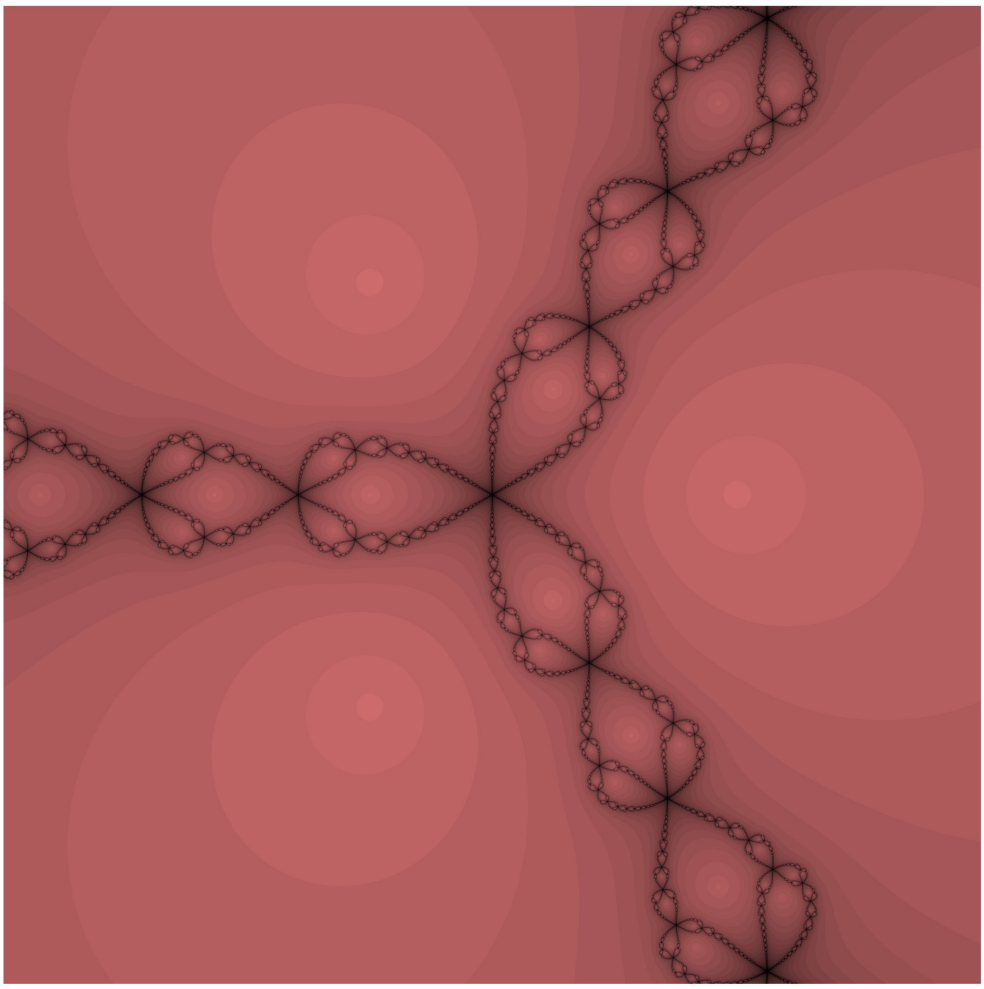
Kompendium zur Lehrveranstaltung

Numerische Mathematik

Mechthild Thalhammer



Leopold–Franzens-Universität Innsbruck

Studienjahr 2022/23

Die Mathematik ist mehr ein Tun als eine Lehre.

Mathematik zu lernen heißt, sie immer wieder neu zu erfinden.

Frustration und Euphorie liegen in der Mathematik oft knapp nebeneinander.

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

Wenn das die Lösung ist, will ich mein Problem zurück.

Es geht einfach oder es geht einfach nicht.

(L. Brouwer, D. O'Shea, R. Höfer, R. Taschner, Unbekannt)

Illustration. Auf der Titelseite ist das Konvergenzverhalten des Newtonverfahrens zur näherungsweisen Lösung der nichtlinearen komplexen Gleichung $z^3 = 1$ illustriert.

Das Gebiet der Numerik ist ein Teilgebiet der Mathematik, das sich insbesondere der Konstruktion und Analyse von Algorithmen widmet.

Das vorliegende Kompendium faßt die im Rahmen der Lehrveranstaltung **Numerische Mathematik** (VO 3 & PS 2) im Studienjahr 2022/23 an der Universität Innsbruck behandelten Themen zusammen. Ohne Anspruch auf Allgemeinheit und Vollständigkeit werden Methoden zur näherungsweisen Lösung von grundlegenden Problemen der Analysis und Linearen Algebra angegeben. Als Illustrationen werden einfache Modellprobleme und Implementierungen betrachtet.

Das Kompendium basiert vorwiegend auf den Vorlesungsskripten

> MATHIAS RICHTER
> *Numerische Mathematik I & II* (2010/2011)

und Unterlagen, welche im darauffolgenden Studienjahr 2011/12 an der Fakultät für Luft- und Raumfahrttechnik der Universität der Bundeswehr München erstellt wurden. Die damals gewählte Strukturierung und entsprechende Verweise wurden beibehalten.

Als ergänzende Literaturquellen werden die Vorlesungsskripten

> HERMANN SCHICHL
> *Numerik 1 & 2* (2000/2001/2011)
> www.mat.univie.ac.at/~herman/skripten
>
> ERNST HAIRER, GERHARD WANNER
> *Introduction á l'Analyse Numérique* (2005)
> www.unige.ch/~hairer/polycop

sowie dort erwähnte Standardwerke empfohlen. Was eine umfassende Einführung in die Software MATLAB betrifft, wird zudem auf

> PETER ARBENZ
> *Einführung in* MATLAB (2007/2008)
> people.inf.ethz.ch/arbenz/MatlabKurs/matlabintro.pdf

verwiesen.

Im Internet zu findende Unterlagen bieten die Vorteile kompakter Darstellungen und freier Verfügbarkeit, sollten jedoch mit einem kritischen Blick auf inhaltliche Richtigkeit und mögliche Druckfehler verwendet werden.

Themenüberblick I

1. Einführung

- **Numerische Mathematik**: Konkrete Lösung mathematischer Probleme, d.h. konstruktive Beschaffung von Lösungen mittels Zahlenrechnungen.

  Theoretische Resultate und Formelmanipulationen sind von Nutzen.

- **Numerische Verfahren der Linearen Algebra**:
  - Verfahren zur Lösung linearer Gleichungssysteme.

    Inbesondere diese Grundaufgabe der Numerik ist nach wie vor Gegenstand aktueller Forschung.
  - Verfahren zur Berechnung von Eigenwerten und Eigenvektoren einer Matrix.

  Anwendungen:
  - Verfahren zur Lösung nichtlinearer Gleichungssysteme
  - Verfahren zur Lösung von Optimierungsproblemen
  - Verfahren zur Lösung gewöhnlicher Differentialgleichungen
  - Verfahren zur Lösung partieller Differentialgleichungen

- **Numerisches Problem** (Definition 2.7): Funktion, die zulässigen Eingabedaten ein Ergebnis zuordnet
$$p : D \subset \mathbb{R}^n \to \mathbb{R}^m : x \mapsto y = p(x)\,.$$

  Numerisches Verfahren bzw. **Algorithmus** zur Lösung eines Problems $p : D \subset \mathbb{R}^n \to \mathbb{R}^m$: Endliche Folge von Teilproblemen (d.h. von elementaren Operationen), deren Reihenfolge beim Ablauf eindeutig festliegt
$$\left(p^{(1)}, \ldots, p^{(k)}\right).$$
  Die schrittweise Anwendung der Teilprobleme $p^{(i)}$ für $i = 1, 2, \ldots, k$ führt auf Zwischenergebnisse bzw. das Endergebnis
$$y^{(i)} = \left(p^{(i)} \circ \cdots \circ p^{(1)}\right)(x)\,, \qquad 1 \le i \le k\,.$$
  - **Direktes Verfahren**: Berechnung der exakten Lösung eines Problems in endlich vielen Rechenschritten (im Prinzip möglich), d.h. es ist
$$p^{(k)} \circ \cdots \circ p^{(1)} = p\,.$$

    Beispiele:
    * Formel von Vieta zur Lösung einer quadratischen Gleichung
    * Gaußsches Eliminationsverfahren zur Lösung eines linearen Gleichungssystems

– **Näherungsweises Verfahren** bzw. **Approximationsverfahren**: Berechnung einer Näherungslösung eines Problems, d.h. es ist

$$p^{(k)} \circ \cdots \circ p^{(1)} \approx p\,.$$

Beispiele:

* Approximation unendlicher Reihen durch endliche Reihen
* Approximation bestimmter Integrale durch Riemann-Summen
* Approximation von Ableitungen durch Differenzenquotienten
* Approximation mittels Iterationsverfahren

Verfahrensfehler bzw. **Approximationsfehler**: Fehler der Näherungslösung, d.h. Differenz zwischen näherungsweiser und exakter Lösung

$$y^{(k)} - y = \left(p^{(k)} \circ \cdots \circ p^{(1)}\right)(x) - p(x)\,.$$

- **Beispiele** (Probleme, Algorithmen):
  - **Elementare arithmetische Operationen** $* \in \{+, -, \times, /\}$

$$p : D \subset \mathbb{R} \times \mathbb{R} \to \mathbb{R} : (x, y) \mapsto x * y\,.$$

  - **Auswerten einer Polynomfunktion**

$$p : \mathbb{R}^{n+1} \times \mathbb{R} \to \mathbb{R} : (c, x) \mapsto \sum_{i=0}^{n} c_i x^i\,.$$

Algorithmus basierend auf dem Horner-Schema

$$\sum_{i=0}^{n} c_i x^i = \Big(\cdots\big((c_n x + c_{n-1})x + c_{n-2}\big)x + \cdots\Big)x + c_0\,.$$

  - **Berechnung der Nullstellen eines quadratischen Polynoms**, z.B. Berechnung der Lösungen $x_{1,2}$ der quadratischen Gleichung $x^2 + 2ax - b = 0$ (unter der Annahme $a, b > 0$ folgt $x_{1,2} \in \mathbb{R}$ mit $x_1 \neq x_2$)

$$p : \mathbb{R}^2 \to \mathbb{R}^2 : (a, b) \mapsto x\,.$$

Algorithmus basierend auf der Formel von Vieta, Algorithmus basierend auf einer Umformulierung

$$x_1 = -a + \sqrt{a^2 + b} \underset{\substack{\text{Erweitern mit} \\ a+\sqrt{a^2+b}}}{=} \frac{b}{a+\sqrt{a^2+b}}\,, \qquad x_2 = -a - \sqrt{a^2 + b}\,.$$

- **Berechnung der Nullstellen eines Polynoms** bzw. Lösung einer polynomialen Gleichung (Annahme $n \in \mathbb{N}$ mit $n \geq 1$ und $c_n \neq 0$)

$$p : \mathbb{C}^{n+1} \to \mathbb{C}^n : c \mapsto z \quad \text{mit } f(z_i) = \sum_{\ell=0}^{n} c_\ell z_i^\ell = 0 .$$

  Bemerkungen: Identifikation von $\mathbb{C}$ mit $\mathbb{R}^2$. Im allgemeinen ist keine explizite Lösungsformel bekannt.

- **Berechnung der Eigenwertzerlegung einer Matrix**, d.h. Berechnung der Eigenwerte $\lambda_1, \ldots, \lambda_n$ einer (diagonalisierbaren) Matrix $A \in \mathbb{C}^{n \times n}$ und zugehöriger Eigenvektoren $v = (v_1 | \cdots | v_n)$

$$p : \mathbb{C}^{n \times n} \to \mathbb{C}^n \times \mathbb{C}^{n \times n} : A \mapsto (\lambda, v) .$$

  Algorithmus basierend auf der Berechnung der Nullstellen des charakteristischen Polynoms und der Lösung linearer Gleichungssysteme

$$\chi(\lambda_i) = \det(\lambda_i I - A) = 0 , \qquad (\lambda_i I - A) v_i = 0 , \quad v_i \neq 0 , \qquad 1 \leq i \leq n .$$

  Bemerkung: Insbesondere für $n \gg 1$ ist die Entwicklung eines alternativen Algorithmus notwendig, vgl. **Illustration1**.

- **Lösung eines linearen Gleichungssystems** $Ax = b$ (unter Annahme $A \in \mathbb{R}^{n \times n}$ invertierbar)

$$p : \mathbb{R}^{n \times n} \times \mathbb{R}^n \to \mathbb{R}^n : (A, b) \mapsto A^{-1} b .$$

  Algorithmus basierend auf dem Eliminationsverfahren nach Gauß.

- Aufgrund der **Verwendung digitaler Rechner** zur Berechnung der Lösungen von (komplexen) Problemen sind gewisse Einschränkungen unvermeidbar.
  - Einschränkung auf elementare Operationen wie arithmetische Operationen, Wurzelziehen und arithmetische Vergleiche

$$* \in \{+, -, \times, /, \sqrt{}, <, \leq, =, \geq, >, \neq\} .$$

  - Einschränkung auf endlich viele darstellbare Zahlen (Maschinenzahlen, im Allgemeinen normalisierte Gleitpunktzahlen zur Basis 2 mit fester Stellenzahl und beschränktem Exponenten).
  - Auftreten von Rundungsfehlern bei der Eingabe von Daten und der Berechnung von Zwischen- bzw. Endergebnissen.

  Auftreten von Datenfehlern, wenn die Eingabedaten beispielsweise das Ergebnis von Messungen mit eingeschränkter Genauigkeit sind.

- **Aufgaben der Numerik**:
  - Konstruktion *guter* Algorithmen beispielsweise hinsichtlich Genauigkeit (z.B. Rate der verbesserten Genauigkeit der Näherungslösung bei höherem technischem Aufwand), Effizienz (z.B. Anzahl der Rechenoperationen bzw. Rechenzeit oder benötigter Speicherplatz) und Stabilität (z.B. Auswirkung von kleinen Änderungen der Eingabedaten auf das Endergebnis).
  - Analyse von Verfahrensfehlern und Herleitung von Abschätzungen für Verfahrensfehler.
  - Analyse der Fortpflanzungen von Fehlern (Rundungsfehler, Datenfehler) in Algorithmen und Auswirkungen auf Endergebnisse.
- **Illustration** (Eigenwertberechnung):
  - Problem: Berechnung aller Eigenwerte einer reellen und symmetrischen Matrix.
  - Theoretisches Resultat sichert, daß alle Eigenwerte reell sind.
  - Algorithmus basierend auf der Berechnung der Nullstellen des charakteristischen Polynoms.
  - Kleine Änderungen der Koeffizienten des charakteristischen Polynoms (vergleichbar mit der Eingabe der Koeffizienten in einfacher Genauigkeit und dabei auftretenden Rundungsfehlern) und Berechnung der zugehörigen Nullstellen.
  - Vergleich der Ergebnisse (Nullstellen, Graphen der Polynome).
  - Vgl. **Illustration1**.

    Vgl. **Illustration1_Modifikation**: Falls die Dimension n der Matrix hinreichend groß ist, ist das Ergebnis auch für kleine relative Änderungen ε der Koeffizienten nicht zufriedenstellend.

    Ergänzungen: Binärdarstellung, Horner-Schema, Verfahren von Bairstow (Faktorisierung von Polynomen)
  - **Schlußfolgerungen**:
    - Problemstellung *Berechnung der Nullstellen eines Polynoms* (höherer Ordnung) bei praktischen Anwendungen nicht sinnvoll. Aufgrund unvermeidbarer (kleiner) Änderungen der Koeffizienten sind (exakt) berechnete Nullstellen wertlos.
    - Notwendigkeit der Entwicklung eines alternativen Algorithmus zur Eigenwertberechnung.
- **Vorsicht!** Algorithmen, die für die Theorie sinnvoll sind, können jedoch für die Numerik wertlos sein.

2. Grundbegriffe der Numerik

- **Fragestellungen**:
  - Welche Zahlen sind am Rechner darstellbar und wie werden elementare Rechenoperationen ausgeführt?
    Maschinenzahlen, Rundung, Gleitpunktoperationen, Rundungsfehleranalyse
  - Welche numerischen Probleme kann man zufriedenstellend lösen?
    Kondition eines Problems
  - Wie kann man die Güte eines numerischen Verfahrens hinsichtlich der erreichbaren Genauigkeit beurteilen?
    Stabilität eines Algorithmus

  Konkrete Anwendung der eingeführten Konzepte in nachfolgenden Kapiteln.

- Vertauschung von Abschnitt 2.4 und Abschnitt 2.3. Verbindung der Abschnitte 2.3 und 2.5.

2.1. Maschinenzahlen und Rundung

- **Körper der reellen Zahlen** $\mathbb{R}$: Gebräuchlichstes Zahlensystem in der Analysis, übliche Rechenregeln für Addition und Multiplikation, vollständige und archimedisch geordnete Menge, Veranschaulichung als unendlich lange lückenlose Linie (Zahlengerade).

  Menge der Maschinenzahlen: Auf einem Rechner exakt darstellbare Zahlen (im Allgemeinen Elemente eines endlichen Systems normalisierter Gleitpunktzahlen, siehe Definition 2.2), endliche und insbesondere beschränkte Menge.

- **Normalisierte Gleitpunktzahl, Basis, Stellenzahl, Signifikand, Exponent** (Definition 2.1): Normalisierte t-stellige Gleitpunktzahl zur Basis B (wobei $B \in \mathbb{N}_{\geq 2}$, $t \in \mathbb{N}_{\geq 1}$)

$$g = 0 \quad \text{oder} \quad g = \pm\,|S| \cdot B^E \in \mathbb{G} = \mathbb{G}_{B,t}\,, \quad |S| \in \mathbb{N}_{\geq 1}\,, \quad B^{t-1} \leq |S| < B^t\,, \quad E \in \mathbb{Z}\,.$$

  Bemerkungen:

  – Die Menge $\mathbb{G}$ ist unendlich.

  – Eindeutigkeit der Darstellung (Signifikand, Exponent) aufgrund der Normalisierung des Signifikanden.

    Beispiel: Darstellung der Zahl $0.012345 = 0.12345 \cdot 10^{-1} = 1.2345 \cdot 10^{-2}$ (etc.) als $12345 \cdot 10^{-6} \in \mathbb{G}_{10,5}$.

  – **Auflösung** in $\mathbb{G}$:

$$\varrho = \max\left\{\left|\tfrac{\tilde{g}-g}{g}\right| : g, \tilde{g} \text{ benachtbarte Zahlen in } \mathbb{G} \setminus \{0\}\right\} = B^{1-t}\,.$$

    Denn: Für eine Zahl $g = \pm|S| \cdot B^E \in \mathbb{G}$ ist die benachtbarte Zahl gegeben durch $\tilde{g} = (\pm|S| + 1) \cdot B^E$ bzw. $\tilde{g} = (\pm|S| - 1) \cdot B^E$ und daher $\left|\tfrac{\tilde{g}-g}{g}\right| = \tfrac{1}{|S|} \leq B^{1-t}$, sofern $g, \tilde{g} \neq 0$. Speziell für $\tilde{g} = 0$ folgt jedoch $\left|\tfrac{\tilde{g}-g}{g}\right| = 1$. ◇

    Bemerkung: Betrachtung relativer Größen anstelle absoluter Größen.

  Maschinenzahl (Definition 2.2): Normalisierte t-stellige Gleitpunktzahl zur Basis B mit beschränktem Exponenten, d.h. $g = 0$ oder (wobei $B \in \mathbb{N}_{\geq 2}$, $t \in \mathbb{N}_{\geq 1}$ und $\alpha, \beta \in \mathbb{Z}$, $\alpha \leq \beta$)

$$g = \pm\,|S| \cdot B^E \in \mathbb{M} = \mathbb{M}_{B,t,\alpha,\beta} \subset \mathbb{G}_{B,t}\,, \quad |S| \in \mathbb{N}_{\geq 1}\,, \quad B^{t-1} \leq |S| < B^t\,, \quad E \in \mathbb{Z}\,, \quad \alpha \leq E \leq \beta\,.$$

  Bemerkungen:

  – Die Menge $\mathbb{M}$ ist endlich.

  – Maschinenzahlen sind (ebenso wie Gleitpunktzahlen) nicht äquidistant verteilt. Die Auflösung in $\mathbb{M}$ stimmt mit der Auflösung in $\mathbb{G}$ überein, d.h. $\varrho = B^{1-t}$.

  – **Kleinste positive Maschinenzahl** $\sigma = B^{t-1+\alpha}$.
    Größte Maschinenzahl $\lambda = (B^t - 1)\, B^\beta \doteq B^{t+\beta}$.

– Übliche Genauigkeitsstufen sind **single precision** bzw. **double precision** bzw. **extended precision**, festgelegt durch den Standard ANSI/IEEE-Std-754-1985 für Gleitpunktarithmetik.

Vgl. **Illustration2_Maschinenzahlen**.

- **Bereichsüberschreitungen** treten auf, wenn Rechenergebnisse außerhalb des definierten Bereichs $[-\lambda,-\sigma]\cup\{0\}\cup[\sigma,\lambda]$ liegen. In mathematischen Software-Packeten werden daher auch zusätzliche **Sonderoperanden** verwendet.

  Vgl. **Illustration2_Sonderoperanden** (Sonderoperanden $\pm\infty$ und quiet NAN not-a-number in MATLAB) und auch **Illustration2_Rundung**.

- **Vereinbarung**: Unter der Annahme, daß Bereichsüberschreitungen vermieden werden können, wird von nun an die Menge der normalisierten Gleitpunktzahlen $\mathbb{G}$ anstelle der Menge der Maschinenzahlen $\mathbb{M}$ betrachtet.

- Eingangsdaten und auch die Ergebnisse von elementaren arithmetischen Operationen liegen im Allgemeinen nicht im Bereich der darstellbaren Zahlen, d.h.

$$x,y\in\mathbb{G} \quad \underset{\text{i.A.}}{\not\Rightarrow} \quad x*y\in\mathbb{G}.$$

  Deshalb besteht die Notwendigkeit der **Rundung** $\mathrm{rd}:\mathbb{R}\to\mathbb{G}$, d.h. der Zuordnung einer reellen Zahl $x\in\mathbb{R}$ dem linken bzw. rechten Nachbarn in $\mathbb{G}$

$$g_L=\max\{g\in\mathbb{G}:g\le x\}\le x\le g_R=\min\{g\in\mathbb{G}:g\ge x\}.$$

  Insbesondere gilt $g_L=x=g_R$ für $x\in\mathbb{G}$.

  Korrektes Runden, gerichtetes Runden (Definition 2.3): Folgende Arten der Rundung sind gebräuchlich

$$\text{Korrektes Runden}: \quad \mathrm{rd}_*:\mathbb{R}\to\mathbb{G}:x\mapsto\begin{cases} g_L & \text{falls } x<\frac{g_L+g_R}{2},\\ g_L \text{ oder } g_R & \text{falls } x=\frac{g_L+g_R}{2},\\ g_R & \text{falls } x>\frac{g_L+g_R}{2},\end{cases}$$

$$\text{Gerichtetes Runden}: \quad \begin{cases} \text{Abrunden}: & \mathrm{rd}_-:\mathbb{R}\to\mathbb{G}:x\mapsto g_L,\\ \text{Aufrunden}: & \mathrm{rd}_+:\mathbb{R}\to\mathbb{G}:x\mapsto g_R,\\ \text{Abhacken}: & \mathrm{rd}_0:\mathbb{R}\to\mathbb{G}:x\mapsto\begin{cases} g_L & \text{falls } x\ge 0,\\ g_R & \text{falls } x<0.\end{cases}\end{cases}$$

- Man unterscheidet den absoluten Rundungsfehler $\mathrm{rd}(x)-x$ für $x\in\mathbb{R}$ und den **relativen Rundungsfehler**

$$\varepsilon_x=\frac{\mathrm{rd}(x)-x}{x}, \qquad 0\neq x\in\mathbb{R}.$$

Für $x = 0$ ist $\mathrm{rd}(x) = x$ und man setzt speziell $\varepsilon_x = 0$.

Maschinengenauigkeit (Satz 2.5): (Relative) Maschinengenauigkeit (wobei $\varrho = B^{1-t}$ Auflösung in $\mathbb{G}$ bzw. $\mathbb{M}$)

$$\varepsilon_{\mathrm{mach}} = \begin{cases} \frac{1}{2}\varrho & \text{für korrektes Runden}, \\ \varrho & \text{für gerichtetes Runden}. \end{cases}$$

Abschätzung des relativen Rundungsfehlers (Satz 2.4): Für $0 \neq x \in \mathbb{R}$ gilt $\varepsilon_x = \frac{\mathrm{rd}(x)-x}{x}$ mit $|\varepsilon_x| \leq \varepsilon_{\mathrm{mach}}$ und damit

$$\mathrm{rd}(x) = x\,(1+\varepsilon_x)\,, \qquad |\varepsilon_x| \leq \varepsilon_{\mathrm{mach}}\,.$$

Denn: Bei korrekter Rundung folgt für $g_L \leq x < \frac{g_L+g_R}{2}$ und damit $\mathrm{rd}(x) = g_L$ die Abschätzung $|\varepsilon_x| = \left|\frac{\mathrm{rd}(x)-x}{x}\right| \leq \frac{1}{2}\left|\frac{g_R-g_L}{g_L}\right| \leq \frac{1}{2}\varrho$. Ähnliche Überlegungen gelten für $\frac{g_L+g_R}{2} \leq x \leq g_R$ sowie gerichtetes Runden. ◇

- **Vorbemerkung**: Summenformel für **geometrische Reihe**

$$\sum_{i=n_0}^{n_1} q^i = \frac{q^{n_0} - q^{n_1+1}}{1-q}\,, \qquad \sum_{i=n_0}^{\infty} q^i = \frac{q^{n_0}}{1-q}\,, \quad |q| < 1\,.$$

Insbesondere gilt (wegen $B \geq 2$)

$$\sum_{i=n_0}^{\infty} B^{-i} = \frac{B^{-n_0}}{1-B^{-1}} = \frac{B^{1-n_0}}{B-1}\,.$$

Bemerkung: Korrekte Rundung einer Zahl $x \in \mathbb{R}$ auf t Stellen bei Kenntnis von $t+1$ Stellen. Mittels der Darstellung (mit Ziffern $0 \leq s_j \leq B-1$, wobei $s_\ell < B-1$ für ein $\ell \leq -2$)

$$x = \pm \sum_{j=-\infty}^{t-1} s_j \cdot B^{j+E} = \pm\, s \cdot B^E\,,$$

$$s = \sum_{j=-\infty}^{t-1} s_j \cdot B^j = \sum_{j=0}^{t-1} s_j \cdot B^j + s_{-1}\cdot B^{-1} + r\,, \qquad 0 \leq r = \sum_{j=-\infty}^{-2} s_j \cdot B^j < (B-1)\sum_{i=2}^{\infty} B^{-i} = B^{-1}\,,$$

ergibt sich für die mathematisch korrekte Rundung

$$\mathrm{rd}(x) = \pm\, B^E \begin{cases} \sum_{j=0}^{t-1} s_j \cdot B^j\,, & \text{falls } s_{-1} < \frac{1}{2}B\,, \\ \sum_{j=0}^{t-1} s_j \cdot B^j + 1\,, & \text{falls } s_{-1} \geq \frac{1}{2}B\,. \end{cases}$$

Beachte, daß die Zifferndarstellung durch die obige Forderung eindeutig ist (beispielsweise identifiziert man im Dezimalsystem $4.999\ldots = 5$). Falls $s_{-1} \geq \frac{1}{2}B$ und $s_0 = B-1$ kommt es zu Überlauf.

Vgl. **Illustration2_Rundung**.

2.2. Gleitpunktoperationen

- Die Ergebnisse der elementaren arithmetischen Operationen $* \in \{+, -, \times, /\}$ liegen im Allgemeinen nicht mehr im Bereich der darstellbaren Zahlen, d.h. für $a, b \in \mathbb{G}$ wird statt des exakten Ergebnisses

$$a * b \underset{\text{i.A.}}{\notin} \mathbb{G}$$

eine Näherungslösung

$$a \dot{*} b \in \mathbb{G}$$

berechnet.

Ideale Arithmetik: Die berechnete Näherungslösung $a \dot{*} b$ ergibt sich durch Rundung des exakten Ergebnisses, d.h. es gilt

$$a \dot{*} b = \mathrm{rd}(a * b) \in \mathbb{G}.$$

Vereinbarung: Entsprechend dem Standard ANSI/IEEE-Std-754-1985 für Gleitpunktarithmetik wird von nun an angenommen, daß die ideale Arithmetik in allen Genauigkeitsstufen für alle elementaren arithmetischen Operationen sowie die Berechnung der Wurzel und alle Rundungsarten gilt.

- **Rundungsfehlerschranken** (Satz 2.5): In idealer Arithmetik gilt

$$a \dot{+} b = \mathrm{rd}(a+b) = (a+b)(1+\varepsilon_1), \qquad a \dot{-} b = \mathrm{rd}(a-b) = (a-b)(1+\varepsilon_2),$$
$$a \dot{\times} b = \mathrm{rd}(a\,b) = a\,b(1+\varepsilon_3), \qquad a \dot{/} b = \mathrm{rd}\left(\tfrac{a}{b}\right) = \tfrac{a}{b}(1+\varepsilon_4),$$

mit relativen Rundungsfehlern $\varepsilon_i = \varepsilon_i(a,b)$ und $|\varepsilon_i| \le \varepsilon_{\text{mach}}$.

Beispiel: Vgl. **Illustration2_Rundung** ($B = 10$, $t = 4$, $\varepsilon_{\text{mach}} = \frac{1}{2} \cdot 10^{1-t}$ bei korrekter Rundung).

- **Beispiel** (Rundungsfehler bei Addition dreier Zahlen, Vorwärtsanalyse):
  - Aufgabe: Berechnung von

$$x = a + b + c = (a+b) + c = a + (b+c)$$

unter dem Einfluß von Rundungsfehlern mittels Satz 2.5.

  - Bestimmung von $(a+b)+c$ unter dem Einfluß von Rundungsfehlern (wobei $|\varepsilon_1|, |\varepsilon_2| \le \varepsilon_{\text{mach}}$)

$$\begin{aligned}
\tilde{x} &= \mathrm{rd}\big(\mathrm{rd}(a+b)+c\big) \\
&= \mathrm{rd}\big((a+b)(1+\varepsilon_1)+c\big) \\
&= \big((a+b)(1+\varepsilon_1)+c\big)(1+\varepsilon_2) \\
&= \big(a+b+c+(a+b)\varepsilon_1\big)(1+\varepsilon_2) \\
&= x\left(1+\tfrac{a+b}{a+b+c}\,\varepsilon_1(1+\varepsilon_2)+\varepsilon_2\right).
\end{aligned}$$

Als relativer Fehler ergibt sich

$$\varepsilon_{\widetilde{x}} = \tfrac{\widetilde{x}-x}{x} = \tfrac{a+b}{a+b+c}\,\varepsilon_1(1+\varepsilon_2)+\varepsilon_2\,.$$

– Bestimmung von $a+(b+c)$ unter dem Einfluß von Rundungsfehlern (wobei $|\varepsilon_3|, |\varepsilon_4| \le \varepsilon_{\text{mach}}$)

$$\begin{aligned}\widehat{x} &= \mathrm{rd}\big(a+\mathrm{rd}(b+c)\big)\\ &= \mathrm{rd}\big(a+(b+c)(1+\varepsilon_3)\big)\\ &= \big(a+(b+c)(1+\varepsilon_3)\big)(1+\varepsilon_4)\\ &= \big(a+b+c+(b+c)\varepsilon_3\big)(1+\varepsilon_4)\\ &= x\big(1+\tfrac{b+c}{a+b+c}\,\varepsilon_3(1+\varepsilon_4)+\varepsilon_4\big)\,.\end{aligned}$$

Als relativer Fehler ergibt sich

$$\varepsilon_{\widehat{x}} = \tfrac{\widehat{x}-x}{x} = \tfrac{b+c}{a+b+c}\,\varepsilon_3(1+\varepsilon_4)+\varepsilon_4\,.$$

– **Schlußfolgerung**: Der relativer Fehler des Ergebnisses bei der Addition von drei Zahlen hängt i.A. von der gewählten Reihenfolge der Operationen ab, d.h. Assoziativgesetz und Distributivgesetz gelten im Allgemeinen nicht mehr. Im Gegensatz zur Addition von zwei Zahlen kann der Fall eintreten, daß der relative Fehler nicht beschränkt ist.

– **Zahlenbeispiel**, vgl. **Illustration2_Rundung**.

Vgl. **Illustration2_Rundung**: Günstige Wahl der Reihenfolge bei der Berechnung von (wobei $n \gg 1$)

$$\sum_{i=1}^{n} \frac{1}{i}\,, \qquad \sum_{i=1}^{n} \frac{(-1)^i}{i}\,.$$

2.4. Kondition

- **Situation**: Lösung eines numerischen Problems $p : D \subset \mathbb{R}^n \to \mathbb{R}^m$, d.h. Berechnung von
$$y = p(x)\,.$$
Annahme p regulär (hinreichend oft differenzierbar).

- **Fragestellung**: Beurteilung der Sinnhaftigkeit der numerischen Lösung eines Problems, d.h. der Berechenbarkeit des Ergebnisses. Untersuchung der Auswirkungen kleiner Änderungen der Eingabedaten auf die Ergebnisse bei Anwendung von $p : D \subset \mathbb{R}^n \to \mathbb{R}^m$

Eingabe:	$x+\xi$	statt	x	mit relativem Fehler	$\frac{\xi}{x}$,
Ergebnis:	$p(x+\xi) = y+\eta$	statt	$y = p(x)$	mit relativem Fehler	$\frac{\eta}{y}$,

d.h. für $\xi \in \mathbb{R}^n$ (*klein* und jedenfalls so gewählt, daß $x+\xi \in D$) bestimme $\eta \in \mathbb{R}^m$ mit
$$\eta = p(x+\xi) - y = p(x+\xi) - p(x)\,.$$
Beispielsweise aufgrund der Darstellung der Eingabedaten als Maschinenzahlen sind solche kleinen Änderungen unvermeidbar.

- **Vorbemerkungen**:
  - Betrachtung von kleinen Änderungen (komponentenweise Abschätzung mit Inkrement $\Delta x \in \mathbb{R}^n$, wobei die Relation $\leq$ komponentenweise zu verstehen ist, oder Abschätzung bzgl. einer Norm mit $\Delta x \in \mathbb{R}$)
$$|\xi| \leq \Delta x \qquad \text{oder} \qquad \|\xi\| \leq \Delta x\,.$$
  - Mittels Taylorreihenentwicklung erhält man
$$\eta = p(x+\xi) - p(x) = p'(x)\,\xi + \mathcal{O}\big(\|\xi\|^2\big)\,.$$
In Komponenten
$$\begin{pmatrix}\eta_1\\ \vdots\\ \eta_m\end{pmatrix} = \begin{pmatrix}\partial_{x_1} p_1(x) & \dots & \partial_{x_n} p_1(x)\\ \vdots & & \vdots\\ \partial_{x_1} p_m(x) & \dots & \partial_{x_n} p_m(x)\end{pmatrix}\begin{pmatrix}\xi_1\\ \vdots\\ \xi_n\end{pmatrix} + \mathcal{O}\big(\|\xi\|^2\big)\,.$$
Für ξ *klein genug* folgt
$$\eta_i \approx \big(p'(x)\,\xi\big)_i = \sum_{j=1}^{n} \partial_{x_j} p_i(x)\,\xi_j\,, \qquad 1 \leq i \leq m\,,$$
und damit für den relativen Fehler
$$\frac{\eta_i}{y_i} \approx \sum_{j=1}^{n} \partial_{x_j} p_i(x)\,\frac{\xi_j}{y_i} = \sum_{j=1}^{n} \frac{x_j}{y_i}\,\partial_{x_j} p_i(x)\,\frac{\xi_j}{x_j}\,.$$

- **Kondition, Konditionszahlen** (Definition 2.8): Unter der Kondition eines numerischen Problems versteht man die Sensitivität des Ergebnisses $y+\eta = p(x+\xi)$ gegenüber *kleinen* Änderungen der Eingangsdaten. Falls *kleine* Änderungen ξ *kleine* Änderungen η im Ergebnis bewirken, spricht man von einem *gut* konditionierten Problem, falls hingegen *kleine* Änderungen ξ *große* Änderungen η im Ergebnis bewirken, spricht man von einem *schlecht* konditionierten Problem. Die Größen (unabhängig von ξ, η)

$$\left|\partial_{x_j} p_i(x)\right| \qquad \text{bzw.} \qquad \left|\frac{x_j}{y_i}\,\partial_{x_j} p_i(x)\right|,$$

bezeichnet man als absolute bzw. relative Konditionszahlen.

Bemerkung: Anstelle der absoluten Konditionszahlen wird auch oft eine Operatornorm $\|p'(x)\|$ betrachtet.

- **Beispiele** (Kondition):
  - **Relative Konditionszahlen elementarer Operationen**: Einsetzen in obige Relation speziell für

$$p : \mathbb{R}^2 \to \mathbb{R} : (a,b) \mapsto y = p(a,b)$$

führt auf (wobei $\xi = (\xi_a, \xi_b)^T$)

$$\frac{\eta}{y} = \frac{a}{y}\,\partial_a p(a,b)\,\frac{\xi_a}{a} + \frac{b}{y}\,\partial_b p(a,b)\,\frac{\xi_b}{b} + \mathcal{O}\left(\xi_a^2 + \xi_b^2\right).$$

    * Addition:

$$y = p(a,b) = a+b\,, \qquad \partial_a p(a,b) = \partial_b p(a,b) = 1\,, \qquad \frac{\eta}{y} = \frac{a}{a+b}\,\frac{\xi_a}{a} + \frac{b}{a+b}\,\frac{\xi_b}{b}\,.$$

      Schlecht konditioniertes Probleme für $a+b \approx 0$ bzw. $b \approx -a$ (Annahme $|a|, |b| > 0$), da dann die relativen Konditionszahlen sehr groß sind, d.h. Verstärkung relativer Eingabefehler.
      Phänomen der Auslöschung signifikanter Stellen, vgl. **Illustration2_Kondition**.
      Bemerkung: Gleichheit gilt aufgrund der Linearität der Funktion.

    * Subtraktion:

$$y = p(a,b) = a-b\,, \qquad \partial_a p(a,b) = 1\,, \quad \partial_b p(a,b) = -1\,, \qquad \frac{\eta}{y} = \frac{a}{a-b}\,\frac{\xi_a}{a} + \frac{b}{a-b}\,\frac{\xi_b}{b}\,.$$

      Bemerkung: Zurückführen auf Addition.

    * Multiplikation, Division:

$$y = p(a,b) = a\,b\,, \qquad \partial_a p(a,b) = b\,, \quad \partial_b p(a,b) = a\,, \qquad \frac{\eta}{y} \approx \frac{\xi_a}{a} + \frac{\xi_b}{b}\,,$$

$$y = p(a,b) = \frac{a}{b}\,, \qquad \partial_a p(a,b) = \frac{1}{b}\,, \quad \partial_b p(a,b) = -\frac{a}{b^2}\,, \qquad \frac{\eta}{y} \approx \frac{\xi_a}{a} - \frac{\xi_b}{b}\,.$$

      Gut konditionierte Probleme, da die relativen Konditionszahlen durch 1 beschränkt sind, d.h. keine Verstärkung relativer Eingabefehler (für die lineare Näherung).

* Wurzelziehen:

$$y = p(a) = \sqrt{a}, \qquad \partial_a p(a) = \frac{1}{2\sqrt{a}}, \qquad \frac{\eta}{y} \approx \frac{1}{2}\frac{\xi_a}{a}.$$

Gut konditioniertes Problem, da die relativen Konditionszahlen durch $\frac{1}{2}$ beschränkt sind, d.h. Verkleinerung relativer Eingabefehler (für lineare Näherung).

– **Relative Konditionszahlen der Formel von Vieta**: Bestimmung einer der beiden Lösungen $x_{1,2} = -a \pm \sqrt{a^2+b}$ der quadratischen Gleichung $x^2 + 2ax - b = 0$ (Annahme $|a|, |b| > 0$, Relationen $x_1 + x_2 = -2a$ und $x_1 x_2 = -b$)

$$y = p(a,b) = -a + \sqrt{a^2+b},$$
$$\partial_a p(a,b) = -1 + \frac{a}{\sqrt{a^2+b}} = -\frac{y}{\sqrt{a^2+b}}, \quad \partial_b p(a,b) = \frac{1}{2\sqrt{a^2+b}},$$
$$\frac{\eta}{y} \approx -\frac{a}{\sqrt{a^2+b}}\frac{\xi_a}{a} + \frac{b}{2y\sqrt{a^2+b}}\frac{\xi_b}{b} = -\frac{a}{\sqrt{a^2+b}}\frac{\xi_a}{a} + \frac{a+\sqrt{a^2+b}}{2\sqrt{a^2+b}}\frac{\xi_b}{b}.$$

Gut konditioniertes Problem beispielsweise für $a > 0$ und $b > 0$, da dann die relativen Konditionszahlen durch 1 beschränkt sind. Falls jedoch $a^2 + b \approx 0$ bzw. $b \approx -a^2$ ($x_1 \approx x_2$, sogenannter *schleifender Schnitt*) ist das Problem schlecht konditioniert. Daran ändert auch die Umformulierung des Problems

$$y = p(a,b) = \frac{b}{a+\sqrt{a^2+b}},$$

nichts (Kettenregel: partielle Ableitungen und damit Konditionszahlen gleich, vgl. **Illustation2_Kondition**).

Aber! (Vorbemerkung zur Stabilität von Algorithmen) Für $a^2 + b \approx a^2$ bzw. $b \approx 0$ ist das Problem gut konditioniert (z.B. $a \gg b > 0$ und $b \approx 0$, relative Konditionszahlen durch 1 beschränkt)

$$(a,b) \xrightarrow[\left|-\frac{a}{\sqrt{a^2+b}}\right|\le 1,\ \left|\frac{a+\sqrt{a^2+b}}{2\sqrt{a^2+b}}\right|\le 1]{p} -a+\sqrt{a^2+b}.$$

Berechnung von $y = -a + \sqrt{a^2+b}$ mit dem kanonischen Algorithmus

$$(a,b) \xrightarrow[1,1]{p^{(1)}} a^2 \xrightarrow[\left|\frac{a^2}{a^2+b}\right|\le 1,\ \left|\frac{b}{a^2+b}\right|\le 1]{p^{(2)}} a^2+b \xrightarrow[\frac{1}{2}]{p^{(3)}} \sqrt{a^2+b} \xrightarrow[\frac{-a}{-a+\sqrt{a^2+b}},\ \frac{\sqrt{a^2+b}}{-a+\sqrt{a^2+b}}]{p^{(4)}} -a+\sqrt{a^2+b}.$$

Die Teilprobleme $p^{(1)}, p^{(2)}, p^{(3)}$ sind in der vorliegenden Situation gut konditioniert, beim letzten Teilproblem $p^{(4)}$ kommt es jedoch zur Auslöschung signifikanter Stellen, d.h. der gewählte Algorithmus führt für ein gut konditioniertes

Problem auf ein unzufriedenstellendes Ergebnis (instabiler Algorithmus). Daher sollte man jedenfalls den (stabilen) Algorithmus

$$(a,b) \xrightarrow[1,1]{p^{(1)}} a^2 \xrightarrow[\left|\frac{a^2}{a^2+b}\right|\leq 1,\left|\frac{b}{a^2+b}\right|\leq 1]{p^{(2)}} a^2+b \xrightarrow[\frac{1}{2}]{p^{(3)}} \sqrt{a^2+b}$$

$$\xrightarrow[\left|\frac{-a}{a+\sqrt{a^2+b}}\right|\leq 1,\left|\frac{\sqrt{a^2+b}}{a+\sqrt{a^2+b}}\right|\leq 1]{p^{(4)}} a+\sqrt{a^2+b} \xrightarrow[1,1]{p^{(5)}} \frac{b}{a+\sqrt{a^2+b}}$$

mit in der vorliegenden Situation gut konditionierten Teilproblemen verwenden.

Vgl. **Illustation2_Kondition**.

Bemerkung: Eine etwas andere Situation liegt für den Fall $b \approx -a^2$ $(a,b \gg 0)$ vor. In dieser Situation ist das Problem schlecht konditioniert und damit dessen numerische Lösung nur eingeschränkt sinnvoll. Außerdem beinhalten beide Algorithmen das schlecht konditionierte Teilproblem $p^{(2)}$.

2.3. Rundungsfehleranalyse

- **Situation**: Anwendung eines Algorithmus $\left(p^{(1)},\ldots,p^{(k)}\right)$ zur Lösung eines numerischen Problems $p : D \subset \mathbb{R}^n \to \mathbb{R}^m$. Betrachtung eines direkten Verfahren, d.h. es gelte

$$p^{(k)} \circ \cdots \circ p^{(1)} = p\,.$$

- **Fragestellung**: Beurteilung der Güte eines Algorithmus. Untersuchung der Auswirkungen von kleinen Änderungen bzw. Rundungen der Eingabedaten sowie von bei der Lösung der Teilprobleme auftretenden Rundungen auf das Endergebnis.

- **Rundungsfehleranalyse**: Untersuchung der Fortpflanzung von Rundungsfehlern auf das Endergebnis.

  - **Vorwärtsanalyse**: Vergleich des unter dem Einfluß von Rundungsfehlern erhaltenen Endergebnisses mit dem exakten Ergebnis und Herleitung einer Relation bzw. Abschätzung für den relativen Fehler.

    Nachteil der Vorwärtsanalyse: Analyse bei komplexeren Algorithmen kompliziert.

  - **Rückwärtsanalyse**: Interpretation des unter dem Einfluß von Rundungsfehlern erhaltenen Endergebnisses als exaktes Ergebnis zu veränderten Eingabedaten. Keine Aussage über die tatsächliche Größe des relativen Rundungsfehlers im Ergebnis.

    Vorteil der Rückwärtsanalyse: Analyse auch bei komplexeren Algorithmen leichter durchführbar als die Vorwärtsanalyse.

- **Beispiele** (Rückwärtsanalyse):

  - **Elementare arithmetische Operationen**: Mit Hilfe von Satz 2.5 und dem Ansatz $\mathrm{rd}(a * b) = \widetilde{a} * \widetilde{b}$ ergibt sich

$$\mathrm{rd}(a * b) = (a * b)(1+\varepsilon) = \widetilde{a} * \widetilde{b}$$

    mit $|\varepsilon| \le \varepsilon_{\mathrm{mach}}$ (Annahme $1 - \varepsilon_{\mathrm{mach}} > 0$). Wähle beispielsweise für die Addition (Subtraktion analog)

$$\mathrm{rd}(a+b) = (a+b)(1+\varepsilon) = \widetilde{a} + \widetilde{b}\,,$$
$$\widetilde{a} = a(1+\varepsilon)\,, \quad \widetilde{b} = b(1+\varepsilon)\,, \qquad \left|\tfrac{\widetilde{a}-a}{a}\right| = \left|\tfrac{\widetilde{b}-b}{b}\right| = |\varepsilon| \le \varepsilon_{\mathrm{mach}}\,,$$

    für die Multiplikation

$$\mathrm{rd}(a\,b) = ab(1+\varepsilon) = \widetilde{a}\,\widetilde{b}\,,$$
$$\widetilde{a} = a\sqrt{1+\varepsilon}\,, \quad \widetilde{b} = b\sqrt{1+\varepsilon}\,, \qquad \left|\tfrac{\widetilde{a}-a}{a}\right| = \left|\tfrac{\widetilde{b}-b}{b}\right| = \left|\sqrt{1+\varepsilon}-1\right| = \tfrac{|\varepsilon|}{\sqrt{1+\varepsilon}+1} \le \varepsilon_{\mathrm{mach}}\,,$$

und für die Division

$$\begin{gathered}\mathrm{rd}\left(\frac{a}{b}\right) = \frac{a}{b}(1+\varepsilon) = \frac{\widetilde{a}}{\widetilde{b}},\\ \widetilde{a} = a\sqrt{1+\varepsilon}, \quad \widetilde{b} = \frac{b}{\sqrt{1+\varepsilon}},\\ \left|\frac{\widetilde{a}-a}{a}\right| \leq \varepsilon_{\mathrm{mach}}, \quad \left|\frac{\widetilde{b}-b}{b}\right| = \frac{|1-\sqrt{1+\varepsilon}|}{\sqrt{1+\varepsilon}} = \frac{|\varepsilon|}{\sqrt{1+\varepsilon}(1+\sqrt{1+\varepsilon})} \leq \frac{\varepsilon_{\mathrm{mach}}}{\sqrt{1-\varepsilon_{\mathrm{mach}}}}.\end{gathered}$$

Schlußfolgerung: Die bei der Anwendung der elementaren arithmetischen Operationen auftretenden Rundungsfehler verfälschen das exakte Ergebnis so wie relative Änderungen der Größenordnung $\varepsilon_{\mathrm{mach}}$ in den Eingangsdaten.

– **Horner-Schema** zur Berechnung des Funktionswertes einer Polynomfunktion:

* **Bernoullische Ungleichung**: Für $x \in \mathbb{R}$ mit $x \geq -1$ und $n \in \mathbb{N}_{\geq 0}$ gilt

$$1 + nx \leq (1+x)^n.$$

Denn: Mittels Induktion folgt

$$(1+x)^{n+1} = (1+x)(1+x)^n \geq (1+x)(1+nx) = 1 + (n+1)x + nx^2 \geq 1 + (n+1)x$$

und damit die Abschätzung. ◇

Abschätzung für Fehlerterme der Rundungsfehleranalyse (Lemma 2.6): Für eine Folge reeller Zahlen $(\varepsilon_i)_{1\leq i\leq n}$ mit $|\varepsilon_i| \leq \varepsilon_{\mathrm{mach}}$ gilt für $0 \leq k \leq n$

$$\left|\prod_{i=1}^{k}(1+\varepsilon_i)\prod_{j=k+1}^{n}\frac{1}{1+\varepsilon_j} - 1\right| \leq \frac{n\varepsilon_{\mathrm{mach}}}{1-n\varepsilon_{\mathrm{mach}}}.$$

Denn: Wegen $|\varepsilon_i| \leq \varepsilon_{\mathrm{mach}}$ ist $-\varepsilon_{\mathrm{mach}} \leq \varepsilon_i \leq \varepsilon_{\mathrm{mach}}$ sowie $-\varepsilon_{\mathrm{mach}} \leq -\varepsilon_i \leq \varepsilon_{\mathrm{mach}}$ für $1 \leq i \leq n$. Zusammen mit der Relation $(1+\varepsilon_i)(1-\varepsilon_i) = 1-\varepsilon_i^2 \leq 1$ folgt (Annahme $n\varepsilon_{\mathrm{mach}} < 1$ und insbesondere $\varepsilon_{\mathrm{mach}} < 1$)

$$1-\varepsilon_{\mathrm{mach}} \underset{-\varepsilon_{\mathrm{mach}}\leq\varepsilon_i}{\leq} 1+\varepsilon_i \underset{(1+\varepsilon_i)(1-\varepsilon_i)\leq 1}{\leq} \frac{1}{1-\varepsilon_i} \underset{-\varepsilon_{\mathrm{mach}}\leq-\varepsilon_i}{\leq} \frac{1}{1-\varepsilon_{\mathrm{mach}}}.$$

Mittels Bernoullischer Ungleichung ergibt sich weiters

$$\begin{aligned}1-\frac{n\varepsilon_{\mathrm{mach}}}{1-n\varepsilon_{\mathrm{mach}}} &\underset{n\varepsilon_{\mathrm{mach}}\leq\frac{n\varepsilon_{\mathrm{mach}}}{1-n\varepsilon_{\mathrm{mach}}}}{\leq} 1-n\varepsilon_{\mathrm{mach}} \underset{\text{Bernoulli}}{\leq} (1-\varepsilon_{\mathrm{mach}})^n\\ &\underset{\text{Relation}}{\leq} \prod_{i=1}^{k}(1+\varepsilon_i)\prod_{j=k+1}^{n}\frac{1}{1+\varepsilon_j} \underset{\text{Relation}}{\leq} \frac{1}{(1-\varepsilon_{\mathrm{mach}})^n}\\ &\underset{\text{Bernoulli}}{\leq} \frac{1}{1-n\varepsilon_{\mathrm{mach}}} = 1 + \frac{n\varepsilon_{\mathrm{mach}}}{1-n\varepsilon_{\mathrm{mach}}}\end{aligned}$$

und damit die Behauptung. ◇

* Für eine Polynomfunktion der Form

$$p(x) = \sum_{i=0}^{n} c_i x^i$$

beruht das Horner-Schema auf der Umformulierung

$$y = c_n x^n + c_{n-1} x^{n-1} + \cdots + c_0 = \Big(\cdots \big((c_n x + c_{n-1}) x + c_{n-2} \big) x + \cdots \Big) x + c_0 \,.$$

Mögliche Implementierung (Pseudo-Code)

$$\begin{aligned} & y = c_n \\ & \text{for i = n-1:-1:0} \\ & \qquad y = y\,x + c_i \\ & \text{end} \end{aligned}$$

Vgl. auch **Illustration1_Modifikation**.

* Unter dem Einfluß von Rundungsfehlern erhält man stattdessen

$$\begin{aligned} & \widetilde{y} = c_n \\ & \text{for i = n-1:-1:0} \\ & \qquad \widetilde{y} = \mathrm{rd}\big(\mathrm{rd}(\widetilde{y}\,x) + c_i\big) = \big(\widetilde{y}\,x\,(1+\varepsilon_i) + c_i\big)\,(1+\widetilde{\varepsilon}_i) \\ & \text{end} \end{aligned}$$

wobei $|\varepsilon_i|, |\widetilde{\varepsilon}_i| \le \varepsilon_{\mathrm{mach}}$. Das berechnete Ergebnis läßt sich in der Form

$$\widetilde{y} = \widetilde{c}_n x^n + \widetilde{c}_{n-1} x^{n-1} + \cdots + \widetilde{c}_0$$

mit Koeffizienten (Induktion)

$$\widetilde{c}_0 = c_0\,(1+\widetilde{\varepsilon}_0)\,, \qquad \widetilde{c}_n = c_n \prod_{\ell=0}^{n-1} (1+\varepsilon_\ell)\,(1+\widetilde{\varepsilon}_\ell)\,,$$

$$\widetilde{c}_i = c_i\,(1+\widetilde{\varepsilon}_i) \prod_{\ell=0}^{i-1} (1+\varepsilon_\ell)(1+\widetilde{\varepsilon}_\ell)\,, \quad 1 \le i \le n-1\,,$$

darstellen. Mittels Lemma 2.6 folgt außerdem (wegen $\frac{k\varepsilon_{\mathrm{mach}}}{1-k\varepsilon_{\mathrm{mach}}} \le \frac{n\varepsilon_{\mathrm{mach}}}{1-n\varepsilon_{\mathrm{mach}}}$ für $0 \le k \le n$)

$$\widetilde{c}_i = c_i\,(1+\delta_i)\,, \quad |\delta_i| \le \tfrac{n\varepsilon_{\mathrm{mach}}}{1-n\varepsilon_{\mathrm{mach}}}\,, \qquad 0 \le i \le n\,.$$

* **Schlußfolgerung**: Die bei der Anwendung des Horner-Schemas auftretenden Rundungsfehler verfälschen das exakte Ergebnis so wie relative Änderungen der Größenordnung $\frac{n\varepsilon_{\mathrm{mach}}}{1-n\varepsilon_{\mathrm{mach}}}$ in den Eingangdaten, d.h. den Koeffizienten des Polynoms.

- **Numerische Stabilität** eines Algorithmus: Ein Algorithmus heißt numerisch stabil bzw. gutartig im Sinne der Rückwärtsanalyse, wenn die Fortpflanzung von Rundungsfehlern zu Änderungen im Endergebnis führt, die in ihrer Größenordnung mit dem unvermeidlichen Fehler aufgrund von Ungenauigkeiten in den Eingangsdaten vergleichbar sind.

- **Spezialfall**: Betrachte speziell einen aus zwei Teilproblemen bestehenden Algorithmus zur Lösung eines Problems $p : \mathbb{R}^n \to \mathbb{R}^m$ (wobei $q : \mathbb{R}^n \to \mathbb{R}^k$, $r : \mathbb{R}^k \to \mathbb{R}^m$)

$$p(x) = \big(r \circ q\big)(x) = r\big(q(x)\big).$$

Die Anwendung der Kettenregel ergibt

$$p'(x) = r'\big(q(x)\big)\, q'(x),$$

$$p' = \begin{pmatrix} \partial_{x_1} p_1 & \dots & \partial_{x_n} p_1 \\ \vdots & & \vdots \\ \partial_{x_1} p_m & \dots & \partial_{x_n} p_m \end{pmatrix}, \quad r' = \begin{pmatrix} \partial_{x_1} r_1 & \dots & \partial_{x_k} r_1 \\ \vdots & & \vdots \\ \partial_{x_1} r_m & \dots & \partial_{x_k} r_m \end{pmatrix}, \quad q' = \begin{pmatrix} \partial_{x_1} q_1 & \dots & \partial_{x_n} q_1 \\ \vdots & & \vdots \\ \partial_{x_1} q_k & \dots & \partial_{x_n} q_k \end{pmatrix},$$

$$\partial_{x_j} p_i(x) = \sum_{\ell=1}^{k} \partial_{x_\ell} r_i\big(q(x)\big)\, \partial_{x_j} q_\ell(x).$$

Damit ergibt sich für die relativen Konditionszahlen

$$\frac{x_j}{y_i}\, \partial_{x_j} p_i(x) = \sum_{\ell=1}^{k} \frac{z_\ell}{y_i}\, \partial_{x_\ell} r_i(z)\, \frac{x_j}{z_\ell}\, \partial_{x_j} q_\ell(x), \qquad z = q(x).$$

Die relativen Konditionszahlen des Problems sind unabhängig von der gewählten Zerlegung, d.h. von der Wahl der Teilprobleme r und q.

Aber! Die Wahl der Teilprobleme wirkt sich auf die Güte des Algorithmus aus, insbesondere bestimmt die Größe der Konditionszahlen die Fortpflanzung von Fehlern und damit die Stabilität des Algorithmus. Untersuchung der Auswirkung kleiner Änderungen der Eingangsdaten auf das Ergebnis des Algorithmus, d.h. insbesondere Hinzunahme auftretender Änderungen (Rundungsfehler) bei der Berechnung des Zwischenergebnisses und des Endergebnisses

Eingabe :	$x + \xi$	statt	x,
Zwischenergebnis :	$q(x+\xi) + \zeta$	statt	$z = q(x)$,
Endergebnis :	$r\big(q(x+\xi)+\zeta\big) + \eta$	statt	$y = r\big(q(x)\big)$.

Mittels Taylorreihenentwicklung von q folgt

$$q(x+\xi) = z + q'(x)\,\xi + \mathcal{O}\big(\|\xi\|^2\big).$$

Eine Taylorreihenentwicklung von r ergibt weiters

$$\begin{aligned} r\big(q(x+\xi)+\zeta\big) &= r\Big(z + q'(x)\,\xi + \mathcal{O}\big(\|\xi\|^2\big) + \zeta\Big) \\ &= y + \mathbf{r'(z)\,q'(x)}\,\xi + \mathbf{r'(z)}\,\zeta + \mathcal{O}\big(\|\xi\|^2\big) + \mathcal{O}\big(\|\zeta\|^2\big). \end{aligned}$$

Falls das Problem gut konditioniert ist, ist

$$\left\| p'(x) \right\| = \left\| r'(z)\, q'(x) \right\|$$

von moderater Größenordnung und damit $\mathbf{r'(z)\, q'(x)}\, \xi$ eine kleine Änderung des exakten Endergebnisses. Ist der Algorithmus jedoch so gewählt, daß das erste Teilproblem sehr gut und das zweite sehr schlecht konditioniert sind (d.h. $\|r'(z)\| >> \|q'(z)\|$) werden unvermeidbare Fehler ζ im Zwischenergebnis (Rundung) sehr verstärkt, d.h. der Beitrag $\mathbf{r'(z)}\, \zeta$ führt auf ein unzufriedenstellendes Endergebnis und somit ist der Algorithmus instabil.

- **Beispiele**:
  - **Formel von Vieta**: Siehe obige Überlegungen. Das Problem ist für $a >> b$ gut konditioniert. Auswirkung von kleinen relativen Änderungen der Eingangsdaten und relativen Rundungsfehlern bei der Anwendung der Formel von Vieta

$$\begin{aligned}
(a,b)\,,(\varepsilon_a,\varepsilon_b) \quad &\xrightarrow[\mathbf{1,1}]{p^{(1)}} \quad a^2,(\varepsilon_a,\varepsilon_1) \\
&\xrightarrow[\left|\frac{\mathbf{a^2}}{\mathbf{a^2+b}}\right|\leq\mathbf{1},\left|\frac{\mathbf{b}}{\mathbf{a^2+b}}\right|\leq\mathbf{1}]{p^{(2)}} \quad a^2+b,(\varepsilon_a,\varepsilon_b,\varepsilon_1,\varepsilon_2) \\
&\xrightarrow[\frac{\mathbf{1}}{\mathbf{2}}]{p^{(3)}} \quad \sqrt{a^2+b},(\varepsilon_a,\varepsilon_b,\varepsilon_1,\varepsilon_2,\varepsilon_3) \\
&\xrightarrow[\frac{\mathbf{-a}}{\mathbf{-a+\sqrt{a^2+b}}},\frac{\mathbf{\sqrt{a^2+b}}}{\mathbf{-a+\sqrt{a^2+b}}}]{p^{(4)}} \quad -a+\sqrt{a^2+b},(\varepsilon_a,\varepsilon_b,\varepsilon_1,\varepsilon_2,\varepsilon_3,\varepsilon_4)\,.
\end{aligned}$$

    Der Algorithmus ist instabil, da der Verstärkungsfaktor der relativen Rundungsfehler im letzten Teilproblem sehr groß ist. Z.B. für $a = 1000$, $b = 0.018000000081$

$$\frac{\sqrt{a^2+b}}{-a+\sqrt{a^2+b}} \approx 10^8\,.$$

    Vgl. **Illustration2_Kondition**.
  - **Eigenwertberechnung**: Das Problem *Berechnung der Eigenwerte einer symmetrischen Matrix* ist sehr gut konditioniert. Der Algorithmus basierend auf der Nullstellenberechnung des charakteristischen Polynoms ist instabil, da das Teilproblem *Berechnung der Nullstellen eines Polynoms* höherer Ordnung sehr schlecht konditioniert ist.

    Siehe **Illustration1**.

- **Schlußfolgerungen**:
  - Die numerische Lösung eines schlecht konditionierten Problems ist nicht oder nur eingeschränkt sinnvoll.

– Da ein einzelnes schlecht konditioniertes Teilproblem das Endergebnis stark verfälschen kann, ist bei der numerischen Lösung eines gut konditionierten Problems darauf zu achten, daß der verwendete Algorithmus gut konditionierte Teilprobleme umfaßt.

2.5. Stabilität

- **Situation**: Anwendung eines Algorithmus $\left(p^{(1)},\ldots,p^{(k)}\right)$ zur Lösung eines numerischen Problems $p : D \subset \mathbb{R}^n \to \mathbb{R}^m$

$$p^{(k)} \circ \cdots \circ p^{(1)} \approx p\,.$$

  Numerische Stabilität eines Algorithmus (im Sinne der Rückwärtanalyse): Änderungen im Endergebnis aufgrund von Rundungsfehlern und Verfahrensfehlern vergleichbar mit *kleinen* Änderungen der Eingabedaten.

  Nun: Präzisierung des Begriffes der numerische Stabilität.

- Betrachtung einer Umgebung $U = U_{x,\varepsilon}$ der exakten Eingabedaten (komponentenweise, bzgl. Norm)

$$U = \left\{x+\xi \in D : |\xi| \le \varepsilon\,|x|\right\} \subset \mathbb{R}^n \qquad \text{oder} \qquad U = \left\{x+\xi \in D : \|\xi\| \le \varepsilon\,\|x\|\right\} \subset \mathbb{R}^n\,.$$

  Elemente der zugehörigen Umgebung des exakten Ergebnisses (Bildmenge)

$$p(U) = \left\{p(x+\xi) : x+\xi \in U\right\} \subset \mathbb{R}^m$$

  (oder nahe bei $p(U)$ liegende Elemente) werden als Näherungslösungen akzeptiert.

  Akzeptable Näherungslösung (Definition 2.9): Eine anstelle des exakten Ergebnisses $y = p(x)$ berechnete Näherungslösung $\widetilde{y} \approx y$ heißt akzeptabel bezüglich der Eingabemenge U, wenn

$$\widetilde{y} \in p(U)$$

  oder die abgeschwächten Bedingungen (komponentenweise, bzgl. Norm)

$$\exists\ z \in p(U) \quad \text{sodaß} \quad |\widetilde{y}-z| \le \mathcal{O}(\varepsilon_{\text{mach}})\,|\widetilde{y}| \quad \text{bzw.} \quad \|\widetilde{y}-z\| \le \mathcal{O}(\varepsilon_{\text{mach}})\,\|\widetilde{y}\|$$

  erfüllt sind.

  Numerische Stabilität eines Algorithmus (Definition 2.11): Ein Algorithmus zur Lösung eines Problems $p : D \subset \mathbb{R}^n \to \mathbb{R}^m$ ist numerisch stabil (im Sinne der Rückwärtsanalyse), wenn für alle zulässigen Eingabedaten $x \in D$ die unter dem Einfluß von Rundungsfehlern und Verfahrensfehlern berechneten Näherungslösungen $\widetilde{y} \approx y = p(x)$ akzeptabel bezüglich der Eingabemenge $U = U_{x,\varepsilon_{\text{mach}}}$ sind.

  Bemerkungen:

  - Die elementaren arithmetischen Operationen und das Wurzelziehen sind numerisch stabil.
  - Kondition eines numerischen Problems: Signifikanz der berechneten Ergebnisse.

    Numerische Stabilität eines Algorithmus (im Sinne der Rückwärtsanalyse): Gutartigkeit eines Verfahrens hinsichtlich der Auswirkung von Rundungsfehlern.

    Je schlechter die Kondition des numerischen Problems ist, desto größere Fehler im Endresultat sind akzeptabel.

- **Lösung eines linearen Gleichungssystems** $Ax = b$:
  - Numerisches Problem (unter der vereinfachenden Annahme, daß das lineare Gleichungssystem eindeutig lösbar ist, d.h. A invertierbar)

$$p : D \subset \mathbb{R}^{n\times n} \times \mathbb{R}^n \to \mathbb{R}^n : (A, b) \mapsto x = A^{-1} b .$$

  - Eingabemenge

$$U = U_{(A,b),\varepsilon} = \left\{(A+\alpha, b+\beta) \in D : |\alpha| \le \varepsilon\,|A|\,, |\beta| \le \varepsilon\,|b|\right\} \subset \mathbb{R}^{n\times n} \times \mathbb{R}^n$$

und zugehörige Bildmenge (Annahme $A+\alpha$ invertierbar, etwa $\|\alpha\| < \|A^{-1}\|^{-1}$, vgl. Satz 4.1)

$$p(U) = \left\{(A+\alpha)^{-1}(b+\beta) : (A+\alpha, b+\beta) \in U\right\}.$$

Akzeptable Näherungslösungen (bzgl. U, im strengen Sinn)

$$\widetilde{x} = (A+\alpha)^{-1}(b+\beta) \in p(U)$$

bzw. $\widetilde{x} = \widetilde{A}^{-1}\widetilde{b}$ mit $\widetilde{A} = A+\alpha$, $|\alpha| \le \varepsilon\,|A|$ und $\widetilde{b} = b+\beta$, $|\beta| \le \varepsilon\,|b|$ bzw.

$$\widetilde{A}\widetilde{x} = \widetilde{b} \qquad \text{mit} \quad |\widetilde{A} - A| \le \varepsilon\,|A| \quad \text{und} \quad |\widetilde{b} - b| \le \varepsilon\,|b|.$$

  - Resultat über die Größe des Residuums beim Einsetzen einer akzeptablen Näherungslösung in $Ax = b$.

Satz von Prager und Oettli (Satz 2.10): Eine Näherungslösung $\widetilde{x}$ des linearen Gleichungssystems $Ax = b$ ist akzeptabel (bzgl. U, im strengen Sinn), genau dann wenn das Residuum $r(\widetilde{x}) = b - A\widetilde{x}$ die folgende Abschätzung erfüllt

$$|A\widetilde{x} - b| \le \varepsilon\left(|A|\,|\widetilde{x}| + |b|\right).$$

Denn: Einerseits: Falls die Näherungslösung $\widetilde{x}$ akzeptabel ist, ergibt sich die Abschätzung (Bezeichnungen $\widetilde{A}$ und $\widetilde{b}$ wie zuvor, komponentenweise Relation $\le$)

$$\begin{aligned} |A\widetilde{x} - b| &\underset{\mp\widetilde{A}\widetilde{x} = -\widetilde{A}\widetilde{x}+\widetilde{b}}{=} \left|(A-\widetilde{A})\,\widetilde{x} + \widetilde{b} - b\right| = |\alpha\,\widetilde{x} + \beta| \le |\alpha|\,|\widetilde{x}| + |\beta| \\ &\le \varepsilon\left(|A|\,|\widetilde{x}| + |b|\right). \end{aligned}$$

Andererseits: Es ist zu zeigen, daß für Näherungslösungen $\widetilde{x}$, welche die Abschätzung $|A\widetilde{x} - b| \le \varepsilon\left(|A|\,|\widetilde{x}| + |b|\right)$ erfüllen, die Relation $\widetilde{A}\widetilde{x} = \widetilde{b}$ mit $|\widetilde{A} - A| \le \varepsilon\,|A|$ und $|\widetilde{b} - b| \le \varepsilon\,|b|$ folgt. Für eine Näherungslösung $\widetilde{x}$ definiere $\widetilde{A}$ und $\widetilde{b}$ durch

$$\widetilde{A} = \left(a_{ij} - \operatorname{sign}(\widetilde{x}_j)\,\varepsilon\,\vartheta_i\,|a_{ij}|\right)_{1\le i,j\le n}, \qquad \widetilde{b} = \left(b_i + \varepsilon\,\vartheta_i\,|b_i|\right)_{1\le i\le n},$$

$$r = A\widetilde{x} - b, \qquad z = \varepsilon\left(|A|\,|\widetilde{x}| + |b|\right), \qquad \vartheta_i = \begin{cases} \frac{r_i}{z_i} & \text{falls} \quad z_i \neq 0 \\ 0 & \text{falls} \quad z_i = 0 . \end{cases}$$

Eine kurze Rechnung zeigt ($\operatorname{sign}(x)\, x = |x|$)

$$\begin{aligned}\widetilde{A}\widetilde{x} - \widetilde{b} &= \Big(\sum_{j=1}^{n} \big(a_{ij} - \operatorname{sign}(\widetilde{x}_j)\, \varepsilon\, \vartheta_i\, |a_{ij}|\big)\, \widetilde{x}_j - \big(b_i + \varepsilon\, \vartheta_i\, |b_i|\big) \Big)_{1 \le i \le n} \\ &= A\widetilde{x} - b - \Big(\underbrace{\vartheta_i\, \varepsilon \sum_{j=1}^{n} \big(|a_{ij}|\, |\widetilde{x}_j| + |b_i|\big)}_{= z_i} \Big)_{1 \le i \le n} = A\widetilde{x} - b - r = 0\,.\end{aligned}$$

Da nach Voraussetzung die Näherungslösung $\widetilde{x}$ die Abschätzung $|r| \le z$ (komponentenweise) erfüllt, folgt außerdem $|\vartheta_i| \le 1$ für $1 \le i \le n$ und damit

$$\begin{aligned}\widetilde{A} - A = \big(-\operatorname{sign}(\widetilde{x}_j)\, \varepsilon\, \vartheta_i\, |a_{ij}| \big)_{1 \le i,j \le n}\,, \qquad \big|\widetilde{A} - A\big| \le \varepsilon\, |A|\,, \\ \widetilde{b} - b = \big(\varepsilon\, \vartheta_i\, |b_i| \big)_{1 \le i \le n}\,, \qquad \big|\widetilde{b} - b\big| \le \varepsilon\, |b|\,.\end{aligned}$$

Damit folgt die Behauptung. ◇

– **Bemerkungen**:

  * Für eine **genaue Näherungslösung** ist die Differenz zur exakten Lösung des linearen Gleichungssystems $|\widetilde{x} - x|$ *klein*.
  * Für eine **akzeptable Näherungslösung** ist das Residuum beim Einsetzen in das lineare Gleichungssystem $|A\widetilde{x} - b|$ *klein*.
  * **Zusammenfassung**: Ein Algorithmus zur Lösung eines linearen Gleichungssystems $Ax = b$ (mit $A \in \mathbb{K}^{n \times n}$ invertierbar) ist numerisch stabil (im Sinne der Rückwärtsanalyse), wenn für alle zulässigen Eingabedaten (A, b) die unter dem Einfluß von Rundungsfehlern und Verfahrensfehlern berechneten Näherungslösungen $\widetilde{x} \approx x = A^{-1}b$ akzeptabel bezüglich der Eingabemenge $U_{(A,b),\varepsilon_{\text{mach}}}$ sind, d.h. es gilt (mittels Satz von Prager und Oettli)

$$\widetilde{x} \in p\big(U_{(A,b),\varepsilon_{\text{mach}}}\big) \quad \Longleftrightarrow \quad |A\widetilde{x} - b| \le \varepsilon_{\text{mach}} \big(|A|\, |\widetilde{x}| + |b|\big)\,.$$

3. Vektoren und Matrizen

- **Inhalte**:
  - Grundlegende Begriffe und Resultate der Linearen Algebra, inbesondere zu Vektoren, Matrizen, Skalarprodukt, Orthogonalität, Norm.
  - QR-Zerlegung einer Matrix, Orthogonalisierungsverfahren nach Gram–Schmidt und Modifikationen.

3.1. Rechnen mit Vektoren und Matrizen

- Endlich dimensionaler **Vektorraum** bzw. **linearer Raum** $\mathbb{K}^n$ mit $\mathbb{K} = \mathbb{R}, \mathbb{C}$ (übliche Bezeichnungen für Komponenten, identifiziere $x \in \mathbb{K}^n$ mit Spalte $x \in \mathbb{K}^{n\times 1}$, Addition und Skalarmultiplikation)

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{K}^n, \qquad x + y = \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{pmatrix}, \quad \lambda x = \begin{pmatrix} \lambda x_1 \\ \vdots \\ \lambda x_n \end{pmatrix}, \quad x, y \in \mathbb{K}^n, \quad \lambda \in \mathbb{K}.$$

- Menge der komplexen bzw. reellen **Matrizen** $\mathbb{K}^{m\times n}$ (Anordnung in Schema mit m Zeilen und n Spalten, komponentenweise Addition und Skalarmultiplikation, **quadratische Matrix** für $m = n$)

$$A = (a_{ij})_{1\le i\le m, 1\le j\le n} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} \in \mathbb{K}^{m\times n}.$$

  Matrizenmultiplikation (assoziativ, nicht kommutativ)

$$C = AB = \Big(\sum_{k=1}^{m} a_{ik} b_{kj} \Big)_{1\le i\le \ell, 1\le j\le n} \in \mathbb{K}^{\ell\times n}, \quad A \in \mathbb{K}^{\ell\times m}, \quad B \in \mathbb{K}^{m\times n}$$
$$ABC = (AB)C = A(BC), \qquad AB \underset{\text{i.A.}}{\neq} BA.$$

- Für $A \in \mathbb{K}^{m\times n}$ ist die zugehörige **lineare Abbildung** (mit Eigenschaften Additivität und Homogenität) gegeben durch

$$f : \mathbb{K}^n \to \mathbb{K}^m : x \mapsto Ax.$$

  Bild der linearen Abbildung bzw. der zugehörigen Matrix

$$\mathscr{R}_A = f(\mathbb{K}^n) = \{ y = Ax \in \mathbb{K}^m : x \in \mathbb{K}^n \} \subset \mathbb{K}^m.$$

  Nullraum der linearen Abbildung bzw. der zugehörigen Matrix

$$\mathscr{N}_A = f^{-1}(0) = \{ x \in \mathbb{K}^n : Ax = 0 \in \mathbb{K}^m \} \subset \mathbb{K}^n.$$

- Läßt sich ein Vektor $w \in \mathbb{K}^n$ in der Form

$$w = \sum_{i=1}^{k} \lambda_i v_i$$

  mit $v_1, \dots, v_k \in \mathbb{K}^n$ und $\lambda_1, \dots, \lambda_k \in \mathbb{K}$ darstellen, heißt w eine **Linearkombination** von $v_1, \dots, v_k$. Die Menge aller Linearkombinationen von $v_1, \dots, v_k$ bezeichnet man als **lineare Hülle** der Vektoren $v_1, \dots, v_k$ bzw. den von $v_1, \dots, v_k$ **aufgespannten Raum**

$$\langle v_1, \dots, v_k \rangle = \Big\{ w = \sum_{i=1}^{k} \lambda_i v_i \in \mathbb{K}^n : \lambda_1, \dots, \lambda_k \in \mathbb{K} \Big\} \subset \mathbb{K}^n.$$

Läßt sich kein Vektor in $\{v_1,\dots,v_k\}$ als Linearkombination der restlichen Vektoren darstellen, d.h. es gilt

$$\sum_{i=1}^{k} \lambda_i v_i = 0 \quad \Longrightarrow \quad \lambda_i = 0, \quad 1 \le i \le k,$$

heißen $v_1,\dots,v_k$ **linear unabhängig**. Folglich ist auch die Darstellung von $w \in \langle v_1,\dots,v_k\rangle$ als Linearkombination der $v_1,\dots,v_k$ eindeutig.

Sind n Vektoren $v_1,\dots,v_n \in \mathbb{K}^n$ linear unabhängig, so bilden sie eine **Basis** des Vektorraumes $\mathbb{K}^n$, d.h. jeder Vektor in $\mathbb{K}^n$ läßt sich als Linearkombination der Basisvektoren darstellen (Erzeugendensystem) und die Darstellung ist eindeutig (lineare Unabhängigkeit).

Die **Standardbasisvektoren** bzw. **kanonischen Einheitsvektoren** $e_1,\dots,e_n \in \mathbb{K}^n$ (definiert durch $(e_i)_j = \delta_{ij}$ für $1 \le i,j \le n$) bilden eine Basis des Vektorraumes $\mathbb{K}^n$. Für $x \in \mathbb{K}^n$ folgt direkt die Darstellung als Linearkombination

$$x = \sum_{i=1}^{n} x_i e_i \in \mathbb{K}^n.$$

Veranschaulichung im $\mathbb{R}^2, \mathbb{R}^3$.

- **Grundlegender Zusammenhang zwischen linearen Gleichungssystemen und Darstellungen als Linearkombinationen**: Angabe und Umformulierung des Matrix-Vektor Produktes $Ax = b$ ergibt (wobei $A \in \mathbb{K}^{m\times n}$, $x \in \mathbb{K}^n$, $b \in \mathbb{K}^m$)

$$Ax = \Big(\sum_{k=1}^{n} a_{ik}x_k\Big)_{1\le i\le m} = \sum_{k=1}^{n} x_k \underbrace{\begin{pmatrix} a_{1k} \\ \vdots \\ a_{ik} \\ \vdots \\ a_{mk} \end{pmatrix}}_{k\text{−te Spalte } a_k \text{ von } A} = b,$$

d.h. die Lösung des linearen Gleichungssystems $Ax = b$ entspricht der Darstellung der rechten Seite b als Linearkombination der Spalten der Matrix A.

Schlußfolgerung: Das lineare Gleichungssystem $Ax = b$ ist genau dann lösbar, wenn sich die rechte Seite b als Linearkombination der Spalten der Matrix A darstellen läßt (somit ist $b \in \mathscr{R}_A$).

- Der **Rang** einer Matrix $A \in \mathbb{K}^{m\times n}$ ist definiert als die Anzahl der linear unabhängigen Spalten (bzw. Zeilen). Es gilt

$$\operatorname{rg}(A) = n - \dim \mathscr{N}_A.$$

Eine Matrix $A \in \mathbb{K}^{m\times n}$ hat **vollen Rang**, wenn $\operatorname{rg}(A) = \min\{m,n\}$.

Falls für eine Matrix $A \in \mathbb{K}^{m\times n}$ die Anzahl der Zeilen größer als die Anzahl der Spalten sind, d.h. es gilt $m \ge n$, sind folgende Aussagen äquivalent:

– Die Matrix $A \in \mathbb{K}^{m\times n}$ hat vollen Rang, d.h. es ist $\operatorname{rg}(A) = n$, d.h. alle Spalten der Matrix sind linear unabhängig.
– Die Abbildung $f : \mathbb{K}^n \to \mathbb{K}^m : x \mapsto Ax$ ist injektiv.
  Denn: Falls alle Spalten der Matrix linear unabhängig sind, folgt

$$Ax = Ay \iff A(x-y) = 0 \iff \sum_{k=1}^{n} (x_k - y_k)\, a_k = 0 \iff x = y$$

  und damit die Injektivität von f. ◇
– Der Nullraum von A ist gegeben durch $\mathcal{N}_A = \{0\} \subset \mathbb{K}^m$.
  Denn: Ähnlich wie zuvor folgt aus der linearen Unabhängigkeit der Spalten von A

$$Ax = 0 \iff \sum_{k=1}^{n} x_k\, a_k = 0 \iff x = 0$$

  und damit die Behauptung. ◇

Für **quadratische Matrizen** gelten inbesondere die folgenden äquivalenten Aussagen:

– Die Matrix $A \in \mathbb{K}^{n\times n}$ hat vollen Rang, d.h. es ist $\operatorname{rg}(A) = n$, d.h. alle Spalten der Matrix sind linear unabhängig.
  Die Spalten von A bilden somit eine Basis von $\mathbb{K}^n$, d.h. jeder Vektor in $\mathbb{K}^n$ läßt sich als Linearkombination der Basisvektoren darstellen und die Darstellung ist eindeutig.
– Die Abbildung $f : \mathbb{K}^n \to \mathbb{K}^n : x \mapsto Ax$ ist bijektiv.
  Die zugehörige inverse Abbildung $f^{-1} : \mathbb{K}^n \to \mathbb{K}^n : x \mapsto A^{-1}x$ ist durch die **inverse Matrix** (bzw. kurz **Inverse**) $A^{-1} \in \mathbb{K}^{n\times n}$ definiert, und es gilt

$$A A^{-1} = I\,.$$

– Der Nullraum von A ist gegeben durch $\mathcal{N}_A = \{0\} \subset \mathbb{K}^m$.

Die Inverse einer Matrix $A \in \mathbb{K}^{n\times n}$ erfüllt die Matrix-Gleichung

$$AX = I\,,$$
$$\left(Ax_1 \,\middle|\, \cdots \,\middle|\, Ax_n\right) = \left(e_1 \,\middle|\, \cdots \,\middle|\, e_n\right),$$
$$\begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} x_{11} & \dots & x_{1n} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{nn} \end{pmatrix} = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix},$$

d.h. die inverse Matrix A^{-1} ergibt sich als Lösung der n linearen Gleichungssysteme

$$Ax_i = e_i\,, \quad 1 \le i \le n\,, \qquad A^{-1} = \left(x_1 \,\middle|\, \cdots \,\middle|\, x_n\right).$$

Bemerkung: Für invertierbare Matrizen $A, B \in \mathbb{K}^{n\times n}$ gilt

$$(AB)^{-1} = B^{-1}A^{-1}\,.$$

- **Eindeutige Lösbarkeit eines linearen Gleichungssystems**: Falls die Matrix $A \in \mathbb{K}^{n \times n}$ invertierbar ist, bilden die Spalten von A eine Basis von $\mathbb{K}^n$. Folglich ist die Darstellung der rechten Seite b als Linearkombination der Spalten der Matrix A eindeutig und damit die Lösung des linearen Gleichungssystems $A\,x = b$ eindeutig bestimmt.
- Vgl. **Illustration3_VektorenMatrizen**.

3.2. Elementare Matrix-Multiplikationen

- **Vorbemerkung**: Die Lösung eines (eindeutig lösbaren) linearen Gleichungssystems $Ax = b$ (mittels Gaußschem Eliminationsverfahren) beruht auf der Transformation der Matrix A auf Dreiecksgestalt. Für theoretische Untersuchungen ist eine Beschreibung der Transformation mittels elementarer Matrix-Multiplikationen vorteilhaft.

- **Elementare Matrix-Umformungen**: Die Multiplikation einer Matrix A von links mit einer elementaren Transformationsmatrix T (d.h. Bildung von TA) wirkt auf die Zeilen der Matrix A.
  - **Skalierung** durch Multiplikation mit einer **Diagonalmatrix**

$$D = \operatorname{diag}(d_1, \ldots, d_n) = \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{pmatrix} \in \mathbb{K}^{n\times n}, \qquad d_i \neq 0, \quad 1 \leq i \leq n.$$

    Es gilt $D^{-1} = \operatorname{diag}\left(\frac{1}{d_1}, \ldots, \frac{1}{d_n}\right)$.
  - **Vertauschung von Zeilen** (oder Spalten) durch Multiplikation mit einer **Permutationsmatrix** (**Standardmatrix** E_{ij} mit Eintrag 1 bei (i,j)-tem Koeffizienten und ansonsten Einträgen 0)

$$P = I - E_{ii} - E_{jj} + E_{ij} + E_{ji} \in \mathbb{K}^{n\times n}, \qquad 1 \leq i, j \leq n, \quad i \neq j.$$

    Es gilt $P^{-1} = P = P^T$.
  - **Addition des skalaren Vielfachen einer Zeile** (oder Spalte) zu einer anderen Zeile (oder Spalte)

$$N_{ij}(\alpha) = I + \alpha E_{ij} \in \mathbb{K}^{n\times n}, \qquad 1 \leq i, j \leq n, \quad i \neq j, \qquad \alpha \in \mathbb{K}.$$

    Es gilt $N_{ij}(\alpha)^{-1} = N_{ij}(-\alpha)$ und beispielsweise für $n = 3$ (Reihenfolge der Matrizen wesentlich!)

$$N_{21}(\alpha_{21})\, N_{31}(\alpha_{31})\, N_{32}(\alpha_{32}) = \begin{pmatrix} 1 & & \\ \alpha_{21} & \ddots & \\ \alpha_{31} & \alpha_{32} & 1 \end{pmatrix} \in \mathbb{K}^{3\times 3}.$$

  Vgl. **Illustration3_VektorenMatrizen**.

3.3. Skalarprodukt und Orthogonalität

- Für eine komplexe Zahl $z = \Re z + \mathrm{i}\,\Im z \in \mathbb{C}$ ist die **komplex konjugierte Zahl** gegeben durch $\overline{z} = \Re z - \mathrm{i}\,\Im z \in \mathbb{C}$.

- Für eine reelle oder komplexe Matrix

$$A = (a_{ij})_{1\le i\le m, 1\le j\le n} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} \in \mathbb{K}^{m\times n}$$

  ist die **transponierte Matrix** (bzw. kurz **Transponierte**) gegeben durch

$$A^T = (a_{ji})_{1\le j\le n, 1\le i\le m} = \begin{pmatrix} a_{11} & \dots & a_{m1} \\ \vdots & & \vdots \\ a_{1n} & \dots & a_{mn} \end{pmatrix} \in \mathbb{K}^{n\times m}$$

  und die **adjungierte Matrix** (bzw. kurz **Adjungierte**) $A^* \in \mathbb{K}^{n\times m}$ gegeben durch

$$A^* = \overline{A}^T = (\overline{a_{ji}})_{1\le j\le n, 1\le i\le m} = \begin{pmatrix} \overline{a_{11}} & \dots & \overline{a_{m1}} \\ \vdots & & \vdots \\ \overline{a_{1n}} & \dots & \overline{a_{mn}} \end{pmatrix} \in \mathbb{K}^{n\times m}.$$

  Offensichtlich gilt (wobei $A, B \in \mathbb{K}^{m\times n}, C \in \mathbb{K}^{n\times k}$, für letzte Relation verwende Annahme $A \in \mathbb{K}^{n\times n}$ invertierbar und z.B. die Relation $I = (A^{-1}A)^* = A^*(A^{-1})^*$)

$$\begin{aligned} &(A+B)^T = A^T + B^T, \quad (\lambda A)^T = \lambda A^T, \quad (AC)^T = C^T A^T, \quad (A^T)^{-1} = (A^{-1})^T, \\ &(A+B)^* = A^* + B^*, \quad (\lambda A)^* = \overline{\lambda} A^*, \quad (AC)^* = C^* A^*, \quad (A^*)^{-1} = (A^{-1})^*, \end{aligned}$$

  und insbesondere $A^* = A^T$ für $A \in \mathbb{R}^{m\times n}$.

- Eine (quadratische) Matrix $A \in \mathbb{K}^{n\times n}$ heißt **symmetrisch**, wenn

$$A^T = A, \qquad a_{ji} = a_{ij}, \quad 1 \le i, j \le n.$$

  Eine (quadratische) Matrix $A \in \mathbb{K}^{n\times n}$ heißt **selbstadjungiert** bzw. **hermitesch**, wenn

$$A^* = A, \qquad \overline{a_{ji}} = a_{ij}, \quad 1 \le i, j \le n.$$

  Offensichtlich ist eine reelle selbstadjungierte Matrix insbesondere symmetrisch.

  Eine selbstadjungierte Matrix $A \in \mathbb{K}^{n\times n}$ heißt **positiv semi-definit** oder **positiv definit**, falls für alle $x \in \mathbb{K}^n$ die Bedingung

$$x^* A x \ge 0 \quad \text{oder} \quad x^* A x > 0$$

erfüllt ist. Beachte, daß die quadratische Form

$$q : \mathbb{K}^n \to \mathbb{R} : x \mapsto x^* A x,$$

$$q(x) = x^* A x = \begin{pmatrix} \overline{x_1} & \dots & \overline{x_n} \end{pmatrix} \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \overline{x_i} x_j,$$

aufgrund der geforderten Selbstadjungiertheit von A reelle Werte annimmt, denn es folgt (für $v, w \in \mathbb{K}^n$ ist $v^T w = w^T v$)

$$\overline{x^* A x} \underset{A^* = A}{=} \overline{x^* A^* x} = \overline{(Ax)^* x} \underset{\overline{(Ax)^*} = (Ax)^T}{=} (Ax)^T \overline{x} \underset{(Ax)^T \overline{x} = \overline{x}^T (Ax)}{=} x^* A x \quad \Longrightarrow \quad x^* A x \in \mathbb{R}.$$

- Das **euklidische Skalarprodukt** ist definiert durch (komplexe Konjugation bzgl. des zweiten Argumentes!)

$$\langle \cdot \,|\, \cdot \rangle_2 : \mathbb{K}^n \times \mathbb{K}^n \to \mathbb{K} : (x, y) \mapsto \langle x \,|\, y \rangle_2 = y^* x = \begin{pmatrix} \overline{y_1} & \dots & \overline{y_n} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \sum_{i=1}^{n} \overline{y_i} x_i .$$

  Die zugehörige **euklidische Norm** ist gegeben durch

$$\| \cdot \|_2 : \mathbb{K}^n \to \mathbb{R}_{\geq 0} : x \mapsto \|x\|_2 = \sqrt{\langle x \,|\, x \rangle_2} = \sqrt{\sum_{i=1}^{n} |x_i|^2}.$$

- Allgemeiner fordert man von einem **Skalarprodukt** bzw. **inneren Produkt** folgende Eigenschaften (**positiv definite hermitesche Sesquilinearform**, insbesondere für $\mathbb{K} = \mathbb{R}$ **positiv definite symmetrische Bilinearform**)

$$\langle \cdot \,|\, \cdot \rangle : \mathbb{K}^n \times \mathbb{K}^n \to \mathbb{K}$$
$$\langle x + \tilde{x} \,|\, y \rangle = \langle x \,|\, y \rangle + \langle \tilde{x} \,|\, y \rangle, \quad \langle \lambda x \,|\, y \rangle = \lambda \langle x \,|\, y \rangle, \qquad \lambda \in \mathbb{K}, \quad x, y \in \mathbb{K}^n,$$
$$\langle y \,|\, x \rangle = \overline{\langle x \,|\, y \rangle}, \qquad x, y \in \mathbb{K}^n,$$
$$\langle x \,|\, x \rangle \geq 0, \quad \langle x \,|\, x \rangle = 0 \Leftrightarrow x = 0, \qquad x \in \mathbb{K}^n.$$

  Die **zugehörige Norm**, definiert durch

$$\| \cdot \| : \mathbb{K}^n \to \mathbb{R}_{\geq 0} : x \mapsto \|x\| = \sqrt{\langle x \,|\, x \rangle},$$

  erfüllt die **Ungleichung von Cauchy–Schwarz**

$$|\langle x \,|\, y \rangle| \leq \|x\| \, \|y\|, \qquad x, y, \in \mathbb{K}^n.$$

  Denn: Insbesondere für den Spezialfall $y = 0$ ist die Behauptung klar. Unter der Annahme $y \neq 0$ und $\mathbb{K} = \mathbb{R}$ führt die Minimierung der quadratischen Funktion

$$f : \mathbb{K} \to \mathbb{R}_{\geq 0} : \lambda \mapsto \|x + \lambda y\|^2 = \langle x + \lambda y \,|\, x + \lambda y \rangle = \|x\|^2 + |\lambda|^2 \|y\|^2 + \left(\overline{\lambda} \langle x \,|\, y \rangle + \lambda \langle y \,|\, x \rangle \right)$$

wegen $f(\lambda) = \|x\|^2 + \lambda^2 \|y\|^2 + 2\lambda \langle x | y \rangle$ auf die Ungleichung

$$f'(\lambda_{\min}) = 2\lambda_{\min} \|y\|^2 + 2\langle x | y \rangle = 0, \quad \lambda_{\min} = -\tfrac{1}{\|y\|^2} \langle x | y \rangle, \qquad f''(\lambda_{\min}) = 2\|y\|^2 > 0,$$
$$0 \le f(\lambda_{\min}) = \|x\|^2 - \tfrac{1}{\|y\|^2} \langle x | y \rangle^2 \quad \Longrightarrow \quad |\langle x | y \rangle| \le \|x\| \|y\|.$$

Für $\mathbb{K} = \mathbb{C}$ ergibt sich die Behauptung direkt durch Einsetzen von $\lambda = -\frac{1}{\|y\|^2} \langle x | y \rangle$

$$0 \le f\big(-\tfrac{1}{\|y\|^2} \langle x | y \rangle\big) = \|x\|^2 - \tfrac{1}{\|y\|^2} |\langle x | y \rangle|^2$$

und Wurzelziehen. ◇

Allgemein fordert man für eine **Norm** $\|\cdot\| : \mathbb{K}^n \to \mathbb{R}_{\ge 0}$ die folgenden Eigenschaften (Positive Definitheit, Homogenität, Sub-Additivität bzw. Dreiecksungleichung)

$$\begin{aligned} \|x\| = 0 \;&\Leftrightarrow\; x = 0, \qquad x \in \mathbb{K}^n, \\ \|\lambda x\| = |\lambda| \|x\|, \qquad &\lambda \in \mathbb{K}, \quad x \in \mathbb{K}^n, \\ \|x + y\| \le \|x\| + \|y\|, \qquad &x, y \in \mathbb{K}^n. \end{aligned}$$

Bemerkung: Die durch ein Skalarprodukt definierte Norm erfüllt die Normeigenschaften. **Denn**: Die Eigenschaften positive Definitheit und Homogenität sind leicht nachzuweisen. Mittels der Ungleichung von Cauchy–Schwarz folgt

$$\begin{aligned} &\langle x | y \rangle + \langle y | x \rangle \le 2 \|x\| \|y\| \\ \Longrightarrow \quad &\langle x | x \rangle + \langle x | y \rangle + \langle y | x \rangle + \langle y | y \rangle \le \|x\|^2 + 2\|x\| \|y\| + \|y\|^2 \\ \Longrightarrow \quad &\|x + y\|^2 \le \|x\|^2 + 2\|x\| \|y\| + \|y\|^2 \\ \Longrightarrow \quad &\|x + y\| \le \|x\| + \|y\| \end{aligned}$$

und damit die Dreiecksungleichung. ◇

- Insbesondere im $\mathbb{R}^2$ ist der von zwei Vektoren $0 \ne x, y \in \mathbb{R}^2$ eingeschlossene **Winkel** gegeben durch
$$\cos \alpha = \frac{\langle x | y \rangle_2}{\|x\|_2 \|y\|_2}.$$
Speziell für $\|x\|_2 = 1$ und mit der Bezeichnung $r = \|y\|_2$ für die Länge von y ergibt sich
$$\langle x | y \rangle_2 = r \cos \alpha,$$
d.h. $\langle x | y \rangle_2$ beschreibt die Projektion von y auf x und $\langle x | y \rangle_2 \, x$ die Komponente von y in Richtung x (vgl. Abbildung, Skriptum, S. 42). Falls die beiden Vektoren einen rechten Winkel einschließen, d.h. $\alpha = \frac{\pi}{2}$, folgt $\langle x | y \rangle_2 = 0$.

Zwei Vektoren $x, y \in \mathbb{K}^n$ heißen **orthogonal**, wenn die Bedingung $\langle x | y \rangle = 0$ (bzw. speziell für das euklidische Skalarprodukt $\langle x | y \rangle_2 = y^* x = 0$) erfüllt ist.

Satz von Pythagoras: Für orthonormale Vektoren $z_1, \ldots, z_k \in \mathbb{K}^n$ und skalare $c_1, \ldots, c_k \in \mathbb{K}$ gilt die folgende Relation

$$\left\| \sum_{i=1}^{k} c_i z_i \right\|^2 = \left\langle \sum_{i=1}^{n} c_i z_i \,\middle|\, \sum_{j=1}^{n} c_j z_j \right\rangle = \sum_{i,j=1}^{n} c_i \overline{c_j} \underbrace{\langle z_i \,|\, z_j \rangle}_{=z_j^* z_i = \delta_{ij}} = \sum_{i=1}^{n} |c_i|^2 .$$

Eine Teilmenge $V = \{v_1, \ldots, v_k\} \subset \mathbb{K}^n \setminus \{0\}$ heißt **orthogonal**, wenn je zwei Elemente orthogonal sind, d.h. es gilt $\langle v_i \,|\, v_j \rangle = 0$ für $1 \le i, j \le k$ mit $i \neq j$. Nach **Satz 3.1** sind die Elemente einer orthogonalen Menge linear unabhängig, denn es gilt

$$\sum_{i=1}^{k} \lambda_i v_i = 0 \quad \Rightarrow \quad \sum_{i=1}^{k} \lambda_i \underbrace{\langle v_i \,|\, v_j \rangle}_{=\|v_j\|^2 \delta_{ij}} = 0, \quad 1 \le j \le k \quad \underset{\|v_j\| \neq 0}{\Longrightarrow} \quad \lambda_j = 0, \quad 1 \le j \le k .$$

Sind außerdem alle Vektoren in V normiert, d.h. $\|v_i\| = 1$ für $1 \le i \le k$, heißt die Menge **orthonormal**.

Die Elemente einer orthonormalen Menge $V = \{v_1, \ldots, v_n\} \subset \mathbb{K}^n$ bilden eine **Orthonormalbasis** von $\mathbb{K}^n$, vgl. **Satz 3.1**. Für einen Vektor $y \in \mathbb{K}^n$ gilt dann die folgende Darstellung bezüglich der Orthonormalbasis

$$y = \sum_{i=1}^{n} \langle y \,|\, v_i \rangle \, v_i .$$

Denn: Da die Vektoren in V eine Basis von $\mathbb{K}^n$ bilden, läßt sich $y \in \mathbb{K}^n$ als Linearkombination von $v_1, \ldots, v_n$ darstellen

$$y = \sum_{i=1}^{n} \lambda_i v_i .$$

Durch Bilden des Skalarproduktes mit den Basisvektoren folgt

$$\langle y \,|\, v_j \rangle = \sum_{i=1}^{n} \lambda_i \underbrace{\langle v_i \,|\, v_j \rangle}_{=\delta_{ij}} = \lambda_j, \quad 1 \le j \le n \quad \Longrightarrow \quad \lambda_j = \langle y \,|\, v_j \rangle, \quad 1 \le j \le n,$$

und damit die angegebene Darstellung. ◇

Zwei Teilmengen $X, Y \subset \mathbb{K}^n$ heissen **orthogonal** zueinander, wenn für jedes Element $x \in X$ und jedes Element $y \in Y$ die Bedingung $\langle x \,|\, y \rangle = 0$ gilt.

- Eine quadratische Matrix $Q \in \mathbb{K}^{n \times n}$, deren Spalten orthonormal sind, heißt eine **unitäre Matrix** (oder speziell **orthonormale Matrix** für $\mathbb{K} = \mathbb{R}$).

  Folglich gilt (Produkt der i-ten Zeile mit der j-ten Spalte von Q ergibt Kronecker-Delta δ_{ij}, d.h. Wert 1 falls $i = j$ und ansonsten Wert 0)

$$Q^* Q = I \quad \text{bzw.} \quad Q^{-1} = Q^* .$$

Die durch eine unitäre Matrix definierte lineare Abbildung $f : \mathbb{K}^n \to \mathbb{K}^n : x \mapsto Qx$ ist **längenerhaltend**, d.h. es gilt

$$\|Qx\|_2 = \|x\|_2$$

wegen $\|Qx\|_2^2 = x^* Q^* Qx = x^* x = \|x\|_2^2$.

- **Zusammenhang zwischen linearen Gleichungssystemen und Darstellungen als Linearkombinationen**: Die Darstellung eines Vektors $b \in \mathbb{K}^n$ als Linearkombination orthogonaler Vektoren $q_1, \ldots, q_k$ entspricht der Lösung eines linearen Gleichungssystems mit unitärer Matrix

$$Qx = b \quad \Longleftrightarrow \quad b = \sum_{i=1}^{n} x_i q_i \,, \qquad Q = \left(q_1 \,\middle|\, \cdots \,\middle|\, q_n\right).$$

Die Lösung erhält man in diesem Fall auf einfache Weise durch Mulitplikation mit der adjungierten Matrix

$$x = Q^* b = \left(\begin{array}{c} \overline{q_1} \\ \hline \vdots \\ \hline \overline{q_n} \end{array}\right) b = \begin{pmatrix} \overline{q_1} b \\ \vdots \\ \overline{q_n} b \end{pmatrix}.$$

3.4. Orthogonalisierungsverfahren nach Gram–Schmidt

- **Vorbemerkungen**:
  - Die Lösung eines linearen Gleichungssystems

$$Ax = b, \qquad A = \left(a_1 \middle| \cdots \middle| a_n\right) \in \mathbb{K}^{n\times n}, \quad x \in \mathbb{K}^n, \quad b \in \mathbb{K}^n,$$

    entspricht der Darstellung der rechten Seite b als Linearkombination der Spaltenvektoren $a_1, \ldots, a_n$.
  - Besonders einfach ist die Lösung eines linearen Gleichungssystems

$$Qy = c, \qquad Q = \left(q_1 \middle| \cdots \middle| q_n\right) \in \mathbb{K}^{n\times n}, \quad y \in \mathbb{K}^n, \quad c \in \mathbb{K}^n,$$

    mit unitärer Matrix Q, d.h. die Spaltenvektoren $q_1, \ldots, q_n$ bilden eine Orthonormalbasis des Vektorraumes $\mathbb{K}^n$. In diesem Fall ist das Gleichungssystem eindeutig lösbar und man erhält die Lösung durch Multiplikation mit der adjungierten Matrix

$$y = Q^* c, \qquad Q^* = \left(\begin{array}{c} \overline{q_1} \\ \hline \vdots \\ \hline \overline{q_n} \end{array}\right).$$

  - Formulierung in Hinblick auf die Lösung **überbestimmter linearer Gleichungssysteme**

$$Ax = b, \qquad A = \left(a_1 \middle| \cdots \middle| a_n\right) \in \mathbb{K}^{m\times n}, \quad x \in \mathbb{K}^n, \quad b \in \mathbb{K}^m, \qquad m \geq n.$$

    Zusätzliche Annahme, daß die Matrix A vollen Rang hat, d.h. die Spaltenvektoren $a_1, \ldots, a_n$ sind linear unabhängig.

- **Problemstellung**: Zu linear unbhängigen Vektoren $a_1, \ldots, a_n \in \mathbb{K}^m$ finde orthonormale Vektoren $q_1, \ldots, q_n \in \mathbb{K}^m$ so, daß die Vektoren $a_1, \ldots, a_k$ und $q_1, \ldots, q_k$ für $1 \leq k \leq n$ denselben Raum aufspannen, d.h. es gelte

$$\langle q_1, \ldots, q_k \rangle = \langle a_1, \ldots, a_k \rangle, \qquad 1 \leq k \leq n.$$

  Dies ist äquivalent zur Berechnung der **reduzierten QR-Zerlegung** der Matrix A

$$\begin{gathered} A = \widehat{Q}\,\widehat{R}, \\ A = \left(a_1 \middle| \cdots \middle| a_n\right) \in \mathbb{K}^{m\times n}, \\ \widehat{Q} = \left(q_1 \middle| \cdots \middle| q_n\right) \in \mathbb{K}^{m\times n}, \\ \widehat{R} = \begin{pmatrix} r_{11} & \ldots & r_{nn} \\ & \ddots & \vdots \\ & & r_{nn} \end{pmatrix} \in \mathbb{K}^{n\times n}, \qquad r_{ii} \neq 0, \quad 1 \leq i \leq n. \end{gathered}$$

Denn: Durch Bestimmung des Produktes $\widehat{Q}\,\widehat{R}$ erhält man wegen $r_{kj} = 0$ für $k \geq j+1$

$$A = \widehat{Q}\,\widehat{R}\,,$$

$$a_{ij} = \sum_{k=1}^{n} q_{ik} r_{kj} = \sum_{k=1}^{j} q_{ik} r_{kj}\,, \qquad 1 \leq i \leq m\,, \quad 1 \leq j \leq n\,,$$

$$a_j = \sum_{k=1}^{j} r_{kj}\, q_k\,, \qquad 1 \leq j \leq n\,,$$

$$a_1 = r_{11} q_1\,, \quad a_2 = r_{12} q_1 + r_{22} q_2\,, \quad \ldots\,, \quad a_n = r_{1n} q_1 + \cdots + r_{n-1,n} q_{n-1} + r_{nn} q_n\,.$$

Die zeigt, daß $a_1, \ldots, a_k \in \langle q_1, \ldots, q_k \rangle$ für $1 \leq k \leq n$. Andererseits ist wegen $r_{ii} \neq 0$ für $1 \leq i \leq n$ die Matrix $\widehat{R}$ invertierbar und deshalb $A = \widehat{Q}\,\widehat{R}$ äquivalent zu $\widehat{Q} = A\,\widehat{R}^{-1}$ (die Inverse von R wird nicht explizit berechnet, Einsetzen der bereits bestimmten Spalten von $\widehat{Q}$). Sukzessives Auflösen der obigen Relationen führt auf

$$q_j = \tfrac{1}{r_{jj}} \Big(a_j - \sum_{i=1}^{j-1} r_{ij}\, q_i \Big)\,, \qquad 1 \leq j \leq n\,,$$

$$q_1 = \tfrac{1}{r_{11}}\, a_1\,, \quad q_2 = \tfrac{1}{r_{22}}\, (a_2 - r_{12} q_1)\,, \quad \text{etc.}\,, \quad q_n = \tfrac{1}{r_{nn}}\, (a_n - r_{1n} q_1 - \cdots - r_{n-1,n} q_{n-1})\,,$$

was zeigt, daß auch $q_1, \ldots, q_k \in \langle a_1, \ldots, a_k \rangle$ für $1 \leq k \leq n$. ◇

Bemerkung: Als (volle) **QR-Zerlegung** einer Matrix $A \in \mathbb{K}^{m \times n}$ bezeichnet man die Darstellung

$$A = Q\,R\,,$$

$$A = \big(a_1 \,\big|\, \cdots \,\big|\, a_n \big) \in \mathbb{K}^{m \times n}\,,$$

$$Q = \big(\widehat{Q} \,\big|\, q_{n+1} \,\big|\, \cdots \,\big|\, q_m \big) \in \mathbb{K}^{m \times m}\,, \qquad R = \begin{pmatrix} \widehat{R} \\ \hline 0 \end{pmatrix} \in \mathbb{K}^{m \times n}\,,$$

wobei die Matrix $\widehat{Q}$ um $m-n$ orthonormale Spalten $q_{n+1}, \ldots, q_m$ zu einer unitären Matrix $Q \in \mathbb{K}^{m \times m}$ ergänzt wird und die Matrix $\widehat{R}$ um $m-n$ Nullzeilen zur Matrix $R \in \mathbb{K}^{m \times n}$ ergänzt wird.

- Das **Orthogonalisierungsverfahren nach Gram–Schmidt** ist ein Verfahren zur Berechnung der reduzierten QR-Zerlegung einer Matrix $A = \big(a_1 \,\big|\, \cdots \,\big|\, a_n \big) \in \mathbb{K}^{m \times n}$, d.h. orthonormaler Vektoren $q_1, \ldots, q_n \in \mathbb{K}^m$ mit

$$q_k = \tfrac{1}{r_{kk}} \Big(a_k - \sum_{i=1}^{k-1} r_{ik}\, q_i \Big)\,, \quad 1 \leq k \leq n\,,$$

wobei $r_{ij} \in \mathbb{K}$ für $1 \leq i, j \leq n$ mit $r_{ij} = 0$ für $i \geq j+1$ und $r_{ii} \neq 0$ für $1 \leq i \leq n$.

Bemerkung: Wähle im Folgenden $\langle \cdot \,|\, \cdot \rangle = \langle \cdot \,|\, \cdot \rangle_2$ und $\| \cdot \| = \| \cdot \|_2$.

Herleitung des Verfahrens:

  - Betrachte die erste Relation $q_1 = \frac{1}{r_{11}}\, a_1$ mit unbekanntem Koeffizienten r_{11}. Die Forderung $\| q_1 \| = 1$ impliziert $|r_{11}| = \| a_1 \|$. Setze etwa $r_{11} = \| a_1 \|$.

– Betrachte die zweite Relation $q_2 = \frac{1}{r_{22}}(a_2 - r_{12}q_1)$ mit unbekannten Koeffizienten r_{12} und r_{22}. Die Forderungen $\langle q_2 \,|\, q_1\rangle = 0$ und $\|q_2\| = 1$ implizieren

$$0 = \langle q_2 \,|\, q_1\rangle = \left\langle \tfrac{1}{r_{22}}(a_2 - r_{12}q_1) \,\middle|\, q_1 \right\rangle = \tfrac{1}{r_{22}}\left(\langle a_2 \,|\, q_1\rangle - r_{12}\right) \underset{r_{22}\neq 0}{\Longrightarrow} r_{12} = \langle a_2 \,|\, q_1\rangle,$$
$$\|q_2\| = \tfrac{1}{|r_{22}|}\|a_2 - r_{12}q_1\| = 1 \quad\Longrightarrow\quad |r_{22}| = \|\widetilde{q}_2\|, \quad \widetilde{q}_2 = a_2 - r_{12}q_1.$$

Setzte etwa $r_{22} = \|\widetilde{q}_2\|$.

– Betrachte allgemein die k-te Relation (beinhaltet auch den Spezialfall $k = 1$)

$$q_k = \tfrac{1}{r_{kk}}\left(a_k - \sum_{i=1}^{k-1} r_{ik}q_i\right), \quad 1 \le k \le n,$$

mit unbekannten Koeffizienten r_{jk} für $1 \le j \le k$. Die Forderungen $\langle q_k \,|\, q_j\rangle = 0$ für $1 \le j \le k-1$ und $\|q_k\| = 1$ implizieren

$$0 = \langle q_k \,|\, q_j\rangle = \left\langle \tfrac{1}{r_{kk}}\left(a_k - \sum_{i=1}^{k-1} r_{ik}q_i\right) \,\middle|\, q_j \right\rangle = \tfrac{1}{r_{kk}}\left(\langle a_k \,|\, q_j\rangle - r_{jk}\right) \underset{r_{kk}\neq 0}{\Longrightarrow} r_{jk} = \langle a_k \,|\, q_j\rangle,$$
$$\|q_k\| = \tfrac{1}{|r_{kk}|}\left\|a_k - \sum_{i=1}^{k-1} r_{ik}q_i\right\| \quad\Longrightarrow\quad |r_{kk}| = \|\widetilde{q}_k\|, \quad \widetilde{q}_k = a_k - \sum_{i=1}^{k-1} r_{ik}q_i.$$

Setze etwa $r_{kk} = \|\widetilde{q}_k\|$.

Die obigen Überlegungen führen auf folgendes Resultat.

Existenz und Eindeutigkeit der QR-Zerlegung einer Matrix (Satz 3.2): Für jede Matrix $A = \left(a_1 \,\middle|\, \cdots \,\middle|\, a_n\right) \in \mathbb{K}^{m\times n}$ von vollem Rang existiert die reduzierte QR-Zerlegung

$$A = \widehat{Q}\,\widehat{R},$$
$$\widehat{Q} = \left(q_1 \,\middle|\, \cdots \,\middle|\, q_n\right) \in \mathbb{K}^{m\times n},$$
$$\widehat{R} = \begin{pmatrix} r_{11} & \cdots & r_{nn} \\ & \ddots & \vdots \\ & & r_{nn} \end{pmatrix} \in \mathbb{K}^{n\times n}, \qquad r_{ii} \neq 0, \quad 1 \le i \le n,$$

mit orthonormalen Vektoren $q_1, \ldots, q_n \in \mathbb{K}^m$. Durch die zusätzliche Forderung $r_{ii} > 0$ für $1 \le i \le n$ ist die reduzierte QR-Zerlegung eindeutig bestimmt.

Das **klassische Orthogonalisierungverfahren nach Gram–Schmidt** zur Berechnung der reduzierten QR-Zerlegung einer Matrix

$$r_{jk} = \langle a_k \,|\, q_j\rangle, \quad 1 \le j \le k-1,$$
$$\widetilde{q}_k = a_k - \sum_{j=1}^{k-1} r_{jk}q_j, \quad r_{kk} = \|\widetilde{q}_k\|, \quad q_k = \tfrac{1}{r_{kk}}\widetilde{q}_k, \qquad 1 \le k \le n,$$

als Pseudo-Code formuliert lautet beispielsweise folgendermaßen

Eingabedaten: a_k für $1 \le k \le n$
for k=1:n
  $\widetilde{q} = a_k$
  for j=1:k-1
    $r_{jk} = \langle a_k \,|\, q_j \rangle$
    $\widetilde{q} = \widetilde{q} - r_{jk}\, q_j$
  end
  $r_{kk} = \|\widetilde{q}\|$
  $q_k = \frac{1}{r_{kk}}\, \widetilde{q}$
end
Ergebnisse: q_k, r_{jk} für $1 \le j \le k$ und $1 \le k \le n$

Vgl. **Illustration3_OrthogonalisierungGramSchmidt**.

Die **geometrische Veranschaulichung** des Orthogonalisierungverfahrens nach Gram–Schmidt ist, daß vom Basisvektor a_k schrittweise die Anteile $r_{jk}\, q_j = \langle a_k \,|\, q_j \rangle\, q_j$ (Erinnerung: $\langle a_k \,|\, q_j \rangle$ ist die Projektion von a_k auf q_j und $\langle a_k \,|\, q_j \rangle\, q_j$ die Komponente von a_k in Richtung von q_j) subtrahiert werden. Der resultierende Vektor $\widetilde{q}_k$ steht dann orthogonal (senkrecht) auf $q_1, \ldots, q_{k-1}$ bzw. die lineare Hülle $\langle q_1, \ldots, q_{k-1} \rangle$. Durch Normierung (Einheitslänge) erhält man den orthonormalen Vektor q_k.

Vorsicht! Speziell für zwei Vektoren $a_1, a_2 \in \mathbb{K}^m$ führt das klassische Orthogonalisierungverfahren nach Gram–Schmidt auf

$$q_1 = \frac{1}{\|a_1\|}\, a_1\,, \qquad \widetilde{q}_2 = a_2 - \langle a_2 \,|\, q_1 \rangle q_1 = a_2 - \frac{1}{\|a_1\|^2} \langle a_2 \,|\, a_1 \rangle\, a_1\,, \quad q_2 = \frac{1}{\|\widetilde{q}_2\|}\, \widetilde{q}_2\,.$$

Falls die beiden Vektoren a_1, a_2 ***fast* linear abhängig** sind, d.h. es ist $a_2 = c\, a_1 + \delta$ mit $\delta \in \mathbb{K}^m$ und $\|\delta\|$ *klein*, folgt

$$\widetilde{q}_2 = \delta - \frac{1}{\|a_1\|^2} \langle \delta \,|\, a_1 \rangle\, a_1\,,$$

d.h. es ist inbesondere $\|\widetilde{q}_2\|$ *klein* und damit der bei q_2 auftretende Faktor $\frac{1}{\|\widetilde{q}_2\|}$ *groß*. Relative Rundungsfehler bei der Berechnung von $\widetilde{q}_2$ werden deshalb bei der Berechnung von q_2 erheblich verstärkt und der Algorithmus ist **numerisch instabil**.

Betrachte das **Beispiel von Läuchli** mit

$$A = \begin{pmatrix} 1 & 1 & 1 \\ \varepsilon & 0 & 0 \\ 0 & \varepsilon & 0 \\ 0 & 0 & \varepsilon \end{pmatrix} \in \mathbb{R}^{4\times 3},$$

vgl. **Illustration3_OrthogonalisierungGramSchmidt**.

- Ein **modifiziertes Orthogonalisierungverfahren nach Gram–Schmidt**

  Eingabedaten: a_k für $1 \le k \le n$
  for k=1:n
  $\quad \widetilde{q}_k = a_k$
  end
  for k=1:n
  $\quad r_{kk} = \|\widetilde{q}_k\|$
  $\quad q_k = \frac{1}{r_{kk}}\,\widetilde{q}_k$
  $\quad$ for j=k+1:n
  $\quad\quad r_{kj} = \langle a_j \,|\, q_k \rangle$
  $\quad\quad \widetilde{q}_j = \widetilde{q}_j - r_{kj}\, q_k$
  $\quad$ end
  end
  Ergebnisse: q_k, r_{jk} für $1 \le j \le k$ und $1 \le k \le n$

  beruht auf einer zeilenweisen Berechnung der Einträge der Matrix R (anstelle einer spaltenweisen Berechnung).

  Eine zweite **Modifikation** lautet

  Eingabedaten: a_k für $1 \le k \le n$
  for k=1:n
  $\quad \widetilde{q} = a_k$
  $\quad$ for j=1:k-1
  $\quad\quad r_{jk} = \langle \widetilde{q} \,|\, q_j \rangle$
  $\quad\quad \widetilde{q} = \widetilde{q} - r_{jk}\, q_j$
  $\quad$ end
  $\quad r_{kk} = \|\widetilde{q}\|$
  $\quad q_k = \frac{1}{r_{kk}}\,\widetilde{q}$
  end
  Ergebnisse: q_k, r_{jk} für $1 \le j \le k$ und $1 \le k \le n$

  Bemerkungen:

  – Das Orthogonalisierungsverfahren nach Gram–Schmidt und auch beide Modifikationen sind in gewissen Situation numerisch instabil, vgl. Beispiel von Läuchli.

    Für die Modifikationen kann man zeigen, daß unter dem Einfluß von Rundungsfehlern Matrizen $\widetilde{Q}$ (im Allgemeinen nicht unitär!) und $\widetilde{R}$ (obere Dreiecksmatrix)

berechnet werden, die von den exakten Matrizen Q und R der QR-Zerlegung einer Matrix $A \in \mathbb{K}^{n\times n}$ folgendermaßen abweichen

$$\left\|\widetilde{Q}\,\widetilde{R} - A\right\|_2 \le C_1(n)\,\varepsilon_{\text{mach}}\,\|A\|_2\,,$$
$$\left\|\widetilde{\mathbf{Q}}^*\widetilde{\mathbf{Q}} - \mathbf{I}\right\|_2 \le \mathbf{C_2(n)}\,\varepsilon_{\text{mach}}\,\boldsymbol{\kappa_2(A)} + \mathcal{O}\left(\varepsilon_{\text{mach}}^{\mathbf{2}}\,\boldsymbol{\kappa_2(A)^2}\right).$$

Für eine *fast* singuläre Matrix A ist die Konditionzahl $\kappa_2(A)$ *groß*. Vgl. Bemerkung im Skriptum, S. 49 sowie **Beispiel von Läuchli**.

Orthogonalisierungsverfahren und Modifikationen werden aber dennoch im Zusammenhang mit iterativen Verfahren (Arnoldi, GMRES) angewendet.

– Eine numerisch stabile Modifikation des Orthogonalisierungsverfahren nach Gram–Schmidt verwendet zusätzlich die Idee der **Re-Orthogonalisierung**. Eine weitere stabile Alternative der QR-Zerlegung einer Matrix mittels Householder-Reflexionen wird in Abschnitt 4.2 besprochen.

- Im **Vergleich** mit dem Orthogonalisierungsverfahren nach Gram–Schmidt speziell für den Fall $n = 3$

$$\begin{aligned}
&k=1: && r_{11} = \|a_1\|\,, \quad q_1 = \tfrac{1}{r_{11}}\|a_1\| \\
&k=2: && \widetilde{q} = a_2\,, \\
&\quad j=1: && r_{12} = \langle a_2 \,|\, q_1\rangle\,, \quad \widetilde{q} = a_2 - r_{12}\,q_1\,, \quad r_{22} = \|\widetilde{q}\|\,, \quad q_2 = \tfrac{1}{r_{22}}\|\widetilde{q}\| \\
&k=3: && \widetilde{q} = a_3\,, \\
&\quad j=1: && r_{13} = \langle a_3 \,|\, q_1\rangle\,, \quad \widetilde{q} = a_3 - r_{13}\,q_1\,, \\
&\quad j=2: && \mathbf{r_{23}} = \langle \mathbf{a_3} \,|\, \mathbf{q_2}\rangle\,, \quad \widetilde{q} = a_3 - r_{13}\,q_1 - r_{23}\,q_2\,, \quad r_{33} = \|\widetilde{q}\|\,, \quad q_3 = \tfrac{1}{r_{33}}\|\widetilde{q}\|
\end{aligned}$$

führt die erste Modifikation auf

$$\begin{aligned}
&k=1: && r_{11} = \|a_1\|\,, \quad q_1 = \tfrac{1}{r_{11}}\|a_1\| \\
&\quad j=2: && r_{12} = \langle a_2 \,|\, q_1\rangle\,, \quad \widetilde{q}_2 = a_2 - r_{12}\,q_1 \\
&\quad j=3: && r_{13} = \langle a_3 \,|\, q_1\rangle\,, \quad \widetilde{q}_3 = a_3 - r_{13}\,q_1 \\
&k=2: && r_{22} = \|\widetilde{q}_2\|\,, \quad q_2 = \tfrac{1}{r_{22}}\|\widetilde{q}_2\| \\
&\quad j=3: && r_{23} = \langle a_3 \,|\, q_2\rangle\,, \quad \widetilde{q}_3 = a_3 - r_{13}\,q_1 - r_{23}\,q_2 \\
&k=3: && r_{33} = \|\widetilde{q}_3\|\,, \quad q_3 = \tfrac{1}{r_{33}}\|\widetilde{q}_3\|
\end{aligned}$$

und die zweite Modifikation auf

$$\begin{aligned}
&k=1: && r_{11} = \|a_1\|\,, \quad q_1 = \tfrac{1}{r_{11}}\|a_1\| \\
&k=2: && \widetilde{q} = a_2 \\
&\quad j=1: && r_{12} = \langle a_2 \,|\, q_1\rangle\,, \quad \widetilde{q} = a_2 - r_{12}\,q_1\,, \quad r_{22} = \|\widetilde{q}\|\,, \quad q_2 = \tfrac{1}{r_{22}}\|\widetilde{q}\| \\
&k=3: && \widetilde{q} = a_3 \\
&\quad j=1: && r_{13} = \langle \widetilde{q} \,|\, q_1\rangle\,, \quad \widetilde{q} = \widetilde{q} - r_{13}\,q_1 \\
&\quad j=2: && \mathbf{r_{23}} = \langle \widetilde{\mathbf{q}} \,|\, \mathbf{q_2}\rangle = \langle \mathbf{a_3} - \mathbf{r_{13}}\,\mathbf{q_1} \,|\, \mathbf{q_2}\rangle\,, \quad \widetilde{q} = \widetilde{q} - r_{23}\,q_2 \\
&\qquad && r_{33} = \|\widetilde{q}\|\,, \quad q_3 = \tfrac{1}{r_{33}}\|\widetilde{q}\|
\end{aligned}$$

3.5. Normen für Vektoren und Matrizen

- Für Vektoren $x \in \mathbb{K}^n$ und Matrizen $A \in \mathbb{K}^{m\times n}$ sind Beträge sowie die arithmetischen Vergleiche **komponentenweise** zu verstehen, d.h. man definiert

$$|x| = \begin{pmatrix} |x_1| \\ \vdots \\ |x_n| \end{pmatrix}, \qquad |A| = \begin{pmatrix} |a_{11}| & \dots & |a_{1n}| \\ \vdots & & \vdots \\ |a_{m1}| & \dots & |a_{mn}| \end{pmatrix},$$

$$A \le B \quad \text{für} \quad A, B \in \mathbb{K}^{m\times n} \quad \Longleftrightarrow \quad a_{ij} \le b_{ij} \quad \text{für alle} \quad 1 \le i \le m, \quad 1 \le j \le n.$$

- **Erinnerung**: Eine **Norm** auf einem Vektorraum $\mathbb{K}^n$ ist eine Funktion $\|\cdot\| : \mathbb{K}^n \to \mathbb{R}_{\ge 0}$, welche die folgenden Eigenschaften erfüllt (positive Definitheit, Homogenität, Sub-Additivität)

$$\begin{aligned} \|x\| = 0 \Leftrightarrow x = 0, &\qquad x \in \mathbb{K}^n, \\ \|\lambda x\| = |\lambda|\,\|x\|, &\qquad \lambda \in \mathbb{K}, \quad x \in \mathbb{K}^n, \\ \|x + y\| \le \|x\| + \|y\|, &\qquad x, y \in \mathbb{K}^n. \end{aligned}$$

  Neben der **euklidischen Norm** $\|\cdot\|_2$ (definiert durch das euklidische Skalarprodukt) werden üblicherweise die **Betragssummennorm** $\|\cdot\|_1$ und die **Maximumsnorm** $\|\cdot\|_\infty$ verwendet

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|, \quad \|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}, \quad \|x\|_\infty = \max\{|x_i| : 1 \le i \le n\}.$$

  Vgl. Definition 3.3 und Veranschaulichungen zum **Abstand** zweier Vektoren im $\mathbb{R}^2$ und der **Normkugel** im $\mathbb{R}^2$ (Skriptum, S. 51)

$$U = U_{0,1} = \{x \in \mathbb{K}^n : \|x\| \le 1\}.$$

  Bemerkungen:

  – Für die Summen- und Maximumsnorm sind die Normeigenschaften leicht nachzuweisen (Eigenschaften des Betrages).

  – Für die euklidische Norm nützt man den Zusammenhang mit dem euklidischen Skalarprodukt (vgl. früher) und insbesondere die Ungleichung von Cauchy–Schwarz zum Beweis der Dreiecksungleichung.

  – Speziell für die euklidische Norm und $\mathbb{K} = \mathbb{R}$ lautet die Ungleichung von Cauchy–Schwarz

$$\Big(\sum_{i=1}^{n} x_i\, y_i\Big)^2 \le \Big(\sum_{i=1}^{n} x_i^2\Big)\Big(\sum_{i=1}^{n} {y_i}^2\Big), \qquad x, y \in \mathbb{R}^n.$$

    Ein direkter Nachweis beruht etwa auf der **arithmetischen-geometrischen Mittelungleichung**

$$\Big(\prod_{i=1}^{n} x_i\Big)^{\frac{1}{n}} \le \frac{1}{n} \sum_{i=1}^{n} x_i, \qquad x \in \mathbb{R}^n.$$

- Der Begriff der Norm läßt sich direkt auf allgemeine Vektorräume V übertragen (Funktion $\|\cdot\| : V \to \mathbb{R}_{\geq 0}$ mit den Eigenschaften positive Definitheit, Homogenität und Sub-Additivität).

  Zwei Normen $\|\cdot\|, |||\cdot||| : V \to \mathbb{R}_{\geq 0}$ auf einem Vektorraum heißen **äquivalent**, wenn Konstanten $c_1, c_2 > 0$ existieren so, daß für alle $x \in V$ die folgende Relation gilt

  $$c_1 \|x\| \leq |||x||| \leq c_2 \|x\| .$$

  In endlichdimensionsionalen Vektorräumen und insbesondere im $\mathbb{K}^n$ sind alle Normen äquivalent (im Gegensatz zu unendlichdimensionalen Vektorräumen).

  Beispielsweise gilt

  $$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\, \|x\|_\infty .$$

  Denn: Der Index $1 \leq \ell \leq n$ sei so gewählt, daß $\|x\|_\infty = \max\{|x_i| : 1 \leq i \leq n\} = |x_\ell|$. Dann folgt einerseits

  $$\|x\|_\infty^2 = |x_\ell|^2 \leq \sum_{i=1}^{\ell-1} |x_i|^2 + |x_\ell|^2 + \sum_{i=\ell+1}^{n} |x_i|^2 = \|x\|_2^2$$

  und andererseits

  $$\|x\|_2^2 = \sum_{i=1}^{n} |x_i|^2 \leq \sum_{i=1}^{n} |x_\ell|^2 = n\, \|x\|_\infty^2 .$$

  Durch Wurzelziehen folgt die Behauptung. ◇

- Für $A \in \mathbb{K}^{m\times n}$ betrachte die lineare Abbildung $f : (\mathbb{K}^n, \|\cdot\|_{\mathbb{K}^n}) \to (\mathbb{K}^m, \|\cdot\|_{\mathbb{K}^m}) : x \mapsto Ax$. Die zugehörige **Operatornorm** $\|\cdot\|_{\mathbb{K}^m \leftarrow \mathbb{K}^n} : \mathbb{K}^{m\times n} \to \mathbb{R}_{\geq 0}$ ist definiert durch

  $$\|A\|_{\mathbb{K}^m \leftarrow \mathbb{K}^n} = \max_{0 \neq x \in \mathbb{K}^n} \frac{\|Ax\|_{\mathbb{K}^m}}{\|x\|_{\mathbb{K}^n}} = \max_{\|x\|_{\mathbb{K}^n} = 1} \|Ax\|_{\mathbb{K}^m} ,$$

  vgl. Definition 3.4.

  Bemerkungen:

  – Nach Festlegung der betrachteten Normen schreibt man kurz $\|A\|$.

  – Wegen der Linearität von A und der Homogenität der Norm gilt $\frac{1}{\|x\|} \|Ax\| = \|A \frac{x}{\|x\|}\|$ für $x \neq 0$ und deshalb die angegebene Identität bei der Definition der Norm.

  – Für die Identität $I : \mathbb{K}^n \to \mathbb{K}^n$ ergibt sich die Operatornorm $\|I\| = 1$.

  In Analogie zu den Eigenschaften einer Norm auf $\mathbb{K}^n$ erfüllt eine Operatornorm $\|\cdot\| : \mathbb{K}^{m\times n} \to \mathbb{R}_{\geq 0}$ folgende **Eigenschaften** (positive Definitheit, Homogenität, Sub-Additivität bzw. Dreiecksungleichung, Sub-Multiplikativität, Konsistenz)

  $$\begin{aligned}
  \|A\| = 0 \;\Leftrightarrow\; A = 0 , &\qquad A \in \mathbb{K}^{m\times n} , \\
  \|\lambda A\| = |\lambda|\, \|A\| , &\qquad \lambda \in \mathbb{K} , \quad A \in \mathbb{K}^{m\times n} , \\
  \|A + B\| \leq \|A\| + \|B\| , &\qquad A, B \in \mathbb{K}^{m\times n} , \\
  \|AB\| \leq \|A\|\, \|B\| , &\qquad A \in \mathbb{K}^{\ell\times m} , \quad B \in \mathbb{K}^{m\times n} \\
  \|Ax\| \leq \|A\|\, \|x\| , &\qquad A \in \mathbb{K}^{m\times n} , \quad x \in \mathbb{K}^n .
  \end{aligned}$$

Denn: Die Eigenschaften positive Definitheit, Homogenität sowie Sub-Additivität folgen aus den Eigenschaften der Norm auf $\mathbb{K}^m$ ($\|A\| = 0 \Leftrightarrow \|Ax\| = 0$ für alle Vektoren $x \in \mathbb{K}^n \Leftrightarrow A = 0$, $\|\lambda Ax\| = |\lambda|\,\|Ax\|$, $\|(A+B)x\| \le \|Ax\| + \|Bx\|$). Die Eigenschaft Sub-Multiplikativität erhält man aus

$$\begin{aligned}\|AB\| &= \max_{0\neq x\in\mathbb{K}^n} \frac{\|ABx\|}{\|x\|} = \max_{0\neq Bx\in\mathbb{K}^m} \frac{\|ABx\|}{\|x\|} = \max_{0\neq Bx\in\mathbb{K}^m} \frac{\|ABx\|}{\|Bx\|}\frac{\|Bx\|}{\|x\|} \\ &\le \max_{0\neq Bx\in\mathbb{K}^m} \frac{\|ABx\|}{\|Bx\|} \max_{0\neq x\in\mathbb{K}^n} \frac{\|Bx\|}{\|x\|} \le \max_{0\neq y\in\mathbb{K}^m} \frac{\|Ay\|}{\|y\|} \max_{0\neq x\in\mathbb{K}^n} \frac{\|Bx\|}{\|x\|} = \|A\|\,\|B\|.\end{aligned}$$

Für $x = 0$ ist die Eigenschaft Konsistenz offensichtlich und wegen

$$\frac{\|Ax\|}{\|x\|} \le \max_{0\neq x\in\mathbb{K}^n} \frac{\|Ax\|}{\|x\|} = \|A\| \quad\Longrightarrow\quad \|Ax\| \le \|A\|\,\|x\|$$

ergibt sich die Behauptung für $0 \neq x \in \mathbb{K}^n$. ◇

- **Berechnung von Operatornormen** (Satz 3.6): Für die Betragssummennorm, euklische Norm und Maximumsnorm ist die zugehörige Operatornorm einer Matrix $A \in \mathbb{K}^{m\times n}$ gegeben durch die größte Spaltenbetragssumme, die Wurzel des maximalen Eigenwertes von $A^*A \in \mathbb{K}^{n\times n}$ und die größte Zeilenbetragssumme

$$\begin{aligned}\|A\|_1 &= \max\Big\{\sum_{i=1}^m |a_{ij}| : 1 \le j \le n\Big\}, \\ \|A\|_2 &= \max\big\{\sqrt{\lambda} : \lambda \text{ Eigenwert von } A^*A\big\}, \\ \|A\|_\infty &= \max\Big\{\sum_{j=1}^n |a_{ij}| : 1 \le i \le m\Big\}.\end{aligned}$$

Denn (Nachweis der Relation für $\|A\|_2$): Beachte, daß die Matrix

$$B = A^*A \in \mathbb{K}^{n\times n}$$

positiv semi-definit ist ($x^*Bx = \|Ax\|_2^2 \ge 0$ für alle $x \in \mathbb{K}^n$) und folglich alle Eigenwerte nicht-negativ sind (jeder Eigenwert $\lambda \in \mathbb{K}$ mit zugehörigem Eigenvektor $v \in \mathbb{K}^n$ erfüllt $0 \le v^*Bv = \lambda\|v\|_2^2$, d.h. $\lambda \ge 0$). Außerdem ist B wegen $B^* = (A^*A)^* = A^*A^{**} = A^*A = B$ insbesondere eine **normale Matrix**, d.h. es gilt die Gleichheit $B^*B = BB^*$. Eine normale Matrix ist **unitär diagonalisierbar**, d.h. es gilt

$$QBQ^* = \Lambda \qquad \text{bzw.} \qquad B = Q^*\Lambda Q$$

mit einer unitären Matrix $Q \in \mathbb{K}^{n\times n}$ und einer Diagonalmatrix $\Lambda \in \mathbb{K}^{n\times n}$. Nun folgt mittels $\|Qx\|_2 = \|x\|_2$

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \max_{\|x\|_2=1} \sqrt{x^*Bx} = \max_{\|x\|_2=1} \sqrt{x^*Q^*\Lambda Qx} \underset{y=Qx}{=} \max_{\|y\|_2=1} \sqrt{y^*\Lambda y}$$

und schließlich durch Einsetzen der Standardbasisvektoren die Behauptung. ◇

Bemerkung: Für den Spezialfall $A \in \mathbb{R}^{2\times 2}$ ergibt sich

$$A = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}, \quad x = \begin{pmatrix} a \\ b \end{pmatrix}, \quad \|x\|_1 = |a| + |b|, \quad Ax = \begin{pmatrix} \alpha\, a + \beta\, b \\ \gamma\, a + \delta\, b \end{pmatrix}, \quad \|A\|_1 = \max_{\|x\|_1 = 1} \|Ax\|_1,$$

und damit die Abschätzung

$$\begin{aligned} \|Ax\|_1 = |\alpha\, a + \beta\, b| + |\gamma\, a + \delta\, b| &\le \big(|\alpha| + |\gamma|\big)\,|a| + \big(|\beta| + |\delta|\big)\,|b| \\ &\le \max\big\{|\alpha| + |\gamma|, |\beta| + |\delta|\big\}\big(|a| + |b|\big), \\ \|A\|_1 \le \max_{|a|+|b|=1} \max\big\{|\alpha| + |\gamma|, |\beta| + |\delta|\big\} &\underbrace{\big(|a| + |b|\big)}_{=1} = \max\big\{|\alpha| + |\gamma|, |\beta| + |\delta|\big\}. \end{aligned}$$

Durch Einsetzen der Standardbasisvektoren folgt weiters (verwende $\|e_1\|_1 = 1 = \|e_2\|_1$)

$$\begin{cases} |\alpha| + |\gamma| = \|Ae_1\|_1 \le \max\limits_{\|x\|_1=1} \|Ax\|_1 = \|A\|_1, \\ |\beta| + |\delta| = \|Ae_2\|_1 \le \|A\|_1, \end{cases} \quad \Longrightarrow \quad \max\big\{|\alpha| + |\gamma|, |\beta| + |\delta|\big\} \le \|A\|_1.$$

Damit folgt die Gleichheit der Operatornorm von A mit dem Maximum der Betragssummen der ersten bzw. zweiten Spalte von A

$$\|A\|_1 = \max\big\{|\alpha| + |\gamma|, |\beta| + |\delta|\big\}.$$

Ähnliche Überlegungen gelten für die Maximumsnorm.

- **Bemerkungen und Veranschaulichungen von Operatornormen**:
  - Die Norm einer Matrix

    $$\|A\| = \max_{\|x\|=1} \|Ax\|, \qquad A \in \mathbb{K}^{m\times n},$$

    gibt für Elemente der Normkugel in $\mathbb{K}^n$, d.h. Argumente $x \in \mathbb{K}^n$ mit $\|x\| = 1$, die maximale Ausdehnung der entsprechenden Bildelemente $\|Ax\|$ an.
  - Speziell für invertierbare Matrizen $A \in \mathbb{K}^{n\times n}$ gilt wegen der eindeutigen Zuordnung $x \leftrightarrow y = Ax$ bzw. $y \leftrightarrow x = A^{-1}x$ für $x, y \in \mathbb{K}^n$ (verwende außerdem die Relation $\max\{x : x \in X\} = \frac{1}{\min\{x : x\in X\}}$ für eine beschränkte Menge $X \subset \mathbb{R}\setminus\{0\}$)

    $$\big\|A^{-1}\big\| = \max_{y\neq 0} \frac{\big\|A^{-1}y\big\|}{\|y\|} = \max_{Ax\neq 0} \frac{\|x\|}{\|Ax\|} = \max_{x\neq 0} \frac{\|x\|}{\|Ax\|} = \frac{1}{\min\limits_{x\neq 0} \frac{\|Ax\|}{\|x\|}} = \frac{1}{\min\limits_{\|x\|=1} \|Ax\|}$$

    und folglich

    $$\frac{1}{\|A^{-1}\|} = \min_{\|x\|=1} \|Ax\| \qquad \text{für} \quad A \in \mathbb{K}^{n\times n} \quad \text{invertierbar}.$$

    Anschaulich bedeutet das, daß die Operatornorm $\frac{1}{\|A^{-1}\|}$ für Argumente $x \in \mathbb{K}^n$ mit $\|x\| = 1$ die minimale Ausdehnung der entsprechenden Bildelemente $\|Ax\|$ angibt.

– Insbesondere im euklidischen Raum $(\mathbb{R}^2, \|\cdot\|_2)$ bilden für Elemente des Einheitskreises (d.h. $x_1 = \cos t$, $x_2 = \sin t$) die zugehörige Bildelemente unter einer linearen Abbildung $x \mapsto Ax$ mit $A \in \mathbb{R}^{2\times 2}$ eine Ellipse, vgl. Abbildung, Skriptum, S. 53.

- Die obigen Überlegungen motivieren die Definition der Konditionzahl einer Matrix.

  Die **Konditionszahl** einer quadratischen Matrix $A \in \mathbb{K}^{n\times n}$ ist definiert durch

$$\kappa(A) = \begin{cases} \|A\|\,\|A^{-1}\|\,, & \text{falls } A \in \mathbb{K}^{n\times n} \text{ invertierbar}\,, \\ \infty\,, & \text{falls } A \in \mathbb{K}^{n\times n} \text{ singulär}\,, \end{cases}$$

  vgl. Definition 3.5. Allgemein definiert man für $A \in \mathbb{K}^{m\times n}$

$$\kappa(A) = \frac{\max\limits_{\|x\|=1} \|Ax\|}{\min\limits_{\|x\|=1} \|Ax\|}\,.$$

  Bemerkungen:

  – Für die Konditionszahl einer Matrix gilt $\kappa(A) \geq 1$ (vgl. Veranschaulichung im $\mathbb{R}^2$).

  – Für unitäre Matrizen $Q \in \mathbb{K}^{n\times n}$ gilt (wegen $Q^*Q = I = QQ^*$ ist $\|Qx\|_2 = \|x\|_2$ und $\|Q^{-1}x\|_2 = \|Q^*x\|_2 = \|x\|_2$)

$$\|Q\|_2 = 1\,, \quad \|Q^{-1}\|_2 = 1\,.$$

    Somit gilt auch

$$\kappa(Q) = 1\,.$$

    Weiters bleibt bei Transformation einer Matrix $A \in \mathbb{K}^{m\times n}$ mittels einer unitären Matrix $Q \in \mathbb{K}^{m\times m}$ wegen $\|QAx\|_2 = \|Ax\|_2$ für $x \in \mathbb{K}^n$ die Operatornorm und folglich auch die Konditionzahl erhalten

$$\|QA\|_2 = \|A\|_2\,, \qquad \kappa(QA) = \kappa(A)\,.$$

- Für $m \times n$ Matrizen ist die **Frobenius-Norm** $\|\cdot\|_F : \mathbb{K}^{m\times n} \to \mathbb{R}_{\geq 0}$ definiert durch

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2}\,, \qquad A \in \mathbb{K}^{m\times n}\,.$$

  Dies entspricht der euklidischen Norm des entsprechenden Vektors $a \in \mathbb{K}^{m\cdot n}$ mit den Komponenten a_{ij} für $1 \leq i \leq m$ und $1 \leq j \leq n$. Die Frobenius-Norm erfüllt die Eigenschaften positive Definitheit, Homogenität, Sub-Additivität, Sub-Multiplikativität und Konsistenz, ist jedoch keine Operatornorm, da $\|I\|_F = \sqrt{n} \neq 1$ für $I \in \mathbb{K}^{n\times n}$ mit $n \geq 2$.

4. Direkte Verfahren für lineare Gleichungssysteme

- Die näherungsweise **Lösung linearer Gleichungssysteme** ist die wichtigste Grundaufgabe der Numerischen Linearen Algebra und findet beispielsweise Anwendung bei
  - Verfahren zur Lösung nichtlinearer Gleichungssysteme,
  - Verfahren zur Lösung von Optimierungsproblemen,
  - Verfahren zur Lösung gewöhnlicher Differentialgleichungen,
  - Verfahren zur Lösung partieller Differentialgleichungen.
- **Inhalte**:
  - Kondition des Problems.
  - QR-Zerlegung einer Matrix mittels Householder-Reflexionen zur direkten Lösung linearer Gleichungssysteme.
  - Gaußsches Eliminationsverfahren bzw. LR-Zerlegung einer Matrix zur direkten Lösung linearer Gleichungssysteme.
  - Numerische Stabilität der Verfahren.

4.1. Kondition linearer Gleichungssysteme

- **Fragestellung**: Bestimmung der Kondition des numerischen Problems der Lösung eines linearen Gleichungssystems $Ax = b$, d.h.

$$p : \mathbb{K}^{m\times n} \times \mathbb{K}^m \to \mathbb{K}^n : (A, b) \mapsto x\,.$$

  Zu untersuchen ist also die Sensitivität der Lösung $x+\xi = p(A+\alpha, b+\beta)$ gegenüber kleinen Änderungen α und β der Eingangsdaten, d.h. das Ziel ist es, die Lösung x des linearen Gleichungssystems $Ax = b$ ist mit der Lösung $x+\xi$ des linearen Gleichungssystems $(A+\alpha)(x+\xi) = b+\beta$ in Relation zu setzen und eine Abschätzung für den relativen Fehler $\frac{\|\xi\|}{\|x\|}$ herzuleiten.

- **Annahme**: Es sei $A \in \mathbb{K}^{n\times n}$ invertierbar und die Änderung $\alpha \in \mathbb{K}^{n\times n}$ so klein, daß $A+\alpha$ ebenfalls invertierbar ist.

  Resultat zur Invertierbarkeit der Matrix $A+\alpha$ (Satz 4.1):

$$\|\alpha\| < \tfrac{1}{\|A^{-1}\|} \quad \Longrightarrow \quad A+\alpha \quad \text{invertierbar}\,.$$

  Denn: Die Matrix $A+\alpha$ ist genau dann invertierbar, wenn das lineare Gleichungssystem $(A+\alpha)x = 0$ nur die triviale Lösung $x = 0$ besitzt. Mit Hilfe der geforderten Invertierbarkeit von A sowie der Bedingung $\|\alpha\| < \frac{1}{\|A^{-1}\|} \Leftrightarrow 1 - \|A^{-1}\|\|\alpha\| > 0$ folgt

$$\begin{aligned}(A+\alpha)x = 0 \quad &\Longrightarrow \quad Ax = -\alpha x \quad \Longrightarrow \quad x = -A^{-1}\alpha x \\ &\Longrightarrow \quad \|x\| \le \|A^{-1}\|\|\alpha\|\|x\| \quad \Longrightarrow \quad (1-\|A^{-1}\|\|\alpha\|)\|x\| \le 0 \quad \Longrightarrow \quad \|x\| = 0\end{aligned}$$

  und damit $x = 0$. ◇

- **Annahme**: Es sei $A \in \mathbb{K}^{n\times n}$ invertierbar und es gelte $\|\alpha\| < \frac{1}{\|A^{-1}\|}$ für $\alpha \in \mathbb{K}^{n\times n}$, d.h. die Matrix $A+\alpha$ ist ebenfalls invertierbar. Weiters sei $0 \neq b \in \mathbb{K}^n$ sowie $0 \neq x \in \mathbb{K}^n$.

  Abschätzung: Unter der Voraussetzung, daß die relativen Fehler der Eingangsdaten den Schranken (wobei $\varepsilon > 0$ hinreichend klein, sodaß $\varepsilon\kappa(A) = \varepsilon\|A\|\|A^{-1}\| < 1$)

$$\frac{\|\alpha\|}{\|A\|} \le \varepsilon\,, \qquad \frac{\|\beta\|}{\|b\|} \le \varepsilon\,,$$

  genügen, ergibt sich folgende Abschätzung für den relativen Fehler des Ergebnisses (beachte $\frac{\|b\|}{\|A\|\|x\|} \le 1$ wegen $\|b\| = \|Ax\| \le \|A\|\|x\|$)

$$\frac{\|\xi\|}{\|x\|} \le \frac{\varepsilon\kappa(A)}{1-\varepsilon\kappa(A)}\left(1+\frac{\|b\|}{\|A\|\|x\|}\right).$$

  Bemerkung: Die Konditionszahl der Matrix A ist ausschlaggebend für die Kondition des Problems. Falls $\varepsilon\kappa(A) \ll 1$ gilt, ist das Problem gut konditioniert. Falls hingegen $\varepsilon\kappa(A) \approx 1$ gilt, ist die Lösung des linearen Gleichungssystems ein schlecht konditioniertes Problem.

Denn: Es gilt (verwende $1 - \|A^{-1}\|\|\alpha\| > 0$)

$$\begin{aligned}
(A+\alpha)(x+\xi) = b+\beta \quad &\Longrightarrow \quad Ax + A\xi + \alpha x + \alpha\xi = b + \beta \\
&\underset{Ax=b}{\Longrightarrow} \quad A\xi = \beta - \alpha x - \alpha\xi \\
&\Longrightarrow \quad \xi = A^{-1}\big(\beta - \alpha x - \alpha\xi\big) \\
&\Longrightarrow \quad \|\xi\| \le \|A^{-1}\|\big(\|\beta\| + \|\alpha\|\|x\| + \|\alpha\|\|\xi\|\big) \\
&\Longrightarrow \quad \big(1 - \|A^{-1}\|\|\alpha\|\big)\|\xi\| \le \|A^{-1}\|\big(\|\beta\| + \|\alpha\|\|x\|\big) \\
&\Longrightarrow \quad \|\xi\| \le \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|\alpha\|}\big(\|\beta\| + \|\alpha\|\|x\|\big).
\end{aligned}$$

Einsetzen der vorausgesetzen Schranken $\|\alpha\| \le \varepsilon\|A\|$ und $\|\beta\| \le \varepsilon\|b\|$ sowie der Konditionszahl $\kappa(A) = \|A\|\|A^{-1}\|$ unter Verwendung der Relation

$$\begin{gathered}
0 \le q = \|A^{-1}\|\|\alpha\| < 1, \qquad q \le \widetilde{q} = \varepsilon\kappa(A) < 1, \\
\frac{1}{1 - \|A^{-1}\|\|\alpha\|} = \frac{1}{1-q} = \sum_{k=0}^{\infty} q^k \le \sum_{k=0}^{\infty} \widetilde{q}^k = \frac{1}{1-\widetilde{q}} = \frac{1}{1 - \varepsilon\kappa(A)},
\end{gathered}$$

ergibt die Abschätzung

$$\|\xi\| \le \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|\alpha\|}\big(\|\beta\| + \|\alpha\|\|x\|\big) \le \frac{\varepsilon\kappa(A)}{1 - \varepsilon\kappa(A)}\left(\frac{\|b\|}{\|A\|} + \|x\|\right),$$

und damit folgt die Behauptung. ◇

- **Bedeutung des Residuums**:
  - Für eine Näherungslösung $\widetilde{x}$ an die Lösung x des linearen Gleichungssystems $Ax = b$ ergibt sich durch Einsetzen das **Residuum**

    $$r(\widetilde{x}) = A\widetilde{x} - b.$$

    Klarerweise gilt $r(x) = Ax - b = 0$.
  - **Aber!** Aus der Größe des Residuums kann man im Allgemeinen nichts über die Güte der Näherungslösung, d.h. die Größe des Fehlers $\|\widetilde{x} - x\|$, ableiten.
  - Es gilt die Abschätzung

    $$\|\widetilde{x} - x\| \le \kappa(A)\frac{\|r(\widetilde{x})\|}{\|A\|},$$

    die aus der Überlegung

    $$\begin{aligned}
    r(\widetilde{x}) = A\widetilde{x} - b = A(\widetilde{x} - x) \quad &\Longrightarrow \quad \widetilde{x} - x = A^{-1}\, r(\widetilde{x}) \\
    &\Longrightarrow \quad \|\widetilde{x} - x\| \le \|A^{-1}\|\,\|r(\widetilde{x})\| = \kappa(A)\frac{\|r(\widetilde{x})\|}{\|A\|}
    \end{aligned}$$

    folgt.

– Aus der Relation

$$A\tilde{x} = b + r(\tilde{x})$$

folgert man, daß bei einem *kleinen* Residuum die Näherungslösung $\tilde{x}$ der exakten Lösung eines linearen Gleichungssystems mit exakter Matrix A und leicht abgeänderter rechter Seite $b + r(\tilde{x})$ entspricht und damit eine akzeptable Näherungslösung ist.

– Für zwei Näherungslösungen $\tilde{x}$ und $\hat{x}$ kann aber folgende Situation eintreten

$$\|\tilde{x} - x\| \approx \|\hat{x} - x\| \qquad \text{und} \qquad \|r(\tilde{x})\| \ll \|r(\hat{x})\|$$

oder sogar

$$\|\tilde{x} - x\| \gg \|\hat{x} - x\| \qquad \text{und} \qquad \|r(\tilde{x})\| \ll \|r(\hat{x})\|\,.$$

– **Beispiel von Kahan**, vgl. **Illustration4_Kahan** und Abbildung, Skriptum, S. 59.

4.2. Lösung über die QR-Zerlegung

- **Situation**: Es sei $A \in \mathbb{K}^{n\times n}$ invertierbar und es bezeichne

$$A = QR,$$

$$\left(a_1 \middle| \cdots \middle| a_n\right) = \left(q_1 \middle| \cdots \middle| q_n\right) \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ & & r_{nn} \end{pmatrix},$$

die (eindeutig bestimmte reduzierte bzw. volle) QR-Zerlegung von A mit unitärer Matrix $Q \in \mathbb{K}^{n\times n}$ (d.h. $Q^*Q = I$) und oberer Dreiecksmatrix $R \in \mathbb{K}^{n\times n}$ mit positiven Diagonaleinträgen $r_{ii} > 0$ für $1 \le i \le n$.

Erinnerung: Die Berechnung der QR-Zerlegung von A mittels des Orthogonalisierungsverfahrens nach Gram–Schmidt basiert auf der Idee, die Spaltenvektoren von A zu orthogonalisieren. Die Forderung $\langle a_1,\dots,a_k\rangle = \langle q_1,\dots,q_k\rangle$ für $1 \le k \le n$ führt auf eine Dreiecksmatrix R. Die numerische Instabilität des Verfahrens in gewissen Situationen erfordert die Konstruktion eines geeigneten alternativen Verfahrens.

Alternativer Zugang: Ein alternatives Verfahren zur Berechnung der QR-Zerlegung einer (invertierbaren) Matrix $A \in \mathbb{K}^{n\times n}$ verwendet die Idee der **Triangulierung** von A durch Multiplikation mit geeignet gewählten unitären Matrizen $Q_n^*,\dots,Q_1^*$ (d.h. $Q_i^* = Q_i$ für $1 \le i \le n$, wegen $(Q_{i+1}Q_i)^*(Q_{i+1}Q_i) = Q_i^*Q_{i+1}^*Q_{i+1}Q_i = I$ ist das Produkt unitärer Matrizen und damit Q^* unitär)

$$\underbrace{Q_1^*\cdots Q_n^*}_{=Q^*}A = R \quad\Longleftrightarrow\quad A = \underbrace{Q_n\cdots Q_1}_{=Q}R.$$

Transformationen zur Triangulierung von A basieren auf der Verwendung von **Householder-Reflexionen** oder Givens-Rotationen.

- Eine **Householder-Reflexion** ist eine Matrix der Form

$$T = I - 2vv^* \in \mathbb{K}^{n\times n}, \qquad v \in \mathbb{K}^n, \quad \|v\|_2 = 1.$$

Bemerkungen zu Householder-Reflexionen:

  - Im allgemeinen ist eine Householder-Reflexion eine *volle* Matrix.

    Der Rang einer Householder-Reflexion ist 1.

    Zur **Berechnung des Produktes** Tx mit $x \in \mathbb{K}^n$ verwendet man **ausschließlich** die folgende Relation

$$Tx = x - 2v\underbrace{v^*x}_{=c\in\mathbb{K}} = x - 2cv.$$

    Zur **Berechnung der transformierten Matrix** TA verwendet man

$$TA = \left(Ta_1 \middle| \cdots \middle| Ta_n\right), \qquad Ta_j = a_j - 2(v^*a_j)v, \quad 1 \le j \le n.$$

– Eine Householder-Reflexion ist eine selbstadjungierte und unitäre Matrix

$$T^* = \left(I - 2\,v v^*\right)^* = I^* - 2\,(v v^*)^* = I - 2\,v^{**} v^* = I - 2\,v v^* = T\,,$$
$$T^* T = T^2 = \left(I - 2\,v v^*\right)\left(I - 2\,v v^*\right) = I - 4\,v v^* + 4\,v \underbrace{v^* v}_{=\|v\|_2^2=1} v^* = I\,.$$

– Die Bezeichnung Householder-Reflexion erklärt sich durch folgende Eigenschaften: Es sei $v, w_1, \ldots, w_{n-1}$ eine Orthonormalbasis des $\mathbb{K}^n$. Einerseits wird v bei Anwendung von T auf $-v$ abgebildet

$$T v = v - 2\,v \underbrace{v^* v}_{=1} = v - 2\,v = -v\,.$$

Andererseits wird jeder auf v orthogonale Vektor $w \in \mathbb{K}^n$, d.h. $w \in \langle w_1, \ldots, w_{n-1}\rangle$ (**Hyperebene** im $\mathbb{K}^n$), auf sich selbst abgebildet

$$T w = w - 2\,v \underbrace{v^* w}_{=0} = w\,.$$

Für einen beliebigen Vektor $x \in \mathbb{K}^n$ folgt mittels eindeutiger Darstellung als Linearkombination der Basisvektoren

$$x = \lambda\, v + w\,, \qquad \lambda = v^* x\,, \quad w \in \langle w_1, \ldots, w_{n-1}\rangle\,,$$
$$T x = T(\lambda\, v + w) = \lambda\, T v + T w = -\lambda\, v + w\,,$$

d.h. die Multiplikation mit der Matrix T bewirkt eine Spiegelung an der zum Vektor v orthogonalen Hyperebene, vgl. Abbildung, Skriptum, S. 62.

– In Hinblick auf die Triangulierung einer Matrix möchte man beispielsweise erreichen, daß durch Multiplikation mit einer Householder-Reflexion der erste Spaltenvektor auf ein Vielfaches des ersten Standardbasisvektors transformiert wird. Die entsprechende Householder-Reflexion wird nun abgeleitet.

Vorbemerkung: Aufgrund der Selbstadjungiertheit von T (d.h. es gilt $T^* = T$) folgt speziell für $y = Tx$

$$x^* y \underset{y=Tx}{=} x^* T x \underset{T^*=T}{=} x^* T^* x = (Tx)^* x \underset{y=Tx}{=} y^* x \underset{\overline{y}^T x = x^T \overline{y} = \overline{\overline{x}^T y}}{=} \overline{x^* y} \quad \Longrightarrow \quad x^* y \in \mathbb{R}\,.$$

Allgemeiner Fall: Eine Householder-Reflexion T soll so bestimmt werden, daß zwei vorgegebene Vektoren gleicher Norm ineinander übergehen

$$y = Tx\,, \qquad x, y \in \mathbb{K}^n\,, \quad x \neq y\,, \quad \|x\|_2 = \|y\|_2\,, \quad x^* y = y^* x \in \mathbb{R}\,.$$

Die obigen Überlegungen ergeben

$$x = \lambda\, v + w\,, \quad y = -\lambda\, v + w \quad \Longrightarrow \quad x - y = 2\lambda\, v$$

und führen damit auf die Wahl (wegen $v = c(x-y)$ und $\|v\|_2 = 1$)

$$v = \tfrac{1}{\|x-y\|_2}(x-y)\,.$$

Man setzt somit

$$T = I - 2\,v v^*\,, \qquad v = \tfrac{1}{\|x-y\|_2}(x-y)\,.$$

Spezialfall: Für ein vorgegebenes Argument $x \in \mathbb{K}^n$ und ein (zu bestimmendes) Vielfaches des ersten Standardbasisvektors als zugehöriges Bildelement $y \in \mathbb{K}^n$ führen die obigen Überlegungen auf (Definition von α in Übereinstimmung mit obigen Bedingungen, Wahl des Vorzeichens von α zur Vermeidung von Auslöschung bei der Berechnung von $x-y = (x_1+\alpha, x_2, \ldots, x_n)$, Norm von $x-y$ vereinfacht sich zu $\|x-y\|_2^2 = \|x\|_2^2 + |\alpha|^2 + 2\,\Re(\alpha\,\overline{x_1}) = 2\,\|x\|_2^2 + 2\,|x_1|\,\|x\|_2$)

$$Tx = y = -\alpha\,e_1\,, \qquad \alpha = \begin{cases} \|x\|_2\,\frac{x_1}{|x_1|}\,, & \text{falls } x_1 \neq 0\,, \\ \|x\|_2\,, & \text{falls } x_1 = 0\,, \end{cases}$$
$$T = I - 2\,v v^*\,, \qquad v = \tfrac{1}{\|x-y\|_2}(x-y)\,.$$

- **Berechnung der QR-Zerlegung** einer Matrix mittels Householder-Reflexionen: Die schrittweise Anwendung der obigen Überlegungen führt auf die Triangulierung von

$$A = \left(a_1 \,\middle|\, \cdots \,\middle|\, a_n\right) \in \mathbb{K}^{n\times n}\,.$$

  - 1. Schritt: Elimination in der ersten Spalte

$$T_1 = I - 2\,v_1 v_1^*\,, \quad v_1 = \tfrac{a_1+\alpha_1 e_1}{\|a_1+\alpha_1 e_1\|_2}\,, \quad \alpha_1 = \begin{cases} \|a_1\|_2\,\frac{a_{11}}{|a_{11}|}\,, & a_{11} \neq 0\,, \\ \|a_1\|_2\,, & a_{11} = 0\,, \end{cases}$$
$$T_1 A = \left(T_1 a_1 \,\middle|\, \cdots \,\middle|\, T_1 a_n\right), \quad T_1 a_1 = -\alpha_1 e_1\,, \quad T_1 a_j = a_j - 2\,(v_1^* a_j)\,v_1\,, \quad 2 \le j \le n\,,$$
$$T_1 A = \begin{pmatrix} -\alpha_1 & * & \ldots & * \\ & * & \ldots & * \\ & \vdots & & \vdots \\ & * & \ldots & * \end{pmatrix} = \begin{pmatrix} -\alpha_1 & * \\ & \\ & A_1 \\ & \end{pmatrix} \in \mathbb{K}^{n\times n}, \quad A_1 \in \mathbb{K}^{(n-1)\times(n-1)}\,.$$

  - 2. Schritt: Elimination in der zweiten Spalte

$$T_2 = \begin{pmatrix} 1 & & \\ & \widetilde{T}_2 & \\ & & \end{pmatrix} \in \mathbb{K}^{n\times n}\,,$$
$$\widetilde{T}_2 = \begin{pmatrix} -\alpha_2 & * \\ & \\ & A_2 \end{pmatrix} \in \mathbb{K}^{(n-1)\times(n-1)}, \quad A_2 \in \mathbb{K}^{(n-2)\times(n-2)}\,,$$
$$T_2 T_1 A = \begin{pmatrix} -\alpha_1 & * & * & \ldots & * \\ & -\alpha_2 & * & \ldots & * \\ & & \vdots & & \vdots \\ & & * & \ldots & * \end{pmatrix}.$$

Insgesamt ergibt sich nach $n-1$ Schritten die QR-Zerlegung von A

$$\underbrace{T_{n-1}\cdots T_1}_{=Q^*} A = R \quad \Longleftrightarrow \quad A = \underbrace{T_1\cdots T_{n-1}}_{=Q} R\,.$$

Als Pseudo-Code formuliert lautet die Berechnung der QR-Zerlegung einer Matrix mittels Householder-Reflexionen folgendermaßen

Eingabedaten: a_{ij} für $1 \le i, j \le n$

for k = 1:n-1

$x = A(k:\text{end}, k)$

$u = x\,, \quad u(1) = u(1) + \|x\|_2 \frac{x_1}{|x_1|}$

$v_k = \frac{1}{\|u\|_2}\, u$

$A(k:\text{end}, k:\text{end}) = A(k:\text{end}, k:\text{end}) - 2\, v_k \left(v_k^* A(k:\text{end}, k:\text{end})\right)$

end

Ergebnisse: a_{ij} für $1 \le i \le j \le n$ und v_{ik} für $1 \le i \le n+1-k$ und $1 \le k \le n-1$

Bemerkungen:

- Beachte, daß $v_k \in \mathbb{K}^{n+1-k}$ für $1 \le k \le n-1$.
- Die Koeffizienten von A können sukzessive durch die neu berechneten Einträge der Dreiecksmatrix R und die Komponenten von v_k überschrieben werden. Zusätzlich werden die Diagonalelemente von R abgespeichert.
- Die Anzahl der zur Berechnung der QR-Zerlegung einer Matrix $A \in \mathbb{K}^{n\times n}$ benötigten Operationen (Addition oder Multiplikation in $\mathbb{K}$) ist

  $$\mathcal{O}\left(\tfrac{4}{3}\, n^3\right),$$

  vgl. Skriptum, S. 65.
- Die Erweiterung des Algorithmus auf Matrizen $A \in \mathbb{K}^{m\times n}$ von vollem Rang verwendet die Hinzunahme von T_n (d.h. im Pseudo-Algorithmus ersetzt man die Schleife for k = 1:n-1 ... end durch for k = 1:n ... end), vgl. Skriptum, S. 66.
- Die Berechnung des Produktes

  $$Q^* b = T_{n-1}\cdots T_1 b$$

  für $b \in \mathbb{K}^n$ benötigt die Kenntnis der Vektoren $v_k \in \mathbb{K}^{n+1-k}$ für $1 \le k \le n-1$. Als Pseudo-Code ergibt sich

  Eingabedaten: b_i für $1 \le i \le n$ und v_{ik} für $1 \le i \le n+1-k$ und $1 \le k \le n-1$

  for k = 1:n-1

  $b(k:\text{end}) = b(k:\text{end}) - 2\, v_k \left(v_k^* b(k:\text{end})\right)$

  end

  Ergebnisse: b_i für $1 \le i \le n$

- **Lösung eines linearen Gleichungssystems mittels QR-Zerlegung**: Falls die QR-Zerlegung der Matrix $A \in \mathbb{K}^{n\times n}$ bekannt ist, ist die Lösung des linearen Gleichungssystems $Ax = b$ (wobei $x \in \mathbb{K}^n$ und $b \in \mathbb{K}^n$) einfach durchzuführen. Einsetzen der QR-Zerlegung von A und Multiplikation des Gleichungssystems mit der adjungierten Matrix Q^* (invertierbar) führt auf ein äquivalentes **gestaffeltes lineares Gleichungssystem**

$$Ax = b \quad \underset{A=QR}{\Longleftrightarrow} \quad QRx = b \quad \underset{Q^*Q=I}{\Longleftrightarrow} \quad Rx = \underbrace{Q^*b}_{=c}$$

dessen Lösung direkt mittels **Rückwärtssubstitution** berechnet werden kann (verwende $r_{ij} = 0$ für $1 \le j < i \le n$ und $r_{ii} > 0$ für $1 \le i \le n$)

$$Rx = c\,,$$

$$\begin{pmatrix} r_{11} & \dots & r_{1n} \\ & \ddots & \vdots \\ & & r_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix},$$

$$\sum_{j=i}^{n} r_{ij}\,x_j = c_i\,, \quad x_i = \frac{1}{r_{ii}}\Big(c_i - \sum_{j=i+1}^{n} r_{ij}\,x_j\Big), \qquad 1 \le i \le n\,,$$

$$x_n = \frac{1}{r_{nn}}\,c_n\,, \quad \dots, \quad x_1 = \frac{1}{r_{11}}\Big(c_1 - \sum_{j=2}^{n} r_{1j}\,x_j\Big).$$

Als Pseudo-Code lautet die Rückwärtssubstitution

Eingabedaten: c_i, r_{ij} für $1 \le i \le j \le n$
for i = n:-1:1
    $x_i = c_i$
    for j = i+1:n
        $x_i = x_i - r_{ij}x_j$
    end
    $x_i = \frac{1}{r_{ii}}\,x_i$
end
Ergebnisse: x_i für $1 \le i \le n$

4.3. Stabilität der Lösungsmethode über die QR-Zerlegung

- **Erinnerung**: Das in Abschnitt 4.2 besprochene Verfahren zur Lösung eines linearen Gleichungssystems $Ax = b$ mit $A \in \mathbb{K}^{n\times n}$ invertierbar und $b \in \mathbb{K}^n$ umfaßt die folgenden Schritte.
  - QR-Zerlegung von $A = QR$ bzw. Triangulierung von A mittels Householder-Reflexionen,
  - Berechnung des Produktes $c = Q^* b$,
  - Rückwärtssubstitution zur Berechnung von x aus $Rx = c$.

  Stabilität des Verfahrens (Satz 4.2): Die unter dem Einfluß von Rundungsfehlern berechnete Näherungslösung $\widetilde{x}$ ist die exakte Lösung eines linearen Gleichungssystems

$$\widetilde{A}\widetilde{x} = b \qquad \text{mit} \qquad \|A - \widetilde{A}\|_2 \le \mathcal{O}(\varepsilon_{\text{mach}})\,\|A\|_2\,,$$

  d.h. das Verfahren ist im strengen Sinn numerisch stabil.

  Bemerkungen:
  - Die Größe $\mathcal{O}(\varepsilon_{\text{mach}})$ hängt von der Dimension n der Matrix A ab.
  - Begründung der Stabilitätsaussage für die Rückwärtssubstitution. Zusätzliche Überlegungen zur Ableitung der Abschätzung.
  - Zusätzliche Berechnung des Residuums $A\widetilde{x} - b$ sichert eine akzeptable Näherungslösung bei kleinem Residuum, vgl. Bemerkung, Skriptum, S. 69.

- **Rundungsfehleranalyse der Rückwärtssubstitution** (im Sinne der Rückwärtsanalyse): Die Rückwärtssubstitution für das lineare Gleichungssystem $Rx = c$ mit invertierbarer oberer Dreiecksmatrix $R \in \mathbb{K}^{n\times n}$ beruht auf der Berechnung der unbekannten Komponenten $x_n, x_{n-1}, \ldots, x_1$ mittels der Relation

$$x_i = \frac{1}{r_{ii}}\Big(c_i - \sum_{j=i+1}^{n} r_{ij}\,x_j\Big), \qquad 1 \le i \le n\,.$$

  Unter dem Einfluß von Rundungsfehlern ergibt sich folgende Näherungslösung (verwende $\text{rd}(a * b) = (a * b)\,(1+\varepsilon)$ mit $|\varepsilon| \le \varepsilon_{\text{mach}}$)

$$\longrightarrow \quad \text{rd}\big(c_i - r_{i,i+1}\,\widetilde{x}_{i+1}\,(1+\widetilde{\varepsilon}_{i1})\big) = c_i\,(1+\varepsilon_{i1}) - r_{i,i+1}\,\widetilde{x}_{i+1}\,(1+\varepsilon_{i1})\,(1+\widetilde{\varepsilon}_{i1})$$

$$\begin{aligned}\longrightarrow \quad & \text{rd}\Big(c_i\,(1+\varepsilon_{i1}) - r_{i,i+1}\,\widetilde{x}_{i+1}\,(1+\varepsilon_{i1})\,(1+\widetilde{\varepsilon}_{i1}) - r_{i,i+2}\,\widetilde{x}_{i+2}\,(1+\widetilde{\varepsilon}_{i2})\Big)\\ &= c_i\,(1+\varepsilon_{i1})\,(1+\varepsilon_{i2}) - r_{i,i+1}\,\widetilde{x}_{i+1}\,(1+\varepsilon_{i1})\,(1+\varepsilon_{i2})\,(1+\widetilde{\varepsilon}_{i1})\\ &\quad - r_{i,i+2}\,\widetilde{x}_{i+2}\,(1+\varepsilon_{i2})\,(1+\widetilde{\varepsilon}_{i2})\end{aligned}$$

$$\begin{aligned}\longrightarrow \quad & \text{rd}\Big(c_i\,(1+\varepsilon_{i1})\,(1+\varepsilon_{i2}) - r_{i,i+1}\,\widetilde{x}_{i+1}\,(1+\varepsilon_{i1})\,(1+\varepsilon_{i2})\,(1+\widetilde{\varepsilon}_{i1})\\ &\quad - r_{i,i+2}\,\widetilde{x}_{i+2}\,(1+\varepsilon_{i2})\,(1+\widetilde{\varepsilon}_{i2})\Big)\\ &= c_i\,(1+\varepsilon_{i1})\,(1+\varepsilon_{i2})\,(1+\varepsilon_{i3}) - r_{i,i+1}\,\widetilde{x}_{i+1}\,(1+\varepsilon_{i1})\,(1+\varepsilon_{i2})\,(1+\varepsilon_{i3})\,(1+\widetilde{\varepsilon}_{i1})\\ &\quad - r_{i,i+2}\,\widetilde{x}_{i+2}\,(1+\varepsilon_{i2})\,(1+\varepsilon_{i3})\,(1+\widetilde{\varepsilon}_{i2})\end{aligned}$$

und insgesamt (zur Vereinfachung werden die Größen $\widetilde{\varepsilon}_{ij}$ nicht unterschieden)

$$\begin{aligned} \dots \quad &\longrightarrow \quad c_i \prod_{\ell=1}^{n-i}(1+\varepsilon_{i\ell}) - \sum_{j=i+1}^{n} r_{ij}\,\widetilde{x}_j \prod_{\ell=j-i}^{n-i}(1+\varepsilon_{i\ell})\,(1+\widetilde{\varepsilon}) \\ &\longrightarrow \quad \widetilde{x}_i = \tfrac{1+\widetilde{\varepsilon}}{r_{ii}}\Big(c_i \prod_{\ell=1}^{n-i}(1+\varepsilon_{i\ell}) - \sum_{j=i+1}^{n} r_{ij}\,\widetilde{x}_j \prod_{\ell=j-i}^{n-i}(1+\varepsilon_{i\ell})\,(1+\widetilde{\varepsilon})\Big) \end{aligned}$$

mit $|\varepsilon_{i\ell}| \le \varepsilon_{\text{mach}}$ für $1 \le \ell \le n-i$ und $|\widetilde{\varepsilon}| \le \varepsilon_{\text{mach}}$. Weiters erhält man

$$\begin{aligned} \widetilde{x}_i = \tfrac{1+\widetilde{\varepsilon}}{r_{ii}}\Big(c_i \prod_{\ell=1}^{n-i}(1+\varepsilon_{i\ell}) - \sum_{j=i+1}^{n} r_{ij}\,\widetilde{x}_j \prod_{\ell=j-i}^{n-i}(1+\varepsilon_{i\ell})\,(1+\widetilde{\varepsilon})\Big) \\ \iff \quad \tfrac{r_{ii}}{1+\widetilde{\varepsilon}}\,\widetilde{x}_i + \sum_{j=i+1}^{n} r_{ij}\,\widetilde{x}_j \prod_{\ell=j-i}^{n-i}(1+\varepsilon_{i\ell})\,(1+\widetilde{\varepsilon}) = c_i \prod_{\ell=1}^{n-i}(1+\varepsilon_{i\ell}) \\ \iff \quad c_i = r_{ii}\,\widetilde{x}_i \prod_{\ell=1}^{n-i}\tfrac{1}{1+\varepsilon_{i\ell}}\,\tfrac{1}{1+\widetilde{\varepsilon}} + \sum_{j=i+1}^{n} r_{ij}\,\widetilde{x}_j \prod_{\ell=1}^{n-i}\tfrac{1}{1+\varepsilon_{i\ell}} \prod_{\ell=j-i}^{n-i}(1+\varepsilon_{i\ell})\,(1+\widetilde{\varepsilon}) \\ \iff \quad c_i = r_{ii}\,\widetilde{x}_i \prod_{\ell=1}^{n-i}\tfrac{1}{1+\varepsilon_{i\ell}}\,\tfrac{1}{1+\widetilde{\varepsilon}} + \sum_{j=i+1}^{n} r_{ij}\,\widetilde{x}_j \prod_{\ell=1}^{j-i-1}\tfrac{1}{1+\varepsilon_{i\ell}}\,(1+\widetilde{\varepsilon})\,. \end{aligned}$$

Mittels Lemma 2.6. folgt damit die Relation (beachte die Abhängigkeit von n)

$$c_i = \sum_{j=i}^{n} r_{ij}\,(1+\delta_{ij})\,\widetilde{x}_j\,, \qquad |\delta_{ij}| \le \tfrac{n\varepsilon_{\text{mach}}}{1-n\varepsilon_{\text{mach}}} = \mathcal{O}(\varepsilon_{\text{mach}})\,, \quad 1 \le i \le j \le n\,.$$

Dies zeigt, daß die Näherungslösung $\widetilde{x}$ die exakte Lösung eines linearen Gleichungssystems ist, dessen Matrix die folgende komponentenweise Abschätzung bzw. Normabschätzung erfüllt (verwende $R = Q^* A$ und folglich $\|R\|_2 = \|A\|_2$)

$$\widetilde{R}\,\widetilde{x} = c\,, \qquad |\widetilde{R} - R| \le \mathcal{O}(\varepsilon_{\text{mach}})\,|R|\,, \quad \|\widetilde{R} - R\|_2 \le \mathcal{O}(\varepsilon_{\text{mach}})\,\|R\|_2 = \mathcal{O}(\varepsilon_{\text{mach}})\,\|A\|_2\,.$$

- **Zusätzliche Überlegungen**: Es seien $\widetilde{v}_1, \dots, \widetilde{v}_{n-1}$ die bei der Triangulierung einer invertierbaren Matrix $A \in \mathbb{K}^{n\times n}$ mittels Householder-Reflexionen berechneten Vektoren und es bezeichnen $\widetilde{T}_k = I - 2\,\widetilde{v}_k\,\widetilde{v}_k^*$ für $1 \le k \le n-1$ und $\widetilde{Q} = \widetilde{T}_1 \cdots \widetilde{T}_{n-1}$ die daraus entstehenden unitären Matrizen sowie $\widetilde{R}$ die berechnete obere Dreiecksmatrix. Dann gilt im Vergleich mit der reduzierten QR-Zerlegung von A die folgende Abschätzung (ohne Begründung)

$$\|\widetilde{Q}\,\widetilde{R} - A\|_2 \le \mathcal{O}(\varepsilon_{\text{mach}})\,\|A\|_2\,.$$

Weiters erfüllt die bei der Berechnung von $\widetilde{Q}^* b$ resultierende Näherungslösung $\widetilde{c}$ die Relation (ohne Begründung)

$$\Delta Q\,\widetilde{c} = b - \widetilde{Q}\,\widetilde{c}\,, \qquad \|\Delta Q\|_2 \le \mathcal{O}(\varepsilon_{\text{mach}})\,.$$

Folglich ergibt sich mit ΔR definiert durch $\Delta R\,\widetilde{x} = \widetilde{c} - \widetilde{R}\,\widetilde{x}$, d.h. $\widetilde{c} = (\widetilde{R} + \Delta R)\,\widetilde{x}$, die Relation

$$b = (\widetilde{Q} + \Delta Q)\,\widetilde{c} = (\widetilde{Q} + \Delta Q)\,(\widetilde{R} + \Delta R)\,\widetilde{x} = \widetilde{A}\,\widetilde{x}\,.$$

Wegen $\widetilde{A} - A = (\widetilde{Q} + \Delta Q)\,(\widetilde{R} + \Delta R) - A = \widetilde{Q}\,\widetilde{R} - A + \Delta Q\,\widetilde{R} + \widetilde{Q}\,\Delta R + \Delta Q\,\Delta R$ ergibt sich weiters (verwende die Abschätzungen $\|\widetilde{R}\|_2 = \|\widetilde{Q}\,\widetilde{R}\|_2 \le (1 + \mathcal{O}(\varepsilon_{\text{mach}}))\,\|A\|_2 \le \mathcal{O}(1)\,\|A\|_2$ und $\|\widetilde{Q}\,\Delta R\|_2 = \|\widetilde{R} - R\|_2 \le \mathcal{O}(\varepsilon_{\text{mach}})\,\|A\|_2$)

$$\begin{aligned}
\|\widetilde{A} - A\|_2 &= \|\widetilde{Q}\,\widetilde{R} - A + \Delta Q\,\widetilde{R} + \widetilde{Q}\,\Delta R + \Delta Q\,\Delta R\|_2 \\
&\le \underbrace{\|\widetilde{Q}\,\widetilde{R} - A\|_2}_{\le \mathcal{O}(\varepsilon_{\text{mach}})\,\|A\|_2} + \underbrace{\|\Delta Q\|_2}_{\le \mathcal{O}(\varepsilon_{\text{mach}})} \underbrace{\|\widetilde{R}\|_2}_{\le \mathcal{O}(1)\,\|A\|_2} + \underbrace{\|\widetilde{Q}\,\Delta R\|_2}_{\le \mathcal{O}(\varepsilon_{\text{mach}})\,\|A\|_2} + \underbrace{\|\Delta Q\,\Delta R\|_2}_{\le \mathcal{O}(\varepsilon_{\text{mach}}^2)\,\|A\|_2} \\
&\le \mathcal{O}(\varepsilon_{\text{mach}})\,\|A\|_2\,.
\end{aligned}$$

Dies führt auf die Abschätzung von Satz 4.2

$$\widetilde{A}\,\widetilde{x} = b\,, \qquad \|\widetilde{A} - A\|_2 \le \mathcal{O}(\varepsilon_{\text{mach}})\,\|A\|_2\,.$$

4.4. Gauß-Elimination und Dreieckszerlegung

- **Situation**: Lösung eines linearen Gleichungssystems

$$Ax = b$$

mit $A \in \mathbb{K}^{n\times n}$ invertierbar und $b \in \mathbb{K}^n$.

Das **Gaußsche Eliminationsverfahren** zur Lösung des linearen Gleichungssystems basiert auf folgender Idee:

  - Eine der n Gleichungen wird explizit nach einer der Unbekannten aufgelöst (z.B. nach x_1). Die resultierende Bedingung (z.B. $x_1 = f_1(x_2,\ldots,x_n)$) wird später zur Bestimmung dieser Unbekannten verwendet und in die restlichen $n-1$ Gleichungen eingesetzt (Elimination z.B. von x_1 führt auf $n-1$ Gleichungen $g_i(x_2,\ldots,x_n)$ für $1 \le i \le n-1$ in den $n-1$ Unbekannten $x_2,\ldots,x_n$). Damit ergibt sich ein Gleichungssystem in $n-1$ Unbekannten, welches $n-1$ Gleichungen umfaßt.
  - Durch induktive Fortführung dieser Idee der Elimination ergibt sich schließlich eine einzige Gleichung in einer Unbekannten, die direkt aufgelöt werden kann.
  - Das sukzessive Einsetzen der bereits bestimmten Unbekannten in die expliziten Gleichungen führt auf die Lösung des linearen Gleichungssystems (Rückwärtssubstitution).

- **Beschreibung des Gaußschen Eliminationsverfahrens** unter der Annahme, daß bei der schrittweisen Elimination der Unbekannten die **natürliche Reihenfolge** $x_1, x_2, \ldots, x_n$ gewählt werden kann.
  - Lineares Gleichungssystem

$$A^{(1)}x = b^{(1)}, \qquad A^{(1)} = A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix} \in \mathbb{K}^{n\times n}, \quad b^{(1)} = b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \in \mathbb{K}^n.$$

  - 1. Schritt: Wahl der ersten Zeile von A als **Pivotzeile**. Elimination der Unbekannten x_1 in der i-ten Gleichung durch Subtraktion des $\frac{a_{i1}}{a_{11}}$-Vielfachen der ersten Zeile von der i-ten Zeile für $2 \le i \le n$ führt auf das äquivalente Gleichungssystem

$$A^{(2)}x = b^{(2)}, \qquad A^{(2)} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ 0 & \widetilde{a}_{22} & \ldots & \widetilde{a}_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & \widetilde{a}_{n2} & \ldots & \widetilde{a}_{nn} \end{pmatrix} \in \mathbb{K}^{n\times n}, \quad b^{(2)} = \begin{pmatrix} b_1 \\ \widetilde{b}_2 \\ \vdots \\ \widetilde{b}_n \end{pmatrix} \in \mathbb{K}^n,$$

$$\ell_{i1} = \tfrac{a_{i1}}{a_{11}}, \quad \widetilde{a}_{ij} = a_{ij} - \ell_{i1}\,a_{1j}, \quad \widetilde{b}_i = b_i - \ell_{i1}\,b_1, \qquad 2 \le i \le n, \quad 1 \le j \le n.$$

Theoretische Beschreibung des ersten Eliminationsschritts durch Multiplikation von A mit den elementaren Matrizen $N_{21}(-\ell_{21}), \ldots, N_{n1}(-\ell_{n1})$, d.h. es ist

$$A^{(2)} = N_{n1}(-\ell_{n1}) \cdots N_{21}(-\ell_{21})\, A^{(1)}\,.$$

– Analoge Elimination der Variablen $x_2, \ldots, x_{n-1}$.
Theoretische Beschreibung der weiteren Eliminationsschritte durch Multiplikation mit elementaren Matrizen, beispielsweise

$$A^{(3)} = N_{n2}(-\ell_{n2}) \cdots N_{32}(-\ell_{32})\, A^{(2)}\,.$$

– Nach $n-1$ Eliminationsschritten ergibt sich ein zu $Ax = b$ äquivalentes lineares Gleichungssystem der Form

$$Rx = c$$

mit oberer Dreiecksmatrix $R \in \mathbb{K}^{n\times n}$, dessen Lösung mittels Rückwärtssubstitution bestimmt wird. Da die Matrix A nach Voraussetzung invertierbar ist, ist R ebenfalls invertierbar, d.h. die Diagonalelemente erfüllen $r_{ii} \neq 0$ für $1 \leq i \leq n$.
Theoretische Beschreibung der Eliminationsschritte durch Multiplikation mit elementaren Matrizen

$$R = \underbrace{N_{n,n-1}(-\ell_{n2}) \cdots N_{n2}(-\ell_{n2}) \cdots N_{32}(-\ell_{32})\, N_{n1}(-\ell_{n1}) \cdots N_{21}(-\ell_{21})}_{=L^{-1}}\, A\,.$$

Dies ist äquivalent zur **LR-Zerlegung** bzw. **Dreieckszerlegung** bzw. **LU-Zerlegung** der Matrix A (verwende die Relation für die Inverse von $N_{ij}(\alpha)$ und erhalte damit $L = N_{21}(\ell_{21}) \cdots N_{n1}(\ell_{n1})\, N_{32}(\ell_{32}) \cdots N_{n2}(\ell_{n2}) \cdots N_{n,n-1}(\ell_{n2})$)

$$A = LR\,,$$

$$L = \begin{pmatrix} 1 & & & & \\ \ell_{21} & 1 & & & \\ \ell_{31} & \ell_{32} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ \ell_{n1} & \ell_{n2} & \ldots & \ell_{n,n-1} & 1 \end{pmatrix}, \qquad R = \begin{pmatrix} r_{11} & r_{12} & \ldots & r_{1n} \\ & r_{22} & \ldots & r_{2n} \\ & & \ddots & \vdots \\ & & & r_{nn} \end{pmatrix},$$

vgl. Skriptum, S. 74/75. Die Lösung des linearen Gleichungssystems $Ax = b$ mittels Gaußschem Eliminationsverfahren in natürlicher Reihenfolge entspricht damit der Berechnung der LR-Zerlegung von A

$$Ax = LRx = b \quad \Longleftrightarrow \quad Ly = b, \quad Rx = y$$

und der Lösung des gestaffelten Gleichungssystems $Ly = b$ mittels **Vorwärtssubstitution**

$$Ly = b\,, \qquad y_i = b_i - \sum_{j=1}^{i-1} \ell_{ij} y_j\,, \qquad 1 \leq i \leq n\,,$$

sowie der Lösung des gestaffelten Gleichungssystems $Rx = y$ mittels **Rückwärtssubstitution**

$$Rx = y\,, \qquad x_i = \frac{1}{r_{ii}}\Big(y_i - \sum_{j=i+1}^{n} r_{ij}\, x_j\Big), \qquad 1 \le i \le n\,.$$

Als Pseudo-Code lautet das Gaußsche Eliminationsverfahren beispielsweise (Überschreiben der Koeffizienten von A und b mit den berechneten Koeffizienten von L, R und c)

Eingabedaten : A, b
for k = 1:n-1
    for i = k+1:n
        $A(i,k) = \frac{A(i,k)}{A(k,k)}$
        for j = k+1:n
            $A(i,j) = A(i,j) - A(i,k)\,A(k,j)$
        end
        $b(i) = b(i) - A(i,k)\,b(k)$
    end
end
Zwischenergebnis : A, b
for i = n:-1:1
    $x(i) = b(i)$
    for j = i+1:n
        $x(i) = x(i) - A(i,j)\,x(j)$
    end
    $x(i) = \frac{x(i)}{A(i,i)}$
end

Vgl. **Illustration4_GaussElimination**.

Bemerkungen:

- Üblichweise wird zuerst die LR-Zerlegung von A berechnet und anschließend die Lösung des linearen Gleichungssystems mittels Vorwärtssubstitution und Rückwärtssubstitution berechnet.
- Die Anzahl der zur Berechnung der LR-Zerlegung von $A \in \mathbb{K}^{n\times n}$ benötigten Operationen (Addition oder Multiplikation in $\mathbb{K}$) ist

$$\mathcal{O}\big(\tfrac{2}{3}\, n^3\big)\,.$$

- Günstige Wahl des Pivotelementes, vgl. Abschnitt 4.6.

4.5. Rundungsfehler-Analyse der Gauß-Elimination

- **Vorüberlegungen**: Das in Abschnitt 4.4 besprochene Verfahren zur Lösung eines linearen Gleichungssystems $Ax = b$ mit $A \in \mathbb{K}^{n\times n}$ invertierbar und $b \in \mathbb{K}^n$ umfaßt die folgenden Schritte.

  – LR-Zerlegung von $A = LR$ bzw. Triangulierung von A mittels elementarer Umformungen (beschrieben durch Multiplikation mit geeignet gewählten Matrizen $N_{ij}(\alpha)$).

  – Vorwärtssubstitution zur Berechnung von y aus $Ly = b$.

  – Rückwärtssubstitution zur Berechnung von x aus $Rx = y$.

  Dieses Verfahren ist äquivalent zum Gaußschen Eliminationsverfahren in natürlicher Reihenfolge. Hinsichtlich einer Rundungsfehleranalyse wird verwendet, daß die Koeffizienten der unteren Dreiecksmatrix L und der oberen Dreiecksmatrix R durch folgende Relationen gegeben sind (zeilenweise Berechnung der Koeffizienten z.B. von L, es ist $\ell_{ik} = 0$ für $k \geq i+1$ und $\ell_{ii} = 1$ für $1 \leq i \leq n$)

$$A = LR,$$

$$L = \begin{pmatrix} 1 & & & & \\ \ell_{21} & 1 & & & \\ \ell_{31} & \ell_{32} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ \ell_{n1} & \ell_{n2} & \dots & \ell_{n,n-1} & 1 \end{pmatrix}, \qquad R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ & r_{22} & \dots & r_{2n} \\ & & \ddots & \vdots \\ & & & r_{nn} \end{pmatrix},$$

$$\begin{cases} \ell_{ik} = \frac{1}{r_{kk}} \left(a_{ik} - \sum_{j=1}^{k-1} \ell_{ij} r_{jk} \right), & 1 \leq k \leq i-1, \\ r_{ik} = a_{ik} - \sum_{j=1}^{i-1} \ell_{ij} r_{jk}, & i \leq k \leq n, \end{cases} \qquad 1 \leq i \leq n.$$

  Vgl. **Illustration4_LRZerlegung**.

- **Rundungsfehleranalyse**: Ähnliche Überlegungen wie in Abschnitt 4.3 im Zusammenhang mit der Stabilität der Rückwärtssubstitution zeigen, daß sich unter dem Einfluß von Rundungsfehlern folgende Näherungslösungen beispielsweise für die Koeffizienten der Matrix L ergeben (verwende wieder $\mathrm{rd}(a * b) = (a * b)(1+\varepsilon)$ mit $|\varepsilon| \leq \varepsilon_{\mathrm{mach}}$)

$$\ell_{ik} = \frac{1}{r_{kk}} \left(a_{ik} - \sum_{j=1}^{k-1} \ell_{ij} r_{jk} \right), \qquad 1 \leq k \leq i-1, \quad 1 \leq i \leq n,$$

$$\widetilde{\ell}_{ik} = \frac{1+\widetilde{\varepsilon}}{\widetilde{r}_{kk}} \left(a_{ik} \prod_{\ell=1}^{k-1} (1+\varepsilon_{i\ell}) - \sum_{j=1}^{k-1} \widetilde{\ell}_{ij} \widetilde{r}_{jk} \prod_{\ell=j}^{k-1} (1+\varepsilon_{i\ell})(1+\widetilde{\varepsilon}) \right), \qquad 1 \leq k \leq i-1, \quad 1 \leq i \leq n,$$

mit $|\varepsilon_{i\ell}| \le \varepsilon_{\text{mach}}$ für $1 \le \ell \le k-1$ und $|\widetilde{\varepsilon}| \le \varepsilon_{\text{mach}}$. Mittels Lemma 2.6. folgt damit die Relation (beachte die Abhängigkeit von n)

$$a_{ik} = \sum_{j=1}^{k} \widetilde{\ell}_{ij}\widetilde{r}_{ik}(1+\delta_{ijk}), \qquad |\delta_{ijk}| \le \tfrac{n\varepsilon_{\text{mach}}}{1-n\varepsilon_{\text{mach}}} = \mathcal{O}(\varepsilon_{\text{mach}}), \quad 1 \le i,k \le n.$$

Ähnliche Überlegungen führen auf folgende Abschätzungen für die berechneten Größen (zusätzliche Abhängigkeit von Dimension n)

$$\begin{aligned} |A - \widetilde{L}\widetilde{R}| &\le \mathcal{O}(\varepsilon_{\text{mach}})|\widetilde{L}||\widetilde{R}|, \\ \Delta L\,\widetilde{y} = b - \widetilde{L}\,\widetilde{y}, &\qquad |\Delta L| \le \mathcal{O}(\varepsilon_{\text{mach}})|\widetilde{L}|, \\ \Delta R\,\widetilde{x} = \widetilde{y} - \widetilde{R}\,\widetilde{x}, &\qquad |\Delta R| \le \mathcal{O}(\varepsilon_{\text{mach}})|\widetilde{R}|. \end{aligned}$$

Daraus folgt weiters

$$b = (\widetilde{L}+\Delta L)\,\widetilde{y} = (\widetilde{L}+\Delta L)\,(\widetilde{R}+\Delta R)\,\widetilde{x} = \widetilde{A}\widetilde{x}$$

$$\widetilde{A} - A = (\widetilde{L}+\Delta L)\,(\widetilde{R}+\Delta R) - A = \widetilde{L}\widetilde{R} - A + \Delta L\widetilde{R} + \widetilde{L}\Delta R + \Delta L\Delta R,$$

$$|\widetilde{A} - A| \le \underbrace{|\widetilde{L}\widetilde{R} - A|}_{\le \mathcal{O}(\varepsilon_{\text{mach}})|\widetilde{L}||\widetilde{R}|} + \underbrace{|\Delta L|}_{\le \mathcal{O}(\varepsilon_{\text{mach}})|\widetilde{L}|}|\widetilde{R}| + |\widetilde{L}|\underbrace{|\Delta R|}_{\le \mathcal{O}(\varepsilon_{\text{mach}})|\widetilde{R}|} + \underbrace{|\Delta L\Delta R|}_{\mathcal{O}(\varepsilon_{\text{mach}}^2)|\widetilde{L}||\widetilde{R}|} \le \mathcal{O}(\varepsilon_{\text{mach}})|\widetilde{L}||\widetilde{R}|,$$

$$r = b - A\widetilde{x} = (\widetilde{A} - A)\widetilde{x}.$$

Schlußfolgerungen:

- Die obigen Überlegungen zeigen, daß die mittels LR-Zerlegung und Substitutionen berechnete Näherungslösung $\widetilde{x}$ die exakte Lösung eines linearen Gleichungssystems ist

$$\widetilde{A}\widetilde{x} = b, \qquad |\widetilde{A} - A| \le \mathcal{O}(\varepsilon_{\text{mach}})|\widetilde{L}||\widetilde{R}|.$$

- **Erinnerung**: Der **Satz von Prager und Oettli** besagt, daß eine Näherungslösung $\widetilde{x}$ des linearen Gleichungssystems $Ax = b$ genau dann akzeptabel ist (bzgl U_ε, im strengen Sinn), wenn das Residuum $A\widetilde{x} - b$ die folgende Abschätzung erfüllt

$$|A\widetilde{x} - b| \le \varepsilon\left(|A|\,|\widetilde{x}| + |b|\right).$$

- Nach dem Satz von Prager und Oettli ist die berechnete Näherungslösung $\widetilde{x}$ akzeptabel (bzgl. $U_{\varepsilon_{\text{mach}}}$), wenn das Residuum die Abschätzung

$$|r| = |\widetilde{A} - A|\,|\widetilde{x}| \le \varepsilon_{\text{mach}}\left(|A|\,|\widetilde{x}| + |b|\right)$$

  erfüllt. Insbesondere ist das Verfahren numerisch stabil, wenn

$$|r| = |\widetilde{A} - A|\,|\widetilde{x}| \le \mathcal{O}(\varepsilon_{\text{mach}})|\widetilde{L}||\widetilde{R}||\widetilde{x}| \le \varepsilon_{\text{mach}}\left(|A|\,|\widetilde{x}| + |b|\right),$$

d.h. $|\widetilde{L}||\widetilde{R}| \approx |A|$.
Allerdings gibt es Situation, in denen

$$|\widetilde{L}||\widetilde{R}| \gg |\widetilde{L}\widetilde{R}| \approx |LR| = |A|$$

gilt, und somit das Verfahren numerisch instabil ist. Um dies (teilweise) zu vermeiden, verwendet man eine **Spalten-Pivotsuche** oder **vollständige Pivotsuche**, vgl. Abschnitt 4.6.

4.6. Pivotwahl bei der Gauß-Elimination

- **Situation**: Lösung eines linearen Gleichungssystems

$$Ax = b$$

mit $A \in \mathbb{K}^{n\times n}$ invertierbar und $b \in \mathbb{K}^n$.

Vorbemerkung: Das **Gaußsche Eliminationsverfahren in natürlicher Reihenfolge** benötigt, daß im k-ten Eliminationsschritt das **natürliche Pivotelement**, d.h. das k-te Diagonalelement der Matrix $A^{(k)}$ ungleich Null ist (notwendige Bedingung für Elimination und Rückwärtssubstitution). Wenn diese Bedingung verletzt ist und auch zur Verbesserung der Stabilität des Verfahrens führt man eine **Pivotsuche** durch.

Spalten-Pivotsuche mit Maxiumums-Strategie: Falls das k-te natürliche Pivotelement gleich Null ist, wählt man unter den Koeffizienten a_{ik} für $k+1 \le i \le n$ ein Pivotelement $a_{\ell k} \neq 0$. Bei Berechnungen mit endlicher Genauigkeit beeinflußt die Wahl des Pivotelements das Ergebnis. Bei der Spalten-Pivotsuche mit Maxiumums-Strategie wählt man im k-ten Eliminationsschritt den betragsmäßig größten Koeffizienten in der k-ten Spalte der aktuellen Matrix als Pivotelement, d.h. jenes Element $a_{\ell k}$ mit

$$|a_{\ell k}| = \max\{|a_{ik}| : k \le i \le n\},$$

und vertauscht die entsprechenden Zeilen. Die Spalten-Pivotsuche mit Maxiumums-Strategie sichert, daß die komponentenweise Abschätzung $|L| \le 1$ erfüllt ist.

Bemerkung: Sollte bei der Durchführung des Gaußschen Eliminationsverfahrens mit Spaltenpivotsuche im k-ten Eliminationsschritt der Fall $a_{ik} = 0$ für alle $k \le i \le n$ eintreten, ist die Voraussetzung $A \in \mathbb{K}^{n\times n}$ invertierbar verletzt und man bricht das Gaußsche Eliminationsverfahren ab.

Vollständige Pivotsuche mit Maxiumums-Strategie: Bei der vollständigen Pivotsuche mit Maxiumums-Strategie wählt man im k-ten Eliminationsschritt das betragsmäßig größte Element der Koeffizienten $a_{i\ell}$ für $k \le i, \ell \le n$ und vertauscht die entsprechenden Zeilen und Spalten (entspricht einer Umnummerierung der Unbekannten). Da die vollständige Pivotsuche vergleichsweise aufwendig ist, wird sie selten angewendet.

- **Bemerkung**: Überlegungen in Abschnitt 4.5 haben gezeigt, daß das Gaußsche Eliminationsverfahren bzw. die Lösung des linearen Gleichungssystems mittels LR-Zerlegung numerisch stabil ist, wenn

$$|r| = |\widetilde{A} - A|\,|\widetilde{x}| \le \mathcal{O}(\varepsilon_{\text{mach}})|\widetilde{L}||\widetilde{R}||\widetilde{x}| \le \varepsilon_{\text{mach}}\left(|A|\,|\widetilde{x}| + |b|\right),$$

wobei $\widetilde{L}$ und $\widetilde{R}$ die berechneten Matrizen bei Durchführung der LR-Zerlegung bezeichnen.

Schlußfolgerung: Bei günstiger Wahl der Pivotelemente sind die Koeffizienten der berechneten Matrix $|\widetilde{L}|\,|\widetilde{R}| \approx |L|\,|R|$ minimal.

Aber! Es gibt Situationen in denen die Wahl des betragsmäßig größten Elementes nicht ideal ist.

Beispiel:

– Für den Spezialfall $A \in \mathbb{R}^{2\times 2}$ führt das Gaußsche Eliminationsverfahren in natürlicher Reihenfolge (unter Annahme $a \neq 0$) auf die LR-Zerlegung

$$A = LR,$$
$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad L = \begin{pmatrix} 1 & 0 \\ \frac{c}{a} & 1 \end{pmatrix}, \quad R = \begin{pmatrix} a & b \\ 0 & \frac{ad-bc}{a} \end{pmatrix},$$
$$|A| = |LR| = \begin{pmatrix} |a| & |b| \\ |c| & |d| \end{pmatrix}, \qquad |L|\,|R| = \begin{pmatrix} |a| & |b| \\ |c| & \frac{|b||c|}{|a|} + \frac{|ad-bc|}{|a|} \end{pmatrix}.$$

Eine kurze Rechnung ergibt (Fallunterscheidung $|a|\,|b| = ab$ oder $|a|\,|b| = -ab$)

$$\begin{aligned} d = 0: \qquad & \alpha = \tfrac{|b||c|}{|a|} + \tfrac{|ad-bc|}{|a|} = 2\,\tfrac{|b||c|}{|a|}, \\ d \neq 0: \qquad & \alpha = \tfrac{|b||c|}{|a|} + \tfrac{|ad-bc|}{|a|} = |d|\left(\tfrac{|b||c|}{|a||d|} + \tfrac{|ad-bc|}{|a||d|}\right) = |d|\left(\tfrac{|b||c|}{|a||d|} + \left|\tfrac{ad}{|a||d|} - \tfrac{bc}{|a||d|}\right|\right) \\ & = |d|\left(\tfrac{|b||c|}{|a||d|} + \left|1 - \tfrac{bc}{ad}\right|\right) = |d|\left(|x| + |1-x|\right), \qquad x = \tfrac{bc}{ad}. \end{aligned}$$

Die Funktion $f : \mathbb{R} \to \mathbb{R} : x \mapsto |1-x| + |x|$ hat für $0 \leq x \leq 1$ den konstanten Wert 1 und wächst ansonsten an. Falls $0 \leq \frac{bc}{ad} \leq 1$ oder $0 \leq \left|\frac{bc}{ad}\right| \leq 1$, folgt $\alpha \leq 3$ und somit

$$|bc| \leq |ad|: \qquad |L|\,|R| \approx |LR|,$$

d.h. das Verfahren ist numerisch stabil. Falls hingegen

$$|bc| \gg |ad|: \qquad |L|\,|R| \gg |LR|$$

ist das Verfahren numerisch instabil.

– Beachte, daß das Stabilitätskriterium **skalierungsinvariant** ist, d.h. Skalierungen der Zeilen oder Spalten von A bewirken keine Änderung der Größe $\frac{bc}{ad}$, denn

$$DA = \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} d_1 a & d_1 b \\ d_2 c & d_2 d \end{pmatrix}, \quad \frac{(d_1 b)\,(d_2 c)}{(d_1 a)\,(d_2 d)} = \frac{bc}{ad},$$
$$AD = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix} = \begin{pmatrix} d_1 a & d_2 b \\ d_1 c & d_2 d \end{pmatrix}, \quad \frac{(d_2 b)\,(d_1 c)}{(d_1 a)\,(d_2 d)} = \frac{bc}{ad}.$$

– Obige Überlegungen zeigen, daß bei einer günstigen Pivolwahl die natürliche Reihenfolge der Zeilen belassen wird, falls $|bc| \leq |ad|$, und ansonsten die beiden Zeilen der Matrix vertauscht werden. Es gibt jedoch Situationen, in denen eine Spaltenpivotsuche (nicht skalierungsinvariant) zu einer ungünstigen Pivolwahl führt.

– **Beispiel von Dahlquist und Björck**, vgl. **Illustration4_Pivotwahl**.

– *Bis heute ist keine Methode bekannt, wie die in numerischer Hinsicht beste Pivotwahl getroffen werden kann — bzw. keine praktikable Skalierungsmethode bekannt, bei der die Maximums-Strategie immer gut wäre.*

Übliche Strategie: Äquilibrierung der Matrix A, d.h. Skalierung von A

$$A \to \widehat{A} = DA = \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{pmatrix} \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix} = \begin{pmatrix} d_1 a_{11} & \dots & d_1 a_{1n} \\ \vdots & & \vdots \\ d_n a_{n1} & \dots & d_n a_{nn} \end{pmatrix}$$

derart, daß die Zeilenbetragssumme konstant ist

$$\sum_{j=1}^{n} |\widehat{a}_{ij}| = 1\,, \quad 1 \le i \le n \quad \Longleftrightarrow \quad d_i = \Big(\sum_{j=1}^{n} |a_{ij}| \Big)^{-1}, \quad 1 \le i \le n$$

und dann Anwendung der Spalten-Pivotsuche mit Maximums-Strategie. Dies entspricht der folgenden Pivotwahl im k-ten Eliminationsschritt

$$\big|a_{\ell k}^{(k)}\big| = \max_{1 \le k \le i \le n} \frac{d_i}{d_\ell} \big|a_{ik}^{(k)}\big|\,.$$

5. Lineare Ausgleichsrechnung

- **Problemstellung**: Näherungsweise Lösung eines **überbestimmten linearen Gleichungssystems** (mehr Gleichungen als Unbekannte, im Allgemeinen existiert keine Lösung)

$$Ax = b, \qquad A \in \mathbb{K}^{m \times n}, \quad x \in \mathbb{K}^n, \quad b \in \mathbb{K}^m, \qquad m > n.$$

  Bestimme eine **Lösung des linearen Ausgleichsproblems**, d.h. eine Lösung im Sinn der **kleinsten Fehlerquadrate** (Gauß)

$$\|Ax - b\|_2 \xrightarrow{!} \min.$$

  Falls $\|Ax - b\|_2 = 0 \Leftrightarrow Ax = b$ erfüllt ist, ist x tatsächlich eine Lösung des linearen Gleichungssystems.

5.1. Ein Beispiel

- **Polynominterpolation**: Bestimme das eindeutig bestimmte Polynom vom Grad $\leq m-1$

$$p : \mathbb{K} \to \mathbb{K} : x \mapsto \sum_{i=0}^{m-1} c_{i+1}\, x^i$$

durch m vorgegebene Datenpunkte $(x_j, y_j) \in \mathbb{K} \times \mathbb{K}$ für $1 \leq j \leq m$, d.h. bestimme die Koeffizienten $c \in \mathbb{K}^m$ des Polynoms durch Lösung des linearen Gleichungssystems

$$p(x_j) = c_1 + c_2\, x_j + \cdots + c_m\, x_j^{m-1} = y_j\,, \quad 1 \leq j \leq m\,,$$

$$\begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{m-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_m & x_m^2 & \dots & x_m^{m-1} \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_m \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}.$$

Vgl. **Illustration5_PolynomInterpolation** (Auswerten mittels Horner-Schema) .

Vgl. Abbildung, Skriptum, S. 86. Für $m > 10$ und äquidistante Stützstellen sind Oszillationen des Interpolationspolynoms insbesondere an den Intervallenden typisch.

Bemerkung: Im Allgemeinen sind Konditionszahlen von **Vandermonde-Matrizen** *groß*. Eine bessere Alternative zur Berechnung der Koeffizienten des Interpolationspolynoms verwendet die Darstellung des Interpolationspolynoms nach Newton, vgl. Numerische Mathematik II.

- **Approximation mittels Ausgleichspolynom**: Bestimme ein Polynom

$$p : \mathbb{K} \to \mathbb{K} : x \mapsto \sum_{i=0}^{n-1} c_{i+1}\, x^i\,, \qquad n < m\,,$$

durch m vorgegebene Datenpunkte $(x_j, y_j) \in \mathbb{K} \times \mathbb{K}$ für $1 \leq j \leq m$, d.h. bestimme die Koeffizienten $c \in \mathbb{K}^n$ des Polynoms durch Lösung des linearen Ausgleichsproblems

$$\|Ac - y\|_2 = \sum_{j=1}^{m} \left| p(x_j) - y_j \right|^2 \xrightarrow{!} \min\,,$$

$$A = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_m & x_m^2 & \dots & x_m^{n-1} \end{pmatrix}, \quad c = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}.$$

Vgl. Abbildung, Skriptum, S. 87. Glatterer Verlauf des approximierenden Polynoms.

5.2. Normalengleichungen

- **Situation**: Näherungsweise Lösung des linearen Ausgleichsproblems

$$\|Ax-b\|_2 \xrightarrow{!} \min, \qquad A\in\mathbb{K}^{m\times n}, \quad x\in\mathbb{K}^n, \quad b\in\mathbb{K}^m, \qquad m>n.$$

- **Mittels Differentialrechung**:
  - Für $\mathbb{K}=\mathbb{C}$ betrachte das äquivalente reelle lineare Gleichungssystem der Dimension $2m\times 2n$.
  - Verwende die äquivalente Formulierung

$$\|Ax-b\|_2^2 \xrightarrow{!} \min.$$

Bestimmte das Minimum mittels Differenzieren (betrachte z.B. den Fall $m=n=2$, für eine symmetrische Matrix $B\in\mathbb{R}^{2\times 2}$ bestimme die erste und zweite Ableitung von $x^TBx=b_{11}x_1^2+2b_{12}x_1x_2+b_{22}x_2^2$, beachte $f(x)=f(x_1,\dots,x_n)\in\mathbb{R}$ und folglich $f'(x)=(\partial_{x_1}f(x),\dots,\partial_{x_n}f(x))\in\mathbb{R}^{1\times n}$)

$$\begin{aligned} f(x) &= \|Ax-b\|_2^2=(Ax-b)^T(Ax-b)=(x^TA^T-b^T)(Ax-b)\\ &= x^TA^TAx-b^TAx-x^TA^Tb+b^Tb \underset{x^TA^Tb=b^TAx\in\mathbb{R}}{=} x^TA^TAx-2b^TAx+b^Tb,\\ f'(x) &= 2x^TA^TA-2b^TA\overset{!}{=}0 \underset{\text{Transponieren}}{\iff} A^TAx=A^Tb,\\ f''(x) &= A^TA \quad \text{positiv semi-definit}. \end{aligned}$$

Folglich erfüllt die Lösung des linearen Ausgleichsproblems die **Normalengleichungen**

$$A^TAx=A^Tb$$

mit symmetrischer Matrix A^TA.

  - Wenn alle Spalten von A linear unabhängig sind, d.h. die Matrix A vollen Rang hat, ist die Lösung der Normalengleichungen eindeutig bestimmt und die Hessematrix $f''(x)=A^TA$ positiv definit.
    Denn: Verwende $x^Tf''(x)x=\|Ax\|_2^2=0\Leftrightarrow Ax=0\Leftrightarrow x=0$ bzw. $x^Tf''(x)x>0$ für alle $x\neq 0$ und $A^TAx=0\Leftrightarrow A^Ty=0, y=Ax\Leftrightarrow 0=y=Ax\Leftrightarrow x=0$. ◇

- **Mittels Ergebnissen der Linearen Algebra**:
  - Es bezeichnen $a_1,\dots,a_n$ die Spalten der Matrix

$$A=\left(a_1\,\middle|\,\cdots\,\middle|\,a_n\right)\in\mathbb{K}^{m\times n}.$$

Früheren Überlegungen zeigten

$$\begin{aligned} &\text{Gleichungssystem}\quad Ax=b\quad\text{lösbar}\quad\Longleftrightarrow\quad b\in\mathscr{R}_A=\langle a_1,\dots,a_n\rangle\qquad\text{bzw.}\\ &\text{Gleichungssystem}\quad Ax=b\quad\text{nicht lösbar}\quad\Longleftrightarrow\quad b\notin\mathscr{R}_A=\langle a_1,\dots,a_n\rangle. \end{aligned}$$

– Die Zerlegung von $\mathbb{K}^m$ in den Unterraum $\mathscr{R}_A$ und den dazu orthogonalen Unterraum U führt auf (wäre $w = b - A\hat{x}$ nicht orthogonal auf $\mathscr{R}_A$, wäre $\|A\hat{x} - b\|_2$ wegen der Dreiecksungleichung nicht minimal)

$$\min_{x\in\mathbb{K}^n} \|Ax - b\|_2 = \hat{x} \quad \Longleftrightarrow \quad b = v + w, \quad v = A\hat{x} \in \mathscr{R}_A, \quad w \in U, \text{ d.h. } A^* w = 0.$$

Multiplikation von $A\hat{x} + w = b$ mit A^* führt auf die **Normalengleichungen**

$$\min_{x\in\mathbb{K}^n} \|Ax - b\|_2 = \hat{x} \quad \Longrightarrow \quad A^*A\hat{x} = A^*b.$$

– Es bleibt zu zeigen, daß für eine Lösung $z \in \mathbb{K}^n$ der Normalengleichungen, d.h. $A^*Az = A^*b$, und einen beliebigen Vektor $x \in \mathbb{K}^n$ die Abschätzung

$$\|Az - b\|_2 \le \|Ax - b\|_2 \quad \Longleftrightarrow \quad \|Az - b\|_2^2 \le \|Ax - b\|_2^2$$

gilt. Unter der Annahme $\|Ax - b\|_2^2 \le \|Az - b\|_2^2$ für ein $x \in \mathbb{K}^n$ folgt

$$\begin{aligned}
x^*A^*Ax - b^*Ax - x^*A^*b + b^*b &= (x^*A^* - b^*)(Ax - b) = \|Ax - b\|_2^2 \\
&\le \|Az - b\|_2^2 = z^*A^*Az - b^*Az - z^*A^*b + b^*b \\
\Longleftrightarrow \quad x^*A^*Ax - \underbrace{b^*A}_{=z^*A^*A}x - x^*\underbrace{A^*b}_{=A^*Az} &\le z^*A^*Az - \underbrace{b^*A}_{=z^*A^*A}z - z^*\underbrace{A^*b}_{=A^*Az} \\
\Longleftrightarrow \quad x^*A^*Ax - z^*A^*Ax - x^*A^*Az + z^*A^*Az &\le 0 \\
\Longleftrightarrow \quad \|A(x - z)\|_2^2 = (x - z)^*A^*A(x - z) &\le 0 \\
\Longleftrightarrow \quad Ax = Az,
\end{aligned}$$

d.h. für Lösungen der Normalengleichungen und des linearen Ausgleichsproblems gilt die Identität $Az = Ax$. Insbesondere ist also das Residuum $Ax - b$ von Lösungen zum linearen Ausgleichsproblem eindeutig bestimmt.

– Falls die Matrix A vollen Rang hat, ist die Lösung des linearen Ausgleichsproblems (wegen $Az = Ax \Leftrightarrow x = y$) eindeutig bestimmt.

- **Resultat zur Lösung des linearen Ausgleichsproblems** (Satz 5.1): Das lineare Ausgleichsproblem ist äquivalent zu den Normalengleichungen

$$\min_{x\in\mathbb{K}^n} \|Ax - b\|_2 = \hat{x} \quad \Longleftrightarrow \quad A^*A\hat{x} = A^*b.$$

Das Residuum ist eindeutig bestimmt, d.h. für zwei Lösungen $x_1, x_2 \in \mathbb{K}^n$ des linearen Ausgleichsproblems gilt $b - Ax_1 = b - Ax_2$. Falls die Matrix A vollen Rang hat, ist die Lösung des linearen Ausgleichsproblems eindeutig bestimmt.

- **Kondition des linearen Ausgleichsproblems**: Es bezeichne $\hat{x}$ die Lösung des (eindeutig lösbaren) linearen Ausgleichsproblems und $\hat{r}$ das entsprechende Residuum

$$\hat{x} = \min_{x\in\mathbb{K}^n} \|Ax - b\|_2, \qquad \hat{r} = A\hat{x} - b.$$

Bei Änderungen α und β der Eingabedaten A und B gilt für den Fehler der zugehörigen Lösung $\widehat{x}+\xi$ und des entsprechenden Residuums $\widehat{r}+\varrho = (A+\alpha)\,(\widehat{x}+\xi)-(b+\beta)$ folgende Abschätzung (ohne Begründung)

$$\|\xi\|_2 \le \frac{\kappa_1}{\|A\|_2} F, \qquad \|\varrho\|_2 \le F,$$
$$F = \|\alpha\|_2\,\|\widehat{x}\|_2 + \|\beta\|_2 + \kappa_1 \frac{\|\alpha\|_2\,\|\widehat{r}\|_2}{\|A\|_2}, \qquad \kappa_1 = \frac{\kappa(A)}{1-\kappa(A)\,\frac{\|\alpha\|_2}{\|A\|_2}},$$

Es ist zu beachten, daß im Vergleich mit einem linearen Gleichungssystem beim linearen Ausgleichsproblem das Quadrat der Kondition der Matrix A (jedoch in Kombination mit dem Residuum) auftritt.

5.3. Cholesky-Zerlegung

- **Vorbemerkung**: Falls die Matrix $A \in \mathbb{K}^{m\times n}$ mit $m > n$ vollen Rang hat, d.h. die Spalten von A linear unabhängig sind, ist die Lösung der Normalengleichungen

$$A^* A x = A^* b$$

eindeutig bestimmt. Beachte, daß die quadratische Matrix $B = A^* A \in \mathbb{K}^{n\times n}$ selbstadjungiert und positiv definit ist

$$B^* = (A^* A)^* = A^* A^{**} = A^* A = B\,, \qquad x^* B x = x^* A^* A x = \|Ax\|_2^2 > 0 \quad \text{für } 0 \neq x \in \mathbb{K}^n\,.$$

Zur Lösung der Normalengleichungen werden üblicherweise

  - das Cholesky-Verfahren (vgl. Abschnitt 5.3) oder
  - Orthogonaltransformationen mittels Householder-Reflexionen (vgl. Abschnitt 5.4)

verwendet.

- **Situation**: Die Durchführung des Gaußschen Eliminationsverfahrens (Implementierung) bzw. der LR-Zerlegung (theoretische Überlegungen) unter den Voraussetzungen

  - $A \in \mathbb{K}^{n\times n}$ selbstadjungiert, d.h. $A^* = A$,
  - A positiv definit, d.h. $x^* Ax > 0$ für alle $0 \neq x \in \mathbb{K}^n$,

führt auf die **Cholesky-Zerlegung** von A. Die effiziente Implementierung der Cholesky-Zerlegung einer Matrix beruht auf der direkten Berechnung der LR-Zerlegung der Matrix (s.u.).

Resultat: Unter den obigen Voraussetzungen an die Matrix A kann das Gaußsche Eliminationsverfahren in natürlicher Reihenfolge durchgeführt werden. Die Eigenschaften der Selbstadjungiertheit und positiven Definitheit bleiben in den entstehenden Teilmatrizen erhalten.
Denn:

  - Im ersten Eliminationsschritt ist die Pivotwahl a_{11} möglich (Anwendung der positiven Definitheit von A mit $x = e_1$)

$$a_{11} = e_1^* A e_1 > 0\,.$$

Elimination in der ersten Spalte von (wegen $A^* = A$ hat A die angegebene Form)

$$A = A^{(1)} = \begin{pmatrix} a_{11} & \alpha^* \\ \alpha & \beta \end{pmatrix}, \qquad \alpha \in \mathbb{K}^{n-1}\,, \quad \beta^* = \beta \in \mathbb{K}^{(n-1)\times(n-1)}\,,$$

führt auf (z.B. Transformation 2. Zeile → 2. Zeile $-\frac{\alpha_1}{a_{11}}\times$ 1. Zeile)

$$A^{(2)} = \begin{pmatrix} a_{11} & \alpha^* \\ 0 & B^{(2)} \end{pmatrix}, \qquad B^{(2)} = \beta - \tfrac{1}{a_{11}}\,\alpha\alpha^* \in \mathbb{K}^{(n-1)\times(n-1)}.$$

Die entstehende Teilmatrix $B^{(2)}$ ist offensichtlich selbstadjungiert ($a_{11} \in \mathbb{R}$)

$$\left(B^{(2)}\right)^* = \left(\beta - \tfrac{1}{a_{11}}\,\alpha\alpha^*\right)^* = \beta^* - \tfrac{1}{a_{11}}\,\alpha^{**}\alpha^* = \beta - \tfrac{1}{a_{11}}\,\alpha\alpha^* = B^{(2)}.$$

Es bleibt zu zeigen, daß $B^{(2)}$ positiv definit ist, d.h. für alle $0 \neq \xi \in \mathbb{K}^{n-1}$ gilt

$$\xi^* B^{(2)}\xi = \xi^*\left(\beta - \tfrac{1}{a_{11}}\,\alpha\alpha^*\right)\xi = \xi^*\beta\xi - \tfrac{1}{a_{11}}\,\xi^*\alpha\,\alpha^*\xi > 0.$$

Aus der positiven Definitheit von A folgt

$$\begin{aligned} 0 < x^*Ax &= \begin{pmatrix} x_1 & \xi^* \end{pmatrix}\begin{pmatrix} a_{11} & \alpha^* \\ \alpha & \beta \end{pmatrix}\begin{pmatrix} x_1 \\ \xi \end{pmatrix} = \begin{pmatrix} x_1 & \xi^* \end{pmatrix}\begin{pmatrix} a_{11}x_1 + \alpha^*\xi \\ x_1\alpha + \beta\xi \end{pmatrix} \\ &= a_{11}x_1^2 + x_1(\alpha^*\xi + \xi^*\alpha) + \xi^*\beta\xi, \qquad 0 \neq x = \begin{pmatrix} x_1 \\ \xi \end{pmatrix} \in \mathbb{K}^n, \end{aligned}$$

und speziell mit $x_1 = -\frac{1}{a_{11}}\,\alpha^*\xi$ ergibt sich die Behauptung

$$\begin{aligned} &0 < a_{11}x_1^2 + x_1(\alpha^*\xi + \xi^*\alpha) + \xi^*\beta\xi = \tfrac{1}{a_{11}}\,(\alpha^*\xi)^2 - \tfrac{1}{a_{11}}\,\alpha^*\xi(\alpha^*\xi + \xi^*\alpha) + \xi^*\beta\xi \\ &= -\tfrac{1}{a_{11}}\,\alpha^*\xi\,\xi^*\alpha + \xi^*\beta\xi = \xi^*B^{(2)}\xi. \end{aligned}$$

Für den Fall $\mathbb{K} = \mathbb{R}$ ist dies motiviert durch (wobei $z = \alpha^*\xi$, $(\Re z)^2 = |z|^2$ für $z \in \mathbb{R}$)

$$\begin{aligned} a_{11}x_1^2 + 2\,\Re z\,x_1 = -\tfrac{1}{a_{11}}\,|z|^2 \quad &\Longleftrightarrow \quad x_1^2 + \tfrac{2}{a_{11}}\,\Re z\,x_1 + \tfrac{1}{a_{11}^2}\,|z|^2 = 0 \\ &\Longleftrightarrow \quad x_1 = -\tfrac{1}{a_{11}}\,\Re z \pm \tfrac{1}{a_{11}}\sqrt{(\Re z)^2 - |z|^2} = -\tfrac{1}{a_{11}}\,\alpha^*\xi. \end{aligned}$$

– Die obigen Überlegungen für A lassen sich direkt auf $B^{(2)}$ anwenden.

Mittels Induktion ergibt sich die Behauptung des Resultates. ◇

Bemerkungen:

– Für selbstadjungierte und positiv definite Matrizen ist das Gaußsche Eliminationsverfahren in natürlicher Reihenfolge ein numerisch stabiler Algorithmus (ohne Begründung).

– Da sämtliche entstehende Teilmatrizen selbstadjungiert sind, ist es ausreichend, die Koeffizienten in und oberhalb der Diagonale abzuspeichern, z.B. für $n = 3$ gilt

$$\begin{aligned} A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = A^* = \begin{pmatrix} \overline{a_{11}} & \overline{a_{21}} & \overline{a_{31}} \\ \overline{a_{12}} & \overline{a_{22}} & \overline{a_{32}} \\ \overline{a_{13}} & \overline{a_{23}} & \overline{a_{33}} \end{pmatrix} \\ \Longleftrightarrow \quad a_{11}, a_{22}, a_{33} \in \mathbb{R}, \quad a_{21} = \overline{a_{12}}, \quad a_{31} = \overline{a_{13}}, \quad a_{32} = \overline{a_{23}}. \end{aligned}$$

Die Anzahl der zur Durchführung des Gaußschen Eliminationsverfahren bzw. äquivalent dazu die Anzahl der zur Berechnung der LR-Zerlegung von $A \in \mathbb{K}^{n\times n}$ benötigten Operationen (Addition oder Multiplikation in $\mathbb{K}$) ist somit (vergleiche dies mit dem Aufwand für die LR-Zerlegung bzw. QR-Zerlegung einer allgemeinen quadratischen Matrix)

$$\mathcal{O}\left(\tfrac{1}{3}\,n^3\right).$$

Rationale Cholesky-Zerlegung und Cholesky-Zerlegung:

* Die geforderte Selbstadjungiertheit von $A \in \mathbb{K}^{n\times n}$ und der Ansatz

$$R = DS, \qquad D = \begin{pmatrix} R_{11} & & \\ & \ddots & \\ & & R_{nn} \end{pmatrix}, \quad S = \begin{pmatrix} 1 & S_{12} & \dots & S_{1n} \\ & \ddots & & \vdots \\ & & \ddots & S_{n-1,n} \\ & & & 1 \end{pmatrix},$$

  zeigt die Relation $S = L^*$. Somit führt die LR-Zerlegung von A auf die **rationale Cholesky-Zerlegung** von A

$$A = LDL^*,$$

$$D = \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{pmatrix} \in \mathbb{R}^{n\times n}, \quad d_i > 0, \quad 1 \le i \le n,$$

$$L = \begin{pmatrix} 1 & & & \\ L_{21} & 1 & & \\ \vdots & \ddots & \ddots & \\ L_{n1} & \dots & L_{n,n-1} & 1 \end{pmatrix} \in \mathbb{K}^{n\times n}.$$

* Setzt man weiters

$$\widehat{L} = LD^{\frac{1}{2}}, \qquad D^{\frac{1}{2}} = \begin{pmatrix} \sqrt{d_1} & & \\ & \ddots & \\ & & \sqrt{d_n} \end{pmatrix} \in \mathbb{R}^{n\times n},$$

  ergibt sich die **Cholesky-Zerlegung** von A

$$A = \widehat{L}\widehat{L}^*,$$

$$\widehat{L} = \begin{pmatrix} \widehat{L}_{11} & & \\ \vdots & \ddots & \\ \widehat{L}_{n1} & \dots & \widehat{L}_{nn} \end{pmatrix} \in \mathbb{K}^{n\times n}.$$

  Die Cholesky-Zerlegung berechnet man direkt durch **sukzessive Bestimmung der Koeffizienten** von $\widehat{L}$.

Vgl. **Illustration5_Cholesky**.

- **Anwendung der Cholesky-Zerlegung** zur Lösung der Normalengleichungen

$$A^* Ax = \widehat{L}\widehat{L}^* x = A^* b.$$

Die Berechnung der Cholesky-Zerlegung von $B = A^* A$ erfolgt gleichzeitig mit der Berechnung von $A^* A$. Ebenso wird die Berechnung von $A^* b$ gleichzeitig mit der Vorwärtssubstitution $\widehat{L}y = A^* b$ ausgeführt.

Bemerkung: Ein Nachteil des Cholesky-Verfahrens zur Lösung der Normalengleichungen liegt darin, daß die Konditionszahl der Matrix $B = A^* A$, d.h. das Quadrat der Konditionszahl der Matrix A auftritt. Falls die Gleichungen nahezu konsistent sind, ist das Residuum klein. In diesem Fall dürften sich Rundungsfehler wie $\kappa(A)$ (aber nicht wie $\kappa(A)^2$) auswirken, d.h. in diesem Fall ist der Algorithmus numerisch instabil. Eine stabile Alternative wird in Abschnitt 5.4 angegeben.

5.4. Lösung über Orthogonaltransformationen

- **Situation**: Es sei $A \in \mathbb{K}^{m\times n}$ mit $m > n$ eine Matrix von vollem Rang, d.h. die Spalten von A seien linear unabhängig.

  Überlegungen: Die Triangulierung der Matrix $A \in \mathbb{K}^{m\times n}$ mittels Householder-Reflexionen $T_1, \ldots, T_n \in \mathbb{K}^{m\times m}$ führt auf ein äquivalentes lineares Gleichungssystem

$$Ax = b \quad \Longleftrightarrow \quad Rx = \underbrace{T_n \cdots T_1}_{=Q^*} Ax = T_n \cdots T_1 b = c,$$

  vgl. Abschnitt 4.2. Diese Transformation entspricht der Bestimmung der (vollen) QR-Zerlegung von A mit unitärer Matrix $Q = (T_n \cdots T_1)^* = T_1^* \cdots T_n^* = T_1 \cdots T_n \in \mathbb{K}^{m\times m}$ (beachte die Selbstadjungiertheit von T_i für $1 \le i \le n$) und oberer Dreiecksmatrix $R \in \mathbb{K}^{m\times n}$

$$A = QR,$$

$$Q = T_1 \cdots T_n \in \mathbb{K}^{m\times m}, \quad Q^*Q = I, \qquad R = \begin{pmatrix} \widehat{R} \\ 0 \end{pmatrix} \in \mathbb{K}^{m\times n}.$$

  Da die Transformation mittels einer unitären Matrix die euklidische Norm erhält

$$\|Ax - b\|_2 = \|Q^*Ax - Q^*b\|_2 = \|Rx - Q^*b\|_2,$$

  gilt für die Lösung des linearen Ausgleichsproblems

$$\|Ax - b\|_2 \xrightarrow{!} \min \quad \Longleftrightarrow \quad \|Rx - Q^*b\|_2 \xrightarrow{!} \min.$$

  Dies entspricht (wobei $R \in \mathbb{K}^{m\times n}$, $\widehat{R} \in \mathbb{K}^{n\times n}$, $x \in \mathbb{K}^n$, $c = Q^*b \in \mathbb{K}^m$, $\widehat{c} \in \mathbb{K}^n$, $\widetilde{c} \in \mathbb{K}^{m-n}$)

$$Rx - c = \begin{pmatrix} \widehat{R} \\ 0 \end{pmatrix} x - \begin{pmatrix} \widehat{c} \\ \widetilde{c} \end{pmatrix} = \begin{pmatrix} \widehat{R}x - \widehat{c} \\ -\widetilde{c} \end{pmatrix},$$

$$\|Rx - Q^*b\|_2 = \|\widehat{R}x - \widehat{c}\|_2 + \|\widetilde{c}\|_2 \xrightarrow{!} \min \quad \Longleftrightarrow \quad \|\widehat{R}x - \widehat{c}\|_2 \xrightarrow{!} \min.$$

  Aufgrund der geforderten Rangbedingung an A ist die quadratische Teilmatrix $\widehat{R} \in \mathbb{K}^{n\times n}$ invertierbar. Deshalb entspricht die Lösung des linearen Ausgleichsproblems gerade der Lösung eines linearen Gleichungssystems ($\|\widehat{R}x - \widehat{c}\|_2 = 0 \Leftrightarrow \widehat{R}x = \widehat{c}$)

$$\|Ax - b\|_2 \xrightarrow{!} \min \quad \Longleftrightarrow \quad \widehat{R}x = \widehat{c}.$$

6. Eigenwerte und SVD (Überblick)

- Die näherungsweise **Lösung des Eigenwertproblems** ist eine wichtige Grundaufgabe der Numerischen Linearen Algebra und findet beispielsweise Anwendung bei
  - Verfahren zur Lösung gewöhnlicher Differentialgleichungen, beispielsweise bei Stabilitätsuntersuchungen für dynamische Systeme,
  - Verfahren zur Lösung partieller Differentialgleichungen, beispielsweise bei der Berechnung von *Basislösungen* (Separationsansatz, Fourieranalyse, Stationäre Lösungen).

- **Gekoppelte Schwingungen**, vgl. Skriptum, S. 93.

  Quantenmechanische Vielteilchenzustände, vgl. Berechnungen von A. Läuchli am Supercomputer MACH der Universitäten Innsbruck und Linz.

6.1. Theoretischer Hintergrund

- **Eigenwerte und Eigenvektoren** (Definition 6.1): Für eine quadratische Matrix $A \in \mathbb{K}^{n\times n}$ heißt $\lambda \in \mathbb{C}$ ein **Eigenwert** von A und $0 \neq v \in \mathbb{C}^n$ ein zugehöriger **Eigenvektor**, wenn folgende Relation gilt

$$Av = \lambda v\,.$$

  Der zum Eigenwert λ zugehörige **Eigenraum** ist gegeben durch (Unterraum von $\mathbb{C}^n$ durch Hinzunahme von $v = 0$)

$$\mathcal{N}_{A-\lambda I} = \left\{v \in \mathbb{C}^n : (A-\lambda I)\, v = 0\right\}.$$

- **Bemerkungen**:
  - Für theoretische Überlegungen wird verwendet, daß die Eigenwerte von A Nullstellen des **charakteristischen Polynoms** $\chi : \mathbb{K} \to \mathbb{K}$ sind (wegen $Av = \lambda v$, $v \neq 0 \Leftrightarrow (A-\lambda I)\, v = 0$, $v \neq 0 \Leftrightarrow \det(A-\lambda I) = 0$)

$$\lambda \text{ Eigenwert von } A \quad \Longleftrightarrow \quad \chi(\lambda) = 0\,, \quad \chi(\lambda) = \det(A-\lambda I) = \sum_{i=0}^{n} c_i\, \lambda^i\,, \quad c_n = (-1)^n\,.$$

    Der Fundamentalsatz der Algebra besagt, daß ein Polynom vom Grad n mit Koeffizienten in $\mathbb{K}$ genau n komplexe Nullstellen besitzt (mit Vielfachheit gezählt) und folglich besitzt eine Matrix $A \in \mathbb{K}^{n\times n}$ genau n komplexe Eigenwerte (mit Vielfachheit gezählt).

    Die Vielfachheit eines Eigenwertes bezeichnet man als **algebraische Multiplizität** $\mu(\lambda)$ des Eigenwertes und die Dimension des zugehörigen Eigenraumes als seine **geometrische Multiplizität** $\nu(\lambda)$. Es gilt $\nu(\lambda) \leq \mu(\lambda)$ (s.u.).

    Beispiel: Die Matrix

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \in \mathbb{R}^{2\times 2}$$

    besitzt den Eigenwert $\lambda = 1 \in \mathbb{R}$ mit algebraischer Multiplizität $\mu(1) = 2$. Wegen

$$(A-\lambda I)\, v = 0\,, \quad v \neq 0 \quad \Longleftrightarrow \quad \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} v_2 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \Longleftrightarrow \quad v = v_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad v_1 \neq 0$$

    sind alle Vielfachen des ersten Standardbasisvektors zugehörige Eigenvektoren. Folglich ist der zugehörige Eigenraum gegeben durch

$$\mathcal{N}_{A-\lambda I} = \langle e_1 \rangle \subset \mathbb{R}^2\,,$$

    d.h. die geometrische Multiplizität des Eigenwertes ist

$$\nu(1) = 1 \leq \mu(1) = 2\,.$$

– Die Menge aller Eigenwerte einer Matrix $A \in \mathbb{K}^{n\times n}$ heißt **Spektrum** von A und der betragsmäßig größte Eigenwert gibt den **Spektralradius** der Matrix an

$$\sigma = \{\lambda \in \mathbb{C} : \lambda \text{ Eigenwert von } A\}, \qquad \varrho = \sup\{|\lambda| : \lambda \in \sigma\}.$$

– Falls für alle Eigenwerte der Matrix $A \in \mathbb{K}^{n\times n}$ algebraische und geometrische Multiplizität übereinstimmen

$$\mu(\lambda) = \nu(\lambda),$$

gibt es eine Basis $v_1, \dots, v_n$ des $\mathbb{C}^n$, die von zugehörigen Eigenvektoren gebildet wird. Wegen

$$AV = V\Lambda,$$

$$A\underbrace{\left(v_1 \middle| \cdots \middle| v_n\right)}_{=V\in\mathbb{C}^{n\times n}} = \left(Av_1 \middle| \cdots \middle| Av_n\right) = \left(\lambda_1 v_1 \middle| \cdots \middle| \lambda_n v_n\right) = \left(v_1 \middle| \cdots \middle| v_n\right) \underbrace{\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}}_{=\Lambda\in\mathbb{C}^{n\times n}},$$

ist die Matrix **diagonalisierbar** mit zugehöriger **Diagonalisierung** $\Lambda \in \mathbb{C}^{n\times n}$ (aufgrund der linearen Unabhängigkeit der Basisvektoren ist V invertierbar)

$$AV = V\Lambda \iff A = V\Lambda V^{-1} \iff \Lambda = V^{-1}AV.$$

Falls die Matrix $A \in \mathbb{K}^{n\times n}$ **unitär diagonalisierbar** ist, d.h. eine Orthonormalbasis aus Eigenvektoren existiert, folgt insbesondere (wegen $V^{-1} = V^*$)

$$AV = V\Lambda \iff A = V\Lambda V^* \iff \Lambda = V^*AV.$$

– Die Transformation

$$F_T : \mathbb{K}^{n\times n} \to \mathbb{K}^{n\times n} : A \mapsto T^{-1}AT$$

mit invertierbarer Matrix $T \in \mathbb{K}^{n\times n}$ heißt eine **Ähnlichkeitstransformation**. Dies entspricht einem Basiswechsel im Definitionsbereich und Bildbereich (verwende $y = Ax$, $\tilde{x} = Tx$, $\tilde{y} = Ty = TAx = TAT^{-1}\tilde{x}$).

Die Matrix A und die transformierte Matrix $T^{-1}AT$ besitzen dieselben Eigenwerte und mittels F_T transformierte Eigenräume

$$\begin{aligned} & Av = \lambda v, \quad 0 \neq v \in \mathcal{N}_{A-\lambda I} = \{v \in \mathbb{K}^n : (A - \lambda I)\, v = 0\} \\ \underset{v=Tw,\, w=T^{-1}v}{\iff} \quad & ATw = \lambda Tw, \quad 0 \neq v \in \mathcal{N}_{A-\lambda I}, \quad 0 \neq w = T^{-1}v \\ \iff \quad & T^{-1}ATw = \lambda w, \quad 0 \neq w \in \mathcal{N}_{T^{-1}AT-\lambda I} = \{w \in \mathbb{K}^n : (T^{-1}AT - \lambda I)\, w = 0\}. \end{aligned}$$

- Nicht jede (qudratische) Matrix ist diagonalisierbar, jedoch kann jede (quadratische) Matrix mittels einer **unitären Ähnlichkeitstransformation auf Dreiecksgestalt** transformiert werden.

Lemma von Schur (Satz 6.2): Für jede Matrix $A \in \mathbb{K}^{n\times n}$ mit (beliebig angeordneten) Eigenwerten $\lambda_1,\ldots,\lambda_n \in \mathbb{C}$ gibt es eine unitäre Matrix $T \in \mathbb{C}^{n\times n}$ (d.h. $T^* T = I$) derart, daß

$$T^* A T = S, \qquad S = \begin{pmatrix} \lambda_1 & S_{12} & \ldots & S_{1n} \\ & \ddots & & \vdots \\ & & \ddots & S_{n-1,n} \\ & & & \lambda_n \end{pmatrix}.$$

Denn: Es sei $\lambda_1 \in \mathbb{C}$ ein Eigenwert der Matrix $A \in \mathbb{K}^{n\times n}$ und $0 \neq v_1 \in \mathbb{C}^n$ ein zugehöriger Eigenvektor

$$A v_1 = \lambda_1 v_1,$$

wobei zusätzlich $\|v_1\|_2 = 1$ und $v_{11} \geq 0$ (v_{11} bezeichne die erste Komponente von v_1) angenommen wird. Frühere Überlegungen zeigten, wie eine Householder-Reflexion $T_1 \in \mathbb{C}^{n\times n}$ konstruiert werden kann, sodaß (Elimination in der ersten Spalte, verwende die Relation $T_1^{-1} = T_1^* = T_1$, explizite Angabe von T_1 wird nicht benötigt)

$$T_1 v_1 = e_1, \qquad T_1 e_1 = T_1^{-1} e_1 = v_1.$$

Mit Hilfe der Relation

$$e_i^T B\, e_j = \sum_{k,\ell=1}^{n} \delta_{ik} B_{k\ell} \delta_{j\ell} = b_{ij}, \qquad 1 \leq i,j \leq n,$$

folgt somit

$$\left(T_1 A T_1^{-1}\right)_{i1} = e_i^T T_1 A T_1^{-1} e_1 = e_i^T T_1 A \underbrace{T_1 e_1}_{=v_1} = e_i^T T_1 \underbrace{A v_1}_{=\lambda_1 v_1} = \lambda_1 e_i^T \underbrace{T_1 v_1}_{=e_1} = \lambda_1 \delta_{i1},$$

$$T_1 A T_1^{-1} = \begin{pmatrix} \lambda_1 & * & \ldots & * \\ 0 & & & \\ \vdots & & A_2 & \\ 0 & & & \end{pmatrix}, \qquad A_2 \in \mathbb{C}^{n-1\times n-1}.$$

Da A und $T_1 A T_1$ ähnliche Matrizen sind, besitzen sie dieselben Eigenwerte. Weiters gilt für das charakteristische Polynom (Determinantenentwicklungssatz)

$$\chi(\lambda) = \det(A - \lambda I) = \det(T_1 A T_1^{-1} - \lambda I) = (\lambda_1 - \lambda)\det(A_2 - \lambda I),$$

d.h. die restlichen Eigenwerte $\lambda_2,\ldots,\lambda_n$ von A sind Eigenwerte von A_2. Induktiv folgt mittels der Transformationsmatrizen

$$T_2 = \begin{pmatrix} 1 & & \\ & \widetilde{T}_2 & \\ & & \end{pmatrix}, \qquad T_3 = \begin{pmatrix} 1 & & \\ & 1 & \\ & & \widetilde{T}_3 \end{pmatrix}, \qquad \text{etc.},$$

die Relation

$$\underbrace{T_{n-1}\cdots T_1}_{=T^*} A \underbrace{T_1^{-1}\cdots T_{n-1}^{-1}}_{=T_1\cdots T_{n-1}=T} = \begin{pmatrix} \lambda_1 & * & \dots & * \\ & \ddots & & \vdots \\ & & \ddots & * \\ & & & \lambda_n \end{pmatrix}$$

und damit die Behauptung. ◇

Bemerkung: Das Lemma von Schur wird vorwiegend für theoretische Überlegungen verwendet (s.u.). Die Berechnung der Matrizen S, T (entsprechend der Herleitung) benötigt die Kenntnis der Eigenwerte von A und zugehöriger Eigenvektoren.

Folgerung: Es sei $\lambda \in \mathbb{C}$ ein Eigenwert von $A \in \mathbb{K}^{n\times n}$ mit algebraischer Multiplizität $\mu(\lambda)$ und geometrischer Multiplizität $\nu(\lambda)$. Es gilt

$$\nu(\lambda) \le \mu(\lambda)\,.$$

Denn: Nach dem Lemma von Schur reicht es aus, die unitär transformierte Matrix

$$T^* A\, T = S, \qquad S = \begin{pmatrix} \lambda_1 & S_{12} & \dots & S_{1n} \\ & \ddots & & \vdots \\ & & \ddots & S_{n-1,n} \\ & & & \lambda_n \end{pmatrix},$$

zu betrachten, wobei die Eigenwerte $\lambda_1, \dots, \lambda_n$ von A derart angeordnet sein sollen, daß $\lambda_1 = \lambda, \dots, \lambda_{\mu(\lambda)} = \lambda$. Betrachte beispielsweise den Fall

$$\mu(\lambda) = n = 3$$

und bestimme den zugehörigen Eigenraum durch Lösung des linearen Gleichungssystems

$$\big(T^* A\, T - \lambda I\big)\, v = \begin{pmatrix} 0 & S_{12} & S_{13} \\ & 0 & S_{23} \\ & & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} S_{12}\, v_2 + S_{13}\, v_3 \\ S_{23}\, v_3 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Die Unterscheidung der folgenden Fälle

$$\begin{aligned}
&S_{12}=0, \quad S_{13}=0, \quad S_{23}=0 \quad \Longrightarrow \quad v_1, v_2, v_3 \text{ beliebig} \quad \Longrightarrow \quad \nu(\lambda)=3,\\
&S_{12}\neq 0, \quad S_{13}=0, \quad S_{23}=0 \quad \Longrightarrow \quad v_2=0,\ v_1, v_3 \text{ beliebig} \quad \Longrightarrow \quad \nu(\lambda)=2,\\
&S_{12}=0, \quad S_{13}\neq 0, \quad S_{23}=0 \quad \Longrightarrow \quad v_3=0,\ v_1, v_2 \text{ beliebig} \quad \Longrightarrow \quad \nu(\lambda)=2,\\
&S_{12}=0, \quad S_{13}=0, \quad S_{23}\neq 0 \quad \Longrightarrow \quad v_3=0,\ v_1, v_2 \text{ beliebig} \quad \Longrightarrow \quad \nu(\lambda)=2,\\
&S_{12}\neq 0, \quad S_{13}\neq 0, \quad S_{23}=0 \quad \Longrightarrow \quad v_2=-\tfrac{S_{13}}{S_{12}}\, v_3,\ v_1, v_3 \text{ beliebig} \quad \Longrightarrow \quad \nu(\lambda)=2,\\
&S_{12}\neq 0, \quad S_{13}=0, \quad S_{23}\neq 0 \quad \Longrightarrow \quad v_2=v_3=0,\ v_1 \text{ beliebig} \quad \Longrightarrow \quad \nu(\lambda)=1,\\
&S_{12}=0, \quad S_{13}\neq 0, \quad S_{23}\neq 0 \quad \Longrightarrow \quad v_3=0,\ v_1, v_2 \text{ beliebig} \quad \Longrightarrow \quad \nu(\lambda)=2,\\
&S_{12}\neq 0, \quad S_{13}\neq 0, \quad S_{23}\neq 0 \quad \Longrightarrow \quad v_2=0, v_3=0,\ v_1 \text{ beliebig} \quad \Longrightarrow \quad \nu(\lambda)=1,
\end{aligned}$$

zeigt $\nu(\lambda) \leq \mu(\lambda)$. Ähnlich Überlegungen zeigen die Behauptung des Resultates im allgemeinen Fall. ◇

- Eine quadratische Matrix $A \in \mathbb{K}^{n\times n}$ heißt **normal**, wenn

$$A^*A = AA^*.$$

Spezialfälle:

- Reelle symmetrische Matrizen

$$A^* = A^T = A \quad \Longrightarrow \quad A^*A = A^2 = AA^*.$$

- Selbstadjungierte Matrizen

$$A^* = A \quad \Longrightarrow \quad A^*A = A^2 = AA^*.$$

- Reelle **antisymmetrische Matrizen** (d.h. $A^T = -A$)

$$A^* = A^T = -A \quad \Longrightarrow \quad A^*A = -A^2 = AA^*.$$

- Reelle orthogonale Matrizen

$$A^* = A^T = A^{-1} \quad \Longrightarrow \quad A^*A = I = AA^*.$$

- Unitäre Matrizen

$$A^* = A^{-1} \quad \Longrightarrow \quad A^*A = I = AA^*.$$

Resultat zur unitären Diagonalisierbarkeit (Korollar 6.3): Eine Matrix $A \in \mathbb{K}^{n\times n}$ ist genau dann normal, wenn es eine unitäre Matrix $T \in \mathbb{C}^{n\times n}$ gibt, derart daß

$$T^*AT = \Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}.$$

Folglich besitzt eine normale Matrix A ein vollständiges System orthonormaler Eigenvektoren und insbesondere gilt $\mu(\lambda) = \nu(\lambda)$ für alle Eigenwerte λ von A.
Denn: Mittels Lemma von Schur folgt die Darstellung

$$T^*AT = S$$

mit unitärer Matrix T und oberer Dreiecksmatrix S und folglich (wegen $T^{-1} = T^*$)

$$A = T^{*-1}ST^{-1} = TST^*,$$
$$A^* = (TST^*)^* = T^{**}S^*T^* = TS^*T^*,$$
$$TS^*ST^* = TS^*T^*TST^* = A^*A = AA^* = TST^*TS^*T^* = TSS^*T^* \Leftrightarrow S^*S = SS^*.$$

Die schrittweise Betrachtung der Koeffizienten der Differenz

$$S^*S - SS^* = 0$$

impliziert $S_{ij} = 0$ für $1 \leq i \leq i+1 \leq j \leq n$ und damit $S = \Lambda$ mit Diagonalmatrix Λ. Ist andererseits $S = \Lambda$ und $T^*AT = \Lambda$ so folgt $A^*A = T\Lambda^2T^* = AA^*$. ◇

Folgerung für selbstadjungierte Matrizen:

- Eine selbstadjungierte Matrix $A \in \mathbb{K}^{n\times n}$ ist unitär diagonalisierbar und alle Eigenwerte sind reell.
- Eine selbstadjungierte Matrix $A \in \mathbb{K}^{n\times n}$ ist genau dann positiv definit, wenn alle Eigenwerte positiv sind.

Denn: Eine selbstadjungierte Matrix (d.h. $A^* = A$) ist insbesondere normal und damit unitär diagonalisierbar

$$T^* A\, T = \Lambda\,, \qquad A = T \Lambda T^*\,.$$

Damit folgt

$$\begin{aligned} T\Lambda^* T^* = T^{**}\Lambda^* T^* = \left(T\Lambda T^*\right)^* = A^* = A = T\Lambda T^* \\ \iff \quad \Lambda^* = \Lambda \quad \iff \quad \overline{\lambda_i} = \lambda_i\,, \quad 1 \le i \le n \quad \iff \quad \lambda_i \in \mathbb{R}, \quad 1 \le i \le n\,. \end{aligned}$$

Die Wahl $0 \neq x = v_i = T_{-i}$ für $1 \le j \le n$ zeigt, daß die Matrix positiv definit ist, wenn

$$0 < x^* A\, x = v_i^* \underbrace{A\, v_i}_{=\lambda_i v_i} = \lambda_i \| v_i \|_2^2 \quad \iff \quad \lambda_i > 0\,, \qquad 1 \le i \le n\,.$$

Gilt andererseits $\lambda_i > 0$ für alle $1 \le i \le n$, so folgt für jedes Element $0 \neq x \in \mathbb{K}^n$ mittels der Darstellung bezüglich der Orthonormalbasis $v_1, \ldots, v_n$

$$0 \neq x = \sum_{i=1}^{n} c_i\, v_i\,, \qquad v_i = T_{-i} \quad 1 \le i \le n\,,$$

die Relation

$$x^* A\, x = \sum_{i=1}^{n} c_i\, x^* A\, v_i = \sum_{i=1}^{n} c_i \lambda_i\, x^* v_i = \sum_{i,j=1}^{n} c_i \overline{c_j} \lambda_i \underbrace{v_j^* v_i}_{=\delta_{ij}} = \sum_{i=1}^{n} |c_i|^2 \lambda_i > 0\,.$$

Dies zeigt die Behauptung. ◇

- **Resultat zur Kondition der Eigenwertberechnung** einer normalen Matrix (Satz 6.4): Es sei $A \in \mathbb{K}^{n\times n}$ eine normale Matrix (d.h. $A^*A = A\,A^*$) mit Eigenwerten $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$. Dann gilt für die entsprechenden Eigenwerte $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_n$ der Matrix $A + \alpha$ die folgende Abschätzung (ohne Begründung)

$$\left|\widetilde{\lambda}_j - \lambda_j\right| \le \|\alpha\|_2\,.$$

Somit ist die Eigenwertberechnung einer normalen Matrix sehr gut konditioniert.

6.2. Singulärwertzerlegung

- **Vorüberlegungen**:

  – Betrachte eine Matrix $A \in \mathbb{K}^{m\times n}$ mit $m \geq n$ und $\operatorname{rg}(A) = k$ mit $1 \leq k \leq n$. Die Matrix $B = A^* A \in \mathbb{K}^{n\times n}$ ist selbstadjungiert und positiv semi-definit

  $$B^* = (A^* A)^* = A^* A^{**} = A^* A = B\,, \qquad x^* Bx = x^* A^* Ax = \|Ax\|_2^2 \geq 0\,, \quad x \in \mathbb{K}^n\,.$$

  Insbesondere ist die Matrix $B = A^* A$ unitär diagonalisierbar mit nicht-negativen (reellen) Eigenwerten, d.h. es existiert eine unitäre Matrix $T \in \mathbb{C}^{n\times n}$ derart, daß

  $$T^* A^* A\, T = \Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}, \qquad T^* T = I\,, \qquad \lambda_i = \sigma_i^2 \geq 0\,, \quad 1 \leq i \leq n\,.$$

  – Wegen $\operatorname{rg}(B) = \operatorname{rg}(A^* A) = \operatorname{rg}(A) = k$ (verwende die Relation $\operatorname{rg}(A) = n - \dim \mathcal{N}_A$ sowie $\operatorname{rg}(A^* A) = n - \dim \mathcal{N}_{A^* A} = n - \dim \mathcal{N}_A = \operatorname{rg}(A)$) folgt nach eventueller Umordnung der Eigenwerte (und entsprechender Anpassung von T)

  $$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0\,, \qquad \lambda_{k+1} = \cdots = \lambda_n = 0\,.$$

  – Die Relation

  $$T^* A^* A\, T = \Lambda \quad \underset{S=AT}{\Longleftrightarrow} \quad S^* S = \Lambda = \begin{pmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_k & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}$$

  zeigt, daß die Spalten der Matrix $AT = S = \left(s_1 \middle| \cdots \middle| s_n\right) \in \mathbb{C}^{m\times n}$ orthogonal aufeinander stehen, d.h. $s_j^* s_i = 0$ für $1 \leq i, j \leq n$ mit $i \neq j$ (die Spalten sind nicht notwendigerweise normiert, es gilt $s_i^* s_i = \lambda_i \geq 0$). Zusätzliche Normierung der ersten k Spalten von S

  $$u_i = \tfrac{1}{\sqrt{\lambda_i}}\, s_i = \tfrac{1}{\sigma_i}\, s_i \in \mathbb{C}^m\,, \quad 1 \leq i \leq k\,,$$

  und Ergänzung mit orthonormalen Vektoren $u_{k+1}, \ldots, u_n \in \mathbb{C}^m$ führt auf die Relation

  $$AT = S = \left(\sigma_1 u_1 \middle| \cdots \middle| \sigma_k u_k \middle| \sigma_{k+1} u_{k+1} \middle| \cdots \middle| \sigma_n u_n\right) = \left(u_1 \middle| \cdots \middle| u_n\right) \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix}$$

und weiters auf die **reduzierte Singulärwertzerlegung** von A

$$A = \hat{U}\hat{\Sigma}T^*,$$
$$A = \left(a_1 \middle| \cdots \middle| a_n\right) \in \mathbb{K}^{m\times n}, \qquad T \in \mathbb{C}^{n\times n}, \quad T^*T = I,$$
$$\hat{U} = \left(u_1 \middle| \cdots \middle| u_n\right) \in \mathbb{C}^{m\times n}, \quad \hat{U}^*\hat{U} = I, \qquad \hat{\Sigma} = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix} \in \mathbb{R}^{n\times n}$$

Die (volle) **Singulärwertzerlegung** von A ergibt sich bei Ergänzung von $\hat{U}$ zu einer Orthonormalbasis des $\mathbb{C}^m$ und entsprechender Ergänzung von $\hat{\Sigma}$ um Nullzeilen

$$A = U\Sigma T^*,$$
$$A = \left(a_1 \middle| \cdots \middle| a_n\right) \in \mathbb{K}^{m\times n}, \qquad T \in \mathbb{C}^{n\times n}, \quad T^*T = I,$$
$$U = \left(u_1 \middle| \cdots \middle| u_m\right) \in \mathbb{C}^{m\times m}, \quad U^*U = I, \qquad \Sigma = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ 0 & \dots & 0 \end{pmatrix} \in \mathbb{R}^{m\times n},$$

vgl. Skriptum, S. 100.

Bemerkungen:

- Die **Erweiterung** auf den Fall $m < n$ ergibt sich durch Betrachtung der Singulärwertzerlegung der adjungierten Matrix $A^* = U\Sigma V^* \in \mathbb{K}^{n\times m}$ mit $n \geq m$ und Adjunktion $A = (U\Sigma V^*)^* = V\Sigma^T U^* \in \mathbb{K}^{m\times n}$.
- Wie üblich werden von nun an die Bezeichnungen $A = U\Sigma V^*$ verwendet.

Resultat zur Singulärwertzerlegung (Satz 6.5): Jede Matrix $A \in \mathbb{K}^{m\times n}$ mit Rang $\operatorname{rg}(A) = k$, wobei $1 \leq k \leq \min\{m,n\}$, besitzt die Darstellung

$$A = U\Sigma V^*,$$
$$U \in \mathbb{C}^{m\times m}, \quad U^*U = I, \qquad V \in \mathbb{C}^{n\times n}, \quad V^*V = I,$$
$$\Sigma = \operatorname{diag}(\sigma_1,\dots,\sigma_{\min\{m,n\}}) \in \mathbb{R}^{m\times n}, \qquad \sigma_1 \geq \cdots \geq \sigma_k > 0, \quad \sigma_{k+1} = \cdots = \sigma_{\min\{m,n\}} = 0,$$

mit unitären Matrizen U, V und Matrix Σ definiert durch die **Singulärwerte** von A.

Mögliche **Anwendung** der Singulärwertzerlegung zur Lösung eines linearen Gleichungssystems (entspricht einer Entkopplung der Gleichungen durch geeignete Basiswechsel)

$$Ax = b \underset{A=U\Sigma V^*}{\iff} U\Sigma V^*x = b \iff \Sigma y = c, \quad y = V^*x, \quad c = U^*b.$$

Falls die Matrix A unitär diagonalisierbar ist, ergibt sich insbesondere

$$Ax = b \underset{AT=T\Lambda}{\iff} T\Lambda T^*x = b \iff \Lambda y = c, \quad y = T^*x, \quad c = T^*b.$$

- **Eigenschaften der Singulärwertzerlegung** (Korollar 6.6): Für die Singulärwertzerlegung $A = U\Sigma V^*$ einer Matrix gelten folgende Eigenschaften:
  - Die Quadrate der Singulärwerte $\{\sigma_1^2, \dots, \sigma_n^2\}$ stimmen mit den Eigenwerten von A^*A und AA^* überein.
  - Es gilt $\mathscr{R}_A = \langle u_1, \dots, u_k \rangle$ und $\mathscr{N}_A = \langle v_{k+1}, \dots, v_n \rangle$.
  - Es gilt (wobei $\kappa(A) = \infty$ für $\sigma_n = 0$)

$$\|A\|_2 = \sigma_1, \qquad \kappa(A) = \frac{\sigma_1}{\sigma_n}.$$

- **Vorüberlegung**: Mittels der Singulärwertzerlegung einer Matrix $A \in \mathbb{K}^{m\times n}$

$$A = U\Sigma V^*,$$
$$V = \left(v_1 \middle| \cdots \middle| v_n\right) \in \mathbb{C}^{n\times n}, \quad V^*V = I, \qquad U = \left(u_1 \middle| \cdots \middle| u_m\right) \in \mathbb{C}^{m\times m}, \quad U^*U = I,$$
$$\Sigma = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ 0 & \dots & 0 \end{pmatrix} \in \mathbb{R}^{m\times n}, \quad m \geq n, \qquad \Sigma = \begin{pmatrix} \sigma_1 & & & 0 \\ & \ddots & & \vdots \\ & & \sigma_n & 0 \end{pmatrix} \in \mathbb{R}^{m\times n}, \quad m \leq n,$$

ergibt sich wegen

$$Ax = U\underbrace{\Sigma V^* x}_{=y} = Uy = \sum_{i=1}^{m} y_i u_i = \sum_{i=1}^{k} \sigma_i v_i^* x\, u_i = \sum_{i=1}^{k} \sigma_i u_i v_i^* x,$$
$$y = \Sigma \begin{pmatrix} v_1^* x \\ \vdots \\ v_n^* x \end{pmatrix} = \begin{pmatrix} \sigma_1 v_1^* x \\ \vdots \\ \sigma_k v_k^* x \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

die Darstellung von A als Summe der Rang 1 Matrizen $u_i v_i^* \in \mathbb{C}^{m\times n}$ für $1 \leq i \leq k$

$$A = \sum_{i=1}^{k} \sigma_i u_i v_i^*.$$

Diese Darstellung wird zur **Approximation** der Matrix A verwendet.

Bemerkung: Für orthonormale Vektoren $z_1, \dots, z_k \in \mathbb{K}^m$ und Skalare $c_1, \dots, c_k \in \mathbb{K}$ gilt die folgende Relation (**Satz von Pythagoras**)

$$\left\| \sum_{i=1}^{k} c_i z_i \right\|_2^2 = \left\langle \sum_{i=1}^{n} c_i z_i \middle| \sum_{j=1}^{n} c_j z_j \right\rangle_2 = \sum_{i,j=1}^{n} c_i \overline{c_j} \underbrace{\langle z_i | z_j \rangle_2}_{=z_j^* z_i = \delta_{ij}} = \sum_{i=1}^{n} |c_i|^2.$$

Resultat zur Approximation einer Matrix durch Matrizen niedrigeren Ranges (Satz 6.7): Für die Approximation der Matrix A durch eine Matrix A_j vom Rang $1 \le j \le k-1$

$$A_j = \sum_{i=1}^{j} \sigma_i u_i v_i^* \approx A = \sum_{i=1}^{k} \sigma_i u_i v_i^*$$

gilt die Relation

$$\|A_j - A\|_2 = \inf\{\|A - B\|_2 : B \in \mathbb{C}^{m\times n}, \operatorname{rg}(B) \le j\} = \sigma_{j+1}\,.$$

Denn: (i) Zeige zunächst die Relation

$$\|A_j - A\|_2 = \sigma_{j+1}\,.$$

Einerseits gilt für $1 \le j \le k-1$ (Anwendung des Satzes von Pythagoras auf die Orthonormalbasis $u_1, \dots, u_m$)

$$A - A_j = \sum_{i=1}^{k} \sigma_i u_i v_i^* - \sum_{i=1}^{j} \sigma_i u_i v_i^* = \sum_{i=j+1}^{k} \sigma_i u_i v_i^*\,,$$

$$\|A - A_j\|_2 = \max_{\|x\|_2=1} \|(A-A_j)x\|_2 = \max_{\|x\|_2=1} \Big\| \sum_{i=j+1}^{k} \sigma_i v_i^* x\, u_i \Big\| = \max_{\|x\|_2=1} \sqrt{\sum_{i=j+1}^{k} \sigma_i^2 |v_i^* x|^2}\,.$$

Mit Hilfe der Darstellung bezüglich der Basis $v_1, \dots, v_n \in \mathbb{C}^n$ folgt weiters für jedes Element $x \in \mathbb{C}^n$ mit $\|x\|_2 = 1$ (wegen $\sigma_i \le \sigma_{j+1}$ für $j+1 \le i \le k$)

$$x = \sum_{i=1}^{n} \xi_i v_i\,, \qquad 1 = \|x\|_2 = \Big\| \sum_{i=1}^{n} \xi_i v_i \Big\|_2 = \sqrt{\sum_{i=1}^{n} |\xi_i|^2}\,, \qquad v_j^* x = \sum_{i=1}^{n} \xi_i \underbrace{v_j^* v_i}_{=\delta_{ij}} = \xi_j\,,$$

$$\|A - A_j\|_2 = \max_{\|x\|_2=1} \sqrt{\sum_{i=j+1}^{k} \sigma_i^2 |v_i^* x|^2} \le \max_{\|x\|_2=1} \sigma_{j+1} \underbrace{\sqrt{\sum_{i=j+1}^{k} |\xi_i|^2}}_{\le 1} \le \sigma_{j+1}\,.$$

Mit der Wahl $x = v_{j+1}$ gilt (beachte, daß $\|v_{j+1}\|_2 = 1$)

$$\sigma_{j+1} = \sqrt{\sum_{i=j+1}^{k} \sigma_i^2 |v_i^* v_{j+1}|^2} \le \|A - A_j\|_2 \le \sigma_{j+1} \quad \Longrightarrow \quad \|A - A_j\|_2 = \sigma_{j+1}\,.$$

(ii) Zeige die Relation

$$\|A_j - A\|_2 = \inf\{\|A - B\|_2 : B \in \mathbb{C}^{m\times n}, \operatorname{rg}(B) \le j\}\,.$$

Unter der Annahme für $B \in \mathbb{C}^{m\times n}$ gilt

$$\operatorname{rg}(B) \le j \Leftrightarrow \dim \mathcal{N}_B \ge n - j\,, \qquad \|A - B\|_2 < \sigma_{j+1}\,.$$

Wählt man einen $(n-j)$-dimensionalen Unterraum $U_1 \subset \mathcal{N}_B \subset \mathbb{C}^n$, so folgt

$$0 \neq u \in U_1: \quad (A-B)u = Au - Bu = Au$$
$$\Longrightarrow \quad \|Au\|_2 = \|(A-B)u\|_2 \leq \|A-B\|_2 \|u\|_2 < \sigma_{j+1}\|u\|_2 .$$

Betrachtet man den $j+1$ dimensionalen Unterraum $U_2 = \langle v_1, \ldots, v_{j+1} \rangle \subset \mathbb{C}^n$, so gilt (beachte, daß $j+1 \leq k$)

$$u = \sum_{i=1}^{j+1} c_i v_i \in U_2 , \qquad A = \sum_{\ell=1}^{k} \sigma_\ell u_\ell v_\ell^* ,$$
$$Au = \sum_{i=1}^{j+1} c_i A v_i = \sum_{i=1}^{j+1} \sum_{\ell=1}^{k} c_i \sigma_\ell u_\ell \underbrace{v_\ell^* v_i}_{=\delta_{i\ell}} = \sum_{i=1}^{j+1} c_i \sigma_i u_i ,$$
$$\|u\|_2 = \sqrt{\sum_{i=1}^{j+1} |c_i|^2} , \qquad \|Au\|_2 = \sqrt{\sum_{i=1}^{j+1} \sigma_i^2 |c_i|^2} \geq \sigma_{j+1} \sqrt{\sum_{i=1}^{j+1} |c_i|^2} = \sigma_{j+1}\|u\|_2 .$$

Wegen $1 \leq j \leq k-1 < k \leq n$ folgt $\dim U_1 \geq n-j \geq 1$ und $\dim U_2 = j+1 \geq 2$ und damit $\dim(U_1 \cap U_2) \geq 1$, d.h. es existiert ein Element $0 \neq u \in U_1 \cap U_2$ mit einerseits $\|Au\|_2 < \sigma_{j+1}\|u\|_2$ und andererseits $\|Au\|_2 \geq \sigma_{j+1}\|u\|_2$. Widerspruch! Somit folgt die Behauptung. ◇

- **Bemerkung**: Niedrigrangapproximationen von Matrizen haben Anwendungen in verschiedenen Bereichen der Numerischen Mathematik (Theorie der Inversen Probleme, Bildkompression).

  Beispiel, vgl. Skriptum S. 104: Ein Bild bestehend aus $m \times n$ Pixel entspricht einer Matrix $A \in \mathbb{R}^{m\times n}$ (Koeffizient a_{ij} entspricht dem Wert des Pixel an der Position (i,j), d.h. einer Farbstufe). Eine Kompression des Bildes entspricht der Approximation der Matrix A durch eine Matrix A_j vom Rang $1 \leq j \leq k-1$

  $$A_j = \sum_{i=1}^{j} \sigma_i u_i v_i^* \approx A = \sum_{i=1}^{k} \sigma_i u_i v_i^* .$$

  Ergebnis für $m = 576$, $n = 768$, $j = 30$.

- **Bemerkungen**:
  - Zur Berechnung der Singulärwertzerlegung $A = U\Sigma V^*$ einer Matrix könnte man im Prinzip die Eigenwertzerlegung von $B = A^*A$ verwenden (zunächst die Relation $V^*BV = \Sigma^2$ zur Berechnung von Σ und V sowie die Relation $AV = U\Sigma$ zur Berechnung von U). Diese Vorgehensweise kann allerdings zu einem numerisch instabilen Algorithmus führen, weil sich kleine Änderungen der Koeffizienten von A stärker auf die Eigenwerte von B auswirken als auf die Singulärwerte von A.

– Zur Berechnung der Singulärwertzerlegung ist es vorteilhafter, die Matrix

$$\boldsymbol{A} = \begin{pmatrix} & A^* \\ A & \end{pmatrix}$$

und die zugehörige Eigenwertrelation (ohne explizite Berechnung von $\boldsymbol{A}$, verwende $A = U\Sigma V^* \Leftrightarrow AV = U\Sigma$ bzw. $A^* = V\Sigma U^* \Leftrightarrow A^*U = V\Sigma$)

$$\boldsymbol{AV} = \boldsymbol{V\Sigma}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \Sigma & \\ & -\Sigma \end{pmatrix}, \quad \boldsymbol{V} = \begin{pmatrix} V & V \\ U & -U \end{pmatrix},$$

$$\begin{pmatrix} A^*U & -A^*U \\ AV & AV \end{pmatrix} = \begin{pmatrix} & A^* \\ A & \end{pmatrix} \begin{pmatrix} V & V \\ U & -U \end{pmatrix} = \begin{pmatrix} V & V \\ U & -U \end{pmatrix} \begin{pmatrix} \Sigma & \\ & -\Sigma \end{pmatrix} = \begin{pmatrix} V\Sigma & -V\Sigma \\ U\Sigma & -U\Sigma \end{pmatrix},$$

zu verwenden.

6.3. Algorithmen zum Eigenwertproblem (Überblick)

- **Vorbemerkungen**:
  - Da es keine explizite Formel für die Nullstellen eines allgemeinen Polynoms höheren Grades gibt, kann es keine explizite Formel für die Eigenwerte einer allgemeinen Matrix höherer Dimension geben. Ein Polynom vom Grad $m \geq 1$ mit Koeffizienten $c_i \in \mathbb{K}$ für $0 \leq i \leq m$ (Leitkoeffizient $c_m = 1$)

    $$p : \mathbb{K} \to \mathbb{K} : z \mapsto p(z) = \sum_{i=0}^{m} c_i z^i$$

    entspricht nämlich dem charakteristischen Polynom der folgenden Matrix (zusätzliches Vorzeichen $(-1)^n$, geeignete Gauß-Elimination zur Berechnung von $\det(A - \lambda I)$ zeigt den Zusammenhang)

    $$A = \begin{pmatrix} & & -c_0 \\ 1 & & -c_1 \\ & \ddots & \vdots \\ & 1 & -c_{m-1} \end{pmatrix} \in \mathbb{K}^{n\times n}, \qquad \chi(\lambda) = \det(A - \lambda I).$$

  - Die obigen Überlegungen zeigen, daß ein numerisches Verfahren zur Berechnung eines (bzw. mehrerer) Eigenwertes λ einer allgemeinen Matrix $A \in \mathbb{K}^{n\times n}$ notwendigerweise ein iteratives Verfahren sein muß, d.h. ein Verfahren der Form

    $$\lambda^{(k+1)} = \Psi\big(\lambda^{(k)}\big), \qquad k \geq 0,$$

    welches eine Folge von Näherungswerten an λ ergibt. Sofern das Verfahren konvergiert, d.h. es ist

    $$\lim_{k\to\infty} \lambda^{(k)} = \lambda,$$

    bricht man die Iteration ab, sobald eine ausreichende Genauigkeit erreicht werden konnte (beispielsweise verwendet man das Abbruchkriterium $|\lambda^{(k+1)} - \lambda^{(k)}| \leq$ TOL)

    $$|\lambda^{(k)} - \lambda| \leq \text{TOL}.$$

  - Das meistverwendete numerische Verfahren zur Berechnung sämtlicher Eigenwerte einer Matrix, der QR-Algorithmus (mit Shift), basiert auf dem Lemma von Schur. Dieses besagt, daß jede Matrix $A \in \mathbb{K}^{n\times n}$ mittels einer unitären Matrix $T \in \mathbb{C}^{n\times n}$ (d.h. $T^* T = I$) auf Dreiecksform transformiert werden kann

    $$A \quad \longrightarrow \quad S = T^* A T = \begin{pmatrix} \lambda_1 & S_{12} & \dots & S_{1n} \\ & \ddots & & \vdots \\ & & \ddots & S_{n-1,n} \\ & & & \lambda_n \end{pmatrix}.$$

Die Diagonalelemente von S entsprechen den Eigenwerten $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ von A (die Eigenwerte von A bleiben unter Ähnlichkeitstransformationen erhalten). Die grundlegende Idee ist es, eine Folge von unitären Matrizen $Q_k \in \mathbb{C}^{n\times n}$ für $k \geq 1$ zu konstruieren, derart daß die transformierten Matrizen gegen S konvergieren

$$\begin{gathered} A_1 = A, \qquad A_{k+1} = Q_k^* A_k Q_k, \qquad k = 1,2,3,\ldots \\ A_1 = A \quad \longrightarrow \quad A_2 = Q_1^* A Q_1 \quad \longrightarrow \quad A_3 = Q_2^* Q_1^* A Q_1 Q_2 \quad \longrightarrow \quad \ldots \\ \lim_{k\to\infty} A_k = S. \end{gathered}$$

Dazu verwendet man die Transformation von $A \in \mathbb{K}^{n\times n}$ auf **obere Hessenberg-Form** sowie die Idee der (inversen) Vektoriteration.

– Beispielsweise mittels Householder-Reflexionen läßt sich eine allgemeinen Matrix $A \in \mathbb{K}^{n\times n}$ auf **obere Hessenberg-Form** transformieren

$$A \quad \longrightarrow \quad H = T^* A T = \begin{pmatrix} H_{11} & \ldots & \ldots & H_{1n} \\ H_{21} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & H_{n,n-1} & H_{nn} \end{pmatrix}.$$

  * 1. Schritt: Konstruktion einer Householder-Reflexion $\widetilde{Q}_1 \in \mathbb{K}^{n-1\times n-1}$ derart, daß die Spaltenelemente $(a_{21}, \ldots, a_{n1})^T \in \mathbb{K}^{n-1}$ auf ein Vielfaches des Standardbasisvektors $e_1 \in \mathbb{R}^{n-1}$ abgebildet werden. Mit

$$Q_1 = \begin{pmatrix} 1 & \\ & \widetilde{Q_1} \end{pmatrix}$$

    folgt somit (beachte, daß die entstehenden Matrizen AQ_1 und $Q_1^* A Q_1$ die gewünschte Form haben)

$$A \quad \longrightarrow \quad Q_1^* A Q_1 = \begin{pmatrix} * & * & \ldots & * \\ * & * & \ldots & * \\ & \vdots & & \vdots \\ & * & \ldots & * \end{pmatrix}.$$

  * Die Anwendung analoger Ideen auf die entstehenden Teilmatrizen führt nach $n-2$ Schritten auf eine Matrix in oberer Hessenberg-Form. Die Anzahl der benötigten Operationen (Addition oder Multiplikation in $\mathbb{K}$) ist

$$\mathcal{O}\left(\tfrac{4}{3} n^3\right).$$

  * Man beachte, daß im Spezialfall einer selbstadjungierten Matrix $A \in \mathbb{K}^{n\times n}$ (d.h. es ist $A^* = A$) die enstehenden Matrizen ebenfalls selbstadjungiert sind.

Somit führt die obige Vorgehensweise auf eine selbstadjungierte **Tridiagonalmatrix**

$$A \quad \longrightarrow \quad H = T^* A T = \begin{pmatrix} H_{11} & H_{12} & & \\ H_{21} & \ddots & \ddots & \\ & \ddots & \ddots & H_{n-1,n} \\ & & H_{n,n-1} & H_{nn} \end{pmatrix},$$

$$H_{ii} \in \mathbb{R}, \quad H_{i+1,i} = \overline{H_{i,i+1}} \in \mathbb{R}, \qquad 1 \le i \le i+1 \le n.$$

– Im Folgenden werden grundlegende Ideen behandelt, die zur Konstruktion von Algorithmen für Eigenwertberechnungen führen. Für die praktische Berechnung von Eigenwerten und Eigenvektoren sollte man ausgeklügelte Software-Pakete verwenden.

6.4. Vektoriteration und inverse Vektoriteration

- **Situation**: Es sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische und reelle (quadratische) Matrix (d.h. es ist $A^T = A$). Folglich sind alle Eigenwerte $\lambda_1, \dots, \lambda_n$ von A reell und es gibt ein vollständiges System von orthonormalen Eigenvektoren $v_1, \dots, v_n \in \mathbb{R}^n$, d.h. es gilt

$$A\,V = V\,\Lambda\,, \qquad V = \big(v_1 \,\big|\, \cdots \,\big|\, v_n\big) \in \mathbb{R}^{n\times n}\,, \quad V^{-1} = V^T\,, \qquad \Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n) \in \mathbb{R}^{n\times n}\,,$$

  vgl. früheres Resultat zur unitären Diagonalisierbarkeit einer normalen Matrix und Folgerung für selbstadjungierte Matrizen.

  Bemerkung: Symmetrische reelle Matrizen sind bedeutsam in Hinblick auf praktische Anwendungen und erlauben gewisse Vereinfachungen. Beispielsweise führen Ähnlichkeitstransformationen mittels Householder-Reflexionen auf symmetrische Tridiagonalmatrizen.

- Für $A \in \mathbb{R}^{n\times n}$ und $x \in \mathbb{R}^n$ ist der **Rayleigh-Quotient** gegeben durch

$$r : \mathbb{R}^n \setminus \{0\} \to \mathbb{R} : x \mapsto r(x) = \frac{x^T A\,x}{x^T x}\,, \qquad x^T A\,x = \sum_{i,j=1}^{n} x_i\,a_{ij}\,x_j\,, \quad x^T x = \sum_{i=1}^{n} x_i^2\,.$$

  Für den Gradienten ergibt sich folgende Relation (verwende Symmetrie $a_{ji} = a_{ij}$ für $1 \le i, j \le n$ und beachte $r'(x) = (\partial_{x_1} r(x), \dots, \partial_{x_n} r(x)) \in \mathbb{R}^{1\times n}$)

$$\begin{aligned}
r'(x) &= 2\,\frac{1}{x^T x}\Big((Ax)^T - r(x)\,x^T\Big)\,,\\
\partial_{x_\ell} r(x) &= \frac{\sum_{i,j=1}^{n} (\delta_{i\ell}\,a_{ij}\,x_j + x_i\,a_{ij}\,\delta_{j\ell})}{x^T x} - 2\,r(x)\,\frac{x_\ell}{x^T x}\\
&= \frac{\sum_{j=1}^{n} a_{\ell j}\,x_j + \sum_{i=1}^{n} x_i\,a_{i\ell}}{x^T x} - 2\,r(x)\,\frac{x_\ell}{x^T x}\\
&= 2\,\frac{(Ax)_{\ell 1}}{x^T x} - 2\,r(x)\,\frac{x_\ell}{x^T x}\\
&= 2\,\frac{1}{x^T x}\big((Ax)_{\ell 1} - r(x)\,x_\ell\big)\,.
\end{aligned}$$

  Ist $0 \neq v \in \mathbb{R}^n$ ein Eigenvektor zum Eigenwert $\lambda \in \mathbb{R}$, folgt insbesondere

$$r(v) = \lambda\,, \qquad r'(v) = 2\,\frac{1}{v^T v}\Big(\lambda\,v^T - \underbrace{r(v)}_{=\lambda}\,v^T\Big) = 0\,.$$

  Taylorreihenentwicklung von r um v zeigt somit

$$r(x) = r(v) + \mathscr{O}\big(\|x - v\|_2^2\big)\,.$$

Dies zeigt, daß der Rayleight-Quotient für eine gute Näherung x an einen Eigenvektor v eine *gute* Approximation an den zugehörigen Eigenwert darstellt

$$r(x) \approx r(v) = \lambda\,, \qquad r(x) - r(v) = \mathcal{O}\big(\|x-v\|_2^2\big)\,.$$

- Die obigen Überlegungen motivieren die (direkte) **Vektoriteration**

$$x^{(k)} = Ax^{(k-1)} = A^k x^{(0)}\,, \qquad \lambda^{(k)} = r\big(x^{(k)}\big) \approx \lambda\,,$$

mit einem *geeignet* gewählten Startvektor $x^{(0)} \in \mathbb{R}^n$ mit $\|x^{(0)}\|_2 = 1$ (mit Ergänzung eines Abbruchkriteriums)

$x = x^{(0)}$
for k = 1,2, ...
  $x = Ax$
  $\lambda = r(x)$
end

Mittels der Darstellung des Startvektors bezüglich der Orthonormalbasis von Eigenvektoren ergibt sich (unter der Annahme $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n| \geq 0$ und $\xi_1 = v_1^T x^{(0)} \neq 0$)

$$x^{(0)} = \sum_{i=1}^n \xi_i v_i\,,$$

$$x^{(k)} = A^k x^{(0)} = \sum_{i=1}^n \xi_i A^k v_i = \sum_{i=1}^n \xi_i \lambda_i^k v_i = \xi_1 \lambda_1^k \Big(v_1 + \sum_{i=2}^n \frac{\xi_i}{\xi_1} \big(\frac{\lambda_i}{\lambda_1}\big)^k v_i \Big)\,,$$

$$\frac{1}{\xi_1 \lambda_1^k} x^{(k)} = v_1 + \sum_{i=2}^n \frac{\xi_i}{\xi_1} \big(\frac{\lambda_i}{\lambda_1}\big)^k v_i \quad \overset{\substack{k\to\infty \\ |\frac{\lambda_i}{\lambda_1}|=\varrho_i<1}}{\longrightarrow} \quad v_1\,.$$

Für den Rayleigh-Quotienten folgt weiters

$$\big\|x^{(k)}\big\|_2^2 = \sum_{i=1}^n \xi_i^2 \lambda_i^{2k} = \xi_1^2\lambda_1^{2k} + \sum_{i=2}^n \xi_i^2 \lambda_i^{2k} = \xi_1^2 \lambda_1^{2k} \Big(1 + \sum_{i=2}^n \frac{\xi_i^2}{\xi_1^2} \big(\frac{\lambda_i}{\lambda_1}\big)^{2k}\Big)\,,$$

$$\big(x^{(k)}\big)^T A x^{(k)} = \Big(\sum_{i=1}^n \xi_j \lambda_j^k v_j\Big)^T \sum_{i=1}^n \xi_i \lambda_i^{k+1} v_i = \sum_{i,j=1}^n \xi_i \xi_j \lambda_i^{k+1} \lambda_j^k \underbrace{v_j^T v_i}_{=\delta_{ij}} = \sum_{i=1}^n \xi_i^2 \lambda_i^{2k+1}\,,$$

$$r\big(x^{(k)}\big) = \frac{1}{\|x^{(k)}\|_2^2} \big(x^{(k)}\big)^T A x^{(k)} = \frac{1}{\|x^{(k)}\|_2^2} \sum_{i=1}^n \xi_i^2 \lambda_i^{2k+1} = \frac{\xi_1^2 \lambda_1^{2k+1} \Big(1 + \sum_{i=2}^n \frac{\xi_i^2}{\xi_1^2} \big(\frac{\lambda_i}{\lambda_1}\big)^{2k+1}\Big)}{\xi_1^2 \lambda_1^{2k} \Big(1 + \sum_{i=2}^n \frac{\xi_i^2}{\xi_1^2} \big(\frac{\lambda_i}{\lambda_1}\big)^{2k}\Big)}$$

$$= \lambda_1 \frac{1 + \sum_{i=2}^n \frac{\xi_i^2}{\xi_1^2} \big(\frac{\lambda_i}{\lambda_1}\big)^{2k+1}}{1 + \sum_{i=2}^n \frac{\xi_i^2}{\xi_1^2} \big(\frac{\lambda_i}{\lambda_1}\big)^{2k}} \quad \overset{k\to\infty}{\longrightarrow} \quad \lambda_1\,.$$

Hier wurden die folgenden Relationen verwendet (geometrische Reihe, für ϱ hinreichend klein)

$$\frac{1}{1-q} = \sum_{\ell=0}^{\infty} q^\ell\,, \quad |q|<1\,, \qquad \frac{1+\mathscr{O}(\varrho)}{1+\mathscr{O}(\varrho)} = 1+\mathscr{O}(\varrho)\,.$$

Dies zeigt, daß die mittels der Vektoriteration berechneten Näherungswerte gegen den dominanten (d.h. betragsmäßig größten) Eigenwert und einen zugehörigen Eigenvektor konvergieren (unter der Annahme $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n| \geq 0$ und der Forderung $v_1^T x^{(0)} = \xi_1 \neq 0$, $c = \pm 1$)

$$\frac{1}{\|x^{(k)}\|_2}\, x^{(k)} \quad \overset{k\to\infty}{\longrightarrow} \quad c\, v_1\,, \qquad r\big(x^{(k)}\big) \quad \overset{k\to\infty}{\longrightarrow} \quad \lambda_1\,.$$

Bemerkungen:

- Beachte, daß der Rayleigh-Quotient skalierungsinvariant ist

$$r(cx) = \frac{(cx)^T A\,(cx)}{(cx)^T (cx)} = \frac{cx^T A\,x}{x^T x} = r(x)\,, \qquad c \neq 0\,.$$

  Eine Normierung der Approximationen $x^{(k)}$ dient dazu, rasch auftretenden Overflow (falls $|\lambda_1|>1$) bzw. Underflow (falls $|\lambda_1|<1$) zu verhindern.

- Die Konvergenzgeschwindigkeit der Vektoriteration wird durch die Größe

$$|\varrho_2| = \left|\frac{\lambda_2}{\lambda_1}\right|$$

  bestimmt. Falls $|\varrho| \approx 1$ ist die Konvergenzrate sehr dürftig.

- Analoge Überlegungen gelten für eine diagonalisierbare Matrix $A \in \mathbb{K}^{n\times n}$.

- **Vorbemerkungen**: Wesentliche Nachteile der direkten Vektoriteration sind, daß
  - nur der betragsmäßig größte Eigenwerte und ein zugehöriger Eigenwert berechnet werden können und
  - die Konvergenzrate unzufriedenstellend ist, falls $|\varrho_2| = \left|\frac{\lambda_2}{\lambda_1}\right| \approx 1$.

  Die **inverse Vektoriteration** basiert auf der Umformulierung der Eigenrelation (wobei $\mu \in \mathbb{R}$ mit $\mu \neq \lambda_i$ für $1 \leq i \leq n$)

$$\begin{aligned} Av_i = \lambda_i v_i\,, \quad 1\leq i\leq n \quad &\Longleftrightarrow \quad (A-\mu I)\,v_i = (\lambda_i-\mu)\,v_i\,, \quad 1\leq i\leq n \\ &\Longleftrightarrow \quad (A-\mu I)^{-1} v_i = \tfrac{1}{\lambda_i-\mu}\,v_i\,, \quad 1\leq i\leq n \\ &\Longleftrightarrow \quad \widetilde{A}_\mu v_i = \widetilde{\lambda}\,v_i\,, \quad B=(A-\mu I)^{-1}, \quad \widetilde{\lambda} = \tfrac{1}{\lambda_i-\mu}\,, \qquad 1\leq i\leq n\,. \end{aligned}$$

Unter der Annahme, daß eine Näherung $\mu \approx \lambda_i$ an den gesuchten Eigenwert der Matrix A bekannt ist, führt man die direkte Vektoriteration für die Matrix $\widetilde{A}_\mu$ durch

$$x^{(k)} = \widetilde{A}_\mu\, x^{(k-1)} = (A-\mu I)^{-1} x^{(k-1)}\,, \qquad \widetilde{\lambda}^{(k)} = r\big(x^{(k)}\big) \approx \widetilde{\lambda} = \tfrac{1}{\lambda_i-\mu}\,, \quad \lambda_i \approx \tfrac{1}{\widetilde{\lambda}^{(k)}} + \mu\,.$$

Dies erfordert in jedem Iterationsschritt die Lösung eines linearen Gleichungssystems (effiziente Umsetzung durch Berechnung z.B. einer LR-Zerlegung der Matrix, pro Iterationsschritt sind dann eine Vorwärtssubstitution und eine Rückwärtssubstitution nötig)

$$(A - \mu I)\, x^{(k)} = x^{(k-1)}\,.$$

Sofern eine gute Näherung $\mu \approx \lambda_i$ bekannt ist und die restlichen Eigenwerte von A deutlich verschieden von λ_i sind, ist die Konvergenzrate der inversen Vektoriteration wegen

$$|\lambda_i - \mu| \ll |\lambda_j - \mu|\,, \quad 1 \le i, j \le n\,, \quad j \ne i \quad \Longleftrightarrow \quad \tfrac{|\lambda_i - \mu|}{|\lambda_j - \mu|} \ll 1\,, \quad 1 \le i, j \le n\,, \quad j \ne i$$

ausgezeichnet.

Bemerkung: Es ist zu beachten, daß bei der inversen Vektoriteration im allgemeinen Matrizen mit großer Konditionszahl auftreten (falls $|\lambda_i - \mu| \approx 0$ ist $(A - \mu I)^{-1}$ nahezu singulär). Dennoch ist die numerische Anwendung sinnvoll. Offene Fragen sind außerdem die Konstruktion geeigneter Startvektoren und optimale Abbruchkriterien.

6.5. Die Grundidee des QR-Algorithmus

- **Vorbemerkung**: Der QR-Algorithmus (genauer der QR-Algorithmus mit Shift) ist ein effizientes numerisch stabiles Verfahren zur Berechnung aller Eigenwerte einer Matrix $A \in \mathbb{K}^{n\times n}$.

- Ausgehend von der bereits auf Hessenberg-Form transformierten Matrix $A^{(0)} = A$, basiert der **QR-Algorithmus** auf einer QR-Zerlegung der Matrix und anschließender Matrizenmultiplikation

$$\begin{cases} \text{QR-Zerlegung:} & Q^{(k)}R^{(k)} = A^{(k)}, \\ \text{Rekombination:} & A^{(k+1)} = R^{(k)}Q^{(k)}, \end{cases} \qquad k \geq 0.$$

  Ohne Einschränkung der Allgemeinheit kann angenommen werden, daß alle Diagonalelemente von $R^{(k)}$ positiv sind (ansonsten erfolgt eine Reduktion des Problems, vgl. Skriptum, S. 110). Beachte, daß der Übergang von A zu $A^{(k)}$ einer unitären Transformation entspricht

$$A^{(k+1)} = \underbrace{R^{(k)}}_{R^{(k)}=(Q^{(k)})^* A^{(k)}} Q^{(k)} = \left(Q^{(k)}\right)^* A^{(k)} Q^{(k)},$$
$$A^{(k)} = \left(X^{(k)}\right)^* A X^{(k)}, \qquad X^{(k)} = Q^{(0)} \cdots Q^{(k-1)}.$$

  Weiters erhält der QR-Algorithmus Eigenschaften wie Selbstadjungiertheit

$$\left(A^{(k)}\right)^* = A^{(k)} \quad \Longrightarrow \quad \left(A^{(k+1)}\right)^* = \left(\left(Q^{(k)}\right)^* A^{(k)} Q^{(k)}\right)^* = \left(Q^{(k)}\right)^* \underbrace{\left(A^{(k)}\right)^*}_{=A^{(k)}} Q^{(k)} = A^{(k+1)}$$

  und auch die Hessenberg-Form einer Matrix.

  Konvergenz des QR-Algorithmus: Unter der Annahme $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0$, konvergieren die beim QR-Algorithmus enstehenden Matrizen gegen eine obere Dreiecksmatrix mit den Eigenwerten der Matrix A als Diagonalelemente

$$\lim_{k\to\infty} A^{(k)} = \begin{pmatrix} \lambda_1 & * & \dots & \dots & * \\ & \ddots & \ddots & & \vdots \\ & & \ddots & \ddots & * \\ & & & \ddots & * \\ & & & & \lambda_n \end{pmatrix}.$$

  Denn: Es reicht aus, die Konvergenz der ersten Spalte und der letzten Zeile von $A^{(k)}$ nachzuweisen

$$\lim_{k\to\infty} A^{(k)} = \begin{pmatrix} \lambda_1 & * & \dots & \dots & * \\ & \vdots & & & \vdots \\ & \vdots & & & \vdots \\ & * & \dots & \dots & * \\ & & & & \lambda_n \end{pmatrix},$$

weil sich dann die Überlegungen auf die entstehende Teilmatrix der Dimension $n-2$ anwenden lassen.

- Zunächst stellt man mittels $X^{(k)} = Q^{(0)} \cdots Q^{(k-1)}$, d.h. es ist $X^{(k+1)} = X^{(k)} Q^{(k)}$, folgenden Zusammenhang her

$$A^{(k)} = \left(X^{(k)}\right)^* A X^{(k)} \quad \Longleftrightarrow \quad A X^{(k)} = X^{(k)} \underbrace{A^{(k)}}_{=Q^{(k)} R^{(k)}} = X^{(k+1)} R^{(k)} .$$

- Speziell für die erste Spalte der entstehenden Matrix ergibt sich aufgrund der oberen Dreiecksform von $R^{(k)}$

$$\begin{aligned} A X^{(k)} &= X^{(k+1)} R^{(k)} , \\ A X^{(k)}_{-,1} = X^{(k+1)} R^{(k)}_{-,1} &= R^{(k)}_{11} X^{(k+1)} e_1 = R^{(k)}_{11} X^{(k+1)}_{-,1} , \end{aligned}$$

  d.h. die Iteration für die erste Spalte entspricht einer direkten Vektoriteration

$$y^{(k+1)} = c^{(k)} A y^{(k)} , \qquad c^{(k)} = \tfrac{1}{R^{(k)}_{11}} > 0 , \quad y^{(k)} = X^{(k)}_{-,1} ,$$

  mit Konvergenz gegen einen Eigenvektor von A zum betragsmäßig größten Eigenwert λ_1.

- Adjunktion und Inversion der obigen Relation ergibt (Transformationsmatrizen sind unitär, d.h. es ist $T^{-1} = T^*$ und $\left(T^{-1}\right)^* = T$)

$$\begin{aligned} A X^{(k)} &= X^{(k+1)} R^{(k)} , \\ \left(X^{(k)}\right)^* A^* = \left(A X^{(k)}\right)^* &= \left(X^{(k+1)} R^{(k)}\right)^* = \left(R^{(k)}\right)^* \left(X^{(k+1)}\right)^* , \\ \left(A^{-1}\right)^* X^{(k)} &= X^{(k+1)} \left(\left(R^{(k)}\right)^*\right)^{-1} . \end{aligned}$$

  Speziell für die letzte Spalte der entstehenden Matrix ergibt sich aufgrund der unteren Dreiecksform von $\left(R^{(k)}\right)^*$

$$\begin{aligned} \left(A^{-1}\right)^* X^{(k)}_{-,n} = X^{(k+1)} \left(\left(R^{(k)}\right)^*\right)^{-1}_{-,n} &= \left(\left(R^{(k)}\right)^*\right)^{-1}_{n,n} X^{(k+1)} e_n = \tfrac{1}{R^{(k)}_{nn}} X^{(k+1)}_{-,n} , \\ z^{(k+1)} = d^{(k)} \left(A^{-1}\right)^* z^{(k)} , \qquad d^{(k)} = R^{(k)}_{nn} > 0 , \quad z^{(k)} &= X^{(k)}_{-,n} , \end{aligned}$$

  d.h. die Iteration für die letzte Spalte entspricht einer inversen Vektoriteration mit Konvergenz gegen einen Eigenvektor z von A^* zum betragsmäßig kleinsten Eigenwert μ. Wegen (Eigenrelation für die adjungierte Matrix und Zusammenhang mit den zugehörigen Eigenwerten und linksseitigen Eigenvektoren)

$$A^* z = \mu z \Leftrightarrow z^* A = \mu z^*$$

  stimmt μ mit λ_n überein.

– Insgesamt ergibt sich somit für die erste Spalte bzw. letzte Zeile von $A^{(k)}$ (vgl. Abschnitt 6.4)

$$A^{(k)} = \left(X^{(k)}\right)^* A X^{(k)},$$

$$A^{(k)}_{-,1} = A^{(k)} e_1 = \left(X^{(k)}\right)^* A X^{(k)} e_1 = \left(X^{(k)}\right)^* \underbrace{A X^{(k)}_{-,1}}_{\to \pm \lambda_1 v_1} \overset{k\to\infty}{\longrightarrow} \lambda_1 e_1,$$

$$A^{(k)}_{n,-} = e_n^T A^{(k)} = \left(A^* X^{(k)} e_n^T\right)^* X^{(k)} = \left(\underbrace{A^* X^{(k)}_{-,n}}_{\to \lambda_n z}\right)^* X^{(k)} \overset{k\to\infty}{\longrightarrow} \lambda_n e_n^T.$$

Dies ergibt die Behauptung. ◇

- **QR-Algorithmus mit Shift**: Im Allgemeinen wird eine Modifikation des QR-Algorithmus verwendet mit einem zusätzlichen **Shift**, ähnlich der Idee der inversen Vektoriteration. Damit verbessert man die Konvergenzrate der letzten Zeile der entstehenden Matrix $A^{(k)}$.

  Beispiel: Speziell für die folgende Matrix

$$\begin{pmatrix} 4 & 3 & 2 & 1 \\ 3 & 3 & 2 & 1 \\ 0 & 2 & 2 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

  ist die Konvergenzrate des QR-Algorithmus sehr dürftig. Ein zufriedenstellendes Ergebnis ergibt sich hingegen mittels des QR-Algorithmus mit Shift.

- **Bemerkung**: Nicht behandelt werden alternative Verfahren zu Eigenwertberechnungen. Für symmetrische und reelle Tridiagonalmatrizen verwendet ein numerisch stabiles Verfahren die **Bisektionsmethode nach Givens** oder **Sturmsche Ketten**. Ein weiteres Verfahren ist das **Verfahren von Arnoldi**.

7. Nichtlineare Gleichungssysteme

- **Problemstellung**: Betrachtet wird eine nichtlineare Funktion (wie üblich sei $\mathbb{R}^n = \mathbb{R}^{n\times 1}$)

$$f : \mathbb{R}^n \to \mathbb{R}^n : x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \mapsto f(x) = \begin{pmatrix} f_1(x_1, \ldots, x_n) \\ \vdots \\ f_n(x_1, \ldots, x_n) \end{pmatrix},$$

  die als hinreichend oft differenzierbar angenommen wird. Gesucht ist eine Näherungslösung an eine (zumindest lokal eindeutige) Lösung $\bar{x} \in \mathbb{R}^n$ des **nichtlinearen Gleichungssystems** (n Gleichungen, n Unbekannte)

$$f(x) = 0\,.$$

  Anwendungen:

  - Verfahren zur Lösung nichtlinearer Optimierungsprobleme
  - Verfahren zur Lösung nichtlinearer gewöhnlicher Differentialgleichungen
  - Verfahren zur Lösung nichtlinearer partieller Differentialgleichungen

  Vgl. auch den Zusammenhang mit technischen Regelkreisen, Skriptum, S. 116.

- **Bemerkungen**:
  - Im Gegensatz zu linearen Gleichungssystemen sind Resultate zur **Existenz und (lokalen) Eindeutigkeit** von Lösungen nichtlinearer Gleichungssysteme nicht allgemein sondern nur in speziellen Situationen gültig. Deshalb wird im Folgenden die (lokale) Existenz und Eindeutigkeit der Lösung des betrachteten nichtlinearen Gleichungssystems angenommen.
  - Hinsichtlich praktischer Anwendungen (komplexe Funktionsvorschrift, zusätzliche Berechnungen zur Funktionsauswertung erforderlich) ist es wesentlich, die Anzahl der Funktionsauswertungen von f möglichst gering zu halten.
  - Für den Spezialfall einer skalaren nichtlinearen Gleichung

$$f(x) = 0\,, \qquad f : \mathbb{R} \to \mathbb{R}\,,$$

    gibt es iterative Verfahren mit ausgezeichneten Konvergenzeigenschaften (sofern die Existenz und lokale Eindeutigkeit einer Lösung gesichert ist und ein einschließendes Intervall bekannt ist).

    In mehreren Dimensionen ist die Lösung eines nichtlinearen Gleichungssystems hingegen ein **schwieriges Problem**, und es gibt kein allgemein anwendbares und mit Sicherheit erfolgreiches numerisches Verfahren.

    Die Konvergenzeigenschaften der verwendeten Iterationsverfahren hängen wesentlich vom gewählten **Startwert** ab. Bisher gibt es für die Berechnung eines geeigneten Startwertes kein allgemein anwendbares Verfahren.

– Beachte, daß

$$f:\mathbb{R}^n\to\mathbb{R}^k:x\mapsto f(x),$$
$$f'(x):\mathbb{R}^n\to\mathbb{R}^k:y\mapsto f'(x)y,\qquad x\in\mathbb{R}^n,$$
$$f''(x):\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}^k:(y,z)\mapsto f''(x)(y,z),\qquad x\in\mathbb{R}^n,$$
$$f(x)=\begin{pmatrix}f_1(x_1,\dots,x_n)\\ \vdots\\ f_k(x_1,\dots,x_n)\end{pmatrix}\in\mathbb{R}^k,\qquad x\in\mathbb{R}^n,$$
$$f'(x)=\begin{pmatrix}\partial_{x_1}f_1(x_1,\dots,x_n) & \dots & \partial_{x_n}f_1(x_1,\dots,x_n)\\ \vdots & & \vdots\\ \partial_{x_1}f_k(x_1,\dots,x_n) & \dots & \partial_{x_n}f_k(x_1,\dots,x_n)\end{pmatrix}\in\mathbb{R}^{k\times n},\qquad x\in\mathbb{R}^n,$$
$$f''(x)(y,z)=\begin{pmatrix}\sum_{i,j=1}^n\partial_{x_ix_j}f_1(x_1,\dots,x_n)\,y_iz_j\\ \vdots\\ \sum_{i,j=1}^n\partial_{x_ix_j}f_k(x_1,\dots,x_n)\,y_iz_j\end{pmatrix}\in\mathbb{R}^k,\qquad x,y,z\in\mathbb{R}^n.$$

Insbesondere für $k=1$ ergibt sich (beachte $\partial_{x_ix_j}f=\partial_{x_jx_i}f$ sofern f zweimal stetig partiell differenzierbar)

$$f:\mathbb{R}^n\to\mathbb{R}:x\mapsto f(x),$$
$$f'(x):\mathbb{R}^n\to\mathbb{R}:y\mapsto f'(x)y,\qquad x\in\mathbb{R}^n,$$
$$f''(x):\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}:(y,z)\mapsto f''(x)(y,z),\qquad x\in\mathbb{R}^n,$$
$$f(x)=f(x_1,\dots,x_n)\in\mathbb{R},\qquad x\in\mathbb{R}^n,$$
$$f'(x)=\big(\partial_{x_1}f(x_1,\dots,x_n),\dots,\partial_{x_n}f(x_1,\dots,x_n)\big)\in\mathbb{R}^{1\times n},\qquad x\in\mathbb{R}^n,$$
$$f'(x)y=\sum_{i=1}^n\partial_{x_i}f(x_1,\dots,x_n)\,y_i\in\mathbb{R},\qquad x,y\in\mathbb{R}^n,$$
$$f''(x)=\begin{pmatrix}\partial_{x_1x_1}f(x_1,\dots,x_n) & \dots & \partial_{x_1x_n}f(x_1,\dots,x_n)\\ \vdots & & \vdots\\ \partial_{x_nx_1}f(x_1,\dots,x_n) & \dots & \partial_{x_nx_n}f(x_1,\dots,x_n)\end{pmatrix}\in\mathbb{R}^{n\times n},\qquad x\in\mathbb{R}^n,$$
$$f''(x)(y,z)=y^Tf''(x)z=\sum_{i,j=1}^n\partial_{x_ix_j}f(x_1,\dots,x_n)\,y_iz_j\in\mathbb{R},\qquad x,y,z\in\mathbb{R}^n.$$

- **Optimierungsprobleme**: Die Minimierung einer differenzierbaren Funktion

$$g(x)\xrightarrow{!}\min,\qquad g:\mathbb{R}^n\to\mathbb{R},$$

führt auf das nichtlineare Gleichungssystem

$$f(x)=0,\qquad f=g':\mathbb{R}^n\to\mathbb{R}^{1\times n}:x\mapsto f(x)=g'(x).$$

In dieser Situation gibt es zusätzlich alternative numerische Verfahren, die zusätzliche Eigenschaften der Ableitung von $f = g'$ ausnützen (beispielsweise die Symmetrie und positive Definitheit der Hessematrix $f'(x) = g''(x)$). Analoge Überlegungen gelten für Maximierungsprobleme (Betrachtung von $-g$).

- **Inhalt**:
  - Grundlegende Begriffe und Resultate
  - Verfahren für eindimensionale Probleme (Bisektionsverfahren und Modifikationen, Newton-Verfahren)
  - Verfahren für mehrdimensionale Probleme (Modifikationen des Newton-Verfahrens)

7.1. Grundbegriffe

- **Situation**: Es sei $f : \mathbb{R}^n \to \mathbb{R}^n$ eine hinreichend oft differenzierbare Funktion. Betrachtet wird das nichtlineare Gleichungssystem

$$f(x) = 0$$

mit (lokal) eindeutig bestimmter Lösung $\bar{x} \in \mathbb{R}^n$.

- **Iterationsverfahren, Fixpunktiteration, Konvergenz**:
  - Numerische Verfahren zur näherungsweisen Berechnung der Lösung des nichtlinearen Gleichungssystems $f(x) = 0$ sind **iterative Vefahren**. Ausgehend von einem (geeignet gewählten) Startwert $x_0 \in \mathbb{R}^n$ wird eine Folge $(x_k)_{k\geq 1}$ von Näherungswerten $x_k \in \mathbb{R}^n$ an $\bar{x}$ mittels einer Rekursion der Form

$$x_{k+1} = \varphi(x_k), \qquad k \geq 0,$$

mit **Iterationsfunktion** $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ berechnet. Dabei gilt es sicherzustellen, daß die Iteration (rasch) gegen die gesuchte Lösung konvergiert

$$\lim_{k\to\infty} x_k = \bar{x}.$$

  - Falls die Iterationsfunktion φ stetig ist und die Folge der Näherungswerte gegen $\bar{x}$ konvergiert, folgt

$$\bar{x} = \lim_{k\to\infty} x_{k+1} = \lim_{k\to\infty} \varphi(x_k) = \varphi\Big(\lim_{k\to\infty} x_k\Big) = \varphi(\bar{x}),$$

d.h. die Lösung des nichtlinearen Gleichungssystems ist ein **Fixpunkt** der Iterationsfunktion

$$f(\bar{x}) = 0 \quad \Longleftrightarrow \quad \varphi(\bar{x}) = \bar{x}.$$

Dies motiviert die Bezeichnung **Fixpunktiteration** für die obige Iteration.

  - Ein Iterationsverfahren heißt **global konvergent**, wenn man eine Menge $D \subset \mathbb{R}^n$ angeben kann, sodaß die Iteration für beliebige Startwerte $x_0 \in D$ gegen den (eindeutig bestimmten) Fixpunkt $\bar{x}$ konvergiert. Konvergiert die Iteration nur für Startwerte x_0, die *hinreichend nahe* beim Fixpunkt $\bar{x}$ liegen, so heißt das Iterationsverfahren **lokal konvergent**.
  - Es sei $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ eine festgelegte Norm.
    Eine Funktion $g : D \subset \mathbb{R}^n \to \mathbb{R}^n$ heißt **Lipschitz-stetig**, wenn es eine Konstante $L > 0$ gibt, sodaß für alle Elemente $x, \tilde{x} \in D$ die folgende Relation gilt

$$\|g(x) - g(\tilde{x})\| \leq L\,\|x - \tilde{x}\|.$$

Eine Lipschitz-stetige Funktion ist insbesondere stetig.

Eine Funktion $g : D \subset \mathbb{R}^n \to \mathbb{R}^n$ heißt eine **kontrahierende Abbildung** (**Kontraktion**), wenn es eine Konstante $0 < \kappa < 1$ gibt, sodaß für alle $x, \tilde{x} \in D$ die folgende Relation gilt (bezüglich einer festgelegten Norm $\|\cdot\|$)

$$\|g(x) - g(\tilde{x})\| \le \kappa \|x - \tilde{x}\|, \qquad 0 < \kappa < 1,$$

d.h. die Funktion g ist insbesondere Lipschitz-stetig mit Konstante $\kappa < 1$.

Resultat zur Existenz und Eindeutigkeit eines Fixpunktes: Für eine auf einer abgeschlossenen Menge definierte kontrahierende Selbstabbildung $g : D \subset \mathbb{R}^n \to \mathbb{R}^n$ (d.h. es gilt $g(D) \subset D$) sichert der **Banachsche Fixpunktsatz** die Existenz und Eindeutigkeit eines Fixpunktes $\bar{x} \in D$. Insbesondere konvergiert die Fixpunktiteration $x_{k+1} = g(x_k)$ für beliebige Startwerte $x_0 \in D$ gegen den Fixpunkt $\bar{x}$.

– Neben **Einschrittverfahren**

$$x_0 \text{ gegeben}, \qquad x_{k+1} = \varphi(x_k), \quad k \ge 0,$$

sind (insbesondere im Zusammenhang mit numerischen Verfahren zur Lösung nichtlinearer Differentialgleichungen) auch **Mehrschrittverfahren** gebräuchlich. Dabei verwendet man mehrere bekannte Approximationswerte zur Bestimmung des neuen Näherungswertes ($x_m = \varphi(x_0, \dots, x_{m-1})$, $x_{m+1} = \varphi(x_1, \dots, x_m)$ etc.)

$$x_0, \dots, x_{m-1} \text{ gegeben}, \qquad x_{m+k} = \varphi(x_k, \dots, x_{m-1+k}), \quad k \ge 0,$$

Mittels der Umformulierung

$$X_0 = \begin{pmatrix} x_0 \\ \vdots \\ x_{m-1} \end{pmatrix}, \qquad X_k = \begin{pmatrix} x_k \\ \vdots \\ x_{k+m-1} \end{pmatrix}, \quad \Phi(X_k) = \begin{pmatrix} x_{k+1} \\ \vdots \\ x_{k+m-1} \\ \varphi(X_k) \end{pmatrix}, \quad k \ge 0,$$

läßt sich jedes Mehrschrittverfahren auf ein Einschrittverfahren zurückführen (für theoretische Untersuchungen)

$$X_0 \text{ gegeben}, \qquad X_{k+1} = \Phi(X_k), \quad k \ge 0.$$

– **Konvergenzordnung einer Iteration** (Definition 7.1): Für alle Startwerte $x_0 \in D$ sei die Iteration konvergent

$$x_{k+1} = \varphi(x_k), \quad k \ge 0, \qquad \lim_{k \to \infty} x_k = \bar{x}.$$

Falls es einen Index $K_0 \in \mathbb{N}$ und eine Konstante $c > 0$ gibt, sodaß die folgende Abschätzung mit $p > 0$ (als Supremum) gilt, heißt p die **Konvergenzordnung** des Iterationsverfahrens

$$\|x_{k+1} - \bar{x}\| \le c \|x_k - \bar{x}\|^p \qquad \text{für alle } k \ge K_0.$$

Im Fall $p = 1$ spricht man von **linearer Konvergenz** und die Konstante c heißt **Konvergenzfaktor**. Falls zusätzlich

$$\lim_{k\to\infty} \frac{\|x_{k+1} - \bar{x}\|}{\|x_k - \bar{x}\|} = 0\,,$$

spricht man von **superlinearer Konvergenz**. Im Fall $p = 2$ spricht man von **quadratischer Konvergenz**.

Bemerkung: Falls die Iterationsfunktion die folgenden Relationen erfüllt

$$\varphi^{(\ell)}(\bar{x}) = 0\,, \quad 1 \le \ell \le p-1\,, \qquad \varphi^{(p)}(\bar{x}) \neq 0\,,$$

folgt mittels einer Taylorreihenentwicklung (beispielsweise für den skalaren Fall, ebenso für den allgemeinen Fall)

$$\begin{aligned} x_{k+1} - \bar{x} &= \varphi(x_k) - \varphi(\bar{x}) \\ &= \sum_{\ell=0}^{p} \tfrac{1}{k!}\varphi^{(\ell)}(\bar{x})\,(x_k - \bar{x})^k + \mathcal{O}\big(\|x_k - \bar{x}\|^{p+1}\big) - \varphi(\bar{x}) \\ &= \tfrac{1}{p!}\varphi^{(p)}(\bar{x})\,(x_k - \bar{x})^p + \mathcal{O}\big(\|x_k - \bar{x}\|^{p+1}\big)\,, \\ \|x_{k+1} - \bar{x}\| &\le \Big(\tfrac{1}{p!}\,\|\varphi^{(p)}(\bar{x})\| + \mathcal{O}\big(\|x_k - \bar{x}\|\big)\Big)\,\|x_k - \bar{x}\|^p\,, \end{aligned}$$

und somit ist die Konvergenzordnung des Verfahrens p.

- Eine **Linearisierung** der Funktion f und Iteration führt auf das **Newton-Verfahren**

$$\begin{gathered} 0 = f(\bar{x}) = f(x_k) + f'(x_k)(\bar{x} - x_k) + \mathcal{O}\big(\|\bar{x} - x_k\|^2\big)\,, \\ 0 \approx f(x_k) + f'(x_k)(\bar{x} - x_k) \quad \Longleftrightarrow \quad \bar{x} \approx x_k - \big(f'(x_k)\big)^{-1} f(x_k)\,, \\ x_{k+1} = x_k - \big(f'(x_k)\big)^{-1} f(x_k)\,. \end{gathered}$$

  - Im **eindimensionalen Fall** ergibt sich

$$x_0 \text{ gegeben}\,, \qquad x_{k+1} = x_k - \tfrac{f(x_k)}{f'(x_k)}\,, \quad k \ge 0\,.$$

    Veranschaulichung (ersetze Funktion durch lineare Approximation bei x_k und bestimme Nullstelle), vgl. Skriptum, S. 114.

    Im **mehrdimensionalen Fall** erfordert das Newton-Verfahren in jedem Iterationsschritt die Lösung eines linearen Gleichungssystems (verwende Hilfsbezeichnung $\xi_k = x_k - x_{k+1}$)

$$x_0 \text{ gegeben}\,, \qquad \begin{cases} f'(x_k)\,\xi_k = f(x_k)\,, \\ x_{k+1} = x_k - \xi_k\,, \end{cases} \quad k \ge 0\,.$$

– Sofern die Matrix $f'(x_k)$ für $k \geq 0$ nicht singulär ist (oder die Matrix $f'(\bar{x})$ invertierbar ist und x_k für $k \geq 0$ *nahe genug* bei der gesuchten Lösung $\bar{x}$ liegt) ist das Newton-Verfahren wohldefiniert.

  Die zugehörige Iterationsfunktion lautet

$$\varphi(x) = x - \left(f'(x)\right)^{-1} f(x)\,.$$

  Im Allgemeinen ist das Newton-Verfahren nur lokal konvergent, vgl. auch Abbildung, Skriptum S. 117.

  Im Spezialfall $f : \mathbb{R} \to \mathbb{R}$ ergibt sich (Produktregel, $f(\bar{x}) = 0$)

$$\begin{aligned}
\varphi(x) &= x - \tfrac{f(x)}{f'(x)}\,, \\
\varphi'(x) &= 1 - \tfrac{f'(x)}{f'(x)} + \tfrac{f(x)\,f''(x)}{(f'(x))^2} = \tfrac{f(x)\,f''(x)}{(f'(x))^2}\,, \qquad \varphi'(\bar{x}) = 0\,, \\
\varphi''(x) &= \tfrac{f''(x)}{f'(x)} + \tfrac{f(x)\,f'''(x)}{(f'(x))^2} - 2\,\tfrac{f(x)\,(f''(x))^2}{(f'(x))^3}\,, \qquad \varphi''(\bar{x}) = \tfrac{f''(\bar{x})}{f'(\bar{x})} \underset{\text{i.A.}}{\neq} 0\,,
\end{aligned}$$

  d.h. das **Newton-Verfahren ist quadratisch konvergent** (in einer geeigneten Umgebung einer einfachen Nullstelle $\bar{x}$). Dasselbe Resultat gilt im mehrdimensionalen Fall (ohne Begründung).

  Beispiel: Quadratische Konvergenz entspricht einer **Verdoppelung der korrekten Dezimalziffern** ($1 \to 2 \to 4 \to 8 \to 16$ korrekte Stellen)

$$\begin{aligned}
\|x_j - \bar{x}\| \approx \tfrac{1}{10} \quad &\longrightarrow \quad \|x_{j+1} - \bar{x}\| \approx \tfrac{1}{10^2} \quad \longrightarrow \quad \|x_{j+2} - \bar{x}\| \approx \tfrac{1}{10^4} \\
&\longrightarrow \quad \|x_{j+3} - \bar{x}\| \approx \tfrac{1}{10^8} \quad \longrightarrow \quad \|x_{j+4} - \bar{x}\| \approx \tfrac{1}{10^{16}}\,.
\end{aligned}$$

– Vorteil des Newton-Verfahrens

  * Ausgezeichnetes lokales Konvergenzverhalten (quadratische Konvergenz)

  Nachteile des Newton-Verfahrens

  * Im Allgemeinen keine globale Konvergenz
  * Notwendigkeit der Berechnung von $f'(x_k)$ in jedem Iterationsschritt

- **Kondition des Nullstellenproblems** (skalare nichtlineare Gleichung): Betrachte die skalare nichtlineare Gleichung

$$f(x) = 0\,, \qquad f : \mathbb{R} \to \mathbb{R}\,, \qquad \text{Lösung } \bar{x}\,,$$

  wobei $\bar{x} \in \mathbb{R}$ eine m-fache Nullstelle von f sei

$$f^{(i)}(\bar{x}) = 0\,, \quad 0 \leq i \leq m-1\,, \qquad f^{(m)}(\bar{x}) \neq 0\,.$$

  Betrachte weiters die veränderte skalare nichtlineare Gleichung

$$g(x) = 0\,, \qquad g : \mathbb{R} \to \mathbb{R}\,, \qquad \text{Lösung } \bar{y}\,,$$

unter der Annahme (Abschwächung durch Betrachtung einer Umgebung der Lösung $\bar{x}$)

$$|g(x) - f(x)| \le \varepsilon\,, \qquad x \in \mathbb{R}\,.$$

Zu untersuchen ist, wie groß die Abweichung $|\bar{x} - \bar{y}|$ der zugehörigen Lösungen ist. Mittels einer Taylorreihenentwicklung ergibt sich (verwende, daß $\bar{x}$ eine m-fache Nullstelle von f ist)

$$f(\bar{y}) = \sum_{i=0}^{m} \tfrac{1}{i!} f^{(i)}(\bar{x})\,(\bar{y}-\bar{x})^i + \mathscr{O}\big(|\bar{y}-\bar{x}|^{m+1}\big) = \tfrac{1}{m!} f^{(m)}(\bar{x})\,(\bar{y}-\bar{x})^m + \mathscr{O}\big(|\bar{y}-\bar{x}|^{m+1}\big)\,.$$

Mittels der Relation $g(\bar{y}) = 0$ folgt

$$\begin{aligned}
f(\bar{y}) - g(\bar{y}) &= \tfrac{1}{m!} f^{(m)}(\bar{x})\,(\bar{y}-\bar{x})^m + \mathscr{O}\big(|\bar{y}-\bar{x}|^{m+1}\big)\,,\\
\tfrac{m!}{f^{(m)}(\bar{x})}\big(f(\bar{y}) - g(\bar{y})\big) &= \Big(1 + \mathscr{O}\big(|\bar{y}-\bar{x}|\big)\Big)(\bar{y}-\bar{x})^m\,,\\
(\bar{y}-\bar{x})^m &= \Big(1 + \mathscr{O}\big(|\bar{y}-\bar{x}|\big)\Big)\,\tfrac{m!}{f^{(m)}(\bar{x})}\big(f(\bar{y}) - g(\bar{y})\big)\,,\\
|\bar{y}-\bar{x}| &= \Big(1 + \mathscr{O}\big(|\bar{y}-\bar{x}|\big)\Big)^{\frac{1}{m}} \Big(\tfrac{m!}{|f^{(m)}(\bar{x})|}\Big)^{\frac{1}{m}} \big|f(\bar{y}) - g(\bar{y})\big|^{\frac{1}{m}}\,,\\
|\bar{y}-\bar{x}| &\le \Big(1 + \mathscr{O}\big(|\bar{y}-\bar{x}|\big)\Big)^{\frac{1}{m}} \Big(\tfrac{m!}{|f^{(m)}(\bar{x})|}\Big)^{\frac{1}{m}} \varepsilon^{\frac{1}{m}}\,.
\end{aligned}$$

Dies führt auf die Abschätzung (Vernachlässigung von $\mathscr{O}(|\bar{y}-\bar{x}|)$ für $m \ge 2$)

$$\begin{aligned}
m = 1: &\qquad |\bar{y}-\bar{x}| \le \Big(1 + \tfrac{C\varepsilon}{|f'(\bar{x})|}\Big)\,\tfrac{1}{|f'(\bar{x})|}\,\varepsilon\,,\\
m \ge 2: &\qquad |\bar{y}-\bar{x}| \le C\Big(\tfrac{m!}{|f^{(m)}(\bar{x})|}\Big)^{\frac{1}{m}} \varepsilon^{\frac{1}{m}}\,.
\end{aligned}$$

Daraus folgt, daß die Berechnung mehrfacher Nullstellen ein schlecht konditioniertes Problem ist (für $0 < \varepsilon \ll 1$ ist $\varepsilon^{\frac{1}{m}} \gg \varepsilon$). Ebenso ist die Berechnung einer einfachen Nullstelle schlecht konditioniert, falls $|f^{(m)}(\bar{x})| \approx 0$ (schleifender Schnitt).

Rundungsfehleranalyse von Fixpunktiterationen: Unter dem Einfluß von Rundungsfehlern wird anstelle der Folge $(x_k)_{k\ge0}$ eine Folge $(y_k)_{k\ge0}$ berechnet (mit Startwert $y_0 = x_0 + \delta_0$, die Größen δ_k beschreiben die Rundungsfehler beim Auswerten von φ)

$$x_{k+1} = \varphi(x_k)\,, \qquad y_{k+1} = \varphi(y_k) + \delta_k\,, \quad k \ge 0\,.$$

Beide Fixpunktiterationen seien konvergent mit Grenzwerten $\bar{x}$ und $\bar{y}$, d.h. es gilt insbesondere $\varphi(\bar{x}) = \bar{x}$. Weiters gelte

$$|\varphi'(x)| \le c < 1\,.$$

Mit Hilfe einer Taylorreihenentwicklung (Mittelwertsatz) folgt (für ein $\zeta_k \in (y_k, \bar{x})$ falls

$y_k < \bar{x}$ oder für ein $\zeta_k \in (\bar{x}, y_k)$ falls $y_k > \bar{x}$, Bezeichnung $d_k = y_k - \bar{x}$)

$$\begin{aligned} d_{k+1} &= \underbrace{y_{k+1}}_{=\varphi(y_k)+\delta_k} - \underbrace{\bar{x}}_{=\varphi(\bar{x})} = \varphi(y_k) - \varphi(\bar{x}) + \delta_k = \varphi'(\zeta_k)\,(y_k - \bar{x}) + \delta_k \\ &= \varphi'(\zeta_k)\,(y_k - y_{k+1} + d_{k+1}) + \delta_k\,, \qquad k \geq 0\,, \\ \Longrightarrow \quad d_{k+1} &= \tfrac{1}{1-\varphi'(\zeta_k)} \left(\varphi'(\zeta_k)\,(y_k - y_{k+1}) + \delta_k\right), \qquad k \geq 0\,, \end{aligned}$$

und weiters (Abschwächung der Voraussetzung $|\varphi'(\zeta_k)| \leq c < 1$ für $k \geq 0$ ausreichend, geometrische Reihe)

$$|d_{k+1}| \leq \tfrac{1}{|1-\varphi'(\zeta_k)|} \left(|\varphi'(\zeta_k)|\,|y_k - y_{k+1}| + |\delta_k|\right) \leq \tfrac{c}{1-c}\,|y_k - y_{k+1}| + \tfrac{1}{1-c}\,|\delta_k|\,, \qquad k \geq 0\,.$$

Dies zeigt insbesondere, daß es bei einer Fixpunktiteration zu keiner Akkumulation von Rundungsfehlern kommt (Abhängigkeit vom aktuellen Iterationsschritt)

$$|y_{k+1} - \bar{x}| \leq \tfrac{c}{1-c}\,|y_k - y_{k+1}| + \tfrac{1}{1-c}\,|\delta_k|\,, \qquad k \geq 0\,.$$

7.2. Verfahren für den eindimensionalen Fall

- **Bisektionsverfahren**:
  - **Situation**: Es sei $f : \mathbb{R} \to \mathbb{R}$ eine stetige Funktion und $[a,b] \subset \mathbb{R}$ ein Intervall (**Einschließungsintervall**) derart, daß

    $$f(a)\,f(b) < 0\,.$$

    Unter diesen Voraussetzungen existiert (mindestens) eine Nullstelle $\bar{x} \in (a,b)$ (Zwischenwertsatz). Oft wird zudem angenommen, daß das Intervall so gewählt ist, daß es genau eine Nullstelle in (a,b) gibt.
  - Das **Bisektionsverfahren** ist ein Zweischrittverfahren. Ausgehend von den Intervallgrenzen a,b und den zugehörigen Funktionswerten $f(a), f(b)$ verwendet es die Berechnung des Intervallmittelpunktes $c = \frac{a+b}{2}$ und des zugehörigen Funktionswerts $f(c)$.
    - Falls $f(a)\,f(c) = 0$ ist, ist c die gesuchte Nullstelle.
    - Falls $f(a)\,f(c) < 0$ ist, liegt die Nullstelle im Intervall (a,c) und man setzt $b = c$.
    - Falls $f(a)\,f(c) > 0$ ist, liegt die Nullstelle im Intervall (c,b) und man setzt $a = c$.

    Die Fortführung der Intervallhalbierung ergibt eine Folge von Näherungswerten x_{k+1} (Intervallmittelpunkte) an die gesuchte Nullstelle $\bar{x}$.
  - Zum Nachweis der Konvergenz des Bisektionsverfahrens verwendet man, daß eine streng monotone und beschränkte Folge konvergiert und die Länge der entstehenden Intervalle gegen Null geht. Unter den obigen Voraussetzungen ist das Bisektionsverfahren **global konvergent** mit **Konvergenzordnung** $p = 1$ (lineare Konvergenz) und Konvergenzfaktor $c = \frac{1}{2}$

    $$|x_{k+1} - \bar{x}| \leq \tfrac{1}{2}\,|x_k - \bar{x}|\,.$$

  - Als **Abbruchkriterium** wählt man meist die Länge des Intervalls

    $$b - a < \mathrm{tol}\,,$$

    eventuell zusammen mit dem Kriterium $|f(c)| < \mathrm{tol}$ (insbesondere bei schleifenden Schnitten wäre $|f(c)| < \mathrm{tol}$ als alleiniges Abbruchkriterium jedoch wenig aussagekräftig). Bei der Wahl der Toleranz sollte die Größe der Nullstelle $\bar{x}$ miteinbezogen werden.
  - Vgl. Pseudo-Code, Skriptum, S. 120.

- **Regula falsi**:

– Die Regula falsi ist eine Modifikation des Bisektionsverfahren. Anstelle des Intervallmittelpunktes $c = \frac{a+b}{2}$ wird die Nullstelle der Sekante durch die Intervallgrenzen berechnet, d.h. die Bedingung $0 = g(x) = f(a) + \frac{f(b)-f(a)}{b-a}(x-a)$ (insbesondere ist $g(a) = f(a)$ und $g(b) = f(b)$) führt auf $\frac{f(b)-f(a)}{b-a}(c-a) = -f(a)$ und weiters (wegen $f(a)\,f(b) < 0$ ist Verfahren wohldefiniert)

$$c = a - \frac{b-a}{f(b)-f(a)}\,f(a)\,.$$

– Unter der Annahme, daß die Funktion f hinreichend oft differenzierbar ist, ist die Regula falsi **global konvergent** mit **Konvergenzordnung** $p = 1$ (lineare Konvergenz).

$$|x_{k+1} - \bar{x}| \le \kappa\,|x_k - \bar{x}|\,.$$

Allerdings kann der Fall eintreten, daß der Konvergenzfaktor κ größer als der Konvergenzfaktor $\kappa = \frac{1}{2}$ des Bisektionsverfahrens ist und die Regula falsi somit langsamer konvergiert.

Denn: Zur einfacheren Untersuchung des Konvergenzverhaltens der Regula falsi wird angenommen, daß die Funktion f konkav (d.h. $f'' < 0$, Graph von f oberhalb jeder Sekante, z.B. $f(x) = -x^2$ und $f''(x) = -2 < 0$) bzw. konvex ist (d.h. $f'' > 0$, Graph von f unterhalb jeder Sekante, z.B. $f(x) = x^2$ und $f''(x) = 2 > 0$). In dieser Situation bleibt eine Intervallgrenze unverändert und die Regula falsi vereinfacht sich zu einem Einschrittverfahren. Man beachte, daß der Schnittpunkt der Sekante gegen die gesuchte Nullstelle konvergiert, die Länge des einschließenden Intervalles konvergiert jedoch *nicht* gegen Null. Beispielsweise für den Fall, daß das linke Intervallende a fest bleibt, sind die Iterationswerte gegeben durch

$$x_{k+1} = \varphi(x_k) = a - \frac{(x_k-a)\,f(a)}{f(x_k)-f(a)} = \frac{a\,(f(x_k)-f(a))-(x_k-a)\,f(a)}{f(x_k)-f(a)} = \frac{a\,f(x_k)-x_k\,f(a)}{f(x_k)-f(a)}\,.$$

Die Ableitung der Iterationsfunktion bei $\bar{x}$ beschreibt die Konvergenzrate des Verfahrens (verwende $f(\bar{x}) = 0$)

$$\begin{aligned}
\varphi(x) &= \frac{a\,f(x)-x\,f(a)}{f(x)-f(a)}\,,\\
\varphi'(x) &= \frac{a\,f'(x)-f(a)}{f(x)-f(a)} - \frac{(a\,f(x)-x\,f(a))\,f'(x)}{(f(x)-f(a))^2}\,,\\
\varphi'(\bar{x}) &= \frac{f(a)-a\,f'(\bar{x})}{f(a)} + \frac{\bar{x}\,f(a)\,f'(\bar{x})}{f(a)^2} = \frac{f(a)+(\bar{x}-a)\,f'(\bar{x})}{f(a)}\,.
\end{aligned}$$

Mit Hilfe des Mittelwertsatzes folgt damit die Abschätzung (ohne Begründung)

$$\kappa = \left|1 + \frac{(\bar{x}-a)\,f'(\bar{x})}{f(a)}\right| < 1$$

und damit die Konvergenz des Verfahrens. ◇

- Der **Mittelwertsatz der Differentialrechnung** besagt, daß es für eine auf einem abgeschlossenen Intervall definierte und stetige Funktion $f : [a,b] \to \mathbb{R}$, die im Inneren differenzierbar ist, einen Punkt $\xi \in (a,b)$ gibt, sodaß $(b-a)\,f'(\xi) = f(b) - f(a)$.

- **Sekantenverfahren**:
  - Im Gegensatz zum Bisektionsverfahren und der Regula falsi wird beim **Sekantenverfahren** die Nullstelle der Sekante durch die vorherigen Iterationswerte x_{k-1} und x_k bestimmt, d.h. es gilt

    $$x_{k+1} = x_k - \tfrac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k).$$

    Eine Reduktion dieses Zweischrittverfahrens auf ein Einschrittverfahren ist nicht möglich.
  - Das Sekantenverfahren erfüllt die Relation (mit ξ_k zwischen $x_{k-1}, x_k, \bar{x}$ und ζ_k zwischen x_{k-1} und x_k, ohne Begründung)

    $$x_{k+1} - \bar{x} = \tfrac{f''(\xi_k)}{2f'(\zeta_k)} (x_k - \bar{x})(x_{k-1} - \bar{x}), \qquad k \geq 0.$$

    Sofern die zweite Ableitung von f stetig (und damit beschränkt auf $[a,b]$) und die Nullstelle einfach ist (und somit $f' > 0$ auf $[a,b]$), folgt die Abschätzung

    $$\|x_{k+1} - \bar{x}\| \leq c \|x_k - \bar{x}\| \|x_{k-1} - \bar{x}\|, \qquad k \geq 0.$$

  - Die **Konvergenzordnung** des Sekantenverfahrens ist $p = \lambda_1 = \frac{1}{2}(1+\sqrt{5})$ (insbesondere keine natürliche Zahl), d.h. es gilt

    $$\|x_{k+1} - \bar{x}\| \leq C \|x_k - \bar{x}\|^{\lambda_1}, \qquad k \geq 0.$$

    Denn: Es bezeichne (ohne Einschränkung der Allgemeinheit sei angenommen, daß $\delta_k > 0$ für $k \geq 0$ und $c > 0$)

    $$\delta_k = \|x_k - \bar{x}\|, \quad \eta_k = \ln(c\,\delta_k), \quad \delta_k = \tfrac{1}{c} e^{\eta_k}, \qquad k \geq 0.$$

    Dann gilt (Multiplikation mit $c > 0$, Logarithmieren)

    $$\begin{aligned} \delta_{k+1} &\leq c\,\delta_k \delta_{k-1}, && k \geq 0, \\ c\,\delta_{k+1} &\leq c\,\delta_k\, c\,\delta_{k-1}, && k \geq 0, \\ \eta_{k+1} &\leq \eta_k + \eta_{k-1}, && k \geq 0. \end{aligned}$$

    Die Betrachtung der zugehörigen Gleichung

    $$\zeta_{k+1} = \zeta_k + \zeta_{k-1}, \qquad k \geq 0,$$

    ist ausreichend, weil aus der Abschätzung $\eta_j \leq \zeta_j$ für alle $0 \leq j \leq k$ insbesondere die Relation $\eta_{k+1} \leq \eta_k + \eta_{k-1} \leq \zeta_k + \zeta_{k-1} = \zeta_{k+1}$ folgt. Die Lösung dieser linearen Dreitermrekursion (**Fibonacci-Folge**) verwendet die Umformulierung der Iteration als Einschrittverfahren

    $$X_k = \begin{pmatrix} \zeta_k \\ \zeta_{k+1} \end{pmatrix} = \begin{pmatrix} \zeta_k \\ \zeta_k + \zeta_{k-1} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \zeta_{k-1} \\ \zeta_k \end{pmatrix} = A\,X_{k-1} = A^k X_0, \quad A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \quad k \geq 0,$$

und die Eigenwertzerlegung der Matrix A

$$\chi(\lambda) = \det(A - \lambda I) = \det\begin{pmatrix} -\lambda & 1 \\ 1 & 1-\lambda \end{pmatrix} = \lambda^2 - \lambda - 1 = 0\,, \quad \lambda_{1,2} = \tfrac{1}{2}\big(1 \pm \sqrt{5}\big)\,,$$

$$\lambda_1 = \tfrac{1}{2}\big(1-\sqrt{5}\big): \quad \begin{pmatrix} -\lambda_1 & 1 \\ 1 & 1-\lambda_1 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad v_2 = \lambda_1 v_1\,, \quad v = v_1\begin{pmatrix} 1 \\ \lambda_1 \end{pmatrix},$$

$$\lambda_2 = \tfrac{1}{2}\big(1+\sqrt{5}\big): \quad \begin{pmatrix} -\lambda_2 & 1 \\ 1 & 1-\lambda_2 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad v_2 = \lambda_2 v_1\,, \quad v = v_1\begin{pmatrix} 1 \\ \lambda_2 \end{pmatrix},$$

$$A = V\Lambda V^{-1}\,, \qquad \Lambda = \begin{pmatrix} \lambda_1 & \\ & \lambda_2 \end{pmatrix}, \quad V = \begin{pmatrix} 1 & 1 \\ \lambda_1 & \lambda_2 \end{pmatrix}, \quad V^{-1} = \tfrac{1}{\lambda_2-\lambda_1}\begin{pmatrix} \lambda_2 & -1 \\ -\lambda_1 & 1 \end{pmatrix}.$$

Dies führt auf (beachte $\lambda_1 \approx 1.618$ und $\lambda_2 \approx -0.618$, eventuelle Vergrößerung des kleinsten Index)

$$A^k = V\Lambda^k V^{-1}\,, \quad A^k X_0 = V\Lambda^k V^{-1} X_0\,,$$

$$\zeta_k = \tfrac{1}{\lambda_2-\lambda_1}\left((\lambda_2\zeta_0 - \zeta_1)\,\lambda_1^k + (\zeta_1 - \lambda_1\zeta_0)\,\lambda_2^k\right) \approx \gamma\,\lambda_1^k\,, \quad \gamma = \tfrac{\lambda_2\zeta_0-\zeta_1}{\lambda_2-\lambda_1}\,, \qquad k \geq 0\,.$$

Beachte, daß dies dem Ansatz $\zeta_k = \lambda^k$ entspricht. Es folgen die Relationen (für kleinsten Index hinreichend groß)

$$\zeta_k \approx \gamma\,\lambda_1^k\,, \quad \zeta_{k+1} \approx \lambda_1\zeta_k\,, \qquad k \geq 0\,,$$

$$\widetilde{\delta}_k = \tfrac{1}{c}\,\mathrm{e}^{\zeta_k} \approx \tfrac{1}{c}\,\mathrm{e}^{\gamma\lambda_1^k}\,, \qquad k \geq 0\,,$$

$$\widetilde{\delta}_{k+1} = \tfrac{1}{c}\,\mathrm{e}^{\zeta_{k+1}} \approx \tfrac{1}{c}\,\mathrm{e}^{\gamma\lambda_1^{k+1}} = \tfrac{1}{c}\left(\mathrm{e}^{\gamma\lambda_1^k}\right)^{\lambda_1} = c^{\lambda_1-1}\left(\tfrac{1}{c}\,\mathrm{e}^{\gamma\lambda_1^k}\right)^{\lambda_1} \approx c^{\lambda_1-1}\,\widetilde{\delta}_k^{\lambda_1}\,, \qquad k \geq 0\,.$$

$$\|x_{k+1} - \bar{x}\| \leq c^{\lambda_1-1}\,\|x_k - \bar{x}\|^{\lambda_1}\,, \qquad k \geq 0\,.$$

Somit ergibt sich die Behauptung. ◇

– **Bemerkung**: Ein Vergleich der Konvergenzrate und der benötigten Funktionsauswertungen ergibt, daß das Sekantenverfahrens effizienter als das Newton-Verfahren ist, jedoch ebenfalls nicht global konvergent.

- **Bemerkungen**:
  - Der **Dekker-Algorithmus** kombiniert die Idee des Bisektionsverfahrens und des Sekantenverfahrens und besitzt gute lokale und globale Konvergenzeigenschaften (Sekantenschritte, die das Einschließungsintervall verlassen würden bzw. nahezu verlassen würden, werden durch Bisektionsschritte ersetzt).
  - Das **Verfahren von Muller** mit Konvergenzrate $\kappa \approx 1.84$ erweitert das Sekantenverfahren (quadratisches Polynom durch drei aufeinanderfolgende Iterationswerte $(x_j, f(x_j))$ für $j = k-2, k-1, k$, Nullstelle ist neuer Iterationswert).
  - Das **Verfahren von Brent** verwendet die Idee der inversen Interpolation (quadratisches Polynom durch $(f(x_j), x_j)$ für $j = k-2, k-1, k$, Auswerten bei Null ergibt den neuen Iterationswert) und die des Dekker-Algorithmus.

7.3. Der mehrdimensionale Fall

- **Vorbemerkung**: Bereits bei zwei nichtlinearen Gleichungen in zwei Unbekannten

$$f(x,y)=0\,,\qquad g(x,y)=0\,,$$

zeigen sich die Schwierigkeiten bei der Lösung eines nichtlinearen Gleichungssystems. Vgl. Graphik, Skriptum, S. 124 (Lösungen sind durch den **Schnitt der Höhenlinien** gegeben).

- Sämtliche praktikablen Verfahren zur näherungsweisen Lösung eines nichtlinearen Gleichungssystems sind Modifikationen des **Newton-Verfahrens** mit
  - besseren Konvergenzeigenschaften und
  - verringertem Aufwand bei der Berechnung der ersten Ableitung.

- **Quasi-Newton-Verfahren**:
  - Das Quasi-Newton-Verfahren beruht auf der Idee, das lineare Gleichungssystem

$$f(x)=0\,,\qquad f:\mathbb{R}^n\to\mathbb{R}^n\,,$$

mit dem Minimierungsproblem

$$F(x)=\tfrac{1}{2}\,\|f(x)\|_2^2\,,\qquad F:\mathbb{R}^n\to\mathbb{R}\,,$$

in Verbindung zu bringen. Eine Lösung des linearen Gleichungssystems ist ein globales Minimum von F (wegen $\|f(x)\|_2=0\Leftrightarrow f(x)=0$). Deshalb sollten die gewählten Iterationsschritte in Abstiegsrichtung sein, d.h. es sollte gelten

$$F(x_{k+1})<F(x_k)\,,\qquad k\geq 0\,.$$

Das Newton-Verfahren erfüllt diese Bedingung, denn mit (beachte $F'(x)\in\mathbb{R}^{1\times n}$)

$$F(x)=\tfrac{1}{2}\,\|f(x)\|_2^2=\tfrac{1}{2}\,\big(f(x)\big)^T f(x)=\tfrac{1}{2}\sum_{i=1}^{n}\big(f_i(x)\big)^2\,,$$

$$\partial_{x_\ell}F(x)=\sum_{i=1}^{n}f_i(x)\underbrace{\partial_{x_\ell}f_i(x)}_{=(f'(x))_{i\ell}}=\sum_{i=1}^{n}(f'(x))^T_{\ell i}\,f_i(x)=\big(\big(f'(x)\big)^T f(x)\big)_{\ell 1}\,,$$

$$F'(x)=\big(f(x)\big)^T f'(x)\,,$$

und mittels einer Taylorreihenentwicklung ergibt sich (sofern $\mathcal{O}\big(\|\xi_k\|_2^2\big)$ hinreichend klein)

$$x_{k+1}=x_k-\xi_k\,,\qquad \xi_k=\big(f'(x_k)\big)^{-1}f(x_k)\,,$$

$$F'(x_k)\,\xi_k=(f(x_k))^T f'(x_k)(f'(x_k))^{-1}f(x_k)=\|f(x_k)\|_2^2\,,$$

$$F(x_{k+1})=F(x_k)-F'(x_k)\,\xi_k+\mathcal{O}\big(\|\xi_k\|_2^2\big)=F(x_k)-\|f(x_k)\|_2^2+\mathcal{O}\big(\|\xi_k\|_2^2\big)<F(x_k)\,,$$

d.h. ein Iterationsschritt des Newtonverfahrens entspricht einem Schritt in Richtung abnehmender Funktionswerte von F.

– Obwohl Iterationsschritte des Newtonverfahrens in Abstiegsrichtung erfolgen, kann es passieren, daß die Schrittlänge zu groß ist und somit nicht die optimale Verkleinerung der Funktionswerte von F erreicht wird. Beim **Quasi-Newton-Verfahren** verwendet man deshalb stattdessen die Iterationsfunktion

$$x_{k+1} = x_k - \lambda_k \xi_k\,, \quad \xi_k = \big(f'(x_k)\big)^{-1} f(x_k)\,, \qquad k \geq 0\,.$$

Die zusätzliche **Schrittweite** λ_k wird dabei so bestimmt, daß $F(x_k - \lambda_k \xi_k)$ (zumindest näherungsweise) minimal wird (**Liniensuche**: Quadratischer (oder kubischer) Ansatz $F(x_k - \lambda\,\xi_k) \approx p(\lambda) = a_0 + b_0\,\lambda + c_0\,\lambda^2$ mit Kenntnis der Funktionswerte bei $\lambda = 0,1$ sowie der Ableitung $-F'(x_k - \lambda\,\xi_k)\,\xi_k \approx p'(\lambda) = b_0 + 2\,c_0\,\lambda$ bei $\lambda = 0$ führt auf die berechenbaren Koeffizienten $a_0 \approx F(x_k), b_0 \approx -F'(x_k)\,\xi_k, c_0 \approx F(x_k - \xi_k) - a_0 - b_0$ und die Wahl $\lambda_k = \lambda_{\min}$. **Armijo–Goldstein Bedingung**: $F(x_{k+1}) = F(x_k - \lambda_k \xi_k) \leq F(x_k) - \alpha\,\lambda_k F'(x_k)\,\xi_k$ bei vorgegebenem Parameter α).

– Unter gewissen Voraussetzungen an die Funktion f, den Startwert x_0 und die Schrittweitenfolge $(\lambda_k)_{k\geq 0}$ kann gezeigt werden, daß das Quasi-Newton-Verfahren **global konvergent** ist, vgl. Skriptum, S. 126.

- **Vorbemerkung**: Um den erheblichen Aufwand bei der Berechnung der Jacobimatrix $f'(x_k)$ für $k \geq 0$ zu verringern, verwendet man beim **vereinfachten Newton-Verfahren** die Iteration (dies erfordert beispielsweise nur eine einzige LR-Zerlegung von $f'(x_0)$ und jeweils eine Vorwärts- und Rückwärtssubstitution pro Iterationsschritt)

$$x_{k+1} = x_k - \xi_k\,, \quad f'(x_0)\,\xi_k = f(x_k)\,, \qquad k \geq 0\,.$$

Allerdings ist dann im Allgemeinen die Konvergenz linear und nicht quadratisch.

Das **Verfahren von Broyden** verwendet folgende Iterationsvorschrift im k-ten Schritt für $k \geq 0$ (üblicherweise mit $J_0 = f'(x_0)$)

$$\begin{cases} \text{Für } J_k \approx f'(x_k) \text{ bestimme } \xi_k \text{ aus } J_k\,\xi_k = f(x_k) \text{ und berechne } x_{k+1} = x_k - \xi_k\,. \\ \text{Bestimme } J_{k+1} \text{ aus } J_{k+1}\,\xi_k = -\big(f(x_{k+1}) - f(x_k)\big)\,. \end{cases}$$

Im eindimensionalen Fall entspricht das Verfahren dem Sekantenverfahren und insbesondere J_{k+1} der Steigung der Sekante durch $(x_k, f(x_k))$ und $(x_{k+1}, f(x_{k+1}))$ (wegen $-\xi_k = x_{k+1} - x_k$, sofern $J_k \neq 0$)

$$J_{k+1} = \tfrac{f(x_{k+1}) - f(x_k)}{x_{k+1} - x_k}\,.$$

Im mehrdimensionalen Fall ist die Matrix J_{k+1} nicht eindeutig festgelegt (beispielsweise sind die Koeffizienten einer Matrix $A \in \mathbb{R}^{2\times 2}$ bei Vorgabe von $x \in \mathbb{R}^2$ und $b \in \mathbb{R}^2$ durch die Bedingung $Ax = b \Leftrightarrow a_{11}x_1 + a_{12}x_2 = b_1, a_{21}x_1 + a_{22}x_2 = b_2$ nicht eindeutig festgelegt). Durch die Zusatzforderung (**Rang-1-Modifikation**)

$$J_{k+1} = J_k + u v^T$$

folgt (Einsetzen in $J_{k+1}\xi_k = f(x_k) - f(x_{k+1})$)

$$\underbrace{v^T \xi_k}_{=-\frac{1}{c}\in\mathbb{R}} u = f(x_k) - f(x_{k+1}) - \underbrace{J_k \xi_k}_{=f(x_k)} = -f(x_{k+1}) \quad \Longleftrightarrow \quad u = c\, f(x_{k+1})\,.$$

Die spezielle Wahl $v = -\xi_k$ führt auf ($-\frac{1}{c} = v^T \xi_k = -\|\xi_k\|_2^2 \Leftrightarrow c = \frac{1}{\|\xi_k\|_2^2}$, $u = \frac{1}{\|\xi_k\|_2^2} f(x_{k+1})$)

$$J_{k+1} = J_k - \tfrac{1}{\|\xi_k\|_2^2} f(x_{k+1})\, \xi_k^T\,,$$

das Verfahren von Broyden.

Bemerkung: Falls für eine Matrix $A \in \mathbb{R}^{n\times n}$ die LR-Zerlegung (oder QR-Zerlegung) bekannt ist, benötigt die Berechnung der LR-Zerlegung (oder QR-Zerlegung) einer Rang-1-Modifikation von A

$$A = LR\,, \quad A + uv^T = (L+\ell)\,(R+r) = LR + \ell R + Lr + \ell\, r\,,$$
$$uv^T = \ell R + Lr + \ell\, r\,,$$

lediglich $\mathcal{O}\big(n^2\big)$ Operationen (und nicht $\mathcal{O}\big(n^3\big)$ Operationen).

Themenüberblick II

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 $x \leftrightarrow x - x_0$, 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 $(\alpha_i)_{0\leq i\leq n}$. 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

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

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	$\mathbf{y_0}$			
		$\delta \mathbf{y_{0,1}} = \frac{y_1 - y_0}{x_1 - x_0}$		
x_1	y_1		$\delta^2 \mathbf{y_{0,1,2}} = \frac{\delta y_{1,2} - \delta y_{0,1}}{x_2 - x_0} = \frac{\frac{y_2 - y_1}{x_2 - x_1} - \frac{y_1 - y_0}{x_1 - x_0}}{x_2 - x_0}$	
		$\delta y_{1,2} = \frac{y_2 - y_1}{x_2 - x_1}$		$\delta^3 \mathbf{y_{0,1,2,3}} = \frac{\delta^2 y_{1,2,3} - \delta^2 y_{0,1,2}}{x_3 - x_0}$
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	$\mathbf{-23}$			
		$\mathbf{12}$		
0	1		$\mathbf{-1}$	
		9		$\mathbf{4}$
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)

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

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

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

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

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

- **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

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

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 $f \in \mathscr{C}^n([a,b])$. Dann existiert ein $\xi \in [a,b]$ mit

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

Denn: Zum Beweis der Relation

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

wird verwendet, daß die Differenz $r = p - f$ (zumindest) $n+1$ Nullstellen $x_0, \ldots, x_n$ 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

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

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

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

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

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

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

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.)

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

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 $j = i+1 \leftrightarrow i = j-1$)

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

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

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

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

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

* 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	$g(0)$		
		$2\big(g\big(\frac{1}{2}\big) - g(0)\big)$	
$\frac{1}{2}$	$g\big(\frac{1}{2}\big)$		$2\big(g(0) - 2\,g\big(\frac{1}{2}\big) + g(1)\big)$
		$2\big(g(1) - g\big(\frac{1}{2}\big)\big)$	
1	$g(1)$		

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

und Integration führt auf die Quadraturapproximation

$$\begin{aligned} Q_0(g) &= \int_0^1 \widetilde{g}(\xi)\,\mathrm{d}\xi \\ &= \int_0^1 \Big(g(0) + 2\big(g\big(\tfrac{1}{2}\big) - g(0)\big)\,\xi + \big(g(0) - 2\,g\big(\tfrac{1}{2}\big) + g(1)\big)\big(2\,\xi^2 - \xi\big)\Big)\,\mathrm{d}\xi \\ &= g\big(\tfrac{1}{2}\big) + \tfrac{1}{6}\big(g(0) - 2\,g\big(\tfrac{1}{2}\big) + g(1)\big) \\ &= \tfrac{1}{6}\big(g(0) + 4\,g\big(\tfrac{1}{2}\big) + g(1)\big), \\ s = 3, \quad b_1 = \tfrac{1}{6}, \quad b_2 = \tfrac{2}{3}, \quad b_3 = \tfrac{1}{6}, \quad c_1 = 0, \quad c_2 = \tfrac{1}{2}, \quad c_3 = 1\,. \end{aligned}$$

* 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

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

eine verbesserte Approximation an den Wert des Integrals.

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

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

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)

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

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

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

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**

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

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)

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

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

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

die in der allgemeineren **impliziten Form** enthalten sind

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

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**

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

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

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

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

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

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

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

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

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

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 $x, \widetilde{x} \in D$ die folgende Abschätzung mit **Lipschitz-Konstante** $L > 0$ gilt

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

  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 $v, \widetilde{v} \in \mathbb{R}^d$ gilt die Relation

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

  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 $y, \tilde{y}$ der Differentialgleichung

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

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

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

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

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

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**

$$\begin{array}{c|c} c = (c_i)_{1\le i\le s} & A = (a_{ij})_{1\le i,j\le s} \\ \hline & b = (b_i)_{1\le i\le s} \end{array}$$

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**

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

die **globale Fehlerentwicklung**

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

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 Stabilitäasabschätzungen 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 $h > 0$)

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

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)

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

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 $\lambda \leq 0$ und wegen $h > 0$ ist die Schranke $|1 - h\lambda|^{-1} \leq 1$ immer gültig!)

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

– **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

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

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 $y''(t) = f'(y(t))\, f(y(t))$)

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

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 $0 < h_i \leq \overline{h}$, $0 \leq i \leq N-1$)

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

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

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

  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

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

  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)

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

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 Funktionen 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)

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

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

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

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

    die **Anfangsauslenkung und Anfangsgeschwindigkeit** vorgegeben

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

    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

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

- **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**

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

  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 $r : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^{d+1}$)

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

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
$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} y(t) = f\big(t, y(t)\big), & t \in (t_0, T), \\ y_1(t_0) = y_{01}, \quad y_1(T) = y_{T1}. \end{cases}$$
Beispiele: Schwingungsgleichung, Wurfparabel
  - **Idee**: Die Idee des (einfachen) **Schießverfahrens** ist es, das Randwertproblem auf ein Anfangswertproblem zurückzuführen
$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} y(t) = f\big(t, y(t)\big), & t \in (t_0, T), \\ y(t_0) = y_0 = (y_{01}, y_{02})^T. \end{cases}$$
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 $y : I \to \mathbb{R}^d$
$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} y(t) = f\big(t, y(t)\big), & t \in (t_0, T), \\ r\big(y(t_0), y(T)\big) = 0, \end{cases}$$
beruht auf der Betrachtung des zugehörigen Anfangswertproblems
$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} y(t) = f\big(t, y(t)\big), & t \in (t_0, T), \\ y(t_0) = y_0, \end{cases}$$
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 $y : I \to \mathbb{R}^d$

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

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 $\langle v_1, \ldots, v_K \rangle$ 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)

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} z(t)\big|_{t=t_n} = f\big(t_n, z(t_n)\big), \quad 1 \le n \le N-1, \\ r\big(z(t_0), z(t_N)\big) = 0, \end{cases}$$

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)

$$\frac{y(x+h)-y(x)}{h} = \frac{\mathrm{d}}{\mathrm{d}x} y(x) + \mathscr{O}(h)\,, \qquad \frac{y(x-h)-y(x)}{-h} = \frac{y(x)-y(x-h)}{h} = \frac{\mathrm{d}}{\mathrm{d}x} y(x) + \mathscr{O}(h)\,,$$
$$\frac{y(x+h)-y(x-h)}{2h} = \frac{\mathrm{d}}{\mathrm{d}x} y(x) + \mathscr{O}(\mathbf{h^2})\,,$$
$$\frac{y(x+h)-2\,y(x)+y(x-h)}{h^2} = \frac{\mathrm{d}^2}{\mathrm{d}x^2} y(x) + \mathscr{O}(\mathbf{h^2})\,.$$

Genauer: Falls die Funktion y hinreichend oft differenzierbar ist, führen Taylorreihenentwicklungen auf folgende Relationen für die symmetrischen Differenzen (jeweils mit $\xi \in [x-h, x+h]$)

$$\begin{aligned} y \in \mathscr{C}^3(a,b): &\qquad \tfrac{y(x+h)-y(x-h)}{2h} = \tfrac{\mathrm{d}}{\mathrm{d}x}\, y(x) + \tfrac{1}{6}\, h^2\, \tfrac{\mathrm{d}^3}{\mathrm{d}x^3}\, y(x)\big|_{x=\xi}\,, \\ y \in \mathscr{C}^4(a,b): &\qquad \tfrac{y(x+h)-2\,y(x)+y(x-h)}{h^2} = \tfrac{\mathrm{d}^2}{\mathrm{d}x^2}\, y(x) + \tfrac{1}{12}\, h^2\, \tfrac{\mathrm{d}^4}{\mathrm{d}x^4}\, y(x)\big|_{x=\xi}\,. \end{aligned}$$

Gitterfunktion (Grid function): Zur Vereinfachung werden im Folgenden äquidistante **Gitterpunkte** zur **Gitterweite** $h > 0$ betrachtet (M innere Punkte $x_1, \ldots, x_M$)

$$\overline{\Omega} = \{x_m = a + mh : 0 \le m \le M+1\}\,, \qquad h = \tfrac{b-a}{M+1}\,.$$

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

$$L : z \longmapsto Lz = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} z + \alpha \frac{\mathrm{d}}{\mathrm{d}x} z + \beta z$$

der Differenzenoperator

$$\widetilde{L} : \widetilde{z} = \left(\widetilde{z}_m\right)_{1\le m\le M} \longmapsto \widetilde{L}\widetilde{z} = \left(-\tfrac{\widetilde{z}_{m+1}-2\widetilde{z}_m+\widetilde{z}_{m-1}}{h^2} + \alpha(x_m)\tfrac{\widetilde{z}_{m+1}-\widetilde{z}_{m-1}}{2h} + \beta(x_m)\widetilde{z}_m\right)_{1\le m\le M}.$$

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

$$\left(\beta(x_m)\widetilde{z}_m\right)_{1\le m\le M} = \begin{pmatrix} \beta(x_1)\widetilde{z}_1 \\ \vdots \\ \beta(x_m)\widetilde{z}_m \\ \vdots \\ \beta(x_M)\widetilde{z}_M \end{pmatrix} = \underbrace{\begin{pmatrix} \beta(x_1) & & & & \\ & \ddots & & & \\ & & \beta(x_m) & & \\ & & & \ddots & \\ & & & & \beta(x_M) \end{pmatrix}}_{=A_0} \begin{pmatrix} \widetilde{z}_1 \\ \vdots \\ \widetilde{z}_m \\ \vdots \\ \widetilde{z}_M \end{pmatrix}$$

sowie

$$\begin{aligned} \left(\alpha(x_m)\tfrac{\widetilde{z}_{m+1}-\widetilde{z}_{m-1}}{2h}\right)_{1\le m\le M} &= \frac{1}{2h}\begin{pmatrix} \alpha(x_1)(\widetilde{z}_2-\widetilde{z}_0) \\ \vdots \\ \alpha(x_m)(\widetilde{z}_{m+1}-\widetilde{z}_{m-1}) \\ \vdots \\ \alpha(x_M)(\widetilde{z}_{M+1}-\widetilde{z}_{M-1}) \end{pmatrix} \\ &= \underbrace{\frac{1}{2h}\begin{pmatrix} 0 & \alpha(x_1) & & & \\ & \ddots & & & \\ & -\alpha(x_m) & 0 & \alpha(x_m) & \\ & & & \ddots & \\ & & & -\alpha(x_M) & 0 \end{pmatrix}}_{=A_1} \begin{pmatrix} \widetilde{z}_1 \\ \vdots \\ \widetilde{z}_m \\ \vdots \\ \widetilde{z}_M \end{pmatrix} + \underbrace{\frac{1}{2h}\begin{pmatrix} -\alpha(x_1)\widetilde{z}_0 \\ 0 \\ \vdots \\ 0 \\ \alpha(x_M)\widetilde{z}_{M+1} \end{pmatrix}}_{=b_1} \end{aligned}$$

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. $\left\|\widetilde{L}\widetilde{z}\right\|_{\infty,\Omega} > 0$)

$$\begin{aligned}
\widetilde{L}\,\widetilde{v} &= \pm\widetilde{L}\,\widetilde{z} - \left\|\widetilde{L}\widetilde{z}\right\|_{\infty,\Omega} \widetilde{L}\widetilde{\zeta} = -\left\|\widetilde{L}\widetilde{z}\right\|_{\infty,\Omega} \left(\widetilde{L}\widetilde{\zeta} \mp \tfrac{1}{\|\widetilde{L}\widetilde{z}\|_{\infty,\Omega}} \widetilde{L}\,\widetilde{z}\right), \\
\left(\widetilde{L}\,\widetilde{v}\right)_m &= -\left\|\widetilde{L}\widetilde{z}\right\|_{\infty,\Omega} \underbrace{\left(\left(\widetilde{L}\widetilde{\zeta}\right)_m \mp \tfrac{1}{\|\widetilde{L}\widetilde{z}\|_{\infty,\Omega}} \left(\widetilde{L}\,\widetilde{z}\right)_m\right)}_{\ge 0} \le 0, \qquad 1 \le m \le M.
\end{aligned}$$

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 $\xi \in [x-h, x+h]$)

$$\begin{aligned} y \in \mathscr{C}^3(a,b): &\qquad \tfrac{y(x+h)-y(x-h)}{2h} - \tfrac{\mathrm{d}}{\mathrm{d}x}\, y(x) = \tfrac{1}{6}\, h^2\, \tfrac{\mathrm{d}^3}{\mathrm{d}x^3}\, y(x)\big|_{x=\xi}\,, \\ y \in \mathscr{C}^4(a,b): &\qquad \tfrac{y(x+h)-2\,y(x)+y(x-h)}{h^2} - \tfrac{\mathrm{d}^2}{\mathrm{d}x^2}\, y(x) = \tfrac{1}{12}\, h^2\, \tfrac{\mathrm{d}^4}{\mathrm{d}x^4}\, y(x)\big|_{x=\xi}\,. \end{aligned}$$

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).