Numerische Mathematik und Numerische Lineare Algebra in den Datenwissenschaften

Prof. Dr. rer. nat. Jens Starke
Sommersemester 2025

Inhaltsverzeichnis

Inhaltsverzeichnis

Diese Mitschrift basiert auf der gleichnamigen Vorlesung *Numerische Mathematik und Numerische Lineare Algebra in den Datenwissenschaften*, gehalten im Sommersemester 2025 an der Universität Rostock.

Literaturempfehlungen:

1. Martin Hanku-Bourgeois, Grundlagen der Numerik und des wissenschaftlichen Rechnens, Mathematische Leitfäden, Vieweg + Teubner Verlag Wiesbaden, 2009, DOI: 10.1007/978-3-8348-9309-3

2. Eberhard Zeidler, Nichtlineare Funktionalanalysis und ihre Anwendung, Springer Spektrum Wiesbaden, 2012, DOI: 10.1007/978-3-658-00289-3_3

1 Wiederholung

Wir starten mit einer kurzen Wiederholung zur Fixpunktiteration zum Lösen von Gleichungen der Form $Tx = x$ durch die Iterationsvorschrift $x^{(n+1)} = Tx^{(n)}$.

Satz 1.1 (Banach 1922). Sei M eine abgeschlossene nichtleere Teilmenge in einem vollständig metrischem Raum (X, d) und $T : M \to M$ eine Selbstabbildung (d.h. $T(M) \subset M$) und k-kontraktiv (d.h. $d(Tx, Ty) \leq k \cdot d(x, y)\ \forall x, y \in M$ mit $0 \leq k < 1$). Dann folgt:

1. Existenz und Eindeutigkeit: die Gleichung $Tx = x$ hat genau eine Lösung, d.h. T hat genau einen Fixpunkt in der Menge M.
2. Konvergenz der Iteration $x^{(k+1)} = Tx^{(k)}$. Die Folge $(x^{(k)})_{k\in\mathbb{N}}$ konvergiert gegen den Fixpunkt $\hat{x}$ für einen beliebigen Startpunkt $x^{(0)} \in M$.
3. Fehlerabschätzung: Für alle $n = 0, 1, \ldots$ gilt
   - a-priori: $d(x^{(n)}, \hat{x}) \leq k^n(1 - k)^{-1}d(x^{(0)}, x^{(1)})$
   - a-posteriori: $d(x^{(n+1)}, \hat{x}) \leq k(1 - k)^{-1}d(x^{(n)}, x^{(n+1)})$
4. Konvergenzrate: Für alle $n \in \mathbb{N}$ gilt $d(x^{(n+1)}, \hat{x}) \leq k \cdot d(x^{(n)}, \hat{x})$

Beweis.

2. Wir zeigen, dass $(x^{(n)})$ eine Cauchy-Folge ist: Für den Abstand zweier benachbarter Folgeglieder $x^{(n)}$ und $x^{(n+1)}$ gilt

$$d(x^{(n)}, x^{(n+1)}) = d(Tx^{(n-1)}, Tx^{(n)}) \leq k \cdot d(x^{(n-1)}, x^{(n)}) \leq \ldots \leq k^n \cdot d(x^{(0)}, x^{(1)})$$

   Mehrfache Anwendung der Dreiecksungleichung liefert daher für $n, m \in \mathbb{N}$:

$$\begin{aligned} d(x^{(n)}, x^{(n+m)}) &\leq d(x^{(n)}, x^{(n+1)}) + d(x^{(n+1)}, x^{(n+2)}) + \ldots + d(x^{(n+m-1)}, x^{(n+m)}) \\ &\leq (k^n + k^{n+1} + \ldots + k^{n+m}) \cdot d(x^{(0)}, x^{(1)}) \\ &\leq k^n(1 + k + k^2 + \ldots) \cdot d(x^{(0)}, x^{(1)}) \\ &= k^n \cdot (1 - k)^{-1}d(x^{(0)}, x^{(1)}) \end{aligned}$$

   Demnach folgt $d(x^{(n)}, x^{(n+m)}) \to 0$ für $n \to \infty$ und da X vollständig ist konvergiert $(x^{(n)})$ gegen ein $\hat{x} \in X$.

1. Da T stetig ist (aufgrund der k-Kontraktivität) folgt für die konvergente Folge $(x^{(n)})$, dass

$$\hat{x} = \lim_{n\to\infty} x^{(n+1)} = \lim_{n\to\infty} Tx^{(n)} = T\hat{x}$$

   Da M abgeschlossen ist existiert also ein Fixpunkt in M.

   Dieser ist eindeutig, denn für x, y mit $Tx = x$ und $Ty = y$ gilt $d(x, y) = d(Tx, Ty) \leq kd(x, y)$, also $d(x, y) = 0$ und damit $x = y$ (Definitheit der Metrik).

3. Aus dem Beweis zu 2. haben wir $d(x^{(n)}, x^{(n+m)}) \leq k^n(1-k)^{-1}d(x^{(0)}, x^{(1)})$, wegen der Stetigkeit der Metrik folgt die a-priori-Fehlerabschätzung aus $m \to \infty$.

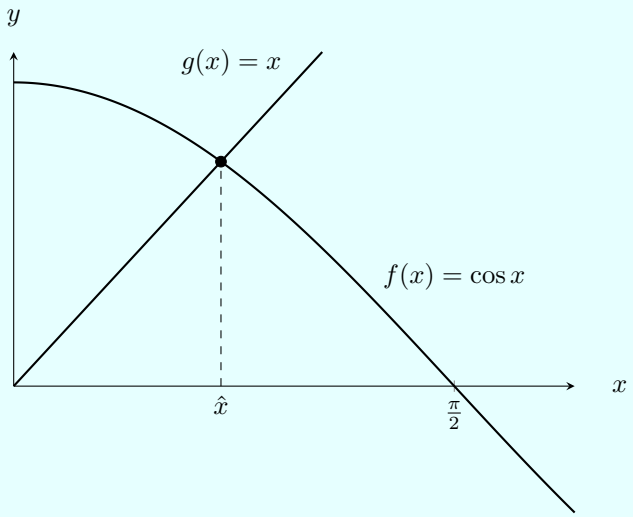
Die a-posteriori-Fehlerabschätzung folgt analog aus dem Ansatz

$$\begin{aligned} d(x^{(n+1)}, x^{(n+1+m)}) &\leq d(x^{(n+1)}, x^{(n+2)}) + \ldots + d(x^{(n+m)}, x^{(n+1+m)}) \\ &\leq (k + \ldots + k^m) \cdot d(x^{(n)}, x^{(n+1)}) \\ &\leq k(1 + k + k^2 + \ldots) \cdot d(x^{(n)}, x^{(n+1)}) \\ &= Ck \cdot (1-k)^{-1} d(x^{(n)}, x^{(n+1)}) \end{aligned}$$

4. Folgt direkt durch $d(x^{(n+1)}, \hat{x}) = d(Tx^{(n)}, T\hat{x}) \leq k \cdot d(x^{(n)}, \hat{x})$

□

Beispiel 1.2. Wir betrachten das Nullstellenproblem $f : \mathbb{R} \to \mathbb{R}, x \mapsto \cos x - x = 0$.
Umformung ergibt $\underbrace{\cos x}_{Tx} = x$ und somit die Fixpunktiteration $x^{(k+1)} = Tx^k = \cos(x^{(k)})$



Prüfung der Voraussetzungen des Banach'schen FP-Satzes:

Wir wählen als Einschränkung $M = [0, 1]$, dies liefert uns eine Selbstabbildung auf einer abgeschlossenen Teilmenge M des vollständig metrischen Raum $\mathbb{R}$ mit der Abstandsfunktion $d(x, y) = |x - y|$.

Weiter ist die Abbildung k-kontraktiv, denn nach dem Mittelwertsatz der Differentialrechnung gilt

$$|\cos x - \cos y| = \underbrace{|\sin \xi|}_{\leq \sin(1)} \cdot |x - y| \leq \underbrace{0,85}_{=:k} \cdot |x - y|, \quad \text{für } \xi \in [0, 1]$$

Wir können also nach Banach die Existenz und Eindeutigkeit eines Fixpunkt $\hat{x}$ folgern, diesen Fixpunkt finden wir durch die konvergente Folge $x^{(k+1)} = \cos x^k$.

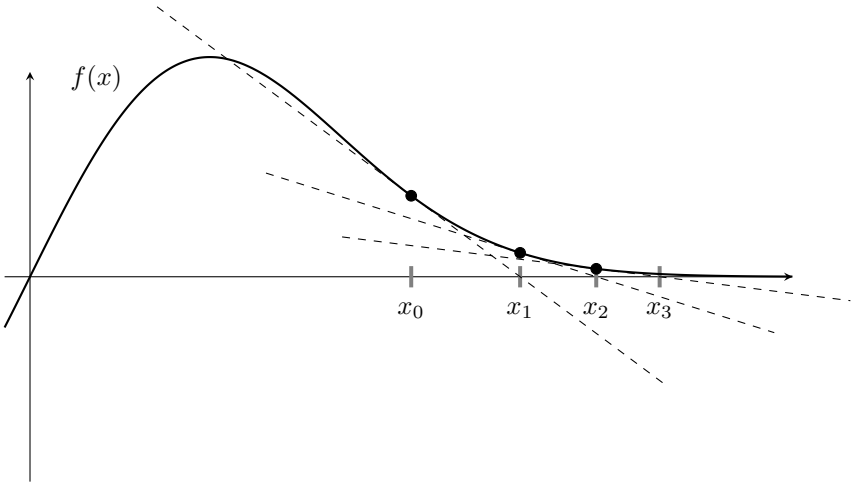
Wir betrachten im folgenden die Idee der Umwandlung eines Nullstellenproblems in Fixpunkt-Gleichung noch etwas allgemeiner. Für eine Gleichung $f(x) = 0$ mit $f : \mathbb{R} \to \mathbb{R}$ haben wir verschiedene Möglichkeiten zur Umformung:

a) Betrachte $Tx := x - f(x)$ gefolgert aus $f(x) = 0 \Leftrightarrow -f(x) = 0 \Leftrightarrow x - f(x) = x$.

b) Betrachte $Tx := x - \omega \cdot f(x)$ mit $\omega \neq 0$ (lineare Relaxation)

c) Betrachte $Tx := x - \omega \cdot g(f(x))$ mit $\omega \neq 0$ und geeigneter Funktion g (nichtlineare Relaxation). Wenn $g(0) \neq 0$ dann betrachte $Tx := x - \omega \cdot (g(f(x)) + g(0))$

d) Betrachte $Tx := x - (f'(x))^{-1} f(x)$ (Newtonverfahren). Newton hat teils Probleme, bei falschen Startwerten:



e) Betrachte $Tx := h^{-1}(f(x) - g(x))$, wobei $f(x) = h(x) + g(x)$ (Splitting-Verfahren)

2 Iteratives Vorgehen zur Lösung linearer Gleichungssysteme

2.1 Splittingverfahren

Gegeben sei das LGS $Ax = b$ für $A \in \mathbb{K}^{n\times n}, b \in \mathbb{K}^n, x \in \mathbb{K}^n$, wobei $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. Wir wollen dieses LGS nun in ein FP-Problem umformen, sei hierfür A nicht singulär (sonst nicht lösbar).

Wir schreiben $A = M - N$, wobei M invertierbar und häufig sogar eine Diagonalmatrix ist (damit M leicht zu invertieren ist). Dies liefert:

$$Ax = b \iff (M-N)x = b \iff Mx = Nx + bx = \underbrace{M^{-1} \cdot (Nx + b)}_{\tilde{T}x}$$

$\tilde{T}$ ist affin-linear, wir erhalten also unser FP-Problem $x = \tilde{T}x = Tx + c$ mit $T = M^{-1}N$ und $c = M^{-1}b$

Algorithmus 1: Splittingverfahren

Initialisierung: $A = M - N$ mit $N \in GL(n, \mathbb{K})$
1 Wähle $x^{(0)} \in \mathbb{K}^n$ beliebig
2 **for** $k = 0, 1, \ldots$
3 | löse $Mx^{(k)} = Nx^{(k-1)} + b$
4 **until stop**

Die Konvergenz dieses Algorithmus folgt aus Banachschen Fixpunktsatz.

Bemerkung 2.1. Nach gleicher Überlegung lässt sich auch unser obiges Splittingverfahren für Nullstellenbestimmung herleiten:

$$f(x) = 0 \iff h(x) + g(x) := f(x) = 0 \iff h(x) = f(x) - g(x) \iff x = h^{-1}(f(x) - g(x))$$

Wiederholung: Eine Matrixnorm ist eine Norm auf dem Vektorraum der Matrizen, d.h. $\|\cdot\| : \mathbb{K}^{n\times n} \to \mathbb{R}$, bereits bekannte Matrixnormen sind:

- Frobeniusnorm: $\|A\|_F := \left(\sum_{i,j} |a_{ij}|^2\right)^{1/2}$
- Spaltensummennorm $\|A\|_1 := \max_j \sum_i |a_{ij}|$
- Zeilensummennorm $\|A\|_\infty := \max_i \sum_j |a_{ij}|$
- Spektralnorm $\|A\|_2 := \sqrt{\lambda_{max}(A^H A)}, \qquad (A^H := \overline{A}^T)$

Im allgemeinen induziert eine Vektornorm auch immer eine Matrixnorm, diese nennen wir Operatornorm:

$$\|A\|_{\text{op}} := \max_{\|x\|=1} \|Ax\|$$

Die oben aufgelisteten Normen $\|\cdot\|_1, \|\cdot\|_2$ und $\|\cdot\|_\infty$ sind die Operatornormen zu den jeweiligen p-Normen.

Eine Norm $\|\cdot\|$ auf $\mathbb{K}^{n\times n}$ heißt submultiplikativ, falls $\|AB\| \leq \|A\| \cdot \|B\|$ und sie heißt verträglich mit einer Vektornorm $\|\cdot\|_V$, falls $\|Ax\|_V \leq \|A\| \cdot \|x\|_V$.

Operatornormen sind immer submultiplikativ und verträglich zu der Vektornorm, aus welcher sie induziert wurden.

Satz 2.2. Ist $|||\cdot|||$ eine Norm auf $\mathbb{K}^{n\times n}$, die mit einer Vektornorm $\|\cdot\|$ verträglich ist, und ist $\left|\left|\left|M^{-1}N\right|\right|\right| < 1$, dann konvergiert der Algorithmus 1 für jedes für jedes $x^{(0)} \in \mathbb{K}^n$ gegen $A^{-1}b$, d.h. gegen die Lösung des linearen Gleichungssystems $Ax = b$.

Beweis. Sei $\tilde{T}(x) := Tx + c$ mit $T = M^{-1}N$ und $c = M^{-1}b$.

Offensichtlich gilt $\tilde{T} : \mathbb{K}^n \to \mathbb{K}^n$, sowie

$$\left\|\tilde{T}(x) - \tilde{T}(y)\right\| = \|Tx - Ty\| \leq |||T||| \cdot \|x - y\|$$

Da $|||T||| = \left|\left|\left|M^{-1}N\right|\right|\right| < 1$, ist $\tilde{T}$ eine k-kontraktive Selbstabbildung und somit konvergiert die Folge (x^k) aus dem Algorithmus gegen den eindeutigen Fixpunkt $\hat{x}$ mit $\tilde{T}(\hat{x}) = \hat{x}$.

Einsetzen der Definition von $\tilde{T}$ liefert:

$$\hat{x} = T\hat{x} + c = M^{-1}(N\hat{x} + b) \implies M\hat{x} = N\hat{x} + b \implies A\hat{x} = (M - N)\hat{x} = b$$

□

Korollar 2.3. Sei A invertierbar, so konvergiert der obige Algorithmus genau dann für alle Startwerte $x^{(0)} \in \mathbb{K}^n$ gegen $\hat{x} = A^{-1}b$, wenn für den Spektralradius $\rho(T) = \max\{|\lambda| : \lambda \in \sigma(T)\}$ die Ungleichung $\rho(T) < 1$ erfüllt ist.

Beweis.

⇐: Falls $\rho(T) < 1$ dann existiert eine Norm $\|\cdot\|_\varepsilon$ auf $\mathbb{K}^n$ und eine dadurch induzierte Operatornorm $|||\cdot|||_\varepsilon$ auf $\mathbb{K}^{n\times n}$ mit $|||T|||_\varepsilon \leq \rho(T) + \varepsilon < 1$ für ε klein genug (vgl. Aufgabe 2.3).
Satz 2.2 liefert dann die Konvergenz des Algorithmus.

⇒: Angenommen $\rho(T) \geq 1$, d.h. es existiert ein Eigenwert λ von T mit $|\lambda| \geq 1$ und zugehörigem Eigenvektor z. Für $x^{(0)} = \hat{x} + z$ und festem k ergibt sich der Iterationsfehler

$$x^{(k)} - \hat{x} = Tx^{(k-1)} + c - \hat{x} = Tx^{(k-1)} - T\hat{x} = T(x^{(k-1)} - \hat{x})$$

Induktiv folgt

$$x^{(k)} - \hat{x} = T^k(x^0 - \hat{x}) = T^k z = \lambda^k z$$

und es gilt $\left\|x^{(k)} - \hat{x}\right\| = |\lambda^k| \cdot \|z\|$.
Für größer werdendes k kann $x^{(k)}$ also nicht gegen $\hat{x}$ konvergieren. □

Satz 2.4. Unter gleichen Voraussetzungen des vorangegangenen Korollars gilt

$$\max_{x^{(0)}\in\mathbb{K}^n} \limsup_{k\to\infty} \left\|\hat{x} - x^{(k)}\right\|^{1/k} = \rho(T)$$

Beweis. Aus dem Beweis von Korollar 2.3 geht hervor, dass

$$\max_{x^{(0)}\in\mathbb{K}^n} \limsup_{k\to\infty} \left\|\hat{x} - x^{(k)}\right\|^{1/k} \geq \limsup_{k\to\infty} \left\|T^k z\right\|^{1/k} = \limsup_{k\to\infty} |\lambda| \cdot \|z\|^{1/k} = |\lambda| = \rho(T)$$

Mit der Norm $\|\cdot\|_\varepsilon$ gilt nun Für jeden Startwert $x^{(0)} \in \mathbb{K}^n$:

$$\left\|x^{(k)} - \hat{x}\right\|_\varepsilon = \left\|T^k(x^{(0)} - \hat{x})\right\|_\varepsilon \leq \|T\|_\varepsilon^k \cdot \left\|x^{(0)} - \hat{x}\right\|_\varepsilon$$

Da im $\mathbb{K}^n$ alle Normen äquivalent sind, also insbesondere auch $\|\cdot\|_\varepsilon$ und $\|\cdot\|$, existiert eine Konstante $c_\varepsilon > 0$ mit

$$\left\|x^{(k)} - \hat{x}\right\|^{1/k} \leq \left(c_\varepsilon \cdot \left\|x^{(k)} - \hat{x}\right\|_\varepsilon\right)^{1/k} \leq \|T\|_\varepsilon \cdot \left(c_\varepsilon \cdot \left\|x^{(0)} - \hat{x}\right\|_\varepsilon\right)^{1/k} \xrightarrow{k\to\infty} \|T\|_\varepsilon$$

Folglich ist

$$\varrho(T) \leq \max_{x^{(0)}} \limsup_{k\to\infty} \left\|x^{(k)} - \hat{x}\right\|^{1/k} \leq \|\|T\|\|_\varepsilon$$

□

Dieser Satz liefert den Konvergenzfaktor und die Konvergenzrate des Splittingverfahren, zur Erinnerung für eine Folge $(x^{(k)})$ mit Grenzwert $\hat{x}$ ist:

- Konvergenzfaktor q: $|x^{(k)} - \hat{x}| \approx C \cdot q^k$ für k groß
- Konvergenzrate r: $r = -\log_{10}(q)$

Korollar 2.5. Die Zahl $\varrho(T)$ ist der (asymptotischer) Konvergenzfaktor von der Iteration $x^{(k)} = Tx^{(k-1)} + c$.
Die (asymptotische) Konvergenzrate lässt sich daher ausdrücken durch $r = -\log_{10} \varrho(T)$

2.2 Zwei einfache Iterationsverfahren

Mittels der Zerlegung $A = D - L - R$, wobei D die Diagonale, $-L$ die untere (linke) Hälfte und $-R$ die obere (rechte) Hälfte der Matrix A sind, erhalten wir einen Spezialfall des Splittingverfahren.

Durch die Wahl $M = D$ und $N = L + R$ ergibt sich

$$x^{(k+1)} = D^{-1}(b + (L + R)x^{(k)})$$

bzw. in algorithmischer Form:

Algorithmus 2: Jacobi / Gesamtschritt Verfahren

Gegeben sei das Lineare Gleichungssystem $Ax = b$ mit $a_{ii} \neq 0$.

Initialisierung: Wähle beliebigen Startvektor $x^{(0)} \in \mathbb{K}^n$

for $k = 0, 1, \ldots$
  for $i = 1, \ldots, n$
    $x_i^{(k+1)} \leftarrow \frac{1}{a_{ii}} \left(b_i - \sum_{i \neq j} a_{ij} x_j^{(k)} \right)$
  end
until stop

Die zugehörige Iterationsmatrix ist hierbei $\mathcal{J} = M^{-1}N = D^{-1}(L+R)$ und nennt sich (beim Jacobi Verfahren) Gesamtschrittoperator.

Einen weitere Version des Splitting-Verfahren ergibt sich durch die Wahl $M = D - L$ und $N = R$. Hierbei bildet $D - L$ eine obere Dreiecksmatrix und die Inversion ergibt sich mittels Vorwärtssubstitution:

Algorithmus 3: Gauss-Seidel / Einzelschritt Verfahren

Gegeben sei das Lineare Gleichungssystem $Ax = b$ mit $a_{ii} \neq 0$.

Initialisierung: Wähle beliebigen Startvektor $x^{(0)} \in \mathbb{K}^n$

for $k = 0, 1, \ldots$
  for $i = 1, \ldots, n$
    $x_i^{(k+1)} \leftarrow \frac{1}{a_{ii}} \left(b_i - \sum_{j<i} a_{ij} x_j^{(k+1)} - \sum_{j>i} a_{ij} x_j^{(k)} \right)$
  end
until stop

Auch hier ergibt sich eine Matrixschreibweise: Durch umstellen erhalten wir:

$$a_{ii} x^{(k+1)} + \sum_{j<i} a_{ij} x_j^{(k+1)} = b_i - \sum_{j>i} a_{ij} x_j^{(k)}$$

und damit (da L und R jeweils die negativen Parts von A sind)

$$(D - L) x^{(k+1)} = b + R x^{(k)}$$

Die hier erhaltene Iterationsmatrix nennen wir Einzelschrittoperator $\mathcal{L} = (D-L)^{-1}R$

Mittels der Zeilensummennorm erhalten wir nun ein leicht prüfbares Konvergenzkriterium:

Satz 2.6. Ist $A \in \mathrm{GL}_n(\mathbb{K})$ strikt diagonaldominant, d.h. $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$, dann konvergieren Gesamtschritt- und Einzelschrittverfahren für alle Startwerte $x^{(0)} \in \mathbb{K}^n$ gegen die eindeutige Lösung von $Ax = b$.

Beweis.
Da A strikt diagonaldominant ist, muss $a_{ii} \neq 0$ und damit sind beide Verfahren wohldefiniert.

Für die Konvergenz wird Satz 2.2 mit der Zeilensummennorm $\|\cdot\|_\infty$ verwendet:

a) Gesamtschrittverfahren: Für die Iterationsmatrix gilt

$$\|\mathcal{J}\|_\infty = \left\|D^{-1}(L+R)\right\|_\infty = \max_{i \in [n]} \frac{1}{|a_{ii}|} \sum_{j \neq i} |a_{ij}| =: q < 1$$

Nach Satz 2.2 folgt damit die Konvergenz des Gesamtschrittverfahren.

b) Einzelschrittverfahren: Um $\|\mathcal{L}\|_\infty < 1$ zu zeigen, nutzen wir, dass die Zeilensummennorm eine durch die Maximumsnorm induzierte Operatornorm induziert ist, d.h.

$$\|\mathcal{L}\|_\infty = \max_{\|x\|_\infty = 1} \|\mathcal{L}x\|_\infty$$

Sei $y = \mathcal{L}x$ für ein $x \in \mathbb{K}^n$ mit $\|x\|_\infty = 1$.

Induktiv folgt nun $y_i \leq q < 1$, wobei der Induktionsanfang aus dem Beweisteil a) folgt.

Unter der Induktionsvoraussetzung gilt für $j < i$, dass $|y_j| \leq q < 1$ und damit ergibt sich aus Zeile 3 von Algorithmus 3 mit $b = 0, x^{(k)} = x$ und $x^{(k+1)} = y$:

$$\begin{aligned}
\|y_i\| &= \left\| \frac{1}{a_{ii}} \Big(0 - \sum_{j<i} a_{ij} y_j - \sum_{j>i} a_{ij} x_j \Big) \right\| \\
&\leq \frac{1}{|a_{ii}|} \Big(\sum_{j<i} |a_{ij}| \cdot \underbrace{|y_j|}_{\leq q < 1} + \sum_{j>i} |a_{ij}| \cdot \underbrace{|x_j|}_{\leq \|x\|_\infty = 1} \Big) \\
&< \frac{1}{|a_{ii}|} \Big(\sum_{j<i} |a_{ij}| + \sum_{j>i} |a_{ij}| \Big) \\
&= \frac{1}{|a_{ii}|} \sum_{j \neq i} |a_{ij}| \\
&= q
\end{aligned}$$

Da dies für alle Einträge von y gilt, folgt $\|y\|_\infty = \|\mathcal{L}x\|_\infty \leq q$ für alle x mit $\|x\|_\infty = 1$ und damit $\|\mathcal{L}\|_\infty \leq q < 1$ □

Beispiel 2.7. Gegeben sei das LGS $Ax = b$ mit

$$A = \begin{pmatrix} 2 & 0 & 1 \\ 1 & -4 & 1 \\ 0 & -1 & 2 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 4 \\ -1 \end{pmatrix}$$

Dieses System hat die eindeutige Lösung $\hat{x} = (1, -1, -1)^T$.

Durch die Wahl $x^{(0)} = (1, 1, 1)^T$ erhalten wir beim Gesamtschritt- bzw. Einzelschrittverfahren:

$$x_{\mathcal{J}}^{(1)} = D^{-1}(b - (L+R)x^{(0)}) = \begin{pmatrix} 2 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & 2 \end{pmatrix}^{-1} \cdot \left[\begin{pmatrix} 1 \\ 4 \\ -1 \end{pmatrix} - \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right] = \begin{pmatrix} 0 \\ -\frac{1}{2} \\ 0 \end{pmatrix}$$

$$x_{\mathcal{L}}^{(1)} = (D-L)^{-1}(b + Rx^{(0)}) = \begin{pmatrix} 2 & 0 & 0 \\ 1 & -4 & 0 \\ 0 & -1 & 2 \end{pmatrix}^{-1} \cdot \left[\begin{pmatrix} 1 \\ 4 \\ -1 \end{pmatrix} - \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right] = \begin{pmatrix} 0 \\ \frac{3}{4} \\ -\frac{7}{8} \end{pmatrix}$$

2.3 Gradientenverfahren

2.3.1 Gradientenverfahren für Optimierung

Eine Funktion $f : \mathbb{R}^n \to \mathbb{R}$ soll minimiert werden. Von einem Startpunkt $x^{(0)}$ ausgehen bewegen wir uns Stück für Stück in Richtung des steilsten Abstiegs, intuitiv sollten wir so ein Minimum finden.

Als Iterationsvorschrift ergibt sich daher

$$x^{(k+1)} = x^{(k)} + \alpha^{(k)} \cdot d^{(k)}, \quad \text{für} k = 0, 1, \ldots$$

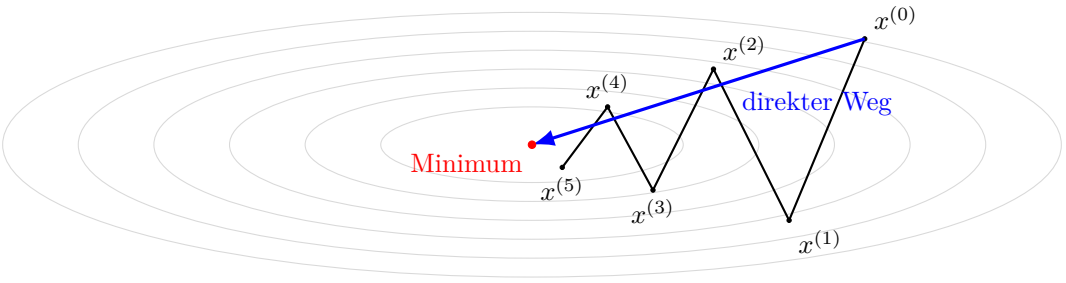
dabei ist $\alpha^{(k)} > 0$ die Schrittweite und $d^{(k)} \in \mathbb{R}^n$ die Abstiegsrichtung. Eine typische Wahl der Abstiegsrichtung ist $d^{(k)} = -\frac{\partial f}{\partial x}(x^{(k)}) = -\nabla f(x^{(k)})$, eine Verfeinerung würde man noch mit $d^{(k)} = -D^{(k)} \cdot \nabla f(x^{(k)})$ erhalten, wobei $D^{(k)} \in \mathbb{R}^n$.

Das Ziel ist des Verfahren ist es, dass sich der Wert von f in jedem Schritt verbessert, d.h. $f(x^{(k+1)}) < f(x^{(k)})$.

Es ergibt sich somit ein eindimensionales Optimierungsproblem um die optimale Schrittweite $\alpha^{(k)}$ zu erhalten:

$$\alpha^{(k+1)} = \min_{\alpha \neq 0} \{ f(x^{(k)} + \alpha \cdot d^{(k)}) \}$$

Ein Nachteil des Verfahrens ist die mögliche Entstehung oszillierender Pfade („Zick-Zack-Verhalten“) aufgrund unvorteilhafter Richtungen (Richtungen noch nicht orthogonal zu Niveaulinien, überarbeiten!):



2.3.2 Das Verfahren der konjugierten Gradienten

Die obige Idee kann zur effizienten Lösung linearer Gleichungssysteme genutzt werden. Gegeben sei das LGS $Ax = b$ mit $A \in \mathbb{K}^{n \times n}$ hermitisch ($a_{ij} = \overline{a_{ji}}$). Das impliziert insbesondere, dass die Hauptdiagonale reell ist, und dass $x^H Ay = y^H Ax$ für $x, y \in \mathbb{K}^n$ gilt.

Zur Lösung wird dieses Mal die Minimierung folgendes quadratischen Funktionals betrachtet

$$\phi(x) = \tfrac{1}{2} x^H Ax - x^H b \tag{1}$$

Denn sollte eine Lösung $\hat{x} = A^{-1}b$ des LGS $Ax = b$ existieren, so gilt für alle $x \in \mathbb{K}^{n \times n}$:

$$\begin{aligned}
&\phi(x) - \phi(\hat{x}) \\
=&\tfrac{1}{2} x^H Ax - x^H b - (\tfrac{1}{2} \hat{x}^H A\hat{x} - \hat{x}^H b) \\
=& \tfrac{1}{2} x^H Ax - x^H b - (\tfrac{1}{2} \hat{x}^H A\hat{x} - \hat{x}^H b) + \overbrace{\tfrac{1}{2}(x-\hat{x})^H A(x-\hat{x}) - \tfrac{1}{2} x^H Ax + \tfrac{1}{2} x^H A\hat{x} + \tfrac{1}{2} \hat{x}^H Ax - \tfrac{1}{2} \hat{x}^H A\hat{x}}^{=0} \\
=& -x^H b + \hat{x}^H b + \tfrac{1}{2}(x-\hat{x})^H A(x-\hat{x}) - \hat{x}^H A\hat{x} + x^H A\hat{x} \\
=& -x^H b + \hat{x}^H b + \tfrac{1}{2}(x-\hat{x})^H A(x-\hat{x}) - \hat{x}^H b + x^H b \\
=& \tfrac{1}{2}(x-\hat{x})^H A(x-\hat{x}) \\
\geq& 0
\end{aligned}$$

Die Funktion hat demnach ein eindeutiges Minimum bei $\hat{x}$.

Definition 2.8. Ist $A \in \mathbb{K}^{n \times n}$ hermitisch und pos. definitiv, dann wird durch $\|x\|_A = \sqrt{x^H Ax}, x \in \mathbb{K}^{n \times n}$ eine Norm in $\mathbb{K}^n$ definiert, die sogenannte Energienorm. Zur Energienorm gehört ein inneres Produkt $\langle x, y \rangle_A = x^H Ay, x, y \in \mathbb{K}^n$. Es ergibt sich die Abweichung des Funktionals von seinem Minimum als:

$$\phi(x) - \phi(\hat{x}) = \tfrac{1}{2} \|x - \hat{x}\|_A^2 \tag{2}$$

geometrische Interpretation: Der Graph von ϕ bezüglich der Energienorm ist ein kreisförmiger Paraboloid, welcher über dem Mittelpunkt $\hat{x}$ liegt.

Idee: Konstruktion eines Verfahrens, welches die Lösung $\hat{x}$ von $Ax = b$ iterativ approximiert, indem das Funktional ϕ sukzessiv minimiert wird:

Zur aktuellen Iteration $x^{(k)}$ wird die Suchrichtung $d^{(k)} \neq 0$ bestimmt, und die neue Iterierte $x^{(k+1)}$ über den Ansatz

$$x^{(k+1)} = x^{(k)} + \alpha \cdot d^{(k)} \tag{3}$$

bestimmt (gleiche Ansatz wie zuvor). Es gilt dann

$$\phi(x^{(k)} + \alpha d^{(k)}) = \phi(x^{(k)}) + \alpha d^{(k)H} A x^{(k)} + \tfrac{1}{2}\alpha^2 d^{(k)H} A d^{(k)} - \alpha d^{(k)H} \cdot b \tag{4}$$

Durch Differentiation nach α und Null setzen der Ableitung ergibt sich unsere Schrittweite $\alpha^{(k)}$:

$$\alpha^{(k)} = \frac{r^{(k)H} d^{(k)}}{d^{(k)H} A d^{(k)}}, \qquad \text{mit } r^{(k)} = b - Ax^{(k)} \tag{5}$$

zusätzlich erhalten wir die Suchrichtung $d^{(k)}$ indem wir die Richtungsableitung von ϕ betrachten, dabei gilt:

$$\begin{aligned}
\frac{\partial}{\partial d^{(k)}}\phi(x^{(k)}) &= \lim_{\alpha\to 0} \frac{\phi(x^{(k)} + \alpha d^{(k)}) - \phi(x^{(k)})}{\alpha} \\
&\overset{(4)}{=} \lim_{\alpha\to 0} \frac{\alpha d^{(k)H} A x^{(k)} + \frac{1}{2}\alpha^2 d^{(k)H} A d^{(k)} - \alpha d^{(k)H} \cdot b}{\alpha} \\
&= d^{(k)H}(Ax^{(k)} - b) + \lim_{\alpha\to 0} \tfrac{1}{2}\alpha d^{(k)H} A d^{(k)} \\
&= -d^{(k)H} r^{(k)}
\end{aligned}$$

Ziel ist es nun die Idee des Verfahren des steilsten Abstiegs zu verbessern, indem wir nicht nur in Richtung $r^{(k)}$ wandern (was zwar der größtmögliche Abstieg wäre, aber unvorteilhaftes „Zick-Zack-Verhalten“ mit sich bringt), sondern stattdessen

$$d^{(k)} = r^{(k)} + \beta^{(k-1)} d^{(k-1)} \tag{6}$$

Dabei sollte $\beta^{(k-1)}$ gerade so gewählt sein, dass $d^{(k)}$ und $d^{(k-1)}$ konjugiert bezüglich A sind, d.h. $\langle d^{(k-1)}, d^{(k)} \rangle = 0$. Aus dieser Bedingung ergibt sich dann

$$\beta^{(k-1)} = -\frac{r^{(k)H} A d^{(k-1)}}{d^{(k-1)H} A d^{(k-1)}} \tag{7}$$

Die Gleichungen (5) und (7) sind wohldefiniert, wenn $d^{(k)^*} A d^{(k)} \neq 0$, aufgrund der positiv Definitheit von A ist dies genau dann der Fall wenn $d^{(k)} \neq 0$.

Nach (6) ist $d^{(k)} = 0$ jedoch nur dann möglich, wenn $r^{(k)}$ und $d^{(k-1)}$ linear abhängig sind, doch nach Konstruktion verläuft die Suchrichtung $d^{(k-1)}$ tangential zur Niveaufläche von ϕ, also orthogonal zum Gradienten $r^{(k)}$.

Somit folgt $d^{(k)} = 0$ nur wenn $r^{(k)} = 0$, was $x^{(k)} = \hat{x}$ implizieren würde.

2.3.3 Eigenschaften des CG-Verfahrens

Wegen der Orthogonalitätsbedingung $\langle d^{(k+1)}, d^{(k)} \rangle_A = 0$ nennt man die Suchrichtungen zueinander A-konjugiert und das Verfahren, Verfahren der konjugierten Gradienten (CG-Verfahren). Zusätzlich zur Konjugiertheit von sukzessiven Suchrichtungen ergeben sich folgende weitere Eigenschaften:

Lemma 2.9. Sei $x^{(0)}$ ein beliebiger Startvektor und $d^{(0)} = r^{(0)} = b - Ax^{(0)}$.
Wenn $x^{(k)} \neq \hat{x}$ für $k = 0, 1, \ldots, m$, mit $A\hat{x} = b$, so gilt:

a) $r^{(m)H} d^{(j)} = 0$ für $0 \leq j < m$

b) $r^{(m)H} r^{(j)} = 0$ für $0 \leq j < m$

b) $\langle d^{(m)}, d^{(j)} \rangle_A = 0$ für $0 \leq j < m$ (10)

Beweis.
Für $k \geq 0$ gilt mit (3), dass $Ax^{(k+1)} = Ax^{(k)} + \alpha^{(k)} Ad^{(k)}$ und somit

$$\begin{aligned} r^{(k+1)} &= b - Ax^{(k+1)} \\ &= b - Ax^{(k)} - \alpha^{(k)} Ad^{(k)} \\ &= r^{(k)} - \alpha^{(k)} Ad^{(k)} \end{aligned} \tag{8}$$

die nach (5) definierte optimale Wahl für α bewirkt dann, dass

$$\begin{aligned} r^{(k+1)H} d^{(k)} &= (r^{(k)} - \alpha^{(k)} Ad^{(k)})^H d^{(k)} \\ &= r^{(k)H} d^{(k)} - \alpha^{(k)} d^{(k)H} Ad^{(k)} \\ &= r^{(k)H} d^{(k)} - \frac{r^{(k)H} d^{(k)}}{d^{(k)H} Ad^{(k)}} d^{(k)H} Ad^{(k)} \\ &\overset{(5)}{=} 0 \end{aligned} \tag{9}$$

Weiter gilt nach Induktion über m:

Induktionsanfang: $m = 1$.
Durch die Wahl $k = 0$ in (9) erhalten wir die Behauptung (a) und nach Setzung $d^{(0)} = r(0)$ auch die Behauptung (b). (c) folgt im Fall $m = 1$ direkt aus der Orthogonalitätsbedingung.

Induktionsschritt: $m \to m + 1$.
Wir nehmen an, dass die Aussagen (a), (b) und (c) für $m \leq \overline{m}$ richtig sind und zeigen damit die Gültigkeit für $m = \overline{m} + 1$.
Zunächst folgt aus (9) mit $k = \overline{m}$, dass $r^{(\overline{m}+1)H} d^{(\overline{m})} = 0$, sowie aus der Darstellungen $r^{(m+1)}$ von (8) mit der Induktionsannahme (a und c):

$$r^{(\overline{m}+1)H} d^{(j)} = r^{(\overline{m})H} d^{(j)} - \alpha^{(\overline{m})} \langle d^{(\overline{m})}, d^{(j)} \rangle_A = 0 \text{ für } 0 \leq j < \overline{m}$$

Damit gilt (a) auch für $m = \overline{m} + 1$.

Weiter ergibt (6) umgestellt $r^{(j)} = d^{(j)} - \beta^{(j-1)} d^{(j-1)}$ und mit $r^{(0)} = d^{(0)}$ folgt daher (b) rekursiv aus (a):

$$r^{(\overline{m}+1)H} r^{(j)} = r^{(\overline{m}+1)H} d^{(j)} - \beta^{(j-1)} \cdot r^{(\overline{m}+1)H} d^{(j-1)} = 0 - \beta^{(j-1)} \cdot 0 = 0$$

Damit (c) gilt muss noch $\alpha^{(j)} \neq 0$ für $j < \overline{m}$ sein (der Fall $j = \overline{m}$ ergibt sich direkt aus der Orthogonalitätsbedingung), denn dann ergibt (8):

$$Ad^{(j)} = \frac{1}{\alpha^{(j)}} \left(r^{(j+1)} - r^{(j)} \right)$$

und somit folgt durch (6):

$$\begin{aligned} \langle d^{(\overline{m}+1)}, d^{(j)} \rangle_A &= \langle r^{(\overline{m}+1)}, d^{(j)} \rangle_A + \beta^{(\overline{m})} \underbrace{\langle d^{(\overline{m})}, d^{(j)} \rangle_A}_{=0} \\ &= r^{(\overline{m}+1)H} Ad^{(j)} \\ &= r^{(\overline{m}+1)H} \cdot \frac{1}{\alpha^{(j)}} \left(r^{(j+1)} - r^{(j)} \right) \\ &= \frac{1}{\alpha^{(j)}} \left(\underbrace{r^{(\overline{m}+1)H} r^{(j+1)}}_{=0} - \underbrace{r^{(\overline{m}+1)H} r^{(j)}}_{=0} \right) \\ &= 0 \end{aligned}$$

Angenommen $\alpha^{(j)} = 0$, dann folgt aus der Definition von $\alpha^{(j)}$ (5), dass auch $r^{(j)^*} d^{(j)} = 0$ und mit (6)

$$0 = r^{(j)H} d^{(j)} = r^{(j)H} \left(r^{(j)} + \beta^{j-1} d^{(j-1)} \right) = r^{(j)H} r^{(j)} + \beta^{(j-1)} \underbrace{r^{(j)H} d^{(j-1)}}_{=0} = \left\| r^{(j)} \right\|_2^2$$

Das impliziert $r^{(j)} = 0$, doch dann wäre aber $x^{(j)} = \hat{x}$ (Widerspruch). □

Das Lemma sagt insbesondere aus, dass alle Suchrichtungen paarweise A-konjugiert. Außerdem bilden die Residuen ein Orthogonalsystem und sind damit linear unabhängig. Es muss sich daher nach spätestens n Schritten $r^{(n)} = 0$, also $x^{(n)} = \hat{x}$ ergeben.

Korollar 2.10. Für $A \in \mathbb{K}^{n\times n}$ hermitisch und positiv definit findet das CG-Verfahren nach höchstens n Schritten die exakte Lösung $x^{(n)} = \hat{x}$.

In der Praxis ist dieses Korollar nicht relevant, da häufig wesentlich weniger Schritte benötigt werden oder die Orthogonalitätsbedingung aufgrund von Rundungsfehlern verloren gehen.

Eine alternative Interpretation als iteratives Verfahren (neben der geometrischen Idee) bietet die Theorie über Krylov-Räume:

Definition 2.11. Sei $A \in \mathbb{K}^{n\times n}$ und $y \in \mathbb{K}^n$. Dann heißt der Unterraum

$$\mathcal{K}_k(A, y) := \operatorname{span}\{y, Ay, \ldots, A^{k-1}y\} \leq \mathbb{K}^n$$

Krylov-Raum der Dimension k von A bezüglich y.

Satz 2.12. Sei $A \in \mathbb{K}^{n\times n}$ hermitisch und positiv definit, $d^{(0)} = r^{(0)}$, und $x^{(k)} \neq \hat{x}$ die k-te Iterierte des CG-Verfahrens. Dann gilt

$$x^{(k)} \in x^{(0)} + \mathcal{K}_k(A, r^{(0)})$$

und $x^{(k)}$ ist in diesem affinen Raum die eindeutige Minimalstelle der Zielfunktion ϕ. (Optimalitätseigenschaft)

Beweis.

a) Wir beginnen damit induktiv zu zeigen, dass

$$d^{(j)} \in \operatorname{span}\{r^{(0)}, \ldots, r^{(j)}\} \qquad \text{für } j = 0, \ldots, k+1 \tag{11}$$

Induktionsanfang: $j = 0$.
Wegen $d^{(0)} = r^{(0)}$ offensichtlich erfüllt.
Induktionsschritt: $j \to j+1$. Folgt direkt aus der Rekursionsvorschrift für d (6):

$$d^{(j+1)} = r^{(j+1)} + \beta^{(j)} \cdot d^{(j)} = r^{(j+1)} + \beta^{(j)} \sum_{i=0}^{j} \delta_i r^{(i)} \in \operatorname{span}\{r^{(0)}, \ldots, r^{(j+1)}\}$$

Es folgt damit $\operatorname{span}\{d^{(0)}, \ldots, r^{(k-1)}\} \subset \operatorname{span}\{r^{(0)}, \ldots, r^{(k-1)}\}$ Zusammen mit dem Lemma 2.9 folgt dass beide Systeme jeweils linear unabhängig sind, also jeweils Dimension $k-1$ haben, wodurch Gleichheit gilt.

$$\operatorname{span}\{d^{(0)}, \ldots, r^{(k-1)}\} = \operatorname{span}\{r^{(0)}, \ldots, r^{(k-1)}\}$$

Aus der Rekursionsformel von x (3) ergibt sich der explizit Ausdruck:

$$x^{(k)} = x^{(0)} + \sum_{j=0}^{k-1} \alpha^{(j)} \cdot d^{(j)} \in x^{(0)} + \operatorname{span}\{r^{(0)}, \ldots, r^{(k-1)}\}, \quad \text{für } j = 0, \ldots, k-1$$

Im nächsten Schritt wird induktiv gezeigt, dass $r^{(j)} \in \mathcal{K}_j(A, r^{(0)})$:

Induktionsanfang: $j = 0$.
Offensichtlich gilt $r^{(0)} \in \operatorname{span}\{r^{(0)}\}$. Induktionsschritt: $j-1 \to j$.
Aus Teil (a) und der Induktionsannahme folgt

$$\begin{aligned} & d^{(j-1)} \in \operatorname{span}\{r^{(0)}, \dots, r^{(j-1)}\} \subset \operatorname{span}\{r^{(0)}, \dots, A^{j-1}r^{(0)}\} \\ \overset{(8)}{\Longrightarrow} \quad & r^{(j)} = r^{(j-1)} - \alpha^{(j-1)} A d^{(j-1)} \in \operatorname{span}\{r^{(0)}, \dots, A^{j}r^{(0)}\} \end{aligned}$$

Damit folgt $\operatorname{span}\{r^{(0)}, \dots, r^{(k-1)}\} \subset \mathcal{K}_j(A, r^{(0)})$. Die Menge $\{r^{(j)}\}_{j=0}^{k-1}$ ist linear unabhängig und daher hat der linke Unterraum die Dimension k, es folgt Gleichheit

$$\operatorname{span}\{d^{(0)}, \dots, r^{(k-1)}\} = \operatorname{span}\{r^{(0)}, \dots, r^{(k-1)}\} = \mathcal{K}_j(A, r^{(0)})$$

und damit auch $x^{(k)} \in x^{(0)} + \mathcal{K}_k(A, r^{(0)})$.

c) Aus Korollar 2.10 folgt die Existenz eines Iterationsindex $m \leq n$ mit

$$\hat{x} = x^{(0)} + \sum_{j=0}^{m-1} \alpha^{(j)} \cdot d^{(j)}$$

Für ein $0 \leq k \leq m$ gilt dann:

$$\hat{x} = x^{(0)} + \sum_{j=0}^{k-1} \alpha^{(j)} \cdot d^{(j)} + \sum_{j=k}^{m-1} \alpha^{(j)} \cdot d^{(j)} = x^{(k)} + \sum_{j=k}^{m-1} \alpha^{(j)} \cdot d^{(j)}$$

Und für ein beliebiges $x \in x^{(0)} + \mathcal{K}_k(A, r^{(0)}) = x^{(0)} + \operatorname{span}\{d^{(0)}, \dots, d^{(k-1)}\}$ demnach

$$\begin{aligned} \hat{x} - x &= \hat{x} - x^{(k)} + x^{(k)} - x \\ &= x^{(0)} + \sum_{j=1}^{m-1} \alpha^{(j)} \cdot d^{(j)} - \sum_{j=0}^{k-1} \alpha^{(j)} \cdot d^{(j)} + \sum_{j=0}^{k-1} \alpha^{(j)} \cdot d^{(j)} - x^{(0)} - \sum_{j=0}^{k-1} \delta_j \cdot d^{(j)} \\ &= \sum_{j=k}^{m-1} \alpha^{(j)} \cdot d^{(j)} + \sum_{j=0}^{k-1} (-\alpha^{(j)} - \delta_j) \cdot d^{(j)} \\ &= \hat{x} - x^{(k)} + \sum_{j=0}^{k-1} (-\alpha^{(j)} - \delta_j) \cdot d^{(j)} \end{aligned}$$

für $\delta_j \in \mathbb{K}$. Da die Suchrichtungen nach Lemma 2.9 A-konjugiert sind folgt:

$$\begin{aligned} \phi(x) - \phi(\hat{x}) &= \tfrac{1}{2}\|\hat{x} - x\|_A^2 \\ &= \frac{1}{2}\left\| \sum_{j=k}^{m-1} \alpha^{(j)} \cdot d^{(j)} + \sum_{j=0}^{k-1} (-\alpha^{(j)} - \delta_j) \cdot d^{(j)} \right\|_A^2 \\ &= \frac{1}{2}\left\| \sum_{j=k}^{m-1} \alpha^{(j)} \cdot d^{(j)} \right\|_A^2 + \frac{1}{2}\left\| \sum_{j=0}^{k-1} (-\alpha^{(j)} - \delta_j) \cdot d^{(j)} \right\|_A^2 \\ &= \tfrac{1}{2}\left\| \hat{x} - x^{(k)} \right\|_A^2 + \frac{1}{2}\left\| \sum_{j=0}^{k-1} (-\alpha^{(j)} - \delta_j) \cdot d^{(j)} \right\|_A^2 \\ &= \phi(\hat{x}) - \phi(x^{(k)}) + \frac{1}{2}\left\| \sum_{j=0}^{k-1} (-\alpha^{(j)} - \delta_j) \cdot d^{(j)} \right\|_A^2 \\ &\geq \phi(\hat{x}) - \phi(x^{(k)}) \end{aligned}$$

Insbesondere gilt Gleichheit bei $x = x^{(k)}$. □

2.3.4 Praktische Aspekte der Implementierung

Bemerkung 2.13. Für eine Implementierung des CG-Verfahren sollte man nicht die Gleichungen (5) und (7) für $\alpha^{(k)}$ und $\beta^{(k)}$ verwenden, sondern lieber folgende Darstellungen, welche numerisch stabiler sind:

$$\alpha^{(k)} = \frac{\left\|r^{(k)}\right\|_2^2}{d^{(k)H} A d^{(k)}} \tag{5'}$$

$$\beta^{(k)} = \frac{\left\|r^{(k+1)}\right\|_2^2}{\left\|r^{(k)}\right\|_2^2} \tag{7'}$$

Die Gleichung (5') folgt aus (3) und Lemma 2.9 (a), nach welchen

$$r^{(k)H} d^{(k)} = r^{(k)H} r^{(k)} + \beta^{(k)} \cdot \underbrace{r^{(k)H} d^{(k-1)}}_{=0} = \left\|r^{(k)}\right\|_2^2$$

Die Gleichung (7') folgt dann aus aus (8), (5') und dem Lemma 2.9 (b):

$$r^{(k+1)H} A d^{(k)} = \frac{1}{\alpha^{(k)}} \left(r^{(k+1)H} r^{(k)} - r^{(k+1)H} r^{(k+1)} \right) = \frac{-\left\|r^{(k+1)}\right\|_2^2}{\alpha^{(k)}} = -\frac{\left\|r^{(k+1)}\right\|_2^2}{\left\|r^{(k)}\right\|_2^2} d^{(k)H} A d^{(k)}$$

Es ergibt sich damit folgender Algorithmus:

Algorithmus 4: CG-Verfahren

Initialisierung: $A \in \mathbb{K}^{n \times n}$ sei hermitisch und positiv definit.
Ergebnis: : $x^{(k)}$ als Approximation für $A^{-1}b$,
$r^{(k)} = b - Ax^{(k)}$ als zugehöriges Residuum.

1 Wähle $x^{(0)} \in \mathbb{K}^n$ beliebig
2 $r^{(0)} \leftarrow b - Ax^{(0)}$
3 $d^{(0)} \leftarrow r^{(0)}$
4 **for** $k = 0, 1, \ldots,$
5 $\quad \alpha^{(k)} \leftarrow \dfrac{\left\|r^{(k)}\right\|_2^2}{d^{(k)^*} A d^{(k)}}$
6 $\quad x^{(k+1)} \leftarrow x^{(k)} + \alpha^{(k)} d^{(k)}$
7 $\quad r^{(k+1)} \leftarrow r^{(k)} - \alpha^{(k)} A d^{(k)}$
8 $\quad \beta^{(k)} \leftarrow \dfrac{\left\|r^{(k+1)}\right\|_2^2}{\left\|r^{(k)}\right\|_2^2}$
9 $\quad d^{(k+1)} \leftarrow r^{(k+1)} + \beta^{(k)} d^{(k)}$
10 **until stop**

Der Aufwand des CG-Verfahrens ergibt sich aus einer Matrix-Vektor Multiplikation in jedem Iterationsschritt und ist damit vergleichbar mit dem Gesamt -und Einzelschritt.

Bemerkung 2.14. Das CG-Verfahren ist typischerweise wesentlich schneller als das Gesamt -bzw. Einzelschrittverfahren, **aber** verlangt, dass die vorausgesetzte Matrix hermitisch ist.
Ein schnelles und einfaches Verfahren für allgemeine Matrizen ist derzeit nicht bekannt, ein komplizierteres Verfahren mit ähnlicher Konvergenzgeschwindigkeit ist das GMRES-Verfahren (Vgl. Hanku-Bourgeois, Kapitel 3, Abschnitt 16).

2.4 Präkonditionierung des CG-Verfahren

Definition 2.15. $\kappa_M(A) = \text{cond}_M(A) = \left\|A^{-1}\right\|_M \cdot \|A\|_M$ wird als Kondition der Matrix A bezüglich der Norm $\|\cdot\|_M$ bezeichnet. Sie beschreibt die schlimmstmögliche Fortpflanzung des Eingangsfehlers beim Lösen eines LGS.

Gegeben sei $Az = b$ mit der Lösung $z = A^{-1}b$. Der Einfluss vom Eingangsfehlers sei Δb:

$$z + \Delta z = A^{-1}(b + \Delta b) = A^{-1}b + A^{-1}\Delta b$$

Die berechnete Lösung erhält dabei den Fortpflanzungsfehler $\Delta z = A^{-1}\Delta b$.

Sei $\|\cdot\|$ die zu $\|\cdot\|_M$ verträgliche Matrixnorm (d.h. $\|Ax\| \leq \|A\|_M \cdot \|x\|$), so ergibt sich als relativer Fehler:

$$\begin{aligned}
\frac{\|\Delta z\|}{\|z\|} &= \frac{\|\Delta z\|}{\|b\|} \cdot \frac{\|b\|}{\|z\|} \\
&= \frac{\left\|A^{-1}\Delta b\right\|}{\|b\|} \cdot \frac{\|Az\|}{\|z\|} \\
&\leq \left\|A^{-1}\right\|_M \cdot \|A\|_M \cdot \frac{\|\Delta b\|}{\|b\|} \cdot \frac{\|z\|}{\|z\|} \\
&= \text{cond}_M A \cdot \frac{\|\Delta b\|}{\|b\|}
\end{aligned}$$

Typischerweise ist die Konvergenz eines numerischen Verfahrens umso langsamer, je schlechter die Matrix A konditioniert ist, d.h. je größer die Konditionszahl A ist.

2.4.1 Präkonditionierung mittels Cholesky

Idee: Gleichungssystem $Ax = b$ in ein äquivalentes LGS umwandeln, sodass die Kondition sich verbessert:

$$M^{-1}Ax = M^{-1}b \tag{$\triangle$}$$

wobei M hermitisch und positiv definit ist.

Problem: Die Matrix $M^{-1}A$ muss nicht notwendig hermitisch sein, daher nutzen wir die Cholesky-Zerlegung $M = LL^H$ und erhalten[1]:

$$L^{-H}L^{-1}Ax = L^{-H}L^{-1}b \implies L^{-1}AL^{-*}z = L^{-1}b \quad \text{mit } x = L^{-*}z$$

Hierbei ist die Koeffizientenmatrix $L^{-1}AL^{-H}$ sicher hermitisch und positiv definit, denn $(L^{-1}AL^{-H})^H = (L^{-H})^H A^H (L^{-1})^H = L^{-1}AL^{-H}$ und für beliebiges $z \in \mathbb{K}^n$ und $x = L^{-H}z$ gilt:

$$z^H L^{-1}AL^{-H}z = x^H L^{-H}z = x^H Ax \geq 0$$

Wir können also CG-Verfahren zum Lösen von $L^{-1}AL^{-*}z = L^{-1}b$ nutzen und aus dieser Lösung dann x bestimmen. In der Praxis werden wir stattdessen durch geschickte Modifikation des klassischen CG-Verfahren direkt x berechnen.

Ziel: Die Konditionszahl von $L^{-1}AL^{-*}$ soll kleiner werden als die Konditionszahl von A, dies liefert schnellere Konvergenz der Iterierten $z^{(k)}$ und der Lösung $x^{(k)} = L^{-*}z$

[1] $L^{-H} = (L^H)^{-1} = (L^{-1})^H$

Bemerkung 2.16. Die Faktorisierung $M = LL^*$ muss nicht explizit berechnet werden, da die Variable z wieder durch x substituiert werden kann.

Für die Berechnung der Koeffizienten $\beta^{(k)}$ und die dafür benötigte Norm $\left\|L^{-1}b - L^{-1}AL^{-H}z^{(k)}\right\|_2^2$ wird neben $r^{(k)} = b - A^{(k)}$ noch ein Hilfsvektor

$$s^{(k)} = M^{-1}r^{(k)} = M^{-1}b - M^{-1}Ax^{(k)}$$

benötigt ($s^{(k)}$ beschreibt das Residuum von ($\triangle$)).

Für die Norm gilt

$$\left\|L^{-1}b - L^{-1}AL^{-H}z^{(k)}\right\|_2^2 = \left\|L^{-1}\underbrace{(b - Ax^{(k)})}_{=r^{(k)}}\right\|_2^2 = (r^{(k)H}\underbrace{L^{-H}}_{=s^{(k)}})(L^{-1}r^{(k)}) = r^{(k)H}s^{(k)}$$

2.4.2 Algorithmus: PCG-Verfahren

Algorithmus 5: Präkonditioniertes CG-Verfahren (PCGV)

Initialisierung: $A, M \in \mathbb{K}^{n\times n}$ seien hermitisch und positiv definit.
Ergebnis: : $x^{(k)}$ als Approximation für $A^{-1}b$,
$r^{(k)} = b - Ax^{(k)}$ Residuum im Schritt k,
$s^{(k)}$ das Residuum von ($\triangle$).

Wähle $x^{(0)} \in \mathbb{K}^n$ beliebig
$r^{(0)} \leftarrow b - Ax^{(0)}$
Löse $Ms^{(0)} = r^{(0)}$
$d^{(0)} \leftarrow s^{(0)}$
for $k = 0, 1, \ldots,$
  $\alpha^{(k)} \leftarrow \dfrac{r^{(k)H}s^{(k)}}{d^{(k)H}Ad^{(k)}}$
  $x^{(k+1)} \leftarrow x^{(k)} + \alpha^{(k)}d^{(k)}$
  $r^{(k+1)} \leftarrow r^{(k)} - \alpha^{(k)}Ad^{(k)}$
  Löse $Ms^{(k+1)} = r^{(k+1)}$
  $\beta^{(k)} \leftarrow \dfrac{r^{(k+1)H}s^{(k+1)}}{r^{(k)H}s^{(k)}}$
  $d^{(k+1)} \leftarrow s^{(k+1)} + \beta^{(k)}d^{(k)}$
until stop

Der Aufwand im Vergleich zum CGV erhöht sich beim PCGV um das Lösen eines LGS $Ms = r$. Die erhoffte schnellere Konvergenz des Iterationsverfahren macht sich also nur bezahlt, wenn das LGS $Ms = r$ entsprechend billig gelöst werden kann.

Da A bei Anwendung des CGV typischerweise dünn besetzt ist, dominieren die Kosten für die Lösung des LGS $Ms = r$ bei dem Gesamtkosten des PCGV.

Satz 2.17. Die k-te Iteriterte $x^{(k)}$ vom Algorithmus des PCGV liegt in dem affin verschobene Krylov-Raum $x^{(0)} + \mathcal{K}_k(M^{-1}A, M^{-1}r^{(0)})$ und ist in dieser Menge die eindeutig bestimmte Minimalstelle des Funktionals $\phi(x) = \frac{1}{2}x^HAx - x^Hb$.

Beweis.

Nach Satz 2.12 liegt die entsprechende Iterierte $z^{(k)} = L^Hx^{(k)}$ in dem affin verschobenen Krylov-

Raum $z^{(0)} + \mathcal{K}_k(L^{-1}AL^{-H}, L^{-1}b - L^{-1}AL^{-1}z^{(0)})$ mit $z^{(0)} = L^H x^{(0)}$ und minimiert in dieser Menge das Fehlerfunktional $\psi(z) = \frac{1}{2}L^{-1}AL^{-H}z - z^H L^{-1}b$.

Durch die Transformation $x = L^{-H}z$ werden die Iterierten und die genannten Krylov-Räume aufeinander abgebildet und es gilt $\psi(z) = \phi(x)$. □

Bemerkung 2.18. Die Konstruktion geeigneter Präkonditionsmatrizen M ist eine schwierige Sache.

2.5 Anwendung und Konvergenzgeschwindigkeit des CG-Verfahren

2.5.1 Lösen von Randwertproblemen mittels CG-Verfahren

Bevor wir zu einer Anwendung derartiger iterierten Verfahren starten, wiederholen wir kurz einen wichtigen Satz der Analysis:

Satz 2.19 (Satz von Taylor). Sei $f : [a, b] \to \mathbb{R}$ eine $(n+1)$-mal stetig differenzierbare Funktion und $x, x_0 \in [a, b]$. Dann existiert ein ξ zwischen x und x_0, so dass

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + \underbrace{\frac{f^{(n+1)}(\xi)}{(n+1)!} \cdot (x - x_0)^{n+1}}_{R_n(x,x_0)}$$

Beweis. Analysis II.

Beispiel 2.20. Wir betrachten das Randwertproblem des Laplace Operators[a]:

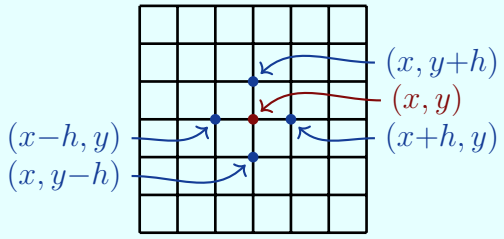
$$-\frac{\partial^2 u(x,y)}{\partial x^2} - \frac{\partial^2 u(x,y)}{\partial y^2} = f(x,y) \quad \text{für } (x,y) \in Q$$

gemeinsam mit der Dirichlet-Randbedingung $u(x, y) = 0$ für $(x, y) \in \partial Q$ auf dem Einheitsquadrat $Q = (0, 1) \times (0, 1) \subset \mathbb{R}^2$.

Die Lösung $u = u(x, y)$ beschreibt z.B. die Auslenkung einer (idealisierten) Membran, die über dem Gebiet Q horizontal gespannt ist und mit einer Kraftdichte f vertikal belastet wird.

Eine Lösung ist im Allgemeinen nicht analytisch angebbar, sodass man auf numerische Näherungslösungen zurückgreifen muss.

Betrachte Q als Quadratgitter:



mit m Knoten und Gitterabstand $h = \frac{1}{m-1}$. Die gesamte Knotenzahl ist $n = m^2$.

Für die Nachbarpunkte um einen Punkte $(x, y) \in Q$ gilt nach Taylorformel:

$$u(x+h,y) \approx u(x,y) + h\frac{\partial u(x,y)}{\partial x} + \tfrac{1}{2}h^2\frac{\partial^2 u(x,y)}{\partial x^2} + \tfrac{1}{6}h^3\frac{\partial^3 u(x,y)}{\partial x^3} + \mathcal{O}(h^4) \tag{1}$$

$$u(x-h,y) \approx u(x,y) - h\frac{\partial u(x,y)}{\partial x} + \tfrac{1}{2}h^2\frac{\partial^2 u(x,y)}{\partial x^2} - \tfrac{1}{6}h^3\frac{\partial^3 u(x,y)}{\partial x^3} + \mathcal{O}(h^4) \tag{2}$$

$$u(x,y+h) \approx u(x,y) + h\frac{\partial u(x,y)}{\partial y} + \tfrac{1}{2}h^2\frac{\partial^2 u(x,y)}{\partial y^2} + \tfrac{1}{6}h^3\frac{\partial^3 u(x,y)}{\partial y^3} + \mathcal{O}(h^4) \tag{3}$$

$$u(x,y-h) \approx u(x,y) - h\frac{\partial u(x,y)}{\partial y} + \tfrac{1}{2}h^2\frac{\partial^2 u(x,y)}{\partial y^2} - \tfrac{1}{6}h^3\frac{\partial^3 u(x,y)}{\partial y^3} + \mathcal{O}(h^4) \tag{4}$$

Das Summieren von (1) und (2), sowie (3) und (4) liefert

$$\begin{aligned}
&u(x+h,y) + u(x-h,y) = 2u(x,y) + h^2\frac{\partial^2 u(x,y)}{\partial x^2} + \mathcal{O}(h^4)\\
\Longrightarrow\ &\frac{\partial^2 u(x,y)}{\partial x^2} = h^{-2}\cdot(u(x+h,y) - 2u(x,y) + u(x-h,y)) + \mathcal{O}(h^2)\\
&u(x,y-h) + u(x,y-h) = 2u(x,y) + h^2\frac{\partial^2 u(x,y)}{\partial y^2} + \mathcal{O}(h^4)\\
\Longrightarrow\ &\frac{\partial^2 u(x,y)}{\partial y^2} = h^{-2}\cdot(u(x,y+h) - 2u(x,y) + u(x,y-h)) + \mathcal{O}(h^2)
\end{aligned}$$

Damit erhalten wir die sogenannte „5-Punktregel“:

$$\begin{aligned}
f(x,y) &= -\frac{\partial^2 u(x,y)}{\partial x^2} - \frac{\partial^2 u(x,y)}{\partial y^2}\\
&= -h^{-2}\cdot(u(x+h,y) + u(x-h,y) + u(x,y+h) + u(x,y-h) - 4u(x,y))
\end{aligned}$$

Durch Berücksichtigung der Randbedingung $u(x, y) = 0$ für $(x, y) \in \partial Q$ ist dies äquivalent zu einem linearen Gleichungssystem $Ax = b$ für den Vektor $x \in \mathbb{R}^n$ der unbekannten Knotenwerte $x_i = u(P_i)$

Die Matrix A hat die Gestalt

$$A = \left.\begin{pmatrix} B & -I & 0 & \dots \\ -I & B & -I & \\ 0 & -I & B & \ddots \\ \vdots & & \ddots & \ddots \end{pmatrix}\right\} n \quad \text{mit } B = \left.\begin{pmatrix} 4 & -1 & 0 & \dots \\ -1 & 4 & -1 & \\ 0 & -1 & 4 & \ddots \\ \vdots & & \ddots & \ddots \end{pmatrix}\right\} m$$

und der Einheitsmatrix I, d.h. $B, I \in \mathbb{R}^{m\times m}$. Die rechte Seite ist $b = h^2(f(P_1), \dots, f(P_n))^T$.

Wir erhalten ein sehr großes LGS mit dünn besetzter Bandmatrix mit Bandbreite $2m+1$, symmetrisch, schwach diagonaldominant, positiv definit. Es biete sich also an unsere iterativen Verfahren zum Lösen anzuwenden.

[a] Sei f eine Funktion in kartesischen Koordinaten (x, y), so ist der Laplace Operator definiert durch

$$\Delta f = \frac{\partial^2 f(x,y)}{\partial x^2} + \frac{\partial^2 f(x,y)}{\partial y^2}$$

2.5.2 Konvergenzgeschwindigkeit des CG-Verfahren

Satz 2.21 (CG-Konvergenz). Sei x die Lösung des linearen Gleichungssystems $Ax = b$. Für das CG-Verfahren gilt die Fehlerabschätzung

$$\left\|x^{(k)} - x\right\|_A \leq 2\left(\frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}}\right)^k \cdot \left\|x^{(0)} - x\right\|$$

wobei $\kappa := \kappa(A) = \text{cond}(A)$ die Konditionszahl von A ist.

Zur Reduktion des Anfangsfehlers um den Faktor ε sind circa

$$k(\varepsilon) \approx \tfrac{1}{2}\sqrt{\kappa(A)} \cdot \ln(\tfrac{2}{\varepsilon})$$

Iterationsschritte erforderlich.

Für den Beweis des Satzes benötigen wir noch einen Hilfssatz:

Hilfssatz 2.22 (polynomiale Normschranke). Sei p ein reelles Polynom k-ten Grades (d.h.$p \in P_k := \mathbb{R}_k[X]$) mit $p(0) = 1$. Wenn p auf einer Menge, die alle Eigenwerte von A enthält beschränkt ist, d.h. $\sup_{\mu \in S} |p(\mu)| \leq M$, dann gilt

$$\left\|x^{(k)} - x\right\|_A \leq M \cdot \left\|x^{(0)} - x\right\|_A$$

Beweis des Hilfssatz.
Durch $\mathcal{K}_k(A, r^{(0)}) = \{p(A)r^{(0)} \,:\, p \in P_{k-1}\}$ erhalten wir

$$\begin{aligned}\left\|x^{(k)} - x\right\|_A &= \min\{\|y - x\|_A : y \in x^{(0)} + \mathcal{K}_k(A, r^{(0)})\} \\ &= \min_{p \in P_{k-1}} \left\|x^{(0)} - x + p(A)r^{(0)}\right\|_A\end{aligned}$$

Wegen $r^{(0)} = Ax^{(0)} - b = A \cdot (x^{(0)} - x)$ folgt dann

$$\begin{aligned}\left\|x^{(k)} - x\right\|_A &= \min_{p \in P_{k-1}} \left\|x^{(0)} - x + p(A) \cdot A \cdot (x^{(0)} - x)\right\| \\ &\leq \min_{p \in P_{k-1}} \|I + A \cdot p(A)\|_A \cdot \left\|x^{(0)} - x\right\|_A\end{aligned}$$

Da $(1 + x \cdot P(x))$ selbst wieder ein Polynom von Grad k mit $p(0) = 1$ ist, lässt sich dies weiter abschätzen zu

$$\left\|x^{(k)} - x\right\|_A \leq \min_{p \in P_k, p(0)=1} \|p(A)\|_A \cdot \left\|x^{(0)} - x\right\|_A$$

mit der von A-Norm (Energienorm)$\|\cdot\|_A$ erzeugten natürlichen Matrixnorm $\|\cdot\|_A$.

Für beliebiges $y \in \mathbb{R}^n$ gilt mit einer Orthonormalbasis $\{w_1, \ldots, w_n\}$ aus Eigenvektoren von A (d.h. $Aw_j = \lambda_j w_j$) die Darstellung:

$$y = \sum_{j=1}^{n} \gamma_j w_j, \quad \gamma_j = \langle y, w_j \rangle$$

und folglich für ein $p(x) = \sum_{i=0}^{k} c_i x^i$:

$$\begin{aligned}p(A)y = p(A)\left(\sum_{j=1}^{n} \gamma_j w_j\right) &= \left(\sum_{j=1}^{n} \gamma_j p(A) w_j\right) \\ = \left(\sum_{j=1}^{n} \gamma_j \sum_{i=0}^{k} c_i A^i w_j\right) &= \left(\sum_{j=1}^{n} \gamma_j \sum_{i=0}^{k} c_i \lambda_j^i w_j\right) = \left(\sum_{j=1}^{n} \gamma_j p(\lambda_j) w_j\right)\end{aligned}$$

Damit ergibt sich:

$$\|p(A)y\|_A^2 = \left\| \Big(\sum_{j=1}^{n} \gamma_j p(\lambda_j) w_j \Big) \right\|_A^2 = \Big\langle \sum_{j=1}^{n} \gamma_j p(\lambda_j) w_j, A \sum_{l=1}^{n} \gamma_l p(\lambda_l) w_l \Big\rangle_2$$
$$= \sum_{j,l=1}^{n} \gamma_j \gamma_l p(\lambda_j) p(\lambda_l) \cdot \langle w_j, A w_l \rangle_2 = \sum_{j=1}^{n} \gamma_j^2 p(\lambda_j)^2 \lambda_j \leq M^2 \sum_{j=1}^{n} \gamma_j^2 \lambda_j = M^2 \|y\|_A^2$$

Dies impliziert

$$\|p(A)\|_A = \sup_{y \in \mathbb{R}^n \setminus \{0\}} \frac{\|p(A)y\|_A}{\|y\|_A} \leq M$$

und damit die Behauptung. □

Beweis zur CG-Konvergenz (Satz 2.21)
Durch Verwendung des Hilfssatz mit $S := [\lambda, \Lambda]$, wobei λ den kleinsten und Λ den größten Eigenwert von A beschreibt, folgt:

$$\left\| x^{(k)} - x \right\|_A \leq \min_{p \in P_k, p(0)=1} \Big(\sup_{\lambda \leq \mu \leq \Lambda} |p(\mu)| \Big) \cdot \left\| x^{(0)} - x \right\|_A$$

Dies ergibt die Behauptung wenn wir noch zeigen können, dass

$$\min_{p \in P_k, p(0)=1} \sup_{\lambda \leq \mu \leq \Lambda} |p(\mu)| \leq 2 \cdot \frac{(1 - \sqrt{\frac{\lambda}{\Lambda}})^k}{(1 + \sqrt{\frac{\lambda}{\Lambda}})^k}$$

Hierbei handelt es sich um ein Problem der Bestapproximation von Polynomen bzgl. der Maximumsnorm (Chebyshev-Approximation).

Die beste Lösung $\overline{p}$ ist gegeben durch

$$\overline{p}(\mu) = \frac{T_k(\frac{\Lambda + \lambda - 2\mu}{\Lambda - \lambda})}{T_k(\frac{\Lambda + \lambda}{\Lambda - \lambda})}$$

wobei T_k das k-te Chebyshev-Polynom auf $[-1, 1]$ ist, für dieses gilt

$$\sup_{\lambda \leq \mu \leq \Lambda} |\overline{p}(\mu)| = T_k \left(\frac{\Lambda + \lambda}{\lambda - \lambda} \right)^{-1}$$

Aus der Darstellungen für die Chebyshev-Polynome

$$T_k(\mu) = \tfrac{1}{2} \left((\mu + \sqrt{\mu^2 - 1})^k + (\mu - \sqrt{\mu^2 - 1})^k \right), \quad \text{für } \mu \in [-1, 1]$$

folgt mittels der Identität

$$\frac{\kappa + 1}{\kappa - 1} + \sqrt{\left(\frac{\kappa + 1}{\kappa - 1} \right)^2 - 1} = \frac{\kappa + 1}{\kappa - 1} + \frac{2\sqrt{\kappa}}{\kappa - 1} = \frac{(\sqrt{\kappa} + 1)^2}{\kappa - 1} = \frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1}$$

die folgende Abschätzung nach unten:

$$T_k \left(\frac{\Lambda + \lambda}{\Lambda - \lambda} \right) = T_k \left(\frac{\kappa + 1}{\kappa - 1} \right) = \frac{1}{2} \left(\left(\frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1} \right)^k + \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \right) \geq \left(\frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1} \right)^k$$

Also ist $\overline{p}(\mu)$ auf $[\lambda, \Lambda]$ durch $2 \left(\frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1} \right)^k$ beschränkt, was die erste Ungleichung des Satzes zeigt.

Für den zweiten Teil betrachten wir die Anzahl der Schritte, so dass

$$2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}-1}\right)^{k(\varepsilon)} < \varepsilon \quad \Longleftrightarrow \quad k(\varepsilon) > \ln(\tfrac{2}{\varepsilon}) \cdot \left(\ln\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}-1}\right)\right)^{-1}$$

Durch die Reihendarstellung $\ln \frac{x+1}{x-1} = 2(\frac{1}{x} + \frac{1}{3}\frac{1}{x^3} + \frac{1}{5}\frac{1}{x^5} + \dots)$ ist die zweite Ungleichung erfüllt wenn

$$k(\varepsilon) > \tfrac{1}{2}\sqrt{\kappa}\ln(\tfrac{2}{\varepsilon})$$

□

3 Eigenwertprobleme

3.1 Einleitung

Aus der linearen Algebra ist bereits das klassische Eigenwertproblem bekannt:

Gegeben sei eine Matrix $A \in \mathbb{K}^{n\times n}$ und gesucht sind $\lambda \in \mathbb{K}$ und $v \in \mathbb{K}^n$, $v \neq 0$ sodass

$$Av = \lambda v$$

Das Umstellen des Eigenwertproblems ergibt das System

$$(A - \lambda I)v = 0 \tag{*}$$

Hierbei muss $A - \lambda I$ singulär sein, sonst ist die eindeutige Lösung des Systems gegeben durch $v = 0$.

Per Hand würden wir hier nun das charakteristische Polynom $\chi_A(\lambda) = \det(A - \lambda I)$ aufstellen und dessen Nullstellen bestimmen, da dies genau die Werte für λ sind, für welche das System (*) nichttriviale Lösungen hat.

Für die numerische Berechnung der Eigenwerte ist dies nicht ratsam, da Nullstellenbestimmung bei Polynomen hochgradig schlecht konditioniert ist.

Wir stellen folgende Zusammenhänge der Berechnung von Eigenwerten und Eigenvektoren fest:

a) Eigenwert bekannt $\implies$ Eigenvektor als Lösung eines LGS (*).

b) Eigenvektor bekannt $\implies$ Eigenwert über Rayleigh-Quotient $\lambda = \dfrac{\langle Av, v\rangle}{\|v\|_2^2}$

3.2 Einschließungssätze und Stabilität

Hilfssatz 3.1. Seien $A, B \in \mathbb{K}^{n\times n}$ beliebige Matrizen und $\|\cdot\|_{\text{op}}$ eine durch $\|\cdot\|$ induzierte Operatornorm. Dann gilt für jeden Eigenwert λ von A, welcher nicht zugleich auch Eigenwert von B ist, die Beziehung

$$\left\|(\lambda I - B)^{-1}(A - B)\right\|_{\text{op}} \geq 1$$

Beweis.
Ist w ein Eigenvektor vom Eigenwert λ von A, so folgt aus der Identität $(A-B)w = (I\lambda - B)w$, wenn λ kein Eigenwert von B ist, d.h. $\lambda I - B$ invertierbar, dass:

$$(\lambda I - B)^{-1}(A - B)w = w$$

Demnach ist also

$$1 \leq \sup_{x\in\mathbb{K}^n\setminus\{0\}} \frac{\left\|(\lambda I - B)^{-1}(A - B)x\right\|}{\|x\|} = \left\|(\lambda I - B)^{-1}(A - B)\right\|_{\text{op}}$$

3.2.1 Gerschgorin-Kreise

Satz 3.2 (Satz von Gerschgorin). Alle Eigenwerte eine Matrix $A \in \mathbb{K}^{n\times n}$ liegen in der Vereinigung der sogenannten Gerschgorin-Kreise

$$K_j := \left\{ z \in \mathbb{C} : |z - a_{jj}| \leq \sum_{k\neq j} |a_{jk}| \right\}, \qquad \text{für } j = 1, \ldots, n$$

Für eine Teilmenge $I \subset \{1, \ldots, n\}$ gilt, sind die Mengen $U = \bigcup_{j\in I} K_j$ und $V = \bigcup_{j\notin I} K_j$ disjunkt, so liegen in U genau $m := |I|$ und in V genau $n - m$ Eigenwerte von A (mehrfache Eigenwerte werden entsprechend ihrer algebraischen Vielfachheit gezählt).

Beweis.
Zur ersten Behauptung:
Wir setzen $B = D := \operatorname{diag}(a_{jj})$ in dem Hilfssatz 3.1 und nehmen $\|\cdot\|_\infty$ als natürliche Matrixnorm.

Für $\lambda \neq a_{jj}$ folgt dann

$$\left\|(\lambda I - D)^{-1}(A - D)\right\|_\infty = \max_{j=1,\ldots,n} \frac{1}{|\lambda - a_{jj}|} \sum_{k\neq j} |a_{jk}| \geq 1,$$

d.h. λ liegt in einem der Gerschgorin-Kreise.

Für den zweiten Teil setzen wir $A_t = D + t(A - D)$, dann hat $A_0 = D$ als Eigenwerte $a_{11}, \ldots, a_{nn}$ die Mittelpunkte von $K_1, \ldots, K_n$ und damit liegen genau m dieser Eigenwerte in U und die restlichen $n - m$ in V.

Die Abbildung $t \mapsto D + t(A - D)$ ist stetig in t und da Eigenwerte einer Matrix selbst stetig sind(d.h. ändert sich die Matrix leicht ab, dann ändern sich auch die Eigenwerte nur leicht ab), folgt, dass beim Wandern von $t = 0$ zu $t = 1$ und damit von $A_0 = D$ zu $A_1 = A$ sich die Eigenwerte stetig durch den Raum $\mathbb{K}$ bewegen. Da die Mengen U und V abgeschlossen und disjunkt sind haben sie einen positiven Abstand, welcher durch die stetige Bewegung der Eigenwerte nicht überschritten werden kann. □

Ein Alternativer Beweis zur ersten Behauptung liefert eine Betrachtung des Eigenwertproblems $Ax = \lambda x$ mit Eigenvektor $x \neq 0$. Offensichtlich existiert ein x_i mit $|x_j| \leq |x_i|$ für alle $j \neq i$.

Die i-te Komponente von Ax ist dann gegeben durch

$$\lambda x_i = (Ax)_i = \sum_{j=1}^{n} a_{jj}x_j = a_{ii}x_i + \sum_{j\neq i} a_{jj}x_j$$

Somit folgt

$$|\lambda - a_{ii}| = \left|\frac{1}{x_i}\sum_{j\neq i} a_{ij}x_j\right| = \left|\sum_{j\neq i} a_{ij}\frac{x_j}{x_i}\right| \leq \sum_{j\neq i} |a_{ij}|$$

Demnach liegt $\lambda \in K_i$. □

Korollar 3.3. Aus Satz 3.2 folgt insbesondere für $m = 1$, dass wenn ein Gerschgorin-Kreis keinen anderen schneidet, dass dieser genau einen Eigenwert enthält. Und wenn alle Gerschgorin-Kreise disjunkt sind, dann ist die Matrix diagonalisierbar.

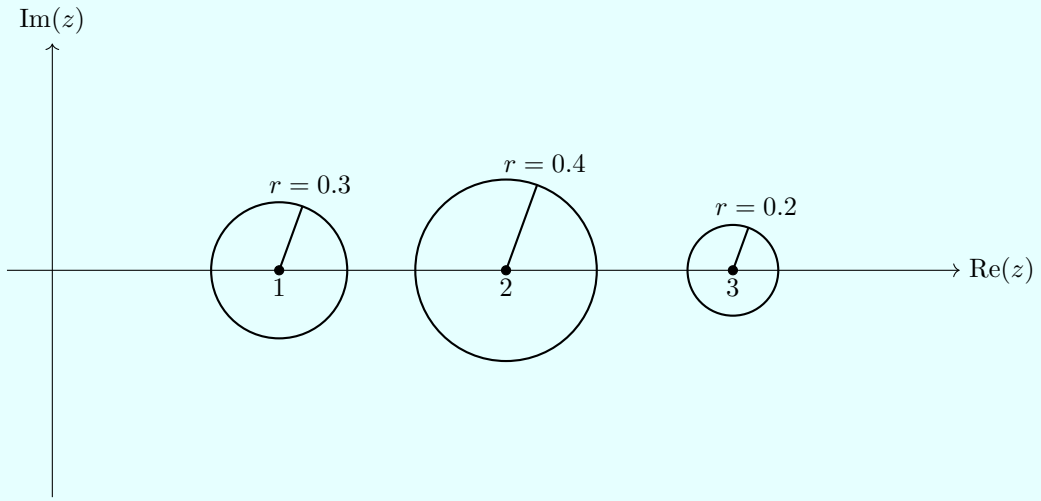
Weiter ist eine Matrix genau dann invertierbar, wenn keiner der Gerschgorin-Kreise die 0 enthält.

Beispiel 3.4. Gegeben sei die Matrix

$$A = \begin{pmatrix} 1 & 0.1 & -0.2 \\ 0 & 2 & 0.4 \\ -0.2 & 0 & 3 \end{pmatrix}$$

Es ergeben sich die folgenden Gerschgorin-Kreise:

$$K_1 = \{z \in \mathbb{C} : |z - 1| \leq 0.3\}$$
$$K_2 = \{z \in \mathbb{C} : |z - 2| \leq 0.4\}$$
$$K_3 = \{z \in \mathbb{C} : |z - 3| \leq 0.2\}$$



3.2.2 Stabilität von Eigenwerten

Satz 3.5 (Stabilitätssatz). Sei $A \in \mathbb{K}^{n\times n}$ eine diagonalisierbare Matrix, d.h. es gibt eine invertierbare Matrix $W = (W^{(1)}|\dots|W^{(2)})$ aus Eigenvektoren mit $A = W\Lambda W^{-1}$ und $\Lambda = \operatorname{diag}(\lambda_i(A))$.

Für eine zweite Matrix $B \in \mathbb{K}^{n\times n}$ gibt es zu jedem Eigenwert $\lambda(B)$ von B einen Eigenwert $\lambda(A)$ von A, sodass

$$|\lambda(A) - \lambda(B)| \leq \operatorname{cond}_2(W) \cdot \|A - B\|_2$$

Beweis.
Sei $\lambda = \lambda(B)$ kein Eigenwert von A, so gilt

$$\begin{aligned} \left\|(\lambda I - A)^{-1}\right\|_2 &= \left\|(\lambda I - W\Lambda W^{-1})^{-1}\right\|_2 \\ &= \left\|W(\lambda I - \Lambda)^{-1}W^{-1}\right\|_2 \\ &\leq \|W\|_2 \cdot \left\|W^{-1}\right\|_2 \cdot \left\|(\lambda I - \Lambda)^{-1}\right\|_2 \\ &= \operatorname{cond}_2(W) \cdot \max_{i=1,\dots,n} |\lambda - \lambda_i(A)|^{-1} \end{aligned}$$

Mit dem Hilfssatz 3.1 folgt dann die Behauptung. □

Für hermitische Matrizen $A \in \mathbb{K}^{n\times n}$ existiert bekannterweise eine Orthonormalbasis des $\mathbb{K}^{n\times n}$ aus Eigenvektoren, sodass die Matrix W als unitär angenommen werden kann, d.h. $W^{-1} = W^H$. In diesem

Fall gilt $\|W\| = \sqrt{\lambda_{\max}(W^H W)} = \sqrt{\lambda_{\max}(I)} = 1$ und analog $\left\|W^H\right\| = 1$, also

$$\mathrm{cond}_2(W) = \left\|W^{-1}\right\|_2 \cdot \|W\|_2 = 1$$

Regel: Allgemein kann man sagen, dass das Eigenwertproblem für hermitische Matrizen gut konditioniert ist, während das allgemeine Eigenwertproblem je nach Größe von $\mathrm{cond}_2(W)$ beliebig schlecht konditioniert sein kann.

3.3 Iterative Verfahren

Im folgenden wollen wir ein iteratives Verfahren zu Lösung des partiellen Eigenwertproblems einer Matrix $A \in \mathbb{K}^{n\times n}$ betrachten.

3.3.1 Potenz-Methode

Definition 3.6. Die Potenzmethode (Von-Mises-Iteration) erzeugt ausgehend von einem Startvektor $z^{(0)} \in \mathbb{C}^n$ mit $\left\|z^0\right\| = 1$ eine Folge von Iterationen $z^{(t)} \in \mathbb{C}^n, t = 1, 2, \ldots$ durch

$$\tilde{z}^{(t)} = Az^{(t-1)} \quad \text{und} \quad z^{(t)} = \frac{\tilde{z}^{(t)}}{\left\|\tilde{z}^{(t)}\right\|}$$

Für einen beliebigen Index $k \in \{1, \ldots, n\}$, (z.B. maximale Komponente von $z^{(k)}$) setzen wir:

$$\lambda^{(t)} = \frac{(Az^{(t)})_k}{(z^{(t)})_k}$$

Zur Normierung wird üblicherweise $\|\cdot\| = \|\cdot\|_2$ oder $\|\cdot\|_\infty$ verwendet.

Zur Analyse des Verfahrens nehmen wir an, dass die Matrix A diagonalisierbar ist. Dies ist äquivalent zu der Tatsache, dass A eine Basis von normierten Eigenvektoren $\{w^{(1)}, \ldots, w^{(n)}\}$ besitzt.

zusätzlich nehmen wir an, dass $z^{(0)}$ eine nicht-triviale Komponente bezüglich $w^{(n)}$ besitzt. *(Dies ist keine wesentliche Einschränkung, da aufgrund des unvermeidbaren Rundungsfehlers dieser Fall der Iteration sicher einmal auftritt)*

Satz 3.7 (Potenz-Methode). Die Matrix A sei diagonalisierbar und ihr betragsgrößter Eigenwert sei separiert von den anderen Eigenwerten, also oBdA $|\lambda_n| > |\lambda_{n-1}| \geq |\lambda_{n-2}| \geq \cdots \geq |\lambda_1|$.

Der Startvektor $z^{(0)}$ habe eine nicht-triviale Komponente bezüglich des zugehörigen Eigenvektors $w^{(n)}$, dann gibt es Zahlen $\delta_t \in \mathbb{C}, |\delta_t| = 1$, sodass

$$\left\|z^{(t)} - \delta_t \cdot w^{(n)}\right\| \to 0 \quad \text{und} \quad \lambda^{(t)} - \lambda_n = \mathcal{O}\left(\left|\frac{\lambda_{n-1}}{\lambda_n}\right|^t\right) \qquad \text{für } t \to \infty$$

Beweis.
Sei $z^{(0)} = \sum_{i=1}^{n} \alpha_i \cdot w^{(i)}$ die Basisdarstellung des Startvektors (mit $\alpha_n \neq 0$). Für die Iterierten gilt:

$$z^{(t)} = \frac{\tilde{z}^{(t)}}{\left\|\tilde{z}^{(t)}\right\|} = \frac{Az^{(t-1)}}{\left\|Az^{(t-1)}\right\|} = \cdots = \frac{A^t z^{(0)}}{\left\|A^t z^{(0)}\right\|} \tag{1}$$

Dabei gilt:

$$A^t z^{(0)} = \sum_{i=1}^{n} \alpha_i \lambda_i^t w^{(i)} = \lambda_n^t \alpha_n \cdot \left(w^{(n)} + \sum_{i=1}^{n-1} \frac{\alpha_i}{\alpha_n} \left(\frac{\lambda_i}{\lambda_n}\right)^t w^{(i)}\right) \tag{2}$$

Wegen $|\frac{\lambda_i}{\lambda_n}| \leq \rho := |\frac{\lambda_{n-1}}{\lambda_n}| < 1$ für $i = 1, \ldots, n-1$ folgt

$$A^t z^{(0)} = \lambda_n^t \alpha_n (w^{(n)} + \mathcal{O}(\rho^t)) \qquad \text{für } t \to \infty$$

Dies ergibt nach (1):

$$z^{(t)} = \frac{A^t z^{(0)}}{\left\|A^t z^{(0)}\right\|} = \frac{\lambda_n^t \alpha_n (w^{(n)} + \mathcal{O}(\rho^t))}{|\lambda_n^t \alpha_n| \cdot \left\|w^{(n)} + \mathcal{O}(\rho)\right\|} = \underbrace{\frac{\lambda_n^t \alpha_n}{|\lambda_n^t \alpha_n|}}_{=:\delta_k} \cdot w^{(n)} + \mathcal{O}(\rho)$$

Dabei ist $\delta_t \in \mathbb{C}$ und $|\delta_t| = 1$, daher folgt die erste Aussage.

Weiter gilt

$$\begin{aligned}
\lambda^{(t)} &= \frac{(Az^{(t)})_k}{(z^{(t)})_k} \\
&\overset{(1)}{=} \frac{(A^{t+1} z^{(0)})_k}{\left\|(A^{t+1} z^{(0)})_k\right\|} \cdot \frac{\left\|(A^{t+1} z^{(0)})_k\right\|}{(A^t z^{(0)})_k} \\
&\overset{(2)}{=} \frac{\lambda_n^{t+1}\left(\alpha_n w_{n,k} + \sum_{i=1}^{n-1} \alpha_i (\frac{\lambda_i}{\lambda_n})^{t+1} w_{i,k}\right)}{\lambda_n^t\left(\alpha_n w_{n,k} + \sum_{i=1}^{n-1} \alpha_i (\frac{\lambda_i}{\lambda_n})^{t} w_{i,k}\right)} \\
&= \lambda_n + \mathcal{O}\left(\left|\frac{\lambda_{n-1}}{\lambda_n}\right|^t\right) \qquad \text{für } t \to \infty
\end{aligned}$$

□

Die Konvergenz der Potenzmethode ist umso besser, je mehr der betragsgrößte Eigenwert λ_n von den übrigen separiert ist. Der Beweis ist verallgemeinerbar für betragsgrößte Eigenwerte, welche mehrfach auftreten, sofern die Matrix diagonalisierbar ist.

3.3.2 Inverse Iteration

Als nächstes wollen wir uns die „Inverse Iteration“ nach Wielandt anschauen.

Wir nehmen an, wir haben bereits eine Näherung $\tilde{\lambda}$ für einen Eigenwert λ_k der regulären Matrix A (z.B. durch Einschließungssätze). Die Näherung sei gut in dem Sinne, dass $|\lambda_k - \tilde{\lambda}| \ll |\lambda_i - \tilde{\lambda}|$ für $i \neq k$.

Wir betrachten das Eigenwertproblem, welches sich für die Matrix $A - \tilde{\lambda} I$ ergibt:

$$(A - \tilde{\lambda} I)v = \xi v \iff (A - \tilde{\lambda} I - \xi I)v = 0 \iff (A - (\tilde{\lambda} + \xi) I)v = 0$$

Wegen $(A - \lambda_k I) = 0$, ist $\xi = \lambda_k - \tilde{\lambda}$ ein Eigenwert von $A - \tilde{\lambda} I$ und damit $\mu = \frac{1}{\xi} = (\lambda_k - \tilde{\lambda})^{-1}$ ein Eigenwert von $(A - \tilde{\lambda} I)^{-1}$. [2]

Allgemeiner hat im Falle $\tilde{\lambda} \neq \lambda_k$ die Matrix $(A - \tilde{\lambda} I)^{-1}$ die Eigenwerte $\mu_i = (\lambda_i - \tilde{\lambda})^{-1}$ für $i = 1, \ldots, n$ und es gilt

$$\left|\frac{1}{\lambda_k - \tilde{\lambda}}\right| \gg \left|\frac{1}{\lambda_i - \tilde{\lambda}}\right| \qquad \text{für } i \neq k$$

[2] $Av = \lambda v \implies v = A^{-1}\lambda v \implies \lambda^{-1} v = A^{-1} v$

Definition 3.8. Die inverse Iteration beruht auf der Anwendung der Potenzmethode auf die Matrix $(A - \tilde{\lambda}I)^{-1}$ mit einer a priori Schätzung $\tilde{\lambda}$ zum gesuchten Eigenwert λ_k.

Ausgehend von einem Startwert $z^{(0)}$ werden Iterierte $z^{(t)}$ als Lösung der Gleichungssysteme

$$(A - \tilde{\lambda}I)\tilde{z}^{(t)} = \tilde{z}^{(t-1)}, \qquad z^{(t)} = \frac{\tilde{z}^{(t)}}{\left\|\tilde{z}^{(t)}\right\|}$$

bestimmt.

Die zugehörige Eigenwertnäherung wird bestimmt durch

$$\mu^{(t)} = \frac{(z^{(t)})_k}{((A - \tilde{\lambda}I)z^{(t)})_k}$$

mit Nenner $\neq 0$ (oder im symmetrischen Fall einfach mit Hilfe der Rayleigh-Quotienten).

Aufgrund der Aussagen über Potenzmethoden liefert die inverse Iteration also für eine diagonalisierbare Matrix jeden Eigenwert, zu dem bereits eine hinreichend gute Näherung bekannt ist.

3.4 Page-Rank-Algorithmus

Das Ziel des Page-Rank-Algorithmus ist die Bestimmung der Ausgabereihenfolge bei Suchergebnissen. Dabei berufen wir uns auf folgende Regeln:

(1) Eine Website erhält eine umso höhere Bewertung, je mehr Links auf sie zeigen.

(2) Links von höher bewerteten Websites soll relevanter sein, als solche von unbedeutenden Websites

(3) Ein Link von einer Website, die wenig Links nach außen hat, soll höher gewichtet werden als der von einer Website mit vielen Links nach außen.

Wir beschreiben unser Model als ein Netz mit n Websites, wobei ein Index k immer für eine Website steht und suchen die Bewertung $x_k \in \mathbb{R}$ der Website k.

Weiter sei L_k die Menge der Websites, welche auf k verlinken, Links auf Websites von sich selbst werden dabei nicht berücksichtigt und n_k sei die Anzahl Websites, auf welche k verlinkt.

Wir modellieren mittels folgendem LGS

$$x_k = \sum_{j \in L_k} \tfrac{1}{n_j} \cdot x_j$$

Die Gleichung $x = Ax$ mit

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \qquad A_{ij} = a_{ij} = \begin{cases} \frac{1}{n_j}, & \text{falls die Seite } j \text{ auf die Seite } i \text{ verlinkt} \\ 0, & \text{sonst} \end{cases}$$

entspricht hierbei der Eigenwertgleichung für den Eigenwert $\lambda = 1$.

Der historische Ansatz von Google ist die Potenzmethode:

3.4.1 Stochastische Vektoren/Matrizen

Definition 3.9. Ein Vektor $p \in \mathbb{R}^n$ heißt stochastischer Vektor, wenn alle Elemente p_i nichtnegativ sind und die Summe der Elemente des Vektors gleich 1 ist, d.h. $\sum_i p_i = 1$.

Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt stochastische Matrix, wenn alle Spalten der Matrix stochastische Vektoren sind, d.h.

$$a_{ij} \geq 0 \ \forall i,j \quad \text{und} \quad \sum_{i=1}^{n} a_{ij} = 1 \ \forall j$$

Lemma 3.10. Sei $A \in \mathbb{R}^{n\times n}$ eine stochastische Matrix und $p \in \mathbb{R}^n$ ein stochastischer Vektor, dann ist das Produkt $Ap \in \mathbb{R}^{n\times n}$ wieder ein stochastischer Vektor.

Beweis.
Offensichtlich $(Ap)_{ij} \geq 0$, weiter es sei a_i die i-te Spalte der Matrix A, d.h. a_i ist ein stochastischer Vektor

$$\begin{aligned} A \cdot p = A \cdot \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix} &= p_1 \cdot a_1 + \cdots + p_n \cdot a_n \\ &= \sum_{i=1}^{n} (p_1 a_{i1} + \cdots + p_n a_{in}) \\ &= p_1 \sum_{i=1}^{n} a_{i1} + \cdots + p_n \sum_{i=1}^{n} a_{in} \\ &= p_1 \cdot 1 + \cdots + p_n \cdot 1 = 1 \end{aligned}$$

□

Korollar 3.11. Seien $A, B \in \mathbb{R}^{n\times n}$ stoch. Matrizen, dann ist das Produkt $A \cdot B$ wieder eine stoch. Matrix.

Beweis. Folgt direkt aus Lemma 3.9.

Satz 3.12. Eine stochastische Matrix A hat immer den Eigenwert 1. Der Betrag aller anderen Eigenwerte ist kleiner oder gleich 1.

Beweis.
Für den ersten Teil nutzen wir aus, dass A und A^T die gleichen Eigenwerte, da A und A^T die gleiche Determinante besitzen und damit die charakteristischen Polynome $\chi_A(\lambda) = \det(A - \lambda I) = \det(A^T - \lambda I) = \chi_{A^T}(\lambda)$ übereinstimmen.

Weiter ist nach Definition einer stochastischen Matrix die Summe jedes Zeilenvektors von A^T ist gleich 1, also ist $e = (1, \ldots, 1)^T$ ein Eigenvektor von A^T mit Eigenwert 1, also besitzt auch die Matrix A den Eigenwert $\lambda = 1$.

Für den zweiten Teil nehmen wir an es existiert ein Eigenvektor v zum Eigenwert λ mit $|\lambda| > 1$, denn dann gilt

$$A^n v = A^{n-1}(Av) = A^{n-1}\lambda v = \lambda A^{n-1} v = \cdots = \lambda^n v$$

Für die Länge dieses Vektors gilt $\|\lambda^n v\| = |\lambda^n| \cdot \|v\|$ ein exponentielles Wachstum in n, da $|\lambda| > 1$.

Daraus folgt, dass für große n ein Element $(A^n)_{ij}$ existiert, welches größer als 1 ist.

Da nach Korollar 3.11 die Matrix A^n stochastisch ist bildet dies einen Widerspruch. □

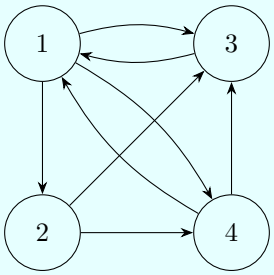
Lemma 3.13. Die Bewertungsmatrix A des Page-Rank-Algorithmus ist eine stochastische Matrix.

Beweis. Offensichtlich gilt $a_{ij} \geq 0$, weiter gilt

$$\sum_{i=1}^{n} a_{ij} = n_j \cdot \frac{1}{n_j} + (n - n_j) \cdot 0 = 1$$

□

Beispiel 3.14. Wir betrachten folgendes einfaches Netz mit 4 Knoten:



Es ergibt sich folgendes Gleichungssystem:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$$

Lösen dieses linearen Gleichungssystems liefert:

$$x \in \text{span} \left\{ \begin{pmatrix} 0.72 \\ 0.24 \\ 0.54 \\ 0.36 \end{pmatrix} \right\}$$

Demnach hat die erste Website die höchste Bewertung.

3.4.2 Vorgehensweite für weitere Eigenwerte/Eigenvektoren

Wir betrachten die Diagonalmatrix

$$A = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}$$

mit den Eigenwerten $\lambda_1 = 3$ und $\lambda_2 = 2$ zu den Eigenvektoren

$$w^{(1)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad w^{(2)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Wir führen folgende Transformation durch:

$$B = \underbrace{\begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}}_{=A} - \underbrace{\begin{pmatrix} 3 & 0 \\ 0 & 0 \end{pmatrix}}_{=(1,0)^T(3,0)} = \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}$$

Idee: Umwandlung der betrachteten Matrix in eine andere Matrix, wobei der betragsgrößte Eigenwert entfernt wird, d.h. durch 0 ersetzt wird. Eine Iterative Anwendung liefert dann alle Eigenwerte.

Herleitung: Sei $A \in \mathbb{R}^{n\times n}$ gegeben mit betragsmäßig absteigenden Eigenwerten, d.h.

$$|\lambda_1| > |\lambda_2| > \cdots > 0$$

Der Eigenvektor zu λ_1 sei gegeben durch $u^{(1)}$.
Wir wählen eine von Null verschiedene Komponente $u_p^{(1)}$ von $u^{(1)}$ und schreiben A_p für die p-te Zeile von A, d.h. $A_p = (A_{p1}, A_{p2}, \ldots, A_{pn})^T$. Betrachte nun die Matrix

$$B = A - \frac{1}{w_p^{(1)}} w^{(1)} \cdot A_p^T \quad \text{mit} \quad B_{ij} = A_{ij} - \frac{1}{w_p^{(1)}} w_i^{(1)} \underbrace{A_{pj}}_{A_{pj}}$$

Aus dem Eigenwertproblem $Aw^{(k)} = \lambda_k w^{(k)}$ ergibt sich

$$\lambda^{(k)} w_p^{(k)} = (Aw^{(k)})_p = \langle A_p, w^{(k)} \rangle$$

Für $k = 1$ ergibt sich:

$$\begin{aligned} Bw^{(1)} &= Aw^{(1)} - \frac{1}{w_p^{(1)}} \cdot w^{(1)} \cdot \langle A_p, w^{(1)} \rangle \\ &= Aw^{(1)} - \frac{1}{w_p^{(1)}} \cdot w^{(1)} \cdot \lambda_1 w_p^{(1)} \\ &= \lambda_1 w^{(1)} - \lambda_1 w^{(1)} \\ &= 0 \end{aligned}$$

d.h. 0 ist ein Eigenwert von B (statt vorher $\lambda^{(1)}$ von A).

Analoge Überlegung für $k = 2, \ldots, n$ liefert:

$$\begin{aligned} Bw^{(k)} &= \lambda_k w^{(k)} - \frac{1}{w_p^{(1)}} \cdot w^{(1)} \cdot \lambda_k w_p^{(k)} \\ &= \lambda_k \cdot \left(w^{(k)} - \frac{w_p^{(k)}}{w_p^{(1)}} \cdot w^{(1)} \right) \end{aligned} \tag{1}$$

Die Eigenwerte bleiben beim Wechsel von A zu B erhalten, da

$$\begin{aligned} Bw^{(k)} + 0 &= Bw^{(k)} + \underbrace{Bw^{(1)}}_{=0} \\ &= Bw^{(k)} + Bw^{(1)} \cdot \frac{-w_p^k}{w_p^{(1)}} \\ &= B \cdot \left(w^{(k)} - \frac{w_p^{(k)}}{w_p^{(1)}} \cdot w^{(1)} \right) \end{aligned} \tag{2}$$

Und damit

$$B \cdot \left(w^{(k)} - \frac{w_p^{(k)}}{w_p^{(1)}} \cdot w^{(1)} \right) \overset{(2)}{=} Bw^{(k)} \overset{(1)}{=} \lambda_k \cdot \left(w^{(k)} - \frac{w_p^{(k)}}{w_p^{(1)}} \cdot w^{(1)} \right)$$

Dies zeigt für $k = 2, 3, \ldots, n$, dass λ_k auch ein Eigenwert zu B ist, wenn auch mit anderem Eigenvektor.

Satz 3.15 (Deflation nach Wielandt). Seien die Eigenwerte von A, betragsmäßig absteigend, d.h. $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$, mit zugehörigen Eigenvektoren $w^{(1)}, \ldots, w^{(n)}$. Dann besitzt die Matrix

$$B = A - \frac{1}{w_p^{(1)}} w^{(1)} \cdot A_p^T \quad \text{mit} \quad u_p^{(1)} \neq 0 \text{ und } A_p = (A_{p1}, A_{p2}, \ldots, A_{pn})^T$$

die Eigenwerte $0, \lambda_2, \ldots, \lambda_n$ mit den zugehörigen Eigenvektoren $w^{(1)}, \tilde{w}^{(2)}, \ldots, \tilde{w}^{(n)}$, wobei

$$\tilde{w}^{(k)} = w^{(k)} - \frac{w_p^{(k)}}{w_p^{(1)}} \cdot w^{(1)} \qquad (*)$$

Den zum Eigenwert λ_2 zugehörigen Eigenvektor $\tilde{w}^{(2)}$ erhält man somit mit der Potenzmethode für die Matrix B nach ihrer Definition.

Der Eigenvektor $w^{(2)}$ zum Eigenwert λ_2 der Matrix A kann dann wie folgt rekonstruiert werden:

a) Lösen des linearen Gleichungssystems (*) bezüglich $w^{(2)}$

b) Lösen des LGS der EW-Gleichung

c) Inverse Iteration nach Wielandt anwenden, um Eigenvektor von A zum zugehörigen Eigenwert λ_2 zu erhalten

Beispiel 3.16. Gesucht seien die Eigenwerte und Eigenvektoren der Matrix

$$A = \begin{pmatrix} -4 & 14 & 0 \\ -5 & 13 & 0 \\ -1 & 0 & 2 \end{pmatrix}$$

Im ersten Schritt verwenden wir die Potenzmethode um den betragsmäßig größten Eigenwert und den zugehörigen Eigenvektor zu bestimmen, wir erhalten:

$$\lambda_1 = 6 \quad , w^{(1)} = \begin{pmatrix} -4 \\ -20/7 \\ 1 \end{pmatrix}$$

Für die Deflation wählen wir nun $p = 1$ mit $w_p^{(1)} \neq 0$ und $A_p = (-4, 14, 0)^T$, die resultierende Matrix B ergibt sich dann durch:

$$B = \begin{pmatrix} -4 & 14 & 0 \\ -5 & 13 & 0 \\ -1 & 0 & 2 \end{pmatrix} - \frac{1}{-4} \cdot \begin{pmatrix} -4 \\ -\frac{20}{7} \\ 1 \end{pmatrix} \cdot (-4, 14, 0) = \begin{pmatrix} 0 & 0 & 0 \\ -15/7 & 3 & 0 \\ -2 & 7/2 & 2 \end{pmatrix}$$

Erneutes Anwenden der Potenzmethode auf die neue Matrix B liefert den zweitgrößten Eigenwert und den zugehörigen Eigenvektor von B:

$$\lambda_2 = 3 \quad , \tilde{w}^{(2)} \begin{pmatrix} 0 \\ 2/7 \\ 1 \end{pmatrix}$$

Eine weitere Deflation mit dem neu gewonnen Eigenvektor und $p = 3$ ergibt

$$C = \begin{pmatrix} 0 & 0 & 0 \\ -15/7 & 3 & 0 \\ -2 & 7/2 & 2 \end{pmatrix} - \frac{1}{1} \cdot \begin{pmatrix} 0 \\ 2/7 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} -2 & 7/2 & 2 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ -11/7 & 2 & -4/7 \\ 0 & 0 & 0 \end{pmatrix}$$

Hierbei ergibt sich der letzte Eigenwert und der zu C zugehörige Eigenvektor

$$\lambda_3 = 2 \quad , \widetilde{\widetilde{w}}^{(3)} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

Nach beliebiger Methodik aus Satz 3.15 lassen sich nun aus $\widetilde{w}^{(2)}$ und $\widetilde{\widetilde{w}}^{(3)}$ die Eigenvektoren $w^{(2)}$ und $w^{(3)}$ von A konstruieren.

4 Krylov-Raum-Methoden für EW-Probleme

Wir verfolgen die gleiche Idee, wie auch schon bei linearen Gleichungssystemen, d.h. die ursprünglich hochdimensionalen Probleme, werden durch geeignete Unterräume (Krylov-Räume) in kleinere Probleme umgewandelt.
Wir erhalten dabei ein iteratives Vorgehen, zu betrachtende Beispiele sind die Arnoldi-Methode und die Lanczos-Methode.

Wir betrachten also die Eigenwertgleichung $Az = \lambda z$ mit $A \in \mathbb{C}^{n\times n}$ (ab jetzt erlauben wir auch komplexe Matrizen), wobei A eine sehr große Matrix ist, typischerweise $n \geq 10^4$.

4.1 Galerkin-Approximation

Eigenwertprobleme können äquivalent in Variationsform (schwache Formulierung) geschrieben werden, diese besagt:
$z \in \mathbb{C}^n$ ist genau dann ein Eigenvektor von A zum Eigenwert $\lambda \in \mathbb{C}$, wenn

$$\langle Az, y\rangle_2 = \lambda\langle z, y\rangle_2 \quad \forall\, y \in \mathbb{C}^n \tag{1}$$

Diese Äquivalenz gilt, da aus $\langle r, y\rangle_2 = 0$ für alle $y \in \mathbb{C}^n$ folgt, dass $r = 0$ sein muss, in unserem Fall ist $r = Az - \lambda z$ das Residuum des Eigenwertproblems.

Sei $K_m = \operatorname{span}\{q^{(1)}, \ldots, q^{(m)}\}$ ein geeignet gewählter Unterraum von $\mathbb{C}^n$ kleiner Dimension, d.h. $\dim K_m = m \ll n$, dann wird das n-dimensionale Eigenwertproblem (1) mit folgendem m-dimensionalem Eigenwertproblem approximiert:

$$\text{suche } z \in K_m,\ \lambda \in \mathbb{C}: \quad \text{mit} \quad \langle Az, y\rangle_2 = \lambda\langle z, y\rangle_2 \quad \forall\, y \in K_m$$

Aufgrund der Bilinearität des Skalarproduktes reicht es auch, wenn wir nur die erzeugenden $q^{(i)}$ betrachten, statt alle $y \in K_m$.

Wir entwickeln die Eigenvektoren $z \in K_m$ bzgl. der gegebenen Basis:

$$z = \sum_{j=1}^{m} \alpha_j q^{(j)}$$

und erhalten somit das Galerkin-System

$$\sum_{j=1}^{k} \alpha_j \langle Aq^{(j)}, q^{(i)}\rangle_2 = \lambda \cdot \sum_{j=1}^{k} \alpha_j \langle q^{(j)}, q^{(i)}\rangle_2 \qquad \forall\, i = 1, \ldots, m$$

Dabei charakterisieren die α_i unser gesuchtes z, wir schreiben dieses System daher typischerweise in kompakter Form als Eigenwertproblem

$$\mathcal{A}\alpha = \lambda\mathcal{M}\alpha$$

mit Vektoren $\alpha = (\alpha_1, \ldots, \alpha_m)$ und Matrizen $\mathcal{A} = (\langle Aq^{(j)}, q^{(i)}\rangle_2)_{i,j=1}^m$, $\mathcal{M} = (\langle q^{(j)}, q^{(i)}\rangle_2)_{i,j=1}^m$.

Im Folgenden betrachten wir immer die *kartesische Repräsentation* der Basisvektoren $q^{(i)} = (q_j^{(i)})_{j=1}^n$ und somit schreibt man das Galerkin-EW-Problem in der Form[3]

$$\sum_{j=1}^{m} \alpha_j \cdot \sum_{k,l=1}^{n} a_{k,l} \cdot q_k^{(j)} \cdot \overline{q}_l^{(i)} = \lambda \cdot \sum_{j=1}^{m} \alpha_j \cdot \sum_{k,l=1}^{n} q_k^{(j)} \cdot \overline{q}_l^{(i)} \quad \forall\, i = 1, \ldots, m$$

[3] Als Erinnerung: Im Komplexen ist das Standardskalarprodukt definiert durch $\langle x, y\rangle_2 = \sum_i x_i \cdot \overline{y}_i$

Mit $\mathcal{Q}^{(m)} = [q^{(1)}, \dots, q^{(m)}] \in \mathbb{C}^{n\times m}$ kann dies in kompakter Form

$$\mathcal{Q}^{(m)H} A \mathcal{Q}^{(m)} \alpha = \lambda \mathcal{Q}^{(m)H} \mathcal{Q}^{(m)} \alpha$$

formuliert werden.

Wenn $\{q^{(1)}, \dots, q^{(m)}\}$ eine Orthonormalbasis von K_m ist, reduziert sich dies zum normalen EW-Problem:

$$\underbrace{\mathcal{Q}^{(m)H} A \mathcal{Q}^{(m)}}_{=:H^{(m)} \in \mathbb{C}^{m\times m}} \alpha = \lambda \alpha \tag{2}$$

Falls $H^{(m)}$ eine spezielle Struktur hat (z. B. Hessenberg-Matrix oder symmetrische Tridiagonalgestalt), dann kann das EW-Problem mit niedriger Dimension (2) mit z.B. QR-Methode gelöst werden.

Seine Eigenwerte, genannt *Ritz-Eigenwerte*, können als Approximationen der dominanten Eigenwerte der ursprünglichen Matrix A betrachtet werden.

Bemerkung 4.1 (Krylov-Methode).

1. Wähle geeignete Unterräume $K_m \in \mathbb{C}^{m\times m}$, $m \ll n$ (Krylov-Räume) durch Verwendung der Matrix A und deren Potenz.
2. Konstruiere eine Orthonormalbasis $\{q^{(1)}, \dots, q^{(m)}\}$ von K_m mit der stabilisierten Version des Gram-Schmidt-Algorithmus und setze $\mathcal{Q}^{(m)} := [q^{(1)}, \dots, q^{(m)}]$.
3. Berechne die Matrix $H^{(m)} := \mathcal{Q}^{(m)H} A \mathcal{Q}^{(m)}$, welche konstruktionsbedingt eine Hessenberg-Matrix oder im hermiteschen Fall eine hermitesche Tridiagonalmatrix ist.
4. Löse das Eigenwertproblem der reduzierten Matrix $H^{(m)} \in \mathbb{C}^{m\times m}$ durch die QR-Methode.
5. Die Eigenwerte von $H^{(m)}$ als Näherung der dominanten (betragsgrößten) Eigenwerte von A. Im Falle des betragskleinsten Eigenwert, muss die Matrix A^{-1} betrachtet werden (Konstruktion der Unterräume K_m kann sehr aufwendig sein).

4.2 Arnoldi-Methode

Idee:
Die Potenzmethode verwendet nur die aktuelle Iterierte $A^m q$ mit $m \ll n$ für den normierten Startvektor $q \in \mathbb{C}^n$ mit $\|q\|_2 = 1$, ignoriert aber die bereits berechneten Iterierten $\{q, Aq, A^2q, \dots, A^{m-1}q\}$.

Wir wollen diese bereits bestimmten Informationen nun nutzen und erstellen eine sogenannte *Krylov-Matrix*:

$$K_m = [q, Aq, A^2q, \dots, A^{m-1}q] \quad \text{mit } 1 \le m \le n$$

Die Spalten dieser Matrix sind jedoch nicht orthogonal zueinander, außerdem konvergiert $A^t q$ gegen den Eigenvektor zum betragsgrößten Eigenwert, d.h. K_m ist schlecht konditioniert[4], da sich die letzten Spalten kaum ändern.

Wie wir sehen werden, ist die Konstruktion in eine orthogonale Basis mit dem Gram-Schmidt-Algorithmus instabil, wir wählen daher als Alternative in der Arnoldi-Methode die Verwendung einer stabilisierten Variante des Gram-Schmidt-Verfahrens um eine Folge orthonormaler Vektoren $\{q^{(1)}, q^{(2)}, \dots\}$ (bezeichnet als Arnoldi-Vektoren) zu erzeugen, sodass für jedes m die Vektoren $\{q^{(1)}, \dots, q^{(m)}\}$ den Krylov-Unterraum K_m aufspannen.

[4] Für nicht-invertierbare Matrizen $A \in \mathbb{C}^{n\times m}$ ist die Konditionszahl über das Pseudoinverse definiert

Definition 4.2. Für das Folgende definieren wir den orthogonalen Projektionsoperator:

$$\mathrm{proj}_u(v) := \frac{\langle v, u\rangle_2}{\|u\|_2^2} \cdot u$$

Dieser projiziert den Vektor v auf $\mathrm{span}\{u\}$.

Mit diesem Operator ergibt sich das *klassische Gram-Schmidt-Orthogonalisierungs-Verfahren* als

$$\begin{aligned} q^{(1)} &= \frac{q}{\|q\|_2}, \\ \text{und für } t &= 2, \dots, m: \\ \tilde{q}^{(t)} &= A^{t-1}q - \sum_{j=1}^{t-1} \mathrm{proj}_{q^{(j)}}(A^{t-1}q), \\ q^{(t)} &= \frac{\tilde{q}^{(t)}}{\left\|\tilde{q}^{(t)}\right\|_2} \end{aligned}$$

Der t-te Schritt projiziert dabei die Komponente von $A^{t-1}q$ in Richtung der bereits bestimmten orthogonalen Vektoren $\{q^{(1)}, \dots, q^{(t-1)}\}$.

Wir betrachten jetzt das *modifizierte Gram-Schmidt-Verfahren* (Nochmal überarbeiten), wobei der t-te Schritt die Komponente von $Aq^{(t)}$ in Richtung $\{q^{(1)}, \dots, q^{(t-1)}\}$ projiziert:

$$\begin{aligned} q^{(1)} &= \frac{q}{\|q\|_2}, \\ \text{und für } t &= 2, \dots, m: \\ \tilde{q}^{(t)} &= Aq^{(t-1)} - \sum_{j=1}^{t-1} \mathrm{proj}_{q^{(j)}}(Aq^{(t-1)}), \\ q^{(t)} &= \frac{\tilde{q}^{(t)}}{\left\|\tilde{q}^{(t)}\right\|_2} \end{aligned}$$

Nach Konstruktion ist $q^{(t)}$ senkrecht zu $\{q^{(j)}\}_{j=1}^{t-1}$ und damit auch zu $\{\mathrm{proj}_{q^{(j)}}(Aq^{(t-1)})\}_{j=1}^{t-1}$, es folgt also

$$\langle q^{(t)}, \tilde{q}^{(j)}\rangle_2 = \left\langle q^{(t)}, Aq^{(j-1)} - \sum_{i=1}^{j-1} \mathrm{proj}_{q^{(i)}}(Aq^{(j-1)}) \right\rangle_2 = \langle q^{(t)}, Aq^{(j-1)}\rangle_2$$

Durch die Setzung $h_{j,t-1} := \langle q^{(j)}, Aq^{(t-1)}\rangle_2$ und $h_{t,t-1} := \|\tilde{q}^t\|$ ergibt sich mit dem modifizierte Gram-Schmidt-Algorithmus dann

$$Aq^{(t-1)} = \sum_{j=1}^{t} h_{j,t-1} q^{(j)}, \qquad t = 2, \dots, m+1 \tag{1}$$

In der Praxis wird der modifizierte Gram-Schmidt-Algorithmus in der folgenden iterierten Form implementiert:

$$\begin{aligned} q^{(1)} &= \|q\|_2^{-1} q, \\ q^{(t,1)} &= Aq^{(t-1)}, \\ q^{(t,j+1)} &= q^{(t,j)} - \mathrm{proj}_{q^{(j)}}(q^{(t,j)}), \\ q^{(t)} &= \left\|q^{(t,t)}\right\|_2^{-1} q^{(t,t)} \end{aligned} \tag{2}$$

Wir erhalten dabei das gleiche Resultat, wie beim klassischen Gram-Schmidt-Verfahren, aber mit kleinerem numerischen Fehler Um dies zu verdeutlichen hilft es das klassische und modifizierte Gram-Schmidt-Verfahren im Allgemeinen zu vergleichen (also nicht auf unsere spezielle Basis bezogen sondern für beliebige linear unabhängige Vektoren $\{v^{(t)}\}_{t=1}^{n}$):

Algorithmus 6 + 7: Gram-Schmidt (klassisch vs. modifiziert)

klassisch:

for $t = 1, \dots, n$
  $q^{(t)} = v^{(t)}$
  for $j = 1, \dots, t-1$
    $h_{j,t-1} \leftarrow \langle q^{(j)}, q^{(t)} \rangle$
  end
  for $j = 1, \dots, t-1$
    $q^{(t)} \leftarrow q^{(t)} - h_{j,t-1} \cdot q^{(j)}$
  end
  $h_{t,t-1} \leftarrow \left\| q^{(t)} \right\|$ $q^{(t)} \leftarrow q^{(t)} \,/\, h_{t,t-1}$
end

modifiziert:

for $t = 1, \dots, n$
  $q^{(t)} = v^{(t)}$
  for $j = 1, \dots, t-1$
    $h_{j,t-1} \leftarrow \langle q^{(j)}, q^{(t)} \rangle$
    $q^{(t)} \leftarrow q^{(t)} - h_{j,t-1} \cdot q^{(j)}$
  end
  $h_{t,t-1} \leftarrow \left\| q^{(t)} \right\|$ $q^{(t)} \leftarrow q^{(t)} \,/\, h_{t,t-1}$
end

Einfaches Nachrechnen zeigt, dass beide Algorithmen bei exakter Arithmetik das gleiche Ergebnis liefern (vgl. Aufgabe 9.2).

Weiter sorgt das verzögerte Berechnen der Koeffizienten $h_{j,t-1}$ dafür, dass sich in $q^{(t)}$ weniger Fehler fortpflanzen. Im j-ten Schritt wurden bereits die Koeffizienten der Basisvektoren $q^{(1)}, \dots, q^{(j-1)}$ von $q^{(t)}$ eliminiert, wohingegen bei der Verwendung von $v^{(t)}$ noch große derartige Koeffizienten vorkommen könnten, welche zu stärkeren Fehlern in $h_{j,t-1}$ sorgen.

Definition 4.3 (Arnoldi-Algorithmus). Für eine beliebige Matrix $A \in \mathbb{C}^{n \times n}$ bestimmt die Arnoldi-Methode eine Folge orthonormaler Vektoren $q^{(t)} \in \mathbb{C}$ für $1 \leq t \leq m \ll n$ (Arnoldi-Basis), durch Anwendung der modifizierten Gram-Schmidt-Methode (2) auf die Basis $\{q, Aq, A^{m-1}q\}$ des Krylov-Unterraums K_m.

Speziell für diese Basis ergibt sich dann folgender Algorithmus:

Algorithmus 8: MGS für Krylov-Räume

Initialisierung: $A \in \mathbb{C}^{n \times n}$ beliebig und $q^{(1)} := q \in \mathbb{C}^n$ mit $\|q\|_2 = 1$
Ergebnis: : $\{q^{(t)}\}_{t=1}^{m}$ als Orthonormalbasis des Krylov-Raum K_m.

for $t = 2, \dots, n$
  $q^{(t)} = Aq^{(t-1)}$
  for $j = 1, \dots, t-1$
    $h_{j,t-1} \leftarrow \langle q^{(j)}, q^{(t)} \rangle$
    $q^{(t)} \leftarrow q^{(t)} - h_{j,t-1} \cdot q^{(j)}$
  end
  $h_{t,t-1} \leftarrow \left\| q^{(t)} \right\|$ $q^{(t)} \leftarrow q^{(t)} \,/\, h_{t,t-1}$
end

Sei $\mathcal{Q}^{(m)} := [q^{(1)}, q^{(2)}, \dots, q^{(m)}]$ die $n \times m$-Matrix aus den ersten Arnoldi-Vektoren und sei $H^{(m)}$ die

obere Hessenberg Matrix $(m \times m)$:

$$H^{(m)} = \begin{pmatrix} h_{11} & h_{12} & h_{13} & \dots & h_{1m} \\ h_{21} & h_{22} & h_{23} & \dots & \vdots \\ 0 & h_{32} & h_{33} & \dots & \vdots \\ \vdots & \ddots & \ddots & \ddots & h_{m-1,m} \\ 0 & \dots & 0 & h_{m,m-1} & h_{m,m} \end{pmatrix}$$

Die Matrizen $\mathcal{Q}^{(m)}$ sind orthonormal und mit (1) ergibt sich die Arnoldi-Beziehung

$$A\mathcal{Q}^{(m)} = \mathcal{Q}^{(m)} H^{(m)} + h_{m,m+1}[0, \dots, 0, q^{(m+1)}] \tag{3}$$

Multiplikation mit $\mathcal{Q}^{(m)H}$ und Verwendung von

$$\mathcal{Q}^{(m)H}\mathcal{Q}^{(m)} = I \quad \text{und} \quad \mathcal{Q}^{(m)H} q^{(m+1)} = 0$$

ergibt die für die Galerkin-Approximation benötigte Darstellung von $H^{(m)}$:

$$H^{(m)} = \mathcal{Q}^{(m)H} A \mathcal{Q}^{(m)}$$

Im Grenzfall $m = n$ ist die Matrix $H^{(n)}$ ähnlich zu A und hat die gleichen Eigenwerte.

Dies legt nahe, dass auch für $m \ll n$ die Eigenwerte der reduzierten Matrix $H^{(m)}$ eine gute Approximation einiger Eigenwerte von A sind. Wenn der Algorithmus endet (in exakter Arithmetik) für $m < n$ mit $h_{m,m+1} = 0$ dann ist der Krylov-Raum K_m ein invarianter Unterraum der Matrix A und die reduzierte Matrix $H^{(m)} = \mathcal{Q}^{(m)H} A \mathcal{Q}^{(m)}$ hat m gemeinsame Eigenwerte mit A, d.h. $\sigma(H^{(m)}) \subset \sigma(A)$ (vgl. Aufgabe 9.3)

Das folgende Lemma liefert a-posteriori Abschätzungen der Genauigkeit für die Approximation der Eigenwerte von A durch $H^{(m)}$.

Lemma 4.4. Sei $\{\mu, w\}$ ein Eigenpaar der Hessenberg-Matrix $H^{(m)}$ und sei $v = \mathcal{Q}^{(m)} w$, sodass $\{\mu, v\}$ ein approximiertes Eigenpaar von A ist, dann gilt

$$\|Av - \mu v\|_2 = |h_{m+1,m}| \cdot |w_m|$$

wobei w_m die letzte Komponente des Eigenvektors w ist.

Beweis. Die Arnoldi-Beziehung (3) liefert

$$\begin{aligned} Av &= A\mathcal{Q}^{(m)} w \\ &= \mathcal{Q}^{(m)} H^{(m)} w + h_{m+1,m} \cdot [0, \dots, 0, q^{(m+1)}] w \\ &= \mu \mathcal{Q}^{(m)} w + h_{m+1,m} \cdot [0, \dots, 0, q^{(m+1)}] w \\ &= \mu v + h_{m+1,m} \cdot [0, \dots, 0, q^{(m+1)}] w \\ &= \mu v + h_{m+1,m} \cdot w_m \cdot q^{(m+1)} \end{aligned}$$

Daraus folgt mit $\left\|q^{(m+1)}\right\|_2 = 1$, dass

$$\|Av - \mu v\|_2 = |h_{m+1,m}| \cdot |w_m|$$

□

Dies liefert keine a-priori-Information der Konvergenz der Eigenwerte von $H^{(m)}$ gegen die von A für $m \to n$, aber liefert eine a-posteriori-Prüfung, ob das erhaltene Paar $\{\mu, w\}$ eine gute Approximation ist, basierend auf den berechneten Größen $h_{m+q,m}$ und w_m.

Bemerkung 4.5. Die Ritz-Eigenwerte konvergieren zu den betragsgrößten Eigenwerten von A. Falls die betragskleinsten Eigenwerte bestimmt werden sollen, muss das diskutierte Verfahren auf die inverse Matrix angewendet werden (Vgl. Inverse Iteration nach Wielandt). In diesem Fall hat man einen großen Aufwand die Krylov-Räume $K_m = \operatorname{span} q, A^{-1}q, \ldots, A^{-m+1}q$ zu bestimmen, da hierfür die linearen Systeme $v^0 := q, Av^1 = v^0, \ldots, Av^m = v^{m-1}$ sukzessiv gelöst werden müssen.

Bemerkung 4.6. Typische Implementierungen der Arnoldi-Methode werden nach einer bestimmten Anzahl von Iterationen neu begonnen. Es kann untersucht werden, dass die Konvergenz sich mit einer größeren Krylov-Unterraum-Dimension m verbessert. Die Größe m, für die eine optimale Konvergenz erhalten wird, ist leider nicht im Voraus bekannt.

Stattdessen verwendet man sogenannte „Switching“ Strategien zum testen, ob ein Neustart sinnvoll ist, um die Konvergenz zu beschleunigen.

Bemerkung 4.7. Die in Algorithmus 7 vorgestellte Methode des modifizierten Gram-Schmidt-Verfahren für allgemeine Basen $\{v^{(t)}\}_{t=1}^n$ ergibt folgende Fehlerabschätzung[a]:

$$\left\| \mathcal{Q}^H \mathcal{Q} - I \right\|_2 \leq \frac{c_1 \cdot \operatorname{cond}_2(A)}{1 - c_2 \cdot \operatorname{cond}_2(A)}$$

wobei $\mathcal{Q} = [q^{(1)}, \ldots, q^{(n)}]$ und $A = [v^{(1)}, \ldots, v^{(n)}]$. Die Konstanten c_1, c_2 kommen aus $\mathcal{O}(n\varepsilon)$ wobei ε die Maschinengenauigkeit beschreibt.

[a] Åke Björck, Christopher C. Paige, Loss and recapture of orthogonality in the modified Gram-Schmidt algorithm, SIAM Journal on Matrix Analysis and Applications, Vol. 13, No. 1, 1992, pp. 176-190, DOI: 10.1137/0613013

Bemerkung 4.8. Andere Methoden zu Orthogonalisierung (wie z.B. Householder Transformation oder Givens-Rotation) sind zum Teil stabiler, als die stabilisierte Gram-Schmidt-Methode, aber aufgrund der iterativen Anwendungsmöglichkeit ist Gram-Schmidt beim Arnoldi-Verfahren vorteilhafter.

4.3 Lanczos-Methode

Für eine hermitesche Matrix A erhalten wir bei der Rekursions-Formel der Arnoldi-Methode:

$$\tilde{q}^{(t)} = Aq^{(t-1)} - \sum_{j=1}^{t-1} \langle Aq^{(t-1)}, q^{(j)} \rangle_2 q^{(j)}, \quad \text{für } t = 2, \ldots, m+1$$

Dabei ist wegen $A^H = A$ aber $\langle Aq^{(t-1)}, q^{(j)} \rangle_2 = \langle q^{(t-1)}, Aq^{(j)} \rangle_2$ und mit $Aq^{(j)} \in \operatorname{span}\{q^{(1)}, \ldots, q^{(j+1)}\}$ ergibt sich für $j = 1, \ldots, t-3$:

$$\begin{aligned} \langle Aq^{(t-1)}, q^{(j)} \rangle_2 &= q^{(j)H} A q^{(t-1)} = q^{(j)H} A^H q^{(t-1)} = (Aq^{(j)})^H q^{(t-1)} \\ &= \langle q^{(t-1)}, Aq^{(j)} \rangle_2 = \left\langle q^{(t-1)}, \sum_{i=1}^{j+1} c_i q^{(i)} \right\rangle_2 = \sum_{i=1}^{j+1} \overline{c_i} \underbrace{\langle q^{(t-1)}, q^{(i)} \rangle_2}_{=0} = 0 \end{aligned}$$

Dies vereinfacht unseren Ausdruck für $\tilde{q}^{(t)}$ zu

$$\begin{aligned} \tilde{q}^{(t)} &= Aq^{(t-1)} - \underbrace{\langle Aq^{(t-1)}, q^{(t-1)} \rangle_2}_{=:\alpha_{t-1}} q^{(t-1)} - \underbrace{\langle Aq^{(t-1)}, q^{t-2} \rangle_2}_{=:\beta_{t-2}} q^{t-2} \\ &= Aq^{(t-1)} - \alpha_{t-1} q^{(t-1)} - \beta_{t-2} q^{t-2} \end{aligned} \tag{1}$$

Da A hermitesch ist, folgt $\alpha_{t-1} \in \mathbb{R}$:

$$\alpha_{t-1} = \langle Aq^{(t-1)}, q^{(t-1)} \rangle_2 = \langle q^{(t-1)}, Aq^{(t-1)} \rangle_2 = \overline{\langle Aq^{(t-1)}, q^{(t-1)} \rangle_2} = \overline{\alpha_{t-1}}$$

Weiter gilt für β_{t-1}:

$$\begin{aligned} \left\| \tilde{q}^{(t)} \right\|_2 &= \langle q^{(t)}, \tilde{q}^{(t)} \rangle_2 \\ &\overset{(1)}{=} \langle q^{(t)}, Aq^{(t-1)} - \alpha_{t-1} q^{(t-1)} - \beta_{t-2} q^{t-2} \rangle_2 \\ &= \langle q^{(t)}, Aq^{(t-1)} \rangle_2 \\ &= \langle Aq^{(t)}, q^{(t-1)} \rangle_2 \\ &= \beta_{t-1} \end{aligned}$$

Daraus folgt, dass auch $\beta_{t-1} \in \mathbb{R}$ und $\beta_{t-1} q^{(t)} = \tilde{q}^{(t)}$. Also erhalten wir

$$Aq^{(t-1)} = \beta_{t-1} q^{(t)} + \alpha_{t-1} q^{(t-1)} + \beta_{t-2} q^{t-2}, \quad \text{für } t = 2, \ldots, m+1 \tag{2}$$

oder in Matrix-Form:

$$A \cdot \mathcal{Q}^{(m)} = \mathcal{Q}^{(m)} \cdot \underbrace{\begin{pmatrix} \alpha_1 & \beta_2 & & & & \\ \beta_2 & \alpha_2 & \beta_3 & & & \\ & \beta_3 & \alpha_3 & \ddots & & \\ & & & \ddots & \beta_{m-1} & \\ & & & \beta_{m-1} & \alpha_{m-1} & \beta_m \\ & & & & \beta_m & \alpha_m \end{pmatrix}}_{=:T^{(m)}} + \beta_m \cdot [0, \ldots, 0, q^{(m+1)}]$$

wobei die Matrix $T^{(m)} \in \mathbb{R}^{m \times m}$ ist reell und symmetrisch. Von dieser sogenannten Lanczos-Beziehung ergibt sich

$$\mathcal{Q}^{(m)H} A \mathcal{Q}^{(m)} = T^{(m)}$$

Definition 4.1 (Lanczos-Algorithmus). Für eine hermitesche Matrix $A \in \mathbb{C}^{n \times n}$ bestimmt die Lanczos-Methode eine Menge von orthogonalen Vektoren $\{q^{(1)}, \ldots, q^{(m)}\}$, $m \ll n$ durch Anwendung der Gram-Schmidt-Methode auf die Basis $\{q, Aq, \ldots, A^{m-1}q\}$ des Krylov-Raumes K_m.

Durch Umstellen von (2) erhalten wir folgenden Algorithmus:

Algorithmus 9: Lanczos-Algorithmus

Initialisierung: Sei $A \in \mathbb{C}^{n \times n}$ hermitesch und $q^{(1)} := q \in \mathbb{C}^n$ mit $\|q\|_2 = 1$

$q^{(0)} = 0$
$\beta_1 = 0$
for $t = 1, \ldots, m-1$
  $r^{(t)} = Aq^{(t)}$
  $\alpha_t = \langle r^{(t)}, q^{(t)} \rangle_2$
  $s^{(t)} = r^{(t)} - \alpha_t q^{(t)} - \beta_t q^{(t-1)}$
  $\beta_{t+1} = \left\| s^{(t)} \right\|$
  $q^{(t+1)} = s^{(t)} / \beta_{t+1}$
end
$r^{(m)} = Aq^{(m)}$
$\alpha_m = \langle r^{(m)}, q^{(m)} \rangle_2$

Nachdem die Matrix $T^{(m)}$ berechnet ist, kann ihr Eigenwert λ_i und der zugehörige Eigenvektor $w^{(i)}$ bestimmt werden (z.B. mit QR-Algorithmus).

Die Eigenwerte und Eigenvektoren von $T^{(m)}$ werden mit dem Aufwand $\mathcal{O}(m^2)$ berechnet und approximieren die der ursprünglichen Matrix A. Die zugehörigen Ritz Eigenvektoren $v^{(i)}$ können dann mit $v^{(i)} = \mathcal{Q}^{(m)} \cdot w^{(i)}$ berechnet werden.

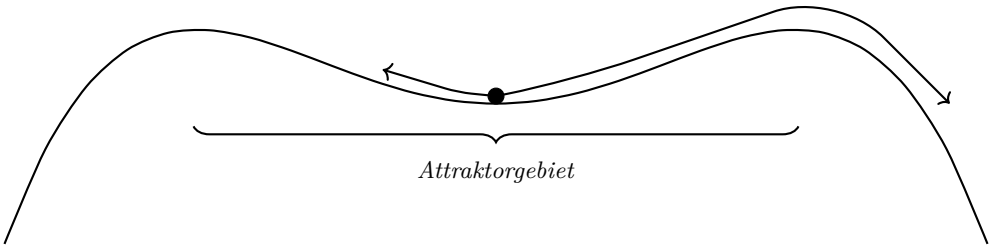
4.4 Pseudospektren

Motivation: Wir stellen uns einen Ball vor, der in einer Landschaft liegt, abhängig von der Form des Untergrundes kann der Ball bei Impulseinwirkung entweder wieder zur Ruhe kommen oder wegrollen:

stabil instabil

Im linken Fall haben wir ein stabiles System, d.h. kleine Störungen (der Ball wird leicht angestoßen) führen nicht dazu, dass sich der Zustand des Systems stark ändert. Rechts hingegen reicht ein kleiner Impuls, sodass der Ball wegrollt, das System ist daher instabil.

Betrachten wir ein scheinbar stabiles System herausgezoomt, so erkennen wir, dass zu große Störungen doch wieder zu einer starken System Änderung sorgt.



Den Bereich, in welchem Störungen kein Problem darstellen, das System also stabil bleibt, nennen wir Attraktorgebiet. In der Praxis sind wir oftmals daran interessiert wie groß das Attraktorgebiet ist, also wie viel Störung verkraftbar ist, ohne dass unser System instabil wird.

Basierend auf dieser Idee wollen wir nun den Begriff der Pseudospektren im Kontext von Matrizen / Eigenwerten einführen.

Die Grundidee des Pseudospektrums kann jedoch auch in viel allgemeineren Situationen eingeführt werden.

Definition 4.9. Für $\varepsilon \in \mathbb{R}_+$, ist das ε-Pseudospektrum $\sigma_\varepsilon(A) \subset \mathbb{C}$ einer Matrix $A \in \mathbb{K}^{n\times n}$ definiert als

$$\sigma_\varepsilon(A) := \{z \in \mathbb{C}\backslash\sigma(A) \,:\, \left\|(A - zI)^{-1}\right\|_2 \geq \tfrac{1}{\varepsilon}\} \cup \sigma(A)$$

Bemerkung 4.10. Die Krylov-Unterraum-Methoden, die bisher diskutiert wurden, lassen sich zur Berechnung des Pseudo-Spektrums einer Matrix verwenden. (z.B. bei der Matrix von diskretisierten partiellen Differentialgleichungen)

Lemma 4.11.

1. Das ε-Pseudospektrum einer Matrix $T \in \mathbb{C}^{n\times n}$ kann definiert werden durch
$$\sigma_\varepsilon(T) := \{z \in \mathbb{C} \,|\, \sigma_{\min}(zI - T) \leq \varepsilon\}$$
wobei $\sigma_{\min}(B)$ den kleinsten Singulärwert der Matrix B bezeichnet, d.h.
$$\sigma_{\min}(B) := \sqrt{\lambda_{\min}(B^H B)}$$
2. Das ε-Pseudospektrum $\sigma_\varepsilon(T)$ einer Matrix $T \in \mathbb{C}^{n\times n}$ ist invariant unter Orhtonormalen Transformationen, d.h. für eine unitäre Matrix $Q \in \mathbb{C}^{n\times n}$ gilt $\sigma_\varepsilon(Q^*TQ) = \sigma_\varepsilon(T)$

Beweis.

1. Es gilt
$$\begin{aligned}\left\|(zI - T)^{-1}\right\|_2 &= \sqrt{\lambda_{\max}((zI - T)^{-H})(zI - T)^{-1}} \\ &= \sigma_{\max}((zI - T)^{-1}) \\ &= \sigma_{\min}((zI - T))^{-1}\end{aligned}$$
Daraus folgt:
$$\begin{aligned}\sigma_\varepsilon(T) &= \left\{z \in \mathbb{C} \;\middle|\; \left\|(zI - T)^{-1}\right\|_2 \geq \tfrac{1}{\varepsilon}\right\} \\ &= \left\{z \in \mathbb{C} \;\middle|\; \sigma_{\min}(zI - T)^{-1} \geq \tfrac{1}{\varepsilon}\right\} \\ &= \{z \in \mathbb{C} \mid \sigma_{\min}(zI - T) \leq \varepsilon\}\end{aligned}$$
2. vgl. Aufgabe 10.1 □

Numerische Berechnung
Zur näherungsweisen Bestimmung von ε-Pseudospektren einer Matrix T betrachtet man in der Praxis eine diskrete Menge von Gitterpunkten $D \subset \mathbb{C}$ und berechnet alle Punkte $z_i \in D$ den kleinsten Singulärwert von $(z_i I - T)$. Diese Werte liefern nach Lemma 4.11 das größtmögliche ε, so dass $z_i \in \sigma_\varepsilon(T)$.

Anschaulich liefern die Pseudospektren eine Art Höhenprofilüber der komplexen Ebene, mittels Interpolation der Gitterwerte erhalten wir eine grafische Darstellung der zugehörigen Höhenlinien.

5 Die schnelle Fourier-Transformation

Im folgenden Abschnitt wollen wir uns mit der schnellen Fourier-Transformation („FFT - fast Fourier transform) als zentrales Werkzeug der Signalverarbeitung und Bildkompression. Um die Idee hinter dem FFT-Algorithmus zu verstehen beginnen wir mit einer kurzen Wiederholung zu Fourier-Reihen.

5.1 Fourier-Reihen

Wir betrachten f eine 2π-periodische Funktion (d.h. $f(x+2\pi) = f(x)$ für alle $x \in \mathbb{R}$) mit dem Ziel eine Annäherung durch Linearkombinationen 2π-periodischen Funktionen $\{\cos(kx)\}_{k=0}^{n}$ und $\{\sin(kx)\}_{k=1}^{n}$ zu finden:

$$g_n(x) = \frac{1}{2}a_0 + \sum_{k=1}^{n} \Big(a_k \cos(kx) + b_k \sin(kx) \Big)$$

Wir suchen eine Approximation im Sinne der L_2 Norm, d.h. wir minimieren den Ausdruck

$$\|g_n(x) - f(x)\|_2 = \left(\int_0^{2\pi} (g_n(x) - f(x))^2 \, \mathrm{d}x \right)^{1/2}$$

Satz 5.1 (Fourier-Koeffizienten). Für trigonometrisches Polynom, d.h. eine Funktion der Form

$$g_n(x) = \frac{1}{2}a_0 + \sum_{k=1}^{n} \Big(a_k \cos(kx) + b_k \sin(kx) \Big)$$

gilt

$$a_k = \frac{1}{\pi} \int_0^{2\pi} g_n(x) \cos(kx) \, \mathrm{d}x, \quad k = 0, 1, \ldots, n$$
$$b_k = \frac{1}{\pi} \int_0^{2\pi} g_n(x) \sin(kx) \, \mathrm{d}x, \quad k = 1, \ldots, n$$

Beweis. Durch Verwendung der Orthogonalitätsbedingungen der trigonometrischen Funktionen ergibt sich für $l = 0$:

$$\begin{aligned}
\frac{1}{\pi} \int_0^{2\pi} g_n(x) \underbrace{\cos(lx)}_{1} \mathrm{d}x &= \frac{1}{\pi} \int_0^{2\pi} \left(\frac{1}{2}a_0 + \sum_{k=1}^{n} \Big(a_k \cos(kx) + b_k \sin(kx) \Big) \right) \cdot 1 \, \mathrm{d}x \\
&= \frac{1}{2\pi} a_0 \cdot \int_0^{2\pi} 1 \, \mathrm{d}x + \sum_{k=1}^{n} \left(\frac{1}{\pi} a_k \cdot \underbrace{\int_0^{2\pi} \cos(kx) \, \mathrm{d}x}_{0} + \frac{1}{\pi} b_k \cdot \underbrace{\int_0^{2\pi} \sin(kx) \, \mathrm{d}x}_{0} \right) \\
&= \frac{1}{2\pi} a_0 \cdot 2\pi = a_0
\end{aligned}$$

und für $1 \leq l \leq n$:

$$\begin{aligned}
&\frac{1}{\pi}\int_0^{2\pi} g_n(x)\cos(lx)\,\mathrm{d}x \\
&\quad= \frac{1}{2\pi}a_0 \cdot \underbrace{\int_0^{2\pi}\cos(lx)\,\mathrm{d}x}_{0} + \sum_{k=1}^{n}\left(\frac{1}{\pi}a_k \cdot \underbrace{\int_0^{2\pi}\cos(kx)\cos(lx)\,\mathrm{d}x}_{\pi \text{ wenn } l=k,\text{ sonst } 0} + \frac{1}{\pi}b_k \cdot \underbrace{\int_0^{2\pi}\sin(kx)\cos(lx)\,\mathrm{d}x}_{0}\right) \\
&\quad= \frac{1}{\pi}a_l \cdot \pi = a_l
\end{aligned}$$

und

$$\begin{aligned}
&\frac{1}{\pi}\int_0^{2\pi} g_n(x)\sin(lx)\,\mathrm{d}x \\
&\quad= \frac{1}{2\pi}a_0 \cdot \underbrace{\int_0^{2\pi}\sin(lx)\,\mathrm{d}x}_{0} + \sum_{k=1}^{n}\left(\frac{1}{\pi}a_k \cdot \underbrace{\int_0^{2\pi}\cos(kx)\sin(lx)\,\mathrm{d}x}_{0} + \frac{1}{\pi}b_k \cdot \underbrace{\int_0^{2\pi}\sin(kx)\sin(lx)\,\mathrm{d}x}_{\pi \text{ wenn } l=k,\text{ sonst } 0}\right) \\
&\quad= \frac{1}{\pi}b_l \cdot \pi = b_l
\end{aligned}$$

□

In Unserem Fall, wo wir f durch g_n annähern wollen verwenden wir daher $f(x)$ bei der Bestimmung unserer Koeffizienten.

Im Allgemeinen ergeben sich für die Fourier-Koeffizienten $\{a_k\}_{k=0}^n$ und $\{b_k\}_{k=1}^n$ keine geschlossenen Formeln, d.h. wir sind auf numerische Integration angewiesen um diese zu bestimmen.

Verenden wir die Trapezregel als Quadraturformel um diese numerische Integration durchzuführen:

Definition 5.2 (Trapezregel). Ein Verfahren zur numerischen Integration einer Funktion $f : [a,b] \to \mathbb{R}$ wird durch die Trapezregel beschrieben. Sie beruht auf der Idee das Intervall $[a,b]$ in kleinere Intervalle $[x_j, x_{j+1}]$ für $j = 0, \ldots, N-1$ mit $a = x_0 < x_1 < \cdots < x_N = b$ aufzuteilen und die Funktion auf jedem dieser Intervalle als linear anzunehmen, dies ermöglicht folgende Annäherung

$$\int_{x_j}^{x_{j+1}} f(x)\,\mathrm{d}x \approx (x_{j+1} - x_j) \cdot \frac{f(x_{j+1}) + f(x_j)}{2}$$

Insbesondere für den Fall von äquidistant gewählten Stützstellen mit Schrittweite $h = \frac{b-a}{N}$ ergibt sich

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{h}{2}\left(f(a) + 2 \cdot \sum_{j=1}^{N-1} f(a + h \cdot j) + f(b)\right)$$

Verwenden wir diese Trapezregel mit $x_j = \frac{2\pi}{N} \cdot j$ um unsere Fourier-Koeffizienten anzunähern ergibt sich die diskrete Fourier-Transformation (DFT):

$$\begin{aligned}
a_k &\approx \frac{1}{N}\left(f(x_0)\cdot\cos(kx_0) + 2\sum_{j=1}^{N-1} f(x_j)\cdot\cos(kx_j) + f(x_N)\cdot\cos(kx_N)\right) \\
b_k &\approx \frac{1}{N}\left(f(x_0)\cdot\sin(kx_0) + 2\sum_{j=1}^{N-1} f(x_j)\cdot\sin(kx_j) + f(x_N)\cdot\sin(kx_N)\right)
\end{aligned}$$

Mit Berücksichtigung der 2π-Periodizität von f ergeben sich für a_k und b_k die Näherungswerte

$$a_k^* := \frac{2}{N}\sum_{j=1}^{N} f(x_j)\cdot\cos(kx_j), \quad k = 0,1,2,\dots$$

$$b_k^* := \frac{2}{N}\sum_{j=1}^{N} f(x_j)\cdot\cos(kx_j), \quad k = 1,2,3\dots$$

Intuition Bilder

Lemma 5.3. Für die diskreten Stützstellen $x_j = \frac{2\pi}{N}\cdot j$ mit $1 \le N$ gilt

$$\sum_{j=1}^{N}\cos(kx_j) = \begin{cases} 0, & \text{falls } \frac{k}{N} \notin \mathbb{Z} \\ N, & \text{falls } \frac{k}{N} \in \mathbb{Z} \end{cases}$$

$$\sum_{j=1}^{N}\sin(kx_j) = 0 \text{ für alle } k \in \mathbb{Z}$$

Beweis.
Wir betrachten die komplexe Kombination beider Ausdrücke und erhalten

$$S_N := \sum_{j=1}^{N}\cos(kx_j) + i\sin(kx_j) = \sum_{j=1}^{N} e^{ikx_j} = \sum_{j=1}^{N} e^{ik\cdot jh}$$

Dies ist eine endliche geometrische Reihe mit komplexem $q := e^{ikh} = e^{2\pi ik/N}$

Ist $\frac{k}{N} \notin \mathbb{Z}$, dann ist $q \neq 1$, und die Summenformel der endlichen geometrischen Reihe liefert

$$S_N = e^{ikh}\frac{e^{ikhN}-1}{e^{ikh}-1} = e^{ikh}\cdot\frac{e^{2\pi ki}-1}{e^{ikh}-1} = 0, \text{wenn } \frac{k}{N} \notin \mathbb{Z}$$

Für $\frac{k}{N} \in \mathbb{Z}$ folgt wegen $q = 1$, dass $S = N$ ist.

Die Unabhängigkeit von Real- und Imaginärteil schließt den Beweis. □

Satz 5.4. Die trigonometrischen Funktionen erfüllen für die äquidistanten Stützstellen x_j die diskreten Orthogonalitätsrelationen:

$$\sum_{j=1}^{N}\cos(kx_j)\cos(lx_j) = \begin{cases} 0, & \text{falls } \frac{k+l}{N} \notin \mathbb{Z} \text{ und } \frac{k-l}{N} \notin \mathbb{Z} \\ N & \text{falls } \frac{k+l}{N} \in \mathbb{Z} \text{ und } \frac{k-l}{N} \in \mathbb{Z} \\ \frac{N}{2} & \text{falls } \frac{k+l}{N} \in \mathbb{Z} \text{ und } \frac{k-l}{N} \notin \mathbb{Z} \\ \frac{N}{2} & \text{falls } \frac{k+l}{N} \notin \mathbb{Z} \text{ und } \frac{k-l}{N} \in \mathbb{Z} \end{cases}$$

und

$$\sum_{j=1}^{N}\sin(kx_j)\sin(lx_j) = \begin{cases} 0, & \text{falls } \frac{k+l}{N} \notin \mathbb{Z} \text{ und } \frac{k-l}{N} \notin \mathbb{Z} \\ 0 & \text{falls } \frac{k+l}{N} \in \mathbb{Z} \text{ und } \frac{k-l}{N} \in \mathbb{Z} \\ -\frac{N}{2} & \text{falls } \frac{k+l}{N} \in \mathbb{Z} \text{ und } \frac{k-l}{N} \notin \mathbb{Z} \\ \frac{N}{2} & \text{falls } \frac{k+l}{N} \notin \mathbb{Z} \text{ und } \frac{k-l}{N} \in \mathbb{Z} \end{cases}$$

und

$$\sum_{j=1}^{N}\cos(kx_j)\sin(lx_j) = 0 \quad \text{für all } k,l \in \mathbb{N}$$

Beweis.
Zur Überprüfung der Orthogonalitätsrelationen werden die trigonometrischen Identitäten

$$\cos(kx_j)\cos(lx_j) = \tfrac{1}{2}\Big(\cos\big((k+l)x_j\big) + \cos\big((k-l)x_j\big)\Big)$$
$$\sin(kx_j)\sin(lx_j) = \tfrac{1}{2}\Big(\cos\big((k-l)x_j\big) - \cos\big((k+l)x_j\big)\Big)$$
$$\cos(kx_j)\sin(lx_j) = \tfrac{1}{2}\Big(\sin\big((k+l)x_j\big) - \sin\big((k-l)x_j\big)\Big)$$

verwendet und das Lemma 5.3 angewandt. □

Satz 5.5. Sei $N = 2n$ mit $n \in \mathbb{N}$. Das Fourier-Polynom

$$g_m^*(x) := \tfrac{1}{2}a_0^* + \sum_{k=1}^{m}\Big(a_k^*\cos(kx) + b_k^*\sin(kx)\Big)$$

von Grad $m < n$ mit Koeffizienten a_k^* und b_k^* approximiert die Funktion $f(x)$ im diskreten quadratischen Mittel der N Stützstellen x_j derart, dass die Summe der quadratischen Abweichungen

$$F := \sum_{j=1}^{N}\Big(g_m^*(x_j) - f(x_k)\Big)^2$$

minimal ist.

ohne Beweis.

Beispiel 5.6. Sei $f(x) = x^2$:
x^2 Plot

5.2 Effiziente Berechnung der Fourier-Koeffizienten

Die näherungsweise Berechnung der Fourier-Koeffizienten a_k^* und b_k^* ist für eine große Anzahl N der Stützstellen sehr aufwendig.

Dies ist vor allem bei der diskreten Fourier-Transformation relevant, die in Ingenieur- und Naturwissenschaften häufig eingesetzt wird, um z.B. die Frequenzen von Vibrationen zu bestimmen.

Zur Berechnung der Summen

$$a_k' := \sum_{j=0}^{N-1} f(x_j)\cos(kx_j), \quad k = 0, 1, \ldots, \tfrac{N}{2}$$
$$b_k' := \sum_{j=0}^{N-1} f(x_j)\sin(kx_j), \quad k = 1, 2, \ldots, \tfrac{N}{2} - 1 \tag{1}$$

werden normalerweise $\propto N^2$ trigonometrische Funktionsauswertungen verlangt. Für den Fall, dass N eine Potenz von 2 ist, kann ein sehr effizienter Algorithmus ($\propto n\log(n)$ Auswertungen) entwickelt werden, indem wir zu einer komplexen Fourier-Transformation übergehen.

Definition 5.7 (Diskrete komplexe Fourier-Transformation). Für eine Folge von komplexen Zahlen $f = (f_0, \dots, f_{n-1})^T \in \mathbb{C}^n$ ergibt sich die diskrete komplexe Fourier-Transformation $\hat{f}$ durch

$$\hat{f}_k := \sum_{j=0}^{n-1} f_j \cdot e^{-2\pi i \cdot \frac{jk}{n}} = \sum_{j=0}^{n-1} f_j \cdot \omega_n^{jk}$$

Dabei sind ω_n die n-ten Einheitswurzeln:

$$\omega_n := e^{-2\pi i \,/\, n} = \cos\left(\frac{2\pi}{n}\right) + i \cdot \sin\left(\frac{2\pi}{n}\right)$$

In Matrix Schreibweise entspricht dies

$$\begin{pmatrix} \hat{f}_0 \\ \hat{f}_1 \\ \hat{f}_2 \\ \vdots \\ \hat{f}_{n-1} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega_n & \omega_n^2 & \cdots & \omega_n^{n-1} \\ 1 & \omega_n^2 & \omega_n^4 & \cdots & \omega_n^{2(n-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega_n^{n-1} & \omega_n^{2(n-1)} & \cdots & \omega_n^{(n-1)^2} \end{pmatrix} \cdot \begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ \vdots \\ f_{n-1} \end{pmatrix}$$

In unserem Fall haben wir eine reellwertige Funktion f und damit auch reellwertige Punkte $f(x_j)$, wir können dennoch eine diskrete komplexe Fourier-Transformation durchführen und uns aus dem Resultat dann unsere reelle diskrete Fourier-Transformation berechnen:

Satz 5.8 (Zusammenhang zwischen reeller und komplexer DFT). Sei $\hat{y} = (\hat{y}_0, \dots, \hat{y}_{n-1})^T$ die komplexe DFT von $y = (y_0, \dots, y_{n-1})^T \in \mathbb{C}^n$, wobei y folgende Darstellung hat

$$y_j := f(x_{2j}) + i \cdot f(x_{2j+1}), \quad j = 0, \dots, n-1$$

Die trigonometrischen Summen a'_k und b'_k (1) sind gegeben durch

$$a'_k - i \cdot b'_k = \tfrac{1}{2}(\hat{y}_k + \overline{\hat{y}_{n-k}}) + \tfrac{1}{2i}(\hat{y}_k - \overline{\hat{y}_{n-k}})e^{-i\pi \cdot \frac{k}{n}}$$
$$a'_{n-k} - i \cdot b'_{n-k} = \tfrac{1}{2}(\overline{\hat{y}_k} + \hat{y}_{n-k}) + \tfrac{1}{2i}(\overline{\hat{y}_k} - \hat{y}_{n-k})e^{i\pi \cdot \frac{k}{n}}$$

für $k = 0, \dots n$, wobei $b'_0 = b'_n = 0$ und $\hat{y}_n = \hat{y}_0$

Beweis. Durch

$$\overline{\omega_n^{j(n-k)}} = \underbrace{\overline{\omega_n^{jn}}}_{1} \cdot \overline{\omega_n^{-jk}} = \overline{e^{-2\pi i \,/\, n \cdot (-jk)}} = e^{-2\pi i \,/\, n \cdot jk} = \overline{\omega_n^{jk}}$$

erhalten wir für die Summanden der oberen Formel

$$\begin{aligned} \tfrac{1}{2}(\hat{y}_k + \overline{\hat{y}_{n-k}}) &= \frac{1}{2} \cdot \sum_{j=0}^{n-1} \left(y_j \cdot \omega_n^{jk} + \overline{y}_j \cdot \overline{\omega_n^{j(n-k)}} \right) \\ &= \frac{1}{2} \cdot \sum_{j=0}^{n-1} \left(y_j + \overline{y}_j \right) \cdot \omega_n^{jk} \\ &= \frac{1}{2} \cdot \sum_{j=0}^{n-1} \left(y_j + \overline{y}_j \right) \cdot e^{-2\pi i \cdot \frac{jk}{n}} \end{aligned}$$

und

$$\begin{aligned}\tfrac{1}{2i}(\hat{y}_k - \overline{\hat{y}_{n-k}})e^{-i\pi\cdot\frac{k}{n}} &= \frac{1}{2i}\cdot\sum_{j=0}^{n-1}\left(y_j\cdot\omega_n^{jk} - \overline{y}_j\cdot\overline{\omega_n^{j(n-k)}}\right)e^{-i\pi\cdot\frac{k}{n}}\\ &= \frac{1}{2i}\cdot\sum_{j=0}^{n-1}\left(y_j-\overline{y}_j\right)\cdot\omega_n^{jk}\cdot e^{-i\pi\cdot\frac{k}{n}}\\ &= \frac{1}{2i}\cdot\sum_{j=0}^{n-1}\left(y_j-\overline{y}_j\right)\cdot e^{-i\pi(2j+1)\frac{k}{n}}\end{aligned}$$

Mit Definition von y_j ergibt sich

$$\begin{aligned}y_j+\overline{y}_j &= 2\cdot\mathrm{Re}(y_j) = 2\cdot f(x_{2j})\\ y_j-\overline{y}_j &= 2i\cdot\mathrm{Im}(y_j) = 2\cdot f(x_{2j+1})\end{aligned}$$

und für die Summe

$$\begin{aligned}&\tfrac{1}{2}(\hat{y}_k+\overline{\hat{y}_{n-k}}) + \tfrac{1}{2i}(\hat{y}_k-\overline{\hat{y}_{n-k}})e^{-i\pi\cdot\frac{k}{n}}\\ &= \sum_{j=0}^{n-1}\left(f(x_{2j})e^{-ijk\frac{2\pi}{n}} + f(x_{2j+1})e^{-i(2j+1)k\frac{\pi}{n}}\right)\\ &= \sum_{j=0}^{n-1}\left(f(x_{2j})\left[\cos(kx_{2j})-i\sin(kx_{2j})\right] + f(x_{2j+1})\left[\cos(kx_{2j+1})-i\sin(kx_{2j+1})\right]\right)\\ &= \sum_{j=0}^{n-1}\left(f(x_{2j})\cos(kx_{2j}) + f(x_{2j+1})\cos(kx_{2j+1})\right)\\ &\quad - i\cdot\sum_{j=0}^{n-1}\left(f(x_{2j})\sin(kx_{2j}) + f(x_{2j+1})\sin(kx_{2j+1})\right)\\ &= a'_k - ib'_k\end{aligned}$$

Die zweite Formel des Satzes ergibt sich durch Substitution von k durch $n-k$. □

Der Vorteil der komplexen DFT ist, dass eine Reduktion gerader Ordnung auf zwei komplexe DFT je der halben Ordnung möglich ist, führen wir diese Reduktion iterativ durch (was bei einer Potenz von 2 möglich ist) erhalten wir einen Divide & Conquer Algorithmus mit linear-logarithmischer Laufzeit.

Satz 5.9. Sei $n = 2m$ mit $m \in \mathbb{N}$. Für die komplexen Fourier-Koeffizienten gilt:

$$\begin{aligned}\hat{f}_{2l} &= \sum_{j=0}^{m-1}(f_j + f_{m+j})\omega_n^{2lj}\\ \hat{f}_{2l+1} &= \sum_{j=0}^{m-1}(f_f - f_{m+j})\omega_n^{j}\cdot\omega_n^{2lj}\end{aligned}$$

Beweis.
Für die Einheitswurzeln gilt

$$\omega_n^{2l(m+j)} = \omega_n^{2lj}\cdot\omega_n^{2lm} = \omega^{2lj}\cdot\left(\omega_n^{2m}\right)^l = \omega^{2lj}\cdot\Big(\underbrace{\omega_n^n}_{1}\Big)^l = \omega_n^{2lj} = \omega_m^{lj}$$

Diese Identität liefert die gewünschte Umformung der Fourier-Koeffizienten:

$$\begin{aligned}\hat{f}_{2l} &= \sum_{j=0}^{2m-1} f_j \omega_n^{2lj} \\ &= \sum_{j=0}^{m-1} f_j \omega_n^{2lj} + \sum_{j=0}^{m-1} f_{m+j} \underbrace{\omega_n^{2l(m+j)}}_{\omega_n^{2lj}} \\ &= \sum_{j=0}^{m-1} (f_j + f_{m+j}) \omega_m^{lj}\end{aligned}$$

und

$$\begin{aligned}\hat{f}_{2l+1} &= \sum_{j=0}^{2m-1} f_j \omega_n^{(2l+1)j} \\ &= \sum_{j=0}^{m-1} f_j \omega_n^{(2l+1)j} + \sum_{j=0}^{m-1} f_{j+m} \omega_n^{(2l+1)(m+j)} \\ &= \sum_{j=0}^{m-1} f_j \omega_n^{2lj} \omega_n^j + \sum_{j=0}^{m-1} f_{j+m} \underbrace{\omega_n^{2l(m+j)}}_{\omega_n^{2lj}} \underbrace{\omega_n^m}_{-1} \omega_n^j \\ &= \sum_{j=0}^{m-1} (f_j - f_{m+j}) \omega_n^j \cdot \omega_m^{lj}\end{aligned}$$

□

Dieser Satz ermöglicht es uns nun die Fourier-Transformierte $\hat{f}$ als Kombination von zwei neuen Fourier-Transformationen zu schreiben, denn für die Hilfswerte

$$g_j := f_j + f_{m+j} \quad \text{und} \quad h_j := (f_j) - (f_{m+j})\omega_n^j$$

gilt

$$\hat{g} = (\hat{f}_0, \hat{f}_2, \ldots, \hat{f}_{2m-1})^T \quad \text{und} \quad \hat{h} = (\hat{f}_1, \hat{f}_3, \ldots, \hat{f}_{2m})^T$$

Wir haben damit die Berechnung einer DFT von $f \in \mathbb{C}^{2m}$ (einmal Ordnung $2m$) auf die Berechnung zweier DFTs von $g \in \mathbb{C}^m$ und $h \in \mathbb{C}^m$ (zweimal Ordnung m).

Beispiel 5.10. Wollen wir eine DFT Ordnung 32 durchführen, so brechen wir dies erst auf die Berechnung von zwei DFTs mit Ordnung 16 herunter. Jede dieser DFTs wird dann wiederum in zwei DFTs mit Ordnung 8 vereinfacht. Wiederholt man dies iterativ so ergibt sich:
$FT_{32} \to 2\ FT_{16} \to 4\ FT_8 \to 8\ FT_4 \to 16\ FT_2 \to 32\ FT_1$

Bemerkung 5.11. Der Rechenaufwand für die Berechnung einer diskreten komplexen Fourier-Transformation nach der Methode des Aufteilens entspricht $\mathcal{O}(n \log(n))$.

5.3 Symmetrische Transformationen

In der Anwendung ist es oftmals hilfreich die symmetrischen Fortsetzungen von reellen Funktionen zu betrachten. Dies ermöglicht die Fourier-Darstellung nur auf Sinus- oder nur auf Kosinusreihen zu reduzieren.

Auch hier geben wir eine kurze Wiederholung zur Fourier-Reihe:

Definition 5.12. Sei $f : \mathbb{R} \to \mathbb{C}$ eine 2π-periodische, über $[0, 2\pi]$ integrierbare Funktion, dann heißt

$$\hat{f}(x) = \sum_{k=-\infty}^{\infty} c_k e^{ikx}$$

Fourier-Reihe von f, wobei sich die Fourier-Koeffizienten c_k durch

$$c_k := \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ikx} \, \mathrm{d}x, \qquad k \in \mathbb{Z}$$

Eine alternative Darstellung ergibt sich durch

$$\hat{f}(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \big(a_k \cos(kx) + b_k \sin(kx) \big)$$

mit

$$a_k := \frac{1}{\pi} \int_0^{2\pi} f(x) \cos(kx) \, \mathrm{d}x$$
$$b_k := \frac{1}{\pi} \int_0^{2\pi} f(x) \sin(kx) \, \mathrm{d}x$$

Satz 5.13. Beide Darstellungen in Definition 5.12 sind äquivalent und für die Umrechnung der Darstellungen gilt

$$c_k = \begin{cases} \frac{1}{2}(a_k - ib_k), & k > 0 \\ \frac{a_0}{2}, & k = 0 \\ \frac{1}{2}(a_{-k} + ib_{-k}), & k < 0 \end{cases} \quad \text{und} \quad a_k = \begin{cases} c_k + c_{-k}, & k > 0 \\ 2c_0 \end{cases}, \quad b_k = i(c_{-k} - c_k)$$

Beweis. Ergibt sich direkt durch die Verwendung der eulerschen Formel und der Euler-Darstellung von sin und cos.

Bemerkung 5.14.

1. Wenn f punktsymmetrisch bzgl. $x = \pi$ ist, d.h. $f(\pi + x) = -f(\pi - x)$, so gilt

$$\begin{aligned} a_k &= \frac{1}{\pi} \int_0^{2\pi} f(x) \cos(kx) \, \mathrm{d}x \\ &= \frac{1}{\pi} \int_{2\pi}^{0} f(2\pi - z) \cos(k(2\pi - z))(-\mathrm{d}z) \qquad (z = 2\pi - x) \\ &= \frac{1}{\pi} \int_0^{2\pi} -f(z) \cos(kz) \, \mathrm{d}z \\ &= -a_n \\ \Rightarrow \quad a_k &= 0 \end{aligned}$$

   Also ist die Fourier-Reihe von f gegeben durch

$$\hat{f}(x) = \sum_{k=1}^{\infty} b_k \sin(kx)$$

2. Wenn f spiegelsymmetrisch bzgl. $x = \pi$ ist, d.h. $f(\pi + x) = f(\pi - x)$, so gilt analog (wegen $\sin(k(2\pi - z)) = -\sin(kz)$), dass die Fourier-Reihe von f gegeben ist durch

$$\hat{f}(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(kx)$$

Sinustransformation:
Sei f eine Funktion, welche auf den Intervallgrenzen von $I = [0, \pi]$ verschwindet, d.h. $f(0) = f(\pi) = 0$, so kann diese punktsymmetrisch bzgl. $x = \pi$ über $[0, 2\pi]$ fortgesetzt werden durch

$$f_u(x) = \begin{cases} f(x), & 0 \le x \le \pi \\ -f(2\pi - x), & \pi < x \le 2\pi \end{cases}$$

und es ergibt sich eine Sinustransformation

$$\hat{f}_u = \sum_{k=1}^{\infty} b_k \sin(kx), \quad 0 \le x \le \pi, \quad \text{mit} \quad b_k = \frac{2}{\pi} \int_0^{\pi} f(x) \sin(kx)\, \mathrm{d}x$$

Kosinustransformation:
Sei f eine beliebige Funktion auf $I = [0, \pi]$, so kann diese spiegelsymmetrisch bzgl. $x = \pi$ über $[0, 2\pi]$ fortgesetzt werden durch

$$f_g(x) = \begin{cases} f(x), & 0 \le x \le \pi \\ -f(2\pi - x), & \pi < x \le 2\pi \end{cases}$$

und es ergibt sich eine Sinustransformation

$$\hat{f}_g = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(kx), \quad 0 \le x \le \pi, \quad \text{mit} \quad a_k = \frac{2}{\pi} \int_0^{\pi} f(x) \cos(kx)\, \mathrm{d}x$$

5.4 diskrete Kosinustransformation

In der Anwendung auf eine diskrete Zeitreihe $x_0, \ldots, x_{N-1}$, zum Beispiel Pixelreihen eines Bildes, ergeben sich verschiedene Möglichkeiten der spiegelsymmetrischen Fortsetzung und somit ergeben sich auch verschiedene Versionen der diskreten Kosinustransformation (DCT):

DCT-I:
Im ersten Fall setzten wir unsere Punkte auf einen Rändern spiegelsymmetrisch bzgl. des Randes fort, d.h. zum Beispiel durch

$$y_0 = x_0, \ldots, y_{N-2} = x_{N-2}, y_{N-1} = x_{N-1}, y_N = x_{N-2}, y_{N+1} = x_{N-3}, \ldots, y_{2N-3} = x_1$$

oder in allgemeiner Schreibweise:

$$y_n = \begin{cases} x_n, & 0 \le n \le N-1 \\ x_{2N-2-n} & N \le n \le 2N-3 \end{cases}$$

Führen wir jetzt eine „normale“ diskrete Fourier-Transformation mit unsere neue Zeitreihe $y_0, \ldots, y_{2N-3}$ durch

$$\hat{y}_k = \sum_{j=0}^{2N-3} y_j \cdot e^{-2\pi i \cdot \frac{jk}{2N-2}}$$

Durch die Symmetry kommt jeder Punkt x_j für $j = 1, \dots, N-2$ doppelt vor und die Ränder x_0 und x_{N-1} einfach. Wir erhalten daher mit analogen Umformungen wie bei der stetigen Sinus- / Kosinustransformation:

$$\begin{aligned}\hat{y}_k &= x_0 e^{-2\pi i \cdot \frac{0\cdot k}{2N-2}} + x_{N-1} \underbrace{e^{-2\pi i \cdot \frac{(N-1)\cdot k}{2N-2}}}_{(e^{-\pi i})^k} + \sum_{j=1}^{N-2} x_j \cdot \left(e^{-2\pi i \cdot \frac{jk}{2N-2}} + e^{-2\pi i \cdot \frac{(2N-2-n)k}{2N-2}} \right) \\ &= x_0 + (-1)^k x_{N-1} + 2 \sum_{j=1}^{N-2} x_j \cos\left(\frac{\pi}{N-1} nk \right)\end{aligned}$$

Nach einer Normierung mit dem Faktor $\frac{1}{2}$ erhalten wir:

$$\hat{x}_k^{(I)} = \frac{1}{2}\left(x_0 + (-1)^k x_{N-1}\right) + \sum_{j=1}^{N-2} x_j \cos\left(\frac{\pi}{N-1} nk \right), \quad k = 0, \dots, N-1 \tag{I}$$

DCT-II: Statt der exakten Fortsetzung auf x_{N-1} setzen wir nun unsere Zeitreihe etwas versetzt fort, also so gesehen an der Stelle $x_{N-1/2}$. Dies hat den Vorteil, dass für unsere neuen Punkte $y_N = x_{N-1}$ und $y_{2N-1} = x_0$ gilt. Damit verschwindet der unschöne $x_0 + (-1)^k x_{N-1}$ aus (I) und wir erhalten:

$$\hat{x}_k^{(II)} = \sum_{j=0}^{N-1} x_j \cos\left(\frac{\pi}{N} \left(n + \frac{1}{2}\right) k \right), \quad k = 0, \dots, N-1 \tag{II}$$

DCT-III: DCT-III basiert auf einer Punktspiegelung, dafür erweitern wir zunächst mit einem neuen Punkt $y_N = 0$ um dann nicht spiegelsymmetrisch fortzusetzen, sondern mit $y_{N+n} = x_{N-n}$ für $n < N-1$. In der Herleitung haben wir daher x_0 einfach und den Rest doppelt. Es ergibt sich

$$\hat{x}_k^{(III)} = \frac{1}{2} x_0 + \sum_{j=1}^{N-1} x_j \cos\left(\frac{\pi}{N} n \left(k + \frac{1}{2}\right) \right), \quad k = 0, \dots, N-1 \tag{III}$$

DCT-IV: Als letztes verbinden wir die Idee von DCT-II und DCT-III indem wir spiegelsymmetrisch bei $x_{N-1/2}$ fortsetzten und bekommen

$$\hat{x}_k^{(III)} = \sum_{j=0}^{N-1} x_j \cos\left(\frac{\pi}{N} \left(n + \frac{1}{2}\right) \left(k + \frac{1}{2}\right) \right), \quad k = 0, \dots, N-1 \tag{III}$$

Satz 5.15. Die DCT-III und DCT-II Verfahren sind (bis auf Skalierung) invers zu einander. DCT-I und DCT-IV sind hingegen selbst-invers.

ohne Beweis.

Insert compare of DCT Versions.

5.5 Mehrdimensionale DCT

In der Anwendung dient DCT-II für die digitale Bildverarbeitung, wir müssen dafür jedoch eine Möglichkeit der mehrdimensionalen DCT finden. Wir nutzen hierfür die Spalten- bzw. Zeilenweise Anwendung von DCT-II und erhalten dann für $x \in \mathbb{R}^{N_1 \times N_2}$ folgende Transformation

$$\hat{x}_{k_1,k_2} = \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} x_{n_1,n_2} \cos\left(\frac{\pi}{N_1} (n_1 + \tfrac{1}{2}) k_1 \right) \cos\left(\frac{\pi}{N_2} (n_2 + \tfrac{1}{2}) k_2 \right)$$

Detaillierte Erklärung zu JPEG und 8x8 Bildern

5.6 Wavelets

Bisher haben wir uns primär mit der Fourier-Transformation und ihren diskreten Varianten (DFT oder DCT) beschäftigt und diese als Werkzeug der Spektralanalyse kennengelernt.

Die dafür verwendeten trigonometrischen Polynome ermöglichen eine einfache Trennung von nieder- und hochfrequenten Anteilen unseres Signals und verwenden dabei glatte Basisfunktionen.

Ein wesentlicher Nachteil der Verwendung von Sinusoiden als Basis liegt in der Eigenschaft der globalen Ausdehnung, also dass jede Basisfunktion über dem gesamten Definitionsbereich „relevant“ ist. Ohne einer lokalen Begrenzung gehen dadurch örtliche Informationen verloren.

Außerdem kommt es bei der Annäherung von unstetigen Funktionen durch eine Fourier-Reihe zu Über- und Unterschwingern an den Unstetigkeitsstellen:

Insert Gibbs phenomenon example for square wave.

Um diesen Nachteilen entgegen zu wirken führen wir nun das Konzept der Wavelets als eine multiskalare Basis ein. Sie erlaubt ähnlich wie die Fourier-Analyse auch eine Frequenzzerlegung gewährleistet gleichzeitig aber auch eine zeitliche Lokalisierung.

Wir wollen uns im folgenden lediglich auf die dikrete Wavelet-Transformation konzentrieren und betrachten dabei das Intervall $[0,1]$ als Definitionsbereich, d.h. wir suchen eine alternative Möglichkeit eine auf $[0,1]$ integrierbare Funktion f approximieren.

Wir beginnen damit absteigende Aufteilung des Intervall $[0,1]$ in äquidistante Teilintervalle zu definieren, für welche wir dann stückweise konstante Funktionen betrachten.

Definition 5.16. Sei $p \in \mathbb{N}$ fest. Für $k = 0, 1, \ldots, p$ ist

$$\Delta_k := \{j \cdot h_k \,|\, j = 0, \ldots, 2^k, k_k = 2^{-k}\} \subset [0,1]$$

ein Gitter von Feinheitsgrad k. Die Unterteilung Δ_{k+1} entsteht dabei durch eine Verfeinerung von Δ_k.

Das feinste Gitter ist Δ_p mit Gitterweite $h_p = 2^{-p}$ und das gröbste ist Δ_0 mit Gitterweite $h_0 = 0$.

Eine Skala von Funktionsräumen ergibt sich durch

$$V_0 \subset V_1 \subset \ldots V_p$$

wobei V_k der Raum aller Treppenfunktionen über Δ_k ist, d.h.

$$V_k := \{f : [0,1] \to \mathbb{R} \,|\, \forall j = 0, \ldots, 2^k - 1 : f \text{ konstant auf } I_{k,j} = [j \cdot h_k, (j+1) \cdot h_k]\}$$

Eine triviale Basis von V_k bilden dabei die charakteristischen Funktionen auf $I_{k,j}$ und eine Skalierung mit $2^{k/2}$ macht diese Basis zu einem Orthonormalsystem:

Satz 5.17. Die Menge $\{\chi_{k,j}\}_{j=0}^{2^k-1}$ mit $\chi = \chi_{[0,1]}$ und

$$\chi_{k,j}(x) = 2^{k/2}\chi(2^k x - j) = \begin{cases} 2^{k/2}, & x \in I_{k,j} \\ 0, & \text{sonst} \end{cases}$$

bildet eine Orthonormalbasis von V_k bzgl. dem L^2-Skalarprodukt.

Beweis.

1. Jede Funktion $f \in V_k$ lässt sich nach Definition von V_k als Treppenfunktionen über Δ_k schreiben,

d.h. sie hat die Darstellung

$$f(x) = \sum_{j=0}^{2^k-1} c_j \cdot \chi_{I_{j,k}}(x) = \sum_{j=0}^{2^k-1} 2^{-k/2} c_j \cdot \chi_{j,k}(x) \in \operatorname{span}\{\chi_{k,j} \,|\, j = 0, \dots, 2^k - 1\}$$

2. Für $j \neq l$ sind $I_{j,k}$ und $I_{l,k}$ disjunkt und es gilt:

$$\begin{aligned}
&\int_0^1 \chi_{j,k}(x) \cdot \chi_{l,k}(x)\,\mathrm{d}x \\
&= 2^k \cdot \int_0^1 \chi_{I_{j,k}}(x) \cdot \chi_{I_{l,k}}(x)\,\mathrm{d}x \\
&= 2^k \cdot \left(\int_{I_{j,k}} \underbrace{\chi_{I_{j,k}}(x)}_{1} \cdot \underbrace{\chi_{I_{l,k}}(x)}_{0}\,\mathrm{d}x + \int_{I_{j,k}} \underbrace{\chi_{I_{j,k}}(x)}_{0} \cdot \underbrace{\chi_{I_{l,k}}(x)}_{1}\,\mathrm{d}x + \int_{[0,1]\setminus(I_{j,k}\cup I_{l,j})} \underbrace{\chi_{I_{j,k}}(x)}_{0} \cdot \underbrace{\chi_{I_{l,k}}(x)}_{0}\,\mathrm{d}x \right) \\
&= 2^k \left(\left((j+1)2^{-k} - j2^{-k} \right) + \left((l+1)2^{-k} - l2^{-k} \right) + 0 \right) \\
&= 0
\end{aligned}$$

Im Fall $l = j$ ergibt sich:

$$\int_0^1 \underbrace{\chi_{j,k}(x)^2}_{0^2=0,1^2=1}\,\mathrm{d}x = 2^k \cdot \int_0^1 \chi_{I_{j,k}}(x)\,\mathrm{d}x = 2^k \cdot \left((j+1)2^{-k} - j2^{-k} \right) = 1$$

Damit bildet die Menge ein Orthonormalsystem, also auch linear unabhängig ist durch 1. eine Orthonormalbasis. □

Bemerkung 5.18. Es gilt $\chi(x) = \chi(2x) + \chi(2x-1)$ und damit folgt

$$\chi_{k,j} = \frac{1}{\sqrt{2}} \left(\chi_{k+1,2j}(x) + \chi_{k+1,2j+1}(x) \right)$$

Die Idee hinter Wavelets ist es nun für eine Festes k das Gitter Δ_k zu Δ_{k+1} zu verfeinern. Da wir wissen, dass V_k ein Unterraum von V_{k+1} ist, lässt sich die Basis $\{\chi_{k,j}\}_{j=0}^{2^k-1}$ zu einer Basis von V_{k+1} erweitern. Durch die Setzung $W_k = V_{k+1} \cap V_k^\perp$ ergibt sich

$$V_{k+1} = V_k \oplus W_k$$

und wir können durch das finden einer Basis von W_k eine Basis von V_k konstruieren. Für die Basis $\{\psi_{k,j}\}_{j=0}^{2^k-1}$ fordern wir erneut die Schreibweise in Abhängigkeit einer Grundfunktion ψ:

$$\psi_{k,j}(x) = 2^{k/2} \psi(2^k x - j) \tag{1}$$

sowie die Bedingung $\psi_{k,j} \in W_k$, d.h.

$$\psi_{k,j} \in V_{k+1} \tag{2}$$

$$\langle \psi(k,j), \chi_{k,l} \rangle_2 = 0 \qquad \text{für alle } l = 0, \dots, 2^k - 1 \tag{3}$$

Außerdem wollen wir auch hier wieder eine Orthonormalbasis, also fordern wir weiter

$$\langle \psi(k,j), \psi_{k,l} \rangle_2 = \delta_{jl} = \begin{cases} 1, & j = l \\ 0, & j \neq l \end{cases} \qquad \text{für alle } l = 0, \dots, 2^k - 1 \tag{4}$$

Lemma 5.19. Für die Funktion

$$\psi(x) := \chi(2x) - \chi(2x-1) = \begin{cases} 1, & 0 \le x \le \frac{1}{2} \\ -1, & \frac{1}{2} \le x \le 1 \\ 0, & \text{sonst} \end{cases}$$

erfüllen $\{\psi_{k,j}\}_{j=0}^{2^k-1}$ aus (1) die Bedingungen (2), (3) und (4).

Beweis.

(2) Für $j = 0, \dots, 2^k - 1$ gilt

$$\begin{aligned} \psi_{k,j} = 2^{k/2}\psi(2^k x - j) &= 2^{k/2}\Big(\chi(2^{k+1}x - j) - \chi(2^{k+1}x - 2j - 1)\Big) \\ &= \frac{1}{\sqrt{2}}\Big(\chi_{k+1,2j}(x) - \chi_{k+1,2j+1}(x)\Big) \end{aligned}$$

Da wir wissen, dass $\{\chi_{k+1,j}\}_{j=0}^{2^{k+1}-1}$ eine Basis von V_{k+1} ist, folgt $\psi_{k,j} \in V_{k+1}$.

(3) Verwenden wir die Darstellungen von $\psi_{k,j}$ und $\chi_{k,j}$ durch $\{\chi_{k+1,j}\}_{j=0}^{2^{k+1}-1}$ erhalten wir durch Satz 5.17 die geforderte Orthogonalität für $j \neq k$:

$$\begin{aligned} \langle \psi_{k,j}, \chi_{k,l} \rangle &= \left\langle \frac{1}{\sqrt{2}}\Big(\chi_{k+1,2j}(x) - \chi_{k+1,2j+1}(x)\Big), \frac{1}{\sqrt{2}}\Big(\chi_{k+1,2l}(x) + \chi_{k+1,2l+1}(x)\Big) \right\rangle_2 \\ &= \frac{1}{2}\Bigg(\underbrace{\langle \chi_{k+1,2j}(x), \chi_{k+1,2l}(x) \rangle_2}_{0} + \underbrace{\langle \chi_{k+1,2j}(x), \chi_{k+1,2l+1}(x) \rangle_2}_{0} \\ &\qquad - \underbrace{\langle \chi_{k+1,2j+1}(x), \chi_{k+1,2l}(x) \rangle_2}_{0} - \underbrace{\langle \chi_{k+1,2j+1}(x), \chi_{k+1,2l+1}(x) \rangle_2}_{0} \Bigg) \\ &= 0 \end{aligned}$$

(4) Analog erhalten wir auch, dass $\{\psi_{k,j}\}_{j=0}^{2^k-1}$ ein Orthonormalsystem bildet.

Bemerkung 5.20. Im allgemeinen werden die Funktion χ und ψ Vater- und Mutter-Wavelet genannt. Für unseren speziellen Fall heißt $\{\psi_{k,j}\}_{j=0}^{2^k-1}$ Haar-Wavelet-Basis.

Neben der Haar-Wavelet-Basis ergeben sich durch andere Wahlen von χ und ψ neue Basismengen, Beispiele hierfür sind Meyer-Wavelet, Mexikanischer Hut oder Daubechies D4-Wavelet:

Insert images of other wavelets.

Das Finden einer Basis für W_k erlaubt das konstruieren einer neuen Basis von V_{k+1}, diese nennen wir auch Zweiskalenbasis. Durch das rekursive Zerlegen von $V_k = V_{k-1} + W_{k-1}$ erhalten wir

$$V_p = V_0 \oplus W_0 \oplus W_1 \oplus \cdots \oplus W_{p-1}$$

mit der Basis

$$\{\chi_{0,0}\} \cup \{\psi_{1,0}, \psi_{1,1}\} \cup \{\psi_{2,0}, \dots, \psi_{2,3}\} \cup \cdots \cup \{\psi_{p-1,0}, \dots, \psi_{p-1,2^{p-1}-1}\}$$

wobei $\chi_{0,0}$ auf $[0,1]$ einfach nur konstant 1 ist.

Bemerkung 5.21. Wie auch schon bei der Fourier-Transformation sind wir nun in der Lage eine Bestapproximation zu f im Raum V_k zu finden, die Wavelet-Transformation. Auch hier lässt sich daraus eine diskrete Wavelet-Transformation und eine schnelle Wavelet-Transformation herleiten.