Numerik II
Sommersemester 2007

Anita Schöbel

12. November 2007

Inhaltsverzeichnis

Kapitel 1

Numerische Integration

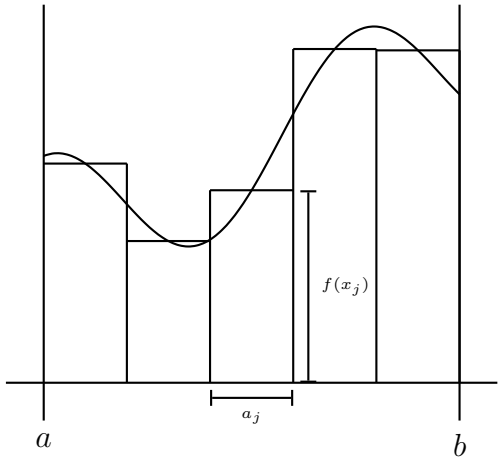
Unser Ziel ist es, eine einfache Formel zur Berechnung von

$$\int_a^b f(x)dx$$

zu finden. Eine Möglichkeit ist die Annäherung durch Rechtecke:

$$\int_a^b f(x)dx \approx \sum_{j=1}^{n} a_j f(x_j).$$

Hierbei ist a_j die Breite des jeweiligen Rechtecks und $f(x_j)$ die Höhe. Diese Summenformel ist „einfach“ zu berechnen.



Notation 1.1 *Sei $x_0, \ldots, x_n \in [a, b]$. Eine Abbildung $Q : \mathbb{R}^{[a,b]} \to \mathbb{R}$ heißt **Quadraturformel** bzgl. $x_0, \ldots, x_n$ falls gilt:*

$$Q(f) = \sum_{j=0}^{n} a_j f(x_j) \text{ für } a_0, \ldots, a_n \in \mathbb{R}.$$

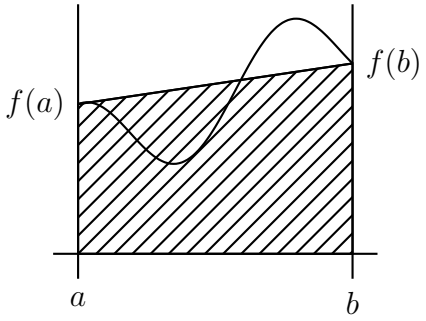
Bemerkung: Q ist eine lineare Abbildung.

Wir versuchen im Folgenden, Quadraturformeln Q zu finden, die Integrale $I(f) := \int_a^b f(x)dx$ annähern, d.h. Q mit $Q(f) \approx I(f)$.

Beispiel: Die Trapezregel ist eine Vorschrift, $\int_a^b f(x)dx$ durch die Fläche eines Trapezes mit den Ecken $(a,0)$, $(a,f(a))$, $(b,f(b))$, $(b,0)$ zu approximieren:

$$\int_a^b f(x)dx \approx \frac{(b-a)}{2}(f(a)+f(b)).$$

Die Trapezregel ist also eine Quadraturformel mit $n=1$, $x_0=a$, $x_1=b$ und $a_0=a_1=\frac{1}{2}(b-a)$. Sie integriert alle affin-linearen Funktionen exakt.



Notation 1.2 *Eine Quadraturformel Q heißt **exakt** für $\mathfrak{F} \subseteq \mathbb{R}^{[a,b]}$, falls $Q(f) = I(f)$ für alle $f \in \mathfrak{F}$ gilt.*

Satz 1.3 *Sei $\mathfrak{F} \subseteq \mathbb{R}^{[a,b]}$ ein endlich-dimensionaler Unterraum von $\mathbb{R}^{[a,b]}$ und $f_0, \ldots, f_N$ eine Basis von $\mathfrak{F}$. Gilt dann $Q(f_i) = I(f_i)$ für alle $i = 0, \ldots, N$ für eine Quadraturformel Q, dann ist Q exakt für $\mathfrak{F}$.*

Beweis: Sei $f \in \mathfrak{F}$. Dann kann man f bezüglich der Basis $f_0, \ldots, f_N$ darstellen als

$$f = \sum_{i=0}^{N} a_i f_i.$$

Es gilt nun

$$\begin{aligned} Q(f) &= Q\left(\sum_{i=0}^{N} a_i f_i\right) = \sum_{i=0}^{N} a_i Q(f_i), \text{ da } Q \text{ linear} \\ &= \sum_{i=0}^{N} a_i I(f_i), \text{ da } Q \text{ exakt für } f_i \text{ und} \\ &= I\left(\sum_{i=0}^{N} a_i f_i\right) = I(f), \text{ da Integrale linear sind.} \end{aligned}$$

QED

Im Folgenden betrachten wir

- Interpolationsquadraturen nach Newton-Côtes,
- Gauß'sche Quadraturformeln und
- die Rombergquadratur.

1.1 Interpolationsquadraturen

Seien $x_0, \ldots, x_n \in [a,b]$ gegeben. Eine Idee ist es, das Integral von f durch das Integral des eindeutig bestimmten Interpolationspolynoms $(L_n f)(x)$ bezüglich der Stützstellen $(x_j, f(x_j))$, $j = 0, \ldots, n$ zu approximieren.

Definition 1.4 *Eine Quadraturformel $Q_n(f) = \sum_{j=0}^{n} a_j f(x_j)$ heißt **Interpolationsquadratur der Ordnung** n, falls für alle $f \in \mathcal{C}[a,b]$ gilt:*

$$Q_n(f) = \sum_{j=0}^{n} a_j f(x_j) = \int_a^b (L_n f)(x)dx = I(L_n f).$$

Dabei ist $L_n f$ das eindeutig bestimmte Interpolationspolynom zu f bezüglich der Stützstellen $x_0, \ldots, x_n$.

Wir erinnern uns an Numerik I, wo wir im Kapitel 6.1 verschiedene Darstellungen für $L_n f$ hergeleitet hatten. Eine war die Lagrange-Darstellung:

$$(L_n f)(x) = \sum_{j=0}^{n} f(x_j) l_j(x) \text{ mit } l_j(x) = \prod_{\substack{k=0 \\ k \neq j}}^{n} \frac{x - x_k}{x_j - x_k}.$$

Wir werden im Folgenden die Koeffizienten a_j von Interpolationsquadraturen der Ordnung n herleiten. Kennt man diese, so kann man alle Polynome $p \in \Pi_n$ exakt integrieren. Erstaunlicherweise gilt auch die Umkehrung dieser Aussage:

Satz 1.5 *Eine Quadraturformel Q_n ist genau dann eine Interpolationsquadratur vom Grad n, wenn alle Polynome $p \in \Pi_n[a,b]$ exakt integriert werden.*

Beweis: „$\Rightarrow$“: Sei $Q_n(f) = \int_a^b (L_n f)(x)dx$ eine Interpolationsquadratur der Ordnung n und sei $f \in \Pi_n[a,b]$. Nach dem Satz 6.4 aus Numerik I gilt dann $f = L_n f$, also ist $Q(f) = \int_a^b f(x)dx = I(f)$ und Q_n ist exakt für alle $f \in \Pi_n[a,b]$.
„$\Leftarrow$“: Sei umgekehrt $Q(f) = \sum_{j=0}^{n} a_i f(x_j)$ eine Quadraturformel die $I(p) = Q(p)$ für alle $p \in \Pi_n[a,b]$ erfüllt. Sei $f \in \mathcal{C}[a,b]$. Dann ist $L_n f \in \Pi_n[a,b]$ und es gilt

$$\begin{aligned} I(L_n f) = Q(L_n f) &= \sum_{j=0}^{n} a_j (L_n f)(x_j) \\ &= \sum_{j=0}^{n} a_j f(x_j) = Q(f), \end{aligned}$$

also ist $Q(f)$ eine Interpolationsquadratur der Ordnung n. QED

Satz 1.6 *Sei $h_{n+1}(x) = \prod_{j=0}^{n}(x - x_j)$. Seien $x_0, \ldots, x_n$ paarweise verschieden aus $[a, b]$. Dann existiert genau eine Interpolationsquadratur der Ordnung n zu $x_0, \ldots, x_n$, die durch die Gewichte*

$$a_j = \frac{1}{h'_{n+1}(x_j)} \int_a^b \frac{h_{n+1}(x)}{x - x_j} dx \text{ mit } j = 0, \ldots, n$$

gegeben ist.

Beweis: Wir zeigen zunächst die Eindeutigkeit. Seien

$$Q_A(f) = \sum_{j=0}^{n} a_j f(x_j) \text{ und } Q_B(f) = \sum_{j=0}^{n} b_j f(x_j)$$

zwei Interpolationsquadraturen der Ordnung n. Dann gilt:

$$Q_A(f) = \int_a^b (L_n f)(x) dx = Q_B(f) \text{ für alle } f \in \mathcal{C}[a, b].$$

Wir wählen nun zu jedem j ein f_j mit $f_j(x_j) \neq 0$ und $f_j(x_k) = 0$ für alle $k \neq j$ – z.B. $f_j = l_j$, die Lagrangepolynome. Dann gilt $Q_A(f_j) = a_j = b_j = Q_B(f_j)$, d.h. die Interpolationsquadraturen sind gleich.
Zum Existenzbeweis: $L_n f$ ist stetig und deshalb integrierbar. Wir gehen über die Lagrange-Darstellung:

$$\int_a^b (L_n f)(x) dx = \int_a^b \sum_{j=0}^{n} f(x_j) l_j(x) dx = \sum_{j=0}^{n} f(x_j) \int_a^b l_j(x) dx,$$

d.h. $\int_a^b (L_n f)(x) dx = \sum_{j=0}^{n} f(x_j) a_j$ mit $a_j = \int_a^b l_j(x) dx$ ist tatsächlich eine Interpolationsquadratur. Weiterhin gilt:

$$a_j = \int_a^b \prod_{\substack{k=0 \\ k \neq j}}^{n} \frac{x - x_k}{x_j - x_k} dx = \frac{1}{h'_{n+1}(x_j)} \int_a^b \frac{h_{n+1}(x)}{x - x_j} dx,$$

wobei

$$h'_{n+1}(x) = \sum_{k=0}^{n} \prod_{\substack{i=0 \\ i \neq k}}^{n} (x - x_i)$$

gilt und insbesondere

$$h'_{n+1}(x_j) = \prod_{\substack{i=0 \\ i \neq j}}^{n} (x_j - x_i).$$

QED

Zur Vereinfachung der Formeln betrachten wir den Fall äquidistanter Stützstellen

$$x_j = a + j \cdot h \text{ mit } j = 0, \ldots, n.$$

Wir bemerken:

$$x_n = a + n \cdot h = b, \text{ also } h = \frac{b-a}{n}.$$

Definition 1.7 *Die Interpolationsquadratur der Ordnung n zu den Stützstellen $x_j = a + j \cdot h$ mit $j = 0, \ldots, n$ mit Schrittweite $h = \frac{b-a}{n}$ heißt* ***Newton-Côtes-Formel*** *der Ordnung n.*

Lemma 1.8 *Die Gewichte der Newton-Côtes-Formel der Ordnung n ergeben sich aus*

$$a_j = h \cdot A_j$$
$$\text{mit } A_j = A_{n-j} = \frac{(-1)^{n-j}}{j!(n-j)!} \int_0^n \prod_{\substack{k=0 \\ k \neq j}}^{n} (z-k)dz \qquad \text{für } j = 0, \ldots, n.$$

Beweis: Übung.

Bemerkung: Die Werte A_j hängen ausschließlich von der Anzahl n der Stützstellen ab, nicht aber von den Werten x_j der Stützstellen und auch nicht von a, b oder h!

Einfacher als die Berechnung der A_j nach Lemma 1.8 ist ihre Ermittlung über die Lösung eines linearen Gleichungssystems. Dazu fordert man speziell für die Monome $p(x) = x^k$ für $k = 0, \ldots, n$, dass

$$\sum_{i=0}^{n} a_i p(x_i) = \int_a^b p(x)dx.$$

Nach Satz 1.3 folgt daraus, dass $\sum_{i=0}^{n} a_i p(x_i) = I(p)$ für alle $p \in \Pi_n$ gilt.
Auf diese Weise berechnen wir nun die Koeffizienten für die Fälle $n = 1$ und $n = 2$.

$\boldsymbol{n = 1}$: Nach der Bemerkung nach Lemma 1.8 können wir o.B.d.A. $a = -1$ und $b = 1$ setzen. Somit gilt $h = 2$ und wir erhalten:

- $p(x) = x^0$:

$$\int_{-1}^{1} x^0 dx = [x]_{-1}^{1} = 2 \text{ und } \sum_{j=0}^{1} a_j p(x_j) = a_0 + a_1,$$

  also $a_0 + a_1 = 2$ als erste Bedingung.

- $p(x) = x$:

$$\int_{-1}^{1} x^1 dx = \left[\tfrac{1}{2}x^2\right]_{-1}^{1} = 0 \text{ und } \sum_{j=0}^{1} a_j p(x_j) = a_0 \cdot (-1) + a_1 \cdot 1,$$

  also $-a_0 + a_1 = 0$ als zweite Bedingung.

Die Lösung des Systems

$$\begin{aligned} a_0 + a_1 &= 2 \\ -a_0 + a_1 &= 0 \end{aligned}$$

ist $a_0 = a_1 = 1$ (bzw. $A_0 = A_1 = \frac{1}{2}$) und man erhält daraus

$$\int_{-1}^{1} f(x)dx \approx 1 \cdot f(-1) + 1 \cdot f(1).$$

Für beliebige Integrationsgrenzen a, b ändern sich A_0 und A_1 nicht, sodass wir die auf Seite 3 schon beschriebene **Trapez-Regel**

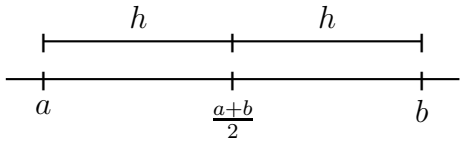
$$\int_a^b f(x)dx \approx \frac{b-a}{2}(f(a) + f(b))$$

erhalten.
$\boldsymbol{n = 2}$: Wie schon im vorherigen Fall wählen wir $a = -1$ und $b = 1$. Daraus ergeben sich nun $h = 1$, $x_0 = -1$, $x_1 = 0$, $x_2 = 1$ und entsprechend die Gleichungen

$$\begin{aligned} \int_{-1}^{1} x^0 dx &= 2 = a_0 + a_1 + a_2, \\ \int_{-1}^{1} x^1 dx &= 0 = -a_0 + a_2, \\ \int_{-1}^{1} x^2 dx &= \tfrac{2}{3} = a_0 + a_2, \end{aligned}$$

woraus man $a_0 = A_0 = \frac{1}{3}$, $a_1 = A_1 = \frac{4}{3}$ und $a_2 = A_2 = \frac{1}{3}$ als eindeutige Lösung errechnet. Daraus erhält man die **Simpson-Regel**

$$\begin{aligned} \int_a^b f(x)dx &\approx \frac{h}{3}(f(x_0) + 4f(x_1) + f(x_2)) \\ &= \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right). \end{aligned}$$



Die folgende Tabelle gibt die Gewichte der ersten fünf Newton-Côtes Formeln an:

n	A_0	A_1	A_2	A_3	A_4	A_5	Bezeichnung
1	$\frac{1}{2}$	$\frac{1}{2}$					Trapez-Regel
2	$\frac{1}{3}$	$\frac{4}{3}$	$\frac{1}{3}$				Simpson-Regel
3	$\frac{3}{8}$	$\frac{9}{8}$	$\frac{9}{8}$	$\frac{3}{8}$			Newton-$\frac{3}{8}$-Regel
4	$\frac{14}{45}$	$\frac{64}{45}$	$\frac{24}{45}$	$\frac{64}{45}$	$\frac{14}{45}$		1. Milne-Regel
5	$\frac{95}{288}$	$\frac{375}{288}$	$\frac{250}{288}$	$\frac{250}{288}$	$\frac{375}{288}$	$\frac{95}{288}$	2. Milne-Regel

Leider tauchen ab $n \geq 8$ auch negative Gewichte auf, die unerwünschte Nebeneffekte haben:

- Auslöschung ist möglich und führt zu numerischer Instabilität und
- es lassen sich positive Funktionen $f \geq 0$ konstruieren, sodass $Q(f) < 0$ gilt.

Wir betrachten nun folgendes Beispiel für die Simpson-Regel:

$$f(x) = \left(x - \frac{a+b}{2}\right)^3.$$

Dann gilt

$$\int_a^b f(x)dx = 0,$$

denn f ist punktsymmetrisch zu $(\frac{a+b}{2}, 0)$:

$$-f(\tfrac{a+b}{2} + x) = -x^3 = (-x)^3 = f(\tfrac{a+b}{2} - x).$$

Wendet man die Simpson-Regel auf f an, so erhält man

$$\begin{aligned} Q_2(f) &= \tfrac{b-a}{6}(f(a) + 4f(\tfrac{a+b}{2}) + f(b)) \\ &= \tfrac{b-a}{6}((\tfrac{a-b}{2})^3 + 4(0)^3 + (\tfrac{b-a}{2})^3) = 0 \end{aligned}$$

also ist die Simpson-Regel für dieses kubische Polynom exakt. Das gilt sogar für alle kubischen Polynome!

Lemma 1.9 *Die Simpson-Regel Q_2 ist exakt für alle $p \in \Pi_3[a, b]$.*

Beweis: Nach Satz 1.3 ist eine Quadraturformel auf Π_3 exakt, wenn sie auf einer Basis von Π_n exakt ist. Wir wählen als Basis

$$(x - \tfrac{a+b}{2})^3,\ (x - \tfrac{a+b}{2})^2,\ x - \tfrac{a+b}{2},\ 1.$$

Im vorangehenden Beispiel haben wir bereits $Q_2((x - \frac{a+b}{2})^3) = I((x - \frac{a+b}{2})^3)$ gezeigt und für die anderen Basisvektoren folgt die Exaktheit aus Satz 1.5, denn alle Polynome aus $\Pi_2[a,b]$ werden von einer Interpolationsquadratur der Ordnung 2 exakt integriert. QED

Man kann diese Aussage weiter verallgemeinern:

Satz 1.10 *Sei $Q_n(f) = \sum_{j=0}^{n} a_j f(x_j)$ eine Newton-Côtes Formel mit geradem n. Dann gilt*

$$Q_n(p) = I(p) \text{ für alle } p \in \Pi_{n+1}[a,b].$$

Beweis: Übung.

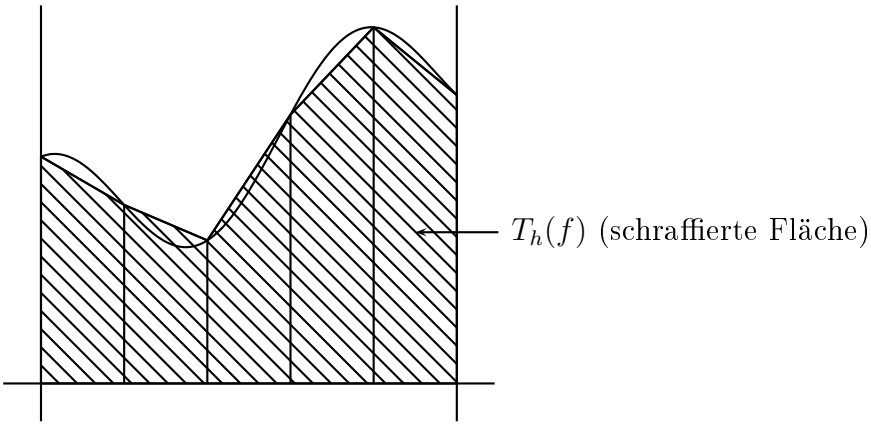
1.2 Zusammengesetzte Newton-Côtes-Formeln

Analog zur Spline-Interpolation zerlegt man bei den zusammengesetzten Newton-Côtes-Formeln das Integrationsintervall in m Teilintervalle und wendet auf je n zusammenhängenden Teilintervallen eine Quadraturformel niederer Ordnung an. Dafür wählen wir m so, dass n Teiler von m ist.

Wir erhalten auf diese Weise die **zusammengesetzte Trapezregel**:

$$\int_a^b f(x)dx \approx T_h(f) := h\left(\tfrac{1}{2}f(x_0) + \sum_{i=1}^{m-1} f(x_i) + \tfrac{1}{2}f(x_m)\right),$$

mit $x_i = a_i + ih$ für $i = 0, \ldots, m$, also $x_0 = a$ und $x_m = b$. Dabei ist $h = \frac{b-a}{m}$.



Analog ergibt sich die **zusammengesetzte Simpsonregel**: Sei dazu $x_0 = a < x_1 < \cdots < x_m = b$ eine äquidistante Zerlegung in eine gerade Anzahl an Teilintervallen. Wir wenden die Simpson-Regel jeweils auf zwei aufeinander folgende Intervalle an und erhalten:

$$\int_a^b f(x)dx \approx S_h(f) = \frac{h}{3} \sum_{j=0}^{\frac{m}{2}-1} \tfrac{1}{3} f(x_{2j}) + \tfrac{4}{3} f(x_{2j+1}) + \tfrac{1}{3} f(x_{2j+2})$$

1.3 Gauß'sche Integrationsformeln

Für die Gauß'schen Integrationsformeln sollen neben den Gewichten $a_0, a_1, \ldots, a_n$ auch die Stützstellen $x_0, \ldots, x_n$ gewählt werden. Man hat also $2(n+1)$ Freiheitsgrade. Entsprechend darf man $2n+2$ Bedingungen stellen, z. B.

$$\sum_{i=0}^{n} a_i p(x_i) = \int_a^b p(x)dx \text{ für } p(x) \in \{1, x, x^2, \ldots, x^{2n+1}\}.$$

Sind diese Bedingungen erfüllt, kann man alle Polynome aus Π_{2n+1} exakt integrieren (Satz 1.3). Wir wollen gerne ohne dieses nichtlineare Gleichungssystem auskommen.
Dabei ist es in verschiedenen Anwendungen günstig, den allgemeineren Fall von Quadraturformeln zu betrachten, nämlich **gemischte Integrale**

$$I_w(f) := \int_a^b \omega(x) f(x) dx$$

mit einer auf (a, b) stetigen und positiven Gewichtsfunktion ω. Weiterhin fordert man, dass

$$\int_a^b \omega(x) x^k dx$$

für alle $k \in \mathbb{N}_0$ existiert.

Typische Beispiele für solche Gewichtsfunktionen sind

- Gauß-Legendre: $\omega(x) = 1$ auf $[a, b]$
- Gauß-Tschebyscheff 1. Art: $\omega(x) = \frac{1}{\sqrt{1-x^2}}$ auf $x \in [-1, 1]$.
- Gauß-Tschebyscheff 2. Art: $\omega(x) = \sqrt{1-x^2}$ auf $x \in [-1, 1]$.
- Gauß-Laguerre: $\omega(x) = e^{-x}$ auf $[0, \infty)$.
- Gauß-Hermite: $\omega(x) = e^{-x^2}$ auf $(-\infty, \infty)$

Definition 1.11 *Eine Quadraturformel* $Q_n(f) = \sum_{i=0}^{n} a_i f(x_i)$ *nennt man* ***Gauß'sche Quadraturformel der Ordnung*** n*, wenn sie alle Polynome* $p \in \Pi_{2n+1}[a,b]$ *exakt integriert, d.h. wenn*

$$Q_n(p) = \int_a^b \omega(x)p(x)dx \text{ für alle } p \in \Pi_{2n+1}[a,b].$$

Lemma 1.12 *Seien* $x_0, \ldots, x_n \in [a,b]$*. Sei* $L_n f$ *das Interpolationspolynom bzgl.* $x_0, \ldots, x_n$ *an die Funktion* f*. Sei weiter* ω *eine zulässige Gewichtsfunktion. Dann ist*

$$\int_a^b \omega(x)(L_n f)(x)dx$$

eine Quadraturformel bzgl. $x_0, \ldots, x_n$*. Genauer gilt:*

$$\int_a^b \omega(x)(L_n f)(x)dx = \sum_{j=0}^{n} a_j f(x_j)$$

mit $a_j = \int_a^b \omega(x) l_j(x)dx$*.*

Beweis: Den Fall $\omega \equiv 1$ haben wir in Satz 1.6 behandelt. Für andere ω verläuft der Beweis analog. QED

Wann ist $Q_n(f)$ eine Gauß'sche Quadraturformel?

Satz 1.13 *Sei* ω *eine zulässige Gewichtsfunktion und seien* $x_0, \ldots, x_n \in [a,b]$ *paarweise verschieden. Sei* $L_n f$ *die Interpolation von* f *bzgl. der Stützstellen* $x_0, \ldots, x_n$*. Sei weiterhin* $h_{n+1}(x) = \prod_{j=0}^{n}(x - x_j)$*. Dann sind die folgenden beiden Aussagen äquivalent:*

1. $Q_n(f) := \int_a^b \omega(x)(L_n f)(x)dx$ *ist eine Gauß'sche Quadraturformel der Ordnung* n*, d.h.* $Q_n(p) = I_w(p)$ *für alle* $p \in \Pi_{2n+1}$*.*

2. $\int_a^b \omega(x) h_{n+1}(x) p(x)dx = 0$ *für alle* $p \in \Pi_n$*.*

Beweis: „1 $\Rightarrow$ 2“: Sei $Q_n(f) = I_w(f)$ für alle $f \in \Pi_{2n+1}$. Sei $p \in \Pi_n$. Dann ist $h_{n+1} \cdot p \in \Pi_{2n+1}$, also gilt nach Lemma 1.12:

$$\int_a^b \omega(x)h_{n+1}(x)p(x)dx = Q_n(h_{n+1} \cdot p) = \sum_{j=0}^{n} a_j \underbrace{h_{n+1}(x_j)}_{=0} p(x_j) = 0,$$

also gilt 2.

„2 $\Rightarrow$ 1“: Sei $p \in \Pi_{2n+1}$. Betrachte $L_n p \in \Pi_n$ bzgl. $x_0, \ldots, x_n$. Dann hat $p - L_n p$ die Nullstellen $x_0, \ldots, x_n$, also gibt es nach dem Hauptsatz der Algebra ein Polynom $q \in \Pi_n$, so dass $p - L_n p = h_{n+1} \cdot q$. Damit gilt

$$\begin{aligned}\int_a^b \omega(x)p(x)dx &= \int_a^b \omega(x)L_n p(x)dx + \underbrace{\int_a^b \omega(x)h_{n+1}(x)q(x)dx}_{=0 \text{ nach 2. weil } q\in\Pi_n} \\ &= Q_n(p),\end{aligned}$$

also ist Q_n Gauß'sche Quadraturformel. QED

Notation 1.14 *Für $f, g \in C([a,b])$ definieren wir*

$$(f,g)_\omega := \int_a^b \omega(x)f(x)g(x)dx.$$

Gilt $(f,g)_\omega = 0$, so bezeichnet man f und g als ω-orthogonal.

Bemerkung: Der Ausdruck $(f,g)_\omega$ existiert für alle Polynome f und g, wenn ω eine zulässige Gewichtsfunktion ist. Es lässt sich sogar zeigen, dass $(f,g)_\omega$ ein Skalarprodukt ist (d.h. bilinear, symmetrisch und positiv für $f = g$, sowie streng positiv für $f = g \neq 0$).

Satz 1.13 lässt sich jetzt folgendermaßen formulieren:

$\int_a^b \omega(x)(L_n f)(x)dx$ ist genau dann eine Gauß'sche Quadraturformel der Ordnung n, wenn $(h_{n+1}, p)_\omega = 0$ für alle $p \in \Pi_n$.

Die gesuchten Stützstellen der Quadraturformel müssen also die Nullstellen eines Polynoms $q(x) = \alpha h_{n+1}(x) \in \Pi_{n+1}$ sein, das ω-orthogonal zu allen $q \in \Pi_n$ ist. Solche Polynome wollen wir im Folgenden konstruieren.

Satz 1.15 *Sei ω eine zulässige Gewichtsfunktion. Dann gilt*

1. *Es existieren Polynome $p_n \in \Pi_n[a,b]$ für alle $n \in \mathbb{N}_0$ mit*

$$(p_n, p_m) = \delta_{n,m} = \begin{cases} 1 & \text{falls } n = m \\ 0 & \text{falls } n \neq m \end{cases} \quad \text{für alle } n, m \in \mathbb{N}_0.$$

2. *Für alle $n \in \mathbb{N}_0$ gilt: Die Nullstellen von p_n sind alle reell und liegen in (a,b).*

Beweis: **ad 1.** Die Folge p_i der gesuchten Polynome lässt sich durch Anwenden des Schmidt'schen Orthonormalisierungsverfahrens auf die Monome konstruieren. Man erhält entsprechend eine Orthonormalbasis. Die Monombasis ist $\{x^0, \ldots, x^n\}$. Man setzt nun

$$p_0(x) = \frac{x^0}{\sqrt{(x^0, x^0)_\omega}} = \frac{1}{\sqrt{\int_a^b \omega(x)dx}}$$

und erhält $(p_0, p_0)_\omega = 1$.
Zur Konstruktion von p_n nehmen wir an, dass $p_0, \ldots, p_{n-1}$ bereits konstruiert sind und dass sie $(p_i, p_j) = \delta_{ij}$ erfüllen. Dann ergibt sich $p_n \in \Pi[a, b]$ aus

$$p_n(x) = \gamma_n \left(x^n - \sum_{i=0}^{n-1} (x^n, p_i)_\omega p_i(x) \right),$$

wobei die $(x^n, p_i(x))_\omega$ die Koeffizienten nach Schmidt sind. Das ist die Lösung, denn

- für $m = 0, \ldots, n-1$ gilt:

$$\begin{aligned}(p_n, p_m) &= \gamma_n \left((x^n, p_m)_\omega - \sum_{i=0}^{n-1} (x^n, p_i)_\omega (p_i, p_m) \right) \\ &= \gamma_n \left((x^n, p_m)_\omega - (x^n, p_m)_\omega \cdot 1 \right) = 0\end{aligned}$$

- und γ_n wird so gewählt, dass $(p_n, p_m)_\omega = 1$.

ad 2. Seien $x_1, \ldots, x_m$ die reellen Nullstellen von p_n in (a, b) mit ungerader Vielfachheit, d.h. genau die Nullstellen mit Vorzeichenwechsel von p_n. Sei

$$q_m(x) = \prod_{i=1}^{m} (x - x_i) \text{ mit } q_0(x) := 1.$$

Wir wollen nun zeigen, dass $m = n$ gilt. Angenommen, es gelte $m < n$. Dann gilt $q_m \in \Pi_m[a, b] \subseteq \Pi_{n-1}[a, b]$. Weiter ist $p_n q_m(x) \geq 0$ für alle $x \in (a, b)$, weil es nur Nullstellen mit gerader Vielfachheit hat. Weil $p_n q_m \neq 0$ folgt

$$(p_n, q_m)_\omega \neq 0.$$

Weil die in Teil 1 konstruierten Polynome $p_0, \ldots, p_m$ eine Basis von Π_m bilden, gilt andererseits

$$q_m = \sum_{i=0}^{m} \lambda_i p_i \text{ mit reellen Koeffizienten } \lambda_i$$

und nach Konstruktion der p_i ist

$$(p_n, q_m)_\omega = \sum_{i=0}^{m} \lambda_i (p_n, p_i)_\omega = 0,$$

denn $(p_n, p_i)_\omega = 0$, weil $m < n$. Das ist ein Widerspruch, also muss $m = n$ gelten und die Vielfachheit jeder Nullstelle ist entsprechend 1. QED

Wir fassen die Ergebnisse in folgendem Existenzsatz zusammen:

Satz 1.16 *Sei $n \in \mathbb{N}$ und ω eine zulässige Gewichtsfunktion. Dann existiert eine Gauß'sche Quadraturformel*

$$Q_n(f) = \sum_{j=0}^{n} a_j f(x_j),$$

wobei $x_0 < x_1 < \cdots < x_n \in (a, b)$ die Nullstellen des in Satz 1.15 konstruierten bzgl. aller $p \in \Pi_n$ ω-orthogonalen Polynoms p_{n+1} sind und

$$a_j = \int_a^b \omega(x) l_j(x) dx$$

gilt.

Beweis: Weil $p_{n+1} \in \Pi_{n+1}$ ist, gilt $p_{n+1} = \alpha h_{n+1}$ mit $\alpha \neq 0$. Nach Lemma 1.12 ist

$$Q_n(f) = \int_a^b \omega(x)(L_n f)(x) dx,$$

welches nach Satz 1.13 eine Gauß'sche Quadraturformel ist, wenn $(h_{n+1}, p) = 0$ für alle $p \in \Pi_n$. Das gilt, weil

$$\begin{aligned}(h_{n+1}, p) &= \frac{1}{\alpha}(p_{n+1}, p) = \frac{1}{\alpha}\left(p_{n+1}, \sum_{i=0}^{n} \lambda_i p_i\right) \\ &= \frac{1}{\alpha} \sum_{i=0}^{n} \lambda_i (p_{n+1}, p_i) = 0.\end{aligned}$$

QED

Es gilt also $Q_n(p) = I_\omega(p)$ für alle $p \in \Pi_{2n+1}$.

Im Gegensatz zu den Newton-Côtes-Formeln gilt für die Gauß'schen Quadraturformeln die folgende numerisch wertvolle Eigenschaft:

Lemma 1.17 *Die Gewichte a_i der Gauß'schen Quadraturformeln sind positiv.*

Beweis: Seien $x_0, \ldots, x_n$ die Stützstellen der Quadraturformel Q_n. Nach Konstruktion sind sie die Nullstellen von p_{n+1} bzgl. ω. Wir definieren

$$h_{n+1}(x) := \prod_{j=0}^{n}(x - x_j) \text{ und } f_i(x) = \left(\frac{h_{n+1}(x)}{x - x_i}\right)^2, \qquad i = 0, \ldots, n.$$

Es ist also $f_i \in \Pi_{2n}[a, b]$ und nach Satz 1.16 gilt

$$0 < \int_a^b \omega(x) f_i(x) dx = Q_n(f_i) = \sum_{j=0}^{n} a_j f_j(x_j) = a_i f_i(x_i).$$

Weil $f_i(x_i) > 0$ folgt auch, dass $a_i > 0$. QED

Übungsaufgabe: Beweisen Sie, dass es keine Quadraturformel der Form $\sum_{j=0}^{n} a_j f_j(x_j)$ geben kann, die auf Π_{2n+2} exakt ist.

Zum Abschluss folgen einige Beispiele für Gauß-Quadraturen.

1. Sei $I = [-1, 1]$ und $\omega \equiv 1$. Die orthogonalen Polynome $p_0, p_1, \ldots$ sind die sogenannten *Legendre-Polynome*

$$L_n(x) := \frac{1}{2^n n!} \frac{d^n}{dx^n}(x^2 - 1)^n,$$

genauer:

$$\begin{aligned} L_0 &= 1 & L_3 &= x^3 - \tfrac{3}{5}x \\ L_1 &= x & L_4 &= x^4 - \tfrac{6}{7}x^2 + \tfrac{3}{35} \\ L_2 &= x^2 - \tfrac{1}{3} & L_5 &= \ldots \end{aligned}$$

L_i ist orthogonal auf Π_{i-1}, d.h. $\int_{-1}^{1} L_i(x) p(x) dx = 0$ für alle $p \in \Pi_{i-1}[-1, 1]$.

Beispiel: Es gilt

$$\int_{-1}^{1} L_2(x) p(x) = 0, \text{ für alle } p \in \Pi_1[-1, 1].$$

Das kann man nachrechnen:

$$\begin{aligned} &\int_{-1}^{1} L_2(x) p(x) dx = \int_{-1}^{1} (x^2 - \tfrac{1}{3})(ax + b) dx \\ &= \int_{-1}^{1} ax^3 + bx^2 - \tfrac{1}{3}ax - \tfrac{1}{3}b dx \\ &= \left[\frac{a}{4}x^4 + \frac{b}{3}x^3 - \frac{a}{6}x^2 - \frac{b}{3}x\right]_{-1}^{1} \\ &= \frac{a}{4} + \frac{b}{3} - \frac{a}{6} - \frac{b}{3} - \left(\frac{a}{4} - \frac{b}{3} - \frac{a}{6} + \frac{b}{3}\right) = 0. \end{aligned}$$

Wie sehen die zugehörigen Quadraturen aus?

$n = 0$

- Die Stützstellen sind die Nullstellen des orthogonalen Polynoms aus Π_1 (Satz 1.16). Die einzige Nullstelle von L_1 ist $x_0 = 0$. Daraus folgt:
$$Q_0(f) = a_0 f(x_0) = a_0 f(0).$$
- Das Gewicht a_0 ergibt sich dadurch, dass z.B. $1 \in \Pi_1 = \Pi_{2n+1}$ exakt integriert wird, also
$$2 = \int_{-1}^{1} 1 dx = a_0 f(0) = a_0.$$

Wir erhalten:
$$Q_0(f) = 2f(0)$$
integriert alle linearen Polynome auf $[-1, 1]$ exakt.

$n = 1$

- Stützstellen sind Nullstellen von L_2, also $\pm\sqrt{\frac{1}{3}}$. Es folgt:
$$Q_1(f) = a_0 f\left(-\sqrt{\tfrac{1}{3}}\right) + a_1 f\left(\sqrt{\tfrac{1}{3}}\right).$$
- Die Gewichte folgen aus den Exaktheitsbedingungen z.B. für $1, x \in \Pi_1 \subseteq \Pi_{2n+1}$:
$$\left.\begin{aligned} 2 &= \int_{-1}^{1} 1 dx = a_0 + a_1 \\ 0 &= \int_{-1}^{1} x dx = -a_0\sqrt{\tfrac{1}{3}} + a_1\sqrt{\tfrac{1}{3}} \end{aligned}\right\} \Rightarrow a_0 = a_1 = 1.$$

Es gilt also:
$$Q_1(f) = f\left(-\sqrt{\tfrac{1}{3}}\right) + f\left(\sqrt{\tfrac{1}{3}}\right)$$
integriert alle Polynome vom Grad bis 3 exakt!

2. $I = [-1, 1]$ und $\omega(x) = \frac{1}{\sqrt{1-x^2}}$. Die orthogonalen Polynome sind die sogenannten *Tschebyscheff-Polynome*
$$T_n(x) := \cos(n \arccos(x)).$$

Mithilfe der Additionstheoreme erhält man die folgende Darstellung:
$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_{n+1}(x) + T_{n-1}(x) &= 2xT_n(x). \end{aligned}$$

Daraus folgt $T_n \in \Pi_n$.

Lemma 1.18 *Es gilt:*

$$\int_{-1}^{1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}dx = \begin{cases} \pi & n = m = 0 \\ \frac{\pi}{2} & n = m > 0 \\ 0 & n \neq m \end{cases}$$

und die Nullstellen von T_n sind:

$$x_i = \cos\left(\frac{2i+1}{2n}\pi\right) \qquad i = 0, \ldots, n-1.$$

Mit diesen Nullstellen erhalten wir die Gauß-Tschebyscheff Quadratur:

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}dx \approx Q_{n-1}(f) = \sum_{i=0}^{n-1} a_i f\left(\cos\frac{2i+1}{2n}\pi\right).$$

Die Gewichte ergeben sich aus den Exaktheitsbedingungen für T_m, $m = 0, \ldots, n-1$ und Lemma 1.18 zu $a_i = \frac{\pi}{n}$, $i = 0, \ldots, n-1$.

Übungsaufgabe: Zeigen Sie, dass $a_i = \frac{\pi}{n}$, $i = 0, \ldots, n-1$ die Gewichte der Gauß-Tschebyscheff-Quadraturformel Q_n sind!

Damit erhält man abschließend

$$Q_{n-1}(f) = \frac{\pi}{n}\sum_{i=0}^{n-1} f\left(\cos\frac{2i+1}{2n}\pi\right), \qquad n \in \mathbb{N}.$$

Bemerkung: Analog zu Kapitel 1.2 kann man auch zusammengesetzte Gauß-Quadraturen (vorzugsweise niedriger Ordnung) betrachten.

1.4 Fehleranalyse

Was für ein Fehler entsteht, wenn man das exakte Integral durch eine Quadraturformel annähert?
Seien zunächst für $n \in \mathbb{N}$ die Stützstellen

$$x_0^{(n)} < x_1^{(n)} < \cdots < x_n^{(n)}$$

mit den zugehörigen Gewichten $a_0^{(n)}, \ldots, a_n^{(n)}$ gegeben und sei

$$Q_n(f) = \sum_{j=0}^{n} a_j^{(n)} f(x_j^{(n)}).$$

Der zugehörige Fehler ist dann:

$$R_n(f) := I_\omega(f) - Q_n(f) = \int_a^b \omega(x) f(x) dx - \sum_{j=0}^{n} a_j^{(n)} f(x_j^{(n)}).$$

Es werfen sich einige Fragen auf:

- Wie groß ist $R_n(f)$ für festes n?
- Konvergiert $R_n(f) \to 0$ für $n \to \infty$?

Satz 1.19 *Sei Q_n eine Folge von Quadraturformeln über dem endlichen Intervall $[a, b]$. Gilt*

1. *$Q_n(p) \to I_\omega(p)$ für $n \to \infty$ für alle Polynome p und*
2. *$\sum_{j=0}^{n} |a_j^{(n)}| \leq C$ für alle $n \in \mathbb{N}$ mit einer Konstante $C > 0$,*

so konvergiert die Folge $Q_n(f)$ gegen $I_\omega(f)$ für jedes $f \in C[a, b]$.

Beweis: Wir verwenden einen Satz aus der Approximationstheorie, nämlich den Satz von Weierstrass: Jede stetige Funktion auf einem kompakten Intervall lässt sich beliebig gut durch Polynome annähern. Genauer gibt es zu jedem $\varepsilon > 0$ ein Polynom p mit $\|f - p\|_\infty := \max_{x \in [a,b]} |f(x) - p(x)| < \varepsilon$.
Sei $\varepsilon > 0$. Wähle p so, dass $\|f - p\|_\infty < \varepsilon$ und N so, dass $|I_\omega(p) - Q_n(p)| < \varepsilon$ für alle $n \geq N$. Dann gilt für alle $n \geq N$:

$$|I_\omega(f - p)| \leq \int_a^b |\omega(x)| \underbrace{|f(x) - p(x)|}_{<\varepsilon} dx < \varepsilon \int_a^b \omega(x) dx$$

$$|Q_n(f - p)| \leq \sum_{j=0}^{n} |a_j^{(n)}| \underbrace{|f(x_j^{(n)}) - p(x_j^{(n)})|}_{<\varepsilon} < \varepsilon \sum_{j=0}^{n} |a_j^{(n)}| \leq \varepsilon C$$

Es ergibt sich:

$$\begin{aligned}
|R_n(f)| &= |R_n((f - p) + p)| \\
&= |R_n(f - p) + R_n(p)| \\
&\leq |R_n(f - p)| + |R_n(p)| \\
&\leq |I_\omega(f - p)| + |Q_n(f - p)| + |I_\omega(p) - Q_n(p)| \\
&\leq \left(\int_a^b \omega(x) dx + C + 1 \right) \cdot \varepsilon
\end{aligned}$$

für alle $n \geq N$, also konvergiert $R_n(f)$ gegen Null. QED

Bedingung (1) des Satzes ist erfüllt, wenn alle Q_n interpolatorische Quadraturformeln sind. Sind weiter alle Gewichte nicht negativ, dann gilt:

$$\sum_{j=0}^{n} |a_j^{(n)}| = \sum_{j=0}^{n} a_j^{(n)} \cdot 1 = \int_a^b \omega(x) \cdot 1 dx = C.$$

Satz 1.20 *Für jeden stetigen Integranden auf dem Intervall $[a, b]$ konvergiert jede Folge von Gauß-Quadraturen Q_n gegen das Integral.*

Beweis: Nach Lemma 1.17 sind die Gewichte der Gauß-Quadraturen positiv, daher ist

$$\sum_{j=0}^{n} |a_j^{(n)}| \leq C$$

erfüllt. Bedingung (1) von Satz 1.19 gilt nach Satz 1.16, also folgt die Behauptung wegen Satz 1.19. QED

Leider ist die Aussage von Satz 1.20 für die Newton-Côtes-Formeln im Allgemeinen falsch und es lassen sich Gegenbeispiele konstruieren. (Der Grund dafür liegt in den negativen a_i, die in den Newton-Côtes-Formeln ab Grad 8 vorkommen.) Wir entwickeln nun Fehlerabschätzungen für die Quadraturformeln auf festen Intervallen $[a, b]$:

$$R_n(f) = I_\omega(f) - Q_n(f) = I_\omega(f) - I_\omega(L_n f) = I_\omega(f - L_n f),$$

wobei $L_n f$ das Interpolationspolynom zu f an den n Stützstellen von Q_n ist. Nun kann man die Fehlerabschätzungen für $f - L_n f$ (Numerik I, Korollar 6.15) heranziehen. Bessere Ergebnisse liefert aber der folgende (allgemeinere) Ansatz, bei dem man ausnutzt, dass für alle $p \in \Pi_m$ gilt:

$$R_n(p) = 0,$$

wobei man m im Falle von Newton-Côtes $\leq n$ wählen muss, falls n ungerade ist und $m = n+1$, falls n gerade ist. Bei der Gauß-Quadratur hingegen ist $m \leq 2n+1$ möglich.

Notation 1.21 *Sei $t \in \mathbb{R}$. Dann bezeichne*

$$z_{t,m}^+(x) = (x-t)_+^m = \begin{cases} (x-t)^m & \text{falls } x \geq t \\ 0 & \text{sonst} \end{cases}$$

und

$$K_m(t) := \frac{1}{m!} R_n(z_{t,m}^+) \text{ mit } t \in [a, b]$$

den Peano-Kern *bzgl. m und R_n.*

Bemerkung: Wir können die $z_{t,m}^+(x)$ sowohl als Funktion in x als auch in t auffassen.

Wir verwenden $z_{t,m}^+$ für folgende Umformulierung:

$$\int_a^x \frac{(x-t)^m}{m!} f^{(m+1)}(t)dt = \int_a^b \frac{(x-t)_+^m}{m!} f^{(m+1)}(t)dt,$$

das heißt, um bei den im Folgenden auftretenden Integralen die obere Grenze von x unabhängig zu machen. Wir erhalten:

Satz 1.22 *Sei Q_n eine auf $\Pi_m(\mathbb{R})$ exakte Quadraturformel. Dann gilt für jedes $f \in C^{m+1}[a,b]$:*

$$R_n(f) = \int_a^b K_m(t) f^{(m+1)}(t)dt.$$

Wechselt K_m auf $[a,b]$ das Vorzeichen nicht, so existiert ein $\xi \in [a,b]$ so, dass

$$\begin{aligned} R_n(f) &= f^{(m+1)}(\xi) \int_a^b K_m(t)dt \\ &= \frac{f^{(m+1)}(\xi)}{(m+1)!} R_n(x^{m+1}). \end{aligned}$$

Beweis: Wir verwenden die Taylor-Entwicklung von f mit Integral-Restglied zum Entwicklungspunkt a:

$$\begin{aligned} f(x) &= \sum_{j=0}^m \frac{f^{(j)}(a)}{j!}(x-a)^j + \int_a^x \frac{(x-t)^m}{m!} f^{(m+1)}(t)dt \\ &= \sum_{j=0}^m \frac{f^{(j)}(a)}{j!}(x-a)^j + \int_a^b \frac{(x-t)_+^m}{m!} f^{(m+1)}(t)dt. \end{aligned}$$

Daraus folgt:

$$\begin{aligned} R_n(f) &= R_n\left(\sum_{j=0}^m \frac{f^{(j)}(a)}{j!}(x-a)^j\right) + R_n\left(\int_a^b \frac{z_{t,m}^+}{m!} f^{(m+1)}(t)dt\right) \\ &= 0 + \int_a^b \frac{1}{m!} R_n(z_{t,m}^+) f^{(m+1)}(t)dt \\ &= \int_a^b K_m(t) f^{(m+1)}(t)dt, \end{aligned}$$

wobei im zweiten Schritt verwendet wurde, dass R_n auf Π_m verschwindet und dass man das Integral mit R_n nach Fubini vertauschen darf und dass Q_n nur aus

Punktauswertungen besteht. Für den Beweis der zweiten Aussage benutzen wir den ersten Mittelwertsatz der Integralrechnung und erhalten

$$R_n(f) = \int_a^b K_m(t) f^{(m+1)}(t) dt = f^{(m+1)}(\xi) \cdot \int_a^b K_m(t) dt \tag{1.1}$$

mit $\xi \in (a, b)$, da $f^{(m+1)}$ und K_m stetig sind und $K_m(t) \neq 0$ für alle $t \in (a, b)$ gilt. Setzt man in (1.1) nun $f(x) = x^{m+1}$ ein, so ergibt sich:

$$R_n(x^{m+1}) = (m+1)! \int_a^b K_m(t) dt,$$

also

$$\frac{R_n(x^{m+1})}{(m+1)!} \cdot f^{(m+1)}(\xi) = f^{(m+1)}(\xi) \cdot \int_a^b K_m(t) dt.$$

QED

Wir betrachten im Folgenden einige Anwendungen von Satz 1.22: Wir leiten Fehlerschranken her für die Trapez-Regel, die Simpson-Regel, die zusammengesetzte Trapez-Regel, die zusammengesetzte Simpson-Regel und für die Gauß-Quadratur. Wir beginnen mit der Trapez-Regel.

Fehlerabschätzung für die Trapez-Regel.

Trapez-Regel: Es ist $n = 1$ und sie ist exakt für $m = 1$, für den Peano-Kern ergibt sich:

$$\begin{aligned}
K_1(t) &= \frac{1}{1!} R_1(z_{t,1}^+) \\
&= \int_t^b (x-t)^1 dx - \frac{b-a}{2} [\underbrace{z_{t,1}(a)}_{=0} + \underbrace{z_{t,1}(b)}_{=(b-t)}] \\
&= \left[\frac{1}{2}(x-t)^2 \right]_t^b - \frac{b-a}{2}(b-t) \\
&= \frac{1}{2} \left[(b-t)^2 - (b-a)(b-t) \right] \\
&= \frac{1}{2}(b-t)[b-t-b+a] \\
&= \frac{1}{2}(b-t)(a-t).
\end{aligned}$$

Da $K_1(t) \leq 0$ für alle $t \in [a,b]$ gilt, können wir den zweiten Teil von Satz 1.22 zum Abschätzen verwenden. Wir erhalten:

$$\begin{aligned} R_1(x^2) &= \int_a^b x^2 dx - \frac{b-a}{2}(a^2+b^2) \\ &= \frac{1}{3}b^3 - \frac{1}{3}a^3 - \frac{1}{2}b^3 + \frac{1}{2}a^3 - \frac{a^2b}{2} + \frac{ab^2}{2} \\ &= \frac{1}{6}a^3 - \frac{1}{6}b^3 - \frac{a^2b}{2} + \frac{ab^2}{2} \\ &= \frac{1}{6}(a-b)^3. \end{aligned}$$

Also existiert zu jedem $f \in C^2[a,b]$ ein $\xi \in [a,b]$ mit

$$R_1(f) = \frac{f''(\xi)}{2} \cdot \frac{1}{6}(a-b)^3 = -\frac{h^3}{12} f''(\xi) \text{ mit } h = \frac{b-a}{1}. \tag{1.2}$$

Fehlerabschätzung für die Simpson-Regel.

Es sind $n = 2$, $m = 3$, dann gilt nach einiger Rechnerei:

$$K_3(t) = \begin{cases} -\frac{(t-a)^3}{72}(a+2b-3t) & \text{für } a \leq t \leq \frac{a+b}{2} \\ -\frac{(b-t)^3}{72}(3t-2a-b) & \text{für } \frac{a+b}{2} \leq t \leq b \end{cases}$$

und $K_3(t) \leq 0$ für alle $t \in [a,b]$. Außerdem gilt:

$$R_2(x^4) = -\frac{(b-a)^5}{120}.$$

Daraus folgt:

$$R_2(f) = -\frac{(b-a)^5}{2880} f^{(4)}(\xi) = -\frac{h^5}{90} f^{(4)}(\xi) \text{ mit } h = \frac{b-a}{2}.$$

Fehlerabschätzung für die zusammengesetzte Trapez-Regel.

Seien nun $x_0, \ldots, x_n$ gegeben. Sei

$$T_h = h \cdot \left(\frac{f(a)}{2} + \sum_{j=1}^{n-1} f(x_j) + \frac{f(b)}{2} \right)$$

die zusammengesetzte Trapez-Regel.

Satz 1.23 *Ist $f \in C^2[a,b]$ und $\Pi_n(f)$ die Näherung an $\int_a^b f(x)dx$ aus der zusammengesetzten Trapezregel (siehe Abschnitt 1.2), so gilt für den Fehler*

$$R(f) = I(f) - T_h(f) = -\frac{h^2(b-a)}{12} f''(\xi)$$

mit einem ξ aus $[a,b]$.

Beweis: Wir benutzen (1.2) auf jedem Teilintervall $[x_j, x_{j+1}]$, das heißt es existiert ein $\xi_j \in [x_j, x_{j+1}]$ mit

$$\begin{aligned}\int_a^b f(x)dx &= \sum_{j=0}^{n-1} \int_{x_j}^{x_{j+1}} f(x)dx \\ &= \sum_{j=0}^{n-1} \left(\frac{h}{2}(f(x_j) + f(x_{j+1})) - \frac{h^3}{12} f''(\xi_j) \right) \\ &= T_h(f) - \frac{h^3}{12} \cdot \sum_{j=0}^{n-1} f''(\xi_j) \\ &= T_h(f) - \frac{h^2(b-a)}{12} \cdot \frac{1}{n} \sum_{j=0}^{n-1} f''(\xi_j)\end{aligned}$$

Sei $\frac{1}{n}\sum_{j=0}^{n-1} f''(\xi_j) =: c$. Weil

$$\min_{j=0,\ldots,n-1} f''(\xi_j) \leq c \leq \max_{j=0,\ldots,n-1} f''(\xi_j)$$

gilt und f'' stetig ist, gibt es nach dem Zwischenwertsatz ein $\xi \in [a, b]$ mit $f(\xi) = c$. Also folgt

$$\int_a^b f(x)dx = T_h(f) - \frac{h^2(b-a)}{12} \cdot f''(\xi)$$

und daraus schließlich

$$R(f) = I(f) - T_h(f) = -\frac{h^2(b-a)}{12} \cdot f''(\xi).$$

QED

Bemerkung: Im Fall der Trapez-Regel liegt also sogar quadratische Konvergenz vor!

Fehlerabschätzung für die zusammengesetzte Simpson-Regel.

Für die zusammengesetzte Simpson-Regel ergibt sich

$$R(f) = I(f) - S_h(f) = -\frac{h^4(b-a)}{180} f^{(4)}(\xi) \text{ mit } \xi \in [a, b].$$

Fehlerabschätzung für die Gauß-Quadratur.

Lemma 1.24 *Sei Q_n die Gauß-Quadratur in $n+1$ Punkten aus $[a,b]$ mit zulässiger Gewichtsfunktion ω. Dann hat der zugehörige Peano-Kern K_m für $0 \leq m \leq 2n+1$ genau $2n+1-m$ Nullstellen in $[a,b]$. Insbesondere wechselt K_{2n+1} auf $[a,b]$ das Vorzeichen nicht.*

Satz 1.25 *Sei Q_n die Gauß-Quadratur in $n+1$ Punkten aus $[a,b]$ mit zulässiger Gewichtsfunktion ω. Zu $f \in C^{2n+2}[a,b]$ gibt es ein $\xi \in [a,b]$ so, dass*

$$R_n(f) = I_\omega(f) - Q_n(f) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_a^b \omega(x)(h_{n+1}(x))^2 dx,$$

wobei wie üblich $h_{n+1}(x) = \prod_{j=0}^{n}(x - x_j)$.

Beweis: Wegen Lemma 1.24 darf man den 2. Teil von Satz 1.22 anwenden. Man erhält

$$R_n(f) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} R_n(x^{2n+2})$$

und rechnet nach, dass

$$R_n(x^{2n+2}) = \int_a^b \omega(x)(h_{n+1}(x))^2$$

gilt. QED

1.5 Romberg-Verfahren

Wir benötigen folgende Begriffe.

Definition 1.26 *Die* Bernoulli-Polynome $B_k \in \Pi_k$ *für* $k = 0, 1, \ldots$ *sind rekursiv definiert durch:*

$$B_0(t) := 1$$
$$\text{und } B_k'(t) = B_{k-1}(t) \text{ und } \int_0^1 B_k(t)dt = 0 \quad k = 1, 2, \ldots.$$

Die Zahlen $b_k := k!B_k(0)$ heißen Bernoulli-Zahlen.

Das $(k+1)$-te Bernoulli-Polynom entsteht also durch Integration aus dem k-ten, wobei die zweite Bedingung die Integrationskonstanten festlegt. Es gilt:

$$B_1'(t) = 1 \Rightarrow B_1(t) = t + C.$$

Wegen

$$\int_0^1 t + C = \left[\frac{1}{2}t^2 + Ct\right]_0^1 = \frac{1}{2} + C \stackrel{!}{=} 0$$

folgt $C = \frac{1}{2}$, also

$$B_1(t) = t - \frac{1}{2}$$

Ähnlich ergibt sich

$$B_2(t) = \frac{1}{2}t^2 - \frac{1}{2}t + \frac{1}{12}.$$

Lemma 1.27 *Es gilt:*

1. $B_k \in \Pi_k$ *für* $k = 1, 2, \ldots$.
2. $B_k(t) = (-1)^k B_k(1-t)$ *für* $k = 0, 1, 2, \ldots$.
3. $B_k(0) = B_k(1)$ *für* $k = 2, 3, \ldots$.
4. *Für* $m = 1, 2, \ldots$ *besitzt das Polynom* $B_{2m} - B_{2m}(0)$ *genau die Nullstellen* 0 *und* 1 *im Intervall* $[0, 1]$ *und das Polynom* B_{2m+1} *genau die Nullstellen* 0, $\frac{1}{2}$ *und* 1.

Mit Hilfe der Bernoulli-Zahlen untersuchen wir nochmals den bei der zusammengesetzten Trapez-Regel entstehenden Fehler in Abhängigkeit der Intervalllänge h. Sei dazu wie bisher

$$T_h(f) = h\left(\tfrac{1}{2}f(a) + \sum_{j=1}^{m-1} f(x_j) + \tfrac{1}{2}f(b)\right)$$

mit $h = \frac{b-a}{n}$ und $x_0 = a$, $x_j = a + jh$ und $x_m = b$ der Wert der zusammengesetzten Trapezregel.

Satz 1.28 (Euler-McLaurinsche Summenformel) *Sei* $l \in \mathbb{N}$ *und* $f \in C^{2l}([a, b])$. *Dann gilt*

$$T_h(f) \;=\; I(f) \;+\; \sum_{j=1}^{l-1} \frac{b_{2j} h^{2j}}{(2j)!}\left[f^{(2j-1)}(b) \;-\; f^{(2j-1)}(a)\right] \;+\; \frac{(b-a) b_{2l} h^{2l}}{(2l)!} f^{(2l)}(\xi)$$

für ein $\xi = \xi(h) \in (a, b)$.

Beweisidee: Partielle Integration von $\int f(t)dt = \int B_0(\frac{t-a}{h}) f(t)$ und Mittelwertsatz der Integralrechnung.

Korollar 1.29 *Ist f periodisch auf $[a,b]$ und genügt den Voraussetzungen aus Satz 1.28, so gibt es ein $\xi \in (a,b)$, so dass*

$$T_h(f) = I(f) + \frac{(b-a)b_{2l}h^{2l}}{(2l)!} f^{(2l)}(\xi).$$

Beweis: Da f periodisch auf $[a,b]$ ist, gilt

$$f^{(i)}(b) = f^{(i)}(a) \text{ für alle } i = 0, 1, \ldots, 2l.$$

QED

Im Wesentlichen besagt die Euler-McLaurinsche Formel also, dass man den Fehler bei der zusammengesetzten Trapezregel schreiben kann als

$$T_h(f) - I(f) = a_2 h^2 + a_4 h^4 + \cdots + a_{2l-2} h^{2l-2} + a_{2l}(h) h^{2l}$$

mit Koeffizienten $a_2, a_4, \ldots, a_{2l-2} \in \mathbb{R}$ und einer Funktion $a_{2l} : \mathbb{R} \to \mathbb{R}$. Genauer gilt:

$$a_{2j} = \frac{b_{2j}}{(2j)!}\left(f^{(2j-1)}(b) - f^{(2j-1)}(a)\right)$$

$$\text{und} \qquad a_{2l}(h) = \frac{(b-a)b_{2l}}{(2l)!} f^{(2l)}(\xi(h)). \tag{1.3}$$

Weil $f^{(2l)}(\xi)$ als stetige Funktion auf $[a,b]$ beschränkt ist, ist auch a_{2l} beschränkt, weshalb gilt

$$\lim_{h \to 0} T_h(f) = I(f).$$

Allerdings geht der Rechenaufwand für $h \to 0$ gegen Unendlich. Die Idee des Romberg-Verfahrens ist es nun, $T_0(f)$ durch „Extrapolation“ folgendermaßen abzuschätzen: Setze $\tau = h^2$ als das Quadrat der Intervalllänge.

1. Sei $g(\tau) := T_{\sqrt{\tau}}(f)$ für alle $\tau \neq 0$, $g(0) := T_0(f)$.
2. Bestimme $g(\tau_0), \ldots, g(\tau_l)$ für $l+1$ Stützstellen $\tau_j := h_j^2$ für Intervalllängen $h_0, \ldots, h_l$ mit $h_j = \frac{b-a}{m_j}$, $m_j \in \mathbb{N}$.
3. Interpoliere g an den $l+1$ Stützstellen durch ein Polynom $p \in \Pi_l$, also mit
$$p(\tau_j) = g(\tau_j) = T_{h_j}(f), \qquad j = 0, \ldots, l.$$
4. Approximiere
$$I(f) = \int_a^b f(x)dx = \lim_{h \to 0} T_h(f) \approx \lim_{h \to 0} p(h^2) = p(0).$$

Weil wir das Interpolationspolynom nicht selbst kennen müssen, sondern nur an seinem Wert an der Stelle 0 interessiert sind, bietet sich zur Berechnung das Verfahren von Neville-Aitken (siehe Numerik I) an.

Nach Satz 6.6 aus Numerik I kann folgende Formel verwendet werden: sei $P_i^k(\tau)$ das Polynom, das g an den Stützstellen $\tau_i, \tau_{i+1}, \ldots, \tau_{i+k}$ interpoliert. Dann gilt:

$$P_i^{k+1}(\tau) = \frac{(\tau - \tau_i)P_{i+1}^k - (\tau - \tau_{i+k+1})P_i^k}{\tau_{i+k+1} - \tau_i}.$$

Für $P_i^k := P_i^k(0)$ gilt somit

$$P_i^k = \frac{\tau_i P_{i+1}^{k-1} - \tau_{i+k} P_i^{k-1}}{\tau_i - \tau_{i+k}}$$

und die Werte lassen sich berechnen durch folgendes Schema:

$$\begin{array}{lllll}
h_0 & P_0^0 = T_{h_0}(f) & \searrow & & \\
 & & P_0^1 & \searrow & \\
h_1 & P_1^0 = T_{h_1}(f) & \nearrow\ \searrow & & P_0^2 \quad \searrow \\
 & & P_1^1 & \nearrow\ \searrow & \qquad\qquad P_0^3 \\
h_2 & P_2^0 = T_{h_2}(f) & \nearrow\ \searrow & & P_1^2 \quad \nearrow \\
 & & P_2^1 & \nearrow & \\
h_3 & P_3^0 = T_{h_3}(f) & \nearrow & &
\end{array}$$

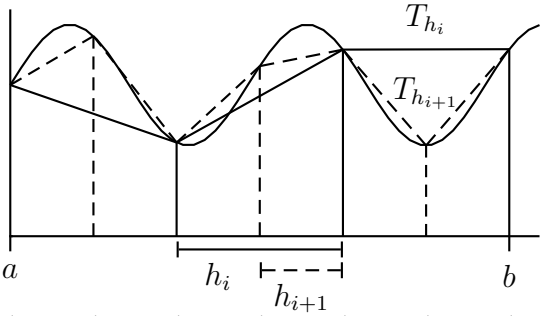
Jeder der Einträge stellt eine eigene Quadraturformel dar. Verwendet man z.B.

$$\tau_i = 2^{-2i} h_0^2 \text{ bzw. } h_i = 2^{-i} h_0,$$

so erhält man aus der ersten Spalte:

$$
\begin{aligned}
P_i^1 &= \frac{2^{-2i}h_0^2 P_{i+1}^0 - 2^{-2i-2}h_0^2 P_i^0}{2^{-2i}h_0^2 - 2^{-2i-2}h_0^2} \\
&= \frac{T_{h_{i+1}}(f) - \frac{1}{4}T_{h_i}(f)}{1-\frac{1}{4}} = \tfrac{4}{3}T_{h_{i+1}}(f) - \tfrac{1}{3}T_{h_i}(f) \\
&= \tfrac{4}{3}\left(\tfrac{1}{2}f(a) + \sum_{j=1}^{2m_i-1} f(a+\tfrac{1}{2}jh_i) + \tfrac{1}{2}f(b)\right)h_{i+1} \\
&\quad - \tfrac{1}{3}\left(\tfrac{1}{2}f(a) + \sum_{j=1}^{m_i-1} f(a+jh_i) + \tfrac{1}{2}f(b)\right)h_i \\
&= h_{i+1}\Bigg(\tfrac{1}{3}f(a) + \tfrac{4}{3}f(a+\tfrac{1}{2}jh_i) + \tfrac{2}{3}f(a+jh_i) \\
&\quad + \tfrac{4}{3}f(a+\tfrac{3}{2}jh_i) + \cdots + \tfrac{1}{3}f(b)\Bigg),
\end{aligned}
$$

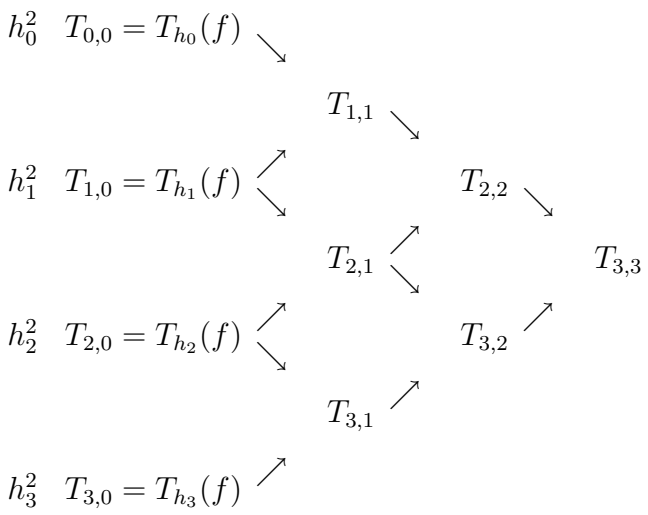
wobei wir im letzten Schritt $\frac{h_i}{h_{i+1}} = 2$ verwendet haben.



Gewichte:							
erster Term:	$\frac{4}{6}$	$\frac{4}{3}$	$\frac{4}{3}$	$\frac{4}{3}$	$\frac{4}{3}$	$\frac{4}{3}$	$\frac{4}{6}$
zweiter Term:	$-\frac{2}{6}$	0	$-\frac{2}{3}$	0	$-\frac{2}{3}$	0	$-\frac{2}{6}$
gesamt:	$\frac{1}{3}$	$\frac{4}{3}$	$\frac{2}{3}$	$\frac{4}{3}$	$\frac{2}{3}$	$\frac{4}{3}$	$\frac{1}{3}$

Man erhält also genau die zusammengesetzte Simpson-Regel.

Bemerkung: In der Literatur wird das Schema anders nummeriert, nämlich durch



mit $T_{i,k} = P_{i-k}^k$ bzw. $P_i^k = T_{i+k,k}$ und

$$T_{i,k} = \frac{h_{i-k}^2 T_{i,k-1} - h_i^2 T_{i-1,k-1}}{h_{i-k}^2 - h_i}.$$

Zum Abschluss untersuchen wir, wann Romberg-Quadraturen exakt sind.

Satz 1.30 *Die Romberg-Quadraturen $P_i^k(f)$ (bzw. $T_{i+k,k}(f)$) sind exakt für Polynome vom Grad kleiner gleich $2k$.*

Beweis: Ist $f \in \Pi_{2k}$, so folgt, dass $f^{(2k)}$ konstant ist, also ist $a_{2k}(h)$ in (1.3) auf Seite 26 konstant. Es gilt $a_{2k}(h) = a_{2k}$. Nach Satz 1.28 erhalten wir, dass

$$T_h(f) = I(f) + a_2 h^2 + a_4 h^4 + \cdots + a_{2k-2} h^{2k-2} + a_{2k} h^{2k}$$

ein Polynom vom Grad kleiner gleich k in der Variablen h^2 ist, beziehungsweise dass

$$g(\tau) = T_{\sqrt{\tau}}(f) = I(f) + a_2 \tau + a_4 \tau^2 + \cdots + a_{2k} \tau^k$$

ein Polynom aus Π_k ist. Aufgrund der Eindeutigkeit der Polynominterpolation folgt

$$p(\tau) \equiv g(\tau).$$

Insbesondere gilt $p(0) = g(0)$, also

$$P_i^k(f) = p(0) = g(0) = T_0(f) = I(f).$$

QED

Wie wählt man die Schrittweiten $h_k = \frac{b-a}{n_k}$? Dazu gibt es die

- klassische Romberg-Folge:

  $$n_k = 2^k \Rightarrow h_k = \tfrac{1}{2} h_{k-1}.$$

  Der Vorteil liegt darin, dass Funktionsauswertungen von einem Schritt i auf den nächsten Schritt $i+1$ wiederverwendet werden können. Der Nachteil ist, dass die Folge sehr schnell wächst!

- harmonische Folge:

  $$n_k = k + 1.$$

  Im Gegensatz zur klassischen Romberg-Folge wächst diese langsamer, doch sind alte Funktionsauswertungen im $(i+1)$-ten Schritt unbrauchbar. Daher wählt man als Kompromiss die

- Burlisch-Folge:

  $$\begin{aligned} n_0 &= 1 \\ n_{2k-1} &= 2^k \\ n_{2k} &= 3 \cdot 2^{k-1}. \end{aligned}$$

1.6 Zusammenfassung

Ziel:

- *Einfache Formel für*

$$I_\omega(f) = \int_a^b \omega(x) f(x) dx$$

- *"einfach": Quadraturformeln*

$$Q_n(f) := \sum_{i=0}^{n} a_i f(x_i)$$

Interpolationsquadraturen

Seien $x_0, \ldots, x_n$ gegeben. Sei

$$Q_n(f) := \int_a^b (L_n f)(x) dx.$$

- Quadratur Q_n Interpolationsquadratur $\Leftrightarrow$ Q_n $\forall p \in \Pi_n$ exakt
- $Q_n(f)$ ist eindeutig bestimmt
- Newton-Côtes Formeln: Trapez-Regel, Simpson-Regel, ...
  - n ungerade: exakt auf Π_n
  - n gerade: exakt auf Π_{n+1}

Zusammengesetzte Newton-Côtes Formeln

$$T_h(f) = h\left(\tfrac{1}{2} f(x_0) + \sum_{i=1}^{m-1} f(x_i) + \tfrac{1}{2} f(x_m)\right)$$

$$S_h(f) = \frac{h}{3}\left(\sum_{j=0}^{\frac{m}{2}-1} f(x_{2j}) + 4f(x_{2j+1}) + f(x_{2j+2})\right)$$

Gauß'sche Quadraturen

Wähle auch $x_0, \ldots, x_n$. Sei $Q_n(f)$ Gauß'sche Quadratur falls exakt $\forall p \in \Pi_{2n+1}$.

- $Q_n(f) := \int_a^b \omega(x)(L_n f)(x) dx$ Gauß'sche Quadratur
  $\Leftrightarrow \int_a^b \omega(x) h_{n+1}(x) p(x) dx (= (h_{n+1}, p)_\omega) = 0.$
- Konstruktion der ω-orthogonalen Polynome (Orthogonalbasis)

- $x_0, \ldots, x_n$ sind Nullstellen des Polynoms $p \in \Pi_n$, das $(p_{n+1}, f) = 0, \forall f \in \Pi_n$ erfüllt.
- Gewichte alle > 0!
  - $I = [-1, 1]$, $\omega \equiv 1 \Rightarrow$ Legendre-Polynome
  - $I = [-1, 1]$, $\omega(x) = \frac{1}{\sqrt{1-x^2}} \Rightarrow$ Tschebyscheff-Polynome

Fehleranalyse

- $Q_n(f) \to I_\omega(f), \forall f \in C[a, b]$, falls $Q_n(p) \to I_\omega(p), \forall p \in \Pi_\infty$ und $\sum_{j=0}^{n} |a_j^{(n)}| \leq C$, $\forall n$.
- Gauß-Quadraturen konvergieren
- Newton-Côtes nicht
- Restglied $R_n(f) = I_\omega(f) - Q_n(f)$
- Peano-Kern: $K_m(t) := \frac{1}{m!} R_n(z_{t,m}^+)$
- Q_n auf Π_m exakt. Dann
  - $R_n(f) = \int_a^b K_m(t) f^{(m+1)} t dt$
  - Wechselt K_m das Vorzeichen nicht, so existiert $\xi \in [a, b]$:

$$R_n(f) = \frac{f^{(m+1)}(\xi)}{(m+1)!} R_n(x^{m+1}).$$

- Trapez-Regel: $R_1(f) = -\frac{h^3}{12} f''(\xi)$
- Simpson-Regel: $R_2(f) = -\frac{h^5}{90} f^{(4)}(\xi)$
- zusammengesetzte Trapez-Regel: $R(f) = -\frac{h^2(b-a)}{12} f''(\xi)$
- zusammengesetzte Simpson-Regel: $R(f) = -\frac{h^4(b-a)}{180} f^{(4)}(\xi)$
- Gauß: $R_n(f) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_a^b \omega(x)(h_{n+1}(x))^2 dx$

Romberg-Verfahren

- Euler McLaurinsche Summenformel:

$$T_h(f) - I(f) = a_2 h^2 + a_4 h^4 + \cdots a_{2l+2} h^{2l-2} + a_{2l}(h) h^{2l}$$

- Extrapolation:

  1. $\tau = h^2$, $g(\tau) := T_h(f)$
  2. Bestimme $g(\tau_0), \ldots, g(\tau_l)$ mit $\tau_j = (\frac{b-a}{m_j})^2$
  3. Interpoliere g durch Polynom p (Neville-Aitken)
  4. $T_0(h) := p(0)$

- Romberg-Quadraturen P_i^k (Interpolationen $\tau_i, \ldots, \tau_{i+k}$)) exakt $\forall p \in \Pi_{2k}$

Kapitel 2

Approximationstheorie

In diesem Kapitel wollen wir eine Funktion f durch eine "einfache" Funktion u (mit $u \in U \subseteq C[a,b]$, z.B. $U = \Pi_n$) annähern. Bei der **Interpolation** sollte u an gegebenen Punkten mit f übereinstimmen (s. Numerik I, Kapitel 6). Bei der **Approximation** soll u die Funktion f im ganzen Definitionsbereich "gut" darstellen. Unter "gut" verstehen wir, dass $\|f-u\|$ klein ist und beschäftigen uns hauptsächlich mit der Tschebyscheff-Norm $\|f\|_\infty := \max_{x\in[a,b]} |f(x)|$.

2.1 Approximationssätze von Weierstraß

In diesem Abschnitt wollen wir den in Abschnitt 1.4 schon benutzten Satz von Weierstraß (Satz 1.19) beweisen. Dazu benutzen wir so genannte *Korovkin-Operatoren*.

Definition 2.1 *Eine Abbildung $K : C[a,b] \to C[a,b]$ heißt **monoton**, falls für alle $f, g \in C[a,b]$ gilt*

$$f(x) \le g(x), \quad \forall x \in [a,b] \quad \Rightarrow \quad Kf(x) \le Kg(x), \quad \forall x \in [a,b].$$

*Eine Folge $K_n : C[a,b] \to C[a,b]$, $n \in \mathbb{N}$ heißt **Korovkin-Folge**, falls*

(a) K_n ist monotoner, linearer Operator für alle $n \in \mathbb{N}$.

(b) $\lim_{n\to\infty} \|K_n f - f\|_\infty = 0$ für $f \in \{\mathbf{1}, x, x^2\}$ (gleichmäßige Konvergenz).

Bemerkung: Ist K_n Korovkin-Folge, so gilt

$$\lim_{n\to\infty} \|K_n f - f\|_\infty = 0, \quad \forall f \in \Pi_2,$$

denn: $f \in \Pi_2$ lässt sich schreiben als $f(x) = \alpha x + \beta x + \gamma \cdot \mathbf{1}$, also ist

$$\begin{aligned} \|K_n f - f\| &= \|\alpha K_n(x^2) + \beta K_n(x) + \gamma K_n(\mathbf{1}) - \alpha x^2 - \beta x - \gamma\|, \text{ da } K_n \text{ linear} \\ &\le \underbrace{|\alpha| \|K_n x^2 - x^2\|}_{\to 0} + \underbrace{|\beta| \|K_n x - x\|}_{\to 0} + \underbrace{|\gamma| \|K_n \mathbf{1} - \mathbf{1}\|}_{\to 0} \\ &\to 0 \text{ für } n \to \infty. \end{aligned}$$

Überraschenderweise folgt aus der gleichmäßigen Konvergenz auf Π_2 sogar die gleichmäßige Konvergenz für alle stetigen Funktionen!

Satz 2.2 *Ist $\{K_n\}$ eine Korovkin-Folge auf $C[a,b]$, so gilt*

$$\lim_{n\to\infty} \|K_n f - f\|_\infty = 0 \text{ für alle } f \in C[a,b].$$

Beweis: Ist f stetig auf $[a,b]$, so ist f sogar gleichmäßig stetig auf $[a,b]$, d.h. zu $\varepsilon > 0$ existiert ein $\delta > 0$, so dass

$$|f(x) - f(y)| \leq \frac{\varepsilon}{3} \text{ für alle } x, y \in [0,1] \text{ mit } |x-y| < \delta.$$

Sei nun $t \in [a,b]$ fest.

- Falls $|x-t| < \delta$ gilt also $|f(x) - f(t)| < \frac{\varepsilon}{3}$.
- Falls $|x-t| \geq \delta$, so gilt

$$\begin{aligned} |f(x) - f(t)| &\leq |f(x)| + |f(t)| \leq 2\|f\|_\infty \\ &\leq 2\|f\|_\infty \underbrace{\left(\frac{x-t}{\delta}\right)^2}_{\geq 1}. \end{aligned}$$

Zusammen erhalten wir

$$\forall x \in [a,b] : |f(x) - f(t)| \leq \underbrace{\frac{\varepsilon}{3}}_{\geq 0} + \underbrace{2\|f\|_\infty \left(\frac{x-t}{\delta}\right)^2}_{\geq 0}. \tag{2.1}$$

Seien nun

$$\begin{aligned} p_t(x) &= f(t) - \frac{\varepsilon}{3} - 2\|f\|_\infty \left(\frac{x-t}{\delta}\right)^2 \\ q_t(x) &= f(t) + \frac{\varepsilon}{3} + 2\|f\|_\infty \left(\frac{x-t}{\delta}\right)^2. \end{aligned}$$

Dann lässt sich (2.1) schreiben als

$$p_t(x) \leq f(x) \leq q_t(x), \quad \forall x \in [a,b]. \tag{2.2}$$

K_n ist monoton für alle n, also gilt $K_n p_t(x) \leq K_n f(x) \leq K_n q_t(x)$. Weil $p_t, q_t \in \Pi_2[a,b]$ konvergiert die Anwendung der K_n auf sie gleichmäßig (in x), d.h.

$$\begin{aligned} |K_n q_t(x) - q_t(x)| &\to 0 \text{ für } n \to \infty \\ |K_n p_t(x) - p_t(x)| &\to 0 \text{ für } n \to \infty \end{aligned}$$

für alle x und für alle t.
Wir möchten nun ein $N \in \mathbb{N}$, so wählen, dass für alle $n \geq N$, für alle $x \in [a,b]$ und für *alle* $t \in [a,b]$ gilt

$$\begin{aligned} |K_n q_t(x) - q_t(x)| &\leq \frac{\varepsilon}{3} \\ |K_n p_t(x) - p_t(x)| &\leq \frac{\varepsilon}{3}. \end{aligned} \tag{2.3}$$

Dazu ist gleichmäßige Konvergenz von $K_n q_t(x) - q_t(x)$ in x und in t nötig. Diese zeigt man für q_t wie folgt:

$$\begin{aligned} q_t(x) &= f(t) + \frac{\varepsilon}{3} + 2\|f\|_\infty \frac{(x-t)^2}{\delta^2} \\ &= f(t) + \frac{\varepsilon}{3} + \frac{2\|f\|_\infty}{\delta^2}(x^2 - 2tx + t^2) \\ &= \mathbf{1}\left(f(t) + \frac{\varepsilon}{3} + \frac{2t^2\|f\|_\infty}{\delta^2} \right) - 4tx\frac{\|f\|_\infty}{\delta^2} + 2x^2\frac{\|f\|_\infty}{\delta^2}. \end{aligned}$$

Man beachte, dass ein Polynom vom Grad zwei in x vorliegt. Aus letzterer Überlegung ergibt sich:

$$\begin{aligned} |K_n q_t(x) - q_t(x)| = \Bigg| &(K_n\mathbf{1} - \mathbf{1})\left[f(t) + \frac{\varepsilon}{3} + \frac{2t^2\|f\|_\infty}{\delta^2} \right] \\ &+ (K_n x - x)\left[\frac{-4t\|f\|_\infty}{\delta^2} \right] + (K_n x^2 - x^2)\left[\frac{2\|f\|_\infty}{\delta^2} \right] \Bigg| \\ \leq \|K_n\mathbf{1} - \mathbf{1}\|_\infty &\left(\|f\|_\infty + \frac{\varepsilon}{3} + \frac{2c^2\|f\|_\infty}{\delta^2} \right) \\ &+ \|K_n x - x\|_\infty \frac{4c\|f\|_\infty}{\delta^2} + \|K_n x^2 - x^2\|_\infty \frac{2\|f\|_\infty}{\delta^2} \end{aligned}$$

mit $c := \max\{|a|, |b|\}$. Dieser Ausdruck hängt weder von x noch von t ab und strebt gleichmäßig gegen Null. Für p_t erhält man analog einen ähnlichen Ausdruck.
Damit finden wir also $N \in \mathbb{N}$, so dass (2.3) gilt und erhalten daraus:

$$p_t(x) - \frac{\varepsilon}{3} \leq K_n f(x) \leq q_t(x) + \frac{\varepsilon}{3}. \tag{2.4}$$

Es folgt für alle x, t und $n > N$:

$$\begin{aligned} p_t(x) - q_t(x) - \frac{\varepsilon}{3} &\leq f(x) - q_t(x) - \frac{\varepsilon}{3}, \text{ denn } p_t(x) \leq f(x) \text{ nach (2.2)} \\ &\leq f(x) - K_n f(x), \text{ weil } K_n f(x) \leq q_t(x) + \tfrac{\varepsilon}{3} \text{ nach (2.4)} \\ &\leq f(x) - p_t(x) + \frac{\varepsilon}{3}, \text{ durch } p_t(x) - \tfrac{\varepsilon}{3} \leq K_n f(x) \text{ aus (2.4)} \\ &\leq q_t(x) - p_t(x) + \frac{\varepsilon}{3}, \text{ da } f(x) \leq q_t(x) \text{ nach (2.2)}. \end{aligned}$$

Insbesondere gilt das auch für $t = x$. Wegen

$$\begin{aligned} p_x(x) - q_x(x) &= f(x) - \tfrac{\varepsilon}{3} - 2\|f\|_\infty \cdot 0 - f(x) - \tfrac{\varepsilon}{3} - 2\|f\|_\infty \cdot 0 \\ &= -\tfrac{2}{3}\varepsilon \end{aligned}$$

gilt also

$$-\tfrac{2}{3}\varepsilon - \tfrac{\varepsilon}{3} \leq f(x) - K_n f(x) \leq \tfrac{2}{3}\varepsilon + \tfrac{\varepsilon}{3}$$

oder

$$|f(x) - K_n f(x)| \leq \varepsilon$$

für alle $n \geq N$ und $x \in [a, b]$. QED

Jetzt kann man zeigen, dass jede stetige Funktion beliebig gut durch Polynome approximiert werden kann, indem man eine Folge von Korovkin-Operatoren

$$K_n : C[a, b] \to \Pi_n$$

angibt, die jede stetige Funktion auf ein Polynom abbilden. Das wird durch die Bernstein-Operatoren erfüllt.

Notation 2.3

$$B_n : C[0, 1] \to \Pi_n(\mathbb{R}),$$

definiert durch

$$B_n f(x) := \sum_{j=0}^{n} \binom{n}{j} f\left(\frac{j}{n}\right) x^j (1 - x)^{n-j}, \quad x \in [0, 1]$$

nennt man ***Bernstein-Operatoren****.*

Satz 2.4 *Die Bernsteinoperatoren bilden eine Korovkin-Folge auf* $C[0, 1]$.

Beweis:

(a) Die B_n sind linear und monoton, da $x \geq 0$ und $1 - x \geq 0$ für alle $x \in [0, 1]$.

(b) Zu zeigen bleibt noch: $B_n f - f \to 0$ für $n \to \infty$ für $f \in \{\mathbf{1}, x, x^2\}$.
Wir betrachten zunächst den Fall $f(x) = \mathbf{1}$:

$$B_n \mathbf{1}(x) = \sum_{j=0}^{n} \binom{n}{j} x^j (1 - x)^{n-j} = 1 = \mathbf{1}(x),$$

nach dem Binomischen Lehrsatz, also ist $B_n \mathbf{1} = \mathbf{1}$.

Sei nun $f(x) = x$:

$$\begin{aligned} B_n x &= \sum_{j=1}^{n} \binom{n}{j} \frac{j}{n} x^j (1-x)^{n-j} \\ &= \sum_{j=0}^{n-1} \binom{n}{j+1} \frac{j+1}{n} x^{j+1} (1-x)^{n-j-1} \\ &= x \underbrace{\sum_{j=0}^{n-1} \binom{n-1}{j} x^j (1-x)^{(n-1)-j}}_{=1}, \text{ denn } \binom{x}{y}\frac{y}{x} = \binom{x-1}{y-1} \\ &= x = f(x). \end{aligned}$$

Wir betrachten abschließend $f(x) = x^2$. Nach etwas Rechnen erhält man

$$B_n f(x) = \frac{n-1}{n} x^2 + \frac{x}{n},$$

und somit

$$\begin{aligned} |f(x) - B_n f(x)| &= \left| x^2 - \frac{n-1}{n} x^2 - \frac{x}{n} \right| = \left| \frac{1}{n} x^2 - \frac{x}{n} \right| \\ &\leq \left| \frac{x^2}{n} \right| + \left| \frac{x}{n} \right| \leq \frac{2}{n} \to 0, \end{aligned}$$

also $\|f - B_n f\|_\infty \to 0$ für $n \to \infty$. QED

Damit folgt der Satz von Weierstraß:

Satz 2.5 (Weierstraß) *Zu jedem $f \in C[a,b]$ und jedem $\varepsilon > 0$ gibt es ein Polynom p so, dass $\|f - p\|_\infty < \varepsilon$.*

Beweis: Für $[a,b] = [0,1]$ folgt die Aussage aus Satz 2.2 und Satz 2.4. Im allgemeinen Fall sei $f \in C[a,b]$. Wir definieren

$$g(s) := f((b-a)s + a) \in C[0,1].$$

Zu g existiert ein Polynom q, so dass $\|g - q\|_\infty < \varepsilon$. Sei weiterhin $p(t) := q(\frac{t-a}{b-a})$, $t \in [a,b]$. Dann ist p ein Polynom und weil $t = (b-a)s + a$ äquivalent ist zu $\frac{t-a}{b-a} = s$ folgt

$$f(t) - p(t) = g\left(\frac{t-a}{b-a}\right) - q\left(\frac{t-a}{b-a}\right)$$

und daraus

$$\|f - p\|_\infty = \|g - q\|_\infty < \varepsilon,$$

also ist p das gesuchte Polynom für f. QED

Bemerkung: Für $f \in C[a,b]$ definiert man die Bernstein-Operatoren vermöge

$$\underline{B}_n f(x) = \sum_{j=0}^{n} \binom{n}{j} f\left(a + (b-a)\frac{j}{n}\right) \underbrace{\left(\frac{x-a}{b-a}\right)^j}_{=:y} \underbrace{\left(\frac{b-x}{b-a}\right)^{n-j}}_{=1-y}$$

$$= \frac{1}{(b-a)^n} \sum_{j=0}^{n} \binom{n}{j} f\left(a + (b-a)\frac{j}{n}\right)(x-a)^j(b-x)^j$$

indem man

$$[a,b] \to [0,1] \quad \text{via} \quad x \to \frac{x-a}{b-a}$$

abbildet.

Übungsaufgabe: Wandeln Sie $B_n f$ so ab, dass sie eine Korovkin-Folge auf $C[a,b]$ erhalten. (Das ist ein alternativer Beweis zum Satz 2.5).

In Satz 2.5 haben wir den Abstand zwischen der Funktion f und ihrer Approximation durch

$$\|f - p\|_\infty := \max_{x \in [a,b]} |f(x) - g(x)|$$

gemessen. Statt der Norm $\|\cdot\|_\infty$ verwenden wir im folgenden Satz die $L_p[a,b]$-Normen, die durch

$$\|f\|_p := \sqrt[p]{\int_a^b |f(x)|^p dx}$$

definiert sind.

Satz 2.6 *Zu jedem $f \in C[a,b]$ und jedem $\varepsilon > 0$ gibt es ein Polynom q so, dass $\|f - q\|_p < \varepsilon$.*

Beweis: Sei $\varepsilon > 0$. Nach Satz 2.5 gibt es ein Polynom q, so dass $\|f - q\|_\infty < \varepsilon' := \frac{\varepsilon}{(b-a)}$. Dann gilt

$$\begin{aligned} \|f - q\|_p^p &= \int_a^b |f(x) - q(x)|^p dx \\ &\leq \|f - q\|_\infty^p \int_a^b 1 dx \\ &= \|f - q\|_\infty^p (b-a) < (\varepsilon')^p (b-a) = \varepsilon^p, \end{aligned}$$

also $\|f - q\|_p \leq \varepsilon$. QED

Von Weierstraß stammt auch das folgende Approximationsresultat für trigonometrische Polynome, das wir ohne Beweis angeben:

Satz 2.7 *Zu jedem $f \in C(\mathbb{R})$ mit Periode 2π und jedem $\varepsilon > 0$ existiert ein trigonometrisches Polynom T so, dass $\|f - T\|_\infty < \varepsilon$ und $\|f - T\|_p < \varepsilon$ für alle $L_p[0, 2\pi]$-Normen.*

2.2 Existenzsätze

Wir verallgemeinern nun den Begriff der Approximation.

Definition 2.8 *Sei V ein normierter Vektorraum und $M \subseteq V$ eine Teilmenge von V. Sei $f \in V$. Dann heißt $u^* \in M$* ***beste Approximation*** *an f, falls*

$$\|f - u^*\| \leq \|f - u\| \text{ für alle } u \in M.$$

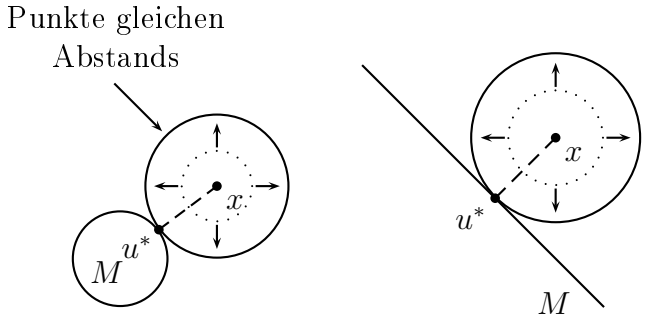
Man nennt $d(f, M) := \inf_{u \in M} \|f - u\|$ den (Minimal-)Abstand von f zu M.

Beispiele:

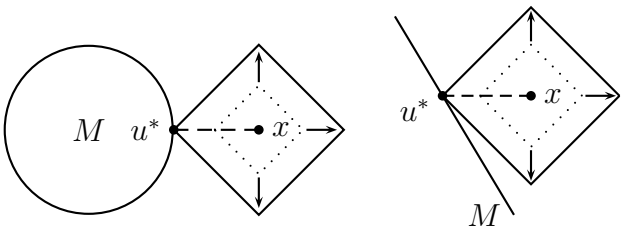
1. $V = C[a, b]$, $M = \Pi_4[a, b]$: Approximation einer stetigen Funktion $f \in V$ durch ein Polynom bis Grad 4.

2. $V = \mathbb{R}^n$, $M \subseteq \mathbb{R}^n$: Approximation eines Punktes durch einen (anderen) Punkt aus M. Hierbei ist $d(x, M)$ der Abstand des Punktes $x \in V$ von der Menge M.

   Für $\|\cdot\| = \|\cdot\|_2$ ist u^* die orthogonale Projektion von x auf M.



   Für $\|\cdot\| = \|\cdot\|_1$ ist u^* wie in der Abbildung.



3. In der linearen Ausgleichsrechnung (Numerik I, Kapitel 4.2) sind $A \in \mathbb{R}^{m,n}$ mit $m > n$ und $b \in \mathbb{R}^m$ gegeben. Gesucht ist ein $x \in \mathbb{R}^n$, sodass $\|Ax - b\|_2$ möglichst klein ist. Wir formulieren das Problem zu einer Approximationsaufgabe um: Seien $V = \mathbb{R}^m$, $M = \{Ax : x \in \mathbb{R}^n\}$ und $b \in V$ gegeben. Finde $u^* \in M$, sodass $\|b - u^*\|$ möglichst klein ist.

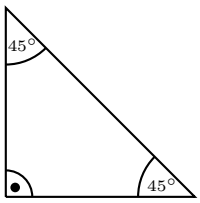
Definition 2.9 *Sei $M \subseteq V$, V normierter Vektorraum. M heißt* ***Existenzmenge****, falls es zu jedem $f \in V$ eine beste Approximation auf f gibt. M heißt Tschebyscheff-Menge, falls es zu jedem $f \in V$ genau eine beste Approximation gibt. M heißt* ***dicht in*** *V, falls $d(f, M) = 0$ für alle $f \in V$.*

Beispiele:

- $\Pi_\infty \subseteq C[a, b]$ ist keine Existenzmenge, aber
- der Satz von Weierstraß (Satz 2.5) besagt, dass für $M = \Pi_\infty$ – den Raum aller Polynome – gilt

$$\begin{aligned} d(f, M) &= \inf_{p \in M} \|f - q\|_\infty \\ &= 0 \text{ für alle } f \in C[a, b], \end{aligned}$$

  also liegt Π_∞ dicht in $C[a, b]$.
- Dagegen ist $\Pi_4[a, b]$ nicht dicht in $C[a, b]$.
- Jede konvexe, kompakte Menge $M \subseteq \mathbb{R}^n$ ist Tschebyscheff-Menge bzgl. $\|\cdot\|_2$.
- Bzgl. $\|\cdot\|_1$ ist z.B. ein gleichschenkliges Dreieck mit achsenparallelen Kanten keine Existenzmenge.



- $\mathbb{Q}$ liegt dicht in $\mathbb{R}$, ist aber keine Existenzmenge.

Lemma 2.10 *Sei M eine kompakte Teilmenge eines normierten Raums V. Dann ist M Existenzmenge.*

Beweis: $\|\cdot\|$ ist stetig, genauer: Sei $f \in V$. Betrachte

$$\begin{aligned} \varphi : V &\to \mathbb{R} \\ v &\mapsto \|f - v\|. \end{aligned}$$

Dann gibt es für jedes $\varepsilon > 0$ ein $\delta := \varepsilon$, sodass

$$|\varphi(v) - \varphi(u)| = |\|f - v\| - \|f - u\|| \leq \|u - v\| \leq \varepsilon$$

für alle u, v mit $\|u - v\| \leq \delta$. Also ist φ eine stetige Funktion auf einer kompakten Menge und nimmt entsprechend ihr Minimum an. QED

Lemma 2.11 *Es gilt:* $|d(f,M) - d(g,M)| \leq \|f-g\|$ *für alle* $f,g \in V$, V *normierter Vektorraum und* $M \subseteq V$, *d.h. der Minimalabstand hängt stetig von dem zu approximierenden Element ab.*

Beweis: Seien $f,g \in V, \varepsilon > 0$. Wähle $u(\varepsilon) \in M$ so, dass $\|g-u(\varepsilon)\| \leq d(g,M)+\varepsilon$. Dann gilt:

$$\begin{aligned} d(f,M) &\leq \|f-u(\varepsilon)\| \leq \|f-g\| + \|g+u(\varepsilon)\| \\ &\leq \|f-g\| + d(g,M) + \varepsilon \end{aligned}$$

also $d(f,M) - d(g,M) \leq \|f-g\| + \varepsilon$. Analog erhält man, wenn man f und g vertauscht:

$$d(g,M) - d(f,M) \leq \|f-g\| + \varepsilon.$$

Zusammen ergibt sich:

$$\begin{aligned} |d(g,M) - d(f,M)| &\leq \|f-g\| + \varepsilon \text{ für alle } \varepsilon > 0 \\ \text{also } |d(g,M) - d(f,M)| &\leq \|f-g\|. \end{aligned}$$

QED

Wir betrachten nun Mengen $M \subseteq V$ mit weiteren Eigenschaften:

1. M konvexe Teilmenge von V.
2. M Unterraum von V.

Wir erinnern uns:

$$M \text{ konvex} \Leftrightarrow \forall x,y \in M, \forall \lambda \in (0,1) : \lambda x + (1-\lambda)y \in M.$$

Es gilt:

- Jeder Unterraum ist konvex.
- $\varnothing$ ist konvex.
- M_1, M_2 konvex $\Rightarrow M_1 \cap M_2$ konvex.

Im Folgenden bezeichnet $\mathcal{U}^*_{(f)}$ die Menge der besten Approximationen an $f \in V$ aus M.

Satz 2.12 *Sei* V *normierter Vektorraum und* $M \subseteq V$ *konvex. Zu* $f \in V$ *existiere eine beste Approximation* $u^* \in M$, *d.h.* $\mathcal{U}^*_{(f)} \neq \varnothing$. *Dann gilt: Entweder* $\mathcal{U}^*_{(f)} = \{u^*\}$ *oder* $|\mathcal{U}^*_{(f)}| = \infty$ *und* $\mathcal{U}^*_{(f)}$ *ist konvex.*

Beweis: Seien u_1, u_2 beides beste Approximationen an f, also

$$d(f, u_1) = \|f - u_1\| = \|f - u_2\| = d(f, u_2).$$

Betrachte $u := tu_1 + (1-t)u_2 \in M$ für beliebiges $t \in [0, 1]$. Dann gilt

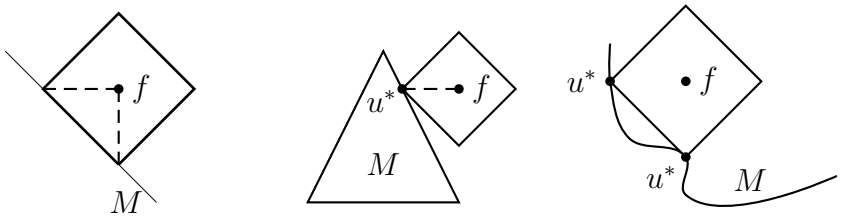
$$\begin{aligned} \|f - u\| &= \|(f - u_1)t + (f - u_2)(1 - t)\| \\ &\leq |t| \underbrace{\|f - u_1\|}_{=d(f,u_1)} + |1 - t| \underbrace{\|f - u_2\|}_{=d(f,u_2)} \\ &= d(f, u_1), \end{aligned}$$

also ist u auch beste Approximation an f und die Menge $\mathcal{U}^*_{(f)}$ ist konvex. Weil

$$|\{tu_1 + (1-t)u_2 : t \in [0, 1]\}| = \infty,$$

hat die Menge aller besten Approximationen – wie jede konvexe Menge mit mehr als einem Element – unendlich viele Elemente. QED

Beispiele: $V = \mathbb{R}^2$, $\|\cdot\| = \|\cdot\|_1$.



konvexe Menge beschreibt beste Approximation

beste Approximation eindeutig

Genau zwei beste Approximationen

Speziell für lineare Unterräume M gilt die folgende Aussage:

Satz 2.13 *Sei U ein endlich-dimensionaler Unterraum eines normierten Vektorraums V. Dann ist U eine Existenzmenge. Weiterhin ist für alle $f \in V$ die Menge der besten Approximationen $\mathcal{U}^*_{(f)}$ konvex und es gilt entweder $|\mathcal{U}^*_{(f)}| = 1$ oder $|\mathcal{U}^*_{(f)}| = \infty$.*

Beweis: Weil U ein Unterraum ist, gilt $0 \in U$. Sei

$$U_0 = \{u \in U : \|f - u\| \leq \|f - 0\|\}$$

die Menge aller Elemente aus U, die f mindestens genauso gut approximieren wie 0. Es ist also $\mathcal{U}^*_{(f)} \subset U_0$.

U_0 ist abgeschlossen (weil $\|\cdot\|$ stetig ist) und beschränkt (weil $\|u\| \leq \|u-f\| + \|f\| \leq 2\|f\|$ für alle $u \in U_0$). Zusammen folgt, dass U_0 eine kompakte Menge ist. Nach Lemma 2.10 existiert also eine beste Approximation an f aus U_0. Diese ist beste Approximation an f aus U.
Weil jeder Unterraum insbesondere konvex ist, folgt der zweite Teil aus Satz 2.12.
QED

Bemerkung: Die Voraussetzung "endlich-dimensional" ist nötig! Betrachte dazu $V = C[a,b]$ mit $\|\cdot\| = \|\cdot\|_\infty$ und $U = \Pi_\infty[a,b]$. Sei $f \in C[a,b] \setminus \Pi_\infty[a,b]$. Dann gilt zwar $d(f,U) = 0$, aber weil f kein Polynom ist, wird dieses Infimum nie angenommen.

Bemerkung: Man kann zeigen, dass die beste Approximation in Euklidischen Räumen – sofern sie existiert – immer eindeutig ist.

2.3 Tschebyscheff-Approximation in $C[a,b]$

Wir untersuchen nun wieder die Approximation einer stetigen Funktion $f \in C[a,b]$ durch $u^* \in U \subseteq C[a,b]$. Dabei betrachten wir als Abstand

$$\|u-f\|_\infty = \max_{x\in[a,b]} |u(x) - f(x)|.$$

Aus Satz 2.13 wissen wir, dass jeder endlich-dimensionale Unterraum $U \subseteq C[a,b]$ eine Existenzmenge ist. Um die Eindeutigkeit zu behandeln, betrachten wir unisolvente Räume (siehe auch Numerik I).

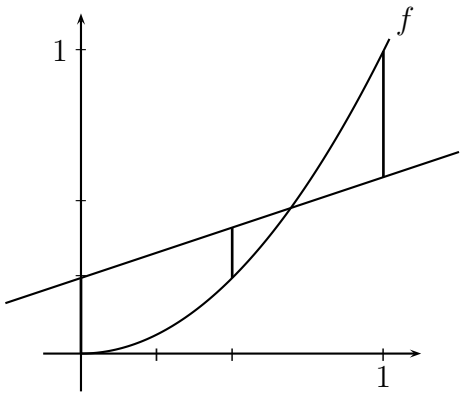
Definition 2.14 *Sei $U \subseteq C[a,b]$ ein Unterraum von $C[a,b]$ mit $\dim(U) = n$. Dann heißt U **Haar'scher Raum** der Dimension n, falls jedes $u \in U \setminus \{0\}$ höchstens $n-1$ Nullstellen in $[a,b]$ hat.*

Bemerkung: Ein Haar'scher Raum ist unisolvent bezüglich jeder Menge $X \subset [a,b]$ mit $|X| \geq n$.

Beispiel: Es ist $\Pi_n \subseteq C[a,b]$ ein Haar'scher Raum der Dimension $n+1$, denn jedes nicht-verschwindende Polynom vom Grad maximal n hat höchstens n Nullstellen in $[a,b]$.

Die Approximation einer Funktion bezüglich der $\|\cdot\|_\infty$-Norm soll zunächst an einem ausführlichen Beispiel demonstriert werden.

Beispiel: Betrachte $I = [0,1]$ und $f(x) = x^2 \in C[0,1]$.
Wir interessieren uns für die beste lineare Approximation, suchen also eine Funktion $u^*(x) = \alpha + \beta x$ mit minimalem Abstand $\|f-u^*\|_\infty$ zu f.



keine optimale Lösung

Für den Fehler gilt:

$$\begin{aligned}\|f - u^*\|_\infty &= \max_{x\in[0,1]} |f(x) - u^*(x)| \\ &= \max_{x\in[0,1]} |x^2 - \beta x - \alpha|.\end{aligned}$$

Es gilt

- $x^2 - \beta x - \alpha$ wird als konvexe Funktion am Rand maximal, also für $x = 0$ oder für $x = 1$ mit Maximalwerten $|\alpha|$ oder $|\beta + \alpha - 1|$.
- $-x^2 + \beta x + \alpha$ wird als konkave und differenzierbare Funktion am Rand maximal, oder falls ihr Gradient gleich Null ist, also falls

  $$-2x + \beta = 0 \Leftrightarrow x = \tfrac{1}{2}\beta.$$

  Der Maximalwert beträgt dann $|\frac{1}{4}\beta^2 - \frac{1}{2}\beta^2 - \alpha| = |\alpha + \frac{1}{4}\beta^2|$.

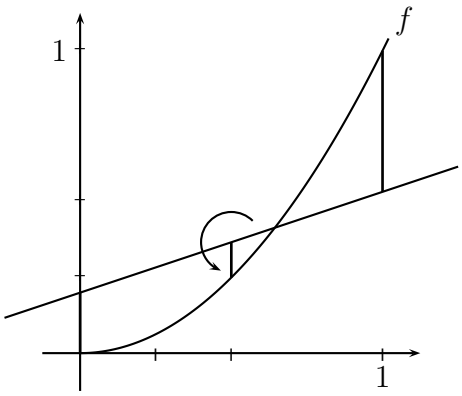
Wir erhalten also

$$\|f - u^*\|_\infty = \max\{|\alpha|, |\beta + \alpha - 1|, |\alpha + \tfrac{1}{4}\beta^2|\}.$$

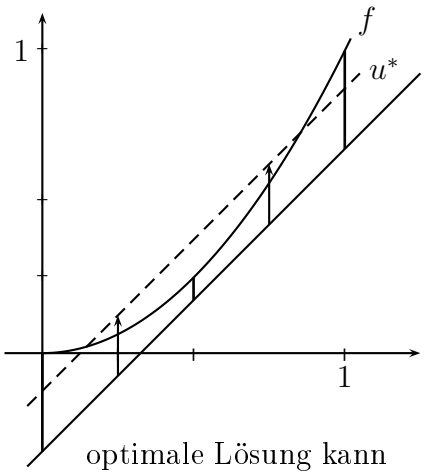
Für welche α, β wird dieser Ausdruck minimal?

- Dazu müssen alle drei Terme den gleichen Wert annehmen. Man kann sich das durch eine Fallunterscheidung leicht klarmachen: Sind die Werte nicht gleich, gilt also zum Beispiel $|\alpha| > |\beta + \alpha + 1|$ und $|\alpha| > |\alpha + \frac{1}{4}\beta^2|$, so kann die Lösung verbessert werden, indem man α (auf Kosten von β) etwas reduziert. (Analog in den anderen Fällen.)
- Außerdem müssen die Vorzeichen der drei Terme alternieren; sonst könnte man die Gerade ebenfalls verbessern (Skizze).

Durch eine Skizze lassen sich beide Aussagen veranschaulichen: Ist eine der drei Strecken länger als die beiden anderen, so kann man sie durch Verschieben und Drehen von u auf Kosten der anderen verkürzen und so u verbessern.



keine optimale Lösung



optimale Lösung kann durch Verschieben erreicht werden

In unserem Beispiel erhält man für den Fall

$$f(0) > u^*(0), \quad f(\tfrac{1}{2}\beta) < u^*(\tfrac{1}{2}\beta) \quad \text{und} \quad f(1) > u^*(1)$$

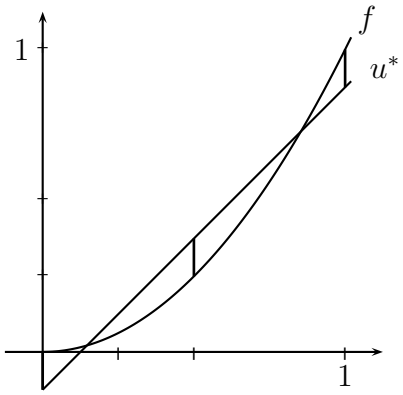
die Gleichungen

$$-\alpha = \alpha + \tfrac{1}{4}\beta^2 = 1 - \alpha - \beta, \text{ d.h. } 2\alpha + \tfrac{1}{4}\beta^2 = 0 \quad \text{und} \quad 1 - \beta = 0.$$

woraus folgt, dass $\beta = 1$, $\alpha = -\frac{1}{8}$ und $\|f - u\|_\infty = \frac{1}{8}$. Da das Vorzeichen alternieren muss, gibt es nur noch einen weiteren Fall: $f(0) < u^*(0), f(\frac{1}{2}\beta) > u^*(\frac{1}{2}\beta)$ und $f(1) < u^*(1)$. Dieser liefert keinen besseren Wert für $\|f - u^*\|_\infty$, also ist

$$u^*(x) = x - \tfrac{1}{8}$$

die beste Approximation.



beste Approximation

Das Beispiel motiviert die folgende Definition:

Definition 2.15 *Sei U ein Haar'scher Raum der Dimension n über $[a,b]$. Eine Menge X von $n+1$ Punkten $a \le x_1 < x_2 < \cdots < x_{n+1} \le b$ heißt* ***Alternante*** *für $f \in C[a,b]$ und $u \in U$, falls*

$$\operatorname{sign}\Big(f(x_j) - u(x_j)\Big) = \sigma(-1)^j, \qquad 1 \le j \le n+1$$

gilt mit einer Konstanten $\sigma \in \{-1,1\}$.

Eine Menge X ist also Alternante für f und u, wenn $f-u$ in den x_j alternierend das Vorzeichen wechselt.

Satz 2.16 *Sei U ein n-dimensionaler Haar'scher Raum über $[a,b]$. Gibt es zu $f \in C[a,b]$ und $u^* \in U$ eine Alternante X mit*

$$|f(x_j) - u^*(x_j)| = \|f - u^*\|_\infty, \qquad 1 \le j \le n+1,$$

so ist u^* eine beste Approximation an f aus U.

Beweis: Sei $X = \{x_1, \ldots, x_n, x_{n+1}\}$ Alternante mit

$$\operatorname{sign}(f(x_j) - u^*(x_j)) = \sigma(-1)^j \quad \text{für alle} \quad 1 \le j \le n+1$$

für ein festes $\sigma \in \{-1,1\}$. Sei $u \in U$. Wir wollen zeigen, dass

$$\|f - u^*\|_\infty \le \|f - u\|_\infty.$$

Dazu rechnen wir

$$\begin{aligned} \|f-u^*\|_\infty &= |f(x_j) - u^*(x_j)| \quad \text{für } j = 1,\ldots,n+1 \\ &= (f(x_j) - u^*(x_j))\sigma(-1)^j \quad \text{für } j = 1,\ldots,n+1 \\ &= (f(x_j) - u(x_j))\sigma(-1)^j + (u(x_j) - u^*(x_j))\sigma(-1)^j \\ &\qquad\qquad \text{für } j = 1,\ldots,n+1 \end{aligned} \tag{2.5}$$

Um diesen Ausdruck weiter abzuschätzen, zeigen wir zunächst, dass es ein $j_0 \in \{1,\ldots,n-1\}$ so gibt, dass

$$(u(x_{j_0}) - u^*(x_{j_0}))(-1)^{j_0}\sigma \le 0. \tag{2.6}$$

Dazu nehmen wir an, dass (2.6) für kein j_0 gültig ist. Das heißt,

$$(u(x_j) - u^*(x_j))(-1)^j\sigma > 0 \quad \text{für alle} \quad j \in \{1,2,\ldots,n,n+1\},$$

also würde $u-u^*$ in jedem der n Intervalle (x_j, x_{j+1}), $j = 1, \ldots, n$ das Vorzeichen wechseln. Nach dem Zwischenwertsatz hätte die stetige Funktion $u - u^*$ also mindestens n Nullstellen. Aber $u - u^* \in U$ und U haben wir als Haar'schen Raum der Dimension n vorausgesetzt. Somit gilt:

$$u - u^* \equiv 0 \text{ oder } u - u^* \text{ hat höchstens } n - 1 \text{ Nullstellen.}$$

Daraus ergibt sich also $u \equiv u^*$; das aber ist ein Widerspruch zu $u(x_j) \neq u^*(x_j)$ an den Punkten $x_1, \ldots, x_{n+1}$.

Wir verwenden die eben gezeigte Aussage, um $\|f-u^*\|$ in (2.5) weiter abzuschätzen, indem wir für j den Index j_0 wählen, der (2.6) erfüllt. Wir erhalten:

$$\begin{aligned}\|f - u^*\|_\infty &= (f(x_{j_0}) - u(x_{j_0}))\sigma(-1)^{j_0} + \underbrace{(u(x_{j_0}) - u^*(x_{j_0}))\sigma(-1)^{j_0}}_{\leq 0 \text{ nach (2.6)}} \\ &\leq |f(x_{j_0}) - u(x_{j_0})| \leq \|f - u\|_\infty.\end{aligned}$$

QED

Um eine beste Approximation zu finden, macht es also Sinn, f zunächst auf einer diskreten Menge $X = (x_1, \ldots, x_{n+1})$ zu approximieren. Wir führen die folgenden Bezeichnungen ein.

Notation: *Einen Vektor $X = (x_1, \ldots, x_{n+1})^T \in \mathbb{R}^{n+1}$ mit $a \leq x_1 < x_2 < \cdots < x_n < x_{n+1} \leq b$ nennen wir **Referenz**. Wir definieren*

$$\|(u - f)|_X\|_\infty := \max_{i=1,\ldots,n+1} |u(x_i) - f(x_i)|.$$

Gilt für $u^* \in U$ dass

$$\|(u^* - f)|_X\|_\infty \leq \|(u - f)|_X\|_\infty$$

*für alle $u \in U$, so nennt man u^* **beste Approximation** an f aus U auf der Referenz X, oder **diskrete Approximation** auf X oder **Tschebyscheff-Approximation** an f aus U auf X.*

Korollar 2.17 *Sei U ein Haar'scher Raum der Dimension n über $[a, b]$. Gibt es zu $f \in C[a, b]$ und $u^* \in U$ eine Alternante X mit*

$$|f(x_j) - u^*(x_j)| = const \quad \text{für alle } 1 \leq j \leq n + 1$$

(das heißt dann, dass $|f(x_j) - u^*(x_j)| = \|(f - u^*)|_X\|_\infty$ für alle $1 \leq j \leq n + 1$), so ist u^* Tschebyscheff-Approximante auf X an f aus U.

Beweis: Der Beweis verläuft genau analog zu dem Beweis von Satz 2.16, nur betrachtet man statt $\|f - u^*\|_\infty$ den Ausdruck $\|(f - u^*)|_X\|_\infty$ beziehungsweise statt $\|f - u\|_\infty$ den Ausdruck $\|(f - u)|_X\|_\infty$. QED

Das ergibt folgende Idee, um eine beste Approximation iterativ anzunähern.

1. Starte mit Referenz X und bestimme $u^* \in U$ so, dass

   - X ist Alternante für f und u^*
   - $|f(x_i) - u^*(x_i)| = \text{const}$.

   Dann ist u^* beste diskrete Approximation an f aus U auf X (nach Korollar 2.17).

2. Gilt zusätzlich, dass $\|f - u^*\|_\infty = \text{const}(= \|(f - u^*)|_X\|_\infty)$, so ist u^* beste Approximante an f aus U auf ganz $[a, b]$ (nach Satz 2.16).

3. Sonst verändere die Referenz X und gehe zu 1.

Wir werden im Folgenden besprechen,

- wie man in Schritt 1 die diskrete Tschebyscheff-Approximante berechnen kann, und
- wie man in Schritt 3 die Referenz X geeignet modifiziert, so dass das Verfahren konvergiert.

Wir beginnen mit der Berechnung der diskreten Tschebyscheff-Approximante.

Sei $u_1, \ldots, u_n$ eine Basis von U. Sei X eine Referenz und bezeichne

$$\rho_X = d_X(f, U) = \inf_{u \in U} \|(f - u)|_X\|_\infty$$

den Minimalabstand von f und U bzgl. der Referenz $X = (x_1, \ldots, x_{n+1})^T$. Wir suchen

$$u^* = \sum_{j=1}^{n} \alpha_j u_j,$$

genauer also die Koeffizienten $\alpha_1, \ldots, \alpha_n$. Sei σ_X das Vorzeichen von $f(x_1) - u^*(x_1)$. Dann müssen die folgenden $n + 1$ Bedingungen erfüllt sein:

$$f(x_i) - u^*(x_i) = \rho_X \sigma_X (-1)^{i-1}, \qquad \forall i = 1, \ldots, n + 1.$$

Das schreiben wir um zu:

$$f(x_i) = \sum_{j=1}^{n} \alpha_j u_j(x_i) + \underbrace{(-1)^{i-1}}_{\substack{:= u_{n+1} \\ \text{bekannt}}} \underbrace{\sigma_X \rho_X}_{\substack{=: \alpha_{n+1} \\ \text{Variable}}} \qquad 1 \le i \le n + 1.$$

Als Gleichungssystem erhält man $n+1$ Gleichungen in $n+1$ Variablen, wobei wir zur Vereinfachung der Schreibweise

$$u_{n+1}(x_i) := (-1)^{i-1}$$

setzen. In Matrixform ergibt sich:

$$\underbrace{\begin{pmatrix} u_1(x_1) & \cdots & u_n(x_1) & u_{n+1}(x_1) \\ u_1(x_2) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ u_1(x_{n+1}) & \cdots & u_n(x_{n+1}) & u_{n+1}(x_{n+1}) \end{pmatrix}}_{=:A} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{n+1} \end{pmatrix} = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{n+1}) \end{pmatrix}. \tag{2.7}$$

Ist dieses Gleichungssystem lösbar? Wir benutzen den Laplaceschen Entwicklungssatz für die letzte Spalte. Sei dazu

$$D_i = \begin{pmatrix} u_1(x_1) & \cdots & u_n(x_1) \\ \vdots & & \vdots \\ u_1(x_{i-1}) & \cdots & u_n(x_{i-1}) \\ u_1(x_{i+1}) & \cdots & u_n(x_{i+1}) \\ \vdots & & \vdots \\ u_1(x_{n+1}) & \cdots & u_n(x_{n+1}) \end{pmatrix} \in \mathbb{R}^{n,n}.$$

Dann gilt:

$$\begin{aligned} \det(A) &= \sum_{i=1}^{n+1} (-1)^i \underbrace{u_{n+1}(x_i)}_{=(-1)^{i-1}} \det(D_i) \\ &= \sum_{i=1}^{n+1} \det(D_i). \end{aligned}$$

D_i ist die zu den Punkten $x_1, \ldots, x_{i-1}, x_{x+1}, \ldots, x_n$ gehörende Interpolationsmatrix, daher ist $\det D_i \neq 0$ für alle i. Man kann sogar zeigen, dass alle $\det D_i$ das gleiche Vorzeichen haben, also gilt $\det A \neq 0$. Außerdem gilt der folgende Satz:

Satz 2.18 *Sei U ein Haar'scher Raum der Dimension n über $[a,b]$ und sei $X = a \leq x_1 < x_2 < \cdots \leq x_{n+1} = b$ eine Referenz. Dann gibt es zu jedem $f \in C[a,b]$ genau eine Lösung der Tschebyscheff-Approximation. Man kann sie durch Lösen des linearen Gleichungssystems* (2.7) *berechnen.*

Beweisskizze:

- Weil (2.7) eindeutig lösbar ist, folgt die Existenz der Tschebyscheff-Approximation.

- Um die Eindeutigkeit nachzuweisen, muss man zeigen, dass jede Tschebyscheff-Approximante auch Lösung von (2.7) ist. Weil (2.7) eindeutig lösbar ist, folgt daraus die Behauptung.

Übungsaufgabe: Sei U ein Haar'scher Raum der Dimension n über $[a,b]$ und sei $X \subseteq [a,b]$ mit $a \leq x_1 \leq \cdots \leq x_n \leq b$, also $|X| = n$. Bestimmen Sie $d_X(f,U)$!

Bevor wir das Remes-Verfahren formulieren, machen wir uns die Idee, die beste Approximation durch eine beste diskrete Approximation anzunähern, an folgendem Lemma klar:

Lemma 2.19 *Sei X eine Referenz. Dann gilt*

$$d_X(f,U) \leq d(f,U).$$

Beweis: Sei $u \in U$. Dann gilt

$$\|(f-u)|_X\|_\infty = \max_{i=1,\ldots,n+1} |f(x_i) - u(x_i)| \leq \|f-u\|_\infty$$

und folglich

$$\inf_{u\in U} \|(f-u)|_X\|_\infty \leq \inf_{u\in U} \|f-u\|_\infty,$$

also $d_X(f,U) \leq d(f,U)$. QED

Wenn man also $d(f,U)$ durch $d_X(f,U)$ annähern möchte, ist die Referenz X dafür besser geeignet als die Referenz X', falls $d_X(f,U) \geq d_{X'}(f,U)$, also falls die Fehlerfunktion $f - u_X^*$ für die Tschebyscheff-Approximante u_X^* bezüglich der Referenz X möglichst *groß* ist! Diese Beobachtung wird im Remes-Verfahren wie folgt ausgenutzt:

Algorithmus 1: Remes-Verfahren

Input: $f \in C[a,b]$, $U \subseteq C[a,b]$ Haar'scher Raum der Dimension n.

Schritt 1: Wähle Startreferenz $X^{(0)} = \{x_1^{(0)}, \ldots, x_{n+1}^{(0)}\}$, $j := 0$,

Schritt 2: Bestimme die Tschebyscheff-Approximation u_j^* auf $X^{(j)}$ an f. Sei $p_j = d_{X^{(j)}}(f,U) = \|(f-u_j^*)|_{X^{(j)}}\|_\infty$.

Schritt 3: Falls $d_{X^{(j)}}(f,U) = \|f-u_j^*\|_\infty$: STOP. Lösung sei v_j^*.

Schritt 4: Bestimme die neue Referenz $X^{(j+1)}$, die den folgenden drei Bedingungen genügt:

a) $\operatorname{sign}(f-u_j^*)(x_k^{(j+1)}) = -\operatorname{sign}(f-u_j^*)(x_{k+1}^{(j+1)})$ für alle $1 \le k \le n$.

b) $|(f-u_j^*)(x_k^{(j+1)})| \ge d_{X^{(j)}}(f,U)$ für alle $1 \le k \le n+1$.

c) $\|(f-u_j^*)|_{X^{(j+1)}}\|_\infty = \|f-u_j^*\|_\infty$.

Setze $j := j+1$ und gehe zu 2.

Zunächst analysieren wir Schritt 4:

- Bedingung a) bedeutet, dass die alte Fehlerfunktion $f-u_j^*$ auch auf der neuen Referenz $X^{(j+1)}$ alternieren soll.
- Bedingung b) besagt, dass die alte Fehlerfunktion $f-u_j^*$, angewendet auf die Punkte der neuen Referenz, nicht kleiner sein darf als an den Punkten der alten Referenz.
- Zusammen mit Bedingung c) heißt das sogar, dass die alte Fehlerfunktion, angewendet auf die neuen Punkte, maximal werden soll, also den Gesamtfehler $\|f-u_j^*\|_\infty$ an einem der neuen Punkte annehmen muss.

Man versucht also, gemäß der Aussage von Lemma (2.19), die neue Referenz so zu wählen, dass ihr Fehler möglichst groß wird. Bevor wir uns mit der Konvergenz des Remes-Verfahrens beschäftigen, zeigen wir, dass es in Schritt 4 immer eine passende neue Referenz gibt.

Lemma 2.20 *Es gibt eine Referenz $X^{(j+1)}$, die den Bedingungen von Schritt 4 genügt, falls $d_{X^{(j)}} < \|f-u_j^*\|_\infty$.*

Beweis: Die Referenz $X^{(j)}$ genügt den Randbedingungen a) und b). Wir werden daher nur einen Punkt aus $X^{(j)}$ gegen einen neuen austauschen. Dazu bestimmen

wir $\tilde{x}$ mit

$$\|f - u_j^*\| = f(\tilde{x}) - u_j^*(\tilde{x})$$

als einen Punkt in $[a, b]$, an dem der Fehler maximal wird. Wegen

$$d_{X^{(j)}} < \|f - u_j^*\|_\infty$$

ist $\tilde{x} \neq x_k$ für $k = 1, \ldots, n+1$. Wir unterscheiden drei Fälle:

1. Existiert ein k, so dass $x_k^{(j)} < \tilde{x} < x_{k+1}^{(j)}$, so setze $x_k^{(j+1)} := \tilde{x}$ falls $\text{sign}(f - u_j^*)(x_k^{(j+1)}) = \text{sign}(f - u_j^*)(\tilde{x})$, sonst $x_{k+1}^{(j)} := \tilde{x}$. Man ersetzt also entweder $x_k^{(j)}$ oder $x_{k+1}^{(j)}$ durch $\tilde{x}$.

2. Falls $\tilde{x} < x_1^{(j)}$, setze $x_1^{(j+1)} := \tilde{x}$. Falls $\text{sign}(f - u_j^*)(x_1) \neq \text{sign}(f - u_j^*)(\tilde{x})$, setze außerdem $x_{k+1}^{(j+1)} := x_k^{(j)}$ für $k = 1, \ldots, n$. (Im ersten Fall wird also $x_1^{(j)}$ aus der Referenz entfernt, im zweiten Fall $x_{n+1}^{(j)}$.)

3. Falls $\tilde{x} > x_{n+1}^{(j)}$ analog zu Fall 2.

QED

In der Praxis ersetzt man meistens mehr als einen Punkt aus $X^{(j)}$.

Jetzt können wir folgenden Satz über das Remes-Verfahren formulieren:

Satz 2.21 *Sei $U \subseteq C[a, b]$ ein Haar-Raum der Dimension n und sei $f \in C[a, b] \setminus U$. Dann existiert genau eine beste Approximation $u^* \in U$ auf f aus U auf $[a, b]$. Ferner bricht das Remes-Verfahren entweder nach endlich vielen Schritten mit u^* ab, oder es liefert Folgen $\{X^{(j)}\}$, $\{u_j^*\}$ und $\{p_j\}$ mit folgenden Eigenschaften:*

- *$\{p_j\}$ konvergiert mindestens linear gegen $\|f - u^*\|$. Genauer existiert eine Konstante $q \in (0, 1)$ mit*

$$\|f - u^*\| - p_{j+1} \leq q(\|f - u^*\| - p_j), \qquad j \in \mathbb{N}_0$$

- *$\{u_j^*\}$ konvergiert gleichmäßig auf I gegen die Lösung u^*.*

Beweis: Bricht das Verfahren nach endlich vielen Schritten ab, so gibt es ein $j \in \mathbb{N}_0$ mit

$$\rho_j = \|(f - u_j^*)|_{X^{(j)}}\|_\infty = \|f - u_j^*\|_\infty,$$

also ist u_j^* beste Approximation an f auf I nach Satz (2.16).
Nehmen wir also an, das Verfahren endet nicht. Dann kann man zeigen, dass die Folge (p_j) aufgrund der Bedingungen a), b) und c) streng monoton wachsend ist. Wegen Lemma 2.19 ist

$$\rho_j \leq d(f, U)$$

also ist die Folge nach oben beschränkt. Daraus folgt Konvergenz.
Der Nachweis der mindestens linearen Konvergenz wird hier nicht beschrieben.
Die Eindeutigkeit folgt folgendermaßen: Sei

$$\{X^{(j)}\} \subseteq \text{Menge aller Referenzen,}$$

dann besitzt diese eine konvergente Teilfolge, die gegen eine Referenz X^* konvergiert. Sei u^* die zugehörige Tschebyscheff-Approximante, die Lösung von $\inf_{u \in U} \|f - u\|_\infty$ ist. Sei $\tilde{u}$ eine weitere Lösung des Approximationsproblems, dann ist $\tilde{u}$ auch eine Tschebyscheff-Approximation an f aus V auf $[a.b]$. Nach Satz 2.18 ist diese eindeutig, also $u^* = \tilde{u}$. QED

Bemerkung: Die beste Approximation u^* ist eindeutig und $u_j^* \to u^*$. Aber die Folge der Referenzen $X^{(j)}$ hat nur eine konvergente Teilfolge, weil es zu u^* mehrere Alternanten geben kann, als Häufungspunkte der Referenzen auftreten können.

Als Folge des letzten Satzes erhalten wir die "Rückrichtung" zu Satz 2.16:

Satz 2.22 (Alternantensatz) *Sei U ein n-dimensionaler Haar'scher Raum über $[a, b]$. Ein Element $u^* \in U$ ist genau dann beste Approximation an $f \in C[a, b]$, wenn es eine Alternate X für f und u^* mit*

$$|f(x_j) - u^*(x_j)| = \|f - u^*\|_\infty, \qquad 1 \le j \le n+1$$

gibt. Die beste Approximation u^* ist eindeutig bestimmt, die Alternante aber nicht.

Durch das Remes-Verfahren haben wir konstruktiv gezeigt, dass jeder Haar'sche Raum eine Tschebyscheff-Menge ist. Der nächste Satz sagt, dass Haar'sche Räume die einzigen Tschebyscheff-Mengen sind.

Satz 2.23 *Sei U ein n-dimensionaler Unterraum von $C[a, b]$. Dann gilt:*

$$U \text{ ist Haar'scher Raum} \Leftrightarrow U \text{ ist Tschebyscheff-Menge.}$$

2.4 Zusammenfassung

Ziel:
- *Nähere ein Objekt f (aus einem VR V) durch ein "einfacheres Objekt" an*
- *"einfacher": Aus einer Menge $M \subseteq V$*
- *"annähern": $\|f - u^*\| \leq \|f - u\|$, $\forall u \in M \Rightarrow u^*$ ist beste Annäherung (beste Approximation)*

Beispiele

- $V = C[a,b]$, $\|\cdot\|_\infty$, M z.B. Π_n: Approximation von Funktionen
- $V = \mathbb{R}^n$, $\|\cdot\|$ bel. Norm, $M \subseteq \mathbb{R}^n$: Projektion
- $V = \mathbb{R}^n$, $M = \{Ax : x \in \mathbb{R}^n\}$, $\|\cdot\|_2$: Ausgleichsrechnung ($A \in \mathbb{R}^{m,n}$)

Existenz- und Tschebyscheff-Mengen

- M Existenzmenge, falls eine beste Approximation $u^* \in M$ existiert
- M Tschebyscheff-Menge, falls genau eine beste Approximation existiert
- M kompakt $\Rightarrow$ M Existenzmenge
- M konvex $\Rightarrow$ $\exists$ keine, genau eine oder unendl. viele beste Approx. und bilden eine konvexe Menge.

Speziell für $C[a,b]$

Satz von Weierstrass

- Satz von Weierstrass:

$$\forall \varepsilon > 0, \forall f \in C[a,b] : \exists p \in \Pi_\infty : \|f - p\|_\infty < \varepsilon$$

- Beweisskizze:
  - $\{K_n\}$ Korovkin-Folge, falls K_n linear und monoton $\forall n$ und $\lim_{n\to\infty} \|K_n f - f\|_\infty = 0$ für $f \in \{\mathbf{1}, x, x^2\}$
  - Ist $\{K_n\}$ Korovkin-Folge, dann gilt $\lim_{n\to\infty} \|K_n f - f\|_\infty = 0$, $\forall f \in C[a,b]$
  - Ziel: Finde Korovkin-Folge $K_n : C[a,b] \to \Pi_\infty[a,b]$
  - Bernstein-Operatoren! $\Rightarrow$ Beweis

- Folge: Satz von Weierstraß gilt auch für $\|\cdot\|_{L_p}$-Normen;

$$\|f\|_{L_p} = \sqrt[p]{\int_a^b |f(x)|^p dx}$$

Tschebyscheff-Approx. in Haar'schen Räumen

- ..., d.h.

$$V = C[a,b], \|f\|_\infty = \max_{x\in[a,b]} |f(x)|, M \text{ ist Haar'scher Raum}$$

- U ist Haar'scher Raum der Dim. n, falls jedes $u \in U \setminus \{0\}$ höchstens $n-1$ Nullstellen hat.
- $X = \{x_1 < x_2 < \cdots < x_{n+1}\} \subseteq [a,b]$ heißt Alternante für f und u, falls $\text{sign}(f(x_j) - u(x_j)) = \sigma(-1)^j$, $\sigma \in \{-1, 1\}$
- Kriterium: $U \subseteq V$ Haar'scher Raum der Dim. n, $f \in [a,b]$, $u^* \in U$: Gibt es eine Alternante X mit $|f(x_j) - u^*(x_j)| = \|f - u^*\|_\infty$, $j = 1, \ldots, n-1$, so ist u^* beste Approx. an f (aus V)

Diskrete Approximation

- Gegeben: $X = \{x_1, \ldots, x_{n+1}\}$. Finde $u^* : \|(f-u^*)|_X\|_\infty \leq \|(f-u)|_X\|_\infty$, $\forall u \in U$.
- Kriterium: $u^* \in U$ ist beste diskrete Approx., falls

$$|f(x_j) - u^*(x_j)| \leq \|(f-u^*)|_X\|_\infty$$

  für eine Alternate X für f und u^*.
- Lösen durch ein Gleichungssystem ($n+1$ Var., $n+1$ Bed.)
- Lösung des diskreten Approx.-Prob. ist immer existent und eindeutig

Remes-Verfahren

- Starte mit einer Referenz $X^{(0)}$, $j = 0$
- diskrete Approx. $u^{(j)}$ an f in $X^{(j)}$
- Falls $u^{(j)}$ beste Approx. an f ist $\rightarrow$ STOP.
- sonst: neue Referenz durch Austauschen (eines) der Punkte in $X^{(j)}$
- Wichtig: $\|(f-u_j^*)|_{X^{(j)}}\|_\infty < \|(f-u_{j+1}^*)|_{X^{(j+1)}}\|_\infty$
- Es gilt: Konvergenz (sublinear) zu eindeutiger Lösung

Alternantensatz

- U Haar'scher Raum der Dim. n, $f \in C[a,b]$, $u^* \in U$ ist beste Approx. an f aus U bzgl. $\|\cdot\|_\infty \Leftrightarrow$ es ex. eine Alternante X mit $\|(f-u^*)|_X\|_\infty = \|f-u^*\|_\infty$
- Beweis: Kriterium + Eindeutigkeit durch Remes-Verfahren
- Bemerkung: Für Unterräume U gilt: Haar'scher Raum $\Leftrightarrow$ Tschebyscheff-Raum (Tschebyscheff-Menge)

Kapitel 3

Numerik gewöhnlicher Differentialgleichungen

3.1 Einführung und Notation

Wir beschäftigen uns in diesem Kapitel hauptsächlich mit gewöhnlichen, expliziten Differentialgleichungen erster Ordnung, gegeben durch

$$x'(t) = f(t, x(t)), \qquad t \in I = [a, b] \tag{3.1}$$

Dabei ist

- $x : I \to \mathbb{R}^d$ eine *gesuchte*, differenzierbare Funktion auf einem Intervall $I = [a, b] \subseteq \mathbb{R}$ (Kurve) und $x'(t) = \begin{pmatrix} x_1'(t) \\ \vdots \\ x_d'(t) \end{pmatrix}$ der Tangentialvektor von x an t.
- $f : D \subseteq (\mathbb{R} \times \mathbb{R}^d) \to \mathbb{R}^d$ eine *gegebene* Funktion.

Wir klären zunächst einige Begriffe.

Notation 3.1

- *Eine Differentialgleichung heißt* ***gewöhnlich****, wenn die unbekannte Funktion x nur von einer reellen Variablen abhängt. Hängt x von mehreren Variablen ab, d.h. gilt*
$$x : B \to \mathbb{R}^d, B \subseteq \mathbb{R}^k,$$
so liegt eine ***partielle*** *Differentialgleichung vor.*
- *Eine Differentialgleichung hat* ***die Ordnung*** *k, falls nur Ableitungen von x bis zur Ordnung k vorkommen. Sie hat die Ordnung 1, falls nur die erste Ableitung von x vorkommt.*

- *Man nennt eine Differentialgleichung* ***explizit****, falls der höchste Ableitungsterm isoliert auftaucht, ansonsten* ***implizit****.*
- *Für $d = 1$ nennt man die Differentialgleichung* ***skalar****, für $d > 1$ spricht man auch von einem* ***System von Differentialgleichungen****.*

Beispiele:

- $F(t, x(t), x'(t)) = 0$, $t \in I = [a, b]$ mit einer gegebenen Funktion $F : \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ ist eine gewöhnliche, implizite Differentialgleichung erster Ordnung.
- $y^{(k)}(t) = g(t, y(t), \ldots, y^{(k-1)}(t))$, $t \in I = [a, b]$ mit einer gesuchten Funktion $y : I \to \mathbb{R}^2$, die k-mal differenzierbar ist, ist eine gewöhnliche, explizite Differentialgleichung der Ordnung k.
- $x'(t) = x(t)$, $t \in [a, b]$ ist eine gewöhnliche, explizite, skalare Differentialgleichung erster Ordnung.

Notation 3.2 *Eine gewöhnliche Differentialgleichung der Form*

$$x'(t) = f(x(t)),$$

bei der die rechte Seite nicht explizit von t abhängt, heißt ***autonom****.*

Wir beschäftigen uns im Wesentlichen mit expliziten, gewöhnlichen Differentialgleichungen.

Beispiel:

$$\begin{aligned} x_1'(t) &= -x_2(t) \\ x_2'(t) &= x_1(t) \end{aligned}$$

ist eine gewöhnliche, explizite und autonome Differentialgleichung (bzw. ein System von Differentialgleichungen) der Form

$$\begin{aligned} x'(t) &= f(t, x(t)) \\ \text{mit } f(t, x(t)) &= \begin{pmatrix} -x_2(t) \\ x_1(t) \end{pmatrix}. \end{aligned}$$

Eine Lösung dieser Differentialgleichung ist

$$x(t) = \begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix},$$

denn

$$\begin{array}{rcccccl} x_1'(t) & = & \cos'(t) & = & -\sin(t) & = & -x_2(t), \\ x_2'(t) & = & \sin'(t) & = & \cos(t) & = & x_1(t). \end{array}$$

Es gibt aber noch weitere Lösungen, nämlich

$$\tilde{x}(t) = C \cdot x(t - t_0) = \begin{pmatrix} C \cdot \cos(t - t_0) \\ C \cdot \sin(t - t_0) \end{pmatrix}$$

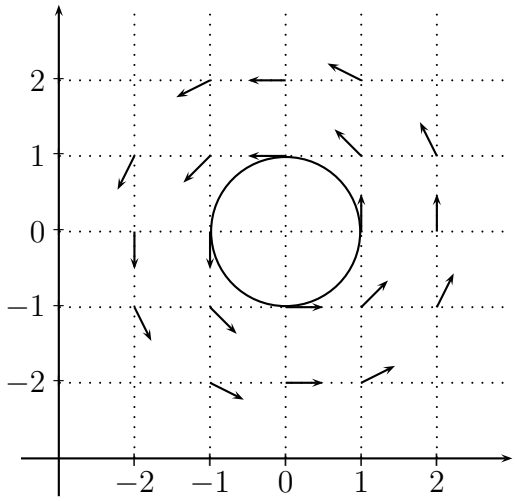
für alle $t_0 \in \mathbb{R}$ und $C \in \mathbb{R}$, denn

$$\tilde{x}'(t) = \begin{pmatrix} -C \cdot \sin(t - t_0) \\ C \cdot \cos(t - t_0) \end{pmatrix} = \begin{pmatrix} -\tilde{x}_2(t) \\ \tilde{x}_1(t) \end{pmatrix}.$$

Veranschaulichung:
Die rechte Seite der Differentialgleichung beschreibt ein Vektorfeld

$$f : \mathbb{R}^2 \to \mathbb{R}^2, \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} -x_2 \\ x_1 \end{pmatrix},$$

das man durch einen Vektor $\alpha \binom{-x_2}{x_1}$ in jedem Punkt $\binom{x_1}{x_2}$ skizzieren kann. Die Lösung $x(t) = \binom{\cos(t)}{\sin(t)}$ beschreibt eine Kurve im $\mathbb{R}^2$, zu der das Vektorfeld in jedem Punkt tangential ist.



Lemma 3.3 *Jede gewöhnliche, explizite Differentialgleichung der Ordnung k kann in eine äquivalente Differentialgleichung erster Ordnung transformiert werden.*

Beweis: Sei

$$y^{(k)}(t) = g(t, y(t), \ldots, y^{(k-1)}(t)), \qquad t \in I$$

mit einer gesuchten, k-mal differenzierbaren Funktion $y : I \to \mathbb{R}^d$ gegeben. Definiere $x_j : I \to \mathbb{R}^d$ durch

$$x_j(t) := y^{(j)}(t) \text{ für } j = 0, \ldots, k - 1.$$

Dann gilt:

$$x_j'(t) = y^{(j)'}(t) = y^{(j+1)}(t) = x_{j+1}(t) \text{ für } j = 0, \ldots, k-2$$

und

$$x_{k-1}'(t) = y^{(k-1)'}(t) = y^{(k)}(t) = g(t, y(t), \ldots, y^{(k-1)}(t)) = g(t, x_0(t), \ldots, x_{k-1}(t)),$$

also erhält man das System

$$\mathbb{R}^{kd} \ni x'(t) \left\{ \begin{array}{l} x_0'(t) = x_1(t) \\ x_1'(t) = x_2(t) \\ \quad \vdots \\ x_{k-2}'(t) = x_{k-1}(t) \\ x_{k-1}'(t) = g(t, x_0(t), \ldots, x_{k-1}(t)), \end{array} \right\} f(t, x(t))$$

in dem nur Ableitungen der Ordnung 1 vorkommen. Sei nun eine Lösung dieses Systems gegeben durch eine differenzierbare Funktionen x_j mit $j = 0, \ldots, k-1$. Dann ist

$$y(t) = x_0(t)$$

k-mal differenzierbar, da

$$y^{(j)}(t) = x_j(t) \text{ für } j = 0, \ldots, k-1$$

gilt und alle x_j mindestens einmal differenzierbar sind. Weiterhin gilt:

$$y^{(k)}(t) = y^{(k-1)'}(t) = x_{k-1}'(t) = g(t, x_0(t), \ldots, x_{k-1}(t)) = g(t, y(t), \ldots, y^{(k-1)}(t)).$$

QED

Die Lösung einer Differentialgleichung ist im Allgemeinen nicht eindeutig bestimmt. In dem Beispiel auf Seite 59 hatten wir zum Beispiel zwei Parameter C und t_0 zu wählen. Um die Eindeutigkeit zu erhalten, müssen die freien Variablen durch zusätzliche Bedingungen festgelegt werden.

Notation 3.4 *Ein **Anfangswertproblem (AWP)** einer gewöhnlichen Differentialgleichung erster Ordnung ist gegeben durch*

$$x'(t) = f(t, x(t)), \quad x(t_0) = x_0. \qquad \text{(AWP)}$$

*Ein **Randwertproblem** einer gewöhnlichen Differentialgleichung zweiter Ordnung ist gegeben durch*

$$x''(t) = f(t, x(t), x'(t)), \quad x(a) = r_a, \quad x(b) = r_b.$$

Dabei sind $x_0, r_a, r_b \in \mathbb{R}^d$.

Bemerkung: Die Gleichung $x(t) = x_0$ besteht aus d Bedingungen, sie legt also d Parameter fest (falls sie eindeutig lösbar ist).

Bemerkung: Die numerische Behandlung von Randwertproblemen und Anfangswertproblemen ist unterschiedlich. In dieser Vorlesung befassen wir uns mit Anfangswertproblemen.

Wir kommen noch einmal auf autonome Differentialgleichungen zurück.

Lemma 3.5 *Sei $x : I \to \mathbb{R}^d$ eine Lösung einer autonomen Differentialgleichung $x'(t) = f(x(t))$. Dann ist*

$$y : I \to \mathbb{R}^d, t \mapsto x(t - t_0)$$

auch eine Lösung der Differentialgleichung und zwar für alle $t_0 \in \mathbb{R}$.

Beweis:

$$y'(t) = x'(t - t_0) = f(x(t - t_0)) = f(y(t))$$

QED

Bemerkung: Im Beispiel auf Seite 59 haben wir die Aussage genutzt, um Lösungen zu erzeugen.

Bemerkung: Oft beschreibt der Parameter t die Zeit. Die Aussage des Lemmas lautet dann: Die Lösung einer autonomen Differentialgleichung ist invariant gegenüber Zeittransformationen.

Lemma 3.6 *Jedes Anfangswertproblem der Form $x'(t) = f(t, x(t))$, $x(t_0) = x_0$ lässt sich in ein äquivalentes, autonomes Anfangswertproblem transformieren.*

Beweis: Definiere $s(t) := t$ und $y(t) = \binom{s(t)}{x(t)}$. Betrachte das autonome System

$$y'(t) = \begin{pmatrix} s'(t) \\ x'(t) \end{pmatrix} = \begin{pmatrix} 1 \\ f(y(t)) \end{pmatrix}, \quad y(t_0) = \begin{pmatrix} s(t_0) \\ x(t_0) \end{pmatrix} = \begin{pmatrix} t_0 \\ x_0 \end{pmatrix}. \tag{3.2}$$

- Sei x eine Lösung von $x'(t) = f(t, x(t))$, $x(t_0) = x_0$. Mit $s(t) := t$ erhalten wir

  $$y'(t) = \begin{pmatrix} s'(t) \\ x'(t) \end{pmatrix} = \begin{pmatrix} 1 \\ f(t, x(t)) \end{pmatrix} = \begin{pmatrix} 1 \\ f(s(t), x(t)) \end{pmatrix} = \begin{pmatrix} 1 \\ f(y(t)) \end{pmatrix},$$

  also eine Lösung von (3.2).

- Sei nun $y(t) = \binom{s(t)}{x(t)}$ eine Lösung von (3.2). Dann gilt:

  $$s'(t) = 1, \text{ setze also } s(t) = t.$$

  Damit ist

  $$x'(t) = f(y(t)) = f(s(t), x(t)) = f(t, x(t))$$

  eine Lösung von $x'(t) = f(t, x(t))$.

Den Übergang eines Anfangswertproblemes zu (3.2) nennt man auch **Autonomisierung** des Anfangswertproblemes.

Zwei praktische Anwendungen

Bewegung eines Massepunktes. Die Bewegung eines Massepunktes zur Zeit t am Ort x kann durch die Differentialgleichung 2. Ordnung

$$m \cdot x''(t) = g(t, x)$$

beschrieben werden. Die Funktion g beschreibt dabei die Wirkung äußerer Kräfte, z.B. erhält man bei einer einseitig gespannten Feder $g(t, x) = -kx$, wobei k die Federkonstante bezeichnet. Weiterhin ist meist der Anfangspunkt $x_0 = x(t_0)$ und die Anfangsgeschwindigkeit $x'_0 = x'(t_0)$ vorgegeben.
Das System kann in das folgende äquivalente System 1. Ordnung verwandelt werden:

$$\begin{aligned} x'_1(t) &= x_2(t) \\ x'_2(t) &= -\tfrac{k}{m} x_1(t), \end{aligned}$$

mit Anfangsbedingungen

$$x_1(t_0) = x_0, \quad x_2(t_0) = x'_0.$$

Dieses System von Differentialgleichungen ist erster Ordnung, linear und autonom. Die Lösung ist gegeben durch

$$\begin{aligned} x(t) &= x_1(t) = x_0 \cos\left(\sqrt{\tfrac{k}{m}}\, t\right) + x'_0 \sin\left(\sqrt{\tfrac{k}{m}}\, t\right) \\ x'(t) &= x_2(t) \end{aligned}$$

Volterra-Lottka Zyklus. Betrachte ein ökologisches System mit zwei Arten, bei denen die eine Art der anderen als Nahrung dient. Entsprechend bezeichnen wir sie als "Jäger" und "Beute". Sei

$$\begin{aligned} x_J(t) &= \text{die Größe der Jäger-Population zur Zeit } t \text{ und} \\ x_B(t) &= \text{die Größe der Beute-Population zur Zeit } t. \end{aligned}$$

Die Wachstumsrate der Populationen ergibt sich aus der Differenz der Geburtenrate und der Sterberate. Dabei nehmen wir an, dass für die Beute-Population genügend Nahrung vorhanden sei, so dass sie sich (im ungestörten Fall) exponentiell vermehren würde, die Geburtenrate also konstant ist. Mit geeigneten Parametern $\alpha, \beta > 0$ ergibt sich dann

$$x'_B(t) = \alpha x_B(t) - \beta x_B(t) x_J(t).$$

Die Gleichung kann wie folgt interpretiert werden:

- das ungestörte eigene Wachstum der Beute-Population resultiert aus einem exponentiellen Wachstum $x_B = e^{\alpha x}$ und ist daher durch $x'_B = \alpha x_B$ beschrieben.
- die Anzahl der durch Jagd gestorbenen Beutetiere ist proportional zur Rate, mit der sich Jäger und Beute treffen, auf einem begrenzten Gebiet also proportional zu x_B und proportional zu x_J.

Für die Jäger-Population ergibt sich

$$x'_J(t) = \gamma x_J(t) x_B(t) - \delta x_J(t),$$

ebenfalls mit geeigneten Parametern $\gamma, \delta > 0$. Die Interpretation dieser Gleichung ist wie folgt:

- Die Jäger-Population wächst exponentiell mit Rate γ und proportional zur Beute-Population x_B,
- die natürliche Sterberate ist (bei exponentiellem Wachstum) $x'_J = -\delta x_J$.

Die Lösung dieses Systems von Differentialgleichungen führt zu periodischen Lösungen, die man auch *Volterra-Lottka-Zyklen* nennt. Bilder dazu finden sich z.B. in der Wikipedia.

Wir beenden diesen einführenden Abschnitt mit einer letzten Notation.

Notation 3.7 *Ein System von Differentialgleichungen heißt* ***linear****, falls*

$$x'(t) = f(t,x) := A(t)x + g(t)$$

gilt, wobei $g : I \to \mathbb{R}^d$ *eine stetige Funktion ist und* $A = (a_{ij})_{i,j=1,\ldots,d}$ *eine* $d \times d$*-Matrix mit stetigen Einträgen* $a_{ij} : I \to \mathbb{R}$*.*

Von den beiden oben beschriebenen Anwendungsbeispielen ist das erste linear, das Volterra-Lottka-System aber nichtlinear.

3.2 Existenz und Eindeutigkeit

In diesem Abschnitt wollen wir die Existenz und die Eindeutigkeit von Lösungen für Anfangswertprobleme der Form

$$\begin{aligned} x'(t) &= f(t, x(t)) \\ x(t_0) &= x_0 \end{aligned}$$

untersuchen. Wir zeigen zunächst zwei Beispiele.

- Das erste Beispiel zeigt, dass die Lösung im Allgemeinen nicht eindeutig sein muss. Sei folgendes Anfangswertproblem

$$\begin{aligned} x'(t) &= |x(t)|^\alpha \\ x(0) &= 0 \end{aligned}$$

für einen Parameter $\alpha \in (0,1)$ gegeben. Die Differentialgleichung hat die folgenden beiden Lösungen $\tilde{x}$ und x:

$$\begin{aligned} \tilde{x}(t) &\equiv 0 \\ x(t) &= \begin{cases} ((1-\alpha)t)^{\frac{1}{1-\alpha}} & \text{für } t \geq 0 \\ 0 & \text{für } t < 0 \end{cases} \end{aligned}$$

Für $\tilde{x}$ sieht man das direkt, für die zweite Lösung x rechnet man nach:

  - $x(0) = 0$,
  - $x'(t) = |x(t)|^\alpha$ für $t \geq 0$ und $x'(t) = 0$ für $t < 0$,
  - und $x(0) = x'(0) = 0$, also ist x stetig und differenzierbar.

- Das zweite Beispiel zeigt, dass keine Lösung auf ganz I existieren muss: Betrachten wir

$$\begin{aligned} x'(t) &= (x(t))^2 \\ x(0) &= 1. \end{aligned}$$

Die Lösung $x(t) = -\frac{1}{t-1}$ ist nur für $t \neq 1$ definiert und kann wegen

$$\lim_{t \to 1} x(t) = \infty$$

nicht als stetige Funktion für $t \geq 1$ fortgesetzt werden. Tatsächlich existiert in diesem Fall keine Lösung des Anfangswertproblemes für *alle* $t > 0$. Der Effekt wird auch "blow up" genannt.

Um die Frage nach Existenz und Eindeutigkeit von Lösungen für Anfangswertprobleme zu beantworten, formulieren wir (AWP) zu einer so genannten *Integralgleichung* um.

Lemma 3.8 *Sei $D \subseteq \mathbb{R}^{d+1}$ offen, $f : D \to \mathbb{R}^d$ stetig, $a \leq t_0 \leq b$ und $x : [a,b] \to \mathbb{R}^d$ eine Funktion. Es gelte*

$$\{(t, x(t)) : t \in [a,b]\} \subseteq D.$$

Dann sind die folgenden Aussagen äquivalent:

1. *x ist stetig differenzierbar und löst das (AWP)*

$$\begin{aligned} x'(t) &= f(t, x(t)),\ t \in [a, b] \\ x(t_0) &= x_0 \end{aligned}$$

2. *x ist stetig und erfüllt die Integralgleichung*

$$x(t) = x_0 + \int_{t_0}^{t} f(\tau, x(\tau))d\tau, \quad t \in [a, b]. \tag{3.3}$$

Beweis: $1 \Longrightarrow 2$: Sei $x'(t) = f(t, x(t))$, $x(t_0) = x_0$ eine Lösung des Anfangswertproblemes. Nach dem Hauptsatz der Differential- und Integralrechnung gilt dann

$$\begin{aligned} x(t) &= x(t_0) + \int_{t_0}^{t} x'(\tau)d\tau \\ &= x_0 + \int_{t_0}^{t} f(\tau, x(\tau))d\tau. \end{aligned}$$

$2 \Longrightarrow 1$: Sei nun $x(t) = x_0 + \int_{t_0}^{t} f(\tau, x(\tau))d\tau$. Da f und x beide stetig sind, ist $\int_{t_0}^{t} f(\tau, x(\tau))d\tau$ stetig nach t differenzierbar. Also ist x stetig differenzierbar und die Ableitung von x ist gegeben durch

$$x'(t) = \frac{d}{dt} \int_{t_0}^{t} f(\tau, x(\tau))d\tau = f(t, x(t))$$

nach dem Hauptsatz der Differential- und Integralrechnung. Weiter gilt:

$$x(t_0) = x_0 + \int_{t_0}^{t_0} f(\tau, x(\tau))d\tau = x_0$$

QED

Wozu hilft uns dieses Lemma? Der Vorteil liegt darin, dass wir durch die Integralgleichung eine Fixpunktgleichung in der unbekannten Funktion x gefunden haben. Diese sieht wie folgt aus:
Wir definieren den Operator F, den wir auf $x : I \to \mathbb{R}^d$ anwenden wollen durch

$$(F(x))(t) := x_0 + \int_{t_0}^{t} f(\tau, x(\tau))d\tau.$$

Dann kann man die Integralgleichung (3.3) schreiben als

$$x(t) = (F(x))(t)$$

oder, kürzer, als

$$x = F(x).$$

Unsere gesuchte Lösung x kann also als die Lösung einer Fixpunktgleichung in einem unendlich dimensionalen Raum aufgefasst werden. Wir wollen darauf nun den Banach'schen Fixpunktsatz anwenden. Dieser wurde in Numerik I behandelt. Zur Wiederholung erinnern wir daran, dass jeder vollständige und normierte Raum ein **Banach-Raum** ist, und dass für eine Teilmenge U eines Banachraumes X eine Abbildung $\Phi : U \to X$ **kontrahierend** ist, falls es einen reellen Kontraktionsfaktor $q < 1$ so gibt, dass

$$\|\Phi(x) - \Phi(y)\| \leq q\|x - y\| \text{ für alle } x, y \in U.$$

Der Banach'sche Fixpunktsatz lautet wie folgt:

Satz 3.9 (Banach'scher Fixpunktsatz) *Sei X ein Banach-Raum mit Norm $\|\cdot\|$ und $U \subseteq X$ eine abgeschlossene Teilmenge von X. Sei weiterhin $F : U \to U$ eine kontrahierende Abbildung mit Kontraktionsfaktor $q < 1$. Dann hat die Fixpunktgleichung $F(x) = x$ einen eindeutigen Fixpunkt x^*.*

Für den Beweis verweisen wir auf die Vorlesung Numerik I.
Im Folgenden bezeichne $\|\cdot\|_2$ die Euklidische Norm. Wir erinnern an die folgende Bezeichnung.

Notation 3.10

- *Sei $f : D \to \mathbb{R}^d$, $D \subseteq \mathbb{R}^{d+1}$. f ist **Lipschitzstetig bezüglich seiner letzten d Variablen**, falls zu jedem $(t_0, x_0) \in D$ eine Umgebung $U := U(t_0, x_0) \subseteq D$ und eine Konstante $L = L(t_0, x_0)$ so existiert, dass*

$$\|f(t, x) - f(t, y)\|_2 \leq L\|x - y\|_2 \text{ für alle } (t, x), (t, y) \in U.$$

- *Sei $f : D \to \mathbb{R}^d$, $D = I \times \mathbb{R}^d$. f ist **global Lipschitzstetig bezüglich seine letzten d Variablen**, falls es eine Konstante $L > 0$ so gibt, dass*

$$\|f(t, x) - f(t, y)\|_2 \leq L\|x - y\|_2 \text{ für alle } t \in I \text{ und } x, y \in \mathbb{R}^d.$$

Damit formulieren wir nun das Hauptergebnis dieses Abschnitts.

Satz 3.11 (Satz von Picard-Lindelöf) *Sei $D \subseteq \mathbb{R}^{d+1}$ offen und sei $f : D \to \mathbb{R}^d$ stetig und bezüglich der letzten d Variablen Lipschitzstetig. Dann existiert zu jedem $(t_0, x_0) \in D$ eine Umgebung I von t_0, auf der das Anfangswertproblem*

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

eindeutig lösbar ist.

Bemerkung: Der Satz liefert nur die *lokale* Existenz von Lösungen, also auf kleinen Intervallen für t um t_0.

Beweis: Seien $(t_0, x_0) \in D$ gegeben. Weil f Lipschitzstetig ist, existiert $\bar{U} := U(t_0, x_0) \subseteq D$ und $L = L(t_0, x_0)$ so, dass

$$\|f(t,x) - f(t,y)\|_2 \leq L\|x - y\|_2, \text{ für alle } (t,x),\ (t,y) \in \bar{U}.$$

Wir wählen nun $\alpha, \beta > 0$ so, dass für

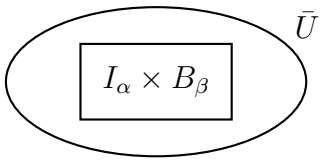
$$\begin{aligned} I_\alpha &= \{t \in \mathbb{R} : |t - t_0| \leq \alpha\} = [t_0 - \alpha, t_0 + \alpha] \\ \text{und} \quad B_\beta &= \{x \in \mathbb{R}^d : \|x - x_0\|_2 \leq \beta\} \end{aligned}$$

gilt:

$$I_\alpha \times B_\beta \subseteq \bar{U}.$$

Da f stetig auf der kompakten Menge $I_\alpha \times B_\beta$ ist, existiert

$$M := \max_{(t,x) \in I_\alpha \times B_\beta} \|f(t,x)\|_2.$$



Wir wählen α^* mit

$$0 < \alpha^* \leq \min\left(\frac{\beta}{M}, \alpha\right),$$

d.h. $\alpha^* > 0$, $\alpha^* \leq \alpha$ und $\alpha^* M \leq \beta$. Sei weiterhin

$$I^* := [-\alpha^* + t_0, \alpha^* + t_0].$$

Wir wollen den Banach'schen Fixpunktsatz anwenden und wählen dazu

- als Banachraum: $X := C(I^*, \mathbb{R}^d)$ als Menge der stetigen Funktionen von I^* nach $\mathbb{R}^d$,
- als Norm
$$\|x\|_B := \sup_{t \in I^*} e^{-2L|t-t_0|}\|x(t)\|_2, \text{ für alle } x \in X,$$
- als Teilmenge U den Unterraum
$$U := \{x \in X : \sup_{t \in I^*} \|x(t) - x_0\|_2 \leq \beta\}$$

- und als Abbildung $F : U \to X$, $x \mapsto F(x)$ den vorhin schon genannten Operator F, der durch

$$(F(x))(t) := x_0 + \int_{t_0}^{t} f(\tau, x(\tau))d\tau$$

definiert ist.

Jetzt überprüfen wir die Voraussetzungen des Banachschen Fixpunktsatzes.

1) X ist Vektorraum. Es ist leicht zu zeigen, dass $\|\cdot\|_B$ Norm auf X ist. Dass X vollständig ist, kann man mit Methoden der Analysis nachrechnen.

2) U ist abgeschlossen. Sei dazu (x_n) mit $x_n \in U$ eine Folge, die bezüglich $\|\cdot\|_B$ (gleichmäßig) gegen $x \in X$ konvergiert.

   Wir wollen zeigen, dass $x \in U$. Dazu berechnen wir:

$$\|x(t) - x_0\|_2 = \|\lim_{n\to\infty} x_n(t) - x_0\|_2 = \lim_{n\to\infty} \|x_n(t) - x_0\|_2,$$

   denn die Normfunktion ist stetig. Weil $x_n, x_0 \in U$ ist, gilt weiter

$$\|x_n(t) - x_0\|_2 \leq \beta \text{ für alle } t \in I^* \text{ und alle } n \in \mathbb{N},$$

   also

$$\|x(t) - x_0\|_2 = \lim_{n\to\infty} \underbrace{\|x_n(t) - x_0\|_2}_{\leq \beta \ \forall n} \leq \beta \text{ für alle } t \in I^*.$$

   Es folgt: $x \in U$.

3) Sei $F : U \to U$, sei $x \in U$. Dann ist $F(x) \in X$. Wir wollen zeigen, dass $F(x) \in U$, d.h.

$$\sup_{t\in I^*} \|(F(x))(t) - x_0\|_2 \leq \beta$$

   und berechnen dazu:

$$\begin{aligned} \|(F(x))(t) - x_0\|_2 &= \left\| \int_{t_0}^{t} f(\tau, x(\tau))d\tau \right\|_2 \\ &\leq |t - t_0| + \max_{\tilde{t}\in I^*, \tilde{x}\in B_\beta} \|f(\tilde{t}, \tilde{x})\|_2 \qquad (3.4) \\ &\leq \underbrace{\alpha^*}_{t,t_0\in I^*} \cdot \underbrace{M}_{\substack{\text{Def.} \\ \text{von } M}} \leq \underbrace{\beta}_{\substack{\text{Def.} \\ \text{von } \alpha^*}} \text{ für alle } t \in I^*. \end{aligned}$$

   In Abschätzung (3.4) darf man über der Menge $\tilde{t} \in I^*, \tilde{x} \in B_\beta$ maximieren, weil

   - mit $t, t_0 \in I^*$ das ganze Intervall zwischen t und t^* in I^* liegt, und weil

- aus $x \in U$ folgt, dass $\|x(\tau) - x_0\|_2 \leq \beta$ und entsprechend $\{x(\tau) : \tau \in [t_0, t]\} \subseteq B_\beta$.

Also folgt: $F(x) \in U$.

4) F ist Kontraktion. Wähle $x, y \in U$ und betrachte

$$e^{-2L|t-t_0|}\|(F(x))(t) - (F(y))(t)\|_2. \tag{3.5}$$

Der Übersicht halber betrachten wir zunächst nur:

$$\begin{aligned}
&\|(F(x))(t) - (F(y))(t)\|_2 \\
&= \left\| \int_{t_0}^{t} f(\tau, x(\tau)) - f(\tau, y(\tau)) d\tau \right\|_2 \\
&\leq \operatorname{sign}(t - t_0) \int_{t_0}^{t} \|f(\tau, x(\tau)) - f(\tau, y(\tau))\|_2 \, d\tau \\
&\leq L \operatorname{sign}(t - t_0) \int_{t_0}^{t} \|x(\tau) - y(\tau)\|_2 d\tau \\
&\leq L \operatorname{sign}(t - t_0) \int_{t_0}^{t} e^{2L|\tau - t_0|} \underbrace{e^{-2L|\tau - t_0|}\|x(\tau) - y(\tau)\|_2}_{\leq \|x - y\|_B} d\tau \\
&\leq L \operatorname{sign}(t - t_0) \int_{t_0}^{t} e^{2L|\tau - t_0|} d\tau \|x - y\|_B \\
&\leq L \operatorname{sign}(t - t_0) \left[\frac{1}{2L} e^{2L|t - t_0|} \right]_{t_0}^{t} \|x - y\|_B \\
&= L \operatorname{sign}(t - t_0) \frac{1}{2L} \operatorname{sign}(t - t_0)(e^{2L|t - t_0|} - 1)\|x - y\|_B \\
&= \frac{1}{2}(e^{2L|t - t_0|} - 1)\|x - y\|_B.
\end{aligned}$$

Dieses setzen wir jetzt in (3.5) ein und erhalten

$$\begin{aligned}
&e^{-2L|t-t_0|}\|(F(x))(t) - (F(y))(t)\|_2 \\
&\qquad \leq \tfrac{1}{2}\|x - y\|_B \underbrace{\left(1 - \underbrace{e^{-2L|t-t_0|}}_{\geq 0}\right)}_{\leq 1} \leq \tfrac{1}{2}\|x - y\|_B,
\end{aligned}$$

also gilt

$$\begin{aligned}
\|F(x) - F(y)\|_B &= \sup_{t \in I^*} e^{-2L|t-t_0|}\|(F(x))(t) - (F(y))(t)\|_2 \\
&\leq \tfrac{1}{2}\|x - y\|_B,
\end{aligned}$$

d.h. F ist Kontraktion mit $q = \frac{1}{2}$.

Somit sind alle Voraussetzungen des Banachschen Fixpunktsatzes erfüllt und wir erhalten:

$$x = F(x) \text{ besitzt eine eindeutige Lösung in } U.$$

Abschließend müssen wir noch ausschließen, dass F noch einen weiteren Fixpunkt $y \in X$ mit $\{(t, y(t)) : t \in I^*\} \subseteq D$ besitzt, der nicht in U liegt. Dazu ersetzen wir im vorhergehenden Beweis β durch $\beta/2$ und erhalten wie unter Punkt 2):

$$\|x(t) - x_0\|_2 \leq \frac{\beta}{2} \text{ für } |t - t_0| \leq \tilde{\alpha} := \min\left(\frac{\beta}{2M}, \alpha\right).$$

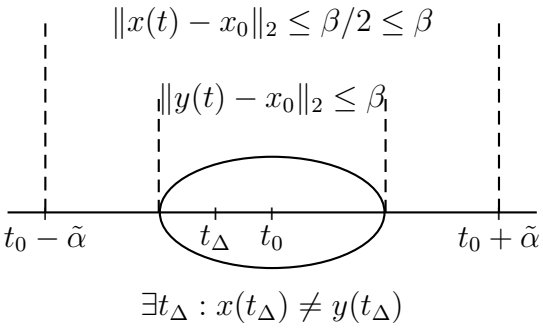
Sei weiterhin unser „neues“ U

$$\tilde{U} = \{x \in X : \sup_{t:|t-t_0|<\tilde{\alpha}} \|x(t) - x_0\|_2 \leq \beta/2\}.$$

Angenommen, so ein Fixpunkt $y \in X \setminus \tilde{U}$ existiert und löst damit (AWP). Wegen $y(t_0) = x_0$ gibt es α^{**} mit $0 < \alpha^{**} < \tilde{\alpha}$ und

$$\|y(t) - x_0\| \leq \beta \text{ für } |t - t_0| < \alpha^{**}.$$

sowie $x(t) \neq y(t)$ für mindestens ein t_Δ mit $|t_\Delta - t_0| < \alpha^{**}$. Weil $\|x(t) - x_0\| \leq \beta/2 \leq \beta$ für alle $|t - t_0| < \tilde{\alpha}$,



gäbe es aber auf $I^{**} := [-\alpha^{**} + t_0, \alpha^{**} + t_0]$ zwei verschiedene Lösungen $x,\ y$ der Fixpunktgleichung, die beide in

$$U^{**} = \{x \in X : \sup_{t \in I^{**}} \|x(t) - x_0\|_2 \leq \beta\}$$

liegen, was nach dem Banachschen Fixpunktsatz nicht sein kann. QED

Die globale Existenz einer Lösung liefert der folgende Satz:

Satz 3.12 *Sei $I \subseteq \mathbb{R}$ ein Intervall, sei $D = I \times \mathbb{R}^d$ und sei $f : D \to \mathbb{R}^d$ bezüglich der letzten d Variablen global Lipschitzstetig. Dann besitzt das Anfangswertproblem $x'(t) = f(t, x(t))$, $x(t_0) = x_0$ für alle $(t_0, x_0) \in D$ eine eindeutige Lösung $x : I \to \mathbb{R}^d$.*

Beweis: Im Beweis des Satzes von Picard-Lindelöf setzen wir $I_\alpha = I_* = I$ und wählen als Teilmenge U den ganzen Banachraum, also $U = X$. Die Konstanten $\alpha, \alpha^*, \beta, M$ werden nun nicht mehr benötigt. Die Details werden hier nicht ausgeführt. QED

Beispiel: Wir untersuchen die Voraussetzungen der Sätze 3.11 und 3.12 am zweiten Beispiel auf Seite 65,

$$\begin{aligned} x'(t) &= (x(t))^2 \\ x(0) &= 1. \end{aligned}$$

Wir erhalten $f(t,x) = x^2$ und entsprechend

$$\|f(t,x) - f(t,y)\|_2 = |x^2 - y^2| = |x+y| \cdot |x-y| \leq L \cdot |x-y| \text{ für alle } x, y \in I$$

falls

$$L \geq |x+y| \text{ für alle } x, y \in I$$

gilt.
Das ist auf jedem beschränkten Intervall erfüllt, nicht aber auf $I = [0, \infty)$ oder auf $I = \mathbb{R}$. Das Anfangswertproblem erfüllt daher die Voraussetzungen von Satz 3.11, aber nicht die von Satz 3.12, was zu dem so genannten "blow up" Effekt führt.

Unter den Voraussetzungen von Satz 3.12 kann dieser "blow up" Effekt nicht auftreten.

Bemerkung: Für lineare Differentialgleichungen

$$x'(t) = A(t)x(t) + g(t)$$

mit $t \in I$ und $f(t,x) = A(t)x + g(t)$ erhält man

$$\begin{aligned} \|f(t,x) - f(t,y)\|_2 &= \|A(t)x + g(t) - A(t)y - g(t)\|_2 \\ &= \|A(t)(x-y)\|_2 \leq \|A(t)\|_2 \|x-y\|_2 \\ &\leq L\|x-y\|_2, \end{aligned}$$

falls $L := \sup_{t \in I} \|A(t)\|_2 < \infty$.
Die Voraussetzungen von Satz 3.12 sind für lineare Differentialgleichungen also erfüllt, falls

$$\sup_{t \in I} \|A(t)\|_2 < \infty.$$

Das gilt insbesondere auf jedem kompakten Intervall I.

Der Banach'sche Fixpunktsatz liefert nicht nur theoretische Aussagen über Existenz und Eindeutigkeit, sondern mit dem Verfahren der sukzessiven Approximation auch ein konvergentes Verfahren zur Bestimmung des Fixpunktes. Dieses Verfahren lässt durch folgenden Iterationsschritt (so genannte *Picard-Iterationen*) auf Anfangswertprobleme anwenden:

$$x^{(n+1)}(t) := x^{(n)}(t_0) + \int_{t_0}^{t} f(\tau, x^{(n)}(\tau))d\tau$$

Als Startwert kann man z.B. $x^{(0)}(t) := x_0$ wählen – das resultierende Verfahren ist allerdings durch die dazu nötige numerische Auswertung der zahlreich auftretenden Integrale ineffizient und wird in der Praxis fast nicht verwendet.

Satz 3.13 (Globale Eindeutigeit) *Sind die Voraussetzungen von Satz 3.11 erfüllt und sind x und y Lösungen des (AWP)*

$$\begin{aligned} x'(t) &= f(t, x(t)) \\ x(t_0) &= x_0. \end{aligned}$$

auf einem beliebigen Intervall I mit $t_0 \in I$, so gilt $x(t) = y(t)$ für alle $t \in I$.

Beweis: Sei $I = [a, b]$, $t_0 \in I$ und seien x und y Lösungen des (AWP). Wähle $I' \subseteq I$ als das längste Intervall mit $x(t) = y(t)$ für alle $t \in I'$.
Wir möchten zeigen, dass $I = I'$.
Angenommen, dies ist gilt nicht, dann sei $I' = [a', b'] \subset I$. Dann ist ohne Beschränkung der Allgemeinheit $b' < b$. Wir betrachten das neue (AWP')

$$\begin{aligned} z'(t) &= f(t, z(t)) \\ z(b') &= x(b'). \end{aligned}$$

Nach Satz 3.11 existiert eine Umgebung $U = (b' - \alpha, b' + \alpha)$ mit $\alpha > 0$ auf der (AWP') eindeutig lösbar ist. Weil x und y beides Lösungen für (AWP') sind, folgt also $x(t) = y(t)$ für alle $t \in U$. Das ist ein Widerspruch zur Maximalität von I'.
QED

Abschließend geben wir noch ein Kriterium an, anhand dessen man die geforderte Lipschitz-Bedingung von Satz 3.11 nachweisen kann.

Lemma 3.14 *Ist $f : I \times \mathbb{R}^d \to \mathbb{R}^d$ bezüglich x stetig partiell differenzierbar, so erfüllt f die Lipschitz-Bedingung des Satzes 3.11 für alle $(t_0, x_0) \in I \times \mathbb{R}^d$.*

Beweis: (Vergleiche auch den Beweis von Lemma 5.7 aus Numerik I).
Weil f bezüglich x stetig partiell differenzierbar ist, existiert der Gradient

$$D_x f(t, x) : \mathbb{R} \times \mathbb{R}^d \to (\mathbb{R}^d)^*$$

und es gilt $L := \sup_{(t,x)\in\bar{U}(t_0,x_0)} \|D_x f(t,x)\|_2 < \infty$, wenn die Umgebung $\bar{U}$ kompakt gewählt wird. Wählt man $\bar{U}$ zusätzlich konvex, so kann man mittels

$$g(\xi) := f(t, x + \xi(y - x))$$

folgern, dass

$$\begin{aligned}
\|f(t,y) - f(t,x)\|_2 &= \|g(1) - g(0)\|_2 = \left\| \int_0^1 g'(\tau) d\tau \right\| \\
&= \left\| \int_0^1 D_x f(t, x + \tau(y-x)) \cdot (y - x) d\tau \right\|_2 \text{ multivariate Kettenregel} \\
&\leq \int_0^1 \|D_x f(t, x + \tau(y-x))\|_2 \cdot \|y - x\|_2 d\tau \\
&\leq \int_0^1 L\|y - x\|_2 d\tau = L\|y - x\|_2.
\end{aligned}$$

QED

Die Aussage von Satz 3.11 nutzen wir nun, um die *Evolution* zu definieren.

Definition 3.15 *Sei $D \subseteq \mathbb{R}^{d+1}$ offen, $f : D \to \mathbb{R}^d$ stetig und Lipschitzstetig bezüglich der letzten d Variablen. Seien $t_0, t \in I$ und $|t - t_0|$ hinreichend klein. Dann definiert man eine zweiparametrige Funktion*

$$\Phi^{t,t_0} : \mathbb{R}^d \to \mathbb{R}^d$$

durch $\Phi^{t,t_0}(x_0) := x(t)$, wobei $x(t)$ die eindeutige (lokale) Lösung des Anfangswertproblemes

$$\begin{aligned}
x'(t) &= f(t, x(t)) \\
x(t_0) &= x_0
\end{aligned}$$

*ist. Man nennt Φ die **Evolution** der Differentialgleichung $x'(t) = f(t, x(t))$.*

Φ^{t,t_0} bildet den Wert der Lösung x zur Zeit t_0 auf den Wert der gleichen Lösung zur Zeit t ab.

Beispiel: Betrachte $x'(t) = (x(t))^2$, also $f(t,x) = x^2$. Dann ist die eindeutige (lokale) Lösung zu (t_0, x_0) mit $t_0 = 0$, $x_0 > 0$ gegeben durch

$$x(t) = \frac{x_0}{1 - t x_0}, \quad \text{für } t < \frac{1}{x_0}.$$

Für die Evolution gilt entsprechend im Fall $t > 0$

$$\Phi^{t,0}(x_0) = \frac{x_0}{1 - t x_0} \text{ für } x_0 < \frac{1}{t}.$$

Lemma 3.16 *Die Evolution Φ der Differentialgleichung $x'(t) = f(t, x(t))$ besitzt die folgenden Eigenschaften:*

(Ev1) $\Phi^{t_0,t_0}(x_0) = x_0$

(Ev2) $\frac{\partial}{\partial\tau}\Phi^{t+\tau,t}(x_0)|_{\tau=0} = f(t, x_0)$

(Ev3) $\Phi^{t_2,t_0}(x_0) = \Phi^{t_2,t_1}(\Phi^{t_1,t_0}(x_0))$

für alle $(t_0, x_0) \in D$ und $|t_1 - t_0|$, $|t_2 - t_0|$ und $|t - t_0|$ hinreichend klein. Weiter ist Φ durch diese drei Bedingungen eindeutig charakterisiert.

Beweis: **(Ev1)** gilt weil $\Phi^{t_0,t_0}(x_0) = x(t_0) = x_0$.

(Ev2) Seien x_0, t fest. Sei x die Lösung des Anfangswertproblemes zum Startwert (t, x_0). Definiere

$$g(\tau) := \Phi^{t+\tau,t}(x_0) = x(t + \tau).$$

Dann gilt

$$\begin{aligned} & \frac{\partial}{\partial\tau}\Phi^{t+\tau,t}(x_0) = g'(\tau) = x'(t + \tau) = f(t + \tau, x(t + \tau)) \\ \Longrightarrow \quad & \frac{\partial}{\partial\tau}\Phi^{t+\tau,t}(x_0)|_{\tau=0} = g'(0) = f(t, x(t)) = f(t, x_0) \end{aligned}$$

(Ev3) Sei x Lösung von

$$\begin{aligned} x'(t) &= f(t, x(t)) \\ x(t_0) &= x_0, \end{aligned}$$

das heißt $\Phi^{t,t_0}(x_0) = x(t)$ für alle t nahe genug an t_0. Damit gilt:

$$\begin{aligned} \Phi^{t_2,t_1}(\Phi^{t_1,t_0}(x_0)) &= \Phi^{t_2,t_1}(x(t_1)) \\ &= x(t_2) = \Phi^{t_2,t_0}(x_0), \end{aligned}$$

wobei die vorletzte Gleichheit gilt, weil für $t_2 - t_0$ hinreichend klein x auch Lösung ist von dem Anfangswertproblem

$$\begin{aligned} y'(t) &= f(t, y(t)) \\ y(t_1) &= x(t_1). \end{aligned}$$

(Eindeutigkeit) Sei $\Psi^{t,t_0} : \mathbb{R}^d \to \mathbb{R}^d$ eine Funktion, die ebenfalls die drei Bedingungen (Ev1),(Ev2) und (Ev3) erfüllt. Sei (t_0, x_0) beliebig. Definiere

$$x(t) := \Psi^{t,t_0}(x_0).$$

Dann gilt

$$\begin{aligned} x'(t) &= \frac{\partial}{\partial \tau} \Psi^{t+\tau,t_0}(x_0)|_{\tau=0} \\ &= \frac{\partial}{\partial \tau} \left(\Psi^{t+\tau,t}\left(\Psi^{t,t_0}(x_0)\right)\right)|_{\tau=0} \text{ wegen (Ev3)} \\ &= f(t, \Psi^{t,t_0}(x_0)) \text{ wegen (Ev2)} \\ &= f(t, x(t)) \end{aligned}$$

und wegen (Ev1) ist außerdem $x(t_0) = \Psi^{t_0,t_0}(x_0) = x_0$. Also ist nach Satz 3.11

$$x(t) = \Phi^{t,t_0}(x_0)$$

die eindeutige (lokale) Lösung des Anfangswertproblemes

$$\begin{aligned} x'(t) &= f(t, x(t)) \\ x(t_0) &= x_0, \end{aligned}$$

und entsprechend gilt $\Psi^{t,t_0}(x_0) = \Phi^{t,t_0}(x_0)$ für alle $(t_0, x_0) \in D$ und alle t mit $|t - t_0|$ hinreichend klein. QED

Abschließend führen wir noch den Begriff der *Stabilität* ein. Dieser gibt an, wie stark sich zwei Lösungen $x(t)$ und $y(t)$ derselben Differentialgleichung unterscheiden, wenn die Anfangswerte $x(t_0)$ und $y(t_0)$ nur wenig voneinander abweichen. Dabei interessieren wir uns für die Zukunft, d.h. nur für Werte $t \geq t_0$.

Definition 3.17 *Sei $D \subset \mathbb{R}^d$, $t_0 \in \mathbb{R}$. Die Funktion $f : [t_0, \infty] \times D$ erfüllt eine einseitige Lipschitz-Bedingung mit Konstante $L^+ = L^+(t) \in \mathbb{R}$, falls*

$$(x - y)^T (f(t, x) - f(t, y)) \leq L^+ \|x - y\|_2^2 \quad \forall x, y \in D$$

und für alle $t \in [t_0, \infty]$. Kann $L^+ \leq 0$ gewählt werden, so nennt man f und die zugehörige Differentialgleichung $x' = f(t, x)$ ***dissipativ****.*

Bemerkung: Aus globaler Lipschitzstetigkeit für $t \geq t_0$ folgt die einseitige Lipschitz-Bedingung.

Dieses zeigen wir im Folgenden.
Sei $\|f(t, x) - f(t, y)\| \leq L \cdot \|x - y\|$ für alle $x, y \in D$ und alle $t \geq t_0$. Dann gilt nach der Cauchy-Schwarzschen Ungleichung:

$$\begin{aligned} (x - y)^T (f(t, x) - f(t, y)) &\leq \|x - y\|_2 \cdot \|f(t, x) - f(t, y)\|_2 \\ &\leq L^+ \cdot \|x - y\|_2^2 \text{ mit } L^+ = L. \end{aligned}$$

Die Umkehrung gilt aber nicht, wie das folgende Beispiel zeigt.

Beispiel: $f(t,x) = -x$ erfüllt die einseitige Lipschitz-Bedingung mit $L^+ = -1$, denn

$$(x-y)(f(t,x) - f(t,y)) = (x-y)(y-x) = -(x-y)^2 = -\|x-y\|_2^2.$$

Dagegen ergibt die globale Lipschitz-Bedingung

$$|f(t,x) - f(t,y)| = |y-x| \leq L \cdot |y-x|,$$

gilt also nur für $L \geq 1$.

Satz 3.18 *Erfüllt $f : [0,\infty] \times D \to \mathbb{R}^d$ eine einseitige Lipschitz-Bedingung mit Konstante L^+, so gilt für die Evolution Φ von $x' = f(t,x)$:*

$$\|\Phi^{t,t_0}(x_0) - \Phi^{t,t_0}(y_0)\|_2 \leq e^{L^+(t-t_0)}\|x_0 - y_0\|_2.$$

Für dissipative Systeme gilt insbesondere, dass

$$\|\Phi^{t,t_0}(x_0) - \Phi^{t,t_0}(y_0)\|_2 \leq \|x_0 - y_0\|_2.$$

Beweis: siehe Übungen.

3.3 Einschritt-Verfahren

3.3.1 Grundlagen

Obwohl eine Lösung bei stetigen Eingangsdaten immer existiert, ist sie im Allgemeinen selbst bei skalaren Differentialgleichungen mit $d = 1$ nicht in geschlossener Form darstellbar. Meist ist f auch nur durch Messwerte gegeben.
Die Grundidee der numerischen Lösung von Anfangswertproblemen ist, die Lösung x näherungsweise an diskreten Punkten zu ermitteln:

> gesucht werden Näherungswerte an den gesuchten Vektor $x(t)$ für $t \in \Delta := \{t_0, t_1, \ldots, t_N\}$ mit $t_0 < t_1 < \ldots < t_N = T$ auf dem Intervall $[t_0, T]$.

Notation 3.19 *$\Delta := \{t_0, t_1, \ldots, t_N\}$ mit $t_0 < t_1 < \ldots < t_N = T$ heißt **Gitter** auf $[t_0, T]$. Die Werte $T_j := t_{j+1} - t_j$ nennt man **Schrittweiten**. Die **Feinheit des Gitters** ist gegeben durch*

$$\tau_\Delta := \max_{j=0,\ldots,N-1} T_j.$$

Gesucht ist dann eine **Gitterfunktion** $x_\Delta : \Delta \to \mathbb{R}^d$, welche die Lösung von $x'(t) = f(t, x(t))$, $x'(t_0) = x_0$ auf dem Gitter möglichst gut approximiert.

Bei **Einschritt-Verfahren** ermittelt man x_Δ durch eine Zwei-Term-Rekursion:

$$x_\Delta(t_j) \to x_\Delta(t_{j+1}),$$

das heißt in die Berechnung von $x_\Delta(t_{j+1})$ geht nur $x_\Delta(t_j)$ ein, keine Werte von t_i mit $i < j$. Dagegen gehen bei Mehr-Term-Rekursionen mehrere Werte in die Berechnung von $x_\Delta(t_{j+1})$ mit ein, genauer für $m \in \mathbb{N}$:

$$x_\Delta(t_j), \ldots, x_\Delta(t_{j-m}) \to x_\Delta(t_{j+1}).$$

Diese Rekursionen führen zu **Mehrschritt-Verfahren**.

Im Folgenden wird die Evolution Φ der Differentialgleichung durch eine **diskrete Evolution** Ψ ersetzt.

korrekte Evolution:	Approximation durch diskrete Evolution:
$x(t_{j+1}) = \Phi^{t_{j+1},t_j}(x(t_j))$	$x_\Delta(t_{j+1}) := \Psi^{t_{j+1},t_j}(x_\Delta(t_j))$
$x(t_0) = x_0$	$x_\Delta(t_0) := x_0$

3.3.2 Beispiele

Um Einschritt-Verfahren herzuleiten benutzt man die Integraldarstellung des Anfangswertproblems aus Lemma 3.8:

$$x(t_0 + \tau) = x_0 + \int_{t_0}^{t_0+\tau} f(t, x(t))dt. \tag{3.6}$$

Explizites Euler-Verfahren

Seien zunächst

$$t_j := t_0 + j \cdot \tau$$

äquidistante Gitterpunkte. Man approximiert $x(t_j)$ aus (3.6) nun iterativ wie folgt:

$$x(t_1) = x(t_0 + \tau) = x_0 + \int_{t_0}^{t_0+\tau} f(t, \underbrace{x(t)}_{\text{unbekannt}})dt,$$

Um das Integral abzuschätzen, verwendet man die Rechteck-Regel mit Funktionsauswertung am linken Randpunkt und erhält:

$$\int_{t_0}^{t_0+\tau} f(t, x(t))dt \approx \tau \cdot f(t_0, x_0).$$

Das ergibt

$$x(t_1) \approx x_0 + \tau \cdot f(t_0, x_0)$$

bzw. für unsere Approximationsfunktion

$$x_\Delta(t_1) = x_0 + \tau \cdot f(t_0, x_0).$$

Diese Formel ergibt sich alternativ auch aus dem Differenzenquotienten durch

$$\tfrac{x(t_0+\tau)-x(t_0)}{\tau} \approx x'(t_0) = f(t_0, x_0).$$

Ist nun $x(t_1)$ approximativ bekannt, erhält man

$$\begin{aligned} x(t_2) &= x(t_1) + \int_{t_1}^{t_1+\tau} f(t, x(t))dt \\ &\approx x_\Delta(t_1) + \tau \cdot f(t_1, x_\Delta(t_1)) =: x_\Delta(t_2) \end{aligned}$$

und rekursiv

$$x_\Delta(t_{j+1}) = x_\Delta(t_j) + \tau \cdot f(t_j, x_\Delta(t_j)).$$

Die diskrete Evolution ergibt sich entsprechend zu

$$\Psi^{t+\tau,t}_{\text{E-Euler}}(x) = x + \tau \cdot f(t, x).$$

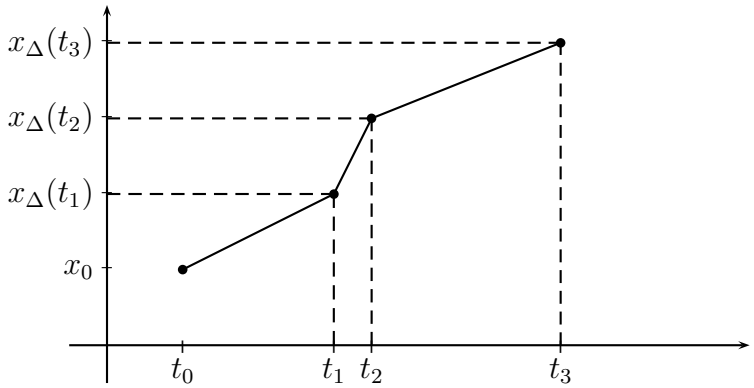
Etwas allgemeiner ist es mit $\tau_i := t_{i+1} - t_i$ nicht mehr nötig, äquidistante Stützstellen zu verwenden. Man erhält

$$x_\Delta(t_{j+1}) = \Psi^{t_{j+1},t_j}_{\text{E-Euler}}(x_\Delta(t_j)) := x_\Delta(t_j) + \tau_j \cdot f(t_j, x_\Delta(t_j)).$$

Interpretation:

> Um den Wert $x_\Delta(t_{j+1})$ an t_{j+1} zu bestimmen, verwendet man den Wert in $x_\Delta(t_j) + \tau_j \cdot x'(t_j, x_\Delta(t_j))$, also den Startwert und die Steigung an dem Ausgangspunkt $(t_j, x_\Delta(t_j))$.

Im skalaren Fall nennt man das explizite Euler-Verfahren daher auch Polygonzug-Verfahren.

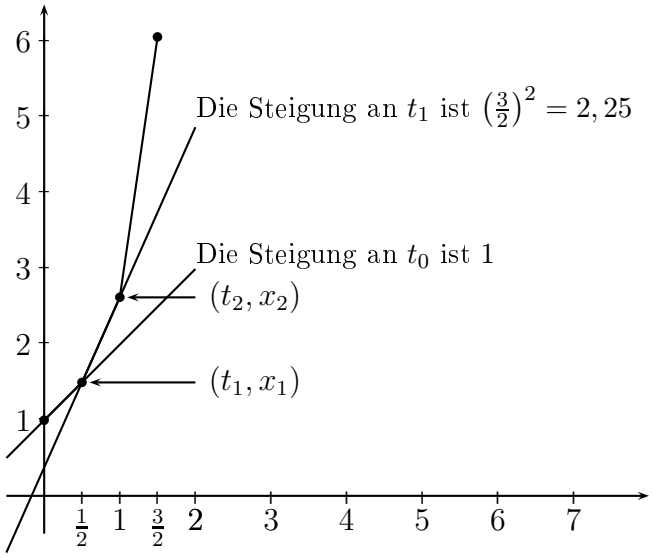


Beispiel: Sei folgendes Problem gegeben:

$$\begin{aligned} x'(t) &= (x(t))^2 \\ x(0) &= 1 \\ f(t,x) &= x^2 \\ \Delta &= \{0, \tfrac{1}{2}, 1, \tfrac{3}{2}\} \end{aligned}$$

Dann erhält man

$$\begin{aligned} x_\Delta(t_1) &= x_0 + \tfrac{1}{2} \cdot f(0, x_0) = 1 + \tfrac{1}{2} \cdot 1 = \tfrac{3}{2} \\ x_\Delta(t_2) &= x_\Delta(t_1) + \tfrac{1}{2} \cdot f(t_1, x_\Delta(t_1)) = \tfrac{3}{2} + \tfrac{1}{2} \cdot \tfrac{9}{4} = \tfrac{12}{8} + \tfrac{9}{8} = \tfrac{21}{8} = 2,625 \\ x_\Delta(t_3) &= x_\Delta(t_2) + \tfrac{1}{2} \cdot f(t_2, x_\Delta(t_2)) = \tfrac{21}{8} + \tfrac{1}{2} \cdot \left(\tfrac{21}{8}\right)^2 \approx 6,07. \end{aligned}$$



Implizites Euler-Verfahren

Das **implizite Euler-Verfahren** entsteht, wenn man das Integral durch die Rechteck-Regel am rechten Randpunkt approximiert:

$$\int_{t_j}^{t_j+\tau_j} f(t, x(t))dt \approx \tau_j \cdot f(t_j + \tau_j, x(t_j + \tau)).$$

Man erhält:

$$x_\Delta(t_{j+1}) = \Psi^{t_{j+1},t_j}_{\text{I-Euler}}(x_\Delta(t_j)) = \underbrace{x_\Delta(t_j)}_{\text{bekannt}} + \tau_j \cdot f(t_{j+1}, \underbrace{x_\Delta(t_{j+1})}_{\text{unbekannt}}).$$

Um $x_\Delta(t_{j+1})$ zu bestimmen, muss also ein (nichtlineares) Gleichungssystem mit d Unbekannten und d Gleichungen gelöst werden – und das in jedem Schritt!

Euler-Heun-Verfahren

Wählt man statt der Rechteck-Regel die Trapez-Regel zur Integralauswertung, so erhält man die Näherung:

$$\int_{t_j}^{t_j+\tau_j} f(t,x(t))dt \approx t_j \cdot \frac{f(t_j,x(t_j)) + f(t_j+\tau_j, x(t_j+\tau_j))}{2}$$

und es ergibt sich

$$\underbrace{x_\Delta(t_{j+1})}_{\text{unbekannt}} = \Psi^{t_{j+1},t_j}_{\text{E-Heun}}(x_\Delta(t_j)) := x_\Delta(t_j) + \frac{\tau_j}{2}(f(t_j,x_\Delta(t_j)) + f(t_{j+1}, \underbrace{x_\Delta(t_{j+1})}_{\text{unbekannt}})). \quad (3.7)$$

Auch dieses Verfahren ist implizit, weil in jedem Zeitschritt der Vektor $x_\Delta(t_{j+1})$ aus einem (nichtlinearen) Gleichungssystem ermittelt werden muss. In diesem Fall kann man dazu das Verfahren der sukzessiven Approximation benutzen:

Lemma 3.20 *Die Funktion $f(t,x)$ sei Lipschitzstetig bezüglich x mit Lipschitzkonstante L. Sei weiter $L \cdot \tau_j < 2$ für alle $j = 0,\ldots,N-1$. Dann lässt sich das Gleichungssystem (3.7) durch sukzessive Approximation*

$$x_\Delta^{(m+1)}(t_{j+1}) := x_\Delta(t_j) + \frac{\tau_j}{2}\left[f(t_j,x_\Delta(t_j)) + f(t_{j+1}, x_\Delta^{(m)}(t_{j+1}))\right], m \in \mathbb{N}_0$$

lösen.

Beweis: Die Fixpunktgleichung lautet $x = g(x)$ mit

$$g(x) = x_\Delta(t_j) + \frac{\tau_j}{2}\left[f(t_j,x_\Delta(t_j)) + f(t_{j+1},x)\right]$$

in jedem Schritt j. Wir müssen nachweisen, dass g eine Kontraktion ist. Es gilt:

$$\begin{aligned}\|g(x) - g(\tilde{x})\|_2 &= \frac{\tau_j}{2}\|f(t_{j+1},x) - f(t_{j+1},\tilde{x})\|_2 \leq \frac{\tau_j}{2} \cdot L \cdot \|x-\tilde{x}\|_2 \\ &= q \cdot \|x-\tilde{x}\|_2 \quad \text{mit } q = \frac{\tau_j}{2} \cdot L < 1.\end{aligned}$$

QED

Prädiktor-Korrektor Variante von Euler-Heun

Hier kombiniert man das explizite Euler-Verfahren und das Euler-Heun Verfahren mit jeweils einem Iterationsschritt der sukzessiven Approximation nach Lemma 3.20 wie folgt:

Im j-ten Schritt:

- bestimme den Startwert (Prädiktor) nach E-Euler:

$$\tilde{x}_\Delta(t_{j+1}) := x_\Delta(t_j) + \tau_j \cdot f(t_j, x_\Delta(t_j)) \qquad \text{(Prädiktor)}$$

- Wähle $\tilde{x}_\Delta(t_{j+1})$ als Startwert für die sukzessive Approximation von Euler-Heun und führe darin genau einen Schritt der sukzessiven Approximation nach Lemma 3.20 aus (Korrektor-Schritt):

$$x_\Delta(t_{j+1}) = \Psi^{t_{j+1},t_j}_{\text{Pre-Kor-V}}(x_\Delta(t_j)) = x_\Delta(t_j) + \frac{\tau_j}{2}\left[f(t_j, x_\Delta(t_j)) + f(t_{j+1}, \underbrace{\tilde{x}_\Delta(t_{j+1})}_{\text{aus (Prädiktor)}})\right].$$

Das Verfahren erreicht gewöhnlich eine höhere Genauigkeit als das explizite Euler-Verfahren.

3.3.3 Konsistenz und Eindeutigkeit

Wir untersuchen nun das Konvergenzverhalten von Einschritt-Verfahren theoretisch. Dazu fordern wir zunächst die ersten beiden der drei Eigenschaften einer Evolution (aus Lemma 3.16) auch für die diskrete Evolution Ψ.

Definition 3.21 *Eine diskrete Evolution Ψ heißt **konsistent** zur Differentialgleichung $x' = f(t,x)$, falls für alle $(t_0, x_0) \in D$ gilt:*

$$\Psi^{t_0,t_0}(x_0) = x_0 \tag{3.8}$$

$$\textit{und} \qquad \frac{d}{d\tau}\Psi^{t_0+\tau,t_0}(x_0)|_{\tau=0} = f(t_0, x_0). \tag{3.9}$$

*Ein Einschritt-Verfahren heißt **konsistent**, falls es jeder hinreichend glatten Funktion f eine konsistente diskrete Evolution $\Psi[f]$ zuordnet.*

Zwei äquivalente Konsistenzkriterien sind die folgenden.

Lemma 3.22 *Die diskrete Evolution $\Psi^{t_0+\tau,t_0}(x_0)$ sei für alle $(t_0, x_0) \in D$ und hinreichend kleines τ differenzierbar. Dann sind die folgenden Aussagen äquivalent:*

(i) *Ψ ist konsistent.*

(ii) *Es gibt eine bezüglich τ stetige Verfahrensfunktion $\phi = \phi(t_0, x_0, \tau)$ mit den Eigenschaften:*

$$\Psi^{t_0+\tau,t_0}(x_0) = x_0 + \tau \cdot \phi(t_0, x_0, \tau) \tag{3.10}$$

$$\phi(t_0, x_0, 0) = f(t_0, x_0) \tag{3.11}$$

(iii) Es gilt:

$$\lim_{\tau\to 0}\frac{1}{\tau}\left\|\Psi^{t_0+\tau,t_0}(x_0)-\Phi^{t_0+\tau,t_0}(x_0)\right\|=0. \tag{3.12}$$

Beweis:

$(i)\Longrightarrow(ii)$**:** Sei Ψ konsistent. Definiere

$$\phi(t_0,x_0,\tau):=\begin{cases}\frac{1}{\tau}\left(\Psi^{t_0+\tau,t_0}(x_0)-x_0\right) & \text{falls } \tau\neq 0.\\ f(t_0,x_0) & \text{falls } \tau=0\end{cases}$$

Dann sind (3.10) und (3.11) direkt erfüllt und es muss nur die Stetigkeit von ϕ gezeigt werden. Dazu betrachten wir

$$\begin{aligned}\lim_{\tau\to 0}\frac{1}{\tau}\left(\Psi^{t_0+\tau,t_0}(x_0)-x_0\right) &= \lim_{\tau\to 0}\frac{\Psi^{t_0+\tau,t_0}(x_0)-\Psi^{t_0,t_0}(x_0)}{\tau}, \text{ wegen (3.8)}\\ &= \tfrac{\partial}{\partial\tau}\Psi^{t_0+\tau,t_0}(x_0)|_{\tau=0}, \text{ wegen (3.9)}\\ &= f(t_0,x_0),\end{aligned}$$

also ist ϕ stetig.

$(ii)\Longrightarrow(iii)$**:** Sei ϕ eine Verfahrensfunktion, die (3.10) und (3.11) erfüllt. Dann gilt

$$\begin{aligned}&\lim_{\tau\to 0}\frac{1}{\tau}\|\Psi^{t_0+\tau,t_0}(x_0)-\Phi^{t_0+\tau,t_0}(x_0)\|\\ &=\lim_{\tau\to 0}\left\|\frac{\Psi^{t_0+\tau,t_0}(x_0)-x_0}{\tau}-\frac{\Phi^{t_0+\tau,t_0}(x_0)-x_0}{\tau}\right\|\\ &=\|\phi(t_0,x_0,0)-f(t_0,x_0)\| \text{ wegen (3.10) und [Ev2] im Lemma 3.16}\\ &=0 \text{ wegen (3.11)}\end{aligned}$$

$(iii)\Longrightarrow(i)$**:** Sei nun (3.12) erfüllt. Eine Taylorentwicklung bis zum Grad 1 liefert wegen [Ev2]

$$\Phi^{t_0+\tau,t_0}(x_0)=x_0+\tau f(t_0,x_0)+o(\tau) \text{ für } \tau\to 0.$$

Weiter ist Ψ nach Voraussetzung für hinreichend kleines τ differenzierbar bezüglich τ. Das ergibt

$$\Psi^{t_0+\tau,t_0}(x_0)=\Psi^{t_0,t_0}(x_0)+\tau\tfrac{\partial}{\partial\tau}\Psi^{t_0+\tau,t_0}(x_0)|_{\tau=0}+o(\tau) \text{ für } \tau\to 0.$$

Für $\tau\to 0$ sind die linken Seiten dieser beiden Gleichungen wegen (3.12) gleich, also auch die rechten Seiten und durch einen Koeffizientenvergleich folgt $x_0=\Psi^{t_0,t_0}(x_0)$ und $f(t_0,x_0)=\frac{\partial}{\partial\tau}\Psi^{t_0+\tau,t_0}|_{\tau=0}$; (3.8) und (3.9) gelten also und Ψ ist konsistent. QED

Ist eine diskrete Evolution konsistent, so ist der lokale Fehler, den wir in jedem Schritt bei der Berechnung der Gitterfunktion machen, klein. Interessanter ist aber der globale Fehler

$$\max_{t\in\Delta} \|x_\Delta(t) - x(t)\|,$$

der möglichst klein sein soll – zumindest wenn das Gitter Δ fein genug ist.

Notation 3.23 *Ein Einschritt-Verfahren heißt* ***konvergent****, falls*

$$\lim_{\tau\to 0} \sup_{\Delta:\tau_\Delta=\tau} \max_{t\in\Delta} \|x_\Delta(t) - x(t)\| = 0$$

Dabei bezeichnet $\tau_\Delta = \max_{j=0,\dots,N-1} t_{j+1} - t_j$ wie schon zu Beginn des Abschnittes 3.3.1 die Feinheit des Gitters $\Delta = \{t_0, \dots, t_N\}$.

Der folgende Satz zeigt, dass aus der Konsistenz unter einer zusätzlichen Stabilitätsannahme die Konvergenz von Einschritt-Verfahren folgt. Dabei müssen wir die Konsistenzbedingung allerdings verstärken: Wir verwenden (3.12) und verlangen, dass die Bedingung gleichmäßig erfüllt ist, also für alle $x(t)$ auf der Lösungskurve.

Satz 3.24 *Die diskrete Evolution Ψ sei in einer Umgebung U der Trajektorie $\{(t, x(t)) : t \in [t_0, T]\}$ definiert und genüge den folgenden Bedingungen.*

Stabilitätsbedingung: *Es gibt Konstanten $L_\Psi \geq 0$ und $\tau_0 > 0$ so, dass*

$$\|\Psi^{t+\tau,t}(x_1) - \Psi^{t+\tau,t}(x_2)\| \leq e^{L_\Psi \tau}\|x_1 - x_2\|$$

für alle $(t, x_1), (t, x_2) \in U$ und alle $0 \leq \tau \leq \tau_0$.

Konsistenzbedingung: *Es gibt eine monoton wachsende Funktion $\mathrm{err} : [0, \tau_0] \to [0, \infty)$ mit $\lim_{\tau\to 0} \mathrm{err}(\tau) = 0$ so, dass*

$$\|\Phi^{t+\tau,t}(x(t)) - \Psi^{t+\tau,t}(x(t))\| \leq \tau\ \mathrm{err}(\tau)$$

für alle $t \in [0, T]$.

Dann gibt es ein $\tau_1 \in [0, \tau_0]$ so, dass für jedes Gitter $\Delta = \{t_0, \dots, t_N\}$ auf $[t_0, T]$ mit Feinheit $\tau_\Delta \leq \tau_1$ die Gitterfunktion x_Δ durch die diskrete Evolution

$$x_\Delta(t_{j+1}) = \Psi^{t_{j+1},t_j}(x_\Delta(t_j)), \qquad x_\Delta(t_0) = x_0$$

wohldefiniert ist, und der Fehler für alle $t \in \Delta$ der Abschätzung

$$\|x_\Delta(t) - x(t)\| \leq r(\tau_\Delta) := \begin{cases} \mathrm{err}(\tau_\Delta)\frac{e^{L_\Psi(t-t_0)}-1}{L_\Psi} & \text{falls } L_\Psi > 0 \\ \mathrm{err}(\tau_\Delta)(t - t_0) & \text{falls } L_\Psi = 0 \end{cases}$$

genügt.

Der Satz sagt auf abstrakter Ebene, dass Konsistenz und Stabilität zusammen Konvergenz ergeben.

Beweis: Wir wählen τ_1 so klein, dass für alle $t \in [0,T]$ und für alle $x_1 \in \mathbb{R}^d$ gilt:

$$\|x_1 - x(t)\| \leq r(\tau_1) \Longrightarrow (t, x_1) \in U.$$

Sei Δ ein beliebiges Gitter mit $\tau_\Delta \leq \tau_1$. Wir möchten nachweisen, dass die Abschätzung

$$\|x_\Delta(t) - x(t)\| \leq r(\tau_\Delta)$$

für alle $t_0, t_1, \ldots, t_N$ des Gitters Δ erfüllt ist.
Insbesondere gilt dann $\|x_\Delta(t) - x(t)\| \leq r(\tau_1)$, woraus wir wegen der Definition von τ_1 folgern, dass $(t_j, x_\Delta(t_j)) \in U$. Entsprechend kann man also $x_\Delta(t_{j+1}) = \Psi^{t_{j+1},t_j}(x_\Delta(t_j))$ berechnen und $x_\Delta(t_j)$ ist wohldefiniert.

Zum Nachweis der Abschätzung verwenden wir Induktion nach j, gehen also der Reihe nach alle Punkte $t_0, t_1, \ldots, t_N$ des Gitters Δ durch.
Für $j = 0$ gilt $x_\Delta(t_0) = x_0 = x(t_0)$, die Abschätzung gilt also wegen $r(\tau_\Delta) \geq 0$.
Sei nun $\|x_\Delta(t_{j'}) - x(t_{j'})\| \leq r(\tau_\Delta)$ für alle $j' \leq j$ erfüllt. Wir betrachten t_{j+1}. Dazu unterscheiden wir zwei Fälle.

Fall 1: Sei $L_\Psi > 0$. Dann gilt

$$\begin{aligned}
& \|x_\Delta(t_{j+1}) - x(t_{j+1})\| \\
= \; & \|\Psi^{t_{j+1},t_j}(x_\Delta(t_j)) - \Phi^{t_{j+1},t_j}(x_\Delta(t_j))\| \\
\leq \; & \|\Psi^{t_{j+1},t_j}(x_\Delta(t_j)) - \Psi^{t_{j+1},t_j}(x(t_j))\| + \|\Psi^{t_{j+1},t_j}(x(t_j)) - \Phi^{t_{j+1},t_j}(x(t_j))\| \\
\leq \; & e^{L_\Psi(t_{j+1}-t_j)}\|x_\Delta(t_j) - x(t_j)\| + (t_{j+1} - t_j)\mathrm{err}(\tau_\Delta) \text{ wegen Stabilität und Konsistenz} \\
\leq \; & \frac{\mathrm{err}(\tau_\Delta)}{L_\Psi}\left(e^{L_\Psi(t_{j+1}-t_j)}\left(e^{L_\Psi(t_j-t_0)} - 1\right) + L_\Psi(t_{j+1} - t_j)\right) \text{ Induktionsannahme} \\
= \; & \frac{\mathrm{err}(\tau_\Delta)}{L_\Psi}\left(e^{L_\Psi(t_{j+1}-t_0)} \underbrace{-e^{L_\Psi(t_{j+1}-t_j)} + L_\Psi(t_{j+1} - t_j)}_{\leq -1, \text{ denn } e^a \geq a+1}\right) \\
\leq \; & \frac{\mathrm{err}(\tau_\Delta)}{L_\Psi}\left(e^{L_\Psi(t_{j+1}-t_0)} - 1\right) = r(\tau_\Delta).
\end{aligned}$$

Fall 2: Sei $L_\Psi = 0$. Dann geht man vor wie oben, allerdings ergibt die Induktionsvoraussetzung, dass

$$\begin{aligned}
\|x_\Delta(t_{j+1}) - x(t_{j+1})\| &\leq \mathrm{err}(\tau_\Delta)(t_j - t_0) + \mathrm{err}(\tau_\Delta)(t_{j+1} - t_j) \\
&= \mathrm{err}(\tau_\Delta)(t_{j+1} - t_0)
\end{aligned}$$

QED

Ein weiterer Begriff ist die Konsistenzordnung, welche hilft, die Konvergenzgeschwindigkeit eines Einschritt-Verfahrens abzuschätzen.

Definition 3.25

- *Eine diskrete Evolution Ψ für eine Differentialgleichung $x'(t) = f(t, x(t))$, $f : D \to \mathbb{R}^d$, besitzt die* ***Konsistenzordnung*** *$p > 0$, falls es für jede kompakte Teilmenge $K \subseteq D$ eine Konstante $C > 0$ so gibt, dass*

  $$\|\Psi^{t+\tau,t}(x) - \Phi^{t+\tau,t}(x)\| \leq C \cdot \tau^{p+1}$$

  für alle $(t, x) \in K$ und alle hinreichend kleinen $\tau \geq 0$.

- *Ein Einschritt-Verfahren besitzt die Konsistenzordnung $p > 0$, falls für jede rechte Seite $f \in C^\infty(D, \mathbb{R}^d)$ die zugeordnete diskrete Evolution $\Psi = \Psi[f]$ die Konsistenzordnung p besitzt.*

- *Ein Einschritt-Verfahren besitzt die Konvergenzordnung $p > 0$, falls für jede Lösung $x : [t_0, T] \to \mathbb{R}^d$ eines Anfangswertproblemes mit rechter Seite $f \in C^\infty(D, \mathbb{R}^d)$ der globale Fehler der durch das Verfahren bestimmten Lösung x_Δ auf einem Gitter Δ mit hinreichend kleiner Gitterfeinheit τ_Δ die Abschätzung*

  $$\max_{t \in \Delta} \|x_\Delta(t) - x(t)\| \leq \tilde{C} \cdot {\tau_\Delta}^p$$

  erfüllt, wobei $\tilde{C}$ nicht von Δ abhängt.

Lemma 3.26 *Besitzt ein Einschritt-Verfahren die Konsistenzordnung p und erfüllt es die Stabilitätsbedingung aus Satz 3.24, so besitzt es die Konvergenzordnung p.*

Beweis: Sei $f \in C^\infty(D, \mathbb{R}^d)$ beliebig. Weil das Verfahren die Konsistenzordnung p hat, gilt für die diskrete Evolution Ψ, dass

$$\|\Psi^{t+\tau,t}(x) - \Phi^{t+\tau,t}(x)\| \leq C \cdot \tau^{p+1}.$$

Die Funktion $\mathrm{err}(\tau) := C \cdot \tau^p$ erfüllt dann wegen $\lim_{\tau \to 0} C \cdot \tau^p = 0$ die Konsistenzbedingung aus Satz 3.24. Wir können also Satz 3.24 anwenden und erhalten

$$\|x_\Delta(t) - x(t)\| \leq r(\tau_\Delta) = \begin{cases} \mathrm{err}(\tau_\Delta) \cdot \frac{e^{L_\Psi (t-t_0)} - 1}{L_\Psi} & L_\Psi > 0 \\ \mathrm{err}(\tau_\Delta) \cdot (t - t_0) & L_\Psi = 0. \end{cases}$$

Es folgt nun

$$\max_{t \in \Delta} \|x_\Delta(t) - x(t)\| \leq \tilde{C} \cdot {\tau_\Delta}^p \text{ mit } \tilde{C} = \begin{cases} C \cdot \frac{e^{L_\Psi (T-t_0)} - 1}{L_\Psi} & L_\Psi > 0 \\ C \cdot (T - t_0) & L_\Psi = 0 \end{cases}.$$

QED

Satz 3.27 *Die diskrete Evolution des expliziten Euler-Verfahrens ist für stetig differenzierbare Seiten f konsistent von der Ordnung 1.*

Beweis: Übung.

3.3.4 Explizite Runge-Kutta-Verfahren

Euler-Verfahren

Approximiere das Integral durch die Rechteckregel, d.h.

$$\int_t^{t+\tau} f(s, \underbrace{\Phi^{s,t}(x)}_{x(s)})ds \approx \tau \cdot f(t,x).$$

Dabei ist der Fehler nach Satz 3.27 von der Größe $O(\tau^2)$.

Verfahren von Runge (explizite Mittelpunktregel)

Die Idee ist, dass man eine Quadraturformel höherer Ordnung verwendet, zum Beispiel die Mittelpunktregel:

$$\int_t^{t+\tau} f(s, \Phi^{s,t}(x))ds \approx \tau \cdot f(t+\tfrac{\tau}{2}, \Phi^{t+\frac{\tau}{2},t}(x))$$

mit einem Fehler von $O(\tau^3)$. Allerdings kennen wir den Wert $\Phi^{t+\frac{\tau}{2},t}(x)$ nicht. Es reicht aber, ihn mit einer Genauigkeit von $O(\tau^2)$ auszuwerten, weil er noch mit τ multipliziert wird. Dazu verwendet man

$$\Phi^{t+\frac{\tau}{2},t}(x) = x + \tfrac{\tau}{2} f(t,x)$$

nach dem Euler-Verfahren mit $O(\tau^2)$. Man erhält

$$\Psi^{t+\tau,t}(x) = x + \tau \cdot f(t+\tfrac{\tau}{2}, x + \tfrac{\tau}{2} f(t,x))$$

oder, algorithmisch:

$$\begin{aligned} k_1 &:= f(t,x) \\ k_2 &:= f(t+\tfrac{\tau}{2}, x + \tfrac{\tau}{2} \cdot k_1) \\ \Psi^{t+\tau,t}(x) &:= x + \tau \cdot k_2 \end{aligned}$$

Dieses Verfahren hat Konsistenzordnung 2.

Runge-Kutta-Verfahren

Seien

$$k_i = k_i(t,x,\tau) = f\left(t + c_i\tau, x + \tau \sum_{j=1}^{i-1} a_{ij} k_j\right), \qquad \text{für } i = 1, \ldots, s$$

$$\Psi^{t+\tau,t}(x) = x + \tau \sum_{j=1}^{s} b_j k_j(t,x,\tau) = x + \tau \sum_{j=1}^{s} b_j k_j.$$

k_i heißt die i-te Stufe des Runge-Kutta-Verfahrens. Man benutzt folgende Notation:

$$A = \begin{pmatrix} 0 & & & & 0 \\ a_{21} & 0 & & & \\ a_{31} & a_{32} & 0 & & \\ \vdots & \vdots & \ddots & \ddots & \\ a_{s1} & \cdots & \cdots & a_{s,s-1} & 0 \end{pmatrix}, \qquad b = \begin{pmatrix} b_1 \\ \vdots \\ b_s \end{pmatrix}, \qquad c = \begin{pmatrix} c_1 \\ \vdots \\ c_s \end{pmatrix}.$$

Mit der Vereinbarung, dass $a_{ij} := 0$ für $j \geq i$ ist, vereinfachen wir die Summenschreibweise. Wir erhalten

$$k_i = f\left(t + c_i\tau, x + \tau \sum_{j=1}^{s} a_{ij} k_j\right), \qquad i = 1, \ldots, s.$$

Dabei heißt s die **Stufenzahl** des Runge-Kutta-Verfahrens und beschreibt die Tiefe der Schachtelungen von f-Auswertungen. Man gibt ein Verfahren oft durch folgendes **Butcher-Schema** an:

$$\begin{array}{c|c} c & A \\ \hline & b^t \end{array}$$

Beispiele:

1. Explizites Euler-Verfahren:

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

   also

$$\begin{aligned} k_1 &= f(t + c_1\tau, x + 0) \\ &= f(t + 0, x + 0) = f(t, x) \end{aligned}$$

   und

$$\Psi^{t+\tau,t}(x) = x + \tau b_1 k_1 = x + \tau f(t, x).$$

2. Verfahren von Runge:

$$\begin{array}{c|cc} 0 & 0 & \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array}$$

3. „Klassisches“ Runge-Kutta-Verfahren der Ordnung 4:

$$\begin{array}{c|cccc} 0 & 0 & & & \\ \frac{1}{2} & \frac{1}{2} & 0 & & \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

Ausführliche Notation:

$$\begin{aligned} k_1 &:= f(t,x) \\ k_2 &:= f(t+\tfrac{1}{2}\tau, x+\tfrac{1}{2}\tau k_1) \\ k_3 &:= f(t+\tfrac{1}{2}\tau, x+\tfrac{1}{2}\tau k_2) \\ k_4 &:= f(t+\tau, x+\tau k_3) \\ \Psi^{t+\tau,t}(x) &:= x+\tau(\tfrac{1}{6}k_1+\tfrac{1}{3}k_2+\tfrac{1}{3}k_3+\tfrac{1}{6}k_4) \end{aligned}$$

Lemma 3.28 *Ein Runge-Kutta-Verfahren (A,b,c) ist genau dann konsistent für alle $f \in C(D,\mathbb{R}^d)$, falls*

$$\sum_{j=1}^{s} b_j = 1.$$

Beweis: Wir benutzen die beiden Bedingungen (3.10) und (3.11) aus Lemma 3.22 und definieren

$$\phi(t,x,\tau) := \sum_{j=1}^{s} b_j k_j(t,x,\tau).$$

Dann gilt (3.10), denn:

$$\begin{aligned} \Psi^{t+\tau,t}(x) &= x+\tau\sum_{j=1}^{s} b_j k_j(t,x,\tau) \\ &= x+\phi(t,x,\tau). \end{aligned}$$

Weiterhin gilt $k_j(t,x,0) = f(t,x)$ für alle j, also

$$\phi(t,x,0) = \sum_{j=1}^{s} b_j k_j(t,x,0) = f(t,x)\sum_{j=1}^{s} b_j.$$

Da die Bedingung (3.11) $\phi(t,x,0) = f(t,x)$ fordert, ist (3.11) genau dann erfüllt, wenn $\sum_{j=1}^{s} b_j = 1$ gilt. QED

Lemma 3.29 *Besitzt ein s-stufiges Runge-Kutta-Verfahren für alle $f \in C^\infty(D,\mathbb{R}^d)$ die Konsistenzordnung p, so gilt $p \le s$.*

Beweis: Betrachte (AWP)

$$x'(t) = x(t), \qquad x(0) = 1.$$

Die Lösung ist

$$\Phi^{\tau,0}(1) = e^\tau 1 = \tau + \tfrac{1}{2!}\tau^2 + \cdots + \tfrac{1}{p!}\tau^p + \mathcal{O}(\tau^{p+1}).$$

Für die Konsistenzordnung wollen wir $\Phi^{t+\tau,t}(x)$ mit $\Psi^{t+\tau,t}(x)$ vergleichen – für $f(t,x) = x(t)$ und $t = 0$, $x = 1$. Die echte Evolution Φ kennen wir schon. Um auch $\Psi^{t+\tau,t}(x)$ zu verstehen, betrachten wir die k_j.

Behauptung: $k_j(0,1,\tau)$ ist ein Polynom in τ vom Grad $\leq j-1$, also $k_j \in \Pi_{j-1}$.

Vollständige Induktion über j:

- $j=1$: $k(0,1,\tau) = f(t+c_1\tau, x) = x \in \Pi_0$, da konstant in τ.
- $j \mapsto j+1$:

$$k_{j+1}(0,1,\tau) = f\left(t + c_j\tau, x + \tau \sum_{l=1}^{j} a_{jl}k_l\right)$$
$$= x + \tau \underbrace{\sum_{l=1}^{j} a_{jl}k_l}_{\in \Pi_{j-1}} \in \Pi_j, \text{ da } k_l \in \Pi_{l-1} \text{ nach der Induktionsannahme.}$$

Also ist $\Psi^{\tau,0}(1) \in \Pi(s)$. Damit erhalten wir:

$$\| \underbrace{\Psi^{\tau,0}(1)}_{\in \Pi_s} - \underbrace{\Phi^{\tau,0}(1)}_{1+\tau+\cdots+\frac{1}{s!}\tau^s+\mathcal{O}(\tau^{s+1})} \| \leq c\,\tau^{s+1}.$$

Folglich kann die Konsistenzordnung höchstens s sein. QED

Bei der Konstruktion vom Runge-Kutta-Verfahren hat man also zunächst viele Wahlmöglichkeiten. Wir stellen aber die folgenden Bedingungen an das Verfahren:

1. Invarianz gegen Autonomisierung und
2. Konsistenzordnung p für vorgegebenes p.

Diese Bedingungen formulieren wir im Folgenden als Bedingungen an die Koeffizienten (A,b,c) des Verfahrens. Wir betrachten zuerst die Invarianz gegen Autonomisierung.
Seien $x'(t) = f(t,x(t))$ ein (AWP) im $\mathbb{R}^d$ und $x(t_0) = x_0$. Nach Lemma 3.6 lässt sich das in ein äquivalentes System im $\mathbb{R}^{d+1}$, nämlich in

$$\widehat{(\text{AWP})} \qquad y'(t) = \begin{pmatrix} 1 \\ f(y(t)) \end{pmatrix}, \qquad y^{(t_0)} = y_0 := \begin{pmatrix} t_0 \\ x_0 \end{pmatrix}$$

umwandeln. Dabei gelten die folgenden Aussagen:

- Ist x Lösung von (AWP), so ist $(t, x(t))^T$ eine Lösung von $\widehat{(\text{AWP})}$.
- Ist $(s,x)^T$ eine Lösung von $\widehat{(\text{AWP})}$, so folgt $s(t) = t$ und x ist Lösung von (AWP).

Formal lässt sich die Äquivalenz der beiden Anfangswertprobleme durch die Evolution $\hat{\Phi}$ von $y' = (1, f(y))^T$ und Φ von $x' = f(t,x)$ folgendermaßen schreiben:

$$\begin{pmatrix} t+\tau \\ \Phi^{t+\tau,t}(x) \end{pmatrix} = \hat{\Phi}^{t+\tau,t} \begin{pmatrix} t \\ x \end{pmatrix}.$$

Diese Eigenschaft soll dann auch für diskrete Evolutionen Ψ und $\hat{\Psi}$ gelten; sie soll also gewissermaßen vererbt werden. Für die Evolution Ψ (bzw. $\hat{\Psi}$ für das erweiterte System) bedeutet

$$\begin{pmatrix} t+\tau \\ \Psi^{t+\tau,t}(x) \end{pmatrix} = \hat{\Psi}^{t+\tau,t} \begin{pmatrix} t \\ x \end{pmatrix} \tag{3.13}$$

dass man das gleiche Ergebnis erhält, egal, ob man ein durch Ψ gegebenes Einschritt-Verfahren direkt auf die gegebene Differentialgleichung anwendet, oder ob man das gleiche Verfahren mittels $\hat{\Psi}$ auf die autonomisierte Differentialgleichung anwendet. Man nennt das Verfahren dann **invariant gegenüber Autonomisierung**.

Lemma 3.30 *Ein explizites Runge-Kutta Verfahren ist genau dann invariant gegen Autonomisierung, wenn es konsistent ist und es*

$$c_i = \sum_{j=1}^{s} a_{ij} \text{ für } j = 1, \ldots, s$$

erfüllt.

Beweis: Sei $y' = \hat{f}(y(t))$ die autonomisierte Differentialgleichung mit

$$\hat{f}\left(\begin{pmatrix} t \\ x \end{pmatrix}\right) = \left(\begin{pmatrix} 1 \\ f(t,x) \end{pmatrix}\right),$$

wobei $y(t) = \left(\begin{pmatrix} t \\ x(t) \end{pmatrix}\right)$ und $\hat{f}$ autonom ist, also $\hat{f}(t, y(t)) = \hat{f}(y(t))$ gilt. Bezeichnen wir nun mit $\hat{K}_i = \left(\begin{pmatrix} \hat{l}_i \\ \hat{k}_i \end{pmatrix}\right)$, $i = 1, \ldots, s$ die Stufen von $\hat{\Psi}$, so gilt:

$$\begin{aligned}
\hat{K}_i &= \hat{f}(t + c_i\tau, y + \tau \sum_{j=1}^{s} a_{ij}\hat{K}_j), \\
&= \hat{f}(y + \tau \sum_{j=1}^{s} a_{ij}\hat{K}_j) \\
&= \hat{f}\left(\begin{pmatrix} t \\ x \end{pmatrix} + \tau \sum_{j=1}^{s} a_{ij} \begin{pmatrix} \hat{l}_j \\ \hat{k}_j \end{pmatrix}\right) \\
&= \begin{pmatrix} 1 \\ f(t + \tau \sum_{j=1}^{s} a_{ij}\hat{l}_j, x + \tau \sum_{j=1}^{s} \hat{k}_j) \end{pmatrix} \quad i = 1, \ldots, s,
\end{aligned}$$

das heißt, $\hat{l}_i = 1$ und $\hat{k}_i = f(t + \tau \sum_{j=1}^{s} a_{ij}\hat{l}_j, x + \tau \sum_{j=1}^{s} a_{ij}\hat{k}_j)$ für $i = 1, \ldots, s$. Für ein Runge-Kutta Verfahren gilt weiter für die diskrete Evolution, dass

$$\hat{\Psi}^{t+\tau,t}\left(\begin{pmatrix} t \\ x \end{pmatrix}\right) = \left(\begin{pmatrix} t \\ x \end{pmatrix}\right) + \tau \sum_{j=1}^{s} b_j \left(\begin{pmatrix} \hat{l}_j \\ \hat{k}_j \end{pmatrix}\right) = \left(\begin{pmatrix} t + \tau \sum_{j=1}^{s} b_j \\ x + \tau \sum_{j=1}^{s} b_j \hat{k}_j \end{pmatrix}\right).$$

Nach (3.13) ist das Verfahren invariant gegen Autonomisierung genau dann wenn

$$\begin{aligned}
&\left(\begin{pmatrix} t+\tau \\ \Psi^{t+\tau,t}(x) \end{pmatrix}\right) = \hat{\Psi}^{t+\tau,t}\left(\begin{pmatrix} t \\ x \end{pmatrix}\right) \\
\iff\ & t + \tau = t + \tau \sum_{j=1}^{s} b_j \text{ und } \Psi^{t+\tau,t}(x) = x + \tau \sum_{j=1}^{s} b_j \hat{k}_j \\
\iff\ & \sum_{j=1}^{s} b_j = 1 \text{ und } x + \tau \sum_{j=1}^{s} b_j k_j = x + \tau \sum_{j=1}^{s} b_j \hat{k}_j \\
\iff\ & \text{konsistent und } k_i = \hat{k}_i \text{ für alle } i = 1, \ldots, s.
\end{aligned}$$

Letzteres ist genau dann der Fall, wenn

$$f\left(t + c_i\tau, x + \tau \sum_{j=1}^{s} a_{ij} k_j\right) = f\left(t + \tau \sum_{j=1}^{s} a_{ij}, x + \tau \sum_{j=1}^{s} a_{ij} k_j\right),$$

also genau dann wenn $c_i = \sum_{j=1}^{s} a_{ij}$. QED

Gegen Autonomisierung invariante Runge-Kutta Verfahren bezeichnen wir kurz mit (A, b) und wir schreiben dann auch $\Psi^{\tau}(x) = \Psi^{t+\tau,t}(x)$, da man c von der Matrix A abhängig ist.

Folgende Bedingungen an die Koeffizienten eines Runge-Kutta Verfahrens haben wir bisher erarbeitet:

- Das Verfahren ist genau dann konsistent, wenn $\sum_{i=1}^{s} b_i = 1$, und
- es ist genau dann invariant gegen Autonomisierung, wenn es konsistent ist und $c_i = \sum_{j=1}^{2} a_{ij}$.

Wir wollen nun den ersten der beiden Punkte verallgemeinern und für gegen Autonomisierung invariante Runge-Kutta-Verfahren genauere Forderungen an die Konsistenzordnung stellen. Diese werden als *Ordnungsbedingungen* bezeichnet.

Satz 3.31 *Ein autonomisierungsinvariantes Runge-Kutta-Verfahren besitzt für jede Differentialgleichung mit p-mal stetig differenzierbarer rechter Seite f die Konsistenzordnung*

- $p = 1$, *falls* $\sum_{i=1}^{s} b_i = 1$.
- $p = 2$, *falls zusätzlich* $\sum_{i=1}^{s} b_i c_i = \frac{1}{2}$.
- $p = 3$, *falls zusätzlich* $\sum_{i=1}^{s} b_i c_i^2 = \frac{1}{3}$ *und* $\sum_{i,j=1}^{s} b_i a_{ij} c_j = \frac{1}{6}$.
- $p = 4$, *falls zusätzlich* $\begin{array}{ll} \sum_{i=1}^{s} b_i c_i^3 = \frac{1}{4} & \sum_{i,j=1}^{s} b_i c_i a_{ij} c_j = \frac{1}{8} \\ \sum_{i,j=1}^{s} b_i a_{ij} c_j^2 = \frac{1}{12} & \sum_{i,j,k=1}^{s} b_i a_{ij} a_{jk} c_k = \frac{1}{24} \end{array}$.

Beweis: Wir geben nur die Grundstruktur des Beweises an: Das Ziel besteht darin, zu zeigen, dass

$$\|\Psi^\tau(x) - \Phi^\tau(x)\| = \mathrm{O}(\tau^{p+1}) \text{ für } \tau \to 0.$$

Dazu geht man in drei Schritten vor:

1. Taylorentwicklung von der exakten Evolution $g_1(\tau) = \Phi^\tau(x)$ bis zur Ordnung p.
2. Taylorentwicklung von der diskreten Evolution $g_2(\tau) = \Psi^\tau(x)$ bis zur Ordnung p.
3. Koeffizientenvergleich der beiden Taylorentwicklungen.

Der Beweis kann z.B in den Skripten von G. Lube oder von T. Hohage nachgelesen werden.

Betrachten wir nun die Ordnungsbedingungen genauer:

$s = 1$**:** Das Schema für $s = 1$ lautet

$$\begin{array}{c|c} c_1 & a_{11} = 0 \\ \hline & b_1 \end{array}$$

Wegen $c_1 = a_{11} = 0$ folgt aus der geforderten Konsistenzordnung von $p = 1$, dass $b_1 = 1$ gelten muss. Das explizite Euler-Verfahren ist also das einzige einstufige, explizite, autonomisierungsinvariante Verfahren der Ordnung 1.

$s = 2$**:** Das Schema für $s = 2$ lautet

$$\begin{array}{c|cc} c_1 & 0 & \\ c_2 & a_{21} & 0 \\ \hline & b_1 & b_2 \end{array}$$

Wegen der Invarianz gegen Autonomisierung sind $c_1 = 0$ und $c_2 = a_{21}$ bereits festgelegt. Als Variablen verbleiben also a_{21}, b_1, b_2, wobei aber die folgenden Bedingungen beachtet werden müssen:

$$\begin{aligned} b_1 + b_2 &= 1 \\ b_1 \underbrace{c_1}_{=0} + b_2 c_2 &= \tfrac{1}{2}. \end{aligned}$$

Aus dem Gleichungssystem ergibt sich

$$\begin{aligned} b_1 &= 1 - b_2 \\ c_2 &= \frac{1}{2b_2}, \text{ falls } b_2 \neq 0. \end{aligned}$$

(Für $b_2 = 0$ ist das System nicht lösbar.) Man erhält das folgende Butcher-Schema für $b \neq 0$.

$$\begin{array}{c|cc} 0 & 0 & \\ \frac{1}{2b} & \frac{1}{2b} & 0 \\ \hline & 1-b & b \end{array}$$

Für $b = 1$ folgt beispielsweise die explizite Mittelpunktsregel und mit $b = \frac{1}{2}$ die explizite Trapezregel, zu der folgendes Butcher-Schema gehört:

$$\begin{array}{c|cc} 0 & 0 & \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

$s = 4$: Für $s = 4$ ergeben sich 10 Unbekannte und 8 Gleichungen, siehe

$$\begin{array}{c|cccc} 0 & 0 & & & \\ c_2 & a_{21} & 0 & & \\ c_3 & a_{31} & a_{32} & 0 & \\ c_4 & a_{41} & a_{42} & a_{43} & 0 \\ \hline & b_1 & b_2 & b_3 & b_4 \end{array}$$

Man kann sich die c_i als Stützstellen der Quadraturformel vorstellen, also für die Simpson-Regel etwa $0, \frac{1}{2}, 1$, was man mit doppelter Stützstelle an $\frac{1}{2}$ als

$$c^t = (0, \tfrac{1}{2}, \tfrac{1}{2}, 1)$$

ausdrücken kann. Eine darauf beruhende Lösung ist das schon vorgestellte *klassische Runge-Kutta Verfahren.*

$s = 10$: In diesem Fall erhält man 1.205 Bedingungen und 55 Variablen.

$s = 20$: Für den Fall $s = 10$ erwarten uns 20.247.374 Bedingungen .

Man erkennt leicht, dass die Anzahl der Bedingungen mit steigendem p immer größer wird.

Beziehung zur numerischen Integration

Wir möchten kurz eine interessante Beziehung zu Kapitel 1 dieses Skriptes erläutern: Man kann die numerische Integration einer Funktion $f \in C([0,1],\mathbb{R})$ auf dem Intervall $[0,1]$ als Spezialfall des folgenden Anfangswertproblemes

$$\begin{aligned} x'(t) &= f(t) \\ x(0) &= 0 \end{aligned}$$

auffassen, denn dessen Lösung ist nach dem Hauptsatz der Differential- und Integralrechnung gegeben durch

$$x(t) = \int_0^t f(\tau)d\tau.$$

Es entspricht also $x(1)$ genau dem gesuchten Integral. Wendet man auf dieses (AWP) ein Runge-Kutta Verfahren an, so erhält man daraus eine Quadraturformel

$$\begin{aligned} \int_0^1 f(\tau)d\tau &= x(1) \approx \Psi^{t_0+\tau,t_0}(x_0) \\ &= \underbrace{x_0}_{=0} + \underbrace{\tau}_{=1} \sum_{j=1}^{s} b_j k_j(t,x,\tau) \\ &= \sum_{i=1}^{s} b_j f(\underbrace{t}_{=0} + c_j \underbrace{\tau}_{=1}) = \sum_{j=1}^{s} b_j f(c_j). \end{aligned}$$

Die jeweils erstgenannten Ordnungsbedingungen aus Satz 3.31 für $p = 1,2,3,4$ entsprechen der Forderung, dass die Monome $1, t, t^2, t^3$ mit Stützstellen c_j und Gewichten b_j exakt integriert werden.

Konvergenz von expliziten Runge-Kutta Verfahren

Bisher haben wir ausschließlich die Konsistenzordnung von Runge-Kutta Verfahren betrachtet. Wir wollen nun die *Konvergenz* der Runge-Kutta Verfahren diskutieren. Auch hierzu benötigen wir in den Voraussetzungen nicht nur die Konsistenz, sondern auch die Stabilität.

Satz 3.32 *Sei $f \in C(D_0,\mathbb{R}^d)$ und genüge der Lipschitz-Bedingung*

$$\|f(x_1) - f(x_2)\| \leq L\|x_1 - x_2\| \text{ für alle } x_1, x_2 \in D_0.$$

Dann erfüllt die diskrete Evolution Ψ eines gegen Autonomisierung invarianten Runge-Kutta-Verfahrens die Stabilitätsbedingung aus Satz 3.24 mit Konstante $L_\Psi = \gamma L$, wobei $\gamma \geq 0$ nur von A und b abhängt. Ist speziell $p \leq 4$ und sind $b_i, a_{ij} \geq 0$ für alle i,j so ist $\gamma = 1$.

Beweis: Der Satz lässt sich durch wiederholtes Anwenden der Lipschitz-Bedingung im Ausdruck

$$\begin{aligned}\|k_i(t,x,\tau) - k_i(t,\tilde{x},\tau)\| &\leq \|f(x+\tau\sum_j a_{ij}k_j(t,x,\tau)) - f(\tilde{x}+\tau\sum_j a_{ij}k_j(t,\tilde{x},\tau))\| \\ &\leq L(\|x-\tilde{x}\| + \tau\sum_j a_{ij}\|k_j(t,x,\tau) - k_j(t,\tilde{x},\tau)\|)\end{aligned}$$

nachrechnen. Auf Details gegen wir hier nicht ein. QED

3.3.5 Implizite Runge-Kutta-Verfahren

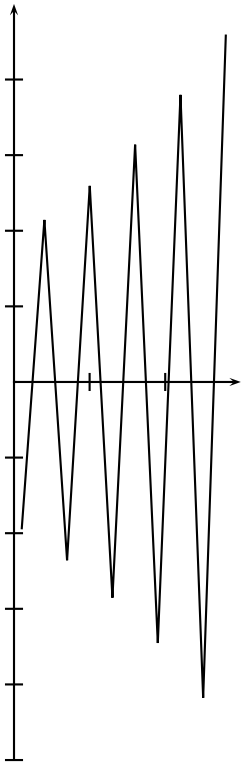
Als „Testproblem“ bekannt ist

$$\begin{aligned}x'(t) &= \lambda x(t) \\ x(0) &= 1\end{aligned}$$

mit Parameter $\lambda \in \mathbb{C}$. Die Lösung ist $x(t) = e^{\lambda x}$. Wir betrachten im Speziellen $\lambda \in \mathbb{R}$. Falls $\lambda < 0$ ist, gilt für $t \to \infty$, dass die Funktion $e^{\lambda t}$ und alle ihre Ableitungen gegen Null konvergieren. Die Hoffnung ist, dass unsere Verfahren schnell konvergieren. Leider ist das nicht so! Das Heun-Verfahren

$$x_\Delta(t_{j+1}) = x_\Delta(t_j) + \tau\lambda x_\Delta(t_j)$$

liefert eine oszillierende, immer weiter ausschlagende Funktion als Lösung.



Euler-Heun Verfahren zu

$$x(t) = -7x(t), \qquad x(0) = 1$$

auf dem Intervall 2 bis 5 mit Schrittweite $h = 0,3$.
Die echte Lösung der DGL ist

$$x(t) = e^{-7t} \approx 0,$$

verläuft also fast entlang der x-Achse. Die Näherung ist unbrauchbar.

Um vernünftige Ergebnisse zu erzielen braucht man *sehr* kleine Schrittweiten. Warum?

Wir erinnern uns an die Idee des Eulerverfahrens: Nutze die Rechteckregel,

$$\int_a^b f(\tau)d\tau \approx (b-a)f(a)$$

um das in der Integralgleichung

$$x(t) = x_0 + \int_a^b f(\tau, x(\tau))d\tau$$

auftretende Integral zu approximieren und erhalte

$$x_\Delta(t_{j+1}) = x_\Delta(t_j) + \tau f(t_j, x_\Delta(t_j)).$$

Benutzen wir stattdessen den rechten Rand des Integrationsintervalls, also

$$\int_a^b f(\tau)d\tau \approx (b-a)f(b),$$

so erhalten wir

$$x_\Delta(t_{j+1}) = x_\Delta(t_j) + \tau f(t_{j+1}, x_\Delta(t_{j+1})),$$

was $e^{\lambda t}$ auch schon für mittlere Schrittweiten recht gut approximiert. Diese Überlegung führt zum *impliziten Euler-Verfahren.*

Graphische Interpretation: Das *explizite* Euler-Verfahren nutzt die Tangente der Lösungskurve im jeweiligen Startpunkt t_j. Das *implizite* Euler-Verfahren nutzt die Tangente der Lösungskurve im jeweiligen Zielpunkt t_{j+1}. Das entspricht dem expliziten Eulerverfahren „von hinten", d.h. mit Startwert t_N.

Wir wenden beide Verfahren auf $f(x) = \lambda x$, $x(0) = 1$ mit $t_j = \tau j$, $j = 0, \ldots, N$ an.

- Explizites Eulerverfahren:

$$x_\Delta(t_{j+1}) = x_\Delta(t_j) + \tau\lambda x_\Delta(t_j).$$

Behauptung: $x_\Delta(t_j) = (1+\lambda\tau)^j$.

Beweis: (Induktion)
$j = 0 \Rightarrow x_\Delta(t_0) = x(0) = 1 = (1+\lambda\tau)^0$
$j \mapsto j+1$:

$$\begin{aligned} x_\Delta(t_{j+1}) &= x_\Delta(t_j) + \tau\lambda x_\Delta(t_j) \\ &= (1+\lambda\tau)^j(1+\lambda\tau) = (1+\lambda\tau)^{j+1}, \end{aligned}$$

woraus die Behauptung folgt.

- Implizites Eulerverfahren

$$x_\Delta(t_{j+1}) = x_\Delta(t_j) + \tau\lambda x_\Delta(t_{j+1}) \Rightarrow x_\Delta(t_{j+1}) = \frac{x_\Delta(t_j)}{1-\lambda\tau}, \qquad \tau\lambda \neq 1.$$

Behauptung: $x_\Delta(t_j) = \left(\frac{1}{1-\lambda\tau}\right)^j$

Beweis: (Induktion)

$j = 0 \Rightarrow x_\Delta(t_0) = x(0) = 1 = \left(\frac{1}{1-\lambda\tau}\right)^0$

$j \mapsto j+1$

$$\begin{aligned} x_\Delta(t_{j+1}) &= \frac{x_\Delta(t_j)}{1-\tau\lambda} = \frac{1}{(1-\lambda\tau)^j}\frac{1}{(1-\tau\lambda)} \\ &= \left(\frac{1}{1-\lambda\tau}\right)^{j+1}, \end{aligned}$$

woraus abermals die Behauptung folgt.

- Zum Vergleich: Die echte Lösung ist $x(t_j) = e^{\lambda t_j}$.

Wir untersuchen unsere Verfahren auf die Eigenschaft $x(t_j) \to 0$ für $j \to \infty$ der echten Lösung $x(t)$ für $\lambda < 0$.

- Im expliziten Euler-Verfahren erhalten wir: $x_\Delta(t_j) \to 0$ falls $|1-\lambda\tau| < 1$. Wegen $|1+\lambda\tau| = |\lambda|\tau - 1$ ist das für $\tau < \frac{2}{|\lambda|}$ erfüllt. Besonders für große λ sind also kleine Schrittweiten erforderlich.

- Im impliziten Eulerverfahren gilt dagegen

$$\left|\frac{1}{1-\lambda\tau}\right| = \frac{1}{|1+|\lambda|\tau|} < 1$$

für alle Schrittweiten τ, also gilt $x_\Delta(t_j) \to 0$ für $j \to \infty$ für jede Schrittweite τ. Das erklärt das bessere Konvergenzverhalten des impliziten Eulerverfahrens.

Bemerkung: Der eben beschriebene Effekt tritt bei dem AWP $x'(t) = \lambda x(t)$, $x(0) = 1$ auch bei allen anderen Runge-Kutta-Verfahren auf, genauer

$$\forall \tau > 0 : \lim_{|\lambda|\to\infty} |\Psi_\lambda^\tau(1)| = \infty,$$

wobei Ψ_λ^τ ein Runge-Kutta-Verfahren zu der Differentialgleichung $f(x) = \lambda x$ ist. (Die Aussage gilt, weil $\Psi_\lambda^\tau(1)$ ein Polynom $\in \Pi_s$ ist.)

Wir erinnern uns daran, dass die exakte Evolution einer Differentialgleichung die Stabilitätsbedingung

$$\|\Phi^{t,t_0}(x_0) - \Phi^{t,t_0}(y_0)\|_2 \leq e^{L_+(t-t_0)}\|x_0 - y_0\|_2$$

erfüllt (Satz 3.18), wobei L_+ die einseitige Lipschitzkonstante ist. Für explizite Runge-Kutta-Verfahren „erbt“ die diskrete Evolution Ψ diese Stabilitätseigenschaften, aber nur mit der Konstanten $L_\Psi = \gamma L$ (Satz 3.32), wobei L die Lipschitzkonstante von f ist. Diese Konstante geht exponentiell in die Fehlerabschätzung aus Satz 3.24

$$\|x_\Delta(t) - x(t)\| \leq r(\tau_\Delta) = \begin{cases} \mathrm{err}(\tau_\Delta)\frac{e^{L_\Psi(t-t_0)}-1}{L_\Psi} & L_\Psi > 0 \\ \mathrm{err}(\tau_\Delta)(t-t_0) & L_\Psi = 0 \end{cases}$$

ein. Daher wäre es gut, wenn $L_\Psi \approx L_+$ gilt.

Das ist bei expliziten Runge-Kutta-Verfahren nicht gegeben, falls $L_+ \ll L$. Solche Differentialgleichungen nennt man **steif**. Für steife Differentialgleichungen liefern explizite Runge-Kutta-Verfahren erst für extrem kleine Schrittweiten verlässliche Ergebnisse und sind daher unbrauchbar. Besser wären Verfahren, bei denen in die Fehlerabschätzung nur die einseitige Lipschitz-Konstante L_+ (und nicht L) einfließt.

Steife Differentialgleichungen treten in der Praxis sehr häufig auf und können (wie in unserem Beispiel) meistens gut mit impliziten Runge-Kutta Verfahren gelöst werden.
Ein s-stufiges implizites Runge-Kutta-Verfahren ist gegeben durch die Vorschrift

$$x_\Delta(t+\tau) := \Psi^{t+\tau,t}(x_\Delta(t)) := x_\Delta(t) + \tau\sum_{j=1}^{s} b_j k_j(t, x_\Delta(t), \tau)$$

mit

$$k_i(t,x,\tau) := f\left(t + c_i\tau, x + \tau\sum_{j=1}^{s} a_{ij}k_j(t,x,\tau)\right).$$

Die Werte c_i nennt man auch **Knoten**, die k_i **Steigungen**.
Das Butcher-Schema lautet:

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array} = \begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \dots & a_{1s} \\ c_2 & a_{21} & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \dots & a_{ss} \\ \hline & b_1 & b_2 & \dots & b_s \end{array}$$

Notation 3.33

- *Für $a_{ij} = 0$, $i \leq j$ ergibt* $\begin{array}{c|c} c & A \\ \hline & b^T \end{array}$ *ein* ***explizites*** *Runge-Kutta-Verfahren*
- *Für $a_{ij} = 0$, $i < j$ erhält man ein* ***diagonal-implizites*** *Runge-Kutta-Verfahren (DIRK). Gilt sogar $a_{ii} = y$, so spricht man von $SDIRK$-Verfahren.*
- *Gibt es ein $j > i$ mit $a_{ij} \neq 0$, so nennt man das Runge-Kutta-Verfahren* ***voll implizit****.*

Bei der Implementation von impliziten Runge-Kutta-Verfahren sind in jedem Schritt die Steigungen k_i durch Lösen von

$$k_i(t, x, \tau) = f\left(t + c_i\tau, x + \tau \sum_{j=1}^{s} a_{ij} k_j(t, x, \tau)\right), \qquad i = 1, \ldots, s$$

zu ermitteln. Leider funktionieren Fixpunktiterationen nur mit Schrittweitenbeschränkungen (vgl. Euler-Heun Verfahren, Lemma 3.20 auf Seite 81). Man benutzt daher das Newton-Verfahren (oder Varianten davon).

Wir wollen nun implizite Runge-Kutta Verfahren höherer Ordnung konstruieren.

- Das implizites Euler-Verfahren $\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$ hat Ordnung 1.
- Das Mittelpunktsverfahren

  $$x_\Delta(t_{j+1}) = x_\Delta(t_j) + \tau f\left(t_j + \frac{\tau}{2}, \frac{x_\Delta(t_j) + x_\Delta(t_{j+1})}{2}\right)$$

  mit dem Butcher-Schema $\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$ hat Konsistenzordnung $p = 2$!

Satz 3.34 *Es gelten sinngemäß die Bedingungen für Konsistenz und Invarianz gegen Autonomisierung sowie die Ordnungsbedingungen auch für implizite Runge-Kutta-Verfahren (Lemma 3.28, Lemma 3.30 und Satz 3.31).*

Zum Festlegen der $s^2 + 2s$ Parameter eines impliziten Runge-Kutta-Verfahrens werden häufig *Kollokationsverfahren* verwendet: Die Idee von Kollokationsverfahren ist es, die Lösung eines gegebenen Anfangswertproblemes durch ein Polynom ω zu approximieren. Dieses soll das Anfangswertproblem an vorgegebenen Stützstellen lösen. Als Stützstellen definiert man *Kollokationspunkte* $t_0 + c_i\tau$, $i = 1, \ldots, s$. Dann verlangt man

$$\omega'(t_0 + c_i\tau) = f(t_0 + c_i\tau, \omega(t_0 + c_i\tau)), \qquad i = 1, \ldots, s \tag{3.14}$$

$$\omega(t_0) = x_0 \tag{3.15}$$

für das vektorwertige Polynom $\omega \in (\Pi_s)^n$. Wir nennen die wesentlichen Resultate:

Lemma 3.35 *Seien für* $0 \leq c_1 < \cdots < c_s \leq 1$ *die Bedingungen* (3.14) *und* (3.15) *eindeutig lösbar. Dann wird durch die diskrete Evolution*

$$\Psi^{t_0+\tau,t}(x_0) := \omega(t_0 + \tau)$$

ein implizites Runge-Kutta-Verfahren definiert, das durch die Parameter

$$\begin{aligned} a_{ij} &= \int_0^{c_i} L_j(\tau)d\tau \text{ für } i,j = 1,\ldots,s \\ b_i &= \int_0^1 L_i(\tau)d\tau \text{ für } i = 1,\ldots,s \end{aligned}$$

gegeben ist.

Lemma 3.36 *Ein durch Kollokation definiertes, implizites Runge-Kutta-Verfahren ist konsistent und invariant gegen Autonomisierung.*

Der Beweis dieser Aussagen lässt sich relativ einfach mit den Standardmitteln dieser Vorlesung zu führen. Dagegen ist der folgende Satz ein etwas tiefliegenderes Ergebnis.

Satz 3.37 *Für gegebene Parameter* $c_1,\ldots,c_s$ *sei die Quadraturformel* $\int_0^1 g(t)dt \approx \sum_{i=1}^s b_i g(c_i)$ *exakt für alle Polynome in* Π_{p-1} *mit* $p \geq s$. *Dann hat das zu* $c_1,\ldots,c_s$ *gehörende, durch Kollokation gewonnene Runge-Kutta-Verfahren die Konsistenzordnung* p.

3.4 Zusammenfassung

Begriffe

- DGL: Differentialgleichung
- AWP: Anfangswerproblem = DGL + Startbedingungen
- gewöhnliche/partielle DGL
- explizite, implizite DGL
- autonom
- linear

Transformationen

- jede gewöhnliche, explizite DGL der Ordnung k kann in eine äquivalente DGL erster Ordnung überführt werden, d Gleichungen $\mapsto k \cdot d$ Gleichungen
- Autonomisierung: Eine gewöhnliche, explizite DGL kann man in eine äquivalente, autonome, gewöhnliche DGL überführen

Eindeutigkeit/Lösbarkeit

- Gegenbeispiel für eindeutige Lösbarkeit
- Gegenbeispiel für Existenz einer Lösung auf ganz I

Äquivalenz: AWP $\Leftrightarrow$ Integralgleichung

$$\begin{aligned} x'(t) &= f(t, x(t)) \\ x(t_0) &= x_0 \end{aligned} \qquad \Leftrightarrow \qquad x(t) = x_0 + \int_{t_0}^{t} f(\tau, x(\tau)) d\tau$$

- Anwendung vom Banach'schen Fixpunktsatz
- Konstruktion von Einschrittverfahren
- Picard-Lindelöf: f stetig + Lipschitzstetig bzgl. der letzten d Variablen. Dann ist jedes AWP auf einer Umgebung U um den Startwert eindeutig lösbar.
  - Die Lösung eines (AWP) ist global eindeutig
  - globale Lösbarkeit auf ganz I, falls Lipschitzstetigkeit global

- Folge: Definition der Evolution Φ einer DGL $x' = f(t,x)$ durch
$$\Phi^{t,t_0}(x_0) = x(t),$$
wenn x die eindeutige Lösung von (AWP)
$$x'(t) = f(t,x)$$
$$x(t_0) = x_0$$
- Evolutionen sind durch drei Eigenschaften eindeutig charakterisiert
- Stabilität einer Evolution
$$\|\Phi^{t,t_0}(x_0) - \Phi^{t,t_0}(x)\| \leq e^{L_+(t-t_0)}\|x_0 - x\|$$

Einschritt-Verfahren

- Gitter $\Delta = \{t_0, \ldots, t_N\}$ gesucht:
$$x_\Delta : \Delta \to \mathbb{R}^d$$
$$x_\Delta(t_{j+1}) := \Psi^{t_{j+1},t_j}(x_\Delta(t_j))$$
- explizites Eulerverfahren
- implizites Eulerverfahren
  - Euler-Heun-Verfahren (implizit, sukzessive Approximation)
  - Prädiktor-Korrektor-Variante
- explizites Runge-Kutta-Verfahren
- implizites Runge-Kutta-Verfahren
- Konsistenz von Ψ: drei äquivalente Bedingungen
$$\|\Psi^{t,t_0}(x_0) - \Psi^{t,t_0}(x_0)\| \to 0 \text{ für } t \to t_0$$
- Konvergenz
$$\|x_\Delta(t) - x(t)\| \to 0 \text{ falls } \tau \to 0, \text{ gleichmäßig}$$
- Konsistenz der Ordnung p + Stabilität $\Rightarrow$ Konvergenz der Ordnung p

Explizite Runge-Kutta-Verfahren

- Butcher-Schema
- Bedingung an Konsistenz und an Invarianz gegen Autonomisierung
- implizite Runge-Kutta-Verfahren für steife DGL
- Kollokationsverfahren

Kapitel 4

Optimierung

4.1 Begriffe und Überblick

Notation 4.1 *Sei $\mathcal{B} \subseteq \mathbb{R}^n$ und sei $f : \mathcal{B} \to \mathbb{R}$. Sei weiter $P \subseteq \mathcal{B}$. Ein Optimierungsproblem ist gegeben durch*

$$(P) \qquad \min_{x \in P} f(x).$$

Man nennt f Zielfunktion, $\mathcal{B}$ Grundmenge und P den zulässigen Bereich von (P).
Schreibweise: (P) wird geschrieben als $\min\{f(x) : x \in P\}$ *oder*

$$\begin{array}{ll} \min & f(x) \\ s.d. & x \in P \end{array}.$$

Bemerkung:

- Es gibt auch Optimierungsprobleme, in denen $\mathcal{B} \subseteq \mathbb{R}^n$ nicht gilt, zum Beispiel bei der Bestimmung einer Funktion.
- $\min_{x \in P} f(x)$ ist äquivalent zu $- - \max_{x \in P} -f(x)$, daher können wir uns o.B.d.A. auf Minimierungsprobleme beschränken.
- Da ein Minimum nicht existieren muss, müsste man eigentlich $\inf_{x \in P} f(x)$ schreiben - die Schreibweise mit min hat sich aber eingebürgert.

Notation 4.2 *Sei $\min_{x \in P} f(x)$ ein Optimierungsproblem.*

- *Jedes $x \in P$ heißt* zulässig.
- *Ist $P = \varnothing$, so nennt man das Optimierungsproblem unzulässig.*
- *$x \in P$ heißt* (global) optimal, *falls $f(x) \leq f(x')$ für alle $x' \in P$ gilt.*

- $x \in P$ *heißt* lokal optimal, *falls es eine „vernünftig definierte" Umgebung* $U(x) \subseteq \mathcal{B}$ *so gibt, dass* $f(x) \leq f(x')$ *für alle* $x' \in U(x)$. *Wenn* $\mathcal{B} = \mathbb{R}^n$ *gilt, so kann man immer* $U(x) = \{x' \in \mathbb{R}^n : \|x - x'\| \leq \varepsilon\}$ *mit einer Norm* $\|\cdot\|$ *wählen.*

Beispiel: Ein aus der Vorlesung schon bekanntes Optimierungsproblem ist die in Kapitel 2 behandelte Approximation in endlich-dimensionalen Räumen:

Gegeben: P „einfache Repräsentanten", $x \in X$

gesucht: $y \in P$ so, dass $\|x - y\|$ klein ist, das heißt

$$\min_{y \in P} f(y),$$

wobei $f(y) = \|x - y\|$.

Wir betrachten jetzt systematisch verschiedene Typen von Optimierungsproblemen.

Nicht-Restringierte, differenzierbare Optimierung

Definition: $\mathcal{B} = P = \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$ differenzierbar.

Ergebnisse:

x^* lokal optimal $\Rightarrow \nabla f(x^*) = 0$.

$\nabla f(x^*) = 0$ und die Hesse-Matrix $H(f)(x^*)$ ist positiv definit $\Rightarrow x^*$ ist lokal optimal.

Verfahren: Verfahren des steilsten Abstiegs („steepest descent"), Newton-Verfahren.

Bemerkung: Globale Optima zu finden ist nicht trivial.

Lineare Optimierung

Definition: $\mathcal{B} = \mathbb{R}^n$, $P \subseteq \mathbb{R}^n$ ist ein Polyeder, $f : \mathcal{B} \to \mathbb{R}$ ist linear.

Ergebnisse: Hat (P) eine Lösung, so gibt es eine Ecke von P, die (global) optimal ist.

Verfahren: Simplex-Verfahren (probiert alle Ecken durch), Innere-Punkte-Verfahren.

Bemerkung: Lineare Optimierung ist weitestgehend verstanden; Effizienz-Steigerung ist aber immer noch sinnvoll.

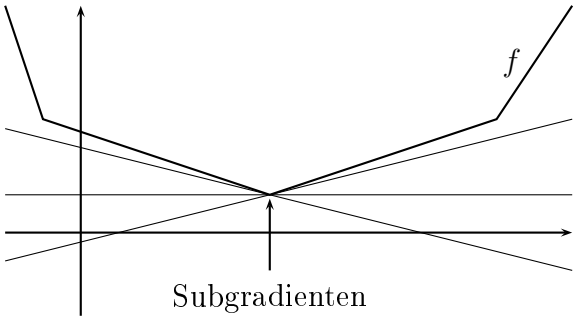
Konvexe Optimierung

Definition: $\mathcal{B} = \mathbb{R}^n$, $P \subseteq \mathbb{R}^n$ konvex, $f : \mathbb{R}^n \to \mathbb{R}$ konvex.

Ergebnisse:

Sei x^* lokales Minimum $\Rightarrow$ x^* ist globales Minimum.

x^* ist (global) optimal auf $\mathbb{R}^n \Leftrightarrow$ Es existiert ein Subgradient $\xi = 0$ an x^*. Auch für $P \subsetneq \mathbb{R}^n$ lassen sich globale Minima durch Subgradienten charakterisieren.



Verfahren: Subgradienten-Verfahren, Volume Algorithmus

Bemerkung: Für spezielle Probleme gibt es effizientere Verfahren.

Konkave Optimierung

Definition: $\mathcal{B} = \mathbb{R}^n$, $P \subseteq \mathbb{R}^n$ konvex, $f : \mathbb{R}^n \to \mathbb{R}$ konkav.

Ergebnisse: Hat (P) eine Lösung, so gibt es einen Extrempunkt von P, der optimal ist. Ist P ein Polyeder, so gibt es eine optimale Ecke. Insbesondere gibt es eine Optimallösung $x^* \in \partial P$.

Verfahren: Auffinden einer endlichen Kandidatenmenge (FDS = finite dominating set).

Ganzzahlige (lineare) Optimierung

Definition: $\mathcal{B} = \mathbb{Z}^n$, $P' \subseteq \mathbb{R}^n$ ist ein Polyeder, $P = P' \cap \mathcal{B}$, $f : \mathbb{R}^n \to \mathbb{R}$ ist linear.

Ergebnisse: Diese liegen vor allem in Spezialfällen vor, zum Beispiel als Ergebnisse im Bereich der Polyeder-Theorie.

Verfahren: Spezialverfahren, welche die Strukturen von P' ausnutzen (TU-Matrizen), ansonsten Gewinnung von oberen Schranken (durch Heuristiken, wie zum Beispiel allgemeine Heuristiken wie Simulated Annealing, genetische Algorithmen, Tabu-Suche) und unteren Schranken (durch Relaxationen).

Bemerkung: Das Problem ist NP-schwer, das heißt ein exaktes Verfahren mit polynomieller Laufzeit ist nicht zu erwarten.

Diskrete Optimierung

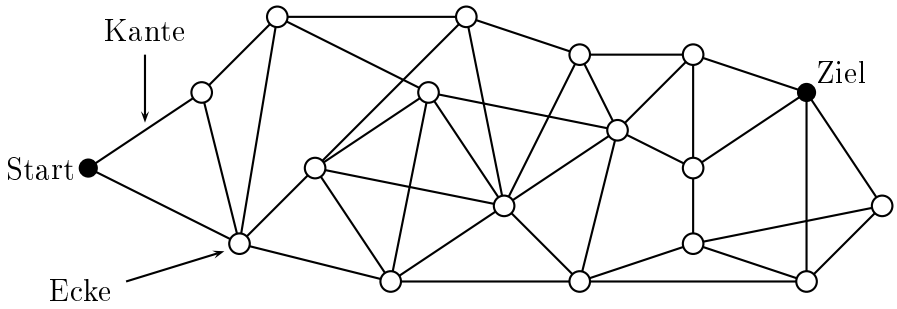
Definition: $\mathcal{B}$ endliche Menge, $P \subseteq \mathcal{B}$ beliebig, $f : \mathcal{B} \to \mathbb{R}$.

Ergebnisse: je nach Problem

Verfahren: je nach Problem oder im Allgemeinen wie Simulated Annealing

Bemerkung: Es gibt effizient lösbare und NP-schwere Probleme.

Beispiel: gegeben ist ein Graph mit Knoten und Kanten, wobei jede Kante eine (positive) Länge hat.



Aufgabe 1: Finde einen kürzesten Weg vom Start zum Ziel.

$\mathcal{B} = \{$alle möglichen Wege vom Start bis zum Ziel$\}$,
$P = \mathcal{B}$,
$f : \mathcal{B} \to \mathbb{R}, \quad f(\text{Weg}) = \text{Länge des Weges} = \sum_{\text{Kanten im Weg}} \text{Länge(Kante)}.$

Dieses Problem ist effizient lösbar in Zeit $\mathcal{O}(n^2)$ (n sei die Anzahl der Knoten im Graph).

Aufgabe 2: Finde den kürzesten Weg vom Start bis zum Ziel, der alle Knoten genau einmal besucht.

$\mathcal{B} = \{$alle möglichen Wege vom Start bis zum Ziel$\}$,
$P \subseteq \mathcal{B}$ enthält die Wege, die alle Knoten genau einmal besuchen
$f : \mathcal{B} \to \mathbb{R}, \quad f(\text{Weg}) = \text{Länge des Weges}.$

Für dieses Problem ist kein effizientes Verfahren bekannt. Es ist schon NP-schwer, herauszufinden, ob $P \neq \varnothing$, d.h. ob es überhaupt einen Weg vom Start bis zum Ziel gibt, der alle Knoten genau einmal besucht.

Die Umgebung eines Weges W kann man z.B. definieren als

$U(W) = \{\text{Wege } W', \text{ die durch Vertauschen von zwei Knoten auf dem Weg } W \text{ entstehen}\}.$

Ein Verfahren, das innerhalb von solchen „benachbarten" zulässigen Lösungen Elemente einer Lösung paarweise tauscht nennt man auch *zwei-opt*. Das Ergebnis eines zwei-opt Verfahrens ist immerhin lokal optimal.

Kontinuierliche, restringierte Optimierung

Definition: $\mathcal{B} = \mathbb{R}^n$, $P \subseteq \mathbb{R}^n$, $f : \mathcal{B} \to \mathbb{R}$.

Ergebnisse: Diese existieren nicht in dieser Allgemeinheit.

Verfahren: Barriere-Verfahren, Penalty-Verfahren (exakt), allgemeine Heuristiken wie Simulated Annealing

Die genannten Klassen von Optimierungsproblemen sind allerdings keineswegs disjunkt. So lassen sich viele diskrete Probleme als ganzzahlige Programme formulieren, oder auch ganzzahlige Programme als nichtlineare Probleme.

Es soll auch nicht unerwähnt bleiben, dass es noch viele weitere Klassen von Optimierungsproblemen gibt. Darunter fallen u.a. quadratische Optimierungsprobleme, die beispielsweise mit dem Verfahren der *konjugierten Gradienten* gelöst werden können.

4.2 Iterative Optimierungsverfahren

In diesem Abschnitt soll auf einige iterative Verfahren zur Lösung von Optimierungsproblemen eingegangen werden. Dabei betrachten wir zuerst differenzierbare Probleme ohne Nebenbedingungen und stellen das Verfahren des steilsten Abstiegs und (kurz) das Newton-Verfahren vor. Danach diskutieren wir Verfahren, die man auf sehr allgemeine restringierte Probleme

$$\min\{f(x) : x \in P\}$$

anwenden kann, nämlich das Strafverfahren und Simulated Annealing.

4.2.1 Differenzierbare, nicht-restringierte Probleme

In diesem Abschnitt betrachten wir die Minimierung einer differenzierbaren Funktion $f : \mathbb{R}^n \to \mathbb{R}$ über dem gesamten $\mathbb{R}^n$. Wir gehen davon aus, dass uns schon Verfahren zur Minimierung von eindimensionalen Funktionen $f : \mathbb{R} \to \mathbb{R}$ zur Verfügung stehen. Solche Verfahren nennt man „Line Search“ Verfahren, darunter sind zum Beispiel

> Intervallhalbierungsverfahren, Dichotomous-Suche, Verfahren des goldenen Schnitts

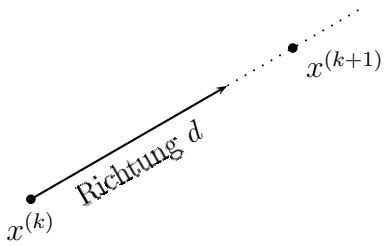
oder, für differenzierbare Funktionen

> Gradienten oder Newton-Verfahren.

Die Methode des steilsten Abstiegs in der mehrdimensionalen Optimierung beruht auf der Idee, eine Lösung $x \in P$ in jedem Schritt entlang einer fest gewählten Richtung d durch Lösen eines eindimensionalen Optimierungsproblems zu verbessern, also durch Lösen des eindimensionalen Problems

$$\min_{\lambda \geq 0} f(x + \lambda d)$$

mit Line Search, wobei f die Zielfunktion darstellt.



Wähle $x^{(k+1)}$ als den besten Punkt entlang der Richtung d. Wir diskutieren zunächst, wie man die Richtung d wählen kann.

Notation 4.3 *Sei $f : \mathbb{R}^n \to \mathbb{R}$ eine Funktion, sei $x \in \mathbb{R}$. Eine Richtung $d \in \mathbb{R}^n$ ist eine* Verbesserungsrichtung *an x bezüglich f, falls es ein $\delta > 0$ so gibt, dass*

$$f(x + \lambda d) < f(x) \quad \textit{für alle } \lambda \in (0, \delta).$$

Das folgende Lemma gibt ein Kriterium, an dem man Verbesserungsrichtungen leicht erkennen kann.

Lemma 4.4 *Sei $f : \mathbb{R}^n \to \mathbb{R}$ eine Funktion, seien $x, d \in \mathbb{R}^n$. Ist die Richtungsableitung $f'(x, d)$ von x in Richtung d echt kleiner als Null, so ist d eine Verbesserungsrichtung.*

Beweis: Es gilt

$$f'(x,d) = \lim_{\lambda \to 0^+} \frac{f(x+\lambda d) - f(x)}{\lambda}.$$

Wegen $f'(x,d) < 0$ gilt also $f'(x+\lambda d) < f(x)$ für alle hinreichend kleinen $\lambda > 0$.

QED

Man kann also jede Richtung d mit negativer Richtungsableitung wählen. Um eine möglichst große Verbesserung zu erzielen, macht es Sinn, eine Richtung d zu wählen, bei der die Richtungsableitung so klein wie möglich ist, also die „Richtung des steilsten Abstiegs“. Das folgende Lemma zeigt, wie man diese Richtung findet. Wir bezeichnen den Gradienten einer Funktion $f : \mathbb{R}^n \to \mathbb{R}$ an $x \in \mathbb{R}^n$ mit $\nabla f(x) \in (\mathbb{R}^n)^*$. Weiterhin sei $\|\cdot\|$ im Folgenden die Euklidische Norm.

Lemma 4.5 *Sei $f : \mathbb{R}^n \to \mathbb{R}$ differenzierbar und sei $\nabla f(x) \neq 0$. Dann ist*

$$\overline{d} = -\frac{(\nabla f(x))^t}{\|\nabla f(x)\|}$$

die normierte Richtung mit kleinster Richtungsableitung, das heißt

$$f'(x,\overline{d}) \leq f'(x,d) \quad \textit{für alle } d \in \mathbb{R}^n \textit{ mit } \|d\| = 1.$$

Beweis: Sei $d \in \mathbb{R}^n$ mit $\|d\| = 1$ beliebig. Dann gilt:

$$\begin{aligned}
|f'(x,d)| &= \left| \lim_{\lambda \to 0^+} \frac{f(x+\lambda d) - f(x)}{\lambda} \right| \\
&= |\nabla f(x) d|, \text{ weil } \lim_{\lambda \to 0} f(x+\lambda d) = f(x) + \lambda \nabla f(x) d \\
&\leq \|\nabla f(x)\| \cdot \|d\| \text{ nach Cauchy-Schwarz} \\
&= \|\nabla f(x)\| = \frac{|\nabla f(x)(\nabla f(x))^t|}{\|\nabla f(x)\|} = |\nabla f(x) \cdot \overline{d}| = |f'(x,\overline{d})|.
\end{aligned}$$

Weiterhin ist $\overline{d}$ eine Abstiegsrichtung wegen

$$f'(x,\overline{d}) = -\frac{\nabla f(x) \cdot (\nabla f(x))^t}{\|\nabla f(x)\|} = -\|\nabla f(x)\| < 0.$$

QED

Algorithmus „Steepest Descent“

Sei $x^{(0)} \in \mathbb{R}^n$ beliebig, $k = 0$.

1. Sei $d^{(k)} := -\nabla f(x^{(k)})$. Falls $\nabla f(x^{(k)}) = 0$, dann STOP.
2. Löse das eindimensionale Optimierungsproblem $\min_{\lambda \geq 0}\{f(x^{(k)} + \lambda d^{(k)})\}$. Sei x^* die Lösung.
3. $x^{(k+1)} := x^*$, gehe zu 1.

Bemerkung: Liegen alle $x^{(k)}$ in einer kompakten Menge, so konvergiert $x^{(k)} \to \overline{x}$ mit $\nabla f(\overline{x}) = 0$. In der Praxis macht das Verfahren in der Nähe des Minimums meistens nur sehr kleine und fast orthogonale Schritte. Man spricht auch von „Zick-Zack-Pfaden“.

Während das Verfahren des steilsten Abstiegs den Gradienten und damit eine lineare Approximation der Funktion f verwendet, nutzt das Newton-Verfahren die quadratische Approximation an die Funktion f. Um einen Punkt x mit $\nabla f(x) = 0$ zu finden, wird f in jedem Schritt durch seine quadratische Approximation ersetzt und eine Nullstelle ihrer Ableitung bestimmt.
Die quadratische Approximation von f an $x^{(k)}$ ist

$$f(x) \approx q(x) = f(x^{(k)}) + \nabla f(x^{(k)})(x - x^{(k)}) + \frac{1}{2}(x - x^{(k)})^t H(x^{(k)})(x - x^{(k)}),$$

wobei $H(x^{(k)})$ die Hesse-Matrix von f an $x^{(k)}$ ist. Wegen

$$\nabla q(x) = \nabla f(x^{(k)}) + (H(x^{(k)})(x - x^{(k)}))^t$$

gilt:

$$(\nabla q(x))^t = 0 \Leftrightarrow (\nabla f(x^{(k)}))^t + H(x^{(k)})(x - x^{(k)}) = 0,$$

also falls

$$x = x^{(k)} - (H(x^{(k)}))^{-1}(\nabla f(x^{(k)}))^t.$$

Entsprechend lautet das Verfahren

Newton-Verfahren

Sei $x^{(0)} \in \mathbb{R}^n$ beliebig, $k = 0$.

1. Falls $\nabla f(x^{(k)}) = 0$, dann STOP.
2. Sonst setze $x^{(k+1)} := x^{(k)} - (H(x^{(k)}))^{-1}(\nabla f(x^{(k)}))^t$, $k := k + 1$, gehe zu 1.

Unter gewissen Voraussetzungen kann quadratische Konvergenz gezeigt werden.

4.2.2 Restringierte Probleme

Wir betrachten

$$\min\{f(x) : x \in \mathcal{B}, x \in P\}.$$

Es gibt mehrere Möglichkeiten, die Verfahren aus dem letzten Abschnitt auch zum Lösen von restringierten Problemen zu nutzen. Eine Idee besteht darin, den berechneten Punkt $x^{(k)}$ in jedem Schritt zulässig zu machen, z.B. durch die Projektion von $x^{(k)}$ auf P, d.h. man wählt den Punkt x aus P, der $\|x - x^k\|$ minimiert. Das wird erfolgreich im Subgradienten-Verfahren für konvexe Probleme eingesetzt. Eine andere Variante ist es, unzulässige Lösungen zu bestrafen. Man spricht von **Strafverfahren**. Betrachten wir dazu

$$P = \{x \in \mathbb{R}^n : g_i(x) \leq 0, i = 1, \ldots, m \text{ und } h_j(x) = 0, j = 1, \ldots, l\}$$

als zulässige Menge unseres Optimierungsproblems. Wie kann man unzulässige Lösungen bestrafen?

Beispiel:

- Das Problem
$$\min f(x), \text{ s.d. } h(x) = 0$$
wird umgewandelt in $\min f(x) + \underbrace{\mu h^2(x)}_{\text{Strafterm}}$, μ groß.

- Das Problem
$$\min f(x), \text{ s.d. } g(x) \leq 0$$
wird umgewandelt in $\min f(x) + \underbrace{\mu(\max\{0, g(x)\})^p}_{\text{Strafterm}}$

Notation: *Sei*

$$(P) \qquad \min\{f(x) : x \in P\} \text{ mit } P \subseteq \mathcal{B}.$$

Dann heißt $\alpha : \mathbb{R}^n \to \mathbb{R}$ ***Straffunktion*** *für* (P)*, falls*

$$\alpha(x) = 0 \text{ für alle } x \in P \text{ und } \alpha(x) > 0 \text{ falls } x \notin P.$$

Das bezüglich α *und* $\mu \geq 0$ *relaxierte Problem ist dann*

$$(P_\mu) \qquad \min\{f(x) + \mu\alpha(x) : x \in \mathcal{B}\}.$$

Weiterhin sei $\theta(\mu) = \inf\{f(x) + \mu\alpha(x) : x \in \mathcal{B}\}$ *für* $\mu \geq 0$ *der Zielfunktionswert von* (P_μ).

Es gilt:

Lemma 4.6

$$\min\{f(x) : x \in P\} \geq \theta(\mu) \text{ für alle } \mu \geq 0 \tag{4.1}$$

Beweis: Sei x^* eine Lösung von P. Dann ist $x^* \in \mathcal{B}$, also für (P_μ) zulässig, und erfüllt

$$f(x^*) = f(x^*) + \mu \underbrace{\alpha(x^*)}_{=0} \geq \min_{x \in \mathcal{B}} f(x) + \mu\alpha(x),$$

also ist die Lösung von (P_μ) mindestens so gut wie x^*. QED

Man kann aber noch mehr zeigen:

Lemma 4.7 *Sei $P \neq \varnothing$ und existiere eine optimale Lösung x_μ von (P_μ) für alle $\mu \geq 0$. Dann gilt:*

- *$\theta(\mu)$ ist monoton wachsend.*
- *$\alpha(X_\mu)$ ist monoton fallend.*
- *$f(X_\mu)$ ist monoton wachsend.*

Beweis: Übung.

Daraus folgt schließlich der folgende Satz:

Satz 4.8 *Sei $P \neq \varnothing$ und existiere eine optimale Lösung x_μ von (P_μ) für alle $\mu \geq 0$ so, dass alle x_μ in einer kompakten Teilmenge von $\mathcal{B}$ enthalten sind. Dann gilt*

$$\min\{f(x) : x \in P\} = \sup_{\mu \geq 0} \theta(\mu) = \lim_{\mu \to \infty} \theta(\mu).$$

Weiter sei $\lambda_k \geq 0$ und $\lambda_k \to \infty$ für $k \to \infty$. Ist $(x_{\lambda_k})_{k \in \mathbb{N}}$ konvergent, dann ist $x := \lim_{k \to \infty} x_{\lambda_k}$ eine optimale Lösung von (P).

Es ergibt sich der folgende Algorithmus:
Sei $\beta > 0$, $\mu_1 > 0$, $k = 1$.

1. Löse
$$(P_{\mu_k}) \qquad \min\{f(x) + \mu_k\alpha(x) : x \in \mathcal{B}\}$$
und erhalte x_{k+1} als optimale Lösung.

2. Falls $\mu_k\alpha(x_{k+1}) < \varepsilon$: x_{k+1} ist zulässig für (P) und nach (4.1) optimal. STOP.
Sonst: $\mu_{k+1} = \beta\mu_k$, $k := k + 1$, gehe zu 1.

Satz 4.8 garantiert Konvergenz zu einer Optimallösung.

Lokale Suche

Hat man bereits eine Lösung $x \in P$ gefunden, besteht die Möglichkeit, diese mittels einer lokalen Suche zu verbessern, um ein lokales Optimum zu erreichen. Dazu sucht man die „Nachbarschaft“ von x ab. Das Verfahren lässt sich auch gut auf diskrete Probleme anwenden.

Beispiel (Nachbarschaften):

- Für
$$\min\{f(x) : x \in P\}, P \subseteq \mathbb{R}^n$$
kann man $N(x) = U_\varepsilon(x) \cap P$ wählen.

- Betrachtet man
$$\min\{f(x) : x \in \{0,1\}^n\}$$
so kann man zum Beispiel

  $N(x) = \{x' \in \{0,1\}^n : x \text{ und } x' \text{ unterscheiden sich nur an höchstens } k \text{ Stellen}\}$

  wählen, wobei k (meistens klein) fest gewählt ist.

- Ist
$$\min\{f(P) : P \text{ Weg in Graph}\},$$
so bietet sich

  $N(P) = \{\text{Wege } P' \text{ die aus } P \text{ durch Vertauschen von zwei Knoten entstehen}\}$

  an.

Für die lokale Suche sei $x \in P$ gegen.

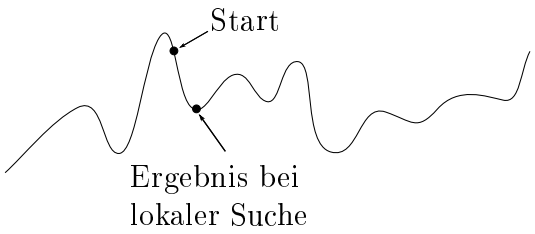
1. Teste, ob es $x' \in N(x)$ mit $f(x') < f(x)$ gibt.
2. Falls ja, setze $x := x'$ und gehe zu 1.

   Sonst: x' lokal optimal, STOP.

Das Verfahren macht Sinn, wenn Schritt 1 leicht zu lösen ist. Oft kann man sogar schnell
$$\min\{f(x') : x' \in N(x)\}$$
lösen, z.B. wenn die Funktion lokal konvex ist, die Mengen $N(x)$ konkave Bereiche sind oder lokale Konvergenz wie beim Newtonverfahren durch Durchprobieren im diskreten Fall vorliegt.

Simulated Annealing

Beim Simulated Annealing versucht man, die lokale Suche so abzuändern, dass man mit hoher Wahrscheinlichkeit ein globales Optimum findet. Man erlaubt dazu auch Schritte, in denen sich der Zielfunktionswert verschlechtert. Dabei soll die Wahrscheinlichkeit für eine Verschlechterung größer sein, wenn die Verschlechterung nur klein ist und im Laufe des Verfahrens abnehmen.



Wir erhalten folgenden Algorithmus:

ALgorithmus: Simulated Annealing
Input: $x \in P$, T_k die „Starttemperatur“, $0 < \alpha < 1$.
Solange T_k groß genug („nicht gefroren“).

1. Wähle zufälliges $x' \in N(x)$.
2. Ist $f(x') < f(x)$, setze $x = x'$ und gehe zu 1.
   Ist $f(x') \geq f(x)$, setze $x := x'$ mit Wahrscheinlichkeit
   $$e^{-\frac{f(x')-f(x)}{T}}.$$
   Setze $T_{k+1} := \alpha T_k$ und gehe zu 1.

Die Idee entstammt chemischen Abkühlungsprozessen, bei denen bei hoher Temperatur eine stabile Molekülbewegung zu beobachten ist, beim Abkühlen aber energieminimale Anordnungen entstehen. Dabei macht das Verfahren nur Sinn, wenn die Nachbarschafts-Definition die folgenden Bedingungen erfüllt:

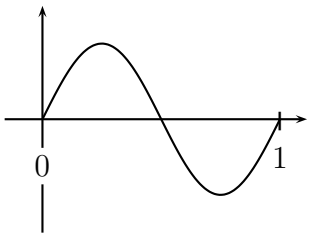
1. $x \in N(x)$, für alle x.
2. $x \in N(x') \Leftrightarrow x' \in N(x)$.
3. Für alle x, x' existiert eine Folge x_k, so dass $x \in N(x_1)$, $x_i \in N(x_{i+1})$, $i = 1, \ldots, k-1$, $x_k \in N(x')$, d.h. jeder Punkt ist von x aus erreichbar.

Kapitel 5

Eigenwertaufgaben

5.1 Motivation

Sei $u(x,t)$ die vertikale Auslenkung einer eingespannten Saite an der Position $x \in [0,1]$ zur Zeit t.



u erfüllt näherungsweise die Wellengleichung

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2}(x,t) &= \frac{1}{c^2} \cdot \frac{\partial^2 u}{\partial x^2}(x,t), \quad x \in (0,1), \quad t \in \mathbb{R} \\ u(0,t) &= u(1,t) = 0, \quad t \in \mathbb{R}, \end{aligned} \tag{5.1}$$

wobei c die Ausbreitungsgeschwindigkeit ist. Wir suchen zeitharmonische Lösungen, das heißt wir machen den Ansatz

$$u(x,t) = \mathrm{Re}(v(x)e^{i\omega t})$$

mit unbekanntem $\omega \in \mathbb{C}$. Einsetzen liefert die gewöhnliche Differentialgleichung

$$\begin{aligned} -v''(x) &= \left(\frac{\omega}{c}\right)^2 v(x), \quad x \in (0,1) \\ v(0) &= v(1) = 0 \end{aligned} \tag{5.2}$$

Das ist ein Eigenwertproblem für den Differentialoperator

$$A : \{u \in C^2([0,1]) | v(0) = v(1) = 0\} \to C([0,1]), \qquad u \mapsto -u''.$$

Diskretisiert man nun dieses Problem, so erhält man ein Matrix-Eigenwertproblem. Betrachten wir hierzu die Gitterpunkte

$$x_j = jh, \quad j = 0, \ldots, N, \qquad h = \frac{1}{N}$$

und approximieren die zweite Ableitung durch den Differenzenquotienten

$$-v''(x_j) \approx \frac{1}{h^2}[-\underbrace{v(x_{j-1})}_{=:v_{j-1}} + 2\underbrace{v(x_j)}_{=:v_j} - \underbrace{v(x_{j+1})}_{=:v_{j+1}}], \quad j = 1, \ldots, N-1.$$

Damit bekommt die Differentialgleichung 5.2 die Form

$$\frac{1}{h^2}(-v_{j-1} + 2v_j - v_{j+1}) = \left(\frac{\omega}{c}\right)^2 v_j, \quad j = 1, \ldots, N-1$$

und man erhält das Problem

$$\frac{c^2}{h^2}\begin{pmatrix} 2 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ 0 & & & -1 & 2 \end{pmatrix}\begin{pmatrix} v_1 \\ \vdots \\ \vdots \\ \vdots \\ v_{N-1} \end{pmatrix} = \omega^2 \begin{pmatrix} v_1 \\ \vdots \\ \vdots \\ \vdots \\ v_{N-1} \end{pmatrix}.$$

Insgesamt haben wir also das Problem 5.1 in ein Matrix-Eigenwertproblem überführt.

5.2 Eigenwerte

Definition 5.1 *Sei $A \in \mathbb{R}^{n\times n}$. Eine Zahl $\lambda \in \mathbb{R}$ heißt Eigenwert zum Eigenvektor $x \in \mathbb{R}^n \setminus \{0\}$, falls*

$$Ax = \lambda x$$

gilt.

Die einfachste Berechnung für den Eigenwert λ benutzt das charakteristische Polynom

$$\varphi(\lambda) = \det(A - \lambda \,\mathrm{Id}).$$

Aus AGLA ist bekannt, dass $\varphi \in \Pi_n$ ein Polynom ist, dessen Wurzeln die Eigenwerte von A sind. Verfahren, die das charakteristische Polynom verwenden, heißen *direkte Verfahren* (z.B. Newton-Verfahren auf φ angewendet). Im Allgemeinen ist die Berechnung des charakteristischen Polynoms durch die Determinante jedoch sehr aufwändig, also werden wir im Folgenden Verfahren betrachten, welche die Berechung von φ vermeiden. Diese Verfahren heißen *iterative Verfahren.*
Grundsätzlich gibt es viele verschiedene Aufgabenstellungen:

- Berechnung des größten bzw. kleinsten Eigewertes
- Berechnung aller Eigenwerte
- Berechnung einiger Eigenwerte mit zugehörigen Eigenvektoren
- Berechnung aller Eigenwerte mit zugehörigen Eigenvektoren

In der Vorlesung werden wir die erste und die vierte Aufgabenstellung betrachten und für diese jeweils ein Beispiel angeben.

5.3 Lokalisierungssatz

Satz 5.2 (Lokalisierungssatz) *Ist $\|\cdot\|$ eine zu einer Vektornorm passende Matrixnorm, so gilt für jeden Eigenwert λ von A die Abschätzung*

$$|\lambda| \leq \rho(A) \leq \|A\| \quad \textit{(siehe Numerik I).}$$

Weiterhin gilt der folgenden Satz.

Satz 5.3 (Gerschgorin) *Für $A = (a_{jk}) \in \mathbb{K}^{n\times n}$ definieren wir die* Gerschgorin-Kreise *als*

$$G_j := \left\{ \lambda \in \mathbb{K} \,\middle|\, |\lambda - a_{jj}| \leq \sum_{\substack{k=1 \\ k\neq j}}^{n} |a_{jk}| \right\}, \quad j = 1, \ldots, n$$

und

$$G_k^* := \left\{ \lambda \in \mathbb{K} \,\middle|\, |\lambda - a_{kk}| \leq \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{jk}| \right\}, \quad k = 1, \ldots, n.$$

Dann gilt für alle Eigenwerte λ von A:

$$\lambda \in \bigcup_{j=1}^{n} G_j \quad \textit{und} \quad \lambda \in \bigcup_{k=1}^{n} G_k^*.$$

Beweis: Sei $Ax = \lambda x$ und $\|x\|_\infty = 1$. Wähle einen Index j mit $|x_j| = \|x\|_\infty = 1$. Dann gilt

$$|\lambda - a_{jj}| = |(\lambda - a_{jj})x_j| = |(Ax)_j - a_{jj}x_j| = \left| \sum_{\substack{k=1 \\ k\neq j}}^{n} a_{jk}x_k \right| \leq \sum_{\substack{k=1 \\ k\neq j}}^{n} |a_{jk}||x_k| \leq \sum_{\substack{k=1 \\ k\neq j}}^{n} |a_{jk}|.$$

Daraus folgt $\lambda \in \bigcup_{j=1}^{n} G_j$. Da A^* die komplex konjugierten Eigenwerte von A besitzt, folgt nun auch $\lambda \in \bigcup_{k=1}^{n} G_k^*$. QED

Im Folgenden wollen wir untersuchen, ob die Eigenwerte von A^* stetig von den Matrixeinträgen abhängen. Zudem werden wir untersuchen, was man über die Lage eines Eigenwertes sagen kann, wenn man „ungefähr“ einen Eigenvektor kennt. Wir werden hier nur den Fall von symmetrischen Matrizen untersuchen.
Die Resultate gelten in ähnlicher Form auch für normale Matrizen ($AA^T = A^TA$). Bei nicht-normalen Matrizen muss man mit extremer Empfindlichkeit der Eigenwerte bei ungenauen Daten rechnen.

Satz 5.4 (Rayleigh) *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch. Seien $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n$ die Eigenwerte von A mit zugehörigen, orthonormalen Eigenvektoren $x_1, \ldots, x_n$. Sei $V_1 = \mathbb{R}^n$ und $V_j = \{x \in \mathbb{R}^n | x^t x_k = 0$ für alle $1 \leq k \leq j-1\}$. Dann gilt*

$$\lambda_j = \max_{\substack{x\in V_j \\ x\neq 0}} \frac{x^tAx}{x^tx} \text{ für alle } 1 \leq j \leq n.$$

Beweis: Sei $x \in V_j \setminus \{0\}$. Dann lässt sich x schreiben als $x = \sum_{k=j}^n c_k x_k$ mit $c_k = x^t x_k$, da der Raum V_j von den $x_j, \ldots, x_n$ aufgespannt wird und die $x_1, \ldots, x_n$ orthonormal sind. Also gelten $x^tx = \sum_{k=j}^n {c_k}^2$ und $Ax = \sum_{k=j}^n c_k\lambda_k x_k$. Man rechnet nun nach, dass

$$\frac{x^tAx}{x^tx} = \frac{\sum_{k=j}^n {c_k}^2\lambda_k}{\sum_{k=j}^n {c_k}^2} \leq \frac{\lambda_j \sum_{k=j}^n {c_k}^2}{\sum_{k=j}^n {c_k}^2} = \lambda_j.$$

Daraus folgt nun, dass

$$\max_{x\in V_j\setminus\{0\}} \frac{x^tAx}{x^tx} \leq \lambda_j$$

gilt. Für den Eigenvektor x_j zu λ_j gilt die Gleichheit, also wird das Maximum auch angenommen. QED

Satz 5.5 (Courant) *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und seien $\lambda_1 \geq \ldots \geq \lambda_n$ die Eigenwerte von A. Dann gilt*

$$\lambda_j = \min_{U_j\in M_j} \max_{\substack{x\in U_j \\ x\neq 0}} \underbrace{\frac{x^tAx}{x^tx}}_{\substack{Rayleigh \\ Quotient}} \text{ für alle } 1 \leq j \leq n,$$

wobei M_j die Menge aller $(n+1-j)$-dimensionalen Unterräume von $\mathbb{R}^n$ bezeichnet.

Beweis: Seien $x_1, \ldots, x_n$ orthogonale Eigenvektoren und die V_j wie in Satz 5.4. Aus $V_j \in M_j$ folgt

$$\min_{U_j\in M_j} \max_{x\in U_j\setminus\{0\}} \frac{x^tAx}{x^tx} \leq \lambda_j.$$

Umgekehrt gibt es für jedes $U_j \in M_j$ ein $x \in U_j \setminus \{0\}$ mit $x^tx_k = 0$ für $j+1 \leq k \leq n$. Also wird das Minimum angenommen.

Korollar 5.6 *Seien $A, B \in \mathbb{R}^{n\times n}$ symmetrisch. Seien $\lambda_1(A), \ldots, \lambda_n(A)$ bzw. $\lambda_1(B), \ldots, \lambda_n(B)$ die zu A bzw. B gehörenden Eigenwerte. Dann gilt für jede beliebige natürliche Matrixnorm:*

$$|\lambda_j(A) - \lambda_j(B)| \leq \|A - B\|.$$

Beweis: Übung.
Tip: Zeige $\lambda_j(A) \leq \lambda_j(B) + \|A - B\|$ und verstausche die Rolle von A und B.

5.4 Verfahren von Mises

Sei $A \in \mathbb{R}^{n\times n}$ diagonalisierbar und habe einen dominanten Eigenwert, das heißt es gilt $|\lambda_1| \gg |\lambda_2| \geq \ldots \geq |\lambda_n|$ für einen Eigenwert λ_1. Sei $x_1, \ldots, x_n$ eine Basis aus Eigenvektoren, dann hat jedes $x \in \mathbb{R}^n$ eine eindeutige Darstellung $x = \sum_{j=1}^n \alpha_j x_j$ mit $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$. Nun gilt

$$A^m x = \sum_{j=1}^n \alpha_j A^m x_j = \sum_{j=1}^n \alpha_j \lambda_j x_j = \lambda_1^m \left(\alpha_1 x_1 + \underbrace{\sum_{j=2}^n \alpha_j \left(\frac{\lambda_j}{\lambda_1}\right)^m x_j}_{=:R_m} \right). \tag{5.3}$$

Man erkennt nun, dass $R_m \to 0$ für $m \to \infty$. Also erhalten wir, falls $\alpha_1 \neq 0$, dass

$$\frac{A^m x}{\lambda_1^m} \to \alpha_1 x_1.$$

Das Problem ist jetzt, dass λ_1 unbekannt ist. Außerdem konvergiert $A^m x$ nur für $|\lambda_1| < 1$. Ein Ausweg aus dieser Situation ist, dass man eine andere Normierung vornimmt. Wir betrachten

$$\|A^m x\|_2 = \left(\sum_{j,k=1}^n \alpha_j \alpha_k \lambda_j^m \lambda_k^m x_j{}^t x_k \right)^{\frac{1}{2}} =: |\lambda_1|^m (|\alpha_1| \|x_1\|_2 + r_m). \tag{5.4}$$

mit $\mathbb{R} \ni r_m \to 0$ für $m \to \infty$. Dann folgt

$$\frac{\|A^{m+1}x\|_2}{\|A^m x\|_2} = \frac{\|A^{m+1}x\|}{|\lambda_1|^{m+1}} \cdot \frac{|\lambda_1|^m}{\|A^m x\|} \cdot |\lambda_1| \to |\lambda_1| \text{ für } m \to \infty. \tag{5.5}$$

Definition 5.7 *Bei dem Mises-Verfahren (auch Potenzmethode genannt) wird ein Startvektor $x^{(0)} = \sum_{j=1}^n \alpha_j x_j$, $\alpha_1 \neq 0$ gewählt und $y^{(0)} = \frac{x^{(0)}}{\|x^{(0)}\|}$ gesetzt. Für $m \geq 1$ wird dann definiert*

$$x^{(m)} = A y^{(m-1)}$$

$$y^{(m)} = \frac{\sigma_m x^{(m)}}{\|x^{(m)}\|} \text{ mit } \sigma_m \in \{-1, 1\} \text{ so, dass } y^{(m)t} y^{(m-1)} \geq 0.$$

Dabei bedeutet die Vorzeichenwahl, dass der Winkel zwischen $y^{(m)}$ und $y^{(m-1)}$ im Intervall $[0, \frac{\pi}{2}]$ liegt, also dass es beim Übergang von $y^{(m-1)}$ zu $y^{(m)}$ keinen Sprung gibt. Um $\alpha_1 \neq 0$ müssen wir uns keine Sorgen machen, denn Rundungsfehler stellen die Bedingung meist sicher.

Satz 5.8 (Konvergenzbeweis für von Mises) *Sei $A \in \mathbb{R}^{n\times n}$ diagonalisierbar und habe einen dominanten Eigenwert λ_1, dann gilt:*

- $\|x^{(m)}\| \to |\lambda_1|$ *für* $m \to \infty$,
- $y^{(m)}$ *konvergiert für* $m \to \infty$ *gegen einen Eigenvektor von A zum Eigenwert* λ_1,
- $\sigma^{(m)} \to \operatorname{sign}(\lambda_1)$, *das heißt* $\sigma^{(m)} = \operatorname{sign}(\lambda_1)$ *für m groß genug.*

Beweis: Durch Induktion kann man zeigen, dass

$$y^{(m)} = \sigma^{(m)} \cdots \sigma^{(1)} \cdot \frac{A^{(m)}x^{(0)}}{\|A^{(m)}x^{(0)}\|_2} \text{ für } m = 1, 2, \ldots.$$

Einsetzen ergibt dann

$$x^{(m+1)} = Ay^{(m)} = \sigma^{(m)} \cdots \sigma^{(1)} \cdot \frac{A^{(m+1)}x^{(0)}}{\|A^{(m)}x^{(0)}\|_2}.$$

Aus (5.5) folgt nun, dass

$$\|x^{(m+1)}\|_2 \to |\lambda_1| \text{ für } m \to \infty$$

gilt. Wir nehmen nun ohne Einschränkungen an, dass $\|x_1\|_2 = 1$, dann gilt:

$$\begin{aligned} y^{(m)} &= \sigma^{(m)} \cdots \sigma^{(1)} \cdot \frac{\lambda_1^m(\alpha_1 x_1 + R_m)}{|\lambda_1|^m(|\alpha_1| + r_m)} \text{ mit (5.4) und (5.3)} \\ &= \sigma^{(m)} \cdots \sigma^{(1)} \cdot \operatorname{sign}(\lambda_1)^m \operatorname{sign}(\alpha_1) x_1 + \rho_m, \end{aligned}$$

wobei $\rho_m \to 0$ für $m \to \infty$. Daraus folgt, wenn $\sigma^{(m)}$ konstant ist für große m, dass $y^{(m)}$ gegen einen Eigenvektor von A zum Eigenwert λ_1 konvergiert. Dieses gilt, weil

$$\begin{aligned} 0 \leq y^{(m-1)^t} y^{(m)} &= \sigma^{(m)}\sigma^{(m-1)} \cdots \sigma^{(1)} \cdot \frac{\lambda_1^{2m-1}(\alpha_1 {x_1}^t + {R_{m-1}}^t)(\alpha_1 x_1 + R_m)}{|\lambda_1|^{(2m-1)}(|\alpha_1| + r_{m-1})(|\alpha_1| + r_m)} \text{ mit (5.4) und (5.3)} \\ &= \sigma^{(m)} \operatorname{sign}(\lambda_1) \cdot \underbrace{\frac{{\alpha_1}^2 + \alpha_1 {x_1}^t R_m + \alpha_1 {R_{m-1}}^t x_1 + {R_{m-1}}^t R_m}{|\alpha_1|^2 + |\alpha_1|(r_{m-1} + r_m) + r_{m-1} r_m}}_{\to 1 \text{ für } m\to\infty} \end{aligned}$$

Wielandt-Verfahren (Inverse Iteration, Nachiteration)

Sei A diagonalisierbar und λ_j ein einfacher Eigenwert von A. Sei λ kein Eigenwert von A und eine Näherung an λ_j, das heißt

$$|\lambda - \lambda_j| \ll |\lambda - \lambda_k| \text{ für } k \neq j.$$

Es folgt: $(A - \lambda \,\mathrm{Id})$ ist nichtsingulär und $(A - \lambda \,\mathrm{Id})^{-1}$ hat die Eigenwerte $\tilde{\lambda}_i$ mit $\tilde{\lambda}_i = \frac{1}{\lambda_i - \lambda}$. Also hat die Matrix $(A - \lambda \,\mathrm{Id})^{-1}$ einen dominanten Eigenwert $\tilde{\lambda}_j$ und die von Mises Iteration ist anwendbar.

Jakobiverfahren

Sei $A \in \mathbb{R}^{m \times m}$ symmetrisch. Wir betrachten die Frobenius-Norm

$$\|A\|_{\mathrm{F}} = \left[\sum_{i,j=1}^{n} |a_{ij}|^2 \right]^{\frac{1}{2}}.$$

Lemma 5.9

1. $\|A\|_F = \mathrm{spur}(A^T A) = \mathrm{spur}(A A^T)$
2. $\|A\|_F = \|Q^T A Q\|_F$, *Q orthogonal*

Beweis:

1. Da für die Spur eines Matrixprodukts AB mit $A, B \in \mathbb{R}^{m \times m}$ gilt

$$\mathrm{spur}(AB) = \sum_{i,j=1}^{n} a_{ij} b_{ji} = \sum_{i,j=1}^{n} b_{ij} a_{ji} = \mathrm{spur}(BA),$$

   bekommen wir insbesondere

$$\mathrm{spur}(A^T A) = \mathrm{spur}(A A^T) = \sum_{i,j=1}^{n} |a_{ij}|^2.$$

2. Wir nutzen die Eigenschaft einer orthogonalen Matrix $Q \in \mathbb{R}^{m \times m}$: $Q^{-1} = Q^T$.

$$\begin{aligned} \|Q^T A Q\|_{\mathrm{F}}^2 &= \mathrm{spur}(Q^T A Q Q^T A^T Q) = \mathrm{spur}(Q^T A A^T Q) = \mathrm{spur}(A^T Q Q^T A) \\ &= \mathrm{spur}(A^T A) = \|A\|_{\mathrm{F}}. \end{aligned}$$

QED

Ist $A \in \mathbb{R}^{m\times m}$ symmetrisch, so lässt sich A nach dem Spektralsatz mit einer orthogonalen Transformation auf Diagonalgestalt bringen. Zusammen mit dem Lemma folgern wir

$$\|A\|_{\mathrm{F}}^2 = \sum_{i,j=1}^{n} |a_{ij}|^2 = \sum_{i=1}^{n} |\lambda_i|^2.$$

Definition 5.10 *Eine* Außennorm *ist eine Abbildung*

$$\begin{aligned} N : \mathbb{R}^{m\times m} &\to \mathbb{R} \\ A &\mapsto \sum_{\substack{i,j=1 \\ i\neq j}}^{n} |a_{ij}|^2 = \|A\|_F^2 - \sum_{i=1}^{n} |a_{ii}|^2. \end{aligned}$$

Bemerkung: Die Außennorm ist keine Norm!

Trivialerweise verschwindet die Außennorm für Diagonalmatrizen. Sei a_{ij} ein Nichtdiagonalelement, d.h. $i \neq j$, ungleich Null. Wir betrachten eine Teilmatrix unserer symmetrischen Matrix A, nämlich:

$$\left(\begin{pmatrix} a_{ii} & a_{ij} \\ a_{ij} & a_{jj} \end{pmatrix} \right).$$

Wir wollen nur mithilfe von Rotationen die Nichtdiagonalelemente eliminieren.

$$\begin{aligned} \begin{pmatrix} b_{ii} & b_{ij} \\ b_{ij} & b_{jj} \end{pmatrix} &:= \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix} \begin{pmatrix} a_{ii} & a_{ij} \\ a_{ij} & a_{jj} \end{pmatrix} \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} \\ &= \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix} \begin{pmatrix} a_{ii}\cos\varphi - a_{ij}\sin\varphi & a_{ij}\cos\varphi - a_{ii}\sin\varphi \\ a_{ij}\cos\varphi + a_{jj}\sin\varphi & a_{jj}\cos\varphi - a_{ij}\sin\varphi \end{pmatrix} \end{aligned}$$

Also lässt sich das transformierte Matrixelement b_{ij} auf folgende Art und Weise berechnen:

$$\begin{aligned} b_{ij} &:= a_{ij}\cos^2\varphi - a_{ii}\cos\varphi\sin\varphi + a_{jj}\sin\varphi\cos\varphi - a_{ij}\sin^2\varphi \\ &= (a_{jj} - a_{ii})\sin\varphi\cos\varphi + a_{ij}(\cos^2\varphi - \sin^2\varphi) \\ &= \tfrac{1}{2}(a_{jj} - a_{ii})\sin 2\varphi + a_{ij}\cos 2\varphi. \end{aligned}$$

Wollen wir es verschwinden lassen, folgt

$$\cot 2\varphi = \frac{a_{ii} - a_{jj}}{2a_{ij}}.$$

Um Kosinus und Sinus als Winkelfunktionen zu vermeiden, definiert man $\tau := \cos(2\varphi) = \cos^2\varphi - \sin^2\varphi$, $\varphi \in [-\pi/4, \pi/4]$. Dann gilt: $\cos\varphi = \sqrt{(1+\tau)/2}$, $\sin\varphi = \sigma\sqrt{(1-\tau)/2}$, $\sigma(\varphi) \in -1, 1$. Sei φ so gewählt, dass

$$a_{ij}\tau + (a_{jj} - a_{ii})\frac{\sigma}{2}\sqrt{1-\tau^2} = 0.$$

Eine Möglichkeit ist

$$\tau = \frac{a_{ii} - a_{jj}}{\sqrt{4a_{ij}^2 + (a_{ii} - a_{jj})^2}}, \qquad \sigma = \operatorname{sign}(a_{ij})$$

wählen. Das Vorzeichen σ ergibt sich wegen des Zählers von τ. Statt der Rotationsmatrix können wir also die Transformationsmatrix

$$\begin{pmatrix} c & -s \\ s & c \end{pmatrix} \quad \text{mit} \quad c = \sqrt{\frac{1+\tau}{2}} \text{ und } s = \sigma\sqrt{\frac{1-\tau}{2}}.$$

verwenden. Wir gehen von der Teilmatrix zur gesamten Matrix über.

Definition 5.11 *Sei $1 \leq i < j \leq n$. Dann nennen wir die Matrix*

$$\begin{pmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ & & & c & & & & -s & & \\ & & & & 1 & & & & & \\ & & & & & \ddots & & & & \\ & & & & & & 1 & & & \\ & & & s & & & & c & & \\ & & & & & & & & 1 & \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{pmatrix}$$

mit
$c = \cos\varphi$,
$s = -\sin\varphi$,
$g_{ii} = g_{jj} = \cos\varphi$,
$g_{ij} = -g_{ji} = \sin\varphi$,
ansonsten Identität

eine Givens-Rotation *bzw. eine* Jakobi-Transformation.

Offensichtlich gilt:

Lemma 5.12 *Die* Givens-Rotation *ist orthogonal.*

Satz 5.13 *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch, $i \neq j$ mit $a_{ij} \neq 0$. Die Matrix*

$$B = G_{ij} A G_{ij}^T$$

1. *ist wieder symmetrisch,*
2. *es gilt $b_{ij} = 0$,*
3. *A und B unterscheiden sich nur in der i-ten bzw. j-ten Spalte/Zeile*
4. *und $N(B) = N(A) - 2a_{ij}^2$.*

Beweis: Aussagen 1-3 folgen aus den bisherigen Überlegungen. Weil die Frobeniusnorm unter orthogonalen Transformationen invariant ist, gilt

$$\left\| \begin{pmatrix} a_{ii} & a_{ij} \\ a_{ij} & a_{jj} \end{pmatrix} \right\|_{\mathrm{F}} = \left\| \begin{pmatrix} b_{ii} & b_{ij} \\ b_{ij} & b_{jj} \end{pmatrix} \right\|_{\mathrm{F}}.$$

Über diese Gleichheit und $b_{ij=0}$ erhalten wir durch Quadrieren $a_{ii}^2 + a_{jj}^2 + 2a_{ij}^2 = b_{ii}^2 + b_{jj}^2$. Wir können – da alle anderen Diagonalelemente von A und B gleich bleiben – auf die Außennorm zurückschließen.

$$\begin{aligned} N(B) &= \|B\|_{\mathrm{F}}^2 - \sum_{k=1}^{n} |b_{kk}|^2 = \|A\|_{\mathrm{F}}^2 - \sum_{k=1}^{n} |b_{kk}|^2 \\ &= N(A) + \sum_{k=1}^{n} (|a_{kk}|^2 - |b_{kk}|^2) \\ &= N(A) - 2a_{ij}^2. \end{aligned}$$

QED

Aus diesen Ergebnissen formulieren wir das Jakobi-Verfahren:

Definition 5.14 *Sei $A \in \mathbb{R}^{n \times n}$ symmetrisch. Im* klassischen Jakobi-Verfahren *wird zunächst $A^{(0)} = A$ gesetzt und für $m = 1, 2, \ldots$ iteriert mit $A^{(m)} = (a_{lk}^{(m)})$:*

1. *Suche $i \neq j$ mit $|a_{ij}^{(m)}| = \max_{l \neq k} |a_{lk}^{(m)}|$ und setze $G^{(m)} := G_{ij}$,*
2. *setze $A^{(m+1)} = G^{(m)} A^{(m)} {G^{(m)}}^T$.*

Das Verfahren sucht also das größte Element aus der Matrix heraus und transformiert es auf Null. Weil wir mit symmetrischen Matrizen arbeiten, müssen wir nur eine obere Dreicksmatrix durchsuchen. Obwohl bei jeder Transformation Nullen wieder verschwinden können, liegt Konvergenz vor.

Satz 5.15 *Das klassische Jakobi-Verfahren konvergiert zumindest linear in der Außennorm.*

Beweis: Wir betrachten ein festes m und erwähnen es daher nicht. Da wir $|a_{ij}| = \max_{l \neq k} |a_{lk}|$ gesetzt haben, können wir damit $N(A)$ abschätzen:

$$N(A) = \sum_{\substack{l,k=1 \\ l \neq k}}^{n} |a_{lk}|^2 \leq n(n-1)|a_{ij}|^2,$$

woraus folgt

$$a_{ij} \geq \frac{N(A)}{n(n-1)}.$$

Jetzt betrachten wir einen Iterationsschritt

$$N(B) = N(A) - 2|a_{ij}|^2 \leq \underbrace{\left(1 - \frac{2}{n(n-1)}\right)^1}_{=:q<1} N(A)$$

und stellen lineare Konvergenz fest, da der Exponent von q gleich 1 ist. QED

Zwar wissen wir nun, dass die Außennormen beim Jakobi-Verfahren gegen Null konvergieren, doch wissen wir nicht, ob dann auf der Diagonalen auch wirklich die Eigenwerte stehen. Dieses Problem wollen wir nun klären:

Korollar 5.16 *Sind $\lambda_1 \geq \cdots \geq \lambda_n$ die Eigenwerte der symmetrischen Matrix $A \in \mathbb{R}^{n\times n}$ und ist $\tilde{a}_{11}^{(m)} \geq \cdots \geq \tilde{a}_{nn}^{(m)}$ eine Umsortierung der Diagonalelemente von $A^{(m)}$, so gilt*

$$|\lambda_i - \tilde{a}_{ii}^{(m)}| \leq \sqrt{N(A^{(m)}} \to 0 \text{ für } m \to \infty.$$

Beweis: Aus Korollar 5.6 mit $A = A^{(m)}$ und $B = \operatorname{diag}(a_{11}^{(m)}, \ldots, a_{nn}^{(m)})$ sowie der euklidischen Norm erhalten wir, da A und $A^{(m)}$ die gleichen Eigenwerte besitzen:

$$|\lambda_i - \tilde{a}_{ii}^{(m)}| = |\lambda_i(A_m) - \lambda_i(B)| \leq \|A^{(m)} - B\|_2 \leq \|A^{(m)} - B\|_{\mathrm{F}} = \sqrt{N(A^{(m)}}.$$

QED

Auf die Eigenvektoren können wir schließen, da sich $A^{(m)}$ schreiben lässt als

$$A^{(m+1)} = G^{(m)} A^{(m)} G^{(m)^T} = \cdots = G^{(m)} \cdot \ldots \cdot G^{(1)} \cdot A \cdot G^{(1)^T} \cdot \ldots \cdot G^{(m)^T} =: Q^{(m)} A Q^{(m)^T},$$

wobei $Q^{(m)}$ orthogonal ist und $A^{(m+1)}$ näherungsweise diagonal. Also bestehen die Zeilen von $Q^{(m)}$ näherungsweise aus Eigenvektoren von A. Es gibt noch weitere Verfeinerungen des Verfahrens:

- Gerade, da das Aufsuchen des Maximums in jedem Schritt mit $n(n-1)$ Vergleichen $\mathcal{O}(n^2)$ wiegt, bei großen Matrizen sehr teuer sein kann. Beispielsweise kann man die Reihenfolge, in der die Paare (i, j) durchlaufen werden, vorher festlegen. Dies nennt man *zyklisches Jakobi-Verfahren.*
- Setzt man zusätzlich einen Schwellenwert, ab dem man sich mit dem a_{ij} zufrieden gibt, spricht man vom *zyklischen Jakobi-Verfahren mit Schwellenwert.*