Numerische Mathematik I

zusammengefaßt von A.A.M.[1]

Das vorliegende Skript ist eine Mitschrift der gleichnamigen Vorlesung von
Univ.-Prof. Dr. H. M. Möller im Wintersemester 1997/98

[1]Rückfragen unter marczok@gmx.de

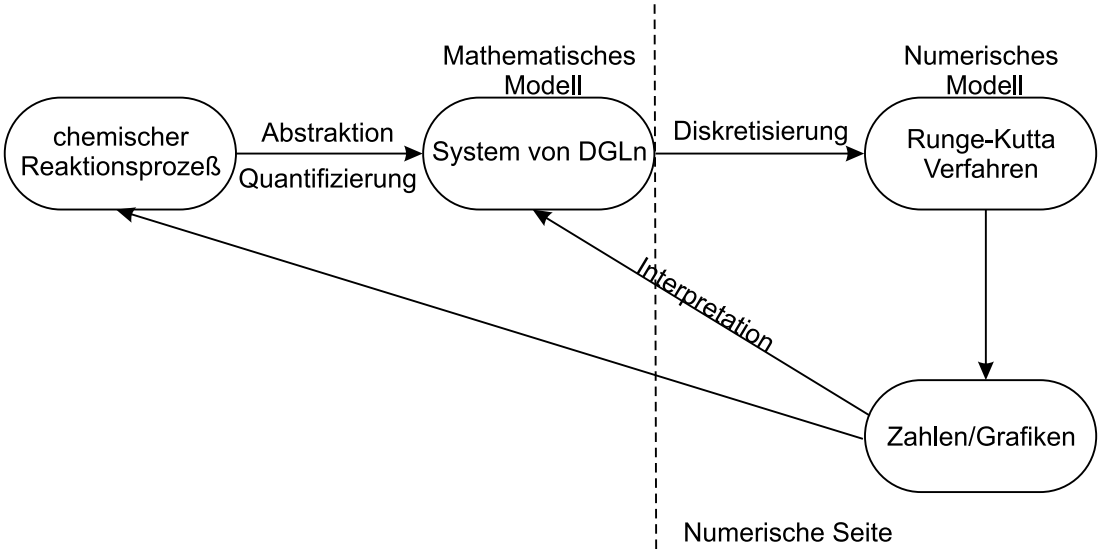
Einige Worte zum Script

Diese Zusammenfassung wurde mit MiKTEX 1.10 – 2.1 unter  erstellt.

INHALTSVERZEICHNIS

Überblick

Beispiel aus der Numerik:



Literatur:

Björck-Dahlquist	*Numerische Methoden 1972 (praxisorientiert)*
Deuflhard-Holmann	*Numerische Mathematik I/II 1991 (anspruchsvoll)*
Reiner	*Grundlagen der Numerischen Mathematik I/II 1980/82 (klassisch)*
Stiefel	*Einführung in die Numerische Mathematik 1970 (handlich, einfach, alt)*
Stoer-Bulirsch	*Einführung in die Numerische Mathematik I/II 1986 (DAS! Numerikbuch)*
Gautschi	*Numerical Analysis 1997 (ausländisch)*

Praktische Mathematik:

- Bleistift und Papier
- Rechenautomaten
- Tabellen

- Rechenschieber

Numerische Mathematik:

- Taschenrechner
- PC

Scientific Computing: (wissenschaftliches Rechnen)

- Rechnen mit riesigen Datenmengen
- Großrechenanlagen

In der Vorlesung wird die Numerische Mathematik behandelt. Als Hilfsmittel werden der Taschenrechner, sowie ein PC eingesetzt, als Software wird MAPLE empfohlen. Eine kleine Rolle spielt immernoch PASCAL.

Programm von Numerik I:

1. Fehleranalyse
2. Iterative Lösung von nichtlinearen Gleichungen
3. Polynome
4. Direkte Lösung von linearen Gleichungssystemen
5. Iterative Lösung von linearen Gleichungssystemen
5. Eigenwertprobleme

In der Vorlesung Numerik II werden die Lösungen von DGLn, Interpolation und Approximation, sowie numerische Integration behandelt.

KAPITEL 1

Fehleranalyse

1.1 Fehler

Fehlertypen:

- Modellfehler:
  - Datenfehler
  - Idealisierungsfehler
- Numerische Fehler
  - Disktretisierungsfehler
  - Abbruchfehler (z.B. bei Reihen)
  - Rundungsfehler

Definition 1.1.1. Sei x eine exakte Größe in einem normierten Raum. Sei $\tilde{x}$ eine Näherung an x. Dann heißen:

$$\varepsilon := x - \tilde{x} \qquad \text{absoluter Fehler}$$

$$\delta := \frac{x - \tilde{x}}{|x|} \qquad \text{relativer Fehler, falls } x \neq 0$$

1.2 Fehlerfortpflanzung und Stabilität

Satz 1.2.1 (Fehlerfortpflanzung bei arithmetischen Operationen). *Die Operanden x_1, x_2 seien näherungsweise bekannt:*

$$x_i = \tilde{x}_i - \varepsilon_i \quad , \quad i = 1, 2$$

δ sei der relative Fehler. η bzw. ξ sei der absolute bzw. relative Fehler des Resultates bei einer arithmetischen Grundoperation. Dann gilt, falls Nenner $\neq 0$:
Addition:

$$\eta = \varepsilon_1 + \varepsilon_2$$

$$\xi = \frac{x_1}{x_1 + x_2}\delta_1 + \frac{x_2}{x_1 + x_2}\delta_2 \qquad \textit{(ungünstig bei Auslöschung)}$$

Multiplikation:

$$\eta \doteq \varepsilon_1 x_1 + \varepsilon_2 x_2 \qquad \textit{„}\doteq\textit{“ bedeutet „in erster Näherung“}$$

$$\xi \doteq \delta_1 + \delta_2$$

Division:

$$\eta \doteq \frac{1}{x_1}\varepsilon_1 - \frac{x_1}{x_2}\varepsilon_2$$
$$\xi \doteq \delta_1 - \delta_2$$

Bei arithmetischen Grundoperationen in $\mathbb{R}$ *oder* ...

$$\delta = \frac{x - \tilde{x}}{x} \qquad \textit{falls } x \neq 0$$

BEWEIS. Addition:

$$z = f(x_1, x_2) = x_1 + x_2$$
$$\tilde{z} = f(\tilde{x}_1, \tilde{x}_2) = \tilde{x}_1 + \tilde{x}_2$$
$$\delta_i = \frac{\varepsilon_i}{x_i}$$

$$\eta = z - \tilde{z} = \varepsilon_1 + \varepsilon_2$$
$$\xi = \frac{\eta}{z} = \frac{\varepsilon_1 + \varepsilon_2}{x_1 + x_2} = \frac{1}{x_1 + x_2}\varepsilon_1 + \frac{1}{x_1 + x_2}\varepsilon_2 = \frac{x_1}{x_1 + x_2}\delta_1 + \frac{x_2}{x_1 + x_2}\delta_2$$

Multiplikation: analog

Division: (schwieriger)

$$z = f(x_1, x_2) = \frac{x_1}{x_2}$$
$$\tilde{z} = \frac{\tilde{x}_1}{\tilde{x}_2}$$

Taylor:

$$\tilde{z} = z + f_{x_1}(x_1, x_2)(-\varepsilon_1) + f_{x_2}(x_1, x_2)(-\varepsilon_2) + \ldots$$
$$\doteq \frac{x_1}{x_2} - \frac{1}{x_1}\varepsilon_2 + \frac{x_1}{x_2^2}\varepsilon_2$$

$$\frac{\eta}{z} = \frac{\tilde{z} - z}{z} \doteq \frac{1}{x_1}\frac{\varepsilon_1}{z} - \frac{x_1}{x_2}\frac{\varepsilon_2}{z} = \frac{1}{x_1}\varepsilon_1 - \frac{1}{x_2}\varepsilon_2 = \delta_1 - \delta_2$$

□

Definition 1.2.1. Ein mathematischer Prozeß heißt *gut konditioniert*, wenn kleine Änderungen der Daten $x_1, \ldots, x_n$ nur kleine Änderungen der (exakten) Lösung bewirken. Sonst heißt der Prozeß *schlecht konditioniert*.

$f : \Omega \to \mathbb{R}$ sei auf $\Omega \subset \mathbb{R}^n$ definiert; Ω sei offen und konvex. f sei auf Ω stetig differenzierbar. Der Datenvektor sei $x = (x_1 \ \cdots \ x_n)^\top \in \Omega$; der Fehlervektor $\varepsilon = (\varepsilon_1 \ \cdots \ \varepsilon_n)^\top$; $\Omega \ni \tilde{x} := x + \varepsilon$. Dann (Taylor):

$$\eta = f(\tilde{x}) - f(x) \doteq \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(x) \cdot \varepsilon_j \qquad \text{absoluter Fehler}$$

$$\xi = \frac{\eta}{f(x)} = \sum_{j=1}^{n} \frac{x_j}{f(x)} \frac{\partial f}{\partial x_j}(x) \cdot \delta_j \qquad \text{relativer Fehler}$$

Definition 1.2.2. Sei $f : \Omega \to \mathbb{R}$ wie oben, dann heißen:

$$\sigma_i = \frac{\partial f}{\partial x_i}(x) \qquad \text{bzw.} \qquad \tau_i = \frac{x_i}{f(x)} \frac{\partial f}{\partial x_i}(x)$$

Konditionszahlen (in bezug auf die i-te Komponente) bezüglich des absoluten bzw. relativen Fehlers.

Ein Prozeß heißt gut konditioniert, wenn die Beträge der Konditionszahlen *klein gegen Eins* sind. Andernfalls schlecht konditioniert.

$\Rightarrow$ „natürliche“ (In)Stabilität.

Gegensatz dazu ist die „numerische“ (In)Stabilität bedingt durch den Rechnungsverlauf.

Beispiel 1.2.1.

$$f(x) = \ln(x - \sqrt{x^2-1}), \quad x = 30$$

exakt: $f(30) = -4,0940\,6666\,8632\ldots$

natürliche Satbilität:

$$f'(x) = \frac{1}{x - \sqrt{x^2-1}} \cdot \left(1 - \frac{x}{\sqrt{x^2-1}}\right) = \frac{-1}{\sqrt{x^2-1}}$$

$$f'(30) = \frac{-1}{\sqrt{899}} = -0,0333\,5\ldots \qquad \text{(absoluter Fehler)}$$

$$\frac{30}{f(30)} \cdot f'(30) = 0,2443\ldots \qquad \text{(relativer Fehler)}$$

numerische Stabilität:

$$x \quad \to \quad \sqrt{x^2-1} \quad \to \quad x - \sqrt{x^2-1} \quad \to \quad \ln(x - \sqrt{x^2-1})$$

Bei vierstelliger Rechnung:

$$\sqrt{30^2-1} = 29,98(33\,2870\ldots)$$
$$30 - \sqrt{899} \approx 0,02$$
$$\ln(0,02) \approx -3,910$$

bei zehnstelliger Rechnung:

$$30 - \sqrt{899} \approx 0,0166\,7130$$
$$\ln(0,0166\,7130) = \underline{-4,0940\,6660}\,1$$

Stabil:

$$\ln(x - \sqrt{x^2-1}) = \ln\left(\frac{1}{x + \sqrt{x^2-1}}\right) = -\ln(x + \sqrt{x^2-1}) \quad \text{(keine Auslöschung)}$$

$$(x + \sqrt{x^2-1})\Big|_{x=30} \approx 59,98$$
$$-\ln(59,98) \approx -4,094$$

Genauere Analyse der numerischen (In)Stabilität:

Zerlegung von f in eine Kette von Elementaralgorithmen:

$$f = \varphi_n \circ \cdots \circ \varphi_2 \circ \varphi_1$$

Ein Fehler, der bei der Berechnung von φ_i auftritt, geht in die Restabbildung

$$\varphi_n \circ \cdots \circ \varphi_{i+1}$$

wie ein Eingabefehler (Datenfehler) ein und wird durch diese Abbildung an das Endresultat weitergegeben.

Definition 1.2.3. Sei $f = \varphi_k \circ \varphi_{k-1} \circ \cdots \circ \varphi_1$ die Zerlegung der Abbildung f in Elementaralgorithmen. Ein Algorithmus $\varphi_k \circ \cdots \circ \varphi_1$ zur Berechnung von f heißt *numerisch stabil*, wenn die Teilalgorithmen

$$h_i = \varphi_n \circ \cdots \circ \varphi_{i+1}, \quad i = 1, \ldots, k-1$$

natürlich stabil sind.

1.3 Rundungsfehler bei Gleitkommaarithmetik

Eine reelle Zahl $x \in \mathbb{R}$ besitzt eine Darstellung als unendlicher Dezimalbruch:

$$\sigma\, 10^e \sum_{i=1}^{\infty} x_i\, 10^{-i}, \quad \sigma \in \{1,-1\},\ e \in \mathbb{Z},\ x_i \in \{0,1,\ldots,9\},\ x_1 \neq 0$$

Allgemeiner ist die g-adische Darstellung (beim Rechner ist $g \in \mathbb{Z}$, gerade):

$$\sigma\, g^e \sum_{i=1}^{\infty} \alpha_i\, g^{-i}, \quad \sigma \in \{1,-1\},\ e \in \mathbb{Z},\ \alpha_i \in \{0,1,\ldots,g-1\},\ \alpha_1 \neq 0$$

Definition 1.3.1. Sei $g \in \mathbb{N}$ gerade, $t \in \mathbb{N}$, $x \in \mathbb{R} \setminus \{0\}$ mit $x = \sigma\, g^e \sum_{i=1}^{\infty} \alpha_i\, g^{-i}$,
$\sigma \in \{1,-1\}$, $e \in \mathbb{Z}$, $\alpha_i \in \{0,1,\ldots,g-1\}$, $\alpha_1 \neq 0$. Dann:

$$\mathrm{rd}_t(x) := \begin{cases} \sigma\, g^e \sum_{i=1}^{t} \alpha_i\, g^{-i} & \text{, falls } \alpha_{t+1} < \frac{g}{2} \\ \sigma\, g^e \left(\sum_{i=1}^{t} \alpha_i\, g^{-i} + \frac{1}{g^t} \right) & \text{, falls } \alpha_{t+1} \geq \frac{g}{2} \end{cases}$$

$\mathrm{rd}_t(x)$ heißt der auf t Stellen gerundete Wert von x.

Satz 1.3.1. *Sei $g \in \mathbb{N}$ gerade, $t \in \mathbb{N}$, $x \neq 0$ mit g-adischer Darstellung wie eben. Dann gilt:*

i) $\mathrm{rd}_t(x) = \sigma\, g^{e'} \sum_{i=1}^{t} \alpha_i'\, g^{-i}$

ii) $|\mathrm{rd}_t(x) - x| \leq \frac{1}{2} g^{e-t}$

iii) $\left| \frac{\mathrm{rd}_t(x) - x}{x} \right| \leq \frac{1}{2} g^{-t+1}$

iv) $\left| \frac{\mathrm{rd}_t(x) - x}{\mathrm{rd}_t(x)} \right| \leq \frac{1}{2} g^{-t+1}$

Bemerkung. Daß sich nach *i)* alle Ziffern ändern können, zeigt $t = 3,\ g = 10$:

$$x = 0,9996 \quad \rightarrow \quad \mathrm{rd}_t(x) = 1,000 = 10^1 \cdot (1 \cdot 10^{-1} + 0 \cdot 10^{-2} + 0 \cdot 10^{-3} + 0 \cdot 10^{-4})$$

BEWEIS. zu *i)*: Nur für $\alpha_{t+1} \geq \frac{g}{2}$

Fall 1:

$$\alpha_1 = \cdots = \alpha_t = g - 1$$
$$\mathrm{rd}_t(x) = \sigma g^{e+1}(1 \cdot g^{-1} + 0 \cdot g^{-2} + \cdots + 0 \cdot g^{-t})$$

Fall 2: Für ein α_i gilt

$$\alpha_i < g - 1 = \alpha_{i+1} = \cdots = \alpha_t$$
$$\mathrm{rd}_t(x) = \sigma g^e(\alpha_1 \cdot g^{-1} + \cdots + \alpha_{i-1} \cdot g^{-i+1} + (\alpha_i + 1) \cdot g^{-i} + 0 \cdot g^{-i-1} + \cdots + 0 \cdot g^{-t})$$

zu *ii)* :

Fall 1:

$$\alpha_{t+1} < \frac{g}{2}$$

$$\begin{aligned} 0 < -\sigma(\mathrm{rd}_t(x) - x) &= g^e \sum_{i=t+1}^{\infty} \alpha_i\, g^{-i} \\ &= g^e \cdot \alpha_{t+1} g^{-t-1} + g^e \sum_{i>t+1}^{\infty} \alpha_i\, g^{-i} \\ &\leq g^e \cdot \left(\frac{g}{2} - 1 \right) g^{-t-1} + g^e (g-1) \sum_{i>t+1}^{\infty} \alpha_i\, g^{-i} \\ &= \frac{1}{2} g^{e-t} \end{aligned}$$

Fall 2:

$$\begin{aligned} \alpha_{t+1} &\geq \frac{g}{2} \\ 0 < \sigma(\mathrm{rd}_t(x) - x) &= g^e g^{-t} - g^e \sum_{i=t+1}^{\infty} \alpha_i \, g^{-i} \\ &= g^{e-t} - g^e \cdot \alpha_{t+1} g^{-t-1} - g^e \sum_{i>t+1}^{\infty} \alpha_i \, g^{-i} \\ &\leq g^{e-t} - g^e \cdot \frac{g}{2} g^{-t-1} \\ &= \frac{1}{2} g^{e-t} \end{aligned}$$

zu *iii*): Wegen $\alpha_i \neq 0$ ist $\alpha_1 \geq 1$. Daher $|x| \geq g^e g^{-1} = g^{e-1}$. Mit *i*) also

$$\left| \frac{\mathrm{rd}_t(x) - x}{x} \right| \leq \frac{1}{2} g^{e-t} \cdot g^{1-e} = \frac{1}{2} g^{1-t}$$

zu *iv*): Rundungsvorschrift gibt:

$$\mathrm{rd}_t(x) \geq \alpha_i g^{e-1} \geq g^{e^1}$$

weiter wie in *iii*). □

Korollar 1.3.1. *Es gelten die Darstellungen*

$$\mathrm{rd}_t(x) = x(1+\varepsilon) = \frac{x}{1+\xi} \qquad \textit{mit} \quad \varepsilon, \xi \in \mathbb{R}: \quad |\varepsilon| \leq \frac{1}{2} g^{-t+1}, \quad |\xi| \leq \frac{1}{2} g^{-t+1}$$

Beweis. Für $x = 0$ setzt man $\varepsilon = \xi = 0$. Sonst

$$\varepsilon = \frac{\mathrm{rd}_t(x) - x}{x} \qquad \text{nach Satz 1.3.1 } \textit{iii})$$

$$|\varepsilon| \leq \frac{1}{2} g^{-t+1}$$

$$\xi = \frac{x - \mathrm{rd}_t(x)}{\mathrm{rd}_t(x)} \qquad \text{nach Satz 1.3.1 } \textit{iv}) \text{ also}$$

$$|\xi| \leq \frac{1}{2} g^{-t+1}$$

□

Die Zahl $\frac{1}{2} g^{-t+1}$ heißt relative Rundungsgenauigkeit det t-stelligen Gleitkommaarithmetik (zur Basis g) oder „Maschinenzahl“

$$\mathrm{eps} := \frac{1}{2} g^{-t+1}$$

Definition 1.3.2. Eine g-adische und t-stellige Gleitkommazahl hat die Form $x = 0$ oder

$$x = \sigma \, g^e \sum_{i=1}^{t} \alpha_i \, g^{-i}, \quad \sigma \in \{1, -1\}, \; e \in \mathbb{Z}, \; \alpha_i \in \{0, 1, \ldots, g-1\}, \; \alpha_1 \neq 0$$

σ heißt Vorzeichen
e heißt Exponent
g heißt Basis
$\sum_{i=1}^{t} \alpha_i \, g^{-i}$ heißt Mantisse von x.

Maschinenzahlen sind g-adische t-stellige Gleitkommazahlen mit beschränkten Exponenten, etwa $-99 \le e \le 99$. Im folgenden wird die Diskussion zu Über- und Unterlauf vernachlässigt.
Die arithmetischen Grundoperationen zwischen Maschinenzahlen

$$\oplus \qquad \ominus \qquad \otimes \qquad \oslash$$

führen i.a. aus dem Bereich der Maschinenzahlen hinaus. Auf den Rechnern werden $\oplus \quad \ominus \quad \otimes \quad \oslash$ i.a. so realisiert, daß

$$x \oplus y = \mathrm{rd}_t(x+y) \quad , \quad x \otimes y = \mathrm{rd}_t(x \cdot y) \quad , \quad x \oslash y = \mathrm{rd}_t\left(\frac{x}{y}\right)$$

Es folgt:

$$\begin{aligned} x \oplus y &= (x+y)(1+\varepsilon) = x(1+\varepsilon) + y(1+\varepsilon) \\ x \otimes y &= x(1+\varepsilon)y \\ x \oslash y &= x(1+\varepsilon)/y \\ |\varepsilon| &\le \mathrm{eps} = \frac{1}{2} g^{1-t} \end{aligned}$$

Wilkinsons backward analysis (Rückwärtsanalyse):

Idee:

Arbeitet man mit Daten, die durch Eingangsquellen verfälscht sind, dann sind Rundungsfehler irrelevant, wenn man das numerisch erhaltene Resultat als exaktes Resultat von verfälschten Daten ansehen kann, deren Störung größenordnungsmäßig unterhalb der Eingangsfehler liegt. In diesem Fall spricht man von einem *„gutartigen" Algorithmus*.

Beispiel 1.3.1. Berechne $S(x_1, \ldots, x_n) := x_1 + x_2 + \cdots + x_n$. Verfahren:

$$S_1 := x_1 \quad , \quad S_k := x_k + S_{k-1}, \qquad k = 2, \ldots, n$$

PASCAL:

```
S:=x[1];   FOR I=2 TO N DO S=S+x[k];
```

In Gleitkommaarithmetik:

$$\begin{aligned} \tilde{S}_1 &= x_1 \\ \tilde{S}_k &= x_k \oplus \tilde{S}_{k-1} \\ &= (x_k + \tilde{S}_{k-1})(1+\varepsilon_k) \\ &= x_k(1+\varepsilon_k) + \tilde{S}_{k-1}(1+\varepsilon_k) \\ &= (S_k - S_{k-1})(1+\varepsilon_k) + \tilde{S}_{k-1}(1+\varepsilon_k) \\ \Rightarrow \tilde{S}_k - S_k &= \varepsilon_k S_k + (1+\varepsilon_k)(\tilde{S}_{k-1} - S_{k-1}) \\ &= \varepsilon_k S_k + (1+\varepsilon_k)[\varepsilon_{k-1} S_{k-1} + (1+\varepsilon_{k-1})(\tilde{S}_{k-2} - S_{k-2})] \\ &= \varepsilon_k S_k + (1+\varepsilon_k)\varepsilon_{k-1} S_{k-1} + (1+\varepsilon_k)(1+\varepsilon_{k-1})(\ldots) \end{aligned}$$

Vollständige Induktion:

$$\begin{aligned} \Rightarrow \tilde{S}_n - S_n &= \varepsilon_n S_n + (1+\varepsilon_n)\varepsilon_{n-1} S_{n-1} + (1+\varepsilon_n)(1+\varepsilon_{n-1})\varepsilon_{n-2} S_{n-2} \\ &\quad + \cdots + (1+\varepsilon_n)(1+\varepsilon_{n-1}) \cdots (1+\varepsilon_3)\varepsilon_2 S_2 \\ &\doteq \varepsilon_n S_n + \varepsilon_{n-1} S_{n-1} + \cdots + \varepsilon_2 S_2 \end{aligned}$$

Folge:

Der Fehler hängt von der Summationsreihenfolge ab! Am besten die Reihenfolge so wählen, daß die S_k nicht zu groß werden! Etwa ($n = 3,\ t = 4,\ g = 10$):

$$x_1 = 0,1234, \quad x_2 = 1997, \quad x_3 = -1998$$

S_2 ist klein, wenn $x_1 = 1997,\ x_2 = -1998$:

$$\begin{aligned} \tilde{S}_2 &= -1,000 \\ \tilde{S}_3 &= -1,000 \oplus 0,1234 \\ &= -0,8766 \qquad \text{gut!} \end{aligned}$$

Andere Reihenfolge:

$$\begin{aligned} \tilde{S}_2 &= 0,1234 \oplus 1997 = 1997 \\ \tilde{S}_3 &= -1998 \oplus 1997 = -1,000 \qquad \text{schlecht!} \end{aligned}$$

$\Rightarrow$ Das Assoziativgesetz der Addition gilt beim numerischen Rechnen nicht!

Am Rande (nicht ernstzunehmen):
Der Fundamentalsatz der Algebra gilt in der Numerik auch nicht, da es zu einem Polynom n-ten Grades immer mehr als n Nullstellen gibt (weil verschiedene Näherungswerte HA! HA!).

KAPITEL 2

Iterative Lösungen nichtlinearer Gleichungen

2.1 Fixpunktiteration

Eine Grundaufgabe der Numerik:

Finde ein (alle) $x \in \Omega$ mit:

$$f(x) = 0, \qquad f : \Omega \to \mathbb{R}$$

x ist Nullstelle von f, x ist Lösung von $f(x) = 0$.

Näherungsweise Lösung:

zu gegebenem $\varepsilon > 0$ finde ein $\tilde{x} \in \Omega$, so daß

$$\exists\, x \in \Omega : \quad f(x) = 0,\ |x - \tilde{x}| < \varepsilon$$

Damit ist die Aufgabe numerisch gelöst!.

Unter Umständen ist eine Umformung zu einer Fixpunktgleichung sinnvoll:

Finde $x \in \Omega$: $\varphi(x) = x$ z.B. :

$$\varphi(x) = x + \lambda f(x), \quad \lambda \neq 0$$

Algorithmus (Fixpunktiteration):

Gegeben: $\varphi : \Omega \to \mathbb{R}, \quad \varepsilon > 0,\ x_0 \in \Omega$

Gesucht: $\tilde{x} \in \Omega$, so daß $|\tilde{x} - x^*| < \varepsilon,\ \varphi(x^*) = x^*$

Start: $k := 0$ ~~$x_0 \in \Omega$ beliebig~~

$(\star)$ $x_{k+1} = \varphi(x_k)$

Ist das Abbruchkriterium erfüllt?

Nein$\longrightarrow$ $k := k + 1$, weiter bei $(\star)$
Ja $\longrightarrow$ fertig!

Für ein Abbruchkriterium braucht man noch Informationen, wie man z.B. aus bekannten $x_1, \ldots, x_k$ (und $\varphi(x_1), \ldots, \varphi(x_k)$) feststellt, ob $|x_k - x^*| < \varepsilon$ gilt.

Fixpunktiteration:

$$x_0 \text{ beliebig}, \quad x_{k+1} := \varphi(x_k),\ k = 0, 1, 2, \ldots$$

Beispiel. $\varphi(x) = \sqrt{x}, \quad \Omega = [0, \infty)$. Fixpunkte bei 1 und 0.

$x_0 = 2$	$x_0 = 0,5$
$x_1 = 1,414$	$x_1 = 0,7071$
$x_2 = 1,189$	$x_2 = 0,8409$
$x_3 = 1,091$	$x_3 = 0,9170$
$x_4 = 1,044$	$x_4 = 0,9576$
$x_5 = 1,021$	$x_5 = 0,9786$
$x_6 = 1,011$	$x_6 = 0,9892$

Die Startpunkte in der Abbildung sind willkürlich gewählt!

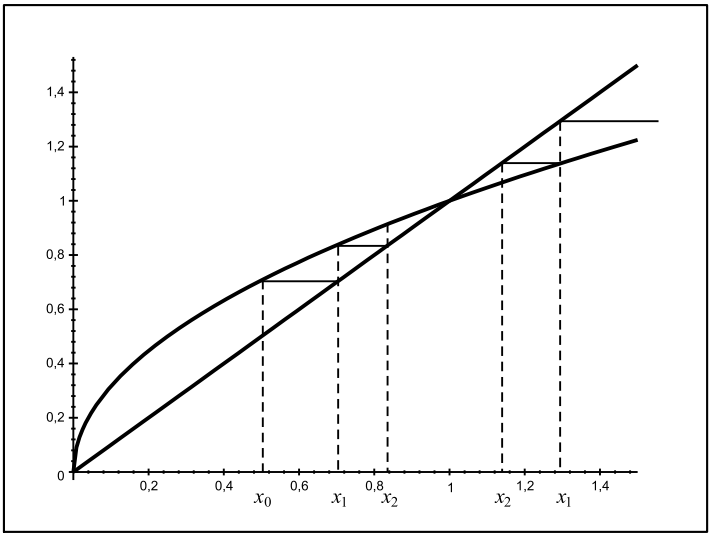


Abbildung 2.1: Iteration der Wurzelfunktion

1 ist anziehender Fixpunkt, 0 ist abstoßender Fixpunkt

2.2 Metrische Räume

Definition 2.2.1. Sei E ein Vektorraum über dem Grundkörper $\mathbb{K}$ (in der Numerik ist $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$). Eine Abbildung $\| \ \| : E \to \mathbb{R}$ heißt *Norm*, wenn sie folgende Eigenschaften erfüllt:

i)	$\|x\| > 0$	für $0 \neq x \in E$
ii)	$\|\alpha x\| = \|\alpha\| \, \|x\|$	für $x \in E, \ \alpha \in \mathbb{K}$
iii)	$\|x + y\| \leq \|x\| + \|y\|$	für $x, y \in E$

Beispiel 2.2.1. $E = \mathbb{R}^n, \ x = (x_1 \cdots x_n)^\top$.

$\|x\|_1 := \sum_{i=1}^{n} \|x_i\|$	$\mathcal{L}_1$-Norm
$\|x\|_2 := \sqrt[2]{\sum_{i=1}^{n} \|x_i\|^2}$	$\mathcal{L}_2$-Norm, Euklidische Norm
$\|x\|_p := \sqrt[p]{\sum_{i=1}^{n} \|x_i\|^p}$	$\mathcal{L}_p$-Norm
$\|x\|_\infty := \max_{i=1}^{n} \|x_i\|$	$\mathcal{L}_\infty$-Norm, Maximum-Norm Tschebyscheff-Norm

Bemerkung. Im $\mathbb{R}^2$ sind die „Einheitskugeln“:

$$S_p := \{(x,y) \in \mathbb{R}^2 \mid \left\|\binom{x}{y}\right\|_p \leq 1\}$$

$$S_1 := \{(x,y) \in \mathbb{R}^2 \mid |x| + |y| \leq 1\}$$

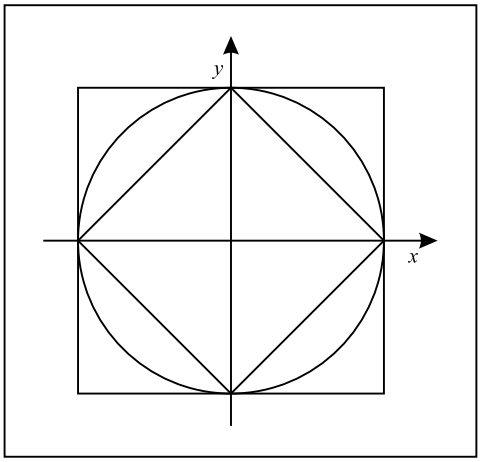


Abbildung 2.2: Verschiedene Normen

$$S_2 := \{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}$$

$$S_\infty := \{(x,y) \in \mathbb{R}^2 \mid \begin{smallmatrix}|x| \leq 1\\ |y| \leq 1\end{smallmatrix}\}$$

Satz 2.2.1. *Auf dem Vektorraum $\mathbb{K}^n$ sei eine Norm $\|\quad\|$ gegeben. Dann sind $\|\quad\|$ und $\|\quad\|_2$ äquivalent, d.h. es gibt Konstanten $\alpha, \beta \in \mathbb{R}: \ 0 < \alpha < \beta < \infty$, so daß für alle $x \in \mathbb{K}^n$:*

$$\alpha \|x\|_2 \leq \|x\| \leq \beta \|x\|_2$$

Beweis. Seien $e_1, \ldots, e_n$ die Einheitsvektoren im $\mathbb{K}^n$. Für beliebige $x = (x_1, \ldots, x_n)^\top, y = (y_1, \ldots, y_n)^\top$ gilt dann:

$$\begin{aligned}
| \|x\| - \|y\| | \leq \|x - y\| &= \left\| \sum_{i=1}^n (x_i - y_i) e_i \right\| \leq \sum_{i=1}^n |x_i - y_i| \, \|e_i\| \\
&\leq \max_{j=0}^n \left\|e_j\right\| \sum_{i=1}^n |x_i - y_i| \\
&\text{Cauchy-Schwarz:} \quad (\textstyle\sum a_k b_k) \leq \sqrt{\sum a_k^2} \sqrt{\sum b_k^2} \text{ mit } a_k = 1, \ b_k = |x_k - y_k| \\
&\leq \max_{j=1}^n \left\|e_j\right\| \cdot \sqrt{n} \, \|x - y\|_2
\end{aligned}$$

$\Rightarrow$ Die Abbildung $x \mapsto \|x\|$ ist stetig! Auf der beschränkten und abgeschlossenen Menge

$$S := \{x \in \mathbb{K} \mid \|x\|_2 = 1\}$$

nimmt $x \mapsto \|x\|$ sein Maximum und Minimun an.

$$\|x\|_2 = 1 \quad \Rightarrow \quad B \leq \|x\| \leq A$$

$y = 0$ ist trivial.

$$\begin{aligned}
0 \neq y \in \mathbb{K}^n \quad & \frac{1}{\|y\|_2} \cdot y \in S \\
& \Rightarrow \quad B \leq \frac{1}{\|y\|_2} \cdot \|y\| \leq A \\
& \Rightarrow \quad B \|y\|_2 \leq \|y\| \leq A \|y\|_2
\end{aligned}$$

□

Korollar. *Alle Normen in $\mathbb{K}^n$ sind untereinander äquivalent.*

BEWEIS. $\|\quad\|, |||\quad|||$ Normen im $\mathbb{K}^n$.

$$\begin{array}{ccccccccc} & & & & \alpha\,\|x\|_2 & \leq & \|x\| & \leq & \beta\,\|x\|_2 \\ & & & & \alpha'\,\|x\|_2 & \leq & |||x||| & \leq & \beta'\,\|x\|_2 \\ \alpha' \cdot \dfrac{1}{\beta}\,\|x\| & \leq & \alpha'\,\|x\|_2 & \leq & |||x||| & \leq & \beta'\,\|x\|_2 & \leq & \beta' \cdot \dfrac{1}{\alpha}\,\|x\| \end{array}$$

□

Abstand $\|x - \tilde{x}\|$

Definition 2.2.2. Sei E nicht leer. E heißt (zusammen mit einer Abbildung $\mathrm{d} : E \times E \to \mathbb{R}$) *metrisch* und d *Metrik* oder *Distanzfunktion*, wenn gilt:

i) $\mathrm{d}(x, y) = 0 \quad \Leftrightarrow \quad x = y$

ii) $\mathrm{d}(x, y) \leq \mathrm{d}(x, z) + \mathrm{d}(y, z)$

Bemerkung. Mit $z = x$ folgt aus *ii)*:

$$\left.\begin{array}{rcl} \mathrm{d}(x, y) & \leq & \mathrm{d}(y, x) \\ \mathrm{d}(y, x) & \leq & \mathrm{d}(x, y) \end{array}\right\} \mathrm{d}(x, y) = \mathrm{d}(y, x)$$

Mit $y = x$ folgt aus *ii)*:

$$0 \leq 2 \cdot \mathrm{d}(x, z) \quad \Rightarrow \quad \mathrm{d}(x, z) \geq 0$$

Definition 2.2.3. Eine Folge $\{x_n\} \subset E$ heißt *Cauchy-Folge*, falls zu jedem $\varepsilon > 0$ ein $N = N(\varepsilon)$ existiert, so daß $\mathrm{d}(x_\mu, x_\nu) < \varepsilon$ für alle $\mu, \nu \geq N$. $\{x_n\} \subset E$ heißt konvergent gegen $x^* \in E$, wenn $\mathrm{d}(x_\nu, x^*) < \varepsilon$ für alle $\nu \geq N$.

Definition 2.2.4. Ein metrischer Raum heißt *vollständig*, wenn jede Cauchy-Folge in E gegen ein $x^* \in E$ konvergiert.

Definition 2.2.5. Ist ein normierter Raum E vollständig bezüglich der Metrik $\mathrm{d}(x, y) = \|x - y\|$, dann ist E mit seiner Norm ein *Banachraum*.

2.3 Der Banachsche Fixpunktsatz (BFS)

Definition 2.3.1. Eine Abbildung φ eines metrischen Raumes E in sich, $\varphi : E \to E$, heißt *lipschitzbeschränkt* mit der Lipschitzkonstanten $L \geq 0$, falls für alle $x, y \in E$ gilt:

$$\mathrm{d}(\varphi(x), \varphi(y)) \leq L\,\mathrm{d}(x, y)$$

φ heißt *Kontraktion* oder *kontrahierend*, falls φ lipschitzbeschränkt ist, mit der Lipschitzkonstanten $L < 1$.

Bemerkung (1). Es reicht nicht $\mathrm{d}(\varphi(x), \varphi(y)) \leq \mathrm{d}(x, y)$

Bemerkung (2). Kontrahierende Abbildungen sind stetig.

Satz 2.3.1 (Banachscher Fixpunktsatz). *Ist $\varphi : E \to E$ eine* kontrahierende *Abbildung eines* vollständigen metrischen *Raumes* in sich*, dann besitzt φ genau einen Fixpunkt $x^* = \varphi(x^*)$. Die Iteration $x_{k+1} = \varphi(x_k)$, $k \geq 0$ konvergiert für jeden Startwert $x_0 \in E$ gegen den Fixpunkt. Bezeichnet L die Kontraktionszahl von φ, dann gelten für $k \geq 1$ die Fehlerabschätzungen:*

i) $\mathrm{d}(x_k, x^*) \leq L \cdot \mathrm{d}(x_{k-1}, x^*)$

ii) $\mathrm{d}(x_k, x^*) \leq \dfrac{L^k}{1 - L} \cdot \mathrm{d}(x_0, x_1)$ *a priori-Abschätzung*

iii) $\mathrm{d}(x_k, x^*) \leq \dfrac{L}{1 - L} \cdot \mathrm{d}(x_k, x_{k-1})$ *a posteriori-Abschätzung*

BEWEIS. Der Beweis erfolgt im mehreren Schritten:

a) $\{x_k\}$ ist eine Cauchy-Folge

b) Der Grenzwert ist ein Fixpunkt

c) Es gibt nur einen Fixpunkt

d) Es gelten die Fehlerabschätzungen

$$d(x_{k+1}, x_{k+2}) = d(\varphi(x_k), \varphi(x_{k+1})) \leq L\, d(x_k, x_{k+1})$$

Mit vollständiger Induktion:

$$\boxed{d(x_{k+s}, x_{k+s+1}) \leq L^s \cdot d(x_k, x_{k+1})}$$

Zu a):

$$\begin{aligned} d(x_k, x_{k+m}) &\leq d(x_k, x_{k+1}) + d(x_{k+1}, x_{k+2}) + \cdots + d(x_{k+m-1}, x_{k+m}) \\ &\leq d(x_k, x_{k+1}) \cdot (1 + L + L^2 + \cdots + L^{m-1}) = \frac{1 - L^m}{1 - L}\, d(x_k, x_{k+1}) \\ &\leq \frac{1}{1 - L}\, d(x_k, x_{k+1}) \\ &\leq \frac{L^k}{1 - L}\, d(x_0, x_1) \end{aligned}$$

Für $\varepsilon > 0$ wähle $N = N(\varepsilon)$ so, daß $\frac{L^N}{1-L}\, d(x_0, x_1) < \varepsilon$ ist. Dann ist $d(x_k, x_{k+m}) < \varepsilon$ für $k, k + m \geq N$.

Zu b):

Die Cauchy-Folge konvergiert gegen ein $x^* \in E$, weil E vollständig ist. φ ist stetig.

$$\varphi(x^*) = \varphi(\lim_{k\to\infty} x_k) = \lim_{k\to\infty} \varphi(x_k) = \lim_{k\to\infty} x_{k+1} = x^*$$

$\Rightarrow x^*$ ist Fixpunkt.

Zu c):

Seien x^* und $\tilde{x}$ Fixpunkte von φ.

$$\begin{aligned} d(x^*, \tilde{x}) &= d(\varphi(x^*), \varphi(\tilde{x})) \\ &\leq L \cdot d(x^*, \tilde{x}) \\ \Rightarrow \quad d(\tilde{x}, x^*) &= 0, \quad \text{also } \tilde{x} = x^* \end{aligned}$$

Zu d):

$$\begin{aligned} &i) \quad d(x_k, x^*) = d(\varphi(x_{k-1}), \varphi(x^*)) \leq L \cdot d(x_{k-1}, x^*) \\ &ii) \quad d(x_k, x_{k+m}) \leq \frac{L^k}{1 - L}\, d(x_0, x_1) \quad \Rightarrow \quad d(x_k, x^*) \leq \frac{L^k}{1 - L}\, d(x_0, x_1) \\ &iii) \quad d(x_k, x_{k+m}) \leq \frac{1}{1 - L}\, d(x_k, x_{k+1}) \leq \frac{L}{1 - L}\, d(x_{k-1}, x_k) \end{aligned}$$

□

Bemerkung. Die Fixpunktfolge kann man auch schreiben als

$$\{x_0, \varphi(x_0), \varphi^2(x_0), \varphi^3(x_0), \dots\}$$

Die Teilfolge $\{x_1, x_2, x_3, \dots\}$ der Fixpunkte ist $\{x_1, \varphi(x_1), \varphi^2(x_1), \dots\}$. Analog ist $\{x_k, x_{k+1}, \dots\}$ die Folge $\{x_k, \varphi(x_k), \varphi^2(x_k), \dots\}$.

Satz 2.3.2. *Sei $I \subseteq \mathbb{R}$ ein Intervall und $\varphi : I \to \mathbb{R}$ stetig differenzierbar. φ ist kontrahierend, wenn für alle $x \in I$ gilt:*

$$\sup_{x \in I} \left|\varphi'(x)\right| \leq L < 1$$

BEWEIS. $\mathrm{d}(x,y) = |x-y|$

$$\begin{aligned}\mathrm{d}(\varphi(x),\varphi(y)) &= |\varphi(x)-\varphi(y)| \\ &\overset{\text{MWS}}{=} \left|\varphi'(\xi)\right| \cdot |x-y|\,, \quad \xi \text{ zwischen } x \text{ und } y\end{aligned}$$

Wenn $|\varphi'(\xi)| \le L < 1$ ist für alle $\xi \in I$, dann ist alles gezeigt. □

Beispiel. $\varphi(x) = 1 + x - \frac{1}{2}x^2, \quad x \in [1,2]$

1) $[1,2]$ ist vollständig (abgeschlossen) und metrisch ($\mathrm{d}(x,y) = |x-y|$).

2) $\varphi'(x) = 1 - x \le 0$ für $x \ge 1 \Rightarrow$ monoton fallend.
$\varphi[1,2] \to [\varphi(2),\varphi(1)] = [1, \frac{3}{2}] \subset [1,2]$

3) $\sup\limits_{x\in I} |\varphi'(x)| = 1 \Rightarrow$ nicht kontrahierend!

Besser: $[1, \frac{3}{2}] \to [\varphi(\frac{3}{2}),\varphi(1)] = [\frac{11}{8}, \frac{3}{2}] \subset [1, \frac{3}{2}]$

$\sup\limits_{x\in I} |\varphi'(x)| = \frac{1}{2} = L < 1$

Jede Folge $\{x_n\}$, $x_0 \in [1, \frac{3}{2}]$, $x_{k+1} = \varphi(x_k)$ konvergiert nach Satz (2.3.1) (BFS) gegen den Fixpunkt $x^* = 1 + x^* - \frac{1}{2}(x^*)^2 \;\Rightarrow\; x^* = \sqrt{2}$

$$|x_k - x^*| \le \frac{L}{1-L}\,|x_k - x_{k+1}| = |x_k - x_{k+1}|$$

In der Praxis ist φ manchmal nicht auf ganz E definiert oder nur in Teilbereichen von E kontrahierend. Daher eine lokale Variante von Satz (2.3.1) (BFS):

$D \subset E$, $D \neq \emptyset$ und Kugel in D

$$K_r(\xi) = \{x \in E \mid \mathrm{d}(x,\xi) \le r\} \subseteq D$$

Mit E ist auch $K_r(\xi)$ ein vollständiger metrischer Raum. Daher hat ein φ, das eine Selbstabbildung auf $K_r(\xi)$ ist und in D kontrahierend ist, einen Fixpunkt, der mit dem Satz (2.3.1) (BFS) berechnet werden kann.

Wann gilt $\varphi : K_r(\xi) \to K_r(\xi)$?
$x \in K_r(\xi)$

$$\begin{aligned}\mathrm{d}(\varphi(x),\xi) &\le \mathrm{d}(\varphi(x),\varphi(\xi)) + \mathrm{d}(\varphi(\xi),\xi) \\ &\le L \cdot \mathrm{d}(x,\xi) + \mathrm{d}(\varphi(\xi),\xi) \\ &\le L \cdot r + \mathrm{d}(\varphi(\xi),\xi)\end{aligned}$$

Wenn man weiß, daß

$$\boxed{\mathrm{d}(\varphi(\xi),\xi) \le r - Lr}$$

erfüllt ist, dann gilt $\mathrm{d}(\varphi(x),\xi) \le r$, d.h. $\varphi(x) \in K_r(\xi)$.

Satz 2.3.3 (lokale Variante des BFS). *φ bilde die Teilmenge D des vollständigen metrischen Raumes E in E ab. Es mögen $\xi \in D$, $r > 0$ und $0 \le L < 1$ existieren mit:*

i) $K_r(\xi) = \{x \in E \mid \mathrm{d}(x,\xi) \le r\} \subseteq D$

ii) *Für $u, v \in K_r(\xi)$:* $\mathrm{d}(\varphi(u),\varphi(v)) \le L \cdot \mathrm{d}(u,v)$

iii) *Es gilt die Kugelbedingung:* $\mathrm{d}(\varphi(\xi),\xi) \le (1-L)\cdot r$

Dann hat φ einen Fixpunkt in $K_r(\xi)$.

Jede Folge $x_{k+1} = \varphi(x_k)$ mit $x_0 \in K_r(\xi)$ konvergiert gegen den Fixpunkt und es gelten die Fehlerabschätzungen.

Definition 2.3.2. Es sei $\varphi : E \to E$ eine Iterationsfunktion in einem metrischen Raum E, x^* sei ein Fixpunkt von φ. Es gebe eine Umgebung $U(x^*)$ von x^*, so daß für alle Startwerte $x_0 \in U(x^*)$ die Iterationsfolge $x_{n+1} = \varphi(x_n)$, $n = 0, 1, \ldots$ gegen x^* konvergiert. Dann heißt das durch φ erzeugte Iterationsverfahren *lokal konvergent*. Ist $U(x^*) = E$ möglich, dann ist das Verfahren *global konvergent*.

2.4 Konvergenzordnung

Definition 2.4.1. Ein durch die Iterationsfunktion $\varphi : E \to E$ mit Fixpunkt $x^* \in E$ gegebenes Iterationsverfahren $x_{n+1} = \varphi(x_n)$ hat mindestens die Konvergenzordnung $s \in \mathbb{R}$. Es ist $s \geq 1$, falls für den absoluten Fehler $0 \neq e_k = \mathrm{d}(x_k, x^*)$ gilt:

$$\limsup_{k\to\infty} \frac{e_{k+1}}{e_k^s} = c \quad \text{mit} \quad \begin{cases} |c| < 1 & \text{für } s = 1 \\ |c| < \infty & \text{für } s > 1 \end{cases}$$

Der Wert c heißt der *asymptotische Fehlerkoeffizient* (kurz *Konvergenzfaktor*). Ein Iterationsverfahren hat genau die Ordnung s, wenn $c \neq 0$ gilt. Die Konvergenz ist:

superlinear, wenn $s = 1,\ c = 0$

linear, wenn $s = 1,\ 0 \neq |c| < 1$

quadratisch, wenn $s = 2,\ c \neq 0$

Satz 2.4.1. *$I \subset \mathbb{R}$ sei ein Intervall, $\varphi : I \to I$ sei s-mal stetig differenzierbar. Es gelte $\varphi(x^*) = x^*$ und*

$$\varphi'(x^*) = \varphi''(x^*) = \cdots = \varphi^{(s-1)}(x^*) = 0 \text{ falls } s > 1$$

bzw. $|\varphi'(x^*)| < 1$, falls $s = 1$. Dann hat das Iterationsverfahren $x_{n+1} = \varphi(x_n),\ x_0 \in I$ geeignet lokal mindestens die Konvergenzordnung s. Gilt zusätzlich $\varphi^{(s)}(x^*) \neq 0$, dann ist s die genaue Konvergenzordnung.

Beweis. Nach Taylor gilt:

$$x_{n+1} = \varphi(x_n) = x^* + 0 + \cdots + 0 + \frac{1}{s!}\varphi^{(s)}(\xi_n) \cdot (x_n - x^*)^s$$

Daher:

$$\frac{x_{n+1} - x^*}{(x_n - x^*)^s} = \frac{1}{s!}\varphi^{(s)}(\xi_n) \text{ mit } \xi_n \text{ zwischen } x^* \text{ und } x_n.$$

$$\begin{aligned} \limsup_{n\to\infty} \frac{\mathrm{d}(x_{n+1}, x^*)}{(\mathrm{d}(x_n, x^*))^s} &= \limsup_{n\to\infty} \frac{1}{s!} \left|\varphi^{(s)}(\xi_n)\right| \\ &= \frac{1}{s!} \left|\varphi^{(s)}(x^*)\right| \end{aligned}$$

□

Bemerkung. Ist $\varphi : I \to I$ stetig differenzierbar mit $|\varphi'(x)| < 1$, dann gibt der MWS, daß die Folge $x_{n+1} = \varphi(x_n)$ monoton konvergiert, falls $\varphi'(x) > 0$ für alle $x \in I$ und und alternierend konvergiert, falls $\varphi'(x) < 0$ für alle $x \in I$.

Beispiel.

$I = [0, \frac{\pi}{2}]$ $\quad \varphi : x \mapsto \cos x$
$\varphi'(x) = -\sin x < 0$
Konvergenzfaktor: $0,6736 = \varphi(x^*)$

$I = [0, \infty)$ $\quad \varphi : x \mapsto \sqrt{x}$
$\varphi'(x) = \frac{1}{2\sqrt{x}}$
Konvergenzfaktor: $\frac{1}{2}$ vgl. Abbildung 2.1 auf Seite 16

2.5 Das Verfahren von Newton und das Sekantenverfahren

Sei x^* eine Nullstelle von $f : \mathbb{R} \to \mathbb{R}$. In einer Umgebung $I_\varepsilon := (x^* - \varepsilon, x^* + \varepsilon)$ sei f zweimal stetig differenzierbar. Es gelte $f'(x^*) \neq 0$. Ist $x_0 \in I_\varepsilon$ eine Näherung an $x^* = x_0 + h$, dann gilt nach Taylor:

$$0 = f(x^*) = f(x_0) + f'(x_0) \cdot h + \underbrace{\frac{1}{2} f''(\xi) h^2}_{\text{wird vernachlässigt}}$$

ξ zwischen x^* und x_0.
Berechne h aus

$$0 = f(x^*) = f(x_0) + f'(x_0) \cdot h \qquad (\Rightarrow h = -\frac{f(x_0)}{f'(x_0)})$$

und eine bessere Näherung:

$$x_1 = x_0 + h = x_0 - \frac{f(x_0)}{f'(x_0)}$$

$\Rightarrow$ Fixpunktiteration:

$$x_{n+1} = \varphi(x_n) \text{ und } \varphi(x) = x - \frac{f(x)}{f'(x)}$$

Tangentensteigung:

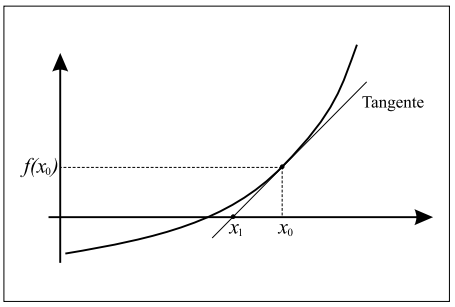


Abbildung 2.3: Geometrische Interpretation des Newton-Verfahrens

$$f'(x_0) = \frac{f(x_0)}{x_0 - x_1} \quad \Rightarrow \quad x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

Algorithmus 2.5.1 (Newtonverfahren).

Gegeben:	$x_0 \in I,\ f \in C^2(I),\ \varepsilon > 0$
Gesucht:	$x_k \in I :\ \lvert x_k - x^* \rvert < \varepsilon,\ f(x^*) = 0$
Iteration:	$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$ *für* $k = 0, 1, 2, \ldots$
	bis das Abbruchkriterium erfüllt ist.

Satz 2.5.1. *Sei x^* eine Nullstelle von f. In einer Umgebung $I_\varepsilon := (x^* - \varepsilon, x^* + \varepsilon),\ \varepsilon > 0$ sei f zweimal stetig differenzierbar und es gelte $f'(x^*) \neq 0$. Dann konvergiert das Newtonverfahren lokal bei x^*, d.h. $\exists I_\delta \subseteq I_\varepsilon$, so daß für jedes $x_0 \in I_\delta$ die Folge $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$ gegen x^* konvergiert.*

BEWEIS. $\varphi(x) = x - \frac{f(x)}{f'(x)}$ falls $f'(x) \neq 0$

$$f' \in C^1(I_\varepsilon),\ f'(x^*) \neq 0 \quad \Rightarrow \quad \exists I_{\delta_1} \subset I_\varepsilon : f'(x) \neq 0$$

$$\varphi' = 1 - \frac{f'f' - ff''}{f'f'} = \frac{ff''}{f'f'}$$

Aus der Stetigkeit:

$$\exists I_{\delta_2} \subset I_{\delta_1}, \text{ so daß } \sup_{x \in I_{\delta_2}} \left|\varphi'(x)\right| \leq L < 1$$

$\delta = \delta_2$, nach Satz (2.3.3) (lokaler BFS) mit $\xi = x^*$, $r = \delta$ (Kugelbedingung trivial erfüllt wegen $\mathrm{d}(\xi, \varphi(\xi)) = \mathrm{d}(x^*, \varphi(x^*)) = 0$) konvergiert die Iterationsfolge in I_δ. □

Satz 2.5.2. *Sei x^* eine Nullstelle von f. In I_ε sei f zweimal differenzierbar und es gelte $f'(x^*) \neq 0$. Dann existiert $I_\delta = (-\delta + x^*, x^* - \delta) \subset I_\varepsilon$, so daß für jedes $x_0 \in I_\delta$ die Folge $\{x_n\}$ des Newtonverfahrens mindestens quadratisch konvergiert.*

BEWEIS. $x^* - x_{n+1} = x^* - x_n + \frac{f(x_n)}{f'(x_n)}$

$$
\begin{aligned}
0 = f(x^*) &= f(x_n) + f'(x_n)(x^* - x_n) + \frac{1}{2} f''(x_n + \vartheta(x^* - x_n)) \\
&\quad 0 < \vartheta < 1,\ \vartheta = \vartheta_n \\
\Rightarrow x^* - x_{n+1} &= x^* - x_n + \frac{f(x_n)}{f'(x_n)} \\
&= -\frac{1}{2} \frac{f''(x_n + \vartheta(x^* - x_n))}{f'(x_n)} (x^* - x_n)^2 \\
\Rightarrow \lim_{n\to\infty} \frac{x^* - x_{n+1}}{(x^* - x_n)^2} &= -\frac{1}{2} \lim_{n\to\infty} \frac{f''(x_n + \vartheta(x^* - x_n))}{f'(x_n)} \\
&= -\frac{1}{2} \frac{f''(x^*)}{f'(x^*)}
\end{aligned}
$$

□

Bemerkung. Wenn $f(x^*) = f'(x^*) = 0$, dann sind die Sätze (2.5.1) und (2.5.2) nicht anwendbar!

Sei $f \in C^{m+1}[a,b],\ m > 1$. In $x^* \in [a,b]$ sei

$$f(x^*) = f'(x^*) = \cdots = f^{(m-1)}(x^*) = 0 \neq f^{(m)}(x^*)$$

Dann gilt für $h := x - x^* \to 0$ nach Taylor:

$$f(x^* + h) = f(x^*) + \underbrace{\sum_{k=1}^{m-1} \frac{1}{k!} f^{(k)}(x^*)\, h^k}_{=0} + \frac{1}{m!} f^{(m)}(x^*)\, h^m + O(h^{m+1})$$

$$f'(x^* + h) = f'(x^*) + \underbrace{\sum_{k=1}^{m-2} \frac{1}{k!} f^{(k+1)}(x^*)\, h^k}_{=0} + \frac{1}{(m-1)!} f^{(m)}(x^*)\, h^{m-1} + O(h^m)$$

$$f''(x^* + h) = f''(x^*) + \underbrace{\sum_{k=1}^{m-3} \frac{1}{k!} f^{(k+2)}(x^*)\, h^k}_{=0} + \frac{1}{(m-2)!} f^{(m)}(x^*)\, h^{m-2} + O(h^{m-1})$$

Daher für $h \to 0$:

$$\begin{aligned}
\frac{f(x^*)}{f'(x^*)} &= \frac{\frac{1}{m!} f^{(m)}(x^*)\, h^m + O(h^{m+1})}{\frac{1}{(m-1)!} f^{(m)}(x^*)\, h^{m-1} + O(h^m)} \\
&= \frac{\frac{1}{m!} f^{(m)}(x^*)\, h + O(h^2)}{\frac{1}{(m-1)!} f^{(m)}(x^*) + O(h)} \\
&= \left[\frac{1}{m!} f^{(m)}(x^*)\, h + O(h^2)\right] \cdot \left[\frac{(m-1)!}{f^{(m)}(x^*)} + O(h)\right] \\
&= \frac{1}{m} \cdot h + O(h) \\
A &:= f^{(m)}(x^*)
\end{aligned}$$

$$\begin{aligned}
\left(\frac{f}{f'}\right)'(x) &= \frac{f'(x) \cdot f'(x) - f(x) \cdot f''(x)}{[f'(x)]^2} \\
&= 1 - \frac{f(x) f''(x)}{[f'(x)]^2} \\
&= 1 - \frac{\left[\frac{1}{m!} A \cdot h^m + O(h^{m+1})\right] \cdot \left[\frac{1}{(m-2)!} A \cdot h^{m-2} + O(h^{m-1})\right]}{\left[\frac{1}{(m-1)!} A \cdot h^{m-1} + O(h^m)\right] \cdot \left[\frac{1}{(m-1)!} A \cdot h^{m-1} + O(h^m)\right]} \\
&= 1 - \frac{\left[\frac{1}{m} \cdot h^m + O(h^{m+1})\right] \cdot \left[(m-1) \cdot h^{m-2} + O(h^{m-1})\right]}{\left[h^{m-1} + O(h^m)\right] \cdot \left[h^{m-1} + O(h^m)\right]} \\
&= 1 - \frac{\left[\frac{1}{m} \cdot h + O(h^2)\right] \cdot \left[(m-1) \cdot h^{-1} + O(h)\right]}{[1 + O(h)] \cdot [1 + O(h)]} \\
&= 1 - \left[\frac{1}{m} + O(h)\right] \cdot [(m-1) + O(h)] \cdot [1 + O(h)]^2 \\
&= 1 - \frac{m-1}{m} + O(h) \\
&= \frac{1}{m} + O(h)
\end{aligned}$$

Folge:

Bei m-facher Nullstelle konvergiert das Newtonverfahren

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = \varphi(x_n)$$

nur linear mit dem Konvergenzfaktor

$$1 - \frac{1}{m} = \lim_{n \to \infty} \varphi'(x_n)$$

Modifiziertes Newtonverfahren:

$$x_{n+1} = x_n - k \cdot \frac{f(x_n)}{f'(x_n)} = \varphi(x_n) =: \varphi_k(x_n) \qquad k \in \mathbb{N}$$

Satz 2.5.3. *Sei $f \in C^{m+1}[a,b]$, $m > 1$ mit m-facher Nullstelle $x^* \in [a,b]$. Sei $1 \geq k \geq m$. Dann existiert ein $I_\delta = (-\delta + x^*, x^* + \delta)$, so daß für jeden Startwert $x_0 \in I_\delta$ die Folge*

$$x_{n+1} = x_n - k \cdot \frac{f(x_n)}{f'(x_n)}$$

für $k < m$ linear mit dem Konvergenzfaktor $1 - \frac{k}{m}$ gegen x^* konvergiert. Für $k = m$ ist die Konvergenz mindestens quadratisch.

Beweis. (analog zum Beweis von Satz (2.5.2))

$$\varphi_k(x) = x - k \cdot \frac{f(x)}{f'(x)}$$

$$\varphi_k'(x) = 1 - k \cdot \left(\frac{f}{f'}\right)'(x)$$

$$= 1 - \frac{k}{m} - O(h) \longrightarrow \varphi_k'(x^*) = 1 - \frac{k}{m} \text{ für } x \to x^*,\ h = x - x^* \to 0$$

$\Rightarrow$ lineare Konvergenz für $k < m$ mit dem Konvergenzfaktor $1 - \frac{k}{m}$, mindestens quadratische Konvergenz für $k = m$. □

Schätzen der Vielfachheit:

$$\begin{aligned} x_n &= x_{n-1} - k \cdot \frac{f(x_{n-1})}{f'(x_{n-1})} \\ x_{n+1} &= x_n - k \cdot \frac{f(x_n)}{f'(x_n)} \\ (x_{n+1} - x_n) - (x_n - x_{n-1}) &= -k \cdot \left[\frac{f(x_n)}{f'(x_n)} - \frac{f(x_{n-1})}{f'(x_{n-1})}\right] \\ &\overset{\text{MWS}}{=} -k \cdot \underbrace{\left[\left(\frac{f}{f'}\right)'(\xi)\right]}_{\approx \frac{1}{m}} \cdot (x_n - x_{n-1}) \end{aligned}$$

$$\text{Also:} \quad m \approx -k \cdot \frac{x_n - x_{n-1}}{(x_{n+1} - x_n) - (x_n - x_{n-1})}$$

Beispiel.

$$\begin{aligned} f(x) &= x(x+1)(2x-3)^2 \\ &= 4x^4 - 8x^3 - 3x^2 + 9x \end{aligned}$$

Iteration: $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$
Start: $x_0 = 2$

n	x_n	$d_n := x_n - x_{n-1}$	$m \approx -\frac{d_n}{d_{n+1} - d_n}$
1	1,7931	−0,2069	2,6633
2	1,6639	−0,1292	2,4226
3	1,5880	−0,0759	2,2481
4	1,5459	−0,0421	2,1369
⋮	⋮	⋮	⋮
9	1,5030	−0,0030	2,0189
10	1,5015	−0,0015	2,0097
11	1,5008	−0,00075	2,0048
⋮	⋮	⋮	⋮
17	1,500016	$-0,1 \cdot 10^{-}5$	2,3722

Iteration: $x_{n+1} = x_n - 2 \cdot \frac{f(x_n)}{f'(x_n)}$
Start: $x_0 = 2$

n	x_n	$\frac{1}{2}m$
1	1,5862	1,2493
2	1,5036	1,0458
3	1,500007040	1

$e_n = x^* - x_n$, $\frac{e_{n+1}}{e_n^2} \approx c$. Wenn $e_n \approx 10^{-3}$, dann ist $e_{n+1} = c \cdot 10^{-6}$.

Nachteile des Newton-Verfahrens:

- In jedem Schritt eine Neuberechnung von $f'(x_n)$ notwendig.
- Manchmal ist f' nicht explizit bekannt.
- f ist nicht differenzierbar.
- f' ist nur mühsam zu berechnen.

Die Idee ist, die Ableitung f' durch den Differenzenquotienten

$$\frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

zu ersetzen.

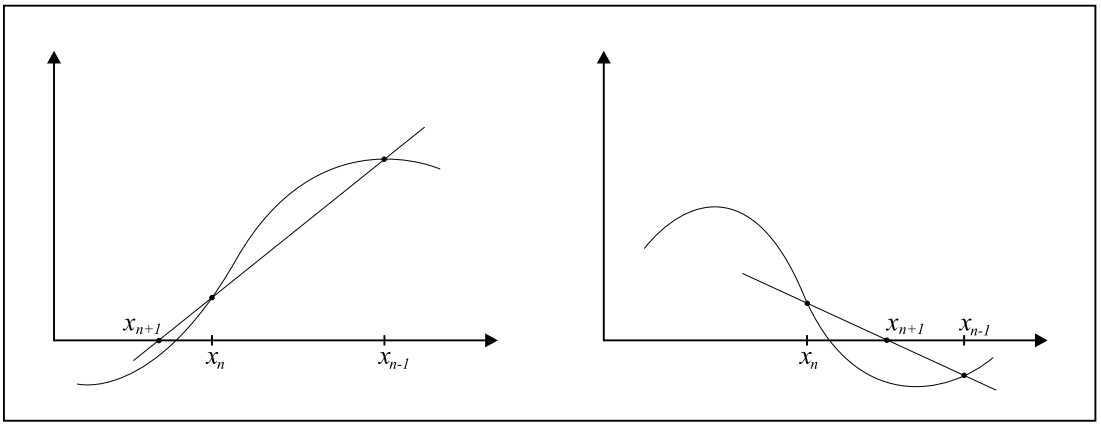


Abbildung 2.4: Sekantenverfahren

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} \cdot f(x_n)$$

Sekantenverfahren, Start mit x_0 und x_1.

Algorithmus 2.5.2 (Sekantenverfahren).

Gegeben:	$x_0, x_1 \in I,\ f \in C(I),\ \varepsilon > 0$
Gesucht:	$x_k \in I :\ \lvert x_k - x^* \rvert < \varepsilon,\ f(x^*) = 0$
Iteration:	$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} \cdot f(x_n)$
	bis das Abbruchktriterium erfüllt ist.

Algorithmus 2.5.3 (Regula falsi).

Gegeben:	$x_0, x_1 \in I,\ f \in C(I),\ \varepsilon > 0$
	$f(x_0) \cdot f(x_1) < 0$
Gesucht:	$x_k \in I :\ \lvert x_k - x^* \rvert < \varepsilon,\ f(x^*) = 0$
Iteration:	$y = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} \cdot f(x_n)$
	$x_{n+1} := y$ *falls* $f(y) \cdot f(x_n) < 0$
	$\left.\begin{array}{rl} x_{n+1} & := y \\ x_n & := x_{n-1} \end{array}\right\}$ *falls* $f(y) \cdot f(x_n) > 0$
	Abbruch sobald $\lvert x_{n+1} - x_n \rvert < \varepsilon$

Literatur: H.R.Schwarz *„Numerische Mathematik“* S.208ff.

Satz 2.5.4. *Es sei $f \in C^2[a,b]$ mit $f(x^*) = 0$ für $x^* = \frac{a+b}{2}$. Außerdem gelte*

1) $f'(x) \neq 0$ *für alle* $x \in [a,b]$

2) $\min\limits_{a\leq x\leq b} |f'(x)| > \frac{b-a}{2} \cdot \max\limits_{a\leq x\leq b} |f''(x)|$

Dann konvergiert das Sekantenverfahren global auf $[a,b]$ gegen x^* mit der Konvergenzordnung

$$\frac{1+\sqrt{5}}{2} \approx 1,618 \quad (f''(x^*) \neq 0)$$

Bemerkung. Ist f in einer Umgebung um x^* streng monoton und strikt konvex, dann hat das Sekantenverfahren die genaue Konvergenzordnung $\frac{1+\sqrt{5}}{2}$ (goldener Schnitt).

Satz 2.5.5. *Für jede stetige Funktion $f : [a,b] \to \mathbb{R}$ konvergiert das Iterationsverfahren der Regula falsi für $x_0 = a, x_1 = b$ und $f(a) \cdot f(b) < 0$ gegen eine Nullstelle x^* von f.*

Bemerkung. Ist $f \in C^2[a,b]$ und $f''(x^*) \neq 0$, dann ist die Konvergenzordnung 1.

Für eine Funktion $F : \Omega \to \mathbb{R}^n,\ \Omega \subseteq \mathbb{R}^n$ gilt

$$F(x) = \begin{pmatrix} F_1(x) \\ \vdots \\ F_n(x) \end{pmatrix},\ F_i(x) : \Omega \to \mathbb{R}^1$$

$x^* \in \mathbb{R}^n : F(x^*) = 0 \Leftrightarrow x^* \in \mathbb{R}^n$ gemeinsame Nullstelle von $F_1, \ldots, F_n$.

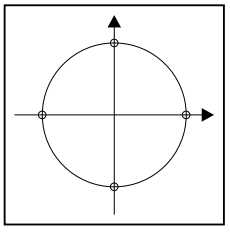
Beispiel. n=2:

$$F(x,y) = \begin{pmatrix} x^2+y^2-1 \\ xy \end{pmatrix}$$

Die Nullstelle (x^*, y^*) ist die gemeinsame Nullstelle von $x^2 + y^2 - 1$ und $x \cdot y$.

$$(x^*)^2 + (y^*)^2 - 1 = 0$$
$$x^* \cdot y^* = 0$$

Nullstellen $(\pm 1, 0), (0, \pm 1)$



Taylor im Punkt $x^{(\nu)} = \begin{pmatrix} x_1^{(\nu)} \\ \vdots \\ x_n^{(\nu)} \end{pmatrix} \in \mathbb{R}^n$

$$F(x) \doteq F(x^{(\nu)}) + \mathrm{J}_F(x^{(\nu)}) \cdot (x - x^{(\nu)}) + \cdots$$

mit

$$\mathrm{J}_F(x^{(\nu)}) := \begin{pmatrix} \frac{\partial F_1}{\partial x_1}(x^{(\nu)}) & \cdots & \frac{\partial F_1}{\partial x_n}(x^{(\nu)}) \\ \vdots & & \vdots \\ \frac{\partial F_n}{\partial x_1}(x^{(\nu)}) & \cdots & \frac{\partial F_n}{\partial x_n}(x^{(\nu)}) \end{pmatrix}$$

Ist x^* eine Nullstelle, dann

$$0 = F(x^{(\nu)}) + \mathrm{J}_F(x^{(\nu)}) \cdot (x^* - x^{(\nu)}) + \cdots$$

Ersetze x^* durch $x^{(\nu+1)}$:

$$\Rightarrow x^{(\nu+1)} = x^{(\nu)} - \underbrace{\mathrm{J}_F^{-1}(x^{(\nu)})\, F(x^{(\nu)})}_{\Delta^{(\nu)}}$$

Newton-Iterationsvorschrift für $F \in C^1(\Omega, \mathbb{R}^n),\ \Omega \subseteq \mathbb{R}^n$.

Numerisch günstiger:

Löse $\mathrm{J}_F(x^{(\nu)}) \cdot \Delta^{(\nu)} = F(x^{(\nu)})$. Dann:

$$x^{(\nu+1)} = x^{(\nu)} - \Delta^{(\nu)}$$

Für multivariables Newton-Verfahren gilt ebenfalls eine lokale Konvergenzaussage und die quadratische Konvergenzordnung, falls x^* eine einfache Nullstelle von F ist.

KAPITEL 3

Polynome

3.1 Polynomiale Approximationen

Taylor:

$f \in C^{n+1}[a,b],\ x_0 \in (a,b)$

$$f(x) = \underbrace{\sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k}_{\text{Polynom}} + R_n(f) \quad \text{mit } \lim_{x\to x_0} \frac{|R_n(f)|}{|x-x_0|} = 0$$

Ist die Reihe alternierend, dann ist die Fehlerabschätzung besser, z.B. :

$$f(x) = \sin x = \sum_{k=0}^{n} (-1)^k \frac{x^{2k+1}}{(2k+1)!} + R_{2n+1}(f)$$

für $x > 0$ ist die Reihe alternierend.

$$R_{2n+1}(f) < \frac{1}{(2k+3)!} \qquad \text{für } x \in [0,1]$$

Ähnlich gutes Konvergenzverhalten für z.B. cos, exp, log in geeigneten Bereichen!
Bei beliebigen Funktionen $f \in C^\infty[a,b]$ treten Probleme auf:

- Die Taylor-Koeffizienten $\frac{f^{(k)}(x_0)}{k!}$ konvergieren oft nur langsam (u.U. auch gar nicht) gegen Null.
- Bei zunehmender Entfernung von x zu x_0 wird die Approximation durch Taylor-Polynome immer schlechter.
- Bei der Rechnung mit endlicher Stellenzahl sind die Polynomkoeffizienten sehr groß, obwohl die Funktionswerte sehr klein sind (Auslöschung!).

Beispiel 3.1.1. $f(x) = e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots + \frac{x^n}{n!} + \cdots$. Diese Reihe ist konvergent für alle $x \in \mathbb{C}$. Für z.B. $x = -10$:

$$e^{-10} = 1 - 10 + \frac{100}{2} - \frac{1000}{6} \pm \cdots$$

Um $e^{-10} = 0,000045399\ldots$ auf drei gültige Stellen auszurechnen, braucht man mindestens 41 Summanden, wenn man 11-stellig rechnet! Der Grund ist die Auslöschung!!

Abhilfe:

$e^{x+y} = e^x \cdot e^y \quad \Rightarrow \quad e^{-10} = \left(\frac{1}{e}\right)^{10}$. Berechne e^{-1} und potenziere.

3.1.1 Kettenbruchentwicklungen

Mit Kettenbrüchen kann man Funktionen mit hoher Genauigkeit darstellen und relativ schnell auswerten.

Definition 3.1.1. Es seien $A = \{a_k\}_{k=1}^{\infty}, B = \{b_k\}_{k=0}^{\infty}$ Folgen reeller oder komplexer Zahlen. Dann nennt man

$$k_n := b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 + \cfrac{\cdots}{\cdots + \cfrac{a_n}{b_n}}}}}$$

den vom A und B erzeugten Kettenbruch der Länge n. Kürzer:

$$k_n = b_0 + \underline{a_1 \big\rceil b_1} + \underline{a_2 \big\rceil b_2} + \cdots + \underline{a_n \big\rceil b_n}$$

Datails: Hardy-Littlewood *Zahlentheorie*

Auswertung:

$$\begin{aligned} z^{(0)} &:= b_n \\ z^{(i)} &:= b_{n-i} + \frac{a_{n-i+1}}{z^{(i-1)}},\ i = 1, \ldots, n \\ k_n &:= z^{(n)} \end{aligned}$$

Es kann passieren, daß $z^{(i-1)} = 0$ ist! Dann Abbruch.

Euler und Wallis:

Für die Berechnung von k_n betrachtet man die Abschnittskettenbrüche (j-te Näherungsbrüche)

$$k_j := b_0 + \underline{a_1 \big\rceil b_1} + \cdots + \underline{a_j \big\rceil b_j} = \frac{P_j}{Q_j},\ j = 0, 1, \ldots, n$$

$$\begin{aligned} P_{-1} &:= 1 \qquad Q_{-1} := 0 \\ P_0 &:= b_0 \qquad Q_0 := 1 \end{aligned}$$

$\frac{P_1}{Q_1} = b_0 + \frac{a_1}{b_1} = \frac{b_0 b_1 + a_1}{b_1}$

$$\begin{aligned} \Rightarrow \quad P_1 &= b_0 b_1 + a_1, & Q_1 &= b_1 \\ &= b_1 P_0 + a_1 P_{-1} & &= b_1 Q_0 + a_1 Q_{-1} \end{aligned}$$

Algorithmus 3.1.1 (Kettenbruchentwicklung).

Gegeben:	$A = \{a_k\}_{k=1}^{n},\ B = \{b_k\}_{k=0}^{n},\ a_k, b_k \in \mathbb{K}$
Gesucht:	$k_n = b_0 + \underline{a_1 \big\rceil b_1} + \cdots + \underline{a_n \big\rceil b_n} \in \mathbb{K}$
Start:	$P_{-1} = 1,\ Q_{-1} = 0,\ P_0 = b_0,\ Q_0 = 1$
Iteration:	*Für* $j = 1, \ldots, n$
	$P_j = b_j P_{j-1} + a_j P_{j-2}, \qquad Q_j = b_j Q_{j-1} + a_j Q_{j-2}$
Ausgabe:	$\frac{P_n}{Q_n} = k_n$

KORREKTHEIT DES ALGORITHMUS. Induktion über die Kettenlänge.

Induktionsannahme:	Algoritmus ist korrekt für Kettenbrüche der Länge $< n$.
Induktionsanfang:	$n = 0,\ n = 1$

$$k_n = b_0 + \underline{a_1 \big\rceil b_1} + \cdots + \underline{a_{n-1} \Big\rceil \left(b_{n-1} + \tfrac{a_n}{b_n}\right)}$$

Kettenbruch der Länge $n-1$.

$$\frac{P_n}{Q_n} = \frac{\left(b_{n-1} + \frac{a_n}{b_n}\right) P_{n-2} + a_{n-1} P_{n-3}}{\left(b_{n-1} + \frac{a_n}{b_n}\right) Q_{n-2} + a_{n-1} Q_{n-3}}$$

$$= \frac{P_{n-1} + \frac{a_n}{b_n} P_{n-2}}{Q_{n-1} + \frac{a_n}{b_n} Q_{n-2}} = \frac{b_n P_{n-1} + a_n P_{n-2}}{b_n Q_{n-1} + a_n Q_{n-2}}$$

□

Beispiel 3.1.2. Für jedes $x \in \mathbb{R}$, $x \geq 1$ gilt:

$$\ln \frac{1-x}{1+x} = \underline{2x \,\big|\, 1} - \underline{1x^2 \,\big|\, 3} - \underline{4x^2 \,\big|\, 5} - \underline{9x^2 \,\big|\, 7} - \underline{16x^2 \,\big|\, 9} - \cdots$$

Wenn man x durch ix ersetzt, dann ergibt sich:

$$\arctan x = \underline{x \,\big|\, 1} + \underline{x^2 \,\big|\, 3} + \underline{4x^2 \,\big|\, 5} + \underline{9x^2 \,\big|\, 7} + \cdots$$

$\arctan(1) = \frac{\pi}{4}$

$$\frac{\pi}{4} = \underline{1 \,\big|\, 1} + \underline{1 \,\big|\, 3} + \underline{4 \,\big|\, 5} + \underline{9 \,\big|\, 7} + \underline{16 \,\big|\, 9} + \cdots \approx 0,7853$$

$$k_0 = \frac{1}{1}, \quad k_1 = \frac{1}{1 + \frac{1}{3}} = \frac{3}{4}, \quad k_2 = \frac{1}{1 + \frac{1}{3 + \frac{4}{5}}} = \frac{1}{1 + \frac{5}{19}} = \frac{19}{24}$$

3.1.2 Gleichmäßige Approximation

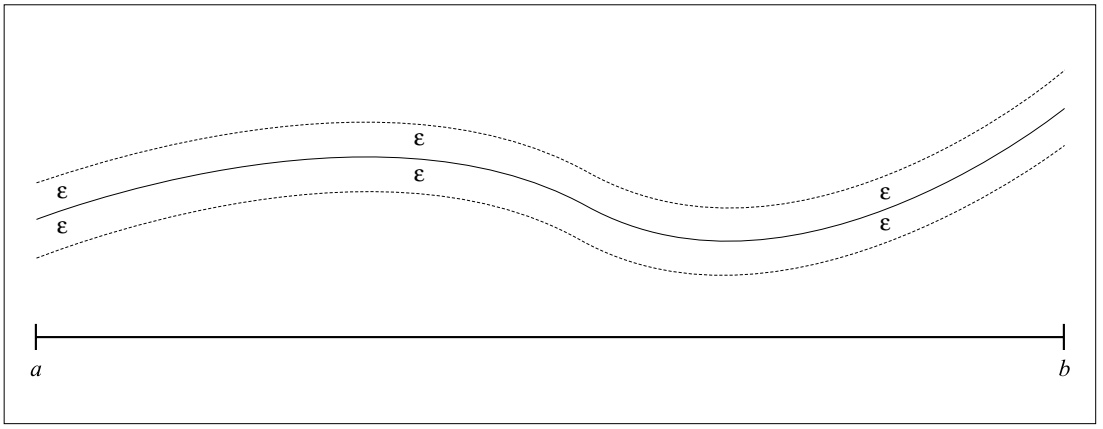


Abbildung 3.1: Gleichmäßige Approximation

Zu $f \in C[a,b]$ gesucht ein Polynom p:

$$\underbrace{\max_{x \in [a,b]} |f(x) - p(x)|}_{\|f-p\|_\infty^{[a,b]}} < \varepsilon, \qquad \| \quad \|_\infty : \text{Maximum-Norm}$$

Approximationssatz von Weierstraß.

$$f \in C[a,b],\ \varepsilon > 0 \quad \Rightarrow \quad \exists\, p \text{ Polynom } : \|f-p\|_\infty^{[a,b]} < \varepsilon$$

Frage: Wie findet man unter allen Polynomen p vom Grad $\leq n$ ein Polynom p^* mit

$$E_n^{[a,b]}(f) := \|f - p^*\|_\infty^{[a,b]} \leq \|f-p\|_\infty^{[a,b]} \quad ?$$

Beispiel 3.1.3. Intervall $[0,\pi]$, $f(x) = \cos x$

$$p(x) = 1 + a_2 x^2 + a_4 x^4, \quad a_2 \approx -0,49670 \quad a_4 \approx 0,03705, \quad \|f-p\|_\infty^{[a,b]} \leq 0,9 \cdot 10^{-4}$$

3.2 Auswertung von Polynomen

$$p_n(x) = a_n x^n + \cdots + a_1 x + a_0$$

Auswertung an der Stelle $x_0 \in \mathbb{K}$

$$p_n(x) = (\cdots \overbrace{(\underbrace{(a_n x + a_{n-1})}_{a_{n-1}^{(1)}} x + a_{n-2})}^{a_{n-2}^{(1)}} x + \cdots + a_1)x + a_0$$

Das Horner-Schema entspricht dieser Kennung.

Algorithmus 3.2.1 (Horner).

Gegeben:	*Koeffizientenvektor* $(a_0, \ldots, a_n) \in \mathbb{K}^{n+1}$ *und* $x_0 \in \mathbb{K}$
Gesucht:	$p_n(x_0)$ *mit* $p_n(x) = \sum_{k=0}^{n} a_k x^k$
Berechnung:	$a_n^{(1)} := a_n$ *für* $k = 1, \ldots, n$
	$a_{n-k}^{(1)} = x_0 \cdot a_{n-k+1}^{(1)} + a_{n-k}$
Ausgabe:	$a_0^{(1)}$
Aufwand:	*n Multiplikationen und n Additionen*

Hornerschema 3.2.1.

	a_n		a_{n-1}		a_{n-2}	$\cdots$		a_1		a_0		
x_0	$-$	$\nearrow$	$x_0 a_n^{(1)}$	$\nearrow$	$x_0 a_{n-1}^{(1)}$	$\cdots$	$\nearrow$	$x_0 a_2^{(1)}$	$\nearrow$	$x_0 a_1^{(1)}$		
	$a_n^{(1)}$	$\searrow$	$a_{n-1}^{(1)}$	$\searrow$	$a_{n-2}^{(1)}$	$\cdots$	$\searrow$	$a_1^{(1)}$	$\searrow$	$a_0^{(1)}$	$=$	$p_n(x_0)$

Bemerkung. $p_n(x) = \sum_{k=0}^{n} a_k x^k = (x - x_0) \sum_{k=0}^{n-1} b_k x^k + p_n(x_0)$

Koeffizientenvergleich:

$$\begin{array}{rcll} a_n & = & b_{n-1} & \Rightarrow \quad b_n = a_n^{(1)} \\ a_{n-1} & = & b_{n-2} - x_0 b_{n-1} & \Rightarrow \quad b_{n-2} = a_{n-1}^{(1)} \\ & \vdots & & \vdots \\ a_1 & = & b_0 - x_0 b_1 & \Rightarrow \quad b_0 = a_1^{(1)} \\ a_0 & = & p(x_0) - x_0 b_0 & \Rightarrow \quad p(x_0) = a_0^{(1)} \end{array}$$

$$\frac{p_n(x) - p_n(x_0)}{x - x_0} = \sum_{k=0}^{n-1} b_k x^k \xrightarrow[x \to x_0]{} \sum_{k=0}^{n-1} b_k x_0^k = p_n'(x_0)$$

	a_n		a_{n-1}	$\cdots$		a_2		a_1		a_0
x_0	$-$	$\nearrow$	$x_0 a_n^{(1)}$	$\cdots$	$\nearrow$	$x_0 a_3^{(1)}$	$\nearrow$	$x_0 a_2^{(1)}$	$\nearrow$	$x_0 a_1^{(1)}$
	$a_n^{(1)}$	$\searrow$	$a_{n-1}^{(1)}$	$\cdots$	$\searrow$	$a_2^{(1)}$	$\searrow$	$a_1^{(1)}$	$\searrow$	$\boxed{p_n(x_0)}$
x_0	$-$	$\nearrow$	$x_0 a_n^{(2)}$	$\cdots$	$\nearrow$	$x_0 a_3^{(2)}$	$\nearrow$	$x_0 a_2^{(2)}$	$\nearrow$	$x_0 a_1^{(2)}$
	$a_n^{(2)}$	$\searrow$	$a_{n-1}^{(2)}$	$\cdots$	$\searrow$	$a_2^{(2)}$	$\searrow$	$a_1^{(2)}$	$\searrow$	$\boxed{p_n'(x_0)}$
x_0	$-$	$\nearrow$	$x_0 a_n^{(3)}$	$\cdots$	$\nearrow$	$x_0 a_3^{(3)}$	$\nearrow$	$x_0 a_2^{(3)}$	$\nearrow$	$x_0 a_1^{(3)}$
	$a_n^{(3)}$	$\searrow$	$a_{n-1}^{(3)}$	$\cdots$	$\searrow$	$a_2^{(3)}$	$\searrow$	$a_1^{(3)}$	$\searrow$	$\boxed{\frac{1}{2!} p_n''(x_0)}$
	$\vdots$		$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\vdots$
	$\cdots$		$\cdots$	$\cdots$		$\cdots$		$\cdots$		$\boxed{a_0^{(n+1)} = \frac{1}{n!} p_n^{(n)}(x_0)}$

Vergleich mit Taylor-Entwicklung:

$$\begin{aligned} p_n(x) &= \sum_{k=0}^{n} \frac{p_n^{(k)}(x_0)}{k!}(x-x_0)^k \\ &= p_n(x_0) + (x-x_0)\cdot\Big(p_n'(x_0) + (x-x_0)\big(\frac{1}{2}p_n''(x_0) + \cdots + (x-x_0)\cdot\frac{1}{n!}p_n^{(n)}(x_0)\big)\Big) \end{aligned}$$

Gibt die Werte $\frac{1}{k!}p_n^{(k)}$ im vollständigen Horner-Schema.

Sei $p_n(x) = \sum\limits_{k=0}^{n} a_k x^k$ mit $a_k \in \mathbb{R}$ und $x_0 \in \mathbb{C}$ eine Nullstelle von p_n.

$$\begin{aligned} p_n(\overline{x}_o) &= \sum_{k=0}^{n} a_k \cdot \overline{x_0^k} = \sum_{k=0}^{n} \overline{a_k x_0^k} \\ &= \overline{\sum_{k=0}^{n} a_k x_o^k} = \overline{p_n(x_0)} = 0 \end{aligned}$$

Um $q_{n-2}(x) = \frac{p_n(x)}{(x-x_0)(x-\overline{x}_0)}$ auszurechnen, kann man entweder erst die Nullstelle x_0 und dann die Nullstelle $\overline{x}_0$ mit dem Horner-Schema abspalten oder beides gleichzeitig.

$$\begin{aligned} (x-x_0)(x-\overline{x}_0) &= x^2 - (x_0+\overline{x}_0)x + x_0\overline{x}_0 \\ &= x^2 - 2\,\mathrm{Re}(x_0)\cdot x + |x_0|^2 \end{aligned}$$

Somit ist q ein Polynom mit *reellen* Koeffizienten. Seien $s := 2\,\mathrm{Re}\,x_0$, $p := -|x_0|^2$. Teilen mit Rest:

$$p_n(x) = (x^2 - sx - p)\cdot q_{n-2}(x) + a_1^{(1)}x + a_0^{(1)}$$

$$\begin{aligned} \sum_{k=0}^{n} a_k x^k &= (x^2 - sx - p)\sum_{k=2}^{n} a_k^{(1)} a^{k-2} + a_1^{(1)}x + a_0^{(1)} \\ &= a_n^{(1)}x^n + (a_{n-1}^{(1)} - sa_n^{(1)})x^{n-1} + \sum_{k=1}^{n-1}(a_k^{(1)} - sa_{k+1}^{(1)} - pa_{k+2}^{(1)})x^k - pa_1^{(1)} + a_0^{(1)} \end{aligned}$$

Der Koeffizientenvergleich gibt eine dreigliedrige Rekursion.

Doppelzeiliges Hornerschema 3.2.2.

Gegeben: $(a_n, \ldots, a_0)$ reeller Koeffizientenvektor von p_n. $s, p \in \mathbb{R}$

Gesucht: Zahlen $a_0^{(1)}, a_1^{(1)} \in \mathbb{R}$ und q_{n-2}, so daß

$$p_n(x) = (x^2 - sx - p)\cdot q_{n-2}(x) + a_1^{(1)}x + a_0^{(1)}$$

Rekursion:

$$\begin{aligned} a_n^{(1)} &= a_n \\ a_{n-1}^{(1)} &= a_{n-1} + sa_n^{(1)} \\ a_k^{(1)} &= a_k + sa_{k+1}^{(1)} + pa_{k+2}^{(1)} \qquad k = n-2, \ldots, 1 \\ a_0^{(1)} &= a_0 + pa_2^{(1)} \end{aligned}$$

Reihenschema 3.2.3.

	a_n		a_{n-1}		a_{n-2}	$\cdots$		a_1		a_0
p	$-$		$-$	$\nearrow$	$pa_n^{(1)}$	$\cdots$	$\nearrow$	$pa_3^{(1)}$	$\nearrow$	$pa_2^{(1)}$
s	$-$	$\nearrow$	$sa_n^{(1)}$	$\nearrow$	$sa_{n-1}^{(1)}$	$\cdots$	$\nearrow$	$sa_2^{(1)}$	$\nearrow$	$-$
	$a_n^{(1)}$	$\nearrow$	$a_{n-1}^{(1)}$	$\nearrow$	$a_{n-2}^{(1)}$	$\cdots$	$\nearrow$	$a_1^{(1)}$	$\nearrow$	$a_0^{(1)}$

Die beiden Schlußkoeffizienten verschwinden genau dann, wenn die Nullstellen von $x^2 - sx - p$ die Nullstellen von p_n sind. $x^2 - sx - p$ hat die Nullstellen

$$\frac{s}{2} \pm \frac{1}{2}\sqrt{s^2 - 4p}$$

Umgekehrt folgt aus $(x-\alpha_1)(x-\alpha_2) = x^2 - (\alpha_1+\alpha_2)x + \alpha_1\alpha_2$ die Beziehung:

$$s = \alpha_1 + \alpha_2, \qquad p = -\alpha_1\alpha_2$$

Satz 3.2.1. *Sind α_1, α_2 Nullstellen des Polynoms p_n, dann lassen sich diese Nullstellen unter Verwendung des doppelzeiligen Hornerschemas mit $s = \alpha_1 + \alpha_2, \quad p = -\alpha_1\alpha_2$ simultan abspalten. Genau in diesem Fall ist $a_0^{(1)} = a_1^{(1)} = 0$.*

3.3 Tschebyscheff-Polynome und -Entwicklungen

In der englischen Literatur (und in MAPLE) Chebyshev, $T_n(x)$.
Sei $1, x, x^2, \ldots, x^n$ die Basis des Raumes aller Polynome vom Grad $\leq n$.
$\mathrm{T}_0, \ldots, \mathrm{T}_n$ (die ersten $n+1$ Tschebyscheff-Polynome) bilden eine andere Basis von $\mathbb{P}_n$.

Definition 3.3.1. Durch die Rekursion

$$\mathrm{T}_0(x) = 1, \ \mathrm{T}_1(x) = x, \qquad \mathrm{T}_{n+1}(x) := 2x\,\mathrm{T}_n(x) - \mathrm{T}_{n-1}(x), \quad n \geq 1$$

werden Polynome definiert. Sie heißen *Tschebyscheff-Polynome (erster Art)*. Durch

$$\mathrm{U}_0(x) = 1, \ \mathrm{U}_1(x) = 2x, \qquad \mathrm{U}_{n+1}(x) := 2x\,\mathrm{U}_n(x) - \mathrm{U}_{n-1}(x), \quad n \geq 1$$

werden Polynome definiert, die *Tschebyscheff-Polynome (zweiter Art)* heißen.

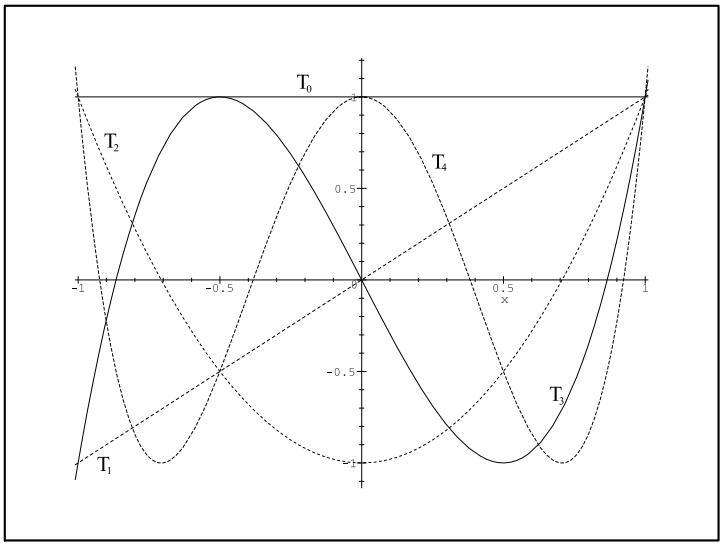


Abbildung 3.2: Tschebyscheff-Polynome

Bemerkung.

$$\mathrm{T}_2(x) = 2x^2 - 1$$
$$\mathrm{T}_3(x) = 4x^3 - 3x$$
$$\mathrm{T}_4(x) = 8x^4 - 8x^2 + 1$$

Leichte Rechnung Zeigt:

$$\mathrm{T}_n(x) = 2^{n-1}x^n \quad + \quad \text{Terme niedriger Ordnung}$$
$$\mathrm{U}_n(x) = 2^n x^n \quad + \quad \text{Terme niedriger Ordnung}$$

Satz 3.3.1. *Die Tschebyscheff-Polynome erster Art haben die Darstellung*

$$\mathrm{T}_n(x) = \cos(n \cdot \arccos x), \quad x \in [-1, 1],\ n \in \mathbb{N}_0$$

und die Eigenschaften:

i) $|\mathrm{T}_n(x)| \leq 1$ *für* $x \in [-1, 1]$

ii) T_n *hat in* $[-1,1]$ *die Extremalpunkte*

$$x_k^{(n)} = \cos\frac{k\pi}{n} \text{ und } \mathrm{T}_n(x_k^{(n)}) = (-1)^n, \quad k = 0,\dots,n$$

iii) T_n *hat n einfache Nullstellen in* $[-1,1]$*, nämlich* $\tilde{x}_k^{(n)} := \cos\left(\frac{2k-1}{2n}\pi\right)$, $k = 1,\dots,n$.

iv) Zwischen je zwei Nullstellen von T_{n+1} *liegt eine Nullstelle von* T_n.

Beweis. Additionstheoreme des Kosinus:

$$\cos(n+1)t = \cos nt\cos t - \sin nt\sin t$$
$$\cos(n-1)t = \cos nt\cos t + \sin nt\sin t$$
$$\cos(n+1)t + \cos(n-1)t = 2\cos nt\cos t$$

Für $\varphi_n(t) := \cos nt$ gilt also die Rekursion:

$$\varphi_{n+1}(t) := 2\cos t\cdot\varphi_n(t) - \varphi_{n-1}(t)$$
$$\varphi_0(t) = 1, \quad \varphi_1(t) = \cos t$$

Die Abbildungen

$$x \mapsto t = \arccos x \qquad [-1,1] \to [0,\pi]$$
$$t \mapsto x = \cos t \qquad [0,\pi] \to [-1,1]$$

sind bijektiv.
Daher sind die Funktionen $\psi_n(x) = \varphi_n(\arccos x) = \cos(n\arccos x)$ auf $[-1,1]$ definiert und genügen der Rekursion

$$\psi_0(x) = 1, \quad \psi_1(x) = x, \quad \psi_{n+1}(x) = 2x\psi_n(x) - \psi_{n-1}(x)$$

Anfangswerte und Rekursion stimmen mit der Folge T_n überein

$$\Rightarrow \mathrm{T}_n(x) = \psi_n(x) \text{ für } x \in [-1,1]$$

$$|\cos nt| \le 1 \quad \forall\, t \in [0,\pi] \Rightarrow i)$$

$$\mathrm{T}_n(x_k^{(n)}) = \cos\left(n\cdot\tfrac{k\pi}{n}\right) = \cos(k\pi) = (-1)^k \Rightarrow ii)$$

$$\mathrm{T}_n(\tilde{x}_k^{(n)}) = \cos\left(n\cdot\tfrac{2k-1}{2n}\cdot\pi\right) = \cos\left(\tfrac{2k-1}{2}\cdot\pi\right) = 0 \Rightarrow iii)$$

$$\tilde{x}_k^{(n)} = \cos\frac{2k+1}{2n+2}\pi < \cos\frac{2k-1}{2n}\pi = \tilde{x}_k^{(n)}$$
$$\tilde{x}_k^{(n)} = \cos\frac{2k-1}{2n+2}\pi < \cos\frac{2k-1}{2k+2}\pi = \tilde{x}_k^{(n+1)} \Rightarrow iv)$$

□

Bemerkung. T_n ist gerade für gerades n und ungerade für ungerades n.

Satz 3.3.2. *Unter allen Polynomen* $p \in \mathbb{P}_n$*, deren Koeffizient bei* x^n *gleich* 1 *ist, hat* $\frac{1}{2^{n-1}}\mathrm{T}_n$ *die kleinste Maximum-Norm im intervall* $[-1,1]$.

$$\min_{p\in\mathbb{P}_n}\|p\|_\infty^{[-1,1]} = \left\|\frac{1}{2^{n-1}}\mathrm{T}_n\right\|_\infty^{[-1,1]} = \frac{1}{2^{n-1}}$$

Beweis. (indirekt) Annahme: Es gibt ein Polynom $p \in \mathbb{P}_n$ mit dem Höchstkoeffizienten 1 und

$$\|p\|_\infty^{[-1,1]} < \frac{1}{2^{n-1}}$$

Dann gilt in den $n+1$ Extremalstellen von T_n:

$$\begin{array}{lcccl} p(x_0^{(n)}) & < & \frac{1}{2^{n-1}} & = & \frac{1}{2^{n-1}}\,\mathrm{T}_n(x_0^{(n)}) \\ p(x_1^{(n)}) & > & -\frac{1}{2^{n-1}} & = & \frac{1}{2^{n-1}}\,\mathrm{T}_n(x_1^{(n)}) \\ p(x_2^{(n)}) & < & \frac{1}{2^{n-1}} & = & \frac{1}{2^{n-1}}\,\mathrm{T}_n(x_2^{(n)}) \\ \vdots & & \vdots & & \vdots \end{array}$$

Das Differenzpolynom $q = p - \frac{1}{2^{n-1}}\,\mathrm{T}_n$ hat ein Minuszeichen in $x_0^{(n)}$, Pluszeichen in $x_1^{(n)}$, Minuszeichen in $x_2^{(n)}$ etc. Daher hat q die Nullstellen in den Intervallen $(x_1^{(n)}, x_0^{(n)})$, $(x_2^{(n)}, x_1^{(n)}), \ldots, (x_n^{(n)}, x_{n-1}^{(n)})$. Also hat q mindestens n verschiedene Nullstellen. Es gilt aber $q \in \mathbb{P}_{n-1}$, daher muß q (wegen Fundametalsatz der Algebra) das Nullpolynom sein. Dies ist ein Widerspruch zu $q(x_0^{(n)}) < 0$, also ist die Annahme falsch! □

Satz 3.3.3. *Die Tschebyscheff-Polynome zweiter Art haben die Darstellung:*

$$\mathrm{U}_n(x) = \frac{\sin\left[(n+1)\arccos x\right]}{\sin(\arccos x)} \qquad x \in [-1,1]$$

und die Eigenschaften:

i) $\mathrm{U}_n(-x) = (-1)^n\,\mathrm{U}_n(x)$, $\mathrm{U}_n(1) = n+1$

ii) $\mathrm{T}_n'(x) = n\,\mathrm{U}_{n-1}(x)$, $n = 1, 2, \ldots$

iii) U_n, $n \geq 1$ *hat n einfache Nullstellen in* $(-1,1)$

$$\tilde{x}_k^{(n)} = \cos\left(\frac{k\pi}{n+1}\right),\ k = 1, \ldots, n$$

es sind die Extremalstellen von T_{n+1} *in* $(-1,1)$.

(ohne Beweis)

Satz 3.3.4. *Die Tschebyscheff-Polynome bilden bezüglich der Gewichtsfunktion*

$$w(x) = \frac{1}{\sqrt{1-x^2}}$$

ein Orthonormalsystem:

$$\int_{-1}^{1} \mathrm{T}_n(x)\,\mathrm{T}_m(x) \cdot \frac{1}{\sqrt{1-x^2}}\,dx = \begin{cases} \pi & \text{für } n = m = 0 \\ \frac{\pi}{2} & \text{für } n = m \neq 0 \\ 0 & \text{für } n \neq m \end{cases}$$

(ohne Beweis)

Konvergiert

$$f(x) = \sum_{k=0}^{\infty} a_k\,\mathrm{T}_k(x)$$

gleichmäßig für alle $x \in [-1,1]$, dann gilt

$$\int_{-1}^{1} f(x)\,\mathrm{T}_m(x) \cdot \frac{1}{\sqrt{1-x^2}}\,dx = \sum_{k=0}^{\infty} a_k \int_{-1}^{1} \mathrm{T}_k(x)\,\mathrm{T}_m(x) \cdot \frac{1}{\sqrt{1-x^2}}\,dx$$

$$= \begin{cases} a_0 \cdot \pi & \text{für } m = 0 \\ a_m \cdot \frac{\pi}{2} & \text{für } m \neq 0 \end{cases}$$

Definition 3.3.2. Sei f stetig in $[-1,1]$. Dann heißt

$$a_k(f) = \frac{2}{\pi} \int_{-1}^{1} f(x)\,\mathrm{T}_k(x) \cdot \frac{1}{\sqrt{1-x^2}}\,dx,\ k = 0,1,2\ldots$$

der k-te *Fourier-Tschebyscheff-Koeffizient* von f. Die formal gebildete Reihe

$$S_f(x) = \frac{a_0(f)}{2} + \sum_{k=1}^{\infty} a_k(f) \cdot \mathrm{T}_k(x)$$

heißt *Fourier-Tschebyscheff-Reihe* von f.

Satz 3.3.5. *Wenn S_f für $f \in C[-1,1]$ auf $[-1,1]$ gleichmäßig konvergiert, d.h.*

$$\lim_{k\to\infty} \left\| \frac{a_0(f)}{2} + \sum_{i=1}^{k} a_i(f) \cdot \mathrm{T}_i(x) - S_f(x) \right\|_{\infty}^{[-1,1]} = 0$$

dann gilt $S_f = f$. (ohne Beweis)

Satz 3.3.6. *Sei $f \in C^2[-1,1]$. Dann gilt*

$$|a_k(f)| \le \frac{C}{k^2},\ k = 1,2,\ldots$$

mit einer nur von f abhängigen Konstanten C. Die Fourier-Tschebyscheff-Reihe konvergiert dann gleichmäßig gegen f. (ohne Beweis)

Auswertung eines Polynoms $p = \sum_{k=0}^{n} a_k\,\mathrm{T}_k \in \mathbb{P}_n$ in $x_0 \in \mathbb{K}$:

Idee 1: Umentwickeln in Darstellung $p(x) = \sum_{k=0}^{n} c_k x^k$
und dann Horner-Algorithmus

Idee 2: Entwickele einen eigenen Algorithmus zur Auswertung
von Tschebyscheff-Darstellung.

$$\begin{aligned}
\mathrm{T}_n(\mathrm{T}_m(x)) &= \mathrm{T}_{n\cdot m}(x)\\
\mathrm{T}_{n+1}(x) &= 2x\,\mathrm{T}_n(x) - \mathrm{T}_{n-1}(x),\ n \ge 1\\
2x\,\mathrm{T}_k(x) &= \mathrm{T}_{k+1}(x) + \mathrm{T}_{k-1}(x),\ k \ge 1\\
2x\,\mathrm{T}_0(x) &= 2\,\mathrm{T}_1(x)
\end{aligned}$$

Ansatz: $p(x) = \sum_{k=0}^{n} a_k\,\mathrm{T}_k(x) = (2x - 2x_0)\sum_{k=0}^{n-1} b_k\,\mathrm{T}_k(x) + p(x_0)$

$$\Big(\mathrm{T}_n(x),\ldots,\mathrm{T}_0(x)\Big) \cdot \begin{pmatrix}
1 & 0 & \cdots & \cdots & \cdots & 0\\
0 & 1 & 0 & & & \vdots\\
1 & 0 & 1 & 0 & & \vdots\\
0 & 1 & 0 & 1 & 0 & \vdots\\
\vdots & & \ddots & \ddots & \ddots & \vdots\\
\vdots & & & 1 & 0 & 2\\
0 & \cdots & \cdots & 1 & 0 & 1
\end{pmatrix} = \Big(2x\,\mathrm{T}_{n-1}(x),\ldots,2x\,\mathrm{T}_0(x),1\Big)$$

$$\Big(\mathrm{T}_n(x),\ldots,\mathrm{T}_0(x)\Big) \cdot \begin{pmatrix}
0 & 0 & \cdots & \cdots & 0\\
-2x_0 & 0 & & & \vdots\\
0 & -2x_0 & 0 & & \vdots\\
\vdots & & \ddots & \ddots & \vdots\\
0 & \cdots & \cdots & -2x_0 & 0
\end{pmatrix} = \Big(-2x_0\,\mathrm{T}_{n-1}(x),\ldots,-2x_0\,\mathrm{T}_0(x),0\Big)$$

$$\Big(\mathrm{T}_n(x),\dots,\mathrm{T}_0(x)\Big)\cdot\begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0\\ -2x_0 & 1 & & & & \cdot\\ 1 & -2x_0 & & & & \cdot\\ \cdot & & \ddots & \ddots & & \cdot\\ \cdot & & & -2x_0 & 2 & \cdot\\ 0 & \dots & \dots & 1 & -2x_0 & 1 \end{pmatrix}$$

$$= \Big(2(x-x_0)\,\mathrm{T}_{n-1}(x),\dots,2(x-x_0)\,\mathrm{T}_0(x),1\Big)$$

$$\Rightarrow 2(x-x_0)\sum_{k=0}^{n-1} b_k\,\mathrm{T}_k(x) + p(x_0) = \Big(2(x-x_0)\,\mathrm{T}_{n-1}(x),\dots,2(x-x_0)\,\mathrm{T}_0(x),1\Big)\cdot\begin{pmatrix} b_{n-1}\\ \vdots\\ b_0\\ p(x_0)\end{pmatrix}$$

$$= \Big(\mathrm{T}_n(x),\dots,\mathrm{T}_0(x)\Big)\cdot\begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0\\ -2x_0 & 1 & & & & \cdot\\ 1 & -2x_0 & & & & \cdot\\ \cdot & & \ddots & \ddots & & \cdot\\ \cdot & & & -2x_0 & 2 & \cdot\\ 0 & \dots & \dots & 1 & -2x_0 & 1 \end{pmatrix}\cdot\begin{pmatrix} b_{n-1}\\ \cdot\\ \cdot\\ \cdot\\ b_0\\ p(x_0)\end{pmatrix}$$

$$\overset{\text{Ansatz}}{=} \Big(\mathrm{T}_n(x),\dots,\mathrm{T}_0(x)\Big)\cdot\begin{pmatrix} a_n\\ \vdots\\ a_1\\ a_0\end{pmatrix}$$

$$= \Big(\mathrm{T}_n(x),\dots,\mathrm{T}_0(x)\Big)\cdot\left\{\begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0\\ -2x_0 & 1 & & & & \cdot\\ 1 & -2x_0 & & & & \cdot\\ \cdot & & \ddots & \ddots & & \cdot\\ \cdot & & & -2x_0 & 2 & \cdot\\ 0 & \dots & \dots & 1 & -2x_0 & 1 \end{pmatrix}\cdot\begin{pmatrix} b_{n-1}\\ \cdot\\ \cdot\\ \cdot\\ b_0\\ p(x_0)\end{pmatrix} - \begin{pmatrix} a_n\\ \cdot\\ \cdot\\ \cdot\\ a_1\\ a_0\end{pmatrix}\right\} = 0$$

$$\Rightarrow \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0\\ -2x_0 & 1 & & & & \cdot\\ 1 & -2x_0 & 1 & & & \cdot\\ \cdot & \ddots & \ddots & \ddots & & \cdot\\ \cdot & & 1 & -2x_0 & 2 & \cdot\\ 0 & \dots & \dots & 1 & -2x_0 & 1 \end{pmatrix}\cdot\begin{pmatrix} b_{n-1}\\ \cdot\\ \cdot\\ \cdot\\ b_0\\ p(x_0)\end{pmatrix} = \begin{pmatrix} a_n\\ \cdot\\ \cdot\\ \cdot\\ a_1\\ a_0\end{pmatrix}$$

$$\begin{aligned} b_{n-1} &= a_n\\ b_{n-2} &= a_{n-1} + 2x_0 b_{n-1}\\ b_{n-k} &= a_{n-k+1} + 2x_0 b_{n-k+1} - b_{n-k+2}, \qquad k = 3,\dots,n-1\\ \beta_0 := 2b_0 &= a_1 + 2x_0 b_1 - b_2\\ p(x_0) &= a_0 + x_0\beta_0 - b_1 \end{aligned}$$

Schema des Clenshaw-Algorithmus:

	a_n		a_{n-1}		a_{n-2}	$\cdots$		a_1		a_0
-1	–		–	↱	$-b_{n-1}$	$\cdots$	↱	$-b_2$	↱	$-b_1$
$2x_0$	–	↱	$2x_0 b_{n-1}$	↱	$2x_0 b_{n-2}$	$\cdots$	↱	$2x_0 b_1$	↱	$x_0\beta_0$
	b_{n-1}		b_{n-2}		b_{n-3}	$\cdots$		β_0		$p(x_0)$

Resultat: Es gilt

$$\sum_{k=0}^{n} a_k\,\mathrm{T}_k(x) = 2(x-x_0)\left[\frac{\beta_0}{2} + \sum_{k=1}^{n-1} b_k\,\mathrm{T}_k(x)\right] + p(x_0)$$

für $p(x) = \sum_{k=0}^{n} a_k\,\mathrm{T}_k(x)$.

Der Clenshaw-Algorithmus erfordert n Multiplikationen und $2n-1$ Additionen.

3.4 Einschließungssätze für Polynomnullstellen

Satz 3.4.1. *Sei $p_n \in \mathbb{P}_n$. Für jedes $z \in \mathbb{C}$ liegt mindestens eine Nullstelle von p_n in dem Kreis um z mit dem Radius*

$$r := n \cdot \left| \frac{p_n(z)}{p_n'(z)} \right|$$

Beweis. Sei $\operatorname{grad}(p_n) = n$

$$p_n(z) = a_n(z - z_1)(z - z_2) \cdots (z - z_n)$$

Für $z = z_k$ ist $p_n(z_k) = 0$ und $r = 0$, d.h. der Satz gilt trivial. Für $z \neq z_k$:

$$\begin{aligned} \frac{p_n'(z)}{p_n(z)} &= \sum_{k=1}^{n} \frac{a_n(z - z_1) \cdots (z - z_{k-1})(z - z_{k+1}) \cdots (z - z_n)}{a_n(z - z_1) \cdots (z - z_k) \cdots (z - z_n)} \\ &= \sum_{k=1}^{n} \frac{1}{z - z_k} \end{aligned}$$

Also:

$$\begin{aligned} \left| \frac{p_n'(z)}{p_n(z)} \right| &= \left| \sum_{k=1}^{n} \frac{1}{z - z_k} \right| \leq \sum_{k=1}^{n} \left| \frac{1}{z - z_k} \right| \\ &= \sum_{k=1}^{n} \frac{1}{\min_{j=1}^{n} |z - z_j|} \leq n \cdot \frac{1}{\min_{j=1}^{n} |z - z_j|} \end{aligned}$$

$$\Rightarrow \quad \min_{j=1}^{n} |z - z_j| \leq n \cdot \left| \frac{p_n(z)}{p_n'(z)} \right|$$

Gilt $\operatorname{grad}(p_n) < n$, dann:

$$\min_{p_n(z)=0} |z - z_k| \leq \operatorname{grad}(p_n) \left| \frac{p_n(z)}{p_n'(z)} \right|$$

□

Lemma 3.4.1 (Frobenius-Begleitmatrix). *Sei*

$$p_n(z) = \sum_{k=0}^{n} a_k z^k \quad \textit{mit} \quad a_n = 1$$

Die Nullstellen von p_n sind die Eigenwerte der Matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & 0 \\ \vdots & & & & \vdots \\ 0 & \cdots & \cdots & 0 & 1 \\ -a_0 & -a_1 & \cdots & \cdots & -a_{n-1} \end{pmatrix} \quad \textit{Frobenius-Begleitmatrix}$$

Zum Eigenwert λ von A gehört der Eigenvektor

$$(1, \lambda, \lambda^2, \ldots, , \lambda^{n-1})^\top$$

Beweis. Mit vollständiger Induktion über n zeigt man daß

$$\det(A - \lambda E) = (-1)^n p_n(\lambda)$$

$n = 1$:

$$p_1(x) = x + a_0, \ (A - \lambda E) = -a_0 - \lambda = -p_1(\lambda) \quad \checkmark$$

$n-1 \Rightarrow n$:

$$A = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ \cdot & & & & \\ \cdot & & \boxed{\tilde{A}} & & \\ \cdot & & & & \\ -a_0 & & & & \end{pmatrix}$$

$\tilde{A}$ ist die Frobenius-Begleitmatrix zum Polynom

$$\tilde{p}_{n-1}(x) = a_1 + a_2 x + \cdots + a_{n-1} x_1 n - 2 + x^{n-1}$$

$$\det(A - \lambda E) = \begin{pmatrix} -\lambda & 1 & 0 & \dots & 0 \\ \cdot & & & & \\ \cdot & & \boxed{\tilde{A} - \lambda E_{n-1}} & & \\ \cdot & & & & \\ -a_0 & & & & \end{pmatrix}$$

Entwicklung nach der ersten Zeile:

$$= -\lambda \cdot \det(\tilde{A} - \lambda E) - \begin{vmatrix} 0 & 1 & 0 & \dots & \dots & 0 \\ 0 & -\lambda & 1 & 0 & \dots & 0 \\ \cdot & & -\lambda & 1 & 0 & \cdot \\ \cdot & & & & & \cdot \\ \cdot & & & & & \cdot \\ \cdot & & & & & \cdot \\ 0 & \dots & \dots & \dots & 0 & 1 \\ -a_0 & -a_2 & \dots & \dots & -a_{n-1} & -\lambda \end{vmatrix}$$

$$= -\lambda \cdot (-1)^{n-1} \tilde{p}_{n-1}(\lambda) - (-1)^n (-a_0) \cdot \begin{vmatrix} 1 & \dots & \dots & 0 \\ -\lambda & 1 & \cdot & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \dots & -\lambda & 1 \end{vmatrix}$$

$$= (-1)^n (a_1 \lambda + a_2 \lambda^2 + \cdots + \lambda^n) + a_0 (-1)^n$$
$$= (-1)^n p_n(\lambda)$$

$$z \begin{pmatrix} 1 \\ z \\ \vdots \\ z^{n-2} \\ z^{n-1} \end{pmatrix} = \begin{pmatrix} z \\ z^2 \\ \vdots \\ z^{n-1} \\ z^n \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & \dots & \dots & 0 & 1 \\ -a_0 & -a_1 & \dots & \dots & -a_{n-1} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ z \\ \vdots \\ z^{n-2} \\ z^{n-1} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ p_n(z) \end{pmatrix} \qquad (\star)$$

Ist z eine Nullstelle von p_n, so ist z ein Eigenwert von A mit dem Eigenvektor

$$(1, z, z^2, \dots, z^{n-1})^\top$$

□

Die folgende Bemerkung, sowie die beiden Beispiele werden im Rahmen dieser Vorlesung nicht gebraucht. Sie können zur Not übergangen werden (Fortsetzung bei Satz 3.4.2).

Bemerkung. Durch gliederweises Differenzieren der Identität ($\star$) kann man zeigen, daß

$$V_k(\lambda) = \frac{1}{k!} \frac{d^{k-1}}{dz^{k-1}} \begin{pmatrix} 1 \\ z \\ \vdots \\ z^{n-2} \\ z^{n-1} \end{pmatrix} \Bigg|_{z=\lambda} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ (k-1)! \\ \vdots \\ \frac{(n-1)!}{(n-k)!} \lambda^{n-k} \end{pmatrix}$$ die Null kommt oben $k-1$ mal vor

Insbesondere $V_1(\lambda) = (1, \lambda, \lambda^2, \ldots, \lambda^{n-1})^\top$, $V_2(\lambda) = (0, 1, 2\lambda, \ldots, (n-1)\lambda^{n-2})^\top$, etc.
Aus $(\star)$ folgt falls $p_n(\lambda) = p_n'(\lambda) = \cdots = p_n^{(k)}(\lambda) = 0$

$$\boxed{V_{k-1}(\lambda) + \lambda V_k(\lambda) = A \cdot V_k(\lambda)}$$

Ist λ_i eine i-fache Nullstelle von p_n, $i = 1, \ldots, s$ $\sum_{i=1}^{s} m_i = n$, dann gilt:

$$A[V_1(\lambda_i), V_2(\lambda_i), \ldots, V_{m_i}(\lambda_i)] = [V_1(\lambda_i), V_2(\lambda_i), \ldots, V_{m_i}(\lambda_i)] \cdot \underbrace{\begin{bmatrix} \lambda_i & 1 & \ldots. & 0 \\ 0 & \lambda_i & 1 \quad \cdot & 0 \\ \cdot & 0 & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & \lambda_i & 1 \\ 0 & 0 & \ldots. & \lambda_i \end{bmatrix}}_{\mathrm{J}_i \text{ : Jordankästchen}}$$

$$A[V_1(\lambda_1), V_2(\lambda_1), \ldots, V_{m_1}(\lambda_1), V_1(\lambda_2), \ldots, V_{m_2}(\lambda_2), \ldots, V_{m_s}(\lambda_s)]$$

$$= \underbrace{[V_1(\lambda_1), \ldots, V_{m_s}(\lambda_s)]}_{V} \cdot \begin{pmatrix} \boxed{\mathrm{J}_1} & & & \\ & \boxed{\mathrm{J}_2} & & \\ & & \ddots & \\ & & & \boxed{\mathrm{J}_s} \end{pmatrix}$$

$$V^{-1} \cdot A \cdot V = \begin{pmatrix} \boxed{\mathrm{J}_1} & & & \\ & \boxed{\mathrm{J}_2} & & \\ & & \ddots & \\ & & & \boxed{\mathrm{J}_s} \end{pmatrix}$$

Beispiel 3.4.1.

$$p_4 = (z-1)(z-2)(z-3)(z-4) = z^4 - 10z^3 + 35z^2 - 50z + 24$$

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -24 & 50 & -35 & 10 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 64 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 64 \\ 1 & 16 & 81 & 256 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 64 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix}$$

Beispiel 3.4.2.

$$p_5 = (z+1)^2 z^2 (z-2) = z^5 - 3z^3 - 2z^2$$

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 2 & 3 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 1 & 0 & 1 \\ -1 & 1 & 0 & 1 & 2 \\ 1 & -2 & 0 & 0 & 4 \\ -1 & 3 & 0 & 0 & 8 \\ 1 & -4 & 0 & 0 & 16 \end{pmatrix} = \begin{pmatrix} -1 & 1 & 0 & 1 & 2 \\ 1 & -2 & 0 & 0 & 8 \\ 1 & 3 & 0 & 0 & 8 \\ 1 & -4 & 0 & 0 & 16 \\ -1 & 5 & 0 & 0 & 32 \end{pmatrix}$$

$$= V \cdot \begin{pmatrix} \boxed{\begin{smallmatrix} -1 & 1 \\ 0 & -1 \end{smallmatrix}} & & \\ & \boxed{\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}} & \\ & & \boxed{2} \end{pmatrix}$$

Satz 3.4.2 (Gerschgorinscher Kreissatz). *Sei $A = (a_{ij})_{i,j=1}^n$ eine Matrix mit $a_{ij} \in \mathbb{K}$. Sei K_i die abgeschlossene Kreisscheibe um a_{ii} mit dem Radius*

$$r_i := \sum_{\substack{j=1 \\ j \neq i}}^{n} \left| a_{ij} \right|, \quad i = 1, \ldots, n$$

Dann liegt jeder Eigenwert von A in mindestens einem K_i, d.h.

$$\{\lambda \in \mathbb{C} \mid \lambda \text{ Eigenwert von } A\} \subset \bigcup_{i=1}^{n} K_i$$

BEWEIS. Zum Eigenwert λ existiert ein Eigenvektor $w = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} \neq 0$. O.B.d.A. gilt $\|w\|_\infty = 1$, $w_i = 1$. Aus $A \cdot w = \lambda w$ folgt für die i-te Zeile

$$a_{ii} \cdot w_i - \lambda w_i = -\sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} w_j$$

$$|a_{ii} - \lambda| \leq \left| \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} w_j \right| \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} \left|a_{ij}\right| \left|w_j\right| \leq r_i$$

Also $\lambda \in K_i$. □

Bemerkung. A und $A^\top$ haben dieselben Eigenwerte.

Satz 3.4.3. *Die Vereinigung der Kreisscheiben*

$$K_i^\top = \{z \in \mathbb{C} \mid |z - a_{ii}| \leq \sum_{\substack{k=1 \\ k \neq i}} |a_{ki}|\}$$

enthält ebenfalls alle Eigenwerte der Matrix A.

Korollar 3.4.3. *Alle Eigenwerte von A liegen in*

$$\left(\bigcup_{i=1}^{n} K_i\right) \cap \left(\bigcup_{i=1}^{n} K_i^\top\right)$$

Anwendung des Gerschgorinschen Kreissatzes auf die Frobeniusmatrix:

Satz 3.4.4. *Die Nullstellen $z_1, \ldots, z_n$ des Polynoms*

$$p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$$

genügen den Abschätzungen:

i) $|z_k| \leq \max\{1, \sum_{j=0}^{n-1} \left|a_j\right|\}$

ii) $|z_k| \leq \max\{|a_0|, 1 + |a_1|, \ldots, 1 + |a_{n-1}|\}$

BEWEIS. Für i):

$$\begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 \ldots & 0 \\ \vdots & & & & \vdots \\ 0 & \ldots & \ldots & 0 & 1 \\ -a_0 & -a_1 & \ldots & \ldots & -a_{n-1} \end{pmatrix} \qquad \begin{aligned} K_1 &= \{|\lambda - 0| \leq 1\} \\ &\cdot \\ &\cdot \\ &\cdot \\ K_{n-1} &= \{\lambda \mid |\lambda| \leq 1\} \end{aligned}$$

$$K_n: \quad |\lambda + a_{n-1}| \leq \sum_{j=0}^{n-2} \left|a_j\right|$$

$$\Rightarrow \quad |\lambda| - |a_{n-1}| \leq |\lambda + a_{n-1}| \leq \sum_{j=0}^{n-2} \left|a_j\right|$$

$$\Rightarrow \quad |\lambda| \leq \sum_{j=0}^{n-1} \left|a_j\right|$$

Für ii): durch Anwendung von Satz 3.4.2 auf

$$\begin{pmatrix} 0 & 0 & 0 & \dots & -a_0 \\ 1 & 0 & 0 & \dots & -a_1 \\ \vdots & & & & \vdots \\ 0 & \dots & \dots & 0 & 1 \\ 0 & \dots & \dots & 1 & -a_{n-1} \end{pmatrix}$$

□

3.5 Sturmsche Ketten und das Bisektionsverfahren

In diesem Abschnitt ist alles reell, d.h. Polynome mit reellen Koeffizienten. Nur reelle Nullstellen von Polynomen werden gezählt.
Sprechweise: $\sigma := (f_0, f_1, \dots, f_n)$ geordnete Liste von Objekten heißt Sequenz.

Definition 3.5.1. Sei $\sigma := (f_0, f_1, \dots, f_n)$ eine Sequenz reeller Polynome. σ heißt Sturmsche Kette auf $[a, b]$, wenn gilt:

i) $f_0(a) \neq 0 \neq f_0(b)$

ii) $f_n(x) \neq 0$ für alle $x \in [a, b]$

iii) $1 \leq k \leq n,\ f_k(\xi) = 0,\ \xi \in [a, b] \Rightarrow f_{k-1}(\xi)\, f_{k+1}(\xi) < 0$

iv) $\xi \in (a, b),\ f_0(\xi) = 0 \Rightarrow f_0'(\xi)\, f_1(\xi) > 0$

Definition 3.5.2. Sei $\sigma = (a_0, a_1, \dots, a_n)$ mit $a_i \in \mathbb{R}$. Dann ist $w(\sigma)$ die Anzahl der Zeichenwechsel in der Sequenz ohne Berücksichtigung der Nullen. Formal:

i) $$w(a_0, a_1, \dots, a_n) = \begin{cases} w(a_1, \dots, a_n) + 1 & \text{falls } a_0\, a_1 < 0 \\ w(a_1, \dots, a_n) & \text{falls } a_0\, a_1 > 0 \text{ oder } a_0 = 0 \\ w(a_0, a_2, \dots, a_n) & \text{falls } a_0 \neq 0,\ a_1 = 0 \end{cases}$$

ii) $w(a_n) = 0$

$w(\sigma, x) = w(f_0(x), \dots, f_n(x))$ für $\sigma = (f_0, \dots, f_n)$
z.B. $w(1, 2, -3, 0, -5, 1) = 2$

Satz 3.5.1. *Sei σ eine Sturmsche Kette auf $[a, b]$. Dann ist $w(\sigma, a) - w(\sigma, b)$ die Anzahl der Nullstellen von f_0 in $[a, b]$.*

Beweis. $x \mapsto w(\sigma, x), \quad [a, b] \to \mathbb{N}_0$

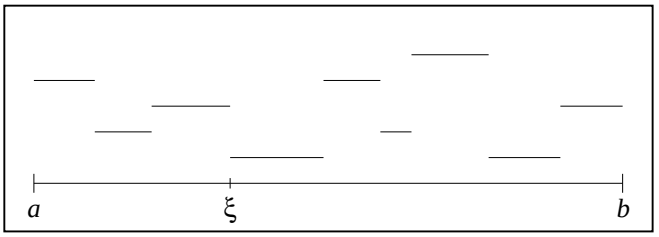


Abbildung 3.3: Die Sprungstellen sind Nullstellen eines f_k

ξ sei Nullstelle von $f_k,\ 1 \leq k < m$

$\tau_k = \{f_{k-1}, f_k, f_{k+1}\}$ Sequenz

x	$f_{k-1}(x)$	$f_k(x)$	$f_{k+1}(x)$	$w(\tau_k, x)$
$\xi - 0$	+	±	−	1
ξ	+	0	−	1
$\xi + 0$	+	±	−	1

oder

x	$f_{k-1}(x)$	$f_k(x)$	$f_{k+1}(x)$	$w(\tau_k, x)$
$\xi - 0$	$-$	$\pm$	$+$	1
ξ	$-$	0	$+$	1
$\xi + 0$	$-$	$\pm$	$+$	1

Fazit: Bei Nullstellen der f_k, $1 \le k < m$ ändert sich die Wechselzahl nicht!

ξ Nullstelle von f_0, $\tau_0(f_0, f_1)$
Wegen $f_0'(\xi)\, f_1(\xi) > 0 \Rightarrow$ Einfache Nullstelle und f_1 hat konstantes Vorzeichen in der Umgebung von ξ.

x	$f_0(x)$	$f_1(x)$	$w(\tau_0, x)$
$\xi - 0$	$+$	$+$	0
ξ	0	$+$	0
$\xi + 0$	$-$	$+$	1

x	$f_0(x)$	$f_1(x)$	$w(\tau_0, x)$
$\xi - 0$	$+$	$-$	1
ξ	0	$-$	0
$\xi + 0$	$-$	$-$	0

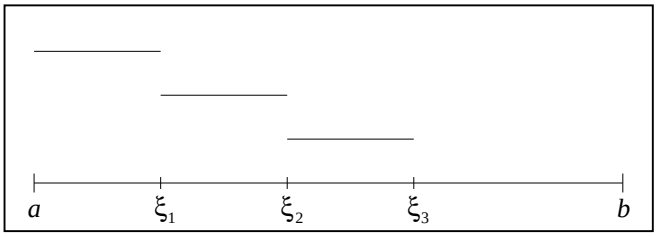
x	$f_0(x)$	$f_1(x)$	$w(\tau_0, x)$
$\xi - 0$	$-$	$+$	1
ξ	0	$+$	0
$\xi + 0$	$+$	$+$	0

x	$f_0(x)$	$f_1(x)$	$w(\tau_0, x)$
$\xi - 0$	$-$	$-$	0
ξ	0	$-$	0
$\xi + 0$	$+$	$-$	1

entweder: f_0 wächst monoton, $f_1 > 0$
oder: f_0 fällt monoton, $f_1 < 0$
Fall 1 und Fall 4 fallen weg

Fazit: Für $f_0(\xi)$ gilt:

$$w(\sigma, \xi - 0) = w(\sigma, \xi + 0) + 1 \qquad \xi \in (a, b)$$



$\Rightarrow w(\sigma, a) - w(\sigma, b) =$ Anzahl der Nullstellen von f_0. □

Konstruktion einer Sturmschen Kette

f_0 ist ein Polynom des Grads n.

$f_1 := f_0'$ Polynom des Grads $n - 1$

Euklidischer Divisionsalgorithmus:

$$f_0 = q_1 f_1 - f_2 \qquad \operatorname{grad}(f_2) < \operatorname{grad}(f_1)$$
$$f_1 = q_2 f_2 - f_3 \qquad \operatorname{grad}(f_3) < \operatorname{grad}(f_2)$$

etc...

$$f_{s-1} = q_s f_s - f_{s+1}$$
$$f_s = q_{s+1} f_{s+1}$$
$$f_{s+1} = \mathrm{ggT}(f_0, f_1) \neq 0$$

$\mathrm{ggT}(f_0, f_0') = \text{const.} \Rightarrow$ alle Nullstellen von f_0 sind einfach

Fall 1: $f_{s+1} = \text{const.}$

Behauptung. $\sigma(f_0, f_1, \ldots, f_{s+1})$ ist Sturmsche Kette für alle $[a, b]$ mit $f_0(a) \neq 0 \neq f_0(b)$.

BEWEIS. i) ist trivial

ii) gilt wegen $f_{s+1} = \text{const.}$

iii) $1 \le k \le s : f_k(\xi) = 0 \Rightarrow f_{k-1}(\xi) = -f_{k+1}(\xi)$

iv) $f_1 = f_0' \Rightarrow f_0'(\xi) \cdot f_1(\xi) > 0$
einfache Nullstelle $\Rightarrow f_0'(\xi) \neq 0$

□

Fall 2: f_{s+1} nicht konstant

$$g_k := \frac{f_k}{f_{s+1}}, \quad 0 \le k \le s+1 \quad \Rightarrow g_{s+1} = 1$$

Behauptung. $\sigma(g_0, g_1, \ldots, g_{s+1})$ ist Sturmsche Kette für alle $[a, b]$ mit $g_0(a) \neq 0 \neq g_0(b)$.

BEWEIS. i) ist trivial

ii) gilt wegen $g_{s+1} = 1$

iii) $f_{k-1}(x) = g_k(x)f_k(x) - f_{k+1}(x) \quad | : f_{s+1}(x)$
$g_{k-1}(x) = g_k(x)g_k(x) - g_{k+1}(x)$

iv) $\left(\frac{f_0}{f_{s+1}}\right)'(\xi) = \frac{f_0'(\xi)}{f_{s+1}(\xi)} \qquad g_1(\xi) = \frac{f_1(\xi)}{f_{s+1}(\xi)} = \frac{f_0'(\xi)}{f_{s+1}(\xi)}$
$\frac{f_0}{f_{s+1}}$ hat nur einfache Nullstellen $\Rightarrow \left(\frac{f_0}{f_{s+1}}\right)'(\xi) \neq 0 \Rightarrow$ iv)

□

Bemerkung. In Fall 2 ist

$$\sigma' = \left(\frac{f_0}{f_{s+1}}, \frac{f_1}{f_{s+1}}, \cdots, \frac{f_{s+1}}{f_{s+1}}\right)$$

eine Sturmsche Kette. $\sigma(f_0, \ldots, f_{s+1})$

$$w(\sigma, x) = w(\sigma', x) \quad \text{für alle } x \text{ mit } f_{s+1}(x) \neq 0$$

Das heißt: zur Berechnung der Wechselzahl im Satz von Sturm reicht es in jedem Fall die Wechselzahl $w(\sigma, a) - w(\sigma, b)$ zu berechnen.

Sturmsche Ketten sind auch gut, wenn man nur eine bestimmte (z.B. „die zweite“) Nullstelle von f_0 ausrechnen will.

Algorithmus 3.5.1.

Gegeben:	*Sturmsche Kette* $(f_0, \ldots, f_m)$, *Intervall* $[a, b]$, *das alle reelle Nullstellen von* f_0 *enthält.* $m := w(\sigma, a) - w(\sigma, b) \qquad k \in \{1, \ldots, m\},\ \varepsilon > 0$
Gesucht:	*Die reelle Nullstelle* ξ_k *von* f_0 *mit* $\xi_1 < \xi_2 < \cdots < \xi_m$, *genauer:* *ein Intervall* $[a', b']$ *mit* $\xi_k \in [a', b']$ *und* $b' - a' < \varepsilon$
Iteration:	*Berechne für* $c := \frac{a+b}{2}$ *die Zahl* $w(\sigma, c)$. *Wenn* $w(\sigma, c) - w(\sigma, a) < k$, *dann ersetze a durch c, d.h.* $a := c$, *sonst ersetze b durch c,* $b := c$. *Abbruch sobald* $b - a < \varepsilon$, *sonst weiter mit Iteration.*

3.6 Anwendung des Newtonverfahrens

Bisher: Algorithmus 2.5.1

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \qquad \text{für } f \in C^1[a, b]$$

sowie

$$x^{(x+1)} = x^k - \mathrm{J}_F^{-1}(x_k)\, F(x_k), \qquad F : \mathbb{R}^n \to \mathbb{R}^n$$

Newton-Verfahren für komplexe Funktionen:

$$z_{k+1} = z_k - \frac{f(z_k)}{f'(z_k)} \qquad \text{für } f \in C^1(G),\ G \subseteq \mathbb{C}$$

Startwertschätzen durch „Höhenlinien“:

$$|f| : z \mapsto |f(z)|$$

Fehlerabschätzung mit Satz 3.4.1 für Polynome $p_n \in \mathbb{P}_n$:

$$|x_k - x^*| \le n \frac{|p_n(x_k)|}{|p_n'(x_k)|}$$

Wie bekommt man alle Nullstellen eines Polynoms?
Im Beweis zum Satz 3.4.1 steht:

$$\frac{p_n'(x_k)}{p_n(x_k)} = \sum_{i=1}^{n} \frac{1}{x_n - \xi_i} \qquad (\xi_1 \cdots \xi \text{ Nullstellen von } p_n)$$

Sei $f = p_n$:

$$x_{k+1} = x_k - \frac{1}{\frac{f'(x_k)}{f(x_k)}} = x_k - \frac{1}{\sum_{i=1}^{n} \frac{1}{x_k - \xi_i}}$$

Wenn $\xi_1, \ldots, \xi_s$ schon bekannt sind, dann (alternative Rechenvorschrift):

$$x_{k+1} = x_k - \frac{1}{\underbrace{\frac{f'(x_k)}{f(x_k)} - \sum_{i=1}^{s} \frac{1}{x_k - \xi_i}}_{= \sum_{i=s+1}^{n} \frac{1}{x_k - \xi_i} = \frac{g'(x_k)}{g(x_k)}}}$$

mit

$$g(x) = a_n(x - \xi_{s+1}) \cdots (x - \xi_n) = \frac{f(x_n)}{(x - \xi_1) \cdots (x - \xi_s)} \qquad \boxed{\text{Deflationsverfahren}}$$

Problem: Finde die komplexen Nullstellen eines Polynoms $p_n(x) = \sum_{k=0}^{n} a_k x^k$ mit $a_k \in \mathbb{R}$ unter Vermeidung der komplexen Arithmetik.
Lösung mit Bairstow-Verfahren. p_n hat mit der Nullstelle x^* auch die Nullstelle $\overline{x}^*$

$$\Leftrightarrow x^2 - 2\,\mathrm{Re}\, x^* - |x^*|^2 \cdot x \qquad \text{teilt } p_n$$

Nach Reihenschema 3.2.3 gilt:

$$\sum_{k=0}^{n} a_k x^k = (x^2 - sx - p) \sum_{k=2}^{n} b_k x^{k-2} + b_1 x + b_0$$

	a_n		a_{n-1}		a_{n-2}	$\cdots$		a_1		a_0
p	$-$		$-$	↱	pb_n	$\cdots$	↱	pb_3	↱	pb_2
s	$-$	↱	sb_n	↱	sb_{n-1}	$\cdots$	↱	sb_2	↱	$-$
	b_n		b_{n-1}		b_{n-2}	$\cdots$		$\|\|b_1$		b_0

$x^2 - sx - p$ ist Teiler von $p_n \Leftrightarrow b_1 = b_2 = 0$. b_1 und b_2 hängen von s und p ab! Lösung mit dem Newton-Verfahren in zwei Variablen:

$$b_1(s, p) = 0$$
$$b_0(s, p) = 0$$

Berechnung von $\frac{\partial b_i}{\partial s}, \frac{\partial b_i}{\partial p},\ i = 0, 1$ für die Iterationvorschrift:

$$\begin{pmatrix} s_{k+1} \\ p_{k+1} \end{pmatrix} = \begin{pmatrix} s_k \\ p_k \end{pmatrix} - \left.\begin{pmatrix} \frac{\partial b_0}{\partial s} & \frac{\partial b_0}{\partial p} \\ \frac{\partial b_1}{\partial s} & \frac{\partial b_1}{\partial p} \end{pmatrix}^{-1}\right|_{(s_k, p_k)} \cdot \begin{pmatrix} b_0(s_k, p_k) \\ b_1(s_k, p_k) \end{pmatrix}$$

Zunächst $c_i := \dfrac{\partial b_{i-2}}{\partial p},\ i = 2, \ldots, n+2$:

$$\begin{aligned}
\frac{\partial b_n}{\partial p} &= 0, \qquad \frac{\partial b_{n-1}}{\partial p} = 0, \\
\frac{\partial b_{n-2}}{\partial p} &= \frac{\partial}{\partial p}\left(a_{n-2} + p\, b_n + s\, b_{n-1}\right) = b_n \\
\frac{\partial b_{n-3}}{\partial p} &= \frac{\partial}{\partial p}\left(a_{n-3} + p\, b_{n-1} + s\, b_{n-2}\right) \\
&= b_{n-1} + s \cdot \frac{\partial b_{n-2}}{\partial p} \\
c_{n-k+2} = \frac{\partial b_{n-k}}{\partial p} &= \frac{\partial}{\partial p}\left(a_{n-k} + p\, b_{n-k+2} + s\, b_{n-k+1}\right) \\
&= b_{n-k+2} + p \cdot \frac{\partial b_{n-k+2}}{\partial p} + s \cdot \frac{\partial b_{n-k+1}}{\partial p} \\
&= b_{n-k+2} + p\, c_{n-k+4} + s\, c_{n-k+3}, \qquad k = 4, \ldots, n-1 \\
c_2 = \frac{\partial b_0}{\partial p} &= p \cdot \frac{\partial b_2}{\partial p} + b_2 \\
&= b_2 + p\, c_4
\end{aligned}$$

	b_n	b_{n-1}	b_{n-2}	$\cdots$	b_3	b_2	b_1	b_0
p	$-$	$-$	$p\, c_n$	$\cdots$	$p\, c_5$	$p\, c_4$	$\cdots$	$\cdots$
s	$-$	$s\, c_n$	$s\, c_{n-1}$	$\cdots$	$s\, c_4$	$-$	$\cdots$	$\cdots$
	c_n	c_{n-1}	c_{n-2}	$\cdots$	c_3	c_2	$\cdots$	$\cdots$

$$\Rightarrow c_3 = \frac{\partial b_1}{\partial p}, \qquad c_2 = \frac{\partial b_0}{\partial p}$$

Entsprechend für die Ableitungen nach s:

$$\frac{\partial b_1}{\partial s} = c_2 + s\, c_3, \qquad \frac{\partial b_0}{\partial s} = p\, c_3$$

Daher die Jacobi-Matrix:

$$\mathrm{J}(s, p) = \begin{pmatrix} \frac{\partial b_0}{\partial s} & \frac{\partial b_0}{\partial p} \\ \frac{\partial b_1}{\partial s} & \frac{\partial b_1}{\partial p} \end{pmatrix} = \begin{pmatrix} p\, c_3 & c_2 \\ c_2 + s\, c_3 & c_3 \end{pmatrix}$$

Algorithmus 3.6.1 (von Bairstow).

Gegeben: *Koeffizientenvektor* $(a_0, \ldots, a_n)^\top \in \mathbb{R}^{n+1}$ *von* $p_n(x) = \sum_{k=0}^{n} a_k x^k$, *Näherung* ξ *an eine komplexe Nullstelle* x^* *von* p_n, $\varepsilon > 0$.

Gesucht: *Näherungen* $\xi_k, \overline{\xi}_k$ *an* $x^*, \overline{x}^*$ *mit* $|\xi_k - x^*| < \varepsilon$ $(\Rightarrow \left|\overline{\xi}_k - \overline{x}^*\right| < \varepsilon)$

Start: $s_0 := 2\,\mathrm{Re}\,\xi,\ p_0 := -|\xi|^2$,

Iteration: *Für* $k = 0, 1, 2, \ldots$:

$$\mathrm{J}(s_k, p_k) = \begin{pmatrix} p_k\, c_{3k} & c_{2k} \\ c_{2k} + s_k\, c_{3k} & c_{3k} \end{pmatrix}$$

$c_{2k} = c_2(s_k, p_k),\ c_{3k} = c_3(s_k, p_k)\ \ldots$

Sei $(\varepsilon_k, \delta_k)^\top$ *Lösung von* $\mathrm{J}(s_k, p_k) \begin{pmatrix} \varepsilon_k \\ \delta_k \end{pmatrix} = \begin{pmatrix} b_0(s_k, p_k) \\ b_1(s_k, p_k) \end{pmatrix}$

und $\left.\begin{array}{l} s_{k+1} = s_k + \varepsilon_k \\ p_{k+1} = p_k + \delta_k \end{array}\right\} \Rightarrow \xi_{k+1}, \overline{\xi}_{k+1} = \dfrac{s_{k+1}}{2} \pm i\sqrt{-p_{k+1} - \dfrac{s_{k+1}^2}{4}}$

Wenn $n \cdot |\xi_{k+1} - \xi_k| < \varepsilon$, *dann Abbruch.*

KAPITEL 4

Direkte Lösung von linearen Gleichungssystemen

$A\,x = b \qquad A : m \times n\text{-Matrix} \quad b : \text{ein } m\text{-Tupel}$

$$\begin{array}{ccccccc} a_{11}\,x_1 & + & \cdots & + & a_{1n}\,x_n - b_1 & = & 0 \\ a_{21}\,x_1 & + & \cdots & + & a_{2n}\,x_n - b_2 & = & 0 \\ \vdots & & & & & & \vdots \\ a_{m1}\,x_1 & + & \cdots & + & a_{mn}\,x_n - b_m & = & 0 \end{array}$$

Beispiel. $y''(x) + y(x) = f(x),\ x \in [0,1]$. Randbedingungen: $y(0) = y(1) = 0$. Numerische Lösung mit Diskretisierung von $[1,0]$.

$x_k = k \cdot h,\ 0 \le k \le N,\ h = \dfrac{1}{N}$

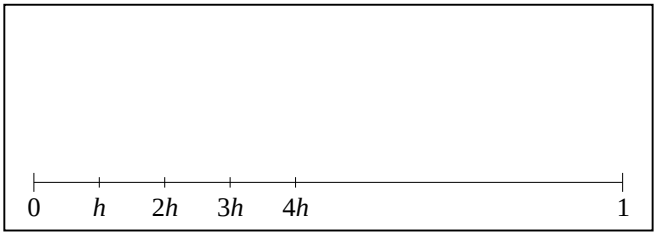


Abbildung 4.1: Einteilung

Ziel: Finde $y_1, \ldots, y_{N-1}$ so, daß $y_k \approx y(x_k),\ k = 1, \ldots, N-1\ y_0 = 0,\ y_N = 0$

$$y''(x) \approx \frac{y(x+h) - 2y(x) + y(x-h)}{h^2}$$

$y''(x_k) + y(x_k) = f(x_k),\ k = 1, \ldots, N-1$. Näherungsweise

$$\frac{y_{k+1} - 2x_k + y_{k-1}}{h^2} + y_k = f(x_k), \qquad k = 1, \ldots, N-1$$

$$\Rightarrow \begin{pmatrix} -2+h^2 & 1 & 0 & \cdot \\ 1 & -2+h^2 & 1 & \cdot \\ 0 & 1 & & \cdot \\ \cdot & & \ddots & 1 \\ \cdot & \ldots\ldots & 1 & -2+h^2 \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_{N-1} \end{pmatrix} = h^2 \begin{pmatrix} f(x_1) \\ \cdot \\ \cdot \\ \cdot \\ f(x_{N-1}) \end{pmatrix}$$

4.1 Das Gaußsche Eliminationsverfahren

Definition 4.1.1. Eine Matrix der Gestalt

$$P_{ik} = \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & 0 & & 1 & & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & 1 & & 0 & & & \\ & & & & & & & 1 & \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{pmatrix}$$

die Nullen sind in der i-ten und k-ten Zeile

heißt *Permutationsmatrix*.

Bemerkung.

$$P_{ik} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_{i-1} \\ x_k \\ x_{i+1} \\ \vdots \\ x_{k-1} \\ x_i \\ x_{k+1} \\ \vdots \\ x_n \end{pmatrix}$$ Vertauschung von i-ter und k-ter Komponente

$$(x_1, \ldots, x_n)\, P_{ik} = (x_1, \ldots, x_{i-1}, x_k, x_{i+1}, \ldots, x_{k-1}, x_i, x_{k+1}, \ldots, x_n)$$
$$P_{ik}\, P_{ik} = E \qquad \text{(Einheitsmatrix)}$$

Definition 4.1.2. Eine Matrix des Typs $L_i \in \mathbb{K}^{n\times n}$

$$L_i = \begin{pmatrix} 1 & & & & & 0 \\ 0 & 1 & & & & \\ \cdot & 0 & 1 & & & \\ \cdot & & l_{i+1,i} & 1 & & \\ \cdot & & \vdots & & \ddots & \\ 0 & \cdots & l_{n\,i} & \cdots\cdots & & 1 \end{pmatrix}$$

heißt *elementare untere Dreiecksmatrix*.

Bemerkung.

$$L_i^{-1} = \begin{pmatrix} 1 & & & & & 0 \\ 0 & 1 & & & & \\ \cdot & 0 & 1 & & & \\ \cdot & & -l_{i+1,i} & 1 & & \\ \cdot & & \vdots & & \ddots & \\ 0 & \cdots & -l_{n\,i} & \cdots\cdots & & 1 \end{pmatrix}$$

Satz 4.1.1. *Jede untere Dreiecksmatrix mit normierter Diagonale, aber*

$$L = \begin{pmatrix} 1 & & & 0 \\ l_{21} & 1 & & \\ \vdots & & \ddots & \\ l_{n1} & & & 1 \end{pmatrix}$$

läßt sich als Produkt von elementaren unteren Dreiecksmatrizen schreiben $L = L_1 L_2 \cdots L_{n-1}$ mit

$$L_i = \begin{pmatrix} 1 & & & & & 0 \\ 0 & 1 & & & & \\ \cdot & 0 & 1 & & & \\ \cdot & & l_{i+1,i} & 1 & & \\ \cdot & & \vdots & & \ddots & \\ 0 & \cdots & l_{ni} & \cdots\cdots & & 1 \end{pmatrix} \qquad i = 1, \ldots, n-1$$

Beweis.

$$L_1 \cdots L_k = \begin{pmatrix} 1 & & & & & 0 \\ l_{21} & \ddots & & & & \\ \cdot & & 1 & & & \\ \cdot & & l_{k+1,k} & 1 & & \\ \cdot & & \vdots & & \ddots & \\ l_{n1} & \cdots & l_{nk} & \cdots\cdots & & 1 \end{pmatrix}$$

Mit vollständiger Induktion: $k = 1$: Definition von L_1.
$k_{n-1} \to n$:

$$L_1 \cdots L_k = \begin{pmatrix} 1 & & & & & 0 \\ l_{21} & \ddots & & & & \\ \cdot & & 1 & & & \\ \cdot & & l_{k,k-1} & 1 & & \\ \cdot & & \vdots & & \ddots & \\ l_{n1} & \cdots & l_{nk-1} & 0 & \cdots & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & & & & & 0 \\ 0 & 1 & & & & \\ \cdot & 0 & 1 & & & \\ \cdot & & l_{k+1,k} & 1 & & \\ \cdot & & \vdots & & \ddots & \\ 0 & \cdots & l_{nk} & \cdots\cdots & & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & & & & & & 0 \\ l_{21} & \ddots & & & & & \\ \cdot & 0 & 1 & & & & \\ \cdot & & l_{k,k-1} & 1 & & & \\ \cdot & & \vdots & l_{k+1,k} & 1 & & \\ \cdot & & \cdot & \cdot & & \ddots & \\ l_{n1} & \cdots & l_{nk-1} & l_{nk} & \cdots & 0 & 1 \end{pmatrix}$$

□

Bemerkung. Untere Dreiecksmatrizen bilden eine Gruppe bezüglich der Multiplikation, erzeugt durch die elementaren unteren Dreiecksmatrizen.

Ziel der Gaußelimination:

$A\,x = b$ umformen zu $R\,x = \tilde{b}$

R : obere Dreiecksmatrix mit gleicher Lösungsmenge.

$$\begin{array}{ccccccc} a_{11}\,x_1 & + & \cdots & + & a_{1n}\,x_n & = & b_1 \\ a_{21}\,x_1 & + & \cdots & + & a_{2n}\,x_n & = & b_2 \\ \vdots & & & & & \vdots & \\ a_{m1}\,x_1 & + & \cdots & + & a_{mn}\,x_n & = & b_m \end{array}$$

Schritt 1: Eliminiere x_1; wenn ein $a_{i1} \neq 0$, dann o.B.d.A. $a_{11} \neq 0$.
Die Zeilenumformung gibt

$$\begin{array}{ccccccccc} a_{11}\,x_1 & + & a_{12}\,x_2 & + & \cdots & + & a_{1n}\,x_n & = & b_1 \\ 0 & + & a'_{22}\,x_2 & + & \cdots & + & a'_{2n}\,x_n & = & b'_2 \\ \vdots & & & & & & & \vdots & \\ 0 & + & a'_{m2}\,x_2 & + & \cdots & + & a'_{mn}\,x_n & = & b'_n \end{array}$$

$$a'_{ik} = a_{ik} - \frac{a_{i1}}{a_{11}}\, a_{1k} \qquad b'_i = b_i - \frac{a_{i1}}{a_{11}}\, b_1 \qquad i = 2, \ldots, m$$

$(A|b)$ sei die Matrix mit den n Spalten von A und b als $n+1$-te Spalte.
Formal:

Schritt 1: Bestimme den Index $p_1 \in \{1, \ldots, n\}$ mit $a_{p_1\,1} \neq 0$.

Schritt 2: Vertausche die erste Zeile von $(A|b)$ mit der Zeile p_1 ($\Rightarrow$ neue Matrix $(\hat{A}|\hat{b})$)

Schritt 3: Subtrahiere für $i = 2, \ldots, n$ das l_{i1}-fache, $l_{i1} = \frac{\hat{a}_{i1}}{\hat{a}_{11}}$, der ersten Zeile von der i-ten.
Resultat: Matrix $(A'|b')$

$(A|b) \to (\hat{A}, \hat{b}) \to (A'|b')$

$(\hat{A}|\hat{b}) = P_1(A|b)$ und $P_1 = P_{1p_1}$ Permutationsmatrix

$(A'|b') = L_1\, P_1(A|b)$

mit

$$L_1 = \begin{pmatrix} 1 & & & \\ -l_{21} & 1 & & \\ \vdots & & \ddots & \\ -l_{n1} & & & 1 \end{pmatrix}$$ elementare untere Dreiecksmatrix

Da P_1 und L_1 regulär sind, besitzen $A\,x = b$ und $A'x = b'$ die gleiche Lösungsgesamtheit.

$$\begin{aligned} A\,x = b \quad &\Rightarrow \quad L_1 P_1\, Ax = L_1 P_1\, b \Leftrightarrow A'x = b' \\ A'x = b' \quad &\Rightarrow \quad P_1 L_1^{-1} A'x = P_1 L_1^{-1} b' \Leftrightarrow A\,x = b \end{aligned}$$

Falls die erste Spalte von A null ist, dann ist $(A'|b') = L_1 P_1 (A|b)$ mit $P_1 = L_1 = E$.

Bemerkung. $a_{p_1 1} = \hat{a}_{11}$ heißt *Pivot-Element*. Entsprechend heißt **Schritt 1** *Pivot-Suche*, genauer: Spaltenpivot-Suche, denn das Pivot-Element wird in der ersten Spalte gesucht. Es reicht $\hat{a}_{11} \neq 0$. Numerisch besser:

$$\left|a_{p_1 1}\right| = \max_{i=1}^{m} |a_{i1}|$$

Numerisch ist es noch besser, die Zeilen am Anfang zu äquilibrieren:

$$\sum_{j=1}^{n} \left|a_{ij}\right| = 1 \qquad \text{für } i = 1, \ldots, m$$

(vergleiche später bei Fehlerabschätzung)
Nächster Eliminationsschritt mit $(\tilde{A}|\tilde{b})$

$$(A'|b') = \left(\begin{array}{c|c} a'_{11} & \cdots \\ 0 & \\ \vdots & \tilde{A} \\ 0 & \end{array} \begin{array}{c} b'_1 \\ \\ \tilde{b} \\ \\ \end{array}\right)$$

$$\left(\begin{array}{c|c|c} a'_{11} & \cdots & b'_1 \\ \hline 0 & & \\ \vdots & L_2 P_2 \tilde{A} & L_2 P_2 \tilde{b} \\ 0 & & \end{array}\right) = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \tilde{L}_2 & \\ 0 & & & \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \tilde{P}_2 & \\ 0 & & & \end{pmatrix} \cdot (A'|b')$$

$(A|b) \to (A'|b') \to (A^{(2)}|b^{(2)}) \to \cdots \to (A^{(s)}|b^{(s)})$

$A^{(s)}$ ist vom Typ:

$$r = \operatorname{Rg} A \left\{ \left(\begin{array}{ccccc|c} * & & & & & \\ & * & & & & \\ & & * & \cdots & & \\ & & & & * & \\ & & 0 & & & * \\ & & & & & \end{array} \right) \right.$$

Ab jetzt $n = m$

D.h. quadratische Matrizen und zusätzlich $\det(A) \neq 0$.
Dann treten beim Eliminationsverfahren keine Nullspalten auf (**Schritt 1** findet *immer* ein $a_{1p_1} \neq 0$), weil jede Matrix vollen Rang hat.
Platzsparende Notation:
Bei der Elimination von x_k aus Zeilen $j = k+1, \ldots, n$

$$a_j^{(k+1)} = a_{jl}^{(k)} - \frac{a_{jk}^{(k)}}{a_{kk}^{(k)}} a_{kl}^{(k)} \qquad l_{jk} := \frac{a_{jk}^{(k)}}{a_{kk}^{(k)}}$$

$$\begin{array}{|cc|c|ccc|c}
\hline
a_{11}^{(1)} & a_{12}^{(1)} & \cdots\cdots & a_{1n}^{(1)} & b_1^{(1)} \\
\hline
l_{21} & a_{22}^{(2)} & \cdots\cdots & a_{2n}^{(2)} & b_2^{(2)} \\
\cline{1-2}
l_{31} & l_{32} & & & b_3^{(3)} \\
\cline{1-3}
\vdots & \vdots & \vdots & \cdots\cdots & \vdots \\
l_{n1} & l_{n2} & l_{n3} & \cdots\cdots & b_n^{(n)} \\
\cline{1-4}
\end{array}$$

Lösung durch Rückwärtseinsetzen.

Beispiel.

$$\begin{array}{rcrcrcr}
-3x_1 & + & 5x_2 & - & 4x_3 & = & 3 \\
2x_1 & - & 6x_2 & + & 12x_3 & = & 2 \\
1x_1 & - & 2x_2 & + & 2x_3 & = & -1
\end{array}$$

$$\begin{array}{|ccc|c}
\hline
-3 & 5 & 4 & 3 \\
\hline
-\frac{2}{3} & -\frac{8}{3} & \frac{28}{3} & 4 \\
-\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} & 0 \\
\hline
\end{array} \quad \Rightarrow \quad \begin{array}{|ccc|c}
\hline
-3 & 5 & 4 & 3 \\
\hline
-\frac{2}{3} & -\frac{8}{3} & \frac{28}{3} & 4 \\
\cline{2-3}
-\frac{1}{3} & \frac{1}{8} & -\frac{1}{2} & -\frac{1}{2} \\
\hline
\end{array} \quad \Rightarrow x_3 = 1,\ x_2 = 2$$

4.2 Die LR-Zerlegung

Definition 4.2.1. Sei $A \in \mathbb{K}^{n\times n}$ und $L, R \in \mathbb{K}^{n\times n}$ mit

$$L = \begin{pmatrix} 1 & & 0 \\ \vdots & \ddots & \\ * & \cdots & 1 \end{pmatrix}, \qquad R = \begin{pmatrix} * & \cdots & * \\ & \ddots & \vdots \\ 0 & & * \end{pmatrix}$$

Gilt $A = L \cdot R$, so heißt diese Darstellung die *LR-Zerlegung* von A.

Satz 4.2.1. *Es sei $A \in \mathbb{K}^{n\times n}$ eine nicht-singuläre Matrix, die nach Durchführung der Gaußelimination auf die obere Dreiecksmatrix R überführt sein möge. Ferner sei*

$$P = P_{n-1} \cdot P_{n-2} \cdots P_2 \cdot P_1$$

das Produkt aller benötigter Permutationsmatrizen. Dann gilt:

$$P \cdot A = L \cdot R \quad \textit{mit} \quad L = \begin{pmatrix} 1 & & & 0 \\ l_{21} & 1 & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & \cdots & 1 \end{pmatrix}$$

Beweis. Zunächst die Annahme, daß keine Permutation notwendig war. Dann gilt:

$$A^{(1)} = A, \qquad A^{(k+1)} = L_k A^{(k)}$$

mit

$$L_k := \begin{pmatrix} 1 & & & & 0 \\ 0 & \ddots & & & \\ \cdot & & 1 & & \\ \cdot & & -l_{k+1\,k} & 1 & \\ \cdot & & \vdots & & \ddots \\ 0 & & -l_{n\,k} & & 1 \end{pmatrix} \quad \text{und} \quad A^{(n)} = R$$

Also

$$\begin{aligned} R = A^{(n)} &= L_{n-1} A^{(n-1)} = L_{n-1} L_{n-2} A^{(n-2)} \qquad \tilde{b} = L_{n-1} \cdots L_1 b \\ &= \cdots = L_{n-1} L_{n-2} \cdots L_1 A \\ \Rightarrow A &= \underbrace{L_1^{-1} \cdots L_{n-2}^{-1} L_{n-1}^{-1}}_{=:L=\begin{pmatrix} 1 & & 0 \\ l_{21} & \ddots & \\ \vdots & \ddots & \ddots \\ l_{n1} & \cdots & l_{nn-1} \; 1 \end{pmatrix}} \cdot R \qquad \text{Nach Satz (4.1.1)} \end{aligned}$$

Bei $PA = P_{n-1} \cdot P_{n-2} \cdots P_2 \cdot P_1 \cdot A$ stehen alle Pivots an richtiger Stelle, d.h. keine Permutation bei PA mehr erforderlich. □

Am Ende der Gaußelimination (ohne Permutation) steht

$$\begin{array}{c|c|c|c} a_{11}^{(1)} & a_{12}^{(1)} & \cdots\cdots & a_{1n}^{(1)} \\ \hline l_{21} & a_{22}^{(2)} & \cdots\cdots & a_{2n}^{(2)} \\ \hline l_{31} & l_{32} & & \\ \hline \vdots & \vdots & \vdots & \ddots \\ \hline l_{n1} & l_{n2} & l_{n3} & \cdots \; a_{nn}^{(n)} \end{array}$$

$$R = \begin{pmatrix} a_{11}^{(1)} & \cdots & a_{1n}^{(1)} \\ & \ddots & \vdots \\ 0 & & a_{nn}^{(n)} \end{pmatrix} \qquad L = \begin{pmatrix} 1 & & & 0 \\ l_{21} & \ddots & & \\ \vdots & \ddots & \ddots & \\ l_{n1} & \cdots & l_{n\,n-1} & 1 \end{pmatrix}$$

Bemerkung.

$$\det(A) = (-1)^k \det(R) = (-1)^k \prod_{j=1}^{n} a_{jj}^{(j)}$$

wobei k die Anzahl der erforderlichen Zeilenvertauschungen ist.

$$\left[\underbrace{\det(P)}_{=(-1)^k} \det(A) = \underbrace{\det(L)}_{=1} \underbrace{\det(R)}_{\prod a_{jj}^{(j)}}, \qquad \det(P_{ij}) = -1 \right]$$

Satz 4.2.2. *Eine reguläre Matrix $A \in \mathbb{K}^{n\times n}$ besitzt genau dann eine LR-Zerlegung, wenn alle Hauptminoren von Null verschieden sind, also* $\det A_k \neq 0$ *für $k = 1 \ldots, n$ mit*

$$A_k = \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix}$$

BEWEIS. Ist die Gaußelimination ohne Permutation durchführbar, dann ist es auch die Gaußelimination für A_k. also

$$\det A_k = a_{11}^{(1)} a_{22}^{(2)} \cdots a_{kk}^{(k)} \neq 0$$

Wenn alle Hauptminoren von Null verschieden sind, dann

$$a_{kk}^{(k)} = \frac{a_{11}^{(1)} a_{22}^{(2)} \cdots a_{kk}^{(k)}}{a_{11}^{(1)} a_{22}^{(2)} \cdots a_{k-1\,k-1}^{(k-1)}} = \frac{\det A_k}{\det A_{k-1}} \neq 0$$

□

Berechnung der *LR*-Zerlegung nach Crout

Für $i = 1, \dots, n$

$$r_{ik} := a_{ik} - \sum_{j=1}^{i-1} l_{ij} r_{jk}, \qquad k = i, i+1, \dots, n$$

$$l_{ki} := \frac{1}{r_{ki}} \left[a_{ki} - \sum_{j=1}^{i-1} l_{kj} r_{ji} \right]_{k=i+1,\dots,n}$$

Satz 4.2.3. *Wenn für $A \in \mathbb{K}^{n\times n}$ mit $\boxed{\det A \neq 0}$ eine LR-Zerlegung existiert, dann ist sie eindeutig.*

Beweis. Seien $A = L_1 \cdot R_1$, $A = L_2 \cdot R_2$ zwei *LR*-Zerlegungen.

$$L_1 R_1 = L_2 R_2 \quad \Rightarrow \quad \underbrace{L_2^{-1} L_1}_{\begin{pmatrix} 1 & & 0 \\ \vdots & \ddots & \\ * & \cdots & 1 \end{pmatrix}} = \underbrace{R_2 R_1^{-1}}_{\begin{pmatrix} * & \cdots & * \\ & \ddots & \vdots \\ 0 & & * \end{pmatrix}} = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} = E$$

$\Rightarrow R_2 = R_1, \quad L_1 = L_2.$ □

Bemerkung. Das Gleichungssystem $Ax = b$ ist leicht lösbar wenn $A = L \cdot R$ oder $PA = L \cdot R$ bekannt ist.

$$PA\,x = Pb \qquad L(Rx) = Pb$$

Schritt 1: Löse $Ly = Pb$ $\quad \downarrow \left(\begin{array}{ccc} 1 & & 0 \\ \vdots & \ddots & \\ * & \cdots & 1 \end{array} \right| = \Big|$ Vorwärts-Einsetzen

Schritt 2: Löse $Rx = y$ $\quad \uparrow \left(\begin{array}{ccc} * & \cdots & * \\ & \ddots & \vdots \\ 0 & & * \end{array} \right| = \Big|$ Rückwärts-Einsetzen

Das Gaußeliminationsverfahren **ist** die *LR*-Zerlegung und Vorwärts- und Rückwärts-Einsetzen mit Vertauschung der Reihenfolge der Rechenoperationen.

$$[y = \tilde{b}, \quad b \text{ nach } (n-1) \text{ Eliminationsschritten}]$$

Komplexität: (Anzahl der Multiplikationen/Divisionen und der Additionen)
LR Berechnung von

$$l_{jk} = \frac{a_{jk}^{(k)}}{a_{kk}^{(k)}}, \qquad j = k+1, \dots, n \qquad \boxed{n-k \text{ Divis.}}$$

$$a_{jk}^{(k+1)} = a_{jl}^{(k)} - l_{jl} a_{kl}^{(k)}, \qquad j, l = k+1, \dots, n \qquad \boxed{(n-k)^2 \text{ Multipl., } k \text{ Additionen}}$$

Berechnung von $A^{(k+1)}$ erfordert daher

$$(n-k) + (n-k)^2 = (n-k)(n-k+1) \text{ Multipl./Divis. und } (n-k)^2 \text{ Additionen}$$

Insgesamt:

$$\begin{aligned}
\sum_{k=1}^{n-1}(n-k)(n-k+1) &\overset{j=n-k}{=} \sum_{j=1}^{n-1} j(j+1) = \sum_{j=1}^{n-1} j^2 + \sum_{j=1}^{n-1} j \\
&= \frac{n(n-1)(2n-1)}{6} + \frac{n(n-1)}{2} = \frac{n(n-1)}{2}\left[\frac{2n-1}{3}+1\right] \\
&= \frac{1}{3}n(n-1)(n+1) = \frac{1}{3}n^3 + O(n^2) \text{ Multipl./Divis.} \\
\sum_{k=1}^{n-1}(n-k)(n-k) &= \sum_{j=1}^{n-1} j^2 = \frac{n(n-1)(2n-1)}{6} \\
&= \frac{1}{3}n^3 + O(n^2) \text{ Additionen}
\end{aligned}$$

Rückwärteinsetzen: $(Rx = y)$

$$x_k = \frac{1}{r_{kk}}\Big(y_k - \sum_{l=k+1}^{n} r_{kl}x_l\Big), \qquad k = n, n-1, \ldots, 2, 1$$

Eine Division und $n-k$ Multiplikationen und Additionen.

Vorwärtseinsetzen: $(Ly = b)$

$$y_k = b_k - \sum_{j=1}^{k-1} l_{kj}y_j \qquad k = 1, 2, \ldots, n$$

$k-1$ Multiplikationen und Additionen.
Hier (Rückwärts- und Vorwärts-Einsetzen):

$$\begin{aligned}
\sum_{k=1}^{n}(n-k+1) + \sum_{k=1}^{n}(k-1) &= \sum_{j=1}^{n-1}(j+1) + \sum_{k=0}^{n-1} k \\
&= \frac{n(n+1)}{2} + \frac{n(n-1)}{2} = n^2 \text{ Multipl./Divis. und } n^2 - n \text{ Additionen}
\end{aligned}$$

Definition 4.2.2. Eine Matrix $A \in \mathbb{K}^{n\times n}$ hat die *obere Bandbreite k* und *untere Bandbreite m*, wenn ihre Elemente a_{ij} für die Indizes i, j mit $j - i \geq k$ und mit $i - j \geq m$ null sind.

$$A = \begin{pmatrix} \overbrace{* * *}^{k} & * & & & \\ \cdot * * & * * & & & \\ \cdot \quad * & * * & * & & \\ * \quad\quad & * \quad * & * * & & \\ \quad * \quad\quad & * \quad * & * * * & & \\ \quad\quad * \quad & * \quad \cdot & * * * & & \\ \quad\quad\quad & * \quad \cdot & * \quad * & & \\ \quad\quad\quad & \underbrace{* \quad\quad *}_{m} & & & \end{pmatrix}$$

Satz 4.2.4. *$A \in \mathbb{K}^{n\times n}$ habe die Zerlegung $A = LR$. Hat A die obere Bandbreite m, dann gilt dies auch für L und R. (L untere Bandbreite m, R obere Bandbreite k). Z.B.:*

$$\begin{pmatrix} * * & & & & \\ * * * & & & & \\ \quad * * * & & & & \\ \quad\quad * * * & & & & \\ \quad\quad\quad * * * & & & & \\ \quad\quad\quad\quad * * & & & & \end{pmatrix} = \begin{pmatrix} * & & & & \\ * * & & & & \\ \quad * * & & & & \\ \quad\quad * * & & & & \\ \quad\quad\quad * * & & & & \\ \quad\quad\quad\quad * * & & & & \end{pmatrix} \cdot \begin{pmatrix} * * & & & & \\ \quad * * & & & & \\ \quad\quad * * & & & & \\ \quad\quad\quad * * & & & & \\ \quad\quad\quad\quad * * & & & & \\ \quad\quad\quad\quad\quad * & & & & \end{pmatrix}$$

4.3 Die Cholesky-Zerlegung

Definition 4.3.1. Eine symmetrische Matrix $A \in \mathbb{K}^{n\times n}$ mit der Eigenschaft

$$x^\top A\, x > 0 \text{ für } x \neq 0$$

heißt *positiv definite Matrix*.

Satz 4.3.1. *Sei $A \in \mathbb{R}^{n\times n}$ symmetrische Matrix. Dann sind äquivalent:*

i) A ist positiv definit

ii) Alle Eigenwerte von A sind positiv

ist A positiv definit, dann gilt $\det A > 0$.

Beweis. Bekannt aus der Linearen Algebra: $\exists$ orthogonale Matrix U, $(U^\top U = E)$

$$U^\top \cdot A \cdot U = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$$

Aus i) folgt: (u_i ist die i-te Spalte von U, $u_i \neq 0$)

$$u_i^\top \cdot A\, u_i = \lambda_i \qquad i = 1, \ldots, n$$

Nun ist λ_i Eigenwert und es gilt:

$$A\, U = U \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

$\Rightarrow A\, u_i = \lambda_i u_i \Rightarrow$ ii)
ii) $y := U^\top x$

$$\begin{aligned} x^\top \cdot A \cdot x &= (x^\top \cdot UU^\top \cdot A \cdot UU^\top \cdot x) \\ &= y^\top U^\top A\, Uy = y \cdot \begin{pmatrix} \lambda_1 y_1 \\ \vdots \\ \lambda_n y_n \end{pmatrix} \\ &= \sum_{i=1}^{n} \lambda_i y_i^2 \geq 0 \end{aligned}$$

Der Fall $x^\top \cdot A \cdot x = 0$ heißt $y = 0 \Rightarrow x = U\, y = 0$. Charakteristisches Polynom:

$$\begin{aligned} \chi(\lambda) = \det(A - \lambda E) &= (-1)^n \cdot \lambda^n + (-1)^{n-1}\Big(\sum_{i=1}^{n} a_{ii}\Big)\lambda^{n-1} + \cdots + \det A \\ &= (-1)^n(\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_n) \end{aligned}$$

$\Rightarrow \det A = \lambda_1 \cdot \lambda_2 \cdots \lambda_n > 0$ □

Beispiel.

$$A = \begin{pmatrix} 2 & -1 & & & & 0 \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & -1 \\ 0 & & & & -1 & 2 \end{pmatrix}$$

A ist reell-symmetrisch. Gerschgorin: Alle Eigenwerte liegen in

$$\{z \in \mathbb{C} \mid |z - 2| \leq 2\}$$

A reell-symmetrisch:

$\Rightarrow$ Alle Eigenwerte sind reell

$\Rightarrow$ Alle Eigenwerte in $[0, 4]$

Null ist ein Eigenwert von $A \Leftrightarrow \det A = 0$.

$$\begin{pmatrix} 2 & -1 & & & & 0 \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & -1 \\ 0 & & & & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix} = 0$$

Sei $x_1 \neq 0$. Dann o.B.d.A. $x_1 = 1$.
Behauptung. $x_k = k$.

BEWEIS. $k = 1$ ✓
$k-1$-te Zeile:

$$-x_{k-2} + 2x_{k-1} - x_k = 0$$
$$\underbrace{-(k-2) + 2(k-1)}_{k} = x_k$$

Letzte Zeile:

$$-x_{k-1} + 2x_n = 0$$
$$\text{aber} \quad -(n-1) + 2n = n + 1 \neq 0 \text{ also } x_1 = 0$$

Wenn $x_1 = \cdots = x_{i-1} = 0, \quad x_i \neq 0$

$$\begin{pmatrix} 2 & -1 & & & & 0 \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & -1 \\ 0 & & & & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix} = 0$$

O.B.d.A. $x_i = 1$, Widerspruch wie eben. Also $Ax = 0 \Rightarrow x = 0$. Also $\det A \neq 0$. Also sind alle Eigenwerte von $A > 0 \Rightarrow A$ positiv definit. □

Satz 4.3.2. *Sei A positiv definit. Dann*

i) alle Hauptuntermatrizen A_k sind positiv definit

ii) $a_{ii} \cdot a_{jj} > a_{ij}^2$ für $1 \leq i < j \leq n$

iii) $a_{ii} > 0$ für $i = 1, \ldots, n$

BEWEIS. Wird dem Leser als Übung überlassen □

Folgerung. $\det A_k > 0$ *für $k = 1, \ldots, n$. Nach Satz (4.2.2) hat A eine LR-Zerlegung $A = L \cdot R$.*

$$R = \underbrace{\begin{pmatrix} r_{11} & & 0 \\ \vdots & \ddots & \\ 0 & \cdots & r_{nn} \end{pmatrix}}_{D} \cdot \underbrace{\begin{pmatrix} 1 & \tilde{r}_{12} & \cdots & \tilde{r}_{1n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \tilde{r}_{n-1\,n} \\ 0 & \cdots\cdots & & 1 \end{pmatrix}}_{\tilde{R}} \qquad \tilde{r}_{ij} = \frac{r_{ij}}{r_{ii}}$$

$$r_{ii} = \frac{\det A_i}{\det A_{i-1}} > 0$$

$$A = L \cdot D \cdot \tilde{R} \Rightarrow A^\top = \tilde{R}^\top \cdot D \cdot L^\top = A$$

$\tilde{R}^\top$: normierte untere Dreiecksmatrix,
$D \cdot L^\top$: obere Dreiecksmatrix

Nach Satz (4.2.3) ist $L = \tilde{R}^\top$, $R = D \cdot L^\top$ und $A = L \cdot D \cdot L^\top$

$$D = \operatorname{diag}(\sqrt{r_{11}}, \ldots, \sqrt{r_{nn}}) \cdot \operatorname{diag}(\sqrt{r_{11}}, \ldots, \sqrt{r_{nn}})$$

Satz 4.3.3 (Cholesky-Zerlegung). *Jede positiv definite Matrix $A \in \mathbb{R}^{n\times n}$ kann eindeutig in die Form*

$$A = G \cdot G^\top$$

zerlegt werden mit einer unteren Dreiecksmatrix G mit positiven Diagonalelementen.

Beweis.

$$\begin{aligned} & G := L \cdot \operatorname{diag}(\sqrt{r_{11}}, \ldots, \sqrt{r_{nn}}) \\ \Rightarrow\ & G\,G^\top = L \cdot \operatorname{diag}(\sqrt{r_{11}}, \ldots, \sqrt{r_{nn}}) \cdot \operatorname{diag}(\sqrt{r_{11}}, \ldots, \sqrt{r_{nn}}) \cdot L^\top \\ & = A \end{aligned}$$

Damit ist die Existenz gesichert.
Eindeutigkeit:

$$\begin{aligned} & A = \hat{G}\,\hat{G}^\top,\ \hat{G} \text{ untere Driecksmatrix },\ \hat{g}_{kk} > 0, \quad k = 1, \ldots, n \\ & \hat{L} := \hat{G} \cdot \operatorname{diag}(\frac{1}{\hat{g}_{11}}, \ldots, \frac{1}{\hat{g}_{nn}}) \text{ ist normierte untere Dreiecksmatrix} \\ & \hat{R} := \operatorname{diag}(\hat{g}_{11}, \ldots, \hat{g}_{nn}) \cdot \hat{G}^\top \\ \Rightarrow\ & \hat{L}\hat{R} = \hat{G} \cdot \operatorname{diag}(1, \ldots, 1) \cdot \hat{G}^\top = A \\ \Rightarrow\ & L = \hat{L} \qquad R = \hat{R} \\ & \hat{G} = L \cdot \operatorname{diag}(\hat{g}_{11}, \ldots, \hat{g}_{nn}) \Rightarrow \hat{G}^\top = \operatorname{diag}(\hat{g}_{11}, \ldots, \hat{g}_{nn}) \cdot L^\top \\ & \hat{G}^\top = \operatorname{diag}(\frac{1}{\hat{g}_{11}}, \ldots, \frac{1}{\hat{g}_{nn}}) \cdot R \\ & \quad = \operatorname{diag}(\frac{1}{\hat{g}_{11}}, \ldots, \frac{1}{\hat{g}_{nn}}) \cdot \operatorname{diag}(r_{11}, \ldots, r_{nn}) \cdot L^\top \\ \Rightarrow\ & \operatorname{diag}(\hat{g}_{11}, \ldots, \hat{g}_{nn}) = \operatorname{diag}(\frac{1}{\hat{g}_{11}}, \ldots, \frac{1}{\hat{g}_{nn}}) \cdot \operatorname{diag}(r_{11}, \ldots, r_{nn}) \\ & \quad = \operatorname{diag}(\frac{r_{11}}{\hat{g}_{11}}, \ldots, \frac{r_{nn}}{\hat{g}_{nn}}) \\ \Rightarrow\ & \hat{g}_{kk}^2 = r_{kk}, \qquad k = 1, \ldots, n \\ \Rightarrow\ & \hat{g}_{kk} = \sqrt{r_{kk}}, \qquad k = 1, \ldots, n \\ \Rightarrow\ & \hat{G} = L \cdot \operatorname{diag}(\sqrt{r_{11}}, \ldots, \sqrt{r_{nn}}) = G \end{aligned}$$

□

Bemerkung. $A = G \cdot G^\top$

$$\Rightarrow x^\top A x = x^\top G G^\top x = (G^\top x)^\top \cdot (G^\top x) \geq 0$$

Also, wenn $\det G \neq 0$, dann

$$G^\top x = 0 \Rightarrow x = 0$$

und A ist positiv definit.

Bemerkung. Aus $A = G \cdot G^\top$ folgt

$$\begin{aligned} & a_{ki} = a_{ik} = \sum_{l=1}^{n} g_{il} \cdot g_{kl}, \qquad 1 \leq i \leq k \leq n \\ & g_{il} = 0 \text{ für } l > i \\ \Rightarrow\ & a_{ik} = \sum_{l=1}^{i} g_{il} \cdot g_{kl} = \sum_{l=1}^{i-1} g_{il} \cdot g_{kl} + g_{ii} \cdot g_{ki} \\ & a_{ii} = g_{i1}^2 + \cdots + g_{ii}^2 \end{aligned}$$

$\Rightarrow$ hieraus die g_{ii} berechnen, $\Rightarrow g_{ki}$ berechnen

Verfahren von Cholesky-Crout

Für $i = 1, \ldots, n$

- berechne $g_{ii} = \sqrt{a_{ii} - \sum_{l=1}^{i-1} g_{il}^2}$
- Für $k = i+1, \ldots, n$

$$g_{ki} = \frac{1}{g_{ii}}\Big(a_{ik} - \sum_{l=1}^{i-1} g_{il} \cdot g_{kl}\Big)$$

Satz 4.3.4. *Bei der Berechnung der Cholesky-Zerlegung einer positiv definiten $n \times n$-Matrix sind*

$$\frac{n(n-1)(n+4)}{6} \quad \textit{Multiplikationen/Divisionen}$$
$$\frac{n(n-1)(n+1)}{6} \quad \textit{Additionen}$$

und n Wurzeln erforderlich. Die Zerlegung $A = L \cdot D \cdot L^\top$ heißt rationale *Cholesky-Zerlegung.*

4.4 Das Gauß-Jordan-Verfahren

$$\begin{array}{lllllll} a_{11}\,x_1 & + & \cdots & + & a_{1n}\,x_n - & = & y_1 \\ a_{21}\,x_1 & + & \cdots & + & a_{2n}\,x_n - & = & y_2 \\ \vdots & & & & & & \vdots \\ a_{n1}\,x_1 & + & \cdots & + & a_{nn}\,x_n - & = & y_n \end{array}$$

Vertausche x_1 und y_1, $a_{11} \neq 0$ Pivot.

$$\begin{aligned} \frac{1}{a_{11}} y_1 + \frac{-a_{12}}{a_{11}} x_2 + \cdots + \frac{-a_{1n}}{a_{11}} x_n &= x_1 \\ \frac{a_{21}}{a_{11}} y_1 + \left(a_{22} - \frac{a_{21}a_{12}}{a_{11}}\right) x_2 + \cdots + \left(a_{2n} - \frac{a_{21}a_{1n}}{a_{11}}\right) x_n &= y_2 \\ \cdots & \end{aligned}$$

$A = A^{(1)}$

$$A^{(1)} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \to A^{(2)} \begin{pmatrix} y_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \to A^{(3)} \begin{pmatrix} y_1 \\ y_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix}$$

$$\to A^{(n+1)} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \qquad \text{gilt für beliebige } (x_1, \ldots, x_n)^\top, (y_1, \ldots, y_n)^\top \in \mathbb{K}^n$$

$$A \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$\Rightarrow A^{(n+1)} = A^{-1}$$

Vertauschen von x_r und y_s:

$$y_r = a_{rs}x_s + \sum_{k=1}^{s-1} a_{rk}x_k + \sum_{k=s+1}^{n} a_{rk}x_k$$

$$\Rightarrow x_s = \sum_{k=1}^{s-1} \frac{-a_{rk}}{a_{rs}} x_k + \frac{1}{a_{rs}} y_r + \sum_{k=s+1}^{n} \frac{-a_{rk}}{a_{rs}} x_k$$

für $i \neq r$

$$y_i = \sum_{k=1}^{s-1} \Big(a_{ik} - \frac{a_{rk}}{a_{rs}} a_{is}\Big) x_k + \frac{a_{is}}{a_{rs}} y_k + \sum_{k=s+1}^{n} \Big(a_{ik} - \frac{a_{rk}a_{is}}{a_{rs}}\Big) x_k$$

Algorithmus (von Gauß-Jordan).

Gegeben: $A = A^{(1)} \in \mathbb{K}^{n\times n}$, *invertierbar*

Gesucht: A^{-1}

Für $l = 1, \ldots, n$ *bilde* $A^{(l)}$ *nach folgenden Regeln:*

1) Pivot-Suche

Bestimme den Zeilenindex $p_l \in \{l+1, \ldots, n\}$ *so, daß*

$$\left|a^{(l)}_{p_l\, l+1}\right| = \max_{l<k\le n} \left|a^{(l)}_{k\, l+1}\right|$$

2) Vertausche $l+1$*-te mit* p_l*-ter Zeile von* $A^{(l)}$*. Dies gibt eine Matrix* $\tilde{A}^{(l)}$ *und die Komponenten* $\tilde{a}_{ij}$*.*

3) Austauschschritt

a) Spaltenregel

$$a^{(l+1)}_{i\, l+1} := \frac{\tilde{a}^{(l)}_{i\, l+1}}{\tilde{a}^{(l)}_{l+1\, l+1}} \qquad i = 1, \ldots, n \quad i \neq l+1$$

b) Zeilenregel

$$a^{(l+1)}_{l+1\, k} := \frac{-\tilde{a}^{(l)}_{l+1\, k}}{\tilde{a}^{(l)}_{l+1\, l+1}} \qquad k = 1, \ldots, n \quad k \neq l+1$$

c) Pivotregel

$$a^{(l+1)}_{l+1\, l+1} := \frac{1}{\tilde{a}^{(l)}_{l+1\, l+1}}$$

d) Rechteckregel

$$a^{(l+1)}_{i\, k} = \tilde{a}^{(l)}_{ik} - \frac{\tilde{a}^{(l)}_{i\, l+1} \cdot \tilde{a}^{(l)}_{l+1\, k}}{\tilde{a}^{(l)}_{l+1\, l+1}} \qquad i, k = 1, \ldots, n \quad i \neq l+1 \neq k$$

$$A^{-1} = A^{(n+1)} \cdot P_{p_{n+1}\, n+1} \cdot P_{p_1\, 1}$$

Pro Austauschschritt gibt es n^2 Multiplikationen/Divisionen sowie $(n-1)^2$ Additionen. Man vertausche immer solange ein x_i gegen ein y_j, bis oben nur noch y_j-Werte stehen:

y_{ν_1}	$\cdots$	y_{ν_n}	
	$\tilde{A}$		x_{μ_1} $\vdots$ x_{μ_n}

$\xrightarrow[\text{Zeilentausch}]{\text{Spalten- und}}$

y_1	$\cdots$	y_n	
	A^{-1}		x_1 $\vdots$ x_n

Darstellung: Gauß-Jordan-Tabellen

Beispiel.

$$A = \begin{pmatrix} -3 & 5 & -4 \\ 2 & -6 & 12 \\ 1 & -2 & 2 \end{pmatrix}$$

x_1	x_2	x_3	
-3	5	-4	y_1
2	-6	12	y_2
1	-2	2	y_3

Rechteckregel $-3 - \dfrac{2\cdot(-4)}{12} = -\dfrac{7}{3}$

x_1	x_2	y_2	
$-\frac{7}{3}$	3	$-\frac{1}{3}$	y_1
$-\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{12}$	x_3
$\frac{2}{3}$	-1	$\frac{1}{6}$	y_3

$\longrightarrow$

x_1	y_1	y_2	
$\frac{7}{9}$	$\frac{1}{3}$	$\frac{1}{9}$	x_2
$\frac{2}{9}$	$\frac{1}{6}$	$\frac{5}{36}$	x_3
$-\frac{1}{9}$	$-\frac{1}{3}$	$\frac{1}{18}$	y_3

$\longrightarrow$

y_3	y_1	y_2	
-7	-2	$\frac{1}{2}$	x_2
-2	$-\frac{1}{2}$	$\frac{1}{4}$	x_3
-9	-3	$\frac{1}{2}$	x_1

$$\begin{array}{cccc|c} y_1 & y_2 & y_3 & & \\ \hline -2 & \frac{1}{2} & -7 & x_2 \\ -\frac{1}{2} & \frac{1}{4} & -2 & x_3 \\ -3 & \frac{1}{2} & -9 & x_1 \end{array} \longrightarrow \begin{array}{cccc|c} y_1 & y_2 & y_3 & & \\ \hline -3 & \frac{1}{2} & -9 & x_1 \\ -2 & \frac{1}{2} & -7 & x_2 \\ -\frac{1}{2} & \frac{1}{4} & -2 & x_3 \end{array} \Longrightarrow A^{-1} = \begin{pmatrix} -3 & \frac{1}{2} & -9 \\ -2 & \frac{1}{2} & -7 \\ -\frac{1}{2} & \frac{1}{4} & -2 \end{pmatrix}$$

Gesamtaufwand von Gauß-Jordan:

n^3 Multipl./Divis.

$n(n-1)^2$ Additionen

Bemerkung. Gauß-Jordan ist nur anwendbar, wenn wirklich A^{-1} erforderlich ist. Im allgemeinen reicht die *LR*-Zerlegung.

4.5 Matrizennormen

Definition 4.5.1. Eine Abbildung $N : \mathbb{C}^{n\times n} \to \mathbb{R}_+ = [0, \infty)$ heißt *Matrixnorm*, wenn

i) $N(A) = 0 \leftrightarrow A = 0 \in \mathbb{C}^{n\times n}$

ii) $\lambda \in \mathbb{C},\ A \in \mathbb{C}^{n\times n} \Rightarrow N(\lambda A) = |\lambda|\ N(A)$

iii) $A, B \in \mathbb{C}^{n\times n} \Rightarrow N(A + B) \leq N(A) + N(B)$

iv) $A, B \in \mathbb{C}^{n\times n} \Rightarrow N(A \cdot B) \leq N(A) \cdot N(B)$

gelten.

Beispiel. i) Zeilensummennorm:

$$A \mapsto \|A\|_\infty = \max_{i=1}^{n} \sum_{k=1}^{n} |a_{ik}|$$

ii) Spaltensummennorm:

$$A \mapsto \|A\|_1 = \max_{k=1}^{n} \sum_{i=1}^{n} |a_{ik}|$$

iii) Frobeniusnorm:

$$A \mapsto \|A\|_F = \sqrt{\sum_{i=1}^{n} \sum_{k=1}^{n} |a_{ik}|^2}$$

Satz 4.5.1. *Ist* $\|\ \ \| : \mathbb{C}^n \to \mathbb{R}_+$ *eine Vektornorm, dann erhält man durch*

$$\mathrm{lub}(A) := \max_{x\neq 0} \frac{\|A\,x\|}{\|x\|} = \max_{\|x\|=1} \|A\,x\|$$

eine Matrixnorm auf $\mathbb{C}^{n\times n}$. *Sie heißt die zu* $\|\ \ \|$ *gehörende* lub-*Norm (***l***east* **u***pper* **b***ound).*

Bemerkung. $\|A\,x\| \leq \mathrm{lub}(A)\ \|x\|$ wenn

$$\|A\,x\| \leq K(A)\ \|x\| \text{ dann } \mathrm{lub}(A) \leq K(A)$$

BEWEIS. (lub ist Matrixnorm)

i)

$$\begin{aligned} \mathrm{lub}(A) = 0 &\Rightarrow \|A\,x\| = 0\ \forall\, x \\ &\Rightarrow A\,x = 0\ \forall\, x \Rightarrow A = 0 \\ &\Rightarrow \mathrm{lub}(A) = 0 \end{aligned}$$

ii) $\operatorname{lub}(\lambda A) = \max\limits_{x\neq 0} \dfrac{\|\lambda\, A\, x\|}{\|x\|} = \max\limits_{x\neq 0} \dfrac{|\lambda|\,\|A\, x\|}{\|x\|} = \lambda \operatorname{lub}(A)$

iii)

$$\begin{aligned}\operatorname{lub}(A+B) &= \max_{x\neq 0} \frac{\|(A+B)\,x\|}{\|x\|} \leq \max_{x\neq 0} \frac{\|A\,x\| + \|B\,x\|}{\|x\|} \\ &\leq \max_{x\neq 0} \frac{\|A\,x\|}{\|x\|} + \max_{x\neq 0} \frac{\|B\,x\|}{\|x\|} = \operatorname{lub}(A) + \operatorname{lub}(B)\end{aligned}$$

iv)

$$\begin{aligned}\operatorname{lub}(A\cdot B) &= \max_{x\neq 0} \frac{\|AB\,x\|}{\|x\|} = \max_{Bx\neq 0} \frac{\|A\,(Bx)\|}{\|Bx\|} \cdot \frac{\|B\,x\|}{\|x\|} \\ &\leq \max_{Bx=y\neq 0} \frac{\|A\,y\|}{\|y\|} \cdot \max_{x\neq 0} \frac{\|B\,x\|}{\|x\|} = \operatorname{lub}(A)\cdot\operatorname{lub}(B)\end{aligned}$$

$B = 0 \Rightarrow A\,B = 0 \Rightarrow \operatorname{lub}(A\,B) = 0 = \operatorname{lub}(A)\cdot\operatorname{lub}(B)$

□

Definition 4.5.2. Seien $A \in \mathbb{C}^{n\times n}$ und $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ die Eigenwerte von A. Dann ist

$$\varrho(A) := \max_{i=1}^{n} |\lambda_i|$$

der *Spektralradius* von A.

Satz 4.5.2. $\operatorname{lub}_2(A) = \sqrt{\varrho(A^{\mathsf{H}}A)}$

Bemerkung. $\|A\|_2 = \operatorname{lub}_2(A)$ heißt *Spektralnorm*.

$$\begin{aligned} A = (a_{ik}) \quad &\Rightarrow \quad A^{\mathsf{H}} = (\overline{a}_{ki}) \qquad A^{\mathsf{H}} = \overline{A^{\top}} \\ (A^{\mathsf{H}}A)^{\mathsf{H}} \quad &= \quad A^{\mathsf{H}}(A^{\mathsf{H}})^{\mathsf{H}} = A^{\mathsf{H}}A \end{aligned}$$

$\Rightarrow$ alle Eigenwerte sind reell.
Wegen $x^{\mathsf{H}}(A^{\mathsf{H}}A)\,x = (Ax)^{\mathsf{H}}\,Ax \geq 0$ sind alle Eigenwerte ≥ 0.

Beweis. $A^{\mathsf{H}}A$ sei hermitesch. Also existiert ein $U \in \mathbb{C}^{n\times n}$ mit $U^{\mathsf{H}}U = E$

$$U^{\mathsf{H}}(A^{\mathsf{H}}A)\,U = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \qquad (*)$$

Sei u_i die i-te Spalte von U

$$\begin{aligned} A^{\mathsf{H}}A\,U = U \cdot \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} &\Rightarrow A^{\mathsf{H}}Au_i = \lambda_i u_i \\ U^{\mathsf{H}}U = E &\Rightarrow u_i^{\mathsf{H}}u_i = 1 \end{aligned}$$

Aus $(*)$ folgt: $u_i^{\mathsf{H}}A^{\mathsf{H}}Au_i = \lambda_i = (Au_i)^{\mathsf{H}}Au_i \geq 0$.

Seien $\quad x \in \mathbb{C}^n, \quad x^{\mathsf{H}}x = 1, \quad y := U^{\mathsf{H}}x \quad (\Rightarrow U\,y = x)$

Einschub:

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad y^{\mathsf{H}} = (\overline{y}_1, \ldots, \overline{y}_n) \qquad y^{\mathsf{H}}y = \sum_{i=1}^{n} \overline{y}_i y_i = \sum_{i=1}^{n} |y_i|^2$$

$$\Rightarrow \|y\|_2 = \sqrt{\sum_{i=1}^{n} |y_i|^2} = \sqrt{y^{\mathsf{H}}y}$$

Ende.

$$\|A\,x\|_2^2 = x^{\mathsf{H}}A^{\mathsf{H}}A\,x = y^{\mathsf{H}}\underbrace{U^{\mathsf{H}}A^{\mathsf{H}}A\,U}_{\operatorname{diag}(\lambda_1,\ldots,\lambda_n)}y = \sum_{i=1}^{n}\lambda_i\,|y_i|^2$$

$$\leq \varrho(A^{\mathsf{H}}A)\sum_{i=1}^{n}|y_i|^2 = \varrho(A^{\mathsf{H}}A)\,y^{\mathsf{H}}y$$

$$= \varrho(A^{\mathsf{H}}A)\cdot x^{\mathsf{H}}UU^{\mathsf{H}}x = \varrho(A^{\mathsf{H}}A) \text{ für } x^{\mathsf{H}}x = 1$$

$$\|A\,x\|_2 \leq \sqrt{\varrho(A^{\mathsf{H}}A)} \quad \forall\, x \in \mathbb{C}^n,\ x^{\mathsf{H}}x = 1$$

$x = u_i \Rightarrow x^{\mathsf{H}}x = 1$; i so, daß $\lambda_i = \varrho(A^{\mathsf{H}}A)$.

$$\|A\,x\|_2^2 = x^{\mathsf{H}}A^{\mathsf{H}}A\,x = u_i^{\mathsf{H}}(A^{\mathsf{H}}A\,u_i) = u_i^{\mathsf{H}}\lambda_i u_i$$
$$= \lambda_i U_i^{\mathsf{H}}u_i = \lambda_i = \varrho(A^{\mathsf{H}}A)$$

□

Satz 4.5.3. *Es gilt:*

i) $\operatorname{lub}_\infty(A) := \max\limits_{\|x\|_\infty=1}\|A\,x\|_\infty = \max\limits_{i=1}^{n}\sum\limits_{k=1}^{n}|a_{ik}|$

ii) $\operatorname{lub}_1(A) := \max\limits_{\|x\|_1=1}\|A\,x\|_1 = \max\limits_{k=1}^{n}\sum\limits_{i=1}^{n}|a_{ik}|$

BEWEIS. Wird dem Leser als Übung überlassen □

Satz 4.5.4. *Für die Frobeniusnorm gilt*

$$\operatorname{lub}_2(A) \leq \|A\|_F \leq \sqrt{\operatorname{Rg}(A)}\,\operatorname{lub}_2(A)$$

(ohne Beweis)

Definition 4.5.3. $N : \mathbb{K}^{n\times n} \to \mathbb{R}_+$ sei Matrixnorm. N heißt *passend zur Vektornorm* $\|\ \| : \mathbb{K}^n \to \mathbb{R}_+$, wenn gilt

$$\|A\,x\| \leq N(A)\,\|x\|, \qquad \forall\, A \in \mathbb{K}^{n\times n}, \quad \forall\, x \in \mathbb{K}^n$$

Bemerkung. $\operatorname{lub}_{\|\ \|}$ ist passend zu $\|\ \|$. Umgekehrt: N Matrixnorm

$$\|x\| := N\begin{pmatrix} x_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ x_n & 0 & \cdots & 0 \end{pmatrix}$$

Wegen

$$x + y \longrightarrow \begin{pmatrix} x_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ x_n & 0 & \cdots & 0 \end{pmatrix} + \begin{pmatrix} y_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ y_n & 0 & \cdots & 0 \end{pmatrix} = \begin{pmatrix} x_1+y_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ x_n+y_n & 0 & \cdots & 0 \end{pmatrix}$$

gilt

$$\underbrace{N\begin{pmatrix} x_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ x_n & 0 & \cdots & 0 \end{pmatrix}}_{\|x\|} + \underbrace{N\begin{pmatrix} y_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ y_n & 0 & \cdots & 0 \end{pmatrix}}_{\|y\|} \geq \underbrace{N\begin{pmatrix} x_1+y_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ x_n+y_n & 0 & \cdots & 0 \end{pmatrix}}_{\|x+y\|}$$

D.h. $\|\ \|$ erfüllt die Dreiecksungleichung. Alle anderen Eigenschaften analog. $\|\ \|$ ist passend zu N, weil

$$A\,x = y \Rightarrow \begin{pmatrix} y_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ y_n & 0 & \cdots & 0 \end{pmatrix} = A\begin{pmatrix} x_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ x_n & 0 & \cdots & 0 \end{pmatrix}$$
$$\Rightarrow \|y\| = \|A\,x\| \leq N(A)\,\|x\|$$

4.6 Fehlerabschätzungen

Gleichungssystem:	$A\,x = b$	
Störung von b:	$A(x + \Delta_1 x) = b + \Delta b$	$\|\Delta_1 x\| \leq ?$
Störung von A:	$(A + \Delta A)(x + \Delta_2 x) = b$	$\|\Delta_2 x\| \leq ?$
Störung von A und b:	$(A + \Delta A)(x + \Delta_3 x) = b + \Delta b$	$\|\Delta_3 x\| \leq ?$

Definition 4.6.1. Für eine reguläre Matrix $A \in \mathbb{K}^{n\times n}$ und eine Matrixnorm $N : \mathbb{K}^{n\times n} \to \mathbb{R}_+$ heißt

$$\operatorname{cond}(A) := N(A) \cdot N(A^{-1})$$

die *Kondition* von A bezüglich N.

Bemerkung. A regulär

$$\begin{aligned} &\Rightarrow 0 < N(A) = N(A \cdot E) \leq N(A) \cdot N(E) \Rightarrow N(E) \geq 1 \\ &\quad 1 \leq N(E) = N(AA^{-1}) \leq N(A) \cdot N(A^{-1}) = \operatorname{cond}(A) \end{aligned}$$

Satz 4.6.1. *Sei $A \in \mathbb{K}^{n\times n}$ regulär und $0 \neq b \in \mathbb{K}^n$. $x + \Delta x$ sei Lösung von $A(x + \Delta x) = b + \Delta b$ und $A\,x = b$. N sei eine zu $\|\ \ \|$ passende Matrixnorm. Dann gilt:*

$$\|\Delta x\| \leq N(A^{-1})\,\|b\| \qquad \textit{absoluter Fehler}$$

$$\frac{\|\Delta x\|}{\|x\|} \leq \operatorname{cond}(A) \cdot \frac{\|\Delta b\|}{\|b\|} \qquad \textit{relativer Fehler}$$

Beweis.

$$\begin{aligned} A\,\Delta x = \Delta b \quad &\Rightarrow \quad \Delta x = A^{-1}\Delta b \\ &\Rightarrow \quad \|\Delta x\| \leq N(A^{-1})\,\|\Delta b\| \\ b = A\,x \quad &\Rightarrow \quad \|b\| \leq N(A)\,\|x\| \\ &\Rightarrow \quad \frac{1}{\|x\|} \leq \frac{N(A)}{\|b\|} \end{aligned}$$

□

Lemma 4.6.2. *Sei $A \in \mathbb{K}^{n\times n}$. Existiert eine Matrixnorm N mit $N(A) < 1$, dann existiert auch $(E + A)^{-1}$ und es gilt*

$$N((E + A)^{-1}) \leq \frac{1}{1 - N(A)}$$

Beweis.

1)

$$\|x\| := N\begin{pmatrix} x_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ x_n & 0 & \cdots & 0 \end{pmatrix} \qquad \text{Vektornorm}$$

$$y := A\,x \Rightarrow \begin{pmatrix} y_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ y_n & 0 & \cdots & 0 \end{pmatrix} = A\begin{pmatrix} x_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ x_n & 0 & \cdots & 0 \end{pmatrix}$$

$$\Rightarrow N\begin{pmatrix} y_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ y_n & 0 & \cdots & 0 \end{pmatrix} \leq N(A) \cdot N\begin{pmatrix} x_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ x_n & 0 & \cdots & 0 \end{pmatrix}$$

$\Rightarrow N$ ist passend zu $\|\ \ \|$.

2) $S_n := E - A + A^2 - A^3 \pm \cdots + (-1)^n A^n$

$$\begin{aligned}
(E+A)S_n &= E - A + A^2 - A^3 \pm \cdots + (-1)^n A^n \\
&\quad + A - A^2 + A^3 \mp \cdots + (-1)^n A^{n+1} \\
&= E + (-1)^n A^{n+1} \qquad x \in \mathbb{K}^n \text{ beliebig} \\
(E+A)S_n x - x &= (-1)^n A^{n+1} x \\
\|(E+A)S_n x - x\| &= \left\|A^{n+1}x\right\| \le N(A^{n+1})\,\|x\| \\
&\le \underbrace{N(A)^{n+1}\,\|x\|}_{\xrightarrow[n\to\infty]{} 0} \\
\Rightarrow (E+A)\underbrace{\lim_{n\to\infty} S_n x}_{S} &= x \\
\underbrace{(E+A)S}_{E}\cdot x &= x \quad \text{für alle } x \in \mathbb{K}^n \qquad \Rightarrow S = (E+A)^{-1}
\end{aligned}$$

3)

$$\begin{aligned}
N(S_n) &\le N(E) + N(A) + N(A)^2 + \cdots + N(A)^n \\
&\le \frac{1}{1 - N(A)} \qquad \forall\, n \in \mathbb{N} \\
\Rightarrow N(S) &\le \frac{1}{1 - N(A)}
\end{aligned}$$

□

Satz 4.6.3. *Sei $A, \Delta A \in \mathbb{K}^{n\times n}$, $b \in \mathbb{K}^n$, A regulär $A\,x = b$, $A + \Delta A)(x + \Delta x) = b$. Sei $\|\ \|$ Norm auf $\mathbb{K}^n$, N passende Matrixnorm und $N(A^{-1})N(\Delta A) < 1$. Dann gilt:*

$$\frac{\|\Delta x\|}{\|x\|} \le \frac{\operatorname{cond}(A)}{1 - \operatorname{cond}(A)\cdot\frac{N(\Delta A)}{N(A)}} \cdot \frac{N(\Delta A)}{N(A)}$$

BEWEIS. $N(A^{-1}\Delta A) \le N(A^{-1})N(\Delta A) < 1$. Nach Lemma (4.6.2) ist $E + A^{-1}\Delta A$ regulär und

$$\begin{aligned}
N((E + A^{-1}\Delta A)^{-1}) &\le \frac{1}{1 - N(A^{-1}\Delta A)} \\
\Rightarrow A + \Delta A &= A(E + A^{-1}\Delta A) \qquad \text{regulär} \\
N((A+\Delta A)^{-1}) &= N((E + A^{-1}\Delta A)^{-1}A^{-1}) \\
&\le \frac{1}{1 - N(A^{-1}\Delta A)} \cdot N(A^{-1})
\end{aligned}$$

$x = A^{-1}b, \quad x + \Delta x = (A + \Delta A)^{-1}b$

$$\begin{aligned}
\Rightarrow \Delta x &= \left[(A+\Delta A)^{-1}b - A^{-1}b\right] \\
&= (A+\Delta A)^{-1}\left[A - (A+\Delta A)\right]A^{-1}b \\
&= -(A+\Delta A)^{-1}\cdot\Delta A\cdot\underbrace{A^{-1}b}_{x} \\
\Rightarrow \|\Delta x\| &\le N((A+\Delta A)^{-1})\,N(\Delta A)\,\|x\| \\
\Rightarrow \frac{\|\Delta x\|}{\|x\|} &\le \frac{N(A^{-1})\,N(\Delta A)}{1 - N(A^{-1}\Delta A)} \\
&\le \frac{N(A^{-1})\,N(\Delta A)}{1 - N(A^{-1})\cdot N(\Delta A)\frac{N(A)}{N(A)}} \cdot \frac{N(\Delta A)}{N(A)}
\end{aligned}$$

□

Bemerkung. Ganz allgemein (Störung mit ΔA und Δb)

$$(A + \Delta A)(x + \Delta x) = b + \Delta b, \qquad A\,x = b$$

Daraus folgt:

$$\Rightarrow \frac{\|\Delta x\|}{\|x\|} \leq \frac{\operatorname{cond}(A)}{1 - \operatorname{cond}(A) \cdot \frac{N(\Delta A)}{N(A)}} \cdot \left(\frac{N(\Delta A)}{N(A)} + \frac{\|\Delta b\|}{\|b\|} \right)$$

Satz 4.6.4. *Ist $A \in \mathbb{K}^{n\times n}$ zeilenäquilibriert*

$$\exists\, c \in \mathbb{R}_+ : \quad \sum_{j=1}^{n} \left|a_{ij}\right| = c \qquad \text{für } i = 1, \ldots, n$$

dann gilt für jede Diagonalmatrix D

$$\operatorname{cond}_\infty(A) \leq \operatorname{cond}_\infty(DA)$$

4.7 Die QR-Zerlegung

Die LR-Zerlegung resp. Gaußelimination

$$A_{\text{neu}} = L^{-1} A_{\text{alt}}$$

z.B. $A_{\text{neu}} = A^{(k+1)},\ A_{\text{alt}} = A^{(k)}$

$$A_{\text{alt}} = A, \qquad A_{\text{neu}} = R$$

$$\begin{aligned} \Rightarrow \operatorname{cond}(A_{\text{neu}}) &= N(A_{\text{neu}}) \cdot N(A_{\text{neu}}^{-1}) \\ &\leq N(L^{-1})\, N(A_{\text{alt}}) \cdot N(A_{\text{alt}}^{-1})\, N(L) \\ &= \operatorname{cond}(L) \cdot \operatorname{cond}(A_{\text{alt}}) \end{aligned}$$

Im allgemeinen ist $\operatorname{cond}(L) > 1$ und eine Konditionszahlverstärkung um den Faktor $\operatorname{cond}(L)$ tritt oft auf.

Gibt es eine Transformation mit $\operatorname{cond}(QA) = \operatorname{cond}(A)$ und QR Dreiecksgestalt?

Q unitär

$$\Leftrightarrow \quad Q^{\mathsf{H}} Q = E \quad \Leftrightarrow \quad Q^{-1} = Q^{\mathsf{H}} \quad \Rightarrow \quad Q Q^{\mathsf{H}} = E \quad \Rightarrow \quad Q^{\mathsf{H}} \text{ unitär}$$

$$\operatorname{lub}_2(Q) = \sqrt{\varrho(Q^{\mathsf{H}} Q)} = \sqrt{\varrho(E)} = 1$$

Satz 4.7.1. *$A \in \mathbb{C}^{n\times n}$, Q unitär*

$$\Rightarrow \operatorname{cond}_2(QA) = \operatorname{cond}_2(A)$$

Beweis.

$$\begin{aligned} \|A\|_2 = \left\|Q^{\mathsf{H}} \cdot QA\right\|_2 &\leq \left\|Q^{\mathsf{H}}\right\|_2 \|QA\|_2 = \|QA\|_2 \\ &\leq \|Q\|_2 \|A\|_2 = \|A\|_2 \\ \Rightarrow \|A\|_2 &= \|QA\|_2 \end{aligned}$$

Genauso für $A^{-1} Q^{\mathsf{H}}$

$$\begin{aligned} \left\|A^{-1}\right\|_2 = \left\|A^{-1} Q^{\mathsf{H}} Q\right\| &\leq \left\|A^{-1} Q^{\mathsf{H}}\right\|_2 \leq \left\|A^{-1}\right\|_2 \\ \left\|A^{-1}\right\| = \left\|A^{-1} Q^{\mathsf{H}}\right\| &\qquad \left\|A^{-1} Q^{\mathsf{H}}\right\| = (QA)^{-1} \\ \|A\|_2 \left\|A^{-1}\right\|_2 &\leq \|QA\|_2 \left\|(QA)^{-1}\right\|_2 \end{aligned}$$

□

Satz 4.7.2. *Jede reguläre Matrix $A \in \mathbb{C}^{n\times n}$ läßt sich schreiben als*

$$A = Q \cdot R$$

mit Q unitär, R obere Dreiecksmatrix mit positiver Diagonalen.

BEWEIS. Mit dem Orthogonalisierungsverfahren von E. Schmidt.
$a_2, \ldots, a_n$ Spalten von A
$q_2, \ldots, q_n$ (noch unbekannte) Spalten von Q
$Q^{\mathsf{H}} \cdot Q = E$ bedeutet

$$q_i^{\mathsf{H}} \cdot q_j = \left\{ \begin{array}{ll} 0 & \text{für } i \neq j \\ 1 & \text{für } i = j \end{array} \right\} = \delta_{ij}$$

$A = QR$, $A \cdot R^{-1} = Q$ ergibt:

$$\langle a_1, \ldots, a_n \rangle = \langle q_1, \ldots, q_n \rangle \qquad k = 1, \ldots, n$$

Induktion über k:

$k = 1 \quad a_1 = r_{11} q_1 \quad r_{11} := \|a_1\|_2 = \sqrt{a_1^{\mathsf{H}} a_1} > 0$

$$q_1 := \frac{1}{r_{11}} a_1 \Rightarrow q_1^{\mathsf{H}} \cdot q_1 = 1$$

$$\langle a_1 \rangle = \langle q_1 \rangle$$

$k - 1 \Rightarrow k$:

$$\tilde{q}_k := a_k - \sum_{l=1}^{k-1} (q_l^{\mathsf{H}} \cdot a_k) \cdot q_l \neq 0 \text{ weil } \langle a_1, \ldots, a_{k-1} \rangle = \langle q_1, \ldots, q_{k-1} \rangle$$

$$r_{kk} := \|\tilde{q}_k\|_2 > 0$$

$$q_k := \frac{1}{r_{kk}} \tilde{q}_k$$

$$\Rightarrow q_k^{\mathsf{H}} q_k = 1 \Rightarrow \langle a_1, \ldots, a_k \rangle = \langle q_1, \ldots, q_k \rangle$$

$$r_{kk} \cdot q_k + \sum_{l=1}^{k-1} (q_l^{\mathsf{H}} \cdot a_k) \cdot q_l = a_k$$

$$r_{kl} := q_l^{\mathsf{H}} a_k \qquad \text{für } l = 1, \ldots, k-1$$

$$\Rightarrow a_k = \sum_{l=1}^{k} r_{kl} q_l$$

$j < k$:

$$\begin{aligned} q_j^{\mathsf{H}} q_k &= \frac{1}{r_{kk}} q_j^{\mathsf{H}} \cdot \tilde{q}_k \\ &= \frac{1}{r_{kk}} \Big(q_j^{\mathsf{H}} \cdot a_k - \sum_{l=1}^{k-1} (q_l^{\mathsf{H}} \cdot a_k) \underbrace{q_j^{\mathsf{H}} q_l}_{\delta_{ij}} \Big) \\ &= \frac{1}{r_{kk}} (q_j^{\mathsf{H}} \cdot a_k - q_j^{\mathsf{H}} a_k) = 0 \end{aligned}$$

$$\Rightarrow q_k^{\mathsf{H}} q_j = 0$$

□

Satz 4.7.3. *Die QR-Zerlegung von einer regulären Matrix A, $A = QR$, Q unitär, R obere Dreiecksmatrix, ist eindeutig, wenn*

$$r_{kk} > 0, \qquad k = 1, \ldots, n$$

BEWEIS. $A = Q_1 \cdot R_1 = Q_2 \cdot R_2$. QR-Zerlegung, wo R_1, R_2 pos. Elemente haben.
$\Rightarrow Q_2^{\mathsf{H}} Q_1 = R_2 R_1^{-1} =: D$ obere Dreiecksmatrix.

$$D^{\mathsf{H}} = Q_1^{\mathsf{H}} \cdot Q_2 = (Q_2^{\mathsf{H}} Q_1)^{-1} = D^{-1}$$

$\Rightarrow D$ ist Diagonalmatrix, alle Diagonalelemente sind positiv.

$$\Rightarrow D^{\mathsf{H}} = D^{-1} = E$$

□

Nachteil des Orthogonalisierungsverfahrens von E. Schmidt

$$\tilde{q}_k := a_k - \sum_{l=1}^{k-1} (q_l^{\mathsf{H}} \cdot a_k) \cdot q_l$$

kann sehr klein sein. Stabilere Methode von Householder.

Definition 4.7.1. Sei $w \in \mathbb{C}^n$, $\|w\|_2 = 1$. Dann heißt

$$H_w = E - 2\,ww^{\mathsf{H}} = \begin{pmatrix} 1-2|w_1|^2 & -2w_1\overline{w}_2 & \cdots & -2w_1\overline{w}_n \\ -2\overline{w}_1 w_2 & 1-2|w_2|^2 & & -2w_2\overline{w}_n \\ \vdots & & \ddots & \vdots \\ -2\overline{w}_1 w_n & \cdots\cdots\cdots\cdots & & 1-2|w_n|^2 \end{pmatrix}$$

elementare hermitesche Matrix bzw. *Householder-Matrix*. Die Abbildung

$$\Phi_n : \mathbb{C}^n \to \mathbb{C}^n : z \mapsto H_w z$$

heißt *Householder-Transformation*.

Satz 4.7.4. *Es gelten die Aussagen:*

i) H_w *ist hermitesch und unitär*

$$H_w^{\mathsf{H}} = H_w, \qquad H_w^{\mathsf{H}} H_w = E$$

ii) H_w *ist involutorisch*

$$H_w^2 = E$$

iii) H_w *hat den Eigenwert* -1 *mit dem Eigenvektor* w *und den* $(n-1)$*-fachen Eigenwert* 1 *mit dem Eigenraum*

$$V = \{u \in \mathbb{C}^n \mid w^{\mathsf{H}} u = 0\} \qquad \textit{Hyperebene mit dem Normalenvektor } w$$

iv) Sei $w = (0, w_2, \ldots, w_n)^\top, \tilde{w} = (w_2, \ldots, w_n)^\top$. *Dann gilt:*

$$H_w = \begin{pmatrix} 1 & 0 \\ 0 & \boxed{H_{\tilde{w}}} \end{pmatrix}$$

Beweis. Der Beweis wird dem Leser als Übung überlassen. □

Bemerkung. Die Householder-Transformation ist eine Spiegelung an der Hyperebene V, denn $\{u_1, \ldots, u_{n-1}\}$ ist eine Basis von V.

$$z \in \mathbb{C}^n \Rightarrow z = \sum_{i=1}^{n-1} c_i u_i + c_n w$$

$$H_w z = \sum_{i=1}^{n-1} c_i H_w u_i + c_n H_w w$$

$$= \sum_{i=1}^{n-1} c_i u_i - c_n w$$

Frage: Gibt es zu $a, b \in \mathbb{C}^n$ ein $w \in \mathbb{C}^n$, $\|w\|_2 = 1$ mit $H_w a = b$?

Antwort: Ja, wenn $\|a\|_2 = \|b\|_2$. Analytischer Beweis für $b = \begin{pmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \alpha \cdot e_1$:

H_w ist unitär

$$\Rightarrow \|a\|_2 = \|H_w a\|_2 = \left\| \begin{pmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{pmatrix} \right\| = |\alpha|$$

$$\alpha \in \mathbb{C} \Rightarrow \alpha = \varrho \cdot e^{i\varphi}, \qquad \varrho = \|a\|_2\,,\ \varphi \in [0, 2\pi) \text{ noch unbek.}$$

$$H_w a = a - 2w(w^{\mathsf{H}} a) = a - 2(w^{\mathsf{H}} a)w = \alpha e_1 \qquad (*)$$

$$\Rightarrow w = \frac{1}{2(w^{\mathsf{H}} a)}(a - \alpha e_1)$$

$$\Rightarrow w = \frac{1}{\|a - \alpha e_1\|}(a - \alpha e_1)$$

Bestimmung von $e^{i\varphi}$, $a = (a_1, \dots a_n)^\top$. Aus $(*)$ folgt durch Multiplikation mit a^{H}:

$$a^{\mathsf{H}} a - 2\underbrace{(w^{\mathsf{H}} a)}_{=\overline{a^{\mathsf{H}} w}} a^{\mathsf{H}} w = \alpha a^{\mathsf{H}} e_1 = \alpha \overline{a}_1$$

$$= \|a\|_2^2 - 2\left|a^{\mathsf{H}} w\right|^2 \in \mathbb{R}$$

Sei $a_1 = |a_1|\, e^{i\sigma} \quad \Rightarrow \quad e^{i\sigma} = \dfrac{a_1}{|a_1|}$

$$\alpha \overline{a}_1 = \|a\|_2\, e^{i\varphi}\ |a_1|\, e^{-i\sigma} = \|a\|_2\, |a_1|\, e^{i(\varphi - \sigma)}$$

$$= \|a\|_2\, |a_1|\, [\cos(\varphi - \sigma) + i \sin(\varphi - \sigma)] \in \mathbb{R}$$

$$\Rightarrow \varphi - \sigma \in \{-\pi, 0, \pi\}$$

Festlegung des Vorzeichens von α :

$$\|a - \alpha e_1\|_2 = \sqrt{|a_1 - \alpha|^2 + |a_2| + \cdots + |a_n|^2}$$

$$= \sqrt{\left||a_1|\ e^{i\sigma} \mp \|a\|_2\ e^{i\sigma}\right|^2 + |a_2| + \cdots}$$

$$= \sqrt{(|a_1| \mp \|a\|_2)^2 + |a_2| + \cdots} \qquad \text{Auslöschung}$$

Auslöschung vermeiden durch die Wahl des unteren Vorzeichens!

$$\alpha = -e^{i\sigma}\ \|a\|_2 = -\frac{\|a\|_2}{|a_1|} a_1 \text{ für } a_1 \neq 0$$

Wenn $a_1 = 0$, dann $\alpha := -\|a\|_2$

Bemerkung. Wenn $a \in \mathbb{R}^n$, dann $\alpha \in \mathbb{R}$ und $w \in \mathbb{R}^n$.

Algorithmus (Householder).

Gegeben: $a \in \mathbb{C}^n, \quad a = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$

Gesucht: $\alpha \in \mathbb{C},\ w \in \mathbb{C}^n : \quad w^{\mathsf{H}} w = 1$
mit $H_w a = \alpha e_1$

$$\alpha := \begin{cases} -\|a\|_2 \cdot \frac{a_1}{|a_1|} & \textit{für } a_1 \neq 0 \\ -\|a\|_2 & \textit{für } a_1 = 0 \end{cases}$$

$$\beta := \|a - \alpha e_1\|_2 = \sqrt{(|a_1| + \|a\|_2)^2 + |a_2|^2 + \cdots + |a_n|^2}$$

$$w = \frac{1}{\beta}(a - \alpha e_1)$$

Satz 4.7.5. *Zu einer beliebigen Matrix $A \in \mathbb{C}^{m\times n}$ existiert eine unitäre Matrix $Q \in \mathbb{C}^{m\times m}$ und eine obere Dreiecksmatrix $R \in \mathbb{C}^{m\times n}$ mit $A = QR$*

Beweis. $A^{(0)} := A$. Sei $a_1^{(0)}$ die erste Spalte von $A^{(0)}$, w_1 so, daß $H_{w_1} a_1^{(0)} = \alpha_1 e_1$

$$H_{w_1} A^{(0)} = \begin{pmatrix} \alpha_1 & * & \cdots & * \\ 0 & & & \\ \vdots & & \tilde{A}^{(1)} & \\ 0 & & & \end{pmatrix} \Bigg\} m-1$$

$$\underbrace{\qquad\qquad}_{n-1}$$

$\tilde{a}_1^{(1)}$ ist erste Spalte von $\tilde{A}^{(1)}$

$$H_{\tilde{w}_1} \tilde{A}^{(1)} = \begin{pmatrix} \alpha_2 & * & \cdots & * \\ 0 & & & \\ \vdots & & \tilde{A}^{(2)} & \\ 0 & & & \end{pmatrix} \qquad w_2 = (0, \tilde{w}_{22}, \ldots, \tilde{w}_{2n})^\top$$

$$H_{w_2} H_{w_1} A = H_{w_2} \begin{pmatrix} \alpha_1 & * & \cdots & * \\ 0 & & & \\ \vdots & & \tilde{A}^{(1)} & \\ 0 & & & \end{pmatrix}$$

$$= \begin{pmatrix} \alpha_1 & * & \cdots\cdots & * \\ 0 & \alpha_2 & * \cdots & * \\ 0 & 0 & & \\ \vdots & \vdots & \tilde{A}^{(2)} & \\ 0 & 0 & & \end{pmatrix}$$

usw. bis

$$H_{w_s} \cdots H_{w_1} A = \begin{pmatrix} \alpha_1 & * & \cdots\cdots\cdots & * \\ 0 & \alpha_2 & * & \cdot \\ \vdots & 0 & \ddots\ * & \cdot \\ 0 & 0 & 0\ \alpha_s\ * & \cdot \\ 0 & \cdots\cdots & 0\ *\ * & * \end{pmatrix} \qquad \text{falls } m < n$$

$$= \begin{pmatrix} \alpha_1 & * & \cdots & * \\ 0 & \alpha_2 & * & \cdots \\ \vdots & 0 & \ddots & * \\ 0 & \cdots & 0 & \alpha_s \\ 0 & \cdots\cdots & 0 & \end{pmatrix} \qquad \text{falls } n \le m$$

$$s = \min\{m-1, n\}$$

$H_{w_s} \cdots H_{w_1} A = R \Rightarrow A = QR$ mit $Q = H_{w_1}^{-1} \cdots H_{w_s}^{-1}$. Q unitär und $Q = H_{w_1} \cdots H_{w_s}$. □

Der Algorithmus ist stabiler und allgemeiner als das Schmidtsche Orthogonalisierungsverfahren!
Wenn $A \in \mathbb{R}^{m\times n}$, dann auch $Q \in \mathbb{R}^{m\times n}$ und $R \in \mathbb{R}^{m\times n}$.

$$A\,x = b \qquad Q^{\mathsf{H}} QR\,x = Q^{\mathsf{H}} b$$
$$R\,x = Q^{\mathsf{H}} b$$

4.8 Lineare Ausgleichsprobleme

Methode der kleinsten Quadrate (Gauß).
Meßpunkte x_i Meßwerte y_i, $\quad i = 1, \ldots, n.$

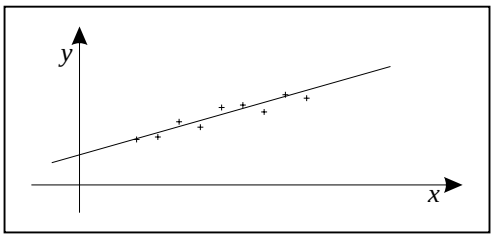


Abbildung 4.2: Meßpunkte

Aus physikalischen oder ökonomischen Gründen muß

$$y_i = \alpha + \beta x_i \qquad i = 1, \ldots, n$$

gelten.

$$r_i = \alpha + \beta x_i - y_i \qquad \text{Residuum}$$
$$r = (r_1, \ldots, r_n)^\top \qquad \text{Residuenvektor}$$

Gesucht sind α^*, β^* mit

$$\min_{\alpha,\beta} \sum_{i=1}^{n} (\alpha + \beta x_i - y_i)^2 = \sum_{i=1}^{n} (\alpha^* + \beta^* x_i - y_i)^2$$

Methode der kleinsten Quadrate (least-square problem).
In Matrixform:

$$r = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} - \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Gesucht: $\min\limits_{(\alpha,\beta)\in\mathbb{R}^2} r^\top r = \min\limits_{(\alpha,\beta)\in\mathbb{R}^2} \|r\|_2^2$. Allgemeiner:
Gesucht: $\min\limits_{x\in\mathbb{R}^n} \|Ax - b\|_2^2$ (lineares Ausgleichsproblem).
Verallgemeinerung der Lösung linearer Gleichungssysteme auf die „Lösung“ von überbestimmten und unlösbaren Problemen. Im folgenden Beschränkung auf $\mathbb{R}^n$.
Im $\mathbb{R}^n$ gilt: $x, y \in \mathbb{R}^n \Rightarrow \langle x, y\rangle := \sum_{i=1}^n x_i y_i$

$\|x\|_2 = \sqrt{\langle x, x\rangle}$, d.h. Euklidische Norm stammt vom inneren Produkt ab.

Satz 4.8.1. *Sei V ein Vektorraum mit innerem Produkt. $U_n \subseteq V$ sei ein n-dimensionaler Unterraum von V. V hat die Norm $\|x\| = \sqrt{\langle x, x\rangle}$. Zu jedem $f \in V$ gibt es genau ein $h^* \in U_n$ mit:*

$(*)$ $\langle f - h^*, h\rangle = 0 \quad \forall\, h \in U_n$

$(**)$ *Für h^* gilt:*

$$\|f - h^*\| = \min_{h\in U_n} \|f - h\| \text{ und}$$
$$\|f - h^*\| = \left\|f - \tilde{h}\right\|, \qquad \tilde{h} \in U_n \Rightarrow h^* = \tilde{h}$$

Beispiel. $V = \{f : [a, b] \to \mathbb{R} \mid \int_a^b f^2(x)\,dx > \infty\ (\approx \mathrm{L}[a, b])\}$

$$\langle f, g\rangle := \int_a^b f(x)g(x)\,dx$$

$$\langle f, f\rangle = \|f\|^2 = 0 \overset{!}{\Rightarrow} f = 0$$

$\mathbb{P}_n$: Menge der Polynome in V

$$\|f - h^*\|^2 = \int_a^b [f(x) - h^*(x)]^2 \, dx$$

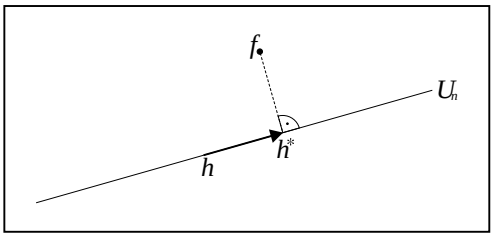


Abbildung 4.3: orthogonale Projektion

Bemerkung. Wegen Satz 4.8.1 $(*)$ heißt h^* *orthogonale Projektion* von f auf U_n. Wegen Satz 4.8.1 $(**)$ heißt h^* die beste Approximation von f auf U_n.

Beweis. $\|x\|$ ist Norm (s. Lineare Algebra). Sei $\{h_1, \ldots, h_n\}$ eine Basis von U_n, $h^* = \sum_{i=1}^n c_i h_i$. Dann ist Satz 4.8.1 $(*)$ äquivalent zu

$$\left\langle f - \sum_{i=1}^n c_i h_i, h_j \right\rangle = 0, \qquad j = 1, \ldots, n$$

Das ist äquivalent zu

$$\sum_{i=1}^n c_i \left\langle h_i, h_j \right\rangle = \left\langle f, h_j \right\rangle, \qquad j = 1, \ldots, n$$

Die Koeffizientenmatrix ist regulär (s. Box).

$$\sum_{i=1}^n c_i \left\langle h_i, h_j \right\rangle = 0, \qquad j = 1, \ldots, n$$
$$\Rightarrow \left\langle \sum_{i=1}^n c_i h_i, h_j \right\rangle = 0, \qquad j = 1, \ldots, n$$
$$\Rightarrow \sum_{i=1}^n c_i h_i = 0 \Rightarrow c_1 = \cdots = c_n = 0$$

Also existieren $c_1, \ldots, c_n$ und sind eindeutig durch Satz 4.8.1 $(*)$ bistimmt. Das gilt dann auch für $h^* = \sum_{i=1}^n c_i h_i$

$$\begin{aligned}
\|f - h\|^2 &= \langle f - h, f - h \rangle = \langle f - h^* + h^* - h, f - h^* + h^* - h \rangle \\
&= \langle f - h^*, f - h^* \rangle + \langle f + h^*, \underbrace{h^* - h}_{\in U_n} \rangle + \langle \underbrace{h^* - h}_{\in U_n}, f - h^* \rangle + \langle h^* - h, h^* - h \rangle \\
&= \|f - h^*\|^2 + 0 + 0 + \|h^* - h\|^2 \\
&\geq \|f - h^*\|^2, \quad \text{Gleichheit nur für } \|h^* - h\| = 0, \text{ d.h. für } h^* = h
\end{aligned}$$

□

Satz 4.8.2. *Sei $a \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ mit $m > n$ gegeben. Betrachte das Ausgleichsproblem*

$$\|Ax^* - b\|_2^2 = \min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 \qquad \odot$$

Dann gilt:

i) x^* *löst* $\odot$ *genau dann, wenn es die Normalgleichungen*

$$A^\top A\, x^* = A^\top b$$

erfüllt

ii) Die Menge

$$L := \{x \in \mathbb{R}^n \mid A^\top A\, x = A^\top b\}$$

ist ein nichtleerer affiner Raum, d.h.

$$x_1, x_2 \in L \Rightarrow (1-\lambda)\, x_1 + \lambda x_2 \in L$$

für alle $\lambda \in \mathbb{R}$. *Ferner gilt* $A\, x_1 = A\, x_2$

iii) $\odot$ *hat genau dann eine eindeutige Lösung, wenn* $\mathrm{Rg}(A) = n$ *(voller Spaltenrang) ist*

iv) Unter allen Lösungen von $\odot$ *gibt es genau eine Lösung* x^+ *mit minimaler Euklidischer Norm*

Bemerkung. x^+ heißt *Pseudonormallösung* von $\odot$ resp. von $A\, x = b$.

BEWEIS. **i):** Nach Satz 4.8.1 mit $V = \mathbb{R}^n$ und $U_n = \mathrm{Bild}(A) = \{Ax \mid x \in \mathbb{R}^n\}, b = f$. $(*)$ ist hier

$$\langle Ax^* - b, a\, y\rangle = 0 \qquad \forall\, y \in \mathbb{R}^n$$

Nebenrechnung:

$$\begin{aligned}\langle u, Av\rangle &= u^\top (Av) = (u^\top A)\, v = (A^\top u)^\top v \\ &= \left\langle A^\top u, v\right\rangle\end{aligned}$$

Dies ist äquivalent zu $A^\top(Ax^* - b) = 0,\ Ax^* = h^*$
ii): $L \neq \emptyset$ weil Ax^* nach Satz 4.8.1 existiert, L : affiner Lösungsraum. $Ax_1 = Ax_2$ weil $Ax_1 = Ax^* = Ax_2$
iii): $\mathrm{Kern}(A^\top A) = \mathrm{Kern}(A)$, also ist $A^\top A$ regulär, genau dann, wenn $\mathrm{Rg}(A) = n$ ist
iiv): $L^* = L \cap \{x \in \mathbb{R}^n \mid \|x\|_2 \leq \|x^*\|_2\}$ ist kompakt (beschränkt und abgeschlossen, $\mathbb{R}^n$ endlichdimensional) und nichtleer wegen $x^* \in L^*$. Die Abbildung $x \mapsto \|x\|_2$ ist stetig. Also existiert ein Minimum von $x \mapsto \|x\|_2$ auf L^*. Es sei bei $x^+ \in L^*$. Dann ist auch $\|x^+\|_2 = \min_{x\in L} \|x\|_2$.
Seien x_1 und x_2 beides Pseudonormallösungen, $\|x_1\|_2 = \|x_2\|_2 = \min_{x\in L} \|x\|_2 =: \varrho$, dann ist

$$\begin{aligned}\left\|\frac{x_1 + x_2}{2}\right\|_2 &\leq \frac{1}{2}\|x_1\|_2 + \frac{1}{2}\|x_2\|_2 \\ &= \frac{1}{2}\varrho + \frac{1}{2}\varrho = \varrho = \left\|\frac{x_1 + x_2}{2}\right\|_2 = \varrho \\ 4\varrho^2 &= \langle x_1 + x_2, x_1 + x_2\rangle = \langle x_1, x_1\rangle + 2\langle x_1, x_2\rangle + \langle x_2, x_2\rangle \\ &= \varrho^2 + 2\langle x_1, x_2\rangle + \varrho^2 \Rightarrow \langle x_1, x_2\rangle = \varrho^2 \\ \|x_1 - x_2\|_2^2 &= \langle x_1 - x_2, x_1 - x_2\rangle = \langle x_1, x_1\rangle - 2\langle x_1, x_2\rangle + \langle x_2, x_2\rangle \\ &= 0 \quad \Rightarrow x_1 = x_2\end{aligned}$$

□

Bemerkung. Ist die QR-Zerlegung von A bekannt, $m \geq n$, $\quad A = Q\, R$, Q orthogonal.

$$R = \begin{pmatrix} \alpha_1 & * & \cdots & * \\ 0 & \alpha_2 & \ddots & \vdots \\ \cdot & 0 & \ddots & * \\ \cdot & 0 & \ddots & \alpha_n \\ \cdot & 0 & & 0 \\ \vdots & & & \vdots \\ 0 & \cdots\cdots & & 0 \end{pmatrix} = \begin{pmatrix} & R_1 & \\ 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \Big\}n \text{ Zeilen}$$

$$\|Ax-b\|_2^2 = \left\|Q^\top(Ax-b)\right\|_2^2 = \left\|\begin{pmatrix} R_1 x \\ 0 \end{pmatrix} - Q^\top b\right\|_2^2, \qquad b = \begin{pmatrix} c \\ \cdots \\ d \end{pmatrix} \begin{matrix} \}n \\ \\ \}m-n \end{matrix}$$

$$= \left\|\begin{pmatrix} Rx \\ 0 \end{pmatrix} - \begin{pmatrix} Qc \\ Q^\top d \end{pmatrix}\right\|_2^2 = \left\|R_1 x - Q^\top c\right\|_2^2 + \left\|Q^\top d\right\|_2^2$$

$$= \left\|R_1 x - Q^\top c\right\|_2^2 + \|d\|_2^2$$

Wenn $\mathrm{Rg}(A) = n$, dann auch $\mathrm{Rg}(R_1) = \mathrm{Rg}(R) = \mathrm{Rg}(Q^\top A) = \mathrm{Rg}(A) = n$ und x^* ist eindeutig bestimmt durch

$$\boxed{R_1 x^* = Q^\top c}$$

Im Fall $\mathrm{Rg}(A) < n$ sind feinere Methoden erforderlich!

Definition 4.8.1. Sei $A \in \mathbb{R}^{m\times n}$. Die Darstellung

$$A = U\Sigma V^\top$$

mit orthogonalen Matrizen $U \in \mathbb{R}^{m\times m}, V \in \mathbb{R}^{n\times n}$ und Diagonalmatrix $\Sigma \in \mathbb{R}^{m\times n}$ heißt *Singulärwertzerlegung* von A.

Lemma 4.8.3. *Sei $A \in \mathbb{R}^{m\times n}$. Die positiven Eigenwerte von $A^\top A$ und $AA^\top$ stimmen überein und haben dieselbe Vielfachheit, d.h.*

$$\dim \mathrm{Kern}(\lambda E_n - A^\top A) = \dim \mathrm{Kern}(\lambda E_n - AA^\top)$$

Bemerkung. $A^\top A$ und $AA^\top$ sind reell symmetrisch $\Rightarrow$ alle Eigenwerte sind reell.

$$A^\top A\, x = \lambda x, \qquad x \neq 0 \qquad \Rightarrow \underbrace{x^\top A^\top}_{(Ax)^\top} Ax = \lambda x^\top x$$

$$\Rightarrow \lambda = \frac{\|Ax\|_2^2}{\|x\|_2^2} \geq 0$$

Analog $AA^\top x = \lambda x, \qquad x \neq 0 \qquad \Rightarrow \lambda \geq 0$

Beweis. $\lambda > 0,\ x \neq 0, \qquad A^\top A = \lambda x \quad \Rightarrow \quad (AA^\top)Ax = \lambda Ax$
$\Rightarrow Ax$ ist Eigenvektor zum Eigenwert λ von $AA^\top$.
$\left[Ax \neq 0 \text{ weil sonst } A^\top Ax = 0 = \lambda x \neq 0\right]$
$\lambda > 0,\ x \neq 0, \qquad AA^\top = \lambda x \quad \Rightarrow \quad (A^\top A)A^\top x = \lambda A\top x$
$\Rightarrow A^\top x$ ist Eigenvektor zum Eigenwert λ von $A^\top A$.
$\left[A^\top x \neq 0 \text{ weil sonst } AA^\top x = 0 = \lambda x \neq 0\right]$
Sei $\{x_1, \ldots, x_k\}$ Orthonormalbasis für den Eigenraum $\mathrm{Kern}(\lambda E_n - A^\top A)$. Insbesondere also

$$x_i^\top x_j = \delta_{ij} \qquad A^\top A x_i = \lambda x_i, \qquad i, j = 1, \ldots, k$$

$\Rightarrow Ax_1, \ldots, Ax_k$ im Eigenraum zum Eigenwert λ von $AA^\top$.

$$(Ax_i)^\top (Ax_j) = x_i^\top A^\top A x_j = \lambda x_i^\top x_j = \lambda \delta_{ij}$$

$\Rightarrow Ax_1, \ldots, Ax_k$ sind zueinander orthogonal, insbesondere linear unabhängig.
$\Rightarrow k = \dim \mathrm{Kern}(\lambda E_n - A^\top A) \leq \dim \mathrm{Kern}(\lambda E_n - AA^\top)$.
Analog $\dim \mathrm{Kern}(\lambda E_n - AA^\top) \leq \dim \mathrm{Kern}(\lambda E_n - A^\top A)$.
Zusammen also

$$\dim \mathrm{Kern}(\lambda E_n - AA^\top) \leq \dim \mathrm{Kern}(\lambda E_n - A^\top A)$$

□

Lemma 4.8.4. *Für $A \in \mathbb{R}^{m\times n}$ ist*

$$\mathrm{Rg}(A) = \mathrm{Rg}(A^\top) = \mathrm{Rg}(AA^\top) = \mathrm{Rg}(A^\top A)$$

die Anzahl der positiven Eigenwerte von $A^\top A$ resp. $AA^\top$ mit der Vielfachheit

$$\dim \mathrm{Kern}(\lambda E_n - A^\top A)$$

gezählt.

BEWEIS. $\mathrm{Rg}(A) = \mathrm{Rg}(A^\top)$ wegen Zeilenrang = Spaltenrang. Ferner bekannt aus der Linearen Algebra:

$$\text{für } B \in \mathbb{R}^{r\times s}: \quad \mathrm{Rg}(B) = s - \dim \mathrm{Kern}(B)$$

Für $B = A^\top A$ und $B = A$ gibt die Differenz

$$\mathrm{Rg}(A^\top A) - \mathrm{Rg}(A) = (n-n) - \dim \mathrm{Kern}(A^\top A) + \dim \mathrm{Kern}(A)$$

Wegen $\mathrm{Kern}(A^\top A) = \mathrm{Kern}(A)$ folgt

$$\mathrm{Rg}(A) = \mathrm{Rg}(A^\top A)$$

Der Rang einer symmetrischen Matrix ist die Anzahl der Eigenwerte $\neq 0$ mit der Vielfachheit gezählt. $A^\top A$ und $AA^\top$ sind symmetrisch $\Rightarrow$ Behauptung. □

Satz 4.8.5. *Sei $A \in \mathbb{R}^{m\times n}$ mit $r = \mathrm{Rg}(A)$. Dann existieren orthogonale Matrizen*

$$U = (u_1, \ldots, u_m) \in \mathbb{R}^{m\times m}, \qquad V = (v_1, \ldots, v_n) \in \mathbb{R}^{n\times n}$$

derart, daß

$$\Sigma := U^\top A\, V = \mathrm{diag}(\sigma_1, \ldots, \sigma_{\min\{n,m\}}) \in \mathbb{R}^{m\times n}$$

mit $\sigma_1, \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_{\min\{n,m\}} = 0$

Bemerkung. Σ hat die Gestalt:

$$\Sigma = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \\ \hline & & & 0 \end{array}\right) \in \mathbb{R}^{m\times n}$$

BEWEIS. $\{v_1, \ldots, v_n\}$ Orthonormalsystem aus Eigenvektoren zu $A^\top A$.

$$A^\top A v_i = \lambda_i v_i \qquad i = 1, \ldots, n$$

O.B.d.A. $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > \lambda_{r+1} = \cdots = \lambda_n = 0$. Sei

$$\sigma_i := \begin{cases} \sqrt{\lambda_i} & i = 1, \ldots, r \\ 0 & i = r+1, \ldots, \min\{m,n\} \end{cases}$$

$$u_i := \frac{1}{\sigma_i} A v_i \qquad i = 1, \ldots, r$$

Dann gilt:

$$AA^\top u_i = \frac{1}{\sigma_i} AA^\top A v_i = \frac{\lambda_i}{\sigma_i} A v_i = \lambda_i u_i$$

$$u_i^\top u_j = \frac{1}{\sigma_i \sigma_j} v_i^\top A^\top A v_i = \frac{\lambda_j}{\sigma_i \sigma_j} \underbrace{v_i^\top \cdot v_j}_{\delta_{ij}}$$

$$= \frac{\lambda_j}{\sigma_i \sigma_j} \delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \quad \text{(Def. von } \sigma_i\text{)} \end{cases}$$

Ergänze $\{u_1, \dots, u_r\}$ durch ein Orthonormalsystem $\{u_{r+1}, \dots, u_m\}$ von Eigenvektoren zum Eigenwert 0 von $AA^\top$.
$U := (u_1, \dots, u_m)$, $V := (v_1, \dots, v_n)$ orthogonal.

$$U^\top AV = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \\ \hline & & & 0 \end{array}\right)$$

$i \le r$, $j \le r$:

$$\begin{aligned}(U^\top AV)_{ij} = u_i^\top A v_j &= \frac{1}{\sigma_i} v_i^\top A^\top A v_j = \frac{\lambda_j}{\sigma_i} v_i^\top v_j \\ &= \frac{\lambda_j}{\sigma_i}\delta_{ij} = \begin{cases} 0 & i \neq j \\ \sigma_i & i = j \end{cases}\end{aligned}$$

$i \le r$, $j > r$:

$$u_i^\top A v_j = \frac{1}{\sigma_i} v_i^\top \underbrace{A^\top A v_j}_{j>r} = \frac{1}{\sigma_i} v_i^\top \cdot 0 = 0$$

$i > r$, j beliebig:

$$\begin{aligned} AA^\top u_i = 0 \cdot u_i = 0 \quad &\Rightarrow \quad u_i \in \operatorname{Kern}(AA^\top) = \operatorname{Kern}(A^\top) \\ \Rightarrow A^\top u_i = 0 \quad &\Rightarrow U_i^\top A^\top = 0 \\ \Rightarrow \underbrace{u_i^\top A}_{=0} v_j &= 0 \end{aligned}$$

Jordan:

$$A := TJT^{-1} \quad \Rightarrow \quad AT = TJ = T \left(\begin{array}{cccc} \square & & & \\ & \square & & \\ & & \square & \\ & & & \square \end{array}\right)$$

$$At_i = \lambda t_i + t_{i-1}$$

□

Definition 4.8.2. Die positiven Eigenwerte von $AA^\top$ (und $A^\top A$) sind eindeutig durch A bestimmt. Damit sind auch $\sigma_1, \dots, \sigma_r$ eindeutig durch A bestimmt. Sie heißen *Singulärwerte* von A. (U und V nicht eindeutig!)

Nun zurück zum linearen Ausgleichsproblem:

$A = U\Sigma V^\top$ **eine** (mögliche) Singulärwertzerlegung

$U = (u_1, \dots, u_m)$. Dann gilt für beliebige $x \in \mathbb{R}^n$

$$\begin{aligned} \|Ax - b\|_2^2 &= \left\|U^\top (Ax - b)\right\|_2^2 \\ &= \left\|U^\top AVV^\top x - U^\top b\right\|_2^2 \\ &= \left\|\Sigma V^\top x - U^\top b\right\|_2^2 \\ &= \sum_{i=1}^{r} \left(\sigma_i (V^\top x)_i - u_i^\top b\right)^2 + \underbrace{\sum_{i=r+1}^{n} (-u_i^\top b)^2}_{\text{const!}} \end{aligned}$$

Minimalwert ist

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 = \sum_{i=r+1} (u_i^\top b)^2$$

Dieses Minimum wird angenommen für alle x mit

$$V^\top x = \left(\frac{u_1^\top b}{\sigma_1}, \frac{u_2^\top b}{\sigma_2}, \ldots, \frac{u_r^\top b}{\sigma_r}, \underbrace{a_{r+1}, \ldots, a_n}_{\text{egal!}} \right)^\top$$

Unter allen Lösungen hat x^+ minimale Euklidische Norm (Satz 4.8.2(iv) im Fall $\underline{m \geq n}$, jetzt aber m, n beliebig).

$$\begin{aligned}
\|x\|_2^2 = \left\|v^\top x\right\|_2^2 &= \sum_{i=1}^{r} \left(\frac{u_i^\top b}{\sigma_i} \right)^\top + \sum_{j=r+1}^{n} a_i^2 \\
\Rightarrow x^\top &= V \cdot \left(\frac{u_1^\top b}{\sigma_1}, \ldots, \frac{u_r^\top b}{\sigma_r}, 0, \ldots, 0 \right)^\top \\
&= V \cdot \left(\begin{array}{ccc|c} \frac{1}{\sigma_1} & & & \\ & \ddots & & \\ & & \frac{1}{\sigma_r} & \\ \hline & & & 0 \end{array} \right) \cdot U^\top b \\
&= V \cdot \Sigma^+ \cdot U^\top b
\end{aligned}$$

Definition 4.8.3. Zu $A \in \mathbb{R}^{m \times n}$ und $\mathrm{Rg}(A) = r$ sei $U^\top A V = \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_{\min\{m,n\}}, 0, \ldots, 0)$ eine nach Satz 4.8.5 existierende Singulärwertzerlegung mit $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_{\min\{m,n\}} = 0$. Mit der Matrix

$$\Sigma^+ = \left(\begin{array}{ccc|c} \frac{1}{\sigma_1} & & & \\ & \ddots & & \\ & & \frac{1}{\sigma_r} & \\ \hline & & & 0 \end{array} \right)$$

heißt

$$A^+ = V \Sigma^+ U^\top \in \mathbb{R}^{n \times m}$$

Pseudoinverse (bzw. Moore-Penrose-Inverse).

Satz 4.8.6. *Sei $A \in \mathbb{R}^{m \times n}$, $r = \mathrm{Rg}(A)$. Dann gilt:*

i) Die eindeutig bestimmte Lösung x^+ minimaler Norm des Ausgleichsproblems

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2$$

ist gegeben durch $x^+ = A^+ b$.

ii) A^+ ist eindeutig durch A bestimmt.

iii) $\mathrm{Rg}(A) = n \Rightarrow A^+ = (A^\top A)^{-1} A^\top$.

iv) $m = n$, A regulär $\Rightarrow A^+ = A^{-1}$.

v) Es gelten die Moore-Penrose Bedingungen:

$$\begin{aligned}
AA^+ &= (AA^+)^\top \\
A^+A &= (A^+A)^\top \\
AA^+A &= A \\
A^+AA^+ &= A^+.
\end{aligned}$$

vi) Erfüllt B die vier Moore-Penrose Bedingungen

$$\begin{aligned}
AB &= (AB)^\top \\
BA &= (BA)^\top \\
ABA &= A \\
BAB &= B,
\end{aligned}$$

dann ist $B = A^+$.

Beweis.

i) eben gezeigt.

ii) Die Abbildung $\Phi : \mathbb{R}^m \to \mathbb{R}^n$, $b \mapsto x^+$ ist wohldefiniert (zu jedem b existiert genau ein x^+) und linear ($x^+ = A^+ b$). Zu Φ gehört die eindeutig bestimmte Matrix A^+.

iii) $\mathrm{Rg}(A) = n \Rightarrow m \geq n$. Satz 4.8.2 (iii)$\Rightarrow$ Normalgleichung $A^\top Ax = A^\top b$. Eindeutig bestimte Lösung; $\mathrm{Rg}(A^\top A) = \mathrm{Rg}(A) = n$. Es folgt die Existenz von $(A^\top A)^{-1}$ und $x = (A^\top A)^{-1} A^\top b$. Mit $x^+ = x = A^+ b$ folgt wg. Eindeutigkeit $(A^\top A^{-1})A^\top = A^\top$

iv) $\mathrm{Rg}(A) = n = m \Rightarrow A^{-1}$ existiert.

(iii)$\Rightarrow A^\top = (A^\top A)^{-1} A^\top = A^{-1}(A^\top)^{-1} A^\top = A^{-1}$

v) $A = U\Sigma V^\top, \quad A^+ = V\Sigma^+ U^\top$

$$\Rightarrow AA^+ = U\Sigma V^\top V\Sigma^+ U^\top = U\Sigma\Sigma^+ U^\top$$
$$= U \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \\ \hline & & 0 \end{pmatrix} U^\top \Rightarrow AA^\top \text{ symmetrisch}$$

$A^+ AA^+ = A^+$ analog.

vi) Wenn B und C die vier Moore-Penrose Bedingungen erfüllen, dann gilt

$$\begin{aligned} B = BAB &= B(AB) = B(AB)^\top = BB^\top A^\top \\ &= B(B^\top A^\top)(C^\top A^\top) \\ &= B(AB)^\top (AC)^\top \\ &= (BAB)AC \\ &= BAC \\ C = (CA)C &= (CA)^\top C = A^\top C^\top C = A^\top B^\top A^\top C^\top C \\ &= (BA)^\top (CA)^\top C \\ &= B(ACA)C \\ &= BAC \end{aligned}$$

$\Rightarrow B = C$

□

Fazit: Die Pseudoinverse kann auf drei Weisen erklärt werden:

i) Matrix der Abbildung $b \mapsto x^+$;

ii) $A^+ = V\Sigma^+ U^\top$;

iii) Durch die Moore-Penrose-Bedingungen.

Satz 4.8.7. *Sei $A \in \mathbb{R}^{m\times n}$ mit Pseudoinverser $A^+ \in \mathbb{R}^{n\times m}$. Dann hat $A^\top$ die Pseudoinverse $(A^+)^\top$. Kurz: $(A^\top)^+ = (A^+)^\top$ (vgl. $(A^{-1})^\top = (A^\top)^{-1} =: A^{-\top}$)*

Beweis. Der Beweis wird dem Leser als Übung überlassen. □

Satz 4.8.8. *Sei $A \in \mathbb{R}^{m\times n}$. Dann ist*

$$x \mapsto AA^+ x \quad \textit{bzw.} \quad x \mapsto A^+ Ax$$

die orthogonale Projektion auf $\mathrm{Bild}(A)$ *bzw.* $\mathrm{Bild}(A^+)$.

BEWEIS. Sei $P_A := AA^+$. Aus der ersten Moore-Penrose Bedingung folgt $P_A = P_A^\top$. Außerdem ist $P_A P_A = AA^+AA^+ = AA^+ = P_A$. Somit ist P_A Projektion.
[Projektion: $x = P_A y \Rightarrow P_A x = P_A P_A y = P_A y = x$]
P_A ist orthogonal.
[$x - P_A x \perp P_A x$ bzw. $(P_A x)^\top (x - P_a x) = x^\top P_A (x - P_A x) = x^\top P_A x - x^\top P_A P_A x = 0$]

Noch zu zeigen: Bild $P_A \subseteq$ Bild A:

$$x \in \text{Bild}\, P_A \Rightarrow x = P_A y = A \underbrace{(A^+ y)}_{=:z} = Az \Rightarrow x \in \text{Bild}\, A$$

Bild $A \subseteq$ Bild P_A

$$x \in \text{Bild}\, A \Rightarrow x = Ay = AA^\top (Ay) \in \text{Bild}\, P_A$$

$P_{A^\top}$ ist orthogonale Projektion auf Bild $A^\top$. Zu zeigen: $P_{A^+} = (A^\top)(A^\top)^+ = A^+A$

$$(A^\top)(A^\top)^+ = A^\top (A^+)^\top = (A^+A)^\top = A^+A$$

□

KAPITEL 5

Iterative Lösungen linearer Gleichungssysteme

5.1 Das Gesamt- und Einzelschrittverfahren

Betrachte spezielles Gleichungssystem:

$$x = Tx + u \qquad T \in \mathbb{K}^{n\times n}, u \in \mathbb{K}^n$$

Iterationsvorschrift:

$$x^{(k+1)} = Tx^{(k)} + u \qquad k = 0, 1, \ldots \qquad x^{(0)} \in \mathbb{K}^n$$
$$\Phi : \mathbb{K}^n \to \mathbb{K}^n \qquad \Phi : x \mapsto Tx + u$$

Φ ist kontrahierend, wenn für ein (passendes) Matrixnorm-Vektornorm-Paar $N, \|\ \|$ gilt:

$$\|\Phi(x) - \Phi(y)\| = \|T(x - y)\| \le N(T)\,\|x - y\|, \qquad N(T) < 1$$

Nach dem Banachschen Fixpunktsatz konvergiert die Iteration für beliebige $x^{(0)} \in \mathbb{K}^n$ gegen den Fixpunkt.
Wie bekommt man nun $Ax = b$ auf auf Fixpunktform?
1. Methode

$$A = D - B \qquad D := \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn}) \qquad B := D - A$$

Hier ist es wichtig, daß $a_{kk} \neq 0$ für $k = 1, \ldots, n$ gilt.

$$\begin{aligned} Ax = b &\Leftrightarrow Dx - Bx = b \\ &\Leftrightarrow x = D^{-1}Bx + D^{-1}b \\ &\Rightarrow T = D^{-1}B \qquad u = D^{-1}b \end{aligned}$$

2. Methode

$$A = D - L - U \qquad D := \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn}) \qquad a_{kk} \neq 0 \text{ für } k = 1, \ldots, n$$

L (für lower): strikte untere Dreiecksmatrix.
U (für upper): strikte obere Dreiecksmatrix.

$$\begin{aligned} Ax = b &\Leftrightarrow Dx - Lx = Ux + b \\ &\Leftrightarrow x = (D - L)^{-1}Ux + (D - L)^{-1}b \\ &\Rightarrow T = (D - L)^{-1}U \qquad u = (D - L)^{-1}b \end{aligned}$$

Gesamtschrittverfahren (GSV, Jacobi-Verfahren):

$$A = D - B \qquad D := \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn}) \quad a_{kk} \neq 0 \,\forall k = 1, \ldots, n \quad x^{(0)} \in \mathbb{K}^n$$

($a_{kk} \neq 0$ ist keine starke Einschränkung, da durch Vertauschung erreichbar)
Iterationvorschrift: $x^{(k+1)} = D^{-1}Bx^{(k)} + D^{-1}b$

Ausführlich:

$$\begin{aligned}
x_1^{(k+1)} &= \frac{1}{a_{11}}(0 - a_{12}x_2^{(k)} - \cdots - a_{1n}x_n^{(k)} + b_1) \\
x_2^{(k+1)} &= \frac{1}{a_{22}}(-a_{21}x_1^{(k)} - 0 - \cdots - a_{2n}x_n^{(k)} + b_2) \\
&\vdots \qquad\qquad \vdots \\
x_n^{(k+1)} &= \frac{1}{a_{nn}}(-a_{n1}x_1^{(k)} - a_{n2}x_2^{(k)} - \cdots - 0 + b_n)
\end{aligned}$$

Iterationsmatrix (GSV):

$$T_{\text{GSV}} = D^{-1}B \qquad \text{Jacobi-Matrix (nicht Funktionalmatrix!!!!)}$$

Einzelschrittverfahren (ESV, Gauß-Seidel-Verfahren):

$$A = D - L - U \qquad D := \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn}) \quad a_{kk} \neq 0 \,\forall k = 1, \ldots, n \quad x^{(0)} \in \mathbb{K}^n$$

L: strikte untere Dreiecksmatrix.
U: strikte obere Dreiecksmatrix.
Iterationsvorschrift: $x^{(k+1)} = (D - L)^{-1}Ux^{(k)} + (D - L)^{-1}b$

Ausführlich:

$$\begin{aligned}
x_1^{(k+1)} &= \frac{1}{a_{11}}(0 - a_{12}x_2^{(k)} - \cdots - a_{1n}x_n^{(k)} + b_1) \\
x_2^{(k+1)} &= \frac{1}{a_{22}}(-a_{21}x_1^{(k)} - 0 - \cdots - a_{2n}x_n^{(k)} + b_2) \\
&\vdots \qquad\qquad \vdots \\
x_n^{(k+1)} &= \frac{1}{a_{nn}}(-a_{n1}x_1^{(k)} - \cdots - a_{nn-1}x_{n-1}^{(k)} - 0 + b_n) \\
x^{(k+1)} &= (D - L)^{-1}Ux^{(k)} + (D - L)^{-1}b \\
Dx^{(k+1)} &= Lx^{(k+1)} + Ux^{(k)} + b \\
x^{(k+1)} &= D^{-1}(Lx^{(k+1)} + Ux^{(k)} + b)
\end{aligned}$$

Iterationsmatrix (ESV):

$$T_{\text{ESV}} = (D - L)^{-1}U$$

Wann konvergieren diese Verfahren?

5.2 Konvergenz von Iterationsverfahren für lineare Gleichungssysteme

$$\begin{aligned}
x &= Tx + u \\
x^{(k+1)} &= Tx^{(k)} + u \\
e^{(k+1)} &:= x^{(k)} - x \qquad \text{Fehler, } x \text{ Lösung} \\
\Rightarrow e^{(k+1)} &= Te^{(k)} = T^k e^{(0)}
\end{aligned}$$

Konvergenz tritt eine, wenn $\lim_{k\to\infty} T^k = 0$ (Nullmatrix) gilt, d.h.

$$\lim_{k\to\infty} T^k_{ij} = 0 \quad \forall i, j$$

Satz 5.2.1. *Das Iterationsverfahren $x^{(k+1)} = Tx^{(k)} + u$ konvergiert für beliebige Startwerte $x^{(0)} \in \mathbb{K}^n$ genau dann gegen die Lösung von $x = Tx + u$, wenn $\varrho(T) < 1$ (ϱ: Spektralradius) ist.*

Beweis. $\varrho(T) \geq 1 \Rightarrow \exists\, \lambda, v \neq 0 : \quad Tv = \lambda v, \quad |\lambda| \geq 1$

$$e^{(0)} = v \Rightarrow e^{(k+1)} = T^k v = \lambda^k v \quad \text{divergent}$$

($\lambda > 1$ divergent, $\lambda = 1$ auf Sphäre) $\varrho(T) < 1$:

$$x = Tx + v \Leftrightarrow (E - T)x = v$$

Eigenwerte von $E - T$ sind $\{1 - \lambda | \lambda$ Eigenwert von $T\}$. Für $\varrho(T) < 1$ liegen also die Eigenwerte von $E - T$ im Kreis um 1 mit dem Radius *kleiner* als 1. $\Rightarrow 0$ ist kein Eigenwert von $E - T$.

$\Rightarrow E - T$ ist regulär.
$\Rightarrow$ Es gibt genau einen Fixpunkt x.

$$J = \begin{pmatrix} J_1 & & 0 \\ & \ddots & \\ 0 & & J_l \end{pmatrix}, \qquad J_i = \begin{pmatrix} \lambda_i & 1 & & 0 \\ & \ddots & \ddots & \vdots \\ & & \ddots & 1 \\ 0 & & & \lambda_i \end{pmatrix}$$

$$\Rightarrow T^k = S J \underbrace{S^{-1} S}_{E} J S^{-1} \cdots S J S^{-1} = S J^k S^{-1}$$

$$T^k \to 0 \Leftrightarrow J^k \to 0 \Leftrightarrow J_i^k \to 0, \quad i = 1, \ldots, l$$

$$J_i^k = \sum_{\nu=0}^{r_i - 1} \binom{k}{\nu} \lambda_i^{k-\nu} I^k \qquad I := \begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \vdots \\ & & \ddots & 1 \\ 0 & & & 0 \end{pmatrix} \in \mathbb{K}^{r_i \times r_i}, \; J_i \in \mathbb{K}^{r_i \times r_i}$$

$I^k = 0$ für $k \geq r_i$

$$\left| \binom{k}{\nu} \lambda_i^{k-\nu} \right| \leq k^\nu |\lambda_i|^{k-\nu} \to 0 \text{ für } k \to \infty, \quad \nu = 0, \ldots, r_i - 1$$

$J_i^k \to 0$ für $k \to \infty \Rightarrow J^k \to 0$ für $k \to \infty$, also $T^k \to 0$ für $k \to \infty$. □

Lemma 5.2.2. *$A \in \mathbb{K}^{n\times n}$, N Matrixnorm. Es gilt $\varrho(A) \leq N(A)$.*

Beweis. Der Beweis wird dem Leser als Übung überlassen. □

Satz 5.2.3. *Ist A* diagonaldominant*, d.h. gilt*

$$|a_{ii}| > \sum_{\substack{k=1 \\ k\neq i}} |a_{ik}|, \quad i = 1, \ldots, n,$$

dann konvergiert das Gesamtschrittverfahren beim beliebigen Startvektor $x^{(0)}$.

Beweis.

$$T_{\text{GSV}} = D^{-1} B \Rightarrow \|T_{\text{GSV}}\|_\infty = \max_{i=1}^{n} \sum_{j=1, i\neq j}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| < 1$$

□

Bemerkung. $|||T|||$ gibt Maß für die Geschwindigkeit, mit der der Fehler gegen Null geht (falls $|||T||| < 1$).

$$e^{(k)} = T^k e^{(0)} \Rightarrow \|e^{(k)}\| \leq |||T|||^k \|e^{(0)}\|$$

$\|\quad\|$ und $|||\quad|||$ passend!

Satz 5.2.4 (Kriterium von Sassenfeld). *Sei $A = E - L - U$, L, U strikte obere bzw. untere Dreiecksmatrix. Ferner gelte $\sum_{k=1, i\neq k}^{n} |a_{ik}| < 1$ für $k = 1, \ldots, n$. Dann folgt*

$$\| \underbrace{(E-L)^{-1}U}_{T_{\text{ESV}}} \|_\infty \leq \| \underbrace{L+U}_{T_{\text{GSV}}} \| < 1$$

d.h. schnellere Konvergenz für T_{ESV} als für T_{GSV}.

BEWEIS. Nach Satz 4.5.3 ist

$$\|(E-L)^{-1}U\|_\infty = \max_{x\neq 0} \frac{\|(E-L)^{-1}Ux\|_\infty}{\|x\|_\infty}$$

Zu zeigen:

$$\|L+U\|_\infty = \max_{i=1}^{n} \sum_{k=1, k\neq i}^{n} |a_{ik}| < 1 \Rightarrow \|(E-L)^{-1}Ux\|_\infty \leq \|L+U\|_\infty \|x\|_\infty \quad \forall\, x \in \mathbb{K}^n, x \neq 0 \qquad (*)$$

Sei $z := (E-L)^{-1}Ux$. Dann gilt $z = Lz + Ux$ oder komponentenweise

$$z_i = \sum_{k=1}^{i-1} a_{ik} z_k + \sum_{k=i+1}^{n} a_{ik} x_k$$

$$|z_i| \leq \sum_{k=1}^{i-1} |a_{ik}|\,|z_k| + \sum_{k=i+1}^{n} |a_{ik}||x_k|$$

Insbesondere:

$$|z_1| \leq \sum_{k=2}^{n} |a_{1k}|\,|x_k| \leq \|x\|_\infty \sum_{k=2}^{n} |a_{1k}| \leq \|x\|_\infty \, \|L+U\|_\infty$$

Wenn schon gezeigt

$$|z_j| \leq \|x\|_\infty \, \|L+U\|_\infty \qquad j = 1, \ldots, i-1,$$

dann folgt

$$\begin{aligned}
|z_i| &\leq \sum_{k=1}^{i-1} |a_{ik}|\,|z_k| + \sum_{k=i+1}^{n} |a_{ik}\,|x_k| \\
&\leq \sum_{k=1}^{i-1} |a_{ik}\,\|x\|_\infty \, \|L+U\|_\infty + \|x\|_\infty \sum_{k=i+1}^{n} |a_{ik}| \\
&= \|x\|_\infty \left(\sum_{k=1}^{i-1} |a_{ik}| \underbrace{\|L+U\|_\infty}_{<1} + \sum_{k=i+1}^{n} |a_{ik}| \right) \\
&\leq \|x\|_\infty \sum_{\substack{k=1 \\ i\neq k}}^{n} |a_{ik}| \leq \|x\|_\infty \, \|L+U\|_\infty \\
\Rightarrow |z|_\infty &\leq \|x\|_\infty \, \|L+U\|_\infty
\end{aligned}$$

□

Satz 5.2.5. *Ist $A \in \mathbb{C}^{n\times n}$ hermitesch und positiv definit, dann konvergiert das ESV für jeden Startvektor $x^{(0)} \in \mathbb{C}^n$.*

Beweis. Siehe z.B. Fernuni-Skript *Numerische Mathematik I.* □

Bemerkung. Das GSV konvergiert für positiv definite, hermitesche Matrizen nicht unbedingt. Beispiel:

$$A = \begin{pmatrix} 1 & \alpha & \alpha \\ \alpha & 1 & \alpha \\ \alpha & \alpha & 1 \end{pmatrix} \qquad \alpha \in \mathbb{R}$$

A hat die Eigenwerte $1-\alpha$ (doppelt) und $1+2\alpha$, d.h. A ist positiv definit für $-\frac{1}{2} < \alpha < 1$. Die Jacobi-Matrix

$$T_{\mathrm{GSV}} = \begin{pmatrix} 0 & -\alpha & -\alpha \\ -\alpha & 0 & -\alpha \\ -\alpha & -\alpha & 0 \end{pmatrix}$$

hat die Eigenwerte α (doppelt) und -2α.
Folge: Für $\frac{1}{2} < \alpha < 1$ ist A positiv definit aber $\varrho(T_{\mathrm{GSV}}) > 1$.

5.3 Relaxation und Nachiteration

Beim GSV $A = D - B$

$$\begin{aligned} x^{(k+1)} &= D^{-1}Bx^{(k)} + D^{-1}b \\ &= x^{(k)} - D^{-1}(D-B)x^{(k)} + D^{-1}b \\ &= x^{(k)} - D^{-1}(Ax^{(k)} - b) \end{aligned}$$

Relaxationsverfahren:

$$x^{(k+1)} = x^{(k)} - wD^{-1}(Ax^{(k)} - b)$$

w ist der Relaxationsfaktor. $w \in \mathbb{R}$: JOR-Verfahren (Jordan-Over… (keine Ahnung)-Relaxation).

Algorithmus 5.3.1 (GSV mit Relaxation, JOR-Verfahren).

Gegeben: $A \in \mathbb{K}^{n\times n}$, $D = \mathrm{diag}(a_{11}, \ldots, a_{nn})$ *regulär*
$b \in \mathbb{K}^n$, *Startwert* $x^{(k)} \in \mathbb{K}^n$

Iteration: $x^{(k+1)} = \left[(1-w)E + wD^{-1}B\right]x^{(k)} + wD^{-1}b$

Komponentenweise:

$$\begin{aligned} x_1^{(k+1)} &= (1-w)x_1^{(k)} - \frac{w}{a_{11}}a_{12}x_2^{(k)} - \cdots - \frac{w}{a_{11}}a_{1n}x_n^{(k)} + \frac{w}{a_{11}}b_1 \\ x_2^{(k+1)} &= -\frac{w}{a_{22}}a_{21}x_1^{(k)} + (1-w)x_2^{(k)} - \cdots - \frac{w}{a_{22}}a_{2n}x_n^{(k)} + \frac{w}{a_{22}}b_2 \\ &\vdots \\ x_n^{(k+1)} &= -\frac{w}{a_{nn}}a_{n1}x_1^{(k)} - \frac{w}{a_{nn}}a_{n2}x_2^{(k)} - \cdots + (1-w)x_n^{(k)} + \frac{w}{a_{11}}b_n \end{aligned}$$

Iterationsmatrix: $T_{\mathrm{GSV}}(w) = (1-w)E + wD^{-1}B$

Satz 5.3.1. *$T_{\mathrm{GSV}} = T_{\mathrm{GSV}}(1)$ habe die Eigenwerte $\lambda_1, \ldots, \lambda_n$. $x^{(1)}, \ldots, x^{(n)}$ seien die zugehörigen Eigenvektoren. Dann hat $T_{\mathrm{GSV}}(w)$ die Eigenwerte*

$$\mu_i(w) = 1 - w + w\cdot\lambda_i$$

mit zugehörigen Eigenvektoren $x^{(i)}$, $\quad i = 1, \ldots, n$.

Beweis. $T_{\mathrm{GSV}}(w)x^{(i)} = (1-w)x^{(i)} + w\underbrace{D^{-1}B}_{T_{\mathrm{GSV}}}x^{(i)} = (1-w+w\lambda_i)x^{(i)}$ □

Satz 5.3.2. *Die Iterationsmatrix des Gesamtschrittverfahrens besitze die reellen Eigenwerte $-1 < \lambda_1 \leq \cdots \leq \lambda_n < 1$. Dann wird der Spektralradius des JOR-Verfahrens minimal für*

$$w^* = \frac{2}{2-\lambda_1-\lambda_n} \qquad \textit{mit } \varrho(T_{\mathrm{GSV}}(w^*)) = \frac{\lambda_n - \lambda_1}{2-\lambda_1-\lambda_n}$$

BEWEIS.

$$\varrho(T_{\mathrm{GSV}}(w)) = \max_{i=1}^{n} |u_i(w)| = \max\{|u_1(w)|, |u_n(w)|\}$$

In w^* gilt: $1 - w^* + w^*\lambda_n = -1 + w^* - w^*\lambda_1$

$$\Rightarrow w^* = \frac{2}{2 - \lambda_1 - \lambda_n}$$

$$|u_1(w^*)| = |1 - w^* + w^*\lambda_1| = \frac{\lambda_n - \lambda_1}{2 - \lambda_1 - \lambda_n}$$

□

Jetzt ESV mit Relaxation
SOR-Verfahren (successive over relaxation)
Ähnlich wie bei GSV Hilfsgröße

$$\tilde{x}_i^{(k+1)} = \frac{1}{a_{ii}}\left(-\sum_{j=1}^{i-1} a_{ij}x_i^{(j+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + b_i\right)$$
$$x_i^{(k+1)} = (1-w)x_i^{(k)} + w\tilde{x}_i^{(k+1)}, \qquad w \in \mathbb{R},\ w = 1:\ \text{ESV}$$

Mit der Zerlegung $A = D - L - U$ hat man

$$\begin{aligned} a_{ii}x_i^{(k+1)} &= (1-w)a_{ii}x_i^{(k)} + wa_{ii}\tilde{x}_i^{(k+1)} \\ &= (1-w)a_{ii}x_i^{(k)} + w\left(-\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + b_i\right) \end{aligned}$$

bzw. in der Matrix-Vektor-Schreibweise:

$$\begin{aligned} Dx^{(k+1)} &= (1-w)x^{(k)} + w(Lx^{(k+1)} + Ux^{(k)} + b) \\ x^{(k+1)} &= (D - wL)^{-1}[(1-w)D + wU]x^{(k)} + (D - wL)^{-1}wb \end{aligned}$$

Iterartionsmatrix:

$$T_{\mathrm{ESV}}(w) = (D - wL)^{-1}[(1-w)D + wU]$$

(Diese Matrix lieber nicht ausrechnen, wenn nicht besonders gefragt.)

Algorithmus 5.3.2 (ESV mit Relaxation, SOR-Verfahren).

Gegeben: $A = D - L - U$, $D = \mathrm{diag}(a_{11}, \ldots, a_{nn})$ *regulär*
L, U *strikte untere bzw. obere* $\triangle$*-Matrix,* $b \in \mathbb{K}^n$, $x^{(0)} \in \mathbb{K}^n$, *Relaxationswert* w

Iteration: *Komponentenweise:*

$$\tilde{x}_i^{(k+1)} = \frac{1}{a_{ii}}\left(-\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + b_i\right), \quad i = 1, \ldots, n$$
$$x_i^{(k+1)} = (1-w)x_i^{(k)} + w\tilde{x}_i^{(k+1)}, \quad i = 1, \ldots, n$$

Iterationsmatrix: $T_{\mathrm{ESV}}(w) = (D - wL)^{-1}[(1-w)D + wU]$

Satz 5.3.3 (Kahan).

$$\varrho(T_{\mathrm{ESV}}(w)) \geq |1 - w|$$

Gleichheit nur dann, wenn alle Eigenwerte von $T_{\mathrm{ESV}}(w)$ *betragsmäßig gleich* $|1 - w|$ *sind.*
Folge: *Konvergenz des SOR-Verfahrens höchstens für* $w \in (0, 2)$.

Beweis. $\lambda_j(w)$ sei Eigenwert von $T_{\mathrm{ESV}}(w)$.

$$\begin{aligned}
\det\left(T_{\mathrm{ESV}}(w) - \lambda E\right) &= (-1)^n \lambda_n + \cdots + \det\left(T_{\mathrm{ESV}}(w)\right) \\
&= (-1)^n (\lambda - \lambda_1(w)) \cdot \cdots \cdot (\lambda - \lambda_n(w)) \\
\Rightarrow \det\left(T_{\mathrm{ESV}}(w)\right) &= \prod_{i=1}^{n} \lambda_i(w) \\
\det\left(T_{\mathrm{ESV}}(w)\right) &= \det\left((D - wL)^{-1}\right) \cdot \det\left((1-w)D + wU\right) \\
&= \det(D^{-1}) \cdot \det\left((1-w)D\right) \\
&= \frac{1}{\det D}(1-w)^n \cdot \det D \\
\Rightarrow \prod_{i=1}^{n} \lambda_i(w) &= (1-w)^n \quad \Rightarrow \quad \prod_{i=1}^{n} |\lambda_i(w)| = |1-w|^n \\
\Rightarrow \underbrace{\max_{i=1}^{n} |\lambda_i(w)|}_{\varrho\left(T_{\mathrm{ESV}}(w)\right)} &\geq |1-w|
\end{aligned}$$

□

Satz 5.3.4 (Reich-Ostrowski). *Ist $A \in \mathbb{R}^{n\times n}$ symmetrisch, dann gilt:*
Das SOR-Verfahren konvergiert genau dann für alle $w \in (0,2)$, wenn A positiv definit ist.

Ohne Beweis. □

Nachiteration:

Statt $A = LR$ mit numerischer Rechnung nur $A = \tilde{L} \cdot \tilde{R} + F$ mit der Fehlermatrix F, $\|F\|_\infty$ sehr klein. „Lösung“ $\tilde{x}$ aus $\tilde{L}\tilde{R}\tilde{x} = b$. Für reguläres A ist $\tilde{R}$ (und $\tilde{L}$) regulär, wenn $\|F\|_\infty$ genügend klein ist.

$$\begin{aligned}
Ax = b \quad &\Leftrightarrow \quad (\tilde{L}\tilde{R} + F)x = b \qquad &&\Leftrightarrow \quad \tilde{L}\tilde{R}x = -Fx + b \\
&\Leftrightarrow \quad \tilde{L}\tilde{R}x = -Fx + \tilde{L}\tilde{R}\tilde{x} \qquad &&\Leftrightarrow \quad x = \tilde{x} - \tilde{R}^{-1}\tilde{L}^{-1}Fx
\end{aligned}$$

Weil $Fx \approx F\tilde{x}$ gilt, ist

$$\begin{aligned}
x &\approx \tilde{x} - \tilde{R}^{-1}\tilde{L}^{-1}F\tilde{x} \\
&= \tilde{x} - \tilde{R}^{-1}\tilde{L}^{-1}(A\tilde{x} - \tilde{L}\tilde{R}\tilde{x}) \\
&= \tilde{x} - \tilde{R}^{-1}\tilde{L}^{-1}(A\tilde{x} - b)
\end{aligned}$$

Zu erwarten ist, daß $\tilde{x} - \tilde{R}^{-1}\tilde{L}^{-1}(A\tilde{x} - b)$ bessere Näherung an x ist als $\tilde{x}$ selbst.

Iterationsvorschrift der Nachiteration:

$$x^{(k+1)} = x^{(k)} - \tilde{R}^{-1}\tilde{L}^{-1}(Ax^{(k)} - b)$$

Iterationsmatrix:

$$T_{\mathrm{N}} = E - \tilde{R}^{-1}\tilde{L}^{-1}A$$

Weil $Ax^{(k)} \approx b$ gilt, tritt Stellenauslöschung bei $Ax^{(k)} - b$ ein. Daher Rechnung mit doppelter Stellenzahl bei

$$r^{(k)} := Ax^{(k)} - b \qquad \text{(sonst normal weiter)}$$

Danach berechnet man die Korrektur:

$$\begin{aligned}
z^{(k)} &= \tilde{R}^{-1}\tilde{L}^{-1}r^{(k)} = x^{(k)} - x^{(k+1)} \\
\tilde{L}v^{(k)} &= r^{(k)} \quad \leftarrow \text{ Vorwärtseinsetzen} \\
\tilde{R}z^{(k)} &= v^{(k)} \quad \leftarrow \text{ Rückwärtseinsetzen}
\end{aligned}$$

Algorithmus 5.3.3 (Nachiteration).

Gegeben: $A \in \mathbb{K}^{n\times n}$ *regulär*
$\tilde{L}\tilde{R}$ *sei LR-Zerlegung von* $A - F$, $\|F\|_\infty$ *sehr klein,*
$\tilde{x}$ *sei Lösung von* $\tilde{L}\tilde{R}\tilde{x} = b$

Iteration: $x^{(0)} := \tilde{x}$. *Für* $k = 0, 1, 2, \ldots$:

i) $r^{(k)} = Ax^{(k)} - b$ *Doppelte Genauigkeit*

ii) $\tilde{L}v^{(k)} = r^{(k)}$ $\leftarrow$ *Vorwärtseinsetzen*
$\tilde{R}z^{(k)} = v^{(k)}$ $\leftarrow$ *Rückwärtseinsetzen*

iii) $x^{(k+1)} = x^{(k)} - z^{(k)}$

solange bis $\|x^{(k+1)} - x^{(k)}\|_\infty < \varepsilon$

KAPITEL 6

Eigenwertprobleme

Motivation:

Gesucht ist eine Folge $\{a_n\}_{n=0}^{\infty}$, $a_n \in \mathbb{N}$, $a_0 = 1$, so daß $\frac{1}{2}(a_n^2 + a_{n+1}^2)$ Quadratzahl ist für alle $n \in \mathbb{N}_0$.
Erstes Tripel $\frac{1}{2}(1^2 + 7^2) = 5^2$.
Wenn Tripel

$$a^2 + b^2 = 2c^2 \qquad (*)$$

gefunden, wie findet man $x, y \in \mathbb{N}$:

$$b^2 + x^2 = 2y^2 \qquad (**)$$

Die Differenzbildung auf beiden Seiten liefert:

$$x^2 - a^2 = 2(y^2 - c^2)$$
$$(x + a)(x - a) = 2(y + c)(y - c)$$

Ansatz:

$$\begin{array}{l} x + a = 2(y - c) \\ x - a = y + c \end{array} \quad \Leftrightarrow \quad \begin{array}{l} y = 2a + 3c \\ x = 3a + 4c \end{array} \quad \Rightarrow b^2 + x^2 = 2y^2$$

$$\begin{pmatrix} b \\ x \\ y \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 3 & 0 & 4 \\ 2 & 0 & 3 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

allgemein

$$\begin{pmatrix} a_{n+1} \\ b_{n+1} \\ c_{n+1} \end{pmatrix} = A \begin{pmatrix} a_n \\ b_n \\ c_n \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & 0 \\ 3 & 0 & 4 \\ 2 & 0 & 3 \end{pmatrix}, \quad \begin{pmatrix} a_n \\ b_n \\ c_n \end{pmatrix} = A^n \begin{pmatrix} 1 \\ 7 \\ 5 \end{pmatrix}$$

Die Frage ist nun: wie hängt a_n von n ab und wie ist das Wachstum von a_n für $n \to \infty$?
A hat die Eigenwerte:

$$\begin{array}{rll} -1 & \text{mit dem Eigenvektor} & (-2, 2, 1)^\top; \\ 2 + \sqrt{3} & \text{mit dem Eigenvektor} & (\sqrt{3} - 1, \sqrt{3} + 1, 2)^\top; \\ 2 - \sqrt{3} & \text{mit dem Eigenvektor} & (\sqrt{3} + 1, \sqrt{3} - 1, -2)^\top. \end{array}$$

$$M := \begin{pmatrix} -2 & \sqrt{3} - 1 & \sqrt{3} + 1 \\ 2 & \sqrt{3} + 1 & \sqrt{3} - 1 \\ 1 & 2 & -2 \end{pmatrix}, \quad \Rightarrow A \cdot M = \begin{pmatrix} -2 & \sqrt{3} - 1 & \sqrt{3} + 1 \\ 2 & \sqrt{3} + 1 & \sqrt{3} - 1 \\ 1 & 2 & -2 \end{pmatrix} \begin{pmatrix} -1 & 0 & 0 \\ 0 & 2 + \sqrt{3} & 0 \\ 0 & 0 & 2 - \sqrt{3} \end{pmatrix}$$

$$= M \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}$$

$$M^{-1}AM = D \quad \Rightarrow \quad \underbrace{(M^{-1}AM)\cdot\dots\cdot(M^{-1}AM)}_{n \text{ Faktoren}} = M^{-1}A^nM$$

$$\Rightarrow \quad M^{-1}\begin{pmatrix}a_n\\b_n\\c_n\end{pmatrix} = \underbrace{M^{-1}AM}_{D^n}M^{-1}\begin{pmatrix}1\\7\\5\end{pmatrix}$$

$$\begin{pmatrix}u_n\\v_n\\w_n\end{pmatrix} := M^{-1}\begin{pmatrix}a_n\\b_n\\c_n\end{pmatrix}, \qquad \begin{pmatrix}u_n\\v_n\\w_n\end{pmatrix} = D^n\begin{pmatrix}u_n\\v_n\\w_n\end{pmatrix}, \qquad M^{-1}\begin{pmatrix}1\\7\\5\end{pmatrix} = \begin{pmatrix}u_0\\v_0\\w_0\end{pmatrix}$$

$$\begin{pmatrix}u_n\\v_n\\w_n\end{pmatrix} = \begin{pmatrix}(-1)^n & 0 & 0\\ 0 & (2+\sqrt{3})^n & 0\\ 0 & 0 & (2-\sqrt{3})^n\end{pmatrix}\begin{pmatrix}u_0\\v_0\\w_0\end{pmatrix} \quad \Rightarrow \quad \begin{array}{ll} u_n = (-1)^n u_0 & \text{alternierend}\\ v_n = (2+\sqrt{3})^n v_0 & (!)\\ w_n = (2-\sqrt{3})^n w_0 & (<1)\end{array}$$

(!) ist der einzige Faktor, der Ausschlag gibt, da nur er richtig wächst.

$$\begin{pmatrix}a_n\\b_n\\c_n\end{pmatrix} = M\begin{pmatrix}u_n\\v_n\\w_n\end{pmatrix} = \begin{pmatrix}-2 & \sqrt{3}-1 & \sqrt{3}+1\\ 2 & \sqrt{3}+1 & \sqrt{3}-1\\ 1 & 2 & -2\end{pmatrix}\begin{pmatrix}(-1)^n u_0\\(2+\sqrt{3})^n v_0\\(2-\sqrt{3})^n w_0\end{pmatrix}$$

Die mittlere Komponenete ist die explizite Darstellung von a_n in Abhängigkeit von n. a_n wächst wie $(\sqrt{3}-1)(2+\sqrt{3})^n v_0$.

$$\begin{pmatrix}u_0\\v_0\\w_0\end{pmatrix} = M^{-1}\begin{pmatrix}1\\7\\5\end{pmatrix} = \begin{pmatrix}-\frac{1}{3} & \frac{1}{3} & -\frac{1}{3}\\ \frac{\sqrt{3}+1}{12} & \frac{\sqrt{3}-1}{12} & \frac{1}{3}\\ \frac{\sqrt{3}-1}{12} & \frac{\sqrt{3}+1}{12} & -\frac{1}{3}\end{pmatrix}\begin{pmatrix}1\\7\\5\end{pmatrix}$$

6.1 Grundbegriffe aus der Algebra

$A \in \mathbb{C}^{n\times n}$, $\lambda \in \mathbb{C}$ ist Eigenwert, wenn ein $x \in \mathbb{C}^n$, $x \neq 0$ existiert mit $Ax = \lambda x$. x heißt Eigenvektor zum Eigenwert λ.
$U := \{x | Ax = \lambda x\}$ heißt Eigenraum zu Eigenwert λ. Die Vielfachheit von λ ist $\dim U$. U ist Unterraum von $\mathbb{C}^n$, der unter A invariant ist, d.h. $x \in U \Rightarrow Ax \in U$. Die Eigenwerte sin die Nullstellen des charakteristischen Polynoms:

$$p(\lambda) := \det(A - \lambda E) = \begin{vmatrix} a_{11}-\lambda & \cdots & a_{1n}\\ \vdots & \ddots & \vdots\\ a_{n1} & \cdots & a_{nn}-\lambda\end{vmatrix}$$

$$= (a_{11}-\lambda)\cdot\dots\cdot(a_{nn}-\lambda) + \cdots = (-\lambda)^n + \text{ Terme mit } \lambda \text{ mit niedriger Ordnung}$$

$$= (-\lambda)^n + (-\lambda)^{n-1}(a_{11}+\dots+a_{nn}) + \cdots + \det(A)$$

$\lambda_1, \dots, \lambda_n$ Eigenwerte von A.

$$p(\lambda) = (-1)^n(\lambda-\lambda_1)\cdot\dots\cdot(\lambda-\lambda_n)$$
$$= (-1)^n\lambda^n + (-1^n)(-\lambda_1 - \cdots - \lambda_n) + \cdots + \lambda_1\cdot\dots\cdot\lambda_n$$

Durch Koeffizientenvergleich ergibt sich

$$\sum_{k=1}^{n} a_{kk} = \sum_{k=1}^{n}\lambda_k \qquad \det A = \prod_{k=1}^{n}\lambda_k$$

Sei $\Phi : \mathbb{C}^n \to \mathbb{C}^n$. Zu Φ gehöre A. Koordinatentransformation $y \mapsto Ty \Rightarrow$ zu Φ gehört $T^{-1}AT$.

$$Ax = \lambda x,\ x = Ty \quad \Rightarrow \quad T^{-1}ATT^{-1}x = \lambda T^{-1}x$$

λ ist auch Eigenwert von $T^{-1}AT$ mit dem Eigenvektor $T^{-1}x$.

Satz 6.1.1 (Komplexe Schur-Zerlegung). *Sei $A \in \mathbb{C}^{n\times n}$. Dann ex. eine unitäre Matrix $U \in \mathbb{C}^{n\times n}$*

$$U^{\mathsf{H}}AU = \begin{pmatrix} \lambda_1 & \cdots & * \\ & \ddots & \vdots \\ 0 & & \lambda_n \end{pmatrix}$$

Dabei sind $\lambda_1, \ldots, \lambda_n$ die Eigenwerte von A (nicht notwendig verschieden).

Beweis. Mit Induktion nach n. $n = 1$ ist trivial. „$n \to n+1$“ : λ_1 sei Eigenwert mit dem Eigenvektor v_1. O.B.d.A. sei $\|v_1\| = 1$. Man ergänze v_1 zu einer Orthonormalbasis $\{v_1, \ldots, v_n\}$ von $\mathbb{C}^n$. $V = (v_1, \ldots, v_n) \in \mathbb{C}^{n\times n}$, $V^{\mathsf{H}}V = E$ und

$$V^{\mathsf{H}}AV = \left(\begin{array}{c|c} \lambda_1 & y^{\top} \\ \hline 0 & A_1 \end{array}\right) \qquad A_1 \in \mathbb{C}^{(n-1)\times(n-1)}$$

Nach Induktionsvoraussetzung existiert ein $W \in \mathbb{C}^{(n-1)\times(n-1)}$, W unitär, mit

$$W^{\mathsf{H}}A_1W = \begin{pmatrix} \lambda_2 & \cdots & * \\ & \ddots & \vdots \\ 0 & & \lambda_n \end{pmatrix} \qquad U := V\left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & W \end{array}\right)$$

$$\Rightarrow U^{\mathsf{H}}U = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & W^{\mathsf{H}} \end{array}\right) V^{\mathsf{H}}V \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & W \end{array}\right) = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & W^{\mathsf{H}}W \end{array}\right) = E$$

$$U^{\mathsf{H}}AU = \begin{pmatrix} 1 & 0 \\ 0 & W^{\mathsf{H}} \end{pmatrix} V^{\mathsf{H}}AV \begin{pmatrix} 1 & 0 \\ 0 & W \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & W^{\mathsf{H}} \end{pmatrix}\begin{pmatrix} \lambda_1 & y^{\top} \\ 0 & A_1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & W \end{pmatrix}$$

$$= \left(\begin{array}{c|c} \lambda_1 & y^{\top} \\ \hline 0 & R_1 \end{array}\right) \quad \Rightarrow \text{ Behauptung}$$

□

Satz 6.1.2 (Reelle Schur-Zerlegung). *Sei $A \in \mathbb{R}^{n\times n}$. Dann ex. eine orthogonale Matrix $Q \in \mathbb{R}^{n\times n}$*

$$Q^{\top}AQ = \begin{pmatrix} R_{11} & \cdots & R_{1s} \\ & \ddots & \vdots \\ 0 & & R_{ss} \end{pmatrix},$$

wobei die R_{kk} reelle 1×1 oder 2×2 Matrizen sind. $R_{kk} = (\lambda_k) \in \mathbb{R}^{1\times 1} \Rightarrow \lambda_k$ Eigenwert von A. $R_{kk} \in \mathbb{R}^{2\times 2}$ $\Rightarrow$

$$R_{kk} = \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}, \quad \alpha \pm i\beta \text{ Eigenwert von } A.$$

Beweis. „leicht“ □

Ist $A = A^{\mathsf{H}}$, dann ist A hermitesch. Es folgt für die Zerlegung aus Satz 6.1.1:
$(U^{\mathsf{H}}AU)^{\mathsf{H}} = U^{\mathsf{H}}A^{\mathsf{H}}U = U^{\mathsf{H}}AU$

$$\Rightarrow \begin{pmatrix} \lambda_1 & \cdots & * \\ & \ddots & \vdots \\ 0 & & \lambda_n \end{pmatrix} = \begin{pmatrix} \overline{\lambda}_1 & & 0 \\ \vdots & \ddots & \\ * & \cdots & \overline{\lambda}_n \end{pmatrix} = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \quad \text{und } \lambda_k = \overline{\lambda}_k, \text{ d.h. } \lambda_k \in \mathbb{R}$$

Satz 6.1.3 (Eigenwertzerlegung hermitescher Matrizen). *Zu jeder hermiteschen Matrix $A \in \mathbb{C}^{n\times n}$ gibt es eine unitäre Matrix $U \in \mathbb{C}^{n\times n}$ mit*

$$U^{\mathsf{H}}AU = D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n).$$

Die i-te Spalte u_i von U ist Eigenvektor zum Eigenwert $\lambda_i \in \mathbb{R}$ von A.

$$Au_i = \lambda_i u_i$$

Bemerkung. Man nennt A *diagonalisierbar* wegen $U^{\mathsf{H}}AU = D$. $u_1, \ldots, u_n$ ist eine Orthonormalbasis von Eigenvektoren von A

$$A = UDU^{\mathsf{H}} = (u_1, \ldots, u_n)\begin{pmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_n\end{pmatrix}\begin{pmatrix}u_1^{\mathsf{H}}\\ \vdots \\ u_n^{\mathsf{H}}\end{pmatrix} = \sum_{k=1}^{n}\lambda_k u_k u_k^{\mathsf{H}}$$

Diese Darstellung heißt die Eigenwertzerlegung von A.

Definition 6.1.1. Eine Matrix $A \in \mathbb{C}^{n\times n}$ heißt *normal*, wenn gilt:

$$AA^{\mathsf{H}} = A^{\mathsf{H}}A.$$

Satz 6.1.4. *$A \in \mathbb{C}^{n\times n}$ ist genau dann normal, wenn es eine unitäre Matrix $U \in \mathbb{C}^{n\times n}$ gibt mit*

$$U^{\mathsf{H}}AU = \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \quad \lambda_k \in \mathbb{C},\ k = 1, \ldots, n$$

Folge: *Normale Matrizen (und nur sie) sind unitär diagonalisierbar. Sie besitzen n Eigenvektoren, die eine Orthonormalbasis des $\mathbb{C}^n$ bilden.*

BEWEIS. Schur-Zerlegung

$$U^{\mathsf{H}}AU = \begin{pmatrix}\lambda_1 & \cdots & *\\ & \ddots & \vdots \\ 0 & & \lambda_n\end{pmatrix} = R = (r_{jk})_{j,k=1}^{n}$$

$$A^{\mathsf{H}}A = AA^{\mathsf{H}} \qquad R^{\mathsf{H}}R = U^{\mathsf{H}}A^{\mathsf{H}}UU^{\mathsf{H}}AU = U^{\mathsf{H}}\underbrace{A^{\mathsf{H}}A}U$$
$$RR^{\mathsf{H}} = U^{\mathsf{H}}AUU^{\mathsf{H}}A^{\mathsf{H}}U = U^{\mathsf{H}}\underbrace{AA^{\mathsf{H}}}U \quad \Rightarrow \quad R^{\mathsf{H}}R = RR^{\mathsf{H}}$$

Erste Zeile, erste Spalte von $R^{\mathsf{H}}R = RR^{\mathsf{H}}$

$$\bar{r}_{11}r_{11} = \sum_{k=1}^{n} r_{1k}\bar{r}_{1k}$$
$$|\lambda_1|^2 = |\lambda_1|^2 + \sum_{k=1}^{n}|r_{1k}|^2 \quad \Rightarrow \quad \sum_{k=1}^{n}|r_{1k}|^2 = 0$$
$$\Rightarrow r_{12} = \cdots = r_{1n} = 0$$

Analog für den Rest. □

6.2 Reduktion auf Tridiagonal- bzw. Hessenberg-Gestalt

Definition 6.2.1. Eine Matrix $B = (b_{ij})_{i,j=1}^{n} \in \mathbb{C}^{n\times n}$ der Form

$$\begin{pmatrix}b_{11} & b_{12} & \cdots & b_{1n}\\ b_{21} & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & b_{n-1\,n}\\ 0 & 0 & b_{n\,n-1} & b_{nn}\end{pmatrix} \qquad \text{d.h. } b_{ij} = 0 \text{ für } i > j+1$$

heißt (obere) *Hessenbergmatrix*. B heißt *nicht-zerfallend*, wenn $b_{j+1\,j} \neq 0$ für $j = 1, \ldots, n-1$ gilt.

Bemerkung. Gilt $B^{\mathsf{H}} = B$, dann ist B *tridiagonal*.

Satz 6.2.1. *Zu $A \in \mathbb{C}^{n\times n}$ existieren eine (obere) Hessenbergmatrix B und Householdermatrizen $H_1, \ldots, H_{n-2} \in \mathbb{C}^{n\times n}$, so daß*

$$B = \underbrace{H_{n-2}\cdot\cdots\cdot H_1}_{U^{\mathsf{H}}} A \underbrace{H_1\cdot\cdots\cdot H_{n-2}}_{U}$$

gilt.

BEWEIS. **Schritt 1:** $\tilde{w}_1 \in \mathbb{C}^{n-1}$, $\|\tilde{w}_1\|_1 = 1$, so daß

$$\underbrace{(E_{n-1} - 2\tilde{w}_1\tilde{w}_1^{\mathsf{H}})}_{H_{\tilde{w}_1}} \begin{pmatrix} a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} = b_{21} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

gilt.

$$w_1 := \begin{pmatrix} 0 \\ \tilde{w}_1 \end{pmatrix} \in \mathbb{C}^n, \quad H_1 = E_n - 2w_1w_1^{\mathsf{H}} = \begin{pmatrix} 1 & 0 \\ 0 & H_{\tilde{w}_1} \end{pmatrix}$$

Betrachte

$$\begin{aligned} H_1AH_1 &= \begin{pmatrix} 1 & 0 \\ 0 & H_{\tilde{w}_1} \end{pmatrix} A \begin{pmatrix} 1 & 0 \\ 0 & H_{\tilde{w}_1} \end{pmatrix} \\ &= \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ b_{21} & & & \\ 0 & & M & \\ \vdots & & & \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & H_{\tilde{w}_1} \end{pmatrix} \\ &= \begin{pmatrix} a_{11} & (a_{12} \cdots a_{1n})H_{\tilde{w}_1} \\ b_{21} & \\ 0 & MH_{\tilde{w}_1} \\ \vdots & \end{pmatrix} = \left(\begin{array}{cc|c} a_{11} & \tilde{a}_{12} & * \\ b_{21} & \tilde{a}_{22} & * \\ \hline 0 & \tilde{a}_{32} & \\ \vdots & \vdots & * \\ 0 & \tilde{a}_{n2} & \end{array}\right) \end{aligned}$$

Nach k Schritten

KOMMT NOCH

□

Bemerkung. Der Beweis ist konstruktiv. Man nennt das resultierende Verfahren „Verfahren von Hyman“ oder auch „Reduktionsverfahren von Householder“.

Wenn $A = A^{\mathsf{H}}$, dann auch $B = B^{\mathsf{H}}$, denn

$$(H_{n-2} \cdot \cdots \cdot H_1AH_1 \cdot \cdots \cdot H_{n-2})^{\mathsf{H}} = H_{n-2} \cdot \cdots \cdot H_1AH_1 \cdot \cdots \cdot H_{n-2} = B$$

Für hermitesche Matrizen liefert Satz 6.2.1 eine Tridiagonalmatrix.

Lemma 6.2.2. *Sei B eine zerf. Hessenbergmatrix $B = (b_{ij}) \in \mathbb{C}^{n\times n}$, $b_{l+1\,l} = 0$ für ein $l \in \{1, \ldots, n-1\}$. Dann gilt*

$$\det(B - \lambda E_n) = \det(B_l - \lambda E_l) \cdot \det(C_{n-l} - \lambda E_{n-l})$$

$$\left(\begin{array}{ccc|ccc} b_{11}-\lambda & * \cdots & * & * & \cdots & * \\ b_{21} & \ddots & \vdots & \vdots & \ddots & \vdots \\ & \ddots & * & \vdots & & \vdots \\ 0 & b_{l\,l-1} & b_{ll}-\lambda & * & \cdots & * \\ \hline 0 & \cdots & 0 & & & \\ \vdots & \ddots & \vdots & & C_{n-l} & \\ 0 & \cdots & 0 & & & \end{array}\right) \qquad C_{n-l} = \begin{pmatrix} b_{l+1\,l+1}-\lambda & * \cdots & * \\ b_{l+2\,l+1} & \ddots & \vdots \\ & \ddots & * \\ 0 & b_{n\,n-1} & b_{nn}-\lambda \end{pmatrix}$$

Ab sofort daher B nicht-zerfallend. $x \neq 0$

$$\begin{aligned}(B-\lambda E)x &= 0\\ -q + (b_{11}-\lambda)x_1 + b_{12}x_2 + \cdots + b_{1n}x_n &= 0\\ b_{21}x_1 + (b_{22}-\lambda)x_2 + \cdots + b_{2n}x_n &= 0\\ &\vdots\\ b_{n\,n-1}x_{n-1} + (b_{nn}-\lambda)x_n &= 0\end{aligned}$$

Wenn $q = q(\lambda) = 0$, dann ist λ Eigenwert und $(x_1, \ldots, x_n)^\top$ ist Eigenvektor. Bei gegebenem x_n haben wir ein Gleichungssystem für $q, x_1, \ldots, x_{n-1}$ mit der Koeffizientenmatrix:

$$\tilde{B}(\lambda) = \begin{pmatrix} -1 & b_{11}-\lambda & b_{12} & \cdots & b_{1\,n-1}\\ 0 & b_{21} & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & \ddots & b_{n-2\,n-1}\\ \vdots & & \ddots & \ddots & b_{n-1\,n-1}-\lambda\\ 0 & \cdots & 0 & b_{n\,n-1} & \end{pmatrix}$$

$$\det\left(\tilde{B}(\lambda)\right) = -b_{12}\cdot\cdots\cdot b_{n\,n-1} \neq 0$$

$$\tilde{B}(\lambda)\begin{pmatrix} q\\ x_1\\ \vdots\\ x_{n-1}\end{pmatrix} = x_n\begin{pmatrix} -b_{1n}\\ -b_{2n}\\ \vdots\\ -(b_{nn}-\lambda)\end{pmatrix}$$

$x_n = 0 \Rightarrow a = x_1 = \cdots = x_{n-1} = 0 \Rightarrow x = 0$ (uninteressant)
$x_n \neq 0$, dann o.B.d.A. $x_n = 1$.

Problem: Finde λ so, daß $q(\lambda) = 0$.

Cramersche Regel:

$$\begin{aligned} q(\lambda) &= \frac{-1}{b_{21}\cdot\ldots\cdot b_{n\,n-1}}\det\begin{pmatrix} -b_{1n} & b_{11}-\lambda & \cdots & \cdots & \cdots & b_{1\,n-1}\\ -b_{2n} & b_{21} & \ddots & & & \vdots\\ \vdots & 0 & & \ddots & & \vdots\\ \vdots & \vdots & & & \ddots & \vdots\\ \vdots & \vdots & & & & b_{n-1\,n-1}-\lambda\\ -(b_{nn}-\lambda) & 0 & \cdots & \cdots & 0 & b_{n\,n-1}\end{pmatrix}\\ &= \frac{-1^n}{b_{21}\cdot\ldots\cdot b_{n\,n-1}}\det\begin{pmatrix} b_{11}-\lambda & & & -b_{1n}\\ b_{21} & \ddots & & \\ & \ddots & \ddots & \\ & & b_{n\,n-1} & -(b_{nn}-\lambda)\end{pmatrix}\\ &= \frac{-1^{n+1}}{b_{21}\cdot\ldots\cdot b_{n\,n-1}}\det(B-\lambda E)\end{aligned}$$

Resultat:
Das charakteristische Polynom einer nicht-zerfallenden oberen Hessenbergmatrix hat sie Darstellung

$$\det(B-\lambda E) = (-1)^{n+1}b_{21}\cdot\ldots\cdot b_{n\,n-1}q(\lambda),$$

wobei $q(\lambda)$ durch das Gleichungssystem

$$-qe_1 + (B-\lambda E)x = 0, \quad x_n = 1 \qquad (*)$$

definiert ist. Für $q(\lambda) = 0$ ist λ Eigenwert und $(x_1(\lambda), \ldots, x_{n-1}(\lambda), 1)^\top$ zugehöriger Eigenvektor.

Berechnung von $q(\lambda)$ mit $\frac{1}{2}n(n-1)$ Multiplikationen und n Divisionen. Für das Newton-Verfahren ist noch $q'(\lambda)$ nötig. Dazu (*) nach λ ableiten:

$$\begin{aligned} -q'(\lambda)e_1 - Ex(\lambda) + (B-\lambda E)x'(\lambda) &= 0 \\ -q'(\lambda)e_1 + (B-\lambda E)x'(\lambda) &= x(\lambda), \quad x'_n = 0 \end{aligned}$$

Komponentenweise sieht es dann so aus:

$$\begin{aligned} -q'(\lambda) + (b_{11}-\lambda)x'_1 + b_{12}x'_2 + \ldots + b_{1\,n-1}x'_{n-1} &= x_1 \\ b_{12}x'_1 + (b_{22}-\lambda)x'_2 + \ldots + b_{2\,n-1}x'_{n-1} &= x_2 \\ \ddots \qquad\qquad \vdots & \\ b_{n\,n-1}x'_{n-1} &= x_n \; (=1) \end{aligned}$$

Algorithmus (Hyman).

Gegeben:	*B obere Hessenbermatrix nicht-zerfallend* *Startwert* $\lambda^{(0)}$
Gesucht:	*Verbesserte Näherung* $\lambda^{(1)}$ *an einen Eigenwert von B und* *Näherung an zugehörigen Eigenvektor.*
Schritt 1:	*Löse* $(B-\lambda^{(0)}E)x - qe_1 = 0, \quad x_n = 1$
Schritt 2:	*Löse* $(B-\lambda^{(0)}E)x' - q'e_1 = x, \quad x'_n = 0$
Schritt 3:	*Löse* $\lambda^{(1)} = \lambda^{(0)} - \dfrac{q(\lambda^{(0)})}{q'(\lambda^{(0)})}$

Jetzt ist B eine obere Hessenbergmatrix, nicht-zerfallend *und* hermitesch. Also ist B hermitesche, nicht-zerfallende Tridiagonalmatrix.

$$B_n = \begin{pmatrix} a_1 & b_1 & 0 & 0 \\ \overline{b}_1 & a_2 & \ddots & 0 \\ 0 & \ddots & \ddots & b_{n-1} \\ 0 & 0 & \overline{b}_{n-1} & a_n \end{pmatrix} \qquad a_k \in \mathbb{R},\ b_k \in \mathbb{C}\setminus\{0\}$$

Satz 6.2.3. *Das charakteristische Polynom* $p_n(\lambda) = \det(B_n - \lambda E)$ *läßt sich mit der Rekursion*

$$\begin{aligned} &p_0(\lambda) = 1, \quad p_1(\lambda) = a_1 - \lambda, \\ &p_k(\lambda) = (a_k-\lambda)p_{k-1}(\lambda) - |b_{k-1}|^2 p_{k-2}(\lambda), \quad k = 2,\ldots,n \end{aligned}$$

berechnen.

Beweis.

$$\begin{aligned} p_k(\lambda) &= (a_k-\lambda)p_{k-1}(\lambda) - b_{k-1}\overline{b}_{k-1}p_{k-2}(\lambda) \\ p_1(\lambda) &= \det(a_1-\lambda) = a_1-\lambda \\ p_2(\lambda) &= (a_1-\lambda)(a_2-\lambda) - |b_1|^2 p_0(\lambda) \quad \text{nach Rekursion} \\ &= (a_1-\lambda)(a_2-\lambda) - |b_1|^2 \quad \text{nach Determinantenformel} \;\Rightarrow\; p_0(\lambda) = 1 \end{aligned}$$

□

$$\begin{aligned} &p_k(\lambda) = (a_k-\lambda)p_{k-1}(\lambda) - |b_{k-1}|^2 p_{k-2}(\lambda) \quad k \geq 2 \\ &p_k(\lambda) = (-1)^l\lambda^k + \ldots \;\Rightarrow\; p(\lambda) > 0 \text{ für „}\lambda \to -\infty\text{“} \end{aligned}$$

Satz 6.2.4. $B_n \in \mathbb{C}^{n\times n}$ *sei eine nicht-zerfallende hermitesche Tridiagonalmatrix.* B_k *ihre k-reihige Hauptuntermatrix. Dann besitzt* p_n*, das charakteristische Polynom von* B_k*, k reelle einfache Nullstellen* $\lambda_j^{(k)}$*,* $j = 1,\ldots,k$*, und für* $k = 1,\ldots,n-1$ *trennen die Nullstellen von* p_k *die* p_{k+1}*, d. h.*

$$\lambda_1^{(k+1)} < \lambda_1^{(k)} < \lambda_2^{(k+1)} < \lambda_2^{(k)} < \ldots < \lambda_k^{(k)} < \lambda_{k+1}^{(k+1)}$$