

Kapitel 4

Lineare Gleichungssysteme

Zu bestimmen sind die Lösungsvektoren $\underline{x} \in \mathbb{R}^n$ einer Familie von linearen Gleichungssystemen

$$A\underline{x} = \underline{f} \tag{4.1}$$

mit regulären, d.h. invertierbaren, Matrizen $A \in \mathbb{R}^{n \times n}$. Die Dimension n widerspiegelt in der Regel einen bei der Approximation auftretenden Diskretisierungsparameter, so daß der Fall $n \rightarrow \infty$ von Interesse ist.

Die Lösungsverfahren können in zwei Klassen eingeteilt werden. Die direkten Verfahren wie zum Beispiel das Gaußsche Eliminationsverfahren oder die LU-Zerlegung liefern bis auf Rundungsfehler die exakte Lösung des linearen Gleichungssystems. Der Aufwand hierfür beträgt im allgemeinen jedoch $\mathcal{O}(n^3)$ wesentliche Operationen, d.h. Multiplikationen bzw. Divisionen: eine Verdoppelung der Freiheitsgrade n verachtfacht die notwendige Rechenzeit zur Lösung.

In vielen Anwendungen resultiert das lineare Gleichungssystem jedoch aus einer Approximationsmethode zur näherungsweise Lösung eines mathematischen Modells. Der Lösungsvektor beschreibt also eine bereits fehlerbehaftete Näherungslösung. Damit ist es in der Regel ausreichend, auch die Lösung des linearen Gleichungssystems mittels eines Iterationsverfahrens nur mit einer hinreichenden Genauigkeit zu berechnen. Zu den klassischen stationären Iterationsverfahren gehören das Jacobi-Verfahren (Gesamtschrittverfahren) und das Gauss-Seidel Verfahren (Einzelschrittverfahren) sowie die daraus resultierenden Relaxationsverfahren. Für den Fall einer symmetrischen und positiv definiten Systemmatrix A soll hier das Verfahren konjugierter Gradienten (CG) behandelt werden, während für den allgemeinen Fall einer regulären Matrix A das Verfahren des verallgemeinerten minimalen Residuums (GMRES) betrachtet wird.

Für weiterführende Betrachtungen zu effizienten Lösungsverfahren für lineare Gleichungssysteme sei hier auf entsprechende weiterführende Vorlesungen bzw. Literatur verwiesen, siehe zum Beispiel [4, 9, 14].

4.1 Direkte Lösungsverfahren

Ist die Matrix A durch eine reguläre obere (oder untere) Dreiecksmatrix gegeben, d.h. $a_{ij} = 0$ für $j < i$, und $a_{ii} \neq 0$ für $i = 1, \dots, n$, so ergibt sich die Lösung unmittelbar durch rückwärtiges Einsetzen,

$$x_i = \frac{1}{a_{ii}} \left[f_i - \sum_{j=i+1}^n a_{ij} x_j \right] \quad \text{für } i = n, n-1, \dots, 1.$$

Für eine allgemein gegebene reguläre Matrix A soll diese zunächst auf obere Dreiecksgestalt transformiert werden. Ausgehend von der ersten Gleichung,

$$a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = f_1,$$

sollen alle weiteren Gleichungen, d.h. für $j = 2, \dots, n$,

$$a_{j1}x_1 + a_{j2}x_2 + \dots + a_{jn}x_n = f_j,$$

so transformiert werden, so dass der Koeffizient vor x_1 zu Null transformiert wird. Offenbar muss hierfür das $\frac{a_{j1}}{a_{11}}$ -fache der ersten Gleichung von der j -ten Gleichung abgezogen werden, wobei ohne Einschränkung der Allgemeinheit $a_{11} \neq 0$ vorausgesetzt wird. Dies ergibt, für $j = 2, \dots, n$,

$$\left(a_{j1} - \frac{a_{j1}}{a_{11}} a_{11} \right) x_1 + \left(a_{j2} - \frac{a_{j1}}{a_{11}} a_{12} \right) x_2 + \dots + \left(a_{jn} - \frac{a_{j1}}{a_{11}} a_{1n} \right) x_n = f_j - \frac{a_{j1}}{a_{11}} f_1,$$

d.h.,

$$\hat{a}_{j2}x_2 + \dots + \hat{a}_{jn}x_n = \hat{f}_j \quad \text{für } j = 2, \dots, n.$$

Dabei sind

$$\hat{a}_{jk} := a_{jk} - \frac{a_{j1}}{a_{11}} a_{1k} \quad \text{für } k = 2, \dots, n, \quad \hat{f}_j := f_j - \frac{a_{j1}}{a_{11}} f_1,$$

welche mit a_{jk} und f_j identifiziert werden können, d.h. es entsteht kein zusätzlicher Speicherbedarf zur Beschreibung der transformierten Gleichungen. Eine rekursive Anwendung für die Transformation der verbleibenden Spalten, d.h. für $i = 2, \dots, n-1$, und Auflösung des resultierenden linearen Gleichungssystems ergibt das in Algorithmus 4.1 angegebene Gaußsche Eliminationsverfahren. Hierbei wird vorausgesetzt, daß das jeweilige Diagonalelement a_{ii} von Null verschieden ist; andernfalls kann eine Pivotisierung durchgeführt werden, d.h. eine geeignete Vertauschung der Freiheitsgrade x_j , so daß a_{ii} das betragsmäßig größte Element der i -ten Zeile ist.

```

For  $i = 1, \dots, n-1$  do
  For  $j = i+1, \dots, n$  do
     $\alpha := \frac{a_{ji}}{a_{ii}}$ 
    For  $k = i, \dots, n$  do
       $a_{jk} := a_{jk} - \alpha a_{ik}$ 
       $f_j := f_j - \alpha f_i$ 
 $x_n := \frac{f_n}{a_{nn}}$ 
For  $i = n-1, \dots, 1$  do
   $\alpha := 0$ 
  For  $j = i+1, \dots, n$  do
     $\alpha := \alpha + a_{ij}x_j$ 
   $x_i := \frac{1}{a_{ii}}[f_i - \alpha]$ 

```

Algorithmus 4.1: Gaußsches Eliminationsverfahren ohne Pivotisierung.

Zur Abschätzung des Rechenaufwandes werden die wesentlichen Rechenoperationen, d.h. Multiplikationen und Divisionen, gezählt. Für die Erstellung der oberen Dreiecksmatrix ergibt dies

$$\begin{aligned}
 \sum_{i=1}^{n-1} \sum_{j=i+1}^n \left[1 + \sum_{k=i}^n 1 + 1 \right] &= \sum_{i=1}^{n-1} \sum_{j=i+1}^n (n+3-i) = \sum_{i=1}^{n-1} (n+3-i)(n-i) \\
 &= \sum_{i=1}^{n-1} [n^2 - (2n+3)i + i^2] \\
 &= n^2(n-1) - (2n+3)\frac{1}{2}(n-1)n + \frac{1}{6}(n-1)n(2(n-1)+1) \\
 &= \frac{1}{3}n(n-1)(n-5) \\
 &= \frac{1}{3}n^3 + \mathcal{O}(n^2),
 \end{aligned}$$

während für das Rückwärtseinsetzen zur Lösung des transformierten Gleichungssystems

$$1 + \sum_{i=1}^{n-1} \left[\sum_{j=i+1}^n 1 + 1 \right] = 1 + \sum_{i=1}^{n-1} (n+1-i) = 1 + (n+1)(n-1) - \frac{1}{2}n(n-1) = \frac{1}{2}n^2 + \frac{1}{2}n$$

folgt. Die Transformation auf obere Dreiecksgestalt ist also von kubischer Komplexität, während die Lösung des Gleichungssystems nur einen quadratischen Aufwand erfordert. Im Folgenden soll dieser Lösungsansatz so umgeschrieben werden, der die Lösung von (4.1) für viele rechte Seiten $\underline{f}$ ermöglicht, aber nur die einmalige Transformation der Matrix A erfordert.

Der erste Schritt des Gaußschen Eliminationsverfahrens kann geschrieben werden als Matrix-Matrix Produkt

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\frac{a_{21}}{a_{11}} & 1 & 0 & \cdots & 0 \\ -\frac{a_{31}}{a_{11}} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{a_{n1}}{a_{11}} & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & \hat{a}_{22} & \hat{a}_{23} & \cdots & \hat{a}_{2n} \\ 0 & \hat{a}_{32} & \hat{a}_{33} & \cdots & \hat{a}_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \hat{a}_{n2} & \hat{a}_{n3} & \cdots & \hat{a}_{nn} \end{pmatrix},$$

bzw. als

$$L_1 A = A_1$$

mit

$$\begin{aligned} L_1 &= \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\frac{a_{21}}{a_{11}} & 1 & 0 & \cdots & 0 \\ -\frac{a_{31}}{a_{11}} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{a_{n1}}{a_{11}} & 0 & 0 & \cdots & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} - \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ \frac{a_{21}}{a_{11}} & 0 & 0 & \cdots & 0 \\ \frac{a_{31}}{a_{11}} & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{a_{n1}}{a_{11}} & 0 & 0 & \cdots & 0 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} - \begin{pmatrix} 0 \\ \frac{a_{21}}{a_{11}} \\ \frac{a_{31}}{a_{11}} \\ \vdots \\ \frac{a_{n1}}{a_{11}} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \end{pmatrix} = I - \underline{a}_1 \underline{b}_1^\top. \end{aligned}$$

Die Transformation der ersten Spalte von A entspricht also der Multiplikation von A mit einer Elementarmatrix $L_1 = I - \underline{a}_1 \underline{b}_1^\top$, d.h. einer Rang 1 Störung der Einheitsmatrix. Durch rekursive Anwendung ergibt sich die obere Dreiecksmatrix

$$R := (I - \underline{a}_{n-1} \underline{b}_{n-1}^\top) \cdots (I - \underline{a}_2 \underline{b}_2^\top) (I - \underline{a}_1 \underline{b}_1^\top) A.$$

Für die Berechnung der Inversen von

$$L = I - \underline{a} \underline{b}^\top$$

mit dem Ansatz

$$L^{-1} = I + \alpha \underline{a} \underline{b}^\top$$

ergibt sich

$$\begin{aligned} L^{-1} L &= (I + \alpha \underline{a} \underline{b}^\top) (I - \underline{a} \underline{b}^\top) \\ &= I + \alpha \underline{a} \underline{b}^\top - \underline{a} \underline{b}^\top + \alpha \underline{a} \underline{b}^\top \underline{a} \underline{b}^\top \\ &= I + \left[\alpha (1 + \underline{b}^\top \underline{a}) - 1 \right] \underline{a} \underline{b}^\top = I \end{aligned}$$

für

$$\alpha = \frac{1}{1 + \underline{b}^\top \underline{a}}, \quad \underline{b}^\top \underline{a} \neq -1.$$

Diese Invertierungsformel ist als Sherman–Morrison Formel bekannt. Wegen

$$\underline{b}_1^\top \underline{a}_1 = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} 0 \\ \frac{a_{21}}{a_{11}} \\ \frac{a_{31}}{a_{11}} \\ \vdots \\ \frac{a_{n1}}{a_{11}} \end{pmatrix} = 0$$

folgt

$$L_1^{-1} = I + \underline{a}_1 \underline{b}_1^\top,$$

und somit

$$A = (I + \underline{a}_1 \underline{b}_1^\top)(I + \underline{a}_2 \underline{b}_2^\top) \cdots (I + \underline{a}_{n-1} \underline{b}_{n-1}^\top) R = LR,$$

mit

$$L = (I + \underline{a}_1 \underline{b}_1^\top)(I + \underline{a}_2 \underline{b}_2^\top) \cdots (I + \underline{a}_{n-1} \underline{b}_{n-1}^\top).$$

Beispiel 4.1. *Gegeben sei die Matrix*

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

für welche sich

$$L_1 = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

ergibt. Dann ist

$$A_1 = L_1 A = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 0 \\ 0 & \frac{7}{2} & 1 \\ 0 & 1 & 2 \end{pmatrix},$$

woraus sich

$$L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{2}{7} & 1 \end{pmatrix}$$

ergibt. Damit folgt

$$R = L_2 A_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{2}{7} & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 0 \\ 0 & \frac{7}{2} & 1 \\ 0 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 0 \\ 0 & \frac{7}{2} & 1 \\ 0 & 0 & \frac{12}{7} \end{pmatrix},$$

und

$$L = L_1^{-1} L_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{2}{7} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ 0 & \frac{2}{7} & 1 \end{pmatrix}.$$

Das Gaußsche Eliminationsverfahren beschreibt also eine Realisierung der bekannten LR Zerlegung einer gegebenen regulären Matrix A in das Produkt $A = LR$ einer regulären unteren Dreiecksmatrix L , und einer regulären oberen Dreiecksmatrix R . Diese kann auch durch einen geeigneten Ansatz realisiert werden, d.h. für $A = LR$ ist

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} = \begin{pmatrix} 1 & & & & \\ l_{21} & 1 & & & \\ l_{31} & l_{32} & 1 & & \\ \vdots & \vdots & & \ddots & \\ l_{n1} & l_{n2} & \cdots & \cdots & 1 \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} & r_{13} & \cdots & r_{1n} \\ & r_{22} & r_{23} & \cdots & r_{2n} \\ & & r_{33} & & \vdots \\ & & & \ddots & \vdots \\ & & & & r_{nn} \end{pmatrix},$$

d.h.

$$a_{ij} = \sum_{k=1}^n l_{ik} r_{kj} = \sum_{k=1}^{\min(i,j)} l_{ik} r_{kj}.$$

Für $i \leq j$ ist dann

$$a_{ij} = \sum_{k=1}^i l_{ik} r_{kj} = \sum_{k=1}^{i-1} l_{ik} r_{kj} + r_{ij}$$

und somit

$$r_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik} r_{kj} \quad \text{für } i \leq j = 1, \dots, n.$$

Für $i > j$ ist

$$a_{ij} = \sum_{k=1}^j l_{ik} r_{kj} = \sum_{k=1}^{j-1} l_{ik} r_{kj} + l_{ij} r_{jj}$$

und somit

$$l_{ij} = \frac{1}{r_{jj}} \left[a_{ij} - \sum_{k=1}^{j-1} l_{ik} r_{kj} \right].$$

Dies ergibt die in Algorithmus 4.2 angegebene LR Zerlegung einer gegebenen regulären Matrix A . Wie beim Gaußschen Eliminationsverfahren können die Matrizen L und R mit der Ausgangsmatrix A identifiziert werden, man beachte $\text{diag } L = I$.

```

For  $j = 1, \dots, n$  do
  For  $k = 1, \dots, j - 1$  do
    For  $i = k + 1, \dots, j - 1$  do
       $a_{ij} := a_{ij} - a_{ik}a_{kj}$ 
    For  $k = 1, \dots, j - 1$  do
      For  $i = j, n$ 
         $a_{ij} = a_{ij} - a_{ik}a_{kj}$ 
      For  $i = j + 1, n$  do
         $a_{ij} = \frac{a_{ij}}{a_{jj}}$ 

```

Algorithmus 4.2: LR Zerlegung ohne Pivotisierung.

Beispiel 4.2. *Betrachtet wird wieder die Matrix*

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{pmatrix}.$$

Dann ergibt sich:

$$\begin{aligned}
a_{11} = 2 = r_{11} &\Rightarrow r_{11} = 2, \\
a_{12} = 1 = r_{12} &\Rightarrow r_{12} = 1, \\
a_{13} = 0 = r_{13} &\Rightarrow r_{13} = 0, \\
a_{21} = 1 = l_{21}r_{11} = 2l_{21} &\Rightarrow l_{21} = \frac{1}{2}, \\
a_{22} = 4 = l_{21}r_{12} + r_{22} = \frac{1}{2} + r_{22} &\Rightarrow r_{22} = \frac{7}{2}, \\
a_{23} = 1 = l_{21}r_{13} + r_{23} = r_{23} &\Rightarrow r_{23} = 1, \\
a_{31} = 0 = l_{31}r_{11} = 2l_{31} &\Rightarrow l_{31} = 0, \\
a_{32} = 1 = l_{31}r_{12} + l_{32}r_{22} = \frac{7}{2}l_{32} &\Rightarrow l_{32} = \frac{2}{7}, \\
a_{33} = 2 = l_{31}r_{13} + l_{32}r_{23} + r_{33} = \frac{2}{7} + r_{33} &\Rightarrow r_{33} = \frac{12}{7},
\end{aligned}$$

d.h.

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ 0 & \frac{2}{7} & 1 \end{pmatrix}, \quad R = \begin{pmatrix} 2 & 1 & 0 \\ 0 & \frac{7}{2} & 1 \\ 0 & 0 & \frac{12}{7} \end{pmatrix}.$$

Für die Lösung eines linearen Gleichungssystems $A\underline{x} = \underline{f}$ mit der Faktorisierung $A = LR$, d.h. $LR\underline{x} = \underline{f}$, ergibt sich mit der Transformation $\underline{z} = R\underline{x}$ die Lösung $\underline{x}$ aus

$$L\underline{z} = \underline{f}, \quad R\underline{x} = \underline{z}.$$

Im Folgenden wird die Transformation einer gegebenen Matrix A auf Dreiecksgestalt unter Verwendung orthogonaler Matrizen

$$P = I - \frac{2}{\underline{v}^\top \underline{v}} \underline{v} \underline{v}^\top$$

betrachtet. Offensichtlich gilt

$$\begin{aligned} P^\top P &= \left(I - \frac{2}{\underline{v}^\top \underline{v}} \underline{v} \underline{v}^\top \right) \left(I - \frac{2}{\underline{v}^\top \underline{v}} \underline{v} \underline{v}^\top \right) \\ &= I - \frac{4}{\underline{v}^\top \underline{v}} \underline{v} \underline{v}^\top + \frac{4}{[\underline{v}^\top \underline{v}]^2} \underline{v} \underline{v}^\top \underline{v} \underline{v}^\top = I. \end{aligned}$$

Weiters ist

$$P \underline{v} = \left(I - \frac{2}{\underline{v}^\top \underline{v}} \underline{v} \underline{v}^\top \right) \underline{v} = -\underline{v},$$

und für $\underline{w} \in \mathbb{R}^n$ mit $\underline{v}^\top \underline{w} = 0$ folgt

$$P \underline{w} = \left(I - \frac{2}{\underline{v}^\top \underline{v}} \underline{v} \underline{v}^\top \right) \underline{w} = \underline{w}.$$

Für einen beliebigen Vektor $\underline{x} \in \mathbb{R}^n$ ergibt sich dann

$$\underline{x} = \frac{\underline{x}^\top \underline{v}}{\underline{v}^\top \underline{v}} \underline{v} + \frac{\underline{x}^\top \underline{w}}{\underline{w}^\top \underline{w}} \underline{w},$$

und somit

$$P \underline{x} = \frac{\underline{x}^\top \underline{v}}{\underline{v}^\top \underline{v}} P \underline{v} + \frac{\underline{x}^\top \underline{w}}{\underline{w}^\top \underline{w}} P \underline{w} = -\frac{\underline{x}^\top \underline{v}}{\underline{v}^\top \underline{v}} \underline{v} + \frac{\underline{x}^\top \underline{w}}{\underline{w}^\top \underline{w}} \underline{w},$$

d.h. die Householder-Transformation P beschreibt eine Spiegelung von $\underline{x}$ an der durch den Vektor $\underline{v}$ beschriebenen Ebene.

Für die Transformation einer gegebenen Matrix A auf obere Dreiecksgestalt betrachten wir zunächst die Transformation der ersten Spalte $\underline{a}$ von A auf ein Vielfaches des ersten Einheitsvektors $\underline{e}$, d.h.

$$P \underline{a} = \left(I - \frac{2}{\underline{v}^\top \underline{v}} \underline{v} \underline{v}^\top \right) \underline{a} = \alpha \underline{e}.$$

Es ist

$$P \underline{a} = \underline{a} - 2 \frac{\underline{v}^\top \underline{a}}{\underline{v}^\top \underline{v}} \underline{v} = \alpha \underline{e},$$

d.h., $\underline{v}$ ist proportional zu $\underline{a} - \alpha \underline{e}$,

$$2 \frac{\underline{v}^\top \underline{a}}{\underline{v}^\top \underline{v}} \underline{v} = \underline{a} - \alpha \underline{e}.$$

Da die Definition von P die Normierung von $\underline{v}$ beinhaltet, kann insbesondere

$$\underline{v} = \underline{a} - \alpha \underline{e}$$

gewählt werden. Aus

$$\begin{aligned}\alpha \underline{e} = P\underline{a} &= \left(I - \frac{2}{\underline{v}^\top \underline{v}} \underline{v} \underline{v}^\top \right) \underline{a} = \underline{a} - 2 \frac{\underline{v}^\top \underline{a}}{\underline{v}^\top \underline{v}} \underline{v} \\ &= \underline{a} - 2 \frac{\underline{v}^\top \underline{a}}{\underline{v}^\top \underline{v}} [\underline{a} - \alpha \underline{e}] = \left[1 - 2 \frac{\underline{v}^\top \underline{a}}{\underline{v}^\top \underline{v}} \right] \underline{a} + 2 \frac{\underline{v}^\top \underline{a}}{\underline{v}^\top \underline{v}} \alpha \underline{e}\end{aligned}$$

folgt dann

$$2 \underline{v}^\top \underline{a} = \underline{v}^\top \underline{v}.$$

Mit

$$\begin{aligned}\underline{v}^\top \underline{v} &= (\underline{a}^\top - \alpha \underline{e}^\top)(\underline{a} - \alpha \underline{e}) \\ &= \underline{a}^\top \underline{a} - \alpha \underline{e}^\top \underline{a} - \alpha \underline{a}^\top \underline{e} + \alpha^2 \underline{e}^\top \underline{e} \\ &= \underline{a}^\top \underline{a} - 2\alpha a_{11} + \alpha^2\end{aligned}$$

und

$$\underline{v}^\top \underline{a} = (\underline{a}^\top - \alpha \underline{e}^\top) \underline{a} = \underline{a}^\top \underline{a} - \alpha \underline{e}^\top \underline{a} = \underline{a}^\top \underline{a} - \alpha a_{11}$$

ergibt sich aus

$$2 [\underline{a}^\top \underline{a} - \alpha a_{11}] = 2 \underline{v}^\top \underline{a} = \underline{v}^\top \underline{v} = \underline{a}^\top \underline{a} - 2\alpha a_{11} + \alpha^2$$

schliesslich

$$\alpha^2 = \underline{a}^\top \underline{a},$$

und zur Vermeidung von Stellenauslöschungen wird

$$\alpha = \begin{cases} -\|\underline{a}\|_2 & \text{für } a_{11} \geq 0, \\ \|\underline{a}\|_2 & \text{für } a_{11} < 0 \end{cases}$$

gesetzt.

Beispiel 4.3. *Betrachtet wird wieder die Matrix*

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 2 \end{pmatrix}$$

mit

$$\underline{a} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}, \quad \|\underline{a}\|_2 = \sqrt{5}, \quad \alpha = -\sqrt{5}.$$

Dann ist

$$\underline{v} = \underline{a} - \alpha \underline{e} = \begin{pmatrix} 2 + \sqrt{5} \\ 1 \\ 0 \end{pmatrix}, \quad \underline{v}^\top \underline{v} = (2 + \sqrt{5})^2 + 1 = 10 + 4\sqrt{5},$$

und

$$\begin{aligned}
 P_1 &= I - \frac{2}{\underline{v}^\top \underline{v}} \underline{v} \underline{v}^\top = I - \frac{2}{10 + 4\sqrt{5}} \begin{pmatrix} 2 + \sqrt{5} \\ 1 \\ 0 \end{pmatrix} \begin{pmatrix} 2 + \sqrt{5} & 1 & 0 \end{pmatrix} \\
 &= I - \frac{1}{5 + 2\sqrt{5}} \begin{pmatrix} (2 + \sqrt{5})^2 & 2 + \sqrt{5} & 0 \\ 2 + \sqrt{5} & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \\
 &= \frac{1}{5 + 2\sqrt{5}} \begin{pmatrix} 5 + 2\sqrt{5} - (2 + \sqrt{5})^2 & -2 - \sqrt{5} & 0 \\ -2 - \sqrt{5} & 5 + 2\sqrt{5} - 1 & 0 \\ 0 & 0 & 5 + 2\sqrt{5} \end{pmatrix} \\
 &= \frac{1}{5 + 2\sqrt{5}} \begin{pmatrix} -4 - 2\sqrt{5} & -2 - \sqrt{5} & 0 \\ -2 - \sqrt{5} & 4 + 2\sqrt{5} & 0 \\ 0 & 0 & 5 + 2\sqrt{5} \end{pmatrix}.
 \end{aligned}$$

Damit ergibt sich

$$\begin{aligned}
 P_1 A &= \frac{1}{5 + 2\sqrt{5}} \begin{pmatrix} -4 - 2\sqrt{5} & -2 - \sqrt{5} & 0 \\ -2 - \sqrt{5} & 4 + 2\sqrt{5} & 0 \\ 0 & 0 & 5 + 2\sqrt{5} \end{pmatrix} \begin{pmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 2 \end{pmatrix} \\
 &= \frac{1}{5 + 2\sqrt{5}} \begin{pmatrix} -10 - 5\sqrt{5} & -12 - 6\sqrt{5} & -2 - \sqrt{5} \\ 0 & 14 + 7\sqrt{5} & 4 + 2\sqrt{5} \\ 0 & 5 + 2\sqrt{5} & 10 + 4\sqrt{5} \end{pmatrix}.
 \end{aligned}$$

Auf die Konstruktion von P_2 soll an dieser Stelle verzichtet werden.

Abschließend sollen die Givens-Rotationen

$$G = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}$$

als orthogonale Transformationen betrachtet werden, d.h., aus

$$G^\top G = \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix} \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix} = \begin{pmatrix} \alpha^2 + \beta^2 & 0 \\ 0 & \alpha^2 + \beta^2 \end{pmatrix}$$

ergibt sich die Forderung

$$\alpha^2 + \beta^2 = 1.$$

Durch die Anwendung von G soll die zweite Komponente eines gegebenen Vektors zu Null transformiert werden, d.h.

$$\begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \bar{x} \\ 0 \end{pmatrix}.$$

Daraus folgt

$$\beta x + \alpha y = 0, \quad \alpha = \frac{x}{\sqrt{x^2 + y^2}}, \quad \beta = -\frac{y}{\sqrt{x^2 + y^2}}, \quad \bar{x} = \alpha x - \beta y = \sqrt{x^2 + y^2}.$$

Beispiel 4.4. Für die Matrix

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 2 \end{pmatrix}$$

betrachten wir zunächst

$$\underline{a} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad x = 2, \quad y = 1, \quad \alpha = \frac{2}{\sqrt{5}}, \quad \beta = -\frac{1}{\sqrt{5}},$$

d.h.

$$G_1 = \frac{1}{\sqrt{5}} \begin{pmatrix} 2 & 1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & \sqrt{5} \end{pmatrix}$$

Damit ist

$$G_1 A = \frac{1}{\sqrt{5}} \begin{pmatrix} 2 & 1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & \sqrt{5} \end{pmatrix} \begin{pmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 2 \end{pmatrix} = \frac{1}{\sqrt{5}} \begin{pmatrix} 5 & 6 & 1 \\ 0 & 7 & 2 \\ 0 & \sqrt{5} & 2\sqrt{5} \end{pmatrix}.$$

Im zweiten Schritt betrachten wir

$$\underline{a} = \frac{1}{\sqrt{5}} \begin{pmatrix} 7 \\ \sqrt{5} \end{pmatrix}, \quad x = \frac{7}{\sqrt{5}}, \quad y = 1, \quad \alpha = \frac{\frac{7}{\sqrt{5}}}{\sqrt{\frac{54}{5}}} = \frac{7}{\sqrt{54}}, \quad \beta = -\frac{\sqrt{5}}{\sqrt{54}},$$

d.h.

$$G_2 = \frac{1}{\sqrt{54}} \begin{pmatrix} \sqrt{54} & 0 & 0 \\ 0 & 7 & \sqrt{5} \\ 0 & -\sqrt{5} & 7 \end{pmatrix}.$$

Damit ergibt sich

$$\begin{aligned} R = G_2 G_1 A &= \frac{1}{\sqrt{220}} \begin{pmatrix} \sqrt{54} & 0 & 0 \\ 0 & 7 & \sqrt{5} \\ 0 & -\sqrt{5} & 7 \end{pmatrix} \begin{pmatrix} 5 & 6 & 1 \\ 0 & 7 & 2 \\ 0 & \sqrt{5} & 2\sqrt{5} \end{pmatrix} \\ &= \frac{1}{\sqrt{220}} \begin{pmatrix} 5\sqrt{54} & 6\sqrt{54} & \sqrt{54} \\ 0 & 54 & 24 \\ 0 & 0 & 12\sqrt{15} \end{pmatrix}. \end{aligned}$$

Somit folgt die Darstellung $A = QR$ mit der orthogonalen Matrix

$$Q = G_1^\top G_2^\top = \frac{1}{\sqrt{220}} \begin{pmatrix} 2 & -1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & \sqrt{5} \end{pmatrix} \begin{pmatrix} \sqrt{54} & 0 & 0 \\ 0 & 7 & -\sqrt{5} \\ 0 & \sqrt{5} & 7 \end{pmatrix}.$$

4.2 Stationäre Iterationsverfahren

Für eine reguläre Matrix $B \in \mathbb{R}^{n \times n}$ und einen positiven reellen Parameter $\alpha \in \mathbb{R}$ ist die Lösung des linearen Gleichungssystems (4.1) äquivalent zur Lösung der Fixpunktgleichung

$$\underline{x} = \underline{x} - \alpha B^{-1}(A\underline{x} - \underline{f}). \quad (4.2)$$

Diese Darstellung motiviert das Iterationsverfahren

$$\underline{x}^{k+1} := \underline{x}^k - \alpha B^{-1}(A\underline{x}^k - \underline{f}) = (I - \alpha B^{-1}A)\underline{x}^k + \alpha B^{-1}\underline{f}, \quad k = 0, 1, 2, \dots \quad (4.3)$$

Dabei ist $\underline{x}^0 \in \mathbb{R}^n$ eine beliebig gewählte Startnäherung. Die Konvergenz des Verfahrens (4.3) der sukzessiven Approximation folgt aus dem Banachschen Fixpunktsatz:

Satz 4.1. *Die Iterationsmatrix des Iterationsverfahrens (4.3) sei eine Kontraktion, das heißt es gilt*

$$\|I - \alpha B^{-1}A\|_M \leq q < 1 \quad (4.4)$$

in einer zu einer Vektornorm $\|\cdot\|_V$ verträglichen Matrixnorm $\|\cdot\|_M$. Dann konvergiert das Iterationsverfahren (4.3) der sukzessiven Approximation gegen die eindeutig bestimmte Lösung $\underline{x} = A^{-1}\underline{f}$ des linearen Gleichungssystems (4.1) und es gelten die a priori Fehlerabschätzung

$$\|\underline{x}^{k+1} - \underline{x}\|_V \leq \frac{q^{k+1}}{1-q} \|\underline{x}^1 - \underline{x}^0\|_V \quad (4.5)$$

sowie die a posteriori Fehlerabschätzung

$$\|\underline{x}^{k+1} - \underline{x}\|_V \leq \frac{q}{1-q} \|\underline{x}^{k+1} - \underline{x}^k\|_V. \quad (4.6)$$

Beweis: Die Lösung $\underline{x} = A^{-1}\underline{f}$ des linearen Gleichungssystems (4.1) ist Lösung der Fixpunktgleichung (4.2). Dann folgt

$$\begin{aligned} \|\underline{x}^{k+1} - \underline{x}\|_V &= \|(I - \alpha B^{-1}A)(\underline{x}^k - \underline{x})\|_V \\ &\leq \|I - \alpha B^{-1}A\|_M \|\underline{x}^k - \underline{x}\|_V \leq q \|\underline{x}^k - \underline{x}\|_V \end{aligned}$$

und durch wiederholtes Anwenden ergibt sich

$$\|\underline{x}^{k+1} - \underline{x}\|_V \leq q^{k+1} \|\underline{x}^0 - \underline{x}\|_V \rightarrow 0 \quad \text{für } k \rightarrow \infty$$

wegen $q < 1$ für jede beliebige Startnäherung $\underline{x}^0 \in \mathbb{R}^n$.

Mit der Dreiecksungleichung

$$\|\underline{x}^{k+1} - \underline{x}\|_V \leq q \|\underline{x}^k - \underline{x}\|_V \leq q (\|\underline{x}^k - \underline{x}^{k+1}\|_V + \|\underline{x}^{k+1} - \underline{x}\|_V)$$

folgt die a posteriori Fehlerabschätzung (4.6). Aus

$$\|\underline{x}^{k+1} - \underline{x}\|_V \leq q^k \|\underline{x}^1 - \underline{x}\|_V$$

und der a posteriori Fehlerabschätzung für $k = 0$,

$$\|\underline{x}^1 - \underline{x}\|_V \leq \frac{q}{1-q} \|\underline{x}^1 - \underline{x}^0\|_V,$$

ergibt sich schließlich die a priori Fehlerabschätzung (4.5). ■

Eine beliebige reguläre Matrix $A \in \mathbb{R}^{n \times n}$ gestattet die Darstellung

$$A = L + D + R \tag{4.7}$$

mit einer unteren Dreiecksmatrix L und einer oberen Dreiecksmatrix R ,

$$L = \begin{pmatrix} 0 & & & & \\ a_{21} & 0 & & & \\ a_{31} & a_{32} & 0 & & \\ \vdots & & \ddots & \ddots & \\ a_{n1} & \cdots & \cdots & a_{nn-1} & 0 \end{pmatrix}, \quad R = \begin{pmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1n} \\ & 0 & a_{23} & & a_{2n} \\ & & \ddots & \ddots & \vdots \\ & & & 0 & a_{n-1n} \\ & & & & 0 \end{pmatrix},$$

sowie einer Diagonalmatrix

$$D = \begin{pmatrix} a_{11} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{nn} \end{pmatrix}.$$

Wegen der Invertierbarkeit von A kann ohne Einschränkung der Allgemeinheit die Invertierbarkeit der Diagonalmatrix D vorausgesetzt werden. Dann ist das lineare Gleichungssystem (4.1) äquivalent zu der Fixpunktgleichung

$$D\underline{x} = \underline{f} - (L + R)\underline{x},$$

woraus das Jacobi-Verfahren (Gesamtschrittverfahren)

$$\underline{x}^{k+1} = D^{-1}[\underline{f} - (L + R)\underline{x}^k] = \underline{x}^k - D^{-1}(A\underline{x}^k - \underline{f}) \tag{4.8}$$

folgt.

Sei $\underline{x}^0 \in \mathbb{R}^n$ eine beliebig gegebene Startnäherung.

Für $k = 0, 1, 2, \dots$ berechne

$$\underline{r}^k = A\underline{x}^k - \underline{f}, \quad \varrho_k = (\underline{r}^k, \underline{r}^k) = \|\underline{r}^k\|_2^2.$$

Stoppe, wenn $\varrho_k \leq \varepsilon^2 \varrho_0$ mit einer vorgegebenen Fehlergenauigkeit ε erreicht ist. Andernfalls berechne die neue Näherungslösung

$$x_i^{k+1} = \frac{1}{a_{ii}} \left[f_i - \sum_{j=1}^{i-1} a_{ij} x_j^k - \sum_{j=i+1}^n a_{ij} x_j^k \right] \quad \text{für } i = 1, \dots, n.$$

Algorithmus 4.3: Jacobi-Verfahren (Gesamtschrittverfahren).

Satz 4.2. Für die Matrix A sei das strenge Zeilensummenkriterium

$$\max_{i=1,\dots,n} \sum_{j=1, j \neq i}^n \frac{|a_{ij}|}{|a_{ii}|} \leq q < 1$$

erfüllt. Dann konvergiert das Jacobi-Verfahren (4.8) für jede beliebige Startnäherung $\underline{x}^0$.

Beweis: In der zur Maximumnorm $\|\cdot\|_V = \|\cdot\|_\infty$ verträglichen Zeilensummennorm $\|\cdot\|_M = \|\cdot\|_\infty$ (vergleiche Lemma 3.2) folgt aus dem strengen Zeilensummenkriterium für die Iterationsmatrix

$$\|I - D^{-1}A\|_\infty = \max_{i=1,\dots,n} \sum_{j=1, j \neq i}^n \frac{|a_{ij}|}{|a_{ii}|} \leq q < 1$$

und somit nach Satz 4.1 die Konvergenz des Jacobi-Verfahrens (4.8). ■

Ausgehend von der Zerlegung (4.7) ist das lineare Gleichungssystem (4.1) auch äquivalent zu der Fixpunktgleichung

$$(D + L)\underline{x} = \underline{f} - R\underline{x},$$

woraus das (vorwärtige) Gauß-Seidel-Verfahren (Einzelschrittverfahren)

$$\underline{x}^{k+1} = (D + L)^{-1}[\underline{f} - R\underline{x}^k] = \underline{x}^k - (D + L)^{-1}[A\underline{x}^k - \underline{f}] \quad (4.9)$$

abgeleitet werden kann.

Sei $\underline{x}^0 \in \mathbb{R}^n$ eine beliebig gegebene Startnäherung.

Für $k = 0, 1, 2, \dots$ berechne

$$\underline{r}^k = A\underline{x}^k - \underline{f}, \quad \varrho_k = (\underline{r}^k, \underline{r}^k) = \|\underline{r}^k\|_2^2.$$

Stoppe, wenn $\varrho_k \leq \varepsilon^2 \varrho_0$ mit einer vorgegebenen Fehlergenauigkeit ε erreicht ist. Andernfalls berechne die neue Näherungslösung

$$x_i^{k+1} = \frac{1}{a_{ii}} \left[f_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i+1}^n a_{ij} x_j^k \right] \quad \text{für } i = 1, \dots, n.$$

Algorithmus 4.4: vorwärtiges Gauß-Seidel-Verfahren (Einzelschrittverfahren).

Satz 4.3. Für die Matrix A sei das strenge Zeilensummenkriterium wie in Satz 4.2 erfüllt. Dann konvergiert das Gauß-Seidel-Verfahren (4.9) für jede beliebige Startnäherung $\underline{x}^0$.

Beweis: Für einen beliebig gegebenen Vektor $\underline{y} \in \mathbb{R}^n$ entspricht die Berechnung von $\underline{z} = (D + L)^{-1}R\underline{y}$ der Lösung des linearen Gleichungssystems

$$(D + L)\underline{z} = R\underline{y}.$$

Dann ist

$$z_1 = \frac{1}{a_{11}} \sum_{j=2}^n a_{1j} y_j$$

und somit

$$|z_1| \leq \sum_{j=2}^n \frac{|a_{1j}|}{|a_{11}|} |y_j| \leq \max_{j=2,n} |y_j| \sum_{j=2}^n \frac{|a_{1j}|}{|a_{11}|} \leq q \|\underline{y}\|_\infty.$$

Es gelte also

$$|z_\ell| \leq q \|\underline{y}\|_\infty \leq \|\underline{y}\|_\infty \quad \text{für } \ell = 1, \dots, k-1.$$

Dann folgt

$$\begin{aligned} |z_k| &= \frac{1}{|a_{kk}|} \left| - \sum_{\ell=1}^{k-1} a_{k\ell} z_\ell + \sum_{\ell=k+1}^n a_{k\ell} y_\ell \right| \\ &\leq \frac{1}{|a_{kk}|} \left[\max_{\ell=1, \dots, k-1} |z_\ell| \sum_{\ell=1}^{k-1} |a_{k\ell}| + \max_{\ell=k+1, \dots, n} |y_\ell| \sum_{\ell=k+1}^n |a_{k\ell}| \right] \\ &\leq \|\underline{y}\|_\infty \sum_{\ell=1, \ell \neq k}^n \frac{|a_{k\ell}|}{|a_{kk}|} \leq q \|\underline{y}\|_\infty \end{aligned}$$

für alle $k = 2, \dots, n$, und somit gilt

$$\|(D + L)^{-1} R \underline{y}\|_\infty = \|\underline{z}\|_\infty \leq q \|\underline{y}\|_\infty.$$

Da die Zeilensummennorm durch die Maximumnorm induziert wird, folgt

$$\|(D + L)^{-1} R\|_\infty = \sup_{\underline{y} \in \mathbb{R}^n} \frac{\|(D + L)^{-1} R \underline{y}\|_\infty}{\|\underline{y}\|_\infty} \leq q < 1,$$

und nach Satz 4.1 ergibt sich die Konvergenz des Gauß–Seidel–Verfahrens. ■

Analog zu (4.9) kann das rückwärtige Gauß–Seidel–Verfahren

$$\underline{x}^{k+1} = (D + R)^{-1} [\underline{f} - L \underline{x}^k] = \underline{x}^k - (D + R)^{-1} [A \underline{x}^k - \underline{f}]$$

hergeleitet und analysiert werden.

Sei $\underline{x}^0 \in \mathbb{R}^n$ eine beliebig gegebene Startnäherung.

Für $k = 0, 1, 2, \dots$ berechne

$$\underline{r}^k = A \underline{x}^k - \underline{f}, \quad \varrho_k = (\underline{r}^k, \underline{r}^k) = \|\underline{r}^k\|_2^2.$$

Stoppe, wenn $\varrho_k \leq \varepsilon^2 \varrho_0$ mit einer vorgegebenen Fehlergenauigkeit ε erreicht ist. Andernfalls berechne die neue Näherungslösung

$$x_i^{k+1} = \frac{1}{a_{ii}} \left[f_i - \sum_{j=1}^{i-1} a_{ij} x_j^k - \sum_{j=i+1}^n a_{ij} x_j^{k+1} \right] \quad \text{für } i = n, n-1, \dots, 2, 1.$$

Algorithmus 4.5: rückwärtiges Gauß–Seidel–Verfahren (Einzelschrittverfahren).

Das Gauß–Seidel–Verfahren (Algorithmus 4.4) konvergiert auch für jede symmetrische und positiv definite Matrix A , siehe hierzu Satz 4.4 und Bemerkung 4.1.

Durch Einfügen eines Relaxationsparameters $\omega \in \mathbb{R}_+$ können die bisher betrachteten Iterationsverfahren abgeändert werden. Für $\omega > 1$ werden diese als Überrelaxationsverfahren, für $\omega < 1$ als Unterrelaxationsverfahren bezeichnet.

Für $\underline{x}^k \in \mathbb{R}^n$ sei zunächst $\hat{\underline{x}}^{k+1} \in \mathbb{R}^n$ die durch das Jacobi–Verfahren (4.8) erhaltene Näherungslösung mit

$$\hat{x}_i^{k+1} = \frac{1}{a_{ii}} \left[f_i - \sum_{j=1, j \neq i}^n a_{ij} x_j^k \right] \quad \text{für } i = 1, \dots, n.$$

Dann kann eine neue Näherungslösung $\underline{x}^{k+1}$ erklärt werden als Konvexkombination zwischen alter Näherungslösung $\underline{x}^k$, und der durch das klassische Jacobi–Verfahren berechneten Näherung $\hat{\underline{x}}^{k+1}$,

$$\begin{aligned} x_i^{k+1} &= (1 - \omega)x_i^k + \omega\hat{x}_i^{k+1} \\ &= (1 - \omega)x_i^k + \frac{\omega}{a_{ii}} \left[f_i - \sum_{j=1, j \neq i}^n a_{ij} x_j^k \right] \\ &= x_i^k - \frac{\omega}{a_{ii}} \left[\sum_{j=1}^n a_{ij} x_j^k - f_i \right]. \end{aligned}$$

Somit ergibt sich die Iterationsvorschrift des ω –Jacobi–Verfahren,

$$\underline{x}^{k+1} = \underline{x}^k - \omega D^{-1} [A \underline{x}^k - \underline{f}]. \quad (4.10)$$

Wird für $\underline{x}^k$ die neue Näherungslösung $\hat{\underline{x}}^{k+1}$ durch das vorwärtige Gauß–Seidel–Verfahren (4.9) berechnet und die Konvexkombination mit der alten Näherungslösung $\underline{x}^k$ gebildet,

$$\hat{x}_i^{k+1} = \frac{1}{a_{ii}} \left[f_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i+1}^n a_{ij} x_j^k \right], \quad x_i^{k+1} = (1 - \omega)x_i^k + \omega\hat{x}_i^{k+1},$$

dann ergibt sich

$$\begin{aligned} x_i^{k+1} &= (1 - \omega)x_i^k + \omega\hat{x}_i^{k+1} \\ &= (1 - \omega)x_i^k + \frac{\omega}{a_{ii}} \left[f_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i+1}^n a_{ij} x_j^k \right] \\ &= x_i^k + \frac{\omega}{a_{ii}} \left[f_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i}^n a_{ij} x_j^k \right]. \end{aligned}$$

In Vektorschreibweise ist dies gleichbedeutend mit

$$\underline{x}^{k+1} = \underline{x}^k + \omega D^{-1} [\underline{f} - L \underline{x}^{k+1} - (D + R) \underline{x}^k]$$

bzw. mit

$$(I + \omega D^{-1}L)\underline{x}^{k+1} = \underline{x}^k + \omega D^{-1}[\underline{f} - (D + R)\underline{x}^k].$$

Durch weiteres Umformen erhält man daraus mit

$$\underline{x}^{k+1} = \underline{x}^k - \omega(D + \omega L)^{-1}[A\underline{x}^k - \underline{f}] \quad (4.11)$$

die Iterationsvorschrift des (vorwärtigen) sukzessiven Überrelaxationsverfahrens (SOR Verfahren).

Satz 4.4. *Sei A eine symmetrische und positiv definite Matrix. Dann konvergiert das SOR-Verfahren (4.11) genau dann, wenn $0 < \omega < 2$ erfüllt ist.*

Beweis: Für eine symmetrische Matrix A kann die Spektralnorm $\|A\|_2$ mit dem Spektralradius $\varrho(A)$ identifiziert werden,

$$\|A\|_2 = \varrho(A) = \max_{k=1,\dots,n} |\lambda_k(A)|.$$

Zu untersuchen sind deshalb die Eigenwerte der Iterationsmatrix

$$M = I - \omega(D + \omega L)^{-1}A.$$

Sei $\lambda \in \mathbb{R}$ ein beliebiger Eigenwert von M und $\underline{z} \in \mathbb{R}^n$ der zugehörige Eigenvektor, d.h. es gilt

$$M\underline{z} = [I - \omega(D + \omega L)^{-1}A]\underline{z} = \lambda \underline{z}$$

bzw.

$$[(D + \omega L) - \omega A]\underline{z} = \lambda(D + \omega L)\underline{z}.$$

Mit der Zerlegung $A = D + L + L^\top$ der symmetrischen Matrix A folgt daraus

$$[(1 - \omega)D - \omega L^\top]\underline{z} = \lambda(D + \omega L)\underline{z}.$$

Durch Bilden des Euklidischen Skalarproduktes dieser Gleichheit mit $\underline{z}$ ergibt sich

$$(1 - \omega)(D\underline{z}, \underline{z}) - \omega(L^\top \underline{z}, \underline{z}) = \lambda[(D\underline{z}, \underline{z}) + \omega(L\underline{z}, \underline{z})].$$

Mit

$$(L^\top \underline{z}, \underline{z}) = (L\underline{z}, \underline{z}) = \frac{1}{2}[(L\underline{z}, \underline{z}) + (L^\top \underline{z}, \underline{z})] = \frac{1}{2}[(A\underline{z}, \underline{z}) - (D\underline{z}, \underline{z})]$$

und durch Multiplikation mit 2 folgt dann die Gleichheit

$$(2 - \omega)(D\underline{z}, \underline{z}) - \omega(A\underline{z}, \underline{z}) = \lambda[(2 - \omega)(D\underline{z}, \underline{z}) + \omega(A\underline{z}, \underline{z})]$$

und somit die Darstellung

$$\lambda = \frac{(2 - \omega)(D\underline{z}, \underline{z}) - \omega(A\underline{z}, \underline{z})}{(2 - \omega)(D\underline{z}, \underline{z}) + \omega(A\underline{z}, \underline{z})}.$$

Aus der positiven Definitheit von A folgt dabei

$$d := (D\underline{z}, \underline{z}) > 0, \quad a := (A\underline{z}, \underline{z}) > 0$$

und die Forderung $|\lambda| < 1$ ist genau dann erfüllt, wenn $\omega \in (0, 2)$ gilt. ■

Bemerkung 4.1. Aus der Konvergenz des SOR-Verfahrens für $\omega = 1$ folgt insbesondere die Konvergenz des Gauß-Seidel-Verfahrens (4.9) für symmetrische und positiv definite Matrizen A .

Analog zum vorwärtigen SOR-Verfahren lautet die Iterationsvorschrift des rückwärtigen SOR-Verfahrens

$$\underline{x}^{k+1} := \underline{x}^k - \omega(D + \omega R)^{-1}[A\underline{x}^k - \underline{f}]. \quad (4.12)$$

Man bemerkt, daß beim SOR-Verfahren selbst für symmetrische Matrizen A eine im allgemeinen nichtsymmetrische Matrix $B = (D + \omega L)$ entsteht. Durch Verknüpfung des vorwärtigen SOR-Verfahrens (4.11) mit der rückwärtigen Variante (4.12),

$$\begin{aligned} \underline{x}^{k+1/2} &= \underline{x}^k - \omega(D + \omega L)^{-1}[A\underline{x}^k - \underline{f}], \\ \underline{x}^{k+1} &= \underline{x}^{k+1/2} - \omega(D + \omega R)^{-1}[A\underline{x}^{k+1/2} - \underline{f}] \end{aligned}$$

entsteht die Iterationsvorschrift des symmetrischen sukzessiven Überrelaxationsverfahrens (SSOR Verfahren),

$$\underline{x}^{k+1} = \underline{x}^k - \omega(2 - \omega)(D + \omega R)^{-1}D(D + \omega L)^{-1}[A\underline{x}^k - \underline{f}]. \quad (4.13)$$

Für eine allgemeinere Konvergenzuntersuchung des stationären Iterationsverfahrens (4.3) der sukzessiven Approximation wird nun zunächst der Fall $B = I$ betrachtet,

$$\underline{x}^{k+1} = \underline{x}^k - \alpha[A\underline{x}^k - \underline{f}]. \quad (4.14)$$

Die Matrix A sei dabei positiv definit,

$$(A\underline{x}, \underline{x}) \geq c_1^A \|\underline{x}\|_2^2 \quad \text{für alle } \underline{x} \in \mathbb{R}^n, \quad (4.15)$$

und beschränkt,

$$\|A\underline{x}\|_2 \leq c_2^A \|\underline{x}\|_2 \quad \text{für alle } \underline{x} \in \mathbb{R}^n. \quad (4.16)$$

Satz 4.5. Sei A eine beschränkte und positiv definite Matrix, das heißt es gelten die Ungleichungen (4.16) und (4.15). Dann konvergiert das Iterationsverfahren (4.14) für den Relaxationsparameter

$$0 < \alpha < 2 \frac{c_1^A}{[c_2^A]^2}$$

und die optimale Konvergenzrate wird angenommen für

$$\alpha^* = \frac{c_1^A}{[c_2^A]^2}.$$

Dann gilt die Fehlerabschätzung

$$\|\underline{x}^{k+1} - \underline{x}\|_2^2 \leq \left[1 - \left(\frac{c_1^A}{c_2^A}\right)^2\right]^{k+1} \|\underline{x}^0 - \underline{x}\|_2^2.$$

Beweis: Durch Übergang zum Euklidischen Skalarprodukt folgt mit den Voraussetzungen (4.15) und (4.16) für die Iterationsvorschrift (4.14)

$$\begin{aligned}
\|\underline{x}^{k+1} - \underline{x}\|_2^2 &= \|(I - \alpha A)(\underline{x}^k - \underline{x})\|_2^2 \\
&= ((I - \alpha A)(\underline{x}^k - \underline{x}), (I - \alpha A)(\underline{x}^k - \underline{x})) \\
&= (\underline{x}^k - \underline{x}, \underline{x}^k - \underline{x}) - 2\alpha (A(\underline{x}^k - \underline{x}), \underline{x}^k - \underline{x}) + \alpha^2 (A(\underline{x}^k - \underline{x}), A(\underline{x}^k - \underline{x})) \\
&= \|\underline{x}^k - \underline{x}\|_2^2 - 2\alpha (A(\underline{x}^k - \underline{x}), \underline{x}^k - \underline{x}) + \alpha^2 \|A(\underline{x}^k - \underline{x})\|_2^2 \\
&\leq (1 - 2\alpha c_1^A + \alpha^2 [c_2^A]^2) \|\underline{x}^k - \underline{x}\|_2^2.
\end{aligned}$$

Damit ergibt sich Konvergenz für

$$1 - 2\alpha c_1^A + \alpha^2 [c_2^A]^2 < 1$$

und somit für

$$0 < \alpha < 2 \frac{c_1^A}{[c_2^A]^2}.$$

Insbesondere für

$$\alpha^* = \frac{c_1^A}{[c_2^A]^2}$$

ergibt sich mit

$$\|\underline{x}^{k+1} - \underline{x}\|_2^2 \leq \left[1 - \left(\frac{c_1^A}{c_2^A} \right)^2 \right] \|\underline{x}^k - \underline{x}\|_2^2$$

die optimale Konvergenzrate. ■

In vielen Anwendungen ist das Verhältnis c_1^A/c_2^A abhängig von der Dimension n der Systemmatrix $A \in \mathbb{R}^{n \times n}$. Insbesondere bei numerischen Näherungsverfahren für gewöhnliche oder partielle Differentialgleichungen gilt oft $c_1^A/c_2^A \rightarrow 0$ für $n \rightarrow \infty$, und somit ergibt sich eine schlechtere Konvergenzrate für großdimensionierte Gleichungssysteme. Durch eine geeignet gewählte Transformation des linearen Gleichungssystems $A\underline{x} = \underline{f}$ sollen Verfahren hergeleitet werden, deren Konvergenzverhalten möglichst unabhängig von der betrachteten Problemgröße, das heißt unabhängig von der Dimension n sind.

Sei $B = B^\top \in \mathbb{R}^{n \times n}$ eine zunächst beliebige symmetrische und positiv definite Matrix mit positiven Eigenwerten $\lambda_k(B)$. Für B gilt dann die Faktorisierung (3.14),

$$B = VDV^\top, \quad D = \text{diag}(\lambda_k(B))_{k=1}^n,$$

mit einer orthogonalen Matrix V , welche durch die zueinander orthogonalen Eigenvektoren von B gebildet wird. Dann ist

$$B^{1/2} = VD^{1/2}V^\top, \quad D^{1/2} = \text{diag}(\sqrt{\lambda_k(B)})_{k=1}^n,$$

und es gilt $B = B^{1/2}B^{1/2}$. Weiter bezeichne $B^{-1/2}$ die inverse Matrix von $B^{1/2}$.

Anstelle des linearen Gleichungssystems $A\underline{x} = \underline{f}$ wird jetzt das dazu äquivalente System

$$B^{-1/2}AB^{-1/2}B^{1/2}\underline{x} = B^{-1/2}\underline{f}$$

betrachtet. Mit

$$\tilde{A} = B^{-1/2}AB^{-1/2}, \quad \tilde{\underline{x}} = B^{1/2}\underline{x}, \quad \tilde{\underline{f}} = B^{-1/2}\underline{f} \quad (4.17)$$

kann zur Lösung des transformierten Systems $\tilde{A}\tilde{\underline{x}} = \tilde{\underline{f}}$ das Iterationsverfahren (4.14) verwendet werden,

$$\tilde{\underline{x}}^{k+1} = \tilde{\underline{x}}^k - \alpha[\tilde{A}\tilde{\underline{x}}^k - \tilde{\underline{f}}].$$

Einsetzen der Transformationen (4.17) und Multiplikation mit $B^{-1/2}$ liefert das vorkonditionierte Iterationsverfahren der sukzessiven Approximation,

$$\underline{x}^{k+1} = \underline{x}^k - \alpha B^{-1}[A\underline{x}^k - \underline{f}]. \quad (4.18)$$

Die Konvergenz des vorkonditionierten Iterationsverfahrens (4.18) folgt nach Satz 4.5 für $\alpha \in (0, 2c_1^{\tilde{A}}/[c_2^{\tilde{A}}]^2)$ mit

$$(\tilde{A}\tilde{\underline{x}}, \tilde{\underline{x}}) \geq c_1^{\tilde{A}} \|\tilde{\underline{x}}\|_2^2, \quad \|\tilde{A}\tilde{\underline{x}}\|_2 \leq c_2^{\tilde{A}} \|\tilde{\underline{x}}\|_2 \quad \text{für alle } \tilde{\underline{x}} \in \mathbb{R}^n.$$

Aus Satz 4.5 ergibt sich die Konvergenzabschätzung

$$\|\tilde{\underline{x}}^{k+1} - \tilde{\underline{x}}\|_2^2 \leq \left[1 - \left(\frac{c_1^{\tilde{A}}}{c_2^{\tilde{A}}} \right)^2 \right]^{k+1} \|\tilde{\underline{x}}^0 - \tilde{\underline{x}}\|_2^2.$$

Einsetzen der Transformationen (4.17) liefert unter Berücksichtigung von

$$\|\tilde{\underline{x}}\|_2^2 = (\tilde{\underline{x}}, \tilde{\underline{x}}) = (B^{1/2}\underline{x}, B^{1/2}\underline{x}) = (B\underline{x}, \underline{x})$$

und

$$\|\tilde{A}\tilde{\underline{x}}\|_2^2 = \|B^{-1/2}A\underline{x}\|_2^2 = (B^{-1/2}A\underline{x}, B^{-1/2}A\underline{x}) = (B^{-1}A\underline{x}, A\underline{x})$$

die Fehlerabschätzung

$$\|\underline{x}^{k+1} - \underline{x}\|_B^2 \leq \left[1 - \left(\frac{c_1^{\tilde{A}}}{c_2^{\tilde{A}}} \right)^2 \right]^{k+1} \|\underline{x}^0 - \underline{x}\|_B^2$$

sowie die Ungleichungen

$$(A\underline{x}, \underline{x}) \geq c_1^{\tilde{A}} (B\underline{x}, \underline{x}), \quad (B^{-1}A\underline{x}, A\underline{x}) \leq [c_2^{\tilde{A}}]^2 (B\underline{x}, \underline{x}) \quad (4.19)$$

für alle $\underline{x} \in \mathbb{R}^n$. Die Matrix B wird als Vorkonditionierung zu A bezeichnet, falls die Ungleichungen (4.19) mit Konstanten $c_i^{\tilde{A}}$ unabhängig von der Dimension n erfüllt sind, und falls die Anwendung der inversen Matrix B^{-1} effizient realisiert werden kann. Für eine positiv definite Matrix A ergibt sich mit $B = \text{diag } A$ gerade das ω -Jacobi-Verfahren (4.10), wobei mit den Ungleichungen (4.19) die Konstanten $c_i^{\tilde{A}}$ und somit die zulässigen Relaxationsparameter fixiert werden können.

Beispiel 4.5. Für das aus einer FEM-Approximation resultierende lineare Gleichungssystem $A_h \underline{u} = \underline{f}$ mit der Steifigkeitsmatrix

$$A_h = \frac{1}{h} \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{(n-1) \times (n-1)}$$

mit den durch

$$\underline{v}_i^\ell = \sin \ell \frac{i\pi}{n} \quad \text{für } i = 1, \dots, n-1, \ell = 1, \dots, n-1,$$

gegebenen Eigenvektoren $\underline{v}^\ell \in \mathbb{R}^{n-1}$ und dem zugehörigen Eigenwert

$$\lambda_\ell(A_h) = \frac{4}{h} \sin^2 \frac{\ell\pi}{2n}$$

wird das ω -Jacobi-Verfahren betrachtet,

$$\underline{u}^{k+1} = \underline{u}^k - \omega D_h^{-1} (A_h \underline{u}^k - \underline{f}).$$

Dabei ist $D_h = \text{diag } A_h = 2h^{-1} I_{n-1}$. Für den zugehörigen Fehler ergibt sich

$$\underline{u}^{k+1} - \underline{u} = (I - \omega \frac{h}{2} A_h) (\underline{u}^k - \underline{u}) = (I - \omega \frac{h}{2} A_h)^{k+1} (\underline{u}^0 - \underline{u}).$$

Wird der Anfangsfehler $\underline{e}^0 = \underline{u}^0 - \underline{u}$ in die Eigenvektoren $\underline{v}^\ell$ der Steifigkeitsmatrix A_h entwickelt,

$$\underline{e}^0 = \sum_{\ell=1}^{n-1} \alpha_\ell \underline{v}^\ell,$$

dann folgt

$$\begin{aligned} \underline{e}^{k+1} &= \sum_{\ell=1}^{n-1} \alpha_\ell \left(I - \omega \frac{h}{2} A_h \right)^{k+1} \underline{v}^\ell = \sum_{\ell=1}^{n-1} \alpha_\ell \left[1 - \omega \frac{h}{2} \lambda_\ell(A_h) \right]^{k+1} \underline{v}^\ell \\ &= \sum_{\ell=1}^{n-1} \alpha_\ell \left[1 - 2\omega \sin^2 \frac{\ell\pi}{2n} \right]^{k+1} \underline{v}^\ell. \end{aligned}$$

In diesem Fall konvergiert das ω -Jacobi-Verfahren für alle $\omega \in (0, 1]$, wobei die Konvergenzrate durch den zu $\ell = 1$ gehörenden Eigenwert

$$1 - 2\omega \sin^2 \frac{\pi}{2n} = 1 - 2\omega \sin^2 \frac{\pi h}{2} = 1 - \mathcal{O}(h^2)$$

dominiert wird. Andererseits ist für $\ell = n-1$

$$1 - 2\omega \sin^2 \frac{\ell\pi}{2n} = 1 - 2\omega \sin^2 \frac{(n-1)\pi}{2n} = 1 - 2\omega - \mathcal{O}(h^2),$$

bzw. für $\ell = n/2$

$$1 - 2\omega \sin^2 \frac{\ell\pi}{2n} = 1 - 2\omega \sin^2 \frac{\pi}{4} = 1 - 2\omega \left(\frac{\sqrt{2}}{2} \right)^2 = 1 - \omega.$$

Für

$$\omega = \frac{2}{3}$$

folgt dann

$$\left| 1 - 2\omega \sin^2 \frac{\ell\pi}{2n} \right| \leq \frac{1}{3} \quad \text{für } \ell = \frac{n}{2}, \dots, n-1.$$

In diesem Fall werden bei der Anwendung des ω -Jacobi-Verfahrens mit $\omega = 2/3$ die Koeffizienten α_ℓ des Anfangsfehlers $\underline{e}^0$ mit $\ell = n/2, \dots, n-1$ besonders stark gedämpft. Diese entsprechen den hochfrequenten Eigenvektoren der Steifigkeitsmatrix A_h . Die verbleibenden Koeffizienten α_ℓ für $\ell = 1, \dots, n/2 - 1$, d.h. für die niedrigfrequenten Eigenvektoren, entsprechen aber gerade wieder dem Ausgangsproblem für eine Diskretisierung mit doppelter Schrittweite. Mit anderen Worten, der verbleibende Fehler nach einigen Schritten des ω -Jacobi-Verfahrens kann auf einem gröberen Gitter dargestellt werden. Die rekursive Anwendung führt dann auf Mehrgitterverfahren [4, 7].

Die Wahl der zulässigen und optimalen Relaxationsparameter α des vorkonditionierten Iterationsverfahrens (4.18) hängt gemäß Satz 4.5 wesentlich von den Konstanten $c_i^{\tilde{A}}$ der Ungleichungen (4.19) ab. Diese sind aber im allgemeinen unbekannt bzw. können nur näherungsweise bestimmt werden. Deshalb sollen nun Verfahren betrachtet werden, die die Berechnung eines optimalen Relaxationsparameters α in jedem Iterationsschritt beinhalten.

4.3 Das Gradientenverfahren des steilsten Abstiegs

Die Lösung des linearen Gleichungssystems $A\underline{x} = \underline{f}$ ist äquivalent zur Minimierung eines Funktionals

$$F(\underline{z}) = \|A\underline{z} - \underline{f}\|_V^2$$

in einer geeigneten Vektornorm $\|\cdot\|_V$. Ist die Matrix A symmetrisch und positiv definit, so kann das Funktional

$$\begin{aligned} F(\underline{z}) &= \|A\underline{z} - \underline{f}\|_{A^{-1}}^2 = (A^{-1}(A\underline{z} - \underline{f}), A\underline{z} - \underline{f})_2 = (\underline{z} - \underline{x}, A\underline{z} - \underline{f})_2 \\ &= (\underline{z} - \underline{x}, A(\underline{z} - \underline{x}))_2 = (A\underline{z}, \underline{z})_2 - 2(A\underline{x}, \underline{z})_2 + (A\underline{x}, \underline{x})_2 \\ &= \|\underline{z} - \underline{x}\|_A^2 = (A\underline{z}, \underline{z})_2 - 2(\underline{f}, \underline{z})_2 + \|\underline{x}\|_A^2 \end{aligned}$$

gewählt werden. Für die Auswertung des Gradienten von F für eine gegebene Näherungslösung $\underline{x}^k$ zur Minimierung des Funktionals ergibt sich

$$\nabla F(\underline{z})|_{\underline{z}=\underline{x}^k} = 2(A\underline{z} - \underline{f})|_{\underline{z}=\underline{x}^k} = 2(A\underline{x}^k - \underline{f}) = 2\underline{r}^k.$$

Zur Bestimmung der neuen Näherungslösung

$$\underline{x}^{k+1} = \underline{x}^k - \alpha_k \underline{r}^k$$

bleibt das Funktional $F(\underline{x}^{k+1})$ bezüglich des reellen Parameters α_k zu minimieren,

$$F(\underline{x}^{k+1}) = F(\underline{x}^k - \alpha_k \underline{r}^k) = \min_{\alpha \in \mathbb{R}} F(\underline{x}^k - \alpha \underline{r}^k).$$

Wegen

$$\begin{aligned} F(\underline{x}^k - \alpha \underline{r}^k) &= \|\underline{x}^k - \alpha \underline{r}^k - \underline{x}\|_A^2 \\ &= (A(\underline{x}^k - \underline{x} - \alpha \underline{r}^k), \underline{x}^k - \underline{x} - \alpha \underline{r}^k) \\ &= (A(\underline{x}^k - \underline{x}), \underline{x}^k - \underline{x}) - 2\alpha (A(\underline{x}^k - \underline{x}), \underline{r}^k) + \alpha^2 (A\underline{r}^k, \underline{r}^k) \\ &= F(\underline{x}^k) - 2\alpha (\underline{r}^k, \underline{r}^k) + \alpha^2 (A\underline{r}^k, \underline{r}^k) \end{aligned}$$

wird das Minimum angenommen für

$$\alpha_k = \frac{(\underline{r}^k, \underline{r}^k)}{(A\underline{r}^k, \underline{r}^k)}.$$

Das resultierende Verfahren ist das Gradientenverfahren des steilsten Abstiegs.

Für eine beliebig gegebene Startnäherung $\underline{x}^0 \in \mathbb{R}^n$ sei $\underline{r}^0 = A\underline{x}^0 - \underline{f}$.
 Für $k = 0, 1, 2, \dots$ berechne
 $\varrho_k = (\underline{r}^k, \underline{r}^k) = \|\underline{r}^k\|_2^2$.
 Stoppe, wenn $\varrho_k \leq \varepsilon^2 \varrho_0$ mit einer vorgegebenen Fehlergenauigkeit ε erreicht ist. Andernfalls bestimme die neue Näherungslösung:
 $\underline{v}^k = A\underline{r}^k, \quad \alpha_k = \frac{(\underline{r}^k, \underline{r}^k)}{(\underline{v}^k, \underline{r}^k)},$
 $\underline{x}^{k+1} = \underline{x}^k - \alpha_k \underline{r}^k, \quad \underline{r}^{k+1} = \underline{r}^k - \alpha_k \underline{v}^k.$

Algorithmus 4.6: Gradientenverfahren des steilsten Abstiegs.

Satz 4.6. Sei A symmetrisch und positiv definit und gelte

$$(A\underline{x}, \underline{x}) \geq c_1^A \|\underline{x}\|_2^2, \quad \|A\underline{x}\|_2 \leq c_2^A \|\underline{x}\|_2$$

für alle $\underline{x} \in \mathbb{R}^n$. Dann konvergiert das Gradientenverfahren des steilsten Abstiegs mit

$$\|\underline{x}^{k+1} - \underline{x}\|_A^2 \leq \left[1 - \left(\frac{c_1^A}{c_2^A} \right)^2 \right]^{k+1} \|\underline{x}^0 - \underline{x}\|_A^2.$$

Beweis: Für die symmetrische und positiv definite Matrix A gilt die Faktorisierung $A = VDV^\top$ mit der durch die orthogonalen Eigenvektoren von A gebildeten Matrix V und mit der Diagonalmatrix $D = \text{diag}(\lambda_k(A))_{k=1}^n$ mit den positiven Eigenwerten $\lambda_k(A)$ von A . Dann ist $A = A^{1/2}A^{1/2}$ mit der symmetrischen und positiv definiten Matrix $A^{1/2} = VD^{1/2}V^\top$, $D^{1/2} = \text{diag}(\sqrt{\lambda_k(A)})_{k=1}^n$.

Aus den Voraussetzungen an die symmetrische und positiv definite Matrix A folgt dann

$$\begin{aligned} (\underline{r}^k, \underline{x}^k - \underline{x})_A &= (\underline{r}^k, A(\underline{x}^k - \underline{x})) = (A(\underline{x}^k - \underline{x}), A(\underline{x}^k - \underline{x})) \\ &= (AA^{1/2}(\underline{x}^k - \underline{x}), A^{1/2}(\underline{x}^k - \underline{x})) \\ &\geq c_1^A \|A^{1/2}(\underline{x}^k - \underline{x})\|_2^2 = c_1^A \|\underline{x}^k - \underline{x}\|_A^2 \end{aligned}$$

sowie

$$\begin{aligned} \|\underline{r}^k\|_A^2 &= (A\underline{r}^k, \underline{r}^k) = (AA(\underline{x}^k - \underline{x}), A(\underline{x}^k - \underline{x})) \\ &= (AA^{1/2}(\underline{x}^k - \underline{x}), AA^{1/2}(\underline{x}^k - \underline{x})) \\ &= \|AA^{1/2}(\underline{x}^k - \underline{x})\|_2^2 \leq (c_2^A)^2 \|A^{1/2}(\underline{x}^k - \underline{x})\|_2^2 = (c_2^A)^2 \|\underline{x}^k - \underline{x}\|_A^2. \end{aligned}$$

Dann ist

$$\begin{aligned} \|\underline{x}^k - \alpha \underline{r}^k - \underline{x}\|_A^2 &= \|\underline{x}^k - \underline{x}\|_A^2 - 2\alpha(\underline{r}^k, \underline{x}^k - \underline{x})_A + \alpha^2 \|\underline{r}^k\|_A^2 \\ &\leq [1 - 2\alpha c_1^A + \alpha^2 (c_2^A)^2] \|\underline{x}^k - \underline{x}\|_A^2. \end{aligned}$$

Für

$$\hat{\underline{x}}^{k+1} = \underline{x}^k - \hat{\alpha}_k \underline{r}^k, \quad \hat{\alpha}_k = \frac{c_1^A}{[c_2^A]^2}$$

gilt dann

$$\|\hat{\underline{x}}^{k+1} - \underline{x}\|_A^2 \leq \left[1 - \left(\frac{c_1^A}{c_2^A}\right)^2\right] \|\underline{x}^k - \underline{x}\|_A^2.$$

Die Behauptung folgt schließlich aus

$$\|\underline{x}^{k+1} - \underline{x}\|_A^2 = \min_{\alpha \in \mathbb{R}} \|\underline{x}^k - \alpha \underline{r}^k - \underline{x}\|_A^2 \leq \|\hat{\underline{x}}^{k+1} - \underline{x}\|_A^2. \quad \blacksquare$$

Beispiel 4.6. Die Anwendung des Gradientenverfahrens des steilsten Abstiegs soll am Beispiel der iterativen Lösung des linearen Gleichungssystems

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 4 \end{pmatrix}$$

mit der exakten Lösung $x_1 = 2$ und $x_2 = 1$ veranschaulicht werden. Für das zu minimierende Funktional ergibt sich dann

$$F(\underline{z}) = \|\underline{z} - \underline{x}\|_A^2 = \left(\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} z_1 - 2 \\ z_2 - 1 \end{pmatrix}, \begin{pmatrix} z_1 - 2 \\ z_2 - 1 \end{pmatrix} \right).$$

In Tabelle 4.1 sind die Näherungslösungen $\underline{x}^k$ und die zugehörigen Werte des Funktionals $F(\underline{x}^k)$ angegeben.

k	x_1^k	x_2^k	$F(\underline{x}^k)$
0	0.000	0.000	14
1	1.680	1.344	2.213 -1
2	1.968	0.984	3.499 -3
3	1.995	1.005	5.530 -5

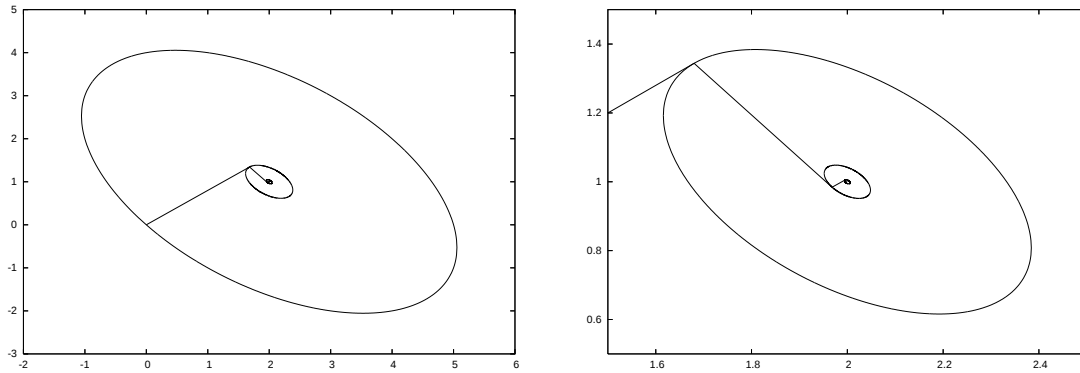
Tabelle 4.1: Näherungslösungen $\underline{x}^k$.

Abbildung 4.1:

Konvergenzverhalten des Gradientenverfahrens des steilsten Abstiegs.

Die Näherungslösungen $\underline{x}^k$ und die dadurch beschriebenen Äquipotentiallinien des Funktionals $F(\underline{z}) = F(\underline{x}^k)$ veranschaulichen in Abbildung 4.1 einen Nachteil von Gradientenverfahren: Zueinander parallele oder fast parallele Suchrichtungen können während der Iteration mehrmals durchlaufen werden. Dies motiviert die Orthogonalisierung von Suchrichtungen.

4.4 Das Verfahren konjugierter Gradienten

Sei $A = A^\top > 0$ symmetrisch und positiv definit. Ein System von linear unabhängigen Vektoren $\{\underline{p}^k\}_{k=0}^{n-1}$ heißt A -orthogonal oder konjugiert, falls

$$(A\underline{p}^k, \underline{p}^\ell) = 0 \quad \text{für } k, \ell = 0, \dots, n-1 \text{ und } k \neq \ell$$

sowie

$$(A\underline{p}^k, \underline{p}^k) > 0 \quad \text{für } k = 0, \dots, n-1$$

erfüllt ist. Die letzte Forderung folgt dabei unmittelbar aus der positiven Definitheit von A . Für ein zunächst beliebig gegebenes System linear unabhängiger Vektoren $\{\underline{w}^k\}_{k=0}^{n-1}$ kann mittels des Gram-Schmidt'schen-Orthogonalisierungsverfahrens ein System orthogonaler Vektoren $\{\underline{p}^k\}_{k=0}^{n-1}$ konstruiert werden:

Setze $\underline{p}^0 := \underline{w}^0$ Für $k = 0, \dots, n-2$ berechne $\underline{p}^{k+1} := \underline{w}^{k+1} - \sum_{\ell=0}^k \beta_{k,\ell} \underline{p}^\ell, \quad \beta_{k,\ell} = \frac{(A\underline{w}^{k+1}, \underline{p}^\ell)}{(A\underline{p}^\ell, \underline{p}^\ell)}$
--

Algorithmus 4.7: Konstruktion A -orthogonaler Vektoren.

Einsetzen des Lösungsansatzes

$$\underline{x} = \underline{x}^0 - \sum_{\ell=0}^{n-1} \alpha_\ell \underline{p}^\ell$$

in das lineare Gleichungssystem $A\underline{x} = \underline{f}$ ergibt

$$A\underline{x} = A\underline{x}^0 - \sum_{\ell=0}^{n-1} \alpha_\ell A\underline{p}^\ell = \underline{f}.$$

Das Bilden des Euklidischen Skalarproduktes mit $\underline{p}^k$ liefert wegen der A -Orthogonalität des Vektorsystems $\{\underline{p}^k\}_{k=0}^{n-1}$

$$(A\underline{x}^0, \underline{p}^k) - \sum_{\ell=0}^{n-1} \alpha_\ell \underbrace{(A\underline{p}^\ell, \underline{p}^k)}_{=0 \text{ für } \ell \neq k} = (\underline{f}, \underline{p}^k)$$

und somit

$$\alpha_k = \frac{(A\underline{x}^0 - \underline{f}, \underline{p}^k)}{(A\underline{p}^k, \underline{p}^k)} \quad \text{für } k = 0, \dots, n-1.$$

Für die Näherungslösung

$$\underline{x}^{k+1} = \underline{x}^0 - \sum_{\ell=0}^k \alpha_\ell \underline{p}^\ell = \underline{x}^0 - \sum_{\ell=0}^{k-1} \alpha_\ell \underline{p}^\ell - \alpha_k \underline{p}^k = \underline{x}^k - \alpha_k \underline{p}^k$$

ist das zugehörige Residuum gegeben durch

$$\underline{r}^{k+1} = A\underline{x}^{k+1} - \underline{f} = A\underline{x}^0 - \sum_{\ell=0}^k \alpha_\ell A\underline{p}^\ell - \underline{f} = A\underline{x}^k - \underline{f} - \alpha_k A\underline{p}^k = \underline{r}^k - \alpha_k A\underline{p}^k$$

und wegen

$$(A\underline{p}^\ell, \underline{p}^k) = 0 \quad \text{für } k \neq \ell$$

folgt

$$(A\underline{x}^0 - \underline{f}, \underline{p}^k) = (A\underline{x}^0 - \sum_{\ell=0}^{k-1} \alpha_\ell A\underline{p}^\ell - \underline{f}, \underline{p}^k) = (\underline{r}^k, \underline{p}^k)$$

und somit

$$\alpha_k = \frac{(\underline{r}^k, \underline{p}^k)}{(A\underline{p}^k, \underline{p}^k)}.$$

Damit ergibt sich für $k = 0, \dots, n-2$ die Iterationsvorschrift

$$\underline{x}^{k+1} = \underline{x}^k - \alpha_k \underline{p}^k, \quad \underline{r}^{k+1} = \underline{r}^k - \alpha_k A\underline{p}^k, \quad \alpha_k = \frac{(\underline{r}^k, \underline{p}^k)}{(A\underline{p}^k, \underline{p}^k)}.$$

Nach Konstruktion gilt

$$(\underline{r}^{k+1}, \underline{p}^k) = (\underline{r}^k - \alpha_k A\underline{p}^k, \underline{p}^k) = 0$$

für alle $k = 0, \dots, n-2$ und somit für $k = 1$ die Induktionsvoraussetzung

$$(\underline{r}^k, \underline{p}^\ell) = (\underline{r}^1, \underline{p}^0) = 0 \quad \text{für } \ell = 0, \dots, k-1.$$

Durch vollständige Induktion folgt, daß dann auch

$$(\underline{r}^{k+1}, \underline{p}^\ell) = 0 \quad \text{für alle } \ell = 0, \dots, k$$

für alle $k = 2, \dots, n-1$ gilt. Zunächst ist

$$(\underline{r}^{k+1}, \underline{p}^k) = 0$$

und für $\ell < k$ folgt aus der Induktionsvoraussetzung und der A -Orthogonalität

$$(\underline{r}^{k+1}, \underline{p}^\ell) = (\underline{r}^k - \alpha_k A\underline{p}^k, \underline{p}^\ell) = (\underline{r}^k, \underline{p}^\ell) - \alpha_k (A\underline{p}^k, \underline{p}^\ell) = 0.$$

Insgesamt gilt also

$$(\underline{r}^{k+1}, \underline{p}^\ell) = 0 \quad \text{für } \ell = 0, \dots, k \text{ und } k = 0, \dots, n-2.$$

Aus der Konstruktion der Suchrichtungen,

$$\underline{p}^\ell = \underline{w}^\ell - \sum_{j=0}^{\ell-1} \beta_{\ell-1,j} \underline{p}^j \quad \text{bzw.} \quad \underline{w}^\ell = \underline{p}^\ell + \sum_{j=0}^{\ell-1} \beta_{\ell-1,j} \underline{p}^j,$$

folgt weiterhin

$$(\underline{r}^{k+1}, \underline{w}^\ell) = (\underline{r}^{k+1}, \underline{p}^\ell) + \sum_{j=0}^{\ell-1} \beta_{\ell-1,j} (\underline{r}^{k+1}, \underline{p}^j) = 0 \quad \text{für } \ell = 0, \dots, k.$$

Damit ist das Residuum $\underline{r}^{k+1}$ orthogonal zu allen Basisvektoren $\underline{w}^\ell, \ell = 0, \dots, k$. Das Vektorsystem

$$\{\underline{w}^0, \underline{w}^1, \dots, \underline{w}^k, \underline{r}^{k+1}\}$$

ist also linear unabhängig, so daß die neue Suchrichtung als

$$\underline{w}^{k+1} = \underline{r}^{k+1} \quad \text{bzw.} \quad \underline{w}^\ell = \underline{r}^\ell \quad \text{für } \ell = 0, \dots, n-1$$

gewählt werden kann. Damit folgt

$$(\underline{r}^{k+1}, \underline{p}^\ell) = (\underline{r}^{k+1}, \underline{r}^\ell) = 0 \quad \text{für } \ell = 0, \dots, k \text{ und } k = 0, \dots, n-2.$$

Für das Orthogonalisierungsverfahren nach Gram-Schmidt ergibt sich dann

$$\underline{p}^0 = \underline{w}^0 = \underline{r}^0, \quad \underline{p}^{k+1} = \underline{r}^{k+1} - \sum_{\ell=0}^k \beta_{k\ell} \underline{p}^\ell \quad \text{für } k = 0, \dots, n-2$$

mit

$$\beta_{k\ell} = \frac{(A\underline{w}^{k+1}, \underline{p}^\ell)}{(A\underline{p}^\ell, \underline{p}^\ell)} = \frac{(\underline{r}^{k+1}, A\underline{p}^\ell)}{(A\underline{p}^\ell, \underline{p}^\ell)}.$$

Ohne Einschränkung der Allgemeinheit sei $\alpha_\ell \neq 0$. Andernfalls würde aus der Rekursion

$$\underline{r}^{\ell+1} = \underline{r}^\ell - \alpha_\ell A\underline{p}^\ell$$

und der Orthogonalität

$$(\underline{r}^{\ell+1}, \underline{r}^\ell) = 0$$

die Gleichheit

$$0 = (\underline{r}^{\ell+1}, \underline{r}^\ell) = (\underline{r}^\ell, \underline{r}^\ell)$$

folgen und somit die Forderung $\underline{r}^\ell = 0$ nach sich ziehen, d.h. $\underline{x}^\ell = \underline{x}$ ist die exakte Lösung des linearen Gleichungssystems $A\underline{x} = \underline{f}$.

Aus

$$\underline{r}^{\ell+1} = \underline{r}^\ell - \alpha_\ell A\underline{p}^\ell$$

folgt dann

$$A\underline{p}^\ell = \frac{1}{\alpha_\ell} [\underline{r}^\ell - \underline{r}^{\ell+1}]$$

und somit ergibt sich für den Zähler von $\beta_{k\ell}$

$$(\underline{r}^{k+1}, A\underline{p}^\ell) = \frac{1}{\alpha_\ell} (\underline{r}^{k+1}, \underline{r}^\ell - \underline{r}^{\ell+1}) = 0 \quad \text{für } \ell = 0, \dots, k-1$$

und daher

$$\beta_{k\ell} = 0 \quad \text{für } \ell = 0, \dots, k-1.$$

Weiterhin ist

$$(\underline{r}^{k+1}, A\underline{p}^k) = \frac{1}{\alpha_k} (\underline{r}^{k+1}, \underline{r}^{k+1}) \quad \text{für } \ell = k.$$

Damit ergibt sich

$$\underline{p}^{k+1} = \underline{r}^{k+1} - \beta_k \underline{p}^k$$

mit

$$\beta_k = \frac{(\underline{r}^{k+1}, A\underline{p}^k)}{(A\underline{p}^k, \underline{p}^k)} = -\frac{1}{\alpha_k} \frac{(\underline{r}^{k+1}, \underline{r}^{k+1})}{(A\underline{p}^k, \underline{p}^k)}.$$

Aus

$$\underline{r}^{k+1} = \underline{r}^k - \alpha_k A\underline{p}^k$$

folgt

$$\alpha_k A\underline{p}^k = \underline{r}^k - \underline{r}^{k+1}$$

und somit

$$\alpha_k (A\underline{p}^k, \underline{p}^k) = (\underline{r}^k - \underline{r}^{k+1}, \underline{p}^k) = (\underline{r}^k, \underline{p}^k) = (\underline{r}^k, \underline{r}^k - \beta_{k-1} \underline{p}^{k-1}) = (\underline{r}^k, \underline{r}^k) = \rho_k.$$

Damit ist schließlich

$$\beta_k = -\frac{\rho_{k+1}}{\rho_k}$$

bzw.

$$\alpha_k = \frac{(\underline{r}^k, \underline{p}^k)}{(A\underline{p}^k, \underline{p}^k)} = \frac{\rho_k}{(A\underline{p}^k, \underline{p}^k)}.$$

Die resultierende Methode ist das Verfahren konjugierter Gradienten (CG), welches auf Hestenes und Stiefel [6] zurückgeht.

Für eine beliebig gegebene Startnäherung $\underline{x}^0 \in \mathbb{R}^n$ sei $\underline{r}^0 = A\underline{x}^0 - \underline{f}$. Setze $\underline{p}^0 := \underline{r}^0$ und berechne $\varrho_0 = (\underline{r}^0, \underline{r}^0)$. Stoppe, falls $\varrho_0 < \varepsilon^2$ mit einer vorgegebenen Fehlergenauigkeit ε erreicht ist.

Berechne für $k = 0, 1, \dots, n-2$:

$$\underline{s}^k = A\underline{p}^k, \quad \sigma_k = (\underline{s}^k, \underline{p}^k), \quad \alpha_k = \frac{\varrho_k}{\sigma_k}$$

$$\underline{x}^{k+1} := \underline{x}^k - \alpha_k \underline{p}^k$$

$$\underline{r}^{k+1} := \underline{r}^k - \alpha_k \underline{s}^k$$

$$\varrho_{k+1} := (\underline{r}^{k+1}, \underline{r}^{k+1})$$

Stoppe, falls $\varrho_{k+1} < \varepsilon^2 \varrho_0$ mit einer vorgegebenen Fehlergenauigkeit ε erreicht ist. Berechne andernfalls die neue Suchrichtung

$$\underline{p}^{k+1} := \underline{r}^{k+1} + \beta_k \underline{p}^k, \quad \beta_k := \frac{\varrho_{k+1}}{\varrho_k}$$

Algorithmus 4.8: Iterationsvorschrift des konjugierten Gradientenverfahrens.

Aus den Induktionsvoraussetzungen

$$\underline{r}^\ell \in \text{span}\{\underline{r}^0\}, \quad \underline{p}^\ell = \underline{r}^\ell \in \text{span}\{\underline{r}^0\}$$

folgt wegen

$$\underline{r}^{\ell+1} = \underline{r}^\ell - \alpha_\ell A\underline{p}^\ell, \quad \underline{p}^{\ell+1} = \underline{r}^{\ell+1} + \beta_\ell \underline{p}^\ell$$

durch vollständige Induktion nach $\ell = 0, \dots, k-1$

$$\underline{p}^k \in \text{span}\{\underline{r}^0, A\underline{r}^0, \dots, A^k \underline{r}^0\} =: S_k(A, \underline{r}^0).$$

Hierbei bezeichnet $S_k(A, \underline{r}^0)$ den k -ten Krylov-Raum der Matrix A zum Anfangsresiduum $\underline{r}^0$. Nach Konstruktion ist durch

$$S_k(A, \underline{r}^0) = \text{span}\{\underline{p}^0, \underline{p}^1, \dots, \underline{p}^k\}$$

eine A -orthogonale Basis von $S_k(A, \underline{r}^0)$ gegeben.

Satz 4.7. *Für eine symmetrische und positiv definite Matrix $A = A^\top > 0$ konvergiert das konjugierte Gradientenverfahren mit der Konvergenzabschätzung*

$$\|\underline{x}^k - \underline{x}\|_A \leq \frac{2q^k}{1 + q^{2k}} \|\underline{e}^0\|_A$$

mit

$$q = \frac{\sqrt{\kappa_2(A)} + 1}{\sqrt{\kappa_2(A)} - 1}, \quad \kappa_2(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

Beweis: Für die durch das konjugierte Gradientenverfahren konstruierte Näherungslösung

$$\underline{x}^k = \underline{x}^0 - \sum_{\ell=0}^{k-1} \alpha_\ell \underline{p}^\ell$$

folgt aus der A -Orthogonalität der Suchrichtungen $\underline{p}^\ell$ für die durch die symmetrische und positiv definite Matrix A induzierte Norm des Fehlers

$$\|\underline{x}^k - \underline{x}\|_A^2 = \left\| \sum_{\ell=0}^{n-1} \alpha_\ell \underline{p}^\ell \right\|_A^2 = \sum_{\ell=0}^{n-1} \sum_{j=0}^{n-1} \alpha_\ell \alpha_j (A \underline{p}^\ell, \underline{p}^j) = \sum_{\ell=0}^{n-1} \alpha_\ell^2 \|\underline{p}^\ell\|_A^2.$$

Für eine beliebige Linearkombination

$$\underline{w} = \sum_{\ell=0}^{k-1} w_\ell \underline{p}^\ell$$

mit beliebigen Koeffizienten $w_0, \dots, w_{k-1}$ folgt analog

$$\|\underline{x}^0 - \underline{w} - \underline{x}\|_A^2 = \sum_{\ell=0}^{k-1} [w_\ell - \alpha_\ell]^2 \|\underline{p}^\ell\|_A^2 + \sum_{\ell=k}^{n-1} \alpha_\ell^2 \|\underline{p}^\ell\|_A^2,$$

und somit

$$\|\underline{x}^k - \underline{x}\|_A \leq \|\underline{x}^0 - \underline{w} - \underline{x}\|_A \quad \text{für alle } \underline{w} \in S_{k-1}(A, \underline{r}^0).$$

Die Näherungslösung $\underline{x}^k$ ist also Lösung des Minimierungsproblems

$$\|\underline{x}^k - \underline{x}\|_A = \min_{\underline{w} \in S_{k-1}(A, \underline{x}^0)} \|\underline{x}^0 - \underline{w} - \underline{x}\|_A.$$

Mit

$$\underline{r}^0 = A\underline{x}^0 - \underline{f} = A(\underline{x}^0 - \underline{x}) = A\underline{e}^0, \quad \underline{e}^0 = \underline{x}^0 - \underline{x}$$

ergibt sich

$$\underline{x}^0 - \underline{w} - \underline{x} = \underline{e}^0 - \sum_{\ell=0}^{k-1} w_\ell A^\ell \underline{r}^0 = \underline{e}^0 - \sum_{\ell=0}^{k-1} w_\ell A^{\ell+1} \underline{e}^0 = \sum_{\ell=0}^k \tilde{w}_\ell A^\ell \underline{e}^0$$

mit $\tilde{w}_0 = 1$ und $\tilde{w}_\ell = w_{\ell-1}$ für $\ell = 1, \dots, k$. Damit gilt

$$\underline{x}^0 - \underline{w} - \underline{x} = p_k(A)\underline{e}^0$$

mit einem Matrix-Polynom $p_k \in \Pi_k^1$, d.h. $\underline{x}^k$ ist Lösung des Minimierungsproblems

$$\|\underline{x}^k - \underline{x}\|_A = \min_{p_k \in \Pi_k^1} \|p_k(A)\underline{e}^0\|_A.$$

Die Eigenvektoren $\{\underline{v}^j\}_{j=1}^n$ der symmetrischen und positiv definiten Matrix A bilden ein Orthonormalsystem. Die zugehörigen Eigenwerte von A seien $\lambda_j(A)$. Dann ergibt sich für den Anfangsfehler

$$\underline{e}^0 = \sum_{j=1}^n (\underline{e}^0, \underline{v}^j) \underline{v}^j$$

und in der Folge

$$p_k(A)\underline{e}^0 = p_k(A) \sum_{j=1}^n (\underline{e}^0, \underline{v}^j) \underline{v}^j = \sum_{j=1}^n (\underline{e}^0, \underline{v}^j) p_k(A) \underline{v}^j = \sum_{j=1}^n (\underline{e}^0, \underline{v}^j) p_k(\lambda_j(A)) \underline{v}^j.$$

Aus der Orthonormalität der Eigenvektoren $\underline{v}^j$ ergibt sich weiterhin

$$\begin{aligned} \|p_k(A)\underline{e}^0\|_A^2 &= (Ap_k(A)\underline{e}^0, p_k(A)\underline{e}^0) \\ &= (A \sum_{j=1}^n (\underline{e}^0, \underline{v}^j) p_k(\lambda_j(A)) \underline{v}^j, \sum_{i=1}^n (\underline{e}^0, \underline{v}^i) p_k(\lambda_i(A)) \underline{v}^i) \\ &= (\sum_{j=1}^n (\underline{e}^0, \underline{v}^j) p_k(\lambda_j(A)) \lambda_j(A) \underline{v}^j, \sum_{i=1}^n (\underline{e}^0, \underline{v}^i) p_k(\lambda_i(A)) \underline{v}^i) \\ &= \sum_{i=1}^n \sum_{j=1}^n (\underline{e}^0, \underline{v}^i) (\underline{e}^0, \underline{v}^j) p_k(\lambda_i(A)) p_k(\lambda_j(A)) \lambda_j(A) (\underline{v}^j, \underline{v}^i) \\ &= \sum_{j=1}^n (\underline{e}^0, \underline{v}^j)^2 [p_k(\lambda_j(A))]^2 \lambda_j(A) \\ &\leq \max_{j=1, \dots, n} [p_k(\lambda_j(A))]^2 \sum_{j=1}^n (\underline{e}^0, \underline{v}^j)^2 \lambda_j(A) = \max_{j=1, \dots, n} [p_k(\lambda_j(A))]^2 \|\underline{e}^0\|_A^2 \end{aligned}$$

und somit

$$\|\underline{x}^k - \underline{x}\|_A \leq \min_{p_k \in \Pi_k^1} \max_{j=1, \dots, n} |p_k(\lambda_j(A))| \|\underline{e}^0\|_A \leq \min_{p_k \in \Pi_k^1} \max_{\lambda \in [\lambda_{\min}(A), \lambda_{\max}(A)]} |p_k(\lambda)| \|\underline{e}^0\|_A.$$

Mit Satz 1.3 ist schließlich

$$\min_{p_k \in \Pi_k^1} \max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} |p_k(\lambda)| = \frac{2q^k}{1 + q^{2k}}, \quad q = \frac{\sqrt{\lambda_{\max}(A)} + \sqrt{\lambda_{\min}(A)}}{\sqrt{\lambda_{\max}(A)} - \sqrt{\lambda_{\min}(A)}} = \frac{\sqrt{\kappa_2(A)} + 1}{\sqrt{\kappa_2(A)} - 1}.$$

■

Die Konvergenzgeschwindigkeit des konjugierten Gradientenverfahrens wird maßgeblich durch die extremalen Eigenwerte der Matrix A bestimmt. Aus der Charakterisierung der extremalen Eigenwerte mittels des Rayleigh-Quotienten,

$$\lambda_{\min}(A) = \min_{\underline{x} \in \mathbb{R}^n} \frac{(A\underline{x}, \underline{x})}{(\underline{x}, \underline{x})} \leq \max_{\underline{x} \in \mathbb{R}^n} \frac{(A\underline{x}, \underline{x})}{(\underline{x}, \underline{x})} = \lambda_{\max}(A),$$

folgt

$$\lambda_{\min}(A) (\underline{x}, \underline{x}) \leq (A\underline{x}, \underline{x}) \leq \lambda_{\max}(A) (\underline{x}, \underline{x}) \quad \text{für alle } \underline{x} \in \mathbb{R}^n.$$

Aus den Spektraläquivalenzungleichungen

$$c_1^A (\underline{x}, \underline{x}) \leq (A\underline{x}, \underline{x}) \leq c_2^A (\underline{x}, \underline{x}) \quad \text{für alle } \underline{x} \in \mathbb{R}^n$$

folgt deshalb für die Abschätzung der spektralen Konditionszahl

$$\kappa_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \leq \frac{c_2^A}{c_1^A}.$$

Abhängig von der jeweiligen Anwendung strebt $\kappa_2(A) \rightarrow \infty$ für $n \rightarrow \infty$. Ziel ist deshalb die Herleitung eines Verfahrens, welches eine beschränkte Anzahl notwendiger Iterationsschritte zum Erreichen einer vorgegebenen relativen Genauigkeit ε unabhängig von der Dimension n gewährleistet.

Eine symmetrische und positiv definite Matrix $B \in \mathbb{R}^{n \times n}$ gestattet die Faktorisierung

$$B = V \operatorname{diag}(\lambda_k(B)) V^\top$$

mit der durch die orthonormalen Eigenvektoren von B gebildeten orthonormalen Matrix V . Die Positivität der Eigenwerte $\lambda_k(B)$ ermöglicht die Definition der symmetrischen und positiv definiten Matrix

$$B^{1/2} = V \operatorname{diag}(\sqrt{\lambda_k(B)}) V^\top$$

mit $B^{1/2} B^{1/2} = B$ und der inversen Matrix $B^{-1/2} = (B^{1/2})^{-1}$.

Anstelle des linearen Gleichungssystems $A\underline{x} = \underline{f}$ wird jetzt das transformierte System

$$B^{-1/2} A B^{-1/2} B^{1/2} \underline{x} = B^{-1/2} \underline{f}$$

bzw. $\tilde{A}\tilde{\underline{x}} = \tilde{\underline{f}}$ mit

$$\tilde{A} = B^{-1/2}AB^{-1/2}, \quad \tilde{\underline{x}} = B^{1/2}\underline{x}, \quad \tilde{\underline{f}} = B^{-1/2}\underline{f}$$

betrachtet. Für die symmetrisch und positiv definite Matrix $\tilde{A}$ kann nun das in Algorithmus 4.8 angegebene konjugierte Gradientenverfahren angewendet werden. Dessen Konvergenzgeschwindigkeit ergibt sich aus der Abschätzung der spektralen Konditionszahl $\kappa_2(\tilde{A})$, welche unmittelbar aus den Spektraläquivalenzungleichungen

$$c_1^{\tilde{A}}(\tilde{\underline{x}}, \tilde{\underline{x}}) \leq (\tilde{A}\tilde{\underline{x}}, \tilde{\underline{x}}) \leq c_2^{\tilde{A}}(\tilde{\underline{x}}, \tilde{\underline{x}}) \quad \text{für alle } \tilde{\underline{x}} \in \mathbb{R}^n$$

folgt. Einsetzen der Transformationen ergibt die dazu äquivalenten Spektraläquivalenzungleichungen

$$c_1^{\tilde{A}}(B\underline{x}, \underline{x}) \leq (A\underline{x}, \underline{x}) \leq c_2^{\tilde{A}}(B\underline{x}, \underline{x}) \quad \text{für alle } \underline{x} \in \mathbb{R}^n.$$

Für die Lösung des linearen Gleichungssystems $\tilde{A}\tilde{\underline{x}} = \tilde{\underline{f}}$ mit der symmetrisch und positiv definiten Matrix $\tilde{A}$ kann die in Algorithmus 4.8 angegebene Iterationsvorschrift angewendet werden. Mit den zugehörigen Transformationen ergibt sich für die Näherungslösung $\tilde{\underline{x}}^k$ sowie für das zugehörige Residuum $\tilde{\underline{r}}^k$

$$\tilde{\underline{x}}^k = B^{1/2}\underline{x}^k, \quad \tilde{\underline{r}}^k = \tilde{A}\tilde{\underline{x}}^k - \tilde{\underline{f}} = B^{-1/2}(A\underline{x}^k - \underline{f}) = B^{-1/2}\underline{r}^k.$$

Aus dem Ansatz

$$\tilde{\underline{x}} = \tilde{\underline{x}}^0 - \sum_{\ell=0}^{n-1} \tilde{\alpha}_\ell \tilde{\underline{p}}^\ell, \quad \tilde{\alpha}_\ell = \frac{(\tilde{\underline{r}}^k, \tilde{\underline{p}}^\ell)}{(\tilde{A}\tilde{\underline{p}}^\ell, \tilde{\underline{p}}^\ell)}$$

für die Lösung $\tilde{\underline{x}}$ des transformierten linearen Gleichungssystems $\tilde{A}\tilde{\underline{x}} = \tilde{\underline{f}}$ folgt durch Multiplikation mit $B^{-1/2}$

$$\underline{x} = \underline{x}^0 - \sum_{\ell=0}^{n-1} \tilde{\alpha}_\ell B^{-1/2} \tilde{\underline{p}}^\ell = \underline{x}^0 - \sum_{\ell=0}^{n-1} \tilde{\alpha}_\ell \underline{p}^\ell$$

mit $\underline{p}^\ell = B^{-1/2} \tilde{\underline{p}}^\ell$ bzw. $\tilde{\underline{p}}^\ell = B^{1/2} \underline{p}^\ell$. Weiter ist

$$\tilde{\varrho}_k = (\tilde{\underline{r}}^k, \tilde{\underline{p}}^k) = (B^{-1}\underline{r}^k, \underline{p}^k)$$

sowie

$$\tilde{\sigma}_k = (\tilde{A}\tilde{\underline{p}}^k, \tilde{\underline{p}}^k) = (A\underline{p}^k, \underline{p}^k).$$

Für die Konstruktion der transformierten Suchrichtungen $\tilde{\underline{p}}^{k+1}$ ergibt sich schließlich

$$\tilde{\underline{p}}^{k+1} = \tilde{\underline{r}}^{k+1} + \tilde{\beta}_k \tilde{\underline{p}}^k$$

bzw. durch Multiplikation mit $B^{-1/2}$

$$\underline{p}^{k+1} = B^{-1}\underline{r}^{k+1} + \tilde{\beta}_k \underline{p}^k.$$

Die resultierende Iterationsvorschrift des vorkonditionierten konjugierten Gradientenverfahrens ist in Algorithmus 4.9 angegeben.

Für eine beliebig gegebene Startnäherung $\underline{x}^0 \in \mathbb{R}^n$ sei $\underline{r}^0 = A\underline{x}^0 - \underline{f}$. Berechne $\underline{v}^0 = B^{-1}\underline{r}^0$, $\underline{p}^0 := \underline{v}^0$, $\varrho_0 = (\underline{v}^0, \underline{r}^0)$. Stoppe, falls $\varrho_0 < \varepsilon^2$ mit einer vorgegebenen Fehlergenauigkeit ε erreicht ist.

Berechne für $k = 0, 1, \dots, n-2$:

$$\underline{s}^k = A\underline{p}^k, \sigma_k = (\underline{s}^k, \underline{p}^k), \alpha_k = \frac{\varrho_k}{\sigma_k}$$

$$\underline{x}^{k+1} := \underline{x}^k - \alpha_k \underline{p}^k$$

$$\underline{r}^{k+1} := \underline{r}^k - \alpha_k \underline{s}^k$$

$$\underline{v}^{k+1} = B^{-1}\underline{r}^{k+1}$$

$$\varrho_{k+1} := (\underline{v}^{k+1}, \underline{r}^{k+1})$$

Stoppe, falls $\varrho_{k+1} < \varepsilon^2 \varrho_0$ mit einer vorgegebenen Fehlergenauigkeit ε erreicht ist. Berechne andernfalls die neue Suchrichtung

$$\underline{p}^{k+1} := \underline{v}^{k+1} + \beta_k \underline{p}^k, \beta_k := \frac{\varrho_{k+1}}{\varrho_k}$$

Algorithmus 4.9: Konjugiertes Gradientenverfahrens mit Vorkonditionierung.

Neben einer Matrix–Vektor–Multiplikation mit A ist in Algorithmus 4.9 pro Iterationsschritt jeweils eine Anwendung der Vorkonditionierung $\underline{v} = B^{-1}\underline{r}$ zu realisieren. Damit muß die Matrix B neben den entsprechenden Spektraläquivalenzungleichungen auch eine effiziente Anwendung von B^{-1} ermöglichen. Sind beide Bedingungen erfüllt, so wird B als Vorkonditionierung zu A bezeichnet. Diese ist in der Regel problemabhängig zu konstruieren und stellt in vielen Anwendungen eine Herausforderung dar.

4.5 Verfahren des minimalen Residuums

Betrachtet wird jetzt das lineare Gleichungssystem $A\underline{x} = \underline{f}$ mit einer invertierbaren Matrix A , d.h. es wird weder die Symmetrie noch die positive Definitheit der Systemmatrix A vorausgesetzt. Wie beim CG–Verfahren wird für eine Anfangsnäherung $\underline{x}^0$ das zugehörige Residuum $\underline{r}^0 = A\underline{x}^0 - \underline{f}$ und der dadurch induzierte Krylov–Raum

$$S_k(A, \underline{r}^0) = \text{span} \{ \underline{r}^0, A\underline{r}^0, \dots, A^k \underline{r}^0 \}$$

eingeführt. Für diesen soll eine Basis $\{\underline{v}^\ell\}_{\ell=0}^{n-1}$ orthonormalen Vektoren mit

$$(\underline{v}^k, \underline{v}^\ell) = \delta_{k\ell}$$

konstruiert werden. Die Anwendung des Gram–Schmidt–Orthogonalisierungsverfahrens führt bei geeigneter Wahl der Ausgangsvektoren auf die Methode von Arnoldi:

Sei mit $\underline{x}^0 \in \mathbb{R}^n$ eine beliebige Startnäherung gegeben.

Berechne $\underline{r}^0 := A\underline{x}^0 - \underline{f}$ und setze

$$\underline{v}^0 = \frac{\underline{r}^0}{\|\underline{r}^0\|_2}.$$

Berechne für $k = 0, 1, \dots, n-2$:

$$\hat{\underline{v}}^{k+1} = A\underline{v}^k - \sum_{\ell=0}^k \beta_{k\ell} \underline{v}^\ell$$

mit

$$\beta_{k\ell} = (A\underline{v}^k, \underline{v}^\ell).$$

Abbruch für $\|\hat{\underline{v}}^k\|_2 = 0$, setze andernfalls

$$\underline{v}^{k+1} = \frac{\hat{\underline{v}}^{k+1}}{\|\hat{\underline{v}}^{k+1}\|_2}.$$

Algorithmus 4.10: Methode von Arnoldi.

Der Ansatz der Näherungslösung

$$\underline{x}^{k+1} = \underline{x}^0 - \sum_{\ell=0}^k \alpha_\ell \underline{v}^\ell$$

ergibt für das zugehörige Residuum

$$\underline{r}^{k+1} = A\underline{x}^{k+1} - \underline{f} = \underline{r}^0 - \sum_{\ell=0}^k \alpha_\ell A\underline{v}^\ell, \quad \underline{r}^0 = A\underline{x}^0 - \underline{f}.$$

Zu bestimmen bleiben die Zerlegungskoeffizienten α_ℓ durch Minimierung des Residuums $\underline{r}^{k+1}$ in der Euklidischen Vektornorm,

$$\|\underline{r}^{k+1}\|_2 = \|\underline{r}^0 - \sum_{\ell=0}^k \alpha_\ell A\underline{v}^\ell\|_2 \rightarrow \min_{\alpha_0, \dots, \alpha_k}.$$

Für eine alternative Darstellung des Residuenvektors $\underline{r}^{k+1}$ folgt aus der Methode von Arnoldi zunächst

$$A\underline{v}^\ell = \hat{\underline{v}}^{\ell+1} + \sum_{j=0}^{\ell} \beta_{\ell,j} \underline{v}^j = \|\hat{\underline{v}}^{\ell+1}\|_2 \underline{v}^{\ell+1} + \sum_{j=0}^{\ell} \beta_{\ell,j} \underline{v}^j = \sum_{j=0}^{\ell+1} \beta_{\ell,j} \underline{v}^j$$

mit

$$\beta_{\ell,j} = \begin{cases} (A\underline{v}^\ell, \underline{v}^j) & \text{für } j = 0, \dots, \ell, \\ \|\hat{\underline{v}}^{\ell+1}\|_2 & \text{für } j = \ell+1. \end{cases}$$

Dann ist

$$\underline{r}^{k+1} = \underline{r}^0 - \sum_{\ell=0}^k \alpha_\ell A\underline{v}^\ell = \underline{r}^0 - \sum_{\ell=0}^k \alpha_\ell \sum_{j=0}^{\ell+1} \beta_{\ell,j} \underline{v}^j = \underline{r}^0 - V_{k+1} H_k \underline{\alpha}$$

mit der durch die orthonormalen Vektoren $\underline{v}^j$ gebildeten Matrix

$$V_{k+1} = (\underline{v}^0, \underline{v}^1, \dots, \underline{v}^{k+1}) \in \mathbb{R}^{n \times (k+2)}$$

und mit der durch

$$H_k[j, \ell] = \begin{cases} \beta_{\ell, j} & \text{für } j \leq \ell + 1, \\ 0 & \text{für } j > \ell + 1 \end{cases}$$

definierten oberen Hessenberg-Matrix $H_k \in \mathbb{R}^{(k+2) \times (k+1)}$. Nach Konstruktion folgt weiterhin

$$\underline{r}^0 = \|\underline{r}^0\|_2 \underline{v}^0 = \|\underline{r}^0\|_2 V_{k+1} \underline{e}^0$$

mit dem ersten Einheitsvektor $\underline{e}^0 = (1, 0, \dots, 0)^\top \in \mathbb{R}^{k+2}$. Mit der Invarianz der Euklidischen Vektornorm bezüglich orthonormalen Matrizen ergibt sich

$$\|\underline{r}^{k+1}\|_2 = \|\underline{r}^0 - V_{k+1} H_k \underline{\alpha}\|_2 = \|V_{k+1} (\|\underline{r}^0\|_2 \underline{e}^0 - H_k \underline{\alpha})\|_2 = \|\|\underline{r}^0\|_2 \underline{e}^0 - H_k \underline{\alpha}\|_2.$$

Wegen $H_k \in \mathbb{R}^{(k+2) \times (k+1)}$ und $\underline{\alpha} \in \mathbb{R}^{k+1}$ sowie $\underline{e}^0 \in \mathbb{R}^{k+2}$ entspricht die Forderung $H_k \underline{\alpha} = \|\underline{r}^0\|_2 \underline{e}^0$ einem überbestimmten linearen Gleichungssystem mit $k+2$ Gleichungen für $k+1$ Unbekannte. Zu bestimmen bleibt jener Lösungsvektor $\underline{\alpha} \in \mathbb{R}^{k+1}$, der das verbleibende Residuum minimiert.

Sei $Q_k \in \mathbb{R}^{(k+2) \times (k+2)}$ eine orthonormale Matrix mit $Q_k^\top Q_k = I_{k+2}$, so daß $Q_k H_k \in \mathbb{R}^{(k+2) \times (k+1)}$ obere Dreiecksgestalt besitzt. Dann gilt

$$\begin{aligned} \|\underline{r}^{k+1}\|_2 &= \|\|\underline{r}^0\|_2 \underline{e}^0 - H_k \underline{\alpha}\|_2 = \|\|\underline{r}^0\|_2 Q_k \underline{e}^0 - Q_k H_k \underline{\alpha}\|_2 \\ &= \|\|\underline{r}^0\|_2 Q_k \underline{e}^0 - R_k \underline{\alpha}\|_2 = \|\underline{r}^0\|_2 |(Q_k \underline{e}^0)_{k+1}|, \end{aligned}$$

falls

$$(R_k \underline{\alpha})_\ell = \|\underline{r}^0\|_2 (Q_k \underline{e}^0)_\ell \quad \text{für } \ell = 0, \dots, k$$

erfüllt ist.

Zu bestimmen bleibt eine orthonormale Matrix $Q_k \in \mathbb{R}^{(k+2) \times (k+2)}$, welche die obere Hessenberg-Matrix

$$H_k = \begin{pmatrix} \beta_{0,0} & \beta_{1,0} & \dots & \beta_{k,0} \\ \beta_{0,1} & \beta_{1,1} & & \vdots \\ 0 & \beta_{1,2} & \ddots & \vdots \\ & 0 & \ddots & \beta_{k,k} \\ & & & \beta_{k,k+1} \end{pmatrix} \in \mathbb{R}^{(k+2) \times (k+1)}$$

in eine obere Dreiecks-Matrix

$$R_k = Q_k H_k = \begin{pmatrix} r_{0,0} & r_{0,1} & \dots & r_{0,k} \\ 0 & r_{1,1} & & \vdots \\ 0 & 0 & \ddots & \vdots \\ & & \ddots & r_{k,k} \\ & & & 0 \end{pmatrix} \in \mathbb{R}^{(k+2) \times (k+1)}$$

transformiert. Ausgehend von dem Spaltenvektor

$$\underline{h}^j = (\beta_{j,0}, \dots, \beta_{j,j-1}, \beta_{j,j}, \beta_{j,j+1}, 0, \dots, 0)^\top \in \mathbb{R}^{k+2}$$

ist zunächst jene orthonormale Matrix G_j gesucht, so daß

$$G_j \underline{h}^j = \left(\beta_{j,0}, \dots, \beta_{j,j-1}, \tilde{\beta}_{j,j}, 0, 0, \dots, 0 \right)^\top.$$

Offenbar ist die Betrachtung einer orthonormalen Transformationsmatrix $\tilde{G}_j \in \mathbb{R}^{2 \times 2}$ mit

$$\tilde{G}_j \begin{pmatrix} \beta_{j,j} \\ \beta_{j,j+1} \end{pmatrix} = \begin{pmatrix} \tilde{\beta}_{j,j} \\ 0 \end{pmatrix}$$

mit der allgemeinen Darstellung

$$\tilde{G}_j = \begin{pmatrix} a_j & b_j \\ -b_j & a_j \end{pmatrix}, \quad a_j^2 + b_j^2 = 1,$$

ausreichend. Die Koeffizienten a_j und b_j ergeben sich aus der Forderung

$$-b_j \beta_{j,j} + a_j \beta_{j,j+1} = 0.$$

Unter Beachtung der Normierung $a_j^2 + b_j^2 = 1$ folgt daraus

$$a_j = \frac{\beta_{j,j}}{\sqrt{\beta_{j,j}^2 + \beta_{j,j+1}^2}}, \quad b_j = \frac{\beta_{j,j+1}}{\sqrt{\beta_{j,j}^2 + \beta_{j,j+1}^2}}$$

und daher

$$\tilde{\beta}_{j,j} = a_j \beta_{j,j} + b_j \beta_{j,j+1} = \sqrt{\beta_{j,j}^2 + \beta_{j,j+1}^2} > 0,$$

wenn $\beta_{j,j+1} > 0$ vorausgesetzt wird. Für $j = 0, \dots, k$ sind die resultierenden Transformationsmatrizen die Givens-Rotationen

$$G_j = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & a_j & b_j & & \\ & & & -b_j & a_j & & \\ & & & & & 1 & \\ & & & & & & \ddots \\ & & & & & & & 1 \end{pmatrix} \in \mathbb{R}^{(k+2) \times (k+2)}$$

mit $G_j[j, j] = G_j[j+1, j+1] = a_j$.

Deren rekursive Anwendung liefert mit

$$\begin{aligned}
G_k G_{k-1} \dots G_2 G_1 G_0 H_k &= G_k G_{k-1} \dots G_2 G_1 G_0 \begin{pmatrix} \beta_{0,0} & \beta_{1,0} & \dots & \beta_{k,0} \\ \beta_{0,1} & \beta_{1,1} & \dots & \beta_{k,1} \\ 0 & \beta_{1,2} & \ddots & \vdots \\ & 0 & \ddots & \beta_{k,k} \\ & & & \beta_{k,k+1} \end{pmatrix} \\
&= G_k G_{k-1} \dots G_2 G_1 \begin{pmatrix} \tilde{\beta}_{0,0} & \tilde{\beta}_{1,0} & \dots & \tilde{\beta}_{k,0} \\ 0 & \tilde{\beta}_{1,1} & \dots & \tilde{\beta}_{k,1} \\ & \beta_{1,2} & \ddots & \vdots \\ & & \ddots & \beta_{k,k} \\ & & & \beta_{k,k+1} \end{pmatrix} \\
&= G_k G_{k-1} \dots G_2 \begin{pmatrix} \tilde{\beta}_{0,0} & \tilde{\beta}_{1,0} & \dots & \tilde{\beta}_{k,0} \\ 0 & \tilde{\beta}_{1,1} & \dots & \tilde{\beta}_{k,1} \\ & 0 & \ddots & \vdots \\ & & \ddots & \beta_{k,k} \\ & & & \beta_{k,k+1} \end{pmatrix} \\
&= \begin{pmatrix} \tilde{\beta}_{0,0} & \tilde{\beta}_{1,0} & \dots & \tilde{\beta}_{k,0} \\ 0 & \tilde{\beta}_{1,1} & \dots & \tilde{\beta}_{k,1} \\ & 0 & \ddots & \vdots \\ & & \ddots & \tilde{\beta}_{k,k} \\ & & & 0 \end{pmatrix} = R_k
\end{aligned}$$

die gewünschte obere Dreiecksmatrix, deren Invertierbarkeit aus der Positivität der Diagonaleinträge $\tilde{\beta}_{j,j}$ folgt. Beim Übergang von H_k zu H_{k+1} , d.h. bei Hinzunahme eines weiteren Spaltenvektors $\underline{h}^{k+1}$ bzw. einer weiteren Suchrichtung $\underline{v}^{k+2}$, und vor der Anwendung der zugehörigen orthonormalen Matrix G_{k+1} sind alle vorherigen Transformationen $G_k, \dots, G_0$ auf $\underline{h}^{k+1}$ anzuwenden.

Nach Konstruktion ist

$$Q_k = G_k G_{k-1} \dots G_1 G_0 \in \mathbb{R}^{(k+2) \times (k+2)}$$

orthonormal und zu untersuchen bleibt die Auswertung von

$$\begin{aligned}
 Q_k \underline{e}^0 &= G_k \dots G_0 \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} = G_k \dots G_1 \begin{pmatrix} a_0 \\ -b_0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} = G_k \dots G_2 \begin{pmatrix} a_0 \\ a_1(-b_0) \\ (-b_0)(-b_1) \\ \vdots \\ 0 \\ 0 \end{pmatrix} \\
 &= \begin{pmatrix} a_0 \\ a_1(-b_0) \\ (-b_0)(-b_1) \\ \vdots \\ a_k(-b_0) \dots (-b_{k-1}) \\ (-b_0) \dots (-b_k) \end{pmatrix} \in \mathbb{R}^{k+2}.
 \end{aligned}$$

Während die ersten $k+1$ Komponenten von $Q_k \underline{e}^0$ die rechte Seite des linearen Gleichungssystems zur Bestimmung des Koeffizientenvektors $\underline{\alpha} \in \mathbb{R}^{k+1}$ bilden, beschreibt die letzte Komponente das verbleibende Residuum

$$\varrho_{k+1} = \|\underline{e}^0\|_2 |(Q_k \underline{e}^0)_{k+1}| = \|\underline{e}^0\|_2 \prod_{j=0}^k b_j.$$

Wegen

$$b_j = \frac{\beta_{j,j+1}}{\sqrt{\beta_{j,j+1}^2 + \beta_{j,j}^2}} = \frac{\|\hat{\underline{v}}^j\|_2}{\sqrt{\|\hat{\underline{v}}^j\|_2^2 + (A\underline{v}^j, \underline{v}^j)^2}}$$

folgt für $(A\underline{v}^j, \underline{v}^j) \neq 0$ wegen $\beta_j < 1$ ein monotonen Abklingverhalten des Fehlers. Insbesondere für die Abbruch-Situation der Methode von Arnoldi, $\|\hat{\underline{v}}^k\|_2 = 0$, folgt $b_k = 0$ und somit

$$\varrho_{k+1} = \|\underline{x}^{k+1}\|_2 = 0,$$

d.h. $\underline{x}^{k+1} = \underline{x}$ ist die exakte Lösung des linearen Gleichungssystems $A\underline{x} = \underline{f}$.

Das resultierende Iterationsverfahren ist das in Algorithmus 4.11 angegebene verallgemeinerte Verfahren des minimalen Residuums (GMRES) [12].

Für eine beliebig gegebene Startnäherung $\underline{x}^0 \in \mathbb{R}^n$ sei $\underline{r}^0 = A\underline{x}^0 - \underline{f}$.
 Berechne $\varrho_0 = \|\underline{r}^0\|_2$. Stoppe, falls $\varrho_0 < \varepsilon$ mit einer vorgegebenen
 Fehlergenauigkeit ε erreicht ist. Setze andernfalls $\underline{v}^0 = \frac{1}{\varrho_0} \underline{r}^0$, $p_0 = \varrho_0$.

Berechne für $k = 0, 1, \dots, n-2$:

$$\underline{w}^k = A\underline{v}^k$$

$$\underline{\tilde{v}}^{k+1} = \underline{w}^k - \sum_{\ell=0}^k \beta_{k\ell} \underline{v}^\ell, \quad \beta_{k\ell} = (\underline{w}^k, \underline{v}^\ell), \quad \beta_{kk+1} = \|\underline{\tilde{v}}^{k+1}\|_2$$

Falls $\beta_{kk+1} = 0$, stoppe und berechne Näherungslösung $\underline{x}^{k+1}$.

$$\underline{v}^{k+1} = \frac{1}{\beta_{kk+1}} \underline{\tilde{v}}^{k+1}$$

Berechne für $\ell = 0, \dots, k-1$:

$$\tilde{\beta}_{k\ell} = a_\ell \beta_{k\ell} + b_\ell \beta_{k\ell+1}$$

$$\tilde{\beta}_{k\ell+1} = -b_\ell \beta_{k\ell} + a_\ell \beta_{k\ell+1}$$

$$a_k = \frac{\beta_{kk}}{\sqrt{\beta_{kk}^2 + \beta_{kk+1}^2}}, \quad b_k = \frac{\beta_{kk+1}}{\sqrt{\beta_{kk}^2 + \beta_{kk+1}^2}}, \quad \tilde{\beta}_{kk} = \sqrt{\beta_{kk}^2 + \beta_{kk+1}^2}$$

$$p_{k+1} = -b_k p_k, \quad p_k = a_k p_k, \quad \varrho_{k+1} = |p_{k+1}|$$

Stoppe, falls $\varrho_{k+1} < \varepsilon \varrho_0$ mit einer vorgegebenen Fehlergenauigkeit ε
 erreicht ist und berechne die Näherungslösung:

Berechne für $\ell = k, k-1, \dots, 0$:

$$\alpha_\ell = \frac{1}{\beta_{\ell\ell}} \left[p_\ell - \sum_{j=\ell+1}^k \beta_{\ell j} \alpha_j \right]$$

$$\underline{x}^{k+1} = \underline{x}^0 - \sum_{\ell=0}^k \alpha_\ell \underline{v}^\ell$$

Algorithmus 4.11: Iterationsvorschrift GMRES Verfahren.