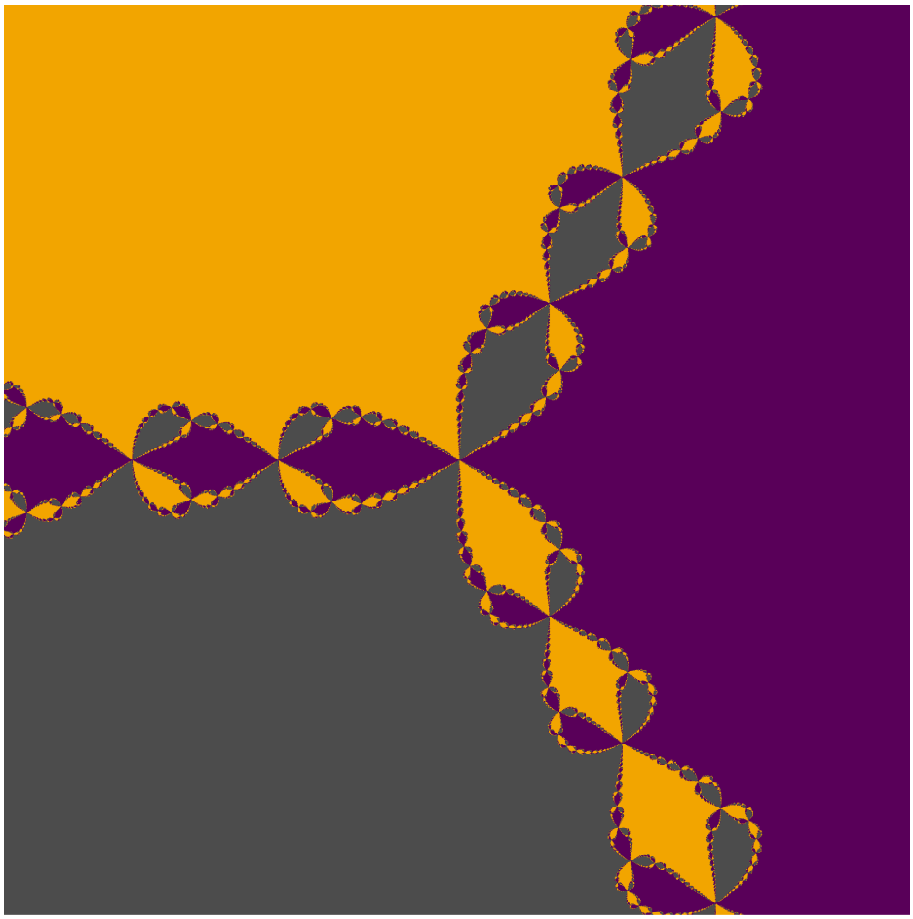
Kompendium zur Lehrveranstaltung

Numerische Mathematik II

Wintertrimester 2012



Mechthild Thalhammer

Nach einem Skriptum von Mathias Richter
zur Numerischen Mathematik II (Wintertrimester 2011)

Mathematik ist schön, aber für viele sperrig. Sie ist einfach nicht sozial*: In ihr kann nur der Einzelne zur Einsicht gelangen.* Rudolf Taschner

Das vorliegende Kompendium faßt die im Rahmen der dreistündigen Lehrveranstaltung **Numerische Mathematik II** im Wintertrimester 2012 an der Fakultät für Luft- und Raumfahrttechnik der Universität der Bundeswehr München besprochenen Themen zusammen. Die Inhalte des Kompendiums entsprechen weitgehend einem Skriptum von Mathias Richter zur Vorlesung Numerische Mathematik II, die im Wintertrimester 2011 abgehalten wurde.

Themenüberblick

1. Polynominterpolation

- Vorbemerkungen:
  - Ein wichtiges Gebiet der Numerischen Mathematik ist die Interpolation und Approximation von Funktionen. Man unterscheidet dabei den Fall univariater Funktionen (d.h. Funktionen in einer Veränderlichen) und den im Allgemeinen um ein Vielfaches schwierigeren Fall multivariater Funktionen (d.h. Funktionen in mehreren Veränderlichen).
  - Thema der Vorlesung ist insbesondere die Interpolation mittels Polynomen für univariate Funktionen. Vorteile der Verwendung von Polynomfunktionen sind ihre einfache Darstellung und mittels Horner-Schema rasche und stabile Auswertung (vgl. Numerische Mathematik I). Ein Nachteil der Polynominterpolation ist jedoch die schlechte Kondition bei höherem Polynomgrad.

    Anwendungen:
    - Numerische Verfahren für nichtlineare Gleichungen (z.B. Interpolation durch Polynome vom Grad 1 zur Herleitung des Sekantenverfahrens)
    - Numerische Integration (Newton–Côtes Formeln)
    - Numerische Verfahren für gewöhnliche Differentialgleichungen (Kollokationsverfahren, Adamsverfahren)
  - Eine vorteilhafte Alternative zur Polynominterpolation ist die Interpolation mittels Splines (stückweise Polynome, Zusatzbedingungen).
  - Eine Alternative zur Interpolation insbesondere bei fehlerbehafteten Daten ist die Approximation (Lineare Regression, Nichtlineare Regression).
- Vorsicht! Notationen unterscheiden sich teilweise von den im Skriptum verwendeten Notationen.

1.1. Aufgabenstellung und Notation

- Problemstellung: Zu vorgegebenen Datenpunkten, d.h. zu gegebenen reellen Stützstellen und zugehörigen reellen Stützwerten, wird eine Polynomfunktion gesucht, welche die Datenpunkte *interpoliert*. Die Stützwerte werden als Funktionswerte oder Ableitungen einer hinreichend oft differenzierbaren Funktion $f : [a,b] \to \mathbb{R}$ interpretiert.

  Vgl. Illustration, Skriptum, S. 15 (Zugrundeliegende Funktion und interpolierende Polynomfunktion bei äquidistanten bzw. nicht äquidistanten Stützstellen sowie interpolierende stückweise kubische Polynomfunktion bei äquidistanten Stützstellen).

- Bezeichnungen: Für $n \geq 0$ bezeichnet $\mathbb{P}_n$ den $(n+1)$-dimensionalen Vektorraum der reellen Polynome vom Grad $\leq n$ (bzw. von der Ordnung $n+1$), d.h. es gilt

$$\mathbb{P}_n = \{p : \mathbb{R} \to \mathbb{R} \text{ Polynom vom Grad } \leq n\}.$$

  Die Monome

$$p_i : \mathbb{R} \to \mathbb{R} : x \mapsto x^i, \qquad 0 \leq i \leq n,$$

  bilden die Taylor-Basis von $\mathbb{P}_n$.

  Bemerkung: Eine (unwesentliche) Verallgemeinerung ist die Betrachtung der Polynomfunktionen $p_i(x) = (x - x_0)^i$ für $0 \leq i \leq n$ (Reduktion durch Variablentransformation $x \mapsto x - x_0$, vgl. Taylorreihenentwicklung).

- Fragestellung: Präzisierung der Problemstellung (Behandlung einfacher und mehrfacher Stützstellen).

  Interpolationsaufgabe (Definition 1.1): Es sei $f : [a,b] \to \mathbb{R}$ eine gegebene Funktion und $x_i \in [a,b]$ für $0 \leq i \leq n$. Ein Polynom $p \in \mathbb{P}_n$ heißt Polynominterpolant von f zu den $n+1$ Stützstellen $(x_i)_{0 \leq i \leq n}$, wenn das Restglied $r = p - f$ die Nullstellen $x_i \in [a,b]$ für $0 \leq i \leq n$ besitzt, d.h. es gilt

$$r(x) = p(x) - f(x) = h(x) \prod_{i=0}^{n} (x - x_i)$$

  mit einer (differenzierbaren) Funktion $h : [a,b] \to \mathbb{R}$. Als Interpolationsaufgabe bezeichnet man das Problem, zu vorgegebener Funktion f und vorgegebenen Stützstellen $(x_i)_{0 \leq i \leq n}$ den zugehörigen Polynominterpolanten p zu bestimmen.

  Bemerkungen:

  - Für theoretische Überlegungen wird manchmal zusätzlich angenommen, daß die Stützstellen angeordnet sind, d.h. es ist $a \leq x_0 \leq \cdots \leq x_n \leq b$.
  - Im Fall einfacher Stützstellen, d.h. paarweise verschiedener Stützstellen

$$x_i \neq x_j, \qquad i \neq j, \quad 0 \leq i, j \leq n,$$

folgt mittels Produktregel

$$\begin{gathered} r(x) = p(x) - f(x) = \widetilde{h}(x)\,(x - x_i)\,, \qquad \widetilde{h}(x_i) \neq 0\,, \\ r(x_i) = p(x_i) - f(x_i) = 0\,, \\ r'(x) = p'(x) - f'(x) = \widetilde{h}'(x)\,(x - x_i) + \widetilde{h}(x)\,, \qquad r'(x_i) = p'(x_i) - f'(x_i) \neq 0\,, \end{gathered}$$

und somit

$$p(x_i) = f(x_i)\,, \qquad 0 \leq i \leq n\,.$$

Die Interpolationsaufgabe besteht also darin, zu vorgegebenen Daten $(x_i, y_i)_{0 \leq i \leq n}$ mit Stützwerten $y_i = f(x_i)$ für $0 \leq i \leq n$ ein Polynom $p \in \mathbb{P}_n$ zu bestimmen, welches die Interpolationsbedingungen erfüllt

$$p(x_i) = y_i\,, \qquad 0 \leq i \leq n\,.$$

– Im Fall zumindest einer mehrfachen Stützstelle spricht man von Hermite-Interpolation. Beispielsweise sei x_i eine m-fache Stützstelle (wobei $m \geq 2$, jedoch keine $(m+1)$-fache Stützstelle), d.h. bei angeordneten Stützstellen gelte

$$x_{i-1} < x_i = \cdots = x_{i+m-1} < x_{i+m}\,.$$

Da in diesem Fall das Restglied den Term $(x - x_i)^m$ enthält, folgt mittels Produktregel bzw. der Regel von Leibniz

$$\begin{gathered} r(x) = p(x) - f(x) = \widetilde{h}(x)\,(x - x_i)^m\,, \qquad \widetilde{h}(x_i) \neq 0\,, \\ r(x_i) = p(x_i) - f(x_i) = 0\,, \\ r'(x) = p'(x) - f'(x) = \widetilde{h}'(x)\,(x - x_i)^m + m\,\widetilde{h}(x)\,(x - x_i)^{m-1}\,, \\ r'(x_i) = p'(x_i) - f'(x_i) = 0\,, \\ r^{(k)}(x) = p^{(k)}(x) - f^{(k)}(x) = \sum_{j=0}^{k} \tfrac{k!\,m!}{j!(k-j)!(m-j)!}\,\widetilde{h}^{(k-j)}(x)\,(x - x_i)^{m-j}\,, \qquad 0 \leq k \leq m-1\,, \\ r^{(k)}(x_i) = p^{(k)}(x_i) - f^{(k)}(x_i) = 0\,, \qquad 0 \leq k \leq m-1\,, \end{gathered}$$

und somit

$$p^{(k)}(x_i) = f^{(k)}(x_i)\,, \qquad 0 \leq k \leq m-1\,.$$

Bei mehrfachen Nullstellen beinhalten die Interpolationsbedingungen neben Funktionswerten auch Ableitungen der zugehörigen Funktion.

- Fragestellung: Aussage zur Existenz und Eindeutigkeit des Polynominterpolanten (Konstruktion).

  Eindeutige Lösbarkeit der Interpolationsaufgabe (Satz 1.2): Die Lösung der Interpolationsaufgabe ist eindeutig bestimmt.

Denn: Zum Nachweis der Existenz und Eindeutigkeit des Interpolationspolynoms p wird die Darstellung nach Newton

$$\begin{aligned} p(x) &= \sum_{i=0}^{n} c_i \prod_{\ell=0}^{i-1} (x - x_\ell) \\ &= c_0 + c_1 (x - x_0) + c_2 (x - x_0)(x - x_1) + \cdots + c_n (x - x_0) \cdots (x - x_{n-1}) \end{aligned}$$

mit zu bestimmenden Koeffizienten $(c_i)_{0 \le i \le n}$ verwendet. In diesem Fall ergibt sich ein lineares Gleichungssystem, dessen eindeutige Lösbarkeit offensichtlich ist. Die *übliche* Darstellung einer Polynomfunktion mittels Taylor-Basis würde auf ein lineares Gleichungssystem für die Koeffizienten führen, dessen Lösbarkeit nicht ersichtlich ist.

– Bei einfachen Stützstellen führen die Interpolationsbedingungen

$$p(x_i) = y_i = f(x_i), \qquad 0 \le i \le n,$$

auf das lineare Gleichungssystem

$$\begin{aligned} y_0 &= p(x_0) = c_0, \\ y_1 &= p(x_1) = c_0 + c_1 (x_1 - x_0), \\ y_2 &= p(x_2) = c_0 + c_1 (x_2 - x_0) + c_2 (x_2 - x_0)(x_2 - x_1), \\ &\vdots \\ y_n &= p(x_n) = c_0 + c_1 (x_n - x_0) + c_2 (x_n - x_0)(x_n - x_1) + \cdots \\ &\qquad + c_n (x_n - x_0) \cdots (x_n - x_{n-1}), \end{aligned}$$

dessen Lösung $c = (c_0, \ldots, c_n)$ schrittweise berechnet werden kann ($n+1$ Gleichungen für $n+1$ Unbekannte, Lösung mittels Vorwärtseinsetzen, Koeffizient bei c_i gegeben durch $(x_i - x_0) \cdots (x_i - x_{i-1}) \neq 0$).

– Der allgemeine Fall mit mehrfachen Nullstellen beruht auf ähnlichen Ideen, allerdings ist es komplizierter, das resultierende lineare Gleichungssystem anzugeben. Beispielsweise bei einfachen Nullstellen x_0, x_2 und einer dreifachen Nullstelle x_1 führt das Einsetzen der Interpolationsbedingungen

$$\begin{aligned} p(x_i) = y_i = f(x_i), &\qquad i = 0, 1, 2, \\ p'(x_1) = y_1' = f'(x_1), &\qquad p''(x_1) = y_1'' = f''(x_1), \end{aligned}$$

in die Darstellung

$$p(x) = c_0 + c_1 (x - x_0) + c_2 (x - x_0)(x - x_1) + c_3 (x - x_0)(x - x_1)^2 + c_4 (x - x_0)(x - x_1)^3$$

auf das lineare Gleichungssystem

$$
\begin{aligned}
y_0 &= p(x_0) = c_0\,,\\
y_1 &= p(x_1) = c_0 + c_1\,(x_1 - x_0)\,,\\
y_1' &= p'(x_1) = c_1 + c_2\,(x_1 - x_0)\,,\\
y_1'' &= p''(x_1) = 2\,c_2 + 2\,c_3\,(x_1 - x_0)\,,\\
y_2 &= p(x_2) = c_0 + c_1\,(x_2 - x_0) + c_2\,(x_2 - x_0)\,(x_2 - x_1) + c_3\,(x_2 - x_0)\,(x_2 - x_1)^2\\
&\qquad + c_4\,(x_2 - x_0)\,(x_2 - x_1)^3\,,
\end{aligned}
$$

dessen Lösung schrittweise bestimmt werden kann (5 Gleichungen für 5 Unbekannte, Lösung mittels Vorwärtseinsetzen, Koeffizient bei c_i ungleich Null). ◇

Bemerkung: Zur Bestimmung der Koeffizienten des Interpolationspolynoms nach Newton verwendet man anstelle der Lösung des im Beweis von Satz 1.2 angegebenen Gleichungssystems üblicherweise das Schema von Aitken–Neville (vgl. Abschnitt 1.2).

- Fragestellung: Angabe des Polynominterpolanten (einfache Stützstellen).

  Lagrange-Polynome (Definition 1.3): Für $n+1$ paarweise verschiedene Stützstellen $(x_i)_{0\le i\le n}$ (d.h. $x_i \neq x_j$ für $i \neq j$ und $0 \le i, j \le n$) sind die Lagrange-Polynome $(L_j)_{0\le j\le n}$ definiert durch

  $$L_j(x) = \prod_{\substack{0\le i\le n\\ i\neq j}} \frac{x - x_i}{x_j - x_i}\,, \qquad 0 \le j \le n\,.$$

  Vgl. Illustration, Skriptum, S. 8 (Äquidistante Stützstellen $x_i = i$ für $0 \le i \le n = 20$, Langrange-Polynome L_{10} und L_{11}).

  Eigenschaften der Lagrange-Polynome:

  – Die Lagrange-Polynome sind Polynome vom Grad n und insbesondere gilt $L_j \in \mathbb{P}_n$ für $0 \le j \le n$.

  – Es gilt $L_j(x_i) = \delta_{ij}$ für $0 \le i, j \le n$. Insbesondere sind die Lagrange-Polynome $(L_j)_{0\le j\le n}$ linear unabhängig und bilden somit eine Basis des Vektorraumes $\mathbb{P}_n$.

  Lagrange-Interpolationsformel (Satz 1.4): Zu einfachen Stützstellen $(x_i)_{0\le i\le n}$ und zugehörigen Stützwerten $y_i = f(x_i)$ für $0 \le i \le n$ ist die eindeutig bestimmte Lösung der Interpolationsaufgabe gegeben durch

  $$p = \sum_{i=0}^{n} y_i\, L_i\,.$$

  Denn: Aufgrund der Basiseigenschaft der Lagrange-Polynome folgt für jedes Polynom $p \in \mathbb{P}_n$ die Darstellung

  $$p = \sum_{i=0}^{n} \alpha_i\, L_i$$

mit eindeutig bestimmten reellen Koeffizienten $(\alpha_i)_{0\le i\le n}$. Einsetzen der Stützstellen

$$y_j = p(x_j) = \sum_{i=0}^{n} \alpha_i \underbrace{L_i(x_j)}_{=\delta_{ij}} = \alpha_j$$

führt auf die angegebene Relation. ◇

Bemerkung: Die obige Darstellung des Interpolationspolynoms ist für theoretische Überlegungen wesentlich. Für praktische Berechnungen bei einer größeren Anzahl an Datenpunkten sind alternative Darstellungen vorzuziehen, da es aufgrund von stark anwachsenden Funktionswerten mit unterschiedlichem Vorzeichen zu numerisch ungünstigen Operationen und insbesondere zur Auslöschung signifikanter Stellen kommen kann.

1.2. Berechnung des Polynominterpolanten nach Aitken–Neville

- Situation: Es sei $f : [a,b] \to \mathbb{R}$ hinreichend oft differenzierbar, und es seien $(x_i)_{0\le i\le n}$ mit $x_i \in [a,b]$ für $0 \le i \le n$ einfache Stützstellen (und zusätzlich angeordnet).

  Fragestellung: Gesucht ist eine für numerische Berechnungen vorteilhafte Darstellung des Polynominterpolanten durch vorgegebene Datenpunkte (nach Satz 1.2 eindeutig bestimmt)

$$p(x_i) = y_i = f(x_i)\,, \qquad 0 \le i \le n\,.$$

  Der in Abschnitt 1.2 behandelte Zugang wird in Abschnitt 1.3 nochmals betrachtet und führt dann auf den bei praktischen Berechnungen verwendeten Zugang (Darstellung nach Newton, Berechnung der Koeffizienten des Interpolationspolynoms mittels Schema der dividierten Differenzen).

  Vorbemerkung: Zur Konstruktion des Polynominterpolanten durch $n+1$ Datenpunkte

$$p(x_i) = y_i = f(x_i)\,, \qquad 0 \le i \le n\,,$$

  wird einerseits der Polynominterpolant durch die ersten n Datenpunkte

$$q(x_i) = y_i = f(x_i)\,, \qquad 0 \le i \le n-1\,,$$

  und andererseits der Polynominterpolant durch die letzten n Datenpunkte

$$\widetilde{q}(x_i) = y_i = f(x_i)\,, \qquad 1 \le i \le n\,,$$

  betrachtet. Der Ansatz ($p \in \mathbb{P}_n$ wegen $q, \widetilde{q} \in \mathbb{P}_{n-1}$)

$$p(x) = c\,(x - x_n)\,q(x) + \widetilde{c}\,(x - x_0)\,\widetilde{q}(x)$$

  und Einsetzen des ersten bzw. letzten Datenpunktes

$$\begin{aligned} y_0 = p(x_0) = c\,(x_0 - x_n)\,q(x_0) = c\,(x_0 - x_n)\,y_0 \quad &\Rightarrow \quad c = -\tfrac{1}{x_n - x_0}\,, \\ y_n = p(x_n) = \widetilde{c}\,(x_n - x_0)\,\widetilde{q}(x_n) = \widetilde{c}\,(x_n - x_0)\,y_n \quad &\Rightarrow \quad \widetilde{c} = \tfrac{1}{x_n - x_0}\,, \end{aligned}$$

  führt auf die Darstellung

$$p(x) = \tfrac{1}{x_n - x_0}\,\big((x - x_0)\,\widetilde{q}(x) - (x - x_n)\,q(x)\big)\,.$$

  Wie gewünscht gilt in den Zwischenpunkten die Identität

$$\begin{aligned} p(x_i) &= \tfrac{1}{x_n - x_0}\,\big((x_i - x_0)\,\widetilde{q}(x_i) - (x_i - x_n)\,q(x_i)\big) \\ &= \tfrac{1}{x_n - x_0}\,\big((x_i - x_0)\,y_i - (x_i - x_n)\,y_i\big) = y_i\,, \qquad 1 \le i \le n-1\,. \end{aligned}$$

  Aufgrund der Eindeutigkeit des Polynominterpolanten ist damit die obige Darstellung nachgewiesen.

Das Schema von Aitken–Neville ist ein konstruktives Verfahren zur Berechnung des Polynominterpolanten, das auf der Idee beruht, den Interpolanten durch $n+1$ Datenpunkte auf zwei Interpolanten durch n Datenpunkte zurückzuführen (siehe Vorbemerkung). Es bezeichne P_j^k den Polynominterpolanten durch die Datenpunkte $(x_i, y_i)_{j \le i \le k}$

$$P_j^k(x_i) = y_i = f(x_i), \quad j \le i \le k.$$

Weiters sei P_j^{k-1} der Polynominterpolant durch die Datenpunkte $(x_i, y_i)_{j \le i \le k-1}$

$$P_j^{k-1}(x_i) = y_i = f(x_i), \quad j \le i \le k-1,$$

und P_{j+1}^k der Polynominterpolant durch die Datenpunkte $(x_i, y_i)_{j+1 \le i \le k}$

$$P_{j+1}^k(x_i) = y_i = f(x_i), \quad j+1 \le i \le k.$$

Wie zuvor führt der Ansatz

$$P_j^k(x) = c\,(x - x_k)\,P_j^{k-1}(x) + \tilde{c}\,(x - x_j)\,P_{j+1}^k(x)$$

und Einsetzen des ersten bzw. letzten Datenpunktes auf die Darstellung

$$P_j^k(x) = \tfrac{1}{x_k - x_j}\left((x - x_j)\,P_{j+1}^k(x) - (x - x_k)\,P_j^{k-1}(x)\right).$$

Wie gewünscht gilt in den Zwischenpunkten die Identität

$$P_j^k(x_i) = y_i, \qquad j+1 \le i \le k-1,$$

und aufgrund der Eindeutigkeit des Polynominterpolanten ist damit die obige Darstellung nachgewiesen. Diese Rekursion kann verwendet werden, den Polynominterpolanten $p = P_0^n$ ausgehend von den konstanten Funktionen $P_i^i = y_i$ für $0 \le i \le n$ schrittweise zu berechnen, vgl. Schema von Aitken–Neville und Algorithmus, Skriptum, S. 10.

Bemerkung: Im Fall einer mehrfachen Nullstelle $x_j = \cdots = x_k$ wird anstelle der zuvor angegebenen Relation die Darstellung von P_j^k mittels Taylorpolynom verwendet

$$P_j^k(x) = \sum_{\ell=0}^{k-j} \tfrac{1}{\ell!}\, f^{(\ell)}(x_j)\,(x - x_j)^\ell$$

und beim Schema von Aitken–Neville an der entsprechenden Stelle eingesetzt, vgl. Skriptum, S. 11.

1.3. Dividierte Differenzen

- Vorbemerkung: Bei Polynomfunktionen $p \in \mathbb{P}_n$ in der *üblichen* Darstellung mittels Taylor-Basis

$$p(x) = \sum_{i=0}^{n} \widetilde{c}_i \, x^i$$

ist die Kenntnis der Koeffizienten $(\widetilde{c}_i)_{0\le i\le n}$ zur Auswertung mittels Horner-Schema ausreichend. Ähnlich ist bei der Darstellung nach Newton

$$p(x) = \sum_{i=0}^{n} c_i \prod_{\ell=0}^{i-1} (x - x_\ell)$$

die Kenntnis der Stützstellen $(x_i)_{0\le i\le n}$ und der Koeffizienten $(c_i)_{0\le i\le n}$ ausreichend. Die in Abschnitt 1.2 hergeleitete Rekursion wird nun verwendet, um die Koeffizienten in dieser Darstellung des Polynominterpolanten zu berechnen, das Polynom wird dann mittels eines modifizierten Horner-Schemas ausgewertet. Bei praktischer Berechnung des Polynominterpolanten (höheren Grades) wird ausschließlich diese Vorgehensweise angewendet, die Überlegungen in Abschnitt 1.2 dienten zu Motivation, alternative Darstellungen z.B. mittels Lagrange-Basisfunktionen dienen vorwiegend theoretischen Überlegungen.

Dividierte Differenzen, Interpolationsformel nach Newton (Definition 1.5, Satz 1.6, Satz 1.7): Die Leitkoeffizienten der im Schema von Aitken–Neville auftretenden Polynome P_j^k heißen dividierte Differenzen. Bei einfachen Stützstellen $(x_i)_{0\le i\le n}$ sind die dividierten Differenzen rekursiv durch

$$\begin{aligned}
\delta y_{i,i+1} &= \tfrac{y_{i+1}-y_i}{x_{i+1}-x_i}, \qquad 0 \le i \le i+1 \le n, \\
\delta^2 y_{i,i+1,i+2} &= \tfrac{\delta y_{i+1,i+2}-\delta y_{i,i+1}}{x_{i+2}-x_i}, \qquad 0 \le i \le i+2 \le n, \\
\delta^{j+1} y_{i,\dots,i+j+1} &= \tfrac{\delta^j y_{i+1,\dots,i+j+1}-\delta^j y_{i,\dots,i+j}}{x_{i+j+1}-x_i}, \qquad 0 \le i \le i+j+1 \le n, \quad 0 \le j \le n-1,
\end{aligned}$$

gegeben; mit $\delta^0 y_i = y_i$ gilt zudem $\delta y_{i,i+1} = \frac{\delta^0 y_{i+1}-\delta^0 y_i}{x_{i+1}-x_i}$. Im Fall einer mehrfachen Stützstelle setzt man

$$x_i = \dots = x_{i+j}: \qquad \delta^j y_{i,\dots,i+j} = \tfrac{1}{(j-i)!} f^{(j-i)}(x_i), \qquad 0 \le i \le i+j \le n, \quad 0 \le j \le n.$$

Das Interpolationspolynom nach Newton ist durch

$$\begin{aligned}
p(x) &= \sum_{i=0}^{n} \delta^i y_{0,\dots,i} \prod_{\ell=0}^{i-1} (x - x_\ell) \\
&= y_0 + \delta y_{0,1}\,(x-x_0) + \delta^2 y_{0,1,2}\,(x-x_0)\,(x-x_1) + \dots + \delta^n y_{0,\dots,n}\,(x-x_0)\cdots(x-x_{n-1})
\end{aligned}$$

gegeben.

Schematische Darstellung:

x_i	y_i	$\delta y_{i,i+1}$	$\delta^2 y_{i,i+1,i+2}$	$\delta^3 y_{i,i+1,i+2,i+3}$
x_0	y_0			
		$\delta y_{0,1} = \frac{y_1 - y_0}{x_1 - x_0}$		
x_1	y_1		$\delta^2 y_{0,1,2} = \frac{\delta y_{1,2} - \delta y_{0,1}}{x_2 - x_0} = \frac{\frac{y_2 - y_1}{x_2 - x_1} - \frac{y_1 - y_0}{x_1 - x_0}}{x_2 - x_0}$	
		$\delta y_{1,2} = \frac{y_2 - y_1}{x_2 - x_1}$		$\delta^3 y_{0,1,2,3} = \frac{\delta^2 y_{1,2,3} - \delta^2 y_{0,1,2}}{x_3 - x_0}$
x_2	y_2		$\delta^2 y_{1,2,3} = \frac{\delta y_{2,3} - \delta y_{1,2}}{x_3 - x_1} = \frac{\frac{y_3 - y_2}{x_3 - x_2} - \frac{y_2 - y_1}{x_2 - x_1}}{x_3 - x_1}$	
		$\delta y_{2,3} = \frac{y_3 - y_2}{x_3 - x_2}$		
x_3	y_3			
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Einfaches Beispiel: Wählt man als Funktion das kubische Polynom

$$f(x) = 4x^3 + 3x^2 + 2x + 1$$

und die Stützstellen $-2, 0, 1, 2$, ergeben sich die Stützwerte

$$f(-2) = -23, \quad f(0) = 1, \quad f(1) = 10, \quad f(2) = 49.$$

Mittels des Schemas der dividierten Differenzen

x_i	y_i	$\delta y_{i,i+1}$	$\delta^2 y_{i,i+1,i+2}$	$\delta^3 y_{i,i+1,i+2,i+3}$
-2	-23			
		12		
0	1		-1	
		9		4
1	10		15	
		39		
2	49			

ergibt sich als Interpolationspolynom in der Darstellung nach Newton

$$p(x) = -23 + 12(x+2) - (x+2)x + 4(x+2)x(x-1),$$

und eine kleine Rechnung verifiziert $f = p$. Zur Auswertung des Interpolationspolynoms wird ein modifiziertes Horner-Schema verwendet. Vgl. Illustration (Polynominterpolation, einfaches Beispiel).

Bemerkung: Für theoretische Überlegungen ist es zweckmäßig , den Fall mehrfacher Nullstellen durch Einführung eines Inkrements ε und Grenzübergang $\varepsilon \to 0$ auf den Fall

einfacher Nullstellen zurückzuführen. Beispielsweise betrachtet man bei einer dreifachen Nullstelle die Stützstellen und Stützwerte (Taylorreihenentwicklung)

$$\begin{aligned} & x_i\,, \quad f(x_i)\,, \\ x_{i+1} = x_i + \varepsilon\,, \qquad & f(x_i+\varepsilon) = f(x_i) + \varepsilon\, f'(x_i) + \tfrac{1}{2}\varepsilon^2 f''(x_i) + \mathscr{O}\big(\varepsilon^3\big)\,, \\ x_{i+2} = x_i + 2\varepsilon\,, \qquad & f(x_i+2\varepsilon) = f(x_i) + 2\varepsilon\, f'(x_i) + 2\varepsilon^2 f''(x_i) + \mathscr{O}\big(\varepsilon^3\big)\,, \end{aligned}$$

und dann den Grenzübergang $\varepsilon \to 0$. Dies ergibt

$$\begin{aligned} \delta y_{i,i+1} &= \frac{y_{i+1}-y_i}{x_{i+1}-x_i} = \frac{f(x_i)+\varepsilon f'(x_i)+\frac{1}{2}\varepsilon^2 f''(x_i)+\mathscr{O}(\varepsilon^3)-f(x_i)}{\varepsilon} \\ &= f'(x_i) + \tfrac{1}{2}\varepsilon f''(x_i) + \mathscr{O}\big(\varepsilon^2\big) \quad \overset{\varepsilon\to 0}{\longrightarrow} \quad f'(x_i)\,, \\ \delta y_{i+1,i+2} &= \frac{y_{i+2}-y_{i+1}}{x_{i+2}-x_{i+1}} = \frac{f(x_i)+2\varepsilon f'(x_i)+2\varepsilon^2 f''(x_i)-f(x_i)-\varepsilon f'(x_i)-\frac{1}{2}\varepsilon^2 f''(x_i)+\mathscr{O}(\varepsilon^3)}{\varepsilon} \\ &= f'(x_i) + \tfrac{3}{2}\varepsilon f''(x_i) + \mathscr{O}\big(\varepsilon^2\big) \quad \overset{\varepsilon\to 0}{\longrightarrow} \quad f'(x_i)\,. \end{aligned}$$

Als Grenzwert der zweiten dividierten Differenz ergibt sich die zweite Ableitung $f''(x_i)$

$$\begin{aligned} \delta^2 y_{i,i+1,i+2} &= \frac{\delta y_{i+1,i+2}-\delta y_{i,i+1}}{x_{i+2}-x_i} = \frac{f'(x_i)+\frac{3}{2}\varepsilon f''(x_i)-f'(x_i)-\frac{1}{2}\varepsilon f''(x_i)+\mathscr{O}(\varepsilon^2)}{2\varepsilon} \\ &= \tfrac{1}{2} f''(x_i) + \mathscr{O}(\varepsilon) \quad \overset{\varepsilon\to 0}{\longrightarrow} \quad \tfrac{1}{2} f''(x_i)\,. \end{aligned}$$

- Restglied der Polynominterpolation: Ziel ist es, eine Aussage über die Güte der Approximation einer hinreichend oft differenzierbaren Funktion $f : [a,b] \to \mathbb{R}$ durch den Polynominterpolanten $p \in \mathbb{P}_n$ zu treffen, d.h. eine Relation bzw. Abschätzung für den Fehler der Polynominterpolation

$$r(x) = p(x) - f(x)\,, \qquad p(x) = \sum_{i=0}^{n} \delta^i y_{0,\dots,i} \prod_{\ell=0}^{i-1} (x - x_\ell)$$

abzuleiten.

Satz von Rolle (vgl. Analysis): Es sei $f : [a,b] \to \mathbb{R}$ stetig und auf (a,b) differenzierbar, und es gelte $f(a) = f(b)$. Dann existiert ein Element $c \in (a,b)$ mit $f'(c) = 0$.

Zusammenhang zwischen dividierten Differenzen und Ableitungen (Erweiterter Mittelwertsatz, Satz 1.8): Es sei $f \in \mathscr{C}^k([a,b])$. Dann existiert ein $\xi \in [a,b]$ mit

$$\delta^j y_{i,\dots,i+j} = \frac{1}{j!} f^{(j)}(\xi)\,, \qquad 0 \le i \le i+j \le n\,, \quad 0 \le j \le n\,.$$

Denn: Zum Beweis der Relation

$$\delta^n y_{0,\dots,n} = \frac{1}{n!} f^{(n)}(\xi)$$

wird verwendet, daß die Differenz $r = p - f$ (zumindest) $n+1$ Nullstellen $x_0, \cdots, x_n$ besitzt. Nach dem Satz von Rolle besitzt die erste Ableitung r' (zumindest) n Nullstellen

$\xi_i \in (x_i, x_{i+1})$ für $0 \le i \le n-1$. Die wiederholte Anwendung des Satzes von Rolle zusammen mit den Relationen

$$p^{(n)}(\xi) = n!\,\delta^n y_{0,\dots,n}\,, \qquad \frac{\mathrm{d}^n}{\mathrm{d}x^n}\,x^i = 0\,, \quad 0 \le i \le n-1\,, \qquad \frac{\mathrm{d}^n}{\mathrm{d}x^n}\,x^n = n!\,,$$

zeigen, daß die n-te Ableitung (zumindest) eine Nullstelle $\xi \in [a,b]$ besitzt, d.h. es gilt

$$0 = r^{(n)}(\xi) = p^{(n)}(\xi) - f^{(n)}(\xi) = n!\,\delta^n y_{0,\dots,n} - f^{(n)}(\xi)\,,$$

und damit folgt die Behauptung. ◇

Fehlerdarstellung der Polynominterpolation (Satz 1.9): Es sei $p \in \mathbb{P}_n$ der Polynominterpolant zu den Datenpunkten $(x_i, f(x_i))_{0\le i\le n}$, und es gelte $f \in \mathscr{C}^{n+1}([x_{\min}, x_{\max}])$ mit $x_{\min} = \min\{x_0,\dots,x_n,\bar{x}\}$ sowie $x_{\max} = \max\{x_0,\dots,x_n,\bar{x}\}$ für ein (fixiertes) Element $\bar{x} \in \mathbb{R}$. Dann existiert ein $\xi \in [x_{\min}, x_{\max}]$ derart, daß das Restglied die folgende Relation erfüllt

$$r(\bar{x}) = p(\bar{x}) - f(\bar{x}) = -\tfrac{1}{(n+1)!}\,f^{(n+1)}(\xi)\,\prod_{i=0}^{n}(\bar{x}-x_i)\,.$$

Denn: Der Polynominterpolant q zu den Datenpunkten $(x_i, f(x_i))_{0\le i\le n+1}$ mit $x_{n+1} = \bar{x}$ ist gegeben durch

$$q(x) = p(x) + \delta^{n+1} y_{0,\dots,n+1}\,\prod_{i=0}^{n}(x-x_i)\,.$$

Für $x = \bar{x}$ und mittels des obigen Resultates für die dividierten Differenzen ergibt sich

$$f(\bar{x}) = q(\bar{x}) = p(\bar{x}) + \tfrac{1}{(n+1)!}\,f^{(n+1)}(\xi)\,\prod_{i=0}^{n}(\bar{x}-x_i)$$

für ein $\xi \in [x_{\min}, x_{\max}]$ und damit die Behauptung. ◇

Bemerkungen:

- Das oben angegebene Resultat zeigt, daß die Größe des Restgliedes einerseits durch den funktionsabhängigen Beitrag

  $$\tfrac{1}{(n+1)!}\,f^{(n+1)}(\xi)$$

  und andererseits durch den von den Stützstellen abhängigen Beitrag

  $$\prod_{i=0}^{n}(\bar{x}-x_i)$$

  bestimmt ist.
- Als Maß für die Länge einer stetigen Funktion oder auch den Abstand zweier stetiger Funktionen betrachtet man die Supremumsnorm

  $$\|f\|_\infty = \max_{a\le x\le b}|f(x)|\,, \quad \|f-\widetilde{f}\|_\infty = \max_{a\le x\le b}|f(x)-\widetilde{f}(x)|\,, \qquad f,\widetilde{f} \in \mathscr{C}([a,b])\,.$$

– Unter der allerdings sehr einschränkenden Voraussetzung, daß sämtliche Ableitungen der Funktion $f : [a,b] \to \mathbb{R}$ beschränkt sind, konvergiert das Interpolationspolynom gleichmäßig gegen f. Genauer, falls eine Konstante $C > 0$ existiert, sodaß für alle $j \geq 0$ gilt

$$\left\| f^{(j)} \right\|_\infty = \max_{x \in [a,b]} \left| f^{(j)}(x) \right| \leq M ,$$

konvergiert das Interpolationspolynom $p_n \in \mathbb{P}_n$ zu (beliebigen) Stützstellen $(x_{ni})_{0 \leq i \leq n}$ mit $x_{ni} \in [a,b]$ für $0 \leq i \leq n$ bezüglich der Supremumsnorm gegen die Funktion f, d.h. es gilt

$$\left\| p_n - f \right\|_\infty = \max_{x \in [a,b]} \left| p_n(x) - f(x) \right| \quad \overset{n \to \infty}{\longrightarrow} \quad 0 .$$

Denn: Aus der oben angegebenen Darstellung für das Restglied folgt die Relation

$$\left| p_n(x) - f(x) \right| = \tfrac{1}{(n+1)!} \left| f^{(n+1)}(\xi) \right| \prod_{i=0}^{n} |x - x_{ni}| \leq C \, \tfrac{(b-a)^{n+1}}{(n+1)!} \quad \overset{n \to \infty}{\longrightarrow} \quad 0 , \qquad x \in [a,b] ,$$

und damit die Behauptung. ◇

– Bei der Interpolation von rationalen Funktionen zu äquidistanten Stützstellen kommt es zum Phänomen von Runge. Im Inneren des durch Polstellen von f festgelegten Intervalls konvergiert der Polynominterpolant $p_n \in \mathbb{P}_n$ für $n \to \infty$ gegen die Funktion f, im Äußeren des Intervalles treten jedoch starke Oszillationen auf und das Interpolationspolynom divergiert. Bekanntes Beispiel ist die rationale Funktion

$$f(x) = \tfrac{1}{1+x^2} , \qquad x \in [-5,5] .$$

Vgl. Illustration (Polynominterpolation, Phänomen von Runge).

– Vorbemerkung: Eine Funktion $f : [a,b] \to \mathbb{R}$ heißt Lipschitz-stetig, wenn eine Konstante $L > 0$ existiert, sodaß für alle Elemente $x, \tilde{x} \in [a,b]$ die Abschätzung

$$\left| f(x) - f(\tilde{x}) \right| \leq L |x - \tilde{x}|$$

gilt. Insbesondere ist jede stetig differenzierbare Funktion $f : [a,b] \to \mathbb{R}$ Lipschitz-stetig mit Konstante

$$L = \left\| f' \right\|_\infty = \max_{x \in [a,b]} \left| f'(x) \right| .$$

Interpolation basierend auf Chebychev-Knoten: Eine optimale Wahl der Stützstellen sind die Chebychev-Knoten

$$x_i = \tfrac{1}{2} \left(a + b + (b-a) \cos \tfrac{(2(i+1)-1)\pi}{2(n+1)} \right) , \qquad 0 \leq i \leq n ,$$

da sie den Beitrag der Stützstellen im Restglied minimieren

$$\max_{x \in [a,b]} \prod_{i=0}^{n} (x - x_i) \quad \to \quad \min .$$

In diesem Fall ist bereits für Lipschitz-stetige Funktionen $f : [a,b] \to \mathbb{R}$ die gleichmäßige Konvergenz des Polynominterpolanten gesichert

$$\left\| p_n - f \right\|_\infty = \max_{x \in [a,b]} \left| p_n(x) - f(x) \right| \quad \overset{n \to \infty}{\longrightarrow} \quad 0 \,.$$

Vgl. Illustration (Polynominterpolation, Chebychev-Knoten).

- Illustration zur Polynominterpolation mit äquidistanten Stützstellen und Chebychev-Knoten, vgl. Skriptum, S. 15.

- Vgl. Skriptum, S. 16: *In der Praxis — das heißt bei der Berechnung des Interpolanten auf einem Computer mit endlicher Stellenzahl — darf man trotzdem keine Polynome zu hohen Grads (größer 30) verwenden, da diese zu schlecht konditioniert sein können. Der Versuch etwa, im Rungeschen Beispiel ein Interpolationspolynom vom Grad 100 zu den Tschebyscheff-Stützstellen zu berechnen, resultiert in Datenschrott.*

1.4. Kondition der Polynominterpolation

- Kondition der Polynominterpolation (Satz 1.10): Die Polynominterpolation ist ein numerisches Problem, das vorgegebenen Stützstellen $(x_i)_{0\le i\le n}$ und Stützwerten $(y_i)_{0\le i\le n}$ mit $y_i = f(x_i)$ für $0 \le i \le n$ sowie einem Argument x den Funktionswert des Polynominterpolanten zuordnet

$$P : \mathbb{R}^{n+1} \times \mathbb{R}^{n+1} \times \mathbb{R} \to \mathbb{R} : (x_0, \dots, x_n, y_0, \dots, y_n, x) \mapsto p(x) .$$

  Die absoluten Konditionszahlen der Polynominterpolation sind durch

$$\partial_x P = p' , \quad \partial_{x_i} P = - p'(x_i)\, L_i , \quad \partial_{y_i} P = L_i , \qquad 0 \le i \le n ,$$

  gegeben.
  Denn: Zur Bestimmung der Konditionszahlen $\partial_{y_i} P$ und $\partial_{x_i} P$ wird die Darstellung mittels Langrange-Basis

$$p(x) = \sum_{i=0}^{n} y_i\, L_i(x) , \qquad L_i(x) = \prod_{\substack{0\le \ell\le n \\ \ell\ne i}} \frac{x - x_\ell}{x_i - x_\ell} , \qquad 0 \le j \le n ,$$

  verwendet. ◇

  Bemerkung: Bestimmend für die Kondition der Polynominterpolation ist die Größe L_i. Im ungünstigsten Fall verstärken sich Änderungen in den Stützwerten um den Faktor

$$\kappa = \sum_{i=0}^{n} |L_i| .$$

  Insbesondere bei äquidistanten Stützstellen kann dieser Faktor (vorallem am Rand des betrachteten Intervalls) große Werte annehmen, mit Chebychev-Knoten läßt sich hingegen ein vorteilhafteres Verhalten erzielen.
  Illustration (Werte von κ in $[0, 20]$ für äquidistante Stützstellen $x_i = i$ für $0 \le i \le 20$ und entsprechende Chebychev-Knoten), vgl. Skriptum, S. 18.

1.5. Vor- und Nachteile der Polynominterpolation

- Vorteile der Polynominterpolation:
  - Aufgrund der einfachen Handhabung sind Polynome für numerische Berechnungen gut geeignet, insbesondere sind die Berechnung und Auswertung des Polynominterpolanten einfach zu realisieren.
  - Polynominterpolanten besitzen gute lokale Approximationseigenschaften.

  Nachteile der Polynominterpolation:
  - Mit $\mathcal{O}(n^2)$ Operationen ist der Rechenaufwand der Polynominterpolation gegenüber $\mathcal{O}(n)$ Operationen bei der Spline-Interpolation erheblich.
  - In vielen Fällen ist die Polynominterpolation schlecht konditioniert.
  - Im Allgemeinen besitzen Polynominterpolanten schlechte globale Approximationseigenschaften.

  Die Berechnung eines Polynominterpolanten vom Grad ≥ 15 sowie ein Auswerten des Interpolanten in Randbereichen des betrachteten Intervalles sollten vermieden werden. Im Allgemeinen ist die Spline-Interpolation eine bessere Alternative.

2. Polynom-Splines

- Eine vorteilhafte Alternative zur Polynominterpolation, insbesondere in Situationen in denen der Polynominterpolant hohen Grad besitzt oder keine Chebychev-Knoten verwendet werden können, ist die Interpolation mittels Polynom-Splines, d.h. mittels stückweisen Polynomfunktionen, die zusätzlichen Stetigkeitsbedingungen genügen.

  Illustration (Vergleich Polynominterpolant und kubischer Splineinterpolant), vgl. Skriptum, S. 20.

- Vorsicht! Notationen unterscheiden sich teilweise von den im Skriptum verwendeten Notationen.

2.1. Kubische Splineinterpolanten

- Splinefunktionen sind stückweise Polynome vom Grad $k \geq 1$ bzw. von der Ordnung $k+1$ auf den jeweiligen Teilintervallen, die global $(k-1)$-mal stetig differenzierbar sind.

  Splinefunktionen (Definition 2.1): Für $m \geq 1$ sei $a = \tau_0 < \cdots < \tau_m = b$ eine Zerlegung des Intervalles $[a,b]$ in m Teilintervalle $[\tau_i, \tau_{i+1}]$ für $0 \leq i \leq m-1$. Der Raum der Splinefunktionen vom Grad $k \geq 0$ (bzw. von der Ordnung $k+1$) zu den $m+1$ Knoten $\tau = \{\tau_0, \ldots, \tau_m\}$ ist gegeben durch

$$\mathbb{P}_{0,\tau} = \left\{ s : [a,b) \to \mathbb{R} : s\big|_{[\tau_i,\tau_{i+1})} \in \mathbb{P}_0 \text{ für } 0 \leq i \leq m-1 \right\},$$
$$\mathbb{P}_{k,\tau} = \left\{ s : [a,b] \to \mathbb{R} \in \mathscr{C}^{k-1}([a,b]) : s\big|_{[\tau_i,\tau_{i+1}]} \in \mathbb{P}_k \text{ für } 0 \leq i \leq m-1 \right\}, \qquad k \geq 1.$$

  Bemerkungen:

  - Die Summe zweier Splinefunktionen $s, \tilde{s} \in \mathbb{P}_{k,\tau}$ und das skalare Vielfache einer Splinefunktion $s \in \mathbb{P}_{k,\tau}$ liegen in $\mathbb{P}_{k,\tau}$, d.h. der Raum der Splinefunktionen $\mathbb{P}_{k,\tau}$ bildet einen reellen Vektorraum. Geeignete Basisfunktionen sind B-Splines, vgl. Abschnitt 2.2.
  - Die stärkere Forderung $s \in \mathbb{P}_{k,\tau} \cap \mathscr{C}^k([a,b])$ würde implizieren, daß die Splinefunktion s global mit einer Polynomfunktion übereinstimmt.
    Denn: Für zwei Polynome vom Grad $k \geq 0$

$$p(x) = \sum_{i=0}^{k} p_i (x-a)^i, \qquad q(x) = \sum_{i=0}^{k} q_i (x-a)^i,$$

    deren Funktionswert und deren j-te Ableitung für $1 \leq j \leq k$ in einem Punkt übereinstimmen, folgt die Gleichheit der Koeffizienten (o.E.d.A. wähle $x = a$, ändere ansonsten die Darstellung der Polynomfunktionen)

$$j!\, p_j = p^{(j)}(a) = q^{(j)}(a) = j!\, q_j, \qquad 0 \leq j \leq k,$$

    und damit die Gleichheit der Funktionen. ◇

- Splinefunktionen vom Grad 0 bzw. von der Ordnung 1: Der Splineraum

$$\mathbb{P}_{0,\tau} = \left\{ s : [a,b) \to \mathbb{R} : s\big|_{[\tau_i,\tau_{i+1})} \in \mathbb{P}_0 \text{ für } 0 \leq i \leq m-1 \right\}$$

  umfaßt stückweise konstante Funktionen

$$s\big|_{[\tau_i,\tau_{i+1})}(x) = \alpha_i, \qquad 0 \leq i \leq m-1,$$

  die bei insgesamt $m+1$ Knoten durch m vorgeschriebene Stützwerte beispielsweise an den Knoten $\tau_0, \ldots, \tau_{m-1}$ bestimmt sind. Im Allgemeinen sind die Splinefunktionen unstetig, mit Sprungstellen an den Knoten; der Funktionswert bei $\tau_m = b$ wird nicht miteinbezogen.

  Illustration (Splinefunktion vom Grad 0), vgl. Skriptum, S. 22.

- Splinefunktionen vom Grad 1 bzw. von der Ordnung 2: Der Splineraum

$$\mathbb{P}_{1,\tau} = \left\{ s : [a,b] \to \mathbb{R} \in \mathscr{C}([a,b]) : s\big|_{[\tau_i,\tau_{i+1}]} \in \mathbb{P}_1 \text{ für } 0 \le i \le m-1 \right\}$$

umfaßt stückweise lineare Polynome

$$s\big|_{[\tau_i,\tau_{i+1}]}(x) = \alpha_i + \beta_i\,(x-\tau_i)\,, \qquad 0 \le i \le m-1\,,$$

die bei insgesamt $m+1$ Knoten durch $m+1$ vorgeschriebene Stützwerte beispielsweise an den Knoten $\tau_0,\ldots,\tau_m$ bestimmt sind. Die Funktionen sind auf dem Intervall $[a,b]$ stetig, im Allgemeinen jedoch an den Knoten nicht differenzierbar.

Illustration (Splinefunktion vom Grad 1), vgl. Skriptum, S. 22.

- Splinefunktionen vom Grad 2 bzw. von der Ordnung 3: Der Splineraum

$$\mathbb{P}_{2,\tau} = \left\{ s : [a,b] \to \mathbb{R} \in \mathscr{C}^1([a,b]) : s\big|_{[\tau_i,\tau_{i+1}]} \in \mathbb{P}_2 \text{ für } 0 \le i \le m-1 \right\}$$

umfaßt stückweise quadratische Polynome

$$s_i(x) = s\big|_{[\tau_i,\tau_{i+1}]}(x) = \alpha_i + \beta_i\,(x-\tau_i) + \gamma_i\,(x-\tau_i)\,(x-\tau_{i+1})\,, \qquad 0 \le i \le m-1\,,$$

die auf dem Intervall $[a,b]$ stetig differenzierbar sind, d.h. an den inneren Knoten die Bedingungen

$$s_{i-1}(\tau_i) = s_i(\tau_i)\,, \quad s'_{i-1}(\tau_i) = s'_i(\tau_i)\,, \qquad 1 \le i \le m-1\,,$$

erfüllen.

Bemerkung: Aufgrund der unerwünschten Eigenschaft, daß sich Änderungen in einem einzelnen Teilintervall (ungedämpft) auf die gesamte Splinefunktion auswirken, werden quadratische Splines selten verwendet. Dies zeigt sich beispielsweise an der Splinefunktion mit Funktionswert Null in sämtlichen äquidistant verteilten Knoten (mit $\tau_i = i\,h$ für $0 \le i \le m = 3$). Einsetzen der Darstellungen (zusätzliche Skalierungen der Koeffizienten vorteilhaft)

$$\begin{aligned}
s_0(x) &= \alpha_0 + \frac{\beta_0}{h}\,x + \frac{\gamma_0}{h}\,x\,(x-h)\,, && x \in [0,h]\,,\\
s_1(x) &= \alpha_1 + \frac{\beta_1}{h}\,(x-h) + \frac{\gamma_1}{h}\,(x-h)\,(x-2\,h)\,, && x \in [h,2\,h]\,,\\
s_2(x) &= \alpha_2 + \frac{\beta_2}{h}\,(x-2\,h) + \frac{\gamma_2}{h}\,(x-2\,h)\,(x-3\,h)\,, && x \in [2\,h,3\,h]\,,
\end{aligned}$$

in die Bedingungen an die Funktionswerte

$$\begin{aligned}
&\alpha_0 = s_0(0) = 0\,, \quad \beta_0 = s_0(h) = 0 = s_1(h) = \alpha_1\,,\\
&\beta_1 = s_1(2\,h) = 0 = s_2(2\,h) = \alpha_2\,, \quad \beta_2 = s_2(3\,h) = 0\,,
\end{aligned}$$

führt auf die Relationen

$$\begin{aligned}
s_0(x) &= \frac{\gamma_0}{h}\,x\,(x-h)\,, && x \in [0,h]\,,\\
s_1(x) &= \frac{\gamma_1}{h}\,(x-h)\,(x-2\,h)\,, && x \in [h,2\,h]\,,\\
s_2(x) &= \frac{\gamma_2}{h}\,(x-2\,h)\,(x-3\,h)\,, && x \in [2\,h,3\,h]\,.
\end{aligned}$$

Die Stetigkeitsbedingungen an die ersten Ableitungen

$$\begin{aligned} s_0'(x) &= \tfrac{\gamma_0}{h}\,(2\,x-h)\,, \qquad x\in[0,h]\,,\\ s_1'(x) &= \tfrac{\gamma_1}{h}\,(2\,x-3\,h)\,, \qquad x\in[h,2\,h]\,,\\ s_2'(x) &= \tfrac{\gamma_2}{h}\,(2\,x-5\,h)\,, \qquad x\in[2\,h,3\,h]\,,\\ \gamma_0 = s_0'(h) = s_1'(h) &= -\gamma_1\,, \quad \gamma_1 = s_1'(2\,h) = s_2'(2\,h) = -\gamma_2\,, \end{aligned}$$

ergeben weiters

$$\begin{aligned} s_0(x) &= \tfrac{\gamma_0}{h}\,x\,(x-h)\,, \qquad x\in[0,h]\,,\\ s_1(x) &= -\tfrac{\gamma_0}{h}\,(x-h)\,(x-2\,h)\,, \qquad x\in[h,2\,h]\,,\\ s_2(x) &= \tfrac{\gamma_0}{h}\,(x-2\,h)\,(x-3\,h)\,, \qquad x\in[2\,h,3\,h]\,. \end{aligned}$$

Die Änderung des Koeffizienten γ_0 bei der Funktion im ersten Teilintervall bewirkt somit eine Änderung der Funktionen auf allen anderen Teilintervallen.

- Splinefunktionen vom Grad 3 bzw. von der Ordnung 4: Der Splineraum

$$\mathbb{P}_{3,\tau} = \left\{ s : [a,b] \to \mathbb{R} \in \mathscr{C}^2([a,b]) : s\big|_{[\tau_i,\tau_{i+1}]} \in \mathbb{P}_3 \text{ für } 0 \le i \le m-1 \right\}$$

umfaßt stückweise kubische Polynome, die auf dem Intervall $[a,b]$ zweimal stetig differenzierbar sind.

  - Ziel ist es, die Funktionen

$$s_i(x) = s\big|_{[\tau_i,\tau_{i+1}]}(x)\,, \qquad 0\le i\le m-1\,,$$

    beispielsweise bei vorgegebenen Funktionswerten in den Knoten und vorgegeben Werten für die erste Ableitung in den Randpunkten

$$s_i(\tau_i) = y_i\,, \quad 0\le i\le m\,, \qquad s_0'(\tau_0) = y_0'\,, \quad s_{m-1}'(\tau_m) = y_m'\,,$$

    zu bestimmen. Dazu verwendet man zunächst, daß das kubische Polynom s_i durch die Funktionswerte und die (noch unbekannten) Werte der ersten Ableitung an den Knoten τ_i und τ_{i+1} für $0\le i\le m-1$ eindeutig festgelegt ist

$$s_i(\tau_i) = y_i\,, \quad s_i(\tau_{i+1}) = y_{i+1}\,, \quad s_i'(\tau_i) = y_i'\,, \quad s_i'(\tau_{i+1}) = y_{i+1}'\,, \qquad 0\le i\le m-1\,.$$

    Damit sind die Stetigkeitsbedingungen an den inneren Knoten für die Splinefunktion und deren erste Ableitung erfüllt. Die Darstellung (wobei $h_i = \tau_{i+1}-\tau_i$ für $0\le i\le m-1$)

$$\begin{aligned} s_i(x) = {} & y_i\left(1+2\,\tfrac{x-\tau_i}{h_i}\right)\left(\tfrac{\tau_{i+1}-x}{h_i}\right)^2 + y_{i+1}\left(3-2\,\tfrac{x-\tau_i}{h_i}\right)\left(\tfrac{x-\tau_i}{h_i}\right)^2\\ & + y_i'\left(\tfrac{\tau_{i+1}-x}{h_i}\right)^2(x-\tau_i) - y_{i+1}'\,\tfrac{\tau_{i+1}-x}{h_i}\,\tfrac{x-\tau_i}{h_i}\,(x-\tau_i)\,, \qquad 0\le i\le m-1\,, \end{aligned}$$

verifiziert man leicht durch Einsetzen

$$z(x) = \tfrac{x-\tau_i}{h_i}, \quad z(\tau_i) = 0, \quad z(\tau_{i+1}) = 1,$$
$$\widetilde{z}(x) = \tfrac{\tau_{i+1}-x}{h_i}, \quad \widetilde{z}(\tau_i) = 1, \quad \widetilde{z}(\tau_{i+1}) = 0,$$
$$s_i = y_i\left(\widetilde{z}^2 + 2\,z\,\widetilde{z}^2\right) + y_{i+1}\left(3\,z^2 - 2\,z^3\right) + y'_i\,h_i\,z\,\widetilde{z}^2 - y'_{i+1}\,h_i\,z^2\,\widetilde{z},$$
$$s_i(\tau_i) = y_i, \quad s_i(\tau_{i+1}) = y_{i+1}, \qquad 0 \le i \le m-1,$$

bzw. Differenzieren und Einsetzen (wegen $z' = \frac{1}{h_i} = -\widetilde{z}'$)

$$s'_i = \tfrac{2}{h_i}\,y_i\left(\widetilde{z}^2 - \widetilde{z} - 2\,z\,\widetilde{z}\right) + \tfrac{6}{h_i}\,y_{i+1}\left(z - z^2\right) + y'_i\left(\widetilde{z}^2 - 2\,z\,\widetilde{z}\right) + y'_{i+1}\left(z^2 - 2\,z\,\widetilde{z}\right),$$
$$s'_i(\tau_i) = y'_i, \quad s'_i(\tau_{i+1}) = y'_{i+1}, \qquad 0 \le i \le m-1.$$

– Aus der obigen Darstellung erhält man folgende Relationen für die Werte der zweiten Ableitung an den inneren Knoten

$$s''_i = \tfrac{2}{h_i^2}\,y_i\left(1 + 2\,z - 4\,\widetilde{z}\right) + \tfrac{6}{h_i^2}\,y_{i+1}\left(1 - 2\,z\right) + \tfrac{2}{h_i}\,y'_i\left(z - 2\,\widetilde{z}\right) + \tfrac{2}{h_i}\,y'_{i+1}\left(2\,z - \widetilde{z}\right),$$
$$s''_i(\tau_i) = \tfrac{6}{h_i^2}\left(y_{i+1} - y_i\right) - \tfrac{2}{h_i}\left(2\,y'_i + y'_{i+1}\right), \qquad 1 \le i \le m-1,$$
$$s''_i(\tau_{i+1}) = \tfrac{6}{h_i^2}\left(y_i - y_{i+1}\right) + \tfrac{2}{h_i}\left(y'_i + 2\,y'_{i+1}\right), \qquad 0 \le i \le m-2.$$

Die Stetigkeitsbedingungen an den inneren Knoten für die zweite Ableitung der Splinefunktion ergeben

$$s''_{i-1}(\tau_i) = s''_i(\tau_i), \qquad 1 \le i \le m-1,$$
$$\tfrac{3}{h_{i-1}^2}\left(y_{i-1} - y_i\right) + \tfrac{1}{h_{i-1}}\left(y'_{i-1} + 2\,y'_i\right) = \tfrac{3}{h_i^2}\left(y_{i+1} - y_i\right) - \tfrac{1}{h_i}\left(2\,y'_i + y'_{i+1}\right), \qquad 1 \le i \le m-1,$$

und weiters

$$\tfrac{1}{h_{i-1}}\,y'_{i-1} + 2\left(\tfrac{1}{h_{i-1}} + \tfrac{1}{h_i}\right)y'_i + \tfrac{1}{h_i}\,y'_{i+1} = \tfrac{3}{h_{i-1}^2}\left(y_i - y_{i-1}\right) + \tfrac{3}{h_i^2}\left(y_{i+1} - y_i\right), \quad 1 \le i \le m-1.$$

Aus diesen Relationen lassen sich bei vorgegebenen Funktionswerten an den Knoten und vorgegeben Werten für die erste Ableitung in den Intervallenden die zu bestimmenden Werte der ersten Ableitung in den inneren Knoten berechnen.

Berechnung eingespannter kubischer Splinefunktionen (Satz 2.2, Satz 2.3): Eine kubische Splinefunktion $s \in \mathbb{P}_{3,\tau}$ ist durch die Darstellung

$$s_i(x) = s\big|_{[\tau_i,\tau_{i+1})}(x) = y_i\left(1 + 2\,\tfrac{x-\tau_i}{h_i}\right)\left(\tfrac{\tau_{i+1}-x}{h_i}\right)^2 + y_{i+1}\left(3 - 2\,\tfrac{x-\tau_i}{h_i}\right)\left(\tfrac{x-\tau_i}{h_i}\right)^2$$
$$+ y'_i\left(\tfrac{\tau_{i+1}-x}{h_i}\right)^2(x-\tau_i) - y'_{i+1}\,\tfrac{\tau_{i+1}-x}{h_i}\,\tfrac{x-\tau_i}{h_i}\,(x-\tau_i), \qquad 0 \le i \le m-1,$$

gegeben. Bei Vorgabe von $m+1$ zu interpolierenden Funktionswerten

$$s_i(\tau_i) = y_i = f(\tau_i), \qquad 0 \le i \le m,$$

lauten die $m-1$ Bedingungen an die $m+1$ Ableitungswerte an den Knoten

$$s_i'(\tau_i) = y_i', \qquad 0 \le i \le m,$$
$$\frac{1}{h_{i-1}}\, y_{i-1}' + 2\left(\frac{1}{h_{i-1}} + \frac{1}{h_i}\right) y_i' + \frac{1}{h_i}\, y_{i+1}' = \frac{3}{h_{i-1}^2}\left(y_i - y_{i-1}\right) + \frac{3}{h_i^2}\left(y_{i+1} - y_i\right), \qquad 1 \le i \le m-1.$$

Beispielsweise bei der Festlegung der Werte der ersten Ableitung an den Intervallenden

$$s_0'(\tau_0) = y_0' = f'(a), \qquad s_{m-1}'(\tau_m) = y_m' = f'(b),$$

ist der Vektor $y' = (y_i')_{1 \le i \le m-1}$ als eindeutige Lösung eines linearen Gleichungssystems gegeben und damit die zugehörige kubische Splinefunktion (eingespannter Spline) eindeutig bestimmt.
Denn: Bei Vorgabe der Funktionswerte $y_i = f(\tau_i)$ für $0 \le i \le m$ und der Ableitungen $y_0' = f'(a)$ und $y_m' = f'(b)$ sind die restlichen Ableitungen $y_1', \ldots, y_{m-1}'$ als Lösungen der linearen Gleichungen

$$\begin{aligned}
i = 1: \quad & 2\left(\tfrac{1}{h_0} + \tfrac{1}{h_1}\right) y_1' + \tfrac{1}{h_1}\, y_2' = \tfrac{3}{h_0^2}\left(y_1 - y_0\right) + \tfrac{3}{h_1^2}\left(y_2 - y_1\right) - \tfrac{1}{h_0}\, y_0', \\
2 \le i \le m-2: \quad & \tfrac{1}{h_{i-1}}\, y_{i-1}' + 2\left(\tfrac{1}{h_{i-1}} + \tfrac{1}{h_i}\right) y_i' + \tfrac{1}{h_i}\, y_{i+1}' = \tfrac{3}{h_{i-1}^2}\left(y_i - y_{i-1}\right) + \tfrac{3}{h_i^2}\left(y_{i+1} - y_i\right), \\
i = m-1: \quad & \tfrac{1}{h_{m-2}}\, y_{m-2}' + 2\left(\tfrac{1}{h_{m-2}} + \tfrac{1}{h_{m-1}}\right) y_{m-1}' = \tfrac{3}{h_{m-2}^2}\left(y_{m-1} - y_{m-2}\right) \\
& \qquad + \tfrac{3}{h_{m-1}^2}\left(y_m - y_{m-1}\right) - \tfrac{1}{h_{m-1}}\, y_m',
\end{aligned}$$

gegeben. In kompakter Form als linearen Gleichungssystems mit symmetrischer Tridiagonalmatrix ergibt sich

$$Ay' = r, \qquad A = (a_{ij})_{1 \le i,j \le m-1} \in \mathbb{R}^{(m-1)\times(m-1)}, \quad y', r \in \mathbb{R}^{m-1},$$
$$a_{ii} = 2\left(\tfrac{1}{h_{i-1}} + \tfrac{1}{h_i}\right), \quad 1 \le i \le m-1, \qquad a_{i,i+1} = a_{i+1,i} = \tfrac{1}{h_i}, \quad 1 \le i \le m-2,$$
$$y = \begin{pmatrix} y_1' \\ \vdots \\ y_{m-1}' \end{pmatrix}, \qquad r = \begin{pmatrix} \frac{3}{h_0^2}\left(y_1 - y_0\right) + \frac{3}{h_1^2}\left(y_2 - y_1\right) \\ \vdots \\ \frac{3}{h_{i-1}^2}\left(y_i - y_{i-1}\right) + \frac{3}{h_i^2}\left(y_{i+1} - y_i\right) \\ \vdots \\ \frac{3}{h_{m-2}^2}\left(y_{m-1} - y_{m-2}\right) + \frac{3}{h_{m-1}^2}\left(y_m - y_{m-1}\right) \end{pmatrix} - \begin{pmatrix} \frac{1}{h_0}\, y_0' \\ 0 \\ \vdots \\ 0 \\ \frac{1}{h_{m-1}}\, y_m' \end{pmatrix}.$$

Da die Matrix A strikt diagonaldominant ist, ist das lineare Gleichungssystem eindeutig lösbar und damit die zugehörige kubische Splinefunktion eindeutig bestimmt. ◇

Bemerkungen:

– Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt strikt diagonaldominant, wenn folgende Relation gilt

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \ne i}}^{n} |a_{ij}| \qquad \text{für alle} \quad 1 \le i \le n.$$

Wegen (mit betragsmäßig größtem Koeffizienten x_ℓ)

$$\begin{aligned}
Ax = 0 \quad &\Longrightarrow \quad \sum_{j=1}^{n} a_{ij} x_j = 0 \quad \text{für} \quad 1 \le i \le n \\
&\Longrightarrow \quad a_{\ell\ell} x_\ell = -\sum_{\substack{j=1 \\ j\neq\ell}}^{n} a_{\ell j}\, x_j \quad \text{mit} \quad |x_\ell| = \max_{1\le i\le n} |x_i| \\
&\Longrightarrow \quad |a_{\ell\ell}|\,|x_\ell| \le \sum_{\substack{j=1 \\ j\neq\ell}}^{n} |a_{\ell j}|\,|x_j| \le |x_\ell| \sum_{\substack{j=1 \\ j\neq\ell}}^{n} |a_{\ell j}| \\
&\Longrightarrow \quad \underbrace{\Big(|a_{\ell\ell}| - \sum_{\substack{j=1 \\ j\neq\ell}}^{n} |a_{\ell j}|\Big)}_{>0} |x_\ell| \le 0 \\
&\Longrightarrow \quad x_\ell = 0 \\
&\Longrightarrow \quad x = 0
\end{aligned}$$

ist jede strikt diagonaldominante Matrix invertierbar.

– Beinhaltet ein lineares Gleichungssystem eine symmetrische und positiv definite Matrix $A \in \mathbb{R}^{n\times n}$, so ist das Gaußsche Eliminationsverfahren ohne Spaltenpivotsuche numerisch stabil durchführbar. Ist die Matrix zudem eine Tridiagonalmatrix, so werden zur Elimination der Subdiagonale $\mathcal{O}(n)$ Operationen benötigt.

– Vgl. Illustration (Eingespannter Spline, Matrix) für den Spezialfall äquidistanter Knoten ($m = 6$). Für die zugehörige Matrix

$$Ay' = r\,, \qquad A = (a_{ij})_{1\le i,j\le 5} \in \mathbb{R}^{(m-1)\times(m-1)}\,,$$
$$a_{ii} = \tfrac{4}{h}\,, \quad 1 \le i \le m-1\,, \qquad a_{i,i+1} = a_{i+1,i} = \tfrac{1}{h}\,, \quad 1 \le i \le m-2\,.$$

Das Gaußsche Eliminationsverfahren (bzw. die Cholesky-Zerlegung) führt auf die Zerlegung $A = \widetilde{L} D \widetilde{L}^T$ mit positiv definiter Diagonalmatrix D (d.h. alle Diagonaleinträge sind positiv) und auf einfache Weise rekursiv berechenbarer Matrix $\widetilde{L}$ (Diagonalelemente von $\widetilde{L}$ gleich Eins, erste Subdiagonale zu berechnen). Wegen $x^T Ax = x^T \widetilde{L} D \widetilde{L}^T x = y^T D y > 0$ mit $y = \widetilde{L}^T x$ folgt die positive Definitheit von A Ähnliche Überlegungen gelten für den allgemeinen Fall.

Fehler der kubischen Splineinterpolation (Satz 2.4): Für $f \in \mathcal{C}^4([a,b])$ sei $s : [a,b] \to \mathbb{R}$ die eindeutig bestimmte kubische Splinefunktion zu den Funktionswerten

$$s_i(\tau_i) = y_i = f(\tau_i)\,, \quad 0 \le i \le m\,, \qquad s'(a) = f'(a)\,, \quad s'(b) = f'(b)\,.$$

Dann gelten folgende Abschätzungen für den Approximationsfehler des Splineinterpolanten (zu $m+1$ Knoten $a = \tau_0 < \cdots < \tau_m = b$ mit $h_i = \tau_{i+1} - \tau_i$ für $0 \le i \le m-1$

und $h_{\max} = \max\{h_i : 0 \le i \le m-1\}$, $\|f\|_\infty = \max\{f(x) : a \le x \le b\}$)

$$
\begin{aligned}
\big|(f-s)(x)\big| &\le \tfrac{3}{64}\, h_i^2\, h_{\max}^2 \big\|f^{(4)}\big\|_\infty, && x \in [\tau_i, \tau_{i+1}], \quad 0 \le i \le m-1, \\
\big|(f-s)'(x)\big| &\le \tfrac{3}{16}\, h_i\, h_{\max}^2 \big\|f^{(4)}\big\|_\infty, && x \in [\tau_i, \tau_{i+1}], \quad 0 \le i \le m-1, \\
\big|(f-s)''(x)\big| &\le \tfrac{3}{8}\, h_{\max}^2 \max_{x\in[a,b]} \big\|f^{(4)}\big\|_\infty, && x \in [\tau_i, \tau_{i+1}], \quad 0 \le i \le m-1, \\
\big|(f-s)'''(x)\big| &\le \tfrac{1}{2}\, h_i \left(1 + \left(\tfrac{h_{\max}}{h_i}\right)^2\right) \big\|f^{(4)}\big\|_\infty, && x \in [\tau_i, \tau_{i+1}], \quad 0 \le i \le m-1.
\end{aligned}
$$

Bemerkungen:

- Die Approximationseigenschaften von kubischen Splinefunktionen für reguläre Funktionen sind insbesondere bei äquidistanten Stützstellen besser als die Eigenschaften von Polynominterpolanten.
- Werden die benötigten Werte der Ableitung an den Intervallenden näherungsweise mittels kubischer Polynominterpolation an vier Knoten bestimmt, bleibt der Approximationsfehler von der Größenordnung $h_{\max}^4$.
- Neben eingespannten Splinefunktionen mit vorgegebenen Ableitungen an den Intervallenden sind beispielsweise auch Splinefunktionen zu natürlichen Randbedingungen $s_0''(a) = 0 = s_{m-1}''(b)$ gebräuchlich, allerdings besitzen diese weniger gute Approximationseigenschaften.
- Splinefunktionen kommen ursprünglich aus dem Schiffsbau und sind wesentlich für Anwendungen aus den Bereichen Computergraphik und Design (u.a. Computeranimationen, Design von Autos und Flugzeugen).

- Vgl. Illustration (Splineinterpolation) sowie Illustration (Approximationsfehler).

Kurveninterpolation

- Die bisher behandelten Ideen lassen sich auch zur Interpolation von Kurven anwenden. Beispielsweise für eine ebene Kurve beruht die Interpolation mittels kubischen Splinefunktionen auf folgendem Zugang:
  - Eine ebene Kurve durch $m+1$ Punkte der Ebene $v_i = (x_i, y_i)_{0 \le i \le m}$ ist durch eine Parametrisierung gegeben

    $$\gamma : [a,b] \to \mathbb{R}^2 : t \mapsto \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}.$$

    Als Knoten $a = \tau_0 < \cdots < \tau_m = b$ wählt man beispielsweise die Zeitpunkte (Zeitinkremente entsprechen der Länge der verbindenden Polygonzüge $\| v_i - v_{i-1} \|_2$)

    $$\tau_0 = a, \qquad \tau_i = \tau_{i-1} + \| v_i - v_{i-1} \|_2, \quad 1 \le i \le m,$$
    $$\| v_i - v_{i-1} \|_2 = \| (x_i - x_{i-1}, y_i - y_{i-1})^T \|_2 = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2},$$

    und setzt dann $b = \tau_m$.

    Vgl. Illustration (Punkte der Ebene, zugehöriges Stützpolygon), Skriptum, S. 27.
  - Man bestimmt einerseits die interpolierende Splinefunktion $s_x : [a,b] \to \mathbb{R}$ zu den Datenpunkten $(\tau_i, x_i)_{0 \le i \le m}$ und andererseits die interpolierende Splinefunktion $s_y : [a,b] \to \mathbb{R}$ zu den Datenpunkten $(\tau_i, y_i)_{0 \le i \le m}$. Die interpolierende Kurve ist dann durch

    $$s : [a,b] \to \mathbb{R}^2 : t \mapsto \begin{pmatrix} s_x(t) \\ s_y(t) \end{pmatrix}$$

    gegeben.

    Vgl. Illustration (Tragflächenprofil), Skriptum, S. 28.

- Vgl. Illustration (Kurveninterpolation).

2.2. B-Splines

- Ziel: Bestimmung einer geeigeneten Basis des Vektorraumes der Splinefunktionen

$$\mathbb{P}_{0,\tau} = \left\{ s : [a,b) \to \mathbb{R} : s\big|_{[\tau_i,\tau_{i+1})} \in \mathbb{P}_0 \text{ für } 0 \le i \le m-1 \right\},$$
$$\mathbb{P}_{k,\tau} = \left\{ s : [a,b] \to \mathbb{R} \in \mathscr{C}^{k-1}([a,b]) : s\big|_{[\tau_i,\tau_{i+1}]} \in \mathbb{P}_k \text{ für } 0 \le i \le m-1 \right\}, \qquad k \ge 1.$$

- Dimension des Raumes der Splinefunktionen und Basisfunktionen (Lemma 2.5, Definition 2.6, Satz 2.7): Bei $m+1$ Knotenpunkten $(\tau_i)_{0\le i\le m}$ hat der Vektorraum $\mathbb{P}_{k,\tau}$ für $k \ge 0$ die Dimension $m+k$, d.h. es gilt

$$\dim \mathbb{P}_{k,\tau} = m+k, \qquad k \ge 0.$$

Die rekursiv definierten B-Splines $(N_{ki})_{-k\le i\le m-1}$ (mit Einführung zusätzlicher Hilfsknoten $\tau_{-k} < \cdots < \tau_{-1} < \tau_0$ und $\tau_m < \tau_{m+1} < \cdots < \tau_{m+k}$)

$$k = 0: \qquad N_{0i}(x) = \begin{cases} 1 & x \in [\tau_i, \tau_{i+1}), \\ 0 & x \notin [\tau_i, \tau_{i+1}), \end{cases} \qquad 0 \le i \le m-1,$$
$$k \ge 1: \qquad N_{ki}(x) = \frac{1}{\tau_{i+k}-\tau_i}(x-\tau_i)\, N_{k-1,i}(x) + \frac{1}{\tau_{i+k+1}-\tau_{i+1}}(\tau_{i+k+1}-x)\, N_{k-1,i+1}(x), \qquad -k \le i \le m-1,$$

bilden eine Basis von $\mathbb{P}_{k,\tau}$. Jede Splinefunktion $s \in \mathbb{P}_{k,\tau}$ ist somit in eindeutiger Weise als Linearkombination der Basisfunktionen

$$s = \sum_{i=-(k-1)}^{m} c_i\, N_{ki}$$

darstellbar. Denn: Vgl. auch Überlegungen in Abschnitt 2.1.

- $k = 0$: Eine stückweise konstante Splinefunktion in $\mathbb{P}_{0,\tau}$ ist durch m Stützwerte an den Knoten $\tau_0, \ldots, \tau_{m-1}$ bestimmt, und folglich ist

$$\dim \mathbb{P}_{0,\tau} = m.$$

Offenbar bilden die m Splinefunktionen $(N_{0i})_{0\le i\le m-1}$, definiert durch

$$N_{0i}(x) = \begin{cases} 1, & x \in [\tau_i, \tau_{i+1}), \\ 0, & x \notin [\tau_i, \tau_{i+1}), \end{cases} \qquad 0 \le i \le m-1,$$

eine Basis von $\mathbb{P}_{0,\tau}$. Da die Splinefunktion N_{0i} am Knoten τ_i den Wert 1 und an den restlichen Knoten den Wert 0 annimmt

$$N_{0i}(\tau_j) = \delta_{ij}, \qquad 0 \le i \le m-1, \quad 0 \le j \le m,$$

ist jede Splinefunktion $s \in \mathbb{P}_{0,\tau}$ in eindeutiger Weise als Linearkombination

$$s = \sum_{i=0}^{m-1} c_i\, N_{0i}, \qquad c_i = s(\tau_i), \quad 0 \le i \le m-1,$$

darstellbar. Vgl. Abbildung, Skriptum, S. 30.

– $k = 1$: Eine stückweise lineare Splinefunktion $s \in \mathbb{P}_{1,\tau}$ ist auf dem ersten Teilintervall $[\tau_0, \tau_1]$ durch $k+1 = 2$ Bedingungen festgelegt, auf jedem weiteren Teilintervall $[\tau_i, \tau_{i+1}]$ für $1 \le i \le m-1$ kommt aufgrund der Stetigkeitsbedingung an s nur jeweils eine Bedingung hinzu. Dies führt auf insgesamt $2 + m - 1 = m + 1$ Bedingungen, und somit ist

$$\dim \mathbb{P}_{1,\tau} = m + 1 .$$

Geeignete Basisfunktionen sind die $m+1$ stückweise linearen Funktionen $(N_{1i})_{0 \le i \le m}$, definiert durch (die Einführung von zwei Hilfsknoten $\tau_{-1} < \tau_0$ und $\tau_{m+1} > \tau_m$, erlaubt die gleichzeitige Behandlung der inneren Teilintervalle $[\tau_i, \tau_{i+1}]$ für $1 \le i \le m-2$ und der Randintervalle $[\tau_0, \tau_1]$, $[\tau_{m-1}, \tau_m]$; auf $[\tau_i, \tau_{i+1}]$ führt der Ansatz $N_{1i}(x) = \alpha + \beta\,(x - \tau_i)$ und die Forderungen $N_{1i}(\tau_i) = 0$ und $N_{1i}(\tau_{i+1}) = 1$ auf die angegebene Darstellung, auf dem Teilintervall $[\tau_{i+1}, \tau_{i+2}]$ führt der analoger Ansatz $N_{1i}(x) = \alpha + \beta(x - \tau_{i+2})$ und die Forderungen $N_{1i}(\tau_{i+1}) = 1$ und $N_{1i}(\tau_{i+2}) = 0$ auf die angegebene Darstellung)

$$N_{1i}(x) = \begin{cases} \frac{1}{\tau_{i+1} - \tau_i}\,(x - \tau_i), & x \in [\tau_i, \tau_{i+1}], \\ \frac{1}{\tau_{i+2} - \tau_{i+1}}\,(\tau_{i+2} - x), & x \in [\tau_{i+1}, \tau_{i+2}], \\ 0, & \text{sonst}, \end{cases} \qquad -1 \le i \le m-1 ,$$

d.h. N_{1i} nimmt am Knotenpunkt τ_{i+1} den Wert 1 und an den restlichen Knotenpunkten (einschließlich der Hilfsknoten) den Wert 0 an

$$N_{1i}(\tau_j) = \delta_{i+1,j} , \qquad -1 \le i \le m-1 , \quad -1 \le j \le m+1 .$$

Mittels der Basisfunktionen $(N_{0i})_{0 \le i \le m}$ ergibt sich die Darstellung

$$N_{1i}(x) = \tfrac{1}{\tau_{i+1} - \tau_i}\,(x - \tau_i)\,N_{0i}(x) + \tfrac{1}{\tau_{i+2} - \tau_{i+1}}\,(\tau_{i+2} - x)\,N_{0,i+1}(x) , \qquad -1 \le i \le m-1 .$$

Offensichtliche ist die Basisfunktion N_{1i} stetig, auf (τ_i, τ_{i+2}) positiv, und es gilt $N_{1i}(x) = 0$ für $x \notin (\tau_i, \tau_{i+2})$, d.h. die Funktion besitzt einen lokalen Träger

$$\operatorname{supp} N_{1i} = \overline{\{x \in \mathbb{R} : N_{1i}(x) \ne 0\}} = [\tau_i, \tau_{i+2}] .$$

Ähnlich wie zuvor folgt für $s \in \mathbb{P}_{1,\tau}$ die Darstellung

$$s = \sum_{i=-1}^{m-1} c_i\,N_{1i} , \qquad c_i = s(\tau_{i+1}) , \quad -1 \le i \le m-1 .$$

Vgl. Abbildung, Skriptum, S. 31.

– $k \ge 2$: Ähnliche Überlegungen gelten für Splinefunktionen vom Grad k. Auf dem ersten Teilintervall $[\tau_0, \tau_1]$ ist $s \in \mathbb{P}_{k,\tau}$ durch $k+1$ Bedingungen festgelegt, und aufgrund der geforderten Stetigkeitsbedingungen an $s^{(j)}(\tau_i)$ für $0 \le j \le k-1$ kommt

auf jedem weiteren Teilintervall $[\tau_i, \tau_{i+1}]$ für $1 \le i \le m-1$ nur jeweils eine Bedingung hinzu, was auf insgesamt $k+1+m-1 = m+k$ Bedingungen führt

$$\dim \mathbb{P}_{k,\tau} = m+k\,.$$

Die Konstruktion geeigneter Basisfunktionen, der B-Splines, basiert auf der Rekursion (Einführung weiterer Hilfsknoten, Relation konsistent mit obigen Überlegungen für den Fall $k = 1$)

$$\begin{aligned} N_{ki}(x) = {} & \tfrac{1}{\tau_{i+k}-\tau_i}\,(x-\tau_i)\,N_{k-1,i}(x) \\ & + \tfrac{1}{\tau_{i+k+1}-\tau_{i+1}}\,(\tau_{i+k+1}-x)\,N_{k-1,i+1}(x)\,, \qquad -k \le i \le m-1\,. \end{aligned}$$

Eine wesentliche Eigenschaft der B-Splines ist die Lokalität (Trägerintervall $[\tau_i, \tau_{i+1}]$ für N_{0i}, $[\tau_i, \tau_{i+2}]$ für N_{1i}, $[\tau_i, \tau_{i+3}]$ für N_{2i} etc.)

$$\operatorname{supp} N_{ki} = \overline{\{x \in \mathbb{R} : N_{ki}(x) \ne 0\}} = [\tau_i, \tau_{i+k+1}]\,.$$

Wie zuvor folgt für $s \in \mathbb{P}_{k,\tau}$ die Darstellung (Berechnung der Koeffizienten, vgl. Abschnitt 2.3)

$$s = \sum_{i=-k}^{m-1} c_i\,N_{ki}\,.$$

Damit folgt die Behauptung. ◇

Vgl. Illustration (Rekursive Berechnung einer quadratischen B-Splinefunktion).

- Bei einer Linearkombination von B-Splinefunktionen führen kleine Änderungen der Koeffizienten zu kleinen Änderungen der Funktionswerte der Splinefunktion. Ebenso bewirken kleine Änderungen der Funktionswerte der Splinefunktion kleine Änderungen der Koeffizienten.

  Kondition der B-Splines (Satz 2.8): Es gilt die Abschätzung

$$\gamma \max_{-k\le i\le m-1} |c_i| \le \|s\|_\infty = \max_{a\le x\le b} \big|s(x)\big| \le \max_{-k\le i\le m-1} |c_i|\,, \qquad s = \sum_{i=-k}^{m-1} c_i\,N_{ki}\,,$$

  mit einer von der Wahl der Knoten und Koeffizienten unabhängigen Konstante γ.

2.3. Linearkombinationen von B-Splines

- Effiziente Berechnung von Splinefunktionen: Die effiziente Berechnung einer Splinefunktion beruht auf der Darstellung als Linearkombination von B-Splines, vgl. Abschnitt 2.2.

  Einsetzen der Rekursion (Kubische Splinefunktionen): Für eine kubische Splinefunktion

  $$s = \sum_{i=-3}^{m-1} c_i N_{3i} \in \mathbb{P}_{3,\tau}$$

  ergeben sich bei schrittweisem Einsetzen der in Abschnitt 2.2 angegebenen Rekursion

  $$\begin{aligned}
  N_{ki}(x) &= \tfrac{1}{\tau_{i+k}-\tau_i}(x-\tau_i)N_{k-1,i}(x) \\
  &\quad + \tfrac{1}{\tau_{i+k+1}-\tau_{i+1}}(\tau_{i+k+1}-x)N_{k-1,i+1}(x), \qquad -k \le i \le m-1, \qquad k \ge 1, \\
  N_{3i}(x) &= \tfrac{x-\tau_i}{\tau_{i+3}-\tau_i}N_{2i}(x) + \tfrac{\tau_{i+4}-x}{\tau_{i+4}-\tau_{i+1}}N_{2,i+1}(x), \qquad -3 \le i \le m-1, \\
  N_{2i}(x) &= \tfrac{x-\tau_i}{\tau_{i+2}-\tau_i}N_{1i}(x) + \tfrac{\tau_{i+3}-x}{\tau_{i+3}-\tau_{i+1}}N_{1,i+1}(x), \qquad -2 \le i \le m-1, \\
  N_{1i}(x) &= \tfrac{x-\tau_i}{\tau_{i+1}-\tau_i}N_{0i}(x) + \tfrac{\tau_{i+2}-x}{\tau_{i+2}-\tau_{i+1}}N_{0,i+1}(x), \qquad -1 \le i \le m-1, \\
  N_{0i}(x) &= \begin{cases} 1 & x \in [\tau_i, \tau_{i+1}), \\ 0 & x \notin [\tau_i, \tau_{i+1}), \end{cases} \qquad 0 \le i \le m-1,
  \end{aligned}$$

  die Relationen (Indexverschiebung $j = i+1 \leftrightarrow i = j-1$)

  $$\begin{aligned}
  s(x) &= \sum_{i=-3}^{m-1} c_i N_{3i}(x) \\
  &= \sum_{i=-3}^{m-1} c_i \tfrac{x-\tau_i}{\tau_{i+3}-\tau_i}N_{2i}(x) + \sum_{i=-3}^{m-1} c_i \tfrac{\tau_{i+4}-x}{\tau_{i+4}-\tau_{i+1}}N_{2,i+1}(x) \\
  &= \sum_{i=-2}^{m-1} \left(c_i \tfrac{x-\tau_i}{\tau_{i+3}-\tau_i} + c_{i-1}\tfrac{\tau_{i+3}-x}{\tau_{i+3}-\tau_i}\right)N_{2i}(x) + \text{Randterme},
  \end{aligned}$$

  $$\begin{aligned}
  s(x) = \sum_{i=-1}^{m-1} &\left(\left(c_i \tfrac{x-\tau_i}{\tau_{i+3}-\tau_i} + c_{i-1}\tfrac{\tau_{i+3}-x}{\tau_{i+3}-\tau_i}\right)\tfrac{x-\tau_i}{\tau_{i+2}-\tau_i} + \left(c_{i-1}\tfrac{x-\tau_{i-1}}{\tau_{i+2}-\tau_{i-1}} + c_{i-2}\tfrac{\tau_{i+2}-x}{\tau_{i+2}-\tau_{i-1}}\right)\tfrac{\tau_{i+2}-x}{\tau_{i+2}-\tau_i}\right)N_{1i}(x) \\
  &+ \text{Randterme},
  \end{aligned}$$

  $$\begin{aligned}
  s(x) = \sum_{i=0}^{m-1} &\left(\left(\left(c_i \tfrac{x-\tau_i}{\tau_{i+3}-\tau_i} + c_{i-1}\tfrac{\tau_{i+3}-x}{\tau_{i+3}-\tau_i}\right)\tfrac{x-\tau_i}{\tau_{i+2}-\tau_i}\right.\right. \\
  &\left. + \left(c_{i-1}\tfrac{x-\tau_{i-1}}{\tau_{i+2}-\tau_{i-1}} + c_{i-2}\tfrac{\tau_{i+2}-x}{\tau_{i+2}-\tau_{i-1}}\right)\tfrac{\tau_{i+2}-x}{\tau_{i+2}-\tau_i}\right)\tfrac{x-\tau_i}{\tau_{i+1}-\tau_i} \\
  &+ \left(\left(c_{i-1}\tfrac{x-\tau_{i-1}}{\tau_{i+2}-\tau_{i-1}} + c_{i-2}\tfrac{\tau_{i+2}-x}{\tau_{i+2}-\tau_{i-1}}\right)\tfrac{x-\tau_{i-1}}{\tau_{i+1}-\tau_{i-1}}\right. \\
  &\left.\left. + \left(c_{i-2}\tfrac{x-\tau_{i-2}}{\tau_{i+1}-\tau_{i-2}} + c_{i-3}\tfrac{\tau_{i+1}-x}{\tau_{i+1}-\tau_{i-2}}\right)\tfrac{\tau_{i+1}-x}{\tau_{i+1}-\tau_{i-1}}\right)\tfrac{\tau_{i+1}-x}{\tau_{i+1}-\tau_i}\right)N_{0i}(x) + \text{Randterme}.
  \end{aligned}$$

Berechnung der Koeffizienten (Allgemeiner Fall): Die Koeffizienten $(c_i)_{-k\le i\le m-1}$ werden so bestimmt, daß die Splinefunktion an Stützstellen $(x_i)_{-k\le i\le m-1}$, welche auch von den Knoten verschieden sein können, eine vorgegebene Funktion f interpoliert

$$s(x_i) = f(x_i)\,, \qquad -k \le i \le m-1\,.$$

Der Satz von Schoenberg–Whitney (Satz 2.9) sichert die Existenz und Eindeutigkeit der Lösung, wenn die Stützstellen die Bedingung $x_i \in (\tau_i, \tau_{i+k})$ für $-k \le i \le m-1$ erfüllen.

Berechnung der Koeffizienten (Kubische Splinefunktionen): Beispielsweise im Fall einer kubischen Splinefunktion führen die Interpolationsbedingungen und die obigen Überlegungen

$$\begin{aligned}
&s(x_i) = f(x_i)\,, \qquad -k \le i \le m-1\,,\\
&s(x) = \sum_{i=0}^{m-1} \Bigg(\Bigg(\Big(c_i \tfrac{x-\tau_i}{\tau_{i+3}-\tau_i} + c_{i-1} \tfrac{\tau_{i+3}-x}{\tau_{i+3}-\tau_i} \Big) \tfrac{x-\tau_i}{\tau_{i+2}-\tau_i}\\
&\qquad + \Big(c_{i-1} \tfrac{x-\tau_{i-1}}{\tau_{i+2}-\tau_{i-1}} + c_{i-2} \tfrac{\tau_{i+2}-x}{\tau_{i+2}-\tau_{i-1}} \Big) \tfrac{\tau_{i+2}-x}{\tau_{i+2}-\tau_i} \Bigg) \tfrac{x-\tau_i}{\tau_{i+1}-\tau_i}\\
&\qquad\quad + \Bigg(\Big(c_{i-1} \tfrac{x-\tau_{i-1}}{\tau_{i+2}-\tau_{i-1}} + c_{i-2} \tfrac{\tau_{i+2}-x}{\tau_{i+2}-\tau_{i-1}} \Big) \tfrac{x-\tau_{i-1}}{\tau_{i+1}-\tau_{i-1}}\\
&\qquad\qquad + \Big(c_{i-2} \tfrac{x-\tau_{i-2}}{\tau_{i+1}-\tau_{i-2}} + c_{i-3} \tfrac{\tau_{i+1}-x}{\tau_{i+1}-\tau_{i-2}} \Big) \tfrac{\tau_{i+1}-x}{\tau_{i+1}-\tau_{i-1}} \Bigg) \tfrac{\tau_{i+1}-x}{\tau_{i+1}-\tau_i} \Bigg) N_{0i}(x) + \text{Randterme}\,,
\end{aligned}$$

auf ein lineares Gleichungssystem für die Koeffizienten $(c_i)_{-3\le i\le m-1}$, welches in der Situation des Satzes von Schoenberg–Whitney eindeutig lösbar ist. Aufgrund der Lokalität der B-Splines besitzt die zugehörige Matrix Bandstruktur (Diagonale und wenige Nebendiagonale beinhalten Elemente, die verschieden von Null sind) und das Gleichungssystem ist somit mittels LR-Zerlegung (sogar ohne Pivotsuche) effizient lösbar.

3. Numerische Integration

- Inhalt: Verfahren zur numerischen Quadratur, d.h. zur näherungsweisen Berechnung bestimmter Integrale der Form

$$I(f) = \int_a^b f(x)\,\mathrm{d}x$$

mit einer (zumindest stetigen) Funktion $f : [a,b] \to \mathbb{R}$.

Bemerkung:

- Die numerische Berechnung bestimmter Integrale ist erforderlich, wenn es nicht möglich ist, eine Stammfunktion des Integranden in geschlossener Form (d.h. mittels elementarer Funktionen) anzugeben. Es gibt aber auch Situationen, in denen eine direkte Auswertung der Stammfunktion aufgrund des Auftretens des Phänomens der Auslöschung signifikanter Stellen numerisch instabil ist, numerische Quadratur hingegen ein zufriedenstellendes Ergebnis ergibt.

  Illustration (Partielle Integration, Instabilität).

- Numerische Quadraturformeln beruhen auf der Idee, das bestimmte Integral

$$I(f) = \int_a^b f(x)\,\mathrm{d}x$$

durch ein einfach zu berechnendes bestimmtes Integral

$$I(\widetilde{f}) = \int_a^b \widetilde{f}(x)\,\mathrm{d}x$$

zu approximieren, d.h. der (komplizierte) Integrand f wird durch eine geeignete Approximation einfacherer Struktur (beispielsweise durch eine (stückweise) Polynomfunktion) ersetzt.

- Falls Funktionswerte des Integranden an beliebigen Stützstellen im Intervall $[a,b]$ berechnet werden können, stehen adaptive Quadraturformeln zur Verfügung, die für hinreichend reguläre Integranden ein Ergebnis hoher Genauigkeit liefern.

- Erinnerung: Ein Maß für die Länge einer stetigen Funktion oder auch den Abstand zweier stetiger Funktionen ist die Supremumsnorm

$$\|f\|_\infty = \max_{a\le x\le b} |f(x)|, \quad \|f-\widetilde{f}\|_\infty = \max_{a\le x\le b} |f(x)-\widetilde{f}(x)|, \qquad f,\widetilde{f} \in \mathscr{C}([a,b]).$$

Kondition der Integration: Zur Untersuchung der Sensibilität des Wertes eines bestimmten Integrals in Abhängigkeit vom Integranden betrachtet man die Differenz

$$I(f) - I(\widetilde{f}) = \int_a^b \big(f(x)-\widetilde{f}(x)\big)\,\mathrm{d}x, \qquad f,\widetilde{f} \in \mathscr{C}([a,b]).$$

Als Abschätzung für die absolute Kondition ergibt sich

$$\begin{aligned}\left|I(f)-I(\widetilde{f})\right| &\leq \int_a^b \left|f(x)-\widetilde{f}(x)\right| \mathrm{d}x \leq \left\|f-\widetilde{f}\right\|_\infty \int_a^b 1 \,\mathrm{d}x \\ &= (b-a)\left\|f-\widetilde{f}\right\|_\infty, \qquad f,\widetilde{f}\in\mathscr{C}([a,b])\,.\end{aligned}$$

Bei der relativen Kondition hingegen

$$\begin{aligned}\left|\frac{I(f)-I(\widetilde{f})}{I(f)}\right| &\leq (b-a)\,\frac{\left\|f-\widetilde{f}\right\|_\infty}{|I(f)|} \\ &= \underbrace{\frac{(b-a)\,\|f\|_\infty}{|I(f)|}}_{=\kappa}\,\frac{\left\|f-\widetilde{f}\right\|_\infty}{\|f\|_\infty}, \qquad f,\widetilde{f}\in\mathscr{C}([a,b])\,,\end{aligned}$$

kann der Fall eintreten, daß der Faktor

$$\kappa = \frac{(b-a)\,\|f\|_\infty}{|I(f)|} \gg 1$$

einen großen Wert annimmt und somit kleine relative Änderungen des Integranden zu großen Änderungen des Ergebnisses führen. Ein typischer Fall eines schlecht konditionierten numerische Problems ist ein stark oszillierender Integrand, der etwa bei Fourier-Integralen auftritt ($I(f)\approx 0$, $(b-a)\,\|f\|_\infty \approx 1$).

3.1. Elementare Quadraturformeln

- Fragestellung: Einführung grundlegender Begriffe und Relationen.

  Eine s-stufige Quadraturformel $(b_i, c_i)_{1\le i\le s}$ ist durch die Knoten $c_i \in [0,1]$ und die zugehörigen Gewichte $b_i \in \mathbb{R}$ für $1 \le i \le s$ gegeben. Eine Quadraturformel $(b_i, c_i)_{1\le i\le s}$ heißt symmetrisch, falls $c_{s+1-i} = 1 - c_i$ und $b_{s+1-i} = b_i$ für $1 \le i \le s$.

  - Lokale Approximation (Einheitsintervall): Für eine Funktion $g : [0,1] \to \mathbb{R}$ erhält man eine Approximation an den Wert des bestimmten Integrals mittels

$$Q_0(g) = \sum_{i=1}^{s} b_i\, g(c_i) \quad \approx \quad I_0(g) = \int_0^1 g(\xi)\,\mathrm{d}\xi\,.$$

    Die Konstruktion von (ersten einfachen) Quadraturformeln beruht auf der Approximation von g durch *einfache* Funktionen $\widetilde{g}$ (z.B. Polynominterpolanten) für welche das bestimmte Integral leicht berechnet werden kann (s.u.)

$$Q_0(g) = \sum_{i=1}^{s} b_i\, g(c_i) = I_0(\widetilde{g}) = \int_0^1 \widetilde{g}(\xi)\,\mathrm{d}\xi \quad \approx \quad I_0(g) = \int_0^1 g(\xi)\,\mathrm{d}\xi\,.$$

    Der (lokale) Verfahrensfehler

$$Q_0(g) - I_0(g) = \sum_{i=1}^{s} b_i\, g(c_i) - \int_0^1 g(\xi)\,\mathrm{d}\xi$$

    ist damit durch die Relation

$$Q_0(g) - I_0(g) = \sum_{i=1}^{s} b_i\, g(c_i) - \int_0^1 g(\xi)\,\mathrm{d}\xi = I_0(\widetilde{g}) - I_0(g) = \int_0^1 (\widetilde{g} - g)(\xi)\,\mathrm{d}\xi$$

    gegeben.

  - Globale Approximation: Für eine Funktion $f : [a,b] \to \mathbb{R}$ erhält man mittels einer Zerlegung des betrachteten Intervalles in *kleine* Teilintervalle

$$a = x_0 < \cdots < x_N = b\,, \qquad h_j = x_{j+1} - x_j\,, \quad 0 \le j \le N-1\,,$$

    und der Variablentransformation $x = x_j + \xi\, h_j \leftrightarrow \xi = \frac{1}{h_j}(x - x_j)$ die Approximation

$$\begin{aligned} I(f) &= \int_a^b f(x)\,\mathrm{d}x = \sum_{j=0}^{N-1} \int_{x_j}^{x_{j+1}} f(x)\,\mathrm{d}x = \sum_{j=0}^{N-1} h_j \int_0^1 f(x_j + \xi\, h_j)\,\mathrm{d}\xi \\ &\approx Q(f) = \sum_{j=0}^{N-1} h_j \sum_{i=1}^{s} b_i\, f(x_j + c_i\, h_j)\,. \end{aligned}$$

    Wesentlich für die Abschätzung des (globalen) Verfahrensfehlers

$$Q(f) - I(f) = \sum_{j=0}^{N-1} h_j \left(\sum_{i=1}^{s} b_i\, f(x_j + c_i\, h_j) - \int_0^1 f(x_j + \xi\, h_j)\,\mathrm{d}\xi \right)$$

sind somit Abschätzungen für den lokalen Verfahrensfehler

$$Q_0(g_j) - I_0(g_j) = \sum_{i=1}^{s} b_i\, g_j(c_i) - \int_0^1 g_j(\xi)\,\mathrm{d}\xi\,, \quad g_j(\xi) = f(x_j + \xi\, h_j)\,, \qquad 0 \le j \le N-1\,.$$

Bemerkung: Es ist wünschenswert, Quadraturformeln mit positiven Gewichten zu verwenden. In diesem Fall ist sichergestellt, daß aus der Positivität des Integranden die Positivität des bestimmten Integrals und ebenso die Positivität der Quadraturapproximation folgt

$$f(x) \ge 0 \quad \text{für } x \in [a,b] \quad \Longrightarrow \quad I(f) \ge 0\,, \quad Q(f) \ge 0\,.$$

- Fragestellung: Untersuchung des Verfahrensfehlers der Quadraturapproximation (Güte der Approximation, Konstruktion von Quadraturformeln)

  – Zur Abschätzung des Verfahrensfehlers der Quadraturapproximation

$$Q(f) - I(f) = \sum_{j=0}^{N-1} h_j \left(\sum_{i=1}^{s} b_i\, f(x_j + c_i\, h_j) - \int_0^1 f(x_j + \xi\, h_j)\,\mathrm{d}\xi \right)$$

betrachtet man zunächst den lokalen Verfahrensfehler. Mittels der Taylorreihenentwicklungen

$$f(x_j + c_i\, h_j) = \sum_{\ell=0}^{p-1} \tfrac{1}{\ell!}\, f^{(\ell)}(x_j)\, c_i^\ell\, h_j^\ell + \tfrac{1}{p!}\, f^{(p)}(\eta_{ij})\, c_i^p\, h_j^p\,,$$
$$f(x_j + \xi\, h_j) = \sum_{\ell=0}^{p-1} \tfrac{1}{\ell!}\, f^{(\ell)}(x_j)\, \xi^\ell\, h_j^\ell + \tfrac{1}{p!}\, f^{(p)}(\widetilde{\eta}_{ij})\, \xi^p\, h_j^p\,,$$

ergibt sich die Relation

$$\begin{aligned}
&\sum_{i=1}^{s} b_i\, f(x_j + c_i\, h_j) - \int_0^1 f(x_j + \xi\, h_j)\,\mathrm{d}\xi \\
&\quad = \sum_{i=1}^{s} b_i \sum_{\ell=0}^{p-1} \tfrac{1}{\ell!}\, f^{(\ell)}(x_j)\, c_i^\ell\, h_j^\ell - \sum_{\ell=0}^{p-1} \tfrac{1}{\ell!}\, f^{(\ell)}(x_j)\, h_j^\ell \underbrace{\int_0^1 \xi^\ell\,\mathrm{d}\xi}_{=\frac{1}{\ell+1}} \\
&\qquad + \sum_{i=1}^{s} b_i\, \tfrac{1}{p!}\, f^{(p)}(\eta_{ij})\, c_i^p\, h_j^p - \tfrac{1}{p!}\, f^{(p)}(\widetilde{\eta}_{ij})\, h_j^p \underbrace{\int_0^1 \xi^p\,\mathrm{d}\xi}_{=\frac{1}{p+1}} \\
&\quad = \sum_{\ell=0}^{p-1} \tfrac{1}{\ell!} \left(\sum_{i=1}^{s} b_i\, c_i^\ell - \tfrac{1}{\ell+1} \right) f^{(\ell)}(x_j)\, h_j^\ell + \tfrac{1}{p!} \left(\sum_{i=1}^{s} b_i\, f^{(p)}(\eta_{ij})\, c_i^p - \tfrac{1}{p+1}\, f^{(p)}(\widetilde{\eta}_{ij}) \right) h_j^p\,.
\end{aligned}$$

Weiters erhält man die Entwicklung

$$Q(f) - I(f) = \sum_{j=0}^{N-1} \sum_{\ell=0}^{p-1} \tfrac{1}{\ell!} \Big(\sum_{i=1}^{s} b_i \, c_i^\ell - \tfrac{1}{\ell+1} \Big) f^{(\ell)}(x_j) \, h_j^{\ell+1} + \tfrac{1}{p!} \sum_{j=0}^{N-1} \Big(\sum_{i=1}^{s} b_i \, f^{(p)}(\eta_{ij}) \, c_i^p - \tfrac{1}{p+1} \, f^{(p)}(\widetilde{\eta}_{ij}) \Big) h_j^{p+1}.$$

– Falls die Ordnungsbedingungen

$$\begin{aligned} \sum_{i=1}^{s} b_i \, c_i^{q-1} &= \tfrac{1}{q}, \qquad 1 \le q \le p, \\ q = 1: \qquad & b_1 + \cdots + b_s = 1 \\ q = 2: \qquad & b_1 \, c_1 + \cdots + b_s \, c_s = \tfrac{1}{2} \\ q = 3: \qquad & b_1 \, c_1^2 + \cdots + b_s \, c_s^2 = \tfrac{1}{3} \\ & \vdots \\ q = p: \qquad & b_1 \, c_1^{p-1} + \cdots + b_s \, c_s^{p-1} = \tfrac{1}{p} \end{aligned}$$

erfüllt sind, ergibt die Quadraturformel eine Approximation der Ordnung $p \ge 1$, d.h. für den lokalen Verfahrensfehler gilt die Relation (sofern $f^{(p)}$ auf $[a,b]$ beschränkt ist)

$$\sum_{i=1}^{s} b_i \, f(x_j + c_i \, h_j) - \int_0^1 f(x_j + \xi \, h_j) \, \mathrm{d}\xi = \mathscr{O}\big(h_{\max}^p\big), \qquad h_{\max} = \max_{0 \le j \le N-1} h_j,$$

und für den globalen Verfahrensfehler folgt die Abschätzung (mit Konstante C abhängig von den Knoten und Gewichten $(b_i, c_i)_{1 \le i \le s}$ sowie der Ordnung p)

$$\big| Q(f) - I(f) \big| \le C \, \big\| f^{(p)} \big\|_\infty \sum_{j=0}^{N-1} h_j^{p+1} \le C \, (b-a) \, h_{\max}^p \big\| f^{(p)} \big\|_\infty, \qquad h_{\max} = \max_{0 \le j \le N-1} h_j.$$

– Für eine Quadraturformel $(b_i, c_i)_{1 \le i \le s}$ der Ordnung p folgt insbesondere, daß die Quadraturapproximation auf dem Einheitsintervall für die Taylor-Basis

$$g_\ell(x) = x^\ell, \qquad 0 \le \ell \le p-1,$$

das exakte Ergebnis liefert (verwende Linearität der Quadraturformel und des bestimmten Integrals)

$$\sum_{i=1}^{s} b_i \, c_i^\ell = \sum_{i=1}^{s} b_i \, g_\ell(c_i) = \int_0^1 g_\ell(\xi) \, \mathrm{d}\xi = \int_0^1 \xi^\ell \, \mathrm{d}\xi = \tfrac{1}{\ell+1}, \qquad 0 \le \ell \le p-1.$$

- Beispiele für (einfache) Quadraturformeln: Mittelpunktsregel, Trapezregel, Simpsonregel

– Mittelpunktsregel

* Die Forderungen $s = 1$ und $p = 2\,s = 2$ führen auf die Knoten und Gewichte der Mittelpunktsregel (Lösung der Ordnungsbedingungen $b_1 = 1$, $b_1\,c_1 = \frac{1}{2}$, jedoch $b_1\,c_1^2 \neq \frac{1}{3}$, symmetrische Quadraturformel, Superkonvergenz $p > s$, Gaußsche Quadraturformel der maximalen Ordnung $p = 2\,s$)

$$s = 1, \quad b_1 = 1\,, \quad c_1 = \tfrac{1}{2}\,, \quad p = 2\,.$$

* Historischer Zugang: Interpolation der Funktion $g : [0,1] \to \mathbb{R}$ im Intervallmittelpunkt $\frac{1}{2}$ mittels eines Polynoms vom Grad 0 und Integration führt auf die Quadraturapproximation

$$\begin{gathered}\widetilde{g}(\xi) = g\big(\tfrac{1}{2}\big) \quad \approx \quad g(\xi)\,, \qquad 0 \le \xi \le 1\,,\\ Q_0(g) = \int_0^1 \widetilde{g}(\xi)\,\mathrm{d}\xi = \int_0^1 g\big(\tfrac{1}{2}\big)\,\mathrm{d}\xi = g\big(\tfrac{1}{2}\big) \quad \approx \quad I_0(g) = \int_0^1 g(\xi)\,\mathrm{d}\xi\,,\\ s = 1\,, \quad b_1 = 1\,, \quad c_1 = \tfrac{1}{2}\,.\end{gathered}$$

Vgl. Abbildung, Skriptum, S. 39 ($a = 0, b = 1$, Rechtecksfläche).

* Insgesamt ergibt sich die Quadraturapproximation (wegen $h_j = x_{j+1} - x_j$ und $x_j + \frac{1}{2}\,h_j = \frac{1}{2}\,(x_j + x_{j+1})$)

$$Q(f) = \sum_{j=0}^{N-1} h_j\, f\big(\tfrac{x_j + x_{j+1}}{2}\big) \quad \approx \quad I(f) = \int_0^1 f(x_j + \xi\,h_j)\,\mathrm{d}\xi\,.$$

– Trapezregel

* Die Forderungen $s = 2$ und $p = s$ sowie $c_1 = 0$ und $c_2 = 1$ führen auf die Gewichte der Trapezregel (Lösung der Ordnungsbedingungen $b_1 + b_2 = 1$, $b_1\,c_1 + b_2\,c_2 = \frac{1}{2}$, jedoch $b_1\,c_1^2 + b_2\,c_2^2 \neq \frac{1}{3}$, äquidistante Knoten in $[0,1]$, Newton–Côtes Formel)

$$s = 2, \quad b_1 = \tfrac{1}{2} = b_2\,, \quad c_1 = 0\,, \quad c_2 = 1\,, \quad p = 2\,.$$

* Historischer Zugang: Interpolation der Funktion $g : [0,1] \to \mathbb{R}$ in den Intervallenden $0, 1$ mittels eines Polynoms vom Grad 1 und Integration führt auf die Quadraturapproximation

$$\begin{gathered}\widetilde{g}(\xi) = g(0) + \big(g(1) - g(0)\big)\,\xi \quad \approx \quad g(\xi)\,, \qquad 0 \le \xi \le 1\,,\\ Q_0(g) = \int_0^1 \widetilde{g}(\xi)\,\mathrm{d}\xi = \int_0^1 \big(g(0) + \big(g(1) - g(0)\big)\,\xi\big)\,\mathrm{d}\xi = g(0) + \tfrac{1}{2}\,\big(g(1) - g(0)\big)\\ = \tfrac{1}{2}\,\big(g(0) + g(1)\big) \quad \approx \quad I_0(g) = \int_0^1 g(\xi)\,\mathrm{d}\xi\,,\\ s = 2\,, \quad b_1 = \tfrac{1}{2} = b_2\,, \quad c_1 = 0\,, \quad c_2 = 1\,.\end{gathered}$$

Vgl. Abbildung, Skriptum, S. 41 ($a = 0, b = 1$, Fläche des Trapezes).

* Insgesamt ergibt sich die Quadraturapproximation

$$\begin{aligned} Q(f) &= \tfrac{1}{2}\sum_{j=0}^{N-1} h_j \left(f(x_j)+f(x_{j+1})\right) \\ &= \tfrac{1}{2}\sum_{j=0}^{N-1} h_j f(x_j) + \tfrac{1}{2}\sum_{j=0}^{N-1} h_j f(x_{j+1}) \\ &= \tfrac{1}{2}\sum_{j=0}^{N-1} h_j f(x_j) + \tfrac{1}{2}\sum_{j=1}^{N} h_{j-1} f(x_j) \\ &= \tfrac{h_0}{2} f(x_0) + \sum_{j=1}^{N-1} \tfrac{h_{j-1}+h_j}{2} f(x_j) + \tfrac{h_{N-1}}{2} f(x_N) \quad \approx \quad I(f) = \int_a^b f(x)\,\mathrm{d}x\,. \end{aligned}$$

* Für periodische Funktionen und äquidistante Stützstellen vereinfacht sich die obige Relation zu ($f(x_N) = f(x_0)$ und $h_j = 0$ für $0 \le j \le N-1$)

$$Q(f) = h\sum_{j=0}^{N-1} f(x_j) \quad \approx \quad I(f) = \int_a^b f(x)\,\mathrm{d}x\,.$$

Die Trapezregel hat spezielle Bedeutung bei der numerischen Berechnung von Fourierreihen (Signalverarbeitung).

– Simpsonregel (Faßregel, Johannes Kepler, 1615)

* Die Forderungen $s = 3$ und $p = s$ sowie $c_1 = 0$, $c_2 = \frac{1}{2}$ und $c_3 = 1$ führen auf die Gewichte der Simpsonregel (Lösung der entsprechenden Ordnungsbedingungen $b_1 + b_2 + b_3 = 1$, $b_1\,c_1 + b_2\,c_2 + b_3\,c_3 = \frac{1}{2}$, $b_1\,c_1^2 + b_2\,c_2^2 + b_3\,c_3^2 = \frac{1}{3}$, äquidistante Knoten in $[0,1]$, Newton–Côtes Formel, aufgrund der Symmetrie der Quadraturformel gilt sogar $b_1\,c_1^3 + b_2\,c_2^3 + b_3\,c_3^3 = \frac{1}{4}$ jedoch $b_1\,c_1^4 + b_2\,c_2^4 + b_3\,c_3^4 = \frac{1}{5}$ und damit $p = 4$)

$$s = 3\,, \quad b_1 = \tfrac{1}{6}\,, \quad b_2 = \tfrac{2}{3}\,, \quad b_3 = \tfrac{1}{6}\,, \quad c_1 = 0\,, \quad c_2 = \tfrac{1}{2}\,, \quad c_3 = 1\,, \quad p = 4\,.$$

* Historischer Zugang: Interpolation der Funktion $g : [0,1] \to \mathbb{R}$ in den Intervallenden $0, 1$ und dem Intervallmittelpunkt $\frac{1}{2}$ mittels eines Polynoms vom Grad 2

0	$g(0)$		
		$2\left(g\left(\frac{1}{2}\right) - g(0)\right)$	
$\frac{1}{2}$	$g\left(\frac{1}{2}\right)$		$2\left(g(0) - 2\,g\left(\frac{1}{2}\right) + g(1)\right)$
		$2\left(g(1) - g\left(\frac{1}{2}\right)\right)$	
1	$g(1)$		

$$\begin{aligned} \widetilde{g}(\xi) &= g(0) + 2\left(g\left(\tfrac{1}{2}\right) - g(0)\right)\xi + 2\left(g(0) - 2\,g\left(\tfrac{1}{2}\right) + g(1)\right)\xi\left(\xi - \tfrac{1}{2}\right) \\ &\approx \quad g(\xi)\,, \qquad 0 \le \xi \le 1\,, \end{aligned}$$

und Integration führt auf die Quadraturapproximation

$$\begin{aligned} Q_0(g) &= \int_0^1 \widetilde{g}(\xi)\,\mathrm{d}\xi \\ &= \int_0^1 \Big(g(0) + 2\big(g\big(\tfrac{1}{2}\big) - g(0)\big)\,\xi + \big(g(0) - 2\,g\big(\tfrac{1}{2}\big) + g(1)\big)\big(2\,\xi^2 - \xi\big)\Big)\,\mathrm{d}\xi \\ &= g\big(\tfrac{1}{2}\big) + \tfrac{1}{6}\big(g(0) - 2\,g\big(\tfrac{1}{2}\big) + g(1)\big) \\ &= \tfrac{1}{6}\big(g(0) + 4\,g\big(\tfrac{1}{2}\big) + g(1)\big), \end{aligned}$$

$$s = 3, \quad b_1 = \tfrac{1}{6}, \quad b_2 = \tfrac{2}{3}, \quad b_3 = \tfrac{1}{6}, \quad c_1 = 0, \quad c_2 = \tfrac{1}{2}, \quad c_3 = 1.$$

* Insgesamt ergibt sich die Quadraturapproximation

$$\begin{aligned} Q(f) &= \tfrac{1}{6}\sum_{j=0}^{N-1} h_j\big(f(x_j) + 4\,f\big(\tfrac{x_j + x_{j+1}}{2}\big) + f(x_{j+1})\big) \\ &= \tfrac{h_0}{6}\,f(x_0) + \sum_{j=1}^{N-1} \tfrac{h_{j-1} + h_j}{6}\,f(x_j) + \tfrac{2}{3}\sum_{j=0}^{N-1} h_j\,f\big(\tfrac{x_j + x_{j+1}}{2}\big) + \tfrac{h_{N-1}}{6}\,f(x_N) \\ &\approx I(f) = \int_a^b f(x)\,\mathrm{d}x. \end{aligned}$$

- Verschiedene Klassen von Quadraturformeln:
  - Allgemeiner erhält man die Newton–Côtes Quadraturformeln bei Vorgabe von s äquidistanten Knoten in $[0,1]$ und Lösung der Ordnungsbedingungen für $p = s$ (eindeutig lösbares lineares Gleichungssystem mit Vandermonde Matrix, allerdings schlecht konditioniert). Dies entspricht gerade der Integration des Polynominterpolanten. Da für höhere Stufenzahlen negative Gewichte auftreten, werden Newton–Côtes Quadraturformeln höherer Ordnung im Allgemeinen vermieden.
  - Bei Verwendung der Chebychev-Knoten ergeben sich die Clenshaw–Curtis Quadraturformeln mit stets positiven Gewichten.
  - Quadraturformeln hoher Ordnung, deren Konstruktion auf orthogonalen Polynomen beruht, sind beispielsweise die Gaußschen Quadraturformeln ($p = 2\,s$) und die Radauschen Quadraturformeln ($p = 2\,s - 1$).

3.2. Best-Approximation des Integrals

- Fragestellung: Zu bestimmen ist die beste Quadraturapproximation an das Integral

$$\widetilde{I}(f) = \int_a^b f(x)\,\mathrm{d}x$$

bei Vorgabe von s Stützstellen und zugehörigen Stützwerten

$$a \le \gamma_1 < \cdots < \gamma_s \le b\,, \qquad \eta_i = f(\gamma_i)\,, \quad 1 \le i \le s\,.$$

Betrachtet werden dabei lokale Quadraturapproximationen der Form (Transformation des Intervalles $[a,b]$ auf das Einheitsintervall $[0,1]$ mittels $x = a + \xi\,(b-a) \leftrightarrow \xi = \frac{x-a}{b-a}$)

$$\begin{aligned}\widetilde{I}(f) &= \int_a^b f(x)\,\mathrm{d}x = (b-a)\int_0^1 f\big(a+\xi\,(b-a)\big)\,\mathrm{d}\xi \\ &\approx \widetilde{Q}(f) = \sum_{i=1}^{s} \underbrace{(b-a)\,b_i}_{\beta_i}\, f\Big(\underbrace{a + c_i\,(b-a)}_{=\gamma_i}\Big) = \sum_{i=1}^{s} \beta_i\, f(\gamma_i)\,.\end{aligned}$$

Als Maß für Funktionen betrachtet man die mittlere quadratische Krümmung (Seminorm, $\nu(f) = \|f''\|_{L^2}$)

$$\nu(f)^2 = \int_a^b \big(f''(x)\big)^2\,\mathrm{d}x\,, \qquad f \in \mathscr{C}^2([a,b])\,.$$

Je *glatter* die Funktion f ist, desto kleiner ist $\nu(f)$, und insbesondere gilt $\nu(f) = 0$ für lineare Polynomfunktionen $f \in \mathbb{P}_1$. Für beliebige Funktionen $f \in \mathscr{C}^2([a,b]) \setminus \mathbb{P}_1$ sollen die Quadraturgewichte $(\beta_i)_{1\le i\le s}$ so bestimmt werden, daß die Größe

$$\tfrac{1}{\nu(f)}\,\big|\widetilde{Q}(f) - \widetilde{I}(f)\big| \quad \longrightarrow \quad \min$$

minimal wird. Als Lösungen des Minimierungsproblems ergeben sich natürliche kubische Splinefunktionen zu den Stützstellen $(\gamma_i)_{1\le i\le s}$ (Variationsrechnung)

$$\widetilde{Q}(f) = \widetilde{I}(s)\,, \qquad s \in \mathbb{P}_{3,\gamma}\,, \quad s''(\gamma_1) = 0 = s''(\gamma_s)\,.$$

3.3. Romberg-Quadratur

- Vorbemerkung: Für die folgenden Überlegungen ist es zweckmäßig, einen hinreichend regulären Integranden f zu fixieren und anstelle der Abhängigkeit der Quadraturapproximation vom Integranden die Abhängigkeit von der Schrittweite h (zu einer äquidistanten Zerlegung des Integrationsintervalles) anzugeben. Ebenso wird der Wert des bestimmten Integrals kurz mit I (oder auch mit α_0) bezeichnet.

  Erinnerung: Bei äquidistanten Stützstellen $a = x_0 < \cdots < x_N = b$ mit $x_j = a + j\,h$ für $0 \le j \le N-1$, wobei $h = \frac{b-a}{N}$, führt die Anwendung der Trapezregel auf die globale Quadraturapproximation

$$T(h) = h\left(\tfrac{1}{2} f(x_0) + \sum_{j=1}^{N-1} f(x_j) + \tfrac{1}{2} f(x_N)\right) \quad \approx \quad I = \int_a^b f(x)\,\mathrm{d}x\,.$$

  Frühere Überlegungen zeigten die Abschätzung

$$\left|T(h) - I\right| \le C\,(b-a)\,h^2 \left\|f''\right\|_\infty .$$

  Eine zusätzliche Analyse des Verfahrensfehlers führt auf eine asymptotische Entwicklung (ohne Begründung). Diese Entwicklung wird dann genützt, um verbesserte Approximationen zu berechnen (Extrapolation, Romberg-Quadratur).

  Asymptotische Entwicklung (Trapezregel) (Satz 3.1): Es sei $f \in \mathscr{C}^{2K}([a,b])$. Der Verfahrensfehler der Trapezregel (mit $h = \frac{b-a}{N}$ und $x_j = a + j\,h$ für $0 \le j \le N-1$)

$$T(h) = h\left(\tfrac{1}{2} f(x_0) + \sum_{j=1}^{N-1} f(x_j) + \tfrac{1}{2} f(x_N)\right) \quad \approx \quad \alpha_0 = \int_a^b f(x)\,\mathrm{d}x$$

  besitzt die asymptotische Entwicklung

$$T(h) - \alpha_0 = \sum_{k=1}^{K-1} \alpha_k\,h^{2k} + \alpha_K(h)\,h^{2K}, \qquad \left|\alpha_K(h)\right| \le C \int_a^b \left|f^{(2K)}(x)\right| \mathrm{d}x,$$

  mit Koeffizienten $\alpha_k \in \mathbb{R}$ für $1 \le k \le K-1$ und einer beschränkten Funktion $\alpha_K : \mathbb{R} \to \mathbb{R}$ (Schranke unabhängig von h).

  Bemerkungen:

  - Aus der obigen Relation und inbesondere wegen der Abschätzung für α_K folgt die Konvergenz der Quadraturapproximation gegen den Wert des bestimmten Integrals für $h \to 0$.
  - Für hinreichend kleine Werte der Schrittweite h ist $\alpha_1\,h^2$ der dominante Beitrag im Verfahrensfehler

$$T(h) - \alpha_0 = \alpha_1 h^2 + \alpha_2 h^4 + \cdots + \alpha_{K-1} h^{2(K-1)} + \alpha_K(h)\,h^{2K} = \alpha_1\,h^2 + \mathscr{O}(h^4)\,.$$

Elimination dominanter Fehlerterme (Richardson-Extrapolation): Die grundlegende Idee der Extrapolation ist, führende Fehlerterme durch geeignete Linearkombinationen zu verschiedenen Schrittweiten zu eliminieren.

– Für zwei verschiedene Schrittweiten $h_1 \neq h_2$ (zu zwei äquidistanten Zerlegungen des Integrationsintervalles) lauten die Entwicklungen

$$\begin{aligned} T(h_1) - \alpha_0 &= \alpha_1 h_1^2 + \alpha_2 h_1^4 + \cdots + \alpha_{K-1} h_1^{2(K-1)} + \alpha_K(h_1)\, h_1^{2K}, \\ T(h_2) - \alpha_0 &= \alpha_1 h_2^2 + \alpha_2 h_2^4 + \cdots + \alpha_{K-1} h_2^{2(K-1)} + \alpha_K(h_2)\, h_2^{2K}. \end{aligned}$$

Elimination der Unbekannten α_1 führt auf (Multiplikation und Subtraktion analog zum Gaußschen Eliminationsverfahren, Bezeichnung für restliche Terme)

$$\begin{aligned} T(h_1) - \alpha_0 &= \alpha_1 h_1^2 + \alpha_2 h_1^4 + R(h_1, 6), \\ T(h_2) - \alpha_0 &= \alpha_1 h_2^2 + \alpha_2 h_2^4 + R(h_2, 6), \\ h_2^2\big(T(h_1) - \alpha_0\big) - h_1^2\big(T(h_2) - \alpha_0\big) &= \alpha_2\, h_1^2\, h_2^2\big(h_1^2 - h_2^2\big) + h_2^2\, R(h_1, 6) - h_1^2\, R(h_2, 6), \\ h_2^2\, T(h_1) - h_1^2\, T(h_2) + \big(h_1^2 - h_2^2\big)\,\alpha_0 &= \alpha_2\, h_1^2\, h_2^2\big(h_1^2 - h_2^2\big) + h_2^2\, R(h_1, 6) - h_1^2\, R(h_2, 6), \\ \frac{h_2^2\, T(h_1) - h_1^2\, T(h_2)}{h_2^2 - h_1^2} &= \alpha_0 - \alpha_2\, h_1^2\, h_2^2 + \frac{h_2^2\, R(h_1,6) - h_1^2\, R(h_2,6)}{h_2^2 - h_1^2}. \end{aligned}$$

Somit ist die aus den (berechenbaren) Quadraturapproximationen $T(h_1), T(h_2)$ gebildete Approximation

$$\frac{h_1^2\, T(h_2) - h_2^2\, T(h_1)}{h_1^2 - h_2^2} = \alpha_0 - \alpha_2\, h_1^2 h_2^2 + \frac{h_1^2 h_2^2}{h_1^2 - h_2^2}\big(R(h_2, 4) - R(h_1, 4)\big).$$

eine verbesserte Approximation an den Wert des Integrals.

– Speziell für die Wahl $h_1 = h$ und $h_2 = \frac{h}{2}$ erhält man

$$\begin{aligned} \tfrac{4\,T(\frac{h}{2}) - T(h)}{3} &= \alpha_0 - \tfrac{1}{4}\,\alpha_2\, h^4 + \tfrac{1}{3}\, h^2\, \big(R(\tfrac{h}{2}, 4) - R(h, 4)\big), \\ \tfrac{4\,T(\frac{h}{2}) - T(h)}{3} - \alpha_0 &= \mathcal{O}\big(h^4\big). \end{aligned}$$

Vgl. Illustration (Extrapolation).

– Im allgemeinen Fall werden die Approximationen zu K paarweise verschiedenen Schrittweiten

$$T(h_i), \qquad 1 \leq i \leq K,$$

geeignet kombiniert. Eine elegante Lösung mittels Polynominterpolation geht auf Romberg zurück (vgl. Satz 3.2). Zur praktischen Durchführung verwendet man das Schema von Aitken–Neville (vgl. Schema der dividierten Differenzen, vgl. Algorithmus zur Romberg-Quadratur, Skriptum, S. 49).

Vgl. Illustration (Extrapolation).

– Eine übliche Wahl der Schrittweiten ist

$$h_1 = h\,, \qquad h_{i+1} = \tfrac{1}{2}\, h_i = \tfrac{1}{2^{i-1}}\, h\,, \quad 1 \le i \le K\,,$$

oder auch die Bulirsch-Folge

$$h_1 = h\,, \qquad h_i = \tfrac{1}{i}\, h\,, \qquad 2 \le k \le K\,.$$

Bemerkung: Das Interpolationspolynom durch einfache Stützstellen $(x_i)_{0\le i\le n}$ und zugehörige Stützwerte $y_i = f(x_i)$ für $0 \le i \le n$ ist durch

$$p = \sum_{i=0}^{n} y_i\, L_i \in \mathbb{P}_n\,, \qquad L_i(x) = \prod_{\substack{0\le j\le n\\ j\ne i}} \frac{x - x_j}{x_i - x_j}\,, \quad 0 \le i \le n\,,$$

gegeben. Im Folgenden werden die Bezeichnungen P (Interpolationspolynom) statt p (Ordnung der Quadraturformel) sowie Indizes $1 \le k \le K$ statt $0 \le i \le n$ verwendet. Die Darstellung des Interpolationspolynoms zu Datenpunkten $(x_i, y_i)_{1\le i\le K}$ lautet dann

$$P = \sum_{i=1}^{K} y_i\, L_i \in \mathbb{P}_{K-1}\,, \qquad L_i(x) = \prod_{\substack{1\le j\le K\\ j\ne i}} \frac{x - x_j}{x_i - x_j}\,, \quad 1 \le i \le K\,.$$

Extrapolation (Satz 3.2): Es sei $f \in \mathscr{C}^{2K}([a,b])$, und es bezeichne $P \in \mathbb{P}_{K-1}$ das Interpolationpolynom zu den Datenpunkten $(h_i^2, T(h_i))_{1\le i\le K}$ mit positiven und paarweise verschiedenen Schrittweiten $h_i > 0$ für $1 \le i \le k$. Dann gilt

$$P(0) - \alpha_0 = \mathscr{O}\big(h_{\max}^{2K}\big)\,, \qquad \alpha_0 = \int_a^b f(x)\,\mathrm{d}x\,, \quad h_{\max} = \max_{1\le i\le K} h_i\,.$$

Denn: Die explizite Darstellung des Interpolationspoynoms $P \in \mathbb{P}_{K-1}$ duch die Datenpunkte $(h_i^2, T(h_i))_{1\le i\le K}$ mittels Langrange-Basispolynomen lautet

$$P(x) = \sum_{i=1}^{K} T(h_i)\, L_i(x) \in \mathbb{P}_{K-1}\,, \qquad L_i(x) = \prod_{\substack{1\le j\le K\\ j\ne i}} \frac{x - h_j^2}{h_i^2 - h_j^2}\,, \quad 1 \le i \le K\,.$$

Da die Polynomfunktionen $Q_k(x) = x^k$ für $0 \le k \le K-1$ exakt durch das Interpolationspolynom zu den Stützstellen $(h_i^2)_{1\le i\le K}$ dargestellt werden, gilt

$$x^k = Q_k(x) = \sum_{j=i}^{K} Q(h_i^2)\, L_i(x) = \sum_{j=i}^{K} h_i^{2k}\, L_i(x)\,, \qquad 0 \le k \le K-1\,.$$

Einsetzen der asymptotischen Entwicklung für $T(h)$ führt auf

$$\begin{aligned} P(x) &= \sum_{i=1}^{K} T(h_i)\, L_i(x) \\ &= \sum_{i=1}^{K} \Big(\sum_{k=0}^{K-1} \alpha_k\, h_i^{2k} + \alpha_K(h_i)\, h_i^{2K} \Big) L_i(x) \\ &= \sum_{k=0}^{K-1} \alpha_k \underbrace{\sum_{i=1}^{K} h_i^{2k} L_i(x)}_{=x^k} + \sum_{i=1}^{K} \alpha_K(h_i)\, h_i^{2K} L_i(x) \\ &= \sum_{k=0}^{K-1} \alpha_k\, x^k + \sum_{i=1}^{K} \alpha_K(h_i)\, h_i^{2K} L_i(x)\,. \end{aligned}$$

Extrapolation bei $x = 0$ ergibt (Auswerten des Interpolationspolynoms an Argumenten außerhalb des betrachteten Intervalles)

$$P(0) = \alpha_0 + \sum_{i=1}^{K} \alpha_K(h_i)\, h_i^{2K} L_i(0)\,, \qquad L_i(0) = \prod_{\substack{1 \le j \le K \\ j \neq i}} \frac{h_j^2}{h_j^2 - h_i^2}\,, \quad 1 \le i \le K\,,$$

und somit die Abschätzung (Schranke für α_K, positive Schrittweiten $h_i > 0$ für $1 \le i \le K$)

$$\begin{aligned} \big| P(0) - \alpha_0 \big| &\le \sum_{i=1}^{K} |\alpha_K(h_i)|\, h_i^{2K} \prod_{\substack{1 \le j \le K \\ j \neq i}} \frac{h_j^2}{|h_j^2 - h_i^2|} \\ &\le h_{\max}^{2K} \cdot C \int_a^b \big| f^{(2K)}(x) \big|\, \mathrm{d}x \sum_{i=1}^{K} \prod_{\substack{1 \le j \le K \\ j \neq i}} \frac{h_j^2}{|h_j^2 - h_i^2|}\,. \end{aligned}$$

Dies zeigt die Behauptung. ◇

3.4. Adaptive Verfahren

- Vorsicht! In Abschnitt 3.3 wurden Schrittweiten $h_1, \dots, h_K$ zu äquidistanten Zerlegungen betrachtet. Im Folgenden geht es um die optimale Wahl der Schrittweiten einer einzigen Zerlegung, und h_j bezeichnet die Schrittweite im j-ten Teilintervall.

- Adaptivität:
  - Das Grundprinzip adaptiver Verfahren ist es, eine Balance zwischen Genauigkeit und Effizienz zu erreichen.
  - Im Zusammenhang mit Numerischer Integration ist es wünschenswert, in Bereichen wo der Integrand wenig variiert relativ große Schrittweiten zuzulassen und in Berechen wo der Integrand hingegen stark variiert relativ kleine Schrittweiten zu wählen. Aufgrund zusätzlicher Funktionsauswertungen und Rechenoperationen hat eine Verkleinerung der Schrittweite eine Verringerung der Effizienz des Verfahrens zur Folge, eine Vergrößerung der Schrittweiten steigert die Effizienz.
    Vgl. Verlauf des Integranden, Skriptum, S. 50.

- Adaptive Verfahren zur numerischen Integration (Trapezregel)
  - Es bezeichnet $a = x_0 < \cdots < x_N = b$ eine Zerlegung des Integrationsintervalles mit zugehörigen Schrittweiten $h_j = x_{j+1} - x_j$ für $0 \le j \le N-1$. Beispielsweise für die Trapezregel ist die Quadraturapproximation durch

$$Q(f) = \tfrac{1}{2} \sum_{j=0}^{N-1} h_j \big(f(x_j) + f(x_{j+1}) \big) \quad \approx \quad I(f) = \int_a^b f(x)\, \mathrm{d}x$$

    gegeben.
  - Für den Verfahrensfehler der Trapezregel gilt die Abschätzung

$$\big| Q(f) - I(f) \big| \le C\,(b-a)\, h_{\max}^p \big\| f^{(p)} \big\|_\infty, \qquad h_{\max} = \max_{0 \le j \le N-1} h_j\,.$$

    Die Vorgabe einer sehr kleinen maximalen Schrittweite würde zwar auf eine sehr gute Approximation führen

$$h_{\max} \ll 1 \quad \Rightarrow \quad \big| Q(f) - I(f) \big| \ll 1\,,$$

    wäre jedoch sehr ineffizient. Das Ziel adaptiver Verfahren ist es sicherzustellen, daß die Schrittweiten $(h_j)_{0 \le j \le N-1}$ so gewählt werden, daß die Approximation eine vorgegebene Toleranz erreicht

$$\big| Q(f) - I(f) \big| \le \mathrm{TOL}\,.$$

    Vgl. Beispiel, Skriptum, S. 58

$$\int_{-1}^{1} \frac{1}{10^{-4} + x^2}\, \mathrm{d}x\,.$$

– Grundlegende Ideen:

  * Ausgehend von einer groben Zerlegung des Intervalles (z.B. ein oder zwei Teilintervalle) bestimmt man jenes Teilintervall $I_j = [x_j, x_{j+1}]$ in welchem der (absolute oder relative) geschätzte Verfahrensfehler

$$Q_j(f) = \tfrac{1}{2}\, h_j \left(f(x_j) + f(x_{j+1}) \right) \quad \approx \quad I_j(f) = \int_{x_j}^{x_{j+1}} f(x)\, \mathrm{d}x$$

    maximal ist. Solange die Forderung (Schätzung $\widetilde{I} \approx I$)

$$\left| Q(f) - \widetilde{I}(f) \right| \leq \mathrm{TOL}$$

    verletzt ist, wird das Teilintervall I_j halbiert und die zugehörigen Quadraturpproximationen berechnet (Berechnung der zusätzlich benötigten Funktionswerte, Berechnung der Quadraturapproximationen sowie der geschätzten Verfahrensfehler auf den beiden Teilintervallen.
  * Zur Schätzung des Verfahrensfehlers verwendet man beispielsweise eine Quadraturformel höherer Ordnung wie etwa die Simpsonregel, die wenig zusätzliche Funktionsauswertungen erfordert. Eine Alternative ist auch Richardson Extrapolation.
  * Bei praktischen Berechnungen verwendet man die bessere Approximation (Unterschätzung der Schrittweite).

  Vgl. Illustration Adaptive Verfahren (Trapezregel, Simpsonregel, ohne Schätzung des Verfahrensfehlers).

4. Anfangswertprobleme für gewöhnliche Differentialgleichungen

- Inhalte:
  - Grundlegende Begriffe und Resultate zur Theorie von Anfangswertproblemen für gewöhnliche Differentialgleichungen
  - Prinzip der Diskretisierung, Verfahrensklassen, Diskretisierungsfehler
  - Konvergenzresultat, Adaptivität
  - A-Stabilität von Lösungsverfahren, Steife Differentialgleichungen
- Grundlagen:
  - Quadraturapproximationen, Interpolation
  - Lösungsverfahren für lineare und nichtlineare Gleichungssysteme

  Weitere Anwendungen:
  - Zeitdiskretisierungen partieller Differentialgleichungen
- Bemerkung: Die Analyse des Diskretisierungsfehlers unterscheidet sich von der im Skriptum gewählten Vorgehensweise.

4.1. Theoretischer Hintergrund

- Anfangswertprobleme für gewöhnliche Differentialgleichungen erster Ordnung: Für vorgegebene Punkte $t_0 \in \mathbb{R}$ (Anfangszeitpunkt) und $T \in \mathbb{R}$ (Endzeitpunkt) sei $I = [t_0, T]$ falls $t_0 < T$ (bzw. $I = [T, t_0]$ falls $t_0 > T$). Weiters sei $f : I \times \mathbb{R}^d \to \mathbb{R}^d : (t, v) \mapsto f(t, v)$ eine vorgegebene stetige Funktion und $y_0 \in \mathbb{R}^d$ ein vorgegebener Anfangswert bei t_0. Eine Lösung des Anfangswertproblems

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} y(t) = f\big(t, y(t)\big), \quad t \in (t_0, T), \qquad (\text{bzw. } (T, t_0)) \\ y(t_0) = y_0, \end{cases}$$

ist eine Funktion $y : I \to \mathbb{R}^d$, welche die (gewöhnliche) Differentialgleichung (erster Ordnung) und die Anfangsbedingung erfüllt. Dabei wird vorausgesetzt, daß y auf I stetig und zumindest in (t_0, T) (bzw. (T, t_0)) differenzierbar ist.

  Beispiel, vgl. Skriptum, S. 60 ($d = 2$, nichtautonome nichlineare Differentialgleichung).

  Bemerkungen:

  - Falls $d > 1$ spricht man auch von einem Differentialgleichungssystem.
  - Im Gegensatz zu gewöhnlichen Differentialgleichungen für Funktionen in einer Veränderlichen geben partielle Differentialgleichungen Zusammenhänge zwischen Funktionen in mehreren Veränderlichen und deren partiellen Ableitungen an.
  - In vielen Anwendungssituationen beschreibt die Variable t die Zeit und die Funktion y den zeitahängigen Zustand eines Systems (z.B. das Wachstum einer Spezies oder die Bahn eines Massenpunktes unter der Einwirkung von Kräften). Zum Zeitpunkt t_0 befindet sich das System in dem bekannten Zustand $y(t_0) = y_0$. Zu bestimmen sind die Zustände $y(t)$ zu späteren (bzw. früheren) Zeitpunkten $t \in (t_0, T)$ (bzw. $t \in (T, t_0)$).
  - Die (komponentenweise) Integration der Differentialgleichung

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t} y(t) = y'(t) &= f\big(t, y(t)\big), \quad t \in (t_0, T), \\ \int_{t_0}^{t} y'(\tau)\,\mathrm{d}\tau &= \int_{t_0}^{t} f\big(\tau, y(\tau)\big)\,\mathrm{d}\tau, \quad t \in (t_0, T), \\ y(t) - y(t_0) &= \int_{t_0}^{t} f\big(\tau, y(\tau)\big)\,\mathrm{d}\tau, \quad t \in (t_0, T), \end{aligned}$$

    und Einsetzen der Anfangsbedingung $y(t_0) = y_0$ führt auf die Integralgleichung

$$y(t) = y(t_0) + \int_{t_0}^{T} f\big(\tau, y(\tau)\big)\,\mathrm{d}\tau, \quad t \in (t_0, T).$$

    Diese wird einerseits für theoretische Untersuchungen (Existenzresultate, Picard–Lindelöf) und andererseits zur Konstruktion numerischer Verfahren (u.a. mittels Quadraturapproximationen bzw. Interpolation) genutzt.

– Eine vorteilhafte Alternative, die eine (semilineare) Formulierung der Differentialgleichung mit zeitunabhängiger Matrix $A \in \mathbb{R}^{d\times d}$ nützt, beruht auf der linearen Variation-der-Konstanten Formel (Nachprüfen durch Einsetzen bei t_0 und Differenzieren)

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t}\, y(t) &= A\, y(t) + g\big(t, y(t)\big)\,, \quad t \in (t_0, T)\,, \\ y(t) &= \mathrm{e}^{(t-t_0)A}\, y(t_0) + \int_{t_0}^{t} \mathrm{e}^{(t-\tau)A} g\big(\tau, y(\tau)\big)\, \mathrm{d}\tau\,, \quad t \in (t_0, T)\,. \end{aligned}$$

– Im Zusammenhang mit Anwendungen (z.B. aus der Mechanik) haben auch Differential-Algebraische Gleichungen besondere Bedeutung

$$\begin{cases} \begin{pmatrix} \tfrac{\mathrm{d}}{\mathrm{d}t}\, y(t) \\ 0 \end{pmatrix} = \begin{pmatrix} f\big(t, y(t), z(t)\big) \\ g\big(t, y(t), z(t)\big) \end{pmatrix}, \quad t \in (t_0, T)\,, \\ \begin{pmatrix} y(t_0) \\ z(t_0) \end{pmatrix} = \begin{pmatrix} y_0 \\ z_0 \end{pmatrix}, \end{cases}$$

die in der allgemeineren impliziten Form enthalten sind

$$\begin{cases} F\big(t, Y(t), \tfrac{\mathrm{d}}{\mathrm{d}t}\, Y(t)\big) = 0\,, \quad t \in (t_0, T)\,, \\ Y(t_0) = Y_0\,. \end{cases}$$

Falls die Funktion F bzgl. des Argumentes $\frac{\mathrm{d}}{\mathrm{d}t}\, Y(t)$ auflösbar ist, ergibt sich eine explizite Differentialgleichung der oben angegebenen Form.

- Autonome Differentialgleichungen: Falls die die Differentialgleichung erster Ordnung definierende Funktion f nicht explizit von der Variable t abhängt, heißt die Differentialgleichung eine autonome Differentialgleichung

$$\begin{cases} \tfrac{\mathrm{d}}{\mathrm{d}t}\, y(t) = f\big(y(t)\big)\,, \quad t \in (t_0, T)\,, \\ y(t_0) = y_0\,. \end{cases}$$

In diesem Fall kann man ohne Einschränkung der Allgemeinheit die Anfangszeit $t_0 = 0$ annehmen.

Reduktion auf autonome Differentialgleichungen: Jedes Anfangswertproblem für eine nichtautonome Differentialgleichung

$$\begin{cases} \tfrac{\mathrm{d}}{\mathrm{d}t}\, y(t) = f\big(t, y(t)\big)\,, \quad t \in (t_0, T)\,, \\ y(t_0) = y_0\,, \end{cases}$$

kann mittels der Definitionen

$$Y(t) = \begin{pmatrix} t \\ y(t) \end{pmatrix}, \qquad Y_0 = \begin{pmatrix} t_0 \\ y_0 \end{pmatrix}, \qquad F\big(Y(t)\big) = \begin{pmatrix} 1 \\ f\big(t, y(t)\big) \end{pmatrix},$$

auf ein Anfangswertproblem für eine autonome Differentialgleichung zurückgeführt werden

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} Y(t) = F\big(Y(t)\big), & t \in (t_0, T), \\ Y(t_0) = Y_0. \end{cases}$$

- Reduktion auf Differentialgleichungen erster Ordnung: Jedes Anfangswertproblem für eine Differentialgleichung k-ter Ordnung

$$\begin{cases} \frac{\mathrm{d}^k}{\mathrm{d}t^k} y(t) = f\big(t, y(t), \frac{\mathrm{d}}{\mathrm{d}t} y(t), \ldots, \frac{\mathrm{d}^{k-1}}{\mathrm{d}t^{k-1}} y(t)\big), & t \in (t_0, T), \\ y(t_0) = y_0, \quad \frac{\mathrm{d}}{\mathrm{d}t} y(t_0) = y_0', \quad \ldots \quad \frac{\mathrm{d}^{k-1}}{\mathrm{d}t^{k-1}} y(t_0) = y_0^{(k-1)}, \end{cases}$$

kann mittels der Definitionen

$$Y(t) = \begin{pmatrix} Y_1(t) \\ Y_2(t) \\ \vdots \\ Y_{k-1}(t) \end{pmatrix} = \begin{pmatrix} y(t) \\ \frac{\mathrm{d}}{\mathrm{d}t} y(t) \\ \vdots \\ \frac{\mathrm{d}^{k-1}}{\mathrm{d}t^{k-1}} y(t) \end{pmatrix}, \qquad Y_0 = \begin{pmatrix} y_0 \\ y_0' \\ \vdots \\ y_0^{(k-1)} \end{pmatrix},$$

$$F\big(t, Y(t)\big) = \begin{pmatrix} Y_2(t) \\ \vdots \\ Y_{k-1}(t) \\ f\big(t, y(t), \frac{\mathrm{d}}{\mathrm{d}t} y(t), \ldots, \frac{\mathrm{d}^{k-1}}{\mathrm{d}t^{k-1}} y(t)\big) \end{pmatrix} = \begin{pmatrix} Y_2(t) \\ \vdots \\ Y_{k-1}(t) \\ f\big(t, Y(t)\big) \end{pmatrix},$$

auf ein Anfangswertproblem für eine Differentialgleichung erster Ordnung zurückgeführt werden

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} Y(t) = F\big(t, Y(t)\big), & t \in (t_0, T), \\ Y(t_0) = Y_0. \end{cases}$$

Beispiel ($d = 2$, autonome lineare Differentialgleichung), vgl. Skriptum, S. 62. Eine Transformation ($Y_1 = y, Y_2 = \frac{\mathrm{d}}{\mathrm{d}t} y$) der Schwingungsgleichung (Newtonsche Bewegungsgleichung, Masse m, Auslenkung aus der Ruhelage y, Geschwindigkeit $\frac{\mathrm{d}}{\mathrm{d}t} y$, Beschleunigung $\frac{\mathrm{d}^2}{\mathrm{d}t^2} y$, Federkraft nach dem Hookschen Gesetz mit Federkonstante k, Reibungskraft durch Dämpfung z.B. in zäher Flüssigkeit mit Reibungskoeffizient r, Vorgabe der Anfangsauslenkung und Anfangsgeschwindigkeit) führt auf

$$m \tfrac{\mathrm{d}^2}{\mathrm{d}t^2} y(t) + r \tfrac{\mathrm{d}}{\mathrm{d}t} y(t) + k\, y(t) = 0,$$

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} Y_1(t) \\ Y_2(t) \end{pmatrix} = \begin{pmatrix} Y_2(t) \\ -\frac{r}{m} Y_2(t) - \frac{k}{m} Y_1(t) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{r}{m} \end{pmatrix} \begin{pmatrix} Y_1(t) \\ Y_2(t) \end{pmatrix}.$$

Vgl. Illustration (Lösung der Schwingungsgleichung mittels Matrixexponentialfunktion).

- Lineare Differentialgleichungen: Falls die rechte Seite der Funktion die spezielle Form $f : I \times \mathbb{R}^d \to \mathbb{R}^d : (t, v) \mapsto f(t, v) = A(t)v + g(t)$ mit zeitabhängiger Matrix $A : I \to \mathbb{R}^{d\times d}$ und Inhomogenität $g : I \to \mathbb{R}^d$ hat (d.h. die Funktion f ist inbesondere linear bzgl. der Variable v), heißt die Differentialgleichung eine inhomogene lineare Differentialgleichung (bzw. eine homogene lineare Differentialgleichung, falls speziell $g = 0$)

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} y(t) = A(t)\, y(t) + g(t)\,, & t \in (t_0, T)\,, \\ y(t_0) = y_0\,. \end{cases}$$

  Ähnlich wie bei Gleichungssystemen ist auch bei Differentialgleichungen der nichtlineare Fall wesentlich schwieriger als der lineare Fall.

- Vorbemerkung: Eine Funktion $f : D \subset \mathbb{R}^n \to \mathbb{R}^m$ heißt Lipschitz-stetig, wenn eine Konstante $L > 0$ existiert, sodaß für alle Elemente $x, \tilde{x} \in D$ die folgende Abschätzung mit Lipschitz-Konstante $L > 0$ gilt

$$\left\| f(x) - f(\tilde{x}) \right\| \leq L \left\| x - \tilde{x} \right\|.$$

  Resultat zur Existenz und Eindeutigkeit (Satz von Picard–Lindelöf): Falls die Funktion $f : I \times \mathbb{R}^d \to \mathbb{R}^d : (t, v) \mapsto f(t, v)$ stetig und bezüglich der Variable v Lipschitz-stetig ist, d.h. für alle $t \in I$ und $v, \tilde{v} \in \mathbb{R}^d$ gilt die Relation

$$\left\| f(t, v) - f(t, \tilde{v}) \right\| \leq L \left\| v - \tilde{v} \right\|$$

  mit Konstante $L > 0$, existiert eine eindeutig bestimmte stetig differenzierbare Funktion $y : I \to \mathbb{R}^d$, welche das Anfangswertproblem

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} y(t) = f\big(t, y(t)\big)\,, & t \in (t_0, T)\,, \\ y(t_0) = y_0\,, \end{cases}$$

  erfüllt.

  Bemerkung: Eine Abschwächung des Satzes von Picard–Lindelöf nützt die lokale Lipschitz-Stetigkeit ein einer Umgebung der Anfangsbedingung (t_0, y_0) und sichert die Existenz und Eindeutigkeit einer lokalen Lösung $y : [t_0, t_0 + \delta] \to \mathbb{R}^d$.

  Beispiel: Ein bekanntes Beispiel einer Funktion, die insbesondere bei $x = 0$ nicht (lokal) Lipschitz-stetig ist, ist die Abbildung

$$f : \mathbb{R} \to \mathbb{R} : x \mapsto f(x) = \sqrt{x}\,, \qquad \frac{f(x) - f(0)}{x - 0} = \frac{1}{\sqrt{x}}\,.$$

  Zudem gilt $f : \mathbb{R} \setminus \{0\} \to \mathbb{R} : x \mapsto \frac{1}{2\sqrt{x}}$ wegen

$$\frac{f(x+\Delta x) - f(x)}{\Delta x} = \frac{\sqrt{x+\Delta x} - \sqrt{x}}{\Delta x} \underset{\substack{\text{Erweitern mit } \sqrt{x+\Delta x}+\sqrt{x}, \\ \text{Verwendung von } (a-b)(a+b)=a^2-b^2}}{=} \frac{1}{\sqrt{x+\Delta x}+\sqrt{x}} \overset{\Delta x \to 0}{\to} f'(x) = \frac{1}{2\sqrt{x}}\,.$$

- Wesentlich in Hinblick auf die Theorie dynamischer Systeme ist das folgende Resultat zur Sensitivität der Lösung eines Anfangswertproblems bezüglich Änderungen in den Anfangswerten.

  Stetige Abhängigkeit der Lösung von den Anfangswerten (Kondition, vgl. Satz 4.2): Falls die Funktion $f : I \times \mathbb{R}^d \to \mathbb{R}^d : (t, v) \mapsto f(t, v)$ stetig und bezüglich der Variable v Lipschitz-stetig mit Konstante $L > 0$ ist, gilt für zwei Lösungen $y, \widetilde{y}$ der Differentialgleichung

  $$\tfrac{\mathrm{d}}{\mathrm{d}t} z(t) = f\big(t, z(t)\big), \qquad t \in (t_0, T),$$

  zu den Anfangswerten $y_0, \widetilde{y}_0$ die Abschätzung

  $$\big\| y(t) - \widetilde{y}(t) \big\| \leq \mathrm{e}^{L|t-t_0|} \big\| y_0 - \widetilde{y}_0 \big\|, \qquad t \in (t_0, T).$$

  Bemerkung: Das exponentielle Auseinanderdriften von Lösungen zeigt sich beispielsweise bei der folgenden skalaren Differentialgleichung mit $\lambda \geq 0$

  $$\tfrac{\mathrm{d}}{\mathrm{d}t} z(t) = \lambda\, z(t), \quad z(t) = \mathrm{e}^{\lambda(t-t_0)} z(t_0), \qquad t \geq t_0.$$

4.2. Diskretisierungen und Diskretisierungsfehler

- Zeitintegrationsverfahren (Zeitdiskretisierungsverfahren) für Anfangswertprobleme der Form

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}\, y(t) = f\big(t, y(t)\big), & t \in (t_0, T), \\ y(t_0) \quad \text{gegeben}, \end{cases}$$

beruhen auf folgendem Zugang (time-stepping approach). Ausgehend von einem näherungsweisen Anfangswert

$$y_0 \approx y(t_0)$$

berechnet man an gewissen Zeitpunkten mit zugehörigen Zeitschrittweiten

$$t_0 < t_1 < \cdots < t_N = T, \qquad h_i = t_{i+1} - t_i, \quad 0 \le i \le N-1,$$

mittels einer Rekursion Näherungswerte an die exakten Lösungswerte

$$y_n \approx y(t_n), \qquad 1 \le n \le N.$$

Bei Einschrittverfahren hat die Rekursion die Form

$$y_n = \Phi(h_{n-1}, t_{n-1}, y_{n-1}), \quad 1 \le n \le N, \qquad y_0 \text{ gegeben}.$$

Zur Konstruktion und Analyse von Zeitintegrationsverfahren für (lineare) Differentialgleichungen ist es vorteilhaft, den numerischen Lösungsoperator Φ mit dem exakten Lösungsoperator E zu vergleichen

$$y_n = \Phi(h_{n-1}, t_{n-1}, y_{n-1}) \approx y(t_n) = E\big(h_{n-1}, t_{n-1}, y(t_{n-1})\big), \quad 1 \le n \le N, \qquad y_0 \text{ gegeben}.$$

Allgemeiner hängt bei Mehrschrittverfahren (k-Schrittverfahren) der numerische Lösungsoperator von zuvor berechneten Approximationen ab

$$y_n = \Phi(h_{n-1}, \ldots, h_{n-k}, t_{n-1}, y_{n-1}, \ldots, y_{n-k}), \qquad k \le n \le N.$$

Neben dem Startwert sind dann geeignete Approximationen $y_1, \ldots, y_{k-1}$ (z.B. mittels eines Einschrittverfahrens) zu berechnen.

- Beispiele für (einfache) Zeitintegrationsverfahren:
  - Explizites Eulerverfahren (*Explicit Euler, Forward Euler*):
    * Das explizite Eulerverfahren beruht auf der Idee, in der Differentialgleichung (Auswerten bei $t = t_{n-1}$)

$$\frac{\mathrm{d}}{\mathrm{d}t}\, y(t)\Big|_{t=t_{n-1}} = f\big(t_{n-1}, y(t_{n-1})\big), \qquad 1 \le n \le N,$$

den Differentialquotienten durch den Differenzenquotienten (Vorwärtsdifferenz, *forward*)

$$\frac{y(t_n) - y(t_{n-1})}{t_n - t_{n-1}} \quad \approx \quad \frac{\mathrm{d}}{\mathrm{d}t}\, y(t)\Big|_{t=t_{n-1}} = \lim_{t \to t_{n-1}} \frac{y(t) - y(t_{n-1})}{t - t_{n-1}}$$

zu ersetzen. Dies führt auf die Rekursion (explizite Relation)

$$y_n = y_{n-1} + h_{n-1}\, f(t_{n-1}, y_{n-1}), \quad 1 \le n \le N, \qquad y_0 \text{ gegeben}.$$

* Eine alternative Herleitung des expliziten Eulerverfahrens beruht auf der Integralgleichung (Integration der Differentialgleichung mit Integrationsintervall $[t_{n-1}, t_n]$)
$$y(t_n) = y(t_{n-1}) + \int_{t_{n-1}}^{t_n} f\big(\tau, y(\tau)\big)\,\mathrm{d}\tau\,,$$
und Polynominterpolation (Grad 0, Stützstelle bei t_{n-1})
$$y(t_n) = y(t_{n-1}) + h_{n-1}\, f\big(t_{n-1}, y(t_{n-1})\big)\,.$$
* Vgl. Skizze, Skriptum, S.66.

– Implizites Eulerverfahren (*Implicit Euler, Backward Euler*):

* Das implizite Eulerverfahren beruht auf der Idee, in der Differentialgleichung (Auswerten bei $t = t_n$)
$$\tfrac{\mathrm{d}}{\mathrm{d}t}\, y(t)\big|_{t=t_n} = f\big(t_n, y(t_n)\big)\,, \qquad 1 \le n \le N\,,$$
den Differentialquotienten durch den Differenzenquotienten (Rückwärtsdifferenz, *backward*)
$$\tfrac{y(t_n)-y(t_{n-1})}{t_n-t_{n-1}} \quad \approx \quad \tfrac{\mathrm{d}}{\mathrm{d}t}\, y(t)\big|_{t=t_n} = \lim_{t\to t_n} \tfrac{y(t)-y(t_n)}{t-t_n}$$
zu ersetzen. Dies führt auf die Rekursion
$$y_n = y_{n-1} + h_{n-1}\, f(t_n, y_n)\,, \quad 1 \le n \le N\,, \qquad y_0 \text{ gegeben}\,.$$
* Im Gegensatz zum expliziten Eulerverfahren ist beim impliziten Eulerverfahren in jedem Zeitschritt ein nichtlineares Gleichungssystem zu lösen (vgl. Numerische Mathematik I). Aufgrund des besseren Stabilitätsverhaltens, lohnt sich im Zusammenhang mit *steifen* Differentialgleichungen (Abschnitt 4.4) der Mehraufwand.
* Eine alternative Herleitung des impliziten Eulerverfahrens beruht auf der Integralgleichung (Integration der Differentialgleichung mit Integrationsintervall $[t_{n-1}, t_n]$)
$$y(t_n) = y(t_{n-1}) + \int_{t_{n-1}}^{t_n} f\big(\tau, y(\tau)\big)\,\mathrm{d}\tau\,,$$
und Polynominterpolation (Grad 0, Stützstelle bei t_n)
$$y(t_n) = y(t_{n-1}) + h_{n-1}\, f\big(t_n, y(t_n)\big)\,.$$

– Mittelpunktsregeln (Einschrittverfahren, Zweischrittverfahren):

* Eine Herleitung einer Mittelpunktsregel beruht ebenfalls auf der Integralgleichung (Integration der Differentialgleichung mit Integrationsintervall $[t_{n-1}, t_n]$)
$$y(t_n) = y(t_{n-1}) + \int_{t_{n-1}}^{t_n} f\big(\tau, y(\tau)\big)\,\mathrm{d}\tau\,,$$

und Polynominterpolation (Grad 1, Mittelpunkt $t_{n-1} + \frac{1}{2} h_{n-1}$, vgl. entsprechende Quadraturapproximation)

$$y(t_n) = y(t_{n-1}) + h_{n-1} f\Big(t_{n-1} + \tfrac{1}{2} h_{n-1}, y\big(t_{n-1} + \tfrac{1}{2} h_{n-1}\big)\Big).$$

Zur Approximation des unbekannten Lösungswertes kann man einerseits das explizite Eulerverfahren verwenden

$$y\big(t_{n-1} + \tfrac{1}{2} h_{n-1}\big) \approx y_{n-1} + \tfrac{1}{2} h_{n-1} f(t_{n-1}, y_{n-1})$$

und erhält dann die Rekursion für die (explizite) Mittelpunktsregel

$$y_n = y_{n-1} + h_{n-1} f\big(t_{n-1} + \tfrac{1}{2} h_{n-1}, y_{n-1} + \tfrac{1}{2} h_{n-1} f(t_{n-1}, y_{n-1})\big), \quad 1 \le n \le N.$$

Verwendet man hingegen das implizite Eulerverfahren (Lösung eines nichtlinearen Gleichungssystems zur Berechnung der Hilfsapproximation $Y_{n-1,1}$)

$$y\big(t_{n-1} + \tfrac{1}{2} h_{n-1}\big) \approx Y_{n-1,1} = y_{n-1} + \tfrac{1}{2} h_{n-1} f(t_{n-1} + \tfrac{1}{2} h_{n-1}, Y_{n-1,1}),$$

ergibt sich die (implizite) Mittelpunktsregel

$$\begin{aligned} Y_{n-1,1} &= y_{n-1} + \tfrac{1}{2} h_{n-1} f(t_{n-1} + \tfrac{1}{2} h_{n-1}, Y_{n-1,1}), \\ y_n &= y_{n-1} + h_{n-1} f\big(t_{n-1} + \tfrac{1}{2} h_{n-1}, Y_{n-1,1}\big), \quad 1 \le n \le N, \qquad y_0 \text{ geg.} \end{aligned}$$

* Bemerkung: Die implizite Mittelpunktsregel ist ein erstes Beispiel eines impliziten Runge–Kutta Verfahrens (Definition, s.u.), das sich speziell für die Wahl der folgenden Verfahrenskoeffizienten ergibt

$$s = 1, \quad c_1 = \tfrac{1}{2}, \quad a_{11} = \tfrac{1}{2}, \quad b_1 = 1.$$

* Betrachtet man die Integralgleichung (Integration der Differentialgleichung mit Integrationsintervall $[t_{n-2}, t_n]$, zur Vereinfachung konstante Zeitschrittweite h)

$$y(t_n) = y(t_{n-2}) + \int_{t_{n-2}}^{t_n} f\big(\tau, y(\tau)\big)\, \mathrm{d}\tau,$$

und Polynominterpolation (Grad 1, Mittelpunkt t_{n-1}, vgl. entsprechende Quadraturapproximation)

$$y(t_n) = y(t_{n-2}) + 2h f\big(t_{n-1}, y(t_{n-1})\big),$$

so ergibt sich ein explizites Zweischrittverfahren (vgl. Skriptum, S. 67)

$$y_n = y_{n-2} + 2h f(t_{n-1}, y_{n-1}), \quad 2 \le n \le N, \qquad y_0 \text{ gegeben}.$$

Zur Bestimmung der unbekannten Approximation y_1 verwendet man üblicherweise ein Einschrittverfahren (z.B. explizites Eulerverfahren mit hinreichend kleiner Zeitschrittweite zur Erhaltung der Konvergenzordnung).

– Trapezregel: Die Trapezregel beruht auf der Integralgleichung (Integration der Differentialgleichung mit Integrationsintervall $[t_{n-1}, t_n]$)

$$y(t_n) = y(t_{n-1}) + \int_{t_{n-1}}^{t_n} f\big(\tau, y(\tau)\big)\,\mathrm{d}\tau\,,$$

und Polynominterpolation (Grad 1, Intervallenden t_{n-1} und t_n als Stützstellen, vgl. entsprechende Quadraturapproximation)

$$y(t_n) = y(t_{n-1}) + \tfrac{1}{2}\,h_{n-1}\Big(f\big(t_{n-1}, y(t_{n-1})\big) + f\big(t_n, y(t_n)\big)\Big).$$

Dies führt auf die Rekursion für die Trapezregel

$$y_n = y_{n-1} + \tfrac{1}{2}\,h_{n-1}\big(f(t_{n-1}, y_{n-1}) + f(t_n, y_n)\big)\,, \quad 1 \le n \le N\,, \qquad y_0 \text{ gegeben}.$$

Die Trapezregel ist ebenfalls ein Beispiel eines (impliziten) Runge–Kutta Verfahrens (s.u.).

- Klassifizierung von gebräuchlichen Zeitintegrationsverfahren (Runge–Kutta Verfahren, Lineare Mehrschrittverfahren):
  – Ein Runge–Kutta Verfahren besitzt die Form

$$\begin{cases} Y'_{n-1,i} = f(t_{n-1} + c_i\,h_{n-1}, Y_{n-1,i})\,, \\ Y_{n-1,i} = y_{n-1} + h_{n-1}\displaystyle\sum_{j=1}^{s} a_{ij}\,Y'_{n-1,j}\,, \end{cases}$$

$$y_n = y_{n-1} + h_{n-1}\sum_{i=1}^{s} b_i\,Y'_{n-1,i}\,, \quad 1 \le n \le N\,, \qquad y_0 \text{ gegeben},$$

mit Hilfsapproximationen $(Y_{n-1,i})_{1\le i\le s}$ und $(Y'_{n-1,i})_{1\le i\le s}$ (Stufen) an die Funktionswerte bzw. Ableitungen der exakten Lösung $Y_{n-1,i} \approx y(t_{n-1} + c_i h_{n-1})$ bzw. $Y'_{n-1,i} \approx y'(t_{n-1} + c_i h_{n-1}) = f\big(t_{n-1} + c_i h_{n-1}, y(t_{n-1} + c_i h_{n-1})\big)$. Ersetzt man die internen Stufen, so ergibt sich die alternative Darstellung

$$Y'_{n-1,i} = f\Big(t_{n-1} + c_i\,h_{n-1}, y_{n-1} + h_{n-1}\sum_{j=1}^{s} a_{ij}\,Y'_{n-1,j}\Big),$$

$$y_n = y_{n-1} + h_{n-1}\sum_{i=1}^{s} b_i\,Y'_{n-1,i}\,, \quad 1 \le n \le N\,, \qquad y_0 \text{ gegeben}.$$

Ein Runge–Kutta Verfahren (und damit die Approximationsgüte der Näherungslösung) ist durch die Angabe der Verfahrenskoeffizienten im (John) Butcher Tableau

$c = (c_i)_{1\le i\le s}$	$A = (a_{ij})_{1\le i,j\le s}$
	$b = (b_i)_{1\le i\le s}$

festgelegt. Das Verfahren ist explizit, falls $a_{ij} = 0$ für alle $1 \le i \le j \le s$, ansonsten implizit.

– Lineare Mehrschrittverfahren benützen die Kenntnis von Approximationen zu früheren Zeitpunkten. Ein lineares k-Schrittverfahren ist durch eine Rekursion der Form (zur Vereinfachung für konstante Zeitschrittweiten h formuliert, bei variablen Schrittweiten hängen die Koeffizienten des Verfahrens von Schrittweitenverhältnissen ab)

$$\sum_{i=0}^{k} \alpha_i \, y_{n-k+i} = h \sum_{i=0}^{k} \beta_i \, f(t_{n-k+i}, y_{n-k+i})$$

mit Verfahrenskoeffizienten $(\alpha_i, \beta_i)_{0\le i\le k}$ gegeben. Falls der Koeffizient β_k nicht auftritt (d.h. $\beta_k = 0$), ist das Verfahren explizit, ansonsten implizit.

Die Konstruktion der bekanntesten Linearen Mehrschrittverfahren beruht auf Interpolation.

* Bei den BDF-Verfahren (Backward Differentiation Formulae) betrachtet man das Interpolationspolynom $P \in \mathbb{P}_k$ durch die Datenpunkte $(t_{n-k+i}, y_{n-k+i})_{0\le i\le k}$ für $k \ge 1$ und bestimmt den unbekannten Approximationswert y_n so, daß die Forderung

$$\tfrac{\mathrm{d}}{\mathrm{d}t} P(t)\big|_{t=t_n} = f\big(t_n, P(t_n)\big) = f(t_n, y_n)$$

erfüllt ist, d.h. das Interpolationspolynom P erfüllt die Differentialgleichung in t_n. Offensichtlich führt dieser Zugang auf ein implizites k-Schrittverfahren.
(i) Für $k = 1$ führt der Ansatz

$$P(t) = y_{n-1} + (t - t_{n-1}) \, \tfrac{1}{h} (y_n - y_{n-1}), \qquad \tfrac{\mathrm{d}}{\mathrm{d}t} P(t) = \tfrac{1}{h} (y_n - y_{n-1}),$$

und die Forderung

$$\tfrac{1}{h} (y_n - y_{n-1}) = \tfrac{\mathrm{d}}{\mathrm{d}t} P(t)\big|_{t=t_n} = f(t_n, y_n)$$

auf das implizite Eulerverfahren

$$y_n = y_{n-1} + h \, f(t_n, y_n), \qquad 1 \le n \le N.$$

(ii) Für $k = 2$ führt beispielsweise der Ansatz (Interpolation nach Newton, Schema dividierter Differenzen)

$$\begin{aligned} P(t) &= y_{n-2} + (t - t_{n-2}) \, \tfrac{1}{h} (y_{n-1} - y_{n-2}) \\ &\quad + (t - t_{n-2}) (t - t_{n-1}) \, \tfrac{1}{h^2} (y_n - 2\, y_{n-1} + y_{n-2}), \\ \tfrac{\mathrm{d}}{\mathrm{d}t} P(t) &= \tfrac{1}{h} (y_{n-1} - y_{n-2}) + (2\, t - t_{n-1} - t_{n-2}) \, \tfrac{1}{2h^2} (y_n - 2\, y_{n-1} + y_{n-2}), \end{aligned}$$

und die Forderung

$$\begin{aligned} &\tfrac{1}{h} (y_{n-1} - y_{n-2}) + (2\, t_n - t_{n-1} - t_{n-2}) \, \tfrac{1}{2h^2} (y_n - 2\, y_{n-1} + y_{n-2}) \\ &\quad = \tfrac{1}{h} (y_{n-1} - y_{n-2}) + \tfrac{3}{2h} (y_n - 2\, y_{n-1} + y_{n-2}) \\ &\quad = \tfrac{1}{2h} \big(3\, y_n - 4\, y_{n-1} + y_{n-2}\big) \\ &\quad = \tfrac{\mathrm{d}}{\mathrm{d}t} P(t)\big|_{t=t_n} = f(t_n, y_n) \end{aligned}$$

auf das bekannte Zweischrittverfahren BDF 2 (zuvor angegebene Form mit Koeffizienten $\alpha_0 = \frac{1}{2}$, $\alpha_1 = -2$, $\alpha_2 = \frac{3}{2}$, $\beta_0 = 0 = \beta_1$, $\beta_2 = 1$)

$$\tfrac{3}{2}\, y_n - 2\, y_{n-1} + \tfrac{1}{2}\, y_{n-2} = h\, f(t_n, y_n)\,, \qquad 2 \le n \le N\,.$$

* Bei den Adamsverfahren betrachtet man die Integralgleichung

$$y(t_n) = y(t_{n-1}) + \int_{t_{n-1}}^{t_n} f\big(\tau, y(\tau)\big)\, \mathrm{d}\tau$$

und ersetzt den Integranden durch ein Interpolationspolynom zu gewissen Datenpunkten. Im Fall der expliziten Adams-Verfahren (Adams–Bashforth Verfahren) bestimmt man das Interpolationspolynom $P \in \mathbb{P}_{k-1}$ durch $\big(t_{n-k+i}, f(t_{n-k+i}, y_{n-k+i})\big)_{0\le i\le k-1}$, was auf ein explizites k-Schrittverfahren der Form

$$y_n = y_{n-1} + h \sum_{i=0}^{k-1} \beta_i\, f(t_{n-k+i}, y_{n-k+i})\,, \qquad k \le n \le N\,,$$

führt. Bei den impliziten Adams-Verfahren (Adams–Moulton Verfahren) bestimmt man das Interpolationspolynom $P \in \mathbb{P}_k$ durch $\big(t_{n-k+i}, f(t_{n-k+i}, y_{n-k+i})\big)_{0\le i\le k}$ (beinhaltet die unbekannte Approximation y_n) und erhält damit ein implizites k-Schrittverfahren der Form

$$y_n = y_{n-1} + h \sum_{i=0}^{k} \beta_i\, f(t_{n-k+i}, y_{n-k+i})\,, \qquad k \le n \le N\,.$$

Bemerkung: Zur näherungsweisen Lösung eines nichtlinearen Gleichungssystems kann man die Idee der Fixpunktiteration verwenden. Ausgehend von einem geeigneten Startwert werden Approximationen an eine Lösung des Problems

$$\eta = F(\eta)$$

mittels der Iteration

$$\eta_i = F(\eta_{i-1})\,, \quad i = 1, 2, \ldots$$

berechnet. Im Zusammenhang mit impliziten Adamsverfahren führt dieser Zugang auf Prädiktor-Korrektor Verfahren, d.h. zur Bestimmung geeigneter Startwerte wendet man das explizite Adamsverfahren an und berechnet anschließend (einige wenige) Werte mittels Fixpunktiteration.

Illustrationen (Explizite Verfahren der Ordnungen $p = 1, 2, 4$ für eine skalare Testgleichung mit bekannter Lösung, implizite Verfahren der Ordnung $p = 1, 2$ für eine skalare lineare Testgleichung, Explizites Zweischrittverfahren der Ordnung $p = 2$).

- Fragestellung: Güte der Approximation mittels Zeitintegrationsverfahren.

  Vorbemerkung: Zur Vereinfachung werden im Folgenden äquidistante Zeitgitter betrachtet, d.h. es gelte

  $$t_n = t_0 + n\,h\,, \quad 0 \le n \le N\,, \qquad h = \tfrac{T-t_0}{N}\,.$$

  Ansonsten ist in den globalen Fehlerabschätzungen die konstante Schrittweite h durch die maximale Schrittweite

  $$h_{\max} = \max_{0\le n\le N-1} h_n\,, \qquad h_n = t_{n+1} - t_n\,, \quad 0 \le n \le N-1\,,$$

  zu ersetzen.

  Globaler und lokaler Fehler:

  – Um die Güte der mittels eines Zeitintegrationsverfahrens bestimmten Approximationen an die exakten Lösungwerte

  $$y_0\,, y_1\,, y_2\,, \ldots\,, y_N\,, \qquad y_n \approx y(t_n)\,, \quad 0 \le n \le N\,,$$

  beurteilen zu können, definiert man den globalen Fehler

  $$e_N = y_N - y(t_N)\,.$$

  Unter der Annahme, daß Anfangswertprobleme betrachtet werden, die durch eine hinreichend reguläre Funktion $f : I \times \mathbb{R}^d \to \mathbb{R}^d$ gegeben sind, heißt ein Zeitintegrationsverfahren konvergent mit Konvergenzordnung $p \ge 1$, falls eine Abschätzung der Form (für hinreichend kleine Schrittweiten $0 < h \le \overline{h}$)

  $$\big\| y_N - y(t_N) \big\| \le C\,h^p$$

  bzw. in Kurzschreibweise die Relation

  $$e_N = \mathcal{O}\big(h^p\big)$$

  gilt. Wesentlich ist dabei, daß die Konstante C weder von der Schrittweite h noch von der Anzahl der Zeitschritte N abhängt. Bei nichtlinearen Problemen ist die Existenz der numerischen Lösung im Allgemeinen nur für Schrittweiten $0 < h \le \overline{h}$ sichergestellt.

  – Zur Analyse des globalen Verfahrensfehlers (und auch zur Konstruktion von Verfahren) ist es im Allgemeinen vorteilhaft, den lokalen Fehler eines Zeitintegrationsverfahrens zu untersuchen. Bei Einschrittverfahren betrachtet man die Differenz zwischen numerischer Approximation und exaktem Lösungwert, ausgehend von einem gemeinsamen (beliebigen) Anfangswert (zur Vereinfachung der Notation wird der Startzeitpunkt t_0 nicht angegeben)

  $$y_1 - y(t_1) = \Phi(h, y_0) - E(h, y_0)\,.$$

Ein Einschrittverfahren hat Konsistenzordnung $p \geq 0$ (im Unterschied zur Konvergenzordnung), wenn

$$y_1 - y(t_1) = \mathcal{O}\big(h^{p+1}\big).$$

Bei einem k-Schrittverfahren geht man von einem (beliebigen) Anfangswert y_0 und Startwerten $y_1, \ldots, y_{k-1}$ aus, die *exakten* Lösungwerten entsprechen (insbesondere gilt $y_{k-1} = y(t_{k-1})$), und betrachtet dann die Differenz

$$y_k - y(t_k) = \Phi(h, y_0, \ldots, y_{k-1}) - E(h, y_{k-1}).$$

– Ein Hauptresultat der Numerik von Zeitintegrationsverfahren besagt, daß unter gewissen Stabilitätsforderungen aus einer lokalen Fehlerentwicklung

$$y_1 - y(t_1) = \mathcal{O}\big(h^{p+1}\big)$$

die globale Fehlerentwicklung

$$y_N - y(t_N) = \mathcal{O}\big(h^{p}\big)$$

folgt (bei hinreichend regulärer Funktion f und geeignet gewählten Startwerten). Kurz gefaßt

Konsistenz + Stabilität = Konvergenz

Ein wesentlicher Schritt ist somit die Analyse des lokalen Fehlers für verschiedene Verfahrensklassen.

Beispiele von lokalen Fehlerentwicklungen für (einfache) Zeitintegrationsverfahren:

– Vorbemerkungen:
  * Bei der Konvergenzanalyse eines Zeitintegrationsverfahrens betrachtet man als ersten Schritt eine skalare lineare Testgleichung (Dahlquist Testgleichung, autonome Differentialgleichung, Anfangszeitpunkt $t_0 = 0$)

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\, y(t) = f\big(t, y(t)\big) = \lambda\, y(t), \quad t \geq 0, \quad y(0) = y_0, \qquad \lambda \in \mathbb{R}, \qquad \text{(bzw. } \lambda \in \mathbb{C}\text{)}$$

  mit exakter Lösung

$$y(t) = \mathrm{e}^{t\lambda} y_0, \qquad t \geq 0.$$

  * Aufgrund der Linearität der Testgleichung ist es zudem ausreichend, den numerischen und exakten Lösungsoperator zu betrachten, d.h. man verwendet die einfache Abhängigkeit der numerischen und exakten Lösung vom Anfangswert $y_1 = \Phi(h, y_0) = \Phi(h)\, y_0$ und $y(h) = E(h, y_0) = E(h)\, y_0$.
  * Eine Abschätzung des globalen Verfahrensfehlers mittels einer Abschätzung für lokale Verfahrensfehler und Stabilitäsabschätzungen basiert auf der

Teleskopsumme (selbes Prinzip für Skalare $\alpha, \beta \in \mathbb{R}$ bzw. Matrizen $\alpha, \beta \in \mathbb{R}^{d\times d}$ bzw. sogar Lösungsoperatoren)

$$\begin{aligned}\alpha^k - \beta^k &= \underbrace{\alpha^k - \alpha^{k-1}\beta}_{=\alpha^{k-1}(\alpha-\beta)} + \underbrace{\alpha^{k-1}\beta - \alpha^{k-2}\beta^2}_{\alpha^{k-2}(\alpha-\beta)\beta} + \underbrace{\alpha^{k-2}\beta^2 - \alpha^{k-3}\beta^3}_{=\alpha^{k-3}(\alpha-\beta)\beta^2} \pm \cdots + \underbrace{\alpha\beta^{k-1} - \beta^k}_{=(\alpha-\beta)\,\beta^{k-1}} \\ &= \alpha^{k-1}(\alpha-\beta) + \alpha^{k-2}(\alpha-\beta)\beta + \alpha^{k-3}(\alpha-\beta)\beta^2 \pm \cdots + (\alpha-\beta)\,\beta^{k-1} \\ &= \sum_{j=0}^{k-1} \alpha^{k-1-j}\,(\alpha-\beta)\,\beta^j\,.\end{aligned}$$

* Vorsicht! Bei der Analyse der lokalen Fehlers nimmt man an, daß die Zeitschrittweite h hinreichend klein ist und somit eine Entwicklung des Wertes der Exponentialfunktion bei h sinnvoll ist (die Anfangsglieder der Entwicklung sind dominant)

$$\mathrm{e}^{h\lambda} = 1 + h\,\lambda + \tfrac{1}{2}\,h^2\lambda^2 + \underbrace{\tfrac{1}{6}\,h^3\lambda^3 + \ldots}_{\text{vergleichsweise klein}}$$

Die Verwendung einer Entwicklung beim Endzeitpunkt T

$$\mathrm{e}^{T\lambda} = 1 + T\,\lambda + \tfrac{1}{2}\,T^2\lambda^2 + \tfrac{1}{6}\,T^3\lambda^3 + \ldots\,.$$

ist im Allgemeinen jedoch nicht sinnvoll (die Anfangsglieder der Entwicklung sind *nicht* dominant).

– Explizites Eulerverfahren:

* Das explizite Eulerverfahren (einzelner Zeitschritt der Länge h)

$$y_1 = y_0 + h\,f(t_0, y_0)$$

angewendet auf die Testgleichung ergibt die Approximation (Taylorreihenentwicklung der Exponentialfunktion)

$$\begin{aligned} y_1 = \Phi(h)\,y_0 \quad &\approx \quad y(h) = E(h)\,y_0\,, \\ \Phi(h) = 1 + h\,\lambda \quad &\approx \quad E(h) = \mathrm{e}^{h\lambda} = 1 + h\,\lambda + \mathscr{O}\big(h^2\big)\,. \end{aligned}$$

* Der lokale Verfahrensfehler des expliziten Eulerverfahrens

$$y_1 - y(h) = \big(\Phi(h) - E(h)\big)\,y_0$$

erfüllt somit die Abschätzung (Konsistenzordnung $p = 1$, d.h. $p + 1 = 2$)

$$\big|y_1 - y(h)\big| \le C\,h^2\,.$$

* Die Approximation zum Endzeitpunkt $t_N = T$ ist durch

$$y_N = \big(1 + h\,\lambda\big)^N\,y_0 \quad \approx \quad y(t_N) = \mathrm{e}^{Nh\lambda}\,y_0$$

gegeben.

* Einfacher als die direkte Analyse des globalen Verfahrensfehlers (insbesondere für allgemeine nichtlineare Differentialgleichungen)

$$y_N - y(t_N) = \big(\Phi(h)^N - E(t_N)\big)\, y_0 = \Big(\big(1+h\,\lambda\big)^N - \mathrm{e}^{Nh\lambda}\Big)\, y_0$$

ist die Verwendung der Relation (Teleskopsumme)

$$\alpha^k - \beta^k = \sum_{j=0}^{k-1} \alpha^{k-1-j}\,(\alpha-\beta)\,\beta^j\,,$$

die auf die Relation (Lady Winderemere Fächer, für den exakten Lösungsoperator gilt die Gleichheit $E(h)^N = E(Nh)$)

$$y_N - y(t_N) = \big(\Phi(h)^N - E(h)^N\big)\, y_0 = \sum_{j=0}^{N-1} \Phi(h)^{N-1-j}\,\big(\Phi(h) - E(h)\big)\,E(jh)\,y_0$$

und damit auf die Abschätzung

$$\begin{aligned}\big|y_N - y(t_N)\big| &\le \sum_{j=0}^{N-1} \underbrace{\big|\Phi(h)\big|^{N-1-j}}_{=|1+h\lambda|^{N-1-j}\le C}\ \underbrace{\big|\Phi(h)-E(h)\big|}_{\le Ch^2}\ \underbrace{\big|E(jh)\,y_0\big|}_{=|y(t_j)|\le C} \le Ch\ \underbrace{\sum_{j=0}^{N-1} h}_{=Nh=T}\\ &\le C\,h\end{aligned}$$

führt.

Bemerkung: Das exponentielle Anwachsen der Lösungen für $\lambda > 0$ spiegelt sich in der Stabilitätsschranke (Annahme $h > 0$)

$$\big|\Phi(h)\big|^n = |1+h\lambda|^n \le \mathrm{e}^{\lambda n h}\,.$$

In dieser Situation ist ein exponentielles Auseinanderdriften (Fehlerschranke im wesentlichen $|y_N - y(t_N)| \le C\,\mathrm{e}^{\lambda T} h^p$) der numerischen und exakten Lösung unvermeidbar. Insbesondere über lange Zeiten $T \gg 1$ ist die Bestimmung genauer Approximationen mit sehr hohem Rechenaufwand verbunden bzw. man kann sich nicht erwarten, daß die Ergebnisse quantitativ korrekt sind (chaotische Systeme, qualitative Theorie).

– Implizites Eulerverfahren: Das implizite Eulerverfahren

$$y_1 = y_0 + h\,f(h, y_1)$$

angewendet auf die Testgleichung ergibt die Approximation (Taylorreihenentwicklung der Exponentialfunktion, geometrische Reihe für $|h\lambda| < 1$ anwendbar)

$$\begin{aligned} y_1 = \Phi(h)\,y_0 \quad &\approx \quad y(h) = E(h)\,y_0\,,\\ \Phi(h) = (1-h\,\lambda)^{-1} = 1 + h\,\lambda + \mathscr{O}\big(h^2\big) \quad &\approx \quad E(h) = \mathrm{e}^{h\lambda} = 1 + h\,\lambda + \mathscr{O}\big(h^2\big)\,.\end{aligned}$$

Der lokale Verfahrensfehler des impliziten Eulerverfahrens erfüllt somit die Abschätzung (Ordnung $p = 1$, d.h. $p+1 = 2$, ebenso wie das explizite Eulerverfahren)

$$\big|y_1 - y(h)\big| \le C\,h^2\,.$$

Die Approximation zum Endzeitpunkt $t_N = T$ ist durch

$$y_N = \big(1 - h\,\lambda\big)^{-N} y_0 \quad \approx \quad y(t_N) = \mathrm{e}^{Nh\lambda}\,y_0$$

gegeben. Wie zuvor beruht die Analyse des globalen Verfahrensfehlers auf der Relation

$$y_N - y(t_N) = \big(\Phi(h)^N - E(h)^N\big)\,y_0 = \sum_{j=0}^{N-1} \Phi(h)^{N-1-j}\,\big(\Phi(h) - E(h)\big)\,E(jh)\,y_0$$

und führt damit auf die Abschätzung (für $\lambda \le 0$ und wegen $h > 0$ ist die Schranke $|1 - h\lambda|^{-1} \le 1$ immer gültig!)

$$\begin{aligned}\big|y_N - y(t_N)\big| &\le \sum_{j=0}^{N-1} \underbrace{\big|\Phi(h)\big|^{N-1-j}}_{=|1-h\lambda|^{-(N-1-j)}\le C}\ \underbrace{\big|\Phi(h) - E(h)\big|}_{\le Ch^2}\ \underbrace{\big|E(jh)\,y_0\big|}_{=|y(t_j)|\le C} \le Ch \underbrace{\sum_{j=0}^{N-1} h}_{=Nh=T}\\ &\le C\,h\,.\end{aligned}$$

– Verallgemeinerung: Zur Konvergenzanalyse von Einschritt- und Mehrschrittverfahren für allgemeine nichtlineare Differentialgleichungen mit (lokal) Lipschitzstetiger definierender Funktion verwendet man dasselbe Prinzip wie beim expliziten und impliziten Eulerverfahren

$$\begin{aligned}\underbrace{\big\|y_N - y(t_N)\big\|}_{\text{Globaler Fehler}} &\le \sum_{j=0}^{N-1} \underbrace{\big\|\Phi(h)\big\|^{N-1-j}}_{\substack{\le C\\ \text{Stabilitätsabschätzung}}}\ \underbrace{\big\|\Phi(h) - E(h)\big\|}_{\substack{\le C\,h^{p+1}\\ \text{Lokale Fehlerabschätzung}}}\ \underbrace{\big\|E(jh)\,y_0\big\|}_{\substack{\le C\\ \text{Exakter Lösungwert}}}\\ &\le C\,h^p\,,\end{aligned}$$

vgl. Resultat zur Konvergenz von Einschrittverfahren (Satz 4.5).

Ähnlich wie bei Quadraturapproximationen beruht die Konstruktion von expliziten und impliziten Runge–Kutta Verfahren auf der Lösung von Ordnungsbedingungen, die sich aus der Forderung

$$y_1 - y(h) = \mathscr{O}\big(h^{p+1}\big)$$

ergeben. Die Herleitung der Ordnungsbedingungen für autonome Differentialgleichungen mit hinreichend oft differenzierbarer Funktion f beruht auf Taylorreihenentwicklungen der exakten Lösung (Differentialgleichung $y'(t) = f(y(t))$, mittels Kettenregel $y''(t) = f'(y(t))\,f(y(t))$)

$$y(h) = y_0 + h\,\underbrace{y'(0)}_{=f(y_0)} + \tfrac{1}{2}\,h^2\ \underbrace{y''(0)}_{=f'(y_0)f(y_0)} + \ldots$$

unter der Verwendung graphentheoretischer Mittel wie *Bäume*) und analoger Entwicklungen der Runge–Kutta Lösung (komplizierter Teil). Beispielsweise führen die Gaußschen Quadraturformeln auf implizite Runge–Kutta Verfahren, die Gaußschen Verfahren.

4.3. Konvergenzresultat für Einschrittverfahren

- Fragestellung: Konvergenzresultat für Zeitintegrationsverfahren (vgl. Abschnitt 4.2)

  Stabilität von Zeitintegrationsverfahren: Ein Zeitintegrationsverfahren (inbesondere ein Einschrittverfahren) mit Verfahrensfunktion Φ heißt stabil, wenn die mehrfache Hintereinanderausführung von Φ durch eine Konstante beschränkt ist, wobei die Konstante unabhängig von der Größe und der Anzahl der Zeitschritte ist.

  Vorsicht! Bei Zeitintegrationsverfahren unterscheidet man die Begriffe Stabilität und u.a. A-Stabilität (vgl. Abschnitt 4.4), weiters zu unterscheiden von verschiedenen Stabilitätsbegriffen bei Differentialgleichungen oder der numerischen Stabilität eines Algorithmus.

  Resultat zur Konvergenz von Einschrittverfahren (Satz 4.5): Die das Anfangswertproblem definierende Funktion f sei hinreichend regulär

  $$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} y(t) = f\big(t, y(t)\big), & t \in (t_0, T), \\ y(t_0) \text{ gegeben}. \end{cases}$$

  Weiters sei das Einschrittverfahren mit Verfahrensfunktion Φ wohldefiniert und stabil (für hinreichend kleine Zeitschrittweiten $0 < h_i \le \overline{h}$, $0 \le i \le N-1$)

  $$\begin{aligned} &t_0 < t_1 < \cdots < t_N = T, \qquad h_i = t_{i+1} - t_i, \quad 0 \le i \le N-1, \\ &y_n = \Phi(h_{n-1}, t_{n-1}, y_{n-1}), \quad 1 \le n \le N, \qquad y_0 \text{ gegeben}, \end{aligned}$$

  und es besitze die Konsistenzordnung $p \ge 1$. Dann gilt die globale Fehlerabschätzung

  $$\big\| y_N - y(t_N) \big\| \le C \big(\| y_0 - y(t_0) \| + h_{\max}^p \big), \qquad h_{\max} = \max_{0 \le i \le N-1} h_i,$$

  mit Konstane $C > 0$ unabhängig von der Größe und der Anzahl der Zeitschritte.

  Bemerkung: Die Aussage des Konvergenzresultates gilt im Wesentlichen auch für Lineare Mehrschrittverfahren. Für ein stabiles k-Schrittverfahren mit Startwerten $y_0, \ldots, y_{k-1}$ ergibt sich die globale Fehlerabschätzung

  $$\big\| y_N - y(t_N) \big\| \le C \big(\max_{0 \le i \le k-1} \| y_i - y(t_i) \| + h_{\max}^p \big).$$

  Da Mehrschrittverfahren durch Mehrschrittrekursionen definiert sind und diese im Allgemeinen instabil sind, sind Mehrschrittverfahren höherer Ordnung zu vermeiden.

  Illustrationen (Einfache stabile Zeitintegrationsverfahren, Numerische Berechnung der Konvergenzordnungen):

  - Explizites Eulerverfahren (Ordnung 1)
    Explizite Mittelpunktsregel (Ordnung 2)
    Explizites Runge–Kutta Verfahren der Ordnung 4
  - Implizites Eulerverfahren (Ordnung 1)
    Implizite Mittelpunktsregel (Ordnung 2)
    (Implizite) Trapezregel (Ordnung 2)
  - Mittelpunktsregel (explizites Zweischrittverfahren, Ordnung 2)

Bemerkungen (Beweis des Konvergenzresultates):

– Die Herleitung eines Konvergenzresultates für Einschrittverfahren angewendet auf nichtlineare Differentialgleichungen beruht auf dem zuvor für das explizite Eulerverfahren angegebenen Zugang. Zur Abschätzung des globalen Fehlers wird die folgende Relation verwendet (zur Vereinfachung Formulierung für lineare Differentialgleichungen und äquidistante Schrittweiten)

$$\underbrace{\big\|y_N - y(t_N)\big\|}_{\text{Globaler Fehler}} \le \sum_{j=0}^{N-1} \underbrace{\big\|\Phi(h)\big\|^{N-1-j}}_{\le C \text{ Stabilitätsabschätzung}} \underbrace{\big\|\Phi(h) - E(h)\big\|}_{\le C\,h^{p+1} \text{ Lokale Fehlerabschätzung}} \underbrace{\big\|E(jh)\,y_0\big\|}_{\le C \text{ Exakter Lösungwert}}$$
$$\le C\,h^p\,.$$

Wesentlich ist es somit, Stabilitätsabschätzungen und lokale Fehlerabschätzungen (Taylorreihenentwicklungen) abzuleiten.

Beispiele:

* Mittels der folgenden Taylorreihenentwicklung der exakten Lösung

$$\frac{\mathrm{d}}{\mathrm{d}t}\,y(t) = f\big(y(t)\big)\,,$$
$$y(h) = y(0) + h\,\underbrace{y'(0)}_{=f(y_0)} + \tfrac{1}{2}\,h^2\,\underbrace{y''(0)}_{=f'(y_0)\,f(y_0)} + \mathscr{O}\big(h^3\big)\,,$$

ergibt sich für das explizite Eulerverfahren die lokale Fehlerentwicklung (Konsistenzordnung $p = 1$)

$$y_1 = y_0 + h\,f(y_0)\,, \qquad y(h) = y_0 + h\,f(y_0) + \mathscr{O}\big(h^2\big)\,,$$
$$y_1 - y(h) = \mathscr{O}\big(h^2\big)\,,$$

und für die explizite Mittelpunktsregel die lokale Fehlerentwicklung (Konsistenzordnung $p = 2$)

$$y_1 = y_0 + h\,\underbrace{f\big(y_0 + \tfrac{h}{2}\,f(y_0)\big)}_{=f(y_0)+\frac{1}{2}\,h\,f'(y_0)\,f(y_0)+\mathscr{O}\left(h^2\right)} = y_0 + h\,f(y_0) + \tfrac{1}{2}\,h^2\,f'(y_0)\,f(y_0) + \mathscr{O}\big(h^3\big)\,,$$
$$y(h) = y_0 + h\,f(y_0) + \tfrac{1}{2}\,h^2\,f'(y_0)\,f(y_0) + \mathscr{O}\big(h^3\big)\,,$$
$$y_1 - y(h) = \mathscr{O}\big(h^3\big)\,.$$

* Diffiziler ist die Vorgehensweise beim impliziten Eulerverfahren

$$y_1 = y_0 + h\,f(y_1)\,,$$

da hier zuerst der Nachweis der Existenz der numerischen Lösung erbracht werden muß. Dazu verwendet man beispielsweise den Banachsche Fixpunktsatz (vgl. Numerische Mathematik I). Für Lipschitz-stetige definierende Funk-

tionen f mit Lipschitz-Konstante $L > 0$ ist für hinreichend kleine Zeitschrittweiten $0 < h \leq \overline{h}$ die Kontraktivität der Abbildung F sichergestellt

$$y_1 = F(y_1) = y_0 + h f(y_1),$$
$$\|F(z) - F(\widetilde{z})\| = \|h\big(f(z) - f(\widetilde{z})\big)\| \leq \underbrace{L\overline{h}}_{=\kappa \overset{!}{<} 1} \|z - \widetilde{z}\|.$$

Damit folgt die Existenz und Eindeutigkeit des Fixpunktes y_1 sowie die Beschränktheit der numerischen Lösung

$$\begin{aligned}
\|y_1\| &\leq \|y_0\| + h\|f(y_1)\| \\
&\leq \|y_0\| + h\|f(y_0)\| + h\|f(y_1) - f(y_0)\| \\
&\leq \|y_0\| + h\|f(y_0)\| + hL\|y_1 - y_0\| \\
&\leq (1 + hL)\|y_0\| + h\|f(y_0)\| + hL\|y_1\| \\
\Longrightarrow \quad \|y_1\| &\leq C = \tfrac{1}{1-\kappa}\big((1 + hL)\|y_0\| + h\|f(y_0)\|\big),
\end{aligned}$$

Nun ist ein Taylorreihenentwicklung der numerischen Lösung sinnvoll und führt auf die lokale Fehlerabschätzung (Konsistenzordnung $p = 1$)

$$\begin{aligned}
y_1 = y_0 + h f(y_1) = y_0 + h f\big(y_0 + h f(y_1)\big) &= y_0 + h f(y_0) + \mathcal{O}\big(h^2\big), \\
y(h) &= y_0 + h f(y_0) + \mathcal{O}\big(h^2\big), \\
y_1 - y(h) &= \mathcal{O}\big(h^2\big),
\end{aligned}$$

– Ein alternativer Zugang verwendet (vgl. Skriptum S. 76)

$$\underbrace{\big\|y_N - y(t_N)\big\|}_{\text{Globaler Fehler}} \leq \sum_{j=0}^{N-1} \underbrace{\big\|E(t_{N-1-j})\big\|}_{\leq C \text{ Abschätzung mittels Satz 4.2}} \underbrace{\big\|\Phi(h) - E(h)\big\|}_{\leq C h^{p+1} \text{ Lokale Fehlerabschätzung}} \underbrace{\big\|\Phi(h)^j y_0\big\|}_{\leq C \text{ Numerische Lösung}}$$
$$\leq C h^p.$$

Stabilitätsabschätzungen gehen hier bei der Abschätzung der numerischen Lösung ein.

– Zur Konvergenanalyse von Linearen Mehrschrittverfahren nützt man die Formulierung als Einschrittverfahren (vgl. Numerische Mathematik I, Abschnitt 7.1).

– Bereits bei einfachen Anfangswertproblemen kann es zu Ordnungsreduktionen kommen, wenn beispielsweise die definierende Funktion f und damit die exakte Lösung y nicht hinreichend oft differenzierbar ist.

Illustration (Ordnungsreduktion): Für das triviale Testbeispiel (direkter Zusammenhang mit bestimmten Integralen und Quadraturapproximationen)

$$\tfrac{\mathrm{d}}{\mathrm{d}t} y(t) = f(t) = \tfrac{3}{2}\sqrt{t}, \quad 0 < t < T, \qquad y(T) = y(0) + \int_0^T f(t)\,\mathrm{d}t = y(0) + T^{\frac{3}{2}},$$

beobachtet man aufgrund der Singularität bei $t = 0$

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\, y(t) = \frac{\mathrm{d}}{\mathrm{d}t}\, f(t) = \frac{3}{4}\,\frac{1}{\sqrt{t}}\,, \qquad 0 < t < T\,,$$

für Ein- und Mehrschrittverfahren der Konsistenzordnung $p \geq 2$ eine Reduktion auf die Konvergenzordnung $\frac{3}{2}$.

– Bei Miteinbeziehung des Einflusses von Rundungsfehlern (zusätzliche Inhomogenität in der Differentialgleichung) ergibt sich insgesamt die Fehlerabschätzung (wegen $Nh = T$ folgt $N = \frac{C}{h}$ für äquidistante Zeitschrittweiten)

$$\text{Gesamtfehler} \leq \underbrace{C\,h^p}_{\text{Verfahrensfehler}} + \underbrace{C\,\varepsilon_{\text{mach}}\,\frac{1}{h}}_{\text{Rundungsfehler}}\,.$$

Bei zu groß gewählten Zeitschrittweiten dominiert der Verfahrensfehler, bei zu klein gewählten Schrittweiten beeinträchtigt der Fehler aufgrund der Akkumulation von Rundungsfehlern die erreichbare Genauigkeit. Für die optimale Wahl (Vernachlässigung der Konstanten, $\varepsilon_{\text{mach}} \approx 10^{-16}$)

$$\begin{gathered}
g(h) = h^p + \varepsilon_{\text{mach}}\,\frac{1}{h} \quad \longrightarrow \quad \min\,, \\
0 = g'(h) = p\,h^{p-1} - \varepsilon_{\text{mach}}\,\frac{1}{h^2} = \frac{1}{h^2}\left(p\,h^{p+1} - \varepsilon_{\text{mach}}\right) \quad \Longrightarrow \quad h = \sqrt[p+1]{\frac{\varepsilon_{\text{mach}}}{p}}\,, \\
g(h) \approx \varepsilon_{\text{mach}}^{\frac{p}{p+1}}\,, \\
p = 1: \qquad g(h) \approx \sqrt{\varepsilon_{\text{mach}}} \approx 10^{-8}\,, \\
p = 2: \qquad g(h) \approx \varepsilon_{\text{mach}}^{\frac{2}{3}} \approx 10^{-11}\,, \\
p = 4: \qquad g(h) \approx \varepsilon_{\text{mach}}^{\frac{4}{5}} \approx 10^{-12}\,, \\
p = 10: \qquad g(h) \approx \varepsilon_{\text{mach}}^{\frac{10}{11}} \approx 10^{-15}\,,
\end{gathered}$$

zeigt sich der Vorteil bei der Anwendung eines Zeitintegrationsverfahrens höherer Ordnung.

- Richardson Extrapolation (Explizites Eulerverfahren): Im Zusammenhang mit Zeitintegrationsverfahren beruhen Extrapolationsverfahren auf der Idee, Linearkombinationen von numerische Approximationen zu verschiedenen Zeitschrittverfahren zu berechnen, die eine höhere Ordnung besitzen. Bestimmt man beispielsweise mittels explizitem Eulerverfahren numerische Approximationen zu den Schrittweiten h und $\frac{1}{2}\,h$ (Ordnung $p = 1$, einzelner Zeitschritt)

$$y_1 = y_0 + h\,f(t_0, y_0)\,, \qquad z_{1/2} = y_0 + \tfrac{1}{2}\,h\,f(t_0, y_0)\,, \quad z_1 = z_{1/2} + \tfrac{1}{2}\,h\,f\left(t_0 + \tfrac{1}{2}\,h, z_{1/2}\right),$$

so zeigt eine einfache Rechnung, daß die Linearkombination

$$2\,z_1 - y_1$$

Ordnung $p+1=2$ besitzt. Der allgemeine Zugang beruht ähnlich wie bei Quadraturapproximationen auf asymptotischen Entwicklungen der numerischen Lösung.

Vgl. Illustration (Extrapolation).

- Entsprechend adaptiven Quadraturformeln verwendet man u.a. Paare von Runge–Kutta Verfahren verschiedener Ordnung zur vorteilhaften Wahl der Zeitschrittweiten (Verfahren zur Integration, Verfahren zur Schätzung des lokalen Fehlers). Optimale Verfahren in Hinblick auf die Anzahl der benötigten Funktionsauswertungen sind eingebettete Runge–Kutta Verfahren der Ordnungen $p \neq \widehat{p}$

$$\begin{array}{c|c} c & A \\ \hline & b \end{array} \qquad \begin{array}{c|c} c & A \\ \hline & \widehat{b} \end{array}$$

mit gleichen Knoten und internen Stufen, jedoch unterschiedlich gewählten Gewichten.

Beispiel: DOPRI (Dormand – Prince Verfahren, 1980), explizites Runge–Kutta Verfahren der Ordnung 4(5), Standard-Löser ODE45

- Adaptive Zeitintegrationsverfahren: Ähnlich dem Prinzip adaptiver Verfahren zur numerischer Berechnung bestimmter Integrale sollen bei adaptiven Zeitintegrationsverfahren die Zeitschrittweiten dem Lösungsverlauf optimal angepaßt werden. Die Idee ist es, bei Anwendung eines Verfahrens der Konsistenzordnung p die Schrittweite so zu modifizieren, daß eine vorgegebene Toleranz erreicht wird

$$\mathrm{ERR}_{\mathrm{lokal}} = \mathrm{err}(h) \approx C\,h^{p+1}, \qquad \mathrm{err}(h_{\mathrm{optimal}}) \approx C\,h_{\mathrm{optimal}}^{p+1} \approx \mathrm{TOL}.$$

Die Division beider Relationen führt auf folgende Faustregel für die optimale Zeitschrittweite

$$\left(\frac{h_{\mathrm{optimal}}}{h}\right)^{p+1} \approx \frac{\mathrm{TOL}}{\mathrm{ERR}_{\mathrm{lokal}}} \quad \Longleftrightarrow \quad h_{\mathrm{optimal}} \approx h\,\sqrt[p+1]{\frac{\mathrm{TOL}}{\mathrm{ERR}_{\mathrm{lokal}}}}.$$

Dabei bezeichnet $\mathrm{ERR}_{\mathrm{lokal}}$ den mittels eines Verfahrens höherer Ordnung (siehe auch Eingebettete Verfahren, Extrapolation) geschätzten lokalen Fehler.

Vgl. Illustration (Schrödingergleichung, Adaptive Verfahren).

4.4. A-Stabilität

- Situation: Betrachtet wird eine autonome Differentialgleichung mit hinreichend regulärer definierender Funktion $f : \mathbb{R}^d \to \mathbb{R}^d$ (unendliches Zeitintervall)

$$\frac{\mathrm{d}}{\mathrm{d}t} y(t) = f\big(y(t)\big), \qquad t > t_0 .$$

  Asymptotisch stabile stationäre Lösungen: Eine Lösung $z : I = [t_0, \infty) \to \mathbb{R}^d$ der Differentialgleichung mit Anfangswert $z_0 = z(t_0)$ heißt asymptotisch stabil, wenn es eine Konstante $\delta > 0$ gibt, sodaß für jede andere Lösung $\widetilde{z} : I \to \mathbb{R}^d$ der Differentialgleichung mit Anfangswert $\widetilde{z}_0 = \widetilde{z}(t_0)$ folgende Eigeschaft gilt

$$\|z_0 - \widetilde{z}_0\| < \delta \quad \Longrightarrow \quad \lim_{t\to\infty} \|z(t) - \widetilde{z}(t)\| = 0 .$$

  Besondere Bedeutung haben asymptotisch stabile stationäre Lösungen (bzw. Gleichgewichtslösungen), d.h. zeitunabhängige Lösungen der Differentialgleichung

$$\frac{\mathrm{d}}{\mathrm{d}t} z(t) = 0 = f\big(z(t)\big), \qquad t > t_0 .$$

  Beispiel: Die skalare lineare Differentialgleichung

$$\frac{\mathrm{d}}{\mathrm{d}t} y(t) = \lambda\, y(t), \quad t \geq 0, \qquad \lambda < 0,$$

  besitzt die asymptotisch stabile stationäre Lösung $z = 0$. Für beliebige Anfangswerte $\widetilde{z}_0 \in \mathbb{R}$ gilt nämlich

$$\widetilde{z}(t) = \mathrm{e}^{\lambda t}\widetilde{z}_0, \quad t \geq 0, \qquad \lim_{t\to\infty} \|\widetilde{z}(t)\| = \lim_{t\to\infty} \mathrm{e}^{\lambda t}\, \|\widetilde{z}_0\| = 0 .$$

- Fragestellung: Approximationsgüte eines Zeitintegrationsverfahrens bei der Berechnung asymptotisch stabiler Lösungen.

  Bemerkung: Aus dem zuvor angegebenen Resultat zum Konvergenzverhalten von Zeitintegrationsverfahren folgt keine Aussage über die Approximationsgüte bei der Berechnung asymptotisch stabiler Lösungen, da die Konvergenzabschätzung nur auf beschränkten Zeitintervallen gültig ist (Aussage für hinreichend kleine Zeitschrittweiten, aufgrund der auftretenden Konstante $C = C(T)$ ist die Abschätzung nur auf vergleichsweise kurzen Zeitintervallen sinnvoll).

- Vorbemerkung: Es sei z eine stationäre Lösung der Differentialgleichung

$$\frac{\mathrm{d}}{\mathrm{d}t} y(t) = f\big(y(t)\big), \qquad t > t_0 .$$

  Theoretische Resultate sichern, daß das asymptotische Verhalten der stationären Lösung durch die linearisierte Differentialgleichung bestimmt ist

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} y(t) &= f\big(y(t)\big) = f(z) + f'(z)\big(y(t) - z\big) + \mathcal{O}\big(\|y(t) - z\|^2\big), \\ \frac{\mathrm{d}}{\mathrm{d}t} \widetilde{y}(t) &= A\,\widetilde{y}(t), \quad \widetilde{y}(t) = \mathrm{e}^{(t-t_0)A}\widetilde{y}(t_0), \quad t \geq t_0, \qquad A = f'(z). \end{aligned}$$

Falls alle Eigenwerte der Matrix A negativen Realteil haben, ist die Lösung asymptotisch stabil, falls jedoch ein Eigenwert mit positivem Realteil existiert, ist die Lösung asymptotisch instabil. Dies rechtfertigt die Betrachtung der Testgleichung von Dahlquist.

Im Folgenden wird wieder die Bezeichnung $y : I \to \mathbb{R}^d$ für die betrachtete Lösung verwendet, und z bezeichnet eine stationäre Lösung der Differentialgleichung.

Bemerkung: Für die folgenden Überlegungen ist es zweckmäßig, die exakte und numerische Lösung mittels des exakten und numerischen Lösungsoperators anzugeben (für die Testgleichung wird die Abhängigkeit vom Zeitpunkt bzw. von der Zeitschrittweite und dem Parameter λ angegeben)

$$y_1 = \Phi(h\lambda)\, y_0 \quad \approx \quad y(h) = E(h\lambda)\, y(0)\,.$$

Testgleichung (Dahlquist): Betrachte die skalare lineare Testgleichung mit bekannter exakter Lösung (vgl. Abschnitt 4.2, Annahme $y_0 \neq 0$)

$$\frac{\mathrm{d}}{\mathrm{d}t}\, y(t) = \lambda\, y(t)\,, \quad t \geq 0\,, \quad y(0) = y_0\,, \qquad \lambda < 0\,, \qquad (\text{bzw. } \lambda \in \mathbb{C} \text{ mit } \Re\lambda < 0)$$
$$y(t) = E(t\lambda)\, y_0 = \mathrm{e}^{t\lambda} y_0\,, \quad t \geq 0\,, \qquad \lim_{t\to\infty} E(\lambda t) = 0\,.$$

A-Stabilität (vgl. Definition 4.8): Ähnlich dem Verhalten der exakten Lösung der Testgleichung fordert man von einem A-stabilen numerischen Verfahren mit Verfahrensfunktion Φ (angewendet mit äquidistanten Zeitschritten $h > 0$), daß in der obigen Situation die Approximationswerte gegen die asymptotisch stabile Lösung konvergieren

$$y_n = \Phi(h\lambda)\, y_{n-1} = \Phi(h\lambda)^n\, y_0\,, \quad n \geq 0\,, \qquad \lim_{n\to\infty} \Phi(h\lambda)^n = 0\,.$$

Dies ist gleichbedeutend mit der Forderung

$$|\Phi(h\lambda)| < 1$$

bzw. auch damit, daß die linke Halbebene im (absoluten) Stabilitätsbereich des Verfahrens enthalten ist

$$\mathbb{C}_{<0} = \{\mu \in \mathbb{C} : \Re\mu < 0\} \subset S = \{\mu \in \mathbb{C} : |\Phi(\mu)| < 1\}\,.$$

Bemerkungen:

- Abhängig von der betrachteten Problemklasse ist auch eine Abschwächung des Begriffes der A-Stabilität ausreichend.
- Bei Anwendung eines Zeitintegrationsverfahrens auf die Testgleichung ist es naheliegend, die Bezeichnungen $\Phi(h\lambda) \approx E(h\lambda)$ zu verwenden, welche die wesentlichen Größen beinhalten.

Beispiele (Explizites Eulerverfahren, Implizites Eulerverfahren):

- Das explizite Eulerverfahren angewendet auf die Testgleichung ist durch

$$\begin{aligned} \Phi(h\lambda) = 1 + h\,\lambda \quad &\approx \quad E(h\lambda) = \mathrm{e}^{h\lambda}, \\ y_n = \Phi(h\lambda)^n y_0 = (1 + h\,\lambda)^n y_0 \quad &\approx \quad y(nh) = E(nh\lambda)\, y_0 = \mathrm{e}^{nh\lambda} y_0, \end{aligned}$$

  gegeben. Somit folgt

$$\lim_{n\to\infty} y_n = 0 \quad \Longleftrightarrow \quad |\Phi(h\lambda)| = |1 + h\lambda| < 1$$

  und weiters ($\lambda < 0$, Fall $|1 - h|\lambda|| = 1 - h|\lambda| < 1$ für $1 - h|\lambda| > 0$ ist abgedeckt)

$$|1 + h\lambda| = \big|h\,|\lambda| - 1\big| = h\,|\lambda| - 1 < 1 \quad \Longleftrightarrow \quad h\,|\lambda| < 2\,.$$

  Z.B. für $\lambda = -10^j$ führt dies auf die Schrittweiteneinschränkung $h < 2 \cdot 10^{-j}$. Bei zu groß gewählten Zeitschrittweiten oszilliert die numerische Lösung und wächst betragsmäßig stark an, folglich ist das numerische Ergebnis wertlos.

  Der Stabilitätsbereich

$$S = \{\mu \in \mathbb{C} : |1 + \mu| < 1\}$$

  beschreibt das Innere eines Kreises mit Mittelpunkt -1 und Radius 1, d.h. das explizite Eulerverfahren ist *nicht* A-stabil.

- Das implizite Eulerverfahren angewendet auf die Testgleichung ist durch

$$\Phi(h\lambda) = \tfrac{1}{1-h\lambda} \quad \approx \quad E(h\lambda) = \mathrm{e}^{h\lambda}$$

  gegeben. Die Bedingung

$$|\Phi(h\lambda)| = \tfrac{1}{|1-h\lambda|} < 1 \quad \Longleftrightarrow \quad 1 < |1 - h\lambda| = 1 + h\,|\lambda|$$

  ist für alle Schrittweiten $h > 0$ erfüllt, und insbesondere ist das implizite Eulerverfahren A-stabil.

Fazit: Bei der numerischen Lösung der Testgleichung mittels explizitem Eulerverfahren sind aus Stabilitätsgründen (für $\lambda \gg 1$ extrem starke) Schrittweiteneinschränkungen erforderlich. Die numerischen Lösung der Testgleichung mittels implizitem Eulerverfahren bleibt hingegen für beliebige Zeitschritte stabil und führt bereits bei vergleichweise großen Schrittweiten auf ein zufriedenstellendes Ergebnis.

Allgemein gilt, daß nur implizite Zeitintegrationsverfahren die Eigenschaft der A-Stabilität besitzen können.

- Implizite Runge–Kutta Verfahren wie Radau IIA Verfahren sind A-stabil.
- Das implizite lineare Mehrschrittverfahren BDF 2 ist A-stabil. Für $3 \le k \le 6$ ist das Verfahren BDF k zwar nicht A-stabil, jedoch A(ϑ)-stabil (Sektor anstelle Halbebene im Stabilitätsbereich enthalten).

- Stabilitätseigenschaften numerischer Verfahren wie A-Stabilität sind im Zusammenhang mit steifen Differentialgleichungen (insbesondere partiellen Differentialgleichungen wie Diffusionsgleichungen) wesentlich.

  > Curtiss & Hirschfelder (1952): *... stiff equations are equations where certain implicit methods ... perform better, usually tremendously better, than explicit ones.*

  Illustration (Zeitintegration der Diffusionsgleichung mittels explizitem und implizitem Eulerverfahren)

5. Randwertprobleme für gewöhnliche Differentialgleichungen

- Inhalte:
  - Problemstellung
  - Schießverfahren (Lösung von Anfangswertproblemen, Newtonverfahren)
    Differenzenverfahren
    Kollokationsverfahren

  Bemerkung: Vertauschen der Abschnitte 5.3 und 5.4

- Ausblick: Ortsdiskretisierungsverfahren für partielle Differentialgleichungen
  - Finite Differenzen Methode
  - Finite Elemente Methode (Galerkin Verfahren)

5.1. Problemstellung

- Erinnerung:
  - Bei einem Anfangswertproblem sind das Zeitintervall $I = [t_0, T]$, die (hinreichend reguläre) definierende Funktion $f : I \times \mathbb{R}^d \to \mathbb{R}^d$ und der Anfangswert $y_0 \in \mathbb{R}^d$ vorgegeben. Gesucht ist eine (hinreichend reguläre) Funktion $y : I \to \mathbb{R}^d$, die die Differentialgleichung sowie die Anfangsbedingung erfüllt

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} y(t) = f\big(t, y(t)\big), & t \in (t_0, T), \\ y(t_0) = y_0. \end{cases}$$

  - Speziell bei der Schwingungsgleichung sind neben der Differentialgleichung

$$m \tfrac{\mathrm{d}^2}{\mathrm{d}t^2} y(t) + r \tfrac{\mathrm{d}}{\mathrm{d}t} y(t) + k\, y(t) = 0, \qquad t \in (t_0, T),$$

    die Anfangsauslenkung und Anfangsgeschwindigkeit vorgegeben

$$y(t_0), \qquad \tfrac{\mathrm{d}}{\mathrm{d}t} y(t)\big|_{t=t_0}.$$

    Bei der Formulierung der Differentialgleichung zweiter Ordung als Differentialgleichungssystem erster Ordnung, entsprechen die erste bzw. zweite Komponente des Vektors $Y(t) = (y(t), \frac{\mathrm{d}}{\mathrm{d}t} y(t))^T$ der Auslenkung bzw. Geschwindigkeit zum aktuellen Zeitpunkt

$$\tfrac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} Y_1(t) \\ Y_2(t) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{r}{m} \end{pmatrix} \begin{pmatrix} Y_1(t) \\ Y_2(t) \end{pmatrix}, \quad t \in (t_0, T), \qquad Y(t_0) \text{ gegeben}.$$

- Vorbemerkung: Bei der Schwingungsgleichung kann man anstelle der Anfangsauslenkung und Anfangsgeschwindigkeit beispielsweise auch die Auslenkung zum Anfangs- und Endzeitpunkt vorgegeben, was auf ein Randwertproblem führt. Allgemeiner ist ein Randwertproblem durch eine gewöhnliche Differentialgleichung und eine algebraische Bedingung für die Randwerte $y(t_0)$ und $y(T)$ gegeben.

  Randwertprobleme für gewöhnliche Differentialgleichungen erster Ordnung: Für vorgegebene Punkte $t_0 \in \mathbb{R}$ und $T \in \mathbb{R}$ sei $I = [t_0, T]$ falls $t_0 < T$ (bzw. $I = [T, t_0]$ falls $t_0 > T$). Weiters seien $f : I \times \mathbb{R}^d \to \mathbb{R}^d : (t, v) \mapsto f(t, v)$ und $r : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d : (v, w) \mapsto r(t, v)$ vorgegebene (reguläre) Funktionen. Eine Lösung des Randwertproblems

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} y(t) = f\big(t, y(t)\big), & t \in (t_0, T), \qquad (\text{bzw. } (T, t_0)) \\ r\big(y(t_0), y(T)\big) = 0, \end{cases}$$

  ist eine Funktion $y : I \to \mathbb{R}^d$, welche die (gewöhnliche) Differentialgleichung (erster Ordnung) und die Randbedingung erfüllt. Dabei wird vorausgesetzt, daß y auf I stetig und zumindest in (t_0, T) (bzw. (T, t_0)) differenzierbar ist.

– Falls die Funktion r durch Matrizen $A, B \in \mathbb{R}^{d\times d}$ und eine Spalte $c \in \mathbb{R}^d$ definiert ist, d.h. es ist $r(v,w) = Av + Bw - c$, spricht man von einer linearen Randbedingung

$$Ay(t_0) + By(T) = c\,.$$

– Falls die Funktion r von der Form $r(v,w) = (\varrho(v), \widetilde{\varrho}(w))$ mit Funktionen $\varrho : \mathbb{R}^d \to \mathbb{R}^k$ und $\widetilde{\varrho} : \mathbb{R}^d \to \mathbb{R}^{d-k}$ für $1 \le k \le d-1$ ist, spricht man von einer separierten Randbedingung. Insbesondere ist dies für Randbedingungen der einfachen Form (nach eventueller Umordnung der Komponenten von y)

$$y_1(t_0) = c_1\,,\ \ldots,\ y_k(t_0) = c_k\,, \qquad y_{k+1}(T) = c_{k+1}\,,\ \ldots,\ y_d(T) = c_d\,,$$

der Fall.

Bemerkung: Im Gegensatz zu Anfangswertproblemen ist es bei (nichtlinearen) Randwertproblemen (noch) schwierig(er), allgemeine Aussagen zur Existenz und Eindeutigkeit zu machen.

– Dies zeigt sich beispielsweise an der einfachen linearen Differentialgleichung (Schwingungsgleichung mit $m = k = 1$, $r = 0$)

$$\tfrac{\mathrm{d}^2}{\mathrm{d}t^2}\, y(t) + y(t) = 0\,, \qquad t \in (0,\pi)\,,$$

mit exakter Lösung

$$y(t) = C_1 \sin t + C_2 \cos t\,, \qquad t \in [0,\pi]\,.$$

Wegen $y(0) = C_2$ und $y(\pi) = -C_2$ existieren bei Vorgabe der Randbedingungen

$$y(0) = 0\,, \quad y(\pi) = 0\,, \qquad y(t) = C_1 \sin t\,,$$

unendlich viele Lösungen, bei Vorgabe der Randbedingungen

$$y(0) = 1\,, \quad y(\pi) = 1\,,$$

existiert jedoch keine Lösung.

– Verallgemeinerung: Die Lösung $y : I \to \mathbb{R}$ einer skalaren linearen Differentialgleichung zweiter Ordnung ist von der folgenden Form mit homogenen Lösungen $y_{h,1}$ und $y_{h,2}$ (zur Differentialgleichung mit $\gamma = 0$) und einer partikulären Lösung y_p (spezielle Lösung).

$$\tfrac{\mathrm{d}^2}{\mathrm{d}t^2}\, y(t) + \alpha(t)\, \tfrac{\mathrm{d}}{\mathrm{d}t}\, y(t) + \beta(t)\, y(t) = \gamma(t)\,, \qquad t \in (t_0, T)\,,$$
$$y = C_1\, y_{h,1} + C_2\, y_{h,2} + y_p\,.$$

Bei der spezieller Wahl der Anfangsbedingungen (Satz von Picard–Lindelöf sichert die Existenz und Eindeutigkeit der Lösungen)

$$y_{h,1}(t_0) = 1\,, \quad \tfrac{\mathrm{d}}{\mathrm{d}t}\, y_{h,1}(t)\big|_{t=t_0} = 0\,, \qquad y_{h,2}(t_0) = 0\,, \quad \tfrac{\mathrm{d}}{\mathrm{d}t}\, y_{h,2}(t)\big|_{t=t_0} = 1\,,$$
$$y_p(t_0) = 0\,, \quad \tfrac{\mathrm{d}}{\mathrm{d}t}\, y_p(t)\big|_{t=t_0} = 0\,,$$

folgen bei Vorgabe der Randbedingungen

$$y(t_0) = y_0\,, \qquad y(T) = y_T\,,$$

die folgenden Relationen für die zu bestimmenden Konstanten C_1, C_2

$$y(t_0) = C_1 \underbrace{y_{h,1}(t_0)}_{=1} + C_2 \underbrace{y_{h,2}(t_0)}_{=0} + \underbrace{y_p(t_0)}_{=0} = C_1 \overset{!}{=} y_0 \qquad \Longrightarrow \quad C_1 = y_0\,,$$

$$y(T) = \underbrace{C_1}_{=y_0}\, y_{h,1}(T) + C_2\, y_{h,2}(T) + y_p(T) = y_0\, y_{h,1}(T) + C_2\, y_{h,2}(T) + y_p(T) = y_T$$

$$\Longrightarrow \quad C_2\, y_{h,2}(T) = y_T - y_p(T) - y_0\, y_{h,1}(T)\,.$$

Somit können drei unterschiedliche Fälle eintreten:

* $y_{h,2}(T) \neq 0$: Eindeutige Lösung mit $C_2 = \frac{1}{y_{h,2}(T)} \big(y_T - y_p(T) - y_0\, y_{h,1}(T)\big)$.
* $y_{h,2}(T) = 0$ und $y(T) = y_p(T) + y_0\, y_{h,1}(T) = y_T$: Unendlich viele Lösungen.
* $y_{h,2}(T) = 0$ und $y(T) = y_p(T) + y_0\, y_{h,1}(T) \neq y_T$: Keine Lösung.

- Freie Randwertprobleme: Beispielsweise beim *Re-Entry Problem* (Wiedereintritt einer Rakete in die Atmosphäre) tritt ein Randwertproblem auf, wo der Endzeitpunkt nicht festgelegt ist und stattdessen eine zusätzliche Randbedingung vorgegeben ist (d.h. es ist $r : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^{d+1}$)

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}\, y(t) = f\big(t, y(t)\big)\,, \quad t \in (t_0, T)\,, \qquad (\text{bzw. } (T, t_0)) \\ r\big(y(t_0), y(T), T\big) = 0\,, \end{cases}$$

Solche freien Randwertprobleme lassen sich mittels einer Transformation des Zeitintervalles auf das Einheitsintervall

$$\begin{aligned} [t_0, T] \quad &\longleftrightarrow \quad [0, 1] \\ t = t_0 + s\,(T - t_0) \quad &\longleftrightarrow \quad s = \tfrac{t - t_0}{T - t_0} \end{aligned}$$

auf die Standardform eines Randwertproblems für den Lösungsvektor $Y(s) = (\widetilde{Y}(s), T)^T$ mit $\widetilde{Y}(s) = y(t)$ reduzieren, nämlich (wegen $\frac{\mathrm{d}}{\mathrm{d}s}\, \widetilde{Y}(s) = \frac{\mathrm{d}}{\mathrm{d}t}\, y(t)\, \frac{\mathrm{d}t}{\mathrm{d}s} = f(t, y(t))\, T$ folgt $\frac{\mathrm{d}}{\mathrm{d}s}\, \widetilde{Y}(s) = T\, f(t_0 + s\,(T - t_0), \widetilde{Y}(s))$, weiters ist $Y(0) = (y(t_0), T)$ und $Y(1) = (y(T), T)$)

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}s}\, Y(s) = \begin{pmatrix} T\, f\big(t_0 + s\,(T - t_0), \widetilde{Y}(s)\big) \\ 0 \end{pmatrix}\,, \quad s \in (0, 1)\,, \\ r\big(Y(0), Y(1)\big) = 0\,. \end{cases}$$

Illustration: Wurfparabel als Anfangswertproblem, Randwertproblem und Freies Randwertproblem (zwei physikalisch sinnvolle Lösungen zu den Endzeitpunkten $T_1, T_2 > 0$)

5.2. Lösung durch Rückführung auf ein Anfangswertproblem (Schießverfahren)

- Einfaches Schießverfahren:
  - Vereinfachung: Zur Vereinfachung wird zunächst ein Randwertproblem für eine Funktion $y = (y_1, y_2)^T : I = [t_0, T] \to \mathbb{R}^2$ mit speziellen separierten Randbedingungen betrachtet
$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} y(t) = f\big(t, y(t)\big), \quad t \in (t_0, T), \\ y_1(t_0) = y_{01}, \quad y_1(T) = y_{T1}. \end{cases}$$
    Beispiele: Schwingungsgleichung, Wurfparabel
  - Idee: Die Idee des (einfachen) Schießverfahrens ist es, das Randwertproblem auf ein Anfangswertproblem zurückzuführen
$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} y(t) = f\big(t, y(t)\big), \quad t \in (t_0, T), \\ y(t_0) = y_0 = (y_{01}, y_{02})^T. \end{cases}$$
    Die unbekannte Komponente des Anfangswertes soll dabei so bestimmt werden, daß die zugehörige exakte Lösung (Angabe der Abhängigkeit des Lösungsoperators E vom aktuellen Zeitpunkt, Anfangszeit und Anfangswert)
$$y(t) = E(t, t_0, y_0)$$
    im Endzeitpunkt T die geforderte Randbedingung erfüllt (dabei bezeichnet $y_1(T) = (E(T, t_0, y_0))_1$ die erste Komponente der exakten Lösung bei $t = T$)
$$y_1(T) = \big(E(T, t_0, y_0)\big)_1 = y_{T1},$$
    d.h. es ist die nichtlineare Gleichung
$$F(y_{02}) = y_1(T) - y_{T1} = \big(E(T, t_0, y_0)\big)_1 - y_{T1} = 0$$
    zu lösen.
    Illustration (Verschiedene Anfangsgeschwindigkeiten und zugehörige Lösungen), vgl. Skriptum, S. 99.
  - Verallgemeinerung: Die Lösung eines Randwertproblems allgemeiner Form für eine Funktion $y : I \to \mathbb{R}^d$
$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} y(t) = f\big(t, y(t)\big), \quad t \in (t_0, T), \\ r\big(y(t_0), y(T)\big) = 0, \end{cases}$$
    beruht auf der Betrachtung des zugehörigen Anfangswertproblems
$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} y(t) = f\big(t, y(t)\big), \quad t \in (t_0, T), \\ y(t_0) = y_0, \end{cases}$$
    und der Lösung der nichtlinearen Gleichung (für die unbekannten Lösungskomponenten)
$$F(y_0) = r\big(y(t_0), y(T)\big) = r\big(y_0, E(T, t_0, y_0)\big) = 0.$$

– Näherungsweise Lösung:

  * Im Allgemeinen wird die nichtlineare Gleichung

$$F(y_0) = r\big(y(t_0), y(T)\big) = 0$$

    mittels eines iterativen Verfahrens näherungsweise gelöst, und eine naheliegende Wahl ist das Newtonverfahren oder Modifikationen davon (vgl. Numerische Mathematik I). Bei Verwendung des Newtonverfahrens wird (zumindest näherungsweise) die erste Ableitung der Funktion F benötigt (Kettenregel, partielle Ableitungen $\partial_1 r$ und $\partial_2 r$, Ableitung der exakten Lösung nach dem Anfangswert $\partial_{y_0} y$)

$$F'(y_0) = \partial_1 r\big(y_0, y(T)\big) + \partial_2 r\big(y_0, y(T)\big)\, \partial_{y_0} y(T)\,.$$

    Zur Bestimmung der Ableitung der exakten Lösung nach dem Anfangswert verwendet man die Variationsgleichung (zeitabhängige Matrix $Y = \partial_{y_0} y$ erfüllt Differentialgleichung $\frac{\mathrm{d}}{\mathrm{d}t}\partial_{y_0} y(t) = \partial_2 f\big(t, y(t)\big)\partial_{y_0} y(t)$ und Anfangsbedingung $\partial_{y_0} y(t_0) = I$, $\partial_2 f(t,v)$ bezeichnet die partielle Ableitung von f bezüglich des zweiten Argumentes v)

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}\, y(t) = f\big(t, y(t)\big)\,, & t \in (t_0, T)\,, \qquad y(t_0) = y_0\,, \\ \frac{\mathrm{d}}{\mathrm{d}t}\, Y(t) = \partial_2 f\big(t, y(t)\big)\, Y(t)\,, & t \in (t_0, T)\,, \qquad Y(t_0) = I\,, \end{cases}$$

    vgl. Dynamische Systeme und insbesondere Stabilität von Differentialgleichungen.

  * Zur näherungsweisen Lösung eines Randwertproblemes wird im Allgemeinen die folgende Vorgehensweise gewählt:

    (i) Wahl eines geeigneten Startwertes y_0.

    (ii) Anwendung eines Zeitintegrationsverfahren zur numerischen Lösung des zugehörigen Anfangswertproblems, und insbesondere Berechnung einer Approximation an den Funktionswert $y_N \approx y(T) = E(T, t_0, y_0)$ (vgl. Abschnitt 4). Berechnung von $\widetilde{F}(y_0) = r\big(y_0, y_N\big) \approx F(y_0)$.

    (iii) Anwendung eines Zeitintegrationsverfahren zur numerischen Lösung der zugehörigen Variationsgleichung, und insbesondere Berechnung einer Approximation an den Funktionswert $Y_N \approx \partial_{y_0} y(T)$. Berechnung von $G(y_0) = \partial_1 r(y_0, y_N) + \partial_2 r(y_0, y_N)\, Y_N \approx F'(y_0)$.

    (iv) In einem Newtonschritt, ersetze y_0 durch $y_0 - G(y_0)^{-1}\widetilde{F}(y_0)$ und iteriere.

  * In Situationen, wo die obige Vorgehensweise versagt, weil das Newtonverfahren schlechte Konvergenzeigenschaften besitzt (Startwert ungeeignet, Differentialgleichung mit exponentiell anwachsenden Lösungen, großes Integrationsintervall) wird anstelle des einfachen Schießverfahrens das mehrfache Schießverfahren (Einfügen zusätzlicher Stützstellen) verwendet, vgl. Skriptum, S. 100.

– Speziell für ein Randwertproblem, das von einem Randwertproblem für eine skalare lineare Differentialgleichung zweiter Ordnung mit speziellen separierten Randbedingungen herstammt (Lösung $z : I \to \mathbb{R}$)

$$\begin{cases} \frac{\mathrm{d}^2}{\mathrm{d}t^2} z(t) + \alpha(t) \frac{\mathrm{d}}{\mathrm{d}t} z(t) + \beta(t) z(t) = \gamma(t), & t \in (t_0, T), \\ z(t_0) = z_0, \quad z(T) = z_T, \end{cases}$$

kann man die gesuchte Komponente des Anfangswertes bestimmen (sofern spezielle homogene und eine spezielle partikuläre Lösung bekannt sind). In diesem Fall ist das zugehörige Anfangswertproblem (setze $y = (y_1, y_2)^T = (z, \frac{\mathrm{d}}{\mathrm{d}t} z)^T : I \to \mathbb{R}^2$)

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \begin{pmatrix} y_2(t) \\ -\beta(t)\, y_1(t) - \alpha(t)\, y_2(t) + \gamma(t) \end{pmatrix}, & t \in (t_0, T), \\ \begin{pmatrix} y_1(t_0) \\ y_2(t_0) \end{pmatrix} = \begin{pmatrix} y_{01} \\ y_{02} \end{pmatrix}. \end{cases}$$

Die unbekannte Komponente y_{02} des Anfangswertes soll dabei so bestimmt werden, daß die zugehörige exakte Lösung im Endzeitpunkt die geforderte Randbedingung erfüllt

$$y_1(T) = z(T) = z_T.$$

Falls homogene Lösungen $z_{h,1}$, $z_{h,2}$ und eine partikuläre Lösung z_p der Differentialgleichung zweiter Ordnung bekannt sind (zu den in Abschnitt 5.1 angegebenen Anfangsbedingungen), folgt die Lösungsdarstellung

$$z = C_1 z_{h,1} + C_2 z_{h,2} + z_p.$$

Einsetzen der bekannten Funktionswerte $y_1(t_0) = z(t_0) = z_0$, $y_1(T) = z(T) = z_T$ bzw. der geforderten Anfangsbedingung $y_2(t_0) = \frac{\mathrm{d}}{\mathrm{d}t} z(t)|_{t=t_0} = y_{02}$ führt auf

$$\begin{aligned} z &= C_1 z_{h,1} + C_2 z_{h,2} + z_p, \\ \tfrac{\mathrm{d}}{\mathrm{d}t} z &= C_1 \tfrac{\mathrm{d}}{\mathrm{d}t} z_{h,1} + C_2 \tfrac{\mathrm{d}}{\mathrm{d}t} z_{h,2} + \tfrac{\mathrm{d}}{\mathrm{d}t} z_p, \\ z(t_0) = C_1 \underbrace{z_{h,1}(t_0)}_{=1} + C_2 \underbrace{z_{h,2}(t_0)}_{=0} + \underbrace{z_p(t_0)}_{=0} &= C_1 = z_0 \quad \Longrightarrow \quad C_1 = z_0, \\ \tfrac{\mathrm{d}}{\mathrm{d}t} z(t)\big|_{t=t_0} = C_1 \underbrace{\tfrac{\mathrm{d}}{\mathrm{d}t} z_{h,1}(t)\big|_{t=t_0}}_{=0} + C_2 \underbrace{\tfrac{\mathrm{d}}{\mathrm{d}t} z_{h,2}(t)\big|_{t=t_0}}_{=1} + \underbrace{\tfrac{\mathrm{d}}{\mathrm{d}t} z_p(t)\big|_{t=t_0}}_{=0} &= C_2 = y_{02} \\ \Longrightarrow \quad C_2 &= y_{02}, \\ z(T) = C_1 z_{h,1}(T) + C_2 z_{h,2}(T) + z_p(T) = z_0 z_{h,1}(T) + y_{02} z_{h,2}(T) + z_p(T) &= z_T \\ \Longrightarrow \quad y_{02} &= \tfrac{z_T - z_p(T) - z_0 z_{h,1}(T)}{z_{h,2}(T)}. \end{aligned}$$

Sofern die Bedingung $z_{h,2}(T) \neq 0$ für die Lösbarkeit des Randwertproblems erfüllt ist, ist die Lösung des Randwertproblems gerade die Lösung des zugehörigen Anfangswertproblems mit

$$y_{02} = \tfrac{z_T - z_p(T) - z_0 z_{h,1}(T)}{z_{h,2}(T)}.$$

– Illustration (Einfaches Schießverfahren für homogene lineare Differentialgleichung $y' = Ay$).

5.4. Kollokationsverfahren

- Situation: Betrachtet wird ein Randwertproblem der Form für eine Funktion $y : I \to \mathbb{R}^d$

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} y(t) = f\big(t, y(t)\big), & t \in (t_0, T), \\ r\big(y(t_0), y(T)\big) = 0. \end{cases}$$

  Bemerkung: Mit Kollokation bezeichnet man die Interpolation von Funktionswerten und Ableitungen.

  Idee: Bei einem Kollokationsverfahren für ein Randwertproblem betrachtet man einen Raum von Funktionen (z.B. Polynomfunktionen oder Splinefunktionen, zusätzliche Eigenschaften) und fordert, daß die Differentialgleichung in gewissen Stützstellen sowie die Randbedingungen erfüllt sind. Im Allgemeinen ist der Funktionenraum als lineare Hülle $\langle v_1, \dots, v_K \rangle$ von Basisfunktionen gegeben und man wählt den Ansatz

$$z = \sum_{i=1}^{K} c_i \, v_i \quad \approx \quad y,$$

  wobei die Koeffizienten $c_i \in \mathbb{R}$ so bestimmt werden, daß für vorgegebene Stützstellen $t_0 < t_1 < \cdots t_{N-1} < t_N = T$ die Bedingungen (Differentialgleichung in den inneren Punkten, Randbedingung)

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} z(t)\big|_{t=t_n} = f\big(t_n, z(t_n)\big), & 1 \le n \le N-1, \\ r\big(z(t_0), z(t_N)\big) = 0, \end{cases}$$

  erfüllt sind.

  Bemerkungen:

  - Oft ist es vorteilhaft, die Basisfunktionen so zu wählen, daß spezielle Randbedingungen wie etwa die homogene Randbedingungen $y(a) = 0 = y(b)$ automatisch erfüllt sind (Linearkombinationen der Basisfunktionen erfüllen dann ebenfalls die homogenen Randbedingungen).
  - Vgl. Konstruktion der BDF-Verfahren.

  Illustration (Einfache Differentialgleichung, Kollokation mittels Polynomfunktionen), vgl. Skriptum, S. 110.

5.3. Differenzenverfahren

- Vorbemerkung: Differenzenverfahren (Relaxationsverfahren) sind *globale* Verfahren (kein *time-stepping approach*) zur näherungsweisen Lösung von Randwertproblemen für gewöhnliche Differentialgleichungen und werden auch zur Ortsdiskretisierung von partiellen Differentialgleichungen verwendet. Aus diesem Grund wird die unabhängige Variable in diesem Abschnitt mit x bezeichnet.

- Situation: Betrachtet wird ein Randwertproblem, das von einem Randwertproblem für eine skalare lineare Differentialgleichung zweiter Ordnung mit speziellen separierten Randbedingungen herstammt (Lösung $y : [a,b] \to \mathbb{R}$, (zumindest) stetige Funktionen $\alpha, \beta, \gamma : [a,b] \to \mathbb{R}$, Voraussetzung $\beta \geq 0$ (punktweise), d.h. $\beta(x) \geq 0$ für $x \in [a,b]$)

$$\begin{cases} -\frac{\mathrm{d}^2}{\mathrm{d}x^2} y(x) + \alpha(x) \frac{\mathrm{d}}{\mathrm{d}x} y(x) + \beta(x)\, y(x) = \gamma(x)\,, & x \in (a,b)\,, \\ y(a) = y_a\,, \quad y(b) = y_b\,. \end{cases}$$

  Kurzschreibweise mittels Differentialoperator: Eine übliche Kurzschreibweise mittels eines (linearen) Differentialoperators (zweiter Ordnung) ist (Bemerkung zu Definitionsbereich, s.u.)

$$L : \mathscr{C}^2(a,b) \longrightarrow \mathscr{C}(a,b) : z \longmapsto Lz = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} z + \alpha \frac{\mathrm{d}}{\mathrm{d}x} z + \beta z\,,$$

  d.h. es ist $(Lz)(x) = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} z(x) + \alpha(x) \frac{\mathrm{d}}{\mathrm{d}x} z(x) + \beta(x)\, z(x)$ (oft auch kurz $Lz(x)$ statt $(Lz)(x)$). Die Differentialgleichung läßt sich dann in der kompakten Form (ähnlich einem linearen Gleichungssystem, s.u.)

$$L y = \gamma \qquad \text{bzw.} \qquad (Ly)(x) = \gamma(x)\,, \quad x \in (a,b)\,,$$

  angeben.

  Differenzenverfahren: Die Idee von Differenzenverfahren (bzw. Finiten Differenzenverfahren) ist es, in einer Differentialgleichung die auftretenden Differentialquotienten durch Differenzenquotienten zu ersetzen. Bei einer linearen Differentialgleichung führt dies auf ein lineares Gleichungssystem für die Näherungswerte an den vorgegebenen Stützwerten.

  Beispiele (Symmetrische Differenzen): Häufig verwendete Approximationen der ersten und zweiten Ableitung sind (Vorwärtsdifferenz, Rückwärtsdifferenz, Symmetrische Differenzen)

$$\frac{y(x+h)-y(x)}{h} = \frac{\mathrm{d}}{\mathrm{d}x} y(x) + \mathscr{O}(h)\,, \qquad \frac{y(x-h)-y(x)}{-h} = \frac{y(x)-y(x-h)}{h} = \frac{\mathrm{d}}{\mathrm{d}x} y(x) + \mathscr{O}(h)\,,$$
$$\frac{y(x+h)-y(x-h)}{2h} = \frac{\mathrm{d}}{\mathrm{d}x} y(x) + \mathscr{O}(h^2)\,,$$
$$\frac{y(x+h)-2\,y(x)+y(x-h)}{h^2} = \frac{\mathrm{d}^2}{\mathrm{d}x^2} y(x) + \mathscr{O}(h^2)\,.$$

Genauer: Falls die Funktion y hinreichend oft differenzierbar ist, führen Taylorreihenentwicklungen auf folgende Relationen für die symmetrischen Differenzen (jeweils mit $\xi \in [x-h, x+h]$)

$$y \in \mathscr{C}^3(a,b): \quad \frac{y(x+h)-y(x-h)}{2h} = \frac{\mathrm{d}}{\mathrm{d}x}\, y(x) + \frac{1}{6}\, h^2\, \frac{\mathrm{d}^3}{\mathrm{d}x^3}\, y(x)\Big|_{x=\xi},$$
$$y \in \mathscr{C}^4(a,b): \quad \frac{y(x+h)-2\,y(x)+y(x-h)}{h^2} = \frac{\mathrm{d}^2}{\mathrm{d}x^2}\, y(x) + \frac{1}{12}\, h^2\, \frac{\mathrm{d}^4}{\mathrm{d}x^4}\, y(x)\Big|_{x=\xi}.$$

Gitterfunktion (Grid function): Zur Vereinfachung werden im Folgenden äquidistante Gitterpunkte zur Gitterweite $h > 0$ betrachtet (M innere Punkte $x_1, \ldots, x_M$)

$$\overline{\Omega} = \{x_m = a + mh : 0 \le m \le M+1\}, \qquad h = \frac{b-a}{M+1}.$$

Als diskretes Analogon der exakten Lösung $y : [a,b] \to \mathbb{R}$ des Randwertproblems ist die Gitterfunktion auf dem Ortsgitter definiert

$$\widetilde{y} : \overline{\Omega} \to \mathbb{R} : x_m \to \widetilde{y}(x_m) = \widetilde{y}_m, \qquad \widetilde{y}_m \approx y(x_m), \quad 0 \le m \le M+1.$$

Die Randwerte $\widetilde{y}_0 = y(x_0) = y_a$ und $\widetilde{y}_{M+1} = y(x_{m+1}) = y_b$ sind vorgegeben (zur Vereinfachung exakte Randwerte), zu bestimmen sind die Werte der Gitterfunktion an den inneren Gitterpunkten

$$\Omega = \{x_m = a + mh : 1 \le m \le M\}.$$

Bemerkungen:

- Im vorliegenden eindimensionalen Fall ist die Gitterfunktion $\widetilde{y}$ durch die Funktionswerte an den (fixierten) inneren Gitterpunkten $x_1, \ldots, x_M$ bestimmt, und deshalb wird auch die Vektorschreibweise (Matrix in 2D, Tensorstruktur in 3D)

  $$\widetilde{y} = (\widetilde{y}_m)_{1 \le m \le M} = (\widetilde{y}_1, \ldots, \widetilde{y}_M)^T$$

  verwendet.

- Nach Einschränkung der exakten Lösung auf die Gitterpunkte (wie zuvor Vektorschreibweise mit $y_m = y(x_m)$ für $0 \le m \le M+1$)

  $$y\big|_\Omega = (y_1, \ldots, y_M)^T,$$

  ist es sinnvoll, die Differenz (Fehler des Differenzenverfahrens, kein Beitrag der Randwerte unter der Annahme $\widetilde{y}_0 = y(x_0) = y_a$ und $\widetilde{y}_{M+1} = y(x_{m+1}) = y_b$)

  $$\widetilde{y} - y\big|_\Omega = (\widetilde{y}_1 - y_1, \ldots, \widetilde{y}_M - y_M)^T$$

  zu betrachten.

Approximation mittels symmetrischer Differenzen: Mittels der oben angegebenen symmetrischen Differenzen ergibt sich als diskretes Analogon des Differentialoperators

$$L : z \longmapsto Lz = -\tfrac{\mathrm{d}^2}{\mathrm{d}x^2} z + \alpha \tfrac{\mathrm{d}}{\mathrm{d}x} z + \beta z$$

der Differenzenoperator

$$\widetilde{L} : \widetilde{z} = \big(\widetilde{z}_m\big)_{1\le m\le M} \longmapsto \widetilde{L}\widetilde{z} = \big(-\tfrac{\widetilde{z}_{m+1}-2\widetilde{z}_m+\widetilde{z}_{m-1}}{h^2} + \alpha(x_m)\tfrac{\widetilde{z}_{m+1}-\widetilde{z}_{m-1}}{2h} + \beta(x_m)\widetilde{z}_m\big)_{1\le m\le M}.$$

Beachte! Der Operator L ist vorerst für auf dem offenen Intervall zweimal differenzierbare Funktionen $z : (a,b) \to \mathbb{R}$ definiert, und es ergibt sich eine Funktion $Lz : (a,b) \to \mathbb{R}$. Durch die Voraussetzung $z \in \mathscr{C}^2(a,b)$ kann man Lz (in eindeutiger Weise) auf das abgeschlossene Intervall $[a,b]$ fortsetzen. Der Operator $\widetilde{L}$ ist für Gitterfunktionen $\widetilde{z} : \overline{\Omega} \to \mathbb{R}$ definiert, und ergibt eine auf den inneren Gitterpunkten definierte Funktion $\widetilde{L}\widetilde{z} : \Omega \to \mathbb{R}$. Schreibt man für $\widetilde{L}\widetilde{z}$ dieselben Randwerte wie für $\widetilde{z}$ vor, erhält man eine auf allen Gitterpunkten definierte Funktion $\widetilde{L}\widetilde{z} : \overline{\Omega} \to \mathbb{R}$, bei der Funktionsvorschrift gibt man üblicherweise jedoch nur die Werte der inneren Gitterpunkte an.

Kompake Schreibweise: Mittels Matrix- und Vektorschreibweise ergibt sich

$$\big(\beta(x_m)\widetilde{z}_m\big)_{1\le m\le M} = \begin{pmatrix} \beta(x_1)\widetilde{z}_1 \\ \vdots \\ \beta(x_m)\widetilde{z}_m \\ \vdots \\ \beta(x_M)\widetilde{z}_M \end{pmatrix} = \underbrace{\begin{pmatrix} \beta(x_1) & & & \\ & \ddots & & \\ & & \beta(x_m) & \\ & & & \ddots \\ & & & & \beta(x_M) \end{pmatrix}}_{=A_0} \begin{pmatrix} \widetilde{z}_1 \\ \vdots \\ \widetilde{z}_m \\ \vdots \\ \widetilde{z}_M \end{pmatrix}$$

sowie

$$\begin{aligned} \big(\alpha(x_m)\tfrac{\widetilde{z}_{m+1}-\widetilde{z}_{m-1}}{2h}\big)_{1\le m\le M} &= \tfrac{1}{2h} \begin{pmatrix} \alpha(x_1)(\widetilde{z}_2-\widetilde{z}_0) \\ \vdots \\ \alpha(x_m)(\widetilde{z}_{m+1}-\widetilde{z}_{m-1}) \\ \vdots \\ \alpha(x_M)(\widetilde{z}_{M+1}-\widetilde{z}_{M-1}) \end{pmatrix} \\ &= \underbrace{\tfrac{1}{2h} \begin{pmatrix} 0 & \alpha(x_1) & & & \\ & \ddots & & & \\ & -\alpha(x_m) & 0 & \alpha(x_m) & \\ & & & \ddots & \\ & & & -\alpha(x_M) & 0 \end{pmatrix}}_{=A_1} \begin{pmatrix} \widetilde{z}_1 \\ \vdots \\ \widetilde{z}_m \\ \vdots \\ \widetilde{z}_M \end{pmatrix} + \underbrace{\tfrac{1}{2h} \begin{pmatrix} -\alpha(x_1)\widetilde{z}_0 \\ 0 \\ \vdots \\ 0 \\ \alpha(x_M)\widetilde{z}_{M+1} \end{pmatrix}}_{=b_1} \end{aligned}$$

und weiters

$$\left(\frac{\widetilde{z}_{m+1}-2\widetilde{z}_m+\widetilde{z}_{m-1}}{h^2}\right)_{1\le m\le M} = \frac{1}{h^2}\begin{pmatrix} \widetilde{z}_2-2\widetilde{z}_1+\widetilde{z}_0 \\ \vdots \\ \widetilde{z}_{m+1}-2\widetilde{z}_m+\widetilde{z}_{m-1} \\ \vdots \\ \widetilde{z}_{M+1}-2\widetilde{z}_M+\widetilde{z}_{M-1} \end{pmatrix}$$

$$= \underbrace{\frac{1}{h^2}\begin{pmatrix} -2 & 1 & & & \\ & \ddots & & & \\ & 1 & -2 & 1 & \\ & & & \ddots & \\ & & & 1 & -2 \end{pmatrix}}_{=A_2}\begin{pmatrix} \widetilde{z}_1 \\ \vdots \\ \widetilde{z}_m \\ \vdots \\ \widetilde{z}_M \end{pmatrix} + \underbrace{\frac{1}{h^2}\begin{pmatrix} \widetilde{z}_0 \\ 0 \\ \vdots \\ 0 \\ \widetilde{z}_{M+1} \end{pmatrix}}_{=b_2}.$$

Insgesamt führt dies auf die kompakte Darstellung (affin-lineare Abbildung)

$$\widetilde{L}:\mathbb{R}^M \longrightarrow \mathbb{R}^M : \widetilde{z} \longmapsto \widetilde{L}\widetilde{z} = A\widetilde{z}+b,$$

mit

$$A = A_0 + A_1 - A_2 \in \mathbb{R}^{M\times M},$$

$$A = \begin{pmatrix} \beta(x_1)+\frac{2}{h^2} & \frac{1}{2h}\alpha(x_1)-\frac{1}{h^2} & & & \\ & \ddots & & & \\ & -\frac{1}{2h}\alpha(x_m)-\frac{1}{h^2} & \beta(x_m)+\frac{2}{h^2} & \frac{1}{2h}\alpha(x_m)-\frac{1}{h^2} & \\ & & & \ddots & \\ & & & -\frac{1}{2h}\alpha(x_M)-\frac{1}{h^2} & \beta(x_M)+\frac{2}{h^2} \end{pmatrix},$$

und (Einsetzen der vorgegebenen Randwerte)

$$b = b_1 - b_2 = \begin{pmatrix} -\left(\frac{1}{2h}\alpha(x_1)+\frac{1}{h^2}\right)y_a \\ 0 \\ \vdots \\ 0 \\ \left(\frac{1}{2h}\alpha(x_M)-\frac{1}{h^2}\right)y_b \end{pmatrix} \in \mathbb{R}^M.$$

Differenzenverfahren für Randwertprobleme: Die näherungsweise Lösung des Randwertproblems

$$Ly = \gamma$$

mittels symmetrischer Differenzen entspricht der Lösung des linearen Gleichungssystems (wegen $\widetilde{L}\widetilde{z} = A\widetilde{z} + b$, mit Tridiagonalmatrix A)

$$\widetilde{L}\,\widetilde{y} = \widetilde{\gamma} \qquad \Longleftrightarrow \qquad A\,\widetilde{y} = \widetilde{\gamma} - b\,,$$

wobei die Approximation $\widetilde{y} = (\widetilde{y}(x_m))_{1\le m\le M} = (\widetilde{y}_m)_{1\le m\le M}$ bei vorgegebener rechter Seite $\widetilde{\gamma} = \big(\widetilde{\gamma}(x_m)\big)_{1\le m\le M}$ zu bestimmen ist.

- Fragestellungen:
  - Lösbarkeit des linearen Gleichungssystems, d.h. Existenz und Eindeutigkeit der diskreten Lösung
  - Konvergenz der diskreten Lösung $\widetilde{y}$ gegen die exakte Lösung y für $h \to 0$, Approximationsgüte

- Vorbemerkung: Anstelle der direkten Betrachtung der Matrix A wird ein Maximumsprinzip verwendet (nützlich in Hinblick auf Verallgemeinerungen für Finite Differenzen Verfahren und Finite Elemente Verfahren zur Ortsdiskretisierung von partiellen Differentialgleichungen).

  Erinnerung: Für eine (zumindest stetige) Funktion $z : [a,b] \to \mathbb{R}$ bezeichnet

  $$\|\widetilde{z}\|_{\infty,[a,b]} = \|\widetilde{z}\|_\infty = \max_{a\le x\le b} |z(x)|\,.$$

  Bezeichnung: Für eine Gitterfunktion $\widetilde{z} : \overline{\Omega} \to \mathbb{R}$ bezeichnet

  $$\|\widetilde{z}\|_{\infty,\overline{\Omega}} = \max_{0\le m\le M+1} |\widetilde{z}_m|\,, \qquad \|\widetilde{z}\|_{\infty,\Omega} = \max_{1\le m\le M} |\widetilde{z}_m|\,.$$

  Diskretes Maximumsprinzip (Lemma 5.1): Es sei $\widetilde{z} : \overline{\Omega} \to \mathbb{R}$ eine Gitterfunktion mit

  $$\big(\widetilde{L}\widetilde{z}\big)_m \le 0\,, \qquad 1 \le m \le M\,.$$

  Weiters sei die Gitterweite $h > 0$ so gewählt, daß die Bedingungen $1 + \frac{h}{2}\,\alpha(x_m) \ge 0$ sowie $1 - \frac{h}{2}\,\alpha(x_m) \ge 0$ für $1 \le m \le M$ erfüllt sind. Falls $\beta \ge 0$ folgt

  $$\|\widetilde{z}\|_{\infty,\overline{\Omega}} = \max\big\{|\widetilde{z}_0|, |\widetilde{z}_{M+1}|\big\}\,.$$

  Denn: (i) Falls $\beta = 0$ gilt (Definition von $\widetilde{L}$)

  $$\big(\widetilde{L}\widetilde{z}\big)_m = -\,\frac{\widetilde{z}_{m+1} - 2\widetilde{z}_m + \widetilde{z}_{m-1}}{h^2} + \alpha(x_m)\,\frac{\widetilde{z}_{m+1} - \widetilde{z}_{m-1}}{2h}\,, \qquad 1 \le m \le M\,.$$

  Sollte das Maximum in einem inneren Punkt angenommen werden

  $$\|\widetilde{z}\|_{\infty,\overline{\Omega}} = |\widetilde{z}_j| \qquad \text{mit} \quad 1 \le j \le M\,,$$

folgt durch Umformen der Relation (Anwendung der Voraussetzungen $(\widetilde{L}\widetilde{z})_j \le 0$ und $1 \pm \frac{h}{2}\,\alpha(x_m) \ge 0$)

$$
\begin{aligned}
&(\widetilde{L}\widetilde{z})_j = -\tfrac{\widetilde{z}_{j+1}-2\widetilde{z}_j+\widetilde{z}_{j-1}}{h^2} + \alpha(x_j)\,\tfrac{\widetilde{z}_{j+1}-\widetilde{z}_{j-1}}{2h} \\
\iff\quad & \tfrac{1}{2}\,h^2\,(\widetilde{L}\widetilde{z})_j = -\tfrac{1}{2}\,\widetilde{z}_{j+1} + \widetilde{z}_j - \tfrac{1}{2}\,\widetilde{z}_{j-1} + \tfrac{1}{4}\,h\,\alpha(x_j)\,(\widetilde{z}_{j+1}-\widetilde{z}_{j-1}) \\
\iff\quad & \tfrac{1}{2}\,h^2\,(\widetilde{L}\widetilde{z})_j = \widetilde{z}_j - \tfrac{1}{2}\,\big(1-\tfrac{h}{2}\,\alpha(x_j)\big)\,\widetilde{z}_{j+1} - \tfrac{1}{2}\,\big(1+\tfrac{h}{2}\,\alpha(x_j)\big)\,\widetilde{z}_{j-1} \\
\iff\quad & \widetilde{z}_j = \underbrace{\tfrac{1}{2}\,h^2\,(\widetilde{L}\widetilde{z})_j}_{\le 0} + \underbrace{\tfrac{1}{2}\,\big(1-\tfrac{h}{2}\,\alpha(x_j)\big)\,\widetilde{z}_{j+1} + \tfrac{1}{2}\,\big(1+\tfrac{h}{2}\,\alpha(x_j)\big)\,\widetilde{z}_{j-1}}_{\le \max\{\widetilde{z}_{j-1},\widetilde{z}_{j+1}\}} \\
\iff\quad & \widetilde{z}_j \le \max\big\{\widetilde{z}_{j-1},\widetilde{z}_{j+1}\big\} \\
\implies\quad & \widetilde{z}_{j-1} = \widetilde{z}_j = \widetilde{z}_{j+1}\,.
\end{aligned}
$$

Eine wiederholte Anwendung der Argumentation zeigt, daß $\widetilde{z}$ notwendigerweise konstant ist (und insbesondere das Maximum am Rand angenommen wird)

$$\widetilde{z}_0 = \widetilde{z}_1 = \cdots = \widetilde{z}_{M+1}\,.$$

(ii) Wie zuvor wird angenommen, daß das Maximum an einem inneren Gitterpunkt angenommen wird, d.h. es gelte $\|\widetilde{z}\|_{\infty,\overline{\Omega}} = |\widetilde{z}_j|$ mit $1 \le j \le M$. Unter der Voraussetzung $\beta \ge 0$ (punktweise), folgt ähnlich wie zuvor

$$
\begin{aligned}
&(\widetilde{L}\widetilde{z})_j = -\tfrac{\widetilde{z}_{j+1}-2\widetilde{z}_j+\widetilde{z}_{j-1}}{h^2} + \alpha(x_j)\,\tfrac{\widetilde{z}_{j+1}-\widetilde{z}_{j-1}}{2h} + \beta(x_j)\,\widetilde{z}_j \\
\iff\quad & \tfrac{1}{2}\,h^2\,(\widetilde{L}\widetilde{z})_j = \big(1+\beta(x_j)\big)\,\widetilde{z}_j - \tfrac{1}{2}\,\big(1-\tfrac{h}{2}\,\alpha(x_j)\big)\,\widetilde{z}_{j+1} - \tfrac{1}{2}\,\big(1+\tfrac{h}{2}\,\alpha(x_j)\big)\,\widetilde{z}_{j-1} \\
\iff\quad & \widetilde{z}_j = \underbrace{\tfrac{1}{2}\,h^2\,(\widetilde{L}\widetilde{z})_j}_{\le 0} + \underbrace{\tfrac{1}{2}\,\big(1-\tfrac{h}{2}\,\alpha(x_j)\big)\,\widetilde{z}_{j+1} + \tfrac{1}{2}\,\big(1+\tfrac{h}{2}\,\alpha(x_j)\big)\,\widetilde{z}_{j-1}}_{\le \max\{\widetilde{z}_{j-1},\widetilde{z}_{j+1}\}} \\
\iff\quad & \widetilde{z}_j \le \max\big\{\widetilde{z}_{j-1},\widetilde{z}_{j+1}\big\} - \beta(x_j)\,\widetilde{z}_j\,.
\end{aligned}
$$

Falls $\widetilde{z}_j \ge 0$ erhält man die Abschätzung

$$\widetilde{z}_j \le \max\big\{\widetilde{z}_{j-1},\widetilde{z}_{j+1}\big\} \underbrace{-\beta(x_j)\,\widetilde{z}_j}_{\le 0} \le \max\big\{\widetilde{z}_{j-1},\widetilde{z}_{j+1}\big\} \quad\implies\quad \widetilde{z}_{j-1} = \widetilde{z}_j = \widetilde{z}_{j+1}\,,$$

und falls $\widetilde{z}_j \le 0$ verwendet man

$$\widetilde{z}_j \le \widetilde{z}_j + \underbrace{\beta(x_j)\,\widetilde{z}_j}_{\ge 0} \le \max\big\{\widetilde{z}_{j-1},\widetilde{z}_{j+1}\big\} \quad\implies\quad \widetilde{z}_{j-1} = \widetilde{z}_j = \widetilde{z}_{j+1}\,.$$

Eine wiederholte Anwendung der Argumentation zeigt wiederum, daß $\widetilde{z}$ notwendigerweise konstant ist (und insbesondere das Maximum am Rand angenommen wird). ◇

Bemerkung: Lemma 5.1 ist das diskrete Analogon zum Maximumsprinzip für Funktionen, welches besagt, daß eine Funktion $z \in \mathscr{C}^2(a,b)$ mit $Lz \le 0$ (punktweise) ihr Maximum am Rand des Definitionsbereiches $[a,b]$ (d.h. im Punkt a oder b) annimmt. Insbesondere für den Differentialoperator $L = -\frac{\mathrm{d}^2}{\mathrm{d}x^2}$ entspricht die Bedingung $Lz \le 0$ der

Eigenschaft, daß die Funktion z konvex ist und die Gültigkeit des Maximumsprinzipes ist dann offensichtlich.
Beispiel: Für die Funktion $z : [a,b] \to \mathbb{R} : x \mapsto z(x) = x^2$ (nach oben geöffnete Parabel mit Maximum am Rand) folgt $\frac{\mathrm{d}^2}{\mathrm{d}x^2} z = 2$ und damit $Lz = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} z \le 0$.

Erinnerung: Im Folgenden wird weiterhin die Voraussetzung $\beta \ge 0$ verwendet.

Abschätzung für Gitterfunktionen (Lemma 5.2): Falls die Gitterweite $h > 0$ hinreichend klein gewählt ist, erfüllt jede Gitterfunktion $\widetilde{z} : \overline{\Omega} \to \mathbb{R}$ die Abschätzung

$$\|\widetilde{z}\|_{\infty,\overline{\Omega}} \le C \left\|\widetilde{L}\widetilde{z}\right\|_{\infty,\Omega} + \max\left\{|\widetilde{z}_0|, |\widetilde{z}_{M+1}|\right\}.$$

Denn: (i) Zur (nichttrivialen) Konstruktion einer Gitterfunktion $\widetilde{\zeta} : \overline{\Omega} \to \mathbb{R} : x_m \to \widetilde{\zeta}_m$, die den folgenden Eigenschaften genügt

$$\widetilde{\zeta}_m \ge 0, \quad 0 \le m \le M+1, \qquad \left(\widetilde{L}\widetilde{\zeta}\right)_m \ge 1, \quad 1 \le m \le M,$$

verwendet man den Ansatz (mit Konstante $\lambda > 0$)

$$\widetilde{\zeta}_m = \mathrm{e}^{\lambda} - \mathrm{e}^{\lambda \frac{x_m - a}{b-a}}, \qquad 0 \le m \le M+1.$$

Die erste Bedingung ist offensichtlich erfüllt (Abschätzung $\frac{x_m - a}{b-a} \le 1$, Monotonie der Exponentialfunktion)

$$\widetilde{\zeta}_m = \mathrm{e}^{\lambda} - \mathrm{e}^{\lambda \frac{x_m - a}{b-a}} \ge 0, \qquad 0 \le m \le M+1.$$

Andererseits folgt (mit $\mu = \frac{\lambda}{b-a}$, verwende $\cosh x = \frac{1}{2}(\mathrm{e}^x + \mathrm{e}^{-x})$ und $\sinh x = \frac{1}{2}(\mathrm{e}^x - \mathrm{e}^{-x})$)

$$\begin{aligned}
\left(\widetilde{L}\widetilde{\zeta}\right)_m &= -\frac{\widetilde{\zeta}_{m+1} - 2\widetilde{\zeta}_m + \widetilde{\zeta}_{m-1}}{h^2} + \alpha(x_m)\frac{\widetilde{\zeta}_{m+1} - \widetilde{\zeta}_{m-1}}{2h} + \beta(x_m)\widetilde{\zeta}_m \\
&= -\frac{1}{h^2}\left(\mathrm{e}^{\lambda} - \mathrm{e}^{\lambda\frac{x_{m+1}-a}{b-a}} - 2\mathrm{e}^{\lambda} + 2\mathrm{e}^{\lambda\frac{x_m-a}{b-a}} + \mathrm{e}^{\lambda} - \mathrm{e}^{\lambda\frac{x_{m-1}-a}{b-a}}\right) \\
&\quad + \alpha(x_m)\frac{1}{2h}\left(\mathrm{e}^{\lambda} - \mathrm{e}^{\lambda\frac{x_{m+1}-a}{b-a}} - \mathrm{e}^{\lambda} + \mathrm{e}^{\lambda\frac{x_{m-1}-a}{b-a}}\right) \\
&\quad + \underbrace{\beta(x_m)\left(\mathrm{e}^{\lambda} - \mathrm{e}^{\lambda\frac{x_m-a}{b-a}}\right)}_{\ge 0} \\
&\ge \mathrm{e}^{\lambda\frac{x_m-a}{b-a}}\left(\frac{1}{h^2}\left(\mathrm{e}^{h\mu} + \mathrm{e}^{-h\mu} - 2\right) - \alpha(x_m)\frac{1}{2h}\left(\mathrm{e}^{h\mu} - \mathrm{e}^{-h\mu}\right)\right) \\
&= \underbrace{\mathrm{e}^{\lambda\frac{x_m-a}{b-a}}}_{\ge 1} \underbrace{\frac{2}{h^2}\left(2\left(\cosh(h\mu) - 1\right) - h\,\alpha(x_m)\sinh(h\mu)\right)}_{\overset{!}{\ge} 1}.
\end{aligned}$$

Bei geeigneter Wahl von $h > 0$ (hinreichend klein) und $\mu > 0$ (d.h. λ hinreichend groß) läßt sich auch die zweite Bedingung erfüllen (setze $x = h\mu$ und verwende die Abschätzung $\alpha(x_m) \le \alpha_{\max}$ durch den Maximalwert)

$$\begin{aligned}
&\frac{2}{h^2}\left(2\left(\cosh(h\mu) - 1\right) - h\,\alpha(x_m)\sinh(h\mu)\right) \\
&\quad \ge \begin{cases} \frac{4\mu^2}{x^2}\left(\cosh x - 1\right), & \alpha \le 0, \\ \frac{2\mu^2}{x^2}\left(2\left(\cosh x - 1\right) - h\,\alpha_{\max}\sinh x\right), & \text{sonst}. \end{cases}
\end{aligned}$$

(ii) Für die Gitterfunktion $\widetilde{z} : \overline{\Omega} \to \mathbb{R}$ definiert man mittels der zuvor konstruierten Gitterfunktion $\zeta : \overline{\Omega} \to \mathbb{R}$ in Abhängigkeit vom Vorzeichen von $\widetilde{z}_0$

$$\widetilde{v} = \widetilde{z} - \left\|\widetilde{L}\widetilde{z}\right\|_{\infty,\Omega} \widetilde{\zeta} \qquad \text{oder} \qquad \widetilde{v} = -\widetilde{z} - \left\|\widetilde{L}\widetilde{z}\right\|_{\infty,\Omega} \widetilde{\zeta},$$

derart daß $|\widetilde{v}_0| = |\widetilde{z}_0 - \|\widetilde{L}\widetilde{z}\|_{\infty,\Omega}\, \mathrm{e}^{\lambda}| \le |\widetilde{z}_0|$ (falls $\widetilde{z}_0 \ge 0$) oder $|\widetilde{v}_0| = |\widetilde{z}_0 + \|\widetilde{L}\widetilde{z}\|_{\infty,\Omega}\, \mathrm{e}^{\lambda}| \le |\widetilde{z}_0|$ (falls $\widetilde{z}_0 \le 0$). Es gilt (verwende Linearität von $\widetilde{L}$, o.E.d.A. $\left\|\widetilde{L}\widetilde{z}\right\|_{\infty,\Omega} > 0$)

$$\widetilde{L}\,\widetilde{v} = \pm\widetilde{L}\,\widetilde{z} - \left\|\widetilde{L}\widetilde{z}\right\|_{\infty,\Omega} \widetilde{L}\widetilde{\zeta} = -\left\|\widetilde{L}\widetilde{z}\right\|_{\infty,\Omega} \left(\widetilde{L}\widetilde{\zeta} \mp \tfrac{1}{\|\widetilde{L}\widetilde{z}\|_{\infty,\Omega}} \widetilde{L}\,\widetilde{z}\right),$$
$$\left(\widetilde{L}\,\widetilde{v}\right)_m = -\left\|\widetilde{L}\widetilde{z}\right\|_{\infty,\Omega} \underbrace{\left(\left(\widetilde{L}\widetilde{\zeta}\right)_m \mp \tfrac{1}{\|\widetilde{L}\widetilde{z}\|_{\infty,\Omega}} \left(\widetilde{L}\,\widetilde{z}\right)_m\right)}_{\ge 0} \le 0, \qquad 1 \le m \le M.$$

Mittels Lemma 5.1 folgt (für $h > 0$ hinreichend klein, Gitterfunktionen $\widetilde{v}$ erfüllen Voraussetzung $\widetilde{L}\,\widetilde{v} \le 0$ (komponentenweise), nach Konstruktion ist $|\widetilde{v}_0| \le |\widetilde{z}_0|$ und wegen $\widetilde{\zeta}_{M+1} = 0$ ist $|\widetilde{v}_{M+1}| = |\widetilde{z}_{M+1}|$)

$$|\widetilde{v}_m| \le \|\widetilde{v}\|_{\infty,\overline{\Omega}} = \max\left\{|\widetilde{v}_0|, |\widetilde{v}_{M+1}|\right\} \le \max\left\{|\widetilde{z}_0|, |\widetilde{z}_{M+1}|\right\},$$

und damit wegen ($\widetilde{z}_m = \pm\widetilde{v}_m \pm \left\|\widetilde{L}\widetilde{z}\right\|_{\infty,\Omega} \widetilde{\zeta}_m$)

$$|\widetilde{z}_m| \le |\widetilde{v}_m| + \left\|\widetilde{L}\widetilde{z}\right\|_{\infty,\Omega} \underbrace{|\widetilde{\zeta}_m|}_{\le C} \le \max\left\{|\widetilde{z}_0|, |\widetilde{z}_{M+1}|\right\} + C\left\|\widetilde{L}\widetilde{z}\right\|_{\infty,\Omega}$$

die Behauptung. ◇

Bemerkung: Lemma 5.2 ist das diskrete Analogon zu einem Resultat, welches besagt, daß eine Funktion $z \in \mathscr{C}^2(a,b)$ die Abschätzung

$$\max_{a\le x\le b} \left|z(x)\right| \le C \max_{a\le x\le b} \left|Lz(x)\right| + \max\left\{\left|z(a)\right|, \left|z(b)\right|\right\}$$

erfüllt. Insbesondere im Zusammenhang mit Randwertproblemen

$$L\,y = \gamma \qquad \text{bzw.} \qquad \widetilde{L}\,\widetilde{y} = \widetilde{\gamma}$$

bezeichnet man die Abschätzung von Lemma 5.2 als Stabilitätsabschätzung (Schranke für die Werte der Lösung in Abhängigkeit von den Eingabedaten, d.h. von den Randwerten und der rechten Seite γ).

Bemerkung: Bei Differentialgleichungen, wo $\beta = 0$ oder $\beta \ge 0$ bzw. $\beta \le 0$ (kein Vorzeichenwechsel) kann man die Einschränkung an die Gitterweite $h > 0$ vermeiden. Im letzteren Fall verwendet man anstelle der symmetrischen ersten Differenzen die einseitigen Differenzen (Rückwärtsdifferenz $\frac{y(x-h)-y(x)}{-h} = \frac{y(x)-y(x-h)}{h} = \frac{\mathrm{d}}{\mathrm{d}x}\, y(x) + \mathscr{O}(h)$ falls $\beta \ge 0$ bzw. Vorwärtsdifferenz $\frac{y(x+h)-y(x)}{h} = \frac{\mathrm{d}}{\mathrm{d}x}\, y(x) + \mathscr{O}(h)$ falls $\beta \le 0$, vgl. Upwindverfahren bei Diffusions-Advektionsgleichungen).

- Eindeutige Lösbarkeit: Das lineare Gleichungssystem

$$\widetilde{L}\,\widetilde{y} = \widetilde{\gamma}$$

besitzt eine eindeutige Lösung (für hinreichend kleine Schrittweiten $h > 0$).
Denn: Frühere Überlegungen zeigten (mit $\widetilde{y} = (\widetilde{y}(x_1), \ldots, \widetilde{y}(x_m))^T$)

$$\widetilde{L}\,\widetilde{y} = A\,\widetilde{y} + b = \widetilde{\gamma}\,.$$

Es reicht aus, das homogene lineare Gleichungssystem

$$A\,\widetilde{y} = 0$$

zu betrachten und zu zeigen, daß nur die triviale Lösung $\widetilde{y} = 0$ existiert. Dies entspricht dem Fall $\widetilde{\gamma} = 0$ und $\widetilde{y}_0 = \widetilde{y}(x_0) = y_a = 0$ sowie $\widetilde{y}_{M+1} = \widetilde{y}(x_{M+1}) = y_b = 0$ (und somit $b = 0$). Mittels Lemma 5.2 folgt

$$\|\widetilde{y}\|_{\infty,\overline{\Omega}} \le C \underbrace{\left\|\widetilde{L}\widetilde{y}\right\|_{\infty,\Omega}}_{=0} + \underbrace{\max\{|\widetilde{y}_0|, |\widetilde{y}_{M+1}|\}}_{=0} = 0 \quad \Longleftrightarrow \quad \widetilde{y}_m = 0\,, \quad 1 \le m \le M\,,$$

was auf $\widetilde{y} = 0$ führt. ◇

Bemerkung: Die Lösung des linearen Gleichungssystems

$$\widetilde{L}\,\widetilde{y} = \widetilde{\gamma}$$

mit Tridiagonalmatrix $\widetilde{L} \in \mathbb{R}^{M\times M}$ benötigt $\mathcal{O}(M)$ Operationen (vgl. Abschnitt 2.1. zu Kubischen Splineinterpolanten).

- Konvergenz der diskreten Lösung: Wie zuvor bezeichnet $y : [a,b] \to \mathbb{R}$ die Lösung des Randwertproblems und $\widetilde{y} : \Omega \to \mathbb{R}$ die Lösung des mittels symmetrischer Differenzen erhaltenen linearen Gleichungssystems (an den inneren Gitterpunkten, zu den exakten Randwerten)

$$Ly = \gamma \qquad \text{bzw.} \qquad \widetilde{L}\,\widetilde{y} = \widetilde{\gamma}\,.$$

Die Differenz (Einschränkung der exakten Lösung auf die inneren Gitterpunkte, in den Randpunkten ist nach Annahme $d_0 = 0 = d_{M+1}$)

$$d = \widetilde{y} - y\big|_\Omega : \Omega \longrightarrow \mathbb{R} : x_m \longmapsto d_m = \widetilde{y}_m - y_m$$

erfüllt an den inneren Gitterpunkten die Relation

$$\begin{aligned}
\widetilde{L}\,d = \widetilde{L}\,\widetilde{y} - \widetilde{L}\,y\big|_\Omega = \widetilde{\gamma} - \widetilde{L}\,y\big|_\Omega = \gamma\big|_\Omega - \widetilde{L}\,y\big|_\Omega = (L\,y)\big|_\Omega - \widetilde{L}\,y\big|_\Omega\,,\\
(L\,y)(x_m) = -\tfrac{\mathrm{d}^2}{\mathrm{d}x^2}\,y(x)\big|_{x=x_m} + \alpha(x_m)\,\tfrac{\mathrm{d}}{\mathrm{d}x}\,y(x)\big|_{x=x_m} + \beta(x_m)\,y(x_m)\,,\\
\left(\widetilde{L}\,y\big|_\Omega\right)_m = -\tfrac{1}{h^2}\left(y(x_{m+1}) - 2\,y(x_m) + y(x_{m-1})\right) + \alpha(x_m)\,\tfrac{1}{2h}\left(y(x_{m+1}) - y(x_{m-1})\right)\\
\qquad + \beta(x_m)\,y(x_m)\,, \qquad 1 \le m \le M\,.
\end{aligned}$$

Frühere Überlegungen zeigten (Taylorreihenentwicklungen symmetrischer Differenzen, jeweils mit $\xi \in [x-h, x+h]$)

$$y \in \mathscr{C}^3(a,b): \qquad \tfrac{y(x+h)-y(x-h)}{2h} - \tfrac{\mathrm{d}}{\mathrm{d}x}\, y(x) = \tfrac{1}{6}\, h^2\, \tfrac{\mathrm{d}^3}{\mathrm{d}x^3}\, y(x)\big|_{x=\xi},$$
$$y \in \mathscr{C}^4(a,b): \qquad \tfrac{y(x+h)-2\,y(x)+y(x-h)}{h^2} - \tfrac{\mathrm{d}^2}{\mathrm{d}x^2}\, y(x) = \tfrac{1}{12}\, h^2\, \tfrac{\mathrm{d}^4}{\mathrm{d}x^4}\, y(x)\big|_{x=\xi}.$$

und da die Koeffizienten α, β auf $[a,b]$ stetig und damit beschränkt sind ergibt sich die Abschätzung

$$\big\| \widetilde{L}\big(\widetilde{y} - y\big|_\Omega\big) \big\|_{\infty,\Omega} \le C\, h^2\, \big\| y^{(4)} \big\|_\infty .$$

Mittels Lemma 5.2 erhält man somit die (globale) Fehlerabschätzung (vergleichsweise einschränkende Regularitätsvoraussetzungen für Ordnung 2)

$$\big\| \widetilde{y} - y\big|_\Omega \big\|_{\infty,\Omega} \le C\, h^2\, \big\| y^{(4)} \big\|_\infty .$$

Bemerkungen:

- Die Idee der Extrapolation mit Lösungen $\widetilde{y}_h$ und $\widetilde{y}_{h/2}$ zu den Gitterweiten h und $\frac{h}{2}$ führt auf die verbesserte Approximation (Ordnung 4)

$$\tfrac{1}{3}\big(4\widetilde{y}_{h/2,m} - \widetilde{y}_{h,m}\big)$$

- Die Idee der Adaptivität führt auf nichtuniforme Gitter.
- Die Anwendung von Differenzenverfahren auf nichtlineare Differentialgleichungen führt auf nichtlineare Gleichungssysteme (Lösbarkeit und Konvergenz deutlich schwieriger zu analysieren).