(1 - \lambda)P_1 + \lambda P_2 : \lambda \in [0, 1]\}$$

liegt in $\partial K(t_1)$, wie man genau wie gerade eben beim Nachweis von $P_1 \in \partial K(t_1)$ zeigt. Sei etwa $P_i = S(u_i)(t_1)$ mit $u_i \in U(t_1)$, $i = 1, 2$. Hierbei ist S natürlich der Steuerungsoperator zum Prozess $\dot{x} = A(t)x + B(t)u$ und der Anfangsbedingung $x(t_0) = x_0$. Jetzt formulieren wir das Lyapunovsche Resultat, das wir nicht beweisen:

- *Sei $y \in L_N^1[t_0, t_1]$. Zu jeder messbaren Teilmenge $E \subset [t_0, t_1]$ definiere man*

$$x(E) := \int_E y(t)\, dt,$$

anschließend definiere man

$$K := \{x(E) : E \subset [t_0, t_1] \text{ messbar}\}.$$

Dann ist $K \subset \mathbb{R}^N$ konvex.

Nun definiere man die $N := 2n$-Vektor-Funktion

$$y(t) := \begin{pmatrix} \Phi(t)^{-1}B(t)u_1(t) \\ \Phi(t)^{-1}B(t)u_2(t) \end{pmatrix},$$

wobei Φ einmal wieder ein normiertes Fundamentalsystem zu $\dot{x} = A(t)x$ ist. Dann ist

$$x(\emptyset) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Wir definieren

$$\begin{pmatrix} r_1 \\ r_2 \end{pmatrix} := x([t_0, t_1]) = \begin{pmatrix} \int_{t_0}^{t_1} \Phi(t)^{-1}B(t)u_1(t)\,dt \\ \int_{t_0}^{t_1} \Phi(t)^{-1}B(t)u_2(t)\,dt \end{pmatrix}.$$

Wegen des Lyapunov-Resultats existiert eine messbare Menge $D \subset [t_0, t_1]$ mit

$$x(D) = \begin{pmatrix} \frac{1}{2}r_1 \\ \frac{1}{2}r_2 \end{pmatrix}.$$

Hieraus folgt

$$x(|t_0, t_1] \setminus D) = x(D) = \begin{pmatrix} \frac{1}{2}r_1 \\ \frac{1}{2}r_2 \end{pmatrix}.$$

Nun definiere man die Steuerungen $\hat{u}_1, \hat{u}_2 \in U(t_1)$ durch

$$\hat{u}_1(t) := \begin{cases} u_1(t), & t \in D, \\ u_2(t), & t \in [t_0, t_1] \setminus D \end{cases}$$

und

$$\hat{u}_2(t) := \begin{cases} u_2(t), & t \in D, \\ u_1(t), & t \in [t_0, t_1] \setminus D. \end{cases}$$

Dann ist

$$S(\hat{u}_1)(t_1) = \Phi(t_1)x_0 + \Phi(t_1)\Big[\int_D \Phi(t)^{-1}B(t)u_1(t)\,dt + \int_{[t_0,t_1]\setminus D} \Phi(t)^{-1}B(t)u_2(t)\,dt\Big].$$

Ferner ist

$$\begin{aligned}\frac{1}{2}S(u_1)(t_1)+\frac{1}{2}S(u_2)(t_1) &= \Phi(t_1)x_0+\Phi(t_1)\Big[\frac{1}{2}\int_D \Phi(t)^{-1}B(t)u_1(t)\,dt\\ &\quad+\frac{1}{2}\int_D \Phi(t)^{-1}B(t)u_2(t)\,dt\\ &\quad+\frac{1}{2}\int_{[t_0,t_1]\setminus D} \Phi(t)^{-1}B(t)u_1(t)\,dt\\ &\quad+\frac{1}{2}\int_{[t_0,t_1]\setminus D} \Phi(t)^{-1}B(t)u_2(t)\,dt\Big]\end{aligned}$$

und daher

$$\begin{aligned}S(\hat{u}_1)(t_1)-\Big[\frac{1}{2}S(u_1)(t_1)+\frac{1}{2}S(u_2)(t_1)\Big] &= \frac{1}{2}\Phi(t_1)\Big[\int_D \Phi(t)^{-1}B(t)u_1(t)\,dt\\ &\quad-\int_{[t_0,t_1]\setminus D} \Phi(t)^{-1}B(t)u_1(t)\,dt\\ &\quad+\int_{[t_0,t_1]\setminus D} \Phi(t)^{-1}B(t)u_2(t)\,dt\\ &\quad-\int_D \Phi(t)^{-1}B(t)u_2(t)\,dt\Big]\\ &= 0 \qquad (\text{wegen } x_D = x_{[t_0,t_1]\setminus D}).\end{aligned}$$

Entsprechend ist auch

$$S(\hat{u}_2)(t_1)=\frac{1}{2}S(u_1)(t_1)+\frac{1}{2}S(u_2)(t_1).$$

Insgesamt ist also

$$S(\hat{u}_1)(t_1)=S(\hat{u}_2)(t_1)=\frac{1}{2}(P_1+P_2)\in[P_1,P_2]\subset\partial K(t_1).$$

Der Prozess wird also durch die beiden auf $[t_0,t_1]$ zulässigen Steuerungen $\hat{u}_1$ und $\hat{u}_2$ vom Anfangszustand x_0 in ein und denselben Randpunkt $\frac{1}{2}(P_1+P_2)$ der in der Zeit erreichbaren Menge $K(t_1)$ gesteuert. Da das Steuerungsproblem nach Voraussetzung normal zur Zeit t_1 ist, ist $\hat{u}_1=\hat{u}_2$ und damit auch $u_1=u_2$ fast überall auf $[t_0,t_1]$. Hieraus folgt aber $P_1=P_2$, ein Widerspruch zu der Annahme, dass P_1,P_2 *verschieden* sind. Also ist $K(t_1)$ strikt konvex und der Beweis des ersten Teils des Satzes ist abgeschlossen.

Jetzt kommen wir zum Beweis des zweiten Teiles des Satzes. Zunächst nehmen wir an, das gegebene Steuerungsproblem sei zur Zeit t_1 normal und haben die angegebene Eindeutigkeitsaussage zu beweisen. Hierzu sei η eine nichttriviale Lösung von $-\dot{\eta}=A(t)^T\eta$, ferner $u_1,u_2\in U(t_1)$ (also auf $[t_0,t_1]$ zulässige Steuerungen) mit

$$\eta(t)^TB(t)u_1(t)=\eta(t)^TB(t)u_2(t)=\max_{u\in\Omega}\eta(t)^TB(t)u$$

fast überall auf $[t_0, t_1]$. Dann ist

$$\begin{aligned}
\eta(t_1)^T S(u_1)(t_1) &= \eta(t_1)^T \Big[\Phi(t_1)x_0 + \Phi(t_1)\int_{t_0}^{t_1} \Phi(t)^{-1}B(t)u_1(t)\,dt\Big] \\
&= \eta(t_1)^T \Big[\Phi(t_1)x_0 + \int_{t_0}^{t_1} \eta(t)^T B(t)u_1(t)\,dt\Big] \\
&= \eta(t_1)^T \Big[\Phi(t_1)x_0 + \int_{t_0}^{t_1} \eta(t)^T B(t)u_2(t)\,dt\Big] \\
&= \eta(t_1)^T S(u_2)(t_1).
\end{aligned}$$

Nun definiere man

$$\gamma := \eta(t_1)^T S(u_1)(t_1) (= \eta(t_1)^T S(u_2)(t_1))$$

und anschließend die Hyperebene

$$H := \{z \in \mathbb{R}^n : \eta(t_1)^T z = \gamma\}.$$

Offenbar ist H eine Stützhyperebene an $K(t_1)$ mit $S(u_1)(t_1), S(u_2)(t_1) \in K(t_1) \cap H$. Wegen des schon bewiesenen ersten Teiles des Satzes folgt aus der vorausgesetzten Normalität des Steuerungsproblems zur Zeit t_1, dass $K(t_1)$ strikt konvex ist. Hieraus wiederum folgt, dass $S(u_1)(t_1) = S(u_2)(t_1)$ und dieser Punkt ist notwendigerweise ein Punkt aus dem Rand $\partial K(t_1)$ von $K(t_1)$. Aus der Normalität des Steuerungsproblems zur Zeit t_1 folgt, dass $u_1(t) = u_2(t)$ für fast alle t aus $[t_0, t_1]$ und die angegebene Eindeutigkeitsaussage ist bewiesen.

Jetzt zeigen wir, dass aus der angegebenen Eindeutigkeitsaussage die Normalität des Steuerungsproblems folgt. Wir nehmen an, die beiden (auf $[t_0, t_1]$ zulässigen) Steuerungen u_1, u_2 mögen den Prozess vom Anfangszustand x_0 in denselben Punkt $P \in \partial K(t_1)$ steuern. Es sei also $P = S(u_1)(t_1) = S(u_2)(t_1)$. Ist dann

$$H := \{z \in \mathbb{R}^n : \eta_1^T z = \eta_1^T P\}$$

mit einem $\eta_1 \in \mathbb{R}^n \setminus \{0\}$ eine Stützhyperebene an $K(t_1)$ (mit $P \in K(t_1) \cap H$) und ist η als Lösung von

$$-\dot{\eta} = A(t)^T \eta, \quad \eta(t_1) = \eta_1$$

definiert, so ist

$$\eta_1^T k \le \eta_1^T S(u_1)(t_1) = \eta_1^T S(u_2)(t_1) \quad \text{für alle } k \in K(t_1).$$

Wie im letzten Teil des Beweises von Satz 8.12 folgt hieraus, dass

$$\eta(t)^T B(t)u_1(t) = \eta(t)^T B(t)u_2(t) = \max_{u\in\Omega} \eta(t)^T B(t)u$$

fast überall auf $[t_0, t_1]$. Nach Annahme folgt $u_1(t) = u_2(t)$ für fast alle t aus $[t_0, t_1]$ und das ist die behauptete Normalität des Steuerungsproblems zur Zeit t_1. □

Beispiel 8.18 Wir hatten uns oben schon überlegt, dass bei einpunktigem Ziel aus der Normalität eines Prozesses bzw. Steuerungsproblems die Eindeutigkeit zeitoptimaler Steuerungen folgt. In Beispiel 8.14 hatten wir ein nicht eindeutig lösbares zeitoptimales Steuerungsproblem mit einem einpunktigen Ziel angegeben, welches also nicht normal sein kann. Es handelte sich um ein Steuerungsproblem mit dem durch

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad \begin{pmatrix} x_1(0) \\ x_2(0) \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$$

gegebenen Prozess bzw. Anfangszustand, dem Steuerbereich $\Omega := \{u \in \mathbb{R}^2 : \|u\|_\infty \leq 1\}$ sowie dem Ziel $G := \{(0,0)\}$. Die adjungierten Gleichungen sind

$$-\dot{\eta} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

also ist eine nichttriviale Lösung der adjungierten Gleichungen durch

$$\eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \text{const}$$

mit $|\eta_1| + |\eta_2| > 0$ gegeben. Nun ist

$$\max_{u:\|u\|_\infty \leq 1} \eta^T u = |\eta_1| + |\eta_2|,$$

aber bei gegebenem $\eta \in \mathbb{R}^2 \setminus \{0\}$ ist ein $u \in \mathbb{R}^2$ mit $\|u\|_\infty$ nur dann aus $\eta^T u = |\eta_1| + |\eta_2|$ eindeutig bestimmbar, wenn η_1 *und* η_2 von Null verschieden sind. Das Steuerungsproblem ist also nicht normal. □

Einfach nachprüfbare Bedingungen für die Normalität eines Steuerungsproblem können wir nur für sehr spezielle Probleme angeben. Wir betrachten nämlich nur autonome Prozesse und nur den Steuerbereich $\Omega = [-1,1]^m := \{u \in \mathbb{R}^m : \|u\|_\infty \leq 1\}$, siehe W. Krabs (1978, Satz 2.5, S. 53).

Satz 8.19 *Ein Steuerungsproblem mit dem autonomen Prozess $\dot{x} = Ax + Bu$ (mit $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$), dem Anfangszustand x_0 zur Zeit $t_0 = 0$ und dem Steuerbereich $\Omega = [-1,1]^m$ ist genau dann (für jede Zeit $t_1 > 0$) normal, wenn für $j = 1, \ldots, m$ die n Vektoren*

$$\{b^j, Ab^j, \ldots, A^{n-1}b^j\} \subset \mathbb{R}^n$$

linear unabhängig sind, wobei b^j die j-te Spalte von B ist, also $B = \begin{pmatrix} b^1 & b^2 & \cdots & b^m \end{pmatrix}$.

Beweis: Im ersten Teil des Beweises nehmen wir an, die Vektoren $\{b^j, Ab^j, \ldots, A^{n-1}b^j\}$ seien für $j = 1, \ldots, m$ linear unabhängig. Wegen Satz 8.17 haben wir zu zeigen: Ist η eine beliebige nichttriviale Lösung von $-\dot{\eta} = A^T\eta$, ist also η durch $\eta(t) = \exp(-A^T t)\eta_0$ mit $\eta_0 \neq 0$ gegeben, so ist $u \in L_m^\infty[0,t_1]$ mit $\|u\|_\infty \leq 1$ für fast alle $t \in [0,t_1]$ eindeutig durch

$$(*) \qquad \eta(t)^T Bu(t) = \max_{u:\|u\|_\infty \leq 1} \eta(t)^T Bu$$

bestimmt. Nun ist

$$\eta(t)^T Bu = \eta_0^T \exp(-A^T t)^T Bu = \eta_0^T \exp(-At)Bu = \sum_{j=1}^{m} \eta_0^T \exp(-At)b^j u_j.$$

Ist daher $\eta_0^T \exp(-At)b^j \neq 0$, $j = 1, \ldots, m$, so ist $u(t)$ aus $(*)$ eindeutig bestimmt und es ist

$$u_j(t) = \operatorname{sign}(\eta_0^T \exp(-At)b^j), \quad j = 1, \ldots, m.$$

Daher ist $u \in L_m^\infty[0, t_1]$ mit $\|u\|_\infty \leq 1$ höchstens dann nicht eindeutig aus $(*)$ bestimmt, wenn es ein $k \in \{1, \ldots, m\}$ und eine Menge $S \subset [0, t_1]$ mit positivem Maß gibt derart, dass $\eta_0^T \exp(-At)b^k = 0$ für alle $t \in S$. Dann ist aber $\eta_0^T \exp(-At)b^k = 0$ für alle t und folglich

$$\eta_0^T b^k = \eta_0^T Ab^k = \cdots = \eta_0^T A^{n-1} b^k = 0.$$

Dann ist span $\{b^k, \ldots, A^{n-1}b^k\}$ ein echter Teilraum des $\mathbb{R}^n$, da das orthogonale Komplement span $\{b^k, \ldots, A^{n-1}b^k\}^\perp$ von span $\{b^k, \ldots, A^{n-1}\}$ nicht nur aus $\{0\}$ besteht, sondern $\eta_0 \neq 0$ enthält. Damit haben wir einen Widerspruch erhalten.

Nun kommen wir zum zweiten Teil des Beweises. Wir setzen also voraus, das Steuerungsproblem sei normal und nehmen im Widerspruch zur Behauptung an, die Vektoren $\{b^k, Ab^k, \ldots, A^{n-1}b^k\}$ seien linear abhängig. Dann gibt es ein $\eta_0 \neq 0$ mit

$$\eta_0^T b^k = \eta_0^T Ab^k = \cdots = \eta_0^T A^{n-1} b^k = 0.$$

Setzt man

$$v(t) := \eta_0^T \exp(-At)b^k,$$

so ist

$$v^{(j)}(t) = \frac{d^j v}{dt^j}(t) = \eta_0^T (-A)^j \exp(-At)b^k, \qquad j = 0, \ldots, n-1,$$

und

$$v^{(j)}(0) = 0, \qquad j = 0, \ldots, n-1.$$

Sei

$$\phi(\lambda) := \det(A - \lambda I) = (-\lambda)^n + a_{n-1}(-\lambda)^{n-1} + \cdots + a_0$$

das charakteristische Polynom zu A. Nach dem Satz von Cayley-Hamilton ist jede quadratische Matrix Nullstelle ihres charakteristischen Polynoms, es ist also

$$\phi(A) = (-A)^n + a_{n-1}(-A)^{n-1} + \cdots + a_1(-A) + a_0 I = 0.$$

Folglich ist

$$v^{(n)}(t) + a_{n-1}v^{(n-1)}(t) + \cdots + a_1 v^{(1)}(t) + a_0 v(t) = 0$$

sowie

$$v^{(j)}(0) = 0, \quad j = 0, \ldots, n-1.$$

Hieraus folgt $v = 0$ bzw. $v(t) = 0$ für alle t. Dies ergibt einen Widerspruch, da $u(t)$ aus

$$\eta(t)^T Bu(t) = \max_{u:\|u\|_\infty \leq 1} \eta(t)^T Bu$$

bzw.

$$\eta_0^T \exp(-At)Bu(t) = \max_{u:\|u\|_\infty \le 1} \eta_0^T \exp(-At)Bu$$

nicht eindeutig bestimmt ist, da die k-te Komponente $u_k(t)$ von $u(t)$ beliebig aus $[-1, 1]$ gewählt werden kann. □

8.4 Notwendige Optimalitätsbedingungen bei diskreten optimalen Steuerungsproblemen

Ein diskretes optimales Steuerungsproblem erhält man durch Diskretisierung eines zugehörigen kontinuierlichen Problems. Genau wie in 8.1 ist auch ein diskretes optimales Steuerungsproblem durch Angabe der Daten Prozess, Anfangszustand, Endzustand, Steuerbereich und Zielfunktion gegeben. In der folgenden Übersicht findet man in der linken Spalte das in 8.1 behandelte kontinuierliche Steuerungsproblem, daneben eine diskretisierte Form des gleichen Problems, welche man dadurch erhält, dass man die Differentialgleichung, die den Prozess beschreibt, z. B. durch die Eulerschen Differenzengleichungen ersetzt, sowie die Zielfunktion z. B. mit Hilfe der Rechteckregel auswertet. In der rechten Spalte wollen wir schließlich das Problem angeben, welches wir im Anschluss weiter behandeln werden. Wir orientieren uns bei der Darstellung an A. Kirsch, W. Warth, J. Werner (1978, S. 111 ff.).

(a)	Der *Prozess* sei gegeben durch $\dot{x} = f(x, u, t)$ auf dem *festen* Zeitintervall $[t_0, T]$. Hierbei ist $f\colon \mathbb{R}^n \times \mathbb{R}^m \times [t_0, T] \longrightarrow \mathbb{R}^n$.	Man zerlege $[t_0, T]$ in k Teilintervalle $t_0 < t_1 < \cdots < t_k = T$, setze $T_i := t_{i+1} - t_i$ und diskretisiere die Differentialgleichung durch $x_{i+1} - x_i = T_i f(x_i, u_i, t_i)$, $i = 0, \ldots, k-1$. Hierbei ist also x_i eine Näherung für $x(t_i)$, $u_i \in \mathbb{R}^m$ eine Näherung für $u(t_i)$.	Der *diskrete Prozess* sei gegeben durch $x_{i+1} - x_i = f_i(x_i, u_i)$, $i = 0, \ldots, k-1$, wobei $f_i\colon \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n$. Setzt man also $f_i(x, u) := T_i f(x, u, t_i)$, so hat man die mittlere Spalte als Spezialfall.
(b)	Der *Anfangszustand* zur Anfangszeit t_0 ist durch $x(t_0) = x_0^*$ fest vorgegeben.	$x_0^* \in \mathbb{R}^n$ gegeben.	$x_0^* \in \mathbb{R}^n$ gegeben.
(c)	Zur vorgegebenen Endzeit T liege der *Endzustand* $x(T)$ auf einer vorgegebenen Fläche $Q_1 := \{x \in \mathbb{R}^n : G(x) = 0\}$. Die Endbedingung sei also $G(x(T)) = 0$, wobei $G\colon \mathbb{R}^n \longrightarrow \mathbb{R}^r$.	$G(x_k)) = 0$.	$G(x_k) = 0$ mit $G\colon \mathbb{R}^n \longrightarrow \mathbb{R}^r$.

(d)	Es sei ein *Steuerbereich* $\Omega \subset \mathbb{R}^m$vorgegeben, d. h. jede zulässige Steuerung muss der Bedingung $u(t) \in \Omega$ für fast alle $t \in [t_0, T]$ genügen.	$u_i \in \Omega$, $i = 0, \ldots, k-1$. Es ist nicht nötig, $u_k \in \Omega$ zu fordern, da u_k im diskreten Prozess nicht auftritt.	$u_i \in \Omega_i$, $i = 0, \ldots, k-1$, mit $\Omega_i \subset \mathbb{R}^m$. Der Steuerbereich darf sich also mit der Zeit ändern.

Durch die angegebenen Daten ist erklärt, was man unter einem *zulässigen Paar* (Zustand, Steuerung) zu verstehen hat. Für das kontinuierliche Problem ist dies in 8.1 schon ausführlich geschehen. Wir wollen uns daher bei der Erläuterung auf das diskrete Problem beschränken. Ein Paar $(x, u) \in \mathbb{R}^{(k+1)n} \times \mathbb{R}^{kn}$ mit $x = (x_0, \ldots, x_k)$ (Zustand, Trajektorie), $u = (u_0, \ldots, u_{k-1})$ (Steuerung) mit $x_i \in \mathbb{R}^n$, $u_i \in \mathbb{R}^m$ heißt *zulässig* für das in der rechten Spalte definierte diskrete Steuerungsproblem, falls

1. $x_{i+1} - x_i = f_i(x_i, u_i)$, $i = 0, \ldots, k-1$.
2. $x_0 = x_0^*$ ist der vorgegebene Anfangszustand.
3. $G(x_k) = 0$.
4. $u_i \in \Omega_i$, $i = 0, \ldots, k-1$.

Die Menge der zulässigen Paare werde mit M bezeichnet.

(e)	Auf der Menge der zulässigen Paare sei eine Zielfunktion durch $F(x,u) = g^0(x(T)) + \int_{t_0}^{T} f^0(x(t), u(t), t)\, dt$ mit $g^0 \colon \mathbb{R}^n \longrightarrow \mathbb{R}$ und $f^0 \colon \mathbb{R}^n \times \mathbb{R}^n \times [t_0, T] \longrightarrow \mathbb{R}$ gegeben.	Eine Diskretisierung der Zielfunktion mit Hilfe der Rechteckregel liefert $F(x,u) = g^0(x_k) + \sum_{i=0}^{k-1} T_i f^0(x_i, u_i, t_i)$.	Auf der Menge M der zulässigen Paare sei eine Zielfunktion gegeben durch $F(x,u) = g^0(x_k) + \sum_{i=0}^{k-1} f_i^0(x_i, u_i)$ mit $g^0 \colon \mathbb{R}^n \longrightarrow \mathbb{R}$, $f_i^0 \colon \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}$, $i = 0, \ldots, k-1$.

In der rechten Spalte haben wir damit vollständig ein diskretes optimales Steuerungsproblem formuliert, welches wir mit (DOSTP) bezeichnen. Unser Ziel ist es, notwendige Optimalitätsbedingungen für (DOSTP) herzuleiten. Bei M. D. Canon, C. D. Cullum Jr., E. Polak (1970, S. 75 ff.) wird ein geringfügig allgemeineres diskretes optimales Steuerungsproblem betrachtet.

Auf das hier vorliegende diskrete optimale Steuerungsproblem wollen wir Satz 7.2, den Satz von F. John, anwenden. Sei

$$C := \{(x,u) \in \mathbb{R}^{(k+1)n} \times \mathbb{R}^{km} : x_0 = x_0^*,\ u = (u_0, \ldots, u_{k-1}) \in \Omega_0 \times \cdots \times \Omega_{k-1}\}$$

und $g \colon \mathbb{R}^{(k+1)n} \times \mathbb{R}^{km} \longrightarrow \mathbb{R}^{kn} \times \mathbb{R}^r$ definiert durch

$$g(x,u) = \begin{pmatrix} (x_{i+1} - x_i - f_i(x_i, u_i))_{i=0,\ldots,k-1} \\ G(x_k) \end{pmatrix}.$$

Das diskrete optimale Steuerungsproblem (DOSTP) kann dann geschrieben werden als

$$\text{(DOSTP)} \qquad \begin{cases} \text{Minimiere} \quad F(x,u) := g^0(x_k) + \displaystyle\sum_{i=0}^{k-1} f_i^0(x_i,u_i) \quad \text{auf} \\ M := \{(x,u) \in \mathbb{R}^{(k+1)n} \times \mathbb{R}^{km} : (x,u) \in C,\ g(x,u) = 0\}. \end{cases}$$

Im folgenden Satz formulieren wir die zu (DOSTP) gehörenden notwendigen Optimalitätsbedingungen.

Satz 8.20 *Gegeben sei das obige diskrete optimale Steuerungsproblem (DOSTP). Sei $(x^*,u^*) \in M$ eine lokale Lösung von (DOSTP). Die folgenden (unnötig starken) Glattheitsvoraussetzungen an die Daten seien erfüllt:*

1. *$f_i\colon \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n$, $i = 0,\ldots,k-1$, ist stetig partiell differenzierbar. Mit f_{ix} bzw. f_{iu} werden die $n \times n$- bzw. $n \times m$-Funktional-Matrix bezüglich des ersten bzw. zweiten Satzes von Variablen bezeichnet.*

2. *$G\colon \mathbb{R}^n \longrightarrow \mathbb{R}^r$ ist stetig partiell differenzierbar. Mit G_x wird die $r \times n$-Funktionalmatrix bezeichnet.*

3. *$g^0\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ und $f_i^0\colon \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}$, $i = 0,\ldots,k-1$, sind stetig partiell differenzierbar. Mit ∇g^0, $\nabla_x f_i^0$ bzw. $\nabla_u f_i^0$ werden die zugehörigen Gradienten bezeichnet.*

Ferner seien $\Omega_i \subset \mathbb{R}^m$ konvex und abgeschlossen, Rang $G_x(x_k^*) = r$ *(also maximal). Dann existieren*

$$\lambda_0 \geq 0, \qquad \eta = (\eta_0,\eta_1,\ldots,\eta_k) \in \mathbb{R}^n \times \mathbb{R}^n \times \cdots \times \mathbb{R}^n, \qquad \mu \in \mathbb{R}^r$$

mit $(\lambda,\eta) \neq (0,0)$ und:

(a) Es gelten die adjungierten Gleichungen

$$-(\eta_{i+1} - \eta_i) = f_{ix}(x_i^*,u_i^*)^T \eta_{i+1} - \lambda_0 \nabla_x f_i^0(x_i^*,u_i^*), \quad i = 0,\ldots,k-1.$$

(b) Es gilt die Transversalitätsbedingung

$$-\eta_k = G_x(x_k^*)^T \mu + \lambda_0 \nabla g^0(x_k^*).$$

(c) Es gilt das lokale Pontryaginsche Maximumprinzip

$$[f_{iu}(x_i^*,u_i^*)^T \eta_{i+1} - \lambda_0 \nabla_u f_i^0(x_i^*,u_i^*)]^T (u_i - u_i^*) \leq 0$$

für alle $u_i \in \Omega_i$, $i = 0,\ldots,k-1$.

Beweis: Wir wollen Satz 7.2, den Satz von F. John, anwenden. Die Zielfunktion F und die Restriktionsabbildung g sind in der lokalen Lösung (x^*, u^*) von (DOSTP) stetig Fréchet-differenzierbar und die Fréchet-Differentiale in (x^*, u^*) sind gegeben durch

$$F'(x^*, u^*)(y, v) = \nabla g^0(x_k^*)^T y_k + \sum_{i=0}^{k-1} [\nabla_x f_i^0(x_i^*, u_i^*)^T y_i + \nabla_u f_i^0(x_i^*, u_i^*)^T v_i]$$

und

$$g'(x^*, u^*)(y, v) = \begin{pmatrix} (y_{i+1} - y_i - f_{ix}(x_i^*, u_i^*) y_i - f_{iu}(x_i^*, u_i^*) v_i)_{i=0,\ldots,k-1} \\ G_x(x_k^*) y_k \end{pmatrix}.$$

Offenbar sind die Voraussetzungen von Satz 7.2 erfüllt: Der Raum $X = \mathbb{R}^{(k+1)n} \times \mathbb{R}^{km}$, "in dem sich alles abspielt", ist ein Banachraum, die Menge

$$C := \{(x, u) \in \mathbb{R}^{(k+1)n} \times \mathbb{R}^{km} : x_0 = x_0^*,\ u = (u_0, \ldots, u_{k-1}) \in \Omega_0 \times \cdots \times \Omega_{k-1}\}$$

der expliziten Restriktionen ist konvex und abgeschlossen, der Bildraum $Y = \mathbb{R}^{kn} \times \mathbb{R}^r$ ist endlichdimensional, es treten nur Gleichungen als implizite Restriktionen auf und die Glattheitsvoraussetzungen an Zielfunktion und Restriktionsabbildung sind erfüllt. Damit erhalten wir die Existenz eines Tripels $(\lambda_0, p, \mu) \in \mathbb{R} \times \mathbb{R}^{kn} \times \mathbb{R}^r \setminus \{(0, 0, 0)\}$ mit $\lambda_0 \geq 0$, $p = (p_1, \ldots, p_k) \in \mathbb{R}^n \times \cdots \times \mathbb{R}^n = \mathbb{R}^{kn}$ und

$$\begin{aligned} 0 \quad \leq \quad & \left[\lambda_0 F'(x^*, u^*) + \begin{pmatrix} p \\ \mu \end{pmatrix}^T g'(x^*, u^*)\right](x - x^*, u - u^*) \\ = \quad & \lambda_0 \Big[\nabla g^0(x_k^*)^T (x_k - x_k^*) + \sum_{i=0}^{k-1} [\nabla_x f_i^0(x_i^*, u_i^*)^T (x_i - x_i^*) + \nabla_u f_i^0(x_i^*, u_i^*)^T (u_i - u_i^*)\Big] \\ & + \sum_{i=0}^{k-1} p_{i+1}^T [(x_{i+1} - x_{i+1}^*) - (x_i - x_i^*) - f_{ix}(x_i^*, u_i^*)(x_i - x_i^*) - f_{iu}(x_i^*, u_i^*)(u_i - u_i^*)] \\ & \quad + \mu^T G_x(x_k^*)(x_k - x_k^*) \end{aligned}$$

für alle $(x, u) \in C$. Setzt man hier speziell $u = u^*$ (und $x_0 = x_0^*$), so erhält man

$$\begin{aligned} 0 \quad \leq \quad & \lambda_0 \Big[\nabla g^0(x_k^*)^T (x_k - x_k^*) + \sum_{i=0}^{k-1} \nabla_x f_i^0(x_i^*, u_i^*)^T (x_i - x_i^*)\Big] \\ & \quad + \sum_{i=0}^{k-1} p_{i+1}^T [(x_{i+1} - x_{i+1}^*) - (x_i - x_i^*) - f_{ix}(x_i^*, u_i^*)(x_i - x_i^*)] \\ & \quad + \mu^T G_x(x_k^*)(x_k - x_k^*) \\ = \quad & \sum_{i=1}^{k-1} [\lambda_0 \nabla_x f_i^0(x_i^*, u_i^*) + p_i - p_{i+1} - f_{ix}(x_i^*, u_i^*)^T p_{i+1}]^T (x_i - x_i^*) \\ & \quad + [\lambda_0 \nabla g^0(x_k^*) + G_x(x_k^*)^T \mu + p_k]^T (x_k - x_k^*) \end{aligned}$$

für alle $(x_1, \ldots, x_k) \in \mathbb{R}^n \times \cdots \times \mathbb{R}^n = \mathbb{R}^{kn}$. Hieraus erhalten wir

1. $-(p_{i+1} - p_i) = f_{ix}(x_i^*, u_i^*)^T p_{i+1} - \lambda_0 \nabla_x f_i^0(x_i^*, u_i^*)$, $i = 1, \ldots, k-1$,

2. $-p_k = G_x(x_k^*)^T \mu + \lambda_0 \nabla g^0(x_k^*)$.

Nun definiere man $\eta = (\eta_0, \eta_1, \ldots, \eta_k) \in \mathbb{R}^n \times \mathbb{R}^n \times \cdots \times \mathbb{R}^n = \mathbb{R}^{(k+1)n}$ durch

$$\eta_i := \begin{cases} p_1 + f_{0x}(x_0^*, u_0^*)^T p_1 - \lambda_0 \nabla_x f_0^0(x_0^*, u_0^*), & i = 0, \\ p_i, & i = 1, \ldots, k. \end{cases}$$

Dann sind offenbar die Bedingungen (a) und (b) erfüllt. Weiter ist $(\lambda_0, \eta) \neq (0, 0)$. Denn andernfalls wäre $(\lambda_0, p) = (0, 0)$, wegen (b) wäre $G_x(x_k^*)^T \mu = 0$ und dann auch $\mu = 0$, da $G_x(x_k^*)$ maximalen Rang besitzt. Wir hätten einen Widerspruch zu $(\lambda_0, p, \mu) \neq (0, 0, 0)$ erhalten. Setzt man in der für alle $(x, u) \in C$ geltenden Ungleichung

$$0 \leq \left[\lambda_0 F'(x^*, u^*) + \begin{pmatrix} p \\ \mu \end{pmatrix}^T g'(x^*, u^*) \right] (x - x^*, u - u^*)$$

speziell $x = x^*$, so erhält man

$$0 \leq \sum_{i=0}^{k-1} [\lambda_0 \nabla_u f_i^0(x_i^*, u_i^*) - f_{iu}(x_i^*, u_i^*)^T p_{i+1}]^T (u_i - u_i^*)$$

für alle $u = (u_0, \ldots, u_{k-1}) \in \Omega_0 \times \cdots \times \Omega_{k-1}$. Setzt man $u_j := u_j^*$ für $j \neq i$, so folgt

$$0 \leq [\lambda_0 \nabla_u f_i^0(x_i^*, u_i^*) - f_{iu}(x_i^*, u_i^*)^T p_{i+1}]^T (u_i - u_i^*) \quad \text{für alle } u_i \in \Omega_i.$$

Wegen $p_{i+1} = \eta_{i+1}$, $i = 0, \ldots, k-1$, ist damit auch (c) und schließlich der ganze Satz bewiesen. □

Beispiel: Wir betrachten ein diskretes optimales Steuerungsproblem mit der Prozessgleichung $x_{i+1} - x_i = u_i$, $i = 0, \ldots, k-1$, dem Anfangszustand $x_0 = 3$, keiner Endbedingung, dem Steuerbereich $\Omega_i = [-1, 1]$, $i = 0, \ldots, k-1$, und der Zielfunktion

$$F(x, u) := \frac{1}{2} \sum_{i=0}^{k} x_i^2 = \frac{1}{2} x_k^2 + \frac{1}{2} \sum_{i=0}^{k-1} x_i^2.$$

Eine Anwendung von Satz 8.20 liefert zu einer Lösung (x^*, u^*) mit

$$x^* = (x_0^*, \ldots, x_k^*), \qquad u^* = (u_0^*, \ldots, u_{k-1}^*)$$

die Existenz von $\lambda_0 \geq 0$, $\eta = (\eta_0, \ldots, \eta_k) \in \mathbb{R}^{k+1}$ mit $(\lambda_0, \eta) \neq (0, 0)$ und

(a) $-(\eta_{i+1} - \eta_i) = -\lambda_0 x_i^*$, $i = 0, \ldots, k-1$,

(b) $-\eta_k = \lambda_0 x_k^*$,

(c) $\eta_{i+1}(u_i - u_i^*) \leq 0$ für alle $u_i \in [-1, 1]$, $i = 0, \ldots, k-1$.

Hier ist notwendig $\lambda_0 > 0$. Denn wäre $\lambda_0 = 0$, so folgt aus (a) und (b), dass $\eta = 0$, ein Widerspruch zu $(\lambda_0, \eta) \neq (0, 0)$. Folglich können wir ohne Einschränkung der Allgemeinheit $\lambda_0 = 1$ annehmen. Zur Lösung (x^*, u^*) existiert also $\eta = (\eta_0, \ldots, \eta_k)$ mit

(a) $-(\eta_{i+1} - \eta_i) = -x_i^*$, $i = 0, \ldots, k-1$,

(b) $-\eta_k = x_k^*$,

(c) $\eta_{i+1}(u_i - u_i^*) \leq 0$ für alle $u_i \in [-1, 1]$, $i = 0, \ldots, k-1$.

Wir wollen uns überlegen, dass diese Bedingungen auch *hinreichend* für die Optimalität einer zulässigen Lösung (x^*, u^*) sind. Wir beziehen uns nur auf unser Beispiel, könnten dieses aber ohne große Schwierigkeiten wesentlich verallgemeinern, da man erwarten kann, dass notwendige Optimalitätsbedingungen bei einer konvexen Optimierungsaufgabe auch hinreichend für Optimalität sind. Sei (x, u) eine weitere zulässige Lösung. Dann ist

$$
\begin{aligned}
F(x,u) - F(x^*,u^*) &= \frac{1}{2}\sum_{i=0}^{k} x_i^2 - \frac{1}{2}\sum_{i=0}^{k}(x_i^*)^2 \\
&\geq \sum_{i=0}^{k} x_i^*(x_i - x_i^*) \\
&\qquad \text{(wegen } \tfrac{1}{2}a^2 - \tfrac{1}{2}b^2 \geq b(a-b)\text{)} \\
&= \sum_{i=0}^{k-1}(\eta_{i+1} - \eta_i)(x_i - x_i^*) - \eta_k(x_k - x_k^*) \\
&\qquad \text{(wegen (a) und (b))} \\
&= \sum_{i=0}^{k-1}\eta_{i+1}(x_i - x_i^*) - \sum_{i=1}^{k-1}\eta_i(x_i - x_i^*) - \eta_k(x_k - x_k^*) \\
&\qquad \text{(wegen } x_0 = x_0^*\text{)} \\
&= \sum_{i=0}^{k-1}\eta_{i+1}(x_i - x_i^*) - \sum_{i=0}^{k-1}\eta_{i+1}(x_{i+1} - x_{i+1}^*) \\
&= \sum_{i=0}^{k-1}\eta_{i+1}((x_i - x_{i+1}) - (x_i^* - x_{i+1}^*)) \\
&= -\sum_{i=0}^{k-1}\underbrace{\eta_{i+1}(u_i - u_i^*)}_{\leq 0} \\
&\qquad \text{(wegen der Prozessgleichung und (c))} \\
&\geq 0.
\end{aligned}
$$

Damit ist nachgewiesen, dass die Bedingungen (a)–(c) auch *hinreichend* für die Optimalität von (x^*, u^*) sind.

Für obige konkrete Daten ($x_0^* = 3$, $\Omega_i = [-1, 1]$) mit $k = 5$ ist eine Lösung gegeben durch

i	0	1	2	3	4	5
x_i^*	3	2	1	0	0	0
u_i^*	-1	-1	-1	0	0	
η_i	-6	-3	-1	0	0	0

Denn offenbar sind die Bedingungen (a)–(c) erfüllt. □

9 Der Brouwersche Abbildungsgrad im $\mathbb{R}^n$

Ziel dieses Abschnitts ist es, den Brouwerschen Abbildungsgrad im $\mathbb{R}^n$ einzuführen und seine wichtigsten Eigenschaften zu beweisen. Der hierbei gewählte analytische Weg stammt von E. Heinz (1959), siehe auch J. M. Ortega, W. C. Rheinboldt (1970, Chapter 6), K. Deimling (1974, 1985), M. Ružička (2004),J. T. Schwartz (1969), L. Nirenberg (1974).

9.1 Definition des Brouwerschen Abbildungsgrades

Der Brouwersche Abbildungsgrad ist eine Funktion d, die einem Tripel (F, Ω, y) bestehend aus einer stetigen Abbildung $F\colon \mathrm{cl}\,(\Omega) \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$, einer offenen, beschränkten Menge $\Omega \subset \mathbb{R}^n$ und einem $y \in \mathbb{R}^n$ mit $y \notin F(\partial\Omega)$, wobei $\partial\Omega := \mathrm{cl}\,(\Omega) \setminus \Omega$ den *Rand* von Ω bedeutet, eine ganze Zahl $d(F, \Omega, y)$ zuordnet. Man ist an Existenzaussagen zu Lösungen eines nichtlinearen Gleichungssystems $F(x) = y$ interessiert. Eine der wichtigsten Eigenschaften des Abbildungsgrades wird daher sein, dass man aus $d(F, \Omega, y) \neq 0$ schließen kann, dass $F(x) = y$ mindestens eine Lösung $x \in \Omega$ besitzt. Weiter wird der Abbildungsgrad die Eigenschaft haben, dass $d(F, \Omega, y) = 1$ ist, falls F die Identität und $y \in \Omega$ ist. Ferner wird $d(\cdot, \Omega, \cdot)$ stetig in (F, y) sein, d. h. kleine Änderungen in F und y ändern den Abbildungsgrad nicht.

Die Definition des Abbildungsgrades ist nicht ganz einfach und wird schrittweise erfolgen. Hierbei benutzen wir die folgenden Bezeichnungen. Bei gegebenem $\alpha > 0$ nennen wir ein Element von

$$W_\alpha := \{\phi \in C[0, \infty) : \text{Es existiert } \delta \in (0, \alpha) \text{ mit } \phi(t) = 0 \text{ für } t \notin [\delta, \alpha]\}$$

eine *Gewichtsfunktion vom Index* α. Ist $\phi \in W_\alpha$, so ist $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}$, definiert durch $g(x) := \phi(\|x\|_2)$, eine stetige Funktion mit *kompaktem Träger* auf dem $\mathbb{R}^n$, und damit sind $\int_{\mathbb{R}^n} \phi(\|x\|_2)\, dx$ und die Menge

$$W_\alpha^1 := \Big\{\phi \in W_\alpha : \int_{\mathbb{R}^n} \phi(\|x\|_2)\, dx = 1\Big\}$$

wohldefiniert. Hierbei bedeutet $\|\cdot\|_2$ natürlich die euklidische Norm auf dem $\mathbb{R}^n$.

Definition 9.1 Sei $F\colon D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig differenzierbar auf der offenen Menge D, Ω eine offene, beschränkte Menge mit $\mathrm{cl}\,(\Omega) \subset D$ und $y \notin F(\partial\Omega)$. Für eine Gewichtsfunktion $\phi \in W_\alpha$ mit

$$0 < \alpha < d(F(\partial\Omega), y) := \min_{x\in\partial\Omega} \|F(x) - y\|_2$$

definiere man die Abbildung $\Phi\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ durch

$$\Phi(x) := \begin{cases} \phi(\|F(x) - y\|_2) \det F'(x), & x \in \Omega, \\ 0, & x \notin \Omega. \end{cases}$$

Hierbei ist $F'(x)$ die Jacobi- oder Funktionalmatrix von F in x. Dann heißt

$$d_\phi(F, \Omega, y) := \int_{\mathbb{R}^n} \Phi(x)\, dx$$

das *Abbildungsgrad-Integral* von F auf Ω bezüglich y und der Gewichtsfunktion ϕ.

Beispiel: Wir betrachten den eindimensionalen Fall, es sei also $n := 1$. Sei $F\colon \mathbb{R} \longrightarrow \mathbb{R}$ stetig differenzierbar, $\Omega := (a, b)$ ein offenes Intervall mit $F'(x) > 0$ für alle $x \in \Omega$ und $F(a) < y := 0 < F(b)$. Sei $\alpha < \min(-F(a), F(b))$ und $\phi \in W_\alpha$. Dann ist

$$d_\phi(F, \Omega, 0) = \int_a^b \phi(|F(x)|)F'(x)\, dx = \int_{F(a)}^{F(b)} \phi(|y|)\, dy = \int_{-\infty}^{\infty} \phi(|y|)\, dy,$$

also ist $d_\phi(F, \Omega, 0) = 1$, falls $\phi \in W_\alpha^1$. □

Unser erstes Ziel wird der Nachweis dafür sein, dass das Abbildungsgrad-Integral d_ϕ für $\phi \in W_\alpha^1$ eine ganze Zahl ist und für hinreichend kleines α nicht von ϕ abhängt. Dies wird dann zur Definition des Abbildungsgrades für stetig differenzierbares F führen.

Satz 9.2 *Sei $F\colon D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig differenzierbar auf der offenen Menge D und Ω eine offene, beschränkte Menge mit $\mathrm{cl}\,(\Omega) \subset D$, ferner sei $y \notin F(\partial\Omega)$. Es wird angenommen, dass die Jacobi-Matrix $F'(x)$ für alle $x \in \Gamma := \{x \in \Omega : F(x) = y\}$ nichtsingulär ist. Dann gilt:*

(a) Γ besteht aus höchstens endlich vielen Punkten.

(b) Es existiert ein $\hat\alpha \in (0, d(F(\partial\Omega), y))$ derart, dass für alle $\phi \in W_\alpha^1$ mit $\alpha \in (0, \hat\alpha)$ gilt:

$$d_\phi(F, \Omega, y) = \begin{cases} \displaystyle\sum_{j=1}^{m} \mathrm{sign}\,(\det F'(x_j)), & \Gamma = \{x_1, \ldots, x_m\}, \\ 0, & \Gamma = \emptyset. \end{cases}$$

Beweis: Im ersten Teil (a) des Beweises zeigen wir, dass Γ aus höchstens endlich vielen Punkten besteht. Angenommen, dies sei nicht der Fall, es existiere also eine unendliche Folge $\{x_j\} \subset \Gamma$. Wegen $\Gamma \subset \Omega \subset \mathrm{cl}\,(\Omega)$ ist Γ als Teilmenge der kompakten Menge $\mathrm{cl}\,(\Omega)$ selbst kompakt. Also besitzt $\{x_j\}$ eine konvergente Teilfolge $\{x_{j_k}\} \subset \{x_j\}$ mit

$x_{j_k} \to x \in \mathrm{cl}\,(\Omega)$. Da F speziell stetig ist, folgt aus $F(x_{j_k}) = y$, dass $y = F(x)$. Da $y \notin F(\partial\Omega)$, ist $x \in \Omega$ und damit $x \in \Gamma$. Nach Voraussetzung ist $F'(x)$ nichtsingulär. Daher können wir den *Satz über inverse Funktionen* (siehe z. B. J. M. Ortega, W. C. Rheinboldt (1970, S. 125)) anwenden:

- Sei $F\colon D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig differenzierbar auf der offenen Menge D. Sei $x \in D$ und $F'(x)$ nichtsingulär. Dann ist F ein lokaler Homöomorphismus in x, d. h. es existieren Umgebungen U bzw. V von x bzw. $F(x)$ derart, dass die Restriktion F_U von F auf U ein Hömöomorphismus zwischen U und V ist[23].

Hiernach folgt insbesondere, dass F eine Umgebung U von x eineindeutig auf eine Umgebung von $y = F(x)$ abbildet. Für alle hinreichend großen k ist $x_{j_k} \in U$, andererseits $y = F(x) = F(x_{j_k})$, also $x_{j_k} = x$ für alle hinreichend großen k, ein Widerspruch. Also besteht Γ in der Tat aus höchstens endlich vielen Punkten.

Zum Beweis des zweiten Teiles (b) nehmen wir zunächst an, es sei $\Gamma = \emptyset$ bzw. $y \notin F(\Omega)$. Da auch $y \notin F(\partial\Omega)$, ist $y \notin F(\mathrm{cl}\,(C))$. Daher ist der Abstand

$$d(F(\mathrm{cl}\,(\Omega)), y) := \min_{x \in \mathrm{cl}\,(\Omega)} \|F(x) - y\|_2$$

von y zur kompakten Menge $F(\mathrm{cl}\,(\Omega))$ positiv. Sei $\hat{\alpha} := d(F(\mathrm{cl}\,(\Omega)), y)$ und $\alpha \in (0, \hat{\alpha})$ sowie $\phi \in W^1_\alpha$. Für $x \in \mathrm{cl}\,(\Omega)$ ist $\|F(x) - y\|_2 \geq \hat{\alpha} > \alpha$, folglich $\phi(\|F(x) - y\|_2) = 0$, damit $\Phi(x) = 0$ für alle $x \in \mathbb{R}^n$ und $d_\phi(F, \Omega, y) = 0$ nach Definition des Abbildungsgrad-Integrals.

Um den Beweis des zweiten Teiles (b) abzuschließen, nehmen wir jetzt an, es sei $\Gamma = \{x_1, \ldots, x_m\} \neq \emptyset$. Wegen des Satzes über inverse Funktionen existieren offene Umgebungen $U(x_j) \subset \Omega$ von x_j und $V_j(y)$ von y derart, dass die Restriktion F_j von F auf $U(x_j)$ ein Homöomorphismus von $U(x_j)$ auf $V_j(y)$ ist, $j = 1, \ldots, m$. O. B. d. A. ist $U(x_j)$ so klein, dass die $U(x_j)$, $j = 1, \ldots, m$, paarweise disjunkt sind und $\mathrm{sign}\,(\det F'(x))$ konstant für $x \in U(x_j)$ ist. Da es nur endlich viele Umgebungen $V_j(y)$ gibt, gibt es eine Kugel um y, die in allen $V_j(y)$, $j = 1, \ldots, m$, enthalten ist. Also existiert ein $\hat{\alpha} \in (0, d(F(\partial\Omega), y))$ mit

$$K := \{z \in \mathbb{R}^n : \|z - y\|_2 \leq \hat{\alpha}\} \subset V_j(y), \quad j = 1, \ldots, m.$$

Man definiere $U_j := F_j^{-1}(K)$, $j = 1, \ldots, m$, und wähle $\alpha \in (0, \hat{\alpha})$, $\phi \in W^1_\alpha$. Für $x \notin U_j$ ist $F(x) \notin K$, also $\|F(x) - y\|_2 > \hat{\alpha} > \alpha$ und damit $\phi(\|F(x) - y\|_2) = 0$. Daher ist $\phi(\|F(x) - y\|_2) = 0$ für $x \notin \bigcup_{j=1}^m U_j$. Folglich ist

$$\begin{aligned} d_\phi(F, \Omega, y) &= \int_\Omega \phi(\|F(x) - y\|_2) \det F'(x)\, dx \\ &= \sum_{j=1}^m \int_{U_j} \phi(\|F(x) - y\|_2) \det F'(x)\, dx \end{aligned}$$

[23] D. h. $F_U\colon U \longrightarrow V$ ist eine eineindeutige Abbildung von U auf V und F_U bzw. F_U^{-1} sind stetig auf U bzw. $V = F(U)$.

$$
\begin{aligned}
&= \sum_{j=1}^{m} \int_{F_j^{-1}(K)} \phi(\|F_j(x) - y\|_2) \det F_j'(x)\, dx \\
&= \sum_{j=1}^{m} \operatorname{sign}(\det F'(x_j)) \int_{F_j^{-1}(K)} \phi(\|F_j(x) - y\|_2) |\det F_j'(x)|\, dx \\
&= \sum_{j=1}^{m} \operatorname{sign}(\det F'(x_j)) \int_{K} \phi(\|x - y\|_2)\, dx \\
&\qquad \text{(Substitutionsregel)} \\
&= \sum_{j=1}^{m} \operatorname{sign}(\det F'(x_j)) \underbrace{\int_{\mathbb{R}^n} \phi(\|x - y\|_2)\, dx}_{=1} \\
&= \sum_{j=1}^{m} \operatorname{sign}(\det F'(x_j)).
\end{aligned}
$$

Damit ist der Satz bewiesen. □

Unser erstes Ziel ist damit fast erreicht. Wir mussten allerdings in Satz 9.2 die Voraussetzung machen, dass $F'(x)$ für $x \in \Gamma$ nichtsingulär ist. Unser nächstes Ziel ist es, den folgenden Satz (siehe J. M. Ortega, W. C. Rheinboldt (1970, 6.1.4. auf S. 151)) zu beweisen.

Satz 9.3 *Sei $F\colon D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig differenzierbar auf der offenen Menge D und Ω eine offene, beschränkte Menge mit* cl $(\Omega) \subset D$. *Ist $y \notin F(\partial\Omega)$, so ist*

$$d_{\phi_1}(F, \Omega, y) = d_{\phi_2}(F, \Omega, y)$$

für alle $\phi_1, \phi_2 \in W_\alpha^1$ mit $\alpha \in (0, d(F(\partial\Omega), y))$.

Haupthilfsmittel für den Beweis von Satz 9.3 ist der folgende Satz (siehe J. M. Ortega, W. C. Rheinboldt (1970, 6.1.3. auf S. 151)). In dessen Beweis steckt die Hauptarbeit. Wir werden technische Einzelheiten weglassen und nur die Idee zu seinem Beweis angeben. Einen ausführlichen Beweis findet man bei J. M. Ortega, W. C. Rheinboldt (1970, S. 169 ff.).

Satz 9.4 *Sei $F\colon D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig differenzierbar auf der offenen Menge D und Ω eine offene, beschränkte Menge mit* cl $(\Omega) \subset D$. *Sei $y \notin F(\partial\Omega)$. Dann gilt: Ist $\alpha \in (0, d(F(\partial\Omega), y))$ und $\phi \in W_\alpha$, so ist $\eta(\phi) := \int_0^\infty t^{n-1}\phi(t)\, dt$ wohldefiniert und $\eta(\phi) = 0$ impliziert $d_\phi(F, \Omega, y) = 0$.*

Beweisidee: Trivialerweise ist $\eta(\phi)$ wohldefiniert, da $\phi(t) = 0$ für $t \geq \alpha$. Der erste Beweisschritt besteht darin sich zu überlegen, dass F o. B. d. A. sogar *zweimal* stetig differenzierbar ist. Haupthilfsmittel (siehe J. M. Ortega, W. C. Rheinboldt (1970, 6.5.5. auf S. 172)), welches wir aber *nicht* beweisen werden, hierbei ist:

(1) Sei $F\colon D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig differenzierbar auf der offenen Menge D und Ω eine offene, beschränkte Menge mit cl $(\Omega) \subset D$. Dann existiert zu jedem $\epsilon > 0$ eine zweimal stetig differenzierbare Abbildung $G\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$ mit

$$\max(\|F(x) - G(x)\|_2, \|F'(x) - G'(x)\|_2) < \epsilon \quad \text{für alle } x \in \text{cl}\,(\Omega).$$

Jetzt kommt der Nachweis, dass wir o. B. d. A. annehmen können, dass F sogar zweimal stetig differenzierbar ist. Sei $\alpha \in (0, d(F(\partial\Omega), y))$. Zu $\epsilon \in (0, d(F(\partial\Omega), y) - \alpha)$ existiert wegen der (hier nicht bewiesenen) Aussage (1) eine zweimal stetig differenzierbare Abbildung $G_\epsilon \colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$ mit

$$(*) \qquad \max(\|F(x) - G_\epsilon(x)\|_2, \|F'(x) - G'_\epsilon(x)\|_2) < \epsilon \quad \text{für alle } x \in \text{cl } (\Omega).$$

Für alle $x \in \partial\Omega$ ist

$$\begin{aligned} \|G_\epsilon(x) - y\|_2 &\geq \|F(x) - y\|_2 - \|F(x) - G_\epsilon(x)\|_2 \\ &\geq \|F(x) - y\|_2 - \epsilon \\ &\geq d(F(\partial\Omega), y) - \epsilon \\ &> \alpha. \end{aligned}$$

Also ist auch $d(G_\epsilon(\partial\Omega), y) > \alpha$. Damit ist $d_\phi(G_\epsilon, \Omega, y)$ für $\phi \in W_\alpha$ wohldefiniert. Ist $\epsilon_1 > 0$ gegeben, so ist wegen der Stetigkeit von ϕ und der stetigen Abhängigkeit der Determinante einer Matrix von deren Elementen

$$|d_\phi(F, \Omega, y) - d_\phi(G_\epsilon, \Omega, y)| < \epsilon_1$$

für alle hinreichend kleinen $\epsilon > 0$, wobei wir $(*)$ benutzt haben. Wenn wir also zeigen können, dass

$$\eta(\phi) = 0 \Longrightarrow d_\phi(G_\epsilon, \Omega, y) = 0,$$

so folgt aus $\eta(\phi) = 0$ auch $|d_\phi(F, \Omega, y)| < \epsilon_1$ und, da ϵ_1 beliebig, sogar $d_\phi(F, \Omega, y) = 0$. Daher ist es keine Beschränkung der Allgemeinheit anzunehmen, dass F sogar zweimal stetig differenzierbar ist.

Der zweite Beweisschritt besteht darin nachzuweisen, dass der Integrand bei der Definition von $d_\phi(F, \Omega, y)$, nämlich

$$\Phi(x) := \begin{cases} \phi(\|F(x) - y\|_2) \det F'(x), & x \in \Omega, \\ 0, & x \notin \Omega, \end{cases}$$

sich als Divergenz darstellen lässt, d. h. dass eine stetig differenzierbare Abbildung $P \colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$, $P(x) = (p_1(x), \ldots, p_n(x))^T$, existiert mit

$$\text{div } P(x) = \sum_{i=1}^{n} \frac{\partial}{\partial x_i} p_i(x) = \Phi(x).$$

Nun sei $F \colon D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ auf D zweimal stetig differenzierbar, $\alpha \in (0, d(F(\partial\Omega), y))$ sowie $\phi \in W_\alpha$ mit $\eta(\phi) = 0$. Man setze

$$\psi(t) := \begin{cases} t^{-n} \displaystyle\int_0^t s^{n-1} \phi(s)\, ds, & t \in (0, \infty), \\ 0, & t = 0. \end{cases}$$

Da nach Definition von W_α zu $\phi \in W_\alpha$ ein $\delta \in (0,\alpha)$ mit $\phi(s) = 0$ für $s \notin [\delta,\alpha]$ existiert, ist $\psi(t) = 0$ für $t \in [0,\delta]$ (wir sagen: ψ verschwindet auf einer Umgebung von 0), $\psi \in C^1[0,\infty)$ und

$$\begin{aligned}\psi'(t) &= -nt^{-n-1}\int_0^t s^{n-1}\phi(s)\,ds + t^{-n}\cdot t^{n-1}\phi(t)\\ &= -nt^{-1}\psi(t) + t^{-1}\phi(t),\end{aligned}$$

folglich

$$t\psi'(t) + n\psi(t) = \phi(t).$$

Wegen $0 = \eta(\phi) = \int_0^\alpha s^{n-1}\phi(s)\,ds$ ist auch $\psi(t) = t^{-n}\int_0^\alpha s^{n-1}\phi(s)\,ds = 0$ für $t \geq \alpha$, also $\psi \in W_\alpha$. Man definiere die Abbildungen $H\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$ und $G\colon D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ durch

$$H(x) := \psi(\|x\|_2)x, \qquad G(x) := H(F(x) - y).$$

Naheliegenderweise sei

$$H(x) = (h_1(x),\ldots,h_n(x))^T, \qquad G(x) = (g_1(x),\ldots,g_n(x))^T.$$

Da ψ in einer Umgebung von $t = 0$ verschwindet und die euklidische Norm $\|\cdot\|_2$ in einem Punkt $x \neq 0$ stetig differenzierbar ist mit

$$\frac{\partial}{\partial x_i}\|x\|_2 = \frac{\partial}{\partial x_i}\Big(\sum_{i=1}^n x_i^2\Big)^{1/2} = \frac{x_i}{\|x\|_2}$$

ist $H\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$ und damit auch $G\colon D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig differenzierbar. Ferner ist

$$\begin{aligned}\operatorname{div} H(x) &= \sum_{i=1}^n \frac{\partial}{\partial x_i}\psi(\|x\|_2)x_i\\ &= \sum_{i=1}^n\Big(\psi'(\|x\|_2)\frac{x_i^2}{\|x\|_2} + \psi(\|x\|_2)\Big)\\ &= \psi'(\|x\|_2)\|x\|_2 + n\psi(\|x\|_2)\\ &= \phi(\|x\|_2).\end{aligned}$$

Nun benötigen wir eine weitere Hilfsaussage:

(2) Sei $F\colon D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ zweimal stetig differenzierbar auf der offenen Menge D,

$$F(x) = (f_1(x),\ldots,f_n(x))^T, \qquad F'(x) = \Big(\frac{\partial f_i}{\partial x_j}(x)\Big).$$

Sei $a_{ij}(x)$ der Kofaktor des (i,j)-Elementes von $F'(x)$, d. h. die Determinante der $(n-1)\times(n-1)$-Untermatrix von $F'(x)$, die durch Streichen der i-ten Zeile und j-ten Spalte entsteht, multipliziert mit $(-1)^{i+j}$. Für alle $x \in D$ gilt dann:

1. $\delta_{kj}\det F'(x) = \sum_{i=1}^n a_{ji}(x)\frac{\partial f_k}{\partial x_i}(x),\ k,j = 1,\ldots,n,$

2. $\displaystyle\sum_{j=1}^{n} \frac{\partial a_{ij}}{\partial x_j}(x) = 0,\ i = 1, \ldots, n.$

Denn: Für den ersten Teil 1. unterscheide man zwischen den Fällen $k = j$ und $k \neq j$. Für $k = j$ wird behauptet, dass

$$\det F'(x) = \sum_{i=1}^{n} a_{ji}(x) \frac{\partial f_j}{\partial x_i}(x),$$

was richtig ist, da es nichts anderes als die Entwicklung von $\det F'(x)$ nach der j-ten Zeile von $F'(x)$ ist. Nun betrachten wir den zweiten Fall $k \neq j$. In $F'(x)$ ersetze man die j-te Zeile durch die k-te Zeile. Die entstehende Matrix hat zwei gleiche Zeilen, ihre Determinante ist also gleich Null. Eine Entwicklung dieser Determinante nach der j-ten Zeile liefert die Behauptung. Für einen Beweis von 2. verweisen wir auf J. M. Ortega, W. C. Rheinboldt (1970, S. 170–171).

Nun setzen wir den zweiten Beweisschritt fort. Dieser besteht im Nachweis dafür, dass eine stetig differenzierbare Abbildung $P\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$ mit $\operatorname{div} P(x) = \Phi(x)$ für alle $x \in \mathbb{R}^n$ existiert. Wir erinnern an die Definition der Abbildungen H und G. Hiernach ist

$$g_j(x) = h_j(F(x) - y) = \psi(\|F(x) - y\|_2) x_j, \quad j = 1, \ldots, n.$$

Für $x \in D$ sei auch im folgenden $a_{ij}(x)$ der Kofaktor des (i,j)-Elementes von $F'(x)$. Wir definieren $P\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$, $P(x) = (p_1(x), \ldots, p_n(x))^T$, durch

$$p_i(x) := \begin{cases} \displaystyle\sum_{j=1}^{n} a_{ji}(x) g_j(x), & x \in \Omega, \\ 0, & x \notin \Omega, \end{cases} \qquad i = 1, \ldots, n.$$

Für $x \in \Omega$ ist dann

$$\begin{aligned}
\operatorname{div} P(x) &= \sum_{i=1}^{n} \frac{\partial p_i}{\partial x_i}(x) \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \Big(\sum_{j=1}^{n} a_{ji}(x) g_j(x) \Big) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial a_{ji}}{\partial x_i}(x) g_j(x) + \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ji}(x) \frac{\partial g_j}{\partial x_i}(x) \\
&= \sum_{j=1}^{n} \Big(\underbrace{\sum_{i=1}^{n} \frac{\partial a_{ji}}{\partial x_i}(x)}_{=0,\ (2)\,2.} \Big) g_j(x) + \sum_{i,j=1}^{n} a_{ji}(x) \frac{\partial g_j}{\partial x_i}(x) \\
&= \sum_{i,j=1}^{n} a_{ji}(x) \sum_{k=1}^{n} \frac{\partial h_j}{\partial x_k}(F(x) - y) \frac{\partial f_k}{\partial x_i}(x)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{n}\sum_{k=1}^{n}\Big(\underbrace{\sum_{i=1}^{n} a_{ji}(x)\frac{\partial f_k}{\partial x_i}(x)}_{(2)\,1.}\Big)\frac{\partial h_j}{\partial x_k}(F(x)-y)\\
&= \sum_{j=1}^{n}\sum_{k=1}^{n}\delta_{kj}\det F'(x)\frac{\partial h_j}{\partial x_k}(F(x)-y)\\
&= \sum_{j=1}^{n}\frac{\partial h_j}{\partial x_j}(F(x)-y)\det F'(x)\\
&= (\operatorname{div} H)(F(x)-y)\det F'(x)\\
&= \phi(\|F(x)-y\|_2)\det F'(x)\\
&= \Phi(x).
\end{aligned}
$$

Da Φ und P außerhalb von Ω verschwinden, ist $\operatorname{div} P(x) = \Phi(x)$ für alle $x \in \mathbb{R}^n$. Der zweite Beweisschritt wird abgeschlossen durch den Nachweis dafür, dass $P\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig differenzierbar ist. Dass P auf Ω stetig differenzierbar ist, ist klar, da F und G auf D stetig differenzierbar sind und $\operatorname{cl}(\Omega) \subset D$ gilt. Wir überlegen uns, dass P auf einer Umgebung des Randes $\partial\Omega$ von Ω verschwindet, woraus dann die Behauptung folgt. Denn wegen der Stetigkeit von F und wegen $\alpha \in (0, d(F(\partial\Omega), y))$ existiert eine Umgebung $U(\partial\Omega)$ von $\partial\Omega$ mit $\|F(x)-y\|_2 > \alpha$ für alle $x \in U(\partial\Omega)$. Wegen $\psi \in W_\alpha$ ist für $x \in U(\partial\Omega)$ dann

$$
G(x) = \underbrace{\psi(\underbrace{\|F(x)-y\|_2}_{>\alpha})}_{=0}(F(x)-y) = 0.
$$

Folglich ist auch $P(x) = 0$ für $x \in U(\partial\Omega)$, womit der zweite Beweisschritt abgeschlossen ist.

Im dritten Beweisschritt überlegen wir uns zunächst die Gültigkeit der folgenden Hilfsaussage:

(3) Sei $P\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$ eine stetig differenzierbare Funktion mit kompaktem Träger. Dann ist
$$
\int_{\mathbb{R}^n} \operatorname{div} P(x)\,dx = 0.
$$

Denn: Der Träger von P sei in $Q := \{x \in \mathbb{R}^n : \|x\|_\infty \le \alpha\}$ enthalten. Dann ist

$$
\begin{aligned}
\int_{\mathbb{R}^n} \operatorname{div} P(x)\,dx &= \int_Q \operatorname{div} P(x)\,dx\\
&= \sum_{i=1}^{n}\int_{-\alpha}^{\alpha}\cdots\int_{-\alpha}^{\alpha}\frac{\partial p_i}{\partial x_i}(x)\,dx_1\cdots dx_n\\
&= 0
\end{aligned}
$$

wegen

$$
\int_{-\alpha}^{\alpha}\frac{\partial p_i}{\partial x_i}(x)\,dx_i = p_i(\alpha) - p_i(-\alpha) = 0.
$$

Da die im zweiten Beweisschritt konstruierte Abbildung den Voraussetzungen der letzten Aussage genügt, ist

$$0 = \int_{\mathbb{R}^n} \operatorname{div} P(x)\, dx = \int_{\mathbb{R}^n} \Phi(x)\, dx = d_\phi(F, \Omega, y).$$

Damit ist Satz 9.4 (mit Lücken) bewiesen. □

Beweis von Satz 9.3: Sei $\alpha \in (0, d(F(\partial\Omega), y))$ und seien $\phi_1, \phi_2 \in W_\alpha^1$. Wie in Satz 9.4 sei $\eta(\phi) := \int_0^\infty s^{n-1}\phi(s)\, ds$ für $\phi \in W_\alpha$. Dann ist

$$\phi := \eta(\phi_1)\phi_2 - \eta(\phi_2)\phi_1 \in W_\alpha$$

und

$$\eta(\phi) = \eta(\phi_1)\eta(\phi_2) - \eta(\phi_2)\eta(\phi_1) = 0.$$

Nun wende man Satz 9.4 mit der Identität $I\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$ (statt F) und

$$\Omega^0 := \{z \in \mathbb{R}^n : \|z - y\|_2 > 2\alpha\}$$

(statt Ω) an. Für $x \in \partial\Omega^0$ ist $\|x - y\|_2 = 2\alpha > \alpha$ und daher ist $y \notin I(\partial\Omega^0) = \partial\Omega^0$ und $\alpha \in (0, d(I(\partial\Omega^0), y))$, daher sind die Voraussetzungen von Satz 9.4 erfüllt und wir erhalten

$$0 = d_\phi(I, \Omega^0, y) = \int_{\mathbb{R}^n} \phi(||x - y||_2)\, dx = \int_{\mathbb{R}^n} \phi(\|x\|_2)\, dx.$$

Damit wird

$$\begin{aligned}
0 &= \int_{\mathbb{R}^n} \phi(\|x\|_2)\, dx \\
&= \eta(\phi_1) \underbrace{\int_{\mathbb{R}^n} \phi_2(\|x\|_2)\, dx}_{=1} - \eta(\phi_2) \underbrace{\int_{\mathbb{R}^n} \phi_1(\|x\|_2)\, dx}_{=1} \\
&\qquad \text{(wegen } \phi_1, \phi_2 \in W_\alpha^1\text{)} \\
&= \eta(\phi_1) - \eta(\phi_2) \\
&= \eta(\phi_1 - \phi_2).
\end{aligned}$$

Wiederum nach Satz 9.4 ist

$$0 = d_{\phi_1 - \phi_2}(F, \Omega, y) = d_{\phi_1}(F, \Omega, y) - d_{\phi_2}(F, \Omega, y).$$

Damit ist Satz 9.3 bewiesen. □

Aus Satz 9.3 folgt insbesondere: Sind $\alpha_1, \alpha_2 \in (0, d(F(\partial\Omega), y))$ und $\phi_i \in W_{\alpha_i}^1$, $i = 1, 2$, so ist $d_{\phi_1}(F, \Omega, y) = d_{\phi_2}(F, \Omega, y)$. Denn sei z. B. $\alpha_1 < \alpha_2$. Dann ist $\phi_1, \phi_2 \in W_{\alpha_2}^1$, aus Satz 9.3 folgt die Behauptung. Damit ist die folgende Definition sinnvoll.

Definition 9.5 Sei $F\colon D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig differenzierbar auf der offenen Menge D und Ω eine offene, beschränkte Menge mit cl $(\Omega) \subset D$. Dann ist der *Brouwersche Abbildungsgrad* $d(F, \Omega, y)$ von F in einem Punkt $y \notin F(\partial\Omega)$ in Bezug auf Ω definiert durch

$$d(F, \Omega, y) := d_\phi(F, \Omega, y),$$

wobei $\phi \in W^1_\alpha$ mit $\alpha \in (0, d(F(\partial\Omega), y))$ beliebig ist[24] .

Der Abbildungsgrad ist bisher nur für stetig differenzierbare Funktionen definiert worden. Wir zeigen zunächst, dass der abbildungsgrad $d(F, \Omega, y)$ stetig von F abhängt und übertragen die Definition des Abbildungsgrades auf stetiges F mit Hilfe des Weierstraßschen Approximationssatzes.

Satz 9.6 *Seien $F, G\colon D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ zwei stetig differenzierbare Abbildungen auf der offenen Menge D und Ω eine offene, beschränkte Menge mit* cl $(\Omega) \subset D$. *Sei ferner $y \notin F(\partial\Omega)$. Ist dann $\alpha \in (0, d(F(\partial\Omega), y))$ und $\sup_{x\in \text{cl }(\Omega)} \|F(x) - G(x)\|_2 < \frac{1}{7}\alpha$, so ist $d(F, \Omega, y) = d(G, \Omega, y)$.*

Beweis: Sei $\alpha_0 := \frac{1}{7}\alpha$ und $\mu\colon [0, \infty) \longrightarrow [0, 1]$ stetig differenzierbar auf $[0, \infty)$ mit $\mu(t) = 1$ für $t \in [0, 2\alpha_0]$ und $\mu(t) = 0$ für $t \geq 3\alpha_0$. Dann ist die Abbildung $H\colon D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$, definiert durch

$$H(x) := (1 - \mu(\|F(x) - y\|_2))F(x) + \mu(\|F(x) - y\|_2)G(x),$$

stetig differenzierbar auf D, was durch Inspektion ersichtlich ist. Ferner ist

$$\|H(x) - F(x)\|_2 = \underbrace{\mu(\|F(x) - y\|_2)}_{\in[0,1]} \|G(x) - F(x)\|_2 < \alpha_0 \quad \text{für alle } x \in \text{cl }(\Omega)$$

und daher

$$\|H(x) - y\|_2 \geq \underbrace{\|F(x) - y\|_2}_{\geq d(F(\partial\Omega),y) > 7\alpha_0} - \underbrace{\|H(x) - F(x)\|_2}_{<\alpha_0} > 6\alpha_0 \quad \text{für alle } x \in \partial\Omega,$$

folglich $d(H(\partial\Omega), y) > 6\alpha_0$. Mit einem beliebigen $\phi \in W^1_{6\alpha_0}$ ist also $d(H, \Omega, y) = d_\phi(H, \Omega, y)$ wohldefiniert. Der Beweis erfolgt nun in zwei Schritten.

(a) Es ist $d(H, \Omega, y) = d(F, \Omega, y)$.

Denn: Man wähle $\phi_1 \in W^1_{5\alpha_0}$ mit $\phi_1(t) = 0$ für $t \in [0, 4\alpha_0]$. Für alle $x \in$ cl (Ω) ist dann

$$\|H(x) - y\|_2 \leq \|F(x) - y\|_2 + \underbrace{\|F(x) - H(x)\|_2}_{<\alpha_0} < \|F(x) - y\|_2 + \alpha_0,$$

sodass $\phi_1(\|F(x) - y\|_2) = \phi_1(\|H(x) - y\|_2) = 0$, falls $\|F(x) - y\|_2 < 3\alpha_0$. Ist dagegen $\|F(x) - y\|_2 \geq 3\alpha_0$, so ist $\mu(\|F(x) - y\|_2) = 0$ (nach Definition von μ) und folglich $H(x) = F(x)$ (nach Definition von H). Daher ist

$$\phi_1(\|H(x) - y\|_2) \det H'(x) = \phi_1(\|F(x) - y\|_2) \det F'(x) \quad \text{für alle } x \in \text{cl }(\Omega)$$

und folglich

$$d(H, \Omega, y) = d_{\phi_1}(H, \Omega, y) = d_{\phi_1}(F, \Omega, y) = d(F, \Omega, y).$$

[24]Es treten (in Vergangenheit, Gegenwart und Zukunft) hoffentlich keine Missverständnisse auf. Der Abstand des Punktes y zu der Menge $F(\partial\Omega)$ bezeichnen wir mit $d(F(\partial\Omega), y)$, während der Abbildungsgrad von F in einem Punkt $y \notin F(\partial\Omega)$ bezüglich Ω durch $d(F, \Omega, y)$ bezeichnet wird.

(b) Es ist $d(H,\Omega,y) = d(G,\Omega,y)$.

Denn: Für alle $x \in \partial\Omega$ ist

$$\|G(x)-y\|_2 \geq \underbrace{\|F(x)-y\|_2}_{\geq d(F(\partial\Omega),y)>7\alpha_0} - \underbrace{\|F(x)-G(x)\|_2}_{<\alpha_0} > \|F(x)-y\|_2 - \alpha_0 > 6\alpha_0.$$

Sei $\phi_2 \in W^1_{\alpha_0}$. Wegen $d(G(\partial\Omega),y) > 6\alpha_0 > \alpha_0$ ist dann $d(G,\Omega,y) = d_{\phi_2}(G,\Omega,y)$. Ist $\|F(x)-y\|_2 > 2\alpha_0$, so ist $\|G(x)-y\|_2 > 2\alpha_0 - \alpha_=\alpha_0$ und entsprechend auch $\|H(x)-y\|_2 > \alpha_0$, folglich

$$\phi_2(\|G(x)-y\|_2) = \phi_2(\|H(x)-y\|_2) = 0.$$

Ist dagegen $\|F(x)-y\|_2 \leq 2\alpha_0$, so ist $\mu(\|F(x)-y\|_2) = 1$ und folglich $H(x) = G(x)$, also

$$\phi_2(\|G(x)-y\|_2)\det G'(x) = \phi_2(\|H(x)-y\|_2)\det H'(x) \quad \text{für alle } x \in \text{cl}\,(\Omega).$$

Daher ist

$$d(H,\Omega,y) = d_{\phi_2}(H,\Omega,y) = d_{\phi_2}(G,\Omega,y) = d(G,\Omega,y).$$

insgesamt ist der Satz bewiesen. □

Wir definieren nun den Brouwerschen Abbildungsgrad $d(F,\Omega,y)$ für *stetiges* F und zeigen im anschließenden Satz, dass diese Definition sinnvoll ist.

Definition 9.7 Sei $F\colon \text{cl}\,(\Omega) \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig, wobei $\Omega \subset \mathbb{R}^n$ eine offene, beschränkte Menge ist. Sei $y \notin F(\partial\Omega)$. Sind dann $F_k\colon D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$, $k \in \mathbb{N}$, Abbildungen, welche stetig differenzierbar auf der offenen Menge $D \supset \text{cl}\,(\Omega)$ sind und für die

$$\lim_{k\to\infty} \|F_k - F\| = \lim_{k\to\infty} \sup_{x\in\text{cl}\,(\Omega)} \|F_k(x) - G_k(x)\|_2 = 0$$

(d. h. $\{F_k\}$ konvergiert auf cl (Ω) gleichmäßig gegen F), so wird der *Brouwersche Abbildungsgrad* von F in y bezüglich Ω definiert durch

$$d(F,\Omega,y) := \lim_{k\to\infty} d(F_k,\Omega,y).$$

Satz 9.8 *Sei $F\colon \text{cl}\,(\Omega) \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig, wobei $\Omega \subset \mathbb{R}^n$ eine offene, beschränkte Menge ist. Sei $y \notin F(\partial\Omega)$. Dann ist der Brouwersche Abbildungsgrad $d(F,\Omega,y)$ wohldefiniert.*

Beweis: Im ersten Teil des Beweises überlegen wir uns, dass eine Folge $\{F_k\}$ von Abbildungen $F_k\colon D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ existiert, welche stetig differenzierbar sind auf der offenen Menge $D \supset \text{cl}\,(\Omega)$ und welche auf cl (Ω) gleichmäßig gegen F konvergiert. Dies ergibt sich aber sofort aus dem *Weierstraßschen Approximationssatz*, der aussagt, dass eine stetige reellwertige Funktion auf einer kompakten Menge cl $(\Omega) \subset \mathbb{R}^n$ gleichmäßig durch eine Folge von Polynomen (in n Variablen) approximiert werden kann. Das entsprechende gilt dann natürlich auch für stetige Vektorfunktionen $F\colon \text{cl}\,(\Omega) \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$. Im zweiten Teil des Beweises zeigen wir, dass $d(F_k,\Omega,y)$ für alle hinreichend großen k

wohldefiniert ist und $\lim_{k\to\infty} d(F_k, \Omega, y)$ existiert. Sei $\alpha \in (0, d(\partial\Omega), y)$. Für alle $x \in \partial\Omega$ ist

$$\|F_k(x) - y\|_2 \geq \|F(x) - y\|_2 - \|F_k(x) - F(x)\|_2 \geq d(F(\partial\Omega), y) - \|F_k(x) - F(x)\|_2$$

und daher $d(F_k(\partial\Omega), y) > \alpha$ für alle hinreichend großen k, etwa alle $k \geq k_0$. Damit ist $d(F_k, \Omega, y)$ für alle $k \geq k_0$ wohldefiniert. Als konvergente Folge ist $\{F_k\}$ eine Cauchyfolge, damit existiert ein $k_1 \geq k_0$ mit

$$\|F_k - F_j\| = \sup_{x\in \mathrm{cl}\,(\Omega)} \|F_k(x) - F_j(x)\|_2 < \frac{1}{7}\alpha \quad \text{für alle } k, j \geq k_1.$$

Wegen Satz 9.6 ist $d(F_k, \Omega, y) = d(F_{k_1}, \Omega, y)$ für alle $k \geq k_1$, sodass $\lim_{k\to\infty} d(F_k, \Omega, y)$ trivialerweise existiert. Im dritten Teil des Beweises bleibt zu zeigen, dass $d(F, \Omega, y)$ von der Wahl der Folge $\{F_k\}$ unabhängig ist. Sei daher $\{G_k\}$ eine weitere Folge von Abbildungen $G_k \colon D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$, die auf der offenen Menge $\subset \mathrm{cl}\,(\Omega)$ stetig differenzierbar sind und für die $\lim_{k\to\infty} \|G_k - F\| = 0$. Dann ist

$$\|F_k - G_j\| \leq \|F_k - F\| + \|F - G_j\| < \frac{1}{7}\alpha$$

für alle hinreichend großen k, j und damit nach Satz 9.6 auch $d(F_k, \Omega, y) = d(G_j, \Omega, y)$ für alle hinreichend großen k, j. Der Satz ist damit bewiesen. □

9.2 Eigenschaften des Brouwerschen Abbildungsgrades

Wichtig ist die Beantwortung der Frage, unter welchen Variationen an die Abbildung F der Abbildungsgrad $d(F, \Omega, y)$ konstant bleibt. Satz 9.6 gab eine Antwort für den Fall, dass F stetig differenzierbar ist. Dieses Ergebnis wird im folgenden Satz auf stetiges F übertragen.

Satz 9.9 *Sei $\Omega \subset \mathbb{R}^n$ offen und beschränkt, $F \colon \mathrm{cl}\,(\Omega) \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig, $y \notin F(\partial\Omega)$ und $\alpha \in (0, d(F(\partial\Omega), y))$. Ist auch $G \colon \mathrm{cl}\,(\Omega) \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig und*

$$\|G - F\| = \sup_{x\in \mathrm{cl}\,(\Omega)} \|G(x) - F(x)\|_2 < \frac{1}{7}\alpha,$$

so ist $d(F, \Omega, y) = d(G, \Omega, y)$.

Beweis: Der Beweis ist einfach. Haupthilfsmittel sind die Dreiecksungleichung und Satz 9.6, siehe auch J. M. Ortega, W. C. Rheinboldt (1970, 6.2.1. auf S. 156). □

Eine wichtige Folgerung ist der folgende Satz.

Satz 9.10 (Homotopie-Invarianz) *Sei $\Omega \subset \mathbb{R}^n$ offen, beschränkt und*

$$H \colon \mathrm{cl}\,(\Omega) \times [0, 1] \subset \mathbb{R}^n \times \mathbb{R} \longrightarrow \mathbb{R}^n$$

stetig. Ferner sei $y \neq H(x, t)$ für alle $(x, t) \in \partial\Omega \times [0, 1]$. Dann ist $d(H(\cdot, t), \Omega, y)$ konstant für $t \in [0, 1]$.

Beweis: Wir definieren

$$J := \{t \in [0,1] : d(H(\cdot,t),\Omega,y) = d(H(\cdot,0),\Omega,y)\}.$$

Dann ist $J \neq \emptyset$, da $0 \in J$. Auf der kompakten Menge cl $(\Omega) \times [0,1]$ ist die stetige Abbildung H gleichmäßig stetig. daher existiert ein $\delta > 0$ derart, dass

$$\|H(\cdot,s) - H(\cdot,t)\|_2 < \tfrac{1}{7}\alpha \quad \text{für alle } s,t \in [0,1] \text{ mit } |s-t| < \delta,$$

wobei

$$0 < \alpha \leq \min_{(x,t)\in\partial\Omega\times[0,1]} \|H(x,t) - y\|_2.$$

Daher ist $\alpha \in (0, d(H(\partial\Omega,t),y))$ für alle $t \in [0,1]$. Wegen Satz 9.9 gilt

$$s,t \in [0,1],\ |s-t] < \delta \Longrightarrow d(H(\cdot,s),\Omega,y) = d(H(\cdot,t),\Omega,y).$$

Hieraus folgt $[0, \frac{1}{2}\delta] \subset J$, $[\frac{1}{2}\delta, \delta] \subset J$ usw. Nach endlich vielen Schritten erhält man, dass $[0,1] \subset J$ und das war zu zeigen. □

Ein Spezialfall von Satz 9.10 ist

Satz 9.11 *Seien F, G: cl $(\Omega) \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ zwei stetige Abbildungen und $\Omega \subset \mathbb{R}^n$ offen und beschränkt. Sei $y \in \mathbb{R}^n$ ein Punkt mit*

$$y \notin \{(1-t)F(x) + tG(x) : (x,t) \in \partial\Omega \times [0,1]\}.$$

Dann ist $d(F,\Omega,y) = d(G,\Omega,y)$.

Beweis: Man definiere H: cl $(\Omega) \times [0,1] \longrightarrow \mathbb{R}^n$ durch $H(x,t) := (1-t)F(x) + tG(x)$ und wende Satz 9.10 an. □

Bemerkungen: 1. Satz 9.11 zeigt, dass der Abbildungsgrad $d(F,\Omega,y)$ nur von den Werten von F auf $\partial\Omega$, nicht aber von den Werten von F in Ω abhängt. Denn offenbar gilt:

- *Sei $\Omega \subset \mathbb{R}^n$ offen, beschränkt und F: cl $(\Omega) \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig. Sei ferner G: cl $(\Omega) \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ eine stetige Abbildung mit $F(x) = G(x)$ für alle $x \in \partial\Omega$. Für jedes $y \notin F(\partial\Omega)$ ist dann $d(F,\Omega,y) = d(G,\Omega,y)$.*

2. Eine weitere Folgerung aus Satz 9.11 ist:

- *Seien F, G: cl $(\Omega) \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ zwei stetige Abbildungen, wobei $\Omega \subset \mathbb{R}^n$ offen, beschränkt. Mit einem $y \in \mathbb{R}^n$ gelte*

  $$\|F(x) - G(x)\|_2 < \|F(x) - y\|_2 \quad \text{für alle } x \in \partial\Omega.$$

  Dann ist $d(F,\Omega,y) = d(G,\Omega,y)$.

Denn: Angenommen, es existiert $(x,t) \in \partial\Omega \times [0,1]$ mit $y = (1-t)F(x) + tG(x)$. Dann ist

$$\|y - F(x)\|_2 = t\,\|F(x) - G(x)\|_2 \leq \|F(x) - G(x)\|_2,$$

ein Widerspruch.

3. Eine weitere Folgerung aus Satz 9.11 ist:

- *Seien f, g: cl $(\Omega) \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ zwei stetige Abbildungen, wobei $\Omega \subset \mathbb{R}^n$ offen, beschränkt. Es sei $|g(x)| < |f(x)|$ für alle $x \in \partial\Omega$. Dann ist $d(f+g, \Omega, 0) = d(f, \Omega, 0)$.*

Denn: In Satz 9.11 setze man $F := f$, $G := f + g$ und $y := 0$. □

Bisher wurde der Einfluss der Abbildung F auf den Abbildungsgrad $d(F, \Omega, y)$, in den nächsten Sätzen wird die dagegen die Abhängigkeit des Abbildungsgrades $d(F, \Omega, y)$ von y und Ω untersucht.

Satz 9.12 *Sei F: cl $(\Omega) \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig und $\Omega \subset \mathbb{R}^n$ eine offene, beschränkte Menge. Ist $y \notin F(\partial\Omega)$ und $z \in \mathbb{R}^n$ beliebig, so ist $d(F - z, \Omega, y - z) = d(F, \Omega, y)$. Hierbei ist die Abbildung $F - z$ natürlich durch $(F - z)(x) := F(x) - z$ definiert.*

Beweis: Es genügt, die Behauptung für eine auf einer offenen Menge $D \supset$ cl (Ω) stetig differenzierbare Abbildung F nachzuweisen. Sei $G := F - z$ und $\phi \in W_\alpha^1$ mit $\alpha \in (0, d(F(\partial\Omega), y))$. Wegen

$$\phi(\|G(x) - (y - z)\|_2) \det G'(x) = \phi(\|F(x) - y\|_2) \det F'(x)$$

folgt die Behauptung. □

Satz 9.13 *Sei F: cl $(\Omega) \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig, $\Omega \subset \mathbb{R}^n$ eine offene, beschränkte Menge. Seien $y^0, y^1 \in \mathbb{R}^n$ zwei Punkte, welche durch eine stetige Kurve p: $[0,1] \subset \mathbb{R} \longrightarrow \mathbb{R}^n$ verbunden werden können (d. h. es ist $p(0) = y^0$, $p(1) = y^1$), die $F(\partial\Omega)$ nicht trifft (d. h. es es ist $p(t) \notin F(\partial\Omega)$ für alle $t \in [0,1]$). Dann ist $d(F, \Omega, y^0) = d(F, \Omega, y^1)$.*

Beweis: Man definiere die stetige Abbildung H: cl $(\Omega) \times [0,1] \subset \mathbb{R}^n \times \mathbb{R} \longrightarrow \mathbb{R}^n$ durch $H(x,t) := F(x) - p(t)$. Dann erhalten wir

$$\begin{aligned}
d(F, \Omega, y^0) &= d(F - y^0, \Omega, 0) \\
&\qquad \text{(Satz 9.12)} \\
&= d(H(\cdot, 0), \Omega, 0) \\
&= d(H(\cdot, 1), \Omega, 0) \\
&\qquad \text{(Satz 9.10)} \\
&= d(F - y^1, \Omega, 0) \\
&= d(F, \Omega, y^1) \\
&\qquad \text{(Satz 9.12)},
\end{aligned}$$

womit die Behauptung bewiesen ist. □

Satz 9.14 *Sei F: cl $(\Omega) \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig und $\Omega \subset \mathbb{R}^n$ wie auch $\Omega_1, \ldots, \Omega_m \subset \Omega$ offene, beschränkte Mengen mit $\Omega_i \cap \Omega_j = \emptyset$ für $i \neq j$ und cl $(\Omega) = \bigcup_{j=1}^m$ cl (Ω_j). Ist dann $y \notin \bigcup_{j=1}^m F(\partial\Omega_j)$, so ist $d(F, \Omega, y) = \sum_{j=1}^m d(F, \Omega_j, y)$.*

Beweis: Aus $y \notin \bigcup_{j=1}^m F(\partial\Omega_j)$ folgt $y \notin F(\partial\Omega_j)$, $j = 1, \ldots, m$. Daher ist $d(F, \Omega_j, y)$ wohldefiniert, $j = 1, \ldots, m$. Andererseits zeigt eine leichte Überlegung, dass $\partial\Omega \subset \bigcup_{j=1}^m \partial\Omega_j$, sodass auch $y \notin F(\partial\Omega)$ und damit $d(F, \Omega, y)$ wohldefiniert ist. Wiederum

genügt es, die Behauptung für eine auf einer offenen Menge $D \supset \mathrm{cl}\,(\Omega)$ stetig differenzierbare Abbildung F nachzuweisen. Wir beachten Definition 9.1, wählen $\phi \in W_\alpha^1$ mit hinreichend kleinem $\alpha > 0$ und benutzen die Additivität des Integrals. □

Der folgende Satz (siehe J. M. Ortega, W. C. Rheinboldt (1970, S. 158)) sagt aus, dass der Abbildungsgrad $d(F, \Omega, y)$ konstant bleibt, wenn man aus Ω eine abgeschlossene Menge Q herausnimmt, die einen leeren Durchschnitt mit $\{x \in \Omega : F(x) = y\}$ hat.

Satz 9.15 *Sei $F\colon \mathrm{cl}\,(\Omega) \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig, wobei $\Omega \subset \mathbb{R}^n$ eine offene, beschränkte Menge ist. Sei $y \notin F(\partial\Omega)$. Dann gilt:*

1. *Ist $Q \subset \mathrm{cl}\,(\Omega)$ eine abgeschlossene Menge mit $y \notin F(Q)$, so ist $d(F, \Omega, y) = d(F, \Omega \setminus Q, y)$.*
2. *Ist $y \notin F(\Omega)$, so ist $d(F, \Omega, y) = 0$.*

Beweis: Die Abbildung F kann als auf einer offenen Menge $D \supset \mathrm{cl}\,(\Omega)$ stetig differenzierbare Abbildung angenommen werden, da der allgemeine Fall wieder durch Approximation von F durch stetig differenzierbare Abbildungen folgt. Da Q kompakt ist und $y \notin F(Q)$, ist $\eta := d(F(Q), y) > 0$. Nun wähle man $\alpha \in (0, \min(\eta, d(F(\partial\Omega), y)))$ und $\phi \in W_\alpha^1$. Dann ist $\phi(\|F(x) - y\|_2) = 0$ für $x \in Q$, folglich $d_\phi(F, \Omega, y) = d_\phi(F, \Omega \setminus Q, y)$ und damit $d(F, \Omega, y) = d(F, \Omega \setminus Q, y)$. Ist $y \notin F(\Omega)$ bzw. dann sogar $y \notin F(\mathrm{cl}\,(\Omega))$, so ist $\phi(\|F(x) - y\|_2) = 0$ für $\phi \in W_\alpha^1$ mit $\alpha \in (0, d(F(\mathrm{cl}\,(\Omega)), y))$, folglich $d_\phi(F, \Omega, y) = 0$ bzw. $d(F, \Omega, y) = 0$ (siehe auch Satz 9.2 (b)). □

Von den wesentlichen Eigenschaften des Abbildungsgrades ist nur noch die Ganzzahligkeit nachzuweisen. Dies wird wieder schwieriger, wobei die Hauptarbeit in dem Beweis eines Lemmas von Sard steckt.

Lemma 9.16 (Sard) *Sei $F\colon \Omega \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ auf der offenen Menge Ω stetig differenzierbar und $C := \{x \in \Omega : F'(x)$ ist singulär$\}$. Dann hat $F(C)$ das Maß Null.*

Beweis: Zur Erinnerung:

- Eine Menge $M \subset \mathbb{R}^n$ hat das *Maß Null*, falls es zu gegebenem $\epsilon > 0$ abzählbar viele Quader Q_j, $j \in \mathbb{N}$, mit Inhalt q_j gibt derart, dass $M \subset \bigcup_{j=1}^\infty Q_j$ und $\sum_{j=1}^\infty q_j \leq \epsilon$. Unter einem *Quader* (oder auch *Hyperrechteck*) versteht man dabei eine Menge Q der Form

$$Q := \Big\{x^0 + \sum_{j=1}^n \alpha_j h^j : \alpha_j \in [0,1],\ j = 1, \ldots, n\Big\},$$

  wobei $x^0 \in \mathbb{R}^n$ und $h^1, \ldots, h^n \in \mathbb{R}^n$ paarweise orthogonal und von Null verschieden sind. Ist $\gamma := \|h^1\|_2 = \cdots = \|h^n\|_2$, so sprechen wir statt von einem Quader von einem *Kubus* der *Seitenlänge* γ. Der *Inhalt* q eines Quaders Q ist definiert durch $q := \prod_{j=1}^n \|h^j\|_2$. In Abbildung 6 veranschaulichen wir uns einen Quader im $\mathbb{R}^2$. Man beachte, dass die abzählbare Vereinigung von Mengen vom Maß Null ebenfalls das Maß Null besitzt.

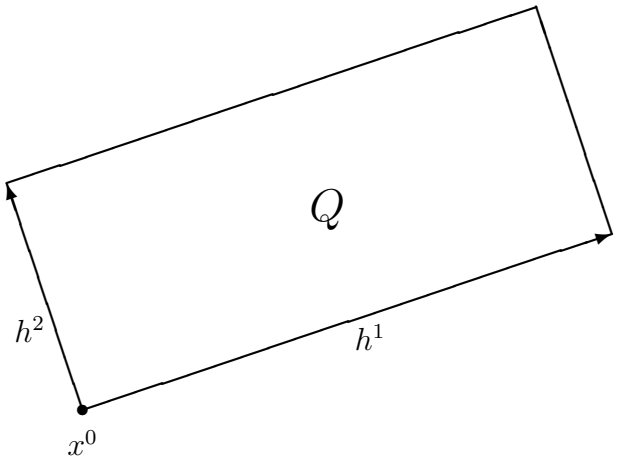

Abbildung 6: Ein Quader im $\mathbb{R}^2$

Die offene Menge $\Omega \subset \mathbb{R}^n$ kann als abzählbare Vereinigung von Kuben $Q_j \subset \Omega$, $j \in \mathbb{N}$, geschrieben werden. Mit $C_j := \{x \in Q_j : F'(x)$ ist singulär genügt es zu zeigen, dass $F(C_j)$ das Maß Null besitzt, $j \in \mathbb{N}$.
Der Gang des Beweises ist nun der folgende.

1. Seien ein Kubus $Q \subset \Omega$ und ein $\epsilon > 0$ gegeben. Sei

$$C := \{x \in Q : F'(x) \text{ ist nichtsingulär}\}.$$

   Unser Ziel besteht darin nachzuweisen, dass $F(C)$ das Maß Null besitzt.

2. Der Kubus Q habe die Seitenlänge γ. Man teile Q in m^n (kleine) Kuben P_j der Seitenlänge γ/m. Dies wird für $n = 2$ und $m = 3$ in Abbildung 7 dargestellt. Sei

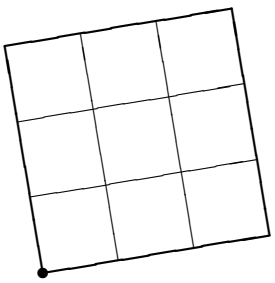

Abbildung 7: Unterteilung eines Kubus

   $u \in C$ beliebig. Dann ist $u \in Q$ in einem der kleinen Kuben der Seitenlänge γ/m enthalten, sei dies etwa P_u. Der Durchmesser von P_u kann durch $\sqrt{n}\gamma/m$ nach oben abgeschätzt werden. Denn ist

$$P_u = \left\{ y^0 + \sum_{j=1}^{n} \alpha_j k^j : \alpha_j \in [0,1],\ j = 1, \ldots, n \right\}$$

mit $y^0 \in \mathbb{R}^n$ und paarweise orthogonalen $k^1, \ldots, k^n$ mit $\|k^1\|_2 = \cdots = \|k^n\|_2 = \gamma/m$, so ist für beliebiges $x \in P_u$ offenbar[25] $\|x-u\|_2 \leq \sqrt{n}\gamma/m$ und folglich

$$P_u \subset \{x \in \mathbb{R}^n : \|x-u\|_2 \leq \sqrt{n}\gamma/m\}.$$

Da F in u differenzierbar ist, gibt es zu $\epsilon > 0$ ein $\delta > 0$ mit

$$\|x-u\|_2 \leq \delta \Longrightarrow \|F(x) - F(u) - F'(u)(x-u)\|_2 \leq \epsilon \|x-u\|_2.$$

Man wähle m nun so groß, dass $\sqrt{n}\gamma/m \leq \delta$. Dann ist

$$\|F(x) - F(u) - F'(u)(x-u)\|_2 \leq \epsilon \|x-u\|_2 \leq \epsilon\sqrt{n}\gamma/m \quad \text{für alle } x \in P_u.$$

3. Unser nächstes Ziel besteht darin, $F(P_u)$ in einem möglichst kleinen Quader (bzw. einer flachen Dose) einzuschließen und damit das Maß von $F(P_u)$ nach oben durch den Inhalt dieses kleinen Quaders abzuschätzen. Hierzu definiere man die affin lineare Abbildung $B\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$ durch $B(x) := F(u) + F'(u)(x-u)$. Da $F'(u)$ singulär ist, liegt $B(P_u)$ in einem affinen Teilraum des $\mathbb{R}^n$ mit einer Dimension $\leq n-1$. Wegen

$$\|F(x) - B(x)\|_2 \leq \epsilon\sqrt{n}\gamma/m \quad \text{für alle } x \in P_u$$

ist $F(P_u)$ in einer $\epsilon\sqrt{n}\gamma/m$-Umgebung von $B(P_u)$ enthalten. Ferner ist

$$\|B(x) - F(u)\|_2 \leq \beta \|x-u\|_2 \leq \beta\sqrt{n}\gamma/m \quad \text{für alle } x \in P_u,$$

wobei $\beta := \max_{x \in Q} \|F'(x)\|$. Da $B(P_u)$ in einem affinen Teilraum des $\mathbb{R}^n$ einer Dimension $\leq n-1$ enthalten ist, existiert eine Hyperebene H im $\mathbb{R}^n$ mit $B(P_u) \subset H$. Wir nehmen an, H habe die Darstellung

$$H = \{x \in \mathbb{R}^n : (\tilde{h}^1)^T x = c\}$$

mit $\|\tilde{h}^1\|_2 = 1$ und $c \in \mathbb{R}$. Man ergänze $\tilde{h}^1$ durch $\tilde{h}^2, \ldots, \tilde{h}^n$ zu einem (vollständigen) Orthonormalsystem des $\mathbb{R}^n$. Sei $x \in P_u$ beliebig. Bezüglich des Orthonormalsystems $\{\tilde{h}^1, \ldots, \tilde{h}^n\}$ hat $F(x) - F(u) \in \mathbb{R}^n$ eine eindeutige Darstellung

$$F(x) - F(u) = \sum_{j=1}^{n} \lambda_j \tilde{h}^j$$

[25] Denn: Seien

$$x = y^0 + \sum_{j=1}^{n} \alpha_j k^j, \qquad u = y^0 + \sum_{j=1}^{n} \beta_j k^j$$

mit $\alpha_j, \beta_j \in [0,1]$, $j = 1, \ldots, n$. Dann ist

$$\|x-u\|_2 = \frac{\gamma}{m}\left(\sum_{j=1}^{n} (\alpha_j - \beta_j)^2\right)^{1/2} \leq \frac{\gamma}{m}\sqrt{n}.$$

mit $\lambda_j \in \mathbb{R}$, $j = 1, \ldots, n$. Wir schätzen $|\lambda_j| = (\tilde{h}^j)^T(F(x) - F(u)|$ ab, $j = 1, \ldots, n$. Der Fall $j = 1$ spielt eine Sonderrolle. Hier berücksichtigen wir nämlich, dass

$$F'(u)(x-u) = B(x) - B(u) \in B(P_u) - B(P_u)$$

als Differenz zweier Elemente aus $B(P_u)$ auf einer zu H parallelen Hyperebene durch den Nullpunkt liegt bzw. $(\tilde{h}^1)^T F'(u)(x-u) = 0$ gilt. Daher ist

$$\begin{aligned} |\lambda_1| &= |(\tilde{h}^1)^T(F(x) - F(u))| \\ &= |(\tilde{h}^1)^T(F(x) - F(u) - F'(u)(x-u)| \\ &\leq \underbrace{\|\tilde{h}^1\|_2}_{=1} \|F(x) - F(u) - F'(u)(x-u)\|_2 \\ &\leq \epsilon\sqrt{n}\gamma/m. \end{aligned}$$

Für $j = 2, \ldots, n$ ist ferner

$$\begin{aligned} |\lambda_j| &= |(\tilde{h}^j)^T(F(x) - F(u)\|_2 \\ &\leq \underbrace{\|\tilde{h}^j\|_2}_{=1} \|F(x) - F(u)\|_2 \\ &\leq \|F(x) - F(u) - F'(u)(x-u)\|_2 + \|F'(u)(x-u)\|_2 \\ &\leq \epsilon\sqrt{n}\gamma/m + \beta\sqrt{n}\gamma/m \\ &= (\epsilon + \beta)\sqrt{n}\gamma/m. \end{aligned}$$

Nun definiere man $h^1, h^2, \ldots, h^n$ durch

$$h^1 := \text{sign}\,(\lambda_1)(\epsilon\sqrt{n}\gamma/m)\tilde{h}^1, \quad h^j := \text{sign}\,(\lambda_j)((\epsilon+\beta)\sqrt{n}\gamma/m)\tilde{h}^j, \quad j = 2, \ldots, n.$$

Hierbei vereinbaren wir, dass $\text{sign}\,(\lambda_j) = 1$, falls $\lambda_j = 0$. Dann lässt sich $F(x) \in F(P_u)$ darstellen als

$$F(x) = F(u) + \sum_{j=1}^{n} \lambda_j \tilde{h}^j = F(u) + \sum_{j=1}^{n} \alpha_j h^j$$

mit $\alpha_j \in [0, 1]$, $j = 1, \ldots, n$. Also ist

$$F(P_u) \subset Q_u := \Big\{ F(u) + \sum_{j=1}^{n} \alpha_j h^j : \alpha_j \in [0,1],\ j = 1, \ldots, n \Big\}.$$

Weiter ist

$$\text{Inhalt}\,(Q_u) = \prod_{j=1}^{n} \|h^j\|_2 = \epsilon(\epsilon+\beta)^{n-1}(\sqrt{n}\gamma/m)^n.$$

Damit ist es uns gelungen, $F(P_u)$ in einer flachen Dose einzuschließen.

4. Zum Finale des Beweises beachten wir, dass C in höchstens m^n solcher Kuben P_u und damit $F(C)$ in höchstens m^n von Quadern Q_u enthalten ist. Daher lässt sich das Maß von $F(C)$ durch

$$m^n \epsilon(\epsilon+\beta)^{n-1}(\sqrt{n}\gamma/m)^n = \epsilon(\epsilon+\beta)^{n-1}(\sqrt{n}\gamma)^n$$

nach oben abschätzen. Mit $\epsilon \to 0+$ erhalten wir, dass $F(C)$ das Maß Null besitzt.

Das Lemma von Sard ist bewiesen. □

Satz 9.17 *Sei $F\colon \mathrm{cl}\,(\Omega) \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig, wobei $\Omega \subset \mathbb{R}^n$ eine offene und beschränkte Menge ist. Dann ist $d(F, \Omega, y)$ für jedes $y \notin F(\partial\Omega)$ ganzzahlig.*

Beweis: Es genügt offenbar wiederum, die Behauptung für den Fall zu beweisen, dass F auf einer offenen Menge $D \supset \mathrm{cl}\,(\Omega)$ stetig differenzierbar ist.
Da $y \notin F(\partial\Omega)$ existiert eine Umgebung $U(y)$ von y mit $U(y) \cap F(\partial\Omega) = \emptyset$. Sei nun $V \subset U(y)$ ebenfalls eine Umgebung von y. Dann ist $V \cap (\mathbb{R}^n \setminus F(C)) \neq \emptyset$, wobei wieder

$$C := \{x \in \Omega : F'(x) \text{ ist singulär}\}.$$

Denn wäre $V \cap (\mathbb{R}^n \setminus F(C) = \emptyset$, so wäre $V \subset F(C)$, ein Widerspruch dazu, dass $F(C)$ wegen des Lemmas von Sard das Maß Null besitzt. In jeder hinreichend kleinen Umgebung von y existieren also Punkte, die weder zu $F(\partial\Omega)$ noch zu $F(C)$ gehören. Also existiert eine Folge $\{y^k\}$ mit $y^k \notin F(\partial\Omega) \cup F(C)$, $k \in \mathbb{N}$, und $\lim_{k\to\infty} y^k = y$. Ferner kann angenommen werden, dass die Folge $\{y^k\}$ in einer Kugel $B(y;\epsilon)$ um y mit dem Radius ϵ liegt, welche einen leeren Durchschnitt mit $F(\partial\Omega)$ hat: $B(y;\epsilon) \cap F(\partial\Omega) = \emptyset$. Nun können y^k und y durch eine stetige Kurve $p_k\colon [0,1] \subset \mathbb{R} \longrightarrow \mathbb{R}^n$ verbunden werden, welche $F(\partial\Omega)$ nicht trifft. Man definiere nämlich $p_k(t) := (1-t)y^k + ty$. Für $t \in [0,1]$ ist $p_k(t) \in B(y;\epsilon)$ und daher $p_k(t) \notin F(\partial\Omega)$, $k \in \mathbb{N}$. Nach Satz 9.13 ist $d(F, \Omega, y^k) = d(F, \Omega, y)$, $k \in \mathbb{N}$. Zu zeigen bleibt also die Ganzzahligkeit von $d(F, \Omega, y^k)$. Diese folgt aber aus Satz 9.2. Ist nämlich $\Gamma_k := \{x \in \Omega : F(x) = y^k\}$, so ist $F'(x)$ nichtsingulär für $x \in \Gamma_k$. Es gibt zwei Möglichkeiten. Ist $\Gamma_k = \emptyset$, so ist $d(F, \Omega, y^k) = 0$. Im zweiten Fall ($\Gamma_k \neq \emptyset$) ist Γ_k notwendigerweise endlich, und $d(F, \Omega, y^k) = \sum_{x\in\Gamma_k} \operatorname{sign} \det F'(x)$ ganzzahlig. Damit ist der Satz bewiesen. □

9.3 Existenzsätze für nichtlineare Gleichungssysteme

Nun kommen wir zu Existenzsätzen über die Lösbarkeit nichtlinearer Gleichungssysteme.

Satz 9.18 (Kronecker) *Sei $F\colon \mathrm{cl}\,(\Omega) \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig, wobei $\Omega \subset \mathbb{R}^n$ eine offene, beschränkte Menge ist. Ist $y \notin F(\partial\Omega)$ und $d(F, \Omega, y) \neq 0$, so besitzt die Gleichung $F(x) = y$ mindestens eine Lösung in Ω.*

Beweis: Hätte $F(x) = y$ keine Lösung in Ω, so wäre $y \notin F(\Omega)$. Der zweite Teil von Satz 9.15 ergibt $d(F, \Omega, y) = 0$, ein Widerspruch. □

Eine direkte Anwendung von Satz 9.18 ist selten möglich, da eine Berechnung des Abbildungsgrades ein nichttriviales Problem ist. Dagegen ist der Satz ein Hilfsmittel für den Beweis weiterer Existenzsätze wie des Brouwerschen Fixpunktsatzes, eines der berühmtesten Sätze der Analysis (fast wörtlich nach J. M. Ortega, W. C. Rheinboldt (1970, S. 161)). Siehe auch Abschnitt 37 bei `http://num.math.uni-goettingen.de/werner/schmankerl.pdf`, wo ein anderer Beweis angegeben wird.

Satz 9.19 (Brouwer) *Sei $G\colon K \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig auf der nichtleeren, kompakten und konvexen Menge K. Die Abbildung G bilde K in sich ab, d. h. es sei $G(K) \subset K$. Dann besitzt G einen Fixpunkt in K, d. h. es existiert ein $x^* \in K$ mit $G(x^*) = x^*$.*

Beweis: Es genügt, den Brouwerschen Fixpunktsatz für den Fall zu beweisen, dass $K := B[0;1]$ die abgeschlossene euklidische Einheitskugel ist, also der Abschluss der offenen euklidischen Einheitskugel $\Omega := B(0;1)$ ist. Dies wird z. B. in Abschnitt 37 bei `http://num.math.uni-goettingen.de/werner/schmankerl.pdf` begründet. Wir werden hier diese Argumentation im wesentlichen noch einmal wiederholen.

1. Sei $K := B[0;r]$ die abgeschlossene Kugel um 0 mit dem Radius $r > 0$und damit $K = \text{cl}\,(\Omega)$ mit $\Omega := B(0;r)$. Man definiere die Homotopie $H\colon K \times [0,1] \subset \mathbb{R}^n \times \mathbb{R} \longrightarrow \mathbb{R}^n$ durch $H(x,t) := x - tG(x)$. Angenommen, G habe keinen Fixpunkt in $K = \text{cl}\,(\Omega)$, es sei also $0 \notin (I-G)(\text{cl}\,(\Omega))$. Wir wollen den Satz 9.10 über die Homotopie-Invarianz anwenden und zeigen hierzu, dass $0 \neq H(x,t)$ für alle $(x,t) \in \partial\Omega \times [0,1]$. Für $t = 1$ ist dies nach Annahme der Fall. Für $(x,t) \in \partial\Omega \times [0,1)$ ist

$$\|H(x,t)\|_2 = \|x - tG(x)\|_2 \geq \|x\|_2 - t\,\|G(x) = r - t\,\|G(x)\|_2 \geq (1-t)r > 0,$$

also sind die Voraussetzungen von Satz 9.10 erfüllt. Wir erhalten

$$\begin{aligned} 0 &= d(I-G,\Omega,0) \\ &\qquad \text{(wegen des zweiten Teiles von Satz 9.15)} \\ &= d(H(\cdot,1),\Omega,0) \\ &= d(H(\cdot,0),\Omega,0) \\ &\qquad \text{(wegen Satz 9.10)} \\ &= d(I,\Omega,0) \\ &= 1, \end{aligned}$$

wobei wir am Schluss die Definition 9.5 des Abbildungsgrades (für stetig differenzierbare Abbildungen) benutzt haben. Damit haben wir den gewünschten Widerspruch erhalten und der Brouwersche Fixpunktsatz ist für den Fall, dass K eine (abgeschlossene) Kugel um den Nullpunkt ist, bewiesen.

2. Sei nun $K \subset \mathbb{R}^n$ nichtleer, konvex und kompakt. Man wähle $r > 0$ so groß, dass $K \subset B[0;r]$. Zu $y \in B[0;r]$ existiert $\hat{G}(y) \in K$ mit $\|\hat{G}(y) - y\|_2 = \min_{x\in K} \|x - y\|_2$, da K kompakt. Da K konvex ist, ist $\hat{G}(y)$ eindeutig bestimmt. Also ist die Abbildung $\hat{G}\colon B[0;r] \longrightarrow K$ wohldefiniert. Weiter ist $\hat{G}$ auf K die Identität. Wir zeigen, dass $\hat{G}$ stetig ist. Sei $\{x_k\} \subset B]0;r]$ eine Folge mit $\lim_{k\to\infty} x_k = x$. Es ist $\lim_{k\to\infty} \hat{G}(x_k) = \hat{G}(x)$ zu zeigen. Hierzu genügt es zu zeigen, dass jede konvergente Teilfolge $\{\hat{G}(x_{k_j})\}$ von $\{\hat{G}(x_k)\}$ den Limes $\hat{G}(x)$ besitzt. Denn angenommen, dies sei der Fall und im Widerspruch zur Behauptung gelte $\hat{G}(x_k) \not\to \hat{(}G(x)$. Dann existiert ein $\epsilon > 0$ und eine Teilfolge $\{\hat{G}(x_{k_j})\}$ mit $\|\hat{G}(x_{k_j}) - \hat{G}(x)\|_2 \geq \epsilon$ für alle j. Da $\{\hat{G}(x_{k_j})\} \subset K$ und K kompakt ist, besitzt wiederum $\{\hat{G}(x_{k_j})\}$ eine konvergente Teilfolge, die nach Annahme gegen $\hat{G}(x)$ konvergiert, was wegen $\|\hat{G}(x_{k_j}) - \hat{G}(x)\|_2 \geq \epsilon$ nicht möglich ist. Sei also $\{\hat{G}(x_{k_j})\} \subset \{\hat{G}(x_k)\}$ eine gegen ein $z \in K$ konvergente Teilfolge. Wir zeigen, dass $\|x - z\|_2 = \|x - \hat{G}(x)\|_2$, woraus $z = \hat{G}(x)$ wegen der Eindeutigkeit einer Minimallösung folgt. Angenommen, dies wäre nicht der Fall, d. h. es wäre $\|x - \hat{G}(x)\|_2 < \|x - z\|_2$. Man wähle ein $\epsilon > 0$ so klein, dass sogar noch $3\epsilon + \|x - \hat{G}(x)\|_2 < \|x - z\|_2$. Für alle

hinreichend großen j ist dann

$$\begin{aligned}\|x-z\|_2 &\leq \underbrace{\|x-x_{k_j}\|_2}_{\leq\epsilon} + \underbrace{\|x_{k_j}-\hat{G}(x_{k_j})\|_2}_{\leq\|x_{k_j}-\hat{G}(x)\|_2} + \underbrace{\|G(x_{k_j})-z\|_2}_{\leq\epsilon}\\ &\leq 2\epsilon + \|x_{k_j}-\hat{G}(x)||_2\\ &\leq 2\epsilon + \|x_{k_j}-x\|_2 + \|x-\hat{G}(x)\|_2\\ &\leq 3\epsilon + \|x-\hat{G}(x)\|_2\\ &< \|x-z\|_2,\end{aligned}$$

ein Widerspruch. Damit ist die Stetigkeit der Abbildung $\hat{G}\colon B[0;r] \longrightarrow K$ gezeigt.

3. Die zusammengesetzte Abbildung $G\circ\hat{G}\colon B[0;r] \longrightarrow K \subset B]0;r]$ ist stetig und besitzt wegen des ersten Teiles des Beweises einen Fixpunkt $x^* \in K$. Da $\hat{G}$ auf K die Identität ist, ist $\hat{G}(x^*) = x^*$ und folglich

$$x^* = G\circ\hat{G}(x^*) = G(G(x^*)) = G(x^*).$$

Also ist $x^* \in K$ ein Fixpunkt von G, der Brouwersche Fixpunktsatz ist bewiesen. □

Bemerkungen: 1. Unter den Voraussetzungen des Brouwerschen Fixpunktsatzes kann natürlich nicht die Eindeutigkeit eines Fixpunktes erwartet werden, wie das Beispiel $G = I$ zeigt.

2. Die Konvexitätsbedingung an K im Brouwerschen Fixpunktsatz kann ein wenig abgeschwächt werden. Es genügt, dass K homöomorph zu einer kompakten, konvexen Menge im $\mathbb{R}^n$ ist. □

Satz 9.20 (Leray-Schauder) *Sei $\Omega \subset \mathbb{R}^n$ eine offene, beschränkte Menge mit $0 \in \Omega$, $G\colon \mathrm{cl}\,(\Omega) \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig. Sei $G(x) \neq \lambda x$ für alle $(x,\lambda) \in \partial\Omega \times (1,\infty)$. Dann besitzt G einen Fixpunkt in* $\mathrm{cl}\,(\Omega)$.

Beweis: Man definiere wieder die Homotopie $H\colon \mathrm{cl}\,(\Omega) \times [0,1] \longrightarrow \mathbb{R}^n$ durch

$$H(x,t) := x - tG(x).$$

Es sind zwei Fälle möglich. Im ersten existiert ein $x \in \partial\Omega$ mit $x = G(x)$, wir sind fertig. Im zweiten Fall ist $x \neq G(x)$ für alle $x \in \partial\Omega$. Um den Satz 9.10 über die Homotopie-Invarianz anwenden zu können, wollen wir $H(x,t) \neq 0$ für alle $(x,t) \in \partial\Omega \times [0,1]$ nachweisen. Dies ist für $t = 0$ richtig. Da nämlich $0 \in \Omega$ und Ω offen ist, ist $0 \notin \partial\Omega$ bzw. $x = H(x,0) \neq 0$ für $x \in \partial\Omega$. Weiter ist

$$H(x,t) = t(\underbrace{t^{-1}x - G(x)}_{\neq 0}) \neq 0 \quad \text{für } (x,t) \in \partial\Omega \times (0,1],$$

es ist also auch für $t \in (0,1]$ richtig. Daher ist

$$1 = d(I,\Omega,0) = d(H(\cdot,0),\Omega,0) = d(H(\cdot,1),\Omega,0) = d(I-G,\Omega,0).$$

Aus dem Satz 9.18 folgt die Behauptung. □

Als letzten Existenzsatz in diesem Unterabschnitt formulieren und beweisen wir

Satz 9.21 *Die Abbildung $G\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$ sei stetig und streng monoton, d. h. es existiere eine Konstante $c > 0$ mit*

$$(G(x_1) - G(x_2))^T (x_1 - x_2) \geq c \, \|x_1 - x_2\|_2^2 \quad \text{für alle } x_1, x_2 \in \mathbb{R}^n.$$

Dann besitzt die Gleichung $G(x) = u$ für jedes $u \in \mathbb{R}^n$ genau eine Lösung.

Beweis: Die Eindeutigkeitsaussage ist trivial. Denn ist $G(x_1) = G(x_2) = u$, so ist

$$0 = (G(x_1) - G(x_2))^T (x_1 - x_2) \geq c \, \|x_1 - x_2\|_2^2,$$

also $x_1 = x_2$. Für den Existenzbeweis definieren wir die Homotopie

$$H(x,t) := (1-t)x + tG(x)$$

und die (offene) Kugel

$$B(0;r) := \{x \in \mathbb{R}^n : \|x\|_2 < r\},$$

wobei $r > 0$ so groß gewählt wird, dass $H(x,t) \neq u$ für alle $(x,t) \in \partial B(0;r) \times [0,1]$. Mit $G_0 := \|G(0)\|_2$ und beliebiges $x \in \mathbb{R}^n$ ist

$$\|G(x)\|_2 \, \|x\|_2 \geq G(x)^T x \geq c \, \|x\|_2^2 + G(0)^T x \geq c \, \|x\|_2^2 - G_0 \, \|x\|_2.$$

Für beliebiges $(x,t) \in \partial B(0;r) \times [0,1]$ und $r > r_0 := G_0/c$, wobei wir o. B. d. A. $c \in (0,1)$ annehmen können, ist dann

$$\begin{aligned}
\|H(x,t)\|_2^2 &= (1-t)^2 \, \|x\|_2^2 + 2t(1-t)G(x)^T x + t^2 \, \|G(x)\|_2^2 \\
&= (1-t)^2 r^2 + 2t(1-t)(G(x) - G(0))^T x + 2t(1-t)G(0)^T x \\
&\qquad + t^2 \, \|G(x)\|_2^2 \\
&\geq (1-t)^2 r^2 + 2t(1-t)cr^2 - 2t(1-t)G_0 r + t^2 \, \|G(x)\|_2^2 \\
&\geq (1-t)^2 r^2 + 2t(1-t)r(cr - G_0) + t^2 (cr - G_0)^2 \\
&= [(1-t)r + t(cr - G_0)]^2 \\
&= [(1-t)r + tc(r - r_0)]^2 \\
&\geq c^2 (r - r_0)^2,
\end{aligned}$$

wobei wir am Schluss $c \in (0,1)$ ausgenutzt haben. Wählt man also $r > r_0$ so groß, dass $c(r - r_0) > \|u\|_2$, so ist $H(x,t) \neq u$ für alle $(x,t) \in \partial B(0;r) \times [0,1]$. Eine Anwendung von Satz 9.10 über die Homotopie-Invarianz liefert

$$d(I, B(0;r), u) = d(H(\cdot,0), B(0;r), u) = d(H(\cdot,1), B(0;r), u) = d(G, B(0;r), u).$$

Nach Wahl von r ist insbesondere $\|u\|_2 < r$, folglich $d(\partial B(0;r), u) = r - \|u\|_2$. Nach Definition 9.5 des Abbildungsgrades für stetig differenzierbare Abbildungen ist

$$d(I, B(0;r), u) = d_\phi(I, B(0;r), u) = \int_{B(0;r)} \phi(\|x - u\|_2) \, dx = 1$$

mit einem beliebigen $\phi \in W_\alpha^1$, wobei $\alpha \in (0, r - \|u\|_2)$. Aus Satz 9.18, dem Satz von Kronecker, folgt die Behauptung. □

9.4 Anwendungen

In diesem Unterabschnitt wollen wir auf einige Anwendungsmöglichkeiten der angegebenen Existenzsätze, speziell des Brouwerschen Fixpunktsatzes, eingehen.

Zunächst soll mit Hilfe des Brouwerschen Fixpunktsatzes ein Teil des Satzes von Perron-Frobenius (siehe Abschnitt 59 bei `http://num.math.uni-goettingen.de/werner/schmankerl.pdf`) bewiesen werden.

Satz 9.22 *Sei $A = (a_{ij})_{i,j=1,\dots,n}$ eine positive $n \times n$-Matrix, d. h. es sei $a_{ij} > 0$, $i, j = 1, \dots, n$. Dann besitzt A einen positiven Eigenwert λ mit einem zugehörigen positiven Eigenvektor x.*

Beweis: Sei $\|\cdot\|_1$ die durch $\|x\|_1 := \sum_{j=1}^n |x_j|$ definierte Betragssummennorm auf dem $\mathbb{R}^n$. Ferner definiere man

$$K := \{x \in \mathbb{R}^n : x \geq 0,\ \|x\|_1 = 1\}.$$

Dann ist $K \subset \mathbb{R}^n$ nichtleer, kompakt und konvex. Sei ferner $F\colon K \longrightarrow \mathbb{R}^n$ definiert durch

$$F(x) := \frac{Ax}{\|Ax\|_1}.$$

Dann ist F auf K stetig und wegen $A > 0$ ist offenbar $F(K) \subset K$. Aus dem Brouwerschen Fixpunktsatz folgt die Existenz von $x \in K$ mit $F(x) = x$. Dann ist $Ax = \|Ax\|_1\, x$. Aus $x \geq 0$ und $A > 0$ folgt $Ax > 0$ und damit $x > 0$. Der Satz ist bewiesen. □

K. P. Hadeler (1971) hat mit Hilfe des Brouwerschen Abbildungsgrades die Existenz einer Lösung einer inversen Eigenwertaufgabe gezeigt. Das Problem ist das folgende:

- *Gegeben* seien eine reelle, symmetrische $n \times n$-Matrix $A = (a_{ij})$ und reelle Zahlen $s_1 > \cdots > s_n$. *Gesucht* ist eine reelle $n \times n$-Diagonalmatrix $V = (v_i \delta_{ij})$ derart, dass $A + V$ die Eigenwerte $s_1, \dots, s_n$ besitzt. O. B. d. A. kann angenommen werden, dass die Diagonalelemente von A gleich Null sind.

Unter etwas schärferen Voraussetzungen als bei Hadeler soll auf etwas einfachere Weise die Existenz einer Lösung des obigen inversen Eigenwertproblems gezeigt werden.

Satz 9.23 *Gegeben seien eine reelle, symmetrische $n \times n$-Matrix $A = (a_{ij})$ mit verschwindenden Diagonalelementen und reelle Zahlen $s_1 > \cdots > s_n$. Sei*

$$g_i := \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| > 0, \quad i = 1, \dots, n$$

und

$$s_i - s_{i+1} > 2(g_i + g_{i+1}), \quad i = 1, \dots, n-1.$$

Dann existiert eine Diagonalmatrix $V = (v_i \delta_{ij})$ derart, dass $A + V$ die Eigenwerte $s_1, \dots, s_n$ besitzt und

$$|v_i - s_i| \leq g_i, \quad i = 1, \dots, n,$$

gil[26].

[26] Von K. P. Hadeler (1971, Satz 1) wird die schwächere Voraussetzung $s_i - s_{i+1} > 2\max(g_i, g_{i+1})$, $i = 1, \dots, n-1$, gemacht. Siehe auch H. Jeggle (1979, S. 115 ff) und K. Deimling (1985, S. 20).

Beweis: Man definiere die Menge

$$\Omega := \{v \in \mathbb{R}^n : |v_i - s_i| < g_i,\ i = 1, \ldots, n\}$$

und eine Abbildung $F\colon \mathrm{cl}\,(\Omega) \times [0,1] \longrightarrow \mathbb{R}^n$ durch

$$F(v,t) := (1-t)v + t\,\mathrm{EW}(A+V).$$

Hierbei bedeutet

$$\mathrm{EW}(A+V) = (\lambda_1(v), \ldots, \lambda_n(v))^T$$

den durch $\lambda_1(v) \geq \cdots \geq \lambda_n(v)$ geordneten Vektor der Eigenwerte von

$$A + V = (a_{ij} + v_i\delta_{ij}).$$

Offenbar ist F stetig. Das Problem ist gelöst, wenn wir zeigen können, dass ein $v \in \mathrm{cl}\,(\Omega)$ mit $F(v,1) = s$ existiert, wobei $s := (s_1, \ldots, s_n)^T$. Es ist $F(v,0) = v \neq s$ für alle $v \in \partial\Omega$. Wir unterscheiden zwei Fälle. Im ersten Fall existiert ein $v \in \partial\Omega$ mit $F(v,1) = s$, dann sind wir fertig, da die Aussage in diesem Fall richtig ist. Im zweiten Fall ist $F(v,1) \neq s$ für alle $v \in \partial\Omega$. Wir wollen zeigen, dass dann auch $F(v,t) \neq s$ für alle $(v,t) \in \partial\Omega \times (0,1)$. Dies zeigen wir durch Widerspruch und nehmen an, es sei $s = F(v,t)$ mit einem Paar $(v,t) \in \partial\Omega \times (0,1)$. Sei $i \in \{1, \ldots, n-1\}$ beliebig. Wegen $v \in \partial\Omega \subset \mathrm{cl}\,(\Omega)$ ist

$$-g_i \leq s_i - v_i \leq g_i, \quad -g_{i+1} \leq v_{i+1} - s_{i+1} \leq g_{i+1}.$$

Durch Addition dieser beiden Ungleichungen erhalten wir

$$-(g_i + g_{i+1}) \leq (s_i - s_{i+1}) - (v_i - v_{i+1}) \leq g_i + g_{i+1}$$

und damit

$$v_i - v_{i+1} \geq (s_i - s_{i+1}) - (g_i + g_{i+1}) > g_i + g_{i+1} > 0.$$

Speziell ist $v_1 > \cdots > v_n$. Wegen $s = F(v,t)$ ist

$$s_i = (1-t)v_i + t\lambda_i(v), \quad i = 1, \ldots, n,$$

und damit

$$\frac{1}{t}|s_i - v_i| = |\lambda_i(v) - v_i|, \quad i = 1, \ldots, n.$$

Wir wollen zeigen, dass $|\lambda_i(v) - v_i| \leq g_i$, $i = 1, \ldots, n$. Ist dies gelungen, so erhalten wir leicht den gewünschten Widerspruch. Denn wegen $v \in \partial\Omega$ existiert ein $i_0 \in \{1, \ldots, n\}$ mit $|v_{i_0} - s_{i_0}| = g_{i_0}$. Damit erhalten wir

$$\frac{1}{t}g_{i_0} = |\lambda_{i_0}(v) - v_{i_0}| \leq g_{i_0},$$

was wegen $t \in (0,1)$ (und $g_{i_0} > 0$) ein Widerspruch ist. Um die gewünschte Abschätzung $|\lambda_i(v) - v_i| \leq g_i$, $i = 1, \ldots, n$, zu erhalten, benötigen wir eine Aussage über *Gerschgorin-Kreise*, die wir nicht beweisen wollen (siehe z. B. J. Werner (1992, S. 3)).

- *Sei $B = (b_{ij})$ eine komplexe $n \times n$-Matrix. Man definiere*

$$g_i := \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| > 0, \quad i = 1, \ldots, n,$$

und die Kreise

$$G_i := \{\lambda \in \mathbb{C} : |\lambda - b_{ii}| \leq g_i\}, \quad i = 1, \ldots, n.$$

Dann gilt:

1. *$\bigcup_{i=1}^{n}$ enthält alle Eigenwerte von B.*
2. *Sind die Kreise $G_1, \ldots, G_n$ paarweise disjunkt, so enthält jedes G_i, $i = 1, \ldots, n$, genau einen Eigenwert von B.*

Bei der Anwendung ist das reelle Intervall (der Gerschgorin-"Kreis" zu einer symmetrischen Matrix ist ein Intervall!)

$$G_i(v) := \{\lambda \in \mathbb{R} : |\lambda - v_i| \leq g_i\}, \quad i = 1, \ldots, n,$$

der i-te Gerschgorin-"Kreis" zu der symmetrischen Matrix $A + V$. Für $v \in \mathrm{cl}\,(\Omega)$ sind die Intervalle $G_1(v), \ldots, G_n(v)$ paarweise disjunkt, wie sofort aus der Voraussetzung $s_i - s_{i+1} > 2(g_i + g_{i+1})$ bzw. $v_{i+1} + g_{i+1} < v_i - g_i$, $i = 1, \ldots, n-1$, folgt. Aus der obigen Aussage über Gerschgorin-Kreise folgt $|\lambda_i(v) - v_i| \leq g_i$, $i = 1, \ldots, n$. Wie wir gesehen haben, ist damit $s \neq F(v,t)$ für alle $(v,t) \in \partial\Omega \times [0,1]$. Aus Satz 9.10 folgt, dass $d(F(\cdot,t), \Omega, s)$ auf $[0,1]$ konstant ist, also

$$d(F(\cdot,0), \Omega, s) = d(I, \Omega, s) = d(F(\cdot,1), \Omega, s)$$

gilt. Wegen Satz 9.2 ist $d(I, \Omega, s) = 1$. Aus Satz 9.18, dem Satz von Kronecker, folgt die Behauptung. □

Beispiel: Das folgende Beispiel soll mehr prinzipieller Natur sein und die Idee skizzieren, wie mit Hilfe des Brouwerschen Fixpunktsatzes die Existenz periodischer Lösungen von Differentialgleichungssystemen nachgewiesen werden kann.

Gegeben sei das Differentialgleichungssystem $\dot{x} = f(x,t)$, wobei

$$f(x,t) = \begin{pmatrix} f_1(x_1, \ldots, x_n, t) \\ \vdots \\ f_n(x_1, \ldots, x_n, t) \end{pmatrix}, \quad f\colon \mathbb{R}^n \times \mathbb{R} \longrightarrow \mathbb{R}^n.$$

Die Abbildung f habe in der letzten Variablen t eine Periode $\omega > 0$, es sei also $f(x,t) = f(x, t+\omega)$ für alle $(x,t) \in \mathbb{R}^n \times \mathbb{R}$. Die Frage ist: Besitzt das Differentialgleichungssystem $\dot{x} = f(x,t)$ eine ω-periodische Lösung?

Der Einfachheit halber (dies lässt sich wesentlich abschwächen) machen wir die folgende Voraussetzung: Die Anfangswertaufgabe

$$\dot{x} = f(x,t), \quad x(0) = x_0$$

besitzt für jedes $x_0 \in \mathbb{R}^n$ eine eindeutige Lösung $x(\cdot; x_0)$ auf $[0, \infty)$, diese hänge stetig von dem Anfangswert x_0 ab.

Man definiere die Abbildung $T\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$ durch $T(x_0) := x(\omega; x_0)$. Angenommen, x_0 ist ein Fixpunkt von T. Dann ist $x(t) := x(t; x_0)$ eine ω-periodische Lösung von $\dot{x} = f(x, t)$. Denn sei $y(t) := x(t + \omega)$. Dann gilt

$$\dot{y}(t) = \dot{x}(t + \omega) = f(x(t + \omega), t + \omega) = f(x(t + \omega), t) = f(y(t), t)$$

und

$$y(0) = x(\omega) = x(\omega; x_0) = T(x_0) = x_0.$$

Wegen der vorausgesetzten eindeutigen Lösbarkeit von Anfangswertaufgaben ist $x(t) = y(t) = x(t+\omega)$. Da T nach Voraussetzung stetig ist, folgt die Existenz eines Fixpunktes aus dem Brouwerschen Fixpunktsatz, falls eine nichtleere, konvexe, kompakte Menge $K \subset \mathbb{R}^n$ existiert, die durch T in sich abgebildet wird. □

Als weitere Anwendung beweisen wir, dass eine gewisse Wachstumseigenschaft einer stetigen Abbildung $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$ ihre Surjektivität impliziert, siehe K. Deimling (1985, S. 19).

Satz 9.24 *Sei $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig und $F(x)^T x/\|x\|_2 \to \infty$ für $\|x\|_2 \to \infty$. Dann ist $F(\mathbb{R}^n) = \mathbb{R}^n$ bzw. F surjektiv.*

Beweis: Sei $y \in \mathbb{R}^n$ gegeben. Wir definieren $H\colon \mathbb{R}^n \times [0, 1] \longrightarrow \mathbb{R}^n$ durch

$$H(x, t) := (1 - t)x + tF(x) - y$$

und mit $r > 0$ die offene Kugel

$$B(0; r) := \{x \in \mathbb{R}^n : \|x\|_2 < r\}.$$

Wir werden wir zeigen, dass wir $r > 0$ so groß wählen können, dass $H(x, t) \neq 0$ für alle $(x, t) \in \partial B(0; r) \times [0, 1]$. Für $(x, t) \in \partial B(0; r) \times [0, 1]$ ist

$$\begin{aligned} H(x,t)^T x &= (1-t)r^2 + tF(x)^T x - y^T x \\ &\geq (1-t)r^2 + tF(x)^T x - \|y\|_2 r \\ p &= r[(1-t)(r - \|y\|_2) + t(F(x)^T x/\|x\|_2 - \|y\|_2)] \\ &> 0, \end{aligned}$$

und daher auch $H(x, t) \neq 0$, falls wir $r > \|y\|_2$ so groß wählen, dass $F(x)^T x/\|x\|_2 > \|y\|_2$. Wegen Satz 9.10, dem Satz über die Homotopie-Invarianz, ist $d(H(\cdot, t), B(0; r), 0)$ konstant für $t \in [0, 1]$. Daher ist

$$\begin{aligned} 1 &= d(I - y, B(0; r), 0) \\ &= d(H(\cdot, 0), B(0; r), 0) \\ &= d(H(\cdot, 1), B(0; r), 0) \\ &= d(F - y, B(0; r), 0), \end{aligned}$$

wobei wir für die erste Gleichung Satz 9.2 zusammen mit Definition 9.5 benutzt haben. Wegen Satz 9.18 besitzt die Gleichung $F(x) = y$ eine Lösung in $B(0; r)$, womit die Behauptung bewiesen ist. □

Als weitere Anwendung geben wir an (siehe z. B. K. Deimling (1985, S. 20)):

Satz 9.25 (Igelsatz) *Sei $\Omega \subset \mathbb{R}^n$ offen, beschränkt mit $0 \in \Omega$ und $F\colon \partial\Omega \longrightarrow \mathbb{R}^n \setminus \{0\}$ stetig. Sei ferner n ungerade. Dann existiert ein $x_0 \in \partial\Omega$ und ein $\lambda \neq 0$ mit $F(x_0) = \lambda x_0$.*

Beweis: Wegen des Fortsetzungssatzes von Tietze[27] können wir o. B. d. A. annehmen, dass F sogar auf cl (Ω) definiert und stetig ist. Da n ungerade ist, ist $d(-I, \Omega, 0) = (-1)^n = -1$ (wegen $0 \in \Omega$ ist $0 \notin \partial\Omega$). Ist $d(F, \Omega, 0) \neq -1$ (beachte, dass $0 \notin F(\partial\Omega)$ nach Voraussetzung), so besitzt die Abbildung $H\colon \text{cl}\,(\Omega) \times [0,1] \longrightarrow \mathbb{R}^n$, definiert durch $H(x,t) := (1-t)F(x) - tx$, eine Nullstelle $(x_0, t_0) \in \partial\Omega \times (0,1)$. Denn andernfalls wäre $d(H(\cdot,t), \Omega, 0)$ wegen Satz 9.10, dem Satz von der Homotopieinvarianz, auf $[0,1]$ konstant, was wegen $d(H(\cdot,0), \Omega, 0) \neq -1$ und $d(H(\cdot,1), \Omega, 0) = -1$ nicht der Fall ist. Dann ist $F(x_0) = t_0(1-t_0)^{-1}x_0$ und die Behauptung ist richtig. Ist dagegen $d(F, \Omega, 0) = -1$, so definiere man $H\colon \text{cl}\,(\Omega) \times [0,1] \longrightarrow \mathbb{R}^n$ durch $H(x,t) := (1-t)F(x) + tx$. Genau wie im eben diskutierten ersten Fall erhalten wir die Existenz von $(x_0, t_0) \in \partial\Omega \times (0,1)$ mit $H(x_0, t_0) = 0$ bzw. $F(x_0) = -t_0(1-t_0)^{-1}x_0$. Die Aussage des Satzes ist also auch in diesem Fall richtig. □

Bemerkung: Im Igelsatz sei speziell $\Omega := B(0;1)$ die offene (euklidische) Einheitskugel im $\mathbb{R}^n$ mit dem Rand $S := \partial B(0;1)$. Ist $F\colon S \longrightarrow \mathbb{R}^n \setminus \{0\}$ stetig und n ungerade, so existiert nach dem Igelsatz ein $x_0 \in S$ mit $F(x_0)^T x_0 \neq 0$. Es gibt für ungerades n also keine stetige Abbildung $F\colon S \longrightarrow \mathbb{R}^n$ mit $F(x) \neq 0$ und $F(x)^T x = 0$ für alle $x \in S$, auf S gibt es also kein stetiges, nichtverschwindendes Tangentenvektorfeld. Oder: In ungerader Raumdimension kann ein Igel nicht stetig gekämmt werden. □

9.5 Der Satz von Borsuk

Die Gleichung $F(x) = y$ besitzt eine Lösung in Ω, falls $d(F, \Omega, y) \neq 0$ (siehe Satz 9.18). Um dies nachzuweisen kann der folgende Satz von Borsuk nützlich sein.

Eine Menge $\Omega \subset \mathbb{R}^n$ heißt *symmetrisch* (bezüglich des Ursprungs), falls $\Omega = -\Omega$ (bzw. mit $x \in \Omega$ auch $-x \in \Omega$). Ferner heißt $F\colon \Omega \longrightarrow \mathbb{R}^n$ *ungerade* auf der symmetrischen Menge Ω, falls $F(-x) = -F(x)$ für alle $x \in \Omega$.

Satz 9.26 (Borsuk) *Sei $\Omega \subset \mathbb{R}^n$ offen, beschränkt und symmetrisch mit $0 \in \Omega$. Die Abbildung $F\colon \text{cl}\,(\Omega) \longrightarrow \mathbb{R}^n$ sei stetig, ungerade auf Ω und $0 \notin F(\partial\Omega)$. Dann ist $d(F, \Omega, 0)$ ungerade.*

Beweis: Wir folgen dem sehr schönen Beweis von W. Gromes (1981). Der Beweis basiert auf der folgenden Hilfsaussage, in deren Beweis die Hauptarbeit steckt.

- Unter den Voraussetzungen des Satzes von Borsuk existiert zu jedem $\epsilon > 0$ eine stetige Abbildung $G\colon \text{cl}\,(\Omega) \longrightarrow \mathbb{R}^n$, welche auf Ω stetig differenzierbar und

[27] Der folgende Spezialfall, der aber für unsere Zwecke ausreicht, wird z. B. bei K. Deimling (1974, S. 21) bewiesen:

- *Es sei (X, d) ein metrischer Raum, $A \subset X$ abgeschlossen, $(Y, \|\cdot\|)$ ein linearer normierter Raum und $F\colon A \longrightarrow Y$ stetig. Dann existiert eine stetige Fortsetzung $\tilde{F}$ von A auf X, d. h. eine stetige Abbildung $\tilde{F}\colon X \longrightarrow Y$ mit $\tilde{F}(x) = F(x)$ für alle $x \in A$.*

ungerade ist (d. h. es ist $G(-x) = -G(x)$ für alle $x \in \Omega$), mit

$$\|F - G\| = \max_{x \in \mathrm{cl}\,(\Omega)} \|F(x) - G(x)\|_2 < \epsilon$$

und der Eigenschaft, dass $G'(x) \in \mathbb{R}^{n \times n}$ für alle $x \in \Omega$ mit $G(x) = 0$ nichtsingulär ist.

Zunächst wollen wir uns davon überzeugen, dass aus dieser Hilfsaussage der Satz von Borsuk folgt. Für hinreichend kleines $\epsilon > 0$ ist $d(F, \Omega, 0) = d(G, \Omega, 0)$, wie sofort aus der Definition und der Ganzzahligkeit des Abbildungsgrades folgt. Wegen Satz 9.2 (a) ist $\Gamma := \{x \in \Omega : G(x) = 0\}$ endlich. Da G ungerade ist, ist $0 \in \Gamma$. Insbesondere ist also $\Gamma \neq \emptyset$. Aus Satz 9.2 (b) folgt daher, dass

$$d(G, \Omega, 0) = \mathrm{sign}\,(\det G'(0)) + \sum_{x \in \Gamma \setminus \{0\}} \mathrm{sign}\,(\det G'(x)).$$

Die rechts stehende Summe ist eine gerade Zahl! Elemente aus $\Gamma \setminus \{0\}$ treten nämlich in Paaren auf, denn mit x ist auch $-x$ ein (von x verschiedenes) Element aus $\Gamma \setminus \{0\}$. Weiter ist G' gerade, und folglich die rechts stehende Summe gerade und der Abbildungsgrad $D(G, \Omega, 0)$ (und damit auch $d(F, \Omega, 0)$) ungerade. Es genügt also, die obige Hilfsaussage zu beweisen.

Zum Beweis der Hilfsaussage überlegen wir uns zunächst, dass wir o. B. d. A. annehmen können, dass F sogar auf Ω stetig differenzierbar und $F'(0)$ nichtsingulär ist. Bei vorgegebenem $\epsilon > 0$ finde man hierzu zunächst eine auf Ω stetig differenzierbare und auf cl (Ω) stetige Funktion $G_1 \colon \mathrm{cl}\,(\Omega) \longrightarrow \mathbb{R}^n$ mit

$$\|F - G_1\| = \max_{x \in \mathrm{cl}\,(\Omega)} \|F(x) - G_1(x)|_2 < \frac{\epsilon}{2}.$$

Zu G_1 bestimme man den ungeraden Teil $G_2 \colon \mathrm{cl}\,(\Omega) \longrightarrow \mathbb{R}^n$, nämlich

$$G_2(x) := \tfrac{1}{2}[G_1(x) - G_1(-x)].$$

Dann ist G_2 ungerade. Da F ungerade ist, ist

$$\begin{aligned} \|F(x) - G_2(x)\|_2 &= \|\tfrac{1}{2}[F(x) - F(-x)] - \tfrac{1}{2}[G_1(x) - G_1(-x)\|_2 \\ &= \|\tfrac{1}{2}[F(x) - G_1(x)] - \tfrac{1}{2}[F(-x) - G_1(-x)]\|_2 \\ &< \frac{\epsilon}{2} \end{aligned}$$

und folglich $\|F - G_2\| < \epsilon/2$. Sei nun $\delta \in \mathbb{R}$, $|\delta|$ hinreichend klein und δ kein Eigenwert von $G_2'(0)$.Dann ist G_δ ungerade, $G_\delta'(0)$ nichtsingulär und

$$\|F - G_\delta\| \leq \|F - G_2\| + |\delta| \max_{x \in \mathrm{cl}\,(\Omega)} \|x\|_2 < \epsilon.$$

Damit ist nachgewiesen, dass wir o. B. d. A. annehmen können, dass F sogar auf Ω stetig differenzierbar und $F'(0)$ nichtsingulär ist.

Zunächst erinnern wir an das Lemma 9.16 von Sard:

- *Sei $H\colon \Omega \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ auf der offenen Menge Ω stetig differenzierbar und*

$$C := \{x \in \Omega : H'(x) \text{ ist singulär}\}.$$

Dann hat $F(C)$ das Maß Null.

Für $k = 1, \ldots, n$ sei

$$S_k := \{x = (x_1, \ldots, x_n) \in \mathbb{R}^n : x_k = 0\}, \qquad \Omega_k := \Omega \setminus (S_1 \cap \cdots \cap S_k).$$

Offenbar ist dann

$$\Omega_k = \{x \in \Omega : x_i \neq 0 \text{ für ein } i \in \{1, \ldots, k\}\}, \qquad \Omega_n = \Omega \setminus \{0\}.$$

Jetzt definieren wir induktiv stetige, ungerade Funktionen $G_k\colon \mathrm{cl}\,(\Omega) \longrightarrow \mathbb{R}^n$, $k = 1, \ldots, n$, die auf Ω stetig differenzierbar sind, für die

$$\|F - G_k\| = \max_{x \in \mathrm{cl}\,(\Omega)} \|F(x) - G_k(x)\|_2 < \epsilon,$$

und für die $G_k'(x)$ nichtsingulär ist für alle $x \in \Omega_k$ mit $G_k(x) = 0$. Dann wird $G := G_n$ die gesuchte Abbildung sein, da wir auch noch nachweisen werden, dass $G_n'(0) = F'(0)$.
Zunächst definiere man $H_1\colon \Omega_1 = \Omega \setminus S_1 \longrightarrow \mathbb{R}^n$ durch

$$H_1(x) := \frac{1}{x_1^3} F(x).$$

Mit

$$C_1 := \{x \in \Omega_1 : H_1'(x) \text{ ist singulär}\}$$

hat die Menge $H_1(C_1)$ wegen des Lemmas von Sard das Maß Null. Insbesondere ist in *jeder* Kugel um den Nullpunkt ein Punkt enthalten, der nicht zu $H_1(C_1)$ gehört. Wir können also ein $y^{(1)} \in \mathbb{R}^n \setminus H_1(C_1)$ finden mit $\|F - G_1\| < \epsilon$, wobei

$$G_1(x) := F(x) - x_1^3 y^{(1)}.$$

Offensichtlich ist $G_1\colon \mathrm{cl}\,(\Omega) \longrightarrow \mathbb{R}^n$ stetig, ungerade und auf Ω stetig differenzierbar. Für $x \in \Omega_1$ mit $G_1(x) = 0$ ist $y^{(1)} = H_1(x)$ und damit $x \notin C_1$ bzw. $H_1'(x)$ nichtsingulär. Wegen $H_1'(x) = (1/x_1^3)G_1'(x)$ ist auch $G_1'(x)$ nichtsingulär.
Für ein $k < n$ sei eine stetige, ungerade Abbildung $G_k\colon \mathrm{cl}\,(\Omega) \longrightarrow \mathbb{R}^n$ gegeben, die auf Ω stetig differenzierbar ist, mit $\|F - G_k\| < \epsilon$ und der Eigenschaft, dass $G_k'(x)$ nichtsingulär ist für alle $x \in \Omega_k$ mit $G_k(x) = 0$. Wir definieren $H_{k+1}\colon \Omega \setminus S_{k+1} \longrightarrow \mathbb{R}^n$ durch

$$H_{k+1}(x) := \frac{1}{x_{k+1}^3} G_k(x).$$

Mit

$$C_{k+1} := \{x \in \Omega \setminus S_{k+1} : H_{k+1}'(x) \text{ ist singulär}\}$$

hat die Menge $H_{k+1}(C_{k+1})$ wegen des Lemmas von Sard das Maß Null. Wie im Induktionsanfang können wir ein $y^{(k+1)} \in \mathbb{R}^n \setminus H_{k+1}(C_{k+1})$ finden mit $\|F - G_{k+1}\| < \epsilon$, wobei

$$G_{k+1}(x) := G_k(x) - x_{k+1}^3 y^{(k+1)}.$$

Offensichtlich ist $G_{k+1}\colon \mathrm{cl}\,(\Omega) \longrightarrow \mathbb{R}^n$ stetig, ungerade und auf Ω stetig differenzierbar. Es bleibt zu zeigen, dass $G'_{k+1}(x)$ für alle $x \in \Omega_{k+1}$ mit $G_{k+1}(x) = 0$ nichtsingulär ist. sei also ein $x \in \Omega_{k+1}$ mit $G_{k+1}(x) = 0$ gegeben. Es ist $\Omega_{k+1} = (\Omega \setminus S_{k+1}) \cup (\Omega_k \cap S_{k+1})$. Daher machen wir eine Fallunterscheidung.

(a) Ist $x \in \Omega \setminus S_{k+1}$, so kann man wie beim Induktionsanfang argumentieren. Wegen $G_{k+1}(x) = 0$ ist dann nämlich $y^{(k+1)} = H_{k+1}(x)$ und damit $x \notin C_{k+1}$ bzw. $H'_{k+1}(x)$ nichtsingulär. Wegen $H'_{k+1}(x) = (1/x_{k+1}^3)G'_{k+1}(x)$ ist auch $G'_{k+1}(x)$ nichtsingulär.

(b) Ist $x \in \Omega_k \cap S_{k+1}$, also insbesondere $x_{k+1} = 0$, so ist $G_{k+1}(x) = G_k(x)$ und $G'_{k+1}(x) = G'_k(x)$. Wegen $x \in \Omega_k$ und $G_k(x) = 0$ ist nach Induktionsannahme $G'_k(x)$ nichtsingulär. Also ist auch $G'_{k+1}(x)$ nichtsingulär.

Durch $G := G_n$ ist also die gesuchte Abbildung gefunden, wenn wir noch zeigen können, dass $G'(0)$ nichtsingulär ist. Wegen $G_1(x) = F(x) - x_1^3 y^{(1)}$ ist $G'_1(0) = F'(0)$ nichtsingulär. Aus $G_{k+1}(x) = G_k(x) - x_{k+1}^3 y^{(k+1)}$, $k = 1, \ldots, n-1$, ist $G'(0) = G'_n(0) = F'(0)$. Am Anfang des Beweises der Hilfsaussage hatten wir uns überlegt, dass o. B. d. A. u.a. angenommen werden kann, dass $F'(0)$ nichtsingulär ist. Damit ist nachgewiesen, dass die konstruierte Abbildung G alle in der Hilfsaussage geforderten Eigenschaften besitzt, diese also bewiesen ist. Damit ist auch der Satz von Borsuk bewiesen. □

Bemerkung: Mit Hilfe des Satzes 9.11 (Homotopie-Invarianz) erhält man aus dem Satz von Borsuk die folgende Variante (siehe K. Deimling (1985, S. 22)):

- *Sei $\Omega \subset \mathbb{R}^n$ offen, beschränkt und symmetrisch (bezüglich des Nullpunkte) und $0 \in \Omega$. Sei $G\colon \mathrm{cl}\,(\Omega) \longrightarrow \mathbb{R}^n$ stetig, $0 \notin G(\partial\Omega)$ und $G(-x) \neq \lambda G(x)$ für alle $x \in \partial\Omega$ und alle $\lambda \geq 1$. Dann ist $d(G, \Omega, 0)$ ungerade.*

Denn: Man definiere $F\colon \mathrm{cl}\,(\Omega) \longrightarrow \mathbb{R}^n$ durch $F(x) := G(x) - G(-x)$. Da F ungerade ist, ist $d(F, \Omega, 0)$ wegen des Satzes von Borsuk ungerade. Für alle $(x, t) \in \partial\Omega \times [0, 1]$ ist $(1-t)F(x) + tG(x) \neq 0$ und daher wegen Satz 9.11 $d(G, \Omega, 0) = d(F, \Omega, 0)$. Denn wäre dies nicht der Fall, existierte also ein Paar $(x, t) \in \partial\Omega \times [0, 1]$ mit

$$\begin{aligned} 0 &= (1-t)F(x) + tG(x) \\ &= (1-t)[G(x) - G(-x)] + tG(x) \\ &= G(x) - (1-t)G(-x), \end{aligned}$$

so wäre $t \in [0, 1)$ wegen $0 \notin G(\partial\Omega)$ und daher

$$G(-x) = \underbrace{\frac{1}{1-t}}_{\geq 1} G(x),$$

ein Widerspruch zu $G(-x) \neq \lambda G(x)$ für alle $(x, \lambda) \in \partial\Omega \times [1, \infty)$. □

Wir geben zwei Anwendungen des Satzes von Borsuk an.

Satz 9.27 (Borsuk-Ulam) *Sei $\Omega \subset \mathbb{R}^n$ offen, beschränkt und symmetrisch mit $0 \in \Omega$. Ist $F\colon \partial\Omega \longrightarrow \mathbb{R}^m$ mit $m < n$ stetig, so existiert ein $x \in \partial\Omega$ mit $F(x) = F(-x)$.*

Beweis: Wir können o. B. d. A. annehmen, dass F stetig auf cl (Ω) ist, siehe den Beweis von Satz 9.25. Man definiere $\tilde{F}\colon \mathrm{cl}\,(\Omega) \longrightarrow \mathbb{R}^m \times \mathbb{R}^{n-m} = \mathbb{R}^n$ durch $\tilde{F}(x) := (F(x), 0)$. Anschließend definieren wir $G\colon \mathrm{cl}\,(\Omega) \longrightarrow \mathbb{R}^n$ durch $G(x) := \tilde{F}(x) - \tilde{F}(-x)$. Dann ist G stetig und auf Ω ungerade. Angenommen, es sei $G(x) \neq 0$ für alle $x \in \partial\Omega$ bzw. $0 \notin G(\partial\Omega)$. Wegen Satz 9.26, dem Satz von Borsuk, ist $d(G, \Omega, 0)$ ungerade und daher insbesondere $d(G, \Omega, 0) \neq 0$. Mit $r \in (0, d(0, G(\partial\Omega)))$ sei $B[0; r]$ die abgeschlossene Kugel um Null mit dem Radius r. Wegen Satz 9.13 ist $d(G, \Omega, 0) = d(G, \Omega, y)$ für jedes $y \in B[0; r]$, da $p(t) := ty \notin G(\partial\Omega)$ für alle $t \in [0, 1]$. Wegen Satz 9.18, dem Satz von Kronecker, ist $y \in G(\Omega)$ für alle $y \in B[0; r]$, also $B[0; r] \subset G(\Omega)$ bzw. der Nullpunkt des $\mathbb{R}^n$ ein innerer Punkt von $G(\Omega)$. Da die letzten $n - m$ Komponenten von Punkten aus $G(\Omega)$ aber verschwinden, hat man einen Widerspruch erreicht. Also existiert ein $x \in \partial\Omega$ mit $G(x) = 0$ bzw. mit $F(x) = F(-x)$. Der Satz ist bewiesen. □

Bemerkung: Ist $\Omega \subset \mathbb{R}^3$ die Erdkugel, $\partial\Omega$ die Erdoberfläche und ist $F\colon \partial\Omega \longrightarrow \mathbb{R}^2$ dadurch gegeben, dass jedem Punkt $P \in \partial\Omega$ ein Paar bestehend aus der in P zu einer bestimmten Zeit herrschenden Temperatur und Luftdruck zugeordnet wird. Unter der Voraussetzung, dass diese Abbildung F stetig ist, sagt der Satz von Borsuk-Ulam aus, dass ein Paar von antipodalen Punkten auf der Erdoberfläche mit gleicher Temperatur und gleichem Luftdruck existiert. □

Ist z. B. $\Omega := \{x \in \mathbb{R}^2 : \|x\|_2 < 1\}$ die offene Kreisscheibe um den Nullpunkt mit dem Radius 1 und $\partial\Omega = \{x \in \mathbb{R}^2 : \|x\|_2 = 1\}$, so benötigt man offenbar mindestens drei abgeschlossene Mengen $A_i \subset \partial\Omega$ um $\partial\Omega$ zu überdecken, wenn die A_i keine antipodalen Punkte enthalten dürfen, d. h. $A_i \cap (-A_i) = \emptyset$ gilt. Allgemein gilt

Satz 9.28 (Lusternik-Schnirelmann-Borsuk) *Sei $\Omega \subset \mathbb{R}^n$ offen, beschränkt und symmetrisch bezüglich $0 \in \Omega$. Sei $\{A_1, \ldots, A_p\}$ eine Überdeckung von $\partial\Omega$ durch abgeschlossene Mengen $A_i \subset \partial\Omega$ mit $A_i \cap (-A_i) = \emptyset$, $i = 1, \ldots, p$. Dann ist $p \geq n + 1$.*

Beweis: Im Widerspruch zur Behauptung nehmen wir an, es sei $p \leq n$. Wir definieren $F_i\colon A_i \cup (-A_i) \longrightarrow \mathbb{R}$, $i = 1, \ldots, p - 1$, durch

$$F_i(x) := \begin{cases} 1, & x \in A_i, \\ -1, & -x \in A_i, \end{cases} \qquad i = 1, \ldots, p - 1,$$

was wegen $A_i \cap (-A_i) = \emptyset$ wohldefiniert ist, und setzen diese Funktionen zu (gleichnamigen) auf cl (Ω) stetigen Funktionen fort. Weiter definieren wir $F_i\colon \mathrm{cl}\,(\Omega) \longrightarrow \mathbb{R}$ durch $F_i(x) := 1$, $i = p, \ldots, n$, und anschließend $F\colon \mathrm{cl}\,(\Omega) \longrightarrow \mathbb{R}^n$ durch

$$F(x) := \begin{pmatrix} F_1(x) \\ \vdots \\ F_n(x) \end{pmatrix}.$$

Wir werden uns überlegen, dass $F(-x) \neq \lambda F(x)$ für alle $(x, \lambda) \in \partial\Omega \times [0, \infty)$. Wegen der Bemerkung im Anschluss an den Beweis des Satzes von Borsuk folgt insbesondere

$d(F, \Omega, 0) \neq 0$, aus Satz 9.18 (Kronecker) folgt die Existenz eines $x \in \Omega$ mit $F(x) = 0$, was wegen $F_n(x) = 1$ für alle $x \in \text{cl}\,(\Omega)$ einen Widerspruch ergibt.
Da $\{A_1, \ldots, A_p\}$ eine Überdeckung von $\partial\Omega$ ist, ist $\partial\Omega \subset \bigcup_{i=1}^{p} A_i$. Ist $x \in A_p$, so ist $-x \notin A_p$ und daher $-x \in A_i$ für ein $i \in \{1, \ldots, p-1\}$. Also ist

$$\partial\Omega \subset \bigcup_{i=1}^{p-1} (A_i \cup (-A_i)).$$

Ist also $x \in \partial\Omega$, so ist $x \in A_i$ mit einem $i \in \{1, \ldots, p-1\}$ und folglich $F_i(x) = 1$, $F_i(-x) = -1$. Also ist $F(-x) \neq \lambda F(x)$ für alle $(x, \lambda) \in \partial\Omega \times [0, \infty)$. Der Satz ist bewiesen. □

10 Der Leray-Schaudersche Abbildungsgrad in linearen normierten Räumen

Ziel dieses Abschnitts ist es, den Brouwerschen Abbildungsgrad vom $\mathbb{R}^n$ auf lineare normierte Räume zu übertragen, um dadurch Existenzsätze für Gleichungen in linearen normierten Räumen zu erhalten. Als Literatur geben wir wieder die Bücher von K. Deimling (1974, 1985), M. Ružička (2004), J. T. Schwartz (1969), L. Nirenberg (1974) an.

10.1 Definition des Abbildungsgrades

Die Konstruktion des Abbildungsgrades für Abbildungen, die einen linearen normierten Raum in sich abbilden, verläuft in folgenden Schritten:

1. Der Abbildungsgrad wird für Abbildungen in einem *endlichdimensionalen* linearen normierten Raum definiert. Wichtig ist hierbei die Invarianz des Abbildungsgrades gegenüber Koordinatentransformationen.

2. Der Abbildungsgrad wird für spezielle Abbildungen auf einem linearen normierten Raum definiert, nämlich "endlichdimensionale Störungen" der Identität.

3. Auf einem linearen normierten Raum wird der Abbildungsgrad für Operatoren definiert, die sich als Grenzwert von endlichdimensionalen Störungen der Identität darstellen lassen. Es wird sich herausstellen, dass Operatoren der Form $I + G$ mit kompaktem G diese Eigenschaft besitzen.

Nach diesem Aufbau wird es ziemlich klar sei, dass die wesentlichen Eigenschaften des so konstruierten Abbildungsgrades erhalten bleiben und auch die Existenzaussagen zu Lösungen von Gleichungen (mit Vorsicht) sich übertragen lassen.

Definition 10.1 Sei E ein n-dimensionaler (reeller) linearer normierter Raum. Durch $\{v_1, \ldots, v_n\} \subset E$ sei eine Basis von E, es sei also $E = \text{span}\,\{v_1, \ldots, v_n\}$. Sei $\Omega \subset E$

offen und beschränkt, $F\colon \mathrm{cl}\,(\Omega) \subset E \longrightarrow E$ stetig sowie $y \in E$ mit $y \notin F(\partial\Omega)$. Dann ist der *Abbildungsgrad* von F auf Ω bezüglich y definiert durch

$$d(F, \Omega, y) := d(\Phi^{-1} F \Phi, \Phi^{-1}(\Omega), \Phi^{-1}(y)).$$

Hierbei ist $\Phi\colon \mathbb{R}^n \longrightarrow E$ der durch

$$\Phi(x) := \sum_{i=1}^{n} x_i v_i$$

definierte lineare Homöomorphismus.

Satz 10.2 *Die Definition 10.1 des Abbildungsgrades in einem endlichdimensionalen linearen normierten Raum ist sinnvoll und von der gewählten Basis unabhängig.*

Beweis: 1. Da $\Phi\colon \mathbb{R}^n \longrightarrow E$ ein Homöomorphismus ist, ist $\Omega_n := \Phi^{-1}(\Omega) \subset \mathbb{R}^n$ offen, $\Phi^{-1}(\partial\Omega) = \partial\Omega_n$ und $\Phi^{-1}(\mathrm{cl}\,(\Omega)) = \mathrm{cl}\,(\Omega_n)$. Die Abbildung

$$F_n := \Phi^{-1} F \Phi\colon \mathrm{cl}\,(\Omega_n) \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

ist stetig und $y_n := \Phi^{-1}(y) \notin F_n(\partial\Omega_n)$. Folglich ist der Brouwersche Abbildungsgrad $d(F_n, \Omega_n, y_n)$ definiert.
2. Sei $\{\hat{v}_1, \ldots, \hat{v}_n\}$ eine weitere Basis von E und $\hat{\Phi}\colon \mathbb{R}^n \longrightarrow E$ der durch $\hat{\Phi}(x) := \sum_{i=1}^{n} x_i \hat{v}_i$ definierte Φ entsprechende lineare Homöomorphismus. Dann ist

$$\Psi := \Phi^{-1}\hat{\Phi}\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

ein linearer Homöomorphismus, es ist also $\psi(x) = Ax$ mit einer nichtsingulären Matrix $A \in \mathbb{R}^{n\times n}$. Definiert man analog zu den obigen Bezeichnungen

$$\hat{\Omega}_n := \hat{\Phi}^{-1}(\Omega), \qquad \hat{F}_n := \hat{\Phi}^{-1} F \hat{\Phi}, \qquad \hat{y}_n := \hat{\Phi}^{-1}(y),$$

so ist zu zeigen, dass $d(F_n, \Omega_n, y_n) = d(\hat{F}_n, \hat{\Omega}_n, \hat{y}_n)$. Nun ist

$$\hat{F}_n = \hat{\Phi}^{-1} F \hat{\Phi} = \Psi^{-1}\Phi^{-1} F \Phi\Psi = \Psi^{-1} F_n \Psi$$

und

$$\hat{\Omega}_n = \hat{\Phi}^{-1}(\Omega) = \Psi^{-1}\Phi^{-1}(\Omega) = \Psi^{-1}(\Omega_n), \qquad \hat{y}_n = \Psi^{-1}(y_n).$$

Also ist zu zeigen, dass

$$d(F_n, \Omega_n, y_n) = d(\Psi^{-1} F_n \Psi, \psi^{-1}(\Omega_n), \Psi^{-1}(y_n)).$$

Es kann o. B. d. A. angenommen werden (siehe den Beweis von Satz 9.17, in dem die Ganzzahligkeit des Abbildungsgrades nachgewiesen wird), dass:

1. F_n ist stetig differenzierbar auf einer $\mathrm{cl}\,(\Omega_n)$ umfassenden offenen Menge D_n,
2. $F_n'(x)$ ist nichtsingulär für alle $x \in \Gamma_n$, wobei

$$\Gamma_n := \{x \in \Omega_n : F_n(x) = y_n\}.$$

Nach Satz 9.2 ist dann

$$d(F_n, \Omega_n, y_n) = \sum_{x \in \Gamma_n} \text{sign det } F_n'(x).$$

Da $\hat{F}_n$ den entsprechenden Voraussetzungen genügt, ist

$$d(\hat{F}_n, \hat{\Omega}_n, \hat{y}_n) = \sum_{z \in \hat{\Gamma}_n} \text{sign det } \hat{F}_n'(z)$$

mit

$$\hat{\Gamma}_n := \{z \in \hat{\Omega}_n : \hat{F}_n(z) = \hat{y}_n\}.$$

Wegen $\hat{F}_n = \Psi^{-1} F_n \Psi$ bzw. $\hat{F}_n(z) = A^{-1} F_n(Az)$ ist $\hat{F}_n'(z) = A^{-1} F_n'(Az) A$ und folglich $\det \hat{F}_n'(z) = \det F_n'(Az)$. Da außerdem $z \in \hat{\Gamma}_n$ genau dann, wenn $\psi(z) = Az \in \Gamma_n$ ist $d(\hat{F}_n, \hat{\Omega}_n, \hat{y}_n) = d(F_n, \Omega_n, y_n)$ bewiesen. □

Nach der Konstruktion des Abbildungsgrades in linearen normierten Räumen ist klar, dass sich alle Aussagen über den Brouwerschen Abbildungsgrad auch für den erweiterten Begriff gelten.

Unser nächstes Ziel ist, den Abbildungsgrad auf "endlichdimensionale Störungen der Identität" zu übertragen.

Definition 10.3 Sei X ein linearer normierter Raum, $D \subset X$ eine Teilmenge. Eine stetige Abbildung $G\colon D \longrightarrow X$ heißt *endlichdimensional*, falls ein endlichdimensionaler linearer Teilraum E von X mit $G(D) \subset E$ existiert. Ist $G\colon D \longrightarrow X$ endlichdimensional, so heißt $F := I + G$ eine *endlichdimensionale Störung der Identität*.

In der folgenden Definition wird der Begriff des Abbildungsgrades auf endlichdimensionale Störungen der Identität übertragen.

Definition 10.4 Sei X ein linearer normierter Raum und $\Omega \subset X$ eine offene, beschränkte Teilmenge. Die Abbildung $G\colon \text{cl}\,(\Omega) \subset X \longrightarrow X$ sei stetig und endlichdimensional, $F := I + G$ und $y \notin F(\partial\Omega)$. Sei E ein endlichdimensionaler linearer Teilraum von X mit $y \in E$, $\Omega \cap E \neq \emptyset$ und $G(\text{cl}\,(\Omega)) \subset E$. Sei

$$F_{\text{cl}\,(\Omega(\cap E} := \text{Rest}|_{\text{cl}\,(\Omega)\cap E} F\colon \text{cl}\,(\Omega) \cap E \subset E \longrightarrow E$$

die Restriktion von F auf $\text{cl}\,(\Omega) \cap E$. Dann ist der *Abbildungsgrad der endlichdimensionalen Störung der Identität F auf Ω bezüglich y* durch

$$d(F, \Omega, y) = d(F_{\text{cl}\,(\Omega)\cap E}, \Omega \cap E, y)$$

definiert.

Satz 10.5 *Die Definition 10.4 des Abbildungsgrades einer endlichdimensionalen Störung der Identität ist sinnvoll und von dem gewählten endlichdimensionalen linearen Teilraum unabhängig.*

Beweis: 1. Die Menge $\Omega \cap E$ ist offen in dem endlichdimensionalen Teilraum E und es ist $\mathrm{cl}\,(\Omega \cap E) = \mathrm{cl}\,(\Omega) \cap E$. Die Abbildung $F_{\mathrm{cl}\,(\Omega)\cap E}\colon \mathrm{cl}\,(\Omega) \cap E \longrightarrow E$ ist stetig und $\partial(\Omega \cap E) \subset \partial\Omega$, sodass $F_{\mathrm{cl}\,(\Omega)\cap E}(\partial(\Omega \cap E)) \subset F(\partial\Omega)$ und $y \notin F_{\mathrm{cl}\,(\Omega)\cap E}(\partial(\mathrm{cl}\,(\Omega) \cap E))$. Nach Definition 10.1 bzw. Satz 10.2 ist also $d(F_{\mathrm{cl}\,(\Omega)\cap E}, \Omega \cap E, y)$ definiert.

2. Sei $\hat{E}$ ein weiterer endlichdimensionaler linearer Teilraum von X mit $y \in \hat{E}$, $\Omega \cap \hat{E} \neq \emptyset$ und $G(\mathrm{cl}\,(\Omega)) \subset \hat{E}$. Wir werden zeigen, dass

$$d(F_{\mathrm{cl}\,(\Omega)\cap E}, \Omega \cap E, y) = d(F_{\mathrm{cl}\,(\Omega)\cap \hat{E}}, \Omega \cap \hat{E}, y).$$

O. B. d. A. ist $E \subset \hat{E}$. Denn ist dies nicht der Fall, so bilde man den E und $\hat{E}$ umfassenden endlichdimensionalen linearen Teilraum $\tilde{E} := \mathrm{span}\,\{E, \hat{E}\}$. Wir können also annehmen, dass $E \subset \hat{E}$. Daher ist $\dim(E) = m \leq n = \dim(\hat{E})$. Da der Abbildungsgrad in einem n-dimensionalen linearen normierten Raum durch Abbildungsgrad im $\mathbb{R}^n$ gegeben ist, genügt es, den Fall $E = \mathbb{R}^m$, $\hat{E} = \mathbb{R}^n$ zu betrachten, wobei die Inklusion $\mathbb{R}^m \subset \mathbb{R}^n$ bedeutet, dass

$$\mathbb{R}^m \equiv \{x = (x_i)_{1\leq i\leq n} \in \mathbb{R}^n : x_{m+1} = \cdots = x_n = 0\},$$

der $\mathbb{R}^m$ also mit dem rechts stehenden linearen Teilraum des $\mathbb{R}^n$ identifiziert wird. Daher genügt es, die folgende Behauptung zu beweisen:

- Sei $\Omega \subset \mathbb{R}^n$ offen und beschränkt, $\mathbb{R}^m \subset \mathbb{R}^n$ (im obigen Sinne), $\Omega \cap \mathbb{R}^m \neq \emptyset$ und $G\colon \mathrm{cl}\,(\Omega) \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$ stetig. Die Abbildung $F\colon \mathrm{cl}\,(\Omega) \longrightarrow \mathbb{R}^n$ sei durch $F := I + G$ definiert. Für jedes $y \in \mathbb{R}^m$ mit $y \notin F(\partial\Omega)$ ist dann

$$f(F, \Omega, y) = d(F_{\mathrm{cl}\,(\Omega)\cap\mathbb{R}^m}, \Omega \cap \mathbb{R}^m, y).$$

Denn: Es ist offenbar $F(\mathrm{cl}\,(\Omega) \cap \mathbb{R}^m) \subset \mathbb{R}^m$, sodass der rechts stehende abbildungsgrad einen Sinn hat. Wie beim Beweis von Satz 10.2 kann wiederum angenommen werden, dass G auf einer $\mathrm{cl}\,(\Omega)$ umfassenden offenen Menge D stetig differenzierbar ist und $F'(x)$ für alle $x \in \Gamma$ nichtsingulär ist, wobei

$$\Gamma := \{x \in \Omega : F(x) = y\}.$$

Ist $x \in \Gamma$, so ist $x + G(x) = y$, also $x = y - G(x) \in \mathbb{R}^m$ und daher

$$\Gamma = \{x \in \Omega : F(x) = y\} = \{x \in \Omega \cap \mathbb{R}^m : F_{\mathrm{cl}\,(\Omega)\cap\mathbb{R}^m}(x) = y\}.$$

Für $x \in \Omega \cap \mathbb{R}^m$ ist

$$F(x_1, \ldots, x_n) = \begin{pmatrix} x_1 + G_1(x_1, \ldots, x_m, 0, \ldots, 0) \\ \vdots \\ x_m + G_m(x_1, \ldots, x_m, 0, \ldots, 0) \\ x_{m+1} \\ \vdots \\ x_n \end{pmatrix}$$

während

$$F_{\mathrm{cl}\,(\Omega)\cap\mathbb{R}^m}(x_1,\ldots,x_m) = \begin{pmatrix} x_1 + G_1(x_1,\ldots,x_m,0,\ldots,0) \\ \vdots \\ x_m + G_m(x_1,\ldots,x_m,0,\ldots,0) \end{pmatrix}.$$

Für $x \in \Omega \cap \mathbb{R}^m$ ist daher

$$F'(x) = \left(\begin{array}{c|c} I + \left(\dfrac{\partial G_i}{\partial x_j}(x)\right)_{1\le i,j\le m} & 0 \\ \hline 0 & I \end{array}\right), \qquad F'_{\mathrm{cl}\,(\Omega)\cap\mathbb{R}^m}(x) = I + \left(\frac{\partial G_i}{\partial x_j}(x)\right)_{1\le i,j\le m}.$$

Folglich ist $\det F'(x) = \det F'_{\mathrm{cl}\,(\Omega)\cap\mathbb{R}^m}(x)$ für alle $x \in \Omega \cap \mathbb{R}^m$, speziell für alle $x \in \Gamma$. Hieraus folgt die Behauptung und der Satz ist bewiesen. □

Bemerkung: Der Abbildungsgrad einer endlichdimensionalen Störung der Identität in einem linearen normierten Raum wird auf den "endlichdimensionalen Abbildungsgrad" zurückgeführt. Damit übertragen sich alle Eigenschaften des Brouwerschen Abbildungsgrades. □

Nun erweitert man den Abbildungsgrad auf diejenigen Abbildungen, die *Limes* von endlichdimensionalen Störungen der Identität sind. Es wird sich zeigen, dass diese Abbildungen sich gerade als die *kompakten Störungen* der Identität ergeben.

Definition 10.6 Seien X, Y lineare normierte Räume und $D \subset X$. Eine Abbildung $G\colon D \subset X \longrightarrow Y$ heißt *kompakt*, wenn sie stetig ist und beschränkte Teilmengen von D in relativ kompakte Teilmengen von Y abbildet.

Bemerkung: Eine relativ kompakte Teilmenge eines linearen normierten Raumes ist insbesondere beschränkt. Eine lineare Abbildung, die beschränkte Mengen in beschränkte Mengen abbildet, ist stetig. Für lineare Abbildungen ist die Stetigkeitsvoraussetzung in der obigen Kompaktheitsdefinition also überflüssig, da sie automatisch erfüllt ist. □

Satz 10.7 *Sei X ein linearer normierter Raum, $S \subset X$ beschränkt und $G\colon S \subset X \longrightarrow X$ eine kompakte Abbildung. Dann gibt es zu jedem $\epsilon > 0$ eine endlichdimensionale kompakte Abbildung $G_\epsilon\colon S \subset X \longrightarrow X_\epsilon \subset X$ mit $\|G_\epsilon(x) - G(x)\| \le \epsilon$ für alle $x \in S$.*

Beweis: Da S beschränkt und G kompakt ist, ist $G(S)$ relativ kompakt und daher cl $(G(S))$ kompakt. Trivialerweise ist cl $(G(S)) \subset \bigcup_{y\in \mathrm{cl}\,(G(S))} B(y;\epsilon)$. Hierbei bedeutet $B(y;\epsilon)$ die offene Kugel um y mit dem Radius ϵ. Aus dieser Überdeckung kann eine endliche Teilüberdeckung ausgewählt werden. Daher existieren $\{y_1,\ldots,y_k\} \subset X$ mit cl $(G(S)) \subset \bigcup_{i=1}^k B(y_i;\epsilon)$. Man definiere den endlichdimensionalen linearen Teilraum $X_\epsilon := \mathrm{span}\,\{y_1,\ldots,y_k\}$ und die stetige Abbildung $\mu_i\colon S \longrightarrow [0,\epsilon]$ durch

$$\mu_i(x) := \max(0, \epsilon - \|G(x) - y_i\|), \quad i = 1,\ldots,k.$$

Ferner existiert zu jedem $x \in S$ ein $i \in \{1,\ldots,k\}$ mit $G(x) \in B(y_i;\epsilon)$ bzw. $\mu_i(x) > 0$. Also wird durch

$$\lambda_i(x) := \frac{\mu_i(x)}{\sum_{j=1}^k \mu_j(x)}, \quad i = 1,\ldots,k,$$

eine stetige Abbildung $\lambda_i: S \longrightarrow [0,1]$ definiert und es ist $\sum_{i=1}^{k} \lambda_i(x) = 1$ für alle $x \in S$. Die gesuchte endlichdimensionale Approximation $G_\epsilon: S \longrightarrow X_\epsilon \subset X$ von G wird nun definiert durch

$$G_\epsilon(x) := \sum_{i=1}^{k} \lambda_i(x) y_i.$$

Dann gilt:

1. Es ist $G_\epsilon(S) \subset X_\epsilon$ und $X_\epsilon \subset X$ ist ein endlichdimensionaler linearer Teilraum, die Abbildung G_ϵ ist also endlichdimensional.

2. Die Abbildung $G_\epsilon: S \longrightarrow X_\epsilon$ ist stetig, da die λ_i, $i = 1, \ldots, k$, stetig sind.

3. Es ist $\|G_\epsilon(x) - G(x)\| \leq \epsilon$ für alle $x \in S$.

   Denn: Für $x \in S$ ist

$$\begin{aligned} \|G_\epsilon(x) - G(x)\| &= \Big\| \sum_{i=1}^{k} \lambda_i(x) y_i - \sum_{i=1}^{k} \lambda_i(x) G(x) \Big\| \\ &\leq \sum_{i=1}^{k} \lambda_i(x) \|y_i - G(x)\| \\ &= \sum_{i: \|y_i - G(x)\| < \epsilon} \lambda_i(x) \|y_i - G(x)\| \\ &\leq \epsilon. \end{aligned}$$

4. Die Abbildung $G_\epsilon: S \longrightarrow X_\epsilon$ ist kompakt.

   Denn: Hierzu braucht nur gezeigt zu werden, dass $G_\epsilon(S)$ beschränkt ist. Denn eine beschränkte Menge in einem endlichdimensionalen linearen normierten Raum ist relativ kompakt. Da $S \subset X$ beschränkt und $G: S \longrightarrow X$ eine kompakte Abbildung ist, ist $G(S)$ relativ kompakt und damit beschränkt, sodass eine Konstante $C_1 > 0$ mit $\|G(x)\| \leq C_1$ für alle $x \in S$ existiert. Für ein beliebiges $x \in S$ ist daher

$$\|G_\epsilon(x)\| \leq \|G(x)\| + \|G_\epsilon(x) - G(x)\| \leq C_1 + \epsilon.$$

Damit ist der Satz bewiesen. □

Hiermit bietet sich die folgende Definition an:

Definition 10.8 Sei X ein linearer normierter Raum, $\Omega \subset X$ offen und beschränkt, $G: \mathrm{cl}\,(\Omega) \subset X \longrightarrow X$ eine kompakte Abbildung, $F := I + G: \mathrm{cl}\,(\Omega) \longrightarrow X$ und $y \notin F(\partial\Omega)$. Dann wird der Abbildungsgrad von F auf Ω bezüglich y durch

$$d(F, \Omega, y) := \lim_{k \to \infty} d(F_k, \Omega, y)$$

definiert. Hierbei ist $F_k := I + G_k$ und $G_k: \mathrm{cl}\,(\Omega) \longrightarrow X$ endlichdimensional mit $\lim_{k\to\infty} \sup_{x \in \mathrm{cl}\,(\Omega)} \|G_k(x) - G(x)\| = 0$.

Der folgende Satz klärt die Wohldefiniertheit des Abbildungsgrades für kompakte Störungen der Identität.

Satz 10.9 *Definition 10.8 ist sinnvoll, d. h. unter den gemachten Voraussetzungen existiert* $\lim_{k\to\infty} d(F_k, \Omega, y)$. *Ferner ist dieser Limes von der Wahl der Folge* $\{F_k\}$ *unabhängig. Ist also* $\hat{F}_k := I + \hat{G}_k$ *mit endlichdimenionalem* $\hat{G}_k\colon \mathrm{cl}\,(\Omega) \longrightarrow X$ *und*

$$\lim_{k\to\infty} \sup_{x\in \mathrm{cl}\,(\Omega)} \|\hat{G}_k(x) - G(x)\| = 0,$$

so ist

$$\lim_{k\to\infty} d(F_k, \Omega, y) = \lim_{k\to\infty} d(\hat{F}_k, \Omega, y).$$

Beweis: 1. Zunächst zeigen wir, dass ein $\alpha > 0$ mit $\|F(x) - y\| \geq \alpha$ für alle $x \in \partial\Omega$ existiert. Dies wird durch Widerspruch gezeigt. Angenommen, es existiert eine Folge $\{x_j\} \subset \partial\Omega$ mit $\lim_{j\to\infty} \|F(x_j) - y\| = 0$. Da mit Ω auch $\partial\Omega$ beschränkt ist, folgt aus der Kompaktheit von G auf $\mathrm{cl}\,(\Omega)$ die Existenz einer Teilfolge $\{x_{j_k}\}$ mit der Eigenschaft, dass $\{G(x_{j_k})\}$ konvergent ist, etwa gegen ein z. Aus

$$y = \lim_{k\to\infty} F(x_{j_k}) = \lim_{k\to\infty} [x_{j_k} + G(x_{j_k})]$$

folgt dann auch die Konvergenz der Folge $\{x_{j_k}\}$, und zwar ist

$$w := \lim_{k\to\infty} x_{j_k} = z - y.$$

Da $\{x_{j_k}\}$ in der abgeschlossenen Menge $\partial\Omega$ enthalten ist, ist auch $w \in \partial\Omega$. Da G als kompakte Abbildung stetig ist, ist schließlich $y = w + G(w) = F(w)$, ein Widerspruch zu $y \notin F(\partial\Omega)$.

2. Nach Satz 10.7 existiert zu $G\colon \mathrm{cl}\,(\Omega) \longrightarrow X$ eine endlichdimensionaler Abbildung $G_\alpha\colon \mathrm{cl}\,(\Omega) \longrightarrow X_\alpha \subset X$ mit $\|G_\alpha(x) - G(x)\| \leq \alpha/2$ für alle $x \in \mathrm{cl}\,(\Omega)$. Durch Adjunktion von y und notfalls eines Punktes von Ω kann X_α so vergrößert werden, dass $y \in X_\alpha$, $\Omega \cap X_\alpha \neq \emptyset$ und $X_\alpha \subset X$ ein endlichdimensionaler linearer Teilraum ist. Mit $F_\alpha := I + G_\alpha$ ist dann der Abbildungsgrad $d(F_\alpha, \Omega, y)$ nach Definition 10.4 bzw. Satz 10.5 wohldefiniert, falls $y \notin F_\alpha(\partial\Omega)$. Dies ist aber der Fall, denn für $x \in \partial\Omega$ ist

$$\begin{aligned}
\|F_\alpha(x) - y\| &= \|x + G_\alpha(x) - y\| \\
&= \|F(x) - y + G_\alpha(x) - G(x)\| \\
&\geq \|F(x) - y\| - \|G_\alpha(x) - G(x)\| \\
&\geq \alpha - \frac{\alpha}{2} \\
&= \frac{\alpha}{2} \\
&> 0.
\end{aligned}$$

Der Satz ist bewiesen, wenn wir zeigen können:

- Ist $G_\beta\colon \mathrm{cl}\,(\Omega) \longrightarrow X_\beta \subset X$ endlichdimensional und gilt $\|G_\beta(x) - G(x)\| \leq \alpha/2$ für alle $x \in \mathrm{cl}\,(\Omega)$, so ist $d(F_\alpha, \Omega, y) = d(F_\beta, \Omega, y)$, wobei $F_\beta := I + G_\beta$.

Denn: Zunächst bette man X_α, X_β in einen gemeinsamen endlichdimensionalen linearen Teilraum $\hat{X} \subset X$ ein. Nach Definition 10.4 ist

$$d(F_\alpha, \Omega, y) = d(F_{\alpha,\mathrm{cl}\,(\Omega)\cap\hat{X}}, \Omega \cap \hat{X}, y), \qquad d(F_\beta, \Omega, y) = d(F_{\beta,\mathrm{cl}\,(\Omega)\cap\hat{X}}, \Omega \cap \hat{X}, y).$$

Nun definiere man die stetige Abbildung $H\colon (\mathrm{cl}\,(\Omega) \cap \hat{X}) \times [0,1] \longrightarrow \hat{X}$ durch

$$H(x,t) := (1-t)F_\alpha(x) + tF_\beta(x).$$

Dann ist $y \neq H(x,t)$ für alle $(x,t) \in \partial(\Omega \cap \hat{X}) \times [0,1] = (\partial\Omega \cap \hat{X}) \times [0,1]$, denn für beliebiges $(x,t) \in (\partial\Omega \cap \hat{X}) \times [0,1]$ ist

$$\begin{aligned}
\|H(x,t) - y\| &= \|F(x) - y + (1-t)[F_\alpha(x) - F(x)] + t[F_\beta(x) - F(x)]\| \\
&\geq \|F(x) - y\| - (1-t)\,\|F_\alpha(x) - F(x)\| - t\,\|F_\beta(x) - F(x)\| \\
&\geq \alpha - (1-t)\frac{\alpha}{2} - t\frac{\alpha}{2} \\
&= \frac{\alpha}{2} \\
&> 0.
\end{aligned}$$

Wegen Satz 9.10 von der Homotopie-Invarianz des Brouwerschen Abbildungsgrades bzw. der Bemerkung im Anschluss an den Beweis von Satz 10.5 folgt $d(F_\alpha, \Omega, y) = d(F_\beta, \Omega, y)$. Damit ist der Satz bewiesen. □

Dem Beweis des vorigen Satzes entnimmt man:

Satz 10.10 *Sei X ein linearer normierter Raum, $\Omega \subset X$ offen, beschränkt, $G\colon \mathrm{cl}\,(\Omega) \longrightarrow X$ eine kompakte Abbildung, $F := I + G\colon \mathrm{cl}\,(\Omega) \longrightarrow X$ und $y \notin F(\partial\Omega)$. Dann gilt:*

1. *Es ist*
$$\gamma := d(F(\partial\Omega), y) = \inf_{x\in\partial\Omega} \|F(x) - y\| > 0.$$

2. *Ist $G_\alpha\colon \mathrm{cl}\,(\Omega) \longrightarrow X_\alpha \subset X$ endlichdimensional mit*
$$\sup_{x\in\mathrm{cl}\,(\Omega)} \|G_\alpha(x) - G(x)\| \leq \alpha < \gamma$$
und $F_\alpha := I + G_\alpha$, so ist $d(F, \Omega, y) = d(F_\alpha, \Omega, y)$.

Nun übertragen sich alle aus dem $\mathbb{R}^n$ bekannten Sätze über den Brouwerschen Abbildungsgrad sinngemäß auf den Abbildungsgrad kompakter Störungen der Identität in einem linearen normierten Raum. Dies soll für den Satz über die Homotopie-Invarianz vorgeführt werden.

Satz 10.11 *Sei X ein linearer normierter Raum, $\Omega \subset X$ offen und beschränkt. Die Abbildung $H(\cdot,t)\colon \mathrm{cl}\,(\Omega) \longrightarrow X$ sei für jedes $t \in [0,1]$ kompakt und bezüglich t gleichmäßig stetig auf* cl (Ω)*, d. h. zu jedem $\epsilon > 0$ existiere ein $\delta = \delta(\epsilon) > 0$ derart, dass*

$$t_1, t_2 \in [0,1],\ |t_1 - t_2| \leq \delta \Longrightarrow \sup_{x\in\mathrm{cl}\,(\Omega)} \|H(x,t_1) - H(x,t_2)\| \leq \epsilon.$$

Ferner sei $y \in X$, $y \neq x + H(x,t)$ für alle $(x,t) \in \partial\Omega \times [0,1]$. Dann ist $d(I + H(\cdot,t), \Omega, y)$ konstant für $t \in [0,1]$.

Beweis: Der Beweis ist ist im wesentlichen genau wie der von Satz 9.10.

1. Es ist

$$\hat{\gamma} := \inf\{\|x + H(x,t) - y\| : (x,t) \in \partial\Omega \times [0,1]\} > 0.$$

Denn: Angenommen, es existiert eine Folge $\{(x_k, t_k)\} \subset \partial\Omega \times [0,1]$ mit $x_k + H(x_k, t_k) \to y$. Aus $\{t_k\} \subset [0,1]$ kann eine konvergente Teilfolge ausgewählt werden. Diese werde der Einfachheit halber wieder mit $\{t_k\}$ bezeichnet. Es sei also $t_0 := \lim_{k\to\infty} t_k$. Die Folge $\{x_k\} \subset \partial\Omega$ ist beschränkt und die Abbildung $H(\cdot, t_0)$ kompakt. Daher kann aus $\{H(x_k, t_0)\}$ eine konvergente Teilfolge ausgewählt werden. Sei also $z := \lim_{j\to\infty} H(x_{k_j}, t_0)$. Wegen der gleichmäßigen Stetigkeit von $H(\cdot, t)$ bezüglich t auf cl (Ω) ist $\lim_{j\to\infty} H(x_{k_j}, t_{k_j}) = z$. Hieraus und aus $x_{k_j} + H(x_{k_j}, t_{k_j}) \to y$ folgt die Konvergenz von $\{x_{k_j}\}$ und es ist $x_{k_j} \to w := y - z \in \partial\Omega$. Dann folgt $y = w + H(w, t_0)$ mit $(w, t_0) \in \partial\Omega \times [0,1]$, ein Widerspruch.

2. Man wähle $\delta > 0$ so klein, dass

$$t_1, t \in [0,1],\ |t_1 - t| \le \delta \Longrightarrow \sup_{x\in \text{cl }(\Omega)} \|H(x,t_1) - H(x,t)\| \le \frac{\hat{\gamma}}{2},$$

was wegen der gleichmäßigen Stetigkeit von $H(\cdot, t)$ bezüglich t auf cl (Ω) möglich ist. Sei $H_\alpha(\cdot, t_1)$: cl $(\Omega) \longrightarrow X_\alpha \subset X$ eine endlichdimensionale kompakte Approximation an $H(\cdot, t_1)$ mit

$$\sup_{x\in \text{cl }(\Omega)} \|H_\alpha(x,t_1) - H(x,t_1)\| \le \frac{\hat{\gamma}}{4}.$$

Dies ist wegen Satz 10.7 möglich. Nach Satz 10.10 ist $d(I + H(\cdot, t_1), \Omega, y) = d(I + H_\alpha(\cdot, t_1), \Omega, y)$. Für $t \in [0,1]$ mit $|t_1 - t| \le \delta$ ist

$$\begin{aligned}
\|H_\alpha(x,t_1) - H(x,t)\| &\le \|H_\alpha(x,t_1) - H(x,t_1)\| + \|H(x,t_1) - H(x,t)\| \\
&\le \frac{\hat{\gamma}}{4} + \frac{\hat{\gamma}}{2} \\
&= \frac{3}{4}\hat{\gamma} \\
&< \gamma
\end{aligned}$$

für alle $x \in$ cl (Ω), sodass wiederum nach Satz 10.10

$$d(I + H(\cdot,t), \Omega, y) = d(I + H_\alpha(\cdot,t_1), \Omega, y) = d(I + H(\cdot,t_1), \Omega, y)$$

falls $|t - t_1| \le \delta$. Da man das Intervall $[0,1]$ durch endlich viele Intervalle der Länge δ überdecken kann, folgt hieraus die Behauptung. □

10.2 Fixpunkt- und Existenzsätze in linearen normierten Räumen

Ziel ist es, Existenzsätze für Gleichungen der Form $x + G(x) = y$ in einem linearen normierten Raum X zu beweisen, wobei wir gelegentlich auch die Vollständigkeit von

X voraussetzen müssen. Wir werden sämtliche Existenzsätze mit Hilfe des Abbildungsgrades beweisen, obwohl z. B. der Beweis des Schauderschen Fixpunktsatzes mit Hilfe des Brouwerschen Fixpunktsatzes auch direkter erbracht werden könnte.

Der erste Existenzsatz entspricht Satz 9.18, dem Satz von Kronecker.

Satz 10.12 *Sei X ein linearer normierter Raum, $\Omega \subset X$ offen und beschränkt, die Abbildung $G\colon \mathrm{cl}\,(\Omega) \longrightarrow X$ sei kompakt sowie $F := I + G$ und $y \in X \setminus F(\partial\Omega)$. Dann existiert ein $x \in \Omega$ mit $F(x) = x + G(x) = y$.*

Beweis: Nach Definition 10.8 des Abbildungsgrades bzw. der Sätze 10.9, 10.10 ist $d(F, \Omega, y) = d(F_k, \Omega, y)$ für alle hinreichend großen k, wenn $F_k = I + G_k$ eine endlichdimensionale Störung der Identität mit $\sup_{x\in \mathrm{cl}\,(\Omega)} \|F_k(x) - F(x)\| \to 0$ für $k \to \infty$. Für alle $k \geq k_0$ mit hinreichend großem k_0 ist also auch $d(F_k, \Omega, y) \neq 0$. Aus dem endlichdimensionalen Satz von Kronecker (Satz 9.18) folgt für $k \geq k_0$ die Existenz von $x_k \in \Omega \cap X_k \subset \Omega$ (hierbei ist $X_k \subset X$ ein endlichdimensionaler linearer Raum mit $G_k(\mathrm{cl}\,(\Omega)) \subset X_k$) mit $F_k(x_k) = x_k + G_k(x_k) = y$. Die Folge $\{x_k\}_{k\geq k_0} \subset \Omega$ ist, beschränkt, die Abbildung $G\colon \mathrm{cl}\,(\Omega) \longrightarrow X$ ist kompakt. Daher existiert eine Teilfolge $\{x_{k_j}\} \subset \{x_k\}_{k\geq k_0}$ derart, dass $z := \lim_{j\to\infty} G(x_{k_j})$ existiert. Sei $x := y - z$. Dann ist

$$\begin{aligned} \|x_{k_j} - x\| &= \|G_{k_j}(x_{k_j} - z\| \\ &\leq \underbrace{\|G_{k_j}(x_{k_j}) - G(x_{k_j})\|}_{\to 0} + \underbrace{\|G(x_{k_j}) - z\|}_{\to 0}, \end{aligned}$$

also $x_{k_j} \to x$. Aus

$$\begin{aligned} y &= F_{k_j}(x_{k_j}) \\ &= x_{k_j} + G_{k_j}(x_{k_j}) \\ &= \underbrace{x_{k_j}}_{\to x} + G(x) + \underbrace{G_{k_j}(x_{k_j}) - G(x_{k_j})}_{\to 0} + \underbrace{G(x_{k_j}) - G(x)}_{\to 0} \end{aligned}$$

folgt $y = x + G(x) = F(x)$. Wegen $\{x_{k_j}\} \subset \Omega$ ist $x \in \mathrm{cl}\,(\Omega)$. Da $y \notin F(\partial\Omega)$ ist $x \notin \partial\Omega$, insgesamt also $x \in \Omega$. Damit ist der Satz bewiesen. □

Satz 10.12 wird, genau wie im endlichdimensionalen Fall, zusammen mit dem Homotopiesatz 10.11 das Haupthilfsmittel zum Beweis der folgenden Existenzsätze sein. Unser erstes Ziel ist der Beweis des Schauderschen Fixpunktsatzes. Beim Beweis des Brouwerschen Fixpunktsatzes (Satz 9.19) gingen wir folgendermaßen vor:

1. Zunächst wird die Aussage für eine (abgeschlossene) Kugel bewiesen.

2. Der allgemeine Fall einer kompakten, konvexen Menge K, die durch die Abbildung G stetig in sich abgebildet wird, wird auf 1. zurückgeführt, indem G stetig auf eine K umfassende Kugel fortgesetzt wird und zwar so, dass auch das Bild der Fortsetzung in K liegt.

Um diesen Zugang zu simulieren, wird man versuchen, den folgenden Satz zu beweisen.

Satz 10.13 *Sei X ein linearer normierter Raum, $K \subset X$ eine nichtleere, kompakte, konvexe Menge. Die Abbildung $G\colon K \longrightarrow X$ sei stetig und $G(K) \subset K$. Dann lässt sich G zu einer stetigen Abbildung $\tilde{G}\colon X \longrightarrow K \subset X$ fortsetzen.*

Beweis: Die Idee des Beweises besteht darin, die Behauptung für die Identität zu zeigen. Denn angenommen, $\tilde{I}\colon X \longrightarrow K$ sei die gesuchte Fortsetzung der Identität. Dann ist $\tilde{G} := G \circ \tilde{I}$ die gesuchte Fortsetzung von G.

Die kompakte Menge $K \subset X$ ist *separabel*, d. h. es existiert eine abzählbare, dichte Teilmenge $\{y_i\}_{i\in\mathbb{N}} \subset K$[28].Für $x \in X \setminus K$ definiere man

$$\lambda_i(x) := \max\Big\{2 - \frac{\|y_i - x\|}{d(x,K)}, 0\Big\}, \quad i \in \mathbb{N},$$

wobei

$$d(x,K) := \inf_{y\in K} \|x - y\|$$

den Abstand von x zu K bedeutet. Dann gilt:

- $\lambda_i\colon X \setminus K \longrightarrow \mathbb{R}$ ist stetig.
- Für $x \in X \setminus K$ und $i \in \mathbb{N}$ ist $0 \le \lambda_i(x) \le 1$.

  Denn es ist $d(x,K) \le \|y_i - x\|$.
- Zu jedem $x \in X \setminus K$ existiert ein $i \in \mathbb{N}$ mit $\lambda_i(x) > 0$.

  Denn: Da K kompakt ist, existiert $y_0 \in K$ mit $\|x - y_0\| = d(x,K)$. Sa $\{y_i\}_{i\in\mathbb{N}}$ dicht in K ist, existiert ein Index $i \in \mathbb{N}$ mit $\|y_i - y_0\| \le \frac{1}{2}d(x,K)$. Dann ist aber $\|y_i - x\| \le \|y_i - y_0\| + \|y_0 - x\| \le \frac{3}{2}d(x,K)$, also $\|y_i - x\|/d(x,K) \le \frac{3}{2}$ und damit $\lambda_i(x) \ge \frac{1}{2}$.

Nun definieren wir $\lambda\colon X \setminus K \longrightarrow \mathbb{R}$ durch

$$\lambda(x) := \sum_{i=1}^{\infty} \frac{1}{2^i}\lambda_i(x).$$

Wegen $\lambda_i(x) \in [0,1]$ ist die Konvergenz der $\lambda(\cdot)$ definierenden Reihe gesichert. Offenbar ist $\lambda(\cdot)$ auf $X \setminus K$ positiv und stetig (als gleichmäßiger Limes stetiger Funktionen, nämlich der Partialsummen). Nun definieren wir $\alpha_i\colon X \setminus K \longrightarrow \mathbb{R}$, $i \in \mathbb{N}$, durch

$$\alpha_i(x) := \frac{\lambda_i(x)}{2^i\lambda(x)}, \quad i \in \mathbb{N}.$$

Für alle $x \in X \setminus K$ ist dann $\alpha_i(x) \ge 0$, $\sum_{i=1}^{\infty} \alpha_i(x) = 1$. Wir werden zeigen:

1. Für alle $x \in X \setminus K$ existiert $\alpha(x) := \lim_{n\to\infty} \sum_{i=1}^{n} \alpha_i(x)y_i$.

[28]Denn für jedes $k \in \mathbb{N}$ ist $\bigcup_{x\in K} B(x;1/k) \cap K$ eine Überdeckung von K, aus der wegen der Kompaktheit von K eine endliche Teilüberdeckung $\bigcup_{l=1}^{l_k} B(x_l^k;1/k) \cap K$ auswählbar ist. Dann ist $\{x_l^k : l = 1,\dots,l_k,\ k \in \mathbb{N}\}$ eine abzählbare, dichte Teilmenge von K.

2. Für alle $x \in X \setminus K$ ist $\alpha(x) \in K$.

3. $\alpha(\cdot)$ ist auf $X \setminus K$ stetig.

Hierzu machen wir zunächst eine Vorbemerkung:

- Eine kompakte Menge K in einem linearen normierten Raum X ist vollständig (d. h. jede Cauchyfolge aus K konvergiert gegen ein Element aus K).

Denn: Sei $\{x_n\} \subset K$ eine Cauchyfolge. Da K kompakt ist, existiert eine konvergente Teilfolge $x_{n_k} \to x \in K$. Dann ist

$$\|x_n - x\| \leq \|x_n - x_{n_k}\| + \|x_{n_k} - x\| \to 0,$$

da $\{x_n\}$ eine Cauchyfolge ist und $\{x_{n_k}\}$ gegen x konvergiert.
Für $x \in X \setminus K$ sei

$$s_n(x) := \sum_{i=1}^{n} \alpha_i(x) y_i + \Big(\sum_{i=n+1}^{\infty} \alpha_i(x) \Big) y_{n+1}.$$

Dann ist $s_n(x)$ eine Konvexkombination der $n+1$ Punkte $\{y_1, \ldots, y_{n+1}\} \subset K$. Da K konvex ist, ist $s_n(x) \in K$. Wir zeigen, dass $\{s_n(x)\}$ eine Cauchyfolge ist. Da K kompakt ist, ist K beschränkt. Insbesondere existiert eine Konstante C mit $\|y_i\| \leq C$, $i = 1, 2, \ldots$. Für $n < m$ ist

$$\|s_n(x) - s_m(x)\| \leq C \Big(\sum_{i=n+1}^{m} \alpha_i(x) + \sum_{i=n+1}^{\infty} \alpha_i(x) + \sum_{i=m+1}^{\infty} \alpha_i(x) \Big),$$

woraus man abliest, dass $\{s_n(x)\}$ eine Cauchyfolge ist. Nach obiger Bemerkung ist die Folge $\{s_n(x)\}$ konvergent, es existiert also

$$s(x) := \lim_{n \to \infty} s_n(x) \in K.$$

Wegen $\Big(\sum_{i=n+1}^{\infty} \alpha_i(x) \Big) y_{n+1} \to 0$ folgt dann

$$s(x) = \alpha(x) = \lim_{n \to \infty} \sum_{i=1}^{n} \alpha_i(x) y_i \in K.$$

Damit sind 1. und 2. bewiesen. Zu zeigen bleibt die Stetigkeit von $\alpha(\cdot)$ auf $X \setminus K$. Diese folgt aber sofort aus der Darstellung

$$\alpha(x) = \frac{1}{\lambda(x)} \sum_{i=1}^{\infty} \frac{\lambda_i(x)}{2^i} y_i.$$

Nun setze man

$$\tilde{I}(x) := \begin{cases} \alpha(x), & x \in X \setminus K, \\ x, & x \in K. \end{cases}$$

Dann ist $\tilde{I}$ auf K und auf $X \setminus K$ stetig. Um die Stetigkeit von $\tilde{I}$ auf ganz X zu beweisen, genügt es zu zeigen:

$$\{x_n\} \subset X \setminus K,\ \lim_{n\to\infty} x_n = x_0 \Longrightarrow \lim_{n\to\infty} \alpha(x_n) = x_0.$$

Sei $\epsilon > 0$ gegeben und x ein beliebiges, festes Element der Folge $\{x_n\}$ mit $\|x - x_0\| < \epsilon$. Sei

$$I := \{i \in \mathbb{N} : \|y_i - x\| < 2\epsilon\}.$$

Für $i \notin I$ ist also $\|y_i - x\| \geq 2\epsilon$, wegen $d(x, K) \leq \|x - x_0\| < \epsilon$ ist also $\lambda_i(x) = 0$ und folglich auch $\alpha_i(x) = 0$. Daher ist $\alpha(x) = \sum_{i\in I} \alpha_i(x) y_i$ und daher

$$\|\alpha(x) - x\| = \Big\|\sum_{i\in I} \alpha_i(x)(y_i - x)\Big\| \leq \sum_{i\in I} \alpha_i(x)\, \|y_i - x\| \leq 2\epsilon.$$

Daher ist

$$\|\alpha(x) - x_0\| \leq \|\alpha(x) - x\| + \|x - x_0\| < 2\epsilon + \epsilon = 3\epsilon.$$

Wegen der Konvergenz der Folge $\{x_n\}$ gegen x_0 ist $\|x_n - x_0\| < \epsilon$ für alle hinreichend großen n und daher auch $\|\alpha(x_n) - x_0\| < 3\epsilon$ für alle hinreichend großen n bzw. $\lim_{n\to\infty} \alpha(x_n) = x_0$. Damit ist die Stetigkeit von $\tilde{I}$ auf ganz X bewiesen. Nach der Bemerkung am Anfang des Beweises ist der Satz vollständig bewiesen. □

Es folgt ein erster Fixpunktsatz.

Satz 10.14 *Sei X ein linearer normierter Raum, $K \subset X$ eine nichtleere, konvexe, kompakte Teilmenge und $G\colon K \longrightarrow X$ stetig mit $G(K) \subset K$. Dann besitzt G einen Fixpunkt $x \in K$.*

Beweis: Als kompakte Menge ist K beschränkt. Daher existiert eine (offene) Kugel $B(0; r)$ um den Nullpunkt in X und dem Radius $r > 0$ mit $K \subset B(0; r)$. Nach Satz 10.13 existiert eine stetige Erweiterung $\tilde{G}\colon B[0; r] \longrightarrow K \subset X$ von G auf die (abgeschlossene) Kugel $B[0; r]$. Nun definiere man $H\colon B[0; r] \times [0, 1] \longrightarrow X$ durch $H(x, t) := -t\tilde{G}(x)$. Dann gilt:

- Für alle $t \in [0, 1]$ ist $H(\cdot, t)\colon B[0; r] \longrightarrow X$ kompakt.

  Denn: Für $t \in [0, 1]$ ist $H(\cdot, t) = -t\tilde{G}(\cdot)$ stetig. Da

  $$H(B[0; r], t) = -t\tilde{G}(B[0; r]) \subset -tK$$

  und $-tK$ kompakt ist, ist $H(B[0; r], t)$ als Teilmenge einer kompakten Menge relativ kompakt.

- Die Abbildung $H(\cdot, t)\colon B[0; r] \longrightarrow X$ ist bezüglich t gleichmäßig stetig auf $B[0; r]$, d. h. zu jedem $\epsilon > 0$ existiert ein $\delta = \delta(\epsilon) > 0$ derart, dass

  $$t_1, t_2 \in [0, 1],\ |t_1 - t_2| \leq \delta \Longrightarrow \sup_{x\in B[0;r]} \|H(x, t_1) - H(x, t_2)\| \leq \epsilon.$$

  Denn: Sei $\epsilon > 0$ gegeben. Ist $t_1, t_2 \in [0, 1]$ und $|t_1 - t_2| \leq \delta(\epsilon) := \epsilon/r$, so ist

  $$\|H(x, t_1) - H(x, t_2)\| = |t_1 - t_2|\, \|\tilde{G}(x)\| \leq \frac{\epsilon}{r} r = \epsilon.$$

- Es ist $0 \neq x + H(x,t) = x - t\tilde{G}(x)$ für alle $(x,t) \in \partial B(0;r) \times [0,1]$.
  Denn: Es ist $\tilde{G}(B[0;r]) \subset K \subset B(0;r)$, woraus die Aussage sofort folgt.

Aus Satz 10.11 erhalten wir

$$1 = d(I, B(0;r), 0) = d(I - \tilde{G}, B(0;r), 0).$$

Eine Anwendung von Satz 10.12 liefert die Existenz eines $x \in B(0;r)$ mit $x - \tilde{G}(x) = 0$. Hieraus folgt $x = \tilde{G}(x) \in K$. Daher ist $x = \tilde{G}(x) = G(x)$ und der Satz ist bewiesen. □

Für Anwendungen ist Satz 10.14 nicht besonders gut geeignet ist, weil es in unendlichdimensionalen linearen normierten Räumen, grob gesprochen, nur wenige[29] kompakte Mengen gibt. Der für viele Anwendungen wichtige Schaudersche Fixpunktsatz wird aber aus Satz 10.14 leicht folgen, wenn wir den nächsten Satz benutzen.

Satz 10.15 *Sei X ein Banachraum, $E \subset X$ relativ kompakt. Dann ist auch die konvexe Hülle* co (E) *von E, also die kleinste E enthaltende konvexe Menge, relativ kompakt.*

Beweis: Man bestätigt sehr leicht, dass die konvexe Hülle co (E) einer Menge E in einem linearen Raum X durch

$$\text{co}\,(E) = \Big\{ \sum_{i=1}^{m} \alpha_i x_i : x_i \in E,\ \alpha_i \geq 0\ (i = 1, \ldots, m),\ \sum_{i=1}^{m} \alpha_i = 1,\ m \in \mathbb{N} \Big\}$$

dargestellt werden kann. Man zeige nun, dass co (E) totalbeschränkt[30] ist. Da in einem vollständigen metrischen Raum, speziell in einem Banachraum, eine totalbeschränkte Teilmenge relativ kompakt ist[31], sind wir dann fertig. Sei hierzu $\epsilon > 0$

[29] Bekanntlich ist die abgeschlossene Einheitskugel in einem linearen normierten Raum genau dann kompakt, wenn X endlichdimensional ist.

[30] Eine Teilmenge A eines linearen normierten (oder metrischen) Raumes X heißt *totalbeschränkt*, wenn es zu jedem $\epsilon > 0$ eine Überdeckung der Menge durch endlich viele offene ϵ-Kugeln gibt, wenn es also zu jedem $\epsilon > 0$ endlich viele $\{x_1, \ldots, x_m\} \subset X$, ein sogenanntes *$\epsilon$-Netz*, mit $A \subset \bigcup_{i=1}^{m} B(x_i;\epsilon)$. Ist $A \subset X$ totalbeschränkt, so existieren zu jedem $\epsilon > 0$ sogar endlich viele $\{y_1, \ldots, y_m\} \subset A$ mit $A \subset \bigcup_{i=1}^{m} B(y_i;\epsilon)$. Denn: Sei $\epsilon > 0$ vorgegeben. Da A totalbeschränkt ist, existieren $\{x_1, \ldots, x_m\} \subset X$ mit $A \subset \bigcup_{i=1}^{m} B(x_i; \frac{1}{2}\epsilon)$. Hierbei können wir annehmen, dass

$$A \cap B(x_i; \tfrac{1}{2}\epsilon) \neq \emptyset, \qquad i = 1, \ldots, m.$$

Denn ist $A \cap B(x_i; \frac{1}{2}\epsilon) = \emptyset$, so trägt die Kugel $B(x_i; \frac{1}{2}\epsilon)$ nichts zur Überdeckung von A bei und kann daher weggelassen werden. Nun bestimme man

$$y_i \in A \cap B(x_i; \tfrac{1}{2}\epsilon), \qquad i = 1, \ldots, m.$$

Dann ist $B(x_i; \frac{1}{2}\epsilon) \subset B(y_i;\epsilon)$, $i = 1, \ldots, m$, wie man sofort mit Hilfe der Dreiecksungleichung nachweist. Damit ist gezeigt, dass die Mittelpunkte der überdeckenden Kugeln o. B. d. A. zu A gehören.

[31] Denn: Sei A eine totalbeschränkte Teilmenge eines Banachraumes X und $\{a_n\} \subset A$ eine Folge. Wir zeigen, dass $\{a_n\}$ eine Cauchyfolge als Teilfolge besitzt. Da X als vollständig vorausgesetzt wurde, ist diese Teilfolge konvergent und die Menge A daher relativ kompakt. Für $k \in \mathbb{N}$ sei $\epsilon_k := 1/k$. Da A totalbeschränkt ist, existieren $\{x_{11}, \ldots, x_{1m_1}\} \subset A$ mit $A \subset \bigcup_{i=1}^{m_1} B(x_{1i};\epsilon_1)$. Wir können annehmen, dass unendlich viele Folgenglieder von $\{a_n\}$ in $B(x_{11};\epsilon_1)$ liegen. Wir wählen nun ein Folgenglied $a_{n_1} \in A_1 := A \cap B(x_{11};\epsilon_1)$. Als Teilmenge der totalbeschränkten Menge A ist auch A_1 totalbeschränkt,

vorgegeben. Als relativ kompakte Menge ist E totalbeschränkt[32]. Daher existieren $\tilde{E} := \{y_1, \ldots, y_n\} \subset E$ mit $E \subset \bigcup_{i=1}^n B(y_i; \frac{1}{2}\epsilon)$, also ein $\frac{1}{2}\epsilon$-Netz für E. Die konvexe Hülle co $(\tilde{E})$ ist als beschränkte Menge in dem durch $\tilde{E}$ erzeugten endlichdimensionalen linearen Raum selbst relativ kompakt und damit totalbeschränkt. Daher existieren endlich viele $\{z_1, \ldots, z_k\} \subset$ co $(\tilde{E})$ mit co $(\tilde{E}) \subset \bigcup_{i=1}^k B(z_i; \frac{1}{2}\epsilon)$, also ein $\frac{1}{2}\epsilon$-Netz für co $(\tilde{E})$. Wir zeigen, dass $\{z_1, \ldots, z_k\}$ ein ϵ-Netz für co (E) ist bzw. co $(E) \subset \bigcup_{i=1}^k B(z_i; \epsilon)$ gilt. Sei hierzu $x \in$ co (E) beliebig. Dann lässt sich x in der Form $x = \sum_{i=1}^m \alpha_i x_i$ mit $\alpha_i \geq 0$, $x_i \in E$, $i = 1, \ldots, m$, mit $m \in \mathbb{N}$ und $\sum_{i=1}^m \alpha_i = 1$ darstellen. Da $\tilde{E} = \{y_1, \ldots, y_n\}$ ein $\frac{1}{2}\epsilon$-Netz für E ist bzw. $E \subset \bigcup_{i=1}^n B(y_i; \frac{1}{2}\epsilon)$ gilt, existiert zu jedem x_i ein y_{k_i} mit $\|x_i - y_{k_i}\| < \frac{1}{2}\epsilon$, $i = 1, \ldots, n$. Dann ist $\tilde{y} := \sum_{i=1}^m \alpha_i y_{k_i} \in$ co $(\tilde{E})$. Da andererseits $\{z_1, \ldots, z_k\}$ ein $\frac{1}{2}\epsilon$-Netz für co $(\tilde{E})$ bildet, existiert ein $z_j \in \{z_1, \ldots, z_k\}$ mit $\|\tilde{y} - z_j\| < \frac{1}{2}\epsilon$. Insgesamt ist dann

$$\begin{aligned} \|x - z_j\| &\leq \|x - \tilde{y}\| + \|\tilde{y} - z_j\| \\ &= \Big\|\sum_{i=1}^m \alpha_i (x_i - y_{k_i})\Big\| + \|\tilde{y} - z_j\| \\ &< \tfrac{1}{2}\epsilon + \tfrac{1}{2}\epsilon \\ &= \epsilon. \end{aligned}$$

Damit ist der Satz bewiesen. □

Nun kommen wir zum *Schauderschen Fixpunktsatz*, siehe J. Schauder (1930).

Satz 10.16 (Schauder) *Sei X ein Banachraum, $A \subset X$ eine nichtleere, abgeschlossene und konvexe Menge. Die Abbildung $F\colon A \subset X \longrightarrow X$ sei stetig und $F(A) \subset A$ sowie $F(A)$ relativ kompakt[33]. Dann besitzt F einen Fixpunkt in A.*

Beweis: Sei $K :=$ cl (co $(F(A))$). Dann gilt:

1. K ist kompakt.

   Denn: Da $F(A)$ nach Voraussetzung relativ kompakt ist, ist co $(F(A))$ nach Satz 10.15 relativ kompakt und damit $K =$ cl (co $(F(A))$) kompakt.

besitzt also eine endliche Überdeckung durch offene ϵ_2-Kugeln. Es existieren also $\{x_{21}, \ldots, x_{2m_2}\} \subset A_1$ mit $A_1 \subset \bigcup_{i=1}^{m_2} B(x_{2i}; \epsilon_2)$. Wir können annehmen, dass in $B(x_{21}; \epsilon_2)$ unendlich viele Folgenglieder von $\{a_n\}$ liegen. Man wähle ein $a_{n_2} \in A_2 := A_1 \cap B(x_{21}; \epsilon_2)$, wobei $n_2 > n_1$. Im k-ten Schritt wählt man $a_{n_k} \in A_k := A_{k-1} \cap B(x_{k1}; \epsilon_k)$ mit $n_k > n_{k-1}$. Dann ist die Teilfolge $\{a_{n_k}\}$ eine Cauchyfolge. Denn für $k, l \geq K$ ist $a_{n_k}, a_{n_l} \in A_K \subset B(x_{K1}; \epsilon_K)$ und daher

$$\|a_{n_k} - a_{n_l}\| \leq \|a_{n_k} - x_{K1}\| + \|x_{K1} - a_{n_l}\| \leq 2\epsilon_k = \frac{2}{K}.$$

Wie wir oben schon begründet haben ist damit die Behauptung bewiesen.

[32] Wäre E nicht totalbeschränkt, so gäbe es ein $\epsilon > 0$ mit der Eigenschaft, dass E nicht durch endlich viele offene ϵ-Kugeln überdeckt werden kann. In naheliegender Weise erhält man hieraus die Existenz einer Folge $\{x_i\} \subset E$ mit $\|x_i - x_j\| \geq \epsilon$ für $i \neq j$. Dann hätte $\{x_i\}$ keine konvergente Teilfolge, ein Widerspruch dazu, dass E relativ (folgen)kompakt ist.

[33] Dies ist z. B. erfüllt, wenn A beschränkt und $F\colon A \longrightarrow X$ kompakt ist.

2. Es ist $K \subset A$.

   Denn: Wegen $F(A) \subset A$ und der Konvexität von A ist $\text{co}\,(F(A)) \subset A$. Da A abgeschlossen ist, ist

$$K = \text{cl}\,(\text{co}\,(F(A))) \subset \text{cl}\,(A) = A.$$

3. Es ist $F(K) \subset K$.

   Denn: Wegen $K \subset A$ ist

$$F(K) \subset F(A) \subset \text{co}\,(F(A)) \subset \text{cl}\,(\text{co}\,(F(A))) = K.$$

Also ist G eine stetige Abbildung, die die kompakte, konvexe (beachte: Der Abschluss einer konvexen Menge ist konvex) Menge K in sich abbildet. Aus Satz 10.14 folgt, dass G einen Fixpunkt in $K \subset A$ besitzt. Damit ist der Schaudersche Fixpunktsatz bewiesen. □

Man kann sich fragen, ob in Satz 10.14 vorausgesetzt werden muss, dass die Menge K kompakt und konvex ist, und ob es nicht genügt vorauszusetzen, dass K konvex, abgeschlossen und konvex ist. Oder: Muss im Schauderschen Fixpunktsatz $G(A)$ als relativ kompakt vorausgesetzt werden? Kann diese Bedingung gestrichen werden oder durch die Voraussetzung ersetzt werden, dass A beschränkt ist? Hierzu geben wir ein Beispiel an.

Beispiel: Sei $X := l^2$ der reelle Hilbertsche Folgenraum, also

$$l^2 := \Big\{x = \{x_i\}_{i\in\mathbb{N}} : x_i \in \mathbb{R},\ (i \in \mathbb{N}),\ \sum_{i=1}^{\infty} x_i^2 < \infty\Big\}, \qquad \|x\| = \Big(\sum_{i=1}^{\infty} x_i^2\Big)^{1/2}.$$

Die Einheitskugel $B[0;1] := \{x \in X : \|x\| \leq 1\}$ ist abgeschlossen, beschränkt und konvex. Wir definieren $F\colon B[0;1] \longrightarrow X$ durch $F(x) := (\sqrt{1 - \|x\|^2}, x_1, x_2, \ldots)$. Dann gilt:

1. Es ist $F(B[0;1]) \subset \partial B[0;1] \subset B[0;1]$.

   Denn: Ist $\|x\| \leq 1$ bzw. $x \in B[0;1]$, so ist $F(x)$ definiert und

$$\|F(x)\|^2 = 1 - \|x\|^2 + \sum_{i=1}^{\infty} x_i^2 = 1 - \|x\|^2 + \|x\|^2 = 1$$

   bzw. $F(x) \in \partial B[0;1]$.

2. Die Abbildung $F\colon B[0;1] \longrightarrow X$ ist stetig.

   Denn: Seien $x, y \in B[0;1]$. Dann ist

$$\|F(x) - F(y)\|^2 = (\sqrt{1 - \|x\|^2} - \sqrt{1 - \|y\|^2})^2 + \|x - y\|^2,$$

   woraus man die Stetigkeit von F abliest.

3. Die Abbildung F besitzt in $B[0;1]$ keinen Fixpunkt.

   Denn: Angenommen, es existiert ein $x^* \in B[0;1]$ mit $F(x^*) = x^*$. Wegen

$$F(B[0;1]) \subset \partial B[0;1]$$

   ist $\|x^*\| = 1$ und daher

$$F(x^*) = (0, x_1^*, x_2^*, \ldots) = (x_1^*, x_2^*, x_3^*, \ldots) = x^*.$$

   Daher ist $x_i^* = 0$, $i \in \mathbb{N}$, bzw. $x^* = 0$, ein Widerspruch zu $\|x^*\| = 1$.

Daher hat eine stetige Abbildung, die eine abgeschlossene, beschränkte, konvexe Teilmenge eines unendlichdimensionalen linearen normierten Raumes in sich abbildet, i. Allg. keinen Fixpunkt. □

Als Vorbereitung für den Beweis des nächsten Fixpunktsatzes formulieren und beweisen wir (siehe z. B. A. E. Taylor (1958, S. 135)):

Lemma 10.17 *Sei $p\colon X \longrightarrow \mathbb{R}$ das durch*

$$p(x) := \inf A_x \quad \textit{mit} \quad A_x := \{\alpha : \alpha > 0,\ x \in \alpha\Omega\}$$

definierte[34] Minkowski-Funktional *zu Ω. Dann gilt*

1. *Es ist $p(0) = 0$ und $p(\lambda x) = \lambda p(x)$ für alle $x \in X$, $\lambda > 0$.*
2. *Es ist $p(x + y) \le p(x) + p(y)$ für alle $x, y \in X$.*
3. *Die Abbildung $p\colon X \longrightarrow \mathbb{R}$ ist stetig.*
4. *Es ist*
$$\Omega = \{x \in X : p(x) < 1\}, \qquad \text{cl}\,(\Omega) = \{x \in X : p(x) \le 1\}.$$

Beweis: Wegen $0 \in \Omega$ ist $p(0) = 0$. Sei $x \in X$ und $\lambda > 0$. Für ein beliebiges $\alpha \in A_x$ ist $x \in \alpha\Omega$, daher $\lambda x \in \lambda\alpha\Omega$ und folglich $p(\lambda x) \le \lambda p(x)$. Anderseits ist

$$p(x) = p(\lambda^{-1}\lambda x) \le \lambda^{-1} p(\lambda x)$$

und insgesamt $p(\lambda x) = \lambda p(x)$. Sind $x, y \in X$ und $\alpha \in A_x$, $\beta \in A_y$, so ist

$$x + y \in \alpha\Omega + \beta\Omega = (\alpha + \beta)\Big(\frac{\alpha}{\alpha + \beta}\Omega + \frac{\beta}{\alpha + \beta}\Omega\Big) \subset (\alpha + \beta)\Omega$$

wegen der Konvexität von Ω und daher $p(x + y) \le p(x) + p(y)$. Da $0 \in \Omega$ und Ω offen ist, existiert ein $\epsilon > 0$ mit $B[0;\epsilon] \subset \Omega$. Für ein beliebiges $x \in X \setminus \{0\}$ ist $\epsilon x/\|x\| \in B[0;\epsilon] \subset \Omega$ bzw. $x \in (\|x\|/\epsilon)\Omega$ und folglich $p(x) \le (1/\epsilon)\|x\|$ für alle $x \in X$. Für beliebige $x, y \in X$ ist

$$p(x) = p(y + (x - y)) \le p(y) + p(x - y)$$

[34] Da $0 \in \Omega$ und Ω beschränkt ist, ist p wohldefiniert.

und daher

$$p(x) - p(y) \leq p(x-y) \leq \frac{1}{\epsilon}\|x-y\|.$$

Vertauscht man hier x und y, so erhält man

$$p(y) - p(x) \leq p(y-x) \leq \frac{1}{\epsilon}\|x-y\|,$$

insgesamt also

$$|p(x) - p(y)| \leq \frac{1}{\epsilon}\|x-y\|.$$

Hieraus liest man die Stetigkeit von p auf X ab. Zu zeigen bleibt, dass

$$\Omega = \{x \in X : p(x) < 1\},$$

denn hieraus folgt wegen der Stetigkeit von p sofort $\mathrm{cl}\,(\Omega) = \{x \in X : p(x) \leq 1\}$. Ist $x \in \Omega$, so ist $\lambda x \in \Omega$ für alle $\lambda \in [0,1]$ wegen $0 \in \Omega$ und der Konvexität von Ω. Da Ω offen ist, existiert ein $\epsilon > 0$ mit $(1+\epsilon)x \in \Omega$. Folglich ist $p(x) \leq 1/(1+\epsilon) < 1$. Ist umgekehrt $p(x) < 1$, so existiert $\alpha \in (0,1)$ mit $x \in \alpha\Omega$ bzw. $(1/\alpha)x \in \Omega$. Wegen der Konvexität von Ω ist dann $(\lambda/\alpha)x \in \Omega$ für alle $\lambda \in [0,1]$. Insbesondere (setze $\lambda = \alpha$) ist $x \in \Omega$. Damit ist das Lemma bewiesen. □

Wir wollen jetzt noch eine Reihe (zumindestens theoretisch) interessanter Existenzaussagen formulieren und beweisen. Der erste stammt von E. Rothe (1938, S. 186), siehe z. B. auch D. R. Smart (1974, S. 27).

Satz 10.18 (Rothe) *Sei X ein Banachraum und $\Omega \subset X$ eine offene, konvexe, beschränkte Menge mit $0 \in \Omega$. Die Abbildung $G\colon \mathrm{cl}\,(\Omega) \subset X \longrightarrow X$ sei kompakt und $G(\partial\Omega) \subset \Omega$. Dann besitzt G einen Fixpunkt in Ω.*

Beweis: Mit p bezeichne man das zu Ω gehörende Minkowski-Funktional. Weiter definiere man $q\colon X \longrightarrow \mathbb{R}$ und die Abbildung $g\colon X \longrightarrow X$ durch

$$q(x) := \max(p(x), 1), \qquad g(x) := \frac{1}{q(x)}x.$$

Dann gilt:

1. Die Abbildung $g\colon X \longrightarrow X$ ist stetig.

   Denn: Dies ist trivial, da q stetig und $q(x) \geq 1$ für alle $x \in X$ ist.

2. Es ist $g(X) \subset \mathrm{cl}\,(\Omega)$.

   Denn: Sei $x \in X$. Dann ist

   $$p(g(x)) = p\Big(\frac{1}{q(x)}x\Big) = \frac{1}{q(x)}p(x) \leq 1$$

   und folglich $g(x) \in \mathrm{cl}\,(\Omega)$.

3. Die Restriktion $\mathrm{Rest}_{\mathrm{cl}\,(\Omega)}g$ von g auf $\mathrm{cl}\,(\Omega)$ ist die Identität.

   Denn: Ist $x \in \mathrm{cl}\,(\Omega)$, so ist $p(x) \leq 1$, also $q(x) = 1$ und daher $g(x) = x$.

Sei nun B eine cl (Ω) und $G(\text{cl}\,(\Omega))$ umfassende Kugel. Da cl (Ω) und $G(\text{cl}\,(\Omega))$ (als relativ kompakte Menge) beschränkt sind existiert eine solche Kugel. Dann ist $g(B) \subset \text{cl}\,(C) \subset B$ und daher ist auch $G \circ g(B) \subset G(\text{cl}\,(\Omega)) \subset B$. Folglich ist $\tilde{G} := G \circ g\colon B \longrightarrow B$ eine stetige Abbildung mit der Eigenschaft, dass $\tilde{G}(B)$ als Teilmenge der relativ kompakten Menge $G(\text{cl}\,(\Omega))$ selbst relativ kompakt ist. Aus dem Schauderschen Fixpunktsatz folgt die Existenz eines $x \in B$ mit $x = \tilde{G}(x)$. Angenommen, es wäre $x \notin \Omega$. Dann wäre $p(x) \geq 1$, folglich $q(x) = p(x)$ und $p(g(x)) = 1$, also $g(x) \in \partial\Omega$. Nach Voraussetzung ist $G(\partial\Omega) \subset \Omega$, also wäre $x = G(g(x)) \in \Omega$, ein Widerspruch. Also ist $x \in \Omega$, $g(x) = x$ und $x = G(x)$. Der Satz ist bewiesen. □

Der nächste Fixpunktsatz stammt von M. [(1957), siehe z. B. auch V. I. Istrătescu (1981, S. 168).

Satz 10.19 (Altman) *Sei X ein linearer normierter Raum, $\Omega \subset X$ offen beschränkt und $0 \in \Omega$. Die Abbildung $G\colon \text{cl}\,(\Omega) \subset X \longrightarrow X$ sei kompakt und es gelte*

$$\|G(x) - x\|^2 \geq \|G(x)\|^2 - \|x\|^2 \quad \textit{für alle } x \in \partial\Omega.$$

Dann besitzt G einen Fixpunkt $x \in \text{cl}\,(\Omega)$.

Beweis: Wie beim Beweis von Satz 10.14 definieren wir $H\colon \text{cl}\,(\Omega) \times [0,1] \longrightarrow X$ durch $H(x,t) := -tG(x)$. Wiederum ist $H(\cdot,t)\colon \text{cl}\,(\Omega) \longrightarrow X$ kompakt für alle $t \in [0,1]$ und bezüglich t auf cl (Ω) gleichmäßig stetig. Wir unterscheiden zwei Fälle.

1. Es existiert ein $x \in \partial\Omega$ mit $0 = x + H(x,1)$ bzw. $x = G(x)$.

   Dann besitzt G einen Fixpunkt auf dem Rand $\partial\Omega$ von Ω.

2. Es ist $0 \neq x + H(x,1)$ für alle $x \in \partial\Omega$.

   Wir wollen zeigen, dass $0 \neq x + H(x,t)$ für alle $(x,t) \in \partial\Omega \times [0,1]$. Angenommen, dies sei nicht der Fall. Dann existiert ein Paar $(x_0,t_0) \in \partial\Omega \times [0,1)$ mit

   $$0 = x_0 + H(x_0,t_0) = x_0 - t_0 G(x_0)$$

   bzw. $t_0 G(x_0) = x_0$. Dann ist notwendig $t_0 \in (0,1)$, da $0 \in \Omega$. Dann ist

   $$\|G(x_0) - x_0\|^2 = \left\| \Big(\frac{1}{t_0} - 1\Big) x_0 \right\|^2 = \Big(\frac{1-t_0}{t_0}\Big)^2 \|x_0\|^2$$

   und

   $$\|G(x_0)\|^2 - \|x_0\|^2 = \frac{1-t_0^2}{t_0^2} \|x_0\|^2.$$

   Nach Voraussetzung ist

   $$\|G(x_0) - x_0\|^2 \geq \|G(x_0)\|^2 - \|x_0\|^2,$$

   also

   $$\frac{(1-t_0)^2}{t_0^2} \|x_0\|^2 \geq \frac{1-t_0^2}{t_0^2} \|x_0\|^2.$$

   Da $x_0 \in \partial\Omega$ ist $x_0 \neq 0$, also $(1-t_0)^2 \geq 1 - t_0^2$, eine Ungleichung, die für $t_0 \in (0,1)$ falsch ist. Insgesamt ist also $0 \neq x + H(x,t)$ für alle $(x,t) \in \partial\Omega \times [0,1]$. Aus Satz 10.11 folgt $d(I,\Omega,0) = d(I - G,\Omega,0)$. Andererseits ist $d(I,\Omega,0) = 1$. Aus Satz 10.12 folgt die Existenz eines Fixpunktes $x \in \Omega$ von G.

In beiden Fällen haben wir die Existenz eines Fixpunktes $x \in \mathrm{cl}\,(\Omega)$ von G nachgewiesen. □

Bemerkung: Angenommen, in Satz 10.19 sei der lineare normierte Raum X sogar ein reeller Prä-Hilbertraum mit dem inneren Produkt $(\cdot,\cdot)$. Was sagt dann die Bedingung

$$\|G(x)-x\|^2 \geq \|G(x)\|^2 - \|x\|^2 \quad \text{für alle } x \in \partial\Omega$$

aus? Wegen

$$\|G(x)-x\|^2 = (G(x)-x, G(x)-x) = \|G(x)\|^2 - 2(G(x),x) + \|x\|^2$$

ist obige Bedingung äquivalent mit

$$(G(x),x) \leq \|x\|^2 \quad \text{für alle } x \in \partial\Omega.$$

□

Der nächste Satz entspricht der Aussage von Satz 9.20.

Satz 10.20 (Leray-Schauder) *Sei X ein linearer normierter Raum, $\Omega \subset X$ offen und beschränkt, $y \in \Omega$. Die Abbildung $G\colon \mathrm{cl}\,(\Omega(\subset X \longrightarrow X$ sei kompakt. Ferner sei $y \neq x - tG(x)$ für alle $(x,t) \in \partial\Omega \times [0,1]$. Dann existiert ein $x \in \mathrm{cl}\,(\Omega)$ mit $y = x - G(x)$.*

Beweis: Die Behauptung folgt offenbar durch eine einfache Kombination von Satz 10.11 und Satz 10.12. □

Als Korollar hierzu erhält man

Satz 10.21 *Sei X ein linearer normierter Raum, $G\colon X \longrightarrow X$ eine kompakte Abbildung und $y \in X$. Es existiere ein $r > 0$ derart, dass*

$$(x,t) \in X \times (0,1),\ y = x - tG(x) \Longrightarrow \|x-y\| < r.$$

Dann besitzt die Gleichung $y = x - G(x)$ wenigstens eine Lösung x mit $\|x-y\| \leq r$.

Beweis: Sei $B(0;r)$ die offene Kugel um den Nullpunkt 0 mit dem Radius r. Nach Voraussetzung ist $y \neq x - tG(x)$ für alle $(x,t) \in \partial B(0;r) \times (0,1)$. Aus Satz 10.20 folgt die Behauptung. □

10.3 Anwendungen der Fixpunkt- und Existenzsätze

In diesem Unterabschnitt wollen wir einige wenige Beispiele für die Anwendung der im letzten Unterabschnitt vorgestellten Fixpunkt- und Existenzsätze geben.

Beispiel: Ein Standardbeispiel für die Anwendung des Schauderschen Fixpunktsatzes ist der Beweis des Existenzsatzes von Peano für Anfangswertaufgaben bei gewöhnlichen Differentialgleichungen. Dieses findet man in fast jedem Buch, in dem der Schaudersche Fixpunktsatz bewiesen wird, also etwa bei L. W. Kantorowitsch, G. P. Akilow (1964, S. 527), R. E. Edwards (1965, S. 164), L. A. Ljusternik, W. I. Sobolew (1968, S. 202) sowie bei E. Zeidler (1986, S. 58) und M. Ružička (2004, S. 28). □

Zum Nachweis der relativen Kompaktheit von Teilmengen eines Raumes stetiger Funktionen spielt der *Satz von Arzela-Ascoli* eine entscheidende Rolle. Wir geben eine verhältnismäßig allgemeine Version dieses Satzes ohne Beweis an:

- *Sei (R, δ) ein kompakter metrischer Raum, (S, Δ) ein vollständiger metrischer Raum und $(C[R, S], d)$ der metrische Raum der von R nach S stetigen Abbildungen mit der Metrik*

$$d(f, g) := \max_{u \in R} \Delta(f(u), g(u)).$$

Sei $K \subset C[R, S]$. Dann ist K in $(C[R, S], d)$ genau dann relativ kompakt, wenn

1. *K gleichgradig stetig ist, wenn es also zu jedem $\epsilon > 0$ ein $\mu(\epsilon) > 0$ mit*

$$u, v \in R,\ f \in K,\ \delta(u, v) \leq \mu(\epsilon) \Longrightarrow \Delta(f(u), f(v)) \leq \epsilon$$

gibt,

2. *$\{f(u) : f \in K\}$ für alle $u \in R$ relativ kompakt in (S, Δ) ist.*

Beispiel: In Satz 9.22 hatten wir gezeigt, dass eine positive $n \times n$-Matrix einen positiven Eigenvektor mit einem zugehörigen positiven Eigenvektor besitzt. In der folgenden Aussage wird dieses Ergebnis auf Integraloperatoren übertragen.

- *Sei $k \in C([a, b] \times [a, b])$ und $k(t, s) > 0$ für alle $(t, s) \in [a, b] \times [a, b]$. Dann besitzt die durch*

$$Kx(t) := \int_a^b k(t, s)x(s)\, ds$$

definierte Abbildung[35] *$K\colon C[a, b] \longrightarrow C[a, b]$ einen positiven Eigenwert mit zugehöriger auf $[a, b]$ stetiger und positiver Eigenfunktion. D. h. es es existiert ein Paar $(x, \lambda) \in C[a, b] \times \mathbb{R}$ mit $x(t) > 0$ für alle $t \in [a, b]$, $\lambda > 0$ und $Kx = \lambda x$ bzw. $Kx(t) = \lambda x(t)$ für alle $t \in [a, b]$.*

Denn: Ähnlich wie in Satz 9.22 der Brouwersche Fixpunktsatz angewandt wurde, wollen wir hier den Schauderschen Fixpunktsatz (Satz 10.16) anwenden. Hierzu setzen wir $X := C[a, b]$, $\| \cdot \| := \| \cdot \|_\infty$ (mit $\|x\|_\infty := \max_{t \in [a,b]} |x(t)|$). Dann ist $(X, \| \cdot \|)$ ein Banachraum. Ferner definiere man auf X die Norm $\|x\|_1 := \int_a^b |x(s)|\, ds$, setze

$$A := \{x \in C[a, b] : x(t) \geq 0 \text{ für alle } t \in [a, b],\ \|x\|_1 = 1\}$$

und definiere $G\colon A \longrightarrow X$ durch

$$G(x) := \frac{1}{\|Kx\|_1} Kx.$$

Für $x \in A$ ist offenbar $\|Kx\|_1 > 0$, sodass diese Definition einen Sinn macht. Offenbar ist $G(A) \subset A$. Ferner ist:

1. A ist konvex und abgeschlossen in X.

   Denn: Trivialerweise ist A eine konvexe Teilmenge von X. Zum Nachweis der Abgeschlossenheit von A beachten wir, dass

$$\|x\|_1 \leq (b - a)\|x\|_\infty \quad \text{für alle } x \in C[a, b].$$

[35] Hier, schon früher, und im folgenden schreiben wir i. Allg. Kx statt $K(x)$, wenn K eine lineare Abbildung ist.

Ist $\{x_k\} \subset A$ und $x_k \to x$ bzw. $x \in C[a,b]$ mit $\|x_k - x\|_\infty \to 0$, so ist natürlich auch x eine auf $[a,b]$ nichtnegative stetige Funktion. Ferner ist

$$|1 - \|x\|_1| = |\|x_k\|_1 - \|x\|_1| \leq \|x_k - x\|_1 \leq (b-a)\|x_k - x\|_\infty \to 0$$

und damit auch $\|x\|_1 = 1$ bzw. $x \in A$.

2. $G(A)$ ist relativ kompakt.

   Denn: Wir wenden den Satz von Arzela-Ascoli (mit $R := [a,b] \subset \mathbb{R}$ und $S := \mathbb{R}$ jeweils mit der durch den Betrag gegebenen Metrik) an und zeigen, dass $G(A)$ beschränkt und gleichgradig stetig ist. Da $k \in C([a,b] \times [a,b])$ und $k(t,s) > 0$ für $(t,s) \in [a,b] \times [a,b]$ existieren positive Konstanten $m < M$ mit

   $$m \leq k(t,s) \leq M \quad \text{für } (t,s) \in [a,b] \times [a,b].$$

   (a) $G(A)$ ist beschränkt.
   Denn: Ist $x \in A$, so ist

   $$|Kx(t)| = \int_a^b k(t,s)x(s)\,ds \leq M \int_a^b x(s)\,ds = M \quad \text{für alle } t \in [a,b],$$

   also $\|Kx\|_\infty \leq M$. Andererseits ist

   $$\|Kx\|_1 = \int_a^b \int_a^b k(t,s)x(s)\,ds\,dt \geq m(b-a)$$

   und damit

   $$\|G(x)\|_\infty = \frac{\|Kx\|_\infty}{\|Kx\|_1} \leq \frac{M}{m(b-a)} \quad \text{für alle } x \in A.$$

   Daher ist $G(A)$ beschränkt.

   (b) $G(A)$ ist gleichgradig stetig.
   Denn: k ist auf $[a,b] \times [a,b]$ gleichmäßig stetig. Daher existiert zu $\epsilon > 0$ ein $\delta > 0$ derart, dass

   $$|k(t_1,s) - k(t_2,s)| \leq \epsilon m(b-a) \quad \text{falls } |t_1 - t_2| \leq \delta,\ s \in [a,b].$$

   Für alle $t_1, t_2 \in [a,b]$ mit $|t_1 - t_2| \leq \delta$ und beliebiges $x \in A$ ist dann

   $$\begin{aligned} |G(x)(t_1) - G(x)(t_2)| &\leq \frac{1}{\|Kx\|_1} \int_a^b |k(t_1,s) - k(t_2,s)|x(s)\,ds \\ &\leq \frac{1}{m(b-a)} \epsilon m(b-a) \int_a^b x(s)\,ds \\ &= \epsilon, \end{aligned}$$

   womit die gleichgradige Stetigkeit von $G(A)$ bewiesen ist.

3. $G\colon A \longrightarrow X$ ist stetig.

   Denn: Sei $\{x_n\} \subset A$ und $\lim_{n\to\infty} \|x_n - x\|_\infty = 0$. Da $K\colon C[a,b] \longrightarrow C[a,b]$ (trivialerweise) stetig ist, gilt $\lim_{n\to\infty} \|Kx_n - Kx\|_\infty = 0$. Ferner gilt

$$\lim_{n\to\infty} \|Kx_n\|_1 = \|Kx\|_1,$$

   da

$$|\|Kx_n\|_1 - \|Kx\|_1| \le \|K(x_n - x)\|_1 \le (b-a)\|Kx_n - Kx\|_\infty \to 0.$$

   Damit ist die Stetigkeit von G auf A nachgewiesen.

Damit sind alle Voraussetzungen von Satz 10.16, dem Schauderschen Fixpunktsatz, nachgewiesen. Es folgt die Existenz von $x \in A$ mit $G(x) = x$ bzw. $Kx = \|Kx\|_1\, x$. Dann ist x eine Eigenfunktion von K mit dem zugehörigen Eigenwert $\lambda := \|K(x)\|_1$. Wegen

$$M(b-a) \ge \|Kx\|_1 = \lambda \ge m(b-a) > 0$$

und

$$x(t) = \frac{Kx(t)}{\|Kx\|_1} \ge \frac{m}{M(b-a)} > 0 \quad \text{für alle } t \in [a,b]$$

ist (x, λ) ein Paar der gesuchten Art.

Eine starke Einschränkung in der letzten Aussage ist die Voraussetzung, dass der Kern k von K auf $[a,b] \times [a,b]$ positiv ist. Genügt es nicht, dass der Kern k nichtnegativ auf $[a,b] \times [a,b]$ ist, wobei natürlich ausgeschlossen werden muss, dass k identisch verschwindet? Die folgende Aussage[36] gibt hierauf eine Antwort.

- *Sei $k \in C([a,b]\times[a,b])$ und $k(t,s) \ge 0$ für alle $(t,s) \in [a,b]\times[a,b]$. Die Abbildung $K\colon C[a,b] \longrightarrow C[a,b]$ sei definiert durch*

$$Kx(t) := \int_a^b k(t,s)x(s)\,ds.$$

  Es existiere $u \in C[a,b] \setminus \{0\}$ mit $u(t) \ge 0$ für $t \in [a,b]$ und eine Zahl $\alpha > 0$ mit $Ku(t) \ge \alpha u(t)$ für alle $t \in [a,b]$. Dann besitzt K wenigstens einen (positiven) Eigenwert $\lambda \ge \alpha$ mit zugehöriger nichtnegativer Eigenfunktion $x \in C[a,b]$.

Denn: Man definiere die Menge

$$A := \{x \in C[a,b] : x \ge 0,\ \|x\|_\infty \le 1\}.$$

Hierbei bedeute $x \ge 0$ für $x \in C[a,b]$, dass $x(t) \ge 0$ für alle $t \in [a,b]$. Offenbar ist A nichtleer, abgeschlossen, konvex und beschränkt. Für $n \in \mathbb{N}$ definiere man die Abbildung $G_n\colon A \longrightarrow C[a,b]$ durch

$$G_n(x) := \frac{K(x + u/n)}{\|K(x+u/n)\|_\infty}.$$

Dann gilt:

[36] Siehe M. A. Krasnoselskii (1964, S. 67).

1. $G_n(A) \subset A$.

2. $G_n(A)$ ist relativ kompakt in $(C[a,b], \|\cdot\|_\infty)$.

   Denn: Natürlich wendet man wieder den Satz von Arzela-Ascoli an. Trivialerweise ist $G_n(A)$ beschränkt. Für $x \in A$ ist

   $$K(x + u/n)(t) \geq \frac{1}{n} Ku(t) \geq \frac{\alpha}{n} u(t) \quad \text{für alle } t \in [a,b]$$

   und daher

   $$0 < \frac{\alpha}{n} \|u\|_\infty \leq \|K(x + u/n)\|_\infty.$$

   Zum Nachweis der gleichgradigen Stetigkeit von $G_n(A)$ geben wir uns ein $\epsilon > 0$ vor. Da der Kern k von K auf $[a,b] \times [a,b]$ gleichmäßig stetig ist, existiert ein $\delta > 0$ mit

   $$|k(t_1, s) - k(t_2, s)| \leq \epsilon \alpha c \quad \text{falls } |t_1 - t_2| \leq \delta,\ s, t_1, t_2 \in [a,b],$$

   wobei $c > 0$ eine noch geeignet zu wählende Konstante ist. Für alle $t_1, t_2 \in [a,b]$ mit $|t_1 - t_2| \leq \delta$ und beliebiges $x \in A$ ist dann

   $$\begin{aligned} & |G_n(x)(t_1) - G_n(x)(t_2)| \\ \leq\ & \frac{1}{\|K_n(x + u/n)\|_\infty} \int_a^b |k(t_1, s) - k(t_2, s)| (x(s) + u(s)/n)\, ds \\ \leq\ & \frac{n\epsilon\alpha c}{\alpha \|u\|_\infty} \int_a^b (x(s) + u(s)/n)\, ds \\ \leq\ & \frac{n\epsilon c}{\|u\|_\infty} (b-a)(1 + \|u\|_\infty/n). \end{aligned}$$

   Wählt man daher

   $$c := \frac{\|u\|_\infty}{(b-a)(n + \|u\|_\infty)},$$

   so ist

   $$|G_n(x)(t_1) - G_n(x)(t_2)| \leq \epsilon \quad \text{falls } t_1, t_2 \in [a,b] \text{ mit } |t_1 - t_2| \leq \delta.$$

   Damit ist auch die gleichgradige Stetigkeit von $G_n(A)$ und wegen des Satzes von Arzela-Ascoli die relative Kompaktheit von $G_n(A)$ bewiesen.

3. $G_n \colon A \longrightarrow C[a,b]$ ist stetig.

   Denn: Die Stetigkeit von G_n auf A folgt aus der Stetigkeit der Abbildung K.

Aus dem Schauderschen Fixpunktsatz folgt für jedes $n \in \mathbb{N}$ die Existenz von $x_n \in A$ mit $G_n(x_n) = x_n$. Dann ist offenbar $\|x_n\|_\infty = 1$. Definiert man $\lambda_n := \|K(x + u/n)\|_\infty$, so gilt[37] also

$$K(x_n + u/n) = \lambda_n x_n.$$

[37] Wegen $K(x + u/n) \neq 0$ ist $\lambda_n > 0$.

Die Abbildung K ist kompakt auf ganz $C[a,b]$, bildet also jede beschränkte Teilmenge von $C[a,b]$ in eine relativ kompakte Menge ab. Die Folge $\{x_n+u/n\}_{n\in\mathbb{N}}$ ist beschränkt. Folglich kann aus $\{K(x_n+u/n)\}_{n\in\mathbb{N}}$ eine konvergente Teilfolge $\{K(x_{n_i}+u/n_i)\}_{i\in\mathbb{N}}$ ausgewählt werden, sei etwa

$$y = \lim_{i\to\infty} K(x_{n_i}+u/n_i).$$

Hieraus folgt

$$\lambda_{n_i} = \|K(x_{n_i}+u/n_i)\|_\infty \to \|y\|_\infty.$$

Angenommen, wir wüssten schon, dass $y \neq 0$. Wegen

$$x_{n_i} = \frac{1}{\lambda_{n_i}} K(x_{n_i}+u/n_i) \to x := \frac{y}{\|y\|_\infty}$$

und

$$Kx \leftarrow K(x_{n_i}+u/n_i) = \lambda_{n_i} x_{n_i} \to \|y\|_\infty\, x$$

ist $Kx = \lambda x$ mit $\lambda := \|y\|_\infty$. Die behauptete Aussage wäre also bewiesen. Zu zeigen bleibt, dass $y \neq 0$. Aus

$$x_n = \frac{1}{\lambda_n} K(x_n+u/n) \geq \frac{1}{\lambda_n n} Ku \geq \frac{\alpha}{\lambda_n n} u$$

schließt man wegen $u \neq 0$ auf die Existenz eines maximalen $\beta_n > 0$ mit $x_n \geq \beta_n u$. Dann ist

$$x_n = \frac{1}{\lambda_n} K(\underbrace{x_n}_{\geq \beta_n u} + u/n) \geq \frac{1}{\lambda_n}(\beta_n + 1/n) Ku \geq \frac{\alpha}{\lambda_n}(\beta_n+1/n)u,$$

wegen der Maximalität von β_n ist also

$$\frac{\alpha}{\lambda_n}(\beta_n+1/n) \leq \beta_n$$

und folglich

$$\lambda_n \geq \alpha + \frac{\alpha}{n\beta_n} > \alpha.$$

Insbesondere ist $\lambda_{n_i} > \alpha$, $i \in \mathbb{N}$, und daher $\|y\| \geq \alpha > 0$ und folglich $y \neq 0$. Daher ist obige Aussage bewiesen. □

Beispiel: Die Idee des *Galerkin-Verfahrens* zur Lösung einer Fixpunktaufgabe

$$x = G(x)$$

besteht in folgendem: Sei $L_1 \subset L_2 \subset \cdots \subset L_n \subset \cdots X$ eine aufsteigende Folge endlichdimensionaler linearer Teilräume eines linearen normierten Raumes X und $\{P_n\}$ eine Folge linearer *Projektionsoperatoren* $P_n\colon X \longrightarrow L_n$, d. h. es sei $P_n^2 = P_n$. Das unendlichdimensionale Problem $x = G(x)$ ersetze man durch die Folge der endlichdimensionalen Probleme $x = P_nG(x)$, $n \in \mathbb{N}$. Wir stellen uns die beiden folgenden Fragen: Angenommen, $x = G(x)$ besitze eine Lösung x.

1. Unter welchen Bedingungen hat das endlichdimensionale Problem $x = P_nG(x)$ wenigstens für alle hinreichend großen n eine Lösung $x_n \in L_n$?

2. Angenommen 1. kann positiv beantwortet werden. Unter welchen Bedingungen kann $x_n \to x$ gezeigt werden?

Hierzu beweisen wir die folgende Aussage, M. A. Krasnoselskii (1964, S. 169 ff.):

- *Sei X ein Banachraum und $G\colon D(G) \subset X \longrightarrow X$ eine kompakte Abbildung. Die Gleichung $x = G(x)$ besitze eine Lösung $x^* \in \operatorname{int}(D(G))$, in der G Fréchet-differenzierbar und 1 kein Eigenwert von $G'(x^*)$ ist (sodass $x = G'(x^*)x$ nur die triviale Lösung $x = 0$ besitzt). Für $n \in \mathbb{N}$ sei $P_n\colon X \longrightarrow L_n \subset X$ mit endlichdimensionalem linearen Teilraum L_n eine lineare, stetige Abbildung. Schließlich konvergiere die Folge $\{P_n\}$ stark gegen die Identität, d. h. es gelte*

$$\lim_{n\to\infty} \|P_nx - x\| = 0 \quad \text{für alle } x \in X.$$

Dann existieren $\sigma > 0$ und $n_0 \in \mathbb{N}$ derart, dass die Gleichung $x = P_nG(x)$ für alle $n \geq n_0$ eine Lösung $x_n \in L_n \cap B[x^*;\sigma]$ besitzt, wobei $B[x^*;\sigma] := \{x \in X : \|x - x^*\| \leq \sigma\} \subset D(G)$.

Der Beweis erfolgt mit Hilfe der folgenden Schritte:

(a) Aus der Voraussetzung, dass 1 kein Eigenwert von $G'(x^*)$ ist, schließe man, dass x^* eine *isolierte* Lösung von $x = G(x)$ ist, d. h. dass es eine abgeschlossene Kugel $B[x^*;\sigma] := \{x \in X : \|x - x^*\| \leq \sigma\}$ gibt, in der $x = G(x)$ keine weitere Lösung besitzt.

(b) Sei $B(x^*;\sigma) := \operatorname{int}(B[x^*;\sigma])$ die offene Kugel um x^* mit dem Radius $\sigma > 0$. Wegen Definition 10.8 bzw. Satz 10.9 ist der Abbildungsgrad $d(I-G, B(x^*;\sigma), 0)$ wohldefiniert. Man zeige, dass $d(I - G, B(x^*;\sigma), 0) \neq 0$.

(c) Die Abbildung $P_nG\colon D(G) \longrightarrow L_n \subset X$ ist (endlichdimensional und) kompakt[38]. Man zeige, dass

$$\lim_{n\to\infty} \sup_{x\in B[x^*;\sigma]} \|G(x) - P_nG(x)\| = 0$$

und schließe mit Satz 10.10, dass $d(I - P_nG, B(x^*;\sigma), 0)$ für alle hinreichend großen n sinnvoll ist und $d(I - G, B(x^*;\sigma), 0) = d(I - P_nG, B(x^*;\sigma), 0)$ gilt.

(d) Wegen (b) ist $d(I - P_nG, B(x^*;\sigma), 0) \neq 0$ für alle hinreichend großen n.

(e) Aus Satz 10.12 folgt, dass $x = P_nG(x)$ für alle hinreichend großen n eine Lösung $x_n \in L_n \cap B(x^*;\sigma)$ besitzt.

Der Beweis von (a) erfolgt in zwei Schritten. Im ersten Schritt überlegen wir uns die Gültigkeit der folgenden Aussage:

[38] Hierzu ist zu zeigen, dass P_nG eine beschränkte Teilmengen von $D(G)$ in eine relativ kompakte Teilmenge von X abbildet. Dies folgt aber sofort aus der Stetigkeit von P_n und der Kompaktheit von G.

- Seien X, Y lineare normierte Räume, $G\colon D(G) \subset X \longrightarrow Y$ kompakt und in $x^* \in \text{int}\,(D(G))$ Fréchet-differenzierbar. Dann ist $G'(x^*)\colon X \longrightarrow Y$ kompakt.

Denn: Angenommen, $G'(x^*)$ wäre nicht kompakt. Dann wäre $G'(x^*)(B)$ mit der Einheitskugel $B := \{x \in X : \|x\| \leq 1\}$ nicht relativ kompakt. Dies bedeutet, dass eine Zahl $\delta > 0$ und eine Folge $\{h_i\} \subset B$ gefunden werden können mit $\|G'(x^*)(h_i - h_j)\| \geq \delta$ für $i \neq j$ (denn dies bedeutet gerade, dass aus $\{G'(x^*)h_i\}$ keine konvergente Teilfolge ausgewählt werden kann. Nach Definition der Fréchet-Differenzierbarkeit von G in x^* existiert zu δ ein $\rho > 0$ mit

$$\|G(x^* + h) - G(x^*) - G'(x^*)h\| \leq \frac{\delta}{3}\|h\| \quad \text{für alle } h \text{ mit } \|h\| \leq \rho.$$

Hieraus folgt aber, dass

$$\begin{aligned}
\|G(x^* + \rho h_i) - G(x^* + \rho h_j)\| &= \|\rho G'(x^*)(h_i - h_j) \\
&\qquad + [G(x^* + \rho h_i) - G(x^*) - \rho G'(x^*)h_i] \\
&\qquad - [G(x^* + \rho h_j) - G(x^*) - \rho G'(x^*)h_j]\| \\
&\geq \rho\,\|G'(x^*)(h_i - h_j)\| \\
&\qquad - \|G(x^* + \rho h_i) - G(x^*) - \rho G'(x^*)h_i\| \\
&\qquad - \|G(x^* + \rho h_j) - G(x^*) - \rho G'(x^*)h_j\| \\
&\geq \rho\delta - \frac{\rho\delta}{3} - \frac{\rho\delta}{3} \\
&= \frac{\rho\delta}{3}
\end{aligned}$$

für $i \neq j$. Dies bedeutet aber, dass aus $\{G(x^* + \rho h_i)\}$ keine konvergente Teilfolge ausgewählt werden kann, ein Widerspruch zur Kompaktheit von G.

Nun folgt der Beweis von (a). Genauer zeigen wir:

- Sei X ein linearer normierter Raum, $G\colon D(G) \subset X \longrightarrow X$ eine kompakte Abbildung und $x^* \in \text{int}\,(D(G))$ eine Lösung von $x = G(x)$. Sei G in x^* Fréchet-differenzierbar und 1 kein Eigenwert von $G'(x^*)$. Dann existiert ein $\sigma > 0$ bzw. eine Kugel $B[x^*; \sigma] := \{x \in X : \|x - x^*\| \leq \sigma\}$ derart, dass $x = G(x)$ in $B[x^*; \sigma]$ nur die Lösung x^* besitzt.

Denn: Nach dem schon bewiesenen Hilfsergebnis ist mit G auch $G'(x^*)$ eine kompakte Abbildung. Hieraus und aus der Voraussetzung, dass 1 kein Eigenwert von $G'(x^*)$ ist folgt, dass $(I - G'(x^*))^{-1}$ auf X existiert und beschränkt ist. Es gilt nämlich (Stichwort: Fredholmsche Alternative):

- Sei X ein linearer normierter Raum und $T\colon X \longrightarrow X$ eine lineare, kompakte Abbildung mit der Eigenschaft, dass 1 kein Eigenwert von T ist. Dann gilt:
  - (i) Bild $(I - T) \subset X$ ist abgeschlossen.
  - (ii) Bild $(I - T) = X$.
  - (iii) $(I - T)^{-1}$ existiert auf X.

(iv) $(I-T)^{-1}: X \longrightarrow X$ ist stetig.

Denn: Zum Beweis von (i) nehmen wir an, es sei $\{y_k\} \subset \text{Bild}\,(I-T)$ eine Folge mit $y_k \to y$. Zu zeigen ist $y \in \text{Bild}\,(I-T)$. Es sei $y_k = (I-T)x_k$. Wir zeigen zunächst, dass die Folge $\{x_k\} \subset X$ beschränkt ist. Angenommen, das sei nicht der Fall. Dann existiert eine Teilfolge $\{x_{k_i}\} \subset \{x_k\}$ mit $\|x_{k_i}\| \to \infty$. Wir definieren die Folge $\{z_{k_i}\}$ durch $z_{k_i} := x_{k_i}/\|x_{k_i}\|$. Dann gilt

$$(I-T)z_{k_i} = \frac{(I-T)x_{k_i}}{\|x_{k_i}\|} = \frac{y_{k_i}}{\|x_{k_i}\|} \to 0.$$

Wegen der Kompaktheit von T kann aus der Folge $\{Tz_{k_i}\}$ eine konvergente Teilfolge ausgewählt werden. Um Schreibarbeit zu sparen, nehmen wir an, die Folge $\{Tz_{k_i}\}$ sei selbst schon konvergent, etwa gegen ein w. Dann gilt

$$z_{k_i} = \underbrace{(I-T)z_{k_i}}_{\to 0} + \underbrace{Tz_{k_i}}_{\to w} \to w.$$

Wegen $\|z_{k_i}\| = 1$ ist auch $\|w\| = 1$. Wegen der Stetigkeit von $(I-T)$ ist $(I-T)w = 0$, also 1 ein Eigenwert von T, was in der Voraussetzung ausgeschlossen wurde. Also ist die Folge $\{x_k\}$ beschränkt. Da T kompakt ist, besitzt $\{Tx_k\}$ eine konvergente Teilfolge $\{Tx_{k_i}\}$, sei etwa $Tx_{k_i} \longrightarrow z$. Wegen $x_{k_i} = y_{k_i} + Tx_{k_i} \to y + z$ ist $z = T(y+z)$ und daher $y = (I-T)(y+z) \in \text{Bild}\,(I-T)$. Damit ist (i) bzw. die Abgeschlossenheit von $(I-T)$ nachgewiesen.

Zum Nachweis von (ii) nehmen wir an, es sei $\text{Bild}\,(I-T) \neq X$ ein echter Teilraum von X. Man definiere $X_k := \text{Bild}\,((I-T)^k)$, $k = 0, 1, \ldots$. Da $(I-T)^k = I - \tilde{T}$ mit einer linearen, kompakten Abbildung $\tilde{T}: X \longrightarrow X$, ist X_k wegen (i) ein abgeschlossener linearer Teilraum von X. Durch vollständige Induktion nach k überlegen wir uns, dass X_{k+1} ein *echter* (abgeschlossener) linearer Teilraum von X_k ist, $k = 0, 1, \ldots$. Der Induktionsanfang liegt bei $k = 0$. Denn nach Annahme ist $X_1 = \text{Bild}\,(I-T)$ ein echter Teilraum von $X_0 = X$. Wir nehmen an, die Aussage sei für $k-1$ richtig, es sei also $X_k = \text{Bild}\,((I-T)^k)$ ein echter Teilraum von $X_{k-1} = \text{Bild}\,((I-T)^{k-1})$. Dann existiert ein $x \in X$ mit $(I-T)^{k-1}x \neq (I-T)^k y$ für alle $y \in X$. Trivialerweise ist $X_{k+1} \subset X_k$. Wir zeigen die Existenz eines Elementes in X_k, welches *nicht* in X_{k+1} liegt. Nun ist $(I-T)^k x \notin X_{k+1}$. Denn andernfalls existiert ein $y \in X$ mit $(I-T)^k x = (I-T)^{k+1} y$. Dies wiederum impliziert

$$(I-T)[\underbrace{(I-T)^{k-1}x - (I-T)^k y}_{\neq 0}] = 0,$$

ein Widerspruch dazu, dass 1 kein Eigenwert von T ist. Wegen des *Lemmas von Riesz* existiert ein $y_k \in X_k$ mit $\|y_k\| = 1$ und $d(y_k, X_{k+1}) \geq \frac{1}{2}$. Denn sei $w \in X_k \setminus X_{k+1}$ (beachte: X_{k+1} ist ein *echter* Teilraum von X_k). Da X_{k+1} abgeschlossen ist, hat w einen positiven Abstand $d(w, X_{k+1})$ zu X_{k+1}. Sei $z \in X_{k+1}$ ein Punkt mit $\|w - z\| \leq 2d(w, X_{k+1})$ und setze

$$y_k := \frac{w-z}{\|w-z\|}.$$

Dann ist $y_k \in X_k$, $\|y_k\| = 1$ und

$$d(y_k, X_{k+1}) \geq \frac{d(w, X_{k+1})}{\|w-z\|} \geq \frac{1}{2}.$$

Jetzt bekommen wir sehr schnell den gewünschten Widerspruch. Denn für $l > k$ ist

$$\begin{aligned} \|Ty_k - Ty_l\| &= \|y_k - \underbrace{(y_l + (I-T)y_k - (I-T)y_l)}_{\in X_{k+1}}\| \\ &\geq d(y_k, X_{k+1}) \\ &\geq \frac{1}{2}. \end{aligned}$$

Dies steht im Widerspruch dazu, dass aus $\{Ty_k\}$ eine konvergente Teilfolge ausgewählt werden kann. Damit ist schließlich (ii) bewiesen.

Wegen (i) und (ii) ist klar, dass $(I-T)^{-1}$ auf X existiert bzw. (iii) gilt. Zu zeigen bleibt, dass $(I-T)^{-1}\colon X \longrightarrow X$ stetig bzw. beschränkt ist. Hierzu zeigen wir zunächst, dass $c := \inf_{x\in X:\|x\|=1} \|(I-T)x\| > 0$. Angenommen, dies sei nicht der Fall. Dann existiert eine Folge $\{x_k\} \subset X$ mit $\|x_k\| = 1$ und $(I-T)x_k \to 0$. Da $T\colon X \longrightarrow X$ kompakt ist, existiert eine gegen ein $w \in X$ konvergente Teilfolge $\{Tx_{k_i}\} \subset \{Tx_k\}$. Wegen

$$x_{k_i} = \underbrace{(I-T)x_{k_i}}_{\to 0} + \underbrace{Tx_{k_i}}_{\to w} \to w$$

und $\|x_{k_i}\| = 1$ ist $\|w\| = 1$ und insbesondere $w \neq 0$. Aus $(I-T)x_{k_i} \to 0$ und der Stetigkeit von $(I-T)$ folgt $(I-T)w = 0$ bzw. $Tw = w$, ein Widerspruch dazu, dass 1 kein Eigenwert von T ist. Damit ist $c := \inf_{x\in X:\|x\|=1} \|(I-T)x\| > 0$ bewiesen. Folglich ist

$$\|(I-T)x\| \geq c\,\|x\| \quad \text{für alle } x \in X.$$

Hieraus erhält man sehr leicht die Stetigkeit von $(I-T)^{-1}$. Denn für ein beliebiges $y \in X$ und $x := (I-T)^{-1}y$ ist

$$\|(I-T)^{-1}y\| = \|x\| \leq \frac{1}{c}\,\|(I-T)x\| = \frac{1}{c}\,\|y\|,$$

womit die Beschränktheit bzw. Stetigkeit von $(I-T)^{-1}$ bewiesen ist.

Nun kommen wir zum Beweis von (a). Wegen der Stetigkeit von $(I-G'(x^*))^{-1}\colon X \longrightarrow X$ existiert $a > 0$ mit

$$\|(I-G'(x^*))h\| \geq a\,\|h\| \quad \text{für alle } h \in X.$$

Da G in x^* Fréchet-differenzierbar ist, existiert ein $\sigma > 0$ mit

$$\|h\| \leq \sigma \Longrightarrow \|G(x^*+h) - G(x^*) - G'(x^*)h\| \leq \frac{a}{2}\,\|h\|.$$

Hierbei sei σ so klein gewählt, dass $B[x^*;\sigma] \subset D(G)$, außerdem können wir o. B. d. A. annehmen, dass $\sigma \leq 1$. Für $\|h\| \leq \sigma$ ist dann

$$\begin{aligned}
\|(x^*+h) - G(x^*+h)\| &= \|\underbrace{x^*}_{=G(x^*)} + (I-G'(x^*))h - [G(x^*+h) - G'(x^*)h]\| \\
&= \|(I-G'(x^*))h - [G(x^*+h) - G(x^*) - G'(x^*)h]\| \\
&\geq \|(I-G'(x^*))h\| - \|G(x^*+h) - G(x^*) - G'(x^*)h\| \\
&\geq \frac{a}{2}\|h\|,
\end{aligned}$$

woraus die Aussage (a) folgt.

Nun kommen wir zum Beweis von (b). Wegen Definition 10.8 bzw. Satz 10.9 ist der Abbildungsgrad $d(I-G, B(x^*;\sigma), 0)$ wohldefiniert, wenn G (bzw. $-G$) eine kompakte Abbildung von cl $(B(x^*;\sigma)) = B[x^*;\sigma] \subset X$ in den linearen normierten Raum X ist und $0 \notin (I-G)(\partial B(x^*;\sigma))$ ist. Dies ist nach Voraussetzung oder wegen (a) (denn $I-G$ hat keine Nullstelle in $B[x^*;\sigma]$ außer x^*, insbesondere keine auf dem Rand von $B(x^*;\sigma)$) erfüllt. Für die Aussage (b) bleibt zu zeigen, dass $d(I-G, B(x^*;\sigma), 0) \neq 0$. Hierzu beweisen wir in (i) zunächst, dass $d(I-G'(x^*), B(x^*;\sigma), x^* - G'(x^*)x^*)) \neq 0$ und anschließend in (ii) mit Hilfe des Homotopiesatzes 10.11, dass

$$d(I-G, B(x^*;\sigma), 0) = d(I-G'(x^*), B(x^*;\sigma), x^* - G'(x^*)x^*)),$$

womit dann (b) nachgewiesen sein wird. In (i) zeigen wir sogar, dass

$$d(I - G'(x^*), B(x^*; \sigma), x^* - G'(x^*)x^*) = \pm 1.$$

Hierzu überlegen wir uns die Gültigkeit der folgenden Aussage:

- *Sei X ein Banachraum, $\{P_n\}$ eine Folge linearer, stetiger Abbildungen $P_n \colon X \longrightarrow X$ mit $\lim_{n\to\infty} P_n x = x$ für alle $x \in X$. Sei $K \colon X \longrightarrow X$ eine lineare kompakte Abbildung. Dann gilt*

  $$\lim_{n\to\infty} \|P_n K - K\| = 0,$$

  d. h. $\{P_n K\}$ konvergiert (gleichmäßig) gegen K.

Denn: Aus dem Prinzip der gleichmäßigen Beschränktheit (die punktweise Beschränktheit einer Folge linearer stetiger Abbildungen von einem Banachraum in einen linearen normierten Raum impliziert deren gleichmäßige Beschränktheit) folgt die Existenz einer Konstanten $M > 0$ mit $\|P_n\| \leq M$ für alle $n \in \mathbb{N}$. Angenommen, es wäre $\|P_n K - K\| \not\to 0$. Dann gibt es eine Teilfolge $\{n_k\} \subset \mathbb{N}$ und ein $\epsilon > 0$ mit

$$\|P_{n_k} K - K\| = \sup_{\|x\|\leq 1} \|(P_{n_k} K - K)x\| \geq \epsilon \quad \text{für alle } k \in \mathbb{N}.$$

Folglich existiert eine Folge $\{x_{n_k}\} \subset X$ mit $\|x_{n_k}\| \leq 1$ und $\|(P_{n_k} - I)Kx_{n_k}\| \geq \frac{1}{2}\epsilon$ für alle $k \in \mathbb{N}$. Da K kompakt ist, besitzt die Folge $\{Kx_{n_k}\}$ eine konvergente Teilfolge. O. B. d. A. ist schon $\{Kx_{n_k}\}$ konvergent, etwa gegen ein $\hat{x} \in X$. Dann gilt

$$\begin{aligned}
\|P_{n_k}\hat{x} - \hat{x}\| &= \|(P_{n_k} - I)\hat{x}\| \\
&= \|(P_{n_k} - I)Kx_{n_k} - (P_{n_k} - I)(Kx_{n_k} - \hat{x})\| \\
&\geq \|(P_{n_k} - I)Kx_{n_k}\| - \|(P_{n_k} - I)(Kx_{n_k} - \hat{x})\| \\
&\geq \frac{\epsilon}{2} - (M+1)\|Kx_{n_k} - \hat{x}\| \\
&\geq \frac{\epsilon}{4}
\end{aligned}$$

für alle hinreichend großen k. Dies ist ein Widerspruch zu $P_{n_k}\hat{x} \to \hat{x}$. Damit ist obige Aussage bewiesen.
Nun definieren wir

$$\gamma := \inf\{\|(I - G'(x^*))x\| : \|x - x^*\| = \sigma\}.$$

Dann ist $\gamma > 0$, siehe den ersten Teil von Satz 10.10. Nun wähle man $\epsilon \in (0, \gamma)$ und anschließend $n \in \mathbb{N}$ so groß, dass $\|G'(x^*) - P_n G'(x^*)\| \leq \epsilon$, was wegen obiger Hilfsaussage möglich ist. Wegen Satz 10.10 ist

$$d(I - G'(x^*), B(x^*; \sigma), x^* - G'(x^*)x^*) = d(I - P_n G'(x^*), B(x^*; \sigma, x^* - G'(x^*)x^*),$$

weiter existiert $(I - P_n G'(x^*))^{-1}: L_n \longrightarrow L_n$. Denn es ist $(I - P_n G'(x^*))z \neq 0$ für jedes $z \neq 0$. Um dies einzusehen, können wir $\|z\| = \sigma$ annehmen. Dann ist

$$\begin{aligned}
\|(I - P_n G'(x^*))z\| &= \|(I - G'(x^*))z + (G'(x^*) - P_n G'(x^*))z\| \\
&\geq \underbrace{\|(I - G'(x^*))z\|}_{\geq \gamma} - \|(G'(x^*) - P_n G'(x^*))z\| \\
&\geq \gamma - \underbrace{\|G'(x^*) - P_n G'(x^*)\|}_{\leq \epsilon} \|z\| \\
&\geq \gamma - \epsilon\sigma \\
&\geq \gamma - \epsilon \qquad (\text{wegen } \sigma \in (0,1]) \\
&> 0 \qquad (\text{wegen } \epsilon \in (0,\gamma),
\end{aligned}$$

woraus wir die Behauptung ableiten. Nun ist aber klar, dass

$$d(I - P_n G'(x^*), B(x^*;\sigma), x^* - G'(x^*)x^*) = \pm 1,$$

da der Abbildungsgrad in diesem Falle gleich dem Vorzeichen der Determinante von $I - P_n G'(x^*)$ bzw. der diese Abbildung darstellenden nichtsingulären Matrix ist. Damit haben wir

$$d(I - G'(x^*), B(x^*;\sigma), x^* - G'(x^*)x^*) = d(I - P_n G'(x^*), B(x^*;\sigma), x^* - G'(x^*)x^*) = \pm 1$$

bewiesen. Jetzt zeigen wir noch, dass

$$d(I - G, B(x^*;\sigma), 0) = d(I - G'(x^*), B(x^*;\sigma), x^* - G'(x^*)x^*),$$

womit dann schließlich die Aussage (b) vollständig bewiesen sein wird. Wegen

$$d(I - G'(x^*), B(x^*;\sigma), x^* - G'(x^*)x^*)) = d(I - G'(x^*) - [x^* - G'(x^*)x^*], B(x^*;\sigma), 0)$$

(siehe Satz 9.12, der sich offenbar auf den unendlichdimensionalen Fall überträgt) genügt es zu zeigen[39], dass

$$d(I - G, B(x^*;\sigma), 0) = d(I - G'(x^*) - [x^* - G'(x^*)x^*], B(x^*;\sigma), 0).$$

Hierzu definiere man die G und $G'(x^*) + (x^* - G'(x^*)x^*)$ verbindende Homotopie $H: B[x^*;\sigma] \times [0,1] \longrightarrow X$ durch

$$H(x,t) := -tG(x) - (1-t)[G'(x^*)(x) + (I - G'(x^*))x^*].$$

[39]Man beachte, dass

$$G(x) \approx \underbrace{G(x^*)}_{=x^*} + G'(x^*)(x - x^*) = G'(x^*)x + x^* - G'(x^*)x^*$$

bzw. $G \approx G'(x^*) + x^* - G'(x^*)x^*$.

Klar ist, dass $H(\cdot,t)$ kompakt und bezüglich t gleichmäßig stetig ist. Angenommen, es existiert $(x,t) \in \partial B(x^*;\sigma) \times [0,1]$ mit $x + H(x,t) = 0$. Mit $h := x - x^*$ ist dann

$$\begin{aligned} 0 &= x + H(x,t) \\ &= \underbrace{x^*}_{=G(x^*)} + h - tG(x^*+h) - (1-t)[G'(x^*)(x^*+h) + (I - G'(x^*))x^* \\ &= G(x^*) - tG(x^*+h) - (1-t)[G'(x^*)h + G(x^*)] \\ &= -t[G(x^*+h) - G(x^*) - G'(x^*)h] + (I - G'(x^*)h. \end{aligned}$$

Wegen $\|h\| = \sigma$ wird dann

$$\begin{aligned} 0 &= \|x + H(x,t)\| \\ &\geq \|(I - G'(x^*))\| - t\|G(x^*+h) - G(x^*) - G'(x^*)h\| \\ &\geq \Big(a - t\frac{a}{2}\Big)\|h\| \\ &\geq \frac{a}{2}\sigma \\ &> 0, \end{aligned}$$

ein Widerspruch. Damit ist gezeigt, dass

$$x + H(x,t) \neq 0 \quad \text{für alle } (x,t) \in \partial B(x^*;\sigma) \times [0,1].$$

Aus dem Homotopiesatz 10.11 folgt, dass $d(I + H(\cdot,t), B(x^*;\sigma), 0)$ auf $[0,1]$ konstant ist. Insbesondere ist daher

$$\begin{aligned} d(I - G, B(x^*;\sigma), 0) &= d(I + H(\cdot,1), B(x^*;\sigma), 0) \\ &= d(I + H(\cdot,0), B(x^*;\sigma), 0) \\ &= d(I - G'(x^*) - [x^* - G'(x^*)x^*], B(x^*;\sigma) \\ &= d(I - G'(x^*), B(x^*;\sigma), x^* - G'(x^*)x^*) \\ &= \pm 1. \end{aligned}$$

Damit ist schließlich die Aussage (b) bewiesen.

Für (c) bleibt wegen Satz 10.10 nachzuweisen, dass

$$\lim_{n\to\infty} \sup_{x \in B[x^*;\sigma]} \|G(x) - P_n G(x)\| = 0.$$

Hierbei beachte man, dass das sup existiert, da $G - P_n G$ kompakt ist, also die beschränkte Menge $B[x^*;\sigma]$ in eine relativ kompakte und daher speziell beschränkte Menge übergeführt wird. Die Behauptung folgt dann aber wörtlich wie der Beweis der entsprechenden Aussage auf S. 187.

Die Aussagen (d) und (e) sind nun offensichtlich richtig und die Aussage ist schließlich bewiesen. □

11 M-Matrizen

11.1 Äquivalente Definitionen einer M-Matrix

Wir wählen als Ausgangsdefinition für eine M-Matrix die folgende (siehe z. B. R. S. Varga (1999, p. 91)):

Definition 11.1 Eine Matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ mit $a_{ij} \leq 0$ für alle $i, j \in \{1, \ldots, n\}$ mit $i \neq j$ heißt eine *M-Matrix*, wenn A nichtsingulär ist und $A^{-1} \geq 0$ bzw. A^{-1} (elementweise) nichtnegativ ist.

In dem folgenden Beweis wird wiederholt eine leichte Folgerung aus dem Satz von Perron (siehe z. B. J. Werner (2013, S. 209)) benutzt:

- *Ist $A \in \mathbb{R}^{n\times n}$ (elementweise) nichtnegativ, so ist der Spektralradius $\rho(A)$ von A ein Eigenwert von A, zu dem ein nichtnegativer Eigenvektor existiert.*

Denn: Sei $\{A_k\} \subset \mathbb{R}^{n\times n}$ eine Folge (elementweise) *positiver* Matrizen mit $A_k \to A$. Wegen des Satzes von Perron existiert insbesondere für jedes $k \in \mathbb{N}$ ein (elementweise) positiver Vektor $x_k \in \mathbb{R}^n$ mit $A_k x_k = \rho(A_k)x_k$ und $\|x_k\| = 1$, wobei $\|\cdot\|$ eine vorgegebene Norm auf dem $\mathbb{R}^n$ ist. Indem wir notfalls zu einer Teilfolge übergehen, können wir $x_k \to x$ annehmen. Dann ist x (elementweise) nichtnegativ und $x \neq 0$ wegen $\|x\| = 1$. Ferner ist $Ax = \rho(A)x$. Damit ist gezeigt, dass der Spektralradius einer (elementweise) nichtnegativen Matrix ein Eigenwert ist, zu dem es einen (elementweise) nichtnegativen Eigenvektor gibt.

Aus dem folgenden Satz folgen äquivalente Definitionen einer M-Matrix.

Satz 11.2 *Sei $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ eine Matrix mit $a_{ij} \leq 0$ für alle $i, j \in \{1, \ldots, n\}$ mit $i \neq j$. Dann sind die folgenden Aussagen äquivalent:*

1. *A ist eine M-Matrix bzw. A nichtsingulär und A^{-1} (elementweise) nichtnegativ.*
2. *Es ist $A = sI - B$ mit einer (elementweise) nichtnegativen Matrix B und $s > \rho(B)$.*
3. *Ist λ ein Eigenwert von A, so ist $\Re(\lambda) > 0$.*
4. *Für jedes $\gamma \geq 0$ ist $A+\gamma I$ eine M-Matrix, also $A+\gamma I$ nichtsingulär und $(A+\gamma I)^{-1}$ (elementweise) nichtnegativ.*

Beweis: 1. $\Rightarrow$ 2.: Sei $s := \max_{i=1,\ldots,n} a_{ii}$ und $B := sI - A$. Dann ist B eine (elementweise) nichtnegative Matrix. Sei x ein (elementweise) nichtnegativer Eigenvektor zum Eigenwert $\rho(B)$ von B. Dann ist $Ax = (s - \rho(B))x$. Da A nichtsingulär und $x \neq 0$, ist $s \neq \rho(B)$. Da x und A^{-1} (elementweise) nichtnegativ sind, folgt $s > \rho(B)$ aus $x = (s - \rho(B))A^{-1}x$.

2. $\Rightarrow$ 3.: Nach Voraussetzung lässt sich A in der Form $A = sI - B$ mit einer (elementweise) nichtnegativen Matrix B und $s > \rho(B)$ darstellen. Sei λ ein Eigenwert von A und daher $s - \lambda$ ein Eigenwert von B. Dann ist

$$|s - \lambda| \leq \rho(B) < s$$

und folglich $\Re(\lambda) > 0$.

3. $\Rightarrow$ 4.: Sei $s := \max_{i=1,\dots,n} a_{ii}$. Dann ist $s > 0$, denn andernfalls wäre A eine (elementweise) nichtpositive Matrix und hätte einen reellen, nichtpositiven Eigenwert. Offensichtlich ist $B := sI - A$ eine (elementweise) nichtnegative Matrix. Da $\rho(B)$ ein Eigenwert von B und folglich $s-\rho(B)$ ein Eigenwert von A ist, ist $s-\rho(B) = \Re(s-\rho(B)) > 0$. Mit vorgegebenem $\gamma \geq 0$ ist auch

$$B_\gamma := \frac{1}{s+\gamma} B$$

eine (elementweise) nichtnegativ, ferner ist

$$\rho(B_\gamma) = \frac{1}{s+\gamma}\rho(B) < \frac{s}{s+\gamma} \leq 1.$$

Daher ist $I - B_\gamma$ nichtsingulär und

$$(I - B_\gamma)^{-1} = \sum_{k=0}^{\infty} B_\gamma^k$$

(elementweise) nichtnegativ. Wegen

$$\frac{1}{s+\gamma}(A + \gamma I) = I - B_\gamma$$

ist auch $A + \gamma I$ nichtsingulär und $(A + \gamma I)^{-1}$ (elementweise) nichtnegativ und damit $A + \gamma I$ eine M-Matrix.

4. $\Rightarrow$ 1.: Diese Richtung ist trivial: Setze $\gamma := 0$. □

Satz 11.2 liefert zwei weitere äquivalente Definitionen einer M-Matrix, die jeweils als Ausgangsdefinition genommen werden könnten. Z. B. ist bei Wikipedia eine M-Matrix eine Matrix, deren Einträge außerhalb der Diagonalen nichtpositiv sind und deren Eigenwerte einen positiven Realteil haben. Bei R. J. Plemmons (1977), wo man (ohne Beweis) ebenso wie bei dem Wikipedia-Artikel viele äquivalente Definitionen einer M-Matrix findet, ist dagegen $A \in \mathbb{R}^{n\times n}$ definiert als eine M-Matrix, wenn A sich darstellen lässt in der Form $A = sI - B$ mit einer (elementweise) nichtnegativen Matrix B und $s > \rho(B)$.

11.2 Der Satz von Krein-Rutman im $\mathbb{R}^n$

Haupthilfsmittel im Beweis von Satz 11.2 war eine Folgerung aus dem Satz von Perron, dass nämlich der Spektralradius einer (elementweise) nichtnegativen Matrix einer ihrer Eigenwerte ist und hierzu ein (elementweise) nichtnegativer Eigenvektor existiert. Eine $n \times n$-Matrix ist offenbar genau dann (elementweise) nichtnegativ, wenn sie den natürlichen Ordnungskegel

$$K_+^n := \{x = (x_j) \in \mathbb{R}^n : x_j \geq 0,\ j = 1, \dots, n\}$$

der (elementweise) nichtnegativen Punkte des $\mathbb{R}^n$ in sich abbildet. Als Verallgemeinerung hierzu betrachten wir nun Matrizen, die einen vorgegebenen Ordnungskegel invariant lassen bzw. diesen in sich abbilden. Hierbei heißt eine Menge $K \subset \mathbb{R}^n$ bekanntlich ein *Kegel* (im $\mathbb{R}^n$), wenn mit $x \in K$ auch $\lambda x \in K$ für jedes $\lambda \geq 0$, wenn also mit jedem Punkt aus K auch der ganze Strahl, ausgehend vom Nullpunkt, durch diesen Punkt zu K gehört. Ein konvexer Kegel K mit $K \cap (-K) = \{0\}$ heißt ein *Ordnungskegel*. Unser Ziel in diesem Unterabschnitt ist es, den folgenden Satz zu beweisen. Diesen Satz bezeichnen wir als den Satz von Krein-Rutmann im $\mathbb{R}^n$. Wir präsentieren einen sehr schönen Beweis, der auf G. Birkhoff (1967) zurückgeht, siehe auch A. Berman, R. J. Plemmons (1979, p. 6).

Satz 11.3 *Sei $K \subset \mathbb{R}^n$ ein abgeschlossener Ordnungskegel mit* int $(K) \neq \emptyset$ *und $A \in \mathbb{R}^{n\times n}$ eine Matrix mit $A(K) \subset K$ bzw. $Ax \in K$ für alle $x \in K$. Dann ist der Spektralradius $\rho(A)$ von A ein Eigenwert von A, zu dem ein Eigenvektor aus K existiert.*

Beweis: Wir erinnern zunächst an die *Jordansche Normalform* von A. Hiernach existiert eine nichtsinguläre Matrix $P \in \mathbb{C}^{n\times n}$ mit $P^{-1}AP = J$, wobei

$$J = \text{diag}\,(J_1, \ldots, J_k)$$

eine Blockdiagonalmatrix ist und die Blöcke J_i die Form

$$J_i = \begin{pmatrix} \lambda_i & 1 & \cdots & 0 \\ 0 & \lambda_i & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & \lambda_i \end{pmatrix} \in \mathbb{C}^{m_i\times m_i}, \qquad i = 1, \ldots, k,$$

haben. Für $i = 1, \ldots, p$ sind λ_i die Eigenwerte von A. Zunächst nehmen wir an, es sei $\rho(A) = 0$, d. h. alle Eigenwerte λ_i, $i = 1, \ldots, k$, von A verschwinden. Wegen $J_i^{m_i} = 0$ ist $A^r = 0$ mit $r := \max_{i=1,\ldots,k} m_i$ und $A^{r-1} \neq 0$. Dann existiert ein $x \in K \setminus \{0\}$ mit $w := A^{r-1}x \neq 0$. Dies ist eine Folge der Voraussetzung int $(K) \neq \emptyset$, welche $\mathbb{R}^n = K - K$ (Kegel mit dieser Eigenschaft nennt man auch *reproduzierend*: Jeder Punkt des $\mathbb{R}^n$ lässt sich als Differenz zweier Punkte aus K darstellen) impliziert, wie man sehr leicht nachweist. Wäre also $A^{r-1}x = 0$ für jedes $x \in K$, so wäre $A^{r-1} = 0$, ein Widerspruch. Dann ist $Aw = A^r x = 0$, also w ein Eigenvektor aus K zum Eigenwert $\rho(A) = 0$.

Nun nehmen wir an, es sei $\rho(A) > 0$. In der Jordanschen Normalform $P^{-1}AP = J$ von A sei $P = (P_1, \ldots, P_k)$ mit

$$P_i = \begin{pmatrix} x_{i1} & \cdots & x_{im_i} \end{pmatrix} \in \mathbb{R}^{n\times m_i}, \qquad i = 1, \ldots, k.$$

Dann ist

$$\begin{pmatrix} AP_1 & \cdots & AP_k \end{pmatrix} = AP = PJ = \begin{pmatrix} P_1J_1 & \cdots & P_kJ_k \end{pmatrix}$$

und folglich

$$AP_i = P_iJ_i, \qquad i = 1, \ldots, k.$$

Mit

$$x_{i0} := 0, \qquad i = 1, \ldots, k,$$

ist also

$$Ax_{ij} = \lambda_i x_{ij} + x_{i,j-1}, \qquad i = 1, \ldots, k, \quad j = 1, \ldots, m_i,$$

und daher insbesondere

$$Ax_{i1} = \lambda_i x_{i1}, \qquad i = 1, \ldots, k.$$

Für $r \in \{j, j+1, \ldots\}$ ist dann

$$A^r x_{ij} = \sum_{s=0}^{j-1} \binom{r}{s} \lambda_i^{r-s} x_{i,j-s}. \tag{1}$$

Dies beweisen wir durch vollständige Induktion nach r. Für $r = 1$ ergibt sich für die rechte Seite

$$\binom{1}{0} \lambda_i^1 x_{ij} + \binom{1}{1} \lambda_i^0 x_{i,j-1} = \lambda_i x_{ij} + x_{i,j-1} = Ax_{ij},$$

also ist $(*)$ für $r = 1$ richtig. Für den Induktionsschluss nehmen wir an, die Aussage (1) sei für r richtig und beachten, dass dann

$$\begin{aligned}
A^{r+1} x_{ij} &= A\Big(\sum_{s=0}^{j-1} \binom{r}{s} \lambda_i^{r-s} x_{i,j-s}\Big) \\
&= \sum_{s=0}^{j-1} \binom{r}{s} \lambda_i^{r-s} A x_{i,j-s} \\
&= \sum_{s=0}^{j-1} \binom{r}{s} \lambda_i^{r-s} (\lambda_i x_{i,j-s} + x_{i,j-s-1}) \\
&= \sum_{s=0}^{j-1} \binom{r}{s} \lambda_i^{r+1-s} x_{i,j-s} + \sum_{s=0}^{j-1} \binom{r}{s} \lambda_i^{r-s} x_{i,j-s-1} \\
&= \sum_{s=0}^{j-1} \Big[\binom{r}{s} + \binom{r}{s-1}\Big] \lambda_i^{r+1-s} x_{i,j-s} \\
&= \sum_{s=0}^{j-1} \binom{r+1}{s} \lambda_i^{r+1-s} x_{i,j-s},
\end{aligned}$$

womit (1) bewiesen ist. Hierbei ist der Binomialkoeffizient

$$\binom{r}{s} = \frac{r(r-1)\cdots(r-(s-1))}{s!}$$

ein Polynom vom Grade s in r. Nun bestimme man ein $z \in \text{int}\,(K)$ derart, dass in der eindeutigen Darstellung

$$z = \sum_{i=1}^{k} \sum_{j=1}^{m_i} c_{ij} x_{ij}$$

alle c_{ij} nicht verschwinden, also $c_{ij} \neq 0$, $i = 1, \ldots, k$, $j = 1, \ldots, m_i$, gilt. Dies erreicht man, indem man sich zunächst ein beliebiges $y \in \text{int}\,(K)$ wählt. Dieses hat eine eindeutige Darstellung

$$y = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \alpha_{ij} x_{ij}.$$

Dann existiert ein $\delta_0 > 0$ derart, dass

$$\sum_{i=1}^{k} \sum_{j=1}^{m_i} (\alpha_{ij} + \delta) x_{ij} \in \text{int}\,(K)$$

für alle δ mit $|\delta| \leq \delta_0$. Nun bestimme man ein $\delta \neq 0$ mit $|\delta| \leq \delta_0$ derart, dass

$$c_{ij} := \alpha_{ij} + \delta \neq 0, \quad i = 1, \ldots, k,\ j = 1, \ldots, m_i.$$

Dass dies möglich ist, erkennt man, indem man die beiden Fälle $\alpha_{ij} = 0$ und $\alpha_{ij} \neq 0$ betrachtet. Also existiert ein $z \in \text{int}\,(K)$ mit der Eigenschaft, dass *alle* Koeffizienten c_{ij} in der Darstellung von z als Linearkombination der x_{ij} von Null verschieden sind. Dann ist

$$A^r z = \sum_{i=1}^{k} \sum_{j=1}^{m_i} c_{ij} A^r x_{ij} = \sum_{i=1}^{k} \sum_{j=1}^{m_i} c_{ij} \sum_{s=0}^{j-1} \binom{r}{s} \lambda_i^{r-s} x_{i,j-s}. \tag{2}$$

Wegen $A(K) \subset K$ ist $\{A^r z\}_{r \in \mathbb{N}} \subset K$. Nun sei

$$m := \max_{i \in \{1,\ldots,k\} : |\lambda_i| = \rho(A)} m_i$$

und

$$L := \{l \in \{1, \ldots, k\} : |\lambda_l| = \rho(A),\ m_l = m\}.$$

Für $l \in L$ sei schließlich $\lambda_l = \rho(A) e^{i\theta_l}$ mit $\theta_l \in [0, 2\pi)$. Den für große r dominierenden Term in (2) erhalten wir für $i \in L$, $j = m$ und $s = m - 1$. Da $\binom{r}{m-1}$ ein Polynom in r vom Grade $m - 1$ ist, ist daher

$$A^r z = r^{m-1} \rho(A)^{r-(m-1)} \Big[\underbrace{\sum_{l \in L} \underbrace{c_{lm} e^{i(r-(m-1))\theta_l}}_{\neq 0} x_{l1}}_{\neq 0} + o(1) \Big], \tag{3}$$

wobei $\lim_{r \to \infty} o(1)/r = 0$. Folglich ist $A^r z \neq 0$ für alle hinreichend großen $r \in \mathbb{N}$, etwa alle $r \in \mathbb{N}_0$. Wir definieren

$$\Omega := \{\omega \in \mathbb{R}^n : \text{Es gibt eine Teilfolge von } \{A^r z / \|A^r z\|\}_{r \in \mathbb{N}_0}, \text{ die gegen } \omega \text{ konvergiert}\},$$

wobei $\|\cdot\|$ eine beliebige Norm auf dem $\mathbb{R}^n$ ist. Offenbar ist Ω eine nichtleere Teilmenge von $K \setminus \{0\}$. Wegen (3) enthält Ω nur von Null verschiedene Elemente ω der Form

$$\omega = \sum_{l \in L} \beta_l x_{l1}. \tag{5}$$

Nun überlegen wir uns, dass das folgende elementare Resultat[40] richtig ist:

- *Sei $\alpha \in \mathbb{C} \setminus \mathbb{R}_+$, also α eine komplexe Zahl, die keine nichtnegative (reelle) Zahl ist. Dann ist α Nullstelle eines Polynoms mit positiven Koeffizienten, d. h. es existieren positive Zahlen $w_0, \dots, w_q$ mit $\sum_{p=0}^{q} w_p \alpha^p = 0$.*

Denn: Sei $\alpha = \rho e^{i\theta}$ mit $\rho > 0$ und $\theta \in (0, 2\pi)$. O. B. d. A. können wir $\rho = 1$ annehmen. Denn ist $\sum_{p=0}^{q} w_p (e^{i\theta})^p = 0$ mit positiven $w_0, \dots, w_q$, so ist $\sum_{p=0}^{q} w_p \rho^{-p} (\rho e^{i\theta})^p = 0$. Daher sei im folgenden $\alpha = e^{i\theta}$ mit $\theta \in (0, 2\pi)$. Dann liegen α und alle Potenzen α^p, $p = 0, \dots, q$, auf dem Einheitskreis. Ist $\theta = (m/n)\pi$ ein rationales Vielfaches von π (also $m, n \in \mathbb{N}$), so ist die Aussage wegen $\sum_{p=0}^{2n-1} \alpha^p = 0$ trivial. Dies ist insbesondere für $\theta = \pi$ der Fall. Weiter können wir $\theta \in (0, \pi)$ annehmen. Ist die Behauptung für $\theta \in (0, \pi)$ richtig und $\theta \in (\pi, 2\pi)$, so ist $2\pi - \theta \in (0, \pi)$, sodass positive $w_0, \dots, w_q$ mit $\sum_{p=0}^{q} w_p (e^{i(2\pi-\theta)})^p = 0$ existieren. Dann ist aber auch

$$\overline{\sum_{p=0}^{q} w_p (e^{i(2\pi-\theta)})^p} = \sum_{p=0}^{q} w_p (e^{i\theta})^p = 0.$$

Daher nehmen wir nun an, $\theta \in (0, \pi)$ sei kein rationales Vielfaches von π. Sei $q := \lceil \pi/\theta \rceil$ die kleinste natürliche Zahl, die größer oder gleich π/θ ist. Da wir annehmen, dass θ

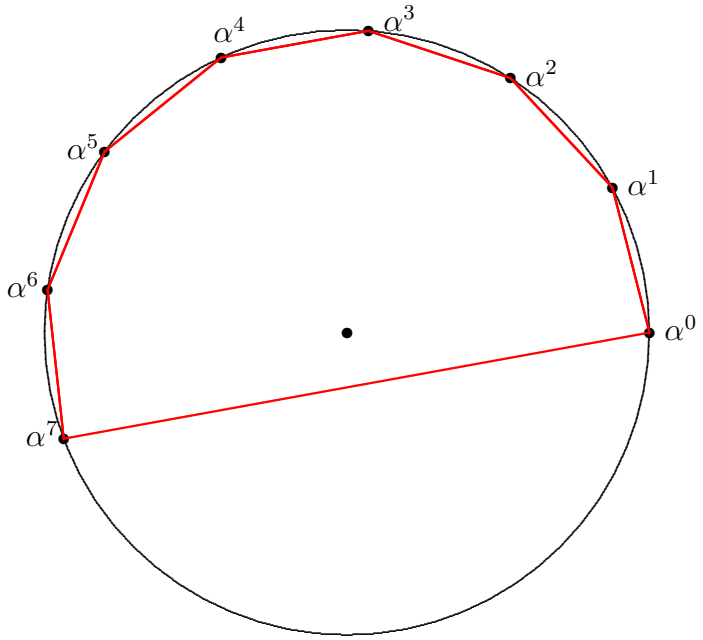

Abbildung 8: Beweis obiger Hilfsaussage

kein rationales Vielfaches von π ist, ist $(q-1)\theta < \pi < q\theta$. Offenbar ist q die kleinste natürliche Zahl mit der Eigenschaft, dass[41] $0 \in \operatorname{co}(\{\alpha^0, \alpha^1, \dots, \alpha^q\})$, dass also der

[40] Dieses wird von G. Birkhoff (1967) und auch von A. Berman, R. J. Plemmons (1979) *nicht* bewiesen.

[41] Hierbei identifizieren wir natürlich die Potenzen α^p der komplexen Zahl $\alpha = e^{i\theta}$ mit den Punkten $(\cos p\theta, \sin p\theta)$ des $\mathbb{R}^2$.

Nullpunkt in der konvexen Hülle von $\{\alpha^0, \ldots, \alpha^p\}$ bzw. der kleinsten die Punkte α^p, $p = 0, \ldots, q$, enthaltenden konvexen Menge liegt. Genauer ist sogar

$$0 \in \text{co}\,(\{\alpha^0, \alpha^{q-1}, \alpha^q\}) \subset \text{co}\,(\{\alpha^0, \alpha^{q-2}, \alpha^{q-1}, \alpha^q\}) \subset \cdots \subset \text{co}\,(\{\alpha^0, \alpha^1, \ldots, \alpha^{q-1}, \alpha^q\}).$$

Hier ist lediglich $0 \in \text{co}\,(\{\alpha^0, \alpha^{q-1}, \alpha^q\})$ zu zeigen, da die weiteren Inklusionen trivialerweise gelten. Man sehe sich hierzu Abbildung 8 an. Dass aber 0 im Innern des Dreiecks mit den Eckpunkten 1, $e^{i(q-1)\theta}$ und $e^{iq\theta}$ liegt, folgt aus $(q-1)\theta < \pi < q\theta$. Daher existieren positive Parameter $\lambda_0, \lambda_{q-1}, \lambda_q$ mit

$$0 = \lambda_0\alpha^0 + \lambda_{q-1}\alpha^{q-1} + \lambda_q\alpha^q$$

(und $\lambda_0 + \lambda_{q-1} + \lambda_q = 1$). Nun ist α^{q-1} enthalten in cone $(\{\alpha^q, \alpha^{q-2}\})$, der *konvexen Kegelhülle* von $\{\alpha^q, \alpha^{q-2}\}$. Daher existieren positive Zahlen a, b mit $\alpha^{q-1} = a\alpha^q + b\alpha^{q-2}$. Folglich ist

$$\begin{aligned}
0 &= \lambda_0\alpha^0 + \lambda_{q-1}\alpha^{q-1} + \lambda_q\alpha^q \\
&= \lambda_0\alpha^0 + \frac{\lambda_{q-1}}{2}\alpha^{q-1} + \frac{\lambda_{q-1}}{2}\alpha^{q-1} + \lambda_q\alpha^q \\
&= \lambda_0\alpha^0 + \frac{\lambda_{q-1}}{2}(a\alpha^q + b\alpha^{q-2}) + \frac{\lambda_{q-1}}{2}\alpha^{q-1} + \lambda_q\alpha^q \\
&= \lambda_0\alpha^0 + \frac{b\lambda_{q-1}}{2}\alpha^{q-2} + \frac{\lambda_{q-1}}{2}\alpha^{q-1} + \Big(\frac{a\lambda_{q-1}}{2} + \lambda_q\Big)\alpha^q \\
&= \mu_0\alpha^0 + \mu_{q-2}\alpha^{q-2} + \mu_{q-1}\alpha^{q-1} + \mu_q\alpha^q
\end{aligned}$$

mit positiven $\mu_0, \mu_{q-2}, \mu_{q-1}, \mu_q$. Nach endlich vielen Schritten erhalten wir, dass α Nullstelle eines Polynoms q-ten Grades mit positiven Koeffizienten ist. Die obige Hilfsaussage ist damit bewiesen.

O. B. d. A. sei $L = \{1, \ldots, h\}$. Ferner seien die Eigenwerte $\lambda_1, \ldots, \lambda_h$ so angeordnet, dass $\lambda_l = \rho(A)e^{i\theta_l}$ mit $0 \le \theta_1 \le \cdots \le \theta_h < 2\pi$, $l = 1, \ldots, h$. Sei

$$\omega_h = \sum_{l=1}^{h} \beta_{lh} x_{l1}$$

ein beliebig aus Ω herausgegriffenes Element. Ist $\lambda_h = \rho(A)$, so ist (hier geht ein, dass $0 \le \theta_1 \le \cdots \le \theta_h < 2\pi$, also $\theta_h = 0$ gilt) $\lambda_1 = \cdots = \lambda_h = \rho(A)$ und folglich $\rho(A)$ ein Eigenwert von A mit dem Eigenvektor $\omega_h \in K$. In diesem Fall ist der Satz schon bewiesen. Daher nehmen wir jetzt an, es sei $\lambda_h \neq \rho(A)$, also $\lambda_h = \rho(A)e^{i\theta_h}$ keine positive Zahl. Wegen obiger Aussage existieren positive Zahlen $w_0, \ldots, w_q$ mit $\sum_{p=0}^{q} w_p\lambda_h^p = 0$. Dann ist

$$\omega_{h-1} := \sum_{p=0}^{q} w_p A^p \omega_h$$

als positive Linearkombination von Elementen aus K selbst ein Element aus K. Es ist $\omega_{h-1} \neq 0$, da andernfalls $w_p A^p \omega_h = 0$, $p = 0, \ldots, q$, was aber wegen $w_0\omega_h \neq 0$ nicht

sein kann. Ferner ist

$$\begin{aligned}\omega_{h-1} &= \sum_{p=0}^{q} w_p A^p \omega_h \\ &= \sum_{p=0}^{q} w_p \sum_{l=1}^{h} \beta_{lh} A^p x_{l1} \\ &= \sum_{l=1}^{h} \Big(\beta_{lh} \sum_{p=0}^{q} w_p \lambda_l^p \Big) x_{l1} \\ &= \sum_{l=1}^{h-1} \beta_{l,h-1} x_{l1}\end{aligned}$$

mit

$$\beta_{l,h-1} := \beta_{lh} \sum_{p=0}^{q} w_p \lambda_l^p, \qquad l \in L,$$

wobei wir

$$\beta_{h,h-1} = \beta_{hh} \underbrace{\sum_{p=0}^{q} w_p \lambda_h^p}_{=0} = 0$$

ausgenutzt haben. So können wir fortfahren. Ist $\lambda_{f+1} \neq \rho(A)$ und $\lambda_f = \rho(A)$, so ist $\rho(A)$ ein Eigenwert von A mit dem Eigenvektor $\omega_f = \sum_{l=1}^{f} \beta_{lf} x_{l1}$, andernfalls existiert $\omega_{f-1} = \sum_{l=1}^{f-1} \beta_{l,f-1} x_{l1} \in K \setminus \{0\}$. Dieser Prozess zeigt, dass (spätestens) $\lambda_1 = \rho(A)$ ein Eigenwert von A mit dem Eigenvektor $\omega_1 = \beta_{11} x_{11}$ ist. Der Satz ist damit bewiesen. □

11.3 M-Matrizen bezüglich eines Ordnungskegels

In Definition 11.1 hatten wir eine nichtsinguläre Matrix $A \in \mathbb{R}^{n\times n}$ eine M-Matrix genannt, wenn ihre Außerdiagonalelemente nichtpositiv sind und die inverse Matrix A^{-1} elementweise nichtnegativ ist. Diese Definition wird nun verallgemeinert.

Definition 11.4 Sei $A \in \mathbb{R}^{n\times n}$ und $K \subset \mathbb{R}^n$ ein Ordnungskegel.

(a) Die Matrix A heißt *quasipositiv bezüglich des Ordnungskegels* K, wenn ein $\alpha \geq 0$ mit $(A + \alpha I)(K) \subset K$ existiert.

(b) Die Matrix A heißt eine *M-Matrix bezüglich des Ordnungskegels* K, wenn A nichtsingulär, $-A$ quasipositiv bezüglich K ist und $A^{-1}(K) \subset K$ gilt.

Offenbar stimmt diese Definition für den natürlichen Ordnungskegel $K = \mathbb{R}^n_+$ mit der in Definition 11.1 überein. Der folgende Satz verallgemeinert Satz 11.2. Es wird sich herausstellen, dass der Beweis fast völlig dem von Satz 11.2 entspricht.

Satz 11.5 *Sei $K \subset \mathbb{R}^n$ ein abgeschlossener Ordnungskegel mit* int $(K) \neq \emptyset$ *und $A \in \mathbb{R}^{n\times n}$ eine Matrix mit der Eigenschaft, dass $-A$ quasipositiv bezüglich K ist. Dann sind die folgenden Aussagen äquivalent:*

1. *A ist eine M-Matrix bezüglich K bzw. A nichtsingulär und $A^{-1}(K) \subset K$.*
2. *Es ist $A = sI - B$ mit einer Matrix B mit $B(K) \subset K$ und $s > \rho(B)$.*
3. *Ist λ ein Eigenwert von A, so ist $\Re(\lambda) > 0$.*
4. *Für jedes $\gamma \geq 0$ ist $A+\gamma I$ eine M-Matrix bezüglich K, also $-(A+\gamma I)$ quasipositiv bezüglich K und nichtsingulär sowie $(A+\gamma I)^{-1}(K) \subset K$.*

Beweis: Da $-A$ quasipositiv bezüglich K ist, existiert $\alpha \geq 0$ mit $(-A+\alpha I)(K) \subset K$.
1. $\Rightarrow$ 2.: Sei $B := \alpha I - A$. Nach Wahl von α ist $B(K) \subset K$. Nach Satz 11.3, dem Satz von Krein-Rutman im $\mathbb{R}^n$, gibt es ein $x \in K \setminus \{0\}$ mit $Bx = \rho(B)x$. Dann ist $Ax = (\alpha - \rho(B))x$. Da A nichtsingulär und $x \neq 0$ ist, $\alpha \neq \rho(B)$. Wegen

$$\underbrace{Bx}_{\in K} = (\alpha - \rho(B))\underbrace{x}_{\in K}$$

ist $\alpha > \rho(B)$, da andernfalls $x \in K \cap (-K) = \{0\}$, ein Widerspruch. Damit gilt 2. mit $s := \alpha$.
2. $\Rightarrow$ 3.: Sei $\lambda \in \mathbb{C}$ ein Eigenwert von A. Dann ist $s - \lambda$ ein Eigenwert von B, daher

$$s - \Re(\lambda) = \Re(s - \lambda) \leq |s - \lambda| \leq \rho(B) < s$$

und folglich $\Re(\lambda) > 0$.
3. $\Rightarrow$ 4.: Sei $\gamma \geq 0$ gegeben. Wir haben zu zeigen, dass $-(A + \gamma I)$ quasipositiv und nichtsingulär ist, ferner $(A + \gamma I)^{-1}(K) \subset K$ gilt. Ersteres ist offenbar wegen

$$(-A + \alpha I)(K) = [-(A + \gamma I) + (\underbrace{\alpha + \gamma}_{\geq 0})I](K) \subset K$$

richtig. Es ist $\alpha > 0$, denn andernfalls wäre

$$-A(K) \subset (-A + \alpha I)(K) - \alpha K \subset K + K \subset K$$

und nach Satz 11.2 hätte $-A$ einen nichtnegativen bzw. A einen nichtpositiven (reellen) Eigenwert, was durch 3. ausgeschlossen ist. Wir definieren $B := \alpha I - A$. Wegen $B(K) \subset K$ ist $\rho(B)$ nach Satz 11.2 ein Eigenwert von B und daher $\alpha - \rho(B)$ ein Eigenwert von A. Wegen 3. ist $\alpha - \rho(B) = \Re(\alpha - \rho(B)) > 0$. Definiert man

$$B_\gamma := \frac{1}{\alpha + \gamma} B,$$

so ist auch $B_\gamma(K) \subset K$ und

$$\rho(B_\gamma) = \frac{1}{\alpha + \gamma}\rho(B) < \frac{\alpha}{\alpha + \gamma} \leq 1.$$

Daher ist $I - B_\gamma$ nichtsingulär. Wegen $(I - B_\gamma)^{-1} = \sum_{k=0}^{\infty} B_\gamma^k$ und der Voraussetzung, dass $K \subset \mathbb{R}^n$ ein abgeschlossener Ordnungskegel ist, ist $(I - B_\gamma)^{-1}(K) \subset K$. Wegen

$$\frac{1}{\alpha + \gamma}(A + \gamma I) = I - B_\gamma$$

ist auch $A + \gamma I$ nichtsingulär und $(A + \gamma I)^{-1}(K) \subset K$.
4. $\Rightarrow$ 1.: Diese Richtung ist trivial: Setze $\gamma := 0$. $\square$

12 Der Zwischenwertsatz

12.1 Der klassische Zwischenwertsatz der Analysis

Der klassische *Zwischenwertsatz* der Analysis sagt aus:

Satz 12.1 *Sei $[a,b] \subset \mathbb{R}$ ein kompaktes Intervall und $f\colon [a,b] \longrightarrow \mathbb{R}$ stetig. Dann existiert zu jedem u zwischen $f(a)$ und $f(b)$ ein $u \in [a,b]$ mit $f(c) = u$. D. h. zu jedem $u \in [f(a), f(b)]$ (falls $f(a) \leq f(b)$) bzw. $u \in [f(b), f(a)]$ (falls $f(b) < f(a)$) gibt es ein $c \in [a,b]$ mit $f(c) = u$.*

Beweis: O. B. d. A. sei $f(a) < f(b)$ und $u \in [f(a), f(b)]$. Man definiere die stetige Funktion $g\colon [a,b] \longrightarrow \mathbb{R}$ durch $g(x) := f(x) - u$. Dann ist $g(a) < g(b)$ und $g(a) \leq 0 \leq g(b)$. Nun definiere man die Folgen $\{a_k\}$ und $\{b_k\}$ durch

- Setze $a_1 := a$, $b_1 := b$.
- Für $k = 1, 2, \ldots$:
  - Berechne $c_k := \frac{1}{2}(a_k + b_k)$.
  - Falls $g(c_k) = 0$, dann: $c := c_k$, STOP.
  - Setze

$$a_{k+1} := \begin{cases} c_k, & \text{falls } g(c_k) < 0, \\ a_k, & \text{sonst,} \end{cases} \qquad b_{k+1} := \begin{cases} b_k, & \text{falls } g(c_k) < 0, \\ c_k, & \text{sonst.} \end{cases}$$

Dann ist $\{a_k\}$ eine monoton nicht fallende, nach oben beschränkte Folge und $\{b_k\}$ eine monoton nicht steigende, nach unten beschränkte Folge. Beide Folgen sind daher konvergent und ihr Limes ist wegen

$$b_k - a_k = \frac{b-a}{2^{k-1}}, \qquad k \in \mathbb{N},$$

gleich, etwa gleich c. Wegen

$$g(a_k) \leq 0 \leq g(b_k), \qquad k \in \mathbb{N},$$

und der Stetigkeit von g ist $g(c) = 0$ bzw. $f(c) = u$. □

Als Folgerung aus dem Zwischenwertsatz 12.1 erhalten wir den *Nullstellensatz von Bolzano*, aus dem man wiederum sehr leicht den Zwischenwertsatz folgert.

Korollar 12.2 *Sei $f\colon [a,b] \longrightarrow \mathbb{R}$ stetig und $f(a)f(b) \leq 0$. Dann existiert ein $x^* \in [a,b]$ mit $f(x^*) = 0$.*

Bemerkung: Wir wollen uns überlegen, dass (für $n = 1$) der Brouwersche Fixpunktsatz aus dem Nullstellensatz folgt. Denn sei $F\colon [a,b] \longrightarrow \mathbb{R}$ stetig und $F([a,b]) \subset [a,b]$. Definiert man die stetige Funktion $f\colon [a,b] \longrightarrow \mathbb{R}$ durch $f(x) := x - f(x)$, so ist $f(a) \leq 0 \leq f(b)$, sodass wegen des Nullstellensatzes eine Nullstelle von f bzw. ein Fixpunkt von F in $[a,b]$ existiert. Auch umgekehrt kann der Nullstellensatz mit Hilfe des Brouwerschen Fixpunktsatzes bewiesen werden, wie wir gleich beim Beweis des verallgemeinerten Zwischenwertsatzes von Poincaré-Miranda sehen werden. □

12.2 Der Satz von Poincaré-Miranda

Von H. Poincaré (1883) stammt die folgende Vermutung (wörtlich zitiert nach F. Browder (1983)):

- Let $\xi_1, \xi_2, \ldots, \xi_n$ be n continuous functions of n variables $x_1, x_2, \ldots, x_n$: the variable x_i is subjected to vary between the limits $+a_i$ and $-a_i$. Let us suppose for $x_i = a_i$, ξ_i is constantly positive, and that for $x_i = -a_i$ constantly negative; I say there will exist a system of values of x for which all the ξ's vanish.

Diese Aussage wurde von C. Miranda (1940) bewiesen. Wir bezeichnen den folgenden Satz (es gibt verschiedene Versionen, wir folgen im wesentlichen V. I. Istrătescu (1981, S. 118)) als *Satz von Poincaré-Miranda*. Etwas andere Versionen werden in der folgenden Bemerkung vorgestellt.

Satz 12.3 *Seien $a = (a_i)$, $b = (b_i) \in \mathbb{R}^n$ mit $a_i < b_i$, $i = 1, \ldots, n$, gegeben. Hiermit sei der Quader*

$$Q := \{x = (x_i) \in \mathbb{R}^n : a_i \le x_i \le b_i,\ i = 1, \ldots, n\}$$

definiert. Für $i = 1, \ldots, n$ seien die Abbildungen $F_i \colon Q \longrightarrow \mathbb{R}$, $i = 1, \ldots, n$, stetig und

$$(*) \qquad \begin{cases} F_i(x_1, \ldots, x_{i-1}, a_i, x_{i+1}, \ldots, x_n) \ge 0 \\ F_i(x_1, \ldots, x_{i-1}, b_i, x_{i+1}, \ldots, x_n) \le 0 \end{cases} \quad \text{für alle } x \in Q.$$

Dann existiert ein $x \in Q$ mit $F_i(x) = 0$, $i = 1, \ldots, n$, bzw. $F(x) = 0$, wobei $F \colon Q \longrightarrow \mathbb{R}^n$ durch $F(x) := (F_1(x), \ldots, F_n(x))^T$ definiert ist.

Beweis: Im ersten Teil des Beweises zeigen wir die Aussage des Satzes unter der stärkeren Voraussetzung, dass $(*)$ für $i = 1, \ldots, n$ sogar mit dem $>$- bzw. $<$-Zeichen erfüllt ist. Wir definieren $f_i \colon Q \longrightarrow \mathbb{R}$ durch $f_i(x) := x_i + \epsilon_i F_i(x)$ mit geeignet zu wählenden $\epsilon_i > 0$, $i = 1, \ldots, n$. Wir setzen

$$m_i := \min_{x \in Q} F_i(x), \qquad M_i := \max_{x \in Q} F_i(x).$$

Wegen der oben gemachten Annahme ist $m_i < 0 < M_i$, $i = 1, \ldots, n$.

(a) Es existiert $\delta_i^{(1)} > 0$ mit

$$x \in Q,\ F_i(x) < 0 \Longrightarrow \delta_i^{(1)} \le x_i - a_i.$$

Denn: Angenommen, es gibt eine Folge $\{x^{(k)}\} \subset Q$ mit $F_i(x^{(k)}) < 0$ und $x_i^{(k)} - a_i \to 0$. Da Q kompakt ist, existiert eine konvergente Teilfolge $\{x^{(k_j)}\} \subset \{x^{(k)}\}$, sei etwa $x^{(k_j)} \to \hat{x}$. Dann ist $\hat{x}_i = a_i$ und $F_i(\hat{x}) \le 0$, ein Widerspruch zu der Annahme, dass $(*)$ mit $>$ erfüllt ist.

(b) Es existiert $\delta_i^{(2)} > 0$ mit

$$x \in Q,\ F_i(x) > 0 \Longrightarrow \delta_i^{(2)} \le b_i - x_i.$$

Der Beweis für diese Aussage verläuft entsprechend dem Beweis von (a).
Nun setze man

$$\epsilon_i := \min\Big(-\frac{\delta_i^{(1)}}{m_i}, \frac{\delta_i^{(2)}}{M_i}\Big), \qquad i = 1, \ldots, n.$$

Wir zeigen, dass $f = (f_i)$ mit $f_i(x) := x_i + \epsilon_i F_i(x)$, $i = 1, \ldots, n$, den Quader Q in sich abbildet.

1. Ist $x \in Q$ und $F_i(x) = 0$, so ist $a_i \leq x_i = f_i(x) \leq b_i$.

2. Ist $x \in Q$ und $F_i(x) < 0$, so ist

$$\begin{aligned} a_i &\leq x_i - \delta_i^{(1)} \\ &= f_i(x) - \epsilon_i F_i(x) - \delta_i^{(1)} \\ &\leq f_i(x) \underbrace{-\epsilon_i m_i - \delta_i^{(1)}}_{\leq 0} \\ &\leq f_i(x) \\ &= x_i + \epsilon_i \underbrace{F_i(x)}_{<0} \\ &\leq b_i. \end{aligned}$$

3. Ist $x \in Q$ und $F_i(x) > 0$, so ist

$$\begin{aligned} a_i &\leq x_i + \epsilon_i \underbrace{F_i(x)}_{>0} \\ &= f_i(x) \\ &\leq b_i - \delta_i^{(2)} + \epsilon_i F_i(x) \\ &\leq b_i \underbrace{-\delta_i^{(2)} + \epsilon_i M_i}_{\leq 0} \\ &\leq b_i. \end{aligned}$$

Damit ist gezeigt, dass die Abbildung f den Quader Q in sich abbildet. Aus dem Brouwerschen Fixpunktsatz 9.19 folgt die Existenz eines Fixpunktes x von $f = (f_i)$ bzw. einer Nullstelle von $F = (F_i)$. Ist also die Bedingung $(*)$ für $i = 1, \ldots, n$ strikt erfüllt, so ist die Behauptung bewiesen.
Im zweiten Teil des Beweises nehmen wir an, $(*)$ sei erfüllt und definieren für $m \in \mathbb{N}$ die Abbildung $F_i^{(m)}: Q \longrightarrow \mathbb{R}$, $i = 1, \ldots, n$, durch

$$F_i^{(m)}(x) := F_i(x) - \frac{1}{m}\Big(x_i - \frac{a_i + b_i}{2}\Big).$$

Dann ist

$$\begin{aligned} F_i^{(m)}(x_1, \ldots, x_{i-1}, a_i, x_{i+1}, \ldots, x_n) &= F_i(x_1, \ldots, x_{i-1}, a_i, x_{i+1}, \ldots, x_n) \\ &\qquad + \frac{1}{m} \cdot \frac{b_i - a_i}{2} \\ &> 0 \end{aligned}$$

und entsprechend

$$\begin{aligned} F_i^{(m)}(x_1,\ldots,x_{i-1},b_i,x_{i+1},\ldots,x_n) &= F_i(x_1,\ldots,x_{i-1},b_i,x_{i+1},\ldots,x_n) \\ &\quad - \frac{1}{m}\cdot\frac{b_i-a_i}{2} \\ &< 0 \end{aligned}$$

für alle $x \in Q$. Für $F_i^{(m)}$ ist die Bedingung $(*)$ sogar strikt erfüllt, sodass wegen des schon bewiesenen ersten Teiles des Beweises $x^{(m)} \in Q$ mit $F_i^{(m)}(x^{(m)}) = 0$ existieren, $i = 1,\ldots,n$. Aus $\{x^{(m)}\} \subset Q$ kann eine gegen ein $x \in Q$ konvergente Teilfolge $\{x^{(m_j)}\}$ ausgewählt werden. Offensichtlich ist $F_i(x) = 0$, $i = 1,\ldots,n$, bzw. $F(x) = 0$. Der Satz ist bewiesen. □

Bemerkung: Von M. N. Vrahatis (1989) ist ein kurzer Beweis (einer etwas spezielleren Version) des Satzes von Poincaré-Miranda mit Hilfe des Abbildungsgrades und eine Verallgemeinerung angegeben worden. Genauer wird die folgende Aussage bei Vrahatis bewiesen:

- *Sei* $Q := \{x \in \mathbb{R}^n : |x_i| \le L,\ i = 1,\ldots,n\}$. *Die Abbildung*

$$F = (F_1,\ldots,F_n)\colon Q \longrightarrow \mathbb{R}^n$$

sei stetig und $F(x) \neq 0$ *für alle* $x \in \partial Q$. *Für alle* $x \in Q$ *sei ferner*

$$\begin{cases} F_i(x_1,\ldots,x_{i-1},-L,x_{i+1},\ldots,x_n) \ge 0, \\ F_i(x_1,\ldots,x_{i-1},+L,x_{i+1},\ldots,x_n) \le 0, \end{cases} \qquad i = 1,\ldots,n.$$

Dann existiert ein $x \in \text{int}\,(Q)$ *mit* $F(x) = 0$.

Zum Beweis mit Hilfe des Homotopie-Satzes 9.10 definieren wir $H\colon Q \times [0,1] \longrightarrow \mathbb{R}^n$ durch $H(x,t) := (1-t)F(x) + t(-x)$. Um den Satz über die Homotopie-Invarianz anwenden zu können, haben wir $H(x,t) \neq 0$ für alle $(x,t) \in \partial Q \times [0,1]$ nachzuweisen. Nun ist nach Voraussetzung $H(x,0) = F(x) \neq 0$ für alle $x \in \partial Q$. Weiter ist $H(x,1) = -x \neq 0$ für alle $x \in \partial Q$. Es ist aber auch $H(x,t) \neq 0$ für alle $(x,t) \in \partial Q \times (0,1)$. Denn ist $(x,t) \in \partial Q \times (0,1)$ vorgegeben, so existiert wegen $x \in \partial Q$ ein $i \in \{1,\ldots,n\}$ mit $x_i = -L$ oder $x_i = +L$. Bezeichnen wir mit $H_i(x,t)$ die i-te Komponente von $H(x,t)$, so ist im ersten Fall

$$H_i(x,t) = \underbrace{(1-t)F_i(x_1,\ldots,x_{i-1},-L,x_{i+1},\ldots,x_n)}_{\ge 0} + \underbrace{tL}_{>0} > 0,$$

während im zweiten Fall

$$H_i(x,t) = \underbrace{(1-t)F_i(x_1,\ldots,x_{i-1},L,x_{i+1},\ldots,x_n)}_{\le 0} + \underbrace{t(-L)}_{<0} < 0.$$

Daher ist

$$\begin{aligned} d(F,\text{int}\,(Q),0) &= d(H(\cdot,0),\text{int}\,(Q),0) \\ &= d(H(\cdot,1),\text{int}\,(Q),0) \\ &= d(-I,\text{int}\,(Q),0) \\ &= (-1)^n \\ &\neq 0. \end{aligned}$$

Aus dem Satz 9.18 von Kronecker folgt die Behauptung der obigen Aussage.
Von F. Stenger (1975, S. 37) ist mit Hilfe des Abbildungsgrades die folgende (verglichen mit den obigen Aussagen) schwache Version des Satzes von Poincaré-Miranda bewiesen worden:

- *Sei* $Q := \{x \in \mathbb{R}^n : |x_i| \leq 1\}$. *Die Abbildung*

$$F = (F_1, \ldots, F_n) \colon Q \longrightarrow \mathbb{R}^n$$

sei stetig. Für alle $x \in Q$ *sei ferner*

$$\begin{cases} F_i(x_1, \ldots, x_{i-1}, -1, x_{i+1}, \ldots, x_n) < 0, \\ F_i(x_1, \ldots, x_{i-1}, +1, x_{i+1}, \ldots, x_n) > 0, \end{cases} \qquad i = 1, \ldots, n.$$

Dann existiert ein $x \in \text{int}\,(Q)$ *mit* $F(x) = 0$.

Offenbar ist dies ein Spezialfall der vorigen Aussage. □

12.3 Einschließungssätze bei nichtlinearen Randwertaufgaben zweiter Ordnung

Als Prototyp eines *Einschließungssatzes* bei einer nichtlinearen Randwertaufgabe wollen wir uns überlegen, dass der folgende Satz richtig ist. In diesem wird ausgesagt, dass zwischen einer Unter- und einer Oberlösung eine Lösung einer Randwertaufgabe liegt. Daher kann der folgende Satz als eine Art Zwischenwertsatz angesehen werden.

Satz 12.4 *Gegeben sei die nichtlineare Randwertaufgabe (zweiter Ordnung)*

$$\text{(P)} \qquad -u'' = f(u,t), \qquad u(a) = u_a,\ u(b) = u_b.$$

Hierbei seien $u_a,\ u_b \in \mathbb{R}$ *vorgeben und* $f \colon \mathbb{R} \times [a,b] \longrightarrow \mathbb{R}$ *stetig sowie bezüglich der ersten Variablen stetig differenzierbar. Es mögen* $\alpha, \beta \in C^2[a,b]$ *existieren mit*

1. $\alpha(t) \leq \beta(t)$ *für alle* $t \in [a,b]$,
2. $-\alpha''(t) - f(\alpha(t),t) \leq 0 \leq -\beta''(t) - f(\beta(t),t)$ *für alle* $t \in [a,b]$, *d. h.* α *ist eine Unterlösung und* β *eine Oberlösung von (P),*
3. $\alpha(a) \leq u_a \leq \beta(a)$, $\alpha(b) \leq u_b \leq \beta(b)$.

Dann existiert eine Lösung $u \in C^2[a,b]$ *von (P) mit* $\alpha(t) \leq u(t) \leq \beta(t)$ *für alle* $t \in [a,b]$.

Beweis: Wir werden den Schauderschen Fixpunktsatz 10.16 anwenden. Der Raum $(X, \|\cdot\|) := (C[a,b], \|\cdot\|_\infty)$ ist ein Banachraum. Die Menge

$$A := \{x \in C[a,b] : \alpha(t) \leq x(t) \leq \beta(t) \ \text{ für alle } t \in [a,b]\}$$

ist nichtleer, abgeschlossen und konvex. Die Idee besteht darin, eine stetige Abbildung $F \colon A \subset C[a,b] \longrightarrow C[a,b]$ mit $F(A) \subset A$ und relativ kompaktem $F(A)$ zu bestimmen, deren nach dem Schauderschen Fixpunktsatz in A existierender Fixpunkt eine Lösung der nichtlinearen Randwertaufgabe (P) ist. Hierzu überlegen wir uns:

- Seien $r \in C[a,b]$, $\lambda \geq 0$ sowie $u_a, u_b \in \mathbb{R}$ gegeben. Dann besitzt die (lineare, inhomogene) Randwertaufgabe

  $$(*) \qquad -u'' + \lambda^2 u = r(t), \qquad u(a) = u_a,\ u(b) = u_b$$

  eine eindeutige Lösung, welche durch

  $$u(t) = u_a v_a^{(\lambda)}(t) + u_b v_b^{(\lambda)}(t) + \int_0^1 G^{(\lambda)}(t,s) r(s)\, ds$$

  gegeben ist. Hierbei ist

  $$v_a^{(0)}(t) := \frac{b-t}{b-a}, \qquad v_b^{(0)}(t) := \frac{t-a}{b-a}$$

  und

  $$G^{(0)}(t,s) := \frac{1}{b-a} \begin{cases} (s-a)(b-t), & a \leq s \leq t \leq b, \\ (b-s)(t-a), & a \leq t \leq s \leq b, \end{cases}$$

  während für $\lambda > 0$

  $$v_a^{(\lambda)}(t) := \frac{\sinh \lambda(b-t)}{\sinh \lambda(b-a)}, \qquad v_b^{(\lambda)}(t) := \frac{\sinh \lambda(t-a)}{\sinh \lambda(b-a)}$$

  und

  $$G^{(\lambda)}(t,s) := \frac{1}{\lambda \sinh \lambda(b-a)} \begin{cases} \sinh \lambda(s-a)\ \sinh \lambda(b-t), & a \leq s \leq t \leq b, \\ \sinh \lambda(b-s)\ \sinh \lambda(t-a), & a \leq t \leq s \leq b. \end{cases}$$

Denn: Durch Nachrechnen weist man nach, dass die für $\lambda = 0$ und $\lambda > 0$ angegebenen Funktionen Lösungen von $(*)$ sind. Die Eindeutigkeit erhält man sehr leicht, indem man nachweist, dass das homogene Problem (homogene Differentialgleichung und homogene Randbedingungen) nur trivial lösbar ist.
Nun sei

$$K := \{(x,t) \in \mathbb{R} \times [a,b] : \alpha(t) \leq x \leq \beta(t)\},$$

anschließend wähle man $\lambda \geq 0$ so groß, dass

$$(**) \qquad \min_{(x,t) \in K} \frac{\partial}{\partial x} f(x,t) + \lambda^2 \geq 0.$$

Mit diesem λ definiere man die Abbildung $F\colon A \longrightarrow C[a,b]$ durch

$$F(x)(t) := u_a v_a^{(\lambda)}(t) + u_b v_b^{(\lambda)}(t) + \int_a^b G^{(\lambda)}(t,s)[f(x(s),s) + \lambda^2 x(s)]\, ds,$$

wobei $v_a^{(\lambda)}(t), v_b^{(\lambda)}(t)$ und $G^{(\lambda)}(t,s)$ die oben für $\lambda = 0$ bzw. $\lambda > 0$ definierten Funktionen sind. Wichtig ist, dass $v_a^{(\lambda)}(\cdot), v_b^{(\lambda)}(\cdot)$ und vor allem die sogenannte *Greensche Funktion* $G^{(\lambda)}(\cdot,\cdot)$ auf $[a,b]$ bzw. $[a,b] \times [a,b]$ *nichtnegative* Funktionen sind.

Zunächst zeigen wir $F(A) \subset A$, danach die relative Kompaktheit von $F(A)$. Für $x \in A$ und beliebiges $t \in [a,b]$ ist

$$\begin{aligned} F(x)(t) &= u_a v_a^{(\lambda)}(t) + u_b v_b^{(\lambda)}(t) + \int_a^b \underbrace{G^{(\lambda)}(t,s)}_{\geq 0}[f(x(s),s) + \lambda^2 x(s)]\,ds \\ &\geq u_a v_a^{(\lambda)}(t) + u_b v_b^{(\lambda)}(t) + \int_a^b G^{(\lambda)}(t,s)[f(\alpha(s),s) + \lambda^2\alpha(s)]\,ds \\ &\qquad \text{(nach Wahl von } \lambda) \\ &\geq \alpha(a)v_a^{(\lambda)}(t) + \alpha(b)v_b^{(\lambda)}(t) + \int_a^b G^{(\lambda)}(t,s)[-\alpha''(s) + \lambda^2\alpha(s)]\,ds \\ &= \alpha(t). \end{aligned}$$

Entsprechend ist $F(x)(t) \leq \beta(t)$ für alle $t \in [a,b]$, womit insgesamt $F(A) \subset A$ bewiesen ist. Weiter ist $F(A) \subset C[a,b]$ relativ kompakt, wie man mit Hilfe des Satzes von Arzela-Ascoli (siehe Seite 178) feststellt. Wegen des Schauderschen Fixpunktsatzes besitzt F also einen Fixpunkt $u \in A$. Dieser Fixpunkt ist eine Lösung der nichtlinearen Randwertaufgabe (P). Damit ist der Satz bewiesen. □

Bemerkung: In Satz 12.4 betrachteten wir für eine gewöhnliche Differentialgleichung zweiter Ordnung die *erste Randwertaufgabe*, bei welcher die Werte der gesuchten Differentialgleichungslösung auf dem *Rande* des Intervalls vorgeben sind. Denkbar sind aber auch andere Randbedingungen, für die ebenfalls die Nichtnegativität der zugehörigen Greenschen Funktion gesichert werden kann. Hierauf wollen wir nicht näher eingehen, sondern nur z. B. auf L. Collatz (1964, S. 300) (monotone Art bei Sturmschen Randwertaufgaben) in Verbindung mit z. B. J. Werner (2001, S. 205 ff.) (Existenz der Greenschen Funktion bei Sturmschen Randwertaufgaben) verweisen. Dafür soll aber auf den Fall *periodischer Randbedingungen* näher eingegangen werden. Hierbei beschränken wir uns auf $[0,1]$ als zu Grunde liegendes Intervall, was durch eine Variablentransformation erreicht werden kann. Als Ersatz für die zu Beginn des Beweises von Satz 12.4 gemachte Aussage über die Darstellung der Lösung einer linearen, inhomogenen Randwertaufgabe erster Ordnung gilt:

- Seien $r \in C[0,1]$, $\lambda > 0$ sowie $u_0, u_1 \in \mathbb{R}$ gegeben. Dann besitzt die Randwertaufgabe zweiter Ordnung mit periodischen Randbedingungen

  $$(*) \qquad -u'' + \lambda^2 u = r(t), \qquad u(1) - u(0) = u_0,\ u'(1) - u'(0) = u_1$$

  eine eindeutige Lösung, welche durch

  $$u(t) = u_0 v_0^{(\lambda)}(t) + u_1 v_1^{(\lambda)}(t) + \int_0^1 G^{(\lambda)}(t,s) r(s)\,ds$$

  gegeben ist. Hierbei ist

  $$\begin{aligned} v_0^{(\lambda)}(t) &:= \frac{1}{2\sinh\lambda}(\sinh\lambda t - \sinh\lambda(1-t)), \\ v_1^{(\lambda)}(t) &:= \frac{1}{\lambda(\cosh\lambda - 1)}(\sinh\lambda t + \sinh\lambda(1-t)), \end{aligned}$$

weiter ist $G^{(\lambda)}$ die Greensche Funktion zu (L_λ, R), wobei

$$L_\lambda u := -u'' + \lambda^2 u, \qquad Ru := \begin{pmatrix} R_1 u \\ R_2 u \end{pmatrix} := \begin{pmatrix} u(1) - u(0) \\ u'(1) - u'(0) \end{pmatrix}.$$

Die Greensche $G^{(\lambda)}$ Funktion existiert zu (L_λ, R) (da die die homogene Aufgabe $L_\lambda u = 0$, $Ru = 0$ nur trivial lösbar ist) und ist bekanntlich durch die folgenden Eigenschaften charakterisiert:

1. $G^{(\lambda)}\colon [0,1] \times [0,1] \longrightarrow \mathbb{R}$ ist stetig.
2. In jedem der beiden Dreiecke

   $$\Delta_1 := \{(t,s) : 0 \le s \le t \le 1\}, \qquad \Delta_2 := \{(t,s) : 0 \le t \le s \le 1\}$$

   existieren die partiellen Ableitungen $G_t^{(\lambda)}$, $G_{tt}^{(\lambda)}$ und sind stetig, wobei auf der Diagonalen des Quadrats $[0,1] \times [0,1]$ die dem Dreieck entsprechende einseitige Ableitung zu nehmen ist.
3. Bei festem $s \in [0,1]$ ist $LG^{(\lambda)}(t,s) = 0$ für alle $t \in [0,1] \setminus \{s\}$.
4. Es gilt die Sprungbeziehung

   $$G_t^{(\lambda)}(t+0,t) - G_t^{(\lambda)}(t-0,t) = -1 \quad \text{für alle } t \in (0,1).$$

5. Es ist $RG^{(\lambda)}(\cdot, s) = 0$ für alle $s \in (0,1)$.

Die Greensche Funktion $G^{(\lambda)}$ ist durch

$$G^{(\lambda)}(t,s) = \frac{1}{2\lambda(\cosh\lambda - 1)} \begin{cases} \sinh\lambda(t-s) + \sinh\lambda(1-t+s), & (s,t) \in \Delta_1, \\ \sinh\lambda(s-t) + \sinh\lambda(1-s+t), & (s,t) \in \Delta_2, \end{cases}$$

gegeben, da die charakterisierenden Eigenschaften 1.–5. erfüllt sind.

Man beachte, dass die Greensche Funktion $G^{(\lambda)}$ auf $[0,1] \times [0,1]$ offensichtlich für jedes $\lambda > 0$ positiv ist. Die Funktion $v_0^{(\lambda)}$ wechselt auf dem Intervall $[0,1]$ das Vorzeichen, während $v_1^{(\lambda)}$ auf $[0,1]$ positiv ist. Daher kann die folgende Aussage genau wie Satz 12.4 mit Hilfe des Schauderschen Fixpunktsatzes bewiesen werden.

- *Gegeben sei die nichtlineare Randwertaufgabe zweiter Ordnung mit periodischen Randbedingungen*

  (P) $$-u'' = f(u,t), \qquad u(0) - u(1) = u_0,\ u'(0) - u'(1) = u_1.$$

  Hierbei seien $u_0, u_1 \in \mathbb{R}$ vorgegeben, ferner sei $f\colon \mathbb{R} \times [0,1] \longrightarrow \mathbb{R}$ stetig sowie bezüglich der ersten Variablen stetig differenzierbar. Es mögen $\alpha, \beta \in C^2[0,1]$ existieren mit

  1. *$\alpha(t) \le \beta(t)$ für alle $t \in [0,1]$,*
  2. *$-\alpha''(t) - f(\alpha(t),t) \le 0 \le -\beta''(t) - f(\beta(t),t)$ für alle $t \in [0,1]$,*

3. Es ist

$$\alpha(1) - \alpha(0) = u_0 = \beta(1) - \beta(0)$$

und

$$\alpha'(1) - \alpha'(0) \leq u_1 \leq \beta'(1) - \beta'(0).$$

Dann existiert eine Lösung $u \in C^2[0,1]$ von (P) mit $\alpha(t) \leq u(t) \leq \beta(t)$ für alle $t \in [0,1]$.

Ganz entscheidend wurde in Satz 12.4 und obiger Bemerkung ausgenutzt, dass die nichtlineare Funktion $f(u,t)$ in der Randwertaufgabe (P) *nicht* von der Ableitung u' abhängt. Durch eine Anwendung der Ergebnisse im folgenden Unterabschnitt können wir uns von dieser Einschränkung zum Teil befreien. □

12.4 Nichtlineare Randwertaufgaben bei Systemen von Differentialgleichungen erster Ordnung

Wir betrachten in diesem Unterabschnitt, in dem wir uns im wesentlichen an J. Werner (1969) halten, nichtlineare Randwertaufgaben für ein System von n Differentialgleichungen erster Ordnung der Form

$$\text{(P)} \qquad -z' = f(z,t), \qquad Rz := Mz(a) + Nz(b) = \gamma.$$

Hierbei sind $M, N \in \mathbb{R}^{n\times n}$,

$$f(z,t) = \begin{pmatrix} f_1(z_1,\ldots,z_n,t) \\ \vdots \\ f_n(z_1,\ldots,z_n,t) \end{pmatrix}, \qquad \gamma = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{pmatrix}$$

vorgegeben. Die Idee bei der Behandlung von (P) besteht darin, auf beiden Seiten des Differentialgleichungssystem einen geeigneten linearen Term $C(t)z$ (mit $C \in C_{n\times n}[a,b]$, es sei also $C\colon [a,b] \longrightarrow \mathbb{R}^{n\times n}$ stetig) dazu zu addieren, um auf diese Weise eine Art Monotonie der rechten Seite zu erzwingen, und (P) anschließend in eine Fixpunktaufgabe umzuformulieren. Dies ist möglich, falls die homogene Aufgabe

$$-z' + Cz = 0, \quad Rz = 0$$

nur die triviale Lösung $z = 0$ besitzt. Grundlage ist der folgende Satz, siehe z. B. E. A. Coddington, N. Levinson (1955, S. 204) oder auch H. Werner, H. Arndt (1986, S. 262 ff.).

Satz 12.5 *Gegeben sei die lineare Randwertaufgabe*

$$\text{(L)} \qquad Lz := -z' + Cz = f, \qquad Rz := Mz(a) + Nz(b) = \gamma,$$

wobei $\in C_{n\times n}[a,b]$, $f \in C_n[a,b]$ (die Abbildung $f\colon [a,b] \longrightarrow \mathbb{R}^n$ sei also stetig), $M, N \in \mathbb{R}^{n\times n}$ und $\gamma \in \mathbb{R}^n$. Die homogene Aufgabe $Lz = 0$, $Rz = 0$ besitze nur die triviale Lösung $z = 0$. Sei $\Phi \in C^1_{n\times n}[a,b]$ ein Fundamentalsystem von $Lz = 0$, die Spalten

von Φ also linear unabhängige Lösungen von $Lz = 0$. Dann besitzt (L) eine eindeutige Lösung, welche in der Form

$$(*) \qquad z(t) = g(t) + \int_a^b G(t,s) f(s)\, ds$$

dargestellt werden kann. Hierbei ist

$$g(t) := \Phi(t)[M\Phi(a) + N\Phi(b)]^{-1}\gamma$$

die Lösung von $Lz = 0$, $Rz = \gamma$ und

$$G(t,s) := \Phi(t) \left\{ \begin{array}{c} [M\Phi(a) + N\Phi(b)]^{-1} N\Phi(b) - I \\ [M\Phi(a) + N\Phi(b)]^{-1} N\Phi(b) \end{array} \right\} \Phi(s)^{-1} \quad \begin{array}{l} s < t, \\ t < s. \end{array}$$

die Greensche Matrix zu (L, R).

Beweis: Zunächst müssen wir uns überlegen, dass $M\Phi(a) + N\Phi(b)$ nichtsingulär ist, weil andernfalls g und die Greensche Matrix G nicht definiert wären. Dies ist aber einfach einzusehen. Denn wenn ein $x \in \mathbb{R}^n \setminus \{0\}$ mit $[M\Phi(a) + N\Phi(b)]x = 0$ existieren würde, so wäre $\Phi(t)x$ eine nichttriviale Lösung von $Lz = 0$, $Rz = 0$, ein Widerspruch zur Annahme. Dass unter dieser Annahme die inhomogene Randwertaufgabe höchstens eine Lösung besitzen kann, ist klar. Daher bleibt zu zeigen, dass durch $(*)$ eine Lösung von (L) gegeben ist. Wegen $Lg = 0$, $Rg = \gamma$ genügt es hierfür wiederum nachzuweisen, dass

$$x(t) := \int_a^b G(t,s) f(s)\, ds$$

eine Lösung von $Lz = f$, $Rz = 0$ ist. Wegen

$$\begin{aligned} x(t) &= \int_a^b G(t,s) f(s)\, ds \\ &= \int_a^t G(t,s) f(s)\, ds + \int_t^b G(t,s) f(s)\, ds \\ &= \Phi(t)\{[M\Phi(a) + N\Phi(b)]^{-1} N\Phi(b) - I\} \int_a^t \Phi(s)^{-1} f(s)\, ds \\ &\qquad + \Phi(t)[M\Phi(a) + N\Phi(b)]^{-1} N\Phi(b) \int_t^b \Phi(s)^{-1} f(s)\, ds \end{aligned}$$

ist

$$\begin{aligned} Lx(t) &= -x'(t) + C(t)x(t) \\ &= -\Phi'(t)\{[M\phi(a) + N\Phi(b)]^{-1} N\Phi(b) - I\} \int_a^t \Phi(s)^{-1} f(s)\, ds \\ &\qquad - \Phi(t)\{[M\phi(a) + N\Phi(b)]^{-1} N\Phi(b) - I\}\Phi(t)^{-1} f(t) \\ &\qquad - \Phi'(t)[M\Phi(a) + N\Phi(b)]^{-1} N\Phi(b) \int_t^b \Phi(s)^{-1} f(s)\, ds \end{aligned}$$

$$
\begin{aligned}
&+ \Phi(t)[M\Phi(a) + N\Phi(b)]^{-1} N\Phi(b)\Phi(t)^{-1} f(t) \\
&+ C(t)x(t) \\
= \; & -C(t)\Phi(t)\{[M\phi(a) + N\Phi(b)]^{-1} N\Phi(b) - I\} \int_a^t \Phi(s)^{-1} f(s)\, ds \\
&- C(t)\Phi(t)[M\Phi(a) + N\Phi(b)]^{-1} N\Phi(b) \int_t^b \Phi(s)^{-1} f(s)\, ds \\
&+ C(t)x(t) + f(t) \\
= \; & -C(t)x(t) + C(t)x(t) + f(t) \\
= \; & f(t)
\end{aligned}
$$

und damit $Lx = f$. Weiter ist

$$
\begin{aligned}
Rx &= Mx(a) + Nx(b) \\
&= M\Phi(a)[M\Phi(a) + N\Phi(b)]^{-1} N\Phi(b) \int_a^b \Phi(s)^{-1} f(s)\, ds \\
&\quad + N\Phi(b)\{[M\Phi(a) + N\Phi(b)]^{-1} N\Phi(b) - I\} \int_a^b \Phi(s)^{-1} f(s)\, ds \\
&= [M\Phi(a) + N\Phi(b)][M\Phi(a) + N\Phi(b)]^{-1} N\Phi(b) \int_a^b \Phi(s)^{-1} f(s)\, ds \\
&\quad - N\Phi(b) \int_a^b \Phi(s)^{-1} f(s)\, ds \\
&= 0.
\end{aligned}
$$

Damit ist der Satz bewiesen. □

Ist also die homogene Aufgabe $-z' + Cz = 0$, $Mz(a) + Nz(b) = 0$ nur trivial lösbar, so ist die nichtlineare Randwertaufgabe

$$
\text{(P)} \qquad -z' = f(z,t), \qquad Mz(a) + Nz(b) = \gamma
$$

äquivalent zu der Fixpunktaufgabe für die Abbildung $F\colon C_n[a,b] \longrightarrow C_n[a,b]$ definiert durch

$$
F(x)(t) := g(t) + \int_a^b G(t,s)[\underbrace{f(z(s),s) + C(s)z(s)}_{F_C(z)(s)}]\, ds.
$$

Mit $Lz := -z' + Cz$, $Rz := Mz(a) + Nz(b)$ ist g die Lösung von $Lz = 0$, $Rz = \gamma$ und G die Greensche Funktion zu (L,R). Ferner definieren wir die Abbildung $F_C\colon C_n[a,b] \longrightarrow C_n[a,b]$ durch

$$
F_C(z)(t) := f(z(t),t) + C(t)z(t).
$$

Die Greensche Funktion $G(t,s)$ wird i. Allg. auf $[a,b] \times [a,b]$ nicht elementweise von einem Vorzeichen sein, d. h. die durch

$$
(*) \qquad G(z)(t) := g(t) + \int_a^b G(t,s)z(s)\, ds
$$

definierte (affin lineare) Abbildung $G\colon C_n[a,b] \longrightarrow C_n[a,b]$ wird i. Allg. nicht bezüglich der natürlichen Halbordnung $\leq$ (siehe das nächste Beispiel) auf $C_n[a,b]$ isoton bzw. monoton wachsend sein. Bezüglich einer anderen Halbordnung ist dies aber sehr wohl denkbar.

Definition 12.6 Ist X ein linearer normierter Raum, so heißt eine Menge $O \subset X$ ein *Ordnungskegel*, wenn gilt:

(i) O ist konvex und $O \neq \{0\}$;

(ii) aus $\lambda \geq 0$ und $u \in O$ folgt $\lambda u \in O$;

(iii) aus $u \in O$ und $-u \in O$ folgt $u = 0$, d. h. es ist $O \cap (-O) = \{0\}$.

Der Ordnungskegel O induziert in kanonischer Weise eine *Halbordnung* $\leq$, indem wir $u \leq v$ anstelle von $v - u \in O$ schreiben. Weiter schreiben wir gelegentlich $v \geq u$ statt $u \leq v$. Wenn wir die Abhängigkeit der Halbordnung $\leq$ vom Ordnungskegel O explizit hervorheben wollen, so schreiben wir $\leq_O$ statt $\leq$. Ein Ordnungskegel $O \subset X$ heißt *normal*, wenn eine Konstante $c > 0$ mit

$$0 \leq u \leq v \Longrightarrow \|u\| \leq c\,\|v\|$$

existiert. Sind $u_0, v_0 \in X$ zwei Elemente mit $u_0 \leq v_0$, so heißt

$$[u_0, v_0] := \{w \in X : u_0 \leq w \leq v_0\}$$

das durch u_0, v_0 gegebene *Intervall*[42]. Wenn wir die Abhängigkeit vom Ordnungskegel O bzw. der Halbordnung $\leq$ deutlich machen wollen, so schreiben wir $[u_0, v_0]_O$ bzw. $[u_0, v_0]_\leq$ statt $[u_0, v_0]$.

Beispiel: Eine "natürliche Halbordnung" im $\mathbb{R}^n$ ist durch die komponentenweise Halbordnung $\leq$ gegeben, welche durch den nichtnegativen Orthanten

$$\mathbb{R}^n_+ := \{x = (x_i) \in \mathbb{R}^n : x_i \geq 0\}$$

als Ordnungskegel erzeugt wird. Auf $C_n[a,b]$, dem mit der Maximumnorm versehenen linearen normierten Raum der stetigen Abbildungen von $[a,b]$ in den $\mathbb{R}^n$, ist hierdurch in naheliegender Weise der Ordnungskegel

$$C_{n,+}[a,b] := \{x \in C_n[a,b] : x(t) \in \mathbb{R}^n_+ \text{ für alle } t \in [a,b]\}$$

gegeben. Dieser ist offenbar abgeschlossen und normal. Die zugehörige Halbordnung wird ebenfalls mit $\leq$ bezeichnet.

Natürlich gibt es in $\mathbb{R}^n$ und $C_n[a,b]$ neben $\mathbb{R}^n_+$ bzw. $C_{n,+}[a,b]$ weitere abgeschlossene und normale Ordnungskegel. Ist z. B. $H \in \mathbb{R}^{n\times n}$ nichtsingulär, so ist

$$O_H := \{x \in \mathbb{R}^n : Hx \in \mathbb{R}^n_+\}$$

[42]Ist der Ordnungskegel $O \subset X$ normal, so ist das Intervall $[u_0, v_0]_O$ offenbar beschränkt. Ist der Ordnungskegel $O \subset X$ abgeschlossen, so ist das Intervall $[u_0, v_0]_O$ abgeschlossen.

ein normaler, abgeschlossener Ordnungskegel im $\mathbb{R}^n$. Ist weiter $H\colon C_n[a,b] \longrightarrow C_n[a,b]$ linear, stetig und bijektiv (dann existiert $H^{-1}\colon C_n[a,b] \longrightarrow C_n[a,b]$ und ist natürlich linear und wegen Satz 6.15, dem Satz über die Beschränktheit der Inversen einer linearen, stetigen und bijektiven Abbildung zwischen Banachräumen selbst stetig), so ist

$$O_H := \{x \in C_n[a,b] : H(x) \in C_{n,+}[a,b]\}$$

offenbar ein abgeschlossener, normaler Ordnungskegel in $C_n[a,b]$. Die zugehörige Halbordnung im $\mathbb{R}^n$ bzw. in $C_n[a,b]$ wird jeweils mit $\leq_H$ bezeichnet. Wir sagen, dass die Halbordnung $\leq_H$ in $\mathbb{R}^n$ bzw. $C_n[a,b]$ durch die nichtsinguläre Matrix $H \in \mathbb{R}^{n\times n}$ bzw. die lineare, stetige, bijektive Abbildung $H\colon C_n[a,b] \longrightarrow C_n[a,b]$ *induziert* sei. □

Mit Hilfe des nächsten Satzes über Matrizen (siehe J. Werner (1969, Hilfssatz 3)) kann eine Bedingung dafür angegeben werden, dass es lineare, stetige und bijektive Abbildungen H, K von $C_n[0,1]$ auf sich gibt derart, dass HGK^{-1} bezüglich der natürlichen Halbordnung isoton ist. Hierbei ist $G\colon C_n[0,1] \longrightarrow C_n[0,1]$ durch $(*)$ definiert.

Satz 12.7 *Sei $A \in \mathbb{R}^{n\times n}$. Ist μ ein Eigenwert von A, so sei $\mu \in \mathbb{R}$ und $\mu \notin (0,1)$. Dann existieren nichtsinguläre Matrizen $\mathcal{H}, \mathcal{K} \in \mathbb{R}^{n\times n}$ mit*

$$\mathcal{H}A\mathcal{K}^{-1} \geq 0, \qquad \mathcal{H}(A-I)\mathcal{K}^{-1} \geq 0,$$

d. h. alle Einträge der jeweiligen Matrizen sind nichtnegativ.

Beweis: Durch eine nichtsinguläre Matrix $\tilde{\mathcal{H}} \in \mathbb{R}^{n\times n}$ lässt sich $A \in \mathbb{R}^{n\times n}$ auf Jordansche Normalform transformieren, d. h. es ist

$$\tilde{\mathcal{H}}A\tilde{\mathcal{H}}^{-1} = \operatorname{diag}(J_0, J_1, \ldots, J_s, J_{s+1}, \ldots, J_t).$$

Hierbei ist

$$J_0 := \operatorname{diag}(\mu_1, \ldots, \mu_q)$$

eine $q \times q$-Diagonalmatrix und

$$J_i := \begin{pmatrix} \mu_{q+i} & 1 & \cdots & 0 \\ 0 & \mu_{q+i} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & \mu_{q+i} \end{pmatrix} \in \mathbb{R}^{r_i \times r_i}, \qquad i = 1, \ldots, t.$$

Die μ_j sind die (nach Voraussetzung reellen) Eigenwerte von A und liegen nach Voraussetzung nicht in $(0,1)$. Es sei $\mu_{q+i} \geq 1$, $i = 1, \ldots, s$, und $\mu_{q+i} \leq 0$, $i = s+1, \ldots, t$. Mit I_0 bezeichnen wir die $q\times q$-Einheitsmatrix, mit I_i für $i = 1, \ldots, t$ die $r_i \times r_i$-Einheitsmatrix. Ferner definiere man $M_i \in \mathbb{R}^{r\times r_i}$, $i = s+1, \ldots, t$, als diejenige Matrix, die außer in der Hauptdiagonalen und der oberen Nebendiagonalen nur Nullen als Einträge besitzt, während die Hauptdiagonale aus $(1, -1, \ldots, \mp 1, \pm 1)$ und die obere Nebendiagonale aus $(1, -1, \ldots, \mp 1)$ besteht. Man setze

$$\tilde{J}_i := M_i J_i M_i^{-1}, \quad i = s+1, \ldots, t,$$

und anschließend

$$\mathcal{H} := \text{diag}\,(I_0, I_1, \ldots, I_s, M_{s+1}, \ldots, M_t)\tilde{\mathcal{H}}.$$

Dann ist

$$\mathcal{H}A\mathcal{H}^{-1} = \text{diag}\,(J_0, J_1, \ldots, J_s, \tilde{J}_{s+1}, \ldots, \tilde{J}_t).$$

Schließlich definiere man

$$s_j := \begin{cases} \text{sign}\,(\mu_j) & \text{falls } \mu_j \neq 0, \\ -1 & \text{falls } \mu_j = 0, \end{cases} \qquad j = 1, \ldots, q,$$

setze

$$S := \text{diag}\,(s_1, \ldots, s_q)$$

und gewinne $\mathcal{K}$ aus

$$\mathcal{K} := \text{diag}\,(S^{-1}, I_1, \ldots, I_s, -I_{s+1}, \ldots, -I_t)\mathcal{H},$$

d. h. es ist

$$\mathcal{K}^{-1} = \mathcal{H}^{-1}\text{diag}\,(S, I_1, \ldots, I_s, -I_{s+1}, \ldots, -I_t).$$

Wir wollen uns überlegen, dass mit $\mathcal{H}$ und $\mathcal{K}$ Matrizen gefunden sind, die die gesuchten Eigenschaften besitzen. Es ist

$$\begin{aligned} \mathcal{H}A\mathcal{K}^{-1} &= \mathcal{H}A\mathcal{H}^{-1}\text{diag}\,(S, I_1, \ldots, I_s, -I_{s+1}, \ldots, -I_t) \\ &= \text{diag}\,(J_0, J_1, \ldots, J_s, \tilde{J}_{s+1}, \ldots, \tilde{J}_t)\,\text{diag}\,(S, I_1, \ldots, I_s, -I_{s+1}, \ldots, -I_t) \\ &= \text{diag}\,(J_0 S, J_1, \ldots, J_s, -\tilde{J}_{s+1}, \ldots, -\tilde{J}_t). \end{aligned}$$

Wir wollen uns überlegen, dass in dieser Blockdiagonalmatrix alle Einträge nichtnegativ sind. Zunächst ist

$$J_0 S = \text{diag}\,(\mu_1, \ldots, \mu_q)\,\text{diag}\,(s_1, \ldots, s_q) = \text{diag}\,(|\mu_1|, \ldots, |\mu_q|)$$

trivialerweise nichtnegativ. In den Jordanblöcken $J_1, \ldots, J_s$ sind nur die Haupdiagonale und die obere Nebendiagonale mit von Null verschiedenen Einträgen besetzt. In der Hauptdiagonale sind alle Einträge ≥ 1, während die obere Nebendiagonale mit Einsen besetzt ist. Zu zeigen bleibt also, dass $-\tilde{J}_i \geq 0$, $i = s+1, \ldots, t$. Hierbei ist $\tilde{J}_i = M_i J_i M_i^{-1}$, wobei im Jordanblock J_i die Diagonalelemente durch $\mu_{q+i} \leq 0$ besetzt sind, $i = s+1, \ldots, t$. Um die Notation zu vereinfachen, nehmen wir an, dass J_i und damit auch $\tilde{J}_i$ ein 4×4-Block sei. Dann ist

$$\begin{aligned} -\tilde{J}_i &= -M_i J_i M_i^{-1} \\ &= -\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} \mu_{q+i} & 1 & 0 & 0 \\ 0 & \mu_{q+i} & 1 & 0 \\ 0 & 0 & \mu_{q+i} & 1 \\ 0 & 0 & 0 & \mu_{q+i} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & -1 & -1 & -1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & -1 \end{pmatrix} \\ &= -\begin{pmatrix} \mu_{q+i} & \mu_{q+i}+1 & 1 & 0 \\ 0 & -\mu_{q+i} & -\mu_{q+i}-1 & -1 \\ 0 & 0 & \mu_{q+i} & \mu_{q+i}+1 \\ 0 & 0 & 0 & -\mu_{q+i} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & -1 & -1 & -1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & -1 \end{pmatrix} \end{aligned}$$

$$= -\begin{pmatrix} \mu_{q+i} & -1 & 0 & 0 \\ 0 & \mu_{q+i} & -1 & 0 \\ 0 & 0 & \mu_{q+i} & -1 \\ 0 & 0 & 0 & \mu_{q+i} \end{pmatrix}.$$

Wegen $\mu_{q+i} \leq 0$ ist dies eine Matrix, deren Einträge sämtlich nichtnegativ sind. Der allgemeine Fall ergibt sich analog. Damit ist $\mathcal{H}A\mathcal{K}^{-1} \geq 0$. Weiter ist

$$\begin{aligned} \mathcal{H}(A-I)\mathcal{K}^{-1} &= \mathcal{H}A\mathcal{K}^{-1} - \mathcal{H}\mathcal{K}^{-1} \\ &= \operatorname{diag}(J_0 S, J_1, \ldots, J_s, -\tilde{J}_{s+1}, \ldots, -\tilde{J}_t) \\ &\qquad - \operatorname{diag}(S, I_1, \ldots, I_s, -I_{s+1}, \ldots, -I_t) \\ &= \operatorname{diag}((J_0 - I_0)S, J_1 - I_1, \ldots, J_s - I_s, I_{s+1} - \tilde{J}_{s+1}, \ldots, I_t - \tilde{J}_t). \end{aligned}$$

Alle Einträge dieser Blockdiagonalmatrix sind nichtnegativ! Der erste Block besteht aus der Diagonalmatrix $(J_0 - I_0)S$. Das j-te Diagonalelement d_j dieser Diagonalmatrix ist

$$d_j = (\mu_j - 1)s_j = \begin{cases} \mu_j - 1, & \mu_j \geq 1, \\ 1 - \mu_j, & \mu_j \leq 0, \end{cases} \qquad j = 1, \ldots, q,$$

und folglich nichtnegativ. In den s Blöcken J_i, $i = 1, \ldots, s$, sind die Diagonalelemente ≥ 1 und folglich $J_i - I_i \geq 0$, $i = 1, \ldots, s$. Schließlich wissen wir schon, dass $-\tilde{J}_i \geq 0$, $i = s+1, \ldots, t$, sodass erst recht $I_i - \tilde{J}_i \geq 0$, $i = s+1, \ldots, t$. Also ist $\mathcal{H}(A-I)\mathcal{K}^{-1} \geq 0$. Insgesamt ist der Satz bewiesen. □

Als einfache Folgerung aus dem letzten Satz 12.7 erhalten wir die folgende Aussage:

Satz 12.8 *Seien $C \in C_{n\times n}[a,b]$, $M, N \in \mathbb{R}^{n\times n}$ gegeben. Wie in Satz 12.5 seien hiermit*

$$Lz := -z' + Cz, \qquad Rz := Mz(a) + Nz(b)$$

definiert. Es wird vorausgesetzt, dass die homogene Aufgabe $Lz = 0$, $Rz = 0$ nur trivial lösbar ist. Mit einem Fundamentalsystem Φ von $Lz = 0$ sei, wie in Satz 12.5 angegeben,

$$G(t,s) := \Phi(t) \left\{ \begin{array}{c} [M\Phi(a) + N\Phi(b)]^{-1} N\Phi(b) - I \\ [M\Phi(a) + N\Phi(b)]^{-1} N\Phi(b) \end{array} \right\} \Phi(s)^{-1} \quad \begin{array}{l} s < t, \\ t < s. \end{array}$$

die Greensche Funktion zu (L, R). Es wird vorausgesetzt, dass die Eigenwerte von

$$A := [M\Phi(a) + N\Phi(b)]^{-1} N\Phi(b)$$

reell sind und nicht in $(0,1)$ liegen (was von der Wahl des Fundamentalsystems unabhängig ist). Mit $g \in C_n[a,b]$ sei die Abbildung $G\colon C_n[a,b] \longrightarrow C_n[a,b]$ durch

$$G(z)(t) := g(t) + \int_a^b G(t,s)z(s)\,ds$$

definiert. Dann existieren lineare, stetige und bijektive Abbildungen $H, K\colon C_n[a,b] \longrightarrow C_n[a,b]$ mit der Eigenschaft, dass HGK^{-1} bezüglich der natürlichen Halbordnung in $C_n[a,b]$ isoton ist bzw. die Implikation

$$u \leq_K v \Longrightarrow G(u) \leq_H G(u)$$

gilt.

Beweis: Ist $\tilde{\Phi}$ ebenfalls ein Fundamentalsystem zu $Lz = 0$, so existiert eine nichtsinguläre Matrix $M \in \mathbb{R}^{n\times n}$ mit $\tilde{\Phi} = \Phi X$. Daher hat

$$\tilde{A} := [M\tilde{\Phi}(a) + N\tilde{\Phi}(b)]^{-1}N\tilde{\Phi})b) = M^{-1}AM$$

dieselben Eigenwerte wie A. Wegen Satz 12.7 existieren nichtsinguläre Matrizen $\mathcal{H}, \mathcal{K} \in \mathbb{R}^{n\times n}$ mit $\mathcal{H}A\mathcal{K}^{-1} \geq 0$ und $\mathcal{H}(A-I)\mathcal{K}^{-1} \geq 0$ bzw. der Eigenschaft, dass alle Einträge von $\mathcal{H}A\mathcal{K}^{-1}$ und $\mathcal{H}(A-I)\mathcal{K}^{-1}$ nichtnegativ sind. Nun definiere man $H, K\colon C_n[a,b] \longrightarrow C_n[a,b]$ durch

$$H(z)(t) := \mathcal{H}\Phi(t)^{-1}z(t), \qquad K(z)(t) := \mathcal{K}\Phi(t)^{-1}z(t).$$

Offenbar sind H und K lineare, stetige und bijektive Abbildungen von $C_n[a,b]$ auf $C_n[a,b]$. Sind nun $u, v \in C_n[a,b]$ mit $u \leq v$ (mit der natürlichen Halbordnung $\leq$) und ein beliebiges $t \in [a,b]$ gegeben, so ist

$$\begin{aligned}
HGK^{-1}(u)(t) &= HG\Phi(t)\mathcal{K}^{-1}u(t) \\
&= H\Big(g(t) + \int_a^b G(t,s)\Phi(s)\mathcal{K}^{-1}u(s)\,ds\Big) \\
&= H(g)(t) + \mathcal{H}\Phi(t)^{-1}\int_a^b G(t,s)\Phi(s)\mathcal{K}^{-1}u(s)\,ds \\
&= H(g)(t) + \mathcal{H}\Phi(t)^{-1}\int_a^t G(t,s)\Phi(s)\mathcal{K}^{-1}u(s)\,ds \\
&\qquad + \mathcal{H}\Phi(t)^{-1}\int_t^b G(t,s)\Phi(s)\mathcal{K}^{-1}u(s)\,ds \\
&= H(g)(t) + \mathcal{H}\Phi(t)^{-1}\int_a^t \Phi(t)(A-I)\Phi(s)^{-1}\Phi(s)\mathcal{K}^{-1}u(s)\,ds \\
&\qquad + \mathcal{H}\Phi(t)^{-1}\int_t^b \Phi(t)A\Phi(s)^{-1}\Phi(s)\mathcal{K}^{-1}u(s)\,ds \\
&= H(g)(t) + \underbrace{\mathcal{H}(A-I)\mathcal{K}^{-1}}_{\geq 0}\int_a^t u(s)\,ds + \underbrace{\mathcal{H}A\mathcal{K}^{-1}}_{\geq 0}\int_t^b u(s)\,ds \\
&\leq H(g)(t) + \mathcal{H}(A-I)\mathcal{K}^{-1}\int_a^t v(s)\,ds + \mathcal{H}A\mathcal{K}^{-1}\int_t^b v(s)\,ds \\
&= HGK^{-1}(v)(t).
\end{aligned}$$

Damit ist die Isotonie von HGK^{-1} bezüglich der natürlichen Halbordnung auf $C_n[a,b]$ bewiesen. Sind nun $u, v \in C_n[a,b]$ mit $u \leq_K v$ bzw. $K(u) \leq K(v)$ gegeben, so folgt wegen der Isotonie von HGK^{-1}, dass $HG(u) \leq HG(v)$ bzw. $G(u) \leq_H G(v)$, womit der Satz insgesamt bewiesen ist. □

Die folgenden Aussagen über Einschließungssätze bei nichtlinearen Randwertaufgaben für Systeme von Differentialgleichungen erster Ordnung werden sich als Folgerungen des folgenden allgemeinen Satzes erweisen.

Satz 12.9 *Sei X ein Banachraum, $O \subset X$ ein normaler, abgeschlossener Ordnungskegel, und $\leq$ die hierdurch induzierte Halbordnung. Seien $u_0, v_0 \in X$ zwei Elemente mit $u_0 \leq v_0$ und $[u_0, v_0] \subset X$ das hiervon erzeugte Intervall. Die Abbildung $F\colon [u_0, v_0] \longrightarrow X$ sei kompakt*[43] *und monoton wachsend auf $[u_0, v_0]$, d. h. für $u, v \in [u_0, v_0]$ mit $u \leq v$ ist $F(u) \leq F(v)$. Ferner sei $u_0 \leq F(u_0)$ und $F(v_0) \leq v_0$. Dann existiert mindestens ein $u^* \in [u_0, v_0]$ mit $F(u^*) = u^*$.*

Beweis: Die Menge $[u_0, v_0]$ ist nichtleer, konvex, abgeschlossen und beschränkt (da der Ordnungskegel abgeschlossen und normal ist). Die Existenz mindestens eines Fixpunktes $u^* \in [u_0, v_0]$ von F folgt aus dem Schauderschen Fixpunktsatz (Satz 10.16), wenn wir noch $F([u_0, v_0]) \subset [u_0, v_0]$ zeigen. Ist $u \in [u_0, v_o]$ bzw. $u_0 \leq u \leq v_0$, so folgt

$$u_0 \leq F(u_0) \leq F(u) \leq F(v_0) \leq v_0$$

bzw. $F(u) \in [u_0, v_0]$ und damit $F([u_0, v_0]) \subset [u_0, v_0]$. □

Als eine Anwendung des letzten Satzes überlegen wir uns die Gültigkeit des folgenden Satzes (siehe J. Werner (1969, S. 28)).

Satz 12.10 *Gegeben sei (wie zu Beginn dieses Abschnitts) die nichtlineare Randwertaufgabe*

$$(P) \qquad -z' = f(z,t), \qquad Rz := Mz(a) + Nz(b) = \gamma,$$

mit $f\colon \mathbb{R}^n \times [a,b] \longrightarrow \mathbb{R}^n$. Sei $C \in C_{n\times n}[a,b]$ und die Aufgabe $-z' + C(t)z = 0$, $Rz = 0$ nur trivial lösbar ist. Durch die linearen, stetigen und bijektiven Abbildungen

$$H, K\colon C_n[a,b] \longrightarrow C_n[a,b]$$

seien auf $C_n[a,b]$ die Halbordnungen $\leq_H$ und $\leq_K$ gegeben, durch die nichtsinguläre Matrix $P \in \mathbb{R}^{n\times n}$ auf $\mathbb{R}^n$ die Halbordnung $\leq_P$, wobei die folgenden Eigenschaften erfüllt seien:

(a) Ist $w \in C_n^1[a,b]$, $-w' + Cw \geq_K 0$ und $Rw = 0$, so ist $w \geq_H 0$.

(b) Ist $w \in C_n^1[a,b]$, $-w' + Cw = 0$ und $Rw \geq_P 0$, so ist $w \geq_H 0$.

Weiter mögen $\alpha, \beta \in C_n^1[a,b]$ existieren mit

$$\alpha \leq_H \beta, \qquad -\alpha' - f(\alpha, \cdot) \leq_K 0 \leq_K -\beta' - f(\beta, \cdot), \qquad R\alpha \leq_P \gamma \leq_P R\beta$$

und der Eigenschaft, dass

$$\alpha \leq_H u \leq_H v \leq_H \beta \Longrightarrow F_C(u) \leq_K F_C(v).$$

Die Abbildung f sei auf $\Omega := \{(z,t) \in \mathbb{R}^n \times [a,b] : \alpha \leq_H z \leq_H \beta\}$ stetig[44]*. Dann besitzt (P) eine Lösung z mit $\alpha \leq_H z \leq_H \beta$.*

[43] Siehe Definition 10.6.

[44] Wenn $\alpha \leq_H z \leq_H \beta$ mit $z \in \mathbb{R}^n$ und $\alpha, \beta \in C_n[a,b]$ geschrieben wird, so soll die heißen, dass die konstante Funktion $f(t) := z$ im Intervall $[\alpha, \beta]_{\leq_H}$ liegt.

Beweis: Wir wenden Satz 12.9 an und definieren hierzu den Banachraum $X := C_n[a,b]$ versehen mit der Maximumnorm, $O := O_H$ sei der von $H\colon C_n[a,b] \longrightarrow C_n[a,b]$ erzeugte Ordnungskegel, $u_0 := \alpha$ und $v_0 := \beta$. Die Abbildung $F\colon [u_0,v_0] \longrightarrow X$ sei definiert durch

$$F(z)(t) := g(t) + \int_a^b G(t,s)[\underbrace{f(z(s),s) + C(s)z(s)}_{F_C(z)(s)}]\,ds,$$

wobei g die Lösung von $-z' + C(t)z = 0$, $Rz = \gamma$ und G die Greensche Funktion zu (L,R) (siehe Satz 12.8) ist, wobei $Lz := -z' + Cz$. Da F auf $[u_0,v_0]$ eine kompakte Abbildung ist, bleibt zu zeigen, dass F auf $[u_0,v_0]$ isoton bzw. monoton wachsend bezüglich der Halbordnung $\leq_H$ ist. Seien $u,v \in C_n[0,1]$ mit $u_0 \leq_H u \leq_H v \leq_H v_0$ gegeben. Wir setzen $w := F(v) - F(u)$. Dann ist

$$-w' + Cw = F_C(v) - F_C(u) \geq_K 0, \qquad Rw = 0.$$

Nach Voraussetzung (a) folgt hieraus $w \geq_H 0$ bzw. $F(u) \leq_H F(v)$. Daher ist F isoton auf $[u_0,v_0] = [\alpha,\beta]$. Zu zeigen bleibt, dass $u_0 \leq_H F(u_0)$ bzw. $\alpha \leq_H F(\alpha)$ und $F(v_0) \leq_H v_0$ bzw. $F(\beta) \leq_H \beta$. Zur Abkürzung definiere man $z_\alpha, v_\alpha \in C_n^1[0,1]$ durch

$$z_\alpha(t) := g(t) + \int_a^b G(t,s)[-\alpha'(s) + C(s)\alpha(s)]\,ds, \qquad v_\alpha(t) := F(\alpha)(t) - z_\alpha(t).$$

Dann ist

$$-v_\alpha' + Cv_\alpha = F_C(\alpha) - [-\alpha' + C\alpha] = \alpha' + f(\alpha,\cdot) \geq_K 0, \qquad Rv_\alpha = 0.$$

Wegen Voraussetzung (a) ist $v_\alpha \geq_H 0$ bzw. $z_\alpha \leq_H F(\alpha)$. Ferner ist $\alpha \leq_H z_\alpha$. Denn setzt man $w_\alpha := z_\alpha - \alpha$, so ist $-w_\alpha' + Cw_\alpha = 0$ und $Rw_\alpha = \gamma - R_\alpha \geq_P 0$. Wegen Voraussetzung (b) folgt hieraus $w \geq_K 0$ bzw. $\alpha \leq_K z_\alpha$, womit $\alpha \leq_K F(\alpha)$ bewiesen ist. Da entsprechend auch $F(\beta) \leq_K \beta$ gezeigt werden kann, ist der Satz bewiesen. □

Der nächste Satz ist eine Folgerung des vorigen und gibt an, wann man mit Hilfe eines Fundamentalsystems $\Phi(t)$ zu $-z' + Cz = 0$ geeignete Halbordnungen $\leq_H$, $\leq_K$ und $\leq_P$ finden kann, für die die Bedingungen (a) und (b) aus Satz 12.10 erfüllt sind.

Satz 12.11 *Gegeben sei die nichtlineare Randwertaufgabe*

$$(P) \qquad -z' = f(z,t), \qquad Rz := Mz(a) + Nz(b) = \gamma,$$

mit $f\colon \mathbb{R}^n \times [a,b] \longrightarrow \mathbb{R}^n$. Sei $C \in C_{n\times n}[a,b]$ mit der Eigenschaft, dass die Aufgabe $-z' + C(t)z = 0$, $Rz = 0$ nur trivial lösbar ist. Für ein Fundamentalsystem $\Phi(t)$ von $Lz := -z' + Cz = 0$ habe $A := [M\Phi(a) + N\Phi(b)]^{-1}N\Phi(b)$ nur reelle Eigenwerte, die nicht in $(0,1)$ liegen. Seien $\mathcal{H},\mathcal{K} \in \mathbb{R}^{n\times n}$ nichtsinguläre Matrizen mit $\mathcal{H}A\mathcal{K}^{-1} \geq 0$ und $\mathcal{H}(A-I)\mathcal{K}^{-1} \geq 0$ (siehe Satz 12.7). Hiermit definiere man die linearen, stetigen, bijektiven Abbildungen $H,K\colon C_n[a,b] \longrightarrow C_n[a,b]$ durch

$$H(z)(t) := \mathcal{H}\Phi(t)^{-1}z(t), \qquad K(z)(t) := \mathcal{K}\Phi(t)^{-1}z(t).$$

Mit $\leq_H$ bzw. $\leq_K$ seien die induzierten Halbordnungen auf $C_n[a,b]$ bezeichnet. Schließlich sei die nichtsinguläre Matrix $P \in \mathbb{R}^{n\times n}$ durch

$$P := \mathcal{H}[M\Phi(a) + N\Phi(b)]^{-1}$$

definiert, $\leq_P$ bezeichne die induzierte Halbordung auf $\mathbb{R}^n$. Dann sind die Bedingungen (a) und (b) in Satz 12.10 erfüllt. Erfüllen also α, β sowie f die Voraussetzungen von Satz 12.10, so besitzt (P) eine Lösung z mit $\alpha \leq_H \leq z \leq_H \beta$.

Beweis: Sei $w \in C_n^1[a,b]$, $x := -w' + Cw \geq_K 0$ und $Rw = 0$. Bezeichnet

$$G(t,s) := \Phi(t) \left\{ \begin{array}{c} (A-I) \\ A \end{array} \right\} \Phi(s)^{-1} \qquad \begin{array}{l} s < t, \\ t < s. \end{array}$$

die Greensche Funktion zu (L,R), so ist

$$w(t) = \int_a^b G(t,s)x(s)\,ds,$$

für jedes $t \in [a,b]$ ist daher

$$\begin{aligned} H(w)(t) &= \mathcal{H}\Phi(t)^{-1} \int_a^b G(t,s)x(s)\,ds \\ &= \mathcal{H}\Phi(t)^{-1}\Big(\int_a^t \Phi(t)(A-I)\Phi(s)^{-1}x(s)\,ds + \int_t^b \Phi(t)A\Phi(s)^{-1}x(s)\,ds\Big) \\ &= \int_a^t \underbrace{\mathcal{H}(A-I)\mathcal{K}^{-1}}_{\geq 0}\underbrace{K(x)(s)}_{\geq 0}\,ds + \int_t^b \underbrace{\mathcal{H}A\mathcal{K}^{-1}}_{\geq 0}\underbrace{K(x)(s)}_{\geq 0}\,ds \\ &\geq 0, \end{aligned}$$

also $w \geq_H 0$. Damit ist Bedingung (a) in Satz 12.10 erfüllt. Da $\Phi(\cdot)$ ein Fundamentalsystem zu $-z' + Cz = 0$ ist, existiert ein $\xi \in \mathbb{R}^n$ mit $w(t) = \Phi(t)\xi$. Folglich ist

$$H(w)(t) = \mathcal{H}\Phi(t)^{-1}\Phi(t)\xi = \mathcal{H}\xi = P[M\Phi(a) + N\Phi(b)]\xi = PRw \geq 0,$$

also $w \geq_H 0$. Damit ist auch Bedingung (b) von Satz 12.10 erfüllt und der Satz bewiesen. □

12.5 Nichtlineare erzwungene Schwingungen

Die Ergebnisse des vorigen Unterabschnitts sollen nun auf Differentialgleichungssysteme mit periodischen Randbedingungen angewandt werden. Gesucht sei eine Lösung von

$$\text{(P)} \qquad -z' = f(z,t), \qquad Rz := z(b) - z(a) = \gamma.$$

Im allgemeinen ist hier $\gamma = 0$, die gesuchte Lösung von (P) hat in diesem Fall also an den Intervallenden des vorgegebenem Intervalls $[a,b]$ den gleichen Wert. Wir lassen aber zunächst ein beliebiges $\gamma \in \mathbb{R}^n$ zu.

Satz 12.12 *Gegeben sei das nichtlineare Differentialgleichungssystem mit periodischen Randbedingungen (P), wobei* $f\colon \mathbb{R}^n \times [a,b] \longrightarrow \mathbb{R}^n$ *und* $\gamma \in \mathbb{R}^n$. *Sei* $C \in \mathbb{R}^{n\times n}$ *eine durch* $\mathcal{H} \in \mathbb{R}^{n\times n}$ *diagonalisierbare Matrix mit reellen Eigenwerten* $\lambda_i \neq 0$, $i = 1,\ldots,n$, *also* $\mathcal{H}C\mathcal{H}^{-1} = \operatorname{diag}(\lambda_i)$. *Dann gilt:*

(a) Die homogene Aufgabe $-z' + Cz = 0$, $Rz = 0$ *ist nur trivial lösbar.*

(b) Sei $\mathcal{K} := \operatorname{diag}(\operatorname{sign}(\lambda_i))\mathcal{H}$. *Hiermit definiere man (wie in Satz 12.11) die linearen, stetigen, bijektiven Abbildungen* $H, K\colon C_n[a,b] \longrightarrow C_n[a,b]$ *durch*

$$H(z)(t) := \mathcal{H}e^{-Ct}z(t), \qquad K(z)(t) := \mathcal{K}e^{-Ct}z(t).$$

Die hierdurch induzierten Halbordnungen auf $C_n[a,b]$ *seien mit* $\leq_H$ *bzw.* $\leq_K$ *bezeichnet. Weiter sei*

$$P := \mathcal{H}[e^{Cb} - e^{Ca}]^{-1}$$

und $\leq_P$ *die von* P *induzierte Halbordnung auf dem* $\mathbb{R}^n$. *Sind dann* $\alpha, \beta \in C_n^1[a,b]$ *stetig differenzierbare Vektorfunktionen mit*

$$\alpha \leq_H \beta, \qquad -\alpha' - f(\alpha,\cdot) \leq_K 0 \leq_K -\beta' - f(\beta,\cdot), \qquad R\alpha \leq_P \gamma \leq_P R\beta$$

und der Eigenschaft, dass

$$\alpha \leq_H u \leq_H v \leq_H \beta \Longrightarrow F_C(u) \leq_K F_C(v),$$

wobei $F_C\colon C_n[a,b] \longrightarrow C_n[a,b]$ *definiert ist durch*

$$F_C(u)(t) := f(u(t),t) + Cu(t),$$

ist ferner die Abbildung f *auf* $\Omega := \{(z,t) \in \mathbb{R}^n \times [a,b] : \alpha \leq_H z \leq_H \beta\}$ *stetig, so besitzt (P) eine Lösung* z *mit* $\alpha \leq_H z \leq_H \beta$.

Beweis: Durch $\Phi(t) := e^{Ct}$ ist ein Fundamentalsystem des linearen Differentialgleichungssystems $-z' + Cz = 0$ gegeben. Hätte die homogene Aufgabe $-z' + Cz = 0$, $Rz = 0$ eine nichttriviale Lösung, so existierte ein $\xi \in \mathbb{R}^n \setminus \{0\}$ mit $(e^{C(b-a)} - I)\xi = 0$. Wegen $\mathcal{H}C\mathcal{H}^{-1} = \operatorname{diag}(\lambda_i)$ mit reellen $\lambda_i \neq 0$, $i = 1,\ldots,n$, wäre

$$0 = (e^{C(b-a)} - I)\xi = \mathcal{H}^{-1}\operatorname{diag}(\underbrace{e^{\lambda_i(b-a)} - 1}_{\neq 0})\mathcal{H}\xi$$

und damit $\mathcal{H}\xi = 0$ und $\xi = 0$, ein Widerspruch. Damit ist (a) bewiesen. Nun definieren wir

$$\begin{aligned} A &:= [\Phi(b) - \Phi(a)]^{-1}\Phi(b) \\ &= [e^{Cb} - e^{Ca}]^{-1}e^{Cb} \\ &= [I - e^{-C(b-a)}]^{-1} \\ &= \mathcal{H}^{-1}\operatorname{diag}\Big(\frac{1}{1 - e^{-\lambda_i(b-a)}}\Big)\mathcal{H}. \end{aligned}$$

Die Eigenwerte von A sind also $1/(1-e^{-\lambda_i(b-a)})$, $i=1,\ldots,n$. Diese sind reell und liegen nicht in $(0,1)$. Wegen $\mathcal{K}=\text{diag}\,(\text{sign}\,(\lambda_i))\mathcal{H}$ ist dann

$$\mathcal{H}A\mathcal{K}^{-1}=\text{diag}\,\Big(\frac{\text{sign}\,(\lambda_i}{1-e^{-\lambda_i(b-a)}}\Big)=\text{diag}\,\Big(\frac{1}{|1-e^{-\lambda_i(b-a)}|}\Big)\geq 0$$

und

$$\mathcal{H}(A-I)\mathcal{K}^{-1}=\text{diag}\,\Big(\frac{e^{-\lambda_i(b-a)}}{|1-e^{-\lambda_i(b-a)}|}\Big)\geq 0.$$

Die Behauptung des Satzes folgt dann aus Satz 12.11. □

Bemerkung: Mit den Bezeichnungen von Satz 12.12 ist die Abbildung $H\colon C_n[a,b]\longrightarrow C_n[a,b]$ durch

$$H(z)(t)=\mathcal{H}e^{-Ct}z(t)=\text{diag}\,(e^{-\lambda_i t})\mathcal{H}z(t)$$

gegeben. Für $K\colon C_n[a,b]\longrightarrow C_n[a,b]$ gilt entsprechend

$$K(z)(t)=\mathcal{K}e^{-Ct}z(t)=\text{diag}\,(\text{sign}\,(\lambda_i))\mathcal{H}e^{-Ct}z(t)=\text{diag}\,(e^{-\lambda_i t})\mathcal{K}z(t).$$

Daher werden die Halbordnungen $\leq_H$ bzw. $\leq_K$ auf $C_n[a,b]$ schon von $\mathcal{H}$ bzw. $\mathcal{K}$ erzeugt. Wegen

$$\begin{aligned}P &= \mathcal{H}[e^{Cb}-e^{Ca}]^{-1}\\ &= \text{diag}\,\Big(\frac{1}{e^{\lambda_i b}-e^{\lambda_i a}}\Big)\mathcal{H}\\ &= \text{diag}\,\Big(\frac{e^{-\lambda_i a}}{e^{\lambda_i(b-a)}-1}\Big)\text{diag}\,(\text{sign}\,(\lambda_i))\mathcal{K}\\ &= \text{diag}\,\Big(\frac{e^{-\lambda_i a}}{|e^{\lambda_i(b-a)}-1|}\Big)\mathcal{K}\end{aligned}$$

wird die Halbordnung $\leq_P$ auf dem $\mathbb{R}^n$ schon von $\mathcal{K}$ erzeugt. □

Bei den jetzt folgenden Anwendungen von Satz 12.12 auf Systeme von zwei Differentialgleichungen erster Ordnung bzw. eine Differentialgleichung zweiter Ordnung sei das zugrunde liegende Intervall $[a,b]$ durch $[0,\omega]$ mit $\omega>0$ gegeben.

Satz 12.13 *Gegeben sei die periodische Randwertaufgabe*

$$\text{(P)}\qquad \begin{cases} -z'=-\begin{pmatrix} x'\\ y'\end{pmatrix}=\begin{pmatrix} F(x,t)-y\\ g(x,t)\end{pmatrix}=f(z,t),\\[2ex] Rz:=z(\omega)-z(0)=\begin{pmatrix} x(\omega)-x(0)\\ y(\omega)-y(0)\end{pmatrix}=0.\end{cases}$$

Seien $\alpha,\beta\in C^2[0,\omega]$ zwei Funktionen mit

1. *$\alpha(t)\leq\beta(t)$ für alle $t\in[0,\omega]$.*
2. *$\alpha(0)=\alpha(\omega)$, $\alpha'(0)\geq\alpha'(\omega)$, $\beta(0)=\beta(\omega)$, $\beta'(0)\leq\beta'(\omega)$.*

3. $-\alpha''(t) - \dfrac{d}{dt}F(\alpha(t),t) - g(\alpha(t),t) \le 0 \le -\beta''(t) - \dfrac{d}{dt}F(\beta(t),t) - g(\beta(t),t)$ *für alle* $t \in [0,\omega]$.

F und g seien auf

$$\Omega := \{(x,t) \in \mathbb{R} \times [0,\omega] : \alpha(t) \le x \le \beta(t)\}$$

stetig und nach dem ersten Argument x stetig differenzierbar. Außerdem sei

$$F(\alpha(0),0) = F(\alpha(0),\omega), \qquad F(\beta(0),0) = F(\beta(0),\omega)$$

und

$$F(\alpha(\cdot),\cdot), F(\beta(\cdot),\cdot) \in C^1[0,\omega].$$

Dann besitzt (P) eine Lösung $z = (x,y)^T$ mit $z(0) = z(\omega)$ und $\alpha(t) \le x(t) \le \beta(t)$ für alle $t \in [0,\omega]$.

Beweis: Wir wenden Satz 12.12 unter Berücksichtigung der anschließenden Bemerkung mit

$$C := \begin{pmatrix} \lambda_1+\lambda_2 & 1 \\ -\lambda_1\lambda_2 & 0 \end{pmatrix}, \qquad \mathcal{H} := \begin{pmatrix} \lambda_1 & 1 \\ -\lambda_2 & -1 \end{pmatrix}$$

an, wobei λ_1, λ_2 mit $\lambda_2 < 0 < \lambda_1$ noch passend zu wählende reelle Zahlen sind. In der Tat ist hiermit

$$\mathcal{H}C\mathcal{H}^{-1} = \begin{pmatrix} \lambda_1 & 1 \\ -\lambda_2 & -1 \end{pmatrix} \begin{pmatrix} \lambda_1+\lambda_2 & 1 \\ -\lambda_1\lambda_2 & 0 \end{pmatrix} \frac{1}{\lambda_2-\lambda_1} \begin{pmatrix} -1 & -1 \\ \lambda_2 & \lambda_1 \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}.$$

Dann ist

$$\mathcal{K} := \begin{pmatrix} \operatorname{sign}(\lambda_1) & 0 \\ 0 & \operatorname{sign}(\lambda_2) \end{pmatrix} \mathcal{H} = \begin{pmatrix} \lambda_1 & 1 \\ \lambda_2 & 1 \end{pmatrix}.$$

Wir definieren $A, B \in C_2^1[0,\omega]$ durch

$$A(t) := \begin{pmatrix} \alpha(t) \\ \alpha'(t) + F(\alpha(t),t) \end{pmatrix}, \qquad B(t) := \begin{pmatrix} \beta(t) \\ \beta'(t) + F(\beta(t),t) \end{pmatrix}$$

und zeigen, dass für geeignete λ_1, λ_2 die Voraussetzungen von Satz 12.12 (mit A, B statt α, β) erfüllt sind. Man wähle $\lambda_1 > 0$ so groß und $\lambda_2 < 0$ so klein, dass für alle $(x_1,t), (x_2,t) \in \Omega$ mit $x_1 \le x_2$ gilt:

$$\lambda_1^2(x_2-x_1) + \lambda_1[F(x_2,t) - F(x_1,t)] + g(x_2,t) - g(x_1,t) \ge 0, \tag{1}$$

$$\lambda_2^2(x_2-x_1) + \lambda_2[F(x_2,t) - F(x_1,t)] + g(x_2,t) - g(x_1,t) \ge 0, \tag{2}$$

und

$$\lambda_1[\beta(\omega) - \alpha(\omega)] + [\beta'(\omega) - \alpha'(\omega) + F(\beta(\omega),\omega) - F(\alpha(\omega),\omega)] \ge 0, \tag{3}$$

$$-\lambda_2[\beta(0) - \alpha(0)] - [\beta'(0) - \alpha'(0) + F(\beta(0),0) - F(\alpha(0),0)] \ge 0. \tag{4}$$

Bei (3) und (4) beachte man, dass wegen der von α und β geforderten Randbedingungen in 2. gilt: Es ist $\alpha(0) = \beta(0)$ genau dann, wenn $\alpha(\omega) = \beta(\omega)$. Ist dies der Fall, so ist

$$0 \leq \beta'(0) - \alpha'(0) \leq \beta'(\omega) - \alpha'(\omega) \leq 0,$$

also $\alpha'(0) = \beta'(0)$ und $\alpha'(\omega) = \beta'(\omega)$. Die Bedingungen (3) und (4) sind in diesem Falle also für beliebige λ_1, λ_2 erfüllt.
Nun weisen wir nach, dass bei Wahl von $\lambda_1 > 0$ und $\lambda_2 < 0$ gemäß (1)–(4) die Voraussetzungen von Satz 12.12 erfüllt sind.

(a) Es ist $A \leq_{\mathcal{H}} B$.

Dies gilt genau dann, wenn $\mathcal{H}[B(t) - A(t)] \geq 0$ für alle $t \in [0, \omega]$ bzw.

$$\begin{aligned} \lambda_1[\beta(t) - \alpha(t)] + [\beta'(t) - \alpha'(t) + F(\beta(t), t) - F(\alpha(t), t)] &\geq 0, \\ -\lambda_2[\beta(t) - \alpha(t)] - [\beta'(t) - \alpha'(t) + F(\beta(t), t) - F(\alpha(t), t)] &\geq 0 \end{aligned}$$

für alle $t \in [0, \omega]$. Wir weisen zunächst die erste Ungleichung nach. Für beliebiges $t \in [0, \omega]$ ist $(\alpha(t), t), (\beta(t), t) \in \Omega$ und wegen (1) daher

$$\begin{aligned} & -\lambda_1^2[\beta(t) - \alpha(t)] - \lambda_1[F(\beta(t), t) - F(\alpha(t), t)] \\ \leq\ & g(\beta(t), t) - g(\alpha(t), t) \\ & \qquad \text{(wegen (1))} \\ \leq\ & -[\beta''(t) - \alpha''(t)] - \frac{d}{dt}[F(\beta(t), t) - F(\alpha(t), t)] \\ & \qquad \text{(wegen Voraussetzung 3.).} \end{aligned}$$

Dann ist

$$\begin{aligned} & \int_t^\omega e^{-\lambda_1 s}\{-\lambda_1^2[\beta(s) - \alpha(s)] - \lambda_1[F(\beta(s), s) - F(\alpha(s), s)]\}\, ds \\ \leq\ & \int_t^\omega e^{-\lambda_1 s}\Big\{-[\beta''(s) - \alpha''(s)] - \frac{d}{ds}[F(\beta(s), s) - F(\alpha(s), s)]\Big\}\, ds \\ =\ & e^{-\lambda_1 s}\{-[\beta'(s) - \alpha'(s)] - [F(\beta(s), s) - F(\alpha(s), s)]\}\,\Big|_{s=t}^{s=\omega} \\ & \quad + \lambda_1 \int_t^\omega e^{-\lambda_1 s}\{-[\beta'(s) - \alpha'(s)] - [F(\beta(s), s) - F(\alpha(s), s)]\}\, ds \\ =\ & e^{-\lambda_1 \omega}\{\underbrace{-[\beta'(\omega) - \alpha'(\omega)] - [F(\beta(\omega) - F(\alpha(\omega)]}_{\leq \lambda_1[\beta(\omega) - \alpha(\omega)]}\} \\ & \quad \text{(wegen (3)} \\ & \quad + e^{-\lambda_1 t}\{[\beta')t) - \alpha'(t)] + [F(\beta(t), t) - F(\alpha(t), t)]\} \\ & \quad - \lambda_1 \int_t^\omega e^{-\lambda_1 s}[\beta'(s) - \alpha'(s)]\, ds \\ & \quad - \lambda_1 \int_t^\omega e^{-\lambda_1 s}[F(\beta(s), s) - F(\alpha(s), s)]\, ds \\ \leq\ & \lambda_1 e^{-\lambda_1 \omega}[\beta(\omega) - \alpha(\omega)] + e^{-\lambda_1 t}\{[\beta'(t) - \alpha'(t)] + [F(\beta(t), t) - F(\alpha(t), t)]\} \end{aligned}$$

$$
\begin{aligned}
& - \lambda_1 e^{-\lambda_1 s}[\beta(s) - \alpha(s)] \Big|_{s=t}^{s=\omega} \\
& + \int_t^\omega e^{-\lambda_1 s}\{-\lambda_1^2[\beta(s) - \alpha(s)] - \lambda_1[F(\beta(s), s) - F(\alpha(s), s)]\}\, ds \\
= \;& e^{-\lambda_1 t}\{\lambda_1[\beta(t) - \alpha(t)] + [\beta'(t) - \alpha'(t) + F(\beta(t), t) - F(\alpha(t), t)]\} \\
& \int_t^\omega e^{-\lambda_1 s}\{-\lambda_1^2[\beta(s) - \alpha(s)] - \lambda_1[F(\beta(s), s) - F(\alpha(s), s)]\}\, ds
\end{aligned}
$$

und hieraus folgt die behauptete erste Ungleichung. Die Gültigkeit der zweiten Ungleichung ergibt sich mit Hilfe von (2) und (4) entsprechend. Damit ist $A \leq_{\mathcal{H}} B$ bewiesen.

(b) Es ist $-A' - f(A, \cdot) \leq_{\mathcal{K}} 0 \leq_{\mathcal{K}} -B' - f(B, \cdot)$.

Dies gilt genau dann, wenn

$$
\mathcal{K}[-A'(t) - f(A(t), t)] \leq 0 \leq \mathcal{K}[-B'(t) - f(B(t), t)] \quad \text{für alle } t \in [0, \omega]
$$

bzw.

$$
\begin{pmatrix} \lambda_1 & 1 \\ \lambda_2 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ -\alpha''(t) - \dfrac{d}{dt} F(\alpha(t), t) - g(\alpha(t), t) \end{pmatrix} \leq \begin{pmatrix} 0 \\ 0 \end{pmatrix}
$$

und

$$
\begin{pmatrix} 0 \\ 0 \end{pmatrix} \leq \begin{pmatrix} \lambda_1 & 1 \\ \lambda_2 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ -\beta''(t) - \dfrac{d}{dt} F(\beta(t), t) - g(\beta(t), t) \end{pmatrix}
$$

für alle $t \in [0, \omega]$. Wegen Voraussetzung 3. ist (b) erfüllt.

(c) Es ist $RA \leq_{\mathcal{K}} 0 \leq_{\mathcal{K}} RB$.

Die Behauptung (c) ist gleichwertig mit

$$
\mathcal{K}[A(\omega) - A(0)] \leq 0 \leq \mathcal{K}[B(\omega) - B(0)]
$$

bzw.

$$
\begin{pmatrix} \lambda_1 & 1 \\ \lambda_2 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ \alpha'(\omega) - \alpha'(0) \end{pmatrix} \leq 0 \leq \begin{pmatrix} \lambda_1 & 1 \\ \lambda_2 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ \beta'(\omega) - \beta'(0) \end{pmatrix},
$$

wobei wir

$$
\alpha(0) = \alpha(\omega), \qquad \beta(0) = \beta(\omega)
$$

und

$$
F(\alpha(0), 0) = F(\alpha(0), \omega), \qquad F(\beta(0), 0) = F(\beta(0), \omega)
$$

ausgenutzt haben. Wegen

$$
\alpha'(\omega) - \alpha'(0) \leq 0 \leq \beta'(\omega) - \beta'(0)
$$

ist auch (c) bewiesen.

(d) Für $z_1, z_2 \in C_2[0,\omega]$ mit $A \leq_{\mathcal{H}} z_1 \leq_{\mathcal{H}} z_2 \leq_{\mathcal{H}} B$ ist $F_C(z_1) \leq_{\mathcal{K}} F_C(z_2)$ bzw. $\mathcal{K}F(z_1)(t) \leq \mathcal{K}F(z_2)(t)$ für alle $t \in [0,\omega]$, wobei

$$F_C(z)(t) := f(z(t),t) + Cz(t) = \begin{pmatrix} F(x(t),t) + (\lambda_1+\lambda_2)x(t) \\ g(x(t),t) - \lambda_1\lambda_2 x(t) \end{pmatrix}.$$

Denn: Mit $z_1 = (x_1, y_1)^T$, $z_2 = (x_2, y_2)^T$ folgt aus $A \leq_{\mathcal{H}} z_1 \leq_{\mathcal{H}} z_2 \leq_{\mathcal{H}} B$, dass

$$\alpha(t) \leq x_1(t) \leq x_2(t) \leq \beta(t) \quad \text{für alle } t \in [0,\omega].$$

Wegen der Bedingungen (1) und (2) an λ_1 und λ_2 ist

$$\lambda_1^2[x_2(t) - x_1(t)] + \lambda_1[F(x_2(t),t) - F(x_1(t),t)] + g(x_2(t),t) - g(x_1(t),t) \geq 0,$$

sowie

$$\lambda_2^2[x_2(t) - x_1(t)] + \lambda_2[F(x_2(t),t) - F(x_1(t),t)] + g(x_2(t),t) - g(x_1(t),t) \geq 0.$$

Wie man leicht nachrechnet sind diese beiden Ungleichungen gleichwertig mit

$$\mathcal{K}[F_C(z_2)(t) - F_C(z_1)(t)] \geq 0,$$

womit auch (d) bewiesen ist.

Damit ist der Satz bewiesen. □

Bemerkung: Die Differentialgleichung

$$(*) \qquad x'' + f(x)x' + g(x,t) = 0$$

ordnet sich dem System

$$\begin{aligned} -x' &= F(x) - y \\ -y' &= g(x,t) \end{aligned}$$

unter, wenn man

$$F(x) := \int_0^x f(s)\,ds$$

setzt. □

Beispiel: Gesucht sei eine 2π-periodische Lösung von

$$x'' - \frac{1}{2}(1-x^2)x' + 3x - 4x^3 = \cos t$$

bzw. eine Lösung von

$$\begin{aligned} -x' &= -\tfrac{1}{2}(x - \tfrac{1}{3}x^3) - y, & x(0) &= x(2\pi) \\ -y' &= 3x - 4x^3 - \cos t, & y(0) &= y(2\pi). \end{aligned}$$

Mit $\alpha(t) := \frac{1}{2}$ und $\beta(t) := 1$ sind die Voraussetzungen von Satz 12.13 erfüllt, es existiert also eine 2π-periodische Lösung x von $(*)$ mit $\frac{1}{2} \leq x(t) \leq 1$ für alle $t \in [0, 2\pi]$. □

Weitere Einschließungssätze bei Randwertaufgaben für nichtlineare Differentialgleichungen zweiter Ordnung mit periodischen Randbedingungen kann man mit Hilfe des folgenden Hilfssatzes gewinnen.

Hilfssatz *Seien λ_1, λ_2 reelle Zahlen mit $\lambda_2 < \lambda_1 < 0$ und*

$$C := \begin{pmatrix} \lambda_1 + \lambda_2 & 1 \\ -\lambda_1\lambda_2 & 0 \end{pmatrix}, \qquad Lz := -z' + Cz, \qquad Rz := z(\omega) - z(0).$$

Man definiere die nichtsingulären Matrizen

$$H := \begin{pmatrix} 1 & 0 \\ \lambda_1 & 1 \end{pmatrix}, \qquad K := \begin{pmatrix} -1 & 0 \\ -\lambda_2 & -1 \end{pmatrix}, \qquad P := \begin{pmatrix} -\lambda_1 & -1 \\ \lambda_2 & 1 \end{pmatrix},$$

welche Halbordnungen $\leq_H$, $\leq_K$ und $\leq_P$ auf $C_2[0,\omega]$ bzw. $\mathbb{R}^2$ induzieren. Dann sind die folgenden beiden Eigenschaften (siehe Satz 12.10) erfüllt:

(a) Ist $w \in C_2^1[0,\omega]$, $-w' + Cw \geq_K 0$ und $Rw = 0$, so ist $w \geq_H 0$.

(b) Ist $w \in C_2^1[0,\omega]$, $-w' + Cw = 0$ und $Rw \geq_P 0$, so ist $w \geq_H 0$.

Beweis: Da C eine diagonalisierbare Matrix mit reellen Eigenwerten ist, besitzt $Lz = 0$, $Rz = 0$ nur die triviale Lösung (siehe Aussage (a) von Satz 12.12). Sei $x := -w' + Cw$. Wegen $Rw = 0$ ist

$$w(t) = \int_0^\omega G(t,s)x(s)\,ds.$$

Wegen $x = -w' + Cw \geq_K 0$ ist $Kx(s) \geq 0$ für alle $s \in [0,\omega]$. Da

$$Hw(t) = \int_0^\omega HG(t,s)K^{-1}\underbrace{Kx(s)}_{\geq 0}\,ds$$

genügt es zum Beweis von (a) zu zeigen, dass $HG(t,s)K^{-1} \geq 0$ für alle $(t,s) \in [0,\omega] \times [0,\omega]$. Mit einem Fundamentalsystem $\Phi(\cdot)$ von $-z' + Cz = 0$ ist die Greensche Matrix $G(t,s)$ zu (L,R) gegeben durch

$$G(t,s) := \Phi(t)\left\{ \begin{array}{c} (A-I) \\ A \end{array} \right\}\Phi(s)^{-1} \quad \begin{array}{l} 0 \leq s < t \leq \omega, \\ 0 \leq t < s \leq \omega, \end{array}$$

wobei

$$A := [\Phi(\omega) - \Phi(0)]^{-1}\Phi(\omega) = [I - \Phi(\omega)^{-1}\Phi(0)]^{-1}.$$

Wegen $\mathcal{H}C\mathcal{H}^{-1} = \text{diag}\,(\lambda_1, \lambda_2)$ mit

$$\mathcal{H} := \begin{pmatrix} \lambda_1 & 1 \\ -\lambda_2 & 1 \end{pmatrix}$$

(siehe Beginn des Beweises von Satz 12.13) ist

$$\Phi(t) := e^{Ct} = \mathcal{H}^{-1}\text{diag}\,(e^{\lambda_1 t}, e^{\lambda_2 t})\mathcal{H}$$

und damit

$$\begin{aligned} A &= \mathcal{H}^{-1}\mathrm{diag}\,\Big(\frac{1}{1-e^{-\lambda_1\omega}},\frac{1}{1-e^{-\lambda_2\omega}}\Big)\mathcal{H}, \\ A-I &= \mathcal{H}^{-1}\mathrm{diag}\,\Big(\frac{e^{-\lambda_1\omega}}{1-e^{-\lambda_1\omega}},\frac{e^{-\lambda_2\omega}}{1-e^{-\lambda_2\omega}}\Big)\mathcal{H}. \end{aligned}$$

Folglich ist

$$G(t,s) = \mathcal{H}^{-1}\left\{\begin{array}{l} \mathrm{diag}\,\Big(\frac{e^{\lambda_1(t-s-\omega)}}{1-e^{-\lambda_1\omega}},\frac{e^{\lambda_2(t-s-\omega)}}{1-e^{-\lambda_2\omega}}\Big) \\ \mathrm{diag}\,\Big(\frac{e^{\lambda_1(t-s)}}{1-e^{-\lambda_1\omega}},\frac{e^{\lambda_2(t-s)}}{1-e^{-\lambda_2\omega}}\Big) \end{array}\right\}\mathcal{H} \quad \begin{array}{l} 0\le s<t\le\omega, \\ 0\le t<s\le\omega, \end{array}$$

und damit

$$HG(t,s)K^{-1} = H\mathcal{H}^{-1}\left\{\begin{array}{l} \mathrm{diag}\,\Big(\frac{e^{\lambda_1(t-s-\omega)}}{1-e^{-\lambda_1\omega}},\frac{e^{\lambda_2(t-s-\omega)}}{1-e^{-\lambda_2\omega}}\Big) \\ \mathrm{diag}\,\Big(\frac{e^{\lambda_1(t-s)}}{1-e^{-\lambda_1\omega}},\frac{e^{\lambda_2(t-s)}}{1-e^{-\lambda_2\omega}}\Big) \end{array}\right\}\mathcal{H}K^{-1} \quad \begin{array}{l} 0\le s<t\le\omega, \\ 0\le t<s\le\omega. \end{array}$$

Nun ist

$$H\mathcal{H}^{-1} = \frac{1}{\lambda_1-\lambda_2}\begin{pmatrix} 1 & 1 \\ \lambda_1-\lambda_2 & 0 \end{pmatrix}, \qquad \mathcal{H}K^{-1} = \begin{pmatrix} -(\lambda_1-\lambda_2) & -1 \\ 0 & 1 \end{pmatrix}$$

und folglich

$$HG(t,s)K^{-1} = \frac{1}{\lambda_1-\lambda_2}\begin{pmatrix} -(\lambda_1-\lambda_2)A & B-A \\ -(\lambda_1-\lambda_2)^2A & -(\lambda_1-\lambda_2)A \end{pmatrix},$$

wobei,

$$A := \frac{e^{\lambda_1(t-s-\omega)}}{1-e^{-\lambda_1\omega}}, \qquad B := \frac{e^{\lambda_2(t-s-\omega)}}{1-e^{-\lambda_2\omega}} \qquad (s<t),$$

bzw.

$$A := \frac{e^{\lambda_1(t-s)}}{1-e^{-\lambda_1\omega}}, \qquad B := \frac{e^{\lambda_2(t-s)}}{1-e^{-\lambda_2\omega}} \qquad (t<s).$$

Wegen $\lambda_2<\lambda_1<0$ ist $A<B<0$ und damit $HG(t,s)K^{-1}\ge 0$ für alle $(t,s)\in[0,\omega]\times[0,\omega]$. Damit ist (a) bewiesen.

Beim Beweis von (b) nehmen wir an, es sei $w\in C_2^1[0,\omega]$, $-w'+Cw=0$ und $Rw\ge_P 0$. Da $\Phi(t)=e^{Ct}$ ein Fundamentalsystem zu $-z'+Cz=0$ ist, existiert ein $\xi\in\mathbb{R}^2$ mit $w(t)=e^{Ct}\xi$. Die Voraussetzung $Rw\ge_P 0$ besagt dann, dass

$$\begin{aligned} PRw &= P(w(\omega)-w(0)) \\ &= P(e^{C\omega}-I)\xi \\ &= P\mathcal{H}^{-1}[\mathrm{diag}\,(e^{\lambda_1\omega}-1,e^{\lambda_2\omega}-1)\mathcal{H}\xi \\ &= \mathrm{diag}\,(1-e^{\lambda_1\omega},1-e^{\lambda_2\omega})\mathcal{H}\xi \\ &\ge 0 \end{aligned}$$

und damit $\mathcal{H}\xi \geq 0$. Die Behauptung $w \geq_H 0$ ist dann richtig, da

$$\begin{aligned} Hw(t) &= He^{Ct}\xi \\ &= H\mathcal{H}^{-1}\text{diag}\,(e^{\lambda_1 t}, e^{\lambda_2 t})\mathcal{H}\xi \\ &= \underbrace{\frac{1}{\lambda_1 - \lambda_2}\begin{pmatrix} 1 & 1 \\ \lambda_1 - \lambda_2 & 0 \end{pmatrix}}_{\geq 0} \underbrace{\text{diag}\,(e^{\lambda_1 t}, e^{\lambda_2 t})}_{\geq 0} \underbrace{\mathcal{H}\xi}_{\geq 0} \\ &\geq 0. \end{aligned}$$

Damit ist der Hilfssatz bewiesen. □

Die beiden folgenden Sätze (siehe Sätze 5 und 6 bei J. Werner (1969)) sind Anwendungen von Satz 12.10 in Verbindung mit dem obigen Hilfssatz.

Satz 12.14 *Gegeben sei die Randwertaufgabe mit periodischen Randbedingungen*

$$(*) \qquad x'' + f(x)x' + x = e(t), \qquad x(0) = x(\omega),\ x'(0) = x'(\omega).$$

Sei $e \in C[0,\omega]$, $A := \min_{t\in[0,\omega]} e(t)$, $B := \max_{t\in[0,\omega]} e(t)$ *und* $f \in C[A,B]$. *Es möge reelle Zahlen* λ_1, λ_2 *mit* $\lambda_2 < \lambda_1 < 0$ *geben derart, das*

$$\max\Big(-\lambda_1, -\frac{\lambda_2^2+1}{\lambda_2}\Big) \leq f(x) \leq -(\lambda_1+\lambda_2) \quad \textit{für alle } x \in [A,B].$$

Dann besitzt $(*)$ *eine Lösung* x *mit* $A \leq x(t) \leq B$ *für alle* $t \in [0,\omega]$.

Beweis: Man schreibe die Randwertaufgabe $(*)$ um in das System

$$\begin{aligned} -x' &= F(x) - y \\ -y' &= x - e(t), \end{aligned} \qquad R\begin{pmatrix} x \\ y \end{pmatrix} := \begin{pmatrix} x(\omega) - x(0) \\ y(\omega) - y(0) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

mit

$$F(x) := \int_0^x f(s)\,ds.$$

Wir setzen

$$\alpha(t) := \begin{pmatrix} A \\ F(A) \end{pmatrix}, \qquad \beta(t) := \begin{pmatrix} B \\ F(B) \end{pmatrix}$$

und wenden Satz 12.10 mit

$$C := \begin{pmatrix} \lambda_1 + \lambda_2 & 1 \\ -\lambda_1\lambda_2 & 0 \end{pmatrix}$$

und den durch (siehe obigen Hilfssatz)

$$H := \begin{pmatrix} 1 & 0 \\ \lambda_1 & 1 \end{pmatrix}, \qquad K := \begin{pmatrix} -1 & 0 \\ -\lambda_2 & -1 \end{pmatrix}, \qquad P := \begin{pmatrix} -\lambda_1 & -1 \\ \lambda_2 & 1 \end{pmatrix}$$

gegebenen Halbordnungen $\leq_H$, $\leq_K$ bzw. $\leq_P$ auf $C_2[0,\omega]$ bzw. $\mathbb{R}^2$ an. Um Satz 12.10 anwenden zu können, ist zunächst $\alpha \leq_H \beta$ nachzuweisen. Dies ist gleichwertig mit

$$\begin{aligned} 0 &\leq H[\beta(t) - \alpha(t)] \\ &= \begin{pmatrix} 1 & 0 \\ \lambda_1 & 1 \end{pmatrix}\begin{pmatrix} B - A \\ F(B) - F(A) \end{pmatrix} \\ &= \begin{pmatrix} B - A \\ \lambda_1(B-A) + F(B) - F(A) \end{pmatrix}. \end{aligned}$$

Wegen $B - A \geq 0$ und

$$\lambda_1(B-A) + F(B) - F(A) = (\underbrace{\lambda_1 + f(C)}_{\geq 0})(\underbrace{B-A}_{\geq 0}) \geq 0$$

ist dies richtig. Die zweite Bedingung $-\alpha' - f(\alpha, \cdot) \leq_K 0 \leq_K -\beta' - f(\beta, \cdot)$ in Satz 12.10 besagt in unserem Spezialfall, dass

$$-K \begin{pmatrix} F(A) - F(A) \\ A - e(t) \end{pmatrix} \leq \begin{pmatrix} 0 \\ 0 \end{pmatrix} \leq -K \begin{pmatrix} F(B) - F(B) \\ B - e(t) \end{pmatrix} \quad \text{für alle } t \in [0, \omega]$$

bzw.

$$A - e(t) \leq 0 \leq B - e(t) \quad \text{für alle } t \in [0, \omega],$$

was wegen $A := \min_{t\in[0,\omega]} e(t)$ und $B := \max_{t\in[0,\omega]} e(t)$ richtig ist. Da α und β konstant sind, ist $R\alpha = 0 = R\beta$ und damit die dritte Bedingung in Satz 12.10 trivialerweise erfüllt. Daher ist nur noch zu zeigen, dass für $z_1, z_2 \in C_2[0, \omega]$ die Implikation

$$\alpha \leq_H z_1 \leq_H z_2 \leq_H \beta \Longrightarrow F_C(z_1) \leq_K F_C(z_2)$$

gültig ist. Hierbei ist $F_C \colon C_2[0, \omega] \longrightarrow C_2[0, \omega]$ definiert durch

$$F_C(z)(t) := \begin{pmatrix} F(x) - y \\ x - e(t) \end{pmatrix} + \begin{pmatrix} \lambda_1 + \lambda_2 & 1 \\ -\lambda_1\lambda_2 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} F(x) + (\lambda_1 + \lambda_2)x \\ (1 - \lambda_1\lambda_2)x - e(t) \end{pmatrix}.$$

Nach Definition von H und K ist daher zu zeigen, dass

$$A \leq x_1 \leq x_2 \leq B \Longrightarrow \begin{array}{rcl} -(F(x_2) - F(x_1)) - (\lambda_1 + \lambda_2)(x_2 - x_1) & \geq & 0, \\ -\lambda_2(F(x_2) - F(x_1)) - (1 + \lambda_2^2)(x_2 - x_1) & \geq & 0. \end{array}$$

Wegen $F(x_2) - F(x_1) = f(x)(x_2 - x_1)$ mit einem $x \in [x_1, x_2] \subset [A, B]$ und

$$-\frac{\lambda_2^2 + 1}{\lambda_2} \leq f(x) \leq -(\lambda_1 + \lambda_2) \quad \text{für alle } x \in [A, B]$$

ist dies aber richtig, womit der Satz bewiesen ist. □

Beispiel: Gesucht sei eine Lösung der Randwertaufgabe mit periodischen Randbedingungen

$$x'' + (3 + x)x' + x = \frac{1}{2}\cos t, \qquad x(0) = x(2\pi),\ x'(0) = x'(2\pi).$$

Mit $A := -\frac{1}{2}$, $B := \frac{1}{2}$, $\lambda_1 := -\frac{3}{2}$, $\lambda_2 := -2$ sind die Voraussetzungen von Satz 12.14 erfüllt. Es existiert daher eine Lösung x mit $-\frac{1}{2} \leq x(t) \leq \frac{1}{2}$ für alle $t \in [0, 2\pi]$. □

Satz 12.15 *Gegeben sei die Randwertaufgabe mit periodischen Randbedingungen*

$$(*) \qquad x'' + Dx' + g(x) = e(t), \qquad x(0) = x(\omega),\ x'(0) = x'(\omega),$$

wobei $D > 0$. Es seien $A, B, \lambda_1, \lambda_2$ reelle Zahlen mit

1. *Es ist $A \leq B$ und $\lambda_2 < \lambda_1 < 0$.*

2. *Es ist* $e \in C[0,\omega]$ *und* $g(A) \le e(t) \le g(B)$ *für alle* $t \in [0,\omega]$.

3. *Es ist* $\lambda_1 + \lambda_2 \le -D \le \lambda_1$.

4. *Es ist* $g \in C^1[A,B]$ *und* $g'(x) \le -\lambda_2(\lambda_2 + D)$ *für alle* $x \in [A,B]$.

Dann besitzt $(*)$ *eine Lösung* x *mit* $A \le x(t) \le B$ *für alle* $t \in [0,\omega]$.

Beweis: Im Prinzip verläuft der Beweis genau wie der des vorigen Satzes. Man schreibe zunächst die Randwertaufgabe $(*)$ um in das System

$$\begin{array}{rcl} -x' &=& Dx - y \\ -y' &=& g(x) - e(t), \end{array} \qquad R\begin{pmatrix} x \\ y \end{pmatrix} := \begin{pmatrix} x(\omega) - x(0) \\ y(\omega) - y(0) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Wir setzen

$$\alpha(t) := \begin{pmatrix} A \\ DA \end{pmatrix}, \qquad \beta(t) := \begin{pmatrix} B \\ DB \end{pmatrix}$$

und wenden Satz 12.10 mit

$$C := \begin{pmatrix} \lambda_1 + \lambda_2 & 1 \\ -\lambda_1\lambda_2 & 0 \end{pmatrix}$$

und den durch (siehe obigen Hilfssatz)

$$H := \begin{pmatrix} 1 & 0 \\ \lambda_1 & 1 \end{pmatrix}, \qquad K := \begin{pmatrix} -1 & 0 \\ -\lambda_2 & -1 \end{pmatrix}, \qquad P := \begin{pmatrix} -\lambda_1 & -1 \\ \lambda_2 & 1 \end{pmatrix}$$

gegebenen Halbordnungen $\le_H$, $\le_K$ bzw. $\le_P$ auf $C_2[0,\omega]$ bzw. $\mathbb{R}^2$ an. Es ist $\alpha \le_H \beta$, da

$$H[\beta(t) - \alpha(t)] = \begin{pmatrix} 1 & 0 \\ \lambda_1 & 1 \end{pmatrix}\begin{pmatrix} B - A \\ D(B - A) \end{pmatrix} = \begin{pmatrix} B - A \\ (\lambda_1 + D)(B - A) \end{pmatrix} \ge \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

wegen $A \le B$ und $\lambda_1 + D \ge 0$. Die zweite Bedingung $-\alpha' - f(\alpha,\cdot) \le_K 0 \le_K -\beta' - f(\beta,\cdot)$ in Satz 12.10 besagt in unserem Spezialfall, dass

$$-K\begin{pmatrix} DA - DA \\ g(A) - e(t) \end{pmatrix} \le \begin{pmatrix} 0 \\ 0 \end{pmatrix} \le -K\begin{pmatrix} DB - DB \\ g(B) - e(t) \end{pmatrix} \quad \text{für alle } t \in [0,\omega]$$

bzw.

$$g(A) - e(t) \le 0 \le g(B) - e(t) \quad \text{für alle } t \in [0,\omega].$$

Dies ist aber nach Voraussetzung 2. richtig. Da α und β konstant sind, ist $R\alpha = 0 = R\beta$ und damit die dritte Bedingung in Satz 12.10 trivialerweise erfüllt. Daher ist nur noch zu zeigen, dass für $z_1, z_2 \in C_2[0,\omega]$ die Implikation

$$\alpha \le_H z_1 \le_H z_2 \le_H \beta \Longrightarrow F_C(z_1) \le_K F_C(z_2)$$

gültig ist. Hierbei ist $F_C\colon C_2[0,\omega] \longrightarrow C_2[0,\omega]$ definiert durch

$$F_C(z)(t) := \begin{pmatrix} Dx - y \\ g(x) - e(t) \end{pmatrix} + \begin{pmatrix} \lambda_1 + \lambda_2 & 1 \\ -\lambda_1\lambda_2 & 0 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} (D + \lambda_1 + \lambda_2)x \\ g(x) - \lambda_1\lambda_2 x - e(t) \end{pmatrix}.$$

Nach Definition von H und K ist daher zu zeigen, dass

$$A \le x_1 \le x_2 \le B \Longrightarrow \begin{array}{rcl} -(D+\lambda_1+\lambda_2)(x_2-x_1) & \ge & 0, \\ -(g(x_2)-g(x_1)) - \lambda_2(D+\lambda_2)(x_2-x_1) & \ge & 0. \end{array}$$

Diese beiden Ungleichungen sind wegen $-(D+\lambda_1+\lambda_2) \ge 0$ bzw. $g'(x) \le -\lambda_2(D+\lambda_2)$ für alle $x \in [A,B]$ (Voraussetzung 3. bzw. 4.). richtig. Damit ist der Satz bewiesen. □

Beispiel: Die Randwertaufgabe

$$x'' + 5x' + \frac{x}{1-x} = \frac{1}{2}\cos t, \qquad x(0) = x(2\pi),\ x'(0) = x'(2\pi)$$

besitzt mindestens eine Lösung x mit $-1 \le x(t) \le \frac{1}{3}$. Denn mit $D := 5$, $A := -1$, $B := \frac{1}{3}$, $\lambda_1 := -\frac{1}{2}$ und $\lambda_2 := -\frac{9}{2}$ sind die Voraussetzungen von Satz 12.15 erfüllt. □

12.6 Einschließungssätze für periodische Lösungen der Liénardschen Differentialgleichung

Unter einer (nicht-autonomen) *Liénardschen Differentialgleichung* versteht man eine Differentialgleichung zweiter Ordnung der Form

$$x'' + f(x)x' + g(x) = e(t). \tag{P}$$

Wir interessieren uns für Lösungen von (P), welche den periodischen Randbedingungen

$$x(0) = x(\omega),\ x'(0) = x'(\omega) \tag{$*$}$$

genügen, wobei $\omega > 0$ fest vorgegeben ist. Wir halten uns im wesentlichen an J. Werner (1970). Speziell fragen wir danach, unter welchen Bedingungen zwischen einer *Unter-* und einer *Oberlösung* (jeweils geeignet definiert) eine Lösung von (P) mit den periodischen Randbedingungen ($*$) liegt. Durch Satz 12.13 (siehe auch die anschließende Bemerkung) haben wir bewiesen:

- *Gegeben sei die Liénardsche Differentialgleichung (P). Gesucht sei eine Lösung von (P), welche den periodischen Randbedingungen ($*$) mit vorgegebenem $\omega > 0$ genügt. Seien α, $\beta \in C^2[0,\omega]$ zwei Funktionen mit:*

  1. *Es ist $\alpha(t) \le \beta(t)$ für alle $t \in [0,\omega]$.*
  2. *Es ist*

  $$\alpha(0) = \alpha(\omega),\ \alpha'(0) \ge \alpha'(\omega),\ \beta(0) = \beta(\omega),\ \beta'(0) \le \beta'(\omega).$$

  3. *Es ist*

  $$D(\alpha)(t) \le 0 \le D(\beta)(t) \quad \text{für alle } t \in [0,\omega],$$

  wobei für $y \in C^2[0,\omega]$ der Defekt *$D(y)$ durch*

  $$D(y)(t) := -y''(t) - f(y(t))y'(t) - g(y(t)) + e(t)$$

  gegeben ist.

Mit

$$\alpha_{\min} := \min_{t\in[0,\omega]} \alpha(t), \qquad \beta_{\max} := \max_{t\in[0,\omega]} \beta(t)$$

seien $e \in C[0,\omega]$, $f \in C[\alpha_{\min}, \beta_{\max}]$ *und* $g \in C^1[\alpha_{\min}, \beta_{\max}]$. *Dann besitzt (P) eine Lösung* x, *welche den periodischen Randbedingungen* $(*)$ *genügt, mit* $\alpha(t) \le x(t) \le \beta(t)$ *für alle* $t \in [0,\omega]$.

Bemerkung: Ist $g\colon [\alpha_{\min}, \beta_{\max}] \longrightarrow \mathbb{R}$ echt monoton wachsend, gilt also

$$\alpha_{\min} \le x \le y \le \beta_{\max} \Longrightarrow g(x) \le g(y)$$

und

$$\alpha_{\min} \le x \le y \le \beta_{\max},\ g(x) = g(y) \Longrightarrow x = y,$$

sind ferner $\alpha,\ \beta$ zwei Funktionen, die den Bedingungen 1.–3. der obigen Aussage genügen, so ist notwendigerweise $\alpha = \beta$ eine Lösung von (P), welche den periodischen Randbedingungen $(*)$ genügt. In diesem Fall ist die Aussage also nicht sinnvoll anwendbar. Dies erkennt man an der folgenden Gleichungs-Ungleichungskette, bei der wir $F(x) := \int_0^x f(s)\,ds$ benutzen:

$$\begin{aligned}
0 &\le \int_0^\omega \underbrace{[g(\beta(t)) - g(\alpha(t))]}_{\ge 0}\, dt \\
&= \int_0^\omega \Big\{ -[\beta''(t) - \alpha''(t)] - \frac{d}{dt}[F(\beta(t)) - F(\alpha(t))] - \underbrace{[D(\beta)(t) - D(\alpha)(t)]}_{\ge 0} \Big\}\, dt \\
&\le \int_0^\omega \Big\{ -[\beta''(t) - \alpha''(t)] - \frac{d}{dt}[F(\beta(t)) - F(\alpha(t))] \Big\}\, dt \\
&= \underbrace{[\alpha'(\omega) - \alpha'(0)]}_{\le 0} - \underbrace{[\beta'(\omega) - \beta'(0)]}_{\ge 0} + \underbrace{[F(\alpha(\omega)) - F(\alpha(0))]}_{=0} - \underbrace{[F(\beta(\omega)) - F(\beta(0))]}_{=0} \\
&\le 0.
\end{aligned}$$

Wegen der vorausgesetzten strengen Monotonie von g auf $[\alpha_{\min}, \beta_{\max}]$ erhält man, wie behauptet, $\alpha = \beta$. Eine wichtige Klasse von Liénardschen Differentialgleichungen, für die die obige Aussage daher nicht (sinnvoll) anwendbar ist, besteht z. B. aus Gleichungen der Form

$$x'' + f(x)x' + x = e(t).$$

Um auch für einige Differentialgleichungen dieses und ähnlichen Typs entsprechende Einschließungssätze aufstellen zu können, werden die Bedingungen 2. und 3. der obigen Aussage verändert. □

Satz 12.16 *Gegeben sei die Liénardsche Differentialgleichung*

(P) $$x'' + f(x)x' + g(x) = e(t).$$

Gesucht sei eine Lösung von (P), welche den periodischen Randbedingungen

$(*)$ $$x(0) = x(\omega),\ x'(0) = x'(\omega)$$

mit vorgegebenem $\omega > 0$ *genügt. Seien* $\alpha,\ \beta \in C^2[0,\omega]$ *zwei Funktionen mit:*

1. *Es ist* $\alpha(t) \le \beta(t)$ *für alle* $t \in [0,\omega]$.

2'. *Es ist*

$$\alpha(0) = \alpha(\omega), \quad \alpha'(0) \le \alpha'(\omega), \quad \beta(0) = \beta(\omega), \quad \beta'(0) \ge \beta'(\omega).$$

3'. *Es ist*

$$-D(\alpha)(t) \le 0 \le -D(\beta)(t) \quad \text{für alle } t \in [0,\omega],$$

wobei für $y \in C^2[0,\omega]$ *der* Defekt $D(y)$ *durch*

$$D(y)(t) := -y''(t) - f(y(t))y'(t) - g(y(t)) + e(t)$$

gegeben ist.

Mit

$$\alpha_{\min} := \min_{t\in[0,\omega]} \alpha(t), \qquad \beta_{\max} := \max_{t\in[0,\omega]} \beta(t)$$

seien $e \in C[0,\omega]$, $f \in C[\alpha_{\min},\beta_{\max}]$ *und* $g \in C^1[\alpha_{\min},\beta_{\max}]$. *Sei*

$$G' := \max_{x\in[\alpha_{\min},\beta_{\max}]} g'(x) > 0$$

und

$$2\sqrt{G'} \le f(x) \quad \text{für alle } x \in [\alpha_{\min},\beta_{\max}].$$

Dann besitzt (P) eine Lösung x, *welche den periodischen Randbedingungen* $(*)$ *genügt, mit* $\alpha(t) \le x(t) \le \beta(t)$ *für alle* $t \in [0,\omega]$.

Beweis: Zunächst schreiben wir (P) mit den periodischen Randbedingungen $(*)$ als das System

$$z' = h(z,t), \qquad Rz := z(\omega) - z(0) = 0,$$

wobei

$$z := \begin{pmatrix} x \\ y \end{pmatrix}, \qquad h(z,t) := \begin{pmatrix} -F(x) + y \\ e(t) - g(x) \end{pmatrix}, \qquad F(x) := \int_0^x f(s)\,ds.$$

Mit noch geeignet zu wählenden $\lambda_1, \lambda_2 > 0$ mit $\lambda_1 \ne \lambda_2$ sei

$$Q := \begin{pmatrix} \lambda_1+\lambda_2 & -1 \\ \lambda_1\lambda_2 & 0 \end{pmatrix}, \qquad H := \begin{pmatrix} 1 & 0 \\ -\lambda_1 & 1 \end{pmatrix}, \qquad L_Q z := z' + Qz.$$

Die homogene Aufgabe $L_Q z = 0$, $Rz = 0$ besitzt nur die triviale Lösung, da Q die beiden reellen, voneinander und von Null verschiedenen Eigenwerte λ_1 und λ_2 besitzt (siehe den Anfang des Beweises von Satz 12.12). Dann besitzt die Aufgabe $L_Q z = r$, $Rz = 0$ für jedes $r \in C_2[0,\omega]$ eine eindeutige Lösung z, welche durch

$$z(t) = \int_0^\omega G_Q(t,s) r(s)\,ds$$

mit der Greenschen Funktion G_Q zu (L_Q, R) gegeben ist. Da $\Phi(t) := e^{-Qt}$ ein Fundamentalsystem zu $L_Q z = 0$ ist, ist die Greensche Funktion G_Q durch

$$G_Q(t,s) := -e^{-Qt} \left\{ \begin{array}{c} (A_Q - I) \\ A_Q \end{array} \right\} e^{Qs} \quad \begin{array}{l} 0 \le s < t \le \omega, \\ 0 \le t < s \le \omega, \end{array}$$

gegeben, wobei

$$A_Q := [I - e^{Q\omega}]^{-1}.$$

Man stellt leicht fest, dass

$$G_Q(t,s) = (I - A_Q) \left\{ \begin{array}{c} e^{-Q(t-s)} \\ e^{-Q(\omega+t-s)} \end{array} \right\} \quad \begin{array}{l} 0 \le s < t \le \omega, \\ 0 \le t < s \le \omega. \end{array}$$

Wir wollen nachweisen, dass $HG_Q(t,s)H^{-1}$ für alle $(t,s) \in [0,\omega] \times [0,\omega]$ eine nichtnegative Matrix ist. Wegen $HG_Q(t,s)H^{-1} = G_{HQH^{-1}}(t,s)$ ist es hierzu nützlich, $e^{HQH^{-1}t}$ zu berechnen. Wegen

$$HQH^{-1} = \begin{pmatrix} \lambda_2 & -1 \\ 0 & \lambda_1 \end{pmatrix} = \begin{pmatrix} 1 & -a \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda_2 & 0 \\ 0 & \lambda_1 \end{pmatrix} \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}$$

mit

$$a := -\frac{1}{\lambda_2 - \lambda_1}$$

ist

$$\begin{aligned} e^{HQH^{-1}t} &= \begin{pmatrix} 1 & -a \\ 0 & 1 \end{pmatrix} \begin{pmatrix} e^{\lambda_2 t} & 0 \\ 0 & e^{\lambda_1 t} \end{pmatrix} \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} e^{\lambda_2 t} & a(e^{\lambda_2 t} - e^{\lambda_1 t}) \\ 0 & e^{\lambda_1 t} \end{pmatrix}. \end{aligned}$$

Folglich ist

$$\begin{aligned} A_{HQH^{-1}} &= [I - e^{HQH^{-1}\omega}]^{-1} \\ &= \left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} e^{\lambda_2\omega} & a(e^{\lambda_2\omega} - e^{\lambda_1\omega} \\ 0 & e^{\lambda_1\omega} \end{pmatrix} \right]^{-1} \\ &= \frac{1}{(1 - e^{\lambda_1\omega})(1 - e^{\lambda_2\omega})} \begin{pmatrix} 1 - e^{\lambda_1\omega} & a(e^{\lambda_2\omega} - e^{\lambda_1\omega}) \\ 0 & 1 - e^{\lambda_2\omega} \end{pmatrix} \end{aligned}$$

und daher

$$I - A_{HQH^{-1}} = \frac{1}{(e^{\lambda_1\omega} - 1)(e^{\lambda_2\omega} - 1)} \begin{pmatrix} e^{\lambda_2\omega}(e^{\lambda_1\omega} - 1) & -a(e^{\lambda_2\omega} - e^{\lambda_1\omega}) \\ 0 & e^{\lambda_1\omega}(e^{\lambda_2\omega} - 1) \end{pmatrix}.$$

Nach Voraussetzung sind λ_1, λ_2 positive Zahlen mit $\lambda_1 \neq \lambda_2$. Daher ist

$$-a(e^{\lambda_2\omega} - e^{\lambda_1\omega}) = \frac{e^{\lambda_2\omega} - e^{\lambda_1\omega}}{\lambda_2 - \lambda_1} > 0$$

und folglich $I - A_{HQH^{-1}} \geq 0$. Weiter ist $e^{-HQH^{-1}t} \geq 0$ für alle $t \geq 0$. Für $0 \leq s < t \leq \omega$ ist daher

$$\begin{aligned} HG_Q(t,s)H^{-1} &= G_{HQH^{-1}}(t,s) \\ &= \underbrace{(I - A_{HQH^{-1}})}_{\geq 0} \underbrace{e^{-HQH^{-1}(t-s)}}_{\geq 0} \\ &\geq 0. \end{aligned}$$

Für $0 \leq t < s \leq \omega$ ist entsprechend

$$\begin{aligned} HG_Q(t,s)H^{-1} &= G_{HQH^{-1}}(t,s) \\ &= \underbrace{(I - A_{HQH^{-1}})}_{\geq 0} \underbrace{e^{-HQH^{-1}(\omega+t-s)}}_{\geq 0} \\ &\geq 0. \end{aligned}$$

Damit ist nachgewiesen[45], dass $HG_Q(t,s)H^{-1} = G_{HQH^{-1}}(t,s)$ nichtnegativ ist. Bei gegebenem λ_1 induziert die nichtsinguläre Matrix $H \in \mathbb{R}^{2\times 2}$ auf $C_2[0,\omega]$ eine Halbordnung $\geq_H$ indem man $z \geq_H 0$ für $z \in C_2[0,\omega]$ durch $Hz(t) \geq 0$ (komponentenweise) für alle $t \in [0,\omega]$ definiert. Man wähle $\lambda_1 > 0$ so, dass

$$-\lambda_1^2 + \lambda_1 f(x) - G' \geq 0 \quad \text{für alle } x \in [\alpha_{\min}, \beta_{\max}].$$

Z. B. kann man $\lambda_1 := \sqrt{G'}$ (wegen $2\sqrt{G'} \leq f(x)$ für alle $x \in [\alpha_{\min}, \beta_{\max}]$) wählen. Anschließend wähle man $\lambda_2 > 0$, $\lambda_2 \neq \lambda_1$, so groß, dass $f(x) \leq \lambda_1 + \lambda_2$ für alle $x \in [\alpha_{\min}, \beta_{\max}]$. Wir definieren nun $A, B \in C_2[0,\omega]$ durch

$$A(t) := \begin{pmatrix} \alpha(t) \\ \alpha'(t) + F(\alpha(t)) \end{pmatrix}, \qquad B(t) := \begin{pmatrix} \beta(t) \\ \beta'(t) + F(\beta(t)) \end{pmatrix}.$$

Es wird zunächst gezeigt, dass $A \leq_H B$. Hierzu ist zu zeigen, dass

$$\begin{aligned} 0 &\leq H(B(t) - A(t)) \\ &= \begin{pmatrix} \beta(t) - \alpha(t) \\ -\lambda_1[\beta(t) - \alpha(t)] + \beta'(t) - \alpha'(t) + F(\beta(t)) - F(\alpha(t)) \end{pmatrix} \quad \text{für alle } t \in [0,\omega]. \end{aligned}$$

Es bleibt zu zeigen, dass

$$0 \leq -\lambda_1[\beta(t) - \alpha(t)] + \beta'(t) - \alpha'(t) + F(\beta(t)) - F(\alpha(t)) \quad \text{für alle } t \in [0,\omega].$$

Für beliebiges $t \in [0,\omega]$ ist

$$\begin{aligned} & \frac{d}{dt} e^{\lambda_1 t}\{-\lambda_1[\beta(t) - \alpha(t)] + \beta'(t) - \alpha'(t) + F(\beta(t)) - F(\alpha(t))\} \\ = \; & e^{\lambda_1 t}\{-\lambda_1^2[\beta(t) - \alpha(t)] + \lambda_1[\beta'(t) - \alpha'(t)] + \lambda_1[F(\beta(t)) - F(\alpha(t))] \\ & \quad - \lambda_1[\beta'(t) - \alpha'(t)] + [\beta''(t) - \alpha''(t)] + [f(\beta(t))\beta'(t) - f(\alpha(t))\alpha'(t)]\} \end{aligned}$$

[45] Im folgenden Unterabschnitt werden wir untersuchen, was der eigentliche Grund hierfür ist.

$$
\begin{aligned}
&= e^{\lambda_1 t}\{-\lambda_1^2[\beta(t)-\alpha(t)] + \lambda_1[F(\beta(t))-F(\alpha(t))] - [g(\beta(t))-g(\alpha(t))] \\
&\qquad + \underbrace{[D(\alpha)(t)-D(\beta)(t)]}_{\geq 0}\} \\
&\geq e^{\lambda_1 t}\{-\lambda_1^2[\beta(t)-\alpha(t)] + \lambda_1[F(\beta(t))-F(\alpha(t))] - [g(\beta(t))-g(\alpha(t))]\} \\
&\geq e^{\lambda_1 t}\{-\lambda_1^2[\beta(t)-\alpha(t)] + \lambda_1[F(\beta(t))-F(\alpha(t))] - G'[\beta(t)-\alpha(t)]\} \\
&= e^{\lambda_1 t}\{\underbrace{[-\lambda_1^2 + \lambda_1 f(\gamma(t)) - G']}_{\geq 0}\underbrace{[\beta(t)-\alpha(t)]}_{\geq 0}\} \\
&\qquad \text{(nach Wahl von } \lambda_1 \text{ wegen } \gamma(t) \in [\alpha(t),\beta(t)] \subset [\alpha_{\min},\beta_{\max}]) \\
&\geq 0.
\end{aligned}
$$

Folglich ist

$$
\begin{aligned}
0 &\leq \int_0^\omega \underbrace{\frac{d}{dt} e^{\lambda_1 t}\{-\lambda_1[\beta(t)-\alpha(t)] + \beta'(t) - \alpha'(t) + F(\beta(t)) - F(\alpha(t))\}}_{\geq 0}\, dt \\
&= e^{\lambda_1 \omega}\{-\lambda_1[\beta(\omega)-\alpha(\omega)] + \beta'(\omega) - \alpha'(\omega) + F(\beta(\omega)) - F(\alpha(\omega))\} \\
&\qquad - \{-\lambda_1[\beta(0)-\alpha(0)] + \beta'(0) - \alpha'(0) + F(\beta(0)) - F(\alpha(0))\} \\
&\leq e^{\lambda_1 \omega}\{-\lambda_1[\beta(0)-\alpha(0)] + \beta'(0) - \alpha'(0) + F(\beta(0)) - F(\alpha(0))\} \\
&\qquad - \{-\lambda_1[\beta(0)-\alpha(0)] + \beta'(0) - \alpha'(0) + F(\beta(0)) - F(\alpha(0))\} \\
&= (\underbrace{e^{\lambda_1 \omega} - 1}_{>0})\{-\lambda_1[\beta(0)-\alpha(0)] + \beta'(0) - \alpha'(0) + F(\beta(0)) - F(\alpha(0))\}.
\end{aligned}
$$

Folglich ist

$$
\begin{aligned}
0 &\leq -\lambda_1[\beta(0)-\alpha(0)] + \beta'(0) - \alpha'(0) + F(\beta(0)) - F(\alpha(0))\} \\
&= e^{\lambda_1 t}\{-\lambda_1[\beta(t)-\alpha(t) + \beta'(t) - \alpha'(t) + F(\beta(t)) - F(\alpha(t))\}\Big|_{t=0},
\end{aligned}
$$

wegen

$$
0 \leq \frac{d}{dt} e^{\lambda_1 t}\{-\lambda_1[\beta(t)-\alpha(t)] + \beta'(t) - \alpha'(t) + F(\beta(t)) - F(\alpha(t))\} \quad \text{für alle } t \in [0,\omega]
$$

ist dann auch

$$
0 \leq e^{\lambda_1 t}\{-\lambda_1[\beta(t)-\alpha(t)] + \beta'(t) - \alpha'(t) + F(\beta(t)) - F(\alpha(t))\} \quad \text{für alle } t \in [0,\omega]
$$

bzw.

$$
0 \leq \{-\lambda_1[\beta(t)-\alpha(t)] + \beta'(t) - \alpha'(t) + F(\beta(t)) - F(\alpha(t))\} \quad \text{für alle } t \in [0,\omega].
$$

Damit ist schließlich $A \leq_H B$ nachgewiesen.

Auf

$$
\mathcal{M} := \{z \in C_2[0,\omega] : A \leq_H z \leq_H B\}
$$

definieren wir die Abbildung $T\colon \mathcal{M} \subset C_2[0,\omega] \longrightarrow C_2[0,\omega]$ durch

$$
T(z)(t) := \int_0^\omega G_Q(t,s)[h(z(s),s) + Qz(s)]\, ds.
$$

Dann ist $\mathcal{M} \subset C_2[0,\omega]$ nichtleer, konvex, abgeschlossen und beschränkt, ferner ist die Abbildung T kompakt (siehe Definition 10.6), also stetig und $T(\mathcal{M})$ relativ kompakt. Wenn wir noch $T(\mathcal{M}) \subset \mathcal{M}$ nachweisen, können wir den Schauderschen Fixpunktsatz (Satz 10.16) anwenden und erhalten die Existenz eines Fixpunktes $z \in \mathcal{M}$ von T. Zum Nachweis von $T(\mathcal{M}) \subset \mathcal{M}$ zeigen wir, dass

$$A \leq_H z_1 \leq_H z_2 \leq_H B \Longrightarrow T(z_1) \leq_H T(z_2)$$

und

$$A \leq_H T(A), \qquad T(B) \leq_H B.$$

Sei also

$$A \leq_H z_1 \leq_H z_2 \leq_H B$$

bzw.

$$\begin{aligned}
\begin{pmatrix} \alpha(t) \\ -\lambda_1\alpha(t) + F(\alpha(t)) \end{pmatrix} &\leq \begin{pmatrix} x_1(t) \\ -\lambda_1 x(t) + y_1(t) \end{pmatrix} \\
&\leq \begin{pmatrix} x_2(t) \\ -\lambda_1 x_2(t) + y_2(t) \end{pmatrix} \\
&\leq \begin{pmatrix} \beta(t) \\ -\lambda_1\beta(t) + \beta'(t) + F(\beta(t)) \end{pmatrix}
\end{aligned}$$

für alle $t \in [0,\omega]$, wobei

$$z_i(t) = \begin{pmatrix} x_i(t) \\ y_i(t) \end{pmatrix}, \qquad i = 1, 2.$$

Für beliebiges $t \in [0,\omega]$ ist dann

$$\begin{aligned}
& H[T(z_2)(t) - T(z_1)(t)] \\
= & \int_0^\omega HG_Q(t,s)[h(z_2(s),s) - h(z_1(s),s) + Q(z_2(s) - z_1(s))]\, ds \\
= & \int_0^\omega \underbrace{HQ_Q(t,s)H^{-1}}_{\geq 0} H[h(z_2(s),s) - h(z_1(s),s) + Q(z_2(s) - z_1(s))]\, ds.
\end{aligned}$$

Wegen

$$\begin{aligned}
& H[h(z_2(s),s) - h(z_1(s),s) + Q(z_2(s) - z_1(s))] \\
= & \begin{pmatrix} (\lambda_1+\lambda_2)(x_2(s) - x_1(s)) - [F(x_2(s)) - F(x_1(s))] \\ -\lambda_1^2(x_2(s) - x_1(s)) + \lambda_1[F(x_2(s)) - F(x_1(s))] - [g(x_2(s) - g(x_1(s))] \end{pmatrix} \\
= & \begin{pmatrix} (\lambda_1 + \lambda_2 - f(u(s)))(x_2(s) - x_1(s)) \\ (-\lambda_1^2 + \lambda_1 f(u(s)) - g'(v(s)))(x_2(s) - x_1(s)) \end{pmatrix}
\end{aligned}$$

mit $u(s), v(s) \in [x_1(s), x_2(s)] \subset [\alpha_{\min}, \beta_{\max}]$ und (nach Wahl von λ_1 und λ_2)

$$\lambda_1 + \lambda_2 \geq f(u(s)), \qquad -\lambda_1^2 + \lambda_1 f(u(s)) - g'(v(s)) \geq 0,$$

ist

$$H[h(z_2(s),s) - h(z_1(s),s) + Q(z_2(s) - z_1(s))] \geq 0$$

und daher

$$\begin{aligned} & H[T(z_2)(t) - T(z_1)(t)] \\ = & \int_0^\omega \underbrace{HQ_Q(t,s)H^{-1}}_{\geq 0} \underbrace{H[h(z_2(s),s) - h(z_1(s),s) + Q(z_2(s) - z_1(s))]}_{\geq 0} \, ds \\ \geq & \; 0 \end{aligned}$$

bzw. $T(z_1) \leq_H T(z_2)$. Weiter ist

$$A'(t) + QA(t) = \begin{pmatrix} (\lambda_1 + \lambda_2)\alpha(t) - F(\alpha(t)) \\ \alpha''(t) + f(\alpha(t))\alpha'(t) + \lambda_1\lambda_2\alpha(t) \end{pmatrix}$$

und

$$A(\omega) - A(0) = \begin{pmatrix} 0 \\ \alpha'(\omega) - \alpha'(0) \end{pmatrix},$$

folglich

$$\begin{aligned} A(t) = & \; e^{-Qt}(e^{-Q\omega} - I)^{-1} \begin{pmatrix} 0 \\ \alpha'(\omega) - \alpha'(0) \end{pmatrix} \\ & + \int_0^\omega G_Q(t,s) \begin{pmatrix} (\lambda_1 + \lambda_2)\alpha(s) - F(\alpha(s)) \\ \alpha''(s) + f(\alpha(s))\alpha'(s) + \lambda_1\lambda_2\alpha(s) \end{pmatrix} ds. \end{aligned}$$

Weiter ist

$$T(A)(t) = \int_0^\omega G_Q(t,s) \begin{pmatrix} (\lambda_1 + \lambda_2)\alpha(s) - F(\alpha(s)) \\ e(s) - g(\alpha(s)) + \lambda_1\lambda_2\alpha(s) \end{pmatrix} ds.$$

Für alle $t \in [0,\omega]$ ist daher

$$\begin{aligned} T(A)(t) - A(t) = & \int_0^\omega G_Q(t,s) \begin{pmatrix} 0 \\ D(\alpha)(s) \end{pmatrix} ds \\ & + e^{-Qt}(I - e^{-Q\omega})^{-1} \begin{pmatrix} 0 \\ \alpha'(\omega) - \alpha'(0) \end{pmatrix} \\ = & \int_0^\omega G_Q(t,s) \begin{pmatrix} 0 \\ D(\alpha)(s) \end{pmatrix} ds \\ & + e^{-Qt}[I - (I - e^{Q\omega})^{-1}] \begin{pmatrix} 0 \\ \alpha'(\omega) - \alpha'(0) \end{pmatrix} \end{aligned}$$

und folglich

$$\begin{aligned} H[T(A)(t) - A(t)] = & \int_0^\omega HG_Q(t,s)H^{-1}H \begin{pmatrix} 0 \\ D(\alpha)(s) \end{pmatrix} ds \\ & + e^{-HQH^{-1}t}[I - (I - e^{HQH^{-1}\omega})^{-1}]H \begin{pmatrix} 0 \\ \alpha'(\omega) - \alpha'(0) \end{pmatrix} \end{aligned}$$

$$
\begin{aligned}
&= \int_0^\omega \underbrace{G_{HQH^{-1}}(t,s)}_{\geq 0} \underbrace{\begin{pmatrix} 0 \\ D(\alpha)(s) \end{pmatrix}}_{\geq 0} ds \\
&\qquad + \underbrace{e^{-HQH^{-1}t}}_{\geq 0}[\underbrace{I-(I-e^{HQH^{-1}\omega})^{-1}}_{\geq 0}] \underbrace{\begin{pmatrix} 0 \\ \alpha'(\omega)-\alpha'(0) \end{pmatrix}}_{\geq 0} \\
&\geq 0
\end{aligned}
$$

und daher $T(A) - A \geq_H 0$ bzw. $A \leq_H T(A)$. Da entsprechend $T(B) \leq_H B$ gezeigt werden kann, ist auch $T(\mathcal{M}) \subset \mathcal{M}$ bewiesen. Aus dem Schauderschen Fixpunktsatz folgt die Existenz eines Fixpunktes $z \in \mathcal{M}$ von T. Mit $z = \begin{pmatrix} x \\ y \end{pmatrix}$ ist x eine Lösung der Liénardschen Differentialgleichung (P), welche den periodischen Randbedingungen $(*)$ genügt und für die $\alpha(t) \leq x(t) \leq \beta(t)$ für alle $t \in [0, \omega]$ gilt. Damit ist der Satz bewiesen. □

Beispiel: Gesucht sei eine Lösung x der Liénardschen Differentialgleichung

$$
\text{(P)} \qquad x'' + 4(1-x^2)x' + x + 0.5x^3 = 0.5\sin t,
$$

welche den periodischen Randbedingungen

$$
(*) \qquad x(0) = x(2\pi), \qquad x'(0) = x'(2\pi)
$$

genügt. Wir wollen Satz 12.16 anwenden und setzen

$$
\alpha(t) := -0.125\cos t - 0.01, \qquad \beta(t) := -0.125\cos t + 0.01.
$$

Dann sind die Bedingungen 1 und 2' in Satz 12.16 mit $\omega := 2\pi$ offensichtlich erfüllt. Aber auch die Bedingung 3' gilt, wie man aus

$$
\begin{aligned}
-D(\alpha)(t) &= \alpha''(t) + 4(1-\alpha(t)^2)\alpha'(t) + \alpha(t) + 0.5\alpha(t)^3 - 0.5\sin t \\
&= \underbrace{\alpha''(t) + 4\alpha'(t) + \alpha(t) - 0.5\sin t}_{=-0.01} + 0.5\cdot\alpha(t)^2[\alpha(t) - 8\alpha'(t)] \\
&= -0.01 - 0.5\cdot[0.125\cos t + 0.01]^2[0.125\cos t + 0.01 + \sin t] \\
&< 0
\end{aligned}
$$

und Abbildung 9 erkennt, in der $-D(\alpha)$ über dem Intervall $[0, 2\pi]$ aufgetragen ist. Entsprechend ist

$$
\begin{aligned}
-D(\beta)(t) &= \beta''(t) + 4(1-\beta(t)^2)\beta'(t) + \beta(t) + 0.5\beta(t)^3 - 0.5\sin t \\
&= \underbrace{\beta''(t) + 4\beta'(t) + \beta(t) - 0.5\sin t}_{=0.01} + 0.5\cdot\beta(t)^2[\beta(t) - 8\beta'(t)] \\
&= 0.01 + 0.5\cdot[-0.125\cos t + 0.01]^2[-0.125\cos t + 0.01 + \sin t] \\
&> 0,
\end{aligned}
$$

wie man aus Abbildung 10 entnimmt, in der $-D(\beta)$ über dem Intervall $[0, 2\pi]$ aufge-

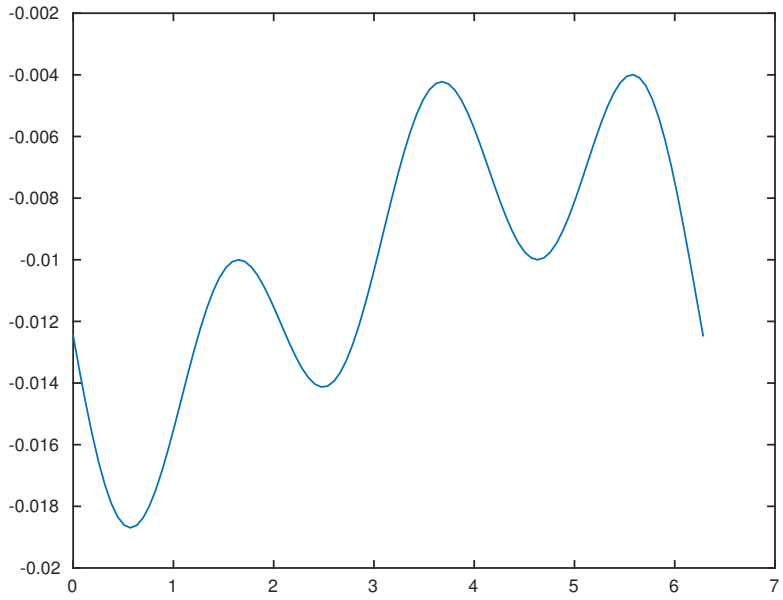

Abbildung 9: $-D(\alpha)$ auf $[0, 2\pi]$

tragen ist. Weiter ist

$$\alpha_{\min} := \min_{t\in[0,2\pi]} \alpha(t) = -0.135, \qquad \beta_{\max} := \max_{t\in[0,2\pi]} \beta(t) = 0.135.$$

Außerdem ist

$$G' := \max_{x\in[\alpha_{\min},\beta_{\max}]} \frac{d}{dx}(x + 0.5x^3) = \max_{|x|\le 0.135} (1 + 1.5x^2) = 1.0273375$$

und folglich

$$2\sqrt{G'} \le 2.02715318 \le 3.9271 \le \min_{|x|\le 0.135} 4(1 - x^2).$$

Damit sind alle Voraussetzungen von Satz 12.16 erfüllt. Daher besitzt (P) eine 2π-periodische Lösung x mit $|x(t) + 0.125 \cos t| \le 0.01$ für alle $t \in [0, 2\pi]$. □

Bisher sind wir bei der Behandlung der Liénardschen Differentialgleichung, einer Differentialgleichung zweiter Ordnung, mit periodischen Randbedingungen

$$x'' + f(x)x' + g(x) = e(t), \qquad x(0) = x(\omega), \quad x'(0) = x'(\omega)$$

so vorgegangen, dass wir zunächst das äquivalente System von zwei Differentialgleichungen erster Ordnung mit periodischen Randbedingungen

$$z' = \begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} -F(x) + y \\ e(t) - g(x) \end{pmatrix} = h(z, t), \qquad z(0) = \begin{pmatrix} x(0) \\ y(0) \end{pmatrix} = \begin{pmatrix} x(\omega) \\ y(\omega) \end{pmatrix} = z(\omega)$$

bildeten, wobei

$$F(x) := \int_0^x f(s)\, ds.$$

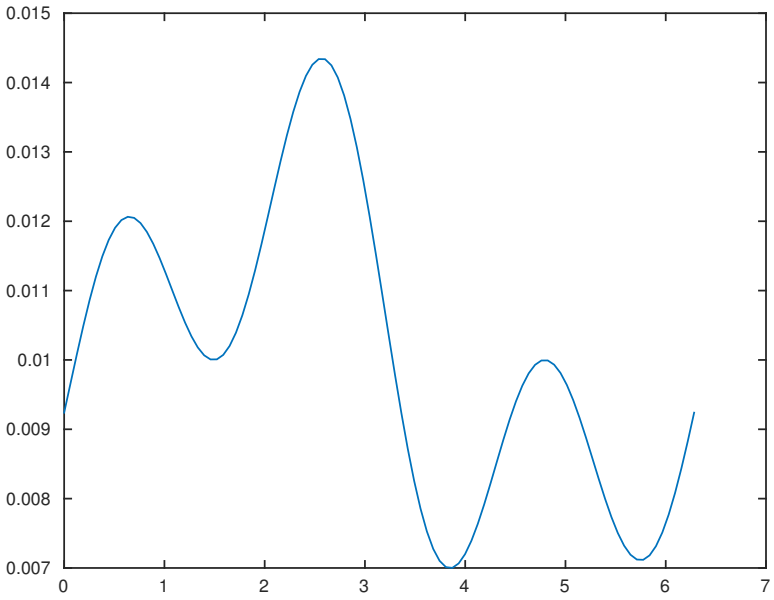

Abbildung 10: $-D(\beta)$ auf $[0, 2\pi]$

Anschließend untersuchten wir, ob durch Addition eines geeigneten linearen Terms Qz auf beiden Seiten des Differentialgleichungssystems, also den Übergang zu der Aufgabe

$$L_Q z := z' + Qz = h(z,t) + Qz, \qquad Rz := z(\omega) - z(0),$$

erreicht werden kann, dass zum einen die homogene Aufgabe

$$L_Q z = 0, \qquad Rz = 0$$

nur trivial lösbar ist und damit die Greensche Matrix $G_Q(t,s)$ zu (L_Q, R) existiert, und zum anderen die durch

$$T(z)(t) := \int_0^\omega G_Q(t,s)[h(z(s),s) + Qz(s)]\,ds$$

definierte Abbildung bezüglich einer geeigneten Halbordnung auf $C_2[0,\omega]$ ein Intervall $\mathcal{M} \subset C_2[0,\omega]$ isoton in sich abgebildet wird.

Ist dagegen der Reibungsterm f konstant, ist die Liénardsche Differentialgleichung also durch

$$x'' + 2Dx' + g(x) = e(t)$$

mit konstantem D gegeben, so liegt es nahe, die äquivalente Gleichung

$$x'' + 2Dx' + C^2 x = e(t) - g(x) + C^2 x$$

mit hinreichend großem C (um die Monotonie der rechten Seite auf einem geeigneten Intervall zu erzwingen) zu betrachten und zunächst zu überlegen, unter welchen

Voraussetzungen die homogene Aufgabe

$$Lx := x'' + 2Dx' + C^2x = 0, \qquad Rx := \begin{pmatrix} x(\omega) - x(0) \\ x'(\omega) - x'(0) \end{pmatrix} = 0$$

nur trivial lösbar ist. Wir setzen voraus, dass $0 < D^2 < C^2$ und definieren $E := \sqrt{C^2 - D^2}$. Dann sind

$$x_1(t) := e^{-Dt}\cos Et, \qquad x_2(t) := e^{-Dt}\sin Et$$

zwei linear unabhängige Lösungen von $Lx = 0$. Eine beliebige Lösung x von $Lx = 0$ lässt sich eindeutig als $x = a_1x_1 + a_2x_2$ mit $a_1, a_2 \in \mathbb{R}$ darstellen. Dann ist

$$\begin{aligned} Rx &= a_1Rx_1 + a_2Rx_2 \\ &= a_1 \begin{pmatrix} e^{-D\omega}\cos E\omega - 1 \\ -e^{-D\omega}(D\cos E\omega + E\sin E\omega) + D \end{pmatrix} \\ &\qquad + a_2 \begin{pmatrix} e^{-D\omega}\sin E\omega \\ e^{-D\omega}(E\cos E\omega - D\sin E\omega) - E \end{pmatrix}. \end{aligned}$$

Da die Matrix

$$A := \begin{pmatrix} e^{-D\omega}\cos E\omega - 1 & e^{-D\omega}\sin E\omega \\ -e^{-D\omega}(D\cos E\omega + E\sin E\omega) + D & e^{-D\omega}(E\cos E\omega - D\sin E\omega) - E \end{pmatrix}$$

aber wegen

$$\begin{aligned} \det A &= [e^{-D\omega}\cos E\omega - 1][e^{-D\omega}(E\cos E\omega - D\sin E\omega) - E] \\ &\qquad - e^{-D\omega}\sin E\omega[-e^{-D\omega}(D\cos E\omega + E\sin E\omega) + D] \\ &= e^{-2D\omega}(E\cos^2 E\omega - D\cos E\omega\sin E\omega) - e^{-D\omega}(2E\cos E\omega - D\sin E\omega) + E \\ &\qquad + e^{-2D\omega}(D\cos E\omega\sin E\omega + E\sin^2 E\omega) - e^{-D\omega}D\sin E\omega \\ &= E(e^{-2D\omega} + 1 - e^{-D\omega}2\cos E\omega) \\ &= E[(e^{-D\omega} - \cos E\omega)^2 + (1 - \cos^2 E\omega)] \\ &> 0 \end{aligned}$$

nichtsingulär ist, folgt aus $Lx = 0$ und $Rx = 0$, dass $x = 0$, die homogene Aufgabe $Lx = 0$, $Rx = 0$ also nur trivial lösbar ist. Wir definieren

$$G(t,s) := \begin{cases} e^{-D(t-s+\frac{\omega}{2})}\Big[a_1\cos E\Big(t - s + \dfrac{\omega}{2}\Big) + a_2\sin E\Big(t - s + \dfrac{\omega}{2}\Big)\Big], & 0 \le t \le s \le \omega, \\ e^{-D(t-s-\frac{\omega}{2})}\Big[a_1\cos E\Big(t - s - \dfrac{\omega}{2}\Big) + a_2\sin E\Big(t - s - \dfrac{\omega}{2}\Big)\Big], & 0 \le s \le t \le \omega, \end{cases}$$

wobei

$$a_1 := \frac{1}{2EF}\sin E\frac{\omega}{2}\cosh D\frac{\omega}{2}, \qquad a_2 := \frac{1}{2EF}\cos E\frac{\omega}{2}\sinh D\frac{\omega}{2}$$

mit

$$F := \cosh D\frac{\omega}{2}\sinh D\frac{\omega}{2}.$$

Wir wollen uns überlegen, dass bei vorgegebenem $r \in C[0,\omega]$ durch

$$x(t) := \int_0^\omega G(t,s) r(s)\, ds$$

die (eindeutige) Lösung von

$$Lx = r, \qquad Rx = 0$$

gegeben ist, d. h. dass $G(t,s)$ die Greensche Funktion zu (L,R) ist. Hierzu beachten wir zunächst, dass $G\colon [0,\omega]\times[0,\omega] \longrightarrow \mathbb{R}$ stetig ist und die partiellen Ableitungen G_t, G_{tt} in jedem der beiden Dreiecke $\{(t,s) : 0 \le t \le s \le \omega\}$ und $\{(t,s) : 0 \le s \le t \le \omega\}$ existieren und stetig sind, wobei natürlich auf der Diagonalen die dem Dreieck entsprechende einseitige Ableitung zu nehmen ist. Weiter gilt:

1. Bei festem $s \in [0,\omega]$ ist $LG(\cdot,s) = 0$.

   Dies ist offensichtlich, da durch

   $$x_1(t) := e^{-Dt}\cos Et, \qquad x_2(t) := e^{-Dt}\sin Et$$

   zwei linear unabhängige Lösungen von $Lx = 0$ gegeben sind.

2. Es gilt die Sprungbedingung

   $$G_t(t+0,t) - G_t(t-0,t) = 1 \quad \text{für alle } t \in (0,\omega).$$

   Denn es ist

$$\begin{aligned}
G_t(t+0,t) - G_t(t-0,t) &= \Big\{-De^{D\frac{\omega}{2}}\Big[a_1\cos E\frac{\omega}{2} - a_2\sin E\frac{\omega}{2}\Big] \\
&\qquad + e^{D\frac{\omega}{2}}\Big[a_1 E\sin E\frac{\omega}{2} + a_2 E\cos E\frac{\omega}{2}\Big]\Big\} \\
&\qquad - \Big\{-De^{-D\frac{\omega}{2}}\Big[a_1\cos E\frac{\omega}{2} + a_2\sin E\frac{\omega}{2}\Big] \\
&\qquad + e^{-D\frac{\omega}{2}}\Big[-a_1 E\sin E\frac{\omega}{2} + a_2 E\cos E\frac{\omega}{2}\Big]\Big\} \\
&= 2\cos E\frac{\omega}{2}\Big[-a_1 D\sinh D\frac{\omega}{2} + a_2 E\cosh D\frac{\omega}{2}\Big] \\
&\qquad + 2\sin E\frac{\omega}{2}\Big[a_2 D\cosh D\frac{\omega}{2} + a_1 E\sinh D\frac{\omega}{2}\Big] \\
&= \frac{1}{EF}\cos E\frac{\omega}{2}\Big[-D\sin E\frac{\omega}{2}\cosh D\frac{\omega}{2}\sinh D\frac{\omega}{2} \\
&\qquad + E\cos E\frac{\omega}{2}\sinh D\frac{\omega}{2}\cosh D\frac{\omega}{2}\Big] \\
&\qquad + \frac{1}{EF}\sin E\frac{\omega}{2}\Big[D\cos E\frac{\omega}{2}\sinh D\frac{\omega}{2}\cosh D\frac{\omega}{2} \\
&\qquad + E\sin E\frac{\omega}{2}\cosh D\frac{\omega}{2}\sinh D\frac{\omega}{2}\Big] \\
&= \frac{1}{F}\cosh D\frac{\omega}{2}\sinh D\frac{\omega}{2}\Big[\underbrace{\cos^2 E\frac{\omega}{2} + \sin^2 E\frac{\omega}{2}}_{=1}\Big] \\
&= 1
\end{aligned}$$

   nach Definition von F.

3. Bei festem $s \in [0,\omega]$ ist $RG(\cdot,s) = 0$.

   Hierzu ist zu zeigen, dass

$$G(\omega,s) - G(0,s) = 0, \qquad G_t(\omega,s) - G_t(0,s) = 0.$$

   Dies ist aber offensichtlich richtig.

Wir zeigen nun, dass bei vorgegebenem $r \in C[0,\omega]$ durch

$$x(t) := \int_0^\omega G(t,s)r(s)\,ds$$

die (eindeutige) Lösung von

$$Lx = r, \qquad Rx = 0$$

gegeben ist. Denn wegen

$$x(t) = \int_0^t G(t,s)r(s)\,ds + \int_t^\omega G(t,s)r(s)\,ds$$

ist

$$\begin{aligned} x'(t) &= G(t,t)r(t) + \int_0^t G_t(t,s)r(s)\,ds - G(t,t)r(t) + \int_t^\omega G_t(t,s)r(s)\,ds \\ &= \int_0^t G_t(t,s)r(s)\,ds + \int_t^\omega G_t(t,s)r(s)\,ds. \end{aligned}$$

Durch erneutes Differenzieren erhält man

$$\begin{aligned} x''(t) &= G_t(t+0,t)r(t) + \int_0^t G_{tt}(t,s)r(s)\,ds \\ &\qquad - G_t(t-0,t)r(t) + \int_t^\omega G_{tt}(t,s)r(s)\,ds \\ &= [\underbrace{G_t(t+0,0) - G_t(t-0,t)}_{=1}]r(t) + \int_0^\omega G_{tt}(t,s)r(s)\,ds \\ &= r(t) + \int_0^\omega G_{tt}(t,s)r(s)\,ds. \end{aligned}$$

Folglich ist

$$\begin{aligned} Lx(t) &= x''(t) + 2Dx'(t) + C^2x(t) \\ &= r(t) + \int_0^\omega \underbrace{LG(t,s)}_{=0} r(s)\,ds \\ &= r(t), \end{aligned}$$

d. h. es ist $Lx = r$. Weiter ist

$$Rx = \begin{pmatrix} x(\omega) - x(0) \\ x'(\omega) - x'(0) \end{pmatrix} = \begin{pmatrix} \int_0^\omega [G(\omega,s) - G(0,s)]r(s)\,ds \\ \int_0^\omega [G_t(\omega,s) - G_t(0,s)]r(s)\,ds \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = 0.$$

Damit ist gezeigt, dass x die Lösung von $Lx = r$, $Rx = 0$ bzw. G die Greensche Funktion zu (L, R) ist.
Nun ist es nicht mehr schwierig, den folgenden Satz zu beweisen.

Satz 12.17 *Gegeben sei die Differentialgleichung*

$$\text{(P)} \qquad x'' + 2Dx' + g(x) = e(t)$$

mit konstantem D. Gesucht sei eine Lösung von (P), welche den periodischen Randbedingungen

$$(*) \qquad x(0) = x(\omega), \quad x'(0) = x'(\omega)$$

mit vorgegebenem $\omega > 0$ genügt. Seien α, $\beta \in C^2[0, \omega]$ zwei Funktionen mit:

1. *Es ist $\alpha(t) \leq \beta(t)$ für alle $t \in [0, \omega]$.*

2'. *Es ist*

$$\alpha(0) = \alpha(\omega), \quad \alpha'(0) \leq \alpha'(\omega), \quad \beta(0) = \beta(\omega), \quad \beta'(0) \geq \beta'(\omega).$$

3'. *Es ist*

$$-D(\alpha)(t) \leq 0 \leq -D(\beta)(t) \quad \text{für alle } t \in [0, \omega],$$

wobei für $y \in C^2[0, \omega]$ der Defekt *$D(y)$ durch*

$$D(y)(t) := -y''(t) - 2Dy'(t) - g(y(t)) + e(t)$$

gegeben ist.

Mit

$$\alpha_{\min} := \min_{t \in [0,\omega]} \alpha(t), \qquad \beta_{\max} := \max_{t \in [0,\omega]} \beta(t)$$

seien $e \in C[0, \omega]$ und $g \in C^1[\alpha_{\min}, \beta_{\max}]$. Ferner sei

$$g'(x) \leq \frac{\pi^2}{4\omega^2} + D^2 \quad \text{für alle } x \in [\alpha_{\min}, \beta_{\max}].$$

Dann besitzt (P) eine Lösung x, welche den periodischen Randbedingungen $(*)$ genügt, mit $\alpha(t) \leq x(t) \leq \beta(t)$ für alle $t \in [0, \omega]$.

Beweis: Wir definieren

$$C := \sqrt{\frac{\pi^2}{4\omega^2} + D^2}.$$

Mit

$$Lx := x'' + 2Dx' + C^2 x, \qquad Rx := \begin{pmatrix} x(\omega) - x(0) \\ x'(\omega) - x'(0) \end{pmatrix}$$

sei G die Greensche Funktion zu (L,R). Wie wir oben nachgewiesen haben, ist

$$G(t,s) := \begin{cases} e^{-D(t-s+\frac{\omega}{2})}\Big[a_1\cos E\Big(t-s+\dfrac{\omega}{2}\Big)+a_2\sin E\Big(t-s+\dfrac{\omega}{2}\Big)\Big], & 0\le t\le s\le\omega,\\ e^{-D(t-s-\frac{\omega}{2})}\Big[a_1\cos E\Big(t-s-\dfrac{\omega}{2}\Big)+a_2\sin E\Big(t-s-\dfrac{\omega}{2}\Big)\Big], & 0\le s<t\le\omega, \end{cases}$$

wobei

$$E := \sqrt{C^2-D^2} = \frac{\pi}{2\omega}$$

sowie

$$a_1 := \frac{1}{2EF}\sin E\frac{\omega}{2}\cosh D\frac{\omega}{2}, \qquad a_2 := \frac{1}{2EF}\cos E\frac{\omega}{2}\sinh D\frac{\omega}{2}$$

mit

$$F := \cosh D\frac{\omega}{2}\sinh D\frac{\omega}{2}.$$

Offenbar ist

$$a_1 = \frac{\sqrt{2}\omega}{2\pi\sinh(D\omega/2)}, \qquad a_2 = \frac{\sqrt{2}\omega}{2\pi\cosh(D\omega/2)}.$$

Wir wollen zeigen, dass $G(t,s)\ge 0$ für alle $(t,s)\in[0,\omega]\times[0,\omega]$, die Greensche Funktion also nichtnegativ ist. Hierzu geben wir uns $(t,s)\in[0,\omega]\times[0,\omega]$ beliebig vor und setzen

$$\xi := \begin{cases} t-s+\dfrac{\omega}{2}, & t\le s,\\ t-s-\dfrac{\omega}{2}, & s<t. \end{cases}$$

Dann ist

$$G(t,s) = e^{-D\xi}[a_1\cos E\xi + a_2\sin E\xi].$$

Offensichtlich ist $\xi\in[-\omega/2,\omega/2]$ und daher $E\xi\in[-\pi/4,\pi/4]$. Mit

$$\sigma := \sin E\xi \in \Big[-\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}\Big]$$

ist dann

$$\begin{aligned} G(t,s) &= e^{-D\xi}[a_1\cos E\xi + a_2\sin E\xi]\\ &= e^{-D\xi}[a_1\sqrt{1-\sigma^2}+a_2\sigma]\\ &\ge e^{-D\xi}[a_1|\sigma|+a_2\sigma]\\ &\qquad\qquad (\text{wegen } |\sigma|\le\sqrt{2}/2)\\ &\ge a_2e^{-D\xi}[|\sigma|+\sigma]\\ &\qquad\qquad (\text{wegen } a_1\ge a_2)\\ &\ge 0. \end{aligned}$$

Damit ist nachgewiesen, dass die Greensche Fuktion G zu (L,R) nichtnegativ ist. Auf dem Banachraum $C[0,\omega]$ (versehen mit der Maximumnorm) definieren wir das (nichtleere, abgeschlossene, konvexe und beschränkte) Intervall

$$\mathcal{M} := \{x\in C[0,\omega] : \alpha(t)\le x(t)\le\beta(t) \ \text{ für alle } t\in[0,\omega]\}$$

sowie die Abbildung $T\colon \mathcal{M} \longrightarrow C[0,\omega]$ durch

$$T(x)(t) := \int_0^\omega G(t,s)[e(s) - g(x(s)) + C^2 x(s)]\,ds.$$

Da $T(\mathcal{M})$ relativ kompakt ist, liefert der Schaudersche Fixpunktsatz 10.16 die Existenz eines Fixpunktes $x \in \mathcal{M}$ von T bzw. einer Lösung von (P), welche den periodischen Randbedingungen $(*)$ genügt, wenn noch $T(\mathcal{M}) \subset \mathcal{M}$ gezeigt werden kann. Wegen $-g'(x) + C^2 \geq 0$ für alle $x \in [\alpha_{\min}, \beta_{\max}]$ (bzw. der Monotonie von $-g(x) + C^2 x$ auf $[\alpha_{\min}, \beta_{\max}]$) und der Nichtnegativität der Greenschen Funktion G ist T auf $\mathcal{M}$ isoton. Aus $x \in \mathcal{M}$ folgt also $T(\alpha)(t) \leq T(x)(t) \leq T(\beta)(t)$ für alle $t \in [0,\omega]$. Zu zeigen bleibt, dass $\alpha(t) \leq T(\alpha)(t)$ und $T(\beta)(t) \leq \beta(t)$ für alle $t \in [0,\omega)$. Zur Abkürzung definieren wir $\alpha_1(t) := T(\alpha)(t)$. Nach Definition der Abbildung T ist

$$\alpha_1''(t) + 2D\alpha_1'(t) + C^2\alpha_1(t) = e(t) - g(\alpha(t)) + C^2\alpha(t) \quad \text{für alle } t \in [0,\omega]$$

sowie

$$\alpha_1(0) = \alpha_1(\omega), \quad \alpha_1'(0) = \alpha_1'(\omega).$$

Wegen Bedingung 3'. ist

$$\alpha''(t) + 2D\alpha'(t) + g(\alpha(t)) - e(t) \leq 0 \quad \text{für alle } t \in [0,\omega].$$

Definiert man daher $\delta(t) := \alpha_1(t) - \alpha(t)$ und berücksichtigt die Bedingung 2'., so erhält man

$$r(t) := \delta''(t) + 2D\delta'(t) + C^2\delta(t) \geq 0, \qquad \delta(0) - \delta(\omega) = 0, \quad \delta'(0) - \delta'(\omega) \geq 0.$$

Also ist

$$\delta(t) = q(t) + \int_0^\omega \underbrace{G(t,s)}_{\geq 0}\underbrace{r(s)}_{\geq 0}\,ds,$$

wobei q die (eindeutige) Lösung von

$$q'' + 2Dq' + C^2 q = 0, \qquad q(0) - q(\omega) = 0, \quad q'(0) - q'(\omega) = \sigma_1 := \delta'(0) - \delta'(\omega)$$

ist. Wir zeigen, dass q und damit auch δ auf $[0,\omega]$ nichtnegativ ist bzw. $\alpha(t) \leq T(\alpha)(t)$ für alle $t \in [0,\omega]$ gilt. Da durch

$$x_1(t) := e^{-D(t-\frac{\omega}{2})}\cos E\Big(t - \frac{\omega}{2}\Big), \qquad x_2(t) := e^{-D(t-\frac{\omega}{2})}\sin E\Big(t - \frac{\omega}{2}\Big)$$

mit

$$E := \sqrt{C^2 - D^2} = \frac{\pi}{2\omega}$$

ein Fundamentalsystem zu $Lx = 0$ gegeben ist, machen wir für q den Ansatz $q = c_1x_1 + c_2x_2$ mit konstanten c_1, c_2. Wegen

$$q(0) - q(\omega) = c_1\sqrt{2}\sinh D\frac{\omega}{2} - c_2\sqrt{2}\cosh D\frac{\omega}{2}$$

und

$$\begin{aligned} q'(0) - q'(\omega) &= c_1\sqrt{2}\Big[-D\sinh D\frac{\omega}{2} + E\cosh D\frac{\omega}{2}\Big] \\ &\qquad + c_2\sqrt{2}\Big[D\cosh D\frac{\omega}{2} + E\sinh D\frac{\omega}{2}\Big] \end{aligned}$$

erhalten wir aus

$$q(0) - q(\omega) = 0, \qquad q'(0) - q'(\omega) = \sigma_1$$

für c_1, c_2 das lineare Gleichungssystem

$$\begin{pmatrix} \sinh D\frac{\omega}{2} & -\cosh D\frac{\omega}{2} \\ -D\sinh D\frac{\omega}{2} + E\cosh D\frac{\omega}{2} & D\cosh D\frac{\omega}{2} + E\sinh D\frac{\omega}{2} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{\sigma_1}{\sqrt{2}} \end{pmatrix}.$$

Mit

$$\Delta := E\Big(\sinh^2 D\frac{\omega}{2} + \cosh^2 D\frac{\omega}{2}\Big) > 0$$

ist dann

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \frac{1}{\Delta} \begin{pmatrix} D\cosh D\frac{\omega}{2} + E\sinh D\frac{\omega}{2} & \cosh D\frac{\omega}{2} \\ D\sinh D\frac{\omega}{2} - E\cosh D\frac{\omega}{2} & \sinh D\frac{\omega}{2} \end{pmatrix} \begin{pmatrix} 0 \\ \frac{\sigma_1}{\sqrt{2}} \end{pmatrix},$$

also

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \frac{\sigma_1}{\sqrt{2}\Delta} \begin{pmatrix} \cosh D\frac{\omega}{2} \\ \sinh D\frac{\omega}{2} \end{pmatrix}.$$

Für $t \in [0,\omega]$ ist $\xi := t - \omega/2 \in [-\omega/2, \omega/2]$, damit $E\xi \in [-\pi/4, \pi/4]$, und daher $\sigma := \sin E\xi \in [-\sqrt{2}/2, \sqrt{2}/2]$. Folglich ist

$$\begin{aligned} q(t) &= c_1x_1(t) + c_2x_2(t) \\ &= \frac{\sigma_1}{\sqrt{2}\Delta} e^{-D(t-\frac{\omega}{2})}\Big[\cosh D\frac{\omega}{2}\cos E\Big(t-\frac{\omega}{2}\Big) + \sinh D\frac{\omega}{2}\sin E\Big(t-\frac{\omega}{2}\Big)\Big] \\ &= \frac{\sigma_1}{\sqrt{2}\Delta} e^{-D\xi}\Big[\cosh D\frac{\omega}{2}\cos E\xi + \sinh D\frac{\omega}{2}\sin E\xi\Big] \\ &= \frac{\sigma_1}{\sqrt{2}\Delta} e^{-D\xi}\Big[\sqrt{1-\sigma^2}\cosh D\frac{\omega}{2} + \sigma\sinh D\frac{\omega}{2}\Big] \\ &\geq \frac{\sigma_1}{\sqrt{2}\Delta} e^{-D\xi}\Big[|\sigma|\cosh D\frac{\omega}{2} + \sigma\sinh D\frac{\omega}{2}\Big] \\ &\qquad (\text{wegen } |\sigma| \leq \sqrt{2}/2) \\ &\geq \frac{\sigma_1}{\sqrt{2}\Delta} e^{-D\xi}\sinh D\frac{\omega}{2}[|\sigma| + \sigma] \\ &\geq 0. \end{aligned}$$

Also ist q und damit auch $\delta := T(\alpha) - \alpha$ auf $[0, \omega]$ nichtnegativ[46]. Da entsprechend auch die Nichtnegativität von $\beta - T(\beta)$ auf $[0, \omega]$ bewiesen werden kann, ist $T(\mathcal{M}) \subset \mathcal{M}$. Wie oben schon angegeben folgt die Existenz eines Fixpunktes $x \in \mathcal{M}$ von T bzw. einer Lösung $x \in \mathcal{M}$, welche den periodischen Randbedingungen $(*)$ genügt, aus dem Schauderschen Fixpunktsatz 10.16. Der Satz ist damit bewiesen. □

Beispiel: Gesucht sei eine Lösung x der Liénardschen Differentialgleichung

$$x'' + 2x' + x + 0.1x^3 = \sin t, \tag{P}$$

welche den periodischen Randbedingungen

$$x(0) = x(2\pi), \qquad x'(0) = x'(2\pi) \tag{$*$}$$

genügt. Mit $\omega := 2\pi$, $D := 1$ und

$$\alpha(t) := -0.5\cos t - 0.012, \qquad \beta(t) := -0.5\cos t + 0.012$$

sind die Bedingungen 1, 2' und 3' der Sätze 12.16 bzw. 12.17 mit $\omega := 2\pi$ erfüllt. Dies ist für die beiden ersten Bedingungen offensichtlich. Aber auch die Bedingung 3' gilt, wie man aus

$$\begin{aligned} -D(\alpha)(t) &= \alpha''(t) + 2\alpha'(t) + \alpha(t) + 0.1\alpha(t)^3 - \sin t \\ &= \underbrace{\alpha''(t) + 2\alpha'(t) + \alpha(t) - \sin t}_{=-0.012} + 0.1\alpha(t)^3 \\ &= -0.012 + 0.2(-0.5\cos t - 0.012)^3 \\ &< 0, \end{aligned}$$

und Abbildung 11 erkennt, in der $-D(\alpha)$ über dem Intervall $[0, 2\pi]$ aufgetragen ist. Entsprechend ist

$$\begin{aligned} -D(\beta)(t) &= \beta''(t) + 2\beta'(t) + \beta(t) + 0.1\beta(t)^3 - \sin t \\ &= \underbrace{\beta''(t) + 2\beta'(t) + \beta(t) - \sin t}_{=0.012} + 0.1\beta(t)^3 \\ &= 0.012 + 0.1(-0.5\cos t + 0.012)^3 \\ &> 0 \quad , \end{aligned}$$

wie man aus Abbildung 12 erkennt, in der $-D(\beta)$ über dem Intervall $[0, 2\pi]$ aufgetragen ist. Die Voraussetzungen von Satz 12.17 sind erfüllt. Denn mit

$$\alpha_{\min} := \min_{t\in[0,2\pi]} \alpha(t) = -0.512, \qquad \beta_{\max} := \max_{t\in[0,2\pi]} \beta(t) = 0.512$$

und

$$g(x) := x + 0.1x^3$$

[46]Man beachte, dass die Nichtnegativität von q ganz ähnlich wie die der Greenschen Funktion G bewiesen wird.

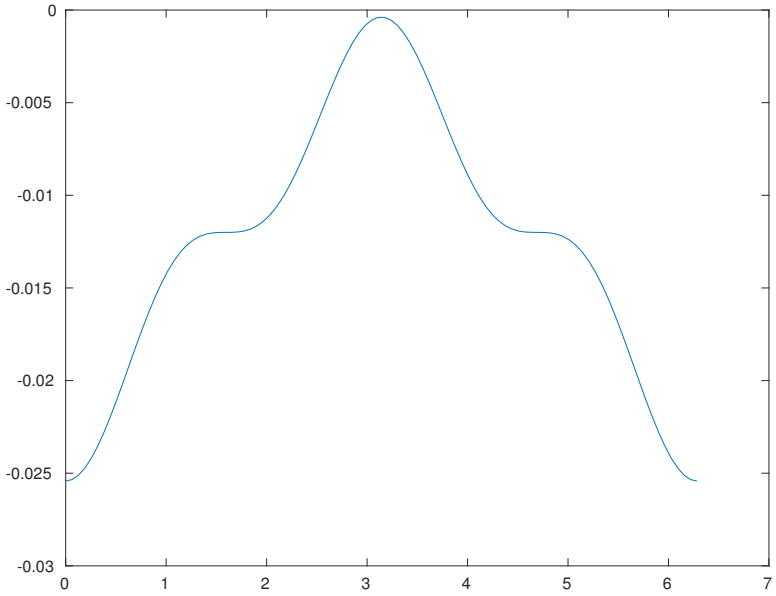

Abbildung 11: $-D(\alpha)$ auf $[0, 2\pi]$

ist

$$\max_{x\in[\alpha_{\min},\beta_{\max}]} g'(x) = 1 + 0.3 \cdot 0.512^2 = 1.0402653184,$$

während

$$\frac{\pi^2}{4\omega^2} + D^2 = \frac{1}{16} + 1 = 1.0625.$$

Dagegen sind die Voraussetzungen von Satz 12.16 *nicht* erfüllt. □

12.7 Nichtnegative Greensche Matrizen bei Randwertaufgaben mit periodischen Randbedingungen

In diesem Unterabschnitt wollen wir die Ergebnisse von J. Werner (1972) darstellen. Wir betrachten das lineare Differentialgleichungssystem mit periodischen Randbedingungen

$$\text{(P)} \qquad x' = A(t)x + r(t), \qquad x(0) = x(\omega),$$

wobei $\omega > 0$ vorgegenen ist und $A \in C_{n\times n}[0,\omega]$ eine auf $[0,\omega]$ stetige $n \times n$-Matrixfunktion ist, mit $A(t) = (a_{ij}(t))$ also $a_{ij} \in C[0,\omega]$, $i,j = 1,\dots,n$, ist. Die Aufgabe (P) besitzt für jede stetige Vektorfunktion $r \in C_n[0,\omega]$ genau dann eine Lösung x, wenn das homogene Problem (hier ist $r := 0$) nur die triviale Lösung besitzt. In diesem Fall lässt sich x mit Hilfe der Greenschen Funktion G in der Form

$$x(t) = \int_0^\omega G(t,s)r(s)\,ds$$

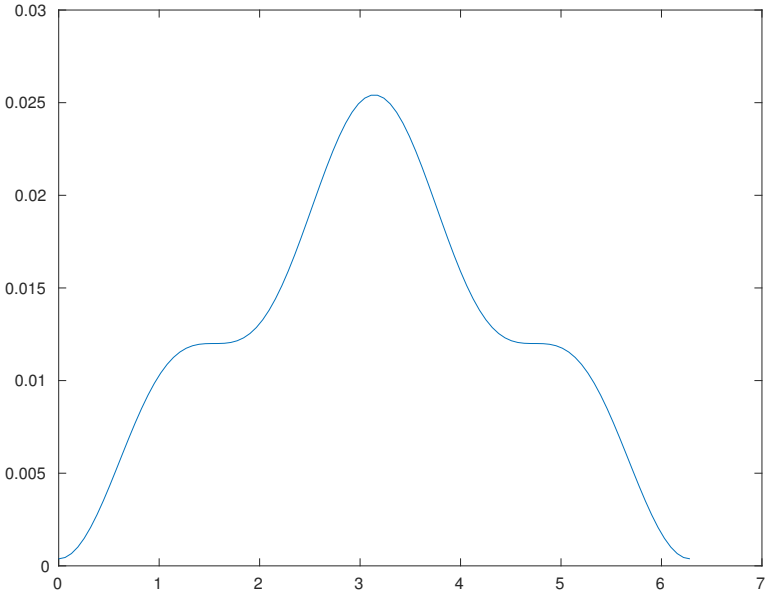

Abbildung 12: $-D(\beta)$ auf $[0, 2\pi]$

darstellen, wobei (siehe Satz 12.5) mit einem durch $\Phi(0) = I$ normierten Fundamentalsystem Φ zu $x' = A(t)x$ die Greensche Matrix durch

$$G(t,s) := \Phi(t) \left\{ \begin{array}{l} (I - \Phi(\omega))^{-1} \\ (I - \Phi(\omega))^{-1}\Phi(\omega) \end{array} \right\} \Phi(s)^{-1}, \qquad \begin{array}{l} s < t, \\ t < s \end{array}$$

gegeben ist.
Sei $K \subset \mathbb{R}^n$ ein Ordnungskegel (siehe Definition 12.6) im $\mathbb{R}^n$. Dieser induziert durch

$$x \leq_K y \quad \text{für } x, y \in \mathbb{R}^n \iff y - x \in K$$

in kanonischer Weise die Halbordnung $\leq_K$ im $\mathbb{R}^n$. Durch K wird mittels

$$K_c := \{x \in C_n[0,\omega] : x(t) \in K \text{ für alle } t \in [0,\omega]\}$$

in $C_n[0,\omega]$ ein Ordnungskegel induziert. Wir wollen hinreichende und zum Teil auch notwendige Bedingungen dafür angeben, dass die durch

$$G(r)(t) := \int_0^\omega G(t,s)r(s)\,ds$$

definierte Abbildung $G\colon C_n[0,\omega] \longrightarrow C_n[0,\omega]$ den Ordnungskegel K_c in sich abbildet.
Mit einer Vektornorm $\|\cdot\|$ auf $\mathbb{R}^n$ sei $C_n[0,\omega]$ durch

$$\|x\| := \max_{t\in[0,\omega]} \|x(t)\|$$

für $x \in C_n[0,\omega]$ normiert.

Definition 12.18 Die Matrixfunktion $A \in C_{n\times n}[0,\omega]$ heißt *quasipositiv* bezüglich des Ordnungskegels $K \subset \mathbb{R}^n$, wenn es eine Konstante $\alpha \geq 0$ mit

$$(A(\cdot) + \alpha I)x \in K_c \quad \text{für alle } x \in K$$

gibt.

Bezeichnet

$$\mathbb{R}^n_+ := \{x = (x_j) \in \mathbb{R}^n : x_j \geq 0,\ j = 1, \ldots, n\}$$

den nichtnegativen Orthanten im $\mathbb{R}^n$, so ist $A = (a_{ij}) \in C_n[0,\omega]$ offenbar genau dann bezüglich $\mathbb{R}^n_+$ quasipositiv, wenn $a_{ij}(t) \geq 0$ für alle $t \in [0,\omega]$ und alle $i, j \in \{1, \ldots, n\}$ mit $i \neq j$. Bei N. J. Higham (2008, S. 260) heißt eine Matrix *essentially nonnegative*, wenn ihre Nebendiagonalelemente nichtnegativ sind.

Lemma 12.19 *Sei $A \in C_{n\times n}[0,\omega]$ quasipositiv bezüglich des abgeschlossenen Ordnungskegels $K \subset \mathbb{R}^n$, $r \in K_c$ und $x_0 \in K$. Bezeichnet x die Lösung der linearen Anfangswertaufgabe*

$$x' = A(t)x + r(t), \qquad x(0) = x_0,$$

so ist $x \in K_c$.

Beweis: Da A nach Voraussetzung quasipositiv (bezüglich K) ist, existiert ein $\alpha \geq 0$ mit $(A(\cdot) + \alpha I)x \in K_c$ für alle $x \in K$. Wir definieren die Abbildung $T\colon C_n[0,\omega] \longrightarrow C_n[0,\omega]$ durch

$$T(z)(t) := e^{-\alpha t}x_0 + \int_0^t e^{-\alpha(t-s)}[(A(s) + \alpha I)z(s) + r(s)]\, ds.$$

Offenbar ist $T(K_c) \subset K_c$. Wir zeigen, dass T bezüglich einer geeigneten Norm kontrahierend (auf dem *gesamten* Raum $C_n[0,\omega]$) ist. Mit einem noch geeignet zu bestimmenden $k > 0$ definiere man die Norm $\|\cdot\|_k$ durch

$$\|x\|_k := \max_{t\in[0,\omega]} e^{-kt}\|x(t)\|.$$

Für beliebige $z_1, z_2 \in C_n[0,\omega]$ ist dann

$$e^{-kt}[T(z_1)(t) - T(z_2)(t)] = \int_0^t e^{-(\alpha+k)(t-s)}(A(s) + \alpha I)e^{-ks}[z_1(s) - z_2(s)]\, ds$$

und daher

$$\|T(z_1) - T(z_2)\|_k \leq \frac{C}{\alpha + k}\|z_1 - z_2\|_k,$$

wobei $C \geq 0$ eine nichtnegative (von z_1, z_2 und k unabhängige) Konstante ist. Für hinreichend großes k ist T also auf $C_n[0,\omega]$ bezüglich der Norm $\|\cdot\|_k$ kontrahierend. Wegen $T(K_c) \subset K_c$, der Abgeschlossenheit von K_c (da K abgeschlossen ist) und des Banachschen Fixpunktsatzes muss also der einzige Fixpunkt x von T in K_c liegen. Da x auch die Lösung der gegebenen Anfangswertaufgabe ist, ist das Lemma bewiesen. □

Wendet man Lemma 12.19 mit $r := 0$ an, so erhält man:

Lemma 12.20 *Sei $A \in C_{n\times n}[0,\omega]$ quasipositiv bezüglich des abgeschlossenen Ordnungskegels $K \subset \mathbb{R}^n$. Sei Φ ein durch $\Phi(0) = I$ normiertes Fundamentalsystem zu $x' = A(t)x$. Dann ist $\Phi(\cdot)x_0 \in K_c$ für alle $x_0 \in K$.*

Ist im obigen Lemma $A \in C_{n\times n}[0,\omega]$ quasipositiv bezüglich des nichtnegativen Orthanten $\mathbb{R}^n_+$, so ist $\Phi(t) \in \mathbb{R}^{n\times n}$ für alle $t \in [0,\omega]$ eine nichtnegative Matrix, d. h. alle Einträge von $\Phi(t)$ sind nichtnegativ. Im Falle einer konstanten Matrixfunktion gilt auch die Umkehrung, wie das nächste Lemma aussagt (siehe auch N. J. Higham (2008, S. 260)).

Lemma 12.21 *Die Matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ ist genau dann quasipositiv bezüglich des nichtnegativen Orthanten $\mathbb{R}^n_+$ (bzw. $a_{ij} \geq 0$ für alle $i,j \in \{1,\ldots,n\}$ mit $i \neq j$), wenn das durch $\Phi(0) = I$ normierte Fundamentalsystem $\Phi(t) = e^{At}$ für alle $t \geq 0$ (elementweise) nichtnegativ ist.*

Beweis: Zu zeigen bleibt:

- Ist e^{At} elementweise nichtnegativ für alle $t \geq 0$, so ist $a_{ij} \geq 0$ für alle $i,j \in \{1,\ldots,n\}$ mit $i \neq j$.

Die Behauptung folgt wegen $e^{At} = I + At + O(t^2)$. □

Im folgenden Satz wird eine hinreichende Bedingung dafür angegeben, dass die Greensche Matrix $G(t,s)$ zur linearen, periodischen Randwertaufgabe (P) existiert und die durch

$$(*) \qquad G(r)(t) := \int_0^\omega G(t,s)r(s)\,ds$$

definierte Abbildung $G\colon C_n[0,\omega] \longrightarrow C_n[0,\omega]$ den Ordnungskegel K_c in sich abbildet. Mit $\rho(C)$ wird der *Spektralradius* der $n \times n$-Matrix C bezeichnet.

Satz 12.22 *Sei $A \in C_{n\times n}[0,\omega]$ quasipositiv bezüglich des abgeschlossenen Ordnungskegels $K \subset \mathbb{R}^n$ und Φ ein durch $\Phi(0) = I$ normiertes Fundamentalsystem zu $x' = A(t)x$. Ist $\rho(\Phi(\omega)) < 1$, so existiert die Greensche Matrix $G(t,s)$ zu (P) und die durch $(*)$ definierte Abbildung $G\colon C_n[0,\omega] \longrightarrow C_n[0,\omega]$ bildet den (abgeschlossenen) Ordnungskegel $K_c \subset C_n[0,\omega]$ in sich ab.*

Beweis: Wegen $\rho(\Phi(\omega)) < 1$ ist $I - \Phi(\omega)$ nichtsingulär. Daher ist das homogene Problem $x' = A(t)x$, $x(0) = x(\omega)$ nur trivial lösbar. Folglich existiert die Greensche Matrix $G(t,s)$ zu (P). Sei $r \in K_c$ gegeben. Dann ist $x := G(r)$ die (eindeutige) Lösung von

$$x' = A(t)x + r(t), \qquad x(0) = x(\omega).$$

Sei y die Lösung von

$$y' = A(t)y + r(t), \qquad y(0) = 0.$$

Wegen Lemma 12.19 ist $y \in K_c$. Definiert man z durch

$$z(t) := \Phi(t)(I - \Phi(\omega))^{-1}y(\omega) + y(t),$$

so ist

$$\begin{aligned} z'(t) &= \Phi'(t)(I-\Phi(\omega))^{-1}y(\omega) + y'(t) \\ &= A(t)\Phi(t)(I-\Phi(\omega))^{-1}y(\omega) + A(t)y(t) + r(t) \\ &= A(t)[\underbrace{\Phi(t)(I-\Phi(\omega))^{-1}y(\omega) + y(t)}_{=z(t)}] + r(t) \\ &= A(t)z(t) + r(t) \end{aligned}$$

und weiter

$$\begin{aligned} z(\omega) - z(0) &= \Phi(\omega)(I-\Phi(\omega))^{-1}y(\omega) + y(\omega) - (I-\Phi(\omega))^{-1}y(\omega) \\ &= [\underbrace{\Phi(\omega)(I-\Phi(\omega))^{-1} + I - (I-\Phi(\omega))^{-1}}_{=0}]y(\omega) \\ &= 0. \end{aligned}$$

Daher ist

$$G(r)(t) = z(t) = \Phi(t)(I-\Phi(\omega))^{-1}y(\omega) + y(t).$$

Wir überlegen uns nun, dass $(I-\Phi(\omega))^{-1}y(\omega) \in K$. Denn wegen $\rho(\Phi(\omega)) < 1$ ist

$$(I-\Phi(\omega))^{-1}y(\omega) = \sum_{i=0}^{\infty} \Phi(\omega)^i y(\omega).$$

Wegen Lemma 12.20 ist $\Phi(\omega)^i y(\omega) \in K$, $i = 0, \ldots$. Wegen $K + K \subset K$ und der Abgeschlossenheit von K folgt $(I-\Phi(\omega))^{-1}y(\omega) \in K$. Damit ist

$$\Phi(\cdot)(I-\Phi(\omega))^{-1}y(\omega) \in K_c.$$

Wegen $y \in K_c$ und $K_c + K_c \subset K_c$ ist damit $G(r) \in K_c$ und der Satz ist bewiesen. □

Nun wollen wir uns dem Fall zuwenden, dass $A(t) = A$ eine konstante Matrix ist. Wir erinnern an die Definition einer M-Matrix bezüglich eines vorgegebenen Ordnungskegels, siehe Definition 11.4.

Lemma 12.23 *Sei $K \subset \mathbb{R}^n$ ein abgeschlossener Ordnungskegel mit* int $(K) \neq \emptyset$. *Ferner sei $A \in \mathbb{R}^{n\times n}$ eine Matrix mit der Eigenschaft, dass $-A$ eine M-Matrix bezüglich K ist. Bei vorgegebenem $\omega > 0$ existiert dann die Greensche Matrix $G(t,s)$ zu*

$$\text{(P)} \qquad x' = Ax + r(t), \qquad x(0) = x(\omega)$$

und die Abbildung $G\colon C_n[0,\omega] \longrightarrow C_n[0,\omega]$, definiert durch

$$(*) \qquad G(r)(t) := \int_0^\omega G(t,s)r(s)\,ds,$$

bildet den Ordnungskegel

$$K_c := \{x \in C_n[0,\omega] : x(t) \in K \text{ für alle } t \in [0,\omega]\}$$

in sich ab.

Beweis: Wir wenden Satz 12.22 an. Da $\Phi(t) := e^{At}$ ein durch $\Phi(0) = I$ normiertes Fundamentalsystem zu $x' = Ax$ ist, genügt es nachzuweisen, dass $\rho(e^{A\omega}) < 1$. Wegen Satz 11.5 (und der Voraussetzung, dass $-A$ eine M-Matrix bezüglich K ist) ist $\Re(\lambda) < 0$ für jeden Eigenwert λ von A. Die Eigenwerte von $e^{A\omega}$ sind durch $e^{\lambda\omega}$ mit einem Eigenwert λ von A gegeben. Für diese gilt $\Re(\lambda) < 0$ und daher ist $\rho(e^{A\omega}) < 1$. Damit ist das Lemma bewiesen. □

Für den Fall, dass $K = \mathbb{R}^n_+$ der natürliche Ordnungskegel ist, können wir eine Umkehrung von Lemma 12.23 beweisen.

Satz 12.24 *Sei $A \in \mathbb{R}^{n\times n}$ eine Matrix mit der Eigenschaft, dass die Greensche Matrix $G(t,s)$ zu*

$$\text{(P)} \qquad x' = Ax + r(t), \qquad x(0) = x(\omega)$$

für jedes $\omega > 0$ existiert und die durch

$$(*) \qquad G(r)(t) := \int_0^\omega G(t,s)r(s)\,ds$$

definierte Abbildung $G\colon C_n[0,\omega] \longrightarrow C_n[0,\omega]$ (komponentenweise) nichtnegative Elemente aus $C_n[0,\omega]$ in sich abbildet. Dann ist $-A$ eine M-Matrix (bezüglich des natürlichen Ordnungskegels $\mathbb{R}^n_+$).

Beweis: Die Greensche Matrix $G(t,s)$ zu (P) existiert bei gegebenem $\omega > 0$ genau dann, wenn die homogene Aufgabe $x' = Ax$, $x(0) = x(\omega)$, nur trivial lösbar bzw. $I - e^{A\omega}$ nichtsingulär ist. Dies ist genau dann der Fall, wenn $e^{\lambda\omega} \neq 1$ bzw. $\lambda\omega \neq 2k\pi i$ für alle $k \in \mathbb{Z}$ und jeden Eigenwert λ von A. Insbesondere folgt aus der Voraussetzung, dass die Greensche Matrix zu (P) für jedes $\omega > 0$ existiert, dass $\lambda = 0$ kein Eigenwert von A bzw. A nichtsingulär ist. Ist ferner $r \in C_n[0,\omega]$ definiert durch $r(t) := r$ mit einem (komponentenweise) nichtnegativen Vektor $r \in \mathbb{R}^n$, so ist $G(r) = -A^{-1}r$ nach Voraussetzung (komponentenweise) nichtnegativ und folglich $-A^{-1}$ (komponentenweise) nichtnegativ. Zu zeigen bleibt, dass A quasipositiv (bezüglich $\mathbb{R}^n_+$) ist bzw. $a_{ij} \geq 0$ für alle $i, j \in \{1,\dots,n\}$ mit $i \neq j$ gilt. Die Greensche Matrix zu (P) ist gegeben durch

$$G(t,s) = e^{At} \left\{ \begin{array}{l} (I - e^{A\omega})^{-1} \\ (I - e^{A\omega})^{-1}e^{A\omega} \end{array} \right\} e^{-As}, \qquad \begin{array}{l} s < t, \\ t < s. \end{array}$$

Da die Abbildung G (komponentenweise) nichtnegative Elemente aus $C_n[0,\omega]$ in sich abbildet, sind die Einträge von $G(t,s)$ für $(t,s) \in [0,\omega] \times [0,\omega]$ und insbesondere von $e^{At}(I - e^{A\omega})^{-1}$ für $t \in [0,\omega]$ und alle $\omega > 0$ nichtnegativ.

Angenommen, wir wüssten schon, dass $\Re(\lambda) < 0$ für jeden Eigenwert λ von A. Dann ist $\lim_{\omega\to+\infty}(I - e^{A\omega})^{-1} = I$ und daher e^{At} (elementweise) nichtnegativ für alle $t \geq 0$. Aus Lemma 12.21 folgt, dass A quasipositiv bezüglich $\mathbb{R}^n_+$ ist.

Zu zeigen bleibt also, dass $\Re(\lambda) < 0$ für jeden Eigenwert λ von A. Hierzu bemerken wir, dass die Eigenwerte der (elementweise) nichtnegativen Matrix $e^{A\omega}(I - e^{A\omega})^{-1}$ durch $e^{\lambda\omega}(1 - e^{\lambda\omega})^{-1}$ mit einem Eigenwert λ von A gegeben sind. Da der Spektralradius von

$e^{A\omega}(I-e^{A\omega})^{-1}$ ein Eigenwert dieser Matrix ist, existiert ein (reeller) negativer Eigenwert Λ von A mit

$$\rho(e^{A\omega}(I-e^{A\omega})^{-1}) = \frac{e^{\Lambda\omega}}{1-e^{\Lambda\omega}}.$$

Für einen beliebigen Eigenwert λ von A ist daher

$$\frac{e^{\Re(\lambda)\omega}}{1+e^{\Re(\lambda)\omega}} \leq \frac{|e^{\lambda\omega}|}{|1-e^{\lambda\omega}|} \leq \frac{e^{\Lambda\omega}}{1-e^{\Lambda\omega}}.$$

Angenommen, es sei $\Re(\lambda) \geq 0$ für einen Eigenwert λ von A. Für alle $\omega > 0$ wäre dann

$$\frac{1}{2} \leq \frac{e^{\Re(\lambda)\omega}}{1+e^{\Re(\lambda)\omega}} \leq \frac{e^{\Lambda\omega}}{1-e^{\Lambda\omega}}.$$

Mit $\omega \to \infty$ erhalten wir einen Widerspruch. Der Satz ist damit bewiesen. □

Literatur

[1] Aigner, M. und G. M. Ziegler (2002) *Das BUCH der Beweise*. 2. Auflage. Springer-Verlag, Berlin-Heidelberg-New York.

[2] Artin, E. (1942) *Galois Theory. Lectures delivered at the University of Notre Dame*. University of Notre Dame Press, Notre Dame-London. Im Internet unter `http://plouffe.fr/simon/math/Artin%20E.%20Galois%20Theory%20%282ed.,%201944%29%28200dpi%29%28T%29%2886s%29.pdf` einzusehen.

[3] Bauer, H. und K. Neumann (1969) *Berechnung optimaler Steuerungen. Maximumprinzip und dynamische Optimierung*. Lecture Notes in Operations Research and Mathematical Systems 17. Springer-Verlag, Berlin-Heidelberg-New York.

[4] Berman, A. and R. J. Plemmons (1979) *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, New York-San Francisco-London.

[5] Birkhoff, G. (1967) Linear transformations with invariant cones. Amer. Math. Monthly 74, 274–276.

[6] Browder, F. (1983) Fixed point theory and nonlinear problems. Bull. Amer. Math. Soc. 9 (1), 1–39. Im Internet unter `https://projecteuclid.org/download/pdf_1/euclid.bams/1183550974` einzusehen.

[7] Canon, M. D., C. D. Cullum Jr. and E. Polak (1970) *Theory of Optimal Control and Mathematical Programming*. McGraw-Hill Book Company, New York-San Francisco-St. Louis-Toronto-London-Sydney-Mexico-Panama.

[8] Childs, L. (1979) *A Concrete Introduction to Higher Algebra*. Springer-Verlag, New York-Heidelberg-Berlin.

[9] Coddington, E. A. and N. Levinson (1955) *Theory of Ordinary Differential Equation*. Mc Graw-Hill Book Company, New York-Toronto-London.

[10] Collatz, L. (1964) *Funktionalanalysis und numerische Mathematik*. Springer-Verlag, Berlin-Heidelberg-New York.

[11] Deimling, K. (1974) *Nichtlineare Gleichungen und Abbildungsgrade*. Springer-Verlag, Berlin-Heidelberg-New York.

[12] Deimling, K. (1985) *Nonlinear Functional Analysis*. Springer-Verlag, Berlin-Heidelberg-New York.

[13] Edwards, R. E. (1965) *Functional Analysis. Theory and Applications*. Holt, Rinehart and Winston, New York-Chicago-San Francisco-Toronto-London.

[14] Eidelheit, M. (1936) Zur Theorie der konvexen Mengen in linearen normierten Räumen. Studia Mathematica, vol. 6, pp. 104-111.

[15] Gromes, W. (1981) Ein einfacher Beweis des Satzes von Borsuk. Math. Z. 178, 399–400.

[16] Güler, O. (2012) *Foundations of Optimization*. Springer-Verlag, New York-Berlin-Heidelberg.

[17] Heinz, E. (1959) An elementary analytic theory of the degree of a mapping in n-dimensional space. J. Math. Mech. 8, 231–247.

[18] Hadeler, K. P. (1971) Existenz- und Eindeutigkeitssätze für inverse Eigenwertaufgaben mit Hilfe des topologischen Abbildungsgrades. Arch. Rat. Mech. Anal. 42, 317–322.

[19] Higham, N. J. (2008) *Functions of Matrices. Theory and Computations*. SIAM, Philadelphia.

[20] Hoffman, A. J. (1952) On approximate solutions of systems of linear inequalities. Journal of Research of the National Bureau of Standards 49, 263–265. Im Internet unter `http://nvlpubs.nist.gov/nistpubs/jres/049/4/V49.N04.A05.pdf` einzusehen.

[21] Holmes, R. B. (1975) *Geometric Functional Analysis and its Applications.* Springer-Verlag, New York-Heidelberg-Berlin.

[22] Istrătescu, V. I. (1981) *Fixed Point Theory. An Introduction*. D. Reidel Publishing Company, Dordrecht-Boston-London.

[23] Jahn, J. (1994) *Introduction to the Theory of Nonlinear Optimization*. Springer-Verlag, Berlin-Heidelberg-New York.

[24] Jeggle, H. (1979) *Nichtlineare Funktionalanalysis.* B. G. Teubner, Stuttgart.

[25] Kantorowitsch, L. W. und G. P. Akilow (1964) *Funktionalanalysis in normierten Räumen*. Akademie-Verlag, Berlin.

[26] Kirsch, A., W. Warth und J. Werner (1978) *Notwendige Optimalitätsbedingungen und ihre Anwendung*. Lecture Notes in Economics and Mathematical Systems 152. Springer-Verlag, Berlin-Heidelberg-New York.

[27] Krabs, W. (1978) *Einführung in die Kontrolltheorie*. Wissenschaftliche Buchgesellschaft, Darmstadt.

[28] Krasnoselskii, M. A. (1964) *Positive Solutions of Operator Equations*. P. Noordhoff, Groningen.

[29] Krasnoselskii, M. A. (1964) *Topological Methods in the Theory of Nonlinear Integral Equations*. Pergamon Press, Oxford-London-New York-Paris.

[30] Kurzweil, H. (2008) *Endliche Körper. Verstehen, Rechnen, Anwenden*. Zweite, überarbeitete Auflage. Springer-Verlag, Berlin-Heidelberg.

[31] Lee, E. B. and L. Markus (1967) *Foundations of Optimal Control Theory*. John Wiley & Sons, New York-London-Sydney.

[32] Ljusternik, L. A. und W. I. Sobolew (1968) *Elemente der Funktionalanalysis*. Akademie-Verlag, Berlin.

[33] Luenberger, D. G. (1969) *Optimization by Vector Space Methods*. John Wiley, New York-London-Sydney-Toronto.

[34] Nirenberg, L. (1974) *Topics in Nonlinear Functional Analysis*. Courant Institute of Mathematical Science, New York University, New York.

[35] Ortega, J. M. and W. C. Rheinboldt (1970) *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York-London.

[36] Plemmons, R. J. (1977) M-Matrix Characterizations. I-Nonsingular M-Matrices. Linear Algebra and its Applications 18, 175–188.

[37] Rothe, E. (1938) Zur Theorie der topologischen Ordnung und der Vektorfelder in Banachschen Räumen. Compositio Math. 5, 177–197. Im Internet unter `http://archive.numdam.org/ARCHIVE/CM/CM_1938__5_/CM_1938__5__177_0/CM_1938__5__177_0.pdf` einzusehen.

[38] Ružička, M. (2004) *Nichtlineare Funktionalanalysis. Eine Einführung*. Springer-Verlag, Berlin-Heidelberg-New York.

[39] Schauder, J. (1930) Der Fixpunktsatz in Funktionalräumen. Studia Mathematica 2, 171–180. Im Internet unter `http://matwbn.icm.edu.pl/ksiazki/sm/sm2/sm2114.pdf` einzusehen.

[40] Schwartz, J. T. *Nonlinear Functional Analysis*. Gordon and Breach, New York-London-Paris.

[41] Serre, J.-P. (1973) *A Course in Arithmetic*. Springer-Verlag, New York-Heidelberg-Berlin.

[42] Smart, D. R. (1974) *Fixed point theorems*. Cambridge University Press, Cambridge.

[43] Stenger, F. (1975) Computing the topological degree of a mapping in R^n. Numer. Math. 25, 23–38.ß

[44] Taylor, A. E. (1958) *Introduction to Functional Analysis*. John Wiley & Sons, New York-London-Sydney.

[45] Varga, R. S. (1999) *Matrix Iterative Analysis. Second Revised and Expanded Edition.* Springer-Verlag, Berlin-Heidelberg-New York.

[46] Vrahatis, M. N. (1989) A short proof and a generalization of Miranda´s existence theorem. Proceedings of the American Mathematical Society Volume 10 Number 3, 701–703.

[47] Wedderburn, J. H. M. (1905) A theorem on finite algebras. Trans. Amer. Math. Soc. 6, 349–352.

[48] Weil, A. (1974) *Basic Number Theory. Third Edition*. Springer-Verlag, Berlin-Heidelberg-New York.

[49] Werner, H. und H. Arndt (1986) *Gewöhnliche Differentialgleichungen. Eine Einführung in Theorie und Praxis*. Springer-Verlag, Berlin-Heidelberg-New York-London-Paris-Tokyo.

[50] Werner, J. (1969) Einschließungssätze bei nichtlinearen gewöhnlichen Randwertaufgaben und erzwungenen Schwingungen. Numer. Math. 13, 24–38.

[51] Werner, J. (1970) Einschließungssätze für perodische Lösungen der Liénardschen Differentialgleichung. Computing 5, 246–252.

[52] Werner, J. (1972) Nichtnegative Greensche Matrizen bei periodischen Randwertaufgaben. Arch. Rational. Mech. Anal. 46, 96–104.

[53] Werner, J. (1984) *Optimization. Theory and Applications*. Friedr. Vieweg & Sohn, Braunschweig-Wiesbaden.

[54] Werner, J. (1988) *Optimierung*. Kurs der Fernuniversität Hagen. Hagen.

[55] Werner, J. (1992) *Numerische Mathematik. Band 1: Lineare und nichtlineare Gleichungssysteme, Interpolation, numerische Integration*. Friedr. Vieweg & Sohn Verlagsgesellschaft mbH, Braunschweig-Wiesbaden.

[56] Werner, J. (1992) *Numerische Mathematik. Band 2: Eigenwertaufgaben, lineare Optimierungsaufgaben, unrestringierte Optimierungsaufgaben*. Friedr. Vieweg & Sohn Verlagsgesellschaft mbH, Braunschweig-Wiesbaden.

[57] Werner, J. (2001) *Gewöhnliche Differentialgleichungen und ihre numerische Behandlung*. Vorlesung, die im Netz unter `http://num.math.uni-goettingen.de/werner/ode.pdf` einzusehen ist.

[58] Werner, J. (2013) *Merkwürdige Mathematik*. Im Internet unter `http://num.math.uni-goettingen.de/werner/schmankerl.pdf` einzusehen.

[59] Witt, E. (1931) Über die Kommutativität endlicher Schiefkörper. Abh. Math. Sem. Univ. Hamburg 8, 413.

[60] Yoshida (1995) *Functional Analysis. Sixth Edition*. Springer-Verlag, Berlin-Heidelberg-New York.

[61] Zeidler, E. (1986) *Nonlinear Functional Analysis and its Applications I: Fixed Point Theorems*. Springer-Verlag, New York-Berlin-Heidelberg.

[62] Zowe, J. and S. Kurcyusz (1979) Regularity and stability for the mathematical programming problem in Banach spaces. Appl. Math. Optim. 5, 49–62.