Praktische Analysis

Mitschrift von www.kuertz.name

Hinweis: Dies ist **kein offizielles Script**, sondern nur eine private Mitschrift. Die Mitschriften sind teweilse **unvollständig, falsch oder inaktuell**, da sie aus dem Zeitraum 2001–2005 stammen. Falls jemand einen Fehler entdeckt, so freue ich mich dennoch über einen kurzen Hinweis per E-Mail – vielen Dank!

Klaas Ole Kürtz (klaasole@kuertz.net)

Inhaltsverzeichnis

Organisatorisches

- **Jan Modersitzki** (jmo@numerik.uni-kiel.de), Homepages:
  - http://www.math.uni-luebeck.de/modersitzki
  - http://www.numerik.uni-kiel.de/ jmo
- Scheinkriterium: 40% eines jeden Übungsblattes (eines muss nicht bearbeitet werden)
- **Inhalt** der Vorlesung: Einführung, Rundungsfehler, Lösung linearer Gleichungssysteme, Interpolation, Quadratur, Nichtlineare Gleichungen, Eigenwertprobleme, Lineare Optimierung, Optimierung, Gewöhnliche Differentialgleichungen
- **Praktische Analysis** (Numerik, Wissenschaftliches Rechnen) beschäftigt sich mit:
  - Rechnen mit Zahlen auf einem Rechner
  - Entwicklung konstruktiver Methoden für Problemklassen
  - *Algorithmus:* durch endlichen Text beschriebene Vorschrift zum Vollzug einer endlichen Reihe von Elementaroperationen. Dazu:
    - abstrakt
    - allgemein: Aufgabenklasse
    - finit: endlich viel Speicher für Programm und Daten
    - terminierend: endliche Laufzeit
    - deterministisch: gleicher Input $\rightarrow$ gleicher Output.
    - Ablauf auf einem Rechner: *Programm*, Umsetzung des Algorithmus in Programmiersprache
    - Komplexität, Rechenzeit, Anzahl der Rechenoperationen
    - Stabilität (bei gestörtem Input, dazu später mehr)
- Praktische Analysis am **Beispiel Räuber-Beute**: Es gebe x kleine und y große Fische. Als Modell für die Änderung der Anzahl der Fische könnte man nehmen

$$\begin{aligned} \dot{x} &= ax - bx \cdot y \\ \dot{y} &= cx \cdot y - dy \end{aligned}$$

Um die Lösunsfunktion zu bestimmen, machen wir folgendes:

$$\begin{aligned} \dot{x} &\approx \frac{x(t+\Delta t) - x(t)}{\Delta t} \\ x(t+\Delta t) &\approx x(t) + \Delta t(a \cdot x(t) - b \cdot x(t) \cdot y(t)) \\ y(t+\Delta t) &\approx \dots \end{aligned}$$

Wie gut ist dieses Verfahren? Was passiert auf dem Rechner? Wie exakt sind die Ergebnisse? Was hat die numerische Lösung mit dem Modell zu tun? Was hat das Modell mit der Wirklichkeit zu tun?

Ergebnisse sind i.a. Approximationen an die Lösung. Für x - Input, y - Output, $y = f(x)$ real, $\hat{y} = \text{Algo}(\hat{x})$ (Näherungen $\hat{x}$ und $\hat{y}$) sei

$$\begin{aligned} \text{absolute Fehler: } \Delta y &:= y - \hat{y} \\ \text{relativer Fehler: } \delta_y &:= \frac{\Delta y}{y},\ y \neq 0 \end{aligned}$$

1 Rundungsfehler

1.1 Zahlendarstellungen

- Analogrechner, Rechenschieber. Genauigkeit ist physikalische Messgrösse, verfolgen wir nicht.
- Digitalrechner. Darstellung im Dualsystem.

$$\begin{aligned} x &= \pm\left[\alpha_n \cdot 2^n + \alpha_{n-1} \cdot 2^{n-1} + \ldots + \alpha_m \cdot 2^m\right],\ n \geq m \in \mathbb{Z}, \alpha_j \in \{0,1\} \\ x &= 18,5 \\ &= 16 + 2 + \frac{1}{2} \\ &= 1 \cdot 2^4 + 0 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 0 \cdot 2^0 + 1 \cdot 2^{-1} \\ &= (10010.1)_2 \end{aligned}$$

  Wichtig auf Rechner: eindeutig und endlich. Problematisch: $\pi, e, \sqrt{2}, \frac{1}{3}$.

- **Fixpunktdarstellung** (z.B. in der Bank): Eine Zahl wird dargestellt z.B. in der Form

$$\underbrace{[\_\ \_\ \_\quad \cdots \quad \_\cdot\_\ \_]}_{16 \text{ Ziffern}}$$

1.2 Gleitpunktdarstellung

Definition: In der Gleitpunktdarstellung Eine Zahl $x \neq 0$ wird dargestellt durch $x = \pm a \cdot B^c$ mit der *Basis* B, der *Mantisse* a mit $\frac{1}{B} \leq a < 1$ und dem *Exponent* c.

Beispiel: $x = 18.5 = (10010.1)_2 = (0.100101)_2 \cdot 2^{(101)_2} = 0.100101_2101$.

Im Speicher haben die Mantisse und der Exponent feste Längen, d.h.

$$x = \pm\left[0.\alpha_1\alpha_2 \ldots \alpha_t\right] \cdot B^{\pm[c_1 \ldots c_E]}$$

mit *Mantissenlänge* t, *Exponentenlänge* E; z.B für $t = 8, E = 4$ ist

$$x = 18.5 = +0.10010100_2 + 0101$$

Definition: Ein *Rechner* ist charakterisiert durch B, t, E, daher beschrieben durch $R(B, t, E)$. Weiterhin sei $A := \{\, x \mid x \text{ ist darstellbar auf Rechner } \}$.

Ganz wichtig: Die Menge A der darstellbaren Zahlen ist endlich!

Problem: Für $x \in \mathbb{R}$ bestimme eine Zahl $\mathrm{Rd}[x] \in A$ mit $|x - \mathrm{Rd}[x]|$ minimal, d.h. $|x - \mathrm{Rd}[x]| \leq |x - y|$ für alle $y \in A$. Lösung: Rundung, z.B. für $t = 4, c = 1, B = 10$ ist

$$\begin{aligned} \mathrm{Rd}[0.142857] &= 0.1429 \\ \mathrm{Rd}[3.141549] &= 0.3142 \cdot 10^1 \\ \mathrm{Rd}[148.842] &= 0.1488 \cdot 10^3 \end{aligned}$$

Rundungsalgorithmus: $y = \tilde{\mathrm{Rd}}[x, t, B]$ [1]. Sei $x \in \mathbb{R}$. Ist $x = 0$, setze $y = 0$. Sonst $x = aB^c \neq 0$ mit $1/B \leq a < 1$.

$$\begin{aligned} |a| &= (0.\alpha_1\alpha_2 \ldots \alpha_t\alpha_{t+1} \ldots)_B \\ \hat{a} &:= \begin{cases} (0.\alpha_1\alpha_2 \ldots \alpha_t)_B, & \text{falls } \alpha_{1+1} < B/2 \\ (0.\alpha_1\alpha_2 \ldots \alpha_t)_B + B^{-t}, & \text{falls } \alpha_{t+1} \geq B/2 \end{cases} \\ y &:= \tilde{\mathrm{Rd}}[x] := \mathrm{sign}(x) \cdot \hat{a} \cdot B^c \end{aligned}$$

Satz: Für $x \in \mathbb{R}, x \neq 0$ gilt:

$$\frac{\left|x - \tilde{\mathrm{Rd}}[x, B, t]\right|}{|x|} = \frac{|a| - \hat{a}}{|a|} \leq 0.5 \cdot B^{1-t}$$

Beweis:

$$\frac{|a| - \hat{a}}{|a|} \leq \frac{0.5 \cdot B^{-t}}{|a|} \leq \frac{0.5 \cdot B^{-t}}{1/B} = 0.5 \cdot B^{1-t}$$

□

Definition: Die Größe $\varepsilon := 0.5 \cdot B^{1-t}$ eines Rechners $R(t, B, E)$ heißt *Maschinengenauigkeit*.

Bemerkung: Falls $\tilde{\mathrm{Rd}}[x] \in A$ gilt $\left|x - \tilde{\mathrm{Rd}}[x]\right| \leq |x - y|$ für alle $y \in A$; aber: auch der Exponent ist endlich auf dem Rechner.

Beispiele:

- auf $R(t, B, E) = R(4, 2, 2)$ ist $\tilde{\mathrm{Rd}}[18.5] = 0.1001_2101 \notin A$
- für $R(t, B, E) = R(4, 10, 1)$ gilt:
  - $\tilde{\mathrm{Rd}}[3.1415_{10}20] = 0.3142_{10}21 \notin A$
  - $\tilde{\mathrm{Rd}}[0.01_{10} - 9] = 0.1000_{10} - 10 \notin A$

[1] $\tilde{\mathrm{Rd}}$ ist nicht das gleiche wie Rd, bei $\tilde{\mathrm{Rd}}$ wird die Endlichkeit des Exponenten nicht berücksichtigt.

– $\tilde{\mathrm{Rd}}[0.99999_{10}9] = 0.1000_{10}10 \notin A$

Mögliche Idee: Falls $\tilde{\mathrm{Rd}}[|x|] < 0.1000_{10} - 9$, so setze $\mathrm{Rd}[|x|] = 0$, aber: der relative Fehler beträgt $\left|\frac{x-\mathrm{Rd}[x]}{x}\right| = 1 = 100\%$!

Über- bzw. Unterläufe sind selten, da in der Regel E sehr groß ist; daher hier ein theoretisches Modell mit $E = \infty$.

Die Rundungsfunktion hat also die Gestalt $\mathrm{Rd} : \mathbb{R} \to A$ mit $\mathrm{Rd}[x] = x(1+\delta_x)$, wobei $|\delta_x| \leq \varepsilon$ für alle $x \in R$. Für absolute und relative Fehler gelten:

$$\begin{aligned} \text{absolute Fehler: } \Delta x &:= x - \mathrm{Rd}[x] \\ \text{relativer Fehler: } \delta_x &:= \frac{x - \mathrm{Rd}[x]}{x},\ x \neq 0 \end{aligned}$$

1.3 Die Operationen +, -, *, /; Auslöschung

Das Resolutat einer arithmetischen Operation muß keine Maschinenzahl sein, selbst wenn die Operanden Maschinenzahlen sind! Seien $x, y \in A$, dann wäre optimal, wenn bei folgenden Operationen $|\varepsilon_i|$ jeweils $\leq \varepsilon$ wäre:

$$\begin{aligned} x\hat{+}y &:= \mathrm{Rd}[x+y] = (x+y)(1+\varepsilon_1) \\ x\hat{-}y &:= \mathrm{Rd}[x-y] = (x-y)(1+\varepsilon_2) \\ x\hat{\cdot}y &:= \mathrm{Rd}[x\cdot y] = (x\cdot y)(1+\varepsilon_3) \\ x\hat{\div}y &:= \mathrm{Rd}[x\div y] = (x\div y)(1+\varepsilon_4) \end{aligned}$$

Diese Operationen haben andere Eigenschaften als die entsprechenden mathematischen Operationen, z.B. gilt $x\hat{+}z = x$ für **alle** z mit $|z| \leq \frac{\varepsilon}{B} \cdot x$ und $(x\hat{+}y)\hat{+}z \neq x\hat{+}(y\hat{+}z)$.

Beispiel: Betrachte auf $R(t, B, E) = R(6, 10, \infty)$ die Zahl $x := 1234.567$, die Rundung ist $\mathrm{Rd}[x] = 1234.57$, der relative Fehler ist also $\frac{x-\mathrm{Rd}[x]}{x} \approx 2.5 \cdot 10^{-6}$. Für $y := -1234.60$ (mit $\mathrm{Rd}[y] = y$). Dann ist:

$$\left|\frac{(x+y) - (\mathrm{Rd}[x] + \mathrm{Rd}[y])}{x+y}\right| = \left|\frac{x - \mathrm{Rd}[x]}{0.033}\right| = \frac{0.003}{0.033} = \frac{1}{11}$$

Das ist ein relativer Fehler von ca. **10%** (bei einer Maschinengenauigkeit von $\varepsilon = 5 \cdot 10^{-6}$). Der relative Fehler bezüglich der Addition ist

$$\frac{(x \pm y) - (\mathrm{Rd}[x] \pm \mathrm{Rd}[y])}{x \pm y} = \frac{\delta_x x}{x \pm y} \pm \frac{\delta_y y}{x \pm y} \overset{y \to \mp x}{\longrightarrow} \infty$$

Dieses Phänomen heißt *Auslöschung*: **vermeiden!** Zum Beispiel berechnet man für $x \approx y$ statt $\log(x) - \log(y)$ besser $\log(\frac{x}{y})$.

1.4 Kondition und Stabilität

Gegeben sei zunächst ein *(reales) Problem*, eine entsprechende *mathematische Formulierung* und einen *numerischer Algorithmus*. Mögliche **Fehlerquellen**:

1. *Modellierungsfehler*, z.B. keine Reibung beachtet in $m \cdot \ddot{x} = F = -dx$
2. *Datenfehler*, Meßwerte ungenau
3. *Parameterfehler* (im obigen Beispiel z.B. die Federkonstante)
4. *Rechenfehler*
5. *Abbruchfehler*, z.B. $e^x = \sum_{k=0}^{???} \frac{x^k}{k!}$
6. *Diskretisierungsfehler*, bei $f'(x) \doteq \frac{f(x+h)-f(x)}{h}$
7. $\pi = 3.14159265358979\ldots$

Statt $y = \text{Algo}(x)$ muß man also $y + \Delta y = \text{Algo}(x + \Delta x)$ erwarten.

Frage: Wie reagiert das mathematische Problem auf die Störung in x? Und wie reagiert der numerische Algorithmus?

Definition:

- Ein mathematisches Problem heißt *schlecht konditioniert*, falls kleine Änderungen der Eingabewerte zu großen Änderungen der Ausgabewerte führen (andernfalls *gut konditioniert*).
- Ein numerischer Algorithmus heißt *numerisch instabil*, falls kleine Änderungen des Inputs zu großen Änderungen des Outputs führen (andernfalls: *numerisch stabil*).

Sei $f : \mathbb{R} \to \mathbb{R}$ stetig differenzierbar. Mit $y = f(x)$, $y + \Delta y = f(x + \Delta x)$ ist

$$\begin{aligned} \Delta y &= f(x + \Delta x) - f(x) = f'(\xi) \cdot \Delta x \text{ mit } \xi \in]x, x + \Delta x[\\ &\doteq f'(x) \cdot \Delta x \quad \text{falls } \Delta x \text{ klein} \end{aligned}$$

Der relative Fehler:

$$|\delta_y| = \left|\frac{\Delta y}{y}\right| \doteq \underbrace{\left|\frac{f'(x)}{f(x)} \cdot x\right|}_{\kappa_f(x)} \cdot |\delta_x|$$

Dabei sei $\kappa_f(x)$ die *Konditionalzahl* von f bei x.

Beispiele:

- Sei $f(x) = (x-1)^{\frac{1}{3}} := \text{sign}(x)\,|x-1|^{\frac{1}{3}}$. Dann ist $f'(x) = \frac{1}{3}(x-1)^{-\frac{2}{3}}$, somit ist

$$\kappa_f(x) = \left|\frac{f'(x)}{f(x)} \cdot x\right| = \left|\frac{x}{3(x-1)}\right| \quad \text{falls } x \neq 1$$

Falls also $x \approx 1$ ist, so ist das Problem schlecht konditioniert:

$$\begin{aligned} f(1.0012) &\doteq 0.1006265 \\ f(1.0015) &\doteq 0.114471 \end{aligned}$$

Somit ist $\kappa_f(1.0012) \doteq 278.1 \doteq 10^3$, also ein Verlust von drei Dezimalstellen bei der Berechnung.

- Für $y = f(x) = \frac{1-\cos(x)}{x}$ ist $f'(x) = \frac{x\cdot\sin(x)-1+\cos(x)}{x^2}$. Nun ist

$$\kappa_f(x) = \frac{x \cdot \sin(x) - 1 + \cos(x)}{1 - \cos(x)} \xrightarrow{x\to 0} 1$$

Dieses Problem ist gut konditioniert für $|x|$ klein. Für $x = 0.12345 \cdot 10^{-4}$ ist $\hat{y} = 0.607533 \cdot 10^{-5}$ auf einem Taschenrechner. Das korrekte Ergebnis ist jedoch $y \approx 0.61725 \cdot 10^{-5}$, Fehlerquelle: Auslöschung (da $cos(x) \approx 1$). Betrachte die Reihe

$$\frac{1-\cos(x)}{x} = \frac{1}{x} \cdot \left(1 - \sum_{j=0}^{\infty} (-1)^j \frac{x^{2j}}{(2j)!}\right)$$

Dies ist eine Leibnitz-Reihe, hier kann man eine Abschätzung anwenden: Da $|x|\, y < 10^{-4}$, ist der Genauigkeitsverlust bei der Addition nur der ersten beiden (!) Reihenglieder kleiner als $\frac{x^2}{12} < 10^{-9}$!

- Sei $I_k := \frac{1}{e}\int_0^1 x^k e^x\,dx$, somit ist $I_0 = \frac{1}{e}\int_0^1 e^x\,dx = \frac{e-1}{e} \doteq 0.6321$. Dann gilt:

$$\begin{aligned} e \cdot I_{k+1} &= \int_0^1 x^{k+1} e^x\,dx \\ &= \left[x^{k+1}e^x\right]_0^1 - (k+1)\cdot\int_0^1 x^k e^x\,dx \\ &= e - (k+1)\cdot e \cdot I_k \\ I_{k+1} &= 1 - (k+1)\cdot I_k \end{aligned}$$

Betrachte nun $\Delta k := \hat{I}_k - I_k$:

$$\begin{aligned} \Delta k &= (1 - k\cdot\hat{I}_{k-1}) - (1 - k\cdot I_{k-1}) \\ &= -k\cdot\Delta I_{k-1} \\ &= (-1)^k \cdot k! \cdot \Delta I_0 \end{aligned}$$

Wenn nun beispielsweise $k = 20$ Schritte ausgeführt werden, so ist $k! = 2.4 \cdot 10^{18}$, d.h. auf einem (Standard-)Rechner mit $\varepsilon \approx 0.5 \cdot 10^{-16}$ sind alle Stellen nutzlos/falsch/weg.

Allgemeines Modell: Es liegen Daten $x \in \mathbb{R}^n$ vor mit einem absoluten Fehler $\Delta x = x - \hat{x} \in \mathbb{R}^n$ und einem komponentenweisen relativen Fehler von $\delta_{x_j} = \frac{x_j - \hat{x}_j}{x_j}$. Betrachte eine Funktion $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ mit $y = \varphi(x)$. Mit einer Taylorentwicklung von φ_k an der Stelle $x + \Delta x$ erhält man:

$$\begin{aligned} \varphi_k(x + \Delta x) &= \varphi_k(x) + \frac{\partial \varphi_k}{\partial x_1}(x)\Delta x_1 + \ldots + \frac{\partial \varphi_k}{\partial x_n}\Delta x_n \\ &= \left(\frac{\partial \varphi_k}{\partial x_1} \ \ldots \ \frac{\partial \varphi_k}{\partial x_n} \right) \Delta x \end{aligned}$$

Somit ist

$$\begin{aligned} \hat{y}_k - y_k &= \varphi_k(\hat{x}) - \varphi_k(x) \\ &\doteq \sum_{j=1}^{n} \underbrace{\frac{\partial \varphi_k}{\partial x_j}(x)}_{PPF} \cdot (\hat{x}_j - x_j) \end{aligned}$$

Dabei ist PPF ein Proportionalitätsfaktor, der besagt, wie y_k auf Änderungen von x_j reagiert. Damit ist der absolute Fehler:

$$\Delta y = \begin{pmatrix} y_1 - \hat{y}_1 \\ \vdots \\ y_m - \hat{y}_m \end{pmatrix} \doteq \begin{pmatrix} \frac{\partial \varphi_1}{\partial x_1} & \ldots & \frac{\partial \varphi_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial \varphi_m}{\partial x_1} & \ldots & \frac{\partial \varphi_m}{\partial x_n} \end{pmatrix} \cdot \begin{pmatrix} x_1 - \hat{x}_1 \\ \vdots \\ x_n - \hat{x}_n \end{pmatrix} = \varphi'(x) \cdot \Delta_x$$

Der relative Fehler ist:

$$\delta_{y_k} = \frac{\hat{y}_k - y_k}{y_k} = \sum_{j=1}^{n} \underbrace{\frac{x_j}{\varphi_k(x)} \cdot \frac{\partial \varphi_k}{\partial x_j}(x)}_{Konditionszahlen} \cdot \delta_{x_j}$$

Beispiele:

- Betrachte eine Funktion $y = \varphi(x_1, x_2, x_3) = x_1 + x_2 + x_3$. Dann ist $\varphi : \mathbb{R}^3 \to \mathbb{R}$. Betrachte die Jacobi-Matrix $\varphi'(x) = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \in \mathbb{R}^{1 \times 3}$. Der relative Fehler beträgt also

$$\begin{aligned} \delta_y &= \sum_{j=1}^{3} \frac{x_j}{\varphi(x)} \cdot \frac{\partial \varphi}{\partial x_j}(x) \cdot \delta_{x_j} \\ &= \frac{1}{x_1 + x_2 + x_3} \left(x_1 \cdot \delta_{x_1} + x_2 \cdot \delta_{x_2} + x_3 \cdot \delta_{x_3} \right) \end{aligned}$$

Der Fehler ist *problemspezifisch*, d.h. ist unabhängig von der Wahl des Algorithmus'.

- Quadratische Gleichungen: $x^2 - 2px + q = 0$, betrachte $\varphi : \mathbb{R}^2 \to \mathbb{R}^2$ mit
$$\varphi(p,q) = \begin{pmatrix} p + \sqrt{p^2 - q} \\ p - \sqrt{p^2 - q} \end{pmatrix}$$
Die Ableitungen in der Jacobi-Matrix:
$$\varphi'(p,q) = \begin{pmatrix} 1 - \frac{2p}{2\sqrt{p^2-q}} & \frac{1}{2\sqrt{p^2-q}} \\ 1 + \frac{2p}{2\sqrt{p^2-q}} & -\frac{1}{2\sqrt{p^2-q}} \end{pmatrix} = \frac{1}{2\sqrt{p^2-q}} \begin{pmatrix} 2\sqrt{p^2-q} - 2p & 1 \\ 2\sqrt{p^2-q} + 2p & -1 \end{pmatrix}$$
Damit läßt sich der relative Fehler berechnen als[2]:
$$\begin{aligned} \delta_{y_1(p,q)} &= \frac{p}{\varphi_1(p,q)} \cdot \frac{\partial \varphi_1(p,q)}{\partial p} \cdot \delta_p + \frac{q}{\varphi_1(p,q)} \cdot \frac{\partial \varphi_1(p,q)}{\partial q} \cdot \delta_q \\ &= \frac{-p}{p - \sqrt{p^2-q}} \cdot \frac{p - \sqrt{p^2-q}}{\sqrt{p^2-q}} \cdot \delta_p + \frac{q}{p - \sqrt{p^2-q}} \cdot \frac{1}{2\sqrt{p^2-q}} \cdot \delta_p \\ &= -\frac{p}{\sqrt{p^2-q}} \cdot \delta_p + \frac{p + \sqrt{p^2-q}}{2\sqrt{p^2-q}} \cdot \delta_q \end{aligned}$$
Schlußfolgerungen:
  - Falls nun also $\left|\frac{p}{\sqrt{p^2-q}}\right|$ und $\left|\frac{p+\sqrt{p^2-q}}{2\sqrt{p^2-q}}\right|$ beide ≤ 1 sind, so ist das Problem gut konditioniert, da dann die Input-Fehler im Laufe der Rechnung gemildert werden.
  - Falls nun $q \approx p^2$ ist, so ist das Problem schlecht konditioniert.
  - Für $q \approx 0$ ist der Algorithmus numerisch instabil, da es bei $p \pm \sqrt{p^2 - q}$ zu Auslöschung kommt.

1.5 Fehlerfortpflanzung (nur exemplarisch)

Idee: Aus endlich vielen Eingabedaten x_j werden endlich vielen Ausgabedaten y_k durch $\varphi : \mathbb{R}^n \to \mathbb{R}^m$. Dabei ist φ eine Funktion, die den Algorithmus beschreibt. Zerlege φ in elementare Schritte (und analysiere diese), d.h. $\varphi = \varphi^{(s)} \circ \ldots \circ \varphi^{(1)} \circ \varphi^{(0)}$.

Beispiele:

[2] Nebenrechnung: $\frac{q}{p - \sqrt{p^2-q}} = \frac{q(p+\sqrt{p^2-q})}{p^2-(p^2-q)} = p + \sqrt{p^2-q}$

- Für $\varphi(x_1, x_2, x_3) = x_1 + x_2 + x_3$ sei die Zerlegung in Elementarschritte:

$$\varphi^{(0)}(x_1, x_2, x_3) = \begin{pmatrix} x_1 + x_2 \\ x_3 \end{pmatrix} \text{ und } \varphi^{(1)}(x_1, x_2) = x_1 + x_2 \quad \text{(Alg1)}$$

$$\varphi^{(0)}(x_1, x_2, x_3) = \begin{pmatrix} x_1 \\ x_2 + x_3 \end{pmatrix} \text{ und } \varphi^{(1)}(x_1, x_2) = x_1 + x_2 \quad \text{(Alg2)}$$

- Für $\varphi(x_1, x_2) = x_1^2 - x_2^2$ sei die Zerlegung in Elementarschritte:

$$\varphi^{(0)}(x_1, x_2) = \begin{pmatrix} x_1^2 \\ x_2^2 \end{pmatrix} \text{ und } \varphi^{(1)}(x_1, x_2) = x_1 - x_2 \quad \text{(Alg3)}$$

$$\varphi^{(0)}(x_1, x_2) = \begin{pmatrix} x_1 + x_2 \\ x_1 - x_2 \end{pmatrix} \text{ und } \varphi^{(1)}(x_1, x_2) = x_1 \cdot x_2 \quad \text{(Alg4)}$$

2 Lineare Gleichungssysteme

Motivation[3]: Finde ein Polynom p zweiten Grades mit $p(x_j) = b_j$ für $x_0 = -1$, $x_1 = 0$, $x_2 = 1$, $b_0 = b_1 = 0$, $b_2 = 2.25$.

Ansatz: $p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2$, also

$$\begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix}$$

Allgemein: Bestimme x in $Ax = b$ mit $A \in \mathbb{R}^{n \times n}$ und $x, b \in \mathbb{R}^n$. Lösungsansätze:

- Gaußsches Eliminationsverfahren
- Cramersche Regel (Problem: Anzahl der Rechenoperationen bei Determinanten ist viel zu groß!)
- direkte Verfahren[4] (später mehr)
- iterative Verfahren

Kriterien: benötigter Speicherplatz, Rechenzeiten, Stabilität, Kondition: $b \to b + \Delta_b \Rightarrow x + \Delta_x$, wie groß ist $\|\Delta_x\|$?

Satz: Sei $A \in \mathbb{R}^{n \times n}, x, b \in \mathbb{R}^n$. Folgende Aussagen sind äquivalent:

1. $Ax = b$ ist eindeutig lösbar für alle $b \in \mathbb{R}^n$
2. A ist invertierbar, d.h. A^{-1} existiert.
3. A ist nicht singulär
4. $\det(A) \neq 0$

2.1 LR-Zerlegung, Gaußsches Eleminationsverfahren

Beispiel:

$$A = \begin{pmatrix} 1 & -1 & 1 & 0 \\ -1 & 1 & 0 & 1 \\ 1 & 2 & 3 & -1 \\ -2 & -2 & 0 & 2 \end{pmatrix} \text{ und } b = \begin{pmatrix} 1 \\ 1 \\ 2 \\ -2 \end{pmatrix}$$

[3] „Man strömt das an mit einem fluidalen Flow…“

[4] „Wenn ich mich einmal verrechne, dann ist der Drops gelutscht!“

Strategie ist nun $Ax = b$ durch elementare Umformungen zu $Rx = c$ mit R obere Dreiecksmatrix:

$$R = \begin{pmatrix} 1 & -1 & 1 & 0 \\ & 2 & 0 & -1 \\ & & 1 & 1 \\ & & & -2 \end{pmatrix} \text{ und } c = \begin{pmatrix} 1 \\ 1 \\ 2 \\ -2 \end{pmatrix}$$

Das Vertauschen von Zeilen kann (theoretisch) durch Permutationsmatrix P beschrieben werden: $Ax = b$ und $PAx = Pb$ mit

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Nochmal für $PAx = Pb$:

				1	−1	1	0	1
				1	1	1	−1	2
				−1	1	0	1	1
				−2	−2	0	2	−2
1	0	0	0	1	−1	1	0	1
−1	1	0	0	0	2	0	−1	1
1	0	1	0	0	0	1	1	2
2	0	0	1	0	−4	2	2	0
1	0	0	0	1	−1	1	0	1
0	1	0	0	0	2	0	−1	1
0	0	1	0	0	0	1	1	2
0	2	0	1	0	0	2	0	2
1	0	0	0	1	−1	1	0	1
0	1	0	0	0	2	0	−1	1
0	0	1	0	0	0	1	1	2
0	0	−2	1	0	0	0	−2	−2

Dabei ist die Struktur obiger Tabelle folgende:

	PA	Pb
M_1	M_1PA	M_1Pb
M_2	M_2M_1PA	M_2M_1Pb
M_3	$M_3M_2M_1PA$	$M_3M_2M_1Pb$

Somit ist $R = M_3 M_2 M_1 PA$. Betrachte noch einmal genauer M_1:

$$M_1 = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ 2 & & & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & & & \\ & 1 & & \\ 1 & & 1 & \\ & & & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & & & \\ -1 & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix}$$

Dabei ist die Berechnung des Inversen für diese Matrizen nicht schwer, da die Einträge außerhalb der Hauptdiagonalen direkt den Zeilenoperationen entsprechen und diese rückgängig gemacht werden können, also:

$$M_1^{-1} = \begin{pmatrix} 1 & & & \\ 1 & 1 & & \\ -1 & & 1 & \\ -2 & & & 1 \end{pmatrix}$$

Sei nun L definiert mit $PA = L \cdot R$:

$$L = M_1^{-1} \cdot M_2^{-1} \cdot M_3^{-1} = \begin{pmatrix} 1 & & & \\ 1 & 1 & & \\ -1 & 0 & 1 & \\ -2 & -2 & 2 & 1 \end{pmatrix}$$

Das Gaußsche Eleminationsverfahren überführt $PAx = Pb$ in die Form $LRx = c$, eine so genannte *LR-Zerlegung von A* (mit L und R nicht singulär und $\det(L) = 1$). Sei nun $y = Rx$, löse zunächst $Ly = c$ und erhalte y („vorwärts lösen“, d.h. mit y_1 anfangen usw.), löse und erhalte danach in einer „Rückwärtsrechnung“ x aus $Rx = y$. Im Beispiel (erster Schritt vorwärts, zweiter Schritt rückwärts):

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -2 & -2 & 2 & 1 \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 1 \\ -2 \end{pmatrix} \implies \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 2 \\ -2 \end{pmatrix}$$

$$\begin{pmatrix} 1 & -1 & 1 & 0 \\ 0 & 2 & 0 & -1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & -2 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 2 \\ -2 \end{pmatrix} \implies \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

Definition: Die Matrix $A \in \mathbb{R}^{n \times n}$ besitzt eine LR-Zerlegung, falls es eine linke untere Dreiecksmatrix L mit Einsen in der Diagonalen und eine rechte obere Dreiecksmatrix R gibt mit $A = LR$ (wobei L, R invertierbar sind).

Satz: $A \in \mathbb{R}^{n\times n}$ besitzt eine LR-Zerlegung genau dann, wenn $\det(A_j) \neq 0 \; \forall \, j = 1, \ldots, n$ (wobei $A_j = A(1:j, 1:j)$) ist.

Beweis: Mit $A = L \cdot R$ ist $A_j = L_j \cdot R_j$ für $j = 1, \ldots, n$ und obere linke $(j \times j)$-Matrizen A_j, L_j, R_j.

„⇒“ A besitze eine LR-Zerlegung, also sind A, L, R invertierbar und damit auch L_j, R_j - und somit auch $A_j = L_j \cdot R_j$ invertierbar, d.h. die Determinante ist von null verschieden.

„⇐“ Über Induktion: Für $n = 1$ ist $A = (a_{11}) = (1) \cdot (a_{11})$ mit $L = (1)$ und $R = (a_{11})$, jeweils invertierbar.

Induktionsschritt: Sei A eine $n \times n$-Matrix, nach Induktionsannahme ist $A_{n-1} = L_{n-1} \cdot R_{n-1}$ mit invertierbaren L_{n-1}, R_{n-1}

$$A = \begin{pmatrix} A_{n-1} & a_{:,n} \\ a_{n,:} & a_{nn} \end{pmatrix} = \begin{pmatrix} L_{n-1} & 0 \\ l & 1 \end{pmatrix} \cdot \begin{pmatrix} R_{n-1} & r \\ 0 & \varrho \end{pmatrix} = \begin{pmatrix} L_{n-1}R_{n-1} & L_{n-1}r \\ lR_{n-1} & l \cdot r + \varrho \end{pmatrix}$$

Es bleibt:

$$\begin{aligned} a_{:,n} = L_{n-1} \cdot r \quad &\Leftrightarrow \quad r = L_{n-1}^{-1} \cdot a_{:,n} \\ a_{n,:} = l \cdot R_{n-1} \quad &\Leftrightarrow \quad a_{n,:} \cdot R_{n-1}^{-1} \\ \text{und } a_{nn} = l \cdot r + \varrho \quad &\Leftrightarrow \quad \varrho = a_{nn} - l \cdot r \end{aligned}$$

Somit ist $A = L \cdot R$, mit invertierbarem $A_n = A$ folgt L, R invertierbar. □

Korollar:

1. $A \in \mathbb{R}^{n\times n}$ strikt diagonaldominant $\Rightarrow$ LR-Zerlegung existiert
2. $A \in \mathbb{R}^{n\times n}$ positiv definit $\Rightarrow$ LR-Zerlegung existiert

Bemerkungen:

- falls Voraussetzungen von obigem Satz erfüllt sind, ist das Gaußsche Eleminationsverfahren durchführbar
- der Rechenaufwand ist $\frac{1}{3}n^3 + O(n^2)$
- effiziente Speichernutzung bei der LR-Zerlegung möglich durch Speicherung von R und L (ohne Hauptdiagonale) im Speicherplatz von A
- effizienteste Methode, A^{-1} zu berechnen: löse $Ax_j = e_j$ für $j = 1, \ldots, n$ und setze $A^{-1} = [x_1, \ldots, x_n]$ - schon ab $n = 1$ sehr dumm! [5]

[5] „Wenn Ihr als Numeriker kommt und schreibt A^{-1} an die Tafel, seid Ihr draußen aus der Prüfung!“

2.2 Die LDL^T-Zerlegung

Satz: Die Matrix $A \in \mathbb{R}^{n\times n}$ besitze eine LR-Zerlegung, dann gibt es untere Dreiecksmatrizen mit Haupdiagonalen 1 L, M und eine Diagonal-Matrix D mit $A = L \cdot D \cdot M^T$.

Beweis: Sei $A = LR$ die LR-Zerlegung von A, D Diagonalmatrix mit Diagonalen $r_{11}, r_{22}, \ldots, r_{nn}$ und $M := R^T D^{-1}$ untere Dreiecksmatrix mit Hauptdiagonalen 1. Also: L, M, D wie gefordert, $A = L \cdot D \cdot M^T$. □

Satz: Sei $A = A^T$ und $A = L \cdot D \cdot M^T$ mit L, M, D wie eben mit D invertierbar. Dann ist $L = M$.

Beweis: Es sei $B := M^{-1}AM^{-T} = B^T$. Zudem ist $B = M^{-1} \cdot L \cdot D \cdot M^T \cdot M^{-T} = M^{-1} \cdot L \cdot D$, also ist B untere Dreiecksmatrix, aber auch symmetrisch, somit eine Diagonalmatrix.

Sei nun $F := B \cdot D^{-1} = M^{-1} \cdot L$, also $L = M \cdot F$. Da aber L, M Matrizen mit 1 auf der Hauptdiagonalen sind, ist $F = E$ die Einheitsmatrix. Damit folgt jedoch: $L = M$. □

Methode zum **Aufbau einer LDL^T-Zerlegung**: Die Existenz einer LDL^T-Zerlegung vorausgesetzt gilt $A = LDL^T$. Dann ist

$$\begin{aligned} A_{jk} &= \sum_{l=1}^{n}\sum_{\mu=1}^{n} L_{jl} \cdot D_{l\mu} \cdot (L^T)_{\mu k} \\ &= \sum_{l=1}^{n} L_{jl} \cdot D_{ll} \cdot L_{kl} \\ &= \sum_{l=1}^{k} L_{jl} \cdot D_{ll} \cdot L_{kl} \\ &= \sum_{l=1}^{k-1} L_{jl} \cdot D_{ll} \cdot L_{kl} + L_{jk} \cdot D_{kk} \end{aligned}$$

Angenommen, L_{pq}, D_{qq} sind bekannt für $q < k$ und alle p. Dann ist $D_{kk} = A_{kk} - \sum_{l=1}^{k-1} L_{kl}^2 D_{ll}$ und

$$L_{jk} = \begin{cases} 0 & \text{für} \quad j = 1, \ldots, k-1 \\ 1 & \text{für} \quad j = k \\ \frac{A_{jk} - \sum_{l=1}^{k-1} L_{jl} D_{ll} L_{kl}}{D_{kk}} & \text{für} \quad j = k+1, \ldots, n \end{cases}$$

2.2.1 Eigenschaften von Matrizen, Norm

Definitionen: Eine Matrix heißt

- symmetrisch: $A = A^T$
- positiv definit: $x^T Ax > 0$ für alle $x \neq 0$
- diagonaldominant: $|a_{jj}| \geq \sum_{\substack{k=1 \\ k \neq j}}^{n} |a_{jk}|$

Bemerkung: Eine Matrix ist symmetrisch und positiv definit genau dann, wenn die Matrix symmetrisch ist und alle Eigenwerte positiv sind.

Definition: Sei V ein $\mathbb{R}$-Vektorraum, $\|\cdot\|$ ist eine *Norm*, falls gilt:

- nicht negativ: $\|x\| \geq 0$ für alle $x \in V$ und $\|x\| = 0 \Leftrightarrow x = 0$
- homogen: $\|\alpha x\| = |A| \, \|x\|$ für alle $x \in V$ und $\alpha \in \mathbb{R}$
- Dreiecksungleichung: $\|x + y\| \leq \|x\| + \|y\|$ für alle $x, y \in V$

Eine *verträgliche* oder *zugeordnete Matrixnorm* $\|\cdot\|_V$ hat die Eigenschaft $\|Ax\|_V \leq \|A\|_M \, \|x\|_V$.

2.3 Cholesky-Zerlegung

Die Cholesky-Zerlegung ist für einen wichtigen Spezialfall eines Linearen Gleichungssystems eine schnellere Lösung.

Satz: Sei A symmetrisch und positiv definit. Dann gibt es genau eine untere Dreiecksmatrix C mit $C_{jj} > 0$ für $j = 1, \ldots, n$ und $A = CC^T$.

Beweis: A besitzt LDL^T-Zerlegung mit $d(j) > 0$ (siehe Übung). Setze $C := L\tilde{D}$ mit $\tilde{d}_j = \sqrt{d_j}$. Dann ist $A = CC^T$ und $c_{jj} > 0$.

Zur Eindeutigkeit: Angenommen, es existiert eine untere Dreiecksmatrix G mit $A = GG^T$. Dann ist $CC^T = GG^T$, also $G^{-1}C = G^T C^{-T}$. Dabei ist die linke Seite eine untere, die rechte eine obere Dreiecksmatrix. Damit muß $G^{-1}C$ eine Diagonalmatrix sein und $G^{-1}C = C^{-1}G = (G^{-1}C)^{-1} =: F$ (mit F Diagonalmatrix). Somit ist $F_{jj} = \pm 1$, da aber C_{jj} und G_{jj} jeweils positiv sind, folgt $F_{jj} = 1$, also $F = E$ und somit auch $G = C$.

2.4 Fehlerabschätzung

Sei z.B.

$$A = \begin{pmatrix} 1+\varepsilon & 1 \\ 1 & 1-\varepsilon \end{pmatrix} \quad b = \begin{pmatrix} 2+\varepsilon \\ 2-\varepsilon \end{pmatrix}$$

Dann ist $\det(A) = -\varepsilon^2 \neq 0$, falls $\varepsilon \neq 0$. Die Lösung ist damit eindeutig bestimmt als $x = A^{-1}b = \binom{1}{1}$. Angenommen, $\varepsilon = 10^{-2}$ und die rechte Seite ist leicht gestört, d.h. $\tilde{b} = \binom{2}{2}$. Dann ist $\tilde{x} = A^{-1}\tilde{b} = \binom{200}{-200}$! Dies ist relativ weit weg von der Lösung des ungestörten Problems!

Falle: Kleine Änderungen der Eingabedaten können große Änderungen der Ausgabedaten bewirken.

2.4.1 Normen und Matrixnomen

Wir betrachten ein realistischeres Modell für $Ax = b$ mit gestörten Daten: $(A+\Delta_A)(x+\Delta_x) = (b+\Delta_b)$. Betrachte in einem $\mathbb{R}$-Vektorraum V die *Norm* $\|\cdot\| : V \to \mathbb{R}$. Dabei ist die *induzierte Norm*: $\|A\|_M := \sup\{|Ax|_V \mid \|x\|_V = 1\}$.

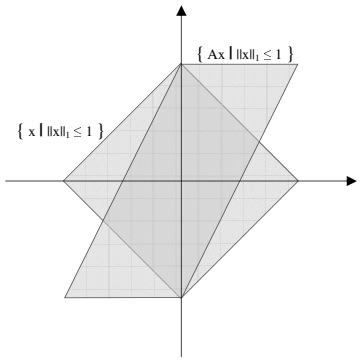
Definition: Im Vektorraum $\mathbb{R}^n$ seien folgende Normen definiert:

$$\begin{aligned} \|x\|_1 &= \sum_{j=1}^{n} |x_j| \\ \|x\|_2 &= \sqrt{\sum_{j=1}^{n} x_j^2} \\ \|x\|_\infty &= \max\{|x_j| \mid j = 1, \ldots, n\} \end{aligned}$$

Entsprechend für Matrizen $A \in \mathbb{R}^{n\times n}$:

$$\begin{aligned} \textit{Spaltensummennorm} \quad \|A\|_1 &:= \sup\{\|Ax\|_1 \mid \|x\|_1 = 1\} \\ \textit{Spektralnorm} \quad \|A\|_2 &:= \sup\{\|Ax\|_2 \mid \|x\|_2 = 1\} \\ \textit{Zeilensummennorm} \quad \|A\|_\infty &:= \max\left\{ \sum_{k=1}^{n} |a_{jk}| \;\middle|\; j = 1, \ldots, n \right\} \end{aligned}$$

Beispiel: Für die Matrix $A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ ist $\|A\|_1 = 2$, da $\|Ax\|_1 \leq 2$ für alle x mit $\|x\|_1 \leq 1$:



Betrachte die einzelnen Normen genauer:

Für die Spaltensummennorm gilt:

$$\|A\|_1 = \max\left\{ \sum_{j=1}^{n} |a_{jk}| \;\middle|\; k = 1, \ldots, n \right\}$$

Beweis: Sei $S := \max\left\{ \sum_{j=1}^{n} |a_{jk}| \;\middle|\; k = 1, \ldots, n \right\}$.

„$\leq$“ Wähle $y \in \mathbb{R}^n$ mit $\|y\|_1 = 1$ und $\|A\|_1 = \|Ay\|_1$, es ist

$$\begin{aligned}
\|A\|_1 &= \|Ay\|_1 \\
&= \sum_{j=1}^{n} |(A \cdot y)_j| = \sum_{j=1}^{n} \left| \sum_{k=1}^{n} a_{jk} \cdot y_{jk} \right| \\
&\leq \sum_{j=1}^{n} \sum_{k=1}^{n} |a_{jk}| \cdot |y_{jk}| = \sum_{k=1}^{n} \sum_{j=1}^{n} |a_{jk}| \cdot |y_{jk}| \\
&= \sum_{k=1}^{n} \left(|a_{jk}| \cdot \sum_{j=1}^{n} \cdot |y_{jk}| \right) \leq \sum_{k=1}^{n} (|a_{jk}| \cdot S) = S
\end{aligned}$$

„$\geq$“ Sei k_0 ein Index, so daß $S = \sum_{j=1}^{n} |a_{jk_0}|$. Sei e_{k_0} entsprechend der k_0-te Einheitsvektor. Es gilt: $\|e_{k_0}\|_1 = 1$ und

$$\|Ae_{k_0}\|_1 = \sum_{j=1}^{n} |a_{jk_0}| = S \leq \|A\|_1$$

Damit ist $\|A\|_1 \leq S \leq \|A\|_1$, also $\|A\|_1 = S$.

Für die Spektralnorm gilt:

$$\|A\|_2 = \sqrt{\varrho(A^T A)} \text{ mit } \varrho(A) := \max\{|\lambda| \mid \lambda \text{ Eigenwert von } A\}$$

Zunächst als Hilfsmittel die **Hauptachsentransformation:**

Sei $A \in \mathbb{R}^{n\times n}$ mit $A = A^T$, dann gibt es $\Lambda, V \in \mathbb{R}^{n\times n}$ mit $A \cdot V = V \cdot \Lambda$ und $V^T V = E$ sowie Λ Diagonalmatrix. Beispielsweise sei $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, dann ist

$$A \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 3 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad A \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 1 \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

$$A \cdot \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$$

Zeige wieder $\|A\|_2^2 \leq \varrho(A^T A)$ und $\|A\|_2^2 \geq \varrho(A^T A)$

„$\leq$“ Wähle $y \in \mathbb{R}^n$ mit $\|y\|_1 = 1$ und $\|A\|_2 = \|Ay\|_2$. Nun kann $A^T A$ reell diagonalisiert werden: $A^T AV = V \cdot \Lambda$, wobei $\lambda_j \geq 0$[6]. Sei nun $z := V^T y \Leftrightarrow y = Vz$. Dann ist

$$\begin{aligned} \|A\|_2^2 &= \|Ay\|_2^2 = y^T A^T Ay = y^T (A^T AV) z = y^T V \Lambda z = z^T \Lambda z \\ &= \sum_{j=1}^{n} \lambda_j z_j^2 \leq \underbrace{\left(\sum_{j=1}^{n} z_j^2\right)}_{z^T z} \cdot \max\{\lambda_p \mid p = 1, \ldots, n\} \\ &= \max\{\lambda_p \mid p = 1, \ldots, n\} = \varrho(A^T A) \end{aligned}$$

„$\geq$“ Ist v_0 ein Eigenvektor zum Eigenwert $\varrho(A^T A)$ von $A^T A$ mit $\|v_0\|_2 = 1$, dann gilt:

$$\|Av_0\|_2^2 = v_0^T A^T Av_0 = \varrho(A^T A) \cdot v_0^T v_0 = \varrho(A^T A)$$

2.4.2 Fehlerabschätzung zur Lösung linearer Gleichungssysteme

Lemma: (Störsatz) Sei A nicht singulär, $Ax = b$, $b \neq 0$ und $A(x + \Delta_x) = b + \Delta_b$. Dann gilt:

$$\frac{\|\Delta_x\|}{\|x\|} \leq \|A\| \cdot \left\|A^{-1}\right\| \cdot \frac{\|\Delta_b\|}{\|b\|}$$

[6] Es ist $A^T Av_j = \lambda_j v_j$, daraus folgt $v_k^T A^T Av_j = \lambda_j v_k^T v_j = \begin{cases} \lambda_j & \text{falls } k = j \\ 0 & \text{sonst} \end{cases}$. Es ist also $\lambda_j = v_j^T A^T Av_j = \|Av_j\|_2^2 \geq 0$

Beweis: Falls $\Delta_b = 0$, so ist die Behauptung wegen $\Delta_x = 0$ richtig. Andernfalls: Aus $Ax = b$ folgt

$$\begin{aligned} & Ax + A \cdot \Delta_x = b + \Delta_b \\ \Rightarrow \quad & A\Delta_x = \Delta_b \\ \Rightarrow \quad & \Delta_x = A^{-1} \cdot \Delta_b \\ \Rightarrow \quad & \|\Delta_x\| = \left\|A^{-1} \cdot \Delta_b\right\| \leq \left\|A^{-1}\right\| \cdot \|\Delta_b\| \end{aligned}$$

Zudem ist $b = Ax \Rightarrow \|b\| \leq \|A\| \cdot \|x\|$, es folgt:

$$\frac{\|\Delta_x\|}{\|x\|} \cdot \frac{\|b\|}{\|\Delta_b\|} \leq \frac{\|A^{-1}\| \cdot \|\Delta_b\|}{\|x\|} \cdot \frac{\|A\|\,\|x\|}{\|\Delta_b\|} = \|A\| \cdot \left\|A^{-1}\right\|$$

Beispiel: Verwende die ∞-Norm, d.h. $\|x\| = \max\{|x_j| \mid j = 1, \ldots, n\}$ und $\|A\| = \max\{\sum_{k=1}^{n} |a_{jk}| \mid j = 1, \ldots, n\}$. Sei nun wieder

$$A = \begin{pmatrix} 1+\varepsilon & 1 \\ 1 & 1-\varepsilon \end{pmatrix} \quad b = \begin{pmatrix} 2+\varepsilon \\ 2-\varepsilon \end{pmatrix} \quad x = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \Delta_b = \begin{pmatrix} -\varepsilon \\ \varepsilon \end{pmatrix} \quad \Delta_x = \begin{pmatrix} 199 \\ -201 \end{pmatrix}$$

Dann ist $\|A\| = 2 + |\varepsilon|$, A^{-1} ist $\frac{1}{\det(A)} \cdot \begin{pmatrix} 1-\varepsilon & -1 \\ -1 & 1+\varepsilon \end{pmatrix}$, also $\|A^{-1}\| = \frac{2+|\varepsilon|}{\varepsilon^2}$, z.B. für $\varepsilon = 10^2$ ist $\|A^{-1}\| = 2.01 \cdot 10^{-4}$ und $\frac{\|\Delta_x\|}{x} = \frac{201}{1} \leq \|A\| \cdot \|A^{-1}\| \cdot \frac{\|\Delta_b\|}{b} = 40401 \cdot \frac{0.01}{2.01}$.

Definition: Für $A \in \mathbb{R}^{n \in n}$ nicht singulär sei $\text{cond}(A) := \|A\| \cdot \|A^{-1}\|$ die *Konditionszahl* von A (bezüglich einer Matrixnorm).

Lemma: Sei $B \in \mathbb{R}^{n \times n}$ mit $\|B\| < 1$, dann gilt:

1. $(E - B)$ nicht singulär
2. $\|(E - B)^{-1}\| \leq \frac{1}{1 - \|B\|}$
3. $(E - B)^{-1} = \sum_{\nu=0}^{\infty} B^{\nu}$

Beweis:

1. Angenommen, $E - B$ sei singulär. Dann existiert $0 \neq x \in \mathbb{R}^n$ mit $(E - B)x = 0$, also $x = Ex = Bx$. Daraus folgt:

   $$\|x\| = \|Bx\| \leq \|B\| \cdot \|x\| < \|x\|$$

   Dies ist offensichtlich ein Widerspruch.

3. Betrachte für alle N:

$$F_N = E - \left((E-B) \cdot \left(\sum_{\nu=0}^{N} B^\nu \right) \right) = E - \left(\sum_{\nu=0}^{N} B^\nu \right) + \left(\sum_{\nu=1}^{N+1} B^\nu \right) = B^{N+1}$$

Dann gilt: $\|F_N\| = \left\|B^{N+1}\right\| \leq \|B\|^{N+1}$, damit ist aber $\lim_{N\to\infty} \|F_N\| = 0$, da $\|B\| < 1$. Damit folgt $\lim_{N\to\infty} F_N = 0 \in \mathbb{R}^{n\times n}$, damit gilt die Behauptung.

2. Es gilt:

$$\left\|(E-B)^{-1}\right\| = \left\| \sum_{\nu=0}^{\infty} B^\nu \right\| \leq \sum_{\nu=0}^{\infty} \|B\|^\nu = \frac{1}{1-\|B\|}$$

Korollar: Sei $A \in \mathbb{R}^{n\times n}$ nicht singulär, $\Delta_A \in \mathbb{R}^{n\times n}$ mit $\|A^{-1}\| \cdot \|\Delta_A\| < 1$, dann gilt:

1. $(A + \Delta_A)$ nicht singulär
2. $\|(A+\Delta_A)^{-1}\| \leq \frac{\left\|A^{-1}\right\|}{1-\|A^{-1}\|\cdot\|\Delta A\|} = \frac{\left\|A^{-1}\right\|}{1-\mathrm{cond}(A)\cdot\frac{\|\Delta A\|}{\|A\|}}$

Beweis: Sei $A + \Delta A = A(E-B)$, also $B := -A^{-1} \cdot \Delta A$. Dann ist $\|B\| = \|A^{-1} \cdot \Delta A\| \leq \|A^{-1}\| \cdot \|\Delta A\| < 1$, d.h. obiges Lemma liefert: $\|(E-B)^{-1}\| \leq \frac{1}{1-\|B\|}$ und $(A+\Delta A)^{-1}$ existiert und es gilt

$$\begin{aligned} (A+\Delta A)^{-1} &= (A\cdot(E-B))^{-1} = (E-B)^{-1}\cdot A^{-1} \\ \left\|(A+\Delta A)^{-1}\right\| &= \left\|(E-B)^{-1}\cdot A^{-1}\right\| \leq \left\|(E-B)^{-1}\right\| \cdot \left\|A^{-1}\right\| \\ &\leq \frac{1}{1-\|B\|} \cdot \left\|A^{-1}\right\| \leq \frac{\|A^{-1}\|}{1-\|A^{-1}\|\cdot\|\Delta A\|} \end{aligned}$$

Satz: Seien $A, \Delta A \in \mathbb{R}^{n\times n}$ mit A nicht singulär. Sei $\|A^{-1}\| \cdot \|\Delta_A\| < 1$. Seien $b, \Delta b \in \mathbb{R}^n$ mit $b \neq 0$. Sei weiter $x \in \mathbb{R}^n$ mit $Ax = b$, sei $\Delta x \in \mathbb{R}^n$ mit $(A+\Delta A)(x+\Delta x) = b + \Delta b$. Dann gilt:

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\mathrm{cond}(A)}{1-\mathrm{cond}(A)\frac{\|\Delta A\|}{\|A\|}} \cdot \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|} \right)$$

Faustregel: Eine Kondition $\mathrm{cond}(A) = 10^k$ führt zum Verlust von k Dezimalstellen im Ergebnis.

Beweis: Nach Korollar ist $(A + \Delta A)$ nicht singulär: $(A + \Delta A)(c + \Delta c) = (b + \Delta b)$ und $(A + \Delta A) \cdot \Delta x = b + \Delta b - (A + \Delta A) \cdot x = \Delta b - \Delta A \cdot x$, d.h. $\Delta x = (A + \Delta A)^{-1}(\Delta b - \Delta A x)$ Damit ist

$$\begin{aligned} \|\Delta x\| &\leq \left\|(A + \Delta A)^{-1}\right\| \cdot (\|\Delta b\| + \|\Delta A x\|) \\ &\leq \frac{\|A^{-1}\|}{1 - \mathrm{cond}(A) \cdot \frac{\|\Delta A\|}{\|A\|}} \cdot (\|\Delta b\| + \|\Delta A\| \cdot \|\Delta x\|) \\ \Longrightarrow \frac{\|\Delta x\|}{\|x\|} &\leq \frac{\|A^{-1}\| \, \|A\|}{1 - \mathrm{cond}(A) \cdot \frac{\|\Delta A\|}{\|A\|}} \cdot \left(\frac{\|\Delta b\|}{\|A\| \, \|x\|} + \frac{\|\Delta A\|}{\|A\|} \right) \end{aligned}$$

Die Behauptung folgt dann mit $\|b\| = \|Ax\| \leq \|A\| \, \|x\|$.

2.4.3 Stabilität der LR-Zerlegung

Sei $M := \{\text{Maschinenzahlen}\}$.

Satz: Für $A \in M^{n \times n}$ existiere eine LR-Zerlegung. Die berechneten Faktoren $\tilde{L}$ und $\tilde{R}$ genügen einer Gleichung $\tilde{L}\tilde{R} = A + H$, wobei

$$|H| \leq 3 \cdot \varepsilon \cdot (n-1) \cdot \left(|A| + \left|\tilde{L}\right| \left|\tilde{R}\right| \right) \text{ mit } |X| = (|x_{j,k}|)_{j,k=1}^{n}$$

Beweis: Per Induktion über n:

- Induktionsanfang: Für $n = 1$ ist $A = (a_{11}) \in M^{1 \times 1}$, dann ist $\tilde{L} = (1)$ under $\tilde{R} = (a_{11})$, damit ist $\tilde{L} \cdot \tilde{R} = A$, also $H = 0 \in \mathbb{R}^{1 \times 1}$.
- Induktionsschritt: Schreibe $A \in M^{n \times n}$ zerlegt in α, $w^T \in M^{n-1 \times 1}$, $v \in M^{1 \times n-1}$ und $B \in M^{n-1 \times n-1}$:

  $$A = \begin{pmatrix} \alpha & w^T \\ v & B \end{pmatrix} \xrightarrow{\text{Gauß}} \begin{pmatrix} \alpha & w^T \\ 0 & A_1 \end{pmatrix}$$

  Beschreibe den Schritt im Gaußschen Eliminationsverfahren mittels $z := \frac{1}{\alpha} v$, $A_1 = B - z w^T$. Numerisch ist dies:

  $$\begin{aligned} \tilde{z} = \frac{1}{\alpha} v + f \quad &\text{mit} \quad |f| \leq \varepsilon \frac{1}{|\alpha|} |v| \\ \tilde{A}_1 = B - \tilde{t} \cdot w^T + F \quad &\text{mit} \quad |F| \leq 2\varepsilon (|B| + |\tilde{z}| \left|w^T\right|) \end{aligned}$$

  Die Induktionsannahme lautet dann, daß eine LR-Zerlegung von $\tilde{A}_1$ existiert mit $\tilde{L}_1 \tilde{R}_1 = \tilde{A}_1 + H_1$, für die gilt:

  $$|H| \leq 3 \cdot \varepsilon \cdot (n-2) \cdot \left(\left|\tilde{A}_1\right| + \left|\tilde{L}_1\right| \left|\tilde{R}_1\right| \right)$$

Setzte nun

$$\tilde{L} = \begin{pmatrix} 1 & 0 \\ \tilde{z} & \tilde{L}_1 \end{pmatrix} \text{ und } \tilde{R} = \begin{pmatrix} \alpha & w^T \\ 0 & \tilde{R}_1 \end{pmatrix}$$

Mit $\tilde{L}_1 \cdot \tilde{R}_1 + \tilde{z}w^T = \tilde{A}_1 + H_1 + \tilde{z}w^T = B + F + H_1$ ist:

$$\begin{aligned} \tilde{L} \cdot \tilde{R} &= \begin{pmatrix} \alpha & w^T \\ \alpha \cdot \tilde{z} & \tilde{L}_1 \cdot \tilde{R}_1 + \tilde{z}w^T \end{pmatrix} \\ &= \begin{pmatrix} \alpha & w^T \\ v + \alpha \cdot f & B + F + H_1 \end{pmatrix} \\ &= A + \begin{pmatrix} 0 & 0^T \\ \alpha \cdot f & F + H_1 \end{pmatrix} \end{aligned}$$

Betrachtet man nun die einzelnen Einträge, so ist $|\alpha f| \leq 3 \cdot \varepsilon \cdot (n-1) \cdot |v|$ und

$$\begin{aligned} |F + H_1| &\leq |H_1| + |F| \\ &\leq 3 \cdot \varepsilon \cdot (n-2) \cdot (\left|\tilde{A}_1\right| + \left|\tilde{L}_1\right| \cdot \left|\tilde{R}_1\right|) + |F| \\ (\star) &\leq 3 \cdot \varepsilon \cdot (n-2) \left((1+2\varepsilon) \left(|B| + |\tilde{z}| \left|w^T\right| + \left|\tilde{L}_1\right| \cdot \left|\tilde{R}_1\right|\right)\right) + 2\varepsilon(|B| + |\tilde{z}| \left|w^T\right|) \\ &\dot{\leq} 3 \cdot \varepsilon \cdot (n-1)(|B| + |\tilde{z}| \left|w^T\right| + \left|\tilde{L}_1\right| \left|\tilde{R}_1\right|) \end{aligned}$$

Dabei folgt $(\star)$ mit

$$\left|\tilde{A}_1\right| \leq |B| + |\tilde{z}| \left|w^T\right| + |F| \leq \left(|B| + |\tilde{z}| \left|w^T\right|\right)(1 + 2\varepsilon)$$

Zusammen ergibt sich

$$|H| \leq 3 \cdot \varepsilon \cdot (n-1) \cdot \left(\begin{bmatrix} |\alpha| & \left|w^T\right| \\ |v| & |B| \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ |\tilde{z}| & \left|\tilde{L}_1\right| \end{bmatrix} \cdot \begin{bmatrix} |\alpha| & \left|w^T\right| \\ 0 & \left|\tilde{R}_1\right| \end{bmatrix} \right)$$

2.5 Lineare Ausgleichsprobleme

Modell: $y = f(t)$ (hier: $f(t) = a_1 + a_2 t$), Meßwerte liegen vor in der Form (t_j, b_j) mit $j = 1, \ldots, m$; aber es gibt keine Wahl von a_1, a_2 mit $b_j = f(t_j)$ für alle $j = 1, \ldots, m$.

Idee: Für $\alpha_1, \alpha_2 \in \mathbb{R}$ berechne Fehler

$$F(\alpha_1, \alpha_2) := \sum_{j=1}^{m} (f(t_j) - b_j)^2$$

also $F : \mathbb{R}^2 \to \mathbb{R}$ und bestimme α_1^*, α_2^* so, daß $F(\alpha_1^*, \alpha_2^*)$ minimal wird. Mit

$$\begin{pmatrix} \alpha_1 + \alpha_2 t_1 \\ \vdots \\ \alpha_1 + \alpha_2 t_m \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_m \end{pmatrix}}_{:=A} \cdot \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

ergibt sich mit $x := \binom{\alpha_1}{\alpha_2}$:

$$F(\alpha_1, \alpha_2) = \left\| \begin{pmatrix} f(t_1) \\ \vdots \\ f(t_m) \end{pmatrix} - \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \right\|_2^2 = \left\| A \binom{\alpha_1}{\alpha_2} - b \right\|_2^2 = \|Ax - b\|_2^2$$

analytischer Ansatz: *Notwendige Bedingung* für ein Minimum von F (stetig differenzierbar)

$$0 = \nabla F(\alpha_1^*, \alpha_2^*) = \begin{pmatrix} \partial_{\alpha_1} F(\alpha_1^*, \alpha_2^*) \\ \partial_{\alpha_2} F(\alpha_1^*, \alpha_2^*) \end{pmatrix}$$

Dies ist äquivalent zu $A^T(Ax^* - b) = 0 \Leftrightarrow A^T A x^* = A^T b$, die so genannte *Normalengleichung.*
Hinreichende Bedingung ist die notwendige Bedingung erster Ordnung und daß $\nabla^2 F(x^*)$ positiv definit ist, d.h.

$$\nabla^2 F(x^*) = \begin{pmatrix} \partial_{x_1,x_1} F(x^*) & \partial_{x_1,x_2} F(x^*) \\ \partial_{x_2,x_1} F(x^*) & \partial_{x_2,x_2} F(x^*) \end{pmatrix} (x^*) = A^T A$$

Behauptung: $\nabla^2 F(x^*)$ ist symmetrisch und positiv definit genau dann, wenn $\operatorname{rg}(A) = n$, also maximal ist (d.h. hier gleich 2). *Beweis:* Es gilt immer: $x^T A^T A x = \|Ax\|_2^2 \geq 0$; es gilt $\|Ax\|_2 = 0$ genau dann, wenn $Ax = 0$ ist, dies ist äquivalent zu $x = 0$, falls A vollen Rang hat.

Das Verfahren ist OK, aber $A^T A$ ist aufwändig, falls A zu groß ist und wegen $\operatorname{cond}(A^T A) \approx \operatorname{cond}(A)^2$ ist der Genauigkeitsverlust recht hoch.

Ansatz mit QR-Zerlegung:

Satz: Sei $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$. Das lineare Ausgleichsproblem $\|Ax - b\| = \min$ mit $x \in \mathbb{R}^n$ besitzt mindestens eine Lösung x. Ist y eine weitere Lösung, dann gilt $Ax = Ay$, d.h. das *Residuum* $r := b - Ax$ ist eindeutig bestimmt, es gilt $A^T r = 0 \in \mathbb{R}^n$. Jede Lösung des Ausgleichsproblems ist auch Lösung der Normalengleichung $A^T A x = A^T b$ und umgekehrt.

Beweis: Betrachte $U := \{ Ax \mid x \in \mathbb{R}^n \} \subseteq \mathbb{R}^m$. Dann ist der Senkrechtraum zu U $U^\perp = \{ t \in \mathbb{R}^m \mid u^T t = 0 \; \forall \; u \in U \}$. Es gilt $\mathbb{R}^m = U \oplus U^\perp$, d.h. für alle b findet man $b = u + u^\perp$ mit $u \in U$ und $u^\perp \in U^\perp$ eindeutig bestimmt. Weiter gibt es ein (nicht notwendigerweise eindeutig bestimmtes) $x \in \mathbb{R}^n$ mit $u = Ax$. Also gilt $A^T b = A^T u + A^T u^\perp$. Dabei ist $A^T u^\perp = 0$, also $A^T b = A^T Ax$, d.h. x ist Lösung der Normalengleichung.

Es gilt: $A^T Ay = A^T b \Leftrightarrow A^T(b - Ay) = 0$ und $Ay \in U$ und $b - Ay \in U^\perp$. Da die Zerlegung $b = u + u^\perp$ eindeutig ist, gilt $Ay = u = Ax$ und $b - Ay = u^\perp = b - Ax$.

Sei nun $z \in \mathbb{R}^n$ beliebig. Es gilt

$$\|b - Az\|_2^2 = \|b - Ax + Ax - Az\|_2^2 \overset{(\star)}{=} \|b - Ax\|_2^2 + \|Ax - Az\|_2^2 \geq \|b - Ax\|_2^2$$

Dabei gilt $(\star)$ wegen $b - Ax \in U^\perp$ und $Ax - Az \in U$.

Zur Berechnung der QR-Zerlegung: Sei $A = QR$ mit $Q^T Q = E$ (d.h. Q Orthogonalmatrix) und R obere Dreiecksmatrix.

$$\|Ax - b\|_2 = \|QRx - b\|_2 = \left\|Q(Rx - Q^{-1}b)\right\|_2 = \left\|Rx - Q^T b\right\|_2$$

Also F minimal, falls $Rx^* = d_1$ und $F(x^*) = \|d_2\|^2$

2.5.1 Householder-Transformation

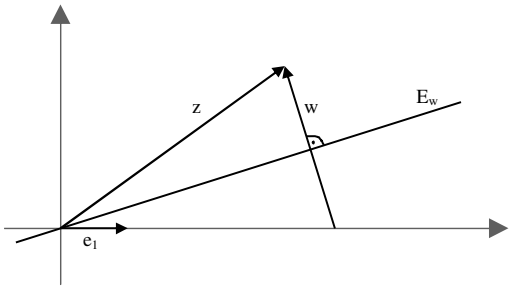
Aufgabe: Gegeben sei $q = m + 1 - j$ und

$$z := \begin{pmatrix} a_{j,j}^{(j)} \\ \vdots \\ a_{m,j}^{(j)} \end{pmatrix} \in \mathbb{R}^q$$

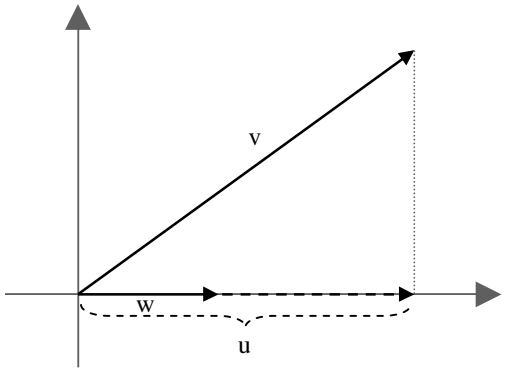
Gesucht sind $H \in \mathbb{R}^{q \times q}$ und $\alpha \in \mathbb{R}$ mit $H^T H = \mathbf{E}_q$ und

$$Hz = \begin{pmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \alpha \mathbf{e}_1$$

Da H orthogonal ist, gilt $\|z\| = \|Hz\| = |\alpha|$, also $\alpha = \pm \|z\|$. Graphisch bedeutet dies:



Dann ist $E_w = \left\{ x \in \mathbb{R}^q \mid w^T x = 0 \right\}$ mit $w = z - \alpha \cdot e_1$, gesucht ist die Matrix-Darstellung der Spiegelung an E_w! Für $v \in \mathbb{R}^q$ ist $u := \frac{v^T w}{w^T w} w$ der Anteil von v längs w (Projektion von v auf $\mathbb{R}w$):



Die Spiegelung erhält man durch $v \mapsto v - 2u$, damit ist

$$H(v) = v - 2\frac{v^T w}{w^T w} w = \mathbf{E}v - \frac{2}{w^T w} w(w^T v) = \underbrace{\left(\mathbf{E} - 2\frac{ww^T}{w^T w} \right)}_{H} v$$

Lemma: Sei $0 \neq w \in \mathbb{R}^q$ ein beliebiger Vektor. Sei $H := \mathbf{E} - \frac{2}{w^T w} ww^T \in \mathbb{R}^{q \times q}$. Dann gilt:

1. $H = H^T$
2. $H^T H = \mathbf{E}_q$
3. $H^{-1} = H$

Beweis:

1. Es gilt:

$$\begin{aligned}\left(\mathbf{E} - \frac{2}{w^T w} w w^T\right)^T &= \mathbf{E} - \frac{2}{w^T w}(w w^T)^T \\ &= \mathbf{E} - \frac{2}{w^T w} w w^T = H\end{aligned}$$

2. Es gilt:

$$\begin{aligned}H^T H &= \left(\mathbf{E} - \frac{2}{w^T w} w w^T\right) \cdot \left(\mathbf{E} - \frac{2}{w^T w} w w^T\right) \\ &= \mathbf{E} - \frac{2}{w^T w} w w^T - \frac{2}{w^T w} w w^T - \frac{4}{(w^T w)^2} w w^T w w^T \\ &= \mathbf{E} - \frac{2}{w^T w} w w^T - \frac{2}{w^T w} w w^T - \frac{4}{w^T w} w w^T = \mathbf{E}\end{aligned}$$

3. Da H orthogonal ist, ist $H^{-1} = H^T = H$.

Für $w = z - \alpha \cdot \mathbf{e}_1$ mit $\alpha \pm \|z\|$ ist $Hz = z - 2\frac{w^T z}{w^T w} \cdot w = z - w$ wegen

$$\begin{aligned}w^T z &= z^T z - \alpha \cdot \mathbf{e_1}^T z \\ &= z^T z - z_1 \alpha \\ w^T w &= z^T z - 2\alpha z^T \mathbf{e}_1 + \alpha^2 \mathbf{e}_1^T \mathbf{e}_1 \\ &= z^T z - 2\alpha z_1 + \alpha^2 \\ &= 2(z^T z - \alpha z_1) \\ &= 2 w^T z\end{aligned}$$

Bemerkung: Es ist $w_j = z_j$ für $j > 1$ und $w_1 = z_1 - \alpha$. Hier kann **Auslöschung** auftreten! Wähle daher das Vorzeichen von α entsprechend so, daß keine Auslöschung auftritt!

Zum **Algorithmus**:

$$\begin{aligned}z &= A(j:m, j) \\ \alpha &= -\operatorname{sign}(z_1)\,\|z\| \\ w &= z - \alpha \mathbf{e}_1 \\ H^{(j)} &= \mathbf{E}_m - \frac{2}{v^T v} v v^T \ \text{ mit } \ v = \begin{pmatrix} 0 \\ w \end{pmatrix} \in \mathbb{R}^m \\ &= \mathbf{E}_m - \frac{2}{v^T v} \begin{pmatrix} 0 & \cdots 0 \cdots \\ \vdots & \\ 0 & w w^T \\ \vdots & \end{pmatrix}\end{aligned}$$

Beispiel: Mit $\alpha = -\|z\| = -3$ im folgenden Beispiel ist

$$A = \begin{pmatrix} 2 & 2 \\ 1 & 6 \\ 2 & 7 \end{pmatrix} \Longrightarrow z = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix} \Longrightarrow w = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix} - \alpha \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 5 \\ 1 \\ 2 \end{pmatrix}$$

In diesem Fall ist $v = w$. Damit ist

$$\begin{aligned} H^{(1)} &= \mathbf{E}_3 - \frac{2}{30} v v^T \Longrightarrow H^{(1)} z = z - 2\frac{z^T w}{w^T w} w \\ &= \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix} - 2\frac{15}{30} \begin{pmatrix} 5 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} -3 \\ 0 \\ 0 \end{pmatrix} = \alpha \mathbf{e}_1 \\ H^{(1)} \begin{pmatrix} 2 \\ 6 \\ 7 \end{pmatrix} &= \begin{pmatrix} 2 \\ 6 \\ 7 \end{pmatrix} - 2\frac{30}{30} \begin{pmatrix} 5 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} -8 \\ 4 \\ 3 \end{pmatrix} \\ H^{(1)} A^{(1)} &= \begin{pmatrix} -3 & -8 \\ 0 & 4 \\ 0 & 3 \end{pmatrix} = A^{(2)} \end{aligned}$$

Schritt 2: $z = \binom{4}{3}$, dann ist $\alpha = -\|z\| = -5$, somit $w = z - \alpha \mathbf{e}_1 = \binom{9}{3}$, dann ist

$$\begin{aligned} v &= \begin{pmatrix} 0 \\ 9 \\ 3 \end{pmatrix} \\ H^{(2)} &= \mathbf{E}_3 - 2\frac{v v^T}{v^T v} = \mathbf{E}_3 - \frac{2}{v^T v} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 81 & 27 \\ 0 & 27 & 9 \end{pmatrix} \\ H^{(2)} A^{(2)} &= \begin{pmatrix} -3 & -8 \\ 0 & -5 \\ 0 & 0 \end{pmatrix} \end{aligned}$$

Der Algorithmus in MatLab:

```
function [alpha, res] = qr_zerl(A, b);
[m,n] = size(A);
A = [A, b];
alpha = zeros(m,1);
for k = 1:n,
      w = A(k:end,k);
      norm_w = norm(w);
```

```
        alpha(k) = (2*(A(k,k)<0)-1)*norm_w;
        w(1) = A(k, k) - alpha(k);
        x2_ww = 1/(norm_w^2-alpha(k)*A(k,k));
        % A --> A - (2/w'*w)*w*(w'A)
        A(k:end,k+1:end) = A(k:end,k+1:end)
                    - w*(w'*A(k:end,k+1:end))*x2_ww;
        A(k:end,k) = w;
end;
```

2.6 Singulärwertzerlegung

Satz: Sei $A \in \mathbb{R}^{m\times n}$. Dann gibt es Matrizen U, Σ, V mit $U \in \mathbb{R}^{m\times m}$, $\Sigma \in \mathbb{R}^{m\times n}$ und $V \in \mathbb{R}^{n\times n}$, $U^TU = \mathbf{E}_m$, $V^TV = \mathbf{E}_n$ und Σ ist Diagonalmatrix mit den Einträgen $\sigma_1, \ldots, \sigma_r$ mit $r = \min\{m, n\}$; dabei ist $\sigma_j \geq 0$ und

$$A = U\Sigma V^T = \sum_{j=1}^{r} \sigma_r u_j v_j^T$$

Beweis: Sei o.B.d.A. $m \geq n$ und $A \neq 0$, Induktion über n:

- Induktionsanfang: Für $n = 1$ ist bei $A^T = a^T = (a_1, \ldots, a_m)$, dann ist $\sigma_1 := \|a\|_2$, $u_1 := \frac{1}{\sigma_r}a$ und $v_1 := (1)$. Dann ist

$$A = \sigma_1 u_1 v_1^T = \underbrace{\begin{pmatrix} \vdots & \vdots \\ u_1 & \hat{U} \\ \vdots & \vdots \end{pmatrix}}_{U} \cdot \underbrace{\begin{pmatrix} \sigma_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{\Sigma} \cdot \underbrace{\begin{pmatrix} 1 \end{pmatrix}^T}_{V^T}$$

- Induktionsschluß: Gelte die Aussage für $n-1$-spaltige Matrizen. Betrachte $\sigma = \|A\|_2 = \max\{\|Ax\|_2 \mid \|x\|_2 = 1\}$, wähle v mit $\|v\|_2 = 1$ und $\sigma = \|Av\|_2$. Es gilt $Av = \sigma u$, d.h. $u = \frac{1}{\sigma} \cdot Av$ (falls $\sigma \neq 0$, setze $u = \mathbf{e}_1$ sonst). Setze nun $V = \begin{pmatrix} v & \hat{V} \end{pmatrix}$, so daß $V^TV = \mathbf{E}_n$; und $U = \begin{pmatrix} u & \hat{U} \end{pmatrix}$, so daß $U^TU = \mathbf{E}_m$. Nun sei $A' := U^TAV$, dann ist

$$\begin{aligned} A' &= \begin{pmatrix} u^T \\ \hat{U}^T \end{pmatrix} \begin{pmatrix} Av & A\hat{V} \end{pmatrix} = \begin{pmatrix} u^TAv & u^TA\hat{V} \\ \hat{U}^TAv & \hat{U}^TA\hat{V} \end{pmatrix} \\ &= \begin{pmatrix} \sigma & \cdots & w^t & \cdots \\ \vdots & & & \\ 0 & & \hat{A} & \\ \vdots & & & \end{pmatrix} \end{aligned}$$

Mit $w \in \mathbb{R}^{n-1}$, $\hat{A} \in \mathbb{R}^{(m-1)\times(n-1)}$. *Behauptung:* $w = 0$. *Angenommen,* $w \neq 0$. Dann ist $z := \binom{\sigma}{w} \neq 0$. Nun ist

$$A'z = \begin{pmatrix} \sigma & w^T \\ 0 & \hat{A} \end{pmatrix} \begin{pmatrix} \sigma \\ w \end{pmatrix} = \begin{pmatrix} \sigma^2 + w^T w \\ \hat{A}w \end{pmatrix}$$

dabei gilt $\|A'z\|_2 = (\sigma^2 + w^T w) + \left\|\hat{A}w\right\|_2^2$. Damit gilt:

$$\sigma = \|A\| = \|A'\|_2 \geq \frac{\|Az\|_2}{\|z\|_2} \geq \frac{\sigma^2 + w^T w}{\sqrt{\sigma^2 + w^T w}} = \sqrt{\sigma^2 + w^T w} > \sigma$$

Also:

$$\begin{aligned} A &= UA'V^T = U \begin{pmatrix} \sigma & 0 \\ 0 & \hat{A} \end{pmatrix} V^T \\ &= U \begin{pmatrix} \sigma & 0 \\ 0 & U_2 \Sigma_2 V_2^T \end{pmatrix} V^T \\ &= \underbrace{U \begin{pmatrix} 1 & 0 \\ 1 & U_2 \end{pmatrix}}_{U_A} \underbrace{\begin{pmatrix} \sigma & 0 \\ 0 & \Sigma_2 \end{pmatrix}}_{\Sigma_A} \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & V_2 \end{pmatrix}^T V^T}_{V_A^T} \end{aligned}$$

Es gilt insgesamt:

$$\begin{aligned} A &= U\Sigma V^T \\ &= \sum_{v=1}^{m} \sum_{\mu=1}^{n} u(:, v) \Sigma_{r\nu} v^T(\mu, :) \\ &= \sum_{j=1}^{r} u(:, j) \sigma_j (v(:, j))^T \\ &= \sum_{j=1}^{r} \sigma_j u_j v_j^T \end{aligned}$$

Beispiel: Seien folgende Matrizen definiert:

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 3 \\ 0 & 0 \end{pmatrix} \quad U = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 3 & 0 \\ 0 & 2 \\ 0 & 0 \end{pmatrix} \quad V = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Dann ist

$$\begin{aligned} A &= U\Sigma V^T = \sum_{j=1}^{2} \sigma_j u_j v_j^T \\ &= 3 \cdot \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 \end{pmatrix} + 2 \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \end{pmatrix} \\ &= 3 \cdot \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} + 2 \cdot \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \end{aligned}$$

Definition: Sei $A \in \mathbb{R}^{m \times n}$, $A = U\Sigma V^T$ mit U, Σ, V wie im Satz über Singulärwertzerlegung und sei $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{m \times n}$ diagonal mit $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_r$. Die Zerlegung heißt *Singulärwertzerlegung (SVD)* und die Zahlen σ_j *Singulärwerte*.

Satz: Sei $A \in \mathbb{R}^{m \times n}$ mit $A = U\Sigma V^T$ eine Singulärwertzerlegung. Es gilt:

1. $\mathrm{rg}(A) = \mathrm{rg}(\Sigma) =: p$, Σ ist eindeutig bestimmt
2. $A = \sum_{j=1}^{p} \sigma_j u_j v_j^T$
3. $\|A\|_2 = \sigma_1$

Beweis:

1. Multiplizieren mit regulären Matrizen verändert den Rang nicht. Wegen $A^T A = V\Sigma^T U^T U \Sigma V^T = V \Sigma^T \Sigma V^T$ und

$$\begin{aligned} \chi_{A^T A}(\lambda) &= \det(\lambda \mathbf{E} - A^T A) \\ &= \det(\lambda V V^T - V \Sigma^T \Sigma V^T) \\ &= \det(V(\lambda \mathbf{E} - \Sigma^T \Sigma) V^T) \\ &= \det(\lambda \mathbf{E} - \Sigma^T \Sigma) = \chi_{\Sigma^T \Sigma}(\lambda) \\ &= \prod_{j=1}^{n} (\lambda - \sigma_j^2) \\ &= \lambda^{n-p} \prod_{j=1}^{p} (\lambda - \sigma_j^2) \end{aligned}$$

   Da die Eigenwerte von $A^T A$ eindeutig bestimmt sind, sind die σ_j eindeutig bestimmt.

Satz: Sei $A = U\Sigma V^T$ eine Singulärwertzerlegung. Die Matrix $A^{(k)} = \sum_{j=1}^{k} \sigma_j u_j v_j^T$ für $k \leq \mathrm{rg}(A)$ ist die beste Rang-k-Approximation an A, d.h.

$$\left\|A - A^{(k)}\right\|_2 = \min\{\|A - X\|_2 \mid \mathrm{rg}(X) = k\}$$

Beweis: Es ist $U^T A(k) V = \mathrm{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0) \in \mathbb{R}^{m\times n}$, d.h. $\mathrm{rg}\, A^{(k)} = k$. Es gilt weiter

$$\begin{aligned}\left\|A - A^{(k)}\right\|_2 &= \left\|U^T(A - A^{(k)})V\right\|_2 \\ &= \|\mathrm{diag}(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_p, 0, \ldots, 0)\| \\ &= \sigma_{k+1}\end{aligned}$$

Sei nun X mit $\mathrm{rg}(X) = k$. Die Dimension des Nullraumes $\dim \mathcal{N}(X) = n - k$. Betrachte

$$W = \mathcal{N}(X) \cap \mathrm{span}\{v_1, \ldots, v_{k+1}\} \neq \{0\}$$

Für $z \in W$ mit $\|z\|_2 = 1$ gilt $Xz = 0$ und es gilt

$$\begin{aligned}\|(A - X)z\|_2 &= \|Az\|_2 = \left\|\sum_{j=1}^{r} \sigma_j u_j v_j^T z\right\|_2 \\ &= \sum_{j=1}^{k+1} (\sigma_j (v_j^T z))^2 \geq \sigma_{k+1}^2 \sum_{j=1}^{k+1} (v_j^T z)^2 \\ &= \sigma_{k+1}^2 \|z\|_2^2 = \sigma_{k+1}^2\end{aligned}$$

Dabei ist $\sum_{j=1}^{k+1} (v_j^T z)^2 = \|z\|_2^2$ wegen $v_j^T z = \beta_j$ für $z = \sum_{j=1}^{k+1} \beta_j v_j$.

Definition: Sei $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, eine Lösung des linearen Ausgleichsproblems, bei der $\|Ax - b\|_2$ minimal ist, heißt *Minimum-Norm-Lösung*, falls $\|x\|_2 \leq \|y\|_2$ für alle Lösungen y des linearen Ausgleichsproblems.

Beispiel: Sei gegeben:

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \quad b = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}$$

Dann ist

$$\|Ax - b\|_2^2 = \left\| \begin{pmatrix} 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} \right\|_2^2 = \left\| \begin{pmatrix} x_2 \\ -1 \\ -2 \end{pmatrix} \right\| = x_2^2 + 5$$

Somit ist $y \in \mathbb{R}^2$ Lösung des linearen Ausgleichsproblems genau dann, wenn $y_2 = 0$ ist. Somit ist $y = \binom{y_1}{0}$ Lösung des Problems und $x = 0$ ist die Minimum-Norm-Lösung.

Satz: Das Lineare Ausgleichsproblem hat genau eine Minimum-Norm-Lösung x. Ist $A = U\Sigma V^T$ eine Singulärwertzerlegung, so gilt $x = \sum_{j=1}^{r} \frac{1}{\sigma_j}(u_j^T b)v_j$ mit $r = \mathrm{rg}(A)$.

Beweis: Sei $A = U\Sigma V^T$ eine Singulärwertzerlegung. Dann ist

$$\|Ax - b\|_2 = \left\| U\Sigma \underbrace{V^T x}_{z} - U \underbrace{U^T b}_{c} \right\|_2 = \|\Sigma z - c\|_2$$

Somit ist x eine Minimum-Norm-Lösung des linearen Ausgleichsproblems $\|Ax - b\| = \min$ genau dann, wenn z Minimum-Norm-Lösung von $\|\Sigma z - c\| = \min$ ist. Es gilt $\|x\| = \|Vz\| = \|z\|$.

$$\begin{aligned} \|\Sigma z - c\|_2^2 &= \left\| \begin{pmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \vdots \end{pmatrix} \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} - \begin{pmatrix} c_1 \\ \\ \vdots \\ \\ c_m \end{pmatrix} \right\|^2 \\ &= \sum_{j=1}^{r} (\sigma_j z_j - c_j)^2 + \sum_{j=r+1}^{m} c_j^2 \end{aligned}$$

Also ist

$$z_j = \begin{cases} \frac{c_j}{\sigma_j} & \text{falls} \quad j = 1, \ldots, r \\ 0 & \text{falls} \quad j = r+1, \ldots, n \end{cases}$$

dann ist $x = Vz = \sum_{j=1}^{r} \frac{c_j}{\sigma_j} v_j = \sum_{j=1}^{r} \frac{1}{\sigma_j}(u_j^T b)v_j$.

Definition: Sei $A \in \mathbb{R}^{m \times n}$ eine Matrix. Die Matrixdarstellung A^+ der Abbildung $\mathcal{A}^+ : \mathbb{R}^m \to \mathbb{R}^n$ mit $b \mapsto x$, wobei x Minimum-Norm-Lösung des linearen Ausgleichsproblems $\|Ax - b\|_2 = \min$ heißt *Pseudoinverse* von A.

Beispiel: Aus $A = \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_n \\ & & \end{pmatrix} \in \mathbb{R}^{m \times n}$ folgt:

$$A^+ = \begin{pmatrix} a_1^+ & & \\ & \ddots & \\ & & a_n^+ \end{pmatrix} \text{ mit } a_j^+ = \begin{cases} \frac{1}{a_j} & \text{falls} \quad a_j \neq 0 \\ 0 & \text{falls} \quad a_j = 0 \end{cases}$$

Satz: Ist $A = U\Sigma V^T$ eine Singulärwertzerlegung von A, dann gilt: $A^+ = V\Sigma^+ U^T$.

Bei einer schlechten Kondition der Matrix betrachte ein Ersatzproblem:

$$\Sigma_\beta^+ = \text{diag}(\sigma_j^{+,\beta}) \text{ mit } \sigma_j^{+,\beta} = \begin{cases} \frac{1}{b_j} & \text{falls} \quad b_j \geq \beta \\ 0 & \text{falls} \quad b_j < \beta \end{cases}$$

3 Interpolation

Finde zu gegebenen Einzelwerten (t_i, y_i) eine Funktion $f : \mathbb{R} \to \mathbb{R}$ mit $f(t_i) = y_i$. Zur Wahl von f: verwende *Polynome*, *Triogonometrische Polynome* oder *stückweise Polynome* (Spline Interpolation). Wozu ist Interpolation sinnvoll?

- zur Tabellierung (z.B. Sinus, Exponentialfunktion etc. auf dem Taschenrechner)
- zur Visualisierung und zum Design
- Ansatz für Quadratur und Differentiation (später mehr dazu)
- Ersatzf einer komplizierten Funktion durch etwas elementares mit „ähnlichen“ Eigenschaften

3.1 Polynomiale Interpolation

Gegeben seien Daten (t_j, y_j) mit $j = 0, \ldots, n$. Gesucht ist ein Polynom $p \in \Pi$ mit $p(t_j) = y_j$ für alle $j = 0, \ldots, n$. Dabei bezeichnet man die Punkte t_j als *Knoten*, die Menge aller Knoten als *Gitter*. Die *Gitterweite* ist der maximale Abstand *benachbarter Knoten*, d.h. Knoten, zwischen denen kein weiterer Knoten liegt.

Satz: Sei $n \in \mathbb{N}_0$, $(t_j, y_j) \in \mathbb{R}^2$ mit $j = 0, \ldots, n$ und die t_j paarweise verschieden. Dann gibt es genau ein Polynom $p_n \in \Pi_n$ mit

$$p_n(x) = \sum_{j=0}^{n} \alpha_j x^j \quad \text{und} \quad p(t_j) = y_j \; \forall \; j = 0, \ldots, n$$

Beweis: Sei φ_j mit $j = 0, \ldots, n$ die monumiale Basis von Π_n, d.h. $\varphi_j(x) = x^j$. Dann ist

$$\left(y_k = p_n(t_k) = \sum_{j=0}^{n} \alpha_j \varphi_j(t_k) = \sum_{j=0}^{n} \alpha_j t_k^j \right) \; \forall \; k = 0, \ldots, n$$

ein lineares Gleichungssystem mit $(n + 1)$ Unbekannten α_j und $(n + 1)$ Gleichungen:

$$\begin{pmatrix} y_0 \\ \vdots \\ y_n \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & t_0^1 & \ldots & t_0^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_n^1 & \ldots & t_n^n \end{pmatrix}}_{V = (t_k^j)_{k,j}} \cdot \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_n \end{pmatrix}$$

Es ist damit:

$$\begin{aligned}
\det(V) &= \left| \begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & t_1 - t_0 & \dots & (t_1 - t_0)t_1^{n-1} \\ 1 & \vdots & \ddots & \vdots \\ 1 & t_n - t_0 & \dots & (t_n - t_0)t_n^{n-1} \end{pmatrix} \right| \\
&= \left| \begin{pmatrix} t_1 - t_0 & \dots & (t_1 - t_0)t_1^{n-1} \\ \vdots & \ddots & \vdots \\ t_n - t_0 & \dots & (t_n - t_0)t_n^{n-1} \end{pmatrix} \right| \\
&= \prod_{\nu=1}^{n}(t_\nu - t_0) \left| \begin{pmatrix} 1 & t_1 & \dots & t_1^{n-1} \\ 1 & \vdots & \ddots & \vdots \\ 1 & t_n & \dots & t_n^{n-1} \end{pmatrix} \right| \\
&= \prod_{\nu=1}^{n}\prod_{\mu=2}^{n}(t_\nu - t_0)(t_\mu - t_1) \left| \begin{pmatrix} 1 & t_2 & \dots & t_2^{n-2} \\ 1 & \vdots & \ddots & \vdots \\ 1 & t_n & \dots & t_n^{n-2} \end{pmatrix} \right| \\
&= \prod_{k<j}(t_j - t_k) \neq 0
\end{aligned}$$

Da die t_j paarweise verschieden sind, ist die Determinante ungleich null und damit die Matrix nicht singulär, d.h. das lineare Gleichungssystem besitzt genau eine Lösung!

Beispiel: Für $n = 2$ seien gegeben

t_j	100	101	103
y_j	1	1	2

Somit ergibt sich für das gesuchte Polynom $p_2(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2$ das Gleichungssystem:

$$\begin{aligned}
y_0 &= \alpha_0 + \alpha_1 t_0 + \alpha_2 t_0^2 \\
y_0 &= \alpha_0 + \alpha_1 t_1 + \alpha_2 t_1^2 \\
y_0 &= \alpha_0 + \alpha_1 t_2 + \alpha_2 t_2^2
\end{aligned}$$

d.h.

$$\begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 & 100 & 100^2 \\ 1 & 101 & 101^2 \\ 1 & 103 & 103^2 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix}$$

Damit ergibt sich $\alpha_2 = \frac{1}{6}$, $\alpha_1 = -33.5$ und $\alpha_0 = 1684.\bar{3}$ und somit beispielsweis $p_2(102) = \frac{4}{3}$.

Zur Anzahl der **FLOPs** ergibt sich:

1. die Lösung des linearen Gleichungssystems kostet $\mathcal{O}(n^3)$

2. Auswertung von $p_n(x)$ kostet naiv $\mathcal{O}(n^2)$, läßt sich verbessern mit dem Horner-Schema zu $\mathcal{O}(n)$

Trotzdem schlecht, da $\text{cond}((t_k^j)_{k,j}) \to \infty$ und insgesamt $\mathcal{O}(n^3)$.

3.1.1 Lagrange-Darstellung

Betrachte die **Lagrange-Darstellung:** Für festes n und paarweise verschiedene Knoten $t_0, \ldots, t_n$ sind die Lagrange-Polynome

$$L_j \in \Pi_n \text{ mit } L_j(t_k) = \begin{cases} 0 & \text{falls} \quad t_j \neq t_k \\ 1 & \text{falls} \quad t_j = t_k \end{cases} = \prod_{\substack{k=0 \\ k \neq j}}^{n} \frac{x - t_k}{t_j - t_k}$$

Betrachte das Polynom $p_n(x)$ in der Basisdarstellung zu den L_j, also $p_n(x) = \sum_{j=0}^{n} \alpha_j L_j(x)$. Dann ist $y_k = \sum_{j=0}^{n} \alpha_j L_j(t_k) = \alpha_k$. Damit ergibt sich

$$\begin{pmatrix} y_0 \\ \vdots \\ y_n \end{pmatrix} = \mathbf{E} \cdot \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_n \end{pmatrix}$$

Die Matrix besitzt die Kondition eins, super einfache Berechnung der Koeffizienten. Auch hier die **FLOPs**:

1. die Lösung des linearen Gleichungssystems: *geschenkt!*

2. Auswertung von $p_n(x)$: $n+1$ Summanden $\alpha_j L_j(x)$, also $\mathcal{O}(n^2)$

Satz: Sei $n \in \mathbb{N}$, $t_1, \ldots, t_n$ paarweise verschieden. Jedes Polynom $p_n \in \Pi_n$ mit $p_n(t_j) = 0$ und $j = 1, \ldots, n$ besitzt die Darstellung

$$p_n(x) = c \prod_{k=1}^{n} (x - t_k)$$

Beweis: Mit $(t_0, y_0) \in \mathbb{R}^2$ mit $t_0 \notin \{t_1, \ldots, t_n\}$ und $y_0 := p_n(t_0)$. Dann ist p_n das eindeutig bestimmte Interpolationspolynom zu den Daten $(t_j, p_n(t_j))$ für $j = 0, \ldots, n$. Es ist

$$p_n(x) = \sum_{j=0}^{n} p_n(t_j) L_j(x) = \frac{p_n(t_0)}{\prod_{k=1}^{n} (t_0 - t_k)} \prod_{k=1}^{n} (x - t_k)$$

3.1.2 Berechnung mit dividierten Differenzen

Definition: Sei $k \in \mathbb{N}_0$, $t_0, \ldots, t_k \in \mathbb{R}$ paarweise verschieden, $y_0, \ldots, y_k \in \mathbb{R}$ und p_k das Interpolationspolynom zu den Daten (t_j, y_j), $j = 0, \ldots, k$,

$$p_k(x) = \alpha_k x^k + \ldots + \alpha_0 \quad p_k(t_j) = y_j \; \forall \; 0, \ldots, k$$

Die Zahl $y[t_0, \ldots, t_k] := \alpha_k$ heißt die *k-te dividierte Differenz.*

Bemerkung: Das Interpolationspolynom ist eindeutig bestimmt, also auch dessen Höchstkoeffizient und damit auch die dividierte Differenz; diese ist auch unabhängig von der Reihenfolge der Knoten.

Beispiele:

1. Für $k = 0$ und (t_0, y_0) ist $p_0(x) = y_0$, also $y[t_0] = y_0$.
2. Für $k = 1$ und $(t_0, y_0), (t_1, y_1)$ ist

$$p_1(x) = \frac{y_1 - y_0}{t_1 - t_0}(x - t_0) + y_0 = y[t_0, t_1](x - t_0) + y[t_0]$$

Satz: Sei $k \in \mathbb{N}_0$, $t_0, \ldots, t_k \in \mathbb{R}$ paarweise verschieden, $y_0, \ldots, y_k \in \mathbb{R}$ und p_k das Interpolationspolynom zu den Daten (t_j, y_j), $j = 0, \ldots, k$ für $k = n - 1, n$. Es gilt:

$$p_n(x) = \underbrace{\left(\frac{y_n - p_{n-1}(t_n)}{\prod_{j=0}^{n-1}(t_n - t_j)} \right)}_{y[t_0, \ldots, t_n]} \left(\prod_{j=0}^{n-1}(x - t_j) \right) + p_{n-1}(x)$$

Beweis: Es ist $p_1 \in \Pi_n$. Es gilt $p_n(t_j) = p_{n-1}(t_j)$ für $j = 0, \ldots, n-1$, zudem ist $p_n(t_n) = y_n - p_{n-1}(t_n) + p_{n-1}(t_n) = y_n$.

KOROLLAR: Sei $n \in \mathbb{N}_0$, $t_0, \ldots, t_n \in \mathbb{R}$ paarweise verschieden, $y_0, \ldots, y_n \in \mathbb{R}$. Das Interpolationspolynom zu (t_j, y_j) für $j = 0, \ldots, n$ besitzt die Darstellung

$$p_n(x) = \sum_{j=0}^{n} \underbrace{y[t_0, \ldots, t_j]}_{\alpha_j} \cdot \underbrace{\prod_{k=0}^{j-1}(x - t_k)}_{\varphi_j(x)}$$

Beweis: Per Induktion über n:

- Für $n = 0$ ist $p_0(x) = y_0 x^0 = y[t_0]$.
- Schritt $n-1 \to n$: Betrachte

$$\sum_{j=0}^{n} y[t_0,\ldots,t_j]\cdot\prod_{k=0}^{j-1}(x-t_k) = y[t_0,\ldots,t_n]\cdot\prod_{k=0}^{n-1}(x-t_k) + \underbrace{\sum_{j=0}^{n-1} y[t_0,\ldots,t_j]\cdot\prod_{k=0}^{j-1}(x-t_k)}_{\text{IPP zu } (t_j,y_j) \text{ mit } j=0,\ldots,n-1}$$

  Dies ist nach Satz das Interpolationspolynom p_n.

Beispiel: Es seien wieder folgende Zahlen für $n = 2$ gegeben:

t_j	100	101	103
y_j	1	1	2

Dann ist

$$p_2(x) = y[t_0] + y[t_0,t_1](x-t_0) + y[t_0,t_1,t_2](x-t_0)(x-t_1)$$

Es ist $p_2(100) = y[t_0] = 1$, dann ist $p_2(101) = y[t_0] + y[t_0,t_1](101-100) = 1$, weiter gilt $p_2(103) = y[t_0] + y[t_0,t_1](103-100) + y[t_0,t_1,t_2](103-100)(103-101) = 2$. Als Gleichungssystem dargestellt ergibt sich

$$\underbrace{\begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 1 & (t_1-t_0) & 0 & \cdots & \cdots & 0 \\ 1 & (t_2-t_0) & (t_2-t_0)(t_2-t_1) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & (t_n-t_0) & \cdots & \cdots & \cdots & \prod_{j=0}^{n-1}(t_n-t_j) \end{pmatrix}}_{:=A} \cdot \begin{pmatrix} y[t_0] \\ y[t_0,t_1] \\ y[t_0,t_1,t_2] \\ \vdots \\ y[t_0,\ldots,t_n] \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Die untere Dreiecksmatrix A ist leidlich konditioniert! Zur **Berechnung:**

Satz: Sei $n \in \mathbb{N}_0$, $t_0,\ldots,t_{n+1} \in \mathbb{R}$ paarweise verschieden, $y_0,\ldots,y_{n+1} \in \mathbb{R}$ und $p_n \in \Pi_n$ das Interpolationspolynom zu den ersten $n+1$ Daten $(0,\ldots,n)$; betrachte $q_n \in \Pi_n$ Interpolationspolynom zu den hinteren $n+1$ Daten $(1,\ldots,n+1)$. Dann gilt für das IPP p_{n-1} zu allen n Daten:

$$p_{n+1}(x) = \frac{p_n(x)(x-t_{n+1}) - q_n(x)(x-t_0)}{t_0 - t_{n+1}}$$

Beweis: Es ist $p_{n+1} \in \Pi_{n+1}$, zudem ist für $j = 1,\ldots,n$: $p_n(t_j) = q_n(t_j) = y_j$ und somit

$$\begin{aligned} p_{n+1}(t_j) &= \frac{p_n(t_j)(t_j-t_{n+1}) - q_n(t_j)(t_j-t_0)}{t_0-t_{n+1}} \\ &= y_j \frac{t_j - t_{n+1} - t_j + t_0}{t_0 - t_{n+1}} = y_j \end{aligned}$$

Weiter gilt für t_0 und t_{n+1}:

$$\begin{aligned} p_{n+1}(t_0) &= \frac{p_n(t_0)(t_0 - t_{n+1}) - q_n(t_0)(t_0 - t_0)}{t_0 - t_{n+1}} = p_n(t_0) = y_0 \\ p_{n+1}(t_{n+1}) &= \frac{p_n(t_{n+1})(t_{n+1} - t_{n+1}) - q_n(t_{n+1})(t_{n+1} - t_0)}{t_0 - t_{n+1}} = q_n(t_{n+1}) = y_{n+1} \end{aligned}$$

Dabei ist die Reihenfolge der Knoten unwichtig.

KOROLLAR: Sei $n \in \mathbb{N}_0, t_0, \ldots, t_{n+1} \in \mathbb{R}$ paarweise verschieden, $y_0, \ldots, y_{n+1} \in \mathbb{R}$. Für $\nu, \mu \in \{0, \ldots, n\}$ mit $\nu \neq \mu$ gilt:

$$y[t_0, \ldots, t_n] = \frac{y[t_0, \ldots, t_{\nu-1}, t\!\!\!/, t_{\nu+1}, \ldots, t_n] - y[t_0, \ldots, t_{\mu-1}, t\!\!\!/, t_{\mu+1}, \ldots, t_n]}{t_\nu - t_\mu}$$

Beweis: Folgt aus dem Satz, da $y[t_0, \ldots, t_k]$ den Höchstkoeffizienten entsprechender Polynome bezeichnet. **Beispiel:** Es seien nochmals folgende Zahlen für $n = 2$ gegeben:

t_j	100	101	103
y_j	1	1	2

Wiederum ist folgende Darstellung gesucht:

$$p_2(x) = y[t_0] + y[t_0, t_1](x - t_0) + y[t_0, t_1, t_2](x - t_0)(x - t_1)$$

Es ergibt sich (mit $y[t_j, t_{j+1}] = \frac{y[t_j] - y[t_{j+1}]}{t_j - t_{j+1}}$):

j	t_j	$y[t_j]$	$y[t_j, t_{j+1}]$	$y[t_j, t_{j+1}, t_{j+2}]$
0	100	1	$\frac{1-1}{101-100} = 0$	$\frac{\frac{1}{2} - 0}{103 - 100} = \frac{1}{6}$
1	101	1	$\frac{2-1}{103-101} = \frac{1}{2}$	
2	103	2		

Damit ist $p_2(x) = 1 + 0(x - 100) + \frac{1}{6}(x - 100)(x - 101)$. **Beispiel:** Die Werte y_j geben die Anzahl von FLOPs zum Lösen eines LGS der Größe t_j an. Es ergibt sich:

j	t_j	$y[t_j]$	$y[t_j, t_{j+1}]$	$y[t_j, ..., t_{j+2}]$	$y[t_j, ..., t_{j+3}]$	$y[t_j, ..., t_{j+4}]$
0	5	145	149	22	$\frac{2}{3}$	0
1	10	830	369	32	$\frac{3}{2}$	
2	15	2735	689	42		
3	20	6180	1109			
4	25	11725				

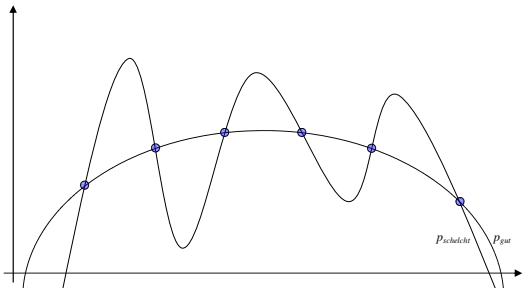
Damit benötigt das Gaußsche Eliminationsverfahren Rechenoperationen in $\mathcal{O}\left(\frac{2}{3}n^3\right)$. Wir betrachten die zum Füllen der Tabelle benötigte **Anzahl an Rechenoperationen**:

- Vorliegend: alle dividierten Differenzen bis zur Nummer k, nun Datum (t_k, y_k)
- k-mal muß man $\frac{d_k - d_{k-1}}{t_k - t_{k-1}}$ berechnen, also $3k$ FLOPs ingesamt.
- Füllen der Tabelle insgesamt: $\sum_{k=0}^{n} 3k = \frac{3}{2}n(n+1) \in \mathcal{O}(n^2)$ FLOPs.
- Anwendung des Horner-Schemas zum Ausrechnen von $p_n(x)$ liefert noch einmal $\mathcal{O}(n)$:

$$p_n(x) = ((\dots(y[t_0,\dots,t_n](x-t_{n-1}) + y[t_0,\dots,t_{n-1}])(x-t_{n-2}))) \dots + y[t_0,t_1](x-t_0) + y[t_0]$$

3.1.3 Fehleranalyse

Eine gegebene Menge von Punkten kann „gut“ oder „schlecht“ interpoliert werden:



Für Daten $y_j = f(t_j)$ und eine Funktion $f\colon \mathbb{R} \to \mathbb{R}$ ist ein Maß für den Fehler die *Fehlerfunktion* $\varepsilon_n(z) = f(z) - p_n(z)$ wobei $p_n \in \Pi_n$ das Interpolationspolynom zu den Daten $(t_j, f(t_j))$ ist.

Satz: Sei $f\colon \mathbb{R} \to \mathbb{R}$ eine Funktion, $p_n \in \Pi_n$ das Interpolationspolynom zu den Daten $(t_j, f(t_j))$ für $j = 0,\dots,n$ paarweise verschieden. Für $z \notin \{t_0,\dots,t_n\}$ gilt:

$$\varepsilon_n(z) = f(z) - p_n(z) = f[t_0,\dots,t_n,z] \prod_{j=0}^{n} (z - t_j)$$

Beweis: Sei p_{n+1} das Interpolationspolynom zu den Daten $(t_j, f(t_j))$ für $j = 0,\dots,n+1$ und $t_{n+1} := z$. Dann ist $p_{n+1}(x) = f[t_0,\dots,t_n,z] \prod_{j=0}^{n}(x - t_j) + p_n(x)$ und $f(z) = p_{n+1}(z)$. Somit folgt die Behauptung.

Satz: Sei $[a,b]$ ein Intervall, $f \in \mathcal{C}^n([a,b])$, $t_0, \ldots, t_n \in [a,b]$ paarweise verschiedene Daten. Dann gibt es ein $\xi \in (a,b)$ mit

$$f[t_0, \ldots, t_n] = \frac{f^{(n)}(\xi)}{n!}$$

Beweis: Sei p_n das Interpolationspolynom zu den Daten $(t_j, f(t_j))$ für $j = 0, \ldots, n$. Damit hat die Funktion $\varepsilon_n(x) = f(x) - p_1(x)$ genau $n+1$ Nullstellen. Damit hat die Ableitung $\varepsilon_n'(x) = f'(x) - p_n'(x)$ noch n Nullstellen usw., somit ist $\varepsilon_n^{(n)}(x) = f^{(n)}(x) - p_n^{(n)}(x)$ genau eine Nullstelle ξ. Somit ist

$$0 = \varepsilon_n^{(n)}(\xi) = f^{(n)}(\xi) - n! \cdot f[t_0, \ldots, t_n]$$

Damit ergibt sich die Behauptung.

Satz: Sei $[a,b]$ ein Intervall, $f \in \mathcal{C}^{n+1}([a,b])$, $t_0, \ldots, t_n \in [a,b]$ paarweise verschiedene Daten. Dann ist für $x \in [a,b]$:

$$\varepsilon_n(x) = f(x) - p_n(x) = \frac{f^{(n+1)(\xi)}}{(n+1)!} \prod_{j=0}^{n} (x - t_j) \text{ wobei } \xi \in (a,b)$$

Damit erfolgt eine Zerlegung des Fehlers in zwei Komponenten:

- In den Term $\frac{f^{(n+1)(\xi)}}{(n+1)!}$ gehen nur die Eigenschaften von f ein,
- in den Term $\prod_{j=0}^{n}(x - t_j)$ gehen nur die Eigenschaften der Knoten ein!

Definition: Seien $t_0, \ldots, t_{n-1} \in [a,b]$ paarweise verschiedene Daten. Das folgende Polynom heißt *Knotenpolynom*:

$$\omega_n(x) := \prod_{j=0}^{n-1} (x - t_j) \in \Pi_n$$

Beispiele:

1. Sei $f(x) = e^{\frac{x+1}{2}}$ durch p_2 auf $[a,b]$. Dann ist

$$\varepsilon_2(x) = e^{\frac{x+1}{2}} - p_2(x) = \frac{f'''(\xi)}{3!}(x - t_0)(x - t_1)(x - t_2)$$

Es gilt $f'(x) = \frac{1}{2}f(x) > 0$, d.h. f ist streng monoton steigend, somit ist $\|f\|_\infty = f(b)$, da f zudem noch positiv ist. Weiter ist $f''(x) = \frac{1}{4}f(x)$ und $f'''(x) = \frac{1}{8}f(x)$. d.h. maximal möglich ist $\frac{f(b)}{8}$.

Z.B. $t_0 = a$, $t_1 = \frac{a+b}{2}$, $t_2 = b$ gilt $\left|\prod_{j=0}^{2}(x - t_j)\right| \leq (b-a)^3$ und somit

$$|\varepsilon_2(x)| \leq \frac{f(b)}{8}(b-a)^3$$

2. Sei $f\colon [a,b] \to \mathbb{R}$ mit $f \in \mathcal{C}^2([a,b])$ und $\|f''\|_\infty \leq M$. Sei p_1 das Interpolationspolynom zu $(t_j, f(t_j))$ für $j = 0, 1$. Dann ist

$$|\varepsilon_1(x)| = |f(x) - p_1(x)| = \frac{|f''(\xi)|}{2}\,|(x-t_0)(x-t_1)| \leq \frac{M}{2}\,|\omega_2(x)|$$

Falls nun $x \notin [t_0, t_1]$[7] ist $|\omega_2(x)|$ groß, also auch $|\varepsilon_1(x)|$, d.h. $p_1 \not\approx f$.

Falls nun $x \in [t_0, t_1]$ ist, so ist

$$\|\omega_2\|_{[t_0,t_1]} = \left|\omega_w\left(\frac{t_0+t_1}{2}\right)\right| = \frac{1}{4}(t_1-t_0)^2$$

Hier gilt also: $|\varepsilon_1(x)| \leq \frac{M}{8}(t_1-t_0)^2$.

3. Erstelle für $f(x) = \sin(x)$ eine Tabelle durch stückweise lineare Interpolation zu $(t_j, \sin(t_j))$ und $(t_{j+1}, \sin(t_{j+1}))$, so daß für $\eta \in [t_j, t_{j+1}]$ gilt: $p_n(\eta) \approx f(\eta)$. Wir suchen für eine vorgegebene Toleranz *toll*

$$|\varepsilon_1(\eta)| = |\sin(\eta) - p_1(\eta)| \leq toll$$

Es gilt: $|\varepsilon_1(\eta)| \leq \frac{|f''(\xi)|}{8}(t_{j+1}-t_j)^2$, weiter ist $|f''(\xi)| = |\sin(\xi)| \leq 1$. Definiere $h_j := t_{j+1} - t_j$, nun verlangen wir $|\varepsilon_1(\eta)| \leq \frac{1}{8}h_j^2 \leq toll$, d.h. $h_j < \sqrt{8toll}$.

Allgemeiner gilt:

$$|\varepsilon_n(\eta)| \leq \frac{\left|f^{(n+1)}(\eta)\right|}{(n+1)!}\,|\omega_{n+1}(\eta)| \leq \frac{1}{(n+1)!}\cdot\pi^{n+1}$$

Somit ist $\lim_{n\to 0}|\varepsilon_n(\eta)| = 0$, denn $\sum_{n=-1}^{\infty}\frac{\pi^{n+1}}{(n+1)!} = e^\pi$, d.h. die Folge ist eine Nullfolge.

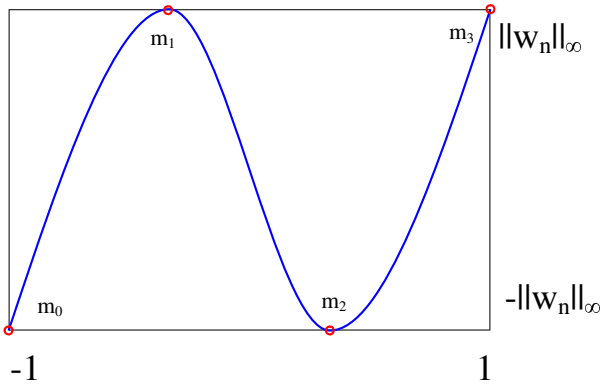
[7] Bei $x \notin [t_0, t_1]$ spricht man von *Extrapolation*, andernfalls von *Interpolation*

3.1.4 Optimale Knoten

Bestimme ein „ausgewogenes“ Knotenpolynom $\hat{\omega}_n = \prod_{j=0}^{n-1}(x - \hat{t}_j)$.

Satz: Das ausgewogene Knotenpolynom $\hat{\omega}_n$ ist optimal, d.h. es gibt kein Knotenpolynom $\omega_n \in \Pi_n$ mit $\|\omega_n\|_\infty < \|\hat{\omega}_n\|_\infty$.

Beweis: Angenommen, es gibt ein ω_n mit $\|\omega_n\|_\infty < \|\hat{\omega}_n\|_\infty$. Die Polynome lassen sich darstellen als $\hat{\omega}_n = x^n \pm \ldots \in \Pi_n$ und ebenso $\omega_n = x^n \pm \ldots \in \Pi_n$, d.h. $q(x) := \hat{\omega}_n(x) - \omega_n(x) \in \Pi_{n-1}$. Sei $m_j \in [-1, 1]$ mit $\hat{\omega}_n(m_j) = (-1)^{n-j} \|\hat{\omega}_n\|_\infty$. Nun gilt $q(m_j) \cdot q(m_{j+1}) < 0$, denn $|\omega_n(m_j)| < |\hat{\omega}_n(m_j)| = \|\hat{\omega}_n\|_\infty$. Also wechselt q das Vorzeichen n mal, damit ist q aber 0 und somit $\omega_n = \hat{\omega}_n$.



Wie finden wir jetzt dieses ausgewogene Polynom? Für $x \in \{m_0, \ldots, m_n\}$ betrachte eine Funktion c_n, die folgende Gleichung erfüllt:

$$(1 - x^2)(c_n'(x))^2 = n^2(c_n^2(x) - 1)$$

Da an den Stellen $m_0, \ldots, m_n$ $c_n(m_i) = \pm 1$ ist, ist an diesen Stellen $(c_n^2 - 1) = 0$, gleichzeitig ist die Ableitung gleich 0 (da an diesen Punkten die Steigung null ist), bringe für die Grade der Polynome noch Korrekturfaktoren $(1 - x^2)$ und n^2 in die Formel. Diese Formel gilt nun für alle x, und es gilt:[8]

$$\begin{aligned}
(1 - x^2)(c_n'(x))^2 &= n^2(1 - c_n^2(x)) \\
\Rightarrow \frac{n}{\sqrt{1 - x^2}} &= \frac{c_n'(x)}{\sqrt{1 - c_n^2(x)}} \\
\Rightarrow \int \frac{n}{\sqrt{1 - x^2}}\, dx &= \int \frac{c_n'(x)}{\sqrt{1 - c_n^2(x)}}\, dx + \mathcal{C}
\end{aligned}$$

[8] „Ich mach mal so 'ne Art Substitution klar...“

Setze $x = \cos\xi$ und $dx = -\sin\xi\, d\xi$, sei zudem $u = c_n(x)$ und $du = c_n'(x)\, dx$, es folgt:

$$\Rightarrow -n \int 1\, d\xi \;=\; \int \frac{1}{\sqrt{1-u^2}}\, du + \mathcal{C}$$

Mit $v = \arccos(u)$ und $du = -\sin v$

$$\Rightarrow n\xi \;=\; v + c$$

Spezielle Lösung für $c = 0$:

$$\begin{aligned} n\xi &= v \\ \Rightarrow n \cdot \arccos(x) &= \cos(n) \\ \Rightarrow \cos(n \cdot \arccos(x)) &= u = c_n(x) \end{aligned}$$

Also: $c_n(x) = \cos(n \cdot \arccos(x))$.

3.1.5 Чебышев-Polynome

Satz: Sei $n \in \mathbb{N}_0$ und $c_n \colon [-1,1] \to \mathbb{R}$ mit $c_n(x) := \cos(n \cdot \arccos(x))$. Dann gilt:

1. $c_n(\hat{t}_j) = 0$ für $\hat{t}_j := \cos\left(\frac{2j+1}{2n}\pi\right)$ für $j = 0, \ldots, n-1$
2. $c_0(x) \equiv 1$, $c_1(x) = x$ und $2x \cdot c_n(x) = c_{n+1}(x) + c_{n-1}(x)$
3. $c_n \in \Pi_n$
4. $c_n = 2^{n-1} \prod_{j=0}^{n-1}(x - \hat{t}_j)$ für $n \geq 1$, $\hat{t}_j$ wie oben
5. $\|c_n\|_\infty = |c_n(m_j)| = 1$ für $m_j = \cos\frac{j\pi}{n}$ (genauer: $c_n(m_j) = (-1)^j$)

Beweis:

1. Es gilt $\cos(z) = 0 \Leftrightarrow z = \pi\frac{2j+1}{2}$, also $\cos(nz) = 0 \Leftrightarrow z = \pi\frac{2j+1}{2n}$. Daher ist $c_n(t) = 0 \Leftrightarrow t = \cos\left(\frac{2j+1}{2n}\pi\right)$. Aufgrund der Periodizität von cos gibt es genau n verschiedene Nullstellen im Intervall $[-1,1]$.

2. Per Induktion über n:

   - Induktionsanfang:

$$\begin{aligned} c_0(x) &= \cos(0 \cdot \arccos(x)) = \cos(0) = 1 \in \Pi_0 \\ c_1(x) &= \cos(1 \cdot \arccos(x)) = x \in \Pi_1 \end{aligned}$$

- Sei die Behauptung richtig für alle $k \leq n$. Setze $y := \arccos(x)$. Dann ist (mit Additionstheorem im dritten Schritt)

$$\begin{aligned} c_{n+1}(x) + c_{n-1}(x) &= \cos((n+1)\cdot y) + \cos((n-1)y) \\ &= \cos(ny+y) + \cos(ny-y) \\ &= 2\cos(ny)\cdot\cos(y) \\ &= 2x \cdot c_n(x) \in \Pi_{n+1} \end{aligned}$$

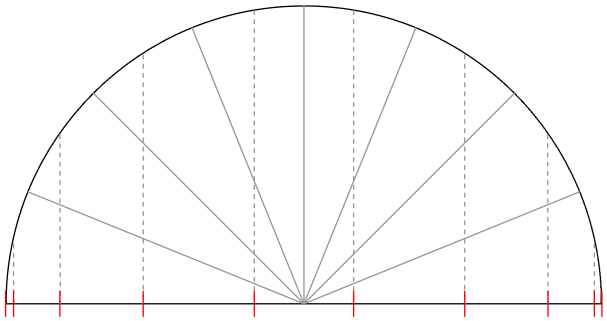
3. siehe eben

4. Für den Höchstkoeffizienten k_n von c_n gilt: $k_0 = 1$, $k_1 = 1$, $k_{n+1} = 2k_n$ (siehe oben)

5. Offensichtlich gilt $|c_n(x)| = |\cos(\bar{x})| \leq 1$. Zudem ist

$$c_n(m_j) = c_n\left(n\cos\left(\frac{j\pi}{n}\right)\right) = \cos(j\pi) = (-1)^j$$

Damit ergeben sich für ein Intervall die optimalen Punkte durch einen Kreisbogen und den Cosinus:



Bemerkung: Man kann auch weitere Parameter mit angeben, suche beispielsweise eine Parabel durch die Punkte $(0,0)$ und $(1,1)$ sowie $f'(0) = 0$. Es ergibt sich:

t_j	$y[t_j]$	$y[t_j, t_j+1]$	$y[t_0,t_1,t_2]$
1	1	$\frac{0-1}{0-1} = 1$	$\frac{0-1}{0-1} = 1$
0	0	$p'(0) = 0$	
0	0		

Damit ist $p_2(x) = 1 + 1(x-1) + 1(x-1)(x-0) = x^2$.

3.2 Spline-Interpolation

Polynome höheren Grades sind nicht unbedingt geeignet für Interpolation, da das Polynom nicht notwendigerweise gegen die Funktion konvergieren und Polynome vor allem oszillieren, je höher der Grad desto schlimmer. **Idee:** Intervall zerlegen! Die Datenpunkte linear zu verbinden ist nicht so gut[9], da die Interpolante nicht differenzierbar ist. Besser:

$$s \in \mathcal{C}^q([a,b]) \text{ und } s|_{[t_j,t_{j+1}]} \in \Pi_n$$

Seien Daten (t_j, y_j) gegeben für $j = 0, \ldots, n$ mit $t_j \in [a,b]$ paarweise verschieden. Sei

$$s_j(x) = \begin{cases} s|_{[a,t_1[}(x) & \text{falls} \quad j = 0 \\ s|_{[t_j,t_{j+1}[}(x) & \text{falls} \quad j = 1, \ldots, n-2 \\ s|_{[t_{n-1},b]}(x) & \text{falls} \quad j = n-1 \end{cases}$$

3.2.1 Lineare Splines

Suche eine Funktion s für ein Intervall $[a,b]$ mit

1. $s \in \mathcal{C}^0([a,b])$ global stetig
2. $s|_{[t_j,t_{j+1}[} \in \Pi_1$ lokal linear
3. $s(t_j) = y_j$ für $j = 0, \ldots, n$ interpolierend

Für die Funktion gilt dann:

2. lokal linear: $s_j(x) = \alpha_j + \beta_j(x - t_j)$
3. interpolierend: $y_j = s(t_j) = s_j(t_j) = \alpha_j + \beta_j(t_j - t_j) = \alpha_j$
1. global stetig: $\lim_{h\to 0} s(t_{j+1} + h) = \lim_{h\to 0} s(t_{j+1} - h$ für $j = 0, \ldots, n-2$

Zudem ist

$$\begin{aligned} y_{j+1} &= s_{j+1}(t_{j+1}) = s(t_{j+1}) = \lim_{h\to 0} s(t_{j+1} - h) \\ &= s_j(t_{j+1}) = \alpha_j + \beta_j(t_{j+1} - t_j) = y_j + \beta_j(t_{j+1} - t_j) \\ y_n &= s_{n-1}(t_n) = \alpha_{n-1} + \beta_{n-1}(t_n - t_{n-1}) \end{aligned}$$

Somit ist

$$\beta_j y[t_j, t_{j+1}] = \frac{y_{j+1} - y_j}{t_{j+1} - t_j} \quad \forall \quad j = 0, \ldots, n-2, n-1$$

[9] „Lada hat mal Autos so gebaut…“

Angenommen, $|f''(x)| \leq M$, dann gilt für $x \in [t_j, t_{j+1}[$:

$$\left| f(x) - s_j^{(n)}(x) \right| = \left| f(x) - s^{(n)}(x) \right| \leq \frac{M}{8}(t_{j+1}^n - t_j^n)^2$$

Zum Beispiel gilt für ein äquidistantes Gitter $t_{j+1}^n - t_j^n = \frac{(b-a)}{n}$:

$$\|f - s^n\|_\infty \leq \max \left\{ \|f - s_j(x)\|_{[t_j^n, t_{j+1}^n]} \;\middle|\; j = 0, \ldots, n-1 \right\} \leq \frac{M}{8} \left(\frac{b-a}{n} \right)^2$$

Damit geht der Fehler gegen 0 für n gegen unendlich.

Satz: Sei $f \in \mathcal{C}^2([a,b])$, $\|f''\|_\infty \leq M$, $x \in [a,b]$, $a = t_0 < \ldots < t_n = b$ und $h_{\max} := \max \{ t_{j+1} - t_j \mid j = 0, \ldots, n-1 \}$. Es gilt:

$$\|f - s\|_\infty \leq h_{\max}^2 \cdot \frac{M}{8}$$

Damit ist $\lim_{h_{\max} \to 0} s = f$.

Zu den **Kosten:** Zum Auswerten des Splines ist notwendig: Wo ist x? Danach: $s_j(x) = y_j + y[t_j, t_{j+1}](x - t_j)$ mit vorberechneten dividierten Differenzen, also nur drei FLOPs!

3.2.2 Quadratische Splines

Forderungen für bessere Ergebnisse: Der Spline soll folgende Eigenschaften haben:

1. $s \in \mathcal{C}^1([a,b])$ global stetig differenzierbar
2. $s|_{[t_j, t_{j+1}[} \in \Pi_2$ lokal quadratisch
3. $s(t_j) = y_j$ für $j = 0, \ldots, n$ interpolierend

Mehr dazu in der Übung

3.2.3 Kubische Splines

Oft wichtig: zweite Ableitung stetig, insgesamt gefordert:

1. $s \in \mathcal{C}^2([a,b])$ global zwei mal stetig differenzierbar
2. $s|_{[t_j, t_{j+1}[} \in \Pi_3$ lokal kubisch

3. $s(t_j) = y_j$ für $j = 0, \ldots, n$ interpolierend

Ansatz: $s_j(x) \in \Pi_3$ hat die Gestalt $s_j(x) = \alpha_j + \beta_j(x - t_j) + \gamma_j(x - t_j)^2 + \delta_j(x - t_j)^3$ Auf jedem der Abschnitte $[t_i, t_{i+1}[$ sind also vier Unbekannte, d.h. insgesamt $4n$ Unbekannte gesucht. Zur Verfügung stehende Gleichungen:

- wegen der Interpolationseigenschaft: $s(t_j) = y_j$ für $j = 0, \ldots, n$ ergibt $n + 1$ Gleichungen
- wegen der Stetigkeit: $s_j(t_{j+1}) = s_{j+1}(t_{j+1})$ für $j = 0, \ldots, n - 2$ ergibt $n - 1$ Gleichungen
- wegen der Stetigkeit der ersten Ableitung: $s'_j(t_{j+1}) = s'_{j+1}(t_{j+1})$ für $j = 0, \ldots, n - 2$ ergibt $n - 1$ Gleichungen
- wegen der Stetigkeit der zweiten Ableitung: $s''_j(t_{j+1}) = s''_{j+1}(t_{j+1})$ für $j = 0, \ldots, n - 2$ ergibt $n - 1$ Gleichungen

Damit ergeben sich $4n - 2$ Gleichungen, d.h. es sind später noch zwei zusätzliche Gleichungen erforderlich. Wir arbeiten hier mit *Momenten* $M_j := s''(t_j)$ (Wert der zweiten Ableitung an der Stelle t_j) und $h_j := t_{j+1} - t_j$. Es ist $s''_j(x)$ eine Gerade mit $s''_j(t_j) = M_j$ und $\lim_{h\to 0} s''_j(t_{j+1} - h) = M_{j+1}$. Unter Berücksichtigung der Stetigkeit von s'' an den Knoten $t_1, \ldots, t_{n-1}$ ist

$$\begin{aligned} s''_j(x) &= \frac{M_j}{h_j}(t_{j+1} - x) + \frac{M_{j+1}}{h_j}(x - t_j) \\ \Rightarrow s'_j(x) &= -\frac{M_j}{2h_j}(t_{j+1} - x)^2 + \frac{M_{j+1}}{2h_j}(x - t_j)^2 + \beta_j \\ \Rightarrow s_j(x) &= \frac{M_j}{6h_j}(t_{j+1} - x)^3 + \frac{M_{j+1}}{6h_j}(x - t_j)^3 + \beta_j(x - t_j) + \alpha_j \end{aligned}$$

Nutze nun die geforderten Eigenschaften der Funktion aus:

1. Interpolationsbedingungen ergeben für $j = 0, \ldots, n - 1$:

   $$y_j = s(t_j) = s_j(t_j) = \frac{M_j}{6}h_j^2 + \alpha_j$$

   Somit ist $\alpha_j = y_j - \frac{M_j \cdot h_j^2}{6}$.

2. Stetigkeit liefert:

   $$\begin{aligned} y_{j+1} &= s(t_{j+1}) = s_j(t_{j+1}) \stackrel{!}{=} \lim_{h\to 0} s(t_{j+1} - h) \\ &= s_j(t_{j+1}) = \frac{M_{j+1}}{6}h_j^2 + \beta_j h_j + \alpha_j \\ y_{j+1} - y_j &= \frac{h_j^2}{6}(M_j - M_{j+1}) + \beta_j h_j \end{aligned}$$

Somit ist $\beta_j = y[t_j, t_{j+1}] + \frac{h_j}{6}(M_{j+1} - M_j)$.

3. Stetigkeit der ersten Ableitung:

$$\begin{aligned} s'(t_j+) &= s'_j(t_j) = -\frac{M_j \cdot h_j}{2} + y[t_j, t_{j+1}] + \frac{h_j}{6}(M_j - M_{j+1}) \\ s'(t_j-) &= s'_{j-1}(t_j) = \frac{M_j \cdot h_{j-1}}{2} + y[t_{j-1}, t_j] + \frac{h_{j-1}}{6}(M_{j-1} - M_j) \end{aligned}$$

Beide Zeilen müssen gleich sein, d.h.

$$\begin{aligned} M_{j-1} \cdot \frac{h_{j-1}}{6} + M_j \left(\frac{h_{j-1}}{2} - \frac{h_{j-1}}{6} + \frac{h_j}{2} - \frac{h_j}{6} \right) + M_{j+1} \cdot \frac{h_j}{6} &= y[t_j, t_{j+1}] - y[t_{j-1}, t_j] \\ \Leftrightarrow M_{j-1} \cdot h_{j-1} + 2 \cdot M_j \cdot (h_{j-1} + h_j) + M_{j+1} \cdot h_j &= 6(y[t_j, t_{j+1}] - y[t_{j-1}, t_j]) \end{aligned}$$

Seien nun

$$\mu_j := \frac{h_{j-1}}{h_{j-1} + h_j} \text{ und } \nu_j := \frac{h_j}{h_{j-1} + h_j} \text{ und } h_j + h_{j-1} = t_{j+1} - t_{j-1}$$

damit ergibt sich

$$\mu_j \cdot M_{j-1} + 2 \cdot M_j + \nu_j \cdot M_{j+1} = 6 \cdot y[t_{j-1}, t_j, t_{j+1}]$$

Als Gleichungssystem ergibt sich:

$$\begin{pmatrix} \mu_1 & 2 & \nu_1 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & \mu_{n-1} & 2 & \nu_{n-1} \end{pmatrix} \cdot \begin{pmatrix} M_0 \\ M_1 \\ \vdots \\ \vdots \\ M_n \end{pmatrix} = 6 \begin{pmatrix} y[t_0, t_1, t_2] \\ y[t_1, t_2, t_3] \\ \vdots \\ y[t_{n-2}, t_{n-1}, t_n] \end{pmatrix}$$

Dies Gleichungs ist unterbestimmt. Zusätzliche zwei Bedingungen für eine eindeutige Lösung:

- vollständiger Spline durch Angabe von $s'(t_q) = f'(t_q)$ oder $s''(t_q) = f''(t_q)$ für $q = 0, n$
- natürlicher Spline: $M_q = s''(t_q) = 0$ für $q = 0, n$.
- periodischer Spline: $s'(t_0) = s'(t_n)$ und $s''(t_0) = s''(t_n)$
- Not-a-Knot-Spline: $s'''_0(t_1) = s'''_1(t_1)$ und $s'''_{n-2}(t_{n-1}) = s'''_{n-1}(t_{n-1})$ ($\mathcal{C}^3$-Bedingung für $s|_{[t_0,t_2[}$ und $s|_{[t_{n-2},t_n]}$).

Satz: Sei $f \in \mathcal{C}^4([a,b])$, $\left\|f^{(4)}\right\|_\infty \leq M$. Dann gilt:

1. $\|f - s\|_\infty \leq \frac{5}{385} h_{\max}^4 M$
2. $\|f' - s'\|_\infty \leq \frac{1}{24} h_{\max}^3 M$
3. $\|f'' - s''\|_\infty \leq \frac{3}{8} h_{\max}^2 M$

wobei $h_{\max} := \max\{|t_{j+1} - t_j|\}$ ist.

Beispiel: Seien folgenden Daten gegeben: $f(-1) = 1$, $f(0) = 0$ und $f(1) = 1$. Damit ergeben sich die Teile s_0 auf $[t_0, t_1[$ und s_1 auf $[t_1, t_2]$ Die natürliche Randbedingung ist $M_0 = M_2 = 0$, Momentensystem: $\mu_1 M_0 + 2M_1 + \nu_1 M_2 = 6y[t_0, t_1, t_2]$. Es ist $y[t_0, t_1, t_2] = 1$. Damit ist $M_1 = 3$. Somit:

$$\begin{aligned}
\alpha_0 &= y_0 = 1 \\
\alpha_1 &= y_1 - \frac{3}{6} = -\frac{1}{2} \\
\beta_0 &= -1 + \frac{1}{6}(0 - 3) = -\frac{3}{2} \\
\beta_1 &= 1 + \frac{1}{6}(3 - 0) = \frac{3}{2}
\end{aligned}$$

Damit ergibt sich s_j aus

$$s_j(x) = \frac{M_j}{6h_j}(t_{j+1} - x)^3 + \frac{M_{j+1}}{6h_j}(x - t_j)^3 + \beta_j(x - t_j) + \alpha_j$$

Insofern ist

$$s(x) = \begin{cases} \frac{1}{2}\left((x+1)^3 - 3(x+1) + 2\right) & \text{falls} \quad x < 0 \\ \frac{1}{2}\left((1 - x^3)^3 + 3x - 1\right) & \text{falls} \quad x \geq 0 \end{cases}$$

Satz: Die Momentenmatrix A des natürlichen Splines ist nicht singulär

$$A = \begin{pmatrix} 2 & \nu_1 & & & \\ \mu_2 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \nu_{n-2} \\ & & & \mu_{n-1} & 2 \end{pmatrix}$$

mit $\mu_1 := v_{n-1} := 0$, $\mu_j + \nu_j \leq 1, \mu_j, \nu_j \geq 0$.

Beweis: Wir zeigen zunächst: Für $Ax = y$ ist $\|x\|_\infty \leq \|y\|_\infty$. Sei $x \in \mathbb{R}^n$ beliebig, $y = Ax$ und r so gegeben, daß $|x_r| = \|x\|_\infty$. Dann ist

$$\begin{array}{lllll} \|y\|_\infty & \geq & |y_r| & = & |\mu_r x_{r-1} + 2x_r + \nu_r x_{r+1}| \\ & \geq & 2\,|x_r| - \mu_r\,|x_{r-1}| - \nu_r\,|x_{r+1}| & \geq & (2 - \mu_r - \nu_r)\,|x_r| \\ & \geq & \|x\|_\infty & & \end{array}$$

Angenommen, A wäre singulär, dann existiert ein $x \neq 0$ mit $Ax = 0 =: y$. Nun gilt aber $0 = \|y\|_\infty \geq \|x\|_\infty > 0$, ein Widerspruch!

3.3 Trigonometrische Interpolation

Hier über $\mathbb{C}$, zur Erinnerung: komplexe Einheit $i^2 = -1$.

Definition: Seien $c_0, \ldots, c_n \in \mathbb{C}$. Ist $c_n \neq 0$, so heißt

$$p\colon [0, 2\pi] \to \mathbb{C} \text{ mit } p(x) := \sum_{j=0}^{n} x_j (e^{ix})^j = \sum_{j=0}^{n} x_j e^{ijx}$$

ein komplexes trigonometrisches Polynom vom Grad n. Mit $\mathbb{T}_n$ bezeichnen wir die Menge aller Polynome vom Grad n.

Satz: Gegeben seien Interpolationsdaten (x_k, y_k) mit $x_k \in [0, 2\pi]$ und $y_k \in \mathbb{C}$ auf einem Gitter

$$G\colon 0 = x_0 < x_1 < \ldots < x_{n-1} < x_n := 2\pi$$

Dann existiert genau ein trigonometrisches Polynom $\mathbb{T}_{n-1}$ mit $p(x_k) = y_k$ für $k = 0, \ldots, n-1$.

Beweis: Mit $z = e^{ix}$ gilt $p(x) = q(z) = \sum_{j=0}^{n-1} c_j z^j$. Existenz und Eindeutigkeit wie bei Polynom-Interpolation.

Sei jetzt speziell $x_j := \frac{2\pi}{n} j$ mit $j = 0, \ldots, n$. Definiere dann $\omega_n := e^{i\frac{2\pi}{n}}$. Verwende zudem als Bezeichnung $z = e^{ix}$ und Polyonme p, q mit $p(x) = q(z)$. Dann gilt:

$$p(x_k) = q(\omega_n^k) = \sum_{j=0}^{n-1} c_j \cdot (\omega_n^k)^j = \sum_{j=0}^{n-1} c_j \cdot \omega_n^{k \cdot j}$$

Sei nun $F_n := (\omega_n^{jk})_{j,k=0}^{n-1}$. *Bemerkung:* Es ist

$$y_k = p(x_k) = q(\omega_n^k) = \sum_{j=0}^{n-1} c_j \omega_n^{jk}$$

Damit ist

$$\begin{pmatrix} y_0 \\ \vdots \\ y_k \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \omega_n^{0\cdot 0} & \cdots & \omega_n^{0\cdot(n-1)} \\ \vdots & \ddots & \vdots \\ \omega_n^{k\cdot 0} & \cdots & \omega_n^{k\cdot(n-1)} \\ \vdots & \ddots & \vdots \\ \omega_n^{(n-1)\cdot 0} & \cdots & \omega_n^{(n-1)^2} \end{pmatrix} \cdot \begin{pmatrix} c_0 \\ \vdots \\ c_{n-1} \end{pmatrix}$$

Naiv könnte man versuchen, $c = F_n^{-1} y$ zu lösen. Dies würde aber $\mathcal{O}(n^3)$ benötigen!

Lemma:

1. Für alle $a \in \mathbb{Z}$ gilt[10]: $\sum_{j=0}^{n-1} (\omega_n^a)^j = n\delta_{a,0}$.
2. Es ist[11] $F_n^H F_n = n\mathbf{E}_n$, d.h. F_n ist bis auf den Faktor n unitär.

Beweis:

1. Falls $a = 0$, so ist die Behauptung klar. Falls $a \neq 0$, so ist[12]

$$\sum_{j=0}^{n-1} (\omega_n^a)^j = \frac{1 - (\omega_n^a)^n}{1 - (\omega_n^a)} = 0$$

2. Es ist

$$\begin{aligned} (F_n^H F_n)_{ab} &= \sum_{j=0}^{n-1} (F_n^H)_{aj} \cdot (F_n)_{jb} = \sum_{j=0}^{n-1} \bar{\omega}_n^{ja} \omega_n^{jb} \\ &= \sum_{j=0}^{n-1} \omega_n^{-ja} \omega_n^{jb} = \sum_{j=0}^{n-1} \left(\omega_n^{b-a}\right)^j = n \cdot \delta_{a,b} \end{aligned}$$

Satz: Die Lösung $c = (c_0, \ldots, c_{n-1})^T \in \mathbb{C}^n$ des obigen Problems mit äquidistanter Knotenwahl lautet

$$c = \frac{1}{n} F_n^H y \text{ bzw. } c_k = \frac{1}{n} \sum_{j=0}^{n-1} y_j \bar{\omega}_n^{jk}$$

[10] mit $\delta_{j,k} = 1$ für $j = k$ und $= 0$ für $j \neq k$
[11] mit $Q^H = (\bar{Q})^T$
[12] mit $\sum_{j=0}^{n-1} q^j = \frac{1-q^n}{1-q}$ und wegen $(\omega_n^a)^n = (\omega_n^n)^a = e^{i2\pi} = 1$

3.3.1 Fast-Fourier-Transformation FFT

Naiv: $\mathcal{O}(n^2)$ ist die optimale Anzahl an Operationen, aber es geht noch viel schneller! Betrachte $n = 2^p, p \in \mathbb{N}$, insbesondere $n = 2^3 = 8$.

$$F_8 = \begin{pmatrix} \omega_n^0 & \omega_n^0 & \omega_n^0 & \omega_n^0 & \omega_n^0 & \omega_n^0 & \omega_n^0 & \omega_n^0 \\ \omega_n^0 & \omega_n^1 & \omega_n^2 & \omega_n^3 & \omega_n^4 & \omega_n^5 & \omega_n^6 & \omega_n^7 \\ \omega_n^0 & \omega_n^2 & \omega_n^4 & \omega_n^6 & \omega_n^8 & \omega_n^{10} & \omega_n^{12} & \omega_n^{14} \\ \omega_n^0 & \omega_n^3 & \omega_n^6 & \omega_n^9 & \omega_n^{12} & \omega_n^{15} & \omega_n^{18} & \omega_n^{21} \\ \omega_n^0 & \omega_n^4 & \omega_n^8 & \omega_n^{12} & \omega_n^{16} & \omega_n^{20} & \omega_n^{24} & \omega_n^{28} \\ \omega_n^0 & \omega_n^5 & \omega_n^{10} & \omega_n^{15} & \omega_n^{20} & \omega_n^{25} & \omega_n^{30} & \omega_n^{35} \\ \omega_n^0 & \omega_n^6 & \omega_n^{12} & \omega_n^{18} & \omega_n^{24} & \omega_n^{30} & \omega_n^{36} & \omega_n^{42} \\ \omega_n^0 & \omega_n^7 & \omega_n^{14} & \omega_n^{21} & \omega_n^{28} & \omega_n^{35} & \omega_n^{42} & \omega_n^{49} \end{pmatrix}$$

Wegen der Eigenschaft $\omega^8 = 1$ ergibt sich:

$$F_8 = \begin{pmatrix} \omega_n^0 & \omega_n^0 & \omega_n^0 & \omega_n^0 & \omega_n^0 & \omega_n^0 & \omega_n^0 & \omega_n^0 \\ \omega_n^0 & \omega_n^1 & \omega_n^2 & \omega_n^3 & \omega_n^4 & \omega_n^5 & \omega_n^6 & \omega_n^7 \\ \omega_n^0 & \omega_n^2 & \omega_n^4 & \omega_n^6 & \omega_n^0 & \omega_n^2 & \omega_n^4 & \omega_n^6 \\ \omega_n^0 & \omega_n^3 & \omega_n^6 & \omega_n^1 & \omega_n^4 & \omega_n^7 & \omega_n^2 & \omega_n^5 \\ \omega_n^0 & \omega_n^4 & \omega_n^0 & \omega_n^4 & \omega_n^0 & \omega_n^4 & \omega_n^0 & \omega_n^4 \\ \omega_n^0 & \omega_n^5 & \omega_n^2 & \omega_n^7 & \omega_n^4 & \omega_n^1 & \omega_n^6 & \omega_n^3 \\ \omega_n^0 & \omega_n^6 & \omega_n^4 & \omega_n^2 & \omega_n^0 & \omega_n^6 & \omega_n^4 & \omega_n^2 \\ \omega_n^0 & \omega_n^7 & \omega_n^6 & \omega_n^5 & \omega_n^4 & \omega_n^3 & \omega_n^2 & \omega_n^1 \end{pmatrix}$$

Sei nun $P_n \in \mathbb{C}^{n \times n}$ die Permutationsmatrix mit

$$P_n = [\mathbf{e}_0, \mathbf{e}_2, \ldots, \mathbf{e}_{n-2}, \mathbf{e}_1, \mathbf{e}_3, \ldots, \mathbf{e}_{n-1}]^T$$

Dann gilt:

$$P_8 F_8 = \begin{pmatrix} \omega_n^0 & \omega_n^0 & \omega_n^0 & \omega_n^0 & \omega_n^0 & \omega_n^0 & \omega_n^0 & \omega_n^0 \\ \omega_n^0 & \omega_n^2 & \omega_n^4 & \omega_n^6 & \omega_n^0 & \omega_n^2 & \omega_n^4 & \omega_n^6 \\ \omega_n^0 & \omega_n^4 & \omega_n^0 & \omega_n^4 & \omega_n^0 & \omega_n^4 & \omega_n^0 & \omega_n^4 \\ \omega_n^0 & \omega_n^6 & \omega_n^4 & \omega_n^2 & \omega_n^0 & \omega_n^6 & \omega_n^4 & \omega_n^2 \\ \omega_n^0 & \omega_n^1 & \omega_n^2 & \omega_n^3 & \omega_n^4 & \omega_n^5 & \omega_n^6 & \omega_n^7 \\ \omega_n^0 & \omega_n^3 & \omega_n^6 & \omega_n^1 & \omega_n^4 & \omega_n^7 & \omega_n^2 & \omega_n^5 \\ \omega_n^0 & \omega_n^5 & \omega_n^2 & \omega_n^7 & \omega_n^4 & \omega_n^1 & \omega_n^6 & \omega_n^3 \\ \omega_n^0 & \omega_n^7 & \omega_n^6 & \omega_n^5 & \omega_n^4 & \omega_n^3 & \omega_n^2 & \omega_n^1 \end{pmatrix}$$

$$P_8F_8 = \begin{pmatrix} \omega_m^0 & \omega_m^0 & \omega_m^0 & \omega_m^0 & \omega_m^0 & \omega_m^0 & \omega_m^0 & \omega_m^0 \\ \omega_m^0 & \omega_m^1 & \omega_m^2 & \omega_m^3 & \omega_m^0 & \omega_m^1 & \omega_m^2 & \omega_m^3 \\ \omega_m^0 & \omega_m^2 & \omega_m^0 & \omega_m^2 & \omega_m^0 & \omega_m^2 & \omega_m^0 & \omega_m^2 \\ \omega_m^0 & \omega_m^3 & \omega_m^2 & \omega_m^1 & \omega_m^0 & \omega_m^3 & \omega_m^2 & \omega_m^1 \\ \omega_n^0 \cdot \omega_m^0 & \omega_n^1 \cdot \omega_m^0 & \omega_n^2 \cdot \omega_m^0 & \omega_n^3 \cdot \omega_m^0 & \omega_n^4 \cdot \omega_m^0 & \omega_n^5 \cdot \omega_m^0 & \omega_n^6 \cdot \omega_m^0 & \omega_n^7 \cdot \omega_m^0 \\ \omega_n^0 \cdot \omega_m^0 & \omega_n^1 \cdot \omega_m^1 & \omega_n^2 \cdot \omega_m^2 & \omega_n^3 \cdot \omega_m^3 & \omega_n^4 \cdot \omega_m^0 & \omega_n^5 \cdot \omega_m^1 & \omega_n^6 \cdot \omega_m^2 & \omega_n^7 \cdot \omega_m^3 \\ \omega_n^0 \cdot \omega_m^0 & \omega_n^1 \cdot \omega_m^2 & \omega_n^2 \cdot \omega_m^0 & \omega_n^3 \cdot \omega_m^2 & \omega_n^4 \cdot \omega_m^0 & \omega_n^5 \cdot \omega_m^2 & \omega_n^6 \cdot \omega_m^0 & \omega_n^7 \cdot \omega_m^2 \\ \omega_n^0 \cdot \omega_m^0 & \omega_n^1 \cdot \omega_m^3 & \omega_n^2 \cdot \omega_m^2 & \omega_n^3 \cdot \omega_m^1 & \omega_n^4 \cdot \omega_m^0 & \omega_n^5 \cdot \omega_m^3 & \omega_n^6 \cdot \omega_m^2 & \omega_n^7 \cdot \omega_m^1 \end{pmatrix}$$

Wobei $n = 2m$. Es gilt also:

$$P_nF_n = \begin{pmatrix} F_m & F_m \\ F_m\Omega_n & -F_m\Omega_m \end{pmatrix} \text{ mit } \Omega_m = \begin{pmatrix} \omega_{2m}^0 & & & \\ & \omega_{2m}^1 & & \\ & & \ddots & \\ & & & \omega_{2m}^m \end{pmatrix}$$

Dabei ist $\omega_8^4 = -1$. Es gilt:

$$\begin{aligned} \begin{pmatrix} F_m & F_m \\ F_m\Omega_n & -F_m\Omega_m \end{pmatrix} &= \begin{pmatrix} F_m & \\ & F_m \end{pmatrix} \cdot \begin{pmatrix} \mathbf{E}_m & \mathbf{E}_m \\ \Omega_m & -\Omega_m \end{pmatrix} \\ &= \begin{pmatrix} F_m & \\ & F_m \end{pmatrix} \cdot \begin{pmatrix} \mathbf{E}_m & \\ & \Omega_m \end{pmatrix} \cdot \begin{pmatrix} \mathbf{E}_m & \mathbf{E}_m \\ \mathbf{E}_m & -\mathbf{E}_m \end{pmatrix} \end{aligned}$$

Beachte: Für $y = F_nc$ ist

$$y \begin{pmatrix} \mathbf{E}_m & \mathbf{E}_m \\ \mathbf{E}_m & -\mathbf{E}_m \end{pmatrix} \longrightarrow \begin{pmatrix} y_0 + y_m \\ \vdots \\ \frac{y_{m-1}+y_{n-1}}{y_0-y_m} \\ \vdots \\ y_{m-1} - y_{n-1} \end{pmatrix} \begin{pmatrix} \mathbf{E}_m & \\ & \Omega_m \end{pmatrix} \longrightarrow \begin{pmatrix} y_0 \\ \vdots \\ — \\ \vdots \\ \omega_n^i y_j^i \end{pmatrix} \begin{pmatrix} F_m & \\ & F_m \end{pmatrix} \longrightarrow \begin{pmatrix} c^0 \\ \vdots \\ — \\ \vdots \\ c^1 \end{pmatrix} = P_nc$$

Satz: Sei $n = 2m$, $m \in \mathbb{N}$. Es gilt:

$$P_nF_n = \begin{pmatrix} F_m & \\ & F_m \end{pmatrix} \cdot \begin{pmatrix} \mathbf{E}_m & \\ & \Omega_m \end{pmatrix} \cdot \begin{pmatrix} \mathbf{E}_m & \mathbf{E}_m \\ \mathbf{E}_m & -\mathbf{E}_m \end{pmatrix}$$

Bemerkung:

1. Die Fourier-Transformation von $y \in \mathbb{C}^n$ wird auf zwei Fourier-Transformationen von y^0, $y^2 \in \mathbb{C}^m$ zurückgespielt. Statt einmal mit F_n zu multiplizieren wird zweimal mit F_m multipliziert. Das benötigt nur halb so viele Rechenoperationen!

2. Multiplikation mit $\begin{pmatrix} \mathbb{E}_m & \mathbb{E}_m \\ \mathbb{E}_m & -\mathbb{E}_m \end{pmatrix}$ erfordert n Rechenoperationen.

3. Multiplikation mit $\begin{pmatrix} \mathbb{E}_m & \\ & \Omega_m \end{pmatrix}$ erfordert m Funktionsauswertungen ω_n^j und m (komplexe) Multiplikationen.

4. Die Permutationsmatrix sortiert gerade/ungerade Indizes.

Implementierung:

```
function c = myFFT(y);
n = length(y);
y = reshape(y, n, 1);
if n > 1 then
      omega = exp(sqrt(-1)*2*pi/2)
      y0 = reshape(y(1:n/2)+y(n/2+1:end),n/2,1);
      y1 = reshape((y(1:n/2)-y(n/2+1:end)).*(omega.^[0:n/2-1])',n/2,1);
      c(1:2:n,1) = myFFT(y0);
      c(2:2:n,1) = myFFT(y1);
else
      c = y;
end;
```

Bemerkungen:

1. Dieser Algorithmus ist nicht effizient und auch nicht unbedingt stabil, da die Potenzen von Omega jeweils neu berechnet werden.

2. im Zweier-System gilt:

$$\begin{pmatrix} P_2 & & & \\ & P_2 & & \\ & & P_2 & \\ & & & P_2 \end{pmatrix} \cdot \begin{pmatrix} \begin{pmatrix} P_4 & \\ & P_4 \end{pmatrix} \cdot \begin{pmatrix} P_8 \begin{pmatrix} 000 \\ 001 \\ 010 \\ 011 \\ 110 \\ 101 \\ 110 \\ 111 \end{pmatrix} \end{pmatrix} \end{pmatrix} = \begin{pmatrix} 000 \\ 100 \\ 010 \\ 110 \\ 001 \\ 101 \\ 011 \\ 111 \end{pmatrix}$$

   D.h. die Permutationsmatrizen bewirken eine Bitumkehr!

4 Quadratur

Seien gegeben $a, b \in \mathbb{R}$ mit $a < b$, eine integrierbare Funktion $f\colon [a,b] \to \mathbb{R}$, gesucht ist

$$I := I[f] := I[f,a,b] := \int_a^b f(x)\,dx$$

4.1 Allgemeines, Newton-Cotes-Formeln

Beispiel:

1. $a = 0, b = 1, f\colon [0,1] \to \mathbb{R}$ mit $f(x) = x^2$, d.h.

$$I[f] = \int_0^1 f(x)\,dx = \left[\frac{1}{3}x^3\right]_0^1 = \frac{1}{3}$$

2. $a = 0, b = 1, f\colon [0,1] \to \mathbb{R}$ mit $f(x) = e^{-x^{-2}}$. Hier ist

$$I[f] = \int_0^1 e^{-x^{-2}}\,dx = ?$$

Kriterien für ein numerisches Verfahren zur Berechnung von I:

1. Glattheit von f
2. Verfügbarkeit von f
3. Genauigkeit
4. Anzahl der zu behandelnden Fälle (finite Elemente)

Kriterien an eine Approximation J von I

1. Für vorgegebenes $\varepsilon > 0$: $|I - J| < \varepsilon$
2. I ist ein lineares Funktional, also sollte J auch eines sein
3. J besitzt eine endliche Rechenvorschrift
4. da f nur an endlich vielen Stellen auswertbar ist (maximal auf allen Maschinenzahlen), besitzt J folgende Gestalt:

$$J[f,a,b] = \sum_{j=0}^{n} \omega_j f(x_j) \text{ mit } x_j \in [a,b] \text{ und } \omega_j \in \mathbb{R}$$

Definition: Die folgende Formel bezeichnet man als *n-Punkt-Formel*:

$$J[f,a,b] = \sum_{j=0}^{n} \omega_j f(x_j) \text{ mit } x_j \in [a,b] \text{ und } \omega_j \in \mathbb{R}$$

Die Stellen x_j heißen *Knoten* und die ω_j *Gewichte*.

Idee: Ist f an gewissen Stellen x_j bekannt, ersetze f durch p_n, das Interpolationspolynom zu den Daten $(x_j, f(x_j))$:

$$p_n(x) = \sum_{j=0}^{n} f(x_j) \underbrace{L_j(x)}_{\text{Lagrange-Polynom}}$$

Dann ist

$$\begin{aligned} I[f] &\approx J[f] := I[p_n] = \int_a^b p_n(x)\,dx \\ &= \int_a^b \sum_{j=0}^{n} f(x_j) L_j(x)\,dx \\ &= \sum_{j=0}^{n} f(x_j) \underbrace{\int_a^b L_j(x)\,dx}_{=:\omega_j} \end{aligned}$$

Beispiel:

0. $n = 0$, $x_0 = \frac{a+b}{2}$. Dann ist $L_0(x) = 1$ und $\omega_0 = \int_a^b L_0(x)\,dx = \int_a^b dx = b - a$. Somit ist

$$R[f] := I_0[f] = \sum_{j=0}^{0} f(x_j) \cdot \omega_j = f\left(\frac{a+b}{2}\right)(b-a)$$

   Bezeichne $R[f]$ als die *Rechtecks-* oder *Mittelpunkts-* oder *Tangentenformel*

1. Für $n = 1$, $x_0 = a$ und $x_1 = b$ ist $L_0(x) = \frac{x-x_1}{x_0-x_1}$ und $L_1(x) = \frac{x-x_0}{x_1-x_0}$, damit ergeben sich $\omega_0 = \omega_1 = \frac{b-a}{2}$.

$$T[f] := I_1[f] = \sum_{j=0}^{1} f(x_j) \cdot \omega_j = \frac{b-a}{2}(f(a) + f(b))$$

   Dabei heißt $T[f]$ die *Trapezformel*.

2. Bei $n = 2$ mit $x_0 = a$, $x_1 = \frac{a+b}{2}$ und $x_2 = b$ ergibt sich $\omega_0 = \frac{b-a}{6} = \omega_2$ und $\omega_1 = \frac{2b-2a}{3}$. Also:

$$S[f] := I_2[f] = \sum_{j=0}^{2} f(x_j)\omega_j = \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right)$$

Hier ist $S[f]$ die *Simpsonformel* oder *Kepplersche Faßregel*.

Definition:

1. Seien $x_0, \ldots, x_n \in [a, b]$ mit $x_0 < \ldots < x_n$, dann heißt $(x_j)_0^n$ *Knotenfolge* (KF)
2. Ist $x_j - x_{j-1} = c$ für alle j, dann heißt die Knotenfolge *äquidistant*
3. Ist $(x_j)_0^n$ eine Knotenfolge und p_n das Interpolationspolynom zu $(x_j, f(x_j))$, dann heißt folgendes *Newton-Cotes-Formel:*

$$NC[f, a, b] = \int_a^b p_n(x)\,dx = \sum_{j=0}^{n} f(x_j)\omega_j \text{ mit } \omega_j = \int_a^b L_j(x)\,dx$$

4. Eine Newton-Cotes-Formel heißt offen (bzw. abgeschlossen), falls $a < x_0$ und $x_n < b$ (bzw. $x_0 = a$ und $x_1 = b$), beispielsweise ist R offen und T und S sind abgeschlossen.

4.1.1 Qualität der Formeln

Definition: Seien $(x_j)_0^n$ eine Knotenfolge und $J_n = J_n[f, a, b] = \sum_{j=0}^{n} f(x_j)\omega_j$ mit $\omega_j \in \mathbb{R}$ beliebig. Dann besitzt J_n die *Ordnung* $k \in \mathbb{N}$, falls $I[p] = J_n[p]$ für alle $p \in \Pi_{k-1}$. Besitzt J_n eine Ordnung $k \geq n + 1$, dann heißt J_n *interpolatorisch*.

Bemerkung: Sei $(\varphi_j, j = 0, \ldots, k-1)$ eine Basis von Π_{k-1}. Dann ist $I[\varphi_j] = J_n[\varphi_j]$ für $j = 0, \ldots, k-1$ genau dann, wenn J_n die Ordnung k besitzt, da J_n linear ist.

Beispiel: Betrachte die Trapezformel $T[f] = \frac{b-a}{2}(f(a) + f(b))$. O.B.d.A. sei

$a = 0$ und $b = 1$. Sei $\varphi_j(x) = x^j$, dann ist $I[\varphi_j] = \frac{1}{j+1}$. Es ist

$$\begin{aligned} T[\varphi_0] &= \frac{1}{2}(\varphi_0(0) + \varphi_0(1)) = 1 = I[\varphi_0] \\ T[\varphi_1] &= \frac{1}{2}(\varphi_1(0) + \varphi_1(1)) = \frac{1}{2} = I[\varphi_1] \\ T[\varphi_2] &= \frac{1}{2}(\varphi_2(0) + \varphi_2(1)) = \frac{1}{2} \neq I[\varphi_2] \end{aligned}$$

Damit hat die Trapezformel die Ordnung 2, die Trapezformel ist interpolatorisch.

Satz: Sei $(x_j)_0^n$ eine Knotenfolge und $J_n[f, a, b] = \sum_{j=0}^n f(x_j)\omega_j$ mit $\omega_j = \int_a^b L_j(x)\,dx$, dann ist J_n interpolatorisch.

Beweis: Für $p \in \Pi_n$ gilt $I[p] = J_n[p]$.

Beispiel: Rechtecksformel: $R[f] = (b-a)f(\frac{a+b}{2})$, sei wieder $a = 0$ und $b = 1$. Dann ist

$$\begin{aligned} R[\varphi_0] &= \varphi_0(\frac{1}{2}) = 1 = I[\varphi_0] \\ R[\varphi_1] &= \varphi_1(\frac{1}{2}) = \frac{1}{2} = I[\varphi_1] \\ R[\varphi_1] &= \varphi_1(\frac{1}{2}) = \frac{1}{2} = I[\varphi_1] \end{aligned}$$

Damit hat die Rechecksformel die Ordnung $k = 2$.

Satz: Sei J_n eine interpolatorische Quadraturformel und $f \in \mathcal{C}^{n+1}([a, b])$. Dann gilt:

$$|I[f] - J_n[f]| \leq \frac{\left\|f^{(n+1)}\right\|_{[a,b]}}{(n+1)!} \int_a^b \prod_{j=0}^n |x - x_j|\,dx$$

Dabei ist $\|g\|_{[a,b]} := \max\{\,|g(x)| \mid x \in [a, b]\}$.

Beweis: Sei p eine Interpolante zu f in $(x_j)_0^n$ und $e = f - p$. Dann ist

$$I[f] - J_n[f] = I[f] - J_n[p] = I[f] - I[p] = I[f - p] = I[e]$$

Dabei ist

$$e(x) \leq \frac{\left\|f^{(n+1)}\right\|_{[a,b]}}{(n+1)!} \prod_{j=0}^n |x - x_j|\,dx$$

Bemerkung: Interpolation mit Polynomen hohen Grades ist schlecht, dies gilt auch für die Quadratur.

Beispiel: (Simpson-Formel) Sei $n = 2$, drei Knoten, $f \in \mathcal{C}^3([a,b])$, es ist $S[f] = \frac{b-a}{2}\left(f(a) + 4f(\frac{a+b}{2}) + f(b)\right)$. Damit ist

$$|I[f] - S[f]| \leq \frac{\|f'''\|_{[a,b]}}{3!} \underbrace{\int_a^b \prod_{j=0}^{2} |x - x_j| \, dx}_{:=Q}$$

Für $x \in [a,b]$ gilt mit $x =: a + (b-a)y$ mit $y \in [0,1]$:

$$\begin{aligned}
Q &= \int_a^b |x-a| \left|x - \frac{a+b}{2}\right| |x-b| \, dx \\
&= \int_0^1 (b-a)y(b-a) \left|y - \frac{1}{2}\right| (b-a)(1-y)(b-a) \, dy \\
&= \int_0^1 (b-a)^4 \cdot y \cdot (1-y) \cdot \left|y - \frac{1}{2}\right| dy \\
&= (b-a)^4 \left(\int_0^{\frac{1}{2}} y(\frac{1}{2} - y)(1-y) \, dy + \int_{\frac{1}{2}}^1 y(y - \frac{1}{2})(1-y) \, dy \right) \\
&= (b-a)^4 \left(\frac{1}{64} + \int_0^{\frac{1}{2}} (1-z)(\frac{1}{2} - z)z \, dz \right) \\
&= \frac{(b-a)^4}{32}
\end{aligned}$$

Damit ergibt sich oben:

$$|I[f] - S[f]| \leq \frac{\|f'''\|_{[a,b]}}{3!} \int_a^b \prod_{j=0}^{2} |x - x_j| \, dx \leq \frac{(b-a)^4}{192} \|f'''\|_{[a,b]}$$

Die Simpsonformel genügt sogar der folgenden Formel, falls $f \in \mathcal{C}^4([a,b])$ ist:

$$|I[f] - S[f]| \leq \frac{(b-a)^5}{2880} \left\|f^{(4)}\right\|_{[a,b]}$$

Dies folgt aus der Symmetrie der Knoten bezüglich der Intervallmitte und $n = 2$.

4.2 Zusammengesetzte Formeln

Besser als ein Polynom hohen Grades sind viele Polynome kleinen Grades (Idee der Splines).

Beispiele:

1. Betrachte die zusammengesetzte Rechtecksformel: Es ist $R[f,a,b] = f(\frac{a+b}{2})(b-a)$, benutze dies für n Intervalle, $y_k := a + k \cdot h$, $h = \frac{b-a}{n}$. Dann ist

$$\begin{aligned} R_n[f,a,b] &= \sum_{k=1}^{n} R[f, y_{k-1}, y_k] \\ &= \sum_{k=1}^{n} (y_k - y_{k-1}) \cdot f(\frac{y_{k-1}+y_k}{2}) \\ &= \sum_{k=1}^{n} h \cdot f(\underbrace{a + \frac{h}{2}(2k-1)}_{:=x_k}) \\ &= \sum_{j=1}^{n} \omega_j f(x_j) \text{ mit } \omega_j = h \end{aligned}$$

   Dies kann man interpretieren als Riemann-Summe (siehe Analysis), insbesondere ist $R_n[f] \to I[f]$, falls f Riemann-integrierbar.

2. Die zusammengesetzte Trapezformel: Es ist $T[f,b,a] = \frac{b-a}{2}(f(a)+f(b))$, zerteile dies in n Intervalle und sei (mit $y_k := a + k \cdot h$ und $h = \frac{b-a}{n}$):

$$T_n[f,a,b] = \sum_{k=1}^{n} T[f, y_{k-1}, y_k]$$

   Es gilt: $\frac{y_k+y_k}{2} = a + h(2k-1)$, $y_k - y_{k-1} = h$. Damit ergibt sich:

$$\begin{aligned} T_n[f,a,b] &= \sum_{k=1}^{n}\sum_{k=1}^{n} \frac{y_k - y_{k-1}}{2}(f(y_{k-1}) + f(y_k)) \\ &= \frac{h}{2}\left(f(y_0) + 2\sum_{k=1}^{n-1} f(y_k) + f(y_n) \right) \\ &= \sum_{j=0}^{n} \omega_j f(x_j) \end{aligned}$$

   mit $x_j := y_k$ und $\omega_j = \frac{h}{2}$ für $j = 0, n$ und $\omega_j = h$ sonst.

4.2.1 Konvergenzuntersuchungen

VORAUSSETZUNG:

1. Knotenfolge $(x_j^h)_0^n$
2. Gewichtsfolge $(\omega_j^n)_0^n$
3. Integrationsformel $J_n[f] = \sum_{j=0}^n \omega_j^n f(x_j^n)$
4. Für alle Polynome φ soll gelten: $J_n[\varphi] \to I[\varphi]$

Satz: Sind obige Voraussetzungen erfüllt und gibt es eine Konstante $c \in \mathbb{R}$ mit

$$\sum_{j=0}^{n} \left|\omega_j^n\right| \leq c \; \forall \; n = 0, 1, \ldots,$$

dann gilt:

$$J_n[f] \longrightarrow I[f] \; \forall \; f \in \mathcal{C}^0([a,b])$$

Beweis: Nach dem Satz von Weierstraß gibt es zu jedem $\varepsilon > 0$ und $f \in \mathcal{C}^0([a,b])$ ein Polynom p mit $\|f - p\|_\infty < \varepsilon$. Weiter gibt es nach Voraussetzung ein $n_0 \in \mathbb{N}$ mit $|I[p] - J_n[P]| \leq \varepsilon$ für alle $n \geq n_0$. Sei $n \geq n_0$, dann gilt:

$$\begin{aligned} |I[f] - J_n[f]| &= |I[f-p] + I[p] - J_n[f-p] - J_n[p]| \\ &\leq \underbrace{|I[f-p]|}_{\leq \varepsilon(b-a)} + \underbrace{|I[p] - J_n[p]|}_{\leq \varepsilon} + \underbrace{|J_n[f-p]|}_{\varepsilon \sum |\omega_j^n|} \\ &\leq c \cdot \varepsilon \;\text{ für eine Konstante }\; c \end{aligned}$$

Also ist $J_n[f] \longrightarrow I[f]$.

KOROLLAR: Gilt $\omega_j^n \geq 0$ für alle $j = 0, \ldots, n$ und $n \in \mathbb{N}$, dann gilt $J_n[f] \longrightarrow I[f]$ für alle $f \in \mathcal{C}^0([a,b])$

Beweis: Es ist

$$J_n[1] = \sum_{j=0}^{n} \omega_j^n = \sum_{j=0}^{n} \left|\omega_j^n\right| \longrightarrow I[1] = b - a$$

Bemerkung: Die Newton-Cotes-Formeln haben nicht notwendigerweise positive Gewichte, d.h. im allgemeinen kann keine Konvergenz gezeigt werden. Die zusammengesetzten Rechteck-, Trapez- und Simpon-Formeln konvergieren, da alle Gewichte positiv sind.

4.3 *Einschub:* Euler-MacLaurinsche Summenformel

Definition: Die Bernoulli-Polynome sind definiert durch

$$\begin{aligned} B_0(x) &= 1 \\ B_k'(x) &= kB_{k-1}(x) \text{ mit } \int_0^1 B_k(x)\,dx = 0 \end{aligned}$$

Bemerkung: Für $k \geq 1$ ist: $B_k(x) = c + k \cdot \int_0^x B_{k-1}(t)\,dt$ mit $c \in \mathbb{R}$ so, daß $\int_0^1 B_k(x)\,dx = 0$.

Beispiel:

$$\begin{aligned} B_1(x) &= c + \int_0^x B_0(t)\,dt = c + \int_0^x dt = c + x \text{ mit } c = -\frac{1}{2} \\ &= x - \frac{1}{2} \\ B_2(x) &= c + 2\int_0^x B_1(t)\,dt = c + \int_0^x \left(t - \frac{1}{2}\right) dt = x^2 - x + c \text{ mit } c = \frac{1}{6} \\ &= x^2 - x + \frac{1}{6} \end{aligned}$$

Satz: Eigenschaften der Bernoulli-Polynome

1. $B_k(0) = B_k(1)$ für $k \geq 2$
2. $P_k(x) := B_k\left(x + \frac{1}{2}\right)$ ist (un)gerade falls k (un)gerade ist
3. $B_{2k+1}(0) = B_{2k+1}(1) = 0$ für $k \geq 1$

Beweis:

1. Es gilt für $k \geq 2$:

$$\begin{aligned} B_k(1) - B_k(0) &= [B_k(x)]_0^1 = \int_0^1 B_k'(x)\,dx \\ &= k \cdot \int_0^1 B_{k-1}(x)\,dx = k \cdot 0 = 0 \end{aligned}$$

2. P_0 ist gerade. *Zeige:* Falls P_{2k} gerade ist, so ist P_{2k+1} ungerade (andere Richtung analog). Sei also P_{2k} eine gerade Funktion. Dann gilt: $P_{2k+1}'(x) = (2k+1)P_{2k}(x)$ und $\int_{-\frac{1}{2}}^{\frac{1}{2}} P_{2k+1}(x)\,dx = 0$. Betrachte $q(x) :=$

$(2k+1)\int_0^x P_{2k}(t)\,dt$, dieses Polynom ist ungerade. Weiter ist $\int_{-\frac{1}{2}}^{\frac{1}{2}} q(x)\,dx = 0$ (da q ungerade), damit ist $P_{2k+1} = q$.

3. Es gilt (da B_{2k+1} ungerade):

$$B_{2k+1}(0) \overset{(1)}{=} B_{2k+1}(0) = P_{2k+1}\left(-\frac{1}{2}\right) = -P_{2k+1}\left(\frac{1}{2}\right) = -B_{2k+1}(0)$$

Also ist $B_{2k+1}(0) = B_{2k+1}(1) = 0$.

Wir definieren $s_k\colon \mathbb{R} \to \mathbb{R}$ durch

$$s_k(x) = B_k(x-j) \text{ falls } x \in [j, j+1[\ \forall\ j \in \mathbb{Z}$$

Eigenschaften von s_k: Es gilt (mit $I_j := [j, j+1[$):

$$x+1 \in I_{j+1} \Longleftrightarrow x \in I_j$$

Für $x \in I_{j-1}$ gilt also:

$$s_k(x+1) = B_k((x+1)-j) = B_k(x-(j-1)) = s_k(x)$$

Damit ist die Funktion 1-periodisch. Für $k \geq 2$ ist s_k stetig auf $\mathbb{R}$, denn:

$$\lim_{x\downarrow j} s_k(x) = \lim_{x\downarrow j} B_k(x-j) = B_k(0) \overset{k\geq 2}{=} B_k(1) = \lim_{x\uparrow j} B_k(x-(j-1)) = \lim_{x\uparrow j} s_k(x)$$

Die Funktion s_1 ist stückweise linear, $B_1(x) = x - \frac{1}{2}$. Weiter ist

$$\begin{aligned} s_k(0) &= s_k(1) = \ldots = s_k(n) \\ s_{2k+1}(0) &= B_{2k+1}(0) = 0 \\ s_{2k}(0) &= B_{2k}(0) := (-1)^{k+1}\beta_k \text{ mit Bernoulli-Zahl } \beta_k \end{aligned}$$

Satz: Euler-MacLaurinsche Summenformel: Sei $f \in \mathcal{C}^{2m+2}([a,b])$, $h = \frac{b-a}{n}$, $n \in \mathbb{N}$. Dann gilt:

$$\int_a^b f(x)\,dx = T_n[f,a,b] - \sum_{p=1}^{m} \underbrace{(-1)^{p+1}\frac{\beta_p}{(2p)!}\left(f^{(2f-1)}(b) - f^{(2p-1)}(a)\right)}_{=:\tau_p} h^{2p} + R_{m+1}$$

wobei

$$R_{m+1} = \frac{2^{2m+2}}{(2m+2)!}\int_a^b s_{2m+2}\left(\frac{x-a}{a}\right) f^{(2m+2)}(x)\,dx$$

Beweis: Zur Vereinfachung: $g\colon [0,n] \to \mathbb{R}$ mit $g(t) = f(a + t \cdot h)$ mit $g \in \mathcal{C}^{2m+2}([0,n])$. Betrachte

$$\begin{aligned}
\int_0^n s_1(t)g'(t)\,dt &= \sum_{j=0}^{n-1} \int_j^{j+1} s_1(t)g'(t)\,dt \\
&= \sum_{j=0}^{n-1} \int_0^1 \underbrace{s_1(t)}_{u} \underbrace{g'(t+j)}_{v'}\,dt \\
&= \sum_{j=0}^{n-1} \left([s_1(t) \cdot g(t+j)]_0^1 - \int_0^1 g(t+j)\,dt \right) \\
&= \left(\sum_{j=0}^{n-1} \frac{1}{2}\left(g(t+j) - g(j)\right) \right) - \left(\int_0^n g(t+j)\,dt \right) \\
&= T_n[g,0,n] - I[g]
\end{aligned}$$

Betrachte weiter (mit $s'_{k+1}(x) = (k+1)s_k(x)$):

$$\frac{1}{k!}\int_0^n \underbrace{s_k(t)}_{u'} \underbrace{g^{(k)}(t)}_{v}\,dt = \frac{1}{(k+1)!}[s_{k+1}(t)g^{(k)}(t)]_0^n - \frac{1}{(k+1)!}\int_0^n s_{k+1}(t)\cdot g^{(k+1)}(t)\,dt$$

$\underline{\text{mit } k = 2p}$

$$\frac{1}{(2p)!}\int_0^n s_{2p}(t)g^{(2p)}(t)\,dt = -\frac{1}{(2p+1)!}\int_0^n s_{2p+1}(t)\cdot g^{(2p+1)}(t)\,dt$$

$\underline{\text{mit } k = 2p-1}$

$$\frac{1}{(2p)!}\int_0^n s_{2p}(t)g^{(2p)}(t)\,dt = \frac{1}{(2p)!}(-1)^{p+1}\beta_p\left(g^{(2p-1)}(n) - g^{(2p-1)}(0)\right) - \frac{1}{(2p)!}\int_0^n s_{2p}(t)\cdot g^{(2p)}(t)\,dt$$

Nach $2m$-maliger partieller Integration ist

$$\begin{aligned}
\int_0^n s_1 g'\,dt &= \sum_{p=1}^{m}(-1)^{p+1}\frac{\beta_p}{(2p)!}(g^{(2p-1)}(n) - g^{(2p-1)}(0)) \\
&\quad + \underbrace{\frac{1}{(2m+1)!}\int_0^n s_{2m+1}(t)g^{(2m+1)}(t)\,dt}_{R_{m+1}}
\end{aligned}$$

Letztmalige partielle Integration liefert

$$R_{m+1} = \frac{1}{(2m+2)!}\underbrace{[s_{2m+2}(t)g^{(2m+1)}(x)]_0^n}_{=0} - \int_0^n s_{2m+2}(t)g^{(2m+2)}(t)\,dt$$

Weiter ist

$$\begin{aligned}\int_0^n g(t)\,dt &= \int_0^n f(a+t\cdot h)\,dt = \frac{1}{h}\int_a^b f(x)\,dx\\ g'(t) &= \frac{d}{dt}[f(a+h\cdot t)] = hf'(x)\\ g^{(k)}(t) &= h^k\cdot f^{(k)}(x)\end{aligned}$$

Schließlich ist

$$\frac{1}{2}\sum_{j=0}^{n-1}(g(j+1)-g(j)) = \frac{1}{2}\sum_{j=0}^{n-1}(f(a+(j+1)h)+f(a+jh)) = T_n[f,a,b]$$

□

4.4 Extrapolation der Maschenweite

(ROMBERG) Angenommen, $J(h) = J_n[f,a,b] = I[f] + c_1h^{k_1} + c_2h^{k_2}$ mit $h = \frac{1}{n}$ und $k_1 < k_2 < \ldots$. Dann gilt:

$$J(\frac{h}{2}) = I[f] + c_1 2^{-k_1}h^{k_1} + c_2\cdot 2^{-k_2}\cdot h^{k_2} + \ldots$$

Also ergibt sich:

$$\begin{aligned}J(h) - I[f] &= c_1h^{k_1} + c_2h^{k_2} + \ldots\\ 2^{k_1}(J(\frac{h}{2}) - I[f]) &= c_1h^{k_1} + c_2 2^{k_1-k_2}h^{k_2} + \ldots\\ 2^{k_1}J(\frac{h}{2}) - J(h) - (2^{k_1}-1)I[f] &= \tilde{c}_2h^{k_2}\end{aligned}$$

Damit besitzt die folgende Formel eine höhere Fehlerordnung:

$$\hat{J}\left(\frac{h}{2}\right) = \frac{2^{k_1}J(\frac{h}{2}) - J(h)}{2^{k_1}-1} = I[f] + d_2h^{k_2} + \ldots$$

Beispiel: Extrapolierte Trapezregel: Mit der Euler-MacLaurinschen Summenformel gilt: Ist $f \in \mathcal{C}^{2n+2}([a,b])$, $h = \frac{b-a}{n}$, $n \in \mathbb{N}$, so gilt:

$$T(h) = T_n[f,a,b] = I[f] + \sum_{p=1}^{m}\tau_p h^{2p} + \alpha_{m+1}(h)\cdot h^{2m+2}$$

wobei τ_j unabhängig von h und die Funktion α ist beschränkt durch

$$|\alpha_{m+1}(h)| \leq \frac{2c}{(2m+2)!} \left\| f^{(2m+2)} \right\|_{[a,b]}$$

Bei Vernachlässigung des Restgliedes:

$$T(h) \dot{\approx} I[f] + \sum_{p=1}^{m} \tau_p h^{2p} \Longrightarrow T(0) \doteq I[f]$$

Idee: Für Werte $h_0, \ldots, h_m > 0$ bestimme $T(h_0), \ldots, T(h_m)$ und das Interpolationspolynom q zu den Daten $(h_j^2, T(h_j))$ für $j = 0, \ldots, m$, dann ist $q(0)$ eine gute Approximation an I.

Beispiel: Für $n_j := 2^j$, $h_j := \frac{b-a}{n_j}$, $T_{j,0} := T_{n_j}[f, a, b]$ berechne:

```
for k = 1 to m
  for j = k to m
    T_{j,k} := T_{j,k-1} + (T_{j,k-1} - T_{j-1,k-1}) · ( h_{j-k}/h_j - 1 )^{-1}
  end
end
T̂ = T_{m,m}
```

4.5 Adaptive Formeln am Beispiel der Simpsonformel

Sei f auf $[a, b]$ mit Genauigkeit $\varepsilon > 0$ zu integrieren. Betrachte die Simpsonformel:

$$S_j := S[f, x_j, x_{j+1}] = \frac{h_j}{6} \left(f(x_j) + 4f(\frac{x_j + x_{j+1}}{2}) + f(x_{j+1}) \right)$$

Besser ist:

$$\begin{aligned} Q_j &:= S_2[f, x_j, x_{j+1}] = S[f, x_j, x_{j+\frac{1}{2}}] + S[f, x_{j+\frac{1}{2}}, x_{j+1}] \\ &= \frac{h_j}{12} \Big(f(x_j) + 4f(x_j + \frac{h_j}{4}) + f(x_j + \frac{h_j}{2}) \\ &\qquad + f(x_j + \frac{h_j}{2}) + 4f(x_j + \frac{3h_j}{4}) + f(x_{j+1}) \Big) \end{aligned}$$

Damit ist

$$\begin{aligned} I_j - S_j &= c h_j^5 + \ldots \\ I_j - Q_j &= 2c \left(\frac{h_j}{2} \right)^5 + \ldots \\ 16(I_j - Q_j) &= c h_j^5 + \ldots \\ &\doteq I_j - S_j = (I_j - Q_j) + (Q_j - S_j) \end{aligned}$$

Damit ist $I_j - Q_j = \frac{1}{15}(Q_j - S_j) + \ldots$. Also:

$$\left|\sum_{j=1}^{L}(I_j - Q_j)\right| \leq \sum_{j=1}^{L}|I_j - Q_j| \leq \frac{1}{15}\sum_{j=1}^{L}|Q_j - S_j| \leq \varepsilon$$

Wählen wir die Knoten so, daß $|Q_j - S_j| \leq \frac{15h_j\varepsilon}{b-a}$ gilt, dann ist

$$\left|I[f] - \sum_{j=1}^{L} Q_f[f]\right| \leq \frac{1}{15}\frac{15\varepsilon}{b-a}\sum_{j=1}^{L} h_j = \varepsilon$$

4.6 Gauß-Quadratur

Wie bisher: $J_n = J_n[f] = J_n[f, a, b] = \sum_{j=0}^{n} f(x_j)\omega_j$ besitzt die Ordnung k genau dann, wenn $J_n[\varphi] = I[\varphi]$ für alle $\varphi \in \Pi_{k-1}$.

Frage: Wann besitzt J_n eine maximale Ordnung?

Bemerkung: Falls $(x_j)_0^n$ Knotenfolge ist mit Gewichten $\omega_j := \int_a^b L_j(x)\,dx$, dann ist die Ordnung von J_n kleinergleich $n+1$. Jetzt soll auch eine optimale Verteilung der Knoten bestimmt werden.

Satz: Die Ordnung von J_n ist kleinergleich $2n+2$.

Beweis: Sei $q(x) := \prod_{j=0}^{n}(x - x_j) \in \Pi_{n+1}$, es ist $q \neq 0$, also $I[q^2] > 0$. Es ist $q^2 \in \Pi_{2n+2}$, nun ist

$$J_n[q^2] = \sum_{j=0}^{n} q^2(x_j) \cdot \omega_j = 0 \neq I[q^2]$$

Satz: Die Ordnung von J_n ist gleich $2n+2$ genau dann, wenn

$$q(x) = \prod_{j=0}^{n}(x - x_j) \perp \Pi_n \text{ und } \omega_j = \int_a^b L_j(x)\,dx$$

Beweis:

„⇒“ Sei die Ordnung von J_n also $2n+2$, d.h. $I[\varphi] = J_n[\varphi]$ für alle $\varphi \in \Pi_{2n+1}$. Also ist $p \in \Pi_n$ und $pq \in \Pi_{2n+1}$, also

$$I[pq] = J_n[pq] = \sum_{j=0}^{n} p(x_j)q(x_j)\omega_j = 0$$

bzw.

$$0 = I[pq] = \int_a^b p(x)q(x)\,dx = \langle p, q\rangle$$

d.h. $q \perp p$ für alle $p \in \Pi_n$

„⇐“ Sei also $\langle p, q\rangle = 0$ für alle $p \in \Pi_n$. Sei $\psi \in \Pi_{2n+1}$, dann gibt es Polynome $p, r \in \Pi_n$ mit $\psi = qp + r$. Es gilt:

$$I[\psi] = I[qp] + I[r] = I[r] \text{ denn } I[qp] = \langle q, p\rangle = 0$$

Nach Konstruktion der ω_j ist J_n interpolatorisch, $J_n[qp] = \sum_{j=0}^n q(x_j)p(x_j)\omega_j = 0$, d.h.

$$I[\psi] = I[r] = J_n[r] = J_n[qp + r] = J_n[\psi]$$

4.6.1 Orthogonale Polynome

Wir betrachten das Skalarprodukt $\langle \cdot, \cdot\rangle : \Pi^2 \to \mathbb{R}$:

$$\langle \varphi, \psi\rangle = \int_a^b \varphi(x)\psi(x)\varrho(x)\,dx$$

wobei $\varrho\colon [a, b] \to \mathbb{R}$ eine Gewichtsfunktion sei.

Definition: Eine Funktion $\varrho\colon [a, b] \to \mathbb{R}$ mit

1. $\varrho(x) \geq 0$ für alle $x \in [a, b]$
2. $\int_a^b \varphi(x)\varrho(x)\,dx$ existiert für alle Polynome ψ
3. Wenn $\psi \in \Pi$ ohne Vorzeichenwechsel ist, so ist $\int_a^b \psi(x)\varrho(x)\,dx \neq 0$

heißt Gewichtsfunktion.

Beispiel:

1. $\varrho(x) = 1$
2. für $a = -1$, $b = 1$: $\varrho(x) = \frac{1}{\sqrt{1-x^2}}$

Definition: Sei $\langle \cdot, \cdot\rangle$ ein Skalarprodukt, p, q zwei Polynome. Gilt $\langle p, q\rangle = 0$, so heißen p, q *orthogonal*, in Zeichen $p \perp q$. Gilt $q \perp p$ für alle $p \in \Pi_n$, dann heißt q orthogonal zu Π_n, in Zeichen $q \perp \Pi_n$. Die Polynome $(\psi_j)_{j\in\mathbb{N}}$ mit $\psi_j \in \Pi_j$ und $\psi_j \perp \Pi_{j-1}$ heißen die *orthogonalen Polynome*, gilt zudem $\langle \psi_j, \psi_j\rangle = 1$; dann heißen die Polynome auch *orthonormal*.

Satz: Sei ψ_n ein Orthogonalpolynom. Dann gilt: $\psi_n = 0$ oder $\psi_n(x) = \prod_{j=1}^{n}(x - x_j)$, wobei die x_j paarweise verschieden sind mit $x_j \in]a, b[$.

Beweis: Sei $\psi_n \neq 0$ und seien $x_1, \dots x_k \in]a, b[$ Nullstellen von ψ_n mit Vorzeichenwechsel. Wir zeigen: $k = n$. Sei dazu $s(x) = \prod_{j=1}^{k}(x - x_j)$ bzw. $s = 1$ falls $k = 0$. Es ist $s \in \Pi_n$ und $\psi \cdot s$ besitzt auf $]a, b[$ keinen Vorzeichenwechsel. Es ist

$$0 \neq \int_a^b \psi_n(x) s(x) \varrho(x)\, dx = \langle \psi_n, s \rangle$$

Damit ist $s \notin \Pi_{n-1}$, es gilt aber $s \in \Pi_n$, also hat s den Grad n.

Bemerkung: Mit dem Gram-Schmidtschen Orthogonalisierungsverfahren (GSOGV) können die Orthogonalpolynome prinzipiell berechnet werden. Es geht aber viel einfacher:

Satz: (Drei-Term-Rekursion) Orthogonalpolynome genügen einer Drei-Term-Rekursion, d.h.

$$x\psi_n = \beta_n \psi_{n-1} + \alpha_n \psi_n + \gamma_n \psi_{n+1}$$

Wobei $\psi_{-1} = 0$ und $\psi_0 = c \neq 0$ sowie $\gamma_n \in \mathbb{R} \setminus \{0\}$. Weiter ist

$$\alpha_n = \frac{\langle x\psi_n, \psi_n \rangle}{\langle \psi_n, \psi_n \rangle} \text{ und } \beta_n = \gamma_{n-1} \frac{\langle \psi_n, \psi_n \rangle}{\langle \psi_{n-1}, \psi_{n-1} \rangle}$$

Beweis: Die Orthogonalpolynome $\psi_0, \dots, \psi_n$ bilden eine Basis von Π_n. Wir orthogonalisieren $x\psi_n(x)$ gegen Π_n:

$$q(x) = x\psi_n(x) - \sum_{j=0}^{n} h_{j,n} \psi_j(x) \text{ mit } h_{j,n} = \frac{\langle x\psi_n, \psi_j \rangle}{\langle \psi_j, \psi_j \rangle}$$

Nun gilt für $j < n - 1$:

$$\langle \psi_j, x\psi_n \rangle = \int_a^b (\psi_j(x)) \psi_n(x) \varrho(x)\, dx = \langle x\psi_j, \psi_n \rangle = 0$$

Es ist $x\psi_j(x) \in \Pi_{n-1}$ und $\psi_n \perp \Pi_{n-1}$, d.h. $h_{j,n} = 0$. Für $j = n - 1$ setze

$$\beta_n := h_{n-1,n} = \frac{\langle \psi_{n-1}, x\psi_n \rangle}{\langle \psi_{n-1}, \psi_{n-1} \rangle}$$

Für $j = n$ setze

$$\alpha_n := h_{n,n} = \frac{\langle \psi_n, x\psi_n \rangle}{\langle \psi_n, \psi_n \rangle}$$

Da $\psi_n \in \Pi_n$ und $\psi_n \notin \Pi_{n-1}$, gilt $q \in \Pi_{n+1}$ und $q \notin \Pi_n$. Also gilt: $\gamma_n \psi_{n+1} = q$ mit $\psi_n \neq 0$. Damit ist

$$\gamma_n \psi_{n+1}(x) = q(x) = x\psi_n(x) - \alpha_n \psi_n(x) - \beta_n \psi_{n-1}(x)$$

Da

$$\begin{aligned} \langle \psi_{n-1}, x\psi_n \rangle &= \langle x\psi_{n-1}, \psi_n \rangle \\ &= \langle \gamma_{n-1}\psi_n + \alpha_{n-1}\psi_{n-1} + \beta_{n-1}\psi_{n-2}, \psi_n \rangle \\ &= \gamma_{n-1} \langle \psi_n, \psi_n \rangle \end{aligned}$$

4.6.2 Stieltjes Algorithmus

Seien $\gamma_0, \ldots,$ gegeben.

$\psi_{-1} := 0;$
$n := 0;$
$\psi_n := \gamma_0 \neq 0;$
$\beta_n := 0;$
`for` $n = 0, 1, \ldots$
  $\alpha_n := \frac{\langle \psi_n, x\psi_n \rangle}{\langle \psi_n, \psi_n \rangle}$
  $\gamma_n \psi_{n+1}(x) = (x - \alpha_n)\varphi_n(x) - \beta_n \varphi_{n-1}(x)$
  $\beta_{n+1} := \gamma_n \frac{\langle \psi_{n+1}, \psi_{n+1} \rangle}{\langle \psi_n, \psi_n \rangle}$
`end`;

Stieltjes Algorithmus für orthonormale Polynome:

$\psi_{-1} := 0;$
$n := 0;$
$\varphi_n := \frac{1}{\sqrt{\langle 1,1 \rangle}}$
`for` $n = 0, 1, \ldots$
  $\alpha_n := \langle \varphi_n, x\varphi_n \rangle$
  $\tilde{\varphi}_{n+1}(x) = (x - \alpha_n)\varphi_n(x) - \beta_n \varphi_{n-1}(x)$
  $\beta_{n+1} := \sqrt{\langle \varphi_{n+1}, \varphi_{n+1} \rangle}$
  $\varphi_{n+1} := \frac{1}{\beta_{n+1}} \tilde{\varphi}_{n+1}$
`end`;

Beispiele:

- Legendre-Polynome: Es sei $\varrho = 1$, d.h. $\langle p, q \rangle := \int_{-1}^{1} p(x)q(x)\,dx$. Wähle $\gamma_j := 1$, dann berechnet Stieltjes Algorithmus: Startwerte $\gamma_j := 0$,

$\beta_0 := 0$, $\psi_0(x) := 1$. Es gilt (da ψ_n^2 symmetrisch):

$$\alpha_n = \langle x\psi_n, \psi_n\rangle = \int_{-1}^{1} x\psi_n^2(x)\,dx = 0$$

Es gilt:

$$\begin{aligned}
\langle\psi_0,\psi_0\rangle &= \int_{-1}^{1} 1\cdot 1\,dx = 2\\
\psi_1(x) &= x\psi_0(x) - \underbrace{\beta_0\psi_{-1}(x)}_{=0} = x\\
\langle\psi_1,\psi_1\rangle &= \int_{-1}^{1} x^2\,dx = \frac{2}{3}\\
\beta_1 &= \frac{\langle\psi_1,\psi_1\rangle}{\langle\psi_0,\psi_0\rangle} = \frac{1}{3}\\
\psi_2(x) &= x\psi_1(x) - \beta_1\psi_0(x) = x^2 - \frac{1}{3}\\
\langle\psi_2,\psi_2\rangle &= \int_{-1}^{1} (x^2-\frac{1}{3})^2\,dx = \frac{8}{45}\\
\beta_2 &= \frac{\langle\psi_2,\psi_2\rangle}{\langle\psi_1,\psi_1\rangle} = \frac{8}{45}\cdot\frac{3}{2} = \frac{4}{15}\\
\psi_3(x) &= x\psi_2(x) - \beta_2\psi_1(x) = x^3 - \frac{3}{5}x
\end{aligned}$$

- Чебышёв-Polynome: Es sei $\varrho(x) = \frac{1}{\sqrt{1-x^2}}$, d.h. $\langle p,q\rangle = \int_{-1}^{1} \frac{p(x)q(x)}{\sqrt{1-x^2}}\,dx$. Die Polynome $c_n(x) := \cos(n\cdot\arccos(x))$ für $n\in\mathbb{N}_0$ bilden ein Orthogonalsystem (mit $x = \cos y$)[13]:

$$\begin{aligned}
\langle c_j, c_k\rangle &= \int_{-1}^{1} \frac{\cos(j\cdot\arccos(x))\cdot\cos(k\cdot\arccos(x))}{\sqrt{1-x^2}}\,dx\\
&= \int_0^{\pi} \cos(jy)\cdot\cos(ky)\,dy\\
&\overset{(\star)}{=} \frac{1}{2}\int_0^{\pi}\cos((j-k)y)\,dy + \frac{1}{2}\int_0^{\pi}\cos((j+k)y)\,dy
\end{aligned}$$

Damit gilt:

$$\langle c_j, c_k\rangle = \begin{cases} 0 & \text{falls} \quad j\neq k\\ \pi & \text{falls} \quad j=k=0\\ \frac{\pi}{2} & \text{falls} \quad j=k\neq 0\end{cases}$$

[13] $(\star)$ mit Additionstheorem $2\cos u\cos c = \cos(u+v)+\cos(u-v)$

Weiter gilt: $\langle xc_n, c_n\rangle = 0$, da der Integrand punktsymmetrisch bezüglich x ist. Die Чебышёв-Polynome genügen der Dreiterm-Rekursion

$$xc_n(x) = \frac{1}{2}c_{n+1}(x) + \frac{1}{2}c_{n-1}$$

Definiere $T_0(x) := 1$ und $T_1(x) := x$ und weiter $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$, dann ist $T_n = \delta_n c_n$ mit $\delta_n \in \mathbb{R} \setminus \{0\}$. Also ist $\gamma_0 = 1$, weiter ist $T_1(x) = xT_0(x)$ und $\beta_n = \gamma_n = \frac{1}{2}$ für $n > 0$; zudem: $\alpha_n = 0$ für alle n.

Die Drei-Term-Rekursion ergibt Gleichungen:

$$\begin{aligned} x\psi_0(x) &= \beta_0\psi_{-1}(x) + \alpha_0\psi_0(x) + \gamma_0\psi_1(x) \\ x\psi_1(x) &= \beta_1\psi_0(x) + \alpha_1\psi_1(x) + \gamma_1\psi_2(x) \\ \vdots &\quad \vdots \\ x\psi_{n-1}(x) &= \beta_{n-1}\psi_{n-2}(x) + \alpha_{n-1}\psi_{n-1}(x) + \gamma_{n-1}\psi_n(x) \end{aligned}$$

Damit ergibt sich ein Gleichungssystem:

$$x\begin{pmatrix} \psi_0(x) \\ \vdots \\ \psi_{n-1}(x) \end{pmatrix} = \begin{pmatrix} \alpha_0 & \gamma_0 & & & \\ \beta_1 & \alpha_1 & \gamma_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \gamma_{n-2} \\ & & & \beta_{n-1} & \alpha_{n-1} \end{pmatrix} \cdot \begin{pmatrix} \psi_0(x) \\ \psi_1(x) \\ \vdots \\ \vdots \\ \psi_{n-1}(x) \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ \gamma_{n-1}\psi_n(x) \end{pmatrix}$$

Mit folgenden Definitionen

$$\Psi^{(n)}(x) = \begin{pmatrix} \psi_0(x) \\ \vdots \\ \psi_{n-1}(x) \end{pmatrix} \in \mathbb{R}^n \text{ und } \hat{J}^{(n)} := \begin{pmatrix} \alpha_0 & \gamma_0 & & & \\ \beta_1 & \alpha_1 & \gamma_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \gamma_{n-2} \\ & & & \beta_{n-1} & \alpha_{n-1} \end{pmatrix}$$

wird aus der Dreiterm-Rekursion

$$\hat{J}^{(n)}\Psi^{(n)}(x) = x\Psi^{(n)}(x) - \gamma_{n-1}\psi_n(x) \cdot \mathbf{e}_n \qquad (\star)$$

Definition: Die aus den Koeffizienten der Drei-Term-Rekursion gebildete Matrix $\hat{J}^{(n)} \in \mathbb{R}^{n\times n}$ heißt *Jacobi-Matrix*.

Satz: Seien $(\psi_j, j = 0, \ldots, n)$ Orthogonalpolynome, $\hat{J}_n \in \mathbb{R}^{n\times n}$ die zugehörige Jacobi-Matrix. Dann gilt

$$\psi_n(\lambda) = 0 \Leftrightarrow \lambda \text{ Eigenwert von } \hat{J}^{(n)}$$

der Vektor $\Psi^{(n)}(\lambda)$ ist Eigenvektor von $\hat{J}^{(n)}$ zum Eigenwert λ.

Beweis: Folgt aus der Darstellung $(\star)$: Mit $\psi_n(\lambda) = 0$ folgt $\hat{J}^{(n)}\Psi^{(n)}(\lambda) = \lambda\Psi^{(n)}(\lambda)$.

4.6.3 Zur Konstruktion von Gauß-Quadraturen

Zur Erinnerung: J_n hat die Ordnung $2n + 2$ genau dann, wenn $q(x) = \prod_{j=0}^{n}(x - x_j) \perp \Pi_n$ gilt, genau dann, wenn $\{x_0, \ldots, x_n\}$ die Eigenwerte der Matrix $\hat{J}^{(n+1)}$ sind.

O.B.d.A betrachten wir Orthonormalpolynome statt Orthogonalpolynome, da hier $\beta_{j+1} = \gamma_j$ und damit $J^{(n)}$ symmetrisch ist.

Satz: Sei $J^{(n+1)} \in \mathbb{R}^{(n+1)\times(n+1)}$ die symmetrische Jacobi-Matrix zu den Orthonormalpolynomen $(\varphi_0, \ldots, \varphi_{n+1})$, seien $V, \Lambda \in \mathbb{R}^{(n+1)\times(n+1)}$ mit

$$J^{(n+1)} \cdot V = V \cdot \Lambda \text{ und } V^T V = E \text{ und } \Lambda \text{ diagonal}$$

Dann bilden die Eigenwerte λ_j (bzw. x_j) als Knoten und die Gewichte $\omega_j = \langle 1, 1\rangle\, V(0, j)^2$ eine Gauß-Quadratur mit Ordnung $2n + 2$.

Beweis: Für die Orthonormalpolynome $(\varphi_j, j = 0, \ldots, n)$ gilt für $k < n$

$$J_n[\varphi_0 \cdot \varphi_k] \;=\; \sum_{j=0}^{n} \varphi_0(x_j) \cdot \varphi_k(x_j) \cdot \omega_j = \begin{cases} 1 & \text{falls} \quad k = 0 \\ 0 & \text{sonst} \end{cases}$$

Also gilt:

$$\underbrace{\begin{pmatrix} \varphi_0(x_0) & \ldots & \varphi_0(x_n) \\ \varphi_1(x_0) & \ldots & \varphi_1(x_n) \\ \vdots & \ddots & \vdots \\ \varphi_n(x_0) & \ldots & \varphi_n(x_n) \end{pmatrix}}_{=:\Phi} \cdot \underbrace{\begin{pmatrix} \varphi_0(x_0)\omega_0 \\ \vdots \\ \varphi_0(x_n) \cdot \omega_n \end{pmatrix}}_{\varphi_0 \cdot w} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Falls $\varphi_{n+1}(x_j) = 0$ ist die j-te Spalte von Φ ein Eigenvektor von $J^{(n+1)}$, d.h.

$$V \cdot \operatorname{diag}\left(\frac{\varphi_0}{V(0, k)}, k = 0, \ldots, n\right) = \Phi$$

Damit ist:

$$V \cdot \begin{pmatrix} \frac{\varphi_0^2}{V(0,0)}\omega_0 \\ \vdots \\ \frac{\varphi_0^2}{V(0,n)}\omega_n \end{pmatrix} = \mathbf{e}_1 \Longleftrightarrow \begin{pmatrix} \frac{\varphi_0^2}{V(0,0)}\omega_0 \\ \vdots \\ \frac{\varphi_0^2}{V(0,n)}\omega_n \end{pmatrix} = V^T\mathbf{e}_1 = \begin{pmatrix} V(0,0) \\ \vdots \\ V(0,n) \end{pmatrix}$$

Es gilt also: $\varphi_0^2\omega_j = V(0,j)^2$ und wegen $\langle\varphi_0,\varphi_0\rangle = 1$ folgt $\varphi_0 = \frac{1}{\sqrt{\langle 1,1\rangle}}$.

4.6.4 Legendre-Polynome

Beispiel: Sei die Gewichtsfunktion $\varrho(x) = 1$. Betrachte Intervall $[-1,1]$, d.h. für $p, q \in \Pi$ ist

$$\langle p, q\rangle = \int_{-1}^{1} p(x)q(x)\,dx \qquad (\star)$$

Definition: Die folgenden Funktionen heißen *Legendre-Polynome* für $n \in \mathbb{N}_0$:

$$P_n(x) := \frac{1}{2^n n!} \cdot \frac{d^n}{dx^n}[(x^2-1)^n]$$

Satz: Die Legendre-Polynome sind Orthogonalpolynome (bezüglich $\langle\cdot,\cdot\rangle$, siehe $(\star)$) vom exakten Grad n,

$$\langle P_m, P_m\rangle = \begin{cases} 0 & \text{falls} \quad m \neq n \\ \frac{2}{2n+1} & \text{falls} \quad m = n \end{cases}$$

und genügen der Dreitermrekursion für $n \in \mathbb{N}$

$$(n+1)P_n(x) = (2n+1)x \cdot P_n(x) - n \cdot P_{n-1}(x) \text{ für } x \in [-1,1]$$

wobei $P_0(x) = 1$ und $P_1(x) = x$.

Beweis:

1. Da $(x^2-1)^n = x^{2n} \pm \ldots \in \Pi_{2n}$ ist, gilt

$$\frac{d^n}{dx^n}\left((x^2-1)^n\right) = \frac{(2n)!}{n!}x^n \pm \ldots \in \Pi_n \setminus \Pi_{n-1}$$

Damit haben die Polynome den „echten“ Grad n.

2. O.B.d.A sei $m \leq n$. Dann ist (mit partieller Integration):

$$\begin{aligned} P_{m,n} &:= 2^m m! \cdot 2^n n! \, \langle P_m, P_n \rangle \\ &= \int_{-1}^{1} \frac{d^m}{dx^m} \left((x^2-1)^m \right) \frac{d^n}{dx^n} ((x^2-1)^n) \, dx \\ &= \left[\frac{d^m}{dx^m} ((x^2-1)^m) \frac{d^{n-1}}{dx^{n-1}} ((x^2-1)^n) \right]_{-1}^{1} \\ &\qquad - \int_{-1}^{1} \frac{d^{m+1}}{dx^{m+1}} ((x^2-1)^m) \frac{d^{n-1}}{dx^{n-1}} ((x^2-1)^n) \, dx \end{aligned}$$

Dabei besitzt das Polynom $(x^2-1)^n$ für ± 1 eine n-fache Nullstelle, d.h.

$$\left[\frac{d^p}{dx^p} ((x^2-1)^n) \right]_{x=\pm 1} = 0 \; \forall \; p = 0, \ldots, n-1$$

Also ist (mit $(n-1)$-facher partieller Integration)

$$\begin{aligned} P_{m,n} &= - \int_{-1}^{1} \frac{d^{m+1}}{dx^{m+1}} ((x^2-1)^m) \cdot \frac{d^{n-1}}{dx^{n-1}} ((x^2-1)^n) \, dx \\ &= (-1)^n \int_{-1}^{1} \frac{d^{m+n}}{dx^{m+n}} ((x^2-1)^m)(x^2-1)^n \, dx \\ &= \end{aligned}$$

Falls nun $m < n$ ist, so ist $\frac{d^{m+n}}{dx^{m+n}} ((x^2-1)^m) = 0$, d.h. $P_{m,n} = 0$ für $m \neq n$. Ist jedoch $m = n$, so ist $\frac{d^{2n}}{dx^{2n}} ((x^2-1)^m) = (2n)!$, also ist (Zwischenschritt mit $(n-1)$-facher partieller Integration)

$$\begin{aligned} P_{n,n} &= (-1)^n (2n)! \int_{-1}^{1} (x-1)^n (x+1)^n \, dx \\ &= (-1)^n (2n)! \Big(\left[(x-1)^n \cdot \frac{(x+1)^{n+1}}{n+1} \right]_{-1}^{1} \\ &\qquad - \int_{-1}^{1} \frac{n}{n+1} (x-1)^{n-1} (x-1)^{n+1} \, dx \Big) \\ &= (-1)^n (-1)^n (2n)! \left(\frac{n}{n+1} \frac{n-1}{n+2} \cdot \ldots \cdot \frac{1}{2n} \right) \int_{-1}^{1} (x+1)^{2n} \, dx \\ &= (n!)^2 \left[\frac{(x+1)^{2n+1}}{2n+1} \right]_{-1}^{1} \\ &= (n!)^2 \frac{2^{2n+1}}{2n+1} \end{aligned}$$

Nach obigem $P_{m,n} = 2^m m! \cdot 2^n n! \langle P_m, P_n \rangle$ ergibt sich $P_{n,n} = \frac{2}{2n+1}$.

3. Es gilt für $n = 0$ und 1:

$$\begin{aligned} P_0(x) &= \frac{1}{2^0 \cdot 0!}((x^2-1)^0) = 1 \\ P_1(x) &= \frac{1}{2^1 \cdot 1!}((x^2-1)^1)' = \frac{1}{2}(x^2-1)' = x \end{aligned}$$

Bezeichne a_n den Höchstkoeffizienten von P_n

$$\begin{aligned} P_n(x) &= \frac{1}{2^n n!}\frac{d^n}{dx^n}\left(x^{2n} - \binom{n}{1}x^{2n-2} \pm \dots\right) \\ &= \underbrace{\frac{(2n)!}{2^n (n!)^2}x^n}_{a_n} - \frac{n(2n-2)!}{2^n n!(n-2)!}x^{n-2} + \dots \end{aligned}$$

Auf Grund der Orthogonalistät bildet $(P_0, \dots, P_n)$ eine Basis von Π_n, d.h.

$$P_{n+1}(x) = \frac{a_{n+1}}{a_n} \cdot x \cdot P_n(x) + \sum_{j=0}^{n} \alpha_j^n P_j(x)$$

Aus der Orthogonalität $\langle P_{n+1}, P_k \rangle = 0$ für $k = 0, \dots, n$ folgt:

(a) Bei $k < n-1$ ist $xP_k(x) \in \Pi_{n-1}$, dann ist $\langle xP_n, P_k \rangle = \langle P_n, xP_k \rangle = 0$, damit gilt:

$$\begin{aligned} P_{n+1}(x) &= \frac{a_{n+1}}{a_n} \cdot x \cdot P_n(x) + \sum_{j=0}^{n} \alpha_j^n P_j(x) \\ \Rightarrow \langle P_{n+1}(x), P_k \rangle &= \frac{a_{n+1}}{a_n} \langle x \cdot P_n(x), P_k \rangle + \alpha_k^n \langle P_k, P_k \rangle \\ \Rightarrow 0 &= 0 + \alpha_k^n \frac{2}{2k+1} \end{aligned}$$

Damit ergibt sich $\alpha_k = 0$ für $k < n-1$

(b) Für $k = n-1$ ist $xP_{n-1}(x) = \frac{a_{n-1}}{a_n}P_n + q$ mit $q \in \Pi_{n-1}$. Damit gilt:

$$\langle xP_n, P_{n-1} \rangle = \langle P_n, xP_{n-1} \rangle = \frac{a_{n-1}}{a_n} \langle P_n, P_n \rangle = \frac{a_{n-1}}{a_n}\frac{2}{2n+1}$$

Also gilt:

$$0 = \frac{a_{n+1}}{a_n}\frac{a_{n-1}}{a_n}\frac{2}{2n+1} + \alpha_{n-1}^n \frac{2}{2n-1}$$

Aufgelöst ergibt sich $\alpha_{n-1}^n = -\frac{n}{n+1}$.

(c) Für $k = n$ ist

$$0 = \frac{a_{n+1}}{a_n} \langle P_n, xP_n \rangle + \alpha_n^n \langle P_n, P_n \rangle$$

und dabei

$$\langle P_n, xP_n \rangle = \int_{-1}^{1} xP_n^2(x)\, dx = 0$$

Damit ist

$$\begin{aligned} P_{n+1}(x) &= \frac{a_{n+1}}{a_n} \cdot x \cdot P_n(x) + \sum_{j=0}^{n} \alpha_j^n P_j(x) \\ &= \frac{(2n+1) \cdot x \cdot P_n(x) - \frac{n}{n+1} P_{n-1}(x)}{n+1} \end{aligned}$$

Aus der Drei-Term-Reduktion folgt für $n > 0$:

$$xP_n(x) = \frac{n+1}{2n+1} P_{n+1} + 0 \cdot P_n + \frac{n}{2n+1} P_{n-1}$$

Damit ergeben sich $\alpha_n = 0$ $(n \in \mathbb{N}_0)$, $\beta_n = \frac{n}{2n+1}$ $(nin\mathbb{N})$ und $\gamma_n = \frac{n+1}{2n+1}$ $(n \in \mathbb{N}_0)$, Um zu den Orthonormalpolynomen φ_n zu gelangen, berechne $\delta_n := \sqrt{\langle P_n, P_n \rangle} \neq 0$. Dann ist $P_j = \delta_j \cdot \varphi_j$, damit ergibt sich

$$x\varphi_n(x) = \underbrace{\frac{\gamma_n \cdot \delta_{n+1}}{\delta_n}}_{\tilde{\gamma}_n} \varphi_{n+1} + \alpha_n \cdot \varphi_n + \underbrace{\frac{\beta_n \cdot \delta_{n-1}}{\delta_n}}_{\tilde{\beta}_n} \varphi_{n-1}$$

Es gilt $\tilde{\gamma}_n = \langle \varphi_{n+1}, x\varphi_n \rangle = \langle x\varphi_{n+1}, \varphi_n \rangle = \tilde{\beta}_{n+1}$, damit ist $\gamma_n \frac{\delta_{n+1}}{\delta_n} = \beta_{n+1} \frac{\delta_n}{\delta_{n+1}}$, d.h.

$$\delta_{n+1} = \pm \sqrt{\frac{\beta_{n+1}}{\gamma_n}} \delta_n \text{ und } \delta_0 = \sqrt{\langle 1, 1 \rangle}$$

Mit $D^{(n)} = \mathrm{diag}(\delta_0, \ldots, \delta_{n-1})$ gilt $P^{(n)} = D^{(n)} \Psi^{(n)}$, damit ist

$$J^{(n+1)} = D^{-1} \hat{J}^{(n+1)} D = \begin{pmatrix} \alpha_0 & \sqrt{\gamma_0 \beta_1} & & & \\ \sqrt{\gamma_0 \beta_1} & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \sqrt{\gamma_{n-1} \beta_n} \\ & & & \sqrt{\gamma_{n-1} \beta_n} & \alpha_n \end{pmatrix}$$

Speziell für die Legendre-Polynome ergibt sich

$$\gamma_{n-1}\beta_n = \frac{n}{(2n-1)}\frac{n}{(2n+1)} = \frac{1}{4-\frac{1}{n^2}}$$

Implementierung:

$[x, w] = \mathrm{GQ} - \mathrm{Legendre}(n)$

$d = \mathrm{sqrt}([1./(4 - 1./[1 : n].hoch2)]);$
$J = \mathrm{diag}(d, -1) + \mathrm{diag}(d, 1);$
$\nu = 2;$ % $\nu = \langle P_0, P_0 \rangle$
$[V, D] = \mathrm{eig}(J);$ % $J \cdot V = V \cdot D$
$[x, I] = \mathrm{sort}(\mathrm{diag}(D));$
$w = \nu \cdot V(1, I).hoch2;$

5 Nichtlineare Gleichungen

5.1 Motivation

Komet: $y = 1 - e^x$, gesucht ist ein Punkt der Menge $G = \{(x, y) \in \mathbb{R}^2 \mid y = 1 - e^x\}$, vom Ursprung den Abstand 2 besitzt. Also: Gesuch ist (x, y) mit $e^x + y = 1$ und $x^2 + y^2 = 2$.

Umformulierung: $f_1 \colon \mathbb{R}^2 \to \mathbb{R}$ mit $(x, y) \mapsto x^2 + y^2 - 2$ und $f_2 \colon \mathbb{R}^2 \to \mathbb{R}$ mit $(x, y) \mapsto e^x + y - 1$, weiter sei $f \colon \mathbb{R}^2 \to \mathbb{R}^2$ mit $(x, y) \mapsto (f_1(x, y), f_2(x, y))$. Gesucht ist

- eine Nullstelle der Funktion $f \colon \mathbb{R}^n \to \mathbb{R}^n$, d.h. ein Vektor $x \in \mathbb{R}^n$ mit $f(x) = 0 \in \mathbb{R}^n$;
- bzw. ein Fixpunkt der Funktion $g \colon \mathbb{R}^n \to \mathbb{R}^n$ mit $x \mapsto f(x) + x$, d.h. ein Vektor $x \in \mathbb{R}^n$ mit $g(x) = x$.

5.2 Fixpunktiteration

Definition: Sei V ein Vektorraum und $g \colon V \to V$ eine Abbildung. Ein Element $\hat{x} \in V$ heißt *Fixpunkt* der Abbildung g, falls $\hat{x} = g(\hat{x})$.

Verwende Startwert $x^{(0)}$ und eine Folge $\left(x^{(k)}\right)_{k \in \mathbb{N}}$ mit $x^{(k+1)} = g(x^{(k)})$ für alle $k \in \mathbb{N}$. Die Frage ist dann, ob $\lim_{k \to \infty} x^{(k)} = \hat{x}$ ist.

Um Konvergenz erklären zu können, benötigen wir einen Banach-Raum $(V, \|\cdot\|)$, d.h.

- V ist ein normierter Vektorraum, dann bedeutet die Konvergenz von $\left(x^{(k)}\right)_{k \in \mathbb{N}}$, daß $\lim_{\to \infty} n \left\|x^{(k)} - \hat{x}\right\| = 0$ ist, d.h. die Folge ist eine Cauchy-Folge:
$$\forall\, \varepsilon > 0\ \exists\, n \in \mathbb{N}_0\ \forall\, m, k \geq n \colon\ \left\|x^{(k)} - x^{(m)}\right\| < \varepsilon$$
- V ist abgeschlossen, d.h. für alle Folgen $\left(x^{(k)}\right)_{k \in \mathbb{N}}$ ist $\lim_{n \to \infty} x^{(n)} \in V$.

Ziel: Bedingungen an g und D (mit $g \colon D \subseteq V \to V$), so daß die Fixpunktiteration (FPI) für belibige $x^{(0)} \in D$ konvergiert, hoffentlich gegen einen Fixpunkt.

Frage ist also: Wann klappt die Fixpunktiteration?

Definition: Sei $(V, \|\cdot\|)$ ein normierter Vektorraum, $g\colon D \subseteq V \to V$ eine Abbildung. g heißt Lipschitz-stetig auf D mit L-Konstante L, wenn $L \in \mathbb{R}$ existiert mit

$$\forall\, x, y \in D\colon\ \|g(y) - g(x)\| \leq L \cdot \|y - x\|$$

Gilt darüberhinaus $L < 1$, so heißt g *kontrahierend* auf D und L heißt *Kontraktions-Konstante* von g auf D.

Bemerkung: Eine Lipschitz-stetige Funktion ist auf D gleichmäßig stetig. Für beliebige $\varepsilon > 0$ gilt mit $\delta := \frac{\varepsilon}{L}$ für alle $x, y \in D$:

$$\|x - y\| < \delta \Rightarrow \|g(x) - g(y)\| \leq L \cdot \|x - y\| \leq L \cdot L \cdot \frac{E}{L} = \varepsilon$$

Beispiele:

1. Sei $D = V = \mathbb{R}$ und $\|\cdot\| = |\cdot|$, sei $g\colon \mathbb{R} \to \mathbb{R}$ definiert als $g\colon x \mapsto \sin x$. Dann ist

$$\begin{aligned} \|g(y) - g(x)\| &= \|\sin(y) - \sin(x)\| \\ &= \left\| \int_x^y \cos(t)\, dt \right\| \\ &= \|\cos(\xi)(y - x)\| \ \text{ für ein } \xi \in [x, y] \\ &\leq \max\{\, |\cos(\xi)| \mid \xi \in [x, y]\} \cdot \|y - x\| \\ &\leq 1 \cdot \|y - x\| \end{aligned}$$

2. Sei nun $D = V = \mathbb{R}^2$ und $g\colon \mathbb{R}^2 \to \mathbb{R}^2$ definiert durch $g\colon x \mapsto Ax$ für ein $A \in \mathbb{R}^{2\times 2}$. Dann ist

$$\|g(x) - g(y)\| = \|Ax - Ay\| = \|A(x - y)\| = \|A\|\, \|x - y\|$$

Satz: Sei $D \in \mathbb{R}^n$ beschränkt, abgeschlossen und konvex und sei $g \in \mathcal{C}^1(D, \mathbb{R}^n)$ mit $L = \max\{\, \|g'(x)\| \mid x \in D\} < 1$. Dann ist g auf D kontrahierend mit einer Kontraktions-Konstante $L < 1$

Bemerkungen:

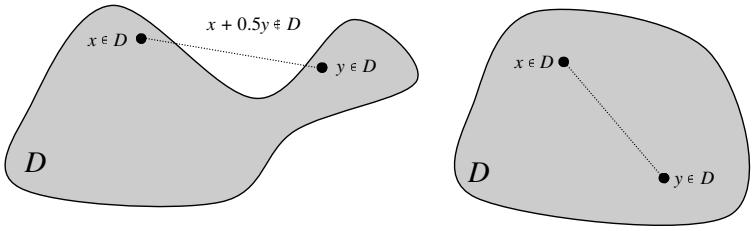
- Seien

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad g(x) = \begin{pmatrix} g_1(x_1, \ldots, x_n) \\ \vdots \\ g_n(x_1, \ldots, x_n) \end{pmatrix}$$

  Dann ist

$$g'(x) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_n} \end{pmatrix}$$

- Eine Menge D ist konvex, wenn mit $x, y \in D$ auch $x + t \cdot (y - x) \in D$ ist für alle $t \in [0, 1]$, Beispiel: die rechte Menge ist konvex, die linke ist es nicht:



Beweis: Benutze den Mittelwertsatz für vektorwertige Funktionen $g\colon D \subseteq \mathbb{R}^p \to \mathbb{R}^q$ mit $g(x)$ wie oben zusammengesetzt aus $g_1, \ldots, g_q$. Definiere für alle $j = 1, \ldots, q$:

$$\varphi_j\colon [0,1] \to \mathbb{R} \text{ mit } \varphi_j\colon t \mapsto g_j(x + t(y - x))$$

Dann ist $\varphi_j(1) - \varphi_j(0) = \int_0^1 \varphi_j(t)\, dt$. Dabei ist

$$\begin{aligned} \varphi_j'(t) &= \frac{\partial g_j}{\partial x_1}(y_1 - x_1) + \ldots + \frac{\partial g_j}{\partial x_p}(y_p - x_p) \\ &= \operatorname{grad} g_j(x + t(y - x)) \cdot (y - x) \end{aligned}$$

Weiter gilt

$$\begin{aligned} g(y) - g(x) &= \begin{pmatrix} \varphi_1(1) - \varphi_1(0) \\ \vdots \quad - \quad \vdots \\ \varphi_q(1) - \varphi_q(0) \end{pmatrix} = \begin{pmatrix} \int_0^1 \varphi_1'(t)\, dt \\ \vdots \\ \int_0^1 \varphi_q'(t)\, dt \end{pmatrix} \\ &= \begin{pmatrix} \int_0^1 \operatorname{grad} g_1(x + t(y - x)) \cdot (y - x)\, dt \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ \int_0^1 \operatorname{grad} g_q(x + t(y - x)) \cdot (y - x)\, dt \end{pmatrix} \\ &= \begin{pmatrix} \sum_{k=1}^p \int_0^1 \frac{\partial g_1}{\partial x_k}(y_k - x_k) \\ \vdots \qquad\qquad \vdots \\ \sum_{k=1}^p \int_0^1 \frac{\partial g_q}{\partial x_k}(y_k - x_k) \end{pmatrix} \\ \overset{14}{=} &\left[\int_0^1 \frac{\partial g_j}{\partial x_k}(x + t(y - x))\right]_{j,k} \cdot (y - x) \\ &= \left[\int_0^1 g'(x + t(y - x))\right] \cdot (y - x) \end{aligned}$$

[14] nur eine andere Schreibweise für die Matrix!

Nun gilt:

$$\begin{aligned}
\|g(y)-g(x)\| &= \left\| \left[\int_0^1 g'(x+t(y-x)) \right] \cdot (y-x) \right\|_V \\
&\leq \left\| \left[\int_0^1 g'(x+t(y-x)) \right] \right\|_V \cdot \|y-x\|_V \\
&\leq \left(\int_0^1 \|g'(x+t(y-x))\|_{M\,[0,1]}\, dt \right) \cdot \|y-x\|_V \\
&\leq \left(\max\{ \|g'(x+t(y-x))\|_M \mid t \in [0,1] \} \cdot \int_0^1 1\, dt \right) \cdot \|y-x\|_V \\
&\leq \max\{ \|g'(x)\|_M \mid x \in D \} \cdot \|y-x\|_V \\
&= L \cdot \|y-x\|_V
\end{aligned}$$

Da zudem $L < 1$ ist nach Voraussetzung ist g auf D kontrahierend.

5.3 Einführung

Wir benötigen einen linearen Vektorraum, der normiert und vollständig ist, also einen Banachraum, Wir betrachten hier aber nur $\mathbb{R}^n$. Gegeben seien $f\colon V \to V$ mit $x \mapsto f(x)$ und $y_0 \in \mathbb{R}^n$. Gesucht ist $x \in \mathbb{R}^n$ mit $f(x) = y_0$. O.b.d.A. sei $f(x) = 0$, ersetze andernfalls einfach $\tilde{f}(x) = f(x) - y_0$. **Beispiele:**

1. Sei $K_2(0)$ gegeben, gesucht ist ein Schnittpunkt mit $y(x) = 1 - e^x$. Genauer: $V = \mathbb{R}^2$, $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ gegeben mit $(x, y) \mapsto (x^2 + y^2 - 2, y - 1 + e^x)$. Gesucht ist $f(x) = 0$.

2. Sei ein LGS $Ax = b$ gegeben, dann ist x gesucht mit $f(x) = Ax - b = 0$.

Zu klärende Fragen sind also:

1. Existiert eine Lösung?

2. Ist die Lösung eindeutig?

3. Wie berechnet man die Lösung?

Grundlegender Gedanke: Iteratives Denken, d.h. Startwert $x^{(0)} \in \mathbb{R}^n$, Berechnungsschleife $x^{(i)} \longrightarrow x^{(i+1)}$ - konvergiert diese Folge?

5.4 Fixpunktiteration

Bisher: $f(x) = 0$. Setze $\varphi(x) = x + f(x)$, d.h. $x = \varphi(x)$.

Definition: Sei $\varphi\colon \mathbb{R}^n \to \mathbb{R}^n$ gegeben.

1. Eine Gleichung der Gestalt $x = \varphi(x)$ heißt Fixpunktgleichung.
2. Jede Lösung der Fixpunktgleichung heißt Fixpunkt von φ.

Fixpunktiteration: Sei $x^{(0)} \in \mathbb{R}^n$ Startwert und benutze $x^{(i+1)} = \varphi(x^{(i)})$. Frage ist nun, ob $\lim_{i\to\infty} x^{(i)} = \hat{x}$ gilt.

Satz: Sei φ stetig, es gelte $\lim_{i\to\infty} x^{(i)} = \hat{x} \in \mathbb{R}^n$. Dann ist $\hat{x}$ Fixpunkt von φ.

Beweis: Es gilt

$$\hat{x} = \lim_{i\to\infty} x^{(i)} = \lim_{i\to\infty} x^{(i+1)} = \lim_{i\to\infty} \varphi(x^{(i)}) = \varphi \lim_{i\to\infty} x^{(i)} = \varphi(\hat{x})$$

Definition: φ heißt *stark kontrahierend*, falls ein $L < 1$ existiert mit

$$\|\varphi(x) - \varphi(y)\| \leq L \cdot \|x - y\| \qquad \forall\, x \neq y \in \mathbb{R}^n$$

φ heißt *schwach kontrahierend*, falls

$$\|\varphi(x) - \varphi(y)\| \leq \|x - y\| \qquad \forall\, x \neq y \in \mathbb{R}^n$$

Bemerkung: Es gilt: Wenn φ stark kontrahierend ist, ist φ schwach kontrahierend.

Satz: Eine schwach kontrahierende Abbildung φ hat höchstens einen Fixpunkt.

Beweis: Seien $\hat{x} \neq \hat{y}$ Fixpunkte von φ. Dann ist

$$\|\hat{x} - \hat{y}\| = \|\varphi(\hat{x}) - \varphi(\hat{y})\| < \|\hat{x} - \hat{y}\|$$

Dies ist ein Widerspruch, also hat φ höchstens einen Fixpunkt.

Satz: Sei $\varphi \in \mathcal{C}(\mathbb{R}^n, \mathbb{R}^n)$ kontrahierend. Dann gilt:

1. φ hat genau einen Fixpunkt $\hat{x}$
2. Für jedes $x^{(0)} \in \mathbb{R}^n$ konvergiert die Fixpunktiteration gegen $\hat{x}$.
3. Es gelten die Fehlerabschätzungen für alle $i \in \mathbb{N}$:

$$\begin{aligned} \left\|x^{(i)} - \hat{x}\right\| &\leq L^i \left\|x^{(0)} - \hat{x}\right\| \\ \left\|x^{(i)} - \hat{x}\right\| &\leq \frac{L^i}{1-L} \left\|x^{(1)} - x^{(0)}\right\| \end{aligned}$$

Beweis:

2. Sei $(x^{(i)})_{i\in\mathbb{N}}$ die Folge der Fixpunkt-Iterierten. Nun betrachte

$$\begin{aligned} \left\|x^{(i+1)} - x^{(i)}\right\| &= \left\|\varphi(x^{(i)}) - \varphi(x^{(i-1)})\right\| \\ &\leq L \cdot \left\|x^{(i)} - x^{(i-1)}\right\| \\ &\leq L^2 \cdot \left\|x^{(i-1)} - x^{(i-2)}\right\| \ldots \\ &\leq L^{i-j} \cdot \left\|x^{(j+1)} - x^{(j)}\right\| \text{ für } 0 \leq j \leq i \end{aligned}$$

   Sei nun $n > m$. Aus obiger Abschätzung für aufeinanderfolgende $x^{(i+1)}$, $x^{(i)}$ ist

$$\begin{aligned} \left\|x^{(n)} - x^{(m)}\right\| &= \left\|x^{(n)} - x^{(n-1)} + x^{(n-1)} - \ldots + x^{(m+1)} - x^{(m)}\right\| \\ &\leq \left\|x^{(n)} - x^{(n-1)}\right\| + \left\|x^{(n-1)} - x^{(n-2)}\right\| + \ldots + \left\|x^{(m+1)} - x^{(m)}\right\| \\ &\leq \left(L^{n-m-1} + \ldots + L^2 + L + 1\right) \cdot \left\|x^{(m+1)} - x^{(m)}\right\| \\ &\leq \sum_{i=1}^{\infty} L^i \cdot \left\|x^{(m+1)} - x^{(m)}\right\| \\ &\leq \frac{1}{1-L} \cdot \left\|x^{(m+1)} - x^{(m)}\right\| \\ &\leq \frac{1}{1-L} \cdot L^m \cdot \left\|x^{(1)} - x^{(0)}\right\| \end{aligned}$$

   Damit ist $\lim_{m\to\infty} \lim_{n\to\infty} \left\|x^{(n)} - x^{(m)}\right\| = 0$. Somit liegt eine Cauchyfolge vor.

1. Folgt mit obigem Satz und dem eben Gezeigten, da eine kontrahierende Funktion auch schwach kontrahierend hat und damit höchstens einen Fixpunkt hat.

3. Es gilt:

$$\lim_{n\to\infty} \left\|x^{(n)} - x^{(m)}\right\| \left\|\lim_{n\to\infty} x^{(n)} - x^{(m)}\right\| \left\|\hat{x} - x^{(m)}\right\| \leq \frac{L^m}{1-L} \left\|x^{(1)} - x^{(0)}\right\|$$

Zudem ist

$$\lim_{n\to\infty} \left\|x^{(i+1)} - \hat{x}\right\| = \left\|\varphi(x^{(i)}) - \varphi(\hat{x})\right\| \leq L^i \left\|x^{(0)} - \hat{x}\right\|$$

Bemerkung: Die Abschätzung $\left\|x^{(i)} - \hat{x}\right\| \leq \frac{L^i}{1-L} \left\|x^{(1)} - x^{(0)}\right\|$ ist eine *a-priori-Abschätzung*. Sei nun $\varepsilon > 0$ gegeben; für welche $k \in \mathbb{N}$ gilt $\left\|x^{(k)} - \hat{x}\right\| < \varepsilon$? Es ist dann

$$\begin{aligned}
\frac{L^k}{1-L} \left\|x^{(1)} - x^{(0)}\right\| &< \varepsilon \\
\Rightarrow L^k &\leq \varepsilon \frac{1-L}{\|x^{(1)} - x^{(0)}\|} \\
\Rightarrow k \cdot \log L &\leq \log \frac{\varepsilon - \varepsilon \cdot L}{\|x^{(1)} - x^{(0)}\|} \\
\Rightarrow k &\geq \left(\log \frac{\varepsilon - \varepsilon \cdot L}{\|x^{(1)} - x^{(0)}\|}\right) \cdot \frac{1}{\log L}
\end{aligned}$$

KOROLLAR: Seien die Voraussetzungen wie oben. Es gilt die *a-posteroi-Abschätzung*

$$\left\|x^{(m)} - \hat{x}\right\| \leq \frac{L}{1-L} \left\|x^{(m)} - x^{(m-1)}\right\|$$

Beweis: Siehe Beweis von obigem Satz, sei $n > m$, dann ist $\left\|x^{(n)} - x^{(m)}\right\| \leq \frac{1}{1-L} \left\|x^{(m+1)} - x^{(m)}\right\|$, nun ist

$$\lim_{n\to\infty} \left\|x^{(n)} - x^{(m)}\right\| \leq \left\|\lim_{n\to\infty} x^{(n)} - x^{(m)}\right\| = \left\|\hat{x} - x^{(m)}\right\|$$

5.4.1 Das Sekantenverfahren

Funktioniert wie das Sehnenverfahren, man achtet aber nicht auf das Vorzeichen.

5.4.2 Das Newton-Raphson-Verfahren

Sei $x^{(0)} \in \mathbb{R}$ gegeben. $x^{(i+1)}$ wird berechnet als Schnitt der Tangente in x^i mit der x-Achse. Analytische Darstellung: Sei x^i gegeben mit $f'(x^{(i)}) \neq 0$.

Gesucht ist x mit $f(x) = 0$. Es folgt mit Taylorentwicklung:

$$\begin{aligned} 0 = f(x^{(i)} + (x - x^{(i)})) &= f(x^{(i)}) + (x - x^{(i)}) f'(x^{(i)}) + \text{Rest} \\ \Longrightarrow x f'(x^{(i)}) &= x^{(i)} f'(x^{(i)}) - f(x^{(i)}) \\ \Longrightarrow x &= x^{(i)} - \frac{f(x^{(i)})}{f'(x^{(i)})} \qquad (\star) \end{aligned}$$

Dabei lassen wir im zweiten Schritt den Rest weg.

Bemerkungen:

1. Die Iteration ist durchführbar solange $f'(x^{(i)}) \neq 0$.
2. $(\star)$ ist eine Fixpunktiteration mit $\phi(x) = x - \frac{f(x)}{f'(x)}$.

Beispiel: Sei $f(x) = x^2 - 2$. Iteration:

$$x^{(i+1)} \quad = \quad x^{(i)} - \frac{(x^{(i)})^2 - 2}{2x^{(i)}} = \frac{x^{(i)}}{2} + \frac{1}{x^{(i)}}$$

Ist $x^{(0)} > 0$, so konvergiert $x^{(i)}$ gegen $\sqrt{2}$, ist $x^{(0)} < 0$, so konvergiert $x^{(i)}$ gegen $-\sqrt{2}$. Falls $x^{(0)} = 0$, Abbruch. Also die Wahl des Startwertes beeinflusst das Ergebnis. Wähle z.B. $x^{(0)} = 2$, dann ist

$$\begin{aligned} x^{(1)} &= \underline{1}.5 \\ x^{(2)} &= \underline{1.41}66 \\ x^{(3)} &= \underline{1.414215}686 \\ x^{(4)} &= \underline{1.4142135}651 \end{aligned}$$

5.5 (5.4) Newton-Verfahren für System von nichtlinearen Gleichungen

5.5.1 (5.4.1) Algorithmus

Wir betrachten $f : U \subseteq \mathbb{R}^n \to \mathbb{R}^n$. Gesucht ist $x \in U$ mit $f(x) = 0$. Wir schreiben

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad f = \begin{bmatrix} f_1 \\ \vdots \\ f_n \end{bmatrix} \quad f'(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}$$

Sei $\|x\|$ die Vektornorm und $\|A\|$ die zugehörige Matrixnorm, d.h. $\|A\| = \sup\limits_{0 \neq x \in \mathbb{R}^n} \frac{\|Ax\|}{\|x\|}$. Sei $x^{(0)} \in U$ gegeben. Wir definieren

$$x^{(i+1)} x^{(i)} - f'(x^{(i)})^{-1} f(x^{(i)})$$

In Wirklichkeit wird es nicht so berechnet (niemals inverse berechnen!). Ein Iterationsschritt besteht aus zwei Teilaufgaben:

1. Löse das lineare Gleichungssystem $f'(x^{(i)})\delta = f(x^{(i)})$, d.h. finde $\delta \in \mathbb{R}^n$.
2. Korrigiere $x^{(i+1)} = x^{(i)} - \delta$.

5.5.2 (5.4.2) Konvergenzaussagen

Bemerkungen:

1. $f(x)$ heißt *affin*, falls $f(x) = Ax + b$. Für beliebiges $x^{(0)} \in \mathbb{R}^n$ ist $x^{(1)}$ eine Lösung.
2. Es kann mehr als eine Lösung von $f(x) = 0$ geben.
3. Ist x^* eine Lösung von $f(x) = 0$, so ist x^* ein Fixpunkt von ϕ mit $\phi(x) = x - f'(x)^{-1}f(x)$.

Definition: Die *Sphere* mit Zentrum $x \in \mathbb{R}^n$ und Radius r ist

$$S_r(x) = \{\, y \in \mathbb{R}^n \mid \|y - x\| < r\}$$

Bemerkung: Falls $\|x\| = \|x\|_2$, so ist $S_r(x)$ eine „Kugel“. Falls $\|x\| = \|x\|_\infty$, so ist $S_r(x)$ ein „Würfel“.

Definition: Ein Iterationsverfahren $x^{(i+1)} = \phi(x^{(i)})$ heißt *quadratisch konvergent* in $U \subseteq \mathbb{R}^n$ gegen $x^* \in U$, falls ein $0 < c \in \mathbb{R}$ existiert mit

$$\|\phi(x) - x^*\| \leq c\,\|x - x^*\|^2$$

Definition: $f \in C^1(U, \mathbb{R}^n)$ hat eine Lipschitz-stetige Ableitung, falls eine Konstante $L < \infty$ existiert mit

$$\|f'(x) - f'(y)\| \leq L\,\|x - y\|$$

Satz: f habe eine Lipschitz-stetige Ableitung in $U = S_r(x^*)$ mit $f(x^*) = 0$ und $f'(x^*)$ sei regulär. Dann gibt es ein $0 < r' \leq r$ so, dass die Newton-Iteration in $S_{r'}(x^*)$ wohldefiniert und quadratisch konvergent ist.

Bemerkung: Viel aussagekräftiger ist der Satz von Newton-Kantorovitsch.

5.5.3 (5.4.3) Modifikationen

1. $f'(x)$ approximieren durch Differentenquotienten:

$$\frac{\partial f_j(x_1, \ldots, x_n)}{\partial x_i} = \lim_{y \to x_i} \frac{f(x_1, \ldots, y, \ldots, x_n) - f(x_1, \ldots, x_i, \ldots, x_n)}{|y - x_i|}$$

2. Man könnte die Jacobi-Matrix „einfrieren“, d.h. z.B. nur einmal pro 10 Iterationen neu berechnen. Dann verliert man allerdings die quadratische Konvergenz und bleibt bei der linearen.