Numerische Mathematik II – Numerische lineare Algebra und Optimierung in den Datenwissenschaften

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

Inhaltsverzeichnis

Inhaltsverzeichnis

Diese Mitschrift basiert auf der gleichnamigen Vorlesung *Numerische Mathematik II - Numerische lineare Algebra und Optimierung 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

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

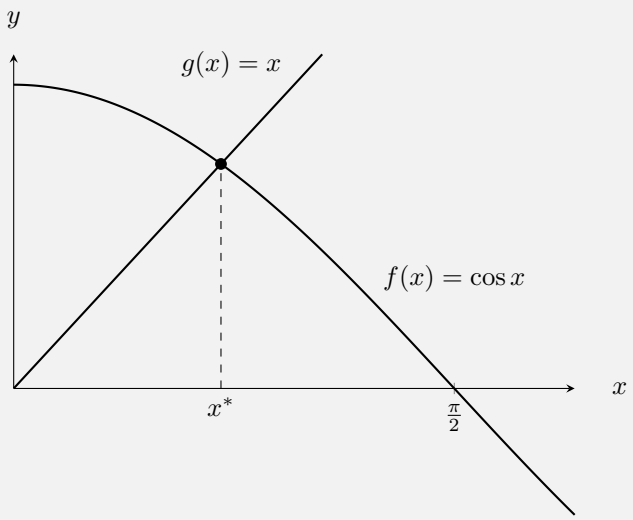
   Da M abgeschlossen ist existiert also ein Fixpunkt in M.
   Dieser ist eindeutig, denn für x, y mit $Tx = x$ und $Ty = y$ gilt $d(x, y) = d(Tx, Ty) \leq kd(x, y)$, also $d(x, y) = 0$.

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 $x_{k+1} = \cos x_k$.

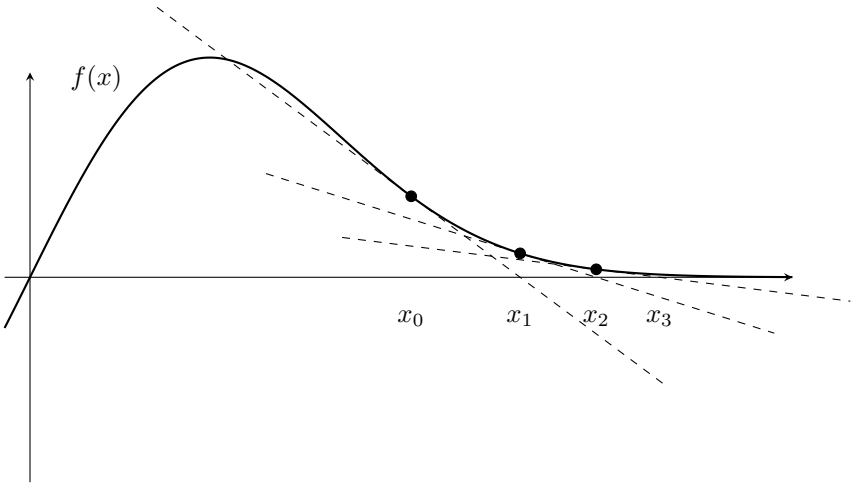
Wir betrachten im folgenden die Idee der Umwandlung eines Nullstellenproblems in Fixpunkt-Gleichung noch etwas allgemeiner. Für eine Gleichung $f(x) = 0$ mit $f : \mathbb{R} \to \mathbb{R}$ haben wir verschiedene Möglichkeiten zur Umformung:

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:

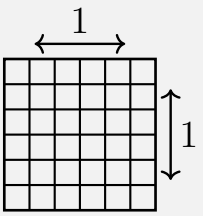
Satz 2.19 (Satz von Taylor). Sei $f : [a, b] \to \mathbb{R}$ eine $(n+1)$-mal stetig differenzierbare Funktion und $x, x_0 \in [a, b]$. Dann existiert ein ξ zwischen x und x_0, so dass

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

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:

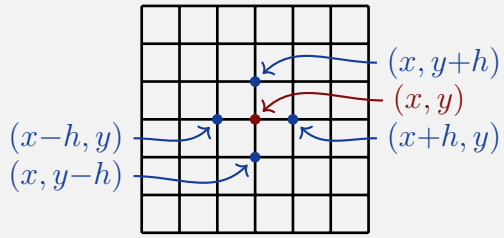
$$\frac{\mathrm{d}^+}{\mathrm{d}x} f(x) \mid_{x=x_0} = \frac{1}{\Delta x} \left(f(x_0 + \Delta x) - f(x_0) \right) \qquad \text{(rechtsseitiger DQ)}$$

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

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

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 Eigenwektor 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

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

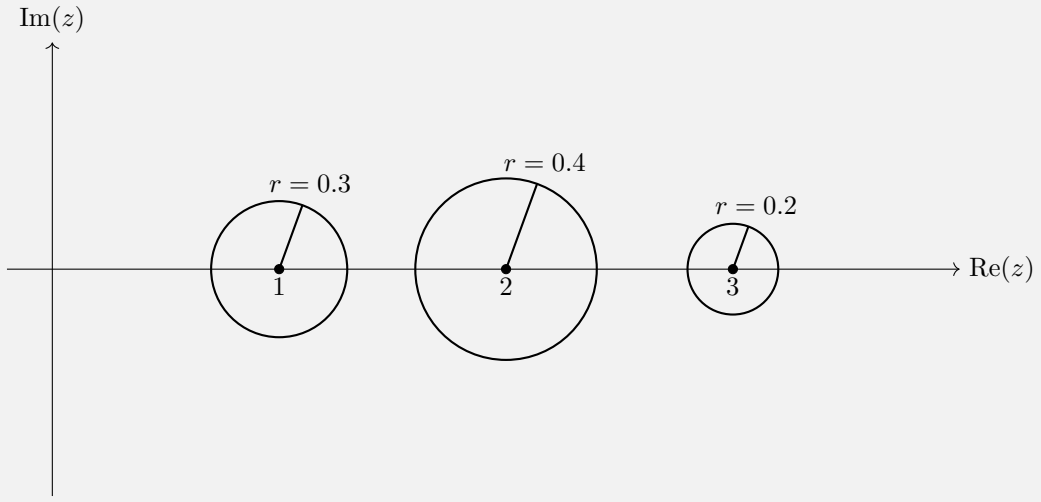
Demnach liegt $\lambda \in K_i$. □

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 $\{w^{(1)}, \ldots, w^{(n)}\}$ 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 $W = (w^{(1)}|\ldots|w^{(n)})$ gilt

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

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

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

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

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

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

Für einen beliebigen Index $k \in \{1, \ldots, n\}$, (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 $\{w^{(1)}, \ldots, w^{(n)}\}$ 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.

Beispiel 3.8 (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 streben ein lineares Model an, da dieses mathematisch einfacher handhabbar ist. 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 die Eigenwertgleichung für den Eigenwert 1.