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.

1 Wiederholung

Wir starten mit einer kurzen Wiederholung zur Fixpunktiteration zum Lösen von Gleichungen der Form $Tx = x$ durch $x_{n+1} = Tx_n$.

Satz 1.1 (Banach 1922). Sei M eine abgeschlossene nichtleere Teilmenge in einem vollständig metrischem Raum (X, d). Sei $T : M \to M$ eine Selbstabbildung 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 M.
2. Konvergenz der Iteration $x_{k+1} = Tx_k$. Die Folge $(x_k)_{k\in\mathbb{N}}$ konvergiert gegen den Fixpunkt x^* für einen beliebigen Startpunkt $x_0 \in M$.
3. Fehlerabschätzung: Für alle $n = 0, 1, \dots$ gilt
   - a-priori: $d(x_n, x^*) \leq k^n(1-k)^{-1}d(x_0, x_1)$
   - a-posteriori: $d(x_{n+1}, x^*) \leq k(1-k)^{-1}d(x_n, x_{n+1})$
4. Konvergenzrate: Für alle $n \in \mathbb{N}$ gilt $d(x_{n+1}, x^*) \leq k \cdot d(x_n, x^*)$

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 $n, m \in \mathbb{N}$:

$$\begin{aligned} d(x_n, x_{n+m}) &\leq d(x_n, x_{n+1}) + d(x_{n+1}, x_{n+2}) + \ldots + d(x_{n+m-1}, x_{n+m}) \\ &\leq (k^n + k^{n+1} + \ldots + k^{n+m}) \cdot d(x_0, x_1) \\ &\leq k^n(1 + k + k^2 + \dots) \cdot d(x_0, x_1) \\ &= k^n \cdot (1-k)^{-1} d(x_0, x_1) \end{aligned}$$

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

1. Da T stetig ist (aufgrund k-Kontraktivität) folgt für die konvergente Folge (x_n), dass

$$x^* = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} Tx_n = Tx^*$$

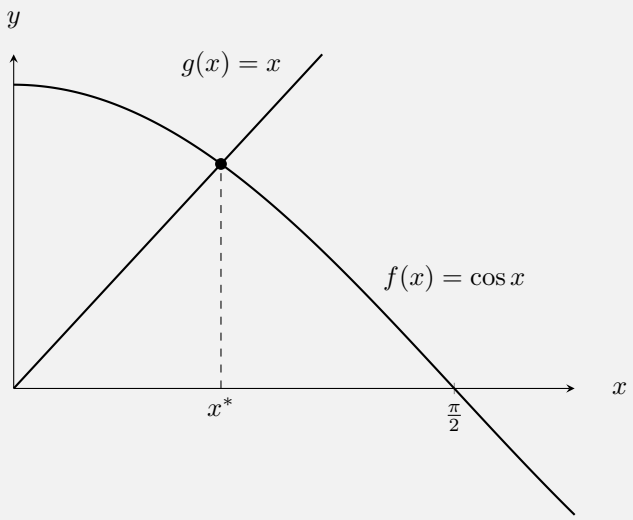
   Da M abgeschlossen ist existiert also ein Fixpunkt in M.
   Dieser ist eindeutig, denn für x, y mit $Tx = x$ und $Ty = y$ gilt $d(x, y) = d(Tx, Ty) \leq kd(x, y)$, also $d(x, y) = 0$.

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

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

4. Folgt direkt durch $d(x_{n+1}, x^*) = d(Tx_n, Tx^*) \leq k \cdot d(x_n, x^*)$

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: Nach 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 x^* folgern, diesen Fixpunkt finden wir durch die konvergente Folge $x_{k+1} = \cos x_k$.

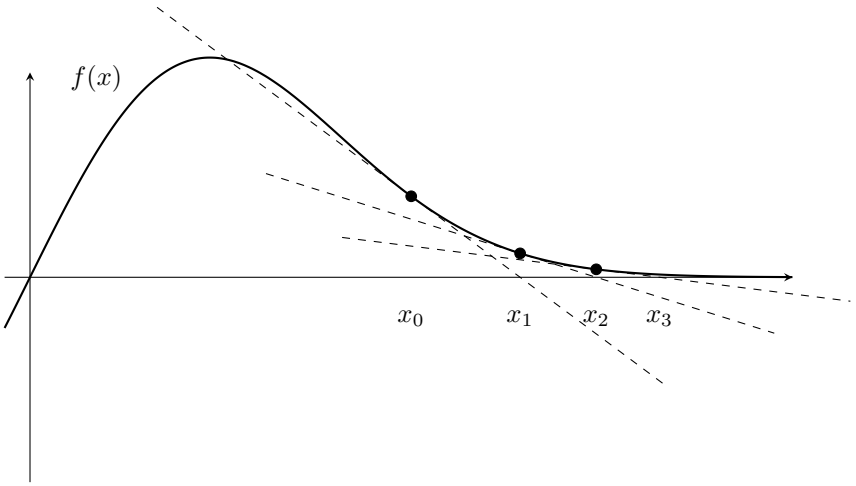
Wir betrachten im folgenden die Idee der Umwandlung eines Nullstellenproblems in Fixpunkt-Gleichung noch etwas allgemeiner. Für eine Gleichung $f(x) = 0$ mit $f : \mathbb{R} \to \mathbb{R}$ haben wir verschiedene Möglichkeiten zur Umformung:

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:

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

$\tilde{T}$ 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 (beliebiges Stopkriterium)*

Konvergenz dieses Algorithmus folgt aus Banachschen Fixpunktsatz.

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

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

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 $\|A\|_1 := \max_j \sum_i |a_{ij}|$
- Zeilensumennorm $\|A\|_\infty := \max_i \sum_j |a_{ij}|$
- 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 auch Operatornorm:

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

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

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

Operatornormen sind immer submultiplikativ und verträglich zu der Vektorrnorm, aus welcher sie abgeleitet wurden.

Satz 2.2. Ist $|||\cdot|||$ eine Norm auf $\mathbb{K}^{n\times n}$, die mit einer Vektornorm verträglich ist, und ist $|||M^{-1}N||| < 1$, dann konvergiert der Algorithmus 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

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

Da $\|T\| = \|M^{-1}N\| < 1$ ist $\tilde{T}$ eine k-kontraktive Selbstabbildung und somit konvergiert die Folge (x^k) aus dem Algorithmus gegen den eindeutigen Fixpunkt x^* mit $\tilde{T}(x^*) = x^*$.
Einsetzen der Definition von $\tilde{T}$ liefert:

$$x^* = Tx + c = M^{-1}(Nx + b) \Rightarrow Mx = Nx + b \Rightarrow Ax = (M - N)x = b$$

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

Beweis.
$\Leftarrow$: 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.
Satz 2.2 liefert dann die Konvergenz des Algorithmus.

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

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

Induktiv folgt dann $x^{(k)} - x^* = T^k(x^0 - x^*) = T^k z = \lambda^k z$, demnach gilt $\|x^{(k)} - x^*\| = |\lambda^k| \cdot \|z\|$. Für größer werdendes k kann $x^{(k)}$ also nicht gegen x^* konvergieren.

Satz 2.4. Unter gleichen Voraussetzungen des obigen Korollars gilt

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

Beweis. Aus dem Beweis von Korollar 2.3 sehen wir

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

Für jeden Startwert $x^{(0)} \in \mathbb{K}^n$ gilt nun

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

Da im $\mathbb{K}^n$ alle Normen äquivalent sind, also inbesondere auch $\|\cdot\|_\varepsilon$ und $\|\cdot\|$, exisitert eine Konstante $c_\varepsilon > 0$, so dass

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

Folglich ist

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

□Dieser Satz ermöglicht es nun einen sinnvollen Begriff der Konvergenzrate zu definieren:

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

Mittels der Zerlegung $A = D + L + R$, wobei D die Dianale, L die untere (linke) Hälfte und R die obere (rechte) Hälfte der Matrix A sind, erhalten wir einen Spezialfall der Splitting-Vefahren.
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 $k = 1, 0, \ldots$
  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 (beliebiges Stopkriterium)*

Die zugehörige Iterationsmatrix ist hierbei $J = M^{-1}N = D^{-1}(L + R)$ 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 $k = 1, 0, \ldots$
  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 (beliebiges Stopkriterium)*

Die hier erhaltene Iterationsmatrix nennen wir Einzelschrittoperator $L = (D - L)^{-1}R$ Mittels der Zeilensumennorm 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 Jordan und Gauss-Seidel Verfahren für alle Startwerte $x^{(0)} \in \mathbb{K}^n$ gegen die eindeutige Lösung von $Ax = b$.

Beweis.
Da A strikt diagonaldominant ist, muss $a_i i \neq 0$ und damit sind beide Verfahren wohldefiniert.

a) Jacobi Verfahren: Für die Iterationsmatrix gilt

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

Nach Satz 2.2 folgt damit die Konvergenz des Jacobi Verfahren.

b) Gauss-Seidel Verfahren: Um $\|L\|_\infty < 1$ zu zeigen, nutzen wir, dass die Zeilensumennorm die Operatornorm induziert durch die Maximumsnorm ist, d.h.

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

Sei nun $y = Lx$ für ein $x \in \mathbb{K}^n$ mit $\|x\|_\infty = 1$.
Induktiv folgt nun $y_i \leq q < 1$, der Induktionsanfang folgt dabei aus dem Beweisteil a).
Unter der Induktionsvoraussetzung gilt für $j < i$, dass $|y_j| \leq q$ und damit:

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

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

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 $x^* = (1, -1, -1)^T$.
Durch die Wahl $x^{(0)} = (1, 1, 1)^T$ erhalten wir beim Jacobi Verfahren:

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

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

$\vdots$

2.2 Gradientenverfahren

2.2.1 Gradientenverfahren für Optimierung

Eine Funtion $f : \mathbb{K}^n \to \mathbb{K}$ soll minimiert werden. Von einem Startpunkt $x^{(0)}$ ausgehen bewegen wir uns nun Stück für Stück in Richtung des steilsten Abstiegs, intuitiv sollten wir so ein Minimum finden.
Als Iterationsvorschrift ergibt sich $x^{(k+1)} = x^{(k)} + \alpha^{(k)} \cdot d^{(k)}, \quad k = 0, 1, \ldots$
dabei ist $\alpha^{(k)} > 0$ die Schrittweite und Abstriegsrichtung $d^{(k)} \in \mathbb{K}^n$. (Eine typische Wahl der Abstriegsrichtung ist $d^{(k)} = -\partial f / \partial x(x^{(k)}) = -\nabla f(x^{(k)})$)

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 ein 1-dim. Optimierungsproblem für die Schrittweite $\alpha^{(k)}$:

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

Ein Nachteil des Verfahrens ist die mögliche Entstehung oszillierender Pfade („Zick-Zack-Verhalten“) aufgrund unvorteilhafter Richtungen.

2.2.2 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, d.h. $a_{ij} = \overline{a_{ji}}$ (hieraus folgt inbesondere, dass die Hauptdiagonale reell ist). Zur Lösung wird hierbei die Minimierung des quaratischen Funktionals

$$\phi(x) = \tfrac{1}{2} x^* A x - x^* b$$

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

$$\begin{aligned} \phi(x) - \phi(\hat{x}) &= \tfrac{1}{2} x^* A x - x^* b - (\tfrac{1}{2} \hat{x}^* A \hat{x} - \hat{x}^* b) \\ &\vdots \\ &= \tfrac{1}{2} (x - \hat{x})^* A (x - \hat{x}) \geq 0 \end{aligned}$$

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 $\|x\|_A = \sqrt{x^* A x}, x \in \mathbb{K}^{n \times n}$ eine Norm in $\mathbb{K}^n$ definiert, die sogenannte Energienorm. Zur Energienorm gehört ein inneres Produkt, nämlich $\langle x, y \rangle_A = x^* A y, x, y \in \mathbb{K}^n$. Mithilfe dieser Definition und obiger Erkentniss ergibt sich die Abweichung des Funktionals von seinem Minimum:

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

geometrische Interpretation: Der Graph von ϕ bezüglich der Energienorm ist ein kreisförmiger Paraboid, 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 zukzessiv 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. Es gilt

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

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

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

Weiter ergibt sich die Suchrichtung $d^{(k+1)}$:

$$d^{(k+1)} = r^{(k+1)} + \beta^{(k)} d^{(k)}, \quad \langle d^{(k+1)}, d^{(k)} \rangle_A = 0 \tag{6}$$

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

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 Definition verläuft die Suchrichtung 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.

2.2.3 Eigenschaften des CG-Verfahrens

Wegen der zusätzlichen 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).

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

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

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

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

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

$$r^{(k+1)} = r^{(k)} - \alpha^{(k)} A d^{(k)} \tag{8}$$

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

$$\begin{aligned} {r^{(k+1)}}^* d^{(k)} &= (r^{(k)} - \alpha^{(k)} A d^{(k)})^* d^{(k)} \\ &= {r^{(k)}}^* d^{(k)} - \alpha^{(k)} {d^{(k)}}^* \underbrace{A^*}_{=A} d^{(k)} \\ &\overset{(5)}{=} 0 \end{aligned} \tag{9}$$

Weiter gilt nach Induktion über m:
Induktionsanfang: $m = 1$. Setzung von $k = 0$ in (9) entspricht der Behauptung (a) und nach Start $d^{(0)} = r(0)$ auch die Behauptung (b). (c) folgt im Fall $m = 1$ direkt aus (6).
Induktionsschritt: $m \to m + 1$. Wir nehmen an, dass die Aussagen (a), (b) und (c) für $\overline{m} < m$ richtig sind und zeigen damit die Gültigkeit für $m + 1$.
Zunächt folgt aus (9) mit $k = m$, dass ${r^{(m+1)}}^* d^{(m)} = 0$, sowie aus (6) mit der Induktionsannahme (a und c):

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

Dies zeigt (a) gilt auch für $m + 1$.
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):

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

Damit (c) gilt muss noch $\alpha^{(j)} \neq 0$ sein, denn dann ergibt (8):

$$\langle d^{(m+1)}, d^{(j+1)} \rangle_A = {d^{(m+1)}}^* A d^{(j)} = \frac{1}{\alpha^j} \cdot \left({d^{(j)}}^* r^{(k)} - {d^{(j)}}^* r^{(k+1)} \right) = 0$$

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

$$0 = {r^{(j)}}^* \left(r^{(j)} + \beta^{j-1} d^{(j-1)} \right) = {r^{(j)}}^* r^{(j)} + \beta^{(j-1)} {r^{(j)}}^* d^{(j-1)}$$

Nach Induktionsannahme ist aber $r^{(j)}d^{(j-1)} = 0$, was $\|r^{(j)}\|_2^2 = 0$ und somit $r^{(j)} = 0$ implizieren würde, dann wäre aber $x^{(j)} = \hat{x}$ (Widerspruch). □

Das Lemma sagt inbesondere aus, dass alle Suchrichtungen paarweise A-konjugiert alle Residuen linear unabhängig sind. Es muss sich daher nach spätestens n (Dimension) 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 Rundingsfehlern verloren gehen.

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

$$\mathcal{K}_k(A, y) = \text{span}\{y, Ay, \ldots, A^{k-1}y\}$$

Krylow-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 $d^{(j)} \in \text{span}\{r^{(0)}, \ldots, r^{(j)}\}$ für $j = 0, \ldots, k+1$ (11):
Induktionsanfang: $j = 0$. Wegen $d^{(0)} = r^{(0)}$ offensichtlich erfüllt.
Induktionsschritt: $j \to j+1$. Folgt direkt aus (6).
Es folgt damit $\text{span}\{d^{(0)}, \ldots, r^{(k-1)}\} \subset \text{span}\{r^{(0)}, \ldots, r^{(k-1)}\}$ Zusammen mit dem Lemma 2.9 folgt dass die beiden Systeme linear unabhängig sind, also gilt Gleichheit:

$$\text{span}\{d^{(0)}, \ldots, r^{(k-1)}\} = \text{span}\{r^{(0)}, \ldots, r^{(k-1)}\} \tag{12}$$

Aus (3) folgt damit:

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

Im nächsten Schritt wird induktiv gezeigt, dass $r^{(j)} \in \mathcal{K}_j(A, r^{(0)})$:
Induktionsanfang: $j = 0$. offensichtlich gilt $r^{(0)} \in \text{span}\{r^{(0)}\}$.
Induktionsschritt: $j-1 \to j$. Aus (11) und der Induktionsannahme folgt

$$\begin{aligned} & d^{(j-1)} \in \text{span}\{r^{(0)}, \ldots, r^{(j-1)}\} \subset \text{span}\{r^{(0)}, \ldots, A^{j-1}r^{(0)}\} \\ \overset{8}{\Rightarrow} \quad & r^{(j)} = r^{(j-1)} - \alpha^{(j-1)}Ad^{(j-1)} \in \text{span}\{r^{(0)}, \ldots, A^{j}r^{(0)}\} \end{aligned}$$

Damit folgt $\text{span}\{r^{(0)}, \ldots, r^{(k-1)}\} \subset \mathcal{K}_j(A, r^{(0)})$. Die Vektoren $r^{(j)}$ sind linear unabhängig und daher hat der linke Unterraum die Dimension k, es folgt Gleichheit (13) und damit auch $x^{(k)} \in x^{(0)} + \mathcal{K}_k(A, r^{(0)})$.

b) 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 nach (3):

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

Und für ein beliebiges $x \in x^{(0)} + \mathcal{K}_k(A, r^{(0)})$ gilt wegen (13)

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

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

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

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

2.2.4 Praktische Aspekte der Implementierung

Bemerkung 2.13. Für eien 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:

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

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

Diese Gleichung (5') folgt aus Lemma 2.9 a) und b), nach welchen

$${r^{(k)}}^* d^{(k)} = {r^{(k)}}^* r^{(k)} + \beta^{(k)} \cdot {r^{(k)}}^* d^{(k-1)} = {r^{(k)}}^* r^{(k)}.$$

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

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

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.

Wähle $x^{(0)} \in \mathbb{K}^n$ beliebig
$r^{(0)} \leftarrow b - Ax^{(0)}$
$d^{(0)} \leftarrow r^{(0)}$
for $k = 0, 1, \ldots,$
  $\alpha^{(k)} \leftarrow \dfrac{\|r^{(k)}\|_2^2}{{d^{(k)}}^* A d^{(k)}}$
  $x^{(k+1)} \leftarrow x^{(k)} + \alpha^{(k)} d^{(k)}$
  $r^{(k+1)} \leftarrow r^{(k)} - \alpha^{(k)} A d^{(k)}$
  $\beta^{(k)} \leftarrow \dfrac{\|r^{(k+1)}\|_2^2}{\|r^{(k)}\|_2^2}$
  $d^{(k+1)} \leftarrow r^{(k+1)} + \beta^{(k)} d^{(k)}$
until *stop (beliebiges Stopkriterium)*

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 Matrixzen ist derzeit nicht bekannt. Ein komplizierteres Verfahren mit ähnlicher Konvergenzgeschwindigkeit ist das GMRES-Verfahren.

2.3 Präkonditionierung des CG-Verfahren

Definition 2.15. $\kappa_M(A) = \text{cond}_M(A) = \|A^{-1}\|_M \cdot \|A\|_M$ 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 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:

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

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

2.3.1 Präkonditionierung mittels Cholesky

Idee: Gleichungssystem $Ax = b$ in ein äquivalentes LGS umwandeln, sodass die Kondition sich verbessert:
$M^{-1}Ax = M^{-1}b$ ($\triangle$), wobei M hermitisch und positiv definit ist.
Problem: Die Matrix $M^{-1}A$ muss nicht notwendig hermitisch sein, daher nutzen wir die Cholesky-Zerlegung $M = CC^*$ un derhalten[1]:

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

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

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

Wir können also CG-Verfahren zum Lösen von $L^{-1}AL^{-*}z = L^{-1}b$ nutzen.

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] $L^{-*} = (L^*)^{-1}$

Bemerkung 2.16. Die Faktorisierung $M = LL^*$ muss nicht explizit berechnet werden, da die Variable z wieder durch x subsituiert werden kann. Man benötigt für das CG-Verfahren die Berechnung der Koeffizienten $\beta^{(k)}$, die Norm $\|L^{-1}b - L^{-1}AL^{-*}z^{(k)}\|$, $r^{(k)} = b - Ax^{(k)}$ und den Hilfsvektor (Residuum der Präkonditionierten Form ($\triangle$)) $s^{(k)} = M^{-1}r^{(k)}$.
Es gilt

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

2.3.2 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 ($\triangle$).

1 Wähle $x^{(0)} \in \mathbb{K}^n$ beliebig
2 $r^{(0)} \leftarrow b - Ax^{(0)}$
3 Löse $Ms^{(0)} = r^{(0)}$
4 $d^{(0)} \leftarrow s^{(0)}$
5 **for** $k = 0, 1, \ldots,$
6 $\quad \alpha^{(k)} \leftarrow \dfrac{{r^{(k)}}^* s^{(k)}}{{d^{(k)}}^* A d^{(k)}}$
7 $\quad x^{(k+1)} \leftarrow x^{(k)} + \alpha^{(k)} d^{(k)}$
8 $\quad r^{(k+1)} \leftarrow r^{(k)} - \alpha^{(k)} A d^{(k)}$
9 $\quad$ Löse $Ms^{(k+1)} = r^{(k+1)}$
10 $\quad \beta^{(k)} \leftarrow \dfrac{{r^{(k+1)}}^* s^{(k+1)}}{{r^{(k)}}^* s^{(k)}}$
11 $\quad d^{(k+1)} \leftarrow s^{(k+1)} + \beta^{(k)} d^{(k)}$
12 **until** *stop (beliebiges Stopkriterium)*

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 typscherweise 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 verschobenene Krylow-Raum $x^{(0)} + \mathcal{K}_k(M^{-1}A, M^{-1}r^{(0)})$ und ist in dieser Menge die eindeuig bestimmte Minimalstelle des Fuktionals $\phi(x) = \frac{1}{2}x^*Ax - x^*b$.

Beweis. Vgl. Beweis zu 2.12

Nach diesem Satz leigt die entsprechende Iterierte $z^{(k)} = L^*x^{(k)}$ in dem affin verschobenen Kylow-Raum $z^{(0)} + \mathcal{K}_k(L^{-1}AL^{-*}, L^{-1}b - L^{-1}AL^{-1}z^{(0)})$ mit $z^{(0)} = L^*x^{(0)}$ und minimiert in dieser Menge das Fehlerfunktional $\psi(z) = \frac{1}{2}L^{-1}AL^{-*}z - z^*L^{-1}b$.
Durch die Transformierte $x = L^{-*}z$ werden die Iterierten und die genannten Krylow-Räume aufeinander abgebildet und es gilt $\psi(z) = \phi(x)$.

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

2.4 Anwendung und Konvergenzgeschwindigkeit des CG-Verfahren

2.4.1 Lösen von Randwertproblemen mittels CG-Verfahren

Bevor wir mit einer Anwendung der neuen Methodik starten, wiederholen wir kurz einen wichtigen Satz der Analysis:

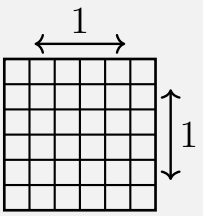
Satz 2.19 (Satz von Taylor). Sei $f : [a, b] \to \mathbb{R}$ eine $(n+1)$-mal stetig differenzierbare Funktion und $x, x_0 \in [a, b]$. Dann existiert ein ξ zwischen x und x_0, so dass

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

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 $n = m^2$.
Die Ableitung von f an der Stelle x_0 sei definiert durch

$$f'(x_0) := \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{(x_0 + \Delta x) - x_0} = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

durch Linearisierung (Tangentengleichung) bzw. für genauere Approximationen Taylorformel ergibt sich:

$$f_L(x) = f(x_0) + f'(x_0)(x - x_0)$$
$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \tfrac{1}{2} f''(x_0)(x - x_0)^2 + \ldots + R_n(x, x_0)$$

und damit

$$f(x + \Delta x) = f(x_0) + f'(x_0)\Delta x + \tfrac{1}{2} f''(x_0)(\Delta x)^2 + R \tag{1}$$
$$f(x - \Delta x) = f(x_0) - f'(x_0)\Delta x + \tfrac{1}{2} f''(x_0)(\Delta x)^2 + \tilde{R} \tag{2}$$

Mit dieser Approximation ergibt sich für die Differenzquotienten:

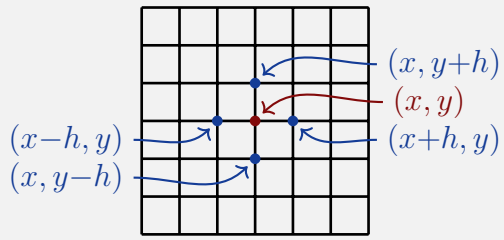
$$\frac{\mathrm{d}^+}{\mathrm{d}x} f(x) \mid_{x=x_0} = \frac{1}{\Delta x} \left(f(x_0 + \Delta x) - f(x_0) \right) \qquad \text{(rechtsseitiger DQ)}$$

$$\frac{\mathrm{d}^-}{\mathrm{d}x} f(x) \mid_{x=x_0} = \frac{1}{\Delta x} \left(f(x_0) - f(x_0 + \Delta x) \right) \qquad \text{(linksseitiger DQ)}$$

$$\frac{\mathrm{d}}{\mathrm{d}x} f(x) \mid_{x=x_0} = \frac{1}{2\Delta x} \left(f(x_0 + \Delta x) - f(x_0 - \Delta x) \right) \qquad \text{(zentraler DQ)}$$

Der zentrale Differenzquotienten approximiert dabei mit einer Ordnung häher als der links- und rechtsseitige Differenzquotient, da quadratische Terme in (1) und (2) sich gegenseitig wegkürzen.

Wir nutzen dies nun um die Laplace-Operator in 2 Dimensionen zu approximieren:



Für innnere Punkte in unserem Quadratgitter gilt die sogenannte „5-Punktregel“:

$$-h^{-2} \left(u(x+h, y) - 2u(x, y) + u(x-h, y) + u(x, y+h) - 2u(x, y) + u(x, y-h) \right) = f(x, y)$$

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

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

2.4.2 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

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

Zur Reduktion des Anfangsfehlers um den Faktor ε sind circa

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

Iterationsschritte erforderlich.

Für den Beweis des Satzes benötigen wir noch einen Hilfssatz:

Hilfssatz 2.22 (polynomiale Normschranke). Für ein Polynom $p \in P_k := \mathbb{R}_k[X]$ mit $p(0) = 1$, gelte auf einer Menge $S \subset \mathbb{R}$, welche alle Eigenwerte von A enthält, $\sup_{\mu \in S} |p(\mu)| \leq M$. Dann gilt

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

Beweis des Hilfssatz. Unter Beachtung der Beziehung

$$\|x^{(k)} - x\|_A = \min\{\|y - x\|_A : y \in x^{(0)} + \mathcal{K}_k(A, r^{(0)})\}$$

erhalten wir

$$\|x^{(k)} - x\|_A = \min_{p \in P_{k-1}} \|x^{(0)} - x + p(A)r^{(0)}\|_A$$

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

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

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:

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

und folglich

$$\|p(A)y\|_A^2 = \sum_{j=1}^{n} \lambda_j p(\lambda_j)^2 \gamma_j^2 \leq M^2 \sum_{j=1}^{n} \lambda_j \gamma_j^2 = M^2 \|y\|_A^2$$

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 von Satz 2.21*Durch Verwendung des Hilfssatz mit $S := [\lambda, \Lambda]$, wobei λ den kleinsten un dΛ den größten Eigenwert von A beschreibt, folgt:

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

Dies ergibt die Behauptung wenn wir noch zeigen können, dass

$$\sup_{\lambda\leq\mu\leq\Lambda}|p(\mu)|\leq 2\cdot\frac{(1-\sqrt{\frac{\lambda}{\Lambda}})^k}{(1+\sqrt{\frac{\lambda}{\Lambda}})^k}$$

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

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

Aus der Darstellungen

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

für die Chebyshev-Polynome 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 wird $\sup_{\lambda\leq\mu\leq\Lambda}\overline{p}(\mu)\leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}-1}\right)^k$, was die erste Ungleichung des Satzes zeigt.
Für den zweiten Teil betrachten wir die Anzahl der Schritte, dass

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

Wegen der 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 genau dann erfüllt wenn

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

□

3 Eigenwertprobleme

3.1 Einleitung

Aus der linearen Algebra ist 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 $(A - \lambda I)v = 0$ (*). Hierbei muss $A - \lambda I$ singulär sein, sonst ist die einedeuige Lösung des Systems gegeben durch $v = 0$.
Per Hand würden wir hier nun das charakteristische Poylnom $\chi_A(\lambda) = \det(A - \lambda I)$ aufstellen und dessen Nullstellen bestimmen, da dies genau die Werte für λ sind, für welche das obige System nichttriviale Lösungen hat.
Für die nuerische Berechung 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-Bestimmung: Eigenvektor über LGS (*).

b) Eigenvektor-Bestimmung: 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 $\|\cdot\|$ eine natürliche Matrixnorm. Dann gilt für jeden Eigenwert λ von A, welcher nicht zugleich auch Eigenwert von B ist, die Beziehung

$$\|(\lambda I - B)^{-1}(A - B)\| \geq 1$$

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

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

Demnach ist also

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

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 $B = diag(a_{jj})$ in dem Hilfssatz 3.1 und nehmen $\|\cdot\|_\infty$ als natürliche Matrixnorm. Für $\lambda \neq a_{jj}$ folgt dann

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

d.h. λ liegt in einem der Gerschgorin-Kreise.
Für den zweiten Teil sei o.B.d.A. $I = \{1, \dots, m\}$.
Setzen wir $A_t = D + t(A - D)$, dann liegen genau m Eigenwerte von $A_0 = D$ in U und $n - m$ Eigenwerte in V. Das selbe folgt auch für $A_1 = A$, da die Eigenwerte von A_t stetige Funktionen in t sind. □

Ein Alternativer Beweis zur ersten Behauptung liefert eine Betrachtung des Eigenwertproblems $Ax = \lambda x$ mit $x \neq 0$. Offensichtlich existiert ein x_i mit $|x_j| \leq |x_i|$ für alle $j \neq i$. Die i-te Komponente von Ax ist gegeben durch

$$\lambda x_i = (Ax)_i = \sum_{j=1}^{m} a_{jj} x_j$$

Somit folgt

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

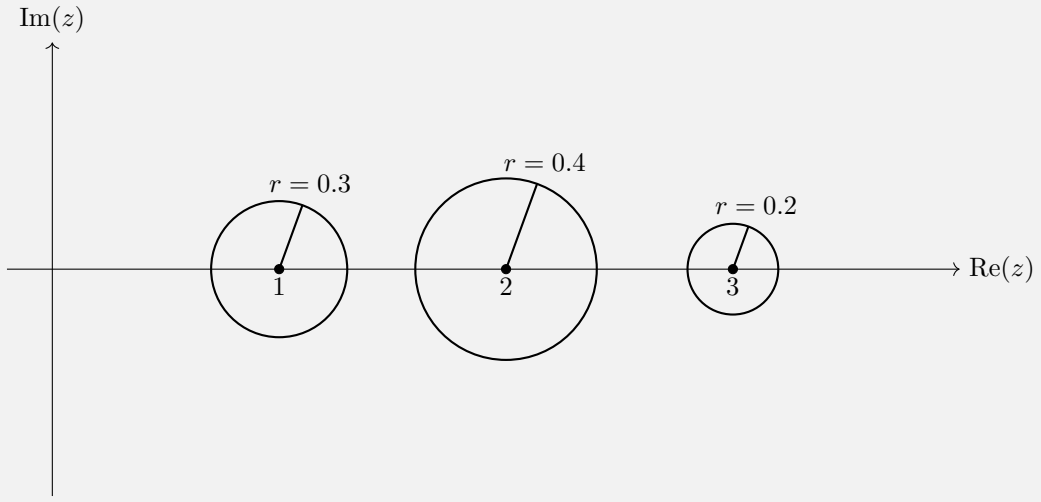
Demnach liegt $\lambda \in K_i$. □

Beispiel 3.3. 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 Gerschgori-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 3.4 (Stabilitätssatz). Sei $A \in \mathbb{K}^{n\times n}$ eine Matrix, zu der es n linear unabhängige Eigenvektoren gibt $\{w^{(1)}, \dots, w^{(n)}\}$ und sei $B \in \mathbb{K}^{n\times n}$ eine zweite Matrix. Dann gibt es zu jedem Eigenwert$\lambda(B)$ von B einen Eigenwert $\lambda(A)$ von A, sodass mit der Matrix $W = (w^{(1)}|\dots|w^{(n)})$ gilt

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

Bewies. Die Eigenwertgleichungen $Aw^{(i)} = \lambda_i(A)w^{(i)}$ lassen sich in der Form $AW = W \cdot \text{diag}(\lambda_i(A))$ schreiben, d.h. $A = W \cdot \text{diag}(\lambda_i(A)) \cdot W^{-1}$ ist ähnlich zu der Diagonalmatrix $\Lambda = \text{diag}(\lambda_i(A))$. Wenn nun $\lambda = \lambda(B)$ kein Eigenwert von A ist, so gilt

$$\|(\lambda I - A)^{-1}\|_2 = \|W(\lambda I - \Lambda)^{-1}W^{-1}\|_2 \leq \|W\|_2 \cdot \|W^{-1}\|_2 \cdot \|(\lambda I - \Lambda)^{-1}\| = \text{cond}_2(W) \cdot \max_{i=1,\dots,n} |\lambda - \lambda_i(A)|^{-1}$$

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 unität angenommen werden kann, d.h. $ww^{-*} = I$. In diesem Fall gilt $\text{cond}_2(W) = \|W^{-*}\|_2 \cdot \|W\|_2 = 1$.

Regel: Allgemein kann man sagen, dass das Eigenwertproblem für hermtische Matrizen gut konditioniert ist, während das allgemeine Eigenwertproblem je nach Größe von $\text{cond}_2(W)$ 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 3.5. Die Potenzmethode (Von-Mises-Iteration) erzeugt ausgehend von einem Startvektor $z^{(0)} \in \mathbb{C}^n$ mit $\|z^0\| = 1$ eine Folge von Iterationen $z^{(t)} \in \mathbb{C}^n, t = 1, 2, \dots$ durch

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

Für einen beliebigen Index $k \in \{1, \dots, n\}$, (z.B. maximale Komponente von z^k) wird gesetzt:

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

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

Zur Analyse des Verfahrens nehmen wir an, dass die Matrix A diagonalisierbar ist, d.h. ähnlich zu einer Diagonalmatrix ist. Dies ist äquivalent zu der Tatsache, dass A eine Basis von Eigenvektoren $\{w^{(1)}, \dots, w^{(n)}\}$ besitzt. Weiter seien diese Eigenvektoren $w^{(i)}$ normiert.
Wir nehmen an, dass $z^{(0)}$ eine nicht-triviale Komponente bezüglich $w^{(n)}$ besitzt. (Dies ist keine wesentliche Einschränkung, da aufgrund des unvermeidbaren Rundungsfhlers dieser Fall der Iteration sicher einmal auftritt)

Satz 3.6 (Potenz-Methode). Die Matrix A sei diagonalisierbar und ihr betragsgrößter Eigenwert sei separiert von den anderen Eigenwerten, d.h. $|\lambda_n| > |\lambda_{n-1}| \geq |\lambda_{n-2}| \geq \ldots \geq |\lambda_1|$.
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 $\|z^{(t)} - \delta_t \cdot w^{(n)}\| \to 0$ für $t \to \infty$ und es gilt

$$\lambda^{(t)} - \lambda_n = \mathcal{O}\left(\left|\frac{\lambda_{n-1}}{\lambda_n}\right|^t\right) \qquad \text{für } t \to \infty$$

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

$$z^{(t)} = \frac{\tilde{z}^{(t)}}{\|\tilde{z}^{(t-1)}\|} = \frac{Az^{(t-1)}}{\|Az^{(t-1)}\|} = \ldots = \frac{A^t z^{(0)}}{\|A^t z^{(0)}\|}$$

Dabei gilt:

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

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:

$$z^{(t)} = \frac{\lambda_n^t \alpha_n (w^{(n)} + \mathcal{O}(1))}{|\lambda_n^t \alpha_n| \cdot \|w^{(n)} + \mathcal{O}(\rho^t)\|} = \underbrace{\frac{\lambda_n^t \alpha_n}{|\lambda_n^t \alpha_n|}}_{=:\delta_k} \cdot (w^{(n)} + \mathcal{O}(\rho^t))$$

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

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

□

Die Konvergenz der Potenzmethode ist umso besser, je mehr der betragsgrößte Eigenwert λ_n von den übrigen betragsmäßig separiert ist. Der Beweis ist verallgemeinerbar für betragsgrößte Eigenwerte, welche merhfach 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, man hat 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 $i \neq k$.
Sei λ ein Eigenwert von A, dann ist λ^{-1} ein Eigenwert von A^{-1}. Wir betrachten das Eigenwertproblem, welches sich für die Matrix $A - \tilde{\lambda} I$ ergibt:

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

Also ist $\xi = \lambda_k - \tilde{\lambda}$ ein Eigenwert von $A - \tilde{\lambda}I$ und folglich ist $\mu = \frac{1}{\xi} = (\lambda_k - \tilde{\lambda})^{-1}$ ein Eigenwert von $(A - \tilde{\lambda}I)^{-1}$.
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 $i = 1, \ldots, n$ un des gilt

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

Definition 3.7. Die inverse Iteration besteht in 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 $z^{(t)}$ bestimmt als Lsg. der Gleichungssysteme

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

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 mit Hilfe der Rayleigh-Quotienten).

Aufgrund der Aussagen über Potenzmethoden liefert die inverse Iteration also für eine diagonalisierbare Matrix jeden Eigenwert, zu dem bereits eien 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 solceh von unbedeutenden

(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 Seiten, wobei ein Index k immer für eine Seite steht. Gesucht ist die Bedeutung einer Seite $x_k \in \mathbb{R}$
L_k sei die Menge der Seiten, die auf k verlinken, Links auf Seiten von sich selbst werden dabei nicht berücksichtigt.
n_k sei die Anzahl der Links, der Webite k nach außen.
Wir modellieren mittels folgendem LGS

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

Die Gleichung $x = Ax$ entspricht hierbei der Eigenwertgleichung für den Eigenwert $\lambda = 1$.
Der historische Ansatz von Google ist die Potenzmethode:

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

3.4.1 Stochastische Vektoren/Matrizen

Definition 3.8. Ein Vektor $p \in \mathbb{R}^n$ heißt stochastischer Vektor, wenn alle Elemente p_i nicht-negativ 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 ale 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 3.9. Sei $A \in \mathbb{R}^{n \times n}$ eine stoch. Matrix und $p \in \mathbb{R}^n$ ein stoch. Vektor, dann ist das Produkt $Ap \in \mathbb{R}^{n \times n}$ wieder ein stoch. Vektor.

Beweis. Es sei a_i die i-te Spalte der Matrix a_i, d.h. a_i ist ein stoch. Vektor

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

Weiter gilt offensichtlich $(Ap)_{ij} \geq 0$. □

Lemma 3.10. 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 3.11. 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 besitzten und damit die charakteristischen Polynome $\chi_A(\lambda) = \det(A - \lambda I) = \det(A^T - \lambda I) = \chi_{A^T}(\lambda)$.
Weiter ist die Summe der Elemente jedes Zeilenvektors von A^T ist gleich 1 (da A stoch.), somit ist $e = (1, \ldots, 1)^T$ ein Eigenvektor von A^T mit Eigenwert 1. Somit besitzt auch die Matrix A den Eigenwert $\lambda = 1$.
Angenommen es existiert ein Eigenvektor v zum Eigenwert λ mit $|\lambda| > 1$, dann gilt

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

Für die Länge dieses Vektors gilt $\|\lambda^n v\| = |\lambda^n| \cdot \|v\|$ ein exponentieller Wachstum in n, da $|\lambda| > 1$.
Daraus folgt, dass für große n ein Element $(A^n)_{ij}$ exisitert, welches größer als 1 ist.
Da nach Lemma 3.10 die Matrix A^n stoch. ist bildet dies einen Widerspruch. □

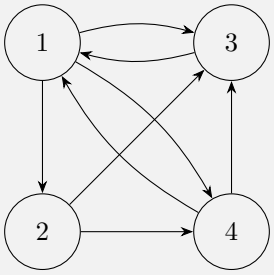
Lemma 3.12. Die Bewertungsmatrix A des Page-Rank-Algorithmus ist eine stoch. Matrix.

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 3.13. Wir betrachten ein 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

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

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.
Iterative Andwendung liefert dann alle Eigenwerte.

Sei $A \in \mathbb{R}^{n \times n}$ gegeben, der Eigenwert $\lambda^{(1)}$ sei der betragsgrößte Eigenwert, d.h.

$$|\lambda^{(1)}| > |\lambda^{(2)}| > \ldots > 0$$

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

$$B = A - \frac{1}{u_p^{(1)}} u^{(1)} \cdot a^T \quad \text{mit} \quad B_{ij} = A_{ij} - \frac{1}{u_p^{(1)}} u_i^{(1)} \underbrace{A_{pj}}_{(a^T)_j}$$

Aus dem Eigenwertproblem $Au^{(k)} = \lambda^{(k)} u^{(k)}$ ergibt sich

$$\lambda^{(k)} u_p^{(k)} = (Au^{(k)})_p = a^T \cdot u^{(k)}$$

Für $k = 1$ ergibt sich:

$$\begin{aligned} Bu^{(1)} &= Au^{(1)} - \frac{1}{u_p^{(1)}} \cdot u^{(1)} \cdot a^T \cdot u^{(1)} \\ &= Au^{(1)} - \frac{1}{u_p^{(1)}} \cdot u^{(1)} \cdot \lambda^{(1)} u_p^{(1)} \\ &= \lambda^{(1)} u^{(1)} - \lambda^{(1)} u^{(1)} = 0 \end{aligned}$$

d.h. 0 ist ein Eigenwert von B (statt vorher $\lambda^{(1)}$ von A).
Analoge Überlegung für $k = 2, \ldots, n$ liefert:

$$\begin{aligned} Bu^{(k)} &= \lambda^{(k)} u^{(k)} - \frac{1}{u_p^{(k)}} \cdot u^{(k)} \cdot \lambda^{(k)} u_p^{(k)} \\ &= \lambda^{(k)} \cdot \left(u^{(k)} - \frac{u_p^{(k)}}{u_p^{(k)}} \cdot u^{(k)} \right) \end{aligned} \tag{1}$$

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

$$\begin{aligned} Bu^{(k)} + 0 &= Bu^{(k)} + \underbrace{Bu^{(1)}}_{=0} \\ &= B^{(k)} + B \cdot \frac{-u_p^k}{u_p^{(1)}} u^{(1)} \\ &= B \cdot \left(u^{(k)} - \frac{u_p^{(k)}}{u_p^{(k)}} \cdot u^{(k)} \right) \overset{(1)}{=} \lambda^{(k)} \cdot \left(u^{(k)} - \frac{u_p^{(k)}}{u_p^{(k)}} \cdot u^{(k)} \right) \end{aligned} \tag{2}$$

Die Gleichung (2) zeigt für $k = 2, 3, \ldots, n$, dass $\lambda^{(k)}$ auch ein Eigenwert zu B ist, wenn auch mit anderem Eigenvektor.

Satz 3.14 (Deflation nach Wielandt). Seien $\lambda_1, \ldots, \lambda_n$ die Eigenwerte von A, betragsmäßig fallend, d.h. $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$, und zugehörigen Eigenvektoren $u^{(1)}, \ldots, u^{(n)}$ Eigenvektoren. Dann besitzt die Matrix

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

die Eigenwerte $\lambda_1, \ldots, \lambda_n$ mit den zugehörigen Eigenvektoren $u^{(1)}, w^{(2)}, \ldots, w^{(n)}$, wobei

$$w^{(k)} = u^{(k)} - \frac{u_p^{(k)}}{u_p^{(1)}} \cdot u^{(1)} \tag{*}$$

Den betragsmäßig zweiten Eigenwert λ_2 zugehörigen Eigenvektor $w^{(2)}$ erhält man somit mit der Potenzmethode für die Matrix B nach ihrer Definition.
Der Eigenvektor $u^{(2)}$ zum Eigenwert λ_2 der Matrix A kann wie folgt rekonstruiert werden:

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

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 3.15. 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:

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

Für die Deflation wählen wir nun $p = 1$ mit $u_p^{(1)} \neq 0$ und $a^T = (-4, -5, -1)$, 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 zweigrößten Eigenwert und den zugehörigen Eigenvektor von B:

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

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

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

Nach beliebiger Methodik aus Satz 3.14 lassen sich nun aus $w^{(2)}$ und $v^{(3)}$ die Eigenvektoren $u^{(2)}$ und $u^{(3)}$ 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 $A \in \mathbb{C}^{n\times n}$, wobei A eine sehr große Matrix, typischerweise $n \geq 10^4$, ist.

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 Unterrraum von $\mathbb{C}^n$ kleiner Dimension, d.h. $\dim K_m = m \ll n$, dann wird das n-dimensionale Eigenwertproblem (1*) mit dem m-dimensionale Eigenwertproblem approximiert:

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

Statt alle $y \in K_m$ zu betrachten, reicht es auch, wenn wir nur die erzeugenden q^i benutzen. 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 Galverkin-System

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

Wir schreiben dieses System typischerweise in kompakter Form als Eigenwertproblem $\mathcal{A}\alpha = \lambda\mathcal{M}\alpha$ mit Vektoren $\alpha = (\alpha_1, \ldots, \alpha_m)$ und Matrix $\mathcal{A} = (\langle Aq^j, q^i\rangle_2)_{i,j=1}^m$, $\mathcal{M} = (\langle q^j, q^i\rangle_2)_{i,j=1}^m$.

Im folgenden betrachten wir immer die *kartesischer Representation* der Basisvektoren $q^i = (q^i_j)_{j=1}^n$ und somit schreibt man das Galerkin-EW-Problem in der Form[2]

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

Mit $\mathcal{Q}^{(m)} = (q \cdot \ldots, q^m) \in \mathbb{C}^{n\times m}$ und dem Vektor $\alpha = (\alpha_j)_{j=1}^m \in \mathbb{C}^m$ kann dies in der kompakten Form

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

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

formuliert werden.

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

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

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

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

Übersicht (Krylov-Methode)

1. Wähle geeignete Unterräume $K_m \in \mathbb{C}^{m \times m}$, $m \ll n$ Krylov-Raum durch Verwendung der Matrix A und deren Potenz.
2. Konstruiere eine ONB $\{q^1, \dots, q^m\}$ 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 $H^{(m)} := (\mathcal{Q}^{(m)})^* A \mathcal{Q}^{(m)}$, welche konstruktionsbedingt eine Hessenberg-Matrix oder im hermitischen Fall hermitische Tridiagonalmatrix ist.
4. Löse das Eigenwertproblem der reduzierten Matrix $H^{(m)} \in \mathbb{C}^{m \times m}$ durch die QR-Methode.
5. Nehme die Eigenwerte von $H^{(m)}$ als die Näherung der dominanten (betragsgrößten) Eigenwerte von A. Im Falle des kleinstgrößten Eigenwert, muss die Matrix A^{-1} betrachtet werden (Konstruktion der Unterräume K_m kann sehr aufwendig sein).

4.2 Arnoldi-Methode

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 $\{q, Aq, A^2 q, \dots, A^{m-1} q\}$.

Idee: Verwendung dieser Informationen und Erstellen einer sogenannten *Krylov-Matrix*

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

Die Spalten dieser Matrix sind nicht orthogonal zueinander. $A^t q$ konvergiert gegen den Eigenvektor zum betragsgrößten Eigenwert, d.h. K_m ist schlecht konditionert (Was heißt schlecht konditionert bei nicht quadratisch?) (da sich die letzten Spalten kaum ändern).
Die Konstruktion in eine orthogonale Basis mit dem Gram-Schmidt-Algorithmus ist instabil.
Wir wählen 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.

Definition 4.1. Wir definieren für das Folgende orthogonalen Projektionsoperator:

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

Dieser projeziert den Vektor v auf $\text{span}\{u\}$.

Damit ergibt sich das klassische *Gram-Schmidt-Orthogonalisierungs-Verfahren* als

$$q^1 = \frac{q}{\|q\|_2}, \quad \text{und} \quad \tilde{q}^t = A^{t-1}q - \sum_{j=1}^{t-1} \text{proj}_{q^j}(A^{t-1}q), \qquad q^t = \frac{\tilde{q}^t}{\|\tilde{q}^t\|_2} \quad \text{für } t = 2, \dots, m$$

Der t-te Schritt projeziert die Komponente von $A^{t-1}q$ in Richtung der bereits bestimmten Orthogonal-Vektoren $\{q^1, \dots, q^{t-1}\}$.
Dies ist numerisch instabil durch Summieren der Rundungsfehler.

Wir betrachten daher *das modifizierte Gram-Schmidt-Verfahren*, wobei der t-te Schritt projeziert die Komponenten von Aq^t in Richtung $\{q^1, \dots, q^{t-1}\}$:

$$q^1 = \frac{q}{\|q\|_2}, \quad \text{und} \quad \tilde{q}^t = Aq^{t-1} - \sum_{j=1}^{t-1} \text{proj}_{q^j}(Aq^{t-1}), \qquad q^t = \frac{\tilde{q}^t}{\|\tilde{q}^t\|_2} \quad \text{für } t = 2, \dots, m \tag{1}$$

Da q^t, $\tilde{q}^t$ in die gleiche Richtung zeigen und $\tilde{q}^t \perp K_t$ erhält man

$$\langle q^t, \tilde{q}^t\rangle_2 = \|\tilde{q}^t\|_2 = \left\langle q^t, Aq^{t-1} - \sum_{j=1}^{t-1} \text{proj}_{q^j}(Aq^{t-1}) \right\rangle_2 = \langle q^t, Aq^{t-1}\rangle_2$$

Mit $h_{i,t-1} := \langle Aq^{t-1}, q^i\rangle_2$ ergibt sich mit dem modifizierte Gram-Schmidt-Algorithmus

$$Aq^{t-1} = \sum_{i=1}^{t} h_{i,t-1} q^i, \qquad t = 2, \dots, m+1$$

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

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

Man erhält das gleiche Resultat, wie beim klassischen Gram-Schmidt-Verfahren, aber mit kleinerem numerischen Fehler.

Definition 4.2 (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 $1 \le t \le m \ll n$ (Arnoldi-Basis), durch Anwendung der modifizierten Gram-Schmidt-Methode (2) auf die Basis $\{q, Aq, A^{m-1}q\}$ des Krylov-Unterraums K_m.

(Algobox)
Startvektor: $q^1 = \|q\|_2^{-1} q$
Iteriere für $2 \le t \le m : q^{t,1} = Aq^{t-1}$
und für $j = 1, \dots, t-1 : h_{j,t} = \langle q^{t,j}, q^j\rangle_2$ und $q^{t,j+1=q^{t,j}-h_{j,t}q^j}$ und $h_{t,t} = \|q^{t,t}\|_2$ und $q^t = h_{t,t}^{-1} \cdot q^{t,t}$

Bezeichne die $n \times m$-Matrix aus den ersten Arnoldi-Vektoren $\{q^1, q^2, \ldots, q^m\}$ mit

$$\mathcal{Q}^{(m)} := [q^1, q^2, \ldots, q^m]$$

und sei $H^{(m)}$ die obere Hessenberg Matrix $(m \times m)$ aus $h_{j,k}$:

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

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

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

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

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

ergibt

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

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 $h_{m,m+1}$ dann ist der Krylov-Raum K_m ein invarianter Unterraum der Matrix A und die reduzierte Matrix $H^{(m)} = {\mathcal{Q}^{(m)}}^* A \mathcal{Q}^{(m)}$ hat m Eigenwerte gemeinsam mit A, d.h. $\sigma(H^{(m)}) \subset \sigma(A)$[3]
Das folgende Lemma liefert a posteriori Abschätzungen der Genauigkeit für die Approximation der Eigenwerte von A durch $H^{(m)}$.

Lemma 4.3. 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

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

wobei w_m die letzte Komponente des Eigenvektors w ist.

Beweis. Multiplikation von (3) mit w ergibt

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

Daraus folgt mit $\|q_{m+1}\|_2 = 1$, 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 a posteriori-Prüfung basierend auf den berechneten Größen $h_{m+q,m}$ und w_m, ob das erhaltene Paar $\{\mu, w\}$ eine gute Approximation ist.

[3] Beweis: Übungsblatt

Bemerkung 4.4. 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 $K_m = \text{span}\{q, A^{-1}q, \ldots, A^{-m+1}q\}$ 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 4.5. 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.
$\Longrightarrow$ „Switching"-Strategien um zu testen, ob ein Neustart sinnvoll ist, um die Konvergenz zu beschleunigen.

Bemerkung 4.6. Der Algorithmus der rekursiven Form des Gram-Schmidt-Verfahrens kann auch für die stabile Orthogonalisierung einer allgemeinen Basis $\{v_1, \ldots, v_m\} \subset \mathbb{C}^n$ verwendet werden:
Algobox
$u^1 = \|v^1\|_2^{-1} \cdot v^1$
Für $t = 2, ..., m$:
Für $j = 1, ..., t-1$:
$u^{t,1} = v^t, u^{t,j+1} = u^{t,j} - proj_{u^j}(u^{t,j})$
$u^t = \|u^{t,t}\|_2^{-1} \cdot u^{t,t}$
Dieser modifizierte Gram-Schmidt-Algorithmus (mit exakter Arithmetik) liefert das gleiche Resultat, wie der klassiche Gram-Schmidt-Algorithmus.
$u^1 = \|v^1\|_2^{-1} \cdot v^1$
Für $t = 2, ..., m$:
$\tilde{u}^t = v^t - \sum_{j=1}^{t-1} proj_{u^j}(v^t)$
$u^t = \|\tilde{u}\|_2^{-1} \cdot \tilde{u}^t$ Beide Algorithmen haben die gleiche arithmetische Komplexität.
In jedem Schritt wird ein zu den vorherigen Vektoren orthogonaler Vektor bestimmt und dieser ist auch orthonormal zu einem eventuellen Fehler, der bei den Berechnungen entstand, was die Stabilität angeht.
Für die Fehlerabschätzung der resultierenden orthonormalen Matrix $u = [u^1, \ldots, u^m]$ gilt:

$$\|u^T u - I\|_2 \leq \frac{c_1 \cdot \text{cond}_2(A)}{1 - c_2 \cdot \text{cond}_2(A)}$$

(Vgl. Björk & Paige)

Bemerkung 4.7. 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

Voraussetzung: A sei hermitisch. Dann erhält man für die Rekursions-Formel der Arnoldi-Methode:

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

da $\langle Aq^{t-1}, q^j\rangle_2 = \langle q^{t-1}, Aq^j\rangle_2 = 0$ für $j = 1, \dots, t-3$.
Die Vereinfachung zu

$$\tilde{q}^t = Aq^{t-1} - \underbrace{\langle Aq^{t-1}, q^{t-1}\rangle_2}_{=:\alpha_{t-1}} q^{t-1} - \underbrace{\langle Aq^{t-1}, q^{t-2}\rangle_2}_{=:\beta_{t-2}} q^{t-2} = Aq^{t-1} - \alpha_{t-1}q^{t-1} - \beta_{t-2}q^{t-2}$$

Da A hermitisch, ist $\alpha_{t-1} \in \mathbb{R}$. Multiplikation mit q^t ergibt

$$\|\tilde{q}^t\|_2 = \langle q^t, \tilde{q}^t\rangle_2 = \langle q^t, Aq^{t-1} - \alpha_{t-1}q^{t-1} - \beta_{t-2}q^{t-2}\rangle_2 = \langle q^t, Aq^{t-1}\rangle_2 = \langle Aq^t, q^{t-1}\rangle_2 = \beta_{t-1}$$

Daraus folgt, dass auch $\beta_{t-1} \in \mathbb{R}$ und $\beta_{t-1}q^t = \tilde{q}^t$. Also erhält man

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

In Matrix-Form:

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

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

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

Definition 4.8 (Lanczos-Algorithmus). Für eine hermitische Matrix $A \in \mathbb{C}^{n\times n}$ bestimmt die Lanczos-Methode eine Menge von orthogonalen Vektoren $\{q^1, \dots, q^m\}$, $m \ll n$ durch Anwendung der Gram-Schmidt-Methode auf die Basis $\{q, Aq, \dots, A^{m-1}q\}$ des Krylov-Raumes K_m:

Algobox:
Startwerte: $q^1 = \|q\|_2^{-1}q, q^0 = 0, \beta_1 = 0$
Iteriere für $1 \le t \le m-1$:

$$\begin{aligned} r^t &= Aq^t, \quad \alpha_t = \langle r^t, q^t\rangle_2 \\ s^t &= r^t - \alpha_t q^t - \beta - tq^{t-1} \\ \beta^{t+1} &= \|s^t\|_2, q^{t+1} = \beta_{t+1}^{-1}s^t \\ r^m &= Aq^m, \alpha_m = \langle r^m, q^m\rangle_2 \end{aligned}$$

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

4.4 Pseudospektren

Motivation: Mit Fußball spielen: Bild einfügen
Bei dem einen stabil, weil Ball bleibt in Kuhle. Bei dem anderen instabil, da Ball wegrollt.
Sauber: Fehler verschwinden / werden schlimmer.

Spezialfall: Bild einfügen

Kleine Auslenkung ist kein Problem, wenn zu groß, dann schon.
Frage ist, wie groß ist das Attraktorgebiet?

Begriff der Pseudospektren geht auf Landau zurück. Viele Resultate von Trefethen etal.
Konzept des Pseudo-Spektrums ist bei normalen Operatoren die Vereinigung der ε-Umgebungen seiner Eigenwerte

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

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

Bemerkung 4.10. Das Konzept des Pseudo-Spektrums kann in viel allgemeineren Situationen eingeführt werden. (z.B. Dunford & Schwartz oder Kato)

Bemerkung 4.11. Krylov-Unterraum-Methoden, die bisher diskutiert wurden, lassen sich zur Berechnung des Pseudo-Spektrums einer Matrix verwenden. (z.B. bei Matrix von diskretisierten partiellen Differentialgleichungen)

Lemma 4.12.

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.

   $$\sigma_{\min(B)} := \min\{\lambda^{1/2} \,|\, \lambda \in \sigma(B^*)\}$$

   mit der adjungierten B^* von 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

   $$\begin{aligned}\|(zI - T)^{-1}\|_2 &= \max\{\mu 1/2 \,|\, \mu \text{ Singulärwert von } (zI - T)^{-1}\} \\ &= \min\{\mu 1/2 \,|\, \mu \text{ Singulärwert von } (zI - T)\}^{-1} \\ &= \sigma_{\min}(zI - T)^{-1}\end{aligned}$$

   Daraus folgt:

   $$\begin{aligned}\sigma_\varepsilon(T) &= \left\{z \in \mathbb{C} \;\middle|\; \|(zI - T)^{-1}\|_2 \geq \varepsilon^{-1}\right\} \\ &= \left\{z \in \mathbb{C} \;\middle|\; \sigma_{\min}(zI - T)^{-1} \geq \varepsilon^{-1}\right\} \\ &= \{z \in \mathbb{C} \mid \sigma_{\min}(zI - T) \leq \varepsilon\}\end{aligned}$$

2. Übungsaufgabe

Betrachtung einer Folge von Gitterpunkten $z_i \in D \subset \mathbb{C}$ für $i = 1, 2, 3, \ldots$ und in jedem Gitterpunkt wir das kleinste ε bestimmt, für das $z_i \in \sigma_\varepsilon(T)$.
Interpolation der erhaltenen Werte, bestimmt ob ein Punkte $z \in \mathbb{C}$ approxmativ zu $\sigma_\varepsilon(T)$ gehört.

5 Die schnelle Fourier-Transformation

(FFT = Fast Fourier Transformation)

5.1 Fourier-Reihen

Sei f eine 2π-periodische Funktion, d.h. $f(x+2\pi) = f(x)$ für alle $x \in \mathbb{R}$. Das Ziel ist eine Annäherung durch Linearkombinationen 2π-periodischen Funktionen:

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

Wir wollen eine Approximation, d.h.

$$\|g_n(x) - f(x)\|_2 = \left(\int_{-\pi}^{\pi} (g_n(x) - f(x))^2 dx \right)^{1/2} \to \min$$

Orthogonalitätsbedingungen der trigonometrischen Funktionen werden verwendet um Fourier-Koeffizienten zu bestimmen:

$$a_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(kx) dx, \quad k = 0, 1, \ldots, n$$
$$b_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(kx) dx, \quad k = 1, \ldots, n$$

Im allg. Fall ergeben sich für die Fourier-Koeffizienten a_k und b_k keine geschlossenen Formeln, d.h. wir sind auf numerische Integration angewiesen
Die Wahl Trapezregel als Quadratur liefert die diskrete Fouriertransformation
Unterteilung des Intervalls $[0, 2\pi]$ in N Teilintervalle. Man hat somit eine Schrittweite $h = \frac{2\pi}{N}$ und Integrationsstützstellen

$$x_j = hj = \frac{2\pi}{N} \cdot j, \quad j = 0, 1, \ldots, N$$

Mit der Trapezregel folgt

$$\begin{aligned} a_k &= \frac{1}{\pi} \int_0^{2\pi} f(x) \cos(kx) dx \\ &\approx \frac{1}{\pi} \cdot \frac{2\pi}{2N} \left(f(x0) \cdot \cos(kx_0) + 2 \sum_{j=1}^{N-1} f(x_j) \cdot \cos(kx_j) + f(x_N) \cdot \cos(kx_N) \right) \end{aligned}$$

Mit Berücksichtigung der 2π-Periodizität von $f(x)$ ergibt sich für a_k und analog für b_k die Näherungswerte

$$a_k^* := \frac{2}{N} \sum j = 1^N f(x_j) \cdot \cos(kx_j), \quad k = 0, 1, 2, \ldots b_k^* \quad := \frac{2}{N} \sum j = 1^N f(x_j) \cdot \cos(kx_j), \quad k = 1, 2, 3 \ldots$$

Satz 5.1. Für die diskreten Stützstellen x_j gelten

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

und

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

Beweis.

$$S := \sum_{j=1}^{N} \cos(kx_j) + i\sin(kx_j) = \sum_{j=1}^{N} e^{ikx_j} = \sum_{j=1}^{N} e^{ijkh}$$

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 geometrischen Reihe ergibt

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

Für $\frac{k}{N} \in \mathbb{Z}$ folgt wegen $q = 1$, dass $S = N$ ist.
Betrachtung on Realteilt und Imaginärteil liefert die Behauptung.
Intuition Bilder

Satz 5.2. Die trigonometrischen Funktionen erfüllen für die äquidistanten Stützstellen x_j die diskreten Orthogonalitätsrelationen

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

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 die Aussagen des folgenden Satzes angewendet

Satz 5.3. Es sei $N = 2n$, mit $n \in \mathbb{N}$. Da Fourier-Polynom

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

von Gram $m < n$ mit Koeffizienten a_k^* und b_k^* approximiert die Funtkion $f(x)$ im diskreten qua-

dratischen Mittel der N Stützstellen x_j derart, dass die Summe der Quadrate der Abweichungen

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

minimal ist.

Beispiel 5.4. 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 vorallem bei der diskreten Fouriertransformation relevant, die in Ingenieur- und Naturwissenschaften häufig eingesetzt wird, um z.B. die Frequenzen von Vibrationen.
$\Longrightarrow$ Aufwand $\propto N^2$ (N^2 trigonometrischen Funktionsauswertungen). Problem bei $N \gg 1000$
Runge 1903, 1905, sowie verchiedene unabhängige Arbeiten vieler Mathematiker.

5.3 Schnelle Fourier-Transformation (Details)

Zur Berechnung der Summe

$$a'_k := \sum_{j=0}^{N-1} f(x_j)\cos(kx_j), \quad k = 0, 1, \ldots, \tfrac{N}{2}$$
$$b'_k := \sum_{j=0}^{N-1} f(x_j)\cos(kx_j), \quad k = 1, 2, \ldots, \tfrac{N}{2} - 1$$

mit $x_j = \frac{2\pi}{N} \cdot j$, kann für den Spezialfall, in dem N eine Potenz von 2 ist, ein sehr effizienter Algorithmus entwickelt werden, wenn man zu einer komplexen Fouriertransformation übergeht.
Aus zwei aufeinanderfolgenden Stützwerten bildet man die $n = N\,/\,2$ komplexen Zahlenwert:

$$y_j := f(x_{2j}) + i \cdot f(x_{2j+1}), \qquad \text{für } j = 0, 1, \ldots, n-1$$

Definition 5.5 (Diskrete komplexe Fouriertransformation der Ordnung n:).

$$c_k := \sum_{j=0}^{n-1} y_j \cdot \exp\left(ijk\frac{2\pi}{n}\right) = \sum_{j=0}^{n-1} y_j \cdot \omega_n^{jk}$$

mit

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

Dabei sind die ω_n die n-ten Einheitswurzeln.

Satz 5.6 (Zusammenhang zwischen reellwertigen und komplexen Fourier-Transformation). Die reellwertigen trigonometrischen Summen a'_k und b'_k sind gegeben durch die komplexen Fourier-Transformierten c_k durch:

$$a'_k - ib'_k = \tfrac{1}{2}(c_k + \overline{c_{n-k}}) + \tfrac{1}{2i}(c_k - \overline{c_{n-k}})e^{\frac{-ik\pi}{n}}$$
$$a'_{n-k} - ib'_{n-k} = \tfrac{1}{2}(\overline{c_k} + c_{n-k}) + \tfrac{1}{2i}(\overline{c_k} - c_{n-k})e^{\frac{ik\pi}{n}}$$

für $k = 0, \dots n$ falls $b'_0 = b'_n = 0$ und $c_n = c_0$

Beweis. Für den ersten Summanden der oberen Formel erhält man

$$\begin{aligned}\frac{1}{2}(c_k + \overline{c_{n-k}}) &= \frac{1}{2} \cdot \sum_{j=0}^{n-1} \left\{ y_j \cdot \omega_n^{jk} + \overline{y_j} \cdot \overline{\omega_n^{j(n-k)}} \right\} \\ &= \frac{1}{2} \cdot \sum_{j=0}^{n-1} (y_j + \overline{y_j}) \cdot \omega_n^{jk}\end{aligned}$$

Für den Ausdruck in Klammern des 2. Summanden prüfen, glaube - statt +

$$\begin{aligned}\frac{1}{2i}(c_k + \overline{c_{n-k}}) &= \frac{1}{2i} \cdot \sum_{j=0}^{n-1} \left\{ y_j \cdot \omega_n^{jk} + \overline{y_j} \cdot \overline{\omega_n^{j(n-k)}} \right\} \\ &= \frac{1}{2} \cdot \sum_{j=0}^{n-1} (y_j + \overline{y_j}) \cdot \omega_n^{jk}\end{aligned}$$

Mit Definition von y_j folgt

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

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

Die Reduktion einer komplexen Fouriertransformation von gerader Ordnung auf zwei Fouriertransformationen je der halben Ordnung ist möglich.
Diese Reduktion der Ordnung wird iterativ durchgeführt:

Es sei $n = 2m$, $m \in \mathbb{N}$. Dann gilt für die komplexe Fouriertransformierte c_k mit geraden Indizes $k = 2l$, $l = 0, 1, \dots, m-1$:

$$c_{2l} = \sum_{j=0}^{2m-1} y_j \omega_n^{2lj} = \sum_{j=0}^{m-1} (y_j + y_{m+j}) \omega_n^{2lj}$$

Dabei wurde die Identität

$$\omega_n^{2l(m+j)} = \omega_n^{2lj} \cdot \omega_n^{2lm} = \omega^{2lj} \cdot \omega_n^{ln} = \omega_n^{2lj} \cdot \left(e^{-i\frac{2\pi}{n}}\right)^{2ln} = \omega_n^{2lj} \cdot \left(e^{-i\cdot 2\pi}\right)^l = \omega_n^{2lj}$$

Mit den m Hilfswerten

$$z_j := y_j + y_{m+j}, \quad j = 0, \dots, m-1$$

und wegen $\omega_n^2 = \omega_m$ sind die Koeffizienten

$$c_{2l} = \sum_{j=0}^{m-1} z_j w_m^{jl}$$

die Fouriertransformierten der Ordnung m der Hilfswerte z_j.
Für die c_k mit ungeraden Indizes $k = 2l + 1$ mit $l = 0, 1, \ldots, m-1$ gilt

$$\begin{aligned} c_{2l+1} &= \sum_{j=0}^{2m-1} y_j \omega_n^{(2l+1)j} \\ &= \sum_{j=0}^{m-1} \left\{ y_j \omega_n^{(2l+1)j} + y_{j+m} \omega_n^{(2l+1)(n+j)} \right\} \\ &= \sum_{j=0}^{m-1} (y_j - y_{m+j}) \omega_n^{(2l+1)j} \\ &= \sum_{j=0}^{m-1} (y_j - y_{m+j}) \omega_n^j \cdot \omega_n^{2lj} \end{aligned}$$

Mit den weiteren m Hilfswerten

$$z_{m+j} := (y_j - y_{m+j}) \omega_n^j, \quad j = 0, 1, \ldots, m-1$$

sind die Koeffizienten

$$c_{2l+1} = \sum_{j=0}^{m-1} z_{m+j} \omega_m^{jl}, \quad l = 0, 1, \ldots, m-1$$

die Fouriertransformierten der Ordnung m der Hilfswerte z_{m+j}.
Die Zurückführung einer komplexen Fouriertransformationen der Ordnung $n = 2m$ auf 2 komplexe Fouriertransformationen der Ordnung m erfordert nach den obigen Formeln als wesentlichen Rechenaufwand m komplexe Multiplikationen.
In die Ordnung $n = 2^\gamma$, $\gamma \in \mathbb{N}$, so kann die Reduktion auf 2 Fouriertransformationen halber Ordnung iterativ durchgeführt werden.

Beispiel 5.7. $FT_{32} \to 2FT_{16} \to 4FT_8 \to 8FT_4 \to 16FT_2 \to 32FT_1$

Da jeder Schritt $n \,/\, 2$ komplexe Multiplikationen fordert,