Einführung in die Numerik (Potschka)

Robin Heinemann

26. Oktober 2017

Inhaltsverzeichnis

0 Einführung

Beispiel 0.1 Simulation einer Pendelbewegung
Modellannahmen:

- Masse m an Stange
- keine Reibung
- Stange: Gewicht 0, starr, Länge l
- Auslenkung ϕ

Erste Fehlerquelle: Modellierungsfehler
Modellgleichungen:

$$F_T(\phi) = -m \cdot g \sin \phi$$

Konsistenzcheck:

$$F_T(0) = 0 \qquad \text{(Ruhelage)}$$
$$F_T\left(\frac{\pi}{2}\right) = F_G = -mg$$

Bewegungsgleichungen:

- Weg $s(t)$
- $\frac{\mathrm{d}s}{\mathrm{d}t} =: v(t)$ Geschwindigkeit
- $\frac{\mathrm{d}v}{\mathrm{d}t} =: a(t)$ Beschleunigung

Beziehungen:

- Bogenlänge $s(t) = l\phi(t)$
- 2. Newton'sches Gesetz ($F = ma$)

$$-mg \sin \phi(t) = m\frac{\mathrm{d}}{\mathrm{d}t}v(t) = m\frac{\mathrm{d}^2}{\mathrm{d}t^2}s(t) = ml\frac{\mathrm{d}^2}{\mathrm{d}t^2}\phi(t)$$

$\Longrightarrow$ DGL 2. Ordnung

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\phi(t) = -\frac{g}{l}\sin\phi(t) \quad t \geq 0$$

Für eindeutige Lösung braucht man zwei Anfangsbedingungen:

$$\phi(0) = \phi_0 \quad \frac{\mathrm{d}}{\mathrm{d}t}\phi(0) = u_0$$

Lösung bei kleiner Auslenkung: Linearisiere um $\phi = 0$

$$\sin\phi = \phi - \frac{1}{3!}\phi^3 + \cdots \approx \phi$$
$$\implies \frac{\mathrm{d}^2}{\mathrm{d}t^2}\phi(t) = -\frac{g}{l}\phi(t)$$

Für $u_0 = 0$ findet man mit dem Ansatz $\phi(t) = A\cos(\omega t)$:

$$-\omega^2 A\cos(\omega t) = -\frac{g}{l}A\cos(\omega t)$$

die Lösung:

$$\phi(t) = \phi_0\cos\left(\sqrt{\frac{g}{l}}t\right)$$

Fehlerquelle: Abschneidefehler.
Numerische Lösung:
Setze $u(t) := \frac{\mathrm{d}}{\mathrm{d}t}\phi(t)$

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\phi\\ u\end{pmatrix} = \begin{pmatrix}u\\ -\frac{g}{l}\sin(\phi)\end{pmatrix}$$

Approximation mit Differenzenquotienten

$$\begin{pmatrix}u(t)\\ -\frac{g}{l}\sin\phi(t)\end{pmatrix} = \frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\phi\\ u\end{pmatrix} = \lim_{\Delta t\to 0}\frac{1}{\Delta t}\begin{pmatrix}\phi(t+\Delta t)-\phi(t)\\ u(t+\Delta t)-u(t)\end{pmatrix} \approx \underbrace{\frac{1}{\Delta t}\begin{pmatrix}\phi(t+\Delta t)-\phi(t)\\ u(t+\Delta t)-u(t)\end{pmatrix}}_{\text{> 0, klein}}$$

Fehlerquelle: Diskretisierungsfehler
Auf Gitter $t_n = n\Delta t$ mit Werten $\phi_n = \phi(n\Delta t), u_n = u(n\Delta t)$:

$$\phi_{n+1} = \phi_n + \Delta t u_n, u_{n+1} = u_n - \Delta t\frac{g}{l}\phi_n$$

Kleinerer Diskretisierungsfehler mit zentralen Differenzen:

$$-\frac{g}{l}\sin\phi(t) = \frac{\mathrm{d}^2}{\mathrm{d}t^2}\phi(t) \approx \frac{\phi(t+\Delta t) - 2\phi(t) + \phi(t-\Delta t)}{\Delta t^2}$$

Rekursionsformel:

$$\phi_{n+1} = 2\phi_n - \phi_{n-1} - \Delta t^2\frac{g}{l}\sin\phi_n, n \geq 1$$

mit $\phi_1 = \phi_0 + \Delta t n_0$ (Expliziter Euler)
Letzte Fehlerquelle: Rundungsfehler

1 Fehleranalyse

1.1 Zahldarstellung und Rundungsfehler

Anforderung: Rechnen mit reellen Zahlen auf dem Computer.
Problem: Speicher endlich ($\implies$ endliche Genauigkeit).
Lösung: Gleitkommazahlen, ein **Kompromiss** zwischen:

- Umfang darstellbarer Zahlen
- Genauigkeit

- Geschwindigkeit einfacher Rechenoperationen (+, -, ·, /)

Alternativen:

- Fixkommazahlen
- logarithmische Zahlen
- Rationalzahlen

Definition 1.1 Eine (normalisierte) Gleitkommazahl zur Basis $b \in \mathbb{N}, b \geq 2$, ist eine Zahl $x \in \mathbb{R}$ der Form

$$x = \pm m \cdot b^{\pm e}$$

mit der Mantisse $m = m_1 b^{-1} + m_2 b^{-2} + \ldots \in \mathbb{R}$ und dem Exponenten $e = e_{s-1} b^{s-1} + \cdots + e_0 b^0 \in \mathbb{N}$, wobei $m_i, e_i \in \{0, \ldots, b-1\}$. Für $x \neq 0$ ist die Darstellung durch die Normierungsvorschrift $m \neq 0$ eindeutig. Für $x = 0$ setzt man $m = 0$.

Beispiel 1.2 ($b = 10$)

- m_i: i -te Nachkommastelle der Mantisse
- e: Verschiebt das Komma um e Stellen.

$$0.314 \times 10^1 = 3.14$$
$$0.123 \times 10^6 = 123\,000$$

Auf dem Rechner stehen nur endlich viele Stellen zur Verfügung:
r Ziffern + 1 Vorzeichen für Mantisse m
s Ziffern + 1 Vorzeichen für Exponenten.
Für $x = \pm[m_1 b^{-1} + \cdots + m_r b^{-r}] \cdot b^{\pm[e_{s-1} b^{s-1} + \cdots + e_0 b^0]}$ muss man also nur $(\pm)[m_1 \ldots m_r](\pm)[e_{s-1} \ldots e_0]$ abspeichern. Wählt man $b = 2$, so gilt $m_i, e_i \in \{0, 1\}$ und x kann mit $2 + r + s$ Bits gespeichert werden (Maschinenzahlen). Maschinenzahlen bilden das numerische Gleitkommagitter $A = A(b, r, s)$

Beispiel 1.3 ($b = 2, r = 3, s = 1$)

$$m = \frac{1}{2} + m_2 \frac{1}{4} + m_3 \frac{1}{8} \in \{\frac{4}{8}, \frac{5}{8}, \frac{6}{8}, \frac{7}{8}\}$$
$$e = e_0 \in \{0, 1\}$$

Da A endlich ist, gibt es eine größte/kleinste darstellbare Zahl:

$$\begin{aligned} x_{\{min/max\}} &= \pm(b-1)[b^{-1} + \cdots + b^{-r}] \cdot b^{(b-1)[b^{s-1} + \cdots + b^0]} \\ &= \pm\left(1 - b^{-r}\right) \cdot b^{(b^s - 1)} \end{aligned}$$

sowie eine kleinste positive/größte negative Zahl:

$$\begin{aligned} x_{posmin/negmax} &= \pm b^{-1} \cdot b^{-(b-1)[b^{s-1} + \cdots + b^0]} \\ &= b^{-b^s} \end{aligned}$$

Das gängigste Format ist das IEEE-Format, das auch hinter dem Python-Datentyp float steht:

$$x = \pm m \cdot 2^{c - 1022}$$

Dieser Datentyp ist 64 Bit (8 Byte) groß (doppelte Genauigkeit, double). Davon speichert 1 Bit das Vorzeichen, 52 Bits die Mantisse $m = 2^{-1} + m_2 2^{-2} + \cdots + m_{53} 2^{-53}$ und 11 Bits die Charakteristik $c = c_0 2^0 + \cdots + c_{10} 2^{10}$, mit $m_i, c_i \in \{0, 1\}$. Es gibt folgende spezielle Werte:

- Alle $c_i, m_i = 0 : x = \pm 0$
- Alle $m_i = 0, c_i = 1 : x = \pm\infty$
- Ein $m_i \neq 0$, alle $c_i = 1$: x = NaN (not a number)

Für c bleibt damit ein Bereich von $\{0, \ldots, 2046\}$ beziehungsweise $c - 1022 \in \{-1022, \ldots, 1024\}$. Damit gilt:

- $x_{max} \approx 2^{1024} \approx 1.8 \times 10^{308}, x_{min} = -x_{max}$
- $x_{posmin} = 2^{-1022} \approx 2.2 \times 10^{-308}, x_{negmax} = -x_{posmin}$

Ausgangsdaten $x \in \mathbb{R}$ einer numerischen Aufgabe und die Zwischenergebnisse einer Rechnung müssen durch Maschinenzahlen dargestellt werden. Für Zahlen des „zulässigen" Bereichs $D = [x_{min}, x_{negmax}] \cup \{0\} [x_{posmin}, x_{max}]$ wird eine Rundungsoperation $\mathrm{rd} : D \to A$ verwendet, die

$$|x - \mathrm{rd}\, x| = \min_{y \in A} |x - y| \forall x \in D$$

erfüllt.

Beispiel 1.4 (Natürliche Rundung im IEEE-Format)

$$\mathrm{rd}(x) = \mathrm{sgn}(x) \cdot \begin{cases} 0, m_1, \ldots, m_{53} \cdot 2^e & m_{54} = 0 \\ (0, m_1, \ldots, m_{53} + 2^{-53}) \cdot 2^e & m_{54} = 1 \end{cases}$$

Rundungsfehler:

- absolut:
$$|x - \mathrm{rd}(x)| \leq \frac{1}{2} b^{-r} b^e$$
- relativ:
$$\left| \frac{x - \mathrm{rd}(x)}{x} \right| \leq \frac{1}{2} \frac{b^{-r} b^e}{|m| b^e} \leq \frac{1}{2} b^{-r+1}$$

Der relative Fehler ist für $x \in D \setminus \{0\}$ beschränkt durch die „Maschienengenauigkeit"

$$eps = \frac{1}{2} b^{-r+1}$$

Für $x \in D$ ist $\mathrm{rd}(x) = x(1 + \varepsilon), |(|\varepsilon) \leq eps$. Für das IEEE-Format (double)

$$eps = \frac{1}{2} 2^{-52} \approx 10^{-16}$$

Arithmetische Grundoperationen

$$* : \mathbb{R} \times \mathbb{R} \to \mathbb{R}, * \in \{x, -, +, /\}$$

werden auf dem Rechner ersetzt durch Maschinenoperationen:

$$\circledast : A \times A \to A$$

Dies ist normalerweise für $x, y \in A$ und $x * y \in D$ realisiert durch

$$x \circledast y := \mathrm{rd}(x * y) = (x * y)(1 + \varepsilon), |\varepsilon| \leq eps$$

Dazu werden die Operationen maschinenintern (unter Verwendung einer längeren Mantisse) ausgeführt, normalisiert und dann gerundet. Im Fall $x * y \notin D$ gibt es eine Fehlermeldung (overflow, underflow) oder das Ergebnis

NaN. Achtung: Das Assoziativ- und Distributivgesetz gilt dann nur näherungsweise. Im Allgemeinen ist für $x, y, z \in A$

$$(x \oplus y) \oplus z \neq x \oplus (y \oplus z)$$
$$(x \oplus y) \odot z \neq (x \odot z) \oplus (y \odot z)$$

Insbesondere gilt für $|y| \leq \frac{|x|}{b} eps$

$$x \oplus y = x$$

Damit ergibt sich eine alternative Charakterisierung der Maschienengenauigkeit: eps ist die kleinste positive Zahl in A, sodass $1 \oplus eps \neq 1$

1.2 Konditionierung numerischer Aufgaben

Eine numerische Aufgabe wird als **gut konditioniert** bezeichnet, wenn eine kleine Störung in den Eingangsdaten (Messfehler, Rundungsfehler) auch nur eine kleine Änderung der Ergebnisse zur Folge hat.

Beispiel 1.5 (Schnittpunkt von Geraden) Zwei Geraden, die sich (annähernd) rechtwinklig treffen sind gut konditioniert.
Zwei Geraden, die sich unter einem stumpfen, oder spitzen Winkel treffen sind schlecht konditioniert.

Beispiel 1.6 (Lineares Gleichungssystem)

$$\begin{pmatrix} 1 & 10^6 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \implies \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & -10^6 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

$$b = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \implies x = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$b = \begin{pmatrix} 1 \\ 10^{-3} \end{pmatrix} \implies x = \begin{pmatrix} -999 \\ 10^{-3} \end{pmatrix} \not\approx \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$\implies$ schlecht konditioniert.

Definition 1.7 Eine **numerische Aufgabe** berechnet aus Eingangsgrößen $x_j \in \mathbb{R}, j = 1, \ldots, m$ unter der funktionellen Vorschrift $f(x_1, \ldots, x_m), i = 1, \ldots, n$ Ausgangsgrößen $y_i = f_i(x_1, \ldots, x_m)$

$$y = f(x), f : \mathbb{R}^m \to \mathbb{R}^n$$

Beispiel 1.8 (Lösung eines LGS) $Ay = x, f(x) = A^{-1}x$

Definition 1.9 Fehlerhafte Eingangsgrößen $x_i + \Delta x_i$ (Δx_i: Rundungsfehler, Maschienenfehler) ergeben fehlerhafte Resultate $y_i + \Delta y_i$. Wir bezeichnen $|\Delta y_i|$ als den absoluten Fehler und $\left|\frac{\Delta y_i}{y_i}\right|$ für $y_i \neq 0$ als den relativen Fehler.

1.2.1 Differentielle Fehleranalyse

Annahmen:

- kleine relative Datenfehler $|\Delta x_i| \ll |x_i|$
- f_i stetig partiell differenzierbar nach allen x_i

Dann gilt:

$$\begin{aligned} y_i = f_i(x_i), y_i + \Delta y_i = f_i(x + \Delta x) \\ \implies \Delta y_i = f_i(x + \Delta x) - f(x) \end{aligned}$$

Taylorentwicklung

$$= \sum_{j=1}^{m} \frac{\partial f_i}{\partial x_j} \Delta x_j + R_i^f(x, \Delta x)$$

mit einem Restglied R_i^f, das für $|\Delta x| = \max_{j=1,\dots,m} |\Delta x_j| \to 0$ schneller gegen 0 geht als $|\Delta x|$. Wenn f sogar zweimal stetig differenzierbar ist, gilt sogar, dass

$$\left| R_i^f(x, \Delta x) \right| \leq c|\Delta x|^2, c \in \mathbb{R}$$

Definition 1.10 (Landau-Notation) Seien $g, h : \mathbb{R}_+ \to \mathbb{R}, t \to 0^+$. Wir schreiben:

- $g(t) = \mathcal{O}(h(t)) : \iff \exists t_0, c \in \mathbb{R}_+ : \forall t \in (0, t_0] : |g(t)| \leq c|h(t)|$
- $gt = \sigma(ht) : \iff \exists t_0 \in \mathbb{R}_+, c : \mathbb{R}_+ \to \mathbb{R}, \lim_{t \to 0^+} c(t) = 0 : \forall t \in (0, t_0] : |g(t)| \leq c(t)|h(t)|$

Bemerkung 1.11

- Analoge Schreibweise für $t \to \infty$
- $\mathcal{O}$ und σ sind Symbole, keine Funktionen

$$\mathcal{O}(t^2) + \mathcal{O}(t^3) + \mathcal{O}(2t^2) = \mathcal{O}(t^2) \not\implies \mathcal{O}(t^3) + \mathcal{O}(2t^2) = 0$$

- $\sigma(t^n)$ ist stärker als $\mathcal{O}(t^n) : \sigma(t^n) + \mathcal{O}(t^n) = \mathcal{O}(t^n)$
- $\mathcal{O}(t^{n+1})$ ist stärker als $\sigma(t^n)$: Wähle $c(t) = t$!

Beispiel 1.12 Ist $g(t)$ zweimal stetig differenzierbar, so gilt mit Taylor

$$g(t + \Delta t) = g(t) + \Delta t g'(t) + \frac{1}{2} \Delta t^2 g''(\tau), \tau \in [t, t + \Delta t]$$

$$\implies \frac{1}{\Delta t}(g(t + \Delta t) - g(t)) = g'(t) + \mathcal{O}(\Delta t)$$

Damit folgt dass Δy_i in erster Näherung, das heißt bis auf eine Größe der Ordnung $\mathcal{O}\left(|\Delta x|^2\right)$ gleich

$$\sum_{j=1}^{m} \frac{\partial f_i}{\partial x_j}(x) \Delta x_j$$

ist. Schreibweise

$$\Delta y_i \doteq \sum_{j=1}^{m} \frac{\partial f_i}{\partial x_j}(x) \Delta x_j$$

Für den komponentenweisen relativen Fehler gilt

$$\frac{\Delta y_i}{y_i} \doteq \sum_{j=1}^{m} \frac{\partial f_i}{\partial x_j}(x) \frac{\Delta x_j}{y_i} = \sum_{j=1}^{m} \underbrace{\frac{\partial f_i}{\partial x_j}(x) \frac{x_j}{f_i(x)}}_{=: k_{ij}(x)} \frac{\Delta x_j}{x_j}$$

Vernachlässigt haben wir dabei

$$\left|\frac{R_i^f(x_j,\Delta x)}{y_i}\right| = \mathcal{O}\left(\frac{|\Delta x|^2}{|y_i|}\right)$$

Diese Vernachlässigung ist nur zulässig falls

$$|\Delta x| = \sigma(|y_i|), i = 1,\dots,n$$

damit

$$\mathcal{O}\left(\frac{|\Delta x|^2}{|y_i|}\right) = \sigma(|\Delta x|)$$

(stärker als $\mathcal{O}(|\Delta x|)$)

Definition 1.13 Die Größen $k_{ij}(x)$ heißen (relative) Konditionszahlen von f im Punkt x. Sie sind Maß dafür, wie sich kleine relative Fehler in den Ausgangsdaten x_j auf das Ergebnis y_i auswirken. Sprechweise:

- $|k_{ij}(x)| \gg 1$: Die Aufgabe $y = f(x)$ ist schlecht konditioniert
- sonst: Die Aufgabe $y = f(x)$ ist gut konditioniert
- $|k_{ij}(x)| < 1$: Fehlerdämpfung
- $|k_{ij}(x)| > 1$: Fehlerverstärkung.

Bemerkung 1.14 Man kann auch Störungen in f betrachten.

Beispiel 1.15 Implizit gegebene Aufgaben. Für $n = m$ sie y die gegebene Eingangsgröße und ein x mit $f(x) = y$ die Ausgabe (zum Beispiel: $f(x) = Ax + b$) Die differentielle Fehleranalyse auf der Umkehrfunktion $x = f^{-1}(y)$ liefert unter geeigneten Annahmen.

$$\frac{\Delta x_i}{x_i} \doteq \sum_{j=1}^{n} k_{ij}^{-1}(y)\frac{\Delta y_j}{y_j}, k_{ij}^{-1} = \frac{\partial f_i^{-1}}{\partial y_j}(y)\frac{y_j}{x_i}$$

Wir definieren die Matrizen

$$K^{-1}(y) = \left(k_{ij}^{-1}\right)_{i,j=1}^{n}, K(x) = (k_{ij}(x))_{i,j=1}^{n}$$

und betrachten deren Produkt:

$$\begin{aligned}
(K^{-1}(y)K(x))_{ij} &= \sum_{l=1}^{n} k_{il}^{-1}(y)k_{lj}(x) \\
&= \sum_{l=1}^{n} \frac{\partial f_i^{-1}}{\partial y_l}(y)\frac{y_l}{x_i}\frac{\partial f_l}{\partial x_j}(x)\frac{x_j}{y_l} \\
&= \frac{x_j}{x_i}\sum_{l=1}^{n} \frac{\partial f_i^{-1}}{\partial y_l}\frac{\partial f_l}{\partial x_j} = \frac{x_j}{x_i}\frac{\mathrm{d}}{\mathrm{d}x_j}\left(f_i^{-1}(f(x))\right) \\
&= \frac{x_j}{x_i}\frac{\mathrm{d}x_i}{\mathrm{d}x_j} = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & \text{sonst} \end{cases}
\end{aligned}$$

K^{-1} ist gerade das Inverse von K.

Wiederholung: Numerische Aufgabe

$$f : x \in \mathbb{R}^m \mapsto y \in \mathbb{R}$$

Konditionszahlen:

$$\frac{\Delta y_i}{y_i} \doteq \sum_{j=1}^{m} k_{ij}(x) \frac{\Delta x_j}{x_j}$$

$$k_{ij}(x) = \frac{\partial f_i}{\partial x_j}(x) \frac{x_j}{f_i(x)}$$

1.2.2 Arithmetische Grundoperationen

Addition: $f(x_1, x_2) = x_1 + x_2, x_1, x_2 \in \mathbb{R} \setminus \{0\}$

$$k_{1j}(x) = \frac{\partial f}{\partial x_j} \frac{x_j}{f} = 1 \frac{x_j}{x_1 + x_2} = \frac{1}{1 + \frac{x_{\bar{j}}}{x_j}}$$

$$\bar{j} = \begin{cases} 2 & j = 1 \\ 1 & j = 2 \end{cases}$$

Die Addition ist schlecht konditioniert für $x_1 \approx -x_2$.

Definition 1.16 (Auslöschung) Unter Auslöschung versteht man den Verlust von Genauigkeit bei der Subtraktion von Zahlen gleichen Vorzeichens.

Beispiel 1.17 $b = 10, r = 4, s = 1$

$$\begin{aligned} x_1 &= 0.112\,587 \times 10^2 & \mathrm{rd}(x_1) &= 0.1126 \times 10^2 \\ x_2 &= 0.112\,448 \times 10^2 & \mathrm{rd}(x_1) &= 0.1124 \times 10^2 \\ x_1 + x_2 &= 0.225\,035 \times 10^2 & \mathrm{rd}(x_1) \oplus \mathrm{rd}(x_2) &= 0.2250 \times 10^2 \\ x_1 - x_2 &= 0.129 \times 10^{-1} & \mathrm{rd}(x_1) \ominus \mathrm{rd}(x_2) &= -0.2 \times 10^{-1} \end{aligned}$$

(Großer Fehler)

Multiplikation: $y = f(x_1, x_2) = x_1 x_2$

$$k_{1j}(x) = \frac{\partial f}{\partial x_j} \frac{x_j}{f} = x_j - \frac{x_j}{x_1 x_2} = 1$$

$\Longrightarrow$ gut konditioniert

Beispiel 1.18 (Lösungen quadratischer Gleichungen) Für $p, q \in \mathbb{R}$ betrachte:

$$0 = y^2 - py + q$$

$$y_{1,2} = y_{1,2}(p, q) = \frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q}$$

nach Vieta $p = y_1 + y_2, q = y_1 \cdot y_2$

$$
\begin{aligned}
1 = \frac{\mathrm{d}p}{\mathrm{d}p} &= \frac{\partial y_1}{\partial p} + \frac{\partial y_2}{\partial p} \\
0 = \frac{\mathrm{d}q}{\mathrm{d}p} &= \frac{\partial y_1}{\partial p} y_2 + y_1 \frac{\partial y_2}{\partial p} \\
\implies (y_2 - y_1)\frac{\partial y_2}{\partial p} &= y_2 \\
\implies \frac{\partial y_2}{\partial p} &= \frac{y_2}{y_2 - y_1} \\
\implies \frac{\partial y_1}{\partial p} &= \frac{y_1}{y_1 - y_2} \\
0 = \frac{\mathrm{d}p}{\mathrm{d}q} &= \frac{\partial y_1}{\partial q} + \frac{\partial y_2}{\partial q} \\
1 = \frac{\mathrm{d}q}{\mathrm{d}q} &= \frac{\partial y_1}{\partial q} y_2 + y_1 \frac{\partial y_2}{\partial q} \\
\implies 1 &= (y_2 - y_1)\frac{\partial y_1}{\partial q} \\
\implies \frac{\partial y_1}{\partial q} &= \frac{1}{y_2 - y_1} \\
\implies \frac{\partial y_2}{\partial q} &= -\frac{1}{y_2 - y_1} \\
k_{11}(x) = \frac{\partial y_1}{\partial p}\frac{p}{y_1} &= \frac{y_1}{y_1 - y_2}\frac{y_1 + y_2}{y_1} = \frac{1 + y_2/y_1}{1 - y_2/y_1} \\
k_{12}(x) = \frac{\partial y_1}{\partial q}\frac{q}{y_1} &= \frac{1}{y_2 - y_1}\frac{y_1 y_2}{y_1} = \frac{1}{1 - y_1/y_2}
\end{aligned}
$$

Analog für k_{21}, k_{22}
Die Berechnung von y_1, y_2 ist schlecht konditioniert $y_1 \approx y_2$.
Konkretes Beispiel: $p = 4, q = 33.999, y_{1,2} = 2 \pm 10 \times 10^{-1}$

$$k_{12} = \frac{y_2}{y_2 - y_1} = \frac{2 - 10^{-2}}{-2 \times 10^{-2}} = -99.5$$

$\implies$ 100-fache Fehlerverstärkung.

1.3 Stabilität numerischer Algorithmen

Gegeben: Numerische Aufgabe $f : x \in \mathbb{R}^m \mapsto y \in \mathbb{R}^n$

Definition 1.19 (Verfahren / Algorithmus) Unter einem Verfahren / Algorithmus zur (gegebenenfalls näherungsweise) Berechnung von y aus x verstehen wir eine endliche Folge von elementaren Abbildungen $\varphi^{(k)}$, die durch sukzessiv Anwendung einen Näherungswert $\tilde{y}$ zu y liefern.

$$x = x^{(0)} \mapsto \varphi^{(1)}\left(x^{(0)}\right) = x^{(1)} \mapsto \ldots \mapsto \varphi^{(k)}\left(x^{(k-1)}\right) \mapsto \tilde{y} \to y$$

Im einfachsten Fall sind die $\varphi^{(i)}$ arithmetische Grundoperationen. Bei der Durchführung des Algorithmus auf dem Rechner treten in jedem Schritt Fehler auf (Rundungsfehler, Auswertungsfehler, …), die sich akkumulieren können.

Definition 1.20 (Algorithmus) Ein Algorithmus heißt stabil, wenn die im Verlauf der Rechnung akkumulierten Fehler den durch die Konditionierung der Aufgabe $y = f(x)$ bedingten unvermeidbaren Problemfehler nicht übersteigen.

Beispiel 1.21 (Lösung quadratischer Gleichungen) Annahme: $0 \neq q < p^2/4$
Für $\left|\frac{y_1}{y_2}\right| \gg 1$, das heißt $q \ll \frac{p^2}{4}$, ist die Aufgabe gut konditioniert. Algorithmus: $u = p^2/4, v = u - q, w = \sqrt{v}$.
Im Fall $p < 0$ wird zur Vermeidung von Auslöschung zunächst $\tilde{y}_2 = p/2 - w$ berechnet.
Fehlerfortpflanzung:

$$w = \sqrt{u - q} \begin{cases} \approx \frac{|p|}{2} & q > 0 \\ > \frac{|p|}{2} & q < 0 \end{cases}$$

$$\begin{aligned} \frac{\Delta y_2}{y_2} &\stackrel{\cdot}{\leq} \left|\frac{\frac{1}{2}p}{\frac{p}{2} - w}\right| \left|\frac{\Delta p}{p}\right| + \left|\frac{-w}{\frac{p}{2} - w}\right| \left|\frac{\Delta w}{w}\right| \\ &= \underbrace{\left|\frac{1}{1 - \frac{2w}{p}}\right|}_{\leq \frac{1}{2}} \left|\frac{\Delta p}{p}\right| + \underbrace{\left|\frac{1}{1 - \frac{p}{2w}}\right|}_{<1} \left|\frac{\Delta w}{w}\right| \end{aligned}$$

Die zweite Wurzel kann so bestimmt werden:

$$A : \tilde{y}_1 = \frac{p}{2} + w, \quad B : \tilde{y}_1 = \frac{q}{\tilde{y}_2}$$

Für $|q| \ll \frac{p^2}{4}$ ist $w \approx \frac{|p|}{2} \implies$ Auslöschung in Variante A

$$\left|\frac{\Delta y_1}{y_1}\right| \stackrel{\cdot}{=} \underbrace{\frac{1}{1 + \frac{2w}{p}}}_{\gg 1} \frac{\Delta p}{p} + \underbrace{\frac{1}{1 + \frac{p}{2w}}}_{\gg 1} \frac{\Delta w}{w}$$

$\implies$ Variante A ist instabil. Variante B ist stabil:

$$\left|\frac{\Delta y_1}{y_1}\right| \stackrel{\cdot}{\leq} \underbrace{\left|\frac{\Delta q}{q}\right|}_{\leq eps} + \underbrace{\left|\frac{\Delta y_2}{y_2}\right|}_{\approx eps}$$

Regel: Bei der Lösung quadratischer Gleichungen sollten nicht beide Wurzeln aus der Lösungsformel berechnet werden.
Konkretes Beispiel: $p = -4, q = 0.01$ (vierstellige Rechnung)

$$u = 4, v = 3.99, w = 1.9974948\ldots, \tilde{y}_2 = -3.997(4981\ldots)$$

$$\tilde{y}_1 = \begin{cases} \text{exakt:} & -0.9925915\ldots \\ A: & -0.003000 \quad \text{(rel. Fehler: 20\%)} \\ B: & -0.002502 \quad \text{(rel. Fehler: } 1.7 \times 10^{-4}\text{)} \end{cases}$$

Auswertung arithmetischer Ausdrücke
Vorwärtsrundungsfehleranalyse: Akkumulation des Rundungsfehlers ausgehend von Startwert.

Beispiel 1.22 $y = f(x_1, x_2) = x_1^2 - x_2^2 = (x_1 - x_2)(x_1 + x_2)$ Konditionierung:

$$\begin{aligned}\left|\frac{\Delta y}{y}\right| &\dot{\leq} \sum_{i=1}^{2} \left|\frac{\partial f}{\partial x_i}\frac{x_i}{f}\right| \left|\frac{\Delta x_i}{x_i}\right| \\ &= \left|2x_1\frac{x_1}{x_1^2 - x_2^2}\right| \left|\frac{\Delta x_1}{x_1}\right| + \left|-2x_2\frac{x_2}{x_1^2 - x_2^2}\right| \left|\frac{\Delta x_2}{x_2}\right| \\ &\leq 2\frac{x_1^2 + x_2^2}{|x_1^2 - x_2^2|} eps = 2\left|\frac{\left(\frac{x_1}{x_2}\right)^2 + 1}{\left(\frac{x_1}{x_2}\right)^2 - 1}\right| eps\end{aligned}$$

$\Longrightarrow$ schlecht konditioniert für $\left|\frac{x_1}{x_2}\right| \approx 1$

Algorithmus A	Algorithmus B
$u = x_1 \odot x_1$	$u = x_1 \oplus x_1$
$v = x_2 \odot x_2$	$v = x_1 \ominus x_2$
$\tilde{q} = u \ominus v$	$\tilde{q} = u \odot v$

Sei $x_1, x_2 \in A$. Für Maschinenoperationen $\circledast$ und $a, b \in A$ gilt

$$a \circledast b = (a * b)(1 + \varepsilon), |(|\varepsilon) \leq eps.$$

Algorithmus A:

$$\begin{aligned}u &= x_1^2(1 + \varepsilon_1), v = x_2^2(1 + \varepsilon_2) \\ \tilde{y} &= \big(x_1^2(1 + \varepsilon_1) - x_2^2(1 + \varepsilon_2)\big)(1 + \varepsilon_3) \\ &= \underbrace{x_1^2 - x_2^2}_{y} + x_1^2\varepsilon_1 - x_2^2\varepsilon_2 + \underbrace{\big(x_1^2 - x_2^2\big)}_{y}\varepsilon_3, |\varepsilon| \leq eps\end{aligned}$$

$$\Longrightarrow \left|\frac{\Delta y}{y}\right| \dot{\leq} eps\frac{x_1^2 + x_2^2 + |x_1^2 - x_2^2|}{|x_1^2 - x_2^2|} = eps\left(1 + \left|\frac{\left(\frac{x_1}{x_2}\right)^2 + 1}{\left(\frac{x_1}{x_2}\right)^2 - 1}\right|\right)$$

Wegen der Konditionierung des Problems

$$\left|\frac{\Delta y}{y}\right| \leq 2\left|\frac{\left(\frac{x_1}{x_2}\right)^2 + 1}{\left(\frac{x_1}{x_2}\right)^2 - 1}\right| eps$$

ist A stabil. Algorithmus B:

$$u = x_1 \oplus x_2, v = x_1 \ominus x_2, y = u \odot v$$

Rundungsfehleranalyse

$$\begin{aligned}u &= (x_1 + x_2)(1 + \varepsilon_1), v = (x_1 - x_2)(1 + \varepsilon_2) \\ \tilde{y} &= (x_1 + x_2)(1 + \varepsilon_1)(x_1 - x_2)(1 + \varepsilon_2)(1 + \varepsilon_3) \\ &= \underbrace{x_1^2 - x_2^2}_{y} + \underbrace{\big(x_1^2 - x_2^2\big)}_{\varepsilon_1 + \varepsilon_2 + \varepsilon_3} + \mathcal{O}\big(eps^3\big)\end{aligned}$$

$$\Longrightarrow \left|\frac{\Delta y}{y}\right| \dot{\leq} |(|\varepsilon_1 + \varepsilon_2 + \varepsilon_3) \leq 3eps$$

$\Longrightarrow$ Algorithmus B ist stabiler als Algorithmus A.

Regel: Bei numerischen Rechnungen sollte man die schlechter konditionierten Operationen möglichst frühzeitig ansetzen.
Wiederholung

- Konditionierung: Eigenschaften einer numerischen Aufgabe
- Stabilität: Eigenschaft eines Verfahrens
  - Auslöschung
- Rundungsfehleranalyse
  - $y = f(x_1, x_2) = x_1^2 - x_2^2 = (x_1 - x_2)(x_1 + x_2)$

Auswertung von Polynomen

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

Als Modellfall betrachten wir

$$p(x) = a_1 x + a_2 x^2 = x(a_1 + a_2 x)$$

Zwei Varianten für $\tilde{y} = p(\xi), \xi \in A$
A: $u = \xi \odot \xi, v = a_2 \odot u, w = a_1 \odot \xi, \tilde{y} = v + w$
B: $u = a_2 \odot \xi, v = a_1 \oplus u, \tilde{y} = \xi \odot v$
B spart eine arithmetische Operation.
Rundungsfehleranalyse A:

$$\begin{aligned}
u &= \xi^2(1+\varepsilon_1), v = a_2\xi^2(1+\varepsilon_1)(1+\varepsilon_2), w = a_1\xi(1+\varepsilon_3) \\
\tilde{y} &= (a_2\xi_2(1+\varepsilon_1)(1+\varepsilon_2) + a_1\xi(1+\varepsilon_3))(1+\varepsilon_4) \\
&= \underbrace{a_2\xi^2 + a_1\xi}_{y} + \underbrace{\left(a_2\xi^2 + a_1\right)}_{y}\varepsilon_4 + a_2\xi^2(\varepsilon_1+\varepsilon_2) + a_1\xi\varepsilon_3 + \mathcal{O}eps^2 \\
\frac{\Delta y}{y} &\doteq \varepsilon_4 \frac{a_2\xi^2(\varepsilon_1\varepsilon_2) + a_1\xi\varepsilon_3}{a_2\xi^2 + a_1\xi} \\
&= \varepsilon_4 + \varepsilon_3 + \frac{a_2\xi^2(\varepsilon_1+\varepsilon_2-\varepsilon_3)}{a_2\xi^2 + a_1\xi} \\
&= \varepsilon_3 + \varepsilon_3 + \frac{\xi}{\frac{a_1}{a_2} + \xi}(\varepsilon_1+\varepsilon_2-\varepsilon_3)
\end{aligned}$$

Variante B:

$$\begin{aligned}
u &= x_2\xi(1+\varepsilon_1), v = (a_1 + a_2\xi(1+\varepsilon_1))(1+\varepsilon_2) \\
\tilde{y} &= \xi \cdot [a_1 + a_2\xi(1+\varepsilon_1)](1+\varepsilon_2)(1+\varepsilon_3) \\
&= \underbrace{\xi(a_1 + a_2\xi)}_{y} + a_1\xi(\varepsilon_2+\varepsilon_3) + a_2\xi^2(\varepsilon_1+\varepsilon_2+\varepsilon_3) + \mathcal{O}\left(eps^2\right) \\
\frac{\Delta y}{y} &= \varepsilon_2 + \varepsilon_3 + \frac{a_2\xi^2}{a_1\xi + a_2\xi}\varepsilon_2 = \varepsilon_2 + \varepsilon_3 + \frac{\xi}{\frac{a_1}{x_2} + \xi}\varepsilon_1
\end{aligned}$$

$\Longrightarrow$ Variante B ist etwas stabiler als A im Fall $\xi \approx -\frac{a_1}{a_2}$ (nahe bei Nullstelle) Allgemein:

$$\begin{aligned}
p(x) &= a_0 + a_1 x + \cdots + a_n x^n \\
&= a_0 + x(a_1 + x(\cdots + x(a_{n-1} + a_n x)\ldots))
\end{aligned}$$

Definition 1.23 (Horner-Schema)

$$b_n = a_n, b_k = a_k + \xi b_{k-1}, k = n-1, \ldots, 0$$

liefert den Funktionswert $p(\xi) = b_0$ des Polynoms an der Stelle $x = \xi$.

Regel: Die Auswertung von Polynomen sollte mit dem Horner-Schema erfolgen.

2 Interpolation und Approximation

Grundproblem:
Darstellung und Auswertung von Funktionen.
Aufgabenstellung:

1. Eine Funktion $f(x)$ ist nur auf einer diskreten Menge von Argumenten $x_0, \ldots, x_n$ bekannt und soll rekonstruiert werden (zum Beispiel für Graph Ausgabe)
2. Eine analytisch gegebene Funktion $f(x)$ soll auf dem Rechner so dargestellt werden, dass jederzeit Funktionswerte zu beliebigen Argument x berechnet werden können.

$\rightarrow$ System mit unendlich vielen Freiheitsgraden $y = f(x)$. „Simulation“ durch endlich viele Datensätze in Klassen P von einfach strukturierten Funktionen

- Polynome: $p(x) = a_0 + a_1 x + \cdots + a_n x^n$
- rationale Funktionen:
$$r(x) = \frac{a_0 + a_1 x + \cdots + a_n x^n}{b_0 + b_1 x + \cdots + b_m x^m}$$
- trigonometrische Funktionen
$$t(x) = \frac{1}{2} a_0 + \sum_{k=1}^{n} (a_k \cos(kx) + b_k \sin(kx))$$
- Exponentialsummen
$$e(x) = \sum_{k=1}^{n} a_k \exp(b_k x)$$

Definition 2.1 Geschieht die Zuordnung eines Elementes $g \in P$ zur Funktion f durch Fixieren von Funktionswerten

$$g(x_i) = y_i = f(x_i), i = 0, \ldots, n$$

so spricht man von **Interpolation**. Ist g im gewissen Sinne die beste Darstellung von f, zum Beispiel:
$\max_{a \le x \le b} |f(x) - g(x)|$ minimal für $g \in P$, oder
$\left(\int_a^b |f(x) - g(x)|^2 \mathrm{d}x \right)^{1/2}$ minimal für $g \in P$ so spricht man von **Approximation**. Die Wahl der Konstruktion von $g \in P$ hängt von der zu erfüllenden Aufgabe ab. Offenbar ist die Interpolation eine Approximation mit

$$\max_{i=0,\ldots,n} |f(x_i) - g(x_i)|$$

für $g \in P$

Wiederholung: Interpolation und Approximation

- Stützstellen x_i mit Werten $y_i, i = 0, \ldots, n$
- Klassen P von Funktion

Polynominterpolation
Wir bezeichnen mit P_n den Vektorraum der Polynome vom Grad $\leq n$:

$$P_n = \{p(x) = a_0 + a_1 x + \cdots + a_n x^n \mid a_i \in \mathbb{R}, i = 0, \ldots, n\}$$

Definition 2.2 (Lagrangesche Interpolationsaufgabe) Die Lagrangesche Interpolationsaufgabe besteht darin zu $x+1$ paarweise verschiedenen Stützstellen (auch Knoten genannt) $x_0, \ldots, x_n \in \mathbb{R}$ und gegebenen Knotenwerten $y_0, \ldots, y_n \in \mathbb{R}$ ein Polynom $p \in P_n$ zu bestimmen mit der Eigenschaft $p(x_i) = y_i$

Satz 2.3 Die Lagrangesche Interpolationsaufgabe ist eindeutig lösbar.

Beweis Eindeutigkeit: Sind $p_1, p_2 \in P_n$ Lösungen, so gilt für $p = p_1 - p_2$, dass

$$p(x_i) = p_1(x_i) - p_2(x_i) = y_i - y_i = 0, i = 0, \ldots, n$$

Also hat p $n + 1$ Nullstellen und ist folglich identisch Null. $\implies p_1 = p_2$
Existenz: Wir betrachten die Gleichungen

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

Dies ist ein lineares Gleichungssystem mit $n + 1$ Gleichungen und $n + 1$ Freiheitsgraden.

$$\begin{pmatrix} x_0^0 & x_0^1 & \ldots & x_0^n \\ x_1^0 & x_1^1 & & x_1^n \\ \vdots & & \ddots & \vdots \\ x_n^0 & x_n^1 & \ldots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Wegen der Eindeutigkeit von p ist $\ker V = \{0\}$. Mit dem Rangsatz ($\dim \mathbb{R}^{n+1} = \dim \ker V + \dim \operatorname{im} V$) liefert V eine surjektive Abbildung. Damit existiert eine Lösung. □

Zur Konstruktion des Interpolationspolynoms $p \in P_n$ verwenden wir die sogenannten Lagrangesche Basispolynome.

$$L_i^{(n)}(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} \in P_n, i = 0, \ldots, n$$

Lemma 2.4 $\{L_i^{(n)}, i = 0, \ldots, n\}$ ist eine Basis von P_n

Beweis Übung. □

Offensichtlich gilt:

$$L_i^{(n)}(x_k) = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

Definition 2.5 Das Polynom

$$p(x) = \sum_{i=0}^{n} y_i L_i^{(n)}(x) \in P_n$$

hat die gewünschten Eigenschaften

$$p(x_j) = y_j, j = 0, \ldots, n$$

und wird die Lagrangesche Darstellung des (Lagrangeschen) Interpolationspolynoms zu dem Stützpunkten $(x_i, y_i), i = 0, \ldots, n$ genannt.

Nachteil: Bei Hinzunahme eines weiteren Stützpunktes (x_{n+1}, y_{n+1}) ändern sich die Basispolynome völlig.
Abhilfe: Newtonsche Basispolynome

$$N_0(x) = 1, N_i(x) = (x - x_{i-1})N_{i-1}(x) = \prod_{j=0}^{i-1}(x - x_j)$$

Für den Ansatz

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

erhält man durch Auswertung von $x_0, \ldots, x_n$ das gestaffelte Gleichungssystem

$$\begin{aligned}
y_0 &= p(x_0) = a_0 \\
y_1 &= p(x_1) = a_0 + a_1(x_1 - x_0) \\
&\vdots \\
y_0 &= p(x_0) = a_0 + a_1(x_1 - x_0) + \cdots + a_n(x_n - x_0) \cdot \ldots \cdot (x_n - x_{n-1})
\end{aligned}$$

aus dem sich die Koeffizienten a_i rekursiv berechnen lassen. Bei Hinzunahme eines weiteren Stützpunktes (x_{n+1}, y_{n+1}) setzt man den Prozess mit der Basisfunktion N_{n+1} fort. In der Praxis verwendet man folgende stabilere und effizientere Methode

Satz 2.6 (Newtonsche Darstellung) Das Lagrangesche Interpolationspolynom zu den Punkten $(x_0, y_0), \ldots, (x_n, y_n)$ lässt sich bezüglich der Newtonschen Polynombasis schreiben in der Form

$$p(x) = \sum_{i=0}^{n} y[x_0, \ldots, x_i] N_i(x)$$

Dabei bezeichnen $y[x_0, \ldots, x_i]$ die zu den Punkten (x_i, y_i) gehörenden „dividierten Differenzen", welche rekursiv definiert sind durch

$$\begin{aligned}
&\text{für } j = 0, \ldots, n : y[x_j] = y_j \\
L\text{für } k = 1, \ldots, j : i = k - j : y\underbrace{[x_i, \ldots, x_{1+k}]}_{k+1} &= \frac{y\underbrace{[x_{i+1}, \ldots, x_{1+k}]}_{k} - y\underbrace{[x_i, \ldots, x_{x_1+k-1}]}_{k}}{x_{i+k} - x_i}
\end{aligned}$$

Beweis Es bezeichne $pi, i + k \in P_k$ das Polynom, welches die Punkte $(x_i, y_i), \ldots, (x_{i+k}, y_{i+k})$ interpoliert. Speziell ist $p_{0,n} = p$ das gesuchte Interpolationspolynom. Wir zeigen

$$p_{i,i+k}(x) = y[x_i] + y[x_i, x_{i+1}](x - x_i) + \cdots + y[x_i, \ldots, x_{i+k}](x - x_i) \cdot \ldots \cdot (x - x_{i+k})$$

was für $i = 0$ und $k = n$ den Satz beweist. Der Beweis wird durch Induktion über die Indexdifferenz k geführt. Für $k = 0$ ist $p_{i,i} = y_i = y[x_i], i = 0, \ldots, n$. Sei die Behauptung richtig für $k - 1 \geq 0$. Nach Konstruktion gilt für ein $a \in \mathbb{R}$

$$p_{i,i+k}(x) = p_{i,i+k-1}(x) + a(x - x_1) \cdot \ldots \cdot (x - x_{i+k-1}) = 0$$

für $x = x_j, j = i, \ldots, i + k - 1$. Zu zeigen: $a = y[x_i, \ldots, x_{i+k}]$. Offenbar ist a der Koeffizient von x^k in $p_{0,i+k}$. Nach Induktionsannahme ist also

$$\begin{aligned} p_{i,i+k-1}(x) &= \cdots + y[x_i, \ldots, x_{i+k-1}]x^{k-1} \\ p_{i+1,i+k-1}(x) &= \underbrace{\ldots}_{\text{Grad} \leq k-2} + y[x_{i+1}, \ldots, x_{i+k}]x^{k-1} \end{aligned}$$

Ansatz:

$$\begin{aligned} q(x) &= \frac{(x - x_i)p_{i+1,i+k}(x) - (x - x_{i+k})p_{i,i+k-1}(x)}{x_{i+k} - x_i} \\ &= p_{i,i+k-1}(x) + \frac{(x - x_i)p_{i+1,i+k}(x) - (x - x_{i+k} + x_{i+k} - xi)p_{i,i+k-1}(x)}{x_{i+k} - x_i} \\ &= p_{i,i+k-1}(x) + (x - x_i)\frac{p_{i+1,i+k}(x) - p_{i,i+k-1}(x)}{x_{i+k} - x_i} \end{aligned}$$

Ex gilt:

$$q(x_i) = y_i, q(x_{i+k}) = \frac{(x_{i+k} - x_i)y_{i+k} + 0}{x_{i+k} - x_i} = y_{1+k}$$

$$q(x_j) = \frac{(x_j - x_i)y_j - (x_j - x_{i+k})y_j}{x_{i+k} - x_i} = y_j, j = i + 1, \ldots, i + k - 1$$

$\implies$ q interpoliert die Stützpunkte $(x_i, y_i), \ldots, (x_{i+k}, y_{i+k})$ $\implies$ $q \equiv p_{i,i+k}$ (Eindeutigkeit des Interpolationspolynoms). Der führende Koeffizient in $p_{i,i+k}(x)$ ist demnach

$$\begin{aligned} q &= \frac{y[x_{i+1}, \ldots, x_{i+k}] - y[x_i, \ldots, x_{i+k-1}]}{x_{i+k} - x_i} \\ &= y[x_i, \ldots, x_{i+k}] \end{aligned}$$

□

Korollar 2.7 Sei $\sigma : \{0, \ldots, n\} \to \{0, \ldots, n\}$ eine beliebige Permutation. Dann gilt für die Stützpunkte $(\tilde{x}_i, \tilde{y}_i) = \left(x_{\sigma(j)}, y_{\sigma(j)}\right)$

$$y[\tilde{x}_0, \ldots, \tilde{x}_n] = y[x_0, \ldots, x_n]$$

Beweis Koeffizient des Monoms x^n ist $y[x_0, \ldots, x_n]$ unabhängig von der Reihenfolge. □

Wiederholung: Lagrange-Interpolation:
Gegeben: $(x_i, y_i), i = 0, \ldots, n$
Suche $p \in P_n : p(x_i) = y_i, i = 0, \ldots, n$
Lösung:

$$\begin{aligned} p(x) &= \sum_{i=0}^{n} y_i L_i^{(n)}(x) \\ &= L_i^{(n)}(x) \qquad\qquad = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} \in P_n \end{aligned}$$

$\implies L_i^{(n)}(x_j) = \delta_{ij}$
Andere Darstellung: Newton-Neville

$$N_i(x) = \prod_{j=0}^{n-1}(x - x_j)$$

$$p(x) = \sum_{i=0}^{N} y[x_0, \ldots, x_i] D_i(x)$$

$$y[x_i] = q_i$$

$$y[x_i, \ldots, x_{i+k}] = \frac{y[x_{i+1}, \ldots, x_{i+k}] - y[x_i, \ldots, x_{i+k-1}]}{x_{i+k} - x_i}$$

Definition 2.8 Das durch die Rekursion $j = 0, \ldots, n, p_{j,j}(x) = y_j$ für $k = 1, \ldots, j : i = k - j$

$$p_{i,i+k}(x) = p_{i,i+k-1}(x) + (x - x_i)\frac{p_{i+1,i+k}(x) - p_{i,i+k-1}(x)}{x_{i+k} - x_i}$$

erzeugte Polynom $p_{0,1}$ ist die sogenannte Nevillsche Darstellung des Interpolationspolynom zu den Stützstellen $(x_0, y_0), \ldots, (x_n, y_n)$

Schema:

	$k=0$		$k=1$		$k=2$	$\ldots$	$k=n-1$	$k=n$	
x_0	y_0	$\rightarrow$	$p_{0,1}$	$\rightarrow$	$p_{0,2}$	$\cdots$	$p_{0,n-1}$	$\rightarrow$	$p_{0,n}$
x_1	y_1	$\nearrow\rightarrow$	$p_{1,2}$	$\nearrow\rightarrow$	$p_{1,3}$	$\cdots$	$p_{1,n}$	$\nearrow$	
x_2	y_2	$\nearrow$							
$\vdots$		$\vdots$	$\ddots$						
x_{n-1}	y_{n-1}	$\rightarrow$	$p_{n-1,n}$						
x_n	y_n	$\nearrow$							

Die Hinzunahme eines weiteren Stützpunktes ist problemlos. Die Auswertung von $p_{0,n}(x)$ an einer Stelle $\xi \neq x_i$ ohne vorige Bestimmung der Koeffizienten der Newton-Darstellung ist damit sehr einfach und numerisch effizient und stabil möglich. Dazu wird im Schema x mit ξ ersetzt.

2.1 Auswertung von Polynomen und deren Ableitungen

Sei $p \in P_n$ gegeben in der Darstellung

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

Wiederholung: Auswertung von $p(\xi)$ mittels Horner-Schema

$$b_k = \begin{cases} a_n & k = n \\ a_k + \xi b_{k+1} & k = n-1, \ldots, 0 \end{cases}$$

$\implies p(\xi) = b_0$.
Zu $p_n = p \in P_n$ und festem ξ wird durch

$$p_{n-1}(x) = b_1 + b_2 x + \cdots + b_n x^{n-1}$$

ein Polynom $p_{n-1} \in P_{n-1}$ definiert. Wegen $a_k = b_k - \xi b_{k+1}, k = 0, \ldots, n-1, a_n = b_n$:

$$\begin{aligned} p_n(x) &= \sum_{k=0}^{n} b_k x^k - \xi \sum_{k=0}^{n-1} b_{k+1} x^k \\ &= b_0 + x \sum_{k=1}^{n} b_k x^{k-1} - \xi \sum_{k=1}^{n} b_k x^{k-1} \\ &= r_0 + (x - \xi) p_{n-1}(x) \quad r_0 = p(\xi) = b_0 \end{aligned}$$

$\Longrightarrow$ Für eine Nullstelle ξ von p_n leistet das Horner-Schema die Abspaltung des Linearfaktors $(x - \xi)$ vom Polynom p_n. Weiter ist dann für $x \neq \xi$

$$\frac{p_n(x) - p_n(\xi)}{x - \xi} = p_{n-1}(x)$$

$x \to \xi$

$$p_n'(\xi) = p_{n-1}(\xi)$$

Zur Berechnung von $p_n'(\xi)$ wird das Horner-Schema auf p_{n-1} angewendet.

$$p_{n-2} \in P_{n-2}, p_{n-1}(x) = r_1 + (x - \xi) p_{n-2}(x), r_1 = p_{n-1}(\xi)$$

Fortsetzen $\to$ endliche Folge von Polynomen $p_n, p_{n-1}, \ldots, p_0$ mit

$$p_{n-j}(x) = (x - \xi) p_{n-j-1}(x) + r_j, \quad j = 0, \ldots, n$$
$$p_n(x) = r_0 + r_1(x - \xi) + \cdots + r_n(x - \xi)^n$$

Vergleich mit der Taylorentwicklung von p_n um ξ ergibt

$$r_j = \frac{1}{j!} p_n^{(j)}(\xi)$$

Die Koeffizienten von p_{n-j} seien

$$p_{n-j}(x) = a_j^{(j)} + a_{j+1}^{(j)} x + \cdots + a_n^{(j)} x^{n-j}, j = 0, \ldots, n$$

Es gilt die Rekursion:

$$a_k^{(j+1)} = \begin{cases} a_n^{(j)} & k = n \\ a_k^{(j)} + \xi a_{k+1}^{(j+1)} \end{cases}$$

und es gilt

$$p^{(j)}(\xi) = j! a_j^{j+1}, j = 0, \ldots, n$$

Dieses „vollständige Horner-Schema“ kann leicht zur Auswertung von Polynomen in Newton-Darstellung modifiziert werden:

$$p(x) = a_0 + a_1(x - x_0) + \cdots + a_n(x - x_0) \cdot \ldots \cdot (x - x_{n-1})$$

2.2 Interpolation von Funktionen

Stützstellen $x_0, \ldots, x_n \in [a, b]$. Werte gegeben durch Funktion $y_i = f(x_i), i = 0, \ldots, n$
Frage: Wie gut approximiert das Interpolationspolynom $p \in P_n$ die Funktion f auf $[a, b]$?
Bezeichnungen:

- $\overline{(x_0, \ldots, x_n)}$ = kleinstes Intervall, das alle x_i enthält.
- $C[a, b]$: Vektorraum der über $[a, b]$ stetigen Funktionen
- $C^k[a, b]$: Vektorraum über $[a, b]$ k-mal stetig differenzierbarer Funktionen.

Satz 2.9 (Interpolationsfehler 1) Sei $f \in C^{n+1}[a, b]$. Dann gibt es zu jedem $x \in [a, b]$ ein $\xi_x \in \overline{(x_0, \ldots, x_n, x)}$, sodass gilt

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi x)}{(n+1)!} \prod_{j=0}^{n} (x - x_j)$$

Beweis Für $x \in \{x_0, \ldots, x_n\}$ ist alles klar. Sei $x \in [a, b] \setminus \{x_0, \ldots, x_n\}$. Wir setzen

$$l(t) = \prod_{j=0}^{n} (t - x_j), \quad c(x) = \frac{f(x) - p(x)}{l(x)}$$

Die Funktion

$$F(t) = f(t) - p(t) - c(x)l(t)$$

besitzt dann mindestens die $n + 2$ Nullstellen $x_0, \ldots, x_n, x$ in $[a, b]$. Durch wiederholte Anwendung des Satzes von Rolle schließt man, dass die Ableitung F^{n+1} eine Nullstelle $\xi_x \in \overline{(x_0, \ldots, x_n, x)}$. Es

$$\begin{aligned} 0 = F^{(n+1)}(\xi_x) &= f^{(n+1)}(\xi) - p^{(n+1)}(\xi) - c(x)l^{(n+1)}(t) \\ &= f^{(n+1)}(\xi) - c(x)(n+1)! \end{aligned}$$

□

Wiederholung:

- Neville-Schema für $p \in P_n$:
$$p(x_i) = y_i, i = 0, \ldots, n$$
- Vollständiges Horner-Schema
- Interpolation von Funktionen $y_i = f(x_i)$

Interpolationsfehler 1: Sei $f \in C^{n+1}[a, b] \implies \forall x \in [a, b] \exists \xi_x \in \overline{(x_0, \ldots, x_n, x)}$:

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \prod_{j=0}^{n} (x - x_j)$$

Satz 2.10 (Interpolationsfehler 2) Sei $f \in C^{n+1}[a, b]$. Dass gilt für $x \in [a, b] \setminus \{x_0, \ldots, x_n\}$:

$$f(x) - p(x) = f[x_0, \ldots, x_n, x] \prod_{j=0}^{n} (x - x_j)$$

mit der Notation

$$f[x_i, \ldots, x_{i+k}] = y[x_i, \ldots, x_{i+k}]$$

und es ist

$$f[x_0, \ldots, x_n, x] = \int_0^1 \int_0^{t_1} \ldots \int_0^{t_n} f^{(n+1)}(x_0 + t_1(x_1 - x_0) + \cdots + t_n(x_n - x_{n-1}) + t(x - x_n)) \mathrm{d}t \mathrm{d}t_n \ldots \mathrm{d}t_1$$

Beweis Per Induktion über n.
IA: $n = 0$:

$$f(x) - p_0(x) = f(x) - f(x_0) = \begin{cases} f[x_0, x](x - x_0) \\ (x - x_0) \int_0^1 f'(x_0 + t(x - x_0))\mathrm{d}t \end{cases}$$

wobei ein

$$\int_0^1 g'(t)\mathrm{d}t = g(1) - g(0)$$

für $g(t) = f(x_0 + t(x - x_0)) \implies g'(t) = f'(t)(x - x_0)$
Sei die Behauptung richtig für $n - 1 \geq 0$. Dann ist

$$\begin{aligned} f(x) - p_n(x) &= f(x) - \sum_{i=0}^{n} f[x_0, \ldots, x_n] \prod_{j=0}^{i-1}(x - x_j) \\ &= f(x) - p_{n-1}(x) - f[x_0, \ldots, x_n] \prod_{j=0}^{n-1}(x - x_j) \\ &= f[x_0, \ldots, x_{n-1}, x] \prod_{j=0}^{n-1}(x - x_j) - f[x_0, \ldots, x_n] \prod_{j=0}^{n-1}(x - x_j) \\ &= (f[x_0, \ldots, x_{n-1}, x] - f[x_0, \ldots, x_n]) \prod_{j=0}^{n-1}(x - x_j) \\ &= \frac{f[x, x_0, \ldots, x_n] - f[x_0, \ldots, x_n]}{x - x_n} \prod_{j=0}^{n-1}(x - x_j) \\ &= f[x_0, \ldots, x_n, x] \prod_{j=0}^{n-1}(x - x_j) \end{aligned}$$

Weiterhin gilt:

$$f[x_0, \ldots, x_{n-1}, x] - f[x_0, \ldots, x_n] = \int_0^1 \int_0^{t_1} \ldots \int_0^{n-1} [f^{(n)}(x_0 + t_1(x_1 - x_0) + \cdots + t_n(x - x_{n+1})) - f^{(n)}(x_0 + t_1(x$$

Setze $g(t) = f^{(n)}(x_0 + t_1(x_1 - x_0) + \cdots + t_n(x_n - x_{n-1}) + t_{x-x_n})$

$$\begin{aligned} &= \int_0^1 \int_0^{t_1} \ldots \int_0^{t_{n-1}} [g(t_n) - g(0)]\mathrm{d}t_n \ldots \mathrm{d}t_1 \\ &= \int_0^1 \int_0^{t_1} \ldots \int_0^{t_{n-1}} \int_0^{t_n} f^{(n+1)}(x_0 + t_1(x_1 - x_0) + \cdots + t_n(x_n - x_{n-1}) + t(x - \\ \implies f[x_0, \ldots, x_n, x] &= \int_0^1 \int_0^{t_1} \ldots \int_0^{t_n} f^{(n+1)}(\ldots)\mathrm{d}t\mathrm{d}t_n \ldots \mathrm{d}t_1 \end{aligned}$$

□

Die Integraldarstellung der dividierten Differenzen gestattet ihre stetige Fortsetzung für den Fall, das Stützstellen zusammenfallen:

$$f[x_0, \ldots, x_r, x_r, \ldots, x_n] = \lim_{\varepsilon \to 0} f[x_0, \ldots, x_r, x_r + \varepsilon, \ldots, x_n]$$

Im Extremfall $x_0 = x_1 = \cdots = x_n$ wird

$$\begin{aligned} f[x_0, \ldots, x_n] &= \int_0^1 \int_0^{t_1} \ldots \int_0^{t_{n-1}} f^{(n)}(x_0) \mathrm{d}t_n \ldots \mathrm{d}t_1 \\ &= \int_0^1 \int_0^{t_1} \ldots \int_0^{t_{n-1}} 1 \mathrm{d}t_n \ldots \mathrm{d}t_1 f^{(n)}(x_0) \\ &= \frac{1}{n!} f^{(n)}(x_0) \end{aligned}$$

Damit geht das Newtonsche Interpolationspolynom über in das Taylorpolynom n-ten Grades von f in x_0. Konstruieren wir die Fehlerdarstellung so erhalten wir für ein $\xi_x \in (x_0, \ldots, x_n, x)$

$$\begin{aligned} \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \prod_{j=0}^{n} (x - x_j) &= f(x) - p(x) \\ &= f[x_0, \ldots, x_n, x] \prod_{j=0}^{n} (x - x_j) \\ \implies f[x_0, \ldots, x_n, x] &= \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \end{aligned}$$

Definition 2.11 (Hermite-Interpolation) Die Hermitesche Interpolationsaufgabe lautet:
Gegeben $x_i, i = 0, \ldots, m$ (paarweise verschieden), $y_i^{(k)}, i = 0, \ldots, m, k = 0, \ldots, \mu_i, \mu \geq 0$.
Gesucht: $p \in P_n, n = m + \sum_{i=0}^{m} \mu_i, p^{(k)}(x_j) = y_i^{(k)}, i = 0, \ldots, m, k = 0, \ldots, \mu_i$, $(\mu_i + 1)$ -fache Stützstellen.

Beispiel 2.12 $x_0 = -1, x_1 = 1, m = 1, y_0^{(0)} = 0, y_1^{(0)} = 1, y_1^{(1)} = 2 \implies \mu_0 = 0, \mu_1 = 1 \implies n = 1 + 0 + 1 = 2 \implies p(x) = x^2$

Analog zur Lagrange-Interpolation:

- Existenz + Eindeutigkeit
- Darstellung des Interpolationsfehlers

Wiederholung: Fehlerdarstellung Lagrange-Interpolation. Sei $f \in C^{n+1}[a, b]$. $\exists \xi_x \in \overline{(x_0, \ldots, x_n, x)}$

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \prod_{j=0}^{n} (x - x_j)$$

$$f(x) - p(x) = f[x_0, \ldots, x_n, x] \prod_{j=0}^{n} (x - x_j)$$

$$f[x_0, \ldots, x_n, x] = \int_0^1 \int_0^{t_1} \ldots \int_0^{t_n} f^{(n+1)}(x_0 + t_1(x_1 - x_0) + \cdots + t_n(x_n - x_{n-1}) + t(x_n - x)) \mathrm{d}t \mathrm{d}t_n \ldots \mathrm{d}t_1$$

Hermite-Interpolation: Such $p \in P_n, n = m + \sum_{i=0}^{m} \mu_i$

$$p^{(k)}(x_i) = y_i^{(k)}, i = 0, \ldots, m, k = 0, \ldots, \mu_i$$

2.3 Richardsonsche Extrapolation zum Limes

Gegeben: Numerischer Prozess mit Werten $a(h), h \in \mathbb{R}_+, h \to 0$.
Gesucht: $a(0) = \lim_{h\to 0} a(h)$
Idee: Für $h_i > 0, i = 0, \dots, n$, interpoliere $(h_i, a(h_i))$ und berechne $p_n(0)$

Beispiel 2.13 (Numerische Differentiation) Sei $f \in C^\infty[a, b], x \in (a, b)$. Nach Taylor gilt

$$a(h) = \frac{f(x+h) - f(x-h)}{2h} = f'(x) + \sum_{i=1}^{\infty} \frac{f^{(2i+1)}(x)}{(2i)!} h^{2i}$$

Satz 2.14 (Extrapolationsfehler) Für $h \in \mathbb{R}_+$ habe $a(n)$ die Entwicklung

$$a(h) = a_0 + \sum_{j=1}^{n} a_j h^{jq} + a_{n+1}(h) h^{(n+1)q}$$

mit $q > 0$, Koeffizienten a_j und

$$a_{n+1}(h) = a_{n+1} + \wr(1)$$

Die Folge $(h_i)_{k\in\mathbb{N}}$ erfülle

$$0 \le \frac{h_{k+1}}{h_k} \le \rho < 1$$

($\implies$ h_k positiv, monoton fallend). Dann gilt für das Interpolationspolynom $p_1^{(k)} \in P_n$ (in h^q) durch

$$\big(h_k^q, a(h_k)\big), \dots, \big(h_{k+n}^q, a(h_{k+n})\big)$$

$$a(0) - p_n^{(k)}(0) = \mathcal{O}\Big(h_k^{(n+1)q}\Big)$$

$(k \to \infty)$

Beweis Abkürzungen $z = h^q, z_k = h_k^q$. Interpoliere $(z_{k+i}, a(h_{k+i})), i = 0, \dots, n$.

$$p_n(z) = \sum_{i=0}^{n} a(h_{k+i}) L_{k+i}^{(n)} I$$

$$L_{k+1}^{(n)}(z) = \prod_{\substack{l=0 \\ l\ neqi}} \frac{z - z_{k+l}}{z_{k+1} - z_{wl}}$$

Übung:

$$\sum_{i=0}^{n} x_{k+1}^n(0) = \begin{cases} 1 & r = 0 \\ 0 & r = 1, \dots, n \\ (-1)^n \prod_{j=0}^{n} z_{k+i} & r = n+1 \end{cases}$$

$$\begin{aligned} p_n(0) &= \sum_{i=0}^{n} \left(a_0 + \sum_{j=1}^{n} a_j z_{k+i}^j + a_{n+1}(h_{k+1}) z_{k+i}^{n+1} \right) L_{k+i}^{(n)}(0) \\ &= a_0 \underbrace{\sum_{i=0}^{n} L_{k+1}^{(n)}}_{=1} + \sum_{j=1}^{n} a_j \underbrace{\sum_{i=0}^{n} z_{k+1}^j L_{k+i}^{(n)}(0)}_{0} \\ &= +a_{n+1} \underbrace{\sum_{i=0}^{n} z_{k+1}^{n+1} L_{k+1}^{(n)}}_{=(-1)^n \prod_{i=0}^{n} z_{k+i}} + \sum_{i=0}^{n} \wr(1) z_{k+i}^{n+1} L_{k+i}^{(n)}(0) \end{aligned}$$

Da man Landau-Symbole nicht ausklammern darf, schätzen wir ab:

$$\begin{aligned}
\left|L_{k+i}^{(n)}(0)\right| &= \prod_{\substack{l=0\\ l\neq i}}^{n}\left|\frac{z_k+l}{z_{k+i}-z_{k+y}}\right| \\
&\leq \prod_{l=0}^{i-1}\left|\frac{z_{n+l}}{z_{k+i}-z_{k+l}}\right| \prod_{l=1+i}^{n}\left|\frac{z_{k+i}}{z_{k+i}-z_{k+l}}\right| \\
&= \prod_{l=0}^{i-1}\frac{1}{\left|\frac{z_{k+i}}{z_{k+y}}-1\right|} \prod_{l=i+1}^{n}\frac{1}{\left|1-\frac{z_{k+l}}{z_{k+i}}\right|} \\
&\leq \frac{1}{(1-\rho^q)^n} \\
\implies p_n(0) &= a_0 + a_{n+1}(-1)^n \prod_{i=0}^{n} z_{k+i} + \wr\left(z_k^{n+1}\right) \\
&= a_0 + \mathcal{O}\left(h_k^{(n+1)q}\right)
\end{aligned}$$

□

2.4 Spline-Interpolation

Problem: Oszillationen des Interpolationspolynoms, wenn man Stützstellen nicht geeignet wählen kann. Abhilfe: Stückweise polynomielle Interpolation:

- Zerlegung: $a = x_0 < x_1 < \cdots < x_n = b$
- Teilintervalle: $I_i = [x_{i-1}, x_i], i = 1, \ldots, n$
- Feinheit: $h = \max_{i=1,\ldots,n} h_i$ mit $h_i = |I_i| = x_i - x_{i-1}$
- Vektorräume stückweise polynomieller Funktionen

$$S_n^{k,r}[a,b] = \{p \in C^r[a,b] \mid p\,|_{I_i} \in P_k(i_i)\}, i = 1, \ldots, n$$

Beispiel 2.15 (Stückweise lineare Interpolation) $\implies p \in S_n^{(1,0)}[a,b]$. Fehlerabschätzung:

$$\max_{x\in[a,b]}|f(x)-p(x)| \leq \frac{1}{2}h^2 \max_{x\in[a,b]}\left|f''(x)\right|$$

Beispiel 2.16 (Splines) Zweimal stetig differenzierbare, stückweise kubische Polynome. Motivation: Biegestab. Minimiere Biegeenergie

$$\int_{x_0}^{x_n} s''(x)^2 \mathrm{d}x$$

Definition 2.17 (Kubischer Spline) Eine Funktion $s_n : [a,b] \to \mathbb{R}$ heißt kubischer Spline bezüglich $a = x_0 < x_1 < \cdots < x_n = b$, wenn gilt

1. $s_n \in C^2[a,b]$
2. $s_n \mid I_i \in P_3, i = 1, \ldots, n$

Gilt zusätzlich

3. $s_n''(a) = s_n''(b) = 0$ so heißt s_n natürlicher Spline.

Existenz des interpolierenden kubischen Spline zu Knotenwerten $s_n(x_i) = y_i, i = 0, \ldots, n$

Satz 2.18 (Spline-Interpolation) Der interpolierende kubische Spline existiert und ist eindeutig bestimmt durch zusätzliche Vorgabe von $s_n''(a), s_n''(b)$

Beweis s hat die Form

$$s(x) \mid_{I_i} = p_i(x), i = 1, \ldots, n, p_i \in P_3(I_i)$$

4 Koeffizienten auf jedem der n Intervalle ergeben $4n$ Freiheitsgrade. Zur Bestimmung:

$s(x_i) = y_i, i = 0, \ldots, n$	$2n$ Gleichungen
$s' \in C[a, b]$	$n - 1$
$s'' \in C[a, b]$	$n - 1$
Zusatzbedingung für $s_n''(a), s_n''(b)$	2
	$4n$

$\Longrightarrow$ quadratisches lineares Gleichungssystem, $4n \times 4n$

$$N = \{\omega \in C^2[a, b] \mid \omega_{x_i} = 0, i = 0, \ldots, n\}$$

Seien $s_n^{(1)}$ und $s_n^{(2)}$ interpolierende Splines $\Longrightarrow$ $s = s_n^{(1)} - s_n^{(2)} \in N$. Für $\omega \in N$ beliebig:

$$\begin{aligned}
\int_a^b s''(x)\omega''(x)\mathrm{d}x &= \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} s''(x)\omega''(x)\mathrm{d}x \\
&= \sum_{i=0}^{n-1} [- \int_{x_i}^{x_{i+1}} s^{(3)}\omega' \mathrm{d}x + s''\omega' \mid_{x_i}^{x_{i+1}}] \\
&= \sum_{i=0}^{n-1} [- \int_{x_i}^{x_{i+1}} s^{(4)}\omega \mathrm{d}x - s^{(3)}\omega \mid_{x_i}^{x_{i+1}} + s''\omega' \mid_{x_i}^{x_{i+1}}] \\
&= \sum_{i=0}^{n-1} s''\omega' \mid_{x_i}^{x_{i+1}} = s''(x)\omega'(x) - s''(a)\omega'(a) \\
&= 0
\end{aligned}$$

Speziell für $\omega = s$

$$\int_a^b \left|s''(x)\right|^2 \mathrm{d}x = 0$$

$\Longrightarrow$ s ist linear $0 = s(a) = s(b) = 0$ □

Wiederholung: Extrapolation $a(h), h_i > 0, a(0) = \lim_{h \to 0} a(h)$ Fehler: Entwicklung

$$a(h) = a_0 + \sum_{j=1}^{n} a_j h^{a_j}$$

$$0 < \frac{h_{k+1}}{h_k} \leq \rho < 1$$

interpolieren $\left(h_{k+1}^a, a(h_{k+1})\right), i = 1, \ldots, n$

$$\Longrightarrow a(0) - p_i^{(k)}(0) = \mathcal{O}\left(h_k^{(n+1)}\right)$$

Splines: $S_h^{(k,r)}[a, b] = \{p \in C^r[a, b] \mid p\big|_{[x_i, x_{i+1}]} \in P_k[x_i, x_{i+1}]\}$ Splines: $s \in S_k^{(n,x)}[a, b]$. Natürliche kubische Splines: $s''(a) = s''(b) = 0$.

Satz 2.19 Für den interpolierenden natürlichen Spline S_n durch $x_0, \ldots, x_n, y_0, \ldots, y_n$ gilt

$$\int_a^b \left|S'(x)\right|^2 \mathrm{d}x \leq \int_a^b \left|g''(x)\right|^2 \mathrm{d}x$$

bezüglich allen Funktionen $g \in C^2[a,b]$ mit $g(x_i) = y_i, i = 0, \ldots, n$

Beweis Sei $N = \{\omega \in C^2[a,b] \mid \omega(x_i) = 0, i = 0, \ldots, n\} \implies \omega = g - I_n \in N$.

$$\begin{aligned}
\implies \int_a^b \left|g''(x)\right|^2 \mathrm{d}x &= \int_a^b \left|S_n''(x) + \omega''(x)\right|^2 \mathrm{d}x \\
&= \int_a^b \left|S_n''(x)\right|^2 \mathrm{d}x + \underbrace{2\int_a^b S_n''(x)\omega''(x)\mathrm{d}x}_{0} + \underbrace{\int_a^b \left|\omega''(x)\right|^2 \mathrm{d}x}_{\geq 0} \\
&\geq \int_a^b \left|S_n''(x)\right|^2 \mathrm{d}x
\end{aligned}$$

□

Satz 2.20 (Approximationsfehler) Sei $f \in C^4[a,b]$. Erfüllt der interpolierende kubische Spline $S_1''(a) = f''(a) \wedge S_n(b) = f''(b)$ so gilt:

$$\max_{x\in[a,b]} |f(x) - S_n(x)| \leq \frac{1}{2}h^4 \max_{x\in[a,b]} \left|f^{(4)}(x)\right|$$

Ohne Beweis.

2.5 Gauß Approximation

Wir betrachten $C[a,b]$, die Menge der stetigen Funktionen auf $[a,b]$ über dem Zahlenkörper $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$, als $\mathbb{K}$ -Vektorraum. Für $f, g \in [a,b]$ erfüllt

$$(f,g) := \int_a^b f(t)\overline{g(t)}\mathrm{d}t$$

die Eigenschaften eines Skalarproduktes:

1. Definitheit:
$$(f,f) = \int_a^b f(t)\overline{f(t)}\mathrm{d}t = \int_a^b |f(t)|^2 \mathrm{d}t \geq 0$$
und $(f,f) = 0 \implies f = 0$
2. $\alpha \in \mathbb{K}, h \in C[a,b]$:
$$(\alpha f + g, h) = \int_a^b (\alpha f(t) + g(t))\overline{h(t)}\mathrm{d}t = \alpha \int_a^b f(t)\overline{h(t)}\mathrm{d}t + \int_a^b g(t)\overline{h(t)}\mathrm{d}t = \alpha(f,h) + (g,h)$$
3. Symmetrie:
$$(f,g) = \int_a^b f(t)\overline{g(t)}\mathrm{d}t = \int_a^b \overline{\overline{f(t)}g(t)}\mathrm{d}t = \int_a^b g(t)\overline{f(t)}\mathrm{d}t = \overline{(g,f)}$$

Durch $\|f\| = \sqrt{(f,f)}$ ist damit eine Norm auf $C[a,b]$ gegeben:

1. Definitheit:
$$\|f\| \geq 0, f = 0 \iff \|f\| = 0$$

2. Sublinearität: Wir benutzen die Cauchy-Schwarz-Ungleichung
$$\begin{aligned} & |(f,g)| \leq \|f\|\|g\| \\ \implies & \|f+g\|^2 = (f+g, f+g) = (f,f) + (f,g) + (g,f) + (g,g) \\ & = \|f\|^2 + \underbrace{2\Re(f,g)}_{\leq |(f,g)|} + \|g\|^2 \\ & \leq \|f\|^2 + 2\|f\|\|g\| + \|g\|^2 = (\|f\| + \|g\|)^2 \\ \implies & \|f+g\| \leq \|f\| + \|g\| \end{aligned}$$ (Dreiecksungleichung)

3. Homogenität:
$$\|\alpha f\| = \sqrt{(\alpha f, \alpha f)} = \sqrt{\alpha\bar{\alpha}(f,f)} = |\alpha|\|f\|$$

Mit diesem Skalarprodukt und dieser Norm ist also $C[a,b]$ ein Prähilbertraum.

Satz 2.21 (Gauß-Approximation) Sei H ein Prähilbertraum und sei $S \subset H$ eine endlichdimensionaler Teilraum. Dann existiert zu jedem $f \in H$ eine eindeutig bestimmte „beste Approximation“ $g \in S$
$$\|f - g\| = \min_{\varphi \in S} \|f - \varphi\|$$

Beweis Vorüberlegung: Wenn $g \in S$ eine beste Approximation ist, so hat für $\varphi \in S$ die Hilfsfunktion
$$F_\varphi(t) = \|f - g - t\varphi\|^2, t \in \mathbb{R}$$
bei $t = 0$ ein Minimum. Somit ist
$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}t} F_\varphi(t)\big|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t}[(f-g-t\varphi, f-g-t\varphi)]\big|_{t=0} \\ &= \frac{\mathrm{d}}{\mathrm{d}t}[(f-g, f-g) - t(\varphi, f-g) - f(f-g,\varphi) + t^2(\varphi,\varphi)]\big|_{t=0} \\ &= 2\Re(f-g,\varphi) \forall \varphi \in S \end{aligned}$$

Falls $\mathbb{K} = \mathbb{C}$ ergibt testen mit $i\varphi$
$$0 = \Re(f-g, i\varphi) = -\Re(f-g,\varphi) = \Im(t-g,\varphi) \implies (f-g,\varphi) = 0 \forall \varphi \in S$$

Interpretation: Der Fehler $f - g$ ist orthogonal zum Teilraum S. Gilt umgekehrt die letzte Gleichung für ein $g \in S$, so gilt für $\varphi \in S$
$$\|f-g\|^2 = (f-g, f-g) = (f-g, f-\varphi) + \underbrace{(f-g,\varphi)}_{0}$$

Cauchy-Schwarz:
$$\begin{aligned} & \leq \|f-g\|\|f-\varphi\| \\ \implies & \|f-g\| \leq \inf_{\varphi \in S} \|f-\varphi\| \end{aligned}$$

$\implies g$ ist Bestapproximation.

Existenz und Eindeutigkeit: Da $n = \dim S < \infty$, besitzt S eine Basis $\{\varphi_1, \ldots, \varphi_n\}$. Jedes $g \in S$ hat eine eindeutige Darstellung

$$g = \sum_{i=1}^{n} \alpha_i \varphi_i$$

$$\implies \left(f - \sum_{i=1}^{n} \alpha_i \varphi, \varphi \right) = (f, \varphi) - \sum_{i=1}^{n} \alpha_i (\varphi_i, \varphi) = 0 \forall \varphi \in S$$

$$\implies \sum_{i=1}^{n} (\varphi_i, \varphi) \alpha_i = (f, \varphi_k), k = 1, \ldots, n$$

Dies ist ein lineares $n \times n$ Gleichungssystem. Notation: $A\alpha = B$ mit $\alpha = \overline{(\alpha_1, \ldots, \alpha_n)}^T \in \mathbb{K}^n, b \in \mathbb{K}^n, b_i = (f, \varphi_i), A \in \mathbb{K}^{n \times n}, A_{ki} = (\varphi_i, \varphi_k)$. A ist hermitesch wegen $(\varphi_i, \varphi_k) = \overline{(\varphi_k, \varphi_i)}$. Sei $\alpha \in \mathbb{K}^n$ beliebig. Wegen

$$\begin{aligned} \alpha^H A \alpha &= \sum_{k=1}^{n} \sum_{i=1}^{n} \bar{\alpha}_k (\varphi_i, \varphi_k) \alpha_i \\ &= \sum_{k=1}^{n} \sum_{i=1}^{n} (\alpha_i, \varphi_i, \alpha_k, \varphi_k) \\ &= \left(\sum_{i=1}^{n} \alpha_i \varphi_i, \sum_{k=1}^{n} \alpha_k \varphi_k \right) = (g, g) > 0 \end{aligned}$$

für $\alpha \neq 0 (\implies g \neq 0)$ ist A also positiv definit und damit invertierbar $\implies$ mit $\alpha = A^{-1} b$ löst das eindeutig bestimmte Gleichungssystem und g ist die Bestapproximation. □

Das lineare Gleichungssystem besitzt besonders einfache Lösung, wenn die Basis $\{\varphi_1, \ldots, \varphi_n\}$ eine Orthogonalbasis ist, das heißt $(\varphi_i, \varphi_j) = \delta_{ij}$

$$\implies \alpha_i = (f, \varphi_i), i = 1, \ldots, n$$

$$\implies g = \sum_{i=1}^{n} (f, \varphi_i) \varphi_i \quad \text{ist Bestapproximation}$$

Lemma 2.22 (Gram-Schmidt-Algorithmus) Zu jeder Basis $\{\psi_1, \ldots, \psi_k\}$ von S lässt sich eine Orthonormalbasis $\{\varphi_1, \ldots, \varphi_n\}$ konstruieren.

$$\tilde{\varphi}_1 = \psi_1, \varphi_1 = \frac{\tilde{\varphi}_1}{\|\tilde{\varphi}_1\|}$$

$$\tilde{\varphi}_k = \psi_k - \sum_{i=1}^{k-1} (\psi_k, \varphi_i), \varphi_k = \frac{\tilde{\varphi}_k}{\|\tilde{\varphi}_k\|}$$

Beweis Per Induktion nach n.
$n = 1$: Da $\psi \neq 0$ gilt $(\varphi_1, \varphi_1) = \frac{|\psi_1|^2}{\|\psi_1\|^2} = 1$.
$n > 1$: Sei $\{\varphi_1, \ldots, \varphi_n\}$ eine Orthonormalbasis. Es gilt

$$0 \neq \tilde{\varphi}_n = \psi_n - \sum_{k=1}^{n-1} (\psi_n, \varphi_k) \varphi_k$$

da sonst $\{\psi_1, \dots, \psi_n\}$ linear abhängig wären. Für $i = 1, \dots, n-1$ gilt

$$(\varphi_n, \varphi_1) = (\psi_n, \varphi_i) - \sum_{k=1}^{n-1} (\psi_n, \varphi_k) \underbrace{(\varphi_k, \varphi_i)}_{\delta_{ik}} = 0$$

und $\|\varphi_n\|^2 = 1$ nach Konstruktion. □

Wiederholung: Gauß-Approximation, Prähilbertraum H, Teilraum $S \subset H, \dim S = n < \infty$

$$\forall f \in H \exists! g \in S : \|f - g\| \leq \min_{\varphi \in S} \|f - \varphi\|$$

Äquivalent: $e := f - g \perp S \iff (f - g, \varphi) = 0 \forall \varphi \in S$
Orthogonalisiere Basis $\{\psi_1, \dots, \psi_n\}$ von S mit Gram-Schmidt

$$\tilde{\varphi}_i = \begin{cases} \psi_i & i = 1 \\ \psi_i - \sum_{j=1}^{i-1} \frac{(\psi_i, \tilde{\varphi}_j)}{\|\tilde{\varphi}_j\|^2} \tilde{\varphi}_j & i = 2, \dots, n \end{cases}$$

Normalisieren: $\varphi_k = \|\tilde{\varphi}_n\|^{-1} \tilde{\varphi}_k$. $(\varphi_1, \dots, \varphi_k)$ Orthogonalbasis $\implies (\varphi_i, \varphi_j) = \delta_{ij}$

$$\implies g = \sum_{k=1}^{n} (f, \varphi_k) \varphi_k$$

Erinnerung:

$$f[x_0, \dots, x_n] = \frac{f[x_1, \dots, x_n] - f[x_0, \dots, x_{n-1}]}{x_n - x_0}, \quad f[g_k] = f(k)$$

3 Numerische Integration

Approximation von bestimmten Integralen reeller Funktionen $f \in C[a, b]$ durch Quadraturformeln

$$I(f) = \int_a^b f(x) \mathrm{d}x \approx I^{(n)}(f) = \sum_{i=1}^{n} \alpha_i f(x_i)$$

mit Stützstellen $a \leq x_0 < x_1 < \dots < x_n \leq b$ und Gewichten $\alpha_i \in \mathbb{R}$.

Beispiel 3.1 (Summierte Rechteckregel)

$$\int_a^b f(x) \mathrm{d}x \approx \sum_{i=0}^{n-1} (x_{i+1} - x_i) f(x_i)$$

Interpolatorische Quadraturformeln.
Idee: Interpoliere f durch ein Interpolationspolynom auf $[a, b]$!

$$p_n(x) = \sum_{i=0}^{n} f(x_i) L_i^{(n)}(x)$$

$$\implies I^{(n)}(f) = \int_a^b p_n(x) \mathrm{d}x = \sum_{i=0}^{n} f(x_i) \int_a^b \underbrace{L_i^{(n)}(x) \mathrm{d}x}_{=\alpha_i}$$

Quadraturgewichte hängen nur von $a, x_0, \dots, x_n, b$ ab.

Satz 3.2 (Lagrange-Quadratur) Für interpolatorische Quadraturformeln gilt

$$I(f) - I^{(n)}(f) = \int_a^b f[x_0, \ldots, x_n, x] \prod_{i=0}^{n} (x - x_i) \mathrm{d}x$$

Beweis Restglieddarstellung der Interpolation. □

Definition 3.3 Eine Quadraturformel $I^{(n)}$ wird „von der Ordnung m" genannt, wenn sie alle $p \in P_{m-1}$ exakt integriert. Das heißt

$$\int_a^b p(x) \mathrm{d}x = I^{(n)}(p) \forall p \in P_{m-1}$$

$\Longrightarrow$ Interpolatorische Quadraturformeln zu $n+1$ Stützstellen sind (mindestens) von der Ordnung $n+1$.

Spezialfall: Äquidistante Stützstellen: Newton-Cotes-Formeln:

1. Abgeschlossene Formeln ($H = \frac{b-a}{n}, x_i = a + iH, a = x_0, b = x_n$)

$$I^{(1)}(f) = \frac{b-a}{2}[f(a) + f(b)] \qquad \text{(Trapezregel)}$$

$$I^{(2)}(f) = \frac{b-a}{6}[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)] \qquad \text{(Simpsonregel, Keplersche Fassregel)}$$

$$I^{(3)}(f) = \frac{b-a}{8}[f(a) + 3f(a+H) + 3f(b-H) + f(b)] \qquad (3/8 \text{ Regel})$$

2. Offene Formeln $\left(H = \frac{b-a}{n+2}, x_i = a + (i+1)H, a < x_0, x_n < b\right)$

$$I^{(0)}(f) = (b-a) f\left(\frac{a+b}{2}\right) \qquad \text{(Mittelpunktregel)}$$

$$I^{(1)}(f) = \frac{(b-a)}{2}(f(a+H) + f(b-H))$$

$$I^{(1)}(f) = \frac{(b-a)}{3}\left(2f(a+H) - f\left(\frac{a+b}{2}\right) + 2f(b-H)\right)$$

Satz 3.4 (Quadraturrestglieder)

1. Trapezregel: Für jedes $f \in C^2[a,b]$ gibt es ein $\xi \in [a,b]$ mit

$$\int_a^b f(x) \mathrm{d}x - \frac{b-a}{2}[f(a) + f(b)] = -\frac{(b-a)^3}{12} f''(\xi)$$

2. Simpson-Regel: Für jedes $f \in C^4[a,b] \exists \xi \in [a,b]$ sodass

$$\int_a^b f(x) \mathrm{d}x - \frac{b-a}{6}[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)] = -\frac{(b-a)^5}{2880} f^{(4)}(\xi)$$

3. Mittelpunktregel: $\forall f \in C^2[a,b] \exists \xi \in [a,b]$ sodass

$$\int_a^b f(x) \mathrm{d}x - (b-a) f\left(\frac{a+b}{2}\right) = \frac{(b-a)^3}{24} f''(\xi)$$

Satz 3.5 (Verallgemeinerter Mittelwertsatz) Sei $f \in C[a,b], g \geq 0$ oder $g \leq 0$ integrierbar. Dann $\exists \xi \in [a,b]$, sodass

$$\int_a^b f(x)g(x)\mathrm{d}x = f(\xi)\int_a^b g(x)\mathrm{d}x$$

Beweis (Beweis der Quadraturrestglieder).

1. Für $x \in [a,b]$ ist $(x-a)(x-b) \leq 0$

$$\implies I(f) - I^{(1)}(f) = \int_a^b f[x_0, x_1, x] \prod_{i=1}^{1}(x - x_i)\mathrm{d}x$$

Verallgemeinerter Mittelwertsatz: $\exists \xi \in [a,b]$, sodass

$$\begin{aligned} &= \frac{f''(\xi)}{2!}\left(-\frac{1}{6}(b-a)^3\right) \\ &= -\frac{f''(\xi)}{12}(b-a)^3 \end{aligned}$$

2.

$$\begin{aligned} I(f) - I^{(2)}(f) &= \int_a^b f[a, \frac{a+b}{2}, b, x](x-a)\left(x - \frac{a+b}{2}\right)(x-b) \\ &= \int_a^b \frac{f[a, \frac{a+b}{2}, b, x] - f[\frac{a+b}{2}, a, \frac{a+b}{2}, b]}{x - \frac{a+b}{2}}(x-a)\left(x - \frac{a+b}{2}\right)^2(x-b)\mathrm{d}x + f[\frac{a+b}{2}, a, \frac{a+b}{2}, b]\int_a \\ &= \frac{f^{(4)}(\xi)}{4!}\int_a^b (x-a)\left(x - \frac{a+b}{2}\right)^2(x-b)\mathrm{d}x \\ &= -\frac{f^{(4)}(\xi)}{2880}(b-a)^5 \end{aligned}$$

3. analog zu 2. □

Probleme:

- negative Gewichte α_i ab $n = 7$ (geschlossen) und $n = 2$ (offen) $\implies$ Auslöschungsgefahr
- Oszillationen des Lagrange-Interpolanten für äquidistante Gitter (Runge-Phänomen), im Allgemeinen $I^{(n)}(f) \not\to I(f), n \to \infty$

Abhilfe: Summierte Quadraturformeln

$$I_n^{(n)}(f) = \sum_{i=1}^{N-1} I_{[x_i, x_i+1]}^{(n)}(f), h = \frac{b-a}{N}, x_i = a + ih$$

Gilt die lokale Fehlerdarstellung:

$$I_{[x_i,x_{i+1}]}(f) - I_{[x_i,x_{i+1}]}^{(n)}(f) = \omega_n h^{n+2} f^{(m+1)}(\xi_i), \quad \xi_i \in [a,b]$$

für $m \geq n$ gilt:

$$\begin{aligned} I(f) - I_n^{(n)}(f) &= \sum_{i=0}^{N-1} [I_{[x_i,x_{i+1}]}(f) - I^{(n)}_{[x_i,x_{i+1}]}(f)] \\ &= \omega_n h^{m+2} N \underbrace{\sum_{i=0}^{N-1} \frac{f^{(m+1)}(\xi_i)}{N}}_{\in [\min_i f^{(m+1)}(\xi_i), \max_i f^{(m+1)}(\xi_i)]} \\ &= \omega_n h^{m+2} N f^{(m+1)}(\xi) \qquad \text{(für ein } \xi \in [a,b] \text{ (Verallg. Mittelwertsatz))} \\ &= \omega_n h^{(m+1)} (b-a) f^{(m+1)}(\xi) \end{aligned}$$

Beispiel 3.6 1. Summierte Trapezregel $(m = 1)$

$$\begin{aligned} I_h^{(1)} &= \sum_{i=0}^{N-1} \frac{x_{i+1} - x_i}{2} [f(x_i) + f(x_{i+1})] \\ &= \frac{h}{2} f(a) + h \sum_{i=1}^{N-1} f(x_i) + \frac{h}{2} f(b) \\ I(f) - I_h^{(n)}(f) &= -\frac{b-a}{12} h^2 f''(\xi), \xi \in [a,b] \end{aligned}$$

2. Summierte Simpson-Regel $(m = 3)$

$$\begin{aligned} I_h^{(2)}(f) &= \sum_{i=0}^{N-1} \frac{x_{i+1} - x_i}{6} [f(x_i) + 4f\left(\frac{x_i + x_{i+1}}{2}\right) + f(x_{i+1})] \\ &= \frac{h}{6} [f(a) + 2 \sum_{i=1}^{N-1} f(x_i) + 4 \sum_{i=0}^{N-1} f\left(\frac{x_i + x_{i+1}}{2}\right) + f(b)] \\ I(f) - I_h^{(2)}(f) &= -\frac{b-a}{2880} h^4 f^{(4)}(\xi), \xi \in [a,b] \end{aligned}$$

3. Summierte Mittelpunktregel $(m = 1)$

$$\begin{aligned} I_h^{(0)}(f) &= \sum_{i=0}^{N-1} (x_{i+1} - x_i) f\left(\frac{x_i + x_{i+1}}{2}\right) = h \sum_{i=0}^{N-1} f\left(\frac{x_i + x_{i+1}}{2}\right) \\ I(f) - I_h^{(0)}(f) &= \frac{b-a}{24} h^2 f''(\xi), \quad \xi \in [a,b] \end{aligned}$$

Wiederholung Quadratur

$$\int_a^b f(x) \mathrm{d}x \approx \sum_{i=0}^{n} \alpha_i f(x_i) = I^{(n)f}$$

- Interpolatorische Quadraturregel, Äquidistante Stützstellen
  $\rightarrow$ Newton-Cotes Formeln (abgeschlossen, offen)
- Summierte Formeln $x_i = a + iH, H > 0$

$$I_H^{(n)}(f) = \sum_{i=1}^{n} I^{(n)}_{[x_{i-1},x_i]}(f)$$

3.1 Gaußsche Quadraturformeln

Frage: Wie wählt man x_i in

$$I^{(n)}(f) = \sum_{i=0}^{N} \alpha_i f(x_i)$$

optimal? Nach Konstruktion ist $I^{(n)}$ mindestens von der Ordnung $n+1$

Lemma 3.7 Interpolatorische Quadraturformeln sind höchstens von der Ordnung $2n+2$

Beweis Wähle

$$p(x) = \prod_{i=0}^{n} (x - x_i)^2 \in P_{2n+2}$$
$$\implies 0 < \int_a^b p(x)\mathrm{d}x = \sum_{i=0}^{n} \alpha_i \underbrace{p(x_i)}_{0} = 0\text{`}$$ □

Gaußsche Quadraturformen erreichen die Maximalordnung $2n+2$ (exakt für $p \in P_{2n+1}$) Herleitung: Für $x_0, \ldots, x_n, x_{n+1}, \ldots, x_{2n+1} \in [a,b]$ betrachte $I^{(n)}(t)$ und Î{(2n + 1)}(f)

$$I(f) - I^{(2n+1)}(f) = I(f) - \sum_{i=0}^{2n+1} f[x_0, \ldots, x_i]\Big|_a^b \int_a^b \prod_{j=0}^{i-1}(x - x_j)\mathrm{d}x$$
$$= I(f) - I^{(n)}(f) - \sum_{i=n+1}^{2n+1} f[x_0, \ldots, x_i] \int_a^b \prod_{j=0}^{i-1}(x - x_j)\mathrm{d}x$$

Für $i > n$ gilt

$$\int_a^b \prod_{j=0}^{i-1}(x - x_j)\mathrm{d}x = \int_a^b \underbrace{\prod_{j=0}^{n}(x - x_j)}_{P_{n+1}} \underbrace{\prod_{j=n+1}^{i-1}(x - x_j)}_{\in P_n} \mathrm{d}x$$

Wähle Stützstellen so, dass

$$0 = \int_a^b \prod_{j=0}^{n}(x - x_j) q(x)\mathrm{d}x = \left(\prod_{j=0}^{n}(x - x_j), q\right) \forall q \in P_n$$
$$I(f) - I^{(n)}(f) = I(f) - I^{(2n+1)}(f)$$

$\implies I^{(n)}$ ist exakt für $p \in P_{2n+1}$, das heißt von Ordnung $2n+2$. Mit einem Orthogonalsystem $\{p_0, \ldots, p_{n+1}\}$ von P_{n+1} sind die Nullstellen $\lambda_0, \ldots, \lambda_n$ von p_{n+1} von Interesse. Frage: Sind die Nullstellen von p_{n+1} reell, einfach und in $[a,b]$?

Satz 3.8 Gegeben sei ein Skalarprodukt auf $C[a,b]$

$$(f,g)_\omega = \int_a^b f(x)g(x)\omega(x)\mathrm{d}x$$

mit integrierbarer Gewichtsfunktion $\omega(x) \geq 0, x \in (a,b)$ mit höchstens endlich vielen Nullstellen. Dann haben die mittels Gram-Schmidt aus $\{1, x^1, \ldots\}$ bezüglich $(\cdot,\cdot)_\omega$ orthogonalisierten Polynome $\{p_0, p_1, \ldots\}$ lauter reelle, einfache Nullstellen in $[a,b]$

Beweis Sei $N_n := \{\lambda \in (a,b) \mid \lambda \text{ Nullstelle ungerader Vielfachheit von } p_n\}$. Setze

$$q(x) = \begin{cases} 1 & N_n \neq \emptyset \\ \prod_{i=1}^{m}(x-\lambda_i) & N_n = \{\lambda_1, \ldots, \lambda_m\}, m > 0 \end{cases}$$

Nach dem Fundamentalsatz der Algebra und wegen $p(x) = x^n - r(x), r \in P_{n-1}$, nach Konstruktion mit Gram-Schmidt (ohne Normalisieren) gilt

$$p_n(x) = \prod_{i=1}^{n}(x-\lambda_i), \lambda_i \in \mathbb{C}, i = 1, \ldots, n$$

Ist λ_I nicht reell, so ist $\bar{\lambda}_i$ auch eine Nullstellen von p_N und

$$(x-\lambda_i)x - \bar{\lambda}_i = (x-\lambda_I)(x-\lambda_i) \implies |x-\lambda_i|^2 \geq 0$$

$\implies p_n q \in P_{n+m}$ ist reell und hat in $[a,b]$ keinen Vorzeichenwechsel.

$$(p_n, q)_\omega = \int_a^b p_n(x)(x)\omega(x)\mathrm{d}k \neq 0$$

Für $m < n$ ist das ein Widerspruch zu $p_n \perp p_{n-1} \implies \mu_n = \{\lambda_1, \ldots, \lambda_n\}$. Für $[a,b] = [-1,1]$ und $\omega \equiv 1$, das heißt $(\cdot,\cdot)_\omega = (\cdot,\cdot)_2$ sind die p_n mittels $p_n(x) = x^n + \ldots$ normierte Legendre-Polynome L_n(x). Wir wählen also die Nullstellen $\zeta_0, \ldots, \lambda_n$ von p_{n+1} beziehungsweise L_{n+2} als Stützstellen einer interpolatorischen Quadraturformel auf $[-1,1]$.

$$I^{(n)}(f) = \sum_{i=9}^{n} \alpha_i f(\lambda_i), \alpha_i = \int_{-1}^{1} \prod_{\substack{j=0 \\ j\neq i}} \frac{x-\lambda_j}{\lambda_i - \lambda_j}\mathrm{d}x$$

□

Satz 3.9 (Gauß-Quadratur) Es gibt genau eine interpolatorische Quadraturformel zu $n+1$ paarweise verschiedenen Stützstellen auf $[-1,b]$ mit Ordnung $2n+2$. Ihre Stützstellen sind gerade die Nullstellen. $\lambda_0, \ldots, \lambda_n \in (-1,1)$ das $(n+1)$ - ten Legendre Polynom $L_{n+1} \in P_{n+1}$ und die Gewichte erfüllen

$$\alpha_i = \int_{-1}^{1} \prod_{\substack{j=0 \\ j\neq i}} \left(\frac{x-\lambda_j}{\lambda_i - \lambda_j}\right)^2 \mathrm{d}x > 0, i = 0, \ldots, n$$

Für $f \in C^{2n+2}[-1,1]$ besitzt des Restglied die Darstellung

$$R^{(n)} = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_{-1}^{1} \prod_{j=0}^{n}(x-\lambda_j)^2 \mathrm{d}x, \xi \in (-1,1)$$

Beweis Existenz: Es gilt $p_{n+1} \perp P_n$ Für $\omega = 1$ und $p_n(x) = \prod_{i=0}^{n}(x-\lambda_i) = x^n + \ldots$

$$\implies I^{(n)}(f) = I^{(2n+1)}(f)$$

$\implies I^{(n)}$ hat Ordnung $2n+2$. Gewichte:

$$L_i^{(x)}(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x-\lambda_j}{\lambda_i - \lambda_j} \in P_n$$

$$\implies \left(L_i^{(n)}(x)\right)^2 \in P_{2n}$$

$$\implies 0 < \int_{-1}^{1} \left(L_i^{(n)}\right)^2 \mathrm{d}x = \sum_{j=0}^{n} \alpha_j \underbrace{\left(L_i^{(n)}(x_i)\right)}_{\delta_{ij}} = \alpha_i$$

Eindeutigkeit: Sei $\tilde{I}^{(n)}(f) = \sum_{i=0}^{n} \tilde{a}_I f\left(\tilde{\lambda}_i\right)$ ebenfalls der Ordnung $2n+2$. Wie oben folgt $\tilde{\alpha}_i > 0$ mithilfe

$$\tilde{L}_i^{(n)}(x) = \prod_{j=0 j\neq i}^{n} \frac{n-\tilde{\lambda}_j}{\tilde{\lambda}_i - \tilde{\lambda}_j}$$

$$\begin{aligned} 0 &= \int_{-1}^{1} \frac{1}{\tilde{\alpha}_i} \tilde{L}_i^{(n)} p_{n+1}(x)\mathrm{d}x \\ &= \sum_{j=0}^{n} \frac{\tilde{\alpha}_i}{\tilde{\alpha}_i} \underbrace{\tilde{L}_i^{(n)}\left(\tilde{\lambda}_j\right)}_{\delta_{ij}} p_{n+1}\left(\tilde{\lambda}_j\right) = p_{n+1}\left(\tilde{\lambda}_i\right), i = 0, \ldots, n \end{aligned}$$

$\implies \tilde{\lambda}_i = \lambda_i$ und $\tilde{\alpha}_i = \alpha_i, i = 1, \ldots, n$.
Restglied: Für $f \in C^{(2n+2)}[-1,1]$ hat der Hermite-Interpolant $h \in P_{2n+1}$ zu den Bedingungen

$$h(\lambda_i) = f(\lambda_i), h'(\lambda_i) = f'(\lambda_i), i = 0, \ldots, n$$

die Darstellung:

$$f(x) - h(x) = f[\lambda_0, \lambda_0, \ldots, \lambda_n, \lambda_n, x] \prod_{i=0}^{n} (x-\lambda_i)^2$$

$$\begin{aligned} \implies I(f) - I^{(f)} &= I(f) - \underbrace{I^{(n)}(h)}_{=I(h)} - \left(I^{(n)}(f) - I^{(n)}(h)\right) \\ &= I(f-h) - I^{(n)}(f-h) \\ &= \int_{-1}^{1} f[\lambda_0, \lambda_0, \ldots, \lambda_n, \lambda_n] \underbrace{\prod_{i=0}^{n} (x-\lambda_i)^2}_{>0} \mathrm{d}x - \underbrace{\sum_{i=0}^{n} \alpha_i [f(\lambda_i) - h(\lambda_i)]}_{0} \end{aligned}$$

Mit verallgemeinertem Mittelwertsatz folgt:

$$= \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_{-1}^{1} \prod_{i=0}^{n} (x-\lambda_i)^2 \mathrm{d}x \qquad \square$$

Die $\lambda_i^{(n)}$ (Nullstellen von p_{n+1}) und die dazugehörigen α_i lassen sich tabellieren. Durch Transformation von $[a,b]$ auf $[-1,1]$ erhält man eine allgemeine Quadraturformel.

Satz 3.10 (Konvergenz der Gauß-Quadratur) Sei $I^{(n)}(f)$ die $(n+1)$ punktige Gauß-Formel zur Berechnung von $I(f) = \int_{-1}^{1} f(x)\mathrm{d}x$. Für jedes $f \in C[-1,1]$ konvergiert $I^{(n)}(f) \xrightarrow{n\to\infty} I(f)$

Beweis Es gilt

$$I^{(n)}(f) = \sum_{i=0}^{n} \alpha_i^{(n)} f\left(\lambda_i^{(n)}\right), \alpha_i^{(n)} > 0, \sum_{i=0}^{n} \alpha_i^{(n)} = 2$$

Sei $\varepsilon > 0$. Nach dem Weierstrassschem Approximationssatz gibt es $p_\varepsilon \in P_n$ mit

$$\max_{x\in[-1,1]} |f(x) - p_\varepsilon(x)| \leq \frac{\varepsilon}{4}$$

Für $n > \frac{1}{2}m - 1$ (das heißt $2n + 2 > m$) gilt

$$\left|I(f) - I^{(n)}(f)\right| \leq \underbrace{|I(f - p_\varepsilon)|}_{\leq \frac{\varepsilon}{4}2} + \underbrace{\left|I(p_\varepsilon) - I^{(n)}(p_\varepsilon)\right|}_{0} + \underbrace{\left|I^{(n)}(f - p_\varepsilon)\right|}_{\leq \frac{\varepsilon}{4}2} \leq \varepsilon$$

□

Wiederholung: Gauß-Quadratur

- $n + 1$ Stützstellen, Ordnung $2n + 2$ (optimal)
- x_i Nullstellen des Legendre Polynoms p_{n+1}
- $I^{(n)}(f) \xrightarrow{n\to\infty} I(f)$ für f stetig
- Verallgemeinerung auf gewichtete Integrale

$$\int_a^b f(x)\omega(x)\mathrm{d}x I(f\omega) I_\omega(f)$$

$\implies$ Orthogonalisiere bezüglich

$$(f, g)_\omega = \int_a^b f(x)g(x)\omega(x)\mathrm{d}x$$

3.2 Praktische Aspekte der Quadratur

Ziel: Möglichst hohe Genauigkeit bei möglichst wenig Funktionsauswertungen. Schwierigkeiten:

- Fehlerabschätzung: $f^{(k)}$ nur schwer zugänglich für $k > 2 \implies$ a-posteriori Fehlerschätzer.

  Beispiel 3.11
  1. Vergleiche $I_n(f)$ und $I_{\frac{n}{2}}(f)$ bei summierten Quadraturformeln
  2. Extrapolationsfehler

- Wiederbenutzung bereits berechneter Werte von f
  - schwierig bei Gauß
  - einfach bei Newton-Cotes

4 Lineare Gleichungssystem

Gegeben: $A \in \mathbb{R}^{m\times n} = (a_{ij}), b \in \mathbb{R}^m$. Gesucht: $x \in \mathbb{R}^n$ mit $Ax = b \implies m$ Gleichungen, n Unbekannte. Das lineare Gleichungssystem $Ax = b$ heißt

- unterbestimmt, falls $m < n$
- überbestimmt falls $m > n$
- quadratisch falls $m = n$

Störungsstheorie:

- Konditionierung von quadratischen linearen Gleichungssystemen
- Fehlereinfluss von Datenfehlern und Rundungsfehlern auf Lösung x

Vektor- und Matrizennormen:
Sei $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$. Erinnerung: Eigenschaften einer Norm: $\|\cdot\| : \mathbb{K}^n \to \mathbb{R}$

- Definitheit: $\|x\| > 0 \forall x \in \mathbb{K}^n \setminus \{0\}$
- Positive Homogenität: $\|\alpha x\| = |\alpha| \|x\| \forall x \in \mathbb{K}^n, \alpha \in \mathbb{K}$
- Subadditivität: $\|x + y\| \leq \|x\| + \|y\| \forall x, y \in \mathbb{K}^n$

Beispiel 4.1 Euklidische Norm: (l_2)

$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2}$$

Maximumsnorm (l_∞)

$$\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$$

l_1 -Norm:

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

l_p -Norm, $p \geq 1, p < \infty$

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

Betrachte Vektorraum der $n \times n$ -Matrizen $A \in \mathbb{K}^{n \times n}$

Definition 4.2 Eine Norm $\|\cdot\|$ auf $\mathbb{K}^{n \times n}$ heißt verträglich mit einer Vektornorm $\|\cdot\|$ auf $\mathbb{K}^n$, wenn gilt:

$$\|Ax\| \leq \|A\| \|x\| \forall x \in \mathbb{K}^n, A \in \mathbb{K}^{n \times n}$$

Sie heißt Matrizennorm, wenn sie submultiplikativ ist

$$\|AB\| \leq \|A\| \|B\| \forall A, B \in \mathbb{K}^{n \times n}$$

Beispiel 4.3 Die Frobeniusnorm

$$\|A\|_{Fr} = \left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2}$$

ist eine mit $\|\cdot\|_2$ verträgliche Matrizennorm.
Die natürliche Matrizennorm

$$\|A\| = \sup_{x \in \mathbb{K}^n \setminus \{0\}} \frac{\|Ax\|}{\|x\|} = \sup_{\substack{x \in \mathbb{K}^n \\ \|x\|=1}} \|Ax\|$$

ist eine mit $\|\cdot\|$ verträgliche Matrizennorm (Übung!). Es gilt

$$\|\mathbb{I}\| = \sup_{\substack{x \in \mathbb{K}^n \\ \|x\|=1}} \|\mathbb{I}x\| = 1$$

Lemma 4.4 Die natürlichen Matrizennormen zu $\|\cdot\|_\infty$ und $\|\cdot\|_1$ sind die „maximale Zeilen-/Spaltensumme“:

$$\|A\|_\infty = \max_{j=1,\dots,n} \sum_{k=1}^{n} |a_{jk}|$$

$$\|A\|_1 = \max_{k=1,\dots,n} \sum_{j=1}^{n} |a_{jk}|$$

Beweis Skript. □

Betrachte: $Ax = b$ und Störung

$$\underbrace{(A + \delta A)}_{\tilde{A}} \underbrace{(x + \delta x)}_{\tilde{x}} = \underbrace{b + \delta b}_{\tilde{b}}$$

Satz 4.5 (Neumann-Reihe) Gilt $\|A\| < 1$, so

$$\mathbb{I} - \mathbb{A} \sum_{k=0}^{\infty} A^k = \mathbb{I}$$

Beweis Für die Partialsummen gilt

$$(\mathbb{I} - A) \sum_{k=0}^{n} A^k = \mathbb{I} - A + A - A^2 + A^2 \cdots - A^{n+1} \xrightarrow{n\to\infty} \mathbb{I}$$

wegen $\left\|A^k\right\| \leq \|A\|^k \xrightarrow{k\to\infty} 0$. □

Wiederholung: Kondition numerischer Aufgabe $y = f(x), y \in \mathbb{R}^n, x \in \mathbb{R}^m$.

$$\frac{\Delta y_i}{y_i} \doteq \sum_{j=1}^{m} \frac{\partial f_i}{\partial x_j}(x) \frac{\Delta x_j}{y_i} = \sum_{j=1}^{m} \underbrace{\frac{\partial f_i}{\partial x_j} \frac{x_j}{f_i(x)}}_{=:k_{ij}(x)} \frac{\Delta x_j}{x_j}$$

Neumann-Reihe:

$$\|A\| < 1 \implies (\mathbb{1} - A_a)^{-1} = \sum_{n=0}^{\infty} A^n$$

Natürliche Matrixnorm:

$$\|A\| = \sup_{\|x\|=1} \|Ax\|$$

$\|A\|_\infty$: „Zeilensummennorm“
$\|A\|_1$: „Spaltensummennorm“
Euklidisches Skalarprodukt auf $\mathbb{K}$

$$(x, y)_2 = \bar{y}^T x$$

Lemma 4.6 (Spektralnorm) Für $A.\mathbb{K}^{n\times n}$ ist

$$\|A\|_2 = \max\{\sqrt{|\lambda|} \mid \lambda \text{ Eigenwert von } \bar{A}^T A\}$$

Für hermitesche $A = \bar{A}^T$ gilt:

$$\|A\|_2 = \max\{|\lambda| \mid \lambda \text{ Eigenwert von } A\}$$

Beweis $B = \bar{A}^T A$ ist hermitesch. $\implies$ B hat n reelle Eigenwerte $\lambda_1, \ldots, \lambda_n$ und eine Orthonormalbasis von Eigenvektoren $\{w_1, \ldots, w_n\} \subset \mathbb{K}^n$ $B\omega_i = \lambda_i \omega_i$. Jedes $x \in \mathbb{K}^n$ hat eine eindeutige Darstellung

$$x = \sum_{i=1}^{n} \alpha_i \omega_i$$

$$\begin{aligned}
\implies \|x\|_2^2 = (x,x)_2 &= \sum_{i,j=1}^{n} \alpha_i \bar{\alpha}_j \underbrace{(\omega_i, \omega_j)_2}_{\delta_{ij}} = \sum_{i=1}^{n} |\alpha_i|^2 \\
\|Ax\|_2^2 = (Bx, Bx)_2 &= \sum_{i,j=1}^{n} \lambda_i \alpha_i \overline{(\lambda_j \alpha_j)} \underbrace{\omega_i, \omega_j}_{\delta_{ij}} \\
&= \sum_{i=1}^{n} |\lambda_i|^2 |\alpha_i|^2 \\
\|B\|_2^2 = \sup_{x \in \mathbb{K}^n \setminus \{0\}} \frac{\|Bx\|_2^2}{\|x\|_2^2} &= \sup_{x \in \mathbb{K}^n \setminus \{0\}} \frac{\sum_{i=1}^{n} \lambda_i^2 |\alpha_i|^2}{\sum_{i=1}^{n} |\alpha_i|^2} \\
&\leq \max_{i=1,\ldots,n} |\lambda_i|^2
\end{aligned}$$

Mit

$$\begin{aligned}
|\lambda_i| &= |\lambda_i| \|\omega_i\|_2 = \|\lambda_i \omega_i\|_2 = \|B\omega_i\|_2 \\
&\leq \|B\|_2 \|\omega_i\|_2 = \|B\|_2, \quad i = 1, \ldots, n
\end{aligned}$$

□

Betrachte $Ax = b$ und Störung

$$\underbrace{(A + \delta A)}_{\tilde{A}} \underbrace{(x + \delta x)}_{\tilde{x}} = \underbrace{b + \delta b}_{\tilde{b}}$$

Satz 4.7 (Störungssatz) Die Matrix $A \in \mathbb{K}^{n \times n}$ sei regulär uns es sei

$$\|\delta A\| \leq \frac{1}{\|A^{-1}\|}$$

Dann ist die gestörte Matrix $\tilde{A} = A + \delta A$ ebenfalls regulär. Für den relativen Fehler der Lösung gilt mit die Konditionszahl von A

$$\operatorname{cond}(A) = \|A\| \left\|A^{-1}\right\|$$

die Ungleichung

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\operatorname{cond}(A)}{1 - \operatorname{cond}(A) \frac{\|\delta A\|}{\|A\|}} \left[\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right]$$

Beweis

$$\left\|A^{-1} \delta A\right\| \leq \left\|A^{-1}\right\| \|\delta A\| < 1$$

Neumann $\implies$ $A + \delta A = A[\not{\text{}} + A^{-1}\delta A]$ ist regulär.
$(A + \delta A)\tilde{x} = b + \delta b, (A + \delta A)x = b + \delta A x$

$$\implies (A + \delta A)\delta x = \delta b - \delta A x$$

$$\begin{aligned}\left\|(A+\delta A)^{-1}\right\| &= \left\|[A(\not{1} + A^{-1})]^{-1}\right\| \\ &= \left\|\left(\not{1} + A^{-1}\delta A\right)^{-1}A^{-1}\right\| \leq \left\|\sum_{n=0}^{\infty}\left(-A^{-1}\delta A\right)^n\right\|\left\|A^{-1}\right\| \\ &\leq \left(\sum_{n=0}^{\infty}\left\|A^{-1}SA\right\|\right)\left\|A^{-1}\right\| = \frac{1}{1-\left\|A^{-1}\delta A\right\|}\left\|A^{-1}\right\|\end{aligned}$$

$$\begin{aligned}\|b\| &= \|Ax\| \leq \|A\|\|x\| \\ \|\delta x\| &\leq \left\|(A+\delta A)^{-1}\right\|[\|\delta b\| + \|\delta A\|\|W\|] \\ &\leq \frac{\left\|A^{-1}\right\|}{1-\|A^{-1}\delta a\|}[\|\delta B\|\|bA\|\|x\|] \\ &\leq \frac{\left\|A^{-1}\right\|}{1-\|A^{-1}\|\|\delta A\|\|A\|\|A\|^{-1}}\left[\frac{\|\delta b\|}{\|x\|} + \frac{\|SA\|}{\|A\|}\right] \\ &\frac{\operatorname{cond}(A)}{1-\operatorname{cond}(A)\frac{\|\delta A\|}{\|A\|}}\left[\frac{\|Sb\|}{\|b\|} + \frac{\|\delta A\|}{A}\right]\|x\|\end{aligned}$$

□

Ist $\operatorname{cond}(A)\|\delta A\| \ll \|A_i\|$, so gilt

$$\frac{\|\delta x\|}{\|x\|} \dot{\leq} \operatorname{cond}(A)\left[\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{A}\right]$$

Die Konditionszahl hängt von der verwendeten Norm ab.

Beispiel 4.8 1. $\operatorname{cond}_\infty(A) = \|A\|_\infty\left\|A^{-1}\right\|_\infty$

2. Für die Spektralnorm gilt:

$$\operatorname{cond}_2(A) = \|A\|_2\left\|A^{-1}\right\| = \sqrt{\frac{|\mu_{max}|}{|\mu_{min}|}}$$

wobei μ_{max}, μ_{min} betragsgrößter beziehungsweise kleinster Eigenvektor von $\bar{A}^T A$. Ist $A = A =$. Ist $A = \bar{A}^T$ so gilt:

$$\operatorname{cond}_2(A) = \frac{|\lambda_{max}|}{|\lambda_{min}|}$$

mit λ_{max} und λ_{min} betragsgrößter beziehungsweise kleinster Eigenvektor von A. Regel: Es gelte $\operatorname{cond}(A) \approx 10^s$

$$\frac{\|\delta A\|}{\|A\|} \approx 10^{-k}, \frac{\|\delta b\|}{\|b\|} \approx 10^{-k}$$

Dann muss ein relativer Fehler von

$$\frac{\|\delta x\|}{\|x\|} \approx 10^{s-k}$$

erwartet werden. Mit $\|\cdot\| = \|\cdot\|_\infty$ verliert man s Stellen Genauigkeit.

Beispiel 4.9

$$A = \begin{pmatrix}1 & 1\\ 0 & \varepsilon\end{pmatrix}, \varepsilon \in (0,1], A^{-1} = \begin{pmatrix}1 & -\varepsilon^{-1}\\ 0 & \varepsilon^{-1}\end{pmatrix}$$

$$\implies \|A\|_\infty = 2, \left\|A^{-1}\right\|_\infty = 1+\varepsilon^{-1}$$

$$\implies \operatorname{cond}_\infty\|A\|\left\|A^{-1}\right\| = 2+\varepsilon^{-1}$$

für $\varepsilon = 10^{-8}$ kann man bereits 8 Stellen Genauigkeit verlieren.

Ist die Abschätzung im Störungssatz scharf? Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch positiv definit mit Eigenwerten $\lambda_1 \geq \cdots \geq \lambda_n$. Wähle: $\delta A = 0, b = \omega_1, \delta B = \varepsilon w_k, \varepsilon \neq 0$

$$\begin{aligned} Ax = b &\implies x = \frac{1}{\lambda_1} w_1 \\ A\tilde{x} = b + \delta b &\implies \tilde{x} = \frac{1}{\lambda_1}\omega_1 + \varepsilon \frac{1}{\lambda_k}\omega_k \\ \implies \frac{\|\delta x\|_2}{\|x\|_2} &= |\varepsilon| \frac{\lambda_1}{\lambda_n} \frac{\|\omega_n\|_2}{\|\omega_1\|_2} \\ &= \operatorname{cond}(A) \frac{\|\delta b\|_2}{\|b\|_2} \end{aligned}$$

4.1 Eliminationsverfahren

Direkte Methode zur Lösung von $Ax = b$, $A \in \mathbb{R}^{n\times n}$. Spezialfall: A obere Dreiecksmatrix $a_{ij} = 0, i > j$

$$\begin{pmatrix} a_{11} & \dots & \dots & \dots & a_{1n} \\ 0 & a_{22} & & & \vdots \\ \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ \vdots \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ \vdots \\ \vdots \\ b_n \end{pmatrix}$$

Ist $a_{ii} \neq 0, i = 1, \dots, n$ löst man durch Rückwärtseinsetzen

$$x_j = \begin{cases} \frac{b_n}{a_{nn}} & j = n \\ \frac{1}{a_{jj}} \left(b_j - \sum_{k=j+1}^{n} a_{jk} x_k \right) & j = n-1, \dots, 1 \end{cases}$$

Arithmetische Operationen:

$$\sum_{j=1}^{n} j = \frac{(n+1)n}{2} = \frac{n^2}{2} + \mathcal{O}(n)$$

Eine Operation: eine Division oder eine Multiplikation und eine Addition. Wiederholung: Konditionszahl einer Matrix

$$A \in \mathbb{R}^{n\times n} : \operatorname{cond}(A) = \|A\| \left\|A^{-1}\right\|$$

Störungssatz: $(A + \delta A)(x + \delta x) = b + \delta b$

$$\frac{\|\delta x\|}{x} \leq \frac{\operatorname{cond}(A)}{1 - \operatorname{cond}(A) \frac{\|\delta A\|}{\|A\|}} \left[\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right]$$

Gaußsches Eliminationsverfahren
Umformung von $Ax = b$ auf $Rx = c$ mit R obere Dreiecksmatrix mittels

- Vertauschen von Gleichungen
- Addition von Vielfachen einer Gleichung zu einer anderen

Annahme: A hat Vollrang

0. Setze $A^{(0)} = A, b^{(0)} = b$

$$\left[\begin{array}{ccc|c} a_{11}^{(0)} & \dots & a_{1n}^{(0)} & b_1^{(0)} \\ \vdots & & \vdots & \vdots \\ a_{n1}^{(0)} & \dots & a_{nn}^{(0)} & b_n^{(0)} \end{array}\right]$$

1. Wähle $r \in \{1, \dots, n\}$ mit $a_{r1}^{(0)} \neq 0$ (Pivotelement) und vertausche 1. und r-te Zeile

$$\left[\begin{array}{ccc|c} \tilde{a}_{11}^{(0)} & \dots & \tilde{a}_{1n}^{(0)} & \tilde{b}_1^{(0)} \\ \vdots & & \vdots & \vdots \\ \tilde{a}_{n1}^{(0)} & \dots & \tilde{a}_{nn}^{(0)} & \tilde{b}_n^{(0)} \end{array}\right] := [\tilde{A}^{(0)} \mid \tilde{b}^{(0)}]$$

2. Für $j = 2, \dots, n$ eliminiere $\tilde{a}_{j1}^{(0)}$ durch Subtraktion von $\frac{\tilde{a}_{j1}^{(0)}}{\tilde{a}_{11}^{(0)}} := q_{j1}$ mal der ersten Zeile von den Zeilen $2, \dots, n$:

$$\left[\begin{array}{cccc|c} \tilde{a}_{11}^{(0)} & \tilde{a}_{12}^{(0)} & \dots & \tilde{a}_{1n}^{(0)} & \tilde{b}_1^{(0)} \\ 0 & \tilde{a}_{22}^{(1)} & \dots & \tilde{a}_{2n}^{(1)} & \tilde{b}_2^{(1)} \\ \vdots & \vdots & & \vdots & \\ 0 & \tilde{a}_{n2}^{(1)} & \dots & \tilde{a}_{nn}^{(1)} & \tilde{b}_n^{(1)} \end{array}\right] := [A^{(1)} \mid b^{(1)}]$$

   Fahre fort auf kleinerem System $\implies [A^{(0)} \mid b^{(0)}] \to [A^{(1)} \mid b^{(1)}] \to \dots \to [A^{(n-1)} \mid b^{(n-1)}] =: [R \mid c]$

Wird im k -ten Schritt $[A^{(k-1)} \mid b^{(n-1)}] \to [\tilde{A}^{(k-1)} \mid \tilde{b}^{(n-1)}] \to [A^{(k)} \mid b^{(k)}]$ das Pivot-Element $q_{r_k k}^{k-1}$ gewählt, so gilt $[\tilde{A}^{(k-1)} \mid \tilde{b}^{(k-1)}] = P_k [A^{(k-1)} \mid b^{(k-1)}]$ mit der Permutationsmatrix

$$P_k = \begin{pmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ & & & 0 & \dots & \dots & \dots & 1 & & \\ & & & & 1 & & & & & \\ & & & & & \ddots & & & & \\ & & & & & & 1 & & & \\ & & & 1 & \dots & \dots & \dots & 0 & & \\ & & & & & & & & 1 & \\ & & & & & & & & & \dots \\ & & & & & & & & & & 1 \end{pmatrix} \quad G_k = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & -q_{k+1,k}^{(k)} & 1 & \\ & \vdots & & \ddots \\ & -q_{n,k}^{(k)} & & 1 \end{pmatrix}$$

Mit den Fehlstellungen von P_k an k und r_k und der Fehlspalte von G_k bei k. Weiterhin gilt: $[A^{(k)} \mid b^{(k)}] = G_k [\tilde{A}^{(k-1)} \mid \tilde{b}^{(k-1)}]$ mit $q_{jk}^{(k)} = \tilde{a}_{jk}^{(k-1)} / \tilde{a}_{kk}^{(k-1)}$. G_k heißt Frobenius Matrix. Wegen $P_k^{-1} = P_k$ und

$$G_k^{-1} = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & q_{k+1,k}^{(k)} & 1 & \\ & \vdots & & \ddots \\ & q_{n,k}^{(k)} & & 1 \end{pmatrix}$$

haben $Ax = b$ und $A^{(k)}x = b^{(k)}$ dieselbe Lösung:

$$Ax = b \iff A^{(k)}x = G_{n-1}P_{n-1} \dots G_1 P_1 Ax = G_{n-1}P_{n-1} \dots G_1 P_1 b = b^{(k)}$$

Wahl des Pivot-Elementes
Ziel: Numerische Stabilität.

1. Spaltenpivotierung:
$$\left|a_{r_k,k}^{(k-1)}\right| = \max_{j=k,\dots,n}\left|a_{jk}^{(k-1)}\right|$$
2. Totalpivotierung
$$\left|a_{r_k,s_k}^{(k-1)}\right| = \max_{i,j=k,\dots,n}\left|a_{ij}^{(k-1)}\right|$$
   - bessere Stabilität
   - teurer
   - Permutationsmatrizen Q_k für x

$$\underbrace{G_kP_k\dots G_1P_1AQ_1\dots Q_k}_{A^{(k)}}\underbrace{Q_k\dots Q_1x}_{Qx} = G_kP_k\dots G_1P_1b$$

Speicherausnutzung
Die $q_{jk}^{(k)}$ können an den eliminierten Stellen im unteren Dreieck von A gespeichert werden. Das obere Dreieck von A wird während der Rechnung ersetzt. Nach k Eliminationsschritten

$$\left[\begin{array}{cccccccc}
r_{11} & r_{12} & \dots & r_{1k} & r_{1,k+1} & \dots & r_{1n} & c_1 \\
\lambda_{21} & r_{22} & \dots & r_{2k} & r_{2,k+1} & \dots & r_{2n} & c_2 \\
\lambda_{31} & \lambda_{32} & \ddots & & \vdots & & \vdots & \vdots \\
\vdots & \vdots & \ddots & & \vdots & & \vdots & \vdots \\
\lambda_{k1} & \dots & \lambda_{kk} & r_{kk} & r_{k,k+1} & \dots & r_{kn} & c_k \\
\lambda_{k1} & \dots & \dots & \lambda_{k+1,k} & a_{k+1,k+1}^{(k)} & \dots & a_{k+1,n}^{(k)} & b_{k+1}^{(k)} \\
\vdots & & & \vdots & \vdots & & \vdots & \vdots \\
\lambda_{n1} & \dots & \dots & \lambda_{n,k} & a_{n,k+1}^{(k)} & \dots & a_{n,n}^{(k)} & b_n^{(k)}
\end{array}\right]$$

mit $\lambda_{i+1,1},\dots,\lambda_{ni}$ Permutationen von $q_{i+1,i}^{(k)},\dots,q_{ni}^{(k)}$. Endresultat ($k = n-1$)

$$\left[\begin{array}{cccc|c}
r_{11} & \dots & \dots & r_{1n} & c_1 \\
l_{21} & r_{22} & \dots & r_{2n} & \vdots \\
\vdots & \ddots & \ddots & \vdots & \vdots \\
l_{n1} & \dots & l_{n,n-1} & r_{nn} & c_n
\end{array}\right]$$

Satz 4.10 (LR-Zerlegung) Die Matrizen

$$L = \begin{bmatrix} 1 & & & \\ l_{21} & 1 & & \\ \vdots & \ddots & \ddots & \\ l_{n1} & \dots & l_{n,n-1} & 1 \end{bmatrix}, R = \begin{bmatrix} r_{11} & \dots & r_{1n} \\ & \ddots & \vdots \\ & & r_{nn} \end{bmatrix}$$

bilden eine LR-Zerlegung der Matrix PA. $PA = LR$, mit $P = P_{n-1}\dots P_1$. Für $P = \not\Vdash$ ist die Zerlegung eindeutig.

Beweis (für $P = \not\Vdash$).
$$R = G_{n-1}\dots G_1A \iff \underbrace{G_1^{-1}\dots G_{n-1}^{-1}}_{L}R = A$$

Eindeutigkeit: Übung. □

Aufwand: k -ter Eliminationsschritt

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} a_{kj}^{(k-1)}$$
$$b_{i}^{(k)} = b_{i}^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} b_{k}^{(k-1)}$$

$i, j = k+1, \ldots, n \implies n-k$ Divisionen, $(n-k) + (n-k)^2$ Multiplikationen und Additionen

$$\implies N_{\text{Gauß}}(n) = \frac{1}{3}n^3 + \backslash^{\in}$$

Gilt für Lösung von $Ax = b$ und für die Berechnung der Zerlegung $PA = LR$

Beispiel 4.11

$$A = \begin{pmatrix} 3 & 1 & 6 \\ 2 & 1 & 3 \\ 1 & 1 & 1 \end{pmatrix}, b = \begin{pmatrix} 2 \\ 7 \\ 4 \end{pmatrix}$$

Pivotierung:

$$\left[\begin{array}{ccc|c} 3 & 1 & 6 & 2 \\ 2 & 1 & 3 & 7 \\ 1 & 1 & 1 & 4 \end{array}\right] \to \left[\begin{array}{ccc|c} 3 & 1 & 6 & 2 \\ 2/3 & 1/3 & -1 & 17/3 \\ 1/3 & 2/3 & -1 & 10/3 \end{array}\right] \to \left[\begin{array}{ccc|c} 3 & 1 & 6 & 2 \\ 1/3 & 2/3 & -1 & 10/3 \\ 2/3 & 1/3 & -1 & 17/3 \end{array}\right] \to \left[\begin{array}{ccc|c} 3 & 1 & 6 & 2 \\ 1/3 & 2/3 & -1 & 10/3 \\ 2/3 & 1/2 & -1/2 & 4 \end{array}\right]$$

$\to$

$$x_3 = -8$$
$$x_2 = \frac{3}{2}\left(\frac{10}{3} + x_3\right) = -7$$
$$x_1 = \frac{1}{3}(2 - x_2 - 6x_3) = 19$$

LR-Zerlegung:

$$P_1 = E_3, P_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

$$PA = \begin{pmatrix} 3 & 1 & 6 \\ 1 & 1 & 1 \\ 2 & 1 & 3 \end{pmatrix} = LR = \begin{pmatrix} 1 & 0 & 0 \\ 1/3 & 1 & 0 \\ 2/3 & 1/2 & 1 \end{pmatrix} \begin{pmatrix} 3 & 1 & 6 \\ 0 & 2/3 & -1 \\ 0 & 0 & -1/2 \end{pmatrix}$$

Für die numerische Stabilität der Gauß-Elimination ist im Allgemeinen eine Pivotierung sehr wichtig.

Rückwärtsanalyse nach Wilkinson $A \in \mathbb{R}^{n \times n}$, löse $Ax = b$ mit Gauß-Elimination mit Spaltenpivotierung. Die berechnete Lösung $\tilde{x}$ ist die exakte Lösung eines gestörten Systems $(A + \delta A)\tilde{x} = b$ mit

$$\frac{\|\delta A\|_\infty}{\|A\|_\infty} \leq 1.01 \cdot 2^{n-1}\left(n^3 + 2n^2\right) eps$$

(ohne Beweis)
Störungssatz $\implies$

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\text{cond}_\infty(A)}{1 - \text{cond}_\infty(A)\|\delta A\|/\|A\|} \cdot 1.012^{n-1}\left(n^3 + 2n^2\right) eps$$

Diese Abschätzung deckt pathologische Fälle ab. In der Praxis ist das Verhalten gutartig, das heißt die Gaußelimination mit Spaltenpivotierung ist ein stabiler Algorithmus. Wiederholung: Gauß-Elimination $Ax = b, A \in \mathbb{R}^{n\times n}$

$$A^{(0)} = A, A^{(k)} = G_k P_k A^{(k-1)}, k = 1, \dots, n-1$$
$$R = A^{(n-1)} a = G_{n-1} P_{n-1} \dots G_1 P_1 A$$

setze

$$\tilde{G}_{n-1} = P_{n-1}, \tilde{G}_k = P_{n-1} \dots P_{k-1} G_k P_{k+1} \dots P_{n-1}$$
$$\implies P_{k+1} \dots P_{n-1} \tilde{G}_k = G_k P_{k+1} \dots p_{n-1}$$
$$\implies R = \underbrace{\tilde{G}_{n+1} \dots \tilde{G}_1}_{L^{-1}} \underbrace{P_{n-1} \dots P_1}_{P} A$$
$$\implies LR = PA$$

Löse $Rx = c$ oder $Ly = b, Rx = y$.

4.2 Nachiteration

Wegen Rundungsfehlern bei der Gauß-Elimination gilt: $PA = LR$ nicht exakt. Damit gilt mit einer Näherungslösung x^0 gewonnen aus $LRx^0 = Pb$ für den sogenannten Defekt

$$d^0 = b - Ax^0 \neq 0$$

Man kann man eine iterative Defektkorrektur betreiben.

$$d^k = b - Ax^k, LR\delta x^k = Pd^k$$
$$x^{k+1} = x^k + \delta x^k, k = 0, 1, \dots$$

Lemma 4.12

$$x^k = \left(\sum_{k=0}^{k} \left(\mathbb{1} - R^{-1} L^{-1} PA \right)^n \right) R^{-1} L^{-1} Pb$$

Beweis per Induktion über k:
$k = 0$ ✓
$k \geq 0$:

$$\delta x^k = R^{-1} L^{-1} \left(b - Ax^k \right)$$
$$x^{k+1} = x^k + \delta x^k = \left(\mathbb{1} - R^{-1} L^{-1} PA \right) x^k + R^{-1} L^{-1} Pb$$
$$= \left(\sum_{n=0}^{k} \left(\mathbb{1} - R^{-1} L^{-1} PA \right)^{n+1} \right) R^{-1} L^{-1} Pb + R^{-1} L^{-1} Pb$$
$$= \left(\sum_{n=0}^{k+1} \left(\mathbb{1} - R^{-1} L^{-1} PA \right)^{n} \right) R^{-1} L^{-1} Pb \qquad \square$$

Ist $\left\| \mathbb{1} - R^{-1} L^{-1} PA \right\| < 1$, so gilt

$$\sum_{n=0}^{\infty} \left(\mathbb{1} - R^{-1} L^{-1} PA \right)^n = \left(\mathbb{1} - \mathbb{1} + R^{-1} L^{-1} PA \right)^{-1}$$
$$= (PA)^{-1} LR$$

$\implies x^k$ konvergiert gegen

$$x^* = (PA)^{-1}LRR^{-1}L^{-1}Pb = A^{-1}b$$

Wichtig: In der Praxis muss der Defekt d^k mit höherer Genauigkeit berechnet werden.

Beispiel 4.13 Skript.

4.3 Determinantenbestimmung

$$A, B \in R^{n\times n} \implies \det(A \cdot B) = \det A \det B$$

$A = P^T LR$

$$\det A = \underbrace{\det\left(P^T\right)}_{\pm 1} \underbrace{L}_{1} \underbrace{\det R}_{\prod_{i=1}^{n} r_{ii}} = \pm \prod_{i=1}^{n} r'_{ii}$$

Bei k Vertauschungen von Zeilen ist das Vorzeichen $(-1)^k$.

4.4 Rangbestimmung

$\rightarrow$ Totalpivotierung $PAQ^T = LR$ Gilt nach dem i -ten Eliminationsschritt

$$a_{k,j}^{(i)} = 0 \forall j, k = i+1, \ldots, n$$

so ist $\mathrm{Rang}(A) = i$ (Geht auch bei nicht quadratischen Matrizen, einfach mit Nullen auffüllen)

4.5 Spezielle Gleichungssysteme

4.5.1 Bandmatrizen

Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt Bandmatrix vom Bandtyp $(m_l, m_r) \in \{0, \ldots, n-1\}^2$, wenn gilt

$$a_{jk} = 0 \forall k < j - m_l \vee k > j + m_r, j, k = 1, \ldots, n$$

Die Größe $m = m_l + m_r$ heißt Bandbreite.

Typ	Name
$(0,0)$	Diagonalmatrix
$(1,1)$	Tridiagonalmatrizen
$(n-1,0)$	Untere Dreiecksmatrix
$(0,n-1)$	Obere Dreiecksmatrix

Satz 4.14 (Bandmatrix) Ist $A \in \mathbb{R}^{n\times n}$ eine Bandmatrix vom Typ (m_l, m_r), Für die Gauß-Elimination $A = LR$ ohne Zeilenvertauschung durchführbar ist, dann sind alle reduzierten Matrizen $A^{(i)}$ desselben Typs und L beziehungsweise R sind vom Typ $(m_l, 0)$ beziehungsweise $(0, m_r)$. Aufwand:

$$N = \frac{1}{3} n m_l m_r + \mathcal{O}(n(m_l + m_r))$$

(Ohne Beweis)

Beispiel 4.15 Typ $(1,1)$:

$$A = \begin{pmatrix} a_1 & b_1 & & \\ c_2 & \ddots & \ddots & \\ & \ddots & \ddots & b_{n-1} \\ & & c_n & a_n \end{pmatrix} = LR, L = \begin{pmatrix} 1 & & & \\ \gamma_1 & \ddots & & \\ & \ddots & \ddots & \\ & & \gamma_n & 1 \end{pmatrix}, R = \begin{pmatrix} \alpha_1 & \beta_1 & & \\ & \ddots & \ddots & \\ & & \ddots & \beta_{n-1} \\ & & & \alpha_n \end{pmatrix}$$

Rekursive Bestimmung

$$\begin{aligned} \alpha_1 = a_1, \beta_1 = b_1 \\ \gamma_i = c_i/\alpha_{i-1}, \alpha_i = a_i - \gamma_i\beta_{i-1}, \beta_i = b_i \\ \gamma_n = c_n/\alpha_{n-1}, \alpha_n = a_n - \gamma_n\beta_{n-1} \end{aligned}$$

Aufwand: $3n-2$ Speicher,$2n-2$ arithmetische Operationen.
Vorsicht: (Beispiel Typ $(4,4)$): Band „füllt auf“

4.5.2 Diagonaldominante Matrizen

Definition 4.16 $A \in \mathbb{R}^{n\times n}$ heißt diagonaldominant, wenn

$$\sum_{\substack{k=1 \\ k\neq j}}^{n} |a_{jk}| \leq |a_{jj}|, \quad j = 1, \ldots, n$$

Satz 4.17 $A \in \mathbb{R}^{n\times n}$ regulär und diagonaldominant $\implies$ $A = LR$ kann mit Gauß-Elimination ohne Zeilenvertauschungen berechnet werden. (Beweis: Skript)

4.5.3 Positiv definite Matrizen

Definition 4.18 $A \in R^{n\times n}$ mit $A^T = A$ heißt positiv definit, wann

$$x^T Ax > 0 \forall x \in \mathbb{R}^n \setminus \{0\}$$

Satz 4.19 $A \in \mathbb{R}^{n\times n}$ symmetrisch positiv definit $\implies$ $A = LR$ kann mit Gauß-Elimination ohne Zeilenvertauschung berechnet werden mit Pivots $a_{ii}^{(i)} > 0$

Beweis

$$0 < e_1^T Ae_1 = a_{11}$$

$$\begin{aligned} a_{jk}^{(1)} = a_{jk} - \frac{a_{j1}}{a_{11}}a_{1k} = a_{kj} - \frac{a_{k1}}{a_{11}}a_{1j} = a_{kj}^{(1)} \\ \implies A^{(1)} = \left(a_{jk}^{(1)}\right)_{j,k=2}^{n} \text{ ist symmetrisch} \end{aligned}$$

Ist $A^{(1)}$ positiv definit, so beweist Induktion die Behauptung. Setze dafür $\tilde{x} = (x_2, \ldots, x_n)^T \in \mathbb{R}^{n-1}, x \in \mathbb{R}^n$, sodass

$$x_1 = -\frac{1}{a_{11}} \sum_{k=2}^{n} a_{1k}x_k$$

$$\begin{aligned}
\implies 0 < x^T Ax &= \sum_{j,k=1}^{n} a_{jk} x_j x_k \\
&= \sum_{j,k=2}^{n} a_{jk} x_j x^k + 2x_1 \sum_{k=2}^{n} a_{1k} x_k + a_{11} x_1^2 - \underbrace{\frac{1}{a_{11}} \sum_{j,k=2}^{n} a_{k1} a_{j1} x_k x_j + \frac{1}{a_{11}} \left(\sum_{k=2}^{n} a_{1k} x_k \right)^2}_{0} \\
&= \underbrace{\sum_{j,k=2}^{n} \left(a_{jk} - \frac{a_{k1} a_{j1}}{a_{11}} \right) x_k x_j}_{=a_{jk}^{(1)}} + a_{11} \underbrace{\left(x_1 + \sum_{1}^{a_{11}} \sum_{k=2}^{n} a_{1k} x_k \right)}_{0}^{2} \\
&= \tilde{x}^T A^{(1)} \tilde{x}. \qquad \square
\end{aligned}$$

$\to A = LR, r_{ii} = a_{ii}^{(i)} > 0$

$$A = A^T = (LR)^T = \left(LD \underbrace{D^{-1} R}_{=:R} \right)^T = \tilde{R}^T D L^T$$

mit $A = \operatorname{diag}(r_1, \ldots, r_{nn})$ und

$$\tilde{R} = \begin{pmatrix} 1 & r_{12}/r_{11} & \ldots & r_{1n}/r_{11} \\ \ddots & & & \vdots \\ & \ddots & & r_{n-1,n}/r_{n-1,n-1} \\ & & & 1 \end{pmatrix}$$

Eindeutigkeit der LR-Zerlegung

$$LR = \tilde{R}^T D L^T \implies L = \tilde{R}^T, R = D L^T$$

Satz 4.20 Jede symmetrisch positiv definite Matrix $A \in \mathbb{R}^{n \times n}$ hat eine sogenannte Cholesky-Zerlegung

$$A = LDL^T = \tilde{L}\tilde{L}^T$$

Aufwand: $N_{\text{Cholesky}}(n) = \frac{n^3}{6} + \mathcal{O}(n^2)$.
Algorithmus von Cholesky:

$$\begin{pmatrix} \tilde{l}_{11} & & \\ \vdots & \ddots & \\ \tilde{l}_{n1} & \ldots & \tilde{l}_{nn} \end{pmatrix} \begin{pmatrix} \tilde{l}_{11} & \ldots & \tilde{l}_{n1} \\ \ddots & & \vdots \\ & & \tilde{l}_{nn} \end{pmatrix} = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \ldots & a_{nn} \end{pmatrix}$$

$$i \geq j : a_{ij} = \sum_{k=1}^{j} \tilde{l}_{ik} \tilde{l}_{jk} = \sum_{k=1}^{j-1} \tilde{l}_{ik} \tilde{l}_{jk} + \tilde{l}_{ij} \tilde{l}_{jj}$$

für $i = 1, \ldots, n$:

$$\tilde{l}_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} \tilde{l}_{ik}^2}$$

Für $j = i+1, \ldots, n$:

$$\tilde{l}_{ij} = \frac{1}{\tilde{l}_{ii}} \left(a_{ij} - \sum_{k=1}^{i-1} \tilde{l}_{ik} \tilde{l}_{jk} \right)$$

Wiederholung: Spezielle Matrizen, LR-Zerlegung

- Bandmatrizen: Nullen nicht speichern / berechnen
- Diagonal-dominante Matrizen: keine Pivotierung notwendig
- Symmetrisch, positiv definite Matrizen keine Pivotierung notwendig

  $$A = \tilde{L}\tilde{L}^T = LDL^T$$

  (billiger als $A = LR$)

4.6 Nicht reguläre Systeme

Wir betrachten $A \in \mathbb{R}^{m \times n}$ (nicht notwendig quadratisch). Das Lineare Gleichungssystem $Ax = b$ hat

- keine Lösung, wenn $b \notin \mathrm{im}(A)$
- unendlich viele Lösungen $\bar{x} + \Delta x$ wenn $A\bar{x} = b, \Delta x \in \ker(A) \neq \{0\}$

Verallgemeinerter Lösungsbegriff: Finde $\bar{x} \in \mathbb{R}^n$ mit minimalem Defekt $d = b - A\bar{x}$ (Für $d = 0$ löst $\bar{x} \quad Ax = b$)

Satz 4.21 (Least-Squares-Lösung) Es gibt immer eine „Lösung“ $\bar{x} \in \mathbb{R}^n$ mit kleinsten Fehlerquadraten:

$$\|A\bar{x} - b\|_2 = \min_{x \in \mathbb{R}^n} \|Ax - b\|_2$$

Das gilt genau dann, wenn

$$A^T A\bar{x} = A^T b$$

(Normalgleichung). Für $\mathrm{Rang}(A) = n$ ist $\bar{x}$ eindeutig bestimmt. Ansonsten hat jede weitere Lösung die Form $\bar{x} + y$ mit $y \in \ker(A)$

Lemma 4.22 Sei $A \in \mathbb{K}^{m \times n}$. Dann ist $\bar{A}^T A$ hermitesch positiv semidefinit. Ist $\mathrm{Rang}(A) = n$, so ist $\bar{A}^T A$ positiv definit.

Beweis

1. $\overline{\bar{A}^T A}^T = \left(A^T \bar{A}\right)^T = \bar{A}^T A$
2. $\bar{x}^T \bar{A}^T A x = \overline{(Ax)}^T (Ax) = \|Ax\|_2^2 \geq 0$
3. $\mathrm{Rang}(A) = n \implies m \geq n$ und $A : \mathbb{R}^n \to \mathbb{R}^m$ injektiv $\implies$ Aus $\|A\|_2 = 0 \implies Ax = 0 \implies x = 0$
   $\implies \bar{x}^T \bar{A}^T A x > 0 \forall x \in \mathbb{K}^n \setminus \{0\}$ □

Beweis

1. Es gelte $A^T A\bar{x} = A^T b$

$$\begin{aligned}
&\implies A^T(A\bar{x} - b) = 0 \\
&\implies \|b - Ax\|_2^2 = \|b - A\bar{x} + A(\bar{x} - x)\|_2^2 \\
&\qquad = \|b - A\bar{x}\|_2^2 + 2(b - A\bar{x}, A(\bar{x} - x))_2 + \|A(\bar{x} - x)\|_2^2 \\
&(A(\bar{x} - b), b - A\bar{x})_2 = (\bar{x} - x)^T A^T (b - A\bar{x}) 0 \\
\implies &\|b - Ax\|_2^2 > \|b - A\bar{x}\|_2^2 + \|A(\bar{x} - x)\|_2^2
\end{aligned}$$

$\Longrightarrow$ $\bar{x}$ ist minimal. Umgekehrt: Sei $\bar{x}$ minimal

$$\begin{aligned} 0 &= \frac{\partial}{\partial x_i}\|Ax-b\|_2^2\big|_{x=\bar{x}} \\ &= \frac{\partial}{\partial x_i}\left(\sum_{j=1}^{m}\left(\sum_{k=1}^{n}a_{jk}x_k - b_j\right)^2\right)\Big|_{x=\bar{x}} \\ &= \sum_{j=1}^{m}2\left(\sum_{k=1}^{n}a_{jk}\bar{x}_k - b_j\right) \\ a_{ji} &= \left(2A^T(A\bar{x}-b)\right)_i \\ \Longrightarrow\ A^TA\bar{x} &= A^Tb \end{aligned}$$

2. Lösbarkeit: Wegen $\mathrm{im}(A)^\perp = \ker\left(A^T\right)$ hat b eine eindeutige Zerlegung $b = r+s, r \in \ker\left(A^T\right), s \in \mathrm{im}(A)$ Sei $\bar{x} \in \mathbb{R}^n$ so, dass $A\bar{x} = s \implies A^TA\bar{x} = A^Ts + A^Tr = A^Tb$
3. $\mathrm{Rang}(A) = n$: A^TA positiv definit $\implies$ $\bar{x}$ eindeutig. $Rang(A) < n : A^TAx_1 = A^Tb$. Wegen

$$b = Ax_1 + (b - A_{x_1}) \in \mathrm{im}\, A + \ker A^T$$

und Eindeutigkeit von $b = r+s$ gilt $A\bar{x} = Ax_1 \forall \bar{x} - x_1 \in \ker A$ □

Numerische Lösung: Cholesky für Normalgleichung. Vorsicht: Im Fall $\mathrm{Rang}(A) = n = m$ gilt

$$\mathrm{cond}_2\left(A^TA\right) = \mathrm{cond}_2(A)^2$$

Merke: Normalgleichungen sind häufig schlecht konditioniert. Abhilfe: QR-Zerlegung von A

Satz 4.23 Sei $A \in \mathbb{K}^{m\times n}$ mit $\mathrm{Rang}(A) = n \le m$. Dann existiert eine eindeutig bestimmte Matrix $Q \in \mathbb{K}^{m\times n}$ mit $\bar{Q}^TQ = E_n$ und eine eindeutig bestimmte obere Dreiecksmatrix $R \in \mathbb{K}^{n\times n}$ mit reellen Diagonalelementen $r_{ii} > 0, i = 1,\dots,n$, sodass

$$A = QR$$

Bezeichnung: Q: orthonormale Matrix ($m = n$: unitär)

Beweis Konstruktion der Spalten q_k von Q mittels Gram-Schmidt aus den Spalten a_k von A

$$q_i = \begin{cases} q_i = \|a_1\|_2^{-1}a_1 & i = 1 \\ q_i = \|\tilde{q}_i\|_2^{-1}\tilde{q}_i, \tilde{q}_i = a_i - \sum_{k=1}^{i-1}(a_i,q_k)_2q_k & i = 2,\dots,n \end{cases}$$

Wegen $\mathrm{Rang}(A) = n$ sind die a_k linear unabhängig und $\|\tilde{q}_k\|_2 \neq 0, k = 1,\dots,n$. Betrachte:

$$\begin{aligned} a_k &= \tilde{q}_k + \sum_{i=1}^{k-1}(a_k,q_i)_2q_i \\ &= \|\tilde{q}_k\|_2q_k + \sum_{i=1}^{k-1}(a_k,q_i)q_i \\ &= \sum_{i=1}^{k}r_{ik}q_i \end{aligned}$$

$r_{kk} = \|\tilde{q}_k\|_2 \in \mathbb{R}_+, r_{ik} = (a_k, q_i)_2$. Setze $r_{ik} = 0, i > k, R = (r_{ik}) \in \mathbb{K}^{n\times n} \implies A = QR$.
Eindeutigkeit: Sei $Q_1R_1 = A = Q_2R_2$. Setze

$$\begin{aligned} Q = \bar{Q}_2^T Q_1 = \bar{Q}_2^T A R_1^{-1} = R_2 R_1^{-1} && \text{(obere Dreiecksmatrix)} \\ \bar{Q}^T = \bar{Q}_1^T Q_2 = \bar{Q}_1^T A R_2^{-1} = R_1 R_2^{-1} && \text{(obere Dreiecksmatrix)} \\ \bar{Q}^T Q = R_1 R_2^{-1} R_2 R_1^{-1} = \mathbb{I} \end{aligned}$$

Q ist orthonormal und diagonal. Ihre Eigenwerte λ_i erfüllen $|\lambda_i| = 1$

$$QR_1 = R_2 R_1^{-1} R_1 = R_2 \implies \lambda_i \underbrace{(R_1)_{ii}}_{>0} = (R_2)_{ii} > 0$$

$$\implies \lambda_i \in \mathbb{R}, \lambda_i = 1$$

$$Q = E_n \implies R_1 = R_2, Q_1 = AR_1^{-1} = AR_2^{-1} = Q_2 \qquad \square$$

Least-Squares-Lösung mit

$$A = Q_1 R = (Q_1 \mid Q_2) \begin{pmatrix} R \\ 0 \end{pmatrix}$$

mit $Q = (Q_1 \mid Q_2) \in \mathbb{R}^{m\times n}, R = \begin{pmatrix} R \\ 0 \end{pmatrix} \in \mathbb{R}^{m\times n}, \operatorname{Rang}(A) = n$

- $\|Qv\|_2^2 = v^T Q^T Q v = \|v\|_2^2$
- $\|Ax - b\|_2^2 = \left\|Q\tilde{R}x - QQ^T b\right\|_2^2 = \left\|Q\left(\tilde{R}x - Q^T b\right)\right\|_2^2$

  $$= \left\|\tilde{R}x - Q^T b\right\|_2^2 = \left\|\vec{R;}0x - \begin{pmatrix} Q_1^T \\ Q_2^T \end{pmatrix} b\right\|_2^2$$
  $$= \left\|Rx - Q_1^T b\right\|_2^2 + \left\|Q_2^T b\right\|_2^2$$

  minimal für $x = R^{-1}Q_1^T b$.
- $A^T A = (Q_1 R)^T Q_1 R = R^T Q_1^T Q_1 R = R^T R$ (Cholesky-Zerlegung)
  $A^T Ax = A^T b = R^T Rx = R^T Q_1^T b$
- Lösung mit R ist besser konditioniert als Lösung mit $R^T R$: $\operatorname{cond}_2\left(R^T R\right) = \operatorname{cond}_2(R)^2$

Wiederholung: $A \in M(n \times m, \mathbb{K})$

- $A = QR = \tilde{Q}\tilde{R}, \tilde{Q} = Q|\tilde{Q}_2, \bar{Q}^T Q = E_n, \bar{\tilde{Q}}^T \tilde{Q} = E_m$
- Eindeutigkeit mit $r_{ij} > 0$
- $\|Ax - b\|_2^2 = \left\|\tilde{Q}\tilde{R}x - \tilde{Q}\bar{\tilde{Q}}^T b\right\|_2^2 = \left\|\tilde{Q}\left(\tilde{R}x - \tilde{Q}^T b\right)\right\|_2^2 = \left\|Rx - Q^T b\right\|_2^2 + \left\|\tilde{Q}_2 b\right\|$. $\operatorname{Rang}(A) = n$: $x = \tilde{R}^{-1}Q^T b$
- verhindert schlechte Konditionierung der Normalgleichung

$$\operatorname{cond}_2\left(A^T A\right) = \operatorname{cond}_2(A)^2 = \operatorname{cond}_2(R)^2$$

Problem: Gram-Schmidt zur Berechnung von Orthogonalbasis ist nicht stabil. Stabile Variante: Householder-Verfahren

Definition 4.24 Für $v \in \mathbb{K}$ mit $\|v\|_2 = 1$ heißt

$$I = E_n - 2v\bar{v}^T \in \mathbb{K}^{n\times n}$$

„Householder-Transformation“.

$$v\bar{v}^T = \begin{pmatrix} v_1\bar{v}_1 & \dots & v_1\bar{v}_n \\ \vdots & \ddots & \vdots \\ v_n\bar{v}_1 & \dots & v_n\bar{v}_n \end{pmatrix}$$

Eigenschaften von S:

- $\bar{S}^T = S$ hermitesch
- $\bar{S}^T S = \big(E_n - 2v\bar{v}^T\big)\big(E_n - 2v\bar{v}^T\big) = E_n - 4v\bar{v}^T + 4v\underbrace{\bar{v}^T v}_{1}\bar{v}^T = E_n$ (unitär)
- Spiegelung: Sei $u \in \mathbb{K}^n$. Zerlege $u = (v,u)_2 v + [u - (v,u)_2 v] = u_1 + u_2$

$$\begin{aligned} Su_1 &= \big(E_n - 2v\bar{v}^T\big)(v,u)_2 v \\ &= (v,u)_2\big(v - 2v\bar{v}^T v\big) = -u_1 \\ SU_2 &= \big(E_n - 2v\bar{v}^T\big)(u - u_1) \\ &= u - 2(v,u)_2 v + u_1 = u - u_1 = u_2 \end{aligned}$$

  v: Normale der Spiegelungshyperebene

Householder-Verfahren:

$$A = A^{(0)} \to \dots \to A^{(i-1)} \to \dots \to A^{(n)} = \tilde{R}$$

mit

$$A^{(i)} = \begin{pmatrix} a_{11}^{(i)} & \dots & \dots & \dots & \dots & a_{1n}^{(i)} \\ & \ddots & & & & \vdots \\ & & a_{ii}^{(i)} & \dots & \dots & a_{in}^{(i)} \\ & 0 & \vdots & & & \vdots \\ & & a_{im}^{(i)} & & & a_{nm}^{(i)} \end{pmatrix}$$

Schritt i: Householder-Transformation

$$S_i A^{(i-1)} = A^{(i)}$$

$$\begin{aligned} &\implies \tilde{R} = A^{(n)} = S_n S_{n-1} \dots S_1 A = \bar{\tilde{Q}}^T A \\ &\implies \tilde{Q}\tilde{R} = A, \tilde{Q} = \bar{S}_1^T \dots \bar{S}_n^T = S_1 \dots S_n \end{aligned}$$

Achtung: $\tilde{r}_{ii} > 0$ wird nicht garantiert. (keine Eindeutigkeit). Bezeichnung: $\tilde{A}^{(i)} = \left(\tilde{a}_i^{(i)} \mid \dots \mid \tilde{a}_n^{(i)}\right)$ Setze

$$v_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \tilde{v}_i \end{pmatrix}, \tilde{v}_i \in \mathbb{K}^{m-i}$$

$$\implies S_i = E_m - 2v_i\bar{v}_i^T = \begin{pmatrix} E_{i-1} & 0 \\ 0 & E_{m-i} - 2\tilde{v}_i\bar{\tilde{v}}_i^T \end{pmatrix} = \begin{pmatrix} \nVdash & 0 \\ 0 & \tilde{S}_i \end{pmatrix}$$

$\Longrightarrow$ Die ersten $i-1$ Zeilen von $A^{(i-1)}$ bleiben unverändert. Wähle $\tilde{v}_i$ so, dass

$$\tilde{S}_i \tilde{s}_i^{(i)} \in \mathrm{Lin}\{e_1^i\}, e_1^i = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{m-i}$$

2 Möglichkeiten:

$$\tilde{v}_i = \frac{\tilde{a}_i^{(i)} - \left\|\tilde{a}_i^{(i)}\right\|_2 e_1}{\left\|\tilde{a}_i^{(i)} - \left\|\tilde{a}_i^{(i)}\right\|_2 e_1\right\|_2}$$

$$\tilde{v}_i = \frac{\tilde{a}_i^{(i)} + \left\|\tilde{a}_i^{(i)}\right\|_2 e_1}{\left\|\tilde{a}_i^{(i)} + \left\|\tilde{a}_i^{(i)}\right\|_2 e_1\right\|_2}$$

Zur Vermeidung von Auslöschung:

$$\tilde{v}_i = \frac{\tilde{a}_i^{(i)} + \mathrm{sgn}\left(\tilde{a}_{ii}^{(i)}\right)\left\|\tilde{a}_i^{(i)}\right\|_2 e_1}{\left\|\tilde{a}_i^{(i)} + \mathrm{sgn}\left(\tilde{a}_{ii}^{(i)}\right)\left\|\tilde{a}_i^{(i)}\right\|_2 e_1\right\|_2}$$

$$\Longrightarrow \tilde{S}_i \tilde{A}^{(i)} = \left.\begin{matrix} \pm\left\|\tilde{a}_i^{(i)}\right\|_2 \\ 0 \\ \vdots \\ 0 \end{matrix}\right| \tilde{a}_i^{(i)} - 2\left(\tilde{a}_{i+j,i}^{(i)}, \tilde{v}_i\right)\tilde{v}_i \quad , j = 2, \ldots, m-i$$

Insgesamt ergibt sich für die Spalten von $A^{(i)} = S_i A^{(i-1)}$

$$\begin{aligned} a_k^{(i)} &= a_k^{(i-1)}, k = 1, \ldots, i-1 \\ a_i^{(i)} &= \left(a_{i,1}^{(i-1)}, \ldots, a_{i-1,i}^{(i-1)}, \left\|\tilde{a}_i^{(i-1)}\right\|_2, 0, \ldots, 2\right)^T \\ a_k^{(i)} &= a_k^{(i-1)} - 2\left(\tilde{a}_k^{(i-1)}, \tilde{v}_i\right)v_i, k = i+1, \ldots, n \end{aligned}$$

4.7 Singulärwertzerlegung

Satz 4.25 Es sei $A \in \mathbb{R}^{m \times n}$. Dann existieren orthogonale Matrizen $V \in \mathbb{R}^{n \times n}$ und $U \in \mathbb{R}^{m \times m}$, sodass

$$U^T A V = \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m \times n}, p = \min(m, n)$$

mit $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$

Beweis Skript, □

$A = U\Sigma V^T$. Nützliche Folgerungen: ($\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0$)

- $\mathrm{Rang}(A) = r$
- $\ker A = \mathrm{Lin}\{v_{r+1}, \ldots, v_n\}$

- $\operatorname{im} A = \operatorname{Lin}\{u_1, \ldots, u_r\}$
- $A = U\Sigma V^T = \sum_{i=1}^{r} \sigma_1 u_i v_i^T$
- $\|A\|_2 = \sigma_1$
- $\operatorname{cond}_2(A) = \frac{\sigma_1}{\sigma_p}$
- $\|A\|_F = \sqrt{\sum_{i=1}^{r} \sigma_i^2} = \left\| \begin{pmatrix} \sigma_1 \\ \vdots \\ \sigma_p \end{pmatrix} \right\|_2$
- $\left\| A - \sum_{i=1}^{k} \sigma_i u_i v_i^T \right\|_2 = \left\| \sum_{i=k+1}^{r} \sigma_i u_i v_i^T \right\|_2 = \sigma_{k+1}$

5 Nichtlineare Gleichungen

5.1 Intervallschachtelung / Bisektion

Sei $f \in C[a,b]$. Suche $x \in [a,b]$ mit $f(x) = 0$. Gilt $a_0, b_0 \in [a,b]$ mit $f(a_0) \cdot f(b_0) < 0$, so hat f eine Nullstelle in $[a_0, b_0]$ (Zwischenwertsatz).

for $k = 0, 1, \ldots$ **do**
  $x_k = 1/2(a_k + b_k)$;
  if $f(a_k)f(x_k) < 0$ **then**
    $a_{k+1} = a_k$;
    $b_{k+1} = x_k$;
  else
    $a_{k+1} = x_k$;
    $b_{k+1} = b_k$;
  end
  if $|b_{k+1} - a_{k+1}| < TOL|a_{k+1}|$ **then**
    Ende Lösung: $1/2(b_{k+1} + a_{k+1})$
  end
end

Konvergenz:

$$a_k \leq a_{k+1} \leq b_{k+1} \leq b_k$$

$$|b_{k+1} - a_{k+1}| = \frac{1}{2}|b_k - a_k| = 2^{-k-1}|b_0 - a_0|$$

Eigenschaften:

- sehr stabil
- langsam
- Erweiterung für $x \in \mathbb{R}^n$ oder $x \in \mathbb{C}$ nicht möglich

5.2 Newton-Verfahren im $\mathbb{R}^n$

Sei $D \subset \mathbb{R}^n$ offen, $f : D \to \mathbb{R}^n$ stetig differenzierbar. Bezeichnung: $J(x) = f'(x) : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ (Jacobi-Matrix). Vorüberlegung: Taylor-Entwicklung von f um eine Näherungslösung $x_k \in D$:

$$f(x_k + \Delta x_k) = f(x_k) + J(x_k)\Delta x_k + \wr(\|\Delta x_k\|) \stackrel{!}{=}$$

Abgeleitete Iterationsvorschrift:

- Löse $J(x_k)\Delta x_k = -f(x_k)$
- Schritt $x_{k+1} = x_k + \Delta x_k$

Insbesondere Fall $n = 1$: $J(x_k)\Delta x_k = -f(x_k) \to \Delta x_k + x_k$ = Nullstelle der Tangente an der Stelle x_k.

5.3 Konvergenzverhalten iterativer Methoden (Spezialfall $n = 1$)

Definition 5.1 Ein Iterationsverfahren zur Berechnung von

$$x_* = \lim_{k\to\infty} x_k$$

hat eine Konvergenz der Ordnung $\alpha, \alpha \geq 1$, wenn mit einem $c > 0$ gilt:

$$|x_{k+1} - x_*| \leq c|x_k - x_*|^\alpha \qquad k = 0, 1, \ldots$$

Im Fall $\alpha = 1$ (lineare Konvergenz) heißt das beste c lineare Kontraktionsrate. Gilt

$$|x_{k+1} - x_k| \leq c_k|x_k - x_*|$$

mit einer Nullfolge $c_k \to 0$, so spricht man von superlinearer Konvergenz.

Definition 5.2 Die Menge $D(x) = \{y \in D \mid \|f(y)\| \leq \|x\|\}$ heißt die Niveaumenge von f zum Punkt x.

Satz 5.3 (Newton-Kantorovich) Für ein $\bar{x} \in D$ gelte

1. $\left\|J^{-1}(x)\right\| \leq \beta, x \in D_f(\bar{x})$
2. $\|J(x) - J(y)\| \leq \gamma\|x - y\|, x, \in D_f(\bar{x})$
3. $x_0 \in D_f(x)$
4. $q := 1/2\alpha\beta\gamma < 1$ mit $\alpha = \left\|J^{-1}(x_0)f(x_0)\right\|$

Dann konvergiert die Folge (x_k) aus der Newtoniteration gegen eine Nullstelle $x_* \in D$ von f, mit der a-priori Fehlerabschätzung

$$\|x_k - x_*\| \leq \frac{\alpha}{1-q} q^{\left(2^k-1\right)}, k \geq 1$$

Beweis Skript □

Wiederholung: $f : D \subset \mathbb{R}^n \to \mathbb{R}^n, J(x) = f'(x) \in \mathbb{R}^{n\times n}$ Suche $x \in D : f(x) = 0$

- $n = 1$: Bisektion, (stabil)
- Newton-Typ-Verfahren: $x_0 \in D, M(x)J(x) \approx E_n$

  $$J(x_k)\Delta x_k = -f(x_k) \mid \Delta x_k = -M(x_k)f(x_k) \to x_{k+1} = x_k + \Delta k$$

  lokal quadratische Konvergenz für $M(x)J(x) = E_n$

Satz 5.4 (Lokaler Kontraktionssatz von Bock) Sei

$$\mathcal{N} := \{(x, y) \in D^2 \mid y = x - M(x)f(x)\}$$

Es Existiere ein $\omega < \infty$ so, dass für alle $(x, y) \in \mathcal{N}, t \in [0, 1]$

$$\|M(y)[J(x + t(y - x)) - J(x)](x - y)\| \leq \omega t\|y - x\|^2$$

und ein $\kappa < 1$ so, dass für alle $(x, y) \in \mathcal{N}$

$$\|M(y)[E_n - J(x)M(x)]f(x)\| \leq \kappa\|y - x\|$$

Mit $c_k := \kappa + \omega/2\|\Delta x_k\|$ gelte $x_0 < 1$ und

$$D_0 := \{y \in \mathbb{R}^n \mid \|y - x_0\| \leq \frac{\|\Delta x_0\|}{1 - c_0}\} \subset D$$

Dann bleibt $x_k \in D_0$ und $\lim_{k\to\infty} x_k = x_*$ existiert. Weiterhin gilt:

$$\|\Delta x_{k+1}\| \leq c_k\|\Delta x_k\| = \kappa\|\Delta x_k\| + \frac{\omega}{2}\|\Delta x_k\|^2$$

die a-priori Fehlerabschätzung

$$\|x_{k+j} - x_*\| \leq \frac{(c_k)^j}{1 - c_k}\|\Delta x_k\| \leq \frac{(c_0)^{k+j}}{1 - c_0}\|\Delta x_0\|$$

und $M(x_*)f(x_*) = 0$. Ist $M(x)$ stetig in x_* und $M(x_*)$ invertierbar, so gilt $f(x_*) = 0$

Beweis $c_0 < 1 \implies x_0, x_1 \in D_0$. Sei $x_{k+1} \in D_0$ und $c_k < 1$. Dann gilt

$$\begin{aligned}
\|\Delta x_k\| &= \|M(x_{k+1})f(x_{k+1})\| \\
&= \|M(x_{k+1})[f(x_k) - J(x_k)M(x_k)f(x_k)] + M(x_{k+1})[f(x_{k+1}) - f(x_k) + J(x_k)M(x_k)\\
&\leq \kappa\|x_{k+1} - x_k\| + \left\|M(x_{k+1})\int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}f(x_k + t\Delta x_k)\mathrm{d}t - J(x_k)\Delta x_k\right\| \\
&\leq \kappa\|\Delta x_k\| + \int_0^1 \|M(x_{k+1})[J(x_k + t(x_{k+1} - x_k)) - J(x_k)]\Delta x_k\|\mathrm{d}t \\
&\leq \kappa\|\Delta x_k\| + \frac{\omega}{2}\|\Delta x_k\|^2 = c_k\|\Delta x_k\| \\
\implies c_{k+1} &= \kappa + \frac{\omega}{2}\|\Delta x_{k+1}\| \leq \kappa + c_k\frac{\omega}{2}\|\Delta x_k\| = c_k - \frac{\omega}{2}\|\Delta x_k\| \\
\implies c_{k+1} &\leq c_k - \underbrace{(1 - c_k)\frac{\omega}{2}\|\Delta x_k\|}_{>0} \leq c_k \\
\implies \|x_{k+2} - x_0\| &= \|x_{k+2} - x_{k+1} + x_{k+1} \cdots - x_0\| \\
&\leq \sum_{j=0}^{k+1}\|\Delta x_j\| \leq \sum_{j=0}^{k+1}(c_0)^j\|\Delta x_0\| \\
&\leq \frac{\|\Delta x_0\|}{1 - c_0}
\end{aligned}$$

$\implies x_k \in D_0, k = 0, 1, \ldots,$ (Induktion)

(x_k) ist Cauchyfolge, wegen

$$\|x_{k+1j} - x_k\| \leq \sum_{i=k}^{k+j-1}\|\Delta x_i\| \leq \sum_{i=0}^{j-1}(c_0)^k\|\Delta x_i\| \qquad \leq (c_0)^k\frac{\|\Delta x_0\|}{1 - c_0}$$

$\implies (x_k)$ konvergiert,

$$\lim_{k\to\infty} x_k = x_*$$

$$\begin{aligned}\|x_{k+j} - x_*\| &\leq \|x_{k+j} - x_{k+j+1} + x_{k+j+1} - \dots x_*\| \\ &\leq \sum_{i=0}^{\infty} \|x_{k+j+1+1} - x_{k+j+i}\| = \sum_{i=0}^{\infty} \|\Delta x_{k+j+1}\| \\ &\leq \sum_{i=0}^{\infty} (c_k)^i \|\Delta x_{k+j}\| \leq \frac{(c_k)^j}{1-c_k} \|\Delta x_k\|\end{aligned}$$

Weiterhin $x^* = x^* - M(x^*)f(x^*) \implies M(x^*)f(x^*) = 0$ □

Diskussion:

- Ist $f(x) = Jx + b$ (affin linear) so ist $\omega = 0$. ω ist ein Maß für die Nichtlinearität von f.
- Für das Newton-Verfahren ($M(x)J(x) = E_n$) gilt $\kappa = 0$, das heißt κ ist ein Maß für die Kompatibilität von M und J
- Das Newton Verfahren für $f(x) = Jx - b$ (J invertierbar) konvergiert in einem Schritt ($\omega = \kappa = 0$)

Sukzessive Approximation
Wahl: $M(x) = C^{-1}$ mit $C \in \mathbb{R}^{n \times n}$. κ -Bedingung: $x - y \in \mathcal{N}$, das heißt $y - x = -C_1^{-1} f(x)$

$$\left\|C^{-1}[E_n - J(x)C^{-1}]f(x)\right\| = \left\|[I_n - C^{-1}J(x)](y-x)\right\| \overset{!}{\leq} \kappa \|y - x\|$$

ist erfüllt für

$$\left\|E_n - C^{-1}J(x)\right\| \leq \kappa < 1$$

Für hinreichend kleines $\|\Delta x_0\|$, das heißt in der Nähe einer Lösung gilt:

$$c_0 = \kappa + \frac{\omega}{2}\|\Delta x_0\| < 1$$

und

$$\|x_k - a_*\| \leq \frac{(c_0)^k}{1-c_0}\|\Delta x_0\|$$

Betrachtung als Fixpunktiteration (FP1)

$$\begin{aligned} g(x) &:= x - C^{-1}f(x) \\ x_{k+1} &= g(x_k) \quad k = 0, 1, \dots \\ \implies g'(x) &= E_n - C^{-1}J(x)\end{aligned}$$

$\implies$ Zu jedem Fixpunkt $x_* \in D$ von x mit $\|g'(x)\| < 1$ gibt es eine Umgebung

$$K_\rho(x_*) = \{x \in \mathbb{R}^n \mid \|x - x_*\| \leq \rho\} \subset D$$

sodass $\kappa \leq c_0 < 1$ auf $K_\rho(x_*)$ (statt D). Wiederholung: $f(x) = 0, x \in \mathbb{R}^n, x_{k+1} = x_k - M(x_k)f(x)$.

- Lokaler Kontraktionssatz
  - ω: Maß für die Nichtlinearität
  - κ: Maß für Kompatibilität von M und $f' := J$

  ist $\|\Delta x_0\|$ klein genug: dann konvergiert $(x_k) \to x^*$
  - $c_k = \kappa + \frac{\omega}{2}\|\Delta x_k\| \overset{!}{<} 1$
  - $\|\Delta x_{k+1}\| \leq c_k \|\Delta x_k\|$
  - apriori Fehlerabschätzung

$$\|x_k - x_x\| \leq \frac{(c_0)^k}{1-c_0}\|\Delta x_0\|$$

- Fixpunktiteration: $M(x) = C^{-1}$

6 Lineare Gleichungssysteme: Iterative Verfahren

Problem direkter Methoden: Speicheraufwand für große n. Alternatives Beispiel: Fixpunktiteration für $Ax = b$ ($A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n$)

$$\implies a_{jj}x_j + \sum_{\substack{k=1\\k\neq j}}^{n} a_{jk}x_k = b_j, j = 1, \dots, n$$

Ist $a_{jj} \neq 0$

$$\iff x_j = \frac{1}{a_{jj}} \left(b_j - \sum_{\substack{k=1\\k\neq j}}^{n} a_{jk}x_k \right), j = 1, \dots, n$$

Gesamtschritt- /Jacobi-Verfahren:

$$x^0 = 0$$
$$x_j^t = \frac{1}{a_{jj}} \left(b_j - \sum_{\substack{k=1\\k\neq j}}^{n} a_{jk}x_k^{t-1} \right)$$
$$j = 1, \dots, n, t = 1, 2, \dots$$

Einzelschritt- /Gauß-Seidel-Verfahren

$$x_j^t = \frac{1}{a_{jj}} \left(b_j - \sum_{k<j} a_{jk}x_k^t - \sum_{k>j} a_{jk}x_k^{t-1} \right)$$
$$j = 1, \dots, n, t = 1, 2, \dots$$

Fixpunktiterationen:

$$A = D + L + R$$

Jacobi:

$$\begin{aligned} x^t &= D^{-1}\big(b - (L + R)x^{t-1}\big) \\ &= \underbrace{-D^{-1}(L + R)}_{=:J} x^{t-1} + D^{-1}b \end{aligned}$$

Gauß-Seidel:

$$\begin{aligned} x^t &= D^{-1}\big(b - Lx^t - Rx^{t-1}\big) \\ \iff Dx^t + Lx^t &= b - Rx^{t-1} \\ \iff x^t &= -(D + L)^{-1}Rx^{t-1} + (D + L)^{-1}b \end{aligned}$$

Gemeinsame Form $x^t = Bx^{t-1} + c$, B: Iterationsmatrix. Konvergiert $\left(x^t\right)$ gegen x, so gilt $x = Bx + c$. Allgemein: Wähle $C \in \mathbb{R}^{n\times n}$ invertierbar

$$\begin{aligned} Ax = b &\iff Cx = Cx - Ax + b \\ &\iff x = x + C^{-1}(b - Ax) \end{aligned}$$

Form der Fixpunktiteration:

$$x^t = \underbrace{\left(E_n - C^{-1}A\right)}_{=:B} x^{t-1} + \underbrace{C^{-1}b}_{=:c}$$

Defektkorrekturiteration:

$$d^{t-1} = b - Ax^{t-1}, C\delta x^{t-1} = d^{t-1}$$
$$x^t = x^{t-1} + \delta x^{t-1}$$

Erinnerung: Lokaler Kontraktionssatz:

$$\kappa = \left\| E_n - C^{-1}A \right\| < 1$$

$\Longrightarrow$ Konvergenz für beliebige Startwerte ($\omega = 0$). Problem: κ ist Norm-abhängig. „Schärfere“ Alternative

$$\operatorname{spr}(B) = \max\{|\lambda| \mid \lambda \in \sigma(B)\}$$

$\sigma(B) \subset \mathbb{C}$: Menge der Eigenwerte von B ($Bx = \lambda x, \lambda \in \mathbb{C}, x \in \mathbb{C}^n, x \neq 0$). Achtung: $\operatorname{spr}(B)$ ist keine Norm. Betrachte

$$\operatorname{spr}\left(\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\right) = 0$$

aber dies ist nicht die Nullmatrix. Für natürliche Matrizennormen gilt

$$\|B\| = \sup_{x \in \mathbb{C}^n \setminus \{0\}} \frac{\|Bx\|}{\|x\|} \geq |\lambda|$$

mit λ ein Eigenwert, wählen x als den zugehörigen Eigenvektor.

$$\Longrightarrow \operatorname{spr}(B) \leq \|B\|$$

Lemma 6.1 Für jede $B \in \mathbb{R}^{n \times n}$ gibt es zu jedem $\varepsilon > 0$ eine natürliche Matrizennorm $\|\cdot\|_\varepsilon$, sodass

$$\operatorname{spr}(B) \leq \|B\|_\varepsilon \leq \operatorname{spr}(B) + \varepsilon$$

Beweis Schnur-Zerlegung $B = T^{-1}R =, T \in \mathbb{C}^{n \times n}$, unitär

$$R = \begin{pmatrix} r_{11} & \dots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{pmatrix}$$
$$\Longrightarrow \operatorname{spr}(B) = \operatorname{spr}(R) = \max_{j=1,\dots,n} |r_{jj}|$$

Für beliebige $\delta \in (0, 1]$, wähle

$$S_\delta = \operatorname{diag}\left(\delta^0, \delta^1, \dots, \delta^{n-1}\right)$$

$$R_0 = \operatorname{diag} r_{11}, r_{22}, \dots, r_{nn}$$

$$Q_\delta = \begin{pmatrix} 0 & r_{12} & \delta r_{13} & \dots & \delta^{n-2} r_{1n} \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & \delta r_{n-2,n} \\ & & & \ddots & r_{n-1,n} \\ 0 & & & & 0 \end{pmatrix}$$

$$R_\delta = S_\delta^{-1} R D_\delta = \begin{pmatrix} r_{11} & \delta r_{12} & \delta^2 r_{13} & \dots & \delta^{n-1} r_{1n} \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & \delta^2 r_{n-2,n} \\ & & & \ddots & \delta r_{n-1,n} \\ 0 & & & & r_{n,n} \end{pmatrix}$$

$$\Longrightarrow R_\delta = R_0 + \delta Q_\delta$$

$S_\delta^{-1}T$ invertierbar

$$\implies \|x\|_\delta = \left\|S_\delta^{-1}Tx\right\|_2$$

ist Vektornorm auf $\mathbb{R}^n$. Mit $B = T^{-1}RT = T^{-1}S_\delta R_\delta S_\delta^{-1}T$ und $y = S_\delta^{-1}Tx$ folgt

$$\begin{aligned}\|Bx\|_\delta &= \left\|T^{-1}S_\delta R_\delta S_\delta^{-1}Tx\right\|_\delta \\ &= \|R_\delta y\|_2 \leq \|R_0 y\|_2 + \delta\|Q_\delta y\|_2 \\ &\leq \left(\max_{i=1,\dots,n}|r_{ii}| + \delta\mu\right)\|y\|_2 \\ &= (\operatorname{spr}(B) + \delta\mu)\|x\|_\delta\end{aligned}$$

mit

$$\begin{aligned}\mu &= \left(\sum_{i,j=1}^{n}|r_{ij}|\right)^{1/2} \\ \|B\|_\delta &= \sup_{x\in\mathbb{R}^n\setminus\{0\}}\frac{\|Bx\|_\delta}{\|x\|_\delta} \\ &\leq \operatorname{spr}(B) + \delta\mu\end{aligned}$$

Wähle $\delta = \varepsilon/\mu$ □

Satz 6.2 (Fixpunktiteration) Die durch

$$x^t = Bx^{t-1} + c$$

erzeugten Iterierten konvergieren genau dann für jeden Startwert $x^0 \in \mathbb{R}^n$ gegen die Lösung von $x = Bx + c$, wenn $\operatorname{spr}(B) < 1$. Asymptotisches Konvergenzverhalten:

$$\sup_{x_0\in\mathbb{R}^n}\limsup_{t\to\infty}\left(\frac{\left\|x^t - x\right\|}{x^0 - x}\right)^{1/t} = \operatorname{spr}(B)$$

Beweis Fehler:

$$\begin{aligned}e^t := x^t - x &= Bx^{t-1} + c - Bx - c = Be^{t1} \\ \implies e^t &= B^t e^0, t \in \mathbb{N}\end{aligned}$$

1. $\operatorname{spr}(B) < 1$. Sei $\varepsilon < 1 - \operatorname{spr}(B)$

$$\implies \exists\|\cdot\|_\varepsilon : \|B\|_\varepsilon \leq \operatorname{spr}(B) + \varepsilon < 1$$

$$\left\|e^t\right\|_\varepsilon = \left\|B^t e^0\right\|_\varepsilon \leq \|B\|_\varepsilon^t\left\|e^0\right\|_\varepsilon \xrightarrow{t\to\infty} 0$$

$\implies x^t \to x$ für $x \to \infty$

2. (Beweis für Fall $B\omega = \lambda\omega, |\lambda| = \operatorname{spr}(B), \omega \in \mathbb{R}^n \setminus \{0\}$). Konvergenz für jeden Startwert. Wähle $x^0 = x + w$

$$\lambda^t\omega = B^t\omega = B^t e^{=} e^t \to 0$$

$\implies |\lambda| < 1 \implies \operatorname{spr}(B) < 1$. Weiterhin:

$$\left(\frac{\|e^t\|}{\|e^0\|}\right)^{1/t} = |\lambda|$$

3. Norm Äquivalenz: $\exists m, M > 0$, sodass

$$\begin{aligned} m\|x\| \leq \|x\|_\varepsilon \leq M\|x\| \quad x \in \mathbb{R}^n \\ \implies \left\|e^t\right\| \leq \frac{1}{m}\left\|e^t\right\|_\varepsilon \leq \frac{1}{m}\|B\|_\varepsilon^t \left\|e^0\right\|_\varepsilon \\ \leq \frac{M}{m}(\operatorname{spr}(B)+\varepsilon)^t \left\|e^0\right\| \end{aligned}$$

Wegen

$$\left(\frac{M}{m}\right)^{1/t} \xrightarrow{t\to\infty} 1$$

$$\limsup_{t\to\infty}\left(\frac{\left\|e^t\right\|}{\left\|e^0\right\|}\right)^{1/t} \leq \sup(B)+\varepsilon \xrightarrow{\varepsilon\to 0} \operatorname{spr}(B)$$

□

Wiederholung: $Ax = b, A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n$

- Fixpunktiteration: $x^{t+1} = Bx^t + c$ konvergiert genau dann $\forall x^0$, wenn $\operatorname{spr}(B) < 1$
- Jacobi: $B = J = -D^{-1}(L+R)$ wobei $A = A + L + R$
- Gauß-Seidel $B = H_1 = -(D+L)^{-1}R$
- Asymptotische Konvergenzrate:

$$\sup_{x^0\in\mathbb{R}^n}\lim_{t\to\infty}\left(\frac{\left\|e^t\right\|}{\left\|e^0\right\|}\right)^{1/t} = \operatorname{spr} B$$

Interpretation: Gewinn von k Dezimalstellen (für große t) $\rho = \operatorname{spr}(B)$. Bestimme t so, dass

$$\rho^t \leq 0,1^t \implies t\log_{10}\rho \leq -k \implies t \geq -\frac{k}{\log_{10}\rho}$$

Beispiel 6.3 ($\rho = 0.99, k = 1$) $t = 230$,

Konstruktion von Iterationsverfahren: Zwei Ziele (Gegenspieler)

1. $\operatorname{spr}\left(\underbrace{E_n - C^{-1}A}_{B}\right)$ klein
2. $C\delta x^{t-1} = d^{t_1}$ leicht lösbar

Jacobi- und Gauß-Seidel_verfahren

Satz 6.4 (Starke Zeilensummenkriterium) Ist $A \in \mathbb{R}^{n\times n}$ strikt diagonaldominant

$$\sum_{\substack{k=1\\k\neq j}}^{n}|a_{jk}| < |a_{jj}|, j = 1,\ldots,n$$

so ist $\operatorname{spr}(J) < 1$ und $\operatorname{spr}(H_1) < 1$ das heißt Jacobi- und Gauß-Seidel-Verfahren konvergieren.

Beweis $0 < |a_{jj}|$. Sei $\lambda \in \sigma(J)$ und $\mu \in \sigma(H_1)$ mit Eigenvektoren $v, w \in \mathbb{C}^n$

$$\|v\|_\infty = \|w\|_\infty = 1$$

das heißt

$$\lambda v = Jv = -D^{-1}(L+R)v$$

und

$$\begin{aligned}
\mu w &= H_1 w = -(D+L)^{-1}Rw \\
\iff \mu w &= -D^{-1}(\mu L + R)w \\
\implies |\lambda| &\leq \left\|D^{-1}(L+R)\right\|_\infty \\
&= \max_{j=1,\ldots,n}\{\frac{1}{|a_{jj}|}\sum_{\substack{k=1\\k\neq j}}^{n}|a_{jk}|\} < 1 \\
|\mu| &\leq \left\|D^{-1}(\mu L + R)\right\|_\infty \\
&\leq \max_{j=1,\ldots,n}\{\frac{1}{|a_{jj}|}\left(\sum_{k<j}|\mu||a_{jk}| + \sum_{k>j}|a_{jk}|\right)\}
\end{aligned}$$

wäre $|\mu| > 1$, so würde

$$|\mu| \leq |\mu|\left\|D^{-1}(L+R)\right\|_\infty < |\mu|\text{`}$$

□

Die Voraussetzungen können abgeschwächt werden (siehe Skript).
SOR-Verfahren (Successive Overrelaxation)

$$\begin{aligned}
\tilde{x}^t &= \frac{1}{a_{jj}}\left(b_j - \sum_{k<j} a_{jk}\tilde{x}_k^t - \sum_{k>j} a_{jk}x_k^{t-1}\right), j = 1,\ldots,n \\
x^t &= \omega\tilde{x}^t + (1-\omega)x^{t-1}, \omega \geq 1
\end{aligned}$$

Für $\omega = 1$ ist SOR gleich Gauß-Seidel ($\omega < 1$: Unterrelaxation)

$$\begin{aligned}
x^t &= -\omega(D+L)^{-1}Rx^{t-1} + (1-\omega)x^{t-1} + \omega(D+L)^{-1}b \\
H_\omega &= (D+\omega L)^{-1}((1-\omega)D - \omega R)
\end{aligned}$$

Lemma 6.5 Für $A \in \mathbb{R}^{n\times n}$ mit D regulär gilt

$$\operatorname{spr}(H_\omega) \geq |\omega - 1|, \omega \in \mathbb{R}$$

Beweis

$$\begin{aligned}
H_\omega &= \left(E_n - \omega\underbrace{D^{-1}L}_{L'}\right)^{-1} B^{-1}DD\left((1-\omega)E_n - \omega\underbrace{D^{-1}R}_{R'}\right) \\
\det(H_\omega) &= \det\left(E_n - \omega L'\right)\cdot\det\left((1-\omega)E_n - \omega R'\right) = (1-\omega)^2
\end{aligned}$$

Wegen

$$\det(H_\omega) = \prod_{\lambda\in\sigma(H_\omega)}\lambda$$

folgt

$$\begin{aligned}\operatorname{spr}(H_\omega) &= \max_{\lambda\in\sigma(H_\omega)} |\lambda| \geq \left(\prod_{\lambda\in\sigma(H_\omega)} |\lambda|\right)^{1/n} \\ &= |1-\omega|\end{aligned}$$

□

Satz 6.6 (SOR) Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch positiv definit. Dann gilt

$$\operatorname{spr}(H_\omega) < 1 \forall \omega \in (0,2)$$

Insbesondere konvergiert Gauß-Seidel.

Beweis A symmetrisch $\implies R = L^T$. $A = D + L + L^T$. Sei $\lambda \in \sigma(H_\omega)$, $\omega \in (0,2)$ mit Eigenvektor $v \in \mathbb{R}^n \setminus \{0\}$, das heißt $H_\omega v = \lambda v$

$$\begin{aligned}\implies \big((1-\omega)D - \omega L^T\big)v &= \lambda(D+\omega L)v \\ \implies \omega\big(D+L^T\big)v &= (1-\lambda)Dv + \lambda\omega Lv \\ \implies \omega Av &= \lambda\omega\big(D+L^T\big)v + \omega Lv\end{aligned}$$

und

$$\begin{aligned}\lambda\omega Av &= \lambda\omega\big(D+L^T\big)v + \lambda\omega Lv \\ &= \lambda\omega\big(D+L^T\big)v + (1-\lambda)Dv - \omega\big(D+L^T\big)v \\ &= (\lambda-1)\omega\big(D+L^T\big)v + (1-\lambda)Dv \\ &= (1-\lambda)(1-\omega)Dv - (1-\lambda)\omega L^T v\end{aligned}$$

Wegen $v\omega TLv = v^T L^T v$ folgt

1. $\omega v^T Av = (1-\lambda)v^T Dv + \omega(1-\lambda)v^T Lv$
2. $\lambda\omega v^T Av = (1-\lambda)(1-\omega)v^T Dv - (1-\lambda)\omega v^T Lv$

$$\implies (1+\lambda)\omega v^T Av = (1-\lambda)\underbrace{(2-\omega)}_{>0} v^T Dv$$

A positiv definit $\implies$ D positiv definit. Also: $v^T Av > 0, v^T Dv > 0$. $\implies \lambda \neq \pm 1$ und

$$\begin{aligned}\mu := \frac{1+\lambda}{1-\lambda} = \frac{2-\omega}{w}\frac{v^T Dv}{v^T Av} > 0 \\ \implies (1-\lambda)\mu = (1+\lambda) \\ (1+\mu)\lambda = -(1-\mu) \\ \implies |\lambda| = \left|\frac{\mu-1}{\mu+1}\right| < 1\end{aligned}$$

□

Wiederholung: SOR $Ax = b, A = D + L + R$

$$x_j^t = (1-\omega)x_j^{t-1} + \frac{\omega}{a_{jj}}\left(b_j - \sum_{k<j} a_{jk}x_k^t - \sum_{k>j} a_{jk}x_k^{t-1}\right) \quad j = 1, \dots, n$$

$$\begin{aligned} &\implies x^t = (1-\omega)x^{t-1} + \omega D^{-1}\big(b - Lx^t - Rx^{t-1}\big) \\ \implies &(D+\omega L)x^t = ((1-\omega)D - \omega R)x^{t-1} + \omega b \\ &\implies x^t = \underbrace{(D+\omega L)^{-1}((1-\omega)D - \omega R)}_{H_\omega} x^{t-1} + \omega(D+\omega L)^{-1}b \end{aligned}$$

- SOR konvergiert für A symmetrisch positiv definit $\omega \in (0, 2)$
- $\omega = 1$: Gauß-Seidel
- ω optimal ist schwer zu finden

Abstiegsverfahren
Vorraussetzung: A symmetrisch, positiv definit.

$$\begin{aligned} \implies &(Ax, y)_2 = (x, Ay)_2 \forall x, y \in \mathbb{R}^n \\ &(Ax, x)_2 > 0 \forall x \in \mathbb{R}^n \setminus \{0\} \end{aligned}$$

Definition 6.7 (A -Skalarprodukt, A -Norm)

$$(x, y)_A = (Ax, y), \|x\|_A = \sqrt{(Ax, x)}$$

Erinnerung: A hat nur reelle Eigenwerte

$$0 < \lambda := \lambda_1 \leq \dots \leq \lambda_n =: \Lambda$$

und die Eigenvektoren $\{\omega_1, \dots, \omega_n\} \subset \mathbb{R}^n$ sind eine Orthonormalbasis von $\mathbb{R}^n$

$$\implies \operatorname{spr}(A) = \Lambda, \operatorname{cond}_2(A) = \frac{\Lambda}{\lambda}$$

Satz 6.8 Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch, positiv definit. Dann gilt $Ax = b$ genau dann, wenn

$$Q(x) \leq Q(y) \forall y \in \mathbb{R}^n \setminus \{x\}$$

mit

$$Q(y) = \frac{1}{2}(Ay, y) - (b, y)$$

Beweis 1. Sei $Ax = b$ für $x \neq y$ folgt

$$\begin{aligned} Q(y) - Q(x) &= \frac{1}{2}((Ay, y) - 2(b, y) - (Ax, x) + 2(b, x)) \\ &= \frac{1}{2}((Ay, y) - 2(Ax, y) + (Ax, x)) \\ &= \frac{1}{2}(A(x-y), x-y) > 0 \end{aligned}$$

2. x ist Minimum von $Q \implies \operatorname{grad} Q(x) = 0$

$$\begin{aligned}\frac{\partial Q}{\partial x_i}(x) &= \frac{1}{2}\frac{\partial}{\partial p_i}\sum_{j,k=1}^{n} a_{jk}x_j x_k - \frac{\partial}{\partial x_i}\sum_{k=1}^{n} b_k x_k \\ &= \sum_{k=1}^{n} a_{ik}x_k - b_i = 0 \quad i = 1,\dots,n \\ \implies Ax &= b \\ \implies \operatorname{grad} Q(y) &= \frac{1}{2}\big(A + A^T\big)y - b = Ay - b \qquad \text{(negativer Defekt)}\end{aligned}$$

□

Iteration:

$$x^{t+1} = x^t + \alpha_t r^t$$

mit Abstiegsrichtung $r^t \in \mathbb{R}^n$
und Schrittweite $\alpha_t \in \mathbb{R}$. Schrittweitenbestimmung: zum Beispiel Liniensuche

$$\begin{aligned}Q\big(x^{t+1}\big) &= \min_{\alpha\in\mathbb{R}} Q\big(x^t + \alpha r^t\big) \\ \implies 0 &\overset{!}{=} \frac{\mathrm{d}}{\mathrm{d}\alpha} Q\big(x^t + \alpha r^t\big) \\ &= \operatorname{grad} Q\big(x^t + \alpha r^t\big) r^t \\ &= \big(A\big(x^t + \alpha r^t\big) - b, r^t\big) \\ &= \big(Ax^t - B, r^t\big) + \alpha\big(Ar^t, r^t\big) \\ \implies \alpha t &= -\frac{\big(t, r^t\big)_2}{\big(r^t, r^t\big)_A} \\ g^t &:= \operatorname{grad} Q\big(x^t\big) = Ax^t - b\end{aligned}$$

Definition 6.9 (Allgemeines Abstiegsverfahren) Gegeben $x^0 \in \mathbb{R}^n$

- Gradint $g^t = Ax^t - b$, Abstiegsrichtung r^t
- Schrittweite
$$\alpha_t = -\frac{\big(g^t, r^t\big)}{\big(Ar^t, r^t\big)}$$
- Iteration: $x^{t+1} = x^t + \alpha_l r^t$

Ökonomischer:

$$\begin{aligned}g^0 &= Ax^0 - b \\ t \geq 0 : \alpha_t &= \frac{\big(g^t, r^t\big)}{\big(Ar^t r^t\big)} \\ x^{t+1} &= x^t + \alpha_t r^t \\ g^{t+1} &= g^t + \alpha_l A r^t\end{aligned}$$

Beobachtung:

$$\begin{aligned}\|y-x\|_A^2 - \|x\|_A^2 = (A(y-x), y-x) - (Ax, x) \\ &= \big(A(y-x), A^{-1}A(y-x)\big) - \big(Ax, A^{-1}Ax\big) \\ &= \|Ay-b\|_{A^{-1}}^2 - \|b\|_{A^{-1}}^2 \\ &= (Ay, y) - (Ay, x) - (Ax, y) \\ &= (Ay, y) - 2(b, y) = 2Q(y)\end{aligned}$$

$\implies$ Minimierung von Q minimiert Defektnorm $\|Ay-b\|_{A^{-1}}$ und Fehlernorm $\|y-x\|_A$.
Gradientenverfahren: Richtung des steilsten Abstiegs

$$r^t = -\operatorname{grad} Q\big(x^t\big) = -g^t$$

Iteration: $x^0 \in \mathbb{R}^n, g^0 = Ax^0 - b. \, t \geq 0$:

$$\begin{aligned}\alpha_t &= \frac{\|g^t\|^2}{(Ag^t, g^t)} \\ x^{t+1} &= x^t - \alpha_t g^t \\ g^{t+1} &= g^t - \alpha_t A g^t\end{aligned}$$

Ist $\big(Ag^t, g^t\big) = 0$ folgt $g^t = 0 \implies Ax^t = b$.

Satz 6.10 (Gradientenverfahren) Ist $A \in \mathbb{R}^{n\times n}$ symmetrisch, positiv definit, so konvergiert das Gradientenverfahren für alle $x^0 \in \mathbb{R}^n$ gegen die Lösung von $Ax = b$

Beweis Fehlerfunktional

$$E(y) = \|y-x\|_A^2 = (y-x, A[y-x]), y \in \mathbb{R}^n$$

Fehler $e^t = x^t - x$

$$\begin{aligned}\implies \frac{E\big(x^t\big) - E\big(x^{t+1}\big)}{E(x^t)} &= \frac{\big(e^t, Ae^t\big) - \big(e^{t+1}, Ae^{t+1}\big)}{(e^t, Ae^t)} \\ &= \frac{\big(e^t, Ae^t\big) - \big(e^t - \alpha_t g^t, A[e^t - \alpha_t g^t]\big)}{(e^t, Ae^t)} \\ &= \frac{2\alpha_t\big(e^t, Ag^t\big) - \alpha_t^2\big(g^t, Ag^t\big)}{(e^t, Ae^t)} \\ &= \frac{2\alpha_t\big\|g^t\big\|^2 - \alpha_t^2\big(g^t, Ag^t\big)}{(g^t, A^{-1}g^t)} \\ &= \frac{2\frac{\|g^t\|^2}{(Ag^t, g^t)}\big\|g^t\big\|^2 - \frac{\|g^t\|^4}{(Ag^t, g^t)}}{(g^t, A^{-1}g^t)} \\ &= \frac{\big\|g^t\big\|^4}{(g^t, Ag^t)(g^t, A^{-1}g^t)}\end{aligned}$$

A symmetrisch, positiv definit $\implies \lambda\|y\|^2 \leq (y, Ay) \leq \Lambda\|y\|^2$

$$\Lambda^{-1}\|y\|^2 \leq \big(y, A^{-1}y\big) \leq \lambda^{-1}\|y\|^2$$

Ist $x^t \neq x$, das heißt $E(x^t) \neq 0$ und $g^t \neq 0$ folgt:

$$\frac{\|g^t\|^4}{(g^t, Ag^t)(g^t, A^{-1}g^t)} \geq \frac{\|g^t\|^4}{\Lambda\|g^t\|^2\lambda^{-1}\|g^t\|^2} = \frac{\lambda}{\Lambda}$$

$\implies E(x^{t+1}) \leq [1-\kappa^{-1}]E(x^t), \kappa := \text{cond}_2(A)$. Wegen $0 < 1-\kappa^{-1} < 1$ konvergiert $E(x^t) \xrightarrow{t\to\infty} 0$ für alle $x_0 \in \mathbb{R}^n \implies x^t \xrightarrow{t\to\infty} x$ □

Lemma 6.11 (Lemma vovn Kantorovich) Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch positiv definit mit $\lambda, \Lambda > 0$ kleinster / gröhter eigenwert. Dann

$$4\frac{\lambda\Lambda}{(\lambda+\Lambda)^2} \leq \frac{\|y^4\|}{(y, Ay)(y, A^{-1}y)}$$

Beweis Skript. □

Satz 6.12 (Fehlerabschätzung) Für das Gradientenverfahren gilt die Fehlerabschätzung

$$\|x^t - x\|_A \leq \left(\frac{1-\kappa^{-1}}{1+\kappa^{-1}}\right)^t \|x^0 - x\|_A, t \in \mathbb{N}$$

Beweis

$$E(x^{t+1}) = \left(1 - \frac{\|g^t\|^4}{(g^t, Ag^t)(g^t, \Lambda^{-1}g^t)}\right)E(x^t)$$

$$\begin{aligned}\implies E(x^{t+1}) &\leq \left(1 - 4\frac{\lambda\Lambda}{(\lambda+\Lambda)^2}\right)E(x^t)\\ &= \frac{\lambda^2 + 2\lambda\Lambda + \Lambda^2 - 4\lambda\Lambda}{(\lambda+\Lambda)^2}E(x^t) = \left(\frac{\lambda-\Lambda}{\lambda+\Lambda}\right)^2 E(x^t)\\ \implies \|x^t - x\|_A^2 &\leq \left(\frac{\Lambda-\lambda}{\Lambda+\lambda}\right)^{2t}\|x^0 - x\|_A^2\end{aligned}$$ □

$Ax = b \iff \underbrace{K^{-1}AK^{-1T}}_{\tilde{A}}\underbrace{K^T x}_{\tilde{x}} = \underbrace{K^{-1}b}_{\tilde{b}}$ Wiederholung: Allgemeines Abstiegsverfahren für $Ax = b, A \in \mathbb{R}^{n\times n}$ symmetrisch positiv definit.

$$\iff \min Q(y) = \frac{1}{2}(y, Ay) - (b, y)$$

$$g^t = Ax^t - b, \alpha_t = -\frac{(g^t, r^t)}{r^t A r^t}$$

$$x^{t+1} = x^= + \alpha_t r^t$$

Gradientenverfharen: $r^t := -g^t$. Fehlerabschätzung:

$$\|x^t - x\|_A \leq \left(\frac{1-\kappa^{-1}}{1+\kappa^{-1}}\right)\|x^0 - x\|_A, \kappa = \Lambda/\lambda = \text{cond}_2(A)$$

Beobachtung:

$$(g^{t+1}, g^t) = (g^t - \alpha_t Ag^t, g^t) = \|g^t\|^2 - \underbrace{\alpha_t(Ag^t, g^t)}_{\|g^t\|^2} = 0$$

$\implies g^{t+1} \perp g^t$. $\implies$ Langsame Konvergenz für $\operatorname{cond}_2 A \gg 1$.

Conjugate-Gradients-Verfahren (CG)

Idee: Wähle Abstiegsrichtung d^t mit $\left(d^i, d^j\right)_A = 0 \forall i \neq j$. ($A$ -orthogonal). Ansatz: $B_t := \operatorname{Lin}\{d^0, \ldots, d^{t-1}\}$.

$$x^t = x^0 + \sum_{i=0}^{t-1} \alpha_i d^i \alpha x^0 + B_t$$

bestimmt durch

$$\begin{aligned} & Q\left(x^t\right) = \min_{y \in x^0 + B_t} Q(y) \\ \iff & \left\|Ax^t - b\right\|_{A^{-1}} = \min_{x^0 + B_j} ! \\ \iff & \left\|x^t - x\right\|_A = \min_{x^0 + B_j} ! \\ \iff & 0 = \frac{\mathrm{d}Q}{\mathrm{d}\alpha_i}\left(x^t\right) = \left(\operatorname{grad} Q\left(x^t\right), d^i\right) \\ \iff & \left(Ax^t - b, d^i\right) = 0, \quad i = 0, \ldots, t-1 \\ \iff & g^t \perp B_t \end{aligned}$$

Wir legen zuerst B_t fest:

$$B_t := K_t\left(d^0; A\right), d^0 = b - Ax^0$$

mit dem Krylov-Raum

$$K_t(v, ; A) = \operatorname{Lin}\{v, Av, A^2 v, \ldots, A^{t-1} v\}$$

Motivation (Lucky breakdown). Wird $K_t\left(d^0; A\right)$ stationär, das heißt gilt

$$A^t d^0 \in K_t\left(d^0; A\right)$$

für ein $t \in \mathbb{N}$, so folgt

$$-g^t = b - Ax^t = d^0 + A\left(x^0 - x^t\right) \in d^0 + AK_t\left(d^0; A\right) \subset K_t\left(d^0; A\right) \subset K_t\left(d^0; A\right)$$

und wegen $g^t \perp K_t\left(d^0; A\right) \implies g^t = 0$, das heißt $Ax^t = b$.

Konstruktion der Richtungen $d^t \in K_{t+1}\left(d^0; A\right)$: Ansatz:

$$d^t = \underbrace{-g^t}_{\notin K_t\left(d^0; A\right)} + \underbrace{\sum_{j=0}^{t-1} \beta_j^{t-1} d^j}_{\in K_t\left(d^0; A\right)} \in K_{t+1}\left(d^0; A\right)$$

A -Orthogonalität:

$$\begin{aligned} 0 \stackrel{!}{=} \left(d^t, Ad^i\right) &= \left(-g^t, Ad^i\right) + \sum_{j=0}^{t-1} \beta_j^{t-1} \underbrace{\left(d^j, Ad^i\right)}_{=0, i \neq j} \\ &= \left(-g^t + \beta_j^{t-1} d^i, Ad^i\right), i = 0, \ldots, t-1 \end{aligned}$$

Wegen $\left(g^t, d^i\right) = 0, i = 0, \ldots, t-1$ folgt

$$\left(g^t, Ad^i\right) = 0, \quad i = 0, \ldots, t-2$$

$\implies \beta_i^{t-1} = 0, i = 0, \ldots, t-2. \, i = t-1$:

$$0 = \left(-g^t, Ad^{t-1}\right) + \beta_{t-1}^{t-1}\left(d^{t-1}, Ad^{t-1}\right)$$

$$\implies \beta_{t_1} := \beta_{t-1}^{t-1} = \frac{\left(g^t, Ad^{t-1}\right)}{\left(d^{t-1}, Ad^{t-1}\right)}$$

$$\begin{aligned} \implies d^t &= -g^t + \beta_{t-1} d^{t-1} \\ \implies g^{t+1} &= Ax^{t+1} - b = Ax^t - b + \alpha_t A d^t \\ &= g^t + \alpha_t A d^t \\ \implies \alpha_t &= -\frac{\left(g^t, d^t\right)}{\left(d^t, Ad^t\right)} \qquad \text{(klassische Form)} \\ \implies x^{t+1} &= x^t + \alpha_t d^t \end{aligned}$$

Vereinfachung:

$$\alpha_t = \frac{\left\|x^t\right\|^2}{\left(d^t, Ad^t\right)}, \beta_t = \frac{\left\|y^{t+1}\right\|^2}{\left\|g^t\right\|^2}$$

Beobachtung: $g^t \neq 0, t = 0, \ldots, n-1$

$$\implies \operatorname{Lin}\{d^0, \ldots, d^{n-1}\} = \mathbb{R}^n \implies x^n = x$$

(Gilt nur in exakter Arithmetik)

Lemma 6.13 Für ein Polynom $p \in P_t, p(0) = 1$ gelte auf einer Menge $S \subset \mathbb{R}$ mit $\sigma(A) \subset S$

$$\sup_{\mu \in S} |p(\mu)| \leq M$$

Dann gilt $\left\|x^t - x\right\|_A \leq M \left\|x^0 - x\right\|_A$

Beweis

$$\left\|x^t - x\right\|_A = \min_{y \in x^0 + B_t} \|y - x\|_A$$

Wegen $B_t = \operatorname{Lin}\{d^0, \ldots, d^{t-1}\} = \operatorname{Lin}\{A^0 g^0, \ldots, A^{t-1}, g^0\}$

$$\begin{aligned} \implies \left\|x^t - x\right\|_A &= \min_{p \in P_{t-1}} \left\|x_0 - x + p(A) g^0\right\|_A \\ &= \min_{p \in P_{t-1}} \left\|[E_n + Ap(A)]\left(x^0 - x\right)\right\|_A \\ &\leq \min_{\substack{p \in P_t \\ p(0)=1}} \|p(A)\|_A \left\|x^0 - x\right\|_A \end{aligned}$$

Orthonormalbasis $\{\omega_1, \ldots, \omega_n\}$ aus Eigenvektoren mit Eigenwerten λ_i

$$y = \sum_{i=1}^n \gamma_i \omega_i, \quad \gamma_i = (y, \omega_i)$$

$$\begin{aligned} \|p(A)y\|_A^2 &= \sum_{i=1}^n \lambda_i p(\lambda_i)^2 \gamma_i^2 \\ &\leq M^2 \sum_{i=1}^n \lambda_i \gamma_i^2 = M^2 \|y\|_A^2 \end{aligned}$$

$$\implies \|p(A)\|_A = \sup_{y \in \mathbb{R}^n \setminus \{0\}} \frac{\|p(A)y\|_A}{\|y\|_A} \leq M$$

□

Satz 6.14 (CG-Konvergenz) Mit $\kappa = \Lambda/\lambda$ gilt

$$\left\|x^t - x\right\|_A \leq \left(\frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}}\right)^t \left\|x^0 - x\right\|_A$$

Beweis $S = [\lambda, \Lambda]$. Lemma $\Longrightarrow$

$$\left\|x^t - x\right\|_A \leq \min_{\substack{p \in P_t \\ p(0)=1}} \sup_{\mu \in [\lambda,\Lambda]} |p(\mu)| \left\|x^0 - x\right\|_A$$

zu zeigen:

$$\min_{\substack{p \in P_t \\ p(0)=1}} \sup_{\mu \in [\lambda,\Lambda]} |p(\mu)| \left\|x^0 - x\right\|_A \leq 2\left(\frac{1 - \sqrt{\lambda/\Lambda}}{1 + \sqrt{\lambda/\Lambda}}\right)^t$$

Problem der Bestapproximation mit Polynomen bezüglich Max-Norm! Lösung: T_t: t -tes Tschebyscheff-Polynom auf $[-1, 1]$

$$\bar{p}(\mu) = T_t\left(\frac{\Lambda + \lambda - 2\mu}{\Lambda - \lambda}\right) T\left(\frac{\Lambda + \lambda}{\Lambda - \lambda}\right)^{-1}$$

$$\Longrightarrow \sup_{\mu \in [\lambda,\Lambda]} |\bar{p}(\mu)| = T\left(\frac{\Lambda + \lambda}{\Lambda - \lambda}\right)^{-1}$$

Darstellung für $|\mu| > 1$

$$T_t(\mu) = \frac{1}{2}\left(\left(\mu + \sqrt{\mu^2 - 1}\right)^t + \left(\mu - \sqrt{\mu^2 - 1}\right)^t\right)$$

$$T_t(\mu) = \frac{1}{2}\left(\left(\frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1}\right)^t + \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^t\right) \geq \frac{1}{2}\left(\frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1}\right)^t \qquad \square$$

Wiederholung: CG, Abstiegsverfahren. $\left(d^i, Ad^j\right) = 0, i \neq j$.

- 2 Parameter

$$\begin{aligned}
\alpha_t &= \frac{\left\|g^t\right\|^2}{(d^t, Ad^t)} \\
\beta_t &= \frac{\left\|g^{t+1}\right\|^2}{\left\|g^t\right\|^2} \\
x^{t+1} &= x^t + \alpha_t d^t \\
g^{t+1} &= g^t + \alpha_t A d^t \\
d^{t+1} &= -g^{t+1} + \beta_t d^t \\
d^0 &= -g^0 = b - Ax^0
\end{aligned}$$

- Axakte Arithmetik: Lösung nach spätestens $n + 1$ Schritten.
-

$$\left\|x^t - x\right\|_A \leq \min_{\substack{p \in P_t \\ p(0)=1}} \max_{\lambda \in \sigma(A)} |p(\lambda)| \cdot \left\|x^0 - x\right\|_A \leq 2\left(\frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}}\right)^t \left\|x^0 - x\right\|_A$$

7 Matrizeneigenwertaufgaben

$A \in \mathbb{K}^{n\times n}, \mathbb{K} = \mathbb{R}$ oder $\mathbb{C}$. Suche $(\lambda, \omega) \in \mathbb{K} \times (\mathbb{K}^n \setminus \{0\})$, sodass

$$A\omega = \lambda\omega$$

7.1 Konditionierung des Eigenwert-Problems.

Lemma 7.1 (Stabilität) Sei $A, B \in \mathbb{K}^{n\times n}$, $\|\cdot\|$ die natürliche Matrizennorm, $\lambda \in \sigma(A) \setminus \sigma(B)$. Dann gilt

$$\left\|(\lambda E_n - B)^{-1}(A - B)\right\| \geq 1$$

Beweis

$$\begin{aligned}
A\omega &= \lambda\omega \\
\implies (A-B)\omega &= (\lambda E_n - B)\omega \\
\implies (\lambda E_n - B)^{-1}(A-B)\omega &= \omega \\
\implies 1 &= \frac{\left\|(\lambda E_n - B)^{-1}(A-B)\omega\right\|}{\|\omega\|} \\
&\leq \sup_{x\in\mathbb{K}^n\setminus\{0\}} \frac{\left\|(\lambda E_n - B)^{-1}(A-B)x\right\|}{\|x\|} \\
&= \left\|(\lambda E_n - B)^{-1}(A-B)\right\|
\end{aligned}$$

□

Satz 7.2 (Satz von Gerschgorin) Alle Eigenwerte von $A \in \mathbb{K}^{n\times n}$ liegen in der Vereinigung der sogenannten Gerschgorin-Kreise $(j = 1, \ldots, n)$

$$K_j = \{z \in \mathbb{C} \mid |z - a_{jj}| \leq \sum_{\substack{k=1\\k\neq j}}^{n} |a_{jk}|\}$$

Sind $U = \cup_{i=1}^{m} K_{ji}$ und $V = \cup_{j=1}^{n} K_j \setminus U$ disjunkt, so liegen in U genau m und in V genau $n - m$ Eigenwerte.

Beweis 1. $B = D = \operatorname{diag}(a_{11}, \ldots, a_{nn})$. Maximale Zeilensummennorm:

$$\begin{aligned}
1 &\leq \left\|(\lambda E_n - D)^{-1}(A-D)\right\|_\infty \\
&= \max_{j=1,\ldots,n} \frac{1}{|\lambda - a_{jj}|} \sum_{\substack{k=1\\k\neq j}}^{n} |a_{jk}|
\end{aligned}$$

für $\lambda \neq a_{jj}, j = 1, \ldots, n \implies \lambda \in K_{j^*}$ mit $j^* = \operatorname{argmax}\{\ldots\}$.

2. Für $t \in [0, 1]$ setze $A_t = (1-t)D + tA$. m Eigenwerte von A_0 liegen in U und $n - m$ Eigenwerte in V. Wegen Stetigkeit der Eigenwerte von A_t bezüglich t folgt die Behauptung für $A_1 = A$ □

Satz 7.3 (Stabilitätssatz) Sei $A \in \mathbb{K}^{n\times n}$ mit n linear unabhängigen Eigenvektoren $\{\omega_1, \ldots, \omega_n\}$ und $B \in \mathbb{K}^{n\times n}$. Dann gibt es zu jedem Eigenvert $\lambda(B)$ von B einen Eigenwert $\lambda(A)$ von A mit

$$|\lambda(A) - \lambda(B)| \leq \operatorname{cond}_2(W)\|A - B\|_2$$

wobei $W = (\omega_1, \ldots, \omega_n) \in \mathbb{K}^{n\times n}$

Beweis $AW = W\Lambda, \Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n) \implies A = W\Lambda W^{-1}$. Sei $\lambda \in \sigma(B) \setminus \sigma(A)$

$$\begin{aligned}
\implies \left\|(\lambda E_n - A)^{-1}\right\|_2 &= \left\|W(\lambda E_n - \Lambda)^{-1}W^{-1}\right\|_2 \\
&\leq \|W\|_2 \left\|W^{-1}\right\|_2 \left\|(\lambda E_n - \Lambda)^{-1}\right\|_2 \\
&= \operatorname{cond}_2(W) \max_{i=1,\dots,n} |\lambda - \lambda_i|^{-1} \\
\implies 1 &\leq \left\|(\lambda E_n - A)^{-1}\right\|_2 \|A - B\|_2 \\
&\leq \operatorname{cond}_2(W)\|A - B\|_2 \max_{i=1,\dots,n} |\lambda - \lambda_i|^{-1} \\
\implies |_{i=1,\dots,n}\ |\lambda - \lambda_i| &\leq \operatorname{cond}_2(W)\|A - B\|_2
\end{aligned}$$

□

A hermitesch: W orthonormal $\implies \operatorname{cond}_2(W) = 1$. Regel: Das Eigenwert-Problem hermitescher Matrizen ist gut konditioniert, während das allgemeine ja nach $\operatorname{cond}_2(W)$ beliebig schlecht konditioniert ist.

7.2 Iterative Methoden

Verfahren um einen (nicht alle) Eigenwert zu finden. Potenzmethode (von Mises) $z^0 \in \mathbb{C}^n, \left\|z^0\right\| = 0. t \geq 1$:

$$\tilde{z}^t = Az^{t-1}, z^t = \frac{\tilde{z}^t}{\|\tilde{z}\|^t}$$

$$\lambda^{(t)} := \frac{\tilde{z}_k^t}{z_k^t}, z_k^t \neq 0$$

Satz 7.4 Sei A diagonalisierbar mit $|\lambda_n| > |\lambda_i|, i = 1, \dots, n$. z^0 habe eine nichtverschwindende Komponente bezüglich Eigenvektor ω_n. Dann gibt es $\sigma_t \in \mathbb{C}, |\sigma_t| = 1$, sodass

$$\left\|z^t - \sigma_t\omega_n\right\| \xrightarrow{t\to\infty} 0$$

und

$$\lambda^{(t)} - \lambda_n = \mathcal{O}\left(\left|\frac{\lambda_{n-1}}{\lambda_n}\right|^t\right)$$

Beweis Skript. □

Hermitesches A:

$$\lambda^{(t)} = \frac{\left(z^t, Az^t\right)_2}{\left(z^t, z^t\right)_2}$$

(Rayleigh-Quotient)

$$\to \lambda^{(t)} = \lambda_n + \mathcal{O}\left(\left|\frac{\lambda_{n+1}}{\lambda_n}\right|^{2t}\right)$$

Inverse Iteration.
Annahme: Gute Näherung $\tilde{\lambda}$ für λ_k verfügbar. Beobachtung: Ist $\tilde{\lambda} \notin \sigma(A)$ so hat $\left(A - \tilde{\lambda}E_n\right)^{-1}$ die Eigenwerte $\mu_i = \left(\lambda_i - \tilde{\lambda}\right)^{-1}$. Idee: Potenzmethode für $\left(A - \tilde{\lambda}E_n\right)^{-1}$. Löse $\left(A - \tilde{\lambda}E_n\right)\tilde{z}^t = z^{t-1}$. Normiere

$$z^t = \frac{\tilde{z}^t}{\|\tilde{z}^t\|}$$

Wiederholung: $Ax = \lambda x$

- Potenzmethode: $z^0 \in \mathbb{C}^n$

$$\begin{aligned} \tilde{z}^{(t)} &= Az^{(t)} \\ z^{(t+1)} &= \frac{\tilde{z}^{(t)}}{\left\|\tilde{z}^{(t)}\right\|} \\ \lambda^{(t)} &= \frac{\tilde{z}_t^{(t)}}{z_k^{(t)}} \\ \lambda^{(t)} &= \lambda_n + \mathcal{O}\left(\left|\frac{\lambda_{n-1}}{\lambda_n}\right|^t\right) \end{aligned}$$

- Inverse Iteration = Potentzmethode auf $\left(A - \tilde{\lambda}E_n\right)^{-1}$

7.3 Reduktionsmethoden

Definition 7.5 $A, B \in \mathbb{C}^{n\times n}$ heißen ähnlich ($A \sim B$), wenn $\exists T \in C^{n\times n}$ invertierbar mit $A = T^{-1}BT$.

Beobachtungen:

$$\begin{aligned} \det(A - zE_n) = \det\big(T^{-1}(B - zE_n)T\big) & \\ &= \det\big(T^{-1}\big)\det(B - zE_n)\det(T) \\ &= \det(B - zE_n) \\ \implies \sigma(A) &= \sigma(B) \end{aligned}$$

$$A\omega = \lambda\omega \implies T^{-1}BT\omega = \lambda W \implies B\underbrace{T\omega}_{\tilde{\omega}} = \lambda\underbrace{T\omega}_{\tilde{\omega}}$$

Reduktionsmethode: Benutze Ähnlichkeitstransformationen, um A auf eine Gestalt zu bringen, in der men die Eigenwerte leicht ablesen kann:

$$A =: A^{(0)} = T_1^{-1}A^{(1)}T_1 = \cdots = T_i^{-1}A^{-1}T_i = \ldots$$

Definition 7.6 (Jordansche Normalform) Jede Matrix $A \in \mathbb{C}^{n\times n}$ ist ähnlich zu einer Blockmatrix der Form ($\lambda_i \neq \lambda_j, i \neq j$)

$$\operatorname{diag}\Big(C_{r_1^{(1)}}(\lambda_1), \ldots, C_{r_{\rho_1}^{(1)}}(\lambda_1), \ldots, C_{r_1^{(m)}}(\lambda_m)\Big), \ldots, C_{r_{\rho_m}^{(m)}}(\lambda_m)$$

mit Jordan-Blöcken

$$C_r(\lambda) = \begin{pmatrix} \lambda & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{pmatrix} \in C^{r\times r}$$

Beobachtung:

- $\sigma(C_r(\lambda)) = \{\lambda\}$
- $\ker(C_r(\lambda) - \lambda E_n) = \operatorname{Lin}\{e_1\}$

$\implies$

- Algebraische Vielfachkeit von λ_i:

$$\sum_{j=1}^{\rho_i} r_j^{(i)} = \sigma_i$$

- Geometrische Vielfachheit von λ_i : ρ_i

Achtung: Jordan-Zerlegung numerisch nicht sinnvoll ($\mathrm{cond}(T) \gg 1/eps$)

Lemma 7.7 Für $A \in \mathbb{C}^{n\times n}$ ist äquivalent:

1. A ist diagonalisierbar
2. $\exists$ Basis von $\mathbb{C}^n$ aus Eigenvektoren von A
3. $\sigma_i = \rho_i, i = 1, \ldots, m$

Schursche Normalform
Sei $A \in \mathbb{C}^{n\times n}$. Dann $\exists U \in \mathbb{C}^{n\times n}$ unitär, sodass

$$\bar{U}^T A U = \begin{pmatrix} \lambda_1 & & \\ & \ddots & * \\ 0 & & \lambda_n \end{pmatrix}$$

Folgerung: Wenn $A = \bar{A}^T$, dann ist $\bar{U}^T A U$ hermitesch. $\implies$ A diagonalisierbar. Fehlefortpflanzung bei Reduktionsmethoden. Sei $B \sim A$, das heißt $B = T^{-1}AT$

$$\begin{aligned} B + \delta B &= T^{-1}(A + \delta A)T \\ \implies \delta A &= T\delta B T^{-1} \\ \implies \|B\| &\le \mathrm{cond}\, T \|A\| \\ \|\delta A\| &\le \mathrm{cond}(T)\|\delta B\| \\ \implies \frac{\|\delta A\|}{\|A\|} &\le \mathrm{cond}(T)^2 \frac{\|\delta B\|}{\|B\|} \end{aligned}$$

Wegen $\mathrm{cond}(T) = \mathrm{cond}(T_1 \ldots T_m) \le \mathrm{cond}(T_1) \cdot \ldots \cdot \mathrm{cond}(T_m)$ muss $\mathrm{cond}(T_i)$ klein gewährleistet sei. T unitär $\implies$ $\mathrm{cond}_2(T_i) = 1$. Reeller Fall

Definition 7.8 $A \in \mathbb{R}^{n\times n}$ heißt Hessenberg-Matrix, wenn $a_{ij} = 0 \forall i > j + 1$

Satz 7.9 (Hessenbergsche Normalform) Für jede Matrix $A \in \mathbb{R}^{n\times n}$ existieren Householdertransformationen $T_1, \ldots, T_{n-2}$ so, dass mit $T = T_{n-2} \cdot \ldots \cdot T_1$ die Matriz TAT^T Hessenberg ist. Für $A = A^T$ ist TAT^T tridiagonal.

Beweis $A = [a_1, \ldots, a_n]$. Wähle $a_1 = (0, u_{12}, \ldots, 2_{1n})^T \in \mathbb{R}^n, \|u_1\|_2 = 1$ sodass

$$T_1 a_1 = \left(E_n - 2u_1 u_1^T\right) a_1 \in \mathrm{Lin}\{e_1, e_2\}$$

$$\implies A^{(1)} = T_1 A T_1^T = \left(\begin{array}{c|ccc} a_{11} & a_{12} & \ldots & a_{1n} \\ \hline * & & & \\ 0 & & & \\ \vdots & & * & \\ 0 & & & \end{array}\right) \cdot \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & * \\ & \end{array}\right)$$

$$= \left(\begin{array}{c|ccc} a_{11} & * & \ldots & * \\ \hline * & & & \\ 0 & & & \\ \vdots & & \tilde{A}^{(1)} & \\ 0 & & & \end{array}\right)$$

Fahre fort auf $\tilde{A}^{(1)}$ für $n-2$ Schritte $\implies A^{(n-2)}$ Hessenberg. □

Verfahren für Hessenberg-/Tridiagonalmatrizen: QR-Verfahren:

$$\begin{aligned} A^{(t)} &=: Q^{(t)}R^{(t)}, A^{(t+1)} = R^{(t)}Q^{(t)} \\ A^{(t+1)} &= \underbrace{\left(Q^{(t)}\right)^T\left(Q^{(t)}\right)}_{E_n} R^{(t)}Q^{(t)} = \left(Q^{(t)}\right)^T A^{(t)}Q^{(t)} \\ \implies A^{(t+1)} &\sim A^{(t)} \end{aligned}$$

Satz 7.10 Für die Eigenwerte λ_i von A gelte $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$. Dann gilt für $A^{(t)} = \left(a_{jk}^{(t)}\right)$ aus dem QR-Verfahren

$$\lim_{t\to\infty} a_{jj}^{(t)} = \lambda_j, \quad j = 1, \ldots, n$$