Numerische Lineare Algebra

Jochen Werner

Sommersemester 1998, 2001

Inhaltsverzeichnis

Kapitel 1

Einführung

1.1 Grundlegende Probleme

Als grundlegende Probleme der numerischen linearen Algebra bezeichnen wir die folgenden Aufgaben:

- *Lineare Gleichungssysteme*: Bestimme einen n-Vektor x mit $Ax = b$, wobei A eine (reelle oder komplexe) nichtsinguläre $n \times n$-Matrix und b ein n-Vektor ist.

- *Lineare Ausgleichsprobleme*: Bestimme einen n-Vektor x, der $\|Ax-b\|_2$ minimiert. Hierbei sind die $m \times n$-Matrix A und der m-Vektor b gegeben, ferner ist die euklidische Norm $\|\cdot\|_2$ durch $\|y\|_2 := \sqrt{\sum_{i=1}^{m} |y_i|^2}$ gegeben. Gewöhnlich ist hier $m \geq n$, so dass man das lineare Ausgleichsproblem auch als die Aufgabe auffassen kann, das (i. Allg. überbestimmte) lineare Gleichungssystem $Ax = b$ nach der Methode der kleinsten Quadrate zu lösen.

- *Eigenwertprobleme*: Gegeben sei eine $n \times n$-Matrix A. Man bestimme einen Skalar λ und einen zugehörigen nichtverschwindenden n-Vektor x mit $Ax = \lambda x$. Speziell werden wir auf *Singulärwertprobleme* eingehen. Hier ist eine $m \times n$-Matrix A gegeben, zu bestimmen ist ein Skalar λ und ein zugehöriger nichtverschwindender n-Vektor x mit $A^T Ax = \lambda x$. Diese spezielle Klasse von (symmetrischen) Eigenwertproblemen verdient es, wie wir sehen werden, gesondert behandelt zu werden.

Viele Varianten zu diesen Aufgaben sind denkbar. Diese beziehen sich vor allem darauf, dass die jeweils auftretende Koeffizientenmatrix eine spezielle Struktur besitzt. Z. B. kann die Symmetrie von A vorteilhaft bei der Lösung zugehöriger linearer Gleichungssysteme oder Eigenwertaufgaben ausgenutzt werden. Ist bei einem linearen Ausgleichsproblem der Rang der Koeffizientenmatrix $A \in \mathbb{R}^{m \times n}$ maximal, also Rang $(A) = n \leq m$, so besitzt die zugehörige Aufgabe genau eine Lösung. Andernfalls (man spricht dann von rang-defizienten Problemen) gibt es unendlich viele Lösungen, von denen genau eine als eine mit minimaler euklidischer Norm ausgezeichnet ist. Es ist klar, dass a priori Informationen dieser Art auf die Wahl eines entsprechenden Verfahrens Einfluss haben müssen. Weitere Varianten ergeben sich dadurch, dass bei einem linearen Gleichungssystem die Dimension der auftretenden Koeffizientenmatrix A groß sein kann. Ist dies

der Fall, so ist A i. Allg. dünn besetzt (engl.: sparse), hat also nur "wenige" von Null verschiedene Einträge. Als Verfahren kommen vor allem Iterationsverfahren in Frage, bei denen ausgenutzt werden kann, dass die Multiplikation von A mit einem Vektor "einfach" ist. Wir werden daher zwischen direkten und iterativen Verfahren bei linearen Gleichungssystemen unterscheiden. Bei einem *direkten Verfahren* zur Lösung eines linearen Gleichungssystems erhält man, wenn man von den (allerdings unausweichlichen) Rundungsfehlern absieht, nach endlich vielen Schritten (in denen jeweils nur elementare arithmetische Operationen durchgeführt werden) die exakte Lösung. Dagegen wird bei einem iterativen Verfahren eine Folge von (hoffentlich immer besseren) Näherungslösungen $x_0, x_1, \ldots$ berechnet, wobei man aus dem Algorithmus aussteigt, wenn man durch einen geeigneten Test feststellt, dass die zuletzt berechnete Näherungslösung "gut genug" ist.

Die numerische Behandlung der oben angegebenen grundlegenden Probleme der numerischen linearen Algebra werden Inhalt dieser Vorlesung sein. Eine gewisse Schwierigkeit besteht darin, dass die Hörerinnen und Hörer Vorkenntnisse aus der numerischen Mathematik haben sollten, andererseits diese unterschiedlich sein werden. Wir werden daher, wenn immer nötig, über das absolute Grundwissen hinausgehende benötigte Begriffe und Aussagen aus der numerischen Mathematik bereitstellen.

Die oben angegebenen "Grundprobleme" der numerischen linearen Algebra treten häufig als Hilfsprobleme in anderen Bereichen der numerischen Mathematik auf, häufig in einer "inneren Schleife". Dies liegt daran, dass nichtlineare Aufgaben i. Allg. durch sukzessives Linearisieren gelöst werden. Man denke nur an das Newton-Verfahren, das die Lösung eines nichtlinearen Gleichungssystems auf die Lösung einer Folge linearer Gleichungssysteme zurückführt oder an das Gauß-Newton-Verfahren, bei welchem zur Lösung eines nichtlinearen Ausgleichsproblems eine Folge von linearen Problemen zu lösen ist. Ähnlich muss man bei sogenannten Innere-Punkt-Verfahren in der Optimierung in jedem Schritt ein lineares Gleichungssystem mit einer Koeffizientenmatrix der Form

$$W := \begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix}$$

lösen, wobei $A \in \mathbb{R}^{m\times n}$ mit $\text{Rang}\,(A) = m$ und $Q \in \mathbb{R}^{n\times n}$ symmetrisch und auf $\text{Kern}\,(A)$ positiv definit ist[1].

Die Vorlesung wird aus den folgenden Teilen bestehen:

- Direkte Verfahren bei linearen Gleichungssystemen.
- Lineare Ausgleichsprobleme.
- Eigenwertaufgaben.
- Iterationsverfahren bei linearen Gleichungssystemen.

An Vorkenntnissen wird von einem Hörer erwartet, dass sie zumindestens eine Vorlesung "Numerische Mathematik I" gehört hat und bereit ist, eventuell nicht vorhandene

[1] Gleichungssysteme mit einer solchen Koeffizientenmatrix sind nichtsingulär (Beweis?) und treten z. B. auch bei der Interpolation mit radialen Basisfunktionen auf.

Vorkenntnisse aus "Numerische Mathematik II" (relevant ist hier nur der Teil über Eigenwertaufgaben) aufzuarbeiten.

1.2 Literatur

In letzter Zeit sind vorzügliche Bücher über numerische lineare Algebra erschienen. Wir nennen nur einige.

- G. H. Golub and C. F. Van Loan (1996) *Matrix Computations. 3rd Edition.*[2] John Hopkins University Press, Baltimore.
- J. W. Demmel (1997) *Applied Numerical Linear Algebra.* SIAM, Philadelphia.
- Å. Björk (1996) *Numerical Methods for Least Squares Problems.* SIAM, Philadelphia.
- N. J. Higham (1996) *Accuracy and Stability of Numerical Algorithms.* SIAM, Philadelphia.
- L. N. Trefethen and D. Bau, III (1997) *Numerical Linear Algebra.* SIAM, Philadelphia.
- Y. Saad (1996) *Iterative Methods for Sparse Linear Systems.* PWS Publishing Company, Boston.
- A. Greenbaum (1997) *Iterative Methods for Solving Linear Systems.* SIAM, Philadelphia.
- P. C. Hansen (1998) *Rank-Deficient and Discrete Ill-Posed Problems. Numerical Aspects of Linear Inversion.* SIAM, Philadelphia.
- G. W. Stewart (1998) *Afternotes goes to Graduate School.* SIAM, Philadelphia.

An (nicht ganz neuer und daher zumindest drucktechnisch veralteter) deutscher Literatur wollen wir nur nennen:

- W. Bunse und A. Bunse-Gerstner (1985) *Numerische lineare Algebra.* Teubner, Stuttgart.
- W. Hackbusch (1993) *Iterative Lösung großer schwachbesetzter Gleichungssysteme.* Teubner, Stuttgart.

Teile der numerischen linearen Algebra erscheinen natürlich in jedem Lehrbuch der numerischen Mathematik. Aus naheliegenden Gründen seien hier nur genannt:

- J. Werner (1992a) *Numerische Mathematik 1. Lineare und nichtlineare Gleichungssysteme, Interpolation, numerische Integration.* Vieweg, Braunschweig-Wiesbaden.

[2] Wir benutzen allerdings nur die ersten beiden Auflagen.

- J. Werner (1992b) *Numerische Mathematik 2. Eigenwertaufgaben, lineare Optimierungsaufgaben, unrestringierte Optimierungsaufgaben.* Vieweg, Braunschweig-Wiesbaden.

Was die Auswahl des Stoffes angeht halten wir uns etwa an das Buch von J. W. Demmel (1997), sind aber auch von dem ausgezeichneten Buch von N. J. Higham (1996) beeinflußt, insbesondere in den nächsten beiden Kapiteln. Dem Kapitel über Iterationsverfahren bei linearen Gleichungssystemen liegt vor allem das Buch von Y. Saad (1996) zugrunde.

1.3 Notationen

Bei den Notationen sollte man sich nach der Mehrheit der Autoren auf dem entsprechenden Gebiet richten und Bezeichnungen vermeiden, die allzu exotisch[3] sind. So werden Matrizen i. Allg. mit großen, Vektoren mit kleinen lateinischen Buchstaben bezeichnet, kleine griechische Buchstaben bezeichnen Skalare. Ein Vektor wird immer als Spaltenvektor aufgefasst. Ist A eine $m \times n$-Matrix, so bezeichnet A^T die transponierte und (dies ist nur relevant, wenn A komplex ist) A^H die konjugierte (man gehe in jedem Eintrag zum konjugiert komplexen über) und transponierte Matrix. Entsprechende Bezeichnungen werden für Vektoren benutzt. Ferner bezeichnet a_{ij} das Element in der i-ten Zeile und j-ten Spalte von A (um dies deutlich zu machen, schreiben wir auch $(A)_{ij} = a_{ij}$), entsprechend ist x_j die j-te Komponente des Vektors x. Ist A eine $m \times n$-Matrix, so bezeichnet $|A|$ die (reelle) $m \times n$-Matrix, deren Einträge gerade die Absolutbeträge der Einträge von A sind (also: $(|A|)_{ij} := |a_{ij}|$). Entsprechende Bezeichnungen werden für Vektoren benutzt. Mit $\mathbb{R}$ bezeichnen wir die Menge der reellen Zahlen, $\mathbb{R}^n$ ist die Menge der reellen n-Vektoren, entsprechend $\mathbb{R}^{m\times n}$ die Menge der reellen $m \times n$-Matrizen. Im komplexen Fall sind $\mathbb{C}$, $\mathbb{C}^n$ und $\mathbb{C}^{m\times n}$ entsprechend zu verstehen. Die Einheitsmatrix wird mit I bezeichnet, zur Verdeutlichung der Dimension n schreiben wir gelegentlich auch I_n für die $n\times n$-Einheitsmatrix. Die $\leq$-Relation zwischen Vektoren (oder auch Matrizen) ist immer komponentenweise (oder koeffizientenweise) zu verstehen. So bedeutet z. B. $A \geq 0$, dass A eine Matrix mit nichtnegativen Einträgen ist. Der i-te Einheitsvektor (nur die i-te Komponente ist gleich 1, alle anderen verschwinden) ist e_i. Mit e wird schließlich der Vektor bezeichnet, dessen Komponenten alle gleich eins sind. Ohne Kommentar wird das Kronecker-Symbol δ_{ij} benutzt, wobei $\delta_{ij} := 0$ für $i \neq j$ und $\delta_{ij} := 1$ für $i = j$.

[3] Natürlich ist eigentlich nichts einzuwenden gegen einen Satz wie: "Für alle hinreichend großen natürlichen Zahlen ϵ gilt ...." Trotzdem sollte man so etwas vermeiden.

Kapitel 2

Direkte Verfahren bei linearen Gleichungssystemen

2.1 Einige Grundlagen

Wie schon in der Einleitung angegeben, werden wir einige Kenntnisse aus der numerischen Mathematik voraussetzen. Z. B. werden wir nicht noch einmal definieren, was eine (Vektor-) *Norm* $\|\cdot\|$ auf $\mathbb{R}^n$ (wir beschränken uns auf den reellen Fall, weil es praktisch keine Unterschiede zum komplexen gibt) ist, ebenso wird der Begriff eines *inneren Produktes* $\langle\cdot,\cdot\rangle$ auf $\mathbb{R}^n \times \mathbb{R}^n$ als bekannt vorausgesetzt. Wie gewohnt, bezeichne $\|\cdot\|_p$ für $1 \le p \le \infty$ (interessant sind vor allem die Fälle $p = 1, 2, \infty$) die p-Norm auf dem $\mathbb{R}^n$, also

$$\|x\|_p := \Big(\sum_{j=1}^{n} |x_j|^p\Big)^{1/p} \quad (1 \le p < \infty), \qquad \|x\|_\infty := \max_{j=1,\dots,n} |x_j|.$$

Es sollte bekannt sein, dass alle Normen auf dem $\mathbb{R}^n$ *äquivalent* sind (und was das genau heißt). Unter einer *Matrixnorm* auf dem $\mathbb{R}^{m\times n}$ verstehen[1] wir eine Norm (also eine rellwertige Abbildung, die nichtnegativ und definit, "homogen" ist und der Dreiecksungleichung genügt) auf dem mn-dimensionalen Raum der $m \times n$-Matrizen[2]. Sind Normen auf $\mathbb{R}^m$ und dem $\mathbb{R}^n$ gegeben (beide werden mit $\|\cdot\|$ bezeichnet, da aus dem Zusammenhang hervorgeht, auf welchem Raum diese jeweils zu verstehen sind), so ist die diesen Normen *zugeordnete* (bzw. durch sie *induzierte*) *Matrixnorm* auf $\mathbb{R}^{m\times n}$ definiert durch

$$\|A\| := \max_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|.$$

Diese zugeordnete Matrixnorm besitzt die Normeigenschaften (Nichtnegativität und Definitheit, Homogenität, Dreiecksungleichung). Ferner ist sie *submultiplikativ* in dem

[1] In diesem Kapitel betrachten wir zwar fast nur quadratische Matrizen, für spätere Anwendungen ist es aber zweckmäßig, auch den Fall $m \neq n$ zuzulassen.

[2] Hier ist die Bezeichnung nicht ganz einheitlich in der Literatur. Zumindestens im quadratischen Fall wird oft auch die Submultiplikativität verlangt.

folgenden Sinne: Ist $\|\cdot\|$ eine Norm auf $\mathbb{R}^m$, $\mathbb{R}^n$ bzw. $\mathbb{R}^p$ und wird mit $\|\cdot\|$ die zugeordnete Matrixnorm auf $\mathbb{R}^{m\times n}$, $\mathbb{R}^{n\times p}$ bzw. $\mathbb{R}^{m\times p}$ bezeichnet, so ist

$$\|AB\| \le \|A\|\,\|B\| \qquad \text{für alle } A\in\mathbb{R}^{m\times n},\ B\in\mathbb{R}^{n\times p}.$$

Bekannt sollten die zugeordneten Matrixnormen für den Fall sein, dass auf dem $\mathbb{R}^m$ und dem $\mathbb{R}^n$ ein und dieselbe p-Norm mit $p=1,2,\infty$ gegeben ist:

$$\begin{aligned}
\|A\|_1 &:= \max_{x\neq 0}\frac{\|Ax\|_1}{\|x\|_1} &&= \max_{j=1,\dots,n}\sum_{i=1}^{m}|a_{ij}| && \text{(maximale SBSN)},\\
\|A\|_2 &:= \max_{x\neq 0}\frac{\|Ax\|_2}{\|x\|_2} &&= \sqrt{\lambda_{\max}(A^TA)} && \text{(Spektralnorm)},\\
\|A\|_\infty &:= \max_{x\neq 0}\frac{\|Ax\|_\infty}{\|x\|_\infty} &&= \max_{i=1,\dots,m}\sum_{j=1}^{n}|a_{ij}| && \text{(maximale ZBSN)}.
\end{aligned}$$

Hierbei bedeutet $\lambda_{\max}(A^TA)$ den größten Eigenwert der (symmetrischen und positiv semidefiniten) Matrix A^TA. Wichtig ist noch die *Frobenius-Norm* auf $\mathbb{R}^{m\times n}$, die durch

$$\|A\|_F := \Big(\sum_{i=1}^{m}\sum_{j=1}^{n}a_{ij}^2\Big)^{1/2}$$

definiert[3] ist. Die Frobeniusnorm $\|\cdot\|_F$ ist eine Matrixnorm, sie entsteht dadurch, dass man Matrizen aus $\mathbb{R}^{m\times n}$ als Vektoren in $\mathbb{R}^{mn}$ auffasst und hierauf die euklidische Norm anwendet. Ferner ist die Frobeniusnorm submultiplikativ, denn für $A\in\mathbb{R}^{m\times n}$ und $B\in\mathbb{R}^{n\times p}$ ist unter Benutzung der Cauchy-Schwarzschen Ungleichung

$$\begin{aligned}
\|AB\|_F^2 &= \sum_{i=1}^{m}\sum_{j=1}^{p}(AB)_{ij}^2\\
&= \sum_{i=1}^{m}\sum_{j=1}^{p}\Big(\sum_{k=1}^{n}a_{ik}b_{kj}\Big)^2 \le \sum_{i=1}^{m}\sum_{j=1}^{p}\Big(\sum_{k=1}^{n}a_{ik}^2\sum_{k=1}^{n}b_{kj}^2\Big)\\
&= \|A\|_F^2\,\|B\|_F^2.
\end{aligned}$$

Schließlich ist der *Spektralradius* $\rho(A)$ einer (quadratischen) Matrix $A\in\mathbb{R}^{n\times n}$ definiert als das Maximum der Beträge der Eigenwerte von A, also

$$\rho(A) := \max\{|\lambda| :\ \lambda \text{ ist Eigenwert von } A\}.$$

Der Kreis in der komplexen Zahlenebene um Null mit dem Radius $\rho(A)$ ist also der kleinste Kreis um den Nullpunkt, der alle Eigenwerte von A enthält.

Ferner werden natürlich die Grundbegriffe der linearen Algebra vorausgesetzt und nicht noch einmal erläutert. So sollten die Begriffe symmetrische, hermitesche, orthogonale, unitäre, positiv (semi)definite Matrix bekannt sein, außerdem sollten Sie wissen, dass sich jede (quadratische) hermitesche Matrix durch eine unitäre Ähnlichkeitstransformation auf Diagonalgestalt transformieren läßt usw. Die Aufgaben sollen testen, ob Sie mit diesen Grundbegriffen bzw. grundlegenden Aussagen umgehen können.

[3] Im Komplexen muss man natürlich a_{ij}^2 durch $|a_{ij}|^2$ ersetzen.

2.1.1 MATLAB-Ergänzungen

Durch `help norm` kann man sich über die MATLAB-Funktion `norm` informieren:

```
norm   Matrix or vector norm.
    norm(X,2) returns the 2-norm of X.
    norm(X) is the same as norm(X,2).
    norm(X,1) returns the 1-norm of X.
    norm(X,Inf) returns the infinity norm of X.
    norm(X,'fro') returns the Frobenius norm of X.

 In addition, for vectors...
    norm(V,P) returns the p-norm of V defined as SUM(ABS(V).^P)^(1/P).
    norm(V,Inf) returns the largest element of ABS(V).
    norm(V,-Inf) returns the smallest element of ABS(V).

 By convention, NaN is returned if X or V contains NaNs.

 See also cond, rcond, condest, normest, hypot.
```

Im Gegensatz zur Matrixnorm steht also die Vektornorm $\|\cdot\|_p$ für beliebiges $p \in (0, \infty)$ durch `NORM(V,P)` zur Verfügung (ist aber bekanntlich für $p \in (0, 1)$ keine Norm). Mit Hilfe des Skript M-files

```
x=rand(500,1);
p=linspace(2,10);
for i=1:100
  y(i)=norm(x,p(i));
end;
plot(p,y);
xlabel('p','FontSize',15);
title('norm(x,p) fuer p zwischen 2 und 10','FontSize',15);
```

erhalten wir den Plot in Abbildung 2.1, in dem mit einen Vektor $x \in \mathbb{R}^{500}$, dessen Komponenten zufällig aus $(0, 1)$ gewählt sind, $\|x\|_p$ für $p \in [2, 10]$ geplottet wird.

2.1.2 Aufgaben

1. Sei $\|\cdot\|$ eine Norm auf $\mathbb{R}^n$ bzw. die zugeordnete Matrixnorm auf $\mathbb{R}^{n\times n}$ und $C \in \mathbb{R}^{n\times n}$ nichtsingulär. Man zeige, dass durch $\|x\|_C := \|Cx\|$ eine Norm auf dem $\mathbb{R}^n$ gegeben ist und berechne die zugeordnete Matrixnorm.

2. Ist $A \in \mathbb{R}^{n\times n}$ symmetrisch, so ist

$$\lambda_{\min}(A)\,\|x\|_2^2 \le x^T A x \le \lambda_{\max}(A)\,\|x\|_2^2 \qquad \text{für alle } x \in \mathbb{R}^n,$$

   wobei $\lambda_{\min}(A)$ den kleinsten und $\lambda_{\max}(A)$ den größten Eigenwert von A bedeutet.

3. Sei $A \in \mathbb{R}^{m\times n}$. Dann gilt:

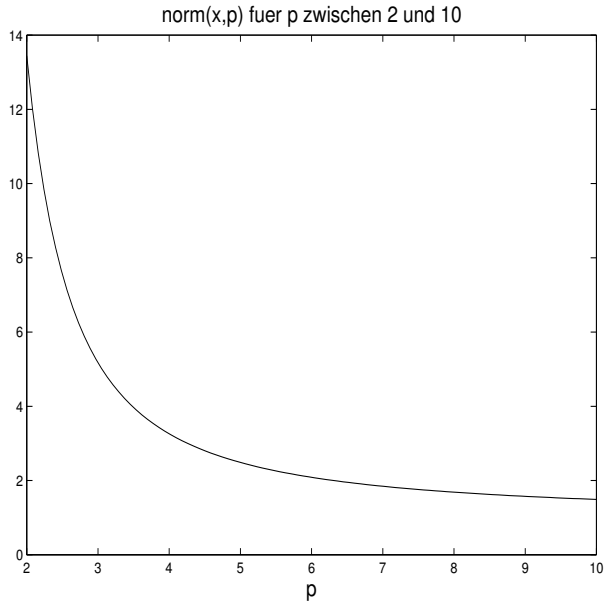


Abbildung 2.1: $\|x\|_p$ für $p \in [2, 10]$

(a) $\|A\|_2 = \|A^T\|_2 \leq \|A\|_F = \|A^T\|_F$,

(b) $\|A\|_\infty \leq \sqrt{n}\,\|A\|_2$,

(c) $\|A\|_2 \leq \sqrt{m}\|A\|_\infty$.

(d) $\|A\|_2 \leq \sqrt{\|A\|_1\,\|A\|_\infty}$.

4. Sei $A \in \mathbb{R}^{m\times n}$, ferner $Q \in \mathbb{R}^{m\times m}$ und $Z \in \mathbb{R}^{n\times n}$ orthogonal. Dann ist $\|QAZ\|_2 = \|A\|_2$ und $\|QAZ\|_F = \|A\|_F$.

5. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit. Man zeige, dass es positive Konstanten c_0, C_0 mit

$$c_0\,\|Ax\|_2^2 \leq x^T Ax \leq C_0\,\|Ax\|_2^2 \qquad \text{für alle } x \in \mathbb{R}^n$$

gibt. Insbesondere gilt: Ist $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit, so folgt aus $x^T Ax = 0$, dass $Ax = 0$.

6. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv (semi)definit. Man zeige, dass es eine symmetrische und positiv (semi)definite Matrix $B \in \mathbb{R}^{n\times n}$ mit $B^2 = A$ gibt. (Man nennt B dann (nichtnegative bzw. positive) Quadratwurzel aus A und schreibt hierfür $A^{1/2}$.)

7. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit. Dann gilt die *verallgemeinerte Cauchy-Schwarzsche Ungleichung*, dass nämlich für alle $x, y \in \mathbb{R}^n$ die Ungleichung

$$(x^T Ay)^2 \leq (x^T Ax)(y^T Ay)$$

besteht.

8. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit. Es sei

$$U^T AU = \Lambda = \text{diag}\,(\lambda_1, \ldots, \lambda_n),$$

wobei $U \in \mathbb{R}^{n\times n}$ orthogonal ist und $\lambda_1, \ldots, \lambda_n$ die (nichtnegativen) Eigenwerte von A sind. Mit $\Lambda^{1/2} := \operatorname{diag}(\lambda_1^{1/2}, \ldots, \lambda_n^{1/2})$ sei die symmetrische, positiv semidefinite Matrix $A^{1/2}$ definiert durch $A^{1/2} := U\Lambda^{1/2}U^T$. Sei $B \in \mathbb{R}^{n\times n}$ eine weitere, symmetrische und positiv semidefinite Matrix mit $B^2 = A$. Man zeige:

(a) Die Matrizen B und A sind vertauschbar.

(b) B ist mit jedem Polynom in A vertauschbar.

(c) B und $A^{1/2}$ sind vertauschbar.

(d) Es ist $B = A^{1/2}$. Insgesamt existiert zu einer symmetrischen, positiv semidefiniten Matrix genau eine symmetrische, positiv semidefinite Quadratwurzel.

9. Seien $A, B \in \mathbb{R}^{n\times n}$ symmetrisch, positiv (semi)definit und $AB = BA$, also A und B vertauschbar. Dann ist auch AB symmetrisch und positiv (semi)definit.

10. Seien $A, B \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Dann existiert genau eine symmetrische, positiv definite Matrix $X \in \mathbb{R}^{n\times n}$ mit $AX + XA = B$. Weiter zeige man: Sind A und B zusätzlich miteinander vertauschbar, so ist $X := \frac{1}{2}BA^{-1}$ symmetrisch, positiv definit und genügt der Gleichung $AX + XA = B$.

    Hinweis: Diese Aufgabe scheint nicht ganz einfach zu lösen zu sein. Wer findet einen "elementaren" Beweis?

11. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Man entwickle das Newton-Verfahren zur Berechnung der positiven Quadratwurzel $A^{1/2}$ und zeige für einen Startwert X_0, der symmetrisch, positiv definit und mit A vertauschbar ist (also z. B. $X_0 := I$ oder $X_0 := A$) die quadratische Konvergenz. Als Beispiel berechne man die symmetrische, positiv definite Quadratwurzel aus

$$A := \begin{pmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{pmatrix}.$$

    Hinweis: Auch für diese Aufgabe werden Vorkenntnisse benötigt, da man insbesondere wissen sollte, was das Newton-Verfahren ist.

12. Sei $s \in \mathbb{R}^n \setminus \{0\}$ und $E \in \mathbb{R}^{n\times n}$. Man zeige, dass

$$\left\| E\left(I - \frac{ss^T}{s^Ts}\right)\right\|_F^2 = \|E\|_F^2 - \frac{\|Es\|_2^2}{\|s\|_2^2}.$$

2.2 Störungstheorie für lineare Gleichungssysteme

In der Störungstheorie interessiert man sich dafür, wie die Lösung eines Problems sich verändert, wenn die Daten gestört werden. Die hier gestellte Frage hat also nichts mit einem Verfahren zu tun, sondern nur mit der "Kondition" des Problems.

Beispiel: Sei $f \in C^1[a, b]$. Das Problem bestehe darin, $f(x)$ für ein $x \in (a, b)$ zu berechnen. Nun kennt man x aber nicht exakt, oder, anders gewendet, wird x durch

einen (im Computer fast immer unumgänglichen) Eingabefehler zu $x + \delta x$ verfälscht (wobei man i. Allg. eine Schranke für $|\delta x|$ kennt). Ohne weitere Informationen können wir nur $f(x + \delta x)$ berechnen und versuchen, den absoluten Fehler $|f(x + \delta x) - f(x)|$ abzuschätzen bzw. zu schätzen. Durch eine lineare Entwicklung erhält man $f(x+\delta x) \approx f(x)+f'(x)\delta x$ und damit $|f(x+\delta x)-f(x)| \approx |\delta x|\,|f'(x)|$. Wir nennen $|f'(x)|$ die *absolute Konditionszahl* von f in x. Ist $|f'(x)|$ hinreichend groß, so kann der Fehler sogar dann groß sein, wenn $|\delta x|$ klein ist. Dann nennen wir f *schlecht konditioniert* in x. Aus

$$\frac{|f(x + \delta x) - f(x)|}{|f(x)|} \approx \frac{|\delta x|}{|x|} \frac{|f'(x)|\,|x|}{|f(x)|}$$

erhält man, dass man $|f'(x)|\,|x|/|f(x)|$ als *relative Konditionszahl* oder einfach nur *Kondition* von f in x bezeichnen könnte. Hier wird der relative Fahler durch ein Vielfaches des relativen Eingabefehlers geschätzt. Etwas genauere Definitionen für die absolute und die relative Konditionszahl findet man bei L. N. Trefethen, D. Bau, III (1997, S. 90). □

Diese Überlegungen werden gleich auf lineare Gleichungssysteme übertragen.

Vorher wollen wir aber noch andeuten, wann wir ein Verfahren zur Lösung eines Problems *rückwärts stabil* nennen. Grob gesagt ist dies deann der Fall, wenn das Verfahren die richtige Antwort für ein nur wenig gestörtes Problem liefert. Wir werden hierauf zurückkommen.

2.2.1 Störungslemma, Kondition, Störungssätze

In diesem Unterabschnitt geben wir das wohlbekannte Störungslemma an, definieren die Kondition einer (nichtsingulären) Matrix und zitieren einen aus der numerischen Mathematik wohlbekannten Störungssatz. Im folgenden sei $\|\cdot\|$ eine beliebige Norm auf dem $\mathbb{R}^n$ bzw. die zugeordnete Matrixnorm.

Lemma 2.1 *Sei $A \in \mathbb{R}^{n\times n}$ nichtsingulär. Ist $\delta A \in \mathbb{R}^{n\times n}$ und $\|A^{-1}\|\,\|\delta A\| < 1$, so ist auch $A + \delta A$ nichtsingulär und*

$$\|(A + \delta A)^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\,\|\delta A\|}.$$

Beweis: Siehe z. B. J. Werner (1992a, S. 25). □

Definition 2.2 Sei $A \in \mathbb{R}^{n\times n}$ nichtsingulär. Dann heißt $\kappa(A) := \|A\|\,\|A^{-1}\|$ die *Kondition* von A bezüglich[4] der Norm $\|\cdot\|$.

Bemerkung: In dem (hoffentlich) motivierenden Beispiel zu Beginn dieses Abschnitts betrachteten wir das Beispiel, eine Funktion an einer Stelle x zu auszuwerten, obwohl nur $f(x + \delta x)$ berechnet werden kann. Entsprechend betrachten wir die Aufgabe, die

[4]Wenn wir die Abhängigkeit der Kondition von der Norm verdeutlichen wollen, machen wir dies mit Hilfe eines Index. So ist z. B. $\kappa_2(A) = \|A\|_2\,\|A^{-1}\|_2$.

nichtsinguläre Matrix A zu invertieren, obwohl nur $(A+\delta A)^{-1}$ berechnet werden kann. Ist $\|A^{-1}\|\,\|\delta A\| < 1$, so ist (Neumannsche Reihe!)

$$(A+\delta A)^{-1} = (I+A^{-1}\delta A)^{-1}A^{-1} = \Big(\sum_{i=0}^{\infty}(-1)^i(A^{-1}\delta A)^i\Big)A^{-1} = A^{-1}-A^{-1}(\delta A)A^{-1}+\cdots.$$

Daher ist

$$\frac{\|A^{-1}-(A+\delta A)^{-1}\|}{\|A^{-1}\|} \approx \frac{\|A^{-1}(\delta A)A^{-1}\|}{\|A^{-1}\|} \approx \kappa(A)\,\frac{\|\delta A\|}{\|A\|}.$$

Die oben definierte Kondition $\kappa(A)$ von A würde man also eigentlich besser relative Kondition von A bezüglich der Matrixinversion nennen. □

Satz 2.3 *Sei $A \in \mathbb{R}^{n\times n}$ nichtsingulär und $\delta A \in \mathbb{R}^{n\times n}$ eine Matrix mit $\|A^{-1}\|\,\|\delta A\| < 1$. Mit vorgegebenen $b \in \mathbb{R}^n \setminus \{0\}$, $\delta b \in \mathbb{R}^n$ seien $x, \delta x$ definiert durch $Ax = b$ bzw. $(A+\delta A)(x+\delta x) = b+\delta b$. Dann ist*

$$\frac{\|\delta x\|}{\|x\|} \le \frac{\kappa(A)}{1-\kappa(A)\,\|\delta A\|/\|A\|}\Big\{\frac{\|\delta A\|}{\|A\|}+\frac{\|\delta b\|}{\|b\|}\Big\}.$$

Beweis: Siehe z. B. J. Werner (1992a, S. 26). □

Im nun folgenden Satz (siehe J. W. Demmel (1997, S. 33)) wird die Bedingung $\|A^{-1}\|\,\|\delta A\| < 1$ bzw. $\|\delta A\|/\|A\| < 1/\kappa(A)$ motiviert. Wir beschränken uns auf die euklidische Norm bzw. ihre zugeordnete Matrixnorm.

Satz 2.4 *Sei $A \in \mathbb{R}^{n\times n}$ nichtsingulär. Dann ist*

$$\min\Big\{\frac{\|\delta A\|_2}{\|A\|_2} : A+\delta A \ \textit{ singulär}\Big\} = \frac{1}{\kappa_2(A)}.$$

Beweis: Offenbar genügt es zu zeigen, dass

$$\min\{\|\delta A\|_2 : A+\delta A \ \text{ singulär}\} = \frac{1}{\|A^{-1}\|_2}.$$

Zunächst beachten wir, dass wegen des Störungslemmas $\|\delta A\|_2 \ge 1/\|A^{-1}\|_2$ für jedes δA, für das $A+\delta A$ singulär ist, das obige Minimum ist also $\ge 1/\|A^{-1}\|_2$. Um zu zeigen, dass das Minimum gleich $1/\|A^{-1}\|$ ist, konstruieren wir ein δA mit $\|\delta A\|_2 = 1/\|A^{-1}\|$ derart, dass $A+\delta A$ singulär ist. Nach Definition von $\|A^{-1}\|_2$ existiert ein $x \in \mathbb{R}^n$ mit $\|x\|_2 = 1$ und $\|A^{-1}\|_2 = \|A^{-1}x\|_2$. Man definiere

$$y := \frac{A^{-1}x}{\|A^{-1}\|_2}, \qquad \delta A := -\frac{xy^T}{\|A^{-1}\|_2}.$$

Nach Wahl von x ist $\|y\|_2 = 1$. Ferner ist

$$\|\delta A\|_2 = \max_{z\neq 0}\frac{\|(y^Tz)x\|_2}{\|A^{-1}\|_2\,\|z\|_2} = \frac{\|x\|_2}{\|A^{-1}\|_2}\max_{z\neq 0}\frac{|y^Tz|}{\|z\|_2} = \frac{1}{\|A^{-1}\|_2}.$$

Da schließlich $A + \delta A$ wegen

$$(A + \delta A)y = Ay - \frac{x}{\|A^{-1}\|_2} = 0$$

singulär ist, ist der Satz bewiesen. □

Bemerkung: Angenommen, $\hat{x}$ sei eine Näherungslösung des linearen Gleichungssystems $Ax = b$, etwa das Resultat eines gewissen Algorithmus. Wir fragen uns nach der in der Spektralnorm kleinsten Störung δA mit $(A + \delta A)\hat{x} = b$, also nach dem in der Koeffizientenmatrix am wenigsten gestörten linearen Gleichungssystem, welches $\hat{x}$ als exakte Lösung besitzt. Wegen $(\delta A)\hat{x} = b - A\hat{x}$ ist zunächst (hier ist die Norm noch beliebig)

$$\frac{\|b - A\hat{x}\|}{\|\hat{x}\|} \leq \|\delta A\|.$$

Man erkennt leicht (jetzt sei $\|\cdot\|$ die euklidische Norm bzw. die zugeordnete Spektralnorm), dass durch

$$\delta A := \frac{(b - A\hat{x})\hat{x}^T}{\|\hat{x}\|_2^2}$$

diese untere Schranke angenommen wird. □

2.2.2 Die Schätzung der Kondition

Um die Kondition $\kappa(A) = \|A\|\,\|A^{-1}\|$ einer nichtsingulären Matrix $A \in \mathbb{R}^{n\times n}$ abzuschätzen oder zu schätzen, kommt es darauf an, $\|A^{-1}\|$ (ab)zuschätzen, da $\|A\|$ i. Allg. (man denke etwa an die Maximumnorm) einfach zu berechnen ist. Eine direkte Berechnung von A^{-1} und $\|A^{-1}\|$ kommt nicht in Frage, da die Anzahl der Operationen proportional zu n^3 wäre. Gesucht ist ein "billigeres" Verfahren zur *Schätzung* von $\|A^{-1}\|$, die Anzahl der Operationen sollte durch $O(n^2)$ gegeben sein.

Wir schildern ein Verfahren, das auf W. W. Hager (1984)[5] zurückgeht. Hingewiesen sei auch auf einen Übersichtsartikel von N. J. Higham (1987)[6]. Das Verfahren geht davon aus, dass $\|B\|_1$ geschätzt werden soll (genauer sollen möglichst gute untere Schranken bestimmt werden), wobei wir Bx und B^Ts für beliebige x und s berechnen können. Uns interessiert natürlich der Fall, dass $B = A^{-1}$. Wir gehen davon aus (im nächsten Abschnitt werden die entsprechenden, aus der numerischen Mathematik bekannten Begriffe und Aussagen wiederholt), dass eine LU-Zerlegung von A berechnet ist, also eine Permutationsmatrix P sowie eine untere Dreiecksmatrix L mit Einsen in der Diagonalen und eine (nichtsinguläre) obere Dreiecksmatrix R mit $PA = LU$. Wie beim Gaußschen Eliminationsverfahren gewinnt man dann $Bx = A^{-1}x$ durch Vorwärts- und Rückwärtseinsetzen, entsprechendes gilt für die Berechnung von $B^Ts = A^{-T}s$. Beides benötigt jeweils n^2 flops.

[5] W. W. Hager (1984) "Condition estimators." SIAM J. Sci. Statist. Comput. 5, 311–316.

[6] N. J. Higham (1987) "A survey of condition number estimation for triangular matrices." SIAM Review 29, 575–596.

Es ist

$$\|B\|_1 = \max_{x\neq 0} \frac{\|Bx\|_1}{\|x\|_1} = \max_{\|x\|_1\leq 1} \|Bx\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |b_{ij}|.$$

Das Maximum auf der rechten Seite wird für ein gewisses j_0 angenommen, es ist dann $\|B\|_1 = \|Be_{j_0}\|_1$. Es wäre zu teuer, alle Spalten von B zu berechnen und die mit maximaler Länge zu bestimmen. Daher wendet man ein Iterationsverfahren auf die Aufgabe an, die Funktion $f(x) := \|Bx\|_1$ unter der Nebenbedingung $\|x\|_1 \leq 1$ zu maximieren. Die Funktion f ist in jedem x mit $(Bx)_i \neq 0$, $i = 1, \ldots, n$, (alle Komponenten von Bx sind von Null verschieden) stetig differenzierbar und der Gradient in einem solchen Punkt ist gegeben durch

$$\nabla f(x) = B^T s(x) \qquad \text{mit} \qquad s(x) := \operatorname{sign}(Bx).$$

Hierbei ist sign angewandt auf einen Vektor, dessen Komponenten alle von Null verschieden sind, natürlich komponentenweise zu verstehen (in MATLAB wird das automatisch so verstanden). Wir geben nun einen Schritt des Verfahrens zur Schätzung von $\|B\|_1$ an, wobei $B \in \mathbb{R}^{n\times n}$ gegeben ist und davon ausgegangen wird, dass Bx und $B^T s$ "billig" berechnet werden können. Als Startwert für das Iterationsverfahren bietet sich $x = (1/n)e$ an, weil man dadurch sozusagen keine Spalte von B einer anderen gegenüber bevorzugt.

- Gegeben $x \in \mathbb{R}^n$ mit $\|x\|_1 = 1$.
- Berechne $w := Bx$, $s := \operatorname{sign}(w)$ und $z := B^T s$.
- Falls $\|z\|_\infty \leq z^T x$, dann: STOP, $\gamma := \|w\|_1$ ist untere Schranke von $\|B\|_1$.
- Andernfalls wähle $j \in \{1, \ldots, n\}$ mit $|z_j| = \|z\|_\infty$ und setze $x_+ := e_j$.

Um diesen Algorithmus zu rechtfertigen, überlegen wir uns:

(1) Sei $x \in \mathbb{R}^n$ mit $\|x\|_1 = 1$ und $(Bx)_i \neq 0$, $i = 1, \ldots, n$, gegeben. Mit $z := B^T \operatorname{sign}(Bx)$ sei $\|z\|_\infty \leq z^T x$. Dann ist in x eine lokale Lösung der Aufgabe, f auf der Einheitskugel (bezüglich der Betragssummennorm) zu maximieren.

Denn: Da nach Voraussetzung alle Komponenten von Bx nicht verschwinden, gibt es eine Umgebung U von x mit $\operatorname{sign}(Bu) = \operatorname{sign}(Bx)$ für alle $u \in U$. Für $u \in U$ ist $f(u) = f(x) + z^T(u - x)$. Sei nun $u \in U$ mit $\|u\|_1 \leq 1$ gegeben. Wir zeigen, dass $z^T(u - x) \leq 0$ und damit $f(u) \leq f(x)$. Denn

$$z^T(u - x) \leq \|z\|_\infty \|u\|_1 - z^T x \leq \|z\|_\infty - z^T x \leq 0.$$

Damit ist die obige Abbruchbedingung gerechtfertigt. Die Wahl von x_+, der neuen und hoffentlich verbesserten Näherung, wird durch die nächste Aussage motiviert.

(2) Sei $x \in \mathbb{R}^n$ mit $\|x\|_1 = 1$ und $(Bx)_i \neq 0$, $i = 1, \ldots, n$, gegeben. Mit $z := B^T \operatorname{sign}(Bx)$ sei $\|z\|_\infty > z^T x$. Ist dann $j \in \{1, \ldots, n\}$ so gewählt, dass $|z_j| = \|z\|_\infty$ und $x_+ := e_j$ gesetzt, so ist $f(x_+) > f(x)$.

Denn: Wir setzen $\tilde{x} := \operatorname{sign}(z_j)e_j$. Wegen $f(x_+) = f(\tilde{x})$ ist dann

$$\begin{aligned} f(x_+) - f(x) &= f(\tilde{x}) - f(x) \\ &\geq \nabla f(x)^T(\tilde{x} - x) \\ &\qquad \text{(da } f \text{ konvex)} \\ &= z^T(\tilde{x} - x) \\ &= |z_j| - z^T x \\ &= \|z\|_\infty - z^T x \\ &> 0, \end{aligned}$$

womit auch die zweite Aussage bewiesen ist. Bemerkt sei noch: Wendet man obigen Algorithmus auf B^T statt B an, so erhält man eine untere Schranke für $\|B\|_\infty$.

Bei N. J. Higham (1996, S. 295) findet man eine etwas verbesserte Version des letzten Algorithmus. Bei dieser werden insbesondere mindestens zwei und höchstens fünf Iterationsschritte durchgeführt.

2.2.3 MATLAB-Ergänzungen

Zur Berechnung der Kondition einer Matrix steht in MATLAB die Funktion `cond` zur Verfügung. Nach `help cond` erhält man:

```
cond   Condition number with respect to inversion.
   cond(X) returns the 2-norm condition number (the ratio of the
   largest singular value of X to the smallest).  Large condition
   numbers indicate a nearly singular matrix.

   cond(X,P) returns the condition number of X in P-norm:

      NORM(X,P) * NORM(INV(X),P).

   where P = 1, 2, inf, or 'fro'.
```

Mit `type cond` kann man sich sogar die in MATLAB geschriebene Funktion `cond` genauer ansehen. Man stellt fest, dass für $p \neq 2$ und $m = n$, wobei `[m,n]=size(A)`, die Kondition durch

```
cond(A,p)=norm(A, p)*norm(inv(A), p)
```

realisiert wird. Die bekanntesten schlecht konditionierten Matrizen sind die Hilbert-Matrizen. Nach

```
A=hilb(7); c=cond(A,1)
```

erhält man `c=9.8519e+08`. Es gibt auch andere Zahldarstellungen. Einen Überblick hierüber erhält man nach dem Befehl `help format`. Nach `format long g` und Eingabe

von c erhält man z. B. die Darstellung `c=985194889.71947`, nach dem Umschaltbefehl `format long e` ist `c=9.851948897194699e+08`.

Zur Schätzung der Kondition einer Matrix steht die Funktion `condest` zur Verfügung. Eine Beschreibung der Funktion ist gegeben durch:

```
condest 1-norm condition number estimate.
    C = condest(A) computes a lower bound C for the 1-norm condition
    number of a square matrix A.

    C = condest(A,T) changes T, a positive integer parameter equal to
    the number of columns in an underlying iteration matrix. Increasing
    the  number of columns usually gives a better condition estimate
    but increases the cost.  The default is T = 2, which almost always
    gives an estimate correct to within a factor 2.
```

Nach `type condest` kann man sich die in MATLAB realisierte Funktion näher anschauen und versuchen, die Funktionsweise zu verstehen. Wir wollen eine einfachere Version `kondest` vorstellen, bei welcher maximal drei Schritte des oben angegebenen Verfahrens durchgeführt werden, und diese mit `condest` und `cond` vergleichen.

```
  function est=kondest(A);
%***********************************************
% Pre:
%   A    n-by-n matrix
% Post:
%   est  estimate for cond_1(A)
%***********************************************
  [n,n]=size(A);  x=(1/n)*ones(n,1);  [L,U]=lu(A);
  anorm=norm(A,1);
  for k=1:3
     w=U\(L\x);  s=sign(w);  z=L'\(U'\s);
     [big,j]=max(abs(z));
     if big<=z'*x
        gamma=norm(w,1);
        break;
     end;
     x=zeros(n,1);  x(j)=1;
  end;
  est=norm(w,1)*anorm;
```

Wir erhalten

```
condest(hilb(7))=9.851948869986799e+08,
kondest(hilb(7))=9.851948869986801e+08,
cond(hilb(7))= 4.753673562966472e+08.
```

MATLAB stellt ferner die Funktion `normest` zur Schätzung der Spektralnorm einer Matrix zur Verfügung. Diesem Schätzer liegt das Verfahren der Vektoriteration zugrunde. Wir kommen darauf später zurück.

2.2.4 Aufgaben

1. Was ist die Kondition für das Problem, $\sqrt{x}$ für $x > 0$ zu berechnen?

2. Sei $\|\cdot\|$ eine *absolute Norm* auf dem $\mathbb{R}^n$, d. h. $\|x\| = \| \, |x| \, \|$ für alle $x \in \mathbb{R}^n$, mit $\|\cdot\|$ sei auch die zugeordnete Matrixnorm bezeichnet. Die Matrix $A \in \mathbb{R}^{n\times n}$ sei nichtsingulär und x die Lösung von $Ax = b$ mit $b \in \mathbb{R}^n \setminus \{0\}$. Mit $\epsilon > 0$ sei $\delta A \in \mathbb{R}^{n\times n}$ eine Matrix mit $|\delta A| \le \epsilon \, |A|$. Sei $(A + \delta A)\hat{x} = b$. Man zeige, dass

$$\frac{\|\hat{x} - x\|}{\|\hat{x}\|} \le \epsilon \, \| \, |A^{-1}| \, |A| \, \|.$$

3. Sei $U = (u_{ij}) \in \mathbb{R}^{n\times n}$ eine nichtsinguläre obere Dreieckmatrix. Dann ist

$$\left(\min_{i=1,\ldots,n} |u_{ii}| \right)^{-1} \le \|U^{-1}\|_p, \qquad p = 1, 2, \infty.$$

4. Sei $A \in \mathbb{R}^{n\times n}$ nichtsingulär. Ziel ist es, (möglichst gute) untere Schranken für $\|A\|_2$ zu erhalten, wobei nur ausgenutzt werden soll, dass wir Ax und $A^T y$ für beliebige x, y berechnen können. (Liegt eine LU-Zerlegung von A vor, so kann auf diese Weise eine untere Schranke für $\kappa_2(A)$ gewonnen werden). Man betrachte ein Verfahren, von dem wir einen Schritt angeben:

   - Gegeben $x \in \mathbb{R}^n$ mit $\|x\|_2 = 1$.
   - Berechne $y := Ax$, $\gamma := \|y\|_2$ und $z := (1/\gamma)A^T y$.
   - Falls $\|z\|_2 \le z^T x$, dann: STOP, Andernfalls: Berechne $x_+ := z/\|z\|$.

   Man zeige:

   (a) Ist die Abbruchbedingung $\|z\|_2 \le z^T x$ erfüllt, so ist x eine stationäre Lösung für die Aufgabe (jetzt ist x als variabel aufzufassen), $f(x) := \|Ax\|_2^2$ unter der Nebenbedingung $\|x\|_2 \le 1$ zu maximieren.

   (b) Ist die Abbruchbedingung nicht erfüllt, so ist $f(x_+) > f(x)$.

5. Sei $R \in \mathbb{R}^{n\times n}$ eine nichtsinguläre obere Dreiecksmatrix. Man definiere die (ebenfalls nichtsinguläre) obere Dreiecksmatrix $U \in \mathbb{R}^{n\times n}$ durch

$$u_{ij} := \begin{cases} |r_{ii}|, & \text{für} \quad i = j, \\ -|r_{ij}|, & \text{für} \quad i \neq j. \end{cases}$$

   Man zeige:

   (a) Es ist $U^{-1} \ge 0$.

   (b) Es ist $\|R^{-1}\|_\infty \le \|U^{-1}\|_\infty$.
   Hinweis: Man zeige, dass $|R^{-1}| \le U^{-1}$.

(c) Es ist $\|U^{-1}\|_\infty = \|U^{-1}e\|_\infty$, wobei $e := (1, \ldots, 1)^T$.

(d) Durch den folgenden Algorithmus wird eine obere Schranke γ_U für $\|R^{-1}\|_\infty$ berechnet:

- Für $i = n, \ldots, 1$:
  - $z_i := (1 + \sum_{j=i+1}^{n} |r_{ij}|\, z_j)/|r_{ii}|$.
- $\gamma_U := \|z\|_\infty$.

6. Das lineare Gleichungssystem $Ax = b$ ist äquivalent zu $DAx = Db$, wenn D nichtsingulär ist. Interessant sind hier vor allem Diagonalmatrizen, da dann die Berechnung von DA noch verhältnismäßig billig ist (man spricht auch von einer *Skalierung* des Gleichungssystems). Je kleiner die Kondition der Koeffizientenmatrix eines linearen Gleichungssystems desto bessere Ergebnisse erwarten wir. Dies ist der Hintergrund der folgenden Aufgabe.

   Sei $A \in \mathbb{R}^{n\times n}$ nichtsingulär und $\mathcal{D}_n$ die Menge der nichtsingulären $n \times n$-Diagonalmatrizen. Man zeige:

   (a) Es ist
   $$\| \, |A^{-1}| \, |A| \, \|_\infty \le \min_{D\in\mathcal{D}_n} \kappa_\infty(DA).$$
   Weshalb dürfen wir hier "min" statt "inf" schreiben?

   (b) Definiert man $D^* = \operatorname{diag}(d_1^*, \ldots, d_n^*) \in \mathcal{D}_n$ durch
   $$d_i^* := 1 \Big/ \sum_{j=1}^{n} |a_{ij}|, \qquad i = 1, \ldots, n,$$
   so ist $\| \, |A^{-1}| \, |A| \, \|_\infty = \kappa_\infty(D^*A)$.

   Beide Teile zusammen zeigen, dass eine sogenannte Zeilen-Äquilibrierung bezüglich der Maximumnorm eine optimale Kondition liefert.

7. Sei
   $$A = \begin{pmatrix} a_1^T \\ \vdots \\ a_n^T \end{pmatrix} \in \mathbb{R}^{n\times n}$$
   nichtsingulär. Man zeige:

   (a) Es ist $\max_{i=1,\ldots,n} \|a_i\|_2 \le \|A\|_2 \le \sqrt{n}\, \max_{i=1,\ldots,n} \|a_i\|_2$.

   (b) Definiert man
   $$\hat{d}_i := \frac{1}{\|a_i\|_2} \quad (i = 1, \ldots, n), \qquad \hat{D} := \operatorname{diag}(\hat{d}_1, \ldots, \hat{d}_n),$$
   so ist
   $$\kappa_2(\hat{D}A) \le \sqrt{n} \min_{D\in\mathcal{D}_n} \kappa_2(DA),$$
   wobei $\mathcal{D}_n$ wieder die Menge der nichtsingulären $n \times n$-Diagonalmatrizen bedeutet.

2.3 Gauß-Elimination

Das Gaußche Eliminationsverfahren ist aus der numerischen Mathematik wohlbekannt. Wir werden uns daher kurz fassen und uns auf Dinge konzentrieren, die normalerweise in einer Vorlesung über numerische Mathematik kaum gebracht werden. Trotzdem werden wir einiges wiederholen müssen, also etwas Geduld!

2.3.1 Die LU-Zerlegung

Das Gaußsche Eliminationsverfahren mit Spaltenpivotsuche basiert bekanntlich darauf, die gegebene nichtsinguläre Koeffizientenmatrix $A \in \mathbb{R}^{n\times n}$ eines linearen Gleichungssystems in $n-1$ Schritten abwechselnd von links mit *Vertauschungsmatrizen* und *Gauß-Matrizen* durchzumultiplizieren, um am Schluss als Resultat eine obere Dreiecksmatrix zu erhalten. Hierdurch ist es möglich, eine sogenannte LU-Zerlegung von A zu berechnen, d. h. eine Permutationsmatrix P, eine untere (**l**ower) Dreiecksmatrix L mit Einsen in der Diagonalen und eine obere (**u**pper) Dreiecksmatrix U derart, dass $PA = LU$.

Eine *Permutationsmatrix* ist, grob gesagt, eine Einheitsmatrix, bei der die Zeilen (bzw. die Spalten) permutiert sind. Ist genauer $p = (p(1), \ldots, p(n))$ eine Permutation der Zahlen $1, \ldots, n$, so heißt

$$P = \begin{pmatrix} e_{p(1)}^T \\ \vdots \\ e_{p(n)}^T \end{pmatrix} \in \mathbb{R}^{n\times n}$$

eine $n \times n$-Permutationsmatrix. Offenbar ist

$$\begin{pmatrix} e_{p(1)}^T \\ \vdots \\ e_{p(n)}^T \end{pmatrix} = \begin{pmatrix} e_{q(1)} & \cdots & e_{q(n)} \end{pmatrix},$$

wobei $q := p^{-1}$ die zu p inverse Permutation ist. Permutationsmatrizen sind orthogonal.

Eine *Vertauschungsmatrix* ist eine Einheitsmatrix, bei der zwei Zeilen (bzw. Spalten) vertauscht sind. Werden mit $1 \le r \le s \le n$ die r-te und die s-te Zeile in der Einheitsmatrix miteinander vertauscht, so gilt das entsprechende auch für die Spalten und die zugehörige Vertauschungsmatrix hat die Form

$$P_{rs} = \begin{pmatrix} e_1 & \cdots & e_{r-1} & e_s & e_{r+1} & \cdots & e_{s-1} & e_r & e_{s+1} & \cdots & e_n \end{pmatrix} = P_{rs}^T.$$

Offenbar ist P_{rs} eine symmetrische Permutationsmatrix, die sich in der Form

$$P_{rs} = I - (e_r - e_s)(e_r - e_s)^T$$

schreiben läßt. Eine Multiplikation einer n-zeiligen Matrix A von links bewirkt eine Vertauschung der r-ten und der s-ten Zeile von A, entsprechend ist die Multiplikation einer n-spaltigen Matrix von rechts mit P_{rs} gleichbedeutend mit einer Vertauschung der r-ten und der s-ten Spalte.

Eine Matrix der Form

$$M_k := I - l_k e_k^T \qquad \text{mit} \qquad l_k := (\underbrace{0, \ldots, 0}_{k}, l_{k+1,k}, \ldots, l_{nk})^T \in \mathbb{R}^n$$

mit $1 \leq k < n$ heißt eine *Gauß-Matrix*. Offenbar unterscheidet sich M_k nur in der k-ten Spalte, und dort auch nur in den Elementen unterhalb des Diagonalelements, von der Identität. Die Matrix M_k ist nichtsingulär, ihre Inverse ist wegen $l_k^T e_k = 0$ durch $M_k^{-1} = I + l_k e_k^T$ gegeben. Multipliziert man eine n-zeilige Matrix von links mit M_k, so bleiben die ersten k Zeilen unverändert, während man für $i = k+1, \ldots, n$ die neue i-te Zeile erhält, indem man von der alten i-ten Zeile das l_{ik}-fache der k-ten Zeile subtrahiert. Ist ein Vektor a_k gegeben, dessen k-te Komponente nicht verschwindet, so existiert offenbar genau eine Gauß-Matrix M_k mit der Eigenschaft, dass die Komponenten $(M_k a_k)_i$, $i = k+1, \ldots, n$, verschwinden.

Im Prinzip erfolgt die Berechnung der LU-Zerlegung folgendermaßen:

- Gegeben nichtsinguläre Matrix $A \in \mathbb{R}^{n \times n}$.
- Für $k = 1, \ldots, n-1$:
  - Bestimmung des Pivotelementes und der Vertauschungsmatrix.
    Bestimme $r = r(k) \in \{k, \ldots, n\}$ mit $|a_{rk}| = \max_{i=k,\ldots,n} |a_{ik}|$, setze $P_k := P_{k,r(k)}$ und berechne $A := P_k A$.
  - Bestimmung der Gauß-Matrix.
    Man definiere

    $$l_k := (\underbrace{0, \ldots, 0}_{k}, l_{k+1,k}, \ldots, l_{nk})^T \qquad \text{mit} \quad l_{ik} := \frac{a_{ik}}{a_{kk}}, \quad i = k+1, \ldots, n,$$

    und hiermit die Gauß-Matrix $M_k := I - l_k e_k^T$. Anschließend berechne $A := M_k A$.

Nach Abschluss des Verfahrens (welches bei nichtsingulärem A natürlich durchführbar ist) ist A eine nichtsinguläre, obere Dreiecksmatrix U. Bezeichnet man mit A wieder die Ausgangsmatrix, so ist also

$$M_{n-1} P_{n-1} \cdots M_1 P_1 A = U.$$

Da die Vertauschungsmatrizen $P_1, \ldots, P_{n-1}$ symmetrisch und orthogonal sind, kann man die letzte Gleichung auch in der Form

$$M'_{n-1} \cdots M'_1 P_{n-1} \cdots P_1 A = U$$

schreiben, wobei

$$M'_k := P_{n-1} \cdots P_{k+1} M_k P_{k+1} \cdots P_{n-1}, \qquad k = 1, \ldots, n-1.$$

Dann ist aber

$$\begin{aligned} M'_k &= P_{n-1}\cdots P_{k+1}(I - l_k e_k^T)P_{k+1}\cdots P_{n-1} \\ &= I - P_{n-1}\cdots P_{k+1} l_k \underbrace{e_k^T P_{k+1}\cdots P_{n-1}}_{=e_k^T} \\ &= I - l'_k e_k^T \end{aligned}$$

ebenfalls eine Gauß-Matrix mit

$$l'_k := P_{n-1}\cdots P_{k+1} l_k, \qquad k = 1, \ldots, n-1.$$

Mit

$$P := P_{n-1}\cdots P_1$$

ist daher

$$PA = (I + l'_1 e_1^T)\cdots(I + l'_{n-1} e_{n-1}^T)U = \Big(I + \sum_{k=1}^{n-1} l'_k e_k^T\Big)U = LU,$$

wobei

$$L := I + \sum_{k=1}^{n-1} l'_k e_k^T.$$

Dies bedeutet: Macht man im k-ten Schritt bei der Berechnung der Gauß-Matrix M_k die Zuweisung $a_{ik} := a_{ik}/a_{kk}$, $i = k+1, \ldots, n$, schreibt man also die relevanten Daten von M_k in die gerade frei gewordenen Positionen in der k-ten Spalte von A unterhalb des Diagonalelements, so steht nach Abschluss in der unteren Hälfte von A die untere Hälfte von L, in der oberen Hälfte von A (einschließlich der Diagonalen) steht die obere Dreiecksmatrix U.

Wir fassen die wesentlichen Ergebnisse über Existenz und Eindeutigkeit einer LR-Zerlegung zusammen.

Satz 3.1 *Ist $A \in \mathbb{R}^{n\times n}$ nichtsingulär, so existiert eine Permutationsmatrix $P \in \mathbb{R}^{n\times n}$, eine untere Dreiecksmatrix $L \in \mathbb{R}^{n\times n}$ mit Einsen auf der Diagonalen und eine obere Dreiecksmatrix $U \in \mathbb{R}^{n\times n}$ mit $PA = LU$. Eine Darstellung $A = LU$ (L, U gleiche Eigenschaften wie oben) ist eindeutig und genau dann möglich, wenn das Gaußsche Eliminationsverfahren ohne Pivotsuche durchführbar ist. Dies wiederum ist genau dann der Fall, wenn alle*[7] *Hauptabschnittsdeterminanten von A nichtsingulär sind. Ist A (symmetrisch und) positiv definit oder diagonal dominant, so ist dies der Fall.*

Der Aufwand zur Berechnung einer LU-Zerlegung (hierbei werden Vergleiche und Vertauschungen nicht gezählt) ist durch

$$G(n) = \sum_{k=1}^{n-1}\sum_{i=k+1}^{n}\Big(1 + \sum_{j=k+1}^{n} 1\Big) = \sum_{k=1}^{n-1}\sum_{i=k+1}^{n}(n-k+1) = \sum_{k=1}^{n-1}(n-k)(n-k+1) = \tfrac{1}{3}n^3 + \cdots$$

[7] Streng genommen brauchen nur die ersten $n-1$ Hauptabschnittsdeterminanten von Null verschieden zu sein, um die Existenz einer LR-Zerlegung zu sichern.

"flops" gegeben. Ist ein lineares Gleichungssystem $Ax = b$ gegeben und liegt eine LU-Zerlegung von A durch $PA = LU$ vor, so erhält man die Lösung x aus $LUx = Pb$ durch Vorwärtssubstitution (bestimme $y = Ux$ aus $Ly = Pb$) und Rückwärtssubstitution (bestimme x aus $Ux = y$). Dies ist so wohlbekannt, dass wir nicht noch einmal darauf eingehen wollen.

Bemerkung: Die Rundungsfehleranalyse beim Gaußschen Eliminationsverfahren ist eine diffizile Angelegenheit, mit der sich so berühmte Mathematiker wie A. Turing, J. von Neumann und J. H. Wilkinson beschäftigt haben. Die Frage ist die folgende: Gegeben sei ein lineares Gleichungssystem $Ax = b$ mit einer nichtsingulären Koeffizientenmatrix A. Auf dieses lineare Gleichungssystem werde das Gaußsche Eliminationsverfahren mit Spaltenpivotsuche angewandt (also zunächst Berechnung einer LU-Zerlegung, anschließend Vorwärts- und Rückwärtssubstitution). Man erhalte hiermit eine numerische Lösung $\hat{x}$ (ein Vektor aus Gleitkommazahlen, also das, was man auf dem Bildschirm sieht bzw. was ausgedruckt wird). Ist $\hat{x} \neq 0$ (und davon können wir ausgehen), so existiert ein $\delta A \in \mathbb{R}^{n\times n}$ mit $(A + \delta A)\hat{x} = b$, z. B.

$$\delta A := \frac{(b - A\hat{x})\hat{x}^T}{\|\hat{x}\|_2^2}.$$

Die Matrix δA ist durch den Vektor $\hat{x}$ aber natürlich nicht eindeutig gegeben. Was ist die minimale Norm eines δA mit $(A + \delta A)\hat{x} = b$, das ist die Frage bei der Rückwärtsanalyse des Gaußschen Eliminationsverfahrens. Den relativen Fehler zwischen der berechneten und der exakten Lösung kann man dann aus dem Störungssatz für lineare Gleichungssysteme schätzen, wenn eine Schätzung der Kondition von A vorliegt.

Bei der Rundungsfehleranalyse wird die Kenntnis einer Zahl u (unit roundoff, Maschinengenauigkeit) angenommen mit der Eigenschaft, dass

- Für alle $x \in \mathbb{R}$ existiert ein ϵ mit $|\epsilon| \leq u$ derart, dass $\mathrm{fl}(x) = x(1 + \epsilon)$ (hierbei bedeutet $\mathrm{fl}(x)$ die x zugeordnete Gleitkommazahl),
- Für alle Gleitkommazahlen x, y existiert ein ϵ mit $|\epsilon| \leq u$ und $\mathrm{fl}(x * y) = (x * y)(1 + \epsilon)$, wobei $*$ eine der vier arithmetischen Grundoperationen bedeutet. Jede Operation von Gleitkommazahlen sei also bis auf einen relativen Fehler u exakt.

Mit diesen "Axiomen" an die Gleitkomma-Arithmetik kann man sich an die (Rückwärts-) Rundungsfehleranalyse für das Gaußsche Eliminationsverfahren mit Spaltenpivotsuche (partial pivoting) machen. Wir wollen es uns hier leicht machen und zitieren nur ein Ergebnis von Wilkinson (siehe N. J. Higham (1996, S. 177)).

- Sei $A \in \mathbb{R}^{n\times n}$ nichtsingulär. Das Gaußsche Eliminationsverfahren mit Spaltenpivotsuche zur Lösung von $Ax = b$ liefere eine berechnete Lösung $\hat{x}$. Dann ist

$$(A + \delta A)\hat{x} = b \qquad \text{mit} \quad \|\delta A\|_\infty \leq 2n^2\gamma_n\rho_n \|A\|_\infty,$$

  wobei

$$\gamma_n := \frac{nu}{1 - nu}, \qquad \rho_n := \frac{\max_{i,j,k} |a_{ij}^{(k)}|}{\max_{i,j} |a_{ij}|}.$$

Hierbei wird stillschweigend $nu < 1$ angenommen (da etwa $u \approx 10^{-16}$ bei doppelt genauer Rechnung), ferner ist $A^{(k)}$ die im k-ten Schritt erhaltene transformierte Matrix, also (in obiger Sprechweise) $A^{(k)} = M_k P_k \cdots M_1 P_1 A$.

Entscheidend in der Abschätzung für δA ist der sogenannte *Wachstums-Faktor* ρ_n. Es ist leicht einzusehen (Induktion!), dass $\rho_n \leq 2^{n-1}$. Diese Schranke kann sogar für gewisse Matrizen angenommen werden. Denn z. B. ist

$$\begin{pmatrix} 1 & & & & 1 \\ -1 & 1 & & & 1 \\ -1 & -1 & 1 & & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & & & & \\ -1 & 1 & & & \\ -1 & -1 & 1 & & \\ -1 & -1 & -1 & 1 & \\ -1 & -1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & & & & 1 \\ & 1 & & & 2 \\ & & 1 & & 4 \\ & & & 1 & 8 \\ & & & & 16 \end{pmatrix}.$$

Auf Matrizen dieser Art bezieht sich die Aussage bei L. N. Trefethen, D. Bau, III (1997, S. 163), dass das Gaußsche Eliminationsverfahren "explosiv instabil" bei einer gewissen Klasse von Matrizen sein kann, aber stabil in der Praxis ist.

Mit diesen wenigen Bemerkungen zur Fehleranalyse des Gaußschen Eliminationsverfahrens wollen wir es genug sein lassen. □

Einige wenige Bemerkungen sollen noch zum Gaußschen Eliminationsverfahren bei einem linearen Gleichungssystem mit einer dünn besetzten (sparse) Koeffizientenmatrix gemacht werden (siehe J. W. Demmel (1997, S. 83 ff.)). Hier kann es sehr darauf ankommen, in welcher Reihenfolge die Gleichungen aufgeschrieben und die Variablen numeriert sind.

Beispiel: Wir gehen aus von einer $n \times n$-Matrix der Form

$$A = \begin{pmatrix} * & & & & * \\ & * & & & * \\ & & \ddots & & \vdots \\ & & & * & * \\ * & * & \cdots & * & * \end{pmatrix},$$

wobei $*$ ein (möglicherweise) von Null verschiedenes Element bedeutet. In A stehen also (höchstens) $3n - 2$ von Null verschiedene Elemente. Angenommen, A besitzt eine LU-Zerlegung (ohne Zeilenvertauschungen), es sei also $A = LU$ mit einer (normierten) unteren Dreiecksmatrix L und einer oberen Dreiecksmatrix U. Die Frage ist, ob L und U die gleiche Struktur besitzen, ob also

$$\begin{pmatrix} * & & & & * \\ & * & & & * \\ & & \ddots & & \vdots \\ & & & * & * \\ * & * & \cdots & * & * \end{pmatrix} = \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ * & * & \cdots & * & 1 \end{pmatrix} \begin{pmatrix} * & & & & * \\ & * & & & * \\ & & \ddots & & \vdots \\ & & & * & * \\ & & & & * \end{pmatrix}.$$

Durch genauere Inspektion sieht man, dass dies möglich ist. Ist z. B.

$$A = \begin{pmatrix} 1 & & & & 0.1 \\ & 1 & & & 0.1 \\ & & 1 & & 0.1 \\ & & & 1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \end{pmatrix},$$

so liefert in MATLAB die Anweisung `[L,U,P]=lu(A)`, dass $P = I$ (keine Zeilenvertauschungen nötig) und

$$A = \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ 0.1 & 0.1 & 0.1 & 0.1 & 1 \end{pmatrix} \begin{pmatrix} 1 & & & & 0.1 \\ & 1 & & & 0.1 \\ & & 1 & & 0.1 \\ & & & 1 & 0.1 \\ & & & & 0.06 \end{pmatrix}.$$

Die Faktoren L und U verschwinden hier also auch dort, wo die Ausgangsmatrix A es tut. Ganz anders sind die Verhältnisse, wenn man die Spalten und Zeilen so umnumeriert, dass die letzten die ersten, die vorletzten die zweiten usw. werden. Man erhält dann

$$A = \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 1 & & & \\ 0.1 & & 1 & & \\ 0.1 & & & 1 & \\ 0.1 & & & & 1 \end{pmatrix}.$$

Bei der Berechnung einer LU-Zerlegung stellt man wieder fest, dass keine Vertauschungen nötig sind, auf vier Dezimalstellen gerundet erhält man

$$L = \begin{pmatrix} 1.0000 & & & & \\ 1.0000 & 1.0000 & & & \\ 1.0000 & -0.1111 & 1.0000 & & \\ 1.0000 & -0.1111 & -0.1250 & 1.0000 & \\ 1.0000 & -0.1111 & -0.1250 & -0.1429 & 1.0000 \end{pmatrix}$$

und

$$U = \begin{pmatrix} 0.1000 & 0.1000 & 0.1000 & 0.1000 & 0.1000 \\ & 0.9000 & -0.1000 & -0.1000 & -0.1000 \\ & & 0.8889 & -0.1111 & -0.1111 \\ & & & 0.8750 & -0.1250 \\ & & & & 0.8571 \end{pmatrix}.$$

Hier sind also L und U vollbesetzt, das Gaußsche Eliminationsverfahren benötigt im wesentlichen genau so viele Operationen, als wenn die Matrix A vollbesetzt wäre. Es kommt also sehr darauf an, durch geeignete Zeilen- und Spaltenvertauschungen das richtige "Muster" in der Koeffizientenmatrix zu erzeugen. Wir werden später sehen,

dass Bandmatrizen das "richtige" Muster haben und machen jetzt noch ein numerisches Experiment mit einem "Schachbrettmuster". Sei

$$A = \begin{pmatrix} 1 & & 0.1 & & 0.1 \\ & 1 & & 0.1 & \\ 0.2 & & 1 & & 0.2 \\ & 0.2 & & 1 & \\ 0.2 & & 0.1 & & 1 \end{pmatrix}.$$

Wir wissen hier a priori, dass das Gaußche Eliminationsverfahren mit Spaltenpivotisierung keine Vertauschungen machen wird, da A spaltenweise diagonal dominant ist (Beweis?). Es ist $A = LU$ mit (auf vier Dezimalstellen angegeben)

$$L = \begin{pmatrix} 1.0000 & & & & \\ & 1.0000 & & & \\ 0.2000 & & 1.0000 & & \\ & 0.2000 & & 1.0000 & \\ 0.2000 & & 0.0816 & & 1.0000 \end{pmatrix}$$

und

$$U = \begin{pmatrix} 1.0000 & & 0.1000 & & 0.1000 \\ & 1.0000 & & 0.1000 & \\ & & 1.0000 & & 0.1800 \\ & & & 0.9800 & \\ & & & & 0.9653 \end{pmatrix}.$$

Also haben L und U (unter Berücksichtigung der Tatsache, dass beide Dreiecksmatrizen sind) dasselbe Muster wie A. Ist dies ein Zufall? Es muss allerdings zugegeben werden, dass das letzte Beispiel kein gutes Beispiel ist. Denn man stellt leicht fest, dass A *zerfallend* ist. Hierbei heißt eine Matrix $A \in \mathbb{R}^{n\times n}$ zerfallend (zerlegbar, reduzibel), wenn es nichtleere Teilmengen N_1, N_2 von $N := \{1, \ldots, n\}$ mit $N_1 \cap N_2 = \emptyset$, $N_1 \cup N_2 = N$ sowie $a_{i,j} = 0$ für alle $(i, j) \in N_1 \times N_2$ gibt. Wählt man im obigen Beispiel $N_1 := \{1, 3, 5\}$ und $N_2 := \{2, 4\}$, so erkennt man, dass A zerfällt. Indem man die ungeraden Gleichungen und die Variablen mit einem ungeraden Index zuerst, danach die geraden nimmt, erkennt man, dass ein lineares Gleichungssystem mit der zugehörigen Koeffizientenmatrix in ein 3×3- und ein 2×2-Gleichungssystem zerfällt. □

Abgesehen von der Schwierigkeit, bei einem linearen Gleichungssystem mit einer dünn besetzten Koeffizientenmatrix die Gleichungen und Unbekannten "richtig" zu numerieren, gibt es etliche weitere Schwierigkeiten, die das "dünne" Gaußsche Eliminationsverfahren wesentlich komplexer machen als das dichte Gegenstück. Zunächst braucht man geeignete Datenstrukturen, um die von Null verschiedenen Elemente von A lokalisieren zu können. Hierauf können wir nicht eingehen, sondern verweisen auf I. S. Duff et al. . (1986)[8]. Bei J. W.Demmel (1996, S. 91) wird auf verfügbare Software zur Lösung dünn besetzter linearer Gleichungssysteme mit direkten Verfahren hingewiesen.

[8] I. S. Duff, A. M. Erisman and J. K. Reid (1986) *Direct Methods for Sparse Matrices.* Clarendon Press, Oxford.

2.3.2 Die Cholesky-Zerlegung einer symmetrischen, positiv definiten Matrix

Bei N. J. Higham (1996, S. 204) steht so schön:

- *Symmetric positive definiteness is one of the highest accolades to which a matrix can aspire. Symmetry confers major advantages and simplifications in the eigenproblem and, ..., positive definiteness permits economy and numerical stability in the solution of linear systems.*

Die Lösung linearer Gleichungssysteme mit einer symmetrischen, positiv definiten Koeffizientenmatrix ist das Thema dieses Unterabschnitts. Hierbei heißt eine symmetrische Matrix $A \in \mathbb{R}^{n\times n}$ bekanntlich[9] positiv definit, wenn $x^T Ax > 0$ für alle $x \in \mathbb{R}^n \setminus \{0\}$. Äquivalente Bedingungen sind bekanntlich, dass alle Eigenwerte von A positiv bzw. alle Hauptabschnittsdeterminanten positiv sind. Der folgende Satz sagt aus, dass jede symmetrische, positiv definite Matrix eine eindeutige sogenannte *Cholesky-Zerlegung* besitzt.

Satz 3.2 *Ist $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, so existiert genau eine untere Dreiecksmatrix $L \in \mathbb{R}^{n\times n}$ mit positiven Diagonalelementen derart, dass $A = LL^T$ (Cholesky-Zerlegung).*

Beweis: Die Behauptung wird durch vollständige Induktion nach n bewiesen. Für $n = 1$ ist die Aussage offenbar richtig. Wir nehmen an, dass jede symmetrische, positiv definite $(n-1)\times(n-1)$-Matrix eine eindeutige Cholesky-Zerlegung besitzt. Die $n \times n$-Matrix denke man sich zerlegt:

$$A = \begin{pmatrix} A_{n-1} & a \\ a^T & \alpha \end{pmatrix}.$$

Hierbei ist $A_{n-1} \in \mathbb{R}^{(n-1)\times(n-1)}$ symmetrisch und positiv definit. Für die untere Dreiecksmatrix L mache man den Ansatz

$$L = \begin{pmatrix} L_{n-1} & 0 \\ l^T & \beta \end{pmatrix}$$

mit einer unteren Dreiecksmatrix $L_{n-1} \in \mathbb{R}^{(n-1)\times(n-1)}$. Dann ist

$$LL^T = \begin{pmatrix} L_{n-1} & 0 \\ l^T & \beta \end{pmatrix} \begin{pmatrix} L_{n-1}^T & l \\ 0^T & \beta \end{pmatrix} = \begin{pmatrix} L_{n-1}L_{n-1}^T & L_{n-1}l \\ (L_{n-1}l)^T & l^T l + \beta^2 \end{pmatrix} = \begin{pmatrix} A_{n-1} & a \\ a^T & \alpha \end{pmatrix} = A$$

genau dann, wenn

$$L_{n-1}L_{n-1}^T = A_{n-1}, \qquad L_{n-1}l = a, \qquad l^T l + \beta^2 = \alpha.$$

Also ist L_{n-1} notwendigerweise der nach Induktionsannahme eindeutig existierende Cholesky-Faktor (untere Dreiecksmatrix mit positiven Diagonalelementen) von A_{n-1}

[9] Wir beschränken uns weiter auf reelle Matrizen, da die Unterschiede zum komplexen Fall marginal sind.

und es bleibt die eindeutige Existenz eines Vektors $l \in \mathbb{R}^{n-1}$ und einer positiven Zahl β mit

$$L_{n-1}l = a, \qquad l^T l + \beta^2 = \alpha$$

zu zeigen. Da L_{n-1} nichtsingulär ist, ist l durch $L_{n-1}l = a$ eindeutig festgelegt. Zu zeigen bleibt, dass $\alpha - l^T l$ positiv ist, da dann genau ein positives β mit $l^T l + \beta^2 = \alpha$ existiert. Nun ist aber

$$\begin{aligned} \alpha - l^T l &= \alpha - a^T L_{n-1}^{-T} L_{n-1}^{-1} a \\ &= \alpha - a^T A_{n-1}^{-1} a \\ &= \begin{pmatrix} -A_{n-1}^{-1}a \\ 1 \end{pmatrix}^T \begin{pmatrix} A_{n-1} & a \\ a^T & \alpha \end{pmatrix} \begin{pmatrix} -A_{n-1}^{-1}a \\ 1 \end{pmatrix} \\ &> 0 \quad , \end{aligned}$$

womit alles gezeigt ist. □

Die Darstellung einer (symmetrischen und) positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$ in der Form $A = LL^T$ mit einer unteren Dreiecksmatrix L, deren Diagonalelemente positiv sind, heißt *Cholesky-Zerlegung* von A. Ein Verfahren zur Berechnung von L erhält man sehr leicht aus der Bestimmungsgleichung $A = LL^T$ durch Koeffizientenvergleich. Wegen der Symmetrie genügt es, die unteren Hälften zu vergleichen. Für $i \geq j$ ist

$$a_{ij} = (A)_{ij} = (LL^T)_{ij} = \sum_{k=1}^{j} l_{ik}l_{jk} = \sum_{k=1}^{j-1} l_{ik}l_{jk} + l_{ij}l_{jj}.$$

Dies ergibt:

$$\begin{aligned} i = j: \quad & a_{jj} = l_{jj}^2 + \sum_{k=1}^{j-1} l_{jk}^2 \qquad \text{bzw.} \qquad l_{jj} := \Big(a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2\Big)^{1/2}, \\ i > j: \quad & a_{ij} = l_{ij}l_{jj} + \sum_{k=1}^{j-1} l_{ik}l_{jk} \qquad \text{bzw.} \qquad l_{ij} := \Big(a_{ij} - \sum_{k=1}^{j-1} l_{ik}l_{jk}\Big)\Big/ l_{jj}. \end{aligned}$$

Wegen dieser Gleichungen kann man sukzessive spalten- oder zeilenweise die Matrix L berechnen. So lautet z. B. eine Spaltenversion des resultierenden *Cholesky-Verfahrens*:

- Input: Gegeben sei die symmetrische, positiv definite Matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$. Benutzt wird nur die untere Hälfte von A, d. h. Elemente a_{ij} mit $i \geq j$.
- Für $j = 1, \ldots, n$:
  $a_{jj} := (a_{jj} - \sum_{k=1}^{j-1} a_{jk}^2)^{1/2}$.
  Für $i = j+1, \ldots, n$:
  $a_{ij} := (a_{ij} - \sum_{k=1}^{j-1} a_{ik}a_{jk})/a_{jj}$.
- Output: Die untere Hälfte von A (einschließlich der Diagonalen) ist mit der gesuchten unteren Dreiecksmatrix L überschrieben.

Bevor das Pivotelement durch Wurzelziehen gebildet wird, prüft man natürlich, ob $a_{kk} - \sum_{j=1}^{k-1} a_{kj}^2 \geq \epsilon$ ist, wobei $\epsilon > 0$ eine vorgegebene kleine Zahl ist. Zur Berechnung von L benötigt man

$$\sum_{k=1}^{n} \sum_{i=k+1}^{n} \sum_{j=1}^{k-1} 1 = \sum_{k=1}^{n} (n-k)(k-1) = \tfrac{1}{6} n^3 + \cdots$$

flops, ferner müssen n Wurzeln gezogen werden.

Man kann die Matrix L auch zeilenweise bestimmen, was den Vorteil hat, dass man schneller erkennt, ob die gegebene Matrix nicht (numerisch) positiv definit ist. In Pseudocode könnte dies folgendermaßen aussehen:

- Input: Gegeben sei die symmetrische, positiv definite Matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$. Benutzt wird nur die untere Hälfte von A, d. h. Elemente a_{ij} mit $i \geq j$.
- Für $i = 1, \ldots, n$:

  Für $j = 1, \ldots, i$:

  $s := a_{ji} - \sum_{k=1}^{j-1} a_{jk} a_{ik}$.

  $$a_{ij} := \begin{cases} s/a_{jj}, & j < i, \\ \sqrt{s}, & j = i. \end{cases}$$
- Output: Die untere Hälfte von A (einschließlich der Diagonalen) ist mit der gesuchten unteren Dreiecksmatrix L überschrieben.

Will man prüfen[10] , ob eine gegebene (symmetrische) Matrix A (numerisch) positiv definit ist, so wendet man (spalten- oder zeilenweise) das Cholesky-Verfahren an und überprüft, ob der Radikand größer oder gleich einem kleinen ϵ ist. Angenommen, im k-ten Schritt wird festgestellt, dass $t_k := a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2 < \epsilon$. Wir stellen uns die Frage:

- Welche (möglichst kleine) positive Zahl muss man mindestens zu den Diagonalelementen von A hinzuaddieren, um eine positiv definite Matrix zu erhalten?

Wir definieren die Matrix

$$\hat{L} := \left(\begin{array}{cccc|ccc} l_{11} & & & & 0 & \cdots & 0 \\ l_{21} & l_{22} & & & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots & \ddots & \vdots \\ l_{k-1,1} & l_{k-1,2} & \cdots & l_{k-1,k-1} & 0 & \cdots & 0 \\ \hline l_{k1} & l_{k2} & \cdots & l_{k,k-1} & 0 & \cdots & 0 \\ \hline 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{array}\right) =: \left(\begin{array}{c|c} L_{k-1} & 0 \\ \hline l_k^T & 0 \\ \hline 0 & 0 \end{array}\right).$$

[10] Bei N. J. Higham (1996, S. 225) liest man: An excellent way to test whether a given symmetric matrix A is positive definite is to attempt to compute a Cholesky factorization. This test is less expensive than computing the eigenvalues and is numerically stable.

Zumindestens dann, wenn man die Cholesky-Zerlegung zeilenweise berechnet, haben wir in die Matrix $\hat{L}$ alle bisher berechneten Einträge hineingesteckt. Ferner ist

$$\hat{L}\hat{L}^T = \begin{pmatrix} L_{k-1}L_{k-1}^T & L_{k-1}l_k & 0 \\ (L_{k-1}l_k)^T & \|l_k\|_2^2 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

und

$$(L_{k-1}L_{k-1}^T)_{ij} = a_{ij} \quad (1 \le i,j \le k-1), \qquad L_{k-1}l_k = \begin{pmatrix} a_{1k} \\ \vdots \\ a_{k-1,k} \end{pmatrix}, \qquad a_{kk} - \|l_k\|_2^2 = t_k.$$

Man definiere $z \in \mathbb{R}^n$ durch

$$z := \begin{pmatrix} u \\ 1 \\ 0 \end{pmatrix} \qquad \text{mit} \quad L_{k-1}^T u = -l_k.$$

Da die Diagonalelemente von L_{k-1}^T alle $\ge \epsilon$ sind, gibt es bei der Berechnung von u keine Schwierigkeiten. Denkt man sich nun A partitioniert durch

$$A = \begin{pmatrix} A_{k-1} & a_k & * \\ a_k^T & a_{kk} & * \\ * & * & * \end{pmatrix},$$

so erkennt man, dass

$$\begin{aligned} z^T A z &= u^T A_{k-1} u + 2a_k^T u + a_{kk} \\ &= u^T L_{k-1} L_{k-1}^T u + 2(L_{k-1}l_k)^T u + t_k + \|l_k\|_2^2 \\ &= t_k. \end{aligned}$$

Folglich ist

$$z^T \Big[A + \Big(\epsilon - \frac{t_k}{\|z\|_2^2} \Big) I \Big] z = \epsilon \, \|z\|_2^2.$$

Addiert man daher $\epsilon - t_k/\|z\|_2^2$ (dies ist eine positive Zahl, denn $t_k < \epsilon$ und $\|z\|_2 \ge 1$) zu allen Diagonalelementen von A, so ist eine notwendige Bedingung dafür erfüllt, dass alle Eigenwerte der so abgeänderten Matrix $\ge \epsilon$ sind.

Eine ausführliche Fehleranalyse zur Cholesky-Zerlegung findet man bei N. J. Higham (1996, S. 203 ff.), hierauf wollen wir nicht eingehen. Ein Indiz für die Stabilität der Cholesky-Zerlegung $A = LL^T$ einer symmetrischen, positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$ liest man daraus ab, dass $|l_{ij}| \le \sqrt{a_{ii}}$ für $i \ge j$, so dass die Einträge der unteren Dreiecksmatrix L auf einfache Weise durch die der Ausgangsmatrix A beschränkt sind.

2.3.3 Die LDL^T-Zerlegung einer symmetrischen Matrix

Einige wenige Bemerkungen sollen noch zur Zerlegung symmetrischer, aber nicht notwendig definiter Matrizen gemacht werden. Ziel ist es stets, die Symmetrie so auszunutzen, dass mit halbem Aufwand verglichen mit dem Gaußschen Eliminationsverfahren eine (symmetrische) LU-Zerlegung berechnet werden kann.

Ist $A \in \mathbb{R}^{n\times n}$ eine Matrix, deren Hauptabschnittsdeterminanten von Null verschieden sind (die also insbesondere selbst nichtsingulär ist), so existiert eine eindeutige Zerlegung $A = LU$ mit einer unteren Dreiecksmatrix L mit diag $(L) = I$ (also Einsen auf der Diagonalen) und einer oberen Dreiecksmatrix U. Definiert man $D := \text{diag}\,(U)$ und $M := U^T D^{-1}$, so ist diag $(M) = I$ und $LDM^T = LU = A$. Ist A darüberhinaus symmetrisch, so ist $M = L$ und wir haben die eindeutige Existenz einer LDL^T-Zerlegung $A = LDL^T$ einer symmetrischen Matrix, deren Hauptabschnittsdeterminanten sämtlich von Null verschieden sind. Hierbei bedeutet L natürlich eine untere Dreiecksmatrix mit Einsen auf der Diagonalen und D eine nichtsinguläre Diagonalmatrix. Interessant ist, dass man mit Hilfe einer LDL^T-Zerlegung von A (wenn sie denn existiert) die *Trägheit* von A bestimmen kann, also das Tripel (i_+, i_-, i_0), wobei i_+ die Anzahl positiver, i_- die Anzahl negativer und i_0 die Anzahl verschwindender Eigenwerte von A bedeutet. Grundlage hierfür ist Sylvesters Trägheitssatz (siehe z. B. J. W. Demmel (1997, S. 202)):

Satz 3.3 *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und $X \in \mathbb{R}^{n\times n}$ nichtsingulär. Dann haben A und X^TAX dieselbe Trägheit.*

Beweis: Wir nehmen an, A habe i_- negative Eigenwerte, aber X^TAX habe $i'_- < i_-$ negative Eigenwerte. Sei N der i_--dimensionale Eigenraum von A, der von (orthonormalen) Eigenvektoren zu den i_- negativen Eigenwerten von A aufgespannt ist. Für jedes $x \in N\setminus\{0\}$ ist daher $x^TAx < 0$. Ferner sei P der $(n-i'_-)$-dimensionale Eigenraum von X^TAX, der von den $n - i'_-$ (orthonormalen) Eigenvektoren zu den nichtnegativen Eigenwerten von X^TAX aufgespannt ist. Dann ist $x^TX^TAXx \geq 0$ für alle $x \in P$. Da X nichtsingulär ist, ist $n - i'_- = \dim(P) = \dim(XP)$. Wegen der Dimensionsformel für Teilräume endlichdimensionaler Vektorräume[11] ist

$$\begin{aligned}\dim(N \cap XP) &= \dim(N) + \dim(XP) - \dim(N + XP)\\ &= i_- + n - i'_- - \dim(N + XP)\\ &\geq i_- - i'_-\\ &\geq 1.\end{aligned}$$

Daher existiert ein $x \in N \cap XP \setminus \{0\}$, woraus einerseits $x^TAx < 0$ wegen $x \in N \setminus \{0\}$, andererseits $x^TAx \geq 0$ wegen $x \in XP$ folgt. Dies ist ein Widerspruch und wir haben $i'_- \geq i_-$ bewiesen. Indem man die Rollen von A und X^TAX vertauscht, erhält man auch $i'_- \leq i_-$, insgesamt also $i'_- = i_-$. Entsprechend ist auch die Anzahl positiver Eigenwerte von A und X^TAX gleich, so dass auch die Anzahl verschwindender Eigenwerte von A und X^TAX übereinstimmen. Der Sylvestersche Trägheitssatz ist damit bewiesen. □

[11] Siehe z. B. M. Koecher (1983, S. 50) *Lineare Algebra und analytische Geometrie.* Springer, Berlin.

In Pseudocode könnte ein Verfahren zur Berechnung der LDL^T-Zerlegung einer symmetrischen Matrix $A \in \mathbb{R}^{n\times n}$ (wenn sie existiert) folgendermaßen aussehen:

- Input: Gegeben sei eine symmetrische Matrix $A \in \mathbb{R}^{n\times n}$, von der aber nur die untere Hälfte benutzt (und überschrieben) wird.
- Für $j = 1, \ldots, n$:

  Für $k = 1, \ldots, j-1$:

  $r_k := a_{jk}a_{kk}$.

  $d_j := a_{jj} - \sum_{k=1}^{j-1} a_{jk}r_k$.

  $a_{jj} := d_j$.

  Falls $d_j = 0$, dann STOP, Verfahren nicht durchführbar.

  Andernfalls:

  Für $i = j+1, \ldots, n$:

  $a_{ij} := (a_{ij} - \sum_{k=1}^{j-1} a_{ik}r_k)/d_j$.
- Output: Die strikte untere Hälfte von A wird mit L (ausgenommen der Diagonalen) überschrieben, in der Diagonalen von A stehen (bei erfolgreichem Ausgang) die Diagonalelemente von D.

Die Berechnung der Einträge von L erfolgt spaltenweise, die Diagonalelemente von D werden in den Diagonalelementen von A gespeichert. Angenommen, $d_1, \ldots, d_{j-1}$ seien schon berechnet und nach $a_{11}, \ldots, a_{j-1,j-1}$ gespeichert, ferner seien auch die ersten $j-1$ Spalten von L berechnet. Aus

$$a_{jj} = (A)_{jj} = (LDL^T)_{jj} = d_j + \sum_{k=1}^{j-1} l_{jk}d_k l_{jk} = d_j + \sum_{k=1}^{j-1} a_{jk}\underbrace{a_{jk}a_{kk}}_{=r_k}$$

erhält man zunächst

$$d_j := a_{jj} - \sum_{k=1}^{j-1} a_{jk}r_k.$$

Danach berechnet man die j-te Spalte von L so, dass für $i > j$ gilt

$$a_{ij} = (A)_{ij} = (LDL^T)_{ij} = l_{ij}d_j + \sum_{k=1}^{j-1} l_{ik}d_k l_{jk} = l_{ij}d_j + \sum_{k=1}^{j-1} a_{ik}\underbrace{a_{kk}a_{jk}}_{=r_k},$$

was auf

$$l_{ij} := \Big(a_{ij} - \sum_{k=1}^{j-1} a_{ik}r_k\Big) \Big/ d_j$$

führt. Das oben in Pseudocode angegebene Verfahren tut also das Verlangte.

Dieser Algorithmus ist nun leider nur für positiv definites A ohne Einschränkungen zu empfehlen. Denn z. B. ist

$$\begin{pmatrix} \epsilon & 1 \\ 1 & \epsilon \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1/\epsilon & 1 \end{pmatrix} \begin{pmatrix} \epsilon & 0 \\ 0 & \epsilon - 1/\epsilon \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1/\epsilon & 1 \end{pmatrix}^T.$$

Für $0 < \epsilon \ll 1$ werden Einträge in der Zerlegung beliebig groß. Z. B. bei N. J. Higham (1996, S. 218) werden stabile Methoden zur Berechnung einer LDL^T-Zerlegung einer symmetrischen Matrix A geschildert. Die Idee besteht darin, eine Block-LDL^T-Zerlegung der Form

$$PAP^T = LDL^T$$

zu berechnen, wobei P eine Permutationsmatrix, L eine untere Dreiecksmatrix mit Einsen in der Diagonalen und D eine Block-Diagonalmatrix mit 1×1 oder 2×2 Diagonalblöcken ist. Dass zu so einer Darstellung 2×2-Blöcke nötig sind, erkennt man an der Matrix

$$A := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Denn einerseits ist $PAP^T = A$ für "alle" 2×2-Permutationsmatrizen, andererseits besitzt A keine (reine) LU-Zerlegung, so dass keine Darstellung $PAP^T = LDL^T$ mit einer 2×2-Diagonalmatrix D möglich ist.

Zu Beginn der Rechnung wähle man eine Permutationsmatrix Π und ein $s \in \{1, 2\}$ derart, dass

$$\Pi A \Pi^T = \begin{pmatrix} E & C^T \\ C & B \end{pmatrix}$$

mit einem nichtsingulären $E \in \mathbb{R}^{s \times s}$, dem Pivotblock. Wenn A ungleich der Null-Matrix ist, so ist dies möglich. Wegen der Symmetrie von A sind natürlich auch E und $B \in \mathbb{R}^{(n-s)\times(n-s)}$ symmetrisch. Es gilt dann die Faktorisierung

$$\Pi A \Pi^T = \begin{pmatrix} I_s & 0 \\ CE^{-1} & I_{n-s} \end{pmatrix} \begin{pmatrix} E & 0 \\ 0 & B - CE^{-1}C^T \end{pmatrix} \begin{pmatrix} I_s & 0 \\ CE^{-1} & I_{n-s} \end{pmatrix}^T.$$

Anschließend wird derselbe Prozeß auf das *Schur-Komplement*

$$\tilde{A} := B - CE^{-1}C^T$$

angewandt. Es kommt hier sehr darauf an, wie die Pivotisierung vorgenommen wird. Auf Einzelheiten wollen wir aber nicht mehr eingehen.

2.3.4 Bandmatrizen

Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt eine *Bandmatrix* mit *unterer Bandbreite* p und *oberer Bandbreite* q, wenn $a_{ij} = 0$ für $i > j + p$ und $j > i + q$, also A die Form

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1,q+1} & & & 0 \\ \vdots & & & a_{2,q+2} & & \\ a_{p+1,1} & & & & \ddots & \\ & a_{p+2,2} & & & & a_{n-q,n} \\ & & \ddots & & & \vdots \\ 0 & & & a_{n,n-p} & \cdots & a_{nn} \end{pmatrix}$$

besitzt. Ist hier $p = q = 1$, so nennt man A bekanntlich eine *Tridiagonalmatrix*. Die Dreiecksfaktoren in einer LU-Zerlegung einer Bandmatrix sind ebenfalls Bandmatrizen, wie der folgende Satz aussagt (siehe G. H. Golub, C. F. Van Loan (1996, Theorem 4.3.1).

Satz 3.4 *Die Matrix $A \in \mathbb{R}^{n\times n}$ besitze eine LU-Zerlegung $A = LU$ mit einer unteren Dreiecksmatrix L mit Einsen auf der Diagonalen und einer oberen Dreiecksmatrix U. Ist A eine Bandmatrix mit unterer Bandbreite p und oberer Bandbreite q, so hat L untere Bandbreite p und U obere Bandbreite q.*

Beweis: Der Beweis erfolgt durch vollständige Induktion nach n. Der Induktionsanfang liegt bei $n = \min(p, q) + 1$. Hier ist gar nichts zu zeigen. Angenommen, die Aussage sei für $(n-1)\times(n-1)$-Bandmatrizen mit unterer Bandbreite p und oberer Bandbreite q richtig. Sei also $A \in \mathbb{R}^{n\times n}$ eine (p, q)-Bandmatrix, die eine LU-Zerlegung $A = LU$ besitzt. Schreibt man

$$A = \begin{pmatrix} \alpha & w^T \\ v & B \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ v/\alpha & I_{n-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & B - vw^T/\alpha \end{pmatrix} \begin{pmatrix} \alpha & w^T \\ 0 & I_{n-1} \end{pmatrix},$$

so ist es klar, dass $B - vw^T/\alpha$ eine $(n-1)\times(n-1)$-Bandmatrix mit unterer Bandbreite p und oberer Bandbreite q ist. Denn nur die ersten p Komponenten von v und die ersten q Komponenten von w können nicht verschwinden. Ferner besitzt $B - vw^T/\alpha$ eine LU-Zerlegung $B - vw^T/\alpha = L_1U_1$, wobei

$$L = \begin{pmatrix} 1 & 0 \\ * & L_1 \end{pmatrix}, \qquad U = \begin{pmatrix} * & * \\ 0 & U_1 \end{pmatrix}.$$

Nach Induktionsannahme ist L_1 eine untere Dreiecksmatrix der unteren Bandbreite p und U_1 eine obere Dreiecksmatrix der oberen Bandbreite q. Daher ist

$$A = \begin{pmatrix} 1 & 0 \\ v/\alpha & L_1 \end{pmatrix} \begin{pmatrix} \alpha & w^T \\ 0 & U_1 \end{pmatrix},$$

wobei die beiden Faktoren die geforderte Bandstruktur haben. □

In Pseudocode könnte das Verfahren, das die LU-Zerlegung einer Bandmatrix (wenn sie existiert) berechnet, folgendermaßen aussehen:

- Input: Gegeben sei eine Bandmatrix $A \in \mathbb{R}^{n\times n}$ mit unterer Bandbreite p und oberer Bandbreite q. Es wird vorausgesetzt, dass A eine LU-Zerlegung $A = LU$ besitzt.

- Für $k = 1, \ldots, n-1$:

  Für $i = k+1, \ldots, \min(k+p, n)$:

  $a_{ik} := a_{ik}/a_{kk}$.

  Für $i = k+1, \ldots, \min(k+p, n)$:

  Für $j = k+1, \ldots, \min(k+q, n)$:

  $a_{ij} := a_{ij} - a_{ik}a_{kj}$.

- Output: Es wird die LU-Zerlegung $A = LU$ von A berechnet, wobei a_{ij} mit l_{ij} für $i > j$ und mit u_{ij} für $i \leq j$ überschrieben wird.

Es ist völlig klar, wie sich das Vorwärts- bzw. Rückwärtseinsetzen vereinfacht, wenn die untere Dreiecksmatrix L bzw. die obere Dreiecksmatrix R Bandstruktur besitzt.

Nicht ganz so einfach ist es, die gegebene Bandstruktur der Matrix A richtig auszunutzen, wenn man das Gaußsche Eliminationsverfahren mit Spaltenpivotsuche anwendet. Wir formulieren das Ergebnis als Satz (siehe G. H. Golub, C. F. Van Loan (1996, Theorem 4.3.2).

Satz 3.5 *Sei $A \in \mathbb{R}^{n\times n}$ eine nichtsinguläre Bandmatrix mit unterer Bandbreite p und oberer Bandbreite q. Es werde das Gaußsche Eliminationsverfahren mit Spaltenpivotsuche angewandt, also Permutationsmatrizen (genauer: Vertauschungsmatrizen) $P_1, \ldots, P_{n-1}$ und Gauß-Matrizen $M_1, \ldots, M_{n-1}$ mit $M_k = I - l_k e_k^T$, $k = 1, \ldots, n-1$, berechnet derart, dass*

$$M_{n-1}P_{n-1}\cdots M_1P_1A = U$$

eine obere Dreiecksmatrix ist. Dann hat R obere Bandbreite $p+q$ und es ist $(l_k)_i = 0$ für $i \leq k$ (das ist sowieso klar) und $i > k+p$, $k = 1, \ldots, n-1$. Daher hat l_k höchstens p von Null verschiedene Komponenten und die untere Dreiecksmatrix L in jeder Spalte höchstens $p+1$ von Null verschiedene Einträge.

Beweis: Sei $PA = LU$ die durch das Gaußsche Eliminationsverfahren mit Spaltenpivotsuche berechnete Zerlegung. Man erinnere sich daran, dass $P = P_{n-1}\cdots P_1$. Sei

$$P = (\ e_{s_1} \ \cdots \ e_{s_n}\ )^T,$$

wobei $\{s_1, \ldots, s_n\}$ eine Permutation von $\{1, \ldots, n\}$ ist. Angenommen, für ein $i \in \{1, \ldots, n\}$ wäre $s_i > i + p$. Dann wäre

$$(PA)_{ij} = a_{s_i,j} = 0, \qquad j = 1, \ldots, i \leq s_i - p - 1,$$

also die $i \times i$-Hauptuntermatrix von PA singulär, weil sie eine Nullzeile enthält. Dies wiederum impliziert, dass der entsprechende $i\times i$-Block in R und damit auch A singulär ist, ein Widerspruch. Also ist $s_i \leq i + p$, $i = 1, \ldots, n$. Da A die obere Bandbreite q besitzt, ist $a_{s_i,j} = 0$ für $j > s_i + q$ und damit $(PA)_{ij} = 0$ für $j > i + p + q$. Daher

hat PA die obere Bandbreite $p+q$, ferner besitzt PA nach Konstruktion eine LU-Zerlegung. Aus Satz 3.4 folgt, dass U die obere Bandbreite $p+q$ besitzt. Angenommen, $M_{k-1}P_{k-1}\cdots M_1P_1A$ sei schon berechnet. Im unteren $(n-k+1)\times(n-k+1)$-Block stehen in permutierter Reihenfolge die $k,\ldots,n$-Komponenten von Zeilen der Ausgangsmatrix A. In der ersten Spalte dieses Blockes können von Null verschiedene Komponenten nur in den Positionen $(k+1,k),\ldots,(k+p,k)$ auftreten, womit auch die letzte Behauptung bewiesen ist. □

Besonders einfach ist der Spezialfall einer *oberen Hessenberg-Matrix*. Dies ist eine Matrix mit unterer Bandbreite $p=1$, d. h. alle Einträge unterhalb der Subdiagonalen verschwinden. Die Vertauschungsmatrizen P_k, $k=1,\ldots,n-1$, im Gaußschen Eliminationsverfahren mit Spaltenpivotsuche sind entweder die Identität oder die "Identität", bei der die Zeilen k und $k+1$ vertauscht sind, die Gauß-Matrizen M_k, $k=1,\ldots,n-1$, sind in der k-ten Spalte außerhalb der Diagonalen nur in der Position $(k+1,k)$ mit einem i. Allg. von Null verschiedenen Element besetzt. Bei symmetrischen Bandmatrizen (dann stimmen natürlich untere und obere Bandbreite überein) vereinfacht sich natürlich auch die Berechnung einer LDL^T-Zerlegung, bei symmetrischen und positiv definiten Bandmatrizen die einer Cholesky-Zerlegung. Bei kleiner Bandbreite (insbesondere bei Tridiagonalmatrizen) einer symmetrischen, positiv definiten Matrix ist die Berechnung der (dann natürlich eindeutig existierenden) LDL^T einer Cholesky-Zerlegung vorzuziehen, da die Berechnung von n Quadratwurzeln verglichen mit dem Gesamtaufwand unverhältnismäßig teuer ist.

2.3.5 MATLAB-Ergänzungen

Zur Berechnung einer LU-Zerlegung einer quadratischen Matrix steht in MATLAB die Funktion `lu` zur Verfügung.

```
 LU     LU factorization.
    [L,U] = LU(X) stores an upper triangular matrix in U and a
    "psychologically lower triangular matrix" (i.e. a product
    of lower triangular and permutation matrices) in L, so
    that X = L*U.  X must be square.

    [L,U,P] = LU(X) returns lower triangular matrix L, upper
    triangular matrix U, and permutation matrix P so that
    P*X = L*U.
```

Die Funktion `lu` ist eine sogenannte *built-in function*, man kann sie sich also nicht ansehen. Man kann aber natürlich eine entsprechende Funktion auch selbst in MATLAB schreiben. Eine Möglichkeit ist die folgende Funktion, die wir C. F. Van Loan (1997, S. 214) entnommen haben:

```
function  [L,U,piv]=GEpiv(A);
%************************************************************
%Pre:
%  A    n-by-n matrix
```

```
%Post:
%
%  L    n-by-n unit lower triangular matrix
%  U    n-by-n upper triangular matrix
%  piv  integer n-vector that is a permutation of 1:n
%       A(piv,:)=LU
%*************************************************************
   [n,n]=size(A);
   piv=1:n;
   for k=1:n-1
       [maxv,q]=max(abs(A(k:n,k)));
       r=q+k-1;
       piv([k r])=piv([r k]);
       A([k r],:)=A([r k],:);
       if A(k,k)~=0
          A(k+1:n,k)=A(k+1:n,k)/A(k,k);
          A(k+1:n,k+1:n)=A(k+1:n,k+1:n)-A(k+1:n,k)*A(k,k+1:n);
       end
   end
   L=eye(n,n)+tril(A,-1);
   U=triu(A);
```

Um mit Hilfe einer LU-Zerlegung ein lineares Gleichungssystem zu lösen, benötigt man noch Funktionen zum Vorwärts- und Rückwärtseinsetzen. Bei C. F. Van Loan (1997, S. 196–197) findet man die Funktionen `LTriSol` und `UTriSol`, von denen wir nur letztere angeben:

```
function x=UTriSol(U,b);
%*********************************************************
% Pre:
%   U    n-by-n nonsingular upper triangular matrix
%   b    n-by-1
% Post:
%   x    Ux=b
%*********************************************************
   n=length(b);
   x=zeros(n,1);
   x(n)=b(n)/U(n,n);
   for i=n-1:-1:1
      x(i)=(b(i)-U(i,i+1:n)*x(i+1:n))/U(i,i);
   end;
```

Die Maschinengenauigkeit u (kleinste positive Zahl mit $1 + u \neq 1$) wird häufig durch das folgende Programm approximiert:

```
u=1;
while 1+u~=1
```

```
   u=u/2;
end;
u=2*u;
```

Als Resultat erhalten wir (mit `format long`) `u=2.220446049250313e-16`, was genau mit dem Resultat der MATLAB-Funktion `eps` übereinstimmt. Durch `help eps` erfahren wir u. a., dass

```
eps, with no arguments, is the distance from 1.0 to the next larger
double  precision number, that is eps with no arguments returns 2^(-52).
```

Natürlich gibt es in MATLAB auch eine Funktion zur Bestimmung der Cholesky-Zerlegung einer symmetrischen, positiv definiten Matrix. Nach `help chol` erhält man u. a. die folgenden Informationen:

```
 chol   Cholesky factorization.
    chol(A) uses only the diagonal and upper triangle of A.
    The lower triangle is assumed to be the (complex conjugate)
    transpose of the upper triangle.  If A is positive definite, then
    R = chol(A) produces an upper triangular R so that R'*R = A.
    If A is not positive definite, an error message is printed.

    L = chol(A,'lower') uses only the diagonal and the lower triangle
    of A to produce a lower triangular L so that L*L' = A.  If
    A is not positive definite, an error message is printed.  When
    A is sparse, this syntax of chol is typically faster.
```

Bei C. F. Van Loan (1997, S. 242 ff.) findet man fünf Implementationen (in MATLAB) zur Berechnung der Cholesky-Zerlegung einer voll besetzten, symmetrischen und positiv definiten Matrix. Eine direkte Umsetzung der zeilenweise Berechnung eines Cholesky-Faktors ist gegeben durch (dies ist die erste Implementation bei C. Van Loan (1997, S. 243)):

```
function L=CholScalar(A);
%*********************************************************
%Pre:    A is a symmetric and positive definit matrix
%Post:   L is a lower triangular matrix so A=L*L'
%*********************************************************
[n,n]=size(A);L=zeros(n,n);
for i=1:n
     for j=1:i
  s=A(j,i);
          for k=1:j-1
  s=s-L(j,k)*L(i,k);
          end;
          if j<i
      L(i,j)=s/L(j,j);
```

```
            else
        L(i,i)=sqrt(s);
            end;
        end;
end;
```

Bei einer effizienten Implementation sollten (im Gegensatz zu der obigen Implementation) möglichst viele Operationen der Form

$$\text{Vektor} + \text{Skalar} \cdot \text{Vektor} \mapsto \text{Vektor},$$

sogenannte *saxpy* (**sc**alar **a** **x** **p**lus **y**) Operationen, oder sogar

$$\text{Vektor} + \text{Matrix} \times \text{Vektor} \mapsto \text{Vektor},$$

sogenannte *gaxpy* (**g**eneral **A** **x** **p**lus **y**) Operationen, vorkommen. Hier ist die bei den Tests bei weitem schnellste Version:

```
function L=CholGax(A);
%***********************************************************
%Pre:   A is a symmetric and positive definite matrix
%Post:  L is lower triangular so A=L*L'
%***********************************************************
[n,n]=size(A);
L=zeros(n,n);
s=zeros(n,1);
for j=1:n
   if j==1
      s(j:n)=A(j:n,j);
   else
      s(j:n)=A(j:n,j)-L(j:n,1:j-1)*L(j,1:j-1)';
   end
   L(j:n,j)=s(j:n)/sqrt(s(j));
end
```

Nun vergleichen wir die Funktionen `CholScalar` und `CholGax`, indem wir `A=hilb(n)` mit $n = 20, 30, 40$ setzen und jeweils 1000 Mal den Cholesky-Faktor von A berechnen. Die dabei verstrichene Zeit messen wir mit `tic` und `toc` und erhalten (die Zeit ist in Sekunden angegeben):

n	`CholScalar`	`CholGax`
20	6.111733	3.413082
30	14.593566	5.306421
40	27.896249	7.457058

In (neueren Versionen von) MATLAB gibt es die Funktion `ldl` zur Berechnung der LDL^T-Zerlegung einer symmetrischen bzw. hermiteschen Matrix. Eine direkte Umsetzung des oben angegebenen Verfahrens ohne Pivotisierung könnte folgendermaßen aussehen:

```
function  [L,D,Info]=LDLT(A);
%Pre:   A is a symmetric matrix, only its lower part being used
%*********************************************************************
%Post:  L is a unit lower triangular matrix, D a diagonal matrix
%       with A=L*D*L' (upon completion).
%       Info=1, if algorithm successful, otherwise Info=0
%*********************************************************************
[n,n]=size(A);d=zeros(n,1);r=zeros(n,1);
for j=1:n
   for k=1:j-1
      r(k)=A(j,k)*A(k,k);
   end;
   d(j)=A(j,j)-A(j,1:j-1)*r(1:j-1);
   A(j,j)=d(j);
   if d(j)==0, Info=0, break, end;
   A(j+1:n,j)=(A(j+1:n,j)-A(j+1:n,1:j-1)*r(1:j-1))/d(j);
end;
L=eye(n,n)+tril(A,-1);
D=diag(d);
Info=1;
```

Die Berechnung der LU-Zerlegung einer Bandmatrix (wenn sie existiert) kann durch die folgende Funktion erfolgen, welche eine direkte Übertragung des in Pseudocode angegebenen Verfahrens bei G. H. Golub, C. F. Van Loan (1996, Algorithm 4.3.1) ist.

```
function [L,U]=GEband(A,p,q);
%*********************************************************************
%Pre:   A is n-by-n banded matrix with lower bandwidth p
%       and upper bandwidth q.
%Post:  L is unit lower triangular n-by-n matrix with bandwidth p
%       U is upper  triangular n-by-n matrix with bandwidth q
%       with A=LU (if LU decomposition exists).
%*********************************************************************
[n,n]=size(A);
for k=1:n-1
    for i=k+1:min([k+p n])
       A(i,k)=A(i,k)/A(k,k);
    end;
    for i=k+1:min([k+p n])
       for j=k+1:min([k+q n])
           A(i,j)=A(i,j)-A(i,k)*A(k,j);
       end;
    end;
end;
L=eye(n,n)+triu(tril(A,-1),-p); U=tril(triu(A),q);
```

Natürlich ist es auch einfach, zugehörige Funktionen zum Vorwärts- und Rückwärtseinsetzen zu schreiben, siehe G. H. Golub, C. F. Van Loan (1996, Algorithm 4.3.2 und 4.3.3).

```
function x=GELband(L,p,b);
%************************************************************************
%Pre     L unit lower triangular matrix having lower bandwidth p.
%        b given n-vector.
%Post:   x solves L*x=b
%************************************************************************
n=size(b);
for i=1:n
   for j=max([i-p 1]):i-1
       b(i)=b(i)-L(i,j)*b(j);
   end;
end;
x=b;
```

und

```
function x=GEUband(U,q,b);
%************************************************************************
%Pre     U  upper triangular matrix having upper bandwidth q.
%        b given n-vector.
%Post:   x solves U*x=b
%************************************************************************
n=size(b);
for j=n:-1:1
   b(j)=b(j)/U(j,j);
   for i=max([1 j-q]):j-1
       b(i)=b(i)-U(i,j)*b(j);
   end;
end;
x=b;
```

sein. Die obigen Algorithmen nehmen an, dass die Bandmatrix A konventionell in einem $n \times n$-Feld gespeichert ist. In der Praxis sollte ein Programm zur Lösung linearer Gleichungssysteme, bei denen die Koeffizientenmatrix Bandstruktur hat, eine Datenstruktur benutzen, welche die vielen Nullen in der Bandmatrix ausnutzt. Hierauf wollen wir nicht mehr eingehen.

2.3.6 Aufgaben

1. Sei $A \in \mathbb{R}^{n\times n}$ (strikt zeilenweise) diagonal dominant, d. h.

$$\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| < |a_{ii}|, \qquad i = 1, \ldots, n.$$

Dann ist A nichtsingulär, ferner besitzt A eine LU-Zerlegung $A = LU$ mit einer unteren Dreiecksmatrix L, die Einsen auf der Diagonalen hat, und einer oberen Dreiecksmatrix U.

2. Sei A spaltenweise diagonal dominant, d. h. es sei

$$\sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{ij}| < |a_{jj}|, \qquad j = 1, \ldots, n.$$

Man zeige:

(a) Die Matrix A ist nichtsingulär.

(b) Die Matrix A besitzt nicht nur eine LU-Zerlegung der Form $A = LU$, sondern mehr noch: Das Gaußsche Eliminationsverfahren mit Spaltenpivotsuche benutzt keine Vertauschungen.

3. Sei $A \in \mathbb{R}^{n \times n}$ spaltenweise (strikt) diagonal dominant, d. h.

$$\sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{ij}| < |a_{jj}|, \qquad j = 1, \ldots, n.$$

Das Gaußsche Eliminationsverfahren mit Spaltenpivotsuche macht nach Aufgabe 2 keine Zeilenvertauschungen und erzeugt mit $A^{(1)} := A$ Matrizen $A^{(2)}, \ldots, A^{(n-1)}$. Man zeige, dass der Wachstums-Faktor ρ_n durch 2 nach oben beschränkt ist, d. h. dass

$$\max_{k \leq i,j \leq n} |a_{ij}^{(k)}| \leq 2 \max_{1 \leq i,j \leq n} |a_{ij}|, \qquad k = 1, \ldots, n-1.$$

4. Sei $A \in \mathbb{R}^{n \times n}$ nichtsingulär und $PA = LU$ eine durch das Gaußsche Eliminationsverfahren mit Spaltenpivotsuche gewonnene LU-Zerlegung von A (dabei P, L und U wie üblich). Man zeige, dass

$$\frac{\|A^{-1}\|_\infty}{2^{n-1}} \leq \|U^{-1}\|_\infty \leq n \, \|A^{-1}\|_\infty.$$

5. Gegeben sei die Matrix

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 3 & 6 & 10 & 15 & 21 \\ 1 & 4 & 10 & 20 & 35 & 56 \\ 1 & 5 & 15 & 35 & 70 & 126 \\ 1 & 6 & 21 & 56 & 126 & 252 \end{pmatrix}.$$

(a) Welche Matrix würden Sie erhalten, wenn Sie die 6×6-Matrix A um eine Zeile und eine Spalte zu einer 7×7-Matrix erweitern müssten?

(b) Mit Hilfe des Gaußschen Eliminationsverfahrens mit Spaltenpivotsuche berechne man eine LU-Zerlegung von A. Anschließend berechne man eine untere Schranke von $\kappa_1(A)$.

6. Sei $A \in \mathbb{R}^{n\times n}$ und $A(\sigma) := A - \sigma I$. Für wie viele Werte von $\sigma \in \mathbb{R}$ höchstens ist das Gaußsche Eliminationsverfahren ohne Spaltenpivotsuche nicht durchführbar bzw. existiert keine LU-Zerlegung von A?

7. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Man zeige:

   (a) Alle Diagonalelemente von A sind positiv.

   (b) Es ist $|a_{ij}| \le \sqrt{a_{ii}a_{jj}}$ für alle $i \neq j$.

   (c) Ein betragsgrößter Eintrag von A liegt auf der Diagonalen.

8. Ist $A = \begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix} \in \mathbb{R}^{n\times n}$, so gilt die Hadamardsche Determinanten-Ungleichung:

$$|\det(A)| \le \prod_{j=1}^{n} \|a_j\|_2.$$

   Hinweis: O. B. d. A. kann man annehmen, dass A nichtsingulär und daher A^TA (symmetrisch und) positivdefinit ist. Man mache eine Cholesky-Zerlegung von A^TA.

9. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Dann ist

$$\|A\|_{\infty,1} := \max_{x\neq 0} \frac{\|Ax\|_1}{\|x\|_\infty} = \max\{x^TAx : \|x\|_\infty = 1\}.$$

   Hinweis: Man mache eine Cholesky-Zerlegung von A.

10. Man berechne die Cholesky-Zerlegung von

$$A := \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 3 & 6 & 10 & 15 & 21 \\ 1 & 4 & 10 & 20 & 35 & 56 \\ 1 & 5 & 15 & 35 & 70 & 126 \\ 1 & 6 & 21 & 56 & 126 & 252 \end{pmatrix}.$$

    Hat man die untere Dreiecksmatrix L mit positiven Diagonalelementen und $A = LL^T$ berechnet, so berechne man anschließend die Cholesky-Zerlegung von $L + L^T - I$.

11. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, ferner $A = LL^T$ die Cholesky-Zerlegung von A. Man zeige, dass $\|L\|_2 = \|A\|_2^{1/2}$.

12. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit, ferner $r := \text{Rang}\,(A)$. Man zeige, dass es eine Permutationsmatrix P gibt derart, dass

$$PAP^T = LL^T, \qquad L = \begin{pmatrix} L_{11} & 0 \\ L_{21} & 0 \end{pmatrix},$$

    wobei $L_{11} \in \mathbb{R}^{r\times r}$ eine obere Dreiecksmatrix mit positiven Diagonalelementen ist.

13. Gegeben sei die symmetrische (indefinite) Matrix

$$A := \begin{pmatrix} 2 & 2 & 3 & 0 & 1 & 2 \\ 2 & 4 & 5 & -1 & 0 & 3 \\ 3 & 5 & 6 & -2 & -3 & 0 \\ 0 & -1 & -2 & 1 & 2 & 3 \\ 1 & 0 & -3 & 2 & 4 & 5 \\ 2 & 3 & 0 & 3 & 5 & 6 \end{pmatrix}.$$

    Man berechne mit oder ohne Pivotsuche eine LDL^T-Zerlegung von A.

14. Die symmetrische, positiv definite Matrix $A \in \mathbb{R}^{n\times n}$ besitze die Bandbreite p. In Pseudocode gebe man ein Verfahren zur Berechnung der Cholesky-Zerlegung von A an.

15. Sei $A \in \mathbb{R}^{n\times n}$ eine Tridiagonalmatrix und (zeilenweise) diagonal dominant. Sei $A = LU$ die (eindeutig existierende) LU-Zerlegung von A. Man gebe in Pseudocode ein Verfahren zur Berechnung von L und U an und zeige, dass $|L|\,|U| \leq 3\,|A|$.

Kapitel 3

Lineare Ausgleichsprobleme

3.1 Problemstellung, Grundlagen

Unter einem *linearen Ausgleichsproblem* (linear least square problem) verstehen wir die Aufgabe, bei gegebenen $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$ eine Lösung von

$$\text{(P)} \qquad \text{Minimiere} \quad \|Ax - b\|_2, \quad x \in \mathbb{R}^n$$

zu bestimmen. Als Literatur zu dieser Aufgabenstellung muss vor allem das vorzügliche Lehrbuch von Å. Björk (1996) genannt werden.

Beispiel: Eine typische Anwendung der linearen Ausgleichsrechnung ist das "curve fitting". Hier sind m Paare $(t_1, b_1), \ldots, (t_m, b_m)$ von reellen Zahlen gegegeben. Bei einem "fit" mit Polynomen wird ein Polynom p vom Grad $\leq n-1$, $p(t) = \sum_{j=1}^{n} x_j t^{j-1}$, gesucht, für das $\sum_{i=1}^{m}(p(t_i) - b_i)^2$ minimal ist. In Abhängigkeit von $x \in \mathbb{R}^n$, des Vektors der Polynomkoeffizienten, ist hier also die euklidische Norm von

$$\begin{pmatrix} p(t_1) \\ p(t_2) \\ \vdots \\ p(t_m) \end{pmatrix} - \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} = \begin{pmatrix} 1 & t_1 & \cdots & t_1^{n-1} \\ 1 & t_2 & \cdots & t_2^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_m & \cdots & t_m^{n-1} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} - \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} = Ax - b$$

zu minimieren. Hierbei ist i. allg. $m \gg n$, da es nicht sinnvoll ist, mit Polynomen eines zu hohen Grades zu arbeiten. Natürlich sind andere Ansatzfunktionen als die Monombasis der Polynome vom Grad $\leq n - 1$ denkbar. □

Im folgenden werden wir immer davon ausgehen, dass die Anzahl n der zu bestimmenden Parameter $x_1, \ldots, x_n$ nicht größer als die Anzahl m der (mit Messfehlern behafteten) Beobachtungen $b_1, \ldots, b_m$ ist. Das lineare Ausgleichsproblem (P) kann man dann auch als die Aufgabe auffassen, das i. allg. überbestimmte lineare Gleichungssystem $Ax = b$ in dem Sinne zu "lösen", dass der Defekt bezüglich der euklidischen Vektornorm minimal wird.

Im folgenden Satz werden einfache Existenz- und Eindeutigkeitsaussagen für das lineare Ausgleichsproblem (P) zusammengestellt. Wir überlassen den einfachen Beweis der Leserin.

Satz 1.1 *Gegeben sei das lineare Ausgleichsproblem*

(P) $$\textit{Minimiere} \quad \|Ax - b\|_2, \quad x \in \mathbb{R}^n.$$

Hierbei sei $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ und $m \geq n$. Dann gilt:

1. *Das lineare Ausgleichsproblem (P) besitzt eine Lösung.*
2. *Ein $x^* \in \mathbb{R}^n$ ist genau dann eine Lösung von (P), wenn x^* eine Lösung der sogenannten Normalgleichungen $A^TAx = A^Tb$ ist.*
3. *Eine Lösung von (P) ist genau dann eindeutig, wenn* Rang $(A) = n$.
4. *Unter allen Lösungen von (P) gibt es genau eine mit minimaler euklidischer Norm.*

Beweis: Siehe Aufgabe 1. □

Zur numerischen Lösung eines linearen Ausgleichsproblems stehen die folgenden direkten Methoden zur Verfügung:

1. Bilde die Normalgleichungen und löse diese (wenn Rang $(A) = n$) mit Hilfe des Cholesky-Verfahrens.

2. Bestimme eine QR-Zerlegung von A, d. h. eine orthogonale Matrix $Q \in \mathbb{R}^{m\times m}$ und eine obere Dreiecksmatrix $R \in \mathbb{R}^{m\times n}$ mit

$$A = QR = Q \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix},$$

   wobei die obere Dreiecksmatrix $\hat{R} \in \mathbb{R}^{n\times n}$ nichtsingulär ist bzw. von Null verschiedene Diagonalelemente besitzt, wenn Rang $(A) = n$. Wegen

$$\|Ax - b\|_2^2 = \|Rx - Q^Tb\|_2^2 = \left\| \begin{pmatrix} \hat{R}x \\ 0 \end{pmatrix} - \begin{pmatrix} c \\ d \end{pmatrix} \right\|_2^2 = \|\hat{R}x - c\|_2^2 + \|d\|_2^2,$$

   wobei wir

$$\begin{pmatrix} c \\ d \end{pmatrix} := Q^Tb$$

   gesetzt haben, erhält man die (unter obiger Rangvoraussetzung) eindeutige Lösung des linearen Ausgleichsproblems durch Rückwärtseinsetzen aus $\hat{R}x = c$. Man beachte, dass es genügen würde, eine Matrix $\hat{Q} \in \mathbb{R}^{m\times n}$ mit $\hat{Q}^T\hat{Q} = I$ (d. h. die n Spalten von $\hat{Q}$ bilden ein Orthonormalsystem im $\mathbb{R}^m$) und eine (nichtsinguläre) obere Dreiecksmatrix $\hat{R} \in \mathbb{R}^{n\times n}$ mit $A = \hat{Q}\hat{R}$ zu bestimmen. Man spricht dann von einer *reduzierten* QR-Zerlegung im Gegensatz zu der oben angegebenen *vollen* QR-Zerlegung.

3. Bestimme eine *Singulärwertzerlegung* von A, d. h. orthogonale Matrizen $U \in \mathbb{R}^{m\times m}$ und $V \in \mathbb{R}^{n\times n}$ sowie eine Diagonalmatrix $\Sigma \in \mathbb{R}^{m\times n}$ mit

$$A = U\Sigma V^T, \qquad \Sigma = \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix},$$

wobei

$$\hat{\Sigma} = \text{diag}\,(\sigma_1, \ldots, \sigma_n), \qquad \sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0.$$

Hier heißen $\sigma_1, \ldots, \sigma_n$ die *singulären Werte* von A, ferner ist $r = \text{Rang}\,(A)$. Mit

$$U = (\ u_1 \ \cdots \ u_m\ ), \qquad V = (\ v_1 \ \cdots \ v_m\ )$$

ist

$$\|Ax - b\|_2^2 = \|\Sigma V^T x - U^T b\|_2^2 = \sum_{i=1}^{r} [\sigma_i (V^T x)_i - u_i^T b]^2 + \sum_{i=r+1}^{m} (u_i^T b)^2$$

für ein beliebiges $x \in \mathbb{R}^n$. Daher ist $x \in \mathbb{R}^n$ genau dann eine Lösung des linearen Ausgleichsproblems, wenn

$$(*) \quad V^T x = (u_1^T b/\sigma_1, \ldots, u_r^T b/\sigma_r, \alpha_{r+1}, \ldots, \alpha_n)^T \quad \text{mit} \quad \alpha_{r+1}, \ldots, \alpha_n \in \mathbb{R}.$$

Dies wiederum liefert: Ist x eine Lösung des linearen Ausgleichsproblems, ist also $(*)$ erfüllt, so ist

$$\|x\|_2^2 = \|V^T x\|_2^2 = \sum_{i=1}^{r} \Big(\frac{u_i^T b}{\sigma_i}\Big)^2 + \sum_{i=r+1}^{n} \alpha_i^2.$$

Die nach Satz 1.1 eindeutige Lösung minimaler euklidischer Norm ist daher durch

$$x_{LS} := V(u_1^T b/\sigma_1, \ldots, u_r^T b/\sigma_r, 0, \ldots, 0)^T = \sum_{i=1}^{r} \frac{u_i^T b}{\sigma_i} v_i$$

gegeben. Oben haben wir die *volle* Singulärwertzerlegung einer Matrix $A \in \mathbb{R}^{m\times n}$ definiert. Eine Darstellung der Form $A = \hat{U}\hat{\Sigma}V^T$ mit einer Matrix $\hat{U} \in \mathbb{R}^{m\times n}$, für die $\hat{U}^T\hat{U} = I$, einer orthogonalen Matrix $V \in \mathbb{R}^{n\times n}$ und einer Diagonalmatrix

$$\hat{\Sigma} = \text{diag}\,(\sigma_1, \ldots, \sigma_r, \underbrace{0, \ldots, 0}_{n-r}) \in \mathbb{R}^{n\times n}, \qquad \sigma_1 \geq \cdots \geq \sigma_r > 0$$

heißt eine *reduzierte* Singulärwertzerlegung. Offensichtlich würde diese genügen, um das lineare Ausgleichsproblem vollständig zu lösen.

Die erste Methode ist die schnellste, aber am wenigsten genaue. Sie ist nur zu empfehlen, wenn Rang $(A) = n$ und die Kondition von A (die Kondition einer $m \times n$-Matrix wird noch zu erklären sein) klein ist. Auch die zweite Methode kann nur angewandt werden, wenn Rang $(A) = n$. Ihr Aufwand ist etwa doppelt so groß wie der der ersten Methode, dafür ist sie wesentlicher zuverlässiger und i. Allg. die "method of choice". Die dritte Methode ist die bei weitem aufwendigste, dafür bei Rang-defizienten Problemen (also Rang $(A) < n$) oder schlecht gestellten Problemen (Kondition von A ist groß) die bei weitem beste Methode. Wir werden auf die letzten beiden Methoden ausführlich eingehen.

3.1.1 MATLAB-Ergänzungen

Als Beispiel für das Auftreten linearer Ausgleichsprobleme wurde das "curve fitting" genannt. In MATLAB gibt es hierzu die Funktion `polyfit`, über die man u. a. die folgenden Informationen erhält:

```
polyfit Fit polynomial to data.
   P = polyfit(X,Y,N) finds the coefficients of a polynomial P(X) of
   degree N that fits the data Y best in a least-squares sense. P is a
   row vector of length N+1 containing the polynomial coefficients in
   descending powers, P(1)*X^N + P(2)*X^(N-1) +...+ P(N)*X + P(N+1).
```

Zum Auswerten von Polynomen steht in MATLAB die Funktion `polyval` zur Verfügung.

```
polyval Evaluate polynomial.
   Y = polyval(P,X) returns the value of a polynomial P evaluated at X. P
   is a vector of length N+1 whose elements are the coefficients of the
   polynomial in descending powers:
        Y = P(1)*X^N + P(2)*X^(N-1) + ... + P(N)*X + P(N+1)
```

Als Beispiel approximieren wir die Fehlerfunktion

$$\operatorname{erf}(t) := \frac{2}{\sqrt{\pi}} \int_0^t e^{-s^2}\, ds$$

auf $[0, 2.5]$ durch ein Polynom vom Grade 2.

```
t=(0:0.1:2.5)';
b=erf(t);
p=polyfit(t,b,2);
z=polyval(p,t);
plot(t,b,'*',t,z);
title('Polynomial fit of error function','FontSize',20);
xlabel('t','FontSize',20);
```

Wir erhalten den in Abbildung 3.1 dargestellten Plot. Hierbei sind zunächst die Punkte $(t_j, \operatorname{erf}(t_j))$ durch $*$ markiert, dann wird die ausgleichende Parabel eingezeichnet.

3.1.2 Aufgaben

1. Man beweise Satz 1.1.

2. Sei $A \in \mathbb{R}^{m\times n}$ mit Rang $(A) = n$, $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n$ gegeben. Dann besitzen die beiden Aufgaben

$$\text{Minimiere} \quad f(x) := \tfrac{1}{2}\|Ax - b\|_2^2 + c^T x, \quad x \in \mathbb{R}^n$$

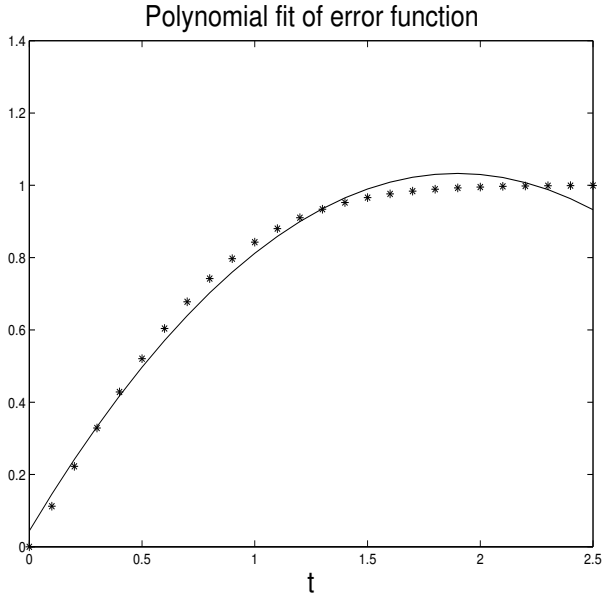


Abbildung 3.1: Eine Ausgleichsparabel zur Fehlerfunktion

und

$$\text{Minimiere} \quad g(y) := \tfrac{1}{2}\|y-b\|_2^2, \quad A^T y = c$$

jeweils genau eine Lösung, die man durch Lösen von

$$\begin{pmatrix} I & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix}$$

erhalten kann.

3. Sei $C \in \mathbb{R}^{m\times m}$ symmetrisch und positiv definit, ferner sei auf dem $\mathbb{R}^m$ durch $\|y\|_C := (y^T C y)^{1/2}$ eine Norm definiert. Mit $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$ betrachte man das (gewichtete) lineare Ausgleichsproblem

   (P) $$\text{Minimiere} \quad \|Ax-b\|_C, \quad x \in \mathbb{R}^n.$$

   Man formuliere die Normalgleichungen zu dieser Aufgabe, gebe also notwendige und hinreichende Optimalitätsbedingungen für (P) an.

4. Gegeben sei das lineare Ausgleichsproblem

   (P) $$\text{Minimiere} \quad \|Ax-b\|_2, \quad x \in \mathbb{R}^n,$$

   wobei $A \in \mathbb{R}^{m\times n}$ mit Rang $(A) = n$ und $b \in \mathbb{R}^m$ mit $b \notin \text{Bild}\,(A)$ gegeben sind. Man bilde die Matrix

   $$\bar{A} := (\ A \quad b\ ) \in \mathbb{R}^{m\times(n+1)}$$

   und zeige:

   (a) Es ist Rang $(\bar{A}) = n+1$. Daher besitzt $\bar{A}^T\bar{A}$ eine Cholesky-Zerlegung $\bar{A}^T\bar{A} = \bar{L}\bar{L}^T$.

(b) Ist

$$\bar{L} = \begin{pmatrix} L & 0 \\ l^T & \lambda \end{pmatrix},$$

erhält man ferner x durch Rückwärtseinsetzen aus $L^T x = l$, so ist x die Lösung von (P) und $\|Ax - b\|_2 = \lambda$.

3.2 Die QR-Zerlegung

In diesem Abschnitt ist $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ gegeben. Ziel ist es stets, eine reduzierte oder volle QR-Zerlegung von A zu bestimmen. Hierzu gehen wir ein auf

- Gram-Schmidt-Verfahren,
- Householder-Verfahren mit und ohne Pivotisierung,
- Givens-Verfahren.

Da einiges hiervon schon aus der numerischen Mathematik bekannt ist, werden wir uns kurz fassen.

3.2.1 Gram-Schmidt-Verfahren

Beim Gram-Schmidt-Verfahren sind linear unabhängige Vektoren $\{a_1, \ldots, a_n\} \subset \mathbb{R}^m$ bzw. eine Matrix $A = (\, a_1 \ \cdots \ a_n \,) \in \mathbb{R}^{m\times n}$ gegeben. Ziel ist es, die Vektoren $\{a_1, \ldots, a_n\}$ sukzessive zu orthonormieren, also ein Orthonormalsystem $\{q_1, \ldots, q_n\}$ mit $\operatorname{span}\{q_1, \ldots, q_k\} = \operatorname{span}\{a_1, \ldots, a_k\}$, $k = 1, \ldots, n$, zu bestimmen. Ist dies gelungen, so ist $\hat{Q} = (\, q_1 \ \cdots \ q_n \,) \in \mathbb{R}^{m\times n}$ eine Matrix mit $\hat{Q}^T\hat{Q} = I$, ferner lässt sich a_k in eindeutiger Weise als Linearkombination von $\{q_1, \ldots, q_k\}$, $k = 1, \ldots, n$, darstellen:

$$a_k = \sum_{i=1}^{k} r_{ik} q_i, \qquad k = 1, \ldots, n.$$

Dies wiederum ist gleichwertig mit einer reduzierten QR-Zerlegung $A = \hat{Q}\hat{R}$, wobei die obere Dreiecksmatrix $\hat{R} \in \mathbb{R}^{n\times n}$ für $i \leq k$ den Eintrag r_{ik} besitzt. Dies erkennt man sehr einfach, wenn man auf beiden Seiten der behaupteten Gleichung die k-te Spalte betrachtet. Denn es ist

$$\hat{Q}\hat{R}e_k = (\, q_1 \ \cdots \ q_k \ \ q_{k+1} \ \cdots \ q_n \,) \begin{pmatrix} r_{1k} \\ \vdots \\ r_{kk} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \sum_{i=1}^{k} r_{ik} q_i = a_k = Ae_k.$$

Das *klassische Gram-Schmidt-Verfahren* (CGS)[1] (oft auch Orthonormierungsverfahren nach E. Schmidt genannt und in jeder Vorlesung über Lineare Algebra oder Hilbertraumtheorie vorkommend) bzw. das *modifizierte Gram-Schmidt-Verfahren* (MGS) verlaufen folgendermaßen (sie unterscheiden sich nur in einem Statement):

- Input: Gegeben $A = (\ a_1 \ \cdots \ a_n\ ) \in \mathbb{R}^{m\times n}$ mit $m \geq n$.
- Für $k = 1, \ldots, n$:

  Setze $q'_k := a_k$.

  Für $i = 1, \ldots, k-1$:

  Berechne $r_{ik} := q_i^T a_k$ (CGS) bzw. $r_{ik} := q_i^T q'_k$ (MGS).

  Berechne $q'_k := q'_k - r_{ik} q_i$.

  Berechne $r_{kk} := \|q'_k\|_2$.

  Falls $r_{kk} = 0$, dann: STOP, da $a_k \in \operatorname{span}\{a_1, \ldots, a_{k-1}\}$.

  Andernfalls: Berechne $q_k := q'_k / r_{kk}$.
- Output: Wenn das (genauer: die) Verfahren wegen Rang $(A) < n$ nicht vorzeitig abbricht (abbrechen), werden eine Matrix $\hat{Q} = (\ q_1 \ \cdots \ q_n\ ) \in \mathbb{R}^{m\times n}$ mit $\hat{Q}^T\hat{Q} = I$ und eine obere Dreiecksmatrix $\hat{R}$ mit r_{ik} für $i \leq k$ als (i,k)-Eintrag erzeugt, die eine reduzierte QR-Zerlegung von A bilden, für die also $A = \hat{Q}\hat{R}$.

Die Matrizen $\hat{Q}$ und $\hat{R}$ werden spaltenweise berechnet. Es wäre einfach, im obigen Programm die Matrix A durch $\hat{Q}$ zu überspeichern. Dass die beiden Verfahren CGS und MGS bei exakter Arithmetik äquivalent sind, also dieselben Matrizen $\hat{Q}$ und $\hat{R}$ berechnen, wird als Aufgabe gestellt. Das klassische CGS ist für die Praxis unbrauchbar, wovon man sich am besten durch ein Beispiel selbst überzeugen sollte (siehe Aufgabe 2). Das MGS hat gegenüber den später zu besprechenden Householder- und Givens-Verfahren den Vorteil, dass es explizit die Matrix $\hat{Q}$ liefert, während dies bei den beiden anderen Verfahren in faktorisierter Form geschieht. Allerdings geht in der Praxis die Orthonormalität der Spalten sukzessive immer mehr verloren.

Bemerkung: Das CGS ist sehr einfach zu motivieren. Angenommen, man hat schon ein Orthonormalsystem $\{q_1, \ldots, q_{k-1}\}$ mit

$$a_k \notin \operatorname{span}\{a_1, \ldots, a_{k-1}\} = \operatorname{span}\{q_1, \ldots, q_{k-1}\}$$

konstruiert. Um $a_k \in \operatorname{span}\{q_1, \ldots, q_k\}$ zu sichern, macht man den Ansatz

$$a_k = \sum_{j=1}^{k-1} r_{jk} q_j + r_{kk} q_k$$

[1] In Kapitel 5 über Iterationsverfahren bei linearen Gleichungssystemen werden wir ein Verfahren zur Lösung eines linearen Gleichungssystems kennenlernen, das ebenfalls mit CGS abgekürzt wird, wobei hier die Abkürzung für **C**onjugate **G**radient **S**quared steht. Also keine Verwechslung!

mit noch unbekannten $r_{1k}, \ldots, r_{kk}$ sowie q_k. Aus diesem Ansatz erhält man $r_{ik} = q_i^T a_k$, $i = 1, \ldots, k-1$. Mit $q_k' := a_k - \sum_{j=1}^{k-1} r_{jk} q_j \neq 0$ ist (jedenfalls, wenn $\hat{R}$ positive Diagonalelemente haben soll) notwendigerweise $r_{kk} = \|q_k'\|_2$ und $q_k = q_k'/r_{kk}$. Das sind genau die im obigen Verfahren angegebenen Formeln. □

Bemerkung: Die Motivation für das modifizierte Gram-Schmidt-Verfahren ist ähnlich einfach. Sei $A^{(1)} := A \in \mathbb{R}^{m \times n}$ mit Rang $(A) = n$ gegeben. Wir nehmen an, die ersten $k-1$ Spalten der Matrizen $\hat{Q}$ und $\hat{R}$ einer reduzierten QR-Zerlegung von A seien schon berechnet. Im nächsten Schritt hat man die Forderungen

$$a_k^{(1)} = \sum_{j=1}^{k} r_{jk} q_j, \qquad q_j^T q_k = \delta_{jk} \quad (j = 1, \ldots, k).$$

Es ist $r_{1k} = q_1^T a_k^{(1)}$ und

$$a_k^{(2)} := a_k^{(1)} - r_{1k} q_1 = \sum_{j=2}^{k} r_{jk} q_j.$$

Hieraus erhält man $r_{2k} = q_2^T a_k^{(2)}$. Man berechnet also $r_{1k}, \ldots, r_{k-1,k}$ durch den folgenden Algorithmus:

- Setze $a_k^{(1)} := a_k$.
- Für $i = 1, \ldots, k-1$:
  Berechne $r_{ik} := q_i^T a_k^{(i)}$.
  Berechne $a_k^{(i+1)} := a_k^{(i)} - r_{ik} q_i$.

Es ist unschwer zu erkennen, dass dies genau die beim MGS vorkommende Iterationsvorschrift ist. Der Rest ist klar: Man setze $r_{kk} := \|a_k^{(k)}\|_2$ und $q_k := a_k^{(k)}/r_{kk}$. □

Bemerkung: Man kann das MGS (und auch das CGS) auffassen als ein Verfahren, die gegebene Matrix $A^{(1)} := A$ durch sukzessive Multiplikation von rechts mit oberen Dreiecksmatrizen in eine orthogonale Matrix zu überführen. Wir setzen

$$A^{(k)} := \begin{pmatrix} q_1 & \cdots & q_{k-1} & a_k & \cdots & a_n \end{pmatrix}$$

und nehmen an, die k-te Spalte

$$r_k := (r_{1k}, \ldots, r_{kk}, \underbrace{0, \ldots, 0}_{n-k})^T$$

sei (mit CGS oder MGS) berechnet. Dann ist

$$R_k := I - \frac{(r_k - e_k) e_k^T}{r_{kk}}$$

eine obere Dreiecksmatrix, die nur in der k-ten Spalte von der Identität abweicht:

$$(R_k)_{ik} = \begin{cases} -r_{ik}/r_{kk}, & i = 1, \ldots, k-1, \\ 1/r_{kk} & i = k. \end{cases}$$

Die Matrix $A^{(k+1)} := A^{(k)}R_k$ unterscheidet sich von $A^{(k)}$ nur in der k-ten Spalte. Diese ist aber

$$A^{(k+1)}e_k = A^{(k)}R_k e_k = \frac{1}{r_{kk}}\Big(a_k - \sum_{i=1}^{k-1} r_{ik}q_i\Big) = q_k.$$

Folglich ist $AR_1 \cdots R_n = \hat{Q}$ bzw. $A = \hat{Q}\hat{R}$ mit $\hat{R} = R_n^{-1} \cdots R_1^{-1}$. □

Bemerkung: Sei das lineare Ausgleichsproblem

$$\text{(P)} \qquad \text{Minimiere} \quad \|Ax - b\|_2, \quad x \in \mathbb{R}^n,$$

gegeben, wobei $A \in \mathbb{R}^{m\times n}$ mit Rang $(A) = n$ und $b \in \mathbb{R}^m$. Die naheliegende Methode zur Berechnung der eindeutig existierenden Lösung besteht darin, eine reduzierte QR-Zerlegung $A = \hat{Q}\hat{R}$ etwa mit dem modifizierten Gram-Schmidt-Verfahren zu berechnen, danach $c := \hat{Q}^T b$ und dann x durch Rückwärteinsetzen aus $\hat{R}x = c$. *Besser* ist es, im Prinzip (genaue Formulierung folgt) folgendermaßen vorzugehen (siehe Å. Björk (1996, S. 64)):

- Berechne eine reduzierte QR-Zerlegung

$$\begin{pmatrix} A & b \end{pmatrix} = \begin{pmatrix} \hat{Q} & q_{n+1} \end{pmatrix} \begin{pmatrix} \hat{R} & z \\ 0 & \rho \end{pmatrix}$$

  durch Anwendung des MGS auf $\begin{pmatrix} A & b \end{pmatrix}$.

- Berechne die Lösung x von (P) durch Rückwärtseinsetzen aus $\hat{R}x = z$, ferner ist $b - Ax = \rho q_{n+1}$ und insbesondere $\|Ax - b\|_2 = \rho$.

Dass diese Aussagen korrekt sind, kann man leicht einsehen. Denn aus der obigen Darstellung folgt $A = \hat{Q}\hat{R}$ und $b = \hat{Q}z + \rho q_{n+1}$. Mit $\hat{R}x = z$ ist daher

$$A^T(Ax - b) = \hat{R}^T\hat{Q}^T(\hat{Q}\underbrace{\hat{R}x}_{=z} - \hat{Q}z - \rho q_{n+1}) = -\rho\hat{R}^T\underbrace{\hat{Q}^T q_{n+1}}_{=0} = 0,$$

d. h. x genügt den Normalgleichungen, ist also die Lösung von (P). Die restliche Aussage ist genauso einfach nachzuweisen. Will man nur die Lösung x bestimmen, so interessiert nicht q_{n+1} oder ρ, sondern nur die letzte Spalte $z \in \mathbb{R}^n$ im zweiten Faktor einer reduzierten QR-Zerlegung von $\begin{pmatrix} A & b \end{pmatrix}$. In Pseudocode erhält man die Lösung x von (P) also folgendermaßen:

- Berechne die reduzierte QR-Zerlegung $A = \hat{Q}\hat{R}$ mit $\hat{Q} = \begin{pmatrix} q_1 & \cdots & q_n \end{pmatrix}$ mit Hilfe von MGS.
- Setze $b' := b$.
- Für $i = 1, \ldots, n$:
  Berechne $z_i := q_i^T b'$.
  Berechne $b' := b' - z_i q_i$.
- Löse $\hat{R}x = z$.

□

3.2.2 Das Householder-Verfahren

Wieder sei eine Matrix $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ vorgegeben. Die Idee des Householder-Verfahrens zur Berechnung einer vollen QR-Zerlegung von A ist aus der numerischen Mathematik bekannt: Man führe A durch Multiplikation mit $\min(n, m-1)$ (also $n-1$ für $m = n$ und n für $m > n$) sogenannten Householder-Spiegelungen (spezielle orthogonale Matrizen) in eine obere Dreiecksmatrix über. Wir können uns kurz fassen und geben die wichtigsten Tatsachen ohne Beweis an.

1. Eine Matrix der Form
$$Q := I - \frac{2}{u^T u} uu^T$$
mit $u \in \mathbb{R}^m \setminus \{0\}$ heißt eine *Householder-Matrix*. Householder-Matrizen sind symmetrisch und orthogonal.

2. Ist $a \in \mathbb{R}^m \setminus \{0\}$ und definiert man $u := a + \operatorname{sign}(a_1)\|a\|_2\, e_1$, so ist durch
$$Q := I - \frac{2}{u^T u} uu^T = I - \beta uu^T \qquad \text{mit} \quad \beta := \frac{2}{u^T u} = \frac{1}{\|a\|_2(\|a\|_2 + |a_1|)}$$
eine Householder-Matrix gegeben, die a in ein Vielfaches des ersten Einheitsvektors überführt. Genauer ist $Qa = -\operatorname{sign}(a_1)\|a\|_2\, e_1$. Hierbei sei $\operatorname{sign}(0) = 1$.

3. Das Householder-Verfahren ohne Spaltenpivotsuche basiert genau darauf, dass man einen vom Nullvektor verschiedenen Vektor mit Hilfe einer Householder-Matrix in ein Vielfaches des ersten Einheitsvektors überführen kann. Im Prinzip sieht es folgendermaßen aus:

   - Gegeben sei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$.
   - Für $k = 1, \ldots, \min(n, m-1)$:
     Bestimme $(m-k+1) \times (m-k+1)$-Householder-Matrix $\tilde{Q}_k$ mit
     $\tilde{Q}_k(a_{kk}, \ldots, a_{mk})^T = (*, \underbrace{0, \ldots, 0}_{n-k-1})^T$.
     Setze $Q_k := \operatorname{diag}(I_{k-1}, \tilde{Q}_k)$ und berechne $A := Q_k A$.
   - Die Ausgangsmatrix wird in n Schritten (für $m = n$ genügen $n-1$ Schritte, da in diesem Falle der letzte Schritt trivial ist) in eine obere Dreiecksmatrix $R := Q_n \cdots Q_1 A$ transformiert. Hierdurch und durch $Q := Q_1 \cdots Q_n$ ist eine vollständige QR-Zerlegung von A erhalten.

   Einzelheiten einer möglichen Implementation sind ausführlich bei J. Werner (1992a, S. 55 ff.) besprochen. Wichtig ist, dass die Matrix Q als Produkt der Householder-Matrizen nicht explizit gebildet wird. Stattdessen speichert man sie in faktorisierter Form $Q = Q_1 \cdots Q_n$ in den gerade frei gewordenen Spalten der Matrix A. Als Output erhält man dann eine Matrix $A \in \mathbb{R}^{m\times n}$ und Vektoren

$r, \beta \in \mathbb{R}^n$ mit

$$A = \begin{pmatrix} u_{11} & r_{12} & \cdots & r_{1n} \\ u_{21} & u_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{n1} & u_{n2} & \cdots & u_{nn} \\ \vdots & \vdots & & \vdots \\ u_{m1} & u_{m2} & \cdots & u_{mn} \end{pmatrix}, \qquad r = \begin{pmatrix} r_{11} \\ r_{22} \\ \vdots \\ r_{nn} \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}.$$

Hierbei ist

$$Q_k = I - \beta_k u_k u_k^T \qquad \text{mit} \quad u_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ u_{kk} \\ \vdots \\ u_{mk} \end{pmatrix}$$

der k-te Faktor der orthogonalen Matrix $Q = Q_1 \cdots Q_n$. In der strikten oberen Hälfte von A wird die strikte obere Hälfte der oberen Dreiecksmatrix R gespeichert, deren Diagonale im Vektor r festgehalten wird. Ist Rang $(A) = n$, sind also die n Spalten von A linear unabhängig, ist ferner

$$R = \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix},$$

so ist die obere Dreiecksmatrix $\hat{R} \in \mathbb{R}^{n \times n}$ nichtsingulär.

Wenn ein lineares Ausgleichsproblem

$$\text{(P)} \qquad \text{Minimiere} \quad \|Ax - b\|_2, \quad x \in \mathbb{R}^n,$$

gegeben ist, so wird man sich den orthogonalen Faktor Q nur dann merken, wenn (P) mit ein und derselben Koeffizientenmatrix A und verschiedenen "rechten Seiten" b zu lösen ist. Ansonsten multipliziert man sukzessive b mit den Householder-Matrizen Q_k, die danach ihre Schuldigkeit getan und vergessen werden können.

Wenn man weiß (und wir wissen es jetzt), wie die Berechnung einer QR-Zerlegung nach Householder einer Matrix $A \in \mathbb{R}^{m \times n}$ mit Rang $(A) = n$ erfolgt, so ist die Modifikation auf den möglicherweise Rang-defizienten Fall einfach. Wir nehmen an, es seien schon $k-1$ Householder-Matrizen $Q_1, \ldots, Q_{k-1} \in \mathbb{R}^{m \times m}$ und Vertauschungsmatrizen $\Pi_1, \ldots, \Pi_{k-1} \in \mathbb{R}^{n \times n}$ mit

$$Q_{k-1} \cdots Q_1 A \Pi_1 \cdots \Pi_{k-1} = R^{(k-1)} = \begin{pmatrix} R_{11}^{(k-1)} & R_{12}^{(k-1)} \\ 0 & R_{22}^{(k-1)} \end{pmatrix}$$

gefunden, wobei $R_{11}^{(k-1)} \in \mathbb{R}^{(k-1)\times(k-1)}$ eine nichtsinguläre obere Dreiecksmatrix ist und

$$R_{12}^{(k-1)} \in \mathbb{R}^{(k-1)\times(n-k+1)}, \qquad R_{22}^{(k-1)} = \begin{pmatrix} r_k^{(k-1)} & \cdots & r_n^{(k-1)} \end{pmatrix} \in \mathbb{R}^{(m-k+1)\times(n-k+1)}.$$

Für $k = 1$, also den Anfang, wird nichts vorausgesetzt. Man bestimme nun einen Index $p \in \{k, \ldots, n\}$ mit

$$\|r_p^{(k-1)}\| = \max(\|r_k^{(k-1)}\|_2, \ldots, \|r_n^{(k-1)}\|_2).$$

Ist $\|r_p^{(k-1)}\|_2 = 0$, so ist $R_{22}^{(k-1)} = 0$ und Rang $(A) = k - 1$. Ist dies nicht der Fall, so vertausche man in $R^{(k-1)}$ die k-te und die p-te Spalte. Dies erreicht man bekanntlich dadurch, dass man $R^{(k-1)}$ von rechts mit der Vertauschungsmatrix $\Pi_k := P_{kp}$ multipliziert. Anschließend bestimme man eine Householder-Matrix $\tilde{Q}_k \in \mathbb{R}^{(m-k+1)\times(m-k+1)}$ derart, dass $\tilde{Q}_k r_p^{(k-1)}$ ein Vielfaches des ersten Einheitsvektors ist. Mit $Q_k := \text{diag}\,(I_{k-1}, \tilde{Q}_k)$ erhält man auf diese Weise

$$Q_k \cdots Q_1 A \Pi_1 \cdots \Pi_k = R^{(k)} = \begin{pmatrix} R_{11}^{(k)} & R_{12}^{(k)} \\ 0 & R_{22}^{(k)} \end{pmatrix}$$

mit einer nichtsingulären oberen Dreiecksmatrix $R_{11}^{(k)} \in \mathbb{R}^{k\times k}$. Damit ist ein Iterationsschritt des Householder-Verfahrens mit Spaltenpivotsuche beschrieben. Insgesamt berechnet man eine orthogonale Matrix $Q \in \mathbb{R}^{m\times m}$ (i. allg. nur in faktorisierter Form) und eine Permutationsmatrix Π derart, dass

$$A\Pi = Q \begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix},$$

wobei $R_{11} \in \mathbb{R}^{r\times r}$ mit $r = \text{Rang}\,(A)$ eine nichtsinguläre obere Dreiecksmatrix ist. In Aufgabe 5 ist angegeben, wie man mit dem Householder-Verfahren mit Spaltenpivotisierung auch im Rang-defizienten Fall eine Lösung des gegebenen linearen Ausgleichsproblems berechnen kann. Allerdings ist diese Lösung, die sogenannte *Basislösung*, nicht notwendig die Lösung minimaler euklidischer Norm.

3.2.3 Das Givens-Verfahren

Sehr kurz wollen wir auf das Givens-Verfahren zur Berechnung einer vollen QR-Zerlegung einer Matrix $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ eingehen. Das Givens-Verfahren ist vor allem für speziell strukturierte Matrizen geeignet. In Kurzform fassen wir das wichtigste zusammen.

Eine *Givens-Rotation* $G_{ik} \in \mathbb{R}^{m\times m}$ mit $1 \leq i < k \leq m$ unterscheidet sich von der Einheitsmatrix nur in den Positionen (i,i), (i,k), (k,i) und (k,k) in denen sie die Werte

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \qquad \text{mit} \quad c^2 + s^2 = 1$$

besitzt. Offenbar sind Givens-Rotationen spezielle orthogonale Matrizen. Ist $x \in \mathbb{R}^m$ und $y := G_{ik}x$, so ist

$$y_j = \begin{cases} cx_i + sx_k & \text{für} \quad j = i, \\ -sx_i + cx_k & \text{für} \quad j = k, \\ x_j & \text{für} \quad j \neq i, k, \end{cases} \qquad j = 1, \ldots, m.$$

Dies bedeutet, dass der Vektor x bei der Multiplikation mit G_{ik} nur in der i-ten und der k-ten Komponente verändert wird, diese werden um den Winkel θ im Uhrzeigersinn gedreht, wobei $c = \cos\theta$ und $s = \sin\theta$. Multipliziert man eine Matrix $A \in \mathbb{R}^{m\times n}$ von links mit der Givens-Rotation G_{ik}, so bewirkt dies lediglich eine Veränderung der i-ten und der k-ten Zeile. Die neuen Zeilen sind eine Linearkombination der alten und gegeben durch

$$(G_{ik}A)_{ij} = ca_{ij} + sa_{kj}, \qquad (G_{ik}A)_{kj} = -sa_{ij} + ca_{kj} \qquad (j = 1, \ldots, n).$$

Grundlegend für das Givens-Verfahren zur Berechnung einer QR-Zerlegung von $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ ist, dass man bei vorgegebenen $1 \leq i < k \leq m$ und $x \in \mathbb{R}^n$ eine Givens-Rotation G_{ik} mit $(G_{ik}x)_k = 0$ bestimmen kann. Durch eine Rotation in der (i,k)-Ebene kann also die k-te Komponente von $y := G_{ik}x$ zu Null gemacht werden, wobei außer der i-ten alle anderen unverändert bleiben. Hierzu ist es praktisch, sich eine Funktion givrot (siehe Å. Björk (1996, S. 54) oder auch J. Werner (1992a, S. 59)) zu definieren, die bei einem vorgegebenen Paar (α, β) reeller Zahlen ein Tripel (c, s, γ) mit

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \gamma \\ 0 \end{pmatrix}, \qquad c^2 + s^2 = 1$$

bestimmt. Dies kann folgendermaßen geschehen:

- Input: Gegeben $\alpha, \beta \in \mathbb{R}$.
- Falls $\beta = 0$, dann: $c := 1, \quad s := 0, \quad \gamma := \alpha$.

  Andernfalls:

  Falls $|\beta| \geq |\alpha$, dann:

  $t := \alpha/\beta, \quad u := (1 + t^2)^{1/2}, \quad s := 1/u, \quad c := st, \quad \gamma := \beta u.$

  Andernfalls:

  $t := \beta/\alpha, \quad u := (1 + t^2)^{1/2}, \quad c := 1/u, \quad s := ct, \quad \gamma := \alpha u.$
- Output: Für $\text{givrot}(\alpha, \beta) := (c, s, \gamma)$ gilt $c^2 + s^2 = 1$ und

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \gamma \\ 0 \end{pmatrix}.$$

Das Givens-Verfahren zur Berechnung einer QR-Zerlegung besteht darin, die gegebene Matrix sukzessive von links mit Givens-Rotationen zu multiplizieren, dabei jeweils ein Element zu annullieren und darauf zu achten, dass zu Null gemachte Einträge auch weiter verschwinden. Man hat einige naheliegende Möglichkeiten, in welcher Reihenfolge man Einträge zu Null macht. Z. B. kann man spaltenweise von oben nach unten vorgehen: Durch $G_{1n} \cdots G_{12}A$ hat man bei geeigneten Rotationen die erste Spalte von A in ein Vielfaches des ersten Einheitsvektors überführt, anschließend kann man sich entsprechend die zweite Spalte vornehmen. Natürlich kann man aber auch von unten beginnen. Bei einer oberen Hessenberg-Matrix (nur die Subdiagonale ist hier besetzt), kann man die Subdiagonale sukzessive zu Null machen. Da bei vollbesetzter

Matrix A das Givens-Verfahren etwa doppelt so teuer wie das Householder-Verfahren ist, wollen wir den dicht-besetzten Fall nicht weiter betrachten, uns aber in Zukunft daran erinnern, dass Givens-Rotationen ein elegantes Hilfsmittel sind, wie mit einem Seziermesser störende Elemente (bzw. besser: Einträge) zu entfernen. Sehr häufig ist es nicht nötig, sich die orthogonalen Matrizen (hier: Givens-Rotationen), mit denen A von links multipliziert wird, zu merken oder gar zu akkumulieren. Denn ist etwa ein lineares Ausgleichsproblem mit den Daten $(\; A \quad b \;)$ gegeben, so multipliziert man auch gleich die "rechte Seite" b mit der entsprechenden orthogonalen Matrix durch und kann sie danach (jedenfalls dann, wenn man nur mit *einem* b ein lineares Ausgleichsproblem zu lösen hat) vergessen. Gelegentlich ist es aber doch nötig, sich den orthogonalen Anteil einer QR-Zerlegung wenigstens in faktorisierter Form zu merken. Beim Householder-Verfahren wurden die relevanten Informationen über die benutzten Householder-Matrizen im wesentlichen in den gerade geräumten Spalten von A (und einem zusätzlichen Vektor) gespeichert. Beim Givens-Verfahren wird in jedem Schritt im wesentlichen nur ein Eintrag zu Null, an dieser Stelle sollte die Information (also (c, s), nicht aber (i, k)) über die benutzte Givens-Rotation sozusagen in chiffrierter Form hinterlegt werden. Hierzu speichere man

$$\rho := \begin{cases} 1, & \text{falls} \quad c = 0, \\ \operatorname{sign}(c)s, & \text{falls} \quad |s| < |c|, \\ \operatorname{sign}(s)/c, & \text{falls} \quad |c| \leq |s|. \end{cases}$$

Bei gegebenem ρ kann man c und s bis auf einen gemeinsamen Faktor ± 1 zurückgewinnen:

$$\text{Falls} \quad \begin{cases} \rho = 1, \\ |\rho| < 1, \\ |\rho| > 1, \end{cases} \qquad \text{dann} \quad \begin{cases} c := 0, & s := 1, \\ s := \rho, & c := (1 - s^2)^{1/2}, \\ c := 1/\rho, & s := (1 - c^2)^{1/2}. \end{cases}$$

Das ist schon alles, was wir zum Givens-Verfahren sagen wollen.

3.2.4 MATLAB-Ergänzungen

Wir beginnen mit dem klassischen Gram-Schmidt-Verfahren. Als MATLAB-Funktion könnte dieses Verfahren folgendermaßen umgesetzt werden.

```
function [Q,R]=CGS(A);
%Classical Gram-Schmidt
%*************************************************
%Pre:  A is m-by-n with rank(A)=n<=m
%Post: Columns of Q orthogonal, i.e. Q'*Q=I,
%      R upper triangular with Q*R=A
%*************************************************
[m,n]=size(A);Q=zeros(m,n);R=zeros(n,n);
for k=1:n
    R(1:k-1,k)=Q(1:m,1:k-1)'*A(1:m,k);
    z=A(1:m,k)-Q(1:m,1:k-1)*R(1:k-1,k);
```

```
    R(k,k)=norm(z);
    Q(1:m,k)=z/R(k,k);
end;
```

Nimmt man für A die 8×8-Hilbertmatrix, berechnet man `[Q,R]=CGS(A)`, so erhält man, dass das $(7,6)$-Element von Q^TQ ungefähr -0.0824 ist, so dass man kaum von Orthogonalität bei der Matrix Q sprechen kann. Zwar ist der Defekt $A - QR$ klein (es ist $\|A - QR\|_\infty = 5.5511 \cdot 10^{-17}$) , die Matrix Q ist aber nicht orthogonal (es ist $\|Q^TQ - I\|_\infty = 1.0834$). Das ist ein typisches Verhalten beim klassischen Gram-Schmidt-Verfahren, wenn die Spalten der Matrix A fast linear abhängig sind.

Es folgt das modifizierte Gram-Schmidt-Verfahren.

```
function [Q,R]=MGS(A);
%Modified Gram-Schmidt
%*************************************************
%Pre:  A is m-by-n with rank(A)=n<=m
%Post: Columns of Q orthogonal, i.e. Q'*Q=I,
%      R upper triangular with Q*R=A
%*************************************************
[m,n]=size(A);Q=zeros(m,n);R=zeros(n,n);
for k=1:n
    R(k,k)=norm(A(1:m,k));
    Q(1:m,k)=A(1:m,k)/R(k,k);
    for j=k+1:n
        R(k,j)=Q(1:m,k)'*A(1:m,j);
        A(1:m,j)=A(1:m,j)-Q(1:m,k)*R(k,j);
    end;
end;
```

Macht man das gleiche Experiment wie oben, berechnet man also mit `[Q,R]=MGS(A)` eine QR-Zerlegung der 8×8-Hilbertmatrix A, so erhält man $\|A-QR\|_\infty = 9.7145\cdot10^{-17}$ und $\|Q^TQ - I\|_\infty = 2.7528 \cdot 10^{-7}$, das Resultat ist also wesentlich besser als beim klassischen Gram-Schmidt-Verfahren.

Ähnlich wie bei Golub-Van Loan formulieren wir zunächst zwei Hilfsfunktionen `House` und `RowHouse`.

```
function v=House(x);
%*****************************************************************
%Given an n-vector x, this function computes an n-vector
%v with v(1)=1 (nonstandard normalization!) such that
%((I-2*v*v'/(v'*v))*x is zero in all but the first component
%*****************************************************************
n=length(x);mu=norm(x);v=x;
if mu~=0
   if x(1)==0
        beta=mu;
   else
```

```
        beta=x(1)+sign(x(1))*mu;
   end
   v(2:n)=v(2:n)/beta;
end
v(1)=1;
```

Man beachte, dass wir die Normierung $v(1) = 1$ benutzen (anders als in obiger Darstellung!).

```
function A=RowHouse(A,v);
%************************************************************************
%Given an m-by-n matrix A and a m-vector v with v(1)=1, the
%following algorithm overwrites A with PA where P=I-2*v*v'/(v'*v).
%************************************************************************
beta=2/(v'*v);
w=beta*A'*v;
A=A-v*w';
```

In der folgenden Funktion wird eine QR-Zerlegung nach Householder durchgeführt. Wieder benutzen wir bis auf Kleinigkeiten Funktionen bei Golub-van Loan.

```
function [Q,R]=HouseholderQR(A);
%*********************************************************************
%Given m-by-n matrix A with rank(A)=n<=m
%the following algorithm computes a full QR-decomposition
%of A with Householder method without partial pivoting.
%*********************************************************************
[m,n]=size(A);v=zeros(m,1);
for j=1:n
    v(j:m)=House(A(j:m,j));
    A(j:m,j:n)=RowHouse(A(j:m,j:n),v(j:m));
    if j<m
        A(j+1:m,j)=v(j+1:m);
    end;
end;
R=triu(A);
Q=eye(m);
for j=n:-1:1
    v(j)=1;v(j+1:m)=A(j+1:m,j);
    Q(j:m,j:m)=RowHouse(Q(j:m,j:m),v(j:m));
end;
```

Ist A die 8×8-Hilbertmatrix, so erhält man als Resultat des Aufrufs `HouseholderQR(A)` ein Paar (Q, R), wobei R eine obere Dreiecksmatrix und $\|A - QR\|_\infty = 1.3600 \cdot 10^{-15}$ sowie $\|Q^TQ - I\|_\infty = 1.0270 \cdot 10^{-15}$, $\|QQ^T - I\|_\infty = 1.1935 \cdot 10^{-15}$. Im Gegensatz zu den entsprechenden Beispielen mit `CGS` und `MGS` kann man hier die gewonnene Matrix wirklich als orthogonal bezeichnen.

Auf das Householder-Verfahren mit Spaltenpivotsuche wollen wir hier nicht mehr eingehen sondern gleich zum Givens-Verfahren übergehen. Bei dem Householder-Verfahren haben wir zunächst Funktionen `House` und `RowHouse` aufgestellt, die dann in `HouseholderQR` benutzt werden. Hier gehen wir wir etwas anders vor und definieren nur eine Funktions `Givens`, danach schon `GivensQR`

```
function [c,s]=Givens(a,b);
%****************************************************************
%Given scalars a and b this function computes c and s so
%   [ c  s][a] [r]
%   [     ][ ]=[ ]
%   [-s  c][b] [0]
%****************************************************************
if b==0
    c=1;s=0;
else
    if  abs(b)>abs(a)
         t=a/b;s=1/sqrt(1+t^2);c=s*t;
    else
         t=b/a;c=1/sqrt(1+t^2);s=c*t;
    end;
end;
```

Jetzt folgt das Analogon zu `householderQR` (siehe C. F. Van Loan (1997, S. 233)).

```
function [Q,R]=GivensQR(A);
%****************************************************************
%Given m-by-n matrix A with rank(A)=n<=m
%the following algorithm computes a QR-decomposition
%of A with Givens method
%****************************************************************
[m,n]=size(A);
Q=eye(m);
for j=1:n
    for i=m:-1:j+1
         [c,s]=Givens(A(i-1,j),A(i,j));
         A(i-1:i,j:n)=[c s;-s c]*A(i-1:i,j:n);
         Q(:,i-1:i)=Q(:,i-1:i)*[c -s;s c];
    end;
end;
R=triu(A);
```

Wir testen die letzte Funktion wieder an der 8×8-Hilbertmatrix `A=hilb(8)`. Wir erhalten diesmal $\|A - QR\|_\infty = 7.4940 \cdot 10^{-16}$, weiter $\|Q^TQ - I\|_\infty = 1.8249 \cdot 10^{-15}$, $\|QQ^T - I\|_\infty = 2.0504 \cdot 10^{-15}$.

Ist nur *ein* lineares Ausgleichsproblem mit den Daten (A, b) lösen, so wird man die orthogonale Matrix Q gar nicht berechnen, sondern den Vektor b sukzessive mit den

benutzten Givens-Rotationen (oder auch Householder-Matrizen) durchmultiplizieren. Wir erhalten z. B. die folgende Funktion (siehe van Loan):

```
function xLS=GivensLS(A,b);
%*********************************************************
%Pre:     A m-by-n with rank(A)=n
%         b m-by-1.
%Post:    xLS n-by-1, minimizes the 2-norm of Ax-b
%*********************************************************
[m,n]=size(A);
for j=1:n
    for i=m:-1:j+1
         [c,s]=Givens(A(i-1,j),A(i,j));
         A(i-1:i,j:n)=[c s;-s c]*A(i-1:i,j:n);
         b(i-1:i)=[c s;-s c]*b(i-1:i);
    end;
end;
xLS=UTriSol(A(1:n,1:n),b(1:n));
```

Ein ähnliches Programm könnte man natürlich leicht auch für das Householder-Verfahren schreiben.

Zum Schluss des Abschnitts über die QR-Zerlegung geben wir noch einen Teil der Informationen wieder, die man nach `help qr` erhält.

```
qr     Orthogonal-triangular decomposition.
  [Q,R] = qr(A), where A is m-by-n, produces an m-by-n upper triangular
  matrix R and an m-by-m unitary matrix Q so that A = Q*R.

  [Q,R] = qr(A,0) produces the "economy size" decomposition.
  If m>n, only the first n columns of Q and the first n rows of R are
  computed. If m<=n, this is the same as [Q,R] = qr(A).

  If A is full:

  [Q,R,E] = qr(A) produces unitary Q, upper triangular R and a
  permutation matrix E so that A*E = Q*R. The column permutation E is
  chosen so that ABS(DIAG(R)) is decreasing.
```

Wichtig ist vor allem, dass man durch `[Q,R] = qr(A,0)` die reduzierte QR-Zerlegung erhält.

3.2.5 Aufgaben

1. Man zeige, dass die in Pseudocode angegebenen Verfahren CGS und MGS bei exakter Arithmetik (also ohne auftretende Rundungsfehler) äquivalent sind, also denselben Output liefern.

2. Man programmiere das klassische und das modifizierte Gram-Schmidt-Verfahren und teste es daran, eine reduzierte QR-Zerlegung $A = \hat{Q}\hat{R}$ der 7×5-Hilbertmatrix $A := (1/(i+j-1)) \in \mathbb{R}^{7\times 5}$ zu berechnen. Insbesondere teste man die "Orthogonalität" von $\hat{Q}$ durch Berechnung von $\hat{Q}^T\hat{Q}$.

3. Sei
$$A := \begin{pmatrix} 1 & 2 & 3 \\ 1 & 5 & 6 \\ 1 & 8 & 9 \\ 1 & 11 & 12 \end{pmatrix}.$$
Mit dem Householder-Verfahren ohne und mit Spaltenpivotsuche berechne man eine volle QR-Zerlegung von A.

4. Gegeben sei das lineare Gleichungssystem $Ax = b$ mit $A \in \mathbb{R}^{m\times n}$ und Rang $(A) = m \leq n$ und $b \in \mathbb{R}^m$, d. h. wir haben ein *unterbestimmtes* lineares Gleichungssystem. Man zeige, wie man mit Hilfe einer QR-Zerlegung von A^T die eindeutige Lösung von $Ax = b$ mit minimaler euklidischer Norm berechnen kann.

5. Gegeben sei das lineare Ausgleichsproblem
$$\text{(P)} \qquad \text{Minimiere} \quad \|Ax - b\|_2, \quad x \in \mathbb{R}^n,$$
wobei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$, $b \in \mathbb{R}^m$. Durch das Householder-Verfahren mit Spaltenpivotisierung sei die Zerlegung
$$A\Pi = QR = Q \begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix}$$
berechnet, wobei $Q \in \mathbb{R}^{m\times m}$ orthogonal, $\Pi \in \mathbb{R}^{n\times n}$ eine Permutationsmatrix, $R_{11} \in \mathbb{R}^{r\times r}$ eine nichtsinguläre obere Dreiecksmatrix und $R_{12} \in \mathbb{R}^{r\times(n-r)}$. Sei
$$Q^T b =: \begin{pmatrix} c \\ d \end{pmatrix} \qquad \text{mit} \quad c \in \mathbb{R}^r,\ d \in \mathbb{R}^{m-r}.$$
Sei
$$x_B := \Pi \begin{pmatrix} R_{11}^{-1} c \\ 0 \end{pmatrix}$$
die sogenannte *Basislösung* von (P). Man zeige:
   (a) Die Basislösung x_B ist eine Lösung von (P).
   (b) Ist $R_{12} = 0$, so ist x_B die eindeutige Lösung minimaler Norm von (P).

6. Was sind die Eigenwerte einer Householder-Matrix, was die einer Givens-Rotation?

7. Sei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ gegeben. Es sei eine volle QR-Zerlegung
$$A = Q \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix}$$
von A bekannt. Wie kann man dann mit $O(m^2)$ flops eine volle QR-Zerlegung von $\tilde{A} := A + uv^T$ bestimmen, wobei $u \in \mathbb{R}^m$, $v \in \mathbb{R}^n$ gegeben sind?

8. Seien $A \in \mathbb{R}^{m\times n}$ mit Rang $(A) = n$, $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n$ gegeben. Hiermit betrachte man die beiden Aufgaben (siehe auch Aufgabe 2 in Abschnitt 3.1)

$$\text{Minimiere} \quad f(x) := \tfrac{1}{2}\|Ax - b\|_2^2 + c^T x, \quad x \in \mathbb{R}^n$$

und

$$\text{Minimiere} \quad g(y) := \tfrac{1}{2}\|y - b\|_2^2, \quad A^T y = c.$$

Mit Hilfe einer QR-Zerlegung von A gebe man eine Methode an, die eindeutig existierenden Lösungen zu berechnen.

3.3 Die Singulärwertzerlegung

3.3.1 Definition und grundlegende Eigenschaften

Wir gehen weiter von einer Matrix $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ aus. Die Übertragung ins Komplexe ist fast immer offensichtlich, ist $m < n$, so betrachte man die transponierte Matrix A^T. Wir hatten schon angegeben, was wir unter einer vollen bzw. reduzierten Singulärwertzerlegung (SVD, singular value decomposition) verstehen. Eine Darstellung der Form

$$A = U\Sigma V^T \qquad \text{mit} \quad \Sigma = \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} \in \mathbb{R}^{m\times n}$$

heißt eine *volle Singulärwertzerlegung* von A, wenn $U \in \mathbb{R}^{m\times m}$ und $V \in \mathbb{R}^{n\times n}$ orthogonal sind und $\hat{\Sigma} = \text{diag}\,(\sigma_1, \ldots, \sigma_n)$, wobei für die sogenannten *singulären Werte* gilt

$$\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0,$$

dagegen heißt die Darstellung

$$A = \hat{U}\hat{\Sigma}V^T,$$

wobei $\hat{U} \in \mathbb{R}^{m\times n}$ mit $\hat{U}^T\hat{U} = I$, $V \in \mathbb{R}^{n\times n}$ orthogonal und $\hat{\Sigma} \in \mathbb{R}^{n\times n}$ wie oben, eine *reduzierte Singulärwertzerlegung*. Natürlich erhält man aus einer vollen Singulärwertzerlegung eine reduzierte, indem man für $\hat{U}$ die ersten n Spalten von U nimmt, entsprechend erhält man aus einer reduzierten eine volle Singulärwertzerlegung, indem man die in $\hat{U}$ zusammengefassten orthonormierten Spalten zu einem Orthonormalsystem des $\mathbb{R}^m$ ergänzt. Ist

$$\hat{U} = (\ u_1 \ \cdots \ u_n\ ), \qquad V = (\ v_1 \ \cdots \ v_n\ ),$$

so heißt u_i linker und v_i rechter Singulärvektor zum Singulärwert σ_i, $i = 1, \ldots, n$. Man beachte, dass die singulären Werte zu A eindeutig bestimmt sind. Denn hat A etwa die reduzierte Singulärwertzerlegung $A = \hat{U}\hat{\Sigma}V^T$, so ist

$$A^T A = V\hat{\Sigma}^T \underbrace{\hat{U}^T\hat{U}}_{=I}\hat{\Sigma}V^T = V\hat{\Sigma}^2 V^T,$$

d. h. es ist $\sigma_i = \lambda_i(A^T A)^{1/2}$, $i = 1, \ldots, n$, wobei

$$\lambda_1(A^T A) \geq \cdots \geq \lambda_n(A^T A)$$

die (nichtnegativen) Eigenwerte der (symmetrischen und positiv semidefiniten) Matrix $A^TA \in \mathbb{R}^{n\times n}$ sind. Dasselbe Ergebnis erhält man natürlich auch für eine volle Singulärwertzerlegung. Offenbar ist

$$\text{Rang}\,(A) = n - \dim \text{Kern}\,(A) = n - \dim \text{Kern}\,(A^TA) = \text{Rang}\,(A^TA) = r,$$

wobei[2] r die Anzahl der positiven Singulärwerte bezeichnet. Nun erhält man leicht die Existenz einer Singulärwertzerlegung. Hierzu seien

$$\lambda_1 \geq \cdots \geq \lambda_r > \lambda_{r+1} = \cdots = \lambda_n = 0$$

die Eigenwerte von A^TA und $\{v_1, \ldots, v_n\}$ ein zugehöriges System von Eigenvektoren. Man definiere $\sigma_i := \lambda_i^{1/2}$, $i = 1, \ldots, n$, und hiermit

$$\hat{\Sigma} := \text{diag}\,(\sigma_1, \ldots, \sigma_n), \qquad V := (\ v_1 \ \cdots \ v_n\ ).$$

Dann ist $V \in \mathbb{R}^{n\times n}$ natürlich eine orthogonale Matrix. Weiter definiere man

$$u_i := \frac{1}{\sigma_i} Av_i, \qquad i = 1, \ldots, r.$$

Dann ist $\{u_1, \ldots, u_r\}$ ein Orthonormalsystem von Eigenvektoren zu $AA^T \in \mathbb{R}^{m\times m}$ mit zugehörigen positiven Eigenwerten $\lambda_1, \ldots, \lambda_r$. Man ergänze $\{u_1, \ldots, u_r\}$ durch $u_{r+1}, \ldots, u_m$ zu einem vollständigen System von Eigenvektoren der Matrix $AA^T \in \mathbb{R}^{m\times m}$, wobei $u_{r+1}, \ldots, u_m$ notwendigerweise Eigenvektoren zum Eigenwert 0 sind. Setzt man nun

$$U := (\ u_1 \ \cdots \ u_m\ ),$$

so ist U orthogonal, ferner

$$(U^TAV)_{ij} = u_i^TAv_j = \sigma_j u_i^T u_j = \sigma_i \delta_{ij}, \qquad 1 \leq i, j \leq r.$$

Wegen

$$AA^Tu_i = 0 \quad (i = r+1, \ldots, m), \qquad A^TAv_j = 0 \quad (j = r+1, \ldots, n)$$

sowie $\text{Kern}\,(A) = \text{Kern}\,(A^TA)$ und $\text{Kern}\,(A^T) = \text{Kern}\,(AA^T)$, ist

$$(U^TAV)_{ij} = 0 \qquad \text{falls } i \in \{r+1, \ldots, m\} \text{ oder } j \in \{r+1, \ldots, n\}.$$

Folglich ist durch

$$A = U \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} V^T$$

die gesuchte volle und damit auch reduzierte Singulärwertzerlegung gefunden.

In dem nächsten Satz fassen wir einige grundlegende Eigenschaften der Singulärwertzerlegung, die wir zum Teil eben schon bewiesen haben, zusammen. Den einfachen Beweis übergehen wir.

[2]Hierbei haben wir benutzt, dass $\text{Kern}\,(A) = \text{Kern}\,(A^TA)$. Beweis?

Satz 3.1 *Sei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ gegeben. Dann existiert eine volle bzw. reduzierte Singulärwertzerlegung*

$$A = U \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} V^T \qquad \text{bzw.} \qquad A = \hat{U}\hat{\Sigma}V^T.$$

Hierbei ist

$$U = (\ u_1 \ \cdots \ u_m\ ), \qquad V = (\ v_1 \ \cdots \ v_n\ ), \qquad \hat{U} = (\ u_1 \ \cdots \ u_n\ )$$

und

$$\hat{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n) \qquad \text{mit} \qquad \sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0,$$

wobei natürlich $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ orthogonal und die Spalten von $\hat{U} \in \mathbb{R}^{m\times n}$ orthonormiert sind, also $\hat{U}^T\hat{U} = I$. Ferner gilt:

1. *Es ist $\sigma_i = \lambda_i^{1/2}$, $i = 1, \ldots, n$, wobei $\lambda_1 \geq \cdots \geq \lambda_n$ die Eigenwerte von A^TA sind. Die Anzahl r positiver singulärer Werte stimmt mit dem Rang von A überein.*
2. *Es ist* $\operatorname{Bild}(A) = \operatorname{span}\{u_1, \ldots, u_r\}$ *und* $\operatorname{Kern}(A) = \operatorname{span}\{v_{r+1}, \ldots, v_n\}$.
3. *Es ist* $\operatorname{Bild}(A^T) = \operatorname{span}\{v_1, \ldots, v_r\}$ *und* $\operatorname{Kern}(A^T) = \operatorname{span}\{u_{r+1}, \ldots, u_m\}$.
4. *Es ist $\|A\|_2 = \sigma_1$ und $\|A\|_F = (\sigma_1^2 + \cdots + \sigma_r^2)^{1/2}$.*
5. *Es ist*
$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T.$$
6. *Ist $b \in \mathbb{R}^m$, so ist die eindeutige Lösung x_{LS} minimaler euklidischer Norm des linearen Ausgleichsproblems*

   (P) *Minimiere* $\|Ax - b\|_2, \quad x \in \mathbb{R}^n,$

   durch
$$x_{LS} := \sum_{i=1}^{r} \frac{u_i^T b}{\sigma_i} v_i$$
   gegeben.

Beispiel: Die Singulärwertzerlegung kann zur Bildkompression benutzt werden. Sei ein Bild mit $m \times n$ Pixeln gegeben. Man erhält eine Matrix $A \in \mathbb{R}^{m\times n}$ (die Bedingung $m \geq n$ ist unerheblich), indem als Eintrag an der Position (i, j) die Helligkeit des Bildes angegeben wird, etwa zwischen 0 (schwarz) und 1 (weiß). Will man das Bild bearbeiten oder übermitteln, so kann es sinnvoll sein, dass Bild zu komprimieren und nicht alle $m \cdot n$ Einträge zu benutzen. Hierzu mache man eine Singulärwertzerlegung $A = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$. In Aufgabe 1 kann bewiesen werden, dass bezüglich der Spektralnorm die Matrix $A_k := \sum_{i=1}^{k} \sigma_i u_i v_i^T$ für alle $k < r$ die beste Approximation an A bezüglich aller $m \times n$-Matrizen vom Rang k ist. Die Matrix A_k ist durch $(m + n)k$ Zahlen bestimmt, was i. allg. wesentlich weniger als mn ist. Bei J. W. Demmel (1997, S. 114–116) findet man ein Beispiel. □

3.3.2 Die Pseudoinverse

Die Pseudoinverse verallgemeinert den Begriff der Inversen einer nichtsingulären, quadratischen Matrix auf beliebige rechteckige Matrizen. Ist $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ und

$$A = U \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} V^T$$

mit orthogonalen $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ sowie

$$\hat{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n) \quad \text{mit} \quad \sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0$$

eine Singulärwertzerlegung von A, so nennt man

$$A^+ := V \begin{pmatrix} \hat{\Sigma}^+ & 0 \end{pmatrix} U^T \in \mathbb{R}^{n\times m} \qquad \text{mit} \quad \hat{\Sigma}^+ := \operatorname{diag}(1/\sigma_1, \ldots, 1/\sigma_r, 0, \ldots, 0) \in \mathbb{R}^{n\times n}$$

die (Moore-Penrose) *Pseudoinverse* von A. (Ist $n > m$, so erkläre man A^+ durch $A^+ := ((A^T)^+)^T$.) Die Wohldefiniertheit von A^+ folgt sofort daraus, dass A^+b für jedes $b \in \mathbb{R}^m$ die eindeutige Lösung minimaler euklidischer Norm zum linearen Ausgleichsproblem mit den Daten $\begin{pmatrix} A & b \end{pmatrix}$ ist. Hieraus folgt auch ohne weitere Rechnung, dass $A^+ = A^{-1}$ für nichtsinguläres quadratisches A, ferner $A^+ = (A^TA)^{-1}A^T$ für $\operatorname{Rang}(A) = n$. Natürlich hätten wir auch von der reduzierten Singulärwertzerlegung $A = \hat{U}\hat{\Sigma}V^T$ ausgehen und die Pseudoinverse durch $A^+ := V\hat{\Sigma}^+\hat{U}^T$ definieren können.

Einige Eigenschaften der Pseudoinversen werden in den Aufgaben angegeben. Wir verweisen insbesondere auf die Aufgaben 2a und 2b, deren Aussagen wir wiederholt ohne nähere Erläuterung benutzen werden.

3.3.3 MATLAB-Ergänzungen

Zunächst geben wir einen Teil der Informationen wieder, die man nach `svd` erhält:

```
svd    Singular value decomposition.
   [U,S,V] = svd(X) produces a diagonal matrix S, of the same
   dimension as X and with nonnegative diagonal elements in
   decreasing order, and unitary matrices U and V so that
   X = U*S*V'.

   S = svd(X) returns a vector containing the singular values.

   [U,S,V] = svd(X,0) produces the "economy size"
   decomposition. If X is m-by-n with m > n, then only the
   first n columns of U are computed and S is n-by-n.
   For m <= n, svd(X,0) is equivalent to svd(X).

   [U,S,V] = svd(X,'econ') also produces the "economy size"
   decomposition. If X is m-by-n with m >= n, then it is
   equivalent to svd(X,0). For m < n, only the first m columns
   of V are computed and S is m-by-m.
```

Um z. B. die eindeutige Lösung minimaler euklidischer Norm zu einem linearen Ausgleichsproblem mit den Daten (A, b) mit Hilfe einer Singulärwertzerlegung von A zu berechnen, kann man entweder `pinv(A)*b` berechnen, oder die folgende Funktion benutzen:

```
function xMEN=MinEuLS(A,b);
%*****************************************************************
%Pre:      A m-by-n with m>=n
%          b m-by-1
%Post:     xMEN solution with minimal euclidean norm of
%          the least square problem to minimize norm(A*x-b);
%*****************************************************************
[m,n]=size(A);
[U,S,V]=svd(A,0);
s=diag(S);
tol=m*max(s)*eps;
r=sum(s>tol);temp=U'*b;
xMEN=V(:,1:r)*(temp(1:r)./s(1:r));
```

Es ist klar, wie diese Funktion arbeitet. Zunächst wird die reduzierte Singulärwertzerlegung berechnet, danach der Rang r. Dann wird die Darstellung

$$x_{\mathrm{MEN}} = \sum_{i=1}^{r} \frac{u_i^T b}{\sigma_i} v_i$$

benutzt. Bei einfachen Beispielen, bei denen A und b z. B. durch `A=rand(7,5)` und `b=rand(7,1)` erzeugt (mit Wahrscheinlichkeit 1 ist der Rang von A dann allerdings maximal) sind, haben wir nach `format long` keinen Unterschied zwischen `MinEuLS(A,b)` sowie `A\b` und `pinv(A)*b` feststellen können.

Die Pseudoinverse A^+ einer Matrix A kann in Matlab mit Hilfe von `pinv(A)` berechnet werden. Als Information erhält man:

```
 pinv   Pseudoinverse.
    X = pinv(A) produces a matrix X of the same dimensions
    as A' so that A*X*A = A, X*A*X = X and A*X and X*A
    are Hermitian. The computation is based on SVD(A) and any
    singular values less than a tolerance are treated as zero.
    The default tolerance is MAX(SIZE(A)) * NORM(A) * EPS(class(A)).

    pinv(A,TOL) uses the tolerance TOL instead of the default.
```

Da `pinv` keine built-in function ist, kann man sie sich mit Hilfe von `type pinv` oder `edit pinv` ansehen. Man wird erkennen, dass zunächst eine reduzierte Singulärwertzerlegung und anschließend Rang (A) berechnet wird (wozu die Anzahl der Singulärwerte gezählt wird, die größer als eine Berechnete oder vorgegebene Toleranz ist), danach wird schließlich die Definition benutzt. Ist `tol` eine vorgegebene Toleranz, so wird `X=pinv(A)` bis auf Ausartungsfälle durch

```
[U,S,V] = svd(A,0);
s=diag(S);r = sum(s > tol);
s = diag(ones(r,1)./s(1:r));
X = V(:,1:r)*s*U(:,1:r)';
```

berechnet.

3.3.4 Aufgaben

1. Sei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ und $r := \text{Rang}\,(A)$ gegeben, ferner sei mit den üblichen Bezeichnungen $A = U\Sigma V^T$ eine Singulärwertzerlegung von A. Mit $\mathbb{R}_k^{m\times n}$ werde die Menge der reellen $m \times n$-Matrizen vom Rang k bezeichnet. Für jedes $k < r$ ist dann

$$\sigma_{k+1} = \|A - A_k\|_2 = \min_{X\in\mathbb{R}_k^{m\times k}} \|A - X\|_2,$$

   wobei

$$A_k := \sum_{i=1}^{k} \sigma_i u_i v_i^T.$$

2. Sei $A \in \mathbb{R}^{m\times n}$ und A^+ die zugehörige Pseudoinverse. Man zeige:

   (a) Es ist

$$AA^+A = A, \quad A^+AA^+ = A^+, \quad (AA^+)^T = AA^+, \quad (A^+A)^T = A^+A.$$

   (b) Sei $L \subset \mathbb{R}^n$ ein linearer Teilraum. Eine lineare Abbildung $P_L\colon \mathbb{R}^n \longrightarrow L \subset \mathbb{R}^n$ heißt *orthogonale Projektion* des $\mathbb{R}^n$ auf L, wenn $(I - P_L)x \perp L$ für alle $x \in \mathbb{R}^n$. Dann ist AA^+ orthogonale Projektion des $\mathbb{R}^m$ auf Bild (A) und A^+A orthogonale Projektion des $\mathbb{R}^n$ auf Bild (A^T).

3. Sei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ und $A^+ \in \mathbb{R}^{n\times m}$ die zugehörige Pseudoinverse. Man zeige, dass A^+ Lösung der Aufgabe

$$\text{Minimiere} \quad \|AX - I\|_F, \quad X \in \mathbb{R}^{n\times m},$$

   ist.

4. Sei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ und $r := \text{Rang}\,(A)$ gegeben. Man definiere $B\colon (0,\infty) \longrightarrow \mathbb{R}^{n\times m}$ durch

$$B(\lambda) := (A^TA + \lambda I)^{-1}A^T, \qquad \lambda > 0.$$

   Man zeige, dass

$$\|B(\lambda) - A^+\|_2 = \frac{\lambda}{\sigma_r(\sigma_r^2 + \lambda)},$$

   wobei σ_r kleinster singulärer Wert von A ist. Insbesondere gilt $\lim_{\lambda\to 0+} B(\lambda) = A^+$.

5. Seien $A \in \mathbb{R}^{m\times n}$ mit $\text{Rang}\,(A) = m < n$ und $b \in \mathbb{R}^m$ gegeben. Die Aufgabe

   (P) $$\text{Minimiere} \quad \|Ax - b\|_2, \quad x \in \mathbb{R}^n,$$

   heißt ein *unterbestimmtes lineares Ausgleichsproblem*. Man zeige, dass (P) eine $(n-m)$-dimensionale affin lineare Lösungsmenge hat, dass (P) genau eine Lösung minimaler euklidischer Norm besitzt und überlege sich, wie man diese mit Hilfe einer QR-Zerlegung von A^T berechnen kann.

6. Sei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ gegeben. Man zeige, dass $\|I - AA^+\|_2 = \min(1, m-n)$.

3.4 Die Berechnung der Singulärwertzerlegung

I. allg. wird über die Berechnung der Singulärwertzerlegung einer Matrix $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ erst dann etwas ausgesagt, wenn entsprechende Vorkenntnisse über die numerische Behandlung von (insbesondere symmetrischen) Eigenwertaufgaben vorliegen. Wir wollen versuchen (ähnlich wie Å. BJÖRK (1996, S. 81 ff.)) die notwendigen Vorkenntnisse hier schon bereit zu stellen.

Natürlich könnte man verhältnismäßig naiv folgendermaßen vorgehen:

- Bilde $A^TA \in \mathbb{R}^{n\times n}$ und berechne die (nichtnegativen) Eigenwerte $\lambda_1 \geq \cdots \geq \lambda_n$ dieser Matrix sowie ein vollständiges Orthonormalsystem $\{v_1, \ldots, v_n\}$ zugehöriger Eigenvektoren. Mit

$$\Lambda := \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \qquad V := \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix}$$

  ist dann $A^TA = V\Lambda V^T$.

- Die singulären Werte von A sind durch $\sigma_i := \lambda_i^{1/2}$, $i = 1, \ldots, n$, gegeben. Sei

$$\hat{\Sigma} := \operatorname{diag}(\sigma_1, \ldots, \sigma_n).$$

- Man wende das Householder-Verfahren mit Spaltenpivotisierung auf AV an und berechne eine orthogonale Matrix $\tilde{U} \in \mathbb{R}^{m\times m}$, eine Permutationsmatrix $\Pi \in \mathbb{R}^{n\times n}$ und eine obere Dreiecksmatrix $\hat{R} \in \mathbb{R}^{n\times n}$ mit nichtnegativen Diagonalelementen derart, dass

$$\tilde{U}^TAV\Pi = R = \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix}.$$

  Dann ist

$$\hat{R}^T\hat{R} = R^TR = \Pi^TV^TA^T\tilde{U}\tilde{U}^TAV\Pi = \Pi^T\Lambda\Pi$$

  eine Diagonalmatrix. Notwendigerweise ist die obere Dreiecksmatrix $\hat{R}$ sogar eine Diagonalmatrix und $\hat{R} = \Pi^T\hat{\Sigma}\Pi$. Folglich ist

$$A = \tilde{U}\begin{pmatrix} \hat{R} \\ 0 \end{pmatrix}\Pi^TV^T = \tilde{U}\begin{pmatrix} \Pi^T\hat{\Sigma}\Pi \\ 0 \end{pmatrix}\Pi^TV^T = \tilde{U}\begin{pmatrix} \Pi^T & 0 \\ 0 & I \end{pmatrix}\begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix}V^T$$

  die gesuchte Singulärwertzerlegung von A.

Mit Hilfe eines Verfahrens zur Berechnung von Eigenwerten und zugehörigen (orthonormierten) Eigenvektoren einer symmetrischen Matrix wäre dieser Zugang *theoretisch* durchführbar. In der *Praxis* verbietet er sich aber, weil er nicht effizient und stabil genug ist. Dies liegt vor allem daran, dass es nicht ratsam ist, A^TA zu bilden und als Grundlage für die weiteren Rechnungen zu nehmen.

Wir werden das Verfahren schildern, das heute für dichtbesetzte Matrizen "method of choice" ist, und auf Golub-Kahan (1965) und Golub-Reinsch (1970) zurückgeht.

Wir werden uns auf Methoden zur Berechnung einer *reduzierten* Singulärwertzerlegung beschränken, was für alle praktischen Anforderungen ausreichend ist, und nur in einer kurzen Bemerkung im Anschluss an Lemma 4.1 darauf eingehen, wie man mit den vorgestellten Methoden auch eine *volle* Singulärwertzerlegung berechnen kann.

3.4.1 Transformation auf obere Bidiagonalform

Der erste Schritt im Golub-Reinsch-Verfahren ist ein Reduktionsschritt, der von Golub-Kahan stammt. In ihm wird die Aufgabe, eine Singulärwertzerlegung der i. allg. voll besetzten Matrix $A \in \mathbb{R}^{m\times n}$ (mit $m \geq n$) auf die Berechnung der Singulärwertzerlegung einer oberen Bidiagonalmatrix zurückgeführt, also einer $n \times n$-Matrix, die außerhalb der Diagonalen und der oberen Nebendiagonale verschwindende Einträge besitzt. Grundlage für den Reduktionsschritt ist das folgende einfache Lemma.

Lemma 4.1 *Sei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ gegeben, ferner seien*

$$P = \begin{pmatrix} p_1 & \cdots & p_m \end{pmatrix} \in \mathbb{R}^{m\times m}, \qquad Q \in \mathbb{R}^{n\times n}$$

orthogonale Matrizen derart, dass

$$P^T A Q = \begin{pmatrix} B \\ 0 \end{pmatrix} \qquad \text{mit} \quad B \in \mathbb{R}^{n\times n}.$$

Dann gilt:

1. *Die Matrizen $A \in \mathbb{R}^{m\times n}$ und $B \in \mathbb{R}^{n\times n}$ haben dieselben singulären Werte.*
2. *Sei*
$$B = U_B \mathrm{diag}\,(\sigma_1, \ldots, \sigma_n) V_B^T$$
mit orthogonalen $U_B \in \mathbb{R}^{n\times n}$, $V_B \in \mathbb{R}^{n\times n}$ eine Singulärwertzerlegung von B. Dann ist durch
$$A = \hat{U}\mathrm{diag}\,(\sigma_1, \ldots, \sigma_n) V^T \qquad \text{mit} \qquad \hat{U} := \begin{pmatrix} p_1 & \cdots & p_n \end{pmatrix} U_B, \quad V := Q V_B$$
eine (reduzierte) Singulärwertzerlegung von A gegeben.

Beweis: Der Beweis erfolgt durch einfaches Nachrechnen. So ist z. B.

$$B^T B = \begin{pmatrix} B \\ 0 \end{pmatrix}^T \begin{pmatrix} B \\ 0 \end{pmatrix} = (P^T A Q)^T (P^T A Q) = Q^T (A^T A) Q,$$

also ist $B^T B$ orthogonal ähnlich zu $A^T A$. Daher stimmen die Eigenwerte von $A^T A$ und $B^T B$ und damit auch die singulären Werte von A und B überein. Wegen

$$A = P \begin{pmatrix} B \\ 0 \end{pmatrix} Q^T = \begin{pmatrix} p_1 & \cdots & p_n \end{pmatrix} B Q^T = \begin{pmatrix} p_1 & \cdots & p_n \end{pmatrix} U_B \mathrm{diag}\,(\sigma_1, \ldots, \sigma_n)(Q V_B)^T$$

folgt der Rest der Behauptungen. □

Bemerkung: Im letzten Lemma 4.1 ist angegeben, wie die Berechnung einer reduzierten Singulärwertzerlegung einer Matrix $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ auf die Berechnung einer Singulärwertzerlegung (reduziert oder voll, das spielt bei quadratischen Matrizen keine Rolle) einer Matrix $B \in \mathbb{R}^{n\times n}$ zurückgeführt werden kann. Die Berechnung einer *vollen* Singulärwertzerlegung von $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ ist ebenso einfach. Denn ist

$$A = P \begin{pmatrix} B \\ 0 \end{pmatrix} Q^T$$

mit orthogonalen $P \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$, und

$$B = U_B \text{diag}\,(\sigma_1, \ldots, \sigma_n) V_B^T$$

mit orthogonalen $U_B \in \mathbb{R}^{n\times n}$, $V_B \in \mathbb{R}^{n\times n}$ eine Singulärwertzerlegung von $B \in \mathbb{R}^{n\times n}$, so ist durch

$$\begin{aligned} A &= P \begin{pmatrix} U_B \text{diag}\,(\sigma_1, \ldots, \sigma_n) V_B^T \\ 0 \end{pmatrix} Q^T \\ &= \underbrace{P \begin{pmatrix} U_B & 0 \\ 0 & I_{m-n} \end{pmatrix}}_{=:U} \begin{pmatrix} \text{diag}\,(\sigma_1, \ldots, \sigma_n) \\ 0 \end{pmatrix} (\underbrace{Q V_B}_{=:V})^T \\ &= U \begin{pmatrix} \text{diag}\,(\sigma_1, \ldots, \sigma_n) \\ 0 \end{pmatrix} V^T \end{aligned}$$

eine volle Singulärwertzerlegung von A gegeben. □

Der Reduktionsschritt besteht darin, Householder-Matrizen $P_1, \ldots, P_n \in \mathbb{R}^{m\times m}$ und $Q_1, \ldots, Q_{n-2} \in \mathbb{R}^{n\times n}$ derart zu bestimmen, dass

$$P_n \cdots P_1 A Q_1 \cdots Q_{n-2} = \begin{pmatrix} B \\ 0 \end{pmatrix}$$

mit einer oberen Bidiagonalmatrix

$$B = \begin{pmatrix} d_1 & f_2 & 0 & \cdots & 0 \\ & d_2 & f_3 & \ddots & \vdots \\ & & \ddots & \ddots & 0 \\ & & & \ddots & f_n \\ & & & & d_n \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

Grundlegend ist wieder, dass man einen vom Nullvektor verschiedenen Vektor durch Multiplikation mit einer geeigneten Householder-Matrix in ein Vielfaches des ersten Einheitsvektors überführen kann. Im Prinzip sieht dieses *Bidiagonalisierungsverfahren* folgendermaßen aus:

- Input: Gegeben $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$.

- Für $k = 1, \ldots, n$:

  Bestimme Householder-Matrix $\overline{P}_k \in \mathbb{R}^{(m-k+1)\times(m-k+1)}$ mit

  $$\overline{P}_k(a_{kk}, \ldots, a_{mk})^T = (*, 0, \ldots, 0)^T.$$

  Setze $P_k := \text{diag}\,(I_{k-1}, \overline{P}_k)$, berechne $A := P_k A$.

  Falls $k \leq n-2$, dann:

  Bestimme Householder-Matrix $\overline{Q}_k \in \mathbb{R}^{(n-k)\times(n-k)}$ mit

  $$\overline{Q}_k(a_{k,k+1}, \ldots, a_{kn})^T = (*, 0, \ldots, 0)^T.$$

  Setze $Q_k := \text{diag}\,(I_k, \overline{Q}_k)$, berechne $A := AQ_k$.

- Output: A wird mit einer oberen Bidiagonalmatrix überschrieben.

In einer Implementation des Verfahrens wird man das $m \times n$-Feld A (sowie zwei Felder der Länge n bzw. $n-2$) dazu benutzen, um die relevanten Informationen über die Matrizen $P_1, \ldots, P_n$ und $Q_1, \ldots, Q_{n-2}$ aufzunehmen, während die Haupt- und Superdiagonalelemente der Bidiagonalmatrix B in Feldern d und f gespeichert werden. Anschließend kann man die ersten n Spalten von $P_1 \cdots P_n$ und das Produkt $Q_1 \cdots Q_{n-2}$ bilden und diese in $m \times n$- bzw. $n \times n$-Feldern $\hat{U}$ bzw. V speichern. Ist dies geschehen, so hat man zu der Ausgangsmatrix $A \in \mathbb{R}^{m\times n}$ Matrizen $\hat{U} \in \mathbb{R}^{m\times n}$, $V \in \mathbb{R}^{n\times n}$ mit $\hat{U}^T\hat{U} = I$, $V^TV = I$ sowie eine obere Bidiagonalmatrix $B \in \mathbb{R}^{n\times n}$ mit $A = \hat{U}BV^T$ bestimmt. Wäre B sogar eine Diagonalmatrix, so hätten wir die gesuchte (reduzierte) Singulärwertzerlegung von A schon gewonnen.

An einer 4×3-Matrix veranschaulichen wir uns den Prozess, eine gegebene Matrix durch Multiplikation von links und rechts mit Householder-Matrizen auf obere Bidiagonal-Gestalt zu transformieren. Festbleibende Elemente werden mit $\bullet$, sich verändernde durch $*$ gekennzeichnet.

$$\begin{pmatrix} \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \end{pmatrix} \xrightarrow{P_1} \begin{pmatrix} * & * & * \\ & * & * \\ & * & * \\ & * & * \end{pmatrix} \xrightarrow{Q_1} \begin{pmatrix} \bullet & * & \\ & * & * \\ & * & * \\ & * & * \end{pmatrix} \xrightarrow{P_2} \begin{pmatrix} \bullet & \bullet & \\ & * & * \\ & & * \\ & & * \end{pmatrix} \xrightarrow{P_3} \begin{pmatrix} \bullet & \bullet & \\ & \bullet & \bullet \\ & & * \\ & & \end{pmatrix}$$

Wir haben jetzt das Problem darauf reduziert, eine Singulärwertzerlegung der oberen Bidiagonalmatrix $B \in \mathbb{R}^{n\times n}$ zu berechnen. Die Hauptdiagonalelemente von B seien $d_1, \ldots, d_n$, die Superdiagonalelemente seien $f_2, \ldots, f_n$. Die Matrix $C := B^TB$ (die wir nicht etwa explizit bilden!) ist eine (symmetrische) $n \times n$-Tridiagonalmatrix, ihre Diagonal- bzw. Nebendiagonalelemente sind

$$c_{11} = d_1^2, \quad c_{ii} = d_i^2 + f_i^2 \quad (i = 2, \ldots, n), \quad c_{i,i+1} = c_{i+1,i} = d_i f_{i+1} \quad (i = 1, \ldots, n-1).$$

Daher ist B^TB genau dann unreduziert (alle Nebendiagonalelemente sind von Null verschieden), wenn $d_1 \cdots d_{n-1} \neq 0$ und $f_2 \cdots f_n \neq 0$. Ist dies nicht der Fall, so kann die Berechnung eine Singulärwertzerlegung der oberen Bidiagonalmatrix $B \in \mathbb{R}^{n\times n}$ auf die von zwei niederdimensionalen oberen Bidiagonalmatrizen B_1 und B_2 zurückgeführt werden. Grundlage hierfür ist das folgende Lemma.

Lemma 4.2 *Die Matrix $B \in \mathbb{R}^{n\times n}$ sei eine obere Bidiagonalmatrix mit Hauptdiagonalelementen $d_1, \ldots, d_n$ sowie Superdiagonalelementen $f_2, \ldots, f_n$.*

1. *Sei $f_{k+1} = 0$ für ein $k \in \{1, \ldots, n-1\}$, also*

$$B = \begin{pmatrix} B_1 & 0 \\ 0 & B_2 \end{pmatrix}$$

   zerfallend in zwei obere Bidiagonalmatrizen $B_1 \in \mathbb{R}^{k\times k}$, $B_2 \in \mathbb{R}^{(n-k)\times(n-k)}$. Dann lässt sich die Berechnung eine Singulärwertzerlegung von B auf die Berechnung von Singulärwertzerlegungen von B_1 und B_2 zurückführen.

2. *Ist $d_k = 0$ und $f_{k+1} \neq 0$ für ein $k \in \{1, \ldots, n-1\}$, so kann man $n-k$ Givens-Rotationen $G_{k,k+1}, \ldots, G_{k,n}$ derart bestimmen, dass $B_+ := G_{k,n} \cdots G_{k,k+1} B$ eine obere Bidiagonalmatrix mit $(B_+)_{k,k+1} = 0$ ist.*

Beweis: Seien

$$B_1 = U_1 \mathrm{diag}\,(\tau_1, \ldots, \tau_k) V_1^T, \qquad B_2 = U_2 \mathrm{diag}\,(\tau_{k+1}, \ldots, \tau_n) V_2^T$$

Singulärwertzerlegungen von B_1 bzw. B_2. Man bestimme eine Permutationsmatrix $\Pi \in \mathbb{R}^n$ mit

$$\mathrm{diag}\,(\sigma_1, \ldots, \sigma_n) = \Pi^T \mathrm{diag}\,(\tau_1, \ldots, \tau_n)\Pi, \qquad \sigma_1 \geq \cdots \geq \sigma_n.$$

Dann ist

$$B = \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix} \Pi \mathrm{diag}\,(\sigma_1, \ldots, \sigma_n) \Pi^T \begin{pmatrix} V_1 & 0 \\ 0 & V_2 \end{pmatrix}^T$$

die gesuchte Singulärwertzerlegung von B.

Die Idee für den Beweis des zweiten Teiles besteht darin, das in der k-ten Zeile störende Superdiagonalelement f_{k+1} sukzessive durch Multiplikation mit einer Givens-Rotation $G_{k,j}$, $j = k+1, \ldots, n$, in der Zeile nach rechts zu schieben und schließlich aus der Matrix zu drängen. Statt eines formalen Beweises machen wir uns die Strategie anhand einer 5×5-Matrix klar, bei der das zweite Diagonalelenent verschwindet. Wieder werden bei einer Transformation festbleibende Elemente durch $\bullet$, sich verändernde durch $*$ gekennzeichnet.

$$B = \begin{pmatrix} \bullet & \bullet & & & \\ & 0 & \bullet & & \\ & & \bullet & \bullet & \\ & & & \bullet & \bullet \\ & & & & \bullet \end{pmatrix} \xrightarrow{G_{23}} \begin{pmatrix} \bullet & \bullet & & & \\ & 0 & 0 & * & \\ & & * & * & \\ & & & \bullet & \bullet \\ & & & & \bullet \end{pmatrix} \xrightarrow{G_{24}} \begin{pmatrix} \bullet & \bullet & & & \\ & 0 & 0 & 0 & * \\ & & \bullet & \bullet & \\ & & & * & * \\ & & & & \bullet \end{pmatrix} \xrightarrow{G_{25}}$$

und man erhält schließlich

$$B_+ = \begin{pmatrix} \bullet & \bullet & & & \\ & 0 & & & \\ & & \bullet & \bullet & \\ & & & \bullet & \bullet \\ & & & & * \end{pmatrix}.$$

Die Givens-Rotation $G_{k,j}(c,s)$ ist so zu bestimmen, dass $c^2+s^2=1$ und

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix}\begin{pmatrix} f \\ d_j \end{pmatrix} = \begin{pmatrix} 0 \\ * \end{pmatrix},$$

wobei f der aktuelle Eintrag in der Position (k,j) ist. Es besteht also ein kleiner Unterschied zu der in Unterabschnitt 3.2.3 (Givens-Verfahren) angegebenen Funktion "givrot". □

Jetzt geben wir den eben skizzierten Algorithmus genauer an.

- Input: Sei $B \in \mathbb{R}^{n\times n}$ eine obere Bidiagonalmatrix mit $d_1,\ldots,d_n$ und $f_2,\ldots,f_n$ als Haupt- bzw. Superdiagonalelementen. Für ein $k \in \{1,\ldots,n-1\}$ sei $d_k = 0$, $f_{k+1} \neq 0$. Gegeben sei ferner eine Matrix $\hat{U} = (u_{ij}) \in \mathbb{R}^{m\times n}$. Mit der Ausgangsmatrix $A \in \mathbb{R}^{m\times n}$ sei z. B. $A = \hat{U}BV^T$.

- Setze $c := 0, \quad s := 1.$

- Für $j = k+1,\ldots,n$:

  Berechne $f := sf_j, \quad f_j := cf_j, \quad u := (f^2+d_j^2)^{1/2}.$

  Berechne $c := d_j/u, \quad s := -f/u, \quad d_j := u.$

  Für $i = 1,\ldots,m$:

  Berechne $(u_{ik},u_{ij}) = (u_{ik},u_{ij})\begin{pmatrix} c & -s \\ s & c \end{pmatrix}.$

- Output: Die Felder d und f enthalten die Haupt- bzw. Superdiagonalelemente einer oberen Bidiagonalmatrix $B_+ := G_{k,n}\cdots G_{k,k+1}B$, die also aus B durch Multiplikation von links mit Givens-Rotationen $G_{k,k+1},\ldots,G_{k,n}$ hervorgeht. Hierbei ist $f_{k+1}=0$, das k-te Superdiagonalelement von B_+ verschwindet also, B_+ zerfällt also. Ferner wird $\hat{U} \in \mathbb{R}^{m\times n}$ mit $\hat{U}_+ := \hat{U}G_{k,k+1}^T\cdots G_{k,n}^T$ überschrieben. War vorher $A = \hat{U}BV^T$, so ist nach Abschluss $A = \hat{U}_+B_+V^T$.

Wir haben jetzt das Problem, eine (reduzierte) Singulärwertzerlegung der Matrix $A \in \mathbb{R}^{m\times n}$ zu finden darauf reduziert, eine Singulärwertzerlegung einer oberen Bidiagonalmatrix $B \in \mathbb{R}^{n\times n}$ zu bestimmen, für die B^TB unreduziert ist.

3.4.2 Das Verfahren von Golub-Reinsch

Sei jetzt $B \in \mathbb{R}^{n\times n}$ bei gegebenem $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ eine obere Bidiagonalmatrix, für die (die Tridiagonalmatrix) $C := B^TB$ unreduziert ist, ferner gelte die Darstellung $A = \hat{U}BV^T$ mit $\hat{U} \in \mathbb{R}^{m\times n}$, $V \in \mathbb{R}^{n\times n}$ und $\hat{U}^T\hat{U} = I$, $V^TV = I$. Ziel ist es, eine neue Darstellung $A = \hat{U}_+B_+V_+^T$ zu berechnen, bei der wieder B_+ eine obere Diagonalmatrix ist und $\hat{U}_+ \in \mathbb{R}^{m\times n}$, $V_+ \in \mathbb{R}^{n\times n}$ orthonormierte Spalten hat bzw. eine orthogonale Matrix, aber zusätzlich zumindestens das letzte Superdiagonalelement von B_+ wesentlich kleiner als das entsprechende in B ist. Grundlage für die Vorgehensweise ist das folgende Ergebnis, das wir allgemeiner formulieren als wir es jetzt gleich benötigen. Wir

erinnern vorher daran, dass eine *obere Hessenberg-Matrix* eine Matrix ist, bei der alle Einträge unterhalb der Subdiagonalen verschwinden. Man nennt sie *unreduziert*, wenn kein Subdiagonalelement verschwindet. Eine symmetrische obere Hessenberg-Matrix ist natürlich eine Tridiagonalmatrix.

Satz 4.3 *Seien*

$$Q = \begin{pmatrix} q_1 & \cdots & q_n \end{pmatrix} \in \mathbb{R}^{n\times n}, \qquad P = \begin{pmatrix} p_1 & \cdots & p_n \end{pmatrix} \in \mathbb{R}^{n\times n}$$

orthogonale Matrizen, die eine Matrix $C \in \mathbb{R}^{n\times n}$ jeweils in eine obere Hessenberg-Matrix $H := Q^TCQ$ bzw. $G := P^TCP$ transformieren, wobei G unreduziert sei. Ist $q_1 = \pm p_1$, stimmen die ersten Spalten von Q und P also (eventuell bis auf einen Faktor -1) überein, so ist auch H eine unreduzierte obere Hessenberg-Matrix und Q und P sowie H und G sind im wesentlichen gleich, d. h. $D := P^TQ$ ist eine orthogonale Diagonalmatrix (besitzt also nur $+1$ oder -1 als Diagonalelemente) und $H = D^TGD$.

Beweis: Man definiere die orthogonale Matrix

$$D := P^TQ = \begin{pmatrix} d_1 & \cdots & d_n \end{pmatrix}.$$

Da $q_1 = \pm p_1$ vorausgesetzt wurde, ist

$$d_1 = P^TQe_1 = P^Tq_1 = \pm P^Tp_1 = \pm e_1,$$

die erste Spalte von D also (eventuell bis auf das Vorzeichen) der erste Einheitsvektor. Angenommen, es sei $d_i = \pm e_i$, $i = 1, \ldots, k$, und $h_{i+1,i} \neq 0$, $i = 1, \ldots, k-1$. Dies ist für $k = 1$ sicherlich richtig. Aus

$$GD = P^TCPP^TQ = P^TCQ = P^TQQ^TCQ = P^TQH = DH$$

erhält man beim Vergleich der k-ten Spalte $GDe_k = DHe_k$, dass

$$(*) \qquad \pm Ge_k = GDe_k = DHe_k = h_{k+1,k}d_{k+1} + \sum_{i=1}^{k} h_{ik}d_i.$$

Multipliziert man diese Gleichung von links mit $d_i^T = \pm e_i^T$, so erhält man $\pm g_{ik} = h_{ik}$, $i = 1, \ldots, k$. Hierbei gilt natürlich entweder für alle i das $-$ oder für alle i das $+$-Zeichen. Eine erneute Betrachtung von $(*)$ liefert

$$\pm \begin{pmatrix} g_{1k} \\ \vdots \\ g_{kk} \\ g_{k+1,k} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = h_{k+1,k}d_{k+1} \pm \begin{pmatrix} g_{1k} \\ \vdots \\ g_{kk} \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

bzw. $g_{k+1,k}e_{k+1} = h_{k+1,k}d_{k+1}$. Da G unreduziert, ist $g_{k+1,k} \neq 0$ und dann auch $h_{k+1,k} \neq 0$. Da $\|e_{k+1}\|_2 = \|d_{k+1}\|_2 = 1$, ist $|h_{k+1,k}| = |g_{k+1,k}|$ und folglich auch $d_{k+1} = \pm e_{k+1}$. Damit ist gezeigt, dass auch H unreduziert und D eine orthogonale Diagonalmatrix ist. Der Satz ist bewiesen. □

Jetzt beschreiben wir einen Schritt des QR-Verfahrens zur Berechnung der Eigenwerte der symmetrischen Matrix $C := B^TB \in \mathbb{R}^{n\times n}$, sagen aber gleich, dass im Verfahren von Golub-Reinsch die Matrix B^TB nur implizit gebildet wird und dort eine Bidiagonal-Matrix B_+ berechnet wird, die dieselben singulären Werte wie B besitzt und von der man sich erhofft, dass das letzte Superdiagonalelement sehr viel kleiner geworden ist:

- Bestimme den *Shift-Parameter* $\sigma \in \mathbb{R}$ als den Eigenwert des unteren 2×2-Blockes

$$\begin{pmatrix} c_{n-1,n-1} & c_{n-1,n} \\ c_{n-1,n} & c_{nn} \end{pmatrix}$$

  von C, der c_{nn} am nächsten liegt.

- Bestimme QR-Zerlegung $C - \sigma I = QR$ und berechne anschließend

$$C_+ := RQ + \sigma I = Q^TCQ.$$

Wie wir schon betonten, wird im Verfahren von Golub-Reinsch die Matrix B^TB nicht explizit gebildet. Der erste Schritt zur Berechnung von B_+ sieht folgendermaßen aus:

- Input: Sei $B \in \mathbb{R}^{n\times n}$ eine obere Bidiagonalmatrix mit den Haupt- bzw. Nebendiagonalelementen $d_1, \ldots, d_n$ bzw. $f_2, \ldots, f_n$. Es sei $d_1 \cdots d_{n-1} \neq 0$ und $f_2 \cdots f_n \neq 0$, d. h. $C := B^TB$ eine unreduzierte, symmetrische Tridiagonalmatrix. Weiter nehmen wir an, dass Matrizen $\hat{U} \in \mathbb{R}^{m\times n}$, $V \in \mathbb{R}^{n\times n}$ mit $\hat{U}^T\hat{U} = I$, $V^TV = I$ und $A = \hat{U}BV^T$ gegeben sind.

- Bestimme den Shift-Parameter $\sigma \in \mathbb{R}$ als den Eigenwert des unteren 2×2-Blockes

$$\begin{pmatrix} c_{n-1,n-1} & c_{n-1,n} \\ c_{n-1,n} & c_{nn} \end{pmatrix} = \begin{pmatrix} d_{n-1}^2 + f_{n-1}^2 & d_{n-1}f_n \\ d_{n-1}f_n & d_n^2 + f_n^2 \end{pmatrix}$$

  in C, der $c_{nn} = d_n^2 + f_n^2$ am nächsten liegt.

- Bestimme Givens-Rotation $T_{12} = T_{12}(c, s)$ mit

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} d_1^2 - \sigma \\ d_1f_2 \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}.$$

  Berechne $B := BT_{12}^T$, $V := VT_{12}^T$.

Nach Konstruktion ist $T_{12}(B^TB - \sigma I)e_1$, also die erste Spalte von $T_{12}(B^TB - \sigma I)$, ein Vielfaches des ersten Einheitsvektors. Daher stimmt die erste Spalte von T_{12}^T mit der ersten Spalte von Q, dem orthogonalen Anteil in einer QR-Zerlegung $B^TB - \sigma I = QR$ von $B^TB - \sigma I$, (eventuell bis auf den Faktor -1) überein.

In der transformierten Matrix BT_{12}^T verändert sich nur der obere 2×2-Block, dieser ist nach der Transformation gegeben durch

$$\begin{pmatrix} d_1 & f_2 \\ 0 & d_2 \end{pmatrix} \begin{pmatrix} c & -s \\ s & c \end{pmatrix} = \begin{pmatrix} d_1 c + f_2 s & -d_1 s + f_2 c \\ d_2 s & d_2 c \end{pmatrix}.$$

In BT_{12}^T wird die obere Bidiagonalgestalt also nur durch das Element in der Position $(2,1)$ gestört. Die Idee besteht jetzt darin, dieses störende Element durch Multiplikation von links mit einer geeigneten Givens-Rotation S_{12} zu annullieren. Die neuen ersten beiden Zeilen sind eine Linearkombination der alten, so dass nach der Transformation ein die Bidiagonal-Gestalt störendes Element in der Position $(1,3)$ steht. Dieses Element wiederum wird durch Multiplikation von rechts mit einer Givens-Rotation T_{23}^T annulliert und auf die Position $(3,2)$ gedrängt. Nun ist klar, dass man durch abwechselnde Multiplikaton von links und rechts mit geeigneten Givens-Rotationen das die Bidiagonal-Gestalt störende Element "in Rössel-Sprüngen" aus der Matrix drängen kann. Genauer sehen diese weiteren Schritte folgendermaßen aus:

- Für $k = 1, \ldots, n-1$:

  Bestimme Givens-Rotation $S_{k,k+1} = S_{k,k+1}(c,s)$ mit

  $$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} b_{kk} \\ b_{k+1,k} \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}.$$

  Berechne $B := S_{k,k+1} B$, $\hat{U} := \hat{U} S_{k,k+1}^T$.

  Falls $k < n-1$, dann:

  Bestimme Givens-Rotation $T_{k+1,k+2} = T_{k+1,k+2}(c,s)$ mit

  $$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} b_{k,k+1} \\ b_{k,k+2} \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}.$$

  Berechne $B := BT_{k+1,k+2}^T$, $V := VT_{k+1,k+2}^T$.

- Output: Die Ausgangsmatrix $B \in \mathbb{R}^{n \times n}$ ist mit der oberen Bidiagonalmatrix

  $$B_+ := S_{n-1,n} \cdots S_{12} B T_{12}^T \cdots T_{n-1,n}^T$$

  überschrieben. Ferner sind die Matrizen $\hat{U} \in \mathbb{R}^{m \times n}$, $V \in \mathbb{R}^{n \times n}$ mit

  $$\hat{U}_+ := \hat{U} S_{12}^T \cdots S_{n-1,n}^T, \qquad V_+ := V T_{12}^T \cdots T_{n-1,n}^T$$

  überschrieben. Daher ist $\hat{U}_+ B_+ V_+^T = \hat{U} B V^T$ und $\hat{U}_+^T \hat{U}_+ = I$, $V_+^T V_+ = I$.

Nun wollen wir uns davon überzeugen, dass es sich bei dem obigen Verfahren im wesentlichen um die Realisierung eines auf $B^T B$ angewandten QR-Schrittes mit dem Shift-Parameter $\sigma \in \mathbb{R}$ handelt. Mit den eben eingeführten Bezeichnungen ist

$$B_+^T B_+ = (T_{n-1,n} \cdots T_{12})(B^T B)(T_{n-1,n} \cdots T_{12})^T.$$

Wegen

$$(T_{n-1,n}\cdots T_{12})^T e_1 = T_{12}^T\cdots T_{n-1,n}^T e_1 = T_{12}^T e_1$$

ist die erste Spalte von $(T_{n-1,n}\cdots T_{12})^T$ genau die erste Spalte von T_{12}^T, die wiederum nach Konstruktion mit der ersten Spalte von Q, dem orthogonalen Anteil in einer QR-Zerlegung von $B^TB-\sigma I$, (eventuell bis auf einen Faktor -1) übereinstimmt. Wir nehmen an, die Matrix $B_+^TB_+$ sei unreduziert (andernfalls zerfällt die Berechnung der Singulärwertzerlegung) und wenden Satz 4.3 mit $C := B^TB$, $P := (T_{n-1,n}\cdots T_{12})^T$ an, ferner sei Q der orthogonale Anteil einer QR-Zerlegung von $B^TB-\sigma I$. Wie wir angenommen haben, ist die Matrix

$$G := B_+^TB_+ = P^T(B^TB)P = P^TCP$$

eine unreduzierte Tridiagonalmatrix, insbesondere obere Hessenberg-Matrix. Auch die Matrix

$$H := Q^T(B^TB)Q = Q^TCQ$$

ist[3] eine Tridiagonalmatrix, insbesondere eine obere Hessenberg-Matrix. Die ersten Spalten von Q und P stimmen (eventuell bis auf das Vorzeichen) überein. Satz 4.3 liefert, dass auch H unreduziert ist und P und Q sowie G und H im wesentlichen übereinstimmen. Damit ist gezeigt, dass es sich bei dem obigen Verfahren um die Durchführung eines impliziten QR-Schrittes mit Shift-Parameter σ auf B^TB handelt, wobei der Witz darin besteht, dass man B^TB nicht bildet.

Anhand einer 4×4-Bidiagonalmatrix veranschaulichen wir uns obigen Algorithmus. Abwechselnd wird mit $T_{k,k+1}^T$ von rechts, mit $S_{k,k+1}$ von links multipliziert, $k = 1,\ldots,n-1$. Bei der Transformation festbleibende Elemente werden mit $\bullet$, sich verändernde mit $*$ bezeichnet.

$$\begin{pmatrix} \bullet & \bullet & & \\ & \bullet & \bullet & \\ & & \bullet & \bullet \\ & & & \bullet \end{pmatrix} \xrightarrow{T_{12}} \begin{pmatrix} * & * & & \\ * & * & \bullet & \\ & & \bullet & \bullet \\ & & & \bullet \end{pmatrix} \xrightarrow{S_{12}} \begin{pmatrix} * & * & * & \\ & * & * & \\ & & \bullet & \bullet \\ & & & \bullet \end{pmatrix} \xrightarrow{T_{23}} \begin{pmatrix} \bullet & * & & \\ & * & * & \\ & * & * & \bullet \\ & & & \bullet \end{pmatrix}$$

$$\xrightarrow{S_{23}} \begin{pmatrix} \bullet & \bullet & & \\ & * & * & * \\ & & * & * \\ & & & \bullet \end{pmatrix} \xrightarrow{T_{34}} \begin{pmatrix} \bullet & \bullet & & \\ & \bullet & * & \\ & & * & * \\ & & * & * \end{pmatrix} \xrightarrow{S_{34}} \begin{pmatrix} \bullet & \bullet & & \\ & \bullet & \bullet & \\ & & * & * \\ & & & * \end{pmatrix}$$

Bemerkung: Einer Implementation des obigen Golub-Reinsch-Verfahrens sollte nun nichts mehr im Wege stehen. Hierbei kann die in Unterabschnitt 3.2.3 eingeführte Funktion "givrot" benutzt werden. Berücksichtigt werden sollte, dass die im Verlauf des Verfahrens transformierte Matrix B immer nur in (höchstens) einem Element von der oberen Bidiagonal-Gestalt abweicht. Nach jedem Schritt $B \longrightarrow B_+$ sollte getestet werden, ob die neue Bidiagonalmatrix B_+ zerfällt, ob also z. B. das letzte Superdiagonalelement betragsmäßig kleiner oder gleich einer vorgegebenen Toleranz ist.

[3] Dies werden wir erst bei der (kurzen) Untersuchung des QR-Verfahrens bei symmetrischen Matrizen genau erkennen.

Eine Bemerkung soll noch zur Berechnung des Shift-Parameters σ gemacht werden. Dieser sollte als derjenige Eigenwert von

$$\begin{pmatrix} c_{n-1,n-1} & c_{n-1,n} \\ c_{n-1,n} & c_{nn} \end{pmatrix} = \begin{pmatrix} d_{n-1}^2 + f_{n-1}^2 & d_{n-1}f_n \\ d_{n-1}f_n & d_n^2 + f_n^2 \end{pmatrix}$$

bestimmt werden, der $c_{nn} = d_n^2 + f_n^2$ am nächsten liegt. Die beiden Eigenwerte sind

$$\begin{aligned}
\lambda_{1,2} &= \frac{c_{n-1,n-1} + c_{nn}}{2} \pm \sqrt{\Big(\frac{c_{n-1,n-1} - c_{nn}}{2}\Big)^2 + c_{n-1,n}^2} \\
&= c_{nn} + \underbrace{\frac{c_{n-1,n-1} - c_{nn}}{2}}_{=:\delta} \pm \sqrt{\Big(\frac{c_{n-1,n-1} - c_{nn}}{2}\Big)^2 + c_{n-1,n}^2} \\
&= c_{nn} + \delta \pm \sqrt{\delta^2 + c_{n-1,n}^2} \\
&= c_{nn} + \frac{(\delta \pm \sqrt{\delta^2 + c_{n-1,n}^2})(\delta \mp \sqrt{\delta^2 + c_{n-1,n}^2})}{\delta \mp \sqrt{\delta^2 + c_{n-1,n}^2}} \\
&= c_{nn} - \frac{c_{n-1,n}^2}{\delta \mp \sqrt{\delta^2 + c_{n-1,n}^2}} \\
&= c_{nn} - \frac{\operatorname{sign}(\delta) c_{n-1,n}^2}{|\delta| \mp \operatorname{sign}(\delta)\sqrt{\delta^2 + c_{n-1,n}^2}}.
\end{aligned}$$

Bei G. H. Golub, C. F. Van Loan (1989, S. 423) wird daher empfohlen, diesen sogenannten *Wilkinson-Shift-Parameter* folgendermaßen zu berechnen:

$$\delta := \frac{c_{n-1,n-1} - c_{nn}}{2}, \qquad \sigma := c_{nn} - \frac{\operatorname{sign}(\delta) c_{n-1,n}^2}{|\delta| + \sqrt{\delta^2 + c_{n-1,n}^2}}.$$

Mit diesen Bemerkungen zu einer Implementation des Golub-Reinsch-Verfahren wollen wir uns begnügen. □

3.4.3 MATLAB-Ergänzungen

Wir wollen zunächst die "naive" Methode umsetzen, die wir am Anfang des Abschnittes 3.4 geschildert haben. Die resultierende Funktion ist etwa

```
function [U,S,V]=NaiveSVD(A);
%*************************************************************
%[U,S,V] = NaiveSVD(A) produces a diagonal matrix S, of the same
%    dimension as A and with nonnegative diagonal elements in
%    decreasing order, and unitary matrices U and V so that
%    A = U*S*V'. We are using the ''naive'' method described at
%    the beginning of section 3.4.
```

```
%*************************************************************
[m,n]=size(A);B=A'*A;
[V,D]=eig(B);d=diag(D);
[lambda,I]=sort(d);V=V(:,I);
lambda=lambda(n:-1:1);V=V(:,n:-1:1);
s=sqrt(lambda);
[tildeU,R,Pi]=qr(A*V);
for j=1:n
    if R(j,j)<0
         R(j,j:n)=-R(j,j:n);tildeU(:,j)=-tildeU(:,j);
    end;
end;
S=[diag(s);zeros(m-n,n)];
U=tildeU*[Pi' zeros(n,m-n);zeros(m-n,n) eye(m-n)];
```

Man sollte sich selbst über die benutzten Funktionen `eig` und `sort` In gut konditionierten Fällen sind die Ergebnisse von `NaiveSVD` zufriedenstellend. Mit `A=rand(12,7)` und `[U,S,V]=NaiveSVD(A)` ist z. B. $\|A-USV^T\|_\infty = 5.7454\cdot 10^{-15}$, während $\|U^TU-I\|_\infty = 1.2184\cdot 10^{-15}$ und $\|V^TV-I\|_\infty = 2.5292\cdot 10^{-15}$. informieren.

Der erste Schritt bei der Berechnung der Singulärwertzerlegung einer Matrix $A \in \mathbb{R}^{m\times n}$ ist ein Reduktionsschritt. Grundlage ist Lemma 4.1. Bevor wir die entsprechende Funktion bereit stellen, geben wir noch ein Analogon zur Funktion `RowHouse` an, mit der die Multiplikation einer Matrix von *rechts* mit einer Householder-Matrix durchgeführt werden kann.

```
function A=ColHouse(A,v);
%*****************************************************************
%Given an m-by-n matrix A and a m-vector v with v(1)=1, the
%following algorithm overwrites A with AP where P=I-2*v*v'/(v'*v).
%*****************************************************************
beta=2/(v'*v);
w=beta*A*v;
A=A-w*v';
```

Nun folgt die Funktion, mit der man die Berechnung einer Singulärwertzerlegung einer beliebigen Matrix $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ auf die Berechnung der Singulärwertzerlegung einer oberen Bidiagonalmatrix $B \in \mathbb{R}^{n\times n}$ zurückführen kann. Grundlage hierfür ist Lemma 4.1 und der folgende Algorithmus.

```
function [B,P,Q]=Bidiagonal(A);
%*************************************************************
%Given m-by-n matrix A with m>=n the following function
%computes orthogonal m-by-m resp. n-by-n matrices P and Q
%such that
%           P'*A*Q=[B]
%                  [0],
```

```
%where B is an n-by-n upper bidiagonal matrix.
%************************************************************
[m,n]=size(A);v=zeros(m,1);B=zeros(n);
for k=1:n
    v(k:m)=House(A(k:m,k));
    A(k:m,k:n)=RowHouse(A(k:m,k:n),v(k:m));
    A(k+1:m,k)=v(k+1:m);
    if k<=n-2
        v(k+1:n)=House(A(k,k+1:n)');
        A(k:m,k+1:n)=ColHouse(A(k:m,k+1:n),v(k+1:n));
        A(k,k+2:n)=v(k+2:n)';
    end;
end;
B=triu(tril(A(1:n,1:n),1));
P=eye(m);
for k=n:-1:1
    v(k)=1;v(k+1:m)=A(k+1:m,k);
    P(k:m,k:m)=RowHouse(P(k:m,k:m),v(k:m));
end;
Q=eye(n);
for k=n-2:-1:1
    v(k+1)=1;v(k+2:n)=A(k,k+2:n)';
    Q(k+1:n,k+1:n)=RowHouse(Q(k+1:n,k+1:n),v(k+1:n));
end;
```

Wir sind großzügig mit dem Speicherplatz umgegangen, da wir z. B. die obere Bidiagonalmatrix B in einem $n \times n$-Feld und nicht in zwei Feldern der Länge n bzw. $n-1$ speichern, was natürlich wesentlich ökonomischer wäre. Im folgenden geht es "nur" noch darum, eine (volle oder reduzierte, das ist jetzt egal) Singulärwertzerlegung der quadratischen oberen Bidiagonalmatrix B zu bestimmen.

Auch der folgende Schritt ist ein Reduktionsschritt. Durch ihn wird gerechtfertigt, dass B o. B. d. A. eine unreduzierte Bidiagonalmatrix (d. h. alle Superdiagonalelemente von B sind von Null verschieden) ist, bei der auch alle Diagonalelemente nicht verschwinden bzw. B^TB eine unreduzierte (natürlich symmetrische) Tridiagonalmatrix ist. Auf diesen wollen wir hier nicht mehr eingehen.

Nun noch einige Hinweise zum Verfahren von Golub-Reinsch. Gegeben sei eine obere Bidiagonalmatrix $B \in \mathbb{R}^{n\times n}$ mit der Eigenschaft, dass B^TB unreduziert ist. Die folgende Matlab-Funktion führt einen Schritt des Golub-Reinsch Verfahrens durch. Wir verzichten hierbei auf das Akkumulieren benutzter orthogonaler Matrizen.

```
function B=StepSVD(B)
%********************************************************************
%This function computes one step of the Golub-Reinsch method
%in order to compute the singular values of the upper diagonal
%n-by-n matrix B. It is assumed that B'*B is unreduced.
%We use the  Wilkinson shift strategy.
```

```
%*****************************************************************
n=size(B,1);
%Compute Wilkinson-Shift
if n>2
    c11=B(n-1,n-1)^2+B(n-2,n-1)^2;
else
    c11=B(n-1,n-1)^2;
end;
c12=B(n-1,n-1)*B(n-1,n);c22=B(n,n)^2+B(n-1,n)^2;
delta=0.5*(c11-c22);
sigma=c22-sign(delta)*c12^2/(abs(delta)+sqrt(delta^2+c12^2));
%First Step
[c,s]=Givens(B(1,1)^2-sigma,B(1,1)*B(1,2));
B(:,1:2)=B(:,1:2)*[c -s;s c];
for k=1:n-1
    [c,s]=Givens(B(k,k),B(k+1,k));
    B(k:k+1,:)=[c s;-s c]*B(k:k+1,:);
    if k<n-1
        [c,s]=Givens(B(k,k+1),B(k,k+2));
        B(:,k+1:k+2)=B(:,k+1:k+2)*[c -s;s c];
    end;
end;
```

Die singulären Werte selbst können dann mit Hilfe von

```
function s=SVD(A);
%This function computes the singular values of the m-by-n matrix A.
%Here we suppose m>=n.

%Bidiagonalize A.
[m,n]=size(A);v=zeros(m,1);B=zeros(n);
for k=1:n
    v(k:m)=House(A(k:m,k));
    A(k:m,k:n)=RowHouse(A(k:m,k:n),v(k:m));
    A(k+1:m,k)=v(k+1:m);
    if k<=n-2
        v(k+1:n)=House(A(k,k+1:n)');
        A(k:m,k+1:n)=ColHouse(A(k:m,k+1:n),v(k+1:n));
        A(k,k+2:n)=v(k+2:n)';
    end;
end;
B=triu(tril(A(1:n,1:n),1));
s=zeros(n,1);
while n>1
     if abs(B(n-1,n))>eps*(abs(B(n,n))+abs(B(n-1,n-1)))
          B=StepSVD(B);
```

```
    else
        s(n)=B(n,n);n=n-1;B=B(1:n,1:n);
    end;
end;
s(1)=B(1,1);
s=-sort(-abs(s));
```

berechnet werden. Es sei betont, dass obiges Verfahren nicht immer funktioniert, weil "keine Reduktion im reduzierten Fall" vorgenommen wird. Auf weitere Feinheiten wollen wir hier nicht eingehen.

Es wurde die Möglichkeit erwähnt, die Singulärwertzerlegung einer Matrix zur Bildkompression zu benutzen. Im Buch von J. W. Demmel (1997, S. 114) wird hierzu ein explizites Beispiel gebracht, das man mit Matlab verfolgen kann. Durch `clown.mat` wird eine 200×320-Matrix X definiert, wobei die ganzzahligen Einträge zwischen 1 (sehr dunkel) und 81 (sehr hell) schwanken und der Rang maximal gleich 200 ist. Für die Speicherung des Bildes benötigt man also 64 000 Zahlen. Durch `[U,S,V]=svd(X)` wird eine volle Singulärwertzerlegung von X berechnet. Mit einem gegebenen $k \in \{1, \ldots, 200\}$ kann man das Bild komprimieren, indem man

$$X = \sum_{i=1}^{n} \sigma_i u_i v_i^T$$

durch

$$X_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

ersetzt. Dies kann durch

```
colormap('gray');
image(U(:,1:k)*S(1:k,1:k)*V(:,1:k)')
```

geschehen. In Abbildung 3.2 steht links das Originalbild ($k = 200$), rechts eine Approximation mit $k = 20$. In der nächsten Abbildung 3.3 sind die entsprechenden Approxi-

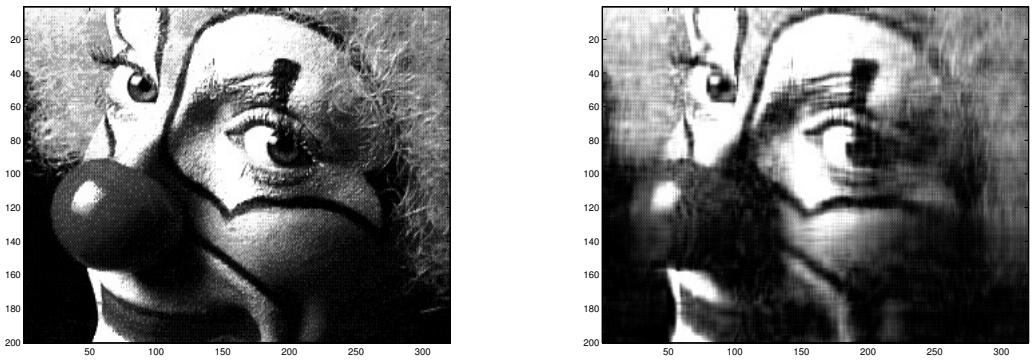


Abbildung 3.2: Das Originalbild und eine Approximation mit $k = 20$

mationen mit $k = 10$ und $k = 5$ dargestellt. Man erkennt, dass die Approximation mit

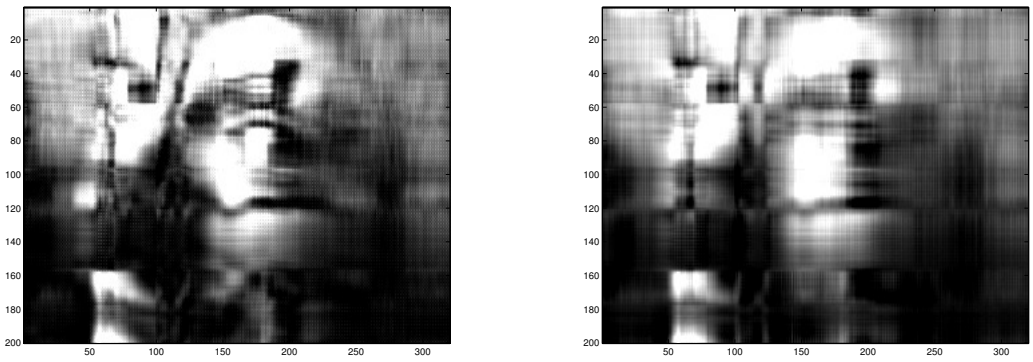


Abbildung 3.3: Approximationen mit $k = 10$ und $k = 5$

$k = 20$ durchaus noch akzeptable Ergebnisse liefert. Zur Speicherung von X_k benötigt man $(m + n + 1)k$ Zahlen, für $k = 20$, $m = 200$ und $n = 320$ sind dies 10 420 Zahlen. Es sei aber darauf hingewiesen, dass es bessere Methoden zur Bildkompression gibt.

3.4.4 Aufgaben

1. Sei $B \in \mathbb{R}^{n\times n}$ eine quadratische Matrix mit der Singulärwertzerlegung $B = U\Sigma V^T$. Sei $C \in \mathbb{R}^{2n\times 2n}$ definiert durch

$$C := \begin{pmatrix} 0 & B^T \\ B & 0 \end{pmatrix}.$$

   Man zeige:

   (a) Die Matrix

$$W := \frac{1}{\sqrt{2}} \begin{pmatrix} V & V \\ U & -U \end{pmatrix}$$

   ist orthogonal.

   (b) Es ist

$$W^T C W = \begin{pmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{pmatrix}.$$

   Sind also $\sigma_1, \ldots, \sigma_n$ die singulären Werte von B, so sind $\pm\sigma_i$, $i = 1, \ldots, n$, die Eigenwerte von C.

2. Sei $B \in \mathbb{R}^{n\times n}$ eine obere Bidiagonalmatrix, bei der alle Haupt- und Superdiagonalelemente nicht verschwinden. Man zeige, dass alle singulären Werte $\sigma_1, \ldots, \sigma_n$ von B positiv und einfach sind, also $\sigma_1 > \cdots > \sigma_n > 0$ gilt.

   Hinweis: Man darf benutzen, dass die Eigenwerte einer symmetrischen, unreduzierten Tridiagonalmatrix einfach sind (siehe z. B. J. Werner (1992b, S. 57)).

3. Sei $R \in \mathbb{R}^{n\times n}$ eine obere Dreiecksmatrix. Unter Benutzung von Givens-Rotationen gebe man ein Verfahren an, welches die Berechnung einer Singulärwertzerlegung von R auf die einer oberen Bidiagonalmatrix $B \in \mathbb{R}^{n\times n}$ reduziert. Ferner mache man sich das Verfahren für $n = 4$ klar.

4. Gegeben sei eine Matrix $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$. Man betrachte die beiden folgenden Möglichkeiten, die Berechnung einer Singulärwertzerlegung von A auf die einer Bidiagonalmatrix $B \in \mathbb{R}^{n\times n}$ zurückzuführen:

   (a) Berechne Householder-Matrizen $P_1, \ldots, P_n \in \mathbb{R}^{m\times m}$ und $Q_1, \ldots, Q_{n-2} \in \mathbb{R}^{n\times n}$ mit
   $$P_n \cdots P_1 A Q_1 \cdots Q_{n-2} = \begin{pmatrix} B \\ 0 \end{pmatrix},$$
   wobei $B \in \mathbb{R}^{n\times n}$ eine obere Bidiagonalmatrix ist.

   (b) Berechne zunächst mit Hilfe des Householder-Verfahrens eine obere Dreiecksmatrix $R \in \mathbb{R}^{n\times n}$, für welche mit orthogonalem $Q \in \mathbb{R}^{m\times m}$ gilt, dass
   $$A = Q \begin{pmatrix} R \\ 0 \end{pmatrix}.$$
   Anschließend berechne man (siehe Aufgabe 3) mit Hilfe von Givens-Rotationen eine obere Bidiagonalmatrix B mit $B = SRT^T$, wobei $S, T \in \mathbb{R}^{n\times n}$ orthogonal sind.

   Wieviele flops benötigt man jeweils im wesentlichen (niedere Terme in m und n können unberücksichtigt bleiben) zur Berechnung der oberen Bidiagonalmatrix B, für welche m, n ist welche Methode vorzuziehen?

5. Seien
$$A := \begin{pmatrix} 1 & 0.04 & 0.0016 \\ 1 & 0.32 & 0.1024 \\ 1 & 0.51 & 0.2601 \\ 1 & 0.73 & 0.5329 \\ 1 & 1.03 & 1.0609 \\ 1 & 1.42 & 2.0164 \\ 1 & 1.60 & 2.5600 \end{pmatrix}, \qquad b := \begin{pmatrix} 2.63 \\ 1.18 \\ 1.16 \\ 1.54 \\ 2.65 \\ 5.41 \\ 7.67 \end{pmatrix}$$
   gegeben. Man berechne mit dem Golub-Reinsch-Verfahren eine reduzierte Singulärwertzerlegung von A und anschließend die (eindeutige) Lösung des linearen Ausgleichsproblems mit den Daten $\begin{pmatrix} A & b \end{pmatrix}$.

6. Seien
$$A := \begin{pmatrix} 22 & 10 & 2 & 3 & 7 \\ 14 & 7 & 10 & 0 & 8 \\ -1 & 13 & -1 & -11 & 3 \\ -3 & -2 & 13 & -2 & 4 \\ 9 & 8 & 1 & -2 & 4 \\ 9 & 1 & -7 & 5 & -1 \\ 2 & -6 & 6 & 5 & 1 \\ 4 & 5 & 0 & -2 & 2 \end{pmatrix}, \qquad B := \begin{pmatrix} -1 & 1 & 0 \\ 2 & -1 & 1 \\ 1 & 10 & 11 \\ 4 & 0 & 4 \\ 0 & -6 & -6 \\ -3 & 6 & 3 \\ 1 & 11 & 12 \\ 0 & -5 & -5 \end{pmatrix}$$
   gegeben. Man berechne eine reduzierte Singulärwertzerlegung von A. Anschließend löse man die linearen Ausgleichsprobleme mit A als Koeffizientenmatrix und den Spalten von B als rechter Seite, genauer bestimme man jeweils die Lösung mit minimaler euklidischer Norm.

7. Gegeben sei die obere Bidiagonalmatrix

$$B := \begin{pmatrix} d_1 & f_2 & & & \\ & d_2 & f_3 & & \\ & & \ddots & \ddots & \\ & & & d_{n-1} & f_n \\ & & & & d_n \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

Man bilde die Matrix

$$C := \begin{pmatrix} 0 & B^T \\ B & 0 \end{pmatrix} \in \mathbb{R}^{2n\times 2n}$$

und zeige:

(a) Es gibt eine Permutationsmatrix $P \in \mathbb{R}^{2n\times 2n}$ mit

$$T := P^T C P = \begin{pmatrix} 0 & d_1 & & & & \\ d_1 & 0 & f_2 & & & \\ & f_2 & 0 & d_2 & & \\ & & d_2 & 0 & \ddots & \\ & & & \ddots & \ddots & d_n \\ & & & & d_n & 0 \end{pmatrix}.$$

(b) Gegeben sei die obere Bidiagonalmatrix $B \in \mathbb{R}^{n\times n}$ und hiermit die spezielle Tridiagonalmatrix T aus 7a. Man betrachte den folgenden Algorithmus, von dem wir annehmen, dass er durchführbar ist.

- Gegeben sei $\sigma > 0$.
- Setze $\delta := -\sigma$, $\omega := 0$.
- Für $k = 2, \ldots, 2n$:

  Setze

$$\alpha := \begin{cases} d_{k/2}, & \text{falls } k \text{ gerade,} \\ f_{(k+1)/2}, & \text{falls } k \text{ ungerade.} \end{cases}$$

  Berechne

$$\delta := -\sigma - \frac{\alpha^2}{\delta}.$$

  Falls $\delta > 0$, dann: $\omega := \omega + 1$.

Man zeige, dass nach Abschluss des Verfahrens ω die Anzahl der singulären Werte von B angibt, die größer als σ sind.

Hinweis: Man entwickle ein Verfahren zur Berechnung der LDL^T-Zerlegung von $T-\sigma I$ und denke an den Sylvesterschen Trägheitssatz (siehe Satz 3.3 in Abschnitt 2.3).

3.5 Störungstheorie für lineare Ausgleichsprobleme

In diesem Abschnitt wollen wir untersuchen, wie sich die (wir werden den Rangdefizienten Fall im wesentlichen unberücksichtigt lassen) Lösung eine linearen Gleichungssystems bei einer Störung der Daten verändern kann.

3.5.1 Kondition, Störungslemma

Zunächst übertragen wir, zumindestens für die Spektralnorm, die Definition der Kondition einer quadratischen, nichtsingulären Matrix auf beliebige Matrizen.

Definition 5.1 Sei $A \in \mathbb{R}^{m\times n}$ und $A^+ \in \mathbb{R}^{n\times m}$ die zugehörige Pseudoinverse. Dann heißt

$$\kappa_2(A) := \|A\|_2 \, \|A^+\|_2$$

die *Kondition* von A (bezüglich der Spektralnorm).

Bemerkung: Offensichtlich ist $\kappa_2(A) = \sigma_1/\sigma_r$, wobei $r :=$ Rang (A) und $\sigma_1 \geq \cdots \geq \sigma_r$ die positiven singulären Werte von A (bzw. A^T) sind. □

In Lemma 2.1 in Abschnitt 2.2 haben wir das bekannte Störungslemma zitiert. Es sagt aus: Ist $A \in \mathbb{R}^{n\times n}$ nichtsingulär und $\delta A \in \mathbb{R}^{n\times n}$ eine Störung mit

$$\|A^{-1}\| \, \|\delta A\| = \kappa(A) \, \frac{\|\delta A\|}{\|A\|} < 1,$$

so ist auch $A + \delta A$ nichtsingulär und

$$\|(A+\delta A)^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \, \|\delta A\|}.$$

Das nächste Lemma verallgemeinert, jedenfalls bei zu grunde gelegter Spektralnorm, diese Aussage auf rechteckige Matrizen maximalen Rangs.

Lemma 5.2 *Sei $A \in \mathbb{R}^{m\times n}$ mit* Rang $(A) = n \leq m$ *gegeben. Ist $\delta A \in \mathbb{R}^{m\times n}$ eine Störung von A mit $\|A^+\|_2 \, \|\delta A\|_2 < 1$, so ist* Rang $(A + \delta A) = n$ *und*

$$\|(A+\delta A)^+\|_2 \leq \frac{\|A^+\|_2}{1 - \|A^+\|_2 \, \|\delta A\|_2}.$$

Beweis: Zunächst zeigen wir, dass Rang $(A + \delta A) = n$. Sei $x \in$ Kern $(A + \delta A)$. Dann ist $Ax = -\delta Ax$, nach Multiplikation mit A^T folgt $A^TAx = -A^T\delta Ax$ und danach $x = -A^+\delta Ax$. Wegen $\|A^+\|_2 \, \|\delta A\|_2 < 1$ folgt $x = 0$. Also sind die n Spalten von $A + \delta A$ linear unabhängig und folglich Rang $(A + \delta A) = n$.

Sei x ein durch $\|x\|_2 = 1$ normierter Eigenvektor zum kleinsten Eigenwert $\lambda_n(A + \delta A)$ von $(A + \delta A)^T(A + \delta A)$. Dann ist

$$\begin{aligned}
\frac{1}{\|(A+\delta A)^+\|_2} &= \sigma_n(A+\delta A) \\
&= \lambda_n(A+\delta A)^{1/2} \\
&= [x^T(A+\delta A)^T(A+\delta A)x]^{1/2} \\
&= \|(A+\delta A)x\|_2 \\
&\geq \|Ax\|_2 - \|\delta Ax\|_2 \\
&\geq (x^TA^TAx)^{1/2} - \|\delta A\|_2 \\
&\geq \sigma_n(A) - \|\delta A\|_2 \\
&= \frac{1}{\|A^+\|_2} - \|\delta A\|_2 \\
&> 0,
\end{aligned}$$

wobei wir für die letzte Ungleichung die Voraussetzung $\|A^+\|_2 \, \|\delta A\|_2 < 1$ benutzt haben. Hiermit folgt die Behauptung. □

Bemerkung: Setzt man im letzten Lemma statt Rang $(A) = n$ voraus, dass

$$\text{Rang } (A + \delta A) \leq \text{Rang } (A),$$

der Rang von A durch die Störung δA also nicht erhöht wird, behält man aber die Voraussetzung $\|A^+\|_2 \, \|\delta A\|_2 < 1$ bei, so kann immer noch

$$\text{Rang } (A + \delta A) = \text{Rang } (A), \qquad \|(A + \delta A)^+\|_2 \leq \frac{\|A^+\|_2}{1 - \|A^+\|_2 \, \|\delta A\|_2}$$

gezeigt werden[4]. □

3.5.2 Störungssätze

Zuerst wollen wir den bei weitem einfachsten Fall betrachten, dass nämlich nur an der "rechten Seite" b "gewackelt" wird. Wir erinnern vorher an die (einfache) Aufgabe 2b im Abschnitt 3.3. Hiernach ist $AA^+ = P_{\text{Bild}\,(A)}$ bei gegebenem $A \in \mathbb{R}^{m \times n}$ die orthogonale Projektion des $\mathbb{R}^m$ auf Bild (A). Die entsprechende Aussage ist:

Lemma 5.3 *Seien $A \in \mathbb{R}^{m \times n}$ mit $m \geq n$ sowie $b, \delta b \in \mathbb{R}^m$ gegeben. Seien x bzw. $x + \delta x$ die eindeutigen Lösungen minimaler euklidischer Norm der linearen Ausgleichsprobleme zu den Daten (A, b) bzw. $(A, b + \delta b)$. Dann ist*

$$\|\delta x\|_2 \leq \|A^+\|_2 \, \|P_{\text{Bild}\,(A)} \delta b\|_2, \qquad \frac{\|\delta x\|_2}{\|x\|_2} \leq \kappa_2(A) \frac{\|P_{\text{Bild}\,(A)} \delta b\|_2}{\|P_{\text{Bild}\,(A)} b\|_2}.$$

Beweis: Es ist $x = A^+ b$ und

$$x + \delta x = A^+(b + \delta b) = x + A^+ \delta b,$$

folglich

$$\delta x = A^+ \delta b = A^+ A A^+ \delta b = A^+ P_{\text{Bild}\,(A)} \delta b$$

und daher

$$\|\delta x\|_2 \leq \|A^+\|_2 \, \|P_{\text{Bild}\,(A)} \delta b\|_2,$$

das ist schon der erste Teil des Beweises. Weiter ist

$$x = A^+ b = A^+ A A^+ b = A^+ P_{\text{Bild}\,(A)} b = \sum_{i=1}^{r} \frac{u_i^T P_{\text{Bild}\,(A)} b}{\sigma_i} v_i,$$

wobei wir mit den üblichen Bezeichnungen von einer Singulärwertzerlegung $A = \hat{U} \hat{\Sigma} V^T$ und $r := \text{Rang } (A)$ ausgehen. Hieraus liest man ab, dass

$$\|x\|_2^2 = \sum_{i=1}^{r} \Big(\frac{u_i^T P_{\text{Bild}\,(A)} b}{\sigma_i} \Big)^2 \geq \frac{1}{\sigma_1^2} \sum_{i=1}^{r} [u_i^T P_{\text{Bild}\,(A)} b]^2 = \frac{1}{\sigma_1^2} \|P_{\text{Bild}\,(A)} b\|_2^2.$$

[4] Dies findet man z. B. auf S. 43 in einem der (älteren) Standardwerke über Ausgleichsproble, nämlich C. L. Lawson, R. J. Hanson (1974) *Solving Least Squares Problems.* Prentice-Hall, Englewood Cliffs.

Für den letzten Schluss haben wir

$$P_{\text{Bild}\,(A)} b \in \text{Bild}\,(A) = \text{span}\,\{u_1, \ldots, u_r\}$$

ausgenutzt. Insgesamt ist daher

$$\frac{\|\delta x\|_2}{\|x\|_2} \leq \frac{\sigma_1}{\sigma_r} \frac{\|P_{\text{Bild}\,(A)}\delta b\|_2}{\|P_{\text{Bild}\,(A)} b\|_2} = \kappa_2(A) \frac{\|P_{\text{Bild}\,(A)}\delta b\|_2}{\|P_{\text{Bild}\,(A)} b\|_2}.$$

Damit ist das Lemma bewiesen. □

Anschließend formulieren und beweisen wir ein wesentlich tieferliegendes Ergebnis.

Satz 5.4 *Seien $A \in \mathbb{R}^{m\times n}$ mit $\text{Rang}\,(A) = n$ und eine Störung $\delta A \in \mathbb{R}^{m\times n}$ mit $\|A^+\|_2 \|\delta A\|_2 < 1$ gegeben. Ferner seien $b, \delta b \in \mathbb{R}^m$ gegeben und x bzw. $x + \delta x$ die Lösung zum linearen Ausgleichsproblem zu den Daten (A, b) bzw. $(A + \delta A, b + \delta b)$. Dann ist*

$$\frac{\|\delta x\|_2}{\|x\|_2} \leq \frac{\kappa_2(A)}{1 - \kappa_2(A)\|\delta A\|_2/\|A\|_2} \Big[\frac{\|\delta A\|_2}{\|A\|_2} \Big(1 + \kappa_2(A) \frac{\|Ax - b\|_2}{\|A\|_2 \, \|x\|_2} \Big) + \frac{\|\delta b\|_2}{\|A\|_2 \, \|x\|_2} \Big].$$

Beweis: Es ist

$$\delta x = (A + \delta A)^+ (b + \delta b) - A^+ b = [(A + \delta A)^+ - A^+]b + (A + \delta A)^+ \delta b.$$

Der Beweis wird dadurch erfolgen, dass wir $\|\delta x\|_2$ mit Hilfe der Dreiecksungleichung nach oben abschätzen und anschließend durch $\|x\|_2$ dividieren. Wegen des Störungslemmas 5.2 ist $\text{Rang}\,(A + \delta A) = n$, die Abschätzung des zweiten Terms bereitet keine Mühe. Schwieriger ist es, den ersten Term $\|[(A+\delta A)^+ - A^+]b\|_2$ abzuschätzen. Es wird $\text{Rang}\,(A) = \text{Rang}\,(A + \delta A) = n$ benutzt, woraus

$$A^+ A = (A^T A)^{-1} A^T A = I$$

und entsprechend $(A + \delta A)^+ (A + \delta A) = I$ folgt. Daher ist

$$\begin{aligned}
(A + \delta A)^+ - A^+ &= (A + \delta A)^+ (I - AA^+) + (A + \delta A)^+ AA^+ - A^+ \\
&= (A + \delta A)^+ (I - AA^+) + (A + \delta A)^+ AA^+ \\
&\qquad - \underbrace{(A + \delta A)^+ (A + \delta A)}_{=I} A^+ \\
&= (A + \delta A)^+ (I - AA^+) - (A + \delta A)^+ (\delta A) A^+,
\end{aligned}$$

nach Multiplikation mit b ist folglich

$$\begin{aligned}
[(A + \delta A)^+ - A^+]b &= (A + \delta A)^+ (I - AA^+)b - (A + \delta A)^+ (\delta A) \underbrace{A^+ b}_{=x} \\
&= (A + \delta A)^+ (I - AA^+)(b - Ax) \\
&\qquad + (A + \delta A)^+ \underbrace{(I - AA^+)A}_{=0} x - (A + \delta A)^+ (\delta A) x \\
&= (A + \delta A)^+ (I - AA^+)(b - Ax) - (A + \delta A)^+ (\delta A) x.
\end{aligned}$$

Zur Abschätzung des ersten Summanden dient die folgende erste Zwischenbehauptung.

- Sind $A, \tilde{A} \in \mathbb{R}^{m\times n}$ zwei Matrizen mit Rang $(A) =$ Rang $(\tilde{A}) = n$, so ist

$$\|(\tilde{A}\tilde{A}^+)(I - AA^+)\|_2 = \|AA^+(I - \tilde{A}\tilde{A}^+)\|_2.$$

Denn: Wir gehen von den vollen Singulärwertzerlegungen

$$A = U\Sigma V^T, \qquad \tilde{A} = \tilde{U}\tilde{\Sigma}\tilde{V}^T$$

aus. Dann ist

$$AA^+ = U \begin{pmatrix} I_n & 0 \\ 0 & 0 \end{pmatrix} U^T, \qquad \tilde{A}\tilde{A}^+ = \tilde{U} \begin{pmatrix} I_n & 0 \\ 0 & 0 \end{pmatrix} \tilde{U}^T.$$

Man denke sich die $m \times m$-Matrix $\tilde{U}^T U$ zerlegt durch

$$\tilde{U}^T U = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix},$$

wobei

$$W_{11} \in \mathbb{R}^{n\times n}, \quad W_{12} \in \mathbb{R}^{n\times(m-n)}, \quad W_{21} \in \mathbb{R}^{(m-n)\times n}, \quad W_{22} \in \mathbb{R}^{m\times m}.$$

Man erhält

$$\begin{aligned} \|\tilde{A}\tilde{A}^+(I - AA^+)\|_2 &= \left\|\tilde{U} \begin{pmatrix} I_n & 0 \\ 0 & 0 \end{pmatrix} \tilde{U}^T U \begin{pmatrix} 0 & 0 \\ 0 & I_{m-n} \end{pmatrix} U^T\right\|_2 \\ &= \left\|\begin{pmatrix} I_n & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & I_{m-n} \end{pmatrix}\right\|_2 \\ &= \left\|\begin{pmatrix} 0 & W_{12} \\ 0 & 0 \end{pmatrix}\right\|_2 \\ &= \|W_{12}\|_2. \end{aligned}$$

Entsprechend ist

$$\|AA^+(I - \tilde{A}\tilde{A}^+)\|_2 = \|W_{21}^T\|_2 = \|W_{21}\|_2.$$

Zum Nachweis der ersten Zwischenbehauptung bleibt daher $\|W_{12}\|_2 = \|W_{21}\|_2$ zu zeigen. Sei $z \in \mathbb{R}^{m-n}$ beliebig und

$$x := \begin{pmatrix} 0 \\ z \end{pmatrix} \in \mathbb{R}^m.$$

Dann ist

$$\|z\|_2^2 = \|x\|_2^2 = \|\tilde{U}^T U x\|_2^2 = \|W_{12}z\|_2^2 + \|W_{22}z\|_2^2$$

und folglich

$$\|W_{12}\|_2^2 = \max_{\|z\|=1} \|W_{12}z\|_2^2 = 1 - \min_{\|z\|=1} \|W_{22}z\|_2^2 = 1 - \lambda_{\min}(W_{22}^T W_{22}).$$

Entsprechend folgt aus

$$\|z\|_2^2 = \|x\|_2^2 = \|(\tilde{U}^T U)x\|_2^2 = \|W_{21}^T z\|_2^2 + \|W_{22}^T z\|_2^2,$$

dass

$$\|W_{21}\|_2^2 = \|W_{21}^T\|_2^2 = \max_{\|z\|=1} \|W_{21}^T z\|_2^2 = 1 - \min_{\|z\|=1} \|W_{22}^T z\|_2^2 = 1 - \lambda_{\min}(W_{22}W_{22}^T).$$

Hieraus folgt schließlich die Behauptung, da natürlich $\lambda_{\min}(W_{22}^T W_{22}) = \lambda_{\min}(W_{22}W_{22}^T)$.

Als zweite Zwischenbehauptung zeigen wir:

- Es ist
$$\|(A+\delta A)^+(I - AA^+)\|_2 \le \|A^+\|_2 \, \|(A+\delta A)^+\|_2 \, \|\delta A\|_2.$$

Denn: Zur Abkürzung setzen wir $\tilde{A} := A + \delta A$ und nutzen Rang (A) = Rang $(\tilde{A}) = n$ aus. Es ist

$$\begin{aligned}
\|(A+\delta A)^+(I - AA^+)\|_2 &= \|\tilde{A}^+(I - AA^+)\|_2 \\
&= \|\tilde{A}^+\tilde{A}\tilde{A}^+(I - AA^+)\|_2 \\
&\le \|\tilde{A}^+\|_2 \, \|\tilde{A}\tilde{A}^+(I - AA^+)\|_2 \\
&= \|\tilde{A}^+\|_2 \, \|AA^+(I - \tilde{A}\tilde{A}^+)\|_2 \\
&\qquad \text{(erste Zwischenbehauptung)} \\
&= \|\tilde{A}^+\|_2 \, \|(AA^+)^T(I - \tilde{A}\tilde{A}^+)\|_2 \\
&= \|\tilde{A}^+\|_2 \, \|(A^+)^T A^T(I - \tilde{A}\tilde{A}^+)\|_2 \\
&= \|\tilde{A}^+\|_2 \, \|(A^+)^T(A^T - \tilde{A}^T)(I - \tilde{A}\tilde{A}^+)\|_2 \\
&= \|\tilde{A}^+\|_2 \, \|(A^+)^T(\delta A)^T(I - \tilde{A}\tilde{A}^+)\|_2 \\
&\le \|A^+\|_2 \, \|\tilde{A}^+\|_2 \, \|\delta A\|_2 \\
&= \|A^+\|_2 \, \|(A+\delta A)^+\|_2 \, \|\delta A\|_2.
\end{aligned}$$

Hierbei haben wir wiederholt ausgenutzt, dass die Spektralnorm beim Transponieren einer Matrix invariant bleibt, ferner bei der letzten Ungleichung, dass $\|I - \tilde{A}\tilde{A}^+\|_2 \le 1$ (siehe auch Aufgabe 6 in Abschnitt 3.3). Damit ist auch die zweite Zwischenbehauptung bewiesen und wir können den Beweis zu Ende führen. Es ist

$$\begin{aligned}
\|\delta x\|_2 &\le \|[(A+\delta A)^+ - A^+]b\|_2 + \|(A+\delta A)^+\|_2 \, \|\delta b\|_2 \\
&\le \|(A+\delta A)^+(I - AA^+)\|_2 \, \|b - Ax\|_2 + \|(A+\delta A)^+\|_2 \, \|\delta A\|_2 \, \|x\|_2 \\
&\qquad \|(A+\delta A)^+\|_2 \, \|\delta b\|_2 \\
&\le \|(A+\delta A)^+\|_2 \, [\|\delta A\|_2 \, \|A^+\|_2 \, \|Ax - b\|_2 + \|\delta A\|_2 \, \|x\|_2 + \|\delta b\|_2] \\
&\le \frac{\|A^+\|_2}{1 - \|A^+\|_2 \, \|\delta A\|_2} [\|\delta A\|_2 \, \|A^+\|_2 \, \|Ax - b\|_2 + \|\delta A\|_2 \, \|x\|_2 + \|\delta b\|_2] \\
&= \frac{\kappa_2(A)}{1 - \kappa_2(A)\|\delta A\|_2/\|A\|_2} \Big[\frac{\|\delta A\|_2}{\|A\|_2} (\|A^+\|_2 \, \|Ax - b\|_2 + \|x\|_2) + \frac{\|\delta b\|_2}{\|A\|_2} \Big] \\
&= \frac{\kappa_2(A)}{1 - \kappa_2(A)\|\delta A\|_2/\|A\|_2} \Big[\frac{\|\delta A\|_2}{\|A\|_2} \Big(\kappa_2(A) \, \frac{\|Ax - b\|_2}{\|A\|_2} + \|x\|_2 \Big) + \frac{\|\delta b\|_2}{\|A\|_2} \Big].
\end{aligned}$$

Eine Division durch $\|x\|_2$ liefert die Behauptung. □

Bemerkung: Interessant an der Aussage des letzten Satzes ist das Auftreten von $\kappa_2(A)^2 \|Ax - b\|_2$ bei der Abschätzung der relativen Störung einer Lösung. Für großen Defekt $\|Ax - b\|_2$ ist das der dominierende Term. Grob kann man sagen, dass die Sensitivität eines linearen Ausgleichsproblems mit der Koeffizientenmatrix A durch $\kappa_2(A)^2$ für großen Defekt, sonst durch $\kappa_2(A)$ bestimmt ist. □

Bemerkung: Ist $m = n$ in Satz 5.4, die Matrix A also quadratisch und nichtsingulär, so ist der Defekt $Ax-b$ gleich Null, so dass man in diesem Falle wegen $\|b\|_2 \leq \|A\|_2 \|x\|_2$ aus Satz 5.4 genau die Abschätzung aus Satz 2.3 in Abschnitt 2.2 (für die euklidische bzw. zugeordnete Spektralnorm) erhält. □

3.5.3 MATLAB-Ergänzungen

In MATLAB gibt es die Funktion `cond`, in 2.2.3 haben wir schon darauf hingewiesen. Diese Funktion kann man sich mit Hilfe von `type cond` ansehen. Der wesentliche Teil der Implementation sieht folgendermaßen aus:

```
s = svd(A);
   if any(s == 0)   % Handle singular matrix
       c = Inf;
   else
       c = max(s)./min(s);
   end
```

Hier sieht man einen Unterschied zu unserer Definition. In MATLAB ist Kondition `cond(A)` einer Matrix `A` gleich `Inf`, wenn nur ein singulärer Wert verschwindet, also stets im Rang defizienten Fall. Nach

```
A=[1 0;0 0];cond(A)
```

erhalten wir z. B. `cond(A)=Inf`, während nach unserer Definition die Kondition $\kappa_2(A)$ von A genau 1 ist.

Nun geben wir noch ein Beispiel zu Satz 5.4 an. Hierzu betrachten wir die folgenden Daten zu einem ungestörten und einem gestörten linearen Ausgleichsproblem:

$$A := \begin{pmatrix} 1 & 0 \\ 0 & 10^{-6} \\ 0 & 0 \end{pmatrix}, \quad b := \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \delta A := \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 10^{-8} \end{pmatrix}, \quad \delta b := \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Als Lösung des linearen Ausgleichsproblems zu den Daten (A, b) erhält man $x = (1, 0)^T$, zu dem gestörten Problem mit den Daten $(A + \delta A, b + \delta b)$ (nach `format long`) den Vektor

$$x + \delta x = 10^3 \cdot \begin{pmatrix} 0.00100000000000 \\ 9.99900009999000 \end{pmatrix}.$$

Der Defekt ist $Ax - b = (0, 0, -1)^T$. Dann ist

$$\frac{\|\delta x\|_2}{\|x\|_2} = 9.999000099990000 \cdot 10^3.$$

Es ist $\kappa_2(A) = 10^6$. Wegen $\kappa_2(A)\,\|\delta A\|_2/\|A\|_2 = 10^{-2}$ ist der wichtige Term in der Abschätzung von Satz 5.4 genau

$$\kappa_2(A)^2 \frac{\|\delta A\|_2}{\|A\|_2} = 10^5.$$

Da die Norm $\|Ax - b\|_2$ nicht klein ist, bestimmt also tatsächlich das Quadrat der Kondition ganz wesentlich den relativen Fehler.

3.5.4 Aufgaben

1. Mit $\epsilon \in (0, 1]$ sei

$$A := \begin{pmatrix} 1 & 0 \\ 0 & \epsilon \\ 0 & 0 \end{pmatrix}, \quad \delta A := \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & \epsilon/2 \end{pmatrix}, \quad b := \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \delta b := \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

   Man berechne die zum linearen Ausgleichsproblem mit den Daten (A, b) bzw. $(A + \delta A, b + \delta b)$ gehörende Lösung x bzw. $x + \delta x$ und vergleiche die in Satz 5.4 angegebene Abschätzung von $\|\delta x\|_2/\|x\|_2$ mit dem exakten Wert.

2. Sei $A\colon (a, b) \longrightarrow \mathbb{R}^{m\times n}$ mit $m \geq n$ stetig differenzierbar und Rang $(A(t_0)) = n$ für ein $t_0 \in (a, b)$. Man zeige:

   (a) Es existiert eine Umgebung $I \subset (a, b)$ mit Rang $(A(t)) = n$ für alle $t \in I$.

   (b) Für $t \in I$ sei $x(t) \in \mathbb{R}^n$ die eindeutige Lösung des linearen Ausgleichsproblems

$$\text{Minimiere} \quad \|A(t)x - b\|_2, \quad x \in \mathbb{R}^n,$$

   wobei $b \in \mathbb{R}^m$ vorgegeben ist. Dann ist $x\colon I \longrightarrow \mathbb{R}^n$ stetig differenzierbar und die Ableitung $\dot{x}(t)$ für $t \in I$ die Lösung des quadratischen, nichtsingulären Gleichungssystem

$$A(t)^T A(t) z = -A(t)^T \dot{A}(t) x(t) + \dot{A}(t)^T (b - A(t)x(t)).$$

   (c) Weiter ist $\dot{x}(t)$ für $t \in I$ die Lösung des linearen Ausgleichsproblems

$$\text{Minimiere} \quad \|A(t)z + \dot{A}(t)x(t) - (A(t)^+)^T \dot{A}(t)^T (b - A(t)x(t))\|_2, \quad z \in \mathbb{R}^n.$$

3.6 Weitere lineare Ausgleichsprobleme

Es gibt einige Variationen zu linearen Ausgleichsproblemen, die es verdient hätten, auch hier genauer untersucht zu werden. Wir verweisen einmal wieder auf das vorzügliche Buch von Å. Björk (1996) und werden versuchen, einen kleinen Einblick zu geben. Hingewiesen sei aber auch auf P. C. Hansen (1998), insbesondere im Zusammenhang mit Rang-defizienten und schlecht gestellten Problemen. Viele Resultate sind dort aber leider ohne Beweis angegeben. Auf durch lineare Ungleichungen restringierte lineare Ausgleichsprobleme werden wir nicht eingehen. Aufgaben dieser Art können (auch wenn man es nicht machen sollte) auf (konvexe) quadratische Optimierungsaufgaben zurückgeführt werden.

3.6.1 Tichonov-Regularisierung

Einem linearen Ausgleichsproblem

$$\text{(P)} \qquad \text{Minimiere} \quad \|Ax-b\|_2, \quad x \in \mathbb{R}^n,$$

bei dem $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ sehr schlecht konditioniert oder Rang-defizient (dann ist das lineare Ausgleichsproblem nicht eindeutig lösbar und es kommt darauf an, unter den unendlich vielen Lösungen die "richtige" auszuwählen) ist, kann man mit einem $\tau > 0$ und einer Matrix $L \in \mathbb{R}^{p\times n}$ die sogenannte *Tichonov-Regularisierung*

$$\text{(P}_\tau\text{)} \qquad \text{Minimiere} \quad \|Ax-b\|_2^2 + \tau\,\|Lx\|_2^2, \quad x \in \mathbb{R}^n,$$

zuordnen. Hierbei wird also ein Kompromiss angestrebt, einerseits das Residuum, andererseits einen Glättungsterm, repräsentiert durch $\|Lx\|_2^2$, klein zu machen. Ist z. B.

$$L := \begin{pmatrix} 1 & -1 & & \\ & 1 & -1 & \\ & & \ddots & \ddots \\ & & & 1 \quad -1 \end{pmatrix} \in \mathbb{R}^{(n-1)\times n},$$

so ist

$$\|Lx\|_2^2 = \sum_{j=1}^{n-1} (x_{j+1}-x_j)^2$$

der Zusatzterm. Er kann als Glättungsterm aufgefasst werden, da er klein ist, wenn der Vektor x "wenig oszilliert" und genau dann verschwindet, wenn x ein Vielfaches von e ist, also alle Komponenten gleich sind. Andererseits kann auch $p = n$ und $L = I$ sein, so dass der zweite Term in (P_τ) dann dafür sorgt, dass $\|x\|_2$ klein ist. Die Gewichtung zwischen den beiden Kompromissen wird durch den sogenannten *Regularisierungsparameter* τ gesteuert. Offensichtlich ist dieses sogenannte *gedämpfte lineare Ausgleichsproblem* äquivalent dem linearen Ausgleichsproblem

$$\text{Minimiere} \quad \left\| \begin{pmatrix} A \\ \sqrt{\tau}L \end{pmatrix} x - \begin{pmatrix} b \\ 0 \end{pmatrix} \right\|_2, \quad x \in \mathbb{R}^n.$$

Dieses lineare Ausgleichsproblem ist genau dann eindeutig lösbar, wenn die $(m+p)\times n$-Koeffizientenmatrix den Rang n besitzt. Dies wiederum ist der Fall, wenn $\text{Kern}\,(A) \cap \text{Kern}\,(L) = \{0\}$, also z. B. dann, wenn L eine positive Diagonalmatrix ist. Die Normalgleichungen lauten

$$\begin{pmatrix} A \\ \sqrt{\tau}L \end{pmatrix}^T \begin{pmatrix} A \\ \sqrt{\tau}L \end{pmatrix} x = \begin{pmatrix} A \\ \sqrt{\tau}L \end{pmatrix}^T \begin{pmatrix} b \\ 0 \end{pmatrix}$$

bzw.

$$(A^TA + \tau L^TL)x = A^Tb,$$

so dass für $\text{Kern}\,(A) \cap \text{Kern}\,(L) = \{0\}$ die eindeutige Lösung x_τ von (P_τ) durch

$$x_\tau = (A^TA + \tau L^TL)^{-1}A^Tb$$

gegeben ist. Es ist klar, dass die Lösung x_τ nicht nach dieser Formel berechnet werden sollte, da es immer ungünstig ist, die Matrizen A^TA oder L^TL explizit zu bilden. Außerdem müsste für jedes neue τ eine neue Cholesky-Zerlegung von $A^TA + \tau L^TL$ berechnet werden. Besser ist es, direkt das lineare Ausgleichsproblem

$$\text{Minimiere} \quad \left\| \begin{pmatrix} A \\ \sqrt{\tau}L \end{pmatrix} x - \begin{pmatrix} b \\ 0 \end{pmatrix} \right\|_2, \quad x \in \mathbb{R}^n$$

anzugehen. Wie dies mit Hilfe der *verallgemeinerten Singulärwertzerlegung* (siehe Unterabschnitt 3.6.3) möglich ist, kann in Aufgabe 11 untersucht werden.

Der Fall $L = I$ ist besonders einfach, wir wollen ihn näher betrachten. In diesem Fall ist das Problem (P_τ) also durch

$$(\mathrm{P}_\tau) \qquad \text{Minimiere} \quad \left\| \begin{pmatrix} A \\ \sqrt{\tau}I \end{pmatrix} x - \begin{pmatrix} b \\ 0 \end{pmatrix} \right\|_2, \quad x \in \mathbb{R}^n$$

gegeben. Bei P. C. Hansen (1998, S. 38 ff.) findet man Hinweise darauf, wie man auf diese sogenannte *Standardform* transformieren kann (siehe auch Aufgabe 3). Wir nehmen an, eine reduzierte Singulärwertzerlegung von A sei bekannt, also eine Darstellung $A = \hat{U}\hat{\Sigma}V^T$ mit

$$\hat{U} = \begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix} \in \mathbb{R}^{m\times n}, \qquad V = \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix} \in \mathbb{R}^{n\times n},$$

wobei $\hat{U}^T\hat{U} = I$ und $V^TV = I$, ferner

$$\hat{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{n\times n}, \qquad \sigma_1 \geq \cdots \geq \sigma_n \geq 0.$$

Dann ist

$$\begin{aligned} x_\tau &= (A^TA + \tau I)^{-1}A^Tb \\ &= [V(\hat{\Sigma}^2 + \tau I)V^T]^{-1}V\hat{\Sigma}\hat{U}^Tb \\ &= V(\hat{\Sigma}^2 + \tau I)^{-1}\hat{\Sigma}U^Tb \\ &= \sum_{i=1}^{r} \frac{\sigma_i\, u_i^Tb}{\sigma_i^2 + \tau} v_i, \end{aligned}$$

wobei r der Rang von A bzw. die Anzahl positiver singulärer Werte von A ist. Man erkennt, dass $\lim_{\tau\to 0+} x_\tau = A^+b$, d. h. die Lösung x_τ von (P_τ) (mit $L = I$) konvergiert mit $\tau \to 0+$ gegen die Lösung von (P) mit minimaler euklidischer Norm. Eine einfachere Methode zur Lösung von (P_τ) (mit $L := I$) ohne die Berechnung der Singulärwertzerlegung von A wird in Aufgabe 2 skizziert.

Man sollte sich aber natürlich auch Gedanken über die Wahl des "richtigen" positiven Regularisierungsparameters τ machen. Hier ist man in einem Dilemma. Ist ursprünglich das Rang-defiziente oder schlecht konditionierte lineare Ausgleichsproblem, $\|Ax - b\|_2$ zu minimieren, gegeben, also $x := A^+b$ zu berechnen, so möchte man τ natürlich möglichst klein machen, damit das mit τ regularisierte Problem (P_τ) möglichst in der Nähe dieses Problems ist. Andererseits wird die "Kondition" (bezüglich der euklidischen Norm) von (P_τ) durch

$$\kappa_2 \begin{pmatrix} A \\ \sqrt{\tau}I \end{pmatrix} = \left\| \begin{pmatrix} A \\ \sqrt{\tau}I \end{pmatrix} \right\|_2 \left\| \begin{pmatrix} A \\ \sqrt{\tau}I \end{pmatrix}^+ \right\|_2$$

bestimmt und man vermutet zu recht, dass diese mit wachsendem τ kleiner wird. Sind nämlich

$$\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0$$

die singulären Werte von A und $A = \hat{U}\hat{\Sigma}V^T$ mit $\hat{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$ die entsprechende (reduzierte) Singulärwertzerlegung, so ist

$$\begin{pmatrix} A \\ \sqrt{\tau}I \end{pmatrix}^T \begin{pmatrix} A \\ \sqrt{\tau}I \end{pmatrix} = A^TA + \tau I = V(\hat{\Sigma}^2 + \tau I)V^T$$

und daher

$$\left\| \begin{pmatrix} A \\ \sqrt{\tau}I \end{pmatrix} \right\|_2 = \lambda_{\max}^{1/2}(\hat{\Sigma}^2 + \tau I) = (\sigma_1^2 + \tau)^{1/2}.$$

Weiter ist

$$\begin{pmatrix} A \\ \sqrt{\tau}I \end{pmatrix}^+ = \left[\begin{pmatrix} A \\ \sqrt{\tau}I \end{pmatrix}^T \begin{pmatrix} A \\ \sqrt{\tau}I \end{pmatrix} \right]^{-1} \begin{pmatrix} A \\ \sqrt{\tau}I \end{pmatrix}^T = (A^TA + \tau I)^{-1} \begin{pmatrix} A \\ \sqrt{\tau}I \end{pmatrix}^T,$$

folglich

$$\begin{pmatrix} A \\ \sqrt{\tau}I \end{pmatrix}^+ \left[\begin{pmatrix} A \\ \sqrt{\tau}I \end{pmatrix}^+ \right]^T = (A^TA + \tau I)^{-1} = V(\hat{\Sigma}^2 + \tau I)^{-1}V^T$$

und damit

$$\left\| \begin{pmatrix} A \\ \sqrt{\tau}I \end{pmatrix}^+ \right\|_2 = \lambda_{\max}^{1/2}[(\hat{\Sigma}^2 + \tau I)^{-1}] = \frac{1}{(\sigma_n^2 + \tau)^{1/2}}.$$

Insgesamt ist also

$$\kappa_2(\tau) := \kappa_2 \begin{pmatrix} A \\ \sqrt{\tau}I \end{pmatrix} = \left(\frac{\sigma_1^2 + \tau}{\sigma_n^2 + \tau} \right)^{1/2}.$$

Hieraus liest man ab, dass $\kappa_2(\cdot)$ auf $(0, \infty)$ für $\sigma_n < \sigma_1$ (andernfalls ist A^TA ein nichtnegatives Vielfaches der Identität) strikt monoton fallend ist und wegen $\lim_{\tau \to +\infty} \kappa_2(\tau) = 1$ das lineare Ausgleichsproblem (P_τ) im Unendlichen sozusagen optimal konditioniert ist.

Nun wissen wir aber immer noch nicht, wie wir τ wählen sollen, wir wissen nur, dass wir ein Dilemma haben. Um dieses zu lösen[5], stellen wir uns vor, dass die "rechte Seite" b in (P_τ) durch "Rauschen" für ein vorgegebenes $\delta > 0$ durch ein $b^\delta \in \mathbb{R}^m$ mit $\|b^\delta - b\| \leq \delta$ ersetzt ist. Wir berechnen also eigentlich

$$x_\tau^\delta := \begin{pmatrix} A \\ \sqrt{\tau}I \end{pmatrix}^+ \begin{pmatrix} b^\delta \\ 0 \end{pmatrix} = (A^TA + \tau^2 I)^{-1}A^Tb^\delta$$

[5]Wir halten uns hier an R. Kress (1998, S. 88 ff.) *Numerical Analysis.* Springer-Verlag.

und wollen den Fehler zwischen der berechneten Lösung x_τ^δ und der eigentlich interessierenden Lösung $x = A^+ b$ abschätzen. Wir erhalten

$$\begin{aligned}
\|x_\tau^\delta - x\|_2 &= \|(A^TA + \tau I)^{-1}b^\delta - A^+b\|_2 \\
&\leq \|(A^TA + \tau I)^{-1}A^T - A^+\|_2 \, \|b\|_2 + \|(A^TA + \tau I)^{-1}A^T\|_2 \, \|b^\delta - b\|_2 \\
&= \|V[(\hat\Sigma^2 + \tau I)^{-1} - \hat\Sigma^+]U^T\|_2 \, \|b\|_2 + \|V(\hat\Sigma^2 + \tau I)^{-1}\hat\Sigma U^T\|_2 \, \|b^\delta - b\|_2 \\
&= \|(\hat\Sigma^2 + \tau)^{-1} - \hat\Sigma^+\|_2 \, \|b\|_2 + \|(\hat\Sigma^2 + \tau I)^{-1}\Sigma\|_2 \, \|b^\delta - b\|_2 \\
&= \frac{\tau}{\sigma_r(\sigma_r^2 + \tau)}\|b\|_2 + \left(\max_{i=1,\ldots,r} \frac{\sigma_i}{\sigma_i^2 + \tau}\right) \|b^\delta - b\|_2 \\
&\leq \frac{\tau}{\sigma_r(\sigma_r^2 + \tau)}\|b\|_2 + \left(\max_{i=1,\ldots,r} \frac{\sigma_i}{\sigma_i^2 + \tau}\right) \delta.
\end{aligned}$$

Hier sieht man sehr deutlich den Unterschied der beiden Summanden zur Abschätzung des Gesamtfehlers. Der erste geht mit $\tau \to 0+$ gegen Null, der zweite wird (zumindestens bei kleinem σ_r) für $\tau \to 0+$ groß. Der erste Summand kann als Approximations-, der zweite als Datenfehler bezeichnet werden. Am besten würde man τ natürlich so wählen, dass die angegebene obere Schranke für den Fehler $\|x_\tau^\delta - x\|_2$ minimal ist, die dafür nötigen Informationen fehlen aber. Daher versucht man, τ in Abhängigkeit von dem vorgegebenen $\delta > 0$ nach dem sogenannten *Diskrepanzprinzip* zu wählen. Dieses geht von der Vorstellung aus, dass es keinen Sinn macht, den Defekt kleiner als den Datenfehler in der rechten Seite zu machen. Die Aussagen[6] fassen wir in einem Satz zusammen. In diesem wird allerdings vorausgesetzt, dass das Gleichungssystem $Ax = b$ lösbar ist, was natürlich bei linearen Ausgleichsproblemen eine nicht ganz adäquate Voraussetzung ist.

Satz 6.1 *Seien $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$, $b \in \text{Bild}\,(A) \setminus \{0\}$ gegeben. Für jedes $\delta > 0$ sei weiter $b^\delta \in \mathbb{R}^m$ eine Störung von b mit*

$$\|b^\delta - b\|_2 < \delta < \|b^\delta\|_2.$$

Dann gilt:

1. *Zu jedem $\delta > 0$ gibt es genau ein $\tau = \tau(\delta) > 0$ derart, dass die eindeutige Lösung $x^\delta := x_\tau^\delta$ der Aufgabe, $\|Ax - b^\delta\|_2^2 + \tau\|x\|_2^2$ auf dem $\mathbb{R}^n$ zu minimieren, der Gleichung $\|Ax^\delta - b^\delta\|_2 = \delta$ genügt.*
2. *Es ist $\lim_{\delta\to 0+} x^\delta = A^+b$.*

Beweis: Bei gegebenem $\delta > 0$ definieren wir die Funktion $f^\delta\colon (0,\infty) \longrightarrow \mathbb{R}$ durch

$$f^\delta(\tau) := \|Ax_\tau^\delta - b^\delta\|_2^2 - \delta^2.$$

Wir gehen diesmal von einer vollen Singulärwertzerlegung

$$A = U \begin{pmatrix} \hat\Sigma \\ 0 \end{pmatrix} V^T$$

[6] Siehe R. Kress (1998, Theorem 5.10).

aus, wobei $U \in \mathbb{R}^{m\times m}$ und $V \in \mathbb{R}^{n\times n}$ orthogonal sind, ferner

$$\hat{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n), \qquad \sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0.$$

Dann ist

$$\begin{aligned} Ax_\tau^\delta - b^\delta &= [A(A^TA + \tau I)^{-1}A^T - I]b^\delta \\ &= U \begin{pmatrix} \hat{\Sigma}^2(\hat{\Sigma}^2 + \tau I)^{-1} - I_n & 0 \\ 0 & -I_{m-n} \end{pmatrix} U^T b^\delta \end{aligned}$$

und daher

$$\begin{aligned} f^\delta(\tau) &= \sum_{i=1}^{r} \frac{\tau^2}{(\sigma_i^2+\tau)^2}(u_i^T b^\delta)^2 + \sum_{i=r+1}^{m} (u_i^T b^\delta)^2 - \delta^2 \\ &= \sum_{i=1}^{r} \frac{\tau^2}{(\sigma_i^2+\tau)^2}(u_i^T b^\delta)^2 + \sum_{i=r+1}^{m} [u_i^T(b^\delta - b)]^2 - \delta^2. \end{aligned}$$

Die letzte Gleichung folgt aus der Voraussetzung $b \in \operatorname{Bild}(A) = \operatorname{span}\{u_1, \ldots, u_r\}$, welche $u_i^T b = 0$, $i = r+1, \ldots, n$ nach sich zieht. Damit ist

$$\lim_{\tau\to 0+} f^\delta(\tau) = \sum_{i=r+1}^{m} [u_i^T(b^\delta - b)]^2 - \delta^2 \leq \|b^\delta - b\|_2^2 - \delta^2 < 0$$

und

$$\lim_{\tau\to\infty} f^\delta(\tau) = \sum_{i=1}^{m} (u_i^T b^\delta)^2 - \delta^2 = \|b^\delta\|_2^2 - \delta^2 > 0.$$

Da ferner f^δ offensichtlich auf $(0,\infty)$ monoton wachsend ist, besitzt f^δ für jedes $\delta > 0$ genau eine Nullstelle $\tau = \tau(\delta) > 0$.

Wir setzen $x^\delta := x^\delta_{\tau(\delta)}$. Wir zeigen, dass $\lim_{\delta\to 0+} \tau(\delta) = 0$. Ist uns dies gelungen, so folgt mit $\delta \to 0+$ wegen $b^\delta \to b$, dass

$$x^\delta = \sum_{i=1}^{r} \frac{\sigma_i\, u_i^T b^\delta}{\sigma_i^2 + \tau(\delta)} v_i \to \sum_{i=1}^{r} \frac{u_i^T b}{\sigma_i} v_i = A^+ b.$$

Für jedes $\delta > 0$ ist zunächst

$$0 < \|b^\delta\|_2 - \delta = \|b^\delta\|_2 - \|Ax^\delta - b^\delta\|_2 \leq \|Ax^\delta\|_2.$$

Andererseits erhält man aus

$$(A^TA + \tau(\delta) I)x^\delta = A^T b^\delta$$

nach Multiplikation mit A und Umordnen, dass

$$\tau(\delta)\,\|Ax^\delta\|_2 = \|AA^T(b^\delta - Ax^\delta)\|_2 \leq \|AA^T\|_2\,\|b^\delta - Ax^\delta\|_2 = \|AA^T\|_2\,\delta$$

und folglich

$$\tau(\delta) \leq \frac{\|AA^T\|_2\,\delta}{\|Ax^\delta\|_2} \leq \frac{\|A^TA\|_2\,\delta}{\|b^\delta\|_2 - \delta}.$$

Hieraus liest man schließlich ab, daß $\lim_{\delta\to 0+} \tau(\delta) = 0$. Damit ist auch die zweite Aussage des Satzes bewiesen. □

3.6.2 Quadratische Restriktionen

Wir betrachten ein lineares Ausgleichsproblem mit einer quadratischen skalaren Restriktion (siehe Å. BJÖRK (1996, S. 205 ff.)), nämlich

$$\text{(P)} \qquad \text{Minimiere} \quad \|Ax-b\|_2 \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : \|Cx-d\|_2 \le \gamma\},$$

wobei $A \in \mathbb{R}^{m\times n}$ mit $m \ge n$, $C \in \mathbb{R}^{p\times n}$ und $\gamma > 0$. Probleme dieser Art treten ebenfalls bei Regularisierungsverfahren auf. Zunächst wollen wir[7] uns über die Existenz und Eindeutigkeit einer Lösung von (P) Gedanken machen. Vorher wollen wir aber einen eher trivialen Fall ausräumen. Hierzu sei S die Menge der Lösungen des (unrestringierten) linearen Ausgleichsproblems, $\|Ax-b\|_2$ auf dem $\mathbb{R}^n$ zu minimieren. Wir wissen, dass dies ein affin linearer Teilraum des $\mathbb{R}^n$ ist, der mit Hilfe eine Singulärwertzerlegung von A genau beschrieben werden kann. Sei weiter $x_{A,C} \in S$ eine Lösung (weshalb existiert eine solche?, siehe Aufgabe 4) der Aufgabe

$$\text{Minimiere} \quad \|Cx-d\|_2, \quad x \in S.$$

Ist dann $\|Cx_{A,C}-d\|_2 \le \gamma$, so ist $x_{A,C} \in S$ auch eine Lösung des uns eigentlich interessierenden Problems (P). Die quadratische Restriktion wird dann als *nicht bindend* angesehen, sie heißt also *bindend*, wenn

$$\min_{x\in S} \|Cx-d\|_2 > \gamma.$$

Beispiel: Sei speziell $p = n$, $C = I$ und $d = 0$, so dass es sich bei (P) um ein quadratisch restringiertes lineares Ausgleichsproblem in sogenannter *Standardform*, also

$$\text{Minimiere} \quad \|Ax-b\|_2 \quad \text{unter der Nebenbedingung} \quad \|x\|_2 \le \gamma,$$

handelt. Dann ist $x_{A,I} = A^+b$ die Lösung minimaler euklidischer Norm der Aufgabe, $\|Ax-b\|_2$ zu minimieren. Zur Lösung der obigen Aufgabe in Standardform wird man also zunächst A^+b ausrechnen und nachprüfen, ob $\|A^+b\|_2 \le \gamma$. Ist dies der Fall, so hat man eine Lösung gefunden. □

Satz 6.2 *Gegeben sei das lineare, quadratisch restringierte Ausgleichsproblem (P). Dann gilt:*

1. *Die Aufgabe (P) besitzt genau dann eine Lösung, wenn*

$$\min_{x\in\mathbb{R}^n} \|Cx-d\|_2 \le \gamma.$$

Diese ist eindeutig, wenn darüberhinaus die quadratische Restriktion bindend ist und

$$\text{Rang} \begin{pmatrix} A \\ C \end{pmatrix} = n$$

gilt.

[7] Vor uns tat dies W. GANDER (1981) "Least Squares with a Quadratic Constraint." Numer. Math. 36, 291–307.

2. *Sei $x^* \in \mathbb{R}^n$ zulässig für (P), also $\|Cx^* - d\|_2 \leq \gamma$. Existiert ein $\lambda^* \geq 0$ mit*

$$(A^TA + \lambda^* C^TC)x^* = A^Tb + \lambda^* C^Td, \qquad \lambda^*[\gamma - \|Cx^* - d\|_2] = 0,$$

so ist x^* eine Lösung von (P).

3. *Die quadratische Restriktion sei bindend, außerdem $\min_{x\in\mathbb{R}^n} \|Cx - d\|_2 < \gamma$. Ist dann x^* mit $\|Cx^* - d\|_2 \leq \gamma$ eine Lösung von (P), so ist ist notwendigerweise $\|Cx^* - d\|_2 = \gamma$, ferner existiert ein $\lambda^* > 0$ mit*

$$(A^TA + \lambda^* C^TC)x^* = A^Tb + \lambda^* C^Td.$$

Beweis: Ist $\min_{x\in\mathbb{R}^n} \|Cx - d\|_2 > \gamma$, so ist das Problem (P) nicht zulässig, d. h. es gibt kein x, das der Restriktion $\|Cx - d\|_2 \leq \gamma$ genügt. Daher ist (P) in diesem Fall trivialerweise nicht lösbar. Sei daher $\min_{x\in\mathbb{R}^n} \|Cx - d\|_2 \leq \gamma$ bzw. (P) zulässig. Der Einfachheit[8] halber setzen wir voraus, dass

$$\text{Rang} \begin{pmatrix} A \\ C \end{pmatrix} = n.$$

Sei $\{x_k\}$ eine *Minimalfolge* für (P), d. h. es sei $\{x_k\} \subset M$ bzw. $\|Cx_k - d\|_2 \leq \gamma$ für alle k und $\|Ax_k - b\|_2 \to \inf_{x\in M} \|Ax - b\|_2$. Dann sind $\{Ax_k\}$ und $\{Cx_k\}$ beschränkt. Aus der (für diesen Schluss nötigen, sonst aber unnötigen) Rangvoraussetzung schließen wir, dass auch $\{x_k\}$ beschränkt ist. Jeder der dann existierenden Häufungspunkte von $\{x_k\}$ ist eine Lösung von (P).

Nun nehmen wir zusätzlich an, dass die quadratische Restriktion bindend und die obige Rangvoraussetzung erfüllt ist. Sei $x^* \in M$ eine Lösung von (P). Wir wollen zeigen, dass notwendigerweise $\|Cx^* - d\|_2 = \gamma$. Angenommen, dies wäre nicht der Fall, es wäre also $\|Cx^* - d\|_2 < \gamma$. Da die quadratische Restriktion als bindend vorausgesetzt wurde, ist $\|Cx_{A,C} - d\|_2 > \gamma$ (siehe die oben eingeführte Bezeichnung). Insbesondere ist x^* keine Lösung der (unrestringierten) Aufgabe, $\|Ax - b\|_2$ auf dem $\mathbb{R}^n$ zu minimieren und folglich $\|Ax_{A,C} - b\|_2 < \|Ax^* - b\|_2$. Definiert man $x(t) := (1 - t)x^* + tx_{A,C}$ für $t \in [0, 1]$, so erkennt man die Existenz eines $t_0 \in (0, 1)$ mit $\|Cx(t_0) - d\|_2 = \gamma$. Also ist $x(t_0)$ zulässig für (P). Andererseits ist

$$\|Ax(t_0) - b\|_2 \leq (1 - t_0)\,\|Ax^* - b\|_2 + t_0\,\|Ax_{A,C} - b\|_2 < \|Ax^* - b\|_2,$$

was ein Widerspruch dazu ist, dass x^* eine Lösung von (P) ist. Man beachte, dass hier die Rangvoraussetzung nicht ausgenutzt und damit schon der erste Teil der dritten Aussage bewiesen ist. Angenommen nun, x^{**} sei eine weitere Lösung von (P). Dann ist auch $\frac{1}{2}(x^* + x^{**})$ eine Lösung von (P) und daher nach dem gerade eben bewiesenen nicht nur

$$\|Ax^* - b\|_2 = \|Ax^{**} - b\|_2 = \|A(\tfrac{1}{2}(x^* + x^{**})) - b\|_2,$$

[8] Der allgemeine Fall folgt aus der Beobachtung, dass (P) äquivalent einem konvexen, quadratisch restringierten quadratischen Programm mit endlichem Optimalwert ist. Solche Programme sind lösbar, wie man (allerdings ist dies nichttrivial) aus der Optimierung weiß.

sondern auch

$$\|Cx^* - d\|_2 = \|Cx^{**} - d\|_2 = \|C(\tfrac{1}{2}(x^* + x^{**})) - d\|_2.$$

Folglich ist nach einfacher Rechnung

$$\begin{aligned} 0 &= \|A(\tfrac{1}{2}(x^* + x^{**})) - b\|_2^2 - \tfrac{1}{2}[\|Ax^* - b\|_2^2 + \|Ax^{**} - b\|_2^2] \\ &= \tfrac{1}{8}(x^* - x^{**})^T A^T A(x^* - x^{**}) \\ &= \tfrac{1}{8}\,\|A(x^* - x^{**})\|_2^2, \end{aligned}$$

also $A(x^* - x^{**}) = 0$. Aus demselben Grund ist $C(x^* - x^{**}) = 0$, insgesamt also

$$\begin{pmatrix} A \\ C \end{pmatrix} (x^* - x^{**}) = 0.$$

Aus der Rangvoraussetzung folgt $x^* = x^{**}$ und das ist die behauptete Eindeutigkeit einer Lösung von (P). Damit ist der erste Teil des Satzes bewiesen.

Nun nehmen wir an, dass es zu einem für (P) zulässigen x^* ein $\lambda^* \geq 0$ mit

$$(A^T A + \lambda^* C^T C)x^* = A^T b + \lambda^* C^T d, \qquad \lambda^*[\gamma - \|Cx^* - d\|_2] = 0$$

gibt. Die Aufgabe (P) ist äquivalent zu

$$\begin{cases} \text{Minimiere} \quad f(x) := \tfrac{1}{2}\|Ax - b\|_2^2 \quad \text{unter der Nebenbedingung} \\ \qquad\qquad\qquad g(x) := \tfrac{1}{2}\|Cx - d\|_2^2 - \tfrac{1}{2}\gamma^2 \leq 0. \end{cases}$$

Sei nun $x \in M$ beliebig. Dann ist

$$\begin{aligned} f(x) - f(x^*) &= \nabla f(x^*)^T(x - x^*) + \tfrac{1}{2}\|A(x^* - x^{**})\|_2^2 \\ &\geq [A^T(Ax^* - b)]^T(x - x^*) \\ &= -\lambda^*[C^T(Cx^* - d)]^T(x - x^*) \\ &= -\lambda^* \nabla g(x^*)^T(x - x^*) \\ &\geq -\lambda^*[g(x) - g(x^*)] \\ &= -\lambda^* g(x) \\ &\geq 0, \end{aligned}$$

also x^* eine Lösung von (P). Damit ist auch der zweite Teil des Satzes bewiesen.

Im dritten Teil des Satzes wird vorausgesetzt, dass die quadratische Restriktion in (P) bindend und $x^* \in M$ eine Lösung von (P) ist. Wir haben im ersten Teil schon bewiesen, dass notwendigerweise $\|Cx^* - d\|_2 = \gamma$, insbesondere ist x^* auch eine Lösung der Aufgabe, $\|Ax - b\|_2$ unter der Gleichungs-Nebenbedingung $\|Cx - d\|_2 = \gamma$ zu minimieren. Seien f und g wie gerade eben definiert. Dann ist $\nabla g(x^*) \neq 0$, denn andernfalls wäre

$$\gamma = \|Cx^* - d\|_2 = \min_{x \in \mathbb{R}^n} \|Cx - d\|_2,$$

ein Widerspruch zur Voraussetzung. Die Lagrangesche Multiplikatorenregel, angewandt auf die Aufgabe, $f(x)$ unter der Nebenbedingung $g(x) = 0$ zu minimieren, liefert die

Existenz eines $\lambda^* \in \mathbb{R}$ mit $\nabla f(x^*) + \lambda^* \nabla g(x^*) = 0$. Es ist $\lambda^* \neq 0$, denn andernfalls wäre $x^* \in S$ und damit die quadratische Restriktion nicht bindend. Angenommen, es wäre $\lambda^* < 0$. Wegen $\min_{x \in \mathbb{R}^n} \|Cx - d\|_2 < \gamma$ existiert ein $\hat{x} \in \mathbb{R}^n$ mit $g(\hat{x}) < 0$. Dann ist aber

$$\begin{aligned} \nabla f(x^*)^T(\hat{x} - x^*) &= -\lambda^* \nabla g(x^*)^T(\hat{x} - x^*) \\ &\leq \underbrace{-\lambda^*}_{>0}[\underbrace{g(\hat{x})}_{<0} - \underbrace{g(x^*)}_{=0}] \\ &< 0. \end{aligned}$$

Hieraus folgt aber im Widerspruch zur Voraussetzung, dass x^* keine Lösung von (P) ist. Denn einerseits ist $(1-t)x^* + t\hat{x} \in M$, also zulässig für (P), für alle $t \in [0,1]$, anderseits ist $f(x^* + t(\hat{x} - x^*)) < f(x^*)$ für alle hinreichend kleinen $t > 0$. Damit ist der Satz schließlich bewiesen. □

Bemerkung: Der letzte Satz gibt eine prinzipielle Möglichkeit, eine Lösung des quadratisch restringierten linearen Ausgleichsproblems (P) zu berechnen. Es werde angenommen, dass die quadratische Restriktion bindend ist und die Rangvoraussetzung aus dem ersten Teil von Satz 6.2 erfüllt ist. Für vorgegebenes $\lambda > 0$ besitzt das lineare Gleichungssystem

$$(*) \qquad (A^TA + \lambda C^TC)x = A^Tb + \lambda C^Td$$

genau eine Lösung x_λ. Die Wahl des "richtigen" λ ist dann zurückgeführt auf die Bestimmung einer positiven Nullstelle von

$$f(\lambda) := \|Cx_\lambda - d\|_2 - \gamma,$$

was etwa mit Hilfe des Newton-Verfahrens erfolgen kann. Günstiger ist es i. allg., das Newton-Verfahren auf die Nullstellenaufgabe für

$$h(\lambda) := \frac{1}{\gamma} - \frac{1}{\|Cx_\lambda - d\|_2}$$

anzuwenden. Wichtig, wie immer in diesem Zusammenhang, ist die Bemerkung, dass man x_λ "nicht wörtlich" direkt aus dem linearen Gleichungssystem berechnen sollte. Besser ist es, $(*)$ als die Normalgleichungen zum linearen Ausgleichsproblem

$$\text{Minimiere} \quad \left\| \begin{pmatrix} A \\ \sqrt{\lambda}C \end{pmatrix} - \begin{pmatrix} b \\ \sqrt{\lambda}d \end{pmatrix} \right\|, \quad x \in \mathbb{R}^n$$

zu entlarven und entsprechend zu lösen. Unter der hier gemachten Voraussetzung

$$\text{Rang} \begin{pmatrix} A \\ \sqrt{\lambda}C \end{pmatrix} = n$$

hat die Koeffizientenmatrix dieses linearen Ausgleichsproblems vollen Rang, man könnte es also z. B. mit einer QR-Zerlegung dieser Matrix bestimmen. Da aber nur der untere Teil der Matrix von λ abhängt, besteht die Hoffnung, dass dies effizienter geht. Hinweise findet man (für $C = I$ in Aufgabe 2) und bei Å. BJÖRK (1996, S. 208 ff.). □

Als Korollar formulieren wir die Spezialisierung der Aussagen von Satz 6.2 auf ein quadratisch restringiertes lineares Ausgleichsproblem in Standardform.

Korollar 6.3 *Gegeben sei das quadratisch restringierte lineare Ausgleichsproblem in Standardform*

(P) *Minimiere* $\|Ax - b\|_2$ *unter der Nebenbedingung* $\|x\|_2 \leq \gamma$,

wobei $A \in \mathbb{R}^{m\times n}$ *mit* $m \geq n$, $b \in \mathbb{R}^m$ *und* $\gamma > 0$. *Dann gilt:*

1. *Die Aufgabe (P) besitzt eine Lösung, die eindeutig ist, wenn* $\|A^+ b\|_2 > \gamma$.
2. *Ist* $\|x^*\|_2 \leq \gamma$ *und existiert ein* $\lambda^* \geq 0$ *mit*

$$(A^T A + \lambda^* I)x^* = A^T b, \qquad \lambda^*(\gamma - \|x^*\|_2) = 0,$$

   so ist x^* *eine Lösung von (P).*
3. *Sei* $\|A^+ b\|_2 > \gamma$. *Ist dann* x^* *eine Lösung von (P), so ist notwendigerweise* $\|x^*\|_2 = \gamma$, *ferner existiert ein* $\lambda^* > 0$ *mit*

$$(A^T A + \lambda^* I)x^* = A^T b.$$

Für ein quadratisch restringiertes Ausgleichsproblem in Standardform legt das letzte Korollar die folgende Vorgehensweise nahe:

- Berechne eine (reduzierte) Singulärwertzerlegung von A, also $\hat{U} \in \mathbb{R}^{m\times n}$ mit $\hat{U}^T\hat{U} = I$, eine orthogonale Matrix $V \in \mathbb{R}^{n\times n}$ und eine Diagonalmatrix $\hat{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$ mit

$$\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0$$

  derart, dass $A = \hat{U}\hat{\Sigma}V^T$.
- Berechne $b := \hat{U}^T b$ (die Matrix $\hat{U}$ kann danach vergessen werden).
- Falls $\sum_{i=1}^r (b_i/\sigma_i)^2 \leq \gamma^2$, dann: $x^* := \sum_{i=1}^r (b_i/\sigma_i)v_i$ ist Lösung des quadratisch restringierten linearen Ausgleichsproblems in Standardform.
- Andernfalls bestimme man eine $\lambda^* > 0$ mit

$$\sum_{i=1}^r \Big(\frac{\sigma_i b_i}{\sigma_i^2 + \lambda^*}\Big)^2 = \gamma^2$$

  und berechne die Lösung

$$x^* := \sum_{i=1}^r \frac{\sigma_i b_i}{\sigma_i^2 + \lambda^*} v_i$$

  von (P).

Jetzt stellt sich natürlich die Frage, wie λ^* als Lösung der angegebenen Gleichung bestimmt werden sollte. Hierzu definieren wir

$$\phi(\lambda) := \Big[\sum_{i=1}^{r}\Big(\frac{\sigma_i b_i}{\sigma_i^2+\lambda}\Big)^2\Big]^{1/2}.$$

Offenbar ist $\phi(0) > \gamma$ (andernfalls wäre man schon im Schritt vorher ausgestiegen) und $\lim_{\lambda\to\infty}\phi(\lambda) = 0 < \gamma$. Außerdem ist

$$\phi'(\lambda) = -\frac{1}{\phi(\lambda)}\sum_{i=1}^{r}\frac{(\sigma_i b_i)^2}{(\sigma_i^2+\lambda)^3},$$

also $\phi(\cdot)$ auf $[0,\infty)$ monoton fallend. Insgesamt hat die Gleichung $\phi(\lambda) = \gamma$ also genau eine Nullstelle $\lambda^* \in (0,\infty)$. Weiter ist

$$\begin{aligned}
\phi''(\lambda) &= \frac{3}{\phi(\lambda)}\sum_{i=1}^{r}\frac{(\sigma_i b_i)^2}{(\sigma_i^2+\lambda)^4} + \frac{\phi'(\lambda)}{\phi(\lambda)^2}\sum_{i=1}^{r}\frac{(\sigma_i b_i)^2}{(\sigma_i^2+\lambda)^3} \\
&= \frac{3}{\phi(\lambda)}\sum_{i=1}^{r}\frac{(\sigma_i b_i)^2}{(\sigma_i^2+\lambda)^4} - \frac{1}{\phi(\lambda)^3}\Big(\sum_{i=1}^{r}\frac{(\sigma_i b_i)^2}{(\sigma_i^2+\lambda)^3}\Big)^2 \\
&= \frac{2}{\phi(\lambda)}\sum_{i=1}^{r}\frac{(\sigma_i b_i)^2}{(\sigma_i^2+\lambda)^4} \\
&\qquad + \frac{1}{\phi(\lambda)^3}\Big[\underbrace{\sum_{i=1}^{r}\frac{(\sigma_i b_i)^2}{(\sigma_i^2+\lambda)^2}}_{=\phi(\lambda)^2}\sum_{i=1}^{r}\frac{(\sigma_i b_i)^2}{(\sigma_i^2+\lambda)^4} - \Big(\sum_{i=1}^{r}\frac{(\sigma_i b_i)^2}{(\sigma_i^2+\lambda)^4}\Big)^2\Big] \\
&\geq \frac{2}{\phi(\lambda)}\sum_{i=1}^{r}\frac{(\sigma_i b_i)^2}{(\sigma_i^2+\lambda)^4} \\
&\qquad \text{(Cauchy-Schwarzsche Ungleichung),}
\end{aligned}$$

also $\phi(\cdot)$ auf $[0,\infty)$ konvex. Es würde nun naheliegen, das Newton-Verfahren auf

$$f(\lambda) := \phi(\lambda) - \gamma = 0$$

anzuwenden, ausgehend etwa vom Startwert $\lambda_0 = 0$. Dies ist aber ineffizient, da ϕ rechts von Null wegen

$$\phi'(0) = -\frac{1}{\phi(0)}\sum_{i=1}^{r}\frac{b_i^2}{\sigma_i^4} = -\sum_{i=1}^{r}\frac{b_i^2}{\sigma_i^4} \Big/ \Big(\sum_{i=1}^{r}\frac{b_i^2}{\sigma_i^2}\Big)^{1/2}$$

insbesondere bei kleinem σ_r eine große negative Steigung besitzt, also sehr stark abfällt. Ein Newton-Schritt würde die Nullstelle von f stark unterschätzen. Besser ist es, bei Vorliegen einer aktuellen Näherung λ die Funktion $f(\cdot)$ in einer Umgebung von λ durch eine rationale Funktion g mit

$$g(t) = \frac{\alpha}{\beta + t} - \gamma$$

zu approximieren und deren Nullstelle als neue aktuelle Näherung λ_+ zu nehmen. Die Parameter α und β kann man etwa durch die Interpolationsforderungen

$$g(\lambda) = f(\lambda)(= \phi(\lambda) - \gamma), \qquad g'(\lambda) = f'(\lambda)(= \phi'(\lambda))$$

bzw.

$$\frac{\alpha}{\beta + \lambda} = \phi(\lambda), \qquad -\frac{\alpha}{(\beta + \lambda)^2} = \phi'(\lambda)$$

bestimmen, was auf

$$\alpha = -\frac{\phi(\lambda)^2}{\phi'(\lambda)}, \qquad \beta = -\lambda - \frac{\phi(\lambda}{\phi'(\lambda)}$$

führt. Bestimmt man anschließend λ_+ als Nullstelle von g, so erhält man

$$\lambda_+ := \lambda - \Big[\frac{\phi(\lambda) - \gamma}{\gamma}\Big] \frac{\phi(\lambda)}{\phi'(\lambda)}.$$

Interessant ist, dass man dieselbe Iterationsvorschrift erhält, wenn man das Newton-Verfahren auf die Nullstellenaufgabe

$$h(\lambda) := \frac{1}{\gamma} - \frac{1}{\phi(\lambda)} = 0$$

anwendet. Man kann zeigen, dass auch h monoton fallend und konvex ist (siehe Aufgabe 6).

3.6.3 Die verallgemeinerte Singulärwertzerlegung

In den letzten beiden Abschnitten betrachteten wir Aufgaben der Form

$$\text{Minimiere} \quad \|Ax - b\|_2^2 + \tau \|Lx\|_2^2, \quad x \in \mathbb{R}^n$$

und

$$\text{Minimiere} \quad \|Ax - b\|_2 \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : \|Cx - d\|_2 \le \gamma\}.$$

Gemeinsam ist beiden Aufgaben, dass zu den Daten *zwei* Matrizen gehören, deren Spaltenzahl übereinstimmt. Die verallgemeinerte Singulärwertzerlegung zu diesem Paar von Matrizen ist hier das richtige Werkzeug. Zunächst geben wir die sogenannte CS-Zerlegung einer Matrix mit orthonormalen Spalten an. Die Darstellung folgt der bei Å. Björk (1996, S. 155).

Satz 6.4 *Die Matrix $Q \in \mathbb{R}^{(m+p)\times n}$ mit $p \le n \le m$ besitze orthonormale Spalten und sei durch*

$$Q = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} \begin{matrix} \}m \\ \}p \end{matrix}$$

zerlegt. Es sei also

$$Q^T Q = Q_1^T Q_1 + Q_2^T Q_2 = I_n.$$

Dann gibt es orthogonale Matrizen $U_1 \in \mathbb{R}^{m\times m}$, $U_2 \in \mathbb{R}^{p\times p}$ *und* $V \in \mathbb{R}^{n\times n}$ *sowie nichtnegative Diagonalmatrizen*

$$C = \operatorname{diag}(c_1, \ldots, c_p), \qquad S = \operatorname{diag}(s_1, \ldots, s_p)$$

mit $C^2 + S^2 = I_p$ *derart, dass*

$$\begin{pmatrix} U_1^T & 0 \\ 0 & U_2^T \end{pmatrix} \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} V = \begin{pmatrix} U_1^T Q_1 V \\ U_2^T Q_2 V \end{pmatrix} = \begin{pmatrix} \Sigma_1 \\ \Sigma_2 \end{pmatrix} \begin{matrix} \}m \\ \}p \end{matrix}$$

mit

$$\Sigma_1 = \begin{pmatrix} C & 0 \\ 0 & I \\ 0 & 0 \end{pmatrix} \begin{matrix} \}p \\ \}n-p \\ \}m-n \end{matrix}, \qquad \Sigma_2 = \begin{pmatrix} S & 0 \end{pmatrix} \}p.$$

Ferner können die Diagonalelemente c_i *und* s_i *von* C *bzw.* S *so bestimmt werden, dass*

$$c_i = \cos(\theta_i), \quad s_i = \sin(\theta_i) \quad (i = 1, \ldots, p), \qquad 0 \le \theta_1 \le \cdots \le \theta_p \le \pi/2.$$

Beweis: Die Matrix $Q_1 \in \mathbb{R}^{m\times n}$ besitzt eine (volle) Singulärwertzerlegung der Form $Q_1 = U_1 \Sigma_1 V^T$ mit orthogonalen $U_1 \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ und einer (rechteckigen) Diagonalmatrix $\Sigma_1 \in \mathbb{R}^{m\times n}$. Wegen $Q_1^T Q_1 + Q_2^T Q_2 = I_n$ liegen die singulären Werte von Q_1 (und auch von Q_2) im Intervall $[0, 1]$, aus dem selben Grund sind von den n singulären Werten von Q_1 mindestens $n - p$ gleich 1. Wir können daher annehmen, dass die singulären Werte c_i von Q_1 so angeordnet sind, dass

$$\Sigma_1 = \begin{pmatrix} C & 0 \\ 0 & I \\ 0 & 0 \end{pmatrix} \begin{matrix} \}p \\ \}n-p \\ \}m-n \end{matrix}.$$

Wir setzen $\tilde{Q}_2 := Q_2 V$. Dann hat die Matrix

$$\begin{pmatrix} \Sigma_1 \\ \tilde{Q}_2 \end{pmatrix} = \begin{pmatrix} U_1^T & 0 \\ 0 & I_p \end{pmatrix} \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} V$$

orthonormale Spalten. Also ist $\Sigma_1^T \Sigma_1 + \tilde{Q}_2^T \tilde{Q}_2 = I_n$, daher ist $\tilde{Q}_2^T \tilde{Q}_2 = I_n - \Sigma_1^T \Sigma_1$ eine Diagonalmatrix, so dass $\tilde{Q}_2 = \begin{pmatrix} \tilde{q}_1^{(2)} & \cdots & \tilde{q}_n^{(2)} \end{pmatrix}$ orthogonale Spalten besitzt. Wir nehmen an, es sei $c_r < c_{r+1} = 1$ mit einem $r \in \{1, \ldots, p\}$ (hier ist auch $r = p$ möglich, nämlich genau dann, wenn Q_1 den $(n-p)$-fachen singulären Wert 1 besitzt). Nun definiere man die Matrix $U_2 = \begin{pmatrix} u_1^{(2)} & \cdots & u_p^{(2)} \end{pmatrix}$ wie folgt. Wegen $\|\tilde{q}_j^{(2)}\|_2^2 = 1 - c_j^2 \neq 0$, $j = 1, \ldots, r$, setze man

$$u_j^{(2)} := \tilde{q}_j^{(2)} / \|\tilde{q}_j^{(2)}\|_2, \qquad j = 1, \ldots, r.$$

Diese r Spalten bilden ein Orthonormalsystem im $\mathbb{R}^p$. Ist $r < p$, so ergänze man sie zu einer Orthonormalbasis des $\mathbb{R}^p$ und erhalte auf diese Weise die orthogonale Matrix $U_2 \in \mathbb{R}^{p\times p}$. Dann ist

$$U_2^T \tilde{Q}_2 = U_2^T Q_2 V = \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix}, \qquad S = \operatorname{diag}(s_1, \ldots, s_p)$$

mit

$$s_j := \begin{cases} (1-c_j^2)^{1/2}, & j = 1, \ldots, r, \\ 0, & j = r+1, \ldots, p. \end{cases}$$

Damit ist der Satz bewiesen. □

Bemerkung: Die Voraussetzung $p \leq n$ haben wir nur der Bequemlichkeit halber und wegen späterer Anwendungen gemacht. Eine etwas allgemeinere Formulierung findet man bei Å. BJÖRK (1996, S. 155 ff.). □

Im folgenden Satz wird die Existenz einer verallgemeinerten Singulärwertzerlegung zu einem Paar von Matrizen gleicher Spaltenzahl nachgewiesen. Wieder betrachten wir bei weitem nicht den allgemeinsten Fall.

Satz 6.5 *Seien $A \in \mathbb{R}^{m\times n}$ und $B \in \mathbb{R}^{p\times n}$ mit $p \leq n \leq m$ gegeben. Sei*

$$M := \begin{pmatrix} A \\ B \end{pmatrix}, \qquad \text{Rang } (M) = n.$$

Dann gibt es orthogonale Matrizen $U_A \in \mathbb{R}^{m\times m}$, $U_B \in \mathbb{R}^{p\times p}$ und eine nichtsinguläre Matrix $Z \in \mathbb{R}^{n\times n}$ derart, dass

$$U_A^T A = \begin{pmatrix} D_A \\ 0 \end{pmatrix} Z, \qquad U_B^T B = \begin{pmatrix} D_B & 0 \end{pmatrix} Z,$$

wobei

$$D_A = \text{diag } (\alpha_1, \ldots, \alpha_n), \qquad D_B = \text{diag } (\beta_1, \ldots, \beta_p)$$

mit

$$0 \leq \alpha_1 \leq \cdots \leq \alpha_n \leq 1, \qquad 1 \geq \beta_1 \geq \cdots \geq \beta_p \geq 0$$

und

$$\alpha_i^2 + \beta_i^2 = 1 \quad (i = 1, \ldots, p), \qquad \alpha_i = 1 \quad (i = p+1, \ldots, n).$$

Ferner stimmen die singulären Werte von Z mit den (notwendigerweise positiven) singulären Werten von M überein.

Beweis: Man bilde die (volle) Singulärwertzerlegung von M, also die Zerlegung

$$M = \begin{pmatrix} A \\ B \end{pmatrix} = Q \begin{pmatrix} \hat{\Sigma}_1 \\ 0 \end{pmatrix} P^T,$$

wobei $Q \in \mathbb{R}^{(m+p)\times(m+p)}$, $P \in \mathbb{R}^{n\times n}$ orthogonal sind und

$$\hat{\Sigma}_1 = \text{diag } (\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{n\times n}, \qquad \sigma_1 \geq \cdots \geq \sigma_n > 0.$$

Hier[9] haben wir die Voraussetzung Rang $(M) = n$ ausgenutzt. Man denke sich Q wie folgt partitioniert:

$$Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \begin{matrix} \}m \\ \}p \end{matrix}, \qquad \begin{matrix} Q_{11} \in \mathbb{R}^{m\times n}, & Q_{12} \in \mathbb{R}^{m\times(m+p-n)} \\ Q_{21} \in \mathbb{R}^{p\times n}, & Q_{22} \in \mathbb{R}^{p\times(m+p-n)} \end{matrix}.$$

[9] Bei Å. BJÖRK (1996, S. 157) wird $k :=$ Rang (M) gesetzt und eine allgemeinere Aussage formuliert. Wir verzichten darauf, weil bei unseren Anwendungen die Matrix M vollen Rang besitzt.

Dann ist

$$\begin{pmatrix} A \\ B \end{pmatrix} P = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \begin{pmatrix} \hat{\Sigma}_1 \\ 0 \end{pmatrix} = \begin{pmatrix} Q_{11} \\ Q_{21} \end{pmatrix} \hat{\Sigma}_1.$$

Nun sei

$$Q_{11} = U_A \begin{pmatrix} C & 0 \\ 0 & I \\ 0 & 0 \end{pmatrix} V^T, \qquad Q_{21} = U_B(\ S \ \ 0 \ )V^T$$

eine CS-Zerlegung von Q_{11} und Q_{21}. Einsetzen in die Singulärwertzerlegung von M liefert

$$AP = U_A \begin{pmatrix} C & 0 \\ 0 & I \\ 0 & 0 \end{pmatrix} V^T \hat{\Sigma}_1$$

und

$$BP = U_B(\ S \ \ 0 \ )V^T \hat{\Sigma}_1.$$

Die Aussage des Satzes folgt dann mit

$$D_A := \begin{pmatrix} C & 0 \\ 0 & I \end{pmatrix}, \qquad D_B := S, \qquad Z := V^T \hat{\Sigma}_1 P^T.$$

Aus der letzten Beziehung liest man ab, dass die singulären Werte von Z und M übereinstimmen. Der Satz ist bewiesen. □

Bemerkung: Seien $A \in \mathbb{R}^{m \times n}$ mit $m \geq n$ und eine nichtsinguläre Matrix $B \in \mathbb{R}^{n \times n}$ gegeben. Natürlich ist dann

$$\text{Rang} \begin{pmatrix} A \\ B \end{pmatrix} = n.$$

Wir wollen die verallgemeinerte Singulärwertzerlegung zum Paar (A, B) und die "normale" Singulärwertzerlegung zu AB^{-1} miteinander vergleichen. Sei etwa

$$U_A^T A = \begin{pmatrix} D_A \\ 0 \end{pmatrix} Z, \qquad U_B^T B = D_B Z$$

die verallgemeinerte Singulärwertzerlegung zu (A, B), also $U_A \in \mathbb{R}^{m \times m}$, $U_B \in \mathbb{R}^{n \times n}$ orthogonal, $Z \in \mathbb{R}^{n \times n}$ nichtsingulär und

$$D_A = \text{diag}\,(\alpha_1, \ldots, \alpha_n), \qquad D_B = \text{diag}\,(\beta_1, \ldots, \beta_n)$$

Diagonalmatrizen mit

$$0 \leq \alpha_1 \leq \cdots \leq \alpha_n \leq 1, \qquad 1 \geq \beta_1 \geq \cdots \geq \beta_n > 0$$

und

$$\alpha_i^2 + \beta_i^2 = 1 \qquad (i = 1, \ldots, n).$$

Dann ist

$$AB^{-1} = U_A \begin{pmatrix} D_A \\ 0 \end{pmatrix} Z Z^{-1} D_B^{-1} U_B^T = U_A \begin{pmatrix} D_A D_B^{-1} \\ 0 \end{pmatrix} U_B^T$$

eine Singulärwertzerlegung von AB^{-1}, wobei allerdings die singulären Werte $\gamma_i := \alpha_i/\beta_i$, $i = 1, \ldots, n$, von AB^{-1} in nichtfallender Anordnung erscheinen. □

Auf die numerische Berechnung einer verallgemeinerten Singulärwertzerlegung wollen wir hier nicht mehr eingehen, Hinweise (insbesondere neuere Literatur) findet man bei Å. Björk (1996, S. 159). Ferner sei angemerkt, dass in MATLAB 5.2 als neuer Befehl `[U,V,X,C,S]=gsvd(A,B)` aufgenommen wurde, durch den zu zwei Matrizen $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{p\times n}$ eine verallgemeinerte Singulärwertzerlegung berechnet werden kann.

Bemerkung: Sei eine verallgemeinerte Singulärwertzerlegung zum Paar (A, B) wie in Satz 6.5 gegeben. Mit

$$W := Z^{-1} = \begin{pmatrix} w_1 & \cdots w_n \end{pmatrix}$$

ist dann

$$(AW)^T(AW) = D_A^2, \qquad (BW)^T(BW) = \begin{pmatrix} D_B^2 & 0 \\ 0 & 0 \end{pmatrix}.$$

Ist Rang $(B) = p$, so ist D_B nichtsingulär bzw. $\beta_p > 0$, wie aus der Zerlegung

$$B = U_B \begin{pmatrix} D_B & 0 \end{pmatrix} Z$$

sofort folgt. Hieraus wiederum folgt leicht, dass

$$A^TAw_i = \left(\frac{\alpha_i}{\beta_i}\right)^2 B^TBw_i, \qquad i = 1, \ldots, p,$$

d. h. mit $\gamma_i := \alpha_i/\beta_i$ ist (γ_i^2, w_i), $i = 1, \ldots, p$, ein verallgemeinertes Eigenwert-Eigenvektorpaar zum Matrizenpaar (A^TA, B^TB). □

Als Anwendung der verallgemeinerten Singulärwertzerlegung betrachten wir die Aufgabe

(P) Minimiere $\|Ax - b\|_2$ auf $M := \{x \in \mathbb{R}^n : \|Cx - d\|_2 \le \gamma\}$.

Hierbei seien $A \in \mathbb{R}^{m\times n}$, $C \in \mathbb{R}^{p\times n}$ mit $p \le n \le m$ gegeben. Ferner sei

$$\text{Rang } (C) = p, \qquad \text{Rang } \begin{pmatrix} A \\ C \end{pmatrix} = n.$$

Die verallgemeinerte Singulärwertzerlegung

$$U_A^TA = \begin{pmatrix} D_A \\ 0 \end{pmatrix} Z, \qquad U_C^TC = \begin{pmatrix} D_C & 0 \end{pmatrix} Z$$

des Paares (A, C) sei bekannt. Hier sind also $U_A \in \mathbb{R}^{m\times m}$, $U_C \in \mathbb{R}^{p\times p}$ orthogonal, $Z \in \mathbb{R}^{n\times n}$ nichtsingulär, ferner stehen in

$$D_A = \text{diag } (\alpha_1, \ldots, \alpha_n), \qquad D_C = \text{diag } (\beta_1, \ldots, \beta_p)$$

die verallgemeinerten singulären Werte zu (A, C). Ferner kann angenommen werden (obwohl es nur zum Teil benutzt wird), dass

$$0 \le \alpha_1 \le \cdots \le \alpha_n \le 1, \qquad 1 \ge \beta_1 \ge \cdots \ge \beta_p \ge 0$$

und

$$\alpha_i^2 + \beta_i^2 = 1 \quad (i = 1, \ldots, p), \qquad \alpha_i = 1 \quad (i = p+1, \ldots, n).$$

Wir beachten, dass $\beta_p > 0$, da wir Rang $(C) = p$ vorausgesetzt haben.

Man mache die Variablentransformation

$$y = Zx$$

und berechne

$$\tilde{b} := U_A^T b, \qquad \tilde{d} := U_C^T d.$$

Benutzt man die Orthogonalität von U_A und U_C, so erkennt man, dass (P) äquivalent zur Aufgabe

$$\begin{cases} \text{Minimiere} & \left\| \begin{pmatrix} D_A \\ 0 \end{pmatrix} y - \tilde{b} \right\|_2 \quad \text{unter der Nebenbedingung} \\ & \|(\ D_C \quad 0\ )y - \tilde{d}\|_2 \le \gamma \end{cases}$$

ist. Hieraus liest man ab:

1. Die Menge
$$S := \left\{ x \in \mathbb{R}^n : \|Ax - b\|_2 = \min_{z \in \mathbb{R}^n} \|Az - b\|_2 \right\}$$
kann in der Form
$$S = \{ Z^{-1} y : y_i = \tilde{b}_i / \alpha_i \ \ (\alpha_i \neq 0) \}$$
geschrieben werden.

2. Definiert man $y = y_{A,C}$ durch
$$y_i := \begin{cases} \tilde{b}_i / \alpha_i, & \text{falls} \quad \alpha_i \neq 0, \quad i = 1, \ldots, n, \\ \tilde{d}_i / \beta_i, & \text{falls} \quad \alpha_i = 0, \quad i = 1, \ldots, p, \end{cases}$$
so ist $x_{A,C} := Z^{-1} y_{A,C}$ eine Lösung von
$$\text{Minimiere} \quad \|Cx - d\|_2, \quad x \in S.$$

3. Wir nannten die quadratische Restriktion im quadratisch restringierten Ausgleichsproblem (P) *bindend*, wenn $\|Cx_{A,C} - d\|_2 > \gamma$. Nun ist
$$\begin{aligned} \|Cx_{A,C} - d\|_2^2 &= \|U_C^T C x_{A,C} - U_C^T d\|_2^2 \\ &= \left\| (\ D_C \quad 0\ ) Z x_{A,C} - \tilde{d} \right\|_2^2 \\ &= \left\| (\ D_C \quad 0\ ) y_{A,C} - \tilde{d} \right\|_2^2 \\ &= \sum_{\substack{i=1 \\ \alpha_i \neq 0}}^{p} (\beta_i \tilde{b}_i / \alpha_i - \tilde{d}_i)^2, \end{aligned}$$
so dass man bei Vorliegen der verallgemeinerten Singulärwertzerlegung leicht prüfen kann, ob die quadratische Restriktion bindend ist. Ist dies nicht der Fall, ist also $\|Cx_{A,C} - d\|_2 \le \gamma$, so ist $x_{A,C}$ eine Lösung des quadratisch restringierten linearen Ausgleichsproblems (P).

Nun gilt es, bei vorgegebenem $\lambda > 0$ die Lösung x_λ der verallgemeinerten Normalgleichungen

$$(A^T A + \lambda C^T C)x = A^T b + \lambda C^T d$$

zu bestimmen, um anschließend ein Nullstellenverfahren auf

$$f(\lambda) := \|Cx_\lambda - d\|_2 - \gamma$$

oder

$$h(\lambda) := \frac{1}{\gamma} - \frac{1}{\|Cx_\lambda - d\|_2}$$

anzuwenden. Nach wie vor gehen wir natürlich davon aus, dass die obige Singulärwertzerlegung vorliegt. Einsetzen liefert nach Multiplikation mit Z^{-T} von links das äquivalente lineare Gleichungssystem

$$\left[D_A^2 + \lambda \begin{pmatrix} D_C^2 & 0 \\ 0 & 0 \end{pmatrix}\right] Zx = \begin{pmatrix} D_A & 0 \end{pmatrix} U_A^T b + \lambda \begin{pmatrix} D_C \\ 0 \end{pmatrix} U_C^T d.$$

Mit den schon oben vorgenommenen Abkürzungen ist dies gleichwertig mit

$$\left[D_A^2 + \lambda \begin{pmatrix} D_C^2 & 0 \\ 0 & 0 \end{pmatrix}\right] y = \begin{pmatrix} D_A & 0 \end{pmatrix} \tilde{b} + \lambda \begin{pmatrix} D_C \tilde{d} \\ 0 \end{pmatrix}.$$

Komponentenweise geschrieben bedeutet dies, dass

$$(\alpha_i^2 + \lambda\beta_i^2)y_i = \alpha_i \tilde{b}_i + \lambda\beta_i \tilde{d}_i \quad (i = 1, \ldots, p), \qquad y_i = \tilde{b}_i \quad (i = p+1, \ldots, n).$$

Die Lösung x_λ der verallgemeinerten Normalgleichungen ist daher durch $x_\lambda := Z^{-1} y_\lambda$ gegeben, wobei

$$(y_\lambda)_i := \begin{cases} \dfrac{\alpha_i \tilde{b}_i + \lambda\beta_i \tilde{d}_i}{\alpha_i^2 + \lambda\beta_i^2}, & i = 1, \ldots, p, \\ \tilde{b}_i, & i = p+1, \ldots, n. \end{cases}$$

Daher ist

$$\begin{aligned} \|Cx_\lambda - d\|_2^2 &= \|U_C^T C x_\lambda - U_C^T d\|_2^2 \\ &= \|\begin{pmatrix} D_C & 0 \end{pmatrix} y_\lambda - \tilde{d}\|_2^2 \\ &= \sum_{i=1}^{p} [\beta_i (y_\lambda)_i - \tilde{d}_i]^2 \\ &= \sum_{i=1}^{p} \Big(\alpha_i \frac{\beta_i \tilde{b}_i - \alpha_i \tilde{d}_i}{\alpha_i^2 + \lambda\beta_i^2}\Big)^2. \end{aligned}$$

Man ist also jetzt dank der verallgemeinerten Singulärwertzerlegung in fast der gleichen Situation wie bei einem quadratisch restringierten linearen Ausgleichsproblem in Standardform. Ähnlich wie bei einem solchen definiere man hier die Funktion $\phi\colon [0, \infty)$ durch

$$\phi(\lambda) := \Big[\sum_{\substack{i=1 \\ \alpha_i \neq 0}}^{p} \Big(\alpha_i \frac{\beta_i \tilde{b}_i - \alpha_i \tilde{d}_i}{\alpha_i^2 + \lambda\beta_i^2}\Big)^2\Big]^{1/2},$$

so stellt man fest, dass

$$\phi(0) = \Big[\sum_{\substack{i=1\\ \alpha_i\neq 0}} \Big(\frac{\beta_i \tilde{b}_i}{\alpha_i} - \tilde{d}_i\Big)^2\Big]^{1/2}.$$

Ist also die quadratische Restriktion in (P) bindend, so ist $\phi(0) > \gamma$. Da weiter $\lim_{\lambda\to\infty} \phi(\lambda) = 0$ und ϕ auf $[0,\infty)$ monoton fallend ist, hat die Gleichung $\phi(\lambda) = \gamma$ genau eine Lösung.

3.6.4 MATLAB-Ergänzungen

Wir wollen eine MATLAB-Funktion schreiben, die bei gegebenen $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ mit $m \geq n$ und einem Regularisierungsparameter $\tau > 0$ die Lösung x_τ von

$$(\mathrm{P}_\tau) \qquad \text{Minimiere} \quad \|Ax-b\|_2^2 + \tau\|x\|_2^2, \quad x \in \mathbb{R}^n,$$

bestimmt. Hierbei wird ausgenutzt (siehe 3.6.1), dass

$$x_\tau = \sum_{i=1}^{r} \frac{\sigma_i\, u_i^T b}{\sigma_i^2 + \tau} v_i,$$

wobei $A = \hat{U}\hat{\Sigma}V^T$ eine reduzierte Singulärwertzerlegung von A ist.

```
function x=Tichonov(A,b,tau);
%************************************************
%Pre:    A is an m-by-n matrix
%        b is an m-vector
%        tau>0 is a regularization parameter
%Post:   x solves ||Ax-b||_2^2+tau ||x||_2^2-->min
%************************************************
[m,n]=size(A);
[U,S,V]=svd(A,0);
s=diag(S);
tol=m*max(s)*eps;
r=sum(s>tol);temp=U'*b;
x=V(:,1:r)*(s(1:r).*temp(1:r)./(s(1:r).^2+tau));
```

Wir geben eine MATLAB-Funktion an, welche das quadratisch restringierte Ausgleichsproblem in Standardform

$$(\mathrm{P}) \qquad \text{Minimiere} \quad \|Ax-b\|_2 \quad \text{unter der Nebenbedingung} \quad \|x\|_2 \leq \gamma,$$

löst, wobei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$, $b \in \mathbb{R}^m$ und $\gamma > 0$ gegeben sind. Wie wir gesehen haben (siehe 3.6.2), kommt es entscheidend darauf an, eine Nullstelle von $f(\lambda) := \phi(\lambda) - \gamma$ zu bestimmen, wobei

$$\phi(\lambda) := \Big[\sum_{i=1}^{r} \Big(\frac{\sigma_i u_i^T b_i}{\sigma_i^2 + \lambda}\Big)^2\Big]^{1/2}.$$

Hierbei ist natürlich $A = \hat{U}\hat{\Sigma}V^T$ eine reduzierte Singulärwertzerlegung von A. Durch das folgende M-file wird (bei speziellen A und b) die Funktion $\phi(\cdot)$ geplottet.

```
A=[22 10 2 3 7;14 7 10 0 8;-1 13 -1 -11 3;-3 -2 13 -2 4;
9 8 1 -2 4;9 1 -7 5 -1;2 -6 6 5 1;4 5 0 -2 2];
b=[-1;2;1;4;0;-3;1;0];
[m,n]=size(A);[U,S,V]=svd(A,0);s=diag(S);
tol=m*max(s)*eps;
r=sum(s>tol);b=U'*b;
lambda=linspace(0,2000);
for i=1:100
    phi(i)=sqrt(sum(((s(1:r).*b(1:r))./(s(1:r).^2+lambda(i))).^2));
end;
plot(lambda,phi);
```

Den Plot findet man in der nächsten Abbildung 3.4. Man erkennt sehr deutlich, wie

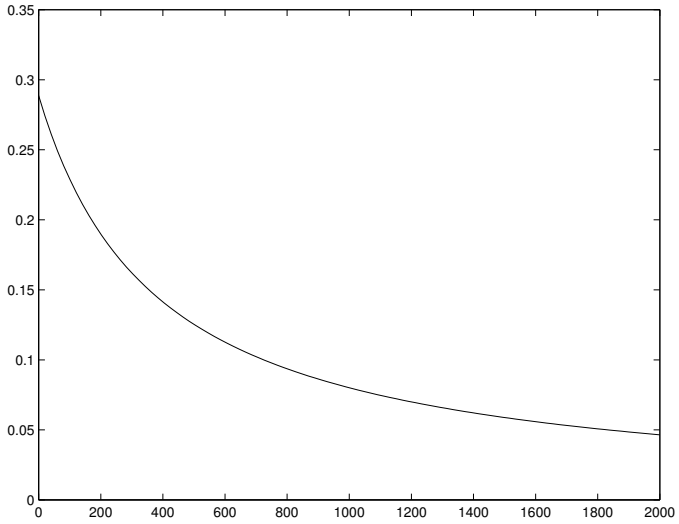


Abbildung 3.4: Die Funktion ϕ

schnell $\phi(\cdot)$ rechts von $\lambda = 0$ fällt, was der Grund dafür ist, dass das Newton-Verfahren, angewandt auf $f(\lambda) := \phi(\lambda) - \gamma$ und startend mit $\lambda_0 := 0$, die Nullstelle weit unterschätzen wird. Wesentlich günstiger ist es, das Newton-Verfahren auf die Funktion

$$h(\lambda) := \frac{1}{\gamma} - \frac{1}{\phi(\lambda)}$$

anzuwenden. In Abbildung 3.5 stellen wir die Funktion $-1/\phi(\cdot)$ dar. In der folgenden MATLAB-Funktion wird das Verfahren zur Lösung von (P) implementiert.

```
function [x,lambda,iter]=QuadRes(A,b,gamma);
%************************************************************
%Pre:     A is m-by-n matrix
```

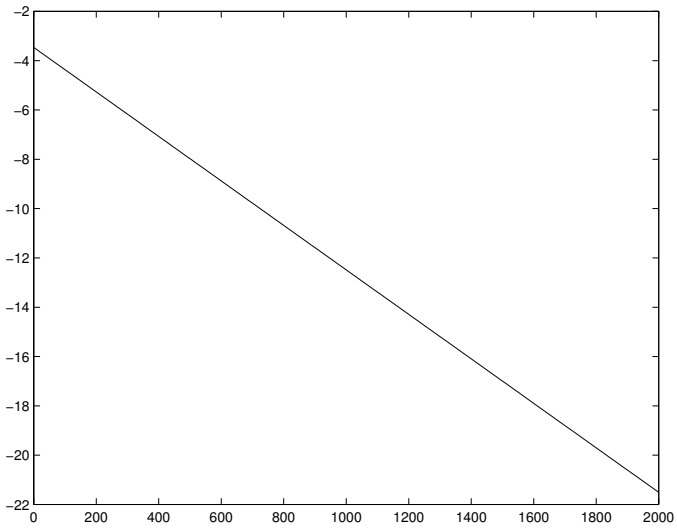


Abbildung 3.5: Die Funktion $-1/\phi$

```
%         b is m-vector
%         gamma>0
%Post:    x solves ||Ax-b||_2-->min subject to ||x||<=gamma
%         lambda is a corresponding multiplier
%         iter is the number of Newton iterates
%*************************************************************
[m,n]=size(A);
[U,S,V]=svd(A,0);
s=diag(S);tol=m*max(s)*eps;
r=sum(s>tol);c=U'*b;
if sum((c(1:r)./s(1:r)).^2)<=gamma^2
    x=V(:,1:r)*(c(1:r)./s(1:r));
    lambda=0;iter=0;
    return
end;
lambda=0;iter=0;
phil=sqrt(sum(((s(1:r).*c(1:r))./(s(1:r).^2+lambda)).^2));
while abs(phil-gamma)>eps
    iter=iter+1;
    if iter>50
        error('Not converged after 50 iterations.')
    end;
    phistr=-(1/phil)*sum(((s(1:r).*c(1:r)).^2)./(s(1:r).^2+lambda).^3);
    lambda=lambda-((phil-gamma)/gamma)*(phil/phistr);
    phil=sqrt(sum(((s(1:r).*c(1:r))./(s(1:r).^2+lambda)).^2));
    end;
x=V(:,1:r)*((s(1:r).*c(1:r))./(s(1:r).^2+lambda));
```

Wegen Abbildung 3.5 ist es nicht überraschend, dass bei obigen Daten für A und b für unterschiedliche γ jeweils nur eine Iteration benötigt wurde. Für das "normale" Newton-Verfahren benötigt man mehr Iterationen bis zum Abbruch, typischerweise 5 bis 10.

Nur erwähnt werden soll, dass es in MATLAB die Funktion `gsvd` zur Berechnung der verallgemeinerten Singulärwertzerlegung gibt. Nach `help gsvd` erfährt man u. a.:

```
gsvd   Generalized Singular Value Decomposition.
   [U,V,X,C,S] = gsvd(A,B) returns unitary matrices U and V,
   a (usually) square matrix X, and nonnegative diagonal matrices
   C and S so that

      A = U*C*X'
      B = V*S*X'
      C'*C + S'*S = I

   A and B must have the same number of columns, but may have
   different numbers of rows.  If A is m-by-p and B is n-by-p, then
   U is m-by-m, V is n-by-n and X is p-by-q where q = min(m+n,p).
```

3.6.5 Aufgaben

1. Gegeben sei die Aufgabe

   $$(\mathrm{P}_\tau) \qquad \text{Minimiere} \quad \|Ax-b\|_2^2 + \tau\,\|x\|_2^2, \quad x \in \mathbb{R}^n,$$

   wobei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$, $b \in \mathbb{R}^m$ und $\tau > 0$ gegeben sind. Sei x_τ die eindeutige Lösung von (P_τ). Mit Hilfe einer Singulärwertzerlegung von A berechne man

   $$f(\tau) := \|x_\tau\|_2, \qquad g(\tau) := \|Ax_\tau - b\|_2.$$

   Man zeige, dass g auf $[0,\infty)$ monoton nicht fallend und i. allg. monoton wachsend ist. Anschließend zeige man, dass $h := f \circ g^{-1}$ auf dem Existenzintervall eine monoton fallende Funktion ist.

2. Man präzisiere die folgende Vorgehensweise zur Lösung von

   $$(\mathrm{P}_\tau) \qquad \text{Minimiere} \quad \|Ax-b\|_2^2 + \tau\,\|x\|_2^2, \quad x \in \mathbb{R}^n$$

   bzw.

   $$(\mathrm{P}_\tau) \qquad \text{Minimiere} \quad \left\| \begin{pmatrix} A \\ \sqrt{\tau} I \end{pmatrix} x - \begin{pmatrix} b \\ 0 \end{pmatrix} \right\|_2, \quad x \in \mathbb{R}^n,$$

   wobei $\tau > 0$ und $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$, $b \in \mathbb{R}^m$.

   - Durch ein auf A angewandtes Bidiagonalisierungsverfahren reduziere man die Aufgabe (P_τ) auf das lineare Ausgleichsproblem

     $$(\hat{\mathrm{P}}_\tau) \qquad \text{Minimiere} \quad \left\| \begin{pmatrix} B \\ \sqrt{\tau} I \end{pmatrix} y - \begin{pmatrix} c \\ 0 \end{pmatrix} \right\|_2, \quad y \in \mathbb{R}^n,$$

wobei $B \in \mathbb{R}^{n\times n}$ eine obere Bidiagonalmatrix und $c \in \mathbb{R}^n$ ist. Dieser Teil ist also von der Wahl von $\tau > 0$ unabhängig.

- Man überlege sich, wie man die Aufgabe $(\hat{\mathrm{P}}_\tau)$ effizient lösen kann, indem man die Koeffizientenmatrix durch Multiplikation mit geeigneten Givens-Rotationen in eine obere Bidiagonalmatrix überführt.

3. Gegeben sei die Aufgabe

$$\text{Minimiere} \quad \|Ax - b\|_2^2 + \tau \|Lx\|_2^2, \quad x \in \mathbb{R}^n,$$

wobei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$, $b \in \mathbb{R}^m$, $L \in \mathbb{R}^{p\times n}$ mit Rang $(L) = p \leq n$ sowie

$$\text{Rang} \begin{pmatrix} A \\ L \end{pmatrix} = n$$

und $\tau > 0$ gegeben sind. Wir wissen, dass diese Aufgabe auch als lineares Ausgleichsproblem

$$(\mathrm{P}_\tau) \qquad \text{Minimiere} \quad \left\| \begin{pmatrix} A \\ \sqrt{\tau}L \end{pmatrix} x - \begin{pmatrix} b \\ 0 \end{pmatrix} \right\|_2, \quad x \in \mathbb{R}^n,$$

geschrieben werden kann.

(a) Man zeige, dass die folgenden Schritte[10] durchführbar sind:

- Berechne eine QR-Zerlegung von $L^T \in \mathbb{R}^{n\times p}$, also

$$L^T = V \begin{pmatrix} R \\ 0 \end{pmatrix}, \qquad V = \begin{pmatrix} V_1 & V_2 \end{pmatrix}.$$

  Hierbei sind natürlich $V_1 \in \mathbb{R}^{n\times p}$ und $V_2 \in \mathbb{R}^{n\times(n-p)}$.
- Berechne $AV_2 \in \mathbb{R}^{m\times(n-p)}$ und eine QR-Zerlegung dieser Matrix:

$$AV_2 = Q \begin{pmatrix} U \\ 0 \end{pmatrix}, \qquad Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}.$$

  Hierbei sind natürlich $Q_1 \in \mathbb{R}^{m\times(n-p)}$ und $Q_2 \in \mathbb{R}^{m\times(m-n+p)}$. Man zeige, daß U nichtsingulär bzw. Rang $(AV_2) = n - p$ ist.
- Berechne

$$\tilde{A} := Q_2^T A V_1 R^{-T} \in \mathbb{R}^{(m-n+p)\times p}, \qquad \tilde{b} := Q_2^T b \in \mathbb{R}^{m-n+p}.$$

(b) Man zeige: Ist $\tilde{x}$ die Lösung des Problems

$$(\tilde{\mathrm{P}}_\tau) \qquad \text{Minimiere} \quad \left\| \begin{pmatrix} \tilde{A} \\ \sqrt{\tau}I \end{pmatrix} \tilde{x} - \begin{pmatrix} \tilde{b} \\ 0 \end{pmatrix} \right\|_2, \quad \tilde{x} \in \mathbb{R}^p$$

in Standardform, so ist

$$x := V_1 R^{-T} \tilde{x} + V_2 U^{-1} Q_1^T (b - A V_1 R^{-T} \tilde{x})$$

die Lösung von (P_τ).

[10] Siehe L. Eldén (1977) "Algorithms for the regularization of ill-conditioned least squares problems." BIT 17, 134–145.

4. Gegeben sei das quadratisch restringierte lineare Ausgleichsproblem

   (P) Minimiere $\|Ax-b\|_2$ auf $M := \{x \in \mathbb{R}^n : \|Cx-d\|_2 \le \gamma\}$,

   wobei $A \in \mathbb{R}^{m\times n}$ mit $m \ge n$, $C \in \mathbb{R}^{p\times n}$ und $\gamma > 0$. Sei S die Menge der Lösungen des unrestringierten linearen Ausgleichsproblems, $\|Ax-b\|_2$ auf dem $\mathbb{R}^n$ zu minimieren. Man zeige, dass die Aufgabe

   $$\text{Minimiere} \quad \|Cx-d\|_2, \quad x \in S$$

   mindestens eine Lösung $x_{A,C} \in S$ besitzt.

5. Gegeben sei das quadratisch restringierte lineare Ausgleichsproblem

   (P) Minimiere $\|Ax-b\|_2$ auf $M := \{x \in \mathbb{R}^n : \|Cx-d\|_2 \le \gamma\}$,

   wobei $A \in \mathbb{R}^{m\times n}$ mit $m \ge n$, $C \in \mathbb{R}^{p\times n}$ und $\gamma > 0$. Man zeige, dass die Voraussetzung

   $$\text{Rang} \begin{pmatrix} A \\ C \end{pmatrix} = n$$

   notwendig für die Eindeutigkeit einer Lösung von (P) ist.

6. Seien $A \in \mathbb{R}^{m\times n}$ mit $m \ge n$, $b \in \mathbb{R}^m$ und $\gamma > 0$ gegeben. Man definiere $\phi\colon [0,\infty) \longrightarrow \mathbb{R}$ durch

   $$\phi(\lambda) := \begin{cases} \|A^+ b\|_2, & \text{falls} \quad \lambda = 0, \\ \|(A^T A + \lambda I)^{-1} A^T b\|_2, & \text{falls} \quad \lambda > 0 \end{cases}$$

   und anschließend

   $$h(\lambda) := \frac{1}{\gamma} - \frac{1}{\phi(\lambda)}.$$

   Man zeige, dass h auf $[0,\infty)$ monoton fallend und konvex ist.

7. Man betrachte das quadratisch restringierte Ausgleichsproblem

   (P) Minimiere $\|Ax-b\|_2$ auf $M := \{x \in \mathbb{R}^n : \|Cx-d\|_2 \le \gamma\}$.

   Hierbei seien $A \in \mathbb{R}^{m\times n}$, $C \in R^{p\times n}$ mit $p \le n \le m$ gegeben, ferner seien $b \in \mathbb{R}^m$, $d \in \mathbb{R}^p$ und $\gamma > 0$. Ist dann Rang $(C) = p$, so besitzt (P) eine Lösung.

8. Gegeben sei das durch eine lineare Gleichung restringierte lineare Ausgleichsproblem

   (P) Minimiere $\|Ax-b\|_2$ auf $M := \{x \in \mathbb{R}^n : Bx = d\}$.

   Hierbei seien $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{p\times n}$ mit $p \le n \le m$ gegeben. Man zeige:

   (a) Ist $M \ne \emptyset$, also (P) zulässig, so besitzt (P) eine Lösung.

   (b) Ist

   $$\text{Rang } (B) = p, \qquad \text{Rang} \begin{pmatrix} A \\ B \end{pmatrix} = n,$$

   so ist (P) eindeutig lösbar.

(c) Man zeige, dass unter der die eindeutige Lösbarkeit von (P) garantierenden Rangvoraussetzung im letzten Aufgabenteil die Lösung von (P) mit Hilfe einer verallgemeinerten Singulärwertzerlegung zu (A, B) berechnet werden kann.

9. Gegeben seien die Matrizen $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{p\times n}$ mit $p \le n \le m$. Es wird vorausgesetzt, dass

$$\text{Rang}\,(B) = p, \qquad \text{Rang}\begin{pmatrix} A \\ B \end{pmatrix} = n.$$

Mit einem $\gamma > 0$ und vorgegebenen Vektoren $b \in \mathbb{R}^m$, $d \in \mathbb{R}^p$ betrachte man das lineare Ausgleichsproblem

$$(\mathrm{P}_\gamma) \qquad \text{Minimiere} \quad \left\| \begin{pmatrix} A \\ \gamma B \end{pmatrix} - \begin{pmatrix} b \\ \gamma d \end{pmatrix} \right\|_2, \quad x \in \mathbb{R}^n.$$

Man zeige:

(a) Die Aufgabe (P_γ) besitzt eine eindeutige Lösung x_γ.

(b) Der Limes $x^* = \lim_{\gamma\to\infty} x_\gamma$ existiert und ist die (nach Aufgabe 8b) eindeutige Lösung von

$$(\mathrm{P}) \qquad \text{Minimiere} \quad \|Ax - b\|_2 \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : Bx = d\}.$$

10. Seien

$$A := \begin{pmatrix} 22 & 10 & 2 & 3 & 7 \\ 14 & 7 & 10 & 0 & 8 \\ -1 & 13 & -1 & -11 & 3 \\ -3 & -2 & 13 & -2 & 4 \\ 9 & 8 & 1 & -2 & 4 \\ 9 & 1 & -7 & 5 & -1 \\ 2 & -6 & 6 & 5 & 1 \\ 4 & 5 & 0 & -2 & 2 \end{pmatrix}, \qquad b := \begin{pmatrix} -1 \\ 2 \\ 1 \\ 4 \\ 0 \\ -3 \\ 1 \\ 0 \end{pmatrix}$$

gegeben. Für $\gamma := 0.05, 0.15, 0.25$ berechne man eine Lösung des quadratisch restringierten linearen Ausgleichsproblems

$$(\mathrm{P}) \qquad \text{Minimiere} \quad \|Ax - b\|_2 \quad \text{unter der Nebenbedingung} \quad \|x\|_2 \le \gamma$$

in Standardform.

11. Man betrachte die Aufgabe

$$(\mathrm{P}) \qquad \text{Minimiere} \quad \|Ax - b\|_2^2 + \tau \|Lx\|_2^2, \quad x \in \mathbb{R}^n.$$

Hierbei seien $A \in \mathbb{R}^{m\times n}$, $L \in \mathbb{R}^{p\times n}$ mit $p \le n \le m$ gegeben. Ferner sei

$$\text{Rang}\,(L) = p, \qquad \text{Rang}\begin{pmatrix} A \\ L \end{pmatrix} = n.$$

Schließlich seien $\tau > 0$ und $b \in \mathbb{R}^m$ gegeben. Mit Hilfe einer verallgemeinerten Singulärwertzerlegung des Paares (A, B) berechne man die Lösung von (P).

Kapitel 4

Eigenwertaufgaben

Wir werden uns verhältnismäßig kurz fassen und die wichtigsten Methoden zur Lösung von Eigenwertaufgaben schildern. Hierbei werden wir uns i. allg. auf die Lösung des *reellen* Eigenwertproblems beschränken, also die Bestimmung (aller oder nur einiger) der Eigenwerte und zugehöriger Eigenvektoren einer Matrix $A \in \mathbb{R}^{n\times n}$. Im ersten Abschnitt betrachten wir unsymmetrische (oder besser: nicht notwendig symmetrische), danach symmetrische Matrizen.

4.1 Das unsymmetrische Eigenwertproblem

4.1.1 Theoretische Grundlagen

Wir nehmen an, dass die Leserin mit den gebräuchlichsten Vokabeln im Zusammenhang mit (Matrix-) Eigenwertaufgaben vertraut ist: Eigenwert, Eigenvektor, ähnliche Matrix, charakteristisches Polynom.

In diesem Unterabschnitt über theoretische Grundlagen zu nicht notwendig symmetrischen Eigenwertaufgaben beginnen wir mit kanonischen Formen bzw. Normalformen. Hier fragen wir danach, in welche *Normalform* die gegebene Matrix $A \in \mathbb{R}^{n\times n}$ durch eine Ähnlichkeitstransformation überführt werden kann. Die aus der linearen Algebra bekannteste Normalform für i. allg. unsymmetrische Matrizen ist die *Jordansche Normalform*. Das entsprechende Ergebnis zitieren wir hier nur aus Vollständigkeitsgründen und verzichten naheliegenderweise auf einen Beweis.

Satz 1.1 *Sei $A \in \mathbb{C}^{n\times n}$. Dann existiert eine nichtsinguläre Matrix $S \in \mathbb{C}^{n\times n}$ derart, dass $S^{-1}AS = J$ eine Blockdiagonalmatrix*

$$J = \operatorname{diag}\left(J_{n_1}(\lambda_1), \ldots, J_{n_r}(\lambda_r)\right)$$

mit

$$J_{n_i}(\lambda_i) := \begin{pmatrix} \lambda_i & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_i \end{pmatrix} \in \mathbb{C}^{n_i\times n_i}, \qquad i = 1, \ldots, r,$$

ist. Hierbei ist J bis auf die Reihenfolge der Blöcke eindeutig durch A bestimmt.

Für uns, insbesondere im Zusammenhang mit dem QR-Verfahren, ist die *Schursche Normalform* wesentlich wichtiger. Hier unterscheiden wir zwischen einer komplexen und einer reellen Version. Auch diese beiden Sätze geben wir ohne Beweis an (Beweise findet man z. B. bei J. W. Demmel (1997, S. 146 ff.) oder auch J. Werner (1992, S. 12 ff.)).

Satz 1.2 *Sei $A \in \mathbb{C}^{n\times n}$. Dann existiert eine unitäre Matrix $U \in \mathbb{C}^{n\times n}$ derart, dass die zu A unitär ähnliche Matrix $U^H AU$ eine obere Dreiecksmatrix (mit den Eigenwerten von A als Diagonalelementen) ist.*

Ein komplettes reelles Analogon (ersetze $\mathbb{C}$ durch $\mathbb{R}$ und unitär durch orthogonal) zu diesem Satz kann es nicht geben, da eine reelle Matrix auch komplexe Eigenwerte besitzen kann, die dann allerdings in konjugiert komplexen Paaren auftreten.

Satz 1.3 *Sei $A \in \mathbb{R}^{n\times n}$. Dann existiert eine orthogonale Matrix $Q \in \mathbb{R}^{n\times n}$ mit*

$$Q^T AQ = \begin{pmatrix} R_{11} & R_{12} & \cdots & R_{1m} \\ 0 & R_{22} & \cdots & R_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & R_{mm} \end{pmatrix},$$

wobei die Diagonalblöcke R_{ii} entweder 1×1-Matrizen oder 2×2-Matrizen mit einem Paar konjugiert komplexer (nicht reeller) Eigenwerte sind. Insbesondere ist eine reelle Matrix, die nur reelle Eigenwerte besitzt, orthogonal ähnlich zu einer oberen Dreiecksmatrix.

Die reelle Schursche Normalform lässt uns hoffen, dass man unter geeigneten Voraussetzungen an eine reelle Matrix $A \in \mathbb{R}^{n\times n}$ eine Folge $\{Q_k\}$ orthogonaler Matrizen bestimmen kann mit der Eigenschaft, dass $\{Q_k^T AQ_k\}$ gegen eine obere Block-Dreiecksmatrix konvergiert, deren Diagonalblöcke entweder 1×1-Matrizen (die reellen Eigenwerte von A) oder 2×2-Matrizen mit einem Paar konjugiert komplexer Eigenwerte (komplexe Eigenwerte von A) sind.

Beispiel: In MATLAB wird durch `[Q,R]=schur(A)` der Matrix A die Schursche Normalform $Q^H AQ = R$ zugeordnet, und zwar die komplexe bzw. die reelle, je nach dem, ob A komplex oder reell ist. Ist z. B.

$$A := \begin{pmatrix} 1 & 2 & 3 & 5 \\ 2 & 4 & 1 & 6 \\ 1 & 2 & -1 & 3 \\ 2 & 0 & 1 & 3 \end{pmatrix},$$

so erhält man

$$Q = \begin{pmatrix} -0.5329 & 0.2148 & 0.7880 & -0.2212 \\ -0.7384 & -0.4672 & -0.2561 & 0.4133 \\ -0.3114 & -0.0075 & -0.4443 & -0.8400 \\ -0.2716 & 0.8576 & -0.3407 & 0.2733 \end{pmatrix}$$

und

$$R = \begin{pmatrix} 8.0716 & -6.0137 & 2.9657 & -1.7722 \\ & 1.5879 & 0.4083 & -1.4707 \\ & & -1.5826 & -1.7953 \\ & & 0.0936 & -1.0769 \end{pmatrix}.$$

An dem 2×2 Block in R erkennt man, dass A ein Paar konjugiert komplexer Eigenwerte besitzt. □

Ohne Beweis (siehe z. B. J. WERNER (1992b, S. 3)) zitieren wir jetzt einen Satz von Gerschgorin.

Satz 1.4 *Sei $A \in \mathbb{C}^{n\times n}$. Für $i = 1, \ldots, n$ definiere man die sogenannten Gerschgorin-Kreise*

$$G_i := \{z \in \mathbb{C} : |z - a_{ii}| \le r_i\} \qquad \textit{mit} \quad r_i := \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|.$$

Dann gilt:

1. *Ist λ ein Eigenwert von A, so ist $\lambda \in \bigcup_{i=1}^{n} G_i$. Alle Eigenwerte von A sind also in der Vereinigung aller Gerschgorin-Kreise enthalten.*
2. *Hat die Vereinigung $\hat{G}$ von $m < n$ Kreisen G_i einen leeren Durchschnitt mit den restlichen $n - m$ Kreisen, so enthält $\hat{G}$ genau m Eigenwerte von A (jeden entsprechend seiner algebraischen Vielfachheit gezählt, d. h. seiner Vielfachheit als Nullstelle des charakteristischen Polynoms).*

Wir beenden diesen kurzen Unterabschnitt mit einigen wenigen Störungsaussagen. Die Frage ist also, wie sich Störungen in der Koeffizientenmatrix auf die zugehörigen Eigenwerte auswirken. Auch den folgenden Satz von Bauer-Fike geben wir ohne Beweis an (siehe z. B. J. WERNER (1992b, S. 8)).

Satz 1.5 *Sei $A \in \mathbb{C}^{n\times n}$ eine diagonalisierbare Matrix, d. h. es existiere eine nichtsinguläre Matrix $P \in \mathbb{C}^{n\times n}$ mit*

$$P^{-1}AP = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) =: D.$$

Ferner sei $\delta A \in \mathbb{C}^{n\times n}$ und λ ein Eigenwert von $A + \delta A$. Dann ist

$$\min_{j=1,\ldots,n} |\lambda - \lambda_j| \le \|P^{-1}(\delta A)P\| \le \kappa(P)\,\|\delta A\|.$$

Hierbei sei $\|\cdot\|$ die einer absoluten[1] Vektornorm zugeordnete Matrixnorm, ferner bedeute $\kappa(P) := \|P\|\,\|P^{-1}\|$ die Kondition von P bezüglich dieser Matrixnorm.

Der Satz von Bauer Fike sagt aus: Ist $A \in \mathbb{C}^{n\times n}$ durch eine Ähnlichkeitstransformation mit der nichtsingulären Matrix P diagonalisierbar (bzw. besitzt A ein System von n linear unabhängigen Eigenvektoren) und ist λ ein Eigenwert der gestörten Matrix

[1] Eine Vektornorm $\|\cdot\|$ auf $***^n$ heißt *absolut*, wenn $\||x|\| = \|x\|$ für alle $x \in \mathbb{C}^n$.

$A + \delta A$, so existiert ein Eigenwert λ_i von A mit $|\lambda - \lambda_i| \leq \kappa(P)\,\|\delta A\|$. Die Kondition der Matrix P, in deren Spalten die (linear unabhängigen) Eigenvektoren von A stehen, bestimmt also die "Störanfälligkeit" der Eigenwerte von A. Hier erkennen wir schon, dass die Aufgabe, die Eigenwerte einer hermitischen bzw. symmetrischen Matrix zu bestimmen, gut konditioniert ist. Denn diese sind durch eine unitäre bzw. orthogonale Ähnlichkeitstransformation diagonalisierbar und unitäre bzw. orthogonale Matrizen haben bezüglich der Spektralnorm minimal mögliche Kondition.

In den Aufgaben findet man weitere Störungssätze für die Eigenwerte nicht notwendig symmetrische Matrizen.

4.1.2 Vektoriteration, inverse Iteration, orthogonale Iteration

Vom Konzept her sehr einfach ist die Vektoriteration nach v. Mises (power method) zur Berechnung eines dominanten Eigenwertes und zugehörigen Eigenvektors. Hierzu nehmen wir an, die Matrix $A \in \mathbb{C}^{n\times n}$ sei diagonalisierbar, also

$$X^{-1}AX = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$$

mit nichtsingulärem $X = (\ x_1 \ \cdots \ x_n\ )$. Ferner sei

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$$

(bzw. λ_1 ein dominanter Eigenwert). Es gibt verschiedene Versionen der Vektoriteration, wir geben die folgende an (siehe J. W. Demmel (1997, S. 154)):

- Gegeben $x^{(0)} \in \mathbb{C}^n$, es wird vorausgesetzt, dass der zu x_1 gehörende Koeffizient einer Darstellung von $x^{(0)}$ als Linearkombination von $x_1, \ldots, x_n$ nicht verschwindet.
- Für $k = 0, 1, \ldots$:
  - Berechne $y^{(k+1)} := Ax^{(k)}$.
  - Berechne $x^{(k+1)} := y^{(k+1)}/\|y^{(k+1)}\|_2$.
  - Berechne $\lambda^{(k+1)} := (x^{(k+1)})^T Ax^{(k+1)}$.

Ist

$$x^{(0)} = \sum_{i=1}^{n} \alpha_i x_i$$

mit $\alpha_1 \neq 0$ die Darstellung des Startvektors $x^{(0)}$ als Linearkombination des linear unabhängigen Systems von Eigenvektoren $x_1, \ldots, x_n$, so ist

$$A^k x^{(0)} = \alpha_1 \lambda_1^k \Big[x_1 + \sum_{i=2}^{n} \frac{\alpha_i}{\alpha_1} \Big(\frac{\lambda_i}{\lambda_1}\Big)^k x_i \Big].$$

Weiter ist

$$x^{(k)} = A^k x^{(0)}/\|A^k x^{(0)}\|_2 \in \operatorname{span}\{A^k x^{(0)}\}, \qquad k = 1, 2, \ldots,$$

was man ganz leicht durch vollständige Induktion einsieht, ferner $\|x^{(k)}\|_2 = 1$, $k = 1, \ldots$. Dann ist

$$\begin{aligned}
\operatorname{dist}(x^{(k)}, \operatorname{span}\{x_1\}) &= \min_{\beta \in ***} \|x^{(k)} - \beta x_1\|_2 \\
&= \min_{\beta \in ***} \left\| \frac{A^k x^{(0)}}{\|A^k x^{(0)}\|_2} - \beta x_1 \right\| \\
&= \min_{\beta \in ***} \left\| \operatorname{sign}(\alpha_1 \lambda_1^k) \frac{x_1 + \sum_{i=2}^n (\alpha_i/\alpha_1)(\lambda_i/\lambda_1)^k x_i}{\|x_1 + \sum_{i=2}^n (\alpha_i/\alpha_1)(\lambda_i/\lambda_1)^k x_i\|} - \beta x_1 \right\| \\
&\leq \frac{1}{\|x_1 + \sum_{i=2}^n (\alpha_i/\alpha_1)(\lambda_i/\lambda_1)^k x_i\|} \sum_{i=2}^n \left| \frac{\alpha_i}{\alpha_1} \right| \left(\frac{|\lambda_i|}{|\lambda_1|} \right)^k \|x_i\|_2 \\
&= O((|\lambda_2|/|\lambda_1|)^k).
\end{aligned}$$

Wir können daher sagen, dass die Folge $\{x^{(k)}\}$ "der Richtung nach" gegen x_1 konvergiert. Ob dies Verfahren nützlich ist, hängt von dem Quotienten $|\lambda_2|/|\lambda_1|$ ab. Ist dies eine nahe bei 1 gegene Zahl, so ist die Konvergenz zu schlecht, andernfall kann es eventuell benutzt werden, vor allem dann, wenn eine gute Näherung zum Eigenvektor eines dominanten Eigenwertes bekannt ist. Ein Nachteil ist natürlich auch, dass Konvergenz, wenn überhaupt, nur gegen ein dominantes Eigenwert-Eigenvektor Paar stattfindet. Nicht so entscheidend ist die Voraussetzung, dass $x^{(0)}$ eine Komponente bezüglich x_1 haben sollte, weil dies sozusagen mit Wahrscheinlichkeit 1 der Fall ist.

Während die Vektoriteration eigentlich mehr als Motivation für das QR-Verfahren denn als praktische Methode von Bedeutung ist, ist die inverse Iteration nach Wielandt eine auch für die Praxis wichtige Methode. Wieder gehen wir von einer diagonalisierbaren Matrix $A \in \mathbb{C}^{n \times n}$ aus, es sei also wieder

$$X^{-1} A X = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$$

mit nichtsingulärem $X = (\, x_1 \ \cdots \ x_n \,)$. Das Verfahren der inversen Iteration kann folgendermaßen formuliert werden:

- Gegeben $x^{(0)} \in \mathbb{C}^n$ und eine Zahl $\sigma \in \mathbb{C}$, die kein Eigenwert von A ist.
- Für $k = 0, 1, \ldots$:
  - Berechne $y^{(k+1)} := (A - \sigma I)^{-1} x^{(k)}$.
  - Berechne $x^{(k+1)} := y^{(k+1)} / \|y^{(k+1)}\|_2$.
  - Berechne $\lambda^{(k+1)} := (x^{(k+1)})^T A x^{(k+1)}$.

Man beachte, dass man zur Berechnung von $y^{(k+1)}$ ein lineares Gleichungssystem mit der Koeffizientenmatrix $A - \sigma I$ zu lösen hat. Daher wird man zu Beginn zunächst mit Hilfe des Gaussschen Eliminationsverfahrens mit Spaltenpivotsuche eine LR-Zerlegung von $A - \sigma I$, den Vektor $y^{(k+1)}$ gewinnt man dann durch Vorwärts- und Rückwärtseinsetzen.

Wir werden zwar keine Konvergenzaussage für das Verfahren der inversen Iteration beweisen, wollen aber doch motivieren, weshalb es sich um ein gutes Verfahren

zur Berechnung bestimmter Eigenwerte und zugehöriger Eigenvektoren handelt. Es sei wieder

$$x^{(0)} = \sum_{i=1}^{n} \alpha_i x_i$$

die Darstellung des Startvektors $x^{(0)}$ als Linearkombination der Eigenvektoren. Offenbar ist $x^{(k)} \in \operatorname{span}\{(A-\sigma I)^{-k} x^{(0)}\}$, ferner

$$(A-\sigma I)^{-k} x^{(0)} = \sum_{i=1}^{n} \frac{\alpha_i}{(\lambda_i - \sigma)^k} x_i.$$

Wir nehmen nun an, dass der vorgebene Paraneter σ sehr viel näher beim Eigenwert λ_j als bei den übrigen Eigenwerten λ_i, $i \neq j$, liegt, dass also

$$|\lambda_j - \sigma| \ll |\lambda_i - \sigma|, \qquad i \neq j$$

bzw.

$$\frac{1}{|\lambda_i - \sigma|} \ll \frac{1}{|\lambda_j - \sigma|}, \qquad i \neq j.$$

Aus der Darstellung von $(A-\sigma I)^{-k} x^{(0)}$ liest man ab, dass der j-te Summand die übrigen Terme stark dominiert, so dass $x^{(k)}$ eine gute Näherung für einen zu λ_j gehörenden Eigenvektor sein wird. Bei J. W. Demmel (1997, S. 156) kann man nachlesen:

- The advantage of inverse iteration over the power method is the ability to converge to any desired eigenvalue (the one nearest the shift σ). By choosing σ very close to a desired eigenvalue, we can converge very quickly and thus not be as limited by the proximity of nerby eigenvalues as is the original power method. The method is particularly effective when we have a good approximation to an eigenvalue and want only its corresponding eigenvector. ...

Das nächste Verfahren, das Verfahren der orthogonalen Iteration, hat den Vorteil, dass Konvergenz gegen einen p-dimensionalen, unter A invarianten Teilraum mit $p > 1$ unter geeigneten Voraussetzungen ermöglicht wird. Eine mögliche Version ist die folgende:

- Gegeben sei eine Matrix $X^{(0)} \in \mathbb{C}^{n\times p}$ mit $1 \leq p < n$, wobei wir annehmen, dass die Spalten von $X^{(0)}$ orthonormiert sind, also $(X^{(0)})^H X^{(0)} = I$ gilt.
- Für $k = 0, 1, \ldots$:
  - Berechne $Y^{(k+1)} := AX^{(k)} \in \mathbb{C}^{n\times p}$.
  - Berechne eine (reduzierte) QR-Zerlegung von $Y^{(k+1)}$, also die Zerlegung $Y^{(k+1)} = X^{(k+1)} R^{(k+1)}$ mit einer Matrix $X^{(k+1)} \in \mathbb{C}^{n\times p}$, deren Spalten orthonormiert sind, und einer oberen Dreiecksmatrix $R^{(k+1)} \in \mathbb{C}^{p\times p}$.

Ist $p = 1$, so erhält man offenbar noch einmal das Verfahren der Vektoriteration. Wieder nehmen wir an, dass A diagonalisierbar ist, dass also

$$X^{-1} A X = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$$

mit einer nichtsingulären Matrix $X = (\ x_1\ \ \cdots\ \ x_n\ )$ gilt. Weiter seien die Eigenwerte von A so numeriert, dass

$$|\lambda_1| \geq \cdots \geq |\lambda_n|,$$

wobei wir diesmal voraussetzen, dass $|\lambda_p| > |\lambda_{p+1}|$ und zusätzlich $\lambda_n \neq 0$ bzw. A nichtsingulär ist. Wir wollen uns überlegen, dass unter vernünftigen Voraussetzungen die Folge $\{X^{(k)}\} \subset \mathbb{C}^{n\times p}$ gegen $\mathcal{X}_p := \operatorname{span}\{x_1, \ldots, x_p\}$ konvergiert, den unter A invarianten Teilraum, der von den ersten p Eigenvektoren aufgespannt wird. Wir müssen näher erklären, was wir darunter verstehen. Denn was soll es heißen, dass eine Folge von Matrizen aus $\mathbb{C}^{n\times p}$ gegen einen p-dimensionalen linearen Teilraum des $\mathbb{C}^n$ konvergiert? Hierzu definieren wir, was wir unter dem *Abstand* zwischen zwei p-dimensionalen linearen Teilräumen des $\mathbb{C}^n$ verstehen.

Definition 1.6 Seien

$$\mathcal{U} := \operatorname{span}\{u_1, \ldots, u_p\}, \qquad \mathcal{V} := \operatorname{span}\{v_1, \ldots, v_p\}$$

zwei p-dimensionale lineare Teiläme des $\mathbb{C}^n$. Man definiere die Matrizen

$$U := (\ u_1\ \ \cdots\ \ u_p\ ) \in \mathbb{C}^{n\times p}, \qquad V := (\ v_1\ \ \cdots\ \ v_p\ ) \in \mathbb{C}^{n\times p}.$$

Dann ist der *Abstand* zwischen $\mathcal{U}$ und $\mathcal{V}$ durch

$$d(\mathcal{U}, \mathcal{V}) := \|U(U^HU)^{-1}U^H - V(V^HV)^{-1}V^H\|_2$$

definiert.

A priori ist nicht klar, ob der definierte Abstand zwischen $\mathcal{U}$ und $\mathcal{V}$ unabhängig von der Wahl einer Basis von $\mathcal{U}$ und $\mathcal{V}$ ist. Dies ist aber einfach einzusehen. Denn ist auch

$$\mathcal{U} = \operatorname{span}\{\hat{u}_1, \ldots, \hat{u}_p\},$$

so existiert eine nichtsinguläre Matrix $S \in \mathbb{C}^{p\times p}$ mit $U = \hat{U}S$. Daher ist

$$U(U^HU)^{-1}U^H = \hat{U}S(S^H\hat{U}^H\hat{U}S)^{-1}S^H\hat{U}^H = \hat{U}(\hat{U}^H\hat{U})^{-1}\hat{U}.$$

Daher ist die Matrix

$$P_{\mathcal{U}} := U(U^HU)^{-1}U^H$$

von der Wahl einer Basis des linearen Teilraumes $\mathcal{U}$ unabhängig. Das ist auch kein Wunder, denn für ein beliebiges $x \in \mathbb{C}^n$ ist $P_{\mathcal{U}}x$ die eindeutige orthogonale Projektion von x auf $\mathcal{U}$. Nachdem wir uns gerade von der Wohldefiniertheit des Abstandes zwischen zwei p-dimensionalen linearen Teilräumes des $\mathbb{C}^n$ überzeugt haben, wollen wir im folgenden Lemma das Wort "Abstand" in der vorigen Definition rechtfertigen.

Lemma 1.7 *Durch $d(\cdot,\cdot)$ ist auf der Menge der p-dimensionalen linearen Teilräume des $\mathbb{C}^n$ eine Metrik (bzw. ein Abstand) definiert.*

Beweis: Natürlich ist der Abstand nichtnegativ. Zum Nachweis der Definitheit nehmen wir an, es sei $d(\mathcal{U},\mathcal{V}) = 0$. Mit den gerade eben definierten Projektionsmatrizen ist $P_{\mathcal{U}} = P_{\mathcal{V}}$. Ist $x = Uy \in \mathcal{U}$, so ist $x = P_{\mathcal{U}}x = P_{\mathcal{V}}x \in \mathcal{V}$, so dass $\mathcal{U} \subset \mathcal{V}$. Aus Symmetriegründen ist auch $\mathcal{V} \subset \mathcal{U}$, insgesamt also $\mathcal{U} = \mathcal{V}$. Die Symmetrie und die Dreiecksungleichung sind offensichtlich erfüllt und das Lemma damit bewiesen. □

Nun ordnen wir der Matrix $X^{(k)} = \begin{pmatrix} x_1^{(k)} & \cdots & x_p^{(k)} \end{pmatrix}$ den linearen Teilraum

$$\mathcal{X}^{(k)} := \operatorname{span}\{x_1^{(k)}, \ldots, x_p^{(k)}\}$$

zu. Im folgenden Satz werden hinreichende Bedingungen dafür angegeben, dass die Folge $\{\mathcal{X}^{(k)}\}$ gegen $\mathcal{X}_p := \operatorname{span}\{x_1, \ldots, x_p\}$ konvergiert.

Satz 1.8 *Sei $A \in \mathbb{C}^{n\times n}$ diagonalähnlich mit Eigenwerten $\lambda_1, \ldots, \lambda_n$ und zugehörigen Eigenvektoren $x_1, \ldots, x_n$. Es sei $|\lambda_1| \geq \cdots \geq |\lambda_n| > 0$, ferner $|\lambda_p| > |\lambda_{p+1}|$ mit $1 \leq p < n$. Das Verfahren der orthogonalen Iteration starte mit einer Matrix $X^{(0)} \in \mathbb{C}^{n\times p}$, deren Spalten orthonormiert sind und deren lineare Hülle, also $\mathcal{X}^{(0)}$, einen leeren Durchnitt mit* $\operatorname{span}\{x_{p+1}, \ldots, x_n\}$ *besitzt*[2]*. Dann gibt es eine Konstante $C > 0$ derart, dass*

$$d(\mathcal{X}^{(k)}, \mathcal{X}_p) \leq C \left|\frac{\lambda_{p+1}}{\lambda_p}\right|^k$$

für alle k, d. h. die Folge $\{\mathcal{X}^{(k)}\}$ konvergiert mit der Konvergenzrate $|\lambda_{p+1}/\lambda_p|$ gegen den Teilraum $\mathcal{X}_p$, der von den ersten p Eigenvektoren aufgespannt wird.

Beweis: Wir machen uns den Beweis relativ einfach, indem wir zunächst $\mathcal{X}^{(k)} = A^k(\mathcal{X}^{(0)})$ nachweisen und anschließend auf ein Ergebnis bei J. Werner (1992b, Satz 2.7) verweisen[3]. Die Beziehung $\mathcal{X}^{(k)} = A^k(\mathcal{X}^{(0)})$ zeigen wir durch vollständige Induktion nach k, wobei der Induktionsanfang bei $k = 0$ trivialerweise richtig ist. Wir nehmen an, es sei $\mathcal{X}^{(k)} = A^k(\mathcal{X}^{(0)})$. Im $(k+1)$-ten Schritt wird zunächst $Y^{(k+1)} := AX^{(k)}$ berechnet. Man beachte, dass die p Spalten von $Y^{(k+1)}$ linear unabhängig sind. In der (reduzierten) QR-Zerlegung $Y^{(k+1)} = X^{(k+1)}R^{(k+1)}$ ist daher die obere Dreiecksmatrix $R^{(k+1)} \in \mathbb{C}^{p\times p}$ nichtsingulär und folglich $X^{(k+1)} = AX^{(k)}(R^{(k+1)})^{-1}$. Dann ist aber

$$\begin{aligned}
\mathcal{X}^{(k+1)} &= \{X^{(k+1)}z : z \in \mathbb{C}^p\} \\
&= \{AX^{(k)}(R^{(k+1)})^{-1}z : z \in \mathbb{C}^p\} \\
&= \{AX^{(k)}z : z \in \mathbb{C}^p\} \\
&= \{AA^kX^{(0)}z : z \in \mathbb{C}^p\} \\
&\qquad \text{(nach Induktionsvoraussetzung)} \\
&= A^{k+1}(\mathcal{X}^{(0)}),
\end{aligned}$$

[2]Diese Voraussetzung stimmt für $p = 1$ mit der Voraussetzung bei der Vektoriteration überein, dass der Startvektor $x^{(0)}$ eine nichtverschwindende erste Komponente bezüglich der Basis $\{x_1, \ldots, x_n\}$ besitzt.

[3]Siehe auch D. S. Watkins (1982, Theorem 2.1) "Understanding the QR algorithm". SIAM Review 24, 427–440.

womit die Induktionsbehauptung bewiesen ist. Wie man in der zitierten Literatur nachlesen kann, existiert dann eine Konstante $C > 0$ mit

$$d(\mathcal{X}^{(k)}, \mathcal{X}_p) = d(A^k(\mathcal{X}^{(0)}), \mathcal{X}_p) \leq C \left| \frac{\lambda_{p+1}}{\lambda_p} \right|^k$$

für alle k. □

4.1.3 Das QR-Verfahren

Wir werden uns in diesem Unterabschnitt auf die Berechnung der Eigenwerte (und zugehöriger Eigenvektoren) einer *reellen* Matrix beschränken. Ziel ist es, das QR-Verfahren in seinen Grundzügen zu schildern, wobei wir uns aber möglichst kurz fassen werden.

Im Prinzip sieht das QR-Verfahren folgendermaßen aus:

- In einem Reduktionsschritt führe man die gegebene Matrix $A \in \mathbb{R}^{n\times n}$ in eine orthogonal ähnliche obere Hessenberg-Matrix A_0 über.
- Für $k = 0, 1, \ldots,$:
  - Wähle bzw. bestimme einen "Shift-Parameter" $\sigma_k \in \mathbb{R}$.
  - Bestimme den orthogonalen Anteil Q_k einer QR-Zerlegung $A_k - \sigma_k I = Q_k R_k$ und berechne $A_{k+1} := Q_k^T A_k Q_k$.

Zunächst beschreiben wir den Reduktionsschritt. Der folgende Algorithmus transformiert die gegebene Matrix $A \in \mathbb{R}^{n\times n}$ mittels $n-2$ Ähnlichkeitstransformationen mit Householder-Matrizen in eine orthogonal ähnliche obere Hessenberg-Matrix.

- Input: Gegeben $A \in \mathbb{R}^{n\times n}$.
- Für $k = 1, \ldots, n-2$:
  - Falls $(a_{k+1,k}, \ldots, a_{n,k})^T \neq 0$, dann:
    - Bestimme Householder-Matrix $\overline{P}_k \in \mathbb{R}^{(n-k)\times(n-k)}$ mit
    $$\overline{P}_k (a_{k+1,k}, \ldots, a_{n,k})^T = (*, 0, \ldots, 0)^T.$$
    - Setze $P_k := \operatorname{diag}(I_k, \overline{P}_k)$ und berechne $A := P_k A P_k$.
  - Andernfalls: Setze $P_k := I$.
- Output: Die Ausgangsmatrix A wird in $n-2$ Schritten mit der orthogonal ähnlichen oberen Hessenberg-Matrix $P^T A P$ überschrieben, wobei $P := P_1 \cdots P_{n-2}$.

Auf nähere Einzelheiten einer möglichen Implementation wollen wir nicht eingehen. Man kann sich relativ einfach überlegen, wie man die relevanten Informationen über die benutzten Householder-Matrizen im wesentlichen in dem gerade frei gemachten Platz in der Matrix A speichern kann. Damit ist der Reduktionsschritt beschrieben. Er dient im wesentlichen dazu, die Berechnung der folgenden QR-Zerlegungen billiger zu machen.

Nun kommen wir zu einer Realisierung eines Schrittes des QR-Verfahrens. Dieser sieht, wie wir schon am Anfang angaben, folgendermaßen aus:

- Input: Gegeben sei die (obere) Hessenberg-Matrix $A \in \mathbb{R}^{n\times n}$ und ein Shift-Parameter $\sigma \in \mathbb{R}$.
- Bestimme eine orthogonale Matrix $Q \in \mathbb{R}^{n\times n}$ derart, dass $A - \sigma I = QR$ mit einer oberen Dreiecksmatrix. Anschließend berechne man $A_+ := Q^T AQ$. Hierbei sollte A_+ ebenfalls eine Hessenberg-Matrix sein, damit man das entsprechende Verfahren im nächsten Schritt auch auf A_+ anwenden kann.
- Output: Es wird eine zu A orthogonal ähnliche Hessenberg-Matrix A_+ berechnet.

Von A_+ erhofft man sich, dass das Element in der Position $(n, n-1)$ wesentlich kleiner geworden ist und durch das Element in der Position (n, n) eine gute Näherung für einen Eigenwert gefunden ist. Ist dies nicht der Fall, so könnte das daran liegen, dass in der Schurschen Normalform der letzte Block ein 2×2-Block ist. In diesem Fall macht man einen sogenannten Doppelschritt. Aber hier greifen wir schon vor, zunächst müssen wir auf die Realisierung der obigen Aufgabe eingehen.

Es ist klar, dass man die obere Hessenberg-Matrix $A-\sigma I$ durch Multiplikation mit $n-1$ Givens-Rotationen $G_{12}, G_{23}, \ldots, G_{n-1,n}$ auf obere Dreiecksgestalt R transfomieren kann:

$$\underbrace{G_{n-1,n}\cdots G_{12}}_{=:Q^T}(A-\sigma I) = R.$$

Nun ist man aber eigentlich nicht an der oberen Dreiecksmatrix R, sondern an

$$A_+ = Q^T AQ = RQ + \sigma I = RG_{12}^T \cdots G_{n-1,n}^T + \sigma I$$

interessiert. Außerdem bleibt noch zu zeigen, dass A_+ wieder eine Hessenberg-Matrix ist. Letzteres ist aber leicht einzusehen. Denn multipliziert man die obere Dreiecksmatrix von rechts mit G_{12}^T, so verändern sich nur die ersten beiden Spalten, und zwar sind die neuen eine Linearkombination der alten. Hierdurch erscheint in der ersten Spalte ein Subdiagonalelement. Entsprechendes gilt offenbar für die nächsten $n-1$ Schritte, A_+ ist in der Tat eine Hessenberg-Matrix. Es wäre nun aber unpraktisch, zunächst

$$G_{n-1,n}\cdots G_{12}(A-\sigma I) = R$$

und anschließend

$$A_+ = RG_{12}^T \cdots G_{n-1,n}^T + \sigma I$$

zu berechnen, da man sich dann alle relevanten Informationen über die $n-1$ Givens-Rotationen merken müsste. Besser ist es, nach der Berechnung von $G_{23}G_{12}(A-\sigma I)$ schon mit G_{12}^T von rechts zu multiplizieren, da die weiteren Multiplikationen von links mit $G_{34}, \ldots, G_{n-1,n}$ die ersten beiden Spalten nicht mehr verändern, und danach abwechselnd von links und rechts zu multiplizieren. Mit $\tilde{A} := A - \sigma I$ erhalten wir also:

$$G_{23}G_{12}\tilde{A}G_{12}^T \to G_{34}G_{23}G_{12}\tilde{A}G_{12}^T \to G_{34}G_{23}G_{12}\tilde{A}G_{12}^T G_{23}^T \to \cdots.$$

Nach der Berechnung von $G_{n-1,n}\cdots G_{12}\tilde{A}G_{12}^T\cdots G_{n-2,n-1}^T$ muss am Schluss noch mit $G_{n-1,n}^T$ von rechts multipliziert werden. Auf weitere Feinheiten wollen wir nicht eingehen. Eine ausführlichere Darstellung findet man bei J. Werner (1992b, S. 28).

Jetzt müssen wir noch etwas über geeignete Shift-Strategien und den sogenannten QR-Doppelschritt sagen. Mit Hilfe des folgenden Lemmas werden wir erkennen, dass die Parameterwahl $\sigma := a_{nn}$ i. allg. eine vernünftige Wahl ist.

Lemma 1.9 *Sei $A \in \mathbb{R}^{n\times n}$ eine unreduzierte Hessenberg-Matrix und $\lambda \in \mathbb{R}$ ein Eigenwert von A. Ist $A_+ := Q^T AQ$, wobei $A - \lambda I = QR$ eine QR-Zerlegung von $A - \lambda I$ ist, so ist $(A_+)_{nj} = 0$, $j = 1, \ldots, n-1$, und $(A_+)_{nn} = \lambda$.*

Beweis: Da A eine unreduzierte Hessenberg-Matrix ist, sind die ersten $n-1$ Spalten von $A - \lambda I$ linear unabhängig. In der QR-Zerlegung $QR = A - \lambda I$ ist daher $r_{ii} \neq 0$, $i = 1, \ldots, n-1$. Da λ ein Eigenwert von A ist, ist $A - \lambda I$ und damit auch R singulär, also notwendig $r_{nn} = 0$. Daher ist die letzte Zeile von RQ eine Nullzeile, woraus die Behauptung folgt. □

Das letzte Lemma zeigt, dass das QR-Verfahren mit einem reellen Eigenwert als Shift-Parameter in einem Schritt zu einer Reduktion der Eigenwertaufgabe führt. Da die Kenntnis eines exakten Eigenwertes eine irreale Annahme ist, nehmen wir nun an, der untere 2×2-Block der (unreduzierten) Hessenberg-Matrix A habe $\epsilon := a_{n,n-1}$ als letztes Subdiagonalelement. Ist ϵ "klein", so wird man a_{nn} als Näherung für einen Eigenwert ansehen und einen Schritt des QR-Verfahrens mit $\sigma = a_{nn}$ als Shift-Parameter durchführen. Wir wollen uns überlegen, wie sich das Element ϵ in der Position $(n, n-1)$ hierbei verändert. Durch Multiplikation von links mit den ersten $n-2$ Givens-Rotationen $G_{12}, \ldots, G_{n-2,n-1}$ erhält man aus $A - a_{nn}I$ eine Matrix der Form

$$G_{n-2,n-1} \cdots G_{12}(A - a_{nn}I) = \left(\begin{array}{c|cc} R_{n-2} & * & \\ \hline & a & b \\ 0 & & \\ & \epsilon & 0 \end{array}\right)$$

mit einer oberen Dreiecksmatrix $R_{n-2} \in \mathbb{R}^{(n-2)\times(n-2)}$. Die letzte Zeile von $A - a_{nn}I$ hat sich hierbei noch nicht verändert. Wir nehmen an, es sei $|\epsilon| \leq |a|$. Dann ist die letzte Givens-Rotation durch die Parameter (siehe die Funktion "givrot")

$$c_{n-1} = \frac{|a|}{(a^2+\epsilon^2)^{1/2}}, \qquad s_{n-1} = \frac{\operatorname{sign}(a)\epsilon}{(a^2+\epsilon^2)^{1/2}}$$

gegeben. Daher ist $R = G_{n-1,n} \cdots G_{12}(A - a_{nn}I)$ eine obere Dreiecksmatrix, deren unterer 2×2-Block

$$\begin{pmatrix} r_{n-1,n-1} & r_{n-1,n} \\ 0 & r_{nn} \end{pmatrix} = \frac{1}{(a^2+\epsilon^2)^{1/2}} \begin{pmatrix} a|a| & |a|b \\ 0 & -\operatorname{sign}(a)\epsilon b \end{pmatrix}$$

ist. Bei der anschließenden Multiplikation von rechts mit $G_{12}^T, \ldots, G_{n-2,n-1}^T$ verändert sich die letzte Spalte von R nicht, ferner bleibt die Null in der Position $(n, n-1)$ erhalten. Also ist

$$RG_{12}^T \cdots G_{n-2,n-1}^T = \left(\begin{array}{cc|cc} H_{n-2} & & * & \\ \hline & * & * & r_{n-1,n} \\ 0 & & & \\ & 0 & 0 & r_{nn} \end{array}\right)$$

mit einer oberen Hessenberg-Matrix $H_{n-2} \in \mathbb{R}^{(n-2)\times(n-2)}$. Nach der abschließenden Multiplikation mit $G_{n-1,n}^T$ von rechts erhält man wegen

$$\begin{pmatrix} 0 & r_{nn} \end{pmatrix} \begin{pmatrix} c_{n-1} & -s_{n-1} \\ s_{n-1} & c_{n-1} \end{pmatrix} = -\frac{1}{a^2+\epsilon^2} \begin{pmatrix} \epsilon^2 b & \epsilon a b \end{pmatrix}$$

in

$$A_+ := G_{n-1,n} \cdots G_{12}(A - a_{nn}I)G_{12}^T \cdots G_{n-1,n}^T + a_{nn}I$$

als neues Element in der $(n, n-1)$-Position gerade

$$(A_+)_{n,n-1} = -\frac{\epsilon^2 b}{a^2+\epsilon^2}.$$

Daher kann sagen: Ist in der Hessenberg-Matrix A das Subdiagonalelement ϵ in der $(n-1)$-ten Spalte klein, so ist das entsprechende Element nach einem Schritt des QR-Verfahrens mit dem Shift-Parameter $\sigma = a_{nn}$ von der Größenordnung ϵ^2, also wesentlich kleiner.

Die einfache Shift-Strategie $\sigma_k = a_{nn}^{(k)}$ verliert ihre Berechtigung, wenn das Element $a_{n,n-1}^{(k)}$ nicht klein ist verglichen mit $a_{n-1,n-1}^{(k)}$ und $a_{nn}^{(k)}$. Daher wird i. allg. vorgezogen, die beiden Eigenwerte σ_k und τ_k des unteren 2×2-Blocks

$$\begin{pmatrix} a_{n-1,n-1}^{(k)} & a_{n-1,n}^{(k)} \\ a_{n,n-1}^{(k)} & a_{nn}^{(k)} \end{pmatrix}$$

zu berechnen, und anschließend einen sogenannten QR-Doppelschritt zu machen. Hierbei wird A_{k+2} aus A_k durch den folgenden Prozess berechnet:

$$A_k - \sigma_k I = Q_k R_k, \qquad A_{k+1} := R_k Q_k + \sigma_k I,$$

d. h. man mache einen QR-Schritt mit σ_k als Shift, anschließend einen Schritt des QR-Verfahrens mit τ_k als Shift-Parameter:

$$A_{k+1} - \tau_k I = Q_{k+1} R_{k+1}, \qquad A_{k+2} := R_{k+1} Q_{k+1} + \tau_k I.$$

Nun können σ_k und τ_k (konjugiert) komplex sein, so dass auch die unitären Matrizen Q_k und Q_{k+1} sowie die oberen Dreiecksmatrizen R_k und R_{k+1} i. allg. komplex sein werden, obwohl A_k reell ist. Wir wollen uns überlegen, dass man A_{k+2} aus A_k durch eine reelle Rechnung erhalten kann. Denn

$$\begin{aligned} Q_k Q_{k+1} R_{k+1} R_k &= Q_k (A_{k+1} - \tau_k I) R_k \\ &= Q_k (A_{k+1} - \tau_k I) Q_k^H (A_k - \sigma_k I) \\ &= Q_k (R_k Q_k + \sigma_k I - \tau_k I) Q_k^H (A_k - \sigma_k I) \\ &= (A_k - \tau_k I)(A_k - \sigma_k I) \\ &= A_k^2 - (\sigma_k + \tau_k) A_k + \sigma_k \tau_k I \\ &=: \tilde{A}_k \end{aligned}$$

ist eine *reelle* Matrix, da

$$\sigma_k + \tau_k = a^{(k)}_{n-1,n-1} + a^{(k)}_{nn}, \qquad \sigma_k\tau_k = a^{(k)}_{n-1,n-1}a^{(k)}_{nn} - a^{(k)}_{n-1,n}a^{(k)}_{n,n-1}$$

reell sind. Eine reelle Matrix besitzt eine reelle QR-Zerlegung, so dass die (komplex) unitären Matrizen Q_k und Q_{k+1} so gewählt werden können, dass Q_kQ_{k+1} (reell) orthogonal ist. Wegen

$$A_{k+2} = Q^H_{k+1}A_{k+1}Q_{k+1} = (Q_kQ_{k+1})^H A_k(Q_kQ_{k+1})$$

muss es daher möglich sein, A_{k+2} aus A_k auch bei (konjugiert) komplexen Shift-Parametern σ_k und τ_k durch eine reelle, orthogonale Ähnlichkeitstransformation zu erhalten. Eine naive Vorgehensweise mit einem Aufwand an Multiplikationen, der im wesentlichen proportional zu n^3 ist, würde darin bestehen, die Matrix $\tilde{A}_k$ zu bilden (was alleine schon im wesentlichen einen zu n^3 proportionalen Aufwand erfordert), hiervon eine (reelle) QR-Zerlegung $\tilde{A}_k = \tilde{Q}_k\tilde{R}_k$ zu berechnen und $A_{k+2} := \tilde{Q}^T_kA_k\tilde{Q}_k$ zu setzen. Grundlage für ein wesentlich effizienteres Verfahren zur Durchführung dieses sogenannten QR-Doppelschrittes ist Satz 4.3 in Abschnitt 3.4, den wir der Übersichtlichkeit halber hier mit geringfügig anderer Notation noch einmal angeben.

Satz 1.10 *Seien*

$$Q = (\ q_1\ \ \cdots\ \ q_n\ ) \in \mathbb{R}^{n\times n}, \qquad V = (\ v_1\ \ \cdots\ \ v_n\ ) \in \mathbb{R}^{n\times n}$$

orthogonale Matrizen, die eine Matrix $A \in \mathbb{R}^{n\times n}$ jeweils in eine obere Hessenberg-Matrix $H := Q^TAQ$ bzw. $G := V^TAV$ transformieren, wobei G unreduziert sei. Ist $q_1 = \pm v_1$, stimmen die ersten Spalten von Q und V also (eventuell bis auf einen Faktor -1) überein, so ist auch H eine unreduzierte obere Hessenberg-Matrix und Q und V sowie H und G sind im wesentlichen gleich, d. h. $D := V^TQ$ ist eine orthogonale Diagonalmatrix (besitzt also nur $+1$ oder -1 als Diagonalelemente) und $H = D^TGD$.

Nun kommen wir zu einer effizienten Lösung der folgenden Aufgabenstellung.

- Input: Gegeben sei eine obere Hessenberg-Matrix $A \in \mathbb{R}^{n\times n}$. Seien σ und τ die beiden Eigenwerte des unteren 2×2-Blocks von A, also von

$$\begin{pmatrix} a_{n-1,n-1} & a_{n-1,n} \\ a_{n,n-1} & a_{nn} \end{pmatrix}.$$

  Ferner sei

$$\tilde{A} := A^2 - (\underbrace{a_{n-1,n-1} + a_{nn}}_{=\sigma+\tau})A + (\underbrace{a_{n-1,n-1}a_{nn} - a_{n-1,n}a_{n,n-1}}_{=\sigma\tau})I.$$

  Dies ist nur eine *Bezeichnung* und soll nicht suggerieren, dass $\tilde{A}$ berechnet werden soll.

- Output: Die Matrix A wird mit einer Hessenberg-Matrix überschrieben, die "im wesentlichen" (im Sinne von Satz 1.10, also bis auf eine Ähnlichkeitstransformation mit einer orthogonalen Diagonalmatrix) mit der Hessenberg-Matrix $A_+ := Q^T A Q$ übereinstimmt. Hierbei ist Q der orthogonale Anteil einer QR-Zerlegung von $\tilde{A}$.

Man beachte, dass in der ersten Spalte von $\tilde{A} = (A - \sigma I)(A - \tau I)$ nur die ersten drei Elemente von Null verschieden sind. Diese sind gegeben durch

$$\begin{aligned} \tilde{a}_{11} &= a_{11}^2 - (\sigma + \tau) a_{11} + \sigma\tau + a_{12} a_{21}, \\ \tilde{a}_{21} &= a_{21}[a_{11} + a_{22} - (\sigma + \tau)], \\ \tilde{a}_{31} &= a_{21} a_{32}. \end{aligned}$$

Damit ist die erste Spalte des orthogonalen Anteils Q in einer QR-Zerlegung von $\tilde{A}$ bekannt, denn diese stimmt bis auf eine Normierung mit der ersten Spalte von $\tilde{A}$ überein. Die Lösung der obigen Aufgabenstellung erfolgt durch die folgenden Schritte.

- Gegeben sei die Hessenberg-Matrix $A \in \mathbb{R}^{n \times n}$, ferner seien die ersten drei Komponenten $\tilde{a}_{11}, \tilde{a}_{21}, \tilde{a}_{31}$ der ersten Spalte von $\tilde{A}$ berechnet.
- Bestimme Householder-Matrix $\overline{P}_0 \in \mathbb{R}^{3 \times 3}$ mit $\overline{P}_0 (\tilde{a}_{11}, \tilde{a}_{21}, \tilde{a}_{31})^T = (*, 0, 0)^T$, setze $P_0 := \text{diag}\,(\overline{P}_0, I_{n-3})$ und berechne $A := P_0 A P_0$.

Nach Konstruktion ist P_0 eine Householder-Matrix, welche die erste Spalte von $\tilde{A}$ in ein Vielfaches des ersten Einheitsvektors überführt. Daher stimmt die erste Spalte von P_0 (eventuell bis auf einen Faktor -1) mit der ersten Spalte von Q, dem orthogonalen Anteil einer QR-Zerlegung von $\tilde{A}$, überein. In der orthogonal ähnlichen, transformierten Matrix $A := P_0 A P_0$ wird die Hessenberg-Gestalt nur durch drei Elemente gestört, und zwar denen in den Positionen $(3,1)$, $(4,1)$ und $(4,2)$. Die Idee für die weiteren Schritte besteht darin, die drei störenden Elemente unterhalb der Subdiagonalen sukzessive nach unten zu schieben und schließlich aus der Matrix zu verdrängen.

- Für $k = 1, \ldots, n-2$:
  - Falls $k \leq n-3$, dann:
    - Bestimme Householder-Matrix $\overline{P}_k \in \mathbb{R}^{3 \times 3}$ mit

      $$\overline{P}_k (a_{k+1,k}, a_{k+2,k}, a_{k+3,k})^T = (*, 0, 0)^T.$$

      Setze $P_k := \text{diag}\,(I_k, \overline{P}_k, I_{n-k+3})$, berechne $A := P_k A P_k$.
  - Andernfalls:
    - Bestimme Householder-Matrix $\overline{P}_{n-2} \in \mathbb{R}^{2 \times 2}$ mit

      $$\overline{P}_{n-2} (a_{n-1,n-2}, a_{n,n-2})^T = (*, 0)^T.$$

      Setze $P_{n-2} := \text{diag}\,(I_{n-2}, \overline{P}_{n-2})$, berechne $A := P_{n-2} A P_{n-2}$.

- Output: Mit der orthogonalen Matrix $V := P_0 P_1 \cdots P_{n-2}$ wird die Ausgangsmatrix A mit der orthogonal ähnlichen Hessenberg-Matrix $V^T AV$ überschrieben.

Für $k = 1, \ldots, n-2$ ist e_1 die erste Spalte der Householder-Matrizen P_k. Die erste Spalte von $V := P_0 P_1 \cdots P_{n-2}$ ist daher $P_0 e_1$, die erste Spalte von P_0. Diese wiederum stimmt nach Wahl von P_0 (eventuell bis auf den Faktor -1) mit der ersten Spalte von Q, dem orthogonalen Anteil in einer QR-Zerlegung von $\tilde{A}$, überein. Ist daher $V^T AV$ unreduziert, so sind V und Q sowie $V^T AV$ und $Q^T AQ = A_+$ wegen Satz 1.10 im wesentlichen (und nur hierauf kommt es an) gleich.

Diesen Prozess, der von der Hessenberg-Matrix A zunächst zu $P_0 AP_0$ und danach in $n-2$ weiteren Ähnlichkeitstransformationen mit Householder-Matrizen zu der orthogonal ähnlichen Hessenberg-Matrix $V^T AV$ führt, wollen wir uns für $n = 6$ veranschaulichen. Wieder werden bei einer Transformation fest bleibende Elemente mit $\bullet$, sich verändernde mit $*$ bezeichnet. Wir erhalten

$$\begin{pmatrix} \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ & \bullet & \bullet & \bullet & \bullet & \bullet \\ & & \bullet & \bullet & \bullet & \bullet \\ & & & \bullet & \bullet & \bullet \\ & & & & \bullet & \bullet \end{pmatrix} \xrightarrow{P_0} \begin{pmatrix} * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & \bullet & \bullet & \bullet \\ & & & \bullet & \bullet & \bullet \\ & & & & \bullet & \bullet \end{pmatrix} \xrightarrow{P_1} \begin{pmatrix} \bullet & * & * & * & \bullet & \bullet \\ * & * & * & * & * & * \\ & * & * & * & * & * \\ & * & * & * & * & * \\ & * & * & * & \bullet & \bullet \\ & & & & \bullet & \bullet \end{pmatrix} \xrightarrow{P_2}$$

und weiter

$$\begin{pmatrix} \bullet & \bullet & * & * & * & \bullet \\ \bullet & \bullet & * & * & * & \bullet \\ & * & * & * & * & * \\ & & * & * & * & * \\ & & * & * & * & * \\ & & * & * & * & \bullet \end{pmatrix} \xrightarrow{P_3} \begin{pmatrix} \bullet & \bullet & \bullet & * & * & * \\ \bullet & \bullet & \bullet & * & * & * \\ & \bullet & \bullet & * & * & * \\ & & * & * & * & * \\ & & & * & * & * \\ & & & * & * & * \end{pmatrix} \xrightarrow{P_4} \begin{pmatrix} \bullet & \bullet & \bullet & \bullet & * & * \\ \bullet & \bullet & \bullet & \bullet & * & * \\ & \bullet & \bullet & \bullet & * & * \\ & & \bullet & \bullet & * & * \\ & & & * & * & * \\ & & & & * & * \end{pmatrix}.$$

Hiermit ist ein QR-Doppelschritt beschrieben, von dem man leicht nachweist, dass sein Aufwand im wesentlichen proportional zu n^2 ist.

Während man noch verhältnismaßig einfach Konvergenzaussagen für das einfache (d. h. es werden keine Shifts durchgeführt) QR-Verfahren beweisen kann, ist dies für das "praktikable" QR-Verfahren mit Shifts viel schwieriger, insbesondere in dem bisher betrachteten i. allg. unsymmetrischen Fall.

4.1.4 Aufgaben

1. Die Matrix $A \in {*}{*}{*}^{n \times n}$ habe die einfachen Eigenwerte $\lambda_1, \ldots, \lambda_n$, sei also insbesondere diagonalisierbar. Die zugehörigen Eigenvektoren $x_1, \ldots, x_n$ seien durch $\|x_i\|_2 = 1$, $i = 1, \ldots, n$, normiert. Ferner seien $y_1, \ldots, y_n$ zugehörige *linke Eigenvektoren*, d. h. $y_i \neq 0$ und $y_i^H A = \lambda_i y_i^H$ (oder $A^H y_i = \overline{\lambda_i} y_i$), $i = 1, \ldots, n$, ebenfalls durch $\|y_i\|_2 = 1$, $i = 1, \ldots, n$, normiert. Weiter wird vorausgesetzt, dass $y_i^H x_i \neq 0$, $i = 1, \ldots, n$. Man zeige[4]:

[4] Man findet diese Aufgabe bei J. W. Demmel (1997, Theorem 4.5). Dort wird die Aussage als Satz von Bauer-Fike bezeichnet. Die Voraussetzung, dass $y_i^H x_i \neq 0$, $i = 1, \ldots, n$, ist dort nicht angegeben und es stellt sich die Frage, ob sie bei einfachen Eigenwerten automatisch erfüllt ist.

(a) Es ist $y_i^H x_j = 0$ für $i \neq j$.

(b) Definiert man

$$P := (\; x_1 \quad \cdots \quad x_n \;),$$

so ist

$$P^{-1} = \begin{pmatrix} y_1^H / y_1^H x_1 \\ \vdots \\ y_n^H / y_n^H x_n \end{pmatrix}.$$

(c) Ist $\delta A \in ***^{n\times n}$ und $\lambda \in ***$ ein Eigenwert von $A + \delta A$, so existiert ein $i \in \{1, \ldots, n\}$ mit

$$|\lambda - \lambda_i| \leq n \frac{\|\delta A\|_2}{|y_i^H x_i|}.$$

2. Sei $A \in ***^{n\times n}$ und $Q^H A Q = \Lambda + N$ mit unitärem $Q \in ***^{n\times n}$, der Diagonalmatrix $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ und der strikten oberen Dreiecksmatrix $N \in ***^{n\times n}$. Ist $\delta A \in ***^{n\times n}$, $\lambda \in ***$ ein Eigenwert von $A + \delta A$, so ist[5]

$$\min_{i=1,\ldots,n} |\lambda - \lambda_i| \leq \max(\theta, \theta^{1/n}),$$

wobei

$$\theta := \|\delta A\|_2 \sum_{k=0}^{n-1} \|N\|_2^k.$$

3. Man wende das Verfahren der Vektoriteration an, um Näherungen für den dominanten Eigenwert und zugehörigen Eigenvektor von

$$A := \begin{pmatrix} 1 & 2 & 3 & 5 \\ 2 & 4 & 1 & 6 \\ 1 & 2 & -1 & 3 \\ 2 & 0 & 1 & 3 \end{pmatrix}$$

zu berechnen. Hierbei starte man mit $x^{(0)} := (1, 1, 1, 1)^T$.

4. Mit Hilfe der inversen Iteration verbessere man die Näherung $\lambda_1 \approx 8$ für einen Eigenwert der Matrix

$$A := \begin{pmatrix} 1 & 2 & 3 & 5 \\ 2 & 4 & 1 & 6 \\ 1 & 2 & -1 & 3 \\ 2 & 0 & 1 & 3 \end{pmatrix}.$$

Man starte mit $x^{(0)} := (1, 1, 1, 1)^T$ und berechne auch den zugehörigen Eigenvektor.

5. Man wende das Verfahren der orthogonalen Iteration auf die Matrix

$$A := \begin{pmatrix} -261 & 209 & -49 \\ -530 & 422 & -98 \\ -800 & 631 & -144 \end{pmatrix}$$

[5]Diese Aussage findet man bei G. H. Golub, C. F. van Loan (1983, S. 201).

an, wobei man $p := 2$ und

$$X^{(0)} := \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$$

wähle[6].

6. Sei $A \in \mathbb{R}^{n\times n}$ eine Matrix mit reellen Eigenwerten $\lambda_1 \geq \cdots \geq \lambda_n$. Man zeige:

   (a) Ist $\sigma < \frac{1}{2}(\lambda_1 + \lambda_n)$ und $\lambda_1 > \lambda_2$, so ist $\lambda_1 - \sigma$ ein dominanter Eigenwert von $A - \sigma I$.

   (b) Ist $\sigma > \frac{1}{2}(\lambda_1 + \lambda_n)$ und $\lambda_n < \lambda_{n-1}$, so ist $\lambda_n - \sigma$ ein dominanter Eigenwert von $A - \sigma I$.

   Daher kann man bei der Anwendung der Vektoriteration auf $A - \sigma I$ nur Konvergenz gegen λ_1 oder λ_n (bzw. $\lambda_1 - \sigma$ oder $\lambda_n - \sigma$) erwarten.

7. Man schreibe ein Programm, das eine Matrix $A \in \mathbb{R}^{n\times n}$ mit Hilfe von $n-2$ Ähnlichkeitstransformationen mit Householder-Matrizen $P_k = \operatorname{diag}(I_k, \overline{P}_k)$ in eine obere Hessenberg-Matrix H überführt. Als Output sollte man erhalten:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{1,n-1} & a_{1n} \\ a_{21} & a_{22} & \cdots & \cdots & a_{2,n-1} & a_{2n} \\ u_3^1 & a_{32} & \cdots & \cdots & a_{3,n-1} & a_{3n} \\ u_4^1 & u_4^2 & \ddots & & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ u_n^1 & u_n^2 & \cdots & u_n^{n-2} & a_{n,n-1} & a_{nn} \end{pmatrix}, \quad d = \begin{pmatrix} u_2^1 \\ u_3^2 \\ \vdots \\ u_{n-1}^{n-2} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{n-2} \end{pmatrix},$$

   wobei

$$H = \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{1,n-1} & a_{1n} \\ a_{21} & a_{22} & \cdots & \cdots & a_{2,n-1} & a_{2n} \\ & a_{32} & \cdots & \cdots & a_{3,n-1} & a_{3n} \\ & & \ddots & & \vdots & \vdots \\ & & & \ddots & \vdots & \vdots \\ & & & & a_{n,n-1} & a_{nn} \end{pmatrix}$$

   die gesuchte, orthogonal ähnliche obere Hessenberg-Matrix ist und

$$\overline{P}_k = I_{n-k} - \beta_k \begin{pmatrix} u_{k+1}^k \\ \vdots \\ u_n^k \end{pmatrix} \begin{pmatrix} u_{k+1}^k \\ \vdots \\ u_n^k \end{pmatrix}^T.$$

   Man teste das Programm an der Matrix

$$A := \begin{pmatrix} 1 & 2 & 3 & 5 \\ 2 & 4 & 1 & 6 \\ 1 & 2 & -1 & 3 \\ 2 & 0 & 1 & 3 \end{pmatrix}.$$

[6]Diese Matrix ist G. H. Golub, C. F. van Loan (1983, S. 210) entnommen worden.

8. Zu einer Matrix $A \in \mathbb{R}^{n\times n}$ sei eine orthogonale Matrix $Q \in \mathbb{R}^{n\times n}$ derart bestimmt worden, dass $Q^T AQ = R$ eine obere Dreiecksmatrix ist. In der Diagonalen von R stehen also die (notwendigerweise reellen) Eigenwerte von A. Wir nehmen an, dass diese sogar paarweise verschieden voneinander sind. Wie kann aus Q und R das vollständige System von Eigenvektoren zu A berechnet werden?

9. Seien $A \in \mathbb{R}^{n\times n}$ und $z \in \mathbb{R}^n \setminus \{0\}$ gegeben. Man zeige, dass es eine orthogonale Matrix $Q \in \mathbb{R}^{n\times n}$ gibt derart, dass $Q^T AQ$ eine obere Hessenberg-Matrix und $Q^T z$ ein Vielfaches des ersten Einheitsvektors e_1 ist.

10. Eine Matrix $A \in \mathbb{C}^{n\times n}$ heißt *normal*, wenn $A^H A = AA^H$.

    (a) Man zeige mit Hilfe des (komplexen) Schurschen Zerlegungssatzes, dass eine Matrix $A \in \mathbb{C}^{n\times n}$ genau dann unitär ähnlich einer Diagonalmatrix ist, wenn A normal ist.

    (b) Seien $\lambda_1, \ldots, \lambda_n$ die Eigenwerte einer Matrix $A \in \mathbb{C}^{n\times n}$. Man zeige, dass A genau dann normal ist, wenn $\sum_{i=1}^n |\lambda_i|^2 = \|A\|_F^2$.

4.2 Das symmetrische Eigenwertproblem

In diesem Abschnitt ist eine symmetrische (reelle) Matrix $A \in \mathbb{R}^{n\times n}$ gegeben, gesucht sind die (notwendigerweise reellen) Eigenwerte und Eigenvektoren. Neben der bisher genannten Literatur geben wir vor allem B. N. Parlett (1980) an[7].

4.2.1 Theoretische Grundlagen

Die zum Grundwissen der linearen Algebra gehörenden Ergebnisse (dass z. B. eine symmetrische Matrix orthogonal ähnlich einer Diagonalmatrix und damit insbesondere diagonalisierbar ist) führen wir nicht extra auf. Stattdessen geben wir die wichtigsten Variationsprinzipien und Störungssätze an, wobei wir aus Zeitgründen auf die Beweise verzichten werden. Wir werden uns (unnötigerweise) auf reelle Matrizen beschränken und beachten, dass die folgenden Sätze auch im Komplexen gelten, wenn man die Eigenschaft "symmetrisch" durch "hermitesch" ersetzt.

Wir beginnen mit dem Rayleighschen Maximum-Prinzip.

Satz 2.1 *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch. Seien $\lambda_1 \geq \cdots \geq \lambda_n$ die Eigenwerte von A und $\{u_1, \ldots, u_n\} \subset \mathbb{R}^n$ ein zugehöriges Orthonormalsystem von Eigenvektoren, also*

$$Au_i = \lambda_i u_i, \qquad u_i^T u_j = \delta_{ij}, \qquad 1 \leq i \leq j \leq n.$$

Für $j = 1, \ldots, n$ definiere man den $(n+1-j)$-dimensionalen linearen Teilraum

$$M_j := \{x \in \mathbb{R}^n : u_i^T x = 0, \ \ (i = 1, \ldots, j-1)\}.$$

Dann ist

$$\lambda_j = \max_{0\neq x\in M_j} \frac{x^T Ax}{x^T x}, \qquad j = 1, \ldots, n.$$

[7] B. N. Parlett (1980) *The symmetric eigenvalue problem.* Prentice-Hall, Englewood Cliffs.

Bemerkung: Insbesondere ist

$$\lambda_1 = \max_{x \neq 0} \frac{x^T A x}{x^T x},$$

was wir schon lange wissen (siehe Aufgabe 2 in Abschnitt 2.1). Ersetzt man A durch $-A$ so erhält man

$$-\lambda_n = \max_{x \neq 0} \Big[-\frac{x^T A x}{x^T x} \Big] \qquad \text{bzw.} \qquad \lambda_n = \min_{x \neq 0} \frac{x^T A x}{x^T x}.$$

Durch $\rho(x) := x^T Ax / x^T x$ ist der zur symmetrischen Matrix $A \in \mathbb{R}^{n \times n}$ gehörende *Rayleigh-Quotient* in $x \in \mathbb{R}^n \setminus \{0\}$ definiert. Die bei festem $x \in \mathbb{R}^{n \times n} \setminus \{0\}$ durch $f(\lambda) := \frac{1}{2}\|Ax - \lambda x\|_2^2$ gegebene Abbildung $f\colon \mathbb{R} \longrightarrow \mathbb{R}$ nimmt in $\lambda = \rho(x)$ ihr Minimum an. Ist daher x näherungsweise ein Eigenvektor von A, so kann man erwarten, dass $\rho(x)$ eine gute Näherung für einen zugehörigen Eigenwert ist. Die Matrix

$$A = \begin{pmatrix} 1 & 3 & 5 \\ 3 & 5 & 7 \\ 5 & 7 & 9 \end{pmatrix}$$

hat nach MATLAB den größten Eigenwert $\lambda_1 = 16.4582364335845$ mit zugehörigem Eigenvektor

$$u_1 = \begin{pmatrix} 0.351625143112725 \\ 0.553356177984084 \\ 0.755087212855444 \end{pmatrix}.$$

Dann erhalten wir z. B.

$$x = \begin{pmatrix} 0.3 \\ 0.6 \\ 0.7 \end{pmatrix}, \qquad \rho(x) = 16.3404255319149$$

und

$$x = \begin{pmatrix} 0.35 \\ 0.55 \\ 0.75 \end{pmatrix}, \qquad \rho(x) = 16.4582278481013.$$

Diese Eigenschaft des Rayleigh-Quotienten wird später beim Rayleigh-Ritz-Verfahren ausgenutzt. □

Es folgt das Courantsche Minimum-Maximum-Prinzip.

Satz 2.2 *Sei $A \in \mathbb{R}^{n \times n}$ symmetrisch mit Eigenwerten $\lambda_1 \geq \cdots \geq \lambda_n$. Für $j = 1, \ldots, n$ sei*

$$\mathcal{N}_j := \{N_j \subset \mathbb{R}^n : \ N_j \text{ ist linearer Teilraum mit } \dim(N_j) = n + 1 - j\}.$$

Dann ist

$$\lambda_j = \min_{N_j \in \mathcal{N}_j} \max_{0 \neq x \in N_j} \frac{x^T A x}{x^T x}, \qquad j = 1, \ldots, n.$$

Eine wichtige Folgerung aus dem Courantschen Minimum-Maximum-Prinzip ist ein Ergebnis, das auf H. Weyl zurückgeht.

Satz 2.3 *Seien $A, B \in \mathbb{R}^{n\times n}$ symmetrisch. Dann genügen die Eigenwerte $\lambda_j(A), \lambda_j(B)$, $j = 1, \ldots, n$, von A bzw. B in der Anordnung*

$$\lambda_1(A) \geq \cdots \geq \lambda_n(A), \qquad \lambda_1(B) \geq \cdots \geq \lambda_n(B)$$

für jede natürliche Matrixnorm (d. h. eine Matrixnorm, die durch eine Vektornorm induziert ist) der Abschätzung

$$|\lambda_j(A) - \lambda_j(B)| \leq \|A - B\|, \qquad j = 1, \ldots, n.$$

Setzt man im vorigen Satz $B := \operatorname{diag}(A)$ und $\|\cdot\| := \|\cdot\|_\infty$ bzw. $\|\cdot\| := \|\cdot\|_F$ (die Frobenius-Norm $\|\cdot\|_F$ ist zwar keine natürliche Matrixnorm, aber es ist $\|A\|_2 \leq \|A\|_F$), so erhält man:

Korollar 2.4 *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und $\{a'_{11}, \ldots, a'_{nn}\}$ eine Permutation der Diagonalelemente $\{a_{11}, \ldots, a_{nn}\}$ mit $a'_{11} \geq \cdots \geq a'_{nn}$. Für die Eigenwerte $\lambda_1 \geq \cdots \geq \lambda_n$ von A gelten dann die Abschätzungen*

$$|\lambda_j - a'_{jj}| \leq \max_{i=1,\ldots,n} \sum_{\substack{k=1\\k\neq i}}^{n} |a_{ik}|, \qquad j = 1, \ldots, n,$$

und

$$|\lambda_j - a'_{jj}| \leq \Big(\sum_{\substack{i,k=1\\i\neq k}}^{n} a_{ik}^2\Big)^{1/2}, \qquad j = 1, \ldots, n.$$

Beweise zu allen diesen Sätzen findet man z. B. bei J. Werner (1992b, S. 9 ff.). Ein weiteres interessantes Ergebnis ist ein auf Hoffman-Wielandt (1953) zurückgehendes Resultat[8].

Satz 2.5 *Seien $A, B \in \mathbb{R}^{n\times n}$ symmetrisch. Dann genügen die Eigenwerte $\lambda_j(A), \lambda_j(B)$, $j = 1, \ldots, n$, von A bzw. B in der Anordnung*

$$\lambda_1(A) \geq \cdots \geq \lambda_n(A), \qquad \lambda_1(B) \geq \cdots \geq \lambda_n(B)$$

der Abschätzung

$$\Big(\sum_{j=1}^{n} (\lambda_j(A) - \lambda_j(B))^2\Big)^{1/2} \leq \|A - B\|_F.$$

[8] Einen elementaren, aber nicht einfachen Beweis findet man bei J. H. Wilkinson (1965, S. 104 ff.) *The Algebraic Eigenvalue Problem*. Clarendon Press Oxford.

4.2.2 Das QR-Verfahren

Wendet man das QR-Verfahren aus Unterabschnitt 4.1.3 auf eine symmetrische Matrix $A \in \mathbb{R}^{n\times n}$ an, so erhält man einige Vereinfachungen gegenüber dem allgemeinen Fall. Im Reduktionsschritt erhält man durch $n-2$ Ähnlichkeitstransformationen mit Householder-Matrizen eine symmetrische obere Hessenberg-Matrix, also eine (symmetrische) Tridiagonalmatrix. Nähere Ausführungen zu einer möglichen Implementation, welche die Symmetrie ausnutzt, findet man z. B. bei J. Werner (1992b, S. 23). Einige Ausführungen sollen nun noch zur Lösung der folgenden Aufgabenstellung gemacht werden:

- Input: Die unreduzierte, symmetrische Tridiagonalmatrix $A \in \mathbb{R}^{n\times n}$ mit den Haupt- bzw. Nebendiagonalelementen $\delta_1, \ldots, \delta_n$ bzw. $\gamma_1, \ldots, \gamma_{n-1}$ sei gegeben. Ferner ist ein Shift-Parameter $\sigma \in \mathbb{R}$ vorgegeben.

- Output: Die Matrix A wird mit einer orthogonal ähnlichen (symmetrischen) Tridiagonalmatrix überschrieben, die im wesentlichen mit $A_+ = Q^T AQ$ übereinstimmt, wobei Q der orthogonale Anteil einer QR-Zerlegung von $A - \sigma I$ ist.

Im Prinzip geht man ganz ähnlich wie im QR-Doppelschritt vor. Der Schlüssel für eine effiziente Berechnung ist wieder durch Satz 4.3 in Abschnitt 3.4 bzw. Satz 1.10 gegeben.

- Bestimme eine Givens-Rotation $G_{12} = G_{12}(c_1, s_1)$ mit

$$\begin{pmatrix} c_1 & s_1 \\ -s_1 & c_1 \end{pmatrix} \begin{pmatrix} \delta_1 - \sigma \\ \gamma_1 \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}$$

  und berechne $A := G_{12} A G_{12}^T$.

Die Givens-Rotation G_{12} transformiert die erste Spalte von $A - \sigma I$ in ein Vielfaches des ersten Einheitsvektors, es ist also $G_{12}(A - \sigma I)e_1 = \alpha e_1$ bzw. $(A - \sigma I)e_1 = \alpha G_{12}^T e_1$. Daher stimmen die ersten Spalten von G_{12}^T und Q, dem orthogonalen Anteil in einer QR-Zerlegung von $A - \sigma I$, (eventuell bis auf den Faktor -1) überein. Ferner wird die Tridiagonalgestalt der transformierten Matrix $G_{12} A G_{12}^T$ nur in der Position $(3,1)$ (und symmetrisch dazu in $(1,3)$) gestört. Die weitere Idee besteht darin, das die Tridiagonalgestalt störend Element sukzessive durch Ähnlichkeitstransformationen mit Givens-Rotationen von der Position $(3,1)$ über $(4,2)$ nach $(n, n-2)$ zu vertreiben und in einem letzten Schritt auch noch das Element an der Stelle $(n, n-2)$ zu annullieren. Diese weiteren Schritte sehen also wie folgt aus:

- Für $k = 2, \ldots, n-1$:

  - Bestimme Givens-Rotation $G_{k,k+1} = G_{k,k+1}(c_k, s_k)$ mit

$$\begin{pmatrix} c_k & s_k \\ -s_k & c_k \end{pmatrix} \begin{pmatrix} a_{k,k-1} \\ a_{k+1,k-1} \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}$$

    und berechne $A := G_{k,k+1} A G_{k,k+1}^T$.

- Output: Die Ausgangsmatrix A wird mit der orthogonal ähnlichen (symmetrischen) Tridiagonalmatrix V^TAV überschrieben, wobei $V := G_{12}^T \cdots G_{n-1,n}^T$.

Die erste Spalte von $G_{23}^T, \ldots, G_{n-1,n}^T$ ist offenbar jeweils der erste Einheitsvektor e_1, so dass die erste Spalte von V genau die erste Spalte von G_{12}^T ist. Nach Konstruktion stimmt diese mit der ersten Spalte von Q, dem orthogonalen Anteil einer QR-Zerlegung von $A-\sigma I$, (eventueel bis auf das Vorzeichen) überein. Aus Satz 1.10 folgt, dass V^TAV und $A_+ = Q^TAQ$ im wesentlichen (d. h. bis auf eine Ähnlichkeitstransformation mit einer orthogonalen Diagonalmatrix) gleich sind, wenn V^TAV unreduziert ist.

Für $n = 6$ wollen wir uns den Prozess, der von der symmetrischen Tridiagonalmatrix A zunächst zu $G_{12}AG_{12}^T$ und dann nach weiten Ähnlichkeitstransformationen mit Givens-Rotationen $G_{23}, \ldots, G_{n-1,n}$ zur Tridiagonalmatrix V^TAV führt, verdeutlichen. Wieder bezeichne $\bullet$ ein bei der Transformation festbleibendes und $*$ ein sich veränderndes Element.

$$\begin{pmatrix} \bullet & \bullet & & & & \\ \bullet & \bullet & \bullet & & & \\ & \bullet & \bullet & \bullet & & \\ & & \bullet & \bullet & \bullet & \\ & & & \bullet & \bullet & \bullet \\ & & & & \bullet & \bullet \end{pmatrix} \xrightarrow{G_{12}} \begin{pmatrix} * & * & * & & & \\ * & * & * & & & \\ * & * & \bullet & \bullet & & \\ & & \bullet & \bullet & \bullet & \\ & & & \bullet & \bullet & \bullet \\ & & & & \bullet & \bullet \end{pmatrix} \xrightarrow{G_{23}} \begin{pmatrix} \bullet & * & & & & \\ * & * & * & * & & \\ & * & * & * & & \\ & * & * & \bullet & \bullet & \\ & & & \bullet & \bullet & \bullet \\ & & & & \bullet & \bullet \end{pmatrix} \xrightarrow{G_{34}}$$

$$\begin{pmatrix} \bullet & \bullet & & & & \\ \bullet & \bullet & * & & & \\ & * & * & * & * & \\ & & * & * & * & \\ & & * & * & \bullet & \bullet \\ & & & & \bullet & \bullet \end{pmatrix} \xrightarrow{G_{45}} \begin{pmatrix} \bullet & \bullet & & & & \\ \bullet & \bullet & \bullet & & & \\ & \bullet & \bullet & * & & \\ & & * & * & * & * \\ & & & * & * & * \\ & & & * & * & \bullet \end{pmatrix} \xrightarrow{G_{56}} \begin{pmatrix} \bullet & \bullet & & & & \\ \bullet & \bullet & \bullet & & & \\ & \bullet & \bullet & \bullet & & \\ & & \bullet & \bullet & * & \\ & & & * & * & * \\ & & & & * & * \end{pmatrix}$$

Offenbar ist der Aufwand zur Berechnung von V^TAV aus A im wesentlichen proportional zu n. Ein Programm in Pseudo-Code findet man bei J. Werner (1992b, S. 62 ff.), wir wollen hierauf nicht näher eingehen. Bei der Wahl eines Shift-Parameters σ werden im wesentlichen zwei Strategien benutzt. Gegeben sei jeweils die symmetrische Tridiagonalmatrix $A \in \mathbb{R}^{n\times n}$ mit den Haupt- bzw. Nebendiagonalelementen $\delta_1, \ldots, \delta_n$ bzw. $\gamma_1, \ldots, \gamma_{n-1}$.

(a) $\sigma := \delta_n$.

(b) σ wird als derjenige Eigenwert von

$$\begin{pmatrix} \delta_{n-1} & \gamma_{n-1} \\ \gamma_{n-1} & \delta_n \end{pmatrix}$$

bestimmt, der näher bei δ_n liegt. Diese Wahl wird als *Wilkinson-Shift* bezeichnet. Es wird empfohlen, den Wilkinson-Shift folgendermaßen zu berechnen:

$$d := \frac{\delta_{n-1} - \delta_n}{2}, \qquad \sigma := \delta_n - \frac{\operatorname{sign}(d)\gamma_{n-1}^2}{|d| + (d^2 + \gamma_{n-1}^2)^{1/2}}.$$

Das QR-Verfahren erzeugt eine Folge $\{A^{(k)}\}$ symmetrischer Tridiagonalmatrizen mit den Haupt- bzw. den Nebendiagonalelementen $\delta_1^{(k)}, \ldots, \delta_n^{(k)}$ bzw. $\gamma_1^{(k)}, \ldots, \gamma_{n-1}^{(k)}$. Man spricht von *globaler Konvergenz* des QR-Verfahrens, wenn $\lim_{k\to\infty} \gamma_{n-1}^{(k)} = 0$ und sagt, es sei *quadratisch* bzw. *kubisch* konvergent, wenn darüber hinaus eine Konstante $c > 0$ derart existiert, dass $|\gamma_{n-1}^{(k+1)}| \leq c\,|\gamma_{n-1}^{(k)}|^2$ bzw. $|\gamma_{n-1}^{(k+1)}| \leq c\,|\gamma_{n-1}^{(k)}|^3$ für alle hinreichend großen k. Es liegt nahe, diesen modifizierten Konvergenzbegriff zu benutzen, da man ja an (schneller) Konvergenz der Folge $\{\gamma_{n-1}^{(k)}\}$ interessiert ist, um (möglichst schnell) durch Streichen der letzten Zeile und letzten Spalte zu einem in der Dimension reduzierten Problem zu gelangen. Die grundlegende Arbeit zur Konvergenz des QR-Verfahrens bei symmetrischen Tridiagonalmatrizen stammt von Wilkinson[9]. Für die Shift-Strategie (a) zeigt Wilkinson, dass $|\gamma_{n-1}^{(k+1)}| \leq |\gamma_{n-1}^{(k)}|$ für alle k. Hieraus folgt die Konvergenz der Folge $\{\gamma_{n-1}^{(k)}\}$. Ist der Limes L gleich Null, liegt also globale Konvergenz vor, so ist die Konvergenz sogar kubisch. Ist dagegen $L > 0$ (dann liegt keine Konvergenz des Verfahrens im obigen Sinn vor), so konvergiert wenigstens die Folge $\{\gamma_{n-1}^{(k)}\}$ (eventuell langsam) gegen Null. In diesem Fall kann man daher schließlich die letzten beiden Zeilen und Spalten streichen. Für das QR-Verfahren mit der Shift-Strategie (b) zeigt Wilkinson zunächst die globale und dann die quadratische Konvergenz. Von Ausnahmen abgesehen kann sogar die kubische Konvergenz nachgewiesen werden.

4.2.3 Das Rayleigh-Quotienten-Verfahren

Das Rayleigh-Quotienten-Verfahren ist ein Verfahren zur Bestimmung eines Eigenwertes und eines zugehörigen Eigenvektors einer symmetrischen Matrix $A \in \mathbb{R}^{n\times n}$, welches sich durch außerordentliche Konvergenzeigenschaften auszeichnet. Im folgenden sei generell $A \in \mathbb{R}^{n\times n}$ eine symmetrische Matrix mit den Eigenwerten $\lambda_1, \ldots, \lambda_n$ und einem zugehörigen Orthonormalsystem $z_1, \ldots, z_n$ von Eigenvektoren. Für ein $x \in \mathbb{R}^n \setminus \{0\}$ sei $\rho(x) := x^T Ax/x^T x$ der zugehörige Rayleigh-Quotient. Als Literatur nennen wir vor allem B. N. Parlett (1980, S. 70 ff.).

Zur Motivation des Verfahrens erinnern wir an die beiden folgenden (vage formulierten) Aussagen:

- Ist $x \neq 0$ eine "gute" Näherung für einen Eigenvektor von A, so ist $\rho(x)$ eine "gute" Näherung für einen zugehörigen Eigenwert.

Genauer: Die Funktion $f\colon \mathbb{R} \longrightarrow \mathbb{R}$, definiert durch $f(\lambda) := \frac{1}{2}\,\|Ax - \lambda x\|_2^2$, nimmt ihr eindeutiges Minimum auf $\mathbb{R}$ in $\rho(x)$ an.

- Sei λ eine "gute" Näherung für einen Eigenwert von A, selbst aber kein Eigenwert, und $x \in \mathbb{R}^n$ mit $\|x\|_2 = 1$ eine Näherung für einen zugehörigen Eigenvektor. Dann erhält man durch
$$x^+ := \frac{(A - \lambda I)^{-1} x}{\|(A - \lambda I)^{-1} x\|_2}$$
eine i. allg. verbesserte Näherung für einen (normierte) Eigenvektor.

[9] Siehe J. H. Wilkinson (1968) "Global convergence of tridiagonal QR algorithm with origin shifts." Linear Algebra and its Applications 1, 409-420.

Genauer: Sei $|\lambda_j - \lambda| \ll |\lambda_i - \lambda|$ für $i \neq j$, also λ eine wesentlich bessere Näherung für den Eigenwert λ_j als für die anderen Eigenwerte. Ferner sei $x = \sum_{i=1}^n \alpha_i z_i$. Dann ist

$$\begin{aligned} x^+ &:= \frac{(A-\lambda I)^{-1}x}{\|(A-\lambda I)^{-1}x\|_2} \\ &= \frac{\sum_i \alpha_i(\lambda_i-\lambda)^{-1}z_i}{[\sum_i \alpha_i^2(\lambda_i-\lambda)^{-2}]^{1/2}} \\ &= \frac{\operatorname{sign}(\lambda_j-\lambda)\alpha_j z_j + \sum_{i\neq j}\alpha_i[(\lambda_i-\lambda)/|\lambda_j-\lambda|]^{-1}z_i}{[\alpha_j^2+\sum_{i\neq j}\alpha_i^2[(\lambda_i-\lambda)/(\lambda_j-\lambda)]^{-2}]^{1/2}}, \end{aligned}$$

was wegen wegen $|\lambda_j - \lambda| \ll |\lambda_i - \lambda|$ für $i \neq j$ eine i. allg. bessere Näherung für z_j als x ist.

Nun ist das folgende Verfahren, das sogenannte *Rayleigh-Quotienten-Verfahren*, sehr naheliegend. In ihm wird zu einer aktuellen Näherung für einen Eigenvektor zunächst der zugehörige Rayleigh-Quotient berechnet, anschließend mit diesen beiden Daten ein Schritt der inversen Iteration durchgeführt.

- Gegeben sei $x_0 \in \mathbb{R}^n$ mit $\|x_0\|_2 = 1$.
- Für $k = 0, 1, \ldots$:
  - Berechne $\rho_k := \rho(x_k)$.
  - Falls $A - \rho_k I$ singulär, dann:
    - Bestimme x_{k+1} mit $(A - \rho_k I)x_{k+1} = 0$, $\|x_{k+1}\|_2 = 1$, STOP.
  - Andernfalls:
    - Berechne $y_{k+1} := (A - \rho_k I)^{-1}x_k$, $x_{k+1} := y_{k+1}/\|y_{k+1}\|_2$.

Man beachte, dass man im Gegensatz zur inversen Iteration beim Rayleigh-Quotienten-Verfahren in jedem Iterationsschritt ein lineares Gleichungssystem mit einer *neuen* Koeffizientenmatrix zu lösen hat.

Im folgenden "Konvergenzsatz" wird die Monotonie des Defektes $\|(A - \rho_k I)x_k\|_2$ nachgewiesen.

Satz 2.6 *Die Folge $\{(x_k, \rho_k)\}$ sei durch das auf die symmetrische Matrix $A \in \mathbb{R}^{n\times n}$ angewandte Rayleigh-Quotienten-Verfahren erzeugt. Dann ist*

$$\|(A - \rho_{k+1}I)x_{k+1}\|_2 \leq \|(A - \rho_k I)x_k\|_2, \qquad k = 0, 1, \ldots.$$

Hier gilt Gleichheit genau dann, wenn $\rho_k = \rho_{k+1}$ und x_k ein Eigenvektor von $(A-\rho_k I)^2$ ist.

Beweis: Der Beweis benutzt die zur Motivation angesprochene Tatsache, dass $\|(A - \lambda I)x\|_2$ als Funktion in λ bei gegebenem $x \neq 0$ in $\rho(x)$ minimal ist. Daher ist

$$\begin{aligned} \|(A - \rho_{k+1}I)x_{k+1}\|_2 &\leq \|(A - \rho_k I)x_{k+1}\|_2 \\ &= |x_k^T(A - \rho_k I)x_{k+1}| \end{aligned}$$

(da $x_k = \alpha(A - \rho_k I)x_{k+1}$ und $\|x_k\|_2 = 1$ tritt Gleichheit in der Cauchy-Schwarzschen Ungleichung ein)

$$= |x_{k+1}^T(A - \rho_k I)x_k|$$

$$\leq \|(A - \rho_k I)x_k\|_2$$

(wegen Cauchy-Schwarz und $\|x_{k+1}\|_2 = 1$).

Tritt hier Gleichheit ein, so auch in den beiden auftretenden Ungleichungen. Ist die erste Ungleichung sogar eine Gleichheit, so ist $\rho_{k+1} = \rho_k$, da die Funktion $\|(A-\lambda I)x\|_2$ in $\rho(x)$ ihr *eindeutiges* Minimum annimmt. Gleichheit in der zweiten Ungleichung impliziert, dass $(A-\rho_k I)x_k$ ein Vielfaches von x_{k+1} ist und dies wiederum, dass x_k ein Eigenvektor von $(A - \rho_k I)^2$ ist. Denn ist

$$(A - \rho_k I)x_k = \alpha_k x_{k+1}, \qquad x_{k+1} = \frac{(A - \rho_k I)^{-1}x_k}{\|(A - \rho_k I)^{-1}x_k\|_2},$$

so folgt

$$(A - \rho_k I)^2 x_k = \frac{\alpha_k}{\|(A - \rho_k I)^{-1}x_k\|_2} x_k.$$

Damit ist der Satz bewiesen. □

Damit haben wir die schöne Aussage erhalten, dass die Folge $\{\|(A-\rho_k I)x_k\|_2\}$ monoton nicht wachsend ist. Insbesondere existiert $\tau := \lim_{k\to\infty} \|(A - \rho_k)x_k\|_2$.

Nun kommen wir zum lokalen Konvergenzsatz für das Rayleigh-Quotientenverfahren.

Satz 2.7 *Das Rayleigh-Quotienten-Verfahren, angewandt auf die symmetrische Matrix $A \in \mathbb{R}^{n\times n}$, erzeuge die Folge $\{(x_k, \rho_k)\}$, wobei $\{x_k\}$ gegen einen durch $\|z\|_2 = 1$ normierten Eigenvektor z von A mit Eigenwert λ konvergiere. Dann konvergiert die Folge $\{x_k\}$ kubisch gegen z. Ferner existiert eine Konstante $c > 0$ mit $|\rho_k - \lambda| \leq c\,\|x_k - z\|^2$ für alle k.*

Beweis: Die Idee besteht darin, den Winkel $\phi_k := \angle(x_k, z)$ einzuführen und nachzuweisen, dass die Folge $\{\sin\phi_k\}$ kubisch gegen Null konvergiert. Da

$$\sin\phi_k = \sqrt{1 - (x_k^T z)^2} = \frac{1}{\sqrt{2}}\,\|x_k - z\|_2,$$

folgt hieraus die kubische Konvergenz von $\{x_k\}$ gegen z. Wir nehmen im folgenden o. B. d. A. an, dass das Verfahren nicht schon nach endlich vielen Schritten abbricht, also insbesondere x_k kein Eigenvektor von A ist. Wegen $\mathbb{R}^n = \operatorname{span}\{z\} \oplus \operatorname{span}\{z\}^\perp$ lässt zich x_k eindeutig in der Form $x_k = \alpha_k z + v_k$ mit $\alpha_k \in \mathbb{R}$ und $v_k^T z = 0$ darstellen. Dann ist $\alpha_k = x_k^T z = \cos\phi_k$, ferner ist $\|v_k\|_2 = \sin\phi_k$. Mit $u_k := v_k/\|v_k\|_2$ erhält man damit die eindeutige Darstellung

$$(*) \qquad x_k = \cos\phi_k\, z + \sin\phi_k\, u_k,$$

wobei $u_k^T z = 0$ und $\|u_k\|_2 = 1$. Eine Multiplikation von $(*)$ mit $(A - \rho_k I)^{-1}$ liefert

$$y_{k+1} = (A - \rho_k I)^{-1}x_k = \frac{\cos\phi_k}{\lambda - \rho_k} z + \sin\phi_k (A - \rho_k I)^{-1}u_k.$$

Wegen $x_{k+1} = y_{k+1}/\|y_{k+1}\|_2$ ist

$$x_{k+1} = \frac{\cos\phi_k}{(\lambda-\rho_k)\,\|y_{k+1}\|_2} z + \frac{\sin\phi_k}{\|y_{k+1}\|_2}(A-\rho_k I)^{-1}u_k.$$

Ein Vergleich mit

$$x_{k+1} = \cos\phi_{k+1}\, z + \sin\phi_{k+1}\, u_{k+1}, \quad u_{k+1}^T z = 0, \quad \|u_{k+1}\|_2 = 1$$

liefert wegen $z^T(A-\rho_k I)^{-1}u_k = 0$ (man beachte: z ist Eigenvektor von $(A-\rho_k I)^{-1}$ und $u_k^T z = 0$), dass

$$\cos\phi_{k+1} = \frac{\cos\phi_k}{(\lambda-\rho_k)\,\|y_{k+1}\|_2}, \qquad u_{k+1} = \frac{(A-\rho_k I)^{-1}u_k}{\|(A-\rho_k I)^{-1}u_k\|_2}$$

und

$$\sin\phi_{k+1} = \frac{\sin\phi_k\,\|(A-\rho_k I)^{-1}u_k\|_2}{\|y_{k+1}\|_2}.$$

Aus

$$\begin{aligned}
\lambda-\rho_k &= \lambda-\rho(x_k)\\
&= \lambda-(\cos\phi_k\, z+\sin\phi_k u_k)^T A(\cos\phi_k\, z+\sin\phi_k\, u_k)\\
&= \lambda-(\cos\phi_k\, z+\sin\phi_k u_k)^T(\lambda\cos\phi_k\, z+\sin\phi_k\, Au_k)\\
&= \lambda-(\lambda\cos^2\phi_k+\sin^2\phi_k\, u_k^T Au_k)
\end{aligned}$$

erhalten wir

$$(**) \qquad \lambda-\rho_k = [\lambda-\rho(u_k)]\sin^2\phi_k.$$

Weiter ist

$$\tan\phi_{k+1} = \frac{\sin\phi_{k+1}}{\cos\phi_{k+1}} = (\lambda-\rho_k)\,\|(A-\rho_k I)^{-1}u_k\|_2\,\tan\phi_k,$$

wegen $(**)$ also

$$(***) \qquad \tan\phi_{k+1} = [\lambda-\rho(u_k)]\,\|(A-\rho_k I)^{-1}u_k\|_2\,\tan\phi_k\,\sin^2\phi_k.$$

Da der Rayleigh-Quotient zwischen dem kleinsten und dem größten Eigenwert liegt, ist die Folge $\{\lambda-\rho(u_k)\}$ trivialerweise beschränkt. Wir wollen uns überlegen, dass auch die Folge $\{\|(A-\rho_k I)^{-1}u_k\|\}$ beschränkt ist. Wichtigstes Hilfsmittel hierfür ist die folgende Aussage.

- Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch mit Eigenwerten $\lambda_1,\ldots,\lambda_n$ und einem zugehörigen Orthonormalsystem von Eigenvektoren $z_1,\ldots,z_n$. Sei λ ein Eigenwert von A und

  $$J := \{j\in\{1,\ldots,n\} : \lambda_j \neq \lambda\}.$$

  Ist dann $u \in \operatorname{span}\{z_j : j\in J\}$, $\|u\|_2 = 1$ und $\rho\in\mathbb{R}$ kein Eigenwert von A, so ist

  $$\|(A-\rho I)^{-1}u\|_2 \le \frac{1}{\gamma} \qquad \text{mit} \quad \gamma := \min_{j\in J}|\lambda_j-\rho|.$$

Denn: Sei $u = \sum_{j\in J} \alpha_j z_j$. Dann ist $\sum_{j\in J} \alpha_j^2 = 1$ wegen $\|u\|_2 = 1$ und folglich

$$\|(A-\rho I)^{-1}u\|_2 = \Big(\sum_{j\in J} \frac{\alpha_j^2}{(\lambda_j-\rho)^2}\Big)^{1/2} \le \frac{1}{\gamma}.$$

Um dieses Resultat auf die obige Situation anwenden zu können, müssen wir mit den eben eingeführten Bezeichnungen uns überlegen, dass $u_k \in \operatorname{span}\{z_j : j \in J\}$ für alle k. Da u_{k+1} ein skalares Vielfaches von $(A-\rho_k I)^{-1}u_k$ ist, und damit $u_{k+1} \in \operatorname{span}\{z_j : j \in J\}$ aus $u_k \in \operatorname{span}\{z_j : j \in J\}$ folgt, genügt es, $u_0 \in \operatorname{span}\{z_j : j \in J\}$ nachzuweisen. Zur Abkürzung setzen wir $M := \operatorname{span}\{z_i : i \notin J\}$, so dass $M^\perp = \operatorname{span}\{z_j : j \in J\}$. Bezeichnet man mit $P_M\colon \mathbb{R}^n \longrightarrow M$ die orthogonale Projektion des $\mathbb{R}^n$ auf M, so ist $P_M(x_k)$ für jedes k ein Vielfaches von $P_M(x_0)$. Denn sei

$$x_k = \underbrace{\sum_{i\notin J} \alpha_i^{(k)} z_i}_{=P_M(x_k)} + \sum_{j\in J} \alpha_j^{(k)} z_j.$$

Dann ist

$$\begin{aligned} x_{k+1} &= \frac{1}{\|(A-\rho_k I)^{-1}x_k\|_2}(A-\rho_k I)^{-1}x_k \\ &= \frac{1}{\|(A-\rho_k I)^{-1}\|_2\,(\lambda-\rho_k)} \sum_{i\notin J} \alpha_i^{(k)} z_i + \frac{1}{\|(A-\rho_k I)^{-1}x_k\|_2} \sum_{j\in J} \frac{\alpha_j^{(k)}}{\lambda_j-\rho_k} z_j \end{aligned}$$

und daher

$$P_M(x_{k+1}) = \frac{1}{\|(A-\rho_k I)^{-1}\|_2\,(\lambda-\rho_k)}\, P_M(x_k),$$

folglich $P_M(x_k)$ ein Vielfaches von $P_M(x_0)$. Nach Voraussetzung konvergiert $\{x_k\}$ gegen $z \in M$, wegen der Stetigkeit der orthogonalen Projektion ist z ein Vielfaches von $P_M(x_0)$, etwa $z = \alpha P_M(x_0)$. Wir wollen $u_0 \in M^\perp$ bzw. $P_M(u_0) = 0$ nachweisen. Wir gehen aus von der Darstellung

$$x_0 = \cos\phi_0\, z + \sin\phi_0\, u_0$$

des Startvektors, wobei wir natürlich $\sin\phi_0 \neq 0$ voraussetzen können, da andernfalls x_0 ein Eigenvektor zum Eigenwert λ wäre und das Rayleigh-Quotienten-Verfahren schon nach einem Schritt stoppen würde. Wegen der Linearität der orthogonalen Projektion ist

$$\begin{aligned} P_M(x_0) &= \cos\phi_0\, P_M(z) + \sin\phi_0\, P_M(u_0) \\ &= \cos\phi_0\, z + \sin\phi_0\, P_M(u_0) \\ &= \underbrace{\alpha\cos\phi_0}_{=1}\, P_M(x_0) + \sin\phi_0\, P_M(u_0) \\ &= P_M(x_0) + \sin\phi_0\, P_M(u_0), \end{aligned}$$

folglich $P_M(u_0) = 0$. Hierbei haben wir ausgenutzt, dass

$$1 = z^T z = \alpha P_M(x_0)^T z = \alpha x_0^T P_M(z) = \alpha x_0^T z = \alpha \cos\phi_0.$$

Damit sind die Voraussetzungen der obigen Hilfsbehauptung schließlich nachgewiesen und wir erhalten, dass

$$\|(A - \rho_k I)^{-1} u_k\|_2 \leq \frac{1}{\min_{j:\lambda_j \neq \lambda} |\lambda - \rho_k|}.$$

Ähnlich wie oben setze man $\gamma := \min_{j:\lambda_j \neq \lambda} |\lambda_j - \lambda|$. Wegen $x_k \to z$ gilt $\rho_k = \rho(x_k) \to \lambda$. Für alle hinreichend großen k ist daher $\min_{j:\lambda_j \neq \lambda} |\lambda_j - \rho_k| \geq \gamma/2$, folglich

$$\|(A - \rho_k I)^{-1} u_k\|_2 \leq \frac{2}{\gamma} \qquad \text{für alle hinreichend großen } k.$$

Wegen $x_k \to z$ ist $\cos\phi_k \to 1$, für alle hinreichend großen k ist daher $\cos\phi_k \geq 1/2$. Aus $(***)$ folgt sofort die kubische Konvergenz von $\{\sin\phi_k\}$ gegen Null, damit auch (wie am Anfang des Beweises begründet) die kubische Konvergenz von $\{x_k\}$ gegen z. Aus $(**)$, also

$$\lambda - \rho_k = [\lambda - \rho(u_k)] \sin^2\phi_k,$$

folgt wegen der Beschränktheit von $\{\lambda - \rho(u_k)\}$ und $\sin^2\phi_k = \frac{1}{2}\|x_k - z\|_2^2$ die Existenz einer Konstanten $c > 0$ mit $|\lambda - \rho_k| \leq c\,\|x_k - x^*\|_2^2$ für alle k. Damit ist der lokale Konvergenzsatz schließlich bewiesen. □

Die globale Konvergenz der durch das Rayleigh-Quotienten-Verfahren erzeugten Folge $\{x_k\}$ gegen einen Eigenvektor z von A (also die Voraussetzung für den lokalen Konvergenzsatz) ist noch wesentlich schwerer zu zeigen (siehe B. N. Parlett (1980, S. 76–79)). Wir wollen uns mit einem verhältnismäßig einfachen Resultat begnügen.

Satz 2.8 *Das auf die symmetrische Matrix $A \in \mathbb{R}^{n\times n}$ angewandte Rayleigh-Quotienten-Verfahren erzeuge die Folge $\{(x_k, \rho_k)\}$. Es sei τ der wegen Satz 2.6 existierende Limes der Folge $\{\|(A - \rho_k I)x_k\|_2\}$. Dann gilt:*

1. *Die Folge $\{\rho_k\}$ konvergiert.*
2. *Ist $\tau = 0$, so konvergiert die Folge $\{x_k\}$ gegen einen (normierten) Eigenvektor von A.*

Beweis: Zunächst zeigen wir die Konvergenz der Folge $\{\rho_k\}$ (ohne weitere Voraussetzungen *nicht* notwendig gegen einen Eigenwert von A). Der Beweis hierfür zerfällt in zwei Teile. Wir zeigen, dass einerseits $\lim_{k\to\infty}(\rho_{k+1} - \rho_k) = 0$, und dass andererseits die beschränkte Folge $\{\rho_k\}$ nur endlich viele Häufungspunkte besitzen kann. Aus einem bekannten Satz der Analysis[10] folgt hieraus die Konvergenz der gesamten Folge $\{\rho_k\}$.

[10] siehe z. B. J. M. Ortega, W. C. Rheinboldt (1969, S. 476) *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York-London.

Zur Abkürzung definieren wir $r_k := (A - \rho_k I)x_k$ und $\theta_k := \angle(r_k, x_{k+1})$. Dann ist

$$\begin{aligned}
\|r_{k+1}\|_2^2 &= \|(A-\rho_k I)x_{k+1} + (\rho_k - \rho_{k+1})x_{k+1}\|_2^2 \\
&= \|(A-\rho_k I)x_{k+1}\|_2^2 + 2(\rho_k - \rho_{k+1})x_{k+1}^T(A-\rho_k I)x_{k+1} + (\rho_k - \rho_{k+1})^2 \\
&= \|(A-\rho_k I)x_{k+1}\|_2^2 + 2(\rho_k - \rho_{k+1})(\rho_{k+1} - \rho_k) + (\rho_k - \rho_{k+1})^2 \\
&= [x_k^T(A-\rho_k I)x_{k+1}]^2 - (\rho_{k+1} - \rho_k)^2 \\
&\qquad \text{(siehe Anfang des Beweises von Satz 2.6)} \\
&= \|r_k\|_2^2 \cos^2\theta_k - (\rho_{k+1} - \rho_k)^2.
\end{aligned}$$

Damit wird

$$(\rho_{k+1} - \rho_k)^2 \le \|r_k\|_2^2 - \|r_{k+1}\|_2^2.$$

Wegen der Monotonie der Folge $\{\|r_k\|_2\}$ konvergiert die rechte Seite dieser Ungleichung und damit auch $\{\rho_{k+1} - \rho_k\}$ gegen Null. Das erste Ziel zum Beweis der Konvergenz von $\{\rho_k\}$ ist damit erreicht.

Nun nehmen wir an, ρ^* sei ein Häufungspunkt der Folge $\{\rho_k\}$. Ist $\tau = 0$, so ist ρ^* offenbar notwendig ein Eigenwert von A, wovon es nur endlich viele gibt. Wir können daher davon ausgehen, dass $\tau > 0$. Aus

$$\underbrace{\|r_{k+1}\|_2^2}_{\to\tau^2} = \underbrace{\|r_k\|_2^2}_{\to\tau^2} \cos^2\theta_k - \underbrace{(\rho_{k+1} - \rho_k)^2}_{\to 0}$$

und

$$0 < \frac{1}{\|(A-\rho_k I)^{-1}x_k\|_2} = r_k^T x_{k+1} = \|r_k\|_2 \cos\theta_k$$

folgt $\cos\theta_k \to 1$. Hieraus wiederum folgt

$$\|r_k - \|r_k\|_2 x_{k+1}\|_2^2 = 2\|r_k\|_2^2(1 - \cos\theta_k) \to 0.$$

Ferner ist

$$\begin{aligned}
\|[(A-\rho_k I)^2 - \|r_k\|_2^2 \cos\theta_k I]x_k\|_2 &= \|(A-\rho_k I)(r_k - \|r_k\|_2 x_{k+1})\|_2 \\
&\le \|A - \rho_k I\|_2 \, \|r_k - \|r_k\|_2 x_{k+1}\|_2 \\
&\to 0.
\end{aligned}$$

Daher ist $\det[(A-\rho^* I)^2 - \tau^2] = 0$, womit bewiesen ist, dass die Folge $\{\rho_k\}$ nur endlich viele Häufungspunkte haben kann. Der erste Teil des Satzes ist bewiesen.

Nun kommen wir zum Beweis des zweiten Teiles. Sei (z, λ) ein Häufungspunkt der Folge $\{(x_k, \rho_k)\}$. Da wir $\tau = 0$ voraussetzen, ist $\|(A - \lambda I)z\| = 0$ und daher z ein Eigenvektor von A mit zugehörigem Eigenwert λ. Wegen des schon bewiesenen ersten Teiles des Satzes ist $\lim_{k\to\infty} \rho_k = \lambda$. Eine Inspektion des Beweises des lokalen Konvergenzsatzes zeigt, dass die gesamte Folge $\{x_k\}$ gegen z konvergiert. □

Bemerkung: Bei B. N. Parlett (1980, S. 76) wird bei der globalen Konvergenzanalyse auch der Fall $\tau > 0$ untersucht. □

4.2.4 Jacobi-Verfahren, Bisektionsverfahren

Sehr kurz wollen wir auf das Jacobi-Verfahren und das Bisektionsverfahren zur Bestimmung aller oder einiger Eigenwerte einer symmetrischen Matrix eingehen. Wir können uns mit Andeutungen begnügen, da diese Verfahren i. allg. ausführlich in einer Vorlesung Numerische Mathematik II (siehe z. B. J. WERNER (1992b)) besprochen werden.

Das Jacobi-Verfahren startet nicht wie das QR-Verfahren mit einem Reduktionsschritt. Ist die symmetrische Matrix $A \in \mathbb{R}^{n\times n}$ gegeben, so wird beim Jacobi-Verfahren $A^{(0)} := A$ gesetzt und $A^{(k+1)}$ aus $A^{(k)}$ durch eine Ähnlichkeitstransformation mit einer gewissen Givens-Rotation (in diesem Zusammenhang oft auch *Jacobi-Rotation* genannt, siehe z. B. J. W. DEMMEL (1997, S. 232)) Q_k gewonnen: $A^{(k+1)} := Q_k^T A^{(k)} Q_k$. Es ist das Ziel, dass die Folge $\{A^{(k)}\}$ gegen eine Diagonalmatrix konvergiert, die die Eigenwerte von A in der Diagonalen enthält. Als Maß für die Abweichung der symmetrischen Matrix A von der Diagonalgestalt wird

$$N(A) := \sum_{\substack{i,j=1\\ i\neq j}}^{n} a_{ij}^2$$

genommen. Die Idee des *klassischen Jacobi-Verfahrens* ist einfach. Im k-ten Schritt bestimmt man ein Paar (p,q) mit $1 \le p < q \le n$ und $|a_{pq}^{(k)}| = \max_{1\le i<j\le n} |a_{ij}^{(k)}|$, also ein dem Betrage nach maximales Außerdiagonalelement, und anschließend eine Givens-Rotation $Q_k = G_{pq}$ derart, dass in $A^{(k+1)} := Q_k^T A^{(k)} Q_k$ das (p,q)-Element annulliert wird. Dann kann leicht die Existenz einer von k unabhängigen Konstanten $c \in (0,1)$ (z. B. $c := 1 - 2/(n^2 - n)$) mit $N(A^{(k+1)}) \le N(A^{(k)})$ gezeigt. Also gilt $\lim_{k\to\infty} a_{ij}^{(k)} = 0$ für $i \neq j$, ferner kann z. B. mit Hilfe des Satzes von Gerschgorin gezeigt werden, dass die Diagonalelemente $a_{ii}^{(k)}$, $i = 1,\ldots,n$, (eventuell nach einer geeigneten Umnumerierung) gegen die Eigenwerte von A konvergieren.

Da die Suche nach einem betragsmaximalen Element verhältnismäßig aufwendig ist, geht man in der Praxis vom klassischen zum *zyklischen Jacobi-Verfahren* kombiniert mit einer sogenannten *Schwellenmethode* über. Hierbei werden die Außerdiagonalelemente (etwa in der strikten oberen Hälfte) in einer festen Reihenfolge durchlaufen, etwa zeilenweise von $(1,2)$ über $(1,n)$, $(2,3)$ über $(2,n)$ bis $(n-1,n)$, eine transformation aber nur dann durchgeführt, wenn das zu annullierende Element nicht kleiner als ein vorgegebener Schwellenwert $\epsilon > 0$ ist. Sobald alle Außerdiagonalelemente betragsmäßig kleiner als der Schwellenwert ϵ sind, so wird dieser heruntergesetzt und das Verfahren entsprechend fortgesetzt. Das ist schon alles was wir zum Jacobi-Verfahren aussagen wollen.

Auch auf das Bisektionsverfahren wollen wir nur sehr kurz eingehen. Gewöhnlich wird es auf unreduzierte (symmetrische) Tridiagonalmatrizen angewandt, am Anfang wird also notfalls ein Reduktionsschritt gemacht. Grundlage des Verfahrens sind die folgenden beiden Sätze, die wir ohne Beweis zitieren (siehe z. B. J. WERNER (1992b, S. 56 ff.)).

Satz 2.9 *Bei vorgegebenen reellen Zahlen $\delta_1,\ldots,\delta_n$ und von Null verschiedenen Zahlen $\gamma_1,\ldots,\gamma_{n-1}$ sei A_i für $i = 1,\ldots,n$ die unreduzierte, symmetrische Tridiagonalmatrix mit den Hauptdiagonalelementen $\delta_1,\ldots,\delta_i$ und den Nebendiagonalelementen*

$\gamma_1, \ldots, \gamma_{i-1}$. *Sei* $p_i(\mu) := \det(A_i - \mu I)$ *das zugehörige charakteristische Polynom. Dann gilt die Rekursionsformel*

$$p_i(\mu) = (\delta_i - \mu)p_{i-1}(\mu) - \gamma_{i-1}^2 p_{i-2}(\mu), \qquad i = 2, \ldots, n,$$

mit $p_0(\mu) := 1$ *und* $p_1(\mu) := \delta_1 - \mu$. *Ferner gilt:*

1. p_n *besitzt nur einfache (reelle) Nullstellen* $\lambda_1 > \cdots > \lambda_n$. *Insbesondere besitzt eine unreduzierte, symmetrische Tridiagonalmatrix paarweise verschiedene Eigenwerte.*

2. *Für* $j = 1, \ldots, n$ *ist* $\operatorname{sign} p_{n-1}(\lambda_j) = -\operatorname{sign} p_n'(\lambda_j)$.

3. *Für* $i = 1, \ldots, n-1$ *folgt aus* $p_i(\xi) = 0$, *dass* $p_{i+1}(\xi)p_{i-1}(\xi) < 0$.

Satz 2.10 *Gegeben sei die unreduzierte, symmetrische Tridiagonalmatrix* $A \in \mathbb{R}^{n\times n}$ *mit den Hauptdiagonalelementen* $\delta_1, \ldots, \delta_n$ *und den (von Null verschiedenen) Nebendiagonalelementen* $\gamma_1, \ldots, \gamma_{n-1}$. *Die Polynome* $p_i \in \Pi_i$, $i = 0, \ldots, n$, *seien durch die Rekursionsformel*

$$p_0(\mu) := 1, \quad p_1(\mu) := \delta_1 - \mu, \quad p_i(\mu) = (\delta_i - \mu)p_{i-1}(\mu) - \gamma_{i-1}^2 p_{i-2}(\mu), \quad i = 2, \ldots, n,$$

definiert. Für $\xi \in \mathbb{R}$ *sei* $N_n(\xi)$ *die Anzahl aufeinanderfolgender Vorzeichen Übereinstimmungen in* $(p_0(\xi), p_1(\xi), \ldots, p_n(\xi))$. *Hierbei wird vereinbart: Ist* $p_k(\xi) = 0$ *für ein* $k \in \{1, \ldots, n\}$, *so erhält* $p_k(\xi)$ *das Vorzeichen von* $p_{k-1}(\xi)$. *Dann gibt es genau* $N_n(\xi)$ *Eigenwerte von* A *(bzw. Nullstellen von* p_n*), welche größer oder gleich* ξ *sind.*

Es resultiert das folgende Bisektionsverfahren zur Berechnung des j-ten Eigenwertes λ_j einer unreduzierten, symmetrischen Tridiagonalmatrix $A \in \mathbb{R}^{n\times n}$. Genauer werden Intervalle $[a_k, b_k]$, $k = 1, 2, \ldots$, gewonnen, die garantiert den j-ten Eigenwert enthalten, so dass man das Verfahren abbricht, wenn die Intervalllänge $b_k - a_k$ hinreichend klein ist.

- Input: Gegeben seien die Hauptdiagonalelemente $\delta_1, \ldots, \delta_n$ und (von Null verschiedene) Nebendiagonalelemente $\gamma_1, \ldots, \gamma_{n-1}$ einer unreduzierten, symmetrischen Tridiagonalmatrix $A \in \mathbb{R}^{n\times n}$, ferner ein $j \in \{1, \ldots, n\}$ und ein Intervall $[a_1, b_1]$, welches den gesuchten j-ten Eigenwert λ_j von A enthält.

- Für $k = 1, 2, \ldots$:

  - Berechne $\xi_k := (a_k + b_k)/2$.
  - Berechne wie in Satz 2.10 die Zahl $N_n(\xi_k)$ aufeinanderfolgender Vorzeichenübereinstimmungen in $(p_0(\xi_k), \ldots, p_n(\xi_k))$.
  - Falls $N_n(\xi_k) < j$, dann: $a_{k+1} := a_k$, $b_{k+1} := \xi_k$.
  - Andernfalls: $a_{k+1} := \xi_k$, $b_{k+1} := b_k$.

Für genauere Hinweise zur Implementation des Bisektionsverfahrens sei auf J. H. Wilkinson, C. Reinsch (1971, S. 249 ff.) verwiesen[11]. Bemerkt sei ferner, dass man mit Hilfe der inversen Vektoriteration i. allg. mit wenig Aufwand gute Näherungen für zugehörige Eigenvektoren erhält.

Ein weiterer Ansatz für ein Bisektionsverfahren beruht auf dem Sylvesterschen Trägheitssatz (siehe Satz 3.3 in Abschnitt 2.3, wir halten uns hier an J. W. Demmel (1997, S. 228))). Die Trägheit einer symmetrischen Matrix $A \in \mathbb{R}^{n\times n}$ ist hierbei definiert durch $\text{inertia}(A) = (i_+, i_-, i_0)$, wobei i_+ die Anzahl positiver, i_- die Anzahl negativer und i_0 die Anzahl verschwindender Eigenwerte von A bedeutet. Der Sylvestersche Trägheitssatz sagt aus, dass $\text{inertia}(A) = \text{inertia}(X^T AX)$ für jede nichtsinguläre Matrix $X \in \mathbb{R}^{n\times n}$. Angenommen nun, dass mit Hilfe Gaussscher Elimination eine Faktorisierung $A - zI = LDL^T$ mit einer nichtsingulären Matrix $L \in \mathbb{R}^{n\times n}$ (etwa einer unteren Dreiecksmatrix mit Einsen in der Diagonalen) und einer Diagonalmatrix $D \in \mathbb{R}^{n\times n}$ berechnet werden kann. Dann ist $\text{inertia}(A - zI) = \text{inertia}(D)$, ferner ist die Trägheit einer Diagonalmatrix trivial zu berechnen. Sei nun $z_1 < z_2$, ferner seien $\text{inertia}(A - z_1 I)$ und $\text{inertia}(A - z_2 I)$ berechnet. Dann kennt man auch die Anzahl der Eigenwerte von A in $[z_1, z_2)$, denn diese ist die Anzahl der Eigenwerte von A, die kleiner als z_2 sind (bzw. die Anzahl negativer Eigenwerte von $A - z_2 I$) vermindert um die Anzahl der Eigenwerte von A, die kleiner als z_1 sind (bzw. die Anzahl negativer Eigenwerte von $A - z_1 I$).

Die Berechnung einer LDL^T-Zerlegung ist bei einer vollbesetzten (symmetrischen) Matrix zu teuer, da sie $O(n^3)$ flops benötigt und wiederholt vorgenommen müsste. Günstiger ist es, wenn A eine Tridiagonalmatrix ist. Ein Ansatz

$$
\begin{aligned}
A - zI &= \begin{pmatrix} \delta_1 - z & \gamma_1 & & \\ \gamma_1 & \delta_2 - z & \ddots & \\ & \ddots & \ddots & \gamma_{n-1} \\ & & \gamma_{n-1} & \delta_n - z \end{pmatrix} \\
&= \begin{pmatrix} 1 & & & \\ l_1 & \ddots & & \\ & \ddots & \ddots & \\ & & l_{n-1} & 1 \end{pmatrix} \begin{pmatrix} d_1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & d_n \end{pmatrix} \begin{pmatrix} 1 & l_1 & & \\ & \ddots & \ddots & \\ & & \ddots & l_{n-1} \\ & & & 1 \end{pmatrix}
\end{aligned}
$$

liefert $\delta_1 - z = d_1$, $d_1 l_1 = \gamma_1$ und danach

$$l_{i-1}^2 d_{i-1} + d_i = \delta_i - z, \quad d_i l_i = \gamma_i \qquad (i = 2, \ldots, n).$$

Da wir nur die Trägheit von $A - zI$ berechnen, sind wir nur an der Diagonalmatrix D, nicht aber an der unteren Dreiecksmatrix L interessiert. Man erhält wegen

$$d_i = \delta_i - z - l_{i-1}^2 d_{i-1} = \delta_i - z(\gamma_{i-1}/d_{i-1})^2 d_{i-1} = \delta_i - z - \frac{\gamma_{i-1}^2}{d_{i-1}}$$

[11] J. H. Wilkinson and C. Reinsch (1971) *Handbook for Automatic Computation. Vol. II, Linear Algebra*. Springer-Verlag, Berlin-Heidelberg-New York.

die Rekursionsformel

$$d_1 := \delta_1 - z, \quad d_i := \delta_i - z - \frac{\gamma_{i-1}^2}{d_{i-1}} \qquad (i = 2, \ldots, n).$$

Bei J. W. DEMMEL (1997, S. 230) findet man Bemerkungen zur Stabilität dieses Verfahrens.

4.2.5 Aufgaben

1. Seien $A, B \in \mathbb{R}^{n\times n}$ symmetrisch mit Eigenwerten $\lambda_1 \geq \cdots \lambda_n$ bzw. $\mu_1 \geq \cdots \geq \mu_n$. Die Eigenwerte von $A + B$ seien $\nu_1 \geq \cdots \geq \nu_n$. Mit Hilfe des Courantschen Minimum-Maximum-Prinzips zeige man, dass

$$\lambda_j + \mu_n \leq \nu_j \leq \lambda_j + \mu_1, \qquad j = 1, \ldots, n.$$

2. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch mit Eigenwerten $\lambda_1, \ldots, \lambda_n$. Für alle $\lambda \in \mathbb{R}$ und alle $x \in \mathbb{R}^n \setminus \{0\}$ ist dann

$$\min_{j=1,\ldots,n} |\lambda - \lambda_j| \leq \frac{\|\lambda x - Ax\|_2}{\|x\|_2}.$$

   Hinweis: Man setze $\delta A := (\lambda x - Ax)x^T/\|x\|_2^2$ und wende den Satz von Bauer-Fike an.

3. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch. Für $x \in \mathbb{R}^n \setminus \{0\}$ bezeichne $\rho(x) := x^T Ax/x^T x$ den zugehörigen Rayleigh-Quotienten. Seien $\lambda_1, \ldots, \lambda_n$ die Eigenwerte von A. Dann gilt:

   (a) Für jedes $x \in \mathbb{R}^n \setminus \{0\}$ ist

$$\min_{j=1,\ldots,n} |\rho(x) - \lambda_j| \leq \left[\left(\frac{\|Ax\|_2}{\|x\|_2}\right)^2 - \rho(x)^2\right]^{1/2}.$$

   (b) Ist $x \in \mathbb{R}^n \setminus \{0\}$ und $\rho(x)$ kein Eigenwert von A, so ist

$$\min_{j=1,\ldots,n} |\rho(x) - \lambda_j| \leq \frac{\|x\|_2}{\|[A - \rho(x)I]^{-1}x\|_2}.$$

   Hinweis: Man wende Aufgabe 2 an.

4. Sei $M \in \mathbb{R}^{n\times m}$ symmetrisch mit Eigenwerten $\mu_1 \geq \cdots \geq \mu_n$. Die Matrix $X \in \mathbb{R}^{n\times(n-1)}$ habe orthonormierte Spalten, es sei also $X^T X = I$. Bezeichnet man mit $\nu_1 \geq \cdots \geq \nu_{n-1}$ die Eigenwerte von $X^T MX \in \mathbb{R}^{(n-1)\times(n-1)}$, so ist

$$\mu_{j+1} \leq \nu_j \leq \mu_j, \qquad j = 1, \ldots, n-1.$$

5. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch, $a \in \mathbb{R}^n$ und $M := A + \alpha\, aa^T$ mit $\alpha \neq 0$. Seien $\lambda_1 \geq \cdots \geq \lambda_n$ die Eigenwerte von A und $\mu_1 \geq \cdots \geq \mu_n$ die Eigenwerte von M. Dann gilt:

   (a) Ist $\alpha > 0$, so ist $\lambda_n \leq \mu_n \leq \lambda_{n-1} \leq \cdots \leq \lambda_1 \leq \mu_1$.

   (b) Ist $\alpha < 0$, so ist $\mu_n \leq \lambda_n \leq \mu_{n-1} \leq \cdots \leq \mu_1 \leq \lambda_1$.

6. Sei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$. Sind $\sigma_1 \geq \cdots \geq \sigma_n$ die singulären Werte von A, so ist

$$\sigma_j = \min_{N_j \in \mathcal{N}_j} \max_{0 \neq x \in N_j} \frac{\|Ax\|_2}{\|x\|_2}, \qquad j = 1, \ldots, n,$$

wobei

$$\mathcal{N}_j := \{N_j \subset \mathbb{R}^n : \ N_j \text{ ist linearer Teilraum mit } \dim(N_j) = n + 1 - j\}.$$

7. Seien $A, B \in \mathbb{R}^{m\times n}$ mit $m \geq n$ und singulären Werten $\sigma_1 \geq \cdots \geq \sigma_n$ bzw. $\tau_1 \geq \cdots \geq \tau_n$ gegeben. Dann ist

$$|\sigma_j - \tau_j| \leq \|A - B\|_2, \qquad j = 1, \ldots, n.$$

8. Man berechne zu

$$A := \begin{pmatrix} 5 & 4 & 1 & 1 \\ 4 & 5 & 1 & 1 \\ 1 & 1 & 4 & 2 \\ 1 & 1 & 2 & 4 \end{pmatrix}$$

eine orthogonal ähnliche Tridiagonalmatrix. Mit Hilfe des QR-Verfahrens berechne man anschließend alle Eigenwerte von A. Hierbei wende man zum Vergleich beide angegebenen Shift-Strategien an.

9. Man wende das Rayleigh-Quotienten-Verfahren auf die Matrix

$$A := \begin{pmatrix} 5 & 4 & 1 & 1 \\ 4 & 5 & 1 & 1 \\ 1 & 1 & 4 & 2 \\ 1 & 1 & 2 & 4 \end{pmatrix}$$

an, wobei man mit $x_0 := \frac{1}{2}(1, 1, 1, 1)^T$ starte. Gegen welchen Eigenwert konvergiert das Verfahren?

10. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und $a_{pq} \neq 0$ mit $1 \leq p < q \leq n$. Auf die folgende Weise berechne man eine Givens-Rotation $G_{pq} = G_{pq}(c, s)$.

- Berechne
$$\theta := \frac{a_{qq} - a_{pp}}{2a_{pq}}.$$
- Berechne
$$t := \frac{\operatorname{sign}(\theta)}{|\theta| + \sqrt{1 + \theta^2}}.$$
- Berechne
$$c := \frac{1}{\sqrt{1 + t^2}}, \qquad s := ct.$$

Man definiere $A^+ := G_{pq}^T A G_{pq}$ und zeige:

(a) Es ist $(A^+)_{pq} = (A^+)_{qp} = 0$.

(b) Es ist $N(A^+) = N(A) - 2a_{pq}^2$, wobei

$$N(A) := \sum_{\substack{i,j=1 \\ i \neq j}}^{n} a_{ij}^2.$$

(c) Ist (p, q) ein Indexpaar mit $1 \leq p < q \leq n$ und $|a_{pq}| = \max_{1 \leq i < j \leq n} |a_{ij}|$, so ist

$$N(A^+) \leq \left(1 - \frac{2}{n^2 - n}\right) N(A).$$

Kapitel 5

Iterationsverfahren bei linearen Gleichungssystemen

Iterationsverfahren zur Lösung eines linearen Gleichungsssystems $Ax = b$ werden dann benutzt, wenn direkte Methoden aus Zeit- oder Speicherplatzgründen nicht in Frage kommen. Bei vielen, durch die Diskretisierung von Randwertaufgaben bei partiellen Differentialgleichungen gewonnenen linearen Gleichungssystemen ist dies der Fall. An neuerer (schon in der Einleitung angegebener) Literatur seien vor allem A. Greenbaum (1997), Y. Saad (1996) und W. Hackbusch (1993) genannt, aber auch an J. W. Demmel (1997) werden wir uns orientieren. Auf die aus einer Vorlesung Numerische Mathematik I bekannten Tatsachen werden wir nur sehr kurz eingehen.

5.1 Das Modellproblem und elementare Iterationsverfahren

5.1.1 Die Poisson-Gleichung als Modellproblem

Wir betrachten die Poisson-Gleichung mit homogenen Dirichlet Randbedingungen auf dem Einheitsquadrat in der Ebene, also die Aufgabe, eine Funktion $u \in C^2(\Omega) \cap C(\overline{\Omega})$ zu finden mit

$$-\Delta u(x,y) = f(x,y) \quad ((x,y) \in \Omega), \qquad u(x,y) = 0 \quad ((x,y) \in \partial\Omega),$$

wobei

$$\Omega := \{(x,y) \in \mathbb{R}^2 : 0 < x, y < 1\}$$

das offene Einheitsquadrat, $\overline{\Omega}$ (oft auch als $\mathrm{cl}\,(\Omega)$ bezeichnet) das abgeschlossene Einheitsquadrat, $\partial\Omega$ dessen Rand und der Differentialoperator Δ den Laplace Operator bedeutet, also durch

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

gegeben ist. Bei einer Diskretisierung mittels finiter Differenzen legt man über $\overline{\Omega}$ ein Gitter und ersetzt den Laplace Operator in den Gitterpunkten durch einen Finite Differenzen Operator. Nimmt man in x- und y-Richtung dieselbe Maschenweite, so definiert

man zunächst $h := 1/(N+1)$ mit einem $N \in \mathbb{N}$ und anschließend die Gitterpunkte

$$(x_i, y_j) = (ih, jh), \qquad 0 \le i, j \le N+1.$$

Dann ist bei hinreichend glattem u nach leichter Rechnung

$$-\Delta u(x_i, y_j) = \frac{4u_{ij} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1}}{h^2} + \tau_{ij},$$

wobei $u_{ij} = u(x_i, y_j)$ und $\tau_{ij} = O(h^2)$. Das diskrete Analogon der Poisson Gleichung mit homogenen Randbedingungen ist daher

$$(*) \qquad 4u_{ij} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1} = h^2 f_{ij}, \qquad 1 \le 1, j \le N,$$

wobei $f_{ij} := f(x_i, y_j)$ und wegen der Randbedingungen

$$u_{0,j} = u_{N+1,j} = u_{i,0} = u_{i,N+1} = 0, \qquad 0 \le i, j \le N+1.$$

Daher ist $(*)$ ein lineares Gleichungssystem in $n := N^2$ Unbekannten und Gleichungen. Es gibt mindestens zwei Möglichkeiten, $(*)$ als eine Matrix-Gleichung zu schreiben[1].

- Man fasse die Unbekannten u_{ij}, $1 \le i, j \le N$, zu einer Matrix $U = (u_{ij}) \in \mathbb{R}^{N\times N}$ zusammen, entsprechend sei $F := (f_{ij}) \in \mathbb{R}^{N\times N}$.

Definiert man dann die Tridiagonalmatrix $T_N \in \mathbb{R}^{N\times N}$ durch

$$T_N := \begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix},$$

beachtet man ferner, dass

$$(T_N U)_{ij} = 2u_{ij} - u_{i-1,j} - u_{i+1,j}, \qquad (UT_N)_{ij} = 2u_{ij} - u_{i,j-1} - u_{i,j+1},$$

so erkennt man, dass man $(*)$ in der Form

$$T_N U + U T_N = h^2 F$$

schreiben kann. Dies ist ein lineares Gleichungssystem für die unbekannten Einträge in der Matrix U, auch wenn es nicht in dem übliche "$Ax = b$" Format geschrieben ist, in dem die Unbekannten in einem *Vektor* (und nicht einer Matrix) zusammengefasst sind. Wir können aber hieraus die Eigenwerte und Eigenvektoren der zu grunde liegenden Matrix A berechnen, da $Ax = \lambda x$ äquivalent zu $T_N U + U T_N = \lambda U$ ist. Angenommen, (λ_i, z_i) und (λ_j, z_j) seien zwei Eigenpaare von T_N, ferner $U := z_i z_j^T$. Dann ist

$$\begin{aligned} T_N U + U T_N &= (T_N z_i) z_j^T + z_i (z_j^T T_N) \\ &= (T_N z_i) z_j^T + z_i (T_N z_j)^T \\ &= (\lambda_i + \lambda_j) z_i z_j^T \\ &= (\lambda_i + \lambda_j) U, \end{aligned}$$

[1] Wir folgen hier fast wörtlich der Darstellung bei J. W. Demmel (1997, S. 271).

so dass $U = z_i z_j^T$ ein "Eigenvektor" und $\lambda_i + \lambda_j$ ein Eigenwert. Es kommt also darauf an, die Eigenwerte und Eigenvektoren von T_N zu bestimmen. Es liegt nahe, hier zum kontinuierlichen Analogon zurückzugehen und die Eigenwertaufgabe

$$-u''(x) = \lambda u(x) \quad (x \in (0,1)), \qquad u(0) = u(1) = 0$$

zu lösen. Eigenfunktionen hierfür sind offenbar $z_i(x) := \sin(i\pi x)$, $i = 1, 2, \ldots$. Für einen Eigenvektor z_i von T_N machen wir daher den Ansatz

$$z_i = (\sin(i\pi x_1), \ldots, \sin(i\pi x_N))^T.$$

Für $j = 1, \ldots, N$ ist daher

$$\begin{aligned}(T_N z_i)_j &= -(z_i)_{j-1} + 2(z_i)_j - (z_i)_{j+1} \\ &= -\sin(i(j-1)\pi h) + 2\sin(ij\pi h) - \sin(i(j+1)\pi h) \\ &= 2[1 - \cos(i\pi h)]\sin(ij\pi h) \\ &= 2[1 - \cos(i\pi/(N+1))](z_i)_j,\end{aligned}$$

mit

$$\lambda_i := 2[1 - \cos(i\pi/(N+1))], \qquad z_i := \begin{pmatrix} \sin(i\pi/(N+1)) \\ \vdots \\ \sin(i\pi N/(N+1)) \end{pmatrix}, \qquad i = 1, \ldots, N,$$

hat man die (Eigenwert, Eigenvektor)-Paare von T_N bestimmt. Dies sind alle, denn die Eigenwerte λ_i der symmetrischen Matrix T_N sind paarweise verschieden, die zugehörigen Eigenvektoren z_i bilden daher nach entsprechender Normierung (Vorfaktor $\sqrt{2/(N+1)}$) ein Orthonormalsystem. Hieraus folgt leicht, dass $U^{ij} := z_i z_j^T$, $1 \le i, j \le N$, linear unabhängig sind (hierbei nehmen wir an, die z_i seien normiert). Denn ist $\sum_{i,j=1}^N a_{ij} z_i z_j^T = 0$, so folgt nach Multiplikation mit z_k, dass $\sum_{i=1}^N a_{ik} z_i = 0$ und hieraus wiederum $a_{ik} = 0$, $1 \le i, k \le N$. Also sind $\lambda_i + \lambda_j$, $1 \le i, j \le N$ alle Eigenwerte der $N^2 \times N^2$-Matrix A.

Die zweite Möglichkeit, $(*)$ als lineares Gleichungssystem zu schreiben, besteht in der folgenden naheliegenden Idee:

- Man fasse die Unbekannten u_{ij}, $1 \le i, j \le N$, zu einem Vektor $u = (u_{ij}) \in \mathbb{R}^{N^2}$ zusammen. Hierzu muss man die u_{ij} anordnen, z. B. zeilenweise von unten nach oben (oder spaltenweise von oben nach unten), hier gibt es viele Möglichkeiten.

Ordnet man die Unbekannten z. B. zeilenweise von links nach rechts und von unten nach oben an, so erhält man den Vektor

$$u = (u_{1,1}, \ldots, u_{N,1}, u_{1,2}, \ldots, u_{N,2}, \ldots, u_{1,N}, \ldots, u_{N,N})^T \in \mathbb{R}^{N^2}.$$

Entsprechend ordne man auch (f_{ij}) zu einem Vektor der Länge N^2 an. Man erhält dann das lineare Gleichungssystem

$$T_{N\times N} u = h^2 f,$$

wobei

$$T_{N\times N} := \begin{pmatrix} T_N + 2I_N & -I_N & & \\ -I_N & \ddots & \ddots & \\ & \ddots & \ddots & -I_N \\ & & -I_N & T_N + 2I_N \end{pmatrix}$$

und nach der für T_N gegebenen Definition

$$T_N + 2I_N = \begin{pmatrix} 4 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{pmatrix}.$$

5.1.2 Elementare Iterationsverfahren

In diesem Unterabschnitt wollen wir sehr kurz auf die aus einer Vorlesung Numerische Mathematik I her bekannten Iterationsverfahren, wie etwa dem Jacobi- (Gesamtschrittverfahren), Gauß-Seidel- (Einzelschrittverfahren), sukzessiven Overrelaxations- (SOR) und der Tschebyscheff-Beschleunigung mit symmetrischer sukzessiver Overrelaxation (SSOR).

Gegeben sei stets ein lineares Gleichungssystem $Ax = b$, wobei $A \in \mathbb{C}^{n\times n}$ und $b \in \mathbb{C}^n$. Die Darstellung $A = M - N$ mit nichtsingulärem $M \in \mathbb{C}^{n\times n}$ nennen wir eine *Zerlegung* von A. Mit Hilfe einer Zerlegung von A kann das lineare Gleichungssystem $Ax = b$ als äquivalente Aufgabe $Mx = Nx + b$ geschrieben werden, welche wiederum Anlass zu einem Iterationsverfahren

$$Mx^{(k+1)} = Nx^{(k)} + b \quad \text{bzw.} \quad x^{(k+1)} := M^{-1}Nx^{(k)} + M^{-1}b$$

gibt. Hierbei sollte die Matrix M nicht nur nichtsingulär, sondern auch so beschaffen sein, dass lineare Gleichungssysteme mit M als Koeffizientenmatrix "leicht" lösbar sind. Dies ist z. B. dann der Fall, wenn M eine Diagonalmatrix (Aufwand ist n) oder eine untere Dreiecksmatrix (Aufwand ist n^2) ist. Weiter ist intuitiv klar, dass für eine gute Konvergenz die Matrix M "möglichst nahe" bei A sein sollte. Mit $R := M^{-1}N$ bezeichnen wir die zu obigen Iterationsverfahren gehörende *Iterationsmatrix*.

Ein bekannter Konvergenzsatz ist das folgende Ergebnis (siehe z. B. J. W. Demmel (1997, S. 280), Y. Saad (1996, S. 104) J. Werner (1992a, S. 123)). In diesem kommt der Begriff des *Spektralradius* $\rho(A)$ einer Matrix $A \in \mathbb{C}^{n\times n}$ vor, der bekanntlich das Maximum aller Beträge der Eigenwerte von A ist. Ferner erinnern wir daran, dass $\rho(A) \le \|A\|$ für jede natürliche Matrixnorm.

Satz 1.1 *Sei $A = M - N$ mit nichtsingulärem $M \in \mathbb{C}^{n\times n}$ eine Zerlegung von $A \in \mathbb{C}^{n\times n}$, ferner sei $b \in \mathbb{C}^n$ gegeben. Dann gilt:*

1. *Ist $\rho(M^{-1}N) < 1$, so ist A nichtsingulär und die aus*

   $$(*) \qquad Mx^{(k+1)} = Nx^{(k)} + b$$

   gewonnene Folge $\{x^{(k)}\}$ konvergiert für jedes $x^{(0)} \in \mathbb{C}^n$ gegen die (eindeutige) Lösung x^* von $Ax = b$.

2. *Ist A nichtsingulär und konvergiert die aus $(*)$ gewonnene Folge $\{x^{(k)}\}$ für jedes $x^{(0)} \in \mathbb{C}^n$ gegen die eindeutige Lösung x^* von $Ax = b$, so ist $\rho(M^{-1}N) < 1$.*

Durch den letzten Satz haben wir eine quantitative Bedingung, nämlich $\rho(M^{-1}N) < 1$, für die qualitative Aussage, M sollte "in der Nähe" von A sein, erhalten.

Beim Jacobi-, Gauß-Seidel- und dem SOR-Verfahren geht man von einer Darstellung $A = A_L + A_D + A_R$ aus, wobei A_L den strikten linken Teil von A, $A_D = \text{diag}\,(A)$ die Diagonale von A und A_R den strikten rechten Teil von A bedeutet. Wir geben rasch die Zerlegungen von A an, die zu den entsprechenden Verfahren führen. Wir nehmen hierzu an, dass A_D nichtsingulär ist, was bei nichtsingulärem A durch Umordnen erreicht werden kann.

- Jacobi-Verfahren: Setze $M := A_D$, die Iterationsmatrix ist

$$R_J := -A_D^{-1}(A_L + A_R).$$

  Komponentenweise bedeutet dies die Iterationsvorschrift, dass $x^{(k+1)}$ aus $x^{(k)}$ folgendermaßen berechnet wird:

  Für $i = 1, \ldots, n$:

  – $x_i^{(k+1)} := \dfrac{1}{a_{ii}}\Big(b_i - \displaystyle\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j^{(k)}\Big)$.

- Gauß-Seidel-Verfahren: Setze $M := A_D + A_L$, die Iterationsmatrix ist

$$R_{GS} := -(A_D + A_L)^{-1}A_R.$$

  Komponentenweise bedeutet dies die Iterationsvorschrift, dass $x^{(k+1)}$ aus $x^{(k)}$ folgendermaßen berechnet wird:

  Für $i = 1, \ldots, n$:

  – $x_i^{(k+1)} := \dfrac{1}{a_{ii}}\Big(b_i - \displaystyle\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\Big)$.

- SOR-Verfahren: Mit $\omega \neq 0$ setze $M(\omega) := (1/\omega)A_D + A_L$, die Iterationsmatrix ist

$$R_{SOR}(\omega) := (A_D + \omega A_L)^{-1}[(1-\omega)A_D - \omega A_R].$$

  Komponentenweise bedeutet dies die Iterationsvorschrift, dass $x^{(k+1)}$ aus $x^{(k)}$ folgendermaßen berechnet wird:

  Für $i = 1, \ldots, n$:

  – $x_i^{(k+1)} := x_i^{(k)} + \dfrac{\omega}{a_{ii}}\Big(b_i - \displaystyle\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i} a_{ij}x_j^{(k)}\Big)$.

Für einige bekannte Konvergenzaussagen verweisen wir auf die Aufgaben, Beweise findet man in fast allen Büchern, die sich mit Iterationsverfahren bei linearen Gleichungssystemen beschäftigen.

Bevor wir im nächsten Unterabschnitt ein wenig auf die optimale Wahl des Relaxationsparameters eingehen, wollen wir hier noch die Idee der Tschebyscheff-Beschleunigung schildern, z. B. angewandt auf die sogenannte symmetrische sukzessive Overrelaxation (SSOR). Angenommen, das lineare Gleichungssystem $Ax = b$ sei in eine äquivalente Fixpunktaufgabe $x = Rx + c$ transformiert worden, wobei $\rho(R) < 1$. Für ein beliebiges $x^{(0)}$ konvergiert dann die aus $x^{(k+1)} := Rx^{(k)} + c$ gewonnene Folge $\{x^{(k)}\}$ gegen den Fixpunkt x. Es stellt sich die Frage, ob eine geeignete Linearkombination von $x^{(0)}, \ldots, x^{(k)}$, $k = 1, 2, \ldots$, eine verbesserte Näherung für x ist. Hierzu machen wir den Ansatz $y^{(k)} := \sum_{i=0}^{k} \alpha_{ki} x^{(i)}$, wobei wir von den Koeffizienten fordern, dass $\sum_{i=0}^{k} \alpha_{ki} = 1$, damit aus $x^{(0)} = x$ (dies impliziert $x^{(i)} = x$, $i = 0, \ldots, k$) auch $y^{(k)} = x$ folgt. Dann ist

$$y^{(k)} - x = \sum_{i=0}^{k} \alpha_{ki} x^{(i)} - x = \sum_{i=0}^{k} \alpha_{ki}(x^{(i)} - x) = \sum_{i=0}^{k} \alpha_{ki} R^i (x^{(0)} - x) = p_k(R)(x^{(0)} - x),$$

wobei $p_k(R) := \sum_{i=0}^{k} \alpha_{ki} R^i$ ein Polynom vom Grad k ist. Angenommen wir wüssten, dass die Eigenwerte von R reell sind und im Intervall $[-\rho, \rho]$ liegen, wobei $\rho < 1$ (dies ist der Fall, wenn R nur reelle Eigenwerte besitzt und $\rho(R) < 1$ ist). Es liegt nahe, unter allen Polynomen $p_k \in \Pi_k$ mit $p_k(1) = 1$ eines zu finden, für welches $\max_{t \in [-\rho,\rho]} |p_k(t)|$ minimal ist. Denn ist dies gelungen, so ist auch $\rho(p_k(R))$ klein. Eine explizite Lösung dieser Aufgabe kann mit Hilfe von *Tschebyscheff-Polynomen* angegeben werden. Und zwar ist die Lösung durch

$$p_k(t) := \frac{T_k(t/\rho)}{T_k(1/\rho)}$$

gegeben, wobei $T_k(\cdot)$ das k-te Tschebyscheff-Polynom erster Art bedeutet. Diese können z. B. durch die dreigliedrige Rekursionsformel

$$T_0(t) := 1, \quad T_1(t) := t, \quad T_k(t) := 2tT_{k-1}(t) - T_{k-2}(t)$$

definiert werden. Setzt man

$$\mu_k := \frac{1}{T_k(1/\rho)},$$

so ist $p_k(R) = \mu_k T_k(R/\rho)$, ferner gilt die Rekursionsformel

$$\mu_0 := 1, \quad \mu_1 := \rho, \quad \mu_k := \frac{1}{2/(\rho\mu_{k-1}) - 1/\mu_{k-2}}.$$

Dann ist

$$\begin{aligned} y^{(k)} - x &= p_k(R)(x^{(0)} - x) \\ &= \mu_k T_k(R/\rho)(x^{(0)} - x) \\ &= \mu_k [2(R/\rho)T_{k-1}(R/\rho) - T_{k-2}(R/\rho)](x^{(0)} - x) \end{aligned}$$

$$
\begin{aligned}
&= \mu_m\Big[2(R/\rho)\frac{p_{k-1}(R/\rho)}{\mu_{k-1}} - \frac{p_{k-2}(R/\rho)}{\mu_{k-2}}\Big](x^{(0)} - x)\\
&= \mu_k\Big[2(R/\rho)\frac{y^{(k-1)} - x}{\mu_{k-1}} - \frac{y^{(k-2)} - x}{\mu_{k-2}}\Big].
\end{aligned}
$$

Hieraus folgt

$$
y^{(k)} = \frac{2\mu_k}{\mu_{k-1}}(R/\rho)y^{(k-1)} - \frac{\mu_k}{\mu_{k-2}}y^{(k-2)} + d^{(k)},
$$

wobei

$$
\begin{aligned}
d^{(k)} &:= x - \frac{2\mu_k}{\mu_{k-1}}(R/\rho)x + \frac{\mu_k}{\mu_{k-2}}x\\
&= x - \frac{2\mu_k}{\mu_{k-1}}\Big(\frac{x-c}{\rho}\Big) + \frac{\mu_k}{\mu_{k-2}}x\\
&\qquad (\text{da } x = Rx + c)\\
&= \mu_k\Big[\mu_k - \frac{2}{\rho\mu_{k-1}} + \frac{1}{\mu_{k-2}}\Big]x + \frac{2\mu_k}{\rho\mu_{k-1}}c\\
&= \frac{2\mu_k}{\rho\mu_{k-1}}c
\end{aligned}
$$

wegen der Rekusionsformel für μ_k. Damit haben wir schließlich

$$
y^{(k)} = \frac{2\mu_k}{\rho\mu_{k-1}}Ry^{(k-1)} - \frac{\mu_k}{\mu_{k-2}}y^{(k-2)} + \frac{2\mu_k}{\rho\mu_{k-1}}c
$$

erhalten. Die Tschebyscheff-Beschleunigung der Iterationsvorschrift $x^{(k+1)} = Rx^{(k)} + c$ lautet also wie folgt:

- Input: Die Matrix $R \in \mathbb{R}^{n\times n}$ habe nur reelle Eigenwerte, von denen vorausgesetzt wird, dass sie in $[-\rho, \rho]$ liegen. Gegeben ferner $c \in \mathbb{R}^n$ und $x^{(0)} \in \mathbb{R}^n$.
- Setze bzw. berechne $\mu_0 := 1$, $\mu_1 := \rho$, $y^{(0)} := x^{(0)}$ und $y^{(1)} := Rx^{(0)} + c$.
- Für $k = 2, 3, \ldots$:
  - Berechne $\mu_k := \dfrac{1}{2/(\rho\mu_{k-1}) - 1/\mu_{k-2}}$.
  - Berechne $y^{(k)} := \dfrac{2\mu_k}{\rho\mu_{k-1}}Ry^{(k-1)} - \dfrac{\mu_k}{\mu_{k-2}}y^{(k-2)} + \dfrac{2\mu_k}{\rho\mu_{k-1}}c$.

Man beachte, dass dies Verfahren nicht wesentlich aufwendiger als das Ausgangsverfahren $x^{(k+1)} := Rx^{(k)} + c$ ist, denn entscheidend für den Aufwand wird in den meisten Fällen die Multiplikation von R mit einem Vektor sein, was in beiden Verfahren genau einmal auftritt.

Die Tschebyscheff-Beschleunigung kann leider i. allg. nicht auf das SOR-Verfahren angewandt werden, da die Iterationsmatrix nicht nur reelle Eigenwerte haben wird. Eine erfolgreiche Idee, jedenfalls dann, wenn die zugrunde liegende Matrix $A \in \mathbb{R}^{n\times n}$ symmetrisch ist, besteht darin, das SOR-Verfahren zu symmetrisieren. Das sieht folgendermaßen aus.

- Berechne $x^{(k+1/2)}$ aus $x^{(k)}$ mittels

$$\Big[\frac{1}{\omega}A_D + A_L\Big]x^{(k+1/2)} = \Big[\Big(\frac{1}{\omega} - 1\Big)A_D - A_R\Big]x^{(k)} + b,$$

  d. h. mache von $x^{(k)}$ ausgehend einen normalen SOR-Schritt.

- Berechne $x^{(k+1)}$ aus $x^{(k+1/2)}$ mittels

$$\Big[\frac{1}{\omega}A_D + A_R\Big]x^{(k+1)} = \Big[\Big(\frac{1}{\omega} - 1\Big)A_D - A_L\Big]x^{(k+1/2)} + b,$$

  d. h. mache einen SOR-Schritt bei dem die Rollen von A_L und A_R vertauscht sind. Hier werden also die Komponenten von $x^{(k+1)}$ in fallender Reihenfolge bestimmt.

Schreiben wir dieses Verfahren als $x^{(k+1)} = R_{SSOR}(\omega)x^{(k)} + c$, indem wir beide Schritte zu einem zusammenfassen, setzen wir ferner

$$\tilde{A}_L := A_D^{-1}A_L, \qquad \tilde{A}_R := A_D^{-1}A_R,$$

so erhalten wir als Iterationsmatrix

$$\begin{aligned} R_{SSOR}(\omega) &= \Big[\frac{1}{\omega}A_D + A_R\Big]^{-1}\Big[\Big(\frac{1}{\omega} - 1\Big)A_D - A_L\Big]\Big[\frac{1}{\omega}A_D + A_L\Big]^{-1}\Big[\Big(\frac{1}{\omega} - 1\Big)A_D - A_R\Big] \\ &= (A_D + \omega A_R)^{-1}[(1-\omega)A_D - \omega A_L](A_D + \omega A_L)^{-1}[(1-\omega)A_D - \omega A_R] \\ &= (I + \omega\tilde{A}_R)^{-1}[(1-\omega)I - \omega\tilde{A}_L](I + \omega\tilde{A}_L)^{-1}[(1-\omega)I - \omega\tilde{A}_R]. \end{aligned}$$

Wir nehmen nun an, es sei $A \in \mathbb{R}^{n\times n}$ symmetrisch, insbesondere also $\tilde{A}_R = \tilde{A}_L^T$. Wir wollen uns überlegen, dass $R_{SSOR}(\omega)$ in diesem Falle einer symmetrischen Matrix ähnlich ist und daher nur reelle Eigenwerte besitzt. Definiert man nämlich

$$\tilde{R}_{SSOR}(\omega) := (I + \omega\tilde{A}_R)R_{SSOR}(\omega)(I + \omega\tilde{A}_R)^{-1},$$

so ist

$$\begin{aligned} \tilde{R}_{SSOR}(\omega) &= [(1-\omega)I - \omega\tilde{A}_L](I + \omega\tilde{A}_L)^{-1}[(1-\omega)I - \omega\tilde{A}_L^T](I + \omega\tilde{A}_L^T)^{-1} \\ &= [(1-\omega)I - \omega\tilde{A}_L](I + \omega\tilde{A}_L)^{-1}(I + \omega\tilde{A}_L^T)^{-1}[(1-\omega)I - \omega\tilde{A}_L^T] \\ &= [(1-\omega)I - \omega\tilde{A}_L](I + \omega\tilde{A}_L)^{-1}([(1-\omega)I - \omega\tilde{A}_L](I + \omega\tilde{A}_L)^{-1})^T, \end{aligned}$$

womit sogar gezeigt ist, dass alle Eigenwerte von $R_{SSOR}(\omega)$ positiv sind.

5.1.3 Optimaler Relaxationsparameter

In den meisten Fälle wird man einen für das SOR-Verfahren optimalen Relaxationsparameter nicht geschlossen angeben können. Hierbei verstehen wir unter dem optimalen Relaxationsparameter ein $\omega_{\text{opt}} \in (0, 2)$ mit

$$\rho(R_{SOR}(\omega_{\text{opt}})) = \min_{\omega\in(0,2)} \rho(R_{SOR}(\omega)).$$

I. allg. ist man also auf heuristische Methoden bei der Steuerung des Relaxationsparameyters angewiesen. Zumindest aber beim Modellproblem kann wesentlich mehr ausgesagt werden. Hierüber kurz zu berichten ist das Ziel dieses Unterabschnitts.

Definition 1.2 Eine Matrix A (mit der Zerlegung $A = A_D + A_L + A_R$ und nichtsingulärer Diagonale A_D) heißt *konsistent geordnet*, wenn die Eigenwerte von

$$C(\alpha) := -A_D^{-1}\Big(\alpha A_L + \frac{1}{\alpha} A_R\Big), \qquad \alpha \neq 0,$$

von α unabhängig sind.

Beispiel: Eine Tridiagonalmatrix (mit nicht verschwindenden Diagonalelementen) ist konsistent geordnet. Denn ist

$$C(1) = -A_D^{-1}(A_L + A_R) = \begin{pmatrix} 0 & b_1 & & & \\ a_2 & 0 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & b_{N-1} \\ & & & a_N & 0 \end{pmatrix},$$

so setze man

$$S(\alpha) := \operatorname{diag}(1, \alpha, \ldots, \alpha^{N-1}).$$

Dann ist

$$C(\alpha) = S(\alpha) C(1) S(\alpha)^{-1}.$$

Also sind $C(\alpha)$ und $C(1)$ ähnlich, haben also dieselben Eigenwerte. □

Beispiel: Auch die Blocktridiagonalmatrix A, die man bei der Diskretisierung des Modellproblems erhält, ist konsistent geordnet. Denn sei

$$A := \begin{pmatrix} A_N & -I_N & & \\ -I_N & A_N & \ddots & \\ & \ddots & \ddots & -I_N \\ & & -I_N & A_N \end{pmatrix} \quad \text{mit} \quad A_N := \begin{pmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{pmatrix}.$$

Die Matrix

$$C(\alpha) := -A_D^{-1}\Big(\alpha A_L + \frac{1}{\alpha} A_R\Big), \qquad \alpha \neq 0,$$

besitzt die von α unabhängigen Eigenwerte

$$\lambda_{pq} = \frac{1}{2}\Big(\cos\frac{p\pi}{N+1} + \cos\frac{q\pi}{N+1}\Big), \qquad 1 \leq p, q \leq n,$$

mit zugehörigen Eigenvektoren

$$u^{pq} = (u_{ij}^{pq}) := \Big(\alpha^{i+j} \sin\frac{ip\pi}{N+1} \sin\frac{jq\pi}{N+1}\Big),$$

wie man aus

$$\frac{1}{4}\Big\{\alpha u_{i-1,j}^{pq} + \alpha u_{i,j-1}^{pq} + \frac{1}{\alpha} u_{i+1,j}^{pq} + \frac{1}{\alpha} u_{i,j+1}^{pq}\Big\} = \frac{1}{2}\Big(\cos\frac{p\pi}{N+1} + \cos\frac{q\pi}{N+1}\Big) u_{ij}^{pq}$$

erkennt. □

Der folgende Satz gibt bei konsistent geordneten Matrizen eine Verbindung zwischen den Eigenwerten der Iterationsmatrix $R_{SOR}(\omega)$ des SOR-Verfahrens und denen der Iterationsmatrix R_J des Jacobi-Verfahrens.

Satz 1.3 *Sei A eine konsistent geordnete Matrix und $\omega \neq 0$. Dann gilt:*

1. *Ist λ ein Eigenwert der Iterationsmatrix $R_J := -A_D^{-1}(A_L + A_R)$ des Jacobi-Verfahren so auch $-\lambda$.*

2. *Ist $\lambda \in \mathbb{C}$ ein Eigenwert von R_J und $\mu \in \mathbb{C}$ eine Zahl mit*

$$(\mu + \omega - 1)^2 = \mu\omega^2\lambda^2,$$

   so ist μ ein Eigenwert von $R_{SOR}(\omega)$.

3. *Ist umgekehrt $\mu \neq 0$ ein Eigenwert von $R_{SOR}(\omega)$ und λ eine Zahl mit $(\mu+\omega-1)^2 = \mu\omega^2\lambda^2$, so ist λ ein Eigenwert von R_J.*

Beweis: Da A nach Voraussetzung konsistent geordnet ist, sind die Eigenwerte von

$$C(\alpha) := -A_D^{-1}\Big(\alpha A_L + \frac{1}{\alpha}A_R\Big)$$

von α unabhängig. Insbesondere stimmen die Eigenwerte von $C(1) = R_J$ und $C(-1) = -R_J$ überein, womit schon der erste Teil des Satzes bewiesen ist.

Sei $\lambda \in \mathbb{C}$ ein Eigenwert von R_J und $\mu \in \mathbb{C}$ eine Zahl mit $(\mu + \omega - 1)^2 = \mu\omega^2\lambda^2$. Hierbei können wir annehmen, dass $\mu \neq 0$, denn andernfalls ist $\omega = 1$ und $\mu = 0$ ein Eigenwert von $R_{SOR}(1) = R_{GS}$. Es ist

$$\lambda = \frac{\mu + \omega - 1}{\omega\sqrt{\mu}}$$

bei richtiger Interpretation von $\sqrt{\mu}$. Mit λ ist auch $-\lambda$ ein Eigenwert von $R_J = C(1)$, dann aber auch von $C(-\sqrt{\mu})$. Folglich ist λ ein Eigenwert von $-C(-\sqrt{\mu})$ und daher

$$\begin{aligned}
0 &= \det(\lambda I + C(-\sqrt{\mu})) \\
&= \det\Big[\lambda I + A_D^{-1}\Big(\sqrt{\mu}A_L + \frac{1}{\sqrt{\mu}}A_R\Big)\Big] \\
&= \det\Big[\frac{\mu+\omega-1}{\omega\sqrt{\mu}} I + A_D^{-1}\Big(\sqrt{\mu}A_L + \frac{1}{\sqrt{\mu}}A_R\Big)\Big] \\
&= (\omega\sqrt{\mu})^n \det((\mu+\omega-1)I + A_D^{-1}(\omega\mu A_L + \omega A_R)) \\
&= (\omega\sqrt{\mu})^n \det\{(I + \omega A_D^{-1}A_L)[\mu I - (I + \omega A_D^{-1}A_L)^{-1}((1-\omega)I - \omega A_D^{-1}A_R)]\} \\
&= (\omega\sqrt{\mu})^n \det(I + \omega A_D^{-1}A_L)\det(\mu I - (A_D + \omega A_L)^{-1}((1-\omega)I - \omega A_R)) \\
&= \underbrace{(\omega\sqrt{\mu})^n \det(I + \omega A_D^{-1}A_L))}_{\neq 0} \det(\mu I - R_{SOR}(\omega))
\end{aligned}$$

und daher μ ein Eigenwert von $R_{SOR}(\omega)$. Die Umkehrung erhält man, indem man die obige Gleichungskette von hinten nach vorne liest. Damit ist der Satz bewiesen. □

Im folgenden Satz wird in einem speziellen Fall, der aber auf jeden Fall das im Modellproblem auftretende lineare Gleichungssystem enthält, ein optimaler Relaxationsparameter angegeben. Auch in neueren Büchern über Iterationsverfahren bei linearen Gleichungssystemen ist dieser Satz zu finden, z. B. bei A. Greenbaum (1997, S. 152).

Satz 1.4 *Sei A eine konsistent geordnete Matrix. Die Eigenwerte der Iterationsmatrix R_J des Jacobi-Verfahrens seien reell und $\rho(R_J) < 1$. Dann gilt:*

1. *Mit*
$$\omega_{\text{opt}} := \frac{2}{1+\sqrt{1-\rho(R_J)^2}}$$
ist
$$\rho(R_{SOR}(\omega)) = \begin{cases} \left(\frac{\omega\rho(R_J)}{2} + \sqrt{\frac{\omega^2\rho(R_J)^2}{4} + 1 - \omega}\right)^2, & \omega \in (0, \omega_{\text{opt}}), \\ \omega - 1, & \omega \in [\omega_{\text{opt}}, 2). \end{cases}$$

2. *Das SOR-Verfahren konvergiert für alle $\omega \in (0, 2)$.*

3. *Der Spektralradius der Iterationsmatrix $R_{SOR}(\omega)$ wird für den Relaxationsparameter ω_{opt} minimal. Insbesondere ist*
$$\rho(R_{SOR}(\omega_{\text{opt}})) = \frac{1-\sqrt{1-\rho(R_J)^2}}{1+\sqrt{1-\rho(R_J)^2}}.$$

Beweis: Sei $\omega \in (0, 2)$ und λ ein (nach Voraussetzung reeller und in $(-1, 1)$ enthaltener) Eigenwert von R_J. Wegen Satz 1.3 sind
$$\mu_{1,2}(\omega) := \left(\frac{\omega\lambda}{2} \pm \sqrt{\frac{\omega^2\lambda^2}{4} + 1 - \omega}\right)^2$$
jeweils Eigenwerte von $R_{SOR}(\omega)$, ferner kann jeder von Null verschiedene Eigenwert von $R_{SOR}(\omega)$ in dieser Weise dargestellt werden. Das Argument der Wurzel ist negativ, wenn
$$\omega_0(\lambda) := \frac{2}{1+\sqrt{1-\lambda^2}} < \omega < 2.$$
In diesem Falle sind $\mu_{1,2}(\omega)$ komplex und es ist
$$|\mu_{1,2}(\omega)| = \frac{\omega^2\lambda^2}{4} + \omega - 1 - \frac{\omega^2\lambda^2}{4} = \omega - 1.$$
Ist dagegen $\omega \in (0, \omega_0(\lambda)]$, so sind $\mu_{1,2}(\omega)$ reell und nichtnegativ, die größere dieser beiden Zahlen ist
$$\mu(\omega) := \left(\frac{\omega|\lambda|}{2} + \sqrt{\frac{\omega^2\lambda^2}{4} + 1 - \omega}\right)^2,$$
ferner ist $\mu(\omega) \geq \omega - 1$. Die Behauptungen folgen dann nach elementarer Argumentation. □

Beispiel: Sei
$$A := \begin{pmatrix} A_N & -I_N & & \\ -I_N & \ddots & \ddots & \\ & \ddots & \ddots & -I_N \\ & & -I_N & A_N \end{pmatrix} \in \mathbb{R}^{N^2 \times N^2},$$

wobei

$$A_N := \begin{pmatrix} 4 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{pmatrix} \in \mathbb{R}^{N\times N}.$$

Wir wissen schon, dass A konsistent geordnet ist. Die Eigenwerte von R_J sind

$$\lambda_{pq} = \frac{1}{2}\Big(\cos\frac{p\pi}{N+1} + \cos\frac{q\pi}{N+1}\Big), \qquad 1 \le p, q \le n,$$

sie sind insbesondere reell, ferner ist

$$\rho(R_J) = \cos\frac{\pi}{N+1}.$$

Der beste Relaxationsparameter ist also

$$\omega_{\text{opt}} = \frac{2}{1+\sin\pi/(N+1)}.$$

Insbesondere ist

$$\rho(R_{SOR}(\omega_{\text{opt}})) = \frac{1-\sin\pi/(N+1)}{1+\sin\pi/(N+1)} \approx \frac{1-\pi/(N+1)}{1+\pi/(N+1)} \approx 1 - \frac{2\pi}{N+1}.$$

Dies ist natürlich verglichen mit

$$\rho(R_J) = \cos\frac{\pi}{N+1} \approx 1 - \frac{\pi^2}{2(N+1)^2}, \qquad \rho(R_{GS}) = \cos^2\frac{\pi}{N+1} \approx 1 - \frac{\pi^2}{(N+1)^2}$$

eine wesentliche Verbesserung. □

5.1.4 MATLAB-Ergänzungen

Die Koeffizientenmatrix $A = T_{N\times N}$ zum Modellproblem können wir durch die folgende Funktion generieren.

```
function A=Modell(N);
%*********************************************************************
% A is the N^2-by-N^2 coefficient matrix corresponding to the
% model problem
%*********************************************************************
I=eye(N);
S=4*eye(N)+diag(-ones(N-1,1),1)+diag(-ones(N-1,1),-1);
A=zeros(N^2,N^2);
for i=1:N
   A((i-1)*N+1:i*N,(i-1)*N+1:i*N)=S;
end;
for i=1:N-1
   A((i-1)*N+1:i*N,i*N+1:(i+1)*N)=-I;
   A(i*N+1:(i+1)*N,(i-1)*N+1:i*N)=-I;
end;
```

Wollen wir ganz konventionell die Aufgabe $-\Delta u = 1$ im Einheitsquadrat mit homogenen Randbedingungen lösen und $U = (U_{ij})$ und $-U$ über dem Einheitsquadrat darstellen, so könnte dies mit

```
tic
N=40;
A=Modell(N);
b=ones(N^2,1);
u=(1/N)^2*(A\b);
for i=1:N
    for j=1:N
        U(i,j)=u((i-1)*N+j);
    end;
end;
x=(1/(N+1))*(1:N);
y=(1/(N+1))*(1:N);
subplot(221);surf(x,y,U); subplot(222);surf(x,y,-U);
subplot(223);surfc(x,y,U); subplot(224);surfc(x,y,-U);
toc
```

geschehen. Hier ist immerhin ein lineares Gleichungssystem mit 1600 Unbekannten zu lösen. Einschließlich der Plots benötigt MATLAB hierfür nur 17.87 sec (wobei diese Zahl nicht konstant ist und von der Auslastung der Maschine abhängt). Man erhält die in 5.1 dargestellte Abbildung.

In den "Templates for the solution of Linear Systems: Building Blocks for Iterative Methods" (leicht über das Internet erreichbar) findet man u. a. MATLAB-Funktionen für Jacobi und SOR. Wir geben diese nun nur leicht abgeändert an.

```
function [x,error,iter,flag]=Jacobi(A,x,b,max_it,tol)
%*********************************************************************
%  -- Iterative template routine --
%     Univ. of Tennessee and Oak Ridge National Laboratory
%     October 1, 1993
%     Details of this algorithm are described in "Templates for the
%     Solution of Linear Systems: Building Blocks for Iterative
%     Methods", Barrett, Berry, Chan, Demmel, Donato, Dongarra,
%     Eijkhout, Pozo, Romine, and van der Vorst, SIAM Publications,
%     1993. (ftp netlib2.cs.utk.edu; cd linalg; get templates.ps).
%
% [x, error, iter, flag]  = jacobi(A, x, b, max_it, tol)
%
% Jacobi.m solves the linear system Ax=b using the Jacobi Method.
%
% input   A        REAL matrix
%         x        REAL initial guess vector
%         b        REAL right hand side vector
```

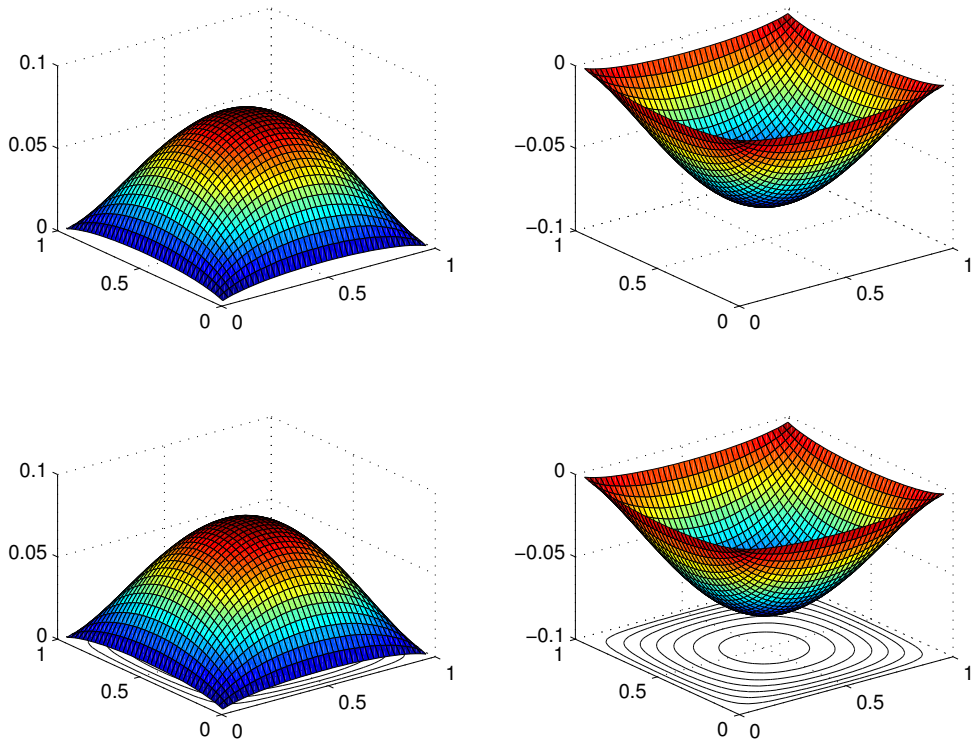


Abbildung 5.1: Die Lösung von $-\Delta u = 1$ mit homogenen Randdaten

```
%          max_it   INTEGER maximum number of iterations
%          tol      REAL error tolerance
%
% output   x        REAL solution vector
%          error    REAL error norm
%          iter     INTEGER number of iterations performed
%          flag     INTEGER: 0 = solution found to tolerance
%                            1 = no convergence given max_it
%*********************************************************************
  iter = 0;flag = 0;

  bnrm2=norm( b );
  if  (bnrm2 == 0.0), bnrm2 = 1.0; end

  r = b - A*x;
  error = norm(r)/bnrm2;
  if ( error < tol ) return, end

  [m,n]=size(A);
  M=diag(diag(A));N=M-A;

  for iter = 1:max_it
```

```
    x_1 = x;
    x   = M\(N*x + b);
    error = norm( x - x_1 ) / norm( x );
    if ( error <= tol ), break, end
  end

  if ( error > tol ) flag = 1; end

% END Jacobi.m
```

Wendet man das Jacobi-Verfahren z. B. auf das Modellproblem an, so erhält man z. B. nach

```
A=Modell(10);b=(1/10)^2*ones(100,1);
max_it=1000;tol=0.0000001;x0=zeros(100,1);
[x,error,iter,flag]=Jacobi(A,x0,b,max_it,tol);
```

einen Abbruch nach 314 Iterationen, dies dauerte 0.1906 Sekunden. In den Templates findet man auch ein SOR-Verfahren. Dies sieht im wesentlichen folgendermaßen aus:

```
function [x,error,iter,flag]=Sor(A,x,b,w,max_it,tol)
%*************************************************************
%  -- Iterative template routine --
%     Univ. of Tennessee and Oak Ridge National Laboratory
%     October 1, 1993
%     Details of this algorithm are described in "Templates for the
%     Solution of Linear Systems: Building Blocks for Iterative
%     Methods", Barrett, Berry, Chan, Demmel, Donato, Dongarra,
%     Eijkhout, Pozo, Romine, and van der Vorst, SIAM Publications,
%     1993. (ftp netlib2.cs.utk.edu; cd linalg; get templates.ps).
%
% [x, error, iter, flag]  = Sor(A, x, b, w, max_it, tol)
%
% sor.m solves the linear system Ax=b using the
% Successive Over-Relaxation Method (Gauss-Seidel method when w =1).
%
% input   A        REAL matrix
%         x        REAL initial guess vector
%         b        REAL right hand side vector
%         w        REAL relaxation scalar
%         max_it   INTEGER maximum number of iterations
%         tol      REAL error tolerance
%
% output  x        REAL solution vector
%         error    REAL error norm
%         iter     INTEGER number of iterations performed
%         flag     INTEGER: 0 = solution found to tolerance
```

```
%                              1 = no convergence given max_it
%*********************************************************************
  flag = 0;iter = 0;

  bnrm2 = norm(b);
  r = b - A*x;
  error = norm( r ) / bnrm2;
  if ( error < tol ) return, end

  b=w*b;
  M=diag(diag(A))+w*tril(A,-1);
  N=(1.0-w)*diag(diag(A))-w*triu(A,1);

  for iter = 1:max_it

     x_1 = x;
     x   = M \ ( N*x + b );                    % update approximation

     error = norm( x - x_1 ) / norm( x );      % compute error
     if ( error <= tol ), break, end           % check convergence

  end
  b = b / w;                                   % restore rhs

  if ( error > tol ) flag = 1; end;            % no convergence

% END Sor.m
```

Dasselbe Experiment wie oben liefert für Gauss-Seidel 166 Iterationen in 0.1084 Sekunden. Für das Modellproblem mit $N = 30$ (also 900 Gleichungen mit ebenso vielen Unbekannten) erhält man mit dem optimalen Relaxationsparameter einen Abbruch nach 91 Iterationen, wofür 10.8286 Sekunden benötigt werden, für Gauss-Seidel werden 1123 Iterationen in 129.5965 Sekunden benötigt.

5.1.5 Aufgaben

1. Man berechne die Eigenwerte und Eigenvektoren der Matrix

$$S_N := \begin{pmatrix} \alpha & \beta & & \\ \beta & \ddots & \ddots & \\ & \ddots & \ddots & \beta \\ & & \beta & \alpha \end{pmatrix} \in \mathbb{R}^{N\times N}.$$

2. Sei $A \in \mathbb{C}^{n\times n}$ (strikt) zeilenweise dominant, d. h.

$$\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| < |a_{ii}|, \qquad i = 1, \ldots, n.$$

Dann sind das Jacobi- und das Gauß-Seidel-Verfahren für jeden Startvektor $x^{(0)} \in \mathbb{C}^n$ konvergent.

3. Die Matrix $A \in \mathbb{C}^{n\times n}$ heißt bekanntlich *zerlegbar* (oder auch *zerfallend*, *reduzibel*), wenn es nichtleere Teilmengen N_1, N_2 von $N := \{1, \ldots, n\}$ mit $N_1 \cap N_2 = \emptyset$, $N_1 \cup N_2 = N$ sowie $a_{ij} = 0$ für alle $(i, j) \in N_1 \times N_2$ gibt, andernfalls *unzerlegbar* (oder auch *nichtzerfallend, irreduzibel*). Weiter heißt A *schwach zeilenweise diagonal dominant* (oder *genügt dem schwachen Zeilensummenkriterium*), wenn

$$\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \begin{cases} \leq |a_{ii}| & \text{für } i = 1, \ldots, n, \\ < |a_{ii}| & \text{für mindestens ein } i = i_0. \end{cases}$$

Man zeige: Ist $A \in \mathbb{C}^{n\times n}$ schwach zeilenweise diagonal dominant und unzerlegbar, so sind alle Diagonalelemente von A von Null verschieden. Ferner sind das Jacobi- und das Gauß-Seidel-Verfahren für jeden Startvektor konvergent, die Matrix A nichtsingulär.

4. Die Diagonalelemente von $A \in \mathbb{C}^{n\times n}$ seien ungleich Null. Dann ist $\rho(R_{SOR}(\omega)) \geq |\omega - 1|$, so dass das SOR-Verfahren (für jeden Startwert) nur konvergieren kann, wenn $\omega \in (0, 2)$.

5. Sei $A \in \mathbb{C}^{n\times n}$ hermitesch und positiv definit. Dann konvergiert das SOR-Verfahren für alle $\omega \in (0, 2)$, speziell also das Gauß-Seidel-Verfahren.

6. Sei

$$T_{N\times N} := \begin{pmatrix} A_N & -I_N & & \\ -I_N & \ddots & \ddots & \\ & \ddots & \ddots & -I_N \\ & & -I_N & A_N \end{pmatrix} \in \mathbb{R}^{N^2\times N^2},$$

wobei

$$A_N := \begin{pmatrix} 4 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{pmatrix} \in \mathbb{R}^{N\times N}.$$

Man zeige, dass $T_{N\times N}$ schwach zeilenweise diagonal dominant, unzerlegbar und (symmetrisch und) positiv definit ist, so dass wegen der Aussagen in den Aufgaben 3 und 5 das Jacobi- und das Gauß-Seidel-Verfahren sowie das SOR-Verfahren für alle $\omega \in (0, 2)$ konvergent sind. Weiter zeige man, dass $T_{N\times N}^{-1}$ eine nichtnegative Matrix ist.

7. Sei A eine Blocktridiagonalmatrix der Form

$$A = \begin{pmatrix} D_1 & H_1 & & & & \\ K_1 & D_2 & H_2 & & & \\ & K_2 & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & D_{M-1} & H_{M-1} \\ & & & & K_{M-1} & D_M \end{pmatrix},$$

wobei D_i, $i = 1, \ldots, M$, quadratische, nichtsinguläre Diagonalmatrizen sind. Man zeige, dass A konsistent geordnet ist.

8. Ist $A \in \mathbb{R}^{n\times n}$ konsistent geordnet und bezeichnen R_J bzw. R_{GS} die Iterationsmatrizen des Jacobi- bzw. Gauß-Seidel-Verfahrens, so ist $\rho(R_{GS}) = \rho(R_J)^2$.

9. Seien $A, M, N \in \mathbb{R}^{n\times n}$ mit $A = M - N$. Das Paar (M, N) heißt eine *reguläre Zerlegung* von A, wenn M nichtsingulär ist und M^{-1} und N nichtnegative Matrizen sind. Man zeige: Ist (M, N) eine reguläre Zerlegung von A, so ist $\rho(M^{-1}N) < 1$ genau dann, wenn A nichtsingulär und A^{-1} nichtnegativ ist[2].

   Hinweis: Hierbei kann man benutzen, dass der Spektralradius einer nichtnegativen Matrix ein Eigenwert ist, zu dem es einen nichtnegativen Eigenvektor gibt (Perron-Frobenius).

10. Gegeben sei das lineare Gleichungssystem

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix}.$$

    Hierbei sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, $B \in \mathbb{R}^{n\times m}$ habe den Rang m, ferner seien $b \in \mathbb{R}^n$ und $c \in \mathbb{R}^m$. Man zeige:

    (a) Das obige lineare Gleichungssystem ist eindeutig lösbar.

    (b) Die Matrix $S := B^T A^{-1} B$ ist positiv definit.

    (c) Mit einem Relaxationsparameter $\omega \neq 0$ betrachte man zur numerischen Lösung des obigen linearen Gleichungssystems das folgende Iterationsverfahren (Uzawa-Verfahren):

    - Wähle $(x_0, y_0) \in \mathbb{R}^n \times \mathbb{R}^m$.
    - Für $k = 0, 1, \ldots$:
      - $x_{k+1} := A^{-1}(b - By_k)$.
      - $y_{k+1} := y_k + \omega(B^T x_{k+1} - c)$.

    Dieses konvergiert genau dann, wenn $\omega \in (0, 2/\lambda_{\min}(S))$, ferner ist der optimale Relaxationsparameter gegeben durch

$$\omega_{\text{opt}} := \frac{2}{\lambda_{\min}(S) + \lambda_{\max}(S)}.$$

[2]Siehe Y. Saad (1996, S. 107).

5.2 Krylov-Verfahren

An Literatur nennen wir hier vor allem die schon in der Einführung angegebenen neueren Lehrbücher von Y. Saad (1996) und A. Greenbaum (1997), aber auch das außerordentlich anregende Buch von G. W. Stewart (1998)[3].

Grundproblem ist wieder das lineare Gleichungssystem $Ax = b$ mit nichtsingulärem $A \in \mathbb{R}^{n\times n}$ und $b \in \mathbb{R}^n$. I. allg. wird nur angenommen, dass ein Matrix-Vektor Produkt Ax bzw. eventuell (bei nichtsymmetrischem A) $A^T y$ "einfach" zu berechnen ist. Daher können schwach besetzte, hochdimensionale Probleme mit den vorzustellenden Verfahren behandelt werden. Wir werden vor allem auf die in MATLAB implementierten Verfahren eingehen.

5.2.1 Krylov-Teilräume

Bei gegebenen $A \in \mathbb{R}^{n\times n}$ und $v \in \mathbb{R}^n \setminus \{0\}$ heißt

$$\mathcal{K}_k(A, v) := \text{span}\,\{v, Av, \ldots, A^{k-1}v\}$$

ein *Krylov-Teilraum*. Offenbar ist

$$\mathcal{K}_k(A, v) = \{p(A)v : p \in \Pi_{k-1}\}.$$

Wir werden im folgenden häufig $\mathcal{K}_k$ statt $\mathcal{K}_k(A, v)$ schreiben, wenn dadurch keine Mißverständnisse auftreten können.

Beispiel: Ist x_0 eine Näherungslösung des linearen Gleichungssystems $Ax = b$ mit nichtsingulärer Koeffizientenmatrix $A \in \mathbb{R}^{n\times n}$, so definieren wir $r_0 := b - Ax_0$ und können davon ausgehen, dass $r_0 \neq 0$. Die Idee der meisten Krylov-Verfahren ist es, im affinen Teilraum $x_0 + \mathcal{K}_k(A, r_0)$ nach einer geeigneten Näherungslösung x_k zu suchen. Wenn dies gelungen ist, wird entweder ein Restart gemacht, also $x_0 := x_k$, $r_0 := b - Ax_0$ gesetzt und das gleiche Verfahren noch einmal angewandt, oder es wird k erhöht. Zur Bestimmung einer geeigneten Näherungslösung in $x_0 + \mathcal{K}_k(A, r_0)$ sind vor allem zwei Prinzipien wichtig. Im ersten wird als Zusatzforderung gestellt, dass der Defekt (oft nennt man diesen auch das Residuum) auf $\mathcal{K}_k(A, r_0)$ oder einem anderen Krylov-Teilraum senkrecht steht (man nennt dies eine Galerkin-Bedingung) oder der Defekt der euklidischen Norm nach auf $\mathcal{K}_k(A, r_0)$ oder einem anderen Krylov-Raum minimiert wird. □

Es ist einleuchtend, dass die Dimension eines Krylov-Teilraumes von Bedeutung ist. Wir definieren:

$$\text{grad}(v) := \min\{j : \text{Es existiert } p \in \Pi_j \setminus \{0\} \text{ mit } p(A)v = 0\}.$$

Der *Satz von Caley-Hamilton* sagt aus, dass A eine Wurzel des eigenen charakteristischen Polynoms ist. Ist also $p(\lambda) := \det(A - \lambda I)$ das charakteristische Polynom zu A (offenbar ein Polynom vom Grad n), so ist $p(A) = 0$. Daher ist $\text{grad}(v) \leq n$ für jedes $v \in \mathbb{R}^n$. In dem folgenden Satz wird unter anderem die Dimension eines Krylov-Teilraumes angegeben.

[3] G. W. Stewart (1998) *Afternotes goes to Graduate School. Lectures on Advanced Numerical Analysis.* SIAM, Philadelphia.

Satz 2.1 *Seien $A \in \mathbb{R}^{n\times n}$ und $v \in \mathbb{R}^n$ gegeben, ferner sei $\mu := \operatorname{grad}(v)$. Dann gilt:*

1. *Der Krylov-Teilraum $\mathcal{K}_\mu$ ist invariant unter A, d. h. es ist $A(\mathcal{K}_\mu) \subset \mathcal{K}_\mu$, ferner ist $\mathcal{K}_k = \mathcal{K}_\mu$ für alle $k \geq \mu$.*
2. *Es ist $\dim(\mathcal{K}_k) = k$ genau dann, wenn $\mu \geq k$.*
3. *Es ist $\dim(\mathcal{K}_k) = \min(k, \mu)$.*

Beweis: Sei $x \in \mathcal{K}_\mu$, also

$$x = \sum_{i=0}^{\mu-1} \alpha_i A^i v.$$

Weiter existieren $\beta_0, \ldots, \beta_\mu$, welche nicht alle verschwinden, mit

$$0 = \sum_{i=0}^{\mu} \beta_i A^i v.$$

Hierbei ist $\beta_\mu \neq 0$, denn andernfalls wäre $\operatorname{grad}(v) \leq \mu - 1$. Dann ist aber

$$Ax = \sum_{i=1}^{\mu} \alpha_{i-1} A^i v - \frac{\alpha_{\mu-1}}{\beta_\mu} \underbrace{\sum_{i=0}^{\mu} \beta_i A^i v}_{=0} = -\frac{\alpha_{\mu-1}}{\beta_\mu}\beta_0 v + \sum_{i=1}^{\mu-1} \Big(\alpha_{i-1} - \frac{\alpha_{\mu-1}}{\beta_\mu}\beta_i\Big) A^i v \in \mathcal{K}_\mu.$$

Also ist $A(\mathcal{K}_\mu) \subset \mathcal{K}_\mu$ bzw. $\mathcal{K}_\mu$ invariant unter A. Ist $k \geq \mu$, so ist selbstverständlich $\mathcal{K}_\mu \subset \mathcal{K}_k$. Sei daher jetzt $k > \mu$ und $x \in \mathcal{K}_k$. Dann ist

$$x = \sum_{i=0}^{k-1} \alpha_i A^i v,$$

weiter existieren wieder $\beta_0, \ldots, \beta_\mu$ mit $\beta_\mu \neq 0$ und

$$0 = \sum_{i=0}^{\mu} \beta_i A^i v.$$

Dann ist

$$x = \sum_{i=0}^{k-1} \alpha_i A^i v - \frac{\alpha_{k-1}}{\beta_\mu} \underbrace{\sum_{i=0}^{\mu} \beta_i A^{i+k-1-\mu} v}_{=0} \in \mathcal{K}_{k-1}.$$

In dieser Weise kann fortgefahren werden bis $x \in \mathcal{K}_\mu$ erhalten ist. Damit ist der erste Teil des Satzes bewiesen.

Die Vektoren $\{v, Av, \ldots, A^{k-1}v\}$ bilden genau dann eine Basis von $\mathcal{K}_k$, wenn für jede Menge $\{\alpha_0, \ldots, \alpha_{k-1}\}$ von k Skalaren, wobei wenigstens ein α_i von Null verschieden ist, $\sum_{i=0}^{k-1} \alpha_i A^i v \neq 0$. Dies ist äquivalent der Bedingung, dass das einzige Polynom $p \in \Pi_{k-1}$ mit $p(A)v = 0$ das Nullpolynom ist, was wiederum äquivalent zu $\mu = \operatorname{grad}(v) \geq k$ ist.

Wegen des zweiten Teiles des Satzes ist

$$\dim(\mathcal{K}_k) = k = \min(k, \mu),$$

wenn $\mu \geq k$. Ist dagegen $\mu < k$, so ist wegen des ersten Teiles $\mathcal{K}_k = \mathcal{K}_\mu$ und folglich

$$\dim(\mathcal{K}_k) = \dim(\mathcal{K}_\mu) = \mu.$$

Damit ist der einfache Satz bewiesen. □

5.2.2 Das Arnoldi-Verfahren zur Bestimmung einer Orthonormalbasis eines Krylov-Teilraumes

Wir betrachten den Krylov-Teilraum

$$\mathcal{K}_k := \operatorname{span}\{v, Av, \ldots, A^{k-1}v\},$$

wobei typischerweise $k \ll n$, und stellen uns die Aufgabe, hierzu eine Orthonormalbasis zu berechnen. Hierbei denkt man natürlich an das Gram-Schmidt-Verfahren bzw. das bei exakter Arithmetik äquivalente, aber numerisch günstigere, modifizierte Gram-Schmidt-Verfahren. Genau dieses führt auf das *Arnoldi-Verfahren* und sieht folgendermaßen aus:

- Input: Gegeben $A \in \mathbb{R}^{n\times n}$ und $v \in \mathbb{R}^n \setminus \{0\}$, ferner $k \in \mathbb{N}$.
- Berechne $q_1 := v/\|v\|_2$.
- Für $j = 1, \ldots, k$:
  - $w := Aq_j$.
  - Für $i = 1, \ldots, j$:
    - $h_{ij} := q_i^T w$.
    - $w := w - h_{ij} q_i$.
  - $h_{j+1,j} := \|w\|_2$.
  - Falls $h_{j+1,j} = 0$, dann: STOP.
  - $q_{j+1} := w/h_{j+1,j}$.
- Output: Falls kein vorzeitiger Abbruch erfolgt, werden eine Matrix

$$Q_k := \begin{pmatrix} q_1 & \cdots & q_k \end{pmatrix} \in \mathbb{R}^{n\times k}$$

und eine Matrix

$$\tilde{H}_k := \begin{pmatrix} h_{11} & h_{12} & \cdots & \cdots & h_{1k} \\ h_{21} & h_{22} & \ddots & & h_{2k} \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & h_{k-1,k} \\ & & & h_{k,k-1} & h_{kk} \\ & & & & h_{k+1,k} \end{pmatrix} \in \mathbb{R}^{(k+1)\times k}$$

ausgegeben. Schließlich wird auch noch ein Vektor $q_{k+1} \in \mathbb{R}^n$ berechnet, so dass auch $Q_{k+1} := (\ Q_k \quad q_{k+1}\ )$ definiert ist. Eigenschaften dieser Matrizen werden in dem folgenden Satz formuliert und anschließend bewiesen.

Den nächsten Satz findet man im wesentlichen bei Y. Saad (1996, Proposition 6.4, 6.5), wobei wir aber gleich den für praktische Zwecke wesentlich günstigeren Algorithm 6.2 (S. 149) (MGS-Arnoldi) angegeben haben, siehe auch J. W. Demmel (1997, S. 303), L. N. Trefethen, D. Bau (1997, S. 252) und A. Greenbaum (1997, S. 38).

Satz 2.2 *Das obige Arnoldi-Verfahren breche nicht vorzeitig ab und liefere die Matrizen $Q_k \in \mathbb{R}^{n\times k}$, $Q_{k+1} = (\ Q_k \quad q_{k+1}\ ) \in \mathbb{R}^{n\times(k+1)}$ und $\tilde{H}_k \in \mathbb{R}^{(k+1)\times k}$. Die Matrix $H_k \in \mathbb{R}^{k\times k}$ bezeichne die Matrix, die man aus $\tilde{H}_k$ durch Weglassen der letzten Zeile erhält. Dann gilt:*

1. *Die Spalten $q_1, \ldots, q_k$ von Q_k bilden eine Orthonormalbasis des Krylov-Teilraumes*
$$\mathcal{K}_k := \operatorname{span}\{v, Av, \ldots, A^{k-1}v\}.$$
2. *Es ist $AQ_k = Q_{k+1}\tilde{H}_k$*
3. *Es ist $Q_k^T A Q_k = H_k$.*

Beweis: Durch vollständige Induktion nach j zeigen wir, dass

$$\{q_1, \ldots, q_j\}, \qquad j = 1, \ldots, k+1,$$

ein Orthonormalsystem ist. Der Induktionsanfang bei $j = 1$ ist wegen $q_1 := v/\|v\|_2$ trivial. Wir nehmen an, $\{q_1, \ldots, q_j\}$ sei ein Orthonormalsystem. Da $\|q_{j+1}\|_2 = 1$ offensichtlich ist, bleibt zu zeigen, dass $q_l^T q_{j+1} = 0$, $l = 1, \ldots, j$. Um dies einzusehen, schreiben wir die Berechnung von q_{j+1} in der folgenden Form:

- $w^{(0)} := Aq_j$.
- Für $i = 1, \ldots, j$:
  - $w^{(i)} := w^{(i-1)} - q_i^T w^{(i-1)} q_i$.
- $q_{j+1} := w^{(j)}/\|w^{(j)}\|_2$.

Für $l = 1, \ldots, j$ ist dann

$$q_l^T w^{(j)} = q_l^T w^{(j-1)} - q_j^T w^{(j-1)} q_l^T q_j = q_l^T w^{(j-1)} - q_l^T w^{(j-1)} \delta_{lj}.$$

Hierbei haben wir $q_l^T q_j = \delta_{lj}$ bzw. die Induktionsvoraussetzung ausgenutzt. Folglich ist $q_l^T w^{(j)} = 0$ für $l = j$, andernfalls, also für $l < j$, ist $q_l^T w^{(j)} = q_l^T w^{(j-1)}$. Wegen

$$q_l^T w^{(j-1)} = q_l^T w^{(j-2)} - q_{j-1}^T w^{(j-2)} q_l^T q_{j-1} = q_l^T w_j^{(j-2)} - q_l^T w^{(j-2)} \delta_{l,j-1}$$

kann hier genauso geschlossen werden. So kann fortgefahren werden und man erhält, dass $w^{(j)}$ und damit auch q_{j+1} auf $q_1, \ldots, q_j$ senkrecht steht. Daher ist $\{q_1, \ldots, q_{j+1}\}$ ein Orthonormalsystem.

Nun zeigen wir, dass $\mathcal{K}_k = \text{span}\{q_1, \ldots, q_k\}$. Hierzu zeigen wir durch vollständige Induktion nach j, dass $q_j = p_{j-1}(A)v$ mit einem Polynom $p_{j-1} \in \Pi_{j-1}$. Es ist $q_1 = v/\|v\|_2$, die Aussage für $j = 1$ ist also richtig mit $p_0(t) := 1/\|v\|_2$. Für den Induktionsschritt beachten wir, dass

$$q_{j+1} = \frac{w}{\|w\|_2} = \frac{1}{\|w\|_2}\Big(Aq_j - \sum_{i=1}^{j} h_{ij}q_i\Big) = \frac{1}{\|w\|_2}\Big(Ap_{j-1}(A)v - \sum_{i=1}^{j} h_{ij}p_{i-1}(A)v\Big) = p_j(A)v,$$

wenn man $p_j \in \Pi_j$ durch

$$p_j(t) := \frac{1}{\|w\|_2}\Big(tp_{j-1}(t) - \sum_{i=1}^{j} h_{ij}p_{i-1}(t)\Big)$$

definiert. Der Induktionsschritt ist abgeschlossen. Hieraus folgt $\text{span}\{q_1, \ldots, q_k\} \subset \mathcal{K}_k$ und dann auch, dass $\{q_1, \ldots, q_k\}$ eine Orthonormalbasis von $\mathcal{K}_k$ ist.

Es ist

$$AQ_k e_j = Aq_j = \sum_{i=1}^{j+1} h_{ij}q_i = Q_{k+1}\tilde{H}_k e_j, \qquad j = 1, \ldots, k,$$

und daher $AQ_k = Q_{k+1}\tilde{H}_k$. Hieraus folgt

$$Q_k^T AQ_k = Q_k^T Q_{k+1}\tilde{H}_k.$$

Wir überlegen uns, dass $Q_k^T Q_{k+1}\tilde{H}_k = H_k$, womit dann auch die letzte Aussage in dem Satz bewiesen sein wird. Denn es ist

$$\begin{aligned} Q_k^T Q_{k+1}\tilde{H}_k &= Q_k^T \begin{pmatrix} Q_k & q_{k+1} \end{pmatrix} \begin{pmatrix} H_k \\ h_{k+1,k}e_k^T \end{pmatrix} \\ &= \begin{pmatrix} I & 0 \end{pmatrix} \begin{pmatrix} H_k \\ h_{k+1,k}e_k^T \end{pmatrix} \\ &= H_k \end{aligned}$$

und damit, wie behauptet, $Q_k^T AQ_k = H_k$. □

Im folgenden Satz werden notwendige und hinreichende Bedingungen dafür angegeben, dass das Arnoldi-Verfahren im j-ten Schritt abbricht.

Satz 2.3 *Das Arnoldi-Verfahren stoppt im j-ten Schritt genau dann, wenn* $\text{grad}(v) = j$. *Insbesondere ist dann $\mathcal{K}_j$ ein unter A invarianter Teilraum.*

Beweis: Sei $\text{grad}(v) = j \le m$. Wegen Satz 2.1 ist $\dim(\mathcal{K}_j) = j$, das Verfahren kann also insbesondere nicht vor dem j-ten Schritt abgebrochen sein. Im j-ten Schritt bricht das Verfahren aber ab, denn andernfalls könnte der normierte, und auf $q_1, \ldots, q_j$ senkrecht stehende Vektor q_{j+1} berechnet werden, es wäre insbesondere $\dim(\mathcal{K}_{j+1}) = j + 1$, ein Widerspruch zum ersten Teil von Satz 2.1. Umgekehrt nehmen wir an, das Arnoldi-Verfahren breche im j-ten Schritt ab. Dann ist $\text{grad}(v) \le j$ nach Definition des Grades. Ferner ist $\text{grad}(v) = j$, denn andernfalls hätte das Verfahren schon in einem früheren Schritt gestoppt. □

Auf ein weiteres Verfahren zur Berechnung einer Orthonormalbasis eines Krylov-Teilraumes wird in Aufgabe 2 eingegangen.

5.2.3 Das Arnoldi-Verfahren bei linearen Gleichungssystemen (FOM)

Gegeben sei das lineare Gleichungssystem $Ax = b$, wobei die nichtsinguläre Matrix $A \in \mathbb{R}^{n\times n}$ und $b \in \mathbb{R}^n$ gegeben sind. Mit einer Anfangsnäherung $x_0 \in \mathbb{R}^n$ berechnen wir den Defekt $r_0 := b - Ax_0$ und definieren den zu r_0 gehörenden Krylov-Teilraum

$$\mathcal{K}_k(A, r_0) := \operatorname{span}\{r_0, Ar_0, \ldots, A^{k-1}r_0\}.$$

Zur Abkürzung setzen wir wieder $\mathcal{K}_k := \mathcal{K}_k(A, r_0)$. Zunächst formulieren und beweisen wir ein kleines Lemma, das Grundlage des Arnoldi-Verfahrens bei linearen Gleichungssystemen ist. Bei diesem wird nämlich $x_k \in x_0 + \mathcal{K}_k$ so bestimmt, dass $b - Ax_k \perp \mathcal{K}_k$.

Lemma 2.4 *Es sei* $\dim(\mathcal{K}_k) = k$*, ferner sei (etwa mit MGS-Arnoldi) eine Matrix* $Q_k = (\ q_1 \ \cdots \ q_k\ ) \in \mathbb{R}^{n\times k}$ *und eine unreduzierte obere Hessenberg-Matrix* $H_k \in \mathbb{R}^{k\times k}$ *mit*

$$Q_k^T Q_k = I, \qquad \mathcal{K}_k = \operatorname{span}\{q_1, \ldots, q_k\}, \qquad Q_k^T A Q_k = H_k$$

berechnet, wobei $q_1 = r_0/\|r_0\|_2$*. Zusätzlich sei* H_k *nichtsingulär*[4]*. Definiert man dann*

$$x_k := x_0 + Q_k H_k^{-1}(\|r_0\|_2 e_1),$$

so ist

$$x_k \in x_0 + \mathcal{K}_k, \qquad b - Ax_k \perp \mathcal{K}_k.$$

Beweis: Da die Spalten von Q_k eine Basis von $\mathcal{K}_k$ bilden, ist offensichtlich $x_k \in x_0 + \mathcal{K}_k$. Wegen

$$\mathcal{K}_k = R(Q_k) = N(Q_k^T)^\perp$$

ist $b - Ax_k \perp \mathcal{K}_k$ genau dann, wenn $Q_k^T(b - Ax_k) = 0$. Nun ist

$$\begin{aligned} Q_k^T(b - Ax_k) &= Q_k^T r_0 - \underbrace{Q_k^T A Q_k H_k^{-1}}_{=I}(\|r_0\|_2 e_1) \\ &= Q_k^T r_0 - \|r_0\|_2\, e_1 \\ &= Q_k^T[r_0 - \|r_0\|_2\, Q_k e_1] \\ &= Q_k^T[r_0 - \|r_0\|_2\, q_1] \\ &= 0. \end{aligned}$$

Damit ist das Lemma bewiesen. □

Eine Kombination mit MGS-Arnoldi liefert das folgende Verfahren (FOM, **F**ull **O**rthogonalization **M**ethod) zur Berechnung einer Näherungslösung x_k des linearen Gleichungssystems $Ax = b$.

- Mit einer Näherungslösung x_0 berechne man den Defekt $r_0 := b - Ax_0$. Berechne $q_1 := r_0/\|r_0\|_2$. Setze $H_k = (h_{ij})_{1\le i,j\le k} := 0$.

[4] Bei Y. Saad (1996, S. 153) wird dies nicht vorausgesetzt, ist es automatisch erfüllt? Klar ist, dass die ersten $k-1$ Spalten der unreduzierten oberen Hessenberg-Matrix H_k linear unabhängig sind.

- Für $j = 1, \ldots, k$:
  - $w := Aq_j$.
  - Für $i = 1, \ldots, j$:
    - $h_{ij} := q_i^T w$.
    - $w := w - h_{ij} q_i$.
  - $h_{j+1,j} := \|w\|_2$.
  - Falls $h_{j+1,j} = 0$, dann: Setze $k := j$ und gehe zum letzten $\bullet$.
  - $q_{j+1} := w / h_{j+1,j}$.
- Setze
$$Q_k = (\ q_1 \ \cdots \ q_k\ ), \qquad H_k := (h_{ij})_{1 \le i,j \le k}$$
und berechne
$$x_k := x_0 + Q_k H_k^{-1}(\|r_0\|_2 e_1).$$

Die Hauptarbeit besteht natürlich in der Lösung des (niederdimensionalen) linearen Gleichungssystems $H_k y = \|r_0\|_2 e_1$. Dieses kann effizient mit Hilfe des Gaußschen Eliminationsverfahrens oder mit Hilfe von Givens-Rotationen geschehen.

Bei Y. Saad (1996, S. 154 ff.) werden noch einige Varianten angegeben, hierauf wollen wir nicht mehr eingehen.

5.2.4 GMRES

Die Abkürzung GMRES steht für "**G**eneralized **M**inimum **Res**idual Method". Wieder gehen wir von einer Näherungslösung x_0 des linearen Gleichungssystems $Ax = b$ aus, setzen $r_0 := b - Ax_0$ und betrachten den Krylov-Teilraum $\mathcal{K}_k = \mathcal{K}_k(A, r_0)$. Als Näherung x_k sucht man diesmal ein $x_k \in x_0 + \mathcal{K}_k$, welches die Aufgabe

$$\text{Minimiere} \quad \|b - Ax\|_2, \quad x \in x_0 + \mathcal{K}_k$$

löst. Mit Hilfe einer durch MGS-Arnoldi berechneten Orthonormalbasis $\{q_1, \ldots, q_k\}$ von $\mathcal{K}_k$ bzw. der Matrix $Q_k = (\ q_1 \ \cdots \ q_k\ )$ kann diese Aufgabe auch in der Form

$$\text{Minimiere} \quad J(y) := \|b - A(x_0 + Q_k y)\|_2 = \|r_0 - AQ_k y\|_2, \quad y \in \mathbb{R}^k$$

geschrieben werden. Nach Satz 2.2 ist $AQ_k = Q_{k+1}\tilde{H}_k$ mit der im MGS-Arnoldi berechneten Matrix $\tilde{H}_k \in \mathbb{R}^{(k+1)\times k}$. Die erste Spalte q_1 von Q_k (oder auch Q_{k+1}) ist durch $r_0/\|r_0\|_2$ gegeben. Daher ist

$$r_0 - AQ_k y = Q_{k+1}[\|r_0\|_2\, e_1 - \tilde{H}_k y].$$

Da die Spalten von Q_{k+1} orthonormiert sind, ist das zu lösende lineare Least Square Problem durch

$$\text{Minimiere} \quad J(y) := \|\|r_0\|_2 e_1 - \tilde{H}_k\|_2, \quad y \in \mathbb{R}^k$$

gegeben. Ist

$$\tilde{H}_k = (h_{ij})_{\substack{1\le i\le k+1\\ 1\le j\le k}}$$

eine unreduzierte obere Hessenberg-Matrix, ist also $h_{j+1,j} \neq 0$, $j = 1, \ldots, k$, so ist Rang $(H_k) = k$ und damit das obige lineare Least Square Problem *eindeutig* lösbar. Zusammenfassend sieht das GMRES-Verfahren folgendermaßen aus (siehe Algorithm 6.9 bei Y. SAAD (1996, S. 159)):

- Mit einer Näherungslösung x_0 berechne man den Defekt $r_0 := b - Ax_0$. Berechne $q_1 := r_0/\|r_0\|_2$. Setze
$$\tilde{H}_k = (h_{ij})_{\substack{1\le i\le k+1\\ 1\le j\le k}} = 0.$$
- Für $j = 1, \ldots, k$:
  - $w := Aq_j$.
  - Für $i = 1, \ldots, j$:
    - $h_{ij} := q_i^T w$.
    - $w := w - h_{ij} q_i$.
  - $h_{j+1,j} := \|w\|_2$.
  - Falls $h_{j+1,j} = 0$, dann: Setze $k := j$ und gehe zum letzten •.
  - $q_{j+1} := w/h_{j+1,j}$.
- Berechne die Lösung y_k des linearen Ausgleichsproblems
$$\text{Minimiere} \quad J(y) := \|\|r_0\|_2 e_1 - \tilde{H}_k y\|_2, \quad y \in \mathbb{R}^k,$$
und berechne anschließend $x_k := x_0 + Q_k y_k$, wobei $Q_k := (\ q_1 \ \cdots \ q_k\ )$.

Nun stellt sich natürlich die Frage, wie das i. allg. niederdimensionale lineare Ausgleichsproblem mit der Koeffizientenmatrix $\tilde{H}_k$ gelöst werden sollte. Wegen Rang $(\tilde{H}_k) = k$ liegt es nahe, eine QR-Zerlegung von $\tilde{H}_k$ zu berechnen, wegen der speziellen Struktur von $\tilde{H}_k$ sind hierzu natürlich Givens-Rotationen besonders geeignet. Man multipliziert also $(\ \tilde{H}_k \ \ \|r_0\|_2 e_1\ )$ sukzessive mit Givens-Rotationen $G_{12}, \ldots, G_{k,k+1}$, wobei $G_{j,j+1} = G_{j,j+1}(c_j, s_j)$ die Aufgabe hat, in der aktuellen Matrix das Element in der Position $(j+1, j)$, $j = 1, \ldots, k$, zu annullieren. Nach k Schritten hat man

$$(\ \tilde{R}_k \ \ \tilde{g}_k\ ) := F_k(\ \tilde{H}_k \ \ \|r_0\|_2 e_1\ )$$

berechnet, wobei wir zur Abkürzung

$$F_k := G_{k,k+1} \cdots G_{12}$$

gesetzt haben. Sei $R_k \in \mathbb{R}^{k\times k}$ die Matrix, die man aus $\tilde{R}_k \in \mathbb{R}^{(k+1)\times k}$ durch Weglassen der letzten Zeile (eine Nullzeile!) erhält, entsprechend entstehe $g_k \in \mathbb{R}^k$ aus $\tilde{g}_k \in \mathbb{R}^{k+1}$ durch Weglassen der letzten Komponente. Ferner sei γ_i die i-te Komponente von $\tilde{g}_k$. Die gesuchte Lösung y_k ist dann durch $y_k = R_k^{-1} g_k$ gegeben, wobei R_k wegen Rang $(\tilde{H}_k) = k$ nichtsingulär ist. Dann können wir folgendes feststellen:

- Es ist
$$b - Ax_k = Q_{k+1}[\|r_0\|_2 e_1 - \tilde{H}_k y_k] = Q_{k+1} F_k^T(\gamma_{k+1} e_{k+1})$$
und folglich
$$\|b - Ax_k\|_2 = |\gamma_{k+1}|.$$

Denn: Es ist

$$b - Ax_k = Q_{k+1}[\|r_0\|_2 e_1 - \tilde{H}_k y_k] = Q_{k+1} F_k^T(\tilde{g}_k - \tilde{R}_k y_k) = Q_{k+1} F_k^T(\gamma_{k+1} e_{k+1}),$$

und folglich (die Spalten von Q_{k+1} sind orthonormiert, ferner ist F_k orthogonal)

$$\|b - Ax_k\|_2 = |\gamma_{k+1}|.$$

Die Berechnung des Vektors $\tilde{g}_k = (\gamma_i)_{1 \le i \le k+1}$ ist außerordentlich einfach. Hierzu beachten wir, dass

$$\tilde{g}_k = F_k(\|r_0\|_2 e_1) = G_{k,k+1} \cdots G_{12}(\|r_0\|_2 e_1),$$

wobei

$$G_{j,j+1} = G_{j,j+1}(c_j, s_j) = \begin{pmatrix} c_j & s_j \\ -s_j & c_j \end{pmatrix}, \qquad j = 1, \ldots, k,$$

die benutzten Givens-Rotationen sind. Dann erhält man $\gamma_1, \ldots, \gamma_{k+1}$ aus

- $\gamma_1 := \|r_0\|_2$.
- Für $j = 1, \ldots, k$:
$$\begin{pmatrix} \gamma_j \\ \gamma_{j+1} \end{pmatrix} := \begin{pmatrix} c_j & s_j \\ -s_j & c_j \end{pmatrix} \begin{pmatrix} \gamma_j \\ 0 \end{pmatrix}.$$

Insbesondere ist $\gamma_{k+1} = -s_k \gamma_k$. Hiermit wollen wir uns überlegen (siehe Y. Saad (1996, S. 165)):

- Sei $A \in \mathbb{R}^{n \times n}$ nichtsingulär. Dann bricht das GMRES-Verfahren im j-ten Schritt wegen $h_{j+1,j} = 0$ genau dann ab, wenn x_j die Lösung des gegebenen linearen Gleichungssystems $Ax = b$ ist.

Denn: Zunächst nehmen wir an, es sei $h_{j+1,j} = 0$. Im Algorithmus wird dann $k := j$ gesetzt. Die letzte Givens-Rotation $G_{k,k+1}$ ist die Identität, da das zu annullierende Element schon Null ist. Also ist $s_k = 0$, folglich $\gamma_{k+1} = 0$ und daher (siehe obige "Feststellung") $Ax_k = b$. Die (nicht ganz so wichtige) Umkehrung beweist man entsprechend.

Wir haben bisher nur das Verfahren GMRES geschildert, wissen aber nichts über die Güte der neuen Näherung x_k. Sei x^* die Lösung von $Ax = b$. Nach Konstruktion ist x_k die Lösung der Aufgabe

$$\text{Minimiere} \quad \|b - Ax\|_2 = \|A(x^* - x)\|_2, \quad x \in x_0 + \mathcal{K}_k.$$

Folglich ist natürlich $\|b - Ax_k\|_2 \le \|b - Ax_0\|_2$, der Defekt kann sich also jedenfalls nicht vergrößern. Für positiv definites A kann mehr gezeigt werden (siehe Y. Saad (1996, S. 195)):

Lemma 2.5 *Sei $A \in \mathbb{R}^{n\times n}$ positiv definit (aber nicht notwendig symmetrisch, also $x^T Ax > 0$ für alle $x \in \mathbb{R}^n \setminus \{0\}$) und ist x_k aus GMRES mit Startwert x_0 gewonnen, so ist*

$$\|b - Ax_k\|_2 \le \Big(1 - \frac{\mu^2}{\sigma^2}\Big)^{1/2} \|b - Ax_0\|_2,$$

wobei

$$\mu := \lambda_{\min}(A + A^T)/2, \qquad \sigma := \|A\|_2.$$

Beweis: Zur Abkürzung setzen wir $r_0 := b - Ax_0$, $r_k := b - Ax_k$. Für jedes $\alpha \in \mathbb{R}$ ist $x_0 + \alpha r_0 \in x_0 + \mathcal{K}_k$. Nach Definition von GMRES ist daher

$$\|r_k\|_2^2 \le \|b - A[x_0 + \alpha r_0]\|_2^2 = \|r_0 - \alpha A r_0\|_2^2 \qquad \text{für alle } \alpha \in \mathbb{R}.$$

Setzt man hier speziell

$$\alpha := \frac{r_0^T A r_0}{\|Ar_0\|_2^2},$$

so erhält man

$$\begin{aligned}
\|r_k\|_2^2 &\le \|r_0\|_2^2 - \Big(\frac{r_0^T A r_0}{\|Ar_0\|_2}\Big)^2 \\
&= \|r_0\|_2^2 \Big[1 - \underbrace{\Big(\frac{r_0^T A r_0}{\|r_0\|_2^2}\Big)^2}_{\ge \mu} \underbrace{\Big(\frac{\|r_0\|_2}{\|Ar_0\|_2}\Big)^2}_{\ge 1/\sigma}\Big] \\
&\le \|r_0\|_2^2 \Big(1 - \frac{\mu^2}{\sigma^2}\Big),
\end{aligned}$$

woraus die Behauptung folgt. □

Bemerkung: Lemma 2.5 zeigt, dass GMRES mit Restart für eine positiv definite Koeffizientenmatrix eine gegen die Lösung x^* konvergente Folge von Näherungslösungen liefert. □

Im nächsten Ergebnis (siehe Y. Saad (1996, S. 195)) befreien wir uns von der Voraussetzung, dass die Koeffizientenmatrix $A \in \mathbb{R}^{n\times n}$ positiv definit ist, setzen aber immer noch die Diagonalisierbarkeit von A voraus.

Satz 2.6 *Sei $A \in \mathbb{R}^{n\times n}$ diagonalisierbar, also $A = X\Lambda X^{-1}$ mit $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Man definiere*

$$\epsilon^{(k)} := \min_{p\in\Pi_k, p(0)=1} \max_{i=1,\ldots,n} |p(\lambda_i)|.$$

Wendet man mit einer Näherungslösung x_0 bzw. dem Anfangsdefekt r_0 GMRES an, um x_k bzw. den zugehörigen Defekt r_k zu erhalten, so ist

$$\|r_k\|_2 \le \kappa_2(X)\epsilon^{(k)} \|r_0\|_2,$$

wobei $\kappa_2(X) := \|X\|_2 \|X^{-1}\|_2$ die Kondition von X ist.

Beweis: Nach Definition ist x_k Lösung der Aufgabe

$$\text{Minimiere} \quad \|b - Ax\|_2, \quad x \in x_0 + \mathcal{K}_k,$$

wobei natürlich $\mathcal{K}_k := \mathcal{K}_k(A, r_0)$. Für ein beliebiges $x \in x_0 + \mathcal{K}_k$ ist $x = x_0 + q(A)r_0$ mit $q \in \Pi_{k-1}$ und daher

$$b - Ax = b - A(x_0 + q(A)r_0 = [I - Aq(A)]r_0 = p(A)r_0$$

mit $p \in \Pi_k$, $p(0) = 1$. Mit diesen Bezeichnungen ist

$$\begin{aligned} \|r_k\|_2 &= \|b - Ax_k\|_2 \\ &\leq \|b - Ax\|_2 \\ &= \|p(A)r_0\|_2 \\ &= \|Xp(\Lambda)X^{-1}r_0\|_2 \\ &\leq \|X\|_2 \, \|X^{-1}\|_2 \, \|p(\Lambda)\|_2 \, \|r_0\|_2 \\ &= \kappa_2(X) \max_{i=1,\ldots,n} |p(\lambda_i)| \, \|r_0\|_2. \end{aligned}$$

Hier kann nun unter allen Polynomen $p \in \Pi_k$ mit $p(0) = 1$ eines gewählt werden, für welches $\min_{i=1,\ldots,n} |p(\lambda_i)|$ minimal ist. Diese Beobachtung liefert die Behauptung. □

Bemerkung: Selbst wenn man wüsste, wo die Eigenwerte $\lambda_1, \ldots, \lambda_n$ liegen (bei Y. Saad (1996, S. 196 ff.) werden Abschätzungen für $\epsilon^{(k)}$ unter der Annahme gemacht, dass die Eigenwerte λ_i in einer gewissen Ellipse in der komplexen Ebene liegen, welchen den Nullpunkt nicht enthält) kann obige Abschätzung schlecht sein, wenn $\kappa_2(X)$ groß ist. Man beachte aber, dass eine normale Matrix A unitär ähnlich einer Diagonalmatrix ist, woraus $\kappa_2(X) = 1$ folgt. Man kann also erwarten, dass GMRES gut konvergiert, wenn die Matrix A normal ist und sich die Eigenwerte von A um einen Punkt außerhalb des Ursprungs häufen. Ist dies Anfangs nicht der Fall, so versucht man die Situation durch eine *Präkonditionierung* zu verbessern. □

5.2.5 Das symmetrische Lanczos-Verfahren

Das symmetrische Lanczos-Verfahren kann als eine Vereinfachung des Arnoldi-Verfahrens für den Spezialfall angesehen werden, dass die gegebene Matrix $A \in \mathbb{R}^{n \times n}$ symmetrisch ist. Wegen $Q_k^T A Q_k = H_k$ ist die symmetrische obere Hessenberg-Matrix H_k in diesem Fall eine symmetrische Tridiagonalmatrix. Statt H_k wird diese Matrix mit T_k bezeichnet, es sei

$$T_k = \begin{pmatrix} \alpha_1 & \beta_2 & & \\ \beta_2 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_k \\ & & \beta_k & \alpha_k \end{pmatrix}.$$

Da man schon vorher weiß, dass man nur die Subdiagonalelemente $\beta_{j+1} = h_{j+1,j}$ und die Diagonalelemente $\alpha_j = h_{jj}$ zu berechnen braucht, kann dies im Algorithmus gleich berücksichtigt werden. Die Spezialisierung des MGS-Arnoldi-Verfahrens zur Berechnung

einer Orthonormalbasis des Krylov-Teilraumes $\mathcal{K}_k(A, v)$ sieht dann folgendermaßen aus:

- Input: Seien $A \in \mathbb{R}^{n\times n}$ symmetrisch, $v \in \mathbb{R}^n \setminus \{0\}$ und $k \leq n$ gegeben.
- Berechne $q_1 := v/\|v\|_2$, setze $\beta_1 := 0$ und $q_0 := 0$.
- Für $j = 1, \ldots, k$:
  - $w := Aq_j - \beta_j q_{j-1}$ (oder auch $w := Aq_j$).
  - $\alpha_j := q_j^T w$, $w := w - \alpha_j q_j$, $\beta_{j+1} := \|w\|_2$.
  - Falls $\beta_{j+1} = 0$, dann: STOP, andernfalls: $q_{j+1} := w/\beta_{j+1}$.
- Output: Berechnet wird eine Orthonormalbasis $\{q_1, \ldots, q_k\}$ von $\mathcal{K}_k$, hiermit die Matrix $Q_k = \begin{pmatrix} q_1 & \cdots & q_k \end{pmatrix}$ und $Q_{k+1} = \begin{pmatrix} Q_k & q_{k+1} \end{pmatrix}$ sowie die Tridiagonalmatrizen

$$T_k = \begin{pmatrix} \alpha_1 & \beta_2 & & \\ \beta_2 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_k \\ & & \beta_k & \alpha_k \end{pmatrix}, \qquad \tilde{T}_k = \begin{pmatrix} \alpha_1 & \beta_2 & & \\ \beta_2 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_k \\ & & \beta_k & \alpha_k \\ & & & \beta_{k+1} \end{pmatrix}$$

mit

$$Q_k^T A Q_k = T_k, \qquad AQ_k = Q_{k+1}\tilde{T}_k.$$

Bemerkung: Umgekehrt führen die im Output gemachten Aussagen leicht auf den angegebenen Algorithmus. Aus dem Ansatz

$$AQ_k = Q_{k+1}\tilde{T}_k$$

erhält man durch Betrachten der j-ten Spalte

$$Aq_j = \beta_j q_{j-1} + \alpha_j q_j + \beta_{j+1} q_{j+1}, \qquad j = 1, \ldots, k,$$

wobei $\beta_1 q_0 = 0$. Die geforderte Orthonormalität der q_j liefert $\alpha_j = q_j^T A q_j$ oder, und diese Version haben wir bevorzugt, $\alpha_j = q_j^T(Aq_j - \beta_j q_{j-1})$. Es ist

$$w = Aq_j - \beta_j q_{j-1} - \alpha_j q_j = \beta_{j+1} q_{j+1},$$

was auf $\beta_{j+1} = \|w\|_2$ führt. Damit erhalten wir, genau wie oben angegeben, $\beta_{j+1} = \|w\|_2$ und danach für $\beta_{j+1} \neq 0$ auch $q_{j+1} = w/\beta_{j+1}$. □

Das Lanczos-Verfahren zur Berechnung einer Näherungslösung eines linearen Gleichungssystems $Ax = b$ mit symmetrischem $A \in \mathbb{R}^{n\times n}$ ist eine Spezialisierung (auf eine symmetrische Matrix) der "Full Orthogonalization Method" (FOM), welche darauf beruht, bei einer gegebenen Näherungslösung x_0 ein $x_k \in x_0 + \mathcal{K}_k$ mit $b - Ax_k \perp \mathcal{K}_k$ zu bestimmen.

- Mit einer Näherungslösung x_0 berechne man den Defekt $r_0 := b - Ax_0$. Berechne $q_1 := r_0/\|r_0\|_2$. Setze $\beta_1 := 0$, $q_0 := 0$.
- Für $j = 1, \ldots, k$:
  - $w := Aq_j - \beta_j q_{j-1}$.
  - Berechne $\alpha_j := q_j^T w$, $w := w - \alpha_j q_j$, $\beta_{j+1} := \|w\|_2$.
  - Falls $\beta_{j+1} = 0$, dann: Setze $k := j$ und gehe zum letzten $\bullet$.
  - $q_{j+1} := w/\beta_{j+1}$.
- Setze $Q_k = (\ q_1 \ \cdots \ q_k\ )$,

$$T_k := \begin{pmatrix} \alpha_1 & \beta_2 & & \\ \beta_2 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_k \\ & & \beta_k & \alpha_k \end{pmatrix},$$

  und berechne $x_k := x_0 + Q_k T_k^{-1}(\|r_0\|_2 e_1)$.

Die Hauptarbeit besteht in der Berechnung von $y_k := T_k^{-1}(\|r_0\|_2 e_1)$ bzw. der Lösung eines linearen Gleichungssystem mit der symmetrischen Tridiagonalmatrix T_k als Koeffizientenmatrix und der rechten Seite $\|r_0\|_2\, e_1$. Wir nehmen an, dieses sei mit Hilfe des Gaußschen Eliminationsverfahren ohne Pivotsuche möglich[5]. Dies wiederum ist äquivalent dazu, dass T_k eine LU-Zerlegung $T_k = L_k U_k$ mit

$$L_k = \begin{pmatrix} 1 & & & \\ \lambda_2 & 1 & & \\ & \ddots & \ddots & \\ & & \lambda_k & 1 \end{pmatrix}, \qquad U_k = \begin{pmatrix} \eta_1 & \beta_2 & & \\ & \eta_2 & \ddots & \\ & & \ddots & \beta_k \\ & & & \eta_k \end{pmatrix}$$

besitzt. Bekanntlich lassen sich die noch unbekannten $\lambda_2, \ldots, \lambda_k$ und $\eta_1, \ldots, \eta_k$ sukzessive berechnen. Wegen

$$(L_k U_k)_{j,j-1} = \lambda_j \eta_{j-1}, \qquad (L_k U_k)_{jj} = \lambda_j \beta_j + \eta_j$$

geschieht dies mit dem folgenden Verfahren:

- $\eta_1 := \alpha_1$.
- Für $j = 2, \ldots, k$:
  - $\lambda_j := \beta_j/\eta_{j-1}$.
  - $\eta_j := \alpha_j - \lambda_j \beta_j$.

[5] Ist A nicht nur symmetrisch, sondern auch positiv definit, so ist auch $T_k = Q_k^T A Q_k$ symmetrisch und positiv definit. Folglich ist in diesem Falle das Gaußsche Eliminationsverfahren ohne Pivotsuche durchführbar bzw. besitzt T_k eine LU-Zerlegung.

Ziel für das weitere ist es, ein Lanczos-Verfahren zur sukzessiven Berechnung von $x_1, x_2, \ldots$ anzugeben. Die durch das obige Lanczos-Verfahren berechnete Näherungslösung x_k ist (unter der Voraussetzung, dass die Tridiagonalmatrix T_k eine LU-Zerlegung besitzt) durch

$$x_k = x_0 + Q_k R_k^{-1} L_k^{-1}(\|r_0\|_2 e_1)$$

gegeben. Wir setzen

$$P_k := \begin{pmatrix} p_0 & \cdots & p_{k-1} \end{pmatrix} := Q_k R_k^{-1}, \qquad z_k := L_k^{-1}(\|r_0\|_2 e_1),$$

so dass $x_k = x_0 + P_k z_k$. Zunächst überlegen wir uns:

- Es ist
$$z_k = \begin{pmatrix} z_{k-1} \\ \zeta_{k-1} \end{pmatrix}, \qquad k = 1, \ldots,$$
wobei
$$\zeta_0 := \|r_0\|_2, \qquad \zeta_{k-1} := -\lambda_k \zeta_{k-2}, \qquad k = 2, \ldots.$$

Denn: Wir zeigen die Behauptung durch vollständige Induktion nach k. Für $k = 1$ ist $z_1 = (\zeta_0)$, die Behauptung also richtig. Nun nehmen wir an, es sei $L_{k-1} z_{k-1} = \|r_0\|_2 e_1$ (wobei e_1 der erste Einheitsvektor im $\mathbb{R}^{k-1}$ ist), die letzte (also $(k-1)$-te) Komponente von z_{k-1} sei ζ_{k-2} und es sei $\zeta_{k-1} := -\lambda_k \zeta_{k-2}$. Dann ist

$$L_k \begin{pmatrix} z_{k-1} \\ \zeta_{k-1} \end{pmatrix} = \begin{pmatrix} L_{k-1} & 0 \\ \lambda_k e_{k-1}^T & 1 \end{pmatrix} \begin{pmatrix} z_{k-1} \\ \zeta_{k-1} \end{pmatrix} = \begin{pmatrix} L_{k-1} z_{k-1} \\ \lambda_k \zeta_{k-2} + \zeta_{k-1} \end{pmatrix} = \begin{pmatrix} \|r_0\|_2 e_1 \\ 0 \end{pmatrix},$$

womit die Induktionsbehauptung bewiesen ist.

Folglich ist

$$\begin{aligned} x_k &= x_0 + P_k z_k \\ &= x_0 + \begin{pmatrix} P_{k-1} & p_{k-1} \end{pmatrix} \begin{pmatrix} z_{k-1} \\ \zeta_{k-1} \end{pmatrix} \\ &= x_0 + P_{k-1} z_{k-1} + \zeta_{k-1} p_{k-1} \\ &= x_{k-1} + \zeta_{k-1} p_{k-1}. \end{aligned}$$

Wegen $P_k R_k = Q_k$ ist (vergleiche die k-te Spalte)

$$\begin{pmatrix} p_0 & \cdots & p_{k-2} & p_{k-1} \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ \beta_k \\ \eta_k \end{pmatrix} = q_k$$

und daher $\beta_k p_{k-2} + \eta_k p_{k-1} = q_k$ bzw.

$$p_{k-1} = \frac{1}{\eta_k}[q_k - \beta_k p_{k-2}].$$

Hier merken wir uns, dass p_{k-1} eine Linearkombination von q_k und p_{k-2} ist, wobei der Koeffizient von q_k nicht verschwindet.

Insgesamt erhält man den folgenden Algorithmus, sozusagen eine direkte Version des symmetrischen Lanczos. Hierbei muss natürlich aus dem Verfahren ausgestiegen werden, wenn versucht wird, durch Null zu dividieren.

- Mit einer Näherungslösung x_0 berechne man den Defekt $r_0 := b - Ax_0$. Ferner berechne man $q_1 := r_0/\|r_0\|_2$, $\zeta_0 := \|r_0\|_2$ und setze $\beta_1 := 0$, $q_0 := 0$ sowie $\lambda_1 := 0$ und $p_{-1} := 0$.
- Für $k = 1, 2, \ldots$:
  - $w := Aq_k - \beta_k q_{k-1}$, $\alpha_k := q_k^T w$.
  - Falls $k > 1$, dann: $\lambda_k := \beta_k/\eta_{k-1}$, $\zeta_{k-1} := -\lambda_k \zeta_{k-2}$.
  - $\eta_k := \alpha_k - \lambda_k \beta_k$.
  - $p_{k-1} := (q_k - \beta_k p_{k-2})/\eta_k$.
  - $x_k := x_{k-1} + \zeta_{k-1} p_{k-1}$.
  - Wenn x_k einem geeigneten Abbruchkriterium genügt, dann: STOP.
  - $w := w - \alpha_k q_k$, $\beta_{k+1} := \|w\|_2$, $q_{k+1} := w/\beta_{k+1}$.

Abbrechen wird man dieses Verfahren, wenn der Defekt $\|b - Ax_k\|_2$ oder der relative Defekt $\|b - Ax_k\|_2/\|b\|_2$ hinreichend klein ist. Daher ist es wichtig, $\|b - Ax_k\|_2$ effizient zu berechnen, d. h. mit möglichst wenig Extraarbeit. Nun ist

$$\begin{aligned} b - Ax_k &= b - A[x_0 + Q_k T_k^{-1}(\|r_0\|_2 e_1)] \\ &= r_0 - Q_{k+1}\tilde{T}_k T_k^{-1}(\|r_0\|_2 e_1)] \\ &= r_0 - \begin{pmatrix} Q_k & q_{k+1} \end{pmatrix} \begin{pmatrix} T_k \\ \beta_{k+1} e_k^T \end{pmatrix} T_k^{-1}(\|r_0\|_2 e_1)] \\ &= r_0 - (Q_k T_k + \beta_{k+1} q_{k+1} e_k^T) T_k^{-1}(\|r_0\|_2 e_1) \\ &= \underbrace{r_0 - \|r_0\|_2 Q_k e_1}_{=0} - \beta_{k+1} e_k^T y_k q_{k+1}, \end{aligned}$$

wobei wir zur Abkürzung

$$y_k := T_k^{-1}(\|r_0\|_2 e_1)$$

gesetzt haben. Insbesondere ist das Residuum $r_k := b - Ax_k$ ein Vielfaches von q_{k+1}, woraus auch folgt, dass $r_0, r_1, \ldots$ aufeinander senkrecht stehen. Ferner ist

$$\|b - Ax_k\|_2 = |\beta_{k+1}|\,|e_k^T y_k|.$$

Schließlich gilt noch (siehe Y. Saad (1996, S. 178)):

- Die im obigen direkten Lanczos-Verfahren berechneten Vektoren $p_0, p_1, \ldots$ sind A-konjugiert, d. h. es ist $p_i^T A p_j = 0$ für $i \neq j$.

Denn: Da $p_0, \ldots, p_{k-1}$ die Spalten von $P_k = Q_k R_k^{-1}$ sind, haben wir zu zeigen, dass $P_k^T A P_k$ eine Diagonalmatrix ist. Es ist

$$P_k^T A P_k = R_k^{-T} Q_k^T A Q_k R_k^{-1} = R_k^{-T} T_k R_k^{-1} = R_k^{-T} L_k.$$

Nun beachte man, dass $R_k^{-T} L_k$ eine symmetrische untere Dreiecksmatrix und folglich eine Diagonalmatrix ist. Daher ist die obige Aussage richtig.

Bemerkung: Wir sind in diesem Kapitel zwar hauptsächlich an linearen Gleichungssystemen interessiert, trotzdem ist hier eine Bemerkung zur Anwendung des Lanczos-Verfahrens auf symmetrische Eigenwertaufgaben angebracht. Sei also $A \in \mathbb{R}^{n\times n}$ symmetrisch, durch das Lanczos-Verfahren sei die Zerlegung $Q_k^T A Q_k = T_k$ mit der symmetrischen Tridiagonalmatrix T_k und der Matrix Q_k, deren Spalten orthonormal sind, berechnet. Die Eigenwerte von T_k seien $\lambda_1(T_k) \geq \cdots \geq \lambda_k(T_k)$, die von A seien $\lambda_1(A) \geq \cdots \geq \lambda_n(A)$. Wegen des Rayleigh-Prinzips ist

$$\lambda_1(T_k) = \max_{y\in\mathbb{R}^k\setminus\{0\}} \frac{y^T T_k y}{y^T y} = \max_{y\in\mathbb{R}^k\setminus\{0\}} \frac{(Q_k y)^T A (Q_k y)}{(Q_k y)^T (Q_k y)} \leq \max_{x\in\mathbb{R}^n\setminus\{0\}} \frac{x^T A x}{x^T x} = \lambda_1(A)$$

und entsprechend $\lambda_k(T_k) \geq \lambda_n(A)$. □

5.2.6 MINRES

Im unsymmetrischem Fall war GMRES ein Verfahren, bei dem mit einer Startnäherung x_0 eine Lösung x_k der Aufgabe

$$\text{Minimiere} \quad \|b - Ax\|_2, \quad x \in x_0 + \mathcal{K}_k$$

gesucht wurde. Hierbei war $A \in \mathbb{R}^{n\times n}$ i. allg. unsymmetrisch. Eine Spezialisierung auf den symmetrischen Fall liefert MINRES (**Min**imal **Res**idual Algorithm). Hierauf wollen wir kurz eingehen (siehe z. B. auch A. Greenbaum (1997, S. 41 ff.)). Wieder nehmen wir an, x_0 sei eine Startnäherung und folglich $r_0 := b - Ax_0$ das zugehörige Residuum. Außerdem seien die Spalten von $Q_k = (\ q_1 \ \cdots \ q_k\ )$ eine Orthonormalbasis des Krylov-Teilraumes $\mathcal{K}_k = \mathcal{K}_k(A, r_0)$. Wie im GMRES wird x_k als Lösung der Aufgabe, $\|b-Ax\|_2$ auf $x_0+\mathcal{K}_k$ zu minimieren, bestimmt. Anders gewendet (siehe GMRES) bestimmt man die Lösung y_k des linearen Ausgleichsproblems, $\|\|r_0\|_2 e_1 - \tilde{T}_k y\|_2$ auf dem $\mathbb{R}^k$ zu minimieren und berechnet $x_k = x_0 + Q_k y_k$. Ähnlich wie beim direkten Lanczos-Verfahren wollen wir nun eine Rekursionsformel für x_k aufstellen. Wir erinnern daran, dass

$$\tilde{T}_k = \begin{pmatrix} \alpha_1 & \beta_2 & & \\ \beta_2 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_k \\ & & \beta_k & \alpha_k \\ & & & \beta_{k+1} \end{pmatrix}.$$

Wie bei GMRES sei

$$(\ \tilde{R}_k \quad \tilde{g}_k\ ) := F_k(\ \tilde{T}_k \quad \|r_0\| e_1\ ),$$

wobei wieder

$$F_k = G_{k,k+1} \cdots G_{12}$$

das Produkt geeigneter Givens-Rotationen ist. Auch $R_k \in \mathbb{R}^{k\times k}$ und g_k erhalte man wieder aus $\tilde{R}_k$ bzw. $\tilde{g}_k$ durch Weglassen der letzten Zeile bzw. Komponente. Dann ist

$$x_k = x_0 + P_k g_k,$$

wobei wir, ähnlich wie beim direkten Lanczos-Verfahren,

$$P_k := Q_k R_k^{-1}, \qquad P_k = \begin{pmatrix} p_0 & \cdots & p_{k-1} \end{pmatrix}$$

gesetzt haben. Weiter bemerken wir, dass in R_k in jeder Zeile höchstens 3 Einträge von Null verschieden sind, nämlich die in den Positionen (i,i), $(i,i+1)$ und $(i,i+2)$. Aus $P_k R_k = Q_k$ erhält man durch Vergleich der k-ten Spalte, dass

$$\begin{pmatrix} 0 & \cdots & 0 & p_{k-3} & p_{k-2} & p_{k-1} \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ r_{k-2,k} \\ r_{k-1,k} \\ r_{kk} \end{pmatrix} = q_k$$

und folglich

$$p_{k-1} = \frac{1}{r_{kk}}(q_k - r_{k-1,k} p_{k-2} - r_{k-2,k} p_{k-3}).$$

Dann ist

$$x_k = x_0 + P_k g_k = x_0 + \begin{pmatrix} P_{k-1} & p_{k-1} \end{pmatrix} \begin{pmatrix} g_{k-1} \\ \gamma_k \end{pmatrix} = x_{k-1} + \gamma_k p_{k-1}.$$

Mit Hilfe dieser Bemerkungen ist es nicht schwierig, den entsprechenden Algorithmus aufzustellen (siehe A. Greenbaum (1997, S. 44)).

5.2.7 CG-Verfahren

Wie wir beim direkten (symmetrischen) Lanczos-Verfahren gesehen haben, kann zur Lösung des linearen Gleichungssystems $Ax = b$ mit symmetrischem $A \in \mathbb{R}^{n\times n}$ ein Iterationsverfahren der Form $x_k = x_{k-1} + \alpha_{k-1} p_{k-1}$ bzw. (verschiebe den Iterationsindex um einen nach oben)

$$x_{k+1} := x_k + \alpha_k p_k, \qquad k = 0, 1, \ldots,$$

angegeben werden, welches die folgenden Eigenschaften hat:

- Die Richtungen $p_0, p_1, \ldots$ sind A-konjugiert, d. h.

$$p_i^T A p_j = 0, \qquad i \neq j.$$

- Die Defekte $r_0, r_1, \ldots$ stehen paarweise aufeinander senkrecht, d. h.

$$r_i^T r_j = 0, \qquad i \neq j.$$

- Mit einer Orthonormalbasis $\{q_1, \ldots, q_k\}$ des Krylov-Teilraumes $\mathcal{K}_k := \mathcal{K}_k(A, r_0)$ ist p_{k-1} eine Linearkombination von q_k und p_{k-2}, wobei der Koeffizient von q_k nicht verschwindet. Ferner ist r_{k-1} ein Vielfaches von q_k. Nach eventueller Skalierung (zu Lasten der Schrittweite α_{k-1}) kann also angenommen werden, dass $p_{k-1} = r_{k-1} + \beta_{k-2} p_{k-2}$ bzw.

$$p_{k+1} = r_{k+1} + \beta_k p_k, \qquad k = 1, 2, \ldots.$$

Für den Fall, dass $A \in \mathbb{R}^{n\times n}$ nicht nur symmetrisch, sondern sogar positiv definit ist, wollen wir hieraus das auf Hestenes-Stiefel (1952) zurückgehende CG-Verfahren ableiten. Es stimmt mit dem Lanczos-Verhren überein, hat aber den Vorteil, dass auf die Berechnung einer Orthonormalbasis der entsprechenden Krylov-Teilräume verzichtet wird. Aus $x_{k+1} = x_k + \alpha_k p_k$ folgt

$$r_{k+1} = b - Ax_{k+1} = b - Ax_k - \alpha_k Ap_k = r_k - \alpha_k Ap_k.$$

Aus der Orthogonalitätsforderungen an die Residuen erhält man

$$0 = r_{k+1}^T r_k = r_k^T r_k - \alpha_k r_k^T Ap_k,$$

damit

$$\alpha_k = \frac{r_k^T r_k}{r_k^T Ap_k}.$$

Dann ist

$$r_k^T Ap_k = (p_k - \beta_{k-1}p_{k-1})^T Ap_k = p_k^T Ap_k,$$

damit

$$\alpha_k = \frac{r_k^T r_k}{p_k^T Ap_k}.$$

Weiter ist

$$0 = p_{k+1}^T Ap_k = r_{k+1}^T Ap_k + \beta_k p_k^T Ap_k,$$

woraus

$$\beta_k = -\frac{r_{k+1}^T Ap_k}{p_k^T Ap_k}$$

folgt. Wegen $Ap_k = -(1/\alpha_k)(r_{k+1} - r_k)$ kann man für β_k auch schreiben:

$$\beta_k = \frac{1}{\alpha_k}\frac{r_{k+1}^T(r_{k+1} - r_k)}{p_k^T Ap_k} = \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}.$$

Damit lautet das CG-Verfahren, wobei wir allerdings auf eine Abfrage, ob der Defekt verschwindet, verzichten:

- Mit einer Näherung x_0 berechne man $r_0 := b - Ax_0$ und setze $p_0 := r_0$.
- Für $k = 0, 1, \ldots$:
  - $\alpha_k := r_k^T r_k / p_k^T Ap_k$.
  - $x_{k+1} := x_k + \alpha_k p_k$.
  - $r_{k+1} := r_k - \alpha_k Ap_k$.
  - $\beta_k := r_{k+1}^T r_{k+1} / r_k^T r_k$.
  - $p_{k+1} := r_{k+1} + \beta_k p_k$.

Hierbei sollte betont werden, dass die Skalare α_k, β_k nicht mit den entsprechend bezeichneten Skalaren im Lanczos-Verfahren übereinstimmen. Der Algorithmus benötigt Speicherplatz für die aktuelle Näherung x, für die Richtung p, für Ap (dieses Matrix-Vektorprodukt wird in jeder Schleife natürlich nur einmal berechnet) und für das Residuum r.

Es sei noch einmal betont, dass das CG-Verfahren mit dem direkten Lanczos-Verfahren übereinstimmt, dass aber vermieden wird, eine Basis für die entsprechenden Krylov-Teilräume zu berechnen. Daher ist es klar, dass $x_k \in x_0 + \mathcal{K}_k$, wobei $\mathcal{K}_k = \mathcal{K}_k(A, r_0)$. Damit ist auch klar, dass

$$\mathcal{K}_k = \operatorname{span}\{r_0, \ldots, r_{k-1}\} = \operatorname{span}\{p_0, \ldots, p_{k-1}\}.$$

Neben der euklidischen Norm $\|\cdot\|_2$ ist eine weitere, durch die symmetrische und positiv definite Matrix $A \in \mathbb{R}^{n\times n}$ erzeugte Norm, die wir mit $\|\cdot\|_A$ bezeichnen, von Bedeutung. Hierbei ist

$$\|x\|_A := (x^T Ax)^{1/2} = \|A^{1/2}x\|_2.$$

Wir wollen zeigen:

- Wird $\{x_k\}$ durch das CG-Verfahren erzeugt und ist x^* die Lösung von $Ax = b$, so ist x_k die Lösung der Aufgabe

  $$\text{Minimiere} \quad \|x^* - x\|_A, \quad x \in x_0 + \mathcal{K}_k.$$

  Als Element von $x_0 + \mathcal{K}_k$ hat x_k die Darstellung $x_k = x_0 + q_{k-1}(A)r_0$ mit einem Polynom $q_{k-1} \in \Pi_{k-1}$ und es ist

  $$\|(I - Aq_{k-1}(A))(x^* - x_0)\|_A = \min_{q\in\Pi_{k-1}} \|(I - Aq(A))(x^* - x_0)\|_A.$$

Denn: Definiert man $f(x) := \|x^* - x\|_A$, so ist x_k genau dann eine Lösung der Aufgabe, f auf $x_0 + \mathcal{K}_k$ zu minimieren, wenn $(x^* - x_k)^T Ax = 0$ für alle $x \in \mathcal{K}_k$. Für ein beliebiges $x \in \mathcal{K}_k = \operatorname{span}\{r_0, \ldots, r_{k-1}\}$ ist aber

$$(x^* - x_k)^T Ax = x^T A(x^* - x_k) = x^T(b - Ax_k) = x^T r_k = 0,$$

womit die erste Aussage bewiesen ist. Die zweite Aussage ist eine direkte Folgerung aus der ersten, wenn man beachtet, dass

$$\begin{aligned}
\|x^* - x_k\|_A &= \|x^* - [x_0 + q_{k-1}(A)r_0]\|_A \\
&= \|x^* - [x_0 + q_{k-1}(A)(b - Ax_0)]\|_A \\
&= \|x^* - [x_0 + q_{k-1}(A)A(x^* - x_0)]\|_A \\
&= \|(I - Aq_{k-1}(A))(x^* - x_0)\|_A.
\end{aligned}$$

Ist $q \in \Pi_{k-1}$, so ist $p(t) := 1 - tp(t)$ ein Polynom vom Grad $\le k$ mit $p(0) = 1$. Daher ist

$$\|x^* - x_k\|_A = \min_{p\in\Pi_k, p(0)=1} \|p(A)(x^* - x_0)\|_A.$$

Seien nun λ_i, $i = 1, \ldots, n$, die Eigenwerte der symmetrischen, positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$, mit $\lambda_{\min}$ bezeichnen wir den kleinsten, mit $\lambda_{\max}$ den größten Eigenwert von A. Ferner sei $\{u_1, \ldots, u_n\}$ ein vollständiges Orthonormalsystem von Eigenvektoren und

$$x^* - x_0 = \sum_{i=1}^{n} \alpha_i u_i.$$

Für ein beliebiges Polynom $p \in \Pi_k$ ist dann

$$\begin{aligned} \|p(A)(x^* - x_0)\|_A^2 &= \sum_{i=1}^{n} \lambda_i p(\lambda_i)^2 \alpha_i^2 \\ &\leq \max_{i=1,\ldots,n} p(\lambda_i)^2 \sum_{i=1,\ldots,n} \lambda_i \alpha_i^2 \\ &= \max_{i=1,\ldots,n} p(\lambda_i)^2 \, \|x^* - x_0\|_A^2 \\ &\leq \max_{\lambda\in[\lambda_{\min},\lambda_{\max}]} p(\lambda)^2 \, \|x^* - x_0\|_A^2. \end{aligned}$$

Folglich ist

$$\|x^* - x_k\|_A \leq \min_{p\in\Pi_k,\, p(0)=1} \max_{\lambda\in[\lambda_{\min},\lambda_{\max}]} |p(\lambda)| \, \|x^* - x_0\|_A.$$

Nun benutzen wir die folgende Aussage, die in einem Spezialfall bei der Tschebyscheff-Beschleunigung eines Iterationsverfahrens schon eine Rolle spielte.

- Sei $[\alpha, \beta]$ ein nichtleeres Intervall in $\mathbb{R}$ und $\gamma \in \mathbb{R}$ beliebig mit $\gamma \notin [\alpha, \beta]$. Dann ist eine Lösung der Aufgabe

  $$\text{Minimiere} \quad \max_{t\in[\alpha,\beta]} |p(t)|, \quad p \in \Pi_k,\ p(\gamma) = 1$$

  gegeben durch

  $$p_k^*(t) := \frac{T_k(1 + 2(\alpha - t)/(\beta - \alpha))}{T_k(1 + 2(\alpha - \gamma)/(\beta - \alpha))}.$$

  Weiter ist

  $$\min_{p\in\Pi_k,\, p(\gamma)=1} \max_{t\in[\alpha,\beta]} |p(t)| = \frac{1}{|T_k(1 + 2(\alpha - \gamma)/(\beta - \alpha))|}.$$

  Hierbei bedeutet T_k das k-te Tschebyscheff-Polynom erster Art.

Eine Spezialisierung auf den uns interessierenden Fall liefert, dass der Fehler im k-ten Schritt des CG-Verfahrens der Abschätzung

$$\|x^* - x_k\|_A \leq \frac{1}{|T_k(1 + 2\lambda_{\min}/(\lambda_{\max} - \lambda_{\min}))|} \, \|x^* - x_0\|_A$$

genügt. Dies kann weiter abgeschätzt werden, wobei wir benutzen, dass für $|t| \geq 1$ die Darstellung

$$T_k(t) = \frac{1}{2}\Big[\Big(t + \sqrt{t^2 - 1}\Big)^k + \Big(t + \sqrt{t^2 - 1}\Big)^{-k}\Big]$$

gilt. Mit

$$\eta := \frac{\lambda_{\min}}{\lambda_{\max} - \lambda_{\min}}, \qquad \kappa := \frac{\lambda_{\max}}{\lambda_{\min}}$$

erhalten wir:

$$\begin{aligned} |T_k(1 + 2\lambda_{\min}/(\lambda_{\max} - \lambda_{\min}))| &= |T_k(1 + 2\eta)| \\ &\geq \frac{1}{2}\Big(1 + 2\eta + \sqrt{(1+2\eta)^2 - 1}\Big)^k \\ &= \frac{1}{2}\Big(1 + 2\eta + 2\sqrt{\eta(\eta+1)}\Big)^k \\ &= \frac{1}{2}[\sqrt{\eta} + \sqrt{\eta+1}]^{2k} \\ &= \frac{1}{2}\Big(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\Big)^k. \end{aligned}$$

Folglich ist

$$\|x^* - x_k\|_A \leq 2\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^k \|x^* - x_0\|_A.$$

Man erkennt, dass die Kondition der Koeffizientenmatrix A entscheidend die "Konvergenzgüte" beeinflusst.

5.2.8 Lanczos-Biorthogonalisierungsverfahren

Beim symmetrischen Lanczos-Verfahren wird zu einer symmetrischen Matrix $A \in \mathbb{R}^{n\times n}$ eine Matrix $Q_k \in \mathbb{R}^{n\times k}$ mit orthogonalen Spalten, die den Krylov-Teilraum $\mathcal{K}_k$ aufspannen, und eine (symmetrische) Tridiagonalmatrix $T_k \in \mathbb{R}^{k\times k}$ mit $Q_k^T A Q_k = T_k$ berechnet. Ferner konnte (ohne dass wir an die entsprechenden Bezeichnungen erinnern) $AQ_k = Q_{k+1}\tilde{T}_k$ nachgewiesen werden. Bei einem Biorthogonalisierungsverfahren versucht man stattdessen, zu einer (nicht notwendig symmetrischen) Matrix $A \in \mathbb{R}^{n\times n}$ Matrizen

$$V_{k+1} = (\ V_k \quad v_{k+1}\ ) \in \mathbb{R}^{n\times(k+1)}, \qquad W_{k+1} = (\ W_k \quad w_{k+1}\ ) \in \mathbb{R}^{n\times(k+1)}$$

zu berechnen mit

$$V_k^T W_k = I$$

(das ist die Biorthogonalisierungsbedingung) und

$$AV_k = V_{k+1}\tilde{T}_k, \qquad A^T W_k = W_{k+1}\tilde{S}_k, \qquad T_k = S_k^T = W_k^T A V_k.$$

Hierbei sind $\tilde{S}_k, \tilde{T}_k \in \mathbb{R}^{(k+1)\times k}$ Tridiagonalmatrizen, ferner ist $T_k = S_k^T$ durch Weglassen der letzten Zeile von $\tilde{T}_k$ (oder durch Weglassen der letzten Spalte von $\tilde{S}_k$) erhalten. Zur Berechnung von

$$V_{k+1} = (\ v_1 \quad \cdots \quad v_k \quad v_{k+1}\ ), \qquad W_{k+1} = (\ w_1 \quad \cdots \quad w_k \quad w_{k+1}\ )$$

sowie

$$T_k = \begin{pmatrix} \alpha_1 & \beta_2 & & & \\ \delta_2 & \alpha_2 & \beta_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \delta_{k-1} & \alpha_{k-1} & \beta_k \\ & & & \delta_k & \alpha_k \end{pmatrix}$$

betrachten wir das folgende *Lanczos-Biorthogonalisierungsverfahren*:

- Input: Gegeben sei die (i. allg. unsymmetrische) Matrix $A \in \mathbb{R}^{n\times n}$, ferner zwei Vektoren $v_1, w_1 \in \mathbb{R}^n$ mit $v_1^T w_1 = 1$.
- Setze $\beta_1 := 0$, $\delta_1 := 0$ sowie $v_0 := 0$, $w_0 := 0$.
- Für $j = 1, \ldots, k$:
  - $\alpha_j := (Av_j)^T w_j$.
  - $\hat{v}_{j+1} := Av_j - \alpha_j v_j - \beta_j v_{j-1}$, $\hat{w}_{j+1} := A^T w_j - \alpha_j w_j - \delta_j w_{j-1}$.
  - Falls $\hat{v}_{j+1}^T \hat{w}_{j+1} = 0$, dann: STOP, andernfalls wähle man (von Null verschiedene) reelle Zahlen $\delta_{j+1}, \beta_{j+1}$ mit $\delta_{j+1}\beta_{j+1} = \hat{v}_{j+1}^T \hat{w}_{j+1}$.
  - $v_{j+1} = \hat{v}_{j+1}/\delta_{j+1}$, $w_{j+1} := \hat{w}_{j+1}/\beta_{j+1}$.
- Output: Falls kein vorzeitiger Abbruch erfolgt, werden Matrizen

$$V_k := \begin{pmatrix} v_1 & \cdots & v_k \end{pmatrix} \in \mathbb{R}^{n\times k}, \qquad V_{k+1} := \begin{pmatrix} V_k & v_{k+1} \end{pmatrix} \in \mathbb{R}^{n\times(k+1)}$$

sowie

$$W_k := \begin{pmatrix} w_1 & \cdots & w_k \end{pmatrix} \in \mathbb{R}^{n\times k}, \qquad W_{k+1} := \begin{pmatrix} W_k & w_{k+1} \end{pmatrix} \in \mathbb{R}^{n\times(k+1)}$$

berechnet, ferner Tridiagonalmatrizen

$$T_k := \begin{pmatrix} \alpha_1 & \beta_2 & & & \\ \delta_2 & \alpha_2 & \beta_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \delta_{k-1} & \alpha_{k-1} & \beta_k \\ & & & \delta_k & \alpha_k \end{pmatrix} \in \mathbb{R}^{k\times k}, \qquad \tilde{T}_k := \begin{pmatrix} T_k \\ \delta_{k+1} e_k^T \end{pmatrix} \in \mathbb{R}^{(k+1)\times k}$$

sowie

$$S_k := \begin{pmatrix} \alpha_1 & \delta_2 & & & \\ \beta_2 & \alpha_2 & \delta_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{k-1} & \alpha_{k-1} & \delta_k \\ & & & \beta_k & \alpha_k \end{pmatrix} \in \mathbb{R}^{k\times k}, \qquad \tilde{S}_k := \begin{pmatrix} S_k \\ \beta_{k+1} e_k^T \end{pmatrix} \in \mathbb{R}^{(k+1)\times k}.$$

Bemerkung: Bei Y. SAAD (1996, S. 206) wird $\delta_{j+1} := (\hat{v}_{j+1}^T \hat{w}_{j+1})^{1/2}$ und $\beta_{j+1} := (\hat{v}_{j+1}^T \hat{w}_{j+1})^{1/2}$ gesetzt, wodurch natürlich die Normalisierungsbedingung $\delta_{j+1}\beta_{j+1} = \hat{v}_{j+1}^T \hat{w}_{j+1}$ erfüllt ist. □

Im folgenden Satz wird gezeigt, dass ohne vorzeitigen Abbruch das oben angegebene Ziel erreicht wird.

Satz 2.7 *Das obige Lanczos Biorthogonalisierungsverfahren breche nicht vorzeitig ab und liefere die im Output genannten Matrizen. Dann bilden die Vektoren $\{v_1, \ldots, v_k\}$ und $\{w_1, \ldots, w_k\}$ ein Biorthogonalsystem, d. h. es ist*

$$V_k^T W_k = I \qquad \text{bzw.} \qquad v_i^T w_j = \delta_{ij} \qquad (1 \le i, j \le k).$$

Ferner ist

$$AV_k = V_{k+1}\tilde{T}_k, \qquad A^T W_k = W_{k+1}\tilde{S}_k, \qquad T_k = S_k^T = W_k^T A V_k.$$

Schließlich bildet $\{v_1, \ldots, v_k\}$ eine Basis von $\mathcal{K}_k(A, v_1)$ und $\{w_1, \ldots, w_k\}$ eine Basis von $\mathcal{K}_k(A^T, w_1)$.

Beweis: Wir zeigen zunächst, dass $\{v_1, \ldots, v_k\}$ und $\{w_1, \ldots, w_k\}$ ein Biorthogonalsystem bilden. Hierzu zeigen wir durch vollständige Induktion nach j, dass $\{v_1, \ldots, v_j\}$ und $\{w_1, \ldots, w_j\}$, $j = 1, \ldots, k$, ein Biorthogonalsystem bilden. Wegen $v_1^T w_1 = 1$ ist dies für $j = 1$ richtig. Wir nehmen an, es sei für $j < k$ richtig. Wegen $v_{j+1}^T w_{j+1} = 1$ bleibt im Induktionsschritt zu zeigen, dass $v_i^T w_{j+1} = 0$ und $v_{j+1}^T w_i = 0$ für $1 \le i \le j$. Wir begnügen uns mit dem Nachweis des ersten Teiles, da der des anderen völlig analog verläuft. Zunächst betrachten wir den Fall $i = j$. Es ist

$$\begin{aligned} v_j^T w_{j+1} &= \frac{1}{\beta_{j+1}} v_j^T (A^T w_j - \alpha_j w_j - \delta_j w_{j-1}) \\ &= \frac{1}{\beta_{j+1}} [(Av_j)^T w_j - \alpha_j] \qquad \text{(Ausnutzung der Induktionsannahme)} \\ &= 0 \qquad \text{(Definition von } \alpha_j\text{)}. \end{aligned}$$

Nun betrachten wir den Fall, dass $i < j$. Dann ist

$$\begin{aligned} v_i^T w_{j+1} &= \frac{1}{\beta_{j+1}} v_i^T (A^T w_j - \alpha_j w_j - \delta_j w_{j-1}) \\ &= \frac{1}{\beta_{j+1}} [(Av_i)^T w_j - \delta_j v_i^T w_{j-1}] \\ &= \frac{1}{\beta_{j+1}} [(\hat{v}_{i+1} - \alpha_i v_i - \beta_i v_{i-1})^T w_j - \delta_j v_i^T w_{j-1}] \\ &= \frac{1}{\beta_{j+1}} [(\delta_{i+1} v_{i+1} - \alpha_i v_i - \beta_i v_{i-1})^T w_j - \delta_j v_i^T w_{j-1}]. \end{aligned}$$

Für $i < j - 1$ verschwindet wegen der Induktionsannahme jedes der obigen inneren Produkte. Für $i = j - 1$ ist dagegen

$$v_{j-1}^T w_{j+1} = \frac{1}{\beta_{j+1}} [(\delta_j v_j - \alpha_{j-1} v_{j-1} - \beta_{j-1} v_{j-2})^T w_j - \delta_j \underbrace{v_{j-1}^T w_{j-1}}_{=1}]$$

$$\begin{aligned} &= \frac{1}{\beta_{j+1}}[\delta_j v_j^T w_j - \delta_j] \\ &= 0. \end{aligned}$$

Damit ist der Induktionsbeweis abgeschlossen.

Nun kommen wir zum Nachweis der restlichen Relationen. Hierzu vergleichen wir die j-ten Spalten der jeweiligen Matrizen. Es ist

$$\begin{aligned} V_{k+1}\tilde{T}_k e_j &= V_{k+1}(\beta_j e_{j-1} + \alpha_j e_j + \delta_{j+1} e_{j+1}) \\ &= \beta_j v_{j-1} + \alpha_j v_j + \delta_{j+1} v_{j+1} \\ &= \beta_j v_{j-1} + \alpha_j v_j + \hat{v}_{j+1} \\ &= A v_j \\ &= A V_k e_j, \end{aligned}$$

also $AV_k = V_{k+1}\tilde{T}_k$. Entsprechend zeigt man, dass $A^T W_k = W_{k+1}\tilde{S}_k$. Schließlich ist

$$\begin{aligned} W_k^T A V_k &= W_k^T V_{k+1} \tilde{T}_k \\ &= W_k^T \begin{pmatrix} V_k & v_{k+1} \end{pmatrix} \begin{pmatrix} T_k \\ \delta_{k+1} e_k^T \end{pmatrix} \\ &= \begin{pmatrix} I & 0 \end{pmatrix} \begin{pmatrix} T_k \\ \delta_{k+1} e_k^T \end{pmatrix} \\ &= T_k. \end{aligned}$$

Hierbei haben wir benutzt, dass $W_k^T V_k = I$ und $W_k^T v_{k+1} = 0$.

Zu zeigen bleibt, dass $\{v_1, \ldots, v_k\}$ bzw. $\{w_1, \ldots, w_k\}$ eine Basis von $\mathcal{K}_k(A, v_1)$ bzw. $\mathcal{K}_k(A^T, w_1)$ ist. Zunächst ist klar, dass $\{v_1, \ldots, v_k\}$ und $\{w_1, \ldots, w_k\}$ jeweils linear unabhängige Vektoren sind, da es sich um ein Biorthogonalsystem handelt. Durch vollständige Induktion nach j zeigt man ferner leicht, dass $v_j \in \mathcal{K}_j(A, v_1) \subset \mathcal{K}_k(A, v_1)$ und entsprechend $w_j \in \mathcal{K}_j(A^T, w_1) \subset \mathcal{K}_k(A^T, w_1)$, $j = 1, \ldots, k$. Damit ist der Satz schließlich bewiesen. □

Bemerkung: Der obige Algorithmus bricht im j-ten Schritt ab, wenn $\hat{v}_{j+1}^T \hat{w}_{j+1} = 0$. Ein Grund hierfür könnte sein, dass einer der Vektoren $\hat{v}_{j+1}$ oder $\hat{w}_{j+1}$ verschwindet. Angenommen, es ist $\hat{v}_{j+1} = 0$. Dann ist $\mathcal{K}_j(A, v_1)$ ein unter A invarianter Teilraum, entsprechend ist für $\hat{w}_{j+1} = 0$ der Krylov-Teilraum $\mathcal{K}_j(A^T, w_1)$ unter A^T invariant. Wir kommen hierauf im Zusammenhang mit der Lösung linearer Gleichungssysteme zurück. Insgesamt kann man diese Situation als "lucky breakdowns" beschreiben. Andererseits ist es auch möglich, dass $\hat{v}_{j+1}^T \hat{w}_{j+1} = 0$ ohne dass einer der beiden Faktoren in diesem inneren Produkt verschwinden. Dies wird ein "serious breakdown" genannt. Es würde zu weit gehen, hierauf näher einzugehen. Hinweise findet man bei Y. Saad (1996, S. 208 ff.). □

5.2.9 BiCG, QMR

BiCG (**Bic**onjugate **G**radient) ist ein Verfahren zur Lösung des linearen Gleichungssystems $Ax = b$, wobei $A \in \mathbb{R}^{n \times n}$ nicht notwendig symmetrisch ist, welches dem

Lanczos-Verfahren im symmetrischen Fall entspricht. Grundlage ist das folgende einfache Lemma.

Lemma 2.8 *Gegeben sei das lineare Gleichungssystem $Ax = b$ mit der nichtsingulären, nicht notwendig symmetrischen Matrix $A \in \mathbb{R}^{n\times n}$. Sei $x_0 \in \mathbb{R}^n$ eine Näherungslösung, $r_0 := b - Ax_0 \neq 0$ und $v_1 := r_0/\|r_0\|_2$. Sei $w_1 \in \mathbb{R}^n$ ein Vektor mit $v_1^T w_1 = 1$, z. B. $w_1 := v_1$. Das Lanczos Biorthogonalisierungsverfahren breche nicht vorzeitig ab und liefere die im Output des Verfahrens genannten Matrizen, deren Eigenschaften in Satz 2.7 formuliert wurden. Ist die Tridiagonalmatrix T_k nichtsingulär, so ist*

$$x_k := x_0 + V_k T_k^{-1}(\|r_0\|_2\, e_1)$$

eine Lösung der Aufgabe, ein $x \in \mathbb{R}^n$ mit

$$x \in x_0 + \mathcal{K}_k(A, v_1), \qquad b - Ax \perp \mathcal{K}_k(A^T, w_1)$$

zu finden.

Beweis: Wegen $R(V_k) = \mathcal{K}_k(A, v_1)$ ist es klar, dass $x_k := x_0 + V_k T_k^{-1}(\|r_0\|_2\, e_1)$ ein Element von $x_0 + \mathcal{K}_k(A, v_1)$ ist. Weiter ist

$$W_k^T(b - Ax_k) = W_k^T[r_0 - AV_k T_k^{-1}(\|r_0\|_2\, e_1)] = W_k^T r_0 - \|r_0\|_2\, e_1 = 0,$$

wobei wir berücksichtigt haben, dass wegen der Biorthogonalität

$$W_k^T r_0 = \|r_0\|_2\, W_k^T v_1 = \|r_0\|_2\, e_1.$$

Wegen

$$(\mathcal{K}_k(A^T, w_1))^\perp = (R(W_k))^\perp = N(W_k^T)$$

ist dann auch $b - Ax_k \perp \mathcal{K}_k(A^T, w_1)$ und der Satz bewiesen. □

Bemerkung: Angenommen, das Lanczos-Biorthogonalisierungsverfahren bricht im j-ten Schritt ab, da $\hat{v}_{j+1} = 0$. Dann ist $x_j := x_0 + V_j T_j^{-1}(\|r_0\|_2\, e_1)$ Lösung von $Ax = b$, so dass man zu Recht von einem "lucky breakdown" sprechen kann. Denn wegen $\hat{v}_{j+1} = 0$ ist $\mathcal{K}_j(A, v_1)$ ein unter A invarianter Teilraum, genauer ist $AV_j = V_j T_j$. Dann ist aber

$$Ax_j = Ax_0 + AV_j T_j^{-1}(\|r_0\|_2\, e_1) = Ax_0 + V_j(\|r_0\|_2\, e_1) = Ax_0 + r_0 = b,$$

also x_j schon die gesuchte Lösung. □

Bemerkung: Zum letzten Lemma kann man sozusagen ein duales Ergebnis angeben. Denkt man sich nämlich ein weiteres Gleichungssystem $A^T x^* = b^*$ gegeben, ist x_0^* eine Näherungslösung und $r_0^* := b^* - A^T x_0^* \neq 0$, $w_1 := r_0^*/\|r_0^*\|_2$ und v_1 mit $v_1^T w_1 = 1$, so ist

$$x_k^* := x_0^* + W_k T_k^{-T}(\|r_0^*\|_2\, e_1)$$

eine Lösung der Aufgabe

$$x^* \in x_0^* + \mathcal{K}_k(A^T, w_1), \qquad b^* - A^T x^* \perp \mathcal{K}_k(A, v_1).$$

Der Beweis ist völlig analog dem obigen. □

Jetzt kann man ähnlich wie beim symmetrischen Lanczos-Verfahren vorgehen. Ziel ist es, ein Verfahren zu entwickeln, das dem CG-Verfahren im symmetrischen (und positiv definiten) Fall ähnelt und welches, wenn man will, auch gleich noch eine Näherungslösung zu einer dualen Aufgabe $A^T x^* = b^*$ findet. Hierzu machen wir die Annahme, dass die Tridiagonalmatrix T_k eine LU-Zerlegung besitzt, dass also $T_k = L_k U_k$ mit einer unteren Dreiecksmatrix L_k mit Einsen in der Diagonalen und einer (nichtsingulären) oberen Dreiecksmatrix U_k. Dann ist

$$x_k = x_0 + P_k z_k, \qquad x_k^* = x_0^* + P_k^* z_k^*,$$

wobei wir zur Abkürzung

$$P_k = (\ p_0 \ \cdots \ p_{k-1}\ ) := V_k R_k^{-1}, \qquad P_k^* := (\ p_0^* \ \cdots \ p_{k-1}^*\ ) := W_k L_k^{-T}$$

sowie

$$z_k := L_k^{-1}(\|r_0\|_2\, e_1), \qquad z_k^* := U_k^{-T}(\|r_0^*\|_2\, e_1)$$

setzen. Wegen

$$(P_k^*)^T A P_k = L_k^{-1} W_k^T A V_k U_k^{-1} = L_k^{-1} T_k U_k^{-1} = I$$

sind die Spalten von P_k^* und Spalten von P_k also A-konjugiert. Weiter ist das Residuum $r_k := b - Ax_k$ ein Vielfaches von v_{k+1} und entsprechend $r_k^* := b^* - A^T x_k^*$ ein Vielfaches von w_{k+1}, wobei die Argumentation genau der beim symmetrischen Lanczos entspricht. Insbesondere bilden die Residuen ein (unnormiertes) Biorthogonalsystem. Wegen $P_k R_k = V_k$ ist p_{k-1} eine Linearkombination aus v_k und p_{k-2}, wobei der Koeffizient von v_k nicht verschwindet, bzw. eine Linearkombination von r_{k-1} und p_{k-2}. Entsprechendes gilt wegen $P_k^* L_k^T = W_k$ für die Richtung p_{k-1}^*. Aus diesen Gründen machen wir den folgenden Ansatz für ein Verfahren, bei dem (implizit) angenommen wird, dass auch noch ein Gleichungssystem $A^T x^* = b^*$ zu lösen ist. Ist dies nicht der Fall, so sind entsprechende Teile im Algorithmus zu modifizieren bzw. wegzulassen:

- Gegeben Näherungen x_0 für $Ax = b$ und x_0^* für $A^T x^* = b^*$. Berechne die Defekte $r_0 := b - Ax_0$ und $r_0^* := b^* - A^T x_0^*$. Es sei $r_0^T r_0^* \neq 0$.
- Setze $p_0 := r_0$, $p_0^* := r_0^*$.
- Für $k = 0, 1, \ldots$:
  - Berechne
    $$x_{k+1} := x_k + \alpha_k p_k, \qquad x_{k+1}^* := x_k^* + \alpha_k p_k^*$$
    mit geeignetem α_k.
  - Berechne die neuen Defekte
    $$r_{k+1} := r_k - \alpha_k A p_k, \qquad r_{k+1}^* := r_k^* - \alpha A^T p_k^*.$$

– Berechne die neuen Richtungen

$$p_{k+1} := r_{k+1} + \beta_k p_k, \qquad p^*_{k+1} := r^*_{k+1} + \beta_k p^*_k$$

mit geeignetem β_k.

Die Forderungen an die Parameter $\alpha_0, \alpha_1, \ldots$ sowie $\beta_0, \beta_1, \ldots$ sind dann naheliegenderweise:

- Die Residuen $r_0, r_1 \ldots$ und $r^*_0, r^*_1, \ldots$ bilden ein (unnormiertes) Bidiagonalsystem, d. h. es ist
$$r_i^T r^*_j = 0, \qquad i \neq j.$$
- Die Richtungen $p^*_0, p^*_1, \ldots$ und $p_0, p_1, \ldots$ sind A-konjugiert, d. h. es ist
$$(p^*_i)^T A p_j = 0, \qquad i \neq j.$$

Hieraus erhält man die Parameter α_k und β_k ähnlich wie beim CG-Verfahren. Aus

$$0 = r_{k+1}^T r^*_k = (r_k - \alpha_k A p_k)^T r^*_k$$

erhält man zunächst

$$\alpha_k = \frac{r_k^T r^*_k}{(Ap_k)^T r^*_k}.$$

Ferner ist

$$(Ap_k)^T r^*_k = (Ap_k)^T (p^*_k - \beta_{k-1} p^*_{k-1}) = (Ap_k)^T p^*_k,$$

und damit

$$\alpha_k = \frac{r_k^T r^*_k}{(Ap_k)^T p^*_k}.$$

Aus

$$0 = (p^*_{k+1})^T A p_k = (r^*_{k+1} + \beta_k p^*_k)^T A p_k$$

erhält man zunächst

$$\beta_k = -\frac{(r^*_{k+1})^T A p_k}{(p^*_k)^T A p_k}.$$

Wegen

$$Ap_k = -\frac{1}{\alpha_k}(r_{k+1} - r_k)$$

ist

$$\beta_k = \frac{1}{\alpha_k} \frac{(r^*_{k+1})^T (r_{k+1} - r_k)}{(p^*_k)^T A p_k} = \frac{r_{k+1}^T r^*_{k+1}}{r_k^T r^*_k}.$$

Das BiCG-Verfahren lautet damit schließlich:

- Gegeben Näherungen x_0 für $Ax = b$ und x^*_0 für $A^T x^* = b^*$. Berechne die Defekte $r_0 := b - Ax_0$ und $r^*_0 := b^* - A^T x^*_0$. Es sei $r_0^T r^*_0 \neq 0$.
- Setze $p_0 := r_0$, $p^*_0 := r^*_0$.

- Für $k = 0, 1, \ldots$:
  - Berechne $\alpha_k := r_k^T r_k^* / (Ap_k)^T p_k^*$.
  - Berechne $x_{k+1} := x_k + \alpha_k p_k$, $x_{k+1}^* := x_k^* + \alpha_k p_k^*$.
  - Berechne $r_{k+1} := r_k - \alpha_k A p_k$, $r_{k+1}^* := r_k^* - \alpha_k A^T p_k^*$.
  - Berechne $\beta_k := r_{k+1}^T r_{k+1}^* / r_k^T r_k^*$.
  - Berechne $p_{k+1} := r_{k+1} + \beta_k p_k$, $p_{k+1}^* := r_{k+1}^* + \beta_k p_k^*$.

QMR (**Q**uasi-**m**inimal **r**esidual)[6] ist ebenfalls ein Verfahren zur Lösung eines linearen Gleichungssystems $Ax = b$ mit nicht notwendig symmetrischer Matrix $A \in \mathbb{R}^{n\times n}$. Wieder sei x_0 eine Näherung und $r_0 := b - Ax_0$ das Anfangsresiduum. Man setze $v_1 := r_0/\|r_0\|_2$ und gewinne mit einem w_1 mit $v_1^T w_1 = 1$ und dem Lanczos Biorthogonalisierungsverfahren den dort angegebenen Output. Die folgende Idee entspricht völlig der des GMRES, wobei dort allerdings entscheidend ausgenutzt wurde, dass man durch das Arnoldi-Verfahren eine *Orthogonalbasis* des entsprechenden Krylov-Raumes berechnet hat.

Für ein beliebiges

$$x = x_0 + V_k y \in x_0 + \mathcal{K}_k(A, v_1)$$

ist dann

$$\begin{aligned} b - Ax &= b - A(x_0 + V_k y) \\ &= r_0 - AV_k y \\ &= \|r_0\|_2\, v_1 - V_{k+1}\tilde{T}_k y \\ &= V_{k+1}(\|r_0\|_2\, e_1 - \tilde{T}_k y). \end{aligned}$$

Wären die Spalten von V_{k+1} orthonormal, so wäre

$$\|b - Ax\|_2 = \|\|r_0\|_2\, e_1 - \tilde{T}_k y\|_2.$$

Auch wenn dies nicht der Fall ist, scheint es sinnvoll zu sein, die Lösung y_k des linearen Ausgleichsproblems

$$\text{Minimiere} \quad \|\|r_0\|_2\, e_1 - \tilde{T}_k y\|_2, \quad y \in \mathbb{R}^k,$$

zu bestimmen und als neue Näherung $x_k := x_0 + V_k y_k$ zu wählen. Entsprechend GMRES (dort war die Koeffizientenmatrix im linearen Ausgleichsproblem eine obere Hessenberg-Matrix) wird man auch hier $(\ \tilde{T}_k \quad \|r_0\|_2\, e_1\ )$ von links mit Givens-Rotationen multiplizieren, um die Berechnung von y_k auf die Lösung eines linearen Gleichungssystems mit einer oberen Dreiecksmatrix als Koeffizientenmatrix zurückzuführen. Wie beim Übergang vom symmetrischen Lanczos zum direkten Lanczos (oder auch der Entwicklung von BiCG) kann man auch eine Iterationsvorschrift angeben, wie

[6] Sucht man im Netz nach QMR, so erhält man QMR auch als Abkürzung für Quality Management (Media) Resources, Quick Medical Reference. Das Verfahrens ist verhältnismäßig kürzlich angegeben worden, und zwar von R. W. Freund, N. M. Nachtigal (1991) "QMR: a quasi-minimal residual method for non-Hermitian linear systems." Numer. Math. 60, 315–339.

x_k aus x_{k-1} zu berechnen ist. Dies wird ausfühlich bei A. GREENBAUM (1997, S. 80 ff.) und etwas kürzer bei Y. SAAD (1996, S. 213) beschrieben, wir gehen kurz darauf ein.

Mit Givens-Rotationen $G_{i,i+1} = G_{i,i+1}(c_i, s_i)$, $i = 1, \ldots, k$, wird

$$G_{k,k+1} \cdots G_{12}(\begin{array}{cc} \tilde{T}_k & \|r_0\|_2 \, e_1 \end{array}) = (\begin{array}{cc} \tilde{R}_k & \tilde{g}_k \end{array}),$$

wobei

$$\tilde{R}_k = \begin{pmatrix} R_k \\ 0^T \end{pmatrix}, \qquad \tilde{g}_k = \begin{pmatrix} g_k \\ \gamma_{k+1} \end{pmatrix},$$

und $R_k \in \mathbb{R}^{k \times k}$ eine nichtsinguläre obere Dreiecksmatrix ist, bei der offenbar nur die Hauptdiagonale und die beiden oberen Nebendiagonalen besetzt sind. Also hat R_k die Form

$$R_k = \begin{pmatrix} \rho_1 & \sigma_2 & \tau_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \tau_k \\ & & & \ddots & \sigma_k \\ & & & & \rho_k \end{pmatrix}.$$

Dann ist $y_k = R_k^{-1} g_k$ die Lösung des obigen linearen Ausgleichsproblems und folglich

$$x_k = x_0 + V_k R_k^{-1} g_k = x_0 + P_k g_k,$$

wobei wir zur Abkürzung

$$P_k := (\begin{array}{ccc} p_0 & \cdots & p_{k-1} \end{array}) := V_k R_k^{-1}$$

gesetzt haben. Daher ist

$$x_k = x_0 + (\begin{array}{cc} P_{k-1} & p_{k-1} \end{array}) \begin{pmatrix} g_{k-1} \\ \gamma_k \end{pmatrix} = x_0 + P_{k-1} g_{k-1} + \gamma_k p_{k-1} = x_{k-1} + \gamma_k p_{k-1}.$$

Durch Vergleich der k-ten Spalte von $P_k R_k = V_k$ erhält man

$$p_{k-1} = \frac{1}{\rho_k}(v_k - \sigma_k p_{k-2} - \tau_k p_{k-3}).$$

Weiter müssen wir uns überlegen, wie man aus einer QR-Zerlegung von $\tilde{T}_{k-1}$ eine solche von $\tilde{T}_k$ berechnet. Zur Berechnung der QR-Zerlegung von $\tilde{T}_k$ werden k Givens-Rotationen benötigt. Die ersten $k-1$ sind alleine durch $\tilde{T}_{k-1}$ bestimmt, von diesen wiederum braucht man sich nur die letzten beiden zu merken und die letzte Spalte von $\tilde{T}_k$ damit zu multiplizieren. Erst die letzte Rotation dient dazu, das Element in der Position $(k+1, k)$ zu annullieren.

Insgesamt resultiert[7] das folgende Verfahren, wobei wir mögliche "breakdowns" nicht abfangen. Wir benutzen wieder die Funktion givrot (siehe Unterabschnitt 3.2.3), die zu einem Paar (α, β) reeller Zahlen ein Tripel (c, s, γ) mit

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \gamma \\ 0 \end{pmatrix}, \qquad c^2 + s^2 = 1$$

bestimmt.

[7] Siehe A. GREENBAUM (1997, S. 83), wobei wir nur etwas andere Bezeichnungen benutzen.

- Sei x_0 Näherungslösung von $Ax = b$, $r_0 := b - Ax \neq 0$ und $v_1 := r_0/\|r_0\|_2$. Wähle w_1 mit $v_1^T w_1 = 1$, z. B. $w_1 := v_1$.

- Setze $\beta_1 := 0$, $\delta_1 := 0$ sowie $v_0 := 0$, $w_0 := 0$. Ferner $\sigma_1 := 0$, $\tau_1 := 0$ und $\tau_2 := 0$. Setze $\gamma_1 := \|r_0\|_2$.

- Für $k = 1, 2, \ldots$:

  - $\alpha_k := (Av_k)^T w_k$.
  - $\hat{v}_{k+1} := Av_k - \alpha_k v_k - \beta_k v_{k-1}$, $\hat{w}_{k+1} := A^T w_k - \alpha_k w_k - \delta_k w_{k-1}$.
  - Falls $\hat{v}_{k+1}^T \hat{w}_{k+1} = 0$, dann: STOP, andernfalls wähle man (von Null verschiedene) reelle Zahlen $\delta_{k+1}, \beta_{k+1}$ mit $\delta_{k+1}\beta_{k+1} = \hat{v}_{k+1}^T \hat{w}_{k+1}$.
  - $v_{k+1} := \hat{v}_{k+1}/\delta_{k+1}$, $w_{k+1} := \hat{w}_{k+1}/\beta_{k+1}$.
  - Falls $k > 2$, dann: Berechne

$$\begin{pmatrix} \tau_k \\ \beta_k \end{pmatrix} := \begin{pmatrix} c_{k-2} & s_{k-2} \\ -s_{k-2} & c_{k-2} \end{pmatrix} \begin{pmatrix} 0 \\ \beta_k \end{pmatrix}.$$

  - Falls $k > 1$, dann: Berechne

$$\begin{pmatrix} \sigma_k \\ \alpha_k \end{pmatrix} := \begin{pmatrix} c_{k-1} & s_{k-1} \\ -s_{k-1} & c_{k-1} \end{pmatrix} \begin{pmatrix} \beta_k \\ \alpha_k \end{pmatrix}.$$

  - Berechne $(c_k, s_k, \rho_k) := \text{givrot}(\alpha_k, \delta_{k+1})$.
  - Berechne

$$\begin{pmatrix} \gamma_k \\ \gamma_{k+1} \end{pmatrix} := \begin{pmatrix} c_k & s_k \\ -s_k & c_k \end{pmatrix} \begin{pmatrix} \gamma_k \\ 0 \end{pmatrix}.$$

  - Berechne

$$p_{k-1} := \frac{1}{\rho_k}(v_k - \sigma_k p_{k-2} - \tau_k p_{k-3}).$$

  - Berechne

$$x_k := x_{k-1} + \gamma_k p_{k-1}.$$

Bemerkung: Wir haben gesehen, dass GMRES und QMR fast auf demselben Prinzip aufgebaut sind, nämlich den Defekt der euklidischen Norm nach auf einem Krylov-Teilraum zu minimieren. Im ersten Fall steht eine Orthonormalbasis des entsprechenden Krylov-Raumes zur Verfügung und die Aufgabe wird zu einem linearen Ausgleichsproblem mit einer oberen Hessenberg-Matrix als Koeffizientenmatrix. Im zweiten Fall kann, da die Basis des Krylov-Teilraumes i. allg. nicht orthonormal ist, der Defekt nur näherungsweise minimiert werden. Das zugehörige lineare Ausgleichsproblem hat eine Tridiagonalmatrix als Koeffizientenmatrix, was zu erheblichen Speicherplatzeinsparungen führt. Während man bei GMRES i. allg. auf einen Restart angewiesen ist, ist dies bei QMR nicht der Fall. □

5.2.10 CGS, BiCGSTAB

BiCGS und QMR benutzen Multiplikationen sowohl mit der Koeffizientenmatrix A des gegebenen linearen Gleichungssystems als auch mit A^T. Dies bedeutet einen Extraaufwand, außerdem ist die (effiziente) Multiplikation mit A^T oft wesentlich schwieriger als die mit A. Daher ist es wünschenswert, ein Iterationsverfahren zu entwickeln, welches mit Multiplikationen mit A auskommt, jedenfalls dann, wenn neben $Ax = b$ nicht auch noch ein duales System $A^T x^* = b^*$ zu lösen ist[8].

Zur Motivation von CGS (**C**onjugate **G**radient **S**quared) erinnern wir noch einmal an BiCG. Es lautet:

- Gegeben Näherungen x_0 für $Ax = b$ und x_0^* für $A^T x^* = b^*$. Berechne die Defekte $r_0 := b - Ax_0$ und $r_0^* := b^* - A^T x_0^*$. Es sei $r_0^T r_0^* \neq 0$.
- Setze $p_0 := r_0$, $p_0^* := r_0^*$.
- Für $k = 0, 1, \ldots$:
  - Berechne $\alpha_k := r_k^T r_k^* / (Ap_k)^T p_k^*$.
  - Berechne $x_{k+1} := x_k + \alpha_k p_k$, $x_{k+1}^* := x_k^* + \alpha_k p_k^*$.
  - Berechne $r_{k+1} := r_k - \alpha_k A p_k$, $r_{k+1}^* := r_k^* - \alpha_k A^T p_k^*$.
  - Berechne $\beta_k := r_{k+1}^T r_{k+1}^* / r_k^T r_k^*$.
  - Berechne $p_{k+1} := r_{k+1} + \beta_k p_k$, $p_{k+1}^* := r_{k+1}^* + \beta_k p_k^*$.

Hierbei sind wir davon ausgegangen, dass auch noch das Gleichungssystem $A^T x^* = b^*$ zu lösen ist. Ist dies nicht der Fall, so wähle man r_0^* beliebig und lasse das Updaten von x_k^* fort.

Offenbar gibt es Polynome $\phi_k \in \Pi_k$ mit $\phi_k(0) = 1$ sowie $\pi_k \in \Pi_k$ derart, dass

$$r_k = \phi_k(A) r_0, \qquad r_k^* = \phi_k(A^T) r_0^*$$

und

$$p_k = \pi_k(A) r_0, \qquad p_k^* = \pi_k(A^T) r_0^*.$$

Dies erkennt man sofort, wenn man die Rekursionsformeln

$$\phi_{k+1}(t) = \phi_k(t) - \alpha_k t \pi_k(t), \qquad \pi_{k+1}(t) = \phi_{k+1}(t) + \beta_k \pi_k(t)$$

berücksichtigt. Dann ist

$$\alpha_k = \frac{r_k^T r_k^*}{(Ap_k)^T p_k^*} = \frac{(\phi_k(A) r_0)^T \phi_k(A^T) r_0^*}{(A\pi_k(A) r_0)^T \pi_k(A^T) r_0^*} = \frac{(\phi_k^2(A) r_0)^T r_0^*}{(A\pi_k^2(A) r_0)^T r_0^*}.$$

Hieran erkennt man, dass man α_k alleine durch Multiplikationen mit A (und nicht mit A^T) berechnen können müsste. Etwas entsprechendes gilt auch für β_k, wie man aus

$$\beta_k = \frac{(\phi_{k+1}^2(A) r_0)^T r_0^*}{(\phi_k^2(A) r_0)^T r_0^*}$$

[8] Wir halten uns hier ziemlich eng an Y. Saad (1996, S. 214 ff.).

entnimmt. Aus den Rekursionsformeln für (ϕ_k, π_k) erhält man (wir lassen die unabhängige Variable t weg, wenn dies möglich ist)

$$\begin{aligned}\phi_{k+1}^2 &= \phi_k^2 - 2\alpha_k t\phi_k\pi_k + \alpha_k^2 t^2\pi_k^2,\\ \pi_{k+1}^2 &= \phi_{k+1}^2 + 2\beta_k\phi_{k+1}\pi_k + \beta_k^2\pi_k^2.\end{aligned}$$

Wenn nicht die gemischten Terme $\phi_k\pi_k$ und $\phi_{k+1}\pi_k$ auftreten würden, so hätte man schon eine Rekursionsformel für ϕ_k^2 und π_k^2. Der Trick besteht jetzt darin, einen der beiden gemischten Terme, nämlich $\phi_{k+1}\pi_k$, sozusagen als drittes Mitglied in die Rekursion mit aufzunehmen. Den anderen Term $\phi_k\pi_k$ kann man dann aus

$$\phi_k\pi_k = \phi_k(\phi_k + \beta_{k-1}\pi_{k-1}) = \phi_k^2 + \beta_{k-1}\phi_k\pi_{k-1}$$

rekursiv berechnen. Man erhält die folgenden Rekursionsformeln:

$$\begin{aligned}\phi_{k+1}^2 &= \phi_k^2 - \alpha_k t(2\phi_k^2 + 2\beta_{k-1}\phi_k\pi_{k-1} - \alpha_k t\pi_k^2),\\ \phi_{k+1}\pi_k &= \phi_k^2 + \beta_{k-1}\phi_k\pi_{k-1} - \alpha_k t\pi_k^2,\\ \pi_{k+1}^2 &= \phi_{k+1}^2 + 2\beta_k\phi_{k+1}\pi_k + \beta_k^2\pi_k^2.\end{aligned}$$

Diese Rekursionsformeln sind Grundlage von CGS. Definieren wir

$$r_k := \phi_k^2(A)r_0, \qquad p_k := \pi_k^2(A)r_0, \qquad q_k := \phi_{k+1}(A)\pi_k(A)r_0,$$

so erhält man aus obigen Rekursionsformeln

$$\begin{aligned}r_{k+1} &= r_k - \alpha_k A(2r_k + 2\beta_{k-1}q_{k-1} - \alpha_k Ap_k),\\ q_k &= r_k + \beta_{k-1}q_{k-1} - \alpha_k Ap_k,\\ p_{k+1} &= r_{k+1} + 2\beta_k q_k + \beta_k^2 p_k.\end{aligned}$$

Ferner definieren wir noch

$$u_k := r_k + \beta_{k-1}q_{k-1}.$$

Berücksichtigt man, dass

$$\begin{aligned}2r_k + 2\beta_{k-1}q_{k-1} - \alpha_k Ap_k &= u_k + q_k,\\ q_k &= u_k - \alpha_k Ap_k,\\ p_{k+1} &= u_{k+1} + \beta_k(q_k + \beta_k p_k),\end{aligned}$$

so erhält man die folgende Form von CGS:

- Sei x_0 eine Näherung von $Ax = b$, $r_0 := b - Ax_0 \neq$. Sei r_0^* mit $r_0^T r_0^* \neq 0$ beliebig.
- Setze $p_0 := r_0$, $u_0 := r_0$.
- Für $k = 0, 1, \ldots$:
  - Berechne $\alpha_k := r_k^T r_0^*/(Ap_k)^T r_0^*$, $q_k := u_k - \alpha_k Ap_k$.
  - Berechne $x_{k+1} := x_k + \alpha_k(u_k + q_k)$, $r_{k+1} := r_k - \alpha_k A(u_k + q_k)$.

- Berechne $\beta_k := r_{k+1}^T r_0^* / r_k^T r_0^*$
- Berechne $u_{k+1} := r_{k+1} + \beta_k q_k$, $p_{k+1} := u_{k+1} + \beta_k(q_k + \beta_k p_k)$.

Offenbar ist r_k der Defekt zu x_k, $k = 0, 1, \ldots$, ferner treten keine Matrix-Vektor-Produkte mit A^T auf, stattdessen wird "nützlichere" Arbeit verrichtet. Durch das Quadrieren der Polynome ist der Algorithmus allerdings anfälliger gegen Rundungsfehler als BiCG. Diesen Nachteil versucht BiCGSTAB (**Bi**conjugate **G**radient **Stab**ilized) aufzuheben. Hier wird eine Folge $\{x_k\}$ konstruiert, deren Defekte sich in der Form

$$r_k = \phi_k(A)\psi_k(A)r_0$$

darstellen lassen, wobei $\phi_k \in \Pi_k$ wie oben beim CGS gegeben ist, und ψ_{k+1} durch die einfache Rekursionsformel

$$\psi_{k+1}(t) := (1 - \omega_k t)\psi_k(t)$$

berechnet wird. Auf die Wahl von ω_k wird erst am Schluss der Herleitung von BiCGSTAB eingegangen. Wir nehmen an, dass $\psi_0(t)$ konstant und damit ψ_k ein Polyno0m vom Grad $\leq k$ ist. Jetzt besteht die Aufgabe darin, für $\phi_k\psi_k$ eine Rekursionsformel zu entwickeln. Es ist

$$\begin{aligned} \phi_{k+1}\psi_{k+1} &= (1 - \omega_k t)\phi_{k+1}\psi_k \\ &= (1 - \omega_k t)(\phi_k\psi_k - \alpha_k t\psi_k\pi_k), \end{aligned}$$

so dass eine Rekursionsformel für $\psi_k\pi_k$ gefunden werden muss. Auch hier ist π_k natürlich wie in der Herleitung von CGS gegeben. Hierzu beachten wir, dass

$$\begin{aligned} \psi_k\pi_k &= \psi_k(\phi_k + \beta_{k-1}\pi_{k-1}) \\ &= \phi_k\psi_k + \beta_{k-1}(1 - \omega_{k-1}t)\psi_{k-1}\pi_{k-1}. \end{aligned}$$

Definiert man nun

$$r_k := \phi_k(A)\psi_k(A)r_0, \qquad p_k := \psi_k(A)\pi_k(A)r_0,$$

so erhält man hierfür die Rekursionsformeln

$$r_{k+1} = (I - \omega_k A)(r_k - \alpha_k A p_k), \qquad p_{k+1} = r_{k+1} + \beta_k(I - \omega_k A)p_k.$$

Jetzt müssen noch die Konstanten α_k und β_k berechnet werden. Im "originalen" BiCG ist $\beta_k = \rho_{k+1}/\rho_k$ mit

$$\rho_k := (\phi_k(A)r_0)^T\phi_k(A^T)r_0^* = (\phi_k^2(A)r_0)^T r_0^*.$$

Direkt ist ρ_k nicht berechenbar, da weder $\phi_k(A)r_0$ und $\phi_k(A^T)r_0$ noch $\phi_k^2(A)r_0$ zur Verfügung stehen. Andererseits kann, wie wir sehen werden, ρ_k zu

$$\begin{aligned} \tilde{\rho}_k &:= (\phi_k(A)r_0)^T\psi_k(A^T)r_0^* \\ &= (\psi_k(A)\phi_k(A)r_0)^T r_0^* \\ &= (\phi_k(A)\psi_k(A)r_0)^T r_0^* \\ &= r_k^T r_0^* \end{aligned}$$

in Verbindung gesetzt werden. Da $\phi_k(A)r_0$ auf $\mathcal{K}_k(A^T, r_0^*)$ senkrecht steht und $\psi_k \in \Pi_k$, ist

$$\begin{aligned}
\tilde{\rho}_k &:= (\phi_k(A)r_0)^T \psi_k(A^T) r_0^* \\
&= (\phi_k(A)r_0)^T [\eta_1^{(k)} (A^T)^k r_0^* + \eta_2^{(k)} (A^T)^{k-1} r_0^* + \cdots] \\
&= \eta_1^{(k)} (\phi_k(A)r_0)^T (A^T k) r_0^* \\
&= \frac{\eta_1^{(k)}}{\gamma_1^{(k)}} (\phi_k(A)r_0)^T \phi_k(A^T) r_0^* \\
&= \frac{\eta_1^{(k)}}{\gamma_1^{(k)}} \rho_k,
\end{aligned}$$

wobei $\gamma_1^{(k)}$ der führende Koeffizient im Polynom ϕ_k ist. Betrachtet man die Rekursionsformeln für ϕ_{k+1} und ψ_{k+1}, also

$$\phi_{k+1} = \phi_k - \alpha_k t \pi_k = \phi_k - \alpha_k t(\phi_k + \beta_{k-1}\pi_{k-1})$$

und

$$\psi_{k+1} = (1 - \omega_k t)\psi_k,$$

so erkennt man die Gültigkeit der Rekursionsformeln

$$\gamma_1^{(k+1)} = -\alpha_k \gamma_1^{(k)}, \qquad \eta_1^{(k+1)} = -\omega_k \eta_1^{(k)}.$$

Dies liefert

$$\frac{\tilde{\rho}_{k+1}}{\tilde{\rho}_k} = \frac{\omega_k}{\alpha_k} \frac{\rho_{k+1}}{\rho_k},$$

anschließend ist

$$\beta_k = \frac{\rho_{k+1}}{\rho_k} = \frac{\tilde{\rho}_{k+1}}{\tilde{\rho}_k} \frac{\alpha_k}{\omega_k} = \frac{r_{k+1}^T r_0^*}{r_k^T r_0^*} \frac{\alpha_k}{\omega_k}.$$

Bei der Berechnung von α_k berücksichtigen wir, dass $\phi_k(A)r_0$ auf $\mathcal{K}_k(A^T, r_0^*)$ senkrecht steht und die höchsten Koeffizienten von ϕ_k und ψ_k übereinstimmen, was man aus $\phi_0 = 1$, $\pi_0 = 1$ und

$$\phi_{k+1} = \phi_k - \alpha_k t \pi_k, \qquad \pi_{k+1} = \phi_{k+1} + \beta_k \pi_k$$

sofort erkennt. Weiter ist $\pi_k(A)r_0$ offenbar A-konjugiert zu $(A^T)^j r_0^*$, $j = 0, \ldots, k-1$. Damit wird

$$\begin{aligned}
\alpha_k &= \frac{(\phi_k(A)r_0)^T \phi_k(A^T) r_0^*}{(A\pi_k(A)r_0)^T \pi_k(A^T) r_0^*} \\
&= \frac{(\phi_k(A)r_0)^T \phi_k(A^T) r_0^*}{(A\pi_k(A)r_0)^T \phi_k(A^T) r_0^*} \\
&= \frac{(\phi_k(A)r_0)^T \psi_k(A^T) r_0^*}{(A\pi_k(A)r_0)^T \psi_k(A^T) r_0^*} \\
&= \frac{(\psi_k(A)\phi_k(A)r_0)^T r_0^*}{(A\psi_k(A)\pi_k(A)r_0)^T r_0^*}
\end{aligned}$$

$$= \frac{\tilde{\rho}_k}{(Ap_k)^T r_0^*}$$
$$= \frac{r_k^T r_0^*}{(Ap_k)^T r_0^*}.$$

Nun kommt es darauf an, den noch freien Parameter ω_k geeignet zu wählen. Naheliegend ist es, ω_k als Lösung der Aufgabe

$$\text{Minimiere} \quad \|(I - \omega A)(r_k - \alpha_k A p_k)\|_2, \quad \omega \in \mathbb{R}$$

zu wählen. Mit der Abkürzung

$$s_k := r_k - \alpha_k A p_k$$

erhält man

$$\omega_k = \frac{s_k^T A s_k}{\|A s_k\|_2^2}.$$

Schließlich muß man sich noch überlegen, wie x_k upzudaten ist, damit r_{k+1} der Defekt in x_{k+1} ist, vorausgesetzt r_k ist der Defekt in x_k. Nun ist

$$r_{k+1} = s_k - \omega_k A s_k = r_k - A(\alpha_k p_k + \omega_k s_k),$$

so dass wir

$$x_{k+1} = x_k + \alpha_k p_k + \omega_k s_k$$

setzen werden. Damit erhalten wir schließlich die folgende Version von BiCGSTAB.

- Sei x_0 eine Näherung von $Ax = b$, $r_0 := b - Ax_0 \neq$. Sei r_0^* mit $r_0^T r_0^* \neq 0$ beliebig.
- Setze $p_0 := r_0$.
- Für $k = 0, 1, \ldots$:
  - $\alpha_k := r_k^T r_0^* / (Ap_k)^T r_0^*$, $s_k := r_k - \alpha_k A p_k$, $\omega_k := (As_k)^T s_k / \|As_k\|_2^2$.
  - $x_{k+1} := x_k + \alpha_k p_k + \omega_k s_k$, $r_{k+1} := s_k - \omega_k A s_k$.
  - $\beta_k := [r_{k+1}^T r_0^* / r_k^T r_0^*]\, \alpha_k / \omega_k$.
  - $p_{k+1} := r_{k+1} + \beta_k (p_k - \omega_k A p_k)$.

5.2.11 MATLAB-Ergänzungen

Das Arnoldi-Verfahren zur Berechnung einer Orthonormalbasis eines Krylov-Teilraumes überträgt sich wörtlich von Pseudocode in MATLAB:

```
function [Q,H]=Arnoldi(A,v,k);
%*************************************************
%Pre:   A is n-by-n matrix
%       v is n-vector, not equal zero.
%       k is natural number
```

```
%Post:  Q is n-by (k+1) Matrix, the first k
%         columns are an orthogonal basis of
%         the Krylov subspace K_k(A,v)
%         H is an (k+1)-by-k upper Hessenbergmatrix
%*************************************************
H=zeros(k+1,k);[n,n]=size(A);Q=zeros(n,k+1);
Q(:,1)=v/norm(v);
for j=1:k
   w=A*Q(:,j);
   for i=1:j
      H(i,j)=Q(:,i)'*w;
      w=w-H(i,j)*Q(:,i);
   end;
   H(j+1,j)=norm(w);
   if H(j+1,j)==0
      return
   end;
   Q(:,j+1)=w/H(j+1,j);
end;
```

Nun kommen wir zu FOM, dem Arnoldi-Verfahren für lineare Gleichungssysteme. Zunächst geben wir eine MATLAB-Funktion an, mit der die Lösung eines linearen Gleichungssystems berechnet werden kann, bei dem die Koeffizientenmatrix eine (nichtsinguläre) obere Hessenberg-Matrix ist.

```
function x=Hessen(H,b);
%*************************************************
%Pre:  H is an n-by-n upper Hessenberg matrix
%      b is an n-vector
%Post: x solves H*x=b
%Es werden die frueher definierten Funktionen Givens
%und UTriSol benutzt.
%*************************************************
[n,n]=size(H);
for k=1:n-1
   [c,s]=Givens(H(k,k),H(k+1,k));
   H(k:k+1,k:n)=[c s;-s c]*H(k:k+1,k:n);
   b(k:k+1)=[c s;-s c]*b(k:k+1);
end;
x=UTriSol(H,b);
```

Es folgt eine MATLAB-Funktion für das Arnoldi-Verfahren bei linearen Gleichungssystemen (FOM).

```
function [x,Q]=Fom(A,b,x,k);
%*************************************************
%Pre:   A is n-by-n matrix
```

```
%      b is n-vector
%      x is initial iterate
%      k is natural number
%Post: x ist Naeherungsloesung von Ax=b
%      Q is n-by-k, the columns of Q being an
%      orthonormal  basis of Krylov subspace K_k
%************************************************
r=b-A*x;Q(:,1)=r/norm(r);H=zeros(k+1,k);
for j=1:k
   w=A*Q(:,j);
   for i=1:j
       H(i,j)=Q(:,i)'*w;
       w=w-H(i,j)*Q(:,i);
   end;
   H(j+1,j)=norm(w);
   if H(j+1,j)==0
        k=j;break
   end;
   Q(:,j+1)=w/H(j+1,j);
end;
Q=Q(:,1:k);
c=zeros(k,1);c(1)=norm(r);
d=Hessen(H(1:k,1:k),c);
x=x+Q*d;
```

Macht man mit dieser einfachen Version von FOM Experimente, so erkennt man, dass die Ergebnisse sehr schlecht sein können.

In MATLAB gibt es die Funktion `gmres`. Mit `help gmres` erfährt man u. a.:

```
 gmres   Generalized Minimum Residual Method.
    X = gmres(A,B) attempts to solve the system of linear equations A*X = B
    for X.  The N-by-N coefficient matrix A must be square and the right
    hand side column vector B must have length N. This uses the unrestarted
    method with MIN(N,10) total iterations.

    X = gmres(AFUN,B) accepts a function handle AFUN instead of the matrix
    A. AFUN(X) accepts a vector input X and returns the matrix-vector
    product A*X. In all of the following syntaxes, you can replace A by
    AFUN.
```

Die Funktion `gmres` ist in MATLAB geschrieben, man kann sie sich daher mittels `type gmres` ansehen. Wir geben eine ähnliche Version aus den sogenannten “Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods” an.

```
function [x,error,iter,flag] = Gmres(A,x,b,M,restrt,max_it,tol)
%*********************************************************************
%  -- Iterative template routine --
```

```
%     Univ. of Tennessee and Oak Ridge National Laboratory
%     October 1, 1993
%     Details of this algorithm are described in "Templates for the
%     Solution of Linear Systems: Building Blocks for Iterative
%     Methods", Barrett, Berry, Chan, Demmel, Donato, Dongarra,
%     Eijkhout, Pozo, Romine, and van der Vorst, SIAM Publications,
%     1993. (ftp netlib2.cs.utk.edu; cd linalg; get templates.ps).
%
% [x,error,iter,flag] = Gmres(A,x,b,M,restrt,max_it,tol)
%
% gmres.m solves the linear system Ax=b using the Generalized Minimal
% residual (GMRESm) method with restarts.
%
% input   A        REAL nonsymmetric positive definite matrix
%         x        REAL initial guess vector
%         b        REAL right hand side vector
%         M        REAL preconditioner matrix
%         restrt   INTEGER number of iterations between restarts
%         max_it   INTEGER maximum number of iterations
%         tol      REAL error tolerance
%
% output  x        REAL solution vector
%         error    REAL error norm
%         iter     INTEGER number of iterations performed
%         flag     INTEGER: 0 = solution found to tolerance
%                           1 = no convergence given max_it
%*********************************************************************
iter = 0;flag = 0;[n,n]=size(A);
bnrm2=norm(b);
if bnrm2==0
   x=zeros(n,1);
   error=0;
   return
end;
r=M\(b-A*x);

error = norm(r)/bnrm2;
if (error<tol) return, end;

k=restrt;
Q(1:n,1:k+1) = zeros(n,k+1);
H(1:k+1,1:k) = zeros(k+1,k);

for iter = 1:max_it
    r=M\(b-A*x);
```

```
    Q(:,1) = r/ norm(r);
    g=zeros(k+1,1);g(1)=norm(r);
    for j=1:k,
        w=M\(A*Q(:,j));
        for i=1:j  H(i,j)= w'*Q(:,i);w = w - H(i,j)*Q(:,i); end;
        H(j+1,j)=norm(w); Q(:,j+1)=w/H(j+1,j);
    end;
    for j=1:k
       [c,s]=Givens(H(j,j),H(j+1,j));
       H(j:j+1,:)=[c s;-s c]*H(j:j+1,:);
       g(j+1)=-s*g(j);g(j)=c*g(j);
    end;
    R=triu(H(1:k,1:k));
    error=abs(g(k+1))/bnrm2;
    y =R\g(1:k);
    x = x + Q(:,1:k)*y;
    if (error<=tol)
       return;
    end;
end;

if ( error > tol ) flag = 1; end;

% END of Gmres.m
```

Wie man am Kommentar erkennt, wird ein sogenannter Präkonditionierer M benutzt. Hierauf gehen wir im folgenden Abschnitt 5.3 ein. Wir benutzen zunächst $M = I$, also keine Präkonditionierung; Als Test führen wir das folgende Programm durch:

```
N=100;
A=sparse(Modell(N));
b=(1/N)^2*ones(N^2,1);
x0=zeros(N^2,1);
max_it=100;
tol=0.0000001;
M=speye(N^2);
tic;
[x,error,iter,flag]=Gmres(A,x0,b,M,100,max_it,tol);
toc;
```

Wir sind erfolgreich (flag=0), es ist `error=7.8753e-10`, es wurden nur drei Iterationen durchgeführt und hierfür wurden 33.9319 Sekunden benötigt. Es handelt sich hier immerhin um ein Gleichungssystem mit 10 000 Gleichungen und ebenso vielen Unbekannten.

In den oben angesprochenen Templates gibt es eine MATLAB-Funktion `cg`. Diese benutzt (wie `gmres`) einen Präkonditionierer. Eine nur geringfügig veränderte Version, in welcher die Bezeichnungen unserer Vorlesung benutzt werden, geben wir nun an.

```
function [x, error, iter, flag] = Cg(A, x, b, M, max_it, tol)

%  -- Iterative template routine --
%     Univ. of Tennessee and Oak Ridge National Laboratory
%     October 1, 1993
%     Details of this algorithm are described in "Templates for the
%     Solution of Linear Systems: Building Blocks for Iterative
%     Methods", Barrett, Berry, Chan, Demmel, Donato, Dongarra,
%     Eijkhout, Pozo, Romine, and van der Vorst, SIAM Publications,
%     1993. (ftp netlib2.cs.utk.edu; cd linalg; get templates.ps).
%
%  [x, error, iter, flag] = Cg(A, x, b, M, max_it, tol)
%
% cg.m solves the symmetric positive definite linear system Ax=b
% using the Conjugate Gradient method with preconditioning.
%
% input   A        REAL symmetric positive definite matrix
%         x        REAL initial guess vector
%         b        REAL right hand side vector
%         M        REAL preconditioner matrix
%         max_it   INTEGER maximum number of iterations
%         tol      REAL error tolerance
%
% output  x        REAL solution vector
%         error    REAL error norm
%         iter     INTEGER number of iterations performed
%         flag     INTEGER: 0 = solution found to tolerance
%                           1 = no convergence given max_it

  flag = 0;iter = 0;[n,n]=size(A); bnrm2=norm(b);
  if bnrm2==0  x=zeros(n,1);  error=0;  return end;

  r = b - A*x;
  error = norm( r ) / bnrm2;
  if ( error < tol ) return, end

  for iter = 1:max_it
     z  = M \ r;  rho = (r'*z);
     if ( iter > 1 ),
        beta = rho / rho_1; p = z + beta*p;
     else
        p = z;
     end
     q = A*p;alpha = rho / (p'*q );x = x + alpha * p;r = r - alpha*q;
     error = norm( r ) / bnrm2;
```

```
    if ( error <= tol ), break, end
    rho_1 = rho;
  end
  if ( error > tol ) flag = 1; end
% END Cg.m
```

Zum Testen benutzen wir wieder das Modellproblem. Genauer verwenden wir das folgende Script-File.

```
N=100;
A=sparse(Modell(N));b=(1/N)^2*ones(N^2,1);
x0=zeros(N^2,1);max_it=1000;tol=0.0000001;
M=speye(N^2);
tic;
[x,error,iter,flag]=Cg(A,x0,b,M,max_it,tol);
toc;
```

Wir sind erfolgreich. Nach 170 Iterationen und 2.2762 Sekunden ist `error=9.5582e-08`. Wie wir am obigen Programm sehen, haben wir keinen Präkonditionierer benutzt.

Zunächst geben wir die in den Templates implementierte Funktion `bicg` an.

```
function [x, error, iter, flag] = bicg(A, x, b, M, max_it, tol)
%*********************************************************************
%  -- Iterative template routine --
%     Univ. of Tennessee and Oak Ridge National Laboratory
%     October 1, 1993
%     Details of this algorithm are described in "Templates for the
%     Solution of Linear Systems: Building Blocks for Iterative
%     Methods", Barrett, Berry, Chan, Demmel, Donato, Dongarra,
%     Eijkhout, Pozo, Romine, and van der Vorst, SIAM Publications,
%     1993. (ftp netlib2.cs.utk.edu; cd linalg; get templates.ps).
%
%  [x, error, iter, flag] = bicg(A, x, b, M, max_it, tol)
%
% bicg.m solves the linear system Ax=b using the
% BiConjugate Gradient Method with preconditioning.
%
% input   A        REAL matrix
%         M        REAL preconditioner matrix
%         x        REAL initial guess vector
%         b        REAL right hand side vector
%         max_it   INTEGER maximum number of iterations
%         tol      REAL error tolerance
%
% output  x        REAL solution vector
%         error    REAL error norm
%         iter     INTEGER number of iterations performed
%         flag     INTEGER: 0 = solution found to tolerance
```

```
%                            1 = no convergence given max_it
%                           -1 = breakdown
%*********************************************************************
  iter = 0;                                % initialization
  flag = 0;

  bnrm2 = norm( b );
  if  ( bnrm2 == 0.0 ), bnrm2 = 1.0; end

  r = b - A*x;
  error = norm( r ) / bnrm2;
  if ( error < tol ) return, end

  r_tld  = r;

  for iter = 1:max_it                      % begin iteration

     z = M \ r;
     z_tld = M' \ r_tld;
     rho   = ( z'*r_tld );
     if ( rho == 0.0 ),
        break
     end

     if ( iter > 1 ),                      % direction vectors
        beta = rho / rho_1;
        p   = z  + beta*p;
        p_tld = z_tld + beta*p_tld;
     else
        p = z;
        p_tld = z_tld;
     end

     q = A*p;                              % compute residual pair
     q_tld = A'*p_tld;
     alpha = rho / (p_tld'*q );

     x = x + alpha*p;                      % update approximation
     r = r - alpha*q;
     r_tld = r_tld - alpha*q_tld;

     error = norm( r ) / bnrm2;            % check convergence
     if ( error <= tol ), break, end

     rho_1 = rho;

  end
```

```
  if ( error <= tol ),                    % converged
     flag =  0;
  elseif ( rho == 0.0 ),                  % breakdown
     flag = -1;
  else
     flag = 1;                            % no convergence
  end

% END bicg.m
```

Wir benutzen eine nur geringfügig veränderte Version an. Der Unterschied besteht vor allem darin, dass wir mit einem Präkonditionierer $M = M_1 M_2$ arbeiten, wobei wir annehmen, dass die Auflösung linearer Gleichungssysteme mit M_1, M_2, M_1^T, M_2^T "einfach" ist. Davon einmal abgesehen sind Input und Output derselbe wie der in obiger Funktion.

```
function [x,error,iter,flag]=Bicg(A,x,b,M1,M2,max_it,tol);

iter=0;flag=0;
bnrm2=norm(b);n=length(x);
if bnrm2==0
   x=zeros(n,1);
   error =0;
   return
end;
r=b-A*x;error=norm(r)/bnrm2;
if (error<tol) return, end;

r_star=r;
for iter=1:max_it
    y=M1\r;z=M2\y;
    y_star=M2'\r_star;z_star=M1'\y_star;
    rho=z'*r_star;
    if (rho==0.0) break, end;
    if (iter>1)
        beta=rho/rho1;
        p=z+beta*p;
        p_star=z_star+beta*p_star;
    else
        p=z;p_star=z_star;
    end;
    q=A*p;q_star=A'*p_star;
    alpha=rho/(p_star'*q);
    x=x+alpha*p;
    r=r-alpha*q;r_star=r_star-alpha*q_star;
    error=norm(r)/bnrm2;
    if (error<tol), break, end;
    rho1=rho;
end;
```

```
if (error<tol),
    flag=0;
elseif (rho==0.0),
    flag=-1;
else
    flag=1;
end;
```

Als Beispiel betrachten wir die Auflösung eines linearen Gleichunggssystems mit einer 479×479 Koeffizientenmatrix (mit 1887 von Null verschiedenen Elementen), wobei die rechte Seite so gewählt ist, dass der Vektor aus lauter Einsen Lösung ist.

```
load west0479;
A=west0479;x0=zeros(479,1);
b=sum(A,2);
[L,U]=luinc(A,1e-6);
tic;
[x,error,iter,flag]=Bicg(A,x0,b,L,U,50,1e-10);
toc;
```

Nach 0.0624 Sekunden erfolgt Abbruch, es ist `error=6.2685e-13`, allerdings ist der absolute Fehler lediglich `2.1872e-04`. Setzen wir allerdings `tol=1e-15`, so vermindert sich der absolute Fehler auf `3.0193e-10`. Ohne Präkonditionierung ist `bicg` nicht erfolgreich. Mit dem Programm

```
N=50;A=sparse(Modell(N));x0=zeros(N^2,1);
b=sum(A,2);
[L,U]=luinc(A,1e-6);
tic;
[x,error,iter,flag]=Bicg(A,x0,b,L,U,50,1e-15);
toc;
```

erhalten wir einen erfolgreichen Abbruch nach 0.1885 Sekunden (ohne Präkonditionierung erfolgt kein erfolgreicher Abbruch innerhalb von 50 Iterationen).

In den Templates ist auch eine MATLAB-Funktion `qmr` angegeben. Hierauf wollen wir nicht mehr eingehen. Es gibt aber auch eine MATLAB-Funktion `qmr`. Nach `help qmr` erfährt man u. a.:

```
 qmr   Quasi-Minimal Residual Method.
    X = qmr(A,B) attempts to solve the system of linear equations A*X=B for
    X. The N-by-N coefficient matrix A must be square and the right hand
    side column vector B must have length N.

    X = qmr(AFUN,B) accepts a function handle AFUN instead of the matrix A.
    AFUN(X,'notransp') accepts a vector input X and returns the
    matrix-vector product A*X while AFUN(X,'transp') returns A'*X. In all
    of the following syntaxes, you can replace A by AFUN.
```

5.2.12 Aufgaben

1. Sei $A \in \mathbb{C}^{n\times n}$. Dann ist A genau dann normal (also $A^H A = AA^H$), wenn ein Polynom $q \in \Pi_{n-1}$ mit $A^H = q(A)$ existiert.

2. Sei $A \in \mathbb{R}^{n\times n}$, $v \in \mathbb{R}^n \setminus \{0\}$ und $k \in \mathbb{N}$. Man betrachte den folgenden Algorithmus, den man als Householder-Arnoldi bezeichnen kann:

   - Setze $z^{(1)} := v$.
   - Für $j = 1, \ldots, k+1$:
     - Bestimme Householder-Matrix $\overline{P}_j \in \mathbb{R}^{(n-j+1)\times(n-j+1)}$ mit
       $$\overline{P}_j (z_j^{(j)}, z_{j+1}^{(j)}, \ldots, z_n^{(j)})^T = (*, 0, \ldots, 0)^T$$
       und setze $P_j := \operatorname{diag}(I_{j-1}, \overline{P}_j)$.
     - Berechne $h_{j-1} := P_j z^{(j)}$.
     - Berechne $q_j := P_1 P_2 \cdots P_j e_j$.
     - Falls $j \le k$, berechne $z^{(j+1)} := P_j P_{j-1} \cdots P_1 A q_j$.

   Man zeige[9] für diesen Algorithmus:

   (a) Es ist
   $$P_k \cdots P_1 (\ v \quad Aq_1 \quad Aq_2 \quad \cdots \quad Aq_k\ ) = (\ h_0 \quad h_1 \quad \cdots \quad h_k\ ).$$

   (b) Die Vektoren $\{q_1, \ldots, q_k\}$ bilden eine Orthonormalbasis des Krylov-Teilraumes $\mathcal{K}_k$, wenn $(h_j)_{j+1} \neq 0$, $j = 1, \ldots, k$.

3. Es seien k Schritte des Arnoldi-Verfahrens zur Berechnung einer Orthonormalbasis des Krylov-Teilraumes $\mathcal{K}(A, r_0)$ durchführbar, der Output sei wie üblich bezeichnet. Es wird vorausgesetzt, dass die obere Hessenberg-Matrix H_k nichtsingulär ist. Seien $G_{j,j+1} = G_{j,j+1}(c_j, s_j)$, $j = 1, \ldots, k$, Givens-Rotationen, die $\tilde{H}_k$ sukzessive in eine obere Dreiecksmatrix transformieren. Durch Anwendung von FOM bzw. GMRES erhalte man die Näherungen x_k^F bzw. x_k^G. Mit r_k^F bzw. r_k^G seien die entsprechenden Defekte bezeichnet. Man zeige:

   (a) Es ist $\|r_k^F\|_2 = |h_{k+1,k}|\,|e_k^T y_k^F|$, wobei $y_k^F := H_k^{-1}(\|r_0\|_2\, e_1)$.

   (b) Es ist
   $$\|r_k^F\|_2 = \|r_k^G\|_2 \sqrt{1 + \frac{h^2}{\xi^2}},$$
   wobei
   $$\xi := (G_{k-1,k} \cdots G_{12} \tilde{H}_k)_{kk}, \qquad h := h_{k+1,k}.$$

4. Es seien k Schritte des Arnoldi-Verfahrens zur Berechnung einer Orthonormalbasis des Krylov-Teilraumes $\mathcal{K}(A, r_0)$ durchführbar, der Output sei wie üblich bezeichnet. Es wird vorausgesetzt, dass die obere Hessenberg-Matrix H_k nichtsingulär ist. Durch Anwendung von FOM bzw. GMRES erhalte man die Näherungen x_k^F bzw. x_k^G. Mit r_k^F bzw. r_k^G seien die entsprechenden Defekte bezeichnet. Man zeige: Ist $x_k^F = x_k^G$, so ist $r_k^F = r_k^G = 0$, d. h. sowohl FOM als auch GMRES liefern die exakte Lösung.

[9] Siehe Y. Saad (1996, S. 150 ff.).

5. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Es sei bekannt, dass A nur $m \le n$ paarweise verschiedene Eigenwerte besitzt. Dann bricht das CG-Verfahren nach höchstens m Schritten ab.

6. Auf das lineare Gleichungssystem $Ax = b$ werde GMRES und QMR angewandt, r_k^G und r_k^Q seien die entsprechenden Residuen im k-ten Schritt. Dann ist

$$\|r_k^Q\|_2 \le \kappa_2(V_{k+1})\,\|r_k^G\|_2,$$

wobei V_{k+1} die durch das Lanczos-Biorthogonalisierungsverfahren berechnete Matrix der Basisvektoren von $\mathcal{K}_{k+1}(A, r_0)$ ist und $\kappa_2(\cdot)$ die Kondition bezüglich der Spektralnorm bedeutet.

Hinweis: Ist $W \in \mathbb{R}^{n\times m}$ eine Matrix mit vollem Spaltenrang m, so ist $\|Wy\|_2 \ge \sigma_{\min}(W)\,\|y\|_2$, wobei $\sigma_{\min}(W)$ den kleinsten singulären Wert von W bedeutet.

7. Gegeben sei das lineare Gleichungssystem $Ax = b$ mit nichtsingulärem $A \in \mathbb{R}^{n\times n}$. Mit der Variablentransformation $x = A^T u$ wende man auf $AA^T u = b$ das CG-Verfahren an. Im resultierenden Verfahren (CGNE) eliminiere man die Variable u und gewinne eine Näherungsfolge $\{x_k\}$.

8. Auf das lineare Gleichungssystem $Ax = b$ mit

$$A := \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ 1 & & & & 0 \end{pmatrix} \in \mathbb{R}^{n\times n}, \qquad b := \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \in \mathbb{R}^n$$

wende man das Verfahren CGNE aus Aufgabe 7 an, wobei man mit $x_0 := 0$ starte[10]. Was kann über die Konvergenz ausgesagt werden?

9. Gegeben sei das lineare Gleichungssystem $Ax = b$, bei dem die Koeffizientenmatrix A ein (nichttriviales) Vielfaches einer orthogonalen Matrix ist. Man zeige, dass das Verfahren CGNE aus Aufgabe 7 mit einem beliebigen Startvektor x_0 nach einem Schritt die Lösung bestimmt.

10. Auf das lineare Gleichungssystem $Ax = b$ aus Aufgabe 8 wende man GMRES an, wobei man wieder mit $x_0 := 0$ starte. Was kann über die Konvergenz ausgesagt werden?

5.3 Präkonditionierung

5.3.1 Einleitung

Es ist klar, dass alle Verfahren in Abschnitt 5.2 zur Lösung des linearen Gleichungssystems $Ax = b$ schnell konvergieren, wenn A nur eine kleine Störung der Identität ist, da

[10] Dieses Beispiel ist der Arbeit

N. M. Nachtigal, S. C. Reddy, L. N. Trefethen (1992) "How fast are nonsymmetric matrix iterations?" SIAM J. Matrix Anal. 13, 778–795

entnommen.

sie in einem Schritt für $A = I$ die Lösung liefern. Dies ist in der Praxis i. allg. nicht der Fall. Daher spielt eine *Präkonditionierung* des gegebenen linearen Gleichungssystems eine wichtige Rolle. Zwei Methoden, die auch kombiniert eingesetzt werden können, sind von Bedeutung. In der ersten denke man sich das gegebene lineare Gleichungssystem mit der Inversen M^{-1} einer nichtsingulären Matrix M von links durchmultipliziert (left preconditioning). Ziel ist es, $M^{-1}A \approx I$ zu erreichen, dabei sollten lineare Gleichungssysteme mit M als Koeffizientenmatrix "leicht" gelöst werden können. In der zweiten Methode denke man sich die Variablentransformation $x = M^{-1}u$ gemacht (right preconditioning) und löse das lineare Gleichungssystem $AM^{-1}u = b$. An M werden dieselben Forderungen wie beim right preconditioning gestellt, genauer sollte diesmal $AM^{-1} \approx I$ sein. Natürlich kann dies mittels $x = M_2^{-1}u$ und $M_1^{-1}AM_2^{-1}u = M_1^{-1}b$ auch kombiniert werden. Ist etwa A symmetrisch und positiv definit, so wird man auch von der Koeffizientenmatrix des präkonditionierten linearen Gleichungssystems erwarten, dass sie dieselbe angenehme Eigenschaft besitzt. Dies kann erreicht werden, indem man $M_1 := L$ und $M_2 := L^T$ wählt. Z. B. könnte die nichtsinguläre untere Dreiecksmatix L eine (dünn besetzte) Näherung an einen Cholesky-Faktor von A sein.

Wir werden zunächst sozusagen generische Präkonditionierer annehmen, d. h. die Matrix M nicht spezifizieren und die Auswirkung eines Präkonditionierers auf das entsprechende Krylov-Verfahren untersuchen. Erst später werden wir auf spezielle Präkonditionierer eingehen.

5.3.2 Präkonditioniertes CG-Verfahren

In diesem Unterabschnitt betrachten wir das lineare Gleichungssystem $Ax = b$ mit symmetrischem und positiv definitem $A \in \mathbb{R}^{n\times n}$. Wir erinnern zunächst an das CG-Verfahren ohne Präkonditionierung:

- Mit einer Näherung x_0 berechne man $r_0 := b - Ax_0$ und setze $p_0 := r_0$.
- Für $k = 0, 1, \ldots$:
  - $\alpha_k := r_k^T r_k / p_k^T A p_k$.
  - $x_{k+1} := x_k + \alpha_k p_k$.
  - $r_{k+1} := r_k - \alpha_k A p_k$.
  - $\beta_k := r_{k+1}^T r_{k+1} / r_k^T r_k$.
  - $p_{k+1} := r_{k+1} + \beta_k p_k$.

Wir nehmen an, ein Präkonditionierer M stehe zur Verfügung, von dem wir nur voraussetzen, dass er ebenfalls symmetrisch und positiv definit ist. Aus praktischen Gründen sollte ein lineares Gleichungssystem mit M als Koeffizientenmatrix "leicht" zu lösen sein, ferner sollte M die gegebene Koeffizientenmatrix A in einem gewissen Sinne approximieren. Es gibt mehrere Möglichkeiten, das gegebene Gleichungssystem zu präkonditionieren. Die naheliegendsten sind

$$M^{-1}Ax = M^{-1}b,$$

oder

$$AM^{-1}u = b, \qquad x = M^{-1}u,$$

oder

$$L^{-1}AL^{-T}u = L^{-1}b, \qquad x = L^{-T}u.$$

Der ins Auge fallende Vorteil der letzten Möglichkeit besteht darin, dass die Koeffizientenmatrix $L^{-1}AL^{-T}$ wieder symmetrisch und positiv definit ist. Dies ist aber eigentlich nur ein scheinbarer Vorteil, da wir zeigen können:

- Die Matrix $M^{-1}A$ ist symmetrisch und positiv definit bezüglich des durch

  $$\langle x, y\rangle_M := (Mx)^T y$$

  definierten inneren Produktes $\langle\cdot,\cdot\rangle_M$.

- Die Matrix AM^{-1} ist symmetrisch und positiv definit bezüglich des durch

  $$\langle x, y\rangle_{M^{-1}} := (M^{-1}x)^T y$$

  definierten inneren Produktes $\langle\cdot,\cdot\rangle_{M^{-1}}$.

Denn: Es ist

$$\langle M^{-1}Ax, y\rangle_M = (Ax)^T y = x^T Ay = x^T M(M^{-1}A)y = (Mx)^T M^{-1}Ay = \langle x, M^{-1}Ay\rangle_M,$$

womit die Symmetrie von $M^{-1}A$ bezüglich $\langle\cdot,\cdot\rangle_M$ gezeigt ist. Die positive Definitheit von $M^{-1}A$ bezüglich desselben inneren Produktes folgt sofort aus $\langle M^{-1}Ax, x\rangle_M = (Ax)^T x$ und der positiven Definitheit von A. Da auch $M^{-1}AM^{-1}$ symmetrisch und positiv definit ist, folgt auch die zweite Aussage.

Wendet man das CG-Verfahren auf $M^{-1}Ax = M^{-1}b$ unter Benutzung des inneren Produktes $\langle\cdot,\cdot\rangle_M$ an, bezeichnet man ferner mit $r_k := b - Ax_k$ den "Originaldefekt", mit $z_k := M^{-1}r_k$ den Defekt im präkonditionierten System, so sieht der k-te Schritt folgendermaßen aus:

- $\alpha_k := \langle z_k, z_k\rangle_M / \langle (M^{-1}Ap_k, p_k\rangle_M$.
- $x_{k+1} := x_k + \alpha_k p_k$.
- $r_{k+1} := r_k - \alpha_k Ap_k$, $z_{k+1} := M^{-1}r_{k+1}$.
- $\beta_k := \langle z_{k+1}, z_{k+1}\rangle_M / \langle z_k, z_k\rangle_M$.
- $p_{k+1} := z_{k+1} + \beta_k p_k$.

Berücksichtigt man nun noch, dass

$$\alpha_k = \frac{\langle z_k, z_k\rangle_M}{\langle M^{-1}Ap_k, p_k\rangle_M} = \frac{r_k^T z_k}{(Ap_k)^T p_k}$$

und

$$\beta_k = \frac{\langle z_{k+1}, z_{k+1}\rangle_M}{\langle z_k, z_k\rangle_M} = \frac{r_{k+1}^T z_{k+1}}{r_k^T z_k},$$

so erhält man die folgende Fassung eines (links) präkonditionierten CG-Verfahrens:

- Mit einer Näherung x_0 berechne man $r_0 := b - Ax_0$, berechne $z_0 := M^{-1}r_0$ und setze $p_0 := z_0$.
- Für $k = 0, 1, \ldots$:
  - $\alpha_k := r_k^T z_k / p_k^T A p_k$.
  - $x_{k+1} := x_k + \alpha_k p_k$.
  - $r_{k+1} := r_k - \alpha_k A p_k$, $z_{k+1} := M^{-1} r_{k+1}$.
  - $\beta_k := r_{k+1}^T z_{k+1} / r_k^T z_k$.
  - $p_{k+1} := z_{k+1} + \beta_k p_k$.

Man beachte, dass pro Iteration ein Matrix-Vektor-Produkt zu bilden und ein lineares Gleichungssystem mit M als Koeffizientenmatrix zu lösen ist.

Nun wollen wir das CG-Verfahren auf $\hat{A}u = \hat{b}$ mit $\hat{A} := L^{-1}AL^{-T}$ und $\hat{b} := L^{-1}b$ anwenden. Eine direkte Übertragung liefert:

- Mit einer Näherung u_0 berechne man $\hat{r}_0 := \hat{b} - \hat{A}u_0$ und setze $\hat{p}_0 := \hat{r}_0$.
- Für $k = 0, 1, \ldots$:
  - $\alpha_k := \hat{r}_k^T \hat{r}_k / \hat{p}_k^T \hat{A} \hat{p}_k$.
  - $u_{k+1} := u_k + \alpha_k \hat{p}_k$.
  - $\hat{r}_{k+1} := \hat{r}_k - \alpha_k \hat{A} \hat{p}_k$.
  - $\beta_k := \hat{r}_{k+1}^T \hat{r}_{k+1} / \hat{r}_k^T \hat{r}_k$.
  - $\hat{p}_{k+1} := \hat{r}_{k+1} + \beta_k \hat{p}_k$.

Mit $x_k := L^{-T}u_k$, $p_k := L^{-T}\hat{p}_k$ und $r_k := L\hat{r}_k = b - Ax_k$ erhält man die folgende Version, die den Vorteil hat, dass sie die Originaldaten benutzt.

- Mit einer Näherung x_0 berechne man $r_0 := b - Ax_0$, berechne $\hat{r}_0 := L^{-1}r_0$ und $p_0 := L^{-T}\hat{r}_0$.
- Für $k = 0, 1, \ldots$:
  - $\alpha_k := \hat{r}_k^T \hat{r}_k / p_k^T A p_k$.
  - $x_{k+1} := x_k + \alpha_k p_k$.
  - $\hat{r}_{k+1} := \hat{r}_k - \alpha_k L^{-1} A p_k$.
  - $\beta_k := \hat{r}_{k+1}^T \hat{r}_{k+1} / \hat{r}_k^T \hat{r}_k$.
  - $p_{k+1} := L^{-T}\hat{r}_{k+1} + \beta_k \hat{p}_k$.

Wieder tritt in jeder Iteration nur eine Matrix-Vektor-Multiplikation auf, ferner sind zwei lineare Gleichungssysteme pro Iteration zu lösen, einmal mit L, das andere mal mit L^T als Koeffizientenmatrix. Ist $M = LL^T$ und wird in beiden Verfahren derselbe Startwert genommen, so produzieren die beiden bisher vorgestellten präkonditionierten CG-Verfahren dieselbe Iterationsfolge $\{x_k\}$. Dies ist einfach zu zeigen.

Nun gehen wir noch kurz auf die Präkonditionierung von rechts ein, also auf die Anwendung des CG-Verfahrens auf $AM^{-1}u = b$, wobei natürlich wieder M symmetrisch und positiv definit sei. Wir beachten, dass AM^{-1} symmetrisch und positiv definit bezüglich des inneren Produktes $\langle \cdot, \cdot \rangle_{M^{-1}}$ ist. Der k-te Iterationsschritt besteht dann naheliegenderweise aus den folgenden Berechnungen:

- $\alpha_k := \langle r_k, r_k \rangle_{M^{-1}} / \langle AM^{-1}p_k, p_k \rangle_{M^{-1}}$.
- $u_{k+1} := u_k + \alpha_k p_k$.
- $r_{k+1} := r_k - \alpha_k AM^{-1}p_k$.
- $\beta_k := \langle r_{k+1}, r_{k+1} \rangle_{M^{-1}} / \langle r_k, r_k \rangle_{M^{-1}}$.
- $p_{k+1} := r_{k+1} + \beta_k p_k$.

Mit $x_k := M^{-1}u_k$, $q_k := M^{-1}p_k$ und $z_k := M^{-1}r_k$ lauten diese Schritte:

- $\alpha_k := z_k^T r_k / (Aq_k)^T q_k$.
- $x_{k+1} := x_k + \alpha_k q_k$.
- $r_{k+1} := r_k - \alpha_k Aq_k$, $z_{k+1} := M^{-1}r_{k+1}$.
- $\beta_k := z_{k+1}^T r_{k+1} / z_k^T r_k$.
- $q_{k+1} := z_{k+1} + \beta_k q_k$.

Offensichtlich erhält man dieselbe Iterationsfolge wie bei der Präkonditionierung von links.

5.3.3 Präkonditioniertes GMRES

Gegeben ist das lineare Gleichungssystem $Ax = b$ mit nicht notwendig symmetrischer Koeffizientenmatrix $A \in \mathbb{R}^{n \times n}$. Wir betrachten zunächst GMRES mit Präkonditionierung von links, also die Anwendung von GMRES auf $M^{-1}Ax = M^{-1}b$. Dies führt auf:

- Mit einer Näherungslösung x_0 berechne man den präkonditionierten Defekt $z_0 := M^{-1}(b - Ax_0)$. Berechne $q_1 := z_0 / \|z_0\|_2$. Setze

$$\tilde{H}_k = (h_{ij})_{\substack{1 \le i \le k+1 \\ 1 \le j \le k}} = 0.$$

- Für $j = 1, \ldots, k$:
  - $w := M^{-1}Aq_j$.
  - Für $i = 1, \ldots, j$:
    - $h_{ij} := q_i^T w$.
    - $w := w - h_{ij} q_i$.

  - $h_{j+1,j} := \|w\|_2$.
  - Falls $h_{j+1,j} = 0$, dann: Setze $k := j$ und gehe zum letzten $\bullet$.
  - $q_{j+1} := w/h_{j+1,j}$.

- Berechne die Lösung y_k des linearen Ausgleichsproblems

$$\text{Minimiere} \quad J(y) := \|\|z_0\|_2 e_1 - \tilde{H}_k y\|_2, \quad y \in \mathbb{R}^k,$$

  und berechne anschließend $x_k := x_0 + Q_k y_k$, wobei $Q_k := (\ q_1 \ \cdots \ q_k\ )$.

Im ersten Teil wird eine Orthonormalbasis zum Krylov-Teilraum $\mathcal{K}_k(M^{-1}A, z_0)$ berechnet, wobei der Defekt präkonditioniert ist. Ein (leichter) Nachteil besteht hier darin, dass man keinen einfachen Zugang auf den nicht präkonditionierten Defekt hat. Auch hier wird man i. allg. einen Restart benutzen. Das gleiche gilt auch für das von rechts präkonditionierte GMRES:

- Mit einer Näherungslösung x_0 berechne man den Defekt $r_0 := b - Ax_0$. Berechne $q_1 := r_0/\|r_0\|_2$. Setze

$$\tilde{H}_k = (h_{ij})_{\substack{1 \le i \le k+1 \\ 1 \le j \le k}} = 0.$$

- Für $j = 1, \ldots, k$:
  - $w := AM^{-1}q_j$.
  - Für $i = 1, \ldots, j$:
    - $h_{ij} := q_i^T w$.
    - $w := w - h_{ij} q_i$.
  - $h_{j+1,j} := \|w\|_2$.
  - Falls $h_{j+1,j} = 0$, dann: Setze $k := j$ und gehe zum letzten $\bullet$.
  - $q_{j+1} := w/h_{j+1,j}$.

- Berechne die Lösung y_k des linearen Ausgleichsproblems

$$\text{Minimiere} \quad J(y) := \|\|r_0\|_2 e_1 - \tilde{H}_k y\|_2, \quad y \in \mathbb{R}^k,$$

  und berechne anschließend $x_k := x_0 + M^{-1}Q_k y_k$, wobei $Q_k := (\ q_1 \ \cdots \ q_k\ )$.

Diesmal wird im ersten Teil eine Orthonormalbasis für $\mathcal{K}_k(AM^{-1}, r_0)$ berechnet, wobei r_0 der nicht präkonditionierte Defekt ist. Im folgenden Lemma wird der prinzipielle Unterschied zwischen dem von links bzw. von rechts präkonditionierten GMRES aufgezeigt.

Lemma 3.1 *Die Näherungslösung x_k, die durch das von links bzw. von rechts präkonditionierte GMRES erhalten wird, ist im ersten Fall Lösung der Aufgabe*

$$\textit{Minimiere} \quad \|M^{-1}(b - Ax)\|_2, \quad x \in x_0 + \mathcal{K}_k(M^{-1}A, z_0),$$

im zweiten Fall ist x_k Lösung der Aufgabe

$$\textit{Minimiere} \quad \|b - Ax\|_2, \quad x \in x_0 + M^{-1}\mathcal{K}_k(AM^{-1}, r_0).$$

Hierbei ist $r_0 := b - Ax_0$ und $z_0 := M^{-1}r_0$. In beiden Fällen[11] *hat x_k eine Darstellung*

$$x_k = x_0 + s_{k-1}(M^{-1}A)z_0 = x_0 + M^{-1}s_{k-1}(AM^{-1})r_0,$$

wobei $s_{k-1} \in \Pi_{k-1}$.

Beweis: Die erste Aussage ist völlig klar, denn x_k wird durch auf $M^{-1}Ax = M^{-1}b$ angewandtes GMRES gewonnen. Außerdem existiert hier (d. h. bei Präkonditionierung von links) ein Polynom $s_{k-1} \in \Pi_{k-1}$ mit

$$x_k = x_0 + s_{k-1}(M^{-1}A)z_0 = x_0 + s_{k-1}(M^{-1}A)M^{-1}r_0 = x_0 + M^{-1}s_{k-1}(AM^{-1})r_0.$$

Bei der letzten Gleichung haben wir ausgenutzt, dass

$$(M^{-1}A)^j M^{-1} = M^{-1}(AM^{-1})^j, \qquad j = 0, \ldots, k-1,$$

wie man leicht durch vollständige Induktion beweist. Im zweiten Fall (d. h. einer Präkonditionierung von rechts) ist $x_k = M^{-1}u_k$ und u_k Lösung der Aufgabe

$$\text{Minimiere} \quad \|b - AM^{-1}u\|_2, \qquad u \in u_0 + \mathcal{K}(AM^{-1}, r_0),$$

wobei $x_0 := M^{-1}u_0$ und $r_0 := b - Ax_0$. Mit der Variablentransformation $x = M^{-1}u$ erhält man dann auch die zweite Aussage. □

5.3.4 Jacobi, SOR und SSOR Präkonditionierer

Bisher sind wir von einem fast beliebigen Präkonditionierer M ausgegangen (im Zusammenhang mit dem CG-Verfahren haben wir vorausgesetzt, dass er symmetrisch und positiv definit ist). Ziel ist es, eine Matix M mit $M^{-1}A \approx I$ (oder $AM^{-1} \approx I$) derart zu finden, dass lineare Gleichungssysteme mit M als Koeffizientenmatrix "leicht" zu lösen sind.

Wir wollen die gebräuchlichsten Präkonditionierer schildern, wobei wir in diesem Unterabschnitt mit den einfachsten (Jacobi, SOR und SSOR) beginnen werden.

Bei den elementaren Iterationsverfahren zur Lösung des linearen Gleichungssystems $Ax = b$ sind wir von einer Zerlegung $A = M - N$ ausgegangen, wobei M nichtsingulär ist und ein Gleichungssystem mit M als Koeffizientenmatrix "leicht" lösbar sein sollte. Das zugehörige Iterationsverfahren ist

$$x_{k+1} := M^{-1}Nx_k + M^{-1}b$$

bzw.

$$x_{k+1} := Rx_k + c,$$

[11] Das soll nicht suggerieren, dass die beiden Näherungen gleich sind!

wobei $c := M^{-1}b$ und

$$R := M^{-1}N = M^{-1}(M - A) = I - M^{-1}A.$$

Mit der Standardzerlegung $A = A_D + A_L + A_R$ ist

$$R_J = I - \underbrace{A_D}_{M_J}{}^{-1}A$$

die Iterationsmatrix des Jacobi-Verfahrens und

$$R_{GS} = I - \underbrace{(A_D + A_L)}_{M_{GS}}{}^{-1}A$$

die des Gauß-Seidel-Verfahrens. Beim SOR-Verfahren mit dem Relaxationsparameter ω ist

$$R_{SOR}(\omega) = I - \underbrace{[(1/\omega)A_D + A_L]}_{M_{SOR}(\omega)}{}^{-1}A.$$

Schließlich wollen wir noch an das symmetrische SOR-Verfahren (SSOR) erinnern. Es sieht bekanntlich folgendermaßen aus:

- Berechne $x_{k+1/2}$ aus x_k mittels

  $$\Big[\frac{1}{\omega}A_D + A_L\Big]x_{k+1/2} = \Big[\Big(\frac{1}{\omega} - 1\Big)A_D - A_R\Big]x_k + b,$$

  d. h. mache von x_k ausgehend einen normalen SOR-Schritt.

- Berechne x_{k+1} aus $x_{k+1/2}$ mittels

  $$\Big[\frac{1}{\omega}A_D + A_R\Big]x_{k+1} = \Big[\Big(\frac{1}{\omega} - 1\Big)A_D - A_L\Big]x_{k+1/2} + b,$$

  d. h. mache einen SOR-Schritt, bei dem die Rollen von A_L und A_R vertauscht sind. Hier werden also die Komponenten von x_{k+1} in fallender Reihenfolge bestimmt.

Schreibt man dies in der Form

$$M_1 x_{k+1/2} = N_1 x_k + b, \qquad M_2 x_{k+1} = N_2 x_{k+1/2} + b,$$

so ist

$$M_2 x_{k+1} = N_2 M_1^{-1} N_1 x_k + (I + N_2 M_1^{-1})b$$

bzw.

$$(I + N_2 M_1^{-1})^{-1} M_2 x_{k+1} = (I + N_2 M_1^{-1})^{-1} N_2 M_1^{-1} N_1 x_k + b.$$

Der hierzu gehörende Präkonditionierer ist daher (sinnvoll ist nur ein Relaxationsparameter $\omega \in (0, 2)$)

$$\begin{aligned} M_{SSOR}(\omega) &= \Big\{I + \Big[\Big(\frac{1}{\omega} - 1\Big)A_D - A_L\Big]\Big[\frac{1}{\omega}A_D + A_L\Big]^{-1}\Big\}^{-1}\Big[\frac{1}{\omega}A_D + A_R\Big] \\ &= \frac{\omega}{2-\omega}\Big(\frac{1}{\omega}A_D + A_L\Big)A_D^{-1}\Big(\frac{1}{\omega}A_D + A_D\Big). \end{aligned}$$

Offensichtlich ist für symmetrisches A auch der SSOR-Präkonditionierer $M_{SSOR}(\omega)$ symmetrisch. Für $\omega = 1$ erhält man den symmetrischen Gauß-Seidel-Präkonditionierer

$$M_{SGS} = (A_D + A_L)A_D^{-1}(A_D + A_R).$$

Wegen $M_{SGS} = LU$ mit der unteren Dreiecksmatrix mit Einsen in der Diagonalen $L := I + A_L A_D^{-1}$ und der oberen Dreiecksmatrix $U := A_D + A_R$ wird ein lineares Gleichungssystem mit M_{SGS} als Koeffizientenmatrix durch Vorwärts- und Rückwärtseinsetzen gelöst. Entsprechendes gilt natürlich auch für $M_{SSOR}(\omega)$.

Das Iterationsverfahren $x_{k+1} = Rx_k + f$ versucht die Gleichung $(I - R)x = f$ bzw. $M^{-1}Ax = M^{-1}b$ zu lösen, es kann also als eine auf ein präkonditioniertes System angewandte Fixpunktiteration aufgefaßt werden.

5.3.5 Unvollständige Zerlegungen

Gegeben sei eine dünn besetzte Matrix (wir werden diesen Begriff später genauer fassen) $A \in \mathbb{R}^{n\times n}$. Eine *unvollständige LU-Zerlegung* von A berechnet eine dünn besetzte untere Dreiecksmatrix L und eine dünn besetzte obere Dreiecksmatrix U derart, dass der Fehler $E := LU - A$ gewissen Bedingungen genügt, etwa verschwindende Einträge in bestimmten Positionen hat. Zunächst werden wir uns auf sogenannte M-Matrizen spezialisieren.

Definition 3.2 Eine nichtsinguläre Matrix $A \in \mathbb{R}^{n\times n}$ heißt eine *M-Matrix*, wenn die Diagonalelemente von A positiv ($a_{ii} > 0$, $i = 1, \ldots, n$), die Außerdiagonalelemente von A nichtpositiv ($a_{ij} \leq 0$, $i, j = 1, \ldots, n$, $i \neq j$) sind und die inverse Matrix A^{-1} nichtnegativ ist.

Die Bedingungen hier sind redundant, da die Positivität der Diagonalelemente aus den übrigen Bedingungen geschlossen werden kann, siehe Aufgabe 4. Diese Tatsache werden wir im weiteren ohne nähere Begründung benutzen.

Der folgende Satz (siehe Y. Saad (1996, S. 270) ist Grundlage für ein Verfahren zur Berechnung einer unvollständigen LU-Zerlegung, von welchem wir zeigen werden können, dass es bei einer M-Matrix durchführbar ist.

Satz 3.3 *Sei $A \in \mathbb{R}^{n\times n}$ eine M-Matrix. Da insbesondere $a_{11} \neq 0$ kann ein Schritt des Gaußschen Eliminationsverfahrens ohne Spaltenpivotsuche durchgeführt werden. Mit*

$$l_1 := \frac{1}{a_{11}} \begin{pmatrix} 0 \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix}, \qquad M_1 := I - l_1 e_1^T$$

liefere der erste Schritt des Gaußschen Eliminationsverfahrens die Matrix $A_1 := M_1 A$. Dann gilt:

1. *Auch A_1 ist eine M-Matrix.*

2. *Die $(n-1)\times(n-1)$-Matrix, die man aus A_1 durch Streichen der ersten Zeile und der ersten Spalte erhält, ist ebenfalls eine M-Matrix.*

Beweis: Da die Gauß-Matrix M_1 und A nichtsingulär sind, ist es auch $A_1 = M_1 A$. Die Außerdiagonalelemente von A_1 in der ersten Zeile sind dieselben wie die von A, insbesondere also nichtpositiv, die in der ersten Spalte verschwinden. Für $i, j = 2, \ldots, n$, $i \neq j$, ist

$$a_{ij}^{(1)} = \underbrace{a_{ij}}_{\leq 0} - \underbrace{\frac{a_{i1}a_{1j}}{a_{11}}}_{\geq 0},$$

die Außerdiagonalelemente von A_1 sind also sämtlich nichtpositiv. Schließlich ist

$$A_1^{-1} = A^{-1}M_1^{-1} = A^{-1}(I + l_1 e_1^T).$$

Folglich ist

$$A_1^{-1}e_1 = A^{-1}(e_1 + l_1) = \frac{1}{a_{11}}A^{-1}\begin{pmatrix} a_{11} \\ \vdots \\ a_{n1} \end{pmatrix} = \frac{1}{a_{11}}e_1 \geq 0,$$

die erste Spalte von A_1^{-1} ist also nichtnegativ. Für $j = 2, \ldots, n$ ist

$$A_1^{-1}e_j = A^{-1}e_j \geq 0,$$

insgesamt ist $A_1^{-1} \geq 0$ und die erste Behauptung ist bewiesen.

Nun sei

$$A_1 = \begin{pmatrix} a_{11} & a^T \\ 0 & \hat{A}_1 \end{pmatrix}, \qquad a := \begin{pmatrix} a_{12} \\ \vdots \\ a_{1n} \end{pmatrix}.$$

Da $a_{11} > 0$ und A_1 nichtsingulär, ist auch $\hat{A}_1$ nichtsingulär. Klar ist, dass die Außerdiagonalelemente von $\hat{A}_1$ nichtpositiv sind. Zu zeigen bleibt also, dass $\hat{A}_1^{-1} \geq 0$. Nun ist aber

$$A_1^{-1} = \begin{pmatrix} 1/a_{11} & -(1/a_{11})a^T\hat{A}_1^{-1} \\ 0 & \hat{A}_1^{-1} \end{pmatrix},$$

so dass aus $A_1^{-1} \geq 0$ auch $\hat{A}_1^{-1} \geq 0$ folgt. Der Satz ist damit bewiesen. □

Durch einen Schritt des Gaußschen Eliminationsverfahrens sei aus A die Matrix A_1 erhalten. In A_1 setze man gewisse (nichtpositive) Außerdiagonalelemente auf Null und erhalte damit eine Matrix $\tilde{A}_1 \geq A_1$. Dann ist auch $\tilde{A}_1$ eine M-Matrix (siehe Aufgabe 5). Auf den unteren $(n-1)\times(n-1)$-Block von $\tilde{A}_1$ kann wieder ein Schritt des Gaußschen Eliminationsverfahren gemacht werden und das Verfahren in dieser Weise fortgesetzt werden. Startet man mit einer M-Matrix, so kann das Verfahren nicht vorzeitig abbrechen.

Nun geben wir uns ein "Nullmuster" (zero pattern)

$$P := \{(i,j) : i \neq j,\ 1 \leq i, j \leq n\}$$

vor. Ziel ist es, eine Zerlegung $A = LU - E$ zu gewinnen, bei der die Einträge der unteren bzw. oberen Dreiecksmatrix L bzw. U in Positionen aus P verschwinden und die Einträge von E in Positionen, die nicht zu P gehören, verschwinden. Hierzu betrachten wir den folgenden Algorithmus:

- Input: Gegeben ist eine Matrix $A \in \mathbb{R}^{n\times n}$, ferner ein Nullmuster P (das die Diagonale nicht enthält).
- Für $k = 1, \ldots, n-1$:
  - Für $i = k+1, \ldots, n$ und $(i,k) \notin P$:
    - $a_{ik} := a_{ik}/a_{kk}$.
    - Für $j = k+1, \ldots, n$ und $(i,j) \notin P$:
      - $a_{ij} := a_{ij} - a_{ik}a_{kj}$.
- Output: Ausgegeben wird eine untere Dreiecksmatrix L mit

$$l_{ij} := \begin{cases} \delta_{ij} & \text{für } i \le j, \\ a_{ij} & \text{für } i > j,\ (i,j) \notin P, \\ 0 & \text{für } i > j,\ (i,j) \in P, \end{cases}$$

  und eine obere Dreiecksmatrix R mit

$$r_{ij} := \begin{cases} a_{ij} & \text{für } i \le j,\ (i,j) \notin P, \\ 0 & \text{für } i \le j,\ (i,j) \in P, \\ 0 & \text{für } i > j. \end{cases}$$

Im folgenden Satz[12] (siehe auch Y. Saad (1996, S. 271) und A. Greenbaum (1997, S. 173)) wird gezeigt, dass durch den obigen Algorithmus das angegebene Ziel erreicht wird, jedenfalls dann, wenn die Ausgangsmatrix A eine M-Matrix ist.

Satz 3.4 *Sei A eine M-Matrix und P ein Nullmuster, das die Diagonale nicht enthält. Dann ist das obige Verfahren durchführbar. Mit den ausgegebenen Matrizen L und U sowie $E := LU - A$ ist $e_{ij} = 0$ für $(i,j) \notin P$. Schließlich ist $A = LU - E$ eine reguläre Zerlegung von A, d. h. LU ist nichtsingulär und $(LU)^{-1}$ und U sind nichtnegative Matrizen.*

Beweis: Der oben angegebene Algorithmus entspricht genau dem Gaußschen Eliminationsverfahren ohne Pivotisierung, wobei nur im k-ten Schritt Einträge in den Positionen $(i,k) \in P$, $i = k+1, \ldots, n$, und $(i,j) \in P$, $i,j = k+1, \ldots, n$, auf Null gesetzt werden. Dies werden wir ausnutzen, um eine Matrix-Version des Verfahrens anzugeben.

- Setze $A_0 := A$. Die Matrix $\tilde{A}_1$ gewinne man aus A_0 dadurch, dass Einträge in Positionen aus P auf Null gesetzt werden und die anderen unverändert bleiben.

[12] Er ist zu finden (Theorem 2.3) in der Arbeit

J. A. Meijerink, H. A. van der Horst (1977) "An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix". Mathematics of Computation 31, 148–162.

- Für $k = 1, \ldots, n-1$:
  - Sei $M_k := I - l_k e_k^T$ mit

$$l_k := \frac{1}{\tilde{a}_{kk}^{(k)}}(\underbrace{0, \ldots, 0}_{k}, \tilde{a}_{k+1,k}^{(k)}, \ldots, \tilde{a}_{nk}^{(k)})^T.$$

  - Berechne $A_k := M_k \tilde{A}_k$. Die Matrix $\tilde{A}_{k+1}$ gewinne man aus A_k dadurch, dass Einträge in Positionen aus P auf Null gesetzt werden und die anderen unverändert bleiben.

Aus Satz 3.3 und obiger Argumentation folgt, dass die Matrizen A_k und $\tilde{A}_k$ jeweils M-Matrizen sind, ferner ist $E_k := \tilde{A}_k - A_{k-1} \geq 0$. Dann ist

$$\begin{aligned} A_{n-1} &= M_{n-1}\tilde{A}_{n-1} \\ &= M_{n-1}(A_{n-2} + E_{n-1}) \\ &\vdots \\ &= \Big(\prod_{j=1}^{n-1} M_{n-j}\Big)A_0 + \sum_{k=1}^{n-1}\Big(\prod_{j=1}^{n-k} M_{n-j}\Big)E_k \\ &= \Big(\prod_{j=1}^{n-1} M_{n-j}\Big)\Big(A + \sum_{k=1}^{n-1} E_k\Big). \end{aligned}$$

Hierbei haben wir in der letzten Gleichung benutzt, dass

$$\Big(\prod_{j=1}^{n-1} M_{n-j}\Big)E_k = \Big(\prod_{j=1}^{n-k} M_{n-j}\Big)E_k.$$

Dies liegt daran, dass $(E_k)_{ij} = 0$, $1 \leq i, j \leq k$. Mit

$$U := A_{n-1}, \qquad E := \sum_{k=1}^{n-1} E_k$$

sowie

$$L := (M_{n-1} \cdots M_1)^{-1} = (I + l_1 e_1^T) \cdots (I + l_{n-1} e_{n-1}^T) = I + \sum_{k=1}^{n-1} l_k e_k^T$$

ist dann $A = LU - E$ (wobei L und U genau die im Output des obigen Algorithmus angegebenen Dreiecksmatrizen sind, sie haben also das gewünschte Nullmuster). In den Positionen, die nicht zu P gehören, verschwinden die Einträge von E, da dies für alle E_k gilt. Schließlich sind L und U und damit auch LU nichtsingulär, es ist $E \geq 0$ und $(LU)^{-1} \geq 0$. Letzteres erkennt man daran, dass $LU = A + E \geq A$ und A eine M-Matrix ist. Damit ist der Satz bewiesen. □

Bemerkung: Wir haben konstruktiv gezeigt, dass es zu jeder M-Matrix A und jedem die Diagonale nicht enthaltenden Nullmuster P eine untere Dreiecksmatrix L mit Einsen in der Diagonalen und eine obere Dreiecksmatrix U gibt derart, dass $A = LU - E$ eine reguläre Zerlegung von A ist und

$$l_{ij} = 0 \quad ((i,j) \in P), \qquad u_{ij} = 0 \quad ((i,j) \in P), \qquad e_{ij} = 0 \quad ((i,j) \notin P).$$

Es kann gezeigt werden, dass durch diese Forderungen die Faktoren L und U eindeutig bestimmt sind (siehe Aufgabe 6). □

Eine symmetrische M-Matrix ist positiv definit (siehe Aufgabe 8). Da die Berechnung der Cholesky-Zerlegung einer symmetrischen, positiv definiten Matrix im Prinzip äquivalent der Berechnung einer LR-Zerlegung mit dem Gaußchen Iterationsverfahren ohne Pivotisierung ist, erhalten wir aus Satz 3.4 für eine symmetrische M-Matrix A die Existenz einer unvollständigen Cholesky-Zerlegung $A = LL^T - E$, und zwar bezüglich eines beliebigen *symmetrischen* Nullmusters (d. h. aus $(i,j) \in P$ folgt auch $(j,i) \in P$). Genauer existiert eine untere Dreiecksmatrix L mit $l_{ij} = 0$, $(i,j) \in P$, und $A = LL^T - E$ mit $e_{ij} = 0$, $(i,j) \notin P$. Dies führt dann auf die *unvollständige Cholesky-Zerlegung*.

Auf viele weitere Feinheiten wird bei Y. Saad (1996, S. 269 ff.) eingegangen, wir verzichten darauf.

5.3.6 Aufgaben

1. Seien die Matrix A und ihr Präkonditionierer M symmetrisch und positiv definit. Man zeige, dass $M^{-1}A$ symmetrisch und positiv definit bezüglich des durch $\langle x, y\rangle_A := (Ax)^T y$ gegebenen inneren Produktes $\langle \cdot, \cdot \rangle_A$ ist. Hierauf aufbauend entwickele man ein CG-Verfahren zur Lösung des präkonditionierten linearen Gleichungssystems $M^{-1}Ax = M^{-1}b$ entsprechend dem von links präkonditionierten CG-Verfahrens. Dieses sollte mit nur einer Matrix-Vektor-Multiplikation pro Iterationsschritt auskommen.

2. Gegeben sei das lineare Gleichungssystem $Ax = b$ mit nichtsingulärem $A \in \mathbb{R}^{n\times n}$. Man gebe zu dem in Aufgabe 7 in Abschnitt 5.2 entwickelten Verfahren CGNE ein von links mit einer symmetrischen, positiv definiten Matrix $M \in \mathbb{R}^{n\times n}$ präkonditioniertes CG-Verfahren an.

3. Sei $M = LR$ ein (nichtsingulärer) Präkonditionierer für die nichtsinguläre Matrix A (hierbei müssen L und R nicht notwendigerweise untere bzw. obere Dreiecksmatrizen sein). Man zeige, dass die Matrizen $M^{-1}A$, AM^{-1} und $L^{-1}AR^{-1}$ dieselben Eigenwerte besitzen.

4. Die nichtsinguläre Matrix $A \in \mathbb{R}^{n\times n}$ habe nichtpositive Außerdiagonalelemente und eine nichtnegative Inverse. Man zeige, dass die Diagonalelemente von A positiv sind.

5. Seien $A, B \in \mathbb{R}^{n\times n}$ zwei Matrizen mit $A \le B$ und $b_{ij} \le 0$ für $i \neq j$. Ist dann A eine M-Matrix, so ist auch B eine M-Matrix.

   Hinweis: Man kann benutzen, dass der Spektralradius einer nichtnegativen Matrix ein Eigenwert ist, zu dem ein nichtnegativer Eigenvektor existiert (Perron-Frobenius).

6. Wir haben konstruktiv gezeigt, dass es zu jeder M-Matrix A und jedem die Diagonale nicht enthaltenden Nullmuster P eine untere Dreiecksmatrix L mit Einsen in der Diagonalen und eine obere Dreiecksmatrix R gibt derart, dass $A = LR - E$ eine reguläre Zerlegung von A ist und

$$l_{ij} = 0 \quad ((i,j) \in P), \qquad r_{ij} = 0 \quad ((i,j) \in P), \qquad e_{ij} = 0 \quad ((i,j) \notin P).$$

   Man zeige, dass durch diese Forderungen die Faktoren L und R eindeutig bestimmt sind.

7. Sei A eine M-Matrix und L, R und E wie in Satz 3.4 gegeben. Dann konvergiert die durch

$$LRx_{k+1} = Ex_k + b$$

   definierte Folge $\{x_k\}$ für jeden Startvektor $x_0 \in \mathbb{R}^n$ gegen die Lösung des linearen Gleichungssystems $Ax = b$.

8. Eine symmetrische M-Matrix ist positiv definit.

9. Eine symmetrische, positiv definite Matrix $A \in \mathbb{R}^{n\times n}$ mit $a_{ij} \leq 0$, $i \neq j$, heißt eine *Stieltjes-Matrix*. Man zeige, dass eine Stieltjes-Matrix eine M-Matrix ist.

10. Bei gegebenem nichtsingulärem $A \in \mathbb{R}^{m\times n}$ betrachte man die durch

$$f(M) := \tfrac{1}{2} \|I - AM\|_F^2$$

    definierte Abbildung $f\colon \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}$. Wie lautet das Verfahren des steilsten Abstiegs mit exakter Schrittweite?

11. Sei $A \in \mathbb{R}^{n\times n}$, $B := A^{-1}$ und $M = (\ m_1 \ \cdots \ m_n\ ) \in \mathbb{R}^{n\times n}$. Ist dann

$$|b_{ij}| > \|e_j - Am_j\|_1 \max_{k=1,\ldots,n} |b_{ik}|,$$

    so ist $m_{ij} \neq 0$.

Kapitel 6

Lösungen zu den Aufgaben

6.1 Aufgaben in Kapitel 2

6.1.1 Aufgaben in Abschnitt 2.1

1. Sei $\|\cdot\|$ eine Norm auf $\mathbb{R}^n$ bzw. die zugeordnete Matrixnorm auf $\mathbb{R}^{n\times n}$ und $C \in \mathbb{R}^{n\times n}$ nichtsingulär. Man zeige, dass durch $\|x\|_C := \|Cx\|$ eine Norm auf dem $\mathbb{R}^n$ gegeben ist und berechne die zugeordnete Matrixnorm.

   Beweis: ss $\|\cdot\|_C$ eine Norm ist, ist offensichtlich (die Definitheit ist wegen der Nichtsingularität von C gesichert). Dann ist aber

   $$\|A\|_C := \max_{x\neq 0} \frac{\|Ax\|_C}{\|x\|_C} = \max_{x\neq 0} \frac{\|CAx\|}{\|Cx\|} = \max_{y\neq 0} \frac{\|CAC^{-1}y\|}{\|y\|} = \|CAC^{-1}\|.$$

   Also ist durch $\|A\|_C = \|CAC^{-1}\|$ die zugeordnete Matrixnorm gegeben. □

2. Ist $A \in \mathbb{R}^{n\times n}$ symmetrisch, so ist

   $$\lambda_{\min}(A)\,\|x\|_2^2 \leq x^T Ax \leq \lambda_{\max}(A)\,\|x\|_2^2 \qquad \text{für alle } x \in \mathbb{R}^n,$$

   wobei $\lambda_{\min}(A)$ den kleinsten und $\lambda_{\max}(A)$ den größten Eigenwert von A bedeutet.

   Beweis: Seien

   $$\lambda_{\max}(A) = \lambda_1 \geq \cdots \geq \lambda_n = \lambda_{\min}(A)$$

   die Eigenwerte von A und $U \in \mathbb{R}^{n\times n}$ eine orthogonale Matrix mit

   $$U^T AU = \Lambda := \operatorname{diag}(\lambda_1, \ldots, \lambda_n).$$

   Sei $x \in \mathbb{R}^n$ beliebig. Dann ist

   $$x^T Ax = (U^T x)^T \Lambda U^T x = \sum_{j=1}^{n} \lambda_j (U^T x)_j^2 \leq \lambda_{\max}(A)\,\|U^T x\|_2^2 = \lambda_{\max}(A)\,\|x\|_2^2,$$

   wobei wir benutzt haben, dass die euklidische Norm eines Vektors bei Transformation mit einer orthogonalen Matrix invariant bleibt. Die linke Ungleichung kann entsprechend bewiesen werden. □

3. Sei $A \in \mathbb{R}^{m\times n}$. Dann gilt:

(a) $\|A\|_2 = \|A^T\|_2 \le \|A\|_F = \|A^T\|_F$,

(b) $\|A\|_\infty \le \sqrt{n}\,\|A\|_2$,

(c) $\|A\|_2 \le \sqrt{m}\|A\|_\infty$.

(d) $\|A\|_2 \le \sqrt{\|A\|_1\,\|A\|_\infty}$.

Beweis: Wir wissen, dass $\|A\|_2 = \sqrt{\lambda_{\max}(A^TA)}$ und folglich $\|A^T\|_2 = \sqrt{\lambda_{\max}(AA^T)}$. Zu zeigen ist also, dass die größten Eigenwerte von A^TA und AA^T übereinstimmen. Sei hierzu λ ein Eigenwert von AA^T, wir zeigen, dass $\lambda \le \lambda_{\max}(A^TA)$. O. B. d. A. ist $\lambda > 0$, mit x wird ein zugehöriger Eigenvektor bezeichnet. Wegen $AA^Tx = \lambda x$ und $\lambda \ne 0$ ist $A^Tx \ne 0$. Daher ist λ auch Eigenwert von AA^T (mit zugehörigem Eigenvektor A^Tx). Daher ist $\lambda \le \lambda_{\max}(A^TA)$ für jeden Eigenwert λ von AA^T, damit $\lambda_{\max}(AA^T) \le \lambda_{\max}(A^TA)$. Aus Symmetriegründen folgt auch die umgekehrte Ungleichung. Damit ist die erste Gleichung in (a) bewiesen. Für beliebiges $x \in \mathbb{R}^n$ ist unter Benutzung der Cauchy-Schwarzschen Ungleichung

$$\|Ax\|_2^2 = \sum_{i=1}^m (Ax)_i^2 = \sum_{i=1}^m \Big(\sum_{j=1}^n a_{ij}x_j\Big)^2 \le \sum_{i=1}^m \Big(\sum_{j=1}^n a_{ij}^2 \sum_{j=1}^n x_j^2\Big) = \|A\|_F^2\,\|x\|_2^2.$$

Dann ist aber

$$\|A\|_2 := \max_{x\ne 0} \frac{\|Ax\|_2}{\|x\|_2} \le \|A\|_F.$$

Die letzte Beziehung in (a) ist trivial.

Für beliebiges $x \in \mathbb{R}^n$ ist

$$\|Ax\|_\infty \le \|Ax\|_2 \le \|A\|_2\,\|x\|_2 \le \sqrt{n}\,\|A\|_2\,\|x\|_\infty$$

und daher $\|A\|_\infty \le \sqrt{n}\,\|A\|_2$. (c) folgt entsprechend.

Der Beweis von (d) ist etwas schwieriger[1]. Wir führen die Zeilen- und Betragssummen ein:

$$r_i := \sum_{j=1}^n |a_{ij}| \quad (i = 1,\dots,m), \qquad s_j := \sum_{i=1}^m |a_{ij}| \quad (j = 1,\dots,n).$$

Dann ist

$$\|A\|_\infty = \max_{i=1,\dots,m} r_i, \qquad \|A\|_1 := \max_{j=1,\dots,n} s_j.$$

Definiert man λ_j bei festem $i \in \{1,\dots,m\}$ mit $r_i \ne 0$ durch $\lambda_j := |a_{ij}|/r_i$, so ist natürlich $\lambda_j \ge 0$, $j = 1,\dots,n$, und $\sum_{j=1}^n \lambda_j = 1$. Mit einem beliebigen $x \in \mathbb{R}^n$ ist wegen der Konvexität der Funktion $f(t) := t^2$ für alle $i \in \{1,\dots,m\}$ mit $r_i \ne 0$ offenbar

$$\left(\frac{|(Ax)_i|}{r_i}\right)^2 \le \left(\sum_{j=1}^n \frac{|a_{ij}|}{r_i}|x_j|\right)^2 \le \sum_{j=1}^n \frac{|a_{ij}|}{r_i}x_j^2,$$

so dass

$$(Ax)_i^2 \le r_i \sum_{j=1}^n |a_{ij}|\,x_j^2 \le [\max_{i=1,\dots,m} r_i] \sum_{j=1}^n |a_{ij}|\,x_j^2 = \|A\|_\infty \sum_{j=1}^n |a_{ij}|\,x_j^2.$$

[1] Wir spezialisieren den Beweisgang bei J. Todd (1976, S. 25–26) *Basic Numerical Mathematics Vol. 2.* Birkhäuser, Basel.

Eine Summation über alle i (die letzte Beziehung ist auch für $r_i = 0$ richtig) liefert

$$\|Ax\|_2^2 \le \|A\|_\infty \sum_{i=1}^m \sum_{j=1}^n |a_{ij}|\, x_j^2 = \|A\|_\infty \sum_{j=1}^n \Big(\underbrace{\sum_{i=1}^m |a_{ij}|}_{=c_j} \Big) x_j^2 \le \|A\|_1 \, \|A\|_\infty, \|x\|_2^2.$$

Hieraus folgt die Behauptung. □

4. Sei $A \in \mathbb{R}^{m\times n}$, ferner $Q \in \mathbb{R}^{m\times m}$ und $Z \in \mathbb{R}^{n\times n}$ orthogonal. Dann ist $\|QAZ\|_2 = \|A\|_2$ und $\|QAZ\|_F = \|A\|_F$.

   Beweis: Es ist

$$\|QAZ\|_2^2 = \lambda_{\max}(Z^T A^T Q^T QAZ) = \lambda_{\max}(Z^T A^T AZ) = \lambda_{\max}(A^T A) = \|A\|_2^2.$$

   Die entsprechende Aussage für die Frobeniusnorm erkennt man am einfachsten, wenn beachtet, dass $\|A\|_F = \operatorname{tr}(A^T A)$ für ein beliebiges $A \in \mathbb{R}^{m\times n}$, wobei $\operatorname{tr}(B)$ für eine quadratische Matrix B die Spur von B, also die Summe der Diagonalelemente von B bezeichnet und außerdem beachtet, dass die Spur einer Matrix, die ja bekanntlich die Summe aller Eigenwerte der Matrix ist, unter einer Ähnlichkeitstransformation invariant bleibt. □

5. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit. Man zeige, dass es positive Konstanten c_0, C_0 mit

$$c_0 \|Ax\|_2^2 \le x^T Ax \le C_0 \|Ax\|_2^2 \qquad \text{für alle } x \in \mathbb{R}^n$$

   gibt. Insbesondere gilt: Ist $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit, so folgt aus $x^T Ax = 0$, dass $Ax = 0$.

   Beweis: Wir können o. B. d. A. annehmen, dass $A \neq 0$. Daher besitzt A mindestens einen positiven Eigenwert. Seien $\lambda_1 \ge \cdots \ge \lambda_r$ die positiven, $\lambda_{r+1} = \cdot = \lambda_n = 0$ die restlichen Eigenwerte. Ferner sei $\{u_1, \ldots, u_n\}$ ein zugehöriges Orthonormalsystem von Eigenvektoren. Ein beliebiges $x \in \mathbb{R}^n$ hat die eindeutige Darstellung

$$x = \sum_{j=1}^n \alpha_j u_j.$$

   Dann ist offenbar

$$x^T Ax = \sum_{j=1}^n \lambda_j \alpha_j^2 = \sum_{j=1}^r \lambda_j \alpha_j^2$$

   und

$$\|Ax\|_2^2 = \sum_{j=1}^n \lambda_j^2 \alpha_j^2 = \sum_{j=1}^r \lambda_j^2 \alpha_j^2.$$

   Hieraus liest man ab, dass die Aussage mit $c_0 := 1/\lambda_1$ und $C_0 := 1/\lambda_r$ richtig ist. □

6. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv (semi)definit. Man zeige, dass es eine symmetrische und positiv (semi)definite Matrix $B \in \mathbb{R}^{n\times n}$ mit $B^2 = A$ gibt. (Man nennt B dann (nichtnegative bzw. positive) Quadratwurzel aus A und schreibt hierfür $A^{1/2}$.)

Beweis: Die Existenz der Quadratwurzel aus der symmetrischen, positiv (semi)definiten Matrix A ist leicht einzusehen (der Beweis der Eindeutigkeit ist wesentlich schwieriger, siehe Aufgabe 8). Es existiert eine orthogonale Matrix $U \in \mathbb{R}^{n\times n}$ mit

$$U^T A U = \Lambda := \operatorname{diag}(\lambda_1, \ldots, \lambda_n),$$

wobei $\lambda_1, \ldots, \lambda_n$ die Eigenwerte von A sind. Man definiere

$$\Lambda^{1/2} := \operatorname{diag}(\lambda_1^{1/2}, \ldots, \lambda_n^{1/2})$$

und anschließend

$$B := U\Lambda^{1/2}U^T.$$

Offensichtlich ist B symmetrisch und positiv (semi)definit, ferner ist

$$B^2 = BB = U\Lambda^{1/2}\underbrace{U^T U}_{=I}\Lambda^{1/2}U^T = U\Lambda^{1/2}\Lambda^{1/2}U^T = U\Lambda U^T = A,$$

also B eine Quadratwurzel aus A. □

7. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit. Dann gilt die *verallgemeinerte Cauchy-Schwarzsche Ungleichung*, dass nämlich für alle $x, y \in \mathbb{R}^n$ die Ungleichung

$$(x^T A y)^2 \le (x^T A x)(y^T A y)$$

besteht.

Beweis: Der Beweis ist ganz ähnlich dem für die gewöhnliche Cauchy-Schwarzsche Ungleichung. Für $\lambda \in \mathbb{R}$ definieren wir

$$x_\lambda := x + \lambda(x^T A y)\, y.$$

Dann ist

$$0 \le x_\lambda^T A x_\lambda = x^T A x + 2\lambda\,(x^T A y)^2 + \lambda^2\,(x^T A y)^2\,(y^T A y).$$

O. B. d. A. ist $y^T A y > 0$, denn andernfalls ist $Ay = 0$ und die Ungleichung trivialerweise richtig. Setzt man nun $\lambda := -1/(y^T A y)$, so erhält man

$$0 \le x^T A x - \frac{(x^T A y)^2}{y^T A y}$$

und damit die verallgemeinerte Cauchy-Schwarzsche Ungleichung. □

8. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit. Es sei

$$U^T A U = \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n),$$

wobei $U \in \mathbb{R}^{n\times n}$ orthogonal ist und $\lambda_1, \ldots, \lambda_n$ die (nichtnegativen) Eigenwerte von A sind. Mit $\Lambda^{1/2} := \operatorname{diag}(\lambda_1^{1/2}, \ldots, \lambda_n^{1/2})$ sei die symmetrische, positiv semidefinite Matrix $A^{1/2}$ definiert durch $A^{1/2} := U\Lambda^{1/2}U^T$. Sei $B \in \mathbb{R}^{n\times n}$ eine weitere, symmetrische und positiv semidefinite Matrix mit $B^2 = A$. Man zeige:

   (a) Die Matrizen B und A sind vertauschbar.

   (b) B ist mit jedem Polynom in A vertauschbar.

(c) B und $A^{1/2}$ sind vertauschbar.

(d) Es ist $B = A^{1/2}$. Insgesamt existiert zu einer symmetrischen, positiv semidefiniten Matrix genau eine symmetrische, positiv semidefinite Quadratwurzel.

Beweis: Es ist $BA = B^3 = B^2B = AB$, also sind A und B vertauschbar. Dann ist B aber auch mit jeder Potenz von A vertauschbar, also $BA^k = A^kB$ für alle $k \in \mathbb{N}$. Dies sieht man leicht durch vollständige Induktion nach k ein. Der Induktionsanfang bei $k = 1$ ist gerade eben gelegt worden. Wir nehmen an, es sei $BA^k = A^kB$. Dann ist

$$BA^{k+1} = BAA^k = ABA^k = AA^kB = A^{k+1}B,$$

womit der Induktionssschritt vollzogen ist. Nun sei $p_k \in \Pi_k$ ein Polynom vom Grad $\leq k$. Dann ist B natürlich auch mit $p_k(A)$ vertauschbar. Für den dritten Schritt überlegen wir uns, dass es eine Folge $\{p_k\} \subset \Pi$ von Polynomen mit $A^{1/2} = \lim_{k\to\infty} p_k(A)$ gibt. Hierzu beachten wir, dass es wegen des Weierstraßschen Approximationssatzes eine Folge $\{p_k\}$ von Polynomen gibt derart, dass

$$\lim_{k\to\infty} \max_{t\in[0,\lambda_{\max}(A)]} |\sqrt{t} - p_k(t)| = 0,$$

die also auf $[0, \lambda_{\max}(A)]$ gleichmässig gegen $\sqrt{t}$ konvergiert. Dann ist

$$\begin{aligned} \|A^{1/2} - p_k(A)\|_2 &= \|U[\Lambda^{1/2} - p_k(\Lambda)]U^T\|_2 \\ &= \|\Lambda^{1/2} - p_k(\Lambda)\|_2 \\ &= \max_{i=1,\ldots,n} |\lambda_i^{1/2} - p_k(\lambda_i)| \\ &\leq \max_{t\in[0,\lambda_{\max}(A)]} |\sqrt{t} - p_k(t)|, \end{aligned}$$

woraus $\lim_{k\to\infty} p_k(A) = A^{1/2}$ folgt. Wegen (b) ist dann aber

$$BA^{1/2} = B[\lim_{k\to\infty} p_k(A)] = \lim_{k\to\infty} Bp_k(A) = \lim_{k\to\infty} p_k(A)B = [\lim_{k\to\infty} p_k(A)]B = A^{1/2}B.$$

Damit ist auch (c) bewiesen. Nun kommen wir zum Beweis von (d), der Eindeutigkeitsaussage[2]. Hierzu geben wir uns ein $x \in \mathbb{R}^n$ beliebig vor und setzen $y := (A^{1/2} - B)x$. B. Dann ist

$$\begin{aligned} y^TA^{1/2}y + y^TBy &= y^T(A^{1/2} + B)y \\ &= y^T(A^{1/2} + B)(A^{1/2} - B)x \\ &= y^T(A - B^2)x \\ &\qquad \text{(da } B \text{ mit } A^{1/2} \text{ vertauschbar)} \\ &= 0. \end{aligned}$$

Also ist $y^TA^{1/2}y = 0$ und $y^TBy = 0$. Dann ist aber (siehe Aufgabe 5) $A^{1/2}y = 0$ und $By = 0$, folglich $(A^{1/2} - B)y = 0$. Damit ist

$$\|(A^{1/2} - B)x\|_2^2 = x^T(A^{1/2} - B)^2x = x^T(A^{1/2} - B)y = 0,$$

also ist $A^{1/2}x = Bx$ für alle $x \in \mathbb{R}^n$ und $A^{1/2} = B$. □

[2]Der folgende Beweis orientiert sich an einem funktionalanalytischen Beweis für eine entsprechende Aussage im Hilbertraum. Man findet ihn bei

F. Riesz, B. Sz.-Nagy (1956, S. 249 ff.) *Vorlesungen über Funktionalanalysis.* VEB Deutscher Verlag der Wissenschaften, Berlin.

9. Seien $A, B \in \mathbb{R}^{n\times n}$ symmetrisch, positiv (semi)definit und $AB = BA$, also A und B vertauschbar. Dann ist auch AB symmetrisch und positiv (semi)definit.

   Beweis: Zunächst ist AB natürlich symmetrisch, denn

   $$(AB)^T = B^T A^T = BA = AB.$$

   Da A und B miteinander vertauschbar sind, sind auch A und $B^{1/2}$ miteinander vertauschbar (siehe Aufgabe 8c bzw. deren Beweis). Für ein beliebiges $x \in \mathbb{R}^n$ ist dann

   $$x^T ABx = x^T AB^{1/2}B^{1/2}x = x^T B^{1/2}AB^{1/2}x = (B^{1/2}x)^T A(B^{1/2}x) \geq 0,$$

   womit die Behauptung bewiesen ist. □

10. Seien $A, B \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Dann[3] existiert genau eine symmetrische, positiv definite Matrix $X \in \mathbb{R}^{n\times n}$ mit $AX + XA = B$. Weiter zeige man: Sind A und B zusätzlich miteinander vertauschbar, so ist $X := \frac{1}{2}BA^{-1}$ symmetrisch, positiv definit und genügt der Gleichung $AX + XA = B$.

    Hinweis: Diese Aufgabe scheint nicht ganz einfach zu lösen zu sein. Wer findet einen "elementaren" Beweis?

    Beweis: Wir benutzen ohne weiteren Kommentar Hilfsmittel aus der Theorie gewöhnlicher, linearer Differentialgleichungssysteme mit konstanten Koeffizienten. Mit $Y(t) := \exp(-At)$ bezeichnen wir das normierte Fundamentalsystem zum linearen Differentialgleichungsproblem $x' + Ax = 0$. Es ist also

    $$\frac{d}{dt}Y(t) + AY(t) = 0, \qquad Y(0) = 0.$$

    Nun setze man

    $$X := \int_0^\infty \exp(-At)B\exp(-At)\,dt.$$

    Die Matrix X ist wohldefiniert, symmetrisch und positiv definit. Hierbei berücksichtige man, dass

    $$\exp(-At)^T = \exp(-A^T t) = \exp(-At).$$

    Weiter ist

    $$\begin{aligned}
    AX + XA &= \int_0^\infty [A\exp(-At)B\exp(-At) + \exp(-At)B\exp(-At)A]\,dt \\
    &= -\int_0^\infty \left[\left(\frac{d}{dt}\exp(-At)\right)B\exp(-At) + \exp(-At)B\frac{d}{dt}\exp(-At)\right]dt \\
    &\qquad \text{(denn } \exp(-At) \text{ und } A \text{ sind miteinander vertauschbar)} \\
    &= -\int_0^\infty \frac{d}{dt}[\exp(-At)B\exp(-At)]\,dt \\
    &= B,
    \end{aligned}$$

[3] In einem gewissen Sinne gilt auch eine Umkehrung der folgenden Aussage. Ist nämlich $A \in \mathbb{R}^{n\times n}$ symmetrisch und existiert eine symmetrische, positiv definite Matrix X derart, dass $AX + XA$ positiv definit ist, so ist auch A positiv definit. Denn sei λ ein (notwendig reeller) Eigenwert von A mit zugehörigem Eigenvektor x. Dann ist

$$0 < x^T(AX + XA)x = (Ax)^T Xx + (Xx)^T Ax = \lambda x^T Xx + \lambda\,(Xx)^T x = 2\lambda\,\underbrace{x^T Xx}_{>0},$$

also $\lambda > 0$ bzw. A positiv definit.

also ist X eine gesuchte Lösung. Damit ist die Existenzfrage geklärt. Für die Eindeutigkeit genügt es zu zeigen, dass bei vorgegebenem symmetrische, positiv definiten $A \in \mathbb{R}^{n\times n}$ die Gleichung $AX + XA = 0$ die Nullmatrix $X = 0$ als einzige symmetrische Lösung besitzt. Sei $u \in \mathbb{R}^n$ ein Eigenvektor von A mit dem (notwendigerweise positiven) Eigenwert λ. Dann ist

$$0 = (AX + XA)u = AXu + \lambda Xu = (A + \lambda I)Xu,$$

woraus $Xu = 0$ folgt, da $A + \lambda I$ positiv definit ist. Da die Eigenvektoren von A eine Basis des $\mathbb{R}^n$ bilden, ist $X = 0$, womit die Eindeutigkeitsaussage bewiesen ist. Nun wird angenommen, dass die symmetrischen, positiv definiten Matrizen A und B miteinander vertauschbar sind. Dann sind auch B und A^{-1} miteinander vertauschbar. Wegen der Aussage in Aufgabe 9 ist dann $X := \frac{1}{2}BA^{-1}$ symmetrisch und positiv definit. Weiter ist

$$AX + XA - B = \tfrac{1}{2}[ABA^{-1} + B] - B = \tfrac{1}{2}[BAA^{-1} + B] - B = 0,$$

womit die Aussage bewiesen ist. □

11. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Man entwickle das Newton-Verfahren zur Berechnung der positiven Quadratwurzel $A^{1/2}$ und zeige für einen Startwert X_0, der symmetrisch, positiv definit und mit A vertauschbar ist (also z. B. $X_0 := I$ oder $X_0 := A$) die quadratische Konvergenz. Als Beispiel berechne man die symmetrische, positiv definite Quadratwurzel aus

$$A := \begin{pmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{pmatrix}.$$

Hinweis: Auch für diese Aufgabe werden Vorkenntnisse benötigt, da man insbesondere wissen sollte, was das Newton-Verfahren ist.

Beweis: Sei $\mathcal{S}_n$ der lineare Raum der symmetrischen $n \times n$-Matrizen. Wir definieren $f\colon \mathcal{S}_n \longrightarrow \mathcal{S}_n$ durch $f(X) := X^2 - A$ und suchen in der Teilmenge der positiv definiten Matrixen eine Nullstelle dieser Abbildung. Wir berechnen die Fréchet-Ableitung von f in X. Mit einem Inkrement $H \in \mathcal{S}_n$ ist

$$f(X + H) - f(X) = (X + H)^2 - X^2 = XH + HX + H^2,$$

woraus man erkennt, dass die Fréchet-Ableitung durch

$$f'(X)H = XH + HX$$

gegeben ist. Das Newton-Verfahren lautet

$$f'(X_k)(X_{k+1} - X_k) = -f(X_k).$$

Einsetzen von f und f' liefert

$$X_k(X_{k+1} - X_k) + (X_{k+1} - X_k)X_k = -X_k^2 + A$$

bzw.

$$X_kX_{k+1} + X_{k+1}X_k = X_k^2 + A.$$

Ist X_k symmetrisch und positiv definit, so besitzt diese Gleichung wegen Aufgabe 10 genau eine symmetrische und positiv definite Lösung X_{k+1}. Wir betrachten nun speziell den Fall, dass $X_0 \in \mathcal{S}_n$ positiv definit und mit A vertauschbar ist und hiermit das Iterationsverfahren

$$X_k X_{k+1} + X_{k+1} X_k = X_k^2 + A.$$

Wir überlegen uns durch vollständige Induktion nach k, dass X_k mit A vertauschbar ist. Dies ist für $k = 0$ nach Voraussetzung richtig. Ist X_k mit A vertauschbar, so auch mit $X_k^2 + A$. Ferner sind auch A und X_k^{-1} miteinander vertauschbar. Wegen Aufgabe 10 ist

$$X_{k+1} = \tfrac{1}{2}(X_k^2 + A)X_k^{-1} = \tfrac{1}{2}(X_k + AX_k^{-1}).$$

Dann ist

$$\begin{aligned} AX_{k+1} &= \tfrac{1}{2}(AX_k + AAX_k^{-1}) \\ &= \tfrac{1}{2}(X_k A + AX_k^{-1}A) \\ &= \tfrac{1}{2}(X_k + AX_k^{-1})A \\ &= X_{k+1}A. \end{aligned}$$

Damit ist schließlich gezeigt, dass die Folge $\{X_k\}$ (ganz analog zum Fall $n = 1$, also der Bestimmung der positiven Quadratwurzel aus einer positiven, reellen Zahl) durch

$$X_{k+1} := \tfrac{1}{2}(X_k + AX_k^{-1})$$

gegeben ist. Zunächst überlegen wir uns, dass die Folge $\{X_k\}$ konvergiert. Da X_k mit A vertauschbar ist, ist X_k mit jedem Polynom in A vertauschbar und dann auch mit $A^{1/2}$ (siehe den Beweis von Aufgabe 8). Ferner sind auch A und X_k^{-1} miteinander vertauschbar. Dann ist

$$\begin{aligned} \tfrac{1}{2}X_k^{-1}(X_k - A^{1/2})^2 &= \tfrac{1}{2}X_k^{-1}(X_k^2 - X_k A^{1/2} - A^{1/2}X_k + A) \\ &= \tfrac{1}{2}(X_k - 2A^{1/2} + X_k^{-1}A) \\ &= \tfrac{1}{2}(X_k + AX_k^{-1}) - A^{1/2} \\ &= X_{k+1} - A^{1/2}. \end{aligned}$$

Nun ist X_k^{-1} positiv definit, $(X_k - A^{1/2})^2$ positiv semidefinit, ferner sind beide Matrizen miteinander vertauschbar. Wegen Aufgabe 9 ist dann auch $X_{k+1} - A^{1/2}$ symmetrisch und positiv semidefinit. Um Schreibarbeit zu sparen und die Argumentation suggestiver zu machen, führen wir auf $\mathcal{S}_n$ eine Halbordnung $\preceq$ ein, indem wir $X \preceq Y$ genau dann schreiben, wenn $Y - X$ positiv semidefinit ist. Die Beziehung $\succeq$ ist entsprechend zu verstehen. Eben haben wir also bewiesen, dass $A^{1/2} \preceq X_{k+1}$, $k = 0, 1, \ldots$. Weiter ist

$$X_k - X_{k+1} = \tfrac{1}{2}(X_k - AX_k^{-1}) \succeq 0, \qquad k = 1, 2, \ldots.$$

Dies sieht man folgendermaßen ein: Für $k \in \mathbb{N}$ folgt aus $A^{1/2} \preceq X_k$, dass $A \preceq A^{1/2}X_k$ und $AX_k^{-1} \preceq A^{1/2} \preceq X_k$. In der Halbordnung $\preceq$ auf $\mathcal{S}_n$ ist also $\{X_k\}$ spätestens nach dem ersten Schritt eine nach unten durch $A^{1/2}$ beschränkte, monoton nicht wachsende

Folge. Hieraus wollen wir schließen, dass die Folge $\{X_k\}$ konvergiert[4]. Hierzu geben wir uns ein belebiges $u \in \mathbb{R}^n$ vor. Die Folge $\{u^T X_k u\}$ ist nach unten durch $u^T A^{1/2} u$ beschränkt und spätestens nach dem ersten Glied monoton nicht wachsend, folglich konvergent und insbesondere eine Cauchy-Folge, woraus folgen wird, dass auch $\{X_k u\}$ eine Cauchy-Folge ist. Denn für $k \le l$ ist $X_k \succeq X_l$ und daher

$$\begin{aligned}
\|X_k u - X_l u\|_2^4 &= [u^T (X_k - X_l)(X_k - X_l) u]^2 \\
&\le (u^T (X_k - X_l) u)\, ((X_k - X_l) u)^T (X_k - X_l)^2 u \\
&\qquad \text{(verallgemeinerte Cauchy-Schwarzsche Ungleichung)} \\
&= (u^T (X_k - X_l) u)\, \|X_k - X_l\|_2^3\, \|u\|_2^2 \\
&\le [u^T X_l u - u^T X_k u]\, \|X_1\|_2^3\, \|u\|_2^2.
\end{aligned}$$

Definiert man $X^* \in \mathbb{R}^{n\times n}$ durch $X^* u := \lim_{k\to\infty} X_k u$, so ist $X^* \in \mathcal{S}_n$, also X^* eine symmetrische Matrix, ferner ist X^* positiv definit wegen $u^T X^* u \ge u^T A^{1/2} u$. Weiter folgt[5] aus $X_k u \to X^* u$ auch $X_k \to X^*$. Aus der Iterationsvorschrift des Newton-Verfahrens erhalten wir nach dem Grenzübergang $k \to \infty$, dass X^* eine Quadratwurzel aus A ist. Da es nach Aufgabe 8d zu der symmetrischen, positiv definiten Matrix A genau eine symmetrische, positiv definite Quadratwurzel $A^{1/2}$ gibt, ist $X^* = A^{1/2}$. Damit ist die Konvergenz der Folge $\{X_k\}$ bewiesen. Wegen

$$X_{k+1} - A^{1/2} = \tfrac{1}{2} X_k^{-1} (X_k - A^{1/2})^2, \qquad 0 \prec A^{1/2} \preceq X_k$$

ist

$$\|X_{k+1} - A^{1/2}\|_2 \le \tfrac{1}{2}\, \|X_k^{-1}\|_2\, \|X_k - A^{1/2}\|_2^2 \le \tfrac{1}{2}\, \|A^{-1/2}\|_2\, \|X_k - A^{1/2}\|_2^2,$$

womit die quadratische Konvergenz der Folge $\{X_k\}$ gegen $A^{1/2}$ bewiesen ist.

Das Newton-Verfahren zur Berechnung der symmetrischen, positiv definiten Quadratwurzel aus der angegebenen Matrix A starten wir mit $X_0 := I$ (die Wahl $X_0 := A$ liefert natürlich dieselben Ergebnisse) und erhalten

$$X_1 = \begin{pmatrix} 2.5 & -0.5 & -0.5 & 0.0 \\ -0.5 & 2.5 & 0.0 & -0.5 \\ -0.5 & 0.0 & 2.5 & -0.5 \\ 0.0 & -0.5 & -0.5 & 2.5 \end{pmatrix}$$

und schließlich

$$X_5 = \begin{pmatrix} 1.9659258263 & -0.2588190451 & -0.2588190451 & -0.0340741737 \\ -0.2588190451 & 1.9659258263 & -0.0340741737 & -0.2588190451 \\ -0.2588190451 & -0.0340741737 & 1.9659258263 & -0.2588190451 \\ -0.0340741737 & -0.2588190451 & -0.2588190451 & 1.9659258263 \end{pmatrix}.$$

Auf die angebenen Dezimalen stimmt X_5 mit der Matrix überein, die man in MATLAB durch `sqrtm(A)` erhält. □

[4]Wieder orientieren wir uns an einem allgemeineren Resultat für symmetrische Operatoren im Hilbertraum, das man bei

F. Riesz, B. Sz.-Nagy (1956, S. 249 ff.) *Vorlesungen über Funktionalanalysis.* VEB Deutscher Verlag der Wissenschaften, Berlin

findet.

[5]Dies ist ein Schluss, der in einem unendlichdimensionalen Hilbertraum i. allg. nicht richtig ist.

12. Sei $s \in \mathbb{R}^n \setminus \{0\}$ und $E \in \mathbb{R}^{n\times n}$. Man zeige, dass

$$\left\| E\left(I - \frac{ss^T}{s^Ts}\right)\right\|_F^2 = \|E\|_F^2 - \frac{\|Es\|_2^2}{\|s\|_2^2}.$$

Beweis: Es ist

$$\begin{aligned}
\left\| E\left(I - \frac{ss^T}{s^Ts}\right)\right\|_F^2 &= \operatorname{tr}\left(\left(I - \frac{ss^T}{s^Ts}\right)E^TE\left(I - \frac{ss^T}{s^Ts}\right)\right)\\
&= \operatorname{tr}\left(E^TE - \frac{s(E^TEs)^T}{\|s\|_2^2} - \frac{(E^TEs)s^T}{\|s\|_2^2} + \frac{\|Es\|_2^2}{\|s\|_2^4}ss^T\right)\\
&= \operatorname{tr}(E^TE) - \frac{\|Es\|_2^2}{\|s\|_2^2} - \frac{\|Es\|_2^2}{\|s\|_2^2} + \frac{\|Es\|_2^2}{\|s\|_2^2}\\
&= \|E\|_F^2 - \frac{\|Es\|_2^2}{\|s\|_2^2}.
\end{aligned}$$

Hierbei haben wir benutzt, dass die Spur eine lineare Abbildung auf dem linearen Raum aller $n \times n$-Matrizen ist. □

6.1.2 Aufgaben in Abschnitt 2.2

1. Was ist die Kondition für das Problem, $\sqrt{x}$ für $x > 0$ zu berechnen?

Lösung: Mit $f(x) := \sqrt{x}$ ist

$$\frac{|f'(x)|\,|x|}{|f(x)|} = \frac{1}{2},$$

die Aufgabe ist also gut konditioniert.

2. Sei $\|\cdot\|$ eine *absolute Norm* auf dem $\mathbb{R}^n$, d. h. $\|x\| = \|\,|x|\,\|$ für alle $x \in \mathbb{R}^n$, mit $\|\cdot\|$ sei auch die zugeordnete Matrixnorm bezeichnet. Die Matrix $A \in \mathbb{R}^{n\times n}$ sei nichtsingulär und x die Lösung von $Ax = b$ mit $b \in \mathbb{R}^n \setminus \{0\}$. Mit $\epsilon > 0$ sei $\delta A \in \mathbb{R}^{n\times n}$ eine Matrix mit $|\delta A| \le \epsilon\,|A|$. Sei $(A + \delta A)\hat{x} = b$. Man zeige, dass

$$\frac{\|\hat{x} - x\|}{\|\hat{x}\|} \le \epsilon\,\|\,|A^{-1}|\,|A|\,\|.$$

Beweis: Aus $(A + \delta A)\hat{x} = b$ und $Ax = b$ folgt durch Subtraktion $A(\hat{x} - x) = -(\delta A)\hat{x}$ und hieraus

$$\hat{x} - x = -A^{-1}(\delta A)\hat{x}.$$

Geht man komponentenweise zu Absolutbeträgen über und wendet mehrmals die Dreiecksungleichung an, so erhält man

$$|\hat{x} - x| \le |A^{-1}(\delta A)\hat{x}| \le |A^{-1}|\,|\delta A|\,|\hat{x}| \le \epsilon\,|A^{-1}|\,|A|\,|\hat{x}|.$$

Dann ist

$$\|\hat{x} - x\| \le \epsilon\,\|\,|A^{-1}|\,|A|\,|\hat{x}|\,\| \le \epsilon\,\|\,|A^{-1}|\,|A|\,\|\,\|\hat{x}\|,$$

woraus[6] die Behauptung folgt. □

3. Sei $U = (u_{ij}) \in \mathbb{R}^{n\times n}$ eine nichtsinguläre obere Dreieckmatrix. Dann ist

$$\left(\min_{i=1,\ldots,n} |u_{ii}|\right)^{-1} \leq \|U^{-1}\|_p, \qquad p = 1, 2, \infty.$$

Beweis: Als Hilfsbehauptung überlegen wir uns:

- Ist $A \in \mathbb{R}^{n\times n}$, so ist $\|A\|_p \geq |a_{ij}|$ für $i, j = 1, \ldots, n$ und $p = 1, 2, \infty$.

Dies ist für $p = 1, \infty$ offensichtlich, denn die der Betragssummen- bzw. Maximumnorm zugeordnete Matrixnorm ist die maximale Spalten- bzw. Zeilenbetragssummennorm. Es bleibt also der Fall $p = 2$. Es ist $\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)}$. Wegen Aufgabe 2 in Abschnitt 2.1 ist

$$\lambda_{\max}(A^T A) \geq \|Ae_j\|_2^2 = \sum_{i=1}^{n} a_{ij}^2, \qquad j = 1, \ldots, n.$$

Also ist auch hier $\|A\|_2 \geq |a_{ij}|$, $i, j = 1, \ldots, n$. Damit ist die Hilfsbehauptung bewiesen.

Dieses Ergebnis wenden wir an mit $A := U^{-1}$ und $i = j$. Es ist $(U^{-1})_{ii} = 1/u_{ii}$ und folglich $\|U^{-1}\|_p \geq (1/|u_{ii}|)$ für $i = 1, \ldots, n$ und $p = 1, 2, \infty$. Daher ist

$$\|U^{-1}\|_p \geq \max_{i=1,\ldots,n} (1/|u_{ii}|) = \left(\min_{i=1,\ldots,n} |u_{ii}|\right)^{-1}, \qquad p = 1, 2, \infty.$$

□

4. Sei $A \in \mathbb{R}^{n\times n}$ nichtsingulär. Ziel ist es, (möglichst gute) untere Schranken für $\|A\|_2$ zu erhalten, wobei nur ausgenutzt werden soll, dass wir Ax und $A^T y$ für beliebige x, y berechnen können. (Liegt eine LU-Zerlegung von A vor, so kann auf diese Weise eine untere Schranke für $\kappa_2(A)$ gewonnen werden). Man betrachte ein Verfahren, von dem wir einen Schritt angeben:

- Gegeben $x \in \mathbb{R}^n$ mit $\|x\|_2 = 1$.
- Berechne $y := Ax$, $\gamma := \|y\|_2$ und $z := (1/\gamma)A^T y$.
- Falls $\|z\|_2 \leq z^T x$, dann: STOP, Andernfalls: Berechne $x_+ := z/\|z\|$.

Man zeige:

(a) Ist die Abbruchbedingung $\|z\|_2 \leq z^T x$ erfüllt, so ist x eine stationäre Lösung für die Aufgabe (jetzt ist x als variabel aufzufassen), $f(x) := \|Ax\|_2^2$ unter der Nebenbedingung $\|x\|_2 \leq 1$ zu maximieren.

(b) Ist die Abbruchbedingung nicht erfüllt, so ist $f(x_+) > f(x)$.

[6] Beim Beweis haben wir ein wenig geschummelt. Wir haben nämlich benutzt, dass für eine absolute Norm $\|\cdot\|$ aus $|x| \leq |y|$ folgt, dass $\|x\| \leq \|y\|$, bzw. eine absolute Norm *monoton* ist. Das ist aber richtig. Denn seien $x, y \in \mathbb{R}^n$ mit $|x| \leq |y|$ gegeben. Ist $D = \text{diag}\,(d_1, \ldots, d_n)$ eine Diagonalmatrix mit $d_j \in \{1, -1\}$, $j = 1, \ldots, n$, so ist $|\,D|y|\,| = |y|$ und daher $\|\,D|y|\,\| = \|\,|y|\,\| = \|y\|$. Anschaulich gesprochen bedeutet dies, dass die 2^n Eckpunkte des Quaders $Q := \{u \in \mathbb{R}^n : -|y| \leq u \leq |y|\}$ zum Rand der Kugel $K := \{u \in \mathbb{R}^n : \|u\| \leq \|y\|\}$ gehören. Nun ist aber $x \in Q$ und lässt sich daher als Konvexkombination der Eckpunkte von Q darstellen. Daher ist auch $x \in K$ bzw. $\|x\| \leq \|y\|$.

Beweis: Die Kuhn-Tucker Bedingungen (notwendige Bedingungen für ein lokales Minimum) sind in x erfüllt, wenn ein λ mit $A^TAx = \lambda x$ existiert bzw. x Eigenvektor von A^TA ist. Wir wollen zeigen, dass dies der Fall ist, wenn die Abbruchbedingung erfüllt ist. Denn ist $\|z\|_2 \leq z^Tx$, so liefert Einsetzen, dass

$$\|A^TAx\| \leq \|Ax\|_2^2 = x^TA^TAx \leq \underbrace{\|x\|_2}_{=1} \|A^TAx\|_2 = \|A^TAx\|_2.$$

Also sind x und A^TAx linear abhängig, hieraus folgt (a).

Sei $\|z\|_2 > z^Tx$. Es ist

$$\|Ax\|_2 = \|y\|_2 = z^Tx < \|z\|_2 = z^Tx_+ = \frac{y^TAx_+}{\|y\|_2} \leq \|Ax_+\|_2,$$

das ist die Behauptung. □

5. Sei $R \in \mathbb{R}^{n\times n}$ eine nichtsinguläre obere Dreiecksmatrix. Man definiere die (ebenfalls nichtsinguläre) obere Dreiecksmatrix $U \in \mathbb{R}^{n\times n}$ durch

$$u_{ij} := \begin{cases} |r_{ii}|, & \text{für } i = j, \\ -|r_{ij}|, & \text{für } i \neq j. \end{cases}$$

Man zeige:

(a) Es ist $U^{-1} \geq 0$.

(b) Es ist $\|R^{-1}\|_\infty \leq \|U^{-1}\|_\infty$.
Hinweis: Man zeige, dass $|R^{-1}| \leq U^{-1}$.

(c) Es ist $\|U^{-1}\|_\infty = \|U^{-1}e\|_\infty$, wobei $e := (1, \ldots, 1)^T$.

(d) Durch den folgenden Algorithmus wird eine obere Schranke γ_U für $\|R^{-1}\|_\infty$ berechnet:

- Für $i = n, \ldots, 1$:
  - $z_i := (1 + \sum_{j=i+1}^{n} |r_{ij}|\, z_j)/|r_{ii}|$.
- $\gamma_U := \|z\|_\infty$.

Beweis: Um $U^{-1} \geq 0$ zu beweisen, überlegen wir uns, dass $x \geq 0$ aus $Ux \geq 0$ folgt. Dies sieht man aber sofort ein, wenn man an das Rückwärtseinsetzen denkt. Man zeigt also einfach der Reihe nach, dass $x_n, \ldots, x_1 \geq 0$. Damit ist der erste Teil der Aufgabe bewiesen.

Für den Beweis des zweiten Teiles überlege man sich, dass $|R^{-1}| \leq U^{-1}$, woraus dann offenbar die Behauptung sofort folgt. Dies wiederum ist äquivalent damit, dass $|R^{-1}b| \leq U^{-1}b$ für alle $b \geq 0$ (lasse etwa b die Einheitsvektoren durchlaufen). Sei daher $b \geq 0$ vorgegeben, definiere $x := R^{-1}b$ und $y := U^{-1}b$. Beide Vektoren werden durch Rückwärtseinsetzen gewonnen, es gilt daher

$$x_i = \left(b_i - \sum_{j=i+1}^{n} r_{ij}x_j\right) \Big/ r_{ii}, \qquad i = n, \ldots, n,$$

und

$$y_i = \left(b_i + \sum_{j=i+1}^{n} |r_{ij}|\, y_j \right) \Big/ |r_{ii}|, \qquad i = n, \ldots, 1.$$

Offenbar ist $|x_n| = b_n/|r_{nn}| = y_n$. Angenommen, es ist $|x_j| \le y_j$, $j = n, \ldots, i+1$. Dann ist

$$|x_i| \le \left(b_i + \sum_{j=i+1}^{n} |r_{ij}|\, |x_j| \right) \Big/ |r_{ii}| \le \left(b_i + \sum_{j=i+1}^{n} |r_{ij}|\, y_j \right) \Big/ |r_{ii}| = y_i,$$

so dass wir durch vollständige Induktion $|x| \le y$ bzw. $|R^{-1}| \le U^{-1}$ bewiesen haben. Hieraus folgt dann (b), während (c) trivial ist. Auch (d) ist inzwischen trivial, denn offenbar ist $\gamma_U = \|U^{-1}e\|_\infty$. □

6. Das lineare Gleichungssystem $Ax = b$ ist äquivalent zu $DAx = Db$, wenn D nichtsingulär ist. Interessant sind hier vor allem Diagonalmatrizen, da dann die Berechnung von DA noch verhältnismäßig billig ist (man spricht auch von einer *Skalierung* des Gleichungssystems). Je kleiner die Kondition der Koeffizientenmatrix eines linearen Gleichungssystems desto bessere Ergebnisse erwarten wir. Dies ist der Hintergrund der folgenden Aufgabe.

   Sei $A \in \mathbb{R}^{n\times n}$ nichtsingulär und $\mathcal{D}_n$ die Menge der nichtsingulären $n \times n$-Diagonalmatrizen. Man zeige:

   (a) Es ist
   $$\| \, |A^{-1}| \, |A| \, \|_\infty \le \min_{D\in\mathcal{D}_n} \kappa_\infty(DA).$$
   Weshalb dürfen wir hier "min" statt "inf" schreiben?

   (b) Definiert man $D^* = \operatorname{diag}(d_1^*, \ldots, d_n^*) \in \mathcal{D}_n$ durch
   $$d_i^* := 1 \Big/ \sum_{j=1}^{n} |a_{ij}|, \qquad i = 1, \ldots, n,$$
   so ist $\| \, |A^{-1}| \, |A| \, \|_\infty = \kappa_\infty(D^*A)$.

   Beide Teile zusammen zeigen, dass eine sogenannte Zeilen-Äquilibrierung bezüglich der Maximumnorm eine optimale Kondition liefert.

   Beweis: Für ein beliebiges $D \in \mathcal{D}_n$ ist

$$\begin{aligned} \| \, |A^{-1}| \, |A| \, \|_\infty &= \| \, |(DA)^{-1}| \, |DA| \, \|_\infty \\ &\le \| \, |(DA)^{-1}| \, \|_\infty \, \| \, |DA| \, \|_\infty \\ &= \|(DA)^{-1}\|_\infty \, \|DA\|_\infty \\ &= \kappa_\infty(DA) \end{aligned}$$

   und daher

$$\| \, |A^{-1}| \, |A| \, \|_\infty \le \inf_{D\in\mathcal{D}_n} \kappa_\infty(DA) = \min_{D\in\mathcal{D}_n} \kappa_\infty(DA).$$

   Beim letzten Schritt haben wir berücksichtigt, dass $\kappa_\infty(\alpha DA) = \kappa_\infty(DA)$ für jedes $\alpha \ne 0$, so dass das Minimum von $\kappa_\infty(DA)$ z. B. nur über die $D \in \mathcal{D}_n$ genommen zu werden braucht, für die $\|D\|_\infty = 1$.

Zum Beweis des zweiten Teils beachten wir, dass

$$\|D^*A\|_\infty = \max_{i=1,\dots,n} d_i^* \sum_{j=1}^n |a_{ij}| = 1.$$

Daher ist

$$\begin{aligned}
\kappa_\infty(D^*A) &= \|(D^*A)^{-1}\|_\infty \\
&= \max_{i=1,\dots,n}\left(\sum_{j=1}^n \frac{1}{d_j^*}|(A^{-1})_{ij}|\right) \\
&= \max_{i=1,\dots,n}\left(\sum_{j=1}^n\sum_{k=1}^n |a_{jk}|\,|(A^{-1})_{ij}|\right) \\
&= \max_{i=1,\dots,n}\left(\sum_{k=1}^n\sum_{j=1}^n |(A^{-1})_{ij}|\,|a_{jk}|\right) \\
&\qquad \text{(Vertauschung der Summationsreihenfolge)} \\
&= \max_{i=1,\dots,n}\left(\sum_{j=1}^n\sum_{k=1}^n |(A^{-1})_{ik}|\,|a_{kj}|\right) \\
&\qquad \text{(Vertauschung von } j \text{ und } k) \\
&= \max_{i=1,\dots,n}\left(\sum_{j=1}^n (|A^{-1}|\,|A|)_{ij}\right) \\
&= \|\,|A^{-1}|\,|A|\,\|_\infty.
\end{aligned}$$

Damit ist die Behauptung bewiesen. □

7. Sei

$$A = \begin{pmatrix} a_1^T \\ \vdots \\ a_n^T \end{pmatrix} \in \mathbb{R}^{n\times n}$$

nichtsingulär. Man zeige:

(a) Es ist $\max_{i=1,\dots,n}\|a_i\|_2 \le \|A\|_2 \le \sqrt{n}\,\max_{i=1,\dots,n}\|a_i\|_2$.

(b) Definiert man

$$\hat{d}_i := \frac{1}{\|a_i\|_2} \quad (i=1,\dots,n), \qquad \hat{D} := \operatorname{diag}(\hat{d}_1,\dots,\hat{d}_n),$$

so ist

$$\kappa_2(\hat{D}A) \le \sqrt{n}\,\min_{D\in\mathcal{D}_n}\kappa_2(DA),$$

wobei $\mathcal{D}_n$ wieder die Menge der nichtsingulären $n\times n$-Diagonalmatrizen bedeutet.

Beweis: Es ist

$$\|A\|_2 = \|A^T\|_2 = \max_{\|x\|_2=1}\|A^Tx\|_2 \ge \|A^Te_i\|_2 = \|a_i\|_2, \qquad i=1,\dots,n,$$

und daher $\max_{i=1,\dots,n} \|a_i\|_2 \le \|A\|_2$. Mit einem beliebigen $x \in \mathbb{R}^n$ ist mit Hilfe der Cauchy-Schwarzschen Ungleichung

$$\|Ax\|_2^2 = \sum_{i=1}^n (a_i^T x)^2 \le \left(\sum_{i=1}^n \|a_i\|_2^2\right) \|x\|_2^2 \le n \left(\max_{i=1,\dots,n} \|a_i\|_2\right)^2 \|x\|_2^2$$

und daher $\|A\|_2 \le \sqrt{n} \max_{i=1,\dots,n} \|a_i\|_2$.

Benutzt man die zweite Ungleichung in (a), so erhält man

$$\|\hat{D}A\|_2 \le \sqrt{n} \max_{i=1,\dots,n} \hat{d}_i \|a_i\|_2 = \sqrt{n}.$$

Mit einem beliebigen $D \in \mathcal{D}_n$ ist ferner

$$\begin{aligned} \|(\hat{D}A)^{-1}\|_2 &= \|A^{-1}D^{-1}D\hat{D}^{-1}\|_2 \\ &\le \|(DA)^{-1}\|_2 \, \|D\hat{D}^{-1}\|_2 \\ &= \|(DA)^{-1}\|_2 \max_{i=1,\dots,n} (|d_i| \, \|a_i\|_2) \\ &\le \|(DA)^{-1}\|_2 \, \|DA\|_2 \\ &= \kappa_2(DA), \end{aligned}$$

wobei wir zur Herleitung der letzten Ungleichung die erste Ungleichung in (a) benutzt haben. Insgesamt ist daher

$$\kappa_2(\hat{D}A) = \|(\hat{D}A)^{-1}\|_2 \, \|\hat{D}A\|_2 \le \sqrt{n}\kappa_2(DA) \qquad \text{für alle } D \in \mathcal{D}_n,$$

womit die Behauptung bewiesen ist. □

6.1.3 Aufgaben in Abschnitt 2.3

1. Sei $A \in \mathbb{R}^{n\times n}$ (strikt zeilenweise) diagonal dominant, d. h.

$$\sum_{\substack{j=1 \\ j\ne i}}^n |a_{ij}| < |a_{ii}|, \qquad i = 1, \dots, n.$$

Dann ist A nichtsingulär, ferner besitzt A eine LU-Zerlegung $A = LU$ mit einer unteren Dreiecksmatrix L, die Einsen auf der Diagonalen hat, und einer oberen Dreiecksmatrix U.

Beweis: Angenommen, es ist $Ax = 0$. Man wähle $i \in \{1, \dots, n\}$ mit $|x_i| = \|x\|_\infty$. Wegen

$$0 = a_{ii}x_i + \sum_{\substack{j=1 \\ j\ne i}}^n a_{ij}x_j$$

folgt aus der diagonalen Dominanz von A, dass

$$|a_{ii}| \, \|x\|_\infty = |a_{ii}| \, |x_i| = \left|\sum_{\substack{j=1 \\ j\ne i}}^n a_{ij}x_j\right| \le \sum_{\substack{j=1 \\ j\ne i}}^n |a_{ij}| \, |x_j| \le \left(\sum_{\substack{j=1 \\ j\ne i}}^n |a_{ij}|\right) \|x\|_\infty.$$

Daher ist

$$\underbrace{\left(|a_{ii}| - \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|\right)}_{>0} \|x\|_\infty \leq 0$$

und damit ist $x = 0$. Folglich ist eine diagonal dominante Matrix nichtsingulär[7]. Auch die Matrizen $(a_{ij})_{1\leq i,j\leq k}$ sind diagonal dominant und daher nichtsingulär. Folglich sind alle Hauptabschnittsdeterminanten von Null verschieden. Dann aber besitzt A bekanntlich eine LU-Zerlegung. □

2. Sei $A \in \mathbb{R}^{n\times n}$ spaltenweise (strikt) diagonal dominant, d. h. es sei

$$\sum_{\substack{i=1\\ i\neq j}}^{n} |a_{ij}| < |a_{jj}|, \qquad j = 1, \ldots, n.$$

Man zeige:

(a) Die Matrix A ist nichtsingulär.

(b) Die Matrix A besitzt nicht nur eine LU-Zerlegung der Form $A = LU$, sondern mehr noch: Das Gaußsche Eliminationsverfahren mit Spaltenpivotsuche benutzt keine Vertauschungen.

Beweis: Die Matrix A^T ist wegen (a) in Aufgabe 1 nichtsingulär, da sie offenbar zeilenweise diagonal dominant ist. Daher ist auch A nichtsingulär, ferner sind alle Hauptabschnittsdeterminanten von Null verschieden.

Da $|a_{11}|$ in der ersten Spalte von A das betragsgrößte Element ist, wird im ersten Schritt des Gaußschen Eliminationsverfahrens mit Spaltenpivotsuche keine Vertauschung vorgenommen. Ist

$$A = \begin{pmatrix} a_{11} & * \\ * & A_{22} \end{pmatrix}$$

mit $A_{22} = (a_{ij})_{2\leq i,j\leq n} \in \mathbb{R}^{(n-1)\times(n-1)}$, so liefert der nächste Schritt des Gaußschen Eliminationsverfahrens die Matrix

$$A^{(2)} = \begin{pmatrix} a_{11} & * \\ 0 & A_{22}^{(2)} \end{pmatrix}$$

mit

$$a_{ij}^{(2)} := a_{ij} - \frac{a_{i1}a_{1j}}{a_{11}}, \qquad 2 \leq i, j \leq n.$$

Wir zeigen, dass auch $A_{22}^{(2)}$ spaltenweise diagonal dominant ist und erhalten durch Induktion die Behauptung. Denn für $j = 2, \ldots, n$ ist

$$\sum_{\substack{i=2\\ i\neq j}}^{n} |a_{ij}^{(2)}| \;=\; \sum_{\substack{i=2\\ i\neq j}}^{n} \left| a_{ij} - \frac{a_{i1}a_{1j}}{a_{11}} \right|$$

[7] Ein alternativer Beweis könnte folgendermaßen aussehen: Die strikte zeilenweise Diagonaldominanz von A ist äquivalent zu $\|I - A_D^{-1}A\|_\infty < 1$, wobei $A_D := \operatorname{diag}(A)$. Dann ist auch $\rho(I - A_D^{-1}A) < 1$, woraus man sofort erhält, dass der Kern von A nur aus dem Nullvektor besteht.

$$
\begin{aligned}
&\leq \sum_{\substack{i=2\\ i\neq j}}^{n} |a_{ij}| + \frac{|a_{1j}|}{|a_{11}|} \sum_{\substack{i=2\\ i\neq j}}^{n} |a_{i1}| \\
&< |a_{jj}| - |a_{1j}| + \frac{|a_{1j}|}{|a_{11}|}[|a_{11}| - |a_{j1}|] \\
&= |a_{jj}| - \left| \frac{a_{j1}a_{1j}}{a_{11}} \right| \\
&\leq \left| a_{jj} - \frac{a_{j1}a_{1j}}{a_{11}} \right| \\
&= |a_{jj}^{(2)}|,
\end{aligned}
$$

womit alles gezeigt ist. □

3. Sei A spaltenweise (strikt) diagonal dominant, d. h.

$$\sum_{\substack{i=1\\ i\neq j}}^{n} |a_{ij}| < |a_{jj}|, \qquad j = 1, \ldots, n.$$

Das Gaußsche Eliminationsverfahren mit Spaltenpivotsuche macht nach Aufgabe 2 keine Zeilenvertauschungen und erzeugt mit $A^{(1)} := A$ Matrizen $A^{(2)}, \ldots, A^{(n-1)}$. Man zeige, dass der Wachstums-Faktor ρ_n durch 2 nach oben beschränkt ist, d. h. dass

$$\max_{k\leq i,j\leq n} |a_{ij}^{(k)}| \leq 2 \max_{1\leq i,j\leq n} |a_{ij}|, \qquad k = 1, \ldots, n-1.$$

Beweis: Wir zeigen die Aussage zunächst für $k = 2$. Es ist

$$A^{(2)} = \begin{pmatrix} a_{11} & * \\ * & A_{22}^{(2)} \end{pmatrix},$$

wobei auch $A_{22}^{(2)}$ spaltenweise diagonal dominant ist, wie wir aus dem Beweis von Aufgabe 2 wissen. Für $j = 2, \ldots, n$ ist

$$
\begin{aligned}
\sum_{i=2}^{n} |a_{ij}^{(2)}| &= \sum_{i=2}^{n} \left| a_{ij} - \frac{a_{i1}a_{1j}}{a_{11}} \right| \\
&\leq \sum_{i=2}^{n} |a_{ij}| + \frac{|a_{1j}|}{|a_{11}|} \underbrace{\sum_{i=2}^{n} |a_{i1}|}_{<|a_{11}|} \\
&\leq \sum_{i=1}^{n} |a_{ij}|.
\end{aligned}
$$

Da auch $A_{22}^{(2)}$ spaltenweise strikt diagonal dominant ist, kann man dieses Argument fortsetzen und erhält induktiv

$$\sum_{i=k}^{n} |a_{ij}^{(k)}| \leq \sum_{i=1}^{n} |a_{ij}|, \qquad j = k, \ldots, n, \quad k = 2, \ldots, n-1.$$

Für $k = 1, \ldots, n-1$ liefert dies

$$\begin{aligned}
\max_{k\le i,j\le n} |a_{ij}^{(k)}| &\le \max_{j=k,\ldots,n} \sum_{i=k}^{n} |a_{ij}^{(k)}| \\
&\le \max_{j=k,\ldots,n} \sum_{i=1}^{n} |a_{ij}| \\
&= \max_{j=k,\ldots,n} \Bigg[|a_{jj}| + \underbrace{\sum_{\substack{i=1 \\ i\ne j}} |a_{ij}|}_{<|a_{jj}|} \Bigg] \\
&\le 2 \max_{j=k,\ldots,n} |a_{jj}| \\
&\le 2 \max_{1\le i,j\le n} |a_{ij}|,
\end{aligned}$$

das war zu zeigen. □

4. Sei $A \in \mathbb{R}^{n\times n}$ nichtsingulär und $PA = LU$ eine durch das Gaußsche Eliminationsverfahren mit Spaltenpivotsuche gewonnene LU-Zerlegung von A (dabei P, L und U wie üblich). Man zeige, dass

$$\frac{\|A^{-1}\|_\infty}{2^{n-1}} \le \|U^{-1}\|_\infty \le n\,\|A^{-1}\|_\infty.$$

Beweis: Da die Spaltenpivotsuche verwandt wird, ist $|l_{ij}| \le 1$. Ferner ist $U^{-1} = A^{-1}P^TL$ und daher

$$\|U^{-1}\|_\infty \le \|A^{-1}\|_\infty \underbrace{\|P^T\|_\infty}_{=1} \underbrace{\|L\|_\infty}_{\le n} \le n\,\|A^{-1}\|_\infty.$$

Ferner ist

$$\|A^{-1}\|_\infty = \|L^{-1}PU^{-1}\|_\infty \le \|L^{-1}\|_\infty \underbrace{\|P\|_\infty}_{=1} \|U^{-1}\|_\infty = \|L^{-1}\|_\infty \|U^{-1}\|_\infty.$$

Wir zeigen nun noch $\|L^{-1}\|_\infty \le 2^{n-1}$. Sei hierzu $x \in \mathbb{R}^n$ beliebig vorgegeben und $y := L^{-1}x$. Durch vollständige Induktion nach k zeigen wir, dass $|y_k| \le 2^{k-1}\,\|x\|_\infty$, $k = 1, \ldots, n$. Dies ist für $k = 1$ wegen $y_1 = x_1$ richtig. Angenommen, es ist $|y_i| \le 2^{i-1}\,\|x\|_\infty$, $i = 1, \ldots, k-1$. Wegen

$$l_{k1}y_1 + \cdots + l_{k,k-1}y_{k-1} + y_k = x_k$$

ist

$$|y_k| \le (1 + 2^0 + 2^1 + \cdots + 2^{k-2})\|x\|_\infty = 2^{k-1}\|x\|_\infty,$$

die Behauptung also auch für k richtig. Folglich ist

$$\|y\|_\infty = \|L^{-1}x\|_\infty \le 2^{n-1}\|x\|_\infty,$$

damit $\|L^{-1}\|_\infty \le 2^{n-1}$ und die Behauptung der Aufgabe ist schließlich bewiesen. □

5. Gegeben sei die Matrix

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 3 & 6 & 10 & 15 & 21 \\ 1 & 4 & 10 & 20 & 35 & 56 \\ 1 & 5 & 15 & 35 & 70 & 126 \\ 1 & 6 & 21 & 56 & 126 & 252 \end{pmatrix}.$$

   (a) Welche Matrix würden Sie erhalten, wenn Sie die 6×6-Matrix A um eine Zeile und eine Spalte zu einer 7×7-Matrix erweitern müssten?

   (b) Mit Hilfe des Gaußschen Eliminationsverfahrens mit Spaltenpivotsuche berechne man eine LU-Zerlegung von A. Anschließend berechne man eine untere Schranke von $\kappa_1(A)$.

   Ergebnis: Man erhält

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 3 & 6 & 10 & 15 & 21 & 28 \\ 1 & 4 & 10 & 20 & 35 & 56 & 84 \\ 1 & 5 & 15 & 35 & 70 & 126 & 210 \\ 1 & 6 & 21 & 56 & 126 & 252 & 462 \\ 1 & 7 & 28 & 84 & 210 & 462 & 924 \end{pmatrix},$$

   wenn man das Pascalsche Dreieck dreht. Die angegebene Matrix erhält man in MATLAB durch die Anweisung `A=pascal(6)` bzw. `pascal(7)`.

   Wir geben die mit MATLAB erhaltenen Ergebnisse an. Der Befehl `[L,U,P]=lu(A)` liefert (`format short`, d. h. vier Stellen hinter dem Komma)

$$L = \begin{pmatrix} 1.0000 & & & & & \\ 1.0000 & 1.0000 & & & & \\ 1.0000 & 0.6000 & 1.0000 & & & \\ 1.0000 & 0.2000 & 0.6667 & 1.0000 & & \\ 1.0000 & 0.8000 & 0.6667 & -0.5000 & 1.0000 & \\ 1.0000 & 0.4000 & 1.0000 & 0.7500 & -0.5000 & 1.0000 \end{pmatrix}$$

   sowie

$$U = \begin{pmatrix} 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 \\ & 5.0000 & 20.0000 & 55.0000 & 125.0000 & 251.0000 \\ & & -3.0000 & -14.0000 & -41.0000 & -95.6000 \\ & & & 1.3333 & 6.3333 & 18.5333 \\ & & & & -0.5000 & -2.8000 \\ & & & & & -0.1000 \end{pmatrix}$$

   und

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

Der Befehl `condest(A)` liefert $2.0513 \cdot 10^5$ als Schätzung für $\kappa_1(A)$. Mit Hilfe von `cond(A,1)` kann man sich $\kappa_1(A)$ genau ausrechnen und erhält (jedenfalls mit `format short`) dasselbe Ergebnis.

Ein eigenes Programm zur Berechnung einer LU-Zerlegung liefert merkwürdigerweise etwas andere Ergebnisse. Schreibt man L (ohne die Diagonalelemente) und U in eine Matrix, so erhalten wir

$$(L,U) = \begin{pmatrix} 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 \\ 1.0000 & 5.0000 & 20.0000 & 55.0000 & 125.0000 & 251.0000 \\ 1.0000 & 0.4000 & -3.0000 & -13.0000 & -36.0000 & -80.4000 \\ 1.0000 & 0.8000 & 0.6667 & -1.3333 & -7.0000 & -22.2000 \\ 1.0000 & 0.2000 & 0.6667 & -0.5000 & -0.5000 & -2.7000 \\ 1.0000 & 0.6000 & 1.0000 & 0.7500 & -0.5000 & 0.1000 \end{pmatrix}$$

und die Permutationsmatrix

$$P = (\ e_1 \quad e_6 \quad e_3 \quad e_5 \quad e_2 \quad e_4\ ).$$

Beim Hagerschen Konditionsschätzer bricht das Verfahren nach dem zweiten Schritt ab und berechnet (abgesehen von geringfügigen Rundungsfehlern) die richtige Kondition $\kappa_1(A) = 205\,128$.

6. Sei $A \in \mathbb{R}^{n\times n}$ und $A(\sigma) := A - \sigma I$. Für wie viele Werte von $\sigma \in \mathbb{R}$ höchstens ist das Gaußsche Eliminationsverfahren ohne Spaltenpivotsuche nicht durchführbar bzw. existiert keine LU-Zerlegung von A?

   Ergebnis: Es existiert eine LU-Zerlegung von $A(\sigma)$, wenn die ersten $n-1$ Hauptabschnittsdeterminanten von $A(\sigma)$ ungleich Null sind. Also darf σ kein (reeller) Eigenwert der entsprechenden Haupabschnittsmatrizen sein. Man erhält damit höchstens $1 + 2 + \cdots + (n-1) = n(n-1)/2$ Ausnahmewerte. □

7. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Man zeige:

   (a) Alle Diagonalelemente von A sind positiv.

   (b) Es ist $|a_{ij}| \le \sqrt{a_{ii}a_{jj}}$ für alle $i \ne j$.

   (c) Ein betragsgrößter Eintrag von A liegt auf der Diagonalen.

   Beweis: Der erste Teil (a) ist trivial, denn $0 < e_i^T A e_i = a_{ii}$, $i = 1, \ldots, n$.

   Für (b) seien $i \ne j$ vorgeben. Dann ist

   $$0 < (\alpha e_i - e_j)^T A(\alpha e_i - e_j) = \alpha^2 a_{ii} - 2\alpha a_{ij} + a_{jj} \qquad \text{für alle } \alpha \in \mathbb{R}.$$

   Hieraus folgt (b).

   Wegen (b) ist

   $$|a_{ij}| \le \sqrt{a_{ii}a_{jj}} \le \tfrac{1}{2}(a_{ii} + a_{jj}) \le \max(a_{ii}, a_{jj}),$$

   womit auch (c) bewiesen ist. □

8. Ist $A = (\ a_1 \quad \cdots \quad a_n\ ) \in \mathbb{R}^{n\times n}$, so gilt die Hadamardsche Determinanten-Ungleichung:

   $$|\det(A)| \le \prod_{i=1}^{n} \|a_i\|_2.$$

Hinweis: O. B. d. A. kann man annehmen, dass A nichtsingulär und daher A^TA (symmetrisch und) positivdefinit ist. Man mache eine Cholesky-Zerlegung von A^TA.

Beweis: Sei $A^TA = LL^T$ mit einer unteren Dreiecksmatrix L, die positive Diagonalelemente besitzt. Dann ist

$$\|a_i\|_2^2 = (A^TA)_{ii} = \sum_{j=1}^{i-1} l_{ij}^2 + l_{ii}^2 \geq l_{ii}^2, \qquad i = 1, \ldots, n.$$

Folglich ist

$$|\det(A)| = \sqrt{\det(A^TA)} = \sqrt{\det(LL^T)} = \det(L) = \prod_{i=1}^{n} l_{ii} \leq \prod_{i=1}^{n} \|a_i\|_2,$$

womit die Behauptung bewiesen ist. □

9. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Dann ist

$$\|A\|_{\infty,1} := \max_{x\neq 0} \frac{\|Ax\|_1}{\|x\|_\infty} = \max\{x^TAx : \|x\|_\infty = 1\}.$$

Hinweis: Man mache eine Cholesky-Zerlegung von A.

Beweis: Sei $A = LL^T$ eine Cholesky-Zerlegung von A. Es ist

$$\begin{aligned}
\|A\|_{\infty,1} &= \max\{\|Ax\|_1 : \|x\|_\infty = 1\} \\
&= \max\left\{\sum_{i=1}^{n} |(LL^Tx)_i| : \|x\|_\infty = 1\right\} \\
&= \max\{y^TLL^Tx : \|x\|_\infty = 1,\ \|y\|_\infty = 1\} \\
&= \max\{(L^Ty)^T(L^Tx) : \|x\|_\infty = 1,\ \|y\|_\infty = 1\} \\
&= \max\{(L^Tx)^T(L^Tx) : \|x\|_\infty = 1\} \\
&= \max\{x^TAx : \|x\|_\infty = 1\}.
\end{aligned}$$

Es fehlt eine Begründung für die vorletzte Gleichung, wobei natürlich die Ungleichung

$$\max\{(L^Ty)^T(L^Tx) : \|x\|_\infty = 1,\ \|y\|_\infty = 1\} \geq \max\{(L^Tx)^T(L^Tx) : \|x\|_\infty = 1\}$$

trivialerweise richtig ist. Angenommen, es existieren $x^*, y^* \in \mathbb{R}^n$ mit $\|x^*\|_\infty = \|y^*\|_\infty = 1$ und

$$(L^Ty^*)^T(L^Tx^*) > \max\{(L^Tx)^T(L^Tx) : \|x\|_\infty = 1\}.$$

O. B. d. A. ist $\|L^Ty^*\|_2 \leq \|L^Tx^*\|_2$ (andernfalls vertausche man die Rollen von x^* und y^*). Dann ist wegen der Cauchy-Schwarzschen Ungleichung

$$\|L^Tx^*\|_2^2 \geq (L^Ty^*)^T(L^Tx^*) > \max\{(L^Tx)^T(L^Tx) : \|x\|_\infty = 1\},$$

ein Widerspruch. Daher ist die Behauptung bewiesen. □

10. Man berechne die Cholesky-Zerlegung von

$$A := \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 3 & 6 & 10 & 15 & 21 \\ 1 & 4 & 10 & 20 & 35 & 56 \\ 1 & 5 & 15 & 35 & 70 & 126 \\ 1 & 6 & 21 & 56 & 126 & 252 \end{pmatrix}.$$

Hat man die untere Dreiecksmatrix L mit positiven Diagonalelementen und $A = LL^T$ berechnet, so berechne man anschließend die Cholesky-Zerlegung von $L + L^T - I$.

Ergebnis: Wir erhalten mit der spalten- oder zeilenweisen Version des Cholesky-Verfahrens, dass $A = LL^T$ mit

$$L = \begin{pmatrix} 1 & & & & & \\ 1 & 1 & & & & \\ 1 & 2 & 1 & & & \\ 1 & 3 & 3 & 1 & & \\ 1 & 4 & 6 & 4 & 1 & \\ 1 & 5 & 10 & 10 & 5 & 1 \end{pmatrix}.$$

Dieselben Ergebnisse erhält man, wenn man in MATLAB die Anweisung `L=chol(A)'` macht. Der Cholesky-Faktor L ist also selbst wieder eine Pascal-Matrix! Ferner ist L selbst wieder Cholesky-Faktor von $L + L^T - I$. □

11. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, ferner $A = LL^T$ die Cholesky-Zerlegung von A. Man zeige, dass $\|L\|_2 = \|A\|_2^{1/2}$.

Beweis: Es ist

$$\|L\|_2 = \|L^T\|_2 = \rho(LL^T)^{1/2} = \rho(A)^{1/2} = \|A\|_2^{1/2},$$

wobei wir bei der letzten Gleichung ausgenutzt haben, dass $\rho(A) = \|A\|_2$ für symmetrisches A. □

12. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit, ferner $r := \text{Rang}\,(A)$. Man zeige, dass es eine Permutationsmatrix P gibt derart, dass

$$PAP^T = LL^T, \qquad L = \begin{pmatrix} L_{11} & 0 \\ L_{21} & 0 \end{pmatrix},$$

wobei $L_{11} \in \mathbb{R}^{r\times r}$ eine obere Dreiecksmatrix mit positiven Diagonalelementen ist.

Beweis: Wir geben einen konstruktiven Beweis an, welcher ausnutzt, dass eine symmetrische, positiv semidefinite Matrix ihre betragsmäßig größten Einträge (auch) auf der Diagonalen annimmt (siehe Aufgabe 7, dort ist dies allerdings nur für positiv definite Matrizen formuliert). Im folgenden bezeichnen wir für $1 \le k \le s \le n$ mit $P_{k,s}$ eine Vertauschungsmatrix, eine Multiplikation von links bedeutet also eine Vertauschung der k-ten mit der s-ten Zeile, eine Multiplikation von rechts mit $P_{k,s}^T = P_{k,s}$ eine Vertauschung der k-ten und s-ten Spalte.

- Input: Gegeben sei die symmetrische, positiv semidefinite Matrix $A \in \mathbb{R}^{n\times n}$.
- Setze $A^{(1)} := A$, $r := n$.

- Für $k = 1, \ldots, n$:

  Bestimme $s \in \{k, \ldots, n\}$ aus

  $$s := \min\left\{ j \in \{k, \ldots, n\} : a_{jj}^{(k)} = \max_{i=k,\ldots,n} a_{ii}^{(k)} \right\}.$$

  Falls $a_{ss}^{(k)} = 0$, dann: $r := k - 1$, STOP.
  Sei $P_k := P_{k,s}$, $A^{(k)} := P_k A^{(k)} P_k^T$ und $l_i := P_k l_i$, $i = 1, \ldots, k-1$.
  Berechne $l_k = (0, \ldots, 0, l_{kk}, \ldots, l_{nk})^T$ durch

  $$l_{kk} := (a_{kk}^{(k)})^{1/2}, \qquad l_{ik} := a_{ik}^{(k)} / l_{kk} \quad (i = k+1, \ldots, n).$$

- Output: Mit $P := P_{r-1} \cdots P_1$ ist

  $$PAP^T = \sum_{k=1}^{r} l_k l_k^T.$$

  Definiert man $L_{11} \in \mathbb{R}^{r \times r}$, $L_{21} \in \mathbb{R}^{(n-r) \times r}$ und $L \in \mathbb{R}^{n \times n}$ durch

  $$L_{11} := (l_{ij})_{1 \le i,j \le r}, \qquad L_{21} := (l_{ij})_{\substack{r+1 \le i \le n \\ 1 \le j \le n}} \qquad L := \begin{pmatrix} L_{11} & 0 \\ L_{21} & 0 \end{pmatrix},$$

  so ist L_{11} eine untere Dreiecksmatrix mit positiven Diagonalelementen, $PAP^T = LL^T$ und es ist $r = \text{Rang}\,(A)$.

In einer Implementation merkt man sich in einem Vektor die Vertauschungsmatrizen, ferner kann die Matrix A natürlich sofort von den Vektoren $l_1, \ldots, l_r$ in den entsprechenden Spalten überspeichert werden. □

13. Gegeben sei die symmetrische (indefinite) Matrix

$$A := \begin{pmatrix} 2 & 2 & 3 & 0 & 1 & 2 \\ 2 & 4 & 5 & -1 & 0 & 3 \\ 3 & 5 & 6 & -2 & -3 & 0 \\ 0 & -1 & -2 & 1 & 2 & 3 \\ 1 & 0 & -3 & 2 & 4 & 5 \\ 2 & 3 & 0 & 3 & 5 & 6 \end{pmatrix}.$$

Man berechne mit oder ohne Pivotsuche eine LDL^T-Zerlegung von A.

Ergebnis: Ohne Vertauschungen erhalten wir $A = LDL^T$, wobei wir die Diagonalelemente von D in die (ansonsten mit Einsen besetzte) Diagonale von L schreiben:

$$(L, D) = \begin{pmatrix} 2.0000 & & & & & \\ 1.0000 & 2.0000 & & & & \\ 1.5000 & 1.0000 & -0.5000 & & & \\ 0.0000 & -0.5000 & 2.0000 & 2.5000 & & \\ 0.5000 & -0.5000 & 7.0000 & 3.4000 & -1.4000 & \\ 1.0000 & 0.5000 & 8.0000 & 4.6000 & 4.7143 & 13.7143 \end{pmatrix}.$$

14. Die symmetrische, positiv definite Matrix $A \in \mathbb{R}^{n\times n}$ besitze die Bandbreite p. In Pseudocode gebe man ein Verfahren zur Berechnung der Cholesky-Zerlegung von A an.

Ergebnis: Wir machen natürlich den Ansatz $A = LL^T$, wobei $l_{ij} = 0$ für $i > j+p$. Sei $i \geq j$. Aus

$$a_{ij} = (A)_{ij} = (LL^T)_{ij} = \sum_{k=1}^{j} l_{ik}l_{jk} = l_{ij}l_{jj} + \sum_{k=\max(1,i-p)}^{j-1} l_{ik}l_{jk}$$

erhält man bei *zeilenweiser* Berechnung von L und Überschreibung von Einträgen der Matrix A durch entsprechende Einträge von L das folgende Verfahren:

- $a_{11} := a_{11}^{1/2}$.
- Für $i = 2, \ldots, n$:
  Für $j = \max(1, i-p), \ldots, i-1$:
  $a_{ij} := (a_{ij} - \sum_{k=\max(1,i-p)}^{j-1} a_{ik}a_{jk})/a_{jj}$.
  $a_{ii} := (a_{ii} - \sum_{k=\max(1,i-p)}^{i-1} a_{ik}^2)^{1/2}$.

□

15. Sei $A \in \mathbb{R}^{n\times n}$ eine Tridiagonalmatrix und (zeilenweise) diagonal dominant. Sei $A = LU$ die (eindeutig existierende) LU-Zerlegung von A. Man gebe in Pseudocode ein Verfahren zur Berechnung von L und U an und zeige, dass $|L|\,|U| \leq 3\,|A|$.

Beweis: Nur die untere bzw. obere Nebendiagonale von L bzw. U ist besetzt (siehe Satz 3.4). Wegen $A = LU$ ist

$$a_{i,i+1} = u_{i,i+1}, \quad a_{ii} = l_{i,i-1}u_{i-1,i} + u_{ii} = l_{i,i-1}a_{i-1,i} + u_{ii}, \quad a_{i,i-1} = l_{i,i-1}u_{i-1,i-1}.$$

Die noch nicht bekannten $u_{11}, \ldots, u_{nn}$ und $l_{21}, \ldots, l_{n,n-1}$ erhält man durch das folgende Verfahren:

- $u_{11} := a_{11}$.
- Für $i = 2, \ldots, n$:
  $l_{i,i-1} := a_{i,i-1}/u_{i-1,i-1}$, $u_{ii} := a_{ii} - l_{i,i-1}a_{i-1,i}$.

Natürlich wird man eine $n \times n$-Tridiagonalmatrix nicht in einem $n \times n$-Feld speichern. Die Modifikationen sind offensichtlich.

Es ist

$$(|L|\,|U|)_{i,i+1} = |(LU)_{i,i+1}| = |a_{i,i+1}|$$

und

$$(|L|\,|U|)_{i,i-1} = |(LU)_{i,i-1}| = |a_{i,i-1}|.$$

Daher ist im zweiten Teil der Aufgabe nur $(|L|\,|U|)_{ii} \leq 3\,|a_{ii}|$, $i = 1, \ldots, n$, bzw. (da dies für $i = 1$ trivial ist)

$$|l_{i,i-1}|\,|a_{i-1,i}| + |u_{ii}| \leq 3\,|a_{ii}|, \qquad i = 2, \ldots, n,$$

nachzuweisen. Hierzu zeigen wir durch vollständige Induktion nach i, dass

$$|a_{i,i+1}| \leq |u_{ii}| \quad (i = 1, \ldots, n-1), \qquad |u_{ii}| \leq |a_{ii}| + |a_{i,i-1}| \quad (i = 2, \ldots, n).$$

Wir zeigen den ersten Satz von Ungleichungen, der Beweis des zweiten verläuft analog. Für $i = 1$ ist $|u_{11}| = |a_{11}| > |a_{12}|$, die entsprechende Ungleichung also richtig. Sie sei für $i-1$ richtig. Dann ist

$$|u_{ii}| \geq |a_{ii}| - |l_{i,i-1}|\,|a_{i-1,i}| = |a_{ii}| - \frac{|a_{i,i-1}|}{|u_{i-1,i-1}|}|a_{i-1,i}| \geq |a_{ii}| - |a_{i,i-1}| > |a_{i,i+1}|,$$

so dass die Aussage auch für i richtig ist. Dann ist schließlich

$$\begin{aligned} |l_{i,i-1}|\,|a_{i-1,i}| + |u_{ii}| &= \frac{|a_{i,i-1}|}{|u_{i-1,i-1}|}|a_{i-1,i}| + |u_{ii}| \\ &\leq |a_{i,i-1}| + |u_{ii}| \\ &\leq |a_{i,i-1}| + (|a_{ii}| + |a_{i,i-1}|) \\ &\leq 3\,|a_{ii}|, \end{aligned}$$

die Behauptung ist bewiesen. □

6.2 Aufgaben in Kapitel 3

6.2.1 Aufgaben in Abschnitt 3.1

1. Man beweise Satz 1.1, also: Gegeben sei das lineare Ausgleichsproblem

   (P) $$\text{Minimiere} \quad \|Ax - b\|_2, \quad x \in \mathbb{R}^n.$$

   Hierbei sei $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ und $m \geq n$. Dann gilt:

   (a) Das lineare Ausgleichsproblem (P) besitzt eine Lösung.

   (b) Ein $x^* \in \mathbb{R}^n$ ist genau dann eine Lösung von (P), wenn x^* eine Lösung der sogenannten Normalgleichungen $A^TAx = A^Tb$ ist.

   (c) Eine Lösung von (P) ist genau dann eindeutig, wenn Rang $(A) = n$.

   (d) Unter allen Lösungen von (P) gibt es genau eine mit minimaler euklidischer Norm.

   Beweis: Eine Niveaumenge bezüglich der Aufgabe, $\|y - b\|_2$ unter der Nebenbedingung $y \in$ Bild (A) zu minimieren, ist kompakt, woraus die Existenz einer Lösung folgt. Definiert man f durch $f(x) := \frac{1}{2}\|Ax - b\|_2^2$, so ist f konvex, und daher $x^* \in \mathbb{R}^n$ genau dann eine Lösung von (P), wenn x^* eine Lösung des linearen Gleichungssystems $\nabla f(x) = A^T(Ax - b) = 0$ ist. Ist Rang $(A) = n$, so ist A^TA nichtsingulär und daher die Normalgleichungen eindeutig lösbar. Ist umgekehrt Rang $(A) < n$, so ist A^TA singulär und die Normalgleichungen nicht eindeutig lösbar. In jedem Fall ist die Menge der Lösungen von (P) ein nichtleerer, affiner Teilraum im $\mathbb{R}^n$. In diesem Teilraum gibt es genau ein Element mit minimaler euklidischer Norm. □

2. Sei $A \in \mathbb{R}^{m\times n}$ mit Rang $(A) = n$, $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n$ gegeben. Dann besitzen die beiden Aufgaben

   $$\text{Minimiere} \quad f(x) := \tfrac{1}{2}\|Ax - b\|_2^2 + c^Tx, \quad x \in \mathbb{R}^n$$

   und

   $$\text{Minimiere} \quad g(y) := \tfrac{1}{2}\|y - b\|_2^2, \quad A^Ty = c$$

jeweils genau eine Lösung, die man durch Lösen von

$$\begin{pmatrix} I & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix}$$

erhalten kann.

Beweis: Eine Lösung der ersten Aufgabe ist als eine Lösung von

$$\nabla f(x) = A^T(Ax - b) + c = 0$$

charakterisiert, wegen Rang $(A) = n$ ist dieses lineare Gleichungssystem eindeutig durch ein x^* lösbar. Definiert man $y^* := b - Ax^*$, so ist $A^T y^* = c$, d. h. y^* genügt der Nebenbedingung der zweiten Aufgabe. Für ein beliebiges y mit $A^T y = c$ ist dann

$$\begin{aligned} g(y) - g(y^*) &= \tfrac{1}{2}\|y - b\|_2^2 - \tfrac{1}{2}\|y^* - b\|_2^2 \\ &= (y^* - b)^T(y - y^*) + \tfrac{1}{2}\|y - y^*\|_2^2 \\ &= -(x^*)^T \underbrace{A^T(y - y^*)}_{=0} + \tfrac{1}{2}\|y - y^*\|_2^2 \\ &= \tfrac{1}{2}\|y - y^*\|_2^2 \\ &\geq 0, \end{aligned}$$

woraus man abliest, dass y^* die eindeutige Lösung der zweiten Aufgabe ist. Offenbar ist (y^*, x^*) Lösung des angegebenen (eindeutig lösbaren) linearen Gleichungssystems. □

3. Sei $C \in \mathbb{R}^{m\times m}$ symmetrisch und positiv definit, ferner sei auf dem $\mathbb{R}^m$ durch $\|y\|_C := (y^T C y)^{1/2}$ eine Norm definiert. Mit $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$ betrachte man das (gewichtete) lineare Ausgleichsproblem

(P) $$\text{Minimiere} \quad \|Ax - b\|_C, \quad x \in \mathbb{R}^n.$$

Man formuliere die Normalgleichungen zu dieser Aufgabe, gebe also notwendige und hinreichende Optimalitätsbedingungen für (P) an.

Lösung: Am einfachsten ist es wohl, wenn man $\|y\|_C = \|C^{1/2}y\|_2$ berücksichtigt, wobei $C^{1/2}$ die symmetrische, positiv definite Quadratwurzel aus C ist. Daher ist (P) äquivalent zu

$$\text{Minimiere} \quad \|C^{1/2}Ax - C^{1/2}b\|_2, \quad x \in \mathbb{R}^n.$$

Daher ist x genau dann eine Lösung von (P), wenn

$$(C^{1/2}A)^T(C^{1/2}Ax - C^{1/2}b) = 0$$

bzw. $A^T C A x = A^T C b$. □

4. Gegeben sei das lineare Ausgleichsproblem

(P) $$\text{Minimiere} \quad \|Ax - b\|_2, \quad x \in \mathbb{R}^n,$$

wobei $A \in \mathbb{R}^{m\times n}$ mit Rang $(A) = n$ und $b \in \mathbb{R}^m$ mit $b \notin \text{Bild}\,(A)$ gegeben sind. Man bilde die Matrix

$$\bar{A} := (\ A \quad b\ ) \in \mathbb{R}^{m\times(n+1)}$$

und zeige:

(a) Es ist Rang $(\bar{A}) = n+1$. Daher besitzt $\bar{A}^T\bar{A}$ eine Cholesky-Zerlegung $\bar{A}^T\bar{A} = \bar{L}\bar{L}^T$.

(b) Ist

$$\bar{L} = \begin{pmatrix} L & 0 \\ l^T & \lambda \end{pmatrix},$$

erhält man ferner x durch Rückwärtseinsetzen aus $L^T x = l$, so ist x die Lösung von (P) und $\|Ax - b\|_2 = \lambda$.

Beweis: Der erste Teil ist offensichtlich, da wir Rang $(A) = n$ und $b \notin$ Bild (A) vorausgesetzt haben. Es ist

$$\bar{A}^T\bar{A} = \begin{pmatrix} A^TA & A^Tb \\ (A^Tb)^T & b^Tb \end{pmatrix} = \begin{pmatrix} LL^T & Ll \\ (Ll)^T & l^Tl + \lambda^2 \end{pmatrix} = \bar{L}\bar{L}^T.$$

Mit $L^Tx = l$ ist

$$A^TAx = LL^Tx = Ll = A^Tb,$$

daher genügt x den Normalgleichungen und ist eine Lösung von (P). Weiter ist

$$\|Ax - b\|_2^2 = b^T(b - Ax) = l^Tl + \lambda^2 - (A^Tb)^Tx = l^Tl + \lambda^2 - (Ll)^Tx = \lambda^2,$$

wobei wir für die erste Gleichung die Normalgleichungen ausgenutzt haben. Damit ist auch der zweite Teil der Aufgabe bewiesen. □

6.2.2 Aufgaben in Abschnitt 3.2

1. Man zeige, dass die in Pseudocode angegebenen Verfahren CGS und MGS bei exakter Arithmetik (also ohne auftretende Rundungsfehler) äquivalent sind, also denselben Output liefern.

   - Input: Gegeben $A = (\ a_1 \ \cdots \ a_n\ ) \in \mathbb{R}^{m\times n}$ mit $m \geq n$.
   - Für $k = 1, \ldots, n$:
     Setze $q_k' := a_k$.
     Für $i = 1, \ldots, k-1$:
     Berechne $r_{ik} := q_i^Ta_k$ (CGS) bzw. $r_{ik} := q_i^Tq_k'$ (MGS).
     Berechne $q_k' := q_k' - r_{ik}q_i$.
     Berechne $r_{kk} := \|q_k'\|_2$.
     Falls $r_{kk} = 0$, dann: STOP, da $a_k \in \text{span}\,\{a_1, \ldots, a_{k-1}\}$.
     Andernfalls: Berechne $q_k := q_k'/r_{kk}$.
   - Output: Wenn das (genauer: die) Verfahren wegen Rang $(A) < n$ nicht vorzeitig abbricht (abbrechen), werden eine Matrix $\hat{Q} = (\ q_1 \ \cdots \ q_n\ ) \in \mathbb{R}^{m\times n}$ mit $\hat{Q}^T\hat{Q} = I$ und eine obere Dreiecksmatrix $\hat{R}$ mit r_{ik} für $i \leq k$ als (i,k)-Eintrag erzeugt, die eine reduzierte QR-Zerlegung von A bilden, für die also $A = \hat{Q}\hat{R}$.

Beweis: Wir denken uns auf die durch das MGS-Verfahren berechneten Größen jeweils einen Hut gesetzt und beweisen durch vollständige Induktion nach k, dass

(a) Es ist $r_{ik} = \hat{r}_{ik}$, $i = 1, \ldots, k$,

(b) Es ist $q_i = \hat{q}_i$, $i = 1, \ldots, k$,

(c) Es ist $\{q_1, \ldots, q_k\}$ ein Orthonormalsystem, $\operatorname{span}\{q_1, \ldots, q_k\} = \operatorname{span}\{a_1, \ldots, a_k\}$ und $a_k = \sum_{i=1}^{k} r_{ik} q_i$.

Es ist $r_{11} = \hat{r}_{11} = \|a_1\|_2$ und $q_1 = \hat{q}_1 = a_1/\|a_1\|_2$, also (a) und (b) für $k = 1$ richtig. Daher ist auch (c) für $k = 1$ richtig. Wir nehmen an, die Behauptung sei für $k-1$ richtig. Für $i = 1, \ldots, k-1$ ist

$$r_{ik} = q_i^T a_k = \hat{q}_i^T \left(a_k - \sum_{j=1}^{i-1} \hat{r}_{jk} \hat{q}_j \right) = \hat{r}_{ik},$$

wobei wir (b) für $k-1$ und die Annahme, dass $\{q_1, \ldots, q_{k-1}\}$ ein Orthonormalsystem ist, ausgenutzt haben. Weiter ist

$$\hat{q}'_k = a_k - \sum_{i=1}^{k-1} \hat{r}_{ik} \hat{q}_i = a_k - \sum_{i=1}^{k-1} r_{ik} q_i = q'_k$$

und folglich $\hat{r}_{kk} = r_{kk}$ und $\hat{q}_k = q_k$, falls $r_{kk} \neq 0$. Andernfalls ist, wie behauptet,

$$a_k \in \operatorname{span}\{q_1, \ldots, q_{k-1}\} = \operatorname{span}\{a_1, \ldots, a_{k-1}\}.$$

Damit sind (a) und (b) für k bewiesen. Für $i = 1, \ldots, k-1$ ist weiter

$$q_i^T q_k = \frac{1}{\|q'_k\|_2} q_i^T \left(a_k - \sum_{j=1}^{k-1} r_{jk} q_j \right) = \frac{1}{\|q'_k\|_2} (q_i^T a_k - r_{ik}) = 0,$$

ferner ist natürlich $\|q_k\|_2 = 1$, also $\{q_1, \ldots, q_k\}$ ein Orthonormalsystem. Der Rest des Induktionsbeweises für (c) ist offensichtlich. □

2. Man programmiere das klassische und das modifizierte Gram-Schmidt-Verfahren und teste es daran, eine reduzierte QR-Zerlegung $A = \hat{Q}\hat{R}$ der 7×5-Hilbertmatrix $A := (1/(i+j-1)) \in \mathbb{R}^{7\times 5}$ zu berechnen. Insbesondere teste man die "Orthogonalität" von $\hat{Q}$ durch Berechnung von $\hat{Q}^T\hat{Q}$.

Ergebnis: Beim CGS erhalten wir

$$\hat{Q}^T\hat{Q} = \begin{pmatrix} 1.00000000 & -0.00000000 & 0.00000000 & -0.00000000 & 0.00000005 \\ & 1.00000000 & 0.00000000 & -0.00000000 & 0.00000004 \\ & & 1.00000000 & -0.00000010 & 0.00000276 \\ & & & 1.00000000 & 0.00011457 \\ & & & & 1.00000000 \end{pmatrix},$$

dagegen beim MGS

$$\hat{Q}^T\hat{Q} = \begin{pmatrix} 1.00000000 & -0.00000000 & 0.00000000 & -0.00000000 & 0.00000005 \\ & 1.00000000 & 0.00000000 & 0.00000000 & -0.00000001 \\ & & 1.00000000 & 0.00000000 & 0.00000000 \\ & & & 1.00000000 & 0.00000000 \\ & & & & 1.00000000 \end{pmatrix}.$$

Der Unterschied ist deutlich. □

3. Sei
$$A := \begin{pmatrix} 1 & 2 & 3 \\ 1 & 5 & 6 \\ 1 & 8 & 9 \\ 1 & 11 & 12 \end{pmatrix}.$$
Mit dem Householder-Verfahren ohne und mit Spaltenpivotsuche berechne man eine volle QR-Zerlegung von A.

Ergebnis: Wir geben acht Dezimalstellen an und erhalten
$$Q = \begin{pmatrix} -0.50000000 & 0.67082039 & -0.50000000 & 0.22360680 \\ -0.50000000 & 0.22360680 & 0.83333333 & 0.07453560 \\ -0.50000000 & -0.22360680 & -0.16666667 & -0.81989159 \\ -0.50000000 & -0.67082039 & -0.16666667 & 0.52174919 \end{pmatrix}$$
und
$$R = \begin{pmatrix} -2.00000000 & -13.00000000 & -15.00000000 \\ 0.00000000 & -6.70820393 & -6.70820393 \\ 0.00000000 & 0.00000000 & -0.00000000 \end{pmatrix}.$$
Dasselbe Ergebnis erhalten wir, wenn wir in MATLAB die Anweisung `[Q,R]=qr(A)` geben. Machen wir im Householder-Verfahren eine Spaltenpivotsuche, so erhalten wir, dass A den Rang 2 besitzt, außerdem die Zerlegung $A\Pi = QR$ mit
$$\Pi = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \qquad R = \begin{pmatrix} -16.43167673 & -14.60593487 & -1.82574186 \\ 0.00000000 & 0.81649658 & -0.81649658 \\ 0.00000000 & 0.00000000 & 0.00000000 \end{pmatrix}$$
und
$$Q = \begin{pmatrix} -0.18257419 & -0.81649658 & -0.40008743 & -0.37407225 \\ -0.36514837 & -0.40824829 & 0.25463292 & 0.79697056 \\ -0.54772256 & 0.00000000 & 0.69099646 & -0.47172438 \\ -0.73029674 & 0.40824829 & -0.54554195 & 0.04882607 \end{pmatrix}.$$
Gibt man die MATLAB-Anweisung `[Q,R,Pi]=qr(A)`, so erhält man dieselbe Permutationsmatrix Π, dieselbe obere Dreiecksmatrix R und auch die ersten beiden Spalten von Q stimmen überein, die beiden restlichen nicht. Die beiden letzten Spalten von Q sind für die Gültigkeit der Zerlegung irrelevant, und natürlich kann man die ersten beiden Spalten auf verschiedene Weise zu einer orthogonalen Matrix ergänzen. □

4. Gegeben sei das lineare Gleichungssystem $Ax = b$ mit $A \in \mathbb{R}^{m\times n}$ und Rang $(A) = m \leq n$ und $b \in \mathbb{R}^m$, d. h. wir haben ein *unterbestimmtes* lineares Gleichungssystem. Man zeige, wie man mit Hilfe einer QR-Zerlegung von A^T die eindeutige Lösung von $Ax = b$ mit minimaler euklidischer Norm berechnen kann.

Lösung: Sei
$$A^T = Q \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix},$$
wobei $Q \in \mathbb{R}^{n\times n}$ orthogonal und $\hat{R} \in \mathbb{R}^{m\times m}$ eine nichtsinguläre obere Dreiecksmatrix ist. Sei $x \in \mathbb{R}^n$ und $Ax = b$. Dann ist
$$b = Ax = \begin{pmatrix} \hat{R}^T & 0 \end{pmatrix} Q^T x = \hat{R}^T y_1,$$

wobei

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = Q^T x.$$

Die allgemeine Lösung des unterbestimmten linearen Gleichungssystems $Ax = b$ ist also

$$x = Q \begin{pmatrix} \hat{R}^{-T} b \\ y_2 \end{pmatrix}$$

mit beliebigem $y_2 \in \mathbb{R}^{n-m}$. Die Lösung minimaler euklidischer Norm erhält man wegen

$$\|x\|_2^2 = \|\hat{R}^{-T} b\|_2^2 + \|y_2\|_2^2$$

offenbar für $y_2 = 0$, sie ist also durch

$$x^* := Q \begin{pmatrix} \hat{R}^{-T} b \\ 0 \end{pmatrix}$$

gegeben. □

5. Gegeben sei das lineare Ausgleichsproblem

(P) $$\text{Minimiere} \quad \|Ax - b\|_2, \quad x \in \mathbb{R}^n,$$

wobei $A \in \mathbb{R}^{m \times n}$ mit $m \geq n$, $b \in \mathbb{R}^m$. Durch das Householder-Verfahren mit Spaltenpivotisierung sei die Zerlegung

$$A\Pi = QR = Q \begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix}$$

berechnet, wobei $Q \in \mathbb{R}^{m \times m}$ orthogonal, $\Pi \in \mathbb{R}^{n \times n}$ eine Permutationsmatrix, $R_{11} \in \mathbb{R}^{r \times r}$ eine nichtsinguläre obere Dreiecksmatrix und $R_{12} \in \mathbb{R}^{r \times (n-r)}$. Sei

$$Q^T b =: \begin{pmatrix} c \\ d \end{pmatrix} \qquad \text{mit} \quad c \in \mathbb{R}^r,\ d \in \mathbb{R}^{m-r}.$$

Sei

$$x_B := \Pi \begin{pmatrix} R_{11}^{-1} c \\ 0 \end{pmatrix}$$

die sogenannte *Basislösung* von (P). Man zeige:

(a) Die Basislösung x_B ist eine Lösung von (P).

(b) Ist $R_{12} = 0$, so ist x_B die eindeutige Lösung minimaler Norm von (P).

Beweis: Sei $x \in \mathbb{R}^n$ beliebig und

$$\Pi^T x =: \begin{pmatrix} y \\ z \end{pmatrix} \qquad \text{mit} \quad y \in \mathbb{R}^r,\ z \in \mathbb{R}^{n-r}.$$

Dann ist

$$\|Ax - b\|_2^2 = \|(Q^T A\Pi)(\Pi^T x) - Q^T b\|_2^2 = \|R_{11} y - (c - R_{12} z)\|_2^2 + \|d\|_2^2.$$

Für ein beliebiges $z \in \mathbb{R}^{n-r}$ ist daher durch

$$x = \Pi \begin{pmatrix} R_{11}^{-1}(c - R_{12}z) \\ z \end{pmatrix}$$

eine Lösung von (P) gegeben, ferner lässt sich jede Lösung in dieser Weise darstellen. Insbesondere (setze $z := 0$) ist die Basislösung x_B eine Lösung von (P). Ist $R_{12} = 0$, so lässt sich die eindeutige Lösung x_{LS} minimaler euklidischer Norm in der Form

$$x_{LS} = \Pi \begin{pmatrix} R_{11}^{-1}c \\ z_{LS} \end{pmatrix}$$

mit $z_{LS} \in \mathbb{R}^{n-r}$ darstellen. Dann ist aber

$$\|x_{LS}\|_2^2 = \|R_{11}^{-1}c\|_2^2 + \|z_{LS}\|_2^2 \geq \|x_B\|_2^2,$$

woraus die Behauptung folgt. □

6. Was sind die Eigenwerte einer Householder-Matrix, was die einer Givens-Rotation?

   Lösung: Gegeben sei die Householder-Matrix

$$Q = I - \frac{2}{u^T u} uu^T \in \mathbb{R}^{m \times m}$$

mit $u \in \mathbb{R}^m \setminus \{0\}$. Dann ist 1 ein $(m-1)$-facher Eigenwert mit Eigenvektoren aus dem $(m-1)$-dimensionalen linearen Raum $\text{span}\,(u)^\perp$, weiter ist -1 ein Eigenwert mit u als zugehörigem Eigenvektor. Nun sei die Givens-Rotation $G_{jk}(c, s)$ gegeben. Dann ist 1 ein $(m-2)$-facher Eigenwert mit zugehörigem Eigenvektor e_i, $i \neq j, k$. Für die beiden restlichen Eigenvektoren mache man natürlich den Ansatz $x = \alpha e_j + \beta e_k$ und erhält $c \pm is = e^{\pm i\theta}$ als weitere Eigenwerte. □

7. Sei $A \in \mathbb{R}^{m \times n}$ mit $m \geq n$ gegeben. Es sei eine volle QR-Zerlegung

$$A = Q \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix}$$

von A bekannt. Wie kann man dann mit $O(m^2)$ flops eine volle QR-Zerlegung von $\tilde{A} := A + uv^T$ bestimmen, wobei $u \in \mathbb{R}^m$, $v \in \mathbb{R}^n$ gegeben sind?

   Lösung: Es ist

$$\tilde{A} = A + uv^T = Q\left[\begin{pmatrix} \hat{R} \\ 0 \end{pmatrix} + Q^T uv^T\right].$$

Man berechne $w := Q^T u \in \mathbb{R}^m$ (das erfordert schon m^2 flops), anschließend führe man durch sukzessive Multiplikation von w mit Givensrotationen $G_{m-1,m}, \ldots, G_{1,2}$ den Vektor w in ein Vielfaches des ersten Einheitsvektors über:

$$G_{12} \cdots G_{m-1,m} w = \alpha e_1 \qquad \text{mit} \quad \alpha = \pm\|w\|_2.$$

Mit diesen Givens-Rotationen werde auch der obere Dreiecks-Faktor der QR-Zerlegung von A multipliziert (hierzu genügen die letzten n Rotationen, weil die anderen keinen Einfluss haben). Man berechnet also

$$\tilde{H} := G_{12} \cdots G_{m-1,m}\left[\begin{pmatrix} \hat{R} \\ 0 \end{pmatrix} + wv^T\right] = H + \alpha e_1 v^T.$$

Offenbar ist H und dann auch $\tilde{H}$ eine obere Hessenberg-Matrix. Die Subdiagonale von $\tilde{H}$ kann durch weitere n Givens-Rotationen zu Null gemacht werden, es kann also erreicht werden, dass

$$\tilde{G}_{n,n+1} \cdots \tilde{G}_{12} \tilde{H} = \begin{pmatrix} \tilde{R} \\ 0 \end{pmatrix}$$

mit einer oberen Dreiecksmatrix $\tilde{R} \in \mathbb{R}^{n \times n}$. Dann ist $\tilde{A} = \tilde{Q} \begin{pmatrix} \tilde{R} \\ 0 \end{pmatrix}$ mit

$$\tilde{Q} := Q G_{12} \cdots G_{m-1,m} \tilde{G}_{n,n+1} \cdots \tilde{G}_{12}.$$

Es ist leicht einzusehen, dass der benötigte Aufwand durch $O(m^2)$ abgeschätzt werden kann. □

8. Seien $A \in \mathbb{R}^{m \times n}$ mit Rang $(A) = n$, $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n$ gegeben. Hiermit betrachte man die beiden Aufgaben (siehe auch Aufgabe 2 in Abschnitt 3.1)

$$\text{Minimiere} \quad f(x) := \tfrac{1}{2}\|Ax - b\|_2^2 + c^T x, \quad x \in \mathbb{R}^n$$

und

$$\text{Minimiere} \quad g(y) := \tfrac{1}{2}\|y - b\|_2^2, \quad A^T y = c.$$

Mit Hilfe einer QR-Zerlegung von A gebe man eine Methode an, die eindeutig existierenden Lösungen zu berechnen.

Lösung: Sei

$$A = Q \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix}$$

eine QR-Zerlegung von A. Wie bei der Lösung eines linearen Ausgleichsproblems mit Hilfe einer QR-Zerlegung berechne man (wir ändern die Bezeichnungen unwesentlich, da der Vektor c hier schon vergeben ist)

$$\begin{pmatrix} d_1 \\ d_2 \end{pmatrix} := Q^T b \qquad \text{mit} \quad d_1 \in \mathbb{R}^n,\ d_2 \in \mathbb{R}^{m-n}.$$

Mit einem beliebigen $x \in \mathbb{R}^n$ ist dann

$$f(x) = \tfrac{1}{2}\|Ax - b\|_2^2 + c^T x = \tfrac{1}{2}[\|\hat{R}x - d_1\|_2^2 + \|d_2\|_2^2] + c^T x.$$

Die eindeutig existierende Lösung, f auf dem $\mathbb{R}^n$ zu minimieren, erhält man aus

$$0 = \nabla f(x) = \hat{R}^T(\hat{R}x - d_1) + c.$$

Damit erhält man x durch die folgenden beiden Schritte:

- Berechne die Lösung z von $\hat{R}^T z = c$ durch Vorwärtseinsetzen.
- Berechne die Lösung x aus $Rx = d_1 - z$ durch Rückwärtseinsetzen.

Schön ist nun, dass man die Lösung des zweiten Problems fast geschenkt bekommt. Denn sei $y \in \mathbb{R}^m$ mit $A^T y = c$ beliebig und

$$Q^T y = \begin{pmatrix} (Q^T y)_1 \\ (Q^T y)_2 \end{pmatrix}.$$

Wegen

$$c = A^T y = (\ \hat{R}^T \quad 0\ ) Q^T y = \hat{R}^T (Q^T y)_1$$

ist $(Q^T y)_1 = z$, wobei z gerade eben berechnet war. Weiter ist

$$g(y) = \tfrac{1}{2}\|y - b\|_2^2 = \tfrac{1}{2}\|Q^T y - Q^T b\|_2^2 = \tfrac{1}{2}[\|z - d_1\|_2^2 + \|(Q^T y)_2 - d_2\|_2^2].$$

Daher ist die Lösung y des zweiten Problems durch

$$Q^T y = \begin{pmatrix} z \\ d_2 \end{pmatrix}$$

festgelegt, sie ist also durch

$$y := Q \begin{pmatrix} z \\ d_2 \end{pmatrix}$$

gegeben. □

6.2.3 Aufgaben in Abschnitt 3.3

1. Sei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ und $r := \text{Rang}\,(A)$ gegeben, ferner sei mit den üblichen Bezeichnungen $A = U\Sigma V^T$ eine Singulärwertzerlegung von A. Mit $\mathbb{R}_k^{m\times n}$ werde die Menge der reellen $m \times n$-Matrizen vom Rang k bezeichnet. Für jedes $k < r$ ist dann

$$\sigma_{k+1} = \|A - A_k\|_2 = \min_{X \in \mathbb{R}_k^{m\times k}} \|A - X\|_2,$$

wobei

$$A_k := \sum_{i=1}^{k} \sigma_i u_i v_i^T.$$

Beweis: Sei $k < r$. Wegen $U^T A_k V = \text{diag}\,(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$ ist $A_k \in \mathbb{R}_k^{m\times n}$ und

$$\|A - A_k\|_2 = \|U^T(A - A_k)V\|_2 = \|\text{diag}\,(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_r)\|_2 = \sigma_{k+1}.$$

Nun sei $X \in \mathbb{R}_k^{m\times n}$ beliebig. Sei $\{w_1, \ldots, w_{n-k}\}$ eine Orthonormalbasis von $\text{Kern}\,(X)$. Dann ist

$$\text{span}\,\{w_1, \ldots, w_{n-k}\} \cap \text{span}\,\{v_1, \ldots, v_{k+1}\} \neq \{0\}.$$

Man wähle ein z aus diesem Durchschnitt mit $\|z\|_2 = 1$. Wegen $Xz = 0$ und $Az = \sum_{i=1}^{k+1} \sigma_i (v_i^T z) u_i$ ist

$$\|A - X\|_2^2 \geq \|(A - X)z\|_2^2 = \|Az\|_2^2 = \sum_{i=1}^{k+1} \sigma_i^2 (v_i^T z)^2 \geq \sigma_{k+1}^2.$$

Damit ist die Behauptung bewiesen. □

2. Sei $A \in \mathbb{R}^{m\times n}$ und A^+ die zugehörige Pseudoinverse. Man zeige:

   (a) Es ist

$$AA^+A = A, \quad A^+AA^+ = A^+, \quad (AA^+)^T = AA^+, \quad (A^+A)^T = A^+A.$$

(b) Sei $L \subset \mathbb{R}^n$ ein linearer Teilraum. Eine lineare Abbildung $P_L\colon \mathbb{R}^n \longrightarrow L \subset \mathbb{R}^n$ heißt *orthogonale Projektion* des $\mathbb{R}^n$ auf L, wenn $(I - P_L)x \perp L$ für alle $x \in \mathbb{R}^n$. Dann ist AA^+ orthogonale Projektion des $\mathbb{R}^m$ auf Bild (A) und A^+A orthogonale Projektion des $\mathbb{R}^n$ auf Bild (A^T).

Beweis: Wir nehmen zunächst an, es sei $m \geq n$. Es ist

$$A = U \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} V^T, \qquad A^+ = V \begin{pmatrix} \hat{\Sigma}^+ & 0 \end{pmatrix} U^T$$

mit den üblichen Eigenschaften. Dann ist

$$AA^+A = U \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} \underbrace{V^TV}_{=I} \begin{pmatrix} \hat{\Sigma}^+ & 0 \end{pmatrix} \underbrace{U^TU}_{=I} \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} V^T = U \begin{pmatrix} \hat{\Sigma}\hat{\Sigma}^+\hat{\Sigma} \\ 0 \end{pmatrix} V^T = A,$$

da $\hat{\Sigma}\hat{\Sigma}^+\hat{\Sigma} = \hat{\Sigma}$. Für $m < n$ ist

$$AA^+A = A((A^T)^+)^TA = (A^T(A^T)^+A^T)^T = (A^T)^T = A.$$

Ähnlich beweist man auch die anderen Gleichungen.

Sei $b \in \mathbb{R}^m$ beliebig. Wir haben zu zeigen, dass

$$(I - AA^+)b \in \text{Bild}\,(A)^\perp = \text{Kern}\,(A^T) \qquad \text{bzw.} \qquad A^T(I - AA^+) = 0.$$

Nun ist aber

$$[A^T(I - AA^+)]^T = [I - (AA^+)^T]A = (I - AA^+)A = A - AA^+A = 0,$$

wenn man die dritte und die erste der Beziehungen in (a) benutzt. Daher ist $AA^+ = P_{\text{Bild}\,(A)}$ die orthogonale Projektion des $\mathbb{R}^m$ auf Bild (A). Zum Nachweis, dass $A^+A = P_{\text{Bild}\,(A^T)}$ orthogonale Projektion des $\mathbb{R}^n$ auf Bild (A^T) ist

$$(I - A^+A)x \in \text{Bild}\,(A^T)^\perp = \text{Kern}\,(A)$$

für alle $x \in \mathbb{R}^n$ zu zeigen, was wegen der ersten der obigen Beziehungen klar ist. □

3. Sei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ und $A^+ \in \mathbb{R}^{n\times m}$ die zugehörige Pseudoinverse. Man zeige, dass A^+ Lösung der Aufgabe

$$\text{Minimiere} \quad \|AX - I\|_F, \quad X \in \mathbb{R}^{n\times m},$$

ist.

Beweis: Mit einem beliebigen $X \in \mathbb{R}^{n\times m}$ ist

$$\|AX - I\|_F^2 = \sum_{i=1}^{m} \|AXe_i - e_i\|_2^2 \geq \sum_{i=1}^{m} \|AA^+e_i - e_i\|_2^2 = \|AA^+ - I\|_F^2,$$

das ist schon der Beweis. □

4. Sei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ und $r := \text{Rang}\,(A)$ gegeben. Man definiere $B\colon (0,\infty) \longrightarrow \mathbb{R}^{n\times m}$ durch

$$B(\lambda) := (A^TA + \lambda I)^{-1}A^T, \qquad \lambda > 0.$$

Man zeige, dass

$$\|B(\lambda) - A^+\|_2 = \frac{\lambda}{\sigma_r(\sigma_r^2 + \lambda)},$$

wobei σ_r kleinster singulärer Wert von A ist. Insbesondere gilt $\lim_{\lambda\to 0+} B(\lambda) = A^+$.

Beweis: Sei

$$A = U \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} V^T$$

eine Singulärwertzerlegung von A. Dann ist

$$B(\lambda) - A^+ = V(\ \hat{T}(\lambda) \quad 0\ )U^T$$

mit

$$\hat{T}(\lambda) := (\hat{\Sigma}^2 + \lambda I)^{-1}\hat{\Sigma} - \Sigma^+.$$

Also ist

$$T(\lambda) = \text{diag}\,(\tau_1(\lambda), \ldots, \tau_r(\lambda), 0, \ldots, 0)$$

mit

$$\tau_i(\lambda) = \frac{\sigma_i}{\sigma_i^2 + \lambda} - \frac{1}{\sigma_i} = -\frac{\lambda}{\sigma_i(\sigma_i^2 + \lambda)}, \qquad i = 1, \ldots, r$$

und hieraus folgt die Behauptung. □

5. Seien $A \in \mathbb{R}^{m\times n}$ mit $\text{Rang}\,(A) = m < n$ und $b \in \mathbb{R}^m$ gegeben. Die Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad \|Ax - b\|_2, \quad x \in \mathbb{R}^n,$$

heißt ein *unterbestimmtes lineares Ausgleichsproblem*. Man zeige, dass (P) eine $(n-m)$-dimensionale affin lineare Lösungsmenge hat, dass (P) genau eine Lösung minimaler euklidischer Norm besitzt und überlege sich, wie man diese mit Hilfe einer QR-Zerlegung von A^T berechnen kann.

Lösung: Die Aufgabe, den Vektor b auf $\text{Bild}\,(A) = \mathbb{R}^m$ zu projizieren, besitzt (wegen der Rangvoraussetzung) eine eindeutige Lösung $b = A\hat{x}$. Durch $\hat{x}+\text{Kern}\,(A)$ ist dann der $(n-m)$-dimensionale affin lineare Lösungsraum gegeben. Die Projektion des Nullvektors auf diesen Lösungsraum ergibt die eindeutige Lösung minimaler Norm. Es sei

$$A^T = Q \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix}$$

mit einer orthogonalen Matrix $Q \in \mathbb{R}^{n\times n}$ und einer nichtsingulären oberen Dreiecksmatrix $\hat{R} \in \mathbb{R}^{m\times m}$. Wir denken uns Q zerlegt in

$$Q = (\ Q_1 \quad Q_2\ ) \qquad \text{mit} \quad Q_1 \in \mathbb{R}^{n\times m},\ Q_2 \in \mathbb{R}^{n\times(n-m)}.$$

Aus

$$\begin{pmatrix} Q_1^T \\ Q_2^T \end{pmatrix} A^T = \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix}$$

erhält man

$$AQ_1\hat{R}^{-T} = I, \qquad AQ_2 = 0.$$

Die $n-m$ Spalten von Q_2 bilden also eine orthonormierte Basis von Kern (A). Wegen $AQ_1\hat{R}^{-T}b = b$ erhält man $\hat{x} = Q_1\hat{R}^{-T}b$ im wesentlichen durch Vorwärtseinsetzen. Die Lösungsmenge ist folglich $\{Q_1\hat{R}^{-T}b + Q_2u : u \in \mathbb{R}^{n-m}\}$. Eine Lösung der Aufgabe

$$\text{Minimiere} \quad \|Q_1\hat{R}^{-T}b + Q_2u\|_2, \quad u \in \mathbb{R}^{n-m},$$

ist durch die zugehörigen Normalgleichungen

$$0 = Q_2^T(Q_1\hat{R}^{-T}b + Q_2u) = u$$

charakterisiert, wobei wir

$$Q^TQ = \begin{pmatrix} Q_1^T \\ Q_2^T \end{pmatrix} \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} = \begin{pmatrix} Q_1^TQ_1 & Q_1^TQ_2 \\ Q_2^TQ_1 & Q_2^TQ_2 \end{pmatrix} = \begin{pmatrix} I_m & 0 \\ 0 & I_{n-m} \end{pmatrix} = I$$

ausgenutzt haben. Also ist $x^* := Q_1\hat{R}^{-T}b$ die Lösung minimaler euklidischer Norm von (P). □

6. Sei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ gegeben. Man zeige, dass $\|I - AA^+\|_2 = \min(1, m-n)$.

   Beweis: Wir gehen von den Darstellungen

$$A = U\begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} V^T, \qquad A^+ = V\begin{pmatrix} \hat{\Sigma}^+ & 0 \end{pmatrix} U^T$$

aus. Dann ist

$$I - AA^+ = I - U\begin{pmatrix} \hat{\Sigma}\hat{\Sigma}^+ & 0 \\ 0 & 0 \end{pmatrix} U^T = U\left[\begin{pmatrix} I_n - \hat{\Sigma}\hat{\Sigma}^+ & 0 \\ 0 & I_{m-n} \end{pmatrix}\right] U^T.$$

Hieraus liest man die Behauptung sofort ab. □

6.2.4 Aufgaben in Abschnitt 3.4

1. Sei $B \in \mathbb{R}^{n\times n}$ eine quadratische Matrix mit der Singulärwertzerlegung $B = U\Sigma V^T$. Sei $C \in \mathbb{R}^{2n\times 2n}$ definiert durch

$$C := \begin{pmatrix} 0 & B^T \\ B & 0 \end{pmatrix}.$$

   Man zeige:

   (a) Die Matrix

$$W := \frac{1}{\sqrt{2}}\begin{pmatrix} V & V \\ U & -U \end{pmatrix}$$

   ist orthogonal.

   (b) Es ist

$$W^TCW = \begin{pmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{pmatrix}.$$

   Sind also $\sigma_1, \ldots, \sigma_n$ die singulären Werte von B, so sind $\pm\sigma_i$, $i = 1, \ldots, n$, die Eigenwerte von C.

Beweis: Es ist

$$W^TW = \frac{1}{2}\begin{pmatrix} V^T & U^T \\ V^T & -U^T \end{pmatrix}\begin{pmatrix} V & V \\ U & -U \end{pmatrix} = \begin{pmatrix} I_n & 0 \\ 0 & I_n \end{pmatrix},$$

also W orthogonal. Die zweite Aussage zeigt man ebenso leicht. □

2. Sei $B \in \mathbb{R}^{n\times n}$ eine obere Bidiagonalmatrix, bei der alle Haupt- und Superdiagonalelemente nicht verschwinden. Man zeige, dass alle singulären Werte $\sigma_1, \ldots, \sigma_n$ von B positiv und einfach sind, also $\sigma_1 > \cdots > \sigma_n > 0$ gilt.

   Hinweis: Man darf benutzen, dass die Eigenwerte einer symmetrischen, unreduzierten Tridiagonalmatrix einfach sind (siehe z. B. J. Werner (1992b, S. 57)).

   Beweis: Die Matrix B^TB ist eine unreduzierte, symmetrische, positiv definite Tridiagonalmatrix. Da die singulären Werte von B gerade die Quadratwurzeln aus den Eigenwerten von B^TB sind, folgt die Behauptung. □

3. Sei $R \in \mathbb{R}^{n\times n}$ eine obere Dreiecksmatrix. Unter Benutzung von Givens-Rotationen gebe man ein Verfahren an, welches die Berechnung einer Singulärwertzerlegung von R auf die einer oberen Bidiagonalmatrix $B \in \mathbb{R}^{n\times n}$ reduziert. Ferner mache man sich das Verfahren für $n = 4$ klar.

   Lösung: Die Idee besteht darin, R sukzessive von rechts mit der Transponierten einer Givens-Rotation und von links mit einer Givens-Rotation zu multiplizieren. Hierbei zähle $i = 1, \ldots, n-2$ die Zeilen und $j = n, \ldots, i+2$ die Spalten, in welchen die obere Dreiecksmatrix R von einer oberen Bidiagonalmatrix abweicht. Genauer könnte das Verfahren folgendermaßen aussehe:

   - Input: Sei $R \in \mathbb{R}^{n\times n}$ eine obere Dreiecksmatrix.
   - Für $i = 1, \ldots, n-2$:
     Für $j = n, \ldots, i+2$:
     Bestimme Givens-Rotation $T_{j-1,j} = T_{j-1,j}(c,s)$ mit

     $$\begin{pmatrix} c & s \\ -s & c \end{pmatrix}\begin{pmatrix} r_{i,j-1} \\ r_{ij} \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}.$$

     Berechne $R := RT_{j-1,j}^T$.
     Bestimme Givens-Rotation $S_{j-1,j} = S_{j-1,j}(c,s)$ mit

     $$\begin{pmatrix} c & s \\ -s & c \end{pmatrix}\begin{pmatrix} r_{j-1,j-1} \\ r_{j,j-1} \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}.$$

     Berechne $R := S_{j-1,j}R$.
   - Output: Nach Abschluss ist R eine obere Bidiagonalmatrix B.

   Wir machen uns das Verfahren an einer 4×4-Matrix klar. Wie üblich bezeichnen wir bei einer Transformation fest bleibende Einträge mit $\bullet$, sich verändernde mit $*$.

$$R = \begin{pmatrix} \bullet & \bullet & \bullet & \bullet \\ & \bullet & \bullet & \bullet \\ & & \bullet & \bullet \\ & & & \bullet \end{pmatrix} \xrightarrow{T_{34}} \begin{pmatrix} \bullet & \bullet & * & \\ & \bullet & * & * \\ & & * & * \\ & & * & * \end{pmatrix} \xrightarrow{S_{34}} \begin{pmatrix} \bullet & \bullet & \bullet & \\ & \bullet & \bullet & \bullet \\ & & * & * \\ & & & * \end{pmatrix} \xrightarrow{T_{23}} \begin{pmatrix} \bullet & * & & \\ & * & * & \bullet \\ & * & * & \bullet \\ & & & \bullet \end{pmatrix}$$

und weiter

$$\xrightarrow{S_{23}} \begin{pmatrix} \bullet & \bullet & & \\ & * & * & * \\ & & * & * \\ & & & \bullet \end{pmatrix} \xrightarrow{T_{34}} \begin{pmatrix} \bullet & \bullet & & \\ & \bullet & * & \\ & & * & * \\ & & * & * \end{pmatrix} \xrightarrow{S_{34}} \begin{pmatrix} \bullet & \bullet & & \\ & \bullet & \bullet & \\ & & * & * \\ & & & * \end{pmatrix} = B$$

Das ist die gewünschte Bidiagonal-Gestalt. □

4. Gegeben sei eine Matrix $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$. Man betrachte die beiden folgenden Möglichkeiten, die Berechnung einer Singulärwertzerlegung von A auf die einer Bidiagonalmatrix $B \in \mathbb{R}^{n\times n}$ zurückzuführen:

   (a) Berechne Householder-Matrizen $P_1, \ldots, P_n \in \mathbb{R}^{m\times m}$ und $Q_1, \ldots, Q_{n-2} \in \mathbb{R}^{n\times n}$ mit
   $$P_n \cdots P_1 A Q_1 \cdots Q_{n-2} = \begin{pmatrix} B \\ 0 \end{pmatrix},$$
   wobei $B \in \mathbb{R}^{n\times n}$ eine obere Bidiagonalmatrix ist.

   (b) Berechne zunächst mit Hilfe des Householder-Verfahrens eine obere Dreiecksmatrix $\hat{R} \in \mathbb{R}^{n\times n}$, für welche mit orthogonalem $Q \in \mathbb{R}^{m\times m}$ gilt, dass
   $$A = Q \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix}.$$
   Anschließend berechne man (siehe Aufgabe 3) mit Hilfe von Givens-Rotationen eine obere Bidiagonalmatrix B mit $B = S\hat{R}T^T$, wobei $S, T \in \mathbb{R}^{n\times n}$ orthogonal sind.

Wieviele flops benötigt man jeweils im wesentlichen (niedere Terme in m und n können unberücksichtigt bleiben) zur Berechnung der oberen Bidiagonalmatrix B, für welche m, n ist welche Methode vorzuziehen?

Lösung: Zunächst betrachten wir die Methode (a). Auf S. 70 ist das Verfahren genauer beschrieben. Die Berechnung der Householder-Matrizen $\overline{P}_k$, $k = 1, \ldots, n$, und $\overline{Q}_k$, $k = 1, \ldots, n-2$, ist für den operational count unerheblich. Wichtig ist der Aufwand zur Berechnung von $A := P_k A$ und $A := AQ_k$. Zur Berechnung von

$$\overline{P}_k \begin{pmatrix} a_{kk} & \cdots & a_{kn} \\ \vdots & & \vdots \\ a_{mk} & \cdots & a_{mk} \end{pmatrix}$$

benötigt man $(n-k+1)2(m-k+1)$ flops. Für die Multiplikation von links mit den Householder-Matrizen $P_1, \ldots, P_n$ werden also insgesamt

$$2\sum_{k=1}^{n}(n-k+1)(m-k+1) = 2\sum_{j=1}^{n} j(m-n+j) = 2\left[(m-n)\frac{n(n+1)}{2} + \frac{n(n+1)(2n+1)}{6}\right],$$

also im wesentlichen $mn^2 - \frac{1}{3}n^3$ flops. Entsprechend werden zur Berechnung von

$$\begin{pmatrix} a_{k,k+1} & \cdots & a_{k,n} \\ \vdots & & \vdots \\ a_{m,k+1} & \cdots & a_{mn} \end{pmatrix} \overline{Q}_k$$

insgesamt $(m-k+1)(n-k)$ flops benötigt. Für die Multiplikation von rechts mit den Householder-Matrizen $Q_1, \ldots, Q_{n-2}$ braucht man daher noch einmal

$$2\sum_{k=1}^{n-2}(m-k+1)(n-k)$$

bzw. im wesentlichen ebenfalls $mn^2 - \frac{1}{3}n^3$ flops. Die Komplexität der Methode (a) ist also durch $2mn^2 - \frac{2}{3}n^3$ gegeben.

Nun zur Methode (b). Zunächst müssen wir einen operational count für die Berechnung einer QR-Zerlegung mit dem Householder-Verfahren nachholen. Im wesentlichen handelt es sich hier um $mn^2 - \frac{1}{3}n^3$ flops, da im wesentlichen nur der Aufwand bei der Multiplikation von A mit den Householder-Matrizen $P_1, \ldots, P_n$ zu zählen ist. Zu zählen bleibt der Aufwand zur Bidiagonalisierung einer oberen Dreiecksmatrix $\hat{R} \in \mathbb{R}^{n\times n}$ durch eine Multiplikation von rechts und links mit Givens-Rotationen. Die Berechnung dieser Givens-Rotationen spielt wieder keine Rolle, zu zählen ist der Aufwand zur Multiplikation der gegebenen Matrix mit ihnen. Die Annullierung die obere Bidiagonal-Gestalt störender Elemente erfolgt zeilenweise. In der i-ten stören die Elemente in den Spalten $j = n, \ldots, i+2$, der Aufwand zur Annullierung dieser Einträge ist (unabhängig von der Spalte j) durch $4(n+2-i)$ gegeben. Der Gesamtaufwand zur Bidiagonalisierung einer oberen $n \times n$-Dreiecksmatrix ist also

$$4\sum_{i=1}^{n-2}\sum_{j=i+2}^{n}(n+2-i) = \tfrac{4}{3}n^3 + \cdots.$$

Insgesamt ist die Komplexität der Methode (b) also im wesentlichen durch $mn^2 + n^3$ gegeben. Es stellt sich also heraus, dass für $m > \frac{5}{3}n$ die Methode (b) billiger ist. Hierbei ist aber zu berücksichtigen, dass das Akkumulieren der benutzten orthogonalen Matrizen (insbesondere der im zweiten Schritt benutzten Givens-Rotationen) bei der Methode (b) teurer ist. Vorzuziehen wäre diese also nur dann, wenn nur die singulären Werte der gegebenen Matrix zu bestimmen sind. □

5. Seien

$$A := \begin{pmatrix} 1 & 0.04 & 0.0016 \\ 1 & 0.32 & 0.1024 \\ 1 & 0.51 & 0.2601 \\ 1 & 0.73 & 0.5329 \\ 1 & 1.03 & 1.0609 \\ 1 & 1.42 & 2.0164 \\ 1 & 1.60 & 2.5600 \end{pmatrix}, \qquad b := \begin{pmatrix} 2.63 \\ 1.18 \\ 1.16 \\ 1.54 \\ 2.65 \\ 5.41 \\ 7.67 \end{pmatrix}$$

gegeben. Man berechne mit dem Golub-Reinsch-Verfahren eine reduzierte Singulärwertzerlegung von A und anschließend die (eindeutige) Lösung des linearen Ausgleichsproblems mit den Daten $(\ A \quad b\ )$.

Lösung: Bei unserer Implementation wurde zunächst $A = PBQ^T$ mit einer Bidiagonalmatrix B berechnet. Wir erhielten (die Angabe von P und Q lassen wir weg)

$$B = \begin{pmatrix} -2.6457513111 & 3.2649562782 & \\ & -2.7530747251 & 0.5501033275 \\ & & 0.3153203113 \end{pmatrix}.$$

Anschließend berechneten wir die reduzierte Singulärwertzerlegung $A = \hat{U}\hat{\Sigma}V^T$ mit

$$\hat{U} = \begin{pmatrix} -0.1035187665 & -0.5280205209 & 0.7050064236 \\ -0.1492374206 & -0.4852997885 & 0.0740752156 \\ -0.1933499389 & -0.4266057150 & -0.2132222953 \\ -0.2576486550 & -0.3286397726 & -0.4036288071 \\ -0.3681944194 & -0.1431575692 & -0.4172475855 \\ -0.5513472405 & 0.1874834817 & -0.0105563452 \\ -0.6509179049 & 0.3742161983 & 0.3389570178 \end{pmatrix},$$

die singulären Werte in der Diagonalmatrix

$$\hat{\Sigma} = \operatorname{diag}\begin{pmatrix} 4.7962000524 & 1.5962018480 & 0.3000087126 \end{pmatrix}$$

sowie

$$V = \begin{pmatrix} -0.4741700348 & -0.8457725368 & 0.2446049756 \\ -0.5300468424 & 0.0523915339 & -0.8463483160 \\ -0.7030029324 & 0.5309651054 & 0.4731415580 \end{pmatrix}.$$

Bis auf das unterschiedliche Vorzeichen in den ersten Spalten von $\hat{U}$ und V stimmen die Ergebnisse in der angegebenen Genauigkeit genau mit denen überein, die wir mit dem MATLAB-Befehl `[U,Sigma,V]=svd(A,0)` erhalten haben[8]. Als Lösung des zu den Daten $\begin{pmatrix} A & b \end{pmatrix}$ gehörenden linearen Ausgleichsproblems erhalten wir

$$x = \begin{pmatrix} 2.7491976489 \\ -5.9546574777 \\ 5.6072465615 \end{pmatrix}.$$

Dies stimmt wiederum (eventuell bis auf die letzte angegebene Dezimalstelle) mit den durch die MATLAB-Befehle `x=A\b` oder `x=pinv(A)*b` erzielten Ergebnissen überein. Im ersten Fall wird eine QR-Zerlegung mit Spaltenpivotisierung gemacht, im zweiten die Pseudoinverse von A mit b multipliziert. □

6. Seien

$$A := \begin{pmatrix} 22 & 10 & 2 & 3 & 7 \\ 14 & 7 & 10 & 0 & 8 \\ -1 & 13 & -1 & -11 & 3 \\ -3 & -2 & 13 & -2 & 4 \\ 9 & 8 & 1 & -2 & 4 \\ 9 & 1 & -7 & 5 & -1 \\ 2 & -6 & 6 & 5 & 1 \\ 4 & 5 & 0 & -2 & 2 \end{pmatrix}, \qquad B := \begin{pmatrix} -1 & 1 & 0 \\ 2 & -1 & 1 \\ 1 & 10 & 11 \\ 4 & 0 & 4 \\ 0 & -6 & -6 \\ -3 & 6 & 3 \\ 1 & 11 & 12 \\ 0 & -5 & -5 \end{pmatrix}$$

gegeben. Man berechne eine reduzierte Singulärwertzerlegung von A. Anschließend löse man die linearen Ausgleichsprobleme mit A als Koeffizientenmatrix und den Spalten von B als rechter Seite, genauer bestimme man jeweils die Lösung mit minimaler euklidischer Norm.

[8] Siehe auch H. R. Schwarz (1997, S. 364) *Numerische Mathematik. 4. Auflage.* Teubner, Stuttgart.

Lösung: Wir erhalten die reduzierte Singulärwertzerlegung $A = \hat{U}\hat{\Sigma}V^T$ mit

$$\hat{U} = \begin{pmatrix} -0.7071067812 & -0.1581138830 & 0.1767766953 & -0.5152210984 & 0.3669126844 \\ -0.5303300859 & -0.1581138830 & -0.3535533906 & 0.5490119754 & 0.0953509029 \\ -0.1767766953 & 0.7905694151 & -0.1767766952 & 0.2383655928 & 0.1602041608 \\ -0.0000000000 & -0.1581138829 & -0.7071067812 & -0.3739302188 & -0.3588064073 \\ -0.3535533906 & 0.1581138830 & 0.0000000000 & -0.0393036955 & -0.6960409511 \\ -0.1767766953 & -0.1581138831 & 0.5303300859 & 0.2558566523 & -0.4430909157 \\ 0.0000000000 & -0.4743416490 & -0.1767766953 & 0.4116801903 & 0.1219595998 \\ -0.1767766953 & 0.1581138830 & 0.0000000000 & -0.0017663870 & -0.0787347892 \end{pmatrix}$$

und

$$\hat{\Sigma} = \begin{pmatrix} 35.3270434655 & & & & \\ & 19.9999999999 & & & \\ & & 19.5959179423 & & \\ & & & 0.0000000000 & \\ & & & & 0.0000000000 \end{pmatrix}$$

sowie

$$V = \begin{pmatrix} -0.8006407690 & -0.3162277660 & 0.2886751346 & 0.4190954851 & 0.0000000000 \\ -0.4803844614 & 0.6324555320 & 0.0000000001 & -0.4405091230 & -0.4185480638 \\ -0.1601281538 & -0.3162277659 & -0.8660254038 & 0.0520045492 & -0.3487900532 \\ 0.0000000000 & -0.6324555321 & 0.2886751345 & -0.6760591402 & -0.2441530372 \\ -0.3202563076 & 0.0000000000 & -0.2886751346 & -0.4129773028 & 0.8022171224 \end{pmatrix}.$$

Insbesondere haben wir festgestellt, dass die 8×5-Matrix A den Rang 5 besitzt. Bis auf das Vorzeichen in den ersten beiden Spalten von U und V stimmt dies in den ersten drei Spalten von $\hat{U}$ und V (denen zu positiven singulären Werten) genau mit dem MATLAB gewonnenen Ergebnis überein. Als Lösungen x_1, x_2, x_3 minimaler euklidischer Norm der drei Ausgleichsprobleme erhalten wir[9]

x_1	x_2	x_3
−0.0833333333	0.0000000000	−0.0833333333
0.0000000000	0.0000000000	0.0000000000
0.2500000000	0.0000000000	0.2500000000
−0.0833333333	0.0000000000	−0.0833333333
0.0833333333	0.0000000000	0.0833333333

□

7. Gegeben sei die obere Bidiagonalmatrix

$$B := \begin{pmatrix} d_1 & f_2 & & & \\ & d_2 & f_3 & & \\ & & \ddots & \ddots & \\ & & & d_{n-1} & f_n \\ & & & & d_n \end{pmatrix} \in \mathbb{R}^{n \times n}.$$

Man bilde die Matrix

$$C := \begin{pmatrix} 0 & B^T \\ B & 0 \end{pmatrix} \in \mathbb{R}^{2n \times 2n}$$

und zeige:

[9] Siehe auch J. H. Wilkinson, C. Reinsch (1971, S. 149) *Handbook for Automatic Computation. Vol. II, Linear Algebra*. Springer, Berlin-Heidelberg-New York.

(a) Es gibt eine Permutationsmatrix $P \in \mathbb{R}^{2n\times 2n}$ mit

$$T := P^T C P = \begin{pmatrix} 0 & d_1 & & & & \\ d_1 & 0 & f_2 & & & \\ & f_2 & 0 & d_2 & & \\ & & d_2 & 0 & \ddots & \\ & & & \ddots & \ddots & d_n \\ & & & & d_n & 0 \end{pmatrix}.$$

(b) Gegeben sei die obere Bidiagonalmatrix $B \in \mathbb{R}^{n\times n}$ und hiermit die spezielle Tridiagonalmatrix T aus 7a. Man betrachte den folgenden Algorithmus, von dem wir annehmen, dass er durchführbar ist.

- Gegeben sei $\sigma > 0$.
- Setze $\delta := -\sigma$, $\omega := 0$.
- Für $k = 2, \ldots, 2n$:
  Setze
  $$\alpha := \begin{cases} d_{k/2}, & \text{falls } k \text{ gerade,} \\ f_{(k+1)/2}, & \text{falls } k \text{ ungerade.} \end{cases}$$
  Berechne
  $$\delta := -\sigma - \frac{\alpha^2}{\delta}.$$
  Falls $\delta > 0$, dann: $\omega := \omega + 1$.

Man zeige, dass nach Abschluss des Verfahrens ω die Anzahl der singulären Werte von B angibt, die größer als σ sind.

Hinweis: Man entwickle ein Verfahren zur Berechnung der LDL^T-Zerlegung von $T-\sigma I$ und denke an den Sylvesterschen Trägheitssatz (siehe Satz 3.3 in Abschnitt 2.3).

Beweis: Offenbar tut die Permutationsmatrix

$$P := (\ e_{n+1} \quad e_1 \quad e_{n+2} \quad e_2 \quad \cdots \quad e_{2n} \quad e_n\ )$$

das Verlangte. Denn für $1 \le i, j \le 2n$ ist

$$t_{ij} = (Pe_i)^T C (Pe_j),$$

wobei

$$Pe_k = \begin{cases} e_{k/2}, & \text{falls } k \text{ gerade,} \\ e_{n+(k+1)/2}, & \text{falls } k \text{ ungerade.} \end{cases}$$

Durch eine einfache Diskussion erhält man, dass P die richtige Permutationsmatrix ist.

Wir nehmen einmal an, die Matrix $T - \sigma I$ lasse sich in der Form $T - \sigma I = L\Delta L^T$ darstellen, wobei

$$L = \begin{pmatrix} 1 & & & \\ \alpha_2 & 1 & & \\ & \ddots & \ddots & \\ & & \alpha_{2n} & 1 \end{pmatrix}, \qquad \Delta = \operatorname{diag}(\delta_1, \ldots, \delta_{2n}).$$

Es läst sich sehr leicht nachweisen, dass der obige Algorithmus sukzessive die Diagonalelemente von Δ und der unteren Bidiagonalmatrix L berechnet. Die Durchführbarkeit des obigen Algorithmus ist also äquivalent zur Existenz einer LDL^T-Zerlegung von $T - \sigma I$. Nach Abschluss gibt ω die Anzahl positiver Elemente in der Diagonalmatrix Δ an. Nach dem Sylvesterschen Trägheitssatz ist ω gleich der Anzahl positiver Eigenwerte von $T - \sigma I$, bzw. die Anzahl derjenigen Eigenwerte von T, die größer als σ sind. Nun ist T orthogonal ähnlich zu C, hat also dieselben Eigenwerte. Nach Aufgabe 1 sind die Eigenwerte von C aber die singulären Werte von B und deren Negatives. Damit ist die Behauptung bewiesen. □

6.2.5 Aufgaben in Abschnitt 3.5

1. Mit $\epsilon \in (0, 1]$ sei

$$A := \begin{pmatrix} 1 & 0 \\ 0 & \epsilon \\ 0 & 0 \end{pmatrix}, \quad \delta A := \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & \epsilon/2 \end{pmatrix}, \quad b := \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \delta b := \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Man berechne die zum linearen Ausgleichsproblem mit den Daten (A, b) bzw. $(A + \delta A, b + \delta b)$ gehörende Lösung x bzw. $x + \delta x$ und vergleiche die in Satz 5.4 angegebene Abschätzung von $\|\delta x\|_2/\|x\|_2$ mit dem exakten Wert.

Lösung: Es ist $\|A\|_2 = 1$ und

$$A^+ = (A^T A)^{-1} A^T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/\epsilon & 0 \end{pmatrix}, \quad \kappa_2(A) = \|A^+\|_2 = 1/\epsilon.$$

Weiter ist $\|\delta A\|_2 = \epsilon/2$ und damit $\|A^+\|_2 \, \|\delta A\|_2 = 1/2$, die Voraussetzungen von Satz 5.4 sind damit erfüllt. Weiter berechnet man

$$x = A^+ b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \|Ax - b\|_2 = 1, \qquad x + \delta x = (A + \delta A)^+ b = \begin{pmatrix} 1 \\ 2/(5\epsilon) \end{pmatrix}.$$

Satz 5.4 liefert

$$\begin{aligned} \frac{2}{5\epsilon} &= \frac{\|\delta x\|_2}{\|x\|_2} \\ &\leq \frac{\kappa_2(A)}{1 - \kappa_2(A)\|\delta A\|_2/\|A\|_2} \left[\frac{\|\delta A\|_2}{\|A\|_2} \left(1 + \kappa_2(A) \frac{\|Ax - b\|_2}{\|A\|_2 \, \|x\|_2} \right) + \frac{\|\delta b\|_2}{\|A\|_2 \, \|x\|_2} \right] \\ &= \frac{1/\epsilon}{1 - 1/\epsilon \cdot \epsilon/2} \left[\frac{\epsilon}{2} (1 + 1/\epsilon \cdot 1) \right] \\ &= 1 + \frac{1}{\epsilon}. \end{aligned}$$

Die entscheidende Aussage, dass nämlich der relative Fehler mit $\epsilon \to 0+$ beliebig groß wird, zeigt sich also auch in der Abschätzung von Satz 5.4. □

2. Sei $A\colon (a, b) \longrightarrow \mathbb{R}^{m \times n}$ mit $m \geq n$ stetig differenzierbar und Rang $(A(t_0)) = n$ für ein $t_0 \in (a, b)$. Man zeige:

   (a) Es existiert eine Umgebung $I \subset (a, b)$ mit Rang $(A(t)) = n$ für alle $t \in I$.

(b) Für $t \in I$ sei $x(t) \in \mathbb{R}^n$ die eindeutige Lösung des linearen Ausgleichsproblems

$$\text{Minimiere} \quad \|A(t)x - b\|_2, \quad x \in \mathbb{R}^n,$$

wobei $b \in \mathbb{R}^m$ vorgegeben ist. Dann ist $x\colon I \longrightarrow \mathbb{R}^n$ stetig differenzierbar und die Ableitung $\dot{x}(t)$ für $t \in I$ die Lösung des quadratischen, nichtsingulären Gleichungssystem

$$A(t)^T A(t)z = -A(t)^T \dot{A}(t)x(t) + \dot{A}(t)^T(b - A(t)x(t)).$$

(c) Weiter ist $\dot{x}(t)$ für $t \in I$ die Lösung des linearen Ausgleichsproblems

$$\text{Minimiere} \quad \|A(t)z + \dot{A}(t)x(t) - (A(t)^+)^T \dot{A}(t)^T(b - A(t)x(t))\|_2, \quad z \in \mathbb{R}^n.$$

Beweis: Es ist Rang $(A(t)) = n$ genau dann, wenn $A(t)^T A(t)$ nichtsingulär ist. Da dies nach Voraussetzung für $t = t_0$ der Fall ist, gilt dies aus Stetigkeitsgründen auch in einer Umgebung I von t_0. Für $t \in I$ ist

$$x(t) = (A(t)^T A(t))^{-1} A(t)^T b.$$

Hieraus liest man ab, dass x auf I stetig differenzierbar ist. Wegen der Produktregel zur Differentiation ist die Ableitung durch

$$\dot{x}(t) = \left[\frac{d}{dt}(A(t)^T A(t))^{-1}\right] A(t)^T b + (A(t)^T A(t))^{-1} \dot{A}(t)^T b$$

gegeben. Um den ersten Term genauer auszurechnen, setzen wir zur Abkürzung $Y(t) := A(t)^T A(t)$. Dann ist

$$0 = \frac{d}{dt}[Y(t)Y^(t)^{-1}] = \dot{Y}(t)Y(t)^{-1} + Y(t)\frac{d}{dt}[Y(t)^{-1}]$$

und folglich

$$\begin{aligned} \frac{d}{dt}[Y(t)^{-1}] &= -Y(t)^{-1}\dot{Y}(t)Y(t)^{-1} \\ &= -(A(t)^T A(t))^{-1}[\dot{A}(t)^T A(t) + A(t)^T \dot{A}(t)](A(t)^T A(t))^{-1}. \end{aligned}$$

Insgesamt ist dann

$$\begin{aligned} \dot{x}(t) &= -(A(t)^T A(t))^{-1}[\dot{A}(t)^T A(t) + A(t)^T \dot{A}(t)](A(t)^T A(t))^{-1} A(t)^T b \\ &\quad + (A(t)^T A(t))^{-1} \dot{A}(t)^T b \\ &= -(A(t)^T A(t))^{-1}[\dot{A}(t)^T A(t) + A(t)^T \dot{A}(t)]x(t) + (A(t)^T A(t))^{-1} \dot{A}(t)^T b \\ &= (A(t)^T A(t))^{-1} \dot{A}(t)^T(b - A(t)x(t)) - (A(t)^T A(t))^{-1} A(t)^T \dot{A}(t)x(t). \end{aligned}$$

Mit anderen Worten ist $\dot{x}(t)$ die eindeutige Lösung des linearen Gleichungssystems

$$A(t)^T A(t)z = -A(t)^T \dot{A}(t)x(t) + \dot{A}(t)^T(b - A(t)x(t)).$$

Dieses lineare Gleichungssystem kann wiederum wegen $(A(t)^T(A(t)^+)^T = I$ als Normalgleichung zum linearen Ausgleichsproblem

$$\text{Minimiere} \quad \|A(t)z + \dot{A}(t)x(t) - (A(t)^+)^T \dot{A}(t)^T(b - A(t)x(t))\|_2, \quad z \in \mathbb{R}^n$$

interpretiert werden. Damit ist alles gezeigt. □

6.2.6 Aufgaben in Abschnitt 3.6

1. Gegeben sei die Aufgabe

$$(\mathrm{P}_\tau) \qquad \text{Minimiere} \quad \|Ax-b\|_2^2+\tau\,\|x\|_2^2, \quad x\in\mathbb{R}^n,$$

wobei $A\in\mathbb{R}^{m\times n}$ mit $m\geq n$, $b\in\mathbb{R}^m$ und $\tau>0$ gegeben sind. Sei x_τ die eindeutige Lösung von (P_τ). Mit Hilfe einer Singulärwertzerlegung von A berechne man

$$f(\tau):=\|x_\tau\|_2, \qquad g(\tau):=\|Ax_\tau-b\|_2.$$

Man zeige, dass g auf $[0,\infty)$ monoton nicht fallend und i. allg. monoton wachsend ist. Anschließend zeige man, dass $h:=f\circ g^{-1}$ auf dem Existenzintervall eine monoton fallende Funktion ist.

Lösung: Sei $A=\hat{U}\hat{\Sigma}V^T$ mit $\hat{U}=(\,u_1\ \cdots\ u_n\,)$ und $\hat{\Sigma}=\mathrm{diag}\,(\sigma_1,\ldots,\sigma_n)$ eine reduzierte Singulärwertzerlegung von A und r der Rang von A. Die Lösung x_τ von (P_τ) ist als Lösung der Normalgleichungen durch

$$x_\tau=(A^TA+\tau I)^{-1}A^Tb=V(\hat{\Sigma}^2+\tau I)^{-1}\hat{\Sigma}\hat{U}^Tb$$

gegeben. Folglich ist

$$f(\tau)=\|x_\tau\|_2=\|(\hat{\Sigma}^2+\tau)^{-1}\hat{\Sigma}\hat{U}^Tb\|_2=\left[\sum_{i=1}^r\left(\frac{\sigma_i}{\sigma_i^2+\tau}u_i^Tb\right)^2\right]^{1/2}.$$

Weiter ist

$$\begin{aligned}
g(\tau) &= \|Ax_\tau-b\|_2\\
&= \|\hat{U}\hat{\Sigma}^2(\hat{\Sigma}^2+\tau I)^{-1}\hat{U}^Tb-b\|_2\\
&= \|\hat{U}[\hat{\Sigma}^2(\hat{\Sigma}^2+\tau I)^{-1}-I]\hat{U}^Tb+(\hat{U}\hat{U}^T-I)b\|_2\\
&= [\|\hat{U}[\hat{\Sigma}^2(\hat{\Sigma}^2+\tau I)^{-1}-I]\hat{U}^Tb\|_2^2+\|(\hat{U}\hat{U}^T-I)b\|_2^2]^{1/2}\\
&= [\|[\hat{\Sigma}^2(\hat{\Sigma}^2+\tau I)^{-1}-I]\hat{U}^Tb\|_2^2+\|(\hat{U}\hat{U}^T-I)b\|_2^2]^{1/2}\\
&= [\|\tau(\hat{\Sigma}^2+\tau I)^{-1}\hat{U}^Tb\|_2^2+\|(\hat{U}\hat{U}^T-I)b\|_2^2]^{1/2}\\
&= \left[\sum_{i=1}^n\left(\frac{\tau}{\sigma_i^2+\tau}u_i^Tb\right)^2+\|(\hat{U}\hat{U}^T-I)b\|_2^2\right]^{1/2}\\
&= \left[\sum_{i=1}^r\left(\frac{\tau}{\sigma_i^2+\tau}u_i^Tb\right)^2+\sum_{i=r+1}^n(u_i^Tb)^2+\|(\hat{U}\hat{U}^T-I)b\|_2^2\right]^{1/2}.
\end{aligned}$$

Es ist

$$g(0)=\left[\sum_{i=r+1}^n(u_i^Tb)^2+\|(\hat{U}\hat{U}^T-I)b\|_2^2\right]^{1/2}=:\delta_0$$

und

$$\lim_{\tau\to\infty}g(\tau)=\left[\sum_{i=1}^n(u_i^Tb)^2+\|(\hat{U}\hat{U}^T-I)b\|_2^2\right]^{1/2}=:\delta_\infty.$$

Weiter ist

$$g'(\tau)=\frac{\tau}{g(\tau)}\sum_{i=1}^r\frac{\sigma_i^2}{(\sigma_i^2+\tau)^3}(u_i^Tb)^2,$$

woraus man abliest, dass $g'(\tau) > 0$ auf $(0, \infty)$, wenn nicht $u_i^T b = 0$, $i = 1, \ldots, r$. Weiter ist $\delta_0 < \delta_\infty$, wenn nicht $u_i^T b = 0$, $i = r+1, \ldots, n$. Beides wird in Zukunft ausgeschlossen. Dann ist g^{-1} und damit auch $h = f \circ g^{-1}$ auf $(\delta_0, \delta_\infty)$ erklärt. Als Ableitung von f berechnet man

$$f'(\tau) = -\frac{1}{f(\tau)} \sum_{i=1}^{r} \frac{\sigma_i^2}{(\sigma_i^2 + \tau)^3} (u_i^T b)^2,$$

also ist f monoton fallend. Für $\alpha \in (\delta_0, \delta_\infty)$ sei $\tau = g^{-1}(\alpha)$ bzw. $g(\tau) = \alpha$. Dann ist

$$h'(\alpha) = \frac{f'(g^{-1}(\alpha))}{g'(g^{-1}(\alpha))} = -\frac{g(g^{-1}(\alpha))}{f(g^{-1}(\alpha)) g^{-1}(\alpha)} = -\frac{\alpha}{h(\alpha) g^{-1}(\alpha)} < 0,$$

also h auf $(\delta_0, \delta_\infty)$ monoton fallend. Genau das[10] war behauptet.

2. Man präzisiere die folgende Vorgehensweise zur Lösung von

$$(\mathrm{P}_\tau) \qquad \text{Minimiere} \quad \|Ax - b\|_2^2 + \tau \|x\|_2^2, \quad x \in \mathbb{R}^n$$

bzw.

$$(\mathrm{P}_\tau) \qquad \text{Minimiere} \quad \left\| \begin{pmatrix} A \\ \sqrt{\tau} I \end{pmatrix} x - \begin{pmatrix} b \\ 0 \end{pmatrix} \right\|_2, \quad x \in \mathbb{R}^n,$$

wobei $\tau > 0$ und $A \in \mathbb{R}^{m \times n}$ mit $m \geq n$, $b \in \mathbb{R}^m$.

- Durch ein auf A angewandtes Bidiagonalisierungsverfahren reduziere man die Aufgabe (P_τ) auf das lineare Ausgleichsproblem

$$(\hat{\mathrm{P}}_\tau) \qquad \text{Minimiere} \quad \left\| \begin{pmatrix} B \\ \sqrt{\tau} I \end{pmatrix} y - \begin{pmatrix} c \\ 0 \end{pmatrix} \right\|_2, \quad y \in \mathbb{R}^n,$$

  wobei $B \in \mathbb{R}^{n \times n}$ eine obere Bidiagonalmatrix und $c \in \mathbb{R}^n$ ist. Dieser Teil ist also von der Wahl von $\tau > 0$ unabhängig.

- Man überlege sich, wie man die Aufgabe $(\hat{\mathrm{P}}_\tau)$ effizient lösen kann, indem man die Koeffizientenmatrix durch Multiplikation mit geeigneten Givens-Rotationen in eine obere Bidiagonalmatrix überführt.

Lösung: Mit dem Bidiagonalisierungsverfahren von Golub-Kahan berechne man orthogonale Matrizen $P \in \mathbb{R}^{m \times m}$ und $Q \in \mathbb{R}^{n \times n}$ sowie eine obere Bidiagonalmatrix $B \in \mathbb{R}^{n \times n}$ mit

$$A = P \begin{pmatrix} B \\ 0 \end{pmatrix} Q^T.$$

Wir denken uns $P^T b$ durch

$$P^T b = \begin{pmatrix} c \\ d \end{pmatrix} \qquad \text{mit} \quad c \in \mathbb{R}^n, \quad d \in \mathbb{R}^{m-n}$$

[10] Bei P. C. Hansen (1998, S. 85) wird auch noch die Konvexität von h behauptet. Dies konnte ich nicht nachweisen. Die Abbildung bei Hansen spricht auch eher gegen die Gültigkeit einer solchen Aussage.

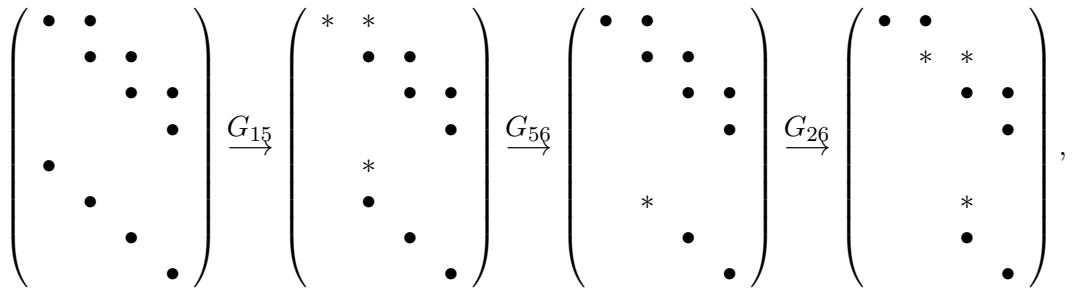
partitioniert. Für ein beliebiges $x \in \mathbb{R}^n$ ist dann

$$\begin{aligned}\left\| \begin{pmatrix} A \\ \sqrt{\tau} I \end{pmatrix} - \begin{pmatrix} b \\ 0 \end{pmatrix} \right\|_2^2 &= \|Ax - b\|_2^2 + \tau \|x\|_2^2 \\ &= \|P^T A Q Q^T x - P^T b\|_2^2 + \tau \|Q^T x\|_2^2 \\ &= \|BQ^T x - c\|_2^2 + \tau \|Q^T x\|_2^2 + \|d\|_2^2 \\ &= \left\| \begin{pmatrix} B \\ \sqrt{\tau} I \end{pmatrix} Q^T x - \begin{pmatrix} c \\ 0 \end{pmatrix} \right\|_2^2 + \|d\|_2^2.\end{aligned}$$

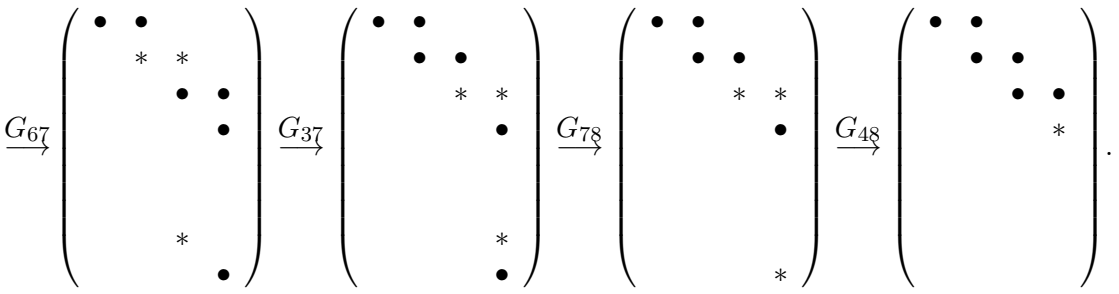
Ist also $y \in \mathbb{R}^n$ die Lösung von

$$(\hat{\mathrm{P}}_\tau) \qquad \text{Minimiere} \quad \left\| \begin{pmatrix} B \\ \sqrt{\tau} I \end{pmatrix} y - \begin{pmatrix} c \\ 0 \end{pmatrix} \right\|_2, \quad y \in \mathbb{R}^n,$$

so ist $x = Qy$ die Lösung von (P_τ). Nun ist die Frage, wie man die Koeffizientenmatrix in $(\hat{\mathrm{P}}_\tau)$ durch Multiplikation von links mit geeigneten Givens-Rotationen auf obere Bidiagonal-Gestalt transformieren kann. Die Vorgehensweise machen wir uns für $n = 4$ klar. Zunächst berechnen wir



anschließend



In den ungeraden Schritten können wir die früher definierte Funktion "givrot" benutzen, in den geraden Schritten sind c, s mit $c^2 + s^2 = 1$ so zu bestimmen, dass

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 0 \\ \gamma \end{pmatrix}.$$

Die allgemeine Vorgehensweise dürfte jetzt klar sein. Wir multiplizieren die gegebene Matrix

$$\begin{pmatrix} B \\ \sqrt{\tau} I \end{pmatrix}$$

abwechseln mit Givens-Rotationen $G_{i,n+i}$, (zum Annullieren des Eintrages in der Position $(n+i, i)$), $i = 1, \ldots, n$, und $G_{n+i,n+i+1}$ (zum Annullieren des gerade vorher eingeführten Eintrages in der Position $(n+i, i+1)$), $i = 1, \ldots, n-1$. □

3. Gegeben sei die Aufgabe

$$\text{Minimiere} \quad \|Ax-b\|_2^2 + \tau \|Lx\|_2^2, \quad x \in \mathbb{R}^n,$$

wobei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$, $b \in \mathbb{R}^m$, $L \in \mathbb{R}^{p\times n}$ mit Rang $(L) = p \leq n$ sowie

$$\text{Rang} \begin{pmatrix} A \\ L \end{pmatrix} = n$$

und $\tau > 0$ gegeben sind. Wir wissen, dass diese Aufgabe auch als lineares Ausgleichsproblem

$$(\mathrm{P}_\tau) \qquad \text{Minimiere} \quad \left\| \begin{pmatrix} A \\ \sqrt{\tau}L \end{pmatrix} x - \begin{pmatrix} b \\ 0 \end{pmatrix} \right\|_2, \quad x \in \mathbb{R}^n,$$

geschrieben werden kann.

(a) Man zeige, dass die folgenden Schritte durchführbar sind:

- Berechne eine QR-Zerlegung von $L^T \in \mathbb{R}^{n\times p}$, also

$$L^T = V \begin{pmatrix} R \\ 0 \end{pmatrix}, \qquad V = \begin{pmatrix} V_1 & V_2 \end{pmatrix}.$$

  Hierbei sind natürlich $V_1 \in \mathbb{R}^{n\times p}$ und $V_2 \in \mathbb{R}^{n\times(n-p)}$.
- Berechne $AV_2 \in \mathbb{R}^{m\times(n-p)}$ und eine QR-Zerlegung dieser Matrix:

$$AV_2 = Q \begin{pmatrix} U \\ 0 \end{pmatrix}, \qquad Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}.$$

  Hierbei sind natürlich $Q_1 \in \mathbb{R}^{m\times(n-p)}$ und $Q_2 \in \mathbb{R}^{m\times(m-n+p)}$. Man zeige, dass U nichtsingulär bzw. Rang $(AV_2) = n-p$ ist.
- Berechne

$$\tilde{A} := Q_2^T A V_1 R^{-T} \in \mathbb{R}^{(m-n+p)\times p}, \qquad \tilde{b} := Q_2^T b \in \mathbb{R}^{m-n+p}.$$

(b) Man zeige: Ist $\tilde{x}$ die Lösung des Problems

$$(\tilde{\mathrm{P}}_\tau) \qquad \text{Minimiere} \quad \left\| \begin{pmatrix} \tilde{A} \\ \sqrt{\tau}I \end{pmatrix} \tilde{x} - \begin{pmatrix} \tilde{b} \\ 0 \end{pmatrix} \right\|_2, \quad \tilde{x} \in \mathbb{R}^p$$

in Standardform, so ist

$$x := V_1 R^{-T} \tilde{x} + V_2 U^{-1} Q_1^T (b - AV_1 R^{-T} \tilde{x})$$

die Lösung von (P_τ).

Beweis: Da wir Rang $(L) = p \leq n$ vorausgesetzt haben[11], ist in der QR-Zerlegung von L^T natürlich $R \in \mathbb{R}^{p\times p}$ nichtsingulär. Nun überlegen wir uns, dass $AV_2 \in \mathbb{R}^{m\times(n-p)}$ den

[11] Ist $p = n$, so kann man sich die restlichen Schritte schenken. Denn dann ist $L = R^T V^T$ und daher

$$\|Ax-b\|_2^2 + \tau\|Lx\|_2^2 = \|\underbrace{AVR^{-T}}_{=\tilde{A}} \underbrace{R^T V^T x}_{=\tilde{x}} - b\|_2^2 + \tau \|\underbrace{R^T V^T x}_{=\tilde{x}}\|_2^2,$$

so dass die Transformation auf Standardform in einfacher Weise gelungen ist.

Rang $n-p$ besitzt, die Spalten dieser Matrix also linear unabhängig sind. Ist $AV_2y_2 = 0$ mit einem $y_2 \in \mathbb{R}^{n-p}$, so ist

$$LV_2y_2 = \begin{pmatrix} R^T & 0 \end{pmatrix} \begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix} V_2y_2 = R^T V_1^T V_2 y_2 = 0,$$

da die Spalten von V_1 senkrecht auf denen von V_2 stehen. Also ist

$$\begin{pmatrix} A \\ L \end{pmatrix} V_2y_2 = 0,$$

wegen der Rangvoraussetzung ist $V_2y_2 = 0$ und damit schließlich $y_2 = 0$, da die Spalten von V_2 ein Orthonormalsystem bilden und daher insbesondere linear unabhängig sind. Daher ist $U \in \mathbb{R}^{(n-p)\times(n-p)}$ nichtsingulär.

Den Rest des Beweises machen wir uns einfach, wobei man dann aber nicht sieht, wie man auf das angegebene Problem in Standardform gelangt. Hierzu sehe man in der Originalliteratur bei L. Eldén (1977) nach.

Da $\tilde{x}$ die Lösung von $(\tilde{\mathrm{P}})$ ist, gelten die entsprechenden Normalgleichungen, also

$$(\tilde{A}^T\tilde{A} + \tau I)\tilde{x} = \tilde{A}^T\tilde{b}.$$

Wir weisen nach, dass

$$x := V_1R^{-T}\tilde{x} + V_2U^{-1}Q_1^T(b - AV_1R^{-T}\tilde{x})$$

den Normalgleichungen zum linearen Ausgleichsproblem (P_τ) genügt. Aus der QR-Zerlegung

$$L^T = \begin{pmatrix} V_1 & V_2 \end{pmatrix} \begin{pmatrix} R \\ 0 \end{pmatrix}$$

folgt

$$L = R^T V_1^T, \qquad LV_1 = R^T, \qquad LV_2 = 0.$$

Insbesondere ist $\tilde{x} = Lx$. Entsprechend folgt aus der QR-Zerlegung

$$AV_2 = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \begin{pmatrix} U \\ 0 \end{pmatrix},$$

dass

$$AV_2 = Q_1U, \qquad Q_1^TAV_2 = U, \qquad V_2^TA^TQ_2 = 0.$$

Dann ist

$$(*) \qquad 0 = L^T[(\tilde{A}^T\tilde{A} + \tau I)\tilde{x} - \tilde{A}^T\tilde{b}] = [(\tilde{A}L)^T\tilde{A}L + \tau L^TL]x - (\tilde{A}L)^T\tilde{b}.$$

Nun beachten wir, dass

$$\tilde{A}L = Q_2^TAV_1R^{-T}L = Q_2^TAV_1V_1^T = Q_2^TA - Q_2^TAV_2V_2^T = Q_2^TA - Q_2^TQ_1UV_2^T = Q_2^TA.$$

Daher ist

$$\begin{aligned} 0 &= [(\tilde{A}L)^T\tilde{A}L + \tau L^TL]x - (\tilde{A}L)^T\tilde{b} \\ &= [A^TQ_2Q_2^TA + \tau L^TL]x - A^TQ_2Q_2^Tb \\ &= (A^TA + \tau L^TL)x - A^Tb + A^TQ_1Q_1^T(Ax - b). \end{aligned}$$

Wir zeigen, dass hier der letzte Summand verschwindet, also x den Normalgleichungen zu (P_τ) genügt und folglich die Lösung von (P_τ) ist. Denn es ist

$$\begin{aligned} A^TQ_1Q_1^T(Ax-b) &= A^TQ_1Q_1^TAV_1R^{-T}\tilde{x} + A^TQ_1Q_1^TAV_2U^{-1}Q_1^T(b-AV_1R^{-T}\tilde{x}) \\ &\quad - A^TQ_1Q_1^Tb \\ &= A^TQ_1Q_1^TAV_1R^{-T}\tilde{x} + A^TQ_1\underbrace{Q_1^TQ_1UU^{-1}}_{=I}Q_1^T(b-AV_1R^{-T}\tilde{x}) \\ &\quad - A^TQ_1Q_1^Tb \\ &= A^TQ_1Q_1^TAV_1R^{-T}\tilde{x} + A^TQ_1Q_1^T(b-AV_1R^{-T}\tilde{x}) \\ &\quad - A^TQ_1Q_1^Tb \\ &= 0. \end{aligned}$$

Damit ist die Behauptung schließlich bewiesen. □

4. Gegeben sei das quadratisch restringierte lineare Ausgleichsproblem

$$\text{(P)} \qquad \text{Minimiere} \quad \|Ax-b\|_2 \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : \|Cx-d\|_2 \le \gamma\},$$

wobei $A \in \mathbb{R}^{m\times n}$ mit $m \ge n$, $C \in \mathbb{R}^{p\times n}$ und $\gamma > 0$. Man zeige, dass die Voraussetzung

$$\text{Rang} \begin{pmatrix} A \\ C \end{pmatrix} = n$$

notwendig für die Eindeutigkeit einer Lösung von (P) ist.

Beweis: Ist

$$\text{Rang} \begin{pmatrix} A \\ C \end{pmatrix} < n,$$

so existiert ein $u \in \mathbb{R}^n \setminus \{0\}$ mit $Au = 0$ und $Cu = 0$. Offensichtlich ist mit einer Lösung x^* von (P) auch $x^* + tu$ für alle $t \in \mathbb{R}$ eine Lösung von (P). □

5. Gegeben sei das quadratisch restringierte lineare Ausgleichsproblem

$$\text{(P)} \qquad \text{Minimiere} \quad \|Ax-b\|_2 \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : \|Cx-d\|_2 \le \gamma\},$$

wobei $A \in \mathbb{R}^{m\times n}$ mit $m \ge n$, $C \in \mathbb{R}^{p\times n}$ und $\gamma > 0$. Sei S die Menge der Lösungen des unrestringierten linearen Ausgleichsproblems, $\|Ax-b\|_2$ auf dem $\mathbb{R}^n$ zu minimieren. Man zeige, dass die Aufgabe

$$\text{Minimiere} \quad \|Cx-d\|_2, \quad x \in S$$

mindestens eine Lösung $x_{A,C} \in S$ besitzt.

Beweis: Wir wissen, dass S die Menge der Lösungen der Normalgleichunge ist, so dass

$$S = \{x \in \mathbb{R}^n : A^T(Ax-b) = 0\}$$

und wollen zeigen, dass das lineare Gleichungssystem

$$(*) \qquad \begin{pmatrix} C^TC & A^TA \\ A^TA & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} C^Td \\ A^Tb \end{pmatrix}$$

eine Lösung $(x^*, y^*) \in \mathbb{R}^{n\times n}$ besitzt. Wegen

$$\text{Bild} \begin{pmatrix} C^TC & A^TA \\ A^TA & 0 \end{pmatrix} = \text{Kern} \begin{pmatrix} C^TC & A^TA \\ A^TA & 0 \end{pmatrix}^{\perp}$$

ist zu zeigen:

$$\begin{pmatrix} C^TC & A^TA \\ A^TA & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{impliziert} \quad \begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} C^Td \\ A^Tb \end{pmatrix} = 0.$$

Dies ist aber einfach einzusehen. Denn zunächst ist (zweite Gleichung nach Multiplikation mit x^T von links) $Ax = 0$, dann (erste Gleichung nach Multiplikation mit x^T von links) $Cx = 0$, schließlich $Ay = 0$ (Multiplikation der ersten Gleichung mit y^T von links), woraus dann schließlich die Behauptung folgt. Wir wollen uns überlegen, dass die erste Komponente x^* des obigen linearen Gleichungssystems $(*)$ eine Lösung der Aufgabe ist, $\|Cx - d\|_2$ auf S zu minimieren. Zunächst ist $x^* \in S$, wie die zweite Gleichung in $(*)$ zeigt. Sei nun $x \in S$ beliebig. Dann ist

$$\begin{aligned} \tfrac{1}{2}\|Cx - d\|_2^2 - \tfrac{1}{2}\|Cx^* - d\|_2^2 &\geq [C^T(Cx^* - d)]^T(x - x^*) \\ &= -(A^TAy^*)^T(x - x^*) \\ &= -(y^*)^TA^TA(x - x^*) \\ &= 0, \end{aligned}$$

da x und x^* den Normalgleichungen genügen. Damit ist die Behauptung bewiesen. □

6. Seien $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$, $b \in \mathbb{R}^m$ und $\gamma > 0$ gegeben. Man definiere $\phi\colon [0, \infty) \longrightarrow \mathbb{R}$ durch

$$\phi(\lambda) := \begin{cases} \|A^+b\|_2, & \text{falls} \quad \lambda = 0, \\ \|(A^TA + \lambda I)^{-1}A^Tb\|_2, & \text{falls} \quad \lambda > 0 \end{cases}$$

und anschließend

$$h(\lambda) := \frac{1}{\gamma} - \frac{1}{\phi(\lambda)}.$$

Man zeige, dass h auf $[0, \infty)$ monoton fallend und konvex ist.

Beweis: Ist $A = U\Sigma V^T$ eine reduzierte Singulärwertzerlegung von A und ersetzt man b durch U^Tb, so erhält man

$$\phi(\lambda) = \left[\sum_{i=1}^{r}\left(\frac{\sigma_i b_i}{\sigma_i^2 + \lambda}\right)^2\right]^{1/2}.$$

Es ist

$$h'(\lambda) = \frac{\phi'(\lambda)}{\phi(\lambda)^2} < 0,$$

da

$$\phi'(\lambda) = -\frac{1}{\phi(\lambda)}\sum_{i=1}^{r}\frac{(\sigma_i b_i)^2}{(\sigma_i^2 + \lambda)^3} < 0.$$

Also ist auch $h(\cdot)$ auf $[0, \infty)$ monoton fallend. In der folgenden Abschätzung (siehe Seite 103) nutzen wir aus, dass

$$\phi''(\lambda) \geq \frac{2}{\phi(\lambda)}\sum_{i=1}^{r}\frac{(\sigma_i b_i)^2}{(\sigma_i^2 + \lambda)^4}.$$

Daher ist

$$\begin{aligned} h''(\lambda) &= \frac{\phi''(\lambda)\phi(\lambda)^2 - 2\phi'(\lambda)^2\phi(\lambda)}{\phi(\lambda)^4} \\ &= \frac{\phi''(\lambda)}{\phi(\lambda)^2} - 2\frac{\phi'(\lambda)^2}{\phi(\lambda)^3} \\ &\geq \frac{2}{\phi(\lambda)^3}\sum_{i=1}^{r}\frac{(\sigma_i b_i)^2}{(\sigma_i^2+\lambda)^4} - \frac{2}{\phi(\lambda)^5}\left(\sum_{i=1}^{r}\frac{(\sigma_i b_i)^2}{(\sigma_i^2+\lambda)^3}\right)^2 \\ &= \frac{2}{\phi(\lambda)^5}\left[\sum_{i=1}^{r}\frac{(\sigma_i b_i)^2}{(\sigma_i^2+\lambda)^2}\sum_{i=1}^{r}\frac{(\sigma_i b_i)^2}{(\sigma_i^2+\lambda)^4} - \left(\sum_{i=1}^{r}\frac{(\sigma_i b_i)^2}{(\sigma_i^2+\lambda)^3}\right)^2\right. \\ &\geq 0 \end{aligned}$$

wegen der Cauchy-Schwarzschen Ungleichung. Daher ist auch $h(\cdot)$ auf $[0,\infty)$ konvex. Insgesamt sind die behaupteten Aussagen bewiesen. □

7. Man betrachte das quadratisch restringierte Ausgleichsproblem

(P) Minimiere $\|Ax-b\|_2$ auf $M := \{x \in \mathbb{R}^n : \|Cx-d\|_2 \leq \gamma\}$.

Hierbei seien $A \in \mathbb{R}^{m\times n}$, $C \in \mathbb{R}^{p\times n}$ mit $p \leq n \leq m$ gegeben, ferner seien $b \in \mathbb{R}^m$, $d \in \mathbb{R}^p$ und $\gamma > 0$. Ist dann Rang $(C) = p$, so besitzt (P) eine Lösung.

Beweis: Wegen des ersten Teils von Satz 6.2 besitzt (P) genau dann eine Lösung, wenn $\min_{x\in\mathbb{R}^n}\|Cx-d\|_2 \leq \gamma$. Nun besitzt das (unterbestimmte) lineare Gleichungssystem $Cx = d$ sogar eine Lösung, d. h. es ist $\min_{x\in\mathbb{R}^n}\|Cx-d\|_2 = 0$. Denn wäre Bild $(C) =$ Kern $(C^T)^\perp$ ein echter Teilraum des $\mathbb{R}^p$, so wäre Kern $(C^T) \neq \{0\}$, was natürlich ein Widerspruch zur Rangvoraussetzung ist. □

8. Gegeben sei das durch eine lineare Gleichung restringierte lineare Ausgleichsproblem

(P) Minimiere $\|Ax-b\|_2$ auf $M := \{x \in \mathbb{R}^n : Bx = d\}$.

Hierbei seien $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{p\times n}$ mit $p \leq n \leq m$ gegeben. Man zeige:

(a) Ist $M \neq \emptyset$, also (P) zulässig, so besitzt (P) eine Lösung.

(b) Ist

$$\text{Rang}\,(B) = p, \qquad \text{Rang}\begin{pmatrix} A \\ B \end{pmatrix} = n,$$

so ist (P) eindeutig lösbar.

(c) Man zeige, dass unter der die eindeutige Lösbarkeit von (P) garantierenden Rangvoraussetzung im letzten Aufgabenteil die Lösung von (P) mit Hilfe einer verallgemeinerten Singulärwertzerlegung zu (A, B) berechnet werden kann.

Beweis: Die Existenzaussage im ersten Teil kann man sicher auf verschiedene Weise zeigen. Wir schlagen den folgenden Beweis vor. Mit einem $x_0 \in M$ ist $M = x_0 +$ Kern (B). Daher ist (P) äquivalent zu

$$\text{Minimiere} \quad \|Az + Ax_0 - b\|_2, \quad z \in \text{Kern}\,(B).$$

Dies ist aber im Prinzip ein "normales" lineares Ausgleichsproblem, wenn man eine Nullraumbasis einführt für B einführt. Z. B. ist $\text{Kern}\,(B) = Cy$ mit einer Matrix $C \in \mathbb{R}^{n\times q}$, deren Spalten eine Basis von $\text{Kern}\,(B)$ bilden. Folglich ist (P) äquivalent dem linearen Ausgleichsproblem

$$\text{Minimiere} \quad \|ACy + Ax_0 - b\|_2, \quad y \in \mathbb{R}^q.$$

Da lineare Ausgleichsprobleme lösbar sind, ist es auch (P).

Für den Beweis der Eindeutigkeitsaussage beachten wir zunächst, dass aus $\text{Rang}\,(B) = p$ die Zulässigkeit von (P) und damit wegen des ersten Teils der Aufgabe die Lösbarkeit von (P) folgt. Danach benutzen wir die Lagrangesche Multiplikatorenregel (angewandt auf die Zielfunktion $f(x) := \frac{1}{2}\|Ax - b\|_2^2$) und schließen, dass es zu einer Lösung $x \in M$ ein $y \in \mathbb{R}^p$ mit $A^T(Ax - b) + B^T y = 0$ gibt. Wir zeigen unter der angegebenen Rangvoraussetzung die eindeutige Lösbarkeit des linearen Gleichungssystems

$$\begin{pmatrix} A^TA & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} A^Tb \\ d \end{pmatrix}$$

(und damit die eindeutige Lösbarkeit von (P)), indem wir nachweisen, dass der Kern der Koeffizientenmatrix trivial ist. Sei also

$$A^TAu + B^Tv = 0, \qquad Bu = 0.$$

Multiplikation der ersten Gleichung von links mit u^T liefert unter Benutzung der zweiten Gleichung $u^TA^TAu = 0$ bzw. $Au = 0$. Aus der ersten Gleichung folgt $B^Tv = 0$, wegen $\text{Rang}\,(B) = p$ ist $v = 0$. Aus

$$\begin{pmatrix} A \\ B \end{pmatrix} u = 0, \qquad \text{Rang} \begin{pmatrix} A \\ B \end{pmatrix} = n$$

folgt auch $u = 0$. Damit ist die Eindeutigkeit bewiesen.

Nun nehmen wir an, wir hätten eine verallgemeinerte Singulärwertzerlegung von (A, B) berechnet. Es sei also

$$U_A^TA = \begin{pmatrix} D_A \\ 0 \end{pmatrix} Z, \qquad U_B^TB = \begin{pmatrix} D_B & 0 \end{pmatrix} Z,$$

wobei $U_A \in \mathbb{R}^{m\times m}$, $U_B \in \mathbb{R}^{p\times p}$ orthogonal und $Z \in \mathbb{R}^{n\times n}$ nichtsingulär ist. Ferner ist

$$D_A = \text{diag}\,(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^{n\times n}, \qquad D_B = \text{diag}\,(\beta_1, \ldots, \beta_p) \in \mathbb{R}^{p\times p}$$

mit

$$0 \le \alpha_1 \le \cdots \le \alpha_n \le 1, \qquad 1 \ge \beta_1 \ge \cdots \ge \beta_p > 0$$

und

$$\alpha_i^2 + \beta_i^2 = 1 \quad (i = 1, \ldots, p), \qquad \alpha_i = 1 \quad (i = p+1, \ldots, n).$$

Wieder machen wir die Variablentransformation $y = Zx$ und setzen $\tilde{b} := U_A^Tb$, $\tilde{d} := U_C^Td$. Dann ist das Problem (P) äquivalent zu

$$\text{Minimiere} \quad \left\| \begin{pmatrix} D_A \\ 0 \end{pmatrix} y - \tilde{b} \right\|_2 \quad \text{unter der Nebembedingung} \quad \begin{pmatrix} D_B & 0 \end{pmatrix} y = \tilde{d}.$$

Für eine Lösung y dieser Aufgabe folgt aus der Nebenbedingung, dass

$$y_i = \frac{\tilde{d}_i}{\beta_i}, \qquad i = 1, \ldots, p.$$

Einsetzen in die Zielfunktion liefert unter Berücksichtigung von $\alpha_i = 1$, $i = p+1, \ldots, n$, dass $y_i = \tilde{b}_i$, $i = p+1, \ldots, n$. Insgesamt ist also mit $W = (\ w_1 \ \cdots \ w_n\ ) = Z^{-1}$ die Lösung von (P) durch

$$x^* := \sum_{i=1}^{p} \frac{\tilde{d}_i}{\beta_i} w_i + \sum_{i=p+1}^{n} \tilde{b}_i w_i$$

gegeben. □

9. Gegeben seien die Matrizen $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{p\times n}$ mit $p \le n \le m$. Es wird vorausgesetzt, dass

$$\text{Rang}\,(B) = p, \qquad \text{Rang}\begin{pmatrix} A \\ B \end{pmatrix} = n.$$

Mit einem $\gamma > 0$ und vorgegebenen Vektoren $b \in \mathbb{R}^m$, $d \in \mathbb{R}^p$ betrachte man das lineare Ausgleichsproblem

$$(\mathrm{P}_\gamma) \qquad \text{Minimiere} \quad \left\| \begin{pmatrix} A \\ \gamma B \end{pmatrix} - \begin{pmatrix} b \\ \gamma d \end{pmatrix} \right\|_2, \quad x \in \mathbb{R}^n.$$

Man zeige:

(a) Die Aufgabe (P_γ) besitzt eine eindeutige Lösung x_γ.

(b) Der Limes $x^* = \lim_{\gamma\to\infty} x_\gamma$ existiert und ist die (nach Aufgabe 8b) eindeutige Lösung von

$$(\mathrm{P}) \qquad \text{Minimiere} \quad \|Ax - b\|_2 \quad \text{auf} \quad M := \{x \in \mathbb{R}^n : Bx = d\}.$$

Beweis: Wegen

$$\text{Rang}\begin{pmatrix} A \\ B \end{pmatrix} = n$$

hat die Koeffizientenmatrix in (P_γ) ebenfalls vollen Rang, so dass (P_γ) eine eindeutige Lösung x_γ besitzt. Diese ist als Lösung der zugehörigen Normalgleichungen

$$(A^TA + \gamma^2 B^TB)x = A^Tb + \gamma^2 B^Td$$

charakterisiert. Um diese zu lösen, gehen wir von einer verallgemeinerten Singulärwertzerlegung zu (A, B) aus. Es sei also

$$U_A^TA = \begin{pmatrix} D_A \\ 0 \end{pmatrix} Z, \qquad U_B^TB = (\ D_B \quad 0\ )Z,$$

wobei $U_A \in \mathbb{R}^{m\times m}$, $U_B \in \mathbb{R}^{p\times p}$ orthogonal und $Z \in \mathbb{R}^{n\times n}$ nichtsingulär ist. Ferner ist

$$D_A = \text{diag}\,(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^{n\times n}, \qquad D_B = \text{diag}\,(\beta_1, \ldots, \beta_p) \in \mathbb{R}^{p\times p}$$

mit

$$0 \le \alpha_1 \le \cdots \le \alpha_n \le 1, \qquad 1 \ge \beta_1 \ge \cdots \ge \beta_p > 0$$

und

$$\alpha_i^2 + \beta_i^2 = 1 \quad (i = 1, \ldots, p), \qquad \alpha_i = 1 \quad (i = p+1, \ldots, n).$$

Mit $y_\gamma := Zx_\gamma$ erhält man nach Einsetzen für y_γ das lineare Gleichungssystem

$$\left[D_A^2 + \gamma^2 \begin{pmatrix} D_B^2 & 0 \\ 0 & 0 \end{pmatrix} \right] y_\gamma = \begin{pmatrix} D_A & 0 \end{pmatrix} \tilde{b} + \gamma^2 \begin{pmatrix} D_B \\ 0 \end{pmatrix} \tilde{d},$$

wobei wieder

$$\tilde{b} := U_A^T b, \qquad \tilde{d} := U_B^T d$$

gesetzt wurde. Hieraus liest man die Lösung y_γ ab, es ist nämlich

$$(y_\gamma)_i = \begin{cases} \dfrac{\alpha_i \tilde{b}_i + \gamma^2 \beta_i \tilde{d}_i}{\alpha_i^2 + \gamma^2 \beta_i^2}, & i = 1, \ldots, p, \\ \tilde{b}_i, & i = p+1, \ldots, n. \end{cases}$$

Definiert man also wieder $W := Z^{-1} = \begin{pmatrix} w_1 & \cdots & w_n \end{pmatrix}$, so ist

$$x_\gamma = \sum_{i=1}^{p} \frac{\alpha_i \tilde{b}_i + \gamma^2 \beta_i \tilde{d}_i}{\alpha_i^2 + \gamma^2 \beta_i^2} w_i + \sum_{i=p+1}^{n} \tilde{b}_i w_i.$$

Offenbar ist

$$\lim_{\gamma \to \infty} x_\gamma = \sum_{i=1}^{p} \frac{\tilde{d}_i}{\beta_i} w_i + \sum_{i=p+1}^{n} \tilde{b}_i w_i =: x^*.$$

Nach der in Aufgabe 8c hergeleiteten Dastellung ist x_∞ die Lösung von (P). Natürlich kann auch noch etwas zur Konvergenzgeschwindigkeit ausgesagt werden. Es ist

$$x_\gamma - x^* = \sum_{i=1}^{p} \frac{\alpha_i(\beta_i \tilde{b}_i - \alpha_i \tilde{d}_i)}{\beta_i(\alpha_i^2 + \gamma^2 \beta_i^2)} w_i.$$

Hier braucht natürlich nur über die i mit $\alpha_i \neq 0$ summiert zu werden. □

10. Seien

$$A := \begin{pmatrix} 22 & 10 & 2 & 3 & 7 \\ 14 & 7 & 10 & 0 & 8 \\ -1 & 13 & -1 & -11 & 3 \\ -3 & -2 & 13 & -2 & 4 \\ 9 & 8 & 1 & -2 & 4 \\ 9 & 1 & -7 & 5 & -1 \\ 2 & -6 & 6 & 5 & 1 \\ 4 & 5 & 0 & -2 & 2 \end{pmatrix}, \qquad b := \begin{pmatrix} -1 \\ 2 \\ 1 \\ 4 \\ 0 \\ -3 \\ 1 \\ 0 \end{pmatrix}$$

gegeben. Für $\gamma := 0.05, 0.15, 0.25$ berechne man eine Lösung des quadratisch restringierten linearen Ausgleichsproblems

(P) Minimiere $\|Ax - b\|_2$ unter der Nebenbedingung $\|x\|_2 \leq \gamma$

in Standardform.

Ergebnis: Wir berechnen eine reduzierte Singulärwertzerlegung $A = \hat{U}\hat{\Sigma}V^T$, setzen anschließend $b := \hat{U}^T b$ und definieren die Funktion $\phi\colon [0,\infty) \longrightarrow \mathbb{R}$ durch

$$\phi(\lambda) := \left[\sum_{i=1}^{r}\left(\frac{\sigma_i b_i}{\sigma_i^2+\lambda}\right)^2\right]^{1/2}.$$

Ist $\phi(0) > \gamma$ (das ist bei unseren Daten der Fall), so ist eine Lösung $\lambda^* > 0$ von $\phi(\lambda) = \gamma$ zu bestimmen. Ist diese gefunden, so ist

$$x^* := \sum_{i=1}^{r} \frac{\sigma_i b_i}{\sigma_i^2+\lambda^*} v_i$$

Lösung von (P). Mit

$$\phi'(\lambda) = -\frac{1}{\phi(\lambda)} \sum_{i=1}^{r} \frac{(\sigma_i b_i)^2}{(\sigma_i^2+\lambda)^3}$$

haben wir mit dem Startwert $\lambda_0 := 0$ das Newton-Verfahren sowohl auf

$$f(\lambda) := \phi(\lambda) - \gamma$$

als auch auf

$$h(\lambda) := \frac{1}{\gamma} - \frac{1}{\phi(\lambda)}$$

angewandt. Die entsprechenden Iterationsvorschriften sind

$$\lambda_{k+1} := \lambda_k - \frac{\phi(\lambda_k)-\gamma}{\phi'(\lambda_k)}$$

bzw.

$$\lambda_{k+1} := \lambda_k - \left[\frac{\phi(\lambda_k)-\gamma}{\gamma}\right]\frac{\phi(\lambda_k)}{\phi'(\lambda_k)}.$$

Im ersten Fall, also der "normalen" Newton-Iteration, erhalten wir ausgehend von $\lambda_0 := 0$ die folgenden Ergebnisse:

	$\gamma = 0.05$	$\gamma = 0.15$	$\gamma = 0.25$
λ_0	0.00000000000	0.00000000000	0.00000000000
λ_1	317.48924898936	184.46774696807	51.44624494678
λ_2	797.02017140581	315.65277348977	59.26215339073
λ_3	1348.90507186384	352.91248057705	59.40496071407
λ_3	1727.31033318842	355.00240058945	59.40500673763
λ_4	1827.98422551868	355.00834451491	59.40500673763
λ_5	1833.01357249732	355.00834456272	59.40500673763
λ_6	1833.02503362891	355.00834456272	59.40500673763
λ_7	1833.02503368816	355.00834456272	59.40500673763

Mit dem zweiten Verfahren erhalten wir erstaunlicherweise in *einem* Schritt, wieder von $\lambda_0 := 0$ ausgehend, in Rechengenauigkeit die Lösung. Als Lösung der angegebenen restringierten linearen Ausgleichsprobleme erhalten wir:

	$\gamma = 0.05$	$\gamma = 0.15$	$\gamma = 0.25$
x^*	−0.01443375672974	−0.04330127018922	−0.07216878364870
	0.00000000000000	0.00000000000000	0.00000000000000
	0.04330127018922	0.12990381056767	0.21650635094611
	−0.01443375672974	−0.04330127018922	−0.07216878364870
	0.01443375672974	0.04330127018922	0.07216878364870

Die Rechnung wurde mit MATLAB gemacht. □

11. Man betrachte die Aufgabe

(P) $$\text{Minimiere} \quad \|Ax-b\|_2^2 + \tau \|Lx\|_2^2, \quad x \in \mathbb{R}^n.$$

Hierbei seien $A \in \mathbb{R}^{m\times n}$, $L \in \mathbb{R}^{p\times n}$ mit $p \le n \le m$ gegeben. Ferner sei

$$\text{Rang}\,(L) = p, \qquad \text{Rang} \begin{pmatrix} A \\ L \end{pmatrix} = n.$$

Schließlich seien $\tau > 0$ und $b \in \mathbb{R}^m$ gegeben. Mit Hilfe einer verallgemeinerten Singulärwertzerlegung des Paares (A, B) berechne man die Lösung von (P).

Lösung: Sei

$$U_A^T A = \begin{pmatrix} D_A \\ 0 \end{pmatrix} Z, \qquad U_L^T L = \begin{pmatrix} D_L & 0 \end{pmatrix} Z$$

eine verallgemeinerte Singulärwertzerlegung mit orthogonalen Matrizen $U_A \in \mathbb{R}^{m\times m}$, $U_L \in \mathbb{R}^{p\times p}$ und einer nichtsingulären Matrix $Z \in \mathbb{R}^{n\times n}$. Ferner sind

$$D_A = \text{diag}\,(\alpha_1, \ldots, \alpha_n), \qquad D_L = \text{diag}\,(\lambda_1, \ldots, \lambda_p)$$

Diagonalmatrizen mit

$$0 \le \alpha_1 \le \cdots \le \alpha_n \le 1, \qquad 1 \ge \lambda_1 \ge \cdots \ge \lambda_p > 0$$

und

$$\alpha_i^2 + \lambda_i^2 = 1 \quad (i = 1, \ldots, p), \qquad \alpha_i = 1 \quad (i = p+1, \ldots, n).$$

Da man (P) auch als lineares Ausgleichsproblem in der Form

$$\text{Minimiere} \quad \left\| \begin{pmatrix} A \\ \sqrt{\tau} L \end{pmatrix} x - \begin{pmatrix} b \\ 0 \end{pmatrix} \right\|_2, \quad x \in \mathbb{R}^n$$

schreiben kann, ist die Lösung von (P) als die Lösung der verallgemeinerten Normalgleichung

$$(A^T A + \tau L^T L)x = A^T b$$

charakterisiert. Einsetzen ergibt nach Multiplikation mit Z^{-T} von links die äquivalente Gleichung

$$\left[D_A^2 + \tau \begin{pmatrix} D_L^2 & 0 \\ 0 & 0 \end{pmatrix} \right] Zx = \begin{pmatrix} D_A & 0 \end{pmatrix} U_A^T b.$$

Mit $\tilde{b} := U_A^T b$ erhält man also $y = Zx$ durch

$$y_i = \frac{\alpha_i \tilde{b}_i}{\alpha_i^2 + \tau \lambda_i^2} \quad (i = 1, \ldots, p), \qquad y_i = \tilde{b}_i \quad (i = p+1, \ldots, n).$$

Wenn also eine verallgemeinerte Singulärwertzerlegung zum Paar (A, B) vorliegt, so ist die Lösung von (P) kein Problem. □

6.3 Aufgaben in Kapitel 4

6.3.1 Aufgaben in Abschnitt 4.1

1. Die Matrix $A \in ***^{n\times n}$ habe die einfachen Eigenwerte $\lambda_1, \ldots, \lambda_n$, sei also insbesondere diagonalisierbar. Die zugehörigen Eigenvektoren $x_1, \ldots, x_n$ seien durch $\|x_i\|_2 = 1$, $i = 1, \ldots, n$, normiert. Ferner seien $y_1, \ldots, y_n$ zugehörige *linke Eigenvektoren*, d. h. $y_i \neq 0$ und $y_i^H A = \lambda_i y_i^H$ (oder $A^H y_i = \overline{\lambda_i} y_i$), $i = 1, \ldots, n$, ebenfalls durch $\|y_i\|_2 = 1$, $i = 1, \ldots, n$, normiert. Weiter wird vorausgesetzt, dass $y_i^H x_i \neq 0$, $i = 1, \ldots, n$. Man zeige:

   (a) Es ist $y_i^H x_j = 0$ für $i \neq j$.

   (b) Definiert man
   $$P := (\ x_1 \ \cdots \ x_n\ ),$$
   so ist
   $$P^{-1} = \begin{pmatrix} y_1^H/y_1^H x_1 \\ \vdots \\ y_n^H/y_n^H x_n \end{pmatrix}.$$

   (c) Ist $\delta A \in ***^{n\times n}$ und $\lambda \in ***$ ein Eigenwert von $A + \delta A$, so existiert ein $i \in \{1, \ldots, n\}$ mit
   $$|\lambda - \lambda_i| \leq n \frac{\|\delta A\|_2}{|y_i^H x_i|}.$$

Beweis: Sei $i \neq j$. Aus $y_i^H A = \lambda_i y_i^H$ und $Ax_j = \lambda_j x_j$ erhält man durch Multiplikatom mit x_j von rechts bzw. y_i^H von links sowie anschließende Subtraktion, dass

$$0 = y_i^H A x_j - y_i^H A x_j = \underbrace{(\lambda_i - \lambda_j)}_{\neq 0} y_i^H x_j,$$

woraus die erste Behauptung folgt. Wegen $y_i^H x_j / y_i^H x_i = \delta_{ij}$ hat P^{-1} die angegebene Form. Auf die Matrix $P^{-1}(A + \delta A)P$ wenden wir den Satz von Gerschgorin an. Es ist

$$P^{-1}(A + \delta A)P = \Lambda + F \qquad \text{mit} \quad \Lambda := \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \quad F := P^{-1}(\delta A)P.$$

Ist nun λ ein Eigenwert von $A + \delta A$ und damit auch von $\Lambda + F$, so sagt der Satz von Gerschgorin aus, dass es ein $i \in \{1, \ldots, n\}$ mit

$$|\lambda - (\lambda_i + f_{ii})| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |f_{ij}|$$

gibt. Folglich ist

$$\begin{aligned} |\lambda - \lambda_i| &\leq \sum_{j=1}^{n} |f_{ij}| \\ &= \|F^T e_i\|_1 \\ &\leq n^{1/2} \|F^T e_i\|_2 \\ &= n^{1/2} \|P^T (\delta A)^T P^{-T} e_i\|_2 \end{aligned}$$

$$\begin{aligned} &\leq n^{1/2} \|P^T\|_2 \|(\delta A)^T\|_2 \|P^{-T} e_i\|_2 \\ &= n^{1/2} \|P\|_2 \|\delta A\|_2 \frac{1}{|y_i^H x_i|} \\ &\leq \frac{n}{|y_i^H x_i|} \|\delta A\|_2. \end{aligned}$$

Hierbei haben wir im letzten Schritt ausgenutzt, dass

$$\|P\|_2 \leq \|P\|_F = \left(\sum_{i=1}^{n} \|x_i\|_2^2 \right)^{1/2} = n^{1/2}.$$

Insgesamt ist die Behauptung bewiesen. □

2. Sei $A \in ***^{n\times n}$ und $Q^H A Q = \Lambda + N$ mit unitärem $Q \in ***^{n\times n}$, der Diagonalmatrix $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ und der strikten oberen Dreiecksmatrix $N \in ***^{n\times n}$. Ist $\delta A \in ***^{n\times n}$, $\lambda \in ***$ ein Eigenwert von $A + \delta A$, so ist

$$\min_{i=1,\ldots,n} |\lambda - \lambda_i| \leq \max(\theta, \theta^{1/n}),$$

wobei

$$\theta := \|\delta A\|_2 \sum_{k=0}^{n-1} \|N\|_2^k.$$

Beweis: Zunächst überlegen wir uns:

- Ist $N \in ***^{n\times n}$ eine strikte obere Dreiecksmatrix, so ist $N^n = 0$. Ferner ist

$$(I - N)^{-1} = \sum_{k=0}^{n-1} N^k.$$

Denn: Wir zeigen durch vollständige Induktion nach k, dass

$$(N^k)_{ij} = 0, \qquad i \geq j - k + 1.$$

Dies ist für $k = 1$ richtig, da N nach Voraussetzung eine strikte obere Dreiecksmatrix ist. Angenommen, es sei für k richtig. Für $i \geq j - k$ ist dann

$$\begin{aligned} (N^{k+1})_{ij} &= \sum_{r=1}^{n} (N^k)_{ir} (N)_{rj} \\ &= \sum_{r=i+k}^{n} (N^k)_{ir} (N)_{rj} \\ &\qquad \text{(wegen der Induktionsannahme ist } (N^k)_{ir} = 0 \text{ für } r \leq i + k - 1) \\ &= 0 \\ &\qquad \text{(wegen } (N)_{rj} = 0 \text{ für } r \geq i + k \geq j.) \end{aligned}$$

Also ist $N^n = 0$ und die erste Behauptung ist bewiesen. Die zweite ist dann wegen

$$(I - N) \sum_{k=0}^{n-1} N^k = \sum_{k=0}^{n-1} N^k - \sum_{k=1}^{n} N^k = \sum_{k=0}^{n-1} N^k - \sum_{k=1}^{n-1} N^k = I$$

trivial.

Nun zum eigentlichen Beweis. Sei

$$\delta := \min_{i=1,\ldots,n} |\lambda - \lambda_i|.$$

Ist $\delta = 0$, so ist λ ein Eigenwert von A und die Aussage trivialerweise richtig. Wir können also annehmen, dass $\delta > 0$ bzw. λ kein Eigenwert von A ist. Ist x ein Eigenvektor zum Eigenwert λ von $A + \delta A$. Dann ist $(\lambda I - A)^{-1}(\delta A)x = x$ und folglich

$$\begin{aligned} 1 &\leq \|(\lambda I - A)^{-1}(\delta A)\|_2 \\ &\leq \|(\lambda I - A)^{-1}\|_2 \, \|\delta A\|_2 \\ &= \|Q^H[(\lambda I - \Lambda) - N]^{-1}Q\|_2 \, \|\delta A\|_2 \\ &= \|[(\lambda I - \Lambda) - N]^{-1}\|_2 \, \|\delta A\|_2. \end{aligned}$$

Da $(\lambda I - \Lambda)$ eine Diagonalmatrix ist, ist mit N auch $(\lambda I - \Lambda)^{-1}N$ eine strikte obere Dreiecksmatrix und daher

$$[(\lambda I - \Lambda) - N]^{-1} = [I - (\lambda I - \Lambda)^{-1}N]^{-1}(\lambda I - \Lambda)^{-1} = \sum_{k=0}^{n-1} [(\lambda I - \Lambda)^{-1}N]^k (\lambda I - \Lambda)^{-1}.$$

Daher können wir die obige Gleichungs-Ungleichungskette fortsetzen und erhalten unter Berücksichtigung von

$$\|(\lambda I - \Lambda)^{-1}\|_2 = \max_{i=1,\ldots,n} \frac{1}{|\lambda - \lambda_i|} = \frac{1}{\delta},$$

dass

$$1 \leq \frac{\|\delta A\|_2}{\delta} \max\left(1, \frac{1}{\delta^{n-1}}\right) \sum_{k=0}^{n-1} \|N\|_2^k.$$

Hieraus folgt: Ist $\delta \geq 1$, so ist $\delta \leq \theta$. Ist dagegen $\delta < 1$, so ist $\delta^n \leq \theta$ und folglich $\delta \leq \theta^{1/n}$. Insgesamt ist die Aussage bewiesen. □

3. Man wende das Verfahren der Vektoriteration an, um Näherungen für den dominanten Eigenwert und zugehörigen Eigenvektor von

$$A := \begin{pmatrix} 1 & 2 & 3 & 5 \\ 2 & 4 & 1 & 6 \\ 1 & 2 & -1 & 3 \\ 2 & 0 & 1 & 3 \end{pmatrix}$$

zu berechnen. Hierbei starte man mit $x^{(0)} := (1, 1, 1, 1)^T$.

Ergebnis: Wir erhalten die folgenden Ergebnisse:

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$x_4^{(k)}$	$\lambda^{(k)}$
0	1.0000	1.0000	1.0000	1.0000	
1	0.5871	0.6939	0.2669	0.3203	8.3105
2	0.5242	0.7351	0.3196	0.2876	8.1903
3	0.5360	0.7366	0.3098	0.2727	8.0748
4	0.5325	0.7383	0.3116	0.2723	8.0772
5	0.5331	0.7384	0.3113	0.2716	8.0716
6	0.5329	0.7384	0.3114	0.2716	8.0718

Die ersten beiden betragsgrößten Eigenwerte von A sind reell und durch

$$\lambda_1 = 8.07158161596004, \qquad \lambda_2 = 1.58793382258425$$

gegeben. □

4. Mit Hilfe der inversen Iteration verbessere man die Näherung $\lambda_1 \approx 8$ für einen Eigenwert der Matrix

$$A := \begin{pmatrix} 1 & 2 & 3 & 5 \\ 2 & 4 & 1 & 6 \\ 1 & 2 & -1 & 3 \\ 2 & 0 & 1 & 3 \end{pmatrix}.$$

Man starte mit $x^{(0)} := (1,1,1,1)^T$ und berechne auch den zugehörigen Eigenvektor.

Ergebnis: Wir erhalten die folgenden Ergebnisse:

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$x_4^{(k)}$	$\lambda^{(k)}$
0	1.0000000000	1.0000000000	1.0000000000	1.0000000000	
1	0.5324620057	0.7401807002	0.3101152341	0.2691566183	8.0541798978
2	0.5329491892	0.7384150990	0.3113648059	0.2715884035	8.0717837164
3	0.5329411749	0.7384352130	0.3113559151	0.2715596332	8.0715792676
4	0.5329412791	0.7384349877	0.3113559774	0.2715599702	8.0715816431
5	0.5329412778	0.7384349902	0.3113559770	0.2715599662	8.0715816156

Die Konvergenz kann als gut bezeichnet werden. □

5. Man wende das Verfahren der orthogonalen Iteration auf die Matrix

$$A := \begin{pmatrix} -261 & 209 & -49 \\ -530 & 422 & -98 \\ -800 & 631 & -144 \end{pmatrix}$$

an, wobei man $p := 2$ und

$$X^{(0)} := \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$$

wähle.

Ergebnis: Wir erhalten die folgenden Ergebnisse:

k	$x_{11}^{(k)}$	$x_{21}^{(k)}$	$x_{31}^{(k)}$	$x_{12}^{(k)}$	$x_{22}^{(k)}$	$x_{32}^{(k)}$
0	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000
1	−0.2624	−0.5329	−0.8044	−0.5751	−0.5830	0.5739
2	−0.2655	−0.5335	−0.8030	0.5743	−0.5815	0.5763
3	−0.2665	−0.5340	−0.8024	−0.5446	−0.5803	0.5771
4	−0.2670	−0.5343	−0.8020	−0.5751	−0.5795	0.5775
5	−0.2671	−0.5344	−0.8019	−0.5756	−0.5789	0.5776
6	−0.2672	−0.5345	−0.8018	−0.5760	−0.5785	0.5776
7	−0.2672	−0.5345	−0.8018	−0.5763	−0.5782	0.5775

Die Konvergenz ist also ziemlich schlecht. Das ist auch nicht anders zu erwarten, denn die Eigenwerte von A sind $\lambda_1 = 10$, $\lambda_2 = 4$ und $\lambda_3 = 3$. Die Konvergenzrate ist $3/4$. Bemerkenswert ist, dass der erste Schritt schon eine relativ gute Näherung liefert. □

6. Sei $A \in \mathbb{R}^{n\times n}$ eine Matrix mit reellen Eigenwerten $\lambda_1 \geq \cdots \geq \lambda_n$. Man zeige:

   (a) Ist $\sigma < \frac{1}{2}(\lambda_1 + \lambda_n)$ und $\lambda_1 > \lambda_2$, so ist $\lambda_1 - \sigma$ ein dominanter Eigenwert von $A - \sigma I$.

   (b) Ist $\sigma > \frac{1}{2}(\lambda_1 + \lambda_n)$ und $\lambda_n < \lambda_{n-1}$, so ist $\lambda_n - \sigma$ ein dominanter Eigenwert von $A - \sigma I$.

   Daher kann man bei der Anwendung der Vektoriteration auf $A - \sigma I$ nur Konvergenz gegen λ_1 oder λ_n (bzw. $\lambda_1 - \sigma$ oder $\lambda_n - \sigma$) erwarten.

   Beweis: Die Matrix $A - \sigma I$ hat natürlich die Eigenwerte $\lambda_i - \sigma$, $i = 1, \ldots, n$. Wir betrachten zunächst den Fall (a) und nehmen an, es sei $\sigma < \frac{1}{2}(\lambda_1 + \lambda_n)$ und $\lambda_1 > \lambda_2$. Wir wollen zeigen, dass $\lambda_1 - \sigma$ ein dominanter Eigenwert von $A - \sigma I$ ist. Hierzu beachten wir zunächst, dass

$$\lambda_1 - \sigma > \lambda_1 - \tfrac{1}{2}(\lambda_1 + \lambda_n) = \tfrac{1}{2}(\lambda_1 - \lambda_n) > 0.$$

   Zu zeigen ist daher, dass

$$-(\lambda_1 - \sigma) < \lambda_i - \sigma < (\lambda_1 - \sigma), \qquad i = 2, \ldots, n.$$

   Die rechte Ungleichung ist wegen $\lambda_2 < \lambda_1$ klar, die linke folgt wegen

$$\lambda_i - \sigma \geq \lambda_n - \sigma > -(\lambda_1 - \sigma).$$

   Den zweiten Teil (b) beweist man praktisch genau so. □

7. Man schreibe ein Programm, das eine Matrix $A \in \mathbb{R}^{n\times n}$ mit Hilfe von $n-2$ Ähnlichkeitstransformationen mit Householder-Matrizen $P_k = \operatorname{diag}(I_k, \overline{P}_k)$ in eine obere Hessenberg-Matrix überführt. Als Output sollte man erhalten:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{1,n-1} & a_{1n} \\ a_{21} & a_{22} & \cdots & \cdots & a_{2,n-1} & a_{2n} \\ u_3^1 & a_{32} & \cdots & \cdots & a_{3,n-1} & a_{3n} \\ u_4^1 & u_4^2 & \ddots & & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ u_n^1 & u_n^2 & \cdots & u_n^{n-2} & a_{n,n-1} & a_{nn} \end{pmatrix}, \quad d = \begin{pmatrix} u_2^1 \\ u_3^2 \\ \vdots \\ u_{n-1}^{n-2} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{n-2} \end{pmatrix},$$

   wobei

$$H = \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{1,n-1} & a_{1n} \\ a_{21} & a_{22} & \cdots & \cdots & a_{2,n-1} & a_{2n} \\ & a_{32} & \cdots & \cdots & a_{3,n-1} & a_{3n} \\ & & \ddots & & \vdots & \vdots \\ & & & \ddots & \vdots & \vdots \\ & & & & a_{n,n-1} & a_{nn} \end{pmatrix}$$

   die gesuchte, orthogonal ähnliche obere Hessenbergmatrix ist und

$$\overline{P}_k = I_{n-k} - \beta_k \begin{pmatrix} u_{k+1}^k \\ \vdots \\ u_n^k \end{pmatrix} \begin{pmatrix} u_{k+1}^k \\ \vdots \\ u_n^k \end{pmatrix}^T.$$

Man teste das Programm an der Matrix

$$A := \begin{pmatrix} 1 & 2 & 3 & 5 \\ 2 & 4 & 1 & 6 \\ 1 & 2 & -1 & 3 \\ 2 & 0 & 1 & 3 \end{pmatrix}.$$

Lösung: Bei unserer Implementation wird A überschrieben mit

$$A = \begin{pmatrix} 1.0000000000 & -5.6666666667 & 1.8633410513 & 1.5546218239 \\ -3.0000000000 & 7.2222222222 & -2.8262510226 & 0.9812861158 \\ 0.5000000000 & 3.3591592129 & 0.2383904692 & 1.5401167031 \\ 1.0000000000 & 1.0000000000 & 0.2067833698 & -1.4606126915 \end{pmatrix},$$

ferner wird

$$d = \begin{pmatrix} 2.5000000000 \\ -1.0474315535 \end{pmatrix}, \qquad \beta = \begin{pmatrix} 0.2666666667 \\ 0.9536921159 \end{pmatrix}$$

ausgegeben.

8. Zu einer Matrix $A \in \mathbb{R}^{n\times n}$ sei eine orthogonale Matrix $Q \in \mathbb{R}^{n\times n}$ derart bestimmt worden, dass $Q^T AQ = R$ eine obere Dreiecksmatrix ist. In der Diagonalen von R stehen also die (notwendigerweise reellen) Eigenwerte von A. Wir nehmen an, dass diese sogar paarweise verschieden voneinander sind. Wie kann aus Q und R das vollständige System von Eigenvektoren zu A berechnet werden?

Lösung: Es genügt, die Eigenvektoren der oberen Dreiecksmatrix R zu bestimmen. Denn ist y ein Eigenvektor von R, so ist Qy ein Eigenvektor von A. Die Eigenwerte von R (und von A) sind natürlich r_{jj}, $j = 1, \ldots, n$. Für einen zu r_{jj} gehörenden Eigenvektor machen wir den Ansatz

$$y = (y_1, \ldots, y_{j-1}, 1, \underbrace{0, \ldots, 0}_{n-j})^T.$$

Dann sind $y_1, \ldots, y_{j-1}$ so zu bestimmen, dass

$$\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1,j-1} & r_{1j} \\ & r_{22} & & r_{2,j-1} & r_{2j} \\ & & \ddots & \vdots & \vdots \\ & & & r_{j-1,j-1} & r_{j-1,j} \\ & & & & r_{jj} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{j-1} \\ 1 \end{pmatrix} = r_{jj} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{j-1} \\ 1 \end{pmatrix}.$$

Dies wiederum ist äquivalent zu

$$\begin{pmatrix} r_{11} - r_{jj} & r_{12} & \cdots & r_{1,j-1} \\ & r_{22} - r_{jj} & & r_{2,j-1} \\ & & \ddots & \vdots \\ & & & r_{j-1,j-1} - r_{jj} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{j-1} \end{pmatrix} = - \begin{pmatrix} r_{1j} \\ r_{2j} \\ \vdots \\ r_{j-1,j} \end{pmatrix}.$$

Die Koeffizientenmatrix dieses linearen Gleichungssystems ist nichtsingulär, da wir vorausgesetzt haben, dass die Eigenwerte von A (bzw. die Diagonalelemente von R) paarweise voneinander verschieden sind. Daher kann das obige lineare Gleichungssystem leicht durch Rückwärtseinsetzen gelöst werden. □

9. Seien $A \in \mathbb{R}^{n\times n}$ und $z \in \mathbb{R}^n \setminus \{0\}$ gegeben. Man zeige, dass es eine orthogonale Matrix $Q \in \mathbb{R}^{n\times n}$ gibt derart, dass $Q^T AQ$ eine obere Hessenberg-Matrix und $Q^T z$ ein Vielfaches des ersten Einheitsvektors e_1 ist.

   Beweis: Man bestimme eine Householdermatrix $P_0 \in \mathbb{R}^{n\times n}$ mit $P_0 z = \alpha e_1$ und berechne $A^{(1)} := P_0 A P_0$. Die Matrix $A^{(1)}$ kann durch $n-2$ Ähnlichkeitstransformationen mit Householder-Matrizen $P_k = \text{diag}\,(I_k, \overline{P}_k)$, $k = 1, \ldots, n-2$, in obere Hessenbergform transformiert werden. Mit $Q := P_0 P_1 \cdots P_{n-2}$ ist dann $Q^T AQ$ eine obere Hessenberg-Matrix und $Qz = \alpha e_1$. □

10. Eine Matrix $A \in \mathbb{C}^{n\times n}$ heißt *normal*, wenn $A^H A = AA^H$.

    (a) Man zeige mit Hilfe des (komplexen) Schurschen Zerlegungssatzes, dass eine Matrix $A \in \mathbb{C}^{n\times n}$ genau dann unitär ähnlich einer Diagonalmatrix ist, wenn A normal ist.

    (b) Seien $\lambda_1, \ldots, \lambda_n$ die Eigenwerte einer Matrix $A \in \mathbb{C}^{n\times n}$. Man zeige, dass A genau dann normal ist, wenn $\sum_{i=1}^n |\lambda_i|^2 = \|A\|_F^2$.

    Beweis: Wir nehmen an, es sei $U^H AU = \Lambda$ mit einer unitären Matrix $U \in \mathbb{C}^{n\times n}$ und einer Diagonalmatrix $\Lambda = \text{diag}\,(\lambda_1, \ldots, \lambda_n)$. Dann ist

$$A^H A - AA^H = U^H (\underbrace{\Lambda^H \Lambda - \Lambda\Lambda^H}_{=0}) U = 0,$$

    also A normal. Umgekehrt genügt es wegen des Schurschen Zerlegungssatzes zu zeigen, dass eine normale obere Dreiecksmatrix $R \in \mathbb{C}^{n\times n}$ notwendigerweise eine Diagonalmatrix ist. Für $i = 1, \ldots, n$ ist dann

$$\begin{aligned} 0 &= (R^H R)_{ii} - (RR^H)_{ii} \\ &= \sum_{k=1}^{n} |r_{ki}|^2 - \sum_{k=1}^{n} |r_{ik}|^2 \\ &= \sum_{k=1}^{i} |r_{ki}|^2 - \sum_{k=i}^{n} |r_{ik}|^2 \\ &= \sum_{k=1}^{i-1} |r_{ki}|^2 - \sum_{k=i+1}^{n} |r_{ik}|^2. \end{aligned}$$

    Durch vollständige Induktion nach i zeigen wir nun, dass $r_{ik} = 0$, $k = i+1, \ldots, n$. Für $i = 1$ erhalten wir aus der obigen Beziehung, dass

$$0 = -\sum_{k=2}^{n} |r_{1k}|^2,$$

    das liefert den Induktionsanfang. In der Induktionsannahme wird vorausgesetzt, dass in allen Zeilen von R bis zur $(i-1)$-ten einschließlich außerhalb der Diagonalen nur verschwindende Einträge vorkommen. Dann ist

$$0 = \sum_{k=1}^{i-1} \underbrace{|r_{ki}|^2}_{=0} - \sum_{k=i+1}^{n} |r_{ik}|^2 = -\sum_{k=i+1}^{n} |r_{ik}|^2,$$

womit der Induktionsbeweis abgeschlossen und die erste Behauptung bewiesen ist.

Nun nehmen wir an, es sei $A \in \mathbb{C}^{n\times n}$ normal. Wegen des ersten Teiles der Aufgabe ist $U_H AU = \Lambda$ mit einer unitären Matrix U und der Diagonalmatrix $\Lambda = \text{diag}\,(\lambda_1, \ldots, \lambda_n)$. Da die Frobeniusnorm unter unitären Transformationen invariant ist, ist

$$\sum_{i=1}^{n} |\lambda_i|^2 = \|\Lambda\|_F^2 = \|Q^H AQ\|_F^2 = \|A\|_F^2.$$

Sei nun umgekehrt $\sum_{i=1}^{n} |\lambda_i|^2 = \|A\|_F^2$. Wegen des Schurschen Zerlegungssatzes existieren eine unitäre Matrix $U \in \mathbb{C}^{n\times n}$ und eine obere Dreiecksmatrix $R \in \mathbb{C}^{n\times n}$ mit $U^H AU = R$. In der Diagonalen von R stehen die Eigenwerte von A, es ist also $R = \Lambda + N$ mit $\Lambda = \text{diag}\,(\lambda_1, \ldots, \lambda_n)$ und einer strikten oberen Dreiecksmatrix N. Dann ist

$$\sum_{i=1}^{n} |\lambda_i|^2 = \|A\|_F^2 = \|R\|_F^2 = \sum_{i=1}^{n} |\lambda_i|^2 + \|N\|_F^2,$$

woraus $N = 0$ folgt. Also ist A einer Diagonalmatrix unitär ähnlich bzw. normal. □

6.3.2 Aufgaben in Abschnitt 4.2

1. Seien $A, B \in \mathbb{R}^{n\times n}$ symmetrisch mit Eigenwerten $\lambda_1 \geq \cdots \lambda_n$ bzw. $\mu_1 \geq \cdots \geq \mu_n$. Die Eigenwerte von $A + B$ seien $\nu_1 \geq \cdots \geq \nu_n$. Mit Hilfe des Courantschen Minimum-Maximum-Prinzips zeige man, dass

$$\lambda_j + \mu_n \leq \nu_j \leq \lambda_j + \mu_1, \qquad j = 1, \ldots, n.$$

Beweis: Sei $j \in \{1, \ldots, n\}$ fest. Sei wieder

$$\mathcal{N}_j := \{N_j \subset \mathbb{R}^n : \ N_j \text{ ist linearer Teilraum mit } \dim(N_j) = n + 1 - j\}.$$

Wegen des Courantschen Minimum-Maximum-Prinzips existiert $L_j \in \mathcal{N}_j$ mit

$$\lambda_j = \max_{0\neq x\in L_j} \frac{x^T Ax}{x^T x}.$$

Weiter ist $\mu_1 \geq x^T Bx/x^T x$ für alle $x \neq 0$. Folglich ist

$$\begin{aligned}
\lambda_j + \mu_1 &= \max_{0\neq x\in L_j} \frac{x^T Ax}{x^T x} + \mu_1 \\
&\geq \max_{0\neq x\in L_j} \frac{x^T (A+B)x}{x^T x} \\
&\geq \min_{N_j\in\mathcal{N}_j} \max_{0\neq x\in N_j} \frac{x^T (A+B)x}{x^T x} \\
&= \nu_j.
\end{aligned}$$

Ganz ähnlich kann man die erste Ungleichung beweisen. Hier wendet man das Courantsche Minimum-Maximum-Prinzip auf $A + B$ an. Es existiert ein $L_j \in \mathcal{N}_j$ mit

$$\nu_j = \max_{0\neq x\in L_j} \frac{x^T (A+B)x}{x^T x},$$

außerdem benutzt man, dass $x^T Bx/x^T x \geq \mu_n$ für alle $x \in \mathbb{R}^n \setminus \{0\}$. Folglich ist

$$\begin{aligned}\nu_j &= \max_{0 \neq x \in L_j} \frac{x^T(A+B)x}{x^T x} \\ &\geq \max_{0 \neq x \in L_j} \frac{x^T Ax}{x^T x} + \mu_n \\ &\geq \min_{N_j \in \mathcal{N}_j} \max_{0 \neq x \in N_j} \frac{x^T Ax}{x^T x} + \mu_n \\ &= \lambda_j + \mu_n.\end{aligned}$$

Damit ist die Behauptung bewiesen. □

2. Sei $A \in \mathbb{R}^{n \times n}$ symmetrisch mit Eigenwerten $\lambda_1, \ldots, \lambda_n$. Für alle $\lambda \in \mathbb{R}$ und alle $x \in \mathbb{R}^n \setminus \{0\}$ ist dann

$$\min_{j=1,\ldots,n} |\lambda - \lambda_j| \leq \frac{\|\lambda x - Ax\|_2}{\|x\|_2}.$$

Hinweis: Man setze $\delta A := (\lambda x - Ax)x^T/\|x\|_2^2$ und wende den Satz von Bauer-Fike an.

Beweis: Man definiere δA wie im Hinweis angegeben. Dann ist λ ein Eigenwert von $A + \delta A$ mit dem zugehörigen Eigenvektor x, wie man aus

$$(A + \delta A)x = Ax + (\lambda x - Ax)\frac{x^T x}{\|x\|_2^2} = \lambda x$$

erkennt. Wendet man den Satz von Bauer-Fike mit $\|\cdot\| := \|\cdot\|_2$ an und benutzt man, dass eine symmetrische Matrix mittels einer orthogonalen Ähnlichkeitstransformation diagonalisiert werden kann, so erhält man

$$\min_{j=1,\ldots,n} |\lambda - \lambda_j| \leq \|\delta A\|_2 = \frac{\|\lambda x - Ax\|_2}{\|x\|_2},$$

was zu zeigen war. □

3. Sei $A \in \mathbb{R}^{n \times n}$ symmetrisch. Für $x \in \mathbb{R}^n \setminus \{0\}$ bezeichne $\rho(x) := x^T Ax/x^T x$ den zugehörigen Rayleigh-Quotienten. Seien $\lambda_1, \ldots, \lambda_n$ die Eigenwerte von A. Dann gilt:

   (a) Für jedes $x \in \mathbb{R}^n \setminus \{0\}$ ist

$$\min_{j=1,\ldots,n} |\rho(x) - \lambda_j| \leq \left[\left(\frac{\|Ax\|_2}{\|x\|_2}\right)^2 - \rho(x)^2\right]^{1/2}.$$

   (b) Ist $x \in \mathbb{R}^n \setminus \{0\}$ und $\rho(x)$ kein Eigenwert von A, so ist

$$\min_{j=1,\ldots,n} |\rho(x) - \lambda_j| \leq \frac{\|x\|_2}{\|[A - \rho(x)I]^{-1}x\|_2}.$$

Hinweis: Man wende Aufgabe 2 an.

Beweis: Aus Aufgabe 2 erhalten wir (setze $\lambda := \rho(x)(x)$), dass

$$\min_{j=1,\ldots,n} |\rho(x) - \lambda_j| \leq \frac{\|\rho(x)x - Ax\|_2}{\|x\|_2}.$$

Nun ist

$$\left(\frac{\|\rho(x)x - Ax\|_2}{\|x\|_2}\right)^2 = \left(\frac{\|Ax\|_2}{\|x\|_2}\right)^2 - \rho(x)^2,$$

woraus die erste Behauptung folgt. Zum Beweis der zweiten Aussage wenden wir wieder Aufgabe 2 an und erhalten, dass

$$\min_{j=1,\ldots,n} |\rho(x) - \lambda_j| \leq \frac{\|(A - \rho(x)(x)I)y\|_2}{\|y\|_2}$$

für alle $y \in \mathbb{R}^n \setminus \{0\}$. Setzt man hier $y := [A - \rho(x)(x)I]^{-1}x$, so erhält man auch den Beweis der zweiten Behauptung. □

4. Sei $M \in \mathbb{R}^{n\times m}$ symmetrisch mit Eigenwerten $\mu_1 \geq \cdots \geq \mu_n$. Die Matrix $X \in \mathbb{R}^{n\times(n-1)}$ habe orthonormierte Spalten, es sei also $X^TX = I$. Bezeichnet man mit $\nu_1 \geq \cdots \geq \nu_{n-1}$ die Eigenwerte von $X^TMX \in \mathbb{R}^{(n-1)\times(n-1)}$, so ist

$$\mu_{j+1} \leq \nu_j \leq \mu_j, \qquad j = 1, \ldots, n-1.$$

Beweis: Sei $j \in \{1, \ldots, n-1\}$ fest. Wie im Courantschen Minimum-Maximum-Prinzip sei

$$\mathcal{N}_j^{(n)} := \{N_j \subset \mathbb{R}^n : \ N_j \text{ ist linearer Teilraum mit } \dim(N_j) = n + 1 - j\}.$$

Eine Anwendung des Courantschen Minimum-Maximum-Prinzips auf X^TMX liefert die Existenz von $L_j \in \mathcal{N}_j^{(n-1)}$ mit

$$\begin{aligned}
\nu_j &= \max_{0\neq y\in L_j} \frac{y^TX^TMXy}{y^Ty} \\
&= \max_{0\neq y\in L_j} \frac{(Xy)^TM(Xy)}{(Xy)^T(Xy)} \\
&= \max_{0\neq x\in X(L_j)} \frac{x^TMx}{x^Tx} \\
&\geq \min_{N_{j+1}\in\mathcal{N}_{j+1}^{(n)}} \max_{0\neq x\in N_{j+1}} \frac{x^TMx}{x^Tx} \\
&= \mu_{j+1}.
\end{aligned}$$

Hirbei haben wir ausgenutzt, dass $X(L_j) \in \mathcal{N}_{j+1}^{(n)}$. Wendet man die gerade eben bewiesene Aussage auf $-M$ statt M an, so erhält man $\nu_j \leq \mu_j$, $j = 1, \ldots, n-1$. □

5. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch, $a \in \mathbb{R}^n$ und $M := A + \alpha\, aa^T$ mit $\alpha \neq 0$. Seien $\lambda_1 \geq \cdots \geq \lambda_n$ die Eigenwerte von A und $\mu_1 \geq \cdots \geq \mu_n$ die Eigenwerte von M. Dann gilt:

   (a) Ist $\alpha > 0$, so ist $\lambda_n \leq \mu_n \leq \lambda_{n-1} \leq \cdots \leq \lambda_1 \leq \mu_1$.

   (b) Ist $\alpha < 0$, so ist $\mu_n \leq \lambda_n \leq \mu_{n-1} \leq \cdots \leq \mu_1 \leq \lambda_1$.

Beweis: Wir betrachten den Fall $\alpha > 0$. Da $M - A = \alpha\, aa^T$ positiv semidefinit ist, folgt aus dem Courantschen Minimum-Maximum-Prinzip, dass $\lambda_j \leq \mu_j$, $j = 1, \ldots, m$.

O. B. d. A. sei $a \neq 0$. Durch $\{x_1, \ldots, x_{n-1}\}$ sei eine Orthonormalbasis von $\{x \in \mathbb{R}^n : a^T x = 0\}$ Hiermit definiere man

$$X := \begin{pmatrix} x_1 & \cdots & x_{n-1} \end{pmatrix} \in \mathbb{R}^{n\times(n-1)}.$$

Dann ist

$$X^T X = I, \qquad X^T M X = X^T A X + \alpha \underbrace{(X^T a)}_{=0}(X^T a)^T = X^T A X.$$

Seien $\nu_1 \geq \cdots \geq \nu_{n-1}$ die Eigenwerte von $X^T M X = X^T A X$. Eine Anwendung der Aussage der vorigen Aufgabe liefert

$$\lambda_{j+1} \leq \nu_j \leq \lambda_j, \qquad \mu_{j+1} \leq \nu_j \leq \mu_j, \qquad j = 1, \ldots, n-1.$$

Daher gilt

$$\lambda_j \leq \mu_j \leq \nu_{j-1} \leq \lambda_{j-1} \leq \mu_{j-1}, \qquad j = 2, \ldots, n,$$

womit die Aussage für $\alpha > 0$ bewiesen ist. Für $\alpha < 0$ verläuft der Beweis analog. □

6. Sei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$. Sind $\sigma_1 \geq \cdots \geq \sigma_n$ die singulären Werte von A, so ist

$$\sigma_j = \min_{N_j \in \mathcal{N}_j} \max_{0\neq x\in N_j} \frac{\|Ax\|_2}{\|x\|_2}, \qquad j = 1, \ldots, n,$$

wobei

$$\mathcal{N}_j := \{N_j \subset \mathbb{R}^n : \ N_j \text{ ist linearer Teilraum mit } \dim(N_j) = n + 1 - j\}.$$

Beweis: Es ist $\sigma_j = \lambda_j^{1/2}$, wobei $\lambda_1 \geq \cdots \geq \lambda_n$ die Eigenwerte der symmetrischen $n \times n$-Matrix $A^T A$ sind. Daher ist wegen des Courantschen Minimum-Maximum-Prinzips

$$\sigma_j = \lambda_j^{1/2} = \left(\min_{N_j \in \mathcal{N}_j} \max_{0\neq x\in N_j} \frac{x^T A^T A x}{x^T x} \right)^{1/2} = \min_{N_j \in \mathcal{N}_j} \max_{0\neq x\in N_j} \frac{\|Ax\|_2}{\|x\|_2}, \qquad j = 1, \ldots, n.$$

Das schließt schon den einfachen Beweis ab. □

7. Seien $A, B \in \mathbb{R}^{m\times n}$ mit $m \geq n$ und singulären Werten $\sigma_1 \geq \cdots \geq \sigma_n$ bzw. $\tau_1 \geq \cdots \geq \tau_n$ gegeben. Dann ist

$$|\sigma_j - \tau_j| \leq \|A - B\|_2, \qquad j = 1, \ldots, n.$$

Beweis: Für ein beliebiges $x \in \mathbb{R}^n \setminus \{0\}$ ist

$$\begin{aligned} \frac{\|Ax\|_2}{\|x\|_2} &= \frac{\|Bx + (A-B)x\|_2}{\|x\|_2} \\ &\leq \frac{\|Bx\|_2 + \|(A-B)x\|_2}{\|x\|_2} \\ &\leq \frac{\|Bx\|_2}{\|x\|_2} + \|A - B\|_2. \end{aligned}$$

Nun sei $j \in \{1, \ldots, n\}$ fest und wieder

$$\mathcal{N}_j := \{N_j \subset \mathbb{R}^n : \ N_j \text{ ist linearer Teilraum mit } \dim(N_j) = n + 1 - j\}.$$

Mit beliebigem $N_j \in \mathcal{N}_j$ ist dann

$$\max_{0\neq x\in N_j} \frac{\|Ax\|_2}{\|x\|_2} \leq \max_{0\neq x\in N_j} \frac{\|Bx\|_2}{\|x\|_2} + \|A-B\|_2$$

und folglich

$$\min_{N_j\in\mathcal{N}_j} \max_{0\neq x\in N_j} \frac{\|Ax\|_2}{\|x\|_2} \leq \min_{N_j\in\mathcal{N}_j} \max_{0\neq x\in N_j} \frac{\|Bx\|_2}{\|x\|_2} + \|A-B\|_2.$$

Aus Aufgabe 6 folgt dann

$$\sigma_j \leq \tau_j + \|A-B\|_2, \qquad j = 1,\ldots,n.$$

Vertauscht man hier die Rollen von A und B, so erhält man auch

$$\tau_j \leq \sigma_j + \|A-B\|_2, \qquad j = 1,\ldots,n.$$

Insgesamt ist die Behauptung bewiesen. □

8. Man berechne zu

$$A := \begin{pmatrix} 5 & 4 & 1 & 1 \\ 4 & 5 & 1 & 1 \\ 1 & 1 & 4 & 2 \\ 1 & 1 & 2 & 4 \end{pmatrix}$$

eine orthogonal ähnliche Tridiagonalmatrix. Mit Hilfe des QR-Verfahrens berechne man anschließend alle Eigenwerte von A. Hierbei wende man zum Vergleich beide angegebenen Shift-Strategien an.

Lösung: Als zu A orthogonal ähnliche Tridiagonalmatrix erhalten wir

$$A^{(1)} = \begin{pmatrix} 5.0000000000 & -4.2426406871 & & \\ -4.2426406871 & 6.0000000000 & 1.4142135624 & \\ & 1.4142135624 & 5.0000000000 & 0.0000000000 \\ & & 0.0000000000 & 2.0000000000 \end{pmatrix}.$$

Diese Matrix zerfällt, einen Eigenwert erhalten wir sofort, nämlich $\lambda = 2$. Wir ersetzen n (bzw. 4) durch $n-1$ (bzw. 3) und machen einen Schritt des QR-Verfahrens mit dem untersten Diagonalelement als Shift-Parameter. Man erhält

$$A^{(2)} = \begin{pmatrix} 6.0000000001 & 4.4721359550 & \\ 4.4721359550 & 5.0000000000 & 0.0000000000 \\ & 0.0000000000 & 5.0000000000 \end{pmatrix}.$$

Schon wieder haben wir Glück, die Matrix zerfällt, und wir haben den Eigenwert 5 erhalten. Wieder können wir die Dimension reduzieren. Die nächsten Schritte des QR-Verfahrens mit derselben Shift-Strategie ergeben

$$A^{(3)} = \begin{pmatrix} 6.9523809525 & -4.2591771000 \\ -4.2591771000 & 4.0476190475 \end{pmatrix},$$

$$A^{(4)} = \begin{pmatrix} 8.9349844197 & 2.9070400817 \\ 2.9070400817 & 2.0650155804 \end{pmatrix},$$

$$A^{(5)} = \begin{pmatrix} 9.9782918225 & -0.4414774664 \\ -0.4414774664 & 1.0217081776 \end{pmatrix},$$
$$A^{(6)} = \begin{pmatrix} 9.9999998728 & 0.0010700079 \\ 0.0010700079 & 1.0000001272 \end{pmatrix},$$
$$A^{(7)} = \begin{pmatrix} 10.0000000000 & 0.0000000000 \\ 0.0000000000 & 1.0000000000 \end{pmatrix}.$$

Damit hat man erhalten, dass A die Eigenwerte 1, 2, 5 und 10 besitzt. Wendet man auf die Tridiagonalmatrix $A^{(1)}$ das QR-Verfahren mit Wilkinson-Shift an, so erhalten wir

$$A^{(2)} = \begin{pmatrix} 7.8421052632 & 3.8287438904 & \\ 3.8287438904 & 3.1783029001 & 0.3774131022 \\ & 0.3774131022 & 4.9795918367 \end{pmatrix},$$
$$A^{(3)} = \begin{pmatrix} 8.4366343428 & -3.4097161601 & \\ -3.4097161601 & 2.5633699270 & 0.0049848158 \\ & 0.0049848158 & 4.9999957302 \end{pmatrix},$$
$$A^{(4)} = \begin{pmatrix} 8.9326997462 & 2.9097375229 & \\ 2.9097375229 & 2.0673002538 & 0.0000000072 \\ & 0.0000000000 & 5.0000000000 \end{pmatrix},$$
$$A^{(5)} = \begin{pmatrix} 9.2864656207 & -2.4315999062 & \\ -2.4315999062 & 1.7135343794 & 0.0000000000 \\ & 0.0000000000 & 5.0000000000 \end{pmatrix}.$$

Jetzt kann man die Aufgabe reduzieren, im nächsten Schritt erhält man die beiden restlichen Eigenwerte. □

9. Man wende das Rayleigh-Quotienten-Verfahren auf die Matrix

$$A := \begin{pmatrix} 5 & 4 & 1 & 1 \\ 4 & 5 & 1 & 1 \\ 1 & 1 & 4 & 2 \\ 1 & 1 & 2 & 4 \end{pmatrix}$$

an, wobei man mit $x_0 := \frac{1}{2}(1,1,1,1)^T$ starte. Gegen welchen Eigenwert konvergiert das Verfahren?

Lösung: Wir erhalten die folgende Folge $\{\rho_k\}$.

k	ρ_k
0	9.50000000000000
1	9.99315068493151
2	9.99999998709413
3	10.00000000000000

Die Rechnung wurde mit MATLAB gemacht. □

10. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und $a_{pq} \neq 0$ mit $1 \leq p < q \leq n$. Auf die folgende Weise berechne man eine Givens-Rotation $G_{pq} = G_{pq}(c,s)$.

- Berechne
$$\theta := \frac{a_{qq} - a_{pp}}{2a_{pq}}.$$

- Berechne
$$t := \frac{\operatorname{sign}(\theta)}{|\theta| + \sqrt{1+\theta^2}}.$$
- Berechne
$$c := \frac{1}{\sqrt{1+t^2}}, \qquad s := ct.$$

Man definiere $A^+ := G_{pq}^T A G_{pq}$ und zeige:

(a) Es ist $(A^+)_{pq} = (A^+)_{qp} = 0$.

(b) Es ist $N(A^+) = N(A) - 2a_{pq}^2$, wobei
$$N(A) := \sum_{\substack{i,j=1 \\ i \neq j}}^{n} a_{ij}^2.$$

(c) Ist (p,q) ein Indexpaar mit $1 \leq p < q \leq n$ und $|a_{pq}| = \max_{1 \leq i < j \leq n} |a_{ij}|$, so ist
$$N(A^+) \leq \left(1 - \frac{2}{n^2 - n}\right) N(A).$$

Beweis: Zunächst ist
$$c^2 + s^2 = \frac{1}{1+t^2} + \frac{t^2}{1+t^2} = 1,$$
also G_{pq} eine Givens-Rotation. Sei $\theta \neq 0$. Es ist
$$\begin{pmatrix} a_{pp}^+ & a_{pq}^+ \\ a_{qp}^+ & a_{qq}^+ \end{pmatrix} = \begin{pmatrix} c & -s \\ s & c \end{pmatrix} \begin{pmatrix} a_{pp} & a_{pq} \\ a_{pq} & a_{qq} \end{pmatrix} \begin{pmatrix} c & s \\ -s & c \end{pmatrix}$$
und daher
$$\begin{aligned}
a_{pq}^+ &= (a_{pp} - a_{qq})cs + a_{pq}(c^2 - s^2) \\
&= a_{pq}[-2\theta cs + c^2 - s^2] \\
&= \frac{a_{pq}}{1+t^2}[-2\theta t - 1 + t^2] \\
&= \frac{a_{pq}}{1+t^2}\left[\frac{2|\theta|}{|\theta| + \sqrt{1+\theta^2}} - 1 + \frac{1}{(|\theta| + \sqrt{1+\theta^2})^2}\right] \\
&= \frac{a_{pq}}{(1+t^2)(|\theta| + \sqrt{1+\theta^2})}[2|\theta|(|\theta| + \sqrt{1+\theta^2}) - (|\theta| + \sqrt{1+\theta^2})^2 + 1] \\
&= \frac{a_{pq}}{(1+t^2)(|\theta| + \sqrt{1+\theta^2})}[2\theta^2 + 2|\theta|\sqrt{1+\theta^2} - \theta^2 - 2|\theta|\sqrt{1+\theta^2} - (1+\theta^2) + 1] \\
&= 0.
\end{aligned}$$

Man überzeugt sich leicht davon, dass die Aussage auch für $\theta = 0$ richtig ist, wenn man z. B. $\operatorname{sign}(0) = 1$ definiert. Die Aussage (a) ist damit bewiesen. Zum Beweis von (b) beachte man zunächst, dass
$$\begin{pmatrix} a_{pp}^+ & 0 \\ 0 & a_{qq}^+ \end{pmatrix} = \begin{pmatrix} c & -s \\ s & c \end{pmatrix} \begin{pmatrix} a_{pp} & a_{pq} \\ a_{pq} & a_{qq} \end{pmatrix} \begin{pmatrix} c & s \\ -s & c \end{pmatrix},$$

daher ist wegen der Invarianz der Frobeniusnorm unter orthogonalen Transformationen

$$(a_{pp}^+)^2 + (a_{qq}^+)^2 = a_{pp}^2 + a_{qq}^2 + 2a_{pq}^2.$$

Folglich ist

$$\begin{aligned} N(A^+) &= \|A^+\|_F^2 - \sum_{i=1}^n (a_{ii}^+)^2 \\ &= \|A\|_F^2 - \sum_{i=1}^n (a_{ii}^+)^2 \\ &= \|A\|_F^2 - \sum_{\substack{i=1 \\ i\neq p,q}}^n a_{ii}^2 - (a_{pp}^+)^2 - (a_{qq}^+)^2 \\ &= \|A\|_F^2 - \sum_{i=1}^n a_{ii}^2 - 2a_{pq}^2 \\ &= N(A) - 2a_{pq}^2. \end{aligned}$$

Damit ist auch (b) bewiesen. Wählt man (p,q) so, dass $|a_{pq}| = \max_{1\le i<j\le n} |a_{ij}|$, so ist $a_{ij}^2 \le a_{pq}^2$ für $i \neq j$ und daher $N(A) \le (n^2-n)a_{pq}^2$. Folglich ist unter Benutzung von (b)

$$N(A^+) = N(A) - 2a_{pq}^2 \le \left(1 - \frac{2}{n^2-n}\right) N(A),$$

auch (c) ist bewiesen. □

6.4 Aufgaben in Kapitel 5

6.4.1 Aufgaben in Abschnitt 5.1

1. Man berechne die Eigenwerte und Eigenvektoren der Matrix

$$S_N := \begin{pmatrix} \alpha & \beta & & \\ \beta & \ddots & \ddots & \\ & \ddots & \ddots & \beta \\ & & \beta & \alpha \end{pmatrix} \in \mathbb{R}^{N\times N}.$$

Lösung: Für die Eigenvektoren von S_N machen wir wie im Falle $\alpha = 2$, $\beta = -1$, den Ansatz

$$z_i := (\sin(i\pi/(N+1)), \ldots, \sin(i\pi N/(N+1)))^T, \qquad i = 1, \ldots, N.$$

Für $j = 1, \ldots, N$ ist daher

$$\begin{aligned} (S_N z_i)_j &= \beta[(z_i)_{j-1} + (z_i)_{j+1}] + \alpha(z_i)_j \\ &= \beta[\sin(i(j-1)\pi/(N+1)) + \sin(i(j+1)\pi/(N+1))] + \alpha \sin(ij\pi/(N+1)) \\ &= [\alpha + 2\beta\cos(i\pi/(N+1))]\sin(ij\pi/(N+1)) \\ &= [\alpha + 2\beta\cos(i\pi/(N+1))](z_i)_j. \end{aligned}$$

Daher sind die Eigenwerte von S_N durch $\lambda_i := \alpha + 2\beta\cos(i\pi/(N+1))$, $i = 1, \ldots, N$, gegeben. □

2. Sei $A \in \mathbb{C}^{n\times n}$ (strikt) zeilenweise dominant, d. h.

$$\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| < |a_{ii}|, \qquad i = 1, \ldots, n.$$

Dann sind das Jacobi- und das Gauß-Seidel-Verfahren für jeden Startvektor $x^{(0)} \in \mathbb{C}^n$ konvergent.

Beweis: Es ist

$$(R_J)_{ij} = -\frac{1}{a_{ii}} a_{ij}(1-\delta_{ij}), \qquad 1 \le i, j \le n,$$

und daher

$$\rho(R_J) \le \|R_J\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |(R_J)_{ij}| = \max_{i=1,\ldots,n} \left(\frac{1}{|a_{ii}|} \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \right) < 1,$$

also das Jacobi-Verfahren konvergent. Nun nehmen wir an, λ sei ein Eigenwert von $R_{GS} = -(A_D + A_L)^{-1} A_R$ und x ein durch $\|x\|_\infty = 1$ normierter zugehöriger Eigenvektor. Sei etwa $|x_i| = \|x\|_\infty$. Aus $R_{GS}x = \lambda x$ erhalten wir $-A_R x = \lambda(A_D + A_R)x$. Betrachtet man hier jeweils die i-te Komponente, so folgt

$$|\lambda| = \frac{|\sum_{j=i+1}^{n} a_{ij}x_j|}{|\sum_{j=1}^{i} a_{ij}x_j|} \le \frac{\sum_{j=i+1}^{n} |a_{ij}|}{|a_{ii}| - \sum_{j=1}^{i-1} |a_{ij}|} < 1,$$

also ist $\rho(R_{GS}) < 1$, das Gauß-Seidel-Verfahren also konvergent[12]. □

3. Die Matrix $A \in \mathbb{C}^{n\times n}$ heißt bekanntlich *zerlegbar* (oder auch *zerfallend*, *reduzibel*), wenn es nichtleere Teilmengen N_1, N_2 von $N := \{1, \ldots, n\}$ mit $N_1 \cap N_2 = \emptyset$, $N_1 \cup N_2 = N$ sowie $a_{ij} = 0$ für alle $(i, j) \in N_1 \times N_2$ gibt, andernfalls *unzerlegbar* (oder auch *nichtzerfallend, irreduzibel*). Weiter heißt A *schwach zeilenweise diagonal dominant* (oder *genügt dem schwachen Zeilensummenkriterium*), wenn

$$\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \begin{cases} \le |a_{ii}| & \text{für } i = 1, \ldots, n, \\ < |a_{ii}| & \text{für mindestens ein } i = i_0. \end{cases}$$

Man zeige: Ist $A \in \mathbb{C}^{n\times n}$ schwach zeilenweise diagonal dominant und unzerlegbar, so sind alle Diagonalelemente von A von Null verschieden. Ferner sind das Jacobi- und das Gauß-Seidel-Verfahren für jeden Startvektor konvergent, die Matrix A nichtsingulär.

Beweis: Da A schwach zeilenweise diagonal dominant ist, können nicht alle Diagonalelemente von A verschwinden. Angenommen, wenigstens eines würde verschwinden. Dann ist durch

$$N_1 := \{i \in N : a_{ii} = 0\}, \qquad N_2 := \{j \in N : a_{jj} \neq 0\}$$

[12] Übrigens kann man sogar zeigen, dass $\|R_{GS}\|_\infty \le \|R_J\|_\infty < 1$, siehe z. B. J. W. Demmel (1997, S. 287).

eine nichttriviale Zerlegung von $N := \{1, \ldots, n\}$. Ist $i \in N_1$, so ist $a_{ij} = 0$ für $j = 1, \ldots, n\}$, da A dem schwachen Zeilensummenkriterium genügt. Insbesondere ist $a_{ij} = 0$ für $(i, j) \in N_1 \times N_2$, ein Widerspruch dazu, dass A unzerlegbar ist. Also sind das Jacobi- und das Gauß-Seidelverfahren durchführbar.

Da A schwach zeilenweise diagonal dominant ist, ist $\rho(R_J) \leq \|R_J\|_\infty \leq 1$. Angenommen, es gibt einen Eigenwert λ von R_J mit $|\lambda| = 1$, Sei x ein durch $\|x\|_\infty = 1$ normierter Eigenvektor zu λ. Man definiere die nichtleere Menge $N_1 := \{i \in N : |x_i| = 1\}$. Wegen

$$1 = |x_i| = |\lambda|\,|x_i| = |(R_J x)_i| = \frac{1}{|a_{ii}|}\left|\sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}x_j\right| \leq \frac{1}{|a_{ii}|}\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \leq 1, \qquad i \in N_1,$$

ist

$$\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| = |a_{ii}|, \qquad i \in N_1.$$

Daher ist $N_2 := N \setminus N_1 \neq \emptyset$, da ein Index i_0, für den beim schwachen Zeilensummenkriterium das strikte Ungleichheitszeichen gilt, nicht zu N_1 gehören kann. Da A unzerlegbar ist, existiert ein Indexpaar $(k, l) \in N_1 \times N_2$ mit $a_{kl} \neq 0$, insbesondere ist $|a_{kl}|\,|x_l| < |a_{kl}|$. Dann ist aber

$$1 = |\lambda|\,|x_k| = |(R_J x)_k| \leq \frac{1}{|a_{kk}|}\sum_{\substack{j=1\\ j\neq k}}^{n} |a_{kj}|\,|x_j| < \frac{1}{|a_{kk}|}\sum_{\substack{j=1\\ j\neq k}}^{n} |a_{kj}| \leq 1,$$

ein Widerspruch. Damit ist $\rho(R_J) < 1$ und die Konvergenz des Jacobi-Verfahrens bewiesen.

Zum Beweis des Gauß-Seidel-Verfahrens überlegen wir uns zunächst, dass $\|R_{GS}\|_\infty \leq 1$. Die Iterationsmatrix des Gauß-Seidel-Verfahrens ist $R_{GS} = -(A_D + A_L)^{-1}A_R$. Sei $x \in \mathbb{C}^n$ mit $\|x\|_\infty = 1$ beliebig vorgegeben und $y := R_{GS}x$ gesetzt. Dann ist $y = -A_D^{-1}(A_L y + A_R x)$. Komponentenweise bedeutet dies, dass

$$y_i = -\frac{1}{a_{ii}}\left(\sum_{j=1}^{i-1} a_{ij}y_j + \sum_{j=i+1}^{n} a_{ij}x_j\right), \qquad i = 1, \ldots, n.$$

Wegen $\|x\|_\infty = 1$ und des schwachen Zeilensummenkriteriums folgt folgt

$$|y_i| \leq \frac{1}{|a_{ii}|}\left(\sum_{j=1}^{i-1} |a_{ij}|\,|y_j| + \sum_{j=i+1}^{n} |a_{ij}|\right) \leq 1 - \frac{1}{|a_{ii}|}\sum_{j=1}^{i-1} |a_{ij}|\,(1 - |y_j|), \qquad i = 1, \ldots, n.$$

Durch vollständige Induktion nach i folgt hieraus, dass $|y_i| \leq 1$, $i = 1, \ldots, n$, bzw. $\|y\|_\infty \leq 1$. Damit ist $\|R_{GS}\|_\infty \leq 1$ bewiesen. Um $\rho(R_{GS}) < 1$ nachzuweisen können wir anfangen, wie beim Jacobi-Verfahren zu argumentieren. Angenommen, λ sei ein Eigenwert von R_{GS} mit $|\lambda| = 1$, $x \in \mathbb{C}^n$ sei ein zugehöriger, durch $\|x\|_\infty = 1$ normierter Eigenvektor. Dann ist $-(A_D + AL)^{-1}A_R x = \lambda x$ bzw. $\lambda x = -A_D^{-1}(\lambda A_L + A_R)x$. Wieder definiere man die nichtleere Menge $N_1 := \{i \in N : |x_i| = 1\}$. Wegen

$$1 = |x_i| = |\lambda|\,|x_i| = \frac{1}{|a_{ii}|}\left|\lambda\sum_{j=1}^{i-1} a_{ij}x_j + \sum_{j=i+1}^{n} a_{ij}x_j\right| \leq \frac{1}{|a_{ii}|}\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \leq 1, \qquad i \in N_1,$$

ist

$$\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| = |a_{ii}|, \qquad i \in N_1.$$

Daher ist wieder $N_2 := N \setminus N_1 \neq \emptyset$, wie oben folgt hieraus ein Widerspruch zur vorausgesetzten Unzerlegbarkeit von A.

Dass A nichtsingulär ist, ist nun klar. Denn wir haben nachgewiesen, dass $\rho(R_J) = \rho(I - A_D^{-1}A) < 1$, woraus die nichtsingularität von A unmittelbar folgt (andernfalls wäre 1 ein Eigenwert von R_J). □

4. Die Diagonalelemente von $A \in \mathbb{C}^{n\times n}$ seien ungleich Null. Dann ist $\rho(R_{SOR}(\omega)) \geq |\omega - 1|$, so dass das SOR-Verfahren (für jeden Startwert) nur konvergieren kann, wenn $\omega \in (0, 2)$.

   Beweis: Die Iterationsmatrix zum SOR-Verfahren mit dem Relaxationsparameter $\omega \neq 0$ ist

   $$R_{SOR}(\omega) = (A_D + \omega A_L)^{-1}[(1-\omega)A_D - \omega A_R] = (I + \omega A_D^{-1}A_L)^{-1}[(1-\omega)I - \omega A_D^{-1}A_R].$$

   Bekanntlich ist die Determinante einer Matrix gleich dem Produkt ihrer Eigenwerte. Daher ist

   $$|1-\omega|^n \leq |\det(R_{SOR}(\omega))| \leq \rho(R_{SOR}(\omega))^n,$$

   woraus die Behauptung folgt. □

5. Sei $A \in \mathbb{C}^{n\times n}$ hermitesch und positiv definit. Dann konvergiert das SOR-Verfahren für alle $\omega \in (0, 2)$, speziell also das Gauß-Seidel-Verfahren.

   Beweis: Das SOR-Verfahren wird von der Zerlegung $A = M(\omega) - N(\omega)]$ mit

   $$M(\omega) := \frac{1}{\omega}A_D + A_L$$

   erzeugt. Mit A ist auch die Diagonalmatrix A_D positiv definit. Wir wollen zeigen, dass $\rho(I - M(\omega)^{-1}A) < 1$. Sei hierzu λ ein Eigenwert von $R_{SOR}(\omega) = I - M(\omega)^{-1}A$ und x ein zugehöriger Eigenvektor. Dann ist

   $$[I - M(\omega)^{-1}A]x = \lambda x \qquad \text{bzw.} \qquad Ax = (1-\lambda)M(\omega)x.$$

   Da A als positiv definite Matrix nichtsingulär ist, ist $\lambda \neq 1$. Aus

   $$\frac{1}{1-\lambda} = \frac{x^H M(\omega)x}{x^H Ax}$$

   folgt

   $$2\Re\left(\frac{1}{1-\lambda}\right) = \frac{x^H[M(\omega) + M(\omega)^H]x}{x^H Ax} = 1 + \underbrace{\left(\frac{2}{\omega} - 1\right)}_{>0}\underbrace{\frac{x^H A_D x}{x^H Ax}}_{>0} > 1.$$

   Ist $\lambda = \alpha + i\beta$, so ist also

   $$1 < 2\Re\left(\frac{1}{1-\lambda}\right) = 2\Re\left(\frac{1}{1-\alpha-i\beta}\right) = \frac{2(1-\alpha)}{(1-\alpha)^2 + \beta^2}$$

   und daher $|\lambda|^2 = \alpha^2 + \beta^2 < 1$. Folglich ist $\rho(R_{SOR}(\omega)) < 1$, die Behauptung ist bewiesen. □

6. Sei

$$T_{N\times N} := \begin{pmatrix} A_N & -I_N & & \\ -I_N & \ddots & \ddots & \\ & \ddots & \ddots & -I_N \\ & & -I_N & A_N \end{pmatrix} \in \mathbb{R}^{N^2\times N^2},$$

wobei

$$A_N := \begin{pmatrix} 4 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{pmatrix} \in \mathbb{R}^{N\times N}.$$

Man zeige, dass $T_{N\times N}$ schwach zeilenweise diagonal dominant, unzerlegbar und (symmetrisch und) positiv definit ist, so dass wegen der Aussagen in den Aufgaben 3 und 5 das Jacobi- und das Gauß-Seidel-Verfahren sowie das SOR-Verfahren für alle $\omega \in (0,2)$ konvergent sind. Weiter zeige man, dass $T_{N\times N}^{-1}$ eine nichtnegative Matrix ist.

Beweis: Dass $T_{N\times N}$ schwach zeilenweise diagonal dominant ist, ist offensichtlich. Die Matrix ist auch unzerlegbar. Denn angenommen, N_1 und N_2 seien disjunkte Zerlegungen von $\{1,\dots,N^2\}$ mit $(T_{N\times N})_{ij} = 0$ für alle $(i,j) \in N_1 \times N_2$. O. B. d. A. nehmen wir an, es sei $1 \in N_1$. Da A_N eine Tridiagonalmatrix mit nichtverschwindenden Nebendiagonalelementen ist, ist dann offenbar $\{1,\dots,N\} \subset N_1$. Da weiter das Element in der Position $(1, N+1)$ nicht verschwindet (nämlich eine -1 ist), ist auch $N+1 \in N_1$, danach auch $\{1,\dots,2N\} \subset N_1$. So kann man fortfahren und erhält, dass $N_1 = \{1,\dots,N^2\}$ bzw. $N_2 = \emptyset$. Also ist $T_{N\times N}$ nicht zerfallend. Nun zeigen wir, dass $T_{N\times N}$ positiv definit ist. Wegen Aufgabe 3 ist $T_{N\times N}$ jedenfalls nichtsingulär, es kann also Null kein Eigenwert sein. Wegen des Satzes von Gerschgorin sind alle Eigenwerte in einem Kreis um 4 mit dem Radius 4 enthalten, sie liegen also (da $T_{N\times N}$ symmetrisch ist, sind sie reell) in $(0,8]$, insbesondere sind sie positiv und daher $T_{N\times N}$ positiv definit[13]. Insbesondere existiert $T_{N\times N}^{-1}$. Um nachzuweisen, dass $T_{N\times N}^{-1}$ eine nichtnegative Matrix ist, genügt es nachzuweisen, dass $T_{N\times N}u = b$ für jedes $b \geq 0$ eine nichtnegative Lösung besitzt (man nehme für b der Reihe nach alle Einheitsvektoren). Nun beachte man, dass die Iterationsmatrix zum Jacobi-Verfahren nichtnegativ ist. Startet man das Jacobi-Verfahren mit einem nichtnegativen Vektor, so erhält man eine Folge nichtnegativer Näherungslösungen. Da wir schon wissen, dass es konvergent ist, ist auch die Lösung nichtnegativ, womit alles bewiesen ist. □

7. Sei A eine Blocktridiagonalmatrix der Form

$$A = \begin{pmatrix} D_1 & H_1 & & & & \\ K_1 & D_2 & H_2 & & & \\ & K_2 & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & D_{M-1} & H_{M-1} \\ & & & & K_{M-1} & D_M \end{pmatrix},$$

[13] Da wir die Eigenwerte explizit kennen, hätten wir dies schneller zeigen können. Der hier eingeschlagene Weg kann aber in anderen Situationen auch zum Ziel führen.

wobei D_i, $i = 1, \ldots, M$, quadratische, nichtsinguläre Diagonalmatrizen sind. Man zeige, dass A konsistent geordnet ist.

Beweis: Der Beweis verläuft ganz ähnlich wie der entsprechende bei Tridiagonalmatrizen. Denn definiert man wieder

$$C(\alpha) := -A_D^{-1}\left(\alpha A_L + \frac{1}{\alpha} A_R\right),$$

so ist

$$C(\alpha) = \begin{pmatrix} 0 & (1/\alpha)\tilde{H}_1 & & & & \\ \alpha\tilde{K}_1 & 0 & (1/\alpha)\tilde{H}_2 & & & \\ & \alpha\tilde{K}_2 & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & 0 & (1/\alpha)\tilde{H}_{M-1} \\ & & & & \alpha\tilde{K}_{M-1} & 0 \end{pmatrix},$$

wobei $\tilde{H}_i := -D_i^{-1} H_i$ und $\tilde{K}_i := -D_{i+1}^{-1} K_i$, $i = 1, \ldots, M-1$. Man definiere die Blockdiagonalmatrix

$$S(\alpha) := \text{diag}\,(I, \alpha I, \ldots, \alpha^{M-1} I),$$

wobei die Dimension der Einheitsmatrizen mit der Dimension der entsprechenden Diagonalmatrizen D_i übereinstimmt. Dann ist offenbar

$$C(\alpha) = S(\alpha) C(1) S(\alpha)^{-1},$$

so dass $C(\alpha)$ und $C(1)$ dieselben Eigenwerte besitzen. □

8. Ist $A \in \mathbb{R}^{n \times n}$ konsistent geordnet und bezeichnen R_J bzw. R_{GS} die Iterationsmatrizen des Jacobi- bzw. Gauß-Seidel-Verfahrens, so ist $\rho(R_{GS}) = \rho(R_J)^2$.

   Beweis: Das Gauß-Seidel-Verfahren erhält man aus dem SOR-Verfahren, indem man für den Relaxationsparameter $\omega = 1$ setzt. Aus Satz 1.3 folgt: Ist $\mu \neq 0$ ein Eigenwert von $R_{GS} = R_{SOR}(1)$ und $\mu = \lambda^2$, so ist λ ein Eigenwert von R_J. Ist umgekehrt λ ein Eigenwert von R_J und $\mu^2 = \mu\lambda^2$, so ist μ ein Eigenwert von R_{GS}. Insgesamt ist die Aussage bewiesen. □

9. Seien $A, M, N \in \mathbb{R}^{n \times n}$ mit $A = M - N$. Das Paar (M, N) heißt eine *reguläre Zerlegung* von A, wenn M nichtsingulär ist und M^{-1} und N nichtnegative Matrizen sind. Man zeige: Ist (M, N) eine reguläre Zerlegung von A, so ist $\rho(M^{-1}N) < 1$ genau dann, wenn A nichtsingulär und A^{-1} nichtnegativ ist.

   Hinweis: Hierbei kann man benutzen, dass der Spektralradius einer nichtnegativen Matrix ein Eigenwert ist, zu dem es einen nichtnegativen Eigenvektor gibt (Perron-Frobenius).

   Beweis: Sei zunächst $\rho(M^{-1}N) < 1$. Ist $Ax = 0$, so folgt $M^{-1}Nx = x$. Da 1 kein Eigenwert von $M^{-1}N$ ist, ist $x = 0$ und daher A nichtsingulär. Um nachzuweisen, dass A^{-1} nichtnegativ ist, überlegen wir uns, dass die eindeutige Lösung von $Ax = b$ für jedes $b \geq 0$ nichtnegativ ist. Die Gleichung $Ax = b$ ist äquivalent zu $x = M^{-1}Nx + M^{-1}b$, das Iterationsverfahren

$$x^{(k+1)} := M^{-1}Nx^{(k)} + M^{-1}b$$

ist wegen $\rho(M^{-1}N) < 1$ für jedes $x^{(0)} \in \mathbb{R}^n$ konvergent gegen die eindeutige Lösung von $Ax = b$, insbesondere für den Startvektor $x^{(0)} := 0$. Wegen $M^{-1}N \geq 0$ und $M^{-1}b \geq 0$ ist $\{x^{(k)}\}$ eine Folge nichtnegativer Vektoren, folglich auch der Limes nichtnegativ.

Sei umgekehrt A nichtsingulär und $A^{-1} \geq 0$. Aus $A = M(I - M^{-1}N)$ erhält man, dass $I - M^{-1}N$ nichtsingulär ist. Ferner ist

$$(*) \qquad A^{-1}N = [M(I - M^{-1}N)]^{-1}N = (I - M^{-1}N)^{-1}M^{-1}N.$$

Da $M^{-1}N$ nichtnegativ ist, ist $\rho(M^{-1}N)$ ein Eigenwert von $M^{-1}N$, zu dem es einen nichtnegativen Eigenvektor x gibt:

$$M^{-1}Nx = \rho(M^{-1}N)x.$$

Hier ist $\rho(M^{-1}N) \neq 1$, da $I - M^{-1}N$ nichtsingulär ist. Da $A^{-1}N$ nichtnegativ ist, folgt aus $(*)$, dass

$$0 \leq A^{-1}Nx = \frac{\rho(M^{-1}N)}{1 - \rho(M^{-1}N)}x.$$

Da x als nichtnegativer Eigenvektor wenigstens eine positive Komponente besitzt, ist

$$0 \leq \frac{\rho(M^{-1}N)}{1 - \rho(M^{-1}N)}.$$

Hieraus folgt, wie behauptet, $\rho(M^{-1}N) < 1$. □

10. Gegeben sei das lineare Gleichungssystem

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix}.$$

Hierbei sei $A \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit, $B \in \mathbb{R}^{n \times m}$ habe den Rang m, ferner seien $b \in \mathbb{R}^n$ und $c \in \mathbb{R}^m$. Man zeige:

(a) Das obige lineare Gleichungssystem ist eindeutig lösbar.

(b) Die Matrix $S := B^T A^{-1} B$ ist positiv definit.

(c) Mit einem Relaxationsparameter $\omega \neq 0$ betrachte man zur numerischen Lösung des obigen linearen Gleichungssystems das folgende Iterationsverfahren (Uzawa-Verfahren):

- Wähle $(x_0, y_0) \in \mathbb{R}^n \times \mathbb{R}^m$.
- Für $k = 0, 1, \ldots$:
  - $x_{k+1} := A^{-1}(b - By_k)$.
  - $y_{k+1} := y_k + \omega(B^T x_{k+1} - c)$.

Dieses konvergiert genau dann, wenn $\omega \in (0, 2/\lambda_{\min}(S))$, ferner ist der optimale Relaxationsparameter gegeben durch

$$\omega_{\text{opt}} := \frac{2}{\lambda_{\min}(S) + \lambda_{\max}(S)}.$$

Beweis: Angenommen, es ist

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \text{bzw.} \qquad \begin{aligned} Ax + By &= 0, \\ B^T x &= 0. \end{aligned}$$

Eine Multiplikation der ersten Gleichung von links mit x^T von links liefert unter Berücksichtigung der zweiten Gleichung, dass

$$0 = x^T Ax + x^T By = x^T Ax + \underbrace{(B^T x)^T y}_{=0} = x^T Ax,$$

da A positiv definit ist, ist $x = 0$. Aus der ersten Gleichung folgt $By = 0$, wegen der Rangvoraussetzung an B folgt $y = 0$. Also ist die Koeffizientenmatrix des obigen Gleichungssystems nichtsingulär.

Mit A ist auch A^{-1} symmetrisch und positiv definit. Für ein beliebiges $y \in \mathbb{R}^m$ ist daher $y^T Sy = (By)^T A^{-1}(By) \geq 0$, wobei hier das Gleichheitszeichen genau dann gilt, wenn $By = 0$ bzw. $y = 0$. Daher ist S positiv definit.

Die Iterationsvorschrift für y_{k+1} ist

$$y_{k+1} = y_k + \omega[B^T A^{-1}(b - By_k) - c] = (I - \omega S)y_k + \omega(B^T A^{-1} b - c).$$

Daher ist die Folge $\{y_k\}$ genau dann für jeden Startwert y_0 konvergent, wenn der Spektralradius $\rho(I - \omega S)$ der Iterationsmatrix $I - \omega S$ kleiner als Eins ist. Hieran erkennt man, dass notwendigerweise $\omega > 0$. Denn andernfalls wären wegen der positiven Definitheit von S alle Eigenwerte der Iterationsmatrix ≥ 1. Wenn $\omega > 0$, so sind alle Eigenwerte von $I - \omega S$ kleiner als 1, daher $\rho(I - \omega S) < 1$ genau dann, wenn $-1 < 1 - \omega\lambda_{\max}(S)$ bzw. $\omega < 2/\lambda_{\max}(S)$. Weiter ist

$$\rho(I - \omega S) = \max(|1 - \omega\lambda_{\min}(S)|, |1 - \omega\lambda_{\max}(S)|).$$

Diese Funktion in ω ist minimal, wenn

$$1 - \omega\lambda_{\min}(S) = -(1 - \omega\lambda_{\max}(S)),$$

woraus man

$$\omega_{\text{opt}} := \frac{2}{\lambda_{\min}(S) + \lambda_{\max}(S)}$$

erhält. □

6.4.2 Aufgaben in Abschnitt 5.2

1. Sei $A \in \mathbb{C}^{n\times n}$. Dann ist A genau dann normal (also $A^H A = AA^H$), wenn ein Polynom $q \in \Pi_{n-1}$ mit $A^H = q(A)$ existiert.

   Beweis: Sei $A \in \mathbb{C}^{n\times n}$ normal. Dann ist A einer Diagonalmatrix unitär ähnlich, also $A = U\Lambda U^H$, wobei $U \in \mathbb{C}^{n\times n}$ unitär ist und $\Lambda = \text{diag}\,(\lambda_1, \ldots, \lambda_n)$ eine die Eigenwerte von A enthaltende Diagonalmatrix ist. Man wähle ein Polynom $q \in \Pi_{n-1}$ mit

   $$q(\lambda_i) = \overline{\lambda}_i, \qquad i = 1, \ldots, n.$$

Ein solches existiert, da man offenbar o. B. d. A. annehmen kann, dass die λ_i paarweise verschieden sind. Dann ist

$$q(A) = q(U\Lambda U^H) = Uq(\Lambda)U^H = U\overline{\Lambda}U^H = A^H.$$

Ist umgekehrt $A^H = q(A)$ mit einem Polynom $q \in \Pi_{n-1}$, so ist $A^H A = AA^H$ bzw. A normal, da $q(A)A = Aq(A)$. □

2. Sei $A \in \mathbb{R}^{n\times n}$, $v \in \mathbb{R}^n \setminus \{0\}$ und $m \in \mathbb{N}$. Man betrachte den folgenden Algorithmus, den man als Householder-Arnoldi-Verfahren bezeichnen kann:

   - Setze $z^{(1)} := v$.
   - Für $j = 1\ldots, k+1$:
     - Bestimme Householder-Matrix $\overline{P}_j \in \mathbb{R}^{(n-j+1)\times(n-j+1)}$ mit

       $$\overline{P}_j(z_j^{(j)}, z_{j+1}^{(j)}, \ldots, z_n^{(j)})^T = (*, 0, \ldots, 0)^T$$

       und setze $P_j := \operatorname{diag}(I_{j-1}, \overline{P}_j)$.
     - Berechne $h_{j-1} := P_j z^{(j)}$.
     - Berechne $q_j := P_1 P_2 \cdots P_j e_j$.
     - Falls $j \le k$, berechne $z^{(j+1)} := P_j P_{j-1} \cdots P_1 A q_j$.

   Man zeige[14] für diesen Algorithmus:

   (a) Es ist
   $$P_k \cdots P_1(\ v \quad Aq_1 \quad Aq_2 \quad \cdots \quad Aq_k\ ) = (\ h_0 \quad h_1 \quad \cdots \quad h_k\ ).$$

   (b) Die Vektoren $\{q_1, \ldots, q_k\}$ bilden eine Orthonormalbasis des Krylov-Teilraumes $\mathcal{K}_k$, wenn $(h_j)_{j+1} \neq 0$, $j = 1, \ldots, k$.

   Beweis: Zur Abkürzung definieren wir die orthogonale Matrix

   $$Q_j := P_j \cdots P_1 \in \mathbb{R}^{n\times n}, \qquad j = 1, \ldots, k+1.$$

   Dann lässt sich die Definition von $z^{(j+1)}$ auch als $z^{(j+1)} = Q_j A q_j$ schreiben. Weiter ist

   $$h_j = P_{j+1} z^{(j+1)} = P_{j+1} Q_j A q_j = Q_{j+1} A q_j.$$

   Nach Definition der Householder-Matrix P_{j+1} verschwinden alle Komponenten von h_j ab der Position $j+2$. Daher ist $P_i h_j = h_j$, $i = j+2, \ldots, k$. Folglich ist

   $$h_j = P_k \cdots P_{j+2} h_j = P_k \cdots P_{j+2} P_{j+1} A q_j = Q_k A q_j, \qquad j = 1, \ldots, k.$$

   Berücksichtigt man nun noch, dass

   $$Q_k v = P_k \cdots P_1 z^{(1)} = P_1 z^{(1)} = h_0,$$

   so hat man die Zerlegung

   $$Q_k(\ v \quad Aq_1 \quad Aq_2 \quad \cdots \quad Aq_k\ ) = (\ h_0 \quad h_1 \quad \cdots \quad h_k\ )$$

[14] Siehe Y. Saad (1996, S. 150 ff.).

bewiesen.

Nun wollen wir zeigen, dass $\{q_1, \ldots, q_k\}$ eine Orthonormalbasis des Krylov-Teilraumes $\mathcal{K}_k$ ist, wenn $(h_j)_{j+1} \neq 0$, $j = 1, \ldots, k$. Offensichtlich haben alle q_j die euklidische Länge 1. Sei nun $1 \leq i < j \leq k$. Unter Berücksichtigung der Tatsache, dass Householder-Matrizen symmetrisch und orthogonal sind, ist

$$q_i^T q_j = (P_1 \cdots P_i e_i)^T P_1 \cdots P_j e_j = e_i^T P_{i+1} \cdots P_j e_j = (P_j \cdots P_{i+1} e_i)^T e_j = e_i^T e_j = 0,$$

also ist $\{q_1, \ldots, q_k\}$ ein Orthonormalsystem. Nun zeigen wir, dass

$$(*) \qquad Aq_j = \sum_{i=1}^{j+1} (h_j)_i q_i, \qquad j = 1, \ldots, k.$$

Denn es ist

$$Aq_j = Q_{j+1}^T h_j = Q_{j+1}^T \sum_{i=1}^{j+1} (h_j)_i e_i = \sum_{i=1}^{j+1} (h_j)_i Q_{j+1}^T e_i.$$

Da weiter

$$Q_{j+1}^T e_i = P_1 \cdots P_{j+1} e_i = P_1 \cdots P_i \underbrace{P_{i+1} \cdots P_{j+1} e_i}_{=e_i} = P_1 \cdots P_i e_i = q_i,$$

ist $(*)$ bewiesen. Den Beweis für $\operatorname{span}\{q_1, \ldots, q_m\} \subset \mathcal{K}_m$ führen wir ähnlich wie beim MGS-Arnoldi-Verfahren. Wir zeigen nämlich durch vollständige Induktion nach j, dass $q_j = p_{j-1}(A)v$ mit einem Polynom $p_{j-1} \in \Pi_{j-1}$. Für den Induktionsanfang beachten wir, dass die Householder-Matrix P_1 den Vektor $z^{(1)} = v$ in ein Vielfaches des ersten Einheitsvektors überführt. Es ist also $P_1 v = \alpha_1 e_1$, wobei $\alpha_1 \neq 0$, da $v \neq 0$. Daher ist

$$q_1 = P_1 e_1 = \frac{1}{\alpha_1} v,$$

so dass der Induktionsanfang mit $p_0(t) := 1/\alpha_1$ gesichert ist. Für den Induktionsschritt beachten wir, dass man aus $(*)$ und der Induktionsannahme die Darstellung

$$\begin{aligned} q_{j+1} &= \frac{1}{(h_j)_{j+1}} \left(Aq_j - \sum_{i=1}^{j} (h_j)_i q_i \right) \\ &= \frac{1}{(h_j)_{j+1}} \left(Ap_{j-1}(A)v - \sum_{i=1}^{j} (h_j)_i p_{i-1}(A)v \right) \\ &= p_j(A)v, \end{aligned}$$

wenn man $p_j \in \Pi_j$ durch

$$p_j(t) := \frac{1}{(h_j)_{j+1}} \left[t p_{j-1}(t) - \sum_{i=1}^{j} (h_j)_i p_{i-1}(t) \right]$$

definiert. Damit ist die Aussage vollständig bewiesen. □

3. Es seien k Schritte des Arnoldi-Verfahrens zur Berechnung einer Orthonormalbasis des Krylov-Teilraumes $\mathcal{K}(A, r_0)$ durchführbar, der Output sei wie üblich bezeichnet. Es wird vorausgesetzt, dass die obere Hessenberg-Matrix H_k nichtsingulär ist. Seien $G_{j,j+1} = G_{j,j+1}(c_j, s_j)$, $j = 1, \ldots, k$, Givens-Rotationen, die $\tilde{H}_k$ sukzessive in eine obere Dreiecksmatrix transformieren. Durch Anwendung von FOM bzw. GMRES erhalte man die Näherungen x_k^F bzw. x_k^G. Mit r_k^F bzw. r_k^G seien die entsprechenden Defekte bezeichnet. Man zeige:

   (a) Es ist $\|r_k^F\|_2 = |h_{k+1,k}|\,|e_k^T y_k^F|$, wobei $y_k^F := H_k^{-1}(\|r_0\|_2\, e_1)$.

   (b) Es ist

$$\|r_k^F\|_2 = \|r_k^G\|_2 \sqrt{1 + \frac{h^2}{\xi^2}},$$

   wobei

$$\xi := (G_{k-1,k} \cdots G_{12} \tilde{H}_k)_{kk}, \qquad h := h_{k+1,k}.$$

Beweis: Aus dem Zusammenhang wird im folgenden hervorgehen, ob e_1 den ersten Einheitsvektor im $\mathbb{R}^k$ oder $\mathbb{R}^{k+1}$ bedeutet. Den ersten Teil der Aufgabe beweist man genau wie die entsprechende Aussage für das symmetrische Lanczos-Verfahren, mit der Abkürzung $\beta := \|r_0\|_2$ wiederholen wir den Beweis. Mit y_k^G bezeichnen wir die Lösung des linearen Ausgleichsproblems ist, $\|\beta e_1 - \tilde{H}_k y\|_2$ auf $\mathbb{R}^k$ zu minimieren. Dann ist

$$\begin{aligned} r_k^{F,G} &= r_0 - AQ_k y_k^{F,G} \\ &= r_0 - Q_{k+1}\tilde{H}_k y_k^{F,G} \\ &= Q_{k+1}(\beta e_1 - \tilde{H}_k y_k^{F,G}) \end{aligned}$$

und folglich

$$\|r_k^{F,G}\|_2 = \|\beta e_1 - \tilde{H}_k y_k^{F,G}\|_2.$$

Wegen

$$\beta e_1 - \tilde{H}_k y_k^F = \beta e_1 - \begin{pmatrix} H_k \\ h_{k+1,k} e_k^T \end{pmatrix} H_k^{-1}(\beta e_1) = \begin{pmatrix} 0 \\ -h_{k+1,k} e_k^T y_k^F \end{pmatrix}$$

folgt dann noch einmal der erste Teil der Aufgabe. Es ist

$$G_{k-1,k} \cdots G_{12} \tilde{H}_k = G_{k-1,k} \cdots G_{12} \begin{pmatrix} H_k \\ h e_k^T \end{pmatrix} = \begin{pmatrix} * & * & \cdots & * \\ & * & \cdots & * \\ & & \ddots & \vdots \\ & & & \xi \\ & & & h \end{pmatrix} = \begin{pmatrix} \hat{R}_k \\ h e_k^T \end{pmatrix}.$$

Die letzte Givens-Roration $G_{k,k+1}(c_k, s_k)$ wird so bestimmt, dass das $(k+1, k)$-Element annulliert wird, d. h. es ist

$$c_k = \pm \frac{\xi}{\sqrt{\xi^2 + h^2}}, \qquad s_k = \pm \frac{h}{\sqrt{\xi^2 + h^2}}.$$

Da H_k als nichtsingulär vorausgesetzt ist, ist $\xi \neq 0$ und damit auch $c_k \neq 0$. Ferner ist

$$G_{k,k+1}G_{k-1,k}\cdots G_{12}\tilde{H}_k = \begin{pmatrix} * & * & \cdots & * \\ & * & \cdots & * \\ & & \ddots & \vdots \\ & & & \eta \\ & & & 0 \end{pmatrix} = \begin{pmatrix} R_k \\ 0^T \end{pmatrix},$$

wobei

$$\eta = \pm\sqrt{\xi^2 + h^2}.$$

Hierbei sind die Einträge $*$ unverändert geblieben. Entsprechend ist

$$G_{k-1,k}\cdots G_{12}\begin{pmatrix} \beta e_1 \\ 0 \end{pmatrix} = \begin{pmatrix} \hat{g}_k \\ 0 \end{pmatrix},$$

ferner nach Multiplikation mit der letzten Givens-Rotation

$$G_{k,k+1}G_{k-1,k}\cdots G_{12}\begin{pmatrix} \beta e_1 \\ 0 \end{pmatrix} = \begin{pmatrix} g_k \\ \gamma_{k+1} \end{pmatrix}.$$

Hierbei ist

$$(g_k)_j = (\hat{g}_k)_j \quad (j = 1, \ldots, k-1), \qquad (g_k)_k = c_k(\hat{g}_k)_k, \qquad \gamma_{k+1} = -s_k(\hat{g}_k)_k.$$

Offensichtlich ist

$$y_k^F = \hat{R}_k^{-1}\hat{g}_k, \qquad y_k^G = R_k^{-1}g_k.$$

Folglich ist

$$\|r_k^F\|_2 = |h|\,|e_k^T y_k^F| = |h|\,\frac{|(\hat{g}_k)_k|}{|\xi|}, \qquad \|r_k^G\|_2 = |\gamma_{k+1}| = |s_k|\,|(\hat{g}_k)_k| = \frac{|h|}{\sqrt{\xi^2+h^2}}\,|(\hat{g})_k|.$$

Hieraus folgt schließlich

$$\|r_k^G\|_2\sqrt{1+\frac{h^2}{\xi^2}} = \frac{|h|}{\sqrt{\xi^2+h^2}}\,|(\hat{g})_k|\sqrt{1+\frac{h^2}{\xi^2}} = |h|\,\frac{|(\hat{g}_k)_k|}{|\xi|} = \|r_k^F\|_2,$$

das war zu zeigen[15]. □

[15]Siehe auch Y. Saad (1996, S. 168). Berücksichtigt man noch, dass $|(\hat{g}_k)_k| = \|r_{k-1}^G\|_2$, so folgt aus der obigen Darstellung von $\|r_k^F\|_2$ und $\|r_k^G\|_2$, dass

$$\|r_k^F\|_2 = \frac{\|r_k^G\|_2}{\sqrt{1-(\|r_k^G\|_2/\|r_{k-1}^G\|_2)^2}},$$

ein Resultat, das man in Theorem 2 der Arbeit

J. Cullum, A. Greenbaum (1996) "Relations between Galerkin and norm-minimizing iterative methods for solving linear systems." SIAM J. Matrix Anal. Appl. 17, 223–247.

finden kann.

4. Es seien k Schritte des Arnoldi-Verfahrens zur Berechnung einer Orthonormalbasis des Krylov-Teilraumes $\mathcal{K}(A, r_0)$ durchführbar, der Output sei wie üblich bezeichnet. Es wird vorausgesetzt, dass die obere Hessenberg-Matrix H_k nichtsingulär ist. Durch Anwendung von FOM bzw. GMRES erhalte man die Näherungen x_k^F bzw. x_k^G. Mit r_k^F bzw. r_k^G seien die entsprechenden Defekte bezeichnet. Man zeige: Ist $x_k^F = x_k^G$, so ist $r_k^F = r_k^G = 0$, d. h. sowohl FOM als auch GMRES liefern die exakte Lösung.

   Beweis: Wegen $x_k^F = x_k^G$ ist auch $\|r_k^F\|_2 = \|r_k^G\|_2$, aus Aufgabe 3b folgt, dass $h_{k+1,k} = 0$. Hieraus folgt aber (siehe frühere Bemerkung über Abbruch bei GMRES), dass x_k^G und damit auch x_k^F das gegebene lineare Gleichungssystem lösen. □

5. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Es sei bekannt, dass A nur $m \le n$ paarweise verschiedene Eigenwerte besitzt. Dann bricht das CG-Verfahren nach höchstens m Schritten ab.

   Beweis: Wir haben gezeigt, dass für die durch das CG-Verfahren erzeugte k-te Näherung x_k die Beziehung

$$\|x^* - x_k\|_A = \min_{p\in\Pi_k, p(0)=1} \|p(A)(x^* - x_0)\|_A, \qquad k = 1, 2, \ldots$$

   gilt. Seien $\lambda_1, \ldots, \lambda_m$ die paarweise verschiedenen Eigenwerte von A, welche natürlich alle von Null verschieden sind, da A als positiv definit vorausgesetzt wurde. Man definiere $p \in \Pi_m$ mit $p(0) = 1$ durch

$$p(t) := \prod_{i=1}^{m} \left(1 - \frac{t}{\lambda_i}\right).$$

   Wegen $p(\lambda) = 0$ für jeden Eigenwert λ von A ist $p(A) = 0$ und folglich $x_m = x^*$, womit die Aussage bewiesen ist. □

6. Auf das lineare Gleichungssystem $Ax = b$ werde GMRES und QMR angewandt, r_k^G und r_k^Q seien die entsprechenden Residuen im k-ten Schritt. Dann ist

$$\|r_k^Q\|_2 \le \kappa_2(V_{k+1})\, \|r_k^G\|_2,$$

   wobei V_{k+1} die durch das Lanczos-Biorthogonalisierungsverfahren berechnete Matrix der Basisvektoren von $\mathcal{K}_{k+1}(A, r_0)$ ist und $\kappa_2(\cdot)$ die Kondition bezüglich der Spektralnorm bedeutet.

   Hinweis: Ist $W \in \mathbb{R}^{n\times m}$ eine Matrix mit vollem Spaltenrang m, so ist $\|Wy\|_2 \ge \sigma_{\min}(W)\, \|y\|_2$, wobei $\sigma_{\min}(W)$ den kleinsten singulären Wert von W bedeutet.

   Beweis: Zunächst überlegen wir uns, dass die Aussage im Hinweis richtig ist. Sei $W = U\Sigma V^T$ eine (reduzierte) Singulärwertzerlegung von W. Dann ist $Wy = \sum_{i=1}^m \sigma_i v_i^T y\, u_i$, woraus man

$$\|Wy\|_2^2 = \sum_{i=1}^{m} \sigma_i^2 (v_i^T y)^2 \ge \sigma_{\min}(W)^2 \sum_{i=1}^{m} (v_i^T y)^2 = \sigma_{\min}(W)^2\, \|y\|_2^2$$

   erhält.

Es ist $x_k^Q = x_0 + V_k y_k^Q$, wobei y_k^Q Lösung der Aufgabe, $\|\beta e_1 - \tilde{T}_k y\|_2$ auf $\mathbb{R}^k$ zu minimieren ist. Hierbei haben wir zur Abkürzung wieder $\beta := \|r_0\|_2$ gesetzt. Unter Benutzung der Bezeichnungen und Aussagen von Satz 2.7 ist

$$\begin{aligned} r_k^Q &= b - A(x_0 + V_k y_k^Q) \\ &= r_0 - A V_k y_k^Q \\ &= V_{k+1}(\beta e_1 - \tilde{T}_k y_k^Q) \end{aligned}$$

und folglich

$$\|r_k^Q\|_2 \leq \|V_{k+1}\|_2 \min_{y \in \mathbb{R}^k} \|\beta e_1 - \tilde{T}_k y\|_2.$$

Als Element von $x_0 + \mathcal{K}_k(A, r_0)$ hat auch x_k^G eine Darstellung der Form $x_k^G = x_0 + V_k y_k$. Folglich ist $r_k^G = V_{k+1}(\beta e_1 - \tilde{T}_k y_k)$. Mit Hilfe des Hinweises folgt

$$\|r_k^G\|_2 \geq \sigma_{\min}(V_{k+1}) \, \|\beta e_1 - \tilde{T}_k y_k\|_2 \geq \sigma_{\min}(V_{k+1}) \min_{y \in \mathbb{R}^k} \|\beta e_1 - \tilde{T}_k y\|_2.$$

Insgesamt folgt die Behauptung, wenn man noch

$$\kappa_2(V_{k+1}) = \frac{\sigma_{\max}(V_{k+1})}{\sigma_{\min}(V_{k+1})} = \frac{\|V_{k+1}\|_2}{\sigma_{\min}(V_{k+1})}$$

berücksichtigt. □

7. Gegeben sei das lineare Gleichungssystem $Ax = b$ mit nichtsingulärem $A \in \mathbb{R}^{n \times n}$. Mit der Variablentransformation $x = A^T u$ wende man auf $AA^T u = b$ das CG-Verfahren an. Im resultierenden Verfahren (CGNE) eliminiere man die Variable u und gewinne eine Näherungsfolge $\{x_k\}$.

   Lösung: Das CG-Verfahren, angewandt auf $AA^T u = b$, lautet (die Richtungen nennen wir q_k statt p_k):

   - Mit einer Näherung u_0 berechne man $r_0 := b - AA^T u_0$ und setze $q_0 := r_0$.
   - Für $k = 0, 1, \ldots$:
     - $\alpha_k := r_k^T r_k / q_k^T AA^T q_k$.
     - $u_{k+1} := u_k + \alpha_k q_k$.
     - $r_{k+1} := r_k - \alpha_k AA^T q_k$.
     - $\beta_k := r_{k+1}^T r_{k+1} / r_k^T r_k$.
     - $q_{k+1} := r_{k+1} + \beta_k q_k$.

   Setzt man nun $x_k := A^T u_k$ und $p_k := A^T q_k$, so erhält man den folgenden Algorithmus:

   - Mit einer Näherung x_0 sei $r_0 := b - Ax_0$. Setze $p_0 := A^T r_0$.
   - Für $k = 0, 1, \ldots$
     - $\alpha_k := r_k^T r_k / p_k^T p_k$.
     - $x_{k+1} := x_k + \alpha_k p_k$.
     - $r_{k+1} := r_k - \alpha_k A p_k$.
     - $\beta_k := r_{k+1}^T r_{k+1} / r_k^T r_k$.
     - $p_{k+1} := A^T r_{k+1} + \beta_k p_k$.

Dieses Verfahren heißt auch CGNE (**C**onjugate **G**radient for the **N**ormal **E**quations), siehe z. B. Y. Saad (1996, S. 239). □

8. Auf das lineare Gleichungssystem $Ax = b$ mit

$$A := \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ 1 & & & & 0 \end{pmatrix} \in \mathbb{R}^{n\times n}, \qquad b := \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \in \mathbb{R}^n$$

wende man das Verfahren CGNE aus Aufgabe 7 an, wobei man mit $x_0 := 0$ starte. Was kann über die Konvergenz ausgesagt werden?

Lösung: Die Lösung des gegebenen linearen Gleichungssystems ist offenbar $x^* = e_2$. Am Start ist $r_0 = e_1$, danach $p_0 = e_2$, $\alpha_0 = 1$ und $x_1 = e_2$.Also konvergiert CGNE in einem Schritt. □

9. Gegeben sei das lineare Gleichungssystem $Ax = b$, bei dem die Koeffizientenmatrix A ein (nichttriviales) Vielfaches einer orthogonalen Matrix ist. Man zeige, dass das Verfahren CGNE aus Aufgabe 7 mit einem beliebigen Startvektor x_0 nach einem Schritt die Lösung bestimmt.

Beweis: Sei $A = \gamma Q$ mit $\gamma \neq 0$ und orthogonalem Q. Die Lösung von $Ax = b$ ist dann $x^* = (1/\gamma)Q^T b$. Es ist $r_0 = b - \gamma Q x_0$, danach

$$p_0 = \gamma Q^T (b - \gamma Q x_0) = \gamma (Q^T b - \gamma x_0).$$

Anschließend berechnet man

$$\alpha_0 = \frac{(b - \gamma Q x_0)^T (b - \gamma Q x_0)}{\gamma^2 (Q^T b - \gamma x_0)^T (Q^T b - \gamma x_0)} = \frac{1}{\gamma^2},$$

danach $x_1 = (1/\gamma)Q^T b$. Die Aussage ist bewiesen. □

10. Auf das lineare Gleichungssystem $Ax = b$ aus Aufgabe 8 wende man GMRES an, wobei man wieder mit $x_0 := 0$ starte. Was kann über die Konvergenz ausgesagt werden?

Lösung: Bei GMRES ist x_k die Lösung von

$$\text{Minimiere} \quad f(x) := \|b - Ax\|_2, \quad x \in x_0 + \mathcal{K}_k(A, r_0).$$

In unserem Falle lautet diese Aufgabe

$$\text{Minimiere} \quad f(x) := \left((1 - x_2)^2 + \sum_{\substack{j=1 \\ j\neq 2}}^{n} x_j^2 \right)^{1/2}, \quad x \in \operatorname{span}\{e_1, e_n, e_{n-1}, \ldots, e_{n-k+1}\}.$$

Außer für $k = n - 1$ ist stets $x_k = 0$, es tritt also überhaupt kein Fortschritt ein. □

6.4.3 Aufgaben in Abschnitt 5.3

1. Seien die Matrix A und ihr Präkonditionierer M symmetrisch und positiv definit. Man zeige, dass $M^{-1}A$ symmetrisch und positiv definit bezüglich des durch $\langle x, y\rangle_A := (Ax)^T y$ gegebenen inneren Produktes $\langle\cdot,\cdot\rangle_A$ ist. Hierauf aufbauend entwickele man ein CG-Verfahren zur Lösung des präkonditionierten linearen Gleichungssystems $M^{-1}Ax = M^{-1}b$ entsprechend dem von links präkonditionierten CG-Verfahrens. Dieses sollte mit nur einer Matrix-Vektor-Multiplikation pro Iterationsschritt auskommen.

   Lösung: Es ist

   $$\langle M^{-1}Ax, y\rangle_A = (AM^{-1}Ax)^T y = (Ax)^T M^{-1}Ay = \langle x, M^{-1}Ay\rangle_A,$$

   woraus die behauptete Symmetrie folgt. Da auch $AM^{-1}A$ positiv definit ist, ist $M^{-1}A$ positiv definit bezüglich des inneren Produkts $\langle\cdot,\cdot\rangle_A$.

   Bezeichnet man mit $r_k = b - Ax_k$ den "Originaldefekt" und mit $z_k = M^{-1}r_k$ den präkonditionierten Defekt, so lautet der k-te Schritt im entsprechend präkonditionierten CG-Verfahren offenbar:

   - $\alpha_k := \langle z_k, z_k\rangle_A / \langle M^{-1}Ap_k, p_k\rangle_A$.
   - $x_{k+1} := x_k + \alpha_k p_k$.
   - $r_{k+1} := r_k - \alpha_k Ap_k$, $z_{k+1} := M^{-1}r_{k+1}$.
   - $\beta_k := \langle z_{k+1}, z_{k+1}\rangle_A / \langle z_k, z_k\rangle_A$.
   - $p_{k+1} := z_{k+1} + \beta_k p_k$.

   Berücksichtigt man noch, dass

   $$\alpha_k = \frac{\langle z_k, z_k\rangle_A}{\langle M^{-1}Ap_k, p_k\rangle_A} = \frac{(Az_k)^T z_k}{(M^{-1}Ap_k)^T Ap_k}$$

   und

   $$\beta_k = \frac{\langle z_{k+1}, z_{k+1}\rangle_A}{\langle z_k, z_k\rangle_A} = \frac{(Az_{k+1})^T z_{k+1}}{(Az_k)^T z_k},$$

   so erkennt man, dass man den Algorithmus noch etwas verändern muss, um mit einer Matrix-Vektor-Multiplikation pro Iterationsschritt auszukommen. Dies ist aber möglich, da $p_0 = z_0$ und $Ap_{k+1} = Az_{k+1} + \beta_k Ap_k$. Auf die naheliegenden Details wollen wir nicht mehr eingehen. □

2. Gegeben sei das lineare Gleichungssystem $Ax = b$ mit nichtsingulärem $A \in \mathbb{R}^{n\times n}$. Man gebe zu dem in Aufgabe 7 in Abschnitt 5.2 entwickelten Verfahren CGNE ein von links mit einer symmetrischen, positiv definiten Matrix $M \in \mathbb{R}^{n\times n}$ präkonditioniertes CG-Verfahren an.

   Lösung: Präkonditioniert man das lineare Gleichungssystem $AA^T u = b$ von links mit M, so erhält man als zugehöriges CG-Verfahren:

   - Mit einer Näherung u_0 berechne man $r_0 := b - AA^T u_0$, berechne $z_0 := M^{-1}r_0$ und setze $q_0 := z_0$.
   - Für $k = 0, 1, \ldots$:
     - $\alpha_k := r_k^T z_k / q_k^T AA^T q_k$.

- $u_{k+1} := u_k + \alpha_k q_k$.
- $r_{k+1} := r_k - \alpha_k A A^T q_k$, $z_{k+1} := M^{-1} r_{k+1}$.
- $\beta_k := r_{k+1}^T z_{k+1} / r_k^T z_k$.
- $q_{k+1} := z_{k+1} + \beta_k q_k$.

Mit $p_k := A^T q_k$ und $x_k := A^T u_k$ erhält man die folgende Version:

- Mit einer Näherung x_0 berechne man $r_0 := b - A x_0$, berechne $z_0 := M^{-1} r_0$ und setze $p_0 := A^T z_0$.
- Für $k = 0, 1, \ldots$:
  - $\alpha_k := r_k^T z_k / p_k^T q_k$.
  - $x_{k+1} := x_k + \alpha_k p_k$.
  - $r_{k+1} := r_k - \alpha_k A p_k$, $z_{k+1} := M^{-1} r_{k+1}$.
  - $\beta_k := r_{k+1}^T z_{k+1} / r_k^T z_k$.
  - $p_{k+1} := A^T z_{k+1} + \beta_k p_k$.

Siehe auch Y. Saad (1996, S. 247). □

3. Sei $M = LR$ ein (nichtsingulärer) Präkonditionierer für die nichtsinguläre Matrix A (hierbei müssen L und R nicht notwendigerweise untere bzw. obere Dreiecksmatrizen sein). Man zeige, dass die Matrizen $M^{-1}A$, AM^{-1} und $L^{-1}AR^{-1}$ dieselben Eigenwerte besitzen.

   Beweis: Die Matrix $M^{-1}A$ ist ähnlich zu $M(M^{-1}A)M^{-1} = AM^{-1}$. Entsprechend ist $L^{-1}AR^{-1}$ ähnlich zu

$$L(L^{-1}AR^{-1})L^{-1} = AR^{-1}L^{-1} = AM^{-1},$$

   womit die einfache Aussage schon bewiesen ist. □

4. Die nichtsinguläre Matrix $A \in \mathbb{R}^{n\times n}$ habe nichtpositive Außerdiagonalelemente und eine nichtnegative Inverse. Man zeige, dass die Diagonalelemente von A positiv sind.

   Beweis: Sei $C := A^{-1}$. Dann ist

$$1 = (AC)_{ii} = \sum_{k=1}^{n} a_{ik}c_{ki} = a_{ii}c_{ii} + \sum_{\substack{k=1\\k\neq i}} \underbrace{a_{ik}c_{ki}}_{\leq 0} \leq a_{ii}\underbrace{c_{ii}}_{\geq 0}, \qquad i = 1, \ldots, n,$$

   woraus die Behauptung folgt. □

5. Seien $A, B \in \mathbb{R}^{n\times n}$ zwei Matrizen mit $A \leq B$ und $b_{ij} \leq 0$ für $i \neq j$. Ist dann A eine M-Matrix, so ist auch B eine M-Matrix.

   Hinweis: Man kann benutzen, dass der Spektralradius einer nichtnegativen Matrix ein Eigenwert ist, zu dem ein nichtnegativer Eigenvektor existiert (Perron-Frobenius).

   Beweis: Die Diagonalelemente von B sind positiv, da sie nicht kleiner sind als die positiven Diagonalelemente von A. Ferner sind die Außerdiagonalelemente von B nach Voraussetzung nichtpositiv. Angenommen, wir wüssten schon, dass $\rho(I - B_D^{-1}B) < 1$, wobei $B_D := \operatorname{diag}(B)$ die Diagonale von B ist. Dann ist B nichtsingulär und das das Jacobi-Verfahren

$$B_D x_{k+1} = (B_D - B)x_{k+1} + b$$

ist konvergent (siehe Satz 1.1 in Abschnitt 5.1). Da $B_D > 0$ und $B_D - B \geq 0$ wird für jedes $b \geq 0$, insbesondere die n Einheitsvektoren, eine Folge $\{x_k\}$ nichtnegativer Vektoren erzeugt, was wiederum $B^{-1} \geq 0$ impliziert.

Zu zeigen bleibt daher, dass $\rho(I - B_D^{-1}B) < 1$. Hierzu beachten wir, dass

$$0 \leq I - B_D^{-1}B = B_D^{-1}(B_D - B) \leq A_D^{-1}(B_D - B) \leq A_D^{-1}(A_D - A) = I - A_D^{-1}A.$$

Hieraus wollen wir auf $\rho(I - B_D^{-1}B) \leq \rho(I - A_D^{-1}A)$ schließen. Dies folgt direkt aus (Jordansche Normalform!)

$$\rho(I - B_D^{-1}B) = \lim_{k\to\infty} \|(I - B_D^{-1}B)^k\|_1^{1/k} \leq \lim_{k\to\infty} \|(I - A_D^{-1}A)^k\|_1^{1/k} = \rho(I - A_D^{-1}A).$$

Nun bleibt zu zeigen, dass $\rho(I - A_D^{-1}A) < 1$. Wegen des Satzes von Perron-Frobenius existiert ein $x \geq 0$, $x \neq 0$, mit

$$(I - A_D^{-1}A)x = \rho(I - A_D^{-1}A)x.$$

Da A als M-Matrix nichtsingulär ist, ist $\rho(I - A_D^{-1}A) \neq 1$. Umordnen liefert

$$0 \leq A^{-1}A_D x = \frac{1}{1 - \rho(I - A_D^{-1}A)}x,$$

woraus $\rho(I - A_D^{-1}A) < 1$ folgt. Insgesamt ist die Aussage bewiesen. □

6. Wir haben konstruktiv gezeigt, dass es zu jeder M-Matrix A und jedem die Diagonale nicht enthaltenden Nullmuster P eine untere Dreiecksmatrix L mit Einsen in der Diagonalen und eine obere Dreiecksmatrix R gibt derart, dass $A = LR - E$ eine reguläre Zerlegung von A ist und

$$l_{ij} = 0 \quad ((i,j) \in P), \qquad r_{ij} = 0 \quad ((i,j) \in P), \qquad e_{ij} = 0 \quad ((i,j) \notin P).$$

Man zeige, dass durch diese Forderungen die Faktoren L und R eindeutig bestimmt sind.

Beweis: Wir zitieren hier den entsprechenden Beweis bei J. A. Meijerink, H. A. van der Horst (1977) und (wörtlich gleich) A. Greenbaum (1997, S. 174):

- *The uniqueness of the factors L and R follows from equating the elements of A and LR for $(i,j) \notin P$, and from the fact that L has a unit diagonal.*

Wer kann dies präzisieren? □

7. Sei A eine M-Matrix und L, R und E wie in Satz 3.4 gegeben. Dann konvergiert die durch

$$LRx_{k+1} = Ex_k + b$$

definierte Folge $\{x_k\}$ für jeden Startvektor $x_0 \in \mathbb{R}^n$ gegen die Lösung des linearen Gleichungssystems $Ax = b$.

Beweis: Zu zeigen ist, dass $\rho((LR)^{-1}E) < 1$. Wegen Aufgabe 9 in Abschnitt 5.1 ist dies aber eine unmittelbare Folgerung aus der Tatsache, dass $A = LR - E$ eine reguläre Zerlegung von A ist. □

8. Eine symmetrische M-Matrix ist positiv definit.

   Beweis: Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische M-Matrix. Wir wollen zeigen, dass alle Hauptunterdeterminanten von A positiv sind, woraus bekanntlich die positive Definitheit von A folgt. Wir definieren

   $$\phi(\lambda) := \det(A_D + \lambda(A - A_D)).$$

   Dann ist $\phi(0) > 0$. Angenommen, es wäre $\phi(1) < 1$. Dann existiert ein $\lambda_0 \in (0,1)$ und ein $x \neq 0$ mit

   $$[A_D + \lambda_0(A - A_D)]x = 0.$$

   Umordnen liedert

   $$(I - A_D^{-1}A)x = \frac{1}{\lambda_0}x,$$

   d. h. $1/\lambda_0 \in (1,\infty)$ ist ein Eigenwert von $I - A_D^{-1}A$. Da $I - A_D^{-1}A \geq 0$, zeigt andererseits eine Anwendung des Satzes von Perron-Frobenius (siehe auch den letzten Teil des Beweises von Aufgabe 5, hier wird die Symmetrie von A noch nicht ausgenutzt), dass $\rho(I - A_D^{-1}A) < 1$. Dies ist ein Widerspruch, die Annahme $\phi(1) = \det(A) < 0$ ist widerlegt. Da A als M-Matrix nichtsingulär ist, ist $\det(A) > 0$. Da weiter jede Hauptuntermatrix einer M-Matrix wieder eine M-Matrix ist (Beweis?), sind alle Hauptunterdeterminanten von A positiv und folglich A positiv definit. □

9. Eine symmetrische, positiv definite Matrix $A \in \mathbb{R}^{n\times n}$ mit $a_{ij} \leq 0$, $i \neq j$, heißt eine *Stieltjes-Matrix*. Man zeige, dass eine Stieltjes-Matrix eine M-Matrix ist.

   Beweis: Die Diagonalelemente der symmetrischen, positiv definiten Matrix A sind positiv. Zu zeigen bleibt daher, dass $A^{-1} \geq 0$. Wie zu Beginn des Beweises von Aufgabe 5 gezeigt wurde, genügt es $\rho(I - A_D^{-1}A) < 1$ nachzuweisen. Da $A - A_D^{-1}A$ eine nichtnegative Matrix ist, ist $\rho := \rho(I - A_D^{-1}A)$ nach Perron-Frobenius ein Eigenwert von $I - A_D^{-1}A$. Dies wiederum impliziert, dass $1-\rho$ ein Eigenwert von $A_D^{-1}A$ ist, eine Matrix die ähnlich zu der symmetrischen, positiv definiten $A_D^{-1/2}AA_D^{-1/2}$ ist, also nur positive Eigenwerte besitzt. Folglich ist $\rho < 1$ und die Aussage ist bewiesen. □

10. Bei gegebenem nichtsingulärem $A \in \mathbb{R}^{m\times n}$ betrachte man die durch

    $$f(M) := \tfrac{1}{2}\,\|I - AM\|_F^2$$

    definierte Abbildung $f\colon \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}$. Wie lautet das Verfahren des steilsten Abstiegs mit exakter Schrittweite?

    Lösung: Wir versehen $\mathbb{R}^{n\times n}$ mit dem inneren Produkt $\langle X, Y\rangle := \mathrm{trace}(X^TY)$, welches die Frobenius-Norm $\|\cdot\|_F$ erzeugt. Der Gradient $\nabla f(M) \in \mathbb{R}^{n\times n}$ ist diejenige Matrix, für die

    $$f(M + E) = f(M) + \langle \nabla f(M), E\rangle + o(\|E\|_F).$$

    Nun ist

    $$\begin{aligned}
    f(M+E) - f(M) &= \tfrac{1}{2}\mathrm{trace}[(I - A(M+E))^T(I - A(M+E))] \\
    &\qquad - \tfrac{1}{2}\mathrm{trace}[(I - AM)^T(I - AM)] \\
    &= -\tfrac{1}{2}\mathrm{trace}[(AE)^T(I - AM) + (I - AM)^TAE - (AE)^T(AE)] \\
    &= -\mathrm{trace}[(A^T(I - AM))^TE] + \tfrac{1}{2}\mathrm{trace}[(AE)^T(AE)] \\
    &= -\langle A^T(I - AM), E\rangle + \tfrac{1}{2}\,\|AE\|_F^2 \\
    &= \langle -A^T(I - AM), E\rangle + o(\|E\|_F).
    \end{aligned}$$

Also ist $\nabla f(M) = -A^T(I - AM)$. Zur Bestimmumg der exakten Schrittweite ist die Lösung der eindimensionalen Optimierungsaufgabe

$$\text{Minimiere} \quad \phi(t) := f(M - t\nabla f(M)), \quad t \geq 0,$$

zu bestimmen. Nun ist

$$\begin{aligned} \phi(t) &= \tfrac{1}{2}\,\|I - AM - tAA^T(I - AM)\|_F^2 \\ &= \tfrac{1}{2}\,\|I - AM\|_F^2 - t\|A^T(I - AM)\|_F^2 + \tfrac{1}{2}t^2\,\|AA^T(I - AM)\|_F^2. \end{aligned}$$

Für $I - AM \neq 0$ (wovon ausgegangen werden kann) nimmt ϕ also auf $[0, \infty)$ in

$$t^* := \frac{\|A^T(I - AM)\|_F^2}{\|AA^T(I - AM)\|_F^2}$$

das Minimum an. Ein Schritt des Verfahrens des steilsten Abstiegs mit exakter Schrittweite ist also durch

$$M_+ := M + \frac{\|A^T(I - AM)\|_F^2}{\|AA^T(I - AM)\|_F^2} A^T(I - AM)$$

gegeben. □

11. Sei $A \in \mathbb{R}^{n\times n}$, $B := A^{-1}$ und $M = (\ m_1 \ \cdots \ m_n\ ) \in \mathbb{R}^{n\times n}$. Ist dann

$$|b_{ij}| > \|e_j - Am_j\|_1 \max_{k=1,\ldots,n} |b_{ik}|,$$

so ist $m_{ij} \neq 0$.

Beweis: Wir setzen $R := I - AM = (\ r_1 \ \cdots \ r_n\ )$. Dann ist

$$M = A^{-1} - A^{-1}R = B - BR$$

und daher

$$m_{ij} = b_{ij} - \sum_{k=1}^{n} b_{ik} r_{kj}.$$

Daher ist

$$\begin{aligned} |m_{ij}| &\geq |b_{ij}| - \sum_{k=1}^{n} |b_{ik} r_{kj}| \\ &\geq |b_{ij}| - \max_{k=1,\ldots,n} |b_{ik}| \sum_{k=1}^{n} |r_{kj}| \\ &= |b_{ij}| - \|r_j\|_1 \max_{k=1,\ldots,n} |b_{ik}| \\ &= |b_{ij}| - \|e_j - Am_j\|_1 \max_{k=1,\ldots,n} |b_{ik}| \\ &> 0, \end{aligned}$$

womit die Behauptung bewiesen ist. □