# Numerik II

Vorlesung
von

**Eugen Schäfer**

L M U München

W S 2006/07

 München, 12. Februar 2007

# Inhaltsverzeichnis

# Abbildungsverzeichnis

# Teil I

# Anfangswertaufgaben bei gewöhnlichen Differentialgleichungen

# Kapitel 1

# Allgemeine Eigenschaften von Diskretisierungen und Beispiele

<u>Zusammenfassung von Kapitel 1:</u>

– Beispiele expliziter und impliziter Ein– und Mehrschrittverfahren,
– Konvergenz von konsistenten asymptotisch stabilen Verfahren,
– Konsistenzordnung impliziert Konvergenzordnung bei asymptotisch stabilen Verfahren,
– Realisierung in Gleitpunktarithmetik.

Numerische Methoden für Anfangswertprobleme bei gewöhnlichen Differentialgleichungen bestimmen Näherungswerte für Funktionen, die als Lösungen von Anfangswertaufgaben gewöhnlicher Differentialgleichungen definiert sind. Ich betrachte folgende Form von Anfangswertaufgaben:

**(1.1) Definition** *(Anfangswertaufgabe)*
Gegeben $f : I \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$, $I = [t_0, t_0 + T] \subset \mathbb{R}$, und $x_0 \in \mathbb{R}^n$.
Gesucht $x : I \longrightarrow \mathbb{R}^n, x \in C^1(I)$, so daß gilt

$$(A) \quad \left\{ \begin{aligned} x(t_0) &= x_0 \\ x'(t) &= f(t, x(t)) \quad , \quad t \in I\,; \end{aligned} \right.$$

□

Dabei ist

$$x'(s) := \begin{bmatrix} \dfrac{dx_1}{dt}(s) \\ \vdots \\ \dfrac{dx_n}{dt}(s) \end{bmatrix} \quad \text{für} \quad x(s) = \begin{bmatrix} x_1(s) \\ \vdots \\ x_n(s) \end{bmatrix}$$

komponentenweise erklärt. An den Rändern $t_0$ und $t_0 + T$ des Intervalls $I$ ist der rechts- beziehungsweise linksseitige Differentialquotient von x zu bilden:

$$x'(t_0+) = \lim_{s \to t_0, s > t_0} \frac{x(s) - x(t_0)}{s - t_0} \quad , \quad x'(t_0 + T-) = \lim_{s \to t_0 + T, s < t_0 + T} \frac{x(s) - x(t_0 + T)}{s - (t_0 + T)} \quad .$$

Falls für alle Komponenten $x_i \in C^1(I), 1 \leq i \leq n$, gilt, sagen wir, daß $x \in C^1(I)$ ist; analoges gilt für $x \in C^k(I)$.
Aus der Schar von Lösungen der Differentialgleichung $x' = f(t,x)$ charakterisiert man durch die zusätzliche Anfangsbedingung eine – unter geeigneten Voraussetzungen, vgl. (1.3) – eindeutige Lösung. Verlangt man allgemeiner

$$r(x(t_0), x(t_0+T)) = 0, \qquad r : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}^n,$$

so ergibt sich die (Zweipunkt)Randwertaufgabe (in den zwei Punkten $t_0$ und $t_0+T$), die wir im zweiten Teil behandeln.
Beispiele für Anfangswertaufgaben sind etwa Populationsmodelle, z.B.

**(1.2) Fallstudie** (*Verhulst*–Modell: Populationsmodell mit *sozialem Stress* )

$$x'(t) \equiv \begin{bmatrix} {x_1}'(t) \\ {x_2}'(t) \end{bmatrix} = \begin{bmatrix} d_1x_1 - \alpha_1{x_1}^2 - bx_1x_2 \\ -d_2x_2 - \alpha_2{x_2}^2 + cx_1x_2 \end{bmatrix}, \ \alpha_1, \alpha_2, d_1, d_2, b, c > 0.$$

□

Existenz und Eindeutigkeit zumindest lokal, d.h. in einer Umgebung des Anfangswertes $(t_0, x_0)$ sind für Lipschitzstetige rechte Seiten gesichert:

**(1.3) Satz**
Sei $I = [a, a+\delta],\ \delta > 0$, und $f : I \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig, und Lipschitzstetig in $K[x_0; \varrho] = \{y \in \mathbb{R}^n : \|y - x_0\| \leq \varrho\}, \varrho > 0$, d.h.

$$\exists\, L > 0 \quad \forall\, t \in I \quad \forall\, y, z \in K[x_0;\varrho] \quad \|f(t,y) - f(t,z)\| \leq L\|y - z\|.$$

Ferner sei

$$\delta \max_{t \in I, y \in K[x_0;\varrho]} \|f(t,y)\| \leq \varrho.$$

Dann existiert in $I$ genau eine Lösung $x$ der Anfangswertaufgabe

$$x(t_0) = x_0\ , \quad x'(t) = f(t, x(t))\ ,\ t \in I.$$

<u>Beweis:</u>

z.B. Knobloch, Kappel: Gew. DGL., S. 53, Satz 9.1

□

Da die Bedingung

$$\delta \max_{t \in I, y \in K[x_0;\varrho]} \|f(t,y)\| \leq \varrho$$

durch eventuelle Verkleinerung von $\delta$ bzw. $I$ immer erreichbar ist, ist dies eine 'lokale' Existenzaussage.
Zur knappen Formulierung der Voraussetzung 'Lipschitzstetig' in (1.3) folgende

**(1.3') Bezeichnungen**
Sei immer in Kapitel 1 ohne ausdrückliche Präzisierung $I = [t_0, t_0+T] \subset \mathbb{R}$ mit $-\infty < t_0 < t_0 + T < \infty$ . Wenn nicht näher spezifiziert, ist $\|\cdot\|$ irgendeine gegebene Norm im $\mathbb{R}^n$. Für $f : I \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ ist

(i) $f \in Lip \iff f$ ist stetig und erfüllt die folgende Bedingung $(L)$ :

$$(L) \quad \exists\, L > 0 \quad \forall\, t \in I \quad \forall\, y, z \in \mathbb{R}^n : \ \|f(t,y) - f(t,z)\| \leq L\ \|y - z\|$$

(ii) $f \in Lip_{cp} \iff f$ ist stetig und erfüllt die folgende Bedingung $(L_{cp})$ :

$$(L_{cp}) \quad \forall\, c > 0 \quad \exists\, L_c > 0 \quad \forall\, t \in I \quad \forall\, y, z \in K[0;c] : \ \|f(t,y) - f(t,z)\| \leq L_c \|y - z\|$$

(iii) $f \in Lip_{loc}$ (bzgl. der Lösung $x$) $\iff$ $f$ erfüllt die folgende Bedingung $(L_{loc})$:

$$(L_{loc}) \quad \left\{ \begin{array}{l} \text{es existiert } c > 0 \text{ , so daß mit } U(x;c) := \bigcup_{t \in I} \{t\} \times K[x(t);c] \\ f : U(x;c) \to \mathbb{R}^n \text{ stetig ist, und es existiert } L_{loc} > 0 \text{ mit} \\ \|f(t,y) - f(t,z)\| \leq L_{loc}\|y - z\| \ \text{ für } (t,y)\ ,\ (t,z) \in U(x;c)\ ; \end{array} \right.$$

Dabei ist in (ii) und (iii) $K[u;\varrho] := \{v \in \mathbb{R}^n : \|v - u\| \leq \varrho\}$ die abgeschlossene Kugel um $u$ mit Radius $\varrho$.

□

Will ich in der Bezeichnung schon auf die Lipschitzkonstante Bezug nehmen, verwende ich statt

‘ $f \in Lip$ mit Lipschitzkonstante $L$ ‘ die Bezeichnung ‘ $f \in Lip(L)$ ‘

und analog ‘ $f \in Lip_{loc}(L)$ ‘ für ‘ $f \in Lip_{loc}$ mit der lokalen Lipschitzkonstanten $L$‘ .

**Bemerkung**
(i)$\Longrightarrow$(ii)$\Longrightarrow$(iii) (klar)

□

Daß ’$f \in Lip_{cp}$’ nicht die ’*Existenz im ganzen Intervall* $I$’ sichert, zeigt z.B.

**(1.4) Beispiel** (Explosionsgleichung)

$$x' = x^2 \quad , \quad x(-1) = \frac{1}{\alpha + 1} \quad , \quad I = [-1, 1]\ .$$

Dann erfüllt $f(x) = x^2$ die Bedingung $(L_{cp})$. Die Lösung lautet, falls $\alpha \neq -1$ ist,

$$x(t) = \frac{1}{\alpha - t} \quad .$$

Für $\alpha = -1 + \varepsilon$ und $\varepsilon > 0$ klein, ist das rechtsmaximale Existenzintervall der Lösung der Anfangswertaufgabe nur $[-1, -1 + \varepsilon[$ , also *nicht ganz* $I$.

□

Wir werden im weiteren aber immer die *Existenz der Lösung* der AWA (1.1) *im ganzen Intervall* $I$ *voraussetzen.* Daher kommen wir mit den schwächeren Bedingungen $(L_{loc})$ oder $(L_{cp})$ aus.

Die Methoden zur Berechnung einer Lösung gehen bei $AWA'n$ schrittweise von einem Teilintervall zum nächsten vor ähnlich wie man den Beweis des Existenz- und Eindeutigkeitsatzes von *Picard-Lindelöf* führen kann. Daher erhält man bei der numerischen Durchführung gleichzeitig Hinweise über Nichtexistenz bzw. Existenz der Lösung, vgl. die Abbildungen 1.1 und 1.2.

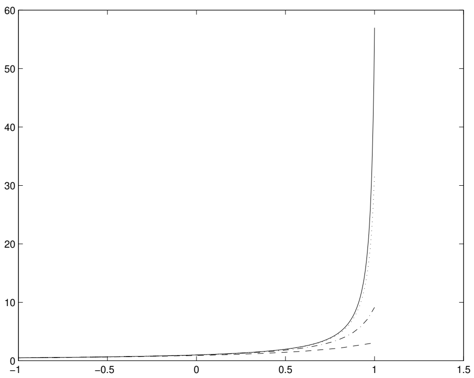


Abbildung 1.1: Verhalten unterschiedlich feiner Diskretisierungen im singulären Fall $\alpha = 0.999$

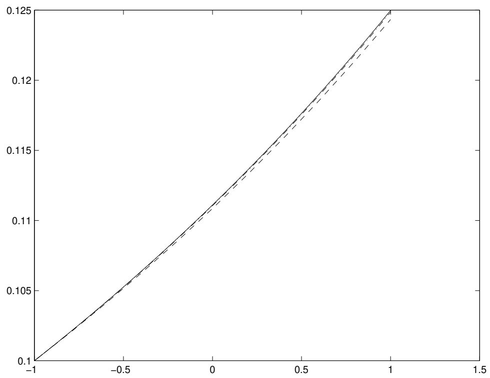


Abbildung 1.2: Verhalten unterschiedlich feiner Diskretisierungen im regulären Fall $\alpha = 9.0$

**Polygonzugverfahren (1.5) für Beispiel (1.4)**
Eine im Intervall $[-1, 1]$ singuläre Lösung erhält man etwa für $\alpha = 0.999$, eine dort reguläre Lösung etwa für $\alpha = 9$.

□

Formal äquivalent zur Lösung von (1.1) ist die Lösung der Integralgleichung

$$x(t) = x_0 + \int_{t_0}^{t} f(s, x(s))\, ds \ , \ t \in I \ .$$

Dabei ist für $f : I \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ das Integral komponentenweise erklärt. Dieser Zusammenhang ist Grundlage und Motivation für einige Verfahren.
Zur näherungsweisen Lösung der Anfangswertaufgabe in $I = [t_0, t_0 + T]$ werden Näherungen für die Lösung in endlich vielen Punkten $I_h$ von $I$ bestimmt

$I_h := \{a = t_0 < t_1 < ... < t_N = b\} \ \subset \ I = [t_0, t_0 + T]$.

Einige Sprechweisen im Zusammenhang mit Gittern $I_h$:

**Bezeichnungen**
(i) Die *Schrittweite* $h = T/N > 0$ , $N \in \mathbb{N}$, liefert das *'äquidistante'* (gleichabständige) Gitter $I_h = \{t_j \in I : t_j = t_0 + jh \ , j = 0, ..., N\}$ bzw. $I_h^{'} := \{t_j \in I_h \ , j = 0, ..., N-1\}$ .

Nicht äquidistante Gitter erhält man durch einen Schrittweitenvektor :

(ii) $h = (h_0, h_1, \ldots, h_{N-1})^t \in \mathbb{R}^N$ mit einem $N \in \mathbb{N}$ heißt Schrittweitenvektor zum Intervall $[t_0, t_0 + T]$, wenn $h_i > 0 \ i = 0, \ldots, N-1$ ; $\sum_{i=0}^{N-1} h_i = T$ . Das zu $h$ gehörige Gitter ist $I_h = \{t_j \in I : t_j = t_0 + \sum_{i=0}^{j-1} h_i \ , j = 0, ..., N\}$ und das halboffene Gitter $I_h^{'} := \{t_j \in I_h \ , j = 0, ..., N-1\}$ .

Umgekehrt bestimmt ein gegebenes Gitter $I_h = \{t_0 < \ldots t_{j-1} < t_j < \ldots t_N\}$ den zugehörigen Schrittweitenvektor $h$ durch $h_j := t_{j+1} - t_j, 0 \le j < N$ .

(iii) $|I_h| := |h| := \max\limits_{0 \le i < N} h_i$ ist die Gitterbreite von $h = (h_0, h_1, \ldots, h_{N-1})^t \in \mathbb{R}^N$ bzw. die Gitterbreite von $I_h$.

(iv) Für eine Folge $(h^{(k)} \in \mathbb{R}^{N_k} \mid k \in \mathbb{N})$ von Schrittweitenvektoren (mit $N_k \to \infty$) ist $\lim_{k\to\infty} h^{(k)} = 0$ definiert durch $\lim_{k\to\infty} |h^{(k)}| = 0$ .

□

Dazu ersetzt man bei *Einschrittverfahren* die Differentialgleichung

$$x'(t) - f(t, x(t)) = 0 \quad , \quad t_j \le t < t_j + h_j$$

auf dem Teilintervall $[t_j, t_j + h_j[$ durch

(a) $x'(t) \doteq [x(t_j + h_j) - x(t_j)]/h_j$;

(b) $f(t, x(t)) \doteq f_h(t_j, x(t_j), t_{j+1}, x(t_{j+1})) \quad , \quad t_j < t < t_{j+1}$

Um in der Fehleranalyse auch die Abhängigkeit der numerischen Lösung von nicht exakten Anfangsdaten – etwa durch Rundung, oder von Anfangsdaten aus ungenauen Beobachtungen oder Messungen – mit zu erfassen, lasse ich mögliche Ungenauigkeiten im Anfangswert zu

(c) $x_0 \doteq x_{0,h}$

Die naheliegendste Wahl in

$$(b) \quad f_h(t_0, x_0, t_1, x_1) := f(t_0, x_0)$$

und in

$$(c) \quad x_{0,h} := x_0 \quad \text{bzw. } x_{0,h} = rd(x_0)$$

stammt von *L. Euler* (*1707 in Basel) – († 1783 in St. Petersburg) und wird als *Eulersches Polygonzugverfahren* bezeichnet. Zur Veranschaulichung vergleiche man die Abb. 1.3 : Näherung auf $I_h$ sind die ∘-Werte.

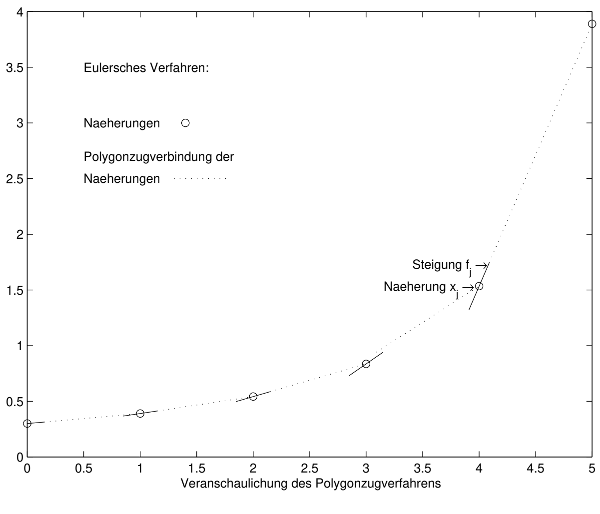


Abbildung 1.3: Veranschaulichung des Polygonzugverfahrens für $n = 1$.

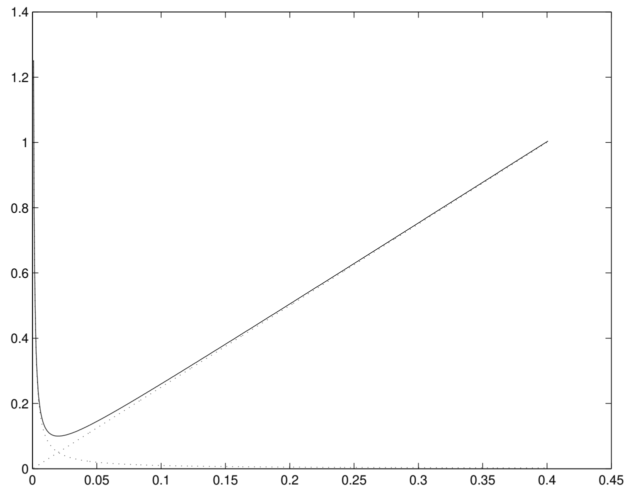

Abbildung 1.4: Verfahrens- versus Rundefehler, $p = 1$.

Eine lineare Verbindung dieser Werte ergibt als Näherung auf ganz I einen *Polygonzug* .

**(1.5) Polygonzugverfahren** (von *Euler*)
Mit der Bezeichnung $x_i \equiv x_h(t_i)$ gilt dabei

$$(A_h) \quad x_{j+1} := x_j + h_j f(t_j, x_j) \ , \ j = 0, 1, \ldots \ .$$

□

Bei Kenntnis von $x_j$ kann man beim Polygonzugverfahren (1.5) $x_{j+1}$ berechnen durch *explizite* Auswertung von $f$ .

Im Gegensatz dazu muß beim folgenden *impliziten* Verfahren ein (im allgemeinen Fall eines beliebigen $f$) nichtlineares Gleichungssystem ($n$ Gleichungen in den $n$ Unbekannten $x_{j+1} \in \mathbb{R}^n$) gelöst werden beim Übergang von $t_j$ zu $t_{j+1}$.

**(1.6) Rückwärtsgenommenes Polygonzugverfahren**
Mit $x_i \equiv x_h(t_i)$ ist

$$(A_h) \quad x_{j+1} := x_j + h_j f(t_{j+1}, x_{j+1}) \ , \ j = 0, 1, \ldots \ ;$$

ein implizites Verfahren, denn $x_{j+1}$ ist nur implizit als Lösung obiger Gleichung definiert.

□

Ich möchte noch eine weitere Verfahrensklasse mit einem Beispiel motivieren. Kennt man z.B. außer dem Anfangswert $x(t_0)$ auch noch den exakten Wert von $x(t_1)$ – oder eine genaue Näherung dafür – , so kann man $x(t_2)$ und alle weiteren Werte hoffentlich genauer bestimmen.

**(1.7) Mittelpunktformel**
Sei $x_0$ und $x_1$ gegeben. Dann erhält man durch

$$x_{j+2} = x_j + 2h f_{j+1} \ , \ j = 0, 1, \ldots \ ;$$

die Näherungswerte der Mittelpunktformel.
Dabei ist $x_i \equiv x_h(t_i)$ und $f_i \equiv f(t_i, x_h(t_i))$ , $i = 0, 1, \ldots$ .

□

Dies ist ein explizites 2-Schrittverfahren: *2–Schritt*–Verfahren, da die Näherungswerte in 2 Intervallen verknüpft werden; *explizit*, da zur Bestimmung des neuen Wertes $x_{i+2}$ nur (explizite) Funktionsauswertungen notwendig sind.

Allgemein formuliert betrachten wir bei Einschrittverfahren für die Anfangswertaufgabe (1.1) – im folgenden immer $(A)$ abgekürzt –

$$(A) \quad \begin{cases} x(t_0) = x_0 \\ x'(t) = f(t, x(t)) \quad , \quad t \in I := [t_0, t_0 + T] \ ; \end{cases}$$

ein Ersatzproblem folgender Bauart:

**(1.8) Einschrittverfahren** (kurz ($ESV$))
Zum Schrittweitenvektor $h = (h_0, \ldots, h_{N-1})^t$ mit zugehörigem Gitter

$$I_h = \{t_j = t_0 + \sum_{i<j} h_i \ : \ 0 \leq j \leq N\}$$

seien

$$f_h : I_h \times \mathbb{R}^n \times I_h \times \mathbb{R}^n \longrightarrow \mathbb{R}^n,$$
$$I_h \times \mathbb{R}^n \times I_h \times \mathbb{R}^n \ni (t_0, x_0, t_1, x_1) \longmapsto f_h\big(t_0, x_0, t_1, x_1\big) \in \mathbb{R}^n$$

und $x_{0,h} \in \mathbb{R}^n$ gegeben. Gesucht ist $x_h : I_h \longrightarrow \mathbb{R}^n$ mit

$$(A_h) \quad \begin{cases} x_h(t_0) = x_{0,h} \\ \big(x_h(t_{j+1}) - x_h(t_j)\big)/h_j = f_h(t_j, x_h(t_j), t_{j+1}, x_h(t_{j+1})) \quad , \quad 0 \leq j < N \end{cases}$$

□

Ein Einschrittverfahren ist genauer eine *Konstruktionsvorschrift* $\Phi$, wie man sich aus jeder gerade betrachteten rechten Seite $f$ bei gegebenem $h$ die Verfahrensfunktion $f_h$ bestimmen soll, d.h.

$$f_h(\cdot,\cdot,\cdot,\cdot) = \Phi(\cdot,\cdot,\cdot,\cdot\,; h, f) \ , \ f \in \mathcal{F} \ ;$$

mit einem Schrittweitenvektor $h$, und $f$ aus einer geeigneten 'Klasse' $\mathcal{F}$ von rechten Seiten. Diese funktionale Abhängigkeit des $f_h$ von $f$ muß man immer vor Augen haben, obwohl ich sie in der Notation im allgemeinen weglasse. Häufig bezeichne ich auch das Einschrittverfahren mit $(A_h)$ - ebenso wie die definierenden Gleichungen.
Für das vorwärtsgenommene bzw. das rückwärtsgenommene Polygonzugverfahren gilt

$$f_h\big(t_0, x_0, t_1, x_1\big) := f(t_0, x_0) \quad \text{bzw.} \quad f_h\big(t_0, x_0, t_1, x_1\big) := f(t_1, x_1) \ .$$

Die verschiedenen Einschrittverfahren unterteilt man nach dem Aussehen der Konstruktionsvorschrift $\Phi$ bzw. der resultierenden Abbildung $f_h$ mit $(s, y, t, z) \longmapsto f_h(s, y, t, z)$ :

| EINSCHRITTVERFAHREN | | |
|---|---|---|
| | explizit | implizit |
| Erklärung für die Namensgebung | die funktionale Abhängigkeit $f_h$ besteht in expliziten Funktionsauswertungen | die Funktion $f_h$ ist implizit durch Gleichungen gegeben |
| Durchführbarkeit | klar | muß analysiert werden: ja, falls $\lvert h\rvert$ hinr. klein |
| Aufwand | nicht sehr hoch | groß, da in jedem Schritt nichtlineares Gleichungs– system zu lösen ist |
| Genauigkeit | gut bei genügend Aufwand | gut bei genügend Aufwand |
| Stabilität | nicht so gut | sehr gut |
| Schrittweiten– Steuerung | einfach | |

*Einschrittverfahren* heißen diese Verfahren deshalb, weil $x_h(t_{j+1})$ berechnet werden kann bei Kenntnis der Näherungslösung $x_h(t_j)$ *einen* Schritt vorher.
Im Gegensatz dazu wird bei *Mehrschrittverfahren* $x_h(t_{j+1})$ bestimmt durch die Näherungen in *mehreren* vorangehenden Gitterpunkten. Eine stichwortartige Bewertung von Mehrschrittverfahren enthält die folgende Tabelle

| MEHRSCHRITTVERFAHREN | | |
|---|---|---|
| | explizit | implizit |
| Namensgebung | wie bei *ESV* | wie bei *ESV* |
| Durchführbarkeit | klar | muß analysiert werden: ja, falls $\lvert h\rvert$ hinr. klein |
| Aufwand | gering | nicht sehr hoch |
| Genauigkeit | sehr gut | sehr gut |
| Stabilität | sehr schlecht | schlecht – nicht so gut |
| Schrittweiten–Steuerung | aufwändig | aufwändig |

Im Rest dieses Abschnitts betrachte ich Ein– und Mehrschrittverfahren gemeinsam. Deshalb sei $(A_h)$ von der Form

**(1.9) Bauart von $(A_h)$**
Sei $m \in \mathbb{N}$. Mit $x_i \equiv x_h(t_i)$ seien $x_0, \ldots, x_{m-1}$ (geeignet) gegeben. Dann werde $x_{j+m}$ für $j = 0, ...., N-m$ bestimmt durch eine Vorschrift der Form

$$(A_h) \qquad \frac{1}{h}\sum_{k=0}^{m} a_k x_{j+k} = f_h\big(t_j, x_j, \ldots, t_{j+m}, x_{j+m}\big) \ , \ j \ = \ 0, 1, ...., N-m \ .$$

□

Dabei ist

$$(\widetilde{D}_h x_h)(t_j) :\equiv \frac{1}{h}\sum_{k=0}^{m} a_k x_h(t_{j+k}) \doteq x'(t_j)$$

und

$$f_h\big(t_j, x_h(t_j), \ldots, t_{j+m}, x_h(t_{j+m})\big) \doteq f\big(t_j, x(t_j)\big).$$

Einschrittverfahren *ESV* erhält man mit der Spezialisierung

$$m = 1 \ , \ a_1 = 1 \ , \ a_0 = -1 \qquad \text{d.h.} \qquad (\widetilde{D}_h x_h)(t_j) \ \equiv \ (D_h x_h)(t_j) \ = \ \big(x_h(t + h(t)) - x_h(t)\big)/h(t)\big|_{t \, = \, t_j} \ .$$

In Analogie zu *ESV* erweitere ich die Definition für $I_h^{'}$ durch $I_h^{'} := \{tj \in I_h : t_{j+m} \in I_h\}$ .

Zur Bewertung der Güte berechneter Näherungen folgendes Maßfür $z_h : I_h \longrightarrow \mathbb{R}^n$

$\|z_h\|_{\infty, I_h} := \max_{t \in I_h} \|z_h(t)\|$ und analog $\|z_h\|_{\infty, I_h^{'}} := \max_{t \in I_h^{'}} \|z_h(t)\|$ .

Minimalziel sind konvergente Verfahren. Dazu

**(1.10) Konvergenz und −ordnung**
Das Verfahren $(A_h)$ der Form (1.9)

(a) heißt *konvergent*, wenn $\lim_{h\to 0} \|x_h - x\|_{\infty,I_h} = 0$ .

(b) hat (mindestens) die *Konvergenzordnung* $p \in \mathbb{N}$, wenn

$$\|x_h - x\|_{\infty,I_h} = O(|h|^p)\ ,\ \ h \to 0\ .$$

Dabei ist $x$ Lösung der Anfangswertaufgabe $(A)$ und $x_h$ die Lösung von $(A_h)$ auf $I_h$. □

Damit $x_h$ eine akzeptable Näherung für $x$ sein kann, muß $(A_h)$ eine passende Näherung für $(A)$ sein. Dieses 'passend sein' nennt man

**(1.11) Konsistenz und −ordnung**
Ein Verfahren $(A_h)$ der Form (1.9)
(a) heißt *konsistent*, wenn

(*i*) $\lim\limits_{h\to 0} \max\limits_{0\le i\le m-1} \|x_h(t_i) - x(t_i)\| = 0$ , , sog. *Konsistenz bzgl. der 'Startwerte'* ;

(*ii*) $\lim\limits_{h\to 0} \max\limits_{t_j\in I_h'} \|\tau_h(t_j)\| = 0$ , wobei $\tau_h(t_j) = \dfrac{1}{h}\sum\limits_{k=0}^{m} a_k x(t_{j+k}) - f_h\big(t_j, x(t_j), \ldots, t_{j+m}, x(t_{j+m})\big)$.

(b) hat (mindestens) die *Konsistenzordnung* $p \in \mathbb{N}$ , wenn

(*i*) $\max\limits_{0\le i\le m-1} \|x_i - x(t_i)\| = O(|h|^p)\ ,\ h \to 0$ ;

(*ii*) $\max\limits_{t_j\in I_h'} \|\tau_h(t_j)\| = O(|h|^p)\ ,\ h \to 0$ .

Dabei ist $x$ die Lösung der Anfangswertaufgabe $(A)$. □

$\tau_h$, der sog. *lokale Diskretisierungsfehler*, ist also der Defekt bei Einsetzen der exakten Lösung der Differentialgleichung in die Differenzengleichung.

**(1.12) Bemerkung** (Konsistenz von *ESV* )
Bei *ESV* ist notwendig und hinreichend für die Konsistenz von $(A_h)$ mit $(A)$, daß gilt

$$\lim_{h\to 0} x_h(t_0) = x(t_0) \quad ; \quad \lim_{h\to 0} \max_{t\in I_h'} \|f_h\big(t, x(t), t+h, x(t+h)\big) - f(t, x(t))\| = 0\ .$$

Dabei ist $x$ die Lösung von $(A)$.

<u>Beweis:</u>

Für $x \in C^1(I)$ gilt immer

$$\begin{aligned} \|(D_h x)(t) - x'(t)\| &\le \frac{1}{h}\int_0^h \|x'(t+\sigma) - x'(t)\| d\sigma \\ &\le \max_{t,t+\delta\in I: 0\le\delta\le|h|} \|x'(t+\delta) - x'(t)\| \overset{h\to 0}{\longrightarrow} 0\ , \end{aligned}$$

wobei diese Konvergenz aus der gleichmäßigen Stetigkeit von $x'$ auf $I$ folgt. Wegen

$$\tau_h(t) = [(D_h x)(t) \underbrace{- x'(t)] + [f(t, x(t))}_{=0} - f_h\big(t, x(t), t+h, x(t+h)\big)]$$

impliziert jede der Bedingungen in (1.12) bzw. (1.11) die andere. □

Wegen $f_h \equiv f$ in der Polygonzugmethode folgt daraus, daß die Polygonzugmethode konsistent ist. Die Konsistenzordnung $p = 1$ der Polygonzugmethode zeige ich im nächsten

**(1.13) Satz**
Das Polygonzugverfahren hat für $f \in C^1(I \times \mathbb{R}^n)$ die Konsistenzordnung $p = 1$. Es gilt mit der Lösung $x$ von $(A)$ für jeden Schrittweitenvektor $h$

$$\|\tau_h\|_{\infty,I_h'} \leq \Big(\frac{1}{2}\|x''\|_{\infty,I}\Big)\ |h|\ .$$

<u>Beweis:</u>

Man beachte

$$x \text{ Lösung der } AWA \implies x \in C^1(I) \implies x'(\cdot) = f(\cdot, x(\cdot)) \in C^1 \text{ n. Vor. über } f \implies x \in C^2(I)\ .$$

Zu $t_j \in I_h'$ ist

$$\begin{aligned}
\tau_h(t_j) &\overset{\text{def}}{=} \frac{1}{h_j}\Big(x(t_j + h_j) - x(t_j)\Big) - f_h(t_j, x(t_j), t_{j+1}, x(t_{j+1})) \\
&\overset{f_h = f}{=} \frac{1}{h_j}\Big(\Big[\ x(t_j) + h_j x'(t_j) + \int_{t_j}^{t_j+h_j} (t_j + h_j - s)x''(s)ds\ \Big] - x(t_j)\Big) - \ f(t_j, x(t_j)) \\
&= \frac{1}{h_j}\int_{t_j}^{t_j+h_j} (t_j + h_j - s)x''(s)ds\ , \quad \text{da } x \text{ Lösung der } AWA\ .
\end{aligned}$$

$\implies$

$$\|\tau_h(t_j)\| \leq \max_{s \in [t_j, t_j+h_j]} \|x''(s)\| \underbrace{\frac{1}{h_j}\int_{t_j}^{t_j+h_j} (t_j + h_j - s)ds}_{= h_j/2}$$

$\implies$

$$\underbrace{\max_{t \in I_h'} \|\tau_h(t)\|}_{=: \|\tau_h\|_{\infty,I_h'}} \leq \underbrace{\max_{s \in I}\left(\frac{1}{2}\|x''(s)\|\right)}_{=: 1/2\|x''\|_{\infty,I}}\ |h|\ .$$

□

Die Konvergenzeigenschaft folgt aus der Konsistenz, falls die Lösung $x_h$ von

$$(A_h) \qquad \frac{1}{h}\sum_{k=0}^{m} a_k x_h(t_{j+k}) = f_h(t_j, x_h(t_j), \ldots, t_{j+m}, x_h(t_{j+m}))\ ,\ j = 0, ..., N - m\ ;$$

stetig abhängt von den Anfangswerten $x_h(t_0), \ldots, x_h(t_{m-1})$ und den rechten Seiten $f_h$. Für die hier betrachteten Verfahren ist dies sogar eine Lipschitzstetige Abhängigkeit. Daher beschränke ich mich bereits in der Definition der 'asymptotischen Stabilität' auf 'Lipschitzstetig'.

**(1.14) Asymptotische Stabilität**
Das Verfahren $(A_h \mid h \in \mathcal{H})$ der Form (1.9) heißt *asymptotisch stabil* , wenn gilt:

$\exists\, H_0 > 0\ \exists\, \varrho_0 > 0\ \exists\, \delta_0 > 0\ \exists\, C > 0\ \forall\, h \in \mathcal{H} : |h| < H_0\ \forall\, x_h : I_h \longrightarrow \mathbb{R}^n :$
$\big(\max_{0 \leq i < m} \|x_h(t_i) - x(t_i)\| \leq \varrho_0 \wedge (\widetilde{D}_h x_h)(t_j) = f_h\big(t_j, x_h(t_j), ., t_{j+m}, x_h(t_{j+m})\big)\ ,\ t_j \in I_h'\big)$

$\forall\, z_h : I_h \longrightarrow \mathbb{R}^n$ :
$\big( \max\limits_{0\le i<m} \|z_h(t_i) - x_h(t_i)\| \le \delta_0 \wedge \max\limits_{t_j \in I_h'} \|\widetilde{D}_h z_h)(t_j) - f_h\big(t_j, z_h(t_j), ., t_{j+m}, z_h(t_{j+m})\big)\| \le \delta_0 \big)$

$$\|z_h - x_h\|_{\infty, I_h} \le C\Big( \max_{0\le i<m} \|z_h(t_i) - x_h(t_i)\| + \max_{t_j \in I_h'} \|(\widetilde{D}_h z_h)(t_j) - f_h\big(t_j, z_h(t_j), ., t_{j+m}, z_h(t_{j+m})\big)\|\Big).$$

Dabei ist $x$ die Lösung von $(A)$.

□

Dies ist gerade die 'Lipschitzstetige Lösbarkeit', d.h. die Lipschitzstetigkeit von

$$\Big( \widetilde{D}_h(\cdot, x_h) - f_h(\cdot, x_h) \Big)^{-1}$$

in der Umgebung von 0, gleichmäßig in $h : |h| \in ]0, H_0]$.
Bei Einschrittverfahren ist diese Stabilitätseigenschaft genauso 'nahezu automatisch' erfüllt wie die Konsistenz nach (1.12). Sei $\mathcal{H}$ eine (verallgemeinerte) Folge von Schrittweitenvektoren mit $h \to 0$ (ihr konkretes Aussehen kann je nach Situation variieren). Dann erhält man die asymptotische Stabilität aus der Lipschitzstetigkeit, gleichmäßig in $h \in \mathcal{H}$, der Verfahrensfunktionen $f_h$ .
Bei Mehrschrittverfahren gilt dies nicht 'automatisch', auch wenn $f_h$ und $f$ Lipschitzstetig sind (s. Übungen).

**(1.15) Satz**
Das Einschrittverfahren $(A_h \mid h \in \mathcal{H})$ sei konsistent mit $(A)$. Die Schar der Verfahrensfunktionen $\big((t, x_0, t+h(t), x_1) \longmapsto f_h(t, x_0, t+h(t), x_1) \mid h \in \mathcal{H}\big)$ sei in der 2. Variablen $x_0 \in U(x;c)$ und 4. Variablen $x_1 \in U(x;c)$ Lipschitzstetig mit in $h \in \mathcal{H}$ und $t \in I_h$ gleichmäßiger Lipschitzkonstante $\widetilde{L}$, wobei $x$ die Lösung von $(A)$ ist. Dann ist das Einschrittverfahren asymptotisch stabil für quasi-uniforme Gitterfolgen, d.h. es gilt $\frac{\max_j h_j}{\min_j h_j} \le c$ für alle $h = (h_j) \in \mathcal{H}$ .

Beweis:

*Wachstumslemma*
Sei $h_j > 0$ , $\delta_j \ge 0$ , $\varepsilon_j \ge 0$ , $j = 0, 1, \dots$ , und mit $M > 0$ gelte

$$\delta_{j+1} \le \delta_0 + M \sum_{i \le j} h_i \delta_i + \varepsilon_j \ , \ j = 0, 1, \dots \ .$$

Dann gilt mit $|h| = \max_{i<j} h_i$

$$\delta_j \ \le \ (1 + |h|M)^j \big(\delta_0 + \max_{i<j} \varepsilon_i\big) \ \le \ e^{Mj|h|}\big(\delta_0 + \max_{i<j} \varepsilon_i\big) \ .$$

*Beweis des Lemma:*
Für $j = 1$ gilt die Abschätzung nach Voraussetzung. Damit folgt nach Induktionsannahme

$$\delta_{j+1} \le \delta_0 + M \sum_{i \le j} h_i (1 + |h|M)^i \big(\delta_0 + \max_{k<i} \varepsilon_k\big) + \varepsilon_j$$

$$\le \delta_0 + M|h|[(1 + |h|M)^{j+1} - 1)/(1 + |h|M - 1)]\big(\delta_0 \ + \ \max_{k \le j} \varepsilon_k\big) + \varepsilon_j$$

$$\le (1 + |h|M)^{j+1}\big(\delta_0 \ + \ \max_{k \le j} \varepsilon_k\big) \ .$$

◇

Sei $d_j := (z_h - x_h)(t_j)$ und $\delta_j := \|(z_h - x_h)(t_j)\|$ , wobei $x_h$ und $z_h$ die Voraussetzungen von (1.14) erfüllen. Subtrahiert man die Gleichungen

$$z_h(t_i + h_i) = z_h(t_i) + h_i f_h(t_i, z_h(t_i), t_{i+1}, z_h(t_{i+1})) + h_i \sigma_h(t_i)$$

$$x_h(t_i + h_i) = x_h(t_i) + h_i f_h(t_i, y(t_i), t_{i+1}, y(t_{i+1})) \ ,$$

und summiert über $i = 0...j$, erhält man für $f_h$ mit Lipschitzkonstante $\widetilde{L}$

$$\delta_{j+1} \le \delta_0 + \widetilde{L} \sum_{i=0}^{j} h_i(\delta_i + \delta_{i+1}) + \sum_{i=0}^{j} h_i \|\sigma_h(t_i)\| \ .$$

Für $|h|\widetilde{L} \le 1/2$ folgt

$$\delta_{j+1} \le 2\delta_0 + 4\widetilde{L} \sum_{i=0}^{j} h_i \delta_i + 2 \sum_{i=0}^{j} h_i \|\sigma_h(t_i)\| \ .$$

Nach obigem Wachstumslemma folgt wegen $j|h| \le c(t_j - t_0)$

$$\begin{aligned} \|(z_h - x_h)(t_j)\| \ &\le \ e^{4\widetilde{L}c(t_j - t_0)} 2\Big( \|(z_h - x_h)(t_0)\| + \sum_{i<j} h_i \|\sigma_h(t_i)\| \Big) \ . \\ &\le \ 2e^{4c\widetilde{L}(t_j - t_0)} \Big( \|(z_h - x_h)(t_0)\| + (t_j - t_0) \max_{i<j} \|\sigma_h(t_i)\| \Big) \ , \end{aligned}$$

d.h. für $[t_0, t_N] = [t_0, t_0 + T]$ gilt

$$\|z_h - x_h\|_{\infty, I_h} \ \le \ 2e^{4c\widetilde{L}T} \big( \|z_h(t_0) - x_h(t_0)\| + T\|\sigma_h\|_{\infty, I'_h} \big) \ .$$

und damit die asymptotische Stabilität unter der stärkeren Voraussetzung

$$' \ f_h \in Lip(\widetilde{L}) \ \text{ für alle } \ h \in \mathcal{H} \ ' \ .$$

Die schwächere Voraussetzung

$$' \ f_h \in Lip_{loc}(\widetilde{L}) \ \text{ für alle } \ h \in \mathcal{H} \ '$$

erfordert eine Zusatzüberlegung: Mit dem Wachstumslemma zeigt man zuerst, daß auf Grund der Konsistenz

$$\|x_h - x\|_{\infty, I_h} \le c \ , \ \|z_h - x\|_{\infty, I_h} \le c \ ,$$

gilt für hinreichend kleine Störungen in den Daten, also $(\widetilde{L}_{loc})$ anwendbar ist. Damit kann als gemeinsame Lipschitzkonstante $\widetilde{L}$ aus der Bedingung $(\widetilde{L}_{loc})$ genommen werden für hinreichend kleines $|h|$. Die asymptotische Stabilität folgt daraus wieder wie im obigen Fall nach dem Wachstumslemma. □

Hat man asymptotisch stabile Verfahren, so lautet der Konvergenzsatz für Ein- und Mehrschrittverfahren

**(1.16) Satz**
Das Verfahren $(A_h \mid h \in \mathcal{H})$ der Form (1.9) sei konsistent mit $(A)$, und asymptotisch stabil. Dann gilt die Konvergenzaussage $\lim_{h\to 0} \|x_h - x\|_{\infty, I_h} = 0$ .
Hat das Verfahren die Konsistenzordnung $p \in \mathbb{N}$, so ist die Konvergenzordnung $p$ .

Beweis:

Wegen 'konsistent' gilt

$$(*)\quad \begin{cases} \forall\, \delta > 0 \;\; \exists\, H(\delta) > 0 \;\; \forall\, h : |h| \le H(\delta) \\ \max\limits_{t_i \in I_h\, :\, i<m} \|x_h(t_j) - x(t_j)\| + \max\limits_{t_j \in I_h'} \|\tau_h(t_j)\| < \delta \end{cases}$$

Wegen 'asymptotisch stabil' gilt

$$(**)\quad \begin{cases} \exists\, H_0 > 0 \;\; \exists\, \varrho_0 > 0 \;\; \exists\, \delta_0 > 0 \;\; \exists\, C > 0 \;\; \forall\, h \in \mathcal{H} \;: \\ \big(\, |h| \le H_0 \wedge \max\limits_{t_i \in I_h\, :\, i<m} \|x_h(t_j) - x(t_j)\| \le \varrho_0 \wedge \max\limits_{t_j \in I_h'} \|\tau_h(t_j)\| \le \delta_0 \big) \\ \left.\begin{array}{r} (\widetilde{D}_h x_h)(t_j) = f_h(t_j, x_h(t_j), ., t_{j+m}, x_h(t_{j+m}))\,,\;\; t_j \in I_h' \\ (\widetilde{D}_h x)(t_j) = f_h(t_j, x(t_j), ., t_{j+m}, x(t_{j+m})) + \tau_h(t_j)\,,\;\; t_j \in I_h' \end{array}\right\} \Longrightarrow \begin{array}{l} \|x - x_h\|_{\infty, I_h} \le \\ C\,\Big( \max\limits_{t_i \in I_h\, :\, i<m} \|x_h(t_j) - x(t_j)\| \\ \quad + \max\limits_{t_j \in I_h'} \|\tau_h(t_j)\| \;\Big) \end{array} \end{cases}$$

Wähle zu $\delta := \min(\varrho_1, \delta_1, \varepsilon / C)$ ein $H(\delta)$ nach $(*)$ und setze

$$\widetilde{H}(\varepsilon) := \min(H_1, H(\delta)) \;.$$

Dann ist für $|h| \le \widetilde{H}(\varepsilon)$ nach $(**)$ $\|x - x_h\|_{\infty, I_h} \le \varepsilon$ .
Die Ordnungsaussage folgt nach $(**)$.

□

Satz (1.16) verlangt geradezu nach Verfahren höherer Konsistenzordnung. *Eine* Konstruktionsmöglichkeit sind *Runge–Kutta–* Verfahren . *Runge* war von 1904-1924 Ordinarius in Göttingen, *Kutta* war von 1911-1935 Ordinarius an der TH Stuttgart.

Mit den bisherigen Überlegungen haben wir prinzipiell das theoretische Verhalten der Näherungen geklärt für $h \to 0$. Wie steht es mit der Realisierung am Rechner?

*Einschrittverfahren in Gleitpunktarithmetik*

Führt man ein Einschrittverfahren in Gleitpunktarithmetik durch, so berechnet man Näherungen $\widetilde{x}_h(t_j)$ an Stelle von $x_h(t_j)$ mit

$$\widetilde{x}_h(t_{j+1}) = \widetilde{x}_h(t_j) + h_j f_h\big(t_j, \widetilde{x}_h(t_j)\big) + \varepsilon_j \;,\; j = 0, ..., N(h) \;,$$

wobei bestenfalls $|\varepsilon_j| \le macheps$ , $j = 0, ..., N(h) - 1$ , gilt. Für die berechneten Näherungen gilt dann (vgl. Übungen) für ein *ESV* der Konvergenzordnung $p$

$$\|\widetilde{x}_h - x\|_{\infty, I_h} \le c_1 |h|^p + c_2 \frac{macheps}{|h|} \;,$$

falls $f_h \in Lip(\widetilde{L})$ und $\frac{\max_j h_j}{\min_j h_j} \le c$ für alle $h = (h_j) \in \mathcal{H}$, d.h. $\mathcal{H}$ quasi-uniform.

Da obige Abschätzung im ungünstigsten Fall eine Identität werden kann, ist nur sinnvoll – vgl. Abb. 1.4 –

$$|h| \geq const\ macheps^{1/(p+1)} ,$$

für das Polygonzugverfahren also

$$|h| \geq const\ \sqrt{macheps} .$$

□

**Bemerkung**
Für Runge–Kutta–Verfahren der Ordnung $p = 4$ bzw. $p = 5$ ergibt obige Schranke bei $macheps = 2^{-52} \doteq 2.2 \cdot 10^{-16}$ (und $const = 1$)

$$h_{min} \doteq 0.0007 \quad \text{bzw. } h_{min} \doteq 0.0025.$$

□

## Literatur zu Anfangswertaufgaben bei gewöhnlichen Differentialgleichungen

*Theorie:*

B. Aulbach [2004]: Gewöhnliche Differentialgleichungen. 2. Auflage, Spektrum Akademischer Verlag 2004, ISBN 3-8274-0204-2.

M. Braun [1991]: Differentialgleichungen und ihre Anwendungen. 2. Auflage Springer-Verlag Berlin u.a.

H.W. Knobloch & F. Kappel [1974]: Gewöhnliche Differentialgleichungen Teubner-Verlag, Stuttgart.

W. Walter [1972]: Gewöhnliche Differentialgleichungen. Eine Einführung. Springer-Verlag, Berlin - Heidelberg - New York. Heidelberger Taschenbücher Band 110

*Numerische Behandlung:*

J.C. Butcher [1987]: The numerical analysis of ordinary differential equations. Runge-Kutta and general linear methods. John Wiley & Sons, New York -London.

K. Dekker & J.G. Verwer [1984]: Stability of Runge-Kutta methods for stiff nonlinear differential equations. North-Holland, Amsterdam - New York - Oxford.

P. Deuflhard, F. Bornemann [1994]: Numerische Mathematik. 2. Integration gewöhnlicher Differentialgleichungen. de Gruyter Verlag 1994. ISBN 3-11-013937-5.

E. Hairer, S.P. Norsett & G. Wanner [1987]: Solving ordinary differential equations I. Nonstiff problems. Springer-Verlag, Berlin - Heidelberg - New York.

E. Hairer & G. Wanner [1991]: Solving ordinary differential equations II. Stiff and differential-algebraic problems. Springer-Verlag, Berlin - Heidelberg - New York.

M. Hanke-Bourgeois [2002]: Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens. Teubner-Verlag, Stuttgart. ISBN 3-519-00356-2. Euro 64,90.

A. Quarteroni & R. Sacco & F. Saleri [2002]: Numerische Mathematik 2. Springer-Verlag, Berlin - Heidelberg - New York. OSBN 3-540-43616-2. 29.95 Euro

J. Stoer & R. Bulirsch [1990]: Numerische Mathematik II. (3. Auflage) Springer-Verlag, Berlin - Heidelberg - New York.

H.R. Schwarz [1988]: Numerische Mathematik. (2.Auflage) Teubner-Verlag, Stuttgart u.a.

# Kapitel 2

# Runge–Kutta–Verfahren ; Stabilitätsbetrachtungen

Neben der Vorstellung dieser Verfahrensklasse möchte ich in diesem Abschnitt Stabilitätsfragen untersuchen.

## 2.1 Runge–Kutta–Verfahren

Die Idee der Runge–Kutta–Verfahren ist, durch eine geeignete Linearkombination von $s$ Näherungen $k_i$ für $x'(t_0+\alpha_i h_0), i=1,...,s$, eine sehr genaue Näherung für $x(t_0+h_0)$ zu konstruieren.

**(2.1.1) Beispiel**
Die 'verbesserte Polygonzugmethode'

$$x_{j+1} \equiv x_j + h_j f\big(t_j + \frac{h_j}{2}, x_j + \frac{h_j}{2} f(t_j, x_j)\big) \ ,$$

mit $x_i \equiv x_h(t_i), i = 0, 1, ...$ , hat die Konsistenzordnung $p=2$, falls $f \in C^2$.

□

Anstelle einer Integralnäherung unter Verwendung des Integranden in dem einen Punkt $t_j + h_j/2$ kann man natürlich auch ein gewichtetes Mittel aus Funktionswerten in mehreren Punkten nehmen.

*Verwendung von $s$ Näherungen für $x'(.) = f(., x(.))$* im Intervall $[t_j, t_j + h_j]$ liefert

**(2.1.2) ($s$–stufige) Runge–Kutta–Verfahren**
Gegeben $s \in \mathbb{N}$, und

$$[\alpha_j \mid 1 \leq j \leq s] : 0 \leq \alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_s \leq 1$$
$$[\beta_{jl} \mid 1 \leq j, l \leq s] \quad , \quad [\gamma_j \mid 1 \leq j \leq s]$$

Zu $(t, y) \in I \times \mathbb{R}^n$ und $\mathbb{R} \ni h > 0$ erfülle $[k_i \mid 1 \leq i \leq s] \equiv [k_i(t, y, h) \mid 1 \leq i \leq s] \in \mathbb{R}^{n \times s}$ das Gleichungssystem

$$k_i = f\big(t + \alpha_i h \ , \ y + h \sum_{l=1}^{s} \beta_{il} k_l\big) \quad , \quad 1 \leq i \leq s.$$

Dann ist mit $x_h(t_i) \equiv x_i \ , \ i = 0, 1, ....$

$$x_0 = x_{0,h} \ ,$$

$$x_{j+1} = x_j + h_j \sum_{l=1}^{s} \gamma_l k_l(t_j, x_j, h_j) \quad , \quad t_j \in I_h^{'} \ ;$$

ein $s$–stufiges Runge–Kutta–Verfahren.

□

Die übliche Notation von Runge–Kutta–Verfahren ist

$$\begin{array}{c|ccc} \alpha_1 & \beta_{11} & \cdots & \beta_{1s} \\ \alpha_2 & \beta_{21} & \cdots & \beta_{2s} \\ \vdots & \vdots & & \vdots \\ \alpha_s & \beta_{s1} & \cdots & \beta_{ss} \\ \hline & \gamma_1 & \cdots & \gamma_s \end{array} \quad \overset{\text{kurz}}{=:} \quad \begin{array}{c|c} \alpha & B \\ \hline & \gamma^t \end{array}$$

sog. *'Butcher–Array'* .

Von dem australischen Mathematiker J.C. Butcher stammen wesentliche Untersuchungen zu Runge–Kutta–Verfahren.

**Explizite Runge–Kutta–Verfahren** haben ein Butcher–Array der Form

$$\begin{array}{r|cccc} \alpha_1 = 0 & 0 & & \cdots & 0 \\ \alpha_2 & \beta_{21} & 0 & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \alpha_s & \beta_{s1} & \cdots & \beta_{s,s-1} & 0 \\ \hline & \gamma_1 & \cdots & \gamma_{s-1} & \gamma_s \end{array}$$

nur *explizite* Funktionsauswertungen

oder ausgeschrieben für einen Schritt: Sei $(t_j, x_j) \in I \times \mathbb{R}^n$ und $h > 0$ gegeben:

$$k_1 = f(t_0 \ , \ x_0)$$

$$k_2 = f(t_0 + \alpha_2 h \ , \ x_0 + h\beta_{21} k_1)$$

$$\vdots$$

$$k_s = f\big(t_0 + \alpha_s h \ , \ x_0 + h(\beta_{s1} k_1 + \beta_{s2} k_2 + \ldots + \beta_{s,s-1} k_{s-1})\big) \ ,$$

und

$$x_1 = x_0 \ + \ h \ ( \ \gamma_1 k_1 + \gamma_2 k_2 + \ldots + \gamma_s k_s \ ) \ .$$

□

Wie hoch die Konsistenzordnung eines Runge–Kutta–Verfahrens ist, ist nicht so leicht zu entscheiden. Durch Taylorabgleich aller Größen, entwickelt um den Punkt $(t, x(t))$, erhält man als Konsistenzbedingungen nichtlineare Gleichungen. Für explizite $s$–stufige Runge–Kutta–Verfahren gilt (vgl.*Hairer, Norsett, Wanner:* Table 2.3, p.153):

**Anzahl der Konsistenzbedingungen** ( für *explizite* $s$–stufige R–K–Verfahren)

| Konsistenzordnung $p$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Zahl der Bedingungsgln. für $\alpha_i,\ \beta_{ij},$ $\gamma_i,\ 1 \le j < i, 1 \le i \le s;$ | 1 | 2 | 4 | 8 | 17 | 37 | 85 | 200 | 486 | 1205 |

□

Die Anzahl der Parameter $\alpha_i,\ \beta_{ij}, 1 \le j < i, 1 \le i \le s;\ \gamma_i,\ 1 \le i \le s;$ ist $s(s+1)/2$.

**Anzahl der Parameter** ( für *explizite* $s$–stufige R–K–Verfahren)

| Stufenzahl $s$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Zahl der Parameter $\alpha_i,\ \beta_{ij},$ $\gamma_i,\ 1 \le j < i, 1 \le i \le s;$ | 1 | 3 | 6 | 10 | 15 | 21 | 28 | 36 | 45 | 55 |

□

Durch zusätzliche Bedingungen erhält man Vereinfachungen in den nichtlinearen Gleichungen; häufige Zusatzbedingungen sind sog. *'vereinfachenden Bedingungen'* wie z. B.

$$\sum_{l=1}^{i-1} \beta_{il} = \alpha_i \quad , \quad 2 \le i \le s; \qquad \text{d.h. gerade } \ k_i = x'(t + \alpha_i h) + O(h^2).$$

Durch Betrachtung von Spezialfällen in den Gleichungen erhält man hinreichende Bedingungen.

Welche Konsistenzordnung man *maximal* erreichen kann, untersuchte Butcher:

**Bemerkung**
Sei $p^*(s)$ die maximale Konsistenzordnung $s$–stufiger *expliziter* Runge–Kutta–Verfahren. Dann gilt

| Stufe $s$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | $s \ge 9$ |
|---|---|---|---|---|---|---|---|---|---|
| $p^*(s)$ | 1 | 2 | 3 | 4 | 4 | 5 | 6 | 6 | $p^*(s) \le s-2$ |

□

Der *Aufwand für einen Schritt* eines expliziten $s$–stufigen Runge–Kutta–Verfahrens wird bestimmt durch die *$s$ Auswertungen der $n$–komponentigen Vektorfunktion $f$* .

Ein Vergleich von $s$ mit $p^*(s)$ zeigt die Attraktivität 4–stufiger Runge–Kutta–Verfahren.

$s = 4:$

$$\begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array} \qquad\qquad \begin{array}{c|cccc} 0 & & & & \\ 1/3 & 1/3 & & & \\ 2/3 & -1/3 & 1 & & \\ 1 & 1 & -1 & 1 & \\ \hline & 1/8 & 3/8 & 3/8 & 1/8 \end{array}$$

*Standard–Runge–Kutta–Formel* , p = 4 ;

3/8–*Formel* , p = 4 .

Diese beiden Formeln haben Konsistenzordnung $p = 4$. Ein Beispiel für *implizite Formeln* sind

**(2.1.3) Beispiele** (2–stufige *implizite* Runge–Kutta–Formeln)

(a)

$$\begin{array}{c|cc} 0 & & \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}$$ sog. *Trapezregel* , $p = 2$ ;

(b)

$$\begin{array}{c|cc} (3-\sqrt{3})/6 & 1/4 & (3-2\sqrt{3})/12 \\ (3+\sqrt{3})/6 & (3+2\sqrt{3})/12 & 1/4 \\ \hline & 1/2 & 1/2 \end{array}$$ sog. *Gauß–Formel* , $p = 4$ .

□

Für $f(t,x) = g(t)$ unabhängig von $x$, ergibt (a) gerade die zusammengesetzte Sehnentrapezregel. Ebenso geht bei (b) das Runge–Kutta–Verfahren in die zusammengesetzte Gauß–Quadraturformel über. Insbesondere sind bei (b) die $\alpha_i$ die Gauß–Stützstellen und die $\gamma_i$ die Gauß–Gewichte, jeweils bezogen auf das Intervall $[0,1]$ .

*Vorteil* von impliziten Runge–Kutta–Verfahren : hohe Konsistenzordnung – maximal $p = 2s$, siehe auch Beispiel (2.1.3)$(b)$.

*Nachteil :* Die $k_i$ müssen iterativ mit einem Verfahren zur Lösung nichtlinearer Gleichungssysteme bestimmt werden.
Dieses $(n \times s)$-Gleichungssystem vereinfacht sich zu $s$ zerfallenden Gleichungsystemen jeweils in $n$ Variablen, falls

$$\beta_{il} = 0 \ , l > i \ .$$

Solche Verfahren nennt man (nur) *diagonal implizite* Runge-Kutta-Verfahren. Die obige *Trapezregel* ist von dieser Bauart.

Die Frage nach der Lösbarkeit der impliziten Runge–Kutta–Gleichungen und der Beschränktheit der Lösungen $[k_i]$ für hinreichen kleines $|h|$ kann man beantworten mit Hilfe des *Fixpunktsatzes von Banach.*

**(2.1.4) Hilfssatz**
Sei $f : I \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ aus $Lip_{loc}$ mit Lipschitzkonstante $\overline{L}$ in $U(x;\overline{c})$, und $[\alpha_i : 1 \le i \le s] \in [0,1]^s$ und $[\beta_{il} : 1 \le i,l \le s] \in \mathbb{R}^{s\times s}$ gegeben. Dann gilt:
$\forall\, c : 0 < c < \overline{c}\ \exists\, H_c > 0\ \exists\, M_c > 0\ \ \forall\, t \in I\ \ \forall\, z \in K[x(t),c]\ \ \forall\, h : 0 < h \le H_c$
$\exists_{\,1} [k_i : 1 \le i \le s] \in \Big\{ [k_i \mid 1 \le i \le s] \in \mathbb{R}^{n\times s} : \max_{1\le i\le s} \|k_i - f(t+\alpha_i h, z)\| \le M_c \Big\}$ :

$$\text{(2.1.4)(i)} \qquad k_i = f\big(t + \alpha_i h, z + h \sum_{l=1}^{s} \beta_{il} k_l\big) \quad , \quad 1 \le i \le s \ .$$

Beweis:

Zu $k = [k_i] \in \mathbb{R}^{n\times s}$ sei $\|k\| := \max_{1\le i\le s} \|k_i\|$ und $\mathbb{R}^{n\times s} \ni k \longrightarrow F(k) \in \mathbb{R}^{n\times s}$ mit $F(k) := [f\big(t + \alpha_i h, z + h\sum_{l=1}^{s} \beta_{il} k_l\big) \mid 1 \le i \le s]$ . Jeder Fixpunkt von F ist gerade Lösung des Gleichungssystems $(i)$; im folgenden zeigen wir, daß in obiger Kugel genau ein Fixpunkt $k \in \mathbb{R}^{n\times s} : F(k) = k$ existiert. Setze

$$\overline{m} := \max_{s\in I, z\in K[x(t),c]} \|f(s,z)\| \ ; \quad \overline{\beta} := \max_i \sum_{l=1}^{s} \mid \beta_{il} \mid \ ; \quad \overline{L} := \text{Lipschitzkonstante von f in } U(x;\overline{c}).$$

Wähle $M_c > 0$ (beliebig) und $H_c$ hinreichend klein, so daß gilt

$$\begin{array}{ll} (\alpha) & c + H_c\overline{\beta}(\overline{m} + M_c) \le \overline{c} \\ (\beta) & \overline{L}H_c\overline{\beta}(\overline{m} + M_c) \le M_c \\ (\gamma) & \overline{L}H_c\overline{\beta} < 1. \end{array}$$

Sei $\|z - x(t)\| \le c$ und $\|k_i - f(t + \alpha_i h, z)\| \le M_c, 1 \le i \le s$. Dann gilt

$$\|k_i\| \le \|f(t + \alpha_i h, z)\| + M_c \le \overline{m} + M_c,$$

und weiter

$$\|(z + h\sum_l \beta_{il}k_l) - x(t)\| \le c + \|h\sum_l \beta_{il}k_l\| \le c + H_c\overline{\beta}(\overline{m} + M_c) \le \overline{c} \quad \text{nach } (\alpha),$$

also

$$z \quad , \quad z + h\sum_l \beta_{il}k_l \quad \in \quad K[x(t), \overline{c}].$$

Damit kann für beide Argumente die Lipschitzkonstante $\overline{L}$ genommen werden. Es folgt $F$ *'Selbstabbildung'* obiger Kugel:

$$\begin{aligned} \|F(k)_i - f(t + \alpha_i h, z)\| &= \|f(t + \alpha_i h, z + h\sum_l \beta_{il}k_l) - f(t + \alpha_i h, z)\| \\ &\le \overline{L}H_c\overline{\beta}(\overline{m} + M_c) \\ &\le M_c \quad \text{nach } (\beta). \end{aligned}$$

$F$ ist außerdem *kontrahierend*:

$$\|F(k) - F(\widetilde{k})\| = \max_i \|f(t + \alpha_i h, z + h\sum_l \beta_{il}k_l) - f(t + \alpha_i h, z + h\sum_l \beta_{il}\widetilde{k}_l)\|$$

$$\le \overline{L}H_c\overline{\beta}\ \|k - \widetilde{k}\| \ =:\ q\|k - \widetilde{k}\| \quad \text{, wobei } q := \overline{L}H_c\overline{\beta} < 1 \text{ nach } (\gamma).$$

Nach dem Fixpunktsatz für kontrahierende Abbildungen folgt die Behauptung.

□

**(2.1.5) Bemerkungen**

Sei $z \longmapsto f(t + \alpha_i h, z) \in Lip(\widetilde{L})$ , $i = 1, .., s$, und $q := h\widetilde{L}\max_i \sum_l |\beta_{il}| < 1$ . Dann gilt

(a) Mit $k^{(-1)} = 0 \in \mathbb{R}^n$ konvergiert sowohl das Gesamtschrittverfahren (zum Beweis vgl. *Numerik I*)

$$k_i^{(r)} := f\big(t + \alpha_i h, z + h\sum_{l=1}^{s} \beta_{il}k_l^{(r-1)}\big) \quad , \quad 1 \le i \le s \ ; \ r = 0, 1, ...$$

als auch das Einzelschrittverfahren (zum Beweis vgl. *Übungen*)

$$k_i^{(r)} := f\big(t + \alpha_i h, z + h\sum_{l<i} \beta_{il}k_l^{(r)} + h\sum_{l\ge i} \beta_{il}k_l^{(r-1)}\big) \quad , \quad 1 \le i \le s \ ; \ r = 0, 1, ...$$

gegen die Lösung $k = [k_i : 1 \le i \le s]$ der Runge-Kutta-Gleichungen

$$k_i = f\big(t + \alpha_i h, z + h \sum_{l=1}^{s} \beta_{il} k_l\big) \quad , \quad 1 \le i \le s \ .$$

(b) Für das Gesamtschrittverfahren gilt die Abschätzung (zum Beweis vgl. *Numerik I*)

$$\|k - k^{(r)}\| \le \frac{q}{1-q} \|k^{(r)} - k^{(r-1)}\| \le \frac{q^r}{1-q} \|k^{(1)} - k^{(0)}\| \ .$$

(c) Wegen $q = h\widetilde{L} \max_i \sum_l |\beta_{il}| = O(|h|)$ sind für implizite Runge–Kutta–Verfahren der Ordnung $p$ also mindestens $p+1$ Iterationen obiger Art nötig mit Startwert
$k^{(0)} := \big[f\big(t + \alpha_i h, x_h(t)\big) \mid 1 \le i \le s\big]$ oder vereinfacht $k^{(0)} := \big[f\big(t, x_h(t)\big) \mid 1 \le i \le s\big]$ ,
um die Konvergenzordnung $p$ nicht zu zerstören.

(d) Für große Lipschitzkonstanten $\widetilde{L}$ erzwingt die Konvergenzbedingung $h\widetilde{L} \max\limits_i \sum\limits_l |\beta_{il}| < 1$, daß die Schrittweiten $h$ sehr klein sind. Da implizite Runge–Kutta–Verfahren aber gerade gemacht sind, um trotz größerer Schrittweiten Stabilität zu erhalten, ist obige Bedingung zu restriktiv. Empfehlenswert ist dann anstelle der Fixpunktiteration das Newtonverfahren (bzw. Varianten davon, vgl. *Numerik I*).

□

Um aus der Konsistenz(ordnung) auf die Konvergenz(ordnung) schließen zu können, benötigen wir die asymptotische Stabilität.

**(2.1.6) Satz**
Sei $f : I \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ aus $Lip_{loc}$. Dann ist für das $s$-stufige Runge–Kutta–Verfahren die zugehörige Schar von Verfahrensfunktionen aus $Lip_{loc}$, gleichmäßig in $h$ für $|h|$ hinreichend klein.
Gilt zusätzlich die Konsistenzbedingung $\sum_{i=1}^{s} \gamma_i = 1$ , so gilt für das Runge–Kutta–Verfahren der Konvergenzsatz (1.16).

<u>Beweis:</u>

*Behauptung:* '$f_h$ *Lipschitzstetig* in $U(x;\overline{c})$, gleichmäßig in $h \in \mathcal{H}$' für $|h|$ hinreichend klein.
*Beweis:* Sei $f$ in $U(x;\overline{c})$ Lipschitzstetig bzgl. der $y$-Variablen mit Lipschitzkonstante $\overline{L}$. Nach Hilfssatz (2.1.4) gilt dann

$$\forall\, 0 < c < \overline{c} \;\; \exists\, H_c, M_c > 0 \;\; \forall\, h : |h| \le H_c \;\; \forall\, t \in I_h \;\; \forall\, z \in K[x(t), c] \;\; \exists_1 \quad [k_i(t,z) \mid 1 \le i \le s] :$$

$$k_i = f\big(t + \alpha_i h(t), z + h(t) \sum_l \beta_{il} k_l\big), 1 \le i \le s \ ; \ \wedge \ \max_i \|k_i - f(t + \alpha_i h(t), z)\| \le M_c \ .$$

$k_i \in Lip$: Für $y, z \in K[x(t), c]$ gilt ja

$$\begin{aligned} & \|k_i(t,y) - k_i(t,z)\| \\ = \;& \|f(t + \alpha_i h(t), y + h(t) \textstyle\sum_l \beta_{il} k_l(t,y)) - f(t + \alpha_i h(t), z + h(t) \sum_l \beta_{il} k_l(t,z))\| \\ \le \;& \overline{L}\big[\|y - z\| \ + \ h(t) \cdot \overline{\beta} \ \max_l \|k_l(t,y) - k_l(t,z)\|\big] \\ \le \;& \overline{L}\|y - z\| \ + \ \overline{L} H_c \overline{\beta} \ \max_l \|k_l(t,y) - k_l(t,z)\|. \end{aligned}$$

Für $\overline{L} H_c \overline{\beta} < 1$ (o.E. immer möglich nach eventueller Verkleinerung von $H_c$) ist

$$\max_i \|k_i(t,y) - k_i(t,z)\| \ \le \ 1/(1 - \overline{L} H_c \overline{\beta}) \ \|y - z\|,$$

also $k_i(t,\cdot)$ Lipschitzstetig mit Konstante $1/(1-\overline{L}H_c\overline{\beta})$ . Wegen

$$f_h(t,z) = \sum_{i=1}^{s} \gamma_i k_i(t,z)$$

folgt, daß $(f_h : |h| \leq H_c)$ auf $\bigcup_{t\in I} \{t\} \times K[x(t),c]$ gleichmäßig Lipschitzstetig ist mit Konstante

$$1/(1-\overline{L}H_c\overline{\beta}) \sum_{l=1}^{s} |\gamma_l|.$$

$\diamond$

*Behauptung:* ' $(A_h)$ *konsistent* '.
*Beweis:* Es ist $(D_h x)(t) = f(t,x(t)) + o(|h|^0)$ , da $x \in C^1$ ;

$$f_h(t,x(t)) = \sum_{i=1}^{s} \gamma_i \underbrace{f\big(t+\alpha_i h(t), x(t) + h(t)\sum_l \beta_{il} k_l(t,x(t))\big)}_{= k_i(t,x(t))} ;$$

also

$$\begin{aligned}
\tau_h(t) &= f\big(t,x(t)\big) + o\big(|h|^0\big) - \sum_{i=1}^{s} \gamma_i f\big(t+\alpha_i h(t), x(t) + h(t)\sum_l \beta_{il} k_l(t,x(t))\big) \\
&= f\big(t,x(t)\big) + o\big(|h|^0\big) - \sum_i \gamma_i f\big(t+\alpha_i h(t), x(t)\big) + O\big(|h|\big) \text{ , denn } \|k_l\| \leq \overline{m} + M_c \text{ ,} \\
&= f\big(t,x(t)\big)\,[1 - \sum_{i=1}^{s} \gamma_i] + o\big(|h|^0\big) \text{ , denn es ist}
\end{aligned}$$

$\sum_i \gamma_i f\big(t+\alpha_i h(t), x(t)\big) = \sum_i \gamma_i f\big(t,x(t)\big) + o\big(|h|^0\big)$ , da $f$ auf dem Kompaktum $\{(t+h, x(t)) \mid 0 \leq h \leq H_c \text{ , } t, t+h \in I\}$ gleichmäßig stetig ist.

Damit ist gezeigt, daß alle Voraussetzungen von (1.16) erfüllt sind.

$\square$

Nach dem letzten Satz wissen wir, daß für $|h| \to 0$ die exakte Lösung einer Anfangswertaufgabe auf einem kompakten Intervall $[a,b]$ beliebig genau approximiert wird *für jedes Runge–Kutta–Verfahren, das konsistent ist.* Damit wird die Konvergenzgüte für unser Ausgangsproblem (1.1) bestimmt durch die Konsistenzordnung.

Ich möchte im nächsten Abschnitt zwei Modifikationen unserer Ausgangsaufgabe (1.1) beschreiben, und die beiden zugehörigen Stabilitätsbegriffe diskutieren.

## 2.2 A– und B–Stabilität

Ich beschreibe hier für RK–Verfahren Stabilitätseigenschaften, die man ähnlich auch für andere Ein– und Mehrschrittverfahren analysieren kann. Bei der A–Stabilität betrachtet man die Testgleichung $x' = qx$ mit $Re\ q < 0$ resp. $x' = Qx$ mit $Re\ S(Q) < 0$.
Bei der B–Stabilität betrachtet man eine etwas verallgemeinerte Testgleichung, sog. einseitig Lipschitzstetige rechte Seiten $f$ mit einseitiger Lipschitzkonstante $\Lambda \leq 0$ (vgl. (2.2.10)).

*A–Stabilität*

Betrachtet man die Realisierung auf einem Rechner, so ist $\inf|h| > 0$. Für festes $|h|$ und sehr große Intervalle $I$, d.h. in der mathematischen Idealisierung:

$$h > 0 \ \text{ fest} \quad , \quad t \to \infty;$$

hängt das Verhalten der Näherung $x_h$ von Eigenschaften des betrachteten Einschrittverfahrens (bzw. des Mehrschrittverfahrens) ab, die *nicht* durch die asymptotische Stabilität beschrieben werden.
Bei der A–Stabilität wird das Verhalten von $x$ und $x_h$ bezüglich der (einfachsten) linearen skalaren Testgleichung

$$x' = qx$$

verglichen. Systeme von linearen Differentialgleichungen mit konstanten Koeffizienten kann man auf den skalaren Fall zurückführen, wenn die Matrix $Q$ des Differentialgleichungssystems

$$x'(t) = Qx(t)$$

diagonalähnlich und unabhängig von $t$ ist.

**(2.2.1) Beispiel**
Für die Anfangswertaufgabe $x' = qx$ , $x(t_0) = x_0 \in \mathbb{R}^1$, mit $q \in \mathbb{R}$ sei

$ESV^1$ das Polygonzugverfahren (1.5) ; ${x_h}^1$ die zugehörige Näherung,

$ESV^2$ das rückwärtsgenommenes Polygonzugverfahren (1.6) ; ${x_h}^2$ die zugehörige Näherung .

Für äquidistante Gitterpunkte $t_j = t_0 + jh$ , $j = 0, 1, \ldots$ , $h > 0$ , gilt für $t_j \to \infty$ im Fall $q < 0$:

$$x(t_j) = x(t_0)e^{q(t_j - t_0)} \quad \longrightarrow 0 \ ;$$

$${x_h}^1(t_j) = x(t_0)(1+qh)^j \begin{cases} \longrightarrow 0 & \text{, falls } |1+qh| < 1, \\ \text{(alternierend) beschränkt} & \text{, falls } |1+qh| = 1, \\ \text{(alternierend) unbeschränkt} & \text{, falls } |1+qh| > 1; \end{cases}$$

$${x_h}^2(t_j) = x(t_0)\frac{1}{(1-qh)^j} \quad \longrightarrow 0 \text{ , denn } 0 < 1/(1-qh) < 1 \ .$$

□

Für die Differentialgleichung ist das Aussehen der Lösung $x$ ja bekannt, für ${x_h}^1$, ${x_h}^2$ folgt dies sofort durch Induktion.

**(2.2.2) Definition**
Das betrachtete Ein– oder Mehrschrittverfahren habe für die Testgleichung $x' = qx$ die Gestalt $x_h(t_j + h) = g(hq)x_h(t_j)$. Dann heißt $g$ die *Stabilitätsfunktion* des Verfahrens und

(*i*) $H_a := \{z \in \mathbb{C}\colon \ | \, g(z) \, | \leq 1\}$

der *Bereich der absoluten Stabilität* des Verfahrens. Das *Verfahren* heißt *absolut stabil* oder *A–stabil* , falls $\{z \in \mathbb{C}\colon Re\ z < 0\} \subset \{z \in \mathbb{C}\colon | \, g(z) \, | < 1 \}$

□

Denn für beliebige Anfangswerte $x_h(t_0)$ gilt für die Testgleichung $x' = qx$ wegen $x_h(t_0 + jh) = g(hq)^j x_h(t_0)$

$$\sup_{j\in\mathbb{N}_0} \mid x_h(t_0 + jh) \mid < \infty \quad \Longleftrightarrow \quad hq \in H_a \ ,$$

und

$$\lim_{j\to\infty} x_h(t_0 + jh) = 0 \quad \Longleftrightarrow \quad hq \in int(H_a) \ .$$

Die exakten Lösungen der Testgleichung $x' = qx$ lauten $\quad x(t) = x(t_0)e^{q(t-t_0)}$ ,

also gilt: $\quad \lim_{t\to\infty} x(t) = 0 \Longleftrightarrow Re\ q < 0$ für $q \in \mathbb{C}$.

*Absolut stabil* ist ein Verfahren also genau dann, wenn die Näherungen bei festem $h > 0$ für $t_j \to \infty$ verschwinden zumindest in den Fällen, in denen die exakte Lösung für $t \to \infty$ verschwindet.

Für die *AWA* $x' = qx$ gilt für die Lösung $x(t_0 + h) = e^{hq}x(t_0)$ . Daher werden auch die Konsistenzeigenschaften – zumindest bezüglich der linearen Testgleichung $x' = qx$ – eines Verfahrens $x_h(t_0 + h) = g(hq)x(t_0)$ , bestimmt dadurch, wie gut $g(z)$ die Funktion $e^z$ approximiert für $z \to 0$: $\quad \tau_h(t) = O\Big(\dfrac{e^{hq} - g(hq)}{h}\Big)x(t)$ .

Die *praktische Bedeutung* von Verfahren mit großem Stabilitätsbereich, insbesondere von *absolut stabilen Verfahren*, erkennt man an *'steifen'* Differentialgleichungssystemen, z.B. linearen steifen Systemen

$$x' = Qx \quad , \quad Q \in \mathbb{R}^{n\times n} \ : \ S(Q) \subset \ ]-\infty, 0[ \ .$$

*Steifheitsrate* : $\quad \kappa := \dfrac{\max\limits_i |\,\lambda_i\,|}{\min\limits_i |\,\lambda_i\,|} \gg 1$ , wobei $S(Q) = \{\lambda_1, \ldots, \lambda_n\}$ die Eigenwerte von $Q$ sind.

**(2.2.3) Beispiel**
Betrachte die Differentialgleichung

$$x' = Qx \ , \ x(0) = \begin{bmatrix} c_1 + c_2 \\ c_1 - c_2 \end{bmatrix} \quad \text{mit} \quad Q = \begin{bmatrix} (\lambda_1 + \lambda_2)/2 & (\lambda_1 - \lambda_2)/2 \\ (\lambda_1 - \lambda_2)/2 & (\lambda_1 + \lambda_2)/2 \end{bmatrix}$$

und $\lambda_1 < \lambda_2 < 0$ , $\lambda_i \in \mathbb{R}$ ; $c_i \in \mathbb{R}$ . Sei $\lambda_1 = -1000$, $\lambda_2 = -1$, das System also 'steif'.

◇

Lösung ist

$$x(t) = c_1 e^{\lambda_1 t} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + c_2 e^{\lambda_2 t} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

denn es ist (nach Ansatz) $\lambda_1$ Eigenwert von $Q$ zum Eigenvektor $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ , $\lambda_2$ Eigenwert von $Q$ zum Eigenvektor $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ .

*Genauigkeitsforderung* für Näherungsverfahren:
Passend zur Lösung kann nach einigen Schritten, wenn $e^{-1000\ t}$ fast Null ist, die Schrittweite $h$ allein nach dem Konsistenzfehler der Komponente $e^{-1\ t}$ gewählt werden, d.h.

$h$ kann groß gewählt werden!

*Stabilitätsforderung* für Näherungsverfahren: $h\lambda_1$ , $h\lambda_2$ $\in H_a$ ,
denn nur dann bleiben die Näherungen beschränkt.
*Bei absolut stabilen Verfahren* gilt dies für $\lambda_i \in S(Q) \subset\ ]-\infty, 0[$ *bei jeder Wahl von* $h > 0$ wegen $h\lambda_i < 0$.

Für *nicht absolut stabile* Verfahren gilt: Obwohl der Genauigkeitswunsch, abgesehen von einer kleinen Umgebung des Anfangspunktes $t = 0$, keine Einschränkung für $h$ bedeutet, bedingt die Forderung der Beschränktheit der Näherungslösung für alle $t$:

$h\lambda_1 \in H_a$ bedingt kleine Schrittweiten bei *nicht abs. stab. Verf.*

Für das Polygonzugverfahren z.B. bedeutet dies für das obige Beispiel (2.2.3)

$h < \frac{2}{1000}$ für Polygonzugverfahren.

□

Zu kleine Schrittweiten machen ein Verfahren nicht nur impraktikabel aufgrund des immensen Rechenaufwandes, sondern u.U. auch völlig wertlos aufgrund zu großer Rundungsfehler, die eine sehr kleine Schrittweite mit sich bringt (vgl. die Bemerkung am Ende von Kapitel 1).

Obige Analyse des Stabilitätsverhaltens von vorwärts– bzw. rückwärtsgenommenem Polygonzugverfahrens basierte auf der expliziten Kenntnis der Stabilitätsfunktion $g(z)$. Für Runge–Kutta–Verfahren gilt aber ganz allgemein

**(2.2.4) Satz**
Jedes $s$–stufige Runge–Kutta–Verfahren mit dem Butcher-Array

$$\begin{array}{c|c} \alpha & B \\ \hline & \gamma^t \end{array}$$

hat die Stabilitätsfunktion

$$g(z) = \det(E - zB + z\mathbb{1}\gamma^t)/\det(E - zB).$$

Hier ist $E \in \mathbb{R}^{s\times s}$ die Einheitsmatrix, $\mathbb{1} := [1 \cdots 1]^t \in \mathbb{R}^s$, und $E - zB$ invertierbar vorausgesetzt.

<u>Beweis:</u>

Betrachtet man die Testgleichung $x' = qx$, dann gilt für einen Runge–Kutta–Schritt von $(t_0, x_0)$ zu $(t_0 + h, x_1)$

$$k_i = q\big(x_0 + h\sum_{l=1}^{s}\beta_{il}k_l\big) \quad , \quad 1 \le i \le s \ ,$$

$$x_1 = x_0 + h\sum_{i=1}^{s}\gamma_i k_i \ .$$

Die Auflösung dieses linearen Gleichungssystems für $k_1, \cdots, k_s, x_1$ nach der Cramerschen Regel ergibt

$$x_1 \ = \ \frac{qx_0\det(E - hqB + hq\mathbb{1}\gamma^t)}{q\det(E - hqB)} \ \equiv \ g(hq)x_0 \ .$$

□

**(2.2.5) Beispiele**

(a) Polygonzugverfahren : $g(z) = 1 + z$

$s = 1$

$$\begin{array}{c|c} \alpha & B \\ \hline & \gamma \end{array} = \begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \qquad \frac{\det(1 - z \cdot 0 + z \cdot 1 \cdot 1)}{\det(1 - z \cdot 0)} = 1 + z$$

*nicht* absolut stabil.

(b) rückwärtsgenommenes Polygonzugverfahren : $g(z) = \dfrac{1}{1-z}$

$s = 1$

$$\begin{array}{c|c} \alpha & B \\ \hline & \gamma \end{array} = \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \qquad \frac{\det(E - zB + z\mathbb{1}\gamma^t)}{\det(E - zB)} = \frac{1 - z + z}{1 - z} = \frac{1}{1 - z}$$

*absolut stabil.*

□

Wie sieht der Stabilitätsbereich expliziter Runge–Kutta–Verfahren aus? Das Aussehen der absoluten Stabilitätsbereiche erkennt man in Abbildung 2.1 (wegen $|g(\bar{z})| = |g(z)|$ ist nur der Teil in der oberen Halbebene gezeichnet).

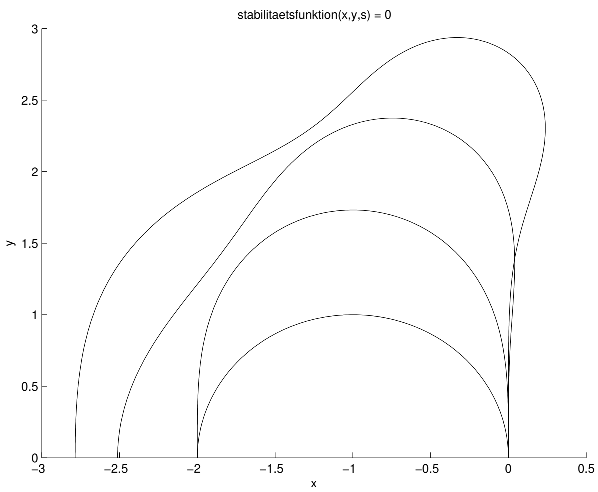


Abbildung 2.1: Stabilitätsbereich (in der oberen Halbebene) der RK-Verfahren mit p=s, s=1,2,3,4.

**(2.2.6) Satz**

Für die expliziten Runge–Kutta–Verfahren, deren Konsistenzordnung gleich der Stufenzahl $s$ ist, ist der Bereich der absoluten Stabilität unabhängig vom speziellen Verfahren. Für reelle $z = qh$ erhält man folgende Stabilitätsintervalle:

| $s$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $H_a \cap \mathbb{R}$ | $[-2, 0]$ | $[-2, 0]$ | $\doteq [-2.51, 0]$ | $\doteq [-2.78, 0]$ |
| $g(z)$ | $1 + z$ | $1 + z + \frac{z^2}{2}$ | $1 + z + \frac{z^2}{2} + \frac{z^3}{6}$ | $1 + \frac{z}{1!} + \frac{z^2}{2!} + \frac{z^3}{3!} + \frac{z^4}{4!}$ |

□

Eine Konsequenz von (2.2.4) ist, daß es keine absolut stabilen expliziten Runge–Kutta–Verfahren geben kann: für explizite RK–Verfahren ist $g$ ein Polynom und $\deg(g) \geq 1$ – sonst ist das Verfahren ja nicht konsistent – , und damit $\lim_{x\to-\infty} |g(x)| = +\infty$.

**(2.2.7) Satz**

(a) Die Gauß–Verfahren (für m=2 vgl. (2.1.3)(*b*)) sind absolut stabil.

(b) Die *Radau–IA*– und *Radau–IIA*–Verfahren sind absolut stabil.

$\square$

Vgl. etwa *Hairer, Wanner*, Theorem 12.9, S. 198. Dort wird für diese Verfahren die hinreichende Eigenschaft *'algebraisch stabil'* gezeigt.

<u>*Linienmethode und B–Stabilität*</u>

Als zweite Verallgemeinerung der Ausgangsaufgabe (1.1) betrachte ich die Stabilität für eine ganze Schar $\mathcal{F}$ von rechten Seiten $f \in \mathcal{F}$ statt für eine einzige rechte Seite $f$. Bei der B–Stabilität sind dies – allgemeiner als bei der A–Stabilität – *einseitig Lipschitzstetige* $f$ mit *einseitiger Lipschitzkonstante* $\Lambda \leq 0$ (s. (2.2.10)). Wo ist eine solche Betrachtungsweise wichtig?

**(2.2.8) Linienmethode** (*für die Wärmeleitungsgleichung*)

Wärmeleitungsgleichung in einer exemplarisch möglichst einfachen Situation:

Die Temperatur $u(s,t)$ im Punkt $s \in \mathbb{R}$ (eine Raumvariable) zur Zeit $t$ erfüllt

$$\textbf{\textit{(i)}} \quad \begin{cases} \dfrac{\partial u}{\partial t}(s,t) = \dfrac{\partial^2 u}{\partial s^2}(s,t) \quad , \quad 0 < s < 1 \quad , \quad t > 0 \\ u(s,0) = u_0(s) \quad , \quad 0 \leq s \leq 1 \quad \text{(Anfangsverteilung)} \\ u(0,t) = 0 = u(1,t) \quad \text{(Ränder konstant gekühlt)} \end{cases}$$

Als Näherung für die 2. Ableitung nehmen wir den Differenzenquotienten 2. Ordnung

$$\frac{\partial^2 u}{\partial s^2}(s,t) \quad \doteq \quad \frac{[u(s-\Delta s,t) - 2u(s,t) + u(s+\Delta s,t)]}{(\Delta s)^2}$$

für $s = s_j := j\Delta s$, $1 \leq j \leq n$ mit $\Delta s := 1/(n+1)$. Sei weiter $s_0 := 0$ , $s_{n+1} := 1$. Für die Approximationen

$$U_j(\cdot) \doteq u(j\Delta s, \cdot) \quad , \quad 1 \leq j \leq n$$

erhält man ein System von gewöhnlichen Differentialgleichungen 1.Ordnung

$$\textbf{\textit{(ii)}} \quad \begin{cases} {U_1}'(t) = [\qquad\quad -2U_1(t) + U_2(t)\ ]/(\Delta s)^2 \ , \ U_0(t) \equiv u(s_0,t) = 0; \\ {U_j}'(t) = [U_{j-1}(t) - 2U_j(t) + U_{j+1}(t)]/(\Delta s)^2 \quad , \quad 1 < j < n \ ; \\ {U_n}'(t) = [U_{n-1}(t) - 2U_n(t) \qquad\quad ]/(\Delta s)^2 \ , \ U_{n+1}(t) \equiv u(s_{n+1},t) = 0; \end{cases}$$

Damit erhält man für $U = [U_j \mid 1 \leq j \leq n,]$

$$\textbf{\textit{(iii)}} \quad \left\{ U' = \underbrace{\left(\frac{1}{(\Delta s)^2}\, A_n\right)}_{=:\, Q_{\Delta s} \in \mathbb{R}^{n\times n}} U \ , \quad \text{wobei} \quad A_n := \begin{bmatrix} -2 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & 1 \\ & & & & 1 & -2 \end{bmatrix} \in \mathbb{R}^{n\times n} \ . \right.$$

Daß $Q_{\Delta s}$ bzgl. $\|\cdot\|_2$ im $\mathbb{R}^n$ einerseits eine sehr große Lipschitzkonstante hat (i.allg. $\geq 10^4$),

$$\textbf{\textit{(iv)}} \quad \begin{cases} \text{Lipschitzkonstante von} \quad Q_{\Delta s} \doteq \dfrac{1}{(\Delta s)^2} \\ \text{einseitige Lipschitzkonstante } \big(\text{ s. Def. (2.2.10)}\big) \;=\; 0 \;, \end{cases}$$

denn $A_n$ und damit $Q_{\Delta s}$ ist negativ definit. Denn nach dem Satz von Gerschgorin (s. *Numerik I*) folgt direkt die für unsere Zwecke ausreichende Aussage

$$S(A_n) \subset [-4, 0] \;,$$

also ist (2.2.10) anwendbar.

□

**(2.2.9) Bemerkung**
Die Steifheitsrate von (2.2.8) *(iii)* ist $O\Big(\dfrac{1}{(\Delta s)^2}\Big)$.

Beweis als Übung.

**(2.2.10) Definition**
$f : I \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ erfüllt eine (globale) *einseitige Lipschitzbedingung*, falls Konstante $\Lambda \in \mathbb{R}$ existiert mit

$$\langle f(t,y) - f(t,z), y - z \rangle \leq \;\Lambda\; \|y - z\|_2{}^2 \quad , \quad \forall\, t \in I \qquad \forall\, y, z \in \mathbb{R}^n$$

Dabei ist $\langle\ ,\ \rangle$ das euklidische Skalarprodukt im $\mathbb{R}^n$ und $\|\cdot\|_2$ die euklidische Norm. Die Konstante $\Lambda$ heißt *einseitige Lipschitzkonstante* .
Analog wird die *lokale* einseitige Lipschitzbedingung in einer Umgebung der Lösung x definiert.

□

**Beispiel**
$f(t,x) = Q_{\Delta s}\, x$ mit $Q_{\Delta s}$ aus (2.2.8) erfüllt eine einseitige Lipschitzbedingung mit Konstante $\Lambda = 0$.

□

**(2.2.11) B–Stabilität** (*Butcher 1975*)
Ein Runge–Kutta–Verfahren mit Verfahrensfunktionen $(f_h \mid h \in \mathcal{H})$ heißt *B–stabil* , wenn für jede rechte Seite $f$, die einer einseitigen Lipschitzbedingung mit $\Lambda = 0$ genügt, für alle $h > 0$ und alle $t, t + h \in I$ und alle $y_0$, $z_0 \in \mathbb{R}^n$ die Runge–Kutta–Näherungen $y_1$ und $z_1$,

$$y_1 = y_0 + h f_h(t, y_0, t+h, y_1) \quad , \quad z_1 = z_0 + h f_h(t, z_0, t+h, z_1) \quad ;$$

erfüllen

$$\|y_1 - z_1\|_2 \leq \|y_0 - z_0\|_2 \;.$$

□

Die Näherungen zeigen dann also ein 'Kontraktionsverhalten' analog zu den exakten Lösungen, denn es gilt

**Bemerkung**
Sei $f : I \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ stetig und einseitig Lipschitzstetig mit Lipschitzkonstante $\Lambda = 0$. Gilt $y'(t) = f(t, y(t))$ und $z'(t) = f(t, z(t))$ in $I$, so folgt für $t \in I$ und $h > 0$

$$\|y(t+h) - z(t+h)\|_2{}^2 \leq \|y(t) - z(t)\|_2{}^2 \ .$$

Beweis:

$$\|y(t+h) - z(t+h)\|_2{}^2 = \|y(t) - z(t)\|_2{}^2 + \int\limits_t^{t+h} \underbrace{\frac{d}{d\tau}\|(y-z)(\tau)\|_2{}^2}_{\leq 0 \ (*)} \, d\tau \leq \|y(t) - z(t)\|_2{}^2 \ ,$$

$(*)$ gilt wegen $\dfrac{d}{dt}\|(y-z)(t)\|_2{}^2 = 2\big\langle f\big(t, y(t)\big) - f\big(t, z(t)\big) \ , \ y(t) - z(t)\big\rangle \leq 0$ wegen $\Lambda = 0$ .

□

Vorbereitend für den Beweis von Satz (2.2.13) beschreibe ich Runge–Kutta–Verfahren in sog. *'integrierter'* Form (da die $X_i \doteq x(t_0 + \alpha_i h)$ sind – Nachweis als Übung – im Gegensatz zu $k_i \doteq x'(t_0 + \alpha_i h)$). Wegen ihrer Analogie zu *linearen MSV'n* heißt diese Darstellung auch *'lineare Form'* ):

$$\begin{aligned} &t_0, x_0, h > 0 \text{ gegeben} \ ; \\ &X_i = x_0 + h \sum_{k=1}^{s} \beta_{ik} f(t_0 + \alpha_k h, X_k) \ , \ 1 \leq i \leq s; \\ &x_1 = x_0 + h \sum_{i=1}^{s} \gamma_i f(t_0 + \alpha_i h, X_i) \ . \end{aligned}$$

Ich möchte zu gegebenen $0 \leq \alpha_1 < \alpha_2 < \ldots < \alpha_{s-1} < \alpha_s \leq 1$ den Zusammenhang der RK–Näherungen mit der *polynomialen Kollokationslösung* $X \in \mathbb{P}^n_{s+1} := X^n_{k=1}\mathbb{P}_{s+1}$

$X : [t_0, t_0 + h] \longrightarrow \mathbb{R}^n$ mit

$$X(t_0) = x_0 \ , \ X'(t_0 + \alpha_i h) = f\big(t_0 + \alpha_i h, X(t_0 + \alpha_i h)\big) \ , \ 1 \leq i \leq s \ ;$$

beschreiben:

**(2.2.12) Hilfssatz**
Sei $0 \leq \alpha_1 < \alpha_2 < \ldots < \alpha_{s-1} < \alpha_s \leq 1$ gegebenen.

(a) Jedes $X \in \mathbb{P}^n_{s+1}$ ist eindeutig bestimmt durch die $s+1$ Werte

$$X(t_0) = x_0 \ , \ X'(t_0 + \alpha_i h) = k_i \ , \ 1 \leq i \leq s \ .$$

(b) Sei $l_j \in \mathbb{P}_s \ : \ l_j(\alpha_i) = \delta_{ij}, 1 \leq i, j \leq s$. Dann gilt mit $l_j^{(-1)}(t) := \int\limits_0^t l_j(s)ds$, daß

$$X(t_0 + s) := x_0 + h \sum_{j=1}^{s} k_j l_j^{(-1)}\Big(\frac{t - t_0}{h}\Big)$$

die Interpolationsaufgabe in (a) löst.

(c) Für $\beta_{ij} := \int\limits_0^{\alpha_i} l_j(s)ds, 1 \leq i, j \leq s$, und $\gamma_i := \int\limits_0^1 l_i(s)ds, 1 \leq i \leq s$, und die RK–Näherung $x_1$ zu $\begin{array}{c|c} \alpha & B \\ \hline & \gamma^t \end{array}$ , ausgehend von $t_0, x_0$ und $h > 0$, zeige man: Ist $X \in \mathbb{P}^n_{s+1}$ Kollokationslösung, so gilt $X(t_0 + h) = x_1$.

□

Beweis als Übung.

**(2.2.13) Satz**
Die impliziten Runge–Kutta–Formeln vom Gauß–Typ (für $s = 2$ siehe Beispiel (2.1.3)(b) ) sind B–stabil.

Beweis:

Ausgehend von $y_0, z_0$ in $t_0$ liefere das RK–Verfahren die Näherungen $y_1, z_1$ in $t_1 = t_0 + h$. Zu zeigen ist $\|y_1 - z_1\|_2^2 \leq \|y_0 - z_0\|_2^2$. Bezeichne $Y(t), Z(t)$ die zugehörigen Kollokationslösungen. Dann gilt für $d(t) := \|Y(t) - Z(t)\|_2{}^2$ die Identität $d(t_1) = d(t_0) + h \int_0^1 d'(t_0 + sh)ds$. Ist $\int_0^1 d'(t_0 + sh)ds \leq 0$, so folgt
$\|y_1 - z_1\|_2{}^2 = \|Y(t_1) - Z(t_1)\|_2{}^2 \leq \|Y(t_0) - Z(t_0)\|_2{}^2 = \|y_0 - z_0\|_2{}^2$.

$$\int_0^1 d'(t_0 + sh)ds$$

$$= 2h \int_0^1 < Y'(t_0 + sh) - Z'(t_0 + sh), Y(t_0 + sh) - Z(t_0 + sh) > ds$$

$$= 2h \sum_{i=1}^{s} \gamma_i < Y'(t_0 + \alpha_i h) - Z'(t_0 + \alpha_i h), Y(t_0 + \alpha_i h) - Z(t_0 + \alpha_i h) > ,$$

da die Gauß–Quadratur exakt ist für $(Y' - Z') \cdot (Y - Z) \in \mathbb{P}_{2s}^n$

$$= 2h \sum_{i=1}^{s} \gamma_i < f\big(t_0 + \alpha_i h, Y(t_0 + \alpha_i h)\big) - f\big(t_0 + \alpha_i h, Z(t_0 + \alpha_i h)\big), Y(t_0 + \alpha_i h) - Z(t_0 + \alpha_i h) > ,$$

nach Definition der Kollokationsnäherungen

$\leq 0$, denn für die Gaußgewichte gilt $\gamma_i > 0$ , $1 \leq i \leq s$ , und es ist

$$< f\big(t_0 + \alpha_i h, Y(t_0 + \alpha_i h)\big) - f\big(t_0 + \alpha_i h, Z(t_0 + \alpha_i h)\big), Y(t_0 + \alpha_i h) - Z(t_0 + \alpha_i h) > \leq 0$$

für $1 \leq i \leq s$ wegen der einseitigen Lischitzkonstante $\Lambda = 0$ von $f$.

□

Die hier speziell im Zusammenhang mit Runge–Kutta–Verfahren betrachteten Stabilitätsprobleme und –eigenschaften kann man natürlich für alle Verfahrenstypen untersuchen.
Noch eine Warnung: Die sehr schönen *Stabilitätseigenschaften* von *impliziten Runge–Kutta–Verfahren* gelten (*theoretisch*) natürlich nur *bei exakter Lösung der impliziten Gleichungen.* In der *praktischen* Durchführung bedeutet dies eine adaptiv *genügend gute näherungsweise Lösung, d.h.* einen a priori unkontrollierbaren, und i.a. *sehr hohen Aufwand. Beschränkt man* von vornherein den *Aufwand* , *z.B.* durch

nur einen Iterationsschritt zur Lösung der Fixpunktgleichung für die $k_i$

*verliert man die Stabilitätseigenschaften* (Verdeutlichung dieses Effekts in den Übungen).

Daß andererseits eine 'Lösung' der RK–Gleichungen mit hinreichend kleinem Fehler auch nur einen kleinen Fehler in der resultierenden 'gestörten' Näherung $\widetilde{x}_h(t+h)$ bewirkt, ist eine zusätzliche Stabilitätseigenschaft, die sog. *BS-Stabilität.* Auch diese Stabilitätseigenschaft haben die *Gauß-Runge-Kutta-* Verfahren.

Eine mit diagonal impliziten Runge–Kutta–Verfahren

$$(*) \quad k_i = f\big(t + \alpha_i h\ ,\ y + h \sum_{l=1}^{i} \beta_{il} k_l\big) \quad , \quad 1 \le i \le s;$$

zusammenhängende Verfahrensklasse sind die Methoden vom *Rosenbrock-Typ.* Im einfachsten Fall führt man dabei, ausgehend von ${k_i}^{(0)} \equiv 0$ nur einen (vereinfachten) Newtonschritt zur Lösung von $(*)$ durch, und erhält ${k_i}^{(1)} \equiv: k_i$ aus dem folgenden linearen Gleichungssystem

$$\Big(E - h\beta_{ii} \frac{\partial f}{\partial x}\big(t, x_0\big)\Big) k_i = f\big(t + \alpha_i h, x_0 + h \sum_{l=1}^{i-1} \beta_{il} k_l\big) \quad , \quad 1 \le i \le s.$$

## 2.3 Fehlerschätzer und Schrittweitensteuerung; Eingebettete Runge–Kutta–Formeln

Im diesem Abschnitt möchte ich Paare sog. *eingebetteter Runge–Kutta–Verfahren* vorstellen im Zusammenhang mit der Fehlerschätzung und einer daraus resultierenden automatischen Anpassung der Schrittweite, der sog. *Schrittweitensteuerung.* Im folgenden betrachte ich Anfangswertaufgaben zu einer gegebenen rechten Seite $f$. Die Abhängigkeit aller im folgenden betrachteten Größen von der rechten Seite $f$ unterdrücke ich. Dagegen möchte ich die Abhängigkeit vom gerade betrachteten Anfangswert $(t, z)$ (anstelle von $(t_0, x_0)$) auch in der Notation zum Ausdruck bringen: $x(\cdot\ ; t, z)$ löst die Anfangswertaufgabe

$$x(t; t, z) = z\ ,\ x'(s; t, z) = f(s, x(s; t, z)) \quad , \quad s \in I\ : s > t.$$

Analog bezeichne $\widetilde{x}(\cdot\ ; t, z, h)$ für $h > 0$ die Näherungslösung, die ein oder mehrere Schritte des betrachteten $ESV'n$ für den Startwert $(t, z)$ liefert, d.h. mit der Verfahrensfunktion $\Phi$ – vgl. (1.8) – gilt

$$\widetilde{x}(t; t, z, h) := z;$$

$$\widetilde{x}(t + (j+1)h; t, z, h) := \widetilde{x}(t + jh; t, z, h) + \Phi(t + jh, \widetilde{x}(t + jh; t, z, h), h, f)\ .$$

Dann ist

$$\widetilde{e}(t + h; t, z, h) := \widetilde{x}(t + h; t, z, h) - x(t + h; t, z)$$

der Fehler nach nur einem Schritt, der sog. *lokale Fehler* . Unter Differenzierbarkeitsvoraussetzungen an $f$ gilt für viele Ein- und auch Mehrschrittverfahren eine sog.

**(2.3.1) Lokale asymptotische Fehlerentwicklung**
Mit der Konsistenzordnung $p$ des Verfahrens gelte mit obigen Bezeichnungen

$$\widetilde{e}(t + h; t, z, h) = \varphi(t, z) h^{p+1} + O\left(h^{p+2}\right)\ ,\ h \to 0\ .$$

□

**(2.3.2) Beispiel** (für eine lokale Fehlerentwicklung)
Für das Polygonzugverfahren gilt für $f \in C^2(I \times \mathbb{R}^n)$ die Entwicklung (2.3.1) mit $p = 1$ und

$$\varphi(t,z) := -\frac{1}{2}\left(\frac{\partial f}{\partial t}(t,z) + \frac{\partial f}{\partial x}(t,z)f(t,z)\right).$$

Beweis:

Wegen $f \in C^2$ ist $x(\cdot\,;t.z) \in C^3([t, t_0 + T]$. Taylorentwicklung von $x(\cdot\,;t,z)$ um $t$ liefert

$$x(t+h;t,z) = \underbrace{x(t;t,z) + \frac{h}{1!}\underbrace{x'(t;t,z)}_{=f(t,z)}}_{=\widetilde{x}(t+h;t,z,h)} + \frac{h^2}{2!}x''(t;t,z) + \int_t^{t+h} \frac{(t+h-s)^2}{2!}x^{(3)}(s;t,z)ds$$

$\Longrightarrow$

$$\widetilde{e}(t+h;t,z) = -\frac{h^2}{2!}x''(t;t,z) - \int_t^{t+h} \frac{(t+h-s)^2}{2!}x^{(3)}(s;t,z)ds$$

Wegen

$$x''(t;t,z) = \frac{d}{ds}f(s,x(s;t,z))\bigg|_{s=t} = (f_t + f_x x')(t,z) = (f_t + f_x f)(t,z)$$

folgt die Behauptung unter Beachtung der Abschätzung

$$\left|\int_t^{t+h} \frac{(t+h-s)^2}{2!}x^{(3)}(s)ds\right| \le \frac{h^3}{3!}\|x^{(3)}\|_{\infty,[t,t+h]} = O(h^3)$$

□

Die lokale asymptotische Entwicklung (2.3.1) führt zu einer Schrittweitenanpassung, die auf einer Schätzung des lokalen Konsistenzfehlers beruht.
Achtung: *keine* Abschätzung , *nur* Schätzung. Es gibt immer Einzelfälle, in denen Schätzungen völlig versagen. Gefragt sind also Strategien, die häufig eine gute Schätzung liefern.

**(2.3.3) Fehlerkontrolle**
Gegeben sei $t_j$, $x_j :\equiv x_h(t_j)$. Gesucht ist $h > 0$ mit

$$\| \,\widetilde{x}(t_j+h;t_j,x_j,h) - x(t_j+h;t_j,x_j)\, \| \;\doteq\; tol\ ,$$

wobei *tol* eine gewünschte Fehlertoleranz ist. Wir steuern $h$ also nach dem lokalen Fehler in $(t_j, x_j)$. Da im Allg. durch obige Genauigkeitsforderung zuviele $h$ als 'günstig' akzeptiert werden – ohne es zu sein –, fordert man eine gewisse Genauigkeitsverschärfung

$$\| \,\widetilde{x}(t_j+h;t_j,x_j,h) - x(t_j+h;t_j,x_j)\, \| \;\doteq\; \beta\ tol \quad \text{mit } 0 < \beta < 1\ ,$$

Sei zu $h_0 > 0$ eine Schätzung *est* für $\widetilde{x}(t_j+h_0;t_j,x_j,h_0) - x(t_j+h_0;t_j,x_j)$ gegeben, die genau ist bis auf Fehler höherer Ordnung:

$$\| \ est \ - \ \widetilde{e}(t_j+h_0;t_j,x_j,h_0)\ \| \ = \ O\big({h_0}^{p+2}\big)\ .$$

Dann impliziert (2.3.1) für $h^* = h_0\Big(\frac{\beta\ tol}{\|est\|}\Big)^{1/(p+1)}$

$$\| \ \widetilde{e}(t_j+h^*;t_j,x_j,h^*)\ \| \ = \ \beta\ tol \ + \ O\big(\max(h_0,h^*)^{p+2}\big)\ .$$

□

Die Wirksamkeit einer Schrittweitensteuerung sieht man deutlich an folgendem 'eingeschränkten Drei–Körper–Problem' (Erde–Mond–Raumschiff; vgl. z.B. *Stroud, A. H.*: Numerical quadrature and solution of ordinary differential equations (1974) SS.232 ff, siehe auch Abbildung 2.2.
Wie erhält man in der Schrittweitensteuerung eine

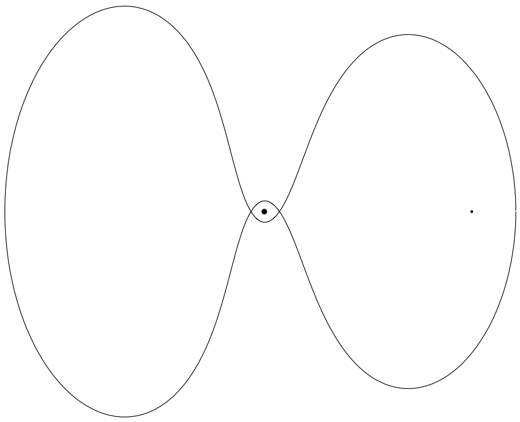

Abbildung 2.2:

Abbildung 2.2: Eingeschränktes Drei–Körper–Problem, berechnet mit dem Dormand/Prince–5(4)–Verfahren. Bei einer Fehlertoleranz $tol = 10^{-12}$ benötigt man *mit Schrittweiten–steuerung: 4563 Schritte* (einschließlich der bei der Fehlerschätzung nicht akzeptierten RK-Schritte). Bildet man – aus Genauigkeitsgründen mit der kleinsten dabei auftretenden Schrittweite – ein *äquidistantes Gitter: 231 620 Schritte.*

$$\text{'}genaue\text{' Schätzung } \; est \stackrel{?}{=} \text{ für den } lokalen\ Fehler\ pro\ Schritt\,?$$

Bei RK–Verfahren liefern *'Eingebettete Formelpaare'* eine *ökonomische Möglichkeit* : Es gibt explizite Runge–Kutta–Formeln der Konsistenzordnung $p = 4$ und der Stufe $s = 6$, die sich zu einer 7–stufigen expliziten Runge–Kutta–Formel der Konsistenzordnung 5 erweitern lassen. Für $\widetilde{x}_1(t + h; ., ., h)$ (Näherung der Ordnung 4) sind dann 6 Auswertungen, und für $x_1(t + h; ., ., h)$ (Näherung der Ordnung 5) nur eine zusätzliche Auswertung von $f$ nötig, also insgesamt 7 Auswertungen von $f$ pro Schritt.
Solche Formeln wurden zuerst von *England* entwickelt (vgl. z.B. *Grigorieff, Bd. 1, S. 20*). Als 'die besten' (zumindest für nichtsteife Probleme) gelten heute die *Dormand/Prince*–Paare (vgl. z.B. *Hairer, Norsett, Wanner [1], S. 171*): für das *5(4)*–Paar ist die verwendete Verfahrensnäherung der Ordnung $p = 5$ $x_1$ ; die Schrittweitensteuerung erfolgt mittels $\widetilde{x}_1$ der Ordnung $\widetilde{p} = 4$ im folgenden Butcher–Array:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0 | | | | | | | |
| 1/5 | 1/5 | | | | | | |
| 3/10 | 3/40 | 9/40 | | | | | |
| 4/5 | 44/45 | −56/15 | 32/9 | | | | |
| 8/9 | 19372/6561 | −25360/2187 | 64448/6561 | −212/729 | | | |
| 1 | 9017/3168 | −355/33 | 46732/5247 | 49/176 | −5103/18656 | | |
| 1 | 35/384 | 0 | 500/1113 | 125/192 | −2187/6784 | 11/84 | |
| $\widetilde{x}_1$ | 35/384 | 0 | 500/1113 | 125/192 | −2187/6784 | 11/84 | 0 |
| $x_1$ | 5179/57600 | 0 | 7571/16695 | 393/640 | −92097/339200 | 187/2100 | 1/40 |

*Dormand–Prince–5(4)–Paar*

Der Stabilitätsbereich dieses Verfahrens ist in Abbildung 2.3 wiedergegeben (wegen $|g(\bar{z})| = |g(z)|$ ist wieder nur der Teil in der oberen Halbebene gezeichnet).

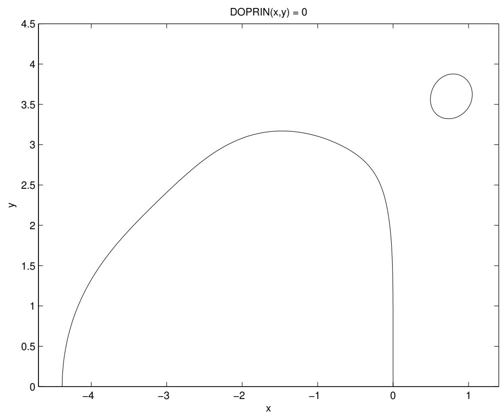


Abbildung 2.3: Stabilitätsbereich (in der oberen Halbebene) des Dormand–Prince–5(4)–Verfahrens

Solche 'eingebetteten' Formeln stellen eine sowohl ökonomische als auch (relativ) sichere Methode dar: *ökonomisch* s.o.; *sicher:* denn die verschiedenen Näherungen verwenden diesselben $t$–, $x_h$– und $f$–Werte oder einen Teil davon. Sie stützen sich *nicht auf völlig unterschiedliche Sätze von Näherungswerten.*

Zusammenfassung von Kapitel 2:

*Vorteile von Einschrittverfahren, insbesondere Runge–Kutta–Verfahren :*

(i) Startwerte sind kein Problem ($x_{0,h} = rd(x_0)$ trivial – nur von Bedeutung im Vergleich zu Mehrschrittverfahren)

(ii) Schrittweitenänderung wie beschrieben leicht durchführbar; die Erhöhung oder Erniedrigung der Ordnung ist ebenfalls leicht zu bewerkstelligen, z.B. durch Extrapolation.

(iii) sehr stabil bei geeigneter Auswahl des Verfahrens, nämlich B-stabil – selbst für bestimmte nichtlineare rechte Seiten.

*Nachteile von Einschrittverfahren, insbesondere Runge–Kutta–Verfahren :*

Sehr aufwendig bei höherer Konsistenzordnung; z.B. erfordern Runge–Kutta–Verfahren höherer Ordnung – und damit notwendig höherer Stufe – *mehrere* Auswertungen von $f$ bei expliziten Verfahren bzw. die Lösung 'großer', i.allg. nichtlinearer Gleichungssysteme bei impliziten Verfahren.

Diese Nachteile vermeiden *Mehrschrittverfahren* , die ich in den Übungen besprechen möchte.

# Kapitel 3

# Anwendung auf Randwertaufgaben: Mehrfach–Schießverfahren

**(3.1) Randwertaufgabe bei gwöhnlichen Differentialgleichungen**
Gegeben

$$f : I \times \mathbb{R}^n \longrightarrow \mathbb{R}^n \quad ; \quad I = [a,b] \subset \mathbb{R} \quad , \quad a < b \quad ;$$
$$r : \mathbb{R}^{2n} \longrightarrow \mathbb{R}^n$$

Gesucht ist $x \in C^1(I)$ (wobei Differenzierbarkeit in den Randpunkten einseitig erklärt ist) mit

$$(i) \qquad x'(t) = f(t, x(t)) \quad , \quad t \in I := [a,b] \qquad (Differentialgleichung) \ ,$$

$$(ii) \qquad r(x(a), x(b)) = 0 \qquad (Randbedingung) \qquad .$$

□

Die $n$ freien Parameter in der allgemeinen Lösung des Differentialgleichungssystems werden hier durch $n$ Gleichungen für die Randwerte festgelegt. Genauer heißt diese Aufgabe *Zweipunkt–Randwertaufgabe.* Häufig sind die *Randbedingungen separiert* , d.h.

$$r(x(a), x(b)) = \begin{bmatrix} r_a(x(a)) \\ r_b(x(b)) \end{bmatrix} \ \text{mit} \ \ r_a : \mathbb{R}^n \longrightarrow \mathbb{R}^{n_a} \quad , \quad r_b : \mathbb{R}^n \longrightarrow \mathbb{R}^{n_b} \quad , \quad n_a + n_b = n \ .$$

*Nicht separiert* sind z.B. *periodische* Randbedingungen

$$x_1(a) = x_1(b) \ , \ x_2(a) = x_2(b) \ , \ldots , \ x_n(a) = x_n(b) \ .$$

Randwertaufgaben ergeben sich vielfach in mechanischen Problemen. Dabei sind eigentlich die Randbedingungen gegeben, die Differentialgleichung – die sog. *Eulersche Variationsgleichung* – ergibt sich dann als 'notwendige Bedingung für ein Extremum'.

**(3.2) Beispiele** (für RWA'n)
(a) Bestimme die Radiusfunktion $x(t)$, $0 \le t \le 1$, eines rotationssymmetrischen Körpers mit Achsenlänge 1 und den 'Deckel'radien

$$x(0) = 1 \quad , \quad x(1) = \beta,$$

derart, daß die Oberfläche minimal ist, d.h. man minimiere unter den Nebenbedingungen $x(0) = 1\ ,\ x(1) = \beta$ das Funktional

$$\int_0^1 x(t)\sqrt{1 + x'(t)^2}dt \longrightarrow \min!$$

Dabei ist $\beta > 0$ gegeben.

Notwendige Bedingung für einen Minimumpunkt $x \in C^2\big([0,1]\big)$ ist

$$1 - 2x(t)x''(t) = 0, t \in [0,1] \quad ; \quad x(0) = 1\ ,\ x(1) = \beta\ .$$

Natürlich kann man diese skalare Differentialgleichung 2. Ordnung in zwei Differentialgleichungen 1. Ordnung umschreiben: $x_1 := x\ ,\ x_2 := x'$ .

$\diamond$

(b) *'lineare Randwertaufgabe 2. Ordnung'* : finde $x : I = [a,b] \longrightarrow \mathbb{R} \quad , \quad x \in C^2(I)$ mit

$$x''(t) + a_1(t)x'(t) + a_0(t)x(t) = c(t) \quad , \quad t \in I$$

$$x(a) = \alpha \quad , \quad x(b) = \beta\ .$$

$\diamond$

(c) analog (b), jedoch nichtlineare Differentialgleichung mit

$$x''(t) = f(t, x(t), x'(t)) \quad , \quad t \in I \quad ; \quad f : I \times \mathbb{R}^2 \longrightarrow \mathbb{R}\ .$$

$\square$

Anders jedoch als bei Anfangswertaufgaben ist die Existenz von Lösungen nicht einmal bei linearen Differentialgleichungen gesichert. Denn die allgemeine Lösung von

$$x'' = -x \quad \text{ist} \quad x(t) = c_1 \sin(t) + c_2 \cos(t)$$

und die Randbedingungen

$$x(0) = 1 \quad , \quad x(\pi) = 0 \quad \Longrightarrow \quad 1 = c_2 = 0\ .$$

$\square$

Für lineare Differentialgleichungen kann man zumindest die Randbedingungen charakterisieren, für die genau eine Lösung existiert, sog. *'Alternative'*.

Im folgenden *setze* ich immer die *Existenz und Eindeutigkeit* einer Lösung der Randwertaufgabe (3.1) *voraus* .

*Prinzipielle Methoden* zur Bestimmung von Näherungen für Lösungen von Randwertaufgaben sind die sog.

(i) *Differenzenverfahren :*
analog zu Anfangswertaufgaben wählt man ein Gitter; Differenzenquotienten-Approximation für $x'$ ergibt auf dem Gitter Differenzengleichungen zusammen mit den Randbedingungen – wie etwa die Diskretisierung der Ortsvariablen $s$ in (2.2.8).

(ii) *Projektions- und Variationsmethoden :*
geeignet z.B. für Aufgaben vom Typ (3.2)(a) : dabei bestimmt man das Minimum auf einem endlichdimensionalen Teilraum von Funktionen; diese Verfahren tragen die Namen von *Galerkin* und *Ritz* . Ich werde diese Verfahrensklasse im folgenden Teil II über partielle Differentialgleichungen behandeln.

Bei gewöhnlichen Differentialgleichungen bei weitem die genauesten Methoden sind

(iii) *Schießverfahren* ( mit sehr trickreich ausgearbeiteten Varianten ) – *prinzipiell* :
Rate in (3.1)$(ii)$ $\quad x(a) = \zeta$;
löse die zugehörige Anfangswertaufgabe mit Lösung $x = x(\cdot\,; a, \zeta)$. Setze $x(b)$ in $r$ ein mit dem Ziel $\quad r(\zeta, x(b; a, \zeta)) \overset{!}{=} 0$ .

Man erhält formal folgenden prinzipiellen Algorithmus

**(3.3) Einfachschießverfahren zur Lösung von (3.1)**

(a) Gegeben $\zeta_{\text{alt}} \in \mathbb{R}^n$ $\quad$ (Ausgangsnäherung für $x(a)$)

(b) Bestimme Lösung $\widetilde{x}$ der Anfangswertaufgabe

$$x'(t) = f(t, x(t)) \ , \ t \in [a, b] \quad ; \quad x(a) = \zeta_{\text{alt}} \ .$$

(c) Falls $r(\widetilde{x}(a), \widetilde{x}(b)) \neq 0$, verbessere $\zeta_{\text{alt}}$ zu $\zeta_{\text{neu}}$; wiederhole (b) mit $\zeta_{\text{alt}} := \zeta_{\text{neu}}$.

□

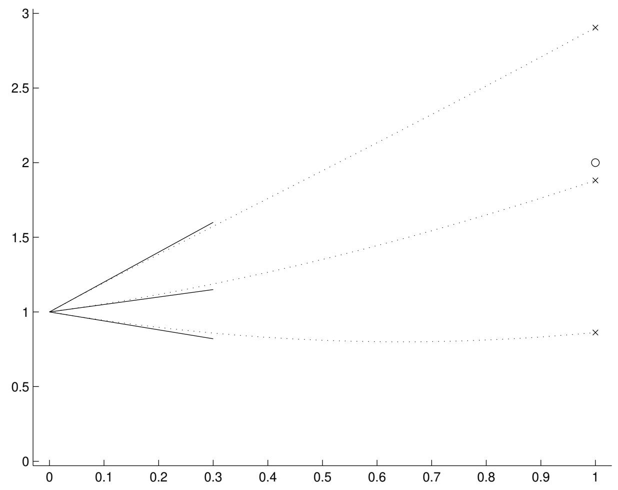


Abbildung 3.1: Veranschaulichung des Schießverfahrens (3.3) für $n = 1$.

Natürlich könnte man auch im rechten Randpunkt den Wert vorgeben bzw. anpassen. Ob die eine oder die andere Möglichkeit günstiger ist, hängt davon ab, ob die Lösungen der Differentialgleichung bzw. ihre numerischen Näherungslösungen $x_h$ für wachsendes oder fallendes $t$ 'Dämpfungs'eigenschaften besitzen.

Daß ein kleiner Fehler in $x(a)$ sich immens vervielfachen kann selbst bei exakter Lösung der Anfangswertaufgabe (3.3)(b), sieht man an

**(3.4) Beispiel** (*Stoer-Bulirsch II, p.177*)
Sei $n = 2$, $[a,b] = [0,10]$;

$$x' = \begin{bmatrix} 0 & 1 \\ 110 & 1 \end{bmatrix} x \quad , \quad x_1(0) = 1 \quad , \quad x_1(10) = 1$$

Sei $x$ die Lösung dieser Randwertaufgabe, $\widetilde{x}$ die Lösung derselben Differentialgleichung mit den gestörten Anfangswerten

$$\widetilde{x}_1(0) = x_1(0) \quad ; \quad \widetilde{x}_2(0) = x_2(0)(1 - 10^{-10}) \ .$$

Dann gilt

$$\widetilde{x}_1(10) \doteq 2.8 \cdot 10^{37} \quad \text{(im Vergleich zu } x_1(10) = 1\text{)}$$

□

Die Erklärung dafür ist natürlich, daß hier der maximale Verstärkungsfaktor bezüglich Störungen in den Daten in etwa angenommen wird.

Abhilfe ist: '*integriere* die (numerische) Lösung *nur über kleine Intervalle*'. Man unterteilt das Intervall in 'kleine' Teilintervalle
$a = \tau_1 < \tau_2 < \ldots < \tau_m < \tau_{m+1} = b$, und wählt mehrere Startwerte $\zeta_1, \ldots, \zeta_m$. Vgl. auch Skizze:

Dazu Bezeichnungen wie früher: bei gegebenem $f$ sei

(3.5)(*) $x(\cdot\,;\tau,\zeta)$ die Lösung der Anfangswertaufgabe

$$x' = f(t,x) \ , \ \tau \le t \le b \ ; \ x(\tau) = \zeta \ \text{ für } \ (\tau,\zeta) \in I \times \mathbb{R}^n \ ;$$

Formal also

**(3.5) Mehrfachschießverfahren** ( *Mehrzielmethode* , *multiple shooting* )
Zur Zerlegung $a = \tau_1 < \tau_2 < \ldots < \tau_m < \tau_{m+1} = b$ bestimme $\zeta_1, \zeta_2, \ldots, \zeta_m \in \mathbb{R}^n$ derart, daß mit der Bezeichnung (3.5)(*) gilt:

$(i)$ für $1 \le j \le m$ löst $x(\cdot\,;\tau_j,\zeta_j)$

$$\frac{d}{dt} x(t;\tau_j,\zeta_j) = f(t, x(t;\tau_j,\zeta_j)) \ , \ \tau_j \le t < \tau_{j+1} \quad ; \quad x(\tau_j;\tau_j,\zeta_j) = \zeta_j \ .$$

$(ii)$

$$r_j := x(\tau_{j+1}-0;\tau_j,\zeta_j) - \zeta_{j+1} = 0 \quad , \quad 1 \le j \le m-1 \ ;$$

$$r_m := r(\zeta_1, x(\tau_{m+1}-0;\tau_m,\zeta_m)) = 0 \ .$$

□

Dieses Vorgehen ist sinnvoll, denn es gilt

**(3.6) Bemerkung**
Sei $\mathsf{Z} := \begin{bmatrix} \zeta_1 & \cdots & \zeta_m \end{bmatrix} \in \mathbb{R}^{n \cdot m}$ Nullstelle der Abbildung $\mathsf{R}$, wobei $\mathsf{R}(\mathsf{Z}) = \begin{bmatrix} r_1 & \cdots & r_m \end{bmatrix} \in \mathbb{R}^{n \cdot m}$ durch (3.5)(i)-(ii) definiert ist. Dann ist $x$, definiert durch

$$x(t) := x(t;\tau_j,\zeta_j) \ , \ \tau_j \le t < \tau_{j+1} \quad ; \quad 1 \le j \le m-1 \ ; \quad x(t) := x(t;\tau_m,\zeta_m) \ , \ \tau_m \le t \le \tau_{m+1} \ ;$$

Lösung der Randwertaufgabe (3.1).

Beweis:

$(3.5)(ii) \Longrightarrow \quad x \in C^0(I).$
$(3.5)(i) \Longrightarrow \quad x'(t) = f(t, x(t)) \quad , \quad t \in I \quad , \quad t \neq \tau_j \quad , \quad 1 < j \leq m \; ;$
in $\tau_j$ gilt aber: $x'(\tau_j - 0) = f(\tau_j, x(\tau_j - 0; \tau_{j-1}, \zeta_{j-1}))$
$= f(\tau_j, \zeta_j) \quad$ , da $x$ stetig
$= x'(\tau_j + 0) \quad$ nach $(3.5)(i)$ ;
$\Longrightarrow \quad x \in C^1(I)$ .

□

Formal kann man (3.5) natürlich wieder als Einfachschießverfahren für den 'langen' Vektor

$$X(t) = [X_j(t)]_{1 \leq i \leq m} \quad , \quad X_j(t) := x(\tau_j + t(\tau_{j+1} - \tau_j)) \quad , \quad 0 \leq t \leq 1 \; ;$$
$$\mathsf{R}(X(0), X(1)) \quad \text{entsprechend (3.5)(ii) definiert ;}$$

auffassen.

Numerisch durchführbar ist nur eine diskretisierte Version von (3.5).

**(3.7) Diskretisiertes Mehrfachschießverfahren**
Sei Zerlegung $a = \tau_1 < \tau_2 < \ldots < \tau_{m+1} = b$ von $[a, b]$, und $(A_h)$ auf zugehörigem Gitter $I_{j,h}$ von $I_j = [\tau_j, \tau_{j+1}]$ gegeben, $1 \leq j \leq m$.
Gesucht sind $\zeta_j \in \mathbb{R}^n$ , $1 \leq j \leq m$ , derart, daß

(i) $x_h(\cdot\, ; \tau_j, \zeta_j)$ löst $(A_h)$ auf $I_{j,h}$ mit dem Anfangswert $x_h(\tau_j; \tau_j, \zeta_j) = \zeta_j \quad , \quad 1 \leq j \leq m.$

(ii) $\begin{bmatrix} r_1 & \cdots & r_m \end{bmatrix} = 0$ , wobei $r_j := x_h(\tau_{j+1} - 0; \tau_j, \zeta_j) - \zeta_{j+1} \quad , \quad 1 \leq j \leq m-1$ ;
$r_m := r(\zeta_1, x_h(b - 0; \tau_m, \zeta_m))$ .

□

Zur iterativen Bestimmung dieser $(\zeta_j : 1 \leq j \leq m) \in \mathbb{R}^{n \cdot m}$ verwendet man beim diskretisierten Mehrfachschießverfahren z.B. das *Broydensche Sekanten-Verfahren* (vgl. *Num. I*). Voraussetzung zur *Anwendung von Newton–Varianten* zur Bestimmung der Nullstelle $\mathsf{Z}$ der Abbildung $\mathsf{R}$ in Bemerkung (3.6) ist, daß $\mathsf{R} \in C^2$ gilt. Dazu

**(3.8) Satz**
Sei $I = [a, b]$ , $f : I \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ aus $C^k$ mit $k \geq 1$, und existiere zu jedem $(\tau, \alpha) \in I \times \mathbb{R}^n$ die (wegen $f \in C^1$ notwendig lokal) eindeutige Lösung $x(\cdot; \tau, \alpha)$ der Anfangswertaufgabe

$$x'(t) = f(t, x(t)) \quad , \quad t \in ]\tau, b[ \quad ,$$
$$x(\tau) = \alpha.$$

Ist zusätzlich $r \in C^k$, so ist die Abbildung $\mathsf{R}$ in Bemerkung (3.6) aus $C^k$.

Beweis:

Nach dem Satz über die stetig differenzierbare Abhängigkeit der Lösung vom Anfangswert (Beweis z.B. *Knobloch, Kappel pp.125-128*) folgt

$$\zeta_j \quad \longmapsto \quad x(t; \tau_j, \zeta_j)\Big|_{t\, =\, \tau_{j+1}} \quad \in C^k \quad , \quad 1 \leq j \leq m \; .$$

und damit

$$r_j \in C^k \quad , \quad 1 \leq j \leq m-1 \; .$$

Da nach Voraussetzung $r \in C^k$ ist, ist auch $r_m \in C^k$ als Komposition zweier $C^k$–Abbildungen.

□

Natürlich sind auch alle Varianten des Newtonverfahrens in (3.6) anwendbar, denn:

$$\mathsf{Z} \longmapsto \mathsf{R}(\mathsf{Z}) \in C^2 \text{ nach Satz (3.8) für } f, r \in C^2 \ .$$

' Existenz von $\mathsf{Z}^*$ ' mit $\mathsf{R}(\mathsf{Z}^*) = 0$ nicht so pauschal beantwortbar.
' $\mathsf{R}'(\mathsf{Z}^*)$ nichtsingulär ? ' nicht so pauschal beantwortbar. Beachte dazu als Behandlungsansatz jedoch:

$$\frac{\partial}{\partial \alpha_k} x(t;\tau,\alpha) =: y(t)$$

löst die Anfangswertaufgabe

$$y'(t) = \frac{\partial f}{\partial x}(t, x(t;\tau,\alpha))y(t) \ , \ t \geq \tau \quad ; \quad y(\tau) = e_k \ .$$

Dabei ist $e_k \in \mathbb{R}^n$ der $k$–te Einheitsvektor.

□

Der *Vorteil* der sehr schnellen *quadratischen (zumindest superlinearen) Konvergenz* der Newton(ähnlichen)–Iteration zur Bestimmung einer Nullstelle von (3.6), kombiniert mit der sehr *hohen Genauigkeit* bei Verwendung von Verfahren $A_h$ hoher Ordnung in (3.7) ergibt ein *schnell konvergierendes sehr genaues Verfahren.*
*Der Nachteil:* nur *'lokale' Konvergenz* , d.h. die Startnäherung für die Nullstelle muß – in nur schwer kontrollierbarer Weise – hinreichend nah an der Nullstelle liegen. In 'realen', 'konkreten', 'interdisziplinären' Anwendungen ist hier der Mathematiker auf die Hinweise dieser Anwender angewiesen.
Andererseits ist die Vorgehensweise dieses Kapitels allgemeiner anwendbar (immer mit diesem letzten großen Nachteil):

**(3.9) 'allgemeine nichtlineare' Eigenwertaufgabe**
Gegeben

$$f : I \times \mathbb{R}^n \times \mathbb{R} \longrightarrow \mathbb{R}^n \quad , \quad \widetilde{r} : \mathbb{R}^{2n} \times \mathbb{R} \longrightarrow \mathbb{R}^n \times \mathbb{R}$$

Gesucht $(x, \lambda) \in C^1(I; \mathbb{R}^n) \times \mathbb{R}$ mit

$$x'(t) = f(t, x(t), \lambda) \quad , \quad t \in I \quad \textit{(Differentialgleichung)}$$

$$\widetilde{r}(x(a), x(b), \lambda) = 0 \quad , \quad \text{wobei}$$

$$\widetilde{r}(x(a), x(b), \lambda) := \begin{bmatrix} r(x(a), x(b), \lambda) \\ \int_I x(t)^2 dt - 1 \end{bmatrix} \quad \begin{array}{l} \textit{(Randbedingungen: häufig von } \lambda \textit{ unabh.)} \\ \textit{(Normierung, damit } x \not\equiv 0) \end{array}$$

□

## Literatur zu Randwertaufgaben bei gewöhnlichen Differentialgleichungen

*Theorie:*

W. Walter [1972]: Gewöhnliche Differentialgleichungen. Eine Einführung. Springer-Verlag, Berlin - Heidelberg - New York. Heidelberger Taschenbücher Band 110

*Numerische Behandlung:*

M. Hanke-Bourgeois [2002]: Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens. Teubner-Verlag, Stuttgart. ISBN 3-519-00356-2. Euro 64,90.

C. Johnson [1995]: Numerical solution of partial differential equations by the finite element method. $6^{th}$ printing 1995.

A. Quarteroni & R. Sacco & F. Saleri [2002]: Numerische Mathematik 2. Springer-Verlag, Berlin - Heidelberg - New York. OSBN 3-540-43616-2. 29.95 Euro

J. Stoer & R. Bulirsch [1990]: Numerische Mathematik II. (3. Auflage) Springer-Verlag, Berlin - Heidelberg - New York.

H.R. Schwarz [1988]: Numerische Mathematik. (2.Auflage) Teubner-Verlag, Stuttgart u.a.

# Teil II

# Partielle Differentialgleichungen

# Kapitel 4

# Einführung in partielle Differentialgleichungen

Werden Funktionen von $m$ Variablen, $m \geq 2$, bestimmt durch Abhängigkeiten zwischen ihren partiellen Ableitungen, spricht man von *partiellen Differentialgleichungen.* Die *Ordnung* einer *partiellen Differentialgleichung* ist die Ordnung der höchsten vorkommenden Ableitung. Einzelgleichungen 1. Ordnung kann man auf gewöhnliche Differentialgleichungen zurückführen (bekannt bzw. s. Übungen).

*Systeme* 1. Ordnung und Einzelgleichungen 2. und höherer Ordnung sind 'schwieriger'. Bereits lineare *partielle Differentialgleichungen* 2. Ordnung liefern eine Fülle von Problemen:

$$Lu(x) = \underbrace{\sum_{i,j=1}^{m} a_{ij}(x)\frac{\partial^2 u}{\partial x_i \partial x_j}(x)}_{=:\, Au(x) \text{ sog. Hauptteil}} + \sum_{j=1}^{m} a_j(x)\frac{\partial u}{\partial x_j}(x) + a_0(x)u(x) \overset{!}{=} f(x) \quad , \quad x \in G \subset \mathbb{R}^m$$

*Gegeben:* $a_{ij}, a_j, f \ : G \longrightarrow \mathbb{R}$ ;

*Gesucht:* $u \ : G \longrightarrow \mathbb{R}$ , $u \in C^2(G)$ als Lösung obiger Gleichung!

Im Prinzip kann dabei $[a_{ij}]_{1\leq i,j\leq m}$ symmetrisch angenommen werden, denn

$$\widetilde{a}_{ij} := \frac{1}{2}(a_{ij} + a_{ji}) \quad , \quad [\widetilde{a}_{ij}] \quad \text{symmetrisch}$$

und

$$\sum \widetilde{a}_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j} = \sum a_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j} \quad , \quad \text{da} \quad \frac{\partial^2 u}{\partial x_i \partial x_j} = \frac{\partial^2 u}{\partial x_j \partial x_i} \quad \text{für} \quad u \in C^2$$

Abhängig von den Eigenwerten der symmetrischen Matrix $[a_{ij}]$ erhält man folgende

*Typeneinteilung:*

$A$ heißt im Punkt $x \in G$

(i) *elliptisch*, wenn die Eigenwerte von $[a_{ij}(x)]_{1\leq i,j\leq m}$ einerlei Vorzeichen haben, o.E. $< 0$

(ii) *parabolisch*, wenn $(m-1)$ Eigenwerte von $[a_{ij}(x)]_{1\leq i,j\leq m} < 0$, ein Eigenwert $= 0$

(iii) *hyperbolisch*, wenn $(m-1)$ Eigenwerte von $[a_{ij}(x)]_{1\leq i,j\leq m} < 0$, ein Eigenwert $> 0$

**(4.1) Standardbeispiele**

$(a)$ $\quad - \triangle u(x) = -\sum_{i=1}^{m} \frac{\partial^2 u}{\partial {x_i}^2}(x) \quad , \quad [a_{ij}] = -E$

mit $E$ Einheitsmatrix; sog. (negativer) *Laplace-Operator.*

bekanntlich gilt: ist $g(x+iy) = u(x,y) + iv(x,y)$ holomorph (mit $(x,y)$ statt $(x_1,x_2)$)
$\Longrightarrow u_x + v_y = 0$ , $v_x - u_y = 0$ Cauchy-Riemannsche Diff.-Gln.
$\Longrightarrow u_{xx} + u_{yy} = 0$ , *i.e.* $\triangle u = 0$ und $\triangle v = 0$ ;
d.h. Real- und Imaginärteil holomorpher Funktionen sind sog. *'harmonische Funktionen'.*

(b) *Wärmeleitung* in einem Körper $G \subset \mathbb{R}^3$:

$$\frac{\partial u}{\partial t} - k\left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \frac{\partial^2 u}{\partial x_3^2}\right) = f(x,t) \ , \ x \in G \subset \mathbb{R}^3 \ , \ t > 0 \ ; \ k > 0 \ :$$

$$[a_{ij}] = \begin{bmatrix} -k & & & 0 \\ & -k & & \\ & & -k & \\ 0 & & & 0 \end{bmatrix}$$

(c) *Wellengleichung*

$$\frac{\partial^2 u}{\partial t^2} - c^2\left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}\right) = f(x,t) \ , \ x \in G \subset \mathbb{R}^2 \ , \ t > 0; \ c > 0 \ :$$

$$[a_{ij}] = \begin{bmatrix} -c^2 & & 0 \\ & -c^2 & \\ 0 & & 1 \end{bmatrix}$$

(bzw. 'schwingende Saite', falls nur eine Raumvariable $x \in \mathbb{R}^1$).

(d) von gemischtem Typ ist z.B.

$$x_2 \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = 0 \ , \ (x_1, x_2) \in \mathbb{R}^2 \ .$$

□

In obigen Beispielen (4.1) $(b)$ und $(c)$ wird die $m$–te$\equiv$ $(d+1)$–te Variable in der Bezeichnung ausgezeichnet:

$$\text{Zeit } t \equiv x_{d+1} \ \text{ im Gegensatz zum 'Ort' } \ x = (x_1, ..., x_d) \in \mathbb{R}^d \ .$$

Die Bezeichnungen 'elliptisch', 'parabolisch', 'hyperbolisch' sind Analogien zu den quadratischen Gleichungen in zwei Variablen.

Durch die Vorgabe von Rand- oder Anfangswerten werden eindeutige Lösungen festgelegt. Für die einzelnen Typen sind unterschiedliche Vorgaben
*sachgemäß:*, i.e. – Existenz einer Lösung
– Eindeutigkeit der Lösung
– stetige Abhängigkeit der Lösung von den Daten .

**(4.2) Anfangswertaufgabe für die eindimensionale Wellengleichung**

$$u_{tt} - c^2 u_{xx} = 0 \ , \ x \in \mathbb{R} \ , \ t > 0;$$

$$u(x,0) = u_0(x) \ , \ u_t(x,0) = u_1(x) \ , \ x \in \mathbb{R} \ .$$

Für $u_0, u_1 \in C^2(\mathbb{R})$ existiert genau eine Lösung obiger Anfangswertaufgabe.

Beweisskizze:

Allgemeine Lösung von $u_{tt} - c^2 u_{xx} = 0$ lautet

$$u(x,t) = \varphi(x - ct) + \psi(x + ct) \ , \ \varphi, \psi \in C^2 \ , \text{ denn:}$$

*u Lösung:* $u_x = \varphi' + \psi' \ , \ u_{xx} = \varphi'' + \psi'' \ ,$
$u_t = -c\varphi' + c\psi' \ , \ u_{tt} = (-c)^2 \varphi'' + c^2 \psi'' \ .$

*Eindeutigkeit:* über sog. 2. Normalform der Wellengleichung: $w_{xt} = 0$.

*Annahme der Anfangswerte:* dies liefert die folgende Linearkombination von Funktionen obiger Bauart $\varphi$ bzw. $\psi$

$$\frac{1}{2}\big(u_0(x + ct) + u_0(x - ct)\big) + \frac{1}{2c}\int_{x-ct}^{x+ct} u_1(s)ds.$$

□

**Veranschaulichung** ( obiger Lösung für $u(x,t) = u_0(x - ct)$)

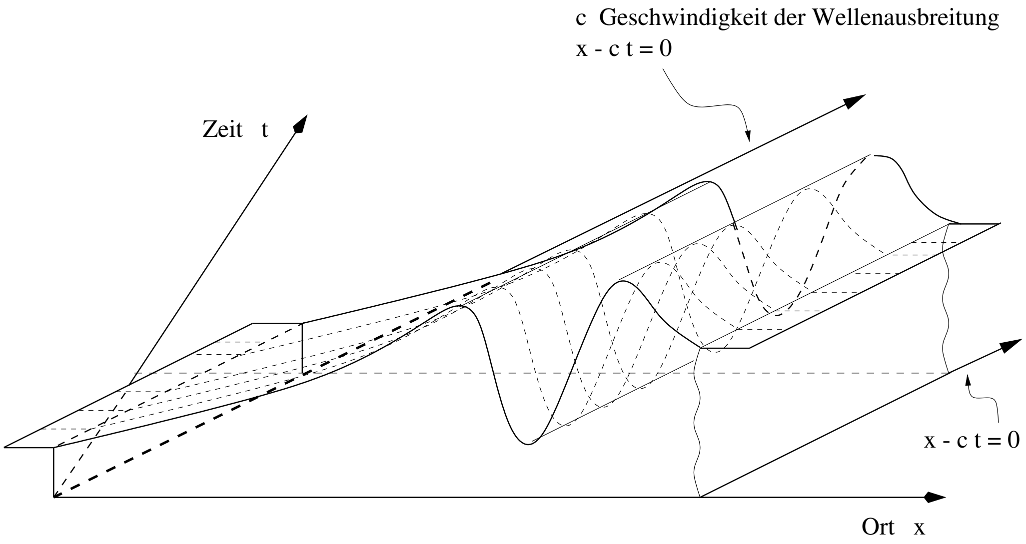


Abbildung 4.1: Wellenausbreitung

Kompliziertere hyperbolische Gleichungen sind nur numerisch lösbar: $\succ$ Definitionsbereich muß beschränkt werden, i.e. Anfangs–Randwertaufgabe auf Kompaktum $\succ$ unerwünschte Reflexionen an den (künstlich eingeführten) Rändern – *Abhilfe:* $\succ$ geeignete sog. *transparente* oder *absorbierende* Randbedingungen derart, daß keine Reflexion auftritt.

Für elliptische Gleichungen sind Randwertaufgaben sachgemäß gestellt.

**(4.3) Randwertaufgaben für elliptische Gleichungen**

$$-\triangle u(x) = f(x)\ ,\ x \in G\ ;$$

mit den sog. 'Dirichlet'-Randbedingungen

$$u(x) = g(x)\ ,\ x \in Rd\ G\ ,\quad ('homogen'\ ,\ \text{falls}\ g(x) \equiv 0)\ ;$$

oder mit den sog. 'Neumann'-Randbedingungen

$$\frac{\partial u}{\partial n}(x) = g(x)\ ,\ x \in Rd\ G\ ,\quad ('homogen'\ ,\ \text{falls}\ g(x) \equiv 0)\ ;$$

bzw. mit den allgemeinen 'Robin'-Randbedingungen

$$u(x) + \sigma(x)\frac{\partial u}{\partial n}(x) = g(x)\ ,\ x \in Rd\ G\ ,\quad ('homogen'\ ,\ \text{falls}\ g(x) \equiv 0)\ .$$

□

Die mathematische Modellierung des sog. *Konstanzer Wasserwunders*, von dem der Stadtschreiber im Jahr 1549 berichtet[1], führt mit dem Ansatz (in komplexer Form) für eine 'stehende Welle'

$$w(x,t) = e^{i\omega t}u(x)$$

auf ein Rand–Eigenwertproblem für $u$,

$$-\triangle u(x) = \lambda u(x)\ ,\ x \in G\quad ,\quad \frac{\partial u}{\partial n}(x) = 0\ ,\ x \in Rd\ G\quad ;\quad \lambda = \omega^2/c^2 > 0\ .$$

Wie kann man (4.3) diskretisieren?

Ist $G = [0,l] \times [0,l]$ ein Quadrat im $\mathbb{R}^2$ mit Kantenlänge $l$, dann ist die naheliegendste Diskretisierung für $-\triangle$ ein *Differenzenverfahren:*

(4.4)(i) *Gitter:*

$$\begin{aligned}
&h = l/(N+1) > 0\ ,\ N \in \mathbb{N}\ ;\\
&\overline{G}_h = \{(ih, jh) \mid i,j \in \mathbb{N}_0\ :\ 0 \le i,j \le N+1\}\ ;\\
&G_h := \overline{G}_h \cap G = \{(ih, jh) \mid i,j \in \mathbb{N}_0\ :\ 1 \le i,j \le N\}\ ;\\
&Rd_h\ G_h := \overline{G}_h \cap Rd\ G = \{(ih, jh) \mid i,j \in \mathbb{N}_0\ :\ i = 0, N+1\ \text{oder}\ j = 0, N+1\}\ ;\ .
\end{aligned}$$

Verwendung des dividierten Differenzenquotienten 2. Ordnung

$$\frac{\partial^2 u}{\partial {x_1}^2}(ih, jh) \doteq \frac{u((i+1)h, jh) - 2u(ih, jh) + u((i-1)h, jh)}{h^2}$$

und analog

$$\frac{\partial^2 u}{\partial {x_2}^2}(ih, jh) \doteq \frac{u(ih, (j+1)h) - 2u(ih, jh) + u(ih, (j-1)h)}{h^2}$$

ergibt insgesamt

$$(4.4)(ii)\qquad -\triangle_h\ u(x) := \frac{-u(x_N) - u(x_E) - u(x_S) - u(x_W) + 4u(x)}{h^2}\ ,$$

wobei die suggestiven Bezeichnungen gerade den sog. *Fünf–Punkt–Differenzenstern* festlegen:

---

[1] zitiert nach Wittum, G.: Mehrgitterverfahren. Spektrum der Wissenschaften, April 1990, SS. 78-90: die Näherung der 10–ten Eigenfunktion hat ein Maximum in der Konstanzer Bucht, die Schwingungsdauer der zugehörigen Amplitudenfunktion beträgt etwa 12 Minuten, was dem Eintrag des Stadschreibers „ ... und das Wasser ist an– und abgelaufen vier– bis fünfmal in der Stund ... “ recht gut entspricht.

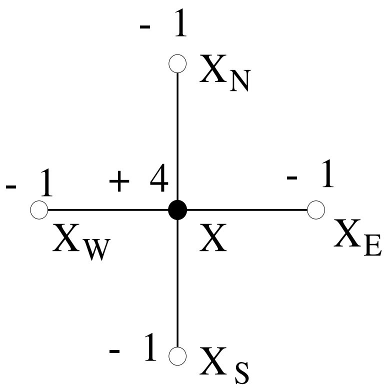


Abbildung 6.2: Fünf–Punkt–Differenzenstern für $-\triangle_h$ im Punkt $x$: angegeben sind die Gewichte für die Funktionswerte (ohne den zusätzlichen Faktor $1/h^2$, wobei $h$ die äquidistante Gitterbreite ist).

**(4.4) Differenzenverfahren** ( für (4.3))
Finde $u_h : \overline{G}_h \longrightarrow \mathbb{R}$ mit

$$-\triangle_h u_h(x) = f(x) \ , \ x \in G_h \ ; \qquad u_h(x) = g(x) \ , \ x \in Rd_h \ G_h \ .$$

Dabei ist $\overline{G}_h$ das Gitter (4.4)(i) und $-\triangle_h$ der Fünf–Punkt–Differenzen–Operator (4.4)(ii).

□

Vor einer Diskussion der Vor– und Nachteile von (4.4) eine hier sehr einfache erste *Konvergenzanalyse.*
Setzt man die Lösung $u$ von (4.3) in (4.4) ein, erhält man den

*Konsistenzfehler:*
$$(-\triangle_h u - f)|_{G_h} = ? \ ;$$
$$(u - g)|_{Rd_h \ G_h} = ? \ .$$

Für glatte Lösungen $u$ erhält man durch Taylorentwicklung

**(4.5) Bemerkung** (Konsistenzordnung des Fünf–Punkt–Differenzenoperators)
Sei die Lösung $u$ von (4.3) aus $C^4(\overline{G})$. Dann gilt für den Fünf–Punkt–Differenzenoperator $-\triangle_h$

$$\max_{x \in G_h} |-\triangle_h u(x) - f(x)| = O(h^2) \ .$$

Beweis:

als Übung – siehe auch *Num. I.*

□

*Konvergenzanalyse: Fehler* $e_h(x) \equiv u_h(x) - u(x) \ = ?$

Die Diskretisierung heißt *stabil* , wenn man von $-\triangle_h(e_h)$ Lipschitzstetig auf $e_h$ schließen kann, gleichmäßig in $h \in \mathcal{H}$. Wegen

$$\begin{aligned} -\triangle_h e_h(x) &= -\triangle_h u_h(x) + \triangle_h u(x) \\ &= f(x) + \triangle_h u(x) \\ &\equiv -\ Konsistenzfehler\ im\ inneren\ Punkt\ x \qquad \text{( klein nach (4.5) ),} \end{aligned}$$

erhält man damit eine Fehlerabschätzung: $\|e_h\|_{\infty,\overline{G}_h} \le L\| - \Delta_h e_h\|_{\infty,G_h} = O(h^2)$ .
Zum Nachweis der Stabilität zeigen wir

**(4.6) Diskretes Maximumprinzip**
Für $v_h : \overline{G}_h \longrightarrow \mathbb{R}$ gilt

$$\max_{x\in\overline{G}_h} |v_h(x)| \;\le\; \max_{x\in Rd_h\ G_h} |v_h(x)| \;+\; \frac{l^2}{2}\max_{x\in G_h} |-\triangle_h v_h(x)| \;.$$

◇

Zum Nachweis dieses diskreten Maximumprinzips zuerst

**(4.7) Hilfssatz**
Für $v_h : \overline{G}_h \longrightarrow \mathbb{R}$ mit $-\triangle_h v_h(x) \le 0$ , $x \in G_h$ , gilt

$$\max_{x\in\overline{G}_h} v_h(x) \;\le\; \max_{x\in Rd_h\ G_h} v_h(x) \;.$$

Beweis:

Durch Widerspruch. Angenommen, für $x \in G_h$ gilt

$$\max_{y\in Rd_h\ G_h} v_h(y) \;<\; v_h(x) \overset{o.E.}{=} \max_{y\in G_h} v_h(y) \;.$$

$$\begin{aligned} -\triangle_h v_h(x) \le 0 \implies v_h(x) \;&\le\; \frac{1}{4}\Big(v_h(x_N) + v_h(x_E) + v_h(x_S) + v_h(x_W)\Big) \\ &\le\; (1/4+1/4+1/4+1/4)\; v_h(x) \text{ , falls alle Nachbarn von } x \text{ in } G_h; \\ &<\; (1/4+1/4+1/4+1/4)\; v_h(x) \text{ , falls ein Nachbar von } x \in Rd\ _h\ G_h. \end{aligned}$$

Sind alle Nachbarn von $x$ in $G_h$ , muß also notwendig gelten:

$$v_h(x) \;=\; v_h(x_N) \;=\; v_h(x_E) \;=\; v_h(x_S) \;=\; v_h(x_W) = \max_{y\in G_h} v_h(y) \;.$$

Wiederholt man diese Schlußweise immer wieder für alle Nachbarn, erhält man schließlich ein $x$, dessen Nachbarn nicht alle in $G_h$ liegen – und damit den Widerspruch '$v_h(x) < v_h(x)$' .

□

Beweis von (4.6):

$\varphi(x_1, x_2) := \frac{1}{2}{x_1}^2 \implies 0 \le \varphi(x) \le \frac{l^2}{2}$ und $\triangle_h\varphi(x) \equiv 1$ ;

betrachte $v_\pm(x) \;:=\; \pm v_h(x) + D\varphi(x)$ , $x \in G_h$ ; mit $D := \max_{x\in G_h} |\triangle_h v_h(x)|$ .

$$\implies -\triangle_h v_\pm(x) \;=\; \mp\, \triangle_h v_h(x) - D \underbrace{\triangle_h\varphi(x)}_{\equiv 1} \;\le\; 0 \text{ , } x \in G_h; \text{ nach Wahl von } D.$$

Nach (4.7) folgt die Behauptung.

□

Aus (4.5) und (4.6) erhält man für den Fünf–Punkt–Differenzen–Stern direkt die Konvergenzordnung 2:

**(4.8) Satz**
Für $\overline{G}_h = [0,l] \times [0,l]$ erhält man für $u_h$ aus (4.4) unter der Voraussetzung, daß die Lösung $u$ von (4.3) aus $C^4(\overline{G})$ ist,

$$\max_{x\in\overline{G}_h} |u_h(x) - u(x)| = O(h^2) \;.$$

□

Eine weitere Diskretisierung von $-\triangle$, für die ebenfalls das diskrete Maximumprinzip (4.6) gilt, ist der *Standard–Neun–Punkt–Differenzenstern* $-\widetilde{\triangle}_h$ (s. Abb. 6.3). Ist die Lösung $u$ von (4.3) aus $C^6(\overline{G})$, so hat dieser Differenzen–Operator die Konsistenzordnung 4, falls man die rechte Seite $f(x)$ (im Fünf–Punkt–Differenzenstern) ersetzt durch einen geeigneten Mittelwert $\widetilde{f}(x) := \frac{1}{12}\big(f(x_N) + f(x_E) + f(x_S) + f(x_W) + 8f(x)\big)$ von Funktionswerten von $f$ um $x$ auf dem Fünf–Punkt–Differenzenstern:

$$-\widetilde{\triangle}_h u(x) \;=\; \widetilde{f}(x)\ ,\ x \in G_h\ .$$

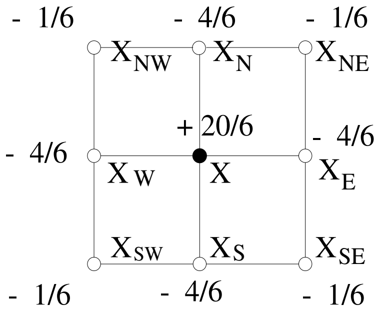


Abbildung 6.3: $-\widetilde{\triangle}_h$ im Punkt $x$; angegeben sind die Gewichte für die Funktionswerte (ohne den zusätzlichen Faktor $1/h^2$, wobei $h$ die äquidistante Gitterbreite ist).

Obige Konvergenzanalyse kann man immer dann anwenden, wenn die Geometrie von $G$ sehr einfach ist, und die Diskretisierung $G_h$ so gewählt werden kann, daß gilt:

$$Rd_h\ G_h\ \subset\ Rd\ G\ ,\ h \in \mathcal{H}\ .$$

Für nicht ganz so regelmäßiges $G$ ergibt sich ein Genauigkeitsverlust!

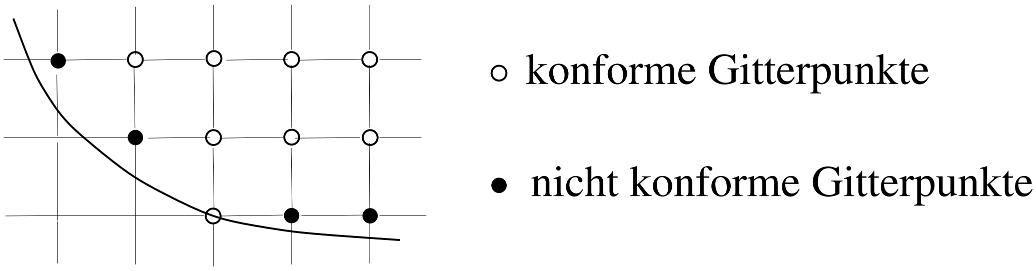


Abbildung 6.4: nicht konforme diskrete Randpunkte

*Zusätzliche Probleme* ergeben sich, wenn die Randbedingungen nicht nur 'einfache' Dirichlet-Daten sind. Wie soll eine Bedingung an die Normalableitung, z.B. die Neumann-Randbedingung, in nichtglatten Randpunkten diskretisiert werden?

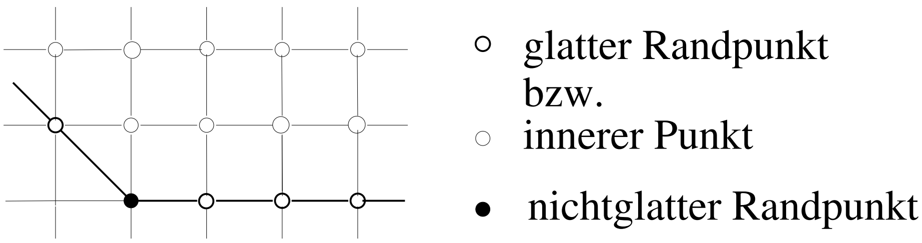


Abbildung 6.5: nichtglatte Randpunkte

Zusammenfassung von Kapitel 4

*Vorteile von Differenzenverfahren :*

(i) konzeptionell einfach,

(ii) in Spezialfällen einfach zu analysieren ( s. (4.8)),

(iii) sehr einfach zu implementieren.

*Nachteile von Differenzenverfahren :*

(i) Lösung muß sehr glatt sein,

(ii) mit 'guter' Konvergenzgüte nur durchführbar bei sehr einfacher Geometrie des Definitionsbereichs $G$,

(iii) nur sehr einfache Randbedingungen sind 'erlaubt'.

Die Vermeidung dieser Nachteile führt zu Projektionsverfahren, insbesondere zu *Finite–Elemente–Verfahren* .

**Literaturhinweise zu *partiellen Differentialgleichungen*:**

*(eher) theoretisch:*

COURANT, R., D. HILBERT:
Methoden der mathematischen Physik. 2. Partielle Differentialgleichungen. Springer, versch. Auflagen.

HELLWIG, G.:
Partielle Differentialgleichungen. Eine Einführung. Teubner 1960.

DiBENEDETTO, E.:
Partial differential equations. Birkhäuser 1995, ISBN 0-8176-3708-7.

STRAUSS, W.A.:
Partielle Differentialgleichungen. Eine Einführung. Vieweg 1995, ISBN 3-528-06604-0.

*(eher) numerisch:*

ich halte mich hier insbesondere an das(die) **fettgedruckte(n)** Zitat(e) – und verwende auch die anderen:

**JOHNSON, C.**:
Numerical Solutions of PDE'S by the finite element method, Cambridge Univ. Pr. 1987

LeVEQUE, R.J.:
Numerical methods for conservation laws. Birkhäuser 1990, ISBN 0-8176-2464-3.

QUARTERONI, A., A. VALLI:
Numerical Approximation of Partial Differential Equations. Springer Series in Computational Mathematics, Vol. 23, Springer 1994, ISBN 3-540-57111-6. (guter Überblick über viele Methoden).

VERFÜHRTH, R.:
A Review of *A Posteriori* Error Estimation and Adaptive Mesh-Refinement Techniques. Wiley-Teubner 1996, ISBN 3-519-02605-8.

Daneben gibt es viele weitere 'Standard'–Werke über FEM. Eine winzige Auswahl von Autoren ist:

BRAESS, D.:
Finite Elemente. Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie. Springer 1997, ISBN: 3-540-61905-4.

BRENNER, S.C.; SCOTT, L.R.:
The mathematical theory of finite element methods.. Springer 1994, ISBN 0-387-94193-2; Texts in applied mathematics ; 15.

**CIARLET, Ph.G.**:
The finite element method for elliptic problems. North-Holland 1978, ISBN 0-444-85028-7. Studies in mathematics and its applications ; 4.

HACKBUSCH, W.:
Theorie und Numerik elliptischer Differentialgleichungen. Teubner 1986, ISBN 3-519-02074-2; Teubner-Studienbücher : Mathematik.

# Kapitel 5

# Finite–Elemente–Verfahren für elliptische Randwertaufgaben

Finite–Elemente–Verfahren sind Projektionsverfahren mit speziellen Ansatzräumen $V_h$, die man oft ebenfalls als Finite–Elemente bezeichnet. In der historischen Entwicklung wurden diese Verfahren zuerst für elliptische Gleichungen entwickelt – und zwar von Ingenieuren. Heute wendet man FE–Methoden auf alle Typen an. Ich beschränke mich auf elliptische Gleichungen, sogar nur auf den negtiven Laplace–Operator $-\triangle$.

## 5.1 Variationsformulierung und Koerzitivitätsungleichungen

**(5.1.1) Aufgabe** (klassisch formuliert)
Sei $G \subset \mathbb{R}^d$ beschränktes Gebiet, und $f : G \longrightarrow \mathbb{R}$, $g : Rd\ G \longrightarrow \mathbb{R}$ gegeben. Gesucht ist $u \in C^2(G) \cap C^0(\overline{G})$ mit

$$(5.1.1)(i) \qquad -\triangle u(x) = f(x)\ ,\ x \in G\ ;$$

$$(5.1.1)(ii) \qquad u(x) = g(x)\ ,\ x \in Rd\ G\ . \qquad \square$$

Für $G \subset \mathbb{R}^2$ bzw. $G \subset \mathbb{R}^d$ kann man 'stückweise stetig differenzierbar' nicht mehr so 'natürlich' definieren wie für $G = [a, b] \subset \mathbb{R}$. Im Hinblick auf die Anwendung des Gaußschen Satzes ist die Zerlegung von $G$ in endlich viele 'Normalgebiete' passend – ein 'Normalgebiet' erlaubt gerade per definitionem die Anwendung des Gaußschen Satzes.

**Bezeichnungen**
(a) $PC^1(G) = \{\ u : G \longrightarrow \mathbb{R} \ |\ $ es existiert Zerlegung $\overline{G} = \bigcup_{i\in I} \overline{G_i}$ , $I$ endlich ; $G_i$ Normalgebiet , $|\overline{G_i} \cap \overline{G_j}| = 0, i \neq j,\ u\big|_{G_i} \in C^1(\overline{G_i}), i \in I\ ;\ u \in C^0(\overline{G})\ \}$ ;

(b) $PC_0^1(G) = \{u \in PC^1(G) \ |\ u\big|_{Rd\ G} = 0\ \}$ .

(c') $PC_{bc}^1(G) = \{u \in PC^1(G) \ |\ u\big|_{Rd\ G} = g\ \}$ zu gegebenem $g : Rd\ G \longrightarrow \mathbb{R}$.

Sind die Randdaten $g$ zu einer Funktion $\widetilde{g} \in PC^1(G)$ fortsetzbar, so gilt

(c) $PC_{bc}^1(G) = \{u \in PC^1(G) \ |\ u - \widetilde{g} \in PC_0^1\}$ . Im weiteren setzen wir (c) voraus, und unterscheiden nicht zwischen $g$ und $\widetilde{g}$.

$\square$

Für $v \in PC_0^1(G)$ und $u \in C^2(\overline{G})$ gilt für die von mir betrachteten $G$

$$\int_G \big( - \triangle u\big)(x)v(x)\ dx = \sum_{i=1}^d \int_G \frac{\partial u}{\partial x_i}(x)\frac{\partial v}{\partial x_i}(x)\ dx$$

Um mathematisch 'angenehme' Räume zu erhalten, nimmt man den Abschluß obiger Räume bzgl.

$$\|u\|_{1,2} := \big(\|u\|_{L_2(G)}{}^2 + \sum_{|\alpha|=1} \|D^\alpha u\|^2_{L_2(G)}\big)^{1/2}$$

und erhält einen Hilbertraum, den *Sobolevraum* $W^{1,2}(G)$. Analog erhält man für $m \geq 1$ und $1 \leq p \leq \infty$ durch Vervollständigung bzgl.

$$\|u\|_{m,p} := \big( \sum_{|\alpha|\leq m} \|D^\alpha u\|^p_{L_p(G)}\big)^{1/p}$$

(mit der üblichen Modifikation für $p = \infty$) den Banachraum $W^{m,p}(G)$. Das Randverhalten wird im verallgemeinerten Sinn ebenfalls durch eine geeignete Approximationsmöglichkeit durch Funktionen mit klassischem Randverhalten beschrieben:

$$W_0^{1,2}(G) \ = \ \overline{C_0^\infty(G)}^{\ \|\cdot\|_{1,2}}$$

sind die $W^{1,2}(G)$–Funktionen, die im verallgemeinerten Sinn auf dem Rand verschwinden. Im folgenden betrachte ich vor allem $p = 2$.

**(5.1.2) Schwache Formulierung** ( von (5.1.1) )
Gegeben $f \in L_2(G)$ und $g \in W^{1,2}(G)$ . Finde $u \in W^{1,2}(G)$ mit

$$(5.1.2)(i) \qquad a(u,v) :\equiv \sum_{i=1}^d \int_G \frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_i}\ dx \ = \ \int_G fvdx \equiv:< f,v > \ , \ v \in W_0^{1,2}(G)\ ;$$

$$(5.1.2)(ii) \qquad u - g \in W_0^{1,2}(G)\ .$$

□

Klar ist, daß jede Lösung $u$ von (5.1.1) auch (5.1.2) löst, wenn $u \in W^{1,2}(G)$ gilt (beachte $C^1(\overline{G}) \subset W^{1,2}(G)$). Zur Existenz einer schwachen Lösung $u$ vergleiche man auch die Übungsaufgaben 8.1 und 8.2.

Wir betrachten (5.1.2) in folgendem abstrakten Rahmen:

$\big(V, <,>_V\big)$ Hilbertraum, $V_0 \subset V$ abgeschlossener Unterraum, $V_{bc} = g + V_0 \subset V$,
$a : V \times V \longrightarrow \mathbb{R}$ bilinear und stetig,
$f : V \longrightarrow \mathbb{R}$ linear und stetig (Schreibweise $f(v) \equiv \ < f, v >$).
Gesucht ist $u \in V_{bc}$ mit

$$a(u,v) =< f,v > \ , \ v \in V_0\ .$$

Konkret haben wir in (5.1.2):

$V = W^{1,2}(G)\ ,\ V_0 = W_0^{1,2}(G)\ ,\ V_{bc} = g + W_0^{1,2}(G)\ ,$

$$a(v,w) :\equiv \sum_{i=1}^{d} \int_G \frac{\partial v}{\partial x_i}\frac{\partial w}{\partial x_i}\, dx \ , \qquad f(v) \equiv < f, v > :\equiv \int_G f(x)v(x)\, dx).$$

Zur Diskretisierung der Aufgabe (5.1.2) betrachten wir einen endlichdimensionalen Teilraum $V_h \subset V = W^{1,2}(G)$. Da u.U. die Randfunktion $g$ nicht exakt darstellbar ist durch Funktionen aus $V_h$, muß man $g$ ersetzen etwa durch die Interpolierende $g_h = \pi_h g \in V_h$, und erhält für die Modifikation

$$V_{h,bc} := g_h + V_{h,0} \ ,$$

daß dann trivialerweise $V_{h,bc} \neq \emptyset$ gilt.

**(5.1.3) Projektionsverfahren** ( für (5.1.2) )
Sei $V_h \subset V$ endlichdimensionaler Teilraum, $V_{h,0} := V_h \cap V_0$ . Finde $u_h \in g_h + V_{h,0}$ mit

$$(5.1.3)(i) \quad a(u_h, v_h) \ = \ < f, v_h > \ , \ v_h \in V_{h,0} \ .$$

□

Eine geeignete Voraussetzung sowohl für die Durchführbarkeit dieses Projektionsverfahrens als auch eine Fehlerabschätzung für $\|u_h - u\|_V$ ist die Koerzitivität:

**(5.1.4) Koerzitive Bilinearformen**
Die Bilinearform $a : V \times V \longrightarrow \mathbb{R}$ heißt *'koerzitiv bzgl.* $V_0$', wenn $a$ stetig ist, d.h. es existiert ein $\overline{c} > 0$ mit $|a(v,w)| \leq \overline{c}\|v\|_V \|w\|_V \ , \ v, w \ \in \ V$ , und wenn ein $\underline{c} > 0$ existiert mit

$$(5.1.4)(i) \quad a(v,v) \geq \underline{c} < v, v >_V \ , \ v \in V_0 \ .$$

□

**(5.1.5) Folgerung**
Gilt (5.1.4)(i), dann gibt es genau ein $u_h \in V_{h,bc}$, das $(5.1.3)(i)$ erfüllt.

Beweis:

Nach Voraussetzung existiert $w_h^{bc} \in V_{h,bc}$. Sei $\{u_1, \ldots, u_N\}$ Basis von $V_{h,0}$.
*Behauptung:* Es existiert genau eine Linearkombination $\sum_{j=1}^N c_j u_j$ derart, daß $u_h := w_h^{bc} + \sum_j c_j u_j$ (5.1.3)(i) löst.
*Beweis:*

$$u_h = w_h^{bc} + \textstyle\sum_j c_j u_j \ \text{ löst } \ (5.1.3)(i)$$

$$\iff \quad u_h - w_h^{bc} = \textstyle\sum c_j u_j \in V_{h,0} \ \text{ und } \ a(\sum c_j u_j, v_h) = \langle f, v_h\rangle - a(w_h^{bc}, v_h) \ , \ v_h \in V_{h,0}$$

$$\iff \quad \textstyle\sum_j c_j a(u_j, u_i) = \langle f, u_i\rangle - a(w_h^{bc}, u_i) \ , \ 1 \leq i \leq N, \text{ denn } \ V_{h,0} \ = \ span(u_1, \ldots, u_N).$$

Dieses lineare Gleichungssystem für $(c_j)$ ist aber eindeutig lösbar, da das zugehörige homogene Gleichungssystem nur die triviale Lösung hat. Denn

$$\textit{Angenommen} \quad \sum_{j=1}^N c_j a(u_j, u_i) = 0 \ , \ 1 \leq i \leq N$$

$$\Longrightarrow \ v_h := \sum_j c_j u_j \ \text{erfüllt} \ \ a(v_h, u_i) = 0 \ , \ 1 \leq i \leq N; \ \Longrightarrow \ 0 = \sum_i c_i a(v_h, u_i) = a(v_h, v_h)$$

$$\overset{(5.1.4)(i)}{\Longrightarrow} \ \|v_h\|_V = 0 \ \Longrightarrow \ v_h = 0 \ \Longrightarrow \ (c_j) = 0 \ , \ \text{da } \{u_j\} \text{ linear unabhängig} \ .$$

□

**(5.1.6) Satz** (Céa's Lemma)
Sei $u$ Lösung von (5.1.2)(i). Unter der Voraussetzung (5.1.4)(i) existiert $c > 0$, so daß für jeden endlichdimensionalen Teilraum $V_h \subset V$ mit $V_h \cap V_{bc} \neq \emptyset$ (d.h. $g_h = g$ in (5.1.3)) gilt: es existiert genau ein $u_h \in V_h \cap V_{bc}$ als Lösung von (5.1.3), und es gilt die Fehlerabschätzung

$$\|u - u_h\|_V \leq c \inf_{v_h \in V_h \cap V_{bc}} \|u - v_h\|_V \ .$$

Beweis:

Wegen (5.1.5) bleibt nur die Fehlerabschätzung zu zeigen. Nach (5.1.4)(i) gilt

$$\begin{aligned}
\|u - u_h\|_V{}^2 &\leq \underline{c}^{-1}\, a(u-u_h, u-u_h)\ , \quad \text{da} \ \ u - u_h \in V_0 \\
&\overset{(*)}{=} \underline{c}^{-1}\, a(u-u_h, u-v_h)\ , \quad v_h \in V_h \cap V_{bc} \\
&= \underline{c}^{-1} \inf_{v_h \in V_h \cap V_{bc}} a(u-u_h, u-v_h) \\
&\leq \underline{c}^{-1}\,\overline{c}\, \|u-u_h\|_V \inf_{v_h \in V_h \cap V_{bc}} \|u - v_h\|_V \quad , \ \text{da } a \ \text{ stetig ist .}
\end{aligned}$$

Division durch $\|u - u_h\|_V$ ($> 0$ o.E.) ergibt die Behauptung.

zu $(*)$ :
$$\begin{aligned}
+\, a(u - u_h, u_h - v_h) &= a(u, \underbrace{u_h - v_h}_{\in V_0}) - a(u_h, \underbrace{u_h - v_h}_{\in V_{h,0}}) \\
&= \langle f, u_h - v_h\rangle - \langle f, u_h - v_h\rangle \quad \text{nach (5.1.2)(i) und (5.1.3)(i)} \\
&= 0\ .
\end{aligned}$$

□

Daß $a$ in (5.1.2) – $V \equiv W^{1,2}(G)$ , $V_0 \equiv W_0^{1,2}(G)$ – koerzitiv ist für homogene Randdaten, d.h. $g \equiv 0$, besagt gerade die

**(5.1.7) Friedrichssche Ungleichung** (Kurt Otto Friedrichs, * 1901 Kiel , † 1982 New Rochelle NY)
Für $G \subset \mathbb{R}^d$ , $G$ beschränkt, existiert $c > 0$, so daß für $v \in W_0^{1,2}(G)$ gilt

$$\sum_{i=1}^{d} \int_G \Big(\frac{\partial v}{\partial x_i}\Big)^2(x)\ dx \ \geq\ c \int_G v(x)^2\ dx\ .$$

Beweis:

Nach Voraussetzung gilt $\overline{G} \subset Q = \{x \in \mathbb{R}^d \ \mid\ |x_i| \leq l, 1 \leq i \leq d\}$. Für $v \in C_0^\infty(G)$ gilt (mit der trivialen Fortsetzung $v\big|_{\mathbb{R}^d \setminus G} \equiv 0$)

$$v(x_1, ..., x_d) = \int_{-l}^{x_1} \frac{\partial v}{\partial x_1}(\xi_1, x_2, .., x_d)\ d\xi_1$$

und damit nach der Schwarzschen Ungleichung

$$|v(x_1, ..., x_d)|^2 \ \leq\ 2l \int_{-l}^{l} |\frac{\partial v}{\partial x_1}(\xi_1, x_2, .., x_d)|^2\ d\xi_1\ .$$

Integration dieser Ungleichung über $Q$ ergibt wegen $v\big|_{\mathbb{R}^d\setminus G} \equiv 0$

$$\begin{aligned}\int_G |v(x)|^2\,dx \;&\leq\; 2l\int_{-l}^{l}\cdots\int_{-l}^{l}\Big(\int_{-l}^{l}|\frac{\partial v}{\partial x_1}(\xi_1,x_2,..,x_d)|^2\,d\xi_1\Big)dx_1...dx_d\;.\\ &=\; 2l\int_{-l}^{l}dx_1\Big(\int_{-l}^{l}\cdots\int_{-l}^{l}|\frac{\partial v}{\partial x_1}(\xi_1,x_2,..,x_d)|^2\,d\xi_1 dx_2...dx_d\Big)\;.\\ &\leq\; (2l)^2\sum_{i=1}^{d}\int_G|\frac{\partial v}{\partial x_i}(x)|^2\,dx\;.\end{aligned}$$

Für beliebiges $v \in W_0^{1,2}(G)$ kann man durch Funktionen aus $C_0^\infty(G)$ gleichzeitig sowohl die linke als auch die rechte Seite der behaupteten Ungleichung beliebig genau approximieren.

□

Damit gilt die Koerzitivitätsungleichung $a(v,v) \geq \underline{c} < v,v >_V$ , $v \in V_0$, für

$$a(v,w) = \sum_{i=1}^{d}\int_G \frac{\partial v}{\partial x_i}(x)\frac{\partial w}{\partial x_i}(x)\,dx \text{ und } V_0 = W_0^{1,2}(G) \subset V = W^{1,2}(G)$$

mit $\underline{c} := 1/\big(1+4l^2\big)$. Für inhomogene Randdaten wurde eine entsprechende Abschätzung von *Henri Poincaré* bewiesen. Dabei benötigt man zusätzliche Voraussetzungen (an die Geometrie des Randes von $G$).

**(5.1.8) Poincarésche Ungleichung** (Jules Henri Poincaré, * 1854 Nancy , † 1912 Paris)
Sei $G$ beschränktes, lokal lipschitzstetig berandetes Gebiet. Dann existiert $c > 0$, so daß gilt

$$\sum_{i=1}^{d}\int_G\Big(\frac{\partial v}{\partial x_i}\Big)^2(x)\,dx \;\geq\; c\int_G v(x)^2\,dx\;, \qquad v \in V_0 := \big\{v \in W^{1,2}(G)\;:\;\int_G v(x)dx = 0\;\big\}.$$

□

Zum Beweis vergleiche man die Literatur. Es reicht auch aus, daß der Rand von $G$ eine gleichmäßige 'Kegelbedingung' erfüllt.

**Bemerkung**
Gilt die Aussage von Satz (5.1.6), so heißt das Verfahren *quasioptimal* bzgl. der $\|\cdot\|_V$-Norm – denn der *Verfahrensfehler* $\|u-u_h\|_V$ ist von derselben Größenordnung wie der *Approximationsfehler* $\inf_{v_h\in V_{h,bc}}\|u-v_h\|_V$ – und besser kann man die (unbekannte) Lösung ja nicht approximieren.

□

Damit ist die Wahl geeigneter Ansatzfunktionen klar vorgezeichnet, um Konvergenz bzw. um eine bestimmte asymptotische Konvergenzordnung zu erzielen: man muß die Wahl so treffen, daß der Approximationsfehler

$$\inf_{v_h\in V_{h,bc}}\|u-v_h\|_V$$

gegen Null konvergiert bzw. darüber hinaus eine bestimmte Konvergenzordnung hat. Solche Räume untersuche ich im nächsten Abschnitt 5.2.

## 5.2 Finite–Elemente–Räume

In diesem Abschnitt führe ich einige Beispiele für Finite Elemente ein und untersuche deren Approximations– und Glattheitseigenschaften.

Zur leichteren Formulierung *und Veranschaulichung* betrachte ich hier

$$(0.i) \quad G \text{ beschränkt, offen, zusammenhängend} \subset \mathbb{R}^d,\ d = 2,$$

und

$$(0.ii) \quad \mathrm{Rd}\, G \text{ Polygonzug.}$$

*Bemerkung:* Lipschitzstetig berandete Gebiete $G$ kann man durch polygonal berandete Gebiete approximieren.

Wir betrachten Zerlegungen von $G$ in 'Elemente', die affine Bilder eines Referenz– oder Standard–Elements sind

**(5.2.1) Triangulierungen**
Gegeben Referenzelement $\widehat{K} :\equiv \widehat{\triangle} = \left\{ x \in \mathbb{R}^d : x_i \geq 0\ ,\ \sum_{i=1}^d x_i \leq 1 \right\}$ , $d = 2$ ; d.h. abgeschlossenes Dreieck mit Ecken (0,0), (1,0), (0,1). Dann heißt $T_h = \{K\}$ 'Triangulierung von $G$', wenn

(i) $K = F_K(\widehat{K})$ mit affinen nichtsingulären Abbildungen $F_K : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ , $K \in T_h$ ;

(ii) $\overline{G} = \bigcup_{K \in T_h} K$ ;

(iii) $$K_1 \cap K_2 = \begin{cases} \emptyset & \\ & \text{oder} \\ \text{Eckpunkt von } K_1 \textit{ und } K_2 & \\ & \text{oder} \\ \text{Dreiecksseite von } K_1 \textit{ und } K_2 & \end{cases} , \quad K_1 \neq K_2 \in T_h \ .$$

□

*Verboten* ist also für eine Triangulierung eine Situation wie in Abbildung 5.2. Aufgrund von (i) heißen die hier betrachteten Finiten Elemente 'affin äquivalente' Familien von Finiten Elementen. Für andere Referenzelemente $\widehat{K}$ sprechen wir von *verallgemeinerten* Triangulierungen oder '$\widehat{K}$'–Zerlegungen von $G$. Das zweite wichtige Referenzelement ist das Einheitsquadrat

$$\widehat{K} = \widehat{Q} = [0,1]^d\ ,\ d = 2\ .$$

**(5.2.2) Finite–Elemente–Räume**
Gegeben sei Triangulierung $T_h = \{K\}$ von $G$. Zu gegebenem $m \in \mathbb{N}$ sei $P \subset C^m(\mathbb{R}^d)$ endlichdimensionaler Teilraum. Dann ist

$$V_h := \{ v : \overline{G} \longrightarrow \mathbb{R}\ :\ v|_{\overset{\circ}{K}} \in P \ \text{ für }\ K \in T_h\ ;\ v \in C^{m-1}(\overline{G})\ \}$$

der Raum der $P$–Elemente zur Triangulierung $T_h$ der verallgemeinerten Differenzierbarkeitsordnung $m$.

□

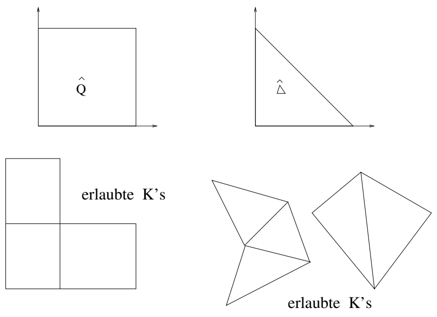


Abbildung 5.1: mögliche Referenzelemente $\widehat{Q}$ und $\widehat{\triangle}$, und erlaubte Triangulierungen

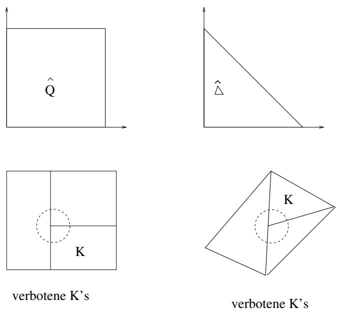


Abbildung 5.2: Nichtkonforme Triangulierung

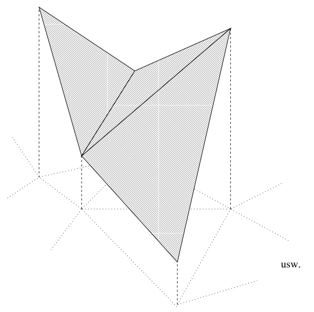


Abbildung 5.3: Aussehen linearer $C^0$–Elemente

Man könnte in naheliegender Weise die elementweise Bauart $P$ der Ansatzfunktionen für jedes $K \in T_h$ unterschiedlich wählen, d.h. $P_K \subset C^m(\mathbb{R}^d), K \in T_h$. Wir beschränken uns in der Darstellung auf eine einzige Bauart.

Polynomiale Finite Elemente: $P \equiv \mathbb{P}_k = span(\{x^\alpha :\equiv \prod_{i=1}^d x_i^{\alpha_i} : \sum_i \alpha_i \equiv: |\alpha| < k\}$ für jedes $\triangle \in T_h$.

Beachte: $\dim \mathbb{P}_k|_{\mathbb{R}^2} = \binom{k+1}{2}$ .

'lineare' Finite Elemente: $P \equiv \mathbb{P}_2 = \operatorname{span}(\{1, x_1, x_2\})$ , $K = \triangle \in T_h$ .
Sowohl zum Nachweis der (verallgemeinerten) Differenzierbarkeit als auch zur rechnerischen Weiterverarbeitung besser geeignet sind 'Lagrange–Basen' $\{\lambda_i\}$ zu 'lokalen Freiheitsgraden' $\{\sigma_j\}$, d.h.

Sei $\Sigma_\triangle$ Basis von $P_\triangle{}'$ , $\Sigma_\triangle = \{\sigma_1, \ldots, \sigma_{dim}\}$ , $dim = \dim P_\triangle$ , und

$$P_\triangle = \operatorname{span}\left(\{\lambda_{\sigma_i} : 1 \le i \le dim\}\right) \text{ mit } \sigma_j(\lambda_{\sigma_i}) = \delta_{ij} = \begin{cases} 1 & , \quad i = j \\ 0 & , \quad \text{sonst} \end{cases} , \ 1 \le i, j \le dim ,$$

dann heißt $\{\lambda_{\sigma_i}, 1 \le i \le dim\}$ die zu $\Sigma_\triangle$ gehörige Lagrange–Basis.

□

**(5.2.3) Beispiel**
Sei $K = \triangle(a^1, a^2, a^3) \subset \mathbb{R}^2$ , $P_K = \mathbb{P}_2$ , $\dim P_K = 3$ ; $\Sigma_K = \{\sigma_1, \sigma_2, \sigma_3\}$ , $\sigma_j(v_h) = v_h(a^j)$ .
Z.B. für

$$\widehat{\triangle} = \triangle(\,(1,0)\,,\,(0,1)\,,\,(0,0))$$

erhält man als Lagrange–Basis (mit $\lambda_{a^i} \equiv \lambda_{\sigma_i}$):

$$\lambda_{(1,0)}(x) = x_1\ ,\ \lambda_{(0,1)}(x) = x_2\ ,\ \lambda_{(0,0)}(x) = 1 - x_1 - x_2 \ \text{ für } x = (x_1, x_2).$$

□

Mit diesen (lokalen) Lagrange–Basen auf jedem Element erhält man dann (globale) Lagrange–Basen von $V_h$:

**(5.2.4) Lineare $C^0$–Elemente**
Sei $T_h = \{\triangle\}$ Triangulierung von $G$ mit den Knoten (= Dreiecks–Eckpunkten) $\{N\}$, und

$$V_h = \{v \in C^0(\overline{G}) \ : \ v|_\triangle \in \mathbb{P}_2 \text{ für } \triangle \in T_h\} \ , \quad \text{sog. lineare } C^0\text{–Finite Elemente} .$$

Dann ist

$$V_h = \operatorname{span}\{\lambda_N \in V_h \ : \ \lambda_N(N') = \delta_{NN'} \text{ für alle Knoten } N'\}$$

□

Beweis:

Setze

$$\lambda_N|_\triangle = \begin{cases} \lambda_{a^j} \text{ aus (5.2.3)} & , \quad N = a^j \in \triangle = \triangle(a^1, a^2, a^3) \ , \\ 0 & , \quad N \notin \triangle \end{cases}$$

Dann gilt

(i) $\lambda_N|_K \in \mathbb{P}_2$ nach Definition (klar)

(ii) $\lambda_N \in C^0(\overline{G})$, denn: $\lambda_N$ ist stetig in allen Knoten $N' \Longrightarrow \lambda_N$ ist stetig auf allen Kanten $K \cap K' = \operatorname{co}(N, N')$ .

(iii) $\{\lambda_N \ : \ N \text{ Knoten}\}$ linear unabhängig:

$$\textstyle\sum_{N' \text{ Knoten}} c_{N'}\lambda_{N'} = 0 \implies 0 = \sum_{N' \text{ Knoten}} c_{N'}\lambda_{N'}(N) = c_N \quad , \quad N \text{ Knoten} .$$

(iv) $\{\lambda_N\}$ ist Erzeugendensystem: $v_h = \sum_N v_h(N)\lambda_N$ denn:

sei $x \in \overline{K} = \triangle(N_1, N_2, N_3) \Longrightarrow v_h(x) = \sum_{i=1}^3 v_h(N_i)\lambda_{N_i}(x) = \sum\limits_{N \text{ Knoten in } T_h} v_h(N)\lambda_N(x)$ ,

da $\lambda_N|_{\overline{K}} \equiv 0$ für $N \notin \{N_1, N_2, N_3\}$ .

□

Ansatzfunktionen derselben Glattheit, jedoch höherer Approximationsgüte, sind

**(5.2.5) $C^0$–Elemente stückweise quadratischer Funktionen**
Sei

$$V_h := \{v \in C^0(\overline{G}) \ : \ v|_\triangle \in \mathbb{P}_3 \ , \ \triangle \in T_h\} \quad ,$$

wobei $T_h = \{\triangle\}$ Triangulierung von $\overline{G}$, und $P_\triangle = \mathbb{P}_3|_\triangle$ . Dann ist $\Sigma_\triangle = \{\sigma_j, \sigma_{ij}\}$ mit

$$\sigma_j(v_h) := v_h(a^j) \quad , \quad \sigma_{ij}(v_h) := v_h(a^{ij})$$

Basis von $P_\triangle{}'$. Dabei sind $a^j$ die Dreiecks–Ecken und $a^{ij}$ die Seitenmittelpunkte zwischen $a^i$ und $a^j$ $(i < j)$ von $\triangle = \triangle(a^1, a^2, a^3)$.

Beweis:

Wegen $|\Sigma_\triangle| = \dim P|_\triangle$ ist alles gezeigt, falls für $v \in P_\triangle$ gilt: $\sigma(v) = 0 \ \forall \sigma \in \Sigma_\triangle \implies v = 0$ .

*Beweis:* o.E. sei $\triangle(a^1, a^2, a^3) = \triangle((1,0), (0,1), (0,0))$.

Da $v|_{<a^2,a^3>} \in \mathbb{P}_3(\mathbb{R})|_{<a^2,a^3>}$, folgt wegen $\sigma_2(v) = \sigma_3(v) = \sigma_{23}(v) = 0$ , daß $v|_{<a^2,a^3>} \equiv 0$.

$\implies v = \lambda_1(x)w_1(x)$ mit $\lambda_1(x)$ aus (5.2.3) und *deg* $w_1 \leq 1$;

analog gilt: $\lambda_2 \mid v$ , $\lambda_2(x)$ aus (5.2.3) $\implies v = \lambda_1(x)\lambda_2(x)w_0$ , $\deg w_0 = 0$ .

Wegen $0 = v(a^{12}) = \lambda_1\lambda_2 \, w_0 \,|_{a^{12}} = \frac{1}{2} \cdot \frac{1}{2} \, w_0$ folgt $w_0 = 0$ , und damit $v \equiv 0$ . □

Als Übung zeige man das Aussehen der elementweisen Lagrange–Basisfunktionen in (5.2.5) zu $\widehat{\triangle} = \triangle\big((1,0),(0,1),(0,0)\big)$ – siehe auch die Abbildungen 5.5 und 5.6 –.

Analog kann man kubische und höhere $C^0$–Elemente einführen.
Ein offensichtlicher Vorteil dieser Lagrange–Basen ist die Eigenschaft eines 'lokalen Trägers'.
Trivial hat z.B. $1|_K$ *nicht* diese Eigenschaft.

**(5.2.6) Bemerkung**
Für die Lagrangebasis $\{\ \lambda_N\}$der linearen und der quadratischen $C^0$–Elemente zur Triangulierung $T_h$, $N$ Knoten oder Seitenmittelpunkt von $\triangle \in T_h$, gilt

$$\operatorname{supp}(\lambda_N) = \bigcup_{K \in T_h \;:\; N \in \overline{K}} \overline{K}$$

□

Zur Veranschaulichung *Skizze* des 'globalen' Verlaufs der *Lagrange–Basis*funktionen bei *linearen* und *quadratischen $C^0$–Elementen* siehe Abbildungen 5.4 – 5.6:

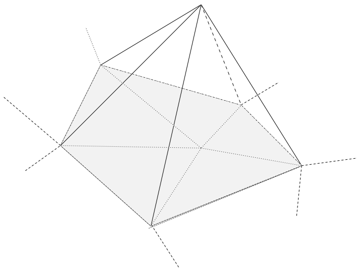

Abbildung 5.4: Träger linearer $C^0$–Lagrange–Basisfunktionen

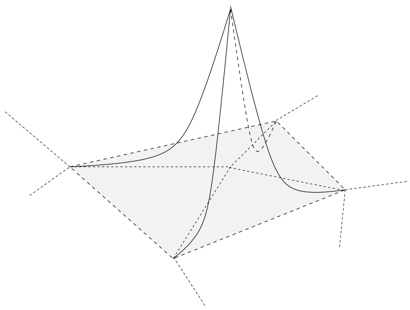

Abbildung 5.5: Träger quadratischer $C^0$–Lagrange–Basisfunktionen zu Eckpunkt

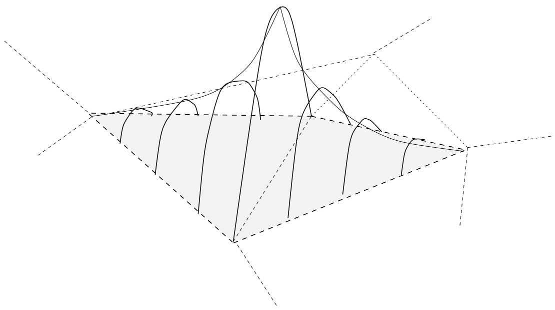

Abbildung 5.6: Träger quadratischer $C^0$–Lagrange–Basisfunktionen zu Kantenmitte

Für das Standard–Referenz–Element

$$\widehat{Q} = [0,1] \times [0,1] \subset \mathbb{R}^2 \ , \ \text{bzw. } \widehat{Q} = [0,1]^d \subset \mathbb{R}^d$$

bieten sich an

$$P_{\widehat{Q}} \equiv \widetilde{P}_k := \mathbb{P}_k \otimes \mathbb{P}_k$$

d.h. Tensorprodukte von Polynomen

$$\widetilde{P}_k = \operatorname{span}\{x^\alpha \equiv \prod_{i=1}^{d} x_i^{\alpha_i} \ : \ \max_{1 \le i \le d} \alpha_i < k\}$$

$$\widetilde{P}_2|_{\mathbb{R}^2} = \operatorname{span}\{1, x_1, x_2, x_1x_2\}$$

$$\widetilde{P}_3|_{\mathbb{R}^2} = \operatorname{span}\{1, x_1, x_2, x_1x_2, x_1^2, x_1^2x_2, x_2^2, x_1x_2^2, x_1^2x_2^2\}$$

Induktion bezüglich $k$ ergibt $\quad \dim \widetilde{P}_k\,|_{\,\mathbb{R}^d} = k^d$ .

Abschließend möchte ich ein $C^1$–Element skizzieren, das *'Argyris'–Element* (siehe Abb. 5.8):

$$P = \mathbb{P}_6|_{\mathbb{R}^2} \;;\; \dim \mathbb{P}_6|_{\mathbb{R}^2} = \binom{7}{2} = 21$$

$\Sigma_K$: Nullte, erste, zweite Ableitungen in $a^i$, $1 \leq i \leq 3$, ergibt $3 \star (1+2+3)$ Bedingungen.
*Zusätzlich* Normalableitungen in

$$a^{ij} = \frac{a^i + a^j}{2} \;,\; i < j$$

ergibt die restlichen 3 Bedingungen.

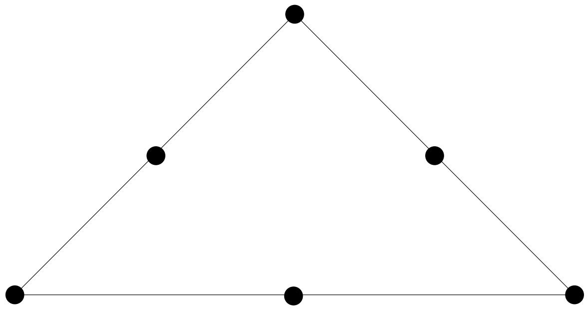

Abbildung 5.7: Quadratisches $C^0$–Element

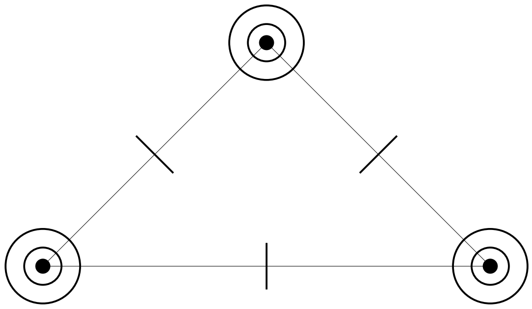

Abbildung 5.8: Argyris–Element

Eine kleine Auswahl vielbenutzter Elemente findet sich in C. Johnson: Numerical solutions of PDE's by the FEM, pp 80.

Ich schätze jetzt noch den Interpolationsfehler für lineare $C^0$–Elemente ab. Dazu folgende *geometrische Größen*:

$$(5.2.7)(i) \quad \begin{cases} h_K = \text{Durchmesser von } K, \text{ d.h. } h_K = \max_{x\,,x' \in K} \|x - x'\|_2; \\ \varrho_K = \text{Radius des größten } K \text{ einbeschriebenen Kreises} \\ \qquad \text{(bei } K = \triangle \text{ gerade Inkreisradius)} \\ h = \max\limits_{K \in T_h} h_K \text{ (dies ist die eigentliche Bedeutung des Index '}h\text{' bei } T_h) \end{cases}$$

**(5.2.7) Satz**
Sei $K = \triangle(a^1, a^2, a^3)$. Zu $v \in C^0(\overline{K})$ sei

$$\pi v \in \mathbb{P}_2 \;:\; \pi v(a^i) = v(a^i) \;,\; 1 \leq i \leq 3$$

Mit den Bezeichnungen (5.2.7)(i) gilt dann für glattes $v$, d.h. $v \in C^2(\overline{K}) \subset W^{2,\infty}(K)$

(a) $\|v - \pi v\|_{L_\infty(K)} \leq 4 \cdot {h_K}^2 \max\limits_{|\alpha|=2} \|D^\alpha v\|_{L_\infty(K)}$ ,

(b) $\max\limits_{|\alpha|=1} \|D^\alpha (v - \pi v)\|_{L_\infty(K)} \leq \left(2 + 3 \cdot \frac{h_K}{\varrho_K}\right) \cdot h_K \cdot \max\limits_{|\alpha|=2} \|D^\alpha v\|_{L_\infty(K)}$ .

Beweis:

Schritt (1): Abschätzung von $\|D^\alpha \pi\|_{L_\infty(K) \to L_\infty(K)}$ :

Es ist

$$\pi v(x) = \sum_{i=1}^{3} v(a^i)\lambda_{a^i}(x) \ , \ x \in K = \triangle(a^1, a^2, a^3) \ .$$

Für $\alpha = 0$ gilt $\|\pi v\|_{\infty,K} \leq \|v\|_{\infty,K}$ , denn :

$$|\pi v(x)| \leq \sum_i |v(a^i)| \cdot |\lambda_{a^i}(x)| \leq \|v\|_{\infty,K} \cdot \underbrace{\sum_{i=1}^{3} |\lambda_{a^i}(x)|}_{= \sum \lambda_{a^i}(x) \equiv 1} \implies$$

$$\|\pi v\|_{\infty,K} \leq \|v\|_{\infty,K} \quad \text{und natürlich gilt} \quad \|\pi v\|_{\infty,K} = \|v\|_{\infty,K} \quad \text{für } v \in \mathbb{P}_2 \ .$$

Für $|\alpha| = 1$ gilt $\|D^\alpha \pi v\|_{\infty,K} \leq \frac{3}{2\varrho_K}\|v\|_{\infty,K}$ , denn :

$$D^\alpha \pi v(x) = \sum_{i=1}^{3} v(a^i) D^\alpha \lambda_{a^i}(x) \implies |D^\alpha \pi v(x)| \leq \|v\|_{\infty,K} \sum_{i=1}^{3} |D^\alpha \lambda_{a^i}(x)| \ .$$

Bezeichnet $\lambda_{a^i}{}'(x;e) =$ die Richtungsableitung von $\lambda_{a^i}$ im Punkt $x$ in Einheitsrichtung $e$, so gilt

$$\begin{aligned} |D^\alpha \lambda_{a^i}(x)| &\leq \max_{e \in \mathbb{R}^2 : \|e\|_2 = 1} |\lambda_{a^i}{}'(x;e)| \overset{(*_1)}{=} \max\left( \left| \frac{1-0}{\|a^i - a^j\|_2} \right|, \left| \frac{1-0}{\|a^i - a^k\|_2} \right| \right) \\ &= \max\left( \frac{1}{\|a^i - a^j\|_2}, \frac{1}{\|a^i - a^k\|_2} \right) \overset{(*_2)}{\leq} \frac{1}{2\varrho_K} \end{aligned}$$

$(*_1)$ gilt, da $\lambda_{a^i}$ linear ist – es ist also der betragsgrößte Differenzenquotient gesucht;
$(*_2)$ gilt, da

$$2\varrho_K \leq \min(\|a^i - a^j\|_2, \|a^i - a^k\|_2) \ .$$

Schritt (2): Fehlerabschätzung über Taylorentwicklung

Es ist $K$ ja konvex; Taylorentwicklung von $v \in C^2(\overline{K})$ um $x_0 \in K$ ergibt dann

$$v(x) = \underbrace{\sum_{|\beta|<2} \frac{1}{\beta!} D^\beta v(x_0)\,(x - x_0)^\beta}_{=: (T_2 v)(x) \in \mathbb{P}_2} + \underbrace{\int_0^1 \sum_{|\beta|=2} \frac{1}{\beta!} D^\beta v(x_0 + t(x - x_0))\,(x - x_0)^\beta \, dt}_{=: (R_2 v)(x)}$$

Zur Erinnerung: $\beta! := \prod_{i=1}^d \beta_i!$ , $x^\beta := \prod_{i=1}^d {x_i}^{\beta_i}$ , $\alpha \geq \beta \Leftrightarrow \alpha_i \geq \beta_i \ \forall\, i$ , $|\alpha| = \sum_i \alpha_i$ ,
$x \in \mathbb{R}^d$ , $\alpha$ , $\beta \in {\mathbb{N}_0}^d$ ; $T_k v$ Taylorpolynom von $v$ um $x_0$ der Ordnung $k$ .

Wegen $T_2 v \in \mathbb{P}_2$ gilt $\pi T_2 v = T_2 v$. Folglich ist $v - \pi v \ = \ v - T_2 v - \pi(v - T_2 v)$ , also folgt

$$\|D^\alpha(v - \pi v)\|_{\infty,K} \leq \|D^\alpha(v - T_2 v)\|_{\infty,K} + \|D^\alpha \pi(v - T_2 v)\|_{\infty,K}$$

$$= \|D^\alpha v - T_{2-|\alpha|} D^\alpha v\|_{\infty,K} + \|D^\alpha \pi(v - T_2 v)\|_{\infty,K} \ , \ D^\alpha T_2 v \overset{\text{ja}}{=} T_{2-|\alpha|} D^\alpha v \, .$$

Schritt (2.i): Abschätzung des Taylorrestglieds mittels $|v|_{2,\infty,K} := \max_{|\beta|=2} \|D^\beta v\|_{\infty,K}$:

$$\|D^\alpha(v-T_2 v)\|_{\infty,K} = \|D^\alpha v - T_{2-|\alpha|} D^\alpha v\|_{\infty,K} \le c^*_{2,\alpha} h_K^{\;2-|\alpha|} |v|_{2,\infty,K}\ ,\ v \in W^{2,\infty}(K)\ ,\ |\alpha| \le 2\ ;$$

Konstante $c^*_{2,\alpha}$ nur von $k = 2$ und $\alpha$ abhängig .
Beweis: klar (eigene Übung), z.B. für $k = 2$ und $\alpha = 0 \equiv (0,0)$:

$$\|v - T_2 v\|_{\infty,K} \le \left( \sum_{|\beta|=2} \frac{1}{\beta!} \right) h_K^{\;2} \sup_{|\beta|=2\ ,\ z\in K} |D^\beta v(z)|$$

$$\Longrightarrow\ c^*_{2,0} = \frac{1}{(2,0)!} + \frac{1}{(1,1)!} + \frac{1}{(0,2)!} = \frac{1}{2} + 1 + \frac{1}{2}\ ,\ \text{also } c^*_{2,0} = 2.$$

Schritt (2.ii): Abschätzung des zweiten Summanden
Nach Schritt (1) gilt die Abschätzung

$$\|D^\alpha \pi(v - T_2 v)\|_{\infty,K} \le \left\{ \begin{array}{c} 1 \\ 3/(2\varrho_K) \end{array} \right\} \|v - T_2 v\|_{\infty,K}\ , \quad \text{falls} \left\{ \begin{array}{l} |\alpha| = 0, \\ |\alpha| = 1. \end{array} \right.$$

Damit folgt mit $c_0 := 2c^*_{2,0}$:

$$\|v - \pi v\|_{\infty,K} \le c_0 h_K^{\;2} |v|_{2,\infty,K} \le c_0 h_K^{\;2} \|v\|_{W^{2,\infty}(K)}\ ,$$

also (a).
(b) folgt wegen $c^*_{2,(1,0)} = \sum_{|\beta|=2-1} \frac{1}{\beta!} = \frac{1}{(1,0)!} + \frac{1}{(0,1)!} = 2 = c^*_{2,(0,1)}$

$$\|D^\alpha(v - \pi v)\|_{\infty,K} \le |v|_{2,\infty,K} \left( c^*_{2,1} h_K^{\;1} + \frac{3}{2\varrho_K} c^*_{2,0} h_K^{\;2} \right) \quad \text{mit } c^*_{2,1} := 2\ .$$

□

Was ich im Satz (5.2.7) für $p = \infty$ gezeigt habe, gilt mit analogen Beweisen auch für $1 \le p < \infty$, insbesondere für $p = 2$.

Betrachtet man Folgen von Triangulierungen,

$$T_h\ ,\ h := \max_{K\in T_h} h_K \in \mathcal{H}\ , \quad \text{mit} \quad h \longrightarrow\ 0\ ,$$

für die mit einem $\beta > 0$ gilt

$$(5.2.8)(i) \quad h_K \le \beta\ \varrho_K\ ,\ K \in T_h\ ,\ h \in \mathcal{H}\ ; \qquad \text{(sog. 'reguläre' Triangulierungsfolgen).}$$

Dann gilt

**(5.2.8) Satz** (Konvergenzsatz)
Sei $G \subset \mathbb{R}^2$ polygonal berandetes beschränktes Gebiet. Die Bilinearform $a : W^{1,2}(G) \times W^{1,2}(G) \longrightarrow \mathbb{R}$ sei $V_0$–koerzitiv mit $W_0^{1,2}(G) \subset V_0 \subset W^{1,2}(G)$. Die Triangulierungsfolge $(T_h)$ von $G$ sei regulär, d.h. es gelte (5.2.8)(i). Ist die Lösung $u$ von (5.2.2) aus $W^{2,\infty}(G)$, dann gibt es $c > 0$, so daß für die Galerkinnäherung $u_h \in V_h \cap V_{bc}$, $V_h$ linearer Finite–Elemente–Raum zu $T_h$, gilt

$$\|u - u_h\|_{W^{1,2}(G)}\ \le\ c\ h\ \|u\|_{W^{2,\infty}(G)}\ .$$

Beweis:

Nach Céa's Lemma (5.1.6) gilt

$$\begin{aligned} \|u-u_h\|_{W^{1,2}(G)} &\le c \inf_{v\in V_{h,bc}} \|u-v_h\|_{W^{1,2}(G)} \le c\,\|u-\pi_h u\|_{W^{1,2}(G)} \\ &\le c'\,\|u-\pi_h u\|_{W^{1,\infty}(G)} \overset{(5.2.7)}{\le} c''\,h\|u\|_{W^{2,\infty}(G)}\ , \end{aligned}$$

wobei $\pi_h u$ Interpolierende $\in V_h$ für $u$ ist.

□

Insbesondere folgt unter den Voraussetzungen von Satz (5.2.8)

$$\lim_{h\to 0}\|u-u_h\|_{L_2(G)} = 0 \qquad \text{und} \qquad \lim_{h\to 0}\left\|\frac{\partial u}{\partial x_i}-\frac{\partial u_h}{\partial x_i}\right\|_{L_2(G)} = 0\ .$$

Die Voraussetzung (5.2.8)(i) ist sicher erfüllt, falls eine Grobtriangulierung verfeinert wird durch Seiten–'halbierung', Seiten–'drittelung', ... . Denn dann sind alle auftretenden Dreiecke ähnlich zu einem Dreieck aus der Ausgangstriangulierung.

**(5.2.9) Bemerkung**
(a) Die Sätze (5.2.7) und (5.2.8) gelten auch analog für $u \in W^{2,p}(G)$ mit $p=2$ anstelle von $p=\infty$.

(b) Unter der zusätzlichen Voraussetzung, daß $(-\triangle)^{-1} : L_2(G) \longrightarrow W^{2,2}(G)$ stetig ist (was z.B. für konvexes polygonal berandetes $G$ gilt), folgt für die Approximationsordnung der Funktionswerte mit dem sog. 'Aubin–Nitsche–Trick'

$$\|u-u_h\|_{L_2(G)} \le \text{ const}\, h^2\,\|u\|_{W^{2,2}(G)}\ .$$

□

Für einen beliebigen koerzitiven Operator $A$ an Stelle von $-\triangle$ muß man in (5.2.9)(b) voraussetzen
'$(A^\star)^{-1} : L_2(G) \longrightarrow W^{2,2}(G)$ stetig '. Die 'richtige' Ordnung für Funktionswerte liefert

**(5.2.10) Satz** (Aubin–Nitsche–'Trick')
Zusätzlich zu den Voraussetzungen von Satz (5.2.8) sei $(A^*)^{-1} : L_2(G) \longrightarrow W^{2,2}(G)$ stetig. Dann gilt

$$\|u-u_h\|_{L_2(G)} \le \text{ const}\, h^2\,\|u\|_{W^{2,2}(G)}$$

Beweis:

$$
\begin{aligned}
\|u-u_h\|_{L_2} &= \sup_{g \in L_2 \;:\; \|g\|_{L_2} \le 1} \langle u-u_h, g\rangle_{L_2} \\
&\overset{(*_1)}{=} \sup_{w \in W_0^{1,2}} a(u-u_h, w)\ , \\
&\overset{(*_2)}{=} \sup_{w \in W_0^{2,2} \;:\; \|w\|_{W^{2,2}} \le c_1} a(u-u_h, w-\pi_h w) \\
&\overset{(*_3)}{\le} \|a\| \cdot \|u-u_h\|_{W^{1,2}} \sup_{w \in W_0^{2,2} \;:\; \|w\|_{W^{2,2}} \le c_1} \|w-\pi_h w\|_{W^{1,2}} \\
&\overset{(*_4)}{\le} \|a\|\ c_2\ h\ \|u\|_{W^{2,2}} \sup_{w \in W_0^{2,2} \;:\; \|w\|_{W^{2,2}} \le c_1} c_3\, h\, \|w\|_{W^{2,2}} \\
&= c\ h^2\ \|u\|_{W^{2,2}}\ .
\end{aligned}
$$

zu $(*_1)$ : löse Randwertaufgabe $A^* w = g$ , $w|_{Rd\ G} = 0$ ; für $g \in L_2(G)$ in Variationsform, i.e. $a(v,w) = < v, g >$ , $v \in W_0^{1,2}$; beachte $u - u_h \in W_0^{1,2}$ .

zu $(*_2)$ : $u_h \in V_{h,bc}$ Galerkinnäherung an $u$, d.h. $a(u-u_h, v_h) = 0,\ v_h \in V_{h,0}$, also speziell auch für $\pi_h w \in V_{h,0}$; außerdem gilt nach der zusätzlichen Voraussetzung für die Lösung $w$ (in Variationsform) von $A^*w = g, w|_{Rd\ G} = 0$ : $\|w\|_{W^{2,2}} \le c_1 \|g\|_{L_2}$ .

zu $(*_3)$ : $a : W^{1,2}(G) \times W^{1,2}(G) \longrightarrow \mathbb{R}$ stetig .

zu $(*_4)$ : nach Céa's Lemma und der Modifikation (5.2.9) der Interpolationsfehlerabschätzung (5.2.7) .

□

Nur einige tabellarische Hinweise zum Aufwand der verschiedenen Möglichkeiten

$$(5.2.11)(i) \quad A\xi = b\ ,\ A \in \mathbb{R}^{M\times M}$$

zu lösen mit

$$M = O(|\{N_h\}|) = O(h^{-d})\ ,\ G \subset \mathbb{R}^d\ ,\ d = 2,3\ ;$$

(zusätzlich sei $A$ auch symmetrisch, um auch das *cg*–Verfahren vergleichen zu können). Unter den in der folgenden Bemerkung gemachten i.a. realistischen Annahmen gilt für den Arbeitsaufwand

**(5.2.11) Bemerkung** (vgl. C. Johnson, S. 139)
Die arithmetische Komplexität zur Lösung von (5.2.11)(i) ist $O(M^\alpha)$ mit $M = O(h^{-d})$ für $G \subset \mathbb{R}^d$ , $d = 2,3$ , wobei der Exponent $\alpha$ in der folgenden Tabelle angegeben ist:

| | Methode | | Exp. $\alpha$ d = 2 | Exp. $\alpha$ d = 3 | Vorteil | Nachteil |
|---|---|---|---|---|---|---|
| direkt | CHOLESKI, wobei Bandbreite von $A$ sei $O(h^{-d+1})$ | Faktorisier. | 2 | 2.33 | | 'dünn besetzt' nicht gut ausnützbar |
| | | back-Subst. | 1.5 | 1.67 | | |
| | geschachtelte Reduktion | Faktorisier. | 1.5 | 2 | | |
| | | back–Subst. | 1 | 1.33 | | |
| iterativ | cg (Conjug. Gradienten) | | 1.5 | 1.33 | leicht program– mierbar | |
| | cg mit Vorkonditionierung | | 1.25 | 1.17 | | 'gute' Vorkond. schwierig |
| | Multigrid | | 1 | 1 | | Programmmier- ung aufwendig |

□

Für sehr spezielle Differentialoperatoren und Gebiete, für die man die Greenfunktion explizit kennt, wie z.B. für $-\triangle$ und die Kugel oder den Halbraum, kann man auch noch den Aufwand der Randelement–Methode vergleichen: $d \longrightarrow d-1$, jedoch $A$ vollbesetzt.

Auch für zeitabhängige Probleme kann man FE–Verfahren formulieren. Ihr großer Vorteil ist, daß sie auch bei 'schwieriger Geometrie' in den räumlichen Variablen 'immer' und (relativ) leicht anzuwenden sind.
Natürlich wurden — und werden auch jetzt noch bei *einfacher Geometrie* und *'einfachen' Randbedingungen* — elliptische Randwertaufgaben auch über Differenzenverfahren gelöst.

## 5.3 Galerkinverfahren für Eigenwertaufgaben

Die Analyse des *Konstanzer Wasserwunders* als stehende Welle (s. die Bemerkung im Anschluß an (4.3)) führt auf das Eigenwertproblem

$$-\triangle u(x) = \lambda u(x)\ ,\ x \in G \quad , \quad \frac{\partial u}{\partial n}(x) = 0\ ,\ x \in Rd\ G\ .$$

Mit $a(u,v) := \int_G \sum_{j=1}^d \frac{\partial u}{\partial x_j}(x)\frac{\partial v}{\partial x_j}(x)dx$ , $< u,v >:= \int_G u(x)v(x)dx$ , erhält man dafür die folgende Variationsformulierung mit $V = W^{1,2}(G)$:

**(5.3.1) Eigenwertaufgabe** (in Variationsformulierung)
Sei $V$ Teilraum mit $W_0^{1,2}(G) \subset V \subset W^{1,2}(G)$ und $a : W^{1,2}(G) \times W^{1,2}(G) \longrightarrow \mathbb{R}$ stetige Bilinearform. Gesucht ist $(u,\lambda) \in V \times \mathbb{C}$, $u \neq 0$, mit

$$a(u,v) = \lambda < u,v > \ ,\ v \in V.$$

□

Die Ersetzung von $V$ durch einen FE-Raum $V_h$ ergibt

**(5.3.2) Galerkin-Diskretisierung** (von (5.3.1))
In der Situation von (5.3.1) sei $V_h \subset V$ ein Finite–Elemente–Unterraum. Gesucht ist $(u_h, \lambda_h) \in V_h \times \mathbb{C}$, $u_h \neq 0$, mit

$$a(u_h, v_h) = \lambda_h < u_h, v_h > \ , \ v_h \in V_h .$$

□

Diese Aufgabe führt auf eine sog. *verallgemeinerte (Matrix–) Eigenwertaufgabe.* Denn mit

$$M = \dim V_h \ , \ V_h = span(\{u_1, u_2, ..., u_M\})$$

gilt:

$$\begin{aligned} & a(u_h, v_h) = \lambda_h < u_h, v_h > \ , \ v_h \in V_h \ ; \\ \Longleftrightarrow \quad & a(u_h, u_i) = \lambda_h < u_h, u_i > \ , \ 1 \leq i \leq M \ ; \\ \Longleftrightarrow \quad & u_h = \sum_j \xi_j u_j \quad , \quad \sum_j a(u_j, u_i)\xi_j = \lambda_h \sum_j < u_j, u_i > \xi_j \ , \ 1 \leq i \leq M \ ; \end{aligned}$$

Mit $x = [\xi_j]_{1 \leq j \leq M} \in \mathbb{R}^M$ *resp.* $\mathbb{C}^M$ , $A = [a(u_j, u_i)]_{1 \leq i,j \leq M} \in \mathbb{R}^{M \times M}$ *resp.* $\mathbb{C}^{M \times M}$ und $B = [< u_j, u_i >]_{1 \leq i,j \leq M} \in \mathbb{R}^{M \times M}$ *resp.* $\mathbb{C}^{M \times M}$ folgt

$$\Longleftrightarrow \quad Ax = \lambda_h Bx, \quad x \in \mathbb{R}^M \ resp. \ \mathbb{C}^M \ : \ x \neq 0 \quad (x \neq 0 \Longleftrightarrow u_h \neq 0) \ .$$

□

$B$ ist als Gramsche Matrix symmetrisch und invertierbar (s. Numerik 1, Satz (3.2.4)). Damit ist diese verallgemeinerte $EWA$ äquivalent zu einer Standard–$EWA$

$$Ax = \lambda_h Bx \quad \Longleftrightarrow \quad B^{-1}Ax = \lambda_h x \quad \Longleftrightarrow \quad AB^{-1}(Bx) = \lambda_h (Bx) \ .$$

Für den Rest dieses Abschnitts betrachte ich folgende

**(5.3.3) Matrix–EWA**
Gegeben $A \in \mathbb{C}^{M \times M}$ , $B \in \mathbb{C}^{M \times M}$ nichtsingulär. Gesucht ist $(x, \lambda) \in \mathbb{C}^M \times \mathbb{C}$, $x \neq 0$ mit

$$Ax = \lambda Bx.$$

$\lambda$ heißt Eigenwert, $x$ zugehöriger Eigenvektor; $S(A, B)$ bezeichnet die Menge aller Eigenwerte der Matrizenschar $A - \lambda B, \lambda \in \mathbb{C}$

□

Für die *symmetrische verallgemeinerte EWA*, d.h. es ist
*A symmetrisch* und *B symmetrisch und positiv definit*;
gilt

**(5.3.4) Satz**
Sei $A$ symmetrisch, $B$ symmetrisch und positiv definit, und $B = LL^t$ die Cholesky–Zerlegung von $B$. Dann ist $\widetilde{A} = L^{-1}A(L^{-1})^t$ symmetrisch und $S(A, B) = S(\widetilde{A})$.

Beweis:

$Ax = \lambda Bx \Longleftrightarrow Ax = \lambda LL^t x \Longleftrightarrow L^{-1}Ax = \lambda L^t x \Longleftrightarrow L^{-1}A(L^t)^{-1}y = \lambda y$ mit $y = L^t x$ .

□

Wendet man die Vektoriteration auf die Matrix $\widetilde{A}$ aus (5.3.4) an,

$$\widetilde{y}^{(k)} := \widetilde{A}y^{(k-1)} \ , \ y^{(k)} := \frac{\widetilde{y}^{(k)}}{\|\widetilde{y}^{(k)}\|} \ ;$$

so muß man – zusätzlich zur Matrix–Vektor–Multiplikation mit $A$ – (nur) zwei Dreiecks–Gleichungssysteme lösen pro Iterationsschritt:

$$L^t z = y_{alt} \quad , \quad L\widetilde{y}_{neu} = Az \quad , \quad y_{neu} = \frac{\widetilde{y}_{neu}}{\|\widetilde{y}_{neu}\|} \ .$$

Die Übertragung des $QR$–Verfahrens für $AB^{-1}$ auf die Matrizenschar $A - \lambda B$, ohne die Matrix $AB^{-1}$ zu bilden, ist gerade das sog. *QZ–Verfahren* .
Ich verweise auf *GOLUB, G.H., van LOAN, Ch.F.*: Matrix Computations. Johns Hopkins Univ. Pr., versch. Auflagen seit 1983, ISBN 0-8018-3011-7 Pbk. oder *BUNSE, W., BUNSE–GERSTNER, A.*: Numerische Lineare Algebra. Teubner 1985, ISBN 3-519-02067-X.

In der Simulation des 'Konstanzer Wasserwunders' (s. Wittum, G.: Mehrgitterverfahren. Spektrum der Wissenschaften, April 1990, S. 78–90), egibt sich für die Näherung der 10. Eigenfunktion von $-\triangle$ eine Schwingungsdauer der zugehörigen stehenden Welle von ca. 12 Minuten, was dem Eintrag des Stadtschreibers „... und das Wasser ist an– und abgelaufen vier– bis fünfmal in der Stund ...“ recht gut entspricht.

# Teil III

# Optimierung im $\mathbb{R}^n$

**Literatur zu Optimierungsaufgaben**

Aus der 'riesigen' Auswahl von Lehrbüchern und Monographien nur eine ganz kleine, persönlich gefärbte Auswahl

*eher theoretisch:*

Bertsekas, Dimitri P.: Constrained optimization and Lagrange multiplier methods. Academic Press 1982 ; Computer science and applied mathematics.
ISBN-Nr.: 0-12-093480-9;

Ciarlet, Philippe G.: Introduction a l'analyse numerique matricielle et a l'optimisation. Masson 1982 ; Collection mathematiques appliquees pour la maitrise.
ISBN-Nr.: 2-226-68893-1.

Luenberger, David G.: Linear and nonlinear programming. 2. ed.; Addison-Wesley 1984, ISBN-Nr.: 0-201-15794-2.

*mehr numerisch:*

Boyd, Stephen; Vandenberghe,Lieven: Convex Optimization. Cambridge Univ. Press 2004, 65.00 US $, ISBN-Nr.: 0521833787 2004.
elektronische Version kostenlos bei http://www.stanford.edu/ boyd/cvxbook.html

Dennis, John E.; Schnabel, Robert B.: Numerical methods for unconstrained optimization and nonlinear equations. Prentice-Hall 1983 ; Prentice-Hall series in computational mathematics, ISBN-Nr.: 0-13-627216-9.

Spellucci, Peter: Numerische Verfahren der nichtlinearen Optimierung. Birkhäuser 1993 ; ISBN-Nr.: 3-7643-2854-1.

Werner, Jochen: Numerische Mathematik 2. Eigenwertaufgaben, lineare Optimierungsaufgaben. Vieweg ; Vieweg Studium : Aufbaukurs Mathematik Vol. ...
ISBN-Nr.: 3-528-07233-4.

# Kapitel 6

# Uneingeschränkte Minimierung

## 6.1 Abstiegsmethoden

**(6.1.1) Allgemeines Minimierungsproblem**
Gegeben sei $f : D \subset \mathbb{R}^n \to \mathbb{R}$ mit $D \neq \emptyset$. Gesucht ist eine *Minimumstelle*
$x^\star \in arg\inf\ f(D)$ , d.h. $x^\star \in D$ mit $f(x^\star) = \inf_{x\in D} f(x)$, also
$f(x^\star) \leq f(x)\ \ \forall\, x \in D.$ □

Häufig ist $D = \mathbb{R}^n$. Man spricht dann von *uneingeschränkter Minimierung.* Zusätzliche Nebenbedingungen an die zulässigen Punkte $\{x\}$ können in Gleichungs- oder/und Ungleichungsform gestellt werden: Sei $G \subset \mathbb{R}^n$ offen und $g : G \to \mathbb{R}^m, h : G \to \mathbb{R}^l$, dann ist
$D := \{x \in G : g(x) \leq 0\ ,\ h(x) = 0\}$,
wobei das '$\leq$' hier komponentenweise zu verstehen ist.

Existenz von $\inf\ f(D)$ (d.h. $f$ nach unten beschränkt) und einer Minimumstelle $x^\star$ werden vorausgesetzt.

**Bemerkung**
(a) $f : \mathbb{R}^n \to \mathbb{R}$ differenzierbar, dann gilt:
$$x^\star \in arg\inf\ f(\mathbb{R}^n) \Longrightarrow \nabla f(x^\star) = 0\ ,\ \text{d.h. } x^\star \textit{ stationärer } \text{Punkt.}$$
(b) $f : \mathbb{R}^n \to \mathbb{R}$ richtungsdifferenzierbar, dann gilt:
$$x^\star \in arg\inf\ f(\mathbb{R}^n) \Longrightarrow f'(x^\star, d) \geq 0\ ,\ d \in \mathbb{R}^n\ ,\ \text{d.h. } x^\star \text{ sog. } \textit{statischer } \text{Punkt.}$$

**(6.1.2) Prinzipieller Abstiegsalgorithmus** ($f$ richtungsdifferenzierbar; keine Nebenbedingungen)
Gegeben: $f : \mathbb{R}^n \to \mathbb{R}$ richtungsdifferenzierbar.
Gesucht: statischer Punkt $x^\star$ ( als Ersatz für '$x^\star \in \arg inf f(\mathbb{R}^n)$' ).

(0) Bestimme Ausgangsnäherung $x_a \in \mathbb{R}^n$ für $x^\star$

(1) Optimalitätstest: $f'(x_a, d) \overset{?}{\geq} 0 \quad \forall\ \ d \in \mathbb{R}^n$

Falls $x_a$ kein statischer Punkt ist, führe die folgenden Schritte aus:

(2) Bestimme eine Abstiegsrichtung $s \neq 0$ mit $f'(x_a, s) < 0$

(3) Bestimme $x_n \in \arg\inf f([x_a, x_a + s >)$ nächstgelegen zu $x_a$

(4) Wiederhole (1) mit $x_a := x_n$.

□

Im Schritt (3) ist dabei $[x_a, x_a + s >= \{x_a + ts : t \geq 0\}$ der Halbstrahl ausgehend von $x_a$ in Richtung $s$.

**(6.1.2') Bemerkung**

(a) $f \in C^1 \Longrightarrow f'(x,d) = < \nabla f(x), d >$;

(b) $f \in C^1$, dann ist die Richtung des 'steilsten Abstiegs' $s^\star \in \; arg \; \min \; \{f'(x_a, d) : \|d\|_2 = 1\}$ gerade

$$s^\star \;=\; -\nabla f(x)/\|\nabla f(x)\|_2;$$

(c) Schritt (3) in (6.1.2) heißt 'eindimensionale Minimierung' oder 'exakte Suche';

(d) Es gibt verschiedene Strategien für eine 'näherungsweise Minimierung' oder 'inexakte Suche' derart, daß auch dann Konvergenz gegen einen statischen Punkt eintritt.

□

Das naheliegendste Abstiegsverfahren ist das

**(6.1.3) Verfahren des steilsten Abstiegs** (oder auch 'Gradientenverfahren')
Sei $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ aus $C^1$. Dann setze in (6.1.2) in

Schritt (1): $\nabla f(x_a) \stackrel{?}{=} 0$;

Schritt (2): $s := -\nabla f(x_a)$;

Schritt (3): $t^\star > 0$ minimal mit $< \nabla f(x_a + t^\star s), s >= 0$.

□

Daß die Strategie, 'lokal den bestmöglichen Abstieg zu realisieren', sich in der Gesamtstrategie als 'zu kurz gesprungen' erweist, sieht man an quadratischen Zielfunktionen. Dazu

**(6.1.4) Typische Testfunktionen** (für Minimierungsalgorithmen)
Sei $f : G \subset \mathbb{R}^n \to \mathbb{R}$ , $G$ offen, eine $C^2$-Funktion und sei $C \subset G$ konvex. Dann gilt:

(a) Ist $f''(x) \equiv H_f(x)$ positiv semidefinit für alle $x \in C$, so ist $f$ konvex in $C$.

(b) Ist $f''(x)$ gleichmäßig positiv definit in $C$ mit der Positivitäts–Konstanten $c > 0$, so ist $f$ mit dieser Konstanten gleichmäßig konvex[1] in $C$, d.h. es gilt
$f(x+d) \geq f(x) + \nabla f(x)^t d + c < d, d >$ , $x, x+d \in C$.

(c) Ist $C$ offen, so gilt in den beiden vorigen Aussagen auch die Umkehrung.

(d) Eine quadratische Funktion $q(x) = c_0 + c^t x + \frac{1}{2} x^t Q x$ mit symmetrischem $Q$ ist gleichmäßig konvex bzw. konvex, wenn $Q$ positiv definit bzw. positiv semidefinit ist.

□

Dies sind alles 'gutartige' Testfunktionen. Für Klassen von 'schwierigen' Testfunktionen s. z.B. die website *http://plato.la.asu.edu/topics/testcases.html.*

**Bemerkung**
Für gleichmäßig konvexes $f$ ist jeder stationäre Punkt (eindeutiger) Minimumpunkt.

□

[1] $f$ ist nach Definition gleichmäßig konvex in $C$, wenn $(1-t)f(x) + tf(y) \geq f\big((1-t)x + ty\big) + ct(1-t)\|x-y\|_2{}^2$ für $x, y \in C$. Obige Bedingung ist äquivalent hierzu.

Verhalten des Gradientenverfahrens für $q(x) = \frac{1}{2}x^tQx = \frac{1}{2}\left({x_1}^2 + 9{x_2}^2\right)$ für $x_{start} = [9,1]^t$.
Bezeichnet man die vom Gradientenverfahren erzeugte Folge mit $x_k$ , $g_k = \nabla q(x_k)$, so gilt $x_{k+1} = x_k - \alpha_k g_k$ mit $\alpha_k =< g_k, g_k > / < g_k, Qg_k >$. Man rechnet nach, daß

$$x_k = 0.8^k[9,(-1)^k]^t \ , \ g_k = Qx_k = 0.8^k[9, 9(-1)^k]^t \ , \ < g_k, g_k >= 0.8^{2k}\left(9^2 + 9^2\right) \ ,$$

$$< g_k, Qg_k >= 0.8^{2k}\left(9^2 + 9^3\right) \ , \ \alpha_k =< g_k, g_k > / < g_k, Qg_k >= 2/10 \ .$$

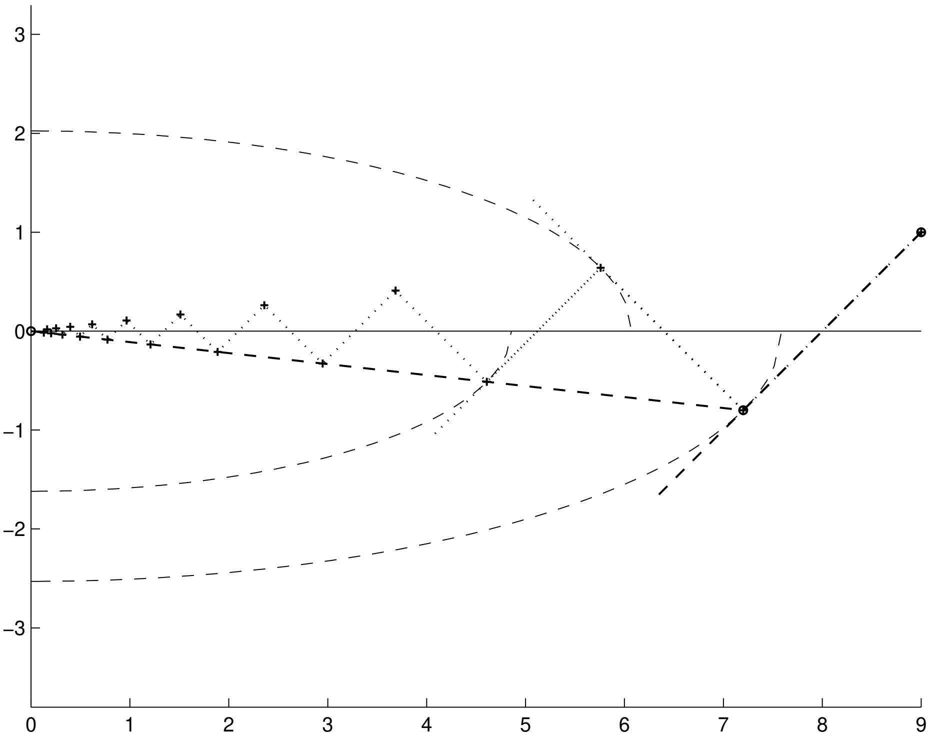


Abbildung 6.1: steilster Abstieg '+' im Vergleich zu konjugierten Gradienten 'o'

Dieses *Zick–Zack–Verhalten* des Gradientenverfahrens hat eine sehr langsame Konvergenz gegen den Minimumpunkt $x^\star = [0,0]^t$ zur Folge.
Beim Gradienten–Verfahren für quadratische Funktionen wird die Genauigkeit pro Iterationsschritt um den Faktor $\mu := \dfrac{cond_2(Q) - 1}{cond_2(Q) + 1}$ verbessert. Es ist aber $\mu << 1$ nur für $cond_2(Q)$ nahe bei 1 .

Ein Grund für die langsame Konvergenz ist die Tatsache, daß das Gradientenverfahren überhaupt nicht die bereits durchlaufenen Richtungen – und damit ein wenig die Krümmung – mit berücksichtigt. Dies geschieht beim Verfahren der konjugierten Richtungen. Dabei sind aufeinander folgende Abstiegsrichtugen *Q–konjugiert*, d.h. es gilt

$$< s_{neu}, Q\, s_{alt} > \ = \ 0 \ .$$

Im folgenden Algorithmus kürze ich $\nabla f$ durch $g$ ab als Bezeichnung für den Gradienten, und $n :\equiv neu$ und $a :\equiv alt$

$$x_a \ :\equiv x_{alt} \ , \ x_n \ :\equiv x_{neu} \ , \ g(x) \ \equiv \ \nabla f(x) \ , \ g_a \ :\equiv \ \nabla f(x_{alt}) \ , \ g_n \ :\equiv \ \nabla f(x_{neu}) \ .$$

Das Verfahren der konjugierten Richtungen oder 'konjugierten Gradienten' wird abkürzend als *cg*–Verfahren bezeichnet.

**(6.1.5) *cg* –Verfahren**
*Gegeben :* $Q \in \mathbb{R}^{n\times n}$ *symmetrisch, positiv definit* , $c \in \mathbb{R}^n$ .

*Gesucht :* $x^\star \in \mathbb{R}^n$ mit $Q\,x^\star + c = 0$ ; bzw.
$x^\star = \arg\min q(x)$ , wobei $q(x) \equiv \frac{1}{2}x^tQx + c^tx$ .

Bestimme eine *Startnäherung* $x_a$ ( gewöhnlich $x_a = 0$ ) , und berechne

$$
\begin{array}{lll}
(1) & g_a := Qx_a + c & ( g_a \equiv \nabla q(x_a) ) \\
 & s_a := -g_a\ ; &
\end{array}
$$

*Solange* $g_a \neq 0$ ( *i.e.* $g_a{}^tg_a = \|g_a\|_2{}^2 \overset{?}{>} 0$ ) , bestimme neue Näherung $x_n$ durch

$$
\begin{array}{llll}
(2) & \alpha & := (g_a{}^t g_a)/(s_a{}^tQs_a) & ( \textit{theoretisch}:\ \alpha := -(g_a{}^t s_a)/(s_a{}^tQs_a) ) \\
(3) & x_n & := x_a + \alpha\, s_a & \\
(4) & g_n & := g_a + \alpha\, Qs_a & ( \textit{theoretisch}:\ g_n := grad\, q(x_n) ) \\
(5) & \beta & := (g_n{}^t\, g_n)/(g_a{}^t\, g_a) & ( \textit{theoretisch}:\ \beta := (g_n{}^tQs_a)/(s_a{}^tQs_a) ) \\
(6) & s_n & := -g_n + \beta\, s_a\ ; &
\end{array}
$$

$$(x_a, g_a, s_a) := (x_n, g_n, s_n)\ .$$

*andernfalls* ist $x^\star := x_a$ Lösung von $Qx + c = 0$. □

**(6.1.6) Bemerkung**
In Algorithmus (6.1.5) gilt

(a) die Wahl von $\alpha$ ergibt sich aus der Bedingung $q(x_a + \alpha s_a) = \min_{t\in\mathbb{R}} q(x_a + ts_a)$ , der 'eindimensionalen Minimierung';

(b) die Wahl von $\beta$ in $s_n = -g_n + \beta s_a$ ergibt sich aus der Bedingung $s_n{}^tQs_a = 0$ , daß $s_n$ und $s_a$ also 'Q-konjugiert' sind.

Beweis:

(a) Mit $\varphi(t) := q(x_a + ts_a)$ gilt notwendig (und hier auch hinreichend)

$$0 = \varphi'(t)\big|_{t=\alpha} = < grad\ q(x_a + ts_a)\big|_{t=\alpha}, s_a > = < g(x_a + \alpha s_a), s_a > = < g_a, s_a > + \alpha < Qs_a, s_a > .$$

(b) Für $s_n = -g_n + \beta s_a$ gilt

$$< s_n, Qs_a > \ = 0 \iff \beta = \ < g_n, Qs_a > / < s_a, Qs_a >$$

□

Übung: Man verifiziere die in (6.1.5) (2), (4) und (5) tatsächlich verwendeten Größen.

Seien die von der iterativen Anwendung des *cg* –Verfahrens erzeugten Größen entsprechend mit $\{x_0, x_1, x_2, ...\}$ , $\{g_0, g_1, s_2, ...\}$ und $\{s_0, s_1, s_2, ...\}$ . Dann gelten die folgenden

**(6.1.7) Eigenschaften des *cg* –Verfahrens**
Für alle $g_k \neq 0$ gilt

(a) $g_k \perp span\big(\{s_0, ..., s_{k-1}\}\big)$ ,

(b) ${s_k}^t Q s_i \;=\; 0 \;,\; 0 \leq i \leq k-1$;

(c) $span\big(\{g_0, g_1, ..., g_k\}\big) = span\big(\{g_0, Qg_0, Q^2 g_0, .., Q^k g_0\}\big)$;

(d) $span\big(\{s_0, s_1, ..., s_k\}\big) = span\big(\{g_0, Qg_0, Q^2 g_0, .., Q^k g_0\}\big)$.

$span\big(\{g_0, Qg_0, Q^2 g_0, ....\}\big)$ ist der sog. *Krylovraum* von $Q$ zu $g_0$.

Beweis:

Wir zeigen (a) – (d) gleichzeitig durch Induktion.
$k = 0$ ist klar wegen $s_0 = -g_0$ in (d).
$k \rightsquigarrow k+1$: sei $g_{k+1} \neq 0$.

(a) ${g_{k+1}}^t s_k = 0$ nach (6.1.6)(a) ;
für $i \leq k-1$ gilt
$${g_{k+1}}^t s_i \;=\; (g_k + \alpha_k Q s_k)^t s_i \;=\; {g_k}^t s_i + \alpha_k {s_k}^t Q s_i \;=\; 0$$
nach Induktionsvoraussetzung (a) und (b).

(b) ${s_{k+1}}^t Q s_k \;=\; -{g_{k+1}}^t Q s_k + \beta_k {s_k}^t Q s_k \;=\; 0$ nach (6.1.6)(b) ;
für $i \leq k-1$ gilt nach Induktionsvoraussetzung (b)
$${s_{k+1}}^t Q s_i \;=\; (-g_{k+1} + \beta_k s_k)^t Q s_i \;=\; -{g_{k+1}}^t Q s_i \;;$$
Unter zweimaliger Anwendung der Induktionsvoraussetzung (d) folgt für $i \leq k-1$
$$Q s_i \in span\big(\{QQ^j g_0 \;:\; 0 \leq j \leq k-1\}\big) \subset span\big(\{s_0, ..., s_k\}\big)$$
und damit ${s_{k+1}}^t Q s_i \;=\; 0$ nach der für $k+1$ bereits gezeigten Aussage (a).

(c) $g_{k+1} = Qx_{k+1} + c = Qx_k + \alpha_k Q s_k + c = g_k + \alpha_k Q s_k$. Also gilt nach Induktionsvoraussetzung (c) und (d) $g_{k+1} \;\in\; span\big(\{Q^j g_0 \;:\; 0 \leq j \leq k\}\big) + span\big(\{QQ^j g_0 \;:\; 0 \leq j \leq k\}\big)$.
Damit gilt
$$span\big(\{g_0, g_1, ..., g_{k+1}\}\big) \subset span\big(\{Q^j g_0 \;:\; 0 \leq j \leq k+1\}\big).$$
Aus der Annahme, daß die beiden Mengen nicht gleich sind, folgt aus Dimensionsgründen, daß
$$g_{k+1} \;\in\; span\big(\{g_0, g_1, ..., g_k\}\big) \;=\; span\big(\{s_0, s_1, ..., s_k\}\big).$$
Da (a) für $k+1$ bereits gezeigt wurde, folgt $g_{k+1} = 0$ im Widerspruch zur Voraussetzung.

(d) $s_{k+1} = -g_{k+1} + \beta_k s_k$ ist nach (c) linear unabhängig von $\{s_0, ..., s_k\}$. Wieder aus Dimensionsgründen gilt damit in $span\big(\{s_0, s_1, ..., s_{k+1}\}\big) \subset span\big(\{Q^j g_0 \;:\; 0 \leq j \leq k+1\}\big)$ die Gleichheit.

□

Theoretisch bricht das cg–Verfahren für quadratische Funktionen wegen (6.1.7)(a) nach höchstens $n$ Schritten ab im Gegensatz zum Gradientenverfahren.
Die Verbesserung pro Iterationsschritt wird bestimmt durch $\nu := 1 - \dfrac{1}{\sqrt{cond_2(Q)}}$. Auch hier ist $\nu^{\,k} \;<<\; \mu^{\,k}$ des Konvergenzfaktors $\mu$ des Gradientenverfahrens. Daher kann man versuchen, das Ausgangsproblem besser zu konditionieren.
*Eine* Zielsetzung einer Vorkonditionierung ist daher $cond_2(Q)$ $klein$ !

*Eine weitere* wünschenswerte Situation ist $dim\big(\{Q^k x_{start} \;:\; k \in \mathbb{N}_0\}\big)$ $klein$ !

Diese wünschenswerten Situationen kann man durch eine Vorkonditionierung zu erreichen versuchen.

**(6.1.8) *cg*-Verfahren mit Vorkonditionierung**

*Gegeben* : $Q \in \mathbb{R}^{n\times n}$ , $c \in \mathbb{R}^n$ wie in (6.1.5);

*Gesucht* : $x^\star \in \mathbb{R}^n$ mit $Q\,x^\star + c = 0$ ; bzw.
$x^\star = \arg\min q(x)$ , wobei $q(x) \equiv \frac{1}{2}x^tQx + c^tx$ .

*Bekannt* : *Vorkonditionierungsmatrix* $C \in \mathbb{R}^{n\times n}$ von der Form $C = H\,H^t$ mit $H \in \mathbb{R}^{n\times n}$ nichtsingulär (z.B. $C = LL^t$ eine näherungsweise Cholesky–Zerlegung von $Q$) ;
wünschenswert : $cond_2(C^{-1}Q) << cond_2(Q)$ oder $rang(C-Q)$ *klein* .

Bestimme eine *Startnäherung* $x_a$ ( gewöhnlich $x_a = 0$ ) , und berechne

| | | |
|---|---|---|
| (1) | $g_a := Qx_a + c$ | ( $g_a \equiv grad\ q(x_a)$ ) |
| (1′) | $C\,z_a = g_a$ | ( *lin. Gl. syst. für* $z_a$ ) |
| | $s_a := -z_a$ | |

*Solange* $g_a \neq 0$ ( *i.e.* $g_a{}^tg_a = \|g_a\|_2{}^2 \overset{?}{>} 0$ ) , bestimme neue Näherung $x_n$ durch

| | | | | |
|---|---|---|---|---|
| (2) | $\alpha$ | $:=$ | $(g_a{}^tz_a)/(s_a{}^tQs_a)$ | |
| (3) | $x_n$ | $:=$ | $x_a + \alpha\,s_a$ | |
| (4) | $g_n$ | $:=$ | $g_a + \alpha\,Qs_a$ | |
| (4′) | $C\,z_n$ | $=$ | $g_n$ | ( *lin. Gl. syst. für* $z_n$ ) |
| (5) | $\beta$ | $:=$ | $(g_n{}^tz_n)/(g_a{}^tz_a)$ | |
| (6) | $s_n$ | $:=$ | $-z_n + \beta\,s_a$ | |

$(x_a, g_a, z_a) \quad := \quad (x_n, g_n, z_n)$ .

andernfalls ist $x^\star := x_a$ Lösung.

□

Die hier betrachtete Vorkonditionierung heißt auf Grund der Form $C = HH^t$ vom 'unvollständigen Cholesky-Typ'.

Sei $C = HH^t$ , $\widetilde{Q} := H^{-1}QH^{-t}$ , $\widetilde{c} = H^{-1}c$ . Die Größen des gewöhnlichen *cg*-Verfahrens zur Lösung von $\widetilde{Q}\widetilde{x} + \widetilde{c} = 0$ seien $\widetilde{x}_a$ , $\widetilde{g}_a$ , $\widetilde{d}_a$ . Dann ergeben sich für die Größen

$x_a := H^{-t}\widetilde{x}_a$ , $s_a := H^{-t}\widetilde{s}_a$ , $z_a := C^{-1}g_a$ , $z_n := C^{-1}g_n$ ,

obige Formeln. Dabei ist

$g_a := grad\ q(x_a)$ mit $q(x) = \frac{1}{2}x^tQx + c^tx$, und $\widetilde{g}_a = grad\ \widetilde{q}(\ \widetilde{x}_a\ )$, mit $\widetilde{q}(\widetilde{x}) = \frac{1}{2}\widetilde{x}^t\widetilde{Q}\widetilde{x} + \widetilde{c}^t\widetilde{x}$.

Es gilt

$H^{-1}\ grad\ q(\ H^{-t}\widetilde{x}\ ) \ = \ grad\ \widetilde{q}(\ \widetilde{x}\ )$ .

**(6.1.9) Beispiel**

Für

$$Q = \begin{bmatrix} 2 & 1 & & & & \\ 1 & 2 & 1 & & & \\ & 1 & 2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & 1 \\ & & & & 1 & 2 \end{bmatrix}$$

ist eine Vorkonditionierungsmatrix $C$ der Form $C = HH^t$ gegeben durch eine Fast–Cholesky–Zerlegung von $Q$ mit

$$H = \begin{bmatrix} 1 & & & & & \\ 1 & 1 & & & & \\ & 1 & 1 & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & 1 & 1 \end{bmatrix} .$$

In diesem 'akademischen' Beispiel gilt $rang(C - Q) = 1$, wie man sofort sehen kann. Damit konvergiert das vorkonditionierte $cg$–Verfahren in höchstens zwei Schritten unabhängig von der Größe $n$ der Matrix $Q \in \mathbb{R}^{n \times n}$ (was man sich natürlich noch genauer überlegen müsste, vgl Aufgabe 10.2).

## 6.2 Quasi–Newton–Verfahren

Grundlage ist wieder die notwendige Bedingung $\nabla f(x^*) = 0$ .

**(6.2.1) Newton–Idee im Rahmen der Minimierung**
Sei $f \in C^2(\mathbb{R}^n)$.

(0) Bestimme Startnäherung $x_a \in \mathbb{R}^n$ für $x^\star$.

(1) Optimalitätstest: $\nabla f(x_a) \stackrel{?}{=} 0$

Falls $x_a$ kein stationärer Punkt ist, führe die folgenden Schritte aus:

(2)
$$\begin{aligned} \widetilde{f}(x_a + \Delta x) &= f(x_a) + f'(x_a)\Delta x + \frac{1}{2} f''(x_a)\Delta x^2 \\ &= f(x_a) + \nabla f(x_a)^t \Delta x + \frac{1}{2}\Delta x^t H_f(x_a)\Delta x \; ; \; H_f(x) = \left[\tfrac{\partial^2 f}{\partial x_i \partial x_j}(x)\right] . \end{aligned}$$

(3) bestimme $\Delta x \in arg \min\limits_{\Delta x \in \mathbb{R}^n} \widetilde{f}(x_a + \Delta x) \quad ; \quad x_n := x_a + \Delta x$ .

(4) Wiederhole (1) mit $x_a := x_n$.

Andernfalls ist $x^* := x_a$ stationärer Punkt.

□

Den Zusammenhang mit der Nullstellensuche des klassischen Newtonverfahrens und die Durchführbarkeit von Algorithmus (6.2.1) unter geeigneten Voraussetzungen zeigt

**(6.2.2) Bemerkung**
Für gleichmässig konvexes $f$ ist (6.2.1) durchführbar. Die Iterierten stimmen mit denen aus der Newtoniteration

$$x_n \quad : \quad \nabla f(x_a) + H_f(x_a)(x_n - x_a) \;=\; 0$$

zur Bestimmung einer Nullstelle von $\nabla f$ überein. Damit ist die Durchführbarkeit und lokal quadratische Konvergenz von Algorithmus (6.2.1) gesichert.

<u>Beweis:</u>

In (6.2.1) ist nach Voraussetzung $\widetilde{f}$ gleichmässig konvex; damit erfüllt $\Delta x$ die notwendige Bedingung $\widetilde{f}'(x_a + \Delta x) = 0$.

$$0 \stackrel{!}{=} \nabla \widetilde{f}(x_a + \Delta x) \qquad \Longleftrightarrow \qquad H_f(x_a)\Delta x + \nabla f(x_a) \;=\; 0 \;.$$

□

Die Nachteile des Newtonverfahrens – Verwendung 2. Ableitungen und 'nur sehr lokale Konvergenz' – werden verbessert durch

– Abstiegsrichtungen der Broydenklasse

– eindimensionale Suche.

Damit erhält man

**(6.2.3) Quasi–Newton–Schritt** (prinzipiell)
Sei $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ differenzierbar mit lipschitzstetiger Ableitung (und $g(x) :\equiv \nabla f(x)$ verfügbar).

(0) Bestimme Startnäherung $x_a \in \mathbb{R}^n$ für $x^\star$, berechne $g_a = \nabla f(x_a)$, und bestimme $B_a \in \mathbb{R}^{n\times n}$ positiv definit und symmetrisch, z.B. $B_a = H_f(x_a)$ oder $B_a = E$.

(1) Optimalitätstest: $g_a \stackrel{?}{=} 0$

Falls $x_a$ kein stationärer Punkt ist, führe die folgenden Schritte aus:

(2) bestimme $s_a$ mit $B_a s_a \;=\; -\, g_a$

(3) bestimme Schrittweite $t_a \;>\; 0$ durch (in-)exakte Suche

(4) berechne $x_n \;:=\; x_a + t_a s_a$ und $g_n = \nabla f(x_n)$

(5) berechne $B_n \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit nach einer Aufdatierungs–Formel $(s, y, B) \longmapsto B_+(s, y, B)$,

$$B_n \;=\; B_+(s_a, y_a, B_a) \qquad \text{mit} \qquad s_a := x_n - x_a \qquad \text{und} \qquad y_a := g_n - g_a \;.$$

Wiederhole (1) mit $(x_a, g_a, B_a) \;:=\; (x_n, g_n, B_n)$.

□

Eigenschaften 'guter' Aufdatierungsformeln:

(6.2.4) (*) *Quasi–Newton–Gleichung* $B_+ s_a \;=\; y_a$ .

**(6.2.4) Bemerkung** (Motivation für die Quasi–Newton–Gleichung)
(a) Für quadratisches $f$, $f(x) = c_0 + c^t x + \frac{1}{2} x^t Q x$ mit symmetrischem $Q$, erfüllt $H_f(x_n)$ die Quasi–Newton–Gleichung.

(b) Sei $B_+$ symmetrisch und positiv definit und erfülle die Quasi–Newton–Gleichung. Dann erfüllt die neue Modellierung $\widetilde{f_n}$ von $f$,

$$\widetilde{f_n}(z) \;:=\; f(x_n) + \nabla f(x_n)(z - x_n) + \frac{1}{2}(z - x_n)^t B_+(z - x_n) \;,$$

die Interpolationsbedingungen

$$\widetilde{f}_n(x_n) \;=\; f(x_n) \quad , \quad \nabla\widetilde{f}_n(x_n) \;=\; \nabla f(x_n) \quad , \quad \nabla\widetilde{f}_n(x_a) \;=\; \nabla f(x_a) \quad .$$

(c) Nit den Bezeichnungen von (6.2.3)(5) gilt: Eine notwendige Bedingung für die Existenz eines $B_+$ mit den Eigenschaften wie in (b) – nämlich $B_+$ symmetrisch, positiv definit und $B_+s_a \;=\; y_a$ mit $s_a \neq 0$ (dies ist gerade der 'Nicht–Abbruch–Fall') – ist

$$y_a{}^t s_a \;>\; 0 \; .$$

Beweis:

(a) $f(x) \equiv q(x) = c_0 + c^t x + \frac{1}{2} x^t Q x \implies \nabla f(x) = c + Qx \implies H_f(x) = Q$ .

$y_a :\equiv \nabla f(x_n) - \nabla f(x_a) = Q(x_n - x_a) = H_f(x_n)s_a$ .

(b) $\nabla\widetilde{f}_n(z)\Big|_{x_a} = g_n + B_+(z - x_n)|_{x_a} = g_n - B_+s_a = g_n - y_a = g_a$ .

(c) $0 < s_a{}^t B_+ s_a = s_a{}^t y_a$ nach der Quasi–Newton–Gleichung.

□

*Die* Aufdatierungsformel ist die von B*royden*, F*letcher*, G*oldfarb* und S*hanno* 1970 unabhängig gefundene BFGS– *Aufdatieungsformel*

$$B_+(s, y, B) \;=\; B - \frac{Bss^tB}{s^tBs} + \frac{yy^t}{y^ts} \; .$$

$B_+$ entsteht aus $B$ also durch eine additive Rang 2 – Modifikation.

**(6.2.5) Eigenschaften der BFGS–Formel**

(a) $B_+(s, y, B)\, s = y$ .

(b) $B$ symmetrisch $\Longrightarrow B_+$ symmetrisch .

(c) $B$ symmetrisch, positiv definit , $y^ts > 0 \Longrightarrow B_+$ symmetrisch, positiv definit .

Beweis:

(a) und (b) sind klar. Den Nachweis von (c) führen wir 'algorithmisch', so daß sich dabei auch gleich eine 'billige' Updating–Prozedur für die Cholesky–Zerlegung von $B_+$ ergibt, vgl. (6.2.6).

Sei $B = LL^t$ mit $L_{ii} > 0, 1 \leq i \leq n$, die Cholesky–Zerlegung von $B$. Setze

$$w \;:=\; \left(\frac{y^ts}{s^tBs}\right)^{1/2} L^ts \quad , \quad J_+ \;:=\; L \;+\; \frac{(y - Lw)w^t}{w^tw} \; .$$

Dann ist $J_+$ nach dem folgenden Hilfssatz (6.2.5') nichtsingulär, und es gilt

$J_+J_+{}^t \;=\; B_+(s, y, B)$ , denn:

$$\left( L \;+\; \frac{(y - Lw)w^t}{w^tw} \right) \left( L^t \;+\; \frac{w(y - Lw)^t}{w^tw} \right) \;=$$

$$LL^t \;+\; \frac{1}{w^tw}\Big[(y - Lw)(Lw)^t + Lw(y - Lw)^t\Big] \;+\; \frac{1}{(w^tw)^{\;2}}(y - Lw)w^tw(y - Lw)^t \;=$$

$$B \;+\; \frac{1}{w^t w}\Big[(y-Lw)(Lw+y-Lw)^t + Lw(y-Lw)^t\Big] \;=$$

$$B \;+\; \frac{1}{w^t w}\Big[yy^t - Lw(Lw)^t\Big] \;=\; B_+(s,y,B) \quad \text{, denn}$$

$$w^t w \;=\; \frac{y^t s}{s^t B s} s^t L L^t s \;=\; y^t s \;\; \text{wegen} \;\; LL^t = B \;,$$

und

$$Lw(Lw)^t \Big/ \frac{1}{w^t w} \;=\; \frac{y^t s}{s^t B s} B s s^t B \Big/ \frac{y^t s}{s^t B s} s^t B s \;=\; \frac{B s s^t B}{s^t B s} \;.$$

$\diamond$

Die positive Definitheit von $B_+(s,y,B)$ folgt aus der Nichtsingularität von $J_+$.

$\square$

**(6.2.5') Hilfssatz**
Sei $A \in \mathbb{R}^{n\times n}$ nichtsingulär. Dann ist $A + uv^t$ nichtsingulär, falls $\sigma := 1 + v^t A^{-1} u \neq 0$ ist.

<u>Beweis:</u>

Sei $B := A + uv^t$ . Nach Voraussetzung existiert die Matrix $B^- := A^{-1} - \frac{1}{\sigma} A^{-1} u v^t A^{-1}$ . Durch Ausmultiplizieren erhält man $B^- B = E$ , d.h. $B^- = B^{-1}$ .

$\square$

**(6.2.6) Bemerkung**
Sei $B = R^t R$, $R$ obere Dreiecksmatrix, und $B_+ = (R + uv^t)^t (R + uv^t)$. Mit $(n-1)$ geeigneten Jacobi–Rotationen $J_{k-1,k}$ in der $span(\{e_{k-1}, e_k\})$–Ebene zur Annulierung des aktuellen $u_k$ erhält man $J_{1,2} * \ldots * J_{n-1,n} * (R + uv^t) = \widetilde{R}_H$ mit einer oberen Hessenbergmatrix $\widetilde{R}_H$. Weitere $(n-1)$ geeignete Jacobi–Rotationen ergeben
$\widetilde{J}_{n-1,n} * \ldots * \widetilde{J}_{1,2} * \widetilde{R}_H = R_+$ mit einer oberen Dreiecksmatrix $R_+$. Der benötigte Aufwand ist $O(n^n)$.

$\square$

Die positive Definitheit von $B_a$ impliziert offensichtlich, daß $s_a$ in (6.2.3)(2) Abstiegsrichtung ist. Wird in (6.2.3)(5) $B_n$ nach der BFGS–Formel berechnet, so ist nach (6.2.5)(c) $B_n$ wieder positiv definit – und damit $s_n$ wieder Abstiegsrichtung –, falls ${y_a}^t s_a > 0$ gilt. Dies ist aber, wie man sich leicht überlegt, bei einer (genügend) exakten Suche für die Schrittweite $t_a$ in (6.2.3)(3) sicher erfüllt.
Insgesamt erhält man mit der *BFGS*–Formel also ein immer durchführbares Quasi–Newton–Verfahren. Die lokale Konvergenzordnung ist superlinear, wie man – allerdings viel aufwändiger – ebenfalls zeigen kann.

# Kapitel 7

# Optimierung unter Nebenbedingungen

## 7.1 Lineare Optimierung; Simplex-Verfahren

Es gibt mehrere Standardaufgabenstellungen, die ineinander überführbar oder äquivalent sind. Ich verwende die folgenden

**Bezeichnungen**
Für $y,\ z \in \mathbb{R}^n$ ist $y \geq z$ , falls $y_i \geq z_i$ , $1 \leq i \leq n$.

Für $A \in \mathbb{R}^{m\times n}$ ist $A_{i\cdot} = [A_{i1}\ A_{i2} \ldots A_{in}] \in \mathbb{R}^{1\times n}$ und $A_{\cdot j} = \begin{bmatrix} A_{1j} \\ \vdots \\ A_{mj} \end{bmatrix} \in \mathbb{R}^{m\times 1}$, □

und betrachte die folgende Standardaufgabenstellung:

**(7.1.1) Lineare Optimierungsaufgabe**
Gegeben sei $c \in \mathbb{R}^n$ , $A \in \mathbb{R}^{m\times n}$ , $m < n, b \in \mathbb{R}^m$ . Gesucht wird
$x^\star \in arg\ \min\left\{c^t x \mid x \in \mathbb{R}^n \ : \ Ax = b\ ,\ x \geq 0\ \right\}$ , d.h.
$x^\star \in \mathbb{R}^n\ : Ax^\star = b, x^\star \geq 0$, so, daß für $x \in \mathbb{R}^n\ : Ax = b, x \geq 0$, gilt: $c^t x^\star \leq c^t x$. □

Die *Nebenbedingungen*

$Ax = b$ heißen *'Gleichungsnebenbedingungen'*
$x \geq 0$ sind *'Vorzeichenbedingungen'*, also spezielle *'Ungleichungsnebenbedingungen'.*

$x$ heißt *zulässig* , wenn $x$ die Nebenbedingungen $Ax = b$ , $x \geq 0$ erfüllt.

Die Nebenbedingungen bewirken i.a., daß die

*lineare Zielfunktion* $x \longrightarrow c^t x$ nach unten beschränkt

ist, und damit ein *Minimum existiert.*

**(7.1.2) Beispiel** (kleines Rechenbeispiel: $n = 5,\ m = 2$;)
Minimiere (die Zielfunktion)

$$3x_1 + x_2 + 9x_3 + x_4$$

unter den Nebenbedingungen $x_i \geq 0,\ 1 \leq i \leq 5$, und

$$\begin{array}{cccccccc} x_1 & +x_2 & +2x_3 & & & = & 4\ , \\ & -x_2 & +x_3 & +x_4 & -x_5 & = & 2\ . \end{array}$$

□

**Bemerkung**
(a) Die Nebenbedingungen $x_i \geq 0,\ 1 \leq i \leq 4$, und

$$\begin{array}{ccccccc} x_1 & +x_2 & +2x_3 & & = & 4\ , \\ & -x_2 & +x_3 & +x_4 & \geq & 2\ . \end{array}$$

sind (unter Verwendung von $x_5 := -x_2+x_3+x_4-2 \geq 0$ , sog. 'künstliche' oder 'Schlupf'variablen) äquivalent zu den Nebenbedingungen $x_i \geq 0,\ 1 \leq i \leq 5$, und

$$\begin{array}{cccccccc} x_1 & +x_2 & +2x_3 & & & = & 4\ , \\ & -x_2 & +x_3 & +x_4 & -x_5 & = & 2\ . \end{array}$$

(b) Analog kann man die Gleichung $A_i.\ x = b_i$ auflösen in zwei Ungleichungen

$$A_i.\ x \leq b_i \quad , \quad -A_i.\ x \leq -b_i\ .$$

(c)

$$y_i \text{ beliebig } \in \mathbb{R} \Longleftrightarrow y_i = (y_i)_+ - (y_i)_- \ ,\ (y_i)_+, (y_i)_- \geq 0\ .$$

□

**(7.1.3) Kostenoptimale Herstellung** ( von $m$ Gütern in $n$ Fabriken )
unter folgenden Modellannahmen:
– die $j$-te Fabrik erzeugt beim Aktivitätslevel $x_j \geq 0$ das $x_j$-fache von $A_{ij}$, der 'Erzeugung pro Einheits-Aktivität' des $i$-ten Produkts in der Fabrik $j$;
– die 'Kosten pro Einheitsaktivität' der $j$-ten Fabrik betragen $c_j$;
– Die geforderte Menge des $i$-ten Produkts ist $b_i$;
– Kosten und Erzeugung sind proprtional zum Aktivitätslevel.
Also ergibt sich folgende Aufgabe:

$$x = [x_j]_{1\leq j\leq n} \in \mathbb{R}^n \ ,\ x \geq 0\ ;\quad \sum_{j=1}^{n} A_{ij}\ x_j \ = \ b_i\ ,\ 1 \leq i \leq m\ ;\quad \sum_{j=1}^{n} c_j x_j \ \longrightarrow min\ !$$

□

Die Rolle, die das *Gaußche Eliminationsverfahren* bei linearen Gleichungssystemen spielt, hat das *Simplexverfahren* von *G. Dantzig* (1947/48) in der linearen Optimierung.

*Anschaulich:* Simplexverfahren ist *iterativer Abstieg von Ecke zu Ecke längs Kanten* eines Polyeders unter Verminderung des Zielfunktionswertes. Wie man zeigen kann, wird der Minimalwert – wenn er existiert – auch in einer Ecke angenommen.

□

*Generelle Voraussetzung* ist

$$(7.1.4)(i) \quad Ax = b\ ;\ A \in \mathbb{R}^{m\times n} \ ,\ rang(A) = m \ < \ n\ .$$

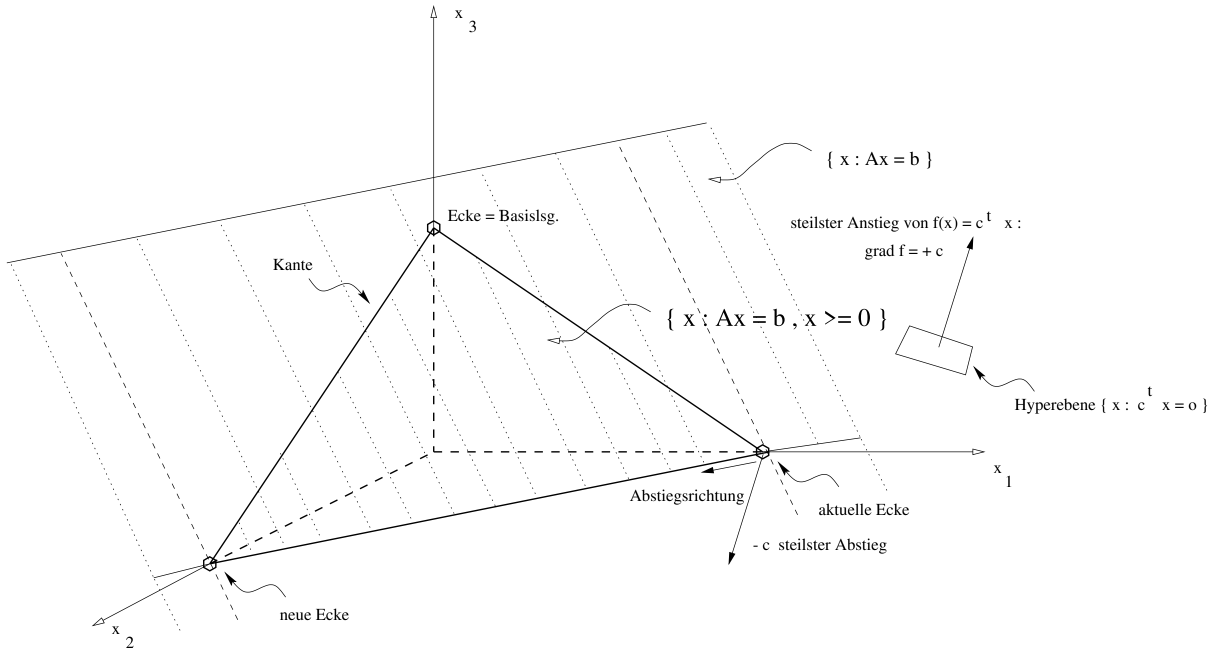


Abbildung 7.1: Simplexaustauschschritt

**(7.1.4) Definition** (Basislösung/Ecke)
$x \in \mathbb{R}^n$ heißt *Basislösung* oder *Ecke* für Problem (7.1.1) unter der Voraussetzung (7.1.4)(i), wenn es eine Permutation $\sigma \in \mathcal{S}_n$ gibt, so daß gilt:

$$Ax = b \quad , \quad x \geq 0 \quad ; \quad x_{\sigma(i)} = 0, m+1 \leq i \leq n \ ; \qquad \text{und}$$

$$\left[ \ A_{\cdot\sigma(1)} \ \middle| \ A_{\cdot\sigma(2)} \ \middle| \ \dots \ \middle| \ A_{\cdot\sigma(m)} \ \right] \in \mathbb{R}^{m\times m} \quad \text{nichtsingulär} \ .$$

□

$x_{\sigma(1)}, x_{\sigma(2)}, \dots, x_{\sigma(m)}$ heißen *Basisvariablen*, $x_{\sigma(i)}, m+1 \leq i \leq n$ heißen *Nichtbasis-Variablen* (bezüglich der betrachteten Permutation $\sigma$).
Eine *Ecke* $x$ von $M$ ist eigentlich geometrisch definiert durch die Bedingung, daß $x$ nicht im relativen Inneren irgendeiner Strecke von $M$ liegen kann. Formal nach Definition (7.1.4) ist eine Ecke $x$ festgelegt durch die Auswahl der $m$ linear unabhängigen Spalten $A_{\cdot\sigma(1)} \dots A_{\cdot\sigma(m)}$ von $A$, d.h. durch die Permutation $\sigma$. Bei *gegebenem* $\sigma$ ordne ich zur Vereinfachung der Notation – unter Unterdrückung der Abhängigkeit von $\sigma$ – alle Größen so, daß die Basis-Variablen an den Stellen $1, 2, \dots m$ stehen:

**(7.1.5) Notation** (bei gegebener Permutation $\sigma$)

$$A\ P_\sigma = \left[ \ B \ \middle| \ N \ \right] \ \text{mit} \ B \equiv [B_1|\cdots\cdots|B_m] = [A_{\cdot\sigma(1)}|\cdots|A_{\cdot\sigma(m)}] \ \text{und}$$
$$N \equiv [N_{m+1}|\cdots|N_n] = [A_{\cdot\sigma(m+1)}|\cdots|A_{\cdot\sigma(n)}] \ .$$

Analog ist

$$x^t\ P_\sigma = \left[ \ {x_B}^t \ \middle| \ 0^t \ \right] \quad \text{bzw.} \quad {P_\sigma}^t\ x = \begin{bmatrix} x_B \\ \hline 0 \end{bmatrix} ,$$

und allgemein für $y \in \mathbb{R}^n$ $\quad y^t\, P_\sigma =: \left[\begin{array}{c|c} {y_B}^t & {y_N}^t \end{array}\right]$ bzw. ${P_\sigma}^t\, y =: \left[\begin{array}{c} y_B \\ \hline y_N \end{array}\right]$ . □

Damit kann man alle Größen sehr kompakt in den entsprechend der jeweiligen Permutation *partitionierten Variablen* schreiben, etwa:

$$Ay = \sum_{j=1}^n y_{\sigma(j)} A_{\cdot\sigma(j)} = A\,\underbrace{(P_\sigma {P_\sigma}^t)}_{=E_n}\,y = \left[\begin{array}{c|c} B & N \end{array}\right] \left[\begin{array}{c} y_B \\ \hline y_N \end{array}\right] = By_B + Ny_N\ , \text{ und}$$

$$c^t y = \sum_{j=1}^n c_{\sigma(j)} y_{\sigma(j)} = (c^t P_\sigma)(y^t P_\sigma)^t = \left[\begin{array}{c|c} {c_B}^t & {c_N}^t \end{array}\right] \left[\begin{array}{c} y_B \\ \hline y_N \end{array}\right] = {c_B}^t y_B + {c_N}^t y_N\ .$$

**(7.1.6) Simplex-Verfahren** (sog. *revidierte Form*)
*Gegeben:* $A \in \mathbb{R}^{m\times n}$ , $m < n$ , $rang(A) = m$ ; $b \in \mathbb{R}^m$ ; $c \in \mathbb{R}^n$ .

*Gesucht:* $x^\star \in \mathbb{R}^n \ :\ x^\star \in arg\ \min\left\{c^t x \mid x \in \mathbb{R}^n \ :\ Ax = b\ ,\ x \geq 0\ \right\}$ .

(1) *Vorbereitungsschritt:* Bestimmung einer (Start–) Basislösung $x$ mit zugehörigem $P_\sigma$ .

(2) *Optimalitätstest:* Gegeben Basislösung $x$ , so daß mit zugehörigem $P_\sigma$

$$B\ x_B = b\,; \quad x \ =\ P_\sigma \left[\begin{array}{c} x_B \\ \hline 0 \end{array}\right]\ ; \quad A\ P_\sigma\ =\ \left[\begin{array}{c|c} B & N \end{array}\right]\ .$$

*Falls* $r_N \ \equiv\ [r_{N\sigma(k)} \mid m+1 \leq k \leq n] \ :=\ c_N - N^t B^{-t} c_B \ \geq\ 0$ , ist
$x \ :\ x_{\sigma(i)} = x_B(i),\ 1 \leq i \leq m\ ,\ x_{\sigma(i)} = 0,\ m+1 \leq i \leq n$ , *eine Lösung* $x^\star$ ;

*sonst* bestimme Abstiegsstrahl $z = \sum_{j=1}^m z_{\sigma(j)} e_{\sigma(j)} \ +\ \zeta\ e_{\sigma(i)} \ :\ Az = b\ ,\ \zeta > 0$, wobei
$i \ \in\ arg\ \min\ \left\{\ r_{N\sigma(k)} \ \mid\ k \in \{m+1, ..., n\}\ \right\}$ ;

(3) *Beschränktheitstest: Falls* $v \ :=\ B^{-1} N_{\cdot\, i} \ \leq\ 0$ , dann
$\inf\ \{c^t y \ \mid\ Ay = b,\ y \geq 0\} \ =\ -\infty$ ; d.h. das Problem hat *keine Lösung* ;
*anderenfalls* bestimme neue Basislösung durch
$j \ \in\ arg\ \min\ \left\{\ \frac{(x_B)_k}{v_k} \ \mid\ k \in \{1, ..., m\} \ :\ v_k > 0\ \right\}$ ;

*Falls* $\min\ \left\{\ \frac{(x_B)_k}{v_k} \ \mid\ k \in \{1, ..., m\} \ :\ v_k > 0\ \right\} \ =\ 0$ , ist $x$ sog.*'entartete' Basislösung* und erfordert *Spezialbehandlung* vgl. (7.1.11).

*Falls* $\min\ \left\{\ \frac{(x_B)_k}{v_k} \ \mid\ k \in \{1, ..., m\} \ :\ v_k > 0\ \right\} \ >\ 0$ , dann

(4) *Austausch:* von $A_{\bullet\sigma(j)}$ durch $A_{\bullet\sigma(i)}$ , bzw. Ersetzung der aktuellen Basisvariablen $x_{\sigma(j)}$ durch die *neue* Basisvariable $x_{\sigma(i)}$ d.h. mit $(i\ j)(i) := j$ , $(i\ j)(j) := i$ , und $(i\ j)(k) := k$ *sonst*, gilt

$$\sigma_+ \;=\; \sigma \circ (i\ j)\ .$$

*Bem.:* Zur Aufdatierung einer Faktorisierung von $B$ zu einer Faktorisierung von $B_+$ ist es günstiger, zuerst durch zyklische Vertauschung die *j-te Spalte von* $B$ als *letzte Spalte* zu haben, d.h.

$$\sigma_+ \;=\; \sigma \circ (i\ m) \circ (m-1\ m) \circ \cdots\ (j+1\ j+2) \circ (j\ j+1)\ .$$

(5) *Bestimmung* der zu $P_{\sigma_+}$ gehörigen *neuen Basislösung*:

$$B_+ \;=\; \left[\ A_{\bullet\sigma(1)}\ \middle|\ \ldots\ \middle|\ A_{\bullet\sigma(j-1)}\ \middle|\ A_{\bullet\sigma(i)}\ \middle|\ A_{\bullet\sigma(j+1)}\ \middle|\ \ldots\ \middle|\ A_{\bullet\sigma(m)}\ \right]\ ;$$

$$B_+\, x_{B_+} \;=\; b\ ; \quad x_+ \;=\; P_{\sigma_+} \left[\begin{array}{c} {B_+}^{-1} b \\ \hline 0 \end{array}\right]\ .$$

Mit $(\ P_\sigma\ ,\ B\ ,\ x\ ) := (\ P_{\sigma_+}\ ,\ B_+\ ,\ x_+\ )$ wiederhole die Schritte (1) – (4) . □

Einige Bemerkungen zu den einzelnen Schritten dieses Algorithmus.

**Bemerkung** (zu Schritt (0))
Äquivalent zu Nebenbedingungen der Form

$$Ay \;\leq\; b\ ,\ y \;\geq\; 0\ ;$$

sind mit $z_i := b_i - \sum_j A_{ij}\ y_j$ , $1 \leq i \leq m$ die Nebenbedingungen

$$Ay \;+\; z \;=\; b\ ,\ \left[\ y^t\ \middle|\ z^t\ \right]^t \;\geq\; 0\ .$$

Gilt $b \;\geq\; 0$ , so ist $\left[\ 0^t\ \middle|\ b^t\ \right]^t \;=:\; x \;\geq\; 0$ eine (zulässige) Startbasislösung für das Simplexverfahren mit Gleichheitsnebenbedingungen, d.h. mit $\sigma(j) = n + j,\ 1 \leq j \leq m$, $\sigma(m+i) = i,\ 1 \leq i \leq n$, liefert $\sigma$ eine zulässige Basislösung. □

Im allgemeinen Fall führt eine ähnliche *Methode der 'künstlichen' (= zusätzlichen) Variablen* zu dem Ziel eine Startbasislösung zu finden. Dazu beachte man, daß bei *Gleichheits*nebenbedingungen nach *eventueller Multiplikation der i–ten Gleichung mit* $-1$ vorausgesetzt werden darf, daß $b \;\geq\; 0$ gilt:

$$\sum_j A_{ij}\ y_j \;=\; b_i\ ,\ b_i \;<\; 0 \iff \sum_j -A_{ij}\ y_j \;=\; -b_i\ ,\ (-\ b_i) \;>\; 0\ .$$

**(7.1.7) Bemerkung** ( zu Schritt (0))
Mit $A \in \mathbb{R}^{m\times n}$ , $m < n$, und $b \geq 0$ sei

$$(i) \quad Ay = b\ ,\ y \geq 0\ ;$$

und mit den zusätzlichen Variablen $z \in \mathbb{R}^m$ sei

$$(ii) \quad \text{minimiere} \quad \sum_{i=1}^{m} z_i \quad \text{unter den Nbd.'n} \quad Ay + z = b \ , \ y \geq 0 \ , \ z \geq 0 \ .$$

Dann gilt

(a) Hat $(i)$ zulässige Punkte, so liefert jede optimale (Basis-)Lösung $\left[\ \widetilde{y}^t \ \middle| \ \widetilde{z}^t \ \right]^t$ von $(ii)$ – mit notwendigerweise $\widetilde{z} = 0$ – einen zulässigen Punkt $\widetilde{y}$ von $(i)$ . Trivial ist natürlich $\left[\ 0^t \ \middle| \ b^t \ \right]^t$ eine Startbasislösung für $(ii)$.

(b) Eventuell in $B$ enthaltene Spalten von $E$, die zu Basisvariablen $z_i = 0$ gehören, muß man nun noch gegen linear unabhängige Spalten $A_{\cdot j}$ mit $y_j = 0$ tauschen, und erhält eine (dann allerdings entartete) Ecke der Ausgangsaufgabe (dieser Austausch versagt für $rang(A) < m$, dann 'Zeilen streichen').

(c) $(i)$ hat keine zulässige (Basis-)Lösung genau dann, wenn in $(ii)$ gilt

$$\min \left\{ \sum_{i=1}^{m} z_i \ \middle| \ Ay + z = b \ , \ y \geq 0 \ , \ z \geq 0 \right\} \ > \ 0 \ .$$

<u>Beweis:</u>

(a) Sei $y$ zulässig in $(i) \Longrightarrow \left[\ y^t \ \middle| \ 0^t \ \right]^t$ ist trivialerweise optimal für $(ii)$ . Also gilt

$$\min \left\{ \sum_{i=1}^{m} z_i \ \middle| \ Ay + z = b \ , \ y \geq 0 \ , \ z \geq 0 \right\} \ = \ 0$$

Damit gilt für jede *optimale* (Basis-)Lösung $\left[\ \widetilde{y}^t \ \middle| \ \widetilde{z}^t \ \right]^t$ von $(ii)$, daß$\widetilde{z} = 0$ gilt. Damit ist $\widetilde{y}$ *zulässige* (Basis-)Lösung von $(i)$. Wegen $b \geq 0$ ist $x := \left[\ 0^t \ \middle| \ b^t \ \right]^t$ (Start-)Basislösung für die 'erweiterte' Minimierungsaufgabe.

$\diamond$

(c) $x := \left[\ 0^t \ \middle| \ b^t \ \right]^t$ ist Basislösung, d.h. die Menge der zulässigen Punkte für $(ii)$ ist nichtleer. Wegen

$$\inf \left\{ \sum_{i=1}^{m} z_i \ \middle| \ Ay + z = b \ , \ y \geq 0 \ , \ z \geq 0 \right\} \ \geq \ 0$$

ist die Zielfunktion beschränkt. Damit liefert der Simplex-Algorithmus eine *optimale* Basislösung, d.h. das Infimum wird angenommen. Wäre

$$\min \left\{ \sum_{i=1}^{m} z_i \ \middle| \ Ay + z = b \ , \ y \geq 0 \ , \ z \geq 0 \right\} \ = \ 0 \ ;$$

$\Longrightarrow$ jede zugehörige optimale Basislösung liefert in der ersten Blockkomponente einen zulässigen Punkt von $(i)$ – im Widerspruch zur Voraussetzung.

$\square$

**Bemerkung**
Die Lösung des Optimierungsproblems (7.1.7)($ii$) mit Hilfe des Simplexverfahrens heißt *Phase I* zur Durchführung des Simplexverfahrensschrittes (7.1.6)(0) . □

Analyse der restlichen Schritte durch Entwicklung von $z$ um die aktuelle Ecke $x$ mit

$$x \ : \ {P_\sigma}^t \ x = \left[\begin{array}{c} x_B \\ \hline 0 \end{array}\right] \ ; \ x_B = B^{-1}b \ , \ A \ P_\sigma = \left[\begin{array}{c|c} B & N \end{array}\right] \ .$$

**(7.1.8) Bemerkung**
Analog zu obigen Bezeichnungen sei für $z \ : \ Az = b$ mit gegebener Permutation $\sigma \in S_n$

$${P_\sigma}^t \ z = \left[\begin{array}{c} z_B \\ \hline z_N \end{array}\right] \ , \ {P_\sigma}^t \ c = \left[\begin{array}{c} c_B \\ \hline c_N \end{array}\right] \ .$$

Dann gilt

$$c^t z \ = \ c^t x \ + \ (c_N \ - \ N^t B^{-t} c_B)^t z_N \ .$$

<u>Beweis:</u>

$$b \ = \ Bz_B + Nz_N \quad \Longrightarrow \quad z_B = B^{-1}b - B^{-1}Nz_N \ ;$$

Eingesetzt in $c^t z = {c_B}^t z_B + {c_N}^t z_N$ folgt

$$\begin{aligned} c^t z \ &= \ \underbrace{{c_B}^t B^{-1} b} \quad -{c_B}^t B^{-1} N z_N + {c_N}^t z_N \\ &= \ c^t x, \ \text{da } x_N = 0 \\ &= \ c^t x \ + \ (c_N \ - \ N^t B^{-t} c_B)^t z_N \ . \end{aligned}$$

□

Wegen $z_N \geq 0$ ist also $x$ optimal, falls $r_N \ := \ (c_N \ - \ N^t B^{-t} c_B) \ \geq \ 0$ ist.
Betrachten wir jetzt im Nicht-Optimalitätsfall $z(\zeta)_N \ := \ \zeta e_{\sigma(i)} \ , \ \zeta \geq 0 \ ,$ und

$${P_\sigma}^t z(\zeta) \ = \ \left[\begin{array}{c} z(\zeta)_B \\ \hline z(\zeta)_N \end{array}\right] \quad \text{mit} \quad Az(\zeta) \ = \ b \ .$$

$\Longrightarrow$

$$b = Bz(\zeta)_B + Nz(\zeta)_N = Bz(\zeta)_B + \zeta N._i$$

$\Longrightarrow$

$$z(\zeta)_B = \underbrace{B^{-1}b}_{= x_B} - \zeta \underbrace{B^{-1}N._i}_{= v}$$

und nach (7.1.8)

$$c^t z(\zeta) = c^t x + \zeta r_{N\sigma(i)}$$

mit

$${r_N}^t = (c_N - N^t B^{-t} c_B)^t \ , \ z_N = \zeta e_{\sigma(i)}.$$

Damit folgt

**(7.1.9) Hilfssatz**
Mit diesen Bezeichnungen und Voraussetzungen gilt

$$z(\zeta) \geq 0 \iff \zeta \;\leq\; \frac{(x_B)_k}{v_k} \quad \text{für alle } k \text{ mit } v_k > 0\ .$$

□

**(7.1.10) Folgerung**
Mit den Bezeichnungen und Voraussetzungen des Simplexverfahrens (7.1.6), und $z(\zeta)$ wie oben, gilt:

(a) $r_N := (c_N - N^t B^{-t} c_B) \geq 0 \Longrightarrow x$ optimal

(b) $v := B^{-1} N_{\cdot i} \;\leq\; 0 \Longrightarrow z(\zeta) \;\geq 0$ für alle $\zeta \geq 0$.

(c) $r_N \;\not\geq\; 0$ und $v \;\leq\; 0 \Longrightarrow \inf\,\{c^t z \mid z \leq 0, Az = b\} \;=\; -\infty$ .

(d) $v \;\not\leq\; 0 \Longrightarrow B_+$ nichtsingulär.

(e) $v \;\not\leq\; 0 \Longrightarrow z(\zeta)$ zulässig für alle $\zeta : 0 \leq \zeta \leq \zeta^* := \min\Big\{(x_B)_k / v_k \;:\; v_k > 0\Big\}$;
$0 < \zeta^* = (x_B)_j / v_j$ nach Wahl von $j$;

(f) $r_N \;\not\geq\; 0\,,\, v \;\not\leq\; 0$, so gilt $c^t x_+ \;\leq\; c^t x$ . Ist $x$ nicht entartet, so gilt $c^t x_+ \;<\; c^t x$ .

(g) $x_+ = z(\zeta^*)$ .

<u>Beweis:</u>

(a) folgt aus (7.1.8), (b) und (e) aus (7.1.9).

◇

(c) $$\inf_{Az=b\ ,\ z\geq 0} c^t z \;\overset{(b)}{\leq}\; \inf_{\zeta\geq 0} c^t z(\zeta) \;\overset{s.o.}{=}\; c^t x \;+\; \inf_{\zeta\geq 0}\; \zeta\; r_{N\sigma(i)} \;\overset{r_{N\sigma(\underline{i})}<0}{=}\; -\infty\ .$$

◇

(d)

$$\begin{aligned} B_+ &= \Big[\; B_{\cdot 1} \;\Big|\; \ldots \;\Big|\; B_{\cdot j-1} \;\Big|\; N_{\cdot i} \;\Big|\; B_{\cdot j+1} \;\Big|\; \ldots \;\Big|\; B_{\cdot m} \;\Big] \\ &= \Big[\; B_{\cdot 1} \;\Big|\; \ldots \;\Big|\; B_{\cdot j-1} \;\Big|\; Bv \;\Big|\; B_{\cdot j+1} \;\Big|\; \ldots \;\Big|\; B_{\cdot m} \;\Big] \\ &= B\; \Big[\; e_1 \;\Big|\; \ldots \;\Big|\; e_{j-1} \;\Big|\; v \;\Big|\; e_{j+1} \;\Big|\; \ldots \;\Big|\; e_m \;\Big]\ . \end{aligned}$$

$B$ ist nach Vor. nichtsingulär; wegen $v_j \;>\; 0$ ist auch $\Big[\; e_1 \;\Big|\; \ldots \;\Big|\; e_{j-1} \;\Big|\; v \;\Big|\; e_{j+1} \;\Big|\; \ldots \;\Big|\; e_m \;\Big]$ nichtsingulär.

◇

(g) $x_+ \;=\; z(\zeta^*)$ , denn

$$\big(z(\zeta^*)_N\big)_k \;=\; 0\,,\; k \neq i\ ;\; \big(z(\zeta^*)_N\big)_j \;\overset{s.o.}{=}\; (x_B - \zeta^* v)_j \;=\; (x_B)_j \;-\; \frac{(x_B)_j}{v_j} \cdot v_j \;=\; 0\ .$$

◇

Damit folgt (f), denn

$c^t x_+ \;=\; c^t z(\zeta^*) \;\overset{s.o.}{=}\; c^t x + \zeta^* r_{N\sigma(i)} \;\leq\; c^t x$ , da $r_{N\sigma(i)} < 0$ nach Wahl von $i$ , und $\zeta^* \geq 0$ ;

und

$c^t x_+ \;<\; c^t x$ wegen $\zeta^* \;>\; 0$ , wenn $x$ nicht entartet ist .

$\square$

**Bemerkung**

$$B_+ \;=\; B \left[\; e_1 \;\middle|\; \ldots \;\middle|\; e_{j-1} \;\middle|\; v \;\middle|\; e_{j+1} \;\middle|\; \ldots \;\middle|\; e_m \;\right]$$

$\Longrightarrow$

$$\begin{aligned} B_+^{-1} \;&=\; \left[\; e_1 \;\middle|\; \ldots \;\middle|\; e_{j-1} \;\middle|\; v \;\middle|\; e_{j+1} \;\middle|\; \ldots \;\middle|\; e_m \;\right]^{-1} B^{-1} \\ &=\; \left[\; e_1 \;\middle|\; \ldots \;\middle|\; e_{j-1} \;\middle|\; -\frac{1}{v_j}\, v \;\middle|\; e_{j+1} \;\middle|\; \ldots \;\middle|\; e_m \;\right] * B^{-1} \end{aligned}$$

sog. *Produktform* der Inversen.

$\square$

**Bemerkung**

Für $\zeta^* \;:=\; (x_B)_j / v_j$ und $j \in arg\ min \left\{ (x_B)_k / v_k \;\mid\; k \in \{1,..,m\} : v_k > 0 \right\}$ ist $z(\zeta^*)$ zulässig , und es gilt

$$\left( P_\sigma{}^t z(\zeta^*) \right)_k \;=\; 0 \;, \quad k \in \left( \{m+1,..,n\} \cup \{j\} \right) \setminus \{i\} \;.$$

$\square$

**(7.1.11) Bemerkung**

Falls $min \left\{ \frac{(x_B)_k}{v_k} \;\mid\; k \in \{1,...,m\} \;:\; v_k > 0 \right\} \;=\; 0$ , gilt nach der Austauschvorschrift (3) in (7.1.6) $x_{neu} \;=\; x_{alt}$ . Diese sog. *'entartete' Basislösung* erfordert *Spezialbehandlung*: Man muß verhindern, daß der Algorithmus in einer Ecke *kreist* :

$$\textit{Austausch} \quad i_1 \rightsquigarrow i_2 \;;\; i_2 \rightsquigarrow i_3 \;;\; \ldots\; i_{l-1} \rightsquigarrow i_l \;;\; i_l \rightsquigarrow i_1 \;!$$

$\square$

Es gibt (konstruierte) Beispiele für das tatsächliche Auftreten solcher Zyklen. Verhindern kann man *Zyklen* z.B. durch

- Markieren bereits bearbeiteter Spaltenindizes, und modifizierte Bestimmung des Abstiegsstrahls $z = \sum_{j=1}^m z_{\sigma(j)} e_{\sigma(j)} \;+\; \zeta\ e_{\sigma(i)} \;:\; Az = b \;,\; \zeta \geq 0$, wobei

  $i \in arg\ min \left\{ r_{N\sigma(k)} \;\mid\; k \in \{m+1,...,n\} \;,\; k\ \textit{noch nicht bearbeitet} \right\}$ ; $\quad i \quad$ *bearbeitet* .

- 'geschicktere' Austauschregeln, z.B. die 'Regel von Bland', vgl. Bland, R.G.: New finite pivoting rules for the simplex method. Math. Oper. Res. 2, 103–107 (1977);

- sog. $\varepsilon$ – Störungen .

**(7.1.12) Folgerung**
Gilt in Problem (7.1.1) $\inf\{c^t z \mid z \geq 0, Az = b\} > -\infty$ , und treten bei der Durchführung des Simplexverfahrens keine entarteten Basislösungen auf, so bricht das Simplexverfahren mit einer optimalen Basislösung ab.

Beweis:

Folgt aus (7.1.10)($f$) , da es höchstens $\binom{n}{m}$ Ecken gibt ($m$ linear unabhängige Spalten aus $n$ möglichen).

□

Daß im ungünstigsten Fall die arithmetische Komplexität des Simplex–Verfahrens exponentiell mit der Anzahl der Variablen wächst, zeigt das folgende Beispiel von *Klee* und *Minty*

**(7.1.13) Beispiel**
Minimiere $-\sum\limits_{i=1}^{m} 10^{m-i} x_i \longrightarrow min!$ unter den Nebenbedingungen $x \geq 0$ und

$$2\Big(\sum_{j=1}^{i-1} 10^{i-j} x_j\Big) + x_i + x_{m+i} = 100^{i-1} \ , \ 1 \leq i \leq m.$$

Ausgehend von der Ecke $x_N \mathrel{\widehat{=}} [x_1, ..., x_m]^t = 0^t$ , $x_B \mathrel{\widehat{=}} [x_{m+1}, ..., x_{2m}]^t = [1, 100, ..., 100^{m-1}]^t$ durchläuft der Simplex–Algorithmus (7.1.6) alle $2^m$ Ecken der Menge der zulässigen Punkte.

□

Für eine experimentelle Bestätigung dieser Tatsache vergleiche man Aufgabe 11.3.

# Anhang A

# Interpolation von Operatornormen

Klassische Fehlerabschätzungen z.Bsp. für die zusammengesetzte $m$–punktige Gauß–Quadraturformel wurden in Numerik 1 hergeleitet, falls der Integrand aus $C^{2m}$ ist. Durch Verwendung sog. 'Interpolationssätze' kann man in vielen Fällen daraus sofort entsprechende Fehlerabschätzungen ableiten auch für nicht so glatte Funktionen, vgl. die Folgerung (A.2.7).

## A.1 Grundlagen

Zur Vertiefung dieses knappen Kapitels – insbesondere bezüglich der nicht ausgeführten Beweise (und auch für die durchgeführten Beweise) – vergleiche man z.B. entweder *Butzer, P.L., H. Berens:* Semi-Groups of Operators and Approximation. Grundlehren der mathematischen Wissenschaften 145, Springer 1967, und/oder *Triebel, H.:* Interpolation Theory, Function Spaces, Differential Operators. North Holland 1978. Weitere Monographien zum Thema dieses Anhangs sind *Bergh, J., J. Löfström:* Interpolation spaces. An introduction. Grundlehren der mathematischen Wissenschaften 223, Springer 1976, und *Bennett, C., R. Sharpley:* Interpolation of operators. Academic Press 1988.

**(A.1.1) Situation**
Zwei Banachräume $X_1,\ X_2$ mit Normen $\|\ .\ \|_i := \|\ \|_{X_i},\ i = 1,\ 2$ heißen *Interpolationspaar* $\big(X_1, X_2\big)$ des Banachraumes[1] $\mathcal{X}$, wenn $X_i \hookrightarrow \mathcal{X}$ stetig eingebettet sind, $i = 1,\ 2$. □
'Stetige Einbettung' symbolisiere ich durch '$\hookrightarrow$'.

**(A.1.2) Bemerkung** ( vgl. Lemma. 2.3.1 Butzer/Berens S. 24 )
In der Situation (A.1.1) gilt:
(a) $X_1 \cap X_2$ mit $\|x\|_{X_1\cap X_2} := \max\left(\|x\|_1,\ \|x\|_2\right)$ ist ein Banachraum.
(b) $X_1 + X_2,\ \|x\|_{X_1+X_2} := \inf\{\|x_1\|_1 + \|x_2\|_2\ : x_1 + x_2 = x,\ x_i \in X_i\ ,\ i = 1, 2\ \}$ ist ein Banachraum.
(c) $X_1 \cap X_2 \hookrightarrow X_i \hookrightarrow X_1 + X_2 \hookrightarrow \mathcal{X}\ ,\ i = 1, 2$ , und alle diese Unterräume sind stetig in $\mathcal{X}$ eingebettet.

<u>Beweis:</u>

Wir zeigen für jede der einfachen Aussagen jeweils eine Eigenschaft:
(a) Dreiecksungleichung:

$$\|x + y\| \ = \ \max\ (\ \|x + y\|_1\ ,\ \|x + y\|_2\ )$$

[1] dies reicht für unsere Zwecke – allgemeiner kann $\mathcal{X}$ ein Hausdorffscher topologischer Vektorraum sein

$$\leq \max\left( \|x\|_1 + \|y\|_1 \ , \ \|x\|_2 + \|y\|_2 \right)$$
$$\leq \max\left(\|x\|_1, \|x\|_2\right) \ + \ \max\left(\|y\|_1, \|y\|_2\right) \ .$$

(b) Definitheit:
Sei $x \in X_1 + X_2 : \|x\|_{X_1+X_2} = 0$. Dann existieren $x_1^n + x_2^n, n \in \mathbb{N}$ mit $x = x_1^n + x_2^n, x_i^n \in X_i, n \in \mathbb{N}$, $\|x_1^n\|_1 + \|x_2^n\|_2 \longrightarrow 0, n \to \infty$.
$\Longrightarrow \ x_i^n \longrightarrow 0$ in $X_i$, also auch in $\mathcal{X} \quad \Longrightarrow \quad x = x_1^n + x_2^n \longrightarrow 0 + 0 = 0$ in $\mathcal{X}$, also $x = 0$.

(c) $X_1 \cap X_2 \hookrightarrow X_i = X_i + 0 \hookrightarrow X_1 + X_2$ ist klar, denn
$$\inf_{x_1+x_2=x} (\|x_1\|_1 + \|x_2\|_2) \leq \|x\|_i + \|0\| \leq \max(\|x\|_1, \|x\|_2) \ , \ i = 1, 2 \ .$$

□

$X_1 \cap X_2$ beziehungweise $X_1 + X_2$ sind der kleinste beziehungsweise größte durch $X_1$ und $X_2$ bestimmte Banachraum. Dazwischen liegen sog. *intermediäre Räume*.

**(A.1.3) Definition** (intermediäre Räume)
Ein Banachraum $X$ heißt *intermediärer Raum* des Interpolationspaares $\big(X_1, X_2\big)$, wenn gilt

$$X_1 \cap X_2 \hookrightarrow X \hookrightarrow X_1 + X_2$$

□

Intermediäre Räume kann man nach *Jaak Peetre* folgendermaßen konstruieren:

***Peetre's K-Funktional:***
Sei $\theta \in [0, 1]$. Für $t > 0$ und $x \in X_1 + X_2$ sei

(A.1.4)(i) $K(t, x) := \inf \big\{ \|x_1\|_1 + t\|x_2\|_2 \ : \ x_1 + x_2 = x; \ x_i \in X_i, \ i = 1, 2 \big\}$ ;

(A.1.4)(ii) $$\|x\|_{(X_1,X_2)_{\theta,q}} := \begin{cases} \left[ \int_0^\infty \left( t^{-\theta} K(t, x) \right)^q \frac{dt}{t} \right]^{1/q} , \ 1 \leq q < \infty \ ; \\ \\ \sup_{t>0} \ t^{-\theta} K(t, x) \ , \ q = \infty. \end{cases}$$

◇

**(A.1.4) Satz** (vgl. Prop. 3.2.5 Butzer/Berens S. 168)
Betrachtet werde die Situation (A.1.1). Mit den Bezeichnungen (A.1.4) (i) – (ii) gilt für $0 < \theta < 1$ und $1 \leq q \leq \infty$

$$(X_1, X_2)_{\theta,q} := \{ x \in X_1 + X_2 \ : \ \|x\|_{(X_1,X_2)_{\theta,q}} < \infty \}$$

ist intermediärer Banachraum zu $\big(X_1, X_2\big)$.

□

Von unmittelbarer Bedeutung für die Numerik ist der folgende Satz, der die Operatornorm bzgl. solcher intermediärer Räume abschätzt im Vergleich zum Interpolationspaar.

**(A.1.5) Satz** (Konvexitätstheorem[2])
Seien $\big(X_1, X_2\big)$ und $\big(Y_1, Y_2\big)$ Interpolationspaare und $A \ : \ X_1 + X_2 \longrightarrow Y_1 + Y_2$ ein linearer Operator mit

$$A(X_i) \subset Y_i, \quad \|Ax_i\|_{Y_i} \leq M_i \|x_i\|_{X_i} \ , \ x_i \in X_i; \ i = 1, 2.$$

Dann ist $A \ : \ (X_1, X_2)_{\theta,q} \longrightarrow (Y_1, Y_2)_{\theta,q}$ stetig, und es gilt für $0 < \theta < 1$ und $1 \leq q \leq \infty$ die Interpolationsungleichung

$$\|Ax\|_{(Y_1,Y_2)_{\theta,q}} \leq {M_1}^{1-\theta} \cdot {M_2}^{\theta} \|x\|_{(X_1,X_2)_{\theta,q}}, \ \ x \in (X_1, X_2)_{\theta,q} \ .$$

[2] Die Aussage dieses Satzes ist ja gerade, daß $\theta \longmapsto \log(\|A\|_{\theta,q})$ konvex ist.

Beweis:

(a) O.E. ist $M_1 > 0$. Dann gilt

$$\begin{aligned} K(t,Ax) &= \inf_{y_1+y_2=Ax\,,\ y_i\in Y_i} \big\{\|y_1\|_{Y_1} + t\|y_2\|_{Y_2}\big\} \\ &\overset{A(X_i)\subset Y_i}{\leq} \inf_{x_1+x_2=x\,,\ x_i\in X_i} \big\{\|Ax_1\|_{Y_1} + t\|Ax_2\|_{Y_2}\big\} \\ &\overset{n.Vor.}{\leq} M_1 \inf_{x_1+x_2=x\,,\ x_i\in X_i} \big\{\|x_1\|_{X_1} + \frac{M_2}{M_1}t\|x_2\|_{X_2}\big\} \;=\; M_1 K\Big(\frac{M_2}{M_1}t, x\Big) \end{aligned}$$

$\diamond$

(b) Für $x \in \big(X_1, X_2\big)_{\theta,q}$ und $q < \infty$ folgt

$$\begin{aligned} \|Ax\|_{\left(Y_1,Y_2\right)_{\theta,q}} &= \left[\int_0^\infty \big(t^{-\theta}K(t,Ax)\big)^q \frac{dt}{t}\right]^{1/q} \\ &\overset{(a)}{\leq} M_1 \left[\int_0^\infty \Big(t^{-\theta}K\big(\frac{M_2}{M_1}t, x\big)\Big)^q \frac{dt}{t}\right]^{1/q} \\ &\overset{\frac{M_2}{M_1}t =: s}{=} M_1 \left[\int_0^\infty \big(M_1^{-\theta}M_2^{\theta}s^{-\theta}K(s,x)\big)^q \frac{ds}{s}\right]^{1/q} \\ &= M_1^{1-\theta}M_2^{\theta}\ \|x\|_{\left(X_1,X_2\right)_{\theta,q}} \end{aligned}$$

$\diamond$

(c) Für $q = \infty$ folgt die Aussage wieder mit $(a)$.

$\square$

## A.2 Anwendungen

Die Anwendungen dieser Theorie ergeben Fehlerabschätzungen etwa für den *Interpolationsfehler*, den *Quadraturfehler* oder den *Verfahrensfehler* – jeweils bei *linearen Methoden*[3] – auch in Fällen, wo die betrachteten Funktionen bzw. Lösungen nicht die klassisch geforderte Glattheit haben.

**(A.2.1) Satz**
Sei $T_h$, $h > 0$, eine Familie von regulären Triangulierungen von $\overline{G} = \bigcup\limits_{\triangle\in T_h} \triangle$, und $V_h$ der Raum der linearen $C^0$–Elemente zu $T_h$. Dann existiert für $0 < \theta < 1$ ein $c_\theta > 0$, so daß für $u \in \big(L_2(G), W^{2,2}(G)\big)_{\theta,2}$ gilt

$$\inf_{v_h\in V_h} \|u - v_h\|_{L_2(G)} \leq c_\theta h^{2\theta}\|u\|_{\left(L_2(G),W^{2,2}(G)\right)_{\theta,2}} .$$

[3] Literaturhinweise zur Übertragung auf gewisse nichtlineare Operatoren findet man z.Bsp. in Section 3.14 von Bergh/Löfström

Beweis:

Sei $P_h : L_2(G) \longrightarrow V_h$ die $L_2(G)$–Orthogonalprojektion und $A := id - P_h \ : \ L_2(G) + W^{2,2}(G) \longrightarrow L_2(G)$. Für $A : L_2(G) \longrightarrow L_2(G)$ gilt wegen

$$\|v\|^2_{L_2(G)} = \|v - P_h v\|^2_{L_2(G)} + \|P_h v\|^2_{L_2(G)} \geq \|v - P_h v\|^2_{L_2(G)}$$

die Abschätzung

$$\|A\|_{L_2(G) \to L_2(G)} \leq 1 \ .$$

Nach Bemerkung (5.2.9) gilt mit von $h$ unabhängigem $c_2 > 0$

$$\|Au\|_{L_2(G)} = \min_{v_h \in V_h} \|u - v_h\|_{L_2(G)} \leq \|u - \pi_h u\|_{L_2(G)} \leq c_2 h^2 \|u\|_{W^{2,2}(G)}, \ h > 0.$$

Damit folgt nach Satz (A.1.5)

$$\|Au\|_{\left(L_2(G), L_2(G)\right)_{\theta,2}} \leq 1^{1-\theta} (c_2 h^2)^\theta \|u\|_{\left(L_2(G), W^{2,2}(G)\right)_{\theta,2}} .$$

□

Wir können dieser Fehlerabschätzung noch eine andere Form geben:

**(A.2.2) Definition**
Für $s \in \mathbb{R} \ : \ k < s < l, \ k,l \in \mathbb{N}_0$, ist

$$W^{s,p}(G) := \left(W^{k,p}(G), W^{l,p}(G)\right)_{\theta,p},$$

wobei $\theta \in ]0,1[$ gerade $(1-\theta)k + \theta l = s$ erfüllt.

□

Im Allgemeinen nimmt man $l = k+1$ in dieser Definition. Jedoch kann man zeigen, daß man für ein festes $s$ immer dieselben Räume (unabhängig von $k$ und $l$) erhält mit zumindest äquivalenten Normen für unterschiedliche Paare $(k,l)$ (sog. *Reiterationssatz*, siehe z.B. Theorem 3.2.20 in Butzer/Berens).
Anwendbar wird diese Definition von $W^{s,p}(G)$ als intermediärer Raum dadurch, daß man auch eine interne Charakterisierung kennt.

**(A.2.3) Besov–Räume** (interne Charakterisierung von $W^{s,p}(G)$)
Für $G \subset \mathbb{R}^d$ beschränkt mit $Rd\,G$ aus $C^1$ gilt für $s = m + \sigma$ mit $m \in \mathbb{N}_0$ , $0 < \sigma < 1$,

$$W^{m+\sigma,p}(G) \ = \ \{u \in W^{m,p}(G) : \ \|u\|_{W^{m+\sigma,p}(G)} \ < \ \infty\}.$$

Dabei ist die Norm

$$\|u\|_{W^{m+\sigma,p}(G)} := \begin{cases} \left( \ \|u\|^p_{W^{m,p}(G)} \ + \displaystyle\sum_{|\alpha|=m} \int_G \int_G \frac{|D^\alpha u(x) - D^\alpha u(y)|^p}{\|x-y\|^{d+\sigma p}} \, dxdy \right)^{\frac{1}{p}} , \ 1 \leq p < \infty \\[2em] \max\left( \ \|u\|_{W^{m,\infty}(G)} \ , \ \displaystyle\max_{|\alpha|=m} \ \operatorname*{ess\,sup^4}_{x,y \,\in\, G \,:\, x \neq y} \ \frac{|D^\alpha u(x) - D^\alpha u(y)|}{\|x-y\|^\sigma} \right) , \ p = \infty \end{cases}$$

äquivalent zur $\left(W^{m,p}(G), W^{m+1,p}(G)\right)_{\sigma,p}$–Norm.

Für $G \ = \ \mathbb{R}^d$ oder $G \ = $ Halbraum im $\mathbb{R}^d$ gilt dieselbe Aussage.

□

[4] für messbares $v : G \longrightarrow \mathbb{R}$ ist das wesentliche Supremum $\operatorname*{ess\,sup}_{x \in G} v(x) = \inf\limits_{\mu(A)=0} \ \sup\limits_{x \in G \setminus A} \ v(x)$ mit dem Lebsgue–Maß $\mu$.

Zum Beweis vergleiche man etwa Adams, R.A.: Sobolew Spaces, Th. 7.48, Remark 7.49.

**(A.2.4) Folgerung**
Betrachtet man eine elliptische Randwertaufgabe mit $W^{1,2}(G)$-stetiger und $W_0^{1,2}(G)$-koerzitiver Bilinearform $a$. Dann gilt für die Diskretisierung durch lineare $C^0$–Elemente $V_h$ zu einer Folge regulärer Triangulierungen $T_h$

$$\|u-u_h\|_{W^{1,2}(G)} \;\leq\; c_s\, h^{s-1}\|u\|_{W^{s,2}(G)}\,,\; h\to 0\;;\; 1\leq r\leq 2.$$

Dabei ist $u\in W^{s,2}(G)$ die Lösung der Randwertaufgabe und $u_h\in V_h$ die Galerkinnäherung.

Beweis:

Wendet man Satz (A.1.5) auf $A(u) := u-u_h$ an, so folgt die Behauptung durch Interpolation der für $r=2$ geltenden Fehlerabschätzung aus Bemerkung (5.2.9) und der folgenden Abschätzung für $r=1$ :

$$\begin{array}{rcll} \underline{c}\,\|u-u_h\|^2_{W^{1,2}(G)} & \leq & a(u-u_h,u-u_h) & ,\ a \text{ koerzitiv };\ u-u_h\in W_0^{1,2}(G) \\ & = & a(u-u_h,u) & ,\ u_h \text{ Galerkinnäherung} \\ & \leq & \overline{c}\,\|u-u_h\|_{W^{1,2}(G)}\,\|u\|_{W^{1,2}(G)} & ,\ a \text{ stetig.} \end{array}$$

Also gilt für $1<s<2$ wegen $s=(1-(s-1))\cdot 1+(s-1)\cdot 2$ mit einer von $h$ unabhängigen Konstanten $c_s>0$ $\quad\|u-u_h\|_{W^{1,2}(G)} \;\leq\; c_s\, h^{s-1}\|u\|_{W^{s,2}(G)}\,.$ □

Die Anwendung der 'Interpolation' von (Fehler) Normen für Funktionenräume $C^m$ von klassisch stetig differenzierbaren Funktionen und $C^{m,\sigma},\sigma\leq 1$, von Funktionen aus $C^m$ mit zum Exponenten $\sigma$ hölderstetigen $m$-ten Ableitungen illustrieren wir an Fehlerabschätzungen für Quadraturformeln. Dabei tritt ein z.B. auch in der Approximationstheorie bekannter Ausnahmefall für $\sigma=1$ auf. Wir definieren

**(A.2.5) Definition**
Sei $G=\mathbb{R}^d$ und $s\geq 0$.
(a) Für $s=m\in\mathbb{N}_0$ ist $C^m(\mathbb{R}^d)$ die Vervollständigung von $C_0^\infty(\mathbb{R}^d)$ bezüglich

$$\|u\|_m \;:=\; \sum_{|\alpha|\leq m} sup_{x\in\mathbb{R}^d}|D^\alpha u(x)|\;.$$

(b) Für $m\in\mathbb{N}_0,\ 0<\sigma\leq 1$, ist

$$C^{m,\sigma}(\mathbb{R}^d) \;=\; \{u\in C^m(\mathbb{R}^d)\ :\ \|u\|_{C^{m,\sigma}} \;<\; \infty\}$$

mit

$$\|u\|_{C^{m,\sigma}} \;:=\; \|u\|_m + \sum_{|\alpha|=m} sup_{x\neq y}\frac{|D^\alpha u(x)-D^\alpha u(y)|}{\|x-y\|^\sigma}\;.$$

(c) Für $m\in\mathbb{N}_0,\ 0<\sigma\leq 1$, ist

$$Z^{m,\sigma}(\mathbb{R}^d) \;=\; \{u\in C^m(\mathbb{R}^d)\ :\ \|u\|_{Z^{m,\sigma}} < \infty\}$$

mit

$$\|u\|_{Z^{m,\sigma}} \;:=\; \|u\|_m + \sum_{|\alpha|=m} sup_{x\neq y}\frac{|D^\alpha u(x)-2D^\alpha u(\frac{x+y}{2})+D^\alpha u(y)|}{\|x-y\|^\sigma}\;.$$

□

Die Räume $C^{m,\sigma}$ und $Z^{m,\sigma}$ unterscheiden sich nur für $\sigma = 1$, d.h.

$$C^{m,\sigma}(\mathbb{R}^d) \;=\; Z^{m,\sigma}(\mathbb{R}^d),\; m \in \mathbb{N}_0, 0 < \sigma < 1 \;;$$

$$C^{m,1}(\mathbb{R}^d) \;\overset{\subset}{\neq}\; Z^{m,1}(\mathbb{R}^d),\; m \in \mathbb{N}_0 \;.$$

**(A.2.6) Satz**

Sei $0 \le r < t < \infty$ , $0 < \theta < 1$ und $s = (1-\theta)r + \theta t$ . Dann gilt

$$\big(C^r(\mathbb{R}^d), C^t(\mathbb{R}^d)\big)_{\theta,\infty} \;=\; \begin{cases} Z^s(\mathbb{R}^d)\;,\; s \notin \mathbb{N}_0\;, \\ \\ Z^{s-1,1}(\mathbb{R}^d)\;,\; s \in \mathbb{N}_0\;. \end{cases}$$

□

Zum Beweis vergleiche man *Triebel, H.*: Interpolation Theory, Function Spaces, Differential Operators, Abschnitt 2.7.2 Theorem 1 auf S. 201.

**(A.2.7) Folgerung**

Betrachtet werde die zusammengesetzte m-punktige Gauß-Quadraturformel $Q_{I,h}$ mit äquidistanten Teilintervallen der Länge $h > 0$ des kompakten Intervalls $I = [a,b] \subset \mathbb{R}$. Dann existiert zu $s \in \mathbb{N}_0$ mit $0 \le s \le 2m$ eine Konstante $c_s > 0$, so daß gilt

$$\left| \int_I u(t)\,dt - Q_{I,h}(u) \right| \;\le\; c_s h^s \|u\|_s,\; u \in C^s(I).$$

Beweis:

Da man Funktionen aus $C^s(I)$ stetig einbetten kann in $C^s(\mathbb{R})$ durch eine $C^s$-Fortsetzung, kann man o.E. betrachten:

$$A(u) \;:=\; \int_I u(t)\,dt - Q_{I,h}(u),\; u \in C^s(\mathbb{R}).$$

Dann gilt für $u \in X_1 \;:=\; (C^0(\mathbb{R}), \|\;\|_0)$

$$|A(u)| \;\le\; |\int_I u(t)\,dt| \;+\; |Q_{I,h}(u)| \;\le\; |I| \max_{t\in I} |u(t)| \;+\; |\sum_j Q_{I_j}(u)|\,,$$

mit $I = \bigcup_{j=0}^{N} I_j,\; I_j = [t_j, t_{j+1}],\; t_j = a + jh,\; h = (b-a)/(N+1)$. $Q_{I_j}(u) \;=\; \sum_{k=1}^{m} w_k^{I_j} u(s_k^{I_j})$ ist die auf das Intervall $I_j$ transformierte m–punktige Gaußformel, $w_k^{I_j}$ und $s_k^{I_j}$ sind die entsprechend transformierten Gewichte und Stützstellen. Wegen

$$|Q_{I_j}(u)| \;\le\; \sum_{k=1}^{m} |\underbrace{w_k^{I_j}}_{>0}| \quad \underbrace{|u(s_k^{I_j})|}_{\le \max_{t\in I} |u(t)| = \|u\|_0} \quad\le\; \|u\|_0 \sum_{k=1}^{m} w_k^{I_j} \;=\; \|u\|_0\, |I_j|$$

folgt

$$|A(u)| \;\le\; 2|I|\; \|u\|_0.$$

Für $X_2 \;=\; (C^{2m}(\mathbb{R}), \|\;\|_{2m})$ gilt nach *Numerik I*

$$|A(u)| \;\le\; c_{2m}\, h^{2m}\, \|u\|_{2m}.$$

Aus dem Konvexitätstheorem (A.1.5) folgt wegen Satz (A.2.6)

$$|A(u)| \;\le\; c_s h^s \|u\|_{Z^s},\;\; u \in Z^s(I).$$

Wegen $C^s(I) \;=\; Z^s(I)$ für $s \notin \mathbb{N}$ bzw. $C^s(I) \hookrightarrow Z^{s-1,1}(I)$ für $s \in \mathbb{N}$ folgt die Behauptung.

□