Iterationsverfahren

Prof. Dr. Andreas Frommer

Sommersemester 2003

Bergische Universität Wuppertal
Fachbereich Mathematik

Inhaltsverzeichnis

1 Krylov-Unterräume

1.1 Definition
Zu $r \in \mathbb{C}^n \setminus \{0\}$ und $A \in \mathbb{C}^{n\times n}$ definieren wir den *Krylov-Unterraum* der Stufe m als

$$\begin{aligned} K_m(A,r) &:= \operatorname{span}\{r, Ar, A^2r, ..., A^{m-1}r\} \\ &= \{y \in \mathbb{C}^n : \; y = \sum_{j=0}^{m-1} \alpha_j A^j r\} \\ &= \{y \in \mathbb{C}^n : \; y = p_{m-1}(A)r, \quad p_{m-1} \in \Pi_{m-1}\}, \end{aligned}$$

wobei Π_{m-1} der Vektorraum der Polynome vom Höchstgrad $m-1$ ist, d.h.

$$\Pi_{m-1} := \left\{ p : \; p = \sum_{j=0}^{m-1} a_j x^j \right\}.$$

Krylov-Unterräume sind geschachtelt. Da sie alle Teilräume von $\mathbb{C}^n$ sind, werden sie irgendwann nicht mehr größer werden. Wir untersuchen dies genauer.

1.2 Lemma
Zu $r \in \mathbb{C}^n$ existiert ein minimales $m^* \in \mathbb{N}_0$, $m^* \leq n$ und ein Polynom $p_{m^*} \neq 0$ vom Grad m^* mit

1. $p_{m^*}(A)r = 0$ (dieses Polynom ist bis auf skalare Vielfache eindeutig; alle Polynome p mit $p(A)r = 0$ bilden ein Ideal).

2. Es gilt

$$\begin{aligned} \dim(K_m(A,r)) &\leq \dim(K_{m+1}(A,r)), \quad m = 1, \dots, m^*, \\ \dim(K_{\hat{m}}(A,r)) &= \hat{m}, \quad \hat{m} = 1, ..., m^*, \\ K_{m^*}(A,r) &= K_{m^*+1}(A,r) = K_{m^*+2}(A,r) = \dots . \end{aligned}$$

Beweis:

1. Das Minimalpolynom p_A von A erfüllt $p_A(A)r = 0$, also existiert auch ein Polynom minimalen Grades mit der geforderten Eigenschaft.

2. Es ist klar, dass

$$\dim(K_m(A,r)) \leq \dim(K_{m+1}(A,r)) \leq \dim(K_m(A,r)) + 1$$

gilt, da

$$K_{m+1}(A,r) = K_m(A,r) + \langle A^m r \rangle .$$

- Für
$$p_{m^*}(t) = \sum_{j=0}^{m^*} c_j t^j, \quad c_{m^*} \neq 0$$
folgt sofort aus $p_{m^*}(A)r = 0$
$$A^{m^*} \in K_{m^*}(A, r) \Rightarrow K_{m^*}(A, r) = K_{m^*+1}(A, r).$$
Jedes Polynom p mit $\deg(p) = m > m^*$ besitzt eine Darstellung $p = q \cdot p_{m^*} + s$ mit $\deg(s) < m^*$. Also ist $p(A)r = s(A)r$ und damit $K_m(A, r) \subseteq K_{m^*}(A, r)$, also $K_m(A, r) = K_{m^*}(A, r)$.
- Es sei $\widetilde{m}$ der erste Index mit $K_{\widetilde{m}}(A, r) = K_{\widetilde{m}+1}(A, r)$, dann ist
$$\begin{aligned} A^{\widetilde{m}} r \in K_{\widetilde{m}} &\Longleftrightarrow \exists \text{ Polynom } p \text{ vom Grad } \widetilde{m} \text{ mit } \widetilde{p}(A)r = 0 \\ &\Longrightarrow \widetilde{m} \geq m^* \Rightarrow \widetilde{m} = m^*. \end{aligned}$$

□

1.3 Satz
Sei $A \in \mathbb{C}^{n \times n}$ regulär und $x^0 \in \mathbb{C}^n,\ r^0 = b - Ax^0$. Dann erfüllt die Lösung $x^* = A^{-1}b$ von
$$Ax = b$$
die Beziehung
$$\begin{aligned} &x^* \in x^0 + K_{m^*}(A, r^0), \\ &x^* \notin x^0 + K_m(A, r^0), \quad m < m^*. \end{aligned}$$

Beweis: Sei $p_{m^*}(t) = \sum\limits_{j=0}^{m^*} c_j t^j$ wie in Lemma 1.2. Dann ist $c_0 \neq 0$, denn sonst gälte
$$\sum_{j=1}^{m^*} c_j A^j r = 0 \Longleftrightarrow A \underbrace{\left(\sum_{j=1}^{m^*} c_j A^{j-1} r \right)}_{\neq 0} = 0,$$
im Widerspruch zur Regularität von A.

Es gilt also
$$\begin{aligned} p_{m^*}(A) r^0 = 0 &\Longleftrightarrow A^{-1} p_{m^*}(A) r^0 = 0 \\ &\Longleftrightarrow A^{-1} \sum_{j=0}^{m^*} c_j A^j r^0 = 0 \end{aligned}$$

$$\begin{aligned}
&\Longleftrightarrow A^{-1}\sum_{j=0}^{m^*} c_j A^j (b - Ax^0) = 0 \\
&\Longrightarrow x^* = A^{-1}b = x^0 + \frac{1}{c_0}\sum_{j=1}^{m^*} c_j A^{j-1} r^0 \\
&\qquad \in x^0 + K_{m^*}(A, r^0).
\end{aligned}$$

Die Annahme $x^* \in x^0 + K_{m^*-1}(A, r^0)$ führt auf ein Polynom $p \neq 0$ mit $\deg(p) < m^*$ und $p(A)r^0 = 0$, im Widerspruch zur Minimalität von m^*. $\square$

1.4 Definition
Ein *Krylov-Unterraum-Verfahren* (KUV) zur Lösung von $Ax = b$ ist ein Unterraumverfahren mit Startwert x^0, Startresiduum $r^0 = b - Ax^0$ und

$$x^m \in x^0 + K_m(A, r^0).$$

Es ist also $x^m \in x^0 + q_{m-1}(A)r^0,\ \deg(q_{m-1}) \leq m - 1$ und

$$\begin{aligned}
r^m = b - Ax^m &= r^0 - Aq_{m-1}(A)r^0 \\
&= p_m(A)r^0, \quad p_m \in \overline{\Pi}_m = \{p \in \Pi_m : \ p(0) = 1\} \\
&\in K_{m+1}(A, r^0).
\end{aligned}$$

p_m und q_{m-1} (mit $p_m(t) = 1 - t \cdot q_{m-1}(t)$) heißen auch die zum KUV gehörigen *Verfahrenspolynome*.

Umgekehrt definiert jede Folge von Polynomen $q_{m-1} \in \Pi_{m-1}$ ein KUV mit $x^m = x^0 + q_{m-1}(A)r^0$ und ebenso jede Folge von Polynomen $p_m \in \overline{\Pi}_m$ (mit $r^m = p_m(A)r^0$). Grundsätzlich besteht immer der Zusammenhang

$$p_m(t) = 1 - tq_{m-1}(t). \tag{1.1}$$

Wir interpretieren nun einfache bekannte Verfahren als KUV.

Erinnerung: Es sei $A = (a_{i,j}) \in \mathbb{C}^{n \times n}$, dann sind $D, L, U \in \mathbb{C}^{n \times n}$ gegeben durch

$$D = \begin{pmatrix} a_{1,1} & & & \\ & \ddots & & 0 \\ & & \ddots & \\ & 0 & & \ddots \\ & & & & a_{n,n} \end{pmatrix} \quad \text{(Diagonalteil}$$

$$L = \begin{pmatrix} 0 & & & & \\ -a_{2,1} & \ddots & & \text{\huge 0} & \\ \vdots & & \ddots & & \\ \vdots & & & \ddots & \\ -a_{n,1} & \dots & \dots & -a_{n,n-1} & 0 \end{pmatrix}$$ (negativer) linker unterer Dreiecksteil

$$U = \begin{pmatrix} 0 & -a_{1,2} & \dots & \dots & -a_{1,n} \\ & \ddots & & & \vdots \\ & & \ddots & & \vdots \\ & \text{\huge 0} & & \ddots & -a_{n-1,n} \\ & & & & 0 \end{pmatrix}$$ (negativer) rechter oberer Dreiecksteil

Es ist also $A = D - L - U$.
Wir betrachten nun die drei Standard-Verfahren:

Jacobi:	$x^{m+1} = D^{-1}((L+U)x^m + b)$
Gauß-Seidel:	$x^{m+1} = (D-L)^{-1}(Ux^m + b)$
SOR:	$x^{m+1} = (\frac{1}{\omega}D - L)^{-1}((\frac{1-\omega}{\omega}D + U)x^m + b) \quad \omega \in \mathbb{R} \setminus \{0\}$

und interpretieren sie folgendermaßen als KUV:

Jacobi: Setze $H := D^{-1}(L+U)$ und $A' = D^{-1}A,\ b' = D^{-1}b$. Mit $r^0 = b' - A'x^0$ erhalten wir

$$\begin{aligned} x^{m+1} &= Hx^m + b' \\ r^{m+1} &= b' - A'x^{m+1} = b' - A'(Hx^m + b') \\ &= b' - A'((I - A')x^m + b') \\ &= b' - A'(x^m + r^m) \\ &= r^m - A'r^m = (I - A')r^m \\ \Rightarrow r^m &= (I - A')^m r^0. \end{aligned}$$

Also ist das Jacobi-Verfahren ein KUV bezüglich $A'x = b'$ mit

$$\begin{aligned} p_m(t) &= (1-t)^m, \\ q_{m-1}(t) &= \frac{1 - p_m(t)}{t} = \sum_{j=0}^{m-1} (-1)^j \binom{m}{j+1} t^j. \end{aligned}$$

Gauß-Seidel: ist KUV bezüglich $\underbrace{(D-L)^{-1}A}_{=A'}x = \underbrace{(D-L)^{-1}b}_{b'}$ mit

$$p_m(t) = (1-t)^m,$$
$$q_{m-1}(t) = \frac{1-p_m(t)}{t} = \sum_{j=0}^{m-1}(-1)^j\binom{m}{j+1}t^j.$$

SOR: ist KUV bezüglich $\underbrace{\left(\frac{1}{\omega}D-L\right)^{-1}A}_{=A'}x = \underbrace{\left(\frac{1}{\omega}D-L\right)^{-1}b}_{b'}$ mit

$$p_m(t) = (1-t)^m,$$
$$q_{m-1}(t) = \frac{1-p_m(t)}{t} = \sum_{j=0}^{m-1}(-1)^j\binom{m}{j+1}t^j.$$

Allgemein gilt folgender

1.5 Satz
Ein Iterationsverfahren

$$x^{m+1} = Hx^m + b' \tag{1.2}$$

zur Lösung von $Ax = b$ mit $A = M - N$, $H = M^{-1}N$, $b' = M^{-1}b$ ist ein KUV (bezüglich des Systems $A'x = b'$, $A' = M^{-1}A$) mit $p_m(t) = (1-t)^m$.

Bemerkung: Die Verfahrenspolynome hängen bei diesen einfachen Iterationen nicht vom Startresiduum ab.

1.6 Definition
Ein KUV zur Lösung von $Ax = b$ mit regulärem A heißt *konvergent*, falls

$$\lim_{k\to\infty} r^k = 0 \iff \lim_{k\to\infty} x^k = x^* = A^{-1}b.$$

1.7 Satz
Ein Iterationsverfahren der Gestalt (1.2) ist konvergent, falls

$$\rho(I - M^{-1}A) < 1$$

gilt.

Beweis: Es gilt $r^m = (I - M^{-1}A)^m r^0$, wegen $\rho(I - M^{-1}A) < 1$ gilt $\lim\limits_{k\to\infty}(I - M^{-1}A) = 0$, also $\lim\limits_{k\to\infty} r^k = 0$. $\square$

1.8 Definition
Das *Richardson-Verfahren erster Ordnung* zur Lösung von $Ax = b$ ist die Iteration

$$x^{m+1} = (I - A)x^m + b = x^m + r^m.$$

Die Richardson-Iteration erster Ordnung ist im Fall $D = I$ mit der Jacobi-Iteration identisch. Für die Verfahrenspolynome gilt wieder

$$p_m(t) = (1 - t)^m.$$

1.9 Korollar
Das Richardson-Verfahren erster Ordnung zur Lösung von $Ax = b$ konvergiert, falls

$$\rho(I - A) < 1,$$

z.B. falls $\operatorname{spek}(A) \subset (0, 2)$.

1.10 Definition
Das *relaxierte Richardson-Verfahren erster Ordnung* zur Lösung von $Ax = b$ ist die Iteration

$$x^{m+1} = x^m + \alpha r^m.$$

Hierbei ist $\alpha \in \mathbb{C}$ ein fester Parameter, der bei geeigneter Wahl die Konvergenz beschleunigt.

Die Verfahrenspolynome für das relaxierte Richardson-Verfahren sind

$$p_m(t) = (1 - \alpha t)^m. \tag{1.3}$$

1.11 Definition
Das *Richardson-Verfahren zweiter Ordnung* zur Lösung von $Ax = b$ ist bei gegebenem x^0 die Iteration

$$\begin{aligned} x^1 &= (I - A)x^0 + b, \\ x^{m+1} &= \alpha((I - A)x^m + b) + (1 - \alpha)x^{m-1}, \quad m \geq 1, \end{aligned}$$

mit $\alpha \in \mathbb{C}$ fest.

Interpretation als KUV:

$$\begin{aligned} r^1 &= (I - A)r^0 \quad \Rightarrow \quad p_1(t) = 1 - t \\ r^{m+1} &= b - Ax^{m+1} \\ &= b - \alpha A((I - A)x^m + b) - (1 - \alpha)Ax^{m-1} \\ &= \alpha r^m - \alpha A r^m + (1 - \alpha)r^{m-1} \\ &= \alpha(I - A)r^m + (1 - \alpha)r^{m-1} \end{aligned}$$

also

$$p_{m+1}(t) = \alpha(1-t)p_m(t) + (1-\alpha)p_{m-1}(t), \quad m \geq 1.$$

Solche *3-Term-Rekursionen* sind typisch für „fortgeschrittene“ KUV. Für eine spätere Analyse halten wir fest, dass gilt

$$(1.4) \qquad \begin{pmatrix} p_{m+1}(t) \\ p_m(t) \end{pmatrix} = \underbrace{\begin{pmatrix} \alpha(1-t) & 1-\alpha \\ 1 & 0 \end{pmatrix}}_{B} \cdot \begin{pmatrix} p_m(t) \\ p_{m-1}(t) \end{pmatrix}.$$

2 Allgemeine Betrachtungen zur Konvergenz von Krylov-Unterraum-Verfahren

Für jedes Krylov-Unterraum-Verfahren zur Lösung von $Ax = b, A \in \mathbb{C}^{n\times n}$, $b \in \mathbb{C}^n$ gilt

$$r^m = b - Ax^m = p_m(A)r^0, \quad p_m \in \overline{\Pi}_m.$$

Ein KUV ist konvergent, wenn

$$\lim_{m\to\infty} p_m(A)r^0 = 0$$

gilt. Bei geeigneter Wahl von p_m wird $p_m(A)r^0 = 0$ bereits für ein endliches $m \leq n$; manche Verfahren erreichen dies auch tatsächlich). Andere Verfahren (wie z.B. das Richardson-Verfahren) benötigen eine unendliche Iterationszahl.

Man unterscheidet zwei verschiedene Ansätze zur Analyse von KUV:

1. parameterfreie KUV: sie benötigen keine (detailliertere) Information über Eigenschaften von A (z.B. das Spektrum). In der Regel bilden sie $p_m(A)$ dann in expliziter Abhängigkeit von $r^0 \hookrightarrow$ später (vergleiche cg-Verfahren),

2. parameterabhängige KUV verwenden detailliertere Informationen über A, z.B. über die Lage des Spektrums. In der Regel bilden sie dann $p_m(A)$ unabhängig von r^0. Vorteil: keine Innenproduktbildung, damit besser parallelisierbar; Nachteil: eventuell weniger effizient.

Frage: Wann gilt $\lim\limits_{m\to\infty} p_m(A) = 0 \in \mathbb{C}^{n\times n}$?
Wäre A diagonalisierbar, d.h. es gibt eine invertierbare Matrix T und eine Diagonalmatrix $\Lambda = diag(\lambda_1, ..., \lambda_n)$, so dass

$$A = T\Lambda T^{-1},$$

dann gilt

$$p_m(A) = Tp_m(\Lambda)T^{-1}.$$

Damit ist dann

$$\lim_{m\to\infty} p_m(A) = 0 \Longleftrightarrow \lim_{m\to\infty} p_m(\lambda) = 0 \quad \forall\, \lambda \in \text{spek}(A).$$

Ist A nicht diagonalisierbar, so werden weitere Bedingungen notwendig.

2.1 Satz
Es sei
$$A = TJT^{-1}$$
mit
$$J = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_k \end{pmatrix} \quad J_\ell = \begin{pmatrix} \lambda_\ell & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_\ell \end{pmatrix} \in \mathbb{C}^{n_\ell \times n_\ell}$$
die Jordan-Normalform von A mit $\lambda_\ell \in \operatorname{spek}(A) = \{\lambda_1, ..., \lambda_k\}$ (aber Mehrfachnennungen möglich). Dann gilt
$$\lim_{m\to\infty} p_m(A) = 0 \iff \lim_{m\to\infty} p_m^{(\nu)}(\lambda_\ell) = 0 \quad \ell = 1, ..., k,\ \nu = 0, ..., n_\ell - 1.$$

Beweis: Es gilt
$$p_m(A) = Tp_m(J)T^{-1}$$
also
$$\lim_{m\to\infty} p_m(A) = 0 \iff \lim_{m\to\infty} p_m(J_\ell) = 0 \quad \ell = 1, ..., k.$$
Wir betrachten also
$$J_\ell = \lambda_\ell \cdot I_{n_\ell} + S, \quad S = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix} \in \mathbb{C}^{n_\ell \times n_\ell} \Rightarrow S^{n_\ell} = 0.$$
Damit ergibt sich (mit $\binom{k}{\nu} = 0$, falls $\nu > k$)
$$J_\ell^k = \sum_{\nu=0}^{n_\ell - 1} \binom{k}{\nu} S^\nu \lambda_\ell^{k-\nu}.$$
Es sei nun $p_m(t) = \sum_{k=0}^{m} c_k t^k$, dann ist
$$\begin{aligned} p_m(J_\ell) &= \sum_{k=0}^{m} c_k \sum_{\nu=0}^{n_\ell-1} \binom{k}{\nu} S^\nu \lambda_\ell^{k-\nu} \\ &= \sum_{\nu=0}^{n_\ell-1} \frac{1}{\nu!} S^\nu \underbrace{\sum_{k=0}^{m} c_k \frac{k!}{(k-\nu)!} \lambda_\ell^{k-\nu}}_{p_m^{(\nu)}(\lambda_\ell)}. \end{aligned}$$

Hierin bezeichnet $p_m^{(\nu)}$ die ν-te Ableitung des Polynoms p_m.
In der Matrix S^ν ist genau die ν-te obere Nebendiagonale (mit Einsen) besetzt. Es gilt also

$$\lim_{m\to\infty} p_m(J_\ell) = 0 \Longleftrightarrow \lim_{m\to\infty} p_m^{(\nu)}(\lambda_\ell) = 0 \quad \nu = 0, ..., n_\ell - 1.$$

□

2.2 Korollar
Sei $A \in \mathbb{C}^{n\times n}$ diagonalisierbar (z.B. hermitesch oder normal[1]). Dann gilt

$$\lim_{m\to\infty} p_m(A) = 0 \Longleftrightarrow \lim_{m\to\infty} p_m(\lambda) = 0 \quad \forall\ \lambda \in \text{spek}(A).$$

Für diagonalisierbares A ist also die Größe $\max\limits_{\lambda\in\text{spek}(A)} |p_m(\lambda)|$ ein Maß für die Größe von $p_m(A)$, sie ist sogar eine Norm von $p_m(A)$:

$$A = SDS^{-1}, \quad D = \text{diag}(\lambda_1, ..., \lambda_n).$$

Nehme

$$\|x\|_S = \|S^{-1}x\|_2,$$

dann ist

$$\begin{aligned}
\|A\|_S = \max_{\|x\|_S\neq 0} \frac{\|Ax\|_S}{\|x\|_S} &= \max_{\|x\|_S\neq 0} \frac{\|DS^{-1}x\|_2}{\|S^{-1}x\|_2} \\
&= \max_{y\neq 0} \frac{\|Dy\|_2}{\|y\|_2} = \|D\|_2 \\
&= \max_{\lambda\in\text{spek}(A)} |\lambda|
\end{aligned}$$

und analog

$$\|p_m(A)\|_S = \max_{\lambda\in\text{spek}(A)} |p_m(\lambda)|.$$

Minimierung von

$$\max_{\lambda\in\text{spek}(A)} |p_m(\lambda)|$$

minimiert also auch $\|p_m(A)\|_S$.

Bemerkung: $\|\cdot\|_S$ ist i.A. eine „schiefe“ Norm. Ist jedoch A normal, so kann man S unitär wählen, so dass $\|x\|_S = \|x\|_2$ und damit $\|A\|_S = \|A\|_2 = \max\limits_{\lambda\in\text{spek}(A)} |p_m(\lambda)|$.

[1]d.h. $AA^* = A^*A$

Im diagonalisierbaren Fall löst damit ein bestmögliches Iterationsverfahren für jedes m die MinMax-Aufgabe

$$\min_{p_m \in \overline{\Pi}_m} \max_{\lambda \in \mathrm{spek}(A)} |p_m(\lambda)|. \tag{2.1}$$

Dies ist jedoch in den meisten Fällen praktisch nicht möglich, schon alleine weil man i.A. die Eigenwerte von A nicht alle kennt.
Parameterabhängige KUV verwenden deshalb nur partielle Informationen über spek(A), um mittels dieser Informationen möglichst „gute“ Polynome $p_m(t)$ zu finden.

Es sei z.B. (mittels Satz von Gerschgorin o.Ä.) bekannt, dass

$$\mathrm{spek}(A) \subset D(\zeta, \rho) = \{z \in \mathbb{C} : \; |z - \zeta| \leq \rho\}.$$

Wie lautet die Lösung von

$$\min_{p_m \in \overline{\Pi}_m} \max_{\lambda \in D(\zeta,\rho)} |p_m(\lambda)|?$$

Beachte: $0 \notin D(\zeta, \rho)$ ist hier eine vernünftige zusätzliche Bedingung, da ansonsten die Lösung $p_m \equiv 1$ lautet für alle m, d.h. es resultiert ein nicht konvergentes KUV.

2.3 Satz
Sei $\rho > 0$, sei $\gamma \in \mathbb{C}$ mit $|\gamma| > \rho$. Dann löst das Polynom

$$p_m(z) = \frac{z^m}{\gamma^m}$$

die Aufgabe

$$\min_{p_m \in \Pi_m,\; p_m(\gamma)=1} \max_{\lambda \in D(0,\rho)} |p_m(\lambda)|,$$

der Wert ist $|\rho/\gamma|^m$.

Bemerkung: $|\rho/\gamma|$ ist klein, wenn γ weit vom Kreis $D(0, \rho)$ entfernt ist.
Zum Beweis des Satzes benötigen wir den Satz von Rouché.

2.4 Lemma (Rouché)
Es seien f, g homolorph im Gebiet $\Omega \subset \mathbb{C}$. Es sei Γ ein einfach geschlossener, in Ω nullhomologer Weg (z.B. der Rand eines Kreises in $\mathbb{C}$), so dass gilt

$$|f(\zeta) - g(\zeta)| < |g(\zeta)| \quad \text{für alle } \zeta \in \Gamma.$$

Dann haben f und g gleich viele Nullstellen im Inneren von Γ.

Vergleiche Remmert, Funktionentheorie I. Grundlehren, Springer, 4. Auflage; Seite 310f.
Beweis: Sei $p_m^*(z) = \frac{z^m}{\gamma^m}$. Sei $p_m \in \Pi_m,\ p_m(\gamma) = 1$ ein Polynom mit

$$\max_{\lambda \in D(0,\rho)} |p_m(\lambda)| < \max_{\lambda \in D(0,\rho)} |p_m^*(\lambda)|.$$

Dann gilt

$$|(p_m^*(z) - p_m(z)) - p_m^*(z)| = |p_m(z)| < \left|\frac{\rho}{\gamma}\right|^m = |p_m^*(z)| \text{ für } z \in \partial D(0,\rho).$$

Also besitzen $p_m^* - p_m$ und p_m^* gleichviele Nullstellen im Inneren von $D(0,\rho)$, nämlich m Stück. Außerdem gilt $(p_m^* - p_m)(\gamma) = 1 - 1 = 0$, also besitzt das Polynom $p := p_m^* - p_m \in \Pi_m$ mindestens $m+1$ Nullstellen, ist also nach dem Fundamentalsatz der Algebra das Nullpolynom. □

2.5 Aufgabe
Finde ausgehend von Satz 2.3 die Lösung von

$$\min_{p_m \in \overline{\Pi}_m,} \max_{\lambda \in D(1,\rho)} |p_m(\lambda)|.$$

Antwort: s. Übung.

Wir erhalten so die folgende Interpretation: Bezüglich der Information

$$\text{spek}(A) \subseteq D(1,\rho)$$

mit $\rho < 1$ ist das Richardson-Verfahren *optimal.*

3 Die Tschebyscheff-Polynome

3.1 Definition
Die *Tschebyscheff-Polynome* $T_m(z)$ sind gegeben durch die Rekursion

$$T_{m+1}(z) = 2zT_m(z) - T_{m-1}(z), \quad m \geq 1,$$

mit

$$T_0(z) = 1, \quad T_1(z) = z.$$

Es ist bekannt: Die $T_m(z)$ sind Orthogonalpolynome bezüglich des Innenproduktes

$$\langle p, q \rangle = \int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} p(x)q(x)\ dx, \qquad \text{auf } \Pi \text{= Menge der Polynome.}$$

3.2 Satz
Die Lösung p_m der Minimierungsaufgabe

$$\min_{p_m \in \Pi_m, p_m(c)=1} \quad \max_{\lambda \in [-1,1]} |p_m(\lambda)|,$$

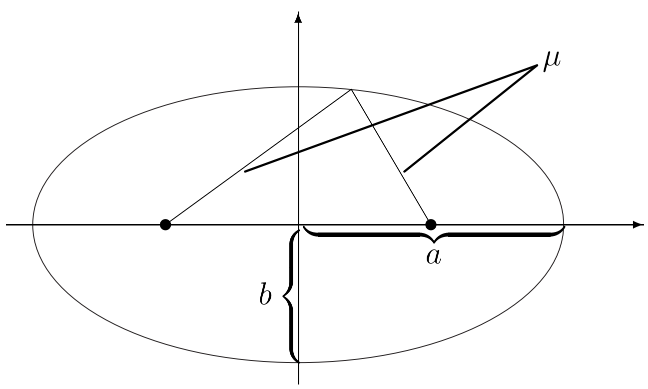
für $c \notin [-1, 1]$ lautet:

$$p_m(z) = \frac{1}{T_m(c)} T_m(z).$$

Beweis: Siehe Vorlesung: Parallele Algorithmen (Vorlesung im Wintersemester 2002/03, Abschnitt 4.4). □

Wir interessieren uns nun für Optimalpolynome auf etwas allgemeineren Gebieten, nämlich Ellipsen.

3.3 Definition
Für $\rho \geq 0$ sei E_ρ die Ellipse mit Brennpunkten in -1 und $+1$ und mit großer Halbachse $a = (1/2) \cdot (\rho + 1/\rho)$ und kleiner Halbachse $b = (1/2) \cdot (\rho - 1/\rho)$. Die „Schnurlänge“ μ beträgt also $\rho + 1/\rho$.



Die hier gegebene Definition einer Ellipse über ihre Brennpunkte und den Parameter ρ ist nicht Standard. Man rechnet aber einfach nach, dass so tatsächlich eine Ellipse gegeben ist, indem man nachweist, dass die Summe der Abstände zu den beiden Brennpunkten für jeden Punkt auf E_ρ gerade die Schnurlänge μ ist.

3.4 Definition
Die *Joukowski-Abbildung* $J : \mathbb{C} \setminus \{0\} \longrightarrow \mathbb{C}$ ist definiert durch

$$J(w) = \frac{1}{2}(w + w^{-1}).$$

Mit Hilfe dieser Abbildung wollen wir Folgendes zeigen:

3.5 Lemma
Es ist $J(D(0, \rho)) = E_\rho$.

Beweis: Wir stellen als erstes die Gleichung für E_ρ auf ($z = x + iy$):

$$\frac{x^2}{\left(\frac{\rho+1/\rho}{2}\right)^2} + \frac{y^2}{\left(\frac{\rho-1/\rho}{2}\right)^2} = 1.$$

Eine Parametrisierung von E_ρ ist dann

$$x(t) = \frac{\rho + 1/\rho}{2} \cdot \cos t, \quad y(t) = \frac{\rho - 1/\rho}{2} \cdot \sin t$$

mit $t \in [0, 2\pi)$. Dann gilt:

$$\begin{aligned} x^2 + y^2 &= \left(\frac{\rho + 1/\rho}{2}\right)^2 \cos^2 t + \left(\frac{\rho - 1/\rho}{2}\right)^2 \sin^2 t \\ &= \frac{1}{4} \left(\rho + (1/\rho)\right)^2 - \sin^2 t. \end{aligned}$$

Andererseits gilt mit $w = \rho e^{i\theta} \in D(0, \rho)$ für das Bild $z = J(w) = \frac{1}{2}(w + w^{-1})$

$$\begin{aligned} z\overline{z} &= \frac{1}{4}(w + w^{-1})(\overline{w} + \overline{w}^{-1}) \\ &= \frac{1}{4}(w\overline{w} + w^{-1}\overline{w} + w\overline{w}^{-1} + w^{-1}\overline{w}^{-1}) \\ &= \frac{1}{4}\left(\rho^2 + e^{-2i\theta} + e^{2i\theta} + \frac{1}{\rho^2}\right) \\ &= \frac{1}{4}\left(\rho^2 + 2\cos(2\theta) + \frac{1}{\rho^2}\right) \end{aligned}$$

$$= \frac{1}{4}\left(\rho^2 - 4\sin^2(\theta) + 2 + \frac{1}{\rho^2}\right)$$
$$= \frac{1}{4}\left(\rho^2 + 2 + \frac{1}{\rho^2}\right) - \sin^2(\theta)$$
$$= \frac{1}{4}\left(\rho + \frac{1}{\rho}\right)^2 - \sin^2(\theta).$$

Vergleich mit der Parameterdarstellung ergibt einfach

$$\theta = t.$$

□

Die Gleichung $z = J(w) = \frac{1}{2}(w + w^{-1})$ hat zwei Lösungen w, die zueinander invers sind. Sei nun $w(z)$ eine dieser Lösungen z.B. die betragsmäßig größere.

3.6 Lemma
Es gilt:

$$T_m(z) = \left[\frac{1}{2}\left(w^m(z) + w^{-m}(z)\right)\right].$$

Beweis: Wir beweisen dies natürlich über die Rekursionsformel. Für $m = 0, 1$ ist die Behauptung offensichtlich richtig und per Induktion gilt:

$$\begin{aligned} T_{m+1}(z) &= z\left(w^m(z) + w^{-m}(z)\right) - \frac{1}{2}\left(w^{m-1}(z) + w^{-(m-1)}(z)\right) \\ &= \frac{1}{2}\left(w(z) + w^{-1}(z)\right)\left(w^m(z) + w^{-m}(z)\right) \\ &\qquad - \frac{1}{2}\left(w^{m-1}(z) + w^{-(m-1)}(z)\right) \\ &= \frac{1}{2}\left(w^{m+1}(z) + w^{-(m+1)}(z)\right). \end{aligned}$$

□

Betrachtet wird nun die MinMax-Aufgabe:

(3.1) $$\min_{p \in \Pi_m, p(\gamma)=1} \max_{\lambda \in E_\rho} |p(\lambda)| \quad \text{für } \gamma \notin E_\rho.$$

3.7 Satz
Sei c_m der Wert des Minimums in (3.1) und $\gamma = \frac{1}{2}(w_\gamma + w_\gamma^{-1})$, wobei w_γ die betragsmäßig größere der beiden möglichen Lösungen ist. Dann gilt:

$$\frac{\rho^m}{|w_\gamma|^m} \le c_m \le \frac{1}{|w_\gamma^m + w_\gamma^{-m}|}\left(\rho^m + \left(\frac{1}{\rho}\right)^m\right),$$

wobei die obere Schranke für das Polynom

$$p_m^*(z) = \frac{1}{T_m(\gamma)} T_m(z)$$

erreicht wird.

Beweis: Wir stellen ein beliebiges Polynom $p_m \in \overline{\Pi}_m$ in der Basis der Tschebyscheff-Polynome dar, d.h.

$$p_m = \sum_{\nu=0}^{m} a_\nu T_\nu.$$

Wegen $p_m(\gamma) = 1$ gilt $1 = \sum_{\nu=0}^{m} a_\nu T_\nu(\gamma)$. Unter Verwendung von $w = w(z)$ wie im Lemma 3.6 ergibt sich

$$\begin{aligned} p_m(z) &= \sum_{\nu=0}^{m} a_\nu \left(\frac{1}{2} \left(w^\nu + w^{-\nu} \right) \right) \\ &= \sum_{\nu=0}^{m} a'_\nu \left(\frac{\frac{1}{2}\left(w^\nu + w^{-\nu} \right)}{\frac{1}{2}\left(w_\gamma^\nu + w_\gamma^{-\nu} \right)} \right). \end{aligned}$$

Hierin ist $\sum_{\nu=0}^{m} a'_\nu = 1$. Wählen wir speziell $a'_m = 1$, $a'_\nu = 0$,$\nu = 0, ..., m-1$, also

$$p_m(z) = p_m^*(z) = \frac{1}{T_m(\gamma)} T_m(z) = \frac{w^m + w^{-m}}{w_\gamma^m + w_\gamma^{-m}},$$

so gilt wegen des Maximums-Prinzips

$$\begin{aligned} \max_{z \in E_\mu} |p_m^*(z)| &= \max_{z \in \partial E_\mu} |p_m^*(z)| \\ &= \max_{w \in \partial D(0,\rho)} \left| \frac{w^m + w^{-m}}{w_\gamma^m + w_\gamma^{-m}} \right| \\ &= \frac{1}{|w_\gamma^m + w_\gamma^{-m}|} \left[\rho^m + \left(\frac{1}{\rho} \right)^m \right]. \end{aligned}$$

Die letzte Gleichheit ist richtig, weil $|w^m + w^{-m}| \leq |w^m| + |w^{-m}| = |\rho^m| + |\rho^{-m}|$, und diese obere Schranke wird für $w = \rho$ auch angenommen.
Für die Bestimmung einer unteren Schranke sei das Optimalpolynom $\widetilde{p}_m(z)$ dargestellt durch

$$\widetilde{p}_m(z) = \sum_{\nu=0}^{m} b_\nu \left(\frac{w^\nu + w^{-\nu}}{w_\gamma^\nu + w_\gamma^{-\nu}} \right), \qquad \sum_{\nu=0}^{m} b_\nu = 1$$

$$= \left(\frac{w^{-m}}{w_\gamma^{-m}}\right) \underbrace{\sum_{\nu=0}^{m} b_\nu \left(\frac{w^{\nu+m} + w^{-\nu+m}}{w_\gamma^{\nu+m} + w_\gamma^{-\nu+m}}\right)}_{\text{Polynom } q \text{ vom Grad } 2m \text{ in } w,\ q(w_\gamma)=1} .$$

Nach Satz 2.3 gilt

$$\max_{w\in D(0,\rho)} \left|\sum_{\nu=0}^{m} b_\nu \left(\frac{w^{\nu+m} + w^{-\nu+m}}{w_\gamma^{\nu+m} + w_\gamma^{-\nu+m}}\right)\right| \geq \left(\frac{\rho}{|w_\gamma|}\right)^{2m},$$

falls w_γ die betragsmäßig größere Lösung ist (und damit außerhalb von $D(0,\rho)$ liegt). Also gilt

$$|\widetilde{p}_m(z)| \geq \left|\frac{w^{-m}}{w_\gamma^{-m}}\right| \left(\frac{\rho}{|w_\gamma|}\right)^{2m} = \frac{|w^{-m}|\,\rho^{2m}}{|w_\gamma|^m} = \frac{\rho^m}{|w_\gamma|^m}.$$

$\square$

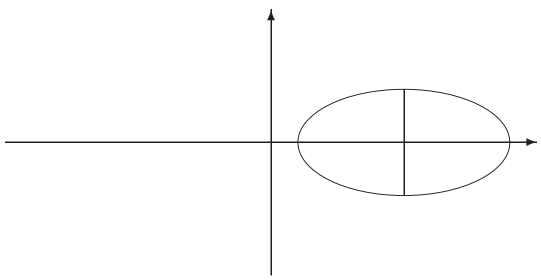
3.8 Bemerkung
Für $\gamma \notin E_\rho$ gilt $|w_\gamma| < \rho^{-1}$ mit $\rho \geq 1$. Also ist sowohl $\lim\limits_{m\to\infty}(1/\rho)^m = 0$ als auch $\lim\limits_{m\to\infty} w_\gamma^{-m} = 0$. Satz 3.7 zeigt so, dass die Tschebyscheff-Polynome *asymptotisch optimal* sind, aber nicht notwendig optimal für endliches m.

3.9 Aufgabe
Bestimme jeweils die (im Sinne von Satz 3.7) asymptotisch optimale Lösung der MinMax-Aufgabe

$$\min_{p_m\in\overline{\Pi}_m} \max_{\lambda\in E}|p(\lambda)|$$

und den zugehörigen Wert des Minimums für die Fälle:

1. $E = a + bE_\rho$, wobei $a, b > 0$ und $a - b\cdot\frac{1}{2}(\rho + 1/\rho) > 0$



2. $E = a + be^{i\frac{\pi}{2}}E_\rho$, wobei $a, b > 0$ und $a - b\cdot\frac{1}{2}(\rho - 1/\rho) > 0$

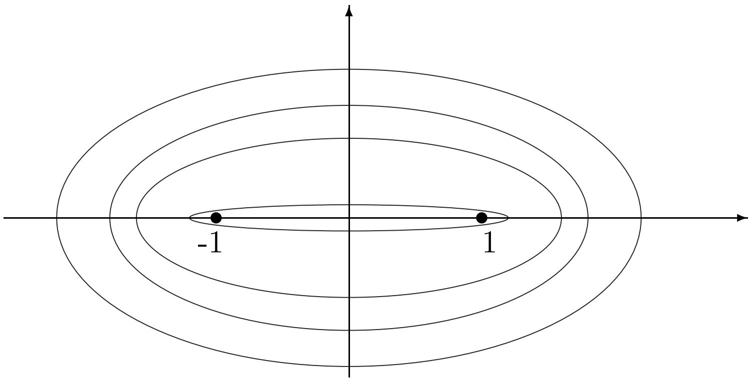
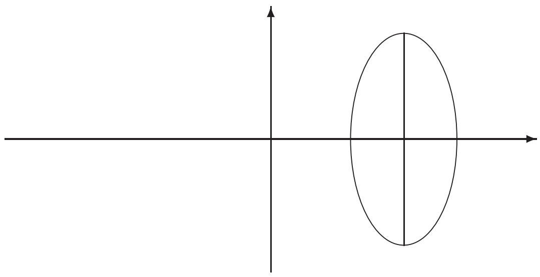
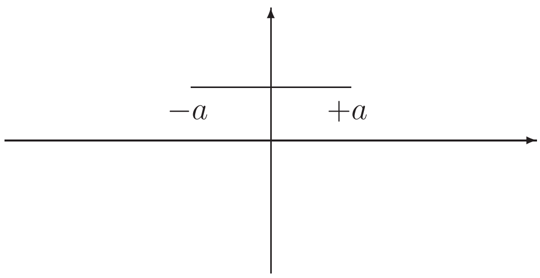


Abbildung 1: Ellipsen E_ρ für verschiedene Werte von ρ



3. E_ρ ist ein Geradenstück parallel zur reellen Achse von $-a$ bis $+a$ mit Achsenabschnitt b auf der imaginären Achse



Sind in diesem Fall die Polynom auch optimal für endliches m?

3.10 Bemerkung
Die Optimalpolynome im Sinne von Satz 3.7 sind unabhängig vom Parameter ρ der Ellipse E_ρ (solange γ immer noch außerhalb der Ellipse liegt). Aus Satz 2.3 ist bekannt, dass für den Kreis $D(0, \rho)$ die Minimax-Aufgabe

$$\min_{p \in \Pi_m, p(\gamma)=1} \max_{\lambda \in D(0,\rho)} |p_m(\lambda)|$$

durch das Polynom $p_m(z) = \frac{z^m}{\gamma^m}$ gelöst wird. Für dessen Ableitungen $p_m^{(k)}(z) = m(m-1)\cdot\ldots\cdot(m-k+1)\frac{z^{m-k}}{\gamma^m}$ gilt

$$\begin{aligned}\max_{z\in D(0,\rho)} \left|p_m^{(k)}(z)\right| &= m\cdot\ldots\cdot(m-k+1)\frac{\rho^{m-k}}{\gamma^m}\\ &= m\cdot\ldots\cdot(m-k+1)\rho^{-k}\left(\frac{\rho}{\gamma}\right)^m \underset{m\to\infty}{\longrightarrow} 0,\end{aligned}$$

wobei die Konvergenz sogar gleichmäßig ist auf $D(0,\rho)$. Dies ist wichtig mit Blick auf Satz 2.1

Für die Tschebyscheffpolynome und Ellipsen gilt ein entsprechendes Resultat.

3.11 Satz
Es sei $E_\rho = J(D(0,\rho))$ mit $\rho > 1$ eine Ellipse. Weiter sei p_m^* das m-te skalierte Tschebyscheffpolynom, $p_m^*(z) = \frac{w^m+w^{-m}}{w_\gamma^m+w_\gamma^{-m}}$, wobei $z = J(w) = \frac{1}{2}(w+w^{-1})$ und $J(w_\gamma) = \gamma$, mit γ außerhalb von E_ρ, d.h. $|w_\gamma| > \rho$. Dann gilt für alle $k \in \mathbb{N}$

$$\left|p_m^{*\,(k)}(z)\right| \underset{m\to\infty}{\longrightarrow} 0 \quad \text{gleichmäßig auf } E_\rho.$$

Beweis: Beachte (im Vorgriff auf die Formulierung des Tschebyscheff-Verfahrens in Algorithmus 4.7): Im Zusammenhang mit Satz 2.1 besagt dieser Satz, dass Diagonalisierbarkeit für ein erfolgreiches Tschebyscheffverfahren nicht vorausgesetzt werden muss.
Mit

$$\begin{aligned}p_m^*(z) &= \frac{w^m + w^{-m}}{w_\gamma^m + w_\gamma^{-m}},\\ \frac{dz}{dw} &= \frac{1}{2}(1-w^{-2}) \text{ und } \frac{dw}{dz}|_{z=J(w)} = \frac{2}{1-w^{-2}},\end{aligned}$$

erhalten wir

$$\begin{aligned}\frac{dp_m^*(z)}{dw}\cdot\frac{dw}{dz} &= \frac{mw^{m-1} - mw^{-(m+1)}}{w_\gamma^m + w_\gamma^{-m}}\cdot\frac{2}{w1w^{-2}}\\ &= 2\frac{mw^{m+1} - mw^{-m+1}}{(w_\gamma^m + w_\gamma^{-m})(w^2-1)}.\end{aligned}$$

Der Term $s_m(w) = m\frac{w^{m+1}}{(w^2-1)(w_\gamma^m+w_\gamma^{-m})}$ wird für $|w| = \rho$ abgeschätzt durch:

$$|s_m(w)| \le m\frac{1}{\rho^2-1}\frac{\rho^{m+1}}{\left|w_\gamma^m + w_\gamma^{-m}\right|} \le \frac{m}{\rho^2-1}\frac{\rho}{|w_\gamma/\rho|^m - |w_\gamma\rho|^{-m}}$$

Hierin ist die rechte obere Schranke unabhängig von $w \in \partial D(0,\rho)$. Wegen des Maximumsprinzips gilt die Schranke sogar für alle $w \in D(0,\rho)$, also für alle $z \in E_\rho$. Aus

$$|w_\gamma/\rho| > 1, \ \ |w_\gamma \rho| > 1,$$

erkennen wir mit der Regel von de l'Hôpital, dass die obere Schranke für $m \to \infty$ gegen 0 geht. Dies beweist den Satz für die erste Ableitung; für höhere Ableitungen geht der Beweis (induktiv) analog. □

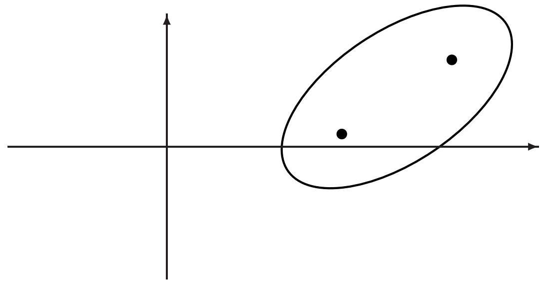


Abbildung 2: spek(A) liegt in einer Ellipse, welche 0 nicht enthält

4 Das Tschebyscheff - Verfahren

4.1 Definition
Mit $E(f_1, f_2, \rho)$ bezeichnen wir die Ellipse mit den Brennpunkten f_1, f_2 und den Halbachsen $\frac{1}{4}|f_1 - f_2|\left(\rho + \frac{1}{\rho}\right)$ und $\frac{1}{4}|f_1 - f_2|\left(\rho - \frac{1}{\rho}\right)$, $\rho \geq 1$.

Wir betrachten nun folgende Situation:

Gegeben: $Ax = b$, A regulär.
Bekannt sei: spek$(A) \subseteq E(f_1, f_2, \rho)$ mit $0 \notin E(f_1, f_2, \rho)$.
Mit Hilfe der Tschebyscheff-Polynome werden wir (asymptotisch optimale) Näherungslösungen für die MinMax-Aufgabe

$$\min_{p_m \in \overline{\Pi}} \max_{\lambda \in E(f_1, f_2, \rho)} |p_m(\lambda)| \tag{4.1}$$

finden. Dafür benötigen wir zunächst die affine Transformation, welche unsere Standard-Ellipse $E_\rho = E(-1, 1, \rho)$ auf $E(f_1, f_2, \rho)$ abbildet.
Wir betrachten hierzu die affinen Grundabbildungen

$$\begin{aligned} D_\phi &: z \mapsto e^{i\phi} z, \phi \in \mathbb{R} \ \text{(Drehung)} \\ T_a &: z \mapsto z + a, a \in \mathbb{C} \ \text{(Transalation)} \\ S_c &: z \mapsto cz, c > 0 \ \text{(Skalierung)} \end{aligned}$$

4.2 Lemma
Die Abbildung

$$\begin{aligned} \varphi &= T_{\frac{1}{2}(f_1+f_2)} \circ D_{\arg(f_2-f_1)} \circ S_{\frac{1}{2}|f_1-f_2|} \\ \varphi &: z \mapsto e^{i\arg(f_2-f_1)} \frac{1}{2}|f_2 - f_1|\, z + \frac{f_1+f_2}{2} = \underbrace{\frac{f_2-f_1}{2}}_{\alpha} z + \underbrace{\frac{f_1+f_2}{2}}_{\beta} \end{aligned}$$

bildet die Ellipse $E(-1, 1, \rho)$ ab auf $E(f_1, f_2, \rho)$.

Damit können wir nun Tschebyscheff Polynome bezüglich allgemeiner Ellipsen $E(f_1, f_2, \rho)$ definieren.

4.3 Definition
Sei $E = E(f_1, f_2, \rho)$ eine Ellipse mit $0 \notin E$. Dann ist

$$T_m^E(z) = T_m\left(\frac{1}{\alpha}(z - \beta)\right)$$

mit $\alpha = \frac{1}{2}(f_2 - f_1)$ und $\beta = \frac{1}{2}(f_1 + f_2)$ das *Tschebyscheff-Polynom* vom Grad m bzgl. E.

Nach Satz 3.7 ist damit klar, dass

$$p_m(z) = \frac{T_m^E(z)}{T_m^E(0)}$$

eine approximative, asymptotisch optimale, Lösung für die MinMax-Aufgabe (4.1) ist.

4.4 Definition
Das zu p_m gehörige KUV heißt *Tschebyscheff-Verfahren*.

Wir werden nun dieses Verfahren konkret herleiten. Hierfür betrachten wir den Fall einer allgemeinen 3-Term-Rekursion

$$p_{m+1}(t) = \alpha_{m+1}(t + \beta_{m+1})p_m(t) + \gamma_{m+1}p_{m-1}(t),\ m \geq 1.$$

Wegen $p_m(0) = 1,\ \forall\ m$, folgt $\gamma_m = 1 - \alpha_m\beta_m$ und insbesondere

$$p_0(t) = 1,\ \ p_1(t) = \alpha_1\left(t + \frac{1}{\alpha_1}\right).$$

Es gilt also für die Residuen $r^m = b - Ax^m = p_m(A)r^0$:

$$\begin{aligned} r^{m+1} &= \alpha_{m+1}(A + \beta_{m+1}I)r^m + (1 - \alpha_{m+1}\beta_{m+1})r^{m-1},\ m \geq 1, \\ r^0 &= r^0, \\ r^1 &= \alpha_1\left(A + \frac{1}{\alpha_1}I\right)r^0 = \alpha_1 Ar^0 + r^0. \end{aligned}$$

Damit erhalten wir für die Iterierten unter Einsetzen von $r^m = b - Ax^m$ die Rekursion

$$\begin{aligned} x^0 &= x^0, \\ x^1 &= x^0 - \alpha_1 r^0, \\ x^{m+1} &= \alpha_{m+1}\beta_{m+1}x^m + (1 - \alpha_{m+1}\beta_{m+1})x^{m-1} - \alpha_{m+1}r^m. \end{aligned}$$

Damit erhalten wir als allgemeine Struktur für ein KUV mit 3-Term-Rekursion für die Residuen den folgenden Algorithmus:

4.5 Algorithmus

wähle x^0, setze $r^0 = b - Ax^0$
$x^1 = x^0 - \alpha_1 r^0$
$r^1 = \alpha_1 A r^0 + r^0$
for $m = 1, 2, ...$ **do**
  $x^{m+1} = \alpha_{m+1}\beta_{m+1}x^m + (1 - \alpha_{m+1}\beta_{m+1})x^{m-1} - \alpha_{m+1}r^m$
  $r^{m+1} = b - Ax^{m+1}$
end for

(Alternativ könnte man r^{m+1} auch aus der 3-Term-Rekursion bestimmen)
Jetzt geht es um die konkrete Bestimmung der Koeffizienten α_m, β_m beim Tschebyscheff-Verfahren.
Mit $E = E(f_1, f_2, \rho)$, $a = \frac{1}{2}(f_2 - f_1)$ und $b = \frac{1}{2}(f_2 + f_1)$ gilt

$$T_m^E = T_m\left(\frac{1}{a}(z-b)\right).$$

Es gilt $T_0 = 1$, $T_1 = z$ und $T_{m+1} = 2zT_m - T_{m-1}$. Also folgt

$$\begin{aligned} T_0^E &= 1, \\ T_1^E &= \frac{1}{a}(z-b) \\ \text{und} \qquad T_{m+1}^E &= T_{m+1}\left(\frac{1}{a}(z-b)\right) = \frac{2}{a}(z-b)T_m^E(z) - T_{m-1}^E(z). \end{aligned}$$

Die Residuenpolynome $p_m(z)$ für das Tschebyscheff-Verfahren sind

$$p_m(z) = \frac{T_m^E(z)}{T_m^E(0)}.$$

Wir benötigen also $c_m := T_m^E(0)$ und finden

$$(4.2) \qquad \left\{ \begin{aligned} c_0 &= 1, \\ c_1 &= -\frac{b}{a} \\ \text{und} \qquad c_{m+1} &= -\frac{2}{a}bc_m - c_{m-1},\ m \geq 1. \end{aligned} \right.$$

Damit erhalten wir

$$\begin{aligned} p_0(z) &= 1, \\ p_1(z) &= -\frac{1}{b}(z-b) \\ \text{und} \qquad p_{m+1}(z) &= \frac{1}{c_{m+1}}\left(\frac{2}{a}(z-b)c_m p_m(z) - c_{m-1}p_{m-1}(z)\right),\ m \geq 1. \end{aligned}$$

4.6 Bemerkung
Es gibt verschiedene andere Formulierungen, mit denen die Koeffizienten aufdatiert werden können. Die Werte c_m werden schnell sehr groß; deshalb ist es aus numerischen Gründen besser, die Berechnung von c_{m+1} zu vermeiden, und dafür die Quotienten $\delta_{m+1} = \frac{c_{m+1}}{c_m}$ aufzudatieren. Wir erhalten

$$\frac{c_{m+1}}{c_m} = -\frac{2}{a}b - \frac{c_{m-1}}{c_m},$$

und damit

$$\begin{aligned} & \delta_1 = c_1 \\ \text{und} \qquad & \delta_{m+1} = -\frac{2}{a}b - \delta_m^{-1},\ m \geq 0. \end{aligned}$$

Wegen $\frac{c_{m-1}}{c_{m+1}} = 1 - \frac{2b}{a}\frac{c_m}{c_{m+1}}$ (dies gilt, da $p_m(0) = 1$ für alle m) ergibt sich

$$\begin{aligned} & p_0(z) = 1, \\ & p_1(z) = -\frac{1}{b}(z - b) \\ \text{und} \qquad & p_{m+1}(z) = \frac{2}{a\delta_{m+1}}(z-b)p_m(z) - \left(1 - \frac{2b}{a\delta_{m+1}}\right)p_{m-1}(z). \end{aligned}$$

Die Koeffizienten der 3-Term-Rekursion sind also

$$\alpha_m = \frac{2}{a\delta_m},\ \beta_m = -b; m \geq 2.$$

Damit erhalten wir den folgenden Algorithmus für das Tschebyscheff-Verfahren zur Lösung von $Ax = c$ (b ist schon vergeben ...):

4.7 Algorithmus

Require: spek$(A) \subseteq E(f_1, f_2, \rho) \not\ni 0$
  setze $a = \frac{1}{2}(f_2 - f_1)$, $b := \frac{1}{2}(f_2 + f_1)$
  setze $\delta_1 = -b/a$
  wähle x^0
  setze $r^0 = c - Ax^0$
  $\alpha_1 = -2/b$
  $x^1 = x^0 - \alpha_1 r^0$
  $r^1 = \alpha_1 A r^0 + r^0$
  for $m = 1, 2, \ldots$ **do**
    $\delta_{m+1} = -\frac{2}{a} b - 1/\delta_m$
    $\alpha_{m+1} = 2/(a \cdot \delta_{m+1})$
    $x^{m+1} = (-\alpha_{m+1} b) x^m + (1 + \alpha_{m+1} b) x^{m-1} - \alpha_{m+1} r^m$
    $r^{m+1} = c - Ax^{m+1}$
  end for

4.8 Bemerkung
Wegen $p_m(0) = 1$ muss auch gelten

$$1 = \frac{c_m}{c_{m+1}} \cdot \frac{2}{a}(-b) - \frac{c_{m-1}}{c_{m+1}},$$

was man mit (4.2) bestätigt.

4.9 Aufgabe
Implementiere das Tschebyscheff-Verfahren für das Modellproblem I für verschiedene Gittergrößen N.
Verwende dabei spek$(A) \subseteq [\xi, \Xi]$ mit

$$\xi = 4 - 4\cos(\pi h), \qquad h = \frac{1}{N+1},$$
$$\text{und} \qquad \Xi = 4 + 4\cos(\pi h).$$

Die zugehörige Ellipse ist also

$$E(\xi, \Xi, 1)$$

- Plotte $\|r^m\|$ gegen m für verschiedene Werte von N.
- Teste auch 2Ξ statt Ξ (langsamere Konvergenz).
- Teste auch $\Xi - \xi$ statt Ξ (keine Konvergenz).
- Teste auch $\xi + 10 \cdot h^2$ statt ξ .
- Teste auch $\frac{\xi+\Xi}{2}$ statt ξ (langsamere Konvergenz).

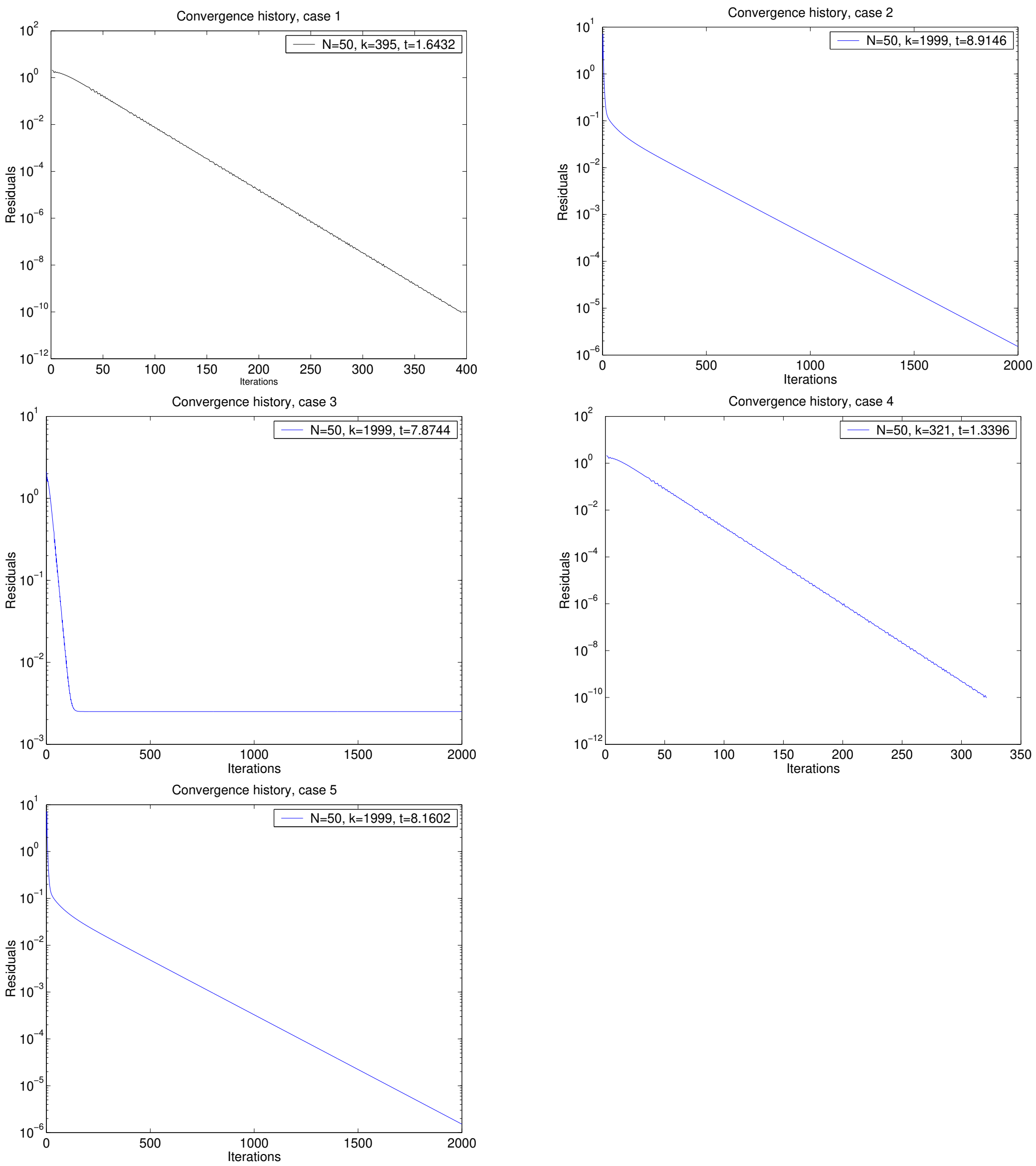


Abbildung 3: Numerische Resultate für das Modellproblem auf einem 50×50-Gitter

5 Analyse von Tschebyscheff- und Richardson-Verfahren

5.1 Satz
Sei p_m das Polynom des Tschebyscheff-Verfahrens bzgl. $E(f_1, f_2, \rho) \not\ni 0$ mit $\operatorname{spek}(A) \subseteq E(f_1, f_2, \rho)$. Dann gilt

$$r^m = p_m(A)r^0$$

und damit für jede Norm $\|\cdot\|$

$$\|r^m\| \leq \|p_m(A)\| \cdot \left\|r^0\right\|.$$

1. Es sei A diagonalisierbar mit $A = SDS^{-1}$ (siehe 2.2) und $\|\cdot\|$ sei die Norm $\|\cdot\|_S$. Dann gilt

$$\|p_m(A)\| \leq \frac{\rho^m + \rho^{-m}}{\left|w_\gamma^m + w_\gamma^{-m}\right|}$$

wobei w_γ (betragsgrößte) Lösung von $\frac{1}{2}(w + w^{-1}) = -b/a$.

2. Ist in 1. sogar S orthogonal (z.B. weil A normal ist), so gilt sogar

$$\|p_m(A)\|_2 \leq \frac{\rho^m + \rho^{-m}}{\left|w_\gamma^m + w_\gamma^{-m}\right|}.$$

Beweis: Ist q_m die approximative Optimallösung für die Ellipse $E(-1, 1, \rho)$ mit $q_m(\gamma) = 1$, dann gilt nach Satz 3.7

$$\max_{t \in E(-1,1,\bar{\rho})} |q_m(t)| = \frac{\bar{\rho}^m + \bar{\rho}^{-m}}{\left|w_\gamma^m + w_\gamma^{-m}\right|}.$$

Mit der affinen Transformation $z \mapsto az+b$ aus Lemma 4.2, welche $E(-1, 1, \rho)$ auf $E(f_1, f_2, \rho)$ abbildet, ergibt sich der Satz. □

Besonders wichtig ist der Fall $\operatorname{spek}(A) \subseteq [\xi, \Xi] \subseteq (0, +\infty)$ und A symmetrisch und positiv definit:

5.2 Korollar
Ist $\operatorname{spek}(A) \subseteq [\xi, \Xi] \subseteq (0, +\infty)$ und A spd, dann gilt für die Residuen des Tschebyscheff-Verfahrens

$$\|r^m\|_2 \leq \|p_m(A)\|_2 \cdot \left\|r^0\right\|_2$$

mit

$$\|p_m(A)\|_2 \leq \frac{2c^m}{1 + c^{2m}} \quad \text{mit } c = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1},\ \kappa = \frac{\Xi}{\xi}.$$

Beweis: Hier ist $f_1 = \xi$, $f_2 = \Xi$, $\rho = 1$, $a = \frac{1}{2}(\Xi - \xi), b = \frac{1}{2}(\Xi + \xi)$ und $\gamma = -b/a$. Aus Satz 5.1 folgt so

$$\|p_m(A)\|_2 \leq \frac{2}{\left|w_\gamma^m + w_\gamma^{-m}\right|}$$

mit

$$\frac{1}{2}(w_\gamma + w_\gamma^{-1}) = -\frac{b}{a} = -\frac{\Xi + \xi}{\Xi - \xi} = -\frac{\kappa + 1}{\kappa - 1}.$$

Auflösen der quadratischen Gleichung ergibt

$$\begin{aligned} w_\gamma &= -\frac{\kappa+1}{\kappa-1} - \sqrt{\left(\frac{\kappa+1}{\kappa-1}\right)^2 - 1} \\ &= -\frac{\kappa+1}{\kappa-1} - \sqrt{\frac{4\kappa}{(\kappa-1)^2}} \\ &= -\frac{\kappa+1}{\kappa-1} - \frac{2\sqrt{\kappa}}{\kappa-1} \\ &= -\frac{\left(\sqrt{\kappa}+1\right)^2}{(\sqrt{\kappa}-1)(\sqrt{\kappa}+1)} \\ &= -\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}. \end{aligned}$$

$\square$

Zum Abschluss (und Vergleich) analysieren wir noch das Richardson-Verfahren 2. Ordnung. Dazu sei A spd mit $\text{spek}(A) \subseteq [\xi, \Xi]$

5.3 Aufgabe
Für das relaxierte Richardson-Verfahren 1. Ordnung

$$x^{m+1} = x^m + \alpha r^m$$

wird für die zugehörigen Polynome $p_m(z) = (1 - \alpha z)^m$ die Größe

$$\max_{z \in [\xi, \Xi]} |p_m(z)|$$

minimal, wenn $\alpha = 2/(\xi + \Xi)$.

5.4 Bemerkung
Das relaxierte Richardson Verfahren 1. Ordnung $Ax = b$ ist äquivalent zum gewöhnlichen Richardson-Verfahren 1. Ordnung für

$$(\alpha A)x = \alpha b,$$

also

$$x^{m+1} = x^m + (\alpha b - \alpha A x^m).$$

Wir gehen ab jetzt davon aus, dass diese „Skalierung“ des LGS (mit $\alpha = \frac{2}{\xi+\Xi}$) bereits erfolgte, d.h. es gilt ab jetzt

$$\operatorname{spek}(A) \subseteq [1-\gamma, 1+\gamma] \text{ mit } \gamma = \frac{\Xi - \xi}{\Xi + \xi} \in [0,1).$$

Für das System

$$Ax = b$$

betrachten wir jetzt das Richardson-Verfahren 2. Ordnung

$$\begin{aligned} x^{m+1} &= \omega\big((I-A)x^m + b\big) + (1-\omega)x^{m-1} \\ &= \omega(x^m + r^m) + (1-\omega)x^{m-1} \end{aligned}$$

mit

$$\begin{aligned} r^{m+1} &= \omega r^m - \omega A r^m + (1-\omega) r^{m-1}, \\ p_{m+1} &= \omega(1-t)p_m(t) + (1-\omega)p_{m-1}(t). \end{aligned}$$

Daraus folgt

$$\begin{pmatrix} p_{m+1}(t) \\ p_m(t) \end{pmatrix} = \underbrace{\begin{pmatrix} \omega(1-t) & 1-\omega \\ 1 & 0 \end{pmatrix}}_{B_\omega(t)} \begin{pmatrix} p_m(t) \\ p_{m-1}(t) \end{pmatrix}.$$

Wir wollen nun das „optimale“ ω ermitteln, d.h. wir suchen die Lösung ω von

$$\min_{\omega\in\mathbb{R}} \max_{t\in[1-\gamma,1+\gamma]} \rho\big(B_\omega(t)\big). \tag{5.1}$$

Zur besseren Darstellung machen wir den Übergang $t \to 1-t$, also

$$\begin{aligned} 1-t &\longrightarrow t \in [-\gamma, \gamma], \\ B_\omega(t) &= \begin{pmatrix} \omega(1-t) & 1-\omega \\ 1 & 0 \end{pmatrix} \longrightarrow \begin{pmatrix} \omega t & 1-\omega \\ 1 & 0 \end{pmatrix}. \end{aligned}$$

Wir suchen damit die Lösung ω der Minimierungsaufgabe

$$\min_{\omega\in\mathbb{R}} \max_{t\in[-\gamma,\gamma]} \rho\big(B_\omega(t)\big).$$

Nun bestimmen wir die Eigenwerte λ von $B_\omega(t)$. Es gilt nämlich

$$\begin{aligned} &(\lambda - \omega t)\lambda - (1-\omega) = 0 \\ \iff\ &\lambda^2 - (\omega t)\lambda - (1-\omega) = 0, \end{aligned}$$

d.h. wir erhalten

$$\lambda_{1,2} = \frac{\omega t}{2} \pm \sqrt{\frac{\omega^2 t^2}{4} + (1-\omega)}\,.$$

Wir machen zunächst eine Fallunterscheidung zur Berechnung von $\rho\big(B_\omega(t)\big)$.
Fall I: $\dfrac{\omega^2 t^2}{4} + (1-\omega) < 0$. Dann gilt

$$\begin{aligned} |\lambda_{1,2}|^2 &= \frac{(\omega t)^2}{4} + \frac{(\omega t)^2}{4} + (1-\omega) \\ &= \frac{(\omega t)^2}{2} + (1-\omega) \end{aligned}$$

Also gilt

$$\rho\big(B_\omega(t)\big) = \left(\frac{(\omega t)^2}{2} + (1-\omega)\right)^{1/2} \tag{5.2}$$

Fall II: $\dfrac{\omega^2 t^2}{4} + (1-\omega) \geq 0$.
(a) $\omega t \geq 0$: In diesem Fall haben wir

$$\rho\big(B_\omega(t)\big) = \frac{\omega t}{2} + \sqrt{\frac{\omega^2 t^2}{4} + (1-\omega)}$$

(b) $\omega t \leq 0$: Dann gilt

$$\rho\big(B_\omega(t)\big) = \left|\frac{\omega t}{2} - \sqrt{\frac{\omega^2 t^2}{4} + (1-\omega)}\right|$$

In beiden Fällen (a) und (b) erhalten wir aber

$$\rho\big(B_\omega(t)\big) = \frac{|\omega t|}{2} + \sqrt{\frac{\omega^2 t^2}{4} + (1-\omega)} \tag{5.3}$$

Sei nun ω fest. Wir bestimmen $\max\limits_{t\in[-\gamma,\gamma]} \rho\big(B_\omega(t)\big) =: \beta_\omega$.

1. Fall: $\omega \le 1$. Es kommt nur Fall II vor. Aus (5.3) erhalten wir, da $\rho\left(B_\omega(t)\right)$ eine gerade Funktion ist und monoton wachsend für $t \in [0, \gamma]$:

$$\beta_\omega = \frac{|\omega|\gamma}{2} + \sqrt{\frac{\omega^2\gamma^2}{4} + (1-\omega)}$$

2. Fall: $\omega > 1$. Jetzt kommen Fall I und Fall II beide vor. Es gilt

$$\frac{\omega^2 t^2}{4} + (1-\omega) \ge 0 \iff |t| \ge \frac{2\sqrt{\omega-1}}{\omega}.$$

Dann ist

$$\rho\big(B_\omega(t)\big) = \frac{\omega|t|}{2} + \sqrt{\frac{\omega^2 t^2}{4} + (1-\omega)}.$$

Im Fall $|t| \le \dfrac{2\sqrt{\omega-1}}{\omega}$ ist

$$\rho\big(B_\omega(t)\big) = \left(\frac{(\omega t)^2}{2} + (1-\omega)\right)^{1/2},$$

also ist

$$\beta_\omega = \max\left\{ \max_{|t| \le \frac{2\sqrt{\omega-1}}{\omega}} \left(\frac{(\omega t)^2}{2} + (1-\omega)\right)^{1/2}, \right.$$
$$\left. \max_{\frac{2\sqrt{\omega-1}}{\omega} \le |t| \le \gamma} \left(\frac{\omega|t|}{2} + \sqrt{\frac{(\omega t)^2}{4} + (1-\omega)}\right)\right\}.$$

Hierin kommt der erste Term nur vor, wenn $\gamma \ge \frac{2\sqrt{\omega-1}}{\omega}$. Aus Monotoniegründen folgt

$$\begin{aligned}
\max_{|t| \le \frac{2\sqrt{\omega-1}}{\omega}} \left(\frac{(\omega t)^2}{2} + (1-\omega)\right)^{1/2} &= \left(\frac{\left(\omega\frac{2\sqrt{\omega-1}}{\omega}\right)^2}{2} + (1-\omega)\right)^{1/2} \\
&= \sqrt{\omega-1}, \\
\max_{\frac{2\sqrt{\omega-1}}{\omega} \le |t| \le \gamma} \left(\frac{\omega|t|}{2} + \sqrt{\frac{(\omega t)^2}{4} + (1-\omega)}\right) &= \frac{\omega\gamma}{2} + \sqrt{\frac{(\omega\gamma)^2}{4} + (1-\omega)}.
\end{aligned}$$

Für $\gamma \ge \frac{2\sqrt{\omega-1}}{\omega}$ gilt

$$\frac{\omega\gamma}{2} + \sqrt{\frac{(\omega\gamma)^2}{4} + (1-\omega)} \ge \frac{\omega\frac{2\sqrt{\omega-1}}{\omega}}{2} + \sqrt{\frac{\left(\omega\frac{2\sqrt{\omega-1}}{\omega}\right)^2}{4} + (1-\omega)}$$

$$= \sqrt{\omega - 1}.$$

Damit gilt

$$\beta_\omega = \begin{cases} \sqrt{\omega - 1} & \text{falls } \gamma < \frac{2\sqrt{\omega-1}}{\omega}, \\ \dfrac{\omega\gamma}{2} + \sqrt{\dfrac{(\omega\gamma)^2}{4} + (1-\omega)} & \text{falls } \gamma \geq \frac{2\sqrt{\omega-1}}{\omega}. \end{cases}$$

Unter Beachtung von

$$\begin{aligned} & \gamma < \frac{2\sqrt{\omega - 1}}{\omega} \\ \Longleftrightarrow\ & \omega^2\gamma^2 < 4(\omega - 1) \\ \Longleftrightarrow\ & \frac{2 - 2\sqrt{1-\gamma^2}}{\gamma^2} < \omega < \frac{2 + 2\sqrt{1-\gamma^2}}{\gamma^2} \\ \Longleftrightarrow\ & \frac{2}{1 + \sqrt{1-\gamma^2}} < \omega < \frac{2}{1 - \sqrt{1-\gamma^2}} \end{aligned}$$

erhalten wir durch Zusammenfassen von 1. Fall und 2. Fall

$$\beta_\omega = \begin{cases} \sqrt{\omega - 1} & \text{falls } \frac{2}{1+\sqrt{1-\gamma^2}} < \omega < \frac{2}{1-\sqrt{1-\gamma^2}}, \\ \frac{\omega\gamma}{2} + \sqrt{\frac{(\omega\gamma)^2}{4} + (1-\omega)} & \text{sonst.} \end{cases}$$

Für $\omega \geq 2/\big(1 - \sqrt{1-\gamma^2}\big)$ (≥ 2) ist $\frac{\omega\gamma}{2} \geq \sqrt{\omega - 1} \geq 1$, also $\beta_\omega \geq 1$. Für $\omega < 2/\big(1 + \sqrt{1-\gamma^2}\big)$ ist

$$\beta_\omega = \frac{\omega\gamma}{2} + \sqrt{\frac{(\omega\gamma)^2}{4} + (1-\omega)}$$

monoton fallend. Die Größe β_ω wird also minimal für $\omega = \omega_0 = \frac{2}{1+\sqrt{1-\gamma^2}}$ und hat dann den Wert $\sqrt{\omega_0 - 1}$.
Zusammenfassend erhalten wir also:

5.5 Satz
Der optimale Parameter ω für das Richardson-Verfahren 2. Ordnung, welcher die Minimierungsaufgabe (5.1) löst, ist

$$\omega_0 = \frac{2}{1 + \sqrt{1-\gamma^2}}\,.$$

Jetzt kann man für $w = w_0$ die Größe

$$\max_{t\in[1-\gamma,1+\gamma]} |p_m(t)|$$

diskutieren unter Beachtung von $p_0 = 1$, $p_1(t) = 1 - t$.

5.6 Aufgabe
Bestimme für alle $t \in [1-\gamma, 1+\gamma]$ die Lösung der Differenzengleichung 2. Ordnung

$$p_{m+1}(t) = w_0 \cdot (1-t)p_m(t) + (1-w_0)p_{m-1}(t)$$

unter den Anfangsbedingungen $p_0 = 0$ und $p_1(t) = 1 - t$ und zeige damit

1. $\max\limits_{t\in[1-\gamma,1+\gamma]} |p_m(t)| = \left(\sqrt{w_0 - 1}\right)^m (1 + m\sqrt{1-\gamma^2}) =: \rho_m$, (Dies ist sehr aufwendig![2])
2. ρ_m fällt monoton in m,
3. $\rho_m > c_m$, $c_m = \frac{2c^m}{1+c^{2m}}$ aus dem Tschebyscheff Verfahren mit $c = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$, $\kappa = \frac{1+\gamma}{1-\gamma}$,
4. es gilt $\lim\limits_{n\to\infty} \frac{\rho_m}{c_m} = 1$.

5.7 Bemerkung
Aus $r^m = p_m(A)r^0$ folgt $e^m = p_m(A)e^0$ mit $e^m = A^{-1}b - x^m$ und damit

$$\|e^m\| \leq \|p_m(A)\|\|e^0\|.$$

Tschebyscheff- und Richardson-Verfahren machen eine Schranke für $\|p_m(A)\|$ klein. Dies wirkt sich in demnach in den Abschätzungen nicht nur für das Residuum, sondern auch für den *Fehler* aus.

[2] Man liest also besser bei den Meistern nach: G. Golub, R. Varga: Chebyshev semi-iterative methods, successive overrelaxation iterative methods and second-order Richardson iterative methods, Part I and II, Numer. Math. **3**, 147-168 (1961).

6 Basen für Krylov-Unterräume

Gegeben: $A \in \mathbb{C}^{n\times n}$ nicht notwendigerweise hermitsch, A regulär,

$$K_m = \langle r, Ar, ..., A^{m-1}r\rangle.$$

Gesucht: $v^1, ..., v^m$ mit „schönen Eigenschaften“ und

$$K_m = \langle v^1, ..., v^m\rangle.$$

Erinnerung: Gram-Schmidt-Orthogonalisierung
Gegeben: $y_1, ..., y_m \in \mathbb{C}^{n\times n}$.

6.1 Algorithmus (Gram-Schmidt-Orthogonalisierung, klassisch)

$v^i = y^i,\ i = 1, \ldots, m$
for $i = 1, \ldots, m$ **do**
  for $j = 1, \ldots, i-1$ **do**
    $v^i = v^i - \langle y^i, v^j\rangle v^j$
  end for
  $v^j = \frac{1}{\|v^j\|} \cdot v^j$
end for

Es ist bekannt, dass diese Variante aber numerisch nicht stabil ist. Günstiger ist es, wenn man bei der Berechnung der Innenprodukte statt y_i den aktuellen Vektor v^i verwendet. Dies ist möglich, da zu jedem Zeitpunkt der j-Schleife der Vektor $y^i - v^i \in \langle v^1, \ldots, v^{j-1}\rangle$ senkrecht auf v^j steht.

6.2 Algorithmus (modifizierte Gram-Schmidt-Orthogonalisierung)

$v^i = y^i\ i = 1, \ldots, m$
for $i = 1, \ldots, m$ **do**
  for $j = 1, \ldots, i-1$ **do**
    $v^i = v^i - \langle v^i, v^j\rangle v^j$
  end for
  $v^j = \frac{1}{\|v^j\|} \cdot v^j$
end for

Unser Ziel ist es, sukzessive eine ONB von $K_m(A, r)$ aufzubauen.

Idee:
$\langle v^1, ..., v^m\rangle$ sei eine ONB in $K_m(A, r)$ mit $v_m \notin K_{m-1}(A, r)$. Dann gilt insbesondere $Av_m \in K_{m+1}(A, r) \setminus K_m(A, r)$. Orthogonalisiere also Av^m gegen $v^1, ..., v^m$.
Durch Anwendung des modifizierten Gram-Schmidt-Verfahrens resultiert das *Arnoldi-Verfahren*.

6.3 Algorithmus (Arnoldi-Verfahren)

$v^1 = r$, $h_{1,0} = \|r\|$, $v^1 = \frac{1}{h_{1,0}} r$
for $i = 2, 3, \dots$ **do**
  $v^i = Av^{i-1}$
  for $j = 1, \dots, i-1$ **do**
    $h_{j,i-1} = \langle v_i, v_j \rangle$
    $v^i = v^i - h_{j,i-1} v^j$
  end for
  $h_{i,i-1} = \|v^i\|$
  $v^i = \frac{1}{h_{i,i-1}} \cdot v^i$
end for

Wir führen folgende Kurznotation ein

$$AV_m = V_{m+1} H_{m+1,m} \tag{6.1}$$

mit

$$V_m = \left[v^1, ..., v^m\right] \in \mathbb{C}^{n \times m}$$

$$H_{m+1,m} = \begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,m} \\ h_{2,1} & h_{2,2} & \ddots & h_{2,m} \\ & \ddots & \ddots & \vdots \\ & & \ddots & h_{m,m} \\ & & & h_{m+1,m} \end{pmatrix} \in \mathbb{C}^{(m+1) \times m}.$$

Im Falle $A = A^H$ vereinfacht sich der Arnoldi-Prozess, da die meisten Einträge in $H_{m,m+1}$ verschwinden. Es bezeichne dazu H_m die Matrix

$$H_m = \begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,m} \\ h_{2,1} & h_{2,2} & \ddots & h_{2,m} \\ & \ddots & \ddots & \vdots \\ & & h_{m-1,m} & h_{m,m} \end{pmatrix} \in \mathbb{C}^{m \times m}.$$

6.4 Lemma
Falls A hermitesch ist, ist H_m eine reelle, symmetrische Tridiagionalmatrix. Es gilt also $h_{i+1,i} = h_{i,i+1}$ für alle i und $h_{i,j} = 0$ für $i > j + 1$.

Beweis: Aus (6.1) folgt

$$
\begin{aligned}
V_m^H A V_m &= V_m^H V_{m+1} H_{m+1,m} \\
&= \left(\begin{array}{cc} & 0 \\ I_m & \vdots \\ & 0 \end{array} \right) H_{m+1,m} \\
&= H_m .
\end{aligned}
$$

Mit $V_m^H A V_m$ ist auch H_m hermitesch, besitzt also Tridiagonalgestalt. Außerdem sind dann die Diagonalelemente reell (und positiv). Die Elemente $h_{i,i-1}$ der ersten Nebendiagonalen ergeben sich nach Algorithmus 6.3 als $\|v^i\|$, sind also auch reell. □

Damit erhalten wir aus dem Arnoldi-Verfahren das Lanczos-Verfahren zur Bestimmung einer ONB von $K_m(A,r)$ im Falle $A = A^H$.

6.5 Algorithmus (Lanczos-Verfahren)

$v^1 = \frac{r}{\|r\|},\ v^0 = 0,\ h_{1,0} = 0$
for $i = 2, 3, \dots$ **do**
  $v^i = Av^{i-1}$
  $v^i = v^i - h_{i-1,i-2} v^{i-2}$
  $h_{i-1,i-1} = \langle v^i, v^{i-1} \rangle$
  $v^i = v^i - h_{i-1,i-1} v^{i-1}$
  $h_{i,i-1} = \|v_i\|$
  $v^i = \frac{1}{h_{i,i-1}} \cdot v^i$
end for

Eine alternative Bezeichnung für die Koeffizienten im Lanczos-Verfahren ist

$$h_{i-1,i-2} = \beta_{i-1},\ \ h_{i-1,i-1} = \alpha_{i-1}.$$

Damit lässt sich der Algorithmus wie folgt formulieren:

6.6 Algorithmus (Lanczos-Verfahren)

$v^1 = \frac{r}{\|r\|},\ v^0 = 0,\ \beta_1 = 0$
for $i = 2, 3, \dots$ **do**
  $q^i = Av^{i-1} - \beta_{i-1}v^{i-2}$
  $\alpha_i = \langle q^i, v^{i-1} \rangle$
  $q^i = q^i - \alpha_{i-1}v^{i-1}$
  $\beta_i = \|q^i\|$
  $v^i = \frac{q^i}{\beta_i}$
end for

Hierin gilt:

$$\beta_i v^i = Av^{i-1} - \alpha_{i-1}v^{i-1} - \beta_{i-1}v^{i-2}. \tag{6.2}$$

6.7 Definition
Zwei Folgen $v^1, \cdots, v^m$ und $w^1, \cdots, w^m \in \mathbb{C}^n$ heißen *bi-orthogonal*, falls gilt

$$\langle v^i, w^i \rangle = \delta_{i,j}.$$

6.8 Bemerkung
v^i, w^i sind nur bis auf reziproke Vielfache eindeutig bestimmt.

Im Falle $A \neq A^H$ konstruieren wir nun eine bi-orthogonale Basis $v^1, \cdots, v^m$ von $K_m(A, r)$ und $w^1, \cdots, w^m$ von $K_m(A^H, \tilde{r})$.

Man berechnet v^1, w^1 mit dem *unsymmetrischen* Lanczos-Verfahren.

Situation:

$v^1, \dots, v^m$ sei eine Basis von $K_m(A, r)$
$w^1, \dots, w^m$ sei eine Basis von $K_m(A^H, \tilde{r})$
die beiden Basen sind bi-orthogonal

und

$$v^m \notin K_{m-1}(A, r)$$
$$w^m \notin K_{m-1}(A^H, \tilde{r})$$

Ansatz:

$$q^{m+1} = Av^m$$
$$p^{m+1} = A^H w^m$$

Wir setzen an:

$$\tilde{v}^{m+1} = q^{m+1} - \sum_{i=1}^{m} \alpha_i v^i, \tag{6.3}$$

$$\tilde{w}^{m+1} = p^{m+1} - \sum_{i=1}^{m} \beta_i w^i. \tag{6.4}$$

Bedingung für Bi-Orthogonalität:

$$\begin{aligned} \langle \tilde{v}^{m+1}, w^j \rangle =& \langle q^{m+1}, w^j \rangle - \alpha_j \stackrel{!}{=} 0, \\ \langle v^j, \tilde{w}^{j+1} \rangle =& \langle v^j, p^{m+1} \rangle - \overline{\beta}_j \stackrel{!}{=} 0. \end{aligned}$$

Damit erhalten wir

$$\begin{aligned} \alpha_j =& \langle q^{m+1}, w^j \rangle, \\ \beta_j =& \langle p^{m+1}, v^j \rangle. \end{aligned}$$

Für $j \leq m - 2$ gilt allerdings weiter

$$\begin{aligned} \alpha_j = \langle q^{m+1}, w^j \rangle &= \langle A v^m, w^j \rangle \\ &= \langle v^m, A^H w^j \rangle \\ &= 0, \end{aligned}$$

da $A^H w^j \in K_{m-1}(A^H, \tilde{r})$. Analoges gilt für die β_j, d.h. es gilt:

$$\alpha_j = \beta_j = 0 \text{ für } j = 1, \ldots, m - 2.$$

Zur Normierung sei

$$\begin{aligned} \delta_{m+1} &= \left(\langle \tilde{v}^{m+1}, \tilde{w}^{m+1} \rangle \right)^{\frac{1}{2}} \\ v^{m+1} &= \frac{1}{\delta_{m+1}} \tilde{v}^{m+1} \\ w^{m+1} &= \frac{1}{\overline{\delta}_{m+1}} \tilde{w}^{m+1} \end{aligned}$$

Schließlich weisen wir noch nach, dass die Koeffizienten α_{m-1} und β_{m-1} vom vorherigen Schritt schon bekannt sind (dafür verwenden wir kurzfristig einen zusätzlichen oberen Index):

$$\begin{aligned} \alpha_{m-1} &= \langle A v^m, w^{m-1} \rangle \\ &= \langle v^m, A^H w^{m-1} \rangle \end{aligned}$$

$$\begin{aligned}
&= \langle v^m, \tilde{w}^m + \beta_{m-1}^{(m-1)} w^{m-1} + \beta_{m-2}^{(m-1)} w^{m-2} \rangle \\
&= \langle v^m, \tilde{w}^m \rangle = \delta_m
\end{aligned}$$

$$\begin{aligned}
\beta_{m-1} &= \langle A^H w^m, v^{m-1} \rangle \\
&= \langle w^m, A v^{m-1} \rangle \\
&= ... \\
&= \langle w^m, \tilde{v}^m \rangle = \overline{\delta}_m
\end{aligned}$$

Insgesamt erhalten wir damit folgendes Verfahren:

6.9 Algorithmus (unsymmetrisches Lanczos-Verfahren)

$v^0 = w^0 = 0$
$\delta_1 = (\langle r, \tilde{r} \rangle)^{\frac{1}{2}}$
$v^1 = \frac{1}{\delta_1} \cdot r$
$w^1 = \frac{1}{\delta_1} \cdot \tilde{r}$
for $m = 1, 2, \ldots$ **do**
  $\alpha_m = \langle A v^m, w^m \rangle$
  $\tilde{v}^{m+1} = A v^m - \alpha_m v^m - \delta_m v^{m-1}$
  $\tilde{w}^{m+1} = A^H w^m - \overline{\alpha}_m w^m - \overline{\delta_m} w^{m-1}$
  $\delta_{m+1} = (\langle \tilde{v}^{m+1}, \tilde{w}^{m+1} \rangle)^{\frac{1}{2}}$
  $v^{m+1} = \frac{1}{\delta_{m+1}} \cdot \tilde{v}^{m+1}$
  $w^{m+1} = \frac{\tilde{w}^{m+1}}{\overline{\delta}_{m+1}} \cdot \tilde{w}^{m+1}$
end for

6.10 Bemerkung
Im unsymmetrischen Lanczos kann Folgendes passieren

$$\delta_m = 0,$$

obwohl $\tilde{v}^m \neq 0$ und $\tilde{w}^m \neq 0$.

Das nennt man *serious breakdown*, da das Verfahren abbricht, ohne dass eine Basis für den maximalen Krylov Unterraum $K_{m^*}(A, r)$ oder $K_{m^{**}}(A^H, \tilde{r})$ berechnet wurde.

7 Krylov-Unterraum-Verfahren mit minimalem Residuum

Gegeben:

$A \in \mathbb{C}^{n\times n}$ mit $A \neq A^H$, A regulär.

Ziel:

Löse $Ax = b$ mit Startvektor x^0 und $r^0 = b - Ax^0$.

Idee:

Bestimme ONB von $K_m(A, r^0)$ mit Arnoldi-Prozess und errechne $x^m \in x^0 + K_m(A, r^0)$, so dass:

$$\|b - Ax^m\|_2 = \min_{x\in x^0+K_m(A,r^0)} \|b - Ax\|_2.$$

Ansatz:

$x = x^0 + V_m z^m$ mit $z^m = (z_1, \cdots, z_m)^T \in \mathbb{C}^m$ sowie $V_m = [v^1, \cdots, v^m]$.

Damit erhalten wir

$$b - Ax = r^0 - AV_m z^m = r^0 - V_{m+1}H_{m+1,m}z^m.$$

Mit $r^0 = h_{1,0}v^1$ und $h_{1,0} = \|r^0\|$ gilt nun

$$b - Ax = V_{m+1}\left[\begin{pmatrix} h_{1,0} \\ 0 \\ \vdots \\ 0 \end{pmatrix} - H_{m+1,m}z^m\right]$$

und somit

$$\|b - Ax\|_2 = \|h_{1,0}e^1 - H_{m+1,m}z^m\|_2.$$

Hierbei ist e^1 der erste Einheitsvektor in $\mathbb{C}^{m+1}$, $H_{m+1,m} \in \mathbb{C}^{(m+1)\times m}$ aber $z \in \mathbb{C}^m$.

Exkurs:

Zu lösen ist also ein überbestimmtes Least Squares Problem $Bx = c$ mit $B \in \mathbb{C}^{k\times\ell},\ k > \ell$. Gesucht ist ein x, welches $\|Bx-c\|_2^2$ minimiert. Dies findet man unter Verwendung der QR-Faktorisierung.

Ansatz:

Faktorisiere $B = QR$ mittels QR-Zerlegung. Dann folgt

$$\begin{aligned}\|Bx-c\|_2^2 &= \|Q^H(Bx-c)\|_2^2\\ &= \|Rx - Q^H c\|_2^2.\end{aligned}$$

Das Minimum wird erreicht für

$$x = R_1^{-1}((Q^H c)(1:\ell)),$$

wobei

$$R = \begin{pmatrix} & R_1 & \\ 0 & \cdots & 0 \end{pmatrix},\quad R_1 \in \mathbb{C}^{\ell\times\ell} \text{ obere Dreiecksmatrix }.$$

In unserem Falle ist $H_{m+1,m}$ eine obere Hessenberg Matrix. Ihre QR-Faktorisierung kann mittels Jacobi Rotationen bestimmt werden. Es kann sogar Q in der QR-Faktorisierung von $H_{m+1,m}$ auf die für $H_{m+2,m+1}$ aufdatiert werden.

Notation:

$$Q_{m+1}H_{m+1,m} = R_{m+1,m} \ \text{ QR-Faktorisierung}$$

mit

$$Q_{m+1} = J_m^{(m+1)} \cdots J_1^{(m+1)},$$

wobei $J_j^{(m+1)}$ eine Jacobi Rotation in $\mathbb{C}^{m+1}$ auf den Zeilen j und $j+1$ ist, welche das 'untere Nebendiagonalelement' $h_{j+1,j}$ eliminiert,

$$J_j^{(m+1)} = \begin{pmatrix} I & & & \\ & \bar{c}_j & \bar{s}_j & \\ & -s_j & c_j & \\ & & & I \end{pmatrix}.$$

Um anzugeben, wie man $J_m^{(m+1)}$ bestimmt, beobachten wir zunächst

$$J_{m-1}^{(m+1)} \cdots J_1^{(m+1)} H_{m+1,m} = J_{m-1}^{(m+1)} \cdots J_1^{(m+1)} \left(\begin{array}{c|c} & h_{1,m} \\ H_{m,m-1} & \vdots \\ & h_{m,m} \\ \hline 0 & h_{m+1,m} \end{array}\right)$$

$$= \begin{pmatrix} R_{m,m-1} & \eta^{(m)} \\ 0 & h_{m+1,m} \end{pmatrix}$$

mit

$$\eta^{(m)} = (r_{1,m}, \dots, r_{m-1,m}, \eta_{m,m})^T,$$

wobei

$$\begin{aligned} r_{1,m} &= \bar{c}_1 h_{1,m} + \bar{s}_1 h_{2,m}, \\ \eta_{2,m} &= -s_1 h_{1,m} + c_1 h_{2,m}, \\ r_{2,m} &= \bar{c}_2 \eta_{2,m} + \bar{s}_2 h_{3,m}, \\ \eta_{2,m} &= -s_2 \eta_{2,m} + c_2 h_{3,m}, \\ &\vdots \\ \eta_{m,m} &= -s_{m-1} \eta_{m-1,m} + c_{m-1} h_{m,m}. \end{aligned}$$

Damit erhalten wir für die Parameter c_m, s_m der Jacobi Rotation $J_m^{(m+1)}$, welche $(0, \dots, 0, \eta_{m,m}, h_{m+1,m})^T$ abbildet auf $(0, \dots, 0, r_{m,m}, 0)^T$ die Werte

$$c_m = \eta_{m,m} / \sqrt{|\eta_{m,m}|^2 + |h_{m+1,m}|^2}, \; s_m = h_{m+1,m} / \sqrt{|\eta_{m,m}|^2 + |h_{m+1,m}|^2} \tag{7.1}$$

und

$$r_{m,m} = \bar{c}_m \eta_{m,m} + \bar{s}_m h_{m+1,m} = \sqrt{|\eta_{m,m}|^2 + |h_{m+1,m}|^2}.$$

Die Lösung des LS-Problems

$$\min \|h_{1,0} e^1 - H_{m+1,m} z^m\|_2$$

lautet damit

$$z^m = h_{1,0} R_m^{-1}((Q_{m+1} e^1)(1:m)), \quad \text{wobei } R_{m+1,m} = \begin{pmatrix} R_m \\ 0 \cdots 0 \end{pmatrix},$$

und der Wert des Minimums ist

$$h_{1,0}(Q_{m+1} e^1)_{m+1}.$$

Dieser Wert kann ohne Kenntnis von z^m wie folgt bestimmt werden: Setze

$$Q_m e^1 = h_{1,0} \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_{m-1} \\ \tilde{\epsilon}_m \end{pmatrix}.$$

Dann ist

$$Q_{m+1}e^1 = J_m^{(m+1)} \begin{pmatrix} Q_m & 0 \\ 0 & 1 \end{pmatrix} e^1 = J_m^{(m+1)} \cdot \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_{m-1} \\ \tilde{\epsilon}_m \\ 0 \end{pmatrix} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_m \\ \tilde{\epsilon}_{m+1} \end{pmatrix},$$

mit

$$\begin{aligned} \epsilon_m &= \tilde{\epsilon}_m \overline{c}_m, \\ \tilde{\epsilon}_{m+1} &= \tilde{\epsilon}_m \overline{s}_m. \end{aligned}$$

Insbesondere gilt

$$\tilde{\epsilon}_{m+1} = \prod_{j=1}^{m} \overline{s}_j,$$

so dass wir für den Wert des Minimums den Ausdruck

$$h_{1,0}\tilde{\epsilon}_{m+1} = \|r^0\| \cdot \tilde{\epsilon}_{m+1}.$$

erhalten.
Wir haben jetzt alle Zutaten für das Verfahren GMRES (Generalized Minimal Residual):

7.1 Algorithmus (GMRES (Saad, Schultz, 1986))

wähle x^0, setze $r^0 = b - Ax^0,\ \rho_0 = h_{1,0} = \|r^0\|_2$
for $m = 1, 2, \dots,$ bis ρ_m klein genug **do**
  bestimme nächsten Vektor v^m und Koeffizienten $h_1^m, \dots, h_{m+1}^m$ aus Arnoldi-Prozess
  bestimme s_m, c_m nach (7.1)
  datiere ϵ_m und $\tilde{\epsilon}_{m+1}$ auf
  setze $\rho_m = \rho_{m-1} \cdot \overline{s}_m \quad \{\ \rho_m = \|r^m\|\}$
  if ρ_m klein genug **then**
    löse $R_m z^m = (\epsilon_1, \dots, \epsilon_m)^T$
    setze $x_m = x_0 + V_m z^m$
  end if
end for

7.2 Bemerkung
Speicher- und Rechenaufwand (ohne MVMs) verhält sich wie $\mathcal{O}(mn + m^2)$
$\curvearrowright$ Verfahren wird schnell impraktikabel.

Es bezeichne x^m die GMRES-Iterierte nach jedem Schritt, also die Lösung von

$$\min_{x \in x^0 + K_m(A, r^0)} \|b - Ax\|_2.$$

Dann gilt natürlich

$$\|r^0\| \geq \|r^1\| \geq \dots .$$

Das folgende Beispiel zeigt, dass hier lange Gleichheit herrschen kann.

7.3 Beispiel
Sei

$$A = \begin{pmatrix} 0 & & & 1 \\ 1 & \ddots & & 0 \\ & \ddots & \ddots & \vdots \\ & & 1 & 0 \end{pmatrix},\ \ x^0 = 0,\ b = e^1 = r^0.$$

Dann ist $K_m(A, r^0) = \operatorname{span}\{e^1, \dots, e^m\}$, $e^i = i$-ter Einheitsvektor. Lösung von

$$\min_{x \in K_m(A, r^0)} \|e^1 - Ax\|_2,$$

ist für $m = 1, 2, \dots, n-1$ wegen $Ax \in \operatorname{span}\{e^2, \dots, e^{m+1}\}$ stets $x^m = 0$. Es gilt also für die GMRES-Iterierten

$$x^0 = x^1 = \dots = x^{n-1} = 0,\ x^n = e^n = A^{-1}b$$

und

$$r^0 = r^1 = \ldots = r^{n-1} = e^1,$$

also

$$\|r^0\| = \ldots = \|r^{n-1}\| = 1.$$

Beachte: $\text{spek}(A) = \{e^{\frac{2\pi ik}{n}},\ k = 0, \ldots, n-1\}$. Ein so um die 0 verteiltes Spektrum ist für KUV prinzipiell schwierig, weil die Verfahrenspolynome ja $p(0) = 1$ erfüllen müssen.

Idee für „praktikable“ GMRES-Variante: „Restarted“GMRES=GMRES(k) mit restart Wert k.

7.4 Algorithmus (GMRES(k))

```
wähle x^0
for m = 1, 2, ... do
   führe k Schritte von GMRES aus mit Startvektor x^{m-1}
   Ergebnis ist x^m
end for
```

Es ist nun

$$x^m \in x^{m-1} + K_k(A, r^{m-1}), r^{m-1} = b - Ax^{m-1},$$

aber es gilt i. Allg. *nicht*

$$x^m \in x^0 + K_{km}(A, r^0).$$

7.5 Bemerkung
Berechnung von x^m aus x^{m-1} kostet k MVM. Speicheraufwand ist $n \cdot k + \mathcal{O}(n + k^2)$, entsprechend für Rechenaufwand.

Auch bei restarted GMRES gilt $\|r^0\| \geq \|r^1\| \geq \ldots$. Jetzt besteht allerdings die Gefahr der *Stagnation*, d.h. $\lim\limits_{m\to\infty} \|r^m\| \neq 0$.

7.6 Beispiel
Seien A, x^0, b wie in Beispiel 7.3. Für alle $k < n$ stagniert GMRES(k) mit $x^m = 0$ für alle m.

Exkurs: Ein Modellproblem (Modellproblem III) mit unsymmetrischer Matrix.

$$\text{(7.2)} \qquad \left\{ \begin{array}{rcl} \Delta u + \epsilon u_x &=& f \text{ für } (x,y) \in \Omega = (0,1) \times (0,1) \\ u &=& g \text{ für } (x,y) \in \partial\Omega \end{array} \right.$$

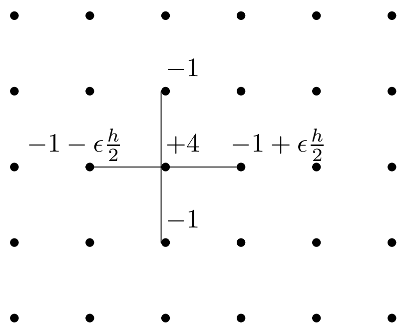


Abbildung 4: 5-Punkte-Stern für die Konvektions-Diffusionsgleichung (7.2)

Man nennt dies eine Konvektions-Diffusions-Gleichung mit Konvektionsterm ϵu_x. Ansatz mit Diskretisierung auf Gitterpunkten $(ih, jh),\ i, j = 1, \ldots, N$, $h = \frac{1}{N+1}$, ergibt ein 5-Punkte-Stern:

$$\begin{aligned}
\Delta u(ih, jh) &\approx \frac{4u_{i,j} - u_{i-1,j} - u_{i,j-1} - u_{i+1,j} - u_{i,j+1}}{h^2}, \\
u_x(ih, jh) &\approx \frac{u_{i+1,j} - u_{i-1,j}}{2h}.
\end{aligned}$$

Wie früher nehmen wir $f = 0$ und

$$\begin{aligned}
g_{0,j} &= jh \\
g_{N+1,j} &= 1 - jh \\
g_{i,0} &= ih \\
g_{i,N+1} &= 1 - ih.
\end{aligned}$$

Die zugehörige Matrix ist unsymmetrisch.

7.7 Aufgabe
Für $\epsilon = 1, 50$ und $N = 10, 20, 40$ berechne man mit MATLAB Folgendes

1. das vollständige Spektrum der Matrix (Plot!)

2. die Konvergenzgeschichte von GMRES(k) für verschiedene Werte von $k = 1, 2, 4, 8, 16, 32, 64$.

Spezialfall für GMRES: $A = A^H$, aber nicht notwendig hpd. Dann: Arnoldi-

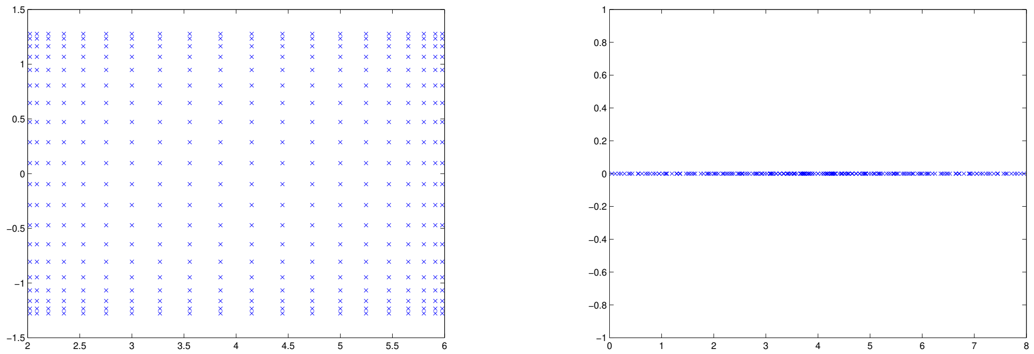


Abbildung 5: Spektrum von A_ϵ für $N = 20$ und $\epsilon = 50$ (links) und $\epsilon = 1$ (rechts)

wird zum Lanczos-Prozess, d.h. kurze Rekursionen werden möglich.

$$H_{m+1,m} \rightarrow T_{m+1,m} = \begin{pmatrix} \alpha_1 & \beta_2 & & \\ \beta_2 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_m \\ & & \beta_m & \alpha_{m+1} \\ & & & \beta_{m+1} \end{pmatrix}.$$

Berechnung der QR-Faktorisierung von $T_{m+1,m}$ vereinfacht sich:

$$\begin{aligned} J_{m-1}^{(m+1)} \cdots J_1^{(m+1)} \cdot T_{m+1,m} &= J_{m-1}^{(m+1)} \cdots J_1^{(m+1)} \cdot \begin{pmatrix} & 0 \\ & \vdots \\ T_{m,m-1} & 0 \\ & \beta_m \\ & \alpha_m \\ 0 \ldots 0 & \beta_{m+1} \end{pmatrix} \\ &= \begin{pmatrix} & 0 \\ & \vdots \\ & 0 \\ R_{m,m-1} & \tilde{\gamma}_m \\ & \tilde{\beta}_m \\ & \tilde{\alpha}_m \\ 0 \cdots 0 & \tilde{\beta}_{m+1} \end{pmatrix} \end{aligned}$$

mit

$$\tilde{\gamma}_m = \overline{s}_{m-2} \beta_m,$$

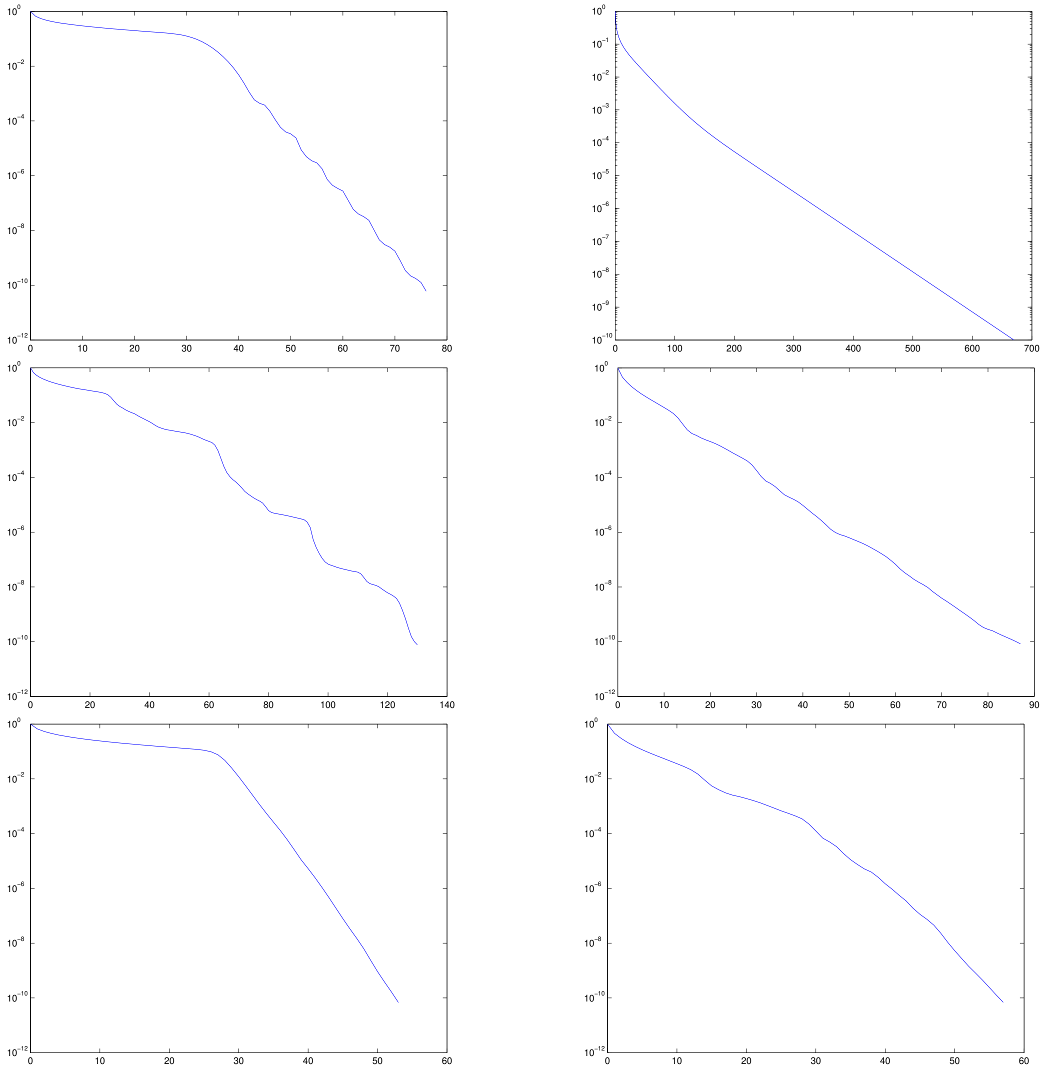


Abbildung 6: Konvergenzgeschichte (Norm des Residuums als Funktion der Iterationszahl) für GMRES(k), $k = 1, 16, 64$). Linke Spalte: $N = 20, \epsilon = 50$, rechte Spalte: $N = 20, \epsilon = 1$

$$
\begin{aligned}
\tilde{\tilde{\beta}}_m &= c_{m-2}\beta_m, \\
\tilde{\beta}_m &= \overline{c}_{m-1}\tilde{\tilde{\beta}}_m + \overline{s}_{m-1}\alpha_m \\
\tilde{\tilde{\alpha}}_m &= -s_{m-1}\tilde{\tilde{\beta}}_m + c_{m-1}\alpha_m
\end{aligned}
$$

Setze nun

$$
\begin{aligned}
\tilde{\alpha}_m &= \sqrt{|\tilde{\tilde{\alpha}}_m|^2 + |\beta_{m+1}|^2}, \\
c_m &= \tilde{\tilde{\alpha}}_m/\tilde{\alpha}_m, \\
s_m &= \beta_{m+1}/\tilde{\alpha}_m.
\end{aligned}
$$

Mit

$$
J_m^{(m+1)} = \begin{pmatrix} \mathrm{I} & 0 & \\ & & \\ 0 & \overline{c}_m & \overline{s}_m \\ & -s_m & c_m \end{pmatrix}
$$

gilt

$$
J_m^{(m+1)} \cdots J_1^{(m+1)} \cdot T_{m+1,m} = \begin{pmatrix} & 0 \\ & \vdots \\ & 0 \\ R_{m,m-1} & \tilde{\gamma}_m \\ & \tilde{\beta}_m \\ & \tilde{\alpha}_m \\ 0 \cdots 0 & 0 \end{pmatrix} = R_{m+1,m}.
$$

$R_{m,m+1}$ hat nur drei besetzte Diagonalen. Die GMRES-Iterierte x^m erfüllt

$$
x^m = x^0 + V_m z^m \text{ mit } R_m z^m = \beta_1(\epsilon_1, \ldots, \epsilon_m)^T.
$$

Frage: Wie datiert man x^m aus x^{m-1} auf, wenn möglich mit kurzer Rekursion? Es ist

$$
x^m = x^0 + V_m z^m \text{ mit } R_m z^m = \beta_1(\epsilon_1, \ldots, \epsilon_m)^T,
$$

also

$$
x^m = x^0 + \underbrace{V_m R_m^{-1}}_{=:W_m} \beta_1(\epsilon_1, \ldots, \epsilon_m)^T,
$$

mit $W_m = [w^1|\dots|w^m]$. Wir setzen also

$$\begin{aligned} W_m &= V_m R_m^{-1} \\ \Longleftrightarrow W_m R_m &= V_m. \end{aligned}$$

Jetzt suchen wir eine Vorschrift zur Aufdatierung von W_m:

$$W_{m+1}R_{m+1} = V_{m+1}$$

$$\Longleftrightarrow W_{m+1} \begin{pmatrix} & & & 0 \\ & & & \vdots \\ & R_m & & 0 \\ & & & \tilde{\gamma}_{m+1} \\ & & & \tilde{\beta}_{m+1} \\ 0 & \dots & 0 & \tilde{\alpha}_{m+1} \end{pmatrix} = [V_m|v^{m+1}]$$

$$\Longrightarrow \qquad W_{m+1} = [W_m|w^{m+1}]$$

mit $\tilde{\alpha}_{m+1}w^{m+1} + \tilde{\beta}_{m+1}w^m + \tilde{\gamma}_{m+1}w^{m-1} = v^{m+1}$, also

$$w^{m+1} = \frac{1}{\tilde{\alpha}_{m+1}} \left(v^{m+1} - \tilde{\beta}_{m+1}w^m - \tilde{\gamma}_{m+1}w^{m-1} \right). \tag{7.3}$$

Damit erhalten wir

$$\begin{aligned} x^{m+1} &= x^0 + W_{m+1}\beta_1(\epsilon_1, \dots, \epsilon_{m+1})^T \\ &= x^0 + W_m\beta_1(\epsilon_1, \dots, \epsilon_m)^T + w^{m+1}\epsilon_{m+1}\beta_1 \\ &= x^m + w^{m+1}\epsilon_{m+1}\beta_1. \end{aligned}$$

Somit ergibt sich die GMRES-Variante für hermitesche Matrizen.

7.8 Algorithmus (MINRES, Paige&Saunders, 1975)

wähle x^0, setze $r^0 = b - Ax^0$, $\beta_1 = \|r^0\|_2$, $v^1 = \frac{1}{\beta_1} r^0$
for $m = 1, 2, \ldots$ bis $\|r^m\|$ klein genug **do**
  bestimme v^{m+1} aus dem Lanczos-Prozess
  bestimme $\epsilon_{m+1}, \tilde{\gamma}_{m+1}, \tilde{\beta}_{m+1}, \tilde{\alpha}_{m+1}$
  bestimme c_m, s_m gemäß (7.1) {alles kurze Rekursionen}
  bestimme w^{m+1} aus (7.3)
  setze $x^{m+1} = x^m + \beta_1 \epsilon_{m+1} w^{m+1}$
end for

Wie kann man die kurzen Rekursionen im Fall $A \neq A^H$ beibehalten?
Idee: Arbeite genau wie im MINRES-Verfahren, wobei diesmal $T_{m+1,m}$ durch den unsymmetrischen Lanczos-Prozess geliefert wird (benutze dabei eine Variante, welche $\|v^j\|_2 = 1$ liefert). Es gilt dann also

$$x^m = x^0 + V_m z^m$$

mit z^m löst

$$\min \|\beta_1 e^1 - T_{m+1,m} z^m\|_2, \ \beta_1 = \|r^0\|_2.$$

7.9 Aufgabe
Formuliere das unsymmetrische Lanczos-Verfahren 6.9 so um, dass $\|v^m\|_2 = 1$ für alle m gilt.

Klar: z^m und x^m berechnen sich algorithmisch *genau gleich* wie in MINRES. Wir verzichten deshalb auf ein explizites Aufschreiben. Was kann man jetzt über die Iterierte x^m sagen?

$$\begin{aligned} \|b - Ax^0\|_2 &= \|r^0 - AV_m z^m\|_2 \\ &= \|r^0 - V_{m+1} T_{m+1,m} z^m\|_2 \\ &= \|V_{m+1}(\beta_1 e^1 - T_{m+1,m} z^m)\|_2 \\ &\leq \underbrace{\|V_{m+1}\|_2}_{\leq \sqrt{m+1}} \cdot \underbrace{\|\beta_1 e^1 - T_{m+1,m} z^m\|_2}_{\text{wird minimiert}}. \end{aligned}$$

Dieses Verfahren heißt QMR („Quasi Minmal Residual“, Freund& Nachtigal, 1990).
Eigenschaften sind:

- kurze Rekursionen
- jeder Schritt erfordert zwei Matrix-Vektor-Multiplikationen (eine mit A, eine mit A^H; doppelter Aufwand wie bei GMRES)

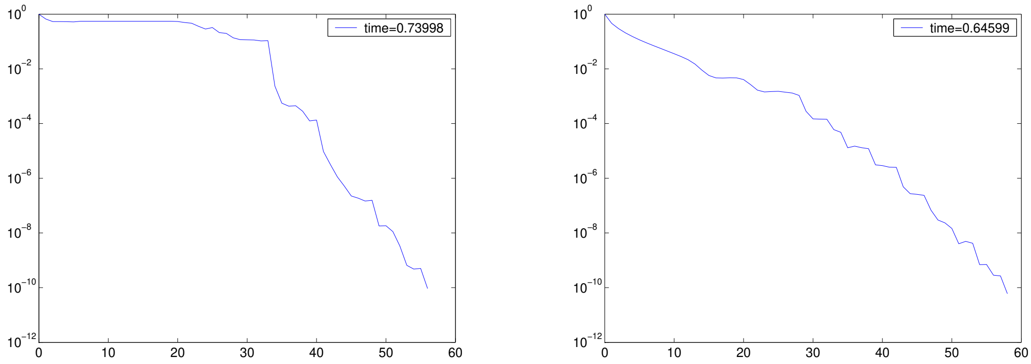


Abbildung 7: Konvergenzgeschichte für QMR. Links:$N = 20, \epsilon = 50$, rechts $N = 20, \epsilon = 1$

- Quasi-Minimierungseigenschaft:

  Statt
  $$\|V_{m+1}(\beta_1 e^1 - T_{m+1,m} z^m)\|_2$$
  wird
  $$\|\beta_1 e^1 - T_{m+1,m} z^m\|_2$$
  minimiert.

7.10 Aufgabe
Führe QMR aus für das lineare Gleichungssystem wie in Aufgabe 7.7 und vergleiche mit GMRES(k) (Konvergenzgeschichte plotten).

8 Konvergenzanalyse für GMRES/MINRES

Gegeben sei $A \in \mathbb{C}^{n\times n},\ Ax = b,\ A$ regulär.

8.1 Definition
Der *numerische Wertebereich* von A („fields of value“) ist definiert durch

$$\begin{aligned} F(A) &= \left\{ \frac{\langle Ax, x\rangle}{\langle x, x\rangle} :\ x \in \mathbb{C}^n,\ x \neq 0 \right\} \\ &= \{\langle Ax, x\rangle :\ \|x\|_2 = 1\}\,. \end{aligned}$$

Mit elementarer Rechnung ergeben sich die folgenden Eigeschaften des numerischen Wertebereichs:

8.2 Lemma

1. Ist A hermitesch mit $\mathrm{spek}(A) = \{\lambda_1 \leq \ldots \leq \lambda_n\}$, dann ist $F(A) = [\lambda_1, \lambda_n]$.
2. $\mathrm{spek}(A) \subseteq F(A)$.
3. Ist A normal, dann ist $F(A) = \mathrm{co}(\mathrm{spek}(A))$ ($\mathrm{co}(M)$ = konvexe Hülle von M).

8.3 Satz

1. Der numerische Wertebereich $F(A)$ ist kompakt, also ist der *numerische Radius*
$$r(A) := \sup\{|\lambda| :\ \lambda \in F(A)\}$$
endlich (und das sup ist ein max).
2. $F(A)$ ist konvex.

Beweis:

1. Klar.
2. Siehe Roger A. Horn, Charles R. Johnson, Topics in Matrix Analysis, Cambridge Univ Press; (1994) (behandelt gerade $F(A)$).

□

8.4 Beispiel
Die Matrix A aus Beispiel 7.3 ist normal. Deshalb ist $F(A)$ die konvexe Hülle der Eigenwerte, also hier ein (ausgefülltes) regelmäßiges n-Eck im Einheitskreis.

8.5 Aufgabe
Zeigen Sie, dass für die Matrix

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

gilt: $F(A) = \{z \in \mathbb{C} : \ |z| \leq 1\}$.

8.6 Lemma
Sei $V_m \in \mathbb{C}^{n \times m}, \ V_m^H V_m = I$. Dann gilt

$$F(V_m^H A V_m) \subseteq F(A).$$

Der Beweis ist wieder trivial.

8.7 Satz
Sei p_m^{GMRES} das Residuen-Polynom aus dem GMRES-Verfahren, d.h.

$$r^m = p_m^{\text{GMRES}}(A) r^0, \quad p_m^{\text{GMRES}} \in \overline{\Pi}_m.$$

Dann gilt

$$p_m^{\text{GMRES}}(\xi) = 0 \Rightarrow \frac{1}{\xi} \in F(W_m^H A^{-1} W_m) \subseteq F(A^{-1}).$$

Dabei ist $W_m \in \mathbb{C}^{n \times m}$ eine Matrix, deren Spalten eine Orthonormalbasis von $A \cdot K_m(A, r^0)$ bilden.

Beweis: Setze $p_m := p_m^{\text{GMRES}}$. Es ist

$$p_m(t) = \prod_{j=1}^{m} \left(1 - \frac{t}{\xi_j}\right).$$

Betrachte $\xi = \xi_m$:

$$p_m(t) = \left(1 - \frac{t}{\xi}\right) \cdot q_{m-1}(t) \text{ mit } q_{m-1}(t) = \prod_{j=1}^{m-1} \left(1 - \frac{t}{\xi_j}\right).$$

Damit ist

$$\|p_m(A) r^0\|_2 = \|(I - \frac{1}{\xi} A) \cdot \underbrace{q_{m-1}(A) \cdot r^0}_{w}\|_2$$

minimal. Also ist ξ so, dass $\|(I - \frac{1}{\xi} A) w\|_2$ minimal ist. Nach Lemma 8.8 (s. unten) gilt deshalb

$$\frac{1}{\xi} = \frac{\langle w, Aw \rangle}{\langle Aw, Aw \rangle} = \frac{\langle A^{-1} y, y \rangle}{\langle y, y \rangle}$$

mit $y = Aw \in A \cdot K_m(A, r^0)$. □

8.8 Lemma
Die Lösung von $\min_\gamma \|w - \gamma y\|_2$ ist

$$\gamma^* = \frac{\langle w, y\rangle}{\langle y, y\rangle}$$

Beweis: Es ist

$$\begin{aligned}\langle w-\gamma y, w-\gamma y\rangle &= \langle w,w\rangle - \gamma\langle y,w\rangle - \bar\gamma\langle w,y\rangle + \gamma\bar\gamma\langle y,y\rangle \\ &= \langle w,w\rangle + \langle y,y\rangle \left(\gamma\bar\gamma - \frac{\langle y,w\rangle}{\langle y,y\rangle}\gamma - \frac{\langle w,y\rangle}{\langle y,y\rangle}\bar\gamma \right. \\ &\qquad \left. + \frac{\langle y,w\rangle\langle w,y\rangle}{\langle y,y\rangle^2}\right) - \frac{\langle y,w\rangle\langle w,y\rangle}{\langle y,y\rangle} \\ &= \langle w,w\rangle + \langle y,y\rangle \left(\gamma - \frac{\langle w,y\rangle}{\langle y,y\rangle}\right)\left(\bar\gamma - \frac{\langle y,w\rangle}{\langle y,y\rangle}\right) \\ &\qquad - \frac{\langle y,w\rangle\langle w,y\rangle}{\langle y,y\rangle}.\end{aligned}$$

Das Minimum wird also für

$$\gamma^* = \frac{\langle w, y\rangle}{\langle y, y\rangle}$$

erreicht.
Schränken wir $A \in \mathbb{C}^{n\times n}$ auf einen Unterraum $V \subseteq \mathbb{C}^n$ ein, so ist nicht notwendig $AV \subseteq V$. Möchte man eine Homomorphismus auf V, muss man AV geeignet projizieren.

8.9 Definition
Die *orthogonale Projektion* von $A \in \mathbb{C}^{n\times n}$ auf den Unterraum $V \subseteq \mathbb{C}^n$ entsteht durch orthogonale Projektion von AV auf V. Bilden die Spalten von $V_m \in \mathbb{C}^{m\times n}$ eine Orthonormalbasis von V ($V_m^H V_m = I$), so ist $V_m^H A V_m$ die Darstellung der orthogonalen Projektion von A auf V in der Basis „Spalten von V_m“.

8.10 Satz
Es sei $0 \notin F(A)$ und $c := \min\{|\lambda| : \lambda \in F(A)\} > 0$. Dann ist $c \le \|A\|_2$ und es gilt

$$\|p_m^{\mathrm{GMRES}}\|_2 \le \left(1 - \tfrac{c^2}{\|A\|_2^2}\right)^{\frac{m}{2}} \cdot \|r^0\|_2. \tag{8.1}$$

Beweis: Es ist

$$c \leq \frac{|\langle Aw, w\rangle|}{\langle w, w\rangle} \leq \frac{\|Aw\|_2 \cdot \|w\|_2}{\langle w, w\rangle} \leq \frac{\|A\|_2 \cdot \|w\|_2^2}{\langle w, w\rangle} = \|A\|_2.$$

Zum Beweis von (8.1) starten wir mit einer Vorüberlegung. Nach Lemma 8.8 gilt für $\frac{1}{\xi} = \frac{\langle w, Aw\rangle}{\langle Aw, Aw\rangle}$

$$\langle w - \tfrac{1}{\xi}Aw, w - \tfrac{1}{\xi}Aw\rangle = \langle w, w\rangle - \tfrac{\langle Aw, w\rangle\langle w, Aw\rangle}{\langle Aw, Aw\rangle}.$$

Für $w \neq 0$ gilt außerdem

$$\begin{aligned} \langle Aw, Aw\rangle &= \|Aw\|_2^2 \leq \|A\|_2^2 \cdot \langle w, w\rangle, \\ |\langle Aw, w\rangle|^2 &\geq \left|\tfrac{\langle Aw, w\rangle}{\langle w, w\rangle}\right|^2 \cdot \langle w, w\rangle^2 \\ &\geq c^2 \cdot \langle w, w\rangle^2. \end{aligned}$$

Also ist für die angegebene Wahl von ξ

$$\langle w - \tfrac{1}{\xi}Aw, w - \tfrac{1}{\xi}Aw\rangle \leq \left(1 - \tfrac{c^2}{\|A\|_2^2}\right) \cdot \langle w, w\rangle, \tag{8.2}$$

mit $\left(1 - \frac{c^2}{\|A\|_2^2}\right) \in [0, 1)$.

Wir schreiben wieder p_m statt p_m^{GMRES} und zeigen (8.1) per Induktion.

$m = 0$: klar.

$m - 1 \to m$: Für $w = p_{m-1}(A)r^0$ ist nach Induktionsvoraussetzung

$$\|w\|_2 \leq \left(1 - \tfrac{c^2}{\|A\|_2^2}\right)^{\frac{m-1}{2}} \cdot \|r^0\|_2.$$

Auf Grund der Minimalitätseigenschaft ist für alle $\xi \in \mathbb{C}\backslash\{0\}$

$$\|p_m(A)r^0\|_2 \leq \|(I - \tfrac{1}{\xi}A)p_{m-1}(A)r^0\|_2 = \|(I - \tfrac{1}{\xi}A)w\|_2.$$

Speziell für die Wahl $\frac{1}{\xi} = \frac{\langle w, Aw\rangle}{\langle Aw, Aw\rangle}$ ergibt sich nach (8.2)

$$\left\|\left(I - \tfrac{1}{\xi}A\right)w\right\|_2^2 \leq \left(1 - \tfrac{c^2}{\|A\|^2}\right) \cdot \|w\|_2^2.$$

$\square$

8.11 Bemerkung
Der Beweis hat gezeigt, dass unter den Voraussetzungen von Satz 8.10 sogar gilt, dass GMRES(k) für alle Restartwerte k mit der selben Abschätzung (8.1) konvergiert.

Wir diskutieren die Nullstellen der GMRES-Polynome jetzt noch genauer.

8.12 Lemma
Für die GMRES-Residuen

$$r^m = p_m^{\mathrm{GMRES}}(A) r^0$$

gilt

$$r^m \perp A \cdot K_m(A, r^0).$$

(Zum Vergleich: Beim cg-Verfahren gilt $r^m \perp K_m(A, r^0)$.)

Beweis: Es gilt

$$r^m = b - Ax^m = r^0 - AV_m z^m, \quad z^m \in \mathbb{C}^m \text{ mit } \|r^0 - AV_m z^m\|_2 \text{ minimal.}$$

Also erfüllt z^m die Normalengleichung (s. Numerik I)

$$\underbrace{V_m^H A^H A V_m}_{M_m} z^m = V_m^H A_m^H r^0$$

mit M_m hpd, da V_m vollen Rang besitzt. Also ist

$$z^m = M_m^{-1} V_m^H A r^0$$

und es gilt für alle $y \in \mathbb{C}^m$

$$\begin{aligned} \langle r^m, AV_m y\rangle &= \langle r^0 - AV_m M_m^{-1} V_m^H A^H r^0, AV_m y\rangle \\ &= \langle r^0, AV_m y\rangle - \langle r^0, AV_m \underbrace{M_m^{-1} V_m^H A^H A V_m}_{=I} y\rangle \\ &= 0. \end{aligned}$$

□

8.13 Satz
Die Nullstellen ξ_i der GMRES-Polynome p_m^{GMRES} sind Eigenwerte des verallgemeinerten Eigenwertproblems

$$(H_{m+1,m}^H H_{m+1,m}) \cdot z = \xi H_m^H \cdot z, \tag{8.3}$$

wobei H_m aus $H_{m+1,m}$ durch Streichen der letzten Spalte entsteht.

Beweis: Man beachte, dass das Produkt $H_{m+1,m}^H H_{m+1,m}$ hpd ist, H_m jedoch singulär sein kann. Sei

$$r^m = p_m(A)r^0 = \left(I - \tfrac{1}{\xi}A\right) \underbrace{q_{m-1}(A)r^0}_{=V_m z}, \; p_m(\xi) = 0.$$

Nach Lemma 8.12 gilt mit $AV_m = V_{m+1}H_{m+1,m}$

$$\begin{aligned}
& r^m \perp AK_m(A, r^0) \\
\iff & \langle r^m, AV_m y\rangle = 0 \quad \forall\, y, \\
\iff & \langle (I - \xi^{-1}A)V_m z, AV_m y\rangle = 0 \quad \forall\, y, \\
\iff & \langle V_m z - \xi^{-1}V_{m+1}H_{m+1,m}z, V_{m+1}H_{m+1,m}y\rangle = 0 \quad \forall\, y, \\
\iff & \langle V_m z, V_{m+1}H_{m+1,m}y\rangle - \xi^{-1}\langle V_{m+1}H_{m+1,m}z, V_{m+1}H_{m+1,m}y\rangle = 0 \quad \forall\, y, \\
\iff & y^H H_{m+1,m}^H V_{m+1}^H V_m z - \xi^{-1} y^H H_{m+1,m}^H \underbrace{V_{m+1}^H V_{m+1}}_{=I} H_{m+1,m} z = 0 \quad \forall\, y.
\end{aligned}$$

Dabei ist $V_{m+1}^H V_m = \begin{pmatrix} & I_m & \\ 0 & \dots & 0 \end{pmatrix}$ und damit $H_{m+1,m}^H V_{m+1}^H V_m = H_m^H$. Wir erhalten also

$$\begin{aligned}
& r^m \perp AK_m(A, r^0) \\
\iff & y^H (H_m^H z - \xi^{-1} H_{m+1,m}^H H_{m+1,m} z) = 0 \\
\iff & H_m^H z - \xi^{-1} H_{m+1,m}^H H_{m+1,m} z = 0.
\end{aligned}$$

□

Bedeutung von (8.3): Betrachte die orthogonale Projektion von A^{-1} auf $A \cdot K_m(A, r^0)$. Eine ONB von $AK_m(A, r^0)$ sind die Spalten von

$$AV_m M_m^{-\frac{1}{2}} \quad \text{mit } M_m = V_m^H A^H A V_m,$$

denn (Basis ist klar)

$$M_m^{-\frac{1}{2}} \underbrace{V_m^H A^H A V_m}_{M_m} M_m^{-\frac{1}{2}} = I.$$

Die orthogonale Projektion von A^{-1} auf $AK_m(A, r^0)$ wird repräsentiert durch

$$\begin{aligned}
M_m^{-\frac{1}{2}} V_m^H (A^H A^{-1} A) V_m M_m^{-\frac{1}{2}} &= M_m^{-\frac{1}{2}} (AV_m)^H V_m M_m^{-\frac{1}{2}} \\
&= M_m^{-\frac{1}{2}} (V_{m+1} H_{m+1,m})^H V_m M_m^{-\frac{1}{2}}
\end{aligned}$$

$$= M_m^{-\frac{1}{2}} H_m^H M_m^{-\frac{1}{2}}.$$

Damit ist ξ^{-1} Eigenwert dieser orthogonalen Projektion, wenn

$$\begin{aligned} & M_m^{-\frac{1}{2}} H_m^H M_m^{-\frac{1}{2}} z = \xi^{-1} z \qquad \text{mit } z \neq 0 \\ \Longleftrightarrow \qquad & H_m^H y = \xi^{-1} M_m y \quad \text{mit } y = M_m^{-\frac{1}{2}} z. \end{aligned}$$

Wir halten fest

8.14 Satz
ξ ist Nullstelle von p_m^{GMRES} genau dann, wenn ξ^{-1} Eigenwert der orthogonalen Projektion von A^{-1} auf $A \cdot K_m(A, r^0)$ ist.

Spezialfall: A normal

$$\begin{aligned} \Rightarrow A &= Q^H \Lambda Q, \quad Q^H Q = I, \quad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \\ \Rightarrow A^{-1} &= Q^H \Lambda^{-1} Q \end{aligned}$$

Damit ist die orthogonale Projektion von A^{-1} auf $AK_m(A, r^0)$ die Matrix

$$W_m^H A^{-1} W_m = W_m^H Q^H \Lambda^{-1} \underbrace{Q W_m}_{\text{orthonormal}} =: B \in \mathbb{C}^{m \times m}$$

mit $W_m^H W_m = I, W_m \in \mathbb{C}^{n \times m}$. B ist selbst normal, $B = P^H D_m P, P \in \mathbb{C}^{m \times m}$ orthonormal, $D_m = \operatorname{diag}(d_1, \ldots, d_m)$.
Also

$$\begin{aligned} & P^H D_m P = W_m^H Q_m^H \Lambda^{-1} Q W_m \\ \Longleftrightarrow \quad & D_m = (P W_m^H Q_m^H) \Lambda^{-1} (\underbrace{Q W_m P_m^H}_{=:Y}) \\ \Rightarrow \quad & d_i = \sum_{j=1}^n \frac{1}{\lambda_i} |y_{ji}|^2 \quad \text{mit} \sum_{j=1}^n |y_{ji}|^2 = 1. \end{aligned}$$

$\frac{1}{\xi}$ ist einer der Eigenwerte d_i von B. Also haben wir

8.15 Satz
Ist A normal (z.B. $A = A^H$), so sind die Nullstellen der GMRES-Polynome gewichtete harmonische Mittel der Eigenwerte von A, d.h.

$$p_m^{\text{GMRES}}(\xi) = 0 \Rightarrow \frac{1}{\xi} = \sum_{j=1}^n \alpha_j \frac{1}{\lambda_j}$$

mit $\alpha_j > 0, \sum \alpha_j = 1, \lambda_j$ Eigenwerte von A.

Man nennt die ξ deshalb auch *harmonische Ritz-Werte.*

8.16 Bemerkung
A sei hermitesch, $spek(A) \cap [-b, a] = \emptyset, a, b > 0$. Dann folgt $\xi \notin [-b, a]$.

Spezielle Resultate für MINRES:

8.17 Satz
Sei $A = A^H \in \mathbb{C}^{n \times n}$ regulär, $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{n^*}$ seien die verschiedenen Eigenwerte von A, $n^* \leq n$.

1. Es gelte $\lambda_1 \leq \ldots \leq \lambda_k < 0$ und $\lambda_{k+1} > 0$. Dann gilt für das Verfahren MINRES:

$$\|r^{m+k}\|_2 \leq \frac{2c^m}{1 + c^{2m}} \prod_{i=1}^{k} \left(1 - \frac{\lambda_{n^*}}{\lambda_i}\right) \|r^0\|_2$$

   mit $c = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ und $\kappa = \frac{\lambda_{n^*}}{\lambda_{k+1}}$

2. Es gelte $\text{spek}(A) \subseteq [-b, -a] \cup [a, b], 0 < a \leq b$. Dann gilt für das Verfahren MINRES:

$$\|r^{2m}\|_2 \leq \frac{2\tilde{c}^m}{1 + \tilde{c}^{2m}} \|r^0\|_2$$

   mit $\tilde{c} = \frac{\tilde{\kappa}-1}{\tilde{\kappa}+1}$ und $\tilde{\kappa} = \frac{b}{a}$

Beweis: Zu 1.: Nehme

$$p_{m+k}(t) = \prod_{i=1}^{k} \left(1 - \frac{t}{\lambda_i}\right) \cdot T_m^{[\lambda_{k+1}, \lambda_{n^*}]}(t)$$

mit $T_m^{[\lambda_{k+1}, \lambda_{n^*}]}$ Tschebyscheff-Polynom für $[\lambda_{k+1}, \lambda_{n^*}]$. Dann gilt mit Korollar 5.2

$$\begin{aligned} \max_{t \in \text{spek}(A)} |p_{m+k}(t)| &= \max_{t \in [\lambda_{k+1}, \lambda_{n^*}]} |p_{m+k}(t)| \\ &\leq \frac{2c^m}{1 + c^{2m}} \prod_{i=1}^{k} \left(1 - \frac{\lambda_{n^*}}{\lambda_i}\right). \end{aligned}$$

Damit gilt

$$\|r^{m+k}\|_2 = \|p_{m+k}^{\text{MINRES}}(A) r^0\|_2$$

$$\begin{aligned}
&\leq \|p_{m+k}(A)r^0\|_2 \\
&\leq \|p_{m+k}(A)\|_2 \cdot \|r^0\|_2 \\
&= \max_{t\in \mathrm{spek}(A)} |p_{m+k}(t)| \cdot \|r^0\|_2 \\
&\leq \prod_{i=1}^{k}\left(1-\frac{\lambda_{n^*}}{\lambda_i}\right)\cdot \frac{2c^m}{1+c^{2m}} \cdot \|r^0\|_2 .
\end{aligned}$$

Zu 2.: Nehme $p_{2m}(t) = T_m^{[a^2,b^2]}(t^2)$. Dann gilt

$$\begin{aligned}
\|r^{2m}\|_2 &\leq \|T_m^{[a^2,b^2]}(A^2)\|_2 \cdot \|r^0\|_2 \\
&\leq \max_{s\in \mathrm{spek}(A^2)} |T_m^{[a^2,b^2]}(s)| \cdot \|r^0\|_2 \\
&\leq \max_{s\in [a^2,b^2]} |T_m^{[a^2,b^2]}(s)| \cdot \|r^0\|_2 \\
&= \frac{2\tilde{c}^m}{1+\tilde{c}^{2m}} \cdot \|r^0\|_2
\end{aligned}$$

$\square$

8.18 Bemerkung

Im 1. Fall ist für kleines k die Konvergenz von MINRES also vergleichbar mit CG (auf $[\lambda_{k+1}, \lambda_{n^*}]$). CG für $A^H Ax = A^H b$ mit $A^H A$ hpd benötigt: 2 MVM pro Schritt (A und A^H) und

$$\kappa(A^H A) = \frac{\max\{\lambda_1^2, \lambda_{n^*}^2\}}{\min\{\lambda_k^2, \lambda_{k+1}^2\}} \geq \kappa^2 .$$

Wenn k klein ist, ist CG für $A^H A$ also 4 mal so aufwendig.

Vergleich mit CG im 2. Fall, wobei a und b jetzt beide Eigenwerte von A seien.
Wegen $\kappa(A^H A) = \frac{b^2}{a^2}$ folgt dann: CG für $A^H A$ ist gleich schnell wie MINRES.

Der Fall $\mathrm{spek}(A) \subseteq [-b, -a] \cup [c, d]$ kann allgemein mit Orthogonalpolynomen bzgl. zueinander disjunkter Intervalle behandelt werden.
Siehe dazu: Bernd Fischer - Polynomial Based Iteration Methods for Symmetric Linear Systems, Wiley (1995).

9 CG und BiCG

Gegeben:

$Ax = b$, $A \in \mathbb{C}^{n \times n}$, A regulär.

Gesucht:

$r^m \in K_m(A, r^0)$, s.d.

$$r^m \perp K_m(A^H, \tilde{r}^0) \tag{9.1}$$

(9.1) heißt *Petrov-Galerkin*-Bedingung.
Erinnerung an unsymmetrischen Lanczos:

$$\left\{ \begin{array}{rcl} AV_m & = & V_{m+1}T_{m+1,m} \\ A^H W_m & = & W_{m+1}\overline{T}_{m+1,m} \end{array} \right. \tag{9.2}$$

mit

$$T_{m+1,m} = \begin{pmatrix} \alpha_1 & \delta_2 & & \\ \delta_2 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \delta_m \\ & & \ddots & \alpha_m \\ & & & \delta_{m+1} \end{pmatrix}$$

mit $\alpha_i = \langle Av^i, w^i \rangle$ und $\delta_i^2 = \langle \tilde{v}_i, \tilde{w}_i \rangle$ aus Algorithmus 6.9. Setze

$$T_m = \begin{pmatrix} \alpha_1 & \delta_2 & & \\ \delta_2 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \delta_m \\ & & \delta_m & \alpha_m \end{pmatrix} \in \mathbb{C}^{m \times m}.$$

9.1 Lemma
Sei $x^m = x^0 + V_m z^m$, $r^m = b - Ax^m$. Dann erfüllt r^m die Bedingung (9.1) genau dann, wenn

$$T_m z^m = \delta_1 e^1 \tag{9.3}$$

gilt.

Beweis: $r^m = r^0 - AV_m z^m$

$$
\begin{aligned}
r^m \perp K_m(A, \tilde{r}^0) &\iff \langle r^m, W_m y\rangle = 0 \qquad \forall y \in \mathbb{C}^n \\
&\iff \langle r^0 - AV_m z^m, W_m y\rangle = 0 \\
&\iff \langle r^0 - V_{m+1} T_{m+1,m} z^m, W_m y\rangle = 0 \\
&\iff \langle V_{m+1}(\delta_1 e^1 - T_{m+1,m} z^m), W_m y\rangle = 0 \\
&\iff \langle W_m^H V_{m+1}(\delta_1 e^1 - T_{m+1,m} z^m), y\rangle = 0 \\
&\iff W_m^H V_{m+1}(\delta_1 e^1 - T_{m+1,m} z^m) = 0 \\
&\iff \left(\begin{array}{c|c} & 0 \\ I_m & \vdots \\ & 0 \end{array}\right) \cdot (\delta_1 e^1 - T_{m+1,m} z^m) = 0 \\
&\iff \delta_1 e^1 - T_m z^m = 0 \\
&\iff T_m z^m = \delta_1 e^1.
\end{aligned}
$$

□

9.2 Bemerkung
Im allgemeinen Fall besitzt (9.3) nicht unbedingt eine Lösung, da T_m singulär werden kann. Ist aber $A = A^H$ hpd und $\tilde{r}^0 = r^0$ im unsymmetrischen Lanczos, so wird der Prozess zum symmetrischen Lanczos und $T_m = V_m^H A V_m$ hpd.

9.3 Definition
Das KUV, welches Iterierte x^m bestimmt mit (9.1), heißt BiCG-Verfahren (bi-konjugierte Gradienten). Im Spezialfall $A = A^H$ hpd, $\tilde{r}^0 = r^0$ heißt es CG-Verfahren.

Ab jetzt diskutieren wir BiCG. Für CG siehe Vorlesung „Parallele Algorithmen WS 02/03“.

9.4 Lemma
Sei $\xi \neq 0$ eine Nullstelle des BiCG-Polynoms, $r^m = p_m^{\mathrm{BiCG}}(A) r^0$. Dann gilt: ξ ist Eigenwert von T_m.

Beweis: Sei $p_m^{\mathrm{BiCG}}(t) = (1 - \frac{t}{\xi}) q_{m-1}(t)$. Wegen (9.1) gilt

$$
\begin{aligned}
& \langle (I - \tfrac{1}{\xi} A) \underbrace{q_{m-1}(A) r^0}_{= V_m z^m}, W_m y\rangle = 0 \qquad \forall y \in \mathbb{C}^m \\
= \quad & \langle (I - \tfrac{1}{\xi} A) V_m z^m, W_m y\rangle = 0 \\
\iff \quad & \langle V_m z^m, W_m y\rangle - \tfrac{1}{\xi} \langle A V_m z^m, W_m y\rangle = 0 \\
\iff \quad & \langle z^m, y\rangle - \tfrac{1}{\xi} \langle V_{m+1} T_{m+1,m} z^m, W_m y\rangle = 0 \\
\iff \quad & \langle z^m, y\rangle - \tfrac{1}{\xi} \langle T_m z^m, y\rangle = 0 \\
\iff \quad & z^m - \tfrac{1}{\xi} T_m z^m = 0 \\
\iff \quad & T_m z^m = \xi z^m.
\end{aligned}
$$

□

Betrachte den Spezialfall $A = A^H$, $r^0 = \tilde{r}^0$. Dann ist

$$\begin{aligned} T_m &= V_m^H A V_m \quad \text{hermitesch}, \\ T_m &= P^H D_m P, \quad P^h P = I, \quad D_m = \operatorname{diag}(d_1, \dots, d_m). \end{aligned}$$

Andererseits

$$A = Q^H \Lambda Q, \quad Q^H Q = I, \quad \Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n).$$

Also

$$\begin{aligned} D_m = P^H T_m P &= P^H V_m^H A V_m P \\ &= P^H V_m^H Q^H \Lambda \underbrace{Q V_m P}_{=:Y}. \end{aligned}$$

Insbesondere gilt

$$d_i = \sum_{j=1}^n \lambda_j |y_{ji}| \quad \text{mit} \quad \sum |y_{ji}|^2 = 1.$$

Die Eigenwerte von T_m sind also gewichtete arithmetische Mittel der Eigenwerte von A, die sog. *Ritz-Werte*.
Wie bei CG (siehe Vorlesung Parallele Algorithmen WS 2002/03) kann man aus dem unsymmetrischen Lanzcos-Verfahren eine Standardform des BiCG-Verfahrens herleiten. Wir geben hier lediglich das Resultat an:

9.5 Algorithmus (BiCG, Fletcher 1973)

wähle x^0, setze $r^0 = b - Ax^0$, wähle $\tilde{r}^0$, $p^0 = r^0$, $\tilde{p}^0 = \tilde{r}^0$
for $k = 0, 1, \dots$ **do**
$\gamma_k = \dfrac{\langle r^k, \tilde{r}^k \rangle}{\langle Ap^k, \tilde{p}^k \rangle}$
$x^{k+1} = x^k + \gamma_k p^k$
$r^{k+1} = r^k - \gamma_k A p^k$
$\rho_k = \dfrac{\langle r^{k+1}, \tilde{r}^{k+1} \rangle}{\langle r^k, \tilde{r}^k \rangle}$
$p^{k+1} = r^{k+1} + \rho_k p^k$
$\tilde{r}^{k+1} = \tilde{r}^k - \overline{\gamma}_k A^H \tilde{p}^k$
$\tilde{p}^{k+1} = \tilde{r}^{k+1} + \overline{\rho}_k \tilde{p}^k$
end for

9.6 Bemerkung

1. 2 MVM pro Schritt (Ap^k und $A^H \tilde{p}^k$)

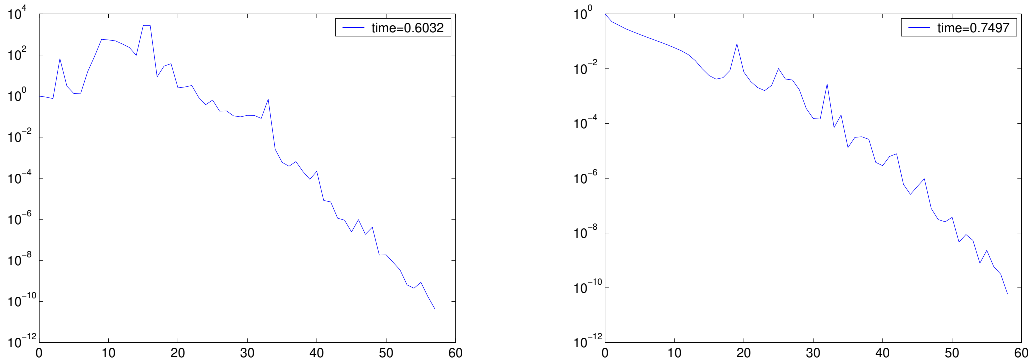


Abbildung 8: Konvergenzgeschichte für BiCG, Modellproblem III. Links:$N = 20, \epsilon = 50$, rechts $N = 20, \epsilon = 1$

2. Breakdowns: Lanczos Breakdown (erkennbar durch $\langle r^k, \tilde{r}^k \rangle = 0$) und BiCG-Breakdown ($T_k$ ist singulär, d.h. es existiert keine Iterierte mit $r^k \perp K_k(A^H, \tilde{r}^0)$, erkennbar durch $\langle Ap^k, \tilde{p}^k \rangle = 0$).

9.7 Aufgabe
Führe BiCG für das Modellproblem III durch.

9.8 Aufgabe
Sei $A \neq A^H$, aber mit einer der folgenden speziellen Eigenschaften:

1. $A^H = \overline{A} \in \mathbb{C}^{n \times n}$ (z.B. bei der Diskretisierung der Maxwell-Gleichungen in der Elektrodynamik)

2. $A^H J = J^H A,\ J \in \mathbb{C}^{n \times n},\ J = J^H$ (man nennt A dann auch J-*hermitesch*)

3. $A^T J = J^T A,\ J \in \mathbb{C}^{n \times n}$ (man nennt A dann auch J-*symmetrisch*)

Zeige jeweils: Für geeignete Wahl von $\tilde{r}^0$ (bzw. w^0) kann man den unsymmetrischen Lanczos-Prozess mit nur einer MVM mit A (und eventuell einer zusätzlichen mit J) durchführen (dies überträgt sich dann auch auf BiCG).

10 BiCGStab

Idee:

1. Glätte Konvergenz in BiCG !
2. Ersetze Multiplikation mit A^H durch eine mit A und erreiche $x^k \in x^0 + K_{2k}(A, r^0)$

Dafür benötigen wir die folgenden BiCG-Verfahrenspolynome:

10.1 Definition
$\varphi_m \in \overline{\Pi}_m,\ \psi_m \in \Pi_m$ sind die Polynome des BiCG-Verfahrens, d.h.

$$\varphi_m(A)r^0 = r^m,\ \psi_m(A)r^0 = p^m.$$

Dann gilt auch

$$\overline{\varphi}_m(A^H)\tilde{r}^0 = \tilde{r}^m,\ \overline{\psi}_m(A^H)\tilde{r}^m = \tilde{p}^m,$$

wobei $\varphi_m(t) = \sum_{i=0}^m \beta_i t^i$ und $\overline{\varphi}_m(t) = \sum_{i=0}^m \overline{\beta}_i t^i$.

10.2 Definition
Auf Π = Menge der Polynome ist die Bilinearform $[\cdot,\cdot]$ erklärt durch:

$$[p,q] = \langle p(A)r^0, \overline{q}(A^H)\tilde{r}^0\rangle.$$

10.3 Lemma
$[\cdot,\cdot]$ ist bilinear und es gilt:

$$[p,q] = [pq, 1]$$

$$[p\vartheta, q] = [p, \vartheta q]\,,\ \text{wobei } \vartheta(t) = t$$

Der Beweis ist trivial.

Anwendung auf BiCG Koeffizienten:

$$\langle r^k, \tilde{r}^k\rangle = [\varphi_k, \varphi_k] = \left[\varphi_k^2, 1\right]$$

$$\langle Ap^k, \tilde{p}^k\rangle = [\vartheta\psi_k, \psi_k] = \left[\vartheta\psi_k^2, 1\right]$$

Setze also:

$$\Phi_k = \varphi_k^2 \in \overline{\Pi}_{2k},\ \Psi_k = \psi_k^2 \in \Pi_{2k},$$

dann gilt:

$$\Phi_{k+1} = (\varphi_k - \gamma_k\vartheta\psi_k)^2 = \Phi_k - 2\gamma_k\vartheta\varphi_k\psi_k + \gamma_k^2\vartheta^2\Psi_k$$

$$\Psi_{k+1} = (\varphi_{k+1} + \rho_k \psi_k)^2 = \Phi_{k+1} + 2\rho\varphi_{k+1}\psi_k + \rho_k^2 \Psi_k$$
$$\varphi_{k+1}\psi_k = \underbrace{\varphi_k\psi_k}_{\Upsilon_k} - \gamma_k \vartheta \Psi_k$$
$$\Upsilon_{k+1} = \varphi_{k+1}\psi_{k+1} = \Phi_{k+1} + \rho\left(\Upsilon_k - \gamma_k \vartheta \Psi_k\right)$$

Zusammenfassung:

1. $\gamma_k = \dfrac{[\Phi_k, 1]}{[\vartheta \Psi_k, 1]}$
2. $\Phi_{k+1} = \Phi_k - \vartheta\left(2\gamma_k \Upsilon_k + \gamma_k^2 \vartheta \Psi_k\right)$
3. $\rho_k = \dfrac{[\Phi_{k+1}, 1]}{[\Phi_k, 1]}$
4. $\Upsilon_{k+1} = \Phi_{k+1} + \rho_k\left(\Upsilon_k + \gamma_k \vartheta \Psi_k\right)$
5. $\Psi_{k+1} = \Phi_{k+1} + 2\rho_k\left(\Upsilon_k - \gamma_k \vartheta \Psi_k\right) + \rho_k^2 \Psi_k$

Unter Verwendung der Vektoren $r^k = \Phi_k(A) r^0$, $p^k = \Psi_k(A) t^0$, $u^k = \Upsilon_k(A) r^0$ und der Iterierten x^k mit $r^k = b - Ax^k$ ergibt dies folgendes neues Verfahren:

10.4 Algorithmus (CGS (CG-squared, Sommerfeld 1989))

wähle x^0, setze $r^0 = b - Ax^0$, wähle $\tilde{r}^0$, $p^0 = r^0 = u^0$
for $k = 0, 1, \ldots$ **do**
  $s^k = Ap^k$
  $\gamma_k = \dfrac{\langle r^k, \tilde{r}^0 \rangle}{\langle s^k, \tilde{r}^0 \rangle}$
  $r^{k+1} = r^k - A\left(2\gamma_k u^k + \gamma_k^2 s^k\right)$
  $x^{k+1} = x^k + 2\gamma_k u^k + \gamma_k^2 s^k$
  $\rho_k = \dfrac{\langle r^{k+1}, \tilde{r}^0 \rangle}{\langle r^k, \tilde{r}^0 \rangle}$
  $u^{k+1} = r^{k+1} + \rho\left(u^k + \gamma_k s^k\right)$
  $p^{k+1} = r^{k+1} + 2\rho_k\left(u^k - \gamma_k s^k\right) + \rho_k^2 p^k$
end for

10.5 Bemerkung

1. $x^k \in x^0 + K_{2k}(A, r^0)$ und 2 MVM mit A pro Schritt.
2. $r^k = \varphi_k^2(A) r^0 = \varphi_k(A) \underbrace{\varphi_k(A) r^0}_{r_k^{\mathrm{BiCG}}}$

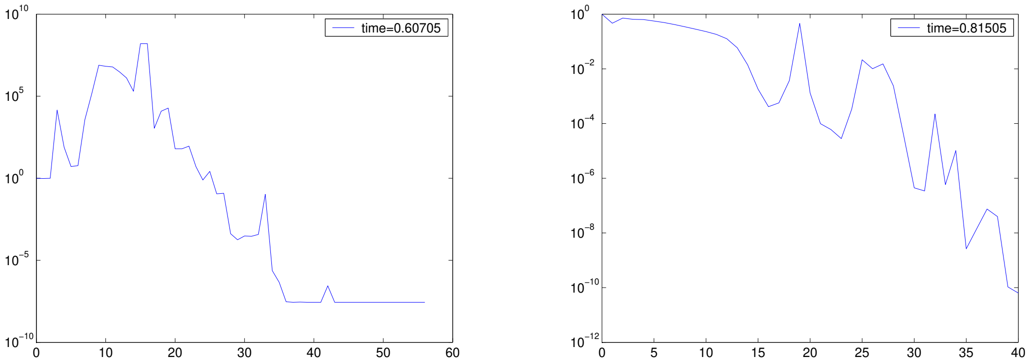


Abbildung 9: Konvergenzgeschichte für CGS, Modellproblem III. Links:$N = 20, \epsilon = 50$, rechts $N = 20, \epsilon = 1$. Man beachte für $\epsilon = 50$ die Stagnation bei 10^{-8}.

Ist das wirklich eine Verbesserung? Ein gutes φ_k für r^0 ist nicht notwendig ein gutes Polynom für $\varphi_k(A)r^0$. Ein schlechtes Polynom φ_k für r^0 ist wahrscheinlich auch schlecht für $\varphi_k(A)r^0$. Folge: CGS schwankt noch extremer als BiCG.

10.6 Aufgabe
Führe CGS für das Modellproblem III durch.

Neue Idee:
Nehme $\Phi_k = \chi_k\varphi_k$ und $\Psi_k = \chi_k\psi_k$, wobei $\chi_k(t) = \prod_{i=1}^k \left(1 - \dfrac{t}{w_i}\right)$, w_i geeignet gewählt zur Stabilisierung des Verfahrens.
Fragen:

1. Wie führt man eine geeignete Wahl von w_i durch ?

2. Wie bekommt man die alten BiCG Koeffizienten aus den neuen Verfahrenspolynom zurückgewonnen ?

Wir versuchen zuerst, die alten BiCG-Koeffizienten aus den neuen Verfahrenspolynomen zurückzugewinnen.

10.7 Lemma
Es gilt:

1. $[\varphi_k, p_{k-1}] = 0 \quad \forall p_{k-1} \in \Pi_{k-1}$

2. $[\psi_k, \vartheta p_{k-1}] = 0 \quad \forall p_{k-1} \in \Pi_{k-1}$

Beweis:

1. Nach Konstruktion des BiCG erfüllt (Petrov-Galerkin-Bedingung)
2. Beweis per Induktion:

$$[\psi_k, \vartheta p_{k-2}] = 0 \quad \text{nach Induktion und Teil 1}$$

$$[\psi_k, \vartheta\varphi_{k-1}] = [\vartheta\psi_k, \varphi_{k-1}] = \left[\frac{[\psi_k, \vartheta\psi_k]}{[\varphi_k, \varphi_k]}\left(\varphi_k - \varphi_{k+1}\right), \varphi_{k-1}\right] = 0$$

□

Folgerung:

$$[\varphi_k, \varphi_k] = \left[\prod_{i=0}^{k-1}(-\gamma_i)\vartheta^k + p_{k-1}, \varphi_k\right] = \left[\prod_{i=0}^{k-1}(-\gamma_i)\vartheta^k, \varphi_k\right] = \prod_{i=0}^{k-1}(-\gamma_i)\left[\vartheta^k, \varphi_k\right].$$

Also gilt:

$$[\Phi_k, 1] = \left[\prod_{i=0}^{k-1}\left(1 - \frac{t}{\omega_i}\right), \varphi_k\right] = \prod_{i=0}^{k-1}\left(-\frac{1}{\omega_i}\right)\left[\vartheta^k, \varphi_k\right].$$

Ebenso rechnet man nach, dass gilt:

$$[\psi_k, \vartheta\psi_k] = [\vartheta\psi_k, \psi_k] = \left[\vartheta\psi_k, \prod_{i=0}^{k-1}(-\gamma_i)\,\vartheta^k\right] = \prod_{i=0}^{k-1}(-\gamma_i)\left[\vartheta\psi_k, \vartheta^k\right],$$

$$[\Psi_k, \vartheta] = [\vartheta\Psi_k, 1] = \left[\vartheta\psi_k, \prod_{i=0}^{k-1}\left(1 - \frac{t}{\omega_i}\right)\right] = \prod_{i=0}^{k-1}\left(-\frac{1}{\omega_i}\right)\left[\vartheta\psi_k, \vartheta^k\right].$$

Daraus leitet sich folgender Zusammenhang her:

$$[\varphi_k, \varphi_k] = [\Phi_k, 1]\prod_{i=0}^{k-1}(\gamma_i\omega_i)$$

$$[\psi_k, \vartheta\psi_k] = [\Psi_k, \vartheta]\prod_{i=0}^{k-1}(\gamma_i\omega_i)$$

und damit

$$\gamma_k = \frac{[\varphi_k, \varphi_k]}{[\psi_k, \vartheta\psi_k]} = \frac{[\Phi_k, 1]}{[\Psi_k, \vartheta]},$$

$$\rho_k = \frac{[\varphi_{k+1}, \varphi_{k+1}]}{[\varphi_k, \varphi_k]} = \gamma_k\omega_k\frac{[\Phi_{k+1}, 1]}{[\Phi_k, 1]}.$$

Es ergeben sich also die Rekursionsvorschriften:

$$\begin{aligned}
\Phi_{k+1} &= \prod_{i=0}^{k}\left(1-\frac{t}{\omega_i}\right)\varphi_{k+1}\\
&= \prod_{i=0}^{k}\left(1-\frac{t}{\omega_i}\right)(\varphi_k-\gamma_k\psi_k)\\
&= \left(1-\frac{t}{\omega_k}\right)(\Phi_k-\gamma_k\vartheta\Psi_k)\,,\\
\Psi_{k+1} &= \Phi_{k+1}+\rho_k\left(1-\frac{t}{\omega_k}\right)\Psi_k
\end{aligned}$$

Jetzt bleibt noch, eine passende Wahl der ω_i zu finden. Setze

$$s^k = \Phi_k(A)r^0 - \gamma_k\vartheta\Psi_k(A)r^0,$$

dann ist s^k ein Residuum zu einer Zwischeniterierten y^k, denn $\Phi_k(0) - \gamma_k\Psi_k(0) = 1$. Wähle ω_k so das $\|\left(1-\frac{1}{\omega_k}A\right)s^k\|_2$ minimal wird, was nach Lemma 8.8 für

$$\omega_k = \frac{\langle As^k, As^k\rangle}{\langle s^k, As^k\rangle}$$

der Fall ist.Damit ergibt sich folgender BiCG-Stab Algorithmus:

10.8 Algorithmus (BiCG-Stab, van der Vorst, 1992)

wähle x^0, setze $r^0 = b - Ax^0$, $p^0 = r^0$, wähle $\tilde{r}^0$

for $k = 0, 1, \dots$ **do**

$u^k = Ap^k$

$\gamma_k = \dfrac{\langle r^k, \tilde{r}^0\rangle}{\langle u^k, \tilde{r}^0\rangle}$

$s^k = r^k - \gamma_k u^k$

$v^k = As^k$

$\omega_k = \dfrac{\langle v^k, v^k\rangle}{\langle s^k, v^k\rangle}$

$r^{k+1} = s^k - \dfrac{1}{\omega_k}v^k$

$\rho_k = \gamma_k\omega_k\dfrac{\langle r^{k+1}, \tilde{r}^0\rangle}{\langle r^k, \tilde{r}^0\rangle}$

$p^{k+1} = r^{k+1} + \rho_k p_k - \dfrac{\rho_k}{\omega_k}u^k$

$x^{k+1} = x^k + \dfrac{1}{\omega_k}s^k + \gamma_k p^k$

end for

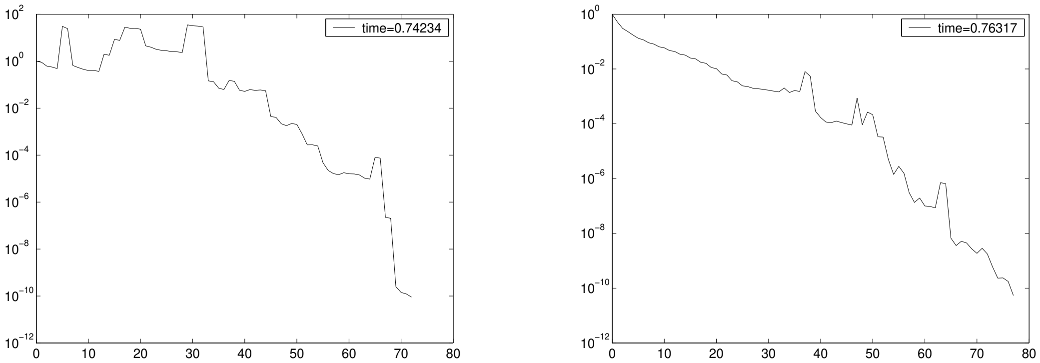


Abbildung 10: Konvergenzgeschichte für BiCGStab, Modellproblem III. Links:$N = 20, \epsilon = 50$, rechts $N = 20, \epsilon = 1$. Man beachte die Verbesserungen gegenüber CGS.

10.9 Aufgabe
Wende BiCG-Stab auf das Modellproblem III an.

10.10 Bemerkung
BiCG-Stab ist sehr beliebt, da 2 MVM mit A pro Schritt nötig sind und $r^k \in K_{2k}(A, r^0)$. Außerdem ist die Konvergenz relativ glatt.

11 Minimales Residuum vs kurze Rekursion

Wir schauen uns normale Matrizen zunächst genauer an.

11.1 Lemma
A ist normal, falls eine der folgenden Bedingungen erfüllt ist:

(i) $AA^H = A^H A$

(ii) $A = Q\Lambda Q^H \quad$ mit Λ diagonal, Q unitär,

(iii) $A^H = p(A), \quad p \in \Pi_n$.

Beweis: (i) $\Longleftrightarrow$ (ii) ist bekannt aus LA.
(ii) $\Longrightarrow$ (iii): $A^H = Q\overline{\Lambda}Q^H = Qp(\Lambda)Q^H = p(A)$, wobei p das Polynom aus Π_k ist mit $p(\lambda_i) = \overline{\lambda_i}, \quad i = 1, \ldots, k$. Dabei sind $\lambda_1, \ldots, \lambda_k$ die *verschiedenen* Eigenwerte von A.
(iii) $\Longrightarrow$ (i): trivial. □

11.2 Definition
Für $A \in \mathbb{C}^{n \times n}$ bezeichnet $d(A)$ die Anzahl der verschiedenen Eigenwerte $\lambda_1, \ldots, \lambda_{d(A)}$.

11.3 Lemma
Sei $\langle \, , \, \rangle_*$ ein Innenprodukt auf $\mathbb{C}^n$. Dann existiert ein $B \in \mathbb{C}^{n \times n}$, B hpd mit

$$\langle x, y \rangle_* = \langle x, By \rangle \;\; (\langle \, , \, \rangle \text{ übliches Skalarprodukt}).$$

Beweis: Übung. □

11.4 Definition
Die Adjungierte A^* zu A bezüglich $\langle \, , \, \rangle_*$ ist die Matrix mit

$$\langle Ax, y \rangle_* = \langle x, A^* y \rangle_* \quad \forall \, x, y.$$

11.5 Definition
A ist normal bzgl. $\langle \, , \, \rangle_*$, falls $A^*A = AA^*$.

11.6 Satz
Sei $\langle \, , \, \rangle_*$ gegeben und B hpd mit $\langle x, y \rangle_* = \langle y, By \rangle$. Dann gilt: A ist normal bzgl. $\langle \, , \, \rangle_* \Longleftrightarrow B^{\frac{1}{2}} A B^{-\frac{1}{2}}$ ist normal bzgl. $\langle \, , \, \rangle$.

Beweis:

$$
\begin{aligned}
& & \langle Ax, y\rangle_* &= \langle x, A^*y\rangle_* \\
&\Longleftrightarrow & \langle Ax, By\rangle &= \langle x, BA^*y\rangle \\
&\Longleftrightarrow & \langle BAx, y\rangle &= \langle x, BA^*y\rangle \quad \forall\, x, y \\
&\Longleftrightarrow & (BA^*)^H &= BA \\
&\Longleftrightarrow & BA^* &= A^H B \\
&\Longleftrightarrow & A^* &= B^{-1}A^H B.
\end{aligned}
$$

Es ist also

$$
\begin{aligned}
& & A^*A &= AA^* \\
&\Longleftrightarrow & B^{-1}A^H BA &= AB^{-1}A^H B \\
&\Longleftrightarrow & B^{-\frac{1}{2}}A^H B^{\frac{1}{2}}B^{\frac{1}{2}}AB^{-\frac{1}{2}} &= B^{\frac{1}{2}}AB^{-\frac{1}{2}}B^{-\frac{1}{2}}A^H B^{\frac{1}{2}} \\
&\Longleftrightarrow & \tilde{A}^H\tilde{A} &= \tilde{A}\tilde{A}^H
\end{aligned}
$$

mit $\tilde{A} = B^{\frac{1}{2}}AB^{-\frac{1}{2}}$. □

Bezeichnung: Statt „A normal bzgl. $\langle\ ,\ \rangle_*$“ notieren wir jetzt: „A ist B-normal“ (B ist die zu $\langle\ ,\ \rangle_*$ gehörige hpd-Matrix).

11.7 Korollar
A ist B-normal $\Longleftrightarrow \exists\, p \in \Pi_n$ mit $A^* = p(A)$.

Beweis:

$$
\begin{aligned}
A \text{ ist } B\text{-normal} &\Longrightarrow & \tilde{A} &= B^{\frac{1}{2}}AB^{-\frac{1}{2}} \text{ normal} \\
&\Longrightarrow & \tilde{A}^H &= p(\tilde{A}) = B^{\frac{1}{2}}p(A)B^{-\frac{1}{2}} \\
&\Longrightarrow & B^{\frac{1}{2}}\tilde{A}^H B^{-\frac{1}{2}} &= p(A) \\
&\Longrightarrow & A^* &= p(A)
\end{aligned}
$$

□

11.8 Definition
A heißt B-normal(t), falls ein $p \in \Pi_t$ existiert mit $A^* = p(A)$.

11.9 Satz
A sei B-normal(t) mit minimalem $t > 1$. Dann besitzt das Minimalpolynom q von A den Grad $\deg(q) = d(A) \le t^2$.

Beweis: Die normale Matrix $\tilde{A}$ ist diagonalisierbar und damit auch A. Der Grad des Minimalpolynoms von A ist also tatsächlich $d(A)$. Sei $A^* = p(A)$, $\deg(p) = t > 1$. Dann ist $\tilde{A}^H = p(\tilde{A})$ mit $\tilde{A} = B^{\frac{1}{2}}AB^{-\frac{1}{2}}$, woraus sich mit $\tilde{A} = Q\Lambda Q^H$ die Beziehung

$$Q\overline{\Lambda}Q^H = Qp(\Lambda)Q^H$$

ergibt. Also ist $p(\lambda_i) = \overline{\lambda}_i$ für $i = 1, \ldots, d(A)$. Wegen $\overline{p}(\overline{\lambda}_i) = \lambda_i$ folgt $\overline{p}(p(\lambda_i)) = \lambda_i, \quad i = 1, \ldots, d(A)$. Das Polynom $\tilde{q}(s) = \overline{p}(p(s)) - s$ hat den Grad t^2 und die Eigenschaft $\tilde{q}(\lambda_i) = 0, \quad i = 1, \ldots, d(A)$.
Also ist $d(A) \leq \deg \tilde{q} = t^2$. $\square$

11.10 Satz
A sei B-normal(1) und $d(A) > 1$. Dann gilt

$$\tilde{A} = e^{i\theta}\left(\frac{r}{2}I + F\right), \tag{11.1}$$

wobei $\theta \in \mathbb{R}, r \in \mathbb{R}, F \in \mathbb{C}^{n\times n}$mit $F^H = -F$ (schiefsymmetrisch).

Beweis: Wir merken zunächst an, dass (11.1) äquivalent formuliert werden kann als

$$\tilde{A} = e^{i\theta'}\left(i\frac{r}{2}I + F'\right) \text{ mit } F' = (F')^H, r \in \mathbb{R}.$$

Jetzt sei $\tilde{A}^H = p(\tilde{A})$ mit $\deg(p) = 1,\ p(s) = as + b$. Wie vorher sei

$$\tilde{q}(s) = \overline{p}(p(s)) - s = \overline{a}(as + b) + \overline{b} - s = (\overline{a}a - 1)s + (\overline{a}b + \overline{b}).$$

Das Polynom $\tilde{q}(s)$ erfüllt $\tilde{q}(A) = 0$, denn es gilt $\tilde{q}(\lambda_i) = 0, \quad i = 1, \ldots, d(A)$.
Also ist $\tilde{q} \equiv 0$ (da $d(A) > 1$) und damit $\overline{a}a - 1 = 0, \quad \overline{a}b + \overline{b} = 0$.
Wir erhalten so

$$a = -\frac{b}{\overline{b}}$$

und damit (mit $b = re^{i\theta}$)

$$p(s) = -e^{i\theta}\left(se^{i\theta} - r\right).$$

Nach Voraussetzung gilt für $i = 1, \ldots, d(A)$

$$\begin{aligned}
& p(\lambda_i) = \overline{\lambda}_i, \quad i = 1, \ldots, d(A) \\
\iff\quad & -e^{i\theta}\left(\lambda_i e^{i\theta} - r\right) = \overline{\lambda}_i \\
\iff\quad & \lambda_i e^{i\theta} - r = -\overline{e^{i\theta}\lambda_i} \\
\iff\quad & \lambda_i e^{i\theta} + \overline{e^{i\theta}\lambda_i} = r
\end{aligned}$$

$$\iff \qquad \Re(\lambda_i e^{i\theta}) = \frac{r}{2}.$$

Also sind alle Eigenwerte λ_j von der Form

$$\lambda_j = \frac{r}{2} e^{-i\theta} + i e^{-i\theta} \mu_j, \quad \mu_j \in \mathbb{R}.$$

Wir erhalten somit für $F = \tilde{A} - \frac{r}{2} e^{-i\theta} I$

$$\begin{aligned} & F = Q^H \Lambda Q \text{ mit } \Lambda = \operatorname{diag}(i\mu_j) \\ \Longrightarrow \quad & F^H = -F. \end{aligned}$$

Also ist

$$\tilde{A} = e^{i\theta} \left(\frac{r}{2} I + F \right)$$

wie behauptet. □

Wir betrachten jetzt Iterations-Verfahren mit kurzer Rekursion der Gestalt

$$x^{k+1} = x^k + \alpha_k p^k, \tag{11.2}$$

$$p^{k+1} = A p^k - \sum_{j=k-s+2}^{k} \beta_{k,j} p^j \quad \text{mit } p^0 = r^0 = b - A x^0. \tag{11.3}$$

Nach eventueller Umformulierung ist CG von dieser Gestalt ($s = 3$) und MINRES ($s = 3$). Das Verfahren soll minimale Residuen erreichen in einer von einem Innenprodukt $\langle\ ,\ \rangle_*$ induzierten Norm. Die zugehörige hpd-Matrix sei B.

Überlegungen: $K_m(A, r^0) = \operatorname{span}\{p^0, \ldots, p^{m-1}\}$. Es gilt für die Fehler $e^k = x^k - A^{-1} b$

$$e^{k+1} = e^k + \alpha_k p^k, \quad k = 0, 1, \ldots. \tag{11.4}$$

Wegen

$$x^k = x^0 + \sum_{j=0}^{k-1} \alpha_j p^j$$

folgt

$$e^k = e^0 + \underbrace{\sum_{j=0}^{k-1} \alpha_j p^j}_{\in K_k(A, r^0)} .$$

Da e^k minimal sein soll, folgt:

$$\langle e^k, p_j \rangle_* = 0, \quad j = 0, \dots, k-1. \tag{11.5}$$

(11.4) in (11.5) gibt:

$$\langle e^{k+1}, p_j \rangle_* = \langle e^k + \alpha_k p_k, p_j \rangle_* = 0, \quad j = 0, \dots, k-1.$$

Für $j < k$ folgt

$$\begin{aligned} & \alpha_k \langle p_k, p_j \rangle_* = 0 \\ \Longrightarrow \quad & \alpha_k = 0 \quad \text{oder} \quad \langle p_k, p_j \rangle_* = 0 \end{aligned} \tag{11.6}$$

Für $j = k$ ergibt sich die definierende Gleichung für α_k

$$\langle e^k, p^k \rangle_* + \alpha_k \langle p_k, p_k \rangle_* = 0.$$

Es ist daher sinnvoll zu fordern:

$$\langle p_k, p_j \rangle_* = 0, \quad j = 1, \dots, k-1, \quad k = 0, 1, \dots. \tag{11.7}$$

Für eine präzise Formulierung des nun anstehenden Satzes von Faber-Manteuffel benötigen wir noch etwas zusätzliche Notation.

11.11 Definition
Sei $A \in \mathbb{C}^{n \times n}$ gegeben.

(i) $d(A)$ ist Grad des Minimal-Polynoms $\psi \in \overline{\Pi}_{d(A)}$ von A. Es gilt also $\psi(A)v = 0$ für alle $v \in \mathbb{C}^n$.

(ii) Sei $V \subseteq \mathbb{C}^n$ Unterraum. Dann ist

$$d_V(A) = \min\{d \in \mathbb{N} : \text{ex. } p \in \Pi_d \text{ mit } p(A)v = 0 \text{ für alle } v \in V\}$$

Das zugehörige Polynom aus $\overline{\Pi}_{d_V(A)}$ notieren wir ψ_V.

(iii) Ist speziell $V = \langle v \rangle$ ein-dimensional, so schreiben wir

$$d_v(A) =: d_v \text{ and } \psi_{\langle v \rangle} =: \psi_v.$$

Folgerungen:

(i) $K_m(A, v) = K_{m+1}(A, v)$ gilt erstmalig für $m = d_v$.

(ii) ψ_V aus Definition 11.11 (ii) ist Teiler von ψ.

11.12 Lemma
Sei $V = V_1 + V_2, \quad V_1, V_2 \subseteq \mathbb{C}^n$ Unterräume. Dann gilt

$$\psi_V = \mathrm{kgV}\,(\psi_{V_1}, \psi_{V_2})\,.$$

11.13 Definition
Ein Verfahren der Gestalt

$$\left\{ \begin{array}{ll} x^{k+1} & = x^k + \alpha_k p^k \\ p^{k+1} & = Ap^k - \sum_{j=k-s+2}^{k} \beta_{k,j} p^j \end{array} \right. \tag{11.8}$$

mit $s \geq 3$ und

$$\langle p^k, p^j \rangle_* = 0, \quad j = 0, \dots, k-1 \quad k = 0, \dots, d_{p^0} - 1 \tag{11.9}$$

für jede Wahl von p^0 (= jede Wahl von x^0 in $p^0 = b - Ax^0$) heißt *s-Term-CG-Verfahren*

Unsere Vorüberlegungen haben bereits gezeigt:

11.14 Lemma
Die Iterierten eines s-Term-CG-Verfahrens besitzen minimalen Fehler

$$\|A^{-1}b - x^k\|_* = \min_{x \in x^0 + K_k(A, r^0)} \|A^{-1}b - x\|_*.$$

Außerdem gilt

$$\langle e^k, p^k \rangle_* + \alpha_k \langle p_k, p_k \rangle_* = 0.$$

11.15 Satz (Faber-Manteuffel, 1984)
Ein s-Term-CG-Verfahren (bzgl. $\langle \cdot, \cdot \rangle_*$) für $A \in \mathbb{C}^{n \times n}$ existiert $\Longleftrightarrow d(A) \leq s$ oder A ist B-normal$(s-2)$ ($B \in \mathbb{C}^{n \times n}$ mit $\langle x, y \rangle_* = \langle x, By \rangle$)

Beweis: „$\Longleftarrow$ “
Setze Verfahren an mit Bauart (11.8).

1. Fall: $d(A) \leq s$
Dann ist $d_{p^0} \leq d(A) \leq s$ für alle Anfangsresiduen p^0. Durch geeignete Bestimmung der $\beta_{k,j}$ in (11.8) kann (mit Gram-Schmidt) deshalb erreicht werden:

$$\langle p_k, p_j \rangle_* = 0 \quad j = 0, \dots, k-1 \quad k = 0, \dots, d_{p^0} - 1.$$

2. Fall: $d(A) > s$ und A ist B-normal(s-2).
Induktion über k. Für $k \leq s-2$ kann (11.9) immer erreicht werden durch Gram-Schmidt Orthogonalisierung). Sei nun $k \geq s-1$. Nun gelte (11.9) bis

$k \geq s-2$. Bestimme $\beta_{k+1,j} \quad j = k-s+2, \ldots, k$ so, dass $\langle p^{k+1}, p^j \rangle_* = 0, \quad j = k-s+2, \ldots, k$. Für $j < k-s+2$ gilt

$$\begin{aligned}\langle p^{k+1}, p^j \rangle_* &= \langle Ap^k, p^j \rangle_* \\ &= \langle p^k, A^* p^j \rangle_* \\ &= \langle p^k, \underbrace{q_{s-2}(A) \cdot p^j}_{\substack{\in K_k(A,p^0) \\ = \langle p^0, \ldots, p^{k-1} \rangle}} \rangle_* \\ &= 0 \quad \text{nach Induktions-Annahme.}\end{aligned}$$

Also gilt (11.9) auch für $k+1$.

„$\Longrightarrow$“ (Beweis aufwendig!)
Es existiere ein s-Term-CG-Verfahren und es sei $d(A) > s$. Zu zeigen ist: A ist B-normal$(s-2)$. Zur Vereinfachung der Notation betrachten wir nur $B = I$, d.h. $\langle\ ,\ \rangle = \langle\ ,\ \rangle_*$. Wir zeigen zuerst, dass A normal, also orthogonal diagonalisierbar ist. Nach Voraussetzung existiert $\tilde{p} \in \mathbb{C}^{n \times n}$ mit $d_{\tilde{p}} = d(A) > s$. Also existiert auch $p \in \mathbb{C}^n$ mit $d_p = s$, denn ist $\psi_{\tilde{p}}(t) = \prod_{i=1}^{d(A)} t - \lambda_i$, so nehme man $p = \prod_{i=1}^{d(A)-s} (A - \lambda_i I)\tilde{p}$. Sei $V = K_s(A, p)$. V ist ein A-invarianter Unterraum mit $d_V = s$. Wir werden zeigen: Eingeschränkt auf V ist A orthogonal diagonalisierbar. Dazu brauchen wir aber zunächst ein Stetigkeitsresultat über die p^k aus (11.8) als Funktionen

$$p^k = F_k(p^0)$$

von p^0. $F_k(p^0)$ ist zunächst nur definiert, falls $p^j \neq 0$ for $j = 0, \ldots, k-1$. $d_{p^0} > k$.

11.16 Lemma
Für $k = 1, \ldots, d(A) - 1$ ist F_k stetig fortsetzbar auf ganz $\mathbb{C}^n$ und es gilt $F_k(p^0) \leq \| A \|^k \cdot \| p^0 \|$.

Beweis: Sei

$$W = \bigcup_{\substack{\psi | \psi(A), \\ \deg(\psi) < d(A)}} \operatorname{kern}(\psi(A)).$$

W ist eine endliche Vereinigung von Unterräumen mit Dimension kleiner als n. Also gilt $\overline{\mathbb{C}^n \backslash W} = \mathbb{C}^n$. Nun sei $k = 1$ und $p^0 \notin W$. Dann gilt

$$F_1(p^0) = Ap^0 - \beta_{1,0} p^0$$

mit

$$\|F_1(p^0)\| \leq \|Ap^0\| \leq \|A\| \|p^0\|.$$

Also ist F_1 stetig und beschränkt auf $\mathbb{C}^n \backslash W$ und damit stetig fortsetzbar auf ganz $\mathbb{C}^n$. Die Fortsetzung heiße wieder F_1.
Für $k > 1$ folgt das Lemma per Induktion genauso. $\square$
Nun sei Q die Orthogonalprojektion auf $V = \langle p, Ap, \ldots, A^{s-1}p \rangle$, insbesondere $Q = Q^H$.

$$((AQ)^* =)\,(AQ)^H = QA^H\,(= QA^*)\,.$$

Setze $\tilde{A} = AQ$. Es gilt $d(\tilde{A}) = s$, da $d_p = s$.
Sei $\tilde{p}$ irgendein Vektor aus V mit $d_{\tilde{p}} = s$ und sei

$$\begin{aligned} f(\tilde{p}) &= \langle F_{s-1}(\tilde{p}), \tilde{A}^*\tilde{p} \rangle \\ &= \langle \tilde{p}^{s-1}, \tilde{A}^*\tilde{p} \rangle \\ &= \langle \tilde{p}^{s-1}, QA^*\tilde{p} \rangle \\ &= \langle Q\tilde{p}^{s-1}, A^*\tilde{p} \rangle \\ &= \langle \tilde{p}^{s-1}, A^*\tilde{p} \rangle \end{aligned}$$

Zwischenbehauptung: Es ist $f(p) = \langle p^{s-1}, p^0 \rangle = 0$ für alle p
Beweis: Da ein s-Term-CG-Verfahren existiert, gilt nach (11.9) für $k = s-1$

$$\langle p^s, p^0 \rangle = \langle Ap^{s-1} - \sum_{j=1}^{s-1} \beta_{kj} p^j, p^0 \rangle = 0, \qquad \text{falls } d(p) \geq s.$$

Falls $d(p) \leq s$, kann $f(p) = \langle AF_{s-1}(p), p \rangle$ stetig fortgesetzt werden auf ganz V und ist damit $= 0$ auf ganz V. Insbesondere gilt also auch $f(p) = 0$ für $d_p = s$. $\square$
Weiter im Beweis: Es gilt $\tilde{A}^*\tilde{p} = QA^*\tilde{p} \in V$. Da $\tilde{p}^0, \ldots, \tilde{p}^{s-1}$ Orthonormalbasis von V ist, gilt wegen der Zwischenbehauptung sogar

$$\tilde{A}^*\tilde{p} \in \langle \tilde{p}^0, \ldots, \tilde{p}^{s-2} \rangle = \langle \tilde{p}^0, A\tilde{p}^0, \ldots, A^{s-2}\tilde{p}^0 \rangle. \tag{11.10}$$

Betrachte

$$W(p) = p \wedge Ap \wedge \cdots \wedge A^{s-2}p \wedge \tilde{A}^*p.$$

Als Multilinearform ist $W(p)$ stetig; als alternierende Multilinearform erfüllt W wegen (11.10) die Beziehung $W(p) = 0$ für alle p mit $d_p = s$. Aus Stetigkeitsgründen gilt dies auch für $d_p < s$. Sei speziell p jetzt so, dass $d_p = s - 1$. Da $p, \ldots, A^{s-2}p$ linear unabhängig sind, folgt aus $W(p) = 0$ also

$$\tilde{A}^*p \in \langle p, \ldots, A^{s-2}p \rangle. \tag{11.11}$$

Sei $V_{s-1} \subseteq V$ ein Unterraum der Dimension $s - 1$. Da $V = \langle p, Ap, \ldots, A^{s-1}p \rangle$ existiert ein (neues) p mit $V_{s-1} = \langle p, Ap, \ldots, A^{s-2}p \rangle$ und $d_p = s - 1$.

Sei $b_1, \ldots, b_{s-1}$ Basis von $V_{s-1}^{(1)}$ mit $d_{b_i} = s-1$ für $i = 1, \ldots, s-1$. Es gilt damit nach (11.11)

$$\tilde{A}^* b_i \in \langle b_i, \ldots, A^{s-2} b_i \rangle = V_{s-1}^{(1)} \; .$$

Also ist $V_{s-1}^{(1)}$ ein $\tilde{A}^*$-invarianter Unterraum.
Sei q der (bis auf skalare Vielfache) eindeutige Vektor aus V mit $q^1 \perp V_{s-1}^{(1)}$. Es gilt

$$\langle Aq^1, y \rangle = \langle \tilde{A} q^1, y \rangle = \langle q^1, \tilde{A}^* y \rangle = 0 \qquad \text{für alle } y \in V_{s-1}^{(1)}.$$

Also gilt

$$Aq^1 = \tilde{A} q^1 = \lambda_1 q^1 \qquad \text{und} \qquad q^1 \perp V_{s-1}^{(1)}.$$

Damit haben wir einen ersten Eigenvektor von A gefunden.
Sei $V_{s-1}^{(2)}$ ein A-invarianter Unterraum von V der Dimension $s-1$ mit $q^1 \in V_{s-1}^{(2)}$. Mit der gleichen Argumentation folgt für $q^2 \perp V_{s-1}^{(2)}$

$$Aq^2 = \lambda_2 q^2 \qquad \text{und} \qquad q^2 \perp V_{s-1}^{(2)},$$

insbesondere $q^2 \perp q^1$.
So fortfahrend erhält man eine Orthonormalbasis $q^1, \ldots, q^s$ von V mit zugehörigen Eigenwerten $\lambda_1, \ldots, \lambda_s$. So gilt $Aq^1 = \lambda_i q^i$ für $i = 1, \ldots, s$.
Jetzt kümmern wir uns noch um ganz $\mathbb{C}^n$. Angenommen, es existieren x und $\lambda \in \mathbb{C}$ mit

$$(A - \lambda I)^2 x = 0 \qquad \text{und} \qquad (A - \lambda I)x \neq 0. \tag{11.12}$$

Wähle einen invarianten Unterraum V_s mit $s-2$ linear unabhängigen Eigenvektoren von A (wir wissen bereits, dass es mindestens s gibt) und x und $(A - \lambda I)x$. V_s ist A-invariant. Es ist $d_{V_s}(A) = s$. Mit der gleichen Argumentation von vorher folgt, dass A auf V_s orthogonal diagonalisierbar ist, im Widerspruch zu (11.12).
Also existieren keine Haupträume der Stufe > 1. Folglich ist A diagonalisierbar. Sind q_i, q_j Eigenvektoren zu verschiedenen Eigenwerten, so ergänze sie zu einem A-invarianten Unterraum der Dimension s. Dann folgt nach dem bisherigen $q_i \perp q_j$.

Wir haben damit gezeigt, dass A normal ist. Noch zu zeigen ist: A ist normal$(s-2)$.
Sei dazu mit $d = d(A)$

$$A^* = q_d(A) = \sum_{i=0}^{d-1} \gamma_i A^i ..$$

Ein solches Polynom existiert als dasjenige mit $q(\lambda_j) = \overline{\lambda}_j$ für die d verschiedenen Eigenwerte λ_j von A. Sei weiter

$$A^d = \sum_{i=0}^{d-1} \beta_i A^i.$$

Wir verwenden

$$f_i(p) = \langle p^i, A^* p\rangle = 0, \qquad s-1 \leq i \leq d(A)-2.$$

Für p mit $d_p < d$ gilt also

$$A^* p \in \langle p, p^1, \ldots, p^{s-2}\rangle.$$

Sei speziell $d_p = d-1$. Alle solche p besitzen Minimalploynome ψ_p der Gestalt

$$\psi_p(t) = \prod_{\substack{i=1\\ i\neq j}}^{d} (t-\lambda_i) \qquad \text{mit } \prod_{i=1}^{d}(t-\lambda_i) \text{ Minimalpolynom von } A,$$

also

$$\psi_p(t) = t^{d-1} - \Big(\sum_{\substack{i=1\\ i\neq j}}^{d} \lambda_i\Big) t^{d-2} + \cdots = t^{d-1} + \sum_{i=0}^{d-2} \delta_i t^i.$$

Die Vektoren $p, Ap, \ldots, A^{d-2}p$ sind linear unabhängig. Es ist

$$A^* p = \sum_{i=0}^{d-1} \gamma_i A^i p.$$

Also gilt

$$\begin{aligned} A^* p &= \sum_{i=0}^{d-2} (\gamma_i - \delta_i \gamma_{d-1}) A^i p \\ &= \left(\gamma_{d-1}\Big(\sum_{i\neq j} \lambda_i\Big) + \gamma_{d-2}\right) A^{d-2} p + \cdots . \end{aligned}$$

Da $A^* p \in \langle p, \ldots, A^{s-2}p\rangle$ folgt

$$\gamma_{d-1}\Big(\sum_{i\neq j} \lambda_i\Big) + \gamma_{d-2} = 0 \qquad \text{für alle } j.$$

Also folgt $\gamma_{d-1} = \gamma_{d-2} = 0$. Aus $\gamma_{d-1} = 0$ folgt aber weiter $\gamma_i = 0$ für $i > s-2$, denn $p, \dots, A^{d-2}p$ sind linear unabhängig und $A^* p \in \langle p, \dots, A^{s-2}p \rangle$.
Dies beendet den Beweis zu Satz 11.15. $\square$
Beachte: Satz 11.15 setzt eine bestimmte Form für kurze Rekursion voraus. Alternative: z.B. gekoppelte kurze Rekursion. Dann könnten doch weitere Verfahren mit minimalem Residuum und kurzer Rekursion existieren.

11.17 Beispiel
Sei $A \in \mathbb{C}^{n\times n}$ unitär, also $AA^H = I$. Es exisitiert eine kurze Rekursion zur Berechnung der Arnoldi-Vektoren

$$h_{k,k+1} v^{k+1} = Av^k - \sum_{j=1}^{k} h_{k,j} v^j \qquad \text{mit } \left\|v^k\right\|_2 = 1 \text{ und } v^0 = \frac{1}{\|r^0\|} r^0.$$

Die Vektoren $v^0, \dots, v^k$ sind eine Orthonormalbasis von $K_{k+1}(A, r^0)$ und $Av^0, \dots, Av^k$ ist Orthonormalbasis von $AK_{k+1}(A, r^0)$.
Damit ist $w^k, Av^0, \dots, Av^{k-1}$ bei geeigneter Wahl von w^k eine Orthonormalbasis von $K_{k+1}(A, r^0)$. Eine geeignete Wahl ergibt sich induktiv über

$$\begin{aligned}
\tilde{w}^k &= w^{k-1} - \sum_{j=0}^{k-1} \frac{\langle w^{k-1}, Av^j \rangle}{\langle Av^j, Av^j \rangle} Av^j \qquad (\langle Av^j, Av^j \rangle = 1) \\
&= w^{k-1} - \sum_{j=0}^{k-1} \langle w^{k-1}, Av^j \rangle Av^j \qquad (w^{k-1} \perp Av^j \text{ für } j = 0, \dots, k-2) \\
&= w^{k-1} - \langle w^{k-1}, Av^{k-1} \rangle Av^{k-1} \\
w^k &= \frac{1}{\|\tilde{w}^k\|} \tilde{w}^k.
\end{aligned}$$

Die Berechnung von v^{k+1} geht jetzt so:

$$\tilde{v}^{k+1} = Av^k - \left(\sum_{j=1}^{k} \beta_{k,j} Av^{j-1} \right) - \beta_{k,0} w^k$$

$$\begin{aligned}
\text{mit} \qquad \beta_{k,j} &= \langle Av^k, Av^{j-1} \rangle = 0 \qquad \text{da } v^k \perp v^{j-1} \\
\beta_{k,0} &= \langle Av^k, w^k \rangle
\end{aligned}$$

Wir erhalten also folgenden Algorithmus 11.18 zur Berechnung der Arnoldi-Vektoren.

11.18 Algorithmus (Unitäres Arnoldi-Verfahren)

$v^0 := \big(1/\|r^0\|\big) r^0$
$w^0 := v^0$
for $k = 0, 1, 2, \dots$ **do**
  $\beta_k = \langle Av^k, w^k \rangle$
  $\tilde{v}^{k+1} := Av^k - \beta_k w^k$
  $v^{k+1} := \big(1/\|\tilde{v}^{k+1}\|\big) \tilde{v}^{k+1}$
  $\tilde{w}^{k+1} := w^k - \overline{\beta}_k Av^k$
  $w^{k+1} := \big(1/\|\tilde{w}^{k+1}\|\big) \tilde{w}^{k+1}$
end for

Jetzt ist $H_{k,k}$ und $H_{k+1,k}$ aus dem Arnoldi-Prozess nicht mehr explizit gegeben. Man kann aber zeigen (Jagels und Reichel, 1994)

1. $H_{k,k}$ wird durch Algorithmus 11.18 in faktorisierter Form (als Produkt von Rotationen) implizit gegeben.

2. Für Systeme der Form
$$(\zeta I + \rho A)x = b$$
mit $A^H A = I$ und $\zeta, \rho \in \mathbb{C}$ ist $K_m(\zeta I + \rho A, r^0) = K_m(A, r^0)$ und es existiert auch eine kurze Rekursion zur Bestimmung der Iterierten x^k mit
$$\big\|b - (\zeta I + \rho A)x^k\big\|_2 = \min_{p_k \in \bar{\Pi}_k} \big\|p_k(\zeta I + \rho A)b\big\|_2.$$
Die Herleitung erfolgt wie bei MINRES unter Ausnutzung der faktorisierten Form von $H_{k+1,k}$

12 Modellproblem IV: Overlap-Fermionen in der Gittereichtheorie

Die Gittereichtheorie befasst sich mit der Diskretisierung der Quantenchromodynamik (QCD). QCD ist die Theorie der starken Wechselwirkung zwischen den Quarks.

12.1 Definition
Die *Wilson-Fermi-Matrix* M beschreibt eine Nächste-Nachbar-Kopplung auf einem äquidistanten 4-dimensionalen Gitter

$$G = \big\{x \in \{1, \dots, N\}^4\big\} \qquad e_1 = (1,0,0,0),\ e_2 = (0,1,0,0),\ \dots$$

durch

$$(M\psi)_x = \psi_x - \kappa\Bigg(\sum_{\mu=1}^{4}\big(U_x^\mu \otimes (I-\gamma_\mu)\big)\psi_{x-e_\mu} + \sum_{\mu=1}^{4}\big(U_{x+e_\mu}^\mu \otimes (I+\gamma_\mu)\big)\psi_{x+e_\mu}\Bigg) \tag{12.1}$$

mit $\psi_x \in \mathbb{C}^{12}$ (12 Unbekannte pro Gitterpunkt), $\gamma_\mu \in \mathbb{C}^{4\times 4}$, $U_x^\mu \in \mathrm{SU}(3)$ (unitäre Matrizen mit Determinante $+1$). Dabei ist $\kappa \in (0,4)$ anpassbar und U_x^μ zufällig erzeugt. Gleichung (12.1) ist an den Rändern von G periodisch zu verstehen.
Es ist also $M \in \mathbb{C}^{n\times n}$ mit $n = 12 \cdot N^4$. Heutzutage (Sommer 2003) liegt N zwischen 32 und 128.

N	n
32	ca. $1.2 \cdot 10^7$
64	ca. $2 \cdot 10^8$
128	ca. $3 \cdot 10^9$

Beim sog. *Overlap-Fermionen-Modell* muss man Systeme der Gestalt

$$\big(I + m\,\Gamma_5\,\mathrm{sgn}(Q)\big)\psi = \phi \tag{12.2}$$

lösen mit Γ_5 einfache Permutations-Matrix, $Q = \Gamma_5 M$ hermitisch, $Q = V\Lambda V^H$ und $m \lesssim 1$.
Dabei ist $\mathrm{sgn}(Q)$ definiert als $\mathrm{sgn}(Q) = V\,\mathrm{sgn}(\Lambda)V^H$ mit

$$\mathrm{sgn}(\Lambda) = \mathrm{diag}(\mathrm{sgn}(\lambda_i))$$

für $\Lambda = \mathrm{diag}(\lambda_i)$.

12.2 Bemerkung

1. Es gilt $\Gamma_5 = \Gamma_5^H = \Gamma_5^{-1}$ (siehe Physik) und $\operatorname{sgn}(Q) \cdot \operatorname{sgn}(Q) = I$ (falls Q regulär) mit $\operatorname{sgn}(Q) = (\operatorname{sgn}(Q))^H$. Also sind Γ_5 und $\operatorname{sgn}(Q)$ unitär, also auch $\Gamma_5 \cdot \operatorname{sgn}(Q)$. Auf (12.2) kann also das Verfahren vom Ende des letzten Abschnittes angewendet werden.

2. $\operatorname{sgn}(Q)$ kann man praktisch nicht explizit berechnen. Der Aufwand wäre $O(n^3)$ für die Zeit und $O(n^2)$ für den Speicher. Für ein KUV zur Lösung von (12.2) genügt es aber, wenn man in jedem Schritt $\operatorname{sgn}(Q)y$ berechnen kann.

$\operatorname{sgn}(Q)y$ kann man mit Hilfe des Lanczos-Verfahrens berechnen:
Es ist

$$\begin{aligned}\operatorname{sgn}(Q) &= (Q^2)^{-1/2}Q\\ \operatorname{sgn}(Q)y &= (Q^2)^{-1/2}Qy.\end{aligned}$$

Sei $V_m = (v^1|\dots|v^m)$ Matrix mit den Lanczos-Vektoren für Q^2 mit Start $v^1 = Qy$ und $\|Qy\| = 1$. Dann gilt

$$Q^2 V_m = V_{m+1} T_{m+1,m}.$$

Daraus folgt $V_m^H Q^2 V_m = T_{m,m}$ (tridiagonal).
Idee: Approximiere

$$\begin{aligned}\operatorname{sgn}(Q)y &= (Q^2)^{-1/2}Qy\\ &\approx V_m(T_{m,m}^{-1/2})\underbrace{V_m^H Qy}_{=e^1}\end{aligned}$$

Dabei kann $T_{m,m}^{-1/2}$ für moderates m explizit durch Berechnung der vollständigen Basis aus Eigenvektoren bestimmt werden.

12.3 Satz
Es gilt

$$\left\|\operatorname{sgn}(Q)y - V_m(T_{m,m}^{-1/2})V_m^H Qy\right\|_2 \le \|r^m\|_2$$

mit r^m = Residuum von CG für $Q^2x = y$, Startwert $x^0 = 0$.

Beweis: Man verwendet

1.

$$a^{-1/2} = \frac{2}{\pi}\int_0^\infty (a + t^2)^{-1}\,dt$$

$$\implies (Q^2)^{-1/2} = \frac{2}{\pi} \int_0^\infty (Q^2 + t^2 I)^{-1}\, dt$$

2. Ist r_t^m das Residuum von CG für $(Q^2+t^2I)x = y$, so gilt $\|r_t^m\|_2 \leq \|r_0^m\|_2$. Es gilt sogar: $r_t^m = \psi_{t,m} r_0^m$ mit $|\psi_{t,m}| < 1$.

Der Beweis wird nur skizziert. Wir leiten zuerst die Abschätzung für die Residuen her:

$$r_i^k = p_{i,k}(Q^2 + t^2 I)b \perp K_k(Q^2).$$

Damit gibt es also ein Polynom $p_{k,0}$ mit

$$p_{k,i}(Q^2 + t^2 I)b = \Phi_{k,t} p_{k,0}(Q^2)b.$$

Wenn man das Polynom nun in t betrachtet ergibt dies

$$\begin{aligned} p_{k,t}(s + t^2) &= \Phi_{k,t} p_k(s) \\ p_{k,t}(s) &= \Phi_{k,t} p_k(s - t^2) \end{aligned}$$

und da $p_{k,t}(0) = 1$, ist $\Phi_{k,t} = (p_k(-t^2))^{-1}$. Die Nullstellen der Polynome sind die Ritz-Werte und daher > 0. Wegen $p(0) = 1$ folgt so $p(s) > 1$ für $s < 0$ und damit $|\Phi_{k,t}| < 1$.

Nun ist noch die Abschätzung zu beweisen. Dazu betrachtet man

$$\begin{aligned} &\operatorname{sgn}(Q)b - QV_m T_m^{-\frac{1}{2}} V_m^H b \\ = \;& \frac{2}{\pi} \int_0^\infty Q(Q^2 + t^2 I)^{-1} b - QV_m (t^2 I + T_m)^{-1} V_m^H b \, dt \\ = \;& \frac{2}{\pi} \int_0^\infty Q(Q^2 + t^2 I)^{-1} \underbrace{\left[b - (Q^2 + t^2 I)V_m (t^2 I + T_m)^{-1} V_m^H b\right]}_{=\psi_{t,m} r^m} \, dt. \end{aligned}$$

Weiter folgt wegen $|\psi_{t,m} \leq 1|$, dass alle Eigenwerte der hermiteschen Matrix

$$\frac{2}{\pi} \int_0^\infty \psi t, m Q(Q^2 + t^2 I)^{-1}\, dt$$

den Betrag ≤ 1 haben. Damit ist auch die 2-Norm ≤ 1. $\square$

13 Unvollständige LU-Zerlegungen

Eine unvollständige LU-Zerlegung (ILU) wird benötig als Präkonditionierer. Starten wollen wir mit einer zeilenorientierten Variante der Gauß-Elimination. Dazu sei $A \in \mathbb{C}^{n \times n}$ regulär.

13.1 Algorithmus (*ikj*-Form der Gauß Elimination)

for $i = 2, \ldots, n$ **do**
  for $k = 1, \ldots, i-1$ **do**
    $a_{i,k} = a_{i,k}/a_{k,k}$
    for $j = k+1, \ldots, n$ **do**
      $a_{i,j} = a_{i,j} - a_{i,k}a_{k,j}$
    end for
  end for
end for

Nach Terminierung enthält das strikte untere Dreieck von A den Faktor L; der Rest ist U.

Vorteil der *ikj*-Form: L und U werden zeilenweise berechnet. Dies ist geeignet für die Datenstruktur bei dünnbesetzten Matrizen.

Idee bei ILU: Lasse Operationen in der Gauß-Elimination weg, so dass die Matrizen weniger gefüllt sind.

13.1 ILU zu einem vorgegebenen Muster

13.2 Definition
$E \subseteq \{1, \ldots, n\} \times \{1, \ldots, n\}$ heißt *Muster*, wenn $\{(i,i), i = 1, \ldots, n\} \subseteq E$.

13.3 Algorithmus (*ikj*-Form der ILU zum Muster E)

for $i = 2, \ldots, n$ **do**
  for $k = 1, \ldots, i-1$ **do**
    if $(i,k) \in E$ **then**
      $a_{i,k} = a_{i,k}/a_{k,k}$
      for $j = k+1, \ldots, n$ **do**
        if $(i,j) \in E \wedge (k,j) \in E$ **then**
          $a_{i,j} = a_{i,j} - a_{i,k}a_{k,j}$
        end if
      end for
    end if
  end for
end for

Für U und L in der LU-Faktorisierung gilt dann

$$u_{ij} = \begin{cases} a_{i,j} & \text{falls } (i,j) \in E \text{ und } j \geq i \\ 0 & \text{sonst} \end{cases}$$

und

$$u_{ij} = \begin{cases} a_{i,j} & \text{falls } (i,j) \in E \text{ und } j \leq i \\ 0 & \text{sonst} \end{cases}$$

13.4 Satz
Sei A eine H-Matrix, dann existiert die ILU für jedes Muster E.

Beweis: s. Vorlesung parallele Algorithmen.
Nun stellt sich natürlich die Frage, was geeignete Muster E sind.

13.5 Definition
Man spricht von ILU(0), falls $E = \{(i,j) : a_{ij} \neq 0\}$.

Frage: Wie findet man weitere, größere geeignete Muster?

Idee: Wir gehen davon aus, dass $\mathrm{diag}(A) = I$ und $|a_{i,j}| < \varepsilon$ für $i \neq j$ (z.B. $\varepsilon = \frac{1}{4}$ bei Modellproblem I). Dann kann man einen „Füll-Level-Index“ auf folgende Weise vergeben.

Initialisierung:

$$\widehat{\ell ev}(i,j) = \begin{cases} 1 & \text{falls } a_{i,j} \neq 0 \\ \infty & \text{falls } a_{i,j} = 0 \end{cases}.$$

Bei jedem Schritt der (unvollständigen) Gauß-Elimination wird $\widehat{\ell ev}(i,j)$ aufdatiert via

$$a_{i,j} = a_{i,j} - a_{i,k}a_{k,j} \Rightarrow \widehat{\ell ev}(i,j) = \min\{\widehat{\ell ev}(i,j), \widehat{\ell ev}(i,k) + \widehat{\ell ev}(k,j)\} \tag{13.1}$$

Geht man davon aus, dass alle $a_{i,i}$ die Größenordnung 1 haben, so hat $a_{i,j}$ die Größenordnung $\mathcal{O}(\varepsilon^{\widehat{\ell ev}(i,j)})$. Je größer $\widehat{\ell ev}(i,j)$, desto kleiner $a_{i,j}$. In einer ILU kann man also das Muster so festlegen, dass die Einträge mit großem Füll-Level auf Null gesetzt werden.
In der Standardterminologie wird $\ell ev(i,j) = \widehat{\ell ev}(i,j) - 1$ verwendet, d.h. (13.1) wird zu

$$\ell ev(i,j) = \min\{\ell ev(i,j), \ell ev(i,k) + \ell ev(k,j) + 1\}. \tag{13.2}$$

In der ILU(p) werden nun die Positionen in das Muster E übernommen, für die $\ell ev(i,j) \leq p$. Dies ist konsistent mit der Bezeichnung ILU(0) von vorher.

13.6 Algorithmus (ikj-Form der ILU(p))

for $i = 2, \ldots, n$ **do**
  for $k = 1, \ldots, i-1$ **do**
    if $\ell ev(i,k) \leq p$ **then**
      $a_{i,k} = a_{i,k}/a_{k,k}$
      for $j = k+1, \ldots, n$ **do**
        $\ell ev(i,j) = \min\{\ell ev(i,j), \ell ev(i,k) + \ell ev(k,j) + 1\}$
        if $\ell ev(i,k) \leq p$ **then**
          $a_{i,j} = a_{i,j} - a_{i,k}a_{k,j}$
        end if
      end for
    end if
  end for
end for

1. Eine gute Implementierung sollte so sein, dass der Aufwand zur Berechnung der Level-Indizes nicht groß wird.

2. Das entstehende Muster E ist unabhängig von den numerischen Werten. Es kann also eine Datenstruktur von L und U vorweg berechnet werden.

13.7 Aufgabe
Sei A symmetrisch und $G(A) = (V, E)$ der Graph von A. Zeige:

(i) Bei der normalen Gauß-Elimination hat die Matrix $(L\backslash U)$ den Graphen $\tilde{G} = (V, \tilde{E})$ mit

$$(i,j) \in \tilde{E} \text{ mit } i < j$$
$$\iff \exists \text{ Kantenzug } (i, i_1, \ldots, i_{\ell-1}, j) \text{ mit } i_\nu \in \{1, \ldots i-1\}.$$

(ii) Bei ILU(p) hat die Matrix $(L \setminus U)$ den Graphen $\tilde{G} = (V, \tilde{E})$ mit

$$(i,j) \in \tilde{E} \text{ mit } i < j$$
$$\iff \exists \text{ Kantenzug der Länge } p\ (i, i_1, \ldots, i_{\ell-1}, j)$$
$$\text{mit } i_\nu \in \{1, \ldots i-1\}.$$

Hinweis: Verwende Induktion über k und formuliere eine geeignete Eigenschaft für die Zwischenresultate $A^{(k)}$.

13.2 Dynamische ILU

Nachteile von ILU(p): Muster ist über ein einfaches Modell, aber ohne wirkliche Rücksicht auf numerische Werte festgelegt. Man hätte aber gerne

$$A = LU - R \text{ ,mit } R \text{ „klein“}$$

Idee: Lege Muster bei ILU dynamisch fest, d.h während der Rechnung.

Typischer Vertreter: ILUT („T“:Threshold = Schwelle)

13.8 Definition
Eine *Nullregel* ist eine Vorschrift, bei der Berechnung einer ILU eine Zahl (oder Teile einer ganzen Zeile) auf 0 zu setzen.

Damit ergibt sich folgendes Gerüst für ILU

13.9 Algorithmus (ikj-Form einer allg. ILU)

1: **for** $i = 2, \ldots, n$ **do**
2:   **for** $k = 1, \ldots, i-1$ **do**
3:     $a_{i,k} = a_{i,k}/a_{k,k}$
4:     wende eine Nullregel auf $a_{i,k}$ an
5:     **if** $a_{i,k} \neq 0$ **then**
6:       **for** $j = k+1, \ldots, n$ **do**
7:         $a_{i,j} = a_{i,j} - a_{i,k} a_{k,j}$
8:       **end for**
9:     **end if**
10:   **end for**
11:   wende eine Nullregel auf $a_{i,\bullet}$ an {i-te Zeile}
12: **end for**

13.10 Bemerkung
ILU(p) passt in das Schema, die Nullregeln basieren nur darauf, ob $(i,j) \in E$

13.11 Definition
ILUT(τ,p) verwendet folgende Regeln:

Für Zeile 4:

$$a_{ik} = \begin{cases} 0 & \text{falls } |a_{ij}| \leq \tau \\ \text{unverändert} & \text{sonst} \end{cases}$$

Für Zeile 11:

1. $a_{i,j} \leftarrow 0$, falls$|a_{ij}| \leq \tau$

2. Behalte danach unter den Einträgen $a_{i1}, \dots, a_{i,i-1}$ ($\simeq L$) nur die p betragsgrößten bei, alle anderen werden 0. Dito mit den Einträgen $a_{i,i+1}, \dots, a_{i,n}$ ($\simeq U$) ($\Rightarrow$ max $2p+1$ Einträge der Zeile sind $\neq 0$)

   Variante für 2.
   Es bezeichne $\ell(i), u(i)$ die Anzahl der Nicht-Nullen in Zeile i von A vor bzw. nach der Diagonalen. Behalte dann $\ell(i)+p$ bzw. $u(i)+p$ Einträge bei.

13.12 Bemerkung
ILUT(τ,p) ist nicht immer eine ILU zu einem Muster E, da in Zeile 4 und Zeile 7 unterschiedliche Nullregeln angewendet werden.

Es gibt nur schwache Aussagen über ILUT, typisch ist die folgende.

13.13 Definition
$A \in \mathbb{R}^{n \times n}$ heißt $\hat{M}$-Matrix, falls gilt:

$$\begin{aligned} &a_{ij} \leq 0 \text{ für } i \neq j, \\ &a_{ii} > 0 \text{ für } i = 1, \dots n-1, \\ &\forall i < n \;\; \exists j_i > i \text{ mit } a_{ij_i} < 0. \end{aligned}$$

13.14 Satz
A sei eine diagonal dominante $\hat{M}$-Matrix, d.h.

$$(a_{i,\bullet}) = \sum_{j=1}^{n} a_{ij} \geq 0 \text{ für } i = 1, \dots n.$$

Die Nullregeln in 13.9 seien beliebig bis auf die Tatsache, dass sie nie auf a_{ij_i} angewendet werden, noch auf a_{ii}. Dann ist Algorithmus 13.9 durchführbar und es gilt nach Terminierung des Algorithmus:

$$\begin{aligned} a_{ii} &> 0 \text{ für } i = 1, \dots, n, \\ a_{nn} &\geq 0, \\ \sum_{j=1}^{n} a_{ij} &\geq 0 \text{ für } i = 1, \dots, n. \end{aligned}$$

Beweis: Wir benötigen Bezeichnungen für die Zwischenwerte in Algorithmus 13.9 und formulieren ihn deshalb nochmal.

13.15 Algorithmus (ikj-Form einer allg. ILU mit Zwischenwerten)

```
for i = 2, ..., n do
  for k = 1, ..., i-1 do
    ℓ_{i,k} = a_{i,k}^{(k)} / a_{k,k}^{(k)}
    Nullregel
    for j = k+1, ..., n do
      a_{i,j}^{(k+1)} = a_{i,j}^{(k)} - ℓ_{i,k} a_{k,j}^{(k)}
    end for
  end for
  Nullregel
end for
```

Wir zeigen, dass für alle i und k gilt:

$$\begin{aligned} a_{ij}^{(k)} &\leq 0 \text{ für } i \neq j, \\ a_{ii}^{(k)} &> 0 \text{ für } i < n, \\ \sum_{j=1}^{n} a_{ij}^{(k)} &\geq 0. \end{aligned}$$

Der Beweis ist eine Induktion über i, innerhalb derer eine Induktion über k gemacht wird. Es reicht deshalb, den Schritt von k nach $k+1$ bei festem i anzugeben.
$k = 1$: Hier ist die Behauptung identisch mit den Voraussetzungen des Satzes.
$k \to k+1$:

$$a_{ij}^{(k)} = \begin{cases} 0 & \text{falls Nullregel zutrifft} \\ a_{i,j}^{(k)} - \underbrace{\ell_{i,k}}_{\leq 0} \underbrace{a_{k,j}^{(k)}}_{\leq 0} & \text{sonst, insbesondere} \leq 0\,. \end{cases}$$

Für die Zeilensummen gilt weiter

$$\sum_{j=i}^{n} a_{i,j}^{(k+1)} = \sum_{j=i}^{n} a_{i,j}^{(k)} - \ell_{i,k} a_{k,j}^{(k)} - r_{ij}^{(k)} \overset{(*)}{=} \underbrace{\sum_{j=i}^{n} a_{i,j}^{(k)}}_{\geq 0} - \underbrace{\ell_{i,k}}_{\leq 0} \underbrace{\sum_{j=i}^{n} a_{k,j}^{(k)}}_{\geq 0} \underbrace{-r_{ij}^{(k)}}_{\geq 0} \geq 0$$

Wobei $(*)$ folgt aus

$$r_{ij}^{(k)} = \begin{cases} 0 & \text{falls Nullregel zutrifft} \\ a_{i,j}^{(k)} - \ell_{i,k} a_{k,j}^{(k)} & \text{sonst, insbesondere} \leq 0 \end{cases}.$$

Zu zeigen bleibt $a_{ii}^{(k+1)} > 0$. Wir wissen

$$a_{ii}^{(k+1)} \geq \sum_{j=i+1}^{n} -a_{ij}^{(k+1)} \geq -a_{ij_i}^{(k+1)} = -(a_{ij_i}^{(k)} - \ell_{i,k} a_{kj_i}^{(k)}) \geq a_{ij_i}^{(k)} > 0.$$

13.3 Approximative Inverse

Die bisherigen Ansätze zielten auf

$$A = LU - R \text{ ,mit } R \text{ „klein“}.$$

Dann ist die präkonditionierte Matrix

$$L^{-1}AU^{-1} = I - L^{-1}RU^{-1}.$$

Dabei hoffen wir, dass mit R auch $L^{-1}RU^{-1}$ „klein“ ist. Diese Hoffnung erfüllt sich nicht, falls L und U schlecht konditioniert sind.

Neuer Ansatz: Suche ein $M \in \mathbb{C}^{n\times n}$ so dass

1. M dünn besetzt,
2. MA nahe bei I.

Ein solches M heißt „sparse approximate inverse“ (dünn besetzte approximative Inverse) von A. Wie kann man M berechnen?

Dazu betrachten wir Verfahren, welche

$$\|I - AM\|_F^2$$

(approximativ) minimieren.

13.16 Definition
Für $A \in \mathbb{C}^{n\times n}$ ist $\|A\|_F^2 = \sum_{i=1}^n \sum_{j=1}^n |a_{ij}|^2$.

13.17 Aufgabe
Man zeige $\|Ax\|_2 \leq \|A\|_F \cdot \|x\|_2$.

13.18 Bemerkung
Es ist $\|A\|_F = \|\vec{(A)}\|_2$ mit

$$vec(A) = \begin{pmatrix} a_{\bullet,1} \\ \vdots \\ a_{\bullet,n} \end{pmatrix}$$

Für $A, B \in \mathbb{C}^{n\times n}$ erhalten wir so auch

$$||A||_F^2 = \langle A, A\rangle,$$

wobei $\langle , \rangle$ die Bilinearform auf $\mathbb{C}^{n\times n}$ ist mit

$$\langle A, B\rangle = \langle vec(A), vec(B)\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\overline{b}_{ij} = \sum_{i=1}^{n}(B^H A)_{i,i} = tr(B^H A).$$

Wir betrachten einfache Minimierungsverfahren mit „line search“

1. $MR(\simeq$ GMRES(1)): Minimierung entlang des Residuums $I - AM$.
2. Steepest descent: Minimierung entlang des Gradienten.

GMRES(1): Ein Iterationsschritt (berechnet M^{k+1} aus M^k mit $R^k = I - AM^k$) lautet (s. Lemma 8.8)

$$M^{k+1} = M^k - \frac{\langle R^k, R^k\rangle}{\langle R^k, AR^k\rangle} R^k = M^k - \frac{tr((R^k)^H R^k)}{tr((AR^k)^H R^k)} R^k.$$

Um die dünne Besetztheit zu erhalten, wenden wir wieder Nullregeln an.

13.19 Algorithmus (Globales MR für dünn besetzte approx. Inverse)

wähle M^0 (z.B. $M^0 = I, M^0 = A^H$)
for $k = 0, 1, \dots$ **do**
  $R^k = I - AM^k$
  $\alpha_k = \frac{tr((R^k)^H R^k)}{tr((AR^k)^H R)}$
  $\tilde{M}^{k+1} = M^k - \alpha_k R^k$
  wende Nullregel auf $\tilde{M}^{k+1}$ an, ergibt M^{k+1}
end for

Steepest descent:

Benötigt wird die Ableitung von $F(M) = \|I - AM\|_F^2$, $F : \mathbb{C}^{n\times n} \to \mathbb{C}$, $F' : \mathbb{C}^{n\times n} \to L(\mathbb{C}^{n\times n}, \mathbb{C})$

$$F(M + H) - F(M) = F'(M) \cdot H + o(\|H\|).$$

Wir berechnen:

$$\begin{aligned} F(M + H) - F(M) = & tr((I - A(M + H))^H (I - A(M + H))) \\ & - tr((I - AM)^H (I - AM)) \end{aligned}$$

$$=2\mathrm{Re}\left(tr((I-AM)^H(AH))\right)+\underbrace{tr((AH)^H(AH))}_{O(\|H\|^2)}.$$

Also ist

$$\begin{aligned}F'(M)\cdot H =& 2\mathrm{Re}\left(tr((I-AM)^H(AH))\right)\\ =& 2\mathrm{Re}\left(\langle AH, I-AM\rangle\right)\\ =& 2\mathrm{Re}\left(\langle H, A^H(I-AM)\rangle\right).\end{aligned}$$

Die Steepest-Descent-Richtung G ist damit gegeben

$$G = A^H(I-AM) = A^H R$$

13.20 Algorithmus (Glob. steepest desc. für dünnbes. approx. Inv.)

wähle M^0
for $k = 0, 1, \dots$ **do**
  $R^k = I - AM^k$
  $G^k = A^H R^k$
  $\alpha_k = \frac{tr((AG^k)^H R^k)}{tr((AG^k)^H (AG^k))}$
  $M^{k+1} = M^k + \alpha_k G^k$
  wende Nullregel auf M^{k+1} an
end for

Es ist für $M = (m_1|\dots|m_n)$

$$F(M) = \sum_{i=1}^{n} \|e_i - Am_i\|_2^2.$$

Die Minimierung von F zerfällt also in n unabhängige Teile. Es ergeben sich damit lokale Versionen der Algorithmen 13.19 und 13.20.

13.21 Algorithmus (lokales MR für dünnbesetzte approx. Inv.)

wähle $M^0 = (m_1, \dots, m_n)$
for $k = 0, 1, \dots$ **do**
  for $i = 1, \dots, n$ **do**
    $r_i^k = e_i - Am_i^k$
    $\alpha_k = \frac{\langle r_i^k, Ar_i^k\rangle}{\langle Ar_i^k, Ar_i^k\rangle}$
    $m_i^{k+1} = m_i^k + \alpha_k r_i^k$
    wende Nullregel auf m_i^{k+1} an
  end for
end for

Analog für Algorithmus 13.20.

13.22 Bemerkung

(i) Schleife über i ist parallelisierbar, da die m_i unabhängig voneinander berechnet werden.

(ii) Es sind verbesserte Varianten möglich ('self preconditioning'), bei denen man die bisher berechnete approximative Inverse als Präkonditionierer bei Bestimmung der nächsten approximativen Inverse verwendet.

Wir untersuchen jetzt die Frage, wann eine dünn besetzte approximierte Inverse M überhaupt regulär ist.

13.23 Lemma
Sei $\|I - AM\| < 1$ in irgendeiner Operatornorm. Dann ist M regulär.

Beweis: $AM = I - (I - AM) \Rightarrow (AM)^{-1} = \sum_{i=0}^{\infty}(I - AM)^i$, da diese Reihe wegen $\|I - AM\| < 1$ konvergiert. Also ist mit A auch M regulär. □

13.24 Bemerkung

a) $M = 0$ erreicht $\|I - AM\| = 1$. In diesem Sinne ist $\|I - AM\| < 1$ eine Minimalforderung.

b) Wegen $\|I - AM\|_2 \leq \|I - AM\|_F$ (s. Aufgabe 13.17) ist auch im Falle $\|I - AM\|_F < 1$ die Matrix M regulär.

Für weitere Resultate betrachten wir jetzt die *1-Normen* $\|e_i - Am_i\|_1$.
Sei

$$\|r_i\| = \|e_i - Am_i\| < \tau_i, \quad i = 1, \ldots, n. \tag{13.3}$$

13.25 Satz
Sei (13.3) erfüllt und es sei $B = (b_{ij}) = A^{-1}$. Weiter gelte für ein i und ein j

$$|b_{ij}| \geq \tau_j \max_{\ell=1,\ldots,n} |b_{i\ell}| > 0.$$

Dann ist $m_{ij} \neq 0$.

Beweis: Es ist $AM = I - R, R = (e_1 - Am_1|\ldots|e_n - Am_n)$. Damit erhalten wir $M = B - BR$, d.h. $m_{ij} = b_{ij} - \sum_{k=1}^{n} b_{ik}r_{kj}$. Dies ergibt

$$\begin{aligned} |m_{ij}| &\geq |b_{ij}| - \sum_{k=1}^{n} |b_{ik}| \cdot |r_{kj}| \\ &\geq |b_{ij}| - \max_k |b_{ik}| \cdot \|r_j\|_1 \\ &> |b_{ij}| - \max_k |b_{ik}| \cdot \tau_j > 0 \end{aligned}$$

□

13.26 Definition
$B \in \mathbb{C}^{n \times n}$ heißt τ-*äquimodular* ($\tau > 0$), falls gilt

$$b_{ij} \neq 0 \Rightarrow |b_{ij}| \geq \tau \cdot \max_{\ell,s=1,\ldots,n} |b_{\ell s}|.$$

Interpretation: $b_{ij}, b_{\ell s} \neq 0 \Rightarrow |\frac{b_{\ell s}}{b_{ij}}| \leq \frac{1}{\tau}$

13.27 Korollar
Gilt für eine dünn besetzte approximative Inverse, dass

$$\|e_i - Am_i\|_1 < \tau \quad \text{für } i = 1, \ldots, n$$

und ist B τ-äquimodular, so folgt

$$b_{ij} \neq 0 \Rightarrow m_{ij} \neq 0.$$

Das Besetztheitsmuster von M enthält also das von B. *Gute* dünnbesetzte approximierte Inversen mit geringerem Muster als B^{-1} sind also nur zu erwarten, wenn die Einträge in B^{-1} stark variieren.

13.4 Faktorisierte dünnbesetzte approximative Inverse

13.28 Definition
Eine *faktorisierte dünnbesetzte approximierte Inverse* ist eine Zerlegung der Gestalt

$$LAU = D + R$$

mit U rechte obere, L linke untere Deiecksmatrix, $diag(U) = diag(L) = I$, D Diagonalmatrix und R „klein“.

13.29 Bemerkung
Faktorisierte dünnbesetzte approximierte Inverse sind eine Möglichkeit, Korollar 13.27 „auszutricksen“:

L, U können dünn besetzt sein, auch wenn $UD^{-1}L$ als approximative Inverse dicht (oder so dicht wie A^{-1}) besetzt ist.

Konkrete Berechnung: Betrachte zunächst Berechnung von L, U, D mit

$$LAU = D.$$

Sei

$$A^k \in \mathbb{C}^{k \times k}, A = \begin{pmatrix} A^{k-1} & y^k \\ x^k & \alpha_k \end{pmatrix}, A^n = A.$$

Gilt $L^kA^kU^k = D^k$, dann gilt

$$\begin{aligned}
&\begin{pmatrix} L^k & 0 \\ \ell^{k+1} & 1 \end{pmatrix} \underbrace{\begin{pmatrix} A^k & y^k \\ x^k & \alpha \end{pmatrix}}_{=A^{k+1}} \begin{pmatrix} U^k & u^{k+1} \\ 0 & 1 \end{pmatrix} \\
= &\begin{pmatrix} L^kA^k & L^ky^k \\ \ell^{k+1}A^k + x^k & \ell^{k+1}y^k + \alpha \end{pmatrix} \begin{pmatrix} U^k & u^{k+1} \\ 0 & 1 \end{pmatrix} \\
= &\begin{pmatrix} L^kA^kU^k & L^kA^ku^{k+1} + L^ky^k \\ \ell^{k+1}A^kU^k + x^kU^k & (\ell^{k+1}A^k + x^k)u^{k+1} + \ell^{k+1}y^k + \alpha \end{pmatrix} = D^{k+1}
\end{aligned}$$

mit

$$\begin{aligned}
L^kA^ku^{k+1} = -L^ky^k &\iff D^k(U^k)^{-1}u^{k+1} = -L^ky^k, \\
\ell^{k+1}A^kU^k = -x^kU^k &\iff (A^kU^k)^T(\ell^{k+1})^T = -(U^k)^T(x^k)^T \\
&\iff [(L^k)^{-1}D^k]^T(\ell^{k+1})^T = -(U^k)^T(x^k)^T.
\end{aligned}$$

Man erhält also:

$$\begin{aligned}
u^{k+1} &= -U^k(D^k)^{-1}L^ky^k, \\
\ell^{k+1} &= -(L^k)^T((D^k)^{-1}(U^k)^T(x^k)^T), \\
d^{k+1} &= (\ell^{k+1}A^k + x^k)u^{k+1} + \ell^{k+1}y^k + \alpha.
\end{aligned}$$

13.30 Bemerkung
Es ist

$$\begin{aligned}
A^ku^{k+1} &= -y^k, && (13.4)\\
\ell^{k+1}A^k &= -x^k, && (13.5)\\
d^{k+1} &= (\ell^{k+1}A^k + x^k)u^{k+1} + \ell^{k+1}y^k + \alpha_k && (13.6)\\
&= \ell^{k+1}y^k + \alpha_k \\
&= x^ku^{k+1} + \alpha_k.
\end{aligned}$$

Da wir aber (13.4), (13.5) nur approximativ lösen, sollte d^{k+1} nach (13.6) bestimmt werden.

13.31 Algorithmus (Berechnung der fakt. dünnbes. approx. Inv.)

for $k = 1, 2, \ldots, n$ **do**
  löse $A^k u^{k+1} = -y^k$ approximativ, z.B. mit $(A^k)^{-1} \approx (U^k)(D^k)^{-1}(L^k)$
  löse $(A^k)^T(\ell^{k+1})^T = -(x^k)^T$ approximativ
  Berechne d^{k+1} nach (13.6)
  wende Nullregel auf u^{k+1}, ℓ^{k+1} an
end for

Wann ist $d^{k+1} \neq 0$?

13.32 Lemma
Sei A *hpd*, die Nullregel usw. sei so, dass $U = L^H$, also $(\ell^{k+1})^H = u^{k+1}$. Dann ist $d^{k+1} > 0$ für alle k.

Beweis: Es ist $L^k A^k U^k = D^k = L^k A^k (L^k)^H$ mit $L^k A^k (L^k)^H$ hpd. $\square$

14 Mehrgitterverfahren

Idee: Ausnutzen mehrerer Diskretisierungslevel bei partiellen Differentialgleichungen.

14.1 Beispiel (Mehrgitterverfahren für Modellproblem I)
Gegeben sei also $N,\ h := \frac{1}{N+1}$ und der zugehörige 5-Punkte-Stern. Des Weiteren sei

$$\Omega_h := \{(ih, jh) :\ i, j = 1,\ \dots,\ N\}$$

das zugehörige Gitter (ohne Ränder). Mit A_h bezeichnen wir die zugehörige Laplace-Matrix zur Diskretisierungsweite h.

Bekannt ist:

14.2 Lemma
Die Matrix A_h besitzt die Eigenwerte

$$\lambda_{k,\ell} := 4 - 2\left(\cos\left(\pi \cdot k \cdot h\right) + \cos\left(\pi \cdot \ell \cdot h\right)\right) \quad k,\ \ell = 1,\ \dots,\ N$$

und die zugehörigen Eigenvektoren sind gegeben durch

$$z_{i,j}^{(k,\ell)} := \sin(i \cdot k \cdot \pi \cdot h) \cdot \sin(j \cdot \ell \cdot \pi \cdot h) \quad k,\ \ell = 1,\ \dots,\ N.$$

14.3 Aufgabe
Plotte $z^{(k,\ell)}$ für verschiedene Werte von k, ℓ mit Matlab (`mesh`), insbesondere für

1. $k = 1,\ \ell = 1,$
2. $k = 2,\ \ell = 1,$
3. $k = 2,\ \ell = 2,$
4. $k = \frac{N}{2},\ \ell = \frac{N}{2}$
5. $k = N,\ \ell = N,$

für $N \approx 32,\ 64.$

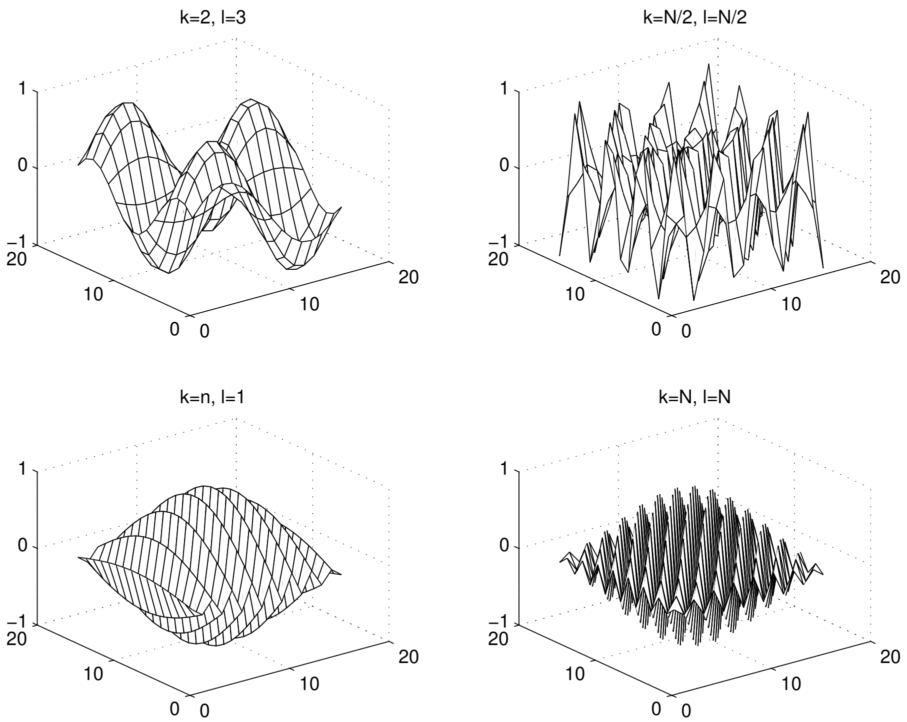
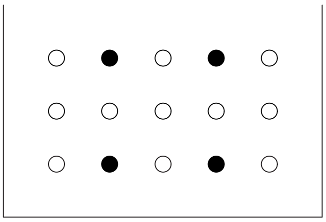


Abbildung 11: Ausgewählte Eigenvektoren für Modellproblem I

14.1 Zweigitter-Verfahren

Die Idee speziell beim Zweigitter-Verfahren ist es jetzt, die Lösung für A_h durch eine Lösung für ein kleineres System $A_{\hat{h}}$ mit größerer Diskretisierungsweite $\hat{h}$ zu approximieren:



mit $N = 2\hat{N} + 1$, also $\hat{h} = 2 \cdot h$.

Wie Aufgabe 14.3 zeigt, gibt es Eigenvektoren, die einer starken Variation unterliegen, und Eigenvektoren, welche eher „glatt" sind. Beim Verkleinern der Punktzahl jedoch können stark variierenden Eigenvektoren nur schlecht approximiert werden. Darum unterteilen wir

$$X := \langle z^{(k,\ell)},\ k,\ \ell = 1,\ \ldots,\ \hat{N}\rangle \quad (\text{„glatte Vektoren"})$$
$$Y := \langle z^{(k,\ell)},\ k > \hat{N} \text{ oder } \ell > \hat{N}\rangle \quad (\text{„oszillierende Vektoren"}).$$

($z^{(k,\ell)}$ Eigenvektoren für A_h.) Es gilt dabei

$$\mathbb{R}^n = X \oplus Y \quad (n = N^2),$$

weil die Eigenvektoren paarweise orthogonal zueinander sind.

Wir entwickeln nun ein Iterationsverfaren (ähnlich dem Schwarz-Verfahren) der folgenden Gestalt:

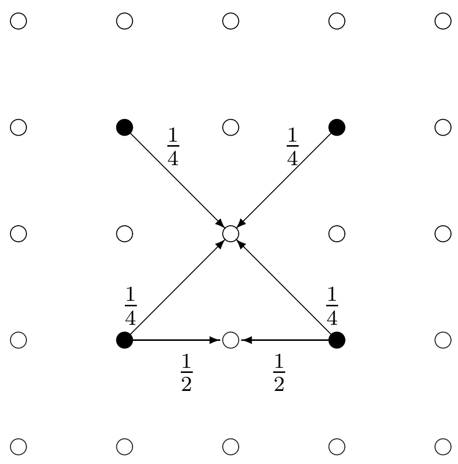
Gesucht ist der Fehler e^k mit $x^* = x^k + e^k$. Statt dessen betrachten wir $r^k = b - A_h x^k$

1. Glätte r^k, d.h. bestimme y^k so dass für $s^k = b - A_h y^k$ gilt
$$s^k \in X.$$
2. Transportiere s^k auf das grobe Gitter und bezeichne das Ergebnis mit $\hat{s}^k$. Löse
$$A_{\hat{h}} \hat{e}^k = \hat{s}^k$$
Transportiere $\hat{e}^k$ auf das feine Gitter in e^k und setze
$$x^{k+1} = y^k + e^k.$$

Anschaulichen entfernen wir aus r^k die stark oszillierenden Komponenten.

14.4 Definition (Transportoperationen)

1. Die *Prolongation* P (Abbildung von $\Omega_{\hat{h}}$ nach Ω_h) ist als gewichtete Mittelung der acht Nachbar-Werte definiert.

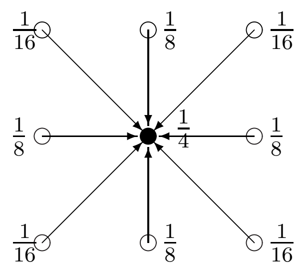


bzw. in Sternform

$$P = \left] \begin{matrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{matrix} \right[\quad .$$

P ist eine $(\hat{n} \times n)$-Matrix, in welcher der Stern über die Zeilen verteilt ist. Dies ist eine bilineare Interpolation.

2. Die Restriktion R als Abbildung von Ω_h nach $\Omega_{\hat{h}}$ ist gegeben durch



bzw. in Sternform

$$R = \frac{1}{4} \cdot \left] \begin{matrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{matrix} \right[\quad .$$

Es gilt also $R = \frac{1}{4} \cdot P^T, \quad R \in \mathbb{R}^{\hat{n} \times n}, \quad P \in \mathbb{R}^{n \times \hat{n}},\ n = N^2,\ \hat{n} = \hat{N}^2.$

14.5 Bemerkung
Die triviale Restriktion

$$R = \left] \begin{matrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{matrix} \right[$$

ist für die Theorie und die Praxis weniger geeignet.

Um die Frage des „Glättens“ zu behandeln, betrachten wir das gedämpfte Gesamschritt-Verfahren

$$y^{\nu+1} = y^{\nu} + \omega \cdot D^{-1} r^{\nu} \quad \omega \text{ Dämpfungsfaktor},$$

wobei $A = D - B,\ D = \operatorname{diag}(A),\ \nu = 0,\ 1,\ \ldots$. Es gilt

$$r^{\nu+1} = ((1-\omega)I + \omega D^{-1} B) r^{\nu}.$$

Setze $J_\omega := (1-\omega)I + \omega D^{-1}B$. Wir wissen: J_ω besitzt die Eigenvektoren $z^{(k,\ell)}$ mit den Eigenwerten

$$\mu_{k,\ell}^{(\omega)} = \frac{\omega}{2}(\cos(\pi \cdot k \cdot h) + \cos(\pi \cdot \ell \cdot h)) + (1-\omega).$$

Für die oszillierenden Eigenvektoren (d.h. $k > \frac{N-1}{2}$ oder $\ell > \frac{N-1}{2}$) gilt:

$$-2 < \cos(\pi \cdot k \cdot h) + \cos(\pi \cdot \ell \cdot h) < 1,$$

also gilt für diese k und ℓ

$$|\mu_{k,\ell}^{(\omega)}| \le \max\left\{ \left|1 - \frac{\omega}{2}\right|,\ |1-2\omega| \right\}, \quad \omega > 0.$$

Die rechte Seite wird minimal für $\omega = \frac{4}{5}$ mit Wert $\frac{3}{5}$.

Wir halten fest

14.6 Lemma
Für $\omega \in \left[0, \frac{4}{5}\right]$ gilt

$$\begin{aligned} \frac{\omega}{2}(\cos(\pi \cdot k \cdot h) + \cos(\pi \cdot \ell \cdot h)) + (1-\omega) &\le \max\left\{ \left|1 - \frac{\omega}{2}\right|,\ |1-2\omega| \right\} \\ &= 1 - \frac{\omega}{2}. \end{aligned}$$

Damit haben wir alle Zutaten für ein Zweigitter-Verfahren beisammen.

14.7 Definition
$y = S_{\nu,\omega}(x)$ bedeute ν Schritte der Glätterungsiteration angewendet auf x.

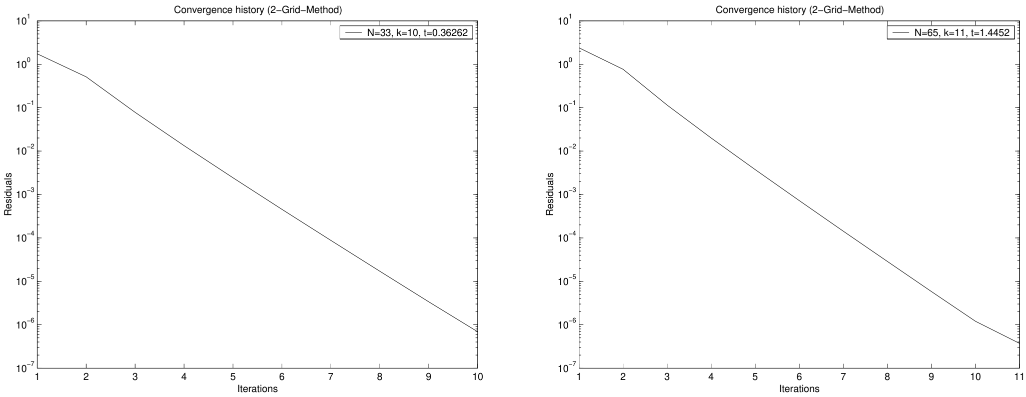


Abbildung 12: Zweigitterverfahrten für Modellproblem 1, $N = 33$ (links), $N = 65$ (rechts)

14.8 Algorithmus (Zweigitter-Verfahren)

wähle x^0, setze $r^0 = b - A_h x^0,\ \omega = \frac{4}{5}$
for $k = 0,\ 1,\ \ldots$ **do**
$y^k = S_{\nu,\omega}(x^k)$
$s^k = b - A_h y^k$ $\{s^k$ ist glatt$\}$
$\hat{s}^k = R s^k$
löse $A_{\hat{h}} \hat{e}^k = \hat{s}^k$
$e^k = P \hat{e}^k$
$x^{k+1} = x^k + e^k$
end for

14.9 Aufgabe
Implementiere das Zweigitter-Verfahren für das Modellproblem I. Verwende direkten Löser für $A_{\hat{h}}$ und messe die Iterationszahlen abhängig von ν und h.

14.2 Mehrgitter-Verfahren

Die Lösung von $A_{\hat{h}} \hat{e}^k = \hat{s}^k$ kann auch rekursiv erfolgen. Man hat dann eine Hierachie:
A_0: zu lösendes System
A_1: kleineres System mit zugehörigen Operatoren P_1 und R_1
$\vdots$
$A_{\ell_{max}}$: kleinstes System mit Operatoren $P_{\ell_{max}}$ und $R_{\ell_{max}}$

Dem entspricht eine Hierachie von Glättern:
$S^0_{\nu,\omega}$

$\vdots$

$S_{\nu,\omega}^{\ell_{max}}$

14.10 Algorithmus (Mehrgitter-Verfahren)

wähle x^0, setze $\ell = 0$
for $k = 0,\ 1,\ \ldots$ **do**
    $x^{k+1} = mgs(x^k, b, \gamma, \ell)$
end for

14.11 Bemerkung

Dabei ist mgs der noch zu definierende Mehrgitter-Schritt.

14.12 Algorithmus (Mehrgitter-Schritt $y = mgs(x, r, \gamma, \ell)$)

1: **if** $\ell = \ell_{max}$ **then**
2:     löse $A_{\ell_{max}} y = r$
3: **else**
4:     $x = S_{\nu,\omega}^{\ell}(x)$ {Vorglättung}
5:     $s = r - A_\ell x$
6:     $\hat{s} = R_\ell s$
7:     $\hat{x} = 0$
8:     **for** $\mu = 1, \ldots, \gamma$ **do**
9:         $\hat{x} = mgs(\hat{x}, \hat{s}, \gamma, \ell + 1)$
10:     **end for**
11:     $e = P\hat{x}$
12:     $y = x + e$
13: **end if**

14.13 Bemerkung

Je größer γ gewählt wird, desto genauer wird die in der Schleife über μ berechnete Approximation $\hat{x}$ für $A_{\ell+1}^{-1}\hat{s}$. Übliche Werte sind $\gamma = 1$ oder $\gamma = 2$. Nach Zeile 11 kann noch ein „Nachglättungsschritt“ $e = S_{\mu,\omega}^{\ell}(e)$ eingeschaltet werden.

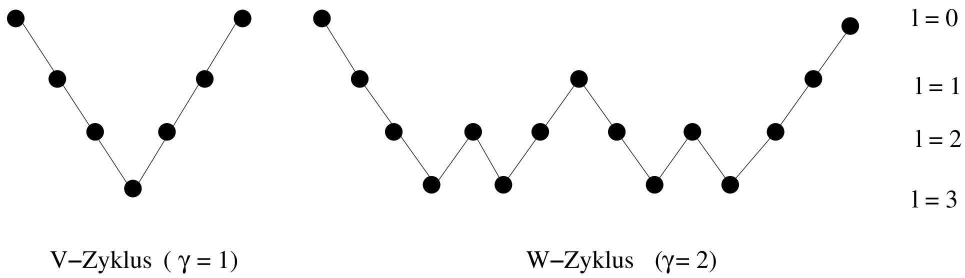


Abbildung 13: V-Zyklus ($\gamma = 1$) und W-Zyklus ($\gamma = 2$) im Mehrgitterverfahren

15 Analyse des Zweigitterverfahrens für Modellproblem I

Iteration des Zweigitterverfahrens:

$$y^k = S_{\nu,\omega}(x^k),$$

$$x^{k+1} = y^k + P\hat{A}^{-1}R(b - Ay^k).$$

Damit gilt für die Fehler $e^k = x^k - x^*$ mit $x^* = A^{-1}b$:

$$e^{k+1} = J_\omega^\nu e^k - P\hat{A}^{-1}RAJ_\omega^\nu e^k$$

mit J_ω Jacobi-Glätter wie früher, also: $J_\omega = (1-\omega)I + \omega D^{-1}B$. Man muss also folgenden Spektralradius untersuchen

$$\rho((I - P\hat{A}^{-1}RA)J_\omega^\nu)$$

Wir verwenden dazu die folgende Faktorisierung:

$$(I - P\hat{A}^{-1}RA)J_\omega^\nu = (A^{-1} - P\hat{A}^{-1}R)(AJ_\omega^\nu)$$

Wir starten mit AJ_ω^ν, $\omega = \frac{4}{5}$. AJ_ω^ν besitzt die Eigenvektoren $z^{(k,\ell)}$ mit den zugehörigen Eigenwerten

$$\mu_\omega^{(k,\ell)} := [4 - 2(\cos(\pi hk) + \cos(\pi h\ell))]\left[\frac{\omega}{2}(\cos(\pi kh) + \cos(\pi \ell h)) + (1-\omega)\right]^\nu$$

Man kann $|\mu_\omega^{(k,\ell)}|$ nach oben abschätzen durch Kurvendiskussion der Funktion $f(t) = (4-2t)\cdot\left[(\frac{\omega}{2}t + (1-\omega)\right]^\nu$ auf dem Intervall $[-2,2]$. Es ist $f(-2) = 8\cdot(1-2\omega)^\nu$, $f(2) = 0$, und f besitzt an der Stelle $(2\omega\nu - 2(1-\omega))/(\omega(1+\nu))$ ein Extremum mit Wert $(4/\omega)\cdot\nu^\nu/(\nu+1)^{\nu+1}$. Damit erhalten wir

15.1 Lemma
Für $\omega = \frac{4}{5}$ und für $\nu = 1, 2$ ist

$$\|AJ_\omega^\nu\|_2 \le 8\cdot\left(\frac{3}{5}\right)^\nu \approx \|A\|_2\cdot\left(\frac{3}{5}\right)^\nu,$$

für $\nu \ge 3$ ist

$$\|AJ_\omega^\nu\|_2 \le 5\cdot\frac{\nu^\nu}{(\nu+1)^{\nu+1}} \approx \|A\|_2\cdot\frac{5}{8}\cdot\frac{\nu^\nu}{(\nu+1)^{\nu+1}}$$

Diese Aussage ist unabhängig von der Gittergröße. Die Schranke wird beliebig klein, wenn man ν groß genug macht.
Jetzt untersuchen wir $A^{-1} - P\hat{A}^{-1}R$. Dafür betrachten wir zuerst $P\hat{A}^{-1}Rz^{(k,\ell)}$.
Es gilt:

$$Rz^{(k,\ell)} = \frac{1}{16} \cdot \left] \begin{matrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{matrix} \right[z^{(k,\ell)}$$

d.h.:

$$\begin{aligned} y_{i,j}^{(k,\ell)} &= \frac{1}{4} z_{2i,2j}^{(k,\ell)} + \frac{1}{8} \cdot \left(z_{2i-1,2j}^{(k,\ell)} + z_{2i+1,2j}^{(k,\ell)} + z_{2i,2j-1}^{(k,\ell)} + z_{2i,2j+1}^{(k,\ell)} \right) \\ &\quad + \frac{1}{16} \left(z_{2i-1,2j-1}^{(k,\ell)} + z_{2i-1,2j+1}^{(k,\ell)} + z_{2i+1,2j-1}^{(k,\ell)} + z_{2i+1,2j+1}^{(k,\ell)} \right) \end{aligned}$$

Jetzt ist unser Ziel

$$(A^{-1} - P\hat{A}^{-1}R)z^{(k,\ell)}$$

in einer noch nicht näher spezifizierten Norm abzuschätzen. Dabei ist

$$z_{i,j}^{(k,\ell)} = \sin(ik\pi h)\sin(i\ell\pi h).$$

Wir betrachten zunächst $P\hat{A}^{-1}Rz^{(k,\ell)}$ näher. Setze

$$\hat{z}^{(k,\ell)} = Rz^{(k,\ell)},$$

so gilt

$$\begin{aligned} \hat{z}_{ij}^{(k,\ell)} &= \sin(2ik\pi h)\sin(2j\ell\pi h) \underbrace{\frac{1}{4}[1+\cos(k\pi h)][1+\cos(\ell\pi h)]}_{=:\tau_{k,\ell}} \\ &= \tau_{k,\ell} \sin(ik\pi\hat{h})\sin(j\ell\pi\hat{h}) \\ &= \tau_{k,\ell}\hat{z}_{ij}^{(k,\ell)} \quad (k, \ell \bmod \hat{n}), \end{aligned}$$

denn

$$\sin((2i-1)k\pi h) + 2\sin(2ik\pi h) + \sin((2i+1)k\pi h) = 2 \cdot \sin(2ik\pi h)[1+\cos(k\pi h)].$$

Mit

$$\begin{aligned} \hat{y}^{(k,\ell)} &= \hat{A}^{-1}\hat{z}^{(k,\ell)} \\ &= \underbrace{\frac{1}{4 - 2[\cos(\pi\hat{h}k) + \cos(\pi\hat{h}\ell)]}}_{=:\sigma_{k,\ell}} \cdot \hat{z}^{(k,\ell)} \end{aligned}$$

ergibt sich für $y^{(k,\ell)} = P\hat{y}^{(k,\ell)}$ komponentenweise

$$\begin{aligned}
y^{(k,\ell)}_{2i,2j} &= \hat{y}^{(k,\ell)}_{i,j},\\
y^{(k,\ell)}_{2i-1,2j} &= \frac{1}{2}\left(\hat{y}^{(k,\ell)}_{i-1,j} + \hat{y}^{(k,\ell)}_{i,j}\right),\\
&\vdots\\
y^{(k,\ell)}_{2i-1,2j-1} &= \frac{1}{4}\left(\hat{y}^{(k,\ell)}_{i-1,j} + \hat{y}^{(k,\ell)}_{i,j} + \hat{y}^{(k,\ell)}_{i,j-1} + \hat{y}^{(k,\ell)}_{i-1,j-1}\right).
\end{aligned}$$

Insgesamt erhalten wir so

$$\begin{aligned}
y^{(k,\ell)}_{2i,2j} &= \sigma_{k,\ell}\tau_{k,\ell}\sin(2ik\pi h)\sin(2j\ell\pi h)\\
y^{(k,\ell)}_{2i-1,2j} &= \sigma_{k,\ell}\tau_{k,\ell}\sin((2i-1)k\pi h)\cos(\ell\pi h)\\
&\vdots\\
y^{(k,\ell)}_{2i-1,2j-1} &= \sigma_{k,\ell}\tau_{k,\ell}\cos(k\pi h)\cos(\ell\pi h)\sin((2j-1)\ell\pi h)\sin((2i-1)k\ell\pi h).
\end{aligned}$$

Wir wissen außerdem

$$A^{-1}z^{(k,\ell)} = \frac{1}{4-[\cos(k\pi h)+\cos(\ell\pi h)]}z^{(k,\ell)}.$$

Damit können wir die Differenz

$$A^{-1}z^{(k,\ell)} - P\hat{A}^{-1}Rz^{(k,\ell)}$$

betrachten. Wir setzen

$$\zeta^{(k,\ell)}_{i,j} := \begin{cases} 1 & i,j \text{ gerade}\\ \cos(k\pi h) & i \text{ ungerade, } j \text{ gerade}\\ \cos(\ell\pi h) & i \text{ gerade, } j \text{ ungerade}\\ \cos(k\pi h)\cos(\ell\pi h) & i \text{ ungerade, } j \text{ ungerade}\end{cases}$$

Es gilt also

$$(A^{-1} - P\hat{A}^{-1}R)z^{(k,\ell)} = \left(\frac{1}{4-2[\cos(k\pi h)+\cos(\ell\pi h)]} - \chi^{(k,\ell)}_{i,j}\right)z^{(k,\ell)} \tag{15.1}$$

mit

$$\chi^{(k,\ell)}_{i,j} = \sigma_{k,\ell}\tau_{k,\ell}\zeta^{(k,\ell)}_{i,j}.$$

Für eine Abschätzung

$$\|A^{-1} - P\hat{A}^{-1}R\|_2 \le C$$

unabhängig von h genügt es, wenn wir für alle Eigenvektoren $z^{(k,\ell)}$ zeigen können, dass

$$\left\|\left(A^{-1} - P\hat{A}^{-1}R\right) z^{(k,\ell}\right\|_2 \leq C\|z^{(k,\ell}\|_2.$$

Der interessante Bereich sind dabei die Eigenvektoren zu *kleinen* Eigenwerten von A (sind die großen von A^{-1}). Wir diskutieren also

$$\begin{aligned}
&\frac{1}{4-2[\cos(k\pi h)+\cos(\ell\pi h)]} - \chi_{i,j}^{(k,\ell)} \\
&\qquad = \frac{1}{4-2[\cos(k\pi h)+\cos(\ell\pi h)]} - \sigma_{k,\ell}\tau_{k,\ell}\zeta_{i,j}^{(k,\ell)} \\
&\qquad = \frac{1}{4-2[\cos(k\pi h)+\cos(\ell\pi h)]} \\
&\qquad\qquad - \frac{\frac{1}{4}[1+\cos(k\pi h)][1+\cos(\ell\pi h)]}{4-2[\cos(k\pi h)+\cos(\ell\pi h)]}\zeta_{i,j}^{(k,\ell)}
\end{aligned}$$

Für k, ℓ klein ist $k\pi h \ll 1$ und $\ell\pi h \ll 1$ und damit

$$\begin{aligned}
4-2[\cos(k\pi h)+\cos(\ell\pi h)] &= (k\pi h)^2 + (\ell\pi h)^2 + \mathcal{O}(h^4) \\
[1+\cos(k\pi h)][1+\cos(\ell k\pi h)] &= 4 - (k\pi h)^2 - (\ell\pi h)^2 + \mathcal{O}h^4)
\end{aligned}$$

Damit ergibt sich

$$\begin{aligned}
&\frac{1}{4-2[\cos(k\pi h)+\cos(\ell\pi h)]} - \frac{\frac{1}{4}[1+\cos(k\pi h)][1+\cos(\ell\pi h)]}{4-2[\cos(k\pi h)+\cos(\ell\pi h)]}\zeta_{i,j}^{(k,\ell)} \\
&\qquad = \left(\frac{1}{(k\pi h)^2+(\ell\pi h)^2+\mathcal{O}(h^4)}\right. \\
&\qquad\qquad \left. - \frac{1-\frac{1}{4}(k\pi h)^2 - \frac{1}{4}(\ell\pi h)^2 + \mathcal{O}(h^4)}{(k\pi h)^2+(\ell\pi h)^2+\mathcal{O}(h^4)}\zeta_{i,j}^{(k,\ell)}\right) \\
&\qquad = \frac{1}{4} + \mathcal{O}(h^2).
\end{aligned}$$

Wegen $\|A\| \approx 8/h^2$ haben wir motiviert (wenn auch nicht streng gezeigt):

15.2 Lemma
Es ist

$$\|A^{-1} - P\hat{A}^{-1}R\|_2 \leq \frac{C}{\|A\|_2}$$

mit einer von der Gitterweite h unabhängigen, kleinen Konstanten C.

Setzen wir die Ergebnisse von Lemma 15.1 und 15.2 zusammen zusammen, erhalten wir

15.3 Satz
Für das Modellproblem I gilt unabhängig von der Gittergröße die folgende Abschätzung für den Iterationsoperator des Zweigitterverfahrens ($\omega = 4/5$):

$$\|(I - P\hat{A}^{-1}RA)S_{\nu,\omega}\|_2 \leq C \cdot \max\left\{\left(\frac{3}{5}\right)^\nu, \frac{5}{8} \cdot \frac{\nu^\nu}{(\nu+1)^{\nu+1}}\right\}.$$

Genügend großes ν erzeugt also Konvergenz. (Tatsächlich genügt $\nu = 1$, aber das sieht man hier nicht.)
Zur Vorbereitung eine Zweigitter-Analyse für allgemeinere Probleme behandeln wir noch eine alternative Betrachtungsweise für die Glättung. Dazu nehmen wir statt des Jacobi-Verfahrens das (relaxierte) Richardson-Verfahren 1. Ordnung.

15.4 Lemma
Sei $X \in \mathbb{C}^{n\times n}$ hpd und $0 \preceq X \preceq I$. Sei $\omega > 0$ fest und betrachte in Abhängigkeit von ν

$$\|X(I - X)^\nu)\| = \eta(\nu).$$

Dann gilt

$$\eta(\nu) = \frac{\nu^\nu}{(\nu+1)^{\nu+1}}.$$

Beweis: Wegen $0 \preceq X \preceq I$, ist $\operatorname{spek}(X) \subset [0,1]$. Da $X(1-X)^\nu$ hermitesch ist gilt

$$\|X - (1 - X)^\nu\|_2 = \max_{\lambda\in\operatorname{spek}(X)} |\lambda(1-\lambda)^\nu| = \eta(\nu).$$

Eindimensionale Kurvendiskussion für $f(\lambda) = \lambda(1-\lambda)^\nu$ liefert

$$f'(\lambda) = (1-\lambda)^\nu - \nu \cdot \lambda(1-\lambda)^{\nu-1}.$$

Also ist

$$\eta(\nu) = \max\left\{|f(0)|, |f(1)|, f\left(\frac{1}{\nu+1}\right)\right\}$$

mit $f(0) = f(1) = 0$ und $f\left(\frac{1}{\nu+1}\right) = \frac{\nu^\nu}{(\nu+1)^{\nu+1}}$. □

15.5 Bemerkung
Mit der Stirlingschen Formel ergibt sich

$$\eta(\nu) = \frac{1}{\mathrm{e}\cdot\nu} + \mathcal{O}(\nu^{-2}).$$

Anwendung: Glättung durch Richardson-Verfahren

$$A = I - (I - A)$$

Richardson-Verfahren (relaxiert)

$$x^{k+1} = (I - \omega A)x^k + \omega b$$

ν Schritte entsprechen einer Matrix

$$(I - \omega A)^\nu =: S_{\omega,\nu} \qquad \text{Glätter.}$$

Verwende $X = \omega A$ und untersuche

$$\left\|AS_{\omega,\nu}\right\| = \left\|\frac{1}{\omega}X(I - X)^\nu\right\| = \frac{1}{\omega}\left\|X(I - X)^\nu\right\|.$$

Wann gilt

$$0 \preceq X \preceq I \iff 0 \preceq \omega A \preceq I$$

15.6 Lemma

1. Sei A hpd. Dann gilt $0 \preceq \omega A \preceq I$ für $\omega \in (0, \frac{1}{\rho(A)}]$ und dann
$$\|AS_{\omega,\nu}\| \leq \frac{1}{\omega}\frac{\nu^\nu}{(\nu + 1)^{\nu+1}}$$
2. Sei A speziell die Matrix des Modellproblems I, $\tilde{A} = \frac{1}{4}A$ (so dass $\operatorname{diag}(\tilde{A}) = I$, $\operatorname{spek}(\tilde{A}) \subseteq (0, 2)$). Dann gilt $0 \preceq \omega\tilde{A} \preceq I$ für $\omega \in (0, \frac{1}{2}]$ und damit
$$\left\|\tilde{A}S_{\omega,\nu}\right\| \leq \frac{1}{\omega}\frac{\nu^\nu}{(\nu + 1)^{\nu+1}}$$

15.7 Bemerkung

Relaxierter Richardson für $\tilde{A}$ bei Modellproblem I ist identisch mit relaxiertem Jacobi für A und es ist $\left\|\tilde{A}S_{\omega,\nu}\right\| = \frac{1}{4}\|AS_{\omega,\nu}\|$

15.8 Korollar

Für das Zweigitterverfahren beim Modellproblem I mit $\omega = 1/2$ gilt die Abschätzung

$$\left\|(I - P\hat{A}^{-1}RA)S_{\omega,\nu}\right\|_2 \leq 2 \cdot C\frac{\nu^\nu}{(\nu + 1)^{\nu+1}}$$

mit C aus Lemma 15.2.

Beweis:
$$e^{k+1} = \big((I - P\hat{A}^{-1}RA)S_{\omega,\nu}\big)e^k$$
mit
$$(I - P\hat{A}^{-1}RA)S_{\omega,\nu} = (A^{-1} - P\hat{A}^{-1}R)AS_{\omega,\nu}$$
und damit
$$\big\|(I - P\hat{A}^{-1}RA)S_{\omega,\nu}\big\|_2 = \big\|(A^{-1} - P\hat{A}^{-1}R)\big\|_2\big\|AS_{\omega,\nu}\big\|_2.$$
Wir wissen für $\omega = \frac{1}{2}$, relaxierter Jacobi-Glätter
$$\begin{aligned} \frac{1}{4}\|AS_{\omega,\nu}\|_2 &\leq 2\frac{\nu^\nu}{(\nu+1)^{\nu+1}} \\ \iff \quad \|AS_{\omega,\nu}\|_2 &\leq 8\frac{\nu^\nu}{(\nu+1)^{\nu+1}} \\ &\leq C \cdot \|A\|_2\frac{\nu^\nu}{(\nu+1)^{\nu+1}} \end{aligned}$$
mit $\|A\|_2 = \big(4 + 4\cos(\pi h)\big) \approx 8$, d.h. $1 \leq c \leq 2$.
Mit Lemma 15.2 gilt
$$\big\|A^{-1} - P\hat{A}^{-1}R\big\|_2 \leq \frac{C}{\|A\|_2}.$$
Damit gilt
$$\big\|(I - P\hat{A}^{-1}RA)S_{\omega,\nu}\big\|_2 \leq 2C \cdot \frac{\nu^\nu}{(\nu+1)^{\nu+1}}.$$
□

Damit haben wir ein Analogon zu Satz 15.3, diesmal für $\omega = 1/2$.

15.9 Korollar
Das Zweigitterverfahren konvergiert mit $\omega = 1/2$, falls ν groß genug ist.

Tatsächlich genügt $\nu = 1$, *unabhängig von* h! Dies kann man auch auf *Mehrgitter* übertragen.

16 Analyse von Zweigitterverfahren allgemein

Gegeben seien $Ax = b$ („feines Gitter“), $\hat{A}$ (grobes Gitter), P, R (Prolongation, Restriktion) und $S_{\omega,\nu}$ (Glätter).
Iterationsoperator des Zweigitterverfahrens: $(I - P\hat{A}^{-1}RA)S_{\omega,\nu}$.

16.1 Definition
$S_{\omega,\nu}$ besitzt die *Glättungseigenschaft*, falls gilt:

$$\|AS_{\omega,\nu}\|_2 \le \|A\|_2 \eta(\nu) \qquad \text{mit } \eta(\nu) \xrightarrow{\nu\to\infty} 0$$

16.2 Definition
P, R, $\hat{A}$ besitzen die *Approximationseigenschaft*, falls gilt:

$$\left\|A^{-1} - P\hat{A}^{-1}R\right\|_2 \le C\frac{1}{\|A\|_2}$$

16.3 Satz
Sind Glättungs- und Approximationseigenschaft erfüllt, so konvergiert das Zweigitterverfahren, wenn ν groß genug ($\eta(\nu)C < 1$).

16.4 Bemerkung
Die Approximationseigenschaft kann man i.A. nur dadurch nachweisen, dass man ausnutzt, dass A eine endlichdimensionale Approximation eines kontinuierlichen Differential-Operators ist, z.B. bei Finiten Elementen (Wir machen das hier nicht).

Für die Glättungseigenschaft haben wir bereits Lemma 15.6. Ein weiteres Resultat ist:

16.5 Lemma
Sei $A = M - N$. Es gelte in einer Norm $\|\cdot\|$

$$\left\|I - 2M^{-1}A\right\| \le 1 \qquad \text{und} \qquad \|M\| \le C\|A\|.$$

Dann gilt mit dem Glätter $S = M^{-1}N = I - M^{-1}A$

$$\left\|AS^\nu\right\| \le \sqrt{\frac{2}{\pi\nu}}\,\left\|A\right\|$$

Beweis: Wir zeigen zunächst: Für B mit $\|B\| < 1$ gilt

$$\left\|(I - B)(I + B)^\nu\right\| \le 2\binom{\nu}{\lfloor \nu/2 \rfloor} \le 2^{\nu-1}\sqrt{\frac{2}{\pi\nu}}$$

Die zweite Ungleichung folgt mit der Stirlingschen Formel. Wir verwenden

$$
\begin{aligned}
(I-B)(I+B)^\nu &= (I-B)\sum_{mu=0}^{nu}\binom{\nu}{\mu}B^\mu \\
&= I + \sum_{\mu=1}^{\nu}\left(\binom{\nu}{\mu}-\binom{\nu}{\mu-1}\right)B^\mu - B^{\nu+1}
\end{aligned}
$$

Mit $\binom{\nu}{\mu}=\binom{\nu}{\nu-\mu}$ und $\binom{\nu}{\mu}\geq\binom{\nu}{\mu-1}$ für $\mu=1,\ldots,\lfloor\nu/2\rfloor$ ergibt sich

$$
\begin{aligned}
\left\|(I-B)(I+B)^\nu\right\| &\leq \underbrace{\left\|I-B^{\nu+1}\right\|}_{\leq 2}+2\sum_{\mu=1}^{\lfloor\nu/2\rfloor}\left(\binom{\nu}{\mu}-\binom{\nu}{\mu-1}\right) \\
&\leq 2+2\left(\binom{\nu}{\lfloor\nu/2\rfloor}-\binom{\nu}{0}\right) \\
&= 2\binom{\nu}{\lfloor\nu/2\rfloor}
\end{aligned}
$$

Zum Beweis des Lemmas: Es ist

$$
M^{-1}AS^\nu = \frac{1}{2^{\nu+1}}(I-B)(I+B)^\nu \qquad \text{mit } B = I-2M^{-1}A
$$

und damit

$$
\|AS^\nu\| \leq \|M\|\cdot\left\|M^{-1}AS^\nu\right\| \leq \|M\|\frac{1}{2^{\nu+1}}2^{\nu+1}\sqrt{\frac{2}{\pi\nu}}
$$

$\square$

A Lösungs-Versuche

A.1 Aufgabe
Finde ausgehend von Satz 2.3 die Lösung von

$$\min_{p_m \in \overline{\Pi}_m,} \max_{\lambda \in D(\gamma,\rho)} |p_m(\lambda)|.$$

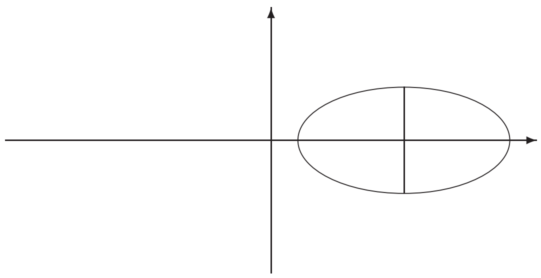
Antwort: $p_m(z) = \frac{(z+\gamma)^m}{\gamma^m}$, wie man leicht mittels linearer Transformation $z \mapsto z - \gamma$ aus Satz 2.3 einsieht.

A.2 Aufgabe
Bestimme jeweils die (im Sinne von Satz 3.7) asymptotisch optimale Lösung der MinMax-Aufgabe

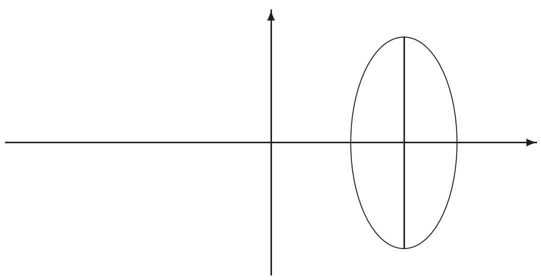
$$\min_{p_m \in \overline{\Pi}_m} \max_{\lambda \in E} |p(\lambda)|$$

und den zugehörigen Wert des Minimums für die Fälle:

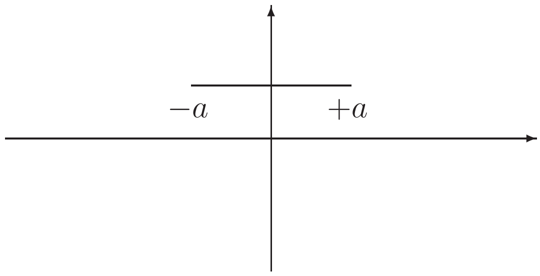
1. $E = a + bE_\rho$, wobei $a, b > 0$ und $a - b \cdot \frac{1}{2}(\rho + 1/\rho) > 0$



2. $E = a + be^{i\frac{\pi}{2}} E_\rho$, wobei $a, b > 0$ und $a - b \cdot \frac{1}{2}(\rho - 1/\rho) > 0$



3. E_ρ ist Geradenstück parallel zur reellen Achse von $-a$ bis $+a$ mit Achsenabschnitt b auf der imaginären Achse



Sind in diesem Fall die Polynom auch optimal für endliches m?

Antwort:

1. und 2.: wurden mit den Tschebyscheff-Polynomen behandelt.

3.:

A.3 Aufgabe
Sei $A \neq A^H$, aber mit einer der folgenden speziellen Eigenschaften:

1. $A^H = \overline{A} \in \mathbb{C}^{n \times n}$ (z.B. bei der Diskretisierung der Maxwell-Gleichungen in der Elektrodynamik)
2. $A^H J = J^H A,\ J \in \mathbb{C}^{n \times n},\ J = J^H$ (man nennt A dann auch J-*hermitesch*)
3. $A^T J = J^T A,\ J \in \mathbb{C}^{n \times n}$ (man nennt A dann auch J-*symmetrisch*)

Zeige jeweils: Für geeignete Wahl von $\tilde{r}^0$ (bzw. w^0) kann man den unsymmetrischen Lanczos-Prozess mit nur einer MVM mit A (und eventuell einer zusätzlichen mit J) durchführen (dies überträgt sich dann auch auf BiCG).

Beweis:

1.
$$AV^m = V^{m+1} T_{m+1,m} \quad \text{und} \quad A^H W^m = W^{m+1} \overline{T}_{m+1,m}$$
$\Longrightarrow A\overline{W}^m = \overline{W}^{m+1} T_{m+1,m}$. Wähle also $w^0 = \overline{v}^0$, dann gilt: $w^m = \overline{v}^m \forall m$

2.
$$AV^m = V^{m+1} T_{m+1,m} \quad \text{und} \quad A^H W^m = W^{m+1} \overline{T}_{m+1,m}$$
Ansatz: $w^i = Jv^i$, dann gilt:
$$A^H W_m = A^H J V_m = JAV_m = \underbrace{JV_{m+1}}_{=W_{m+1}} T_{m+1,m}.$$

Es muß also gelten das $T_{m+1,m} \in \mathbb{R}^{(m+1)\times m}$

$$\alpha_m = \langle Av^m, w^m \rangle = \langle Av^m, Jv^m \rangle = \langle J^H A v^m, v^m \rangle \in \mathbb{R}$$

denn: $(J^H A)^H = A^H J = JA = J^H A$

$$\delta_m = \langle \overline{v}^m, \overline{w}^m \rangle = \langle \overline{v}^m, J\overline{v}^m \rangle \in \mathbb{R}$$

da $J = J^H$

3. analog zu 2 mit dem Ansatz: $w^i = \overline{Jv^i}$

$\square$