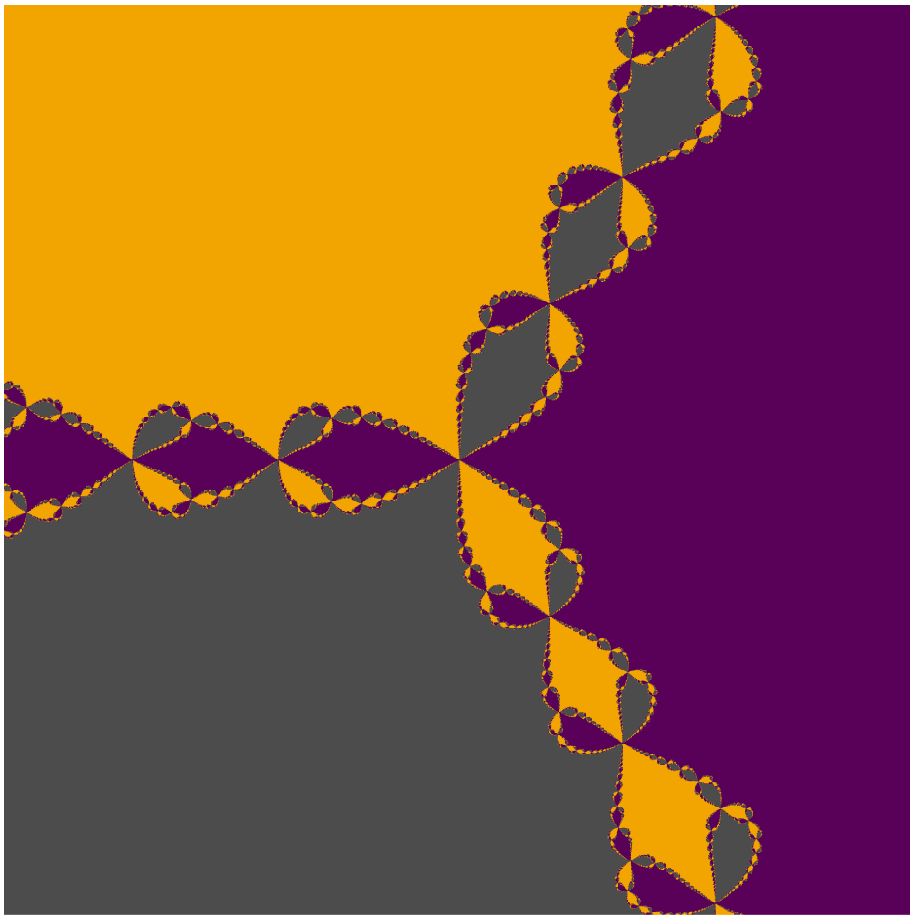
Kompendium zur Lehrveranstaltung

Numerische Mathematik I

Herbsttrimester 2011



Mechthild Thalhammer

Nach einem Skriptum von Mathias Richter
zur Numerischen Mathematik I (Herbsttrimester 2010)

Das vorliegende Kompendium faßt die im Rahmen der dreistündigen Lehrveranstaltung **Numerische Mathematik I** im Herbsttrimester 2011 an der Fakultät für Luft- und Raumfahrttechnik der Universität der Bundeswehr München besprochenen Themen zusammen. Die Inhalte des Kompendiums entsprechen weitgehend einem Skriptum von Mathias Richter zur Vorlesung Numerische Mathematik I, die im Herbsttrimester 2010 abgehalten wurde.

Themenüberblick

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.
  - 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 $\left|\frac{\tilde{g}-g}{g}\right| = \frac{1}{|S|} \leq B^{1-t}$, sofern $g, \tilde{g} \neq 0$. Speziell für $\tilde{g} = 0$ folgt jedoch $\left|\frac{\tilde{g}-g}{g}\right| = 1$. ◇

    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

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

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

- 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 $|\varepsilon_i| \le \varepsilon_{\mathrm{mach}}$.

Beispiel: Vgl. Illustration2_Rundung ($B = 10$, $t = 4$, $\varepsilon_{\mathrm{mach}} = \frac{1}{2} \cdot 10^{1-t}$ 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 $|\varepsilon_1|, |\varepsilon_2| \le \varepsilon_{\mathrm{mach}}$)

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

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 $x_1 x_2 = -b$)

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

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)

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

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

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

$$\xrightarrow[\frac{-a}{-a+\sqrt{a^2+b}}, \frac{\sqrt{a^2+b}}{-a+\sqrt{a^2+b}}]{p^{(4)}} \quad -a + \sqrt{a^2+b}.$$

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

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

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

mit in der vorliegenden Situation gut konditionierten Teilproblemen verwenden.

Vgl. Illustation2_Kondition.

Bemerkung: Eine etwas andere Situation liegt für den Fall $b \approx -a^2$ $(a,b >> 0)$ 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 $\left(p^{(1)},\dots,p^{(k)}\right)$ 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

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

Schlußfolgerung: Die bei der Anwendung der elementaren arithmetischen Operationen auftretenden Rundungsfehler verfälschen das exakte Ergebnis so wie relative Änderungen der Größenordnung $\varepsilon_{\text{mach}}$ 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 $|\varepsilon_i| \leq \varepsilon_{\text{mach}}$ gilt für $0 \leq k \leq n$

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

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

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

    Mittels Bernoullischer Ungleichung ergibt sich weiters

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

    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)

$$
\begin{aligned}
&\widetilde{c}_0 = c_0\,(1+\widetilde{\varepsilon}_0)\,, \qquad \widetilde{c}_n = c_n \prod_{\ell=0}^{n-1} (1+\varepsilon_\ell)\,(1+\widetilde{\varepsilon}_\ell)\,, \\
&\widetilde{c}_i = c_i\,(1+\widetilde{\varepsilon}_i) \prod_{\ell=0}^{i-1} (1+\varepsilon_\ell)(1+\widetilde{\varepsilon}_\ell)\,, \quad 1 \le i \le n-1\,,
\end{aligned}
$$

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

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

Falls das Problem gut konditioniert ist, ist

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

von moderater Größenordnung und damit $r'(z)\,q'(x)\,\xi$ 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. $\|r'(z)\| \gg \|q'(z)\|$) werden unvermeidbare Fehler ζ im Zwischenergebnis (Rundung) sehr verstärkt, d.h. der Beitrag $r'(z)\,\zeta$ führt auf ein unzufriedenstellendes Endergebnis und somit ist der Algorithmus instabil.

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

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

    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 $\left(p^{(1)},\dots,p^{(k)}\right)$ 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

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

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

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 $m = n$)

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

  Matrizenmultiplikation (assoziativ, nicht kommutativ)

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

- 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

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

  Nullraum der linearen Abbildung bzw. der zugehörigen Matrix

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

- 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 $\{v_1, \ldots, v_k\}$ als Linearkombination der restlichen Vektoren darstellen, d.h. es gilt

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

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

Sind n Vektoren $v_1, \ldots, v_n \in \mathbb{K}^n$ 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 $e_1, \ldots, e_n \in \mathbb{K}^n$ (definiert durch $(e_i)_j = \delta_{ij}$ für $1 \leq i, j \leq n$) 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 $A \in \mathbb{K}^{m \times n}$, $x \in \mathbb{K}^n$, $b \in \mathbb{K}^m$)

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

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 $m \geq n$, 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 $Ax = b$ 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 $D^{-1} = \operatorname{diag}(\frac{1}{d_1}, \ldots, \frac{1}{d_n})$.
  - 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

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

  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)

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

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

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

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

$$A = Q\,R,$$

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

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

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

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

mit orthonormalen Vektoren $q_1, \dots, q_n \in \mathbb{K}^m$. 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

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

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 $n = 3$

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

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

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

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

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

- 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 $\|e_1\|_1 = 1 = \|e_2\|_1$)

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

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 $X \subset \mathbb{R} \setminus \{0\}$)

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

    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

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

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

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

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

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

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

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

    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

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

– 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

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

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

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

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

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

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 $y = Tx$

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

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 $\|x-y\|_2^2 = \|x\|_2^2 + |\alpha|^2 + 2\,\Re(\alpha\,\overline{x_1}) = 2\,\|x\|_2^2 + 2\,|x_1|\,\|x\|_2$)

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

- 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

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

  - 2. Schritt: Elimination in der zweiten Spalte

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

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 $\tilde{x}$ ist die exakte Lösung eines linearen Gleichungssystems

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

  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 $A\tilde{x} - b$ 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 $|\varepsilon| \le \varepsilon_{\text{mach}}$)

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

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

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

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. $x_1 = f_1(x_2,\dots,x_n)$) 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 $g_i(x_2,\dots,x_n)$ für $1 \le i \le n-1$ in den $n-1$ Unbekannten $x_2,\dots,x_n$). 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 $x_1, x_2, \dots, x_n$ gewählt werden kann.
  - Lineares Gleichungssystem

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

  - 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

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

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

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

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

– Analoge Elimination der Variablen $x_2, \dots, x_{n-1}$.

  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 $L = N_{21}(\ell_{21}) \cdots N_{n1}(\ell_{n1})\, N_{32}(\ell_{32}) \cdots N_{n2}(\ell_{n2}) \cdots N_{n,n-1}(\ell_{n2})$)

$$A = LR,$$

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

  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

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

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

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

- 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 $1 \leq i \leq n$)

$$A = LR,$$

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

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

  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

$$\begin{gathered} b = (\widetilde{L}+\Delta L)\,\widetilde{y} = (\widetilde{L}+\Delta L)\,(\widetilde{R}+\Delta R)\,\widetilde{x} = \widetilde{A}\widetilde{x} \\ \widetilde{A} - A = (\widetilde{L}+\Delta L)\,(\widetilde{R}+\Delta R) - A = \widetilde{L}\widetilde{R} - A + \Delta L\widetilde{R} + \widetilde{L}\Delta R + \Delta L\Delta R, \\ |\widetilde{A} - A| \le \underbrace{|\widetilde{L}\widetilde{R} - A|}_{\le \mathcal{O}(\varepsilon_{\text{mach}})|\widetilde{L}||\widetilde{R}|} + \underbrace{|\Delta L|}_{\le \mathcal{O}(\varepsilon_{\text{mach}})|\widetilde{L}|}|\widetilde{R}| + |\widetilde{L}|\underbrace{|\Delta R|}_{\le \mathcal{O}(\varepsilon_{\text{mach}})|\widetilde{R}|} + \underbrace{|\Delta L\Delta R|}_{\mathcal{O}(\varepsilon_{\text{mach}}^2)|\widetilde{L}||\widetilde{R}|} \le \mathcal{O}(\varepsilon_{\text{mach}})|\widetilde{L}||\widetilde{R}|, \\ r = b - A\widetilde{x} = (\widetilde{A} - A)\widetilde{x}. \end{gathered}$$

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
$$|A\widetilde{x} - b| \le \varepsilon\big(|A|\,|\widetilde{x}| + |b|\big).$$
- Nach dem Satz von Prager und Oettli ist die berechnete Näherungslösung $\widetilde{x}$ akzeptabel (bzgl. $U_{\varepsilon_{\text{mach}}}$), wenn das Residuum die Abschätzung
$$|r| = |\widetilde{A} - A|\,|\widetilde{x}| \le \varepsilon_{\text{mach}}\big(|A|\,|\widetilde{x}| + |b|\big)$$
erfüllt. Insbesondere ist das Verfahren numerisch stabil, wenn
$$|r| = |\widetilde{A} - A|\,|\widetilde{x}| \le \mathcal{O}(\varepsilon_{\text{mach}})|\widetilde{L}||\widetilde{R}||\widetilde{x}| \le \varepsilon_{\text{mach}}\big(|A|\,|\widetilde{x}| + |b|\big),$$
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

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

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 $0 \le x \le 1$ den konstanten Wert 1 und wächst ansonsten an. Falls $0 \le \frac{bc}{ad} \le 1$ oder $0 \le \left|\frac{bc}{ad}\right| \le 1$, folgt $\alpha \le 3$ und somit

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

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 $|bc| \le |ad|$, 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

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

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

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

5.2. Normalengleichungen

- Situation: Näherungsweise Lösung des linearen Ausgleichsproblems

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

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

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

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 $f'(x)=(\partial_{x_1}f(x),\dots,\partial_{x_n}f(x))\in\mathbb{R}^{1\times n}$)

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

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

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

– 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

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

  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 $R \in \mathbb{K}^{m\times n}$

$$A = Q\,R\,,$$
$$Q = T_1 \cdots T_n \in \mathbb{K}^{m\times m}\,, \quad Q^*Q = I\,, \qquad R = \left(\begin{array}{c} \widehat{R} \\ \hline 0 \end{array}\right) \in \mathbb{K}^{m\times n}\,.$$

  Da die Transformation mittels einer unitären Matrix die euklidische Norm erhält

$$\|A\,x - b\|_2 = \|Q^*A\,x - Q^*b\|_2 = \|R\,x - Q^*b\|_2\,,$$

  gilt für die Lösung des linearen Ausgleichsproblems

$$\|A\,x - b\|_2 \xrightarrow{\;!\;} \min \quad \Longleftrightarrow \quad \|R\,x - Q^*b\|_2 \xrightarrow{\;!\;} \min\,.$$

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

$$R\,x - c = \left(\begin{array}{c} \widehat{R} \\ \hline 0 \end{array}\right) x - \left(\begin{array}{c} \widehat{c} \\ \hline \widetilde{c} \end{array}\right) = \left(\begin{array}{c} \widehat{R}\,x - \widehat{c} \\ \hline -\widetilde{c} \end{array}\right),$$
$$\|R\,x - Q^*b\|_2 = \|\widehat{R}\,x - \widehat{c}\|_2 + \|\widetilde{c}\|_2 \xrightarrow{\;!\;} \min \quad \Longleftrightarrow \quad \|\widehat{R}\,x - \widehat{c}\|_2 \xrightarrow{\;!\;} \min\,.$$

  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 ($\|\widehat{R}\,x - \widehat{c}\|_2 = 0 \Leftrightarrow \widehat{R}\,x = \widehat{c}$)

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

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 $v_1, \ldots, v_n$ 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)

$$AV = V\Lambda \quad\Longleftrightarrow\quad A = V\Lambda V^{-1} \quad\Longleftrightarrow\quad \Lambda = V^{-1}AV.$$

Falls die Matrix $A \in \mathbb{K}^{n\times n}$ unitär diagonalisierbar ist, d.h. eine Orthonormalbasis aus Eigenvektoren existiert, folgt insbesondere (wegen $V^{-1} = V^*$)

$$AV = V\Lambda \quad\Longleftrightarrow\quad A = V\Lambda V^* \quad\Longleftrightarrow\quad \Lambda = V^*AV.$$

– 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

$$\begin{aligned} & Av = \lambda v, \quad 0 \neq v \in \mathcal{N}_{A-\lambda I} = \left\{v \in \mathbb{K}^n : (A - \lambda I)\, v = 0\right\} \\ \underset{v=Tw,\, w=T^{-1}v}{\Longleftrightarrow} \quad & ATw = \lambda Tw, \quad 0 \neq v \in \mathcal{N}_{A-\lambda I}, \quad 0 \neq w = T^{-1}v \\ \Longleftrightarrow \quad & T^{-1}ATw = \lambda w, \quad 0 \neq w \in \mathcal{N}_{T^{-1}AT-\lambda I} = \left\{w \in \mathbb{K}^n : (T^{-1}AT - \lambda I)\, w = 0\right\}. \end{aligned}$$

- 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

$$\left(T^* A\, T - \lambda I\right) v = \begin{pmatrix} 0 & S_{12} & S_{13} \\ & 0 & S_{23} \\ & & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} S_{12}\, v_2 + S_{13}\, v_3 \\ S_{23}\, v_3 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

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. $A^*A = AA^*$) 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 $\mathrm{rg}(A) = k$ mit $1 \leq k \leq n$. Die Matrix $B = A^* A \in \mathbb{K}^{n\times n}$ ist selbstadjungiert und positiv semi-definit

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

  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 $\mathrm{rg}(B) = \mathrm{rg}(A^* A) = \mathrm{rg}(A) = k$ (verwende die Relation $\mathrm{rg}(A) = n - \dim \mathcal{N}_A$ sowie $\mathrm{rg}(A^* A) = n - \dim \mathcal{N}_{A^* A} = n - \dim \mathcal{N}_A = \mathrm{rg}(A)$) 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

$$A = \widehat{U}\widehat{\Sigma}T^*,$$
$$A = \left(a_1 \middle| \cdots \middle| a_n\right) \in \mathbb{K}^{m\times n}, \qquad T \in \mathbb{C}^{n\times n}, \quad T^*T = I,$$
$$\widehat{U} = \left(u_1 \middle| \cdots \middle| u_n\right) \in \mathbb{C}^{m\times n}, \quad \widehat{U}^*\widehat{U} = I, \qquad \widehat{\Sigma} = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix} \in \mathbb{R}^{n\times n}$$

Die (volle) Singulärwertzerlegung von A ergibt sich bei Ergänzung von $\widehat{U}$ zu einer Orthonormalbasis des $\mathbb{C}^m$ und entsprechender Ergänzung von $\widehat{\Sigma}$ 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

$$A = U\Sigma V^*,$$
$$U \in \mathbb{C}^{m\times m}, \quad U^*U = I, \qquad V \in \mathbb{C}^{n\times n}, \quad V^*V = I,$$
$$\Sigma = \operatorname{diag}\left(\sigma_1, \dots, \sigma_{\min\{m,n\}}\right) \in \mathbb{R}^{m\times n}, \qquad \sigma_1 \geq \cdots \geq \sigma_k > 0, \quad \sigma_{k+1} = \cdots = \sigma_{\min\{m,n\}} = 0,$$

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

$$\|A_j - A\|_2 = \inf\{\|A - B\|_2 : B \in \mathbb{C}^{m\times n}, \mathrm{rg}(B) \le j\} = \sigma_{j+1}\,.$$

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 $u_1, \ldots, u_m$)

$$A - A_j = \sum_{i=1}^{k} \sigma_i u_i v_i^* - \sum_{i=1}^{j} \sigma_i u_i v_i^* = \sum_{i=j+1}^{k} \sigma_i u_i v_i^*\,,$$

$$\|A - A_j\|_2 = \max_{\|x\|_2=1} \|(A-A_j)x\|_2 = \max_{\|x\|_2=1} \Big\| \sum_{i=j+1}^{k} \sigma_i v_i^* x\, u_i \Big\| = \max_{\|x\|_2=1} \sqrt{\sum_{i=j+1}^{k} \sigma_i^2 |v_i^* x|^2}\,.$$

Mit Hilfe der Darstellung bezüglich der Basis $v_1, \ldots, v_n \in \mathbb{C}^n$ 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

$$\|A_j - A\|_2 = \inf\{\|A - B\|_2 : B \in \mathbb{C}^{m\times n}, \mathrm{rg}(B) \le j\}\,.$$

Unter der Annahme für $B \in \mathbb{C}^{m\times n}$ gilt

$$\mathrm{rg}(B) \le j \Leftrightarrow \dim \mathcal{N}_B \ge n - j\,, \qquad \|A - B\|_2 < \sigma_{j+1}\,.$$

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

    $$\lambda^{(k+1)} = \Psi\left(\lambda^{(k)}\right), \qquad k \geq 0,$$

    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 $|\lambda^{(k+1)} - \lambda^{(k)}| \leq \text{TOL}$)

    $$|\lambda^{(k)} - \lambda| \leq \text{TOL}.$$

  - 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 $\lambda_1,\dots,\lambda_n \in \mathbb{C}$ 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

$$\begin{gathered} A_1 = A, \qquad A_{k+1} = Q_k^* A_k Q_k, \qquad k = 1,2,3,\dots \\ A_1 = A \quad \longrightarrow \quad A_2 = Q_1^* A Q_1 \quad \longrightarrow \quad A_3 = Q_2^* Q_1^* A Q_1 Q_2 \quad \longrightarrow \quad \dots \\ \lim_{k\to\infty} A_k = S. \end{gathered}$$

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

$$A \quad \longrightarrow \quad H = T^* A T = \begin{pmatrix} H_{11} & \dots & \dots & H_{1n} \\ H_{21} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & H_{n,n-1} & H_{nn} \end{pmatrix}.$$

  * 1. Schritt: Konstruktion einer Householder-Reflexion $\widetilde{Q}_1 \in \mathbb{K}^{n-1\times n-1}$ derart, daß die Spaltenelemente $(a_{21},\dots,a_{n1})^T \in \mathbb{K}^{n-1}$ 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)

$$A \quad \longrightarrow \quad Q_1^* A Q_1 = \begin{pmatrix} * & * & \dots & * \\ * & * & \dots & * \\ & \vdots & & \vdots \\ & * & \dots & * \end{pmatrix}.$$

  * 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

$$A \quad \longrightarrow \quad H = T^* AT = \begin{pmatrix} H_{11} & H_{12} & & \\ H_{21} & \ddots & \ddots & \\ & \ddots & \ddots & H_{n-1,n} \\ & & H_{n,n-1} & H_{nn} \end{pmatrix},$$

$$H_{ii} \in \mathbb{R}, \quad H_{i+1,i} = \overline{H_{i,i+1}} \in \mathbb{R}, \qquad 1 \le i \le i+1 \le n.$$

– 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 $r'(x) = (\partial_{x_1} r(x), \dots, \partial_{x_n} r(x)) \in \mathbb{R}^{1\times n}$)

$$\begin{aligned}
r'(x) &= 2\,\frac{1}{x^T x}\Big((Ax)^T - r(x)\,x^T\Big)\,,\\
\partial_{x_\ell} r(x) &= \frac{\sum\limits_{i,j=1}^{n} (\delta_{i\ell}\,a_{ij}\,x_j + x_i\,a_{ij}\,\delta_{j\ell})}{x^T x} - 2\,r(x)\,\frac{x_\ell}{x^T x}\\
&= \frac{\sum\limits_{j=1}^{n} a_{\ell j}\,x_j + \sum\limits_{i=1}^{n} x_i\,a_{i\ell}}{x^T x} - 2\,r(x)\,\frac{x_\ell}{x^T x}\\
&= 2\,\frac{(Ax)_{\ell 1}}{x^T x} - 2\,r(x)\,\frac{x_\ell}{x^T x}\\
&= 2\,\frac{1}{x^T x}\big((Ax)_{\ell 1} - r(x)\,x_\ell\big)\,.
\end{aligned}$$

  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

$$r(x) = r(v) + \mathcal{O}\big(\|x - v\|_2^2\big)\,.$$

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 $\xi_1 = v_1^T x^{(0)} \neq 0$)

$$x^{(0)} = \sum_{i=1}^n \xi_i v_i\,,$$

$$x^{(k)} = A^k x^{(0)} = \sum_{i=1}^n \xi_i A^k v_i = \sum_{i=1}^n \xi_i \lambda_i^k v_i = \xi_1 \lambda_1^k \Big(v_1 + \sum_{i=2}^n \tfrac{\xi_i}{\xi_1} \big(\tfrac{\lambda_i}{\lambda_1}\big)^k v_i \Big)\,,$$

$$\tfrac{1}{\xi_1 \lambda_1^k} x^{(k)} = v_1 + \sum_{i=2}^n \tfrac{\xi_i}{\xi_1} \big(\tfrac{\lambda_i}{\lambda_1}\big)^k v_i \quad \underset{|\frac{\lambda_i}{\lambda_1}| = \varrho_i < 1}{\overset{k\to\infty}{\longrightarrow}} \quad v_1\,.$$

Für den Rayleigh-Quotienten folgt weiters

$$\big\|x^{(k)}\big\|_2^2 = \sum_{i=1}^n \xi_i^2 \lambda_i^{2k} = \xi_1^2 \lambda_1^{2k} + \sum_{i=2}^n \xi_i^2 \lambda_i^{2k} = \xi_1^2 \lambda_1^{2k} \Big(1 + \sum_{i=2}^n \tfrac{\xi_i^2}{\xi_1^2} \big(\tfrac{\lambda_i}{\lambda_1}\big)^{2k} \Big)\,,$$

$$\big(x^{(k)}\big)^T A x^{(k)} = \Big(\sum_{i=1}^n \xi_j \lambda_j^k v_j \Big)^T \sum_{i=1}^n \xi_i \lambda_i^{k+1} v_i = \sum_{i,j=1}^n \xi_i \xi_j \lambda_i^{k+1} \lambda_j^k \underbrace{v_j^T v_i}_{=\delta_{ij}} = \sum_{i=1}^n \xi_i^2 \lambda_i^{2k+1}\,,$$

$$r\big(x^{(k)}\big) = \tfrac{1}{\|x^{(k)}\|_2^2} \big(x^{(k)}\big)^T A x^{(k)} = \tfrac{1}{\|x^{(k)}\|_2^2} \sum_{i=1}^n \xi_i^2 \lambda_i^{2k+1} = \frac{\xi_1^2 \lambda_1^{2k+1} \Big(1 + \sum_{i=2}^n \tfrac{\xi_i^2}{\xi_1^2} \big(\tfrac{\lambda_i}{\lambda_1}\big)^{2k+1} \Big)}{\xi_1^2 \lambda_1^{2k} \Big(1 + \sum_{i=2}^n \tfrac{\xi_i^2}{\xi_1^2} \big(\tfrac{\lambda_i}{\lambda_1}\big)^{2k} \Big)}$$

$$= \lambda_1 \frac{1 + \sum_{i=2}^n \tfrac{\xi_i^2}{\xi_1^2} \big(\tfrac{\lambda_i}{\lambda_1}\big)^{2k+1}}{1 + \sum_{i=2}^n \tfrac{\xi_i^2}{\xi_1^2} \big(\tfrac{\lambda_i}{\lambda_1}\big)^{2k}} \quad \overset{k\to\infty}{\longrightarrow} \quad \lambda_1\,.$$

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

$$r(cx) = \frac{(cx)^T A\,(cx)}{(cx)^T (cx)} = \frac{c x^T A\, x}{x^T x} = r(x)\,, \qquad c \neq 0\,.$$

  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

$$|\varrho_2| = \left|\tfrac{\lambda_2}{\lambda_1}\right|$$

  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 $|\varrho_2| = \left|\tfrac{\lambda_2}{\lambda_1}\right| \approx 1$.

  Die inverse Vektoriteration basiert auf der Umformulierung der Eigenrelation (wobei $\mu \in \mathbb{R}$ mit $\mu \neq \lambda_i$ für $1 \leq i \leq n$)

$$\begin{aligned} A v_i = \lambda_i v_i\,, \quad 1\leq i\leq n \quad &\Longleftrightarrow \quad (A-\mu I)\, v_i = (\lambda_i - \mu)\, v_i\,, \quad 1\leq i\leq n \\ &\Longleftrightarrow \quad (A-\mu I)^{-1} v_i = \tfrac{1}{\lambda_i-\mu}\, v_i\,, \quad 1\leq i\leq n \\ &\Longleftrightarrow \quad \widetilde{A}_\mu v_i = \widetilde{\lambda}\, v_i\,, \quad B = (A-\mu I)^{-1}\,, \quad \widetilde{\lambda} = \tfrac{1}{\lambda_i-\mu}\,, \qquad 1\leq i\leq n\,. \end{aligned}$$

  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

$$\lim_{k\to\infty} A^{(k)} = \begin{pmatrix} \lambda_1 & * & \cdots & \cdots & * \\ & \ddots & \ddots & & \vdots \\ & & \ddots & \ddots & * \\ & & & \ddots & * \\ & & & & \lambda_n \end{pmatrix}.$$

  Denn: Es reicht aus, die Konvergenz der ersten Spalte und der letzten Zeile von $A^{(k)}$ nachzuweisen

$$\lim_{k\to\infty} A^{(k)} = \begin{pmatrix} \lambda_1 & * & \cdots & \cdots & * \\ & \vdots & & & \vdots \\ & \vdots & & & \vdots \\ & * & \cdots & \cdots & * \\ & & & & \lambda_n \end{pmatrix},$$

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 $R^{(k)}$

$$\begin{gathered} A X^{(k)} = X^{(k+1)} R^{(k)}, \\ A X^{(k)}_{-,1} = X^{(k+1)} R^{(k)}_{-,1} = R^{(k)}_{11} X^{(k+1)} e_1 = R^{(k)}_{11} X^{(k+1)}_{-,1}, \end{gathered}$$

  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 $\left(T^{-1}\right)^* = T$)

$$\begin{gathered} A X^{(k)} = X^{(k+1)} R^{(k)}, \\ \left(X^{(k)}\right)^* A^* = \left(A X^{(k)}\right)^* = \left(X^{(k+1)} R^{(k)}\right)^* = \left(R^{(k)}\right)^* \left(X^{(k+1)}\right)^*, \\ \left(A^{-1}\right)^* X^{(k)} = X^{(k+1)} \left(\left(R^{(k)}\right)^*\right)^{-1}. \end{gathered}$$

  Speziell für die letzte Spalte der entstehenden Matrix ergibt sich aufgrund der unteren Dreiecksform von $\left(R^{(k)}\right)^*$

$$\begin{gathered} \left(A^{-1}\right)^* X^{(k)}_{-,n} = X^{(k+1)} \left(\left(R^{(k)}\right)^*\right)^{-1}_{-,n} = \left(\left(R^{(k)}\right)^*\right)^{-1}_{n,n} X^{(k+1)} e_n = \tfrac{1}{R^{(k)}_{nn}} X^{(k+1)}_{-,n}, \\ z^{(k+1)} = d^{(k)} \left(A^{-1}\right)^* z^{(k)}, \qquad d^{(k)} = R^{(k)}_{nn} > 0, \quad z^{(k)} = X^{(k)}_{-,n}, \end{gathered}$$

  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)

$$A^{(k)} = \left(X^{(k)}\right)^* A X^{(k)},$$

$$A^{(k)}_{-,1} = A^{(k)} e_1 = \left(X^{(k)}\right)^* A X^{(k)} e_1 = \left(X^{(k)}\right)^* \underbrace{A X^{(k)}_{-,1}}_{\to \pm \lambda_1 v_1} \quad \overset{k\to\infty}{\longrightarrow} \quad \lambda_1 e_1,$$

$$A^{(k)}_{n,-} = e_n^T A^{(k)} = \left(A^* X^{(k)} e_n^T\right)^* X^{(k)} = \left(\underbrace{A^* X^{(k)}_{-,n}}_{\to \lambda_n z}\right)^* X^{(k)} \quad \overset{k\to\infty}{\longrightarrow} \quad \lambda_n e_n^T.$$

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 $x, \widetilde{x} \in D$ die folgende Relation gilt (bezüglich einer festgelegten Norm $\|\cdot\|$)

$$\|g(x) - g(\widetilde{x})\| \le \kappa \|x - \widetilde{x}\|, \qquad 0 < \kappa < 1,$$

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

$$\varphi(x) = x - \big(f'(x)\big)^{-1} f(x)\,.$$

  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, $f(\bar{x}) = 0$)

$$\begin{aligned}
\varphi(x) &= x - \frac{f(x)}{f'(x)}\,, \\
\varphi'(x) &= 1 - \frac{f'(x)}{f'(x)} + \frac{f(x)\,f''(x)}{(f'(x))^2} = \frac{f(x)\,f''(x)}{(f'(x))^2}\,, \qquad \varphi'(\bar{x}) = 0\,, \\
\varphi''(x) &= \frac{f''(x)}{f'(x)} + \frac{f(x)\,f'''(x)}{(f'(x))^2} - 2\,\frac{f(x)\,(f''(x))^2}{(f'(x))^3}\,, \qquad \varphi''(\bar{x}) = \frac{f''(\bar{x})}{f'(\bar{x})} \underset{\text{i.A.}}{\neq} 0\,,
\end{aligned}$$

  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 $d_k = y_k - \bar{x}$)

$$\begin{aligned}
d_{k+1} &= \underbrace{y_{k+1}}_{=\varphi(y_k)+\delta_k} - \underbrace{\bar{x}}_{=\varphi(\bar{x})} = \varphi(y_k) - \varphi(\bar{x}) + \delta_k = \varphi'(\zeta_k)\,(y_k - \bar{x}) + \delta_k\\
&= \varphi'(\zeta_k)\,(y_k - y_{k+1} + d_{k+1}) + \delta_k\,, \qquad k \ge 0\,,\\
\Longrightarrow\quad d_{k+1} &= \tfrac{1}{1-\varphi'(\zeta_k)}\big(\varphi'(\zeta_k)\,(y_k - y_{k+1}) + \delta_k\big)\,, \qquad k \ge 0\,,
\end{aligned}$$

und weiters (Abschwächung der Voraussetzung $|\varphi'(\zeta_k)| \le c < 1$ für $k \ge 0$ ausreichend, geometrische Reihe)

$$|d_{k+1}| \le \frac{1}{|1-\varphi'(\zeta_k)|}\left(|\varphi'(\zeta_k)|\,|y_k - y_{k+1}| + |\delta_k|\right) \le \frac{c}{1-c}\,|y_k - y_{k+1}| + \frac{1}{1-c}\,|\delta_k|\,, \qquad k \ge 0\,.$$

Dies zeigt insbesondere, daß es bei einer Fixpunktiteration zu keiner Akkumulation von Rundungsfehlern kommt (Abhängigkeit vom aktuellen Iterationsschritt)

$$|y_{k+1} - \bar{x}| \le \frac{c}{1-c}\,|y_k - y_{k+1}| + \frac{1}{1-c}\,|\delta_k|\,, \qquad k \ge 0\,.$$

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 $c = \frac{1}{2}$
    $$|x_{k+1} - \bar{x}| \le \tfrac{1}{2}\,|x_k - \bar{x}|\,.$$
  - 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).

$$|x_{k+1} - \bar{x}| \leq \kappa \, |x_k - \bar{x}| \, .$$

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

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}\, f(x_k) \, .$$

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)

$$x_{k+1} - \bar{x} = \frac{f''(\xi_k)}{2f'(\zeta_k)} (x_k - \bar{x}) (x_{k-1} - \bar{x}), \qquad k \geq 0.$$

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 $c > 0$)

$$\delta_k = \|x_k - \bar{x}\|, \quad \eta_k = \ln(c\delta_k), \quad \delta_k = \frac{1}{c} e^{\eta_k}, \qquad k \geq 0.$$

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 $F'(x)\in\mathbb{R}^{n\times 1}$)

$$F(x)=\tfrac{1}{2}\,\|f(x)\|_2^2=\tfrac{1}{2}\,\big(f(x)\big)^T f(x)=\tfrac{1}{2}\sum_{i=1}^{n}\big(f_i(x)\big)^2\,,$$

$$\partial_{x_\ell}F(x)=\sum_{i=1}^{n}f_i(x)\,\underbrace{\partial_{x_\ell}f_i(x)}_{=(f'(x))_{i\ell}}=\sum_{i=1}^{n}(f'(x))^T_{\ell i}\,f_i(x)=\Big(\big(f'(x)\big)^T f(x)\Big)_{\ell 1}\,,$$

$$F'(x)=\big(f(x)\big)^T f'(x)\,,$$

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 $F(x_k - \lambda \xi_k) = p(\lambda) = a_0 + b_0 \lambda + c_0 \lambda^2$ mit Kenntnis der Funktionswerte bei $\lambda = 0, 1$ sowie der Ableitung bei $\lambda = 0$ führt auf die berechenbaren Koeffizienten $a_0 = F(x_k), b_0 = F'(x_k)\xi_k, c_0 = F(x_k + \xi_k) - a_0 - b_0$ 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)

$$J_{k+1} = J_k + uv^T$$

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 $\mathscr{O}\big(n^2\big)$ Operationen (und nicht $\mathscr{O}\big(n^3\big)$ Operationen).