Iterative Lösungsverfahren für große lineare Gleichungssysteme

Steffen Börm

Stand 25. September 2020

Inhaltsverzeichnis

1 Einleitung

Im Mittelpunkt dieser Vorlesung stehen Verfahren zur effizienten Behandlung großer linearer Gleichungssysteme der Form

$$\mathbf{A}\mathbf{x} = \mathbf{b} \qquad \text{für } \mathbf{A} = (a_{ij})_{i,j\in\mathcal{I}} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}, \mathbf{x} = (x_i)_{i\in\mathcal{I}}, \mathbf{b} = (b_i)_{i\in\mathcal{I}} \in \mathbb{R}^{\mathcal{I}} \tag{1.1}$$

mit einer n-elementigen allgemeinen Indexmenge $\mathcal{I}$.

Wir haben bereits Algorithmen kennengelernt, mit denen sich derartige Probleme lösen lassen, etwa die Gauß-Elimination (bzw. LR-Faktorisierung), die Cholesky-Faktorisierung (für symmetrische positiv definite Matrizen) oder die Householder-Zerlegung (bzw. QR-Faktorisierung). Im allgemeinen Fall wächst der Rechenaufwand dieser Verfahren kubisch in n, also führt eine Verdopplung der Problemdimension zu einer Verachtfachung des Rechenaufwands. Der Speicherbedarf wächst quadratisch, eine Verdopplung der Dimension bedeutet also eine Vervierfachung des Speicherbedarfs.

Deshalb lassen sich diese Verfahren nur dann auf große Probleme anwenden, wenn sehr schnelle Rechner (heutzutage in der Regel Parallelrechner, da sich die Leistung einzelner Prozessoren nicht beliebig steigern lässt) mit sehr hoher Speicherkapazität zur Verfügung stehen.

Wenn wir ein großes lineares Gleichungssystem lösen wollen, ohne den Gegenwert mehrerer Einfamilienhäuser in modernste Rechnertechnik zu investieren, müssen wir uns also nach alternativen Verfahren umsehen. Dieses Kapitel stellt eine Reihe erfolgreicher Verfahren kurz dar, die detaillierte Behandlung der Algorithmen und die Analyse ihrer Vor- und Nachteile ist späteren Kapiteln vorbehalten.

Dieses Skript orientiert sich eng an dem Buch „Iterative Lösung großer schwachbesetzter Gleichungssysteme“ von Wolfgang Hackbusch, erschienen 1993 im Teubner-Verlag. Dieses Buch bietet wesentlich mehr als den hier gegebenen Überblick und ist jedem an iterativen Verfahren Interessierten wärmstens zu empfehlen. Eine erweiterte Version dieses Buchs ist unter der Web-Adresse `http://www.mis.mpg.de/scicomp/Fulltext/ggl.ps` zu finden.

Danksagung. Ich bedanke mich bei Knut Reimer, Christoph Gerken, Jonas Grams, Jelle Kuiper und Franziska Schwarzenbach für Verbesserungsvorschläge und Korrekturen.

1.1 Direkte Löser für schwachbesetzte Matrizen

Falls die meisten Einträge von $\mathbf{A}$ gleich Null sind, lassen sich die klassischen Lösungsverfahren so modifizieren, dass sie effizienter arbeiten. In diesem Abschnitt sei $\mathcal{I} = [1:n]$

Definition 1.1 (Dreiecksmatrizen) *Eine Matrix* $\mathbf{L} \in \mathbb{K}^{n\times n}$ *nennen wir eine* linke untere Dreiecksmatrix*, falls*

$$\ell_{ij} = 0 \qquad \text{für alle } i,j \in [1:n] \text{ mit } i < j$$

gilt, falls also alle Einträge oberhalb der Diagonalen gleich null sind.

Eine Matrix $\mathbf{R} \in \mathbb{K}^{n\times n}$ *nennen wir eine* rechte obere Dreiecksmatrix*, falls*

$$r_{ij} = 0 \qquad \text{für alle } i,j \in [1:n] \text{ mit } i > j$$

gilt, falls also alle Einträge unterhalb der Diagonalen gleich null sind.

Wir nennen $(\mathbf{L},\mathbf{R})$ *eine* LR-Zerlegung *einer Matrix* $\mathbf{A} \in \mathbb{K}^{n\times n}$*, falls* $\mathbf{L} \in \mathbb{K}^{n\times n}$ *eine linke untere und* $\mathbf{R} \in \mathbb{K}^{n\times n}$ *eine rechte obere Dreiecksmatrix ist und* $\mathbf{A} = \mathbf{LR}$ *gilt.*

Satz 1.2 (Skyline-Struktur) *Sei* $\mathbf{A} \in \mathbb{K}^{n\times n}$ *invertierbar, und sei* $(\mathbf{L},\mathbf{R})$ *eine LR-Zerlegung der Matrix* $\mathbf{A}$*. Dann gelten*

$$\begin{aligned}
(\forall k \in [1:j] \ : \ a_{ik} = 0) &\Rightarrow \ell_{ij} = 0 \qquad && \text{für alle } i,j \in [1:n] \text{ mit } i > j,\\
(\forall k \in [1:i] \ : \ a_{kj} = 0) &\Rightarrow r_{ij} = 0 \qquad && \text{für alle } i,j \in [1:n] \text{ mit } i < j.
\end{aligned}$$

Falls also am Anfang einer Zeile der Matrix $\mathbf{A}$ *Nulleinträge auftreten, bleiben sie auch in der Matrix* $\mathbf{L}$ *erhalten. Falls am Anfang einer Spalte der Matrix* $\mathbf{A}$ *Nulleinträge auftreten, bleiben sie auch in der Matrix* $\mathbf{R}$ *erhalten.*

Beweis. Wir führen den Beweis per Induktion über $n \in \mathbb{N}$.

Induktionsanfang: Für $n = 1$ ist nichts zu zeigen.

Induktionsvoraussetzung: Sei $n \in \mathbb{N}$ so gegeben, dass die Behauptung für alle Matrizen $\mathbf{A} \in \mathbb{K}^{n\times n}$ gilt.

Induktionsschritt: Sei $\mathbf{A} \in \mathbb{K}^{(n+1)\times(n+1)}$ eine Matrix, und sei $(\mathbf{L},\mathbf{R})$ eine LR-Zerlegung dieser Matrix.

Wir zerlegen die Matrizen in

$$\begin{aligned}
\mathbf{A} &= \begin{pmatrix} \mathbf{A}_{**} & \mathbf{A}_{*,n+1} \\ \mathbf{A}_{n+1,*} & a_{n+1,n+1} \end{pmatrix}, && \mathbf{A}_{**} \in \mathbb{K}^{n\times n},\ \mathbf{A}_{*,n+1} \in \mathbb{K}^{n\times 1},\ \mathbf{A}_{n+1,*} \in \mathbb{K}^{1\times n},\\
\mathbf{L} &= \begin{pmatrix} \mathbf{L}_{**} & \\ \mathbf{L}_{n+1,*} & \ell_{n+1,n+1} \end{pmatrix}, && \mathbf{L}_{**} \in \mathbb{K}^{n\times n},\ \mathbf{L}_{n+1,*} \in \mathbb{K}^{1\times n},\\
\mathbf{R} &= \begin{pmatrix} \mathbf{R}_{**} & \mathbf{R}_{*,n+1} \\ & r_{n+1,n+1} \end{pmatrix}, && \mathbf{R}_{**} \in \mathbb{K}^{n\times n},\ \mathbf{R}_{*,n+1} \in \mathbb{K}^{n\times 1},
\end{aligned}$$

und stellen fest, dass

$$\begin{pmatrix} \mathbf{A}_{**} & \mathbf{A}_{*,n+1} \\ \mathbf{A}_{n+1,*} & a_{n+1,n+1} \end{pmatrix} = \mathbf{A} = \mathbf{LR} = \begin{pmatrix} \mathbf{L}_{**} & \\ \mathbf{L}_{n+1,*} & \ell_{n+1,n+1} \end{pmatrix} \begin{pmatrix} \mathbf{R}_{**} & \mathbf{R}_{*,n+1} \\ & r_{n+1,n+1} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{L}_{**}\mathbf{R}_{**} & \mathbf{L}_{**}\mathbf{R}_{*,n+1} \\ \mathbf{L}_{n+1,*}\mathbf{R}_{**} & \mathbf{L}_{n+1,*}\mathbf{R}_{*,n+1} + \ell_{n+1,n+1}r_{n+1,n+1} \end{pmatrix}$$

gilt, also insbesondere

$$\mathbf{A}_{**} = \mathbf{L}_{**}\mathbf{R}_{**}, \qquad \mathbf{A}_{*,n+1} = \mathbf{L}_{**}\mathbf{R}_{*,n+1}, \qquad \mathbf{A}_{n+1,*} = \mathbf{L}_{n+1,*}\mathbf{R}_{**}.$$

Da $\mathbf{A}$ invertierbar ist, müssen auch $\mathbf{L}$ und $\mathbf{R}$ invertierbar sein. Da die beiden letztgenannten Dreiecksmatrizen sind, folgt, dass auch $\mathbf{L}_{**}$ und $\mathbf{R}_{**}$ invertierbar sind.

Seien nun $i, j \in [1 : n+1]$ mit $i > j$ so gegeben, dass $a_{ik} = 0$ für alle $k \in [1 : j]$ gilt. Falls $i \leq n$ gilt, folgt $\ell_{ij} = 0$ mit der Induktionsvoraussetzung, da $(\mathbf{L}_{**}, \mathbf{R}_{**})$ eine LR-Zerlegung der Matrix $\mathbf{A}_{**}$ ist. Wir brauchen also nur den Fall $i = n+1$ zu betrachten. Aus unserer Voraussetzung folgt, dass $\mathbf{A}_{n+1,*}$ dann die Form

$$\mathbf{A}_{n+1,*} = \begin{pmatrix} \mathbf{0} & \mathbf{A}_{n+1,+} \end{pmatrix} \quad \text{mit} \quad \mathbf{A}_{n+1,+} \in \mathbb{K}^{1\times(n-j)}$$

aufweist. Mit der Zerlegung

$$\mathbf{R}_{**} = \begin{pmatrix} \mathbf{R}_{00} & \mathbf{R}_{0+} \\ & \mathbf{R}_{++} \end{pmatrix}, \qquad \mathbf{R}_{00} \in \mathbb{K}^{j\times j},\ \mathbf{R}_{0+} \in \mathbb{K}^{j\times(n-j)},\ \mathbf{R}_{++} \in \mathbb{K}^{(n-j)\times(n-j)},$$

$$\mathbf{L}_{n+1,*} = \begin{pmatrix} \mathbf{L}_{n+1,0} & \mathbf{L}_{n+1,+} \end{pmatrix}, \quad \mathbf{L}_{n+1,0} \in \mathbb{K}^{1\times j},\ \mathbf{L}_{n+1,+} \in \mathbb{K}^{1\times(n-j)}$$

folgt aus

$$\begin{pmatrix} \mathbf{0} & \mathbf{A}_{n+1,+} \end{pmatrix} = \mathbf{A}_{n+1,*} = \mathbf{L}_{n+1,*}\mathbf{R}_{**} = \begin{pmatrix} \mathbf{L}_{n+1,0} & \mathbf{L}_{n+1,+} \end{pmatrix} \begin{pmatrix} \mathbf{R}_{00} & \mathbf{R}_{0+} \\ & \mathbf{R}_{++} \end{pmatrix}$$

unmittelbar $\mathbf{0} = \mathbf{L}_{n+1,0}\mathbf{R}_{00}$. Da $\mathbf{R}_{00}$ als Diagonalblock einer invertierbaren Dreiecksmatrix selbst invertierbar ist, folgen $\mathbf{L}_{n+1,0} = \mathbf{0}$ und damit insbesondere $\ell_{ij} = 0$.

Für den Nachweis zweite Aussage seien $i, j \in [1 : n+1]$ mit $i < j$ so gegeben, dass $a_{kj} = 0$ für alle $k \in [1 : i]$ gilt. Falls $j \leq n$ gilt, folgt $\ell_{ij} = 0$ wie zuvor mit der Induktionsvoraussetzung. Wir betrachten den Fall $j = n+1$. Es gilt

$$\mathbf{A}_{*,n+1} = \begin{pmatrix} \mathbf{0} \\ \mathbf{A}_{+,n+1} \end{pmatrix} \quad \text{mit} \quad \mathbf{A}_{+,n+1} \in \mathbb{K}^{(n-i)\times 1},$$

und mit der Zerlegung

$$\mathbf{L}_{**} = \begin{pmatrix} \mathbf{L}_{00} & \\ \mathbf{L}_{+0} & \mathbf{L}_{++} \end{pmatrix}, \qquad \mathbf{L}_{00} \in \mathbb{K}^{i\times i},\ \mathbf{L}_{+0} \in \mathbb{K}^{(n-i)\times i},\ \mathbf{L}_{++} \in \mathbb{K}^{(n-i)\times(n-i)},$$

$$\mathbf{R}_{*,n+1} = \begin{pmatrix} \mathbf{R}_{0,n+1} \\ \mathbf{R}_{+,n+1} \end{pmatrix} \qquad \mathbf{R}_{0,n+1} \in \mathbb{K}^{i\times 1},\ \mathbf{R}_{+,n+1} \in \mathbb{K}^{(n-i)\times 1}$$

folgt aus

$$\begin{pmatrix} \mathbf{0} \\ \mathbf{A}_{+,n+1} \end{pmatrix} = \mathbf{A}_{*,n+1} = \mathbf{L}_{**}\mathbf{R}_{*,n+1} = \begin{pmatrix} \mathbf{L}_{00} & \\ \mathbf{L}_{+0} & \mathbf{L}_{++} \end{pmatrix} \begin{pmatrix} \mathbf{R}_{0,n+1} \\ \mathbf{R}_{+,n+1} \end{pmatrix}$$

unmittelbar $\mathbf{0} = \mathbf{L}_{00}\mathbf{R}_{0,n+1}$. Da $\mathbf{L}_{00}$ als Diagonalblock einer invertierbaren Dreiecksmatrix selbst invertierbar ist, haben wir $\mathbf{R}_{0,n+1} = \mathbf{0}$, also insbesondere $r_{ij} = 0$. ■

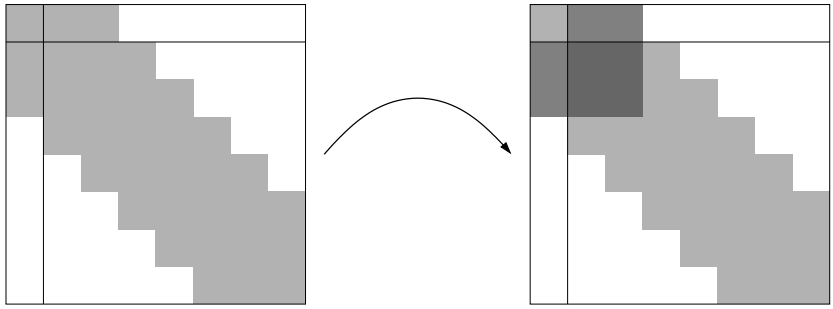

Abbildung 1.1: Konstruktion der Matrix **S** für Bandmatrizen

Definition 1.3 (Bandbreite) *Sei $p \in \mathbb{N}_0$. Falls*

$$|i-j| > p \Rightarrow a_{ij} = 0 \qquad \text{für alle } i,j \in \mathcal{I}$$

gilt, nennen wir p eine Bandbreitenschranke *für* $\mathbf{A}$*. Die kleinste Bandbreitenschranke nennen wir die* Bandbreite *von* $\mathbf{A}$*.*

Falls die Matrix $\mathbf{A}$ eine durch p beschränkte Bandbreite besitzt, gilt dasselbe auch für die Faktoren $\mathbf{L}, \mathbf{R} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ der *LR*-Zerlegung:

Satz 1.4 (Bandbreitenschranke für die *LR*-Zerlegung) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ invertierbar, und sei $(\mathbf{L}, \mathbf{R})$ eine LR-Zerlegung dieser Matrix.*

Sei $p \in \mathbb{N}$ eine Bandbreitenschranke für $\mathbf{A}$. Dann ist p auch eine Bandbreitenschranke für die Matrizen $\mathbf{L}$ und $\mathbf{R}$.

Beweis. Seien $i, j \in [1:n]$ mit $|i-j| > p$ gegeben.

Falls $i < j$ gilt, folgt $\ell_{ij} = 0$ per Definition. Anderenfalls haben wir $i - j > p$, also für alle $k \in [1:j]$ auch $i - k \geq i - j > p$ und damit $a_{ik} = 0$. Mit Satz 1.2 folgt $\ell_{ij} = 0$.

Falls $i > j$ gilt, folgt $r_{ij} = 0$ per Definition. Anderenfalls haben wir $j - i > p$, also für alle $k \in [1:i]$ auch $j - k \geq j - i > p$ und damit $a_{kj} = 0$. Mit Satz 1.2 folgt $r_{ij} = 0$. ■

Für die praktische Konstruktion der LR-Zerlegung empfiehlt sich eine ähnliche Vorgehensweise wie im Beweis des Satzes 1.2: Statt die letzte Zeile und Spalte separat zu behandeln, nehmen wir die erste, wir betrachten also

$$\mathbf{A} = \begin{pmatrix} a_{11} & \mathbf{A}_{1*} \\ \mathbf{A}_{*1} & \mathbf{A}_{**} \end{pmatrix}, \qquad \mathbf{A}_{**} \in \mathbb{K}^{(n-1)\times(n-1)},\ \mathbf{A}_{1*} \in \mathbb{K}^{1\times(n-1)},\ \mathbf{A}_{*1} \in \mathbb{K}^{(n-1)\times 1}$$

und erhalten bei entsprechender Zerlegung der Faktoren $\mathbf{L}$ und $\mathbf{R}$ die Gleichung

$$\begin{pmatrix} a_{11} & \mathbf{A}_{1*} \\ \mathbf{A}_{*1} & \mathbf{A}_{**} \end{pmatrix} = \mathbf{A} = \mathbf{LR} = \begin{pmatrix} \ell_{11} & \\ \mathbf{L}_{*1} & \mathbf{L}_{**} \end{pmatrix} \begin{pmatrix} r_{11} & \mathbf{R}_{1*} \\ & \mathbf{R}_{**} \end{pmatrix} = \begin{pmatrix} \ell_{11} r_{11} & \ell_{11}\mathbf{R}_{1*} \\ \mathbf{L}_{*1} r_{11} & \mathbf{L}_{*1}\mathbf{R}_{1*} + \mathbf{L}_{**}\mathbf{R}_{**} \end{pmatrix},$$

aus der sich die Gleichungen

$$a_{11} = \ell_{11} r_{11}, \qquad \mathbf{A}_{1*} = \ell_{11}\mathbf{R}_{1*}, \qquad \mathbf{A}_{*1} = \mathbf{L}_{*1} r_{11}, \qquad \mathbf{A}_{**} - \mathbf{L}_{*1}\mathbf{R}_{1*} = \mathbf{L}_{**}\mathbf{R}_{**}$$

```
procedure LRZerlegung(n, p, var A, L, R);
for m := 1 to n do
  l_mm ← 1;
  r_mm ← a_mm;
  for i ∈ [m + 1 : min{m + p, n}] do
    l_im ← a_im/r_mm;
    r_mi ← a_mi
  end for;
  for i, j ∈ [m + 1 : min{m + p, n}] do
    a_ij ← a_ij − l_im r_mj
  end for
end for
```

Abbildung 1.2: Berechnung der LR-Zerlegung

ergeben. Mit den ersten drei Gleichungen können wir ℓ_{11}, r_{11}, $\mathbf{R}_{1*}$ und $\mathbf{L}_{*1}$ bestimmen. Die vierte erlaubt es uns dann, $\mathbf{L}_{**}$ und $\mathbf{R}_{**}$ mit einer LR-Zerlegung der Dimension $n-1$ zu berechnen. Indem wir die Bandstruktur ausnutzen, sind für die Berechnung der linken Seite $\mathbf{S} := \mathbf{A}_{**} - \mathbf{L}_{*1}\mathbf{R}_{1*}$ der vierten Gleichung höchstens p^2 Operationen erforderlich, wenn wir $\mathbf{A}_{**}$ mit $\mathbf{S}$ überschreiben. Der resultierende Algorithmus ist in Abbildung 1.2 gegeben. Da wir nur an dem Fall $n \gg p$ interessiert sind, können wir uns die Analyse seiner algorithmischen Komplexität leicht machen:

Satz 1.5 (Komplexität der LR-Zerlegung) *Der in Abbildung 1.2 gegebene Algorithmus benötigt nicht mehr als np Divisionen, np^2 Multiplikationen und np^2 Subtraktionen.*

Beweis. Für ein m führt die erste innere Schleife des Algorithmus höchstens p Divisionen durch, da i nur Werte zwischen $m+1$ und $m+p$ annehmen kann.

In der zweiten inneren Schleife können i und j jeweils nur Werte zwischen $m+1$ und $m+p$ annehmen, also wird diese Schleife höchstens p^2-mal durchlaufen, so dass höchstens p^2 Multiplikationen und Subtraktionen durchzuführen sind.

Da die äußere Schleife genau n-mal durchlaufen wird, erhalten wir direkt das gewünschte Ergebnis. ■

Wir können sehen, dass diese Komplexitätsabschätzung wesentlich günstiger als im allgemeinen Fall ist: Falls die Bandbreite von $\mathbf{A}$ unabhängig von n beschränkt ist, können wir die LR-Zerlegung mit einem Aufwand berechnen, der lediglich *linear* in n wächst.

Sobald die LR-Zerlegung zur Verfügung steht, können wir das Gleichungssystem $\mathbf{b} = \mathbf{A}\mathbf{x} = \mathbf{L}\mathbf{R}\mathbf{x}$ durch Rückwärtseinsetzen in $\mathbf{R}$ und Vorwärtseinsetzen in $\mathbf{L}$ lösen. Da die Bandbreiten von $\mathbf{L}$ und $\mathbf{R}$ durch p beschränkt sind, können diese Berechnungen in $\mathcal{O}(np)$ Operationen durchgeführt werden, wir erhalten also wieder eine lineare Komplexität in der Problemdimension n.

Indem wir die besonderen Eigenschaften der Matrix $\mathbf{A}$ ausnutzen, können wir so eine wesentlich höhere Effizienz erreichen.

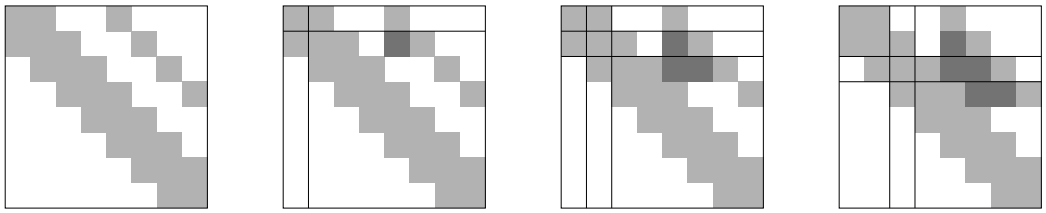

Abbildung 1.3: Auffüllen der Matrix $\mathbf{A}$

Leider ist die Bandbreite vieler in der Praxis auftretender Matrizen nicht beschränkt. Es tritt zwar häufig der Fall auf, dass nur wenige Einträge der Matrix von Null verschieden sind, aber diese Eigenschaft alleine genügt nicht, um ähnliche Aussagen wie in Satz 1.5 zu zeigen.

Ein Beispiel ist in Abbildung 1.3 zu sehen: Obwohl in der Ausgangsmatrix nur jeweils höchstens vier Einträge pro Zeile von Null verschieden sind, füllen die einzelnen Berechnungsschritte die Matrix $\mathbf{A}$ immer weiter auf, bis schließlich die volle Bandbreite von 4 erreicht ist, wir erhalten die von Satz 1.2 vorhergesagte Struktur. In der Praxis treten häufig Matrizen mit einer Bandbreite von $n^{1/2}$ auf, hier würde unser Algorithmus also $\mathcal{O}(n^2)$ Operationen und $\mathcal{O}(n^{3/2})$ Speicherplätze benötigen, selbst wenn die Ausgangsmatrizen fast nur aus Nulleinträgen bestehen.

1.2 Iterationsverfahren

Wir haben gesehen, dass bei Matrizen, die zwar nur wenige von Null verschiedene Einträge, aber trotzdem eine hohe Bandbreite aufweisen, Algorithmen wie die LR-Zerlegung ineffizient werden, weil sich die Matrix nach und nach auffüllt, also Nulleinträge durch die einzelnen Berechnungsschritte überschrieben werden.

Es wäre also sehr wünschenswert, ein Verfahren zu kennen, das die Matrix überhaupt nicht verändert und insbesondere auch keine Nulleinträge überschreibt.

Dieses Ziel erreichen die meisten *iterativen* Lösungsverfahren. Diese Algorithmen berechnen eine Folge $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots$ von Vektoren, ausgehend von einem beliebigen Startvektor $\mathbf{x}^{(0)} \in \mathbb{R}^{\mathcal{I}}$, die gegen die gewünschte Lösung $\mathbf{x}$ konvergieren. Im Gegensatz zu LR-, QR- oder Cholesky-Zerlegungen erhalten wir von diesen Verfahren also keine „exakte“ Lösung, sondern lediglich eine approximative Lösung, die allerdings *beliebig* genau ist.

Dieser scheinbare Nachteil wird durch zwei Beobachtungen relativiert: Erstens ist man bei den meisten Anwendungen ohnehin nicht an einer exakten Lösung interessiert, weil die Ausgangsdaten in der Regel schon gestört sind, etwa durch Rundungsfehler. Zweitens ist es leichter, bei iterativen Verfahren die Fehlerfortpflanzung zu kontrollieren, weil in jedem Schritt lediglich die Ausgangsdaten und die aktuelle Approximation eingehen, aber nicht die vorangehenden Approximationen. Aus diesem Grund werden in der Praxis gelegentlich iterative Verfahren eingesetzt, um die Rundungsfehler zu kompensieren, die bei einer LR- oder QR-Zerlegung auftreten können (dann spricht man von einer

„Nachiteration“).

Ein sehr einfaches Iterationsverfahren ist die *Richardson-Iteration*, bei der die Folge der Vektoren durch die Vorschrift

$$\mathbf{x}^{(m+1)} := \mathbf{x}^{(m)} - \theta(\mathbf{A}\mathbf{x}^{(m)} - \mathbf{b}) \qquad \text{für alle } m \in \mathbb{N}_0$$

gegeben ist. Hierbei ist $\theta \in \mathbb{C}$ ein geeignet zu wählender Parameter, der offensichtlich entscheidenden Einfluss auf das Verhalten des Verfahrens hat, beispielsweise bewirkt das Verfahren in komplizierter Weise nichts, wenn wir $\theta = 0$ setzen.

Bereits bei diesem sehr einfachen Beispiel können wir die Vorteile eines iterativen Verfahrens erkennen: Neben der Matrix $\mathbf{A}$, dem Vektor $\mathbf{b}$ und dem Lösungsvektor $\mathbf{x}$ (in dem wir die einzelnen Iterierten $\mathbf{x}^{(m)}$ unterbringen) ist lediglich ein Hilfsvektor erforderlich, in dem wir den sogenannten „Defekt“ $\mathbf{d}^{(m)} := \mathbf{A}\mathbf{x}^{(m)} - \mathbf{b}$ speichern, wir kommen also mit relativ wenig Speicher aus. Die Matrix $\mathbf{A}$ tritt ausschließlich in Form der Matrix-Vektor-Multiplikation $\mathbf{A}\mathbf{x}^{(m)}$ auf, insbesondere müssen wir ihre Einträge nicht verändern und vor allem keine Nulleinträge überschreiben.

Die Durchführung eines Iterationsschritts, also die Berechnung von $\mathbf{x}^{(m+1)}$ aus $\mathbf{x}^{(m)}$, kann in den meisten Anwendungsfällen sehr effizient gestaltet werden, sehr häufig sogar mit der optimalen Komplexität $\mathcal{O}(n)$.

Leider ist es mit der bloßen Durchführung des Verfahrens nicht getan, wir müssen auch wissen, wieviele Schritte erforderlich sind, um eine gewisse Genauigkeit zu erzielen, und ob das Verfahren für ein bestimmtes Problem überhaupt konvergiert. Die Untersuchung der Konvergenz iterativer Verfahren kann sich beliebig kompliziert gestalten, und der Entwurf eines geeigneten Verfahrens für eine bestimmte Problemklasse kann auch auf dem heutigen Stand der Forschung noch eine große wissenschaftliche Herausforderung darstellen.

Im Falle des Richardson-Verfahrens lässt sich die Frage nach der Konvergenz relativ einfach beantworten: Das Verfahren konvergiert (für ein geeignetes θ), falls alle Eigenwerte der Matrix $\mathbf{A}$ in derselben Hälfte der komplexen Ebene liegen, falls es also ein $z \in \mathbb{C}$ gibt, das $\operatorname{Re}(z\lambda) > 0$ für alle Eigenwerte λ von $\mathbf{A}$ erfüllt.

Wir werden diese Aussage später beweisen, im Augenblick genügt es, zwei Beispiele zu untersuchen. Das erste Beispiel ist das durch

$$\mathbf{A} := \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \qquad \mathbf{b} := \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

gegebene Gleichungssystem $\mathbf{A}\mathbf{x} = \mathbf{b}$. Offensichtlich besitzt es die eindeutige Lösung $\mathbf{x} = \mathbf{0}$. Falls wir das Richardson-Verfahren auf einen beliebigen Startvektor $\mathbf{x}^{(0)} \in \mathbb{R}^2$ anwenden, erhalten wir

$$\mathbf{x}^{(1)} = \begin{pmatrix} (1-\theta)x_1^{(0)} \\ (1-2\theta)x_2^{(0)} \end{pmatrix}, \quad \mathbf{x}^{(2)} = \begin{pmatrix} (1-\theta)^2 x_1^{(0)} \\ (1-2\theta)^2 x_2^{(0)} \end{pmatrix}, \quad \mathbf{x}^{(m)} = \begin{pmatrix} (1-\theta)^m x_1^{(0)} \\ (1-2\theta)^m x_2^{(0)} \end{pmatrix},$$

es folgt somit

$$\|\mathbf{x}^{(m)}\|_2 \le \max\{|1-\theta|^m, |1-2\theta|^m\}\|\mathbf{x}^{(0)}\|_2 \qquad \text{für alle } m \in \mathbb{N}_0,$$

wir erhalten also Konvergenz für alle $\theta \in (0,1)$, und die optimale Wahl $\theta = 2/3$ führt zu

$$\|\mathbf{x}^{(m)}\|_2 \leq \left(\frac{1}{3}\right)^m \|\mathbf{x}^{(0)}\|_2 \qquad \text{für alle } m \in \mathbb{N}_0,$$

also reduziert jeder Schritt des Verfahrens den Fehler um den Faktor $\varrho := 1/3$. Diese Konvergenzrate hängt nicht von der Größe des Gleichungssystems ab, sondern lediglich von dessen Eigenwerten.

Das zweite Beispiel ist das durch

$$\mathbf{A} := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad \mathbf{b} := \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{1.2}$$

gegebene Gleichungssystem $\mathbf{Ax} = \mathbf{b}$. Offensichtlich besitzt es ebenfalls die eindeutige Lösung $\mathbf{x} = \mathbf{0}$. Falls wir das Richardson-Verfahren auf einen Startvektor $\mathbf{x}^{(0)}$ anwenden, erhalten wir diesmal

$$\mathbf{x}^{(1)} = \begin{pmatrix} (1-\theta)x_1^{(0)} \\ (1+\theta)x_2^{(0)} \end{pmatrix}, \qquad \mathbf{x}^{(2)} = \begin{pmatrix} (1-\theta)^2 x_1^{(0)} \\ (1+\theta)^2 x_2^{(0)} \end{pmatrix}, \qquad \mathbf{x}^{(m)} = \begin{pmatrix} (1-\theta)^m x_1^{(0)} \\ (1+\theta)^m x_2^{(0)} \end{pmatrix}.$$

Falls $\mathbf{x}_1^{(0)} \neq 0 \neq \mathbf{x}_2^{(0)}$ gilt, erhalten wir, dass die Folge der Vektoren niemals gegen die korrekte Lösung konvergieren wird, weil

$$\max\{|1-\theta|, |1+\theta|\} = \sqrt{1 + |\theta|^2 + 2|\operatorname{Re}\theta|} \geq 1$$

gilt. Im schlimmsten Fall, für $\theta \neq 0$, wird die Folge sogar divergieren.

Wir können also feststellen, dass die Konvergenz eines iterativen Verfahrens entscheidend von den Eigenschaften der Matrix $\mathbf{A}$ abhängt. Wenn das Verfahren zur Matrix passt, können iterative Verfahren sehr effizient sein, beispielsweise gibt es für einige wichtige Klassen von Gleichungssystemen iterative Verfahren, die auch mehrere Millionen Unbekannte in wenigen Sekunden berechnen.

1.3 Semiiterative Verfahren

Eine Kombination aus direkten und iterativen Verfahren stellen die sogenannten *semiiterativen Verfahren* dar. Diese Algorithmen bestimmen die Lösung in einer festen Anzahl von Schritten (üblicherweise n), berechnen dabei aber Zwischenergebnisse, die auch schon gute Approximationen darstellen.

Ein großer Vorteil der Richardson-Iteration besteht darin, dass sie durchgeführt werden kann, sobald wir eine Möglichkeit besitzen, Matrix-Vektor-Produkte $\mathbf{y} := \mathbf{Ax}$ zu berechnen. Wir wollen nun ein Verfahren skizzieren, das ebenfalls ausschließlich mit derartigen Produkten auskommt.

Nach dem Satz von Cayley-Hamilton gibt es mindestens ein Polynom π mit einem Grad von höchstens n derart, dass

$$\pi(\mathbf{A}) = 0$$

gilt. Wir wählen ein Polynom minimalen Grades p, das diese Gleichung erfüllt, und bezeichnen seine Koeffizienten mit $\pi_0, \ldots, \pi_p \in \mathbb{R}$, so dass die Gleichung die Gestalt

$$\pi_p \mathbf{A}^p + \pi_{p-1}\mathbf{A}^{p-1} + \ldots + \pi_1 \mathbf{A} + \pi_0 \mathbf{I} = 0$$

annimmt. Da π minimalen Grad besitzt, muss $\pi_0 \neq 0$ gelten (sonst könnten wir ein Polynom niedrigeren Grades konstruieren, indem wir die obige Gleichung mit $\mathbf{A}^{-1}$ multiplizieren), und wir erhalten

$$\left(-\frac{\pi_p}{\pi_0}\right)\mathbf{A}^p + \left(-\frac{\pi_{p-1}}{\pi_0}\right)\mathbf{A}^{p-1} + \ldots + \left(-\frac{\pi_1}{\pi_0}\right)\mathbf{A} = \mathbf{I}.$$

Nun können wir $\mathbf{A}$ entweder nach links oder rechts aus der Summe herausziehen und gelangen zu der Gleichung

$$\left[\left(-\frac{\pi_p}{\pi_0}\right)\mathbf{A}^{p-1} + \left(-\frac{\pi_{p-1}}{\pi_0}\right)\mathbf{A}^{p-2} + \ldots + \left(-\frac{\pi_1}{\pi_0}\right)\mathbf{I}\right]\mathbf{A} = \mathbf{I},$$

sowie ihrem Gegenstück

$$\mathbf{A}\left[\left(-\frac{\pi_p}{\pi_0}\right)\mathbf{A}^{p-1} + \left(-\frac{\pi_{p-1}}{\pi_0}\right)\mathbf{A}^{p-2} + \ldots + \left(-\frac{\pi_1}{\pi_0}\right)\mathbf{I}\right] = \mathbf{I}.$$

Beide zusammen implizieren

$$\left(-\frac{\pi_p}{\pi_0}\right)\mathbf{A}^{p-1} + \left(-\frac{\pi_{p-1}}{\pi_0}\right)\mathbf{A}^{p-2} + \ldots + \left(-\frac{\pi_1}{\pi_0}\right)\mathbf{I} = \mathbf{A}^{-1}.$$

Wir haben also die Inverse von $\mathbf{A}$ durch ein Polynom dargestellt.

Für unsere Aufgabe, also das Lösen des Gleichungssystems $\mathbf{A}\mathbf{x} = \mathbf{b}$, bedeutet diese Gleichung, dass wir $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ in der Form

$$\mathbf{x} = \left(-\frac{\pi_p}{\pi_0}\right)\mathbf{A}^{p-1}\mathbf{b} + \left(-\frac{\pi_{p-1}}{\pi_0}\right)\mathbf{A}^{p-2}\mathbf{b} + \ldots + \left(-\frac{\pi_1}{\pi_0}\right)\mathbf{b}$$

darstellen können, wir brauchen also lediglich die Vektoren

$$\mathbf{y}^{(0)} := \mathbf{b}, \quad \mathbf{y}^{(1)} := \mathbf{A}\mathbf{y}^{(0)} = \mathbf{A}\mathbf{b}, \quad \mathbf{y}^{(2)} := \mathbf{A}\mathbf{y}^{(1)} = \mathbf{A}^2\mathbf{b}, \quad \mathbf{y}^{(m)} := \mathbf{A}\mathbf{y}^{(m-1)} = \mathbf{A}^m\mathbf{b}$$

zu berechnen und wissen, dass

$$\mathbf{x} = \alpha_{p-1}\mathbf{y}^{(p-1)} + \alpha_{p-2}\mathbf{y}^{(p-2)} + \ldots + \alpha_0\mathbf{y}^{(0)}$$

mit den Koeffizienten $\alpha_i := -\pi_{i+1}/\pi_0$ gilt.

In der Praxis steht uns das Polynom π nicht zur Verfügung, wir müssen also das richtige p und die richtigen Koeffizienten α_i anders konstruieren. *Krylow-Verfahren* gehen dabei so vor, dass der Reihe nach die Vektoren $\mathbf{y}^{(m)}$ berechnet und dann die Lösung im zugehörigen *Krylow-Raum*

$$\mathcal{K}(\mathbf{b}, m) := \operatorname{span}\{\mathbf{y}^{(0)}, \ldots, \mathbf{y}^{(m)}\} = \operatorname{span}\{\mathbf{b}, \ldots, \mathbf{A}^m\mathbf{b}\}$$

gesucht wird. Die Suche nach der Lösung lässt sich als lineares Ausgleichsproblem formulieren: Wir suchen Koeffizienten $\alpha_0, \ldots, \alpha_m \in \mathbb{R}$ derart, dass der Fehler

$$\epsilon_m := \|\mathbf{x} - \alpha_m \mathbf{y}^{(m)} - \ldots - \alpha_0 \mathbf{y}^{(0)}\|$$

in einer geeigneten Norm (schließlich steht uns $\mathbf{x}$ nicht zur Verfügung) minimiert wird. Dieses lineare Ausgleichsproblem kann effizient gelöst werden, sofern m nicht zu groß ist.

Wir können diesen Prozess so lange wiederholen, wie die Vektoren $\mathbf{y}^{(m)}$ linear unabhängig sind. Wenn sie für ein m nicht mehr linear unabhängig sein sollten, kann man zeigen, dass $\epsilon_m = 0$ gelten muss, dass wir also unsere Lösung bereits gefunden haben.

Da die Krylow-Räume geschachtelt sind (es gilt offensichtlich $\mathcal{K}(\mathbf{b}, m) \subseteq \mathcal{K}(\mathbf{b}, m+1)$), muss auch $\epsilon_m \geq \epsilon_{m+1}$ gelten, und aus dem Satz von Cayley-Hamilton folgt $\epsilon_{n-1} = 0$. Das Krylow-Verfahren berechnet also eine Folge von Vektoren, die monoton gegen die Lösung $\mathbf{x}$ konvergiert und nach spätestens $n-1$ Schritten die exakte Lösung erreicht.

1.4 Eindimensionales Modellproblem

Wie wir gesehen haben, spielen für die verschiedenen skizzierten Verfahren verschiedene Eigenschaften des Gleichungssystems eine Rolle: Bei Band-LR-Zerlegungen ist die Bandbreite der Matrix von entscheidender Bedeutung, bei der Richardson-Iteration sind es die Eigenwerte, und auch die Analyse von Krylow-Verfahren lässt sich, zumindest in wichtigen Spezialfällen, auf die Analyse der Eigenwerte zurückführen.

Deshalb ist es sinnvoll, zum Vergleich verschiedener Verfahren einfache Modellprobleme heranzuziehen, bei denen sich die nötigen Eigenschaften leicht nachweisen lassen. Wir beschränken uns hier auf lineare Gleichungssysteme, die aus der Diskretisierung einer partiellen Differentialgleichung entstehen, denn dieser Ansatz hat den Vorteil, dass wir Matrizen beliebiger Größe mit geringem Aufwand aufstellen können und dass sie viele Eigenschaften aufweisen, die auch praktisch auftretende Probleme besitzen.

Zunächst betrachten wir die eindimensionale partielle Differentialgleichung

$$-u''(x) = f(x) \qquad \text{für alle } x \in \Omega := (0,1), \tag{1.3a}$$

$$u(0) = u(1) = 0, \tag{1.3b}$$

für eine Funktion $u \in C[0,1]$ mit $u|_{(0,1)} \in C^2(0,1)$ und eine Funktion $f \in C(0,1)$. Da das Intervall $[0,1]$ unendlich viele Punkte enthält, können wir die Lösung u in einem Computer mit endlichem Speicher nicht exakt darstellen, müssen sie also approximieren.

Zu diesem Zweck wählen wir ein $N \in \mathbb{N}$ und unterteilen das Intervall $[0,1]$ in $N+1$ Teilintervalle der Länge $h := 1/(N+1)$. Die Anfangs- und Endpunkte der Intervalle sind durch

$$\xi_i := hi \qquad \text{für alle } i \in \{0, \ldots, N+1\}$$

gegeben, und es gilt $0 = \xi_0 < \xi_1 < \ldots < \xi_N < \xi_{N+1} = 1$. Unser Ziel ist es nun, die Werte der Funktion in diesen Punkten zu bestimmen, also

$$u_i := u(\xi_i) \qquad \text{für alle } i \in \{0, \ldots, N+1\}$$

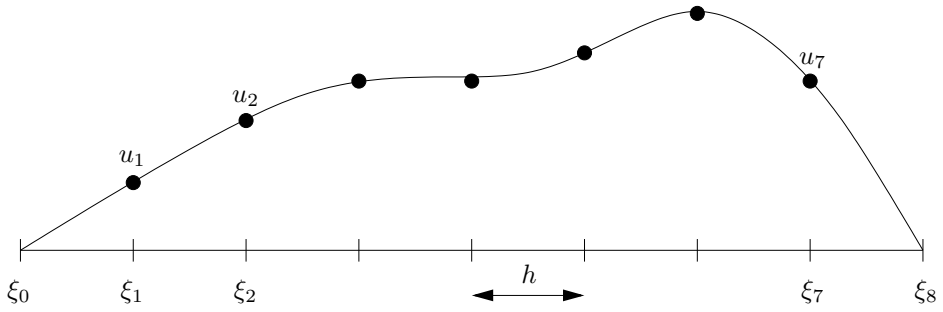

Abbildung 1.4: Eindimensionales Modellproblem

zu berechnen. Wegen $u_0 = u(0) = 0$ und $u_{N+1} = u(1) = 0$ sind nur die Werte $u_1, \ldots, u_N \in \mathbb{R}$ zu bestimmen, also der Vektor $\mathbf{u} = (u_i)_{i\in\mathcal{I}}$ für $\mathcal{I} := \{1, \ldots, N\}$.

Da uns die wahre Lösung u nicht zur Verfügung steht, müssen wir versuchen, ihre Ableitung u'' zu approximieren, indem wir nur die Werte $u_1, \ldots, u_N$ verwenden. Wir nehmen an, dass $u \in C^4[0,1]$ gilt und wenden die Taylor-Entwicklung an, um die Gleichungen

$$\begin{aligned} u(x+h) &= u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u^{(3)}(x) + \frac{h^4}{24}u^{(4)}(\eta_+), \\ u(x-h) &= u(x) - hu'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u^{(3)}(x) + \frac{h^4}{24}u^{(4)}(\eta_-) \end{aligned}$$

für geeignete $\eta_+ \in [x, x+h]$ und $\eta_- \in [x-h, x]$ zu erhalten. Wir addieren beide Gleichungen und finden

$$u(x+h) + u(x-h) = 2u(x) + h^2u''(x) + \frac{h^4}{24}u^{(4)}(\eta_+) + \frac{h^4}{24}u^{(4)}(\eta_-).$$

Wir sortieren die Terme um und stellen fest, dass

$$\left| u''(x) - \frac{1}{h^2}\left(u(x-h) - 2u(x) + u(x+h)\right) \right| \leq \frac{h^2}{12}\|u^{(4)}\|_{\infty,[x-h,x+h]} \tag{1.4}$$

gilt, also können wir

$$\frac{1}{h^2}\left(u(x-h) - 2u(x) + u(x+h)\right) \approx u''(x) \tag{1.5}$$

als Endergebnis festhalten. Angewendet auf die Punkte $\xi_1, \ldots, \xi_N$ bedeutet diese Gleichung

$$\frac{1}{h^2}(u_{i-1} - 2u_i + u_{i+1}) \approx u''(\xi_i).$$

Um die Verbindung zur Differentialgleichung (1.3) herzustellen, führen wir den Vektor $\mathbf{f} = (f_i)_{i\in\mathcal{I}}$ mit

$$f_i = f(\xi_i) \qquad \text{für alle } i \in \mathcal{I}$$

ein und erhalten wegen $f_i = -u''(\xi_i)$ die diskrete Approximation

$$\frac{1}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) = f_i \qquad \text{für alle } i \in \mathcal{I}$$

der Differentialgleichung (1.3). Wie wir sehen, handelt es sich dabei um ein lineares Gleichungssystem

$$\frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \\ u_N \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{N-1} \\ f_N \end{pmatrix},$$

das mit Hilfe der durch

$$L_{ij} = \begin{cases} 2h^{-2} & \text{falls } i = j, \\ -h^{-2} & \text{falls } |i-j| = 1, \\ 0 & \text{sonst} \end{cases} \qquad \text{für alle } i, j \in \mathcal{I}$$

und $\mathbf{L} = (L_{ij})_{i,j\in\mathcal{I}}$ gegebenen Matrix in der kompakten Form

$$\mathbf{Lu} = \mathbf{f}$$

dargestellt werden kann. An der Definition von $\mathbf{L}$ kann man leicht ablesen, dass diese Matrix eine Bandbreite von 1 besitzt, also lässt sich das Gleichungssystem mit den in Abschnitt 1.1 eingeführten Verfahren sehr effizient lösen.

Die Fehlerabschätzung (1.4) legt uns nahe, dass wir nur dann auf eine gute Genauigkeit hoffen dürfen, wenn h klein ist. Wegen $h = 1/(N+1)$ können wir dieses Ziel nur erreichen, indem wir die Anzahl N der Unbekannten relativ groß wählen.

Für die Untersuchung des Richardson-Verfahrens benötigen wir die Eigenwerte der Matrix $\mathbf{L}$, die sich mit Hilfe von etwas trigonometrischer Arithmetik bestimmen lassen.

Lemma 1.6 (Trigonometrische Orthogonalbasis) *Sei $N \in \mathbb{N}$ und $h := 1/(N+1)$. Es gilt*

$$\sum_{j=1}^{N} \sin(\pi jkh)\sin(\pi j\ell h) = \begin{cases} (N+1)/2 & \textit{falls } k = \ell, \\ 0 & \textit{ansonsten} \end{cases} \qquad \textit{für alle } k, \ell \in [1:N].$$

Beweis. Seien $k, \ell \in [1:N]$ gegeben. Mit der Eulerschen Gleichung gilt

$$\begin{aligned} \sum_{j=1}^{N} \sin(\pi jkh)\sin(\pi j\ell h) &= \sum_{j=1}^{N} \left(\frac{e^{i\pi jkh} - e^{-i\pi jkh}}{2i}\right)\left(\frac{e^{i\pi j\ell h} - e^{-i\pi j\ell h}}{2i}\right) \\ &= \frac{1}{4}\sum_{j=1}^{N} \left(e^{i\pi j(k-\ell)h} + e^{-i\pi j(k-\ell)h} - e^{i\pi j(k+\ell)h} - e^{-i\pi j(k+\ell)h}\right), \end{aligned}$$

und wegen $k+\ell \in [2:2N]$ gilt $(k+\ell)h \in (0,2)$, also folgen $q := e^{i\pi(k+\ell)h} \neq 1$ und

$$\sum_{j=1}^{N} e^{i\pi j(k+\ell)h} = \sum_{j=1}^{N} q^j = \frac{q^{N+1}-q}{q-1} = \frac{\sigma - q}{q-1}$$

für $\sigma := q^{N+1} = e^{i\pi(k+\ell)} \in \{-1,1\}$. In derselben Weise erhalten wir für $p := e^{i\pi(k-\ell)h}$ die Gleichung

$$\sum_{j=1}^{N} e^{i\pi j(k-\ell)h} = \sum_{j=1}^{N} p^j = \begin{cases} N & \text{falls } k=\ell, \\ \frac{\sigma-p}{p-1} & \text{ansonsten,} \end{cases}$$

da aus $k+\ell = (k-\ell)+2\ell$ auch $\sigma = e^{i\pi(k-\ell)} = p^{N+1}$ folgt. Für $k=\ell$ haben wir $\sigma = 1$, also

$$\begin{aligned}\sum_{j=1}^{N} \sin(\pi jkh)\sin(\pi j\ell h) &= \frac{1}{4}\left(2N - \frac{\sigma-q}{q-1} - \overline{\frac{\sigma-q}{q-1}}\right) = \frac{N}{2} - \frac{1}{2}\operatorname{Re}\frac{(1-q)(\bar{q}-1)}{|q-1|^2} \\ &= \frac{N}{2} + \frac{1}{2}\operatorname{Re}\frac{(q-1)(\bar{q}-1)}{|q-1|^2} = \frac{N+1}{2}.\end{aligned}$$

Für $k \neq \ell$ müssen wir die Fälle $\sigma = 1$ und $\sigma = -1$ unterscheiden. Im ersten Fall gilt

$$\begin{aligned}\sum_{j=1}^{N} \sin(\pi jkh)\sin(\pi j\ell h) &= \frac{1}{4}\left(\frac{\sigma-p}{p-1} + \overline{\frac{\sigma-p}{p-1}} - \frac{\sigma-q}{q-1} - \overline{\frac{\sigma-q}{q-1}}\right) \\ &= \frac{1}{4}\operatorname{Re}\left(\frac{1-p}{p-1} - \frac{1-q}{q-1}\right) = \frac{1}{4}(-1+1) = 0,\end{aligned}$$

im zweiten Fall erhalten wir

$$\begin{aligned}\sum_{j=1}^{N} \sin(\pi jkh)\sin(\pi j\ell h) &= \frac{1}{4}\left(\frac{\sigma-p}{p-1} + \overline{\frac{\sigma-p}{p-1}} - \frac{\sigma-q}{q-1} - \overline{\frac{\sigma-q}{q-1}}\right) \\ &= \frac{1}{4}\operatorname{Re}\left(\frac{-1-p}{p-1} - \frac{-1-q}{q-1}\right) = \frac{1}{4}\operatorname{Re}\left(\frac{(-1-p)(\bar{p}-1)}{|p-1|^2} - \frac{(-1-q)(\bar{q}-1)}{|q-1|^2}\right) \\ &= \frac{1}{4}\operatorname{Re}\left(\frac{1-|p|^2-\bar{p}+p}{|p-1|^2} - \frac{1-|q|^2-\bar{q}+q}{|q-1|^2}\right) = \frac{1}{4}\operatorname{Re}\left(\frac{p-\bar{p}}{|p-1|^2} - \frac{q-\bar{q}}{|q-1|^2}\right) = 0,\end{aligned}$$

und damit ist die gewünschte Orthonormalitätsbeziehung bewiesen. ■

Lemma 1.7 (Eigenwerte) *Für alle $k \in \mathcal{I}$ definieren wir den Vektor $\mathbf{e}^k \in \mathbb{R}^{\mathcal{I}}$ durch*

$$e_j^k := \sqrt{2h}\sin(\pi jkh) \qquad \text{für alle } j \in \mathcal{I}.$$

Es gilt

$$\mathbf{L}\mathbf{e}^k = \lambda_k \mathbf{e}^k \qquad \text{mit } \lambda_k := 4h^{-2}\sin^2(\pi kh/2) \qquad \text{für alle } k \in \mathcal{I},$$

und die Vektoren erfüllen

$$\langle \mathbf{e}^k, \mathbf{e}^\ell \rangle_2 = \begin{cases} 1 & \textit{falls } k = \ell, \\ 0 & \textit{ansonsten} \end{cases} \qquad \textit{für } k, \ell \in \mathcal{I},$$

bilden also eine aus Eigenvektoren der Matrix $\mathbf{L}$ *bestehende Orthonormalbasis von* $\mathbb{R}^{\mathcal{I}}$.

Beweis. Sei $k \in \mathcal{I}$. Um nachzuweisen, dass $\mathbf{e}^k$ ein Eigenvektor ist, betrachten wir für einen Index $j \in \mathcal{I}$ die j-te Komponente von $\mathbf{L}\mathbf{e}^k$: Wegen $\sin(\pi 0kh) = 0 = \sin(\pi(N+1)kh)$ erhalten wir

$$\begin{aligned} (\mathbf{L}\mathbf{e}^k)_j &= h^{-2}(2e_j^k - e_{j-1}^k - e_{j+1}^k) \\ &= \sqrt{2h}h^{-2}(2\sin(\pi jkh) - \sin(\pi(j-1)kh) - \sin(\pi(j+1)kh)). \end{aligned}$$

Wir wenden das Additionstheorem $\sin(x+y) = \sin(x)\cos(y) + \cos(x)\sin(y)$ an:

$$\begin{aligned} (\mathbf{L}\mathbf{e}^k)_j &= \sqrt{2h}h^{-2}(2\sin(\pi jkh) - \sin(\pi jkh)\cos(-\pi kh) - \cos(\pi jkh)\sin(-\pi kh) \\ &\qquad - \sin(\pi jkh)\cos(\pi kh) - \cos(\pi jkh)\sin(\pi kh)) \\ &= \sqrt{2h}h^{-2}(2 - 2\cos(\pi kh))\sin(\pi jkh). \end{aligned}$$

Nun benutzen wir das Additionstheorem $\cos(2x) = 2\cos^2(x) - 1$:

$$\begin{aligned} (\mathbf{L}\mathbf{e}^k)_j &= \sqrt{2h}h^{-2}(2 - 4\cos^2(\pi kh/2) + 2)\sin(\pi jkh) \\ &= 4\sqrt{2h}h^{-2}(1 - \cos^2(\pi kh/2))\sin(\pi jkh) \\ &= 4\sqrt{2h}h^{-2}\sin^2(\pi kh/2)\sin(\pi jkh) \\ &= \lambda_k e_j^k = (\lambda_k \mathbf{e}^k)_j. \end{aligned}$$

Da $\pi kh/2 = \pi k/(2(N+1)) \in (0, \pi/2)$ für alle $k \in \mathcal{I} = \{1, \ldots, N\}$ gilt und die Sinusfunktion auf diesem Intervall streng monoton wächst, sind alle Eigenwerte verschieden, also müssen alle Eigenvektoren $\mathbf{e}^k$ linear unabhängig sein.

Seien nun $k, \ell \in \mathcal{I}$. Mit Lemma 1.6 erhalten wir

$$\begin{aligned} \langle \mathbf{e}^k, \mathbf{e}^\ell \rangle_2 &= 2h \sum_{j=1}^{N} \sin(\pi jkh)\sin(\pi j\ell h) = 2h \begin{cases} (N+1)/2 & \text{falls } k = \ell, \\ 0 & \text{ansonsten} \end{cases} \\ &= \begin{cases} 1 & \text{falls } k = \ell, \\ 0 & \text{ansonsten.} \end{cases} \end{aligned}$$

■

Untersuchen wir nun das Verhalten der Richardson-Iteration für die Matrix $\mathbf{L}$. Wir betrachten wieder das vereinfachte System $\mathbf{L}\mathbf{u} = \mathbf{0}$, das die exakte Lösung $\mathbf{u} = \mathbf{0}$ besitzt. Den Startvektor $\mathbf{u}^{(0)} \in \mathbb{R}^{\mathcal{I}}$ stellen wir in der Eigenvektorbasis dar und erhalten

$$\mathbf{u}^{(0)} = \sum_{k \in \mathcal{I}} \alpha_k \mathbf{e}^k,$$

so dass sich der Iterationsschritt

$$\mathbf{u}^{(1)} = \mathbf{u}^{(0)} - \theta \mathbf{L}\mathbf{u}^{(0)}$$

in der Form

$$\mathbf{u}^{(1)} = \sum_{k \in \mathcal{I}} \alpha_k (1 - \theta \lambda_k) \mathbf{e}^k \tag{1.6}$$

darstellen lässt. Für die weiteren Iterierten erhalten wir die Darstellung

$$\mathbf{u}^{(m)} = \sum_{k \in \mathcal{I}} \alpha_k (1 - \theta \lambda_k)^m \mathbf{e}^k$$

in der Eigenvektorbasis, also konvergiert die Iteration gegen die korrekte Lösung $\mathbf{u} = \mathbf{0}$, falls

$$\max\{|1 - \theta \lambda_k| \ : \ k \in \mathcal{I}\} < 1 \tag{1.7}$$

gilt. Glücklicherweise genügt es, nur den größten und den kleinsten Eigenwert zu untersuchen, also

$$\lambda_1 = 4h^{-2} \sin^2(\pi h/2), \qquad \lambda_N = 4h^{-2} \sin^2(\pi N h/2) = 4h^{-2} \cos^2(\pi h/2).$$

Wenn wir $\max\{|1 - \theta\lambda_1|, |1 - \theta\lambda_N|\} < 1$ bewiesen haben, gilt auch die Bedingung (1.7). Wir wählen θ so, dass diese Größe minimiert wird, nämlich als Lösung von

$$1 - \theta_{\text{opt}} \lambda_1 = \theta_{\text{opt}} \lambda_N - 1.$$

Diese Lösung ist durch

$$\theta_{\text{opt}} := \frac{2}{\lambda_N + \lambda_1} = \frac{2}{4h^{-2}} = \frac{h^2}{2}$$

gegeben und führt zu der Schranke

$$\max\{|1 - \theta_{\text{opt}} \lambda_k| \ : \ k \in \mathcal{I}\} = \varrho := 1 - \theta_{\text{opt}} \lambda_1 = 1 - \frac{2\lambda_1}{\lambda_N + \lambda_1} = \frac{\lambda_N - \lambda_1}{\lambda_N + \lambda_1} < 1.$$

Mit dieser Schranke und der Dreiecksungleichung erhalten wir

$$\|\mathbf{u}^{(m)}\| \le \varrho^m \sum_{k \in \mathcal{I}} \|\alpha_k \mathbf{e}^k\|$$

für jede beliebige Norm. Für die euklidische Norm gilt sogar

$$\|\mathbf{u}^{(m)}\|_2 \le \varrho^m \|\mathbf{u}^{(0)}\|_2,$$

da wir bereits nachgewiesen haben, dass die Eigenvektoren senkrecht zueinander stehen. Die Richardson-Iteration konvergiert also. Leider konvergiert sie nicht sehr schnell: Wir haben

$$\begin{aligned} \varrho = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} &= \frac{4h^{-2}\cos^2(\pi h/2) - 4h^{-2}\sin^2(\pi h/2)}{4h^{-2}} \\ &= \cos^2(\pi h/2) - \sin^2(\pi h/2) = 1 - 2\sin^2(\pi h/2), \end{aligned} \tag{1.8}$$

und für $h \to 0$, also $N \to \infty$, bedeutet diese Gleichung $\varrho \approx 1 - \pi^2 h^2/2$, die Konvergenzrate wird also schlechter, wenn die Problemdimension wächst.

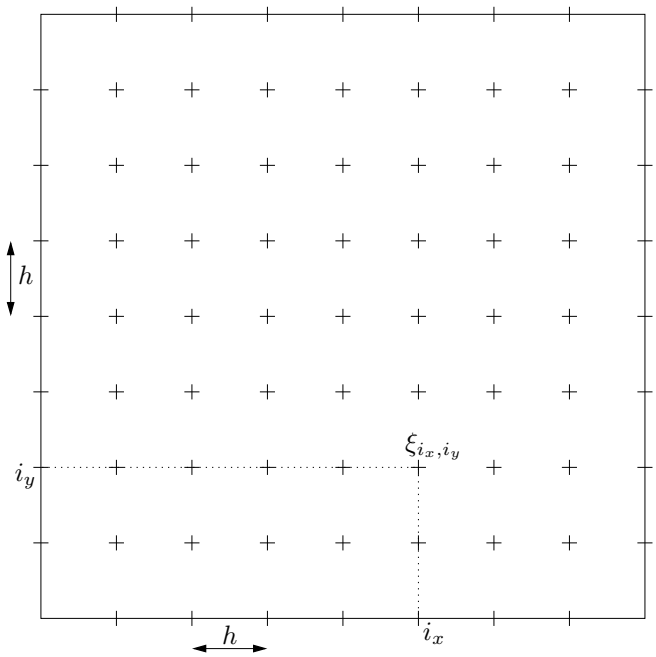

Abbildung 1.5: Zweidimensionales Modellproblem für $N = 7$

1.5 Zweidimensionales Modellproblem

Während das eindimensionale Modellproblem (und auch viele andere eindimensionale Probleme) zu Matrizen geringer Bandbreite führt und somit effizient mit der Band-LR-Zerlegung gehandhabt werden kann, wird die Situation wesentlich schwieriger, wenn wir zu einem zweidimensionalen Problem übergehen: Wir untersuchen die partielle Differentialgleichung

$$-\Delta u(\mathbf{x}) := \left(-\frac{\partial^2}{\partial x_1^2} - \frac{\partial^2}{\partial x_2^2}\right) u(\mathbf{x}) = f(\mathbf{x}) \qquad \text{für alle } \mathbf{x} \in \Omega := (0,1)^2, \tag{1.9a}$$

$$u(\mathbf{x}) = 0 \qquad \text{für alle } \mathbf{x} \in \partial\Omega = \{0,1\} \times [0,1] \cup [0,1] \times \{0,1\}, \tag{1.9b}$$

für eine Funktion $u \in C(\overline{\Omega})$ mit $u|_\Omega \in C^2(\Omega)$ und eine rechte Seite $f \in C(\Omega)$.

Diese Differentialgleichung ist das zweidimensionale Gegenstück zu der Gleichung (1.3). Wir diskretisieren sie, indem wir eine endliche Anzahl von Punkten in Ω fixieren und die zweite Ableitung mit Hilfe der Funktionswerte in diesen Punkten approximieren. Die Punkte konstruieren wir ähnlich wie im eindimensionalen Fall: Wir fixieren $N \in \mathbb{N}$, setzen $h := 1/(N+1)$ und wählen

$$\xi_{i_x,i_y} := (hi_x, hi_y) \qquad \text{für alle } i_x, i_y \in \{0, \ldots, N+1\}.$$

Infolge der Randbedingung sind nur die Werte

$$u_{i_x,i_y} := u(\xi_{i_x,i_y}) \qquad \text{für alle } i_x, i_y \in \{1, \dots, N\}$$

zu bestimmen.

Zur Vereinfachung der Notation fassen wir i_x und i_y zu einem Multiindex $i := (i_x, i_y)$ zusammen und bezeichnen die Menge aller dieser Multiindizes mit

$$\mathcal{I} := \{i = (i_x, i_y) \ : \ i_x, i_y \in \{1, \dots, N\}\}.$$

Die Mächtigkeit der neuen Indexmenge $\mathcal{I}$ beträgt $n := N^2$, und mit ihrer Hilfe können wir die Notationen des eindimensionalen Problems weiterverwenden: $\xi_i = (hi_x, hi_y)$, $u_i = u(\xi_i) = u(hi_x, hi_y)$. Der Vektor der Werte von u schreibt sich als $\mathbf{u} = (u_i)_{i\in\mathcal{I}}$.

Um die Differentialgleichung (1.9) zu diskretisieren, ersetzen wir wieder die zweiten Ableitungen durch die Approximation (1.5) und erhalten

$$\begin{aligned}
&\frac{1}{h^2}\left(u\begin{pmatrix}x_1-h\\x_2\end{pmatrix} - 2u\begin{pmatrix}x_1\\x_2\end{pmatrix} + u\begin{pmatrix}x_1+h\\x_2\end{pmatrix}\right) \approx \frac{\partial^2}{\partial x_1^2}u\begin{pmatrix}x_1\\x_2\end{pmatrix},\\
&\frac{1}{h^2}\left(u\begin{pmatrix}x_1\\x_2-h\end{pmatrix} - 2u\begin{pmatrix}x_1\\x_2\end{pmatrix} + u\begin{pmatrix}x_1\\x_2+h\end{pmatrix}\right) \approx \frac{\partial^2}{\partial x_2^2}u\begin{pmatrix}x_1\\x_2\end{pmatrix} \qquad \text{für alle } \mathbf{x} = \begin{pmatrix}x_1\\x_2\end{pmatrix} \in \Omega.
\end{aligned}$$

Wie bereits im eindimensionalen Fall wenden wir diese Gleichungen auf die Punkte ξ_i an und erhalten

$$\frac{1}{h^2}\left(4u_i - u_{i-(1,0)} - u_{i-(0,1)} - u_{i+(1,0)} - u_{i+(0,1)}\right) \approx -\Delta u(\xi_i) \qquad \text{für alle } i \in \mathcal{I}.$$

Wir führen wieder den Vektor $\mathbf{f} \in \mathbb{R}^{\mathcal{I}}$ mit

$$f_i = f(\xi_i) = f(hi_x, hi_y) \qquad \text{für alle } i \in \mathcal{I}$$

ein und setzen $-\Delta u(\xi_i) = f_i$, um die diskrete Approximation

$$\frac{1}{h^2}\left(4u_i - u_{i-(1,0)} - u_{i-(0,1)} - u_{i+(1,0)} - u_{i+(0,1)}\right) = f_i \qquad \text{für alle } i \in \mathcal{I}$$

der Differentialgleichung (1.9) zu erhalten.

Mit Hilfe der Matrix $\mathbf{L} = (L_{ij})_{i,j\in\mathcal{I}}$, gegeben durch

$$L_{ij} = \begin{cases} 4h^{-2} & \text{falls } i_x = j_x, i_y = j_y,\\ -h^{-2} & \text{falls } |i_x - j_x| = 1, i_y = j_y,\\ -h^{-2} & \text{falls } |i_y - j_y| = 1, i_x = j_x,\\ 0 & \text{sonst} \end{cases} \qquad \text{für alle } i, j \in \mathcal{I},$$

können wir das lineare Gleichungssystem wieder in der kompakten Form

$$\mathbf{Lu} = \mathbf{f}$$

```
procedure MVMModell2D(N, x, var y);
h ← 1/(N + 1);
α ← 4h^{-2}; β ← −h^{-2};
for i_x ∈ {1, ..., N} do
  for i_y ∈ {1, ..., N} do
    y_{i_x,i_y} ← α x_{i_x,i_y};
    if i_x > 1 then
      y_{i_x,i_y} ← y_{i_x,i_y} + β x_{i_x−1,i_y}
    end if;
    if i_x < N then
      y_{i_x,i_y} ← y_{i_x,i_y} + β x_{i_x+1,i_y}
    end if;
    if i_y > 1 then
      y_{i_x,i_y} ← y_{i_x,i_y} + β x_{i_x,i_y−1}
    end if;
    if i_y < N then
      y_{i_x,i_y} ← y_{i_x,i_y} + β x_{i_x,i_y+1}
    end if
  end for
end for
```

Abbildung 1.6: Berechnung von $\mathbf{y} := \mathbf{L}\mathbf{x}$

schreiben. Da $\mathcal{I}$ keine Teilmenge von $\mathbb{N}$ ist, können wir keine direkte Aussage über die Bandbreite von $\mathbf{L}$ treffen. Wir können allerdings versuchen, die Indexmenge $\mathcal{I}$ durch die Menge $\mathbb{N}_n := \{1, \ldots, n\}$ mit $n = |\mathcal{I}| = N^2$ zu ersetzen. Formal geschieht dies durch die Wahl einer bijektiven Abbildung

$$\iota : \mathcal{I} \to \mathbb{N}_n,$$

die die Multiindizes $i = (i_x, i_y) \in \mathcal{I}$ durchnumeriert. Für so ein ι können wir dann Matrizen $\mathbf{L}^{(\iota)} \in \mathbb{R}^{n \times n}$ und Vektoren $\mathbf{u}^{(\iota)}, \mathbf{f}^{(\iota)} \in \mathbb{R}^n$ durch

$$L^{(\iota)}_{\iota(i),\iota(j)} := L_{ij}, \qquad u^{(\iota)}_{\iota(i)} := u_i, \qquad f^{(\iota)}_{\iota(i)} := f_i \qquad \text{für alle } i, j \in \mathcal{I}$$

definieren und feststellen, dass das Gleichungssystem $\mathbf{L}\mathbf{u} = \mathbf{f}$ und das „numerierte“ Gleichungssystem

$$\mathbf{L}^{(\iota)}\mathbf{u}^{(\iota)} = \mathbf{f}^{(\iota)}$$

vollständig gleichwertig sind.

Es stellt sich die Frage, ob es eine von n (also von N) unabhängige Bandbreitenschranke k für $\mathbf{L}^{(\iota)}$ geben kann. Im eindimensionalen Fall war sogar $k = 1$ akzeptabel, im zweidimensionalen Fall ist die Situation komplizierter:

Lemma 1.8 (Minimale Bandbreite) *Sei $k \in \mathbb{N}$ eine Bandbreitenschranke für $\mathbf{L}^{(\iota)}$. Dann gilt $k \geq (N + 1)/4$.*

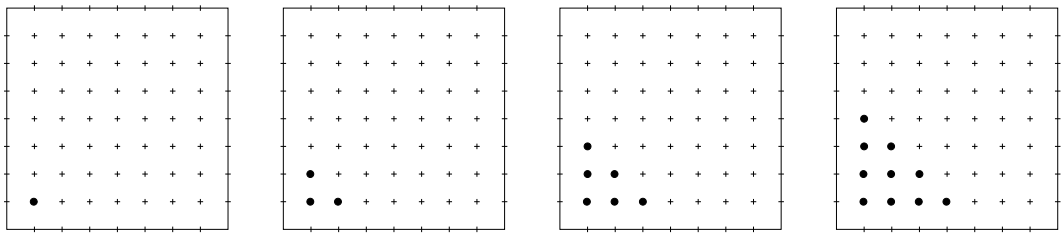

Abbildung 1.7: Mengen $\mathcal{G}_0$, $\mathcal{G}_1$, $\mathcal{G}_2$ und $\mathcal{G}_3$ für den Fall $N = 7$

Beweis. Wir betrachten die Folge

$$\begin{aligned}
\mathcal{G}_0 &:= \{(1,1)\},\\
\mathcal{G}_1 &:= \{j \in \mathcal{I} \ : \ \exists i \in \mathcal{G}_0 \ : \ L_{ij} \neq 0\} = \{(1,1),(2,1),(1,2)\},\\
\mathcal{G}_2 &:= \{j \in \mathcal{I} \ : \ \exists i \in \mathcal{G}_1 \ : \ L_{ij} \neq 0\} = \{(1,1),(2,1),(1,2),(3,1),(2,2),(1,3)\},\\
\mathcal{G}_m &:= \{j \in \mathcal{I} \ : \ \exists i \in \mathcal{G}_{m-1} \ : \ L_{ij} \neq 0\}.
\end{aligned}$$

Man kann sich einfach überlegen, dass

$$|\mathcal{G}_m| = |\mathcal{G}_{m-1}| + m + 1 \qquad \text{für alle } m \in [1 : N-1]$$

gilt (siehe Abbildung 1.7), und die Gauß'sche Summenformel ergibt

$$|\mathcal{G}_m| = \frac{(m+2)(m+1)}{2} \qquad \text{für alle } m \in [1 : N-1]. \tag{1.10}$$

Wir untersuchen nun die Mächtigkeit der Mengen $\iota(\mathcal{G}_m)$. Wir zeigen

$$\iota(\mathcal{G}_m) \subseteq \{j \in \mathbb{N} \ : \ |j - c| \leq km\} \tag{1.11}$$

per Induktion für $c := \iota(1,1)$. Für $m = 0$ ist die Aussage (1.11) offensichtlich.

Nehmen wir nun an, dass (1.11) für ein $m \in \mathbb{N}_0$ gilt. Sei $j \in \mathcal{G}_{m+1}$. Nach Definition gibt es ein $i \in \mathcal{G}_m$ mit $L_{ij} \neq 0$, also gilt auch $L^{(\iota)}_{\iota(i),\iota(j)} \neq 0$. Da k eine Bandbreitenschranke für $\mathbf{L}^{(\iota)}$ ist, folgt $|\iota(i) - \iota(j)| \leq k$. Die Induktionsannahme impliziert $|\iota(i) - c| \leq km$, also erhalten wir $|\iota(j) - c| \leq |\iota(j) - \iota(i)| + |\iota(i) - c| \leq k(m+1)$. Damit ist (1.11) auch für $m + 1$ bewiesen.

Aus (1.11) folgt $\#\iota(\mathcal{G}_m) \leq 2km + 1$, und da ι eine Bijektion ist, erhalten wir

$$\frac{(m+2)(m+1)}{2} = |\mathcal{G}_m| = |\iota(\mathcal{G}_m)| \leq 2km + 1 \qquad \text{für alle } m \in [0 : N-1].$$

Wir haben $k \geq 1$, also folgt

$$\frac{(m+2)(m+1)}{2} \leq 2km + 1 \leq 2km + 2k = 2k(m+1)$$

und damit $k \geq (m+2)/4$. Wir setzen den maximalen Wert $m = N-1$ ein und folgern $k \geq (N+1)/4$. ■

Wenn wir in diesem Beweis $i_0 := (\lfloor (N+1)/2 \rfloor, \lfloor (N+1)/2 \rfloor)$ statt $i_0 := (1,1)$ verwenden, können wir sogar nachweisen, dass $k \geq \lfloor (N+1)/2 \rfloor$ für alle $N > 1$ gelten muss.

Selbst wenn wir also eine Numerierung ι der Indexmenge $\mathcal{I}$ finden können, die dafür sorgt, dass die Bandbreite der Matrix $\mathbf{L}^{(\iota)}$ minimal wird, wären für die Band-LR-Zerlegung immer noch ungefähr $nk^2 \geq nN^2/4 = n^2/4$ Multiplikationen und Subtraktionen erforderlich, die Berechnung hätte also keine lineare Komplexität.

Ein Schritt der Richardson-Iteration, und auch der meisten anderen Iterationsverfahren, hingegen kann mit einem zu n proportionalen Aufwand durchgeführt werden, hier wäre also zu klären, wie schnell die Iterierten gegen die Lösung konvergieren. Zur Untersuchung dieser Frage benötigen wir die Eigenwerte der Matrix $\mathbf{L}$.

Lemma 1.9 (Eigenwerte) *Für alle $k = (k_x, k_y) \in \mathcal{I}$ definieren wir den Vektor $\mathbf{e}^k \in \mathbb{R}^{\mathcal{I}}$ durch*

$$e_i^k := 2h \sin(\pi i_x k_x h) \sin(\pi i_y k_y h) \qquad \text{für alle } i = (i_x, i_y) \in \mathcal{I}.$$

Es gilt

$$\mathbf{L}\mathbf{e}^k = \lambda_k \mathbf{e}^k \qquad \text{mit } \lambda_k := 4h^{-2}(\sin^2(\pi k_x h/2) + \sin^2(\pi k_y h/2))$$

für alle $k = (k_x, k_y) \in \mathcal{I}$, und die Vektoren erfüllen

$$\langle \mathbf{e}^k, \mathbf{e}^\ell \rangle_2 = \begin{cases} 1 & \text{falls } k = \ell, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } k, \ell \in \mathcal{I},$$

bilden also eine aus Eigenvektoren der Matrix $\mathbf{L}$ bestehende Orthonormalbasis von $\mathbb{R}^{\mathcal{I}}$.

Beweis. Sei $k = (k_x, k_y) \in \mathcal{I}$ und $i = (i_x, i_y) \in \mathcal{I}$. Gemäß der Definition von $\mathbf{L}$ haben wir

$$\begin{aligned} (\mathbf{L}\mathbf{e}^k)_i &= h^{-2}(4e_i^k - e_{i-(1,0)}^k - e_{i+(1,0)}^k - e_{i-(0,1)}^k - e_{i+(0,1)}^k) \\ &= h^{-2}(2e_i^k - e_{i-(1,0)}^k - e_{i+(1,0)}^k \\ &\qquad + 2e_i^k - e_{i-(0,1)}^k - e_{i+(0,1)}^k). \end{aligned} \tag{1.12}$$

Wir betrachten zunächst die erste Zeile dieser Gleichung:

$$\begin{aligned} h^{-2}(2e_i^k - e_{i-(1,0)}^k - e_{i+(1,0)}^k) = 2h^{-1} &\sin(\pi i_y k_y h)(2\sin(\pi i_x k_x h) \\ &- \sin(\pi(i_x - 1)k_x h) - \sin(\pi(i_x + 1)k_x h)). \end{aligned}$$

Wir verfahren wie im Beweis von Lemma 1.7:

$$2h^{-1}(2e_i^k - e_{i-(1,0)}^k - e_{i+(1,0)}^k) = 2h^{-1} \sin(\pi i_y k_y h)(2 - 2\cos(\pi k h)) \sin(\pi i_x k_x h)$$

$$
\begin{aligned}
&= 8h^{-1}\sin(\pi i_x k_x h)\sin(\pi i_y k_y h)(1-\cos^2(\pi k_x h/2)) \\
&= 8h^{-1}\sin^2(\pi k_x h/2)e_i^k
\end{aligned}
$$

Für die zweite Zeile der Gleichung (1.12) erhalten wir auf demselben Weg

$$h^{-2}(2e_i^k - e_{i-(0,1)}^k - e_{i+(0,1)}^k) = 8h^{-1}\sin^2(\pi k_y h/2)e_i^k,$$

und die Summe der beiden Zeilen ergibt

$$(\mathbf{L}\mathbf{e}^k)_i = 8h^{-1}(\sin^2(\pi k_x h/2) + \sin^2(\pi k_y h/2))e_i^k = \lambda_k e_i^k = \lambda_k(\mathbf{e}^k)_i.$$

Da wir diese Gleichung für alle $i \in \mathcal{I}$ bewiesen haben, muss $\mathbf{e}^k$ ein Eigenvektor zum Eigenwert λ_k sein.

Der Nachweis der Orthonormalität lässt sich einfach auf den in Lemma 1.7 behandelten Fall zurückführen. ■

Mit denselben Argumenten wie im eindimensionalen Fall können wir auch hier schlussfolgern, dass das Richardson-Verfahren für den optimalen Parameter $\theta_{\text{opt}} = h^2/4$ mit der Rate $\varrho = 1 - 2\sin^2(\pi h/2)$ konvergiert.

Für das zweidimensionale Modellproblem werden also sowohl die Band-LR-Zerlegung als auch die Richardson-Iteration sehr ineffizient, wenn N groß wird.

2 Lineare Iterationsverfahren

Wir haben gesehen, dass das zweidimensionale Modellproblem sich mit einfachen Methoden nur schlecht behandeln lässt: Der Aufwand für die Band-LR-Zerlegung wächst zu schnell, und die Richardson-Iteration konvergiert zu langsam.

Wenn wir das Problem effizient lösen wollen, müssen wir uns also auf die Suche nach anderen Algorithmen begeben, und ein guter Ausgangspunkt sind die iterativen Methoden, weil sie sehr einfach zu programmieren und flexibel sind.

2.1 Allgemeine lineare Iterationsverfahren

Sei $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, und sei $\mathbf{A} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ eine reguläre Matrix. Unser Ziel ist die Berechnung einer beliebig guten Approximation der Lösung $\mathbf{x} \in \mathbb{K}^{\mathcal{I}}$ eines linearen Gleichungssystems $\mathbf{Ax} = \mathbf{b}$ mit der rechten Seite $\mathbf{b} \in \mathbb{K}^{\mathcal{I}}$ (vgl. (1.1)).

Definition 2.1 (Iterationsverfahren) *Eine Abbildung*

$$\Phi : \mathbb{K}^{\mathcal{I}} \times \mathbb{K}^{\mathcal{I}} \to \mathbb{K}^{\mathcal{I}},$$

die im ersten Argument stetig ist, bezeichnen wir als Iterationsverfahren.

Diese Definition ist so zu interpretieren, dass Φ einer aktuellen Iterierten (im ersten Argument) und der rechten Seite (im zweiten) eine neue Iterierte zuordnet.

Selbstverständlich wird bei jedem sinnvollen Iterationsverfahren die Abbildung Φ auch von der Matrix $\mathbf{A}$ abhängen, aber diese Abhängigkeit nehmen wir nicht explizit in die Notation auf, sondern setzen sie implizit voraus.

Definition 2.2 (Iterierte) *Sei Φ ein Iterationsverfahren, und seien $\mathbf{b}, \mathbf{x}^{(0)} \in \mathbb{K}^{\mathcal{I}}$. Die durch*

$$\mathbf{x}^{(m)} := \Phi(\mathbf{x}^{(m-1)}, \mathbf{b}) \qquad \text{für alle } m \in \mathbb{N}$$

definierte Folge $(\mathbf{x}^{(m)})_{m\in\mathbb{N}_0}$ bezeichnen wir als Folge der Iterierten *zu dem Verfahren Φ, der rechten Seite $\mathbf{b}$ und dem Startvektor $\mathbf{x}^{(0)}$.*

Wir sind an Verfahren interessiert, die eine beliebig genaue Approximation der Lösung $\mathbf{x}$ des Gleichungssystems (1.1) berechnen. Damit diese Lösung tatsächlich *beliebig* genau werden kann, muss also die Folge der Iterierten $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots$ konvergieren.

Definition 2.3 (Konvergenz) *Ein Iterationsverfahren Φ heißt* konvergent, *falls für alle $\mathbf{b} \in \mathbb{K}^{\mathcal{I}}$ ein $\mathbf{x}^* \in \mathbb{K}^{\mathcal{I}}$ so existiert, dass für jeden Startvektor $\mathbf{x}^{(0)} \in \mathbb{K}^{\mathcal{I}}$ die Folge $(\mathbf{x}^{(m)})_{m\in\mathbb{N}_0}$ der Iterierten gegen den Grenzwert $\mathbf{x}^* \in \mathbb{K}^{\mathcal{I}}$ konvergiert.*

Bei dieser Definition ist wichtig, dass der Grenzwert unabhängig vom Startvektor sein soll: Im vorangegangenen Kapitel haben wir in (1.2) ein Beispiel für ein lineares Gleichungssystem gesehen, bei dem das Richardson-Verfahren nicht für alle Startvektoren konvergiert, für spezielle (etwa $\mathbf{x}^{(0)} = (1, 0)$ und $\theta = 1$) jedoch sehr wohl.

Es genügt im Allgemeinen nicht, dass die Folge konvergiert, sie muss auch gegen die richtige Lösung $\mathbf{x}$ des Gleichungssystems (1.1) konvergieren. Wir charakterisieren die Grenzwerte der Folgen mit Hilfe von Fixpunkten:

Definition 2.4 (Fixpunkt) *Ein Vektor $\mathbf{x}^* \in \mathbb{K}^{\mathcal{I}}$ heißt* Fixpunkt *eines Iterationsverfahrens Φ zu einem Vektor $\mathbf{b} \in \mathbb{K}^{\mathcal{I}}$, falls $\Phi(\mathbf{x}^*, \mathbf{b}) = \mathbf{x}^*$ gilt.*

Lemma 2.5 *Sei Φ ein Iterationsverfahren. Seien $\mathbf{b}, \mathbf{x}^{(0)} \in \mathbb{K}^{\mathcal{I}}$ so gegeben, dass die Folge $(\mathbf{x}^{(m)})_{m\in\mathbb{N}_0}$ der Iterierten gegen einen Grenzwert $\mathbf{x}^* \in \mathbb{K}^{\mathcal{I}}$ konvergiert. Dann ist $\mathbf{x}^*$ ein Fixpunkt von Φ zu $\mathbf{b}$.*

Beweis. Sei $\epsilon \in \mathbb{R}_{>0}$. Da Φ im ersten Argument stetig ist, finden wir ein $\delta \in \mathbb{R}_{>0}$ so, dass für alle $\mathbf{y} \in \mathbb{K}^{\mathcal{I}}$ mit $\|\mathbf{x}^* - \mathbf{y}\| \leq \delta$ die Ungleichung $\|\Phi(\mathbf{x}^*, \mathbf{b}) - \Phi(\mathbf{y}, \mathbf{b})\| \leq \epsilon$ gilt.

Da die Folge $(\mathbf{x}^{(m)})_{m\in\mathbb{N}_0}$ gegen $\mathbf{x}^*$ konvergiert, gibt es ein $m_0 \in \mathbb{N}_0$ so, dass

$$\|\mathbf{x}^* - \mathbf{x}^{(m)}\| \leq \min\{\epsilon, \delta\} \qquad \text{für alle } m \in \mathbb{N}_{\geq m_0}$$

gilt. Wir wählen ein $m \in \mathbb{N}_{\geq m_0}$ und erhalten

$$\begin{aligned}\|\Phi(\mathbf{x}^*, \mathbf{b}) - \mathbf{x}^*\| &= \|\Phi(\mathbf{x}^*, \mathbf{b}) - \Phi(\mathbf{x}^{(m)}, \mathbf{b}) + \Phi(\mathbf{x}^{(m)}, \mathbf{b}) - \mathbf{x}^*\| \\ &\leq \|\Phi(\mathbf{x}^*, \mathbf{b}) - \Phi(\mathbf{x}^{(m)}, \mathbf{b})\| + \|\mathbf{x}^* - \mathbf{x}^{(m+1)}\| \leq 2\epsilon.\end{aligned}$$

Da ϵ beliebig gewählt wurde, folgt $\Phi(\mathbf{x}^*, \mathbf{b}) = \mathbf{x}^*$. ∎

Wir suchen nach Verfahren, die gegen die Lösung $\mathbf{x}$ konvergieren, wir müssten also eigentlich fordern, dass das Iterationsverfahren genau einen Fixpunkt besitzt, der mit $\mathbf{x}$ übereinstimmt. Diese Eigenschaft ist praktisch nur schwer nachzuprüfen.

Sehr viel brauchbarer ist die folgende Definition:

Definition 2.6 (Konsistenz) *Ein Iterationsverfahren Φ heißt* konsistent*, falls die Lösung $\mathbf{x}$ des Gleichungssystems (1.1) ein Fixpunkt von Φ ist, also die Gleichung $\Phi(\mathbf{x}, \mathbf{b}) = \mathbf{x}$ erfüllt.*

Lemma 2.7 *Sei Φ konsistent und konvergent, sei $\mathbf{b} \in \mathbb{K}^{\mathcal{I}}$. Dann konvergiert die Folge $(\mathbf{x}^{(m)})_{m\in\mathbb{N}_0}$ der Iterierten zu jedem beliebigen Startvektor $\mathbf{x}^{(0)} \in \mathbb{K}^{\mathcal{I}}$ gegen die Lösung $\mathbf{x}$ des Gleichungssystems (1.1).*

Beweis. Da Φ konsistent ist, ist die Folge der Iterierten zum Startvektor $\mathbf{x}$ konstant, besitzt also insbesondere den Grenzwert $\mathbf{x}^* := \mathbf{x}$.

Da Φ auch konvergent ist, muss auch die Folge der Iterierten zu jedem anderen Startvektor $\mathbf{x}^{(0)} \in \mathbb{K}^{\mathcal{I}}$ gegen $\mathbf{x}$ konvergieren. ∎

Da wir ein *lineares* Gleichungssystem lösen möchten, bietet es sich an, auch ein *lineares* Iterationsverfahren zu betrachten, also zu verlangen, dass Φ eine lineare Abbildung ist.

Definition 2.8 (Lineares Iterationsverfahren) *Ein Iterationsverfahren* Φ *heißt* linear*, falls es Matrizen* $\mathbf{M}, \mathbf{N} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ *so gibt, dass*

$$\Phi(\mathbf{x}, \mathbf{b}) = \mathbf{M}\mathbf{x} + \mathbf{N}\mathbf{b} \qquad \textit{für alle } \mathbf{x}, \mathbf{b} \in \mathbb{K}^{\mathcal{I}} \textit{ gilt.} \tag{2.1}$$

Diese Darstellung des Iterationsverfahrens bezeichnet man als erste Normalform *und die Matrix* $\mathbf{M}$ *als* Iterationsmatrix*.*

Umgekehrt gehört zu jedem Paar von Matrizen $\mathbf{M}$ und $\mathbf{N}$ genau ein lineares Iterationsverfahren.

Für lineare Iterationsverfahren lässt sich einfach nachprüfen, ob sie konsistent sind:

Lemma 2.9 *Sei* Φ *ein lineares Iterationsverfahren, und seien* $\mathbf{M}, \mathbf{N} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ *die korrespondierenden Matrizen der ersten Normalform.* Φ *ist genau dann konsistent, wenn* $\mathbf{M} = \mathbf{I} - \mathbf{N}\mathbf{A}$ *gilt.*

Beweis. Sei zunächst Φ konsistent. Wir wählen ein $\mathbf{x} \in \mathbb{K}^{\mathcal{I}}$ und setzen $\mathbf{b} := \mathbf{A}\mathbf{x}$. Da Φ konsistent ist, gilt

$$\mathbf{x} = \Phi(\mathbf{x}, \mathbf{b}) = \mathbf{M}\mathbf{x} + \mathbf{N}\mathbf{b} = \mathbf{M}\mathbf{x} + \mathbf{N}\mathbf{A}\mathbf{x},$$

und diese Identität impliziert $\mathbf{I} - \mathbf{N}\mathbf{A} = \mathbf{M}$.

Falls wir umgekehrt diese Gleichung voraussetzen, erhalten wir für die Lösung $\mathbf{x}$ zu der rechten Seite $\mathbf{b}$ die Gleichung

$$\Phi(\mathbf{x}, \mathbf{b}) = \mathbf{M}\mathbf{x} + \mathbf{N}\mathbf{b} = \mathbf{x} - \mathbf{N}\mathbf{A}\mathbf{x} + \mathbf{N}\mathbf{b} = \mathbf{x} - \mathbf{N}(\mathbf{A}\mathbf{x} - \mathbf{b}) = \mathbf{x},$$

also ist $\mathbf{x}$ ein Fixpunkt von Φ. ■

Aufgrund dieses Lemmas lässt sich jedes konsistente lineare Iterationsverfahren in der *zweiten Normalform*

$$\Phi(\mathbf{x}, \mathbf{b}) = \mathbf{x} - \mathbf{N}(\mathbf{A}\mathbf{x} - \mathbf{b}) \qquad \text{für alle } \mathbf{x}, \mathbf{b} \in \mathbb{K}^{\mathcal{I}} \tag{2.2}$$

darstellen, und offenbar ist jedes in dieser Form darstellbare Iterationsverfahren auch konsistent. Besonders gut sind diejenigen Iterationsverfahren, bei denen $\mathbf{N}$ eine gute Approximation der Inversen von $\mathbf{A}$ ist.

2.2 Hinreichendes Konvergenzkriterium

Die Untersuchung des Konvergenzverhaltens eines linearen Iterationsverfahrens ist etwas komplizierter und erfordert einige vorbereitende Resultate.

Bevor wir uns dem allgemeinen Beweis zuwenden, betrachten wir zunächst einen einfacheren Fall, in dem wir die Normen der Iterationsfehler explizit abschätzen können. Zunächst leiten wir dazu eine Darstellung der einzelnen Iterierten eines linearen Iterationsverfahrens her.

Lemma 2.10 *Sei Φ ein lineares Iterationsverfahren, und seien $\mathbf{M}, \mathbf{N} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ die Matrizen seiner ersten Normalform. Für einen Startvektor $\mathbf{x}^{(0)} \in \mathbb{K}^{\mathcal{I}}$ und eine rechte Seite $\mathbf{b} \in \mathbb{K}^{\mathcal{I}}$ gilt dann*

$$\mathbf{x}^{(m)} = \mathbf{M}^m\mathbf{x}^{(0)} + \sum_{\ell=0}^{m-1} \mathbf{M}^\ell\mathbf{N}\mathbf{b} \qquad \textit{für alle } m \in \mathbb{N}_0.$$

Beweis. Per Induktion über $m \in \mathbb{N}_0$. Für $m = 0$ gilt die Behauptung offenbar. Sei nun $m \in \mathbb{N}_0$ so gewählt, dass die Gleichung gilt. Wir erhalten

$$\begin{aligned}
\mathbf{x}^{(m+1)} &= \Phi(\mathbf{x}^{(m)}, \mathbf{b}) = \mathbf{M}\mathbf{x}^{(m)} + \mathbf{N}\mathbf{b} \\
&= \mathbf{M}\left(\mathbf{M}^m\mathbf{x}^{(0)} + \sum_{\ell=0}^{m-1} \mathbf{M}^\ell\mathbf{N}\mathbf{b}\right) + \mathbf{N}\mathbf{b} = \mathbf{M}^{m+1}\mathbf{x}^{(0)} + \sum_{\ell=1}^{m} \mathbf{M}^\ell\mathbf{N}\mathbf{b} + \mathbf{N}\mathbf{b} \\
&= \mathbf{M}^{m+1}\mathbf{x}^{(0)} + \sum_{\ell=0}^{m} \mathbf{M}^\ell\mathbf{N}\mathbf{b},
\end{aligned}$$

also gilt die Gleichung auch für $m + 1$. ∎

Der Grenzwert der Folge der Iterierten kann also nur dann von der Wahl des Startvektors $\mathbf{x}^{(0)}$ unabhängig sein, falls $\mathbf{M}^m\mathbf{x}^{(0)}$ für alle Startvektoren $\mathbf{x}^{(0)}$ gegen Null konvergiert.

Die Konvergenz gegen einen Fixpunkt lässt sich besonders einfach charakterisieren, indem wir den Iterationsfehler $\mathbf{x}^{(m)} - \mathbf{x}^*$ analysieren:

Lemma 2.11 (Explizite Fehlerdarstellung) *Sei Φ ein lineares Iterationsverfahren, und seien $\mathbf{M}, \mathbf{N} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ die Matrizen seiner ersten Normalform. Seien $\mathbf{x}^{(0)}, \mathbf{b} \in \mathbb{K}^{\mathcal{I}}$ gegeben. Falls $\mathbf{x}^* \in \mathbb{K}^{\mathcal{I}}$ ein Fixpunkt der Iteration Φ zur rechten Seite $\mathbf{b}$ ist, gelten für den Iterationsfehler die Gleichungen*

$$\mathbf{x}^{(m+1)} - \mathbf{x}^* = \mathbf{M}(\mathbf{x}^{(m)} - \mathbf{x}^*) = \mathbf{M}^{m+1}(\mathbf{x}^{(0)} - \mathbf{x}^*) \qquad \textit{für alle } m \in \mathbb{N}_0. \tag{2.3}$$

Beweis. Sei $m \in \mathbb{N}_0$. Da $\mathbf{x}^*$ ein Fixpunkt ist, gilt

$$\mathbf{x}^* = \Phi(\mathbf{x}^*, \mathbf{b}) = \mathbf{M}\mathbf{x}^* + \mathbf{N}\mathbf{b}.$$

Aus der Definition folgt

$$\mathbf{x}^{(m+1)} = \Phi(\mathbf{x}^{(m)}, \mathbf{b}) = \mathbf{M}\mathbf{x}^{(m)} + \mathbf{N}\mathbf{b},$$

also erhalten wir für die Differenz

$$\mathbf{x}^{(m+1)} - \mathbf{x}^* = \mathbf{M}(\mathbf{x}^{(m)} - \mathbf{x}^*).$$

Eine einfache Induktion vervollständigt den Beweis. ∎

Auch hier sehen wir, dass das Verfahren gegen einen eindeutigen Fixpunkt konvergiert, falls $\mathbf{M}^m$ für $m \to \infty$ gegen null konvergiert. Die Fehlerdarstellung bietet aber

auch die Möglichkeit, die Konvergenz auf Teilräumen zu analysieren: Falls beispielsweise der Startfehler $\mathbf{x}^{(0)} - \mathbf{x}^*$ aus einem Eigenraum von $\mathbf{M}$ zu einem Eigenwert $\lambda \in (-1, 1)$ stammt, wird die Folge der Fehler wie λ^m konvergieren, selbst wenn die Folge der Matrizen $\mathbf{M}^m$ divergieren sollte.

In der Regel sind wir daran interessiert, die Konvergenz auch quantitativ zu erfassen, also die Konvergenzgeschwindigkeit zu beschreiben. Dazu verwenden wir eine geeignete Matrixnorm:

Lemma 2.12 (Induzierte Matrixnorm) *Sei $\|\cdot\| : \mathbb{K}^{\mathcal{I}} \to \mathbb{R}_{\geq 0}$ eine Norm. Dann ist die Abbildung*

$$f : \mathbb{K}^{\mathcal{I}\times\mathcal{I}} \to \mathbb{R}_{\geq 0}, \qquad \mathbf{X} \mapsto \sup\left\{ \frac{\|\mathbf{X}\mathbf{y}\|}{\|\mathbf{y}\|} \ : \ \mathbf{y} \in \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\} \right\}$$

eine Norm auf dem Raum $\mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ der Matrizen.

Da Verwechslungen ausgeschlossen sind, verwenden wir die Notation $\|\mathbf{X}\| := f(\mathbf{X})$ für alle $\mathbf{X} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ und bezeichnen diese Norm als die von $\|\cdot\|$ induzierte Matrixnorm.

Beweis. Um zu zeigen, dass f reellwertig ist, stellen wir fest, dass

$$f(\mathbf{X}) = \sup\left\{ \|\mathbf{X}\mathbf{y}\| \ : \ \mathbf{y} \in \mathbb{K}^{\mathcal{I}} \text{ mit } \|\mathbf{y}\| = 1 \right\} \qquad \text{für alle } \mathbf{X} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$$

gilt. Nach dem Satz von Heine-Borel ist die Menge $\mathcal{S} := \{\mathbf{x} \in \mathbb{K}^{\mathcal{I}} \ : \ \|\mathbf{x}\| = 1\}$ kompakt, also besitzt die stetige Abbildung $g_X : \mathcal{S} \to \mathbb{R}_{\geq 0}, \mathbf{y} \mapsto \|\mathbf{X}\mathbf{y}\|$ auf ihr ein endliches Maximum. Damit gilt $f(\mathbf{X}) = \max g_X(\mathcal{S}) \in \mathbb{R}_{\geq 0}$ für alle Matrizen $\mathbf{X} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$.

Die Dreiecksungleichung und die absolute Homogenität der Funktion f folgen aus den entsprechenden Eigenschaften der Norm $\|\cdot\|$. Falls $f(\mathbf{X}) = 0$ gilt, folgt $\|\mathbf{X}\mathbf{y}\| = 0$ für alle $\mathbf{y} \in \mathbb{K}^{\mathcal{I}}$, also $\mathbf{X} = \mathbf{0}$. ∎

Lemma 2.13 (Submultiplikativität) *Sei $\|\cdot\| : \mathbb{K}^{\mathcal{I}} \to \mathbb{R}_{\geq 0}$ eine Norm. Es gilt*

$$\|\mathbf{X}\mathbf{y}\| \leq \|\mathbf{X}\|\,\|\mathbf{y}\| \qquad \textit{für alle } \mathbf{X} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}},\ \mathbf{y} \in \mathbb{K}^{\mathcal{I}}, \tag{2.4a}$$

$$\|\mathbf{X}\mathbf{Y}\| \leq \|\mathbf{X}\|\,\|\mathbf{Y}\| \qquad \textit{für alle } \mathbf{X}, \mathbf{Y} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}. \tag{2.4b}$$

Beweis. Seien eine Matrix $\mathbf{X} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ und ein Vektor $\mathbf{y} \in \mathbb{K}^{\mathcal{I}}$ gegeben. Für $\mathbf{y} = \mathbf{0}$ ist die Aussage trivial. Ansonsten gilt $\|\mathbf{y}\| \neq 0$ und wir haben

$$\|\mathbf{X}\mathbf{y}\| = \frac{\|\mathbf{X}\mathbf{y}\|}{\|\mathbf{y}\|}\|\mathbf{y}\| \leq \sup\left\{ \frac{\|\mathbf{X}\mathbf{z}\|}{\|\mathbf{z}\|} \ : \ \mathbf{z} \in \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\} \right\} \|\mathbf{y}\| = \|\mathbf{X}\|\,\|\mathbf{y}\|.$$

Zum Beweis der Abschätzung (2.4b) seien $\mathbf{X}, \mathbf{Y} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ und ein Vektor $\mathbf{z} \in \mathbb{K}^{\mathcal{I}}$ gegeben. Dank (2.4a) erhalten wir

$$\|\mathbf{X}\mathbf{Y}\mathbf{z}\| \leq \|\mathbf{X}\|\,\|\mathbf{Y}\mathbf{z}\| \leq \|\mathbf{X}\|\,\|\mathbf{Y}\|\,\|\mathbf{z}\|,$$

also auch

$$\begin{aligned}\|\mathbf{XY}\| &= \sup\left\{\frac{\|\mathbf{XYz}\|}{\|\mathbf{z}\|} \;:\; \mathbf{z} \in \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\}\right\} \\ &\leq \sup\left\{\frac{\|\mathbf{X}\|\,\|\mathbf{Y}\|\,\|\mathbf{z}\|}{\|\mathbf{z}\|} \;:\; \mathbf{z} \in \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\}\right\} = \|\mathbf{X}\|\,\|\mathbf{Y}\|.\end{aligned}$$

■

Wie wir in Lemma 2.10 gesehen habe, spielen Potenzen der Iterationsmatrix $\mathbf{M}$ und die zu dieser Matrix gehörende Potenzreihe eine wichtige Rolle bei der Untersuchung des Konvergenzverhaltens. Die Eigenschaften der betreffenden Potenzreihe sind im folgenden Hilfssatz zusammengefasst:

Lemma 2.14 (Neumannsche Reihe) *Sei $\|\cdot\| : \mathbb{K}^{\mathcal{I}} \to \mathbb{R}_{\geq 0}$ eine Norm. Sei $\mathbf{X} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ eine Matrix. Dann gilt*

$$(\mathbf{I} - \mathbf{X}) \sum_{\ell=0}^{m} \mathbf{X}^{\ell} = \mathbf{I} - \mathbf{X}^{m+1} \qquad \textit{für alle } m \in \mathbb{N}_0. \tag{2.5}$$

Falls $\|\mathbf{X}\| < 1$ gilt, ist $\mathbf{I} - \mathbf{X}$ invertierbar und es gelten

$$(\mathbf{I} - \mathbf{X})^{-1} = \sum_{\ell=0}^{\infty} \mathbf{X}^{\ell}, \qquad \|(\mathbf{I} - \mathbf{X})^{-1}\| \leq \frac{1}{1 - \|\mathbf{X}\|}. \tag{2.6}$$

Beweis. Wir wählen ein $m \in \mathbb{N}_0$ und erhalten

$$(\mathbf{I} - \mathbf{X}) \sum_{\ell=0}^{m} \mathbf{X}^{\ell} = \sum_{\ell=0}^{m} \mathbf{X}^{\ell} - \sum_{\ell=1}^{m+1} \mathbf{X}^{\ell} = \mathbf{I} - \mathbf{X}^{m+1}.$$

Das ist gerade (2.5). Gelte nun $\|\mathbf{X}\| < 1$. Wir definieren

$$\mathbf{Y}^{(m)} := \sum_{\ell=0}^{m} \mathbf{X}^{\ell} \qquad \text{für alle } m \in \mathbb{N}_0$$

und erhalten mit der Submultiplikativität (2.4b) der Matrixnorm und der geometrischen Summenformel

$$\|\mathbf{Y}^{(m)}\| = \left\|\sum_{\ell=0}^{m} \mathbf{X}^{\ell}\right\| \leq \sum_{\ell=0}^{m} \|\mathbf{X}\|^{\ell} \leq \sum_{\ell=0}^{\infty} \|\mathbf{X}\|^{\ell} = \frac{1}{1 - \|\mathbf{X}\|}. \tag{2.7}$$

Also ist die Neumannsche Reihe absolut summierbar. Insbesondere ist $(\mathbf{Y}^{(m)})_{m=0}^{\infty}$ eine Cauchy-Folge: Sei $\epsilon \in \mathbb{R}_{>0}$. Wir finden ein $m_0 \in \mathbb{N}$ mit

$$\frac{\|\mathbf{X}\|^{m_0}}{1 - \|\mathbf{X}\|} \leq \epsilon.$$

Für $n, m \in \mathbb{N}$ mit $m_0 \leq m < n$ folgt mit (2.4b) und (2.7) die Abschätzung

$$\begin{aligned}\|\mathbf{Y}^{(n)} - \mathbf{Y}^{(m)}\| &= \left\| \sum_{\ell=m+1}^{n} \mathbf{X}^{\ell} \right\| = \left\| \mathbf{X}^{m+1} \sum_{k=0}^{n-m-1} \mathbf{X}^{k} \right\| = \|\mathbf{X}^{m+1}\mathbf{Y}^{(n-m-1)}\| \\ &\leq \|\mathbf{X}\|^{m+1}\|\mathbf{Y}^{(n-m-1)}\| \leq \frac{\|\mathbf{X}\|^{m+1}}{1-\|\mathbf{X}\|} < \frac{\|\mathbf{X}\|^{m_0}}{1-\|\mathbf{X}\|} \leq \epsilon,\end{aligned}$$

also ist $(\mathbf{Y}^{(m)})_{m=0}^{\infty}$ eine Cauchy-Folge. Da $\mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ ein vollständiger Raum ist, besitzt diese Folge einen Grenzwert, den wir mit

$$\mathbf{Y} := \lim_{m\to\infty} \mathbf{Y}^{(m)} = \sum_{\ell=0}^{\infty} \mathbf{X}^{\ell}.$$

bezeichnen. Wegen $\|\mathbf{X}\| < 1$ und (2.4b) gilt insbesondere auch

$$\|\lim_{m\to\infty} \mathbf{X}^{m}\| = \lim_{m\to\infty} \|\mathbf{X}^{m}\| \leq \lim_{m\to\infty} \|\mathbf{X}\|^{m} = 0,$$

also $\lim_{m\to\infty} \mathbf{X}^{m} = \mathbf{0}$. Zusammen mit (2.5) erhalten wir

$$(\mathbf{I}-\mathbf{X})\mathbf{Y} = (\mathbf{I}-\mathbf{X}) \lim_{m\to\infty} \mathbf{Y}^{(m)} = \lim_{m\to\infty} (\mathbf{I}-\mathbf{X}) \sum_{\ell=0}^{m} \mathbf{X}^{\ell} = \lim_{m\to\infty} \mathbf{I} - \mathbf{X}^{m+1} = \mathbf{I},$$

also ist $\mathbf{Y}$ die Inverse der Matrix $\mathbf{I}-\mathbf{X}$. Aus (2.7) folgt schließlich

$$\|\mathbf{Y}\| = \lim_{m\to\infty} \|\mathbf{Y}^{(m)}\| \leq \frac{1}{1-\|\mathbf{X}\|},$$

und (2.6) ist bewiesen. ■

Lemma 2.15 (Konvergenzkriterium) *Sei Φ ein lineares Iterationsverfahren mit Iterationsmatrix $\mathbf{M} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$. Sei $\|\cdot\|$ eine Norm auf $\mathbb{K}^{\mathcal{I}}$, für die*

$$\|\mathbf{M}\| < 1 \tag{2.8}$$

gilt. Dann ist Φ konvergent, und für eine rechte Seite $\mathbf{b} \in \mathbb{K}^{\mathcal{I}}$ ist der Grenzwert durch $\mathbf{x}^* := (\mathbf{I}-\mathbf{M})^{-1}\mathbf{N}\mathbf{b}$ gegeben.

Beweis. Wir weisen nach, dass $\mathbf{M}$ die Voraussetzungen von Lemma 2.14 erfüllt.

Nach Lemma 2.13 gilt

$$\|(\mathbf{I}-\mathbf{M})\mathbf{x}\| \geq \|\mathbf{x}\| - \|\mathbf{M}\mathbf{x}\| \geq \|\mathbf{x}\| - \|\mathbf{M}\|\,\|\mathbf{x}\| = (1-\|\mathbf{M}\|)\|\mathbf{x}\|,$$

und da $\|\mathbf{M}\| < 1$ vorausgesetzt ist, kann der Kern von $\mathbf{I}-\mathbf{M}$ nur den Nullvektor enthalten. Also muss $\mathbf{I}-\mathbf{M}$ invertierbar sein.

Mit einer einfachen Induktion können wir zeigen, dass die in Lemma 2.13 nachgewiesene Submultiplikativität (2.4b) der induzierten Matrixnorm die Ungleichung

$$\|\mathbf{M}^m\| \leq \|\mathbf{M}\|^m \qquad \text{für alle } m \in \mathbb{N},$$

impliziert, und da $\|\mathbf{M}\| < 1$ vorausgesetzt ist, erhalten wir $\lim_{m\to\infty} \|\mathbf{M}^m\| = 0$.

Also können wir Lemma 2.10 mit Lemma 2.14 kombinieren, um

$$\lim_{m\to\infty} \mathbf{x}^{(m)} = \lim_{m\to\infty} \mathbf{M}^m \mathbf{x}^{(0)} + \sum_{\ell=0}^{m-1} \mathbf{M}^\ell \mathbf{N}\mathbf{b} = \left(\sum_{\ell=0}^{\infty} \mathbf{M}^\ell\right) \mathbf{N}\mathbf{b} = (\mathbf{I} - \mathbf{M})^{-1}\mathbf{N}\mathbf{b}$$

zu erhalten. ■

Dieses Kriterium erlaubt es uns, die Konvergenz eines linearen Iterationsverfahrens nachzuweisen, aber die Charakterisierung mit Hilfe einer induzierten Matrixnorm ist nicht vollständig zufriedenstellend: Wir können eine beliebige Norm $\|\cdot\| : \mathbb{K}^2 \to \mathbb{R}_{\geq 0}$ wählen und die durch

$$\mathbf{M}_\alpha := \begin{pmatrix} 0 & \alpha \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{N}_\alpha := \begin{pmatrix} 1 & -\alpha \\ 0 & 1 \end{pmatrix}$$

gegebene lineare Iteration

$$\Phi_\alpha(\mathbf{x}, \mathbf{b}) := \mathbf{M}_\alpha \mathbf{x} + \mathbf{N}_\alpha \mathbf{b} \qquad \text{für alle } \mathbf{x}, \mathbf{b} \in \mathbb{K}^2$$

untersuchen. Falls wir $\alpha \in \mathbb{R}$ groß genug wählen, folgt $\|\mathbf{M}_\alpha\| > 1$, also sind die Bedingungen von Lemma 2.15 nicht erfüllt, aber wegen $\mathbf{M}_\alpha^2 = \mathbf{0}$ wird das Verfahren trotzdem bereits nach zwei Schritten seinen Fixpunkt erreichen.

2.3 Konvergenz und Spektralradius

Den Schlüssel zu einer besseren Charakterisierung der konvergenten linearen Iterationsverfahren bieten die Eigenwerte. In unserem Beispiel besitzt $\mathbf{M}_\alpha$ lediglich den Eigenwert Null, und der Satz von Cayley-Hamilton impliziert, dass auch eine allgemeine Matrix $\mathbf{M} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$, die nur Null als Eigenwert besitzt, die Gleichung $\mathbf{M}^n = \mathbf{0}$ für $n := \#\mathcal{I}$ erfüllt, so dass die Folge der Iterierten nach endlich vielen Schritten ihren Grenzwert tatsächlich annimmt.

In der Praxis können wir nicht verlangen, dass die Iterationsmatrix nur Null als Eigenwert besitzen darf, aber glücklicherweise genügt es auch, wenn wir die Größe der Eigenwerte beschränken können. Allerdings eignen sich die Eigenwerte nur im komplexwertigen Fall für die Charakterisierung einer Matrix: Die Matrix

$$\mathbf{X} = \begin{pmatrix} 0 & \alpha \\ -\alpha & 0 \end{pmatrix}$$

beispielsweise besitzt für $\alpha \in \mathbb{R}_{>0}$ keinen reellen Eigenwert, allerdings folgt aus

$$\mathbf{X}^2 = \begin{pmatrix} -\alpha^2 & 0 \\ 0 & -\alpha^2 \end{pmatrix}$$

trotzdem, dass für $\alpha < 1$ die Matrizen $\mathbf{X}^m$ für $m \to \infty$ gegen null konvergieren werden, während sie für $\alpha > 1$ unbegrenzt wachsen. Im reellwertigen Fall ermöglichen uns Eigenwerte also keine befriedigenden Aussagen. Deshalb werden wir uns im folgenden Abschnitt auf die Untersuchung des komplexwertigen Falls $\mathbb{K} = \mathbb{C}$ beschränken.

Definition 2.16 (Eigenwerte, Eigenvektoren, Spektralradius) *Sei $\mathbf{X} \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$ eine Matrix. Falls für ein $\lambda \in \mathbb{C}$ die Matrix $\lambda\mathbf{I} - \mathbf{X}$ nicht regulär ist, nennen wir λ einen* Eigenwert*. In diesem Fall ist der Kern von $\lambda\mathbf{I} - \mathbf{X}$ nicht trivial, also gibt es mindestens einen Vektor $\mathbf{e} \in \mathbb{C}^{\mathcal{I}} \setminus \{\mathbf{0}\}$ mit $\mathbf{X}\mathbf{e} = \lambda\mathbf{e}$. Derartige Vektoren nennen wir* Eigenvektoren *zum Eigenwert λ.*

Das Spektrum *der Matrix $\mathbf{X}$ ist die Menge ihrer Eigenwerte, definiert durch*

$$\sigma(\mathbf{X}) := \{\lambda \in \mathbb{C} \ : \ \lambda\mathbf{I} - \mathbf{X} \text{ ist nicht regulär}\}.$$

Das Maximum der Beträge der Eigenwerte nennen wir den Spektralradius *von $\mathbf{X}$, er ist definiert durch*

$$\varrho(\mathbf{X}) := \max\{|\lambda| \ : \ \lambda \in \sigma(\mathbf{X})\}.$$

Eine Schranke für den Spektralradius ist im folgenden Sinne eine schwächere Bedingung als eine Schranke einer induzierten Matrixnorm:

Lemma 2.17 *Sei $\|\cdot\|$ eine Norm auf $\mathbb{C}^{\mathcal{I}}$. Für die von ihr induzierte Matrixnorm gilt*

$$\varrho(\mathbf{X}) \le \|\mathbf{X}\| \qquad \text{für alle } \mathbf{X} \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}.$$

Beweis. Sei $\lambda \in \mathbb{C}$ ein Eigenwert von $\mathbf{X}$, und sei $\mathbf{e} \in \mathbb{C}^{\mathcal{I}} \setminus \{\mathbf{0}\}$ ein zugehöriger Eigenvektor. Nach Definition der induzierten Matrixnorm gilt

$$|\lambda| = \frac{|\lambda|\,\|\mathbf{e}\|}{\|\mathbf{e}\|} = \frac{\|\lambda\mathbf{e}\|}{\|\mathbf{e}\|} = \frac{\|\mathbf{X}\mathbf{e}\|}{\|\mathbf{e}\|} \le \sup\left\{\frac{\|\mathbf{X}\mathbf{x}\|}{\|\mathbf{x}\|} \ : \ \mathbf{x} \in \mathbb{C}^{\mathcal{I}} \setminus \{\mathbf{0}\}\right\} = \|\mathbf{X}\|.$$

Da $\lambda \in \sigma(\mathbf{X})$ beliebig gewählt werden kann, ist die gewünschte Aussage bewiesen. ■

Übungsaufgabe 2.18 (Spektralradius eines Produkts) *Seien $\mathcal{I}, \mathcal{J}$ endliche Indexmengen. Für Matrizen $\mathbf{A} \in \mathbb{K}^{\mathcal{I}\times\mathcal{J}}$ und $\mathbf{B} \in \mathbb{K}^{\mathcal{J}\times\mathcal{I}}$ zeige man*

$$\sigma(\mathbf{AB}) \cup \{0\} = \sigma(\mathbf{BA}) \cup \{0\}.$$

Daraus folgt offenbar $\varrho(\mathbf{AB}) = \varrho(\mathbf{BA})$.

Hinweis: *Der Beweis lässt sich besonders elegant führen, indem man zeigt, dass die Blockmatrizen*

$$\begin{pmatrix} \mathbf{AB} & \mathbf{A} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{0} & \mathbf{BA} \end{pmatrix}$$

ähnlich sind.

Wir werden nun nachweisen, dass aus $\varrho(\mathbf{X}) < 1$ bereits $\lim_{m\to\infty} \mathbf{X}^m = 0$ folgt. Die Idee besteht dabei darin, für ein beliebiges $\epsilon \in \mathbb{R}_{>0}$ eine spezielle Norm $\|\cdot\|_{X,\epsilon}$ auf $\mathbb{C}^{\mathcal{I}}$ zu konstruieren, deren induzierte Matrixnorm die Ungleichung

$$\|\mathbf{X}\|_{X,\epsilon} \leq \varrho(\mathbf{X}) + \epsilon$$

erfüllt. Falls $\varrho(\mathbf{X}) < 1$ gilt, können wir ϵ so klein wählen, dass auch $\|\mathbf{X}\|_{X,\epsilon} < 1$ gilt, so dass sich das Konvergenzkriterium aus Lemma 2.15 anwenden lässt.

Zu dieser Norm gelangen wir in drei Schritten: Zunächst verwenden wir eine Ähnlichkeitstransformation, um $\mathbf{X}$ in eine obere Dreiecksmatrix zu überführen. Dann verwenden wir eine weitere Ähnlichkeitstransformation, um die Dreiecksmatrix „bis auf ϵ“ in die Nähe einer Diagonalmatrix zu bringen. Schließlich konstruieren wir aus den beiden Transformationen eine Norm und weisen nach, dass diese Norm die gewünschte Eigenschaft besitzt.

Definition 2.19 (Orthogonale Matrix) *Eine Matrix $\mathbf{Q} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ nennen wir* orthogonal*, falls $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}$ gilt. In diesem Fall ist $\mathbf{Q}$ regulär und es gilt $\mathbf{Q}^* = \mathbf{Q}^{-1}$.*

Lemma 2.20 (Schur-Normalform) *Sei $\mathbf{X} \in \mathbb{C}^{n\times n}$ eine beliebige Matrix. Dann existieren eine orthogonale Matrix $\mathbf{Q} \in \mathbb{C}^{n\times n}$ und eine rechte obere Dreiecksmatrix $\mathbf{R} \in \mathbb{C}^{n\times n}$ mit*

$$\mathbf{X} = \mathbf{Q}\mathbf{R}\mathbf{Q}^*.$$

Wir können also jede komplexwertige quadratische Matrix mit einer orthogonalen Ähnlichkeitstransformation auf Dreiecksgestalt bringen.

Beweis. Nach dem Fundamentalsatz der Algebra besitzt das charakteristische Polynom

$$\zeta \mapsto \det(\zeta\mathbf{I} - \mathbf{X})$$

eine Nullstelle $\lambda \in \mathbb{C}$. Offenbar ist λ dann ein Eigenwert von $\mathbf{X}$, so dass wir einen passenden Eigenvektor $\mathbf{e} \in \mathbb{C}^{\mathcal{I}} \setminus \{\mathbf{0}\}$ finden können. Sei $\delta_1 \in \mathbb{C}^n$ der erste kanonische Einheitsvektor, und sei $\mathbf{H} \in \mathbb{C}^{n\times n}$ eine Householder-Spiegelung, die $\mathbf{H}\delta_1 = \gamma\mathbf{e}$ für ein $\gamma \in \mathbb{C}$ erfüllt. Dann folgt

$$\mathbf{H}^*\mathbf{X}\mathbf{H}\delta_1 = \mathbf{H}^*\mathbf{X}\gamma\mathbf{e} = \mathbf{H}^*\gamma\lambda\mathbf{e} = \lambda\mathbf{H}^*(\gamma\mathbf{e}) = \lambda\delta_1,$$

also insbesondere

$$\mathbf{H}^*\mathbf{X}\mathbf{H} = \begin{pmatrix} \lambda & \mathbf{C} \\ & \widehat{\mathbf{X}} \end{pmatrix}$$

mit $\mathbf{C} \in \mathbb{C}^{1\times(n-1)}$ und $\widehat{\mathbf{X}} \in \mathbb{C}^{(n-1)\times(n-1)}$. Indem wir induktiv mit $\widehat{\mathbf{X}}$ fortfahren, folgt die Behauptung. ∎

Es lässt sich leicht erkennen, dass $\lambda\mathbf{I} - \mathbf{R}$ genau dann nicht regulär ist, wenn λ mit einem der Diagonalelemente der Dreiecksmatrix übereinstimmt, sämtliche Eigenwerte der Matrix $\mathbf{X}$ lassen sich also an der Diagonale der Matrix $\mathbf{R}$ ablesen. Den Außerdiagonalanteil können wir mit Hilfe einer zweiten Ähnlichkeitstransformation reduzieren, und in einer geeigneten Norm erhalten wir damit die nötige Abschätzung.

Definition 2.21 (Maximumnorm) *Die durch*

$$\|\mathbf{x}\|_\infty := \max\{|x_i| \ : \ i \in \mathcal{I}\} \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^{\mathcal{I}}$$

gegebene Norm nennen wir die Maximumnorm *oder* ℓ^∞-Norm.

Lemma 2.22 (Zeilensummennorm) *Die durch die Maximumnorm induzierte Matrixnorm erfüllt*

$$\|\mathbf{X}\|_\infty = \max_{i\in\mathcal{I}} \sum_{j\in\mathcal{I}} |x_{ij}| \qquad \text{für alle } \mathbf{X} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$$

und wird deshalb als Zeilensummennorm *bezeichnet.*

Beweis. Seien $\mathbf{X} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ und $\mathbf{y} \in \mathbb{K}^{\mathcal{I}}$ gegeben. Dann gilt

$$\|\mathbf{X}\mathbf{y}\|_\infty = \max_{i\in\mathcal{I}} \left| \sum_{j\in\mathcal{I}} x_{ij} y_j \right| \le \max_{i\in\mathcal{I}} \sum_{j\in\mathcal{I}} |x_{ij}||y_j| \le \max_{i\in\mathcal{I}} \sum_{j\in\mathcal{I}} |x_{ij}| \|\mathbf{y}\|_\infty,$$

also folgt

$$\|\mathbf{X}\|_\infty \le \max_{i\in\mathcal{I}} \sum_{j\in\mathcal{I}} |x_{ij}|.$$

Sei nun $k \in \mathcal{I}$ so gewählt, dass

$$\sum_{j\in\mathcal{I}} |x_{kj}| = \max_{i\in\mathcal{I}} \sum_{j\in\mathcal{I}} |x_{ij}|$$

gilt. Wir definieren einen Vektor $\mathbf{y} \in \mathbb{K}^{\mathcal{I}}$ durch

$$y_j := \begin{cases} |x_{kj}|/x_{kj} & \text{falls } x_{kj} \neq 0, \\ 1 & \text{ansonsten} \end{cases} \qquad \text{für alle } j \in \mathcal{I}$$

und stellen fest, dass $\|\mathbf{y}\|_\infty = 1$ gilt. Außerdem gilt

$$\|\mathbf{X}\mathbf{y}\|_\infty \ge |(\mathbf{X}\mathbf{y})_k| = \left| \sum_{j\in\mathcal{I}} x_{kj} y_j \right| = \sum_{j\in\mathcal{I}} |x_{kj}| = \max_{i\in\mathcal{I}} \sum_{j\in\mathcal{I}} |x_{ij}| = \max_{i\in\mathcal{I}} \sum_{j\in\mathcal{I}} |x_{ij}| \|\mathbf{y}\|_\infty,$$

also haben wir auch eine untere Schranke für $\|\mathbf{X}\|_\infty$ gefunden. ∎

Die Zeilensummennorm hängt in besonders einfacher Weise von den einzelnen Koeffizienten der Matrix ab, also bietet es sich an, nach einer Ähnlichkeitstransformation zu suchen, die die Koeffizienten außerhalb der Diagonale reduziert.

Lemma 2.23 (Diagonalskalierung) *Sei* $\mathbf{R} \in \mathbb{R}^{n\times n}$ *eine rechte obere Dreiecksmatrix. Wir zerlegen sie in der Form*

$$\mathbf{R} = \mathbf{D} + \mathbf{N}, \qquad \mathbf{D} = \begin{pmatrix} r_{11} & & & \\ & r_{22} & & \\ & & \ddots & \\ & & & r_{nn} \end{pmatrix}, \qquad \mathbf{N} = \begin{pmatrix} 0 & r_{12} & \dots & r_{1n} \\ & \ddots & \ddots & \vdots \\ & & 0 & r_{n-1,n} \\ & & & 0 \end{pmatrix}$$

in ihren Diagonalanteil $\mathbf{D}$ *und einen Rest* $\mathbf{N}$.

Sei $\epsilon \in (0,1]$ *und sei*

$$\mathbf{E} := \begin{pmatrix} \epsilon & & & \\ & \epsilon^2 & & \\ & & \ddots & \\ & & & \epsilon^n \end{pmatrix}.$$

Dann gilt für die Zeilensummennorm

$$\|\mathbf{E}^{-1}\mathbf{R}\mathbf{E}\|_\infty \le \|\mathbf{D}\|_\infty + \epsilon\|\mathbf{N}\|_\infty \le \|\mathbf{D}\|_\infty + \epsilon\|\mathbf{R}\|_\infty.$$

Beweis. Für jede beliebige Matrix $\mathbf{Y} \in \mathbb{R}^{n\times n}$ gilt

$$(\mathbf{E}^{-1}\mathbf{Y}\mathbf{E})_{ij} = \epsilon^{-i} y_{ij} \epsilon^j = \epsilon^{j-i} y_{ij} \qquad \text{für alle } i,j \in [1:n].$$

In unserem Fall folgt daraus

$$\mathbf{E}^{-1}\mathbf{D}\mathbf{E} = \begin{pmatrix} r_{11} & & & \\ & r_{22} & & \\ & & \ddots & \\ & & & r_{nn} \end{pmatrix}, \qquad \mathbf{E}^{-1}\mathbf{N}\mathbf{E} = \begin{pmatrix} 0 & \epsilon r_{12} & \dots & \epsilon^{n-1} r_{1n} \\ & \ddots & \ddots & \vdots \\ & & 0 & \epsilon r_{n-1,n} \\ & & & 0 \end{pmatrix}$$

und dank Lemma 2.22 folgt wegen $|\epsilon| \le 1$ unmittelbar

$$\|\mathbf{E}^{-1}\mathbf{N}\mathbf{E}\|_\infty \le \epsilon\|\mathbf{N}\|_\infty.$$

Mit der Dreiecksungleichung erhalten wir

$$\|\mathbf{E}^{-1}\mathbf{R}\mathbf{E}\|_\infty = \|\mathbf{E}^{-1}(\mathbf{D}+\mathbf{N})\mathbf{E}\|_\infty \le \|\mathbf{E}^{-1}\mathbf{D}\mathbf{E}\|_\infty + \|\mathbf{E}^{-1}\mathbf{N}\mathbf{E}\|_\infty \le \|\mathbf{D}\|_\infty + \epsilon\|\mathbf{N}\|_\infty,$$

so dass sich mit $\|\mathbf{N}\|_\infty \le \|\mathbf{R}\|_\infty$ die gewünschte Aussage ergibt. ∎

Damit sind die nötigen Vorarbeiten abgeschlossen und wir können die gewünschte Normabschätzung beweisen.

Satz 2.24 (Spektralradius) *Sei* $\mathbf{X} \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$ *eine beliebige Matrix, sei* $\epsilon \in \mathbb{R}_{>0}$. *Dann existiert eine Norm* $\|\cdot\|_{X,\epsilon}$ *auf* $\mathbb{C}^{\mathcal{I}}$ *derart, dass die von ihr induzierte Matrixnorm die Abschätzung*

$$\|\mathbf{X}\|_{X,\epsilon} \le \varrho(\mathbf{X}) + \epsilon$$

erfüllt.

Beweis. Für $\mathbf{X} = \mathbf{0}$ ist die Aussage trivial, also können wir im Folgenden von $\mathbf{X} \neq \mathbf{0}$ ausgehen.

Da wir die Elemente der Indexmenge $\mathcal{I}$ beliebig durchnumerieren können, dürfen wir ohne Beschränkung der Allgemeinheit $\mathbf{X} \in \mathbb{C}^{n\times n}$ annehmen.

Mit Lemma 2.20 finden wir eine orthogonale Matrix $\mathbf{Q} \in \mathbb{C}^{n\times n}$ und eine rechte obere Dreiecksmatrix $\mathbf{R} \in \mathbb{C}^{n\times n}$ mit

$$\mathbf{X} = \mathbf{Q}\mathbf{R}\mathbf{Q}^*.$$

Da $\mathbf{X} \neq \mathbf{0}$ vorausgesetzt ist, muss auch $\mathbf{R} \neq \mathbf{0}$ gelten, also $\|\mathbf{R}\|_\infty > 0$. Indem wir Lemma 2.23 auf $\hat{\epsilon} := \min\{\epsilon/\|\mathbf{R}\|_\infty, 1\}$ anwenden, erhalten wir eine Diagonalmatrix $\mathbf{E} \in \mathbb{C}^{n\times n}$, die

$$\|\mathbf{E}^{-1}\mathbf{R}\mathbf{E}\|_\infty \leq \|\mathbf{D}\|_\infty + \hat{\epsilon}\|\mathbf{R}\|_\infty \leq \|\mathbf{D}\|_\infty + \epsilon$$

erfüllt. Mit Lemma 2.22 folgt

$$\|\mathbf{D}\|_\infty = \max\{|d_{ii}| \ : \ i \in \{1, \ldots, n\}\} = \varrho(\mathbf{D}),$$

und da $\mathbf{D}$ dieselben Eigenwerte wie $\mathbf{R}$ besitzt und $\mathbf{R}$ durch eine Ähnlichkeitstransformation aus $\mathbf{X}$ hervorgegangen ist, erhalten wir

$$\|\mathbf{D}\|_\infty = \varrho(\mathbf{D}) = \varrho(\mathbf{R}) = \varrho(\mathbf{X}).$$

Insgesamt haben wir also bereits

$$\|\mathbf{E}^{-1}\mathbf{Q}^*\mathbf{X}\mathbf{Q}\mathbf{E}\|_\infty = \|\mathbf{E}^{-1}\mathbf{R}\mathbf{E}\|_\infty \leq \varrho(\mathbf{X}) + \epsilon$$

bewiesen und müssen nur noch den Term auf der linken Seite als eine induzierte Matrixnorm identifizieren.

Dazu setzen wir

$$\|\mathbf{y}\|_{X,\epsilon} := \|\mathbf{E}^{-1}\mathbf{Q}^*\mathbf{y}\|_\infty \qquad \text{für alle } \mathbf{y} \in \mathbb{C}^n$$

und stellen fest, dass wir dank der Regularität der Matrix $\mathbf{E}^{-1}\mathbf{Q}^*$ eine Norm konstruiert haben, die

$$\begin{aligned}
\|\mathbf{X}\|_{X,\epsilon} &= \sup\left\{\frac{\|\mathbf{X}\mathbf{y}\|_{X,\epsilon}}{\|\mathbf{y}\|_{X,\epsilon}} \ : \ \mathbf{y} \in \mathbb{C}^n \setminus \{\mathbf{0}\}\right\} \\
&= \sup\left\{\frac{\|\mathbf{E}^{-1}\mathbf{Q}^*\mathbf{X}\mathbf{y}\|_\infty}{\|\mathbf{E}^{-1}\mathbf{Q}^*\mathbf{y}\|_\infty} \ : \ \mathbf{y} \in \mathbb{C}^n \setminus \{\mathbf{0}\}\right\} \\
&= \sup\left\{\frac{\|\mathbf{E}^{-1}\mathbf{Q}^*\mathbf{X}\mathbf{Q}\mathbf{E}\mathbf{z}\|_\infty}{\|\mathbf{z}\|_\infty} \ : \ \mathbf{z} \in \mathbb{C}^n \setminus \{\mathbf{0}\}\right\} = \|\mathbf{E}^{-1}\mathbf{Q}^*\mathbf{X}\mathbf{Q}\mathbf{E}\|_\infty
\end{aligned}$$

erfüllt. In der dritten Zeile haben wir dabei die Substitution $\mathbf{y} = \mathbf{Q}\mathbf{E}\mathbf{z}$ durchgeführt, die dank der Regularität der Matrix nichts an dem Supremum ändert. ■

Mit Hilfe dieses Resultats können wir nun die Konvergenz linearer Iterationsverfahren vollständig charakterisieren und die folgende wesentliche Verallgemeinerung des Lemmas 2.15 herleiten:

Satz 2.25 (Konvergenz) *Sei Φ ein lineares Iterationsverfahren, dessen erste Normalform durch die Matrizen $\mathbf{M}, \mathbf{N} \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$ gegeben ist. Φ ist genau dann konvergent, wenn $\varrho(\mathbf{M}) < 1$ gilt.*

In diesem Fall ist der Grenzwert für eine rechte Seite $\mathbf{b} \in \mathbb{C}^{\mathcal{I}}$ durch die Formel $\mathbf{x}^* := (\mathbf{I} - \mathbf{M})^{-1}\mathbf{N}\mathbf{b}$ gegeben.

Beweis. Sei zunächst $\varrho(\mathbf{M}) \geq 1$. Wir wählen einen Eigenwert $\lambda \in \mathbb{C}$ von $\mathbf{M}$ mit $|\lambda| \geq 1$ und einen zugehörigen Eigenvektor $\mathbf{e} \in \mathbb{C}^{\mathcal{I}} \setminus \{\mathbf{0}\}$. Für die rechte Seite $\mathbf{b} = \mathbf{0}$ und den Startvektor $\mathbf{0}$ ist die Folge der Iterierten konstant $\mathbf{0}$.

Für den Startvektor $\mathbf{x}^{(0)} := \mathbf{e}$ dagegen gilt

$$\mathbf{x}^{(1)} = \Phi(\mathbf{x}^{(0)}, \mathbf{b}) = \mathbf{M}\mathbf{x}^{(0)} = \mathbf{M}\mathbf{e} = \lambda\mathbf{e},$$

und mit Lemma 2.10 erhalten wir

$$\mathbf{x}^{(m)} = \lambda^m \mathbf{e} \qquad \text{für alle } m \in \mathbb{N}_0,$$

also konvergiert diese Folge nicht gegen $\mathbf{0}$. Also ist der Grenzwert nicht vom Startvektor unabhängig, und das Verfahren Φ gemäß Definition 2.3 nicht konvergent.

Sei nun $\varrho(\mathbf{M}) < 1$. Dann können wir ein $\epsilon \in \mathbb{R}_{>0}$ so wählen, dass $\varrho(\mathbf{M}) + \epsilon < 1$ gilt, und erhalten mit der in Satz 2.24 definierten Norm

$$\|\mathbf{M}\|_{M,\epsilon} < 1.$$

Mit Lemma 2.15 folgt, dass Φ konvergent ist und dass der Grenzwert die gewünschte Gestalt aufweist. ∎

Wir haben bereits in Lemma 2.17 gesehen, dass $\varrho(\mathbf{M}) \leq \|\mathbf{M}\|$ für jede beliebige induzierte Matrixnorm gilt, also ist Satz 2.25 tatsächlich eine Verallgemeinerung des einfacheren Kriteriums, das in Lemma 2.15 vorgestellt wurde.

Bemerkung 2.26 (Dritte Normalform) *Falls Φ eine konsistente und konvergente lineare Iteration ist, ist wegen $\varrho(\mathbf{M}) < 1$ insbesondere 1 kein Eigenwert der Matrix $\mathbf{M}$. Also muss $\mathbf{I} - \mathbf{M} = \mathbf{N}\mathbf{A}$ regulär sein. Da $\mathbf{N}$ und $\mathbf{A}$ quadratische Matrizen sind, folgt daraus, dass auch $\mathbf{N}$ regulär sein muss.*

Deshalb lässt sich eine konsistente und konvergente lineare Iteration mit der Matrix $\mathbf{W} := \mathbf{N}^{-1}$ auch in der dritten Normalform

$$\Phi(\mathbf{x}, \mathbf{b}) = \mathbf{x} - \mathbf{W}^{-1}(\mathbf{A}\mathbf{x} - \mathbf{b}) \qquad \textit{für alle } \mathbf{x}, \mathbf{b} \in \mathbb{C}^{\mathcal{I}}$$

darstellen, und die Folge der Iterierten ist durch die Gleichungssysteme

$$\mathbf{W}\left(\mathbf{x}^{(m-1)} - \mathbf{x}^{(m)}\right) = \mathbf{A}\mathbf{x}^{(m-1)} - \mathbf{b} = \mathbf{A}(\mathbf{x}^{(m-1)} - \mathbf{x}^*) \qquad \textit{für alle } m \in \mathbb{N}$$

gegeben. Ein Iterationsverfahren ist also dann besonders gut, wenn $\mathbf{W}$ eine gute Approximation von $\mathbf{A}$ ist und sich diese Gleichungssysteme effizient lösen lassen.

Bemerkung 2.27 (Grenzwert) *Falls Φ konsistent, konvergent und linear ist, erhalten wir für den Grenzwert der Folge der Iterierten die Gleichung*

$$\mathbf{x}^* = (\mathbf{I} - \mathbf{M})^{-1}\mathbf{N}\mathbf{b} = (\mathbf{I} - \mathbf{I} + \mathbf{N}\mathbf{A})^{-1}\mathbf{N}\mathbf{b} = \mathbf{A}^{-1}\mathbf{N}^{-1}\mathbf{N}\mathbf{b} = \mathbf{A}^{-1}\mathbf{b}, \tag{2.9}$$

im Falle einer linearen Iteration können wir also sogar explizit nachrechnen, dass die Folge der Iterierten gegen die Lösung des Gleichungssystems konvergiert.

Bemerkung 2.28 *Falls Φ lediglich konvergent ist, aber nicht die stärkere Bedingung (2.8) erfüllt, gilt gemäß Satz 2.25 immerhin noch $\varrho(\mathbf{M}) < 1$, also können wir nach Satz 2.24 für jedes $\zeta \in (\varrho(\mathbf{M}), 1)$ eine Norm $\|\cdot\|_{M,\epsilon}$ mit*

$$\|\mathbf{M}^m\|_{M,\epsilon} \leq \zeta^m \qquad \text{für alle } m \in \mathbb{N}_0$$

finden. Da in dem endlich-dimensionalen Raum $\mathbb{C}^n$ alle Normen äquivalent sind, folgt auch für jede andere induzierte Matrixnorm

$$\|\mathbf{M}^m\| \leq C\zeta^m \qquad \text{für alle } m \in \mathbb{N}_0$$

mit einer geeigneten Konstanten $C \in \mathbb{R}_{>0}$.

In diesem Sinn ist $\varrho(\mathbf{M})$ eine obere Schranke für die Konvergenzgeschwindigkeit, deshalb wird diese Größe auch als die Konvergenzrate *des Verfahrens Φ bezeichnet.*

2.4 Richardson-Iteration

Wenden wir uns nun der Untersuchung der bereits in Kapitel 1 erwähnten Richardson-Iteration zu, die das einfachste lineare Iterationsverfahren ist.

Definition 2.29 (Richardson-Iteration) *Sei $\theta \in \mathbb{K}$. Das durch*

$$\Phi_{\mathrm{Rich},\theta}(\mathbf{x}, \mathbf{b}) = \mathbf{x} - \theta(\mathbf{A}\mathbf{x} - \mathbf{b}) \qquad \text{für alle } \mathbf{x}, \mathbf{b} \in \mathbb{K}^{\mathcal{I}}$$

gegebene lineare Iterationsverfahren nennen wir die Richardson-Iteration. *Die Konstante θ bezeichnen wir als* Dämpfungsparameter.

Aus der Definition können wir bereits ablesen, dass die Richardson-Iteration konsistent ist, und dass die Matrizen ihrer ersten Normalform durch

$$\mathbf{M}_{\mathrm{Rich},\theta} := \mathbf{I} - \theta\mathbf{A}, \qquad \mathbf{N}_{\mathrm{Rich},\theta} := \theta\mathbf{I}$$

gegeben sind. Da die Konsistenz geklärt ist, wenden wir uns der Untersuchung der Konvergenz des Verfahrens zu.

Gemäß Satz 2.25 können wir an den Eigenwerten der Iterationsmatrix $\mathbf{M}_{\mathrm{Rich},\theta}$ erkennen, ob $\Phi_{\mathrm{Rich},\theta}$ konvergiert.

Lemma 2.30 *Sei $\theta \in \mathbb{C}$. Es gilt*

$$\sigma(\mathbf{M}_{\mathrm{Rich},\theta}) = \{1 - \theta\lambda \ : \ \lambda \in \sigma(\mathbf{A})\}.$$

Beweis. Für $\theta = 0$ ist die Aussage trivial. Sei also nun $\theta \neq 0$.

Wir wählen zunächst $\lambda \in \sigma(\mathbf{A})$. Sei $\mathbf{e} \in \mathbb{C}^{\mathcal{I}} \setminus \{\mathbf{0}\}$ ein Eigenvektor mit $\mathbf{A}\mathbf{e} = \lambda\mathbf{e}$. Die Definition der Iterationsmatrix impliziert

$$\mathbf{M}_{\mathrm{Rich},\theta}\mathbf{e} = (\mathbf{I} - \theta\mathbf{A})\mathbf{e} = \mathbf{e} - \theta\mathbf{A}\mathbf{e} = \mathbf{e} - \theta\lambda\mathbf{e} = (1 - \theta\lambda)\mathbf{e},$$

also ist $\mu := 1 - \theta\lambda$ ein Eigenwert von $\mathbf{M}_{\mathrm{Rich},\theta}$ zum Eigenvektor $\mathbf{e}$.

Sei nun $\mu \in \sigma(\mathbf{M}_{\mathrm{Rich},\theta})$. Wir können einen Eigenvektor $\mathbf{e} \in \mathbb{C}^{\mathcal{I}} \setminus \{\mathbf{0}\}$ mit

$$\mu\mathbf{e} = \mathbf{M}_{\mathrm{Rich},\theta}\mathbf{e} = (\mathbf{I} - \theta\mathbf{A})\mathbf{e} = \mathbf{e} - \theta\mathbf{A}\mathbf{e}$$

finden, also gilt auch

$$\frac{1-\mu}{\theta}\mathbf{e} = \mathbf{A}\mathbf{e}.$$

Demnach ist $\lambda := (1-\mu)/\theta$ ein Eigenwert zum Eigenvektor $\mathbf{e}$ von $\mathbf{A}$, für den $\theta\lambda = 1 - \mu$ und somit $\mu = 1 - \theta\lambda$ gilt. ∎

Unsere Aufgabe besteht also darin, einen Dämpfungsparameter $\theta \in \mathbb{C}$ so zu finden, dass der Spektralradius von $\mathbf{M}_{\mathrm{Rich},\theta}$ kleiner als Eins ist.

Die entsprechende Gleichung

$$|1 - \theta\lambda| < 1 \qquad \text{für alle } \lambda \in \sigma(\mathbf{A})$$

lässt sich in der komplexen Ebene $\mathbb{C}$ geometrisch interpretieren: Die Zahlen $\theta\lambda$ müssen in einem Kreis mit Radius kleiner als Eins um 1 liegen. Da wir für θ komplexe Zahlen zugelassen haben, beschreibt die Multiplikation mit dem Dämpfungsparameter eine Rotation und Skalierung des Spektrums, und wir können das folgende einfache Konvergenzkriterium gewinnen:

Lemma 2.31 (Konvergenzkriterium) *Es gibt genau dann ein $\theta \in \mathbb{C}$, für das die Richardson-Iteration konvergiert, wenn es ein $z_0 \in \mathbb{C}$ und ein $r \in [0, |z_0|)$ so gibt, dass $\sigma(\mathbf{A}) \subseteq \bar{K}(z_0, r)$ gilt, dass also*

$$|\lambda - z_0| \leq r \qquad \textit{für alle } \lambda \in \sigma(\mathbf{A}) \textit{ gilt.}$$

In diesem Fall gilt mit $\theta = 1/z_0$ die Abschätzung $\varrho(\mathbf{M}_{\mathrm{Rich},\theta}) \leq r/|z_0| < 1$.

Beweis. Seien $z_0 \in \mathbb{C}$ und $r \in [0, |z_0|)$ mit der Eigenschaft $\sigma(\mathbf{A}) \subseteq \bar{K}(z_0, r)$ gegeben. Wir setzen $\theta := 1/z_0$ und müssen die Schranke für den Spektralradius von $\mathbf{M}_{\mathrm{Rich},\theta}$ nachweisen.

Sei also $\mu \in \sigma(\mathbf{M}_{\mathrm{Rich},\theta})$. Nach Lemma 2.30 muss es ein $\lambda \in \sigma(\mathbf{A})$ so geben, dass $\mu = 1 - \theta\lambda$ gilt, und wir erhalten

$$|\mu| = |1 - \theta\lambda| = \left|1 - \frac{\lambda}{z_0}\right| = \left|\frac{z_0 - \lambda}{z_0}\right| = \frac{|z_0 - \lambda|}{|z_0|} \leq \frac{r}{|z_0|}$$

wegen $\lambda \in \bar{K}(z_0, r)$. Da μ beliebig gewählt werden kann, folgt aus dieser Abschätzung bereits $\varrho(\mathbf{M}_{\mathrm{Rich},\theta}) \leq r/|z_0|$.

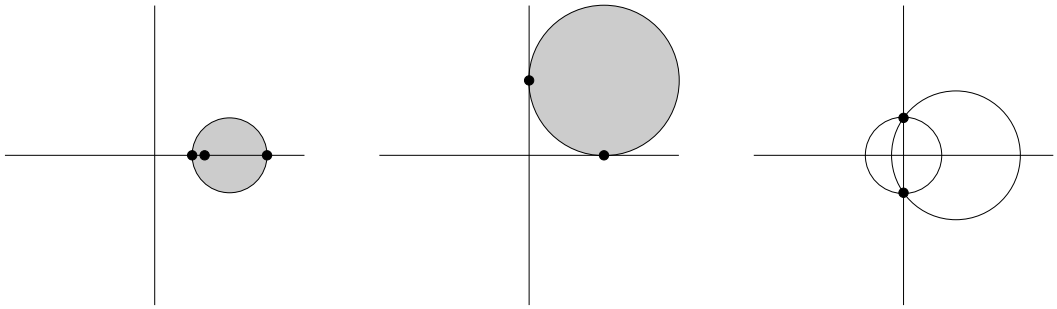

Abbildung 2.1: Spektren der Beispielmatrizen $\mathbf{A}_1$, $\mathbf{A}_2$ und $\mathbf{A}_3$

Sei nun umgekehrt $\theta \in \mathbb{C}$ so gewählt, dass die Richardson-Iteration konvergiert. Nach Satz 2.25 muss dann $\varrho(\mathbf{M}_{\mathrm{Rich},\theta}) < 1$ gelten, also wegen Lemma 2.30 gerade

$$\varrho := \max\{|1 - \theta\lambda| \ : \ \lambda \in \sigma(\mathbf{A})\} < 1.$$

Die Wahl $\theta = 0$ kann nicht zu einem konvergenten Verfahren (vgl. Definition 2.3) führen, also sind $z_0 := 1/\theta \in \mathbb{C}$ und $r := \varrho|z_0| \in [0, |z_0|)$ wohldefiniert. Für $\lambda \in \sigma(\mathbf{A})$ stellen wir fest, dass

$$|z_0 - \lambda| = \left| z_0 - z_0 \frac{\lambda}{z_0} \right| = |z_0 - z_0\theta\lambda| = |z_0|\,|1 - \theta\lambda| \leq |z_0|\varrho = r$$

gilt, also $\lambda \in \bar{K}(z_0, r)$ und somit $\sigma(\mathbf{A}) \subseteq \bar{K}(z_0, r)$. ■

Die Bedingung $r \in [0, |z_0|)$ ist äquivalent dazu, dass der Kreis $\bar{K}(z_0, r)$ die Null nicht enthält, die Richardson-Iteration konvergiert also, falls das Spektrum von $\mathbf{A}$ in einem abgeschlossenen Kreis enthalten ist, der die Null nicht enthält.

Betrachten wir einige Beispiele (vgl. Abbildung 2.1). Für die Matrix

$$\mathbf{A}_1 = \begin{pmatrix} 1/2 & 5 & 6 \\ 0 & 2/3 & 8 \\ 0 & 0 & 3/2 \end{pmatrix}$$

ist die Situation einfach: Ihre Eigenwerte sind durch $\sigma(\mathbf{A}_1) = \{1/2, 2/3, 3/2\}$ gegeben und liegen in einem Kreis mit Radius $1/2$ um $z_0 = 1$, also gilt schon für die einfache Wahl $\theta = 1$ bereits

$$\sigma(\mathbf{M}_{\mathrm{Rich},1}) = \{-1/2, 1/3, 1/2\}$$

und wir erhalten $\varrho(\mathbf{M}_{\mathrm{Rich},1}) = 1/2 < 1$, die Richardson-Iteration wird also konvergieren.

Als nächstes untersuchen wir die Matrix

$$\mathbf{A}_2 = \begin{pmatrix} -5 & 3 \\ -10 - 2i & 6 + i \end{pmatrix}.$$

Ihr charakteristisches Polynom ist $p_2(\lambda) = \det(\lambda\mathbf{I} - \mathbf{A}_2) = (\lambda + 5)(\lambda - 6 - i) + (30 + 6i)$, und seine Nullstellen, also die Eigenwerte von $\mathbf{A}_2$, sind $\sigma(\mathbf{A}_2) = \{1, i\}$. Für den Kreis

um $z_0 = 1 + i$ mit Radius $r = 1$ können wir leicht nachprüfen, dass $\sigma(\mathbf{A}_2) \subseteq \bar{K}(z_0, r)$ gilt, also folgt für $\theta = 1/z_0 = (1-i)/2$ schon $\varrho(\mathbf{M}_{\mathrm{Rich},\theta}) = \sqrt{1/2} < 1$ und wir haben die Konvergenz bewiesen.

Für die Matrix

$$\mathbf{A}_3 = \begin{pmatrix} 0 & 1/2 \\ -1/2 & 0 \end{pmatrix}$$

hingegen ist die Situation hoffnungslos: Ihr charakteristisches Polynom ist durch $p_3(\lambda) = \lambda^2 + 1/4$ gegeben, besitzt also nur die Nullstellen $i/2$ und $-i/2$, und jeder Kreis, der diese beide Eigenwerte einschließt, muss auch die Null enthalten. Also kann es keinen Dämpfungsparameter $\theta \in \mathbb{C}$ geben, mit dem eine Richardson-Iteration für $\mathbf{A}_3$ konvergiert.

Bemerkung 2.32 (Reelles Spektrum) *Falls alle Eigenwerte reell und positiv sind, falls also $\sigma(\mathbf{A}) \subseteq [\alpha, \beta]$ mit $0 < \alpha \leq \beta$ gilt, können wir Lemma 2.31 erheblich vereinfachen: Wir wählen $z_0 = (\alpha + \beta)/2 > 0$ sowie $r = z_0 - \alpha$ und stellen fest, dass $\sigma(\mathbf{A}) \subseteq \bar{K}(z_0, r)$ gilt. Mit $\theta = 1/z_0$ folgt*

$$\varrho(\mathbf{M}_{Rich,\theta}) \leq r/|z_0| = \frac{2(z_0 - \alpha)}{\beta + \alpha} = \frac{\alpha + \beta - 2\alpha}{\beta + \alpha} = \frac{\beta - \alpha}{\beta + \alpha} < 1.$$

Da die Konvergenz der Richardson-Iteration im Wesentlichen durch die Verteilung der Eigenwerte der Matrix $\mathbf{A}$ festgelegt wird, führen geringe Störungen dieser Matrix auch nur zu geringen Störungen im Konvergenzverhalten.

Bemerkung 2.33 (Vorkonditionierer) *Die relativ ausführliche Analyse des in der Praxis eher selten eingesetzten Richardson-Verfahrens ist dadurch gerechtfertigt, dass es als Prototyp für alle anderen konsistenten linearen Iterationsverfahren gesehen werden kann: Falls Φ ein konsistentes und lineares Iterationsverfahren ist, lässt es sich nach Lemma 2.9 in der Form*

$$\Phi(\mathbf{x}, \mathbf{b}) = \mathbf{x} - \mathbf{N}(\mathbf{A}\mathbf{x} - \mathbf{b}) \qquad \textit{für alle } \mathbf{x}, \mathbf{b} \in \mathbb{K}^{\mathcal{I}}$$

darstellen. Ein Richardson-Verfahren für die mit $\mathbf{N}$ vorkonditionierte *Gleichung*

$$\mathbf{N}\mathbf{A}\mathbf{x} = \mathbf{N}\mathbf{b} \tag{2.10}$$

hat gerade die Form

$$\Phi_{\mathrm{Rich},\theta}(\mathbf{x}, \mathbf{b}) = \mathbf{x} - \theta\mathbf{N}(\mathbf{A}\mathbf{x} - \mathbf{b}) \qquad \textit{für alle } \mathbf{x}, \mathbf{b} \in \mathbb{K}^{\mathcal{I}},$$

entspricht also für $\theta = 1$ genau dem allgemeinen Verfahren. Falls $\mathbf{N} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ regulär ist, besitzt das lineare Gleichungssystem (2.10) dieselbe Lösung wie das ursprüngliche System (1.1), also entspricht die Anwendung des allgemeinen Verfahrens auf das ursprüngliche System der Anwendung des Richardson-Verfahrens auf das vorkonditionierte System.

Damit lässt sich die für das Richardson-Verfahren entwickelte Konvergenztheorie auch für das allgemeine Verfahren einsetzen, beispielsweise wird das allgemeine Verfahren genau dann konvergent sein, wenn

$$\sigma(\mathbf{N}\mathbf{A}) \subseteq K(1, 1)$$

gilt. Entsprechendes erhält man für den Fall des gedämpften Verfahrens. Ein konsistentes lineares Iterationsverfahren ist besonders gut, wenn die entsprechende Matrix $\mathbf{N}$ *die Eigenschaft besitzt, dass das Spektrum von* $\mathbf{NA}$ *in einem möglichst kleinen Kreis mit möglichst großem Abstand zum Nullpunkt liegt.*

2.5 Jacobi-Iteration

Wie wir in Bemerkung 2.33 gesehen haben, ist es wünschenswert, ein lineares Iterationsverfahren so zu konstruieren, dass für die Matrix $\mathbf{N} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ der zweiten Normalform das Spektrum von $\mathbf{NA}$ in einem möglichst kleinen und vom Nullpunkt möglichst weit entfernten Kreis liegt. Die beste Wahl wäre natürlich $\mathbf{N} = \mathbf{A}^{-1}$, aber wenn wir die Inverse von $\mathbf{A}$ einfach berechnen könnten, bräuchten wir keine Iterationsverfahren. Gesucht ist also eine Matrix $\mathbf{N}$, die effizient berechnet werden kann und trotzdem das Spektrum positiv beeinflusst.

Aufgrund dieser Beobachtung ist es uns nun möglich, die Probleme erneut zu untersuchen, die mit dem Richardson-Verfahren nicht erfolgreich behandelt werden konnten.

Wir haben bereits gesehen, dass das Richardson-Verfahren nicht konvergiert, wenn die Eigenwerte von $\mathbf{A}$ auf verschiedenen Seiten des Nullpunkts liegen, etwa bei der Matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 3 \\ 0 & -1 \end{pmatrix}.$$

Die Lösung ist hier sehr einfach: Wir verwenden die Inverse der *Diagonalen* von $\mathbf{A}$ als Approximation der vollständigen Inversen, setzen also

$$\mathbf{N} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

und erhalten

$$\mathbf{NA} = \begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix},$$

also $\sigma(\mathbf{NA}) = \{1\}$ und damit sogar schon $\varrho(\mathbf{M}_{\mathrm{Rich},1}) = 0$. Durch eine einfache Diagonalskalierung wird also aus einem schwierigen Problem ein besonders einfaches.

Solange die Diagonale der Matrix $\mathbf{A}$ invertierbar ist, solange also kein Diagonalelement gleich null ist, lässt sich das auf diese Weise definierte Iterationsverfahren durchführen.

Definition 2.34 (Jacobi-Iteration) *Sei* $\mathbf{D} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ *die Diagonale von* $\mathbf{A}$*, definiert durch*

$$D_{ij} = \begin{cases} A_{ii} & \text{falls } i = j, \\ 0 & \text{sonst} \end{cases} \qquad \text{für alle } i, j \in \mathcal{I}.$$

Sei $\mathbf{D}$ *invertierbar, also jedes Diagonalelement von* $\mathbf{A}$ *von Null verschieden. Das durch*

$$\Phi_{\mathrm{Jac}}(\mathbf{x}, \mathbf{b}) = \mathbf{x} - \mathbf{D}^{-1}(\mathbf{Ax} - \mathbf{b}) \qquad \text{für alle } \mathbf{x}, \mathbf{b} \in \mathbb{K}^{\mathcal{I}}$$

gegebene lineare Iterationsverfahren nennen wir die Jacobi-Iteration.

Wie schon im Fall der Richardson-Iteration ist auch die Jacobi-Iteration offensichtlich konsistent. Die Matrizen ihrer ersten Normalform sind durch

$$\mathbf{M}_{\mathrm{Jac}} := \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}, \qquad \mathbf{N}_{\mathrm{Jac}} := \mathbf{D}^{-1}$$

gegeben. Im Gegensatz zur Richardson-Iteration gibt es Matrizen für die die Jacobi-Iteration nicht durchführbar ist: Falls eine Null auf der Diagonalen auftritt, ist $\mathbf{D}$ nicht invertierbar und die Jacobi-Iteration nicht definiert.

Die Analyse der Konvergenz des Jacobi-Verfahrens läßt sich wie in Bemerkung 2.33 auf die Analyse des vorkonditionierten Richardson-Verfahrens zurückführen, wenn wir einen zusätzlichen Dämpfungsparameter zulassen:

Definition 2.35 (Gedämpfte Jacobi-Iteration) *Sei $\theta \in \mathbb{K}$. Unter den Voraussetzungen von Definition 2.34 bezeichnen wir das durch*

$$\Phi_{\mathrm{Jac},\theta}(\mathbf{x}, \mathbf{b}) = \mathbf{x} - \theta\mathbf{D}^{-1}(\mathbf{A}\mathbf{x} - \mathbf{b}) \qquad \textit{für alle } \mathbf{x}, \mathbf{b} \in \mathbb{K}^{\mathcal{I}}$$

definierte lineare Iterationsverfahren als gedämpfte Jacobi-Iteration *mit Dämpfungsparameter θ.*

Auf die gedämpfte Jacobi-Iteration lässt sich unmittelbar Lemma 2.31 anwenden:

Lemma 2.36 (Konvergenzkriterium) *Es gibt genau dann ein $\theta \in \mathbb{K}$, für das die gedämpfte Jacobi-Iteration konvergiert, wenn es ein $z_0 \in \mathbb{K}$ und ein $r \in [0, |z_0|)$ so gibt, dass $\sigma(\mathbf{D}^{-1}\mathbf{A}) \subseteq \bar{K}(z_0, r)$ gilt. In diesem Fall erhalten wir $\varrho(\mathbf{M}_{\mathrm{Jac},\theta}) \leq r/|z_0| < 1$.*

Beweis. Wir wenden Lemma 2.31 auf das vorkonditionierte System $\mathbf{D}^{-1}\mathbf{A}\mathbf{x} = \mathbf{D}^{-1}\mathbf{b}$ an und stellen fest, dass die resultierende Richardson-Iteration dieselbe Iterationsmatrix wie die gedämpfte Jacobi-Iteration für (1.1) besitzt. ∎

Im Allgemeinen ist die Analyse des Spektrums von $\mathbf{D}^{-1}\mathbf{A}$ schwierig, deshalb werden wir uns bis auf Weiteres auf Matrizen konzentrieren, die neben der Regularität zusätzliche Annahmen erfüllen.

Wir haben bereits gesehen, dass für die Konvergenz eines Iterationsverfahrens die Eigenwerte der Iterationsmatrix ausschlaggebend sind. Im Fall des eindimensionalen Modellproblems konnten wir sogar die Konvergenz des Richardson-Verfahrens analysieren, ohne auf die in diesem Kapitel vorgestellten Konvergenzresultate zurückzugreifen, indem wir eine aus Eigenvektoren bestehende Basis des Raums $\mathbb{K}^{\mathcal{I}}$ verwendeten. Wir werden diese Technik in folgenden Beweis häufiger anwenden, deshalb ist es sinnvoll, an dieser Stelle die nötigen Aussagen zusammenzustellen.

Definition 2.37 (Spektralnorm) *Zu dem durch*

$$\langle \mathbf{x}, \mathbf{y} \rangle_2 = \sum_{i \in \mathcal{I}} \bar{x}_i y_i \qquad \textit{für alle } \mathbf{x}, \mathbf{y} \in \mathbb{K}^{\mathcal{I}}$$

gegebenen euklidischen Skalarprodukt *gehört die durch*

$$\|\mathbf{x}\|_2 = \langle \mathbf{x}, \mathbf{x} \rangle_2^{1/2} = \left(\sum_{i \in \mathcal{I}} |x_i|^2 \right)^{1/2} \qquad \textit{für alle } \mathbf{x} \in \mathbb{K}^{\mathcal{I}}$$

definierte euklidische Norm. *Die von dieser Norm induzierte Matrixnorm auf* $\mathbb{K}^{\mathcal{I} \times \mathcal{I}}$ *heißt* Spektralnorm, *sie wird ebenfalls mit* $\|\cdot\|_2$ *bezeichnet.*

Definition 2.38 (Adjungierte) *Sei* $\mathbf{X} \in \mathbb{K}^{\mathcal{I} \times \mathcal{J}}$. *Die durch*

$$Y_{ij} = \bar{X}_{ji} \qquad \textit{für alle } i \in \mathcal{J}, j \in \mathcal{I}$$

definierte Matrix $\mathbf{Y} \in \mathbb{K}^{\mathcal{J} \times \mathcal{I}}$ *heißt die* Adjungierte *von* $\mathbf{X}$ *und wird mit* $\mathbf{X}^*$ *bezeichnet.*

Lemma 2.39 *Sei* $\mathbf{X} \in \mathbb{K}^{\mathcal{I} \times \mathcal{J}}$. *Es gilt*

$$\langle \mathbf{X}\mathbf{x}, \mathbf{y} \rangle_2 = \langle \mathbf{x}, \mathbf{X}^* \mathbf{y} \rangle_2 \qquad \textit{für alle } \mathbf{x} \in \mathbb{K}^{\mathcal{J}}, \mathbf{y} \in \mathbb{K}^{\mathcal{I}}.$$

Beweis. Sei $\mathbf{x} \in \mathbb{K}^{\mathcal{J}}$ und $\mathbf{y} \in \mathbb{K}^{\mathcal{I}}$. Dann gilt

$$\langle \mathbf{X}\mathbf{x}, \mathbf{y} \rangle_2 = \sum_{i \in \mathcal{I}} \overline{\left(\sum_{j \in \mathcal{J}} X_{ij} x_j \right)} y_i = \sum_{j \in \mathcal{J}} \bar{x}_j \sum_{i \in \mathcal{I}} \bar{X}_{ij} y_i = \langle \mathbf{x}, \mathbf{X}^* \mathbf{y} \rangle_2.$$

Das ist die zu beweisende Gleichung. ∎

Definition 2.40 (Selbstadjungiert) *Sei* $\mathbf{X} \in \mathbb{K}^{\mathcal{I} \times \mathcal{I}}$ *eine Matrix. Sie heißt* selbstadjungiert, *wenn* $\mathbf{X} = \mathbf{X}^*$ *gilt.*

Lemma 2.41 *Sei* $\mathbf{X} \in \mathbb{K}^{\mathcal{I} \times \mathcal{I}}$ *selbstadjungiert. Dann gilt*

$$\langle \mathbf{x}, \mathbf{X}\mathbf{x} \rangle_2 = \overline{\langle \mathbf{x}, \mathbf{X}\mathbf{x} \rangle}_2 \qquad \textit{für alle } \mathbf{x} \in \mathbb{K}^{\mathcal{I}},$$

also auch $\sigma(\mathbf{X}) \subseteq \mathbb{R}$.

Beweis. Sei $\mathbf{x} \in \mathbb{K}^{\mathcal{I}}$. Nach Lemma 2.39 gilt

$$\langle \mathbf{x}, \mathbf{X}\mathbf{x} \rangle_2 = \langle \mathbf{X}^* \mathbf{x}, \mathbf{x} \rangle_2 = \langle \mathbf{X}\mathbf{x}, \mathbf{x} \rangle_2 = \overline{\langle \mathbf{x}, \mathbf{X}\mathbf{x} \rangle}_2.$$

Sei nun $\lambda \in \sigma(\mathbf{X})$ und $\mathbf{e} \in \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\}$ ein entsprechender Eigenvektor. Dann haben wir

$$\lambda \|\mathbf{e}\|_2^2 = \lambda \langle \mathbf{e}, \mathbf{e} \rangle_2 = \langle \mathbf{e}, \lambda \mathbf{e} \rangle_2 = \langle \mathbf{e}, \mathbf{X}\mathbf{e} \rangle_2 = \overline{\langle \mathbf{e}, \mathbf{X}\mathbf{e} \rangle}_2 = \overline{\langle \mathbf{e}, \lambda \mathbf{e} \rangle}_2 = \overline{\lambda \langle \mathbf{e}, \mathbf{e} \rangle}_2 = \bar{\lambda} \|\mathbf{e}\|_2^2,$$

also $\lambda = \bar{\lambda}$. Daraus folgt $\lambda \in \mathbb{R}$. ∎

Lemma 2.42 *Sei* $\mathbf{Q} \in \mathbb{K}^{\mathcal{I} \times \mathcal{I}}$ *orthogonal. Dann gilt*

$$\|\mathbf{Q}\mathbf{x}\|_2 = \|\mathbf{x}\|_2 \qquad \textit{für alle } \mathbf{x} \in \mathbb{K}^{\mathcal{I}}.$$

Beweis. Sei $\mathbf{x} \in \mathbb{K}^{\mathcal{I}}$. Nach Definition des euklidischen Skalarprodukts gilt

$$\|\mathbf{Qx}\|_2^2 = \langle \mathbf{Qx}, \mathbf{Qx} \rangle_2 = \langle \mathbf{x}, \mathbf{Q}^*\mathbf{Qx} \rangle_2 = \langle \mathbf{x}, \mathbf{x} \rangle_2 = \|\mathbf{x}\|_2^2,$$

also auch die gewünschte Gleichung. ■

Lemma 2.43 (Reelle Schur-Normalform) *Sei* $\mathbf{X} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ *eine selbstadjungierte Matrix. Dann existieren eine orthogonale Matrix* $\mathbf{Q} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ *und eine reelle Diagonalmatrix* $\mathbf{D} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ *so, dass*

$$\mathbf{X} = \mathbf{QDQ}^*$$

gilt. Offenbar sind die Diagonalelemente von $\mathbf{D}$ *gerade die Eigenwerte von* $\mathbf{X}$ *und die Spalten von* $\mathbf{Q}$ *entsprechende Eigenvektoren, also ist* $\mathbf{X}$ *reell diagonalisierbar.*

Beweis. Per Induktion über $n := \#\mathcal{I}$. Für $n = 1$ ist die Aussage trivial.

Sei nun $n \in \mathbb{N}$ so gegeben, dass die gesuchte Basis für alle Indexmengen $\mathcal{I}$ mit $\#\mathcal{I} = n$ existiert. Sei $\mathcal{I}$ eine Indexmenge mit $\#\mathcal{I} = n + 1$. Wir betrachten die Funktion

$$f : \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\} \to \mathbb{K}, \qquad \mathbf{x} \mapsto \frac{\langle \mathbf{x}, \mathbf{Xx} \rangle_2}{\|\mathbf{x}\|_2^2}$$

(den sogenannten *Rayleigh-Quotienten*) auf dem Gebiet

$$\mathcal{S} := \{x \in \mathbb{K}^{\mathcal{I}} \ : \ 1/2 \le \|x\|_2 \le 2\}.$$

Da $\mathbf{X}$ selbstadjungiert ist, folgt aus Lemma 2.41 $f(\mathcal{S}) \subseteq \mathbb{R}$. Weil $\mathbb{K}^{\mathcal{I}}$ ein endlichdimensionaler Raum ist, ist $\mathcal{S}$ kompakt, also auch $f(\mathcal{S})$. Demzufolge muss es ein $\mathbf{e} \in \mathcal{S}$ geben, das

$$f(\mathbf{e}) \ge f(\mathbf{x}) \qquad \text{für alle } \mathbf{x} \in \mathcal{S}$$

erfüllt. Nach Definition von f gilt $f(\mathbf{e}) = f(\mathbf{e}/\|\mathbf{e}\|_2)$, also können wir ohne Beschränkung der Allgemeinheit $\|\mathbf{e}\|_2 = 1$ annehmen. Da $\mathbf{e}$ ein innerer Punkt von $\mathcal{S}$ ist, in dem f sein Maximum annimmt, muss die erste Ableitung von f in diesem Punkt verschwinden:

$$\begin{aligned}
0 = Df(\mathbf{e}) \cdot \mathbf{y} &= \frac{(\langle \mathbf{e}, \mathbf{Xy} \rangle_2 + \langle \mathbf{y}, \mathbf{Xe} \rangle_2)\|\mathbf{e}\|_2^2 - \langle \mathbf{e}, \mathbf{Xe} \rangle_2(\langle \mathbf{e}, \mathbf{y} \rangle_2 + \langle \mathbf{y}, \mathbf{e} \rangle_2)}{\|\mathbf{e}\|_2^4} \\
&= \langle \mathbf{e}, \mathbf{Xy} \rangle_2 + \langle \mathbf{y}, \mathbf{Xe} \rangle_2 - \langle \mathbf{e}, \mathbf{Xe} \rangle_2(\langle \mathbf{e}, \mathbf{y} \rangle_2 + \langle \mathbf{y}, \mathbf{e} \rangle_2) \\
&= \overline{\langle \mathbf{Xy}, \mathbf{e} \rangle_2} + \langle \mathbf{y}, \mathbf{Xe} \rangle_2 - \langle \mathbf{e}, \mathbf{Xe} \rangle_2(\overline{\langle \mathbf{y}, \mathbf{e} \rangle_2} + \langle \mathbf{y}, \mathbf{e} \rangle_2) \\
&= \overline{\langle \mathbf{y}, \mathbf{Xe} \rangle_2} + \langle \mathbf{y}, \mathbf{Xe} \rangle_2 - \langle \mathbf{e}, \mathbf{Xe} \rangle_2(\overline{\langle \mathbf{y}, \mathbf{e} \rangle_2} + \langle \mathbf{y}, \mathbf{e} \rangle_2) \\
&= 2\,\mathrm{Re}\langle \mathbf{y}, \mathbf{Xe} \rangle_2 - 2\langle \mathbf{e}, \mathbf{Xe} \rangle_2\,\mathrm{Re}\langle \mathbf{y}, \mathbf{e} \rangle_2 \\
&= 2\,\mathrm{Re}\,\langle \mathbf{y}, \mathbf{Xe} - \langle \mathbf{Xe}, \mathbf{e} \rangle_2 \mathbf{e} \rangle_2 \qquad \text{für alle } \mathbf{y} \in \mathbb{K}^{\mathcal{I}}.
\end{aligned}$$

Indem wir $\mathbf{y} := \mathbf{Xe} - \langle \mathbf{Xe}, \mathbf{e} \rangle_2 \mathbf{e}$ setzen, erhalten wir $\|\mathbf{y}\|_2 = 0$, also

$$\mathbf{Xe} = \lambda \mathbf{e}$$

für $\lambda := \langle \mathbf{X}\mathbf{e}, \mathbf{e}\rangle_2 = f(\mathbf{e})$, also ist $\mathbf{e}$ ein Eigenvektor von $\mathbf{X}$ zu dem Eigenwert λ.

Wir können wie im Beweis des Lemmas 2.20 fortfahren, indem wir eine Householder-Spiegelung $\mathbf{H}$ wählen, die $\mathbf{H}\delta_1 = \gamma\mathbf{e}$ für den ersten kanonischen Einheitsvektor δ_1 und ein $\gamma \in \mathbb{K}$ erfüllt. Dann folgt wieder

$$\mathbf{H}^*\mathbf{X}\mathbf{H}\delta_1 = \mathbf{H}^*\mathbf{X}\gamma\mathbf{e} = \lambda\mathbf{H}^*\gamma\mathbf{e} = \lambda\delta_1,$$

also hat die Matrix $\mathbf{H}^*\mathbf{X}\mathbf{H}$ die Form

$$\mathbf{H}^*\mathbf{X}\mathbf{H} = \begin{pmatrix} \lambda & \mathbf{B} \\ & \widehat{\mathbf{X}} \end{pmatrix}$$

für $\widehat{\mathbf{X}} \in \mathbb{K}^{\mathcal{I}_n \times \mathcal{I}_n}$, $\mathbf{B} \in \mathbb{K}^{\{i_1\} \times \mathcal{I}_n}$ und $\mathcal{I}_n := \mathcal{I} \setminus \{i_1\}$. Da $\mathbf{X}$ selbstadjungiert ist, folgt

$$\begin{pmatrix} \lambda & \mathbf{B} \\ & \widehat{\mathbf{X}} \end{pmatrix} = \mathbf{H}^*\mathbf{X}\mathbf{H} = \mathbf{H}^*\mathbf{X}^*\mathbf{H} = (\mathbf{H}^*\mathbf{X}\mathbf{H})^* = \begin{pmatrix} \lambda & \mathbf{B} \\ & \widehat{\mathbf{X}} \end{pmatrix}^* = \begin{pmatrix} \bar{\lambda} & \\ \mathbf{B}^* & \widehat{\mathbf{X}}^* \end{pmatrix},$$

also $\lambda = \bar{\lambda}$, $\mathbf{B} = \mathbf{0}$ und $\widehat{\mathbf{X}}^* = \widehat{\mathbf{X}}$.

Wegen $\#\mathcal{I}_n = \#\mathcal{I} - 1 = n$ können wir die Induktionsvoraussetzung auf $\widehat{\mathbf{X}}$ anwenden und erhalten eine orthogonale Matrix $\widehat{\mathbf{Q}} \in \mathbb{K}^{\mathcal{I}_n \times \mathcal{I}_n}$ und eine Diagonalmatrix $\widehat{\mathbf{D}} \in \mathbb{K}^{\mathcal{I}_n \times \mathcal{I}_n}$ mit $\widehat{\mathbf{X}} = \widehat{\mathbf{Q}}\widehat{\mathbf{D}}\widehat{\mathbf{Q}}^*$, also folgt

$$\begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{Q}}^* \end{pmatrix} \widehat{\mathbf{H}}^*\mathbf{X}\widehat{\mathbf{H}} \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{Q}} \end{pmatrix} = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{Q}}^* \end{pmatrix} \begin{pmatrix} \lambda & \\ & \widehat{\mathbf{X}} \end{pmatrix} \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{Q}} \end{pmatrix} = \begin{pmatrix} \lambda & \\ & \widehat{\mathbf{D}} \end{pmatrix},$$

so dass wir mit

$$\mathbf{Q} := \widehat{\mathbf{H}} \begin{pmatrix} 1 & \\ & \widehat{\mathbf{Q}} \end{pmatrix}, \qquad \mathbf{D} := \begin{pmatrix} \lambda & \\ & \widehat{\mathbf{D}} \end{pmatrix}$$

die Gleichung $\mathbf{Q}^*\mathbf{X}\mathbf{Q} = \mathbf{D}$ erhalten, die offenbar zu $\mathbf{X} = \mathbf{Q}\mathbf{D}\mathbf{Q}^*$ äquivalent ist. ■

Wir haben bereits gesehen, dass sich der Spektralradius durch eine beliebige induzierte Matrixnorm abschätzen lässt. Für die Spektralnorm können wir unter Zuhilfenahme des soeben bewiesenen Lemmas diese Abschätzung wie folgt verbessern:

Lemma 2.44 (Spektralnorm) *Sei $\mathbf{X} \in \mathbb{K}^{\mathcal{I} \times \mathcal{J}}$. Es gilt $\|\mathbf{X}\|_2 = \varrho(\mathbf{X}^*\mathbf{X})^{1/2}$. Falls $\mathbf{X}$ selbstadjungiert ist, gilt sogar $\|\mathbf{X}\|_2 = \varrho(\mathbf{X})$.*

Beweis. Wir untersuchen zunächst den Fall einer selbstadjungierten Matrix $\mathbf{Y} \in \mathbb{K}^{\mathcal{I} \times \mathcal{I}}$. Nach Lemma 2.43 gibt es eine orthogonale Matrix $\mathbf{Q} \in \mathbb{K}^{\mathcal{I} \times \mathcal{I}}$ und eine reelle Diagonalmatrix $\mathbf{D} \in \mathbb{R}^{\mathcal{I} \times \mathcal{I}}$, die $\mathbf{Y} = \mathbf{Q}\mathbf{D}\mathbf{Q}^*$ erfüllen.

Für einen Vektor $\mathbf{x} \in \mathbb{K}^{\mathcal{I}}$ erhalten wir

$$\|\mathbf{Y}\mathbf{x}\|_2^2 = \langle \mathbf{Y}\mathbf{x}, \mathbf{Y}\mathbf{x}\rangle_2 = \langle \mathbf{Q}\mathbf{D}\mathbf{Q}^*\mathbf{x}, \mathbf{Q}\mathbf{D}\mathbf{Q}^*\mathbf{x}\rangle_2 = \langle \mathbf{Q}^*\mathbf{Q}\mathbf{D}\mathbf{y}, \mathbf{D}\mathbf{y}\rangle_2 = \langle \mathbf{D}^2\mathbf{y}, \mathbf{y}\rangle_2$$

mit $\mathbf{y} := \mathbf{Q}^*\mathbf{x}$.

Da $\mathbf{D}$ und $\mathbf{Y}$ ähnlich sind, besitzen sie dasselbe Spektrum. Also folgt $|D_{ii}| \leq \varrho(\mathbf{Y})$ für alle $i \in \mathcal{I}$, und nach Lemma 2.42 und der Definition des euklidischen Skalarprodukts gilt

$$\begin{aligned}\langle \mathbf{y}, \mathbf{D}^2\mathbf{y}\rangle_2 &= \sum_{i\in\mathcal{I}} \bar{y}_i D_{ii}^2 y_i = \sum_{i\in\mathcal{I}} D_{ii}^2 |y_i|^2 \leq \varrho(\mathbf{Y})^2 \sum_{i\in\mathcal{I}} |y_i|^2 \\ &= \varrho(\mathbf{Y})^2 \|\mathbf{y}\|_2^2 = \varrho(\mathbf{Y})^2 \|\mathbf{Q}^*\mathbf{x}\|_2^2 = \varrho(\mathbf{Y})^2 \|\mathbf{x}\|_2^2.\end{aligned}$$

Damit haben wir $\|\mathbf{Y}\|_2 \leq \varrho(\mathbf{Y})$ bewiesen, und aus Lemma 2.17 folgt $\|\mathbf{Y}\|_2 = \varrho(\mathbf{Y})$.

Um den allgemeinen Fall zu behandeln, setzen wir $\mathbf{Y} := \mathbf{X}^*\mathbf{X}$ und erhalten

$$\begin{aligned}\|\mathbf{X}\mathbf{x}\|_2^2 &= \langle \mathbf{X}\mathbf{x}, \mathbf{X}\mathbf{x}\rangle_2 = \langle \mathbf{x}, \mathbf{X}^*\mathbf{X}\mathbf{x}\rangle_2 = \langle \mathbf{x}, \mathbf{Y}\mathbf{x}\rangle_2 \\ &\leq \|\mathbf{Y}\mathbf{x}\|_2 \|\mathbf{x}\|_2 \leq \varrho(\mathbf{Y}) \|\mathbf{x}\|_2^2 \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^{\mathcal{I}}.\end{aligned} \tag{2.11}$$

Damit ist $\|\mathbf{X}\|_2^2 \leq \varrho(\mathbf{Y})$ bewiesen.

Sei $\lambda \in \sigma(\mathbf{Y})$ mit $|\lambda| = \varrho(\mathbf{Y})$, und sei $\mathbf{e} \in \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\}$ ein entsprechender Eigenvektor. Aus (2.11) folgt

$$\|\mathbf{X}\mathbf{e}\|_2^2 = \langle \mathbf{Y}\mathbf{e}, \mathbf{e}\rangle_2 = \langle \lambda\mathbf{e}, \mathbf{e}\rangle_2 = \lambda \|\mathbf{e}\|_2^2,$$

also $\lambda \in \mathbb{R}_{\geq 0}$, somit muss $\lambda = |\lambda| = \varrho(\mathbf{Y})$ gelten und wir erhalten

$$\|\mathbf{X}\|_2^2 = \sup\left\{ \frac{\|\mathbf{X}\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} \ : \ \mathbf{x} \in \mathbb{K}^{\mathcal{I}} \right\} \geq \frac{\|\mathbf{X}\mathbf{e}\|_2^2}{\|\mathbf{e}\|_2^2} = \lambda = \varrho(\mathbf{Y}),$$

also $\|\mathbf{X}\|_2^2 = \varrho(\mathbf{Y})$. ■

Im Fall des Richardson-Verfahrens haben wir gesehen, dass wir nur auf Konvergenz hoffen dürfen, wenn das Spektrum der Matrix $\mathbf{A}$ in einem Kreis enthalten ist, der die Null nicht enthält. Für eine selbstadjungierte Matrix liegt das Spektrum gemäß Lemma 2.41 auf der reellen Achse, das Richardson-Verfahren kann also nur konvergieren, wenn das Spektrum entweder nur aus echt positiven oder echt negativen Eigenwerten besteht. Der Einfachheit halber beschränken wir uns auf den Fall eines positiven Spektrums:

Definition 2.45 (Positiv definit) *Sei $\mathbf{X} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ eine Matrix. Wir nennen $\mathbf{X}$* positiv definit*, falls $\mathbf{X}$ selbstadjungiert ist und die Ungleichung*

$$\langle \mathbf{x}, \mathbf{X}\mathbf{x}\rangle_2 > 0 \qquad \textit{für alle } \mathbf{x} \in \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\} \textit{ gilt.}$$

Falls $\mathbf{X}$ selbstadjungiert ist und

$$\langle \mathbf{x}, \mathbf{X}\mathbf{x}\rangle_2 \geq 0 \qquad \textit{für alle } \mathbf{x} \in \mathbb{K}^{\mathcal{I}} \textit{ gilt,}$$

nennen wir $\mathbf{X}$ positiv semidefinit*.*

Anstelle von „$\mathbf{X}$ ist positiv definit“ schreiben wir kurz $\mathbf{X} > \mathbf{0}$, anstelle von „$\mathbf{X}$ ist positiv semidefinit“ schreiben wir kurz $\mathbf{X} \geq \mathbf{0}$.

In der Literatur sieht man häufig eine Definition des Begriffs „positiv definit“, die auf die Forderung der Selbstadjungiertheit verzichtet. Wir schließen diese Eigenschaft in die Definition ein, weil in praktisch allen von uns hier untersuchten Fällen beide Eigenschaften gleichzeitig auftreten.

Positiv definite Matrizen erlauben es uns, einem bestimmten Problem angepasste Normen zu definieren, die für die Untersuchung von Konvergenzeigenschaften von entscheidender Bedeutung sind.

Definition 2.46 (Energienorm) *Sei* $\mathbf{X} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ *positiv definit. Die durch*

$$f : \mathbb{K}^{\mathcal{I}} \to \mathbb{R}_{\geq 0}, \qquad \mathbf{x} \mapsto \langle \mathbf{x}, \mathbf{X}\mathbf{x}\rangle_2^{1/2},$$

definierte Abbildung ist eine Norm, die wir die zu $\mathbf{X}$ *gehörende* Energienorm *nennen und mit* $\|\cdot\|_X$ *bezeichnen.*

Wir haben bereits gesehen, dass sich die Spektralnorm durch den Spektralradius darstellen lässt, und diese Darstellung erleichtert viele Beweise erheblich. Deshalb ist es nun unser Ziel, eine entsprechende Darstellung für die durch die Energienorm induzierte Matrixnorm zu gewinnen.

Mit Hilfe des Begriffs der positiv definiten Matrix können wir eine Halbordnung auf dem Raum der selbstadjungierten Matrizen definieren:

Definition 2.47 *Für alle selbstadjungierten Matrizen* $\mathbf{X}, \mathbf{Y} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ *definieren wir*

$$\begin{aligned} \mathbf{X} < \mathbf{Y} &:\iff \mathbf{0} < \mathbf{Y} - \mathbf{X}, \\ \mathbf{X} \leq \mathbf{Y} &:\iff \mathbf{0} \leq \mathbf{Y} - \mathbf{X}. \end{aligned}$$

Lemma 2.48 *Seien* $\mathbf{X}, \mathbf{Y} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ *selbstadjungierte Matrizen, die* $\mathbf{X} < \mathbf{Y}$ *erfüllen. Sei* $\mathbf{T} \in \mathbb{K}^{\mathcal{I}\times\mathcal{J}}$ *injektiv. Dann gilt auch*

$$\mathbf{T}^*\mathbf{X}\mathbf{T} < \mathbf{T}^*\mathbf{Y}\mathbf{T}.$$

Beweis. Sei $\mathbf{x} \in \mathbb{K}^{\mathcal{J}} \setminus \{\mathbf{0}\}$. Dann gilt nach Lemma 2.39

$$\langle \mathbf{T}^*(\mathbf{Y} - \mathbf{X})\mathbf{T}\mathbf{x}, \mathbf{x}\rangle_2 = \langle (\mathbf{Y} - \mathbf{X})\mathbf{T}\mathbf{x}, \mathbf{T}\mathbf{x}\rangle_2 = \langle (\mathbf{Y} - \mathbf{X})\mathbf{y}, \mathbf{y}\rangle_2 > 0$$

für $\mathbf{y} := \mathbf{T}\mathbf{x} \neq \mathbf{0}$. ∎

Lemma 2.49 *Sei* $\mathbf{X} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ *selbstadjungiert. Seien* $\alpha, \beta \in \mathbb{R}$ *mit* $\alpha \leq \beta$ *gegeben. Es gelten*

$$\begin{aligned} \alpha\mathbf{I} < \mathbf{X} &\iff \sigma(\mathbf{X}) \subseteq \mathbb{R}_{>\alpha}, \\ \alpha\mathbf{I} \leq \mathbf{X} &\iff \sigma(\mathbf{X}) \subseteq \mathbb{R}_{\geq\alpha}, \\ \alpha\mathbf{I} < \mathbf{X} < \beta\mathbf{I} &\iff \sigma(\mathbf{X}) \subseteq (\alpha, \beta), \\ \alpha\mathbf{I} \leq \mathbf{X} \leq \beta\mathbf{I} &\iff \sigma(\mathbf{X}) \subseteq [\alpha, \beta]. \end{aligned}$$

Beweis. Da $\mathbf{X}$ selbstadjungiert ist, gibt es nach Lemma 2.43 eine orthogonale Matrix $\mathbf{Q} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ und eine reelle Diagonalmatrix $\mathbf{D} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ mit $\mathbf{X} = \mathbf{QDQ}^*$.

Die Matrix $\mathbf{X} - \alpha\mathbf{I}$ ist nach Lemma 2.48 genau dann positiv definit, wenn

$$\mathbf{Q}^*(\mathbf{X} - \alpha\mathbf{I})\mathbf{Q} = \mathbf{D} - \alpha\mathbf{I}$$

es ist, also genau dann, wenn alle Diagonalelemente von $\mathbf{D}$ größer als α sind. Da die Diagonalelemente von $\mathbf{D}$ gerade die Eigenwerte von $\mathbf{X}$ sind, ist die erste Aussage damit bewiesen.

Die restlichen Aussagen lassen sich analog behandeln. ■

Lemma 2.50 (Wurzel einer Matrix) *Sei $\mathbf{X} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ positiv semidefinit. Es gibt genau eine positiv semidefinite Matrix $\mathbf{Y} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$, für die $\mathbf{Y}^2 = \mathbf{X}$ gilt. Diese Matrix nennen wir die* Wurzel *von $\mathbf{X}$ und bezeichnen sie mit $\mathbf{X}^{1/2} := \mathbf{Y}$.*

Falls $\mathbf{X}$ positiv definit ist, ist auch $\mathbf{X}^{1/2}$ positiv definit, und wir bezeichnen seine Inverse mit $\mathbf{X}^{-1/2}$.

Beweis. Zunächst beweisen wir die Existenz von $\mathbf{Y}$. Da $\mathbf{X}$ selbstadjungiert ist, gibt es eine orthogonale Matrix $\mathbf{Q} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ und eine reelle Diagonalmatrix $\mathbf{D} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$, die $\mathbf{X} = \mathbf{QDQ}^*$ erfüllen. Aufgrund der Lemmas 2.48 und 2.49 sind alle Eigenwerte, also Diagonaleinträge, von $\mathbf{D}$ nicht-negativ. Wir definieren $\mathbf{R} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ durch

$$R_{ij} = \begin{cases} D_{ii}^{1/2} & \text{falls } i = j, \\ 0 & \text{sonst} \end{cases} \qquad \text{für alle } i, j \in \mathcal{I},$$

und setzen $\mathbf{Y} := \mathbf{QRQ}^*$. Offenbar gilt

$$\mathbf{Y}^2 = \mathbf{QRQ}^*\mathbf{QRQ}^* = \mathbf{QR}^2\mathbf{Q}^* = \mathbf{QDQ}^* = \mathbf{X}.$$

Damit ist $\mathbf{Y}$ die gesuchte Matrix.

Man kann leicht nachrechnen, dass jeder Eigenvektor $\mathbf{e} \in \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\}$ von $\mathbf{X}$ zum Eigenwert λ auch ein Eigenvektor von $\mathbf{Y}$ zum Eigenwert $\lambda^{1/2}$ ist.

Sei nun $\mathbf{Z} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ eine zweite positiv semidefinite Matrix, die $\mathbf{Z}^2 = \mathbf{X}$ erfüllt. Sei $\mathbf{e} \in \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\}$ ein Eigenvektor von $\mathbf{Z}$ zum Eigenwert μ. Es folgt $\mathbf{Xe} = \mathbf{Z}^2\mathbf{e} = \mu^2\mathbf{e}$, also ist $\mathbf{e}$ ein Eigenvektor von $\mathbf{X}$ zum Eigenwert $\lambda := \mu^2$. Nach der obigen Beobachtung muss $\mathbf{e}$ dann auch ein Eigenvektor von $\mathbf{Y}$ zum Eigenwert $\lambda^{1/2} = \mu$ sein, und wir haben

$$\mathbf{Ze} = \mu\mathbf{e} = \mathbf{Ye}$$

bewiesen. Da $\mathbf{Z}$ selbstadjungiert ist, existiert eine Basis aus derartigen Eigenvektoren, also muss $\mathbf{Z} = \mathbf{Y}$ gelten.

Sei nun $\mathbf{X}$ positiv definit. Dann ist $\mathbf{X}_1 := \mathbf{X}^{1/2}$ positiv semidefinit, also ist auch $\mathbf{X}_2 := \mathbf{X}_1^{1/2}$ positiv semidefinit. Da $(\mathbf{X}_2)^4 = (\mathbf{X}_1)^2 = \mathbf{X}$ regulär ist, muss auch $\mathbf{X}_2$ selbst regulär sein. Sei $\mathbf{x} \in \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\}$, und sei $\mathbf{y} := \mathbf{X}_2{}^{-1}\mathbf{x}$. Aus $\mathbf{y} \neq \mathbf{0}$ folgt

$$\langle \mathbf{X}_1\mathbf{x}, \mathbf{x}\rangle_2 = \langle \mathbf{X}_1\mathbf{X}_2\mathbf{y}, \mathbf{X}_2\mathbf{y}\rangle_2 = \langle \mathbf{X}_2\mathbf{X}_1\mathbf{X}_2\mathbf{y}, \mathbf{y}\rangle_2 = \langle \mathbf{Xy}, \mathbf{y}\rangle_2 > 0,$$

also ist auch $\mathbf{X}_1 = \mathbf{X}^{1/2}$ positiv definit. ■

Mit Hilfe der Wurzel einer Matrix können wir nun die gewünschte Darstellung der Energienorm gewinnen.

Lemma 2.51 *Sei $\mathbf{X} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ positiv definit. Dann gilt*

$$\|\mathbf{x}\|_X = \|\mathbf{X}^{1/2}\mathbf{x}\|_2 \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^{\mathcal{I}}.$$

Für die von der Energienorm induzierte Matrixnorm erhalten wir

$$\|\mathbf{Y}\|_X = \|\mathbf{X}^{1/2}\mathbf{Y}\mathbf{X}^{-1/2}\|_2 \qquad \text{für alle } \mathbf{Y} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}.$$

Beweis. Sei $\mathbf{x} \in \mathbb{K}^{\mathcal{I}}$. Nach Lemma 2.39 gilt

$$\|\mathbf{x}\|_X^2 = \langle \mathbf{x}, \mathbf{X}\mathbf{x}\rangle_2 = \langle \mathbf{x}, \mathbf{X}^{1/2}\mathbf{X}^{1/2}\mathbf{x}\rangle_2 = \langle \mathbf{X}^{1/2}\mathbf{x}, \mathbf{X}^{1/2}\mathbf{x}\rangle_2 = \|\mathbf{X}^{1/2}\mathbf{x}\|_2^2.$$

Zum Beweis der zweiten Aussage können wir direkt die Definition der induzierten Matrixnorm einsetzen: Für $\mathbf{Y} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ gilt

$$\begin{aligned}
\|\mathbf{Y}\|_X &= \sup\left\{\frac{\|\mathbf{Y}\mathbf{x}\|_X}{\|\mathbf{x}\|_X} \;:\; \mathbf{x} \in \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\}\right\} = \sup\left\{\frac{\|\mathbf{X}^{1/2}\mathbf{Y}\mathbf{x}\|_2}{\|\mathbf{X}^{1/2}\mathbf{x}\|_2} \;:\; \mathbf{x} \in \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\}\right\} \\
&= \sup\left\{\frac{\|\mathbf{X}^{1/2}\mathbf{Y}\mathbf{X}^{-1/2}\mathbf{y}\|_2}{\|\mathbf{X}^{1/2}\mathbf{X}^{-1/2}\mathbf{y}\|_2} \;:\; \mathbf{y} \in \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\}\right\} \\
&= \sup\left\{\frac{\|\mathbf{X}^{1/2}\mathbf{Y}\mathbf{X}^{-1/2}\mathbf{y}\|_2}{\|\mathbf{y}\|_2} \;:\; \mathbf{y} \in \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\}\right\} = \|\mathbf{X}^{1/2}\mathbf{Y}\mathbf{X}^{-1/2}\|_2
\end{aligned}$$

infolge der Invertierbarkeit von $\mathbf{X}^{1/2}$. ■

Mit dieser Darstellung der Energienorm können wir uns nun wieder der Analyse des Jacobi-Verfahrens zuwenden: Um seine Konvergenz nachzuweisen, müssen wir den Spektralradius von

$$\mathbf{M}_{\mathrm{Jac},\theta} = \mathbf{I} - \theta\mathbf{D}^{-1}\mathbf{A}$$

abschätzen. Nach Lemma 2.17 genügt es dazu, eine beliebige induzierte Matrixnorm dieser Matrix abzuschätzen. Falls $\mathbf{A} > 0$ gilt, können wir die korrespondierende Energienorm $\|\cdot\|_A$ verwenden und erhalten

$$\|\mathbf{M}_{\mathrm{Jac},\theta}\|_A = \|\mathbf{A}^{1/2}(\mathbf{I} - \theta\mathbf{D}^{-1}\mathbf{A})\mathbf{A}^{-1/2}\|_2 = \|\mathbf{I} - \theta\mathbf{A}^{1/2}\mathbf{D}^{-1}\mathbf{A}^{1/2}\|_2.$$

Da das Argument der Spektralnorm selbstadjungiert ist, können wir die weiteren Abschätzungen dank Lemma 2.44 auf eine Analyse der Eigenwerte reduzieren.

Satz 2.52 (Konvergenz) *Sei Φ ein lineares konsistentes Iterationsverfahren, und seien $\mathbf{M}, \mathbf{N} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ die entsprechenden Matrizen der ersten Normalform.*

Seien $\mathbf{A}$ *und* $\mathbf{N}$ *positiv definit. Falls die Matrix* $\mathbf{W} := \mathbf{N}^{-1}$ *die Bedingung*

$$0 < \mathbf{A} < 2\mathbf{W} \tag{2.12}$$

erfüllt, ist Φ *konvergent mit*

$$\varrho(\mathbf{M}) = \|\mathbf{M}\|_A = \|\mathbf{M}\|_W.$$

Falls für $\alpha, \beta \in \mathbb{R}$ *die Bedingung*

$$\alpha\mathbf{W} \leq \mathbf{A} \leq \beta\mathbf{W} \tag{2.13}$$

gilt, erhalten wir die Schranke

$$\varrho(\mathbf{M}) \leq \max\{|1-\alpha|, |1-\beta|\}$$

für die Konvergenzrate.

Beweis. Da die Matrizen $\mathbf{A}$ und $\mathbf{W}$ positiv definit sind, sind nach Lemma 2.50 ihre Wurzeln $\mathbf{A}^{1/2}$ und $\mathbf{W}^{1/2}$ wohldefiniert und regulär. Die Matrix $\mathbf{M}$ und die transformierten Matrizen

$$\begin{aligned}
\mathbf{M}_A &:= \mathbf{A}^{1/2}\mathbf{M}\mathbf{A}^{-1/2} = \mathbf{A}^{1/2}(\mathbf{I} - \mathbf{N}\mathbf{A})\mathbf{A}^{-1/2} = \mathbf{I} - \mathbf{A}^{1/2}\mathbf{N}\mathbf{A}^{1/2}, \\
\mathbf{M}_W &:= \mathbf{W}^{1/2}\mathbf{M}\mathbf{W}^{-1/2} = \mathbf{I} - \mathbf{W}^{1/2}\mathbf{W}^{-1}\mathbf{A}\mathbf{W}^{-1/2} = \mathbf{I} - \mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2}
\end{aligned}$$

sind ähnlich, also gilt $\sigma(\mathbf{M}) = \sigma(\mathbf{M}_A) = \sigma(\mathbf{M}_W)$.

Da $\mathbf{M}_A$ und $\mathbf{M}_W$ selbstadjungiert sind, erhalten wir mit Hilfe von Lemma 2.51

$$\begin{aligned}
\varrho(\mathbf{M}) &= \varrho(\mathbf{M}_A) = \|\mathbf{M}_A\|_2 = \|\mathbf{A}^{1/2}\mathbf{M}\mathbf{A}^{-1/2}\|_2 = \|\mathbf{M}\|_A, \\
\varrho(\mathbf{M}) &= \varrho(\mathbf{M}_W) = \|\mathbf{M}_W\|_2 = \|\mathbf{W}^{1/2}\mathbf{M}\mathbf{W}^{-1/2}\|_2 = \|\mathbf{M}\|_W.
\end{aligned}$$

Mit Hilfe von Lemma 2.48 folgt aus (2.12) bereits

$$\mathbf{0} < \mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2} < 2\mathbf{I},$$

also nach Lemma 2.49 auch

$$\sigma(\mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2}) \subseteq (0, 2).$$

Für die Iterationsmatrix $\mathbf{M}$ bedeutet diese Inklusion

$$\varrho(\mathbf{M}) = \varrho(\mathbf{M}_W) = \varrho(\mathbf{I} - \mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2}) < 1,$$

also ist Φ nach Satz 2.25 konvergent. Falls (2.13) gilt, erhalten wir analog

$$\sigma(\mathbf{M}_W) \subseteq [1-\beta, 1-\alpha],$$

also die Schranke

$$\varrho(\mathbf{M}) = \varrho(\mathbf{M}_W) \leq \max\{|1-\beta|, |1-\alpha|\}$$

für die Konvergenzrate. ■

Dieses allgemeine Kriterium können wir nun auf die ungedämpfte oder gedämpfte Jacobi-Iteration anwenden:

Lemma 2.53 *Sei* $\mathbf{A}$ *positiv definit. Falls*

$$\mathbf{0} < \mathbf{A} < 2\mathbf{D}$$

gilt, ist das Jacobi-Verfahren Φ_{Jac} *konvergent. Falls*

$$\mathbf{0} < \mathbf{A} < \frac{2}{\theta}\mathbf{D}$$

für ein $\theta \in \mathbb{R}_{>0}$ *gilt, ist das gedämpfte Jacobi-Verfahren* $\Phi_{\mathrm{Jac},\theta}$ *konvergent.*

Es gibt ein $\theta_{\max} \in \mathbb{R}_{>0}$ *so, dass diese Bedingung für alle* $\theta \in (0, \theta_{\max})$ *gilt.*

Beweis. Wir wenden Satz 2.52 auf $\mathbf{N} = \mathbf{D}^{-1}$ beziehungsweise $\mathbf{N} = \theta\mathbf{D}^{-1}$ an.

Um zu beweisen, dass es einen Dämpfungsparameter gibt, der Konvergenz garantiert, setzen wir

$$\lambda_{\max} := \max\{\lambda \ : \ \lambda \in \sigma(\mathbf{A})\}, \qquad d_{\min} := \min\{D_{ii} \ : \ i \in \mathcal{I}\}.$$

Da $\mathbf{A}$ positiv definit ist, gilt $\lambda_{\max} \in \mathbb{R}_{>0}$, und $\mathbf{D}$ ist ebenfalls positiv definit, also muss auch $d_{\min} \in \mathbb{R}_{>0}$ gelten. Wir setzen $\theta_{\max} := 2d_{\min}/\lambda_{\max}$ und erhalten mit Hilfe von Lemma 2.49

$$\mathbf{0} < \mathbf{A} \leq \lambda_{\max}\mathbf{I} = 2\frac{\lambda_{\max}}{2d_{\min}}d_{\min}\mathbf{I} = \frac{2}{\theta_{\max}}d_{\min}\mathbf{I} \leq \frac{2}{\theta_{\max}}\mathbf{D} < \frac{2}{\theta}\mathbf{D},$$

für alle $\theta \in (0, \theta_{\max})$, also ist $\Phi_{\mathrm{Jac},\theta}$ konvergent. ■

Wir stellen also fest, dass wir die Konvergenz des Jacobi-Verfahrens für jede beliebige positiv definite Matrix sicherstellen können, indem wir den Dämpfungsparameter θ klein genug wählen.

2.6 Diagonaldominante Matrizen*

Die Konvergenz des Jacobi-Verfahrens lässt sich unter bestimmten Bedingungen auch für nicht positiv definite Matrizen nachweisen, beispielsweise für *diagonaldominante* Matrizen.

Für die Konvergenz ist der Spektralradius der Iterationsmatrix $\mathbf{M} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}$ ausschlaggebend. Die Koeffizienten dieser Matrix sind durch

$$M_{ij} = \begin{cases} 0 & \text{falls } i = j, \\ -A_{ij}/A_{ii} & \text{ansonsten} \end{cases} \qquad \text{für alle } i, j \in \mathcal{I}$$

gegeben. Aus dieser Darstellung lässt sich bereits eine Aussage über die Eigenwerte der Matrix $\mathbf{M}$ gewinnen.

Satz 2.54 (Gerschgorin) *Sei* $\mathbf{M} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ *eine Matrix. Wir definieren die* Gerschgorin-Kreise

$$\mathcal{D}_i := \left\{ z \in \mathbb{C} \ : \ |z - M_{ii}| \leq \sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |M_{ij|} \right\}.$$

Dann gilt

$$\sigma(\mathbf{M}) \subseteq \bigcup_{i\in\mathcal{I}} \mathcal{D}_i,$$

jeder Eigenwert der Matrix $\mathbf{M}$ *ist also in mindestens einem Gerschgorin-Kreis enthalten.*

Beweis. Sei $\lambda \in \sigma(\mathbf{M})$ ein Eigenwert der Matrix $\mathbf{M}$, und sei $\mathbf{e} \in \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\}$ ein zugehöriger Eigenvektor.

Wir wählen $i \in \mathcal{I}$ als den Index, für den $|e_i|$ maximal wird, es gilt also

$$|e_j| \leq |e_i| \qquad \text{für alle } j \in \mathcal{I}.$$

Aus $\mathbf{e} \neq \mathbf{0}$ folgt $|e_i| > 0$, und mit der Dreiecksungleichung erhalten wir

$$\begin{aligned}
\lambda e_i = (\lambda \mathbf{e})_i = (\mathbf{M}\mathbf{e})_i &= \sum_{j\in\mathcal{I}} M_{ij} e_j, \\
(\lambda - M_{ii}) e_i &= \sum_{\substack{j\in\mathcal{I}\\ j\neq i}} M_{ij} e_j \\
|\lambda - M_{ii}|\,|e_i| &\leq \sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |M_{ij}|\,|e_j| \\
|\lambda - M_{ii}| &\leq \sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |M_{ij}| \frac{|e_j|}{|e_i|} \leq \sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |M_{ij}|,
\end{aligned}$$

also $\lambda \in \mathcal{D}_i$. ∎

In unserem Fall gilt $M_{ii} = 0$ für alle $i \in \mathcal{I}$, die Null ist also der Mittelpunkt aller Gerschgorin-Kreise. Um Konvergenz zu erhalten, müssen wir sicherstellen, dass die Radien aller Kreise echt kleiner als eins sind, dass also die Bedingung

$$1 > \sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |M_{ij}| = \frac{1}{|A_{ii}|} \sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |A_{ij}| \qquad \text{für alle } i \in \mathcal{I} \tag{2.14}$$

erfüllt ist.

Definition 2.55 (Streng diagonaldominante Matrix) *Sei* $\mathbf{A} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ *eine Matrix. Wir nennen sie* streng diagonaldominant*, falls*

$$\sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |A_{ij}| < |A_{ii}| \qquad \textit{für alle } i \in \mathcal{I} \textit{ gilt.}$$

Folgerung 2.56 (Konvergenz) *Sei* $\mathbf{A} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ *eine streng diagonaldominante Matrix. Dann konvergiert das ungedämpfte Jacobi-Verfahren.*

Beweis. Die strenge Diagonaldominanz stellt sicher, dass (2.14) gilt. Mit Satz 2.54 folgt $\varrho(\mathbf{M}_{\mathrm{Jac}}) < 1$, also mit Satz 2.25 die Konvergenz des Jacobi-Verfahrens. ∎

Leider ist strenge Diagonaldominanz eine relativ selten anzutreffende Eigenschaft. Man kann sie allerdings so abschwächen, dass sich immerhin für unsere Modellprobleme eine Aussage gewinnen lässt. Dazu brauchen wir eine weitere Eigenschaft die Matrix $\mathbf{A}$.

Definition 2.57 (Irreduzible Matrix) *Sei* $\mathbf{A} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ *eine Matrix. Eine Folge* $(i_\ell)_{\ell=0}^p$ *nennen wir* Pfad im Graphen der Matrix $\mathbf{A}$ von i_0 zu i_p, *falls*

$$A_{i_{\ell-1},i_\ell} \neq 0 \qquad \text{für alle } \ell \in [1:p]$$

gilt. Die Matrix $\mathbf{A}$ *nennen wir* irreduzibel, *falls für alle* $i, j \in \mathcal{I}$ *ein Pfad im Graphen von* i *zu* j *existiert.*

Falls $\mathbf{A} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ irreduzibel (mit von null verschiedenen Diagonalelementen) ist, gilt dasselbe offenbar auch für $\mathbf{M}$.

Satz 2.58 (Gerschgorin) *Sei* $\mathbf{M} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ *eine irreduzible Matrix. Seien* $\mathbf{D}_i$ *die aus Satz 2.54 bekannten Gerschgorin-Kreise.*

Falls ein Eigenwert $\lambda \in \sigma(\mathbf{M})$ *auf dem Rand der Vereinigung*

$$\bigcup_{i\in\mathcal{I}} \mathcal{D}_i$$

aller Kreise liegt, liegt er auch auf dem Rand aller Kreise.

Beweis. Der Beweis folgt dem Buch „Iterative Methods for Sparse Linear Systems“ von Yousef Saad (SIAM, 2003).

Wir gehen wie im Beweis des Satzes 2.54 vor. Sei $\lambda \in \sigma(\mathbf{M})$, und sei $\mathbf{e} \in \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\}$ ein zugehöriger Eigenvektor. Wir wählen wieder $i \in \mathcal{I}$ mit

$$|e_j| \leq |e_i| \qquad \text{für alle } j \in \mathcal{I}$$

und erhalten $\lambda \in \mathcal{D}_i$.

Wir setzen voraus, dass der Eigenwert λ auf dem Rand der Vereinigung der Gerschgorin-Kreise liegt.

Wir zeigen per Induktion über $p \in \mathbb{N}_0$, dass für alle $k \in \mathcal{I}$, für die ein Pfad der Länge p von i zu k existiert, die Gleichungen

$$|\lambda - M_{kk}| = \sum_{\substack{j\in\mathcal{I}\\ j\neq k}} |M_{kj}|, \qquad |e_k| = |e_i| \tag{2.15}$$

gelten.

Induktionsanfang. Sei $k \in \mathcal{I}$ so gegeben, dass ein Pfad der Länge $p = 0$ von i zu k existiert. Nach Definition gilt $k = i$.

Wenn der Eigenwert λ im Inneren des Kreises $\mathcal{D}_i$ liegen würde, läge er auch im Inneren der Vereinigung, und das haben wir soeben ausgeschlossen. λ muss also auf dem Rand des Kreises $\mathcal{D}_i$ liegen, demnach gilt

$$|\lambda - M_{ii}| = \sum_{\substack{j \in \mathcal{I} \\ j \neq i}} |M_{ij}|.$$

Die zweite Gleichung in (2.15) ist wegen $i = k$ trivial erfüllt.

Induktionsvoraussetzung. Sei $p \in \mathbb{N}_0$ so gegeben, dass für alle $k \in \mathcal{I}$, für die ein Pfad der Länge p von i zu k existiert, die Gleichungen (2.15) gelten.

Induktionsschritt. Sei $k \in \mathcal{I}$ so gegeben, dass ein Pfad $(i_\ell)_{\ell=0}^{p+1}$ der Länge $p+1$ von i zu k existiert. Dann ist $(i_\ell)_{\ell=0}^{p}$ ein Pfad der Länge p von i zu $q := i_p$, und es gilt $M_{qk} \neq 0$. Nach Induktionsvoraussetzung gelten

$$|\lambda - M_{qq}| = \sum_{\substack{j \in \mathcal{I} \\ j \neq q}} |M_{qj}|, \qquad\qquad |e_q| = |e_i|.$$

Wir haben

$$(\lambda - M_{qq}) e_q = \sum_{\substack{j \in \mathcal{I} \\ j \neq q}} M_{qj} e_j,$$

$$\sum_{\substack{j \in \mathcal{I} \\ j \neq q}} |M_{qj}|\,|e_i| = |\lambda - M_{qq}|\,|e_q| \leq \sum_{\substack{j \in \mathcal{I} \\ j \neq q}} |M_{qj}|\,|e_j|,$$

so dass

$$0 \leq \sum_{\substack{j \in \mathcal{I} \\ j \neq q}} |M_{qj}|\,(|e_j| - |e_i|)$$

folgt. Da $|e_i|$ maximal gewählt wurde, folgt $|e_j| - |e_i| \leq 0$ für alle $j \in \mathcal{I}$. Da $M_{qk} \neq 0$ nach Definition gilt, folgt daraus insbesondere $|e_k| = |e_i|$. Wie zuvor erhalten wir

$$|\lambda - M_{kk}|\,|e_k| \leq \sum_{\substack{j \in \mathcal{I} \\ j \neq k}} |M_{kj}|\,|e_j| \leq \sum_{\substack{j \in \mathcal{I} \\ j \neq k}} |M_{kj}|\,|e_k|,$$

also $\lambda \in \mathcal{D}_k$. Da λ nicht im Inneren des Kreises $\mathcal{D}_k$ liegen kann, folgt (2.15).

Da $\mathbf{M}$ irreduzibel ist, können wir jeden Index $k \in \mathcal{I}$ mit einem Pfad erreichen, also ist die gewünschte Aussage bewiesen. ■

Definition 2.59 (Irreduzibel diagonaldominant) *Sei* $\mathbf{A} \in \mathbb{K}^{\mathcal{I} \times \mathcal{I}}$ *eine Matrix. Wir nennen sie* irreduzibel diagonaldominant*, falls sie irreduzibel ist,*

$$\sum_{\substack{j \in \mathcal{I} \\ j \neq i}} |A_{ij}| \leq |A_{ii}| \qquad \textit{für alle } i \in \mathcal{I} \textit{ und} \tag{2.16a}$$

$$\sum_{\substack{j\in\mathcal{I}\\ j\neq i}} |A_{ij}| < |A_{ii}| \qquad \textit{für mindestens ein } i \in \mathcal{I} \textit{ gilt.} \tag{2.16b}$$

Folgerung 2.60 (Konvergenz) *Sei* $\mathbf{A} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ *eine irreduzibel diagonaldominante Matrix. Dann konvergiert das ungedämpfte Jacobi-Verfahren.*

Beweis. Wie zuvor folgt aus (2.16a), dass alle Gerschgorin-Kreise der Matrix $\mathbf{M}$ einen Radius von höchstens eins aufweisen können, ihre Vereinigung ist also in dem Einheitskreis um null enthalten.

Aus (2.16b) folgt, dass der Radius mindestens eines Gerschgorin-Kreises $\mathcal{D}_i$ echt kleiner als eins ist.

Per Kontraposition folgt mit Satz 2.58, dass kein Eigenwert auf dem Rand des Einheitskreises liegen kann, denn dann müsste er auch auf dem Rand des Kreises $\mathcal{D}_i$ liegen, dessen Radius echt kleiner als eins ist. ■

Die Matrizen des ein- und zweidimensionalen Modellproblems sind irreduzibel diagonaldominant, so dass wir Konvergenz des ungedämpften Jacobi-Verfahrens erhalten.

2.7 Gauß-Seidel-Iteration

Wir haben bereits gesehen, dass ein gutes Iterationsverfahren zwei Bedingungen erfüllen muss: Die Matrix $\mathbf{N}$ der zweiten Normalform muss effizient berechnet werden können, und sie muss eine gute Approximation der Inversen von $\mathbf{A}$ darstellen.

Das Richardson-Verfahren verwendet $\theta\mathbf{I}$ als Approximation von $\mathbf{A}^{-1}$, das Jacobi-Verfahren stattdessen $\mathbf{D}^{-1}$, denn die Identität und die Inversen von Diagonalmatrizen lassen sich einfach und effizient berechnet. Wir kennen noch weitere Klassen von Matrizen $\mathbf{W}$, für die die Auswertung von $\mathbf{W}^{-1}$ effizient durchgeführt werden kann, beispielsweise obere und untere Dreiecksmatrizen.

Das Gauß-Seidel-Verfahren basiert darauf, die Matrix $\mathbf{A}$ durch eine untere Dreiecksmatrix zu approximieren, deren Inverse dann mit Hilfe des Vorwärtseinsetzens ausgewertet wird. Um überhaupt definieren zu können, was eine untere Dreiecksmatrix ist, müssen wir unsere Indexmenge $\mathcal{I}$ mit einer totalen Ordnung versehen.

Sei $\iota : \mathcal{I} \to \{1, \dots, n\}$ eine Bijektion mit $n := \#\mathcal{I}$, also eine Numerierung der Indexmenge $\mathcal{I}$. Wir zerlegen $\mathbf{A}$ in die Diagonalmatrix $\mathbf{D} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$, die strikte untere Dreiecksmatrix $\mathbf{E} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ und die strikte obere Dreiecksmatrix $\mathbf{F} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ mit

$$\mathbf{A} = \mathbf{D} - \mathbf{E} - \mathbf{F} \tag{2.17}$$

für

$$D_{ij} = \begin{cases} A_{ii} & \text{falls } i = j, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i, j \in \mathcal{I},$$

$$E_{ij} = \begin{cases} -A_{ij} & \text{falls } \iota(i) > \iota(j), \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i, j \in \mathcal{I},$$

$$F_{ij} = \begin{cases} -A_{ij} & \text{falls } \iota(i) < \iota(j), \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i, j \in \mathcal{I}.$$

Das Jacobi-Verfahren entspricht der Wahl $\mathbf{N} := \mathbf{D}^{-1}$, das Gauß-Seidel-Verfahren der Wahl $\mathbf{N} := (\mathbf{D} - \mathbf{E})^{-1}$:

Definition 2.61 (Gauß-Seidel-Iteration) *Seien $\mathbf{D}, \mathbf{E}, \mathbf{F} \in \mathbb{K}^{\mathcal{I} \times \mathcal{I}}$ wie in (2.17) gegeben. Sei $\mathbf{D}$ invertierbar, also jedes Diagonalelement von $\mathbf{A}$ von Null verschieden. Das durch*

$$\Phi_{\mathrm{GS}}(\mathbf{x}, \mathbf{b}) = \mathbf{x} - (\mathbf{D} - \mathbf{E})^{-1}(\mathbf{A}\mathbf{x} - \mathbf{b}) \qquad \textit{für alle } \mathbf{x}, \mathbf{b} \in \mathbb{K}^{\mathcal{I}}$$

gegebene lineare Iterationsverfahren nennen wir die Gauß-Seidel-Iteration.

Offenbar ist auch die Gauß-Seidel-Iteration ein konsistentes Iterationsverfahren, das in der ersten Normalform durch die Matrizen

$$\begin{aligned} \mathbf{M}_{\mathrm{GS}} &:= \mathbf{I} - (\mathbf{D} - \mathbf{E})^{-1}\mathbf{A} = (\mathbf{D} - \mathbf{E})^{-1}\mathbf{F}, \\ \mathbf{N}_{\mathrm{GS}} &:= (\mathbf{D} - \mathbf{E})^{-1} \end{aligned}$$

beschrieben wird.

Bevor wir uns der Analyse der Konvergenz des Verfahrens zuwenden, werfen wir zunächst einen Blick auf seine praktische Implementierung. Nach Definition gilt $\mathbf{A} = \mathbf{D} - \mathbf{E} - \mathbf{F}$, also

$$\begin{aligned} \Phi_{\mathrm{GS}}(\mathbf{x}, \mathbf{b}) &= \mathbf{x} - (\mathbf{D} - \mathbf{E})^{-1}(\mathbf{A}\mathbf{x} - \mathbf{b}) \\ &= \mathbf{x} - (\mathbf{D} - \mathbf{E})^{-1}(\mathbf{D} - \mathbf{E})\mathbf{x} + (\mathbf{D} - \mathbf{E})^{-1}(\mathbf{F}\mathbf{x} + \mathbf{b}) \\ &= (\mathbf{D} - \mathbf{E})^{-1}(\mathbf{F}\mathbf{x} + \mathbf{b}) \qquad \text{für alle } \mathbf{x}, \mathbf{b} \in \mathbb{K}^{\mathcal{I}}. \end{aligned}$$

Die Berechnung von $\mathbf{x}' := \Phi_{\mathrm{GS}}(\mathbf{x}, \mathbf{b})$ zerfällt also in zwei Schritte: Zunächst müssen wir den Vektor $\mathbf{y} := \mathbf{F}\mathbf{x} + \mathbf{b}$ berechnen, der komponentenweise durch

$$y_i = b_i - \sum_{\substack{j \in \mathcal{I} \\ \iota(j) > \iota(i)}} A_{ij} x_j \qquad \text{für alle } i \in \mathcal{I}$$

gegeben ist, dann müssen wir das Gleichungssystem $(\mathbf{D} - \mathbf{E})\mathbf{x}' = \mathbf{y}$ durch Vorwärtseinsetzen lösen, erhalten also

$$x'_i = \frac{1}{A_{ii}} \left(y_i - \sum_{\substack{j \in \mathcal{I} \\ \iota(j) < \iota(i)}} A_{ij} x'_j \right) \qquad \text{für alle } i \in \mathcal{I}.$$

Indem wir beide Gleichungen kombinieren, erhalten wir

$$x'_i = \frac{1}{A_{ii}} \left(b_i - \sum_{\substack{j \in \mathcal{I} \\ \iota(j) > \iota(i)}} A_{ij} x_j - \sum_{\substack{j \in \mathcal{I} \\ \iota(j) < \iota(i)}} A_{ij} x'_j \right) \qquad \text{für alle } i \in \mathcal{I}. \tag{2.18}$$

```
procedure GaussSeidel(n, ι, b, var x);
for k := 1 to n do
  i ← ι^{-1}(k);
  y ← b_i;
  for j ∈ I \ {i} do
    y ← y − A_{ij} x_j
  end for;
  x_i ← y/A_{ii}
end for
```

Abbildung 2.2: Ein Schritt der Gauß-Seidel-Iteration

Mit Hilfe dieser Darstellung lässt sich das Gauß-Seidel-Verfahren besonders einfach implementieren:

Wir nehmen an, dass der Algorithmus zur Durchführung eines Iterationsschritts die alte Iterierte $\mathbf{x}$ mit der neuen $\mathbf{x}'$ überschreiben soll. Wenn wir die Komponenten von $\mathbf{x}'$ in der durch die Numerierung ι gegebenen Reihenfolge $\iota^{-1}(1), \iota^{-1}(2), \ldots, \iota^{-1}(n)$ durchlaufen, enthalten bei der Berechnung von $i \in \mathcal{I}$ die Komponenten x_j mit $\iota(j) > \iota(i)$ noch die Werte der alten Iterierten, während die Komponenten mit $\iota(j) < \iota(i)$ bereits mit den neuen überschrieben wurden. Der resultierende besonders einfache Algorithmus ist in Abbildung 2.2 gegeben. Anders als im Falle des Richardson- und des Jacobi-Verfahrens, bei denen im Allgemeinen ein Hilfsvektor zur Durchführung eines Iterationsschritts erforderlich ist, kann ein Schritt des Gauß-Seidel-Verfahrens ohne Hilfsspeicher durchgeführt werden.

Wenn wir die Gauß-Seidel-Iteration auf das zweidimensionale Modellproblem anwenden wollen, müssen wir zunächst eine passende Numerierung festlegen. Eine übliche Wahl ist die *lexikographische Numerierung*, die durch

$$\iota_{\mathrm{lx}}(i) := i_x + (i_y - 1)N \qquad \text{für alle } i = (i_x, i_y) \in \mathcal{I}$$

gegeben ist: Jede Zeile des Gitters wird von links nach rechts numeriert, und die Zeilen werden von unten nach oben numeriert. Der resultierende Algorithmus ist in Abbildung 2.3 angegeben. Wie wir sehen können, tritt die Numerierung ι nicht explizit im Algorithmus auf, sie ist implizit durch die Organisation der Schleifen gegeben.

Wenden wir uns nun der Analyse des Konvergenzverhaltens zu. Wie im Fall des Jacobi-Verfahrens betrachten wir lediglich den Fall einer positiv definiten Matrix $\mathbf{A}$.

Satz 2.62 (Konvergenz) *Sei Φ ein lineares konsistentes Iterationsverfahren, und seien $\mathbf{M}, \mathbf{N} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ die entsprechenden Matrizen der ersten Normalform.*

Sei $\mathbf{A}$ positiv definit. Falls die Matrix $\mathbf{W} := \mathbf{N}^{-1}$ die Bedingung

$$0 < \mathbf{A} < \mathbf{W} + \mathbf{W}^* \tag{2.19}$$

erfüllt, ist Φ konvergent mit

$$\varrho(\mathbf{M}) \leq \|\mathbf{M}\|_A < 1.$$

```
procedure GSModell2D(N, b, var x);
for i_y := 1 to N do
  for i_x := 1 to N do
    y ← b_{i_x,i_y};
    if i_x > 1 then
      y ← y + h^{-2} x_{i_x-1,i_y}
    end if;
    if i_x < N then
      y ← y + h^{-2} x_{i_x+1,i_y}
    end if;
    if i_y > 1 then
      y ← y + h^{-2} x_{i_x,i_y-1}
    end if;
    if i_y < N then
      y ← y + h^{-2} x_{i_x,i_y+1}
    end if;
    x_{i_x,i_y} ← y/(4h^{-2})
  end for
end for
```

31	32	33	34	35	36
25	26	27	28	29	30
19	20	21	22	23	24
13	14	15	16	17	18
7	8	9	10	11	12
1	2	3	4	5	6

Abbildung 2.3: Durchführung eines Gauß-Seidel-Iterationsschritts für das zweidimensionale Modellproblem mit lexikographischer Numerierung

Beweis. Wir setzen

$$\mathbf{M}_A := \mathbf{A}^{1/2}\mathbf{M}\mathbf{A}^{-1/2} = \mathbf{I} - \mathbf{A}^{1/2}\mathbf{N}\mathbf{A}^{1/2}.$$

Nach Lemma 2.44 gilt

$$\|\mathbf{M}\|_A = \|\mathbf{A}^{1/2}\mathbf{M}\mathbf{A}^{-1/2}\|_2 = \|\mathbf{M}_A\|_2 = \varrho(\mathbf{M}_A^*\mathbf{M}_A)^{1/2},$$

und mit Hilfe von Lemma 2.48 erhalten wir

$$\begin{aligned}
\mathbf{M}_A^*\mathbf{M}_A &= (\mathbf{I} - \mathbf{A}^{1/2}\mathbf{N}^*\mathbf{A}^{1/2})(\mathbf{I} - \mathbf{A}^{1/2}\mathbf{N}\mathbf{A}^{1/2}) \\
&= \mathbf{I} - \mathbf{A}^{1/2}\mathbf{N}^*\mathbf{A}^{1/2} - \mathbf{A}^{1/2}\mathbf{N}\mathbf{A}^{1/2} + \mathbf{A}^{1/2}\mathbf{N}^*\mathbf{A}\mathbf{N}\mathbf{A}^{1/2} \\
&< \mathbf{I} - \mathbf{A}^{1/2}\mathbf{N}^*\mathbf{A}^{1/2} - \mathbf{A}^{1/2}\mathbf{N}\mathbf{A}^{1/2} + \mathbf{A}^{1/2}\mathbf{N}^*(\mathbf{W} + \mathbf{W}^*)\mathbf{N}\mathbf{A}^{1/2} \\
&= \mathbf{I} - \mathbf{A}^{1/2}\mathbf{N}^*\mathbf{A}^{1/2} - \mathbf{A}^{1/2}\mathbf{N}\mathbf{A}^{1/2} + \mathbf{A}^{1/2}\mathbf{N}\mathbf{A}^{1/2} + \mathbf{A}^{1/2}\mathbf{N}^*\mathbf{A}^{1/2} = \mathbf{I}.
\end{aligned}$$

Lemma 2.49 impliziert also $\sigma(\mathbf{M}_A^*\mathbf{M}_A) \subseteq \mathbb{R}_{<1}$. Da $\mathbf{M}_A^*\mathbf{M}_A$ offenbar auch positiv semidefinit ist, folgt $\sigma(\mathbf{M}_A^*\mathbf{M}_A) \subseteq [0,1)$, also $\|\mathbf{M}\|_A = \varrho(\mathbf{M}_A^*\mathbf{M}_A)^{1/2} < 1$. ∎

Nun können wir das allgemeine Kriterium auf den Fall der Gauß-Seidel-Iteration anwenden:

Lemma 2.63 *Sei* $\mathbf{A}$ *positiv definit. Dann ist das Gauß-Seidel-Verfahren konvergent mit* $\varrho(\mathbf{M}_{\mathrm{GS}}) \leq \|\mathbf{M}_{\mathrm{GS}}\|_A < 1$.

Beweis. Nach Definition 2.61 gilt

$$\mathbf{N}_{\mathrm{GS}} = (\mathbf{D} - \mathbf{E})^{-1}, \qquad \mathbf{W}_{\mathrm{GS}} = \mathbf{D} - \mathbf{E}.$$

Da $\mathbf{A}$ selbstadjungiert ist, erhalten wir

$$\mathbf{A} = \mathbf{D} - \mathbf{E} - \mathbf{E}^* < 2\mathbf{D} - \mathbf{E} - \mathbf{E}^* = (\mathbf{D} - \mathbf{E}) + (\mathbf{D} - \mathbf{E})^* = \mathbf{W}_{\mathrm{GS}} + \mathbf{W}_{\mathrm{GS}}^*,$$

so dass wir Satz 2.62 anwenden können. ■

Im Gegensatz zum Jacobi-Verfahren wird das Gauß-Seidel-Verfahren also für *jede* positiv definite Matrix konvergieren, und es ist nicht erforderlich, einen Dämpfungsparameter einzusetzen.

Wie wir gesehen haben, wird eine Komponente x_i' der neuen Iterierten mit der Formel

$$x_i' = \frac{1}{A_{ii}} \left(b_i - \sum_{\substack{j \in \mathcal{I} \\ \iota(j) > \iota(i)}} A_{ij} x_j - \sum_{\substack{j \in \mathcal{I} \\ \iota(j) < \iota(i)}} A_{ij} x_j' \right)$$

berechnet. Indem wir mit A_{ii} multiplizieren und die Summen umordnen, erhalten wir

$$\sum_{\substack{j \in \mathcal{I} \\ \iota(j) > \iota(i)}} A_{ij} x_j + \sum_{\substack{j \in \mathcal{I} \\ \iota(j) \leq \iota(i)}} A_{ij} x_j' = b_i,$$

also wurde x_i' gerade so gewählt, dass das ursprüngliche lineare Gleichungssystem in der i-ten Komponente exakt gelöst wird. Anstatt zu versuchen, dass System (1.1) direkt zu lösen, löst das Gauß-Seidel-Verfahren eine Folge von eindimensionalen Teilproblemen.

Auf ähnlichem Wege lässt sich nachrechnen, dass für das ungedämpfte Jacobi-Verfahren die Komponente x_i' gerade so bestimmt wird, dass

$$\sum_{j \in \mathcal{I} \setminus \{i\}} A_{ij} x_j + A_{ii} x_i' = b_i$$

gilt, auch hier wird das zum Index i gehörende eindimensionale Teilproblem gelöst.

Sowohl im Jacobi- als auch im Gauß-Seidel-Verfahren wird also versucht, einzelne Freiheitsgrade so zu wählen, dass der Fehler lokal minimiert wird. Wenn man das eindimensionale Modellproblem als physikalische Beschreibung einer gespannten Saite interpretiert, approximiert $(\mathbf{Ax})_i$ gerade die Spannung, unter der der zu i gehörenden Gitterpunkt steht. Das Jacobi- und Gauß-Seidel-Verfahren berechnen x_i' gerade so, dass diese Spannung minimiert wird, sie führen also zu einer lokalen Entspannung. Aus diesem Grund bezeichnet man Verfahren wie die Jacobi- und Gauß-Seidel-Iterationen auch als *Relaxationsverfahren.*

2.8 SOR-Iteration

Eine Variante der Relaxationsverfahren ist das Über-Relaxationsverfahren, bei dem die Diagonale der Matrix $\mathbf{A}$ mit einem zusätzlichen Faktor skaliert wird, so dass die von dem Verfahren durchgeführten Korrekturen verstärkt oder abgeschwächt werden können:

Definition 2.64 (SOR-Iteration) *Seien* $\mathbf{D}, \mathbf{E}, \mathbf{F} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ *wie in (2.17) gegeben. Sei* $\mathbf{D}$ *invertierbar, also jedes Diagonalelement von* $\mathbf{A}$ *von Null verschieden, und sei* $\omega \in \mathbb{R}_{>0}$. *Das durch*

$$\Phi_{\mathrm{SOR},\omega}(\mathbf{x}, \mathbf{b}) = \mathbf{x} - (\omega^{-1}\mathbf{D} - \mathbf{E})^{-1}(\mathbf{A}\mathbf{x} - \mathbf{b}) \qquad \textit{für alle } \mathbf{x}, \mathbf{b} \in \mathbb{K}^{\mathcal{I}}$$

gegebene lineare Iterationsverfahren nennen wir die SOR-Iteration.

Die Matrizen der zweiten Normalform sind durch

$$\mathbf{M}_{\mathrm{SOR},\omega} := \mathbf{I} - (\omega^{-1}\mathbf{D} - \mathbf{E})^{-1}\mathbf{A}, \qquad \mathbf{N}_{\mathrm{SOR},\omega} := (\omega^{-1}\mathbf{D} - \mathbf{E})^{-1}$$

gegeben. Das SOR-Verfahren sollte nicht mit dem gedämpften Gauß-Seidel-Verfahren verwechselt werden.

Ähnlich zur Darstellung (2.18) können wir auch für das SOR-Verfahren eine einfache Formel zur Berechnung der einzelnen Komponenten des Lösungsvektors herleiten: Seien $\mathbf{x}, \mathbf{b} \in \mathbb{K}$. Für $\mathbf{x}' := \Phi_{\mathrm{SOR},\omega}(\mathbf{x}, \mathbf{b})$ gilt

$$\begin{aligned}
\mathbf{x}' &= \mathbf{x} - (\omega^{-1}\mathbf{D} - \mathbf{E})^{-1}(\mathbf{A}\mathbf{x} - \mathbf{b}) = \mathbf{x} - (\omega^{-1}\mathbf{D} - \mathbf{E})^{-1}(\mathbf{D}\mathbf{x} - \mathbf{E}\mathbf{x} - \mathbf{F}\mathbf{x} - \mathbf{b}) \\
&= \mathbf{x} - (\omega^{-1}\mathbf{D} - \mathbf{E})^{-1}(\omega^{-1}\mathbf{D} - \mathbf{E})\mathbf{x} + (\omega^{-1}\mathbf{D} - \mathbf{E})^{-1}((\omega^{-1} - 1)\mathbf{D}\mathbf{x} + \mathbf{F}\mathbf{x} + \mathbf{b}) \\
&= (\mathbf{D} - \omega\mathbf{E})^{-1}((1 - \omega)\mathbf{D}\mathbf{x} + \omega\mathbf{F}\mathbf{x} + \omega\mathbf{b}),
\end{aligned}$$

also können wir die Berechnung von $\mathbf{x}'$ wieder in zwei Schritte aufteilen: Zuerst wird $\mathbf{y} := (1-\omega)\mathbf{D}\mathbf{x}+\omega\mathbf{F}\mathbf{x}+\omega\mathbf{b}$ berechnet, dann wird das Gleichungssystem $(\mathbf{D}-\omega\mathbf{E})\mathbf{x}' = \mathbf{y}$ durch Vorwärtseinsetzen gelöst. In Komponentendarstellung erhalten wir

$$y_i = \omega b_i - (\omega - 1)A_{ii}x_i - \omega \sum_{\substack{j\in\mathcal{I} \\ \iota(j)>\iota(i)}} A_{ij}x_j \qquad \text{für alle } i \in \mathcal{I},$$

und Vorwärtseinsetzen in $\mathbf{D} - \omega\mathbf{E}$ führt zu

$$\begin{aligned}
x_i' &= \frac{1}{A_{ii}}\left(y_i - \omega \sum_{\substack{j\in\mathcal{I} \\ \iota(j)<\iota(i)}} A_{ij}x_j'\right) \\
&= \frac{1}{A_{ii}}\left(\omega b_i - (\omega - 1)A_{ii}x_i - \omega \sum_{\substack{j\in\mathcal{I} \\ \iota(j)>\iota(i)}} A_{ij}x_j - \omega \sum_{\substack{j\in\mathcal{I} \\ \iota(j)<\iota(i)}} A_{ij}x_j'\right)
\end{aligned}$$

```
procedure SOR(n, ι, ω, b, var x);
for k := 1 to n do
  i ← ι^{-1}(k);
  y ← b_i;
  for j ∈ I do
    y ← y − A_{ij} x_j
  end for;
  x_i ← x_i + ωy/A_{ii}
end for
```

Abbildung 2.4: Ein Schritt der SOR-Iteration

$$= x_i - \frac{\omega}{A_{ii}} \left(\sum_{\substack{j \in \mathcal{I} \\ \iota(j) \geq \iota(i)}} A_{ij} x_j + \sum_{\substack{j \in \mathcal{I} \\ \iota(j) < \iota(i)}} A_{ij} x'_j - b_i \right) \qquad \text{für alle } i \in \mathcal{I},$$

wir können also die Iterierten des SOR-Verfahrens berechnen, indem wir eine Komponente des Iterationsvektors nach der anderen aktualisieren. Der resultierende Algorithmus findet sich in Abbildung 2.4.

Lemma 2.65 *Sei* $\mathbf{A}$ *positiv definit, und sei* $\omega \in (0, 2)$. *Dann ist das SOR-Verfahren konvergent mit* $\varrho(\mathbf{M}_{\mathrm{SOR},\omega}) \leq \|\mathbf{M}_{\mathrm{SOR},\omega}\|_A < 1$.

Beweis. Nach Definition 2.64 gilt

$$\mathbf{N}_{\mathrm{SOR}} = (\omega^{-1}\mathbf{D} - \mathbf{E})^{-1}, \qquad \mathbf{W}_{\mathrm{SOR}} = \omega^{-1}\mathbf{D} - \mathbf{E}.$$

Da $\omega < 2$ gilt, haben wir $2/\omega > 1$, also

$$\mathbf{A} = \mathbf{D} - \mathbf{E} - \mathbf{E}^* < 2\omega^{-1}\mathbf{D} - \mathbf{E} - \mathbf{E}^* = (\omega^{-1}\mathbf{D} - \mathbf{E}) + (\omega^{-1}\mathbf{D} - \mathbf{E})^* = \mathbf{W}_{\mathrm{SOR},\omega} + \mathbf{W}^*_{\mathrm{SOR},\omega},$$

so dass wir Satz 2.62 anwenden können. ■

In der Praxis kann das SOR-Verfahren bei geeigneter Wahl des Parameters ω *wesentlich* schneller als das eng verwandte Gauß-Seidel-Verfahren konvergieren.

Lemma 2.66 *Sei* $\mathbf{A}$ *positiv definit und* $\omega \in (0, 2)$. *Dann gilt*

$$\|\mathbf{M}_{\mathrm{SOR},\omega}\|_A = \sqrt{1 - \frac{2/\omega - 1}{\|\mathbf{A}^{-1/2}\mathbf{W}_{\mathrm{SOR},\omega}\mathbf{D}^{-1/2}\|_2^2}}$$

mit $\mathbf{W}_{\mathrm{SOR},\omega} = \mathbf{N}^{-1}_{\mathrm{SOR},\omega} = (1/\omega\mathbf{D} - \mathbf{E})$. *Falls*

$$\mathbf{W}_{\mathrm{SOR},\omega}\mathbf{D}^{-1}\mathbf{W}^*_{\mathrm{SOR},\omega} \leq c\mathbf{A} \tag{2.20}$$

für ein $c \in \mathbb{R}_{>0}$ *gilt, folgen*

$$\|\mathbf{A}^{-1/2}\mathbf{W}_{\mathrm{SOR},\omega}\mathbf{D}^{-1/2}\|_2^2 \leq c, \qquad \|\mathbf{M}_{\mathrm{SOR},\omega}\|_A \leq \sqrt{1 - \frac{2/\omega - 1}{c}}.$$

Beweis. Wie im Beweis von Satz 2.62 verwenden wir

$$\mathbf{M}_A := \mathbf{A}^{1/2}\mathbf{M}_{\mathrm{SOR},\omega}\mathbf{A}^{-1/2} = \mathbf{I} - \mathbf{A}^{1/2}\mathbf{N}_{\mathrm{SOR},\omega}\mathbf{A}^{1/2}.$$

Aus der Gleichung

$$\mathbf{A} - \mathbf{W}_{\mathrm{SOR},\omega} - \mathbf{W}_{\mathrm{SOR},\omega}^* = \mathbf{D} - \mathbf{E} - \mathbf{E}^* - \frac{1}{\omega}\mathbf{D} + \mathbf{E} - \frac{1}{\omega}\mathbf{D} + \mathbf{E}^* = \left(1 - \frac{2}{\omega}\right)\mathbf{D}.$$

folgt dann

$$\begin{aligned}
\mathbf{M}_A^*\mathbf{M}_A &= (\mathbf{I} - \mathbf{A}^{1/2}\mathbf{N}_{\mathrm{SOR},\omega}^*\mathbf{A}^{1/2})(\mathbf{I} - \mathbf{A}^{1/2}\mathbf{N}_{\mathrm{SOR},\omega}\mathbf{A}^{1/2}) \\
&= \mathbf{I} - \mathbf{A}^{1/2}(\mathbf{N}_{\mathrm{SOR},\omega}^* + \mathbf{N}_{\mathrm{SOR},\omega})\mathbf{A}^{1/2} + \mathbf{A}^{1/2}\mathbf{N}_{\mathrm{SOR},\omega}^*\mathbf{A}\mathbf{N}_{\mathrm{SOR},\omega}\mathbf{A}^{1/2} \\
&= \mathbf{I} - \mathbf{A}^{1/2}(\mathbf{N}_{\mathrm{SOR},\omega}^*\mathbf{W}_{\mathrm{SOR},\omega}\mathbf{N}_{\mathrm{SOR},\omega} + \mathbf{N}_{\mathrm{SOR},\omega}^*\mathbf{W}_{\mathrm{SOR},\omega}^*\mathbf{N}_{\mathrm{SOR},\omega})\mathbf{A}^{1/2} \\
&\quad + \mathbf{A}^{1/2}\mathbf{N}_{\mathrm{SOR},\omega}^*\mathbf{A}\mathbf{N}_{\mathrm{SOR},\omega}\mathbf{A}^{1/2} \\
&= \mathbf{I} + \mathbf{A}^{1/2}\mathbf{N}_{\mathrm{SOR},\omega}^*(\mathbf{A} - \mathbf{W}_{\mathrm{SOR},\omega} - \mathbf{W}_{\mathrm{SOR},\omega}^*)\mathbf{N}_{\mathrm{SOR},\omega}\mathbf{A}^{1/2} \\
&= \mathbf{I} + \left(1 - \frac{2}{\omega}\right)\mathbf{A}^{1/2}\mathbf{N}_{\mathrm{SOR},\omega}^*\mathbf{D}\mathbf{N}_{\mathrm{SOR},\omega}\mathbf{A}^{1/2} = \mathbf{I} - \left(\frac{2}{\omega} - 1\right)(\mathbf{X}\mathbf{X}^*)^{-1}
\end{aligned}$$

mit der Matrix

$$\mathbf{X} := \mathbf{A}^{-1/2}\mathbf{N}_{\mathrm{SOR},\omega}^{-1}\mathbf{D}^{-1/2} = \mathbf{A}^{-1/2}\mathbf{W}_{\mathrm{SOR},\omega}\mathbf{D}^{-1/2}.$$

Um die gewünschte Abschätzung zu erhalten, müssen wir den kleinsten Eigenwert von $(\mathbf{X}\mathbf{X}^*)^{-1}$ nach unten abschätzen. Dieser Eigenwert ist gerade

$$\frac{1}{\varrho(\mathbf{X}\mathbf{X}^*)} = \frac{1}{\|\mathbf{X}\|_2^2},$$

also erhalten wir

$$\|\mathbf{M}_{\mathrm{SOR},\omega}\|_A^2 = \|\mathbf{M}_A\|_2^2 = \varrho(\mathbf{M}_A^*\mathbf{M}_A) = 1 - \left(\frac{2}{\omega} - 1\right)\frac{1}{\|\mathbf{X}\|_2^2}.$$

Das ist die gesuchte obere Schranke für die Konvergenzrate.

Jetzt bleibt noch die Abschätzung der Norm zu zeigen. Wir haben

$$\mathbf{X}\mathbf{X}^* = \mathbf{A}^{-1/2}\mathbf{W}_{\mathrm{SOR},\omega}\mathbf{D}^{-1}\mathbf{W}_{\mathrm{SOR},\omega}^*\mathbf{A}^{-1/2} \le c\mathbf{A}^{-1/2}\mathbf{A}\mathbf{A}^{-1/2} = c\mathbf{I},$$

also folgt $\sigma(\mathbf{X}\mathbf{X}^*) \subseteq [0, c]$ und damit $\|\mathbf{X}\|_2^2 = \varrho(\mathbf{X}\mathbf{X}^*) \le c$. ■

Wenn wir schon einen Parameter haben, mit dem wir das Konvergenzverhalten eines Verfahrens beeinflussen können, sind wir natürlich daran interessiert, ihn möglichst geschickt zu wählen, um eine möglichst gute Konvergenzrate zu erhalten. Mit Hilfe der expliziten Darstellung der Norm der Iterationsmatrix $\mathbf{M}_{\mathrm{SOR},\omega}$ aus Lemma 2.66 können wir uns diesem Ziel nähern.

Satz 2.67 (Optimierung von ω) *Seien $\gamma, \Gamma \in \mathbb{R}_{>0}$ mit*

$$\mathbf{0} < \gamma\mathbf{D} \leq \mathbf{A}, \qquad \left(\frac{1}{2}\mathbf{D} - \mathbf{E}\right)\mathbf{D}^{-1}\left(\frac{1}{2}\mathbf{D} - \mathbf{E}^*\right) \leq \frac{\Gamma}{4}\mathbf{A}$$

gegeben. Dann erfüllt

$$c = \frac{\Omega^2}{\gamma} + \Omega + \frac{\Gamma}{4}, \qquad \Omega := \frac{2-\omega}{2\omega} = \frac{1}{\omega} - \frac{1}{2}$$

die Abschätzung (2.20) und wir erhalten

$$\|\mathbf{M}_{\mathrm{SOR},\omega}\|_A \leq \sqrt{1 - \frac{2\Omega}{\Omega^2/\gamma + \Omega + \Gamma/4}}. \tag{2.21}$$

Für die Wahl $\omega_{\mathrm{opt}} := 2/(1+\sqrt{\gamma\Gamma})$ nimmt die rechte Seite ihr Minimum an, und zwar

$$\|\mathbf{M}_{\mathrm{SOR},\omega_{\mathrm{opt}}}\|_A \leq \sqrt{\frac{\sqrt{\Gamma} - \sqrt{\gamma}}{\sqrt{\Gamma} + \sqrt{\gamma}}}.$$

Beweis. Wir stellen die Matrix $\mathbf{W}_{\mathrm{SOR},\omega}$ in der Form

$$\mathbf{W}_{\mathrm{SOR},\omega} = \frac{1}{\omega}\mathbf{D} - \mathbf{E} = \left(\frac{1}{\omega} - \frac{1}{2}\right)\mathbf{D} + \frac{1}{2}\mathbf{D} - \mathbf{E} = \Omega\mathbf{D} + \frac{1}{2}\mathbf{D} - \mathbf{E}$$

dar. Für die uns interessierende Matrix folgt dann

$$\begin{aligned}
\mathbf{W}_{\mathrm{SOR},\omega}\mathbf{D}^{-1}\mathbf{W}^*_{\mathrm{SOR},\omega} &= \left(\Omega\mathbf{D} + \left(\frac{1}{2}\mathbf{D} - \mathbf{E}\right)\right)\mathbf{D}^{-1}\left(\Omega\mathbf{D} + \left(\frac{1}{2}\mathbf{D} - \mathbf{E}^*\right)\right) \\
&= \Omega^2\mathbf{D} + \Omega\left(\frac{1}{2}\mathbf{D} - \mathbf{E} + \frac{1}{2}\mathbf{D} - \mathbf{E}^*\right) + \left(\frac{1}{2}\mathbf{D} - \mathbf{E}\right)\mathbf{D}^{-1}\left(\frac{1}{2}\mathbf{D} - \mathbf{E}^*\right) \\
&\leq \frac{\Omega^2}{\gamma}\mathbf{A} + \Omega\mathbf{A} + \frac{\Gamma}{4}\mathbf{A} = \left(\frac{\Omega^2}{\gamma} + \Omega + \frac{\Gamma}{4}\right)\mathbf{A} = c\mathbf{A},
\end{aligned}$$

und das ist die gewünschte Abschätzung.

Zur Wahl des optimalen Wertes für ω haben wir also die Aufgabe, den Term

$$f(\Omega) := \frac{2\Omega}{\Omega^2/\gamma + \Omega + \Gamma/4}$$

möglichst zu maximieren. Dazu suchen wir nach Nullstellen der Ableitung

$$f'(\Omega) = \frac{2(\Omega^2/\gamma + \Omega + \Gamma/4) - 2\Omega(2\Omega/\gamma + 1)}{(\Omega^2/\gamma + \Omega + \Gamma/4)^2}.$$

Wegen

$$2(\Omega^2/\gamma + \Omega + \Gamma/4) - 2\Omega(2\Omega/\gamma + 1) = -2\Omega^2/\gamma + \Gamma/2$$

erhalten wir $\Omega_{\text{opt}} = \sqrt{\gamma\Gamma}/2$ und das Maximum

$$f(\Omega_{\text{opt}}) = \frac{2\sqrt{\gamma\Gamma}}{\Gamma + \sqrt{\gamma\Gamma}}.$$

Durch Einsetzen in die Ungleichung (2.21) erhalten wir die gewünschte Abschätzung, und indem wir die Gleichung $\Omega_{\text{opt}} = (2 - \omega_{\text{opt}})/(2\omega_{\text{opt}})$ nach ω_{opt} auflösen folgt auch $\omega_{\text{opt}} = 2/(1 + \sqrt{\gamma\Gamma})$. ■

Eine überraschende Eigenschaft des SOR-Verfahrens besteht darin, dass es tatsächlich nicht nur schneller als das Gauß-Seidel-Verfahren sein kann, sondern dass es sogar eine bessere Konvergenzordnung erreicht, sofern ω korrekt gewählt wird.

Bemerkung 2.68 (Modellproblem) *Wir wenden Satz 2.67 auf das eindimensionale Modellproblem aus Abschnitt 1.4 an. Dafür müssen wir zunächst γ so wählen, dass*

$$\mathbf{0} < \gamma\mathbf{D} \leq \mathbf{A}$$

gilt. Im Modellproblem gilt $\mathbf{D} = 2h^{-2}\mathbf{I}$, und wir wissen aus Lemma 1.7, dass der kleinste Eigenwert von $\mathbf{A}$ durch $\lambda_{\min} = 4h^{-2}\sin^2(\pi h/2)$ gegeben ist. Also ist für $\gamma \leq 2\sin^2(\pi h/2)$ die Ungleichung erfüllt.

Außerdem müssen wir Γ so wählen, dass

$$\left(\frac{1}{2}\mathbf{D} - \mathbf{E}\right)\mathbf{D}^{-1}\left(\frac{1}{2}\mathbf{D} - \mathbf{E}^*\right) \leq \frac{\Gamma}{4}\mathbf{A}$$

gilt. Da $\mathbf{A}$ eine Tridiagonalmatrix ist, können wir diese Berechnung explizit durchführen: Es gilt

$$\begin{aligned}
\left(\frac{1}{2}\mathbf{D} - \mathbf{E}\right)\mathbf{D}^{-1}\left(\frac{1}{2}\mathbf{D} - \mathbf{E}^*\right) &= \frac{h^{-2}}{2}\begin{pmatrix} 1 & & & \\ -1 & \ddots & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix}\begin{pmatrix} 1 & -1 & & \\ & \ddots & \ddots & \\ & & \ddots & -1 \\ & & & 1 \end{pmatrix} \\
&= \frac{h^{-2}}{2}\begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \leq \frac{1}{2}\mathbf{A},
\end{aligned}$$

also können wir $\Gamma = 2$ wählen. Nach Satz 2.67 ist dann

$$\omega_{\text{opt}} = \frac{2}{1 + \sqrt{4\sin^2(\pi h/2)}} \approx \frac{2}{1 + \sqrt{4\pi^2 h^2/4}} = \frac{2}{1 + \pi h}$$

die beste Wahl des Parameters und führt zu einer Konvergenzrate von

$$\|\mathbf{M}_{\text{SOR},\omega_{opt}}\|_A \leq \sqrt{\frac{\sqrt{2} - \sqrt{2}\sin(\pi h/2)}{\sqrt{2} + \sqrt{2}\sin(\pi h/2)}} = \sqrt{1 - 2\frac{\sin(\pi h/2)}{1 + \sin(\pi h/2)}}$$

$$\approx \sqrt{1 - 2\frac{\pi h}{2}} \approx \sqrt{1 - 2\frac{\pi h}{2} + \left(\frac{\pi h}{2}\right)^2} = 1 - \frac{\pi}{2}h.$$

Bei Verwendung von ω_{opt} erreicht also das SOR-Verfahren eine Konvergenzrate, die nur linear in h gegen 1 strebt, während sie sich im Richardson-, Jacobi- und Gauß-Seidel-Verfahren wie $1 - ch^2$ verhält, wie wir in den ersten beiden Fällen der Formel (1.8) entnehmen können.

Eine kleine Modifikation des Verfahrens, in diesem Fall eine geschickte Skalierung der Diagonalen, kann also zu einer signifikanten Verbesserung der Effizienz führen.

Im allgemeinen Fall ist es schwierig, den Parameter ω korrekt zu wählen, zumindest sofern man von der Möglichkeit absieht, sich ihm durch geschicktes Ausprobieren anzunähern. Für einige wichtige Spezialfälle gibt es Techniken, mit denen sich ω im Zuge der Iteration optimieren lässt, allerdings ist das resultierende Verfahren dann keine lineare Iteration mehr. Unter diesen Umständen sind oft die im nächsten Kapitel diskutierten Kryloff-Verfahren empfehlenswerter.

Bemerkung 2.69 (Parallelisierung) *Das Gauß-Seidel- und erst recht das SOR-Verfahren konvergieren in der Regel wesentlich schneller als das Jacobi- und das Richardson-Verfahren. Dieser Vorteil ist eine Folge der Tatsache, dass im Zuge der beiden erstgenannten Verfahren der Iterationsvektor schrittweise aktualisiert und die einzelnen Korrekturen bereits auf der Grundlage der vorher erfolgten Korrekturen bestimmt werden, während die beiden letztgenannten Verfahren alle Korrekturen unabhängig voneinander durchführen.*

Dieser Vorteil kann sich als Nachteil erweisen, wenn ein Iterationsschritt auf einem Parallelrechner durchgeführt werden soll: Werden etwa die Freiheitsgrade lexikographisch durchlaufen, so basiert die Berechnung von x'_i für $i = (i_x, i_y) \in \mathcal{I}$ auf dem Wert von x'_j für $j = (i_x - 1, i_y)$, also kann x'_i erst berechnet werden, wenn x'_j berechnet wurde. Per Induktion folgt, dass sich weder Gauß-Seidel- noch SOR-Verfahren bei dieser Anordnung effizient parallelisieren lassen, während Jacobi- und Richardson-Verfahren sich perfekt für eine Parallelisierung eignen.

Glücklicherweise lässt sich die Situation verbessern, indem man auf eine andere Numerierung der Freiheitsgrade zurückgreift. Eine mögliche Lösung für das zweidimensionale Modellproblem ist die sogenannte *Schachbrett-Numerierung*: In Anlehnung an ein Schachbrett werden die Freiheitsgrade in „weiße“ und „schwarze“ Felder eingeteilt. Ein Freiheitsgrad $i = (i_x, i_y) \in \mathcal{I}$ ist „weiß“, falls $i_x + i_y$ gerade ist, anderenfalls ist er „schwarz“:

$$\begin{aligned} \mathcal{W} &:= \{i = (i_x, i_y) \in \mathcal{I} \ : \ i_x + i_y \text{ ist gerade}\}, \\ \mathcal{S} &:= \{i = (i_x, i_y) \in \mathcal{I} \ : \ i_x + i_y \text{ ist ungerade}\}. \end{aligned}$$

Die Numerierung ι wird nun so gewählt, dass kleine Zahlen zu Indizes in $\mathcal{W}$ und große Zahlen zu Indizes in $\mathcal{S}$ gehören:

$$\iota(i) \le \lfloor (N^2 + 1)/2 \rfloor \qquad \text{für alle } i \in \mathcal{W},$$

```
procedure SORModell2D(N, b, var x);
for i ∈ W do
   SORUpdate(N, b, i, ω, x)
end for;
for i ∈ S do
   SORUpdate(N, b, i, ω, x)
end for
end;
procedure SORUpdate(N, b, i, ω, var x);
y ← b_{i_x,i_y} − 4h^{−2} x_{i_x,i_y};
if i_x > 1 then
   y ← y + h^{−2} x_{i_x−1,i_y}
end if;
if i_x < N then
   y ← y + h^{−2} x_{i_x+1,i_y}
end if;
if i_y > 1 then
   y ← y + h^{−2} x_{i_x,i_y−1}
end if;
if i_y < N then
   y ← y + h^{−2} x_{i_x,i_y+1}
end if;
x_{i_x,i_y} ← x_{i_x,i_y} + ωy/(4h^{−2})
end;
```

34	16	35	17	36	18
13	31	14	32	15	33
28	10	29	11	30	12
7	25	8	26	9	27
22	4	23	5	24	6
1	19	2	20	3	21

Abbildung 2.5: Durchführung eines SOR-Iterationsschritts für das zweidimensionale Modellproblem mit Schachbrett-Numerierung

$$\iota(i) > \lfloor (N^2+1)/2 \rfloor \qquad \text{für alle } i \in \mathcal{S}.$$

Wenn wir den in Abbildung 2.2 angegebenen Algorithmus auf diese Numerierung anwenden, sehen wir, dass zur Berechnung von x_i' für ein $i = (i_x, i_y) \in \mathcal{I}$ außer i selbst lediglich die Indizes

$$\{(i_x - 1, i_y), (i_x + 1, i_y), (i_x, i_y - 1), (i_x, i_y + 1)\} \cap \mathcal{I}$$

herangezogen werden. Falls $i \in \mathcal{W}$ gilt, sind alle Elemente dieser Menge in $\mathcal{S}$ enthalten, anderenfalls sind sie alle in $\mathcal{W}$ enthalten.

Von dieser Eigenschaft können wir profitieren, indem wir die Berechnung der neuen Iterierten $\mathbf{x}'$ in zwei Phasen aufteilen: Zuerst berechnen wir die Komponenten x_i' für alle „weißen“ Indizes $i \in \mathcal{W}$, dann behandeln wir die verbliebenen „schwarzen“ Indizes $i \in \mathcal{S}$. Da bei der Berechnung der „weißen“ Indizes außer dem Diagonalelement nur „schwarze“ Indizes verwendet werden, sind die Berechnungen innerhalb der beiden Phasen voneinander völlig unabhängig, lassen sich also gut parallelisieren.

Der entsprechende Algorithmus ist in Abbildung 2.4 angegeben. Um zu betonen, dass die Berechnungen der „weißen“ und „schwarzen“ Phase unabhängig sind, verwenden wir die Notation $i \in \mathcal{W}$ und $i \in \mathcal{S}$ in den beiden zentralen Schleifen. Die eigentliche Berechung einer neuen Komponente x_i' ist in eine separate Prozedur ausgelagert, um den Algorithmus kompakt zu halten.

2.9 Kaczmarz-Iteration

Zum Abschluss dieses Kapitels wenden wir uns einem Iterationsverfahren zu, das gegenüber den bisher erwähnten Methoden den Vorteil bietet, für jede reguläre Matrix zu konvergieren, insbesondere auch für die indefiniten Probleme, die den bisherigen Verfahren Schwierigkeiten bereiten konnten.

Lemma 2.63 besagt, dass das Gauß-Seidel-Verfahren für positiv definite Matrizen immer konvergiert. Falls $\mathbf{A} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ regulär ist, gilt dasselbe für die Adjungierte $\mathbf{A}^*$, also ist die Matrix $\widehat{\mathbf{A}} := \mathbf{A}\mathbf{A}^*$ wegen

$$\langle \mathbf{x}, \widehat{\mathbf{A}}\mathbf{x}\rangle_2 = \langle \mathbf{A}^*\mathbf{x}, \mathbf{A}^*\mathbf{x}\rangle_2 = \langle \mathbf{y}, \mathbf{y}\rangle_2 > 0 \qquad \text{für alle } \mathbf{x} \in \mathbb{K} \setminus \{\mathbf{0}\}, \mathbf{y} := \mathbf{A}^*\mathbf{x} \neq \mathbf{0}$$

positiv definit. Wir untersuchen nun die Gleichung

$$\widehat{\mathbf{A}}\widehat{\mathbf{x}} = \mathbf{b}$$

und stellen fest, dass die Lösung $\widehat{\mathbf{x}} \in \mathbb{K}^{\mathcal{I}}$ dieses Systems wegen

$$\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{A}^*\widehat{\mathbf{x}} = \widehat{\mathbf{A}}\widehat{\mathbf{x}} = \mathbf{b}$$

zu einer Lösung $\mathbf{x} := \mathbf{A}^*\widehat{\mathbf{x}}$ des ursprünglichen Gleichungssystems (1.1) führt.

Also können wir die Gauß-Seidel-Iteration auf das modifizierte positiv definite Gleichungssystem anwenden und aus der Iterierten $\widehat{\mathbf{x}}^{(m)}$ eine Folge von Iterierten $\mathbf{x}^{(m)} := \mathbf{A}^*\widehat{\mathbf{x}}^{(m)}$ für das ursprüngliche System (1.1) gewinnen.

Um eine praktisch brauchbare Darstellung des so definierten *Kaczmarz-Verfahrens* zu gewinnen, nehmen wir zur Vereinfachung an, dass $\mathcal{I} = \{1, \ldots, n\}$ gilt. Die Formel (2.18) ist für $\widehat{\mathbf{x}}$ und $\widehat{\mathbf{A}}$ äquivalent zu

$$\widehat{x}_i' = \widehat{x}_i - \frac{1}{\widehat{A}_{ii}} \left(\sum_{j=i}^{n} \widehat{A}_{ij}\widehat{x}_j + \sum_{j=1}^{i-1} \widehat{A}_{ij}\widehat{x}_j' - b_i \right).$$

Wir führen für jedes $i \in \mathcal{I}$ den durch

$$\widehat{x}_j^{(i)} := \begin{cases} \widehat{x}_j & \text{falls } j \geq i, \\ \widehat{x}_j' & \text{sonst} \end{cases} \qquad \text{für alle } j \in \mathcal{I}$$

gegebenen Hilfsvektor $\widehat{\mathbf{x}}^{(i)} \in \mathbb{K}^{\mathcal{I}}$ ein, mit dem sich die Gleichung in der Form

$$\widehat{x}_i' = \widehat{x}_i - \frac{1}{\widehat{A}_{ii}} \left(\sum_{j=i}^{n} \widehat{A}_{ij}\widehat{x}_j + \sum_{j=1}^{i-1} \widehat{A}_{ij}\widehat{x}_j' - b_i \right) = \widehat{x}_i - \frac{1}{\widehat{A}_{ii}}((\widehat{\mathbf{A}}\widehat{\mathbf{x}}^{(i)})_i - b_i)$$

```
procedure Kaczmarz(n, ι, b, var x);
for k := 1 to n do
   i ← ι^{-1}(k);
   KaczmarzUpdate(n, b, i, x);
end for
end;
procedure KaczmarzUpdate(n, b, i, var x);
y ← b_i;
a ← 0;
for j ∈ I do
   y ← y − A_{ij} x_j;
   a ← a + |A_{ij}|^2
end for;
y ← y/a;
for j ∈ I do
   x_j ← x_j + Ā_{ij} y
end for
end
```

Abbildung 2.6: Ein Schritt der Kaczmarz-Iteration

darstellen lässt. Mit dem i-ten kanonischen Einheitsvektor $\mathbf{e}^{(i)} \in \mathbb{K}^{\mathcal{I}}$ gilt also

$$\widehat{\mathbf{x}}^{(i+1)} = \widehat{\mathbf{x}}^{(i)} - \mathbf{e}^{(i)} \frac{1}{\widehat{A}_{ii}} ((\widehat{\mathbf{A}}\widehat{\mathbf{x}}^{(i)})_i - b_i).$$

Unser eigentliches Interesse gilt nicht $\widehat{\mathbf{x}}$, sondern $\mathbf{x} = \mathbf{A}^*\widehat{\mathbf{x}}$, also definieren wir $\mathbf{a}^{(i)} := \mathbf{A}^*\mathbf{e}^{(i)}$ und $\mathbf{x}^{(i)} := \mathbf{A}^*\widehat{\mathbf{x}}^{(i)}$ und erhalten

$$\begin{aligned}\mathbf{x}^{(i+1)} = \mathbf{A}^*\widehat{\mathbf{x}}^{(i+1)} &= \mathbf{A}^*\widehat{\mathbf{x}}^{(i)} - \mathbf{A}^*\mathbf{e}^{(i)} \frac{1}{\widehat{A}_{ii}} \left((\mathbf{A}\mathbf{A}^*\widehat{\mathbf{x}}^{(i)})_i - b_i \right) \\ &= \mathbf{x}^{(i)} - \mathbf{a}^{(i)} \frac{1}{\|\mathbf{a}^{(i)}\|_2^2} \left((\mathbf{A}\mathbf{x}^{(i)})_i - b_i \right).\end{aligned}$$

Nach Definition gelten $\mathbf{x}^{(1)} = \mathbf{x}$ und $\mathbf{x}^{(n+1)} = \mathbf{x}'$, wir können also die neue Iterierte berechnen, indem wir der Reihe nach die Zwischenergebnisse $\mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n+1)}$ bestimmen.

Indem wir $\mathbf{b} = \mathbf{A}\mathbf{x}^*$ ausnutzen, erhalten wir für einen Einzelschritt die Darstellung

$$\begin{aligned}\mathbf{x}^{(i+1)} &= \mathbf{x}^{(i)} - \mathbf{a}^{(i)} \frac{1}{\|\mathbf{a}^{(i)}\|_2^2} \left(\mathbf{A}(\mathbf{x}^{(i)} - \mathbf{x}^*) \right)_i = \mathbf{x}^{(i)} - \mathbf{a}^{(i)} \frac{1}{\|\mathbf{a}^{(i)}\|_2^2} \langle \mathbf{e}^{(i)}, \mathbf{A}(\mathbf{x}^{(i)} - \mathbf{x}^*) \rangle_2 \\ &= \mathbf{x}^{(i)} - \mathbf{a}^{(i)} \frac{1}{\|\mathbf{a}^{(i)}\|_2^2} \langle \mathbf{A}^*\mathbf{e}^{(i)}, (\mathbf{x}^{(i)} - \mathbf{x}^*) \rangle_2 = \mathbf{x}^{(i)} - \mathbf{a}^{(i)} \frac{1}{\|\mathbf{a}^{(i)}\|_2^2} \langle \mathbf{a}^{(i)}, \mathbf{x}^{(i)} - \mathbf{x}^* \rangle_2,\end{aligned}$$

jeder Einzelschritt berechnet also die orthogonale Projektion des Fehlers $\mathbf{x}^{(i)} - \mathbf{x}^*$ auf den von $\mathbf{a}^{(i)}$ aufgespannten eindimensionalen Teilraum und subtrahiert sie von der aktuellen

Iterierten. Dadurch wird dafür gesorgt, dass der Fehler der neuen Iterierten senkrecht auf $\mathbf{a}^{(i)}$ steht, also optimal bezüglich dieses eindimensionalen Unterraums ist.

Da $\mathbf{a}^{(i)}$ gemäß

$$a_j^{(i)} = (\mathbf{A}^*\mathbf{e}^{(i)})_j = \sum_{k\in\mathcal{I}} \bar{A}_{kj}\delta_{ik} = \bar{A}_{ij}$$

gerade der (gegebenenfalls konjugierten) i-ten Zeile von $\mathbf{A}$ entspricht, lassen sich die einzelnen Berechnungsschritte sehr effizient durchführen, falls, wie im Modellproblem, die meisten Einträge einer Zeile gleich Null sind.

Der resultierende Algorithmus ist in Abbildung 2.6 zusammengefasst. Da bei gängigen Speicherformaten für schwachbesetzte Matrizen der Zugriff auf eine Zeile der Matrix wesentlich effizienter als der Zugriff auf eine Spalte ist, ist der Algorithmus hier so formuliert, dass die Matrixeinträge ausschließlich zeilenweise durchlaufen werden.

Die Hauptprozedur „Kaczmarz“ enthält lediglich eine Schleife über alle Freiheitsgrade, die eigentliche Arbeit leistet die Unterprozedur „KaczmarzUpdate“, die zunächst die Werte $y = b_i - (\mathbf{A}\mathbf{x})_i$ und $a = \|\mathbf{a}^{(i)}\|_2^2$ berechnet und dann den Vektor $\mathbf{x}$ aktualisiert.

3 Semiiterative und Krylow-Raum-Verfahren

Bisher haben wir uns lediglich für das gerade aktuelle Element der Folgen der Iterierten interessiert, also im m-ten Schritt eines Iterationsverfahrens für $\mathbf{x}^{(m)}$. Wir werden jetzt Verfahren untersuchen, bei denen *alle* Iterierten $\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(m)}$ in die Berechnung einer verbesserten Approximation $\mathbf{y}^{(m)}$ der Lösung einfließen können. Erfreulicherweise stellt sich heraus, dass die so konstruierten *semiiterativen* Verfahren ähnlich effizient wie die bisher untersuchten iterativen Verfahren implementiert werden können, aber zum Teil deutlich bessere Konvergenzeigenschaften besitzen.

3.1 Allgemeine lineare semiiterative Verfahren

Die Idee eines semiiterativen Verfahrens besteht darin, aus den im m-ten Schritt eines iterativen Verfahrens zur Verfügung stehenden Iterierten $\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(m)}$ einen Vektor $\mathbf{y}^{(m)} \in \mathbb{K}^{\mathcal{I}}$ zu bestimmen, der möglichst nahe an der Lösung des Gleichungssystems (1.1) liegt.

Definition 3.1 (Semiiteration) *Eine Abbildung*

$$\Sigma : \left(\bigcup_{m \in \mathbb{N}_0} (\mathbb{K}^{\mathcal{I}})^{m+1} \right) \to \mathbb{K}^{\mathcal{I}}$$

bezeichnen wir als Semiiteration *oder als* semiiteratives Verfahren*.*

Für eine Folge $(\mathbf{x}^{(m)})_{m \in \mathbb{N}_0}$ von Iterierten eines der bisher eingeführten Verfahren definiert eine Semiiteration Σ die durch

$$\mathbf{y}^{(m)} := \Sigma(\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(m)}) \qquad \text{für alle } m \in \mathbb{N}_0 \tag{3.1}$$

gegebene Folge $(\mathbf{y}^{(m)})_{m \in \mathbb{N}_0}$ von „Semiiterierten“.

Ein lineares Verfahren haben wir als konsistent bezeichnet, falls die Lösung des linearen Gleichungssystems ein Fixpunkt des Verfahrens ist, die entsprechende Folge der Iterierten also konstant ist. Natürlich erwarten wir von einem sinnvollen semiiterativen Verfahren, dass es in diesem Fall ebenfalls die Lösung berechnet.

Definition 3.2 (Konsistenz) *Ein Semiiterationsverfahren Σ heißt* konsistent*, falls*

$$\Sigma(\underbrace{\mathbf{x}, \ldots, \mathbf{x}}_{\in (\mathbb{K}^{\mathcal{I}})^{m+1}}) = \mathbf{x} \qquad \textit{für alle } \mathbf{x} \in \mathbb{K}^{\mathcal{I}}, m \in \mathbb{N}_0$$

erfüllt ist.

Wie bereits im vorangehenden Kapitel sind wir vor allem an *linearen* semiiterativen Verfahren interessiert:

Definition 3.3 (Lineare Semiiteration) *Eine Semiiteration Σ heißt* linear*, falls es für jedes $m \in \mathbb{N}_0$ Koeffizienten $\theta_0^{(m)}, \ldots, \theta_m^{(m)} \in \mathbb{K}$ so gibt, dass*

$$\Sigma(\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(m)}) = \theta_0^{(m)}\mathbf{x}^{(0)} + \ldots + \theta_m^{(m)}\mathbf{x}^{(m)} \qquad \textit{für alle } \mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(m)} \in \mathbb{K}^{\mathcal{I}} \textit{ gilt.}$$

Im Falle eines linearen Iterationsverfahrens konnten wir die Konsistenz des Verfahrens mit Hilfe einer einfachen Formel (vgl. Lemma 2.9) charakterisieren. Für lineare Semiiterationen ist es ebenfalls möglich, die Konsistenz einfach zu beschreiben:

Lemma 3.4 *Sei Σ eine durch die Koeffizienten $(\theta_i^{(m)})_{m\in\mathbb{N}_0, i\le m}$ beschriebene lineare Semiiteration. Σ ist genau dann konsistent, wenn für alle $m \in \mathbb{N}_0$ die Gleichung $\theta_0^{(m)} + \ldots + \theta_m^{(m)} = 1$ gilt.*

Beweis. Sei zunächst Σ konsistent, und seien $\mathbf{x} \in \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\}$ und $m \in \mathbb{N}_0$. Dann gilt

$$\mathbf{x} = \Sigma(\underbrace{\mathbf{x}, \ldots, \mathbf{x}}_{\in(\mathbb{K}^{\mathcal{I}})^{m+1}}) = \theta_0^{(m)}\mathbf{x} + \ldots + \theta_m^{(m)}\mathbf{x} = (\theta_0^{(m)} + \ldots + \theta_m^{(m)})\mathbf{x}.$$

Wegen $\mathbf{x} \neq \mathbf{0}$ folgt daraus $\theta_0^{(m)} + \ldots + \theta_m^{(m)} = 1$. Der Rest des Beweises ist trivial. ∎

Um herauszufinden, welche Semiiterationen besonders günstig sind, benötigen wir eine Darstellung des Fehlers. Bei linearen Iterationsverfahren lässt er sich nach Lemma 2.11 in der Form

$$\mathbf{x}^{(m)} - \mathbf{x}^* = \mathbf{M}^m(\mathbf{x}^{(0)} - \mathbf{x}^*)$$

ausdrücken, so dass für die Konvergenzanalyse das Verhalten den Potenzen $\mathbf{M}^m$ zu untersuchen ist. Bei linearen Semiiterationen werden die Potenzen durch allgemeinere Polynome ersetzt. Wir bezeichnen den Raum der Polynome höchstens m-ten Grades mit

$$\Pi_m := \left\{ \xi \to \sum_{\ell=0}^{m} \alpha_\ell \xi^\ell \ : \ \alpha_0, \ldots, \alpha_m \in \mathbb{K} \right\}$$

und können in ein $p \in \Pi_m$ mit

$$p(\xi) = \alpha_0 + \alpha_1\xi + \ldots + \alpha_m\xi^m = \sum_{\ell=0}^{m} \alpha_\ell \xi^\ell$$

eine quadratische Matrix $\mathbf{X} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ einsetzen, um die Matrix

$$p(\mathbf{X}) = \alpha_0\mathbf{I} + \alpha_1\mathbf{X} + \ldots + \alpha_m\mathbf{X}^m = \sum_{\ell=0}^{m} \alpha_\ell \mathbf{X}^\ell$$

zu erhalten. Dabei verwenden wir die Konvention $\mathbf{X}^0 = \mathbf{I}$.

Der Einfachheit halber nehmen wir an, dass die Semiiteration auf einem konvergenten Iterationsverfahren aufsetzt:

Lemma 3.5 *Sei Φ ein lineares Iterationsverfahren, gegeben durch die Matrizen $\mathbf{M}, \mathbf{N} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ der ersten Normalform. Sei Σ eine konsistente lineare Semiiteration, gegeben durch die Koeffizienten $(\theta_i^{(m)})_{m\in\mathbb{N}_0, i\le m}$.*

Sei $\mathbf{b} \in \mathbb{K}^{\mathcal{I}}$, und sei $\mathbf{x}^* \in \mathbb{K}^{\mathcal{I}}$ ein Fixpunkt von Φ zu $\mathbf{b}$. Sei $(\mathbf{x}^{(m)})_{m\in\mathbb{N}_0}$ eine Folge von Iterierten von Φ zu einem Startvektor $\mathbf{x}^{(0)}$. Sei $(\mathbf{y}^{(m)})_{m\in\mathbb{N}_0}$ die durch (3.1) gegebene Folge von Semiiterierten und $\mathbf{y}^* := \mathbf{x}^*$. Dann gilt

$$\mathbf{y}^{(m)} - \mathbf{y}^* = p_m(\mathbf{M})\left(\mathbf{y}^{(0)} - \mathbf{y}^*\right) \qquad \textit{für alle } m \in \mathbb{N}_0$$

mit den durch

$$p_m(\xi) := \sum_{\ell=0}^{m} \theta_\ell^{(m)} \xi^\ell \qquad \textit{für alle } m \in \mathbb{N}_0, \xi \in \mathbb{K}$$

gegebenen Polynomen p_m.

Beweis. Da Φ linear ist, gilt

$$\begin{aligned}\mathbf{x}^{(m+1)} - \mathbf{x}^* &= \Phi(\mathbf{x}^{(m)}, \mathbf{b}) - \Phi(\mathbf{x}^*, \mathbf{b}) \\ &= \mathbf{M}\mathbf{x}^{(m)} + \mathbf{N}\mathbf{b} - \mathbf{M}\mathbf{x}^* - \mathbf{N}\mathbf{b} = \mathbf{M}(\mathbf{x}^{(m)} - \mathbf{x}^*) \qquad \text{für alle } m \in \mathbb{N}_0,\end{aligned}$$

und wir erhalten

$$\mathbf{x}^{(m)} - \mathbf{x}^* = \mathbf{M}^m(\mathbf{x}^{(0)} - \mathbf{x}^*) \qquad \text{für alle } m \in \mathbb{N}_0.$$

Sei $m \in \mathbb{N}_0$. Da Σ konsistent ist, folgt aus Lemma 3.4 die Gleichung

$$\begin{aligned}\mathbf{y}^{(m)} - \mathbf{y}^* &= \sum_{\ell=0}^{m} \theta_\ell^{(m)} \mathbf{x}^{(\ell)} - \mathbf{x}^* = \sum_{\ell=0}^{m} \theta_\ell^{(m)} \mathbf{x}^{(\ell)} - \sum_{\ell=0}^{m} \theta_\ell^{(m)} \mathbf{x}^* \\ &= \sum_{\ell=0}^{m} \theta_\ell^{(m)} \left(\mathbf{x}^{(\ell)} - \mathbf{x}^*\right) = \sum_{\ell=0}^{m} \theta_\ell^{(m)} \mathbf{M}^\ell \left(\mathbf{x}^{(0)} - \mathbf{x}^*\right) \\ &= \sum_{\ell=0}^{m} \theta_\ell^{(m)} \mathbf{M}^\ell \left(\mathbf{y}^{(0)} - \mathbf{y}^*\right) = p_m(\mathbf{M})\left(\mathbf{y}^{(0)} - \mathbf{y}^*\right),\end{aligned}$$

die zu beweisen war. ∎

Offenbar ist eine lineare Semiiteration genau dann konsistent, wenn

$$p_m(1) = \sum_{\ell=0}^{m} \theta_\ell^{(m)} = 1 \qquad \text{für alle } m \in \mathbb{N}_0$$

gilt, wenn also 1 ein Fixpunkt aller Polynome p_m ist.

Für die triviale Wahl $p_m(\xi) = \xi^m$, also $\theta_0^{(m)} = \ldots = \theta_{m-1}^{(m)} = 0$ und $\theta_m^{(m)} = 1$, erhalten wir eine konsistente Semiiteration, bei der die Semiiterierten gerade den Iterierten der

zugrundeliegenden linearen Iteration entsprechen, wir können also erwarten, dass eine gut konstruierte Semiiteration nicht schlechter als das zugrundeliegende Verfahren ist.

Wenden wir uns nun der Frage zu, wie wir die Polynome p_m wählen müssen, um den Fehler möglichst schnell zu minimieren. Ideal wäre es, wenn wir die Koeffizienten $\theta_0^{(m)}, \ldots, \theta_m^{(m)}$ so wählen können, dass die Norm

$$\|\mathbf{y}^{(m)} - \mathbf{y}^*\| = \|p_m(\mathbf{M})(\mathbf{x}^{(0)} - \mathbf{x}^*)\|$$

minimiert wird. In diesem Fall müssten die Koeffizienten vom Startvektor $\mathbf{x}^{(0)}$ und von der rechten Seite $\mathbf{b}$ abhängig gewählt werden, wodurch die Analyse verkompliziert wird.

Einfacher ist es, die induzierte Matrixnorm zu minimieren: Wegen Lemma 2.13 gilt

$$\|\mathbf{y}^{(m)} - \mathbf{x}^*\| = \|p_m(\mathbf{M})(\mathbf{x}^{(0)} - \mathbf{x}^*)\| \leq \|p_m(\mathbf{M})\| \, \|\mathbf{x}^{(0)} - \mathbf{x}^*\|, \tag{3.2}$$

also könnten wir nach Polynomen suchen, die die Norm $\|p_m(\mathbf{M})\|$ minimieren. Der Satz von Cayley-Hamilton legt nahe, dass eine gut konstruierte lineare Semiiteration spätestens nach $m = n = \#\mathcal{I}$ Schritten die exakte Lösung $\mathbf{y}^{(m)} = \mathbf{x}^*$ berechnen sollte, denn das charakteristische Polynom p_M erfüllt $p_M(\mathbf{M}) = \mathbf{0}$ und hat die Ordnung n.

3.2 Tschebyscheff-Semiiteration

Im Allgemeinen lassen sich Polynome von Matrizen nur schwer handhaben, also würden wir uns gerne auf Eigenwerte beschränken. Wir wissen aus dem letzten Kapitel bereits, wie sich dieses Ziel erreichen lässt: Wir nehmen an, dass $\mathbf{A}$ und $\mathbf{N}$ positiv definit sind und betrachten die Energienorm des Fehlers. Dank Lemma 2.51 gilt

$$\begin{aligned}
\|p_m(\mathbf{M})\|_A = \|\mathbf{A}^{1/2} p_m(\mathbf{M}) \mathbf{A}^{-1/2}\|_2 &= \left\| \sum_{\ell=0}^{m} \theta_\ell^{(m)} \mathbf{A}^{1/2} (\mathbf{I} - \mathbf{N}\mathbf{A})^\ell \mathbf{A}^{-1/2} \right\|_2 \\
&= \left\| \sum_{\ell=0}^{m} \theta_\ell^{(m)} (\mathbf{I} - \mathbf{A}^{1/2}\mathbf{N}\mathbf{A}^{1/2})^\ell \right\|_2 \\
&= \varrho(p_m(\mathbf{I} - \mathbf{A}^{1/2}\mathbf{N}\mathbf{A}^{1/2})) = \max\{|p_m(\lambda)| \ : \ \lambda \in \sigma(\mathbf{I} - \mathbf{A}^{1/2}\mathbf{N}\mathbf{A}^{1/2})\} \\
&= \max\{|p_m(\lambda)| \ : \ \lambda \in \sigma(\mathbf{I} - \mathbf{N}\mathbf{A})\} = \max\{|p_m(\lambda)| \ : \ \lambda \in \sigma(\mathbf{M})\},
\end{aligned} \tag{3.3}$$

also brauchen wir „nur“ ein Polynom p_m zu finden, dass einerseits die Konsistenzbedingung $p_m(1) = 1$ erfüllt und andererseits auf dem Spektrum der Iterationsmatrix $\mathbf{M}$ möglichst kleine Werte annimmt.

Aus unseren Annahmen folgt bereits, dass das Spektrum von $\mathbf{M}$ eine Teilmenge von $\mathbb{R}$ ist, und da es diskret ist, gibt es ein Intervall $[a, b] \subseteq \mathbb{R}$ derart, dass $\sigma(\mathbf{M}) \subseteq [a, b]$ gilt. Wenn wir ein Polynom p_m finden können, das auf $[a, b]$ besonders kleine Werte annimmt, wird es auch auf $\sigma(\mathbf{M})$ besonders kleine Werte annehmen, also werden auch $\|p_m(\mathbf{M})\|_A$ und der Fehler der m-ten Semiiterierten klein sein.

Falls Φ ein konvergentes Verfahren ist, gilt $\sigma(\mathbf{M}) \subseteq (-1, 1)$, wir können also $[a, b] \subseteq (-1, 1)$ wählen und müssen ein Polynom p_m mit Grad $\leq m$ finden, das $p_m(1) = 1$

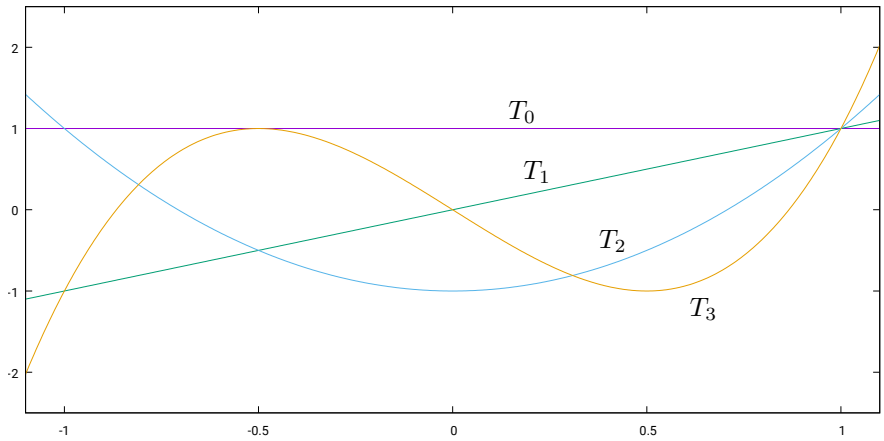

Abbildung 3.1: Tschebyscheff-Polynome $T_0,\ldots,T_3$

erfüllt und auf $[a,b] \subseteq (-1,1)$ möglichst kleine Werte annimmt. Diese Aufgabe lösen die *Tschebyscheff-Polynome*:

Definition 3.6 (Tschebyscheff-Polynome) *Die* Tschebyscheff-Polynome T_m *sind für alle* $m \in \mathbb{N}_0$ *durch die Rekursion*

$$\begin{aligned} T_0(\xi) &= 1, \\ T_1(\xi) &= \xi, \\ T_m(\xi) &= 2\xi T_{m-1}(\xi) - T_{m-2}(\xi) \qquad && \text{für alle } \xi \in \mathbb{K} \end{aligned}$$

definiert. Das Tschebyscheff-Polynom T_m *hat den Grad* m.

Diese Darstellung der Tschebyscheff-Polynome ist gut für die Implementierung geeignet, für Beweise sind die folgenden alternativen Darstellungen nützlicher:

Lemma 3.7 (Alternative Darstellungen) *Für alle* $x \in [-1,1]$ *und alle* $m \in \mathbb{N}_0$ *gilt*

$$c_m(x) = \cos(m \arccos x). \tag{3.4a}$$

Für alle $x \in \mathbb{R} \setminus [-1,1]$ *und alle* $m \in \mathbb{N}_0$ *gilt*

$$c_m(x) = \frac{1}{2}\left(\left(x + \sqrt{x^2-1}\right)^m + \left(x + \sqrt{x^2-1}\right)^{-m}\right). \tag{3.4b}$$

Beweis. Wir verwenden die *Kutta-Schukowski-Transformation*

$$s\colon \mathbb{C} \setminus \{0\} \to \mathbb{C}, \qquad z \mapsto \frac{z + 1/z}{2}.$$

Sie ist surjektiv, denn für jedes $x \in \mathbb{C}$ gilt

$$s(z) = x \iff \frac{z+1/z}{2} = x \iff z^2 - 2xz + 1 = 0,$$

und diese quadratische Gleichung ist in dem Körper $\mathbb{C}$ der komplexen Zahlen immer lösbar. Falls z eine Lösung ist, gilt dasselbe aus Symmetriegründen auch für $1/z$.

Wir definieren für alle $m \in \mathbb{N}_0$ die Funktionen

$$f_m \colon \mathbb{C} \setminus \{0\} \to \mathbb{C}, \qquad z \mapsto \frac{z^m + z^{-m}}{2}$$

und werden die Gleichungen

$$c_m(x) = f_m(z) \qquad \text{für alle } m \in \mathbb{N}_0,\ z \in \mathbb{C} \setminus \{0\},\ x := s(z) \tag{3.5}$$

per abschnittsweiser Induktion beweisen.

Induktionsanfang: Für $m = 0$ gilt $f_m(z) = 1 = c_m(x)$ für alle $z \in \mathbb{C}\setminus\{0\}$ und alle $x \in \mathbb{C}$. Für $m = 1$ haben wir $f_m = s$, also folgt aus $x = s(z)$ bereits $c_1(x) = x = s(z) = f_m(z)$ für alle $z \in \mathbb{C} \setminus \{0\}$.

Induktionsvoraussetzung: Sei $m \in \mathbb{N}_{\geq 1}$ so gegeben, dass die Gleichung $c_n(x) = f_n(z)$ für alle $n \in [0 : m]$, $z \in \mathbb{C} \setminus \{0\}$, $x = s(z)$ gilt.

Induktionsschritt: Sei $z \in \mathbb{C} \setminus \{0\}$, sei $x := s(z)$. Es gilt

$$\begin{aligned} f_{m+1}(z) &= \frac{z^{m+1} + \frac{1}{z^{m+1}}}{2} = \frac{(z + \frac{1}{z})z^m - z^{m-1} + (z + \frac{1}{z})\frac{1}{z^m} - \frac{1}{z^{m-1}}}{2} \\ &= \left(z + \frac{1}{z}\right)\frac{z^m + \frac{1}{z^m}}{2} - \frac{z^{m-1} + \frac{1}{z^{m-1}}}{2} = 2s(z)\, f_m(z) - f_{m-1}(z) \\ &= 2x\, c_m(x) - c_{m-1}(x) = c_{m+1}(x). \end{aligned}$$

Um (3.4a) zu beweisen, wählen wir $x \in [-1, 1]$ und setzen $\xi := \arccos(x) \in [0, \pi]$ sowie $z := e^{i\xi}$. Dann gilt mit der Eulerschen Formel

$$s(z) = \frac{z + 1/z}{2} = \frac{e^{i\xi} + e^{-i\xi}}{2} = \frac{e^{i\xi} + \overline{e^{i\xi}}}{2} = \operatorname{Re}(e^{i\xi}) = \cos(\xi) = x.$$

Mit (3.5) und der Eulerschen Formel folgt

$$\begin{aligned} c_m(x) = f_m(z) &= \frac{z^m + z^{-m}}{2} = \frac{e^{im\xi} + e^{-im\xi}}{2} \\ &= \frac{e^{im\xi} + \overline{e^{im\xi}}}{2} = \operatorname{Re}(e^{im\xi}) = \cos(m\xi) = \cos(m \arccos(x)). \end{aligned}$$

Um (3.4b) zu zeigen, wählen wir $x \in \mathbb{R} \setminus [-1, 1]$ und setzen $z := x + \sqrt{x^2 - 1}$. Dann gilt

$$\frac{1}{z} = \frac{1}{x + \sqrt{x^2 - 1}} = \frac{x - \sqrt{x^2 - 1}}{(x + \sqrt{x^2 - 1})(x - \sqrt{x^2 - 1})} = \frac{x - \sqrt{x^2 - 1}}{x^2 - (x^2 - 1)} = x - \sqrt{x^2 - 1},$$

und es folgt

$$s(z) = \frac{z + 1/z}{2} = \frac{x + \sqrt{x^2 - 1} + x - \sqrt{x^2 - 1}}{2} = \frac{2x}{2} = x.$$

Indem wir in (3.5) einsetzen, folgt direkt (3.4b). ■

Bevor wir diese Polynome zur Lösung unserer Minimierungsaufgabe verwenden können, müssen wir einige ihrer wichtigsten Eigenschaften zusammentragen:

Lemma 3.8 *Sei $m \in \mathbb{N}_0$ und $\xi_0 \in \mathbb{R}_{>1}$. Sei q ein Polynom mit Grad $\leq m$, das*

$$\sup\{|q(\xi)| \ : \ \xi \in [-1,1]\} \leq \sup\{|T_m(\xi)| \ : \ \xi \in [-1,1]\}, \qquad q(\xi_0) = T_m(\xi_0) \tag{3.6}$$

erfüllt. Dann gilt $q = T_m$, also sind die Tschebyscheff-Polynome in diesem Sinne die Lösung einer Minimierungsaufgabe.

Beweis. Sei q ein Polynom mit Grad $\leq m$, dass (3.6) erfüllt. Wir definieren die Punkte $(\zeta_\nu)_{\nu=0}^m$ durch

$$\zeta_\nu := \cos(\pi - \pi\nu/m) \qquad \text{für alle } \nu \in \{0,\dots,m\}.$$

Da der Cosinus auf dem Intervall $[0,\pi]$ streng monoton fällt, folgt $-1 = \zeta_0 < \zeta_1 < \ldots < \zeta_m = 1$. Wegen (3.4a) gelten die Gleichungen

$$\begin{aligned} T_m(\zeta_\nu) &= T_m(\cos(\pi - \pi\nu/m)) = \cos(m\pi - m\pi\nu/m) \\ &= \cos(\pi(m-\nu)) = (-1)^{m-\nu} \qquad \text{für alle } \nu \in \{0,\dots,m\}. \end{aligned}$$

Gemäß unsere Voraussetzung muss $q(\zeta_\nu) \leq 1 = T_m(\zeta_\nu)$ für gerades $m-\nu$ und $q(\zeta_\nu) \geq -1 = T_m(\zeta_\nu)$ für ungerades $m-\nu$ gelten, also folgt für $r := T_m - q$ gerade

$$\begin{aligned} r(\zeta_\nu) &\geq 0 \qquad \text{für alle } \nu \in \{0,\dots,m\} \text{ mit geradem } m-\nu, \\ r(\zeta_\nu) &\leq 0 \qquad \text{für alle } \nu \in \{0,\dots,m\} \text{ mit ungeradem } m-\nu. \end{aligned}$$

Wir wollen beweisen, dass r mindestens m Nullstellen, nach Vielfachheit gezählt, besitzt.

Für jedes $\nu \in \{1,\dots,m\}$ existiert nach dem Zwischenwertsatz mindestens eine Nullstelle des Polynoms r in dem Intervall $[\zeta_{\nu-1},\zeta_\nu]$. Wir wählen eine Nullstelle $\xi_\nu \in [\zeta_{\nu-1},\zeta]$, wobei wir Nullstellen im Inneren gegenüber Randpunkten vorziehen, es soll also $\xi_\nu \in \{\zeta_{\nu-1},\zeta_\nu\}$ nur gelten, falls r keine Nullstelle im offenen Intervall $(\zeta_{\nu-1},\zeta_\nu)$ besitzt.

Nun haben wir m Nullstellen $\xi_1 \leq \xi_2 \leq \ldots \leq \xi_m$ gefunden, allerdings müssen sie nicht alle verschieden sein. In diesem Fall kommt die Vielfachheit ins Spiel: Falls für ein $\nu \in \{1,\dots,m-1\}$ die Gleichung $\xi_\nu = \xi_{\nu+1}$ gilt, folgt $\xi_\nu = \xi_{\nu+1} = \zeta_\nu$ und das Polynom r besitzt nach Definition weder auf $(\zeta_{\nu-1},\zeta_\nu)$ noch auf $(\zeta_\nu,\zeta_{\nu+1})$ Nullstellen. Die Vorzeichen von $r(\zeta_{\nu-1})$ und $r(\zeta_{\nu+1})$ sind gleich, also muss r ein lokales Minimum oder Maximum in $\zeta_\nu = \xi_\nu$ besitzen. Damit folgt $r'(\xi_\nu) = 0$, ξ_ν ist also eine doppelte Nullstelle des Polynoms r.

Wenn wir die Nullstellen nun inklusive ihrer Vielfachheit zählen, haben wir m Nullstellen in $[-1,1]$ gefunden.

Nach unserer Voraussetzung ist $\xi_0 \in \mathbb{R}_{>1}$ eine weitere Nullstelle, also ist r ein Polynom des Grades m mit $m+1$ Nullstellen. Nach dem Identitätssatz für Polynome folgt $r = 0$, also $q = T_m$. ∎

Wir sind nicht an einem Polynom interessiert, dass auf $[-1,1]$ minimal ist, unser Interesse gilt dem Intervall $[a,b]$, also müssen wir das Tschebyscheff-Polynom T_m geeignet transformieren: Die Abbildung

$$\xi \mapsto \frac{2\xi - a - b}{b - a}$$

bildet a auf -1 und b auf 1 ab, muss also infolge ihrer Linearität das Intervall $[a, b]$ bijektiv auf $[-1, 1]$ abbilden. Also können wir das gewünschte Polynom durch

$$p_m(\xi) := \frac{1}{C_m} T_m\left(\frac{2\xi - a - b}{b - a}\right) \qquad \text{für alle } m \in \mathbb{N}_0, \xi \in \mathbb{K} \tag{3.7}$$

definieren. Die Konstante C_m benutzen wir, um sicherzustellen, dass die Konsistenzbedingung $p_m(1) = 1$ gilt:

$$C_m := T_m\left(\frac{2 - a - b}{b - a}\right) \qquad \text{für alle } m \in \mathbb{N}_0.$$

Nach Lemma 3.8 gilt $|T_m(\xi)| \leq 1$ für $\xi \in [-1, 1]$, also folgt $|p_m(\xi)| \leq 1/C_m$ für $\xi \in [a, b]$.

Wenn q ein weiteres Polynom vom Grad $\leq m$ mit $q(1) = 1$ ist, können wir durch Rücktransformation das Polynom

$$\hat{q}(\xi) := C_m q\left(\frac{(b - a)\xi + (a + b)}{2}\right)$$

mit Grad $\leq m$ definieren und erhalten

$$\hat{q}(\xi_0) = \hat{q}\left(\frac{2 - a - b}{b - a}\right) = C_m q(1) = C_m = C_m p(1) = \frac{C_m}{C_m} T_m(\xi_0) = T_m(\xi_0)$$

für den Punkt

$$\xi_0 := \frac{2 - a - b}{b - a} = \frac{b - a + 2 - 2b}{b - a} = 1 + 2\frac{1 - b}{b - a} > 1, \tag{3.8}$$

also impliziert die Minimalitätseigenschaft (3.6) bereits

$$\sup\{\hat{q}(\xi) \ : \ \xi \in [-1, 1]\} \geq \sup\{T_m(\xi) \ : \ \xi \in [-1, 1]\},$$

also auch

$$\sup\{q(\xi) \ : \ \xi \in [a, b]\} \geq \sup\{p_m(\xi) \ : \ \xi \in [a, b]\},$$

somit ist p_m die optimale Wahl und $1/C_m$ die bestmögliche Schranke auf dem für uns relevanten Intervall $[a, b]$.

Um abschätzen zu können, wie sich die Schranke $1/C_m$ verhält, verwenden wir eine weitere alternative Darstellung von T_m:

Lemma 3.9 *Seien $a, b \in \mathbb{R}$ mit $a < b < 1$. Es gilt*

$$\frac{1}{C_m} = \frac{2c^m}{1 + c^{2m}} \leq 2c^m \qquad \textit{für alle } m \in \mathbb{N}_0$$

mit den Konstanten

$$c := \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}, \qquad \kappa := \frac{1 - a}{1 - b} \in \mathbb{R}_{>0}. \tag{3.9}$$

Beweis. Wir setzen ξ_0 wie in (3.8). Es gilt

$$\begin{aligned}
\xi_0^2 - 1 &= \left(1 + 2\frac{1-b}{b-a}\right)^2 - 1 = 1 + 4\frac{1-b}{b-a} + 4\frac{(1-b)^2}{(b-a)^2} - 1 \\
&= 4\frac{(1-b)(b-a) + (1-b)^2}{(b-a)^2} = 4\frac{(1-b)(1-a)}{(b-a)^2} > 0, \\
\xi_0 + \sqrt{\xi_0^2 - 1} &= \frac{(1-a)+(1-b)}{b-a} + \frac{2\sqrt{1-a}\sqrt{1-b}}{b-a} = \frac{(\sqrt{1-a}+\sqrt{1-b})^2}{b-a} \\
&= \frac{1-a}{b-a}\left(1 + \sqrt{\frac{1-b}{1-a}}\right)^2 = \frac{1-a}{(1-a)-(1-b)}\left(1 + \frac{1}{\sqrt{\kappa}}\right)^2 \\
&= \left(1 - \frac{1-b}{1-a}\right)^{-1}\left(1 + \frac{1}{\sqrt{\kappa}}\right)^2 = \left(1 - \frac{1}{\kappa}\right)^{-1}\left(1 + \frac{1}{\sqrt{\kappa}}\right)^2 \\
&= \frac{(1+1/\sqrt{\kappa})^2}{(1+1/\sqrt{\kappa})(1-1/\sqrt{\kappa})} = \frac{1+1/\sqrt{\kappa}}{1-1/\sqrt{\kappa}} = \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1} = \frac{1}{c},
\end{aligned}$$

also erhalten wir mit (3.4b)

$$\frac{1}{C_m} = \frac{1}{T_m(\xi_0)} = \frac{2}{c^{-m} + c^m} = \frac{2c^m}{1 + c^{2m}} \le 2c^m,$$

und das ist die zu beweisende Abschätzung. ■

Natürlich ist unser Ansatz nur dann effizient, wenn sich auch die Semiiterierten $\mathbf{y}^{(m)}$ effizient berechnen lassen.

Sei Φ ein lineares Iterationsverfahren. Sei $\mathbf{b} \in \mathbb{K}^{\mathcal{I}}$ eine rechte Seite, $\mathbf{x}^{(0)} \in \mathbb{K}^{\mathcal{I}}$ ein Startvektor, und $(\mathbf{x}^{(m)})_{m\in\mathbb{N}_0}$ die zugehörige Folge von Iterierten.

Für jedes Polynom p mit der Darstellung

$$p(\xi) = \sum_{\ell=0}^{m} \theta_\ell \xi^\ell \qquad \text{für alle } \xi \in \mathbb{K}$$

definieren wir die entsprechende Semiiterierte durch

$$\mathbf{y}^p := \sum_{\ell=0}^{m} \theta_\ell \mathbf{x}^{(\ell)}.$$

Für diese Semiiterierten gelten Rechenregeln, die es uns ermöglichen werden, die rekursive Definition 3.6 auf die Konstruktion eines effizienten semiiterativen Verfahrens zu übertragen:

Lemma 3.10 *Seien p und q Polynome, und sei $\alpha \in \mathbb{K}$. Dann gilt*

$$\mathbf{y}^{p+\alpha q} = \mathbf{y}^p + \alpha \mathbf{y}^q.$$

Falls die Gleichungen

$$q(1) = 1, \qquad p(\xi) = \xi q(\xi) \qquad \textit{für alle } \xi \in \mathbb{K}$$

erfüllt ist, gilt

$$\mathbf{y}^p = \Phi(\mathbf{y}^q, \mathbf{b}).$$

Beweis. Sei m der maximale Grad von p und q. Dann gibt es Koeffizienten $(\theta_\ell)_{\ell=0}^m$ und $(\omega_\ell)_{\ell=0}^m$ mit

$$p(\xi) = \sum_{\ell=0}^m \theta_\ell \xi^\ell, \qquad q(\xi) = \sum_{\ell=0}^m \omega_\ell \xi^\ell \qquad \text{für alle } \xi \in \mathbb{K},$$

und wir erhalten

$$(p + \alpha q)(\xi) = \sum_{\ell=0}^m (\theta_\ell + \alpha\omega_\ell)\xi^\ell \qquad \text{für alle } \xi \in \mathbb{K},$$

also auch

$$\mathbf{y}^{p+\alpha q} = \sum_{\ell=0}^m (\theta_\ell + \alpha\omega_\ell)\mathbf{x}^{(\ell)} = \sum_{\ell=0}^m \theta_\ell \mathbf{x}^{(\ell)} + \alpha \sum_{\ell=0}^m \theta_\ell \mathbf{x}^{(\ell)} = \mathbf{y}^p + \alpha \mathbf{y}^q.$$

Damit ist die erste Aussage bewiesen.

Wir setzen nun $q(1) = 1$ und $p(\xi) = \xi q(\xi)$ für alle $\xi \in \mathbb{K}$ voraus. Es folgt

$$p(\xi) = \xi q(\xi) = \xi \sum_{\ell=0}^m \omega_\ell \xi^\ell = \sum_{\ell=0}^m \omega_\ell \xi^{\ell+1} = \sum_{\ell=1}^{m+1} \omega_{\ell-1} \xi^\ell,$$

also per Koeffizientenvergleich

$$\theta_\ell = \begin{cases} 0 & \text{falls } \ell = 0, \\ \omega_{\ell-1} & \text{ansonsten} \end{cases} \qquad \text{für alle } \ell \in \{0, \dots, m\}.$$

Wäre $\omega_m \neq 0$, so hätte p den Grad $m+1$ im Gegensatz zu unserer Voraussetzung. Also muss $\omega_m = 0$ gelten.

Definition 2.2 und die Linearität von Φ ergeben

$$\begin{aligned}
\mathbf{y}^p &= \sum_{\ell=0}^m \theta_\ell \mathbf{x}^{(\ell)} = \sum_{\ell=1}^m \omega_{\ell-1} \mathbf{x}^{(\ell)} = \sum_{\ell=1}^m \omega_{\ell-1} \Phi(\mathbf{x}^{(\ell-1)}, \mathbf{b}) = \sum_{\ell=0}^{m-1} \omega_\ell \Phi(\mathbf{x}^{(\ell)}, \mathbf{b}) \\
&= \sum_{\ell=0}^m \omega_\ell \Phi(\mathbf{x}^{(\ell)}, \mathbf{b}) = \sum_{\ell=0}^m \omega_\ell (\mathbf{M}\mathbf{x}^{(\ell)} + \mathbf{N}\mathbf{b}) = \sum_{\ell=0}^m \omega_\ell \mathbf{M}\mathbf{x}^{(\ell)} + \sum_{\ell=0}^m \omega_\ell \mathbf{N}\mathbf{b} \\
&= \mathbf{M}\left(\sum_{\ell=0}^m \omega_\ell \mathbf{x}^{(\ell)}\right) + \left(\sum_{\ell=0}^m \omega_\ell\right)\mathbf{N}\mathbf{b} = \mathbf{M}\mathbf{y}^q + \mathbf{N}\mathbf{b} = \Phi(\mathbf{y}^q, \mathbf{b}),
\end{aligned}$$

wobei wir im vorletzten Schritt $\sum_{\ell=0}^{m} \omega_\ell = q(1) = 1$ ausgenutzt haben. ■

Nun wenden wir diese Rechenregeln auf die transformierten Tschebyscheff-Polynome p_m an. Nach Definition 3.6 gilt

$$C_0 = 1, \quad C_1 = \frac{2-a-b}{b-a}, \quad C_m = 2\frac{2-a-b}{b-a}C_{m-1} - C_{m-2} \qquad \text{für alle } m \in \mathbb{N}_{\geq 2},$$

und gemäß der Formel (3.7) erhalten wir

$$\begin{aligned}
p_0(\xi) &= \frac{1}{C_0} T_0\left(\frac{2\xi-a-b}{b-a}\right) = 1,\\
p_1(\xi) &= \frac{1}{C_1} T_1\left(\frac{2\xi-a-b}{b-a}\right) = \frac{1}{C_1}\frac{2\xi-a-b}{b-a} = \frac{2}{C_1(b-a)}\xi - \frac{b+a}{C_1(b-a)},\\
p_m(\xi) &= \frac{1}{C_m} T_m\left(\frac{2\xi-a-b}{b-a}\right)\\
&= \frac{1}{C_m}\left(2\frac{2\xi-a-b}{b-a}T_{m-1}\left(\frac{2\xi-a-b}{b-a}\right) - T_{m-2}\left(\frac{2\xi-a-b}{b-a}\right)\right)\\
&= \frac{2C_{m-1}}{C_m}\frac{2\xi-a-b}{b-a}p_{m-1}(\xi) - \frac{C_{m-2}}{C_m}p_{m-2}(\xi)\\
&= \frac{4C_{m-1}}{C_m(b-a)}\xi p_{m-1}(\xi) - \frac{2C_{m-1}(b+a)}{C_m(b-a)}p_{m-1}(\xi) - \frac{C_{m-2}}{C_m}p_{m-2}(\xi)\\
&\qquad \text{für alle } \xi \in \mathbb{K}, m \in \mathbb{N}_{\geq 2}.
\end{aligned}$$

Mit Hilfe von Lemma 3.10 können nun die Semiiterierten gemäß der Formeln

$$\mathbf{y}^{(0)} = \mathbf{x}^{(0)}, \tag{3.10a}$$

$$\mathbf{y}^{(1)} = \frac{2}{C_1(b-a)}\Phi(\mathbf{y}^{(0)}, \mathbf{b}) - \frac{b+a}{C_1(b-a)}\mathbf{y}^{(0)}, \tag{3.10b}$$

$$\mathbf{y}^{(m)} = \frac{4C_{m-1}}{C_m(b-a)}\Phi(\mathbf{y}^{(m-1)}, \mathbf{b}) - \frac{2C_{m-1}(b+a)}{C_m(b-a)}\mathbf{y}^{(m-1)} - \frac{C_{m-2}}{C_m}\mathbf{y}^{(m-2)} \tag{3.10c}$$

für alle $m \in \mathbb{N}_{\geq 2}$ berechnet werden.

Damit ist unser Ziel erreicht: Die neue Semiiterierte $\mathbf{y}^{(m)}$ kann aus den beiden unmittelbar vorangehenden Semiiterierten $\mathbf{y}^{(m-1)}$ und $\mathbf{y}^{(m-2)}$ berechnet werden, es ist nicht erforderlich, alle Iterierten $\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(m)} \in \mathbb{K}^{\mathcal{I}}$ abzuspeichern.

Definition 3.11 (Tschebyscheff-Semiiteration) *Sei Φ ein lineares Iterationsverfahren mit der Iterationsmatrix $\mathbf{M} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$, und seien $a, b \in \mathbb{R}$ mit $a < b < 1$ so gegeben, dass $\sigma(\mathbf{M}) \subseteq [a, b]$ gilt.*

Dann definiert (3.10) eine konsistente lineare Semiiteration, die die Tschebyscheff-Semiiteration *genannt wird.*

Indem wir Lemma 3.9 mit der in Lemma 3.5 bereits bewiesenen Schranke

$$\|\mathbf{y}^{(m)} - \mathbf{y}^*\|_A \leq \|p_m(\mathbf{M})\|_A \|\mathbf{y}^{(0)} - \mathbf{y}^*\|_A \leq \max\{p_m(\lambda) \ : \ \lambda \in [a, b]\} \, \|\mathbf{y}^{(0)} - \mathbf{y}^*\|_A$$

$$= \frac{1}{C_m} \|\mathbf{y}^{(0)} - \mathbf{y}^*\|_A$$

für den Fehler der Semiiterierten kombinieren, erhalten wir die folgende Abschätzung für den Iterationsfehler.

Satz 3.12 (Konvergenz) *Sei Φ ein lineares Iterationsverfahren, gegeben durch die Matrizen $\mathbf{M}, \mathbf{N} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ der ersten Normalform. Seien $\mathbf{A}$ und $\mathbf{N}$ positiv definit, und seien $a, b \in \mathbb{R}$ mit $a < b < 1$ und $\sigma(\mathbf{M}) \subseteq [a, b]$. Dann gilt*

$$\|\mathbf{y}^{(m)} - \mathbf{y}^*\|_A \leq \frac{2c^m}{1 + c^{2m}} \|\mathbf{y}^{(0)} - \mathbf{y}^*\|_A \qquad \text{für alle } m \in \mathbb{N}_0$$

mit der in (3.9) definierten Konstanten c.

Beweis. Wir kombinieren Lemma 3.9 mit (3.2) und (3.3). ∎

Wir wenden diese Abschätzung nun auf das optimal gedämpfte Richardson-Verfahren an: Sei $\mathbf{A}$ positiv definit, und seien $\alpha, \beta \in \mathbb{R}_{>0}$ der minimale und maximale Eigenwert. Nach Bemerkung 2.32 ist das Richardson-Verfahren konvergent für $\theta_{\text{opt}} := 2/(\beta + \alpha)$ und erfüllt

$$\varrho(\mathbf{M}_{\text{Rich},\theta_{\text{opt}}}) \leq \frac{\beta - \alpha}{\beta + \alpha},$$

also gilt $\sigma(\mathbf{M}_{\text{Rich},\theta_{\text{opt}}}) \subseteq [a, b]$ für

$$a := \frac{\alpha - \beta}{\beta + \alpha} = 1 - 2\frac{\beta}{\beta + \alpha} < 1, \qquad b := \frac{\beta - \alpha}{\beta + \alpha} = 1 - 2\frac{\alpha}{\beta + \alpha} < 1.$$

Wenn wir p_m für dieses Intervall konstruieren, erhalten wir aus Lemma 3.9 die Abschätzung

$$\|p_m(\mathbf{M}_{\text{Rich},\theta_{\text{opt}}})\|_A = \frac{1}{C_m} = \frac{2c^m}{1 + c^{2m}} \leq 2c^m$$

mit

$$\kappa := \frac{1 - a}{1 - b} = \frac{(\beta + \alpha) - (\alpha - \beta)}{\beta + \alpha} \frac{\beta + \alpha}{(\beta + \alpha) + (\alpha - \beta)} = \frac{\beta}{\alpha},$$
$$c := \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} = \frac{\sqrt{\beta/\alpha} - 1}{\sqrt{\beta/\alpha} + 1} = \frac{\sqrt{\beta} - \sqrt{\alpha}}{\sqrt{\beta} + \sqrt{\alpha}},$$

minimaler und maximaler Eigenwert gehen also bei der Abschätzung des Iterationsfehlers nur noch über die Wurzel ein.

Für das Modellproblem bei großer Problemdimension, also kleinem Gitterparameter h, bedeutet das, dass der Fehler in einem Schritt ungefähr um den Faktor

$$\frac{\sqrt{\beta} - \sqrt{\alpha}}{\sqrt{\beta} + \sqrt{\alpha}} = \frac{2h^{-1}\cos(\pi h/2) - 2h^{-1}\sin(\pi h/2))}{2h^{-1}\cos(\pi h/2) + 2h^{-1}\sin(\pi h/2))} \approx \frac{2h^{-1} - \pi}{2h^{-1} + \pi} = 1 - \frac{2\pi}{2h^{-1} + \pi} \approx 1 - \pi h$$

procedure Tschebyscheff(a, b, Φ, $\mathbf{b}$, **var** $\mathbf{x}$);
$\xi_0 \leftarrow \frac{2-a-b}{b-a}$;
$C_1 \leftarrow 1$; $\quad \mathbf{y}' \leftarrow \mathbf{x}$;
$C_0 \leftarrow \xi_0$;
$\mathbf{x} \leftarrow \Phi(\mathbf{y}', \mathbf{b})$;
$\mathbf{x} \leftarrow \frac{2}{C_0(b-a)}\mathbf{x} - \frac{a+b}{C_0(b-a)}\mathbf{y}'$;
while Fehler zu groß **do**
  $C_2 \leftarrow C_1$; $\quad \mathbf{y}'' \leftarrow \mathbf{y}'$;
  $C_1 \leftarrow C_0$; $\quad \mathbf{y}' \leftarrow \mathbf{x}$;
  $C_0 \leftarrow 2\xi_0 C_1 - C_2$;
  $\mathbf{x} \leftarrow \Phi(\mathbf{y}', \mathbf{b})$;
  $\mathbf{x} \leftarrow \frac{4C_1}{C_0(b-a)}\mathbf{x} - \frac{2C_1(a+b)}{C_0(b-a)}\mathbf{y}' - \frac{C_2}{C_0}\mathbf{y}''$
end while

Abbildung 3.2: Tschebyscheff-Semiiteration

reduziert wird, während im ursprünglichen Verfahren nur der Faktor

$$\frac{\beta-\alpha}{\beta+\alpha} = \frac{4h^{-2}\cos^2(\pi h/2) - 4h^{-2}\sin^2(\pi h/2)}{4h^{-2}\cos^2(\pi h/2) + 4h^{-2}\sin^2(\pi h/2)} \approx \frac{4h^{-2}-\pi^2}{4h^{-2}+\pi^2} = 1 - \frac{2\pi^2}{4h^{-2}+\pi^2} \approx 1 - \pi^2 h^2/2$$

zu erwarten ist. Für große Probleme kann das Tschebyscheff-Verfahren also die Richardson-Iteration wesentlich beschleunigen.

Ein Beispiel für eine sinnvolle Implementierung des Verfahrens ist in Abbildung 3.2 angegeben: Die Koeffizienten C_{m-2}, C_{m-1} und C_m werden in den Variablen C_2, C_1 und C_0 gespeichert, die Vektoren $\mathbf{y}^{(m-2)}$, $\mathbf{y}^{(m-1)}$ und $\mathbf{y}^{(m)}$ in $\mathbf{y}''$, $\mathbf{y}'$ und $\mathbf{x}$. Eine elegantere Implementierung würde die Koeffizienten und die Vektoren nicht in jedem Schritt kopieren, sondern zyklisch überschreiben: Die Semiiterierte $\mathbf{y}^{(3)}$ würde das nicht mehr benötigte $\mathbf{y}^{(0)}$ ersetzen, $\mathbf{y}^{(4)}$ dann $\mathbf{y}^{(1)}$ und so weiter.

Für einen Durchlauf der Schleife werden lediglich ein Iterationsschritt, zwei Kopiervorgänge und die Berechnung einer Linearkombination von drei Vektoren benötigt, also ist die Tschebyscheff-Semiiteration in der hier diskutierten Form effizient durchführbar.

Bei Verfahren in der zweiten Normalform wird in jedem Schleifendurchgang der Defekt $\mathbf{d} := \mathbf{A}\mathbf{x} - \mathbf{b}$ berechnet. Es bietet sich an, $\|\mathbf{d}\|_2$ oder $\|\mathbf{N}\mathbf{d}\|_2$ als Grundlage für das Abbruchkriterium der Schleife zu verwenden: Falls $\|\mathbf{d}\|_2$ klein ist, dürfte auch $\|\mathbf{x}-\mathbf{x}^*\|_2 = \|\mathbf{A}^{-1}\mathbf{d}\|_2 \leq \|\mathbf{A}^{-1}\|_2\|\mathbf{d}\|_2$ klein sein. Für den präkonditionierten Defekt $\mathbf{N}\mathbf{d}$ lässt sich eine ähnliche Abschätzung beweisen.

Bemerkung 3.13 *Um Fehler bei der Implementierung des Tschebyscheff-Verfahrens, insbesondere bei der Berechnung der Koeffizienten der Linearkombinationen, zu vermeiden, ist es nützlich, zu wissen, dass diese Koeffizienten jeweils die Summe* 1 *ergeben: Für* $m = 0$ *ist die Aussage trivial, für* $m = 1$ *gilt*

$$\frac{1}{C_1}\frac{2}{b-a} - \frac{1}{C_1}\frac{a+b}{b-a} = \frac{1}{C_1}\frac{2-a-b}{b-a} = 1,$$

und für $m \geq 2$ haben wir

$$\frac{4C_{m-1}}{b-a} - \frac{C_{m-1}2(a+b)}{b-a} - C_{m-2} = 2C_{m-1}\frac{2-a-b}{b-a} - C_{m-2} = C_m,$$

also auch

$$\frac{C_{m-1}}{C_m}\frac{4}{b-a} - \frac{C_{m-1}}{C_m}\frac{2(b+a)}{b-a} - \frac{C_{m-2}}{C_m} = 1.$$

Bemerkung 3.14 (Spektralschranken) *Seien $\mathbf{A}, \mathbf{N} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ positiv definit. Falls die Matrix $\mathbf{W} := \mathbf{N}^{-1}$ die aus Satz 2.52 bekannte Bedingung*

$$\alpha\mathbf{W} \leq \mathbf{A} \leq \beta\mathbf{W}$$

mit $\beta \geq \alpha > 0$ erfüllt, gilt für die zugehörige Iterationsmatrix $\mathbf{M} = \mathbf{I} - \mathbf{NA}$ die Inklusion

$$\sigma(\mathbf{M}) \subseteq [1-\beta, 1-\alpha],$$

denn die Matrizen $\mathbf{M}_W := \mathbf{W}^{1/2}\mathbf{M}\mathbf{W}^{-1/2} = \mathbf{I} - \mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2}$ und $\mathbf{M}$ sind ähnlich, und aus der Voraussetzung folgt mit Lemma 2.48

$$(1-\beta)\mathbf{I} \leq \mathbf{M}_W \leq (1-\alpha)\mathbf{I},$$

so dass wir Lemma 2.49 anwenden können.

3.3 Gradientenverfahren

Das Tschebyscheff-Verfahren führt zwar zu einer Verbesserung der Konvergenz, erfordert aber die Kenntnis guter Schranken für das Spektrum der Iterationsmatrix $\mathbf{M}$. Bei allgemeinen Anwendungen kann es sich als schwierig erweisen, diese Schranken herzuleiten.

Deshalb werden wir nun unsere Aufmerksamkeit auf Verfahren richten, bei der keine Parameter a priori gewählt werden müssen.

Wir untersuchen zuerst das *Gradientenverfahren*, das eng mit der Richardson-Iteration verwandt ist. Die Grundidee dieses Verfahrens besteht darin, das Lösen des Gleichungssystems (1.1) als Minimierungsproblem zu formulieren, um neue Techniken zum Einsatz bringen zu können.

Sei $\mathbf{A}$ positiv definit, sei $\mathbf{b} \in \mathbb{K}^{\mathcal{I}}$ und $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$. Offenbar ist $\mathbf{x} \in \mathbb{K}^{\mathcal{I}}$ genau dann eine Lösung des Gleichungssystems $\mathbf{Ax} = \mathbf{b}$, wenn $\mathbf{x} = \mathbf{x}^*$ gilt. Das ist äquivalent dazu, dass die Energienorm gleich null ist, dass also

$$\begin{aligned}
\|\mathbf{x}^* - \mathbf{x}\|_A^2 &= \langle(\mathbf{x}^* - \mathbf{x}), \mathbf{A}(\mathbf{x}^* - \mathbf{x})\rangle_2 \\
&= \langle\mathbf{x}^*, \mathbf{Ax}^*\rangle_2 - \langle\mathbf{x}, \mathbf{Ax}^*\rangle_2 - \langle\mathbf{x}^*, \mathbf{Ax}\rangle_2 + \langle\mathbf{x}, \mathbf{Ax}\rangle_2 \\
&= \langle\mathbf{x}^*, \mathbf{Ax}^*\rangle_2 - \langle\mathbf{x}, \mathbf{Ax}^*\rangle_2 - \langle\mathbf{Ax}^*, \mathbf{x}\rangle_2 + \langle\mathbf{x}, \mathbf{Ax}\rangle_2 \\
&= \langle\mathbf{x}^*, \mathbf{Ax}^*\rangle_2 - \langle\mathbf{x}, \mathbf{b}\rangle_2 - \langle\mathbf{b}, \mathbf{x}\rangle_2 + \langle\mathbf{x}, \mathbf{Ax}\rangle_2 \\
&= \langle\mathbf{x}^*, \mathbf{Ax}^*\rangle_2 - \overline{\langle\mathbf{b}, \mathbf{x}\rangle_2} - \langle\mathbf{b}, \mathbf{x}\rangle_2 + \langle\mathbf{x}, \mathbf{Ax}\rangle_2
\end{aligned}$$

$$= \langle \mathbf{x}^*, \mathbf{A}\mathbf{x}^* \rangle_2 - 2\,\mathrm{Re}\langle \mathbf{b}, \mathbf{x} \rangle_2 + \langle \mathbf{x}, \mathbf{A}\mathbf{x} \rangle_2$$

gilt. Die letzten beiden Terme können wir berechnen, ohne $\mathbf{x}^*$ zu kennen, während der erste Term von $\mathbf{x}$ völlig unabhängig ist. Statt direkt die Energienorm zu minimieren, können wir demnach auch die quadratische Funktion

$$f\colon \mathbb{K}^{\mathcal{I}} \to \mathbb{R}, \qquad \mathbf{x} \mapsto \frac{1}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x} \rangle_2 - \mathrm{Re}\langle \mathbf{b}, \mathbf{x} \rangle_2,$$

behandeln, die nach den vorangehenden Betrachtungen die Gleichung

$$\|\mathbf{x}^* - \mathbf{x}\|_A^2 = 2f(\mathbf{x}) + \langle \mathbf{x}^*, \mathbf{A}\mathbf{x}^* \rangle_2 \tag{3.11}$$

erfüllt. Man beachte, dass aus der Selbstadjungiertheit der Matrix $\mathbf{A}$ die Gleichung

$$\langle \mathbf{x}, \mathbf{A}\mathbf{x} \rangle_2 = \langle \mathbf{A}^*\mathbf{x}, \mathbf{x} \rangle_2 = \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle_2 = \overline{\langle \mathbf{x}, \mathbf{A}\mathbf{x} \rangle_2}$$

folgt, so dass der erste Summand in der Definition der Funktion f nur reelle Werte annehmen kann und die Funktion somit wohldefiniert ist.

Lemma 3.15 (Minimierung) *Seien $\mathbf{x}, \mathbf{p} \in \mathbb{K}^{\mathcal{I}}$. Dann gilt die Minimalitätsbedingung*

$$f(\mathbf{x}) \le f(\mathbf{x} + \lambda \mathbf{p}) \qquad \text{für alle } \lambda \in \mathbb{K} \tag{3.12}$$

genau dann, wenn

$$\langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{p} \rangle_2 = 0 \tag{3.13}$$

erfüllt ist. Insbesondere ist $\mathbf{x}$ genau dann ein globales Minimum der Funktion f, wenn es das lineare Gleichungssystem (1.1) löst.

Beweis. Wir berechnen zunächst

$$\begin{aligned}
f(\mathbf{x} + \lambda\mathbf{p}) &= \frac{1}{2}\langle \mathbf{x} + \lambda\mathbf{p}, \mathbf{A}(\mathbf{x} + \lambda\mathbf{p}) \rangle_2 - \mathrm{Re}\langle \mathbf{b}, \mathbf{x} + \lambda\mathbf{p} \rangle_2 \\
&= \frac{1}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x} \rangle_2 + \frac{\lambda}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{p} \rangle_2 + \frac{\bar{\lambda}}{2}\langle \mathbf{p}, \mathbf{A}\mathbf{x} \rangle_2 + \frac{|\lambda|^2}{2}\langle \mathbf{p}, \mathbf{A}\mathbf{p} \rangle_2 \\
&\quad - \mathrm{Re}\langle \mathbf{b}, \mathbf{x} \rangle_2 - \mathrm{Re}\,\lambda\langle \mathbf{b}, \mathbf{p} \rangle_2 \\
&= f(\mathbf{x}) + \frac{\lambda}{2}\langle \mathbf{A}\mathbf{x}, \mathbf{p} \rangle_2 + \frac{\bar{\lambda}}{2}\overline{\langle \mathbf{A}\mathbf{x}, \mathbf{p} \rangle_2} + \frac{|\lambda|^2}{2}\langle \mathbf{p}, \mathbf{A}\mathbf{p} \rangle_2 - \mathrm{Re}\,\lambda\langle \mathbf{b}, \mathbf{p} \rangle_2 \\
&= f(\mathbf{x}) + \mathrm{Re}\,\lambda\langle \mathbf{A}\mathbf{x}, \mathbf{p} \rangle_2 - \mathrm{Re}\,\lambda\langle \mathbf{b}, \mathbf{p} \rangle_2 + \frac{|\lambda|^2}{2}\langle \mathbf{p}, \mathbf{A}\mathbf{p} \rangle_2 \\
&= f(\mathbf{x}) + \mathrm{Re}\,\lambda\langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{p} \rangle_2 + \frac{|\lambda|^2}{2}\langle \mathbf{p}, \mathbf{A}\mathbf{p} \rangle_2.
\end{aligned} \tag{3.14}$$

Falls nun (3.13) gilt, folgt aus dieser Gleichung direkt $f(\mathbf{x} + \lambda\mathbf{p}) \ge f(\mathbf{x})$, also (3.12).

Gelte nun (3.12). Für $\mathbf{p} = \mathbf{0}$ ist die Aussage trivial, also nehmen wir nun $\mathbf{p} \neq \mathbf{0}$ an und wählen

$$\lambda := -\frac{\overline{\langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{p} \rangle_2}}{\langle \mathbf{A}\mathbf{p}, \mathbf{p} \rangle_2}.$$

Da $\mathbf{A}$ positiv definit ist, ist der Nenner ungleich null und λ somit wohldefiniert. Aus (3.14) folgt dann

$$\begin{aligned} f(\mathbf{x}+\lambda\mathbf{p}) &= f(\mathbf{x}) - \operatorname{Re}\frac{\overline{\langle \mathbf{A}\mathbf{x}-\mathbf{b},\mathbf{p}\rangle_2}}{\langle \mathbf{A}\mathbf{p},\mathbf{p}\rangle_2}\langle \mathbf{A}\mathbf{x}-\mathbf{b},\mathbf{p}\rangle_2 + \frac{|\langle \mathbf{A}\mathbf{x}-\mathbf{b},\mathbf{p}\rangle_2|^2}{2\langle \mathbf{A}\mathbf{p},\mathbf{p}\rangle_2^2}\langle \mathbf{A}\mathbf{p},\mathbf{p}\rangle_2 \\ &= f(\mathbf{x}) - \frac{|\langle \mathbf{A}\mathbf{x}-\mathbf{b},\mathbf{p}\rangle_2|^2}{\langle \mathbf{A}\mathbf{p},\mathbf{p}\rangle_2} + \frac{|\langle \mathbf{A}\mathbf{x}-\mathbf{b},\mathbf{p}\rangle_2|^2}{2\langle \mathbf{A}\mathbf{p},\mathbf{p}\rangle_2} \\ &= f(\mathbf{x}) - \frac{|\langle \mathbf{A}\mathbf{x}-\mathbf{b},\mathbf{p}\rangle_2|^2}{2\langle \mathbf{A}\mathbf{p},\mathbf{p}\rangle_2}. \end{aligned} \tag{3.15}$$

Aus der Minimalitätsbedingung (3.12) ergibt sich so

$$\frac{|\langle \mathbf{A}\mathbf{x}-\mathbf{b},\mathbf{p}\rangle_2|^2}{2\langle \mathbf{A}\mathbf{p},\mathbf{p}\rangle_2} \le 0,$$

also insbesondere auch (3.13).

Falls $\mathbf{x} \in \mathbb{K}^{\mathcal{I}}$ eine Lösung des Gleichungssystems (1.1) ist, gilt (3.13) für alle $\mathbf{p} \in \mathbb{K}^{\mathcal{I}}$, und wir haben soeben gezeigt, dass das bedeutet, dass $\mathbf{x}$ ein globales Minimum ist.

Falls umgekehrt (3.13) für alle $\mathbf{p} \in \mathbb{K}^{\mathcal{I}}$ gilt, können wir $\mathbf{p} = \mathbf{A}\mathbf{x} - \mathbf{b}$ einsetzen und erhalten

$$0 = \langle \mathbf{A}\mathbf{x}-\mathbf{b},\mathbf{p}\rangle_2 = \langle \mathbf{A}\mathbf{x}-\mathbf{b},\mathbf{A}\mathbf{x}-\mathbf{b}\rangle_2 = \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2,$$

also insbesondere $\mathbf{A}\mathbf{x} = \mathbf{b}$. ■

Unser Ziel sollte es nun also sein, die Funktion f zu minimieren. Ein naheliegender Ansatz besteht darin, eine *Suchrichtung* $\mathbf{p} \in \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\}$ zu wählen, die zu einer möglichst schnellen Reduktion des Funktionswerts führt. An der Gleichung (3.15) können wir ablesen, dass es besonders günstig ist, wenn

$$\frac{|\langle \mathbf{A}\mathbf{x}-\mathbf{b},\mathbf{p}\rangle_2|^2}{\langle \mathbf{A}\mathbf{p},\mathbf{p}\rangle_2^2}$$

einen möglichst großen Wert annimmt. Aus der Cauchy-Schwarz-Ungleichung folgt, dass der Zähler maximal wird, wenn $\mathbf{p}$ ein Vielfaches des Vektors $\mathbf{A}\mathbf{x} - \mathbf{b}$ ist, und wenn wir für den Moment den Nenner ignorieren, erhalten wir so eine praktisch berechenbare Suchrichtung.

Diese Wahl der Suchrichtung lässt sich geometrisch rechtfertigen: Im reellwertigen Fall ist f eine differenzierbare Funktion, und aus dem Satz von Taylor folgt

$$\begin{aligned} f(\mathbf{x}+\mathbf{p}) &= f(\mathbf{x}) + Df(\mathbf{x})\mathbf{p} + \frac{1}{2}D^2 f(\eta)(\mathbf{p},\mathbf{p}) \\ &= f(\mathbf{x}) + \langle \mathbf{A}\mathbf{x}-\mathbf{b},\mathbf{p}\rangle_2 + \frac{1}{2}\langle \mathbf{A}\mathbf{p},\mathbf{p}\rangle_2 \end{aligned}$$

für einen zwischen $\mathbf{x}$ und $\mathbf{x}+\mathbf{p}$ liegenden Zwischenpunkt $\eta \in \mathbb{K}^{\mathcal{I}}$ (vgl. (3.14)). Falls die Norm des Vektors $\mathbf{p}$ hinreichend klein ist, können wir den quadratischen Term vernachlässigen und stellen fest, dass das *Residuum*

$$\mathbf{r} := \mathbf{b} - \mathbf{A}\mathbf{x}$$

die (lokal) größte Reduktion der Funktion f verspricht.

Mit dieser Wahl der Suchrichtung erhalten wir das *Verfahren des steilsten Abstiegs.* Im reellwertigen Fall definiert man den *Gradienten* $\mathbf{grad}\, f(\mathbf{x})$ der Funktion f im Punkt $\mathbf{x}$ durch die Gleichung

$$\langle \mathbf{grad}\, f(\mathbf{x}), \xi \rangle_2 = Df(\mathbf{x})\xi \qquad \text{für alle } \xi \in \mathbb{K}^{\mathcal{I}},$$

so dass in unserem Fall gerade $\mathbf{r} = -\,\mathbf{grad}\, f(\mathbf{x})$ gilt. Deshalb trägt das beschriebene Verfahren auch den Namen *Gradientenverfahren.* Es ist ein allgemeiner Algorithmus zur Minimierung reellwertiger Funktionen, denn aus dem Satz von Taylor folgt auch im allgemeinen Fall, dass $-\,\mathbf{grad}\, f(\mathbf{x})$ die lokal beste Richtung ist.

Da wir uns nun für eine Suchrichtung entschieden haben, stellt sich die Frage, wie weit wir ihr folgen sollen.

Lemma 3.16 (Optimale Schrittweite) *Sei* $\mathbf{p} \neq \mathbf{0}$. *Mit der Wahl*

$$\lambda_{\text{opt}} := \frac{\langle \mathbf{p}, \mathbf{b} - \mathbf{A}\mathbf{x} \rangle_2}{\langle \mathbf{p}, \mathbf{A}\mathbf{p} \rangle_2}$$

gilt

$$f(\mathbf{x} + \lambda_{\text{opt}}\mathbf{p}) \leq f(\mathbf{x} + \lambda \mathbf{p}) \qquad \textit{für alle } \lambda \in \mathbb{K}.$$

Beweis. Indem wir $\widehat{\mathbf{x}} := \mathbf{x} + \lambda_{\text{opt}}\mathbf{p}$ setzen, können wir diese Minimalitätsbedingung in der Form

$$f(\widehat{\mathbf{x}}) \leq f(\widehat{\mathbf{x}} + \lambda \mathbf{p}) \qquad \text{für alle } \lambda \in \mathbb{K}$$

schreiben und so auf Lemma 3.15 zurückführen. Aus diesem Lemma folgt, dass

$$0 = \langle \mathbf{p}, \mathbf{A}\widehat{\mathbf{x}} - \mathbf{b} \rangle_2 = \langle \mathbf{p}, \mathbf{A}(\mathbf{x} + \lambda_{\text{opt}}\mathbf{p}) - \mathbf{b} \rangle_2 = \langle \mathbf{p}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle_2 + \lambda_{\text{opt}} \langle \mathbf{p}, \mathbf{A}\mathbf{p} \rangle_2$$

gelten muss. Daraus erhalten wir direkt

$$\lambda_{\text{opt}} = -\frac{\langle \mathbf{p}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle_2}{\langle \mathbf{p}, \mathbf{A}\mathbf{p} \rangle_2} = \frac{\langle \mathbf{p}, \mathbf{b} - \mathbf{A}\mathbf{x} \rangle_2}{\langle \mathbf{p}, \mathbf{A}\mathbf{p} \rangle_2},$$

also ist λ_{opt} die optimale Wahl. ■

Im Gradientenverfahren verwenden wir keine allgemeine Suchrichtung, sondern

$$\mathbf{p} = \mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x},$$

so dass sich für die optimale Schrittweite die vereinfachte Formel

$$\lambda_{\text{opt}} = \frac{\langle \mathbf{r}, \mathbf{b} - \mathbf{A}\mathbf{x} \rangle_2}{\langle \mathbf{r}, \mathbf{A}\mathbf{r} \rangle_2} = \frac{\|\mathbf{r}\|_2^2}{\langle \mathbf{r}, \mathbf{A}\mathbf{r} \rangle_2}$$

ergibt. Insbesondere gilt, da $\mathbf{A}$ positiv definit ist, immer $\lambda_{\text{opt}} \in \mathbb{R}_{>0}$. Damit ist das Gradientenverfahren für die Funktion f vollständig beschrieben:

```
procedure Gradienten(b, var x);
r ← b − Ax;
while Fehler zu groß do
  a ← Ar;
  λ_opt ← ||r||_2^2 / ⟨a,r⟩_2;
  x ← x + λ_opt r;
  r ← r − λ_opt a
end while
```

Abbildung 3.3: Gradientenverfahren

Definition 3.17 (Gradientenverfahren) *Sei* $\mathbf{A}$ *positiv definit. Das durch*

$$\Phi_{\mathrm{Grad}}(\mathbf{x},\mathbf{b}) := \mathbf{x} + \frac{\|\mathbf{r}\|_2^2}{\langle \mathbf{r}, \mathbf{A}\mathbf{r}\rangle_2}\mathbf{r}, \qquad \mathbf{r} := \mathbf{b} - \mathbf{A}\mathbf{x} \qquad \textit{für alle } \mathbf{x},\mathbf{b} \in \mathbb{K}^{\mathcal{I}}$$

gegebene Iterationsverfahren nennen wir das Gradientenverfahren.

Offensichtlich ist Φ_{Grad} im Allgemeinen kein *lineares* Iterationsverfahren mehr, lässt sich aber immerhin wegen

$$\Phi_{\mathrm{Grad}}(\mathbf{x},\mathbf{b}) = \mathbf{x} - \lambda_{\mathrm{opt}}(\mathbf{A}\mathbf{x} - \mathbf{b})$$

noch als „nichtlinear gedämpftes“ Richardson-Verfahren interpretieren.

Bei der Implementierung des Gradientenverfahrens ist es möglich, mit einer Matrix-Vektor-Multiplikation pro Iterationsschritt auszukommen, wenn man das Residuum $\mathbf{r}$ speichert und ausnutzt, dass das Residuum $\mathbf{r}' := \mathbf{b} - \mathbf{A}\mathbf{x}'$ zur nächsten Iterierten $\mathbf{x}' := \Phi_{\mathrm{Grad}}(\mathbf{x},\mathbf{b})$ aus dem alten Residuum $\mathbf{r}$ durch

$$\mathbf{r}' = \mathbf{b} - \mathbf{A}\mathbf{x}' = \mathbf{b} - \mathbf{A}(\mathbf{x} + \lambda_{\mathrm{opt}}\mathbf{r}) = \mathbf{b} - \mathbf{A}\mathbf{x} - \lambda_{\mathrm{opt}}\mathbf{A}\mathbf{r} = \mathbf{r} - \lambda_{\mathrm{opt}}\mathbf{A}\mathbf{r}$$

berechnet werden kann. Der entsprechende Algorithmus ist in Abbildung 3.3 gegeben. Er benötigt neben der rechten Seite $\mathbf{b}$ und der Iterierten $\mathbf{x}$ lediglich das Residuum $\mathbf{r}$ und das Produkt aus der Matrix $\mathbf{A}$ und dem Residuum, sein Speicherbedarf ist also mit dem der Tschebyscheff-Semiiteration vergleichbar.

Die Überwachung der Genauigkeit dieses Verfahren ist besonders einfach, da am Ende jedes Schleifendurchlaufs das aktuelle Residuum zur Verfügung steht, so dass sich die Defektnorm $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2 = \|\mathbf{r}\|_2$ einfach berechnen lässt. Da sie ohnehin für die Berechnung des nächsten Wertes von λ_{opt} benötigt wird, entsteht kein zusätzlicher Aufwand.

Lemma 3.18 *Sei* $\theta \in \mathbb{K}$. *Sei* $\mathbf{b} \in \mathbb{K}^{\mathcal{I}}$ *und* $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$. *Dann gilt*

$$\|\Phi_{\mathrm{Grad}}(\mathbf{x},\mathbf{b}) - \mathbf{x}^*\|_A \le \|\Phi_{\mathrm{Rich},\theta}(\mathbf{x},\mathbf{b}) - \mathbf{x}^*\|_A \qquad \textit{für alle } \mathbf{x} \in \mathbb{K}^{\mathcal{I}},$$

das Gradientenverfahren ist also mindestens so schnell wie ein optimal gedämpftes Richardson-Verfahren.

Beweis. Sei nun $\mathbf{x} \in \mathbb{K}^{\mathcal{I}}$ und $\mathbf{r} := \mathbf{b} - \mathbf{A}\mathbf{x}$. Wir haben λ_{opt} gerade so konstruiert, dass

$$f(\mathbf{x} + \lambda_{\text{opt}}\mathbf{r}) \leq f(\mathbf{x} + \lambda\mathbf{r}) \qquad \text{für alle } \lambda \in \mathbb{K}$$

gilt, also gilt nach (3.11) auch

$$\begin{aligned}\|\mathbf{x} + \lambda_{\text{opt}}\mathbf{r} - \mathbf{x}^*\|_A^2 &= 2f(\mathbf{x} + \lambda_{\text{opt}}\mathbf{r}) + \langle \mathbf{x}^*, \mathbf{A}\mathbf{x}^*\rangle_2 \\ &\leq 2f(\mathbf{x} + \lambda\mathbf{r}) + \langle \mathbf{x}^*, \mathbf{A}\mathbf{x}^*\rangle_2 = \|\mathbf{x} + \lambda\mathbf{r} - \mathbf{x}^*\|_A^2 \qquad \text{für alle } \lambda \in \mathbb{K}.\end{aligned}$$

Für $\lambda = \theta$ erhalten wir so

$$\begin{aligned}\|\Phi_{\text{Grad}}(\mathbf{x}, \mathbf{b}) - \mathbf{x}^*\|_A^2 &\leq \|\mathbf{x} + \lambda\mathbf{r} - \mathbf{x}^*\|_A^2 = \|\mathbf{x} - \theta(\mathbf{A}\mathbf{x} - \mathbf{b}) - \mathbf{x}^*\|_A^2 \\ &= \|\Phi_{\text{Rich},\theta}(\mathbf{x}, \mathbf{b}) - \mathbf{x}^*\|_A^2.\end{aligned}$$

Das ist die gesuchte Abschätzung. ■

Satz 3.19 (Konvergenz) *Sei* $\mathbf{A}$ *positiv definit. Seien* $\alpha, \beta \in \mathbb{R}_{>0}$ *gegeben mit* $\sigma(\mathbf{A}) \subseteq [\alpha, \beta]$. *Sei* $\mathbf{b} \in \mathbb{K}^{\mathcal{I}}$, *und sei* $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$. *Dann gilt*

$$\|\Phi_{\text{Grad}}(\mathbf{x}, \mathbf{b}) - \mathbf{x}^*\|_A \leq \frac{\beta - \alpha}{\beta + \alpha}\|\mathbf{x} - \mathbf{x}^*\|_A \qquad \textit{für alle } \mathbf{x} \in \mathbb{K}^{\mathcal{I}}.$$

Für die Folge $(\mathbf{x}^{(m)})_{m\in\mathbb{N}_0}$ *der Iterierten des Gradientenverfahrens* Φ_{Grad} *zu einem Startvektor* $\mathbf{x}^{(0)} \in \mathbb{K}^{\mathcal{I}}$ *folgt daraus*

$$\|\mathbf{x}^{(m)} - \mathbf{x}^*\|_A \leq \left(\frac{\beta - \alpha}{\beta + \alpha}\right)^m \|\mathbf{x}^{(0)} - \mathbf{x}^*\|_A \qquad \textit{für alle } m \in \mathbb{N}_0.$$

Beweis. Für den Dämpfungsparameter

$$\theta := \frac{2}{\beta + \alpha}$$

erhalten wir gemäß Bemerkung 2.32 die Abschätzung

$$\varrho(\mathbf{M}_{\text{Rich},\theta}) \leq \frac{\beta - \alpha}{\beta + \alpha}.$$

Da $\mathbf{A}$ positiv definit und θ reell ist, ist $\mathbf{M}_{\text{Rich},\theta} = \mathbf{I} - \theta\mathbf{A}$ selbstadjungiert, also gilt

$$\|\mathbf{M}_{\text{Rich},\theta}\|_A = \|\mathbf{A}^{1/2}(\mathbf{I} - \theta\mathbf{A})\mathbf{A}^{-1/2}\|_2 = \|\mathbf{I} - \theta\mathbf{A}\|_2 = \varrho(\mathbf{M}_{\text{Rich},\theta}) \leq \frac{\beta - \alpha}{\beta + \alpha} < 1.$$

Aus Lemma 3.18 und Lemma 2.11 erhalten wir

$$\begin{aligned}\|\Phi_{\text{Grad}}(\mathbf{x}, \mathbf{b}) - \mathbf{x}^*\|_A &\leq \|\Phi_{\text{Rich},\theta}(\mathbf{x}, \mathbf{b}) - \mathbf{x}^*\|_A \leq \|\mathbf{M}_{\text{Rich},\theta}\|_A\|\mathbf{x} - \mathbf{x}^*\|_A \\ &\leq \frac{\beta - \alpha}{\beta + \alpha}\|\mathbf{x} - \mathbf{x}^*\|_A.\end{aligned}$$

Für die zu Φ_{Grad} und $\mathbf{b}$ gehörende Folge der Iterierten gilt

$$\|\mathbf{x}^{(m+1)} - \mathbf{x}^*\|_A = \|\Phi_{\text{Grad}}(\mathbf{x}^{(m)}, \mathbf{b}) - \mathbf{x}^*\|_A \leq \frac{\beta - \alpha}{\beta + \alpha}\|\mathbf{x}^{(m)} - \mathbf{x}^*\|_A \quad \text{für alle } m \in \mathbb{N}_0,$$

also können wir die gewünschte Abschätzung mit einer einfachen Induktion erhalten. ■

Das Gradientenverfahren lässt sich also auf beliebige positiv definite Matrizen $\mathbf{A}$ anwenden und erzielt immer eine Konvergenzrate, die der der optimal gedämpften Richardson-Iteration entspricht.

Wir haben in Bemerkung 2.33 gesehen, dass sich beliebige konsistente lineare Iterationsverfahren aus dem Richardson-Verfahren konstruieren lassen, indem man von dem linearen Gleichungssystem (1.1) zu dem vorkonditionierten System

$$\mathbf{NAx} = \mathbf{Nb}$$

übergeht. Wir würden gerne in dieser Weise auch unsere Aussagen über das Gradientenverfahren verallgemeinern. Leider wäre die oben verwendete Matrix $\mathbf{NA}$ im Allgemeinen nicht mehr positiv definit, so dass sich Satz 3.19 nicht mehr anwenden ließe, deshalb müssen wir einen alternativen Zugang wählen:

Wir nehmen an, dass $\mathbf{A}$ und $\mathbf{N}$ positiv definit sind, und definieren

$$\widehat{\mathbf{A}} := \mathbf{N}^{1/2}\mathbf{A}\mathbf{N}^{1/2}, \qquad \widehat{\mathbf{b}} := \mathbf{N}^{1/2}\mathbf{b}, \qquad \widehat{\mathbf{x}} := \mathbf{N}^{-1/2}\mathbf{x}.$$

Dann ist $\widehat{\mathbf{A}}$ positiv definit, und die Gleichung

$$\widehat{\mathbf{A}}\widehat{\mathbf{x}} = \widehat{\mathbf{b}}, \tag{3.16}$$

ist äquivalent zu (1.1).

Wenn wir das Gradientenverfahren auf dieses vorkonditionierte Gleichungssystem anwenden, ist das Residuum nun durch

$$\widehat{\mathbf{r}} := \widehat{\mathbf{b}} - \widehat{\mathbf{A}}\widehat{\mathbf{x}} = \mathbf{N}^{1/2}\mathbf{b} - \mathbf{N}^{1/2}\mathbf{A}\mathbf{N}^{1/2}\mathbf{N}^{-1/2}\mathbf{x} = \mathbf{N}^{1/2}(\mathbf{b} - \mathbf{Ax}) = \mathbf{N}^{1/2}\mathbf{r}$$

gegeben, und der optimale Dämpfungsparameter durch

$$\widehat{\lambda}_{\text{opt}} := \frac{\|\widehat{\mathbf{r}}\|_2^2}{\langle \widehat{\mathbf{r}}, \widehat{\mathbf{A}}\widehat{\mathbf{r}}\rangle_2} = \frac{\|\mathbf{N}^{1/2}\mathbf{r}\|_2^2}{\langle \mathbf{N}^{1/2}\mathbf{r}, \mathbf{N}^{1/2}\mathbf{A}\mathbf{N}^{1/2}\mathbf{N}^{1/2}\mathbf{r}\rangle_2} = \frac{\langle \mathbf{Nr}, \mathbf{r}\rangle_2}{\langle \mathbf{Nr}, \mathbf{ANr}\rangle_2},$$

so dass die nächste Iterierte $\widehat{\mathbf{x}}'$ durch

$$\widehat{\mathbf{x}}' := \widehat{\mathbf{x}} + \frac{\langle \mathbf{Nr}, \mathbf{r}\rangle_2}{\langle \mathbf{Nr}, \mathbf{ANr}\rangle_2}\mathbf{N}^{1/2}\mathbf{r}$$

gegeben ist. Indem wir beide Seiten mit $\mathbf{N}^{1/2}$ multiplizieren, erhalten wir schließlich

$$\mathbf{x}' := \mathbf{N}^{1/2}\widehat{\mathbf{x}}' = \mathbf{x} + \frac{\langle \mathbf{Nr}, \mathbf{r}\rangle_2}{\langle \mathbf{Nr}, \mathbf{ANr}\rangle_2}\mathbf{Nr}.$$

In diesem Ausdruck treten $\mathbf{N}^{1/2}$ und $\mathbf{N}^{-1/2}$ nicht mehr auf, so dass er sich praktisch auswerten lässt.

```
procedure VorkondGradienten(b, var x);
r ← b − Ax;
while Fehler zu groß do
  q ← Nr;
  a ← Aq;
  λ_opt ← ⟨q,r⟩_2 / ⟨q,a⟩_2;
  x ← x + λ_opt q;
  r ← r − λ_opt a
end while
```

Abbildung 3.4: Vorkonditioniertes Gradientenverfahren

Definition 3.20 (Vorkonditioniertes Gradientenverfahren) *Seien* $\mathbf{A}$ *und* $\mathbf{N}$ *positiv definit. Das durch*

$$\Phi_{\mathrm{Grad},N}(\mathbf{x},\mathbf{b}) := \mathbf{x} + \frac{\langle \mathbf{Nr},\mathbf{r}\rangle_2}{\langle \mathbf{Nr},\mathbf{ANr}\rangle_2}\mathbf{Nr}, \qquad \mathbf{r} := \mathbf{b} - \mathbf{Ax} \qquad \textit{für alle } \mathbf{x},\mathbf{b} \in \mathbb{K}^{\mathcal{I}}$$

gegebene Iterationsverfahren nennen wir das vorkonditionierte Gradientenverfahren*. Die Matrix* $\mathbf{N}$ *bezeichnen wir in diesem Kontext als* Vorkonditionierer*.*

Das vorkonditionierte Gradientenverfahren benötigt den zusätzlichen Vektor $\mathbf{q} := \mathbf{Nr}$, der den vorkonditionierten Gradienten aufnimmt. Es ist möglich, die Berechnung so zu organisieren, dass für einen Schritt des neuen Verfahrens lediglich eine Multiplikation mit $\mathbf{A}$ und eine mit $\mathbf{N}$ erforderlich sind. Der entsprechende Algorithmus ist in Abbildung 3.4 angegeben.

Die Konvergenzaussage von Satz 3.19 überträgt sich direkt auf das Gradientenverfahren mit Vorkonditionierer:

Satz 3.21 (Konvergenz) *Seien* $\mathbf{A}$ *und* $\mathbf{N}$ *positiv definit. Seien* $\alpha,\beta \in \mathbb{R}_{>0}$ *mit* $\sigma(\mathbf{NA}) \subseteq [\alpha,\beta]$*. Sei* $\mathbf{b} \in \mathbb{K}^{\mathcal{I}}$*, und sei* $\mathbf{x}^* := \mathbf{A}^{-1}\mathbf{b}$*. Dann gilt*

$$\|\Phi_{\mathrm{Grad},N}(\mathbf{x},\mathbf{b}) - \mathbf{x}^*\|_A \le \frac{\beta-\alpha}{\beta+\alpha}\|\mathbf{x}-\mathbf{x}^*\|_A \qquad \textit{für alle } \mathbf{x} \in \mathbb{K}^{\mathcal{I}}.$$

Für die Folge $(\mathbf{x}^{(m)})_{m\in\mathbb{N}_0}$ *der Iterierten des vorkonditionierten Gradientenverfahrens* $\Phi_{\mathrm{Grad},N}$ *zu einem Startvektor* $\mathbf{x}^{(0)} \in \mathbb{K}^{\mathcal{I}}$ *folgt daraus*

$$\|\mathbf{x}^{(m)} - \mathbf{x}^*\|_A \le \left(\frac{\beta-\alpha}{\beta+\alpha}\right)^m \|\mathbf{x}^{(0)} - \mathbf{x}^*\|_A \qquad \textit{für alle } m \in \mathbb{N}_0.$$

Beweis. Wir bezeichnen mit $\widehat{\Phi}_{\mathrm{Grad}}$ das Gradientenverfahren für das vorkonditionierte Gleichungssystem (3.16). Wir bezeichnen die auf $\widehat{\mathbf{x}}$ folgende Iterierte wieder mit

$$\widehat{\mathbf{x}}' := \widehat{\Phi}_{\mathrm{Grad}}(\widehat{\mathbf{x}},\widehat{\mathbf{b}}) = \widehat{\mathbf{x}} + \frac{\langle \mathbf{Nr},\mathbf{r}\rangle_2}{\langle \mathbf{Nr},\mathbf{ANr}\rangle_2}\mathbf{N}^{1/2}\mathbf{r}$$

und die Lösung des vorkonditionierten Systems mit

$$\widehat{\mathbf{x}}^* := \mathbf{N}^{-1/2}\mathbf{x}^*.$$

Nach Definition 2.46 gilt

$$\begin{aligned}\|\widehat{\Phi}_{\mathrm{Grad}}(\widehat{\mathbf{x}}, \widehat{\mathbf{b}}) - \widehat{\mathbf{x}}^*\|_{\widehat{A}}^2 &= \|\widehat{\mathbf{x}}' - \widehat{\mathbf{x}}^*\|_{\widehat{A}}^2 = \langle \mathbf{N}^{1/2}\mathbf{A}\mathbf{N}^{1/2}(\widehat{\mathbf{x}}' - \widehat{\mathbf{x}}^*), (\widehat{\mathbf{x}}' - \widehat{\mathbf{x}}^*)\rangle_2 \\ &= \langle \mathbf{A}\mathbf{N}^{1/2}(\widehat{\mathbf{x}}' - \widehat{\mathbf{x}}^*), \mathbf{N}^{1/2}(\widehat{\mathbf{x}}' - \widehat{\mathbf{x}}^*)\rangle_2 = \langle \mathbf{A}(\mathbf{x}' - \mathbf{x}^*), (\mathbf{x}' - \mathbf{x}^*)\rangle_2 \\ &= \|\mathbf{x}' - \mathbf{x}^*\|_A^2 = \|\Phi_{\mathrm{Grad},N}(\mathbf{x}, \mathbf{b}) - \mathbf{x}^*\|_A^2,\end{aligned}$$

also erhalten wir die gewünschte Aussage, indem wir Satz 3.19 auf das vorkonditionierte Gleichungssystem (3.16) anwenden. ■

Bemerkung 3.22 *Die Bedingung $\sigma(\mathbf{NA}) \subseteq [\alpha, \beta]$ aus Satz 3.21 lässt sich auf die von der Behandlung des Jacobi-Verfahrens her bekannte Form bringen: Für die positiv definite Matrix $\mathbf{W} := \mathbf{N}^{-1}$ gilt*

$$\begin{aligned}\sigma(\mathbf{NA}) \subseteq [\alpha, \beta] &\iff \sigma(\mathbf{N}^{1/2}\mathbf{A}\mathbf{N}^{1/2}) \subseteq [\alpha, \beta] \iff \alpha\mathbf{I} \leq \mathbf{N}^{1/2}\mathbf{A}\mathbf{N}^{1/2} \leq \beta\mathbf{I} \\ &\iff \alpha\mathbf{N}^{-1} \leq \mathbf{A} \leq \beta\mathbf{N}^{-1} \iff \alpha\mathbf{W} \leq \mathbf{A} \leq \beta\mathbf{W}.\end{aligned}$$

3.4 Verfahren der konjugierten Gradienten

Das Gradientenverfahren hat den Vorteil, dass es für jede positiv definite Matrix $\mathbf{A}$ funktioniert und keinerlei Parameter gewählt werden müssen. Leider sind die in den Sätzen 3.19 und 3.21 gegebenen Abschätzungen für die Konvergenzgeschwindigkeit weniger gut als im Fall der Tschebyscheff-Semiiteration, insofern werden wir uns nun auf die Suche nach einem Verfahren begeben, das die allgemeine Anwendbarkeit des Gradientenverfahrens mit der hohen Geschwindigkeit des Tschebyscheff-Verfahrens kombiniert.

Im Gradientenverfahren wird der Parameter λ_{opt} gerade so gewählt, dass $f(\mathbf{x} + \lambda_{\mathrm{opt}}\mathbf{r})$ minimal ist. Wenn wir die neue Iterierte mit

$$\mathbf{x}' := \mathbf{x} + \lambda_{\mathrm{opt}}\mathbf{r}$$

bezeichnen, bedeutet diese Minimalitätsbedingung nach Lemma 3.15 gerade, dass

$$\langle \mathbf{A}(\mathbf{x}' - \mathbf{x}^*), \mathbf{r}\rangle_2 = \langle \mathbf{A}\mathbf{x}' - \mathbf{b}, \mathbf{r}\rangle_2 = 0 \tag{3.17}$$

gilt. Das neue Residuum

$$\mathbf{r}' := \mathbf{b} - \mathbf{A}\mathbf{x}'$$

erfüllt also die Gleichung

$$\langle \mathbf{r}', \mathbf{r}\rangle_2 = 0.$$

Wir können den Schluss ziehen, dass die im Gradientenverfahren verwendeten Suchrichtungen

$$\mathbf{r}^{(m)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(m)}$$

bezüglich des euklidischen Skalarprodukts senkrecht aufeinander stehen. In vielen Situation führt das dazu, dass die in einem Schritt des Verfahrens erzielte Optimalität bezüglich der gerade gewählten Suchrichtung im nächsten Schritt wieder zunichte gemacht wird.

Unser Ziel ist es nun, das Verfahren so zu verbessern, dass die Optimalität erhalten bleibt. Die erste Iterierte soll also optimal bezüglich der ersten Suchrichtung sein, die zweite bezüglich der ersten beiden Suchrichtungen, dir m-te bezüglich der ersten m.

Den Ausgangspunkt unserer Betrachtung bildet eine Charakterisierung der Optimalität bezüglich einer Suchrichtung.

Definition 3.23 (Optimalität) *Seien* $\mathbf{x}, \mathbf{p} \in \mathbb{K}^{\mathcal{I}}$. *Falls*

$$f(\mathbf{x}) \leq f(\mathbf{x} + \lambda \mathbf{p}) \qquad \text{für alle } \lambda \in \mathbb{K}$$

gilt, nennen wir $\mathbf{x}$ optimal bezüglich der Richtung $\mathbf{p}$.

Aus Lemma 3.15 folgt, dass ein Vektor $\mathbf{x} \in \mathbb{K}^{\mathcal{I}}$ genau dann optimal bezüglich einer Richtung $\mathbf{p} \in \mathbb{K}^{\mathcal{I}}$ ist, wenn die Gleichung

$$\langle \mathbf{p}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle_2 = \langle \mathbf{p}, \mathbf{A}(\mathbf{x} - \mathbf{x}^*) \rangle_2 = 0 \tag{3.18}$$

erfüllt ist.

Wenn $\mathbf{x}$ bereits optimal bezüglich der Richtungen $\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m-1)} \in \mathbb{K}^{\mathcal{I}}$ ist, möchten wir sicherstellen, dass diese Eigenschaft auch nach der Korrektur

$$\mathbf{x}' := \mathbf{x} + \lambda \mathbf{p}^{(m)}$$

noch gegeben ist, dass also

$$\begin{aligned} 0 = \langle \mathbf{p}^{(\ell)}, \mathbf{A}\mathbf{x}' - \mathbf{b} \rangle_2 &= \langle \mathbf{p}^{(\ell)}, \mathbf{A}\mathbf{x} + \lambda \mathbf{A}\mathbf{p}^{(m)} - \mathbf{b} \rangle_2 \\ &= \langle \mathbf{p}^{(\ell)}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle_2 + \lambda \langle \mathbf{p}^{(\ell)}, \mathbf{A}\mathbf{p}^{(m)} \rangle_2 \\ &= \lambda \langle \mathbf{p}^{(\ell)}, \mathbf{A}\mathbf{p}^{(m)} \rangle_2 \quad \text{für alle } \ell \in [0 : m-1] \end{aligned}$$

erfüllt ist. Da die Wahl $\lambda = 0$ zu keiner Verbesserung des Fehlers führen würde, besteht die einzige Lösung darin, sicherzustellen, dass $\mathbf{p}^{(m)}$ die Gleichung

$$\langle \mathbf{p}^{(\ell)}, \mathbf{A}\mathbf{p}^{(m)} \rangle_2 = 0 \qquad \text{für alle } \ell \in [0 : m-1]$$

erfüllt. Diese Gleichung lässt sich geometrisch interpretieren: Da $\mathbf{A}$ eine selbstadjungierte positiv definite Matrix ist, können wir das *Energieskalarprodukt*

$$\langle \mathbf{x}, \mathbf{y} \rangle_A := \langle \mathbf{x}, \mathbf{A}\mathbf{y} \rangle_2 \qquad \text{für alle } \mathbf{x}, \mathbf{y} \in \mathbb{K}^{\mathcal{I}}$$

definieren, das für alle $\mathbf{x} \in \mathbb{K}^{\mathcal{I}}$ über die Gleichung $\|\mathbf{x}\|_A = \langle \mathbf{x}, \mathbf{x} \rangle_A^{1/2}$ mit dem bereits bekannten Energieskalarprodukt in Beziehung steht.

Die Gleichung (3.17) nimmt dann die Form

$$\langle \mathbf{p}^{(\ell)}, \mathbf{p}^{(m)} \rangle_A = 0$$

an, die Richtungen müssen also bezüglich des Energieskalarprodukts senkrecht zueinander sein. Derartige Vektoren nennt man auch *konjugiert* zueinander.

Glücklicherweise lässt sich diese Eigenschaft mit Hilfe der *Gram-Schmidt-Orthogonalisierung* einfach garantieren: Wie zuvor wählen wir als Ausgangspunkt das Residuum $\mathbf{r}^{(m)} := \mathbf{b} - \mathbf{A}\mathbf{x}^{(m)}$ und konstruieren die Suchrichtung $\mathbf{p}^{(m)}$ durch

$$\mathbf{p}^{(m)} := \mathbf{r}^{(m)} - \sum_{k=0}^{m-1} \frac{\langle \mathbf{p}^{(k)}, \mathbf{r}^{(m)} \rangle_A}{\langle \mathbf{p}^{(k)}, \mathbf{p}^{(k)} \rangle_A} \mathbf{p}^{(k)}, \tag{3.19}$$

denn diese Wahl stellt sicher, dass

$$\begin{aligned}
\langle \mathbf{p}^{(\ell)}, \mathbf{p}^{(m)} \rangle_A &= \langle \mathbf{p}^{(\ell)}, \mathbf{r}^{(m)} \rangle_A - \sum_{k=0}^{m-1} \frac{\langle \mathbf{p}^{(k)}, \mathbf{r}^{(m)} \rangle_A}{\langle \mathbf{p}^{(k)}, \mathbf{p}^{(k)} \rangle_A} \langle \mathbf{p}^{(\ell)}, \mathbf{p}^{(k)} \rangle_A \\
&= \langle \mathbf{p}^{(\ell)}, \mathbf{r}^{(m)} \rangle_A - \frac{\langle \mathbf{p}^{(\ell)}, \mathbf{r}^{(m)} \rangle_A}{\langle \mathbf{p}^{(\ell)}, \mathbf{p}^{(\ell)} \rangle_A} \langle \mathbf{p}^{(\ell)}, \mathbf{p}^{(\ell)} \rangle_A \\
&= \langle \mathbf{p}^{(\ell)}, \mathbf{r}^{(m)} \rangle_A - \langle \mathbf{p}^{(\ell)}, \mathbf{r}^{(m)} \rangle_A = 0 \qquad \text{für alle } \ell \in [0 : m-1]
\end{aligned}$$

gilt. Da im Fall $\mathbb{K} = \mathbb{R}$ die Residuen $\mathbf{r}^{(m)}$ gerade den negativen Gradienten $-\nabla f(\mathbf{x}^{(m)})$ der zu minimierenden Funktion entsprechen, nennen wir die so konstruierten konjugierten Richtungen $\mathbf{p}^{(m)}$ auch *konjugierte Gradienten*.

Nun können wir wie bisher fortfahren: Aus Lemma 3.16 folgt, dass für die Suchrichtung $\mathbf{p}^{(m)}$ der Parameter

$$\lambda_{\text{opt}}^{(m)} := \frac{\langle \mathbf{p}^{(m)}, \mathbf{r}^{(m)} \rangle_2}{\langle \mathbf{p}^{(m)}, \mathbf{A}\mathbf{p}^{(m)} \rangle_2}$$

optimal ist, und die nächste Iterierte wird durch

$$\mathbf{x}^{(m+1)} := \mathbf{x}^{(m)} + \lambda_{\text{opt}}^{(m)} \mathbf{p}^{(m)} \tag{3.20}$$

definiert. Abgesehen von der zusätzlichen Orthogonalisierung (3.19) entspricht unser neues Verfahren also dem Gradientenverfahren.

Lemma 3.24 (Lucky Breakdown) *Sei $m \in \mathbb{N}_0$. Falls $\mathbf{p}^{(m)} = \mathbf{0}$ gilt, folgt auch $\mathbf{r}^{(m)} = \mathbf{0}$, also $\mathbf{A}\mathbf{x}^{(m)} = \mathbf{b}$.*

Beweis. Es gelte $\mathbf{p}^{(m)} = \mathbf{0}$. Aufgrund der Konstruktion (3.19) folgt

$$\mathbf{r}^{(m)} \in \operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m-1)}\},$$

also finden wir $\alpha_0, \ldots, \alpha_{m-1} \in \mathbb{K}$ mit

$$\mathbf{r}^{(m)} = \sum_{\ell=0}^{m-1} \alpha_\ell \mathbf{p}^{(\ell)}.$$

Da $\mathbf{x}^{(m)}$ nach Konstruktion optimal bezüglich aller Richtungen $\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m-1)}$ ist, gilt mit (3.18) auch

$$\|\mathbf{r}^{(m)}\|_2^2 = \langle \mathbf{r}^{(m)}, \mathbf{r}^{(m)} \rangle_2 = \sum_{\ell=0}^{m-1} \alpha_\ell \langle \mathbf{r}^{(m)}, \mathbf{p}^{(\ell)} \rangle_2 = \sum_{\ell=0}^{m-1} \alpha_\ell \langle \mathbf{b} - \mathbf{A}\mathbf{x}^{(m)}, \mathbf{p}^{(\ell)} \rangle_2 = 0,$$

und damit $\mathbf{r}^{(m)} = \mathbf{0}$, also $\mathbf{b} = \mathbf{A}\mathbf{x}^{(m)}$. ■

Falls also keine neuen „echten“ Suchrichtungen mehr konstruiert werden können, hat unser Algorithmus bereits die exakte Lösung gefunden, so dass wir ihn guten Gewissens beenden können.

Im Folgenden bezeichnen wir mit

$$m_0 := \max\{m \in \mathbb{N}_0 \ : \ \mathbf{p}^{(\ell)} \neq \mathbf{0} \text{ für alle } \ell \in [0 : m-1]\}$$

die Nummer der ersten Iteration, bei der $\mathbf{p}^{(m_0)} = \mathbf{0}$ auftritt, so dass die Orthogonalisierung nicht weiter durchgeführt werden kann.

Die Definition (3.19) vermittelt den Eindruck, dass wir alle bisherigen Suchrichtungen $(\mathbf{p}^{(\ell)})_{\ell=0}^m$ benötigen, um $\mathbf{p}^{(m+1)}$ berechnen zu können. Wäre das der Fall, würden Speicherbedarf und Rechenzeit unseres Verfahrens wie m^2 wachsen, es wäre also wesentlich aufwendiger als die vorher betrachteten Methoden.

Glücklicherweise können wir dieses Problem vermeiden, indem wir die Optimalität der Vektoren $\mathbf{x}^{(m)}$ und $\mathbf{A} = \mathbf{A}^*$ ausnutzen, um die Berechnung effizienter zu gestalten.

Definition 3.25 (Krylow-Raum) *Sei $\mathbf{y} \in \mathbb{K}^{\mathcal{I}}$ und $m \in \mathbb{N}_0$. Der m-te Krylow-Raum zu $\mathbf{A}$ und $\mathbf{y}$ ist gegeben durch*

$$\mathcal{K}(\mathbf{y}, m) := \operatorname{span}\{\mathbf{y}, \mathbf{A}\mathbf{y}, \ldots, \mathbf{A}^m\mathbf{y}\}.$$

Offenbar besitzt er höchstens die Dimension $m+1$.

Krylow-Räume besitzen eine enge Beziehung zu Polynomen: Ein Vektor $\mathbf{z} \in \mathcal{K}(\mathbf{y}, m)$ lässt sich nach Definition in der Form

$$\mathbf{z} = \alpha_0\mathbf{y} + \alpha_1\mathbf{A}\mathbf{y} + \ldots + \alpha_m\mathbf{A}^m\mathbf{y}$$

schreiben, und mit dem Polynom

$$p(\xi) := \alpha_0 + \alpha_1\xi + \ldots + \alpha_m\xi^m \qquad \text{für alle } \xi \in \mathbb{K}$$

erhalten wir

$$\mathbf{z} = p(\mathbf{A})\mathbf{y},$$

so dass sich der Krylow-Raum auch als

$$\mathcal{K}(\mathbf{y}, m) = \{p(\mathbf{A})\mathbf{y} \ : \ p \in \Pi_m\} \tag{3.21}$$

darstellen lässt. Ausdrücke dieser Form haben wir schon im Kontext des Tschebyscheff-Verfahrens gesehen.

Bevor wir die Konstruktion (3.19) in eine effizientere Form überführen können, benötigen wir eine alternative Darstellung des von den Suchrichtungen $\mathbf{p}^{(\ell)}$ aufgespannten Vektorraums:

Lemma 3.26 *Für alle $m \in [0 : m_0]$ gilt*

$$\operatorname{span}\{\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(m)}\} = \operatorname{span}\{\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(m)}\} = \mathcal{K}(\mathbf{r}^{(0)}, m). \tag{3.22}$$

Beweis. Wir beweisen zunächst

$$\operatorname{span}\{\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(m)}\} \subseteq \operatorname{span}\{\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(m)}\} \tag{3.23}$$

für alle $m \in [0 : m_0]$ per abschnittsweiser Induktion.

Induktionsanfang: Für $m = 0$ ist die Aussage wegen $\mathbf{p}^{(0)} = \mathbf{r}^{(0)}$ trivial.

Induktionsvoraussetzung: Sei nun $\hat{m} \in [0 : m_0 - 1]$ so gegeben, dass (3.23) für alle $m \in [0 : \hat{m}]$ gilt.

Induktionsschritt: Sei $m := \hat{m} + 1$. Gemäß Konstruktion haben wir

$$\mathbf{p}^{(m)} = \mathbf{r}^{(m)} - \sum_{\ell=0}^{\hat{m}} \frac{\langle \mathbf{r}^{(m)}, \mathbf{p}^{(\ell)} \rangle_A}{\langle \mathbf{p}^{(\ell)}, \mathbf{p}^{(\ell)} \rangle_A} \mathbf{p}^{(\ell)}.$$

Die Induktionsvoraussetzung impliziert

$$\mathbf{p}^{(\ell)} \in \operatorname{span}\{\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(\ell)}\} \qquad \text{für alle } \ell \in [0 : \hat{m}],$$

also folgt (3.23) auch für $m = \hat{m} + 1$.

Als nächstes beweisen wir

$$\operatorname{span}\{\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(m)}\} \subseteq \mathcal{K}(\mathbf{r}^{(0)}, m) \tag{3.24}$$

für alle $m \in [0 : m_0]$, und zwar wieder per abschnittsweiser Induktion.

Induktionsanfang: Der Induktionsbeginn $m = 0$ ist wieder trivial.

Induktionsvoraussetzung: Sei nun $\hat{m} \in [0 : m_0 - 1]$ so gegeben, dass (3.24) für alle $m \in [0 : \hat{m}]$ gilt.

Induktionsschritt: Sei $m := \hat{m} + 1$. Nach Konstruktion von $\mathbf{x}^{(m)}$ haben wir

$$\begin{aligned} \mathbf{r}^{(m)} &= \mathbf{b} - \mathbf{A}\mathbf{x}^{(m)} = \mathbf{b} - \mathbf{A}(\mathbf{x}^{(\hat{m})} + \lambda_{\text{opt}}^{(\hat{m})} \mathbf{p}^{(\hat{m})}) \\ &= \mathbf{r}^{(\hat{m})} - \lambda_{\text{opt}}^{(\hat{m})} \mathbf{A}\mathbf{p}^{(\hat{m})}. \end{aligned}$$

Mit (3.23) erhalten wir

$$\mathbf{p}^{(\hat{m})} \in \operatorname{span}\{\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(\hat{m})}\},$$

also mit Hilfe der Induktionsvoraussetzung auch

$$\mathbf{p}^{(\hat{m})} \in \mathcal{K}(\mathbf{r}^{(0)}, \hat{m})$$

und damit

$$\mathbf{A}\mathbf{p}^{(\hat{m})} \in \mathcal{K}(\mathbf{r}^{(0)}, m).$$

Da nach Induktionsvoraussetzung auch $\mathbf{r}^{(\hat{m})} \in \mathcal{K}(\mathbf{r}^{(0)})$ gilt, ist der Induktionsschritt vollendet.

Sei $m \in [0 : m_0]$. Gelte zunächst $m < m_0$. Dann sind die Vektoren $\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m)}$ nach Definition von Null verschieden. Da sie nach Konstruktion (3.19) auch orthogonal sind, müssen sie linear unabhängig sein. Daraus folgt

$$\dim \operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m)}\} = m + 1,$$

und wegen $\dim \mathcal{K}(\mathbf{r}^{(0)}, m) \leq m + 1$ und den Inklusionen (3.23) und (3.24) die Gleichheit (3.22) der drei betrachteten Räume.

Gelte nun $m = m_0$. Falls $m = 0$ gilt, folgt $\mathbf{p}^{(0)} = \mathbf{r}^{(0)} = \mathbf{0}$, und die Aussage ist trivial.

Sei also $m = m_0 > 0$ vorausgesetzt. Wir haben in Lemma 3.24 bereits gesehen, dass daraus $\mathbf{r}^{(m)} = \mathbf{0}$ folgt. Nach Konstruktion gilt

$$\mathbf{0} = \mathbf{r}^{(m)} = \mathbf{r}^{(m-1)} - \lambda_{\text{opt}}^{(m-1)} \mathbf{A}\mathbf{p}^{(m-1)},$$

also

$$\mathbf{r}^{(m-1)} = \lambda_{\text{opt}}^{(m-1)} \mathbf{A}\mathbf{p}^{(m-1)}.$$

Da $\mathbf{p}^{(m-1)} \neq \mathbf{0}$ aus $\mathbf{r}^{(m-1)}$ konstruiert wird, muss auch $\mathbf{r}^{(m-1)} \neq \mathbf{0}$ gelten, also $\lambda_{\text{opt}}^{(m-1)} \neq 0$, und damit

$$\mathbf{A}\mathbf{p}^{(m-1)} \in \operatorname{span}\{\mathbf{r}^{(m-1)}\} \subseteq \mathcal{K}(\mathbf{r}^{(0)}, m-1) = \operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m-1)}\}.$$

Für alle $\ell \in [0 : m-2]$ erhalten wir mit dem bereits Gezeigten $\mathbf{p}^{(\ell)} \in \mathcal{K}(\mathbf{r}^{(0)}, \ell)$, also

$$\mathbf{A}\mathbf{p}^{(\ell)} \in \mathcal{K}(\mathbf{r}^{(0)}, \ell + 1) \subseteq \operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m-1)}\} \qquad \text{für alle } \ell \in [0 : m-2].$$

Wir haben bereits gezeigt, dass

$$\mathbf{A}^{m-1}\mathbf{r}^{(0)} \in \mathcal{K}(\mathbf{r}^{(0)}, m-1) \subseteq \operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m-1)}\}$$

gilt, also folgt

$$\mathbf{A}^{m}\mathbf{r}^{(0)} \in \operatorname{span}\{\mathbf{A}\mathbf{p}^{(0)}, \ldots, \mathbf{A}\mathbf{p}^{(m-1)}\} \subseteq \operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m-1)}\}.$$

Insgesamt haben wir

$$\mathcal{K}(\mathbf{r}^{(0)}, m) \subseteq \operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m-1)}\} = \operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m-1)}, \mathbf{p}^{(m)}\}$$

bewiesen. ■

Aus diesem Lemma folgt insbesondere, dass m_0 gerade die maximale Dimension eines von $\mathbf{r}^{(0)}$ aufgespannten Krylow-Raums sein muss: Falls $m_0 = 0$ gilt, haben wir $\mathbf{r}^{(0)} = \mathbf{p}^{(0)} = \mathbf{0}$, also hat der Krylow-Raum die Dimension null. Ansonsten gilt

$$\begin{aligned}\mathcal{K}(\mathbf{r}^{(0)}, m_0 - 1) &= \operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m_0-1)}\} \\ &= \operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m_0-1)}, \mathbf{p}^{(m_0)}\} = \mathcal{K}(\mathbf{r}^{(0)}, m_0),\end{aligned}$$

so dass wir unmittelbar

$$m_0 = \max\{\dim \mathcal{K}(\mathbf{r}^{(0)}, m) \ : \ m \in \mathbb{N}_0\} \tag{3.25}$$

erhalten. Insbesondere ist m_0 durch die Dimension des Raums $\mathbb{K}^{\mathcal{I}}$ beschränkt, also durch die Anzahl der Unbekannten.

Bemerkung 3.27 (Invarianter Teilraum) *Es ist zu beachten, dass m_0 von der Wahl unseres Startvektors abhängt. Angenommen, das Startresiduum $\mathbf{r}^{(0)}$ liegt in einem invariaten Teilraum $\mathcal{V} \subseteq \mathbb{K}^{\mathcal{I}}$, also in einem Raum, der*

$$\mathbf{A}\mathbf{z} \in \mathcal{V} \qquad \text{für alle } \mathbf{z} \in \mathcal{V}$$

erfüllt. Dann gilt auch

$$\mathcal{K}(\mathbf{r}^{(0)}, m) \subseteq \mathcal{V} \qquad \text{für alle } m \in \mathbb{N}_0,$$

so dass

$$m_0 = \max\{\dim \mathcal{K}(\mathbf{r}^{(0)}, m) \ : \ m \in \mathbb{N}_0\} \leq \dim \mathcal{V}$$

gilt. In dieser Situation wird also die exakte Lösung nach höchstens $k := \dim \mathcal{V}$ Schritten berechnet worden sein.

Wenden wir uns nun wieder der Definition (3.19) der Suchrichtung $\mathbf{p}^{(m)}$ zu. Um die Effizienz unseres Verfahrens zu verbessern, ist es unser Ziel, die Formel

$$\mathbf{p}^{(m)} = \mathbf{r}^{(m)} - \sum_{k=0}^{m-1} \frac{\langle \mathbf{p}^{(k)}, \mathbf{r}^{(m)} \rangle_A}{\langle \mathbf{p}^{(k)}, \mathbf{p}^{(k)} \rangle_A} \mathbf{p}^{(\ell)}$$

zu vereinfachen, insbesondere die Summe zu verkürzen. Dazu untersuchen wir die im Zähler der Summanden auftretenden Energieskalarprodukte

$$\langle \mathbf{p}^{(k)}, \mathbf{r}^{(m)} \rangle_A = \langle \mathbf{p}^{(k)}, \mathbf{A}\mathbf{r}^{(m)} \rangle_2$$

für $k \in [0 : m-1]$. An dieser Stelle ist von entscheidender Bedeutung, dass die Matrix $\mathbf{A}$ selbstadjungiert ist, denn diese Tatsache erlaubt es uns, die Gleichung in die Form

$$\langle \mathbf{p}^{(k)}, \mathbf{A}\mathbf{r}^{(m)} \rangle_2 = \langle \mathbf{A}^* \mathbf{p}^{(k)}, \mathbf{r}^{(m)} \rangle_2 = \langle \mathbf{A}\mathbf{p}^{(k)}, \mathbf{r}^{(m)} \rangle_2$$

zu überführen. Nach Lemma 3.26 gilt $\mathbf{p}^{(k)} \in \mathcal{K}(\mathbf{r}^{(0)}, k)$, und nach Definition der Krylow-Räume muss dann auch

$$\mathbf{A}\mathbf{p}^{(k)} \in \mathcal{K}(\mathbf{r}^{(0)}, k+1)$$

gelten. Für $k < m-1$ folgt $k+1 \leq m-1$ und damit nach Lemma 3.26 auch

$$\mathbf{A}\mathbf{p}^{(k)} \in \operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m-1)}\}.$$

Nach unserer Konstruktion ist $\mathbf{x}^{(m)}$ optimal bezüglich der Richtungen $\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m-1)}$, also gilt dank (3.18) insbesondere

$$\langle \mathbf{y}, \mathbf{r}^{(m)} \rangle_2 = -\langle \mathbf{y}, \mathbf{A}\mathbf{x}^{(m)} - \mathbf{b} \rangle_2 = 0 \qquad \text{für alle } \mathbf{y} \in \operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m-1)}\}.$$

Indem wir diese Gleichung auf $\mathbf{y} = \mathbf{A}\mathbf{p}^{(k)}$ anwenden, erhalten wir

$$\langle \mathbf{p}^{(k)}, \mathbf{r}^{(m)} \rangle_A = \langle \mathbf{A}\mathbf{p}^{(k)}, \mathbf{r}^{(m)} \rangle_2 = 0 \qquad \text{für alle } k \in [0 : m-2],$$

procedure KonjGrad(**b**, **var x**);
$\mathbf{r} \leftarrow \mathbf{b} - \mathbf{A}\mathbf{x}$;
$\mathbf{p} \leftarrow \mathbf{r}$;
while Fehler zu groß und $\mathbf{p} \neq \mathbf{0}$ **do**
  $\mathbf{a} \leftarrow \mathbf{A}\mathbf{p}$;
  $\lambda_{\text{opt}} \leftarrow \frac{\langle \mathbf{p},\mathbf{r}\rangle_2}{\langle \mathbf{p},\mathbf{a}\rangle_2}$;
  $\mathbf{x} \leftarrow \mathbf{x} + \lambda_{\text{opt}}\mathbf{p}$;
  $\mathbf{r} \leftarrow \mathbf{r} - \lambda_{\text{opt}}\mathbf{a}$;
  $\mu \leftarrow \frac{\langle \mathbf{a},\mathbf{r}\rangle_2}{\langle \mathbf{p},\mathbf{a}\rangle_2}$;
  $\mathbf{p} \leftarrow \mathbf{r} - \mu\mathbf{p}$
end while

Abbildung 3.5: Verfahren der konjugierten Gradienten

in unserer Summe werden also alle Summanden mit Ausnahme des letzten wegfallen, und die Formel (3.19) reduziert sich auf

$$\begin{aligned}\mathbf{p}^{(m)} &:= \mathbf{r}^{(m)} - \frac{\langle \mathbf{p}^{(m-1)}, \mathbf{r}^{(m)}\rangle_A}{\langle \mathbf{p}^{(m-1)}, \mathbf{p}^{(m-1)}\rangle_A}\mathbf{p}^{(m-1)} \\ &= \mathbf{r}^{(m)} - \frac{\langle \mathbf{A}\mathbf{p}^{(m-1)}, \mathbf{r}^{(m)}\rangle_2}{\langle \mathbf{p}^{(m-1)}, \mathbf{A}\mathbf{p}^{(m-1)}\rangle_2}\mathbf{p}^{(m-1)}. \end{aligned} \tag{3.26}$$

Es ist also, wie schon bei der Tschebyscheff-Semiiteration, nicht erforderlich, alle Zwischenergebnisse zu speichern, es genügen die Vektoren des unmittelbar vorhergehenden Schrittes.

Da die Suchrichtung des vorhergehenden Schritts explizit in die Berechnung der nächsten Iterierten eingeht, können wir das neue Verfahren nicht direkt als Iterationsverfahren darstellen.

Definition 3.28 (cg-Verfahren) *Sei* $\mathbf{A}$ *positiv definit. Seien* $\mathbf{x}^{(0)} \in \mathbb{K}^{\mathcal{I}}$ *und* $\mathbf{b} \in \mathbb{K}^{\mathcal{I}}$ *gegeben. Die durch*

$$\begin{aligned}\mathbf{r}^{(m)} &:= \mathbf{b} - \mathbf{A}\mathbf{x}^{(m)}, \\ \mathbf{p}^{(m)} &:= \begin{cases} \mathbf{r}^{(0)} & \textit{falls } m = 0, \\ \mathbf{r}^{(m)} - \frac{\langle \mathbf{A}\mathbf{p}^{(m-1)},\mathbf{r}^{(m)}\rangle_2}{\langle \mathbf{p}^{(m-1)},\mathbf{A}\mathbf{p}^{(m-1)}\rangle_2}\mathbf{p}^{(m-1)} & \textit{falls } m > 0 \textit{ und } \mathbf{p}^{(m-1)} \neq \mathbf{0}, \\ \mathbf{0} & \textit{sonst}, \end{cases} \\ \mathbf{x}^{(m+1)} &:= \mathbf{x}^{(m)} + \frac{\langle \mathbf{p}^{(m)}, \mathbf{r}^{(m)}\rangle_2}{\langle \mathbf{p}^{(m)}, \mathbf{A}\mathbf{p}^{(m)}\rangle_2}\mathbf{p}^{(m)} \qquad \textit{für alle } m \in \mathbb{N}_0 \end{aligned}$$

definierte Folge $(\mathbf{x}^{(m)})_{m\in\mathbb{N}_0}$ *bezeichnen wir als die Folge der Semiiterierten des* Verfahrens der konjugierten Gradienten *(oder kurz des* cg-Verfahrens*, von* conjugate gradients*).*

Ein praktischer Algorithmus zur Berechnung der Folge der Semiiterierten ist in Abbildung 3.5 gegeben. Um unnötige Matrix-Vektor-Multiplikationen zu vermeiden, verwenden wir wie im Falle des Gradientenverfahrens einen Hilfsvektor $\mathbf{a}$, der jeweils das

Produkt von $\mathbf{A}$ mit der aktuellen Suchrichtung $\mathbf{p}$ speichert und zur Aktualisierung des Residuums $\mathbf{r}$ sowie zur Berechnung der verschiedenen Energieskalarprodukte benutzt wird. Die gegebene Variante des Verfahrens benötigt also neben der rechten Seite $\mathbf{b}$ und der Approximation $\mathbf{x}$ der Lösung drei Hilfsvektoren: Das Residuum $\mathbf{r}$, die Suchrichtung $\mathbf{p}$ und das Produkt $\mathbf{a}$ von $\mathbf{A}$ mit dieser Suchrichtung.

Wenden wir uns nun der Fehleranalyse für das cg-Verfahren zu. Gemäß unserer Konstruktion ist die m-te „Iterierte“ $\mathbf{x}^{(m)}$ optimal bezüglich der Richtungen $\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(m-1)}$. Diese Aussage lässt sich auf den von diesen Richtungen aufgespannten Unterraum übertragen:

Lemma 3.29 *Sei* $\mathbf{A}$ *positiv definit, sei* $\mathbf{b} \in \mathbb{K}^{\mathcal{I}}$ *und* $\mathbf{x}^* := \mathbf{A}^{-1}\mathbf{b}$. *Sei* $(\mathbf{x}^{(m)})_{m\in\mathbb{N}_0}$ *die Folge der Semiiterierten des cg-Verfahrens zu einem Startvektor* $\mathbf{x}^{(0)} \in \mathbb{K}^{\mathcal{I}}$. *Dann gilt*

$$f(\mathbf{x}^{(m)}) \le f(\mathbf{x}^{(m)} + \mathbf{y}) \qquad \textit{für alle } \mathbf{y} \in \mathcal{K}(\mathbf{r}^{(0)}, m-1) \textit{ und alle } m \in [0 : m_0]. \tag{3.27}$$

Insbesondere gilt

$$\|\mathbf{x}^* - \mathbf{x}^{(m)}\|_A = \min\{\|q(\mathbf{A})(\mathbf{x}^* - \mathbf{x}^{(0)})\|_A \ : \ q \in \Pi_m,\ q(0) = 1\} \qquad \textit{für alle } m \in [0 : m_0]. \tag{3.28}$$

Beweis. Sei $m \in [0 : m_0]$. Die Gleichung (3.27) folgt aus unserer Konstruktion: Wir haben die Suchrichtungen $\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(m-1)}$ gerade so konstruiert, dass sie bezüglich des Energieskalarprodukts paarweise orthogonal sind. Daraus folgt, dass $\mathbf{x}^{(m)}$ optimal bezüglich aller dieser Suchrichtungen ist, also nach (3.18) bereits

$$\langle \mathbf{A}\mathbf{x}^{(m)} - \mathbf{b}, \mathbf{p}^{(\ell)} \rangle_2 = 0 \qquad \text{für alle } \ell \in [0 : m-1]. \tag{3.29}$$

Mit der Linearität des Skalarprodukts überträgt sich die Optimalität auf den gesamten Teilraum $\operatorname{span}\{\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(m-1)}\} = \mathcal{K}(\mathbf{r}^{(0)}, m-1)$.

Nach der Konstruktion unseres Verfahrens existieren $\lambda^{(0)}, \dots, \lambda^{(m-1)} \in \mathbb{K}$ so, dass die Gleichungen

$$\mathbf{x}^{(\ell+1)} = \mathbf{x}^{(\ell)} + \lambda^{(\ell)}\mathbf{p}^{(\ell)} \qquad \text{für alle } \ell \in [0 : m-1],$$

gelten, also insbesondere auch

$$\mathbf{x}^{(m)} - \mathbf{x}^{(0)} = \sum_{\ell=0}^{m-1} \lambda^{(\ell)}\mathbf{p}^{(\ell)} \in \mathcal{K}(\mathbf{r}^{(0)}, m-1).$$

Damit existiert wegen (3.21) ein Polynom $\widehat{q}_m \in \Pi_{m-1}$ mit

$$\mathbf{x}^{(m)} - \mathbf{x}^{(0)} = \widehat{q}_m(\mathbf{A})\mathbf{r}^{(0)},$$

und für den Fehler folgt

$$\mathbf{x}^* - \mathbf{x}^{(m)} = \mathbf{x}^* - \mathbf{x}^{(0)} - \widehat{q}_m(\mathbf{A})\mathbf{r}^{(0)}. \tag{3.30}$$

Mit der Gleichung

$$\mathbf{r}^{(0)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)} = \mathbf{A}\mathbf{x}^* - \mathbf{A}\mathbf{x}^{(0)} = \mathbf{A}(\mathbf{x}^* - \mathbf{x}^{(0)})$$

folgt schließlich

$$\mathbf{x}^* - \mathbf{x}^{(m)} = \mathbf{x}^* - \mathbf{x}^{(0)} - \widehat{q}_m(\mathbf{A})\mathbf{A}(\mathbf{x}^* - \mathbf{x}^{(0)}) = q_m(\mathbf{A})(\mathbf{x}^* - \mathbf{x}^{(0)})$$

mit dem Polynom $q_m(\xi) = 1 - \widehat{q}_m(\xi)\xi$. Dieses Polynom erfüllt $q_m \in \Pi_m$ und $q_m(0) = 1$.

Sei nun $q \in \Pi_m$ beliebig mit $q(0) = 1$. Dann ist null eine Nullstelle des Polynoms $1-q$, die wir herausdividieren können, um ein Polynom $\widehat{q} \in \Pi_{m-1}$ mit $1 - q(\xi) = \widehat{q}(\xi)\xi$, also $q(\xi) = 1 - \widehat{q}(\xi)\xi$ zu erhalten. Es folgt

$$q(\mathbf{A})(\mathbf{x}^* - \mathbf{x}^{(0)}) = \mathbf{x}^* - \mathbf{x}^{(0)} - \widehat{q}(\mathbf{A})\mathbf{A}(\mathbf{x}^* - \mathbf{x}^{(0)}) = \mathbf{x}^* - \mathbf{x}^{(0)} - \widehat{q}(\mathbf{A})\mathbf{r}^{(0)}. \tag{3.31}$$

Indem wir die Gleichung (3.31) von der Gleichung (3.30) subtrahieren, erhalten wir

$$\mathbf{y} := \mathbf{x}^* - \mathbf{x}^{(m)} - q(\mathbf{A})(\mathbf{x}^* - \mathbf{x}^{(0)}) = \widehat{q}(\mathbf{A})\mathbf{r}^{(0)} - \widehat{q}_m(\mathbf{A})\mathbf{r}^{(0)} \in \mathcal{K}(\mathbf{r}^{(0)}, m-1).$$

Mit (3.11) und (3.27) folgt aus dieser Gleichung

$$\begin{aligned}
\|q(\mathbf{A})(\mathbf{x}^* - \mathbf{x}^{(0)})\|_A^2 &= \|\mathbf{x}^* - (\mathbf{x}^{(m)} + \mathbf{y})\|_A^2 = 2f(\mathbf{x}^{(m)} + \mathbf{y}) + \langle \mathbf{x}^*, \mathbf{A}\mathbf{x}^*\rangle_2 \\
&\geq 2f(\mathbf{x}^{(m)}) + \langle \mathbf{x}^*, \mathbf{A}\mathbf{x}^*\rangle_2 = \|\mathbf{x}^* - \mathbf{x}^{(m)}\|_A^2 \\
&= \|q_m(\mathbf{A})(\mathbf{x}^* - \mathbf{x}^{(0)})\|_A^2.
\end{aligned}$$

Also wird das Minimum der rechten Seite in (3.28) für $q = q_m$ angenommen. ■

Der Beweis der Konvergenz des Gradientenverfahrens basiert darauf, dass wir in Lemma 3.18 nachweisen, dass es mindestens so gut wie ein optimal gedämpftes Richardson-Verfahren ist. Lemma 3.29 ermöglicht es uns nun, nachzuweisen, dass das cg-Verfahren mindestens so gut wie eine optimale konsistente Semiiteration ist, die auf dem Richardson-Verfahren fußt. Dieser Ansatz führt direkt zu dem folgenden Konvergenzresultat:

Satz 3.30 (Konvergenz) *Sei* $\mathbf{A}$ *positiv definit. Seien* $\alpha, \beta \in \mathbb{R}_{>0}$ *mit* $\sigma(\mathbf{A}) \subseteq [\alpha, \beta]$. *Sei* $\mathbf{b} \in \mathbb{K}^{\mathcal{I}}$, *und sei* $\mathbf{x}^* := \mathbf{A}^{-1}\mathbf{b}$. *Sei* $(\mathbf{x}^{(m)})_{m\in\mathbb{N}_0}$ *die Folge der Semiiterierten des cg-Verfahrens zu einem Startvektor* $\mathbf{x}^{(0)} \in \mathbb{K}^{\mathcal{I}}$. *Dann gilt*

$$\|\mathbf{x}^{(m)} - \mathbf{x}^*\|_A \leq \frac{2c^m}{1 + c^{2m}}\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_A \qquad \text{für alle } m \in \mathbb{N}_0$$

mit der bereits aus (3.9) bekannten Konstanten

$$c := \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}, \qquad \kappa := \frac{\beta}{\alpha}.$$

Beweis. Wir setzen $\theta := 2/(\beta + \alpha)$ und rechnen wie in Abschnitt 3.2 nach, dass $\sigma(\mathbf{M}_{\mathrm{Rich},\theta}) \subseteq [a, b]$ für

$$a = 1 - 2\frac{\beta}{\beta + \alpha}, \qquad b = 1 - 2\frac{\alpha}{\beta + \alpha}$$

gilt, so dass wir $\kappa = \beta/\alpha$ erhalten. Sei p_m das in (3.7) definierte transformierte Tschebyscheff-Polynom, und sei

$$q_m(\xi) := p_m(1 - \theta\xi) \qquad \text{für alle } \xi \in \mathbb{K}. \tag{3.32}$$

Dann erhalten wir

$$q_m(\mathbf{A}) = p_m(\mathbf{I} - \theta\mathbf{A}) = p_m(\mathbf{M}_{\mathrm{Rich},\theta}),$$

also auch

$$\begin{aligned}\|q_m(\mathbf{A})(\mathbf{x}^{(0)} - \mathbf{x}^*)\|_A &\le \|q_m(\mathbf{A})\|_A \|\mathbf{x}^{(0)} - \mathbf{x}^*\|_A = \|p_m(\mathbf{M}_{\mathrm{Rich},\theta})\|_A \|\mathbf{x}^{(0)} - \mathbf{x}^*\|_A \\ &\le \max\{|p_m(\mu)| \ : \ \mu \in [a, b]\} \|\mathbf{x}^{(0)} - \mathbf{x}^*\|_A = \frac{1}{C_m} \|\mathbf{x}^{(0)} - \mathbf{x}^*\|_A.\end{aligned}$$

Indem wir die in Lemma 3.9 angegebene Abschätzung für C_m einsetzen erhalten wir das gewünschte Resultat. ■

Falls wir das Spektrum der Matrix $\mathbf{A}$ genauer eingrenzen können, lässt ich die Aussage des vorigen Satzes erheblich verbessern.

Bemerkung 3.31 (Extreme Eigenwerte) *Da wir bei der Wahl des Polynoms q in Lemma 3.29 relativ frei sind, können wir beispielsweise einzelne extreme Eigenwerte einer Sonderbehandlung unterziehen. Falls etwa $\sigma(\mathbf{A}) \subseteq [\alpha, \gamma] \cup \{\beta\}$ für ein $\gamma < \beta$ gilt, kann es sich auszahlen, ein Polynom der Form*

$$q_k(\xi) := \frac{\beta - \xi}{\beta} p_{m-1}(1 - \theta\xi) \qquad \textit{für alle } \xi \in \mathbb{K}$$

zu betrachten, wobei das transformierte Tschebyscheff-Polynom p_{m-1} so gewählt wird, dass es lediglich auf dem Intervall $[\alpha, \gamma]$ kleine Werte annimmt, schließlich verschwindet q_k in β ohnehin. Da $(\beta - \xi)/\beta \le 1$ für alle $\xi \in [\alpha, \beta]$ gilt, würde die Konvergenzrate nur von dem kleinen Intervall $[\alpha, \gamma]$ abhängen, aber nicht von dem größeren Intervall $[\alpha, \beta]$, das das gesamte Spektrum einschließt.

Es ist zu beachten, dass diese Konstruktionen ausschließlich für die theoretische Analyse des Verfahrens wichtig sind, denn die Abschätzung aus Lemma 3.29 gilt für alle *Polynome q_m mit $q_m(0) = 1$.*

Wie schon im Falle des Gradientenverfahrens sind wir wieder daran interessiert, das Verfahren der konjugierten Gradienten zu beschleunigen, indem wir das ursprüngliche Gleichungssystem (1.1) durch das vorkonditionierte System (3.16) ersetzen, also durch

$$\widehat{\mathbf{A}}\widehat{\mathbf{x}} = \widehat{\mathbf{b}}$$

für die transformierten Vektoren

$$\widehat{\mathbf{A}} := \mathbf{N}^{1/2}\mathbf{A}\mathbf{N}^{1/2}, \qquad \widehat{\mathbf{b}} := \mathbf{N}^{1/2}\mathbf{b}, \qquad \widehat{\mathbf{x}} := \mathbf{N}^{-1/2}\mathbf{x}.$$

Die einzelnen Iterierten können wir durch

$$\mathbf{x}^{(m)} := \mathbf{N}^{1/2}\widehat{\mathbf{x}}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0$$

rekonstruieren. Wie schon im Fall des vorkonditionierten Gradientenverfahrens (vgl. Definition 3.20) ist das vorkonditionierte cg-Verfahren in der Regel nur dann effizient implementierbar, wenn wir es durchführen können, ohne die Matrizen $\mathbf{N}^{1/2}$ und $\mathbf{N}^{-1/2}$ explizit zur Verfügung zu haben, wir müssen also eine geeignete Umformulierung finden. Zu diesem Zweck führen wir vorkonditionierte Residuen und Suchrichtungen durch

$$\mathbf{q}^{(m)} := \mathbf{N}^{1/2}\widehat{\mathbf{r}}^{(m)} = \mathbf{N}\mathbf{r}^{(m)}, \qquad \mathbf{p}^{(m)} := \mathbf{N}^{1/2}\widehat{\mathbf{p}}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0$$

ein. Für diese Vektoren gilt

$$\begin{aligned}
\langle \widehat{\mathbf{A}}\widehat{\mathbf{p}}^{(m-1)}, \widehat{\mathbf{r}}^{(m)}\rangle_2 &= \langle \mathbf{N}^{1/2}\mathbf{A}\mathbf{N}^{1/2}\mathbf{N}^{-1/2}\mathbf{p}^{(m-1)}, \mathbf{N}^{1/2}\mathbf{r}^{(m)}\rangle_2 = \langle \mathbf{A}\mathbf{p}^{(m-1)}, \mathbf{q}^{(m)}\rangle_2,\\
\langle \widehat{\mathbf{p}}^{(m)}, \widehat{\mathbf{A}}\widehat{\mathbf{p}}^{(m)}\rangle_2 &= \langle \mathbf{N}^{-1/2}\mathbf{p}^{(m)}, \mathbf{N}^{1/2}\mathbf{A}\mathbf{N}^{1/2}\mathbf{N}^{-1/2}\mathbf{p}^{(m)}\rangle_2 = \langle \mathbf{p}^{(m)}, \mathbf{A}\mathbf{p}^{(m)}\rangle_2,\\
\langle \widehat{\mathbf{p}}^{(m)}, \widehat{\mathbf{r}}^{(m)}\rangle_2 &= \langle \mathbf{N}^{-1/2}\mathbf{p}^{(m)}, \mathbf{N}^{1/2}\mathbf{r}^{(m)}\rangle_2 = \langle \mathbf{p}^{(m)}, \mathbf{r}^{(m)}\rangle_2,
\end{aligned}$$

also können wir die im vorkonditionierten cg-Verfahren auftretenden Skalarprodukte berechnen, ohne explizit auf die Matrix $\mathbf{N}^{1/2}$ zurückgreifen zu müssen.

Definition 3.32 (Vorkonditioniertes cg-Verfahren) *Seien* $\mathbf{A}$ *und* $\mathbf{N}$ *positiv definit. Seien* $\mathbf{x}^{(0)} \in \mathbb{K}^{\mathcal{I}}$ *und* $\mathbf{b} \in \mathbb{K}^{\mathcal{I}}$ *gegeben. Die durch*

$$\begin{aligned}
\mathbf{r}^{(m)} &:= \mathbf{b} - \mathbf{A}\mathbf{x}^{(m)},\\
\mathbf{q}^{(m)} &:= \mathbf{N}\mathbf{r}^{(m)},\\
\mathbf{p}^{(m)} &:= \begin{cases} \mathbf{q}^{(0)} & \textit{falls } m = 0,\\ \mathbf{q}^{(m)} - \frac{\langle \mathbf{A}\mathbf{p}^{(m-1)}, \mathbf{q}^{(m)}\rangle_2}{\langle \mathbf{p}^{(m-1)}, \mathbf{A}\mathbf{p}^{(m-1)}\rangle_2}\mathbf{p}^{(m-1)} & \textit{falls } m > 0 \textit{ und } \mathbf{p}^{(m-1)} \neq \mathbf{0},\\ \mathbf{0} & \textit{sonst} \end{cases},\\
\mathbf{x}^{(m+1)} &:= \mathbf{x}^{(m)} + \frac{\langle \mathbf{p}^{(m)}, \mathbf{r}^{(m)}\rangle_2}{\langle \mathbf{p}^{(m)}, \mathbf{A}\mathbf{p}^{(m)}\rangle_2}\mathbf{p}^{(m)} \qquad \textit{für alle } m \in \mathbb{N}_0
\end{aligned}$$

definierte Folge $(\mathbf{x}^{(m)})_{m\in\mathbb{N}_0}$ *bezeichnen wir als die Folge der Semiiterierten des* vorkonditionierten Verfahrens der konjugierten Gradienten *(oder kurz des* vorkonditionierten cg-Verfahrens*). Die Matrix* $\mathbf{N}$ *bezeichnen wir in diesem Kontext als* Vorkonditionierer*.*

Wenn wir wieder den Hilfsvektor $\mathbf{a} := \mathbf{A}\mathbf{p}$ einführen und die Vektoren aus dem vorangehenden Iterationsschritt durch die des aktuellen überschreiben, erhalten wir den in Abbildung 3.6 angegebenen Algorithmus. Im Vergleich zum ursprünglichen cg-Verfahren

procedure VorkondKonjGrad(**b**, **var x**);
$\mathbf{r} \leftarrow \mathbf{b} - \mathbf{A}\mathbf{x}$;
$\mathbf{q} \leftarrow \mathbf{N}\mathbf{r}$;
$\mathbf{p} \leftarrow \mathbf{q}$;
while Fehler zu groß und $\mathbf{p} \neq \mathbf{0}$ **do**
  $\mathbf{a} \leftarrow \mathbf{A}\mathbf{p}$;
  $\lambda_{\text{opt}} \leftarrow \frac{\langle \mathbf{p},\mathbf{r}\rangle_2}{\langle \mathbf{p},\mathbf{a}\rangle_2}$;
  $\mathbf{x} \leftarrow \mathbf{x} + \lambda_{\text{opt}}\mathbf{p}$;
  $\mathbf{r} \leftarrow \mathbf{r} - \lambda_{\text{opt}}\mathbf{a}$;
  $\mathbf{q} \leftarrow \mathbf{N}\mathbf{r}$;
  $\mu \leftarrow \frac{\langle \mathbf{a},\mathbf{q}\rangle_2}{\langle \mathbf{p},\mathbf{a}\rangle_2}$;
  $\mathbf{p} \leftarrow \mathbf{q} - \mu\mathbf{p}$
end while

Abbildung 3.6: Vorkonditioniertes Verfahren der konjugierten Gradienten

benötigt er den zusätzlichen Hilfsvektor $\mathbf{q}$, abgesehen davon ist die Struktur der beiden Verfahren sehr ähnlich.

Wenden wir uns nun der Analyse der Konvergenzgeschwindigkeit des vorkonditionierten cg-Verfahrens zu. Wie schon im Fall des Gradientenverfahrens können wir ausnutzen, dass uns Satz 3.30 bereits eine Abschätzung für den nicht vorkonditionierten Fall zur Verfügung stellt. Es bleibt lediglich nachzuprüfen, dass sich bei der Anwendung dieses Resultats auf den vorkonditionierten Fall die richtigen Normen ergeben.

Satz 3.33 (Konvergenz) *Seien* $\mathbf{A}$ *und* $\mathbf{N}$ *positiv definit. Seien* $\alpha, \beta \in \mathbb{R}_{>0}$ *gegeben mit* $\sigma(\mathbf{N}\mathbf{A}) \subseteq [\alpha, \beta]$. *Sei* $\mathbf{b} \in \mathbb{K}^{\mathcal{I}}$, *und sei* $\mathbf{x}^* := \mathbf{A}^{-1}\mathbf{b}$. *Dann gilt*

$$\|\mathbf{x}^{(m)} - \mathbf{x}^*\|_A \leq \frac{2c^m}{1+c^{2m}}\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_A \qquad \textit{für alle } m \in \mathbb{N}_0$$

mit den Konstanten

$$c := \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}, \qquad \kappa := \frac{\beta}{\alpha}.$$

Beweis. Sei $m \in \mathbb{N}_0$. Wir setzen $\widehat{\mathbf{x}}^* := \mathbf{N}^{-1/2}\mathbf{x}^*$ und erhalten

$$\begin{aligned}\|\mathbf{x}^{(m)} - \mathbf{x}^*\|_A^2 &= \langle \mathbf{A}(\mathbf{x}^{(m)} - \mathbf{x}^*), \mathbf{x}^{(m)} - \mathbf{x}^*\rangle_2 = \langle \mathbf{A}\mathbf{N}^{1/2}(\widehat{\mathbf{x}}^{(m)} - \widehat{\mathbf{x}}^*), \mathbf{N}^{1/2}(\widehat{\mathbf{x}}^{(m)} - \widehat{\mathbf{x}}^*)\rangle_2 \\ &= \langle \mathbf{N}^{1/2}\mathbf{A}\mathbf{N}^{1/2}(\widehat{\mathbf{x}}^{(m)} - \widehat{\mathbf{x}}^*), \widehat{\mathbf{x}}^{(m)} - \widehat{\mathbf{x}}^*\rangle_2 \\ &= \langle \widehat{\mathbf{A}}, (\widehat{\mathbf{x}}^{(m)} - \widehat{\mathbf{x}}^*), \widehat{\mathbf{x}}^{(m)} - \widehat{\mathbf{x}}^*\rangle_2 = \|\widehat{\mathbf{x}}^{(m)} - \widehat{\mathbf{x}}^*\|_{\widehat{A}}^2.\end{aligned}$$

Offenbar gilt $\widehat{\mathbf{A}}\widehat{\mathbf{x}}^* = \mathbf{N}^{1/2}\mathbf{A}\mathbf{N}^{1/2}\mathbf{N}^{-1/2}\mathbf{x}^* = \mathbf{N}^{1/2}\mathbf{b} = \widehat{\mathbf{b}}$, also können wir Satz 3.30 auf das vorkonditionierte Gleichungssystem (3.16) anwenden, um

$$\|\mathbf{x}^{(m)} - \mathbf{x}^*\|_A = \|\widehat{\mathbf{x}}^{(m)} - \widehat{\mathbf{x}}^*\|_{\widehat{A}} \leq \frac{2c^m}{1+c^{2m}}\|\widehat{\mathbf{x}}^{(0)} - \widehat{\mathbf{x}}^*\|_{\widehat{A}} = \frac{2c^m}{1+c^{2m}}\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_A$$

zu erhalten und den Beweis abzuschließen. ■

Auch hier gilt wieder, dass die Bedingung $\sigma(\mathbf{NA}) \subseteq [\alpha, \beta]$ äquivalent zu

$$\alpha\mathbf{W} \leq \mathbf{A} \leq \beta\mathbf{W}$$

mit der Matrix $\mathbf{W} := \mathbf{N}^{-1}$ ist. Wenn also $\mathbf{W}$ eine gute Approximation von $\mathbf{A}$ ist, dürfen wir darauf hoffen, dass das vorkonditionierte cg-Verfahren schnell konvergieren wird. Die Skalierung von $\mathbf{W}$ spielt dabei keine Rolle, da die Konvergenzgeschwindigkeit nur vom Quotienten $\kappa = \beta/\alpha$ abhängt.

3.5 Krylow-Verfahren für nicht positiv definite Matrizen

Bei unseren bisherigen Untersuchungen war es ausschlaggebend, dass die Matrix $\mathbf{A}$, und bei vorkonditionierten Verfahren auch die Matrix $\mathbf{N}$, positiv definit sind. Beispielsweise beziehen sich alle Konvergenzaussagen auf die Energienorm, die nur dann sinnvoll ist, wenn $\mathbf{A}$ positiv definit ist.

Wir werden nun Verfahren einführen, die auch dann noch funktionieren, wenn $\mathbf{A}$ nicht positiv definit ist, die aber so viele der guten Eigenschaften von Krylow-Verfahren wie möglich erhalten.

Die Grundidee ist dieselbe wie im Falle des cg-Verfahrens: Wir benötigen Basisvektoren $\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m-1)}$ für die Krylow-Unterräume $\mathcal{K}(\mathbf{r}^{(0)}, m-1)$, und diese Basen sollten Eigenschaften besitzen, die es uns ermöglichen, die jeweils bestmögliche Approximation $\mathbf{x}^{(m)}$ der Lösung $\mathbf{x}^*$ zu berechnen.

Wir stehen also vor der Aufgabe, für eine geeignete Vektornorm $\|\cdot\|$ einen Vektor $\mathbf{z}^{(m)} \in \mathcal{K}(\mathbf{r}^{(0)}, m-1) = \operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m-1)}\}$ so zu finden, dass für die m-te Iterierte

$$\mathbf{x}^{(m)} := \mathbf{x}^{(0)} + \mathbf{z}^{(m)}$$

die Optimalitätseigenschaft

$$\|\mathbf{x}^{(m)} - \mathbf{x}^*\| \leq \|\mathbf{x}^{(m)} + \mathbf{y} - \mathbf{x}^*\| \qquad \text{für alle } \mathbf{y} \in \mathcal{K}(\mathbf{r}^{(0)}, m-1)$$

erfüllt ist. Da wir $\mathbf{x}^*$ nicht kennen, können wir diese Eigenschaft für allgemeine Normen nicht sicherstellen, weil wir nicht einmal die Norm auswerten können. Wenn wir allerdings zu der *Defektnorm*

$$\mathbf{x} \mapsto \|\mathbf{Ax}\|_2$$

übergehen, wird die Aufgabe lösbar: Für alle $\mathbf{x} \in \mathbb{K}^{\mathcal{I}}$ gilt

$$\|\mathbf{A}(\mathbf{x} - \mathbf{x}^*)\|_2 = \|\mathbf{Ax} - \mathbf{b}\|_2,$$

und die Optimalitätseigenschaft nimmt die Form

$$\|\mathbf{Ax}^{(m)} - \mathbf{b}\|_2 \leq \|\mathbf{Ax}^{(m)} + \mathbf{Ay} - \mathbf{b}\|_2 \qquad \text{für alle } \mathbf{y} \in \mathcal{K}(\mathbf{r}^{(0)}, m-1)$$

an. Wenn wir $\mathbf{x}^{(m)}$ durch $\mathbf{z}^{(m)}$ ausdrücken, erhalten wir

$$\mathbf{A}\mathbf{x}^{(m)} - \mathbf{b} = \mathbf{A}\mathbf{x}^{(0)} + \mathbf{A}\mathbf{z}^{(m)} - \mathbf{b} = \mathbf{A}\mathbf{z}^{(m)} - (\mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}) = \mathbf{A}\mathbf{z}^{(m)} - \mathbf{r}^{(0)},$$

mit dem Residuum $\mathbf{r}^{(0)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$, und die Optimalitätseigenschaft lässt sich als

$$\|\mathbf{A}\mathbf{z}^{(m)} - \mathbf{r}^{(0)}\|_2 \leq \|\mathbf{A}\mathbf{y} - \mathbf{r}^{(0)}\|_2 \qquad \text{für alle } \mathbf{y} \in \mathcal{K}(\mathbf{r}^{(0)}, m-1) \tag{3.33}$$

schreiben, also als ein lineares Ausgleichsproblem. Wir beobachten, dass

$$\mathbf{z}^{(m)} \in \mathcal{K}(\mathbf{r}^{(0)}, m-1), \qquad \mathbf{A}\mathbf{z}^{(m)} - \mathbf{r}^{(0)} \in \mathcal{K}(\mathbf{r}^{(0)}, m) \tag{3.34}$$

gelten, demzufolge müssen wir $\mathbf{z}^{(m)}$ aus dem m-dimensionalen Raum $\mathcal{K}(\mathbf{r}^{(0)}, m-1)$ so wählen, dass eine Differenz in dem $(m+1)$-dimensionalen Raum $\mathcal{K}(\mathbf{r}^{(0)}, m)$ minimiert wird.

Falls $m \ll n = \#\mathcal{I}$ gilt, ist es nicht erstrebenswert, das Ausgleichsproblem direkt zu behandeln. Wesentlich vorteilhafter wäre es, das Problem in eine niedrigdimensionale Darstellung zu überführen, schließlich wissen wir, dass die auftretenden Vektoren aus relativ niedrigdimensionalen Räumen stammen.

Den Wechsel der Darstellung müssen wir so gestalten, dass die im Ausgleichsproblem verwendete Norm leicht berechnet werden kann. Da es sich um die euklidische Norm handelt, bietet es sich an, eine orthogonale Transformation heranzuziehen.

Analog zu (3.25) setzen wir

$$m_0 := \max\{\dim \mathcal{K}(\mathbf{r}^{(0)}, m) \ : \ m \in \mathbb{N}_0\}.$$

Definition 3.34 (Arnoldi-Basis) *Eine Familie* $(\mathbf{p}^{(m)})_{m=0}^{m_0-1}$ *nennen wir eine* Arnoldi-Basis *für die Matrix* $\mathbf{A}$ *und das Anfangsresiduum* $\mathbf{r}^{(0)}$*, falls die Bedingungen*

$$\mathcal{K}(\mathbf{r}^{(0)}, m) = \operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m)}\} \qquad \textit{für alle } m \in [0 : m_0 - 1], \tag{3.35a}$$

$$\langle \mathbf{p}^{(\ell)}, \mathbf{p}^{(m)} \rangle_2 = \begin{cases} 1 & \textit{falls } \ell = m, \\ 0 & \textit{ansonsten} \end{cases} \qquad \textit{für alle } m, \ell \in [0 : m_0 - 1] \tag{3.35b}$$

gelten, falls also für alle $m \in [0 : m_0 - 1]$ *die ersten* $m+1$ *Vektoren jeweils eine Orthonormalbasis des Krylow-Raums* $\mathcal{K}(\mathbf{r}^{(0)}, m)$ *bilden.*

Eine Arnoldi-Basis können wir induktiv konstruieren: Falls $\mathbf{r}^{(0)} = 0$ gilt, haben wir $m_0 = 0$ und brauchen nichts zu tun.

Anderenfalls können wir

$$\mathbf{p}^{(0)} := \frac{\mathbf{r}^{(0)}}{\|\mathbf{r}^{(0)}\|_2}$$

setzen, um die Bedingungen (3.35) zu erfüllen.

Gehen wir nun davon aus, dass wir Vektoren $\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m)}$ mit den gewünschten Eigenschaften konstruiert haben. Wir suchen einen Vektor $\mathbf{p}^{(m+1)} \in \mathbb{K}^{\mathcal{I}}$ mit

$$\mathcal{K}(\mathbf{r}^{(0)}, m+1) = \operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m)}, \mathbf{p}^{(m+1)}\},$$

$$\langle \mathbf{p}^{(\ell)}, \mathbf{p}^{(m+1)} \rangle_2 = \begin{cases} 1 & \text{falls } \ell = m+1, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } \ell \in [0:m+1].$$

Mit der Voraussetzung und dem Basis-Austauschsatz finden wir

$$\begin{aligned} \mathcal{K}(\mathbf{r}^{(0)}, m+1) &= \operatorname{span}\{\mathbf{r}^{(0)}, \mathbf{A}\mathbf{r}^{(0)}, \ldots, \mathbf{A}^m\mathbf{r}^{(0)}, \mathbf{A}^{m+1}\mathbf{r}^{(0)}\} \\ &= \operatorname{span}\{\mathbf{p}^{(0)}, \mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(m)}, \mathbf{A}^{m+1}\mathbf{r}^{(0)}\}. \end{aligned} \tag{3.36}$$

Unter gewissen Annahmen werden die Winkel zwischen den Vektoren $\mathbf{A}^m\mathbf{r}^{(0)}$ sehr klein, so dass Rundungsfehler dazu führen können, dass unser Algorithmus zu einem sehr ungenauen Ergebnis führt. Deshalb ist es unser Ziel, $\mathbf{A}^{m+1}\mathbf{r}^{(0)}$ in (3.36) durch einen günstigeren Vektor zu ersetzen. Nach Voraussetzung haben wir

$$\mathbf{A}^m\mathbf{r}^{(0)} \in \mathcal{K}(\mathbf{r}^{(0)}, m) = \operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m)}\}$$

existieren Koeffizienten $\alpha_0, \ldots, \alpha_m \in \mathbb{K}$ mit

$$\begin{aligned} \mathbf{A}^m\mathbf{r}^{(0)} &= \alpha_0\mathbf{p}^{(0)} + \ldots + \alpha_m\mathbf{p}^{(m)}, \\ \mathbf{A}^{m+1}\mathbf{r}^{(0)} &= \alpha_0\mathbf{A}\mathbf{p}^{(0)} + \ldots + \alpha_m\mathbf{A}\mathbf{p}^{(m)}. \end{aligned}$$

Wieder aufgrund der Voraussetzung gilt

$$\mathbf{A}\mathbf{p}^{(\ell)} \in \mathcal{K}(\mathbf{r}^{(0)}, \ell+1) \subseteq \mathcal{K}(\mathbf{r}^{(0)}, m) \qquad \text{für alle } \ell \in [0:m-1],$$

so dass wir

$$\mathbf{A}^{m+1}\mathbf{r}^{(0)} - \alpha_m\mathbf{A}\mathbf{p}^{(m)} \in \mathcal{K}(\mathbf{r}^{(0)}, m) = \operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m)}\}$$

erhalten. Mit dieser Gleichung folgt aus (3.36) wieder mit dem Basis-Austauschsatz

$$\mathcal{K}(\mathbf{r}^{(0)}, m+1) = \operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m)}, \mathbf{A}\mathbf{p}^{(m)}\},$$

so dass wir lediglich dafür sorgen müssen, dass aus $\mathbf{A}\mathbf{p}^{(m)}$ ein Einheitsvektor wird, der auf $\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m)}$ senkrecht steht. Das lässt sich mit der Gram-Schmidt-Orthogonalisierung erreichen: Wir setzen

$$\widetilde{\mathbf{p}}^{(m+1)} := \mathbf{A}\mathbf{p}^{(m)} - \sum_{k=0}^{m} \langle \mathbf{p}^{(k)}, \mathbf{A}\mathbf{p}^{(m)} \rangle_2 \mathbf{p}^{(k)}$$

und erhalten

$$\begin{aligned} \langle \mathbf{p}^{(\ell)}, \widetilde{\mathbf{p}}^{(m+1)} \rangle_2 &= \langle \mathbf{p}^{(\ell)}, \mathbf{A}\mathbf{p}^{(m)} \rangle_2 - \sum_{k=0}^{m} \langle \mathbf{p}^{(k)}, \mathbf{A}\mathbf{p}^{(m)} \rangle_2 \langle \mathbf{p}^{(\ell)}, \mathbf{p}^{(k)} \rangle_2 \\ &= \langle \mathbf{p}^{(\ell)}, \mathbf{A}\mathbf{p}^{(m)} \rangle_2 - \langle \mathbf{p}^{(\ell)}, \mathbf{A}\mathbf{p}^{(m)} \rangle_2 = 0 \qquad \text{für alle } \ell \in [0:m], \end{aligned}$$

wobei wir im zweiten Schritt die Orthogonalität der bereits bekannten Basisvektoren ausnutzen.

```
procedure Arnoldi(q, var m_0, p^(0), ..., p^(m_0-1));
p̃ ← q;
α ← ‖p̃‖_2;
m ← 0;
while ‖p̃‖_2 > ε_iu α do
   p^(m) ← p̃/‖p̃‖_2;
   p̃ ← Ap^(m);
   α ← ‖p̃‖_2;
   for k ∈ [0 : m] do
      Â_km ← ⟨p^(k), p̃⟩_2;
      p̃ ← p̃ − Â_km p^(k)
   end for;
   Â_{m+1,m} ← ‖p̃‖_2;
   m ← m + 1
end while;
m_0 ← m
```

Abbildung 3.7: Konstruktion einer Arnoldi-Basis

Falls $\widetilde{\mathbf{p}}^{(m+1)} = \mathbf{0}$ gilt, folgt

$$\mathbf{A}\mathbf{p}^{(m)} \in \operatorname{span}\{\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(m)}\} = \mathcal{K}(\mathbf{r}^{(0)}, m),$$

also

$$\mathcal{K}(\mathbf{r}^{(0)}, m+1) = \mathcal{K}(\mathbf{r}^{(0)}, m),$$

und damit $m_0 \leq m$. Wir können also einfach prüfen, ob wir den Krylow-Raum vollständig ausgeschöpft haben.

Falls $\widetilde{\mathbf{p}}^{(m+1)} \neq \mathbf{0}$ gilt, können wir

$$\mathbf{p}^{(m+1)} := \frac{\widetilde{\mathbf{p}}^{(m+1)}}{\|\widetilde{\mathbf{p}}^{(m+1)}\|_2}$$

setzen und sind fertig.

In der Praxis werden Rundungsfehler häufig dazu führen, dass $\widetilde{\mathbf{p}}^{(m_0)}$ nicht exakt gleich Null ist. Deshalb kann es sinnvoll sein, ein Kriterium der Art

$$\|\widetilde{\mathbf{p}}^{(m+1)}\|_2 \leq \epsilon_{\mathrm{iu}} \|\mathbf{A}\mathbf{p}^{(m)}\|_2 \qquad \Longleftrightarrow \qquad m = m_0$$

zu verwenden und $\epsilon_{\mathrm{iu}} \in \mathbb{R}_{>0}$ hinreichend klein zu wählen. Geometrisch bedeutet diese Bedingung, dass der Winkel zwischen der neuen Richtung $\mathbf{A}\mathbf{p}^{(m)}$ und dem Unterraum $\operatorname{span}\{\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(m)}\}$ sehr klein ist, ein neuer Basisvektor die Lösung also nur wenig verbessern würde. Der resultierende Algorithmus ist in Abbildung 3.7 zusammengefasst.

Eine Arnoldi-Basis können wir nun verwenden, um unsere Minimierungsaufgabe (3.33) zu vereinfachen. Dazu führen wir für alle $m \in [0 : m_0 - 1]$ die Matrizen $\mathbf{P}_m \in \mathbb{K}^{\mathcal{I} \times [0:m]}$ mit

$$P_{m,i\ell} := p_i^{(\ell)} \qquad \text{für alle } i \in \mathcal{I},\ \ell \in [0 : m]$$

ein. Die Spalten sind gerade die Vektoren $\mathbf{p}^{(\ell)}$, wir können also auch kurz

$$\mathbf{P}_m = \begin{pmatrix} \mathbf{p}^{(0)} & \dots & \mathbf{p}^{(m)} \end{pmatrix}$$

schreiben. Da die Vektoren eine Orthonormalbasis bilden, gilt

$$\begin{aligned} (\mathbf{P}_m^* \mathbf{P}_m)_{\ell k} &= \sum_{i \in \mathcal{I}} \overline{P_{m,i\ell}} P_{m,ik} = \langle p^{(\ell)}, p^{(k)} \rangle_2 \\ &= \begin{cases} 1 & \text{falls } \ell = k, \\ 0 & \text{sonst} \end{cases} \qquad \text{für alle } \ell, k \in [0:m], \end{aligned}$$

also kurz

$$\mathbf{P}_m^* \mathbf{P}_m = \mathbf{I}. \tag{3.37}$$

Matrizen mit dieser Eigenschaft bezeichnet man als *orthogonal* oder *isometrisch*.

Lemma 3.35 (Projektion und Norminvarianz) *Sei $m \in [0 : m_0 - 1]$. Dann ist $\mathbf{\Pi} := \mathbf{P}_m \mathbf{P}_m^* \in \mathbb{K}^{\mathcal{I} \times \mathcal{I}}$ eine orthogonale Projektion auf* $\operatorname{Bild} \mathbf{P}_m = \mathcal{K}(\mathbf{r}^{(0)}, m)$*, erfüllt also insbesondere*

$$\mathbf{y} = \mathbf{\Pi}\mathbf{y} = \mathbf{P}_m \mathbf{P}_m^* \mathbf{y} \qquad \textit{für alle } \mathbf{y} \in \mathcal{K}(\mathbf{r}^{(0)}, m).$$

Außerdem gilt

$$\|\mathbf{P}_m \widehat{\mathbf{y}}\|_2 = \|\widehat{\mathbf{y}}\|_2 \qquad \textit{für alle } \widehat{\mathbf{y}} \in \mathbb{K}^{[0:m]}.$$

Beweis. Wegen (3.37) gilt

$$\mathbf{\Pi}^2 = \mathbf{P}_m \underbrace{\mathbf{P}_m^* \mathbf{P}_m}_{=\mathbf{I}} \mathbf{P}_m^* = \mathbf{P}_m \mathbf{P}_m^* = \mathbf{\Pi},$$

also ist $\mathbf{\Pi}$ eine Projektion. Es gilt auch

$$\mathbf{\Pi}^* = (\mathbf{P}_m \mathbf{P}_m^*)^* = \mathbf{P}_m^{**} \mathbf{P}_m^* = \mathbf{P}_m \mathbf{P}_m^* = \mathbf{\Pi},$$

also ist $\mathbf{\Pi}$ auch eine orthogonale Projektion.

Offenbar gilt $\operatorname{Bild} \mathbf{\Pi} \subseteq \operatorname{Bild} \mathbf{P}_m$. Sei nun $\mathbf{y} \in \operatorname{Bild} \mathbf{P}_m$. Dann existiert ein $\widehat{\mathbf{y}} \in \mathbb{K}^{[0:m]}$ mit $\mathbf{y} = \mathbf{P}_m \widehat{\mathbf{y}}$, und es folgt aus (3.37) auch

$$\mathbf{\Pi}\mathbf{y} = \mathbf{P}_m \underbrace{\mathbf{P}_m^* \mathbf{P}_m}_{=\mathbf{I}} \widehat{\mathbf{y}} = \mathbf{P}_m \widehat{\mathbf{y}} = \mathbf{y},$$

also erhalten wir $\operatorname{Bild} \mathbf{\Pi} = \operatorname{Bild} \mathbf{P}_m$. Aus der Konstruktion der Matrizen $\mathbf{P}_m$ folgt direkt

$$\mathcal{K}(\mathbf{r}^{(0)}, m) = \operatorname{span}\{\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(m)}\} = \operatorname{Bild} \mathbf{P}_m.$$

Für den Nachweis der Norminvarianz wählen wir ein $\widehat{\mathbf{y}} \in \mathbb{K}^{[0:m]}$ und stellen fest, dass

$$\|\mathbf{P}_m \widehat{\mathbf{y}}\|_2^2 = \langle \mathbf{P}_m \widehat{\mathbf{y}}, \mathbf{P}_m \widehat{\mathbf{y}} \rangle_2 = \langle \underbrace{\mathbf{P}_m^* \mathbf{P}_m}_{=\mathbf{I}} \widehat{\mathbf{y}}, \widehat{\mathbf{y}} \rangle_2 = \langle \widehat{\mathbf{y}}, \widehat{\mathbf{y}} \rangle_2 = \|\widehat{\mathbf{y}}\|_2^2$$

gilt. Damit ist alles bewiesen, was zu beweisen war. ■

Da wir bereits gesehen haben, dass $\mathbf{z}^{(m)}$ und der in (3.33) zu minimierende Term in geeigneten Krylow-Räumen liegen, können wir mit Hilfe der Matrizen $\mathbf{P}_m$ eine kompaktere Darstellung des Ausgleichsproblems finden:

Lemma 3.36 (Vereinfachtes Ausgleichsproblem) *Sei $m \in [1 : m_0 - 1]$. Wir definieren*

$$\widehat{\mathbf{A}}^{(m)} := \mathbf{P}_m^* \mathbf{A} \mathbf{P}_{m-1} \in \mathbb{K}^{[0:m]\times[0:m-1]}, \qquad \widehat{\mathbf{r}}^{(m)} := \mathbf{P}_m^* \mathbf{r}^{(0)} \in \mathbb{K}^{[0:m]}.$$

Ein Vektor $\mathbf{z}^{(m)} \in \mathcal{K}(\mathbf{r}^{(0)}, m-1)$ ist genau dann eine Lösung des Ausgleichsproblems (3.33), falls der Vektor $\widehat{\mathbf{z}}^{(m)} := \mathbf{P}_{m-1}^* \mathbf{z}^{(m)}$ eine Lösung des Ausgleichsproblems

$$\|\widehat{\mathbf{A}}^{(m)} \widehat{\mathbf{z}}^{(m)} - \widehat{\mathbf{r}}^{(m)}\|_2 \le \|\widehat{\mathbf{A}}^{(m)} \widehat{\mathbf{y}} - \widehat{\mathbf{r}}^{(m)}\|_2 \qquad \text{für alle } \widehat{\mathbf{y}} \in \mathbb{K}^{[0:m-1]} \tag{3.38}$$

ist. Dabei gilt $\mathbf{z}^{(m)} = \mathbf{P}_{m-1} \widehat{\mathbf{z}}^{(m)}$.

Beweis. Sei zunächst $\mathbf{z} \in \mathcal{K}(\mathbf{r}^{(0)}, m-1)$ beliebig, und sei $\widehat{\mathbf{z}} := \mathbf{P}_m^* \mathbf{z}$. Aus Lemma 3.35 erhalten wir unmittelbar

$$\mathbf{z} = \mathbf{P}_{m-1} \widehat{\mathbf{z}}. \tag{3.39}$$

Wir haben bereits gesehen, dass sich aus

$$\mathbf{r}^{(0)} \in \mathcal{K}(\mathbf{r}^{(0)}, m-1), \qquad \mathbf{z} \in \mathcal{K}(\mathbf{r}^{(0)}, m-1)$$

unmittelbar

$$\mathbf{A}\mathbf{z} - \mathbf{r}^{(0)} \in \mathcal{K}(\mathbf{r}^{(0)}, m)$$

ergibt. Mit Lemma 3.35 folgen die Gleichungen

$$\begin{aligned} \mathbf{A}\mathbf{z} - \mathbf{r}^{(0)} &= \mathbf{P}_m \mathbf{P}_m^* (\mathbf{A}\mathbf{z} - \mathbf{r}^{(0)}), \\ \|\mathbf{A}\mathbf{z} - \mathbf{r}^{(0)}\|_2 &= \|\mathbf{P}_m \mathbf{P}_m^* (\mathbf{A}\mathbf{z} - \mathbf{r}^{(0)})\|_2 = \|\mathbf{P}_m^* (\mathbf{A}\mathbf{z} - \mathbf{r}^{(0)})\|_2. \end{aligned}$$

Wir setzen (3.39) ein und erhalten

$$\|\mathbf{A}\mathbf{z} - \mathbf{r}^{(0)}\|_2 = \|\mathbf{P}_m^* \mathbf{A} \mathbf{P}_{m-1} \widehat{\mathbf{z}} - \mathbf{P}_m^* \mathbf{r}^{(0)}\|_2 = \|\widehat{\mathbf{A}}^{(m)} \widehat{\mathbf{z}} - \widehat{\mathbf{r}}^{(m)}\|_2. \tag{3.40}$$

Sei nun $\mathbf{z}^{(m)} \in \mathcal{K}(\mathbf{r}^{(0)}, m-1)$ Lösung des Ausgleichsproblems (3.33). Sei $\widehat{\mathbf{z}}^{(m)} := \mathbf{P}_{m-1}^* \mathbf{z}^{(m)}$, und sei $\widehat{\mathbf{y}} \in \mathbb{K}^{[0:m-1]}$ beliebig. Wir setzen $\mathbf{y} := \mathbf{P}_{m-1} \widehat{\mathbf{y}}$. Dann gilt

$$\|\widehat{\mathbf{A}}^{(m)} \widehat{\mathbf{z}}^{(m)} - \widehat{\mathbf{r}}^{(m)}\|_2 = \|\mathbf{A}\mathbf{z}^{(m)} - \mathbf{r}^{(0)}\|_2 \le \|\mathbf{A}\mathbf{y} - \mathbf{r}^{(0)}\|_2 = \|\widehat{\mathbf{A}}^{(m)} \widehat{\mathbf{y}} - \widehat{\mathbf{r}}^{(m)}\|_2,$$

also ist $\widehat{\mathbf{z}}^{(m)}$ Lösung des reduzierten Ausgleichsproblems (3.38).

Sei nun umgekehrt $\widehat{\mathbf{z}}^{(m)} \in \mathbb{K}^{[0:m-1]}$ eine Lösung des reduzierten Ausgleichsproblems (3.38). Wir setzen $\mathbf{z}^{(m)} := \mathbf{P}_{m-1} \widehat{\mathbf{z}}^{(m)}$. Sei $\mathbf{y} \in \mathcal{K}(\mathbf{r}^{(0)}, m-1)$ beliebig, und set $\widehat{\mathbf{y}} := \mathbf{P}_{m-1}^* \mathbf{y}$. Dann gilt

$$\|\mathbf{A}\mathbf{z}^{(m)} - \mathbf{r}^{(0)}\|_2 = \|\widehat{\mathbf{A}}^{(m)} \widehat{\mathbf{z}}^{(m)} - \widehat{\mathbf{r}}^{(m)}\|_2 \le \|\widehat{\mathbf{A}}^{(m)} \widehat{\mathbf{y}} - \widehat{\mathbf{r}}^{(m)}\|_2 = \|\mathbf{A}\mathbf{y} - \mathbf{r}^{(0)}\|_2,$$

also ist $\mathbf{z}^{(m)}$ Lösung des Ausgleichsproblems (3.33). ■

Mit Hilfe dieses Lemmas reduziert sich das lineare Ausgleichsproblem (3.33) auf ein Problem in einem $(m+1)$-dimensionalen Raum: Wir suchen lediglich nach der Lösung $\widehat{\mathbf{z}}^{(m)}$ des Ausgleichsproblems

$$\|\widehat{\mathbf{A}}^{(m)}\widehat{\mathbf{z}}^{(m)} - \widehat{\mathbf{r}}^{(m)}\|_2 \leq \|\widehat{\mathbf{A}}^{(m)}\widehat{\mathbf{y}} - \widehat{\mathbf{r}}^{(m)}\|_2 \qquad \text{für alle } \widehat{\mathbf{y}} \in \mathbb{K}^{[0:m-1]},$$

denn wenn wir sie gefunden haben, ist der Vektor

$$\mathbf{x}^{(m)} = \mathbf{x}^{(0)} + \mathbf{P}^{(m-1)}\widehat{\mathbf{z}}^{(m)}$$

die bezüglich der Defektnorm bestmögliche Approximation der Lösung.

Lineare Ausgleichsprobleme lassen sich besonders leicht lösen, wenn die ihnen zugrundeliegende Matrix injektiv ist. In unserem Fall erbt $\widehat{\mathbf{A}}^{(m)}$ diese Eigenschaft von der Matrix $\mathbf{A}$:

Lemma 3.37 (Injektivität) *Sei $\mathbf{r}^{(0)} \neq \mathbf{0}$. Für alle $m \in [0 : m_0 - 1]$ ist $\widehat{\mathbf{A}}^{(m)}$ injektiv. Für $m = m_0$ ist*

$$\widehat{\mathbf{A}} := \mathbf{P}_{m-1}^* \mathbf{A} \mathbf{P}_{m-1}$$

sogar invertierbar.

Beweis. Sei $m \in [0 : m_0 - 1]$. Zum Nachweis der Injektivität wählen wir einen Vektor $\widehat{\mathbf{y}} \in \mathbb{K}^{[0:m-1]}$ mit $\widehat{\mathbf{A}}^{(m)}\widehat{\mathbf{y}} = \mathbf{0}$. Für alle $\widehat{\mathbf{t}} \in \mathbb{K}^{[0:m]}$ gilt dann

$$0 = \langle \widehat{\mathbf{t}}, \widehat{\mathbf{A}}^{(m)}\widehat{\mathbf{y}} \rangle_2 = \langle \widehat{\mathbf{t}}, \mathbf{P}_m^* \mathbf{A} \mathbf{P}_{m-1}\widehat{\mathbf{y}} \rangle_2 = \langle \mathbf{P}_m \widehat{\mathbf{t}}, \mathbf{A} \mathbf{P}_{m-1}\widehat{\mathbf{y}} \rangle_2. \tag{3.41}$$

Wie wir bereits gesehen haben gilt

$$\mathbf{A}\mathbf{P}_{m-1}\widehat{\mathbf{y}} \in \mathcal{K}(\mathbf{r}^{(0)}, m) = \operatorname{Bild} \mathbf{P}_m,$$

also können wir $\widehat{\mathbf{t}}$ so wählen, dass

$$\mathbf{A}\mathbf{P}_{m-1}\widehat{\mathbf{y}} = \mathbf{P}_m \widehat{\mathbf{t}}$$

gilt. Durch Einsetzen in (3.41) folgt

$$0 = \langle \mathbf{A}\mathbf{P}_{m-1}\widehat{\mathbf{y}}, \mathbf{A}\mathbf{P}_{m-1}\widehat{\mathbf{y}} \rangle_2 = \|\mathbf{A}\mathbf{P}_{m-1}\widehat{\mathbf{y}}\|_2^2,$$

also $\mathbf{A}\mathbf{P}_{m-1}\widehat{\mathbf{y}} = \mathbf{0}$. Da $\mathbf{A}$ und $\mathbf{P}_{m-1}$ injektiv sind, folgt daraus bereits $\widehat{\mathbf{y}} = \mathbf{0}$, also ist $\widehat{\mathbf{A}}^{(m)}$ injektiv.

Für $m = m_0$ wählen wir $\widehat{\mathbf{y}} \in \mathbb{K}^m$ mit $\widehat{\mathbf{A}}\widehat{\mathbf{y}} = \mathbf{0}$ und nutzen aus, dass

$$\mathbf{A}\mathbf{P}_{m-1}\widehat{\mathbf{y}} \in \mathcal{K}(\mathbf{r}^{(0)}, m) = \mathcal{K}(\mathbf{r}^{(0)}, m-1) = \operatorname{Bild} \mathbf{P}_{m-1}$$

gilt, um entsprechend die Injektivität nachzuweisen. Da $\widehat{\mathbf{A}}$ quadratisch ist, impliziert Injektivität auch Regularität. ■

Bemerkung 3.38 (Lucky Breakdown) *Für $m = m_0$ gilt Lemma 3.36 sogar in verschärfter Form: Nach Definition haben wir*

$$\mathbf{A}\mathbf{z}^{(m)} - \mathbf{r}^{(0)} \in \mathcal{K}(\mathbf{r}^{(0)}, m) = \mathcal{K}(\mathbf{r}^{(0)}, m-1),$$

also können wir uns auf das Ausgleichsproblem

$$\|\widehat{\mathbf{A}}\widehat{\mathbf{z}}^{(m)} - \widehat{\mathbf{r}}^{(m-1)}\|_2 \leq \|\widehat{\mathbf{A}}\widehat{\mathbf{y}} - \widehat{\mathbf{r}}^{(m-1)}\|_2, \qquad \text{für alle } \widehat{\mathbf{y}} \in \mathbb{K}^m \tag{3.42}$$

beschränken. Da $\widehat{\mathbf{A}}$ in diesem Fall quadratisch und nach Lemma 3.37 regulär ist, lässt sich $\widehat{\mathbf{z}}^{(m)}$ als Lösung des Gleichungssystems

$$\widehat{\mathbf{A}}\widehat{\mathbf{z}}^{(m)} = \widehat{\mathbf{r}}^{(m-1)} \tag{3.43}$$

berechnen. Es folgen

$$\mathbf{A}\mathbf{P}^{(m-1)}\widehat{\mathbf{z}}^{(m)} = \mathbf{r}^{(0)}, \qquad \mathbf{A}\mathbf{x}^{(m)} = \mathbf{b}.$$

Falls wir also die Arnoldi-Basis nicht mehr erweitern können, ist die Iterierte $\mathbf{x}^{(m)} = \mathbf{x}^*$ bereits die exakte Lösung.

Um das Ausgleichsproblem im allgemeinen Fall $m < m_0$ lösen zu können, bietet es sich an, nach für diesen Zweck nützlichen Eigenschaften der Matrix $\widehat{\mathbf{A}}^{(m)}$ zu suchen.

Lemma 3.39 *Sei $m \in [1 : m_0 - 1]$. Dann gilt*

$$\widehat{A}_{\ell k}^{(m)} = \begin{cases} \langle \mathbf{p}^{(\ell)}, \mathbf{A}\mathbf{p}^{(k)} \rangle_2 & \text{falls } \ell \leq k+1, \\ 0 & \text{sonst} \end{cases} \qquad \text{für alle } \ell \in [0:m],\ k \in [0:m-1],$$

die Matrix $\widehat{\mathbf{A}}^{(m)}$ besitzt also Hessenberg-Gestalt.

Beweis. Nach Definition gilt

$$\begin{aligned} \widehat{A}_{\ell k}^{(m)} &= (\mathbf{P}_m^* \mathbf{A}\mathbf{P}_{m-1})_{\ell k} = \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{I}} \overline{P_{m,i\ell}} A_{ij} P_{m-1,jk} = \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{I}} \overline{p_i^{(\ell)}} A_{ij} p_j^{(k)} \\ &= \langle \mathbf{p}^{(\ell)}, \mathbf{A}\mathbf{p}^{(k)} \rangle_2 \qquad \text{für alle } \ell \in [0:m],\ k \in [0:m-1]. \end{aligned}$$

Für alle $k \in [0 : m-1]$ gilt nach Konstruktion $\mathbf{p}^{(k)} \in \mathcal{K}(\mathbf{r}^{(0)}, k)$, also auch $\mathbf{A}\mathbf{p}^{(k)} \in \mathcal{K}(\mathbf{r}^{(0)}, k+1) \subseteq \operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(k+1)}\}$. Falls nun $\ell > k+1$ gilt, muss $\mathbf{p}^{(\ell)}$ senkrecht zu allen Suchrichtungen $\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(k+1)}$ sein, und es folgt $\langle \mathbf{p}^{(\ell)}, \mathbf{A}\mathbf{p}^{(k)} \rangle_2 = 0$. ■

Auch der Vektor $\widehat{\mathbf{r}}^{(m)}$ ist von besonderer Form: Da $\mathbf{r}^{(0)} \in \mathcal{K}(\mathbf{r}^{(0)}, 0) = \operatorname{span}\{\mathbf{p}^{(0)}\}$ gelten muss, folgt

$$\mathbf{r}^{(0)} = \langle \mathbf{p}^{(0)}, \mathbf{r}^{(0)} \rangle_2 \mathbf{p}^{(0)},$$

und damit dank der Orthogonalität der Basisvektoren auch

$$\langle \mathbf{p}^{(\ell)}, \mathbf{r}^{(0)} \rangle_2 = \langle \mathbf{p}^{(\ell)}, \mathbf{p}^{(0)} \rangle_2 \langle \mathbf{p}^{(0)}, \mathbf{r}^{(0)} \rangle_2 = 0 \qquad \text{für alle } \ell \in [1:m],$$

so dass wir

$$\widehat{\mathbf{r}}^{(m)} = \mathbf{P}_m^* \mathbf{r}^{(0)} = \begin{pmatrix} \langle \mathbf{p}^{(0)}, \mathbf{r}^{(0)} \rangle_2 \\ \langle \mathbf{p}^{(1)}, \mathbf{r}^{(0)} \rangle_2 \\ \vdots \\ \langle \mathbf{p}^{(m)}, \mathbf{r}^{(0)} \rangle_2 \end{pmatrix} = \begin{pmatrix} \langle \mathbf{p}^{(0)}, \mathbf{r}^{(0)} \rangle_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

erhalten, der Vektor $\widehat{\mathbf{r}}^{(m)}$ ist also ein Vielfaches des ersten kanonischen Einheitsvektors.

Gemäß Lemma 3.39 haben die Einträge der Matrix $\widehat{\mathbf{A}}^{(m)}$ die Form

$$\widehat{A}_{\ell k}^{(m)} = \langle \mathbf{p}^{(\ell)}, \mathbf{A}\mathbf{p}^{(k)} \rangle_2.$$

Für $\ell \leq k$ treten genau diese Skalarprodukte auch bei der Berechnung der Arnoldi-Basis auf, für $\ell = k+1$ haben wir

$$\begin{aligned} \widehat{A}_{\ell k}^{(m)} &= \langle \mathbf{p}^{(\ell)}, \mathbf{A}\mathbf{p}^{(k)} \rangle_2 = \langle \mathbf{p}^{(\ell)}, \widetilde{\mathbf{p}}^{(k+1)} \rangle_2 + \sum_{j=0}^{k} \langle \mathbf{p}^{(j)}, \mathbf{A}\mathbf{p}^{(k)} \rangle_2 \langle \mathbf{p}^{(\ell)}, \mathbf{p}^{(j)} \rangle_2 \\ &= \langle \mathbf{p}^{(k+1)}, \widetilde{\mathbf{p}}^{(k+1)} \rangle_2 + \sum_{j=0}^{k} \langle \mathbf{p}^{(j)}, \mathbf{A}\mathbf{p}^{(k)} \rangle_2 \langle \mathbf{p}^{(k+1)}, \mathbf{p}^{(j)} \rangle_2 = \|\widetilde{\mathbf{p}}^{(k+1)}\|_2, \end{aligned}$$

also treten alle zur Konstruktion der Matrix $\widehat{\mathbf{A}}^{(m)}$ benötigten Größen bereits bei der Konstruktion der Arnoldi-Basis auf, so dass sich die Matrix mit minimalem Mehraufwand aufstellen lässt. Der in Abbildung 3.7 dargestellte Algorithmus ist bereits so formuliert, dass die Matrizen $\widehat{\mathbf{A}}^{(m)}$ konstruiert werden.

Aus Stabilitätsgründen empfiehlt es sich, die Orthogonalisierung von $\widetilde{\mathbf{p}}^{(m)}$ wie in Abbildung 3.7 durchzuführen: Die einzelnen Suchrichtungen werden der Reihe nach subtrahiert, und die Skalarprodukte werden nicht für das ursprüngliche $\widetilde{\mathbf{p}}^{(m)}$ berechnet, sondern bereits für das partiell orthogonalisierte. Theoretisch sind beide Ansätze identisch, weil die Orthogonalität der Suchrichtungen impliziert, dass in beiden Fällen dieselben Skalarprodukte berechnet werden. Praktisch wirken sich Rundungsfehler bei der modifizierten Konstruktion weniger stark als bei der ursprünglichen aus.

Übungsaufgabe 3.40 (Biorthogonale Basis) *Seien* $\mathbf{y}, \mathbf{z} \in \mathbb{K}^{\mathcal{I}}$ *mit* $\langle \mathbf{y}, \mathbf{z} \rangle_2 \neq 0$ *gegeben. Neben den Krylow-Räumen*

$$\mathcal{K}(\mathbf{y}, m) := \operatorname{span}\{\mathbf{y}, \mathbf{A}\mathbf{y}, \ldots, \mathbf{A}^m \mathbf{y}\} \qquad \textit{für alle } m \in \mathbb{N}_0$$

können wir auch die zu der adjungierten Matrix $\mathbf{A}^*$ *gehörenden Räume*

$$\mathcal{K}^*(\mathbf{z}, m) := \operatorname{span}\{\mathbf{z}, \mathbf{A}^*\mathbf{z}, \ldots, (\mathbf{A}^*)^m \mathbf{z}\} \qquad \textit{für alle } m \in \mathbb{N}_0$$

betrachten. Wir sind daran interessiert, biorthogonale Basen $(\mathbf{p}^{(m)})_{m=0}^{m_0-1}$ *und* $(\mathbf{q}^{(m)})_{m=0}^{m_0-1}$ *dieser Räume zu konstruieren, es sollen also*

$$\operatorname{span}\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(m)}\} = \mathcal{K}(\mathbf{y}, m),$$

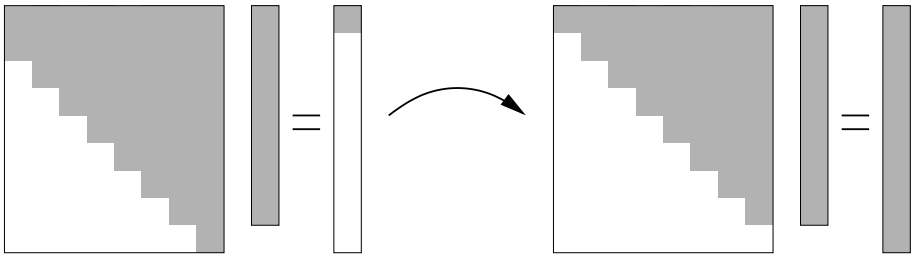

Abbildung 3.8: Hessenberg-Ausgleichsproblem mit Householder-Transformation

$$\begin{aligned} \operatorname{span}\{\mathbf{q}^{(0)}, \dots, \mathbf{q}^{(m)}\} &= \mathcal{K}^*(\mathbf{z}, m) && \textit{für alle } m \in [0 : m_0 - 1], \\ \langle \mathbf{q}^{(\ell)}, \mathbf{p}^{(k)} \rangle_2 &= \begin{cases} 1 & \textit{falls } \ell = k, \\ 0 & \textit{ansonsten} \end{cases} && \textit{für alle } \ell, k \in [0 : m_0 - 1] \end{aligned}$$

gelten, und $m_0 \in \mathbb{N}$ sollte natürlich möglichst groß sein.

(a) Geben Sie einen Algorithmus an, mit dem sich, analog zu der Konstruktion der Arnoldi-Basis, die gesuchten biorthogonalen Basen konstruieren lassen.

(b) Beweisen Sie, dass $\langle \mathbf{q}^{(\ell)}, \mathbf{A}\mathbf{p}^{(k)} \rangle_2 = 0$ für alle $\ell, k \in [0 : m_0 - 1]$ mit $|\ell - k| > 1$ gilt.

(c) Analog zu dem cg-Verfahren suchen wir nach einer Näherung $\mathbf{x}^{(m)} = \mathbf{x}^{(0)} + \alpha_0 \mathbf{p}^{(0)} + \dots + \alpha_{m-1}\mathbf{p}^{(m-1)}$ derart, dass das zugehörige Residuum $\mathbf{r}^{(m)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(m)}$ die Orthogonalitätseigenschaft $\langle \mathbf{q}^{(\ell)}, \mathbf{r}^{(m)} \rangle_2 = 0$ für alle $\ell \in [0 : m - 1]$ besitzt. Geben Sie einen Algorithmus an, mit dem ein solches $\mathbf{x}^{(m)}$ berechnet werden kann.

Die im Zuge der Konstruktion der Arnoldi-Basis berechneten Matrizen $\widehat{\mathbf{A}}^{(m)}$ können wir nun verwenden, um Lösungen $\widehat{\mathbf{z}}^{(m)} \in \mathbb{K}^m$ des linearen Ausgleichsproblems (3.38) und damit auch die Iterierten $\mathbf{x}^{(m)} = \mathbf{x}^{(0)} + \mathbf{P}_{m-1}\mathbf{z}^{(m)}$ zu bestimmen. Nach Lemma 3.37 wissen wir, dass das Ausgleichsproblem immer eindeutig lösbar ist.

Praktisch berechnen können wir die Lösung mit Hilfe einer Householder-Zerlegung: Falls wir eine orthogonale Matrix $\widehat{\mathbf{Q}}^{(m)} \in \mathbb{K}^{(m+1)\times(m+1)}$ und eine obere Dreiecksmatrix $\widehat{\mathbf{R}}^{(m)} \in \mathbb{K}^{(m+1)\times m}$ mit $\widehat{\mathbf{Q}}^{(m)}\widehat{\mathbf{R}}^{(m)} = \widehat{\mathbf{A}}^{(m)}$ konstruieren können, erhalten wir dank der Invarianz der Norm unter der orthogonalen Transformation $\widehat{\mathbf{Q}}^{(m)}$ das transformierte Ausgleichsproblem

$$\|\widehat{\mathbf{R}}^{(m)}\widehat{\mathbf{z}}^{(m)} - (\widehat{\mathbf{Q}}^{(m)})^*\widehat{\mathbf{r}}^{(m)}\|_2 \le \|\widehat{\mathbf{R}}^{(m)}\widehat{\mathbf{y}} - (\widehat{\mathbf{Q}}^{(m)})^*\widehat{\mathbf{r}}^{(m)}\|_2 \qquad \text{für alle } \widehat{\mathbf{y}} \in \mathbb{K}^m.$$

Nach Lemma 3.37 hat die Matrix $\widehat{\mathbf{R}}^{(m)}$ vollen Rang, und ihre letzte Zeile ist Null (vgl. Abbildung 3.8).

Demzufolge können wir das transformierte Ausgleichsproblem lösen, indem wir die Komponenten von $\widehat{\mathbf{z}}^{(m)}$ durch Rückwärtseinsetzen in die oberen m Zeilen der Matrix $\widehat{\mathbf{R}}^{(m)}$ berechnen. Der Betrag der letzten Komponente von $(\widehat{\mathbf{Q}}^{(m)})^*\widehat{\mathbf{r}}^{(m)}$ entspricht gerade dem verbliebenen Approximationsfehler.

procedure FindeGivens(x, y, **var** c, s);
if $|x| \geq |y|$ **then**
  $\tau \leftarrow -y/\bar{x}$;
  $c \leftarrow \frac{\bar{x}}{|x|\sqrt{1+|\tau|^2}}$;
  $s \leftarrow c\tau$
else
  $\tau \leftarrow -\bar{x}/y$;
  $s \leftarrow \frac{-y}{|y|\sqrt{1+|\tau|^2}}$;
  $c \leftarrow s\tau$
end if

procedure Givens(c, s, **var** x, y);
$h \leftarrow x$;
$x \leftarrow ch - \bar{s}y$;
$y \leftarrow sh + \bar{c}y$

Abbildung 3.9: Bestimmung der Koeffizienten einer Givens-Rotation und Anwendung der Rotation auf einen Vektor (x, y)

Unsere Aufgabe besteht also darin, eine geeignete orthogonale Transformation $\widehat{\mathbf{Q}}^{(m)} \in \mathbb{K}^{(m+1)\times(m+1)}$ zu finden. Laut Lemma 3.39 ist $\widehat{\mathbf{A}}^{(m)}$ bereits eine Hessenberg-Matrix, wir müssen also lediglich die untere Nebendiagonale eliminieren. Diese Aufgabe kann mit Hilfe von *Givens-Rotationen* gelöst werden:

Seien $x, y \in \mathbb{K}$. Wir suchen eine orthogonale Matrix

$$\mathbf{Q} = \begin{pmatrix} c & -\bar{s} \\ s & \bar{c} \end{pmatrix},$$

die die untere Komponente des Vektors (x, y) eliminiert. Wenn wir das Matrix-Vektor-Produkt ausschreiben, erhalten wir

$$\mathbf{Q}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} c & -\bar{s} \\ s & \bar{c} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} cx - \bar{s}y \\ sx + \bar{c}y \end{pmatrix},$$

wir müssen also $c, s \in \mathbb{K}$ so bestimmen, dass $sx + \bar{c}y = 0$ gilt und $\mathbf{Q}$ orthogonal ist.

Wir erreichen dieses Ziel, indem wir

$$c := \frac{\bar{x}}{\sqrt{|x|^2 + |y|^2}}, \qquad s := \frac{-y}{\sqrt{|x|^2 + |y|^2}}$$

setzen, denn dann gelten

$$\begin{aligned} cx - \bar{s}y &= \frac{\bar{x}x + \bar{y}y}{\sqrt{|x|^2 + |y|^2}} = \sqrt{|x|^2 + |y|^2}, \\ sx + \bar{c}y &= \frac{-yx + xy}{\sqrt{|x|^2 + |y|^2}} = 0, \end{aligned}$$

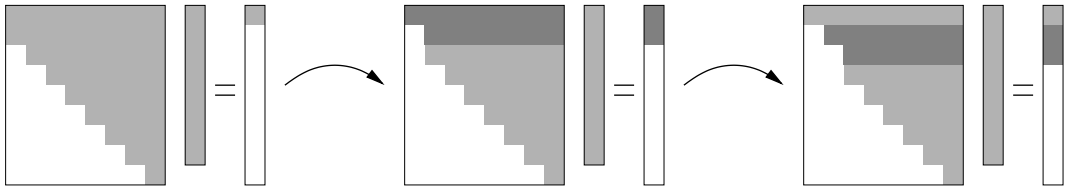

Abbildung 3.10: Transformation einer Hessenberg-Matrix auf obere Dreiecksgestalt mit Hilfe von zeilenweisen Givens-Rotationen

$$
\begin{aligned}
|c|^2 + |s|^2 &= \frac{|x|^2 + |y|^2}{|x|^2 + |y|^2} = 1, \\
-cs + sc &= \frac{\bar{x}y - y\bar{x}}{|x|^2 + |y|^2} = 0.
\end{aligned}
$$

In der Praxis ist es sinnvoll, c und s so zu berechnen, dass Rundungsfehler reduziert werden, etwa indem wir die Fälle $|x| \geq |y|$ und $|x| < |y|$ unterscheiden. Im ersten Fall setzen wir $\tau := -y/\bar{x}$, stellen

$$
\sqrt{|x|^2 + |y|^2} = |x|\sqrt{1 + |y/x|^2} = |x|\sqrt{1 + |\tau|^2}
$$

fest und erhalten

$$
c = \frac{\bar{x}}{\sqrt{|x|^2 + |y|^2}} = \frac{\bar{x}}{|x|\sqrt{1 + |\tau|^2}}, \qquad s = \frac{-y}{\sqrt{|x|^2 + |y|^2}} = \frac{\bar{x}}{|x|\sqrt{1 + |\tau|^2}} \frac{-y}{\bar{x}} = c\tau.
$$

Im zweiten Fall setzen wir $\tau := -\bar{x}/y$, erhalten

$$
\sqrt{|x|^2 + |y|^2} = |y|\sqrt{|x/y|^2 + 1} = |y|\sqrt{1 + |\tau|^2}
$$

und können damit c und s gemäß

$$
s = \frac{-y}{\sqrt{|x|^2 + |y|^2}} = \frac{-y}{|y|\sqrt{1 + |\tau|^2}}, \qquad c = \frac{\bar{x}}{\sqrt{|x|^2 + |y|^2}} = \frac{-y}{|y|\sqrt{1 + |\tau|^2}} \frac{\bar{x}}{-y} = s\tau
$$

berechnen. Die resultierenden Prozeduren zur Bestimmung von c und s und zur Anwendung auf einen Vektor sind in Abbildung 3.9 zusammengefasst.

Bemerkung 3.41 *In der Regel sind wir nur daran interessiert, dass die Matrix* $\mathbf{Q}$ *orthogonal ist und die zweite Zeile des Ausgangsvektors eliminiert. Dabei spielt das Vorzeichen von* $\mathbf{Q}$ *keine Rolle, wir können also in Abbildung 3.9 auch* x *statt* $\bar{x}$ *im ersten Fall und* y *statt* $-y$ *im zweiten Fall verwenden.*

Falls $\mathbb{K} = \mathbb{R}$ *gilt, können wir die Vorzeichen gerade so wählen, dass jeweils nur* $1/\sqrt{1 + |\tau|^2}$ *übrig bleibt, so dass die Berechnung besonders einfach ausfällt.*

```
procedure Ausgleich(m, Â, r̂, var ẑ);
for ℓ = 0 to m − 1 do
   FindeGivens(Â_{ℓ,ℓ}, Â_{ℓ+1,ℓ}, c_ℓ, s_ℓ);
   for k ∈ {ℓ, ..., m − 1} do
      Givens(c_ℓ, s_ℓ, Â_{ℓ,k}, Â_{ℓ+1,k})
   end for;
   Givens(c_ℓ, s_ℓ, r̂_ℓ, r̂_{ℓ+1})
end for;
for ℓ = m − 1 downto 0 do
   q ← r̂_ℓ;
   for k ∈ {ℓ + 1, ..., m − 1} do
      q ← q − Â_{ℓ,k} ẑ_k
   end for;
   ẑ_ℓ ← q/Â_{ℓ,ℓ}
end for
```

Abbildung 3.11: Effizientes Lösen des Ausgleichsproblems (3.38)

Wir können die Givens-Rotation verwenden, um zunächst den ersten Nebendiagonaleintrag $\widehat{A}_{10}^{(m)}$ zu eliminieren: Für $x = \widehat{A}_{00}^{(m)}$ und $y = \widehat{A}_{10}^{(m)}$ erhalten wir

$$\mathbf{Q}\begin{pmatrix}\widehat{A}_{00}^{(m)}\\ \widehat{A}_{10}^{(m)}\end{pmatrix} = \begin{pmatrix}\sqrt{|\widehat{A}_{00}^{(m)}|^2 + |\widehat{A}_{10}^{(m)}|^2}\\ 0\end{pmatrix}.$$

Wenn wir also die orthogonale Matrix

$$\begin{pmatrix}\mathbf{Q} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{pmatrix} \in \mathbb{K}^{(m+1)\times(m+1)}$$

von links mit $\widehat{\mathbf{A}}^{(m)}$ multiplizieren, werden die ersten beiden Zeilen der Matrix gerade so linear kombiniert, dass $\widehat{A}_{10}^{(m)}$ anschließend Null ist. Wir können entsprechend fortfahren, um auch die restlichen Nebendiagonaleinträge zu eliminieren. Am besten ist es, bei dieser Gelegenheit auch gleich die rechte Seite $\widehat{\mathbf{r}}^{(m)}$ zu behandeln (vgl. Abbildung 3.10). Anschließend kann das resultierende Ausgleichsproblem wie beschrieben durch Rückwärtseinsetzen gelöst werden.

Der resultierende Algorithms ist in Abbildung 3.11 zusammengefasst: Er geht von einer Matrix $\widehat{\mathbf{A}}^{(m)}$ und einer rechten Seite $\widehat{\mathbf{r}}^{(m)}$ aus und löst das Ausgleichsproblem (3.38), indem zuerst die Matrix mit Hilfe von Givens-Rotationen auf obere Dreiecksgestalt gebracht und dann $\widehat{\mathbf{z}}^{(m)}$ durch einfaches Rückwärtseinsetzen in die obere Dreiecksmatrix bestimmt wird.

In der Regel wollen wir nicht die vollständige Arnoldi-Basis aufstellen, sondern wir sind lediglich daran interessiert, eine approximative Lösung einer bestimmten Genauigkeit zu konstruieren. Diese Genauigkeit lässt sich mit Hilfe des Ausgleichsproblems steuern,

unsere Aufgabe besteht also darin, die in den Abbildungen 3.7 und 3.11 gegebenen Algorithmen zu kombinieren.

Laut Lemma 3.39 können wir die Matrix $\widehat{\mathbf{A}}^{(m+1)}$ aus der Matrix $\widehat{\mathbf{A}}^{(m)}$ gewinnen, indem wir unten eine Nullzeile und rechts eine weitere Spalte hinzufügen. Wenn wir bereits Givens-Rotationen für die ersten m Spalten konstruiert haben, können wir sie nun auf die letzte Spalte anwenden und dann eine neue Rotation bestimmen, die den letzten Eintrag dieser Spalte eliminiert und die rechte Seite aktualisiert.

Indem wir in dieser Weise die Konstruktion der Arnoldi-Basis mit der Householder-Faktorisierung kombinieren, erhalten wir den in Abbildung 3.12 dargestellten *GMRES-Algorithmus*. Nach der Berechnung jedes neuen Arnoldi-Vektors bestimmt er die neue Defektnorm und bricht ab, sobald sie klein genug oder ein invarianter Unterraum erreicht ist. Anschließend wird durch Rückwärtseinsetzen die neue Iterierte berechnet.

Lemma 3.42 *Sei* $\mathbf{A}$ *regulär, sei* $\mathbf{b} \in \mathbb{K}^{\mathcal{I}}$ *und* $\mathbf{x}^* := \mathbf{A}^{-1}\mathbf{b}$. *Sei* $(\mathbf{x}^{(m)})_{m\in\mathbb{N}_0}$ *die Folge der Semiiterierten des GMRES-Verfahrens zu einem Startvektor* $\mathbf{x}^{(0)} \in \mathbb{K}^{\mathcal{I}}$. *Dann gilt*

$$\|\mathbf{b} - \mathbf{A}\mathbf{x}^{(m)}\|_2 = \min\{\|q_m(\mathbf{A})(\mathbf{b} - \mathbf{A}\mathbf{x}^{(0)})\|_2 \ : \ q_m \text{ ist ein Polynom vom Grad } \le m \text{ mit } q_m(0) = 1\} \qquad \text{für alle } m \in \mathbb{N}_0.$$

Beweis. Nach Konstruktion ist

$$\mathbf{x}^{(m)} = \mathbf{x}^{(0)} + \mathbf{P}_{m-1}\widehat{\mathbf{z}}^{(m)},$$

und $\widehat{\mathbf{z}}^{(m)} \in \mathbb{K}^{[0:m-1]}$ ist so gewählt, dass

$$\|\mathbf{b} - \mathbf{A}\mathbf{x}^{(m)}\|_2 = \|\mathbf{r}^{(0)} - \mathbf{A}\mathbf{P}_{m-1}\widehat{\mathbf{z}}^{(m)}\|_2 \le \|\mathbf{r}^{(0)} - \mathbf{A}\mathbf{P}_{m-1}\widehat{\mathbf{y}}\|_2 \quad \text{für alle } \widehat{\mathbf{y}} \in \mathbb{K}^{[0:m-1]}$$

gilt. Das Bild von $\mathbf{P}_{m-1}$ ist der Aufspann der Vektoren $\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(m-1)}$, und dieser Aufspann ist gerade der Krylow-Raum $\mathcal{K}(\mathbf{r}^{(0)}, m-1)$.

Demzufolge gibt es Koeffizienten $\alpha_0, \dots, \alpha_{m-1} \in \mathbb{K}$ mit

$$\mathbf{P}_{m-1}\widehat{\mathbf{z}}^{(m)} = \sum_{k=0}^{m-1} \alpha_k \mathbf{A}^k \mathbf{r}^{(0)}.$$

Wir definieren die Koeffizienten

$$\beta_\ell := \begin{cases} 1 & \text{falls } \ell = 0, \\ -\alpha_{\ell-1} & \text{sonst} \end{cases} \qquad \text{für alle } \ell \in \{0, \dots, m\}$$

und führen das Polynom q_m durch

$$q_m(\xi) := \sum_{\ell=0}^{m} \beta_\ell \xi^\ell \qquad \text{für alle } \xi \in \mathbb{K},$$

ein, denn dann gelten $q_m(0) = 1$ und

$$\mathbf{b} - \mathbf{A}\mathbf{x}^{(m)} = \mathbf{r}^{(0)} - \mathbf{A}\mathbf{P}_{m-1}\widehat{\mathbf{z}}^{(m)} = \mathbf{r}^{(0)} - \mathbf{A}\sum_{k=0}^{m-1} \alpha_k \mathbf{A}^k \mathbf{r}^{(0)}$$

procedure GMRES(**b**, **var x**);
$\mathbf{r}^{(0)} \leftarrow \mathbf{b} - \mathbf{A}\mathbf{x}$; $\quad \widehat{r}_0 \leftarrow \|\mathbf{r}^{(0)}\|_2$
$\mathbf{p}^{(0)} \leftarrow \mathbf{r}^{(0)}/\widehat{r}_0$; $\quad m \leftarrow 0$;
$\widetilde{\mathbf{p}} \leftarrow \mathbf{A}\mathbf{p}^{(0)}$; $\quad \alpha \leftarrow \|\widetilde{\mathbf{p}}\|_2$;
$\widehat{A}_{0,0} \leftarrow \langle \mathbf{p}^{(0)}, \widetilde{\mathbf{p}} \rangle_2$; $\quad \widetilde{\mathbf{p}} \leftarrow \widetilde{\mathbf{p}} - \widehat{A}_{00}\mathbf{p}^{(0)}$;
$\widehat{A}_{1,0} \leftarrow \|\widetilde{\mathbf{p}}\|_2$; $\quad \beta \leftarrow \widehat{A}_{1,0}$;
FindeGivens($\widehat{A}_{0,0}$, $\widehat{A}_{1,0}$, c_0, s_0);
Givens(c_0, s_0, $\widehat{A}_{0,0}$, $\widehat{A}_{1,0}$);
$\widehat{r}_1 \leftarrow 0$;
Givens(c_0, s_0, $\widehat{r}_0$, $\widehat{r}_1$);
while $\beta > \epsilon_{\mathrm{iu}}\alpha$ **and** $|\widehat{r}_{m+1}| > \epsilon$ **do**
  $\mathbf{p}^{(m+1)} \leftarrow \widetilde{\mathbf{p}}/\beta$; $\quad m \leftarrow m+1$;
  $\widetilde{\mathbf{p}} \leftarrow \mathbf{A}\mathbf{p}^{(m)}$; $\quad \alpha \leftarrow \|\widetilde{\mathbf{p}}\|_2$;
  for $\ell \in \{0, \ldots, m\}$ **do**
    $\widehat{A}_{\ell m} \leftarrow \langle \mathbf{p}^{(\ell)}, \widetilde{\mathbf{p}} \rangle_2$; $\quad \widetilde{\mathbf{p}} \leftarrow \widetilde{\mathbf{p}} - \widehat{A}_{\ell m}\mathbf{p}^{(\ell)}$
  end for;
  $\widehat{A}_{m+1,m} \leftarrow \|\widetilde{\mathbf{p}}\|_2$; $\quad \beta \leftarrow \widehat{A}_{m+1,m}$;
  for $\ell = 0$ **to** $m-1$ **do**
    Givens(c_ℓ, s_ℓ, $\widehat{A}_{\ell,m}$, $\widehat{A}_{\ell+1,m}$)
  end for;
  FindeGivens($\widehat{A}_{m,m}$, $\widehat{A}_{m+1,m}$, c_m, s_m);
  Givens(c_m, s_m, $\widehat{A}_{m,m}$, $\widehat{A}_{m+1,m}$);
  $\widehat{r}_{m+1} \leftarrow 0$;
  Givens(c_m, s_m, $\widehat{r}_m$, $\widehat{r}_{m+1}$)
end while;
for $\ell = m$ **downto** 0 **do**
  $q \leftarrow \widehat{r}_\ell$;
  for $k \in \{\ell+1, \ldots, m\}$ **do**
    $q \leftarrow q - \widehat{A}_{\ell,k}\widehat{z}_k$
  end for;
  $\widehat{z}_\ell \leftarrow q/\widehat{A}_{\ell,\ell}$; $\quad \mathbf{x} \leftarrow \mathbf{x} + \widehat{z}_\ell \mathbf{p}^{(\ell)}$
end for

Abbildung 3.12: GMRES-Algorithmus zum approximativen Lösen eines regulären linearen Gleichungssystems

$$= \mathbf{r}^{(0)} - \sum_{k=0}^{m-1} \alpha_k \mathbf{A}^{k+1}\mathbf{r}^{(0)} = \mathbf{r}^{(0)} + \sum_{\ell=1}^{m} \beta_\ell \mathbf{A}^\ell \mathbf{r}^{(0)} = q_m(\mathbf{A})\mathbf{r}^{(0)}.$$

Damit haben wir

$$\|\mathbf{b} - \mathbf{A}\mathbf{x}^{(m)}\|_2 = \|q_m(\mathbf{A})(\mathbf{b} - \mathbf{A}\mathbf{x}^{(0)})\|_2$$

für unsere Wahl von q_m gezeigt. Jetzt müssen wir nachweisen, dass diese Norm auch minimal ist.

Sei dazu q'_m ein weiteres Polynom mit Grad $\leq m$ und $q'_m(0) = 1$. Wir wählen Koeffizienten $\beta'_0, \ldots, \beta'_m \in \mathbb{K}$ mit

$$q'_m(\xi) = \sum_{\ell=0}^{m} \beta'_\ell \xi^\ell \qquad \text{für alle } \xi \in \mathbb{K}.$$

Wegen $q'_m(0) = 1$ gilt $\beta'_0 = 1$, und wir erhalten

$$\begin{aligned} q'_m(\mathbf{A})\mathbf{r}^{(0)} &= \sum_{\ell=0}^{m} \beta'_\ell \mathbf{A}^\ell \mathbf{r}^{(0)} = \mathbf{r}^{(0)} + \sum_{\ell=1}^{m} \beta'_\ell \mathbf{A}^\ell \mathbf{r}^{(0)} \\ &= \mathbf{r}^{(0)} - \mathbf{A} \sum_{k=0}^{m-1} (-\beta'_{k+1}) \mathbf{A}^k \mathbf{r}^{(0)} = \mathbf{r}^{(0)} - \mathbf{A}\mathbf{y} \end{aligned}$$

für den Vektor

$$\mathbf{y} := -\sum_{k=0}^{m-1} \beta'_{k+1} \mathbf{A}^k \mathbf{r}^{(0)} \in \mathcal{K}(\mathbf{r}^{(0)}, m-1).$$

Da das Bild von $\mathbf{P}_{m-1}$ gerade $\mathcal{K}(\mathbf{r}^{(0)}, m-1)$ ist, muss es einen Vektor $\widehat{\mathbf{y}} \in \mathbb{K}^m$ geben, der $\mathbf{y} = \mathbf{P}_{m-1}\widehat{\mathbf{y}}$ erfüllt, also gilt

$$\begin{aligned} \|q'_m(\mathbf{A})\mathbf{r}^{(0)}\|_2 &= \|\mathbf{r}^{(0)} - \mathbf{A}\mathbf{y}\|_2 = \|\mathbf{r}^{(0)} - \mathbf{A}\mathbf{P}_{m-1}\widehat{\mathbf{y}}\|_2 \\ &\geq \|\mathbf{r}^{(0)} - \mathbf{A}\mathbf{P}_{m-1}\widehat{\mathbf{z}}^{(m)}\|_2 = \|q_m(\mathbf{A})\mathbf{r}^{(0)}\|_2, \end{aligned}$$

demzufolge minimiert q_m tatsächlich die Defektnorm. ■

Mit Hilfe dieser Bestapproximationsaussage lassen sich, wie schon im Fall des Gradienten- und des cg-Verfahrens, Aussagen über die Konvergenz des GMRES-Verfahrens gewinnen:

Satz 3.43 (Konvergenz) *Sei $\mathbf{A}$ regulär und diagonalisierbar, es gebe also eine reguläre Diagonalmatrix $\mathbf{D} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ und eine reguläre Matrix $\mathbf{T} \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}}$ so, dass $\mathbf{A} = \mathbf{T}\mathbf{D}\mathbf{T}^{-1}$ gilt. Sei $\mathbf{b} \in \mathbb{K}^{\mathcal{I}}$, und sei $\mathbf{x}^* := \mathbf{A}^{-1}\mathbf{b}$.*

1. *Falls $\sigma(\mathbf{A}) \subseteq [\alpha, \beta]$ für $\alpha, \beta \in \mathbb{R}_{>0}$ gilt, erfüllen die Iterierten des GMRES-Verfahrens die Abschätzung*

$$\|\mathbf{b} - \mathbf{A}\mathbf{x}^{(m)}\|_2 \leq \|\mathbf{T}\|_2 \|\mathbf{T}^{-1}\|_2 \frac{2c^m}{1 + c^{2m}} \|\mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}\|_2 \qquad \text{für alle } m \in \mathbb{N}_0$$

   und $c := \frac{\sqrt{\beta/\alpha}-1}{\sqrt{\beta/\alpha}+1}$.

2. *Falls $\sigma(\mathbf{A}) \subseteq [-\beta, -\alpha] \cup [\alpha, \beta]$ für $\alpha, \beta \in \mathbb{R}_{>0}$ gilt, erfüllen die Iterierten des GMRES-Verfahrens die Abschätzung*

$$\|\mathbf{b} - \mathbf{A}\mathbf{x}^{(m)}\|_2 \leq \|\mathbf{T}\|_2 \|\mathbf{T}^{-1}\|_2 \frac{2c^{\lfloor m/2 \rfloor}}{1 + c^{2\lfloor m/2 \rfloor}} \|\mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}\|_2 \qquad \text{für alle } m \in \mathbb{N}_0$$

   und $c := \frac{\beta-\alpha}{\beta+\alpha}$.

3. *Ganz allgemein gilt* $\mathbf{b} = \mathbf{A}\mathbf{x}^{(m_0)}$ *für* $m_0 \le n := \#\mathcal{I}$.

Beweis. Die dritte Aussage ist wegen (3.42) trivial.

Zum Beweis der ersten beiden Aussagen benutzen wir Lemma 3.42 und die Submultiplikativität der Spektralnorm, um

$$\|\mathbf{b} - \mathbf{A}\mathbf{x}^{(m)}\|_2 \le \|q_m(\mathbf{A})(\mathbf{b} - \mathbf{A}\mathbf{x}^{(0)})\|_2 \le \|q_m(\mathbf{A})\|_2 \|\mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}\|_2$$

für ein beliebiges Polynom q_m der Ordnung $\le m$ mit $q_m(0) = 1$ zu erhalten. Aus

$$\mathbf{A}^\ell = (\mathbf{T}\mathbf{D}\mathbf{T}^{-1})^\ell = \mathbf{T}\mathbf{D}^\ell\mathbf{T}^{-1}$$

folgt

$$\begin{aligned} \|q_m(\mathbf{A})\|_2 &= \|\mathbf{T}q_m(\mathbf{D})\mathbf{T}^{-1}\|_2 \le \|\mathbf{T}\|_2\|\mathbf{T}^{-1}\|_2\|q_m(\mathbf{D})\|_2 \\ &= \|\mathbf{T}\|_2\|\mathbf{T}^{-1}\|_2 \max\{|q_m(\lambda)| \ : \ \lambda \in \sigma(\mathbf{A})\}. \end{aligned} \tag{3.44}$$

Falls $\sigma(\mathbf{A}) \subseteq [\alpha, \beta]$ gilt, können wir ein transformiertes Tschebyscheff-Polynom p_m verwenden, das

$$\max\{|p_m(\xi)| \ : \ \xi \in [1-\beta, 1-\alpha]\} \le \frac{2c^m}{1+c^{2m}}$$

für $c := \frac{\sqrt{\beta/\alpha}-1}{\sqrt{\beta/\alpha}+1}$ und $p_m(1) = 1$ erfüllt. Wir setzen, wie schon im Beweis von Satz 3.30, wieder $q_m(\xi) := p_m(1-\xi)$ und erhalten

$$\begin{aligned} \max\{|q_m(\lambda)| \ : \ \lambda \in \sigma(\mathbf{A})\} &\le \max\{|q_m(\lambda)| \ : \ \lambda \in [\alpha, \beta]\} \\ &= \max\{|p_m(1-\lambda)| \ : \ \lambda \in [\alpha, \beta]\} \\ &= \max\{|p_m(\xi)| \ : \ \xi \in [1-\beta, 1-\alpha]\} \le \frac{2c^m}{1+c^{2m}} \end{aligned}$$

sowie $q_m(0) = 1$ und damit die gewünschte Abschätzung.

Falls $\sigma(\mathbf{A}) \subseteq [-\beta, -\alpha] \cup [\alpha, \beta]$ gilt, setzen wir $\mu := \lfloor m/2 \rfloor$ und wählen ein transformiertes Tschebyscheff-Polynom p_μ der Ordnung $\le \mu$, das auf dem Intervall $[1-\beta^2, 1-\alpha^2]$ besonders kleine Werte annimmt, also

$$\max\{|p_\mu(\xi)| \ : \ \xi \in [1-\beta^2, 1-\alpha^2]\} \le \frac{2c^\mu}{1+c^{2\mu}}$$

für $c := \frac{\sqrt{\beta^2/\alpha^2}-1}{\sqrt{\beta^2/\alpha^2}+1} = \frac{\beta-\alpha}{\beta+\alpha}$ und $p_\mu(1) = 1$ erfüllt. Wir setzen jetzt $q_m(\xi) := p_\mu(1-\xi^2)$ und erhalten

$$\begin{aligned} \max\{|q_m(\lambda)| \ : \ \lambda \in \sigma(\mathbf{A})\} &\le \max\{|q_m(\lambda)| \ : \ \lambda \in [-\beta, -\alpha] \cup [\alpha, \beta]\} \\ &= \max\{|p_\mu(1-\lambda^2)| \ : \ \lambda \in [-\beta, -\alpha] \cup [\alpha, \beta]\} \\ &= \max\{|p_\mu(\xi)| \ : \ \xi \in [1-\beta^2, 1-\alpha^2]\} \le \frac{2c^\mu}{1+c^{2\mu}}, \end{aligned}$$

also die gewünschte Abschätzung. ■

Wir sehen also, dass das GMRES-Verfahren auch bei indefiniten Problemen konvergiert, solange die Matrix $\mathbf{A}$ diagonalisierbar ist. Allerdings ist die Konvergenzrate in diesem Fall offenbar deutlich reduziert und entspricht ungefähr dem Ergebnis, das man auch für das cg-Verfahren angewendet auf die Normalengleichung

$$\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b}$$

erwarten würde. Falls die Matrix $\mathbf{A}$ *normal* ist, also die Gleichung $\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^*$ erfüllt, kann die Matrix $\mathbf{T}$ in Satz 3.43 orthogonal gewählt werden. In diesem Fall verschwinden die Terme $\|\mathbf{T}\|_2$ und $\|\mathbf{T}^{-1}\|_2 = \|\mathbf{T}^*\|_2$, und wir erhalten eine Abschätzung, die der für das cg-Verfahren entspricht.

Leider gibt es Fälle, in denen sich die ersten beiden Abschätzungen von Satz 3.43 nicht verwenden lassen und wir lediglich auf seine letzte Aussage angewiesen sind. Für ein $n \in \mathbb{N}$ betrachten wir als Beispiel die Matrix $\mathbf{F} \in \mathbb{K}^{n\times n}$, die durch

$$F_{ij} := \begin{cases} 1 & \text{falls } i = j+1 \text{ oder } i+n = j+1, \\ 0 & \text{sonst} \end{cases} \qquad \text{für alle } i,j \in [1:n]$$

gegeben ist. Wir wählen eine rechte Seite $\mathbf{b} \in \mathbb{K}^n$ gemäß

$$b_i := \begin{cases} 1 & \text{falls } i = 1, \\ 0 & \text{sonst} \end{cases} \qquad \text{für alle } i \in [1:n]$$

und wollen das Gleichungssystem $\mathbf{F}\mathbf{x} = \mathbf{b}$ lösen, das die Gestalt

$$\begin{pmatrix} 0 & \dots & 0 & 1 \\ 1 & \ddots & \ddots & 0 \\ & \ddots & \ddots & \vdots \\ & & 1 & 0 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

aufweist. Wir können sehen, dass die Spalten von $\mathbf{F}$ eine Orthogonalbasis bilden, also ist $\mathbf{F}$ nicht nur regulär, sondern sogar orthogonal, also insbesondere auch sehr gut konditioniert.

Wir wollen nun das Verhalten des GMRES-Verfahrens für dieses Gleichungssystem untersuchen. Wir beginnen mit dem Startvektor $\mathbf{x}^{(0)} = \mathbf{0}$, also dem Startdefekt $\mathbf{r}^{(0)} = \mathbf{b}$, und konstruieren die Arnoldi-Basis. Da $\mathbf{F}$ orthogonal und $\mathbf{r}^{(0)}$ der erste kanonische Einheitsvektor ist, sind die Suchrichtungen bereits durch $\mathbf{p}^{(m+1)} = \mathbf{F}\mathbf{p}^{(m)}$ gegeben. Mittels einer einfachen Induktion können wir

$$p_i^{(m)} = \begin{cases} 1 & \text{falls } i = m+1, \\ 0 & \text{sonst} \end{cases} \qquad \text{für alle } m \in [0:n-1],\ i \in [1:n]$$

nachweisen, da $\mathbf{F}$ den m-ten kanonischen Einheitsvektor gerade auf den $(m+1)$-ten abbildet. Darüber hinaus gilt

$$\mathbf{F}\mathbf{p}^{(n-1)} = \mathbf{p}^{(0)},$$

da $\mathbf{p}^{(n-1)}$ der n-te kanonische Einheitsvektor ist, der durch $\mathbf{F}$ gerade auf den ersten Einheitsvektor abgebildet wird.

Für alle $\ell \in [0 : n-2]$ gilt demzufolge

$$(\mathbf{F}\mathbf{p}^{(\ell)})_1 = p_1^{(\ell+1)} = 0.$$

Sei $m \in [1 : n-1]$, und sei $\mathbf{y} \in \mathcal{K}(\mathbf{r}^{(0)}, m-1)$. Dann existieren $\alpha_0, \ldots, \alpha_{m-1} \in \mathbb{K}$ mit

$$\mathbf{y} = \sum_{\ell=0}^{m-1} \alpha_\ell \mathbf{p}^{(\ell)},$$

und wir erhalten

$$(\mathbf{F}\mathbf{y})_1 = \sum_{\ell=0}^{m-1} \alpha_\ell (\mathbf{F}\mathbf{p}^{(\ell)})_1 = \sum_{\ell=0}^{m-1} \alpha_\ell p_1^{(\ell+1)} = 0,$$

also insbesondere

$$\|\mathbf{F}\mathbf{y} - \mathbf{r}^{(0)}\|_2 \geq \|\mathbf{r}^{(0)}\|_2 = 1.$$

Da wir diese Abschätzung für alle $\mathbf{y} \in \operatorname{span} \mathcal{K}(\mathbf{r}^{(0)}, m-1)$ gezeigt haben, folgt insbesondere auch

$$\|\mathbf{F}\mathbf{x}^{(m)} - \mathbf{r}^{(0)}\|_2 = 1 \qquad \text{für alle } m \in [0 : n-1],$$

der Iterationsfehler wird also in den ersten $n-1$ Iterationsschritten völlig unverändert bleiben. Im n-ten Schritt gilt dann

$$\mathbf{F}\mathbf{p}^{(n-1)} = \mathbf{p}^{(0)} = \mathbf{r}^{(0)},$$

und $\mathbf{x}^{(n)}$ ist die exakte Lösung.

Wir können also sehen, dass es Situationen gibt, in denen das GMRES-Verfahren erst im letzten Schritt eine sinnvolle Approximation der Lösung berechnet und zwischendurch völlig unsinnige Suchrichtungen ausprobiert.

Immerhin berechnet GMRES für eine beliebige reguläre Matrix eine Lösung, auch wenn diese Berechnung, wie gesehen, im schlimmsten Fall so lange wie bei einem direkten Verfahren dauern kann. Also werden wir wagemutig: Was passiert, wenn $\mathbf{A}$ nicht regulär ist? Eine elegante Antwort findet sich in dem Artikel "The Idea Behind Krylov Methods" von Ipsen und Meyer: Mit Hilfe der Jordan-Zerlegung können wir $\mathbf{A}$ in der Form

$$\mathbf{A} = \mathbf{T} \begin{pmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{N} \end{pmatrix} \mathbf{T}^{-1}$$

darstellen, wobei $\mathbf{C}$ eine reguläre und $\mathbf{N}$ eine nilpotente Matrix ist, also $\mathbf{N}^i = \mathbf{0}$ für ein $i \in \mathbb{N}$ gilt. Wir beschränken uns wieder auf den Fall $\mathbf{x}^{(0)} = \mathbf{0}$, müssen also klären, ob eine Lösung $\mathbf{x}^*$ von $\mathbf{A}\mathbf{x}^* = \mathbf{b}$ sich in einem Krylow-Raum $\mathcal{K}(\mathbf{r}^{(0)}, m) = \mathcal{K}(\mathbf{b}, m)$ darstellen lässt.

Nehmen wir zunächst an, dass das der Fall ist, dass es also Koeffizienten $\alpha_0, \ldots, \alpha_m \in \mathbb{K}$ mit

$$\mathbf{x}^* = \sum_{\ell=0}^{m} \alpha_\ell \mathbf{A}^\ell \mathbf{b}$$

gibt. Wir führen die transformierten Vektoren

$$\widehat{\mathbf{x}}^* := \mathbf{T}^{-1}\mathbf{x}^*, \qquad \widehat{\mathbf{b}} := \mathbf{T}^{-1}\mathbf{b}$$

ein und erhalten

$$\widehat{\mathbf{x}}^* = \sum_{\ell=0}^{m} \alpha_\ell \begin{pmatrix} \mathbf{C}^\ell & \mathbf{0} \\ \mathbf{0} & \mathbf{N}^\ell \end{pmatrix} \widehat{\mathbf{b}}.$$

Wir zerlegen $\widehat{\mathbf{x}}^*$ und $\widehat{\mathbf{b}}$ passend zur Blockmatrix in

$$\widehat{\mathbf{x}}^* = \begin{pmatrix} \widehat{\mathbf{x}}_1^* \\ \widehat{\mathbf{x}}_2^* \end{pmatrix}, \qquad \widehat{\mathbf{b}} = \begin{pmatrix} \widehat{\mathbf{b}}_1 \\ \widehat{\mathbf{b}}_2 \end{pmatrix}$$

und gelangen zu den Gleichungen

$$\begin{aligned} \widehat{\mathbf{x}}_1^* &= \sum_{\ell=0}^{m} \alpha_\ell \mathbf{C}^\ell \widehat{\mathbf{b}}_1, \\ \widehat{\mathbf{x}}_2^* &= \sum_{\ell=0}^{m} \alpha_\ell \mathbf{N}^\ell \widehat{\mathbf{b}}_2. \end{aligned}$$

Aus $\mathbf{A}\mathbf{x}^* = \mathbf{b}$ folgt $\mathbf{N}\widehat{\mathbf{x}}_2^* = \widehat{\mathbf{b}}_2$, und indem wir die zweite Gleichung mit $\mathbf{N}$ multiplizieren erhalten wir

$$\widehat{\mathbf{b}}_2 = \sum_{\ell=0}^{m} \alpha_\ell \mathbf{N}^{\ell+1} \widehat{\mathbf{b}}_2$$

und somit

$$\mathbf{0} = \left(\mathbf{I} - \sum_{\ell=0}^{m} \alpha_\ell \mathbf{N}^{\ell+1} \right) \widehat{\mathbf{b}}_2.$$

Da $\mathbf{N}$ nilpotent ist, gilt $\sigma(\mathbf{N}) = \{0\}$, also ist die Matrix in dieser Gleichung invertierbar und es folgt $\widehat{\mathbf{b}}_2 = \mathbf{0}$.

Falls umgekehrt $\widehat{\mathbf{b}}_2 = \mathbf{0}$ gilt, löst der durch $\widehat{\mathbf{x}}_1^* := \mathbf{C}^{-1}\widehat{\mathbf{b}}_1$ und $\widehat{\mathbf{x}}_2^* = \mathbf{0}$ gegebene Vektor

$$\mathbf{x} := \mathbf{T} \begin{pmatrix} \widehat{\mathbf{x}}_1^* \\ \widehat{\mathbf{x}}_2^* \end{pmatrix}$$

das Gleichungssystem. Wie schon in Abschnitt 1.3 können wir ein Polynom p konstruieren, das $p(\mathbf{C}) = \mathbf{C}^{-1}$ erfüllt. Daraus folgt

$$p(\mathbf{A})\mathbf{b} = \mathbf{T} \begin{pmatrix} p(\mathbf{C}) & \mathbf{0} \\ \mathbf{0} & p(\mathbf{N}) \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{b}}_1 \\ \widehat{\mathbf{b}}_2 \end{pmatrix} = \mathbf{T} \begin{pmatrix} p(\mathbf{C})\widehat{\mathbf{b}}_1 \\ p(\mathbf{N})\widehat{\mathbf{b}}_2 \end{pmatrix} = \mathbf{T} \begin{pmatrix} \mathbf{C}^{-1}\widehat{\mathbf{b}}_1 \\ \mathbf{0} \end{pmatrix} = \mathbf{T} \begin{pmatrix} \widehat{\mathbf{x}}_1^* \\ \widehat{\mathbf{x}}_2^* \end{pmatrix} = \mathbf{x},$$

also kann das GMRES-Verfahren die Lösung $\mathbf{x}$ tatsächlich bestimmen.

Unser Verfahren funktioniert also genau dann, wenn $\widehat{\mathbf{b}}_2 = \mathbf{0}$ gilt. Wegen

$$\mathbf{A}^i = \mathbf{T}\begin{pmatrix}\mathbf{C}^i & \mathbf{0}\\ \mathbf{0} & \mathbf{N}^i\end{pmatrix}\mathbf{T}^{-1} = \mathbf{T}\begin{pmatrix}\mathbf{C}^i & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{T}^{-1}$$

ist das äquivalent dazu, dass die rechte Seite $\mathbf{b}$ im Bild von $\mathbf{A}^i$ liegt. Da das Bild von $\mathbf{A}^i$ im Allgemeinen eine echte Teilmenge des Bilds von $\mathbf{A}$ ist, kann es rechte Seiten geben, für die zwar eine Lösung existiert, diese Lösung aber nicht mit einem Krylow-Verfahren konstruiert werden kann.

Die Zahl i kann durch eins plus der Differenz zwischen der algebraischen und der geometrischen Vielfachheit des Eigenwerts 0 der Matrix $\mathbf{A}$ beschränkt werden. Für selbstadjungierte Matrizen, bei denen algebraische und geometrische Vielfachheiten übereinstimmen, ist es also ausreichend, dass $\mathbf{b}$ im Bild von $\mathbf{A}$ liegt.

Die Konstruktion der orthogonalen Basen $\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(m)}$ erfordert es, alle Basisvektoren abzuspeichern, so dass für den m-ten Schritt des GMRES-Verfahrens m Vektoren der Länge $n = \#\mathcal{I}$ abgespeichert werden müssen. Da mit allen diesen Vektoren Skalarprodukte berechnet und Linearkombinationen gebildet werden müssen, erfordert der m-te Schritt des Verfahrens einen Rechenaufwand, der zu mn proportional wächst, die Durchführung des Verfahrens wird also immer aufwendiger, je höher m wird.

Im Allgemeinen lässt sich dieses Problem nicht vermeiden. Man kann zwar die Dimension der Basis künstlich beschränken, also bei Erreichen einer Schranke m_+ die Konstruktion beenden, die neue Iterierte berechnen und das Verfahren mit der neuen Iterierten als Ausgangsvektor neu beginnen, aber wir haben bereits gesehen, dass es Gleichungssysteme gibt, bei denen in diesem Fall keinerlei Konvergenz mehr auftritt.

Für den speziellen Fall selbstadjungierter (nicht unbedingt positiv definiter) Matrizen können wir das Verfahren so modifizieren, dass Speicher- und Zeitbedarf proportional zum cg-Verfahren sind, dass also insbesondere ein Schritt mit einem zu n proportionalen Aufwand durchgeführt werden kann: Wenn $\mathbf{A}$ selbstadjungiert ist, gilt

$$\langle \mathbf{p}^{(\ell)}, \mathbf{A}\mathbf{p}^{(k)}\rangle_2 = \langle \mathbf{A}\mathbf{p}^{(\ell)}, \mathbf{p}^{(k)}\rangle_2 = \overline{\langle \mathbf{p}^{(k)}, \mathbf{A}\mathbf{p}^{(\ell)}\rangle_2} \qquad \text{für alle } \ell, k \in \{0, \dots, m_0 - 1\},$$

und Lemma 3.39 impliziert $\widehat{A}_{\ell k}^{(m)} = 0$ für alle $\ell \in \{0, \dots, m-1\}$, $k \in \{0, \dots, m-2\}$ mit $|\ell - k| > 1$, also besitzt $\widehat{\mathbf{A}}^{(m)}$ in diesem Fall nicht nur Hessenberg-, sondern sogar Tridiagonalstruktur.

Die zur Lösung des linearen Ausgleichsproblems verwendeten Givens-Rotationen transformieren die Tridiagonalmatrix in eine obere Dreiecksmatrix der Bandbreite 2, so dass sich das Ausgleichsproblem mit einem Aufwand proportional zu m (statt m^2 im allgemeinen Fall) auflösen lässt.

Indem man die Tatsache ausnutzt, dass in jedem Schritt des Verfahrens jeweils nur eine Spalte zu dieser Matrix hinzugefügt wird, lässt sich eine Umformulierung des Algorithmus finden, bei der die Matrizen $\widehat{\mathbf{A}}^{(m)}$ überhaupt nicht mehr aufgestellt werden müssen, die Struktur ähnelt dann sehr der des cg-Verfahrens, und das resultierende *MINRES-Verfahren* ist ähnlich effizient. Da es dieselben Semiiterierten wie das GMRES-Verfahren berechnet, übertragen sich die Konvergenzaussagen aus Lemma 3.42 und Satz 3.43 direkt auf das neue Verfahren.

3.6 Verfahren für Sattelpunktprobleme

Wie wir bereits gesehen haben, sind die Konvergenzeigenschaften der semiiterativen Verfahren für indefinite Matrizen in der Regel deutlich schlechter als im positiv definiten Fall. Für eine wichtige Klasse indefiniter Probleme, nämlich die *Sattelpunktprobleme*, lassen sich Verfahren konstruieren, die die guten Konvergenzeigenschaften des positiv definiten Falls erreichen.

Bei einem typischen Sattelpunktproblem zerfällt die Indexmenge in zwei disjunkte Teilmengen $\mathcal{I}_1, \mathcal{I}_2$ mit

$$\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2, \qquad \mathcal{I}_1 \cap \mathcal{I}_2 = \emptyset$$

und die Systemmatrix besitzt die Gestalt

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{0} \end{pmatrix},$$

wobei $\mathbf{A}_{11} \in \mathbb{K}^{\mathcal{I}_1 \times \mathcal{I}_1}$ eine selbstadjungierte positiv definite Matrix ist und $\mathbf{A}_{12} = \mathbf{A}_{21}^* \in \mathbb{K}^{\mathcal{I}_1 \times \mathcal{I}_2}$ injektiv ist. Damit ist $\mathbf{A}$ selbstadjungiert. Um einen Eindruck von den Eigenwerten der Matrix $\mathbf{A}$ zu erhalten, verwenden wir eine durch die Gauß-Elimination inspirierte Kongruenztransformation:

$$\begin{pmatrix} \mathbf{I} & \\ -\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ & -\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \end{pmatrix}, \tag{3.45a}$$

$$\begin{pmatrix} \mathbf{I} & \\ -\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \end{pmatrix} \begin{pmatrix} \mathbf{I} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \\ & \mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{11} & \\ & -\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \end{pmatrix}. \tag{3.45b}$$

Da $\mathbf{A}_{12}$ injektiv ist, ist $\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$ positiv definit, also besitzt die transformierte Matrix in der Gleichung (3.45b) sowohl positive Eigenwerte (im linken oberen Diagonalblock) als auch negative (im rechten unteren). Diese Eigenschaft überträgt sich auf $\mathbf{A}$, so dass wir tatsächlich ein indefinites Problem zu untersuchen haben.

Trotzdem können wir ein Lösungsverfahren konstruieren, indem wir die Gleichung (3.45a) verwenden: Das Gleichungssystem

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix}$$

wird durch eine Block-Gauß-Elimination in das System

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ & -\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{b}_1 \end{pmatrix}$$

überführt, das wir per Rückwärtseinsetzen lösen können. Dazu sind die beiden Gleichungen

$$\mathbf{S}\mathbf{x}_2 = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{b}_1 - \mathbf{b}_2 \tag{3.46a}$$

$$\mathbf{A}_{11}\mathbf{x}_1 = \mathbf{b}_1 - \mathbf{A}_{12}\mathbf{x}_2, \tag{3.46b}$$

mit dem *Schur-Komplement* $\mathbf{S} = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$ äquivalent. Das Schur-Komplement ist positiv definit, so dass wir die Gleichung (3.46a) mit den bereits bekannten Verfahren behandeln können. Als Beispiel untersuchen wir das Gradientenverfahren: Für einen Startvektor $\mathbf{x}_2^{(0)} \in \mathbb{K}^{\mathcal{I}_2}$ ist das Residuum durch

$$\begin{aligned}\mathbf{r}_2^{(0)} &:= \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{b}_1 - \mathbf{b}_2 - \mathbf{S}x_2^{(0)} = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{b}_1 - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{x}_2^{(0)} - \mathbf{b}_2\\ &= \mathbf{A}_{21}\mathbf{A}_{11}^{-1}(\mathbf{b}_1 - \mathbf{A}_{12}\mathbf{x}_2^{(0)}) - \mathbf{b}_2\end{aligned} \tag{3.47}$$

gegeben, kann also effizient berechnet werden, sofern sich Gleichungssysteme mit der Matrix $\mathbf{A}_{11}$ schnell lösen lassen.

Wir untersuchen nun die Berechnung der Iterierten $\mathbf{x}_2^{(m+1)}$ aus $\mathbf{x}_2^{(m)}$ für ein $m \in \mathbb{N}_0$. Für die Berechnung des optimalen Dämpfungsparameters benötigen wir den Vektor

$$\mathbf{s}^{(m)} := \mathbf{S}\mathbf{r}_2^{(m)} = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{r}_2^{(m)}, \tag{3.48}$$

wir müssen also ein weiteres Gleichungssystem mit der Matrix $\mathbf{A}_{11}$ lösen. Der Dämpfungsparameter ist gegeben durch

$$\lambda_{\text{opt}}^{(m)} := \frac{\langle \mathbf{r}_2^{(m)}, \mathbf{r}_2^{(m)}\rangle_2}{\langle \mathbf{s}^{(m)}, \mathbf{r}_2^{(m)}\rangle_2},$$

die nächste Iterierte und das nächste Residuum berechnen sich dann gemäß

$$\begin{aligned}\mathbf{x}_2^{(m+1)} &:= \mathbf{x}_2^{(m)} + \lambda_{\text{opt}}^{(m)}\mathbf{r}_2^{(m)},\\ \mathbf{r}_2^{(m+1)} &:= \mathbf{r}_2^{(m)} - \lambda_{\text{opt}}^{(m)}\mathbf{s}^{(m)}.\end{aligned} \tag{3.49}$$

Ein Schritt des Gradientenverfahrens erfordert also die Multiplikation mit den Matrizen $\mathbf{A}_{12}$ und $\mathbf{A}_{21}$ sowie das Lösen eines Gleichungssystems mit der Matrix $\mathbf{A}_{11}$.

Die Idee des *Uzawa-Verfahrens* besteht darin, in jedem Schritt des Gradientenverfahrens auch eine Näherung des Vektors $\mathbf{x}_1$ zu bestimmen, indem die einzelnen Rechenoperationen geschickt arrangiert werden. Beispielsweise sehen wir, dass in der Gleichung (3.47) der Vektor $\mathbf{A}_{11}^{-1}(\mathbf{b}_1 - \mathbf{A}_{12}\mathbf{x}_2)$ als Zwischenergebnis auftritt, der nach (3.46b) gerade $\mathbf{x}_1$ entspricht.

Wenn wir $\mathbf{x}_2^{(m+1)}$ mit der Formel (3.49) aktualisieren, ergibt sich aus (3.46b) die Korrekturgleichung

$$\begin{aligned}\mathbf{x}_1^{(m+1)} &= \mathbf{A}_{11}^{-1}(\mathbf{b}_1 - \mathbf{A}_{12}\mathbf{x}_2^{(m+1)}) = \mathbf{A}_{11}^{-1}(\mathbf{b}_1 - \mathbf{A}_{12}\mathbf{x}_2^{(m)} - \lambda_{\text{opt}}^{(m)}\mathbf{A}_{12}\mathbf{r}_2^{(m)})\\ &= \mathbf{x}_1^{(m)} - \lambda_{\text{opt}}^{(m)}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{r}_2^{(m)},\end{aligned}$$

und der rechte Vektor tritt als Zwischenergebnis bei der Berechnung des Vektors $\mathbf{s}^{(m)}$ in (3.48) auf, steht uns also ebenfalls ohne weiteren Rechenaufwand zur Verfügung.

Um die Konvergenz des Uzawa-Gradientenverfahrens mit Hilfe des Satzes 3.19 untersuchen zu können, ist es erforderlich, Schranken für das Spektrum des Schur-Komplements $\mathbf{S}$ zu finden.

procedure UzawaGrad($\mathbf{b}_1$, $\mathbf{b}_2$, **var** $\mathbf{x}_1$, $\mathbf{x}_2$);
Löse $\mathbf{A}_{11}\mathbf{x}_1 = \mathbf{b}_1 - \mathbf{A}_{12}\mathbf{x}_2$;
$\mathbf{r}_2 \leftarrow \mathbf{A}_{21}\mathbf{x}_1 - \mathbf{b}_2$;
while Fehler zu groß **do**
  Löse $\mathbf{A}_{11}\mathbf{a} = \mathbf{A}_{12}\mathbf{r}_2$;
  $\mathbf{s} \leftarrow \mathbf{A}_{21}\mathbf{a}$;
  $\lambda_{\text{opt}} \leftarrow \frac{\|\mathbf{r}_2\|_2^2}{\langle \mathbf{s}, \mathbf{r}_2 \rangle_2}$;
  $\mathbf{x}_2 \leftarrow \mathbf{x}_2 + \lambda_{\text{opt}}\mathbf{r}_2$;
  $\mathbf{r}_2 \leftarrow \mathbf{r}_2 - \lambda_{\text{opt}}\mathbf{s}$;
  $\mathbf{x}_1 \leftarrow \mathbf{x}_1 - \lambda_{\text{opt}}\mathbf{a}$
end while

Abbildung 3.13: Uzawa-Iteration basierend auf dem Gradientenverfahren

Satz 3.44 (Spektralschranken) *Seien* $\alpha, \beta \in \mathbb{R}_{>0}$ *so gegeben, dass die* inf-sup-Bedingung

$$\sqrt{\alpha}\|\mathbf{x}_2\|_2 \leq \sup\left\{\frac{\langle \mathbf{A}_{12}\mathbf{x}_2, \mathbf{x}_1\rangle_2}{\|\mathbf{x}_1\|_{A_{11}}} \;:\; \mathbf{x}_1 \in \mathbb{K}^{\mathcal{I}_1} \setminus \{\mathbf{0}\}\right\} \qquad \textit{für alle } \mathbf{x}_2 \in \mathbb{K}^{\mathcal{I}_2} \tag{3.50}$$

und die Stetigkeitsbedingung

$$\|\mathbf{A}_{21}\mathbf{x}_1\|_2 \leq \sqrt{\beta}\|\mathbf{x}_1\|_{A_{11}} \qquad \textit{für alle } \mathbf{x}_1 \in \mathbb{K}^{\mathcal{I}_1} \tag{3.51}$$

gelten. Dann folgt $\sigma(\mathbf{S}) \subseteq [\alpha, \beta]$.

Beweis. Sei $\mathbf{x}_2 \in \mathbb{K}^{\mathcal{I}_2}$. Aus (3.50) folgt

$$\begin{aligned}
\sqrt{\alpha}\|\mathbf{x}_2\|_2 &\leq \sup\left\{\frac{\langle \mathbf{A}_{12}\mathbf{x}_2, \mathbf{x}_1\rangle_2}{\|\mathbf{x}_1\|_{A_{11}}} \;:\; \mathbf{x}_1 \in \mathbb{K}^{\mathcal{I}_1} \setminus \{\mathbf{0}\}\right\} \\
&= \sup\left\{\frac{\langle \mathbf{A}_{12}\mathbf{x}_2, \mathbf{x}_1\rangle_2}{\|\mathbf{A}_{11}^{1/2}\mathbf{x}_1\|_2} \;:\; \mathbf{x}_1 \in \mathbb{K}^{\mathcal{I}_1} \setminus \{\mathbf{0}\}\right\} \\
&= \sup\left\{\frac{\langle \mathbf{A}_{12}\mathbf{x}_2, \mathbf{A}_{11}^{-1/2}\mathbf{x}_1\rangle_2}{\|\mathbf{x}_1\|_2} \;:\; \mathbf{x}_1 \in \mathbb{K}^{\mathcal{I}_1} \setminus \{\mathbf{0}\}\right\} \\
&= \sup\left\{\frac{\langle \mathbf{A}_{11}^{-1/2}\mathbf{A}_{12}\mathbf{x}_2, \mathbf{x}_1\rangle_2}{\|\mathbf{x}_1\|_2} \;:\; \mathbf{x}_1 \in \mathbb{K}^{\mathcal{I}_1} \setminus \{\mathbf{0}\}\right\} = \|\mathbf{A}_{11}^{-1/2}\mathbf{A}_{12}\mathbf{x}_2\|_2,
\end{aligned}$$

wobei im letzten Schritt die Cauchy-Schwarz-Ungleichung zum Einsatz kommt. Wir erhalten

$$\begin{aligned}
\alpha\langle \mathbf{x}_2, \mathbf{x}_2\rangle_2 = \alpha\|\mathbf{x}_2\|_2^2 \leq \|\mathbf{A}_{11}^{-1/2}\mathbf{A}_{12}\mathbf{x}_2\|_2^2 &= \langle \mathbf{A}_{11}^{-1/2}\mathbf{A}_{12}\mathbf{x}_2, \mathbf{A}_{11}^{-1/2}\mathbf{A}_{12}\mathbf{x}_2\rangle_2 \\
&= \langle \mathbf{A}_{12}^{*}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{x}_2, \mathbf{x}_2\rangle_2 = \langle \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{x}_2, \mathbf{x}_2\rangle_2 = \langle \mathbf{S}\mathbf{x}_2, \mathbf{x}_2\rangle_2,
\end{aligned}$$

also $\alpha\mathbf{I} \leq \mathbf{S}$.

Aus (3.51) folgt

$$\|\mathbf{A}_{21}\widehat{\mathbf{x}}_1\|_2 \le \sqrt{\beta}\|\widehat{\mathbf{x}}_1\|_{A_{11}} = \sqrt{\beta}\|\mathbf{A}_{11}^{1/2}\widehat{\mathbf{x}}_1\|_2 \qquad \text{für alle } \widehat{\mathbf{x}}_1 \in \mathbb{K}^{\mathcal{I}_1},$$

und mit $\widehat{\mathbf{x}}_1 = \mathbf{A}_{11}^{-1/2}\mathbf{x}_1$ erhalten wir

$$\|\mathbf{A}_{21}\mathbf{A}_{11}^{-1/2}\mathbf{x}_1\|_2 \le \sqrt{\beta}\|\mathbf{x}_1\|_2 \qquad \text{für alle } \mathbf{x}_1 \in \mathbb{K}^{\mathcal{I}_1}.$$

Wir wählen $\mathbf{x}_1 := \mathbf{A}_{11}^{-1/2}\mathbf{A}_{12}\mathbf{x}_2$ und finden

$$\begin{aligned}\|\mathbf{S}\mathbf{x}_2\|_2^2 &= \|\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{x}_2\|_2^2 = \|\mathbf{A}_{21}\mathbf{A}_{11}^{-1/2}\mathbf{x}_1\|_2^2 \le \beta\|\mathbf{x}_1\|_2^2 = \beta\langle\mathbf{x}_1,\mathbf{x}_1\rangle_2\\ &= \beta\langle\mathbf{A}_{11}^{-1/2}\mathbf{A}_{12}\mathbf{x}_2, \mathbf{A}_{11}^{-1/2}\mathbf{A}_{12}\mathbf{x}_2\rangle_2 = \beta\langle\mathbf{A}_{12}^*\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{x}_2,\mathbf{x}_2\rangle_2\\ &= \beta\langle\mathbf{S}\mathbf{x}_2,\mathbf{x}_2\rangle_2 \le \beta\|\mathbf{S}\mathbf{x}_2\|_2\|\mathbf{x}_2\|_2.\end{aligned}$$

Falls $\mathbf{S}\mathbf{x}_2 \ne \mathbf{0}$ gilt, können wir auf beiden Seiten durch $\|\mathbf{S}\mathbf{x}_2\|_2$ dividieren und erhalten

$$\|\mathbf{S}\mathbf{x}_2\|_2 \le \beta\|\mathbf{x}_2\|_2.$$

Mit der Cauchy-Schwarz-Ungleichung folgt daraus

$$\langle\mathbf{S}\mathbf{x}_2,\mathbf{x}_2\rangle \le \|\mathbf{S}\mathbf{x}_2\|_2\|\mathbf{x}_2\|_2 \le \beta\|\mathbf{x}_2\|_2^2 = \beta\langle\mathbf{x}_2,\mathbf{x}_2\rangle_2,$$

also haben wir auch $\mathbf{S} \le \beta\mathbf{I}$ bewiesen. Mit Lemma 2.49 folgt die Behauptung. ■

Folgerung 3.45 (Konvergenz) *Falls die Bedingungen (3.50) und (3.51) erfüllt sind, erfüllen die Iterierten des Uzawa-Gradientenverfahrens die Abschätzung*

$$\|\mathbf{x}_2^{(m+1)} - \mathbf{x}_2^*\|_S \le \frac{\beta-\alpha}{\beta+\alpha}\|\mathbf{x}_2^{(m)} - \mathbf{x}_2^*\|_S \qquad \textit{für alle } m \in \mathbb{N}_0.$$

Beweis. Wir kombinieren Satz 3.44 mit Satz 3.19. ■

Wir haben bereits gesehen, dass das cg-Verfahren eine wesentlich bessere Konvergenz als das Gradientenverfahren aufweist, also bietet es sich an, eine entsprechende Variante der Uzawa-Iteration zu konstruieren. Auch in diesem Fall lässt sich die erste Hälfte $\mathbf{x}_1$ des Lösungsvektors elegant mitführen, indem wir zu jeder Suchrichtung auch ihr Produkt mit $\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$ berechnen. Da dieser Vektor im Zuge der Berechnung des Schur-Komplements ohnehin vorkommt, kann die Aktualisierung des Vektors $\mathbf{x}_1$ mit einer einzigen Linearkombination erfolgen. Die resultierende *Uzawa-cg-Iteration* ist in Abbildung 3.14 zusammengefasst.

Folgerung 3.46 (Konvergenz) *Falls die Bedingungen (3.50) und (3.51) erfüllt sind, erfüllen die Iterierten des Uzawa-cg-Verfahrens die Abschätzung*

$$\|\mathbf{x}_2^{(m+1)} - \mathbf{x}_2^*\|_S \le \frac{2c^m}{1+c^{2m}}\|\mathbf{x}_2^{(m)} - \mathbf{x}_2^*\|_S \qquad \textit{für alle } m \in \mathbb{N}_0$$

mit den Konstanten

$$c := \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}, \qquad \kappa := \frac{\beta}{\alpha}.$$

Beweis. Wir kombinieren Satz 3.44 mit Satz 3.30. ■

procedure UzawaKonjGrad($\mathbf{b}_1$, $\mathbf{b}_2$, **var** $\mathbf{x}_1$, $\mathbf{x}_2$);
Löse $\mathbf{A}_{11}\mathbf{x}_1 = \mathbf{b}_1 - \mathbf{A}_{12}\mathbf{x}_2$;
$\mathbf{r}_2 \leftarrow \mathbf{A}_{21}\mathbf{x}_1 - \mathbf{b}_2$;
$\mathbf{p}_2 \leftarrow \mathbf{r}_2$;
while Fehler zu groß **do**
  Löse $\mathbf{A}_{11}\mathbf{a} = \mathbf{A}_{12}\mathbf{p}_2$;
  $\mathbf{s} \leftarrow \mathbf{A}_{21}\mathbf{a}$;
  $\lambda_{\text{opt}} \leftarrow \frac{\langle \mathbf{p}_2, \mathbf{r}_2 \rangle_2}{\langle \mathbf{s}, \mathbf{p}_2 \rangle_2}$;
  $\mathbf{x}_2 \leftarrow \mathbf{x}_2 + \lambda_{\text{opt}}\mathbf{p}_2$;
  $\mathbf{r}_2 \leftarrow \mathbf{r}_2 - \lambda_{\text{opt}}\mathbf{s}$;
  $\mathbf{x}_1 \leftarrow \mathbf{x}_1 - \lambda_{\text{opt}}\mathbf{a}$;
  $\mu \leftarrow \frac{\langle \mathbf{r}_2, \mathbf{s} \rangle_2}{\langle \mathbf{p}_2, \mathbf{s} \rangle_2}$;
  $\mathbf{p}_2 \leftarrow \mathbf{r}_2 - \mu\mathbf{p}_2$
end while

Abbildung 3.14: Uzawa-Iteration basierend auf dem cg-Verfahren

4 Mehrgitterverfahren

Alle in den vorangehenden Kapiteln vorgestellten Verfahren haben den Nachteil, dass der mit ihnen verbundene Rechenaufwand mit der Konditionszahl der Matrix $\mathbf{A}$ des zu lösenden linearen Gleichungssystems (1.1) wächst. Im Fall des Modellproblems bedeutet das, dass desto mehr Iterationsschritte erforderlich werden, je größer die Problemdimension n wird. Nicht nur wird also die Durchführung eines Schrittes aufwendiger (alle Verfahren benötigen mindestens n Rechenoperationen für einen Schritt), es sind auch immer mehr Schritte erforderlich, um eine hinreichend genaue Lösung zu bestimmen.

Infolge dieser Eigenschaft werden die bisher vorgestellten Verfahren für große Gleichungssysteme sehr zeitaufwendig oder sogar undurchführbar (etwa im Fall des GMRES-Verfahrens wegen des wachsenden Speicherbedarfs).

Wir suchen also nach einem Verfahren, das einerseits ähnlich einfach und effizient wie die bisher vorgestellten ist, andererseits aber ein Konvergenzverhalten zeigt, das möglichst unabhängig von der Konditionszahl ist. Für allgemeine lineare Gleichungssysteme ist kein Verfahren bekannt, das diese Eigenschaften aufweist. Für Gleichungssysteme wie das Modellproblem, die aus der Diskretisierung einer partiellen Differentialgleichung entstehen, lässt sich allerdings eine Klasse von Verfahren angeben, die die gewünschten Eigenschaften besitzen, nämlich die *Mehrgitterverfahren*, denen dieses Kapitel gewidmet ist.

4.1 Motivation: Zweigitterverfahren

Wir suchen nach einem Verfahren, das möglichst einfach zu implementieren sein soll, also ist es naheliegend, ein besonders einfaches Verfahren als Ausgangspunkt unserer Betrachtung zu wählen: Die Richardson-Iteration.

Für das eindimensionale Modellproblem haben wir bereits in Lemma 1.7 nachgerechnet, dass der optimale Dämpungsparameter durch

$$\theta_{\text{opt}} = \frac{h^2}{2}$$

gegeben ist und zu einer Konvergenzrate von

$$\varrho = 1 - 2\sin^2(\pi h/2)$$

führt, die Konvergenzrate wird sich also für $h \to 0$ ungefähr wie $1 - \pi^2 h^2/2$ verhalten.

Aus Gründen, die später klar werden, verwenden wir hier nicht θ_{opt}, sondern den halbierten Wert

$$\theta_{\text{gl}} := \frac{h^2}{4}.$$

Entsprechend unserer Theorie konvergiert das Richardson-Verfahren auch für diesen Dämpfungsparameter. Betrachten wir nun das Verhalten des Iterationsfehlers $\mathbf{x}^{(\ell)} - \mathbf{x}^*$, das bekanntlich durch die Gleichung

$$\mathbf{x}^{(\ell+1)} - \mathbf{x}^* = \mathbf{M}_{\mathrm{Rich},\theta_{\mathrm{gl}}}(\mathbf{x}^{(\ell)} - \mathbf{x}^*)$$

beschrieben ist. Falls wir annehmen, dass der Fehler ein Eigenvektor $\mathbf{e}^k$ der Matrix $\mathbf{L}$ ist, erhalten wir

$$\mathbf{M}_{\mathrm{Rich},\theta_{\mathrm{gl}}}\mathbf{e}^k = (1 - \theta_{\mathrm{gl}}\lambda_k)\mathbf{e}^k = (1 - \sin^2(\pi kh/2))\mathbf{e}^k,$$

falls $\sin^2(\pi kh/2)$ größer als $1/2$ ist, wird der Fehler also um einen Faktor von mindestens $1/2$ reduziert. Diese Bedingung ist für $kh \geq 1/2$, also $k \geq (n+1)/2$ erfüllt. Fehlerkomponenten, die in der „oberen Hälfte“ des Spektrums liegen, werden also durch diese modifizierte Richardson-Iteration um einen Faktor von mindestens $1/2$ reduziert. Dieses Verhalten stellt sich nur ein, wenn wir den Dämpfungsparameter geschickt wählen.

Der Index k des Eigenvektors $\mathbf{e}^k$ beschreibt (vgl. Lemma 1.7) die Frequenz der dem Eigenvektor zugrundeliegenden Sinusfunktion, ein großer Wert von k entspricht also einer *hochfrequenten* Funktion, während ein kleiner Wert zu einer *niedrigfrequenten* gehört.

Dementsprechend teilen wir nun auch die Eigenvektoren in hoch- und niedrigfrequente ein und bezeichnen die von ihnen aufgespannten Räume mit

$$\begin{aligned}
\mathcal{X}_{\mathrm{hf}} &:= \operatorname{span}\{\mathbf{e}^k \ : \ k \in \mathcal{I}_{\mathrm{hf}}\} && \text{mit } \mathcal{I}_{\mathrm{hf}} := \{k \in \mathcal{I} \ : \ k \geq (n+1)/2\},\\
\mathcal{X}_{\mathrm{nf}} &:= \operatorname{span}\{\mathbf{e}^k \ : \ k \in \mathcal{I}_{\mathrm{nf}}\} && \text{mit } \mathcal{I}_{\mathrm{nf}} := \{k \in \mathcal{I} \ : \ k < (n+1)/2\} = \mathcal{I} \setminus \mathcal{I}_{\mathrm{hf}}.
\end{aligned}$$

Diese beiden Räume bilden eine orthogonale Zerlegung des Raums $\mathbb{K}^{\mathcal{I}}$.

Lemma 4.1 (Orthogonale Zerlegung) *Für jeden Vektor $\mathbf{x} \in \mathbb{K}^{\mathcal{I}}$ existieren $\mathbf{y} \in \mathcal{X}_{\mathrm{hf}}$ und $\mathbf{z} \in \mathcal{X}_{\mathrm{nf}}$ mit $\mathbf{x} = \mathbf{y} + \mathbf{z}$.*

Es gilt

$$\langle \mathbf{y}, \mathbf{z} \rangle_2 = 0 \qquad \text{für alle } \mathbf{y} \in \mathcal{X}_{\mathrm{hf}},\ \mathbf{z} \in \mathcal{X}_{\mathrm{nf}}.$$

Beweis. Sei $\mathbf{x} \in \mathbb{K}^{\mathcal{I}}$. Da die Eigenvektoren nach Lemma 1.7 eine Orthonormalbasis des Raums $\mathbb{K}^{\mathcal{I}}$ bilden, existieren Koeffizienten $(\alpha_k)_{k\in\mathcal{I}}$ mit

$$\mathbf{x} = \sum_{k\in\mathcal{I}} \alpha_k \mathbf{e}^k = \underbrace{\sum_{k\in\mathcal{I}_{\mathrm{hf}}} \alpha_k \mathbf{e}^k}_{=:\mathbf{y}} + \underbrace{\sum_{k\in\mathcal{I}_{\mathrm{nf}}} \alpha_k \mathbf{e}^k}_{=:\mathbf{z}},$$

also folgt $\mathbf{x} = \mathbf{y} + \mathbf{z}$ mit $\mathbf{y} \in \mathcal{X}_{\mathrm{hf}}$ und $\mathbf{z} \in \mathcal{X}_{\mathrm{nf}}$.

Seien nun $\mathbf{y} \in \mathcal{X}_{\mathrm{hf}}$ und $\mathbf{z} \in \mathcal{X}_{\mathrm{nf}}$ gegeben. Dann existieren Koeffizienten $(\beta_k)_{k\in\mathcal{I}_{\mathrm{hf}}}$ und $(\gamma_k)_{k\in\mathcal{I}_{\mathrm{nf}}}$ mit

$$\mathbf{y} = \sum_{k\in\mathcal{I}_{\mathrm{hf}}} \beta_k \mathbf{e}^k, \qquad \mathbf{z} = \sum_{\ell\in\mathcal{I}_{\mathrm{nf}}} \gamma_\ell \mathbf{e}^\ell.$$

Indem wir wieder ausnutzen, dass die Eigenvektoren eine Orthonormalbasis bilden, folgt

$$\langle \mathbf{y}, \mathbf{z} \rangle_2 = \langle \sum_{k \in \mathcal{I}_{\mathrm{hf}}} \beta_k \mathbf{e}^k, \sum_{\ell \in \mathcal{I}_{\mathrm{nf}}} \gamma_\ell \mathbf{e}^\ell \rangle_2 = \sum_{k \in \mathcal{I}_{\mathrm{hf}}} \sum_{\ell \in \mathcal{I}_{\mathrm{nf}}} \bar{\beta}_k \gamma_\ell \underbrace{\langle \mathbf{e}^k, \mathbf{e}^\ell \rangle_2}_{=0} = 0.$$

■

Wir können den Raum $\mathbb{K}^{\mathcal{I}}$ in diese beiden Teilräume zerlegen und erhalten die folgenden Abschätzungen für den Iterationsfehler:

Lemma 4.2 (Konvergenz im Teilraum) *Es gilt*

$$\|\mathbf{M}_{\mathrm{Rich},\theta_{\mathrm{gl}}} \mathbf{x}\|_2 \leq \frac{1}{2} \|\mathbf{x}\|_2 \qquad \textit{für alle } \mathbf{x} \in \mathcal{X}_{\mathrm{hf}}.$$

Beweis. Sei $\mathbf{x} \in \mathcal{X}_{\mathrm{hf}}$, und seien Koeffizienten $(\alpha_k)_{k \in \mathcal{I}_{\mathrm{hf}}}$ mit

$$\mathbf{x} = \sum_{k \in \mathcal{I}_{\mathrm{hf}}} \alpha_k \mathbf{e}^k$$

gegeben. Laut Lemma 1.7 bilden die Eigenvektoren eine Orthonormalbasis, also folgt

$$\|\mathbf{x}\|_2^2 = \langle \mathbf{x}, \mathbf{x} \rangle_2 = \sum_{k \in \mathcal{I}_{\mathrm{hf}}} \sum_{\ell \in \mathcal{I}_{\mathrm{hf}}} \alpha_k \bar{\alpha}_\ell \langle \mathbf{e}^k, \mathbf{e}^\ell \rangle_2 = \sum_{k \in \mathcal{I}_{\mathrm{hf}}} |\alpha_k|^2. \tag{4.1}$$

Wir haben in (1.6) bereits

$$\mathbf{M}_{\mathrm{Rich},\theta_{\mathrm{gl}}} \mathbf{e}^k = (1 - \theta_{\mathrm{gl}} \lambda_k) \mathbf{e}^k = (1 - \sin^2(\pi k h/2)) \mathbf{e}^k \qquad \text{für alle } k \in \mathcal{I}$$

bewiesen, so dass wir insgesamt

$$\mathbf{M}_{\mathrm{Rich},\theta_{\mathrm{gl}}} \mathbf{x} = \sum_{k \in \mathcal{I}_{\mathrm{hf}}} \underbrace{(1 - \sin^2(\pi k h/2)) \alpha_k}_{=:\beta_k} \mathbf{e}^k = \sum_{k \in \mathcal{I}_{\mathrm{hf}}} \beta_k \mathbf{e}^k$$

erhalten. Es gilt $\pi k h/2 \geq \frac{\pi}{2} \frac{n+1}{2} \frac{1}{n+1} = \pi/4$. Da die Sinusfunktion auf dem Intervall $[\pi/4, \pi/2]$ monoton von $\sin(\pi/4) = \sqrt{1/2}$ auf $\sin(\pi/2) = 1$ steigt, folgt

$$0 \leq 1 - \sin^2(\pi k h/2) \leq 1 - 1/2 = 1/2, \qquad |\beta_k| \leq |\alpha_k|/2 \qquad \text{für alle } k \in \mathcal{I}_{\mathrm{hf}}.$$

Indem wir (4.1) erst auf $\mathbf{M}_{\mathrm{Rich},\theta_{\mathrm{gl}}} \mathbf{x}$ und dann auf $\mathbf{x}$ anwenden, erhalten wir mit

$$\|\mathbf{M}_{\mathrm{Rich},\theta_{\mathrm{gl}}} \mathbf{x}\|_2^2 = \sum_{k \in \mathcal{I}_{\mathrm{hf}}} |\beta_k|^2 \leq \frac{1}{4} \sum_{k \in \mathcal{I}_{\mathrm{hf}}} |\alpha_k|^2 = \frac{1}{4} \|\mathbf{x}\|_2^2,$$

die gewünschte Abschätzung. ■

Zumindest auf dem von den hochfrequenten Eigenvektoren aufgespannten Teilraum $\mathcal{X}_{\mathrm{hf}}$ besitzt unser Richardson-Verfahren also tatsächlich eine von h, also n, unabhängige Konvergenzrate. Wenn wir somit die Iteration auf den Gesamtfehler anwenden, wird

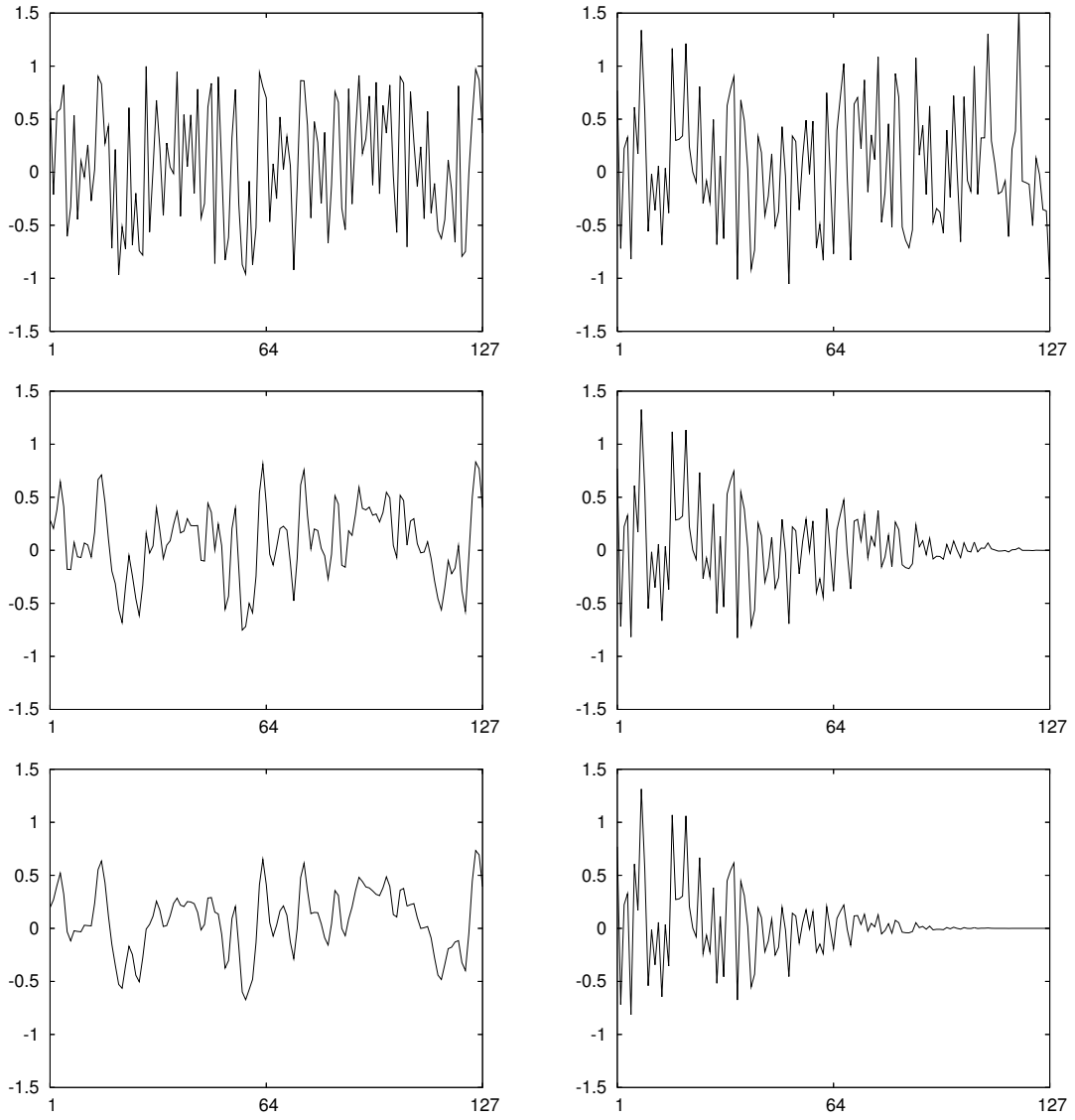

Abbildung 4.1: Fehlerreduktion durch die Richardson-Iteration. Links: Entwicklung des Startfehlers. Rechts: Entwicklung der Anteile der Eigenvektoren.

dessen hochfrequenter Anteil schnell reduziert werden, der Fehler wird mit jedem Iterationsschritt „glatter“ werden. Deshalb bezeichnet man das Richardson-Verfahren in diesem Fall als *Glättungsverfahren*. Zur Illustration ist in Abbildung 4.1 dargestellt, wie sich der Iterationsfehler bei wiederholter Anwendung des Richardson-Verfahrens verhält. In der linken Hälfte findet sich jeweils der Fehler (zur besseren Lesbarkeit sind die einzelnen Punktwerte durch eine Linie verbunden), in der rechten Hälfte die Anteile α_k der einzelnen Eigenvektoren $\mathbf{e}^k$. Man kann deutlich erkennen, dass die hochfrequenten

Anteile sehr schnell reduziert werden.

Wenn wir den Gesamtfehler reduzieren wollen, müssen wir auch eine Möglichkeit finden, die niedrigfrequenten Anteile in den Griff zu bekommen. Da die entsprechenden Funktionen „glatt“ sind, bietet es sich an, sie mit Hilfe eines gröberen Gitter zu approximieren.

Dazu führen wir eine Hierarchie von Indexmengen und Gleichungssystemen ein: Für jedes $\ell \in \mathbb{N}_0$ seien

- eine endliche Indexmenge $\mathcal{I}_\ell$,
- eine reguläre Matrix $\mathbf{A}_\ell \in \mathbb{K}^{\mathcal{I}_\ell \times \mathcal{I}_\ell}$ und
- eine rechte Seite $\mathbf{b}_\ell \in \mathbb{K}^{\mathcal{I}_\ell}$ gegeben.

Die Zahl ℓ bezeichnen wir als *Gitterstufe*, unter $\mathcal{I}_\ell$ stellt man sich häufig ein Punktegitter vor, das das Gebiet annähert, auf dem eine Differentialgleichung gelöst werden soll. Intuitiv sollen die Gleichungssysteme

$$\mathbf{A}_\ell \mathbf{x}_\ell = \mathbf{b}_\ell \qquad \text{für } \ell \in \mathbb{N}_0$$

zu verschieden feinen Auflösungen derselben zugrundeliegenden partiellen Differentialgleichung gehören. Höhere Gitterstufen sollen dabei höheren Genauigkeiten entsprechen, wir nehmen insbesondere an, dass die Indexmenge $\mathcal{I}_\ell$ größer als $\mathcal{I}_{\ell-1}$ ist.

Wenn wir das Gleichungssystem auf einer Stufe $\ell \in \mathbb{N}_0$ lösen wollen, können wir zunächst ausgehend von einem Startvektor $\mathbf{x}_\ell^{(0)}$ einige Schritte eines Glättungsverfahrens durchführen. Wir bezeichnen die daraus resultierende Näherungslösung mit $\mathbf{x}_\ell^{(1/2)}$ und die exakte Lösung des Gleichungssystems mit $\mathbf{x}_\ell^*$, es gilt also $\mathbf{A}_\ell \mathbf{x}_\ell^* = \mathbf{b}_\ell$. Falls das Glättungsverfahren erfolgreich ist, ist der verbliebene Fehler

$$\mathbf{e}_\ell^{(1/2)} := \mathbf{x}_\ell^{(1/2)} - \mathbf{x}_\ell^*$$

in einem geeigneten Sinn „glatt“, also besteht die Hoffnung, ihn auf einem gröberen Gitter noch hinreichend gut approximieren zu können. Wir suchen demnach einen Vektor $\mathbf{x}_{\ell-1} \in \mathbb{K}^{\mathcal{I}_{\ell-1}}$, der in einem geeigneten Sinne eine gute Approximation des Fehlers $\mathbf{e}_\ell^{(1/2)}$ beschreibt. Um eine Verbindung zwischen den beiden Gittern herstellen zu können, führen wir Transferoperatoren ein:

Definition 4.3 (Gittertransfer) *Sei $\ell \in \mathbb{N}$. Eine injektive Matrix*

$$\mathbf{p}_\ell \in \mathbb{K}^{\mathcal{I}_\ell \times \mathcal{I}_{\ell-1}}$$

bezeichnen wir als Prolongation. *Eine surjektive Matrix*

$$\mathbf{r}_\ell \in \mathbb{K}^{\mathcal{I}_{\ell-1} \times \mathcal{I}_\ell}$$

bezeichnen wir als Restriktion.

Mit Hilfe der Prolongation lässt sich nun formulieren, wie der Vektor $\mathbf{x}_{\ell-1} \in \mathbb{K}^{\mathcal{I}_{\ell-1}}$ beschaffen sein soll: $\mathbf{p}_\ell \mathbf{x}_{\ell-1}$ soll eine gute Approximation des Fehlers sein, es sollte also

$$\mathbf{p}_\ell \mathbf{x}_{\ell-1} \approx \mathbf{e}_\ell^{(1/2)}$$

gelten. Der Fehler $\mathbf{e}_\ell^{(1/2)}$ steht uns in der Regel nicht praktisch zur Verfügung, da wir die exakte Lösung $\mathbf{x}_\ell^*$ nicht kennen. Wir können allerdings mit $\mathbf{A}_\ell$ multiplizieren, um den *Defekt* zu erhalten:

$$\mathbf{A}_\ell \mathbf{p}_\ell \mathbf{x}_{\ell-1} \approx \mathbf{A}_\ell \mathbf{e}_\ell^{(1/2)} = \mathbf{A}_\ell(\mathbf{x}_\ell^{(1/2)} - \mathbf{x}_\ell^*) = \mathbf{A}_\ell \mathbf{x}_\ell^{(1/2)} - \mathbf{b}_\ell.$$

Diese „Gleichung“ können wir nicht einfach lösen, um $\mathbf{x}_{\ell-1}$ zu bestimmen, da $\mathbf{A}_\ell \mathbf{p}_\ell$ keine quadratische Matrix ist. Also multiplizieren wir mit der Restriktion $\mathbf{r}_\ell$ und gelangen zu

$$\mathbf{r}_\ell \mathbf{A}_\ell \mathbf{p}_\ell \mathbf{x}_{\ell-1} = \mathbf{r}_\ell(\mathbf{A}_\ell \mathbf{x}_\ell^{(1/2)} - \mathbf{b}_\ell).$$

Da $\mathbf{A}_\ell$ und $\mathbf{A}_{\ell-1}$ denselben Differentialoperator approximieren sollen, dürfen wir

$$\mathbf{A}_{\ell-1} \approx \mathbf{r}_\ell \mathbf{A}_\ell \mathbf{p}_\ell \tag{4.2}$$

erwarten, denn es sollte für „glatte“ Funktionen keinen großen Unterschied machen, ob wir den Operator auf dem feinen oder dem groben Gitter anwenden.

Also bestimmen wir $\mathbf{x}_{\ell-1}$ als Lösung des linearen Gleichungssystems

$$\mathbf{A}_{\ell-1} \mathbf{x}_{\ell-1} = \mathbf{r}_\ell(\mathbf{A}_\ell \mathbf{x}_\ell^{(1/2)} - \mathbf{b}_\ell), \tag{4.3}$$

so dass $\mathbf{p}_\ell \mathbf{x}_{\ell-1} \approx \mathbf{x}_\ell^{(1/2)} - \mathbf{x}_\ell^*$ gilt, und führen anschließend die Korrektur

$$\mathbf{x}_\ell^{(1)} := \mathbf{x}_\ell^{(1/2)} - \mathbf{p}_\ell \mathbf{x}_{\ell-1}$$

durch. Dieser Korrekturschritt lässt sich als Schritt eines linearen Iterationsverfahrens interpretieren.

Definition 4.4 (Grobgitterkorrektur) *Sei $\ell \in \mathbb{N}$. Das durch*

$$\Phi_{\mathrm{GGK},\ell}(\mathbf{x}_\ell, \mathbf{b}_\ell) = \mathbf{x}_\ell - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1} \mathbf{r}_\ell(\mathbf{A}_\ell \mathbf{x}_\ell - \mathbf{b}_\ell) \qquad \textit{für alle } \mathbf{x}_\ell, \mathbf{b}_\ell \in \mathbb{K}^{\mathcal{I}_\ell}$$

definierte lineare Iterationsverfahren nennen wir die Grobgitterkorrektur *auf der Gitterstufe ℓ.*

Offenbar ist $\Phi_{\mathrm{GGK},\ell}$ ein konsistentes lineares Iterationsverfahren mit den Matrizen

$$\mathbf{M}_{\mathrm{GGK},\ell} := \mathbf{I} - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1} \mathbf{r}_\ell \mathbf{A}_\ell, \qquad \mathbf{N}_{\mathrm{GGK},\ell} := \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1} \mathbf{r}_\ell.$$

Die Grobgitterkorrektur ist nur dann effizient zu berechnen, wenn sich das Gleichungssystem (4.3) effizient lösen lässt, wenn also $\mathcal{I}_{\ell-1}$ deutlich kleiner als $\mathcal{I}_\ell$ ist. In diesem Fall ist die Grobgitterkorrektur allerdings nicht mehr konvergent: Da $|\mathcal{I}_{\ell-1}| < |\mathcal{I}_\ell|$ vorausgesetzt

procedure Zweigitter(ℓ, $\mathbf{b}_\ell$, **var** $\mathbf{x}_\ell$);
Glätter($\mathbf{x}_\ell$, $\mathbf{b}_\ell$);
$\mathbf{d}_\ell \leftarrow \mathbf{A}_\ell \mathbf{x}_\ell - \mathbf{b}_\ell$;
$\mathbf{b}_{\ell-1} \leftarrow \mathbf{r}_\ell \mathbf{d}_\ell$;
Löse $\mathbf{A}_{\ell-1}\mathbf{x}_{\ell-1} = \mathbf{b}_{\ell-1}$;
$\mathbf{x}_\ell \leftarrow \mathbf{x}_\ell - \mathbf{p}_\ell \mathbf{x}_{\ell-1}$

Abbildung 4.2: Ein Schritt des Zweigitterverfahrens

ist, kann die Restriktion $\mathbf{r}_\ell$ nicht injektiv sein, es muss also einen Vektor $\mathbf{z}_\ell \in \mathbb{K}^{\mathcal{I}_\ell} \setminus \{\mathbf{0}\}$ geben, der $\mathbf{r}_\ell \mathbf{z}_\ell = \mathbf{0}$ erfüllt. Wir setzen $\mathbf{x}_\ell := \mathbf{A}_\ell^{-1}(\mathbf{z}_\ell + \mathbf{b}_\ell)$ und erhalten

$$\Phi_{\mathrm{GGK},\ell}(\mathbf{x}_\ell, \mathbf{b}_\ell) = \mathbf{x}_\ell - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1} \mathbf{r}_\ell (\mathbf{A}_\ell \mathbf{x}_\ell - \mathbf{b}_\ell) = \mathbf{x}_\ell - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1} \mathbf{r}_\ell \mathbf{z}_\ell = \mathbf{x}_\ell,$$

also wird die Folge der Iterierten für den Startvektor $\mathbf{x}_\ell$ nicht gegen die korrekte Lösung $\mathbf{A}_\ell^{-1}\mathbf{b}_\ell$ konvergieren.

Für sich genommen ist die Grobgitterkorrektur also kein sinnvolles Verfahren, lediglich im Zusammenspiel mit einem geeigneten Glättungsverfahren, das die, typischerweise hochfrequenten, Fehleranteile aus dem Kern der Restriktion $\mathbf{r}_\ell$ handhabt, erhalten wir eine effiziente Methode.

Definition 4.5 (Zweigitterverfahren) *Sei $\ell \in \mathbb{N}$. Sei $\Phi_{\mathrm{Gl},\ell}$ ein lineares konsistentes Iterationsverfahren für die Matrix $\mathbf{A}_\ell$. Das durch*

$$\Phi_{\mathrm{ZGV},\ell}(\mathbf{x}_\ell, \mathbf{b}_\ell) = \Phi_{\mathrm{GGK},\ell}(\Phi_{\mathrm{Gl},\ell}(\mathbf{x}_\ell, \mathbf{b}_\ell), \mathbf{b}_\ell) \qquad \textit{für alle } \mathbf{x}_\ell, \mathbf{b}_\ell \in \mathbb{K}^{\mathcal{I}_\ell}$$

definierte lineare Iterationsverfahren nennen wir das Zweigitterverfahren *auf der Gitterstufe ℓ mit dem* Glättungsverfahren *(oder kurz* Glätter*) $\Phi_{\mathrm{Gl},\ell}$.*

Abstrakt formuliert ist das Zweigitterverfahren eine *Produktiteration*, die sich aus der Kombination der Grobgitterkorrektur $\Phi_{\mathrm{GGK},\ell}$ mit der Glättungsiteration $\Phi_{\mathrm{Gl},\ell}$ ergibt.

Wenn wir die Matrizen der ersten und zweiten Normalform der Glättungsiteration mit $\mathbf{M}_{\mathrm{Gl},\ell}$ und $\mathbf{N}_{\mathrm{Gl},\ell}$ bezeichnen, erhalten wir für das Zweigitterverfahren die Darstellung

$$\begin{aligned}
\Phi_{\mathrm{ZGV},\ell}(\mathbf{x}_\ell, \mathbf{b}_\ell) &= \Phi_{\mathrm{GGK},\ell}(\Phi_{\mathrm{Gl},\ell}(\mathbf{x}_\ell, \mathbf{b}_\ell), \mathbf{b}_\ell) = \mathbf{M}_{\mathrm{GGK},\ell}\Phi_{\mathrm{Gl},\ell}(\mathbf{x}_\ell, \mathbf{b}_\ell) + \mathbf{N}_{\mathrm{GGK},\ell}\mathbf{b}_\ell \\
&= \mathbf{M}_{\mathrm{GGK},\ell}(\mathbf{M}_{\mathrm{Gl},\ell}\mathbf{x}_\ell + \mathbf{N}_{\mathrm{Gl},\ell}\mathbf{b}_\ell) + \mathbf{N}_{\mathrm{GGK},\ell}\mathbf{b}_\ell \\
&= \mathbf{M}_{\mathrm{GGK},\ell}\mathbf{M}_{\mathrm{Gl},\ell}\mathbf{x}_\ell + \mathbf{M}_{\mathrm{GGK},\ell}\mathbf{N}_{\mathrm{Gl},\ell}\mathbf{b}_\ell + \mathbf{N}_{\mathrm{GGK},\ell}\mathbf{b}_\ell,
\end{aligned}$$

also sind die Matrizen der ersten und zweiten Normalform des Zweigitterverfahrens durch

$$\begin{aligned}
\mathbf{M}_{\mathrm{ZGV},\ell} &:= \mathbf{M}_{\mathrm{GGK},\ell}\mathbf{M}_{\mathrm{Gl},\ell}, \\
\mathbf{N}_{\mathrm{ZGV},\ell} &:= \mathbf{M}_{\mathrm{GGK},\ell}\mathbf{N}_{\mathrm{Gl},\ell} + \mathbf{N}_{\mathrm{GGK},\ell}
\end{aligned}$$

gegeben. Die Entwicklung des Iterationsfehlers im Zweigitterverfahren wird durch die Iterationsmatrix $\mathbf{M}_{\mathrm{ZGV},\ell}$ beschrieben, und diese Matrix ist das Produkt der Iterationsmatrizen der Grobgitterkorrektur und des Glättungsverfahrens. Falls also die Grobgitterkorrektur die niedrigfrequenten und das Glättungsverfahren die hochfrequenten Fehleranteile reduziert, dürfen wir darauf hoffen, dass das Zweigitterverfahren den gesamten Fehler reduzieren wird.

Bevor wir uns eingehender mit den Eigenschaften des Zweigitterverfahrens befassen, ist es ratsam, dass wir uns zunächst mit den neuen Begriffen der Prolongation und der Restriktion vertraut machen. Am besten gelingt das anhand des Modellproblems: Wir betrachten zunächst das eindimensionale Modellproblem mit $N_\ell := 2^{\ell+1} - 1$. Für jedes $j \in \mathcal{I}_{\ell-1} = [1 : N_{\ell-1}]$ gilt $\xi_{\ell-1,j} = \xi_{\ell,2j}$, jeder Gitterpunkt des groben Gitters ist also auch im feinen Gitter enthalten. Deshalb ist es naheliegend,

$$(\mathbf{p}_\ell \mathbf{x}_{\ell-1})_{2j} = x_{\ell-1,j} \qquad \text{für alle } j \in \mathcal{I}_{\ell-1}$$

zu setzen. Damit ist bereits fast die Hälfte der Komponenten von $\mathbf{p}_\ell \mathbf{x}_{\ell-1}$ definiert. Die ungeradzahligen Punkte $\xi_{\ell,2j+1}$ liegen für $j \in [0 : N_{\ell-1}]$ jeweils im Mittelpunkt der Verbindungsstrecke von $\xi_{\ell-1,j}$ und $\xi_{\ell-1,j+1}$, also bietet es sich an, den Wert $(\mathbf{p}_\ell \mathbf{x}_{\ell-1})_{2j+1}$ durch Mittelung der Werte $x_{\ell-1,j}$ und $x_{\ell-1,j+1}$ zu gewinnen, also

$$(\mathbf{p}_\ell \mathbf{x}_{\ell-1})_{2j+1} = \frac{x_{\ell-1,j} + x_{\ell-1,j+1}}{2} \qquad \text{für alle } j \in [0 : N_{\ell-1}]$$

zu setzen. Damit ist die Prolongation $\mathbf{p}_\ell$ durch

$$(\mathbf{p}_\ell)_{ij} = \begin{cases} 1 & \text{falls } i = 2j, \\ 1/2 & \text{falls } |i - 2j| = 1, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i \in \mathcal{I}_\ell, j \in \mathcal{I}_{\ell-1}$$

gegeben. Wir können leicht sehen, dass die Prolongation $\mathbf{p}_\ell$ eine schwachbesetzte Matrix ist, und dass sich die Multiplikation eines Vektors mit dieser Matrix sehr einfach durchführen lässt.

Die Restriktion wird in der Regel mit Hilfe der Adjungierten der Prolongation konstruiert: Die Prolongation verteilt den Vektor aus den Grobgitterknoten in die Feingitterknoten, die Restriktion sammelt die Werte aus den Feingitterknoten wieder in den Grobgitterknoten.

In unserem Fall empfiehlt es sich (siehe Definition 4.6), die Restriktion so zu gewichten, dass die Restriktion des konstanten Vektors $\mathbf{1}$ auf dem feinen Gitter wieder den konstanten Vektor $\mathbf{1}$ auf dem groben Gitter ergibt, also

$$\mathbf{r}_\ell := \frac{1}{2}\mathbf{p}_\ell^*$$

zu verwenden. Die Restriktionsmatrix ist dann durch

$$(\mathbf{r}_\ell)_{ij} = \begin{cases} 1/2 & \text{falls } j = 2i, \\ 1/4 & \text{falls } |j - 2i| = 1, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i \in \mathcal{I}_{\ell-1}, j \in \mathcal{I}_\ell$$

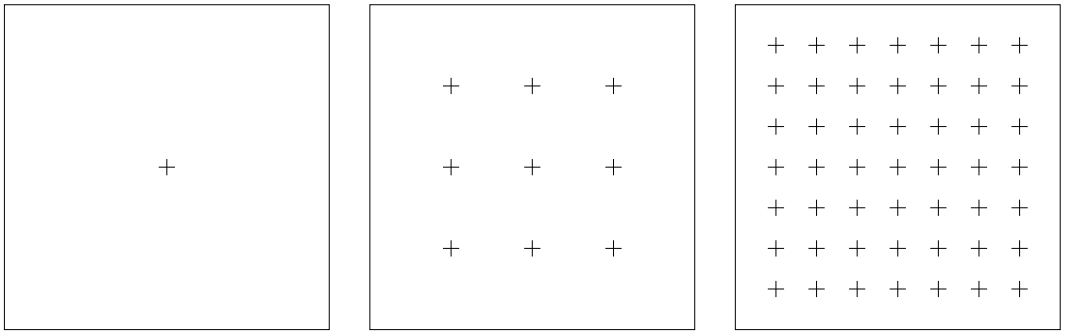

Abbildung 4.3: Gitterhierarchie für das zweidimensionale Modellproblem

gegeben. Entsprechend ihrer Konstruktion ist auch diese Matrix schwachbesetzt und lässt sich einfach mit einem Vektor multiplizieren.

Anschaulich entspricht die Restriktion einer gewichteten Mittelwertbildung, bei der der zentrale Knoten doppelt so stark wie seine beiden Nachbarn gewichtet wird. Da alle Koeffizienten der Restriktionsmatrix nicht negativ sind, kann ein Vektor nur dann in ihrem Kern liegen, wenn die Vorzeichen seiner Komponenten hinreichend schnell wechseln, er also nicht „glatt“ ist.

Wenden wir uns nun dem zweidimensionalen Modellproblem zu. Auch hier setzen wir $N_\ell := 2^{\ell+1} - 1$ und definieren darauf basierend die Indexmenge

$$\mathcal{I}_\ell := \{(i_x, i_y) \ : \ i_x, i_y \in [1 : N_\ell]\}.$$

Wie schon im eindimensionalen Fall gilt für jedes $j = (j_x, j_y) \in \mathcal{I}_{\ell-1}$ die Gleichung $\xi_{\ell-1,j} = \xi_{\ell,2j}$, jetzt allerdings für Multiindizes, also können wir

$$(\mathbf{p}_\ell \mathbf{x}_{\ell-1})_{2j} = x_{\ell-1,j} \qquad \text{für alle } j \in \mathcal{I}_{\ell-1}$$

setzen. Falls der Gitterpunkt ξ_i zu einem Index $i \in \mathcal{I}_\ell$ zwischen zwei Gitterpunkten des groben Gitters liegt, können wir wie im eindimensionalen Fall linear interpolieren und erhalten

$$\begin{aligned}
(\mathbf{p}_\ell \mathbf{x}_{\ell-1})_{2j_x+1,2j_y} &= \frac{x_{\ell-1,(j_x,j_y)} + x_{\ell-1,(j_x+1,j_y)}}{2} \\
&\qquad \text{für alle } j_x \in [0 : N_{\ell-1}],\ j_y \in [1 : N_{\ell-1}], \\
(\mathbf{p}_\ell \mathbf{x}_{\ell-1})_{2j_x,2j_y+1} &= \frac{x_{\ell-1,(j_x,j_y)} + x_{\ell-1,(j_x,j_y+1)}}{2} \\
&\qquad \text{für alle } j_x \in [1 : N_{\ell-1}],\ j_y \in [0 : N_{\ell-1}].
\end{aligned}$$

Falls der Gitterpunkt zwischen vier Punkten des groben Gitters liegt, würde man erwarten, dass wir bilinear zwischen allen vier Punkten interpolieren. Es gibt allerdings einen einfacheren Weg, der auch für die spätere theoretische Untersuchung Vorteile bietet: Wir

```
procedure Prolongation2D(ℓ, x, var y);
N_{ℓ-1} ← 2^ℓ − 1;
y ← 0;
for i_y := 1 to N_{ℓ-1} do
  for i_x := 1 to N_{ℓ-1} do
    y_{2i_x,2i_y} ← y_{2i_x,2i_y} + x_{i_x,i_y};
    y_{2i_x+1,2i_y} ← y_{2i_x+1,2i_y} + x_{i_x,i_y}/2;
    y_{2i_x-1,2i_y} ← y_{2i_x-1,2i_y} + x_{i_x,i_y}/2;
    y_{2i_x,2i_y+1} ← y_{2i_x,2i_y+1} + x_{i_x,i_y}/2;
    y_{2i_x,2i_y-1} ← y_{2i_x,2i_y-1} + x_{i_x,i_y}/2;
    y_{2i_x+1,2i_y+1} ← y_{2i_x+1,2i_y+1} + x_{i_x,i_y}/2;
    y_{2i_x-1,2i_y-1} ← y_{2i_x-1,2i_y-1} + x_{i_x,i_y}/2;
  end for
end for
```

Abbildung 4.4: Durchführung der Prolongation für das zweidimensionale Modellproblem

interpolieren entlang der Diagonalen von „links unten nach rechts oben“. Auf diese Weise erhalten wir

$$(\mathbf{p}_\ell \mathbf{x}_{\ell-1})_{2j_x+1,2j_y+1} = \frac{x_{\ell-1,(j_x,j_y)} + x_{\ell-1,(j_x+1,j_y+1)}}{2} \qquad \text{für alle } j_x, j_y \in [0 : N_{\ell-1}].$$

Damit ist die Prolongation für das zweidimensionale Modellproblem durch

$$(\mathbf{p}_\ell)_{ij} = \begin{cases} 1 & \text{falls } i_x = 2j_x, i_y = 2j_y, \\ 1/2 & \text{falls } |i_x - 2j_x| = 1, i_y = 2j_y, \\ 1/2 & \text{falls } i_x = 2j_x, |i_y - 2j_y| = 1, \\ 1/2 & \text{falls } i_x - 2j_x = i_y - 2j_y \in \{-1, 1\}, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i \in \mathcal{I}_\ell, j \in \mathcal{I}_{\ell-1}$$

definiert. Offenbar ist auch diese Prolongationsmatrix schwachbesetzt und ihre Anwendung auf einen Vektor algorithmisch leicht umzusetzen (vgl. Abbildung 4.4). Wir können sehen, dass pro Feingitterpunkt höchstens zwei Additionen und Multiplikationen durchgeführt werden müssen, die Prolongation hat also einen geringeren Aufwand als die Multiplikation mit der Matrix $\mathbf{A}_\ell$. Abhängig von der zugrundeliegenden Rechnerarchitektur kann es empfehlenswert sein, die Berechnung der Prolongation nicht durch aufeinanderfolgende Additionen zu dem Ergebnisvektor $\mathbf{y}$ durchzuführen, sondern ähnlich wie bei der Multiplikation mit $\mathbf{A}_\ell$ (vgl. Abbildung 1.6) direkt einzelne Einträge zu berechnen und dabei geeignete Fallunterscheidungen zu verwenden. Dadurch wird der Algorithmus allerdings im Allgemeinen nicht unbedingt lesbarer, lediglich schneller.

Die entsprechende Restriktion definieren wir wieder, indem wir die Adjungierte der Prolongation so skalieren, dass die Restriktion des konstanten Vektors wieder dieselbe

```
procedure Restriktion2D(ℓ, x, var y);
N_{ℓ-1} ← 2^ℓ − 1;
y ← 0;
for i_y := 1 to N_{ℓ-1} do
  for i_x := 1 to N_{ℓ-1} do
    y_{i_x,i_y} ← y_{i_x,i_y} + x_{2i_x,2i_y}/4;
    y_{i_x,i_y} ← y_{i_x,i_y} + x_{2i_x+1,2i_y}/8;
    y_{i_x,i_y} ← y_{i_x,i_y} + x_{2i_x-1,2i_y}/8;
    y_{i_x,i_y} ← y_{i_x,i_y} + x_{2i_x,2i_y+1}/8;
    y_{i_x,i_y} ← y_{i_x,i_y} + x_{2i_x,2i_y-1}/8;
    y_{i_x,i_y} ← y_{i_x,i_y} + x_{2i_x+1,2i_y+1}/8;
    y_{i_x,i_y} ← y_{i_x,i_y} + x_{2i_x-1,2i_y-1}/8
  end for
end for
```

Abbildung 4.5: Durchführung der Restriktion für das zweidimensionale Modellproblem

Konstante ergibt, in diesem Fall also durch

$$\mathbf{r}_\ell := \frac{1}{4}\mathbf{p}_\ell^*.$$

Die Restriktionsmatrix ist dann durch

$$(\mathbf{r}_\ell)_{ij} = \begin{cases} 1/4 & \text{falls } j_x = 2i_x, j_y = 2i_y, \\ 1/8 & \text{falls } |j_x - 2i_x| = 1, j_y = 2i_y, \\ 1/8 & \text{falls } j_x = 2i_x, |j_y - 2i_y| = 1, \\ 1/8 & \text{falls } j_x - 2i_x = j_y - 2i_y \in \{-1, 1\}, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i \in \mathcal{I}_{\ell-1}, j \in \mathcal{I}_\ell$$

gegeben und wiederum schwachbesetzt. Sie lässt sich in sehr ähnlicher Weise implementieren (vgl. Abbildung 4.5) und besitzt denselben Rechenaufwand wie die Prolongation. Mit den so definierten Prolongationen und Restriktionen können wir das Zweigitterverfahren (und auch das später definierte Mehrgitterverfahren) für unsere Modellprobleme durchführen.

Eine erste einfache Abschätzung für die Konvergenzgeschwindigkeit des Zweigitterverfahrens lässt sich gewinnen, falls die Matrizen auf den verschiedenen Gitterstufen und die Prolongationen und Restriktionen im Sinne der Gleichung (4.2) zueinander passen:

Definition 4.6 (Galerkin-Eigenschaft) *Falls für alle $\ell \in \mathbb{N}$ die Gleichung*

$$\mathbf{A}_{\ell-1} = \mathbf{r}_\ell \mathbf{A}_\ell \mathbf{p}_\ell$$

gilt, sagen wir, dass die Hierarchie der linearen Gleichungssysteme die Galerkin-Eigenschaft *besitzt.*

Lemma 4.7 *Die Hierarchie der Gleichungssysteme besitze die Galerkin-Eigenschaft. Sei $\ell \in \mathbb{N}$. Dann ist die Iterationsmatrix $\mathbf{M}_{\mathrm{GGK},\ell}$ der Grobgitterkorrektur eine Projektion, es gilt also*

$$\mathbf{M}_{\mathrm{GGK},\ell}^2 = \mathbf{M}_{\mathrm{GGK},\ell}.$$

Falls $\mathbf{A}_\ell$ positiv definit ist und $\mathbf{r}_\ell = c\mathbf{p}_\ell^*$ für ein $c \in \mathbb{R}_{>0}$ gilt, ist diese Matrix bezüglich des Energieskalarprodukts selbstadjungiert, es gilt also

$$\langle \mathbf{M}_{\mathrm{GGK},\ell}\mathbf{x}_\ell, \mathbf{y}_\ell\rangle_A = \langle \mathbf{x}_\ell, \mathbf{M}_{\mathrm{GGK},\ell}\mathbf{y}_\ell\rangle_A \qquad \text{für alle } \mathbf{x}_\ell, \mathbf{y}_\ell \in \mathbb{K}^{\mathcal{I}_\ell}.$$

Daraus folgt, dass die Matrix bezüglich des Energieskalarprodukts eine orthogonale Projektion ist und $\|\mathbf{M}_{\mathrm{GGK},\ell}\|_A \leq 1$ erfüllt.

Beweis. Die erste Gleichung können wir direkt nachrechnen:

$$\begin{aligned}
\mathbf{M}_{\mathrm{GGK},\ell}^2 &= (\mathbf{I} - \mathbf{p}_\ell\mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\mathbf{A}_\ell)(\mathbf{I} - \mathbf{p}_\ell\mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\mathbf{A}_\ell)\\
&= \mathbf{I} - 2\mathbf{p}_\ell\mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\mathbf{A}_\ell + \mathbf{p}_\ell\mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\mathbf{A}_\ell\mathbf{p}_\ell\mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\mathbf{A}_\ell\\
&= \mathbf{I} - 2\mathbf{p}_\ell\mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\mathbf{A}_\ell + \mathbf{p}_\ell\mathbf{A}_{\ell-1}^{-1}\mathbf{A}_{\ell-1}\mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\mathbf{A}_\ell = \mathbf{I} - \mathbf{p}_\ell\mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\mathbf{A}_\ell = \mathbf{M}_{\mathrm{GGK},\ell}.
\end{aligned}$$

Sei nun $\mathbf{A}_\ell$ positiv definit, und sei $c \in \mathbb{R}_{>0}$ so gegeben, dass $\mathbf{r}_\ell = c\mathbf{p}_\ell^*$ erfüllt ist. Wir fixieren $\mathbf{x}_\ell, \mathbf{y}_\ell \in \mathbb{K}^{\mathcal{I}_\ell}$. Da $\mathbf{A}_\ell$ positiv definit und $\mathbf{p}_\ell$ injektiv ist, ist auch

$$\mathbf{A}_{\ell-1} = \mathbf{r}_\ell\mathbf{A}_\ell\mathbf{p}_\ell = c\mathbf{p}_\ell^*\mathbf{A}_\ell\mathbf{p}_\ell$$

positiv definit, und es gilt

$$\begin{aligned}
\langle \mathbf{M}_{\mathrm{GGK},\ell}\mathbf{x}_\ell, \mathbf{y}_\ell\rangle_A &= \langle \mathbf{A}_\ell(\mathbf{I} - \mathbf{p}_\ell\mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\mathbf{A}_\ell)\mathbf{x}_\ell, \mathbf{y}_\ell\rangle_2 = \langle (\mathbf{A}_\ell - c\mathbf{A}_\ell\mathbf{p}_\ell\mathbf{A}_{\ell-1}^{-1}\mathbf{p}_\ell^*\mathbf{A}_\ell)\mathbf{x}_\ell, \mathbf{y}_\ell\rangle_2\\
&= \langle (\mathbf{I} - c\mathbf{A}_\ell\mathbf{p}_\ell\mathbf{A}_{\ell-1}^{-1}\mathbf{p}_\ell^*)\mathbf{A}_\ell\mathbf{x}_\ell, \mathbf{y}_\ell\rangle_2 = \langle \mathbf{A}_\ell\mathbf{x}_\ell, (\mathbf{I} - c\mathbf{p}_\ell\mathbf{A}_{\ell-1}^{-1}\mathbf{p}_\ell^*\mathbf{A}_\ell)\mathbf{y}_\ell\rangle_2\\
&= \langle \mathbf{A}_\ell\mathbf{x}_\ell, (\mathbf{I} - \mathbf{p}_\ell\mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\mathbf{A}_\ell)\mathbf{y}_\ell\rangle_2 = \langle \mathbf{x}_\ell, \mathbf{M}_{\mathrm{GGK},\ell}\mathbf{y}_\ell\rangle_A.
\end{aligned}$$

Mit Hilfe dieser Gleichung und der Cauchy-Schwarz-Ungleichung erhalten wir schließlich

$$\begin{aligned}
\|\mathbf{M}_{\mathrm{GGK},\ell}\mathbf{x}_\ell\|_A^2 &= \langle \mathbf{M}_{\mathrm{GGK},\ell}\mathbf{x}_\ell, \mathbf{M}_{\mathrm{GGK},\ell}\mathbf{x}_\ell\rangle_A = \langle \mathbf{M}_{\mathrm{GGK},\ell}^2\mathbf{x}_\ell, \mathbf{x}_\ell\rangle_A\\
&= \langle \mathbf{M}_{\mathrm{GGK},\ell}\mathbf{x}_\ell, \mathbf{x}_\ell\rangle_A \leq \|\mathbf{M}_{\mathrm{GGK},\ell}\mathbf{x}_\ell\|_A\|\mathbf{x}_\ell\|_A,
\end{aligned}$$

und daraus folgt die gewünschte Abschätzung. ∎

Unter den Voraussetzungen dieses Lemmas (die für viele praktische Anwendungsfälle erfüllt sind) können wir also die Konvergenzgeschwindigkeit des Zweigitterverfahrens einfach durch

$$\|\mathbf{M}_{\mathrm{ZGV},\ell}\|_A = \|\mathbf{M}_{\mathrm{GGK},\ell}\mathbf{M}_{\mathrm{Gl},\ell}\|_A \leq \|\mathbf{M}_{\mathrm{GGK},\ell}\|_A\|\mathbf{M}_{\mathrm{Gl},\ell}\|_A \leq \|\mathbf{M}_{\mathrm{Gl},\ell}\|_A$$

abschätzen, also wird bei den meisten von uns bisher untersuchten Verfahren das Zweigitterverfahren mindestens so schnell wie das Glättungsverfahren konvergieren. Natürlich

sind wir daran interessiert, eine bessere Abschätzung für die Konvergenzrate zu erzielen, aber dafür werden aufwendigere Techniken erforderlich sein, die später eingeführt werden.

Da wir die Prolongation und Restriktion passend gewählt haben, besitzen die oben beschriebenen Hierarchien für das ein- und zweidimensionale Modellproblem die Galerkin-Eigenschaft. Für das eindimensionale Problem ist der Nachweis besonders einfach: Für jedes $i \in \mathcal{I}_\ell$ definieren wir den kanonischen Einheitsvektor $\delta_{\ell,i} \in \mathbb{K}^{\mathcal{I}_\ell}$ durch

$$(\delta_{\ell,i})_j = \begin{cases} 1 & \text{falls } i = j, \\ 0 & \text{sonst} \end{cases} \qquad \text{für alle } j \in \mathcal{I}_\ell.$$

Damit erhalten wir nach Definition von $\mathbf{p}_\ell$, $\mathbf{A}_\ell$, $\mathbf{A}_{\ell-1}$ und $\mathbf{r}_\ell$ die Gleichungskette

$$\begin{aligned}
\mathbf{p}_\ell \delta_{\ell-1,i} &= \frac{1}{2}\delta_{\ell,2i-1} + \delta_{\ell,2i} + \frac{1}{2}\delta_{\ell,2i+1}, \\
\mathbf{A}_\ell \mathbf{p}_\ell \delta_{\ell-1,i} &= h_\ell^{-2}\delta_{\ell-1,2i} - \frac{h_\ell^{-2}}{2}\begin{cases} \delta_{\ell,2i-2} & \text{falls } i > 1, \\ \mathbf{0} & \text{ansonsten} \end{cases} \\
&\quad - \frac{h_\ell^{-2}}{2}\begin{cases} \delta_{\ell,2i+2} & \text{falls } i < N_{\ell-1}, \\ \mathbf{0} & \text{ansonsten} \end{cases}, \\
\mathbf{r}_\ell \mathbf{A}_\ell \mathbf{p}_\ell \delta_{\ell-1,i} &= \frac{h_\ell^{-2}}{2}\delta_{\ell-1,i} - \frac{h_\ell^{-2}}{4}\begin{cases} \delta_{\ell-1,i-1} & \text{falls } i > 1, \\ \mathbf{0} & \text{ansonsten} \end{cases} \\
&\quad - \frac{h_\ell^{-2}}{4}\begin{cases} \delta_{\ell-1,i+1} & \text{falls } i < N_{\ell-1}, \\ \mathbf{0} & \text{ansonsten} \end{cases}, \\
&= 2h_{\ell-1}^{-2}\delta_{\ell-1,i} - h_{\ell-1}^{-2}\begin{cases} \delta_{\ell,i-1} & \text{falls } i > 1, \\ \mathbf{0} & \text{ansonsten} \end{cases} \\
&\quad - h_{\ell-1}^{-2}\begin{cases} \delta_{\ell,i+1} & \text{falls } i < N_{\ell-1}, \\ \mathbf{0} & \text{ansonsten} \end{cases}, \\
&= \mathbf{A}_{\ell-1}\delta_{\ell-1,i},
\end{aligned}$$

womit die gewünschte Gleichung bewiesen ist. Für das zweidimensionale Modellproblem lässt sich der Beweis in ähnlicher Weise führen.

4.2 Mehrgitterverfahren

Der aufwendige Teil des Zweigitterverfahrens ist die Berechnung der Grobgitterkorrektur, denn in der Regel wird die Indexmenge $\mathcal{I}_{\ell-1}$ noch so groß sein, dass die Lösung des Grobgitter-Gleichungssystems $\mathbf{A}_{\ell-1}\mathbf{x}_{\ell-1} = \mathbf{b}_{\ell-1}$ sehr viele Rechenoperationen erfordert.

Glücklicherweise müssen wir dieses System nicht exakt lösen: Das Glättungsverfahren reduziert den Fehler schließlich auch nur um einen gewissen Faktor, nicht in einem Schritt auf Null, also sollte es genügen, auch auf dem groben Gitter nur eine hinreichend gute Approximation der Lösung zu berechnen.

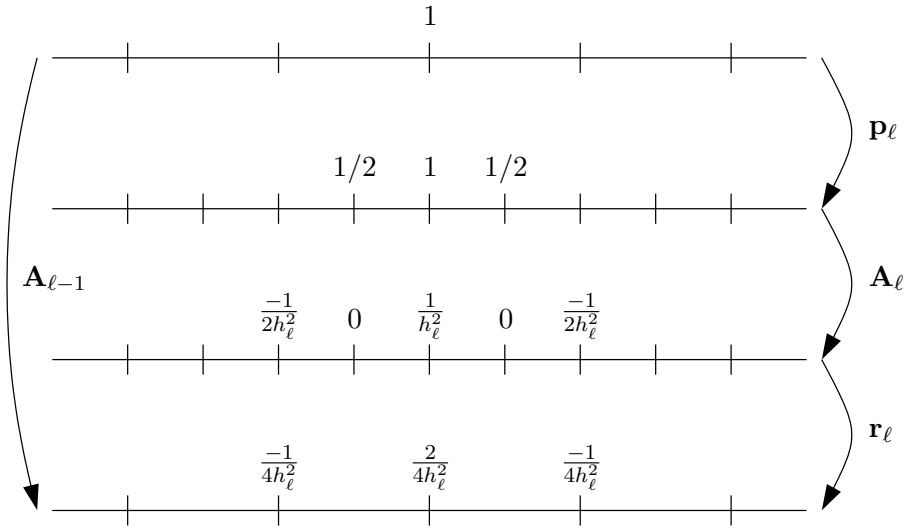

Abbildung 4.6: Nachweis der Galerkin-Eigenschaft $\mathbf{A}_{\ell-1} = \mathbf{r}_\ell \mathbf{A}_\ell \mathbf{p}_\ell$ für das eindimensionale Modellproblem: Wenn wir einen kanonischer Einheitsvektor (erste Zeile) der Reihe nach mit der Prolongationsmatrix $\mathbf{p}_\ell$, der Feingittermatrix $\mathbf{A}_\ell$ und der Restriktionsmatrix $\mathbf{r}_\ell$ multiplizieren, erhalten wir dasselbe wie bei einer Multiplikation mit der Grobgittermatrix $\mathbf{A}_{\ell-1}$.

procedure Mehrgitter1(ℓ, $\mathbf{b}_\ell$, **var** $\mathbf{x}_\ell$);
if $\ell > 0$ **then**
  Glätter($\mathbf{x}_\ell$, $\mathbf{b}_\ell$);
  $\mathbf{d}_\ell \leftarrow \mathbf{A}_\ell \mathbf{x}_\ell - \mathbf{b}_\ell$;
  $\mathbf{b}_{\ell-1} \leftarrow \mathbf{r}_\ell \mathbf{d}_\ell$;
  $\mathbf{x}_{\ell-1} \leftarrow \mathbf{0}$;
  Mehrgitter1($\ell - 1$, $\mathbf{b}_{\ell-1}$, $\mathbf{x}_{\ell-1}$);
  $\mathbf{x}_\ell \leftarrow \mathbf{x}_\ell - \mathbf{p}_\ell \mathbf{x}_{\ell-1}$
else
  $\mathbf{x}_\ell \leftarrow \mathbf{A}_\ell^{-1} \mathbf{b}_\ell$
end if

Abbildung 4.7: Erste Fassung des Mehrgitterverfahrens

Dazu ließe sich im Prinzip jedes der bisher eingeführten Verfahren verwenden, aber im Interesse der Effizienz bietet es sich an, wiederum auf ein Zweigitterverfahren auf den Gitterstufen $\ell - 1$ und $\ell - 2$ zurückzugreifen. Wenn wir in dieser Weise rekursiv fortfahren, bis die gröbste Gitterstufe 0 erreicht ist, erhalten wir eine erste Variante des Mehrgitterverfahrens, die in Abbildung 4.7 gegeben ist.

In der Praxis kann es sehr sinnvoll sein, nicht nur einen einzelnen Schritt des Glättungsverfahrens durchzuführen, und aus Gründen der Symmetrie ist es empfehlenswert, auch nach der Grobgitterkorrektur noch weitere Glättungsschritte vorzusehen. Deshalb führen

```
procedure Mehrgitter(ℓ, b_ℓ, var x_ℓ);
if ℓ > 0 then
  for i := 1 to ν_1 do
    Glätter(x_ℓ, b_ℓ)
  end for;
  d_ℓ ← A_ℓ x_ℓ − b_ℓ;
  b_{ℓ−1} ← r_ℓ d_ℓ;
  x_{ℓ−1} ← 0;
  for i := 1 to γ do
    Mehrgitter(ℓ − 1, b_{ℓ−1}, x_{ℓ−1})
  end for;
  x_ℓ ← x_ℓ − p_ℓ x_{ℓ−1};
  for i := 1 to ν_2 do
    Glätter(x_ℓ, b_ℓ)
  end for
else
  x_ℓ ← A_ℓ^{−1} b_ℓ
end if
```

Abbildung 4.8: Allgemeine Fassung des Mehrgitterverfahrens

wir die Parameter $\nu_1, \nu_2 \in \mathbb{N}_0$ ein, die die Anzahl der *Vor- und Nachglättungsschritte* angeben.

Außerdem kann es passieren, dass ein einzelner Mehrgitterschritt nicht ausreicht, um eine hinreichend gute Approximation der „echten“ Grobgitterkorrektur zu berechnen, also führen wir den Parameter $\gamma \in \mathbb{N}$ ein, der die Anzahl der zur Approximation der Grobgitterlösung verwendeten rekursiven Mehrgitterschritte angibt. Damit erhalten wir das allgemeine Mehrgitterverfahren, das in Abbildung 4.8 angegeben ist. In der Praxis sind vor allem der sogenannte *V-Zyklus*, also der Fall $\gamma = 1$, und der *W-Zyklus*, nämlich der Fall $\gamma = 2$, relevant, die ihre Namen der Form der Struktur der Rekursion verdanken.

Wie wir sehen, setzt sich ein Schritt des Mehrgitterverfahrens im Wesentlichen aus wohlbekannten Bestandteilen zusammen: Als Glätter können viele der im bisherigen Verlauf der Vorlesung vorgestellten iterativen Verfahren eingesetzt werden, besonders populär sind Varianten der Jacobi- und Gauß-Seidel-Iterationen. Abgesehen vom Glätter treten lediglich Multiplikationen zwischen schwachbesetzten Matrizen und Vektoren sowie Linearkombinationen auf, so dass wir erwarten dürfen, dass der Rechenaufwand, sieht man von dem rekursiven Aufruf ab, im Wesentlichen proportional zu n ist.

Dementsprechend setzen wir voraus, dass es Konstanten $C_G \in \mathbb{R}_{>0}$, $C_D \in \mathbb{R}_{>0}$ und $C_P \in \mathbb{R}_{>0}$ so gibt, dass

- die Durchführung eines Glättungsschrittes auf Stufe $\ell \in \mathbb{N}$ höchstens $C_G|\mathcal{I}_\ell|$ Operationen erfordert,
- die Berechnung des Defekts $\mathbf{d}_\ell$ auf Stufe $\ell \in \mathbb{N}$ höchstens $C_D|\mathcal{I}_\ell|$ Operationen

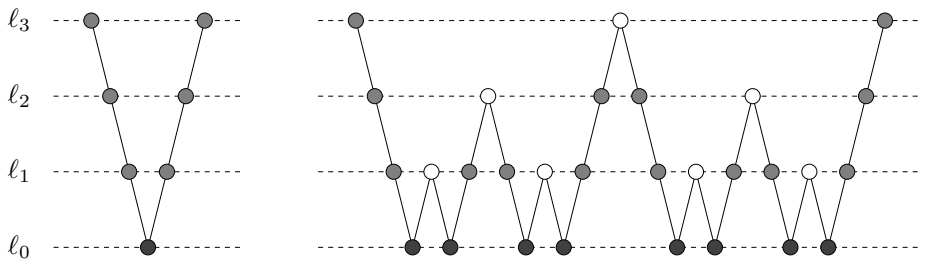

Abbildung 4.9: Struktur der Rekursion für $\gamma = 1$ (links) und $\gamma = 2$ (rechts)

erfordert und

- sowohl die Berechnung von $\mathbf{b}_{\ell-1}$ mit Hilfe der Restriktion als auch die Aktualisierung von $\mathbf{x}_\ell$ mit Hilfe der Prolongation jeweils höchstens $C_P|\mathcal{I}_\ell|$ Operationen erfordern.

Um den Gesamtaufwand für einen Schritt der Mehrgitteriteration abschätzen zu können, müssen wir also einen Weg finden, um seine rekursive Struktur (vgl. Abbildung 4.9) in den Griff zu bekommen.

Ein eleganter Ansatz besteht darin, vorauszusetzen, dass die Anzahl der Freiheitsgrade auf jeder Gitterstufe mindestens um einen gewissen Faktor kleiner als auf der nächstfeineren Stufe ist, dass es also ein $\alpha \in [0, 1)$ so gibt, dass

$$|\mathcal{I}_{\ell-1}| \leq \alpha|\mathcal{I}_\ell| \qquad \text{für alle } \ell \in \mathbb{N}$$

gilt. Wir bezeichnen die Anzahl der auf Stufe $\ell \in \mathbb{N}_0$ benötigten Rechenoperationen mit $C_\ell \in \mathbb{N}_0$. Um einen Schritt auf Stufe $\ell = 1$ durchzuführen, sind zunächst ν_1 Vorglättungsschritte erforderlich, dann muss der Defekt berechnet, auf das Gitter der Stufe $\ell = 0$ transportiert, dort die Grobgittergleichung gelöst und anschließend die Korrektur zur aktuellen Iterierten addiert und ν_2 Nachglättungsschritte durchgeführt werden, so dass sich ein Gesamtaufwand von

$$C_1 \leq C_G(\nu_1 + \nu_2)|\mathcal{I}_1| + C_D|\mathcal{I}_1| + 2C_P|\mathcal{I}_1| + C_0\gamma$$

ergibt. Allgemein erhalten wir die Rekursionsformel

$$C_\ell \leq (C_G(\nu_1 + \nu_2) + C_D + 2C_P)|\mathcal{I}_\ell| + C_{\ell-1}\gamma \qquad \text{für alle } \ell \in \mathbb{N}.$$

Zur Lösung dieser Gleichung verwenden wir den Ansatz

$$C_\ell \leq C_{\mathrm{mg}}|\mathcal{I}_\ell|$$

und erhalten

$$C_\ell \leq (C_G(\nu_1 + \nu_2) + C_D + 2C_P)|\mathcal{I}_\ell| + C_{\ell-1}\gamma$$

$$
\begin{aligned}
&\leq (C_G(\nu_1+\nu_2)+C_D+2C_P)|\mathcal{I}_\ell| + C_{\mathrm{mg}}\gamma|\mathcal{I}_{\ell-1}| \\
&\leq (C_G(\nu_1+\nu_2)+C_D+2C_P)|\mathcal{I}_\ell| + C_{\mathrm{mg}}\gamma\alpha|\mathcal{I}_\ell| \\
&= (C_G(\nu_1+\nu_2)+C_D+2C_P+C_{\mathrm{MG}}\gamma\alpha)|\mathcal{I}_\ell|.
\end{aligned}
$$

Wenn wir nachweisen können, dass

$$C_G(\nu_1+\nu_2)+C_D+2C_P+C_{\mathrm{mg}}\gamma\alpha \leq C_{\mathrm{mg}}$$

gilt, haben wir die gewünschte Aufwandsabschätzung bewiesen. Offensichtlich kann diese Ungleichung nur dann gelten, wenn

$$\gamma\alpha < 1 \tag{4.4}$$

erfüllt ist, wenn also die Anzahl der Freiheitsgrade auf dem groben Gitter hinreichend viel kleiner als auf dem feinen Gitter ist. Wenn wir (4.4) voraussetzen, erhalten wir

$$C_G(\nu_1+\nu_2)+C_D+2C_P \leq (1-\gamma\alpha)C_{\mathrm{mg}},$$

also ist

$$C'_{\mathrm{mg}} := \frac{C_G(\nu_1+\nu_2)+C_D+2C_P}{1-\gamma\alpha}$$

ein guter Kandidat für unsere Aufwandsabschätzung. Um auch den Fall $\ell = 0$ zu berücksichtigen, müssen wir die etwas modifizierte Definition

$$C_{\mathrm{mg}} := \max\left\{\frac{C_G(\nu_1+\nu_2)+C_D+2C_P}{1-\gamma\alpha}, \frac{C_0}{|\mathcal{I}_0|}\right\}$$

verwenden und haben bewiesen, dass der Rechenaufwand des Mehrgitterverfahrens unter der Voraussetzung (4.4) sich tatsächlich proportional zu der Anzahl der Freiheitsgrade verhält.

Im Fall des Modellproblems (und auch in vielen anderen praktisch relevanten Fällen) dürfen wir annehmen, dass die Berechnung des Defekts nicht mehr Operationen als die Durchführung eines Glättungsschrittes erfordert, und dass die Berechnung von Restriktion *und* Prolongation zusammen ebenfalls nicht mehr Operationen als ein Glättungsschritt benötigt, dass also $C_D \leq C_G$ und $2C_P \leq C_G$ gelten. Im eindimensionalen Modellproblem können wir $\alpha = 1/2$, im zweidimensionalen sogar $\alpha = 1/4$ voraussetzen, und in beiden Fällen genügt $\gamma = 1$, um ein schnell konvergierendes Mehrgitterverfahren zu erhalten. Der Rechenaufwand wird also im eindimensionalen Fall ungefähr

$$C_\ell \leq C_{\mathrm{mg}}|\mathcal{I}_\ell| \approx \frac{C_G(\nu_1+\nu_2+2)}{1-\gamma\alpha}|\mathcal{I}_\ell| = 2C_G(\nu_1+\nu_2+2)|\mathcal{I}_\ell|$$

betragen, also schlimmstenfalls (für $\nu_1 = 1$, $\nu_2 = 0$) sechsmal so hoch wie der Aufwand eines einzelnen Glättungsschrittes sein, im zweidimensionalen Fall erhalten wir

$$C_\ell \leq C_{\mathrm{mg}}|\mathcal{I}_\ell| \approx \frac{C_G(\nu_1+\nu_2+2)}{1-\gamma\alpha}|\mathcal{I}_\ell| = \frac{4}{3}C_G(\nu_1+\nu_2+2)|\mathcal{I}_\ell|,$$

also sogar nur einen Faktor von vier.

Der Rechenaufwand des Mehrgitterverfahrens wird also im Wesentlichen durch den Glätter bestimmt, und wie wir bereits gesehen haben, genügt bereits ein sehr einfaches, und damit schnelles, Glättungsverfahren wie die Richardson-Iteration.

4.3 Konvergenzbeweis per Fourier-Analyse

Im einfachen Fall des eindimensionalen Modellproblems ist es möglich, eine Abschätzung für die Konvergenzgeschwindigkeit zumindest des Zweigitterverfahrens explizit nachzurechnen, indem ausgenutzt wird, dass alle Eigenvektoren bekannt sind.

Wir fixieren eine Stufe $\ell \in \mathbb{N}$, auf der wir das Zweigitterverfahren analysieren wollen. Die zu der Matrix $\mathbf{A}_\ell$ gehörenden Eigenvektoren $(\mathbf{e}^k)_{k\in\mathcal{I}_\ell}$ sind (vgl. Lemma 1.7), durch

$$e_j^k := \sqrt{2h_\ell}\sin(\pi jkh_\ell) \qquad \text{für alle } j,k \in \mathcal{I}_\ell$$

gegeben und gehören zu den Eigenwerten

$$\lambda_k := 4h_\ell^{-2}\sin^2(\pi kh_\ell/2) \qquad \text{für alle } k \in \mathcal{I}_\ell.$$

Um die Grobgitterkorrektur analysieren zu können, benötigen wir auch eine Eigenvektorbasis $(\widehat{\mathbf{e}}^k)_{k\in\mathcal{I}_{\ell-1}}$ auf der Stufe $\ell-1$, die durch

$$\widehat{e}_j^k := \sqrt{2h_{\ell-1}}\sin(\pi jkh_{\ell-1}) = \sqrt{4h_\ell}\sin(\pi jkh_{\ell-1}) \qquad \text{für alle } j,k \in \mathcal{I}_{\ell-1}$$

gegeben ist und zu den Eigenwerten

$$\widehat{\lambda}_k := 4h_{\ell-1}^{-2}\sin^2(\pi kh_{\ell-1}/2) = h_\ell^{-2}\sin^2(\pi kh_{\ell-1}/2) \qquad \text{für alle } k \in \mathcal{I}_{\ell-1}$$

der Matrix $\mathbf{A}_{\ell-1}$ gehört. Zur Vereinfachung der Notation setzen wir $\mathcal{I} := \mathcal{I}_\ell$, $\widehat{\mathcal{I}} := \mathcal{I}_{\ell-1}$, $h := h_\ell$ und $\widehat{h} := h_{\ell-1}$ sowie $\mathbf{A} := \mathbf{A}_\ell$, $\widehat{\mathbf{A}} := \mathbf{A}_{\ell-1}$, $\mathbf{r} := \mathbf{r}_\ell$ und $\mathbf{p} := \mathbf{p}_\ell$.

Unsere Aufgabe besteht zunächst darin, die Auswirkungen der Grobgitterkorrektur zu analysieren. Für $k \in \mathcal{I}$ und $j \in \widehat{\mathcal{I}}$ erhalten wir mit Hilfe der Additionstheoreme $\sin(x+y) = \cos(x)\sin(y) + \sin(x)\cos(y)$ und $\cos(x+y) = \cos(x)\cos(y) - \sin(x)\sin(y)$ die Gleichung

$$\begin{aligned}
(\mathbf{re}^k)_j &= \frac{1}{4}e_{2j-1}^k + \frac{1}{2}e_{2j}^k + \frac{1}{4}e_{2j+1}^k \\
&= \frac{\sqrt{2h}}{4}\left(\sin(\pi(2j-1)kh) + 2\sin(\pi(2j)kh) + \sin(\pi(2j+1)kh)\right) \\
&= \frac{\sqrt{2h}}{4}\left(\cos(-\pi kh)\sin(\pi(2j)kh) + \sin(-\pi kh)\cos(\pi(2j)kh) + 2\sin(\pi(2j)kh)\right. \\
&\qquad \left. + \cos(\pi kh)\sin(\pi(2j)kh) + \sin(\pi kh)\cos(\pi(2j)kh)\right) \\
&= \frac{\sqrt{2h}}{4}(2\cos(\pi kh)+2)\sin(\pi(2j)kh) = \frac{\sqrt{2h}}{2}(\cos(\pi kh)+1)\sin(\pi jk\widehat{h}) \\
&= \frac{\sqrt{2h}}{2}(\cos^2(\pi kh/2) - \sin^2(\pi kh/2) + 1)\sin(\pi jk\widehat{h}) \\
&= \sqrt{2h}\cos^2(\pi kh/2)\sin(\pi jk\widehat{h}).
\end{aligned}$$

Für $k = \widehat{N}+1$ ist der letzte Term Null, also folgt

$$\mathbf{re}^{\widehat{N}+1} = \mathbf{0}. \tag{4.5}$$

Falls dagegen $k \in \widehat{\mathcal{I}}$ gilt, haben wir

$$\mathbf{r}\mathbf{e}^k = \sqrt{\frac{1}{2}}\cos^2(\pi kh/2)\widehat{\mathbf{e}}^k$$

nachgewiesen. Anderenfalls gilt $k^* := N + 1 - k \in \widehat{\mathcal{I}}$, und aus

$$\begin{aligned}\sin(\pi jk\widehat{h}) &= \sin(2\pi jkh) = -\sin(2\pi j - 2\pi jkh) = -\sin(2\pi j(N+1)h - 2\pi jkh)\\ &= -\sin(2\pi jk^*h) = -\sin(\pi jk^*\widehat{h})\end{aligned}$$

und der Gleichung

$$\cos(\pi kh/2) = \sin(\pi/2 - \pi kh/2) = \sin(\pi(N+1)h/2 - \pi kh/2) = \sin(\pi k^*h/2)$$

folgt die Gleichung

$$(\mathbf{r}\mathbf{e}^k)_j = \sqrt{2h}\cos^2(\pi kh/2)\sin(\pi jk\widehat{h}) = -\sqrt{2h}\sin^2(\pi k^*h/2)\sin(\pi jk^*\widehat{h}),$$

und somit

$$\mathbf{r}\mathbf{e}^k = -\sqrt{\frac{1}{2}}\sin^2(\pi k^*h/2)\widehat{\mathbf{e}}^{k^*}.$$

Wir führen die Kurznotation

$$s_k := \sin(\pi kh/2), \qquad c_k := \cos(\pi kh/2) \qquad \text{für alle } k \in \widehat{\mathcal{I}}$$

ein und fassen unser erstes Teilergebnis in der Form

$$\mathbf{r}\begin{pmatrix}\mathbf{e}^k & \mathbf{e}^{k^*}\end{pmatrix} = \widehat{\mathbf{e}}^k\begin{pmatrix}\frac{c_k^2}{\sqrt{2}} & -\frac{s_k^2}{\sqrt{2}}\end{pmatrix} \qquad \text{für alle } k \in \widehat{\mathcal{I}}$$

zusammen. Wenden wir uns nun der Prolongation zu. Wir wählen $k \in \widehat{\mathcal{I}}$ und $j \in \mathcal{I}$ und betrachten zunächst den Fall $j = 2m + 1$ für $m \in \{0, \ldots, \widehat{N}\}$. Nach Definition der Prolongation (und mit der Konvention $e_0^k = 0$) gilt

$$\begin{aligned}(\mathbf{p}\widehat{\mathbf{e}}^k)_j &= \frac{1}{2}\widehat{e}_m^k + \frac{1}{2}\widehat{e}_{m+1}^k\\ &= \frac{\sqrt{4h}}{2}\Big(\sin(\pi mk\widehat{h}) + \sin(\pi(m+1)k\widehat{h})\Big)\\ &= \sqrt{h}\,(\sin(\pi(2m)kh) + \sin(\pi(2m+2)kh))\\ &= \sqrt{h}\,(\sin(\pi(j-1)kh) + \sin(\pi(j+1)kh))\\ &= \sqrt{h}\,(\cos(-\pi kh)\sin(\pi jkh) + \sin(-\pi kh)\cos(\pi jkh)\\ &\qquad + \cos(\pi kh)\sin(\pi jkh) + \sin(\pi kh)\cos(\pi jkh))\\ &= 2\sqrt{h}\cos(\pi kh)\sin(\pi jkh) = 2\sqrt{h}(\cos^2(\pi kh/2) - \sin^2(\pi kh/2))\sin(\pi jkh).\end{aligned}$$

Da j ungerade ist, gilt $\sin(\pi j - x) = \sin(x)$ und wir erhalten

$$(\mathbf{p}\widehat{\mathbf{e}}^k)_j = 2\sqrt{h}(\cos^2(\pi kh/2) - \sin^2(\pi kh/2))\sin(\pi jkh)$$

$$
\begin{aligned}
&= 2\sqrt{h}c_k^2 \sin(\pi jkh) - 2\sqrt{h}s_k^2 \sin(\pi j - \pi jkh) \\
&= 2\sqrt{h}c_k^2 \sin(\pi jkh) - 2\sqrt{h}s_k^2 \sin(\pi j(N+1)h - \pi jkh) \\
&= 2\sqrt{h}c_k^2 \sin(\pi jkh) - 2\sqrt{h}s_k^2 \sin(\pi jk^*h),
\end{aligned}
$$

dürfen also als Zwischenergebnis

$$
(\mathbf{p}\widehat{\mathbf{e}}^k)_j = \sqrt{2}c_k^2 e_j^k - \sqrt{2}s_k^2 e_j^{k^*} \qquad \text{für alle } j = 2m+1, m \in \{0, \dots, \widehat{N}\}
$$

festhalten. Wenden wir uns nun dem Fall $j = 2m$ für $m \in \widehat{\mathcal{I}}$ zu. Nach Definition gilt

$$
(\mathbf{p}\widehat{e}^k)_j = \widehat{e}_m^k = \sqrt{4h}\sin(\pi mk\widehat{h}) = \sqrt{4h}\sin(\pi jkh) = \sqrt{4h}(c_k^2 + s_k^2)\sin(\pi jkh).
$$

Da diesmal j gerade ist, gilt $\sin(\pi j - x) = -\sin(x)$ und wir finden

$$
\begin{aligned}
(\mathbf{p}\widehat{e}^k)_j &= \sqrt{4h}(c_k^2 + s_k^2)\sin(\pi jkh) \\
&= \sqrt{4h}c_k^2 \sin(\pi jkh) + \sqrt{4h}s_k^2 \sin(\pi jkh) \\
&= \sqrt{4h}c_k^2 \sin(\pi jkh) - \sqrt{4h}s_k^2 \sin(\pi j - \pi jkh) \\
&= \sqrt{4h}c_k^2 \sin(\pi jkh) - \sqrt{4h}s_k^2 \sin(\pi jk^*h),
\end{aligned}
$$

so dass wir

$$
\mathbf{p}\widehat{\mathbf{e}}^k = \sqrt{2}c_k^2 \mathbf{e}^k - \sqrt{2}s_k^2 \mathbf{e}^{k^*} \qquad \text{für alle } k \in \widehat{\mathcal{I}}
$$

bewiesen haben. In Matrixnotation schreibt sich diese Gleichung als

$$
\mathbf{p}\widehat{\mathbf{e}}^k = \begin{pmatrix} \mathbf{e}^k & \mathbf{e}^{k^*} \end{pmatrix} \begin{pmatrix} \sqrt{2}c_k^2 \\ -\sqrt{2}s_k^2 \end{pmatrix} \qquad \text{für alle } k \in \widehat{\mathcal{I}}.
$$

Nun können wir die Grobgitterkorrektur analysieren: Wir haben

$$
\sin(\pi k^*h/2) = \sin(\pi/2 - \pi kh/2) = \cos(\pi kh/2),
$$

also lässt sich die Anwendung der Matrix $\mathbf{A}$ in der Form

$$
\mathbf{A}\begin{pmatrix} \mathbf{e}^k & \mathbf{e}^{k^*} \end{pmatrix} = \begin{pmatrix} \mathbf{e}^k & \mathbf{e}^{k^*} \end{pmatrix} \begin{pmatrix} 4h^{-2}s_k^2 & \\ & 4h^{-2}c_k^2 \end{pmatrix} \qquad \text{für alle } k \in \widehat{\mathcal{I}}
$$

darstellen und wegen

$$
\sin(\pi k\widehat{h}/2) = \sin(\pi kh) = \sin(2\pi kh/2) = 2\sin(\pi kh/2)\cos(\pi kh/2) = 2s_k c_k
$$

erhalten wir für die Grobgittermatrix

$$
\widehat{\mathbf{A}}\widehat{\mathbf{e}}^k = \widehat{\mathbf{e}}^k 4h^{-2}s_k^2 c_k^2 \qquad \text{für alle } k \in \widehat{\mathcal{I}}.
$$

Nun können wir mit der Analyse der Grobgitterkorrektur beginnen. Es gilt

$$
\mathbf{p}\widehat{\mathbf{A}}^{-1}\mathbf{r}\mathbf{A}\begin{pmatrix} \mathbf{e}^k & \mathbf{e}^{k^*} \end{pmatrix} = \mathbf{p}\widehat{\mathbf{A}}^{-1}\mathbf{r}\begin{pmatrix} \mathbf{e}^k & \mathbf{e}^{k^*} \end{pmatrix} \begin{pmatrix} 4h^{-2}s_k^2 & \\ & 4h^{-2}c_k^2 \end{pmatrix}
$$

$$
\begin{aligned}
&= \mathbf{p}\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{e}}^k \begin{pmatrix} \frac{c_k^2}{\sqrt{2}} & -\frac{s_k^2}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 4h^{-2}s_k^2 & \\ & 4h^{-2}c_k^2 \end{pmatrix} \\
&= \mathbf{p}\widehat{\mathbf{e}}^k \frac{h^2}{4s_k^2c_k^2} \begin{pmatrix} \frac{c_k^2}{\sqrt{2}} & -\frac{s_k^2}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 4h^{-2}s_k^2 & \\ & 4h^{-2}c_k^2 \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{e}^k & \mathbf{e}^{k^*} \end{pmatrix} \begin{pmatrix} \sqrt{2}c_k^2 \\ -\sqrt{2}s_k^2 \end{pmatrix} \frac{h^2}{4s_k^2c_k^2} \begin{pmatrix} \frac{c_k^2}{\sqrt{2}} & -\frac{s_k^2}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 4h^{-2}s_k^2 & \\ & 4h^{-2}c_k^2 \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{e}^k & \mathbf{e}^{k^*} \end{pmatrix} \frac{1}{s_k^2c_k^2} \begin{pmatrix} s_k^2c_k^4 & -s_k^2c_k^4 \\ -s_k^4c_k^2 & s_k^4c_k^2 \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{e}^k & \mathbf{e}^{k^*} \end{pmatrix} \begin{pmatrix} c_k^2 & -c_k^2 \\ -s_k^2 & s_k^2 \end{pmatrix} \qquad \text{für } k \in \widehat{\mathcal{I}},
\end{aligned}
$$

so dass wir für die Iterationsmatrix der Grobgitterkorrektur die Gleichung

$$
(\mathbf{I} - \mathbf{p}\widehat{\mathbf{A}}^{-1}\mathbf{r}\mathbf{A}) \begin{pmatrix} \mathbf{e}^k & \mathbf{e}^{k^*} \end{pmatrix} = \begin{pmatrix} \mathbf{e}^k & \mathbf{e}^{k^*} \end{pmatrix} \begin{pmatrix} s_k^2 & c_k^2 \\ s_k^2 & c_k^2 \end{pmatrix} \qquad \text{für } k \in \widehat{\mathcal{I}}
$$

erhalten. Als Glättungsverfahren verwenden wir die Richardson-Iteration mit Dämpfungsparameter $\theta_{\mathrm{gl}} := h^2/4$, deren Iterationsmatrix die Gleichung

$$
(\mathbf{I} - \theta_{\mathrm{gl}}\mathbf{A}) \begin{pmatrix} \mathbf{e}^k & \mathbf{e}^{k^*} \end{pmatrix} = \begin{pmatrix} \mathbf{e}^k & \mathbf{e}^{k^*} \end{pmatrix} \begin{pmatrix} c_k^2 & \\ & s_k^2 \end{pmatrix} \qquad \text{für } k \in \widehat{\mathcal{I}}
$$

erfüllt. Für die Iterationsmatrix des Zweigitterverfahrens mit $\nu \in \mathbb{N}$ Vorglättungsschritten ergibt sich so die Gleichung

$$
\mathbf{M}_{\mathrm{ZGV}} \begin{pmatrix} \mathbf{e}^k & \mathbf{e}^{k^*} \end{pmatrix} = \begin{pmatrix} \mathbf{e}^k & \mathbf{e}^{k^*} \end{pmatrix} \begin{pmatrix} s_k^2c_k^{2\nu} & c_k^2s_k^{2\nu} \\ s_k^2c_k^{2\nu} & c_k^2s_k^{2\nu} \end{pmatrix} \qquad \text{für } k \in \widehat{\mathcal{I}}. \tag{4.6}
$$

Zusammen mit $\mathbf{e}^{\widehat{N}+1}$ bilden die Vektoren $\mathbf{e}^k, \mathbf{e}^{k^*}$ für $k \in \widehat{\mathcal{I}}$ eine Orthonormalbasis von $\mathbb{K}^{\mathcal{I}}$, also erhalten wir durch Kombination von (4.5) und (4.6) die Gleichung

$$
\varrho(\mathbf{M}_{\mathrm{ZGV}}) = \max \left\{ \varrho \begin{pmatrix} s_k^2c_k^{2\nu} & c_k^2s_k^{2\nu} \\ s_k^2c_k^{2\nu} & c_k^2s_k^{2\nu} \end{pmatrix}, \frac{1}{2^\nu} \ : \ k \in \widehat{\mathcal{I}} \right\}.
$$

Wegen $s_k^2 \in [0, 1/2]$ genügt es, eine obere Schranke für den Spektralradius der Matrizen

$$
\mathbf{X}(\xi) := \begin{pmatrix} \xi(1-\xi)^\nu & (1-\xi)\xi^\nu \\ \xi(1-\xi)^\nu & (1-\xi)\xi^\nu \end{pmatrix} \qquad \text{für alle } \xi \in [0, 1/2]
$$

zu finden. Zwei Eigenvektoren von $\mathbf{X}(\xi)$ lassen sich schnell finden: Es gelten

$$
\begin{aligned}
\mathbf{X}(\xi) \begin{pmatrix} \xi^{\nu-1} \\ (1-\xi)^{\nu-1} \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix} && \text{für alle } \xi \in [0, 1/2], \\
\mathbf{X}(\xi) \begin{pmatrix} 1 \\ 1 \end{pmatrix} &= (\xi(1-\xi)^\nu + \xi^\nu(1-\xi)) \begin{pmatrix} 1 \\ 1 \end{pmatrix} && \text{für alle } \xi \in [0, 1/2],
\end{aligned}
$$

wir müssen also lediglich eine obere Schranke der Funktion

$$\varrho_\nu : [0, 1/2] \to \mathbb{R}_{\geq 0}, \qquad \xi \mapsto \xi(1-\xi)^\nu + \xi^\nu(1-\xi),$$

bestimmen. Für den einfachen Fall $\nu = 1$ erhalten wir sofort $\varrho(\mathbf{M}_{\mathrm{ZGV}}) \leq 1/2$, und diese Aussage bleibt offenbar auch für größere Anzahlen von Glättungsschritten gültig.

Um eine bessere Abschätzung für große Werte von ν zu gewinnen, betrachten wir die beiden Summanden separat: Durch Differenzieren stellen wir fest, dass der erste von ihnen sein Maximum bei $\xi_0 := 1/(\nu+1)$ annimmt, und dieses Maximum lässt sich durch

$$\frac{1}{\nu+1}\left(\frac{\nu}{\nu+1}\right)^\nu = \frac{1}{\nu}\left(\frac{\nu}{\nu+1}\right)^{\nu+1} = \frac{1}{\nu}\left(1+\frac{1}{\nu}\right)^{-(\nu+1)} \leq \frac{1}{e\nu}$$

abschätzen. Der zweite der beiden Terme ist wegen $\xi \leq 1/2$ durch $2^{-\nu}$ zu beschränken, so dass wir

$$\varrho(\mathbf{M}_{\mathrm{ZGV}}) \leq \frac{1}{e\nu} + 2^{-\nu} \qquad \text{für alle } \nu \in \mathbb{N}$$

erhalten, nicht nur ist also die Konvergenzrate von der Stufenzahl ℓ unabhängig, wir können sie auch auf einen beliebig kleinen Wert reduzieren, indem wir hinreichend viele Glättungsschritte durchführen.

4.4 Konvergenz des W-Zyklus-Mehrgitterverfahrens

Im Spezialfall des eindimensionalen Modellproblems haben wir nachweisen können, dass das Zweigitterverfahren für ν Glättungsschritte mit einer Rate von $\sim 1/\nu$ konvergiert, und dass diese Rate insbesondere von der Anzahl der Unbekannten unabhängig ist.

Wir wollen nun ein ähnliches Resultat für das Mehrgitterverfahren beweisen. Der entscheidende Unterschied zwischen dem bereits untersuchten Zweigitter- und dem Mehrgitterverfahren besteht darin, dass im letzteren Fall die Grobgitterkorrektur nicht exakt, sondern lediglich approximativ erfolgt. Zur Lösung dieser Aufgabe leiten wir zunächst ein allgemeines Konvergenzkriterium für das Zweigitterverfahren her, das sich dann auf den Fall der Mehrgitterverfahrens übertragen lässt.

Entscheidend bei der Einführung des Zweigitterverfahrens war die Beobachtung, dass ein einfaches iteratives Verfahren wie etwa die Richardson-Iteration für hochfrequente Anteile des Fehlers eine von der Problemgröße unabhängige Konvergenzrate besitzt.

Da hochfrequente Eigenvektoren, zumindest im Fall des Modellproblems, zu großen Eigenwerten gehören, können wir die „Glattheit“ eines Vektors

$$\mathbf{x}_\ell = \sum_{k \in \mathcal{I}_\ell} \alpha_k \mathbf{e}^k \in \mathbb{K}^{\mathcal{I}_\ell}$$

mit Hilfe der Norm

$$\|\mathbf{A}_\ell \mathbf{x}_\ell\|_2^2 = \sum_{k \in \mathcal{I}_\ell} |\lambda_k|^2 |\alpha_k|^2$$

messen: Wenn diese Norm klein ist, müssen alle Summanden klein sein, also können hochfrequente Eigenvektoren keine große Rolle in $\mathbf{x}_\ell$ spielen, da sie zu großen Eigenwerten gehören.

Also ist $\Phi_{\mathrm{Gl},\ell}$ ein gutes Glättungsverfahren, wenn der Iterationsfehler in der oben angegebenen Norm reduziert wird, wenn also $\|\mathbf{A}_\ell \mathbf{M}_{\mathrm{Gl},\ell}^\nu\|_2$ sich unabhängig von der Gitterstufe ℓ beschränken lässt.

Um das gewünschte Verhalten quantifizieren zu können, untersuchen wir zunächst das Richardson-Verfahren. Für eine positiv definite Matrix $\mathbf{A}_\ell$ konvergiert das Richardson-Verfahren mit einem Dämpfungsparameter von $\theta_\ell := 1/\|\mathbf{A}_\ell\|_2$, und die entsprechende Iterationsmatrix ist durch

$$\mathbf{M}_{\mathrm{Rich},\theta_\ell} := \mathbf{I} - \theta_\ell \mathbf{A}_\ell$$

gegeben. Wenn wir $\mathbf{M}_{\mathrm{Rich},\theta_\ell}$ als Glättungsverfahren einsetzen wollen, müssen wir beweisen, dass

$$\|\mathbf{A}_\ell \mathbf{M}_{\mathrm{Rich},\theta_\ell}^\nu\|_2 = \|\mathbf{A}_\ell(\mathbf{I} - \theta_\ell \mathbf{A}_\ell)^\nu\|_2 = \frac{1}{\theta_\ell}\|\theta_\ell \mathbf{A}_\ell(\mathbf{I} - \theta_\ell \mathbf{A}_\ell)^\nu\|_2$$

sich durch eine von der Stufe ℓ unabhängige Konstante beschränken lässt. Wir setzen $\mathbf{X}_\ell := \theta_\ell \mathbf{A}_\ell$ und stellen fest, dass aus der Selbstadjungiertheit von $\mathbf{A}_\ell$ bereits

$$\|\mathbf{X}_\ell(\mathbf{I} - \mathbf{X}_\ell)^\nu\|_2 = \varrho\left(\mathbf{X}_\ell(\mathbf{I} - \mathbf{X}_\ell)^\nu\right) = \max\{\lambda(1-\lambda)^\nu \ : \ \lambda \in \sigma(\mathbf{X}_\ell)\}$$

folgt. Da $\mathbf{X}_\ell$ positiv definit ist, gilt $\mathbf{X}_\ell > \mathbf{0}$. Da $\varrho(\mathbf{X}_\ell) = \varrho(\mathbf{A}_\ell)/\|\mathbf{A}_\ell\|_2 = 1$ gilt, erhalten wir somit $\sigma(\mathbf{X}_\ell) \in [0,1]$ und

$$\|\mathbf{X}_\ell(\mathbf{I} - \mathbf{X}_\ell)^\nu\|_2 \le \max\{\lambda(1-\lambda)^\nu \ : \ \lambda \in [0,1]\}.$$

Die auf der rechten Seite auftretende Funktion

$$\lambda \mapsto \lambda(1-\lambda)^\nu$$

nimmt ihr Maximum für $\lambda_\nu := 1/(\nu+1)$ an, so dass wir die Schranke

$$\|\mathbf{X}_\ell(\mathbf{I} - \mathbf{X}_\ell)^\nu\|_2 \le \eta(\nu) := \frac{1}{\nu+1}\left(\frac{\nu}{\nu+1}\right)^\nu \qquad \text{für alle } \nu \in \mathbb{N}$$

erhalten. Diese Schranke ist von der Stufe ℓ unabhängig, und eine nähere Untersuchung ergibt

$$\eta(\nu) = \frac{1}{\nu+1}\left(\frac{\nu}{\nu+1}\right)^\nu = \frac{1}{\nu}\left(\frac{\nu}{\nu+1}\right)^{\nu+1} = \frac{1}{\nu}\left(1 + \frac{1}{\nu}\right)^{-(\nu+1)} \le \frac{1}{e\nu},$$

auch im allgemeinen Fall wird also $\eta(\nu)$ für $\nu \to \infty$ gegen Null streben. Wir fassen zusammen: Für das Richardson-Verfahren gilt

$$\|\mathbf{A}_\ell \mathbf{M}_{\mathrm{Rich},\theta_\ell}^\nu\|_2 \le \eta(\nu)\|\mathbf{A}_\ell\|_2 \qquad \text{für alle } \ell \in \mathbb{N}_0, \nu \in \mathbb{N}, \tag{4.7}$$

und $\eta(\nu) \to 0$ für $\nu \to \infty$. Diese Beobachtung führt zu dem ersten Teil des Konvergenzkriteriums für das Zweigitterverfahren:

Definition 4.8 (Glättungseigenschaft) *Sei $\ell \in \mathbb{N}$. Sei Φ_ℓ ein lineares Iterationsverfahren mit Iterationsmatrix $\mathbf{M}_\ell$ für das Gleichungssystem $\mathbf{A}_\ell \mathbf{x}_\ell = \mathbf{b}_\ell$. Falls es eine Funktion $\eta : \mathbb{N} \to \mathbb{R}_{\geq 0}$ mit*

$$\lim_{\nu \to \infty} \eta(\nu) = 0$$

gibt, die die Ungleichung

$$\|\mathbf{A}_\ell \mathbf{M}_\ell^\nu\|_2 \leq \eta(\nu) \|\mathbf{A}_\ell\|_2 \qquad \text{für alle } \ell \in \mathbb{N}, \nu \in \mathbb{N}$$

erfüllt, besitzt Φ_ℓ die Glättungseigenschaft.

Wie wir bereits bewiesen haben, besitzt für positiv definite Matrizen $\mathbf{A}_\ell$ beispielsweise die Richardson-Iteration die Glättungseigenschaft, falls der Dämpfungsparameter θ klein genug gewählt wird. Es lässt sich beweisen, dass unter diesen Bedingungen auch die gedämpfte Jacobi-Iteration, die Gauß-Seidel-Iteration und die Kaczmarz-Iteration die Glättungseigenschaft besitzen. Die ungedämpfte Jacobi-Iteration oder die SOR-Iteration mit großem Überrelaxationsparameter ω hingegen besitzen diese Eigenschaft nicht.

Die Glättungseigenschaft kann auch für semiiterative Verfahren formuliert werden. In diesem Fall zeigt sich, dass auch das Tschebyscheff-Verfahren bei geeigneter Wahl der Parameter eine entsprechend verallgemeinerte Glättungseigenschaft besitzt.

Mit Hilfe der Glättungseigenschaft können wir steuern, wie stark diejenigen Anteile des Fehlers reduziert werden, die zu großen Eigenwerten, also in der Regel zu hochfrequenten Eigenvektoren gehören.

Um auch den Einfluss der Grobgitterkorrektur messen zu können, müssen wir eine Beziehung zwischen den Gleichungssystemen auf verschiedenen Gitterstufen herstellen: Anstatt das System

$$\mathbf{A}_\ell \mathbf{f}_\ell = \mathbf{d}_\ell$$

für den Fehler $\mathbf{f}_\ell := \mathbf{x}_\ell - \mathbf{x}_\ell^*$ zu lösen, lösen wir das Grobgittersystem

$$\mathbf{A}_{\ell-1} \mathbf{f}_{\ell-1} = \mathbf{b}_{\ell-1} = \mathbf{r}_\ell \mathbf{d}_\ell$$

und verwenden $\mathbf{p}_\ell \mathbf{f}_{\ell-1}$ als Approximation von $\mathbf{f}_\ell$. Die Grobgitterkorrektur ist also um so besser, je näher $\mathbf{p}_\ell \mathbf{f}_{\ell-1} = \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1} \mathbf{r}_\ell \mathbf{d}_\ell$ dem tatsächlichen Fehler $\mathbf{f}_\ell = \mathbf{A}_\ell^{-1} \mathbf{d}_\ell$ ist.

Als Maß für die Qualität der Grobgitterkorrektur bietet sich deshalb die Norm

$$\|\mathbf{A}_\ell^{-1} - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1} \mathbf{r}_\ell\|_2$$

an. Für niedrigfrequente Eigenvektoren dürfen wir annehmen, dass beide Terme ähnliche Ergebnisse liefern. Hochfrequente Eigenvektoren dagegen gehören zu großen Eigenwerten von $\mathbf{A}_\ell$ und $\mathbf{A}_{\ell-1}$, also zu besonders kleinen Eigenwerten der Inversen $\mathbf{A}_\ell^{-1}$ und $\mathbf{A}_{\ell-1}^{-1}$, so dass hier die Skalierung zu einer kleinen Norm führt.

Dieses Argument bringt uns zu dem zweiten Teil des Konvergenzkriteriums für das Zweigitterverfahren:

Definition 4.9 (Approximationseigenschaft) *Falls es eine Konstante $C \in \mathbb{R}_{>0}$ so gibt, dass die Ungleichung*

$$\|\mathbf{A}_\ell^{-1} - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1} \mathbf{r}_\ell\|_2 \leq \frac{C}{\|\mathbf{A}_\ell\|_2} \qquad \text{für alle } \ell \in \mathbb{N}$$

gilt, besitzt die Hierarchie der Gleichungssysteme die Approximationseigenschaft.

Die Approximationseigenschaft ist von dem verwendeten Glättungsverfahren vollständig unabhängig und beschreibt den Grad der „Verwandtschaft“ zwischen den einzelnen Gitterstufen.

Der Nachweis der Approximationseigenschaft ist häufig wesentlich aufwendiger als der der Glättungseigenschaft, weil die speziellen Eigenschaften des zugrundeliegenden Problems, beispielsweise der Differentialgleichung und ihrer Diskretisierung, berücksichtigt werden müssen. In der Regel ist es deshalb nicht möglich, diese Eigenschaft allgemein nachzuweisen.

Im Fall des eindimensionalen Modellproblems lässt sie sich allerdings relativ einfach nachrechnen: Wir verifizieren die Abschätzung für die von den Eigenvektoren $\mathbf{e}^k$ aufgespannten invarianten Unterräume. Für den Fall $k = (N_\ell + 1)/2$ lässt sie sich elementar nachrechnen, für die restlichen Unterräume fixieren wir wieder ein $k \in \{1, \ldots, (N_\ell - 1)/2\}$ und betrachten die Eigenvektoren $\mathbf{e}^k$ und $\mathbf{e}^{k^*}$ (mit $k^* = N_\ell + 1 - k$) auf dem feinen Gitter und den Eigenvektor $\widehat{\mathbf{e}}^k$ auf dem groben Gitter. Wie wir bereits in Abschnitt 4.3 gesehen haben, gelten die Gleichungen

$$\begin{aligned}
\mathbf{r}_\ell \begin{pmatrix} \mathbf{e}^k & \mathbf{e}^{k^*} \end{pmatrix} &= \widehat{\mathbf{e}}^k \begin{pmatrix} \frac{c_k^2}{\sqrt{2}} & -\frac{s_k^2}{\sqrt{2}} \end{pmatrix}, \\
\mathbf{p}_\ell \widehat{\mathbf{e}}^k &= \begin{pmatrix} \mathbf{e}^k & \mathbf{e}^{k^*} \end{pmatrix} \begin{pmatrix} \sqrt{2} c_k^2 \\ -\sqrt{2} s_k^2 \end{pmatrix}, \\
\mathbf{A}_\ell \begin{pmatrix} \mathbf{e}^k & \mathbf{e}^{k^*} \end{pmatrix} &= \begin{pmatrix} \mathbf{e}^k & \mathbf{e}^{k^*} \end{pmatrix} \begin{pmatrix} 4h_\ell^{-2} s_k^2 & \\ & 4h_\ell^{-2} c_k^2 \end{pmatrix}, \\
\mathbf{A}_{\ell-1} \widehat{\mathbf{e}}^k &= \widehat{\mathbf{e}}^k 4 h_\ell^{-2} s_k^2 c_k^2,
\end{aligned}$$

so dass wir für den Nachweis der Approximationseigenschaft lediglich das Verhalten der beteiligten Matrizen auf dem von $\mathbf{e}^k$ und $\mathbf{e}^{k^*}$ aufgespannten invarianten Unterraum untersuchen müssen. Für Vektoren aus diesem Unterraum gelten

$$\begin{aligned}
\mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1} \mathbf{r}_\ell \begin{pmatrix} \mathbf{e}^k & \mathbf{e}^{k^*} \end{pmatrix} &= \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1} \widehat{\mathbf{e}}^k \begin{pmatrix} \frac{c_k^2}{\sqrt{2}} & -\frac{s_k^2}{\sqrt{2}} \end{pmatrix} = \mathbf{p}_\ell \widehat{\mathbf{e}}^k \begin{pmatrix} \frac{h_\ell^2}{4 s_k^2 \sqrt{2}} & -\frac{h_\ell^2}{4 c_k^2 \sqrt{2}} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{e}^k & \mathbf{e}^{k^*} \end{pmatrix} \begin{pmatrix} \frac{h_\ell^2 c_k^2}{4 s_k^2} & -\frac{h_\ell^2}{4} \\ -\frac{h_\ell^2}{4} & \frac{h_\ell^2 s_k^2}{4 c_k^2} \end{pmatrix}, \\
\mathbf{A}_\ell^{-1} \begin{pmatrix} \mathbf{e}^k & \mathbf{e}^{k^*} \end{pmatrix} &= \begin{pmatrix} \mathbf{e}^k & \mathbf{e}^{k^*} \end{pmatrix} \begin{pmatrix} \frac{h_\ell^2}{4 s_k^2} & \\ & \frac{h_\ell^2}{4 c_k^2} \end{pmatrix},
\end{aligned}$$

und durch Kombination der beiden Gleichungen erhalten wir

$$\begin{aligned}\left(\mathbf{A}_\ell^{-1} - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\right)\begin{pmatrix}\mathbf{e}^k & \mathbf{e}^{k^*}\end{pmatrix} &= \begin{pmatrix}\mathbf{e}^k & \mathbf{e}^{k^*}\end{pmatrix}\begin{pmatrix}\frac{h_\ell^2(1-c_k^2)}{4s_k^2} & \frac{h_\ell^2}{4} \\ \frac{h_\ell^2}{4} & \frac{h_\ell^2(1-s_k^2)}{4c_k^2}\end{pmatrix} \\ &= \begin{pmatrix}\mathbf{e}^k & \mathbf{e}^{k^*}\end{pmatrix}\begin{pmatrix}\frac{h_\ell^2}{4} & \frac{h_\ell^2}{4} \\ \frac{h_\ell^2}{4} & \frac{h_\ell^2}{4}\end{pmatrix}.\end{aligned}$$

Mit Hilfe der elementaren Abschätzung

$$\varrho\begin{pmatrix}1 & 1 \\ 1 & 1\end{pmatrix} = 2$$

erhalten wir damit das gewünschte Ergebnis

$$\|\mathbf{A}_\ell^{-1} - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\|_2 = \varrho(\mathbf{A}_\ell^{-1} - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell) = 2\frac{h_\ell^2}{4} \le \frac{2}{\|\mathbf{A}_\ell\|_2},$$

das eindimensionale Modellproblem besitzt also die Approximationseigenschaft mit der Konstanten $C = 2$.

Mit Hilfe von Approximations- und Glättungseigenschaft lässt sich nun sehr einfach eine Konvergenzaussage für das Zweigitterverfahren gewinnen:

Satz 4.10 (Konvergenz Zweigitterverfahren) *Für alle $\ell \in \mathbb{N}$ sei $\Phi_{\mathrm{Gl},\ell}$ ein lineares Iterationsverfahren mit der Glättungseigenschaft. Die Hierarchie $(\mathbf{A}_\ell)_{\ell\in\mathbb{N}_0}$ besitze die Approximationseigenschaft. Sei $\varrho \in (0,1)$. Dann können wir ein $\nu \in \mathbb{N}$ so wählen, dass das Zweigitterverfahren mit ν Vorglättungsschritten konvergent ist und die Konvergenzrate die Abschätzung*

$$\|\mathbf{M}_{\mathrm{ZGV},\ell}\|_2 \le \varrho \qquad \textit{für alle } \ell \in \mathbb{N} \textit{ erfüllt.}$$

Beweis. Die Iterationsmatrix des Zweigitterverfahrens ist durch

$$\begin{aligned}\mathbf{M}_{\mathrm{ZGV},\ell} &= \mathbf{M}_{\mathrm{GGK},\ell}\mathbf{M}_{\mathrm{Gl},\ell}^\nu = (\mathbf{I} - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\mathbf{A}_\ell)\mathbf{M}_{\mathrm{Gl},\ell}^\nu \\ &= (\mathbf{A}_\ell^{-1} - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell)\mathbf{A}_\ell\mathbf{M}_{\mathrm{Gl},\ell}^\nu\end{aligned}$$

gegeben, ihre Norm lässt sich durch

$$\|\mathbf{M}_{\mathrm{ZGV},\ell}\|_2 \le \|\mathbf{A}_\ell^{-1} - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\|_2\|\mathbf{A}_\ell\mathbf{M}_{\mathrm{Gl},\ell}^\nu\|_2 \le \frac{C}{\|\mathbf{A}_\ell\|_2}\eta(\nu)\|\mathbf{A}_\ell\|_2 = C\eta(\nu)$$

abschätzen. Dank der Glättungseigenschaft finden wir ein $\nu \in \mathbb{N}$ so, dass $C\eta(\nu) \le \varrho$ gilt, und damit ist der Beweis abgeschlossen. ■

Wenden wir nun unsere Aufmerksamkeit dem Mehrgitterverfahren zu. Es unterscheidet sich vom Zweigitterverfahren im Wesentlichen dadurch, dass die Grobgitterkorrektur nur approximativ berechnet wird: Für

$$\mathbf{d}_\ell := \mathbf{A}_\ell\mathbf{x}_\ell - \mathbf{b}_\ell = \mathbf{A}_\ell(\mathbf{x}_\ell - \mathbf{x}_\ell^*)$$

setzen wir $\mathbf{b}_{\ell-1} := \mathbf{r}_\ell \mathbf{d}_\ell$ und bestimmen die approximative Grobgitterlösung $\mathbf{x}_{\ell-1}$ durch γ Schritte des Mehrgitterverfahrens auf Stufe $\ell - 1$ ausgehend von dem Startvektor $\mathbf{0}$.

Damit ist der Iterationsfehler auf dieser Stufe durch

$$\mathbf{x}_{\ell-1} - \mathbf{x}^*_{\ell-1} = \mathbf{M}^\gamma_{\mathrm{MGV},\ell-1}(\mathbf{0} - \mathbf{x}^*_{\ell-1})$$

gegeben und wir erhalten

$$\mathbf{x}_{\ell-1} = (\mathbf{I} - \mathbf{M}^\gamma_{\mathrm{MGV},\ell-1})\mathbf{x}^*_{\ell-1} = (\mathbf{I} - \mathbf{M}^\gamma_{\mathrm{MGV},\ell-1})\mathbf{A}^{-1}_{\ell-1}\mathbf{b}_{\ell-1}.$$

Die Iterationsmatrix der approximativen Grobgitterkorrektur besitzt demzufolge die Form

$$\widetilde{\mathbf{M}}_{\mathrm{GGK},\ell} = \mathbf{I} - \mathbf{p}_\ell(\mathbf{I} - \mathbf{M}^\gamma_{\mathrm{MGV},\ell-1})\mathbf{A}^{-1}_{\ell-1}\mathbf{r}_\ell\mathbf{A}_\ell = \mathbf{M}_{\mathrm{GGK},\ell} + \mathbf{p}_\ell\mathbf{M}^\gamma_{\mathrm{MGV},\ell-1}\mathbf{A}^{-1}_{\ell-1}\mathbf{r}_\ell\mathbf{A}_\ell,$$

kann also als Störung der exakten Grobgitterkorrektur interpretiert werden, die um so kleiner ausfällt, je schneller das Mehrgitterverfahren auf Stufe $\ell - 1$ konvergiert beziehungsweise je mehr Schritte dieses Verfahrens wir durchführen.

Um den Beweis des Mehrgitterverfahrens auf den des Zweigitterverfahrens zurückführen zu können, müssen wir die Störung abschätzen. Dazu benötigen wir zwei zusätzliche Voraussetzungen: Zunächst muss es eine Konstante $C_S \in \mathbb{R}_{\geq 0}$ so geben, dass

$$\|\mathbf{M}^\nu_{\mathrm{Gl},\ell}\|_2 \leq C_S \qquad \text{für alle } \nu \in \mathbb{N} \tag{4.8}$$

gilt. Diese Ungleichung ist beispielsweise schon erfüllt, wenn $\|\mathbf{M}_{\mathrm{Gl},\ell}\|_2 \leq 1$ gilt, wenn also die Glättungsiteration in der euklidischen Norm konvergiert.

Außerdem muss die Skalierung der Prolongation mit der euklidischen Norm verträglich sein, es muss also eine Konstante $C_P \in \mathbb{R}_{>0}$ so geben, dass

$$C_P^{-1}\|\mathbf{x}_{\ell-1}\|_2 \leq \|\mathbf{p}_\ell\mathbf{x}_{\ell-1}\|_2 \leq C_P\|\mathbf{x}_{\ell-1}\|_2 \qquad \text{für alle } \ell \in \mathbb{N}, \mathbf{x}_{\ell-1} \in \mathbb{K}^{\mathcal{I}_{\ell-1}} \tag{4.9}$$

gilt. Für den eindimensionalen Modellfall ist diese Ungleichung mit $C_P = 2$ erfüllt, für den zweidimensionalen Modellfall gilt sie mit $C_P = 4$.

Mit Hilfe der zusätzlichen Annahmen (4.8) und (4.9) können wir nun eine Konvergenzaussage für das Mehrgitterverfahren beweisen:

Satz 4.11 (Konvergenz Mehrgitterverfahren) *Für alle $\ell \in \mathbb{N}$ sei $\Phi_{\mathrm{Gl},\ell}$ ein lineares Iterationsverfahren mit der Glättungseigenschaft, dessen Iterationsmatrix die Bedingung (4.8) erfüllt. Die Hierarchie $(\mathbf{A}_\ell)_{\ell\in\mathbb{N}_0}$ besitze die Approximationseigenschaft und erfülle (4.9).*

Sei $\varrho \in (0,1)$. Dann gibt es ein $\nu \in \mathbb{N}$ so, dass das Mehrgitterverfahren mit $\nu \in \mathbb{N}$ Vorglättungsschritten und $\gamma \in \mathbb{N}_{\geq 2}$ konvergent ist mit

$$\|\mathbf{M}_{\mathrm{MGV},\ell}\|_2 \leq \varrho \qquad \textit{für alle } \ell \in \mathbb{N}.$$

Beweis. Zunächst müssen wir nachweisen, dass wir die durch die approximative Grobgitterkorrektur verursachte Störung beschränken können. Dank der zusätzlichen Annahmen (4.8) und (4.9) erhalten wir

$$\begin{aligned}\|\mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\mathbf{A}_\ell\mathbf{M}_{\mathrm{Gl},\ell}^\nu\|_2 &\le C_P\|\mathbf{p}_\ell\mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\mathbf{A}_\ell\mathbf{M}_{\mathrm{Gl},\ell}^\nu\|_2\\ &= C_P\|\mathbf{M}_{\mathrm{Gl},\ell}^\nu - (\mathbf{I}-\mathbf{p}_\ell\mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\mathbf{A}_\ell)\mathbf{M}_{\mathrm{Gl},\ell}^\nu\|_2\\ &\le C_P(C_S+\|\mathbf{M}_{\mathrm{ZGV},\ell}\|_2),\end{aligned}$$

so dass wir

$$\begin{aligned}\|\mathbf{M}_{\mathrm{MGV},\ell}\|_2 = \|\widetilde{\mathbf{M}}_{\mathrm{GGK},\ell}\mathbf{M}_{\mathrm{Gl},\ell}^\nu\|_2 &\le \|\mathbf{M}_{\mathrm{ZGV},\ell}\|_2 + \|\mathbf{p}_\ell\mathbf{M}_{\mathrm{MGV},\ell-1}^\gamma\mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\mathbf{A}_\ell\mathbf{M}_{\mathrm{Gl},\ell}^\nu\|_2\\ &\le \|\mathbf{M}_{\mathrm{ZGV},\ell}\|_2 + C_P\|\mathbf{M}_{\mathrm{MGV},\ell-1}\|_2^\gamma C_P(C_S+\|\mathbf{M}_{\mathrm{ZGV},\ell}\|_2)\\ &\le \|\mathbf{M}_{\mathrm{ZGV},\ell}\|_2 + C_P^2(C_S+\|\mathbf{M}_{\mathrm{ZGV},\ell}\|_2)\|\mathbf{M}_{\mathrm{MGV},\ell-1}\|_2^\gamma\end{aligned}$$

erhalten. Um den Faktor $C_P^2(C_S+1)$ kompensieren zu können, müssen wir sicher stellen, dass die Konvergenzrate auf der Gitterstufe $\ell-1$ hinreichend gut ist. Wir verwenden ein später noch genauer zu bestimmendes $\varrho_0 \in (0,\varrho] \subseteq (0,1)$, nehmen an, dass auf der nächstgröberen Stufe bereits

$$\|\mathbf{M}_{\mathrm{MGV},\ell-1}\|_2 \le \varrho_0$$

gilt, und wollen nun zeigen, dass

$$\begin{aligned}\|\mathbf{M}_{\mathrm{MGV},\ell}\|_2 &\le \|\mathbf{M}_{\mathrm{ZGV},\ell}\|_2 + C_P^2(C_S+\|\mathbf{M}_{\mathrm{ZGV},\ell}\|_2)\|\mathbf{M}_{\mathrm{MGV},\ell-1}\|_2^\gamma\\ &\le \|\mathbf{M}_{\mathrm{ZGV},\ell}\|_2 + C_P^2(C_S+\|\mathbf{M}_{\mathrm{ZGV},\ell}\|_2)\varrho_0^\gamma \le \varrho_0\end{aligned}$$

gilt. Nach Satz 4.10 können wir die Konvergenzrate des Zweigitterverfahrens beliebig nahe null wählen. Wir geben ein $\alpha \in (0,1)$ vor und wählen $\nu \in \mathbb{N}$ so, dass

$$\|\mathbf{M}_{\mathrm{ZGV},\ell}\|_2 \le \alpha\varrho_0$$

gilt. Wegen $\alpha\varrho_0 < 1$ folgt

$$\|\mathbf{M}_{\mathrm{MGV},\ell}\|_2 \le \alpha\varrho_0 + C_P^2(C_S+1)\varrho_0^\gamma = \big(\alpha + C_P^2(C_S+1)\varrho_0^{\gamma-1}\big)\varrho_0.$$

Also wählen wir

$$\varrho_0 := \min\left\{\varrho, \left(\frac{1-\alpha}{C_P^2(C_S+1)}\right)^{1/(\gamma-1)}\right\},$$

um

$$\|\mathbf{M}_{\mathrm{MGV},\ell}\|_2 \le (\alpha + C_P^2(C_S+1)\varrho_0^{\gamma-1})\varrho_0 \le \left(\alpha + C_P^2(C_S+1)\frac{1-\alpha}{C_P^2(C_S+1)}\right)\varrho_0 = \varrho_0$$

zu erhalten.

Der Rest ist ein einfacher Induktionsbeweis: Den Induktionsschritt haben wir bereits bewiesen. Der Induktionsanfang folgt aus $\|\mathbf{M}_{\mathrm{MGV},0}\|_2 = 0 \le \varrho_0$, da auf dem gröbsten Gitter exakt gelöst wird. ■

Bis auf die in der Praxis leicht zu erfüllenden zusätzlichen Bedingungen (4.8) und (4.9) genügen also für den Nachweis der Konvergenz des Mehrgitterverfahrens die Approximations- und Glättungseigenschaft, sofern mindestens der W-Zyklus, also $\gamma = 2$, verwendet wird.

4.5 Glättungs- und Approximationseigenschaft bei Finite-Elemente-Verfahren

Bevor wir die Konvergenz von Mehrgitterverfahren weiter untersuchen sollen kurz die Glättungs- und Approximationseigenschaft am Beispiel der Diskretisierung einer elliptischen partiellen Differentialgleichung mit einer Finite-Elemente-Methode illustriert werden. Wir beschränken uns dabei zur Vereinfachung auf den Fall eines reellen und symmetrischen Problems maximaler Regularität.

Wir gehen von der Variationsformulierung der partiellen Differentialgleichung aus: Für einen rellen Hilbertraum V bezeichnen wir den *Dualraum* mit

$$V' := \{\lambda : V \to \mathbb{R} \ : \ \lambda \text{ linear und stetig}\},$$

versehen mit der *Dualnorm*

$$\|\lambda\|_{V'} := \sup \left\{ \frac{|\lambda(v)|}{\|v\|_V} \ : \ v \in V \setminus \{0\} \right\}.$$

Die Elemente des Dualraums nennen wir *Funktionale*.

Sei $\lambda \in V'$ ein Funktional, und sei eine Bilinearform

$$a : V \times V \to \mathbb{R}$$

gegeben. Die Variationsaufgabe besteht darin, ein $u \in V$ mit

$$a(v, u) = \lambda(v) \qquad \text{für alle } v \in V \tag{4.10}$$

zu finden. Im Allgemeinen wird V unendlich-dimensional und damit für die Behandlung auf dem Computer unzugänglich sein. Die Idee des *Galerkin-Verfahrens* besteht darin, zur Approximation eine Hierarchie

$$V_0 \subseteq V_1 \subseteq \ldots \subseteq V$$

endlich-dimensionaler Teilräume einzuführen und nach Näherungslösungen $u_\ell \in V_\ell$ zu suchen, die die endlich-dimensionalen Variationsaufgaben

$$a(v_\ell, u_\ell) = \lambda(v_\ell) \qquad \text{für alle } v_\ell \in V_\ell \tag{4.11}$$

lösen. Üblicherweise werden diese Aufgaben gelöst, indem man für $\ell \in \mathbb{N}_0$ eine geeignete Basis $(\varphi_{\ell,i})_{i \in \mathcal{I}_\ell}$ fixiert und die Darstellung

$$u_\ell = \sum_{j \in \mathcal{I}_\ell} x_{\ell,j} \varphi_{\ell,j} \tag{4.12}$$

mit einem Koeffizientenvektor $\mathbf{x}_\ell \in \mathbb{R}^{\mathcal{I}_\ell}$ wählt, mit der die Gleichung (4.11) die Form

$$\sum_{j \in \mathcal{I}_\ell} x_{\ell,j} a(\varphi_{\ell,i}, \varphi_{\ell,j}) = \lambda(\varphi_{\ell,i}) \qquad \text{für alle } i \in \mathcal{I}_\ell$$

annimmt. Mit der Matrix $\mathbf{A}_\ell \in \mathbb{R}^{\mathcal{I}_\ell \times \mathcal{I}_\ell}$ und dem Vektor $\mathbf{b}_\ell \in \mathbb{R}^{\mathcal{I}_\ell}$, gegeben durch

$$A_{\ell,ij} := a(\varphi_{\ell,i}, \varphi_{\ell,j}), \qquad b_{\ell,i} := \lambda(\varphi_{\ell,i}) \qquad \text{für alle } i,j \in \mathcal{I}_\ell, \tag{4.13}$$

erhalten wir schließlich das lineare Gleichungssystem

$$\mathbf{A}_\ell \mathbf{x}_\ell = \mathbf{b}_\ell, \tag{4.14}$$

dessen Lösung $\mathbf{x}_\ell$ die diskrete Näherung u_ℓ mittels (4.12) definiert.

Um diese Gleichungssysteme mit einem Mehrgitterverfahren behandeln zu können, benötigen wir geeignete Prolongationen und Restriktionen. Dazu definieren wir für jedes $\ell \in \mathbb{N}_0$ den *Koeffizientenisomorphismus* $P_\ell : \mathbb{R}^{\mathcal{I}_\ell} \to V_\ell$ durch

$$P_\ell \mathbf{x}_\ell := \sum_{j \in \mathcal{I}_\ell} (\mathbf{x}_\ell)_j \varphi_{\ell,j} \qquad \text{für alle } \mathbf{x}_\ell \in \mathbb{R}^{\mathcal{I}_\ell}.$$

Für $\ell \in \mathbb{N}$ gilt $V_{\ell-1} \subseteq V_\ell$, also können wir

$$\mathbf{p}_\ell := P_\ell^{-1} P_{\ell-1}$$

definieren und erhalten

$$P_\ell \mathbf{p}_\ell = P_{\ell-1}, \tag{4.15}$$

also berechnet $\mathbf{p}_\ell$ gerade diejenigen Koeffizienten auf Stufe ℓ, die eine auf Stufe $\ell - 1$ gegebene Funktion *exakt* wiedergeben. In diesem Sinn ist $\mathbf{p}_\ell$ die bestmögliche Prolongation.

Bei der Konstruktion der Restriktion müssen wir beachten, dass die Vektoren $\mathbf{b}_\ell$ anders interpretiert werden als die Vektoren $\mathbf{x}_\ell$: Letztere beschreiben Funktionen, erstere dagegen Funktionale. Das Gegenstück des Koeffizientenisomorphismus P_ℓ auf Seiten der Funktionale ist der Operator

$$R_\ell : V' \to \mathbb{R}^{\mathcal{I}_\ell}, \qquad \mu \mapsto (\mu(\varphi_{\ell,i}))_{i \in \mathcal{I}_\ell},$$

der einem Funktional den korrespondierenden Vektor zuordnet. Für einen beliebigen Vektor $\mathbf{y}_\ell \in \mathbb{R}^{\mathcal{I}_\ell}$ und ein beliebiges Funktional $\mu \in V'$ gilt

$$\langle \mathbf{y}_\ell, R_\ell \mu \rangle_2 = \sum_{i \in \mathcal{I}_\ell} y_{\ell,i} \mu(\varphi_{\ell,i}) = \mu \left(\sum_{i \in \mathcal{I}_\ell} y_{\ell,i} \varphi_{\ell,i} \right) = \mu(P_\ell \mathbf{y}_\ell). \tag{4.16}$$

Für $\ell \in \mathbb{N}$ folgt aus dieser Gleichung

$$\langle \mathbf{y}_{\ell-1}, R_{\ell-1}\mu \rangle_2 = \mu(P_{\ell-1}\mathbf{y}_{\ell-1}) = \mu(P_\ell \mathbf{p}_\ell \mathbf{y}_{\ell-1}) = \langle \mathbf{p}_\ell \mathbf{y}_{\ell-1}, R_\ell \mu \rangle_2 = \langle \mathbf{y}_{\ell-1}, \mathbf{p}_\ell^* R_\ell \mu \rangle_2,$$

also mit $\mathbf{r}_\ell := \mathbf{p}_\ell^*$ gerade

$$\mathbf{r}_\ell R_\ell = R_{\ell-1}. \tag{4.17}$$

Damit ist $\mathbf{r}_\ell$ diejenige Matrix, die einem Funktional auf Stufe ℓ dasselbe Funktional auf Stufe $\ell - 1$ zuordnet, also in diesem Sinn die bestmögliche Restriktion.

Bemerkung 4.12 (Dualitätsprodukt) *Als Verallgemeinerung des bisher verwendeten Skalarprodukts verwendet man gelegentlich das* Dualitätsprodukt $\langle \cdot, \cdot \rangle$, *das durch*

$$\langle v, \mu \rangle_{V \times V'} := \mu(v) \qquad \text{für alle } v \in V,\ \mu \in V'$$

definiert ist. Die Gleichung (4.16) können wir dann als

$$\langle \mathbf{y}_\ell R_\ell \mu \rangle_2 = \langle P_\ell \mathbf{y}_\ell, \mu \rangle_{V \times V'} \qquad \text{für alle } \mathbf{y} \in \mathbb{R}^{\mathcal{I}_\ell},\ \mu \in V'$$

schreiben. In Anlehnung an Lemma 2.39 können wir R_ℓ dann als verallgemeinerte Adjungierte des Operators P_ℓ interpretieren und als $R_\ell = P_\ell^*$ schreiben.

Unsere Aufgabe besteht nun darin, die Eigenschaften der durch $\mathbf{A}_\ell$, $\mathbf{p}_\ell$ und $\mathbf{r}_\ell$ beschriebenen Hierarchie von Gleichungssystemen zu untersuchen.

Dafür benötigen wir einige zusätzliche Annahmen funktionalanalytischer Art. Der Raum V wird im Fall des Modellproblems aus Funktionen bestehen, die in einem geeigneten Sinn einmal differenzierbar sind. Um den Approximationsfehler in diesem Raum abschätzen zu können, müssen wir einen weiteren Raum $W \subseteq V$ einführen, unter dem wir uns den Raum der zweimal differenzierbaren Funktionen vorstellen können. Entsprechend brauchen wir auch einen Teilraum des Raums V' der Funktionale, den wir erhalten, indem wir einen Teilraum $U \subseteq V$ wählen und $V' \subseteq U'$ zur Kenntnis nehmen. Unter U können wir uns für die Zwecke unserer Untersuchung den Raum der stetigen Funktionen vorstellen.

Bemerkung 4.13 (Sobolew-Räume) *In der Praxis analysiert man Variationsaufgaben in der Regel in* Sobolew-Räumen *auf einem Gebiet $\Omega \subseteq \mathbb{R}^d$. Ausgehend von $H^0(\Omega) := L^2(\Omega)$ definiert man für jedes $m \in \mathbb{N}$ einen Raum $H^m(\Omega)$, der in einem geeigneten Sinn m-mal differenzierbare Funktionen enthält. Die in unserem Modellproblem verwendeten Null-Randbedingungen werden als Teilräume $H_0^m(\Omega)$ beschrieben.*

Für unsere Zwecke würde man dann

$$U = H^0(\Omega) = L^2(\Omega), \qquad V = H_0^m(\Omega), \qquad W = H^{2m}(\Omega)$$

verwenden und $2m$ als die Ordnung *des zu der Bilinearform a gehörenden Differentialoperators bezeichnen.*

Für unsere Analyse müssen sowohl die Variationsaufgabe geeignet gestellt als auch die für die Diskretisierung verwendeten Räume sinnvoll gewählt sein.

Zunächst müssen wir sicherstellen, dass die Variationsaufgabe (4.10) überhaupt eine Lösung besitzt. Dazu fordern wir, dass die Bilinearform *V-stetig*, *symmetrisch* und *V-elliptisch* ist, dass also

$$|a(v,u)| \le C_S \|v\|_V \|u\|_V \qquad \text{für alle } v, u \in V, \tag{4.18a}$$

$$a(v,u) = a(u,v) \qquad \text{für alle } v, u, \in V, \tag{4.18b}$$

$$C_E a(v,v) \ge \|v\|_V^2 \qquad \text{für alle } v \in V \tag{4.18c}$$

mit geeigneten Konstanten $C_S, C_E \in \mathbb{R}_{>0}$ gelten. Unter diesen Voraussetzungen besagt der *Satz von Lax-Milgram*, dass die Variationsaufgabe (4.10) eine eindeutige Lösung $u \in V$ besitzt. Um Approximationsaussagen formulieren zu können, genügt es nicht, wenn die Lösung lediglich in V liegt. Deshalb fordern wir die *volle Regularität* der Variationsaufgabe, dass also für jedes Funktional $\lambda \in U' \subseteq V'$ die Lösung $u \in V$ sogar in dem Teilraum W liegt und die Regularitätsabschätzung

$$\|u\|_W \leq C_R \|\lambda\|_{U'} \tag{4.19}$$

gilt. Diese Forderung gilt in der Praxis häufig nur, falls das Gebiet, auf dem wir rechnen, glatt berandet oder konvex ist, stellt also eine ernsthafte Einschränkung unserer Theorie dar, die wir im Interesse der Einfachheit der Analyse in Kauf nehmen.

Für jedes ℓ soll eine Zahl $h_\ell \in \mathbb{R}_{>0}$ gewählt sein, die in einem geeigneten Sinn die „Feinheit“ der Diskretisierung beschreibt, in unserem Modellproblem beispielsweise die Maschenweite des Gitters. Wir fordern zunächst, dass die U-Norm mit der euklidischen Norm in geeigneter Weise verträglich ist, dass nämlich die Abschätzungen

$$\|P_\ell \mathbf{x}_\ell\|_U \leq C_P h_\ell^{d/2} \|\mathbf{x}_\ell\|_2, \quad \|\mathbf{x}_\ell\|_2 \leq C_P h_\ell^{-d/2} \|P_\ell \mathbf{x}_\ell\|_U \qquad \text{für alle } \ell \in \mathbb{N}_0,\ \mathbf{x}_\ell \in \mathbb{R}^{\mathcal{I}_\ell} \tag{4.20}$$

mit einer geeigneten Konstanten $C_P \in \mathbb{R}_{>0}$ gelten.

Wir müssen auch fordern, dass sich Funktionen in dem Teilraum $W \subseteq V$ durch Funktionen in V_ℓ approximieren lassen, dass also

$$\min\{\|v - v_\ell\|_V \ : \ v_\ell \in V_\ell\} \leq C_B h_\ell^m \|v\|_W \qquad \text{für alle } \ell \in \mathbb{N}_0,\ v \in W \tag{4.21}$$

mit einer Konstanten $C_B \in \mathbb{R}_{>0}$ gilt.

Um (4.20) mit der Stetigkeit (4.18a) kombinieren zu können, benötigen wir die *inverse Ungleichung*: Wir können die V-Norm durch die schwächere U-Norm abschätzen, wenn wir dazu bereit sind, mit einer negativen Potenz des Diskretisierungsparameters h_ℓ zu „bezahlen“. Genauer gesagt sollen unsere Räume die Abschätzung

$$\|v_\ell\|_V \leq C_I h_\ell^{-m} \|v_\ell\|_U \qquad \text{für alle } \ell \in \mathbb{N}_0,\ v_\ell \in V_\ell \tag{4.22}$$

mit einer Konstanten $C_I \in \mathbb{R}_{>0}$ erfüllen.

Schließlich müssen wir sicherstellen, dass die Approximationseigenschaften unserer Räume sich nicht zu schnell verschlechtern, wenn wir von feineren zu gröberen Räumen wechseln. Mit Hilfe des Diskretisierungsparameters können wir diese Forderung als

$$h_{\ell-1} \leq C_h h_\ell \qquad \text{für alle } \ell \in \mathbb{N} \tag{4.23}$$

präzisieren. Damit sind alle nötigen Voraussetzungen formuliert und wir können uns dem Nachweis der Glättungs- und Approximationseigenschaft widmen.

Lemma 4.7 besagt, dass die Grobgitterkorrektur wünschenswerte Eigenschaften aufweist, falls die Matrizen der Gitterhierarchie die Galerkin-Eigenschaft aufweisen. Es liegt nahe, zu erwarten, dass das Galerkin-Verfahren zu Matrizen führt, die diese Eigenschaft besitzen. Und so ist es auch:

Lemma 4.14 (Galerkin-Eigenschaft) *Die durch (4.13) gegebenen Matrizen* $(\mathbf{A}_\ell)_{\ell=0}^\infty$ *erfüllen*

$$\mathbf{A}_{\ell-1} = \mathbf{r}_\ell \mathbf{A}_\ell \mathbf{p}_\ell \qquad \text{für alle } \ell \in \mathbb{N},$$

so dass unsere Gitterhierarchie die Galerkin-Eigenschaft besitzt.

Beweis. Seien $\ell \in \mathbb{N}$ und $\mathbf{y}_{\ell-1}, \mathbf{z}_{\ell-1} \in \mathbb{R}^{\mathcal{I}_{\ell-1}}$ gegeben. Aus (4.15) folgt

$$\begin{aligned}\langle \mathbf{A}_{\ell-1}\mathbf{y}_{\ell-1}, \mathbf{z}_{\ell-1}\rangle_2 &= a(P_{\ell-1}\mathbf{y}_{\ell-1}, P_{\ell-1}\mathbf{z}_{\ell-1}) = a(P_\ell \mathbf{p}_\ell \mathbf{y}_{\ell-1}, P_\ell \mathbf{p}_\ell \mathbf{z}_{\ell-1}) \\ &= \langle \mathbf{A}_\ell \mathbf{p}_\ell \mathbf{y}_{\ell-1}, \mathbf{p}_\ell \mathbf{z}_{\ell-1}\rangle_2 = \langle \mathbf{p}_\ell^* \mathbf{A}_\ell \mathbf{p}_\ell \mathbf{y}_{\ell-1}, \mathbf{z}_{\ell-1}\rangle_2 \\ &= \langle \mathbf{r}_\ell \mathbf{A}_\ell \mathbf{p}_\ell \mathbf{y}_{\ell-1}, \mathbf{z}_{\ell-1}\rangle_2.\end{aligned}$$

Da wir diese Gleichung für beliebige Vektoren $\mathbf{y}_{\ell-1}$ und $\mathbf{z}_{\ell-1}$ bewiesen haben, folgt $\mathbf{A}_{\ell-1} = \mathbf{r}_\ell \mathbf{A}_\ell \mathbf{p}_\ell$. ■

Wir haben bereits gesehen, dass das Richardson-Verfahren die Glättungseigenschaft besitzt, falls wir den Dämpfungsparameter geeignet wählen. Um das praktisch tun zu können benötigen wir eine Abschätzung der Norm $\|\mathbf{A}_\ell\|_2$.

Lemma 4.15 (Norm von $\mathbf{A}_\ell$) *Es gilt*

$$\|\mathbf{A}_\ell\|_2 \le C_S C_I^2 C_P^2 h_\ell^{d-2m} \qquad \text{für alle } \ell \in \mathbb{N}_0.$$

Beweis. Sei $\ell \in \mathbb{N}_0$. Sei $\mathbf{x}_\ell \in \mathbb{R}^{\mathcal{I}_\ell}$ und $\mathbf{y}_\ell := \mathbf{A}_\ell \mathbf{x}_\ell$. Dann gilt

$$\begin{aligned}\|\mathbf{A}_\ell \mathbf{x}_\ell\|_2^2 &= \langle \mathbf{y}_\ell, \mathbf{A}_\ell \mathbf{x}_\ell\rangle_2 = \sum_{i\in\mathcal{I}_\ell}\sum_{j\in\mathcal{I}_\ell} (\mathbf{y}_\ell)_i (\mathbf{A}_\ell)_{ij} (\mathbf{x}_\ell)_j = \sum_{i\in\mathcal{I}_\ell}\sum_{j\in\mathcal{I}_\ell} (\mathbf{y}_\ell)_i a(\varphi_{\ell,i}, \varphi_{\ell,j})(\mathbf{x}_\ell)_j \\ &= a\left(\sum_{i\in\mathcal{I}_\ell} \varphi_{\ell,i}(\mathbf{y}_\ell)_i, \sum_{j\in\mathcal{I}_\ell} \varphi_{\ell,j}(\mathbf{x}_\ell)_j\right) = a(P_\ell \mathbf{y}_\ell, P_\ell \mathbf{x}_\ell) \le C_S \|P_\ell \mathbf{y}_\ell\|_V \|P_\ell \mathbf{x}_\ell\|_V \\ &\le C_S C_I^2 h_\ell^{-2m} \|P_\ell \mathbf{y}_\ell\|_U \|P_\ell \mathbf{x}_\ell\|_U \le C_S C_I^2 C_P^2 h_\ell^{d-2m} \|\mathbf{y}_\ell\|_2 \|\mathbf{x}_\ell\|_2 \\ &= C_S C_I^2 C_P^2 h_\ell^{d-2m} \|\mathbf{A}_\ell \mathbf{x}_\ell\|_2 \|\mathbf{x}_\ell\|_2,\end{aligned}$$

also insbesondere

$$\|\mathbf{A}_\ell \mathbf{x}_\ell\|_2 \le C_S C_I^2 C_P^2 h_\ell^{d-2m} \|\mathbf{x}_\ell\|_2.$$

Da $\mathbf{x}_\ell$ beliebig gewählt war, folgt daraus die gewünschte Abschätzung. ■

Als nächstes wenden wir uns dem Nachweis der Approximationseigenschaft zu. Sie ergibt sich direkt aus der folgenden Abschätzung für den Diskretisierungsfehler des Finite-Elemente-Verfahrens.

Lemma 4.16 (Fehlerabschätzung) *Sei $f \in U'$, $u \in V$ die korrespondierende Lösung von (4.10), und $u_\ell \in V_\ell$ die entsprechende Lösung von (4.11). Dann gilt*

$$\|u - u_\ell\|_U \le C_N h_\ell^{2m} \|\lambda\|_{U'}, \qquad C_N := C_E C_S^2 C_B^2 C_R^2.$$

Beweis. Sei $\ell \in \mathbb{N}_0$. Aus der *Galerkin-Orthogonalitätsbeziehung*

$$a(u - u_\ell, v_\ell) = a(u, v_\ell) - a(u_\ell, v_\ell) = \lambda(v_\ell) - \lambda(v_\ell) = 0 \qquad \text{für alle } v_\ell \in V_\ell$$

folgt in Verbindung mit (4.18c) und (4.18a)

$$\begin{aligned} \|u - u_\ell\|_V^2 &\le C_E a(u - u_\ell, u - u_\ell) = C_E a(u - u_\ell, u - v_\ell) \\ &\le C_E C_S \|u - u_\ell\|_V \|u - v_\ell\|_V \qquad \text{für alle } v_\ell \in V_\ell. \end{aligned}$$

Indem wir (4.19) und (4.21) kombinieren, erhalten wir

$$\min\{\|u - v_\ell\|_V \ : \ v_\ell \in V_\ell\} \le C_B C_R h_\ell^m \|\lambda\|_{U'},$$

also insbesondere auch

$$\|u - u_\ell\|_V \le C_E C_S C_B C_R h_\ell^m \|\lambda\|_{U'}.$$

Jetzt verwenden wir den sogenannten *Nitsche-Trick*: Wir suchen die Lösung $w \in V$ des Variationsproblems

$$a(v, w) = \langle v, u - u_\ell \rangle_U \qquad \text{für alle } v \in V.$$

Die rechte Seite liegt offenbar in U', und ihre Operatornorm ist dank der Cauchy-Schwarz-Ungleichung durch $\|u - u_\ell\|_U$ beschränkt, also können wir eine Funktion $w_\ell \in V_\ell$ finden, die

$$\|w - w_\ell\|_V \le C_B C_R h_\ell^m \|u - u_\ell\|_U$$

erfüllt. Indem wir wieder die Galerkin-Orthogonalität einsetzen, erhalten wir schließlich

$$\begin{aligned} \|u - u_\ell\|_U^2 &= \langle u - u_\ell, u - u_\ell \rangle_U = a(u - u_\ell, w) = a(w, u - u_\ell) = a(w - w_\ell, u - u_\ell) \\ &\le C_S \|w - w_\ell\|_V \|u - u_\ell\|_V \le C_S C_B C_R h_\ell^m \|u - u_\ell\|_U C_E C_S C_B C_R h_\ell^m \|\lambda\|_{U'} \\ &= C_N h_\ell^{2m} \|\lambda\|_{U'} \|u - u_\ell\|_U. \end{aligned}$$

Indem wir gegebenenfalls durch $\|u - u_\ell\|_U$ dividieren erhalten wir die gesuchte Abschätzung. ■

Unser Ziel ist es, aus dieser Fehlerabschätzung eine Approximationsaussage für die Matrizen $\mathbf{A}_\ell$ zu gewinnen. Dazu setzen wir die Matrizen in Bezug zu dem Lösungsoperator des kontinuierlichen Problems.

Wir definieren den Operator $A : V \to V'$ durch

$$Au := a(\cdot, u) \qquad \text{für alle } u \in V.$$

Die Stetigkeit der Bilinearform a impliziert die Stetigkeit von A mit

$$\|A\|_{V' \leftarrow V} = \sup \left\{ \frac{a(v, u)}{\|v\|_V \|u\|_V} \ : \ v, u \in V \setminus \{0\} \right\}$$

$$\leq \sup\left\{\frac{C_S\|v\|_V\|u\|_V}{\|v\|_V\|u\|_V} \;:\; v,u \in V\setminus\{0\}\right\} \leq C_S,$$

die Elliptizität von a impliziert die Invertierbarkeit von A mit

$$\begin{aligned}
\|A^{-1}\|_{V\leftarrow V'} &= \sup\left\{\frac{\|A^{-1}\mu\|_V^2}{\|\mu\|_{V'}\|A^{-1}\mu\|_V} \;:\; \mu \in V'\setminus\{0\}\right\}\\
&\leq \sup\left\{\frac{C_E a(A^{-1}\mu, A^{-1}\mu)}{\|\mu\|_{V'}\|A^{-1}\mu\|} \;:\; \mu \in V'\setminus\{0\}\right\}\\
&= \sup\left\{\frac{C_E \mu(A^{-1}\mu)}{\|\mu\|_{V'}\|A^{-1}\mu\|} \;:\; \mu \in V'\setminus\{0\}\right\}\\
&= \sup\left\{\frac{C_E\|\mu\|_{V'}\|A^{-1}\mu\|_V}{\|\mu\|_{V'}\|A^{-1}\mu\|} \;:\; \mu \in V'\setminus\{0\}\right\} \leq C_E.
\end{aligned}$$

Um exakte und diskrete Lösungen miteinander vergleichen zu können, bietet es sich an, den diskreten Lösungsoperator als Abbildung von V' nach V darzustellen. Nach Definition der Operatoren $R_\ell : V' \to \mathbb{R}^{\mathcal{I}_\ell}$ und $P_\ell : \mathbb{R}^{\mathcal{I}_\ell} \to V$ gelten

$$\mathbf{b}_\ell = R_\ell\lambda, \qquad u_\ell = P_\ell\mathbf{x}_\ell \qquad \text{für alle } \ell \in \mathbb{N}_0,$$

so dass wir mit $\mathbf{A}_\ell\mathbf{x}_\ell = \mathbf{b}_\ell$ die Gleichung

$$u_\ell = P_\ell\mathbf{A}_\ell^{-1}R_\ell\lambda \qquad \text{für alle } \ell \in \mathbb{N}_0$$

erhalten. Die Abschätzung aus Lemma 4.16 nimmt damit die Form

$$\|(A^{-1} - P_\ell\mathbf{A}_\ell^{-1}R_\ell)\lambda\|_U = \|u - u_\ell\|_U \leq C_N h_\ell^{2m}\|\lambda\|_{U'} \qquad \text{für alle } \ell \in \mathbb{N}_0$$

an und lässt sich kurz als

$$\|A^{-1} - P_\ell\mathbf{A}_\ell^{-1}R_\ell\|_{U\leftarrow U'} \leq C_N h_\ell^{2m} \qquad \text{für alle } \ell \in \mathbb{N}_0 \tag{4.24}$$

schreiben. Diese Abschätzung erinnert bereits sehr an die Approximationseigenschaft, die wir für die Analyse des Mehrgitterverfahrens benötigen, allerdings brauchen wir eine Abschätzung, die die Räume V_ℓ und $V_{\ell-1}$ zueinander Beziehung setzt, nicht V und V_ℓ. Mit Hilfe der Gleichungen (4.15) und (4.17) erhalten wir

$$\begin{aligned}
P_\ell(\mathbf{A}_\ell^{-1} - p_\ell\mathbf{A}_{\ell-1}^{-1}r_\ell)R_\ell &= P_\ell\mathbf{A}_\ell^{-1}R_\ell - P_{\ell-1}\mathbf{A}_{\ell-1}^{-1}R_{\ell-1}\\
&= (P_\ell\mathbf{A}_\ell^{-1}R_\ell - A^{-1}) + (A^{-1} - P_{\ell-1}\mathbf{A}_{\ell-1}^{-1}R_{\ell-1}),
\end{aligned}$$

und die letzten beiden Terme können wir mit (4.24) abschätzen. Um daraus die Approximationseigenschaft zu gewinnen, benötigen wir das folgende Lemma.

Lemma 4.17 (R_ℓ surjektiv) *Sei $\ell \in \mathbb{N}_0$ und $\mathbf{c}_\ell \in \mathbb{R}^{\mathcal{I}_\ell}$. Dann existiert ein Funktional $\mu \in U'$ mit*

$$R_\ell\mu = \mathbf{c}_\ell, \qquad \|\mu\|_{U'} \leq C_P h_\ell^{-d/2}\|\mathbf{c}_\ell\|_2.$$

Beweis. Wir suchen nach einem Funktional der Form

$$\mu = \sum_{j \in \mathcal{I}_\ell} \langle \cdot, \varphi_{\ell,j} \rangle_U y_{\ell,j}$$

mit einem geeignet gewählten Vektor $\mathbf{y}_\ell \in \mathbb{R}^{\mathcal{I}_\ell}$. Damit unsere Gleichung erfüllt ist, müsste

$$c_{\ell,i} = R_{\ell,i}\mu = \sum_{j \in \mathcal{I}_\ell} \langle \varphi_{\ell,i}, \varphi_{\ell,j} \rangle_U y_{\ell,j} \qquad \text{für alle } i \in \mathcal{I}_\ell$$

gelten. Das ist offenbar ein lineares Gleichungssystem, dass wir mit der durch

$$m_{\ell,ij} := \langle \varphi_{\ell,i}, \varphi_{\ell,j} \rangle_U \qquad \text{für alle } i, j \in \mathcal{I}_\ell$$

definierten *Massematrix* $\mathbf{M}_\ell \in \mathbb{R}^{\mathcal{I}_\ell \times \mathcal{I}_\ell}$ in der kompakten Form

$$\mathbf{c}_\ell = \mathbf{M}_\ell \mathbf{y}_\ell$$

schreiben können. Also brauchen wir lediglich die Lösbarkeit dieser Gleichung zu untersuchen. Glücklicherweise gilt

$$\begin{aligned} \langle \mathbf{M}_\ell \mathbf{z}_\ell, \mathbf{z}_\ell \rangle_2 &= \sum_{i \in \mathcal{I}_\ell} \sum_{j \in \mathcal{I}_\ell} m_{\ell,ij} z_{\ell,i} z_{\ell,j} = \sum_{i \in \mathcal{I}_\ell} \sum_{j \in \mathcal{I}_\ell} \langle \varphi_{i,\ell}, \varphi_{j,\ell} \rangle_U \\ &= \langle P_\ell \mathbf{z}_\ell, P_\ell \mathbf{z}_\ell \rangle_U = \|P_\ell \mathbf{z}_\ell\|_U^2 \geq C_P^{-2} h_\ell^d \|\mathbf{z}_\ell\|_2^2 \qquad \text{für alle } \mathbf{z}_\ell \in \mathbb{R}^{\mathcal{I}_\ell}, \end{aligned}$$

also ist $\mathbf{M}_\ell$ positiv definit und damit insbesondere invertierbar. Indem wir $\mathbf{y}_\ell := \mathbf{M}_\ell^{-1} \mathbf{c}_\ell$ einsetzen, erhalten wir

$$\begin{aligned} \|\mathbf{c}_\ell\|_2 \|P_\ell \mathbf{y}_\ell\|_U &\geq \|\mathbf{c}_\ell\|_2 C_P^{-1} \|\mathbf{y}_\ell\|_2 \geq C_P^{-1} h_\ell^{d/2} \langle \mathbf{c}_\ell, \mathbf{y}_\ell \rangle_2 \\ &= C_P^{-1} h_\ell^{d/2} \langle \mathbf{M}_\ell \mathbf{y}_\ell, \mathbf{y}_\ell \rangle_2 = C_P^{-1} h_\ell^{d/2} \|P_\ell \mathbf{y}_\ell\|_U^2, \end{aligned}$$

also insbesonder $\|P_\ell \mathbf{y}_\ell\|_U \leq C_P h_\ell^{-d/2} \|\mathbf{c}\|_2$. Für ein beliebiges $v \in U$ gilt

$$|\mu(v)| = \left| \sum_{j \in \mathcal{I}_\ell} \langle v, \varphi_{\ell,j} y_{\ell,j} \rangle_U \right| = |\langle v, P_\ell \mathbf{y}_\ell \rangle_U| \leq \|v\|_U \|P_\ell \mathbf{y}_\ell\|_U \leq \|v\|_U C_P h_\ell^{-d/2} \|\mathbf{c}_\ell\|_2,$$

so dass wir auch die Normabschätzung bewiesen haben. ∎

Als unmittelbare Konsequenz ergibt sich die folgende Möglichkeit, die für uns relevante Norm abzuschätzen:

Lemma 4.18 *Sei $\ell \in \mathbb{N}_0$ und $\mathbf{X}_\ell \in \mathbb{R}^{\mathcal{I}_\ell \times \mathcal{I}_\ell}$. Dann gilt*

$$\|\mathbf{X}_\ell\|_2 \leq C_P^2 h_\ell^{-d} \|P_\ell \mathbf{X}_\ell R_\ell\|_{U \leftarrow U'}.$$

Beweis. Seien $\mathbf{y}_\ell, \mathbf{z}_\ell \in \mathbb{R}^{\mathcal{I}_\ell}$ gegeben. Nach Lemma 4.17 finden wir $\lambda, \mu \in V'$ mit

$$R_\ell \lambda = \mathbf{y}_\ell, \qquad R_\ell \mu = \mathbf{z}_\ell, \qquad \|\lambda\|_{U'} \le C_P h_\ell^{-d/2} \|\mathbf{y}\|_2, \qquad \|\mu\|_{U'} \le C_P h_\ell^{-d/2} \|\mathbf{z}\|_2.$$

Mit (4.16) folgt daraus

$$\begin{aligned}
|\langle \mathbf{X}_\ell \mathbf{y}_\ell, \mathbf{z}_\ell \rangle_2| &= |\langle \mathbf{X}_\ell R_\ell \lambda, R_\ell \mu \rangle_2| = |\mu(P_\ell \mathbf{X}_\ell R_\ell \lambda)| \le \|\mu\|_{U'} \|P_\ell \mathbf{X}_\ell R_\ell \lambda\|_U \\
&\le \|\mu\|_{U'} \|P_\ell \mathbf{X}_\ell R_\ell\|_{U \leftarrow U'} \|\lambda\|_{U'} \\
&\le C_P h_\ell^{-d/2} \|\mathbf{z}\|_2 \|P_\ell \mathbf{X}_\ell R_\ell\|_{U \leftarrow U'} C_P h_\ell^{-d/2} \|\mathbf{y}\|_2 \\
&= C_P^2 h_\ell^{-d} \|P_\ell \mathbf{X}_\ell R_\ell\|_{U \leftarrow U'} \|\mathbf{y}_\ell\|_2 \|\mathbf{z}_\ell\|_2,
\end{aligned}$$

also die Behauptung. ■

Mit Hilfe dieses Lemmas können wir nun die Approximationseigenschaft nachweisen:

Lemma 4.19 (Approximationseigenschaft für FEM) *Es gilt*

$$\begin{aligned}
\|\mathbf{A}_\ell^{-1} - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1} \mathbf{r}_\ell\|_2 &\le C_N C_P^2 (1 + C_h^{2m}) h_\ell^{2m-d} \\
&\le \frac{C_S C_N C_I C_P^4 (1 + C_h^{2m})}{\|\mathbf{A}_\ell\|_2} \qquad \textit{für alle } \ell \in \mathbb{N}.
\end{aligned}$$

Beweis. Sei $\ell \in \mathbb{N}$. Mit Hilfe von Lemma 4.18 und der Abschätzung (4.24) erhalten wir

$$\begin{aligned}
\|\mathbf{A}_\ell^{-1} - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1} \mathbf{r}_\ell\|_2 &\le C_P^2 h_\ell^{-d} \|P_\ell (\mathbf{A}_\ell^{-1} - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1} \mathbf{r}_\ell) R_\ell\|_{U \leftarrow U'} \\
&= C_P^2 h_\ell^{-d} \|P_\ell \mathbf{A}_\ell^{-1} R_\ell - A^{-1} + A^{-1} - P_{\ell-1} \mathbf{A}_{\ell-1}^{-1} R_{\ell-1}\|_{U \leftarrow U'} \\
&\le C_P^2 h_\ell^{-d} (C_N h_\ell^{2m} + C_N h_{\ell-1}^{2m}) \le C_N C_P^2 h_\ell^{-d} (h_\ell^{2m} + C_h^{2m} h_\ell^{2m}) \\
&= C_N C_P^2 (1 + C_h^{2m}) h_\ell^{2m-d},
\end{aligned}$$

also der erste Teil der gewünschten Aussage.

Mit Hilfe von Lemma 4.15 erhalten wir

$$\frac{1}{\|\mathbf{A}_\ell\|_2} \ge \frac{h_\ell^{2m-d}}{C_S C_I C_P^2}$$

und damit auch den zweiten Teil der Abschätzung. ■

Bemerkung 4.20 (Praktische Durchführung) *Bei der praktischen Durchführung des Mehrgitterverfahrens steht uns im Allgemeinen $\|\mathbf{A}_\ell\|_2$ nicht zur Verfügung, so dass wir nicht, wie im Beispiel, $\theta_\ell := 1/\|\mathbf{A}_\ell\|_2$ als Dämpfungsparameter verwenden können.*

Dank Lemma 4.15 können wir davon ausgehen, dass auch

$$\theta_\ell := \frac{h_\ell^{2m-d}}{C_S C_I C_P^2}$$

zu einer hinreichend starken Reduktion der niedrigfrequenten Anteile des Fehlers führen wird, und die Abschätzung (4.7) nimmt bei dieser Wahl die Form

$$\|\mathbf{A}_\ell \mathbf{M}_{\mathrm{Rich},\theta_\ell}^\nu\|_2 \leq \eta(\nu) C_S C_I C_P^2 h_\ell^{d-2m} \qquad \textit{für alle } \ell \in \mathbb{N}_0, \nu \in \mathbb{N}$$

an, die schwächer als die ursprüngliche Aussage ist.

Da wir dank Lemma 4.19 die Aussage

$$\|\mathbf{A}_\ell^{-1} - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1} \mathbf{r}_\ell\|_2 \leq C_N C_P^2 (1 + C_h^{2m}) h_\ell^{2m-d} \qquad \textit{für alle } \ell \in \mathbb{N}_0$$

zur Verfügung haben, können wir beispielsweise Satz 4.10 auch mit h_ℓ^{2m-d} *statt* $\|\mathbf{A}_\ell\|_2$ *beweisen und erhalten Abschätzungen der Form*

$$\|\mathbf{M}_{\mathrm{ZGV},\ell}\|_2 \leq C_S C_I C_N C_P^4 (1 + C_h^{2m}) \eta(\nu) \qquad \textit{für alle } \ell \in \mathbb{N}, \nu \in \mathbb{N},$$

wir haben also auch bei einer praktisch durchführbaren Wahl des Dämpfungsparameters nichts verloren.

4.6 Symmetrische Mehrgitterverfahren

Die Konvergenzaussage von Satz 4.11 gilt nur für den Fall $\gamma > 1$, also insbesondere nicht für V-Zyklus-Mehrgitterverfahren, die sich in der Praxis allerdings durch besonders hohe Effizienz auszeichnen. Unser Ziel ist es also nun, eine Konvergenzaussage für den Fall $\gamma = 1$ zu finden. Dazu bietet es sich an, das Mehrgitterverfahren *symmetrisch* zu gestalten, so dass sich die üblichen Beweistechniken für selbstadjungierte Matrizen anwenden lassen.

Wir benötigen die folgenden Voraussetzungen:

- Die Matrizen $\mathbf{A}_\ell$ sind für alle $\ell \in \mathbb{N}_0$ positiv definit.
- Es gibt eine Konstante $c \in \mathbb{R}_{>0}$ so, dass $\mathbf{r}_\ell = c\mathbf{p}_\ell^*$ für alle $\ell \in \mathbb{N}$ gilt.
- Für jedes $\ell \in \mathbb{N}$ gibt es eine positiv definite Matrix $\mathbf{W}_\ell \in \mathbb{K}^{\mathcal{I}_\ell \times \mathcal{I}_\ell}$ so, dass der Glätter in der dritten Normalform

  $$\Phi_{\mathrm{Gl},\ell}(\mathbf{x}_\ell, \mathbf{b}_\ell) = \mathbf{x}_\ell - \mathbf{W}_\ell^{-1}(\mathbf{A}_\ell \mathbf{x}_\ell - \mathbf{b}_\ell) \qquad \text{für alle } \mathbf{x}_\ell, \mathbf{b}_\ell \in \mathbb{K}^{\mathcal{I}_\ell}$$

  gegeben ist.
- Die Hierarchie der Gleichungssysteme besitzt die Galerkin-Eigenschaft.

Die meisten dieser Voraussetzungen sind in der Praxis erfüllt, sofern die den Gleichungssystemen zugrundeliegende Differentialgleichung symmetrisch und koerziv ist.

Für die Glättungseigenschaft fordern wir

$$\mathbf{A}_\ell \leq \mathbf{W}_\ell \qquad \text{für alle } \ell \in \mathbb{N}. \tag{4.25}$$

Für die Approximationseigenschaft wählen wir nun eine Variante, die an das Glättungsverfahren gekoppelt ist: Wir setzen voraus, dass eine Konstante $C_A \in \mathbb{R}_{>0}$ so existiert, dass

$$\|\mathbf{W}_\ell^{1/2}(\mathbf{A}_\ell^{-1} - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell)\mathbf{W}_\ell^{1/2}\|_2 \le C_A \qquad \text{für alle } \ell \in \mathbb{N} \tag{4.26}$$

gilt. Wegen

$$\begin{aligned}\|\mathbf{W}_\ell^{1/2}(\mathbf{A}_\ell^{-1} - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell)\mathbf{W}_\ell^{1/2}\|_2 &\le \|\mathbf{W}_\ell^{1/2}\|_2^2 \|\mathbf{A}_\ell^{-1} - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\|_2 \\ &= \|\mathbf{W}_\ell\|_2 \|\mathbf{A}_\ell^{-1} - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\|_2\end{aligned}$$

folgt die Approximationseigenschaft (4.26) aus der in Definition 4.9 eingeführten, falls $\|\mathbf{W}_\ell\|_2 \le C_W \|\mathbf{A}_\ell\|_2$ für eine von ℓ unabhängige Konstante $C_W \in \mathbb{R}_{>0}$ gilt. Falls etwa das Richardson-Verfahren als Glätter verwendet wird, ist diese letzte Abschätzung sogar mit $C_W = 1$ erfüllt.

Wir untersuchen die Mehrgitteriteration mit $\nu_1 = \nu_2 = \nu \in \mathbb{N}$, es werden also soviele Vor- wie Nachglättungsschritte durchgeführt. Unser Ziel ist es, die Energienorm der Iterationsmatrizen abzuschätzen.

Für das Zweigitterverfahren ist die Iterationsmatrix durch

$$\mathbf{M}_{\mathrm{ZGV},\ell} = \mathbf{M}_{\mathrm{Gl},\ell}^\nu (\mathbf{I} - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell \mathbf{A}_\ell)\mathbf{M}_{\mathrm{Gl},\ell}^\nu$$

gegeben, und die Energienorm lässt sich in der Form

$$\|\mathbf{M}_{\mathrm{ZGV},\ell}\|_A = \|\mathbf{A}_\ell^{1/2}\mathbf{M}_{\mathrm{ZGV},\ell}\mathbf{A}_\ell^{-1/2}\|_2$$

schreiben. Deshalb ist es naheliegend, auch die Matrizen der Prolongation, des Restriktion und des Glättungsoperators mit der Wurzel von $\mathbf{A}_\ell$ zu transformieren: Wir setzen

$$\begin{aligned}\widehat{\mathbf{p}}_\ell &= \mathbf{A}_\ell^{1/2}\mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1/2}, & \widehat{\mathbf{r}}_\ell &= \mathbf{A}_{\ell-1}^{-1/2}\mathbf{r}_\ell \mathbf{A}_\ell^{1/2}, \\ \widehat{\mathbf{W}}_\ell &= \mathbf{A}_\ell^{-1/2}\mathbf{W}_\ell \mathbf{A}_\ell^{-1/2}, & \widehat{\mathbf{M}}_{\mathrm{Gl},\ell} &= \mathbf{A}_\ell^{1/2}\mathbf{M}_{\mathrm{Gl},\ell}\mathbf{A}_\ell^{-1/2} = \mathbf{I} - \widehat{\mathbf{W}}_\ell^{-1}\end{aligned}$$

und erhalten die Gleichung

$$\mathbf{A}_\ell^{1/2}\mathbf{M}_{\mathrm{ZGV},\ell}\mathbf{A}_\ell^{-1/2} = \widehat{\mathbf{M}}_{\mathrm{Gl},\ell}^\nu (\mathbf{I} - \widehat{\mathbf{p}}_\ell \widehat{\mathbf{r}}_\ell)\widehat{\mathbf{M}}_{\mathrm{Gl},\ell}^\nu.$$

Die transformierte Grobgitterkorrektur bezeichnen wir mit

$$\widehat{\mathbf{Q}}_\ell := \mathbf{I} - \widehat{\mathbf{p}}_\ell \widehat{\mathbf{r}}_\ell = \mathbf{A}_\ell^{1/2}(\mathbf{I} - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell \mathbf{A}_\ell)\mathbf{A}_\ell^{-1/2}$$

und stellen fest, dass sie wegen der Galerkin-Eigenschaft die Gleichungen

$$\begin{aligned}\widehat{\mathbf{r}}_\ell \widehat{\mathbf{p}}_\ell &= \mathbf{A}_{\ell-1}^{-1/2}\mathbf{r}_\ell \mathbf{A}_\ell^{1/2}\mathbf{A}_\ell^{1/2}\mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1/2} = \mathbf{A}_{\ell-1}^{-1/2}\mathbf{r}_\ell \mathbf{A}_\ell \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1/2} = \mathbf{A}_{\ell-1}^{-1/2}\mathbf{A}_{\ell-1}\mathbf{A}_{\ell-1}^{-1/2} = \mathbf{I}, \\ \widehat{\mathbf{Q}}_\ell^2 &= (\mathbf{I} - \widehat{\mathbf{p}}_\ell \widehat{\mathbf{r}}_\ell)^2 = \mathbf{I} - 2\widehat{\mathbf{p}}_\ell \widehat{\mathbf{r}}_\ell + \widehat{\mathbf{p}}_\ell \widehat{\mathbf{r}}_\ell \widehat{\mathbf{p}}_\ell \widehat{\mathbf{r}}_\ell = \mathbf{I} - 2\widehat{\mathbf{p}}_\ell \widehat{\mathbf{r}}_\ell + \widehat{\mathbf{p}}_\ell \widehat{\mathbf{r}}_\ell = \mathbf{I} - \widehat{\mathbf{p}}_\ell \widehat{\mathbf{r}}_\ell = \widehat{\mathbf{Q}}_\ell\end{aligned}$$

erfüllt, also eine Projektion ist. Da $\widehat{\mathbf{Q}}_\ell$ außerdem selbstadjungiert ist, ist es sogar eine orthogonale Projektion, erfüllt also insbesondere die Ungleichung

$$\mathbf{0} \leq \widehat{\mathbf{Q}}_\ell \leq \mathbf{I}. \tag{4.27}$$

Indem wir die die Glättungseigenschaft definierende Ungleichung (4.25) von links und rechts mit $\mathbf{A}_\ell^{-1/2}$ multiplizieren, erhalten wir

$$\mathbf{I} \leq \mathbf{A}_\ell^{-1/2}\mathbf{W}_\ell\mathbf{A}_\ell^{-1/2} = \widehat{\mathbf{W}}_\ell,$$

und indem wir beide Seiten invertieren folgt

$$\mathbf{0} \leq \widehat{\mathbf{W}}_\ell^{-1} \leq \mathbf{I}. \tag{4.28}$$

Aus der Approximationseigenschaft (4.26) folgt durch Multiplikation mit $\mathbf{W}_\ell^{-1/2}$ die Ungleichung

$$\mathbf{A}_\ell^{-1} - \mathbf{p}_\ell\mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell \leq C_A\mathbf{W}_\ell^{-1},$$

und durch Multiplikation mit $\mathbf{A}_\ell^{1/2}$ folgt

$$\mathbf{0} \leq \widehat{\mathbf{Q}}_\ell \leq C_A\widehat{\mathbf{W}}_\ell^{-1}. \tag{4.29}$$

Mit Hilfe dieser Abschätzungen können wir nun eine verbesserte Konvergenzabschätzung für das Zweigitterverfahren finden:

Satz 4.21 (Konvergenz symmetrisches Zweigitterverfahren) *Im symmetrischen Fall mit der symmetrischen Glättungseigenschaft (4.25) und der symmetrischen Approximationseigenschaft (4.26) gilt*

$$\|\mathbf{M}_{\mathrm{ZGV},\ell}\|_A \leq \begin{cases} \left(1 - \frac{1}{C_A}\right)^{2\nu} & \textit{falls } 2\nu < C_A - 1, \\ \frac{C_A}{2\nu e} & \textit{ansonsten} \end{cases} \qquad \textit{für alle } \ell \in \mathbb{N}_0, \nu \in \mathbb{N}.$$

Beweis. Indem wir (4.27) und (4.29) kombinieren, erhalten wir

$$\mathbf{0} \leq \widehat{\mathbf{Q}}_\ell \leq \alpha C_A\widehat{\mathbf{W}}_\ell^{-1} + (1-\alpha)\mathbf{I} \qquad \text{für alle } \alpha \in [0,1],$$

und die transformierte Iterationsmatrix des Zweigitterverfahrens lässt sich in der Form

$$\mathbf{0} \leq \widehat{\mathbf{M}}_{\mathrm{ZGV},\ell} \leq \widehat{\mathbf{M}}_{\mathrm{Gl},\ell}^\nu(\alpha C_A\widehat{\mathbf{W}}_\ell^{-1} + (1-\alpha)\mathbf{I})\widehat{\mathbf{M}}_{\mathrm{Gl},\ell}^\nu \qquad \text{für alle } \alpha \in [0,1]$$

abschätzen. Wegen

$$\widehat{\mathbf{M}}_{\mathrm{Gl},\ell} = \mathbf{I} - \widehat{\mathbf{W}}_\ell^{-1}$$

erhalten wir schließlich

$$\mathbf{0} \leq \widehat{\mathbf{M}}_{\mathrm{ZGV},\ell} \leq f(\widehat{\mathbf{W}}_\ell^{-1}, \alpha) \qquad \text{für alle } \alpha \in [0,1]$$

mit dem Polynom

$$f(\xi,\alpha) = (1-\xi)^{2\nu}(1-\alpha+\alpha C_A\xi).$$

Wegen der Ungleichung (4.28) haben wir $\sigma(\widehat{\mathbf{W}}^{-1}) \subseteq [0,1]$ und folgern

$$\|\mathbf{M}_{\mathrm{ZGV},\ell}\|_A = \|\widehat{\mathbf{M}}_{\mathrm{ZGV},\ell}\|_2 \le \max\{f(\xi,\alpha) \ : \ \xi \in [0,1]\} \qquad \text{für alle } \alpha \in [0,1].$$

Indem wir verschiedene Werte für α einsetzen, können wir unterschiedliche Abschätzungen für die Konvergenzrate erhalten. Wenn wir $\alpha = 1$ wählen, erhalten wir

$$\begin{aligned}
f(\xi,\alpha) &= (1-\xi)^{2\nu} C_A\xi,\\
\frac{\partial f}{\partial \xi}(\xi,\alpha) &= -2\nu(1-\xi)^{2\nu-1}C_A\xi + (1-\xi)^{2\nu}C_A = (1-\xi)^{2\nu-1}(-2\nu C_A\xi + (1-\xi)C_A)\\
&= C_A(1-\xi)^{2\nu-1}(1-(2\nu+1)\xi),
\end{aligned}$$

und das Maximum ist wir für $\xi_0 = 1/(2\nu+1)$ angenommen (die Randpunkte $\xi = 0$ und $\xi = 1$ sind Nullstellen). Wir erhalten die Abschätzung

$$\begin{aligned}
\|\mathbf{M}_{\mathrm{ZGV},\ell}\|_A \le f(\xi_0,\alpha) &= \left(1-\frac{1}{2\nu+1}\right)^{2\nu}\frac{C_A}{2\nu+1} = \left(\frac{2\nu}{2\nu+1}\right)^{2\nu}\frac{C_A}{2\nu+1}\\
&= \left(\frac{2\nu}{2\nu+1}\right)^{2\nu+1}\frac{C_A}{2\nu} = \frac{1}{\left(1+\frac{1}{2\nu}\right)^{2\nu+1}}\frac{C_A}{2\nu} \le \frac{C_A}{2\nu e},
\end{aligned}$$

die wir schon aus dem allgemeinen Fall kennen. Falls $2\nu < C_A - 1$ gelten sollte, können wir auch $\alpha := 2\nu/(C_A-1)$ setzen. Dann erhalten wir

$$\begin{aligned}
f(\xi,\alpha) &= (1-\xi)^{2\nu}\left(1-\frac{2\nu}{C_A-1}+\frac{C_A\xi 2\nu}{C_A-1}\right) = (1-\xi)^{2\nu}\frac{C_A-1+2\nu(C_A\xi-1)}{C_A-1},\\
\frac{\partial f}{\partial \xi}(\xi,\alpha) &= -2\nu(1-\xi)^{2\nu-1}\frac{C_A-1+2\nu(C_A\xi-1)}{C_A-1} + (1-\xi)^{2\nu}\frac{2\nu C_A}{C_A-1}\\
&= \frac{(1-\xi)^{2\nu-1}}{C_A-1}(-2\nu C_A + 2\nu - 4\nu^2 C_A\xi + 4\nu^2 + 2\nu C_A - 2\nu C_A\xi)\\
&= 2\nu\frac{(1-\xi)^{2\nu-1}}{C_A-1}(1-2\nu C_A\xi + 2\nu - C_A\xi)\\
&= 2\nu(2\nu+1)\frac{(1-\xi)^{2\nu-1}}{C_A-1}(1-C_A\xi),
\end{aligned}$$

also wird das Maximum bei $\xi_0 = 1/C_A$ angenommen (der Randpunkt $\xi = 1$ ist eine Nullstelle, die Ableitung im Randpunkt $\xi = 0$ ist positiv, so dass auch dieser Punkt kein Maximum sein kann), und es ergibt sich die Abschätzung

$$\|\mathbf{M}_{\mathrm{ZGV},\ell}\|_A \le f(\xi_0,\alpha) = \left(1-\frac{1}{C_A}\right)^{2\nu}\left(1-\frac{2\nu}{C_A-1}+\frac{2\nu}{C_A-1}\right) = \left(1-\frac{1}{C_A}\right)^{2\nu}.$$

■

Für kleine Werte von ν nimmt die Konvergenzrate also sogar *exponentiell* ab, wenn wir die Anzahl der Glättungsschritte erhöhen. Insbesondere erhalten wir auch schon für $\nu = 1$ ein konvergentes Verfahren.

Unter den für die symmetrischen Verfahren eingeführten Voraussetzungen lässt sich auch ein Konvergenzbeweis für das V-Zyklus-Mehrgitterverfahren gewinnen.

Satz 4.22 (Konvergenz V-Zyklus-Mehrgitterverfahren) *Im symmetrischen Fall mit der symmetrischen Glättungseigenschaft (4.25) und der symmetrischen Approximationseigenschaft (4.26) gilt*

$$\|\mathbf{M}_{\mathrm{MGV},\ell}\|_A \leq \frac{C_A}{C_A + 2\nu} \qquad \textit{für alle } \ell \in \mathbb{N}_0, \nu \in \mathbb{N}.$$

Beweis. Die Iterationsmatrix der V-Zyklus-Mehrgitterverfahrens mit ν Vor- und ν Nachglättungsschritten ist durch

$$\mathbf{M}_{\mathrm{MGV},\ell} = \mathbf{M}_{\mathrm{ZGV},\ell} + \mathbf{M}_{\mathrm{Gl},\ell}^{\nu}\mathbf{p}_\ell\mathbf{M}_{\mathrm{MGV},\ell-1}\mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\mathbf{A}_\ell\mathbf{M}_{\mathrm{Gl},\ell}^{\nu}$$

gegeben. Wir transformieren sie wieder mit Hilfe von $\mathbf{A}_\ell^{1/2}$ und $\mathbf{A}_\ell^{-1/2}$, um die Gleichung

$$\begin{aligned}
\widehat{\mathbf{M}}_{\mathrm{MGV},\ell} := \mathbf{A}_\ell^{1/2}\mathbf{M}_{\mathrm{MGV},\ell}\mathbf{A}_\ell^{-1/2} &= \widehat{\mathbf{M}}_{\mathrm{ZGV},\ell} + \widehat{\mathbf{M}}_{\mathrm{Gl},\ell}^{\nu}\widehat{\mathbf{p}}_\ell\widehat{\mathbf{M}}_{\mathrm{MGV},\ell-1}\widehat{\mathbf{r}}_\ell\widehat{\mathbf{M}}_{\mathrm{Gl},\ell}^{\nu} \\
&= \widehat{\mathbf{M}}_{\mathrm{Gl},\ell}^{\nu}\left(\mathbf{I} - \widehat{\mathbf{p}}_\ell\widehat{\mathbf{r}}_\ell + \widehat{\mathbf{p}}_\ell\widehat{\mathbf{M}}_{\mathrm{MGV},\ell-1}\widehat{\mathbf{r}}_\ell\right)\widehat{\mathbf{M}}_{\mathrm{Gl},\ell}^{\nu} \\
&= \widehat{\mathbf{M}}_{\mathrm{Gl},\ell}^{\nu}\left(\mathbf{I} - \widehat{\mathbf{p}}_\ell(\mathbf{I} - \widehat{\mathbf{M}}_{\mathrm{MGV},\ell-1})\widehat{\mathbf{r}}_\ell\right)\widehat{\mathbf{M}}_{\mathrm{Gl},\ell}^{\nu}
\end{aligned}$$

zu erhalten. Wegen $\mathbf{I} - \widehat{\mathbf{p}}_\ell\widehat{\mathbf{r}}_\ell \geq 0$ und $\widehat{\mathbf{M}}_{\mathrm{MGV},0} = \mathbf{0}$ können wir mit einer einfachen Induktion $\widehat{\mathbf{M}}_{\mathrm{MGV},\ell} \geq \mathbf{0}$ für alle $\ell \in \mathbb{N}_0$ nachweisen.

Unser Ziel ist es nun, die Ungleichung

$$\widehat{\mathbf{M}}_{\mathrm{MGV},\ell} \leq \zeta\mathbf{I}, \qquad \zeta := \frac{C_A}{C_A + 2\nu} \qquad \text{für alle } \ell \in \mathbb{N}$$

zu beweisen. Dazu gehen wir induktiv vor. Der Induktionsanfang $\ell = 0$ ist trivial. Sei nun $\ell \in \mathbb{N}$ so gewählt, dass $\widehat{\mathbf{M}}_{\mathrm{MGV},\ell-1} \leq \zeta\mathbf{I}$ gilt. Daraus folgt

$$\begin{aligned}
\widehat{\mathbf{M}}_{\mathrm{MGV},\ell} &= \widehat{\mathbf{M}}_{\mathrm{Gl},\ell}^{\nu}\left(\mathbf{I} - \widehat{\mathbf{p}}_\ell(\mathbf{I} - \widehat{\mathbf{M}}_{\mathrm{MGV},\ell-1})\widehat{\mathbf{r}}_\ell\right)\widehat{\mathbf{M}}_{\mathrm{Gl},\ell}^{\nu} \\
&\leq \widehat{\mathbf{M}}_{\mathrm{Gl},\ell}^{\nu}\left(\mathbf{I} - \widehat{\mathbf{p}}_\ell(\mathbf{I} - \zeta\mathbf{I})\widehat{\mathbf{r}}_\ell\right)\widehat{\mathbf{M}}_{\mathrm{Gl},\ell}^{\nu} \\
&= \widehat{\mathbf{M}}_{\mathrm{Gl},\ell}^{\nu}\left(\mathbf{I} - (1-\zeta)\widehat{\mathbf{p}}_\ell\widehat{\mathbf{r}}_\ell\right)\widehat{\mathbf{M}}_{\mathrm{Gl},\ell}^{\nu} \\
&= \widehat{\mathbf{M}}_{\mathrm{Gl},\ell}^{\nu}\left((1-\zeta)\widehat{\mathbf{Q}}_\ell + \zeta\mathbf{I}\right)\widehat{\mathbf{M}}_{\mathrm{Gl},\ell}^{\nu} \\
&\leq \widehat{\mathbf{M}}_{\mathrm{Gl},\ell}^{\nu}\left((1-\zeta)\left(\alpha C_A\widehat{\mathbf{W}}_\ell^{-1} + (1-\alpha)\mathbf{I}\right) + \zeta\mathbf{I}\right)\widehat{\mathbf{M}}_{\mathrm{Gl},\ell}^{\nu} \\
&= (\mathbf{I} - \widehat{\mathbf{W}}_\ell^{-1})^{\nu}\left((1-\zeta)\left(\alpha C_A\widehat{\mathbf{W}}_\ell^{-1} + (1-\alpha)\mathbf{I}\right) + \zeta\mathbf{I}\right)(\mathbf{I} - \widehat{\mathbf{W}}_\ell^{-1})^{\nu}
\end{aligned}$$

für alle $\alpha \in [0,1]$. Mit Hilfe des durch

$$\begin{aligned} f(\xi,\alpha) &:= (1-\xi)^{\nu}((1-\zeta)(\alpha C_A\xi + (1-\alpha)) + \zeta)(1-\xi)^{\nu} \\ &= (1-\xi)^{2\nu}((1-\zeta)(\alpha C_A\xi + (1-\alpha)) + \zeta) \qquad \text{für alle } \alpha \in [0,1],\ \xi \in \mathbb{K} \end{aligned}$$

definierten Polynoms können wir die Abschätzung kurz als

$$\widehat{\mathbf{M}}_{\mathrm{MGV},\ell} \le f(\widehat{\mathbf{W}}_{\ell}^{-1},\alpha) \qquad \text{für alle } \alpha \in [0,1]$$

schreiben. Aus (4.28) folgt $\sigma(\widehat{\mathbf{W}}_{\ell}^{-1}) \subseteq [0,1]$, und wir erhalten

$$\|\mathbf{M}_{\mathrm{MGV},\ell}\|_A = \|\widehat{\mathbf{M}}_{\mathrm{MGV},\ell}\|_2 = \varrho(\widehat{\mathbf{M}}_{\mathrm{MGV},\ell}) \le \max\{f(\xi,\alpha)\ :\ \xi \in [0,1]\}.$$

Um zu der gewünschten Abschätzung zu kommen, setzen wir $\alpha = 1$ und müssen das Maximum des durch

$$f(\xi,\alpha) = (1-\xi)^{2\nu}((1-\zeta)C_A\xi + \zeta) \qquad \text{für alle } \xi \in [0,1]$$

gegebenen Polynoms bestimmen. In den Randpunkten $\xi = 0$ und $\xi = 1$ haben wir $f(0,\alpha) = \zeta$ und $f(1,\alpha) = 0$.

Um lokale Maxima zu finden, suchen wir nach Nullstellen der Ableitung

$$\begin{aligned} \frac{\partial f}{\partial \xi}(\xi,\alpha) &= -2\nu(1-\xi)^{2\nu-1}((1-\zeta)C_A\xi + \zeta) + (1-\xi)^{2\nu}(1-\zeta)C_A \\ &= (1-\xi)^{2\nu-1}((1-\xi)(1-\zeta)C_A - 2\nu(1-\zeta)C_A\xi - 2\nu\zeta) \\ &= (1-\xi)^{2\nu-1}((1-(2\nu+1)\xi)(1-\zeta)C_A - 2\nu\zeta) \\ &= (1-\xi)^{2\nu-1}\left((1-(2\nu+1)\xi)\frac{2\nu C_A}{C_A+2\nu} - \frac{2\nu C_A}{C_A+2\nu}\right) \\ &= (1-\xi)^{2\nu-1}\frac{-2\nu C_A(2\nu+1)}{C_A+2\nu}\xi. \end{aligned}$$

Offenbar gibt es genau zwei Nullstellen, nämlich $\xi_1 = 0$ und $\xi_2 = 1$, und das sind die bereits diskutierten Randpunkte. Das Maximum ist offenbar ζ, also haben wir

$$\|\mathbf{M}_{\mathrm{MGV},\ell}\|_A \le f(0,1) = \zeta$$

bewiesen, und damit die gewünschte Abschätzung. ■

Wir haben also bewiesen, dass das V-Zyklus-Mehrgitterverfahren auch für nur einen Vor- und Nachglättungsschritt konvergiert, und wir haben eine explizite Abschätzung für die Konvergenzrate in Abhängigkeit von ν und C_A.

Das Mehrgitterverfahren lässt sich weiter beschleunigen, indem man Techniken aus dem Bereich der semiiterativen Verfahren einsetzt. Einerseits lässt sich die symmetrische Mehrgitteriteration als Vorkonditionierer für das cg-Verfahren einsetzen, was vor allem in Situationen lohnend ist, in denen die Konvergenzrate der Iteration noch relativ nahe bei 1 liegt.

Andererseits lassen sich auch semiiterative Techniken in die einzelnen Komponenten des Mehrgitterverfahrens integrieren. Eine offensichtliche Möglichkeit besteht darin, den Glätter mit Hilfe einer entsprechend modifizierten Tschebyscheff-Semiiteration zu realisieren. Eine weniger offensichtliche Variante ist die Modifikation der Grobgitterkorrektur: Bisher wurde die Korrektur $\mathbf{x}_{\ell-1}$ prolongiert und direkt von $\mathbf{x}_\ell$ abgezogen, wir hatten also

$$\mathbf{x}_\ell \leftarrow \mathbf{x}_\ell - \mathbf{p}_\ell \mathbf{x}_{\ell-1}.$$

Es sind Situationen denkbar, in denen es sinnvoll ist, den Vektor $\mathbf{q}_\ell := \mathbf{p}_\ell \mathbf{x}_{\ell-1}$ lediglich als *Suchrichtung* in einem entsprechend angepassten Gradientenverfahren zu verwenden. Für diese Suchrichtung ergibt sich der optimale Skalierungsparameter als

$$\lambda_{\text{opt}} := \frac{\langle \mathbf{d}_\ell, \mathbf{q}_\ell \rangle_2}{\langle \mathbf{A}_\ell \mathbf{q}_\ell, \mathbf{q}_\ell \rangle_2},$$

und wir erhalten die *optimal gedämpfte Grobgitterkorrektur*

$$\mathbf{x}_\ell \leftarrow \mathbf{x}_\ell - \lambda_{\text{opt}} \mathbf{q}_\ell.$$

Dank der Galerkin-Eigenschaften kann der Faktor λ_{opt} auch durch

$$\begin{aligned}\lambda_{\text{opt}} &= \frac{\langle \mathbf{d}_\ell, \mathbf{q}_\ell \rangle_2}{\langle \mathbf{A}_\ell \mathbf{q}_\ell, \mathbf{q}_\ell \rangle_2} = \frac{\langle \mathbf{p}_\ell^* \mathbf{d}_\ell, \mathbf{x}_{\ell-1} \rangle_2}{\langle \mathbf{p}_\ell^* \mathbf{A}_\ell \mathbf{p}_\ell \mathbf{x}_{\ell-1}, \mathbf{x}_{\ell-1} \rangle_2} \\ &= \frac{c^{-1} \langle \mathbf{r}_\ell \mathbf{d}_\ell, \mathbf{x}_{\ell-1} \rangle_2}{c^{-1} \langle \mathbf{r}_\ell \mathbf{A}_\ell \mathbf{p}_\ell \mathbf{x}_{\ell-1}, \mathbf{x}_{\ell-1} \rangle_2} = \frac{\langle \mathbf{b}_{\ell-1}, \mathbf{x}_{\ell-1} \rangle_2}{\langle \mathbf{A}_{\ell-1} \mathbf{x}_{\ell-1}, \mathbf{x}_{\ell-1} \rangle_2}\end{aligned}$$

berechnet werden, also ausschließlich mit Hilfe von Operationen auf der gröberen Gitterstufe $\ell - 1$. Neben zwei Skalarprodukten erfordert die Bestimmung von λ_{opt} auch eine zusätzliche Matrix-Vektor-Multiplikation, in praktischen Anwendungen sollte man also sorgfältig abwägen, ob der zusätzliche Rechenaufwand zu einer angemessenen Beschleunigung des Verfahrens führt.

In der Praxis hängen die Matrizen $\mathbf{A}_\ell$ häufig von Parametern ab: Beispielsweise können die Koeffizienten der zugrundeliegenden Differentialgleichung variabel sein, oder das Gebiet, auf dem gerechnet werden soll, ändert sich. In dieser Situation ist man an Verfahren interessiert, die nicht nur für ein spezielles Problem, sondern für eine ganze Familie von Problemen gleichmäßig gut funktionieren.

Im Fall des Mehrgitterverfahrens bedeutet das, dass wir nach Verfahren suchen, bei denen $\|\mathbf{M}_{\text{MGV},\ell}\| \leq \zeta < 1$ nicht nur für alle Gitterstufen $\ell \in \mathbb{N}_0$, sondern auch für alle möglichen Parameter des zugrundeliegenden Problems gilt, wir also die Konvergenzrate gleichmäßig beschränken können. Ein Verfahren mit dieser Eigenschaft bezeichnet man als *robust*.

Die Konstruktion von robusten Mehrgitterverfahren kann relativ kompliziert werden, abhängig davon, wie groß die Auswahl der abzudeckenden Probleme ist. In der Praxis haben sich verschiedene Zugänge bewährt:

Falls das zugrundeliegende Gebiet und die Koeffizienten der Differentialgleichung Tensor-Gestalt besitzen, lassen sich mit Hilfe modifizierter Glätter und modifizierter Gitterhierarchien Verfahren konstruieren, die für eine relativ breite Auswahl von Koeffizienten sehr schnell arbeiten.

Bei variablen Koeffizienten lassen sich spezielle Glätter mit Hilfe von approximativen LR-Zerlegungen konstruieren. Die Konstruktion folgt im Prinzip der üblichen LR-Zerlegung, allerdings mit der Nebenbedingung, dass nur wenige zusätzliche von Null verschiedene Einträge erzeugt werden dürfen. Dadurch wird das bei der klassischen LR-Zerlegung unvermeidliche Auffüllen der Matrixstruktur verhindert, und aus einem direkten Löser wird ein iteratives Verfahren, die *ILU-Iteration*.

Ebenfalls im Fall variabler Koeffizienten kann es sich lohnen, die konventionelle Prolongation, die üblicherweise als eine geeignet definierte Interpolation eingeführt wird, durch eine Abbildung zu ersetzen, bei der die Interpolanten mit Hilfe der Matrixkoeffizienten gewichtet werden. Die so konstruierte *matrixabhängige Prolongation* kann dann zur Konstruktion einer passenden Restriktion verwendet werden, und die Galerkin-Eigenschaft lässt sich sicherstellen, indem die Grobgittermatrizen mit Hilfe von $\mathbf{p}_\ell$ und $\mathbf{r}_\ell$ konstruiert werden, statt direkt aus einer Diskretisierung zu entstehen.

Besonders schwierig wird die Situation, wenn das dem Gleichungssystem zugrundeliegende kontinuierliche Problem unbekannt ist oder das Gleichungssystem sogar keine kontinuierliche Entsprechung hat. In diesem Fall können unter gewissen Bedingungen noch *algebraische Mehrgitterverfahren* zum Einsatz kommen. Derartige Verfahren benötigen lediglich die Matrix $\mathbf{A}$ und konstruieren eine Hierarchie von Gleichungssystemen und passende Glätter ausschließlich aufgrund der in $\mathbf{A}$ enthaltenen Informationen.

Algebraische Mehrgitterverfahren erfreuen sich großer Beliebtheit, weil sie sich besonders einfach in existierende Programme integrieren lassen und in vielen praktischen Anwendungsfällen gute Ergebnisse erzielen. Allerdings ist es aufgrund ihrer allgemeinen Struktur sehr schwierig, konkrete Aussagen über ihre Eigenschaften und insbesondere ihre Konvergenzraten zu beweisen.

4.7 Allgemeine Unterraumverfahren

Sei $\mathbf{A}$ positiv definit.

Ein wesentlicher Bestandteil des Zweigitterverfahrens ist die Grobgitterkorrektur

$$\Phi_{\mathrm{GGK},\ell}(\mathbf{x}_\ell, \mathbf{b}_\ell) = \mathbf{x}_\ell - \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1} \mathbf{r}_\ell (\mathbf{A}_\ell \mathbf{x}_\ell - \mathbf{b}_\ell).$$

Für einen Vektor $\mathbf{f}_\ell \in \mathbb{K}^{\mathcal{I}_\ell}$ und einen Testvektor $\mathbf{y}_{\ell-1} \in \mathbb{K}^{\mathcal{I}_{\ell-1}}$ gilt

$$\begin{aligned}
\langle \mathbf{M}_{\mathrm{GGK},\ell} \mathbf{f}_\ell, \mathbf{p}_\ell \mathbf{y}_{\ell-1} \rangle_A &= \langle (\mathbf{A}_\ell - \mathbf{A}_\ell \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1} \mathbf{r}_\ell \mathbf{A}_\ell) \mathbf{f}_\ell, \mathbf{p}_\ell \mathbf{y}_{\ell-1} \rangle_2 \\
&= c^{-1} \langle (\mathbf{r}_\ell \mathbf{A}_\ell - \mathbf{r}_\ell \mathbf{A}_\ell \mathbf{p}_\ell \mathbf{A}_{\ell-1}^{-1} \mathbf{r}_\ell \mathbf{A}_\ell) \mathbf{f}_\ell, \mathbf{y}_{\ell-1} \rangle_2 \\
&= c^{-1} \langle (\mathbf{r}_\ell \mathbf{A}_\ell - \mathbf{r}_\ell \mathbf{A}_\ell) \mathbf{f}_\ell, \mathbf{y}_{\ell-1} \rangle_2 = 0,
\end{aligned}$$

falls die Galerkin-Eigenschaft und $\mathbf{r}_\ell = c\mathbf{p}_\ell^*$ gelten. Die Grobgitterkorrektur ist also ein Iterationsverfahren, das dafür sorgt, dass der Fehler bezüglich des Energie-Skalarprodukts senkrecht auf dem Bild von $\mathbf{p}_\ell$ steht.

Ein ähnliche Eigenschaft besitzt auch das Gauß-Seidel-Verfahren: Im i-ten Schritt wird

$$\begin{aligned} x_i' &= \frac{1}{A_{ii}} \left(b_i - \sum_{\substack{j \in \mathcal{I} \\ \iota(j) > \iota(i)}} A_{ij} x_j - \sum_{\substack{j \in \mathcal{J} \\ \iota(j) < \iota(i)}} A_{ij} x_j' \right) \\ &= x_i - \frac{1}{A_{ii}} \left(\sum_{\substack{j \in \mathcal{I} \\ \iota(j) \geq \iota(i)}} A_{ij} x_j + \sum_{\substack{j \in \mathcal{J} \\ \iota(j) < \iota(i)}} A_{ij} x_j' - b_i \right) \end{aligned}$$

berechnet, und diese Berechnung lässt sich mit Hilfe des i-ten kanonischen Einheitsvektors $\mathbf{e}^{(i)}$ in der Form

$$\Phi_{\mathrm{GS},i}(\mathbf{x}, \mathbf{b}) = \mathbf{x} - \mathbf{e}^{(i)} \frac{1}{A_{ii}} \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{e}^{(i)} \rangle_2 \qquad \text{für alle } \mathbf{x}, \mathbf{b} \in \mathbb{K}^{\mathcal{I}}$$

schreiben, so dass das vollständige Gauß-Seidel-Verfahren als Hintereinanderausführung mehrerer Einzelschritte

$$\Phi_{\mathrm{GS}}(\mathbf{x}, \mathbf{b}) = \Phi_{\mathrm{GS},i_n}(\Phi_{\mathrm{GS},i_{n-1}}(\ldots, \mathbf{b}), \mathbf{b})$$

geschrieben werden kann. Die Iterationsmatrix für einen Einzelschritt ist durch

$$\mathbf{M}_{\mathrm{GS},i} = \mathbf{I} - \mathbf{e}^{(i)} \frac{1}{A_{ii}} (\mathbf{e}^{(i)})^* \mathbf{A}$$

gegeben und erfüllt für $\mathbf{f} \in \mathbb{K}^{\mathcal{I}}$ die Gleichung

$$\begin{aligned} \langle \mathbf{M}_{\mathrm{GS},i} \mathbf{f}, \mathbf{e}^{(i)} \rangle_A &= \langle (\mathbf{A} - \mathbf{A}\mathbf{e}^{(i)} \frac{1}{A_{ii}} (\mathbf{e}^{(i)})^* \mathbf{A}) \mathbf{f}, \mathbf{e}^{(i)} \rangle_2 \\ &= \langle \mathbf{A}\mathbf{f}, \mathbf{e}^{(i)} \rangle_2 - \langle \mathbf{A}\mathbf{e}^{(i)} \frac{1}{A_{ii}} (\mathbf{e}^{(i)})^* \mathbf{A}\mathbf{f}, \mathbf{e}^{(i)} \rangle_2 \\ &= \langle \mathbf{A}\mathbf{f}, \mathbf{e}^{(i)} \rangle_2 - \frac{1}{A_{ii}} \langle (\mathbf{e}^{(i)})^* \mathbf{A}\mathbf{f}, (\mathbf{e}^{(i)})^* \mathbf{A}\mathbf{e}^{(i)} \rangle_2 \\ &= \langle \mathbf{A}\mathbf{f}, \mathbf{e}^{(i)} \rangle_2 - \frac{1}{A_{ii}} \langle \mathbf{A}\mathbf{f}, \mathbf{e}^{(i)} \rangle_2 A_{ii} \\ &= \langle \mathbf{A}\mathbf{f}, \mathbf{e}^{(i)} \rangle_2 - \langle \mathbf{A}\mathbf{f}, \mathbf{e}^{(i)} \rangle_2 = 0, \end{aligned}$$

nach dem i-ten Schritt wird der Fehler bezüglich des Energie-Skalarprodukts also senkrecht auf dem Aufspann des i-ten Einheitsvektors $\mathbf{e}^{(i)}$ stehen.

Sowohl $\Phi_{\mathrm{GGK},\ell}$ als auch $\Phi_{\mathrm{GS},i}$ sind also Iterationsverfahren, die dafür sorgen, dass der Fehler senkrecht auf einem geeignet gewählten Unterraum von $\mathbb{K}^{\mathcal{I}}$ steht: Im ersten Fall auf dem Bild der Prolongation $\mathbf{p}_\ell$, im zweiten Fall auf dem Aufspann des i-ten Einheitsvektors $\mathbf{e}^{(i)}$.

Derartige Verfahren bezeichnet man als *Unterraumverfahren*. Allgemein sind sie durch eine Familie $(\mathbf{P}_\kappa)_{\kappa \in \mathcal{K}}$ von injektiven Matrizen $\mathbf{P}_\kappa \in \mathbb{K}^{\mathcal{I} \times \mathcal{I}_\kappa}$ definiert, die in Anlehnung an

Mehrgitterverfahren ebenfalls als *Prolongationen* bezeichnet werden. Die Bilder dieser Prolongationen bezeichnen wir mit

$$\mathcal{X}_\kappa := \operatorname{Bild} \mathbf{P}_\kappa \qquad \text{für alle } \kappa \in \mathcal{K},$$

und wir verlangen, dass

$$\mathcal{X} := \mathbb{K}^{\mathcal{I}} = \sum_{\kappa \in \mathcal{K}} \mathcal{X}_\kappa \tag{4.30}$$

gilt, dass also die Summe aller Unterräume den gesamten Raum ergibt.

Zu der Prolongation $\mathbf{P}_\kappa$ führen wir die passende Restriktion $\mathbf{R}_\kappa := \mathbf{P}_\kappa^*$ ein und definieren das entsprechende Galerkin-Produkt durch

$$\mathbf{A}_\kappa := \mathbf{R}_\kappa \mathbf{A} \mathbf{P}_\kappa \qquad \text{für alle } \kappa \in \mathcal{K}.$$

Zu jedem $\kappa \in \mathcal{K}$ definieren wir die Unterraumkorrektur

$$\Phi_{\mathrm{UK},\kappa}(\mathbf{x}, \mathbf{b}) := \mathbf{x} - \mathbf{P}_\kappa \mathbf{A}_\kappa^{-1} \mathbf{R}_\kappa (\mathbf{A}\mathbf{x} - \mathbf{b}) \qquad \text{für alle } \mathbf{x}, \mathbf{b} \in \mathbb{K}^{\mathcal{I}}, \kappa \in \mathcal{K}.$$

Für die Wahl

$$\mathbf{P}_\kappa = \mathbf{p}_\ell, \qquad \mathbf{R}_\kappa = c^{-1} \mathbf{r}_\ell, \qquad \mathbf{A}_\kappa = c^{-1} \mathbf{A}_{\ell-1}$$

erhalten wir die Grobgitterkorrektur.

Um einen Einzelschritt des Gauß-Seidel-Verfahrens als Unterraumkorrektur zu identifizieren, fixieren wir eine Numerierung ι von $\mathcal{I}$ und schreiben die Indexmenge als

$$\mathcal{I} = \bigcup_{\kappa \in \mathcal{K}} \mathcal{I}_\kappa$$

für $\mathcal{K} := \mathcal{I}$ und $\mathcal{I}_\kappa := \{\kappa\}$. Für jedes $\kappa \in \mathcal{K}$ definieren wir die Prolongation $\mathbf{P}_\kappa \in \mathbb{K}^{\mathcal{I} \times \mathcal{I}_\kappa}$ als

$$(P_\kappa)_{ij} = \begin{cases} 1 & \text{falls } i = j \in \mathcal{I}_\kappa, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i \in \mathcal{I}, j \in \mathcal{I}_\kappa$$

und erhalten $\mathbf{A}_\kappa = A_{\kappa\kappa}$. Damit gilt $\Phi_{\mathrm{UK},\kappa} = \Phi_{\mathrm{GS},\kappa}$.

Definition 4.23 (Multiplikatives Unterraumverfahren) *Sei $\iota : \mathcal{K} \to \{1, \dots, K\}$ eine Numerierung von $\mathcal{K}$ mit $K := |\mathcal{K}|$, und sei $\kappa_k := \iota^{-1}(k)$ für alle $k \in \{1, \dots, K\}$. Das durch*

$$\Phi_{\mathrm{Mul}}(\mathbf{x}, \mathbf{b}) = \Phi_{\mathrm{UK},\kappa_k}(\Phi_{\mathrm{UK},\kappa_{k-1}}(\dots, \mathbf{b}), \mathbf{b}) \qquad \textit{für alle } \mathbf{x}, \mathbf{b} \in \mathbb{K}^{\mathcal{I}}$$

definierte lineare Iterationsverfahren nennen wir das (exakte) multiplikative Unterraumverfahren *zu den Prolongationen* $(\mathbf{P}_\kappa)_{\kappa \in \mathcal{K}}$.

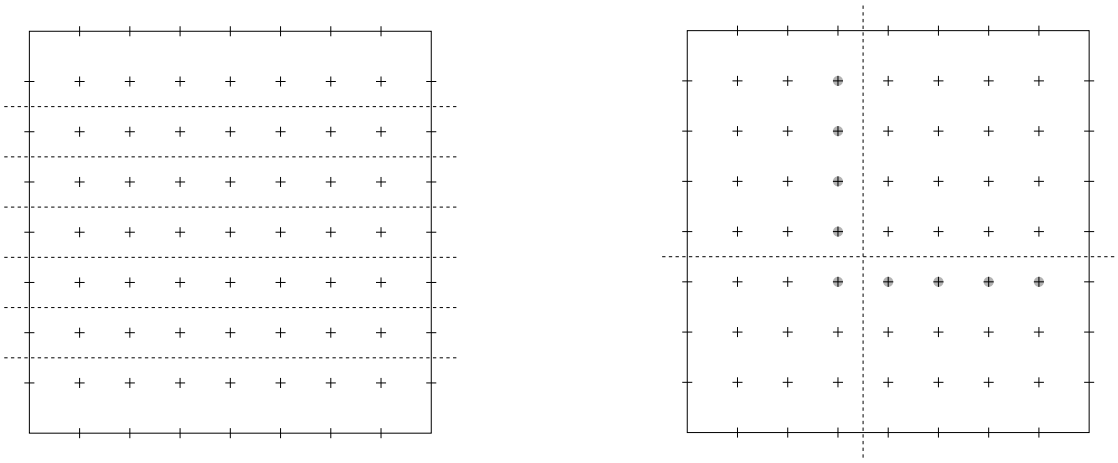

Abbildung 4.10: Spezialfälle des allgemeinen Unterraumverfahrens: Zeilen-Block-Jacobi-Verfahren (links) und Gebietszerlegungsverfahren (rechts)

Wir haben bereits gesehen, dass das Gauß-Seidel-Verfahren sich als multiplikatives Unterraumverfahren mit eindimensionalen Unterräumen interpretieren lässt.

Alternativ können wir die einzelnen Unterraumkorrekturen auch parallel durchführen und die Ergebnisse linear kombinieren.

Definition 4.24 (Additives Unterraumverfahren) *Sei $\theta \in \mathbb{R}_{>0}$. Das durch*

$$\Phi_{\mathrm{Add},\theta}(\mathbf{x},\mathbf{b}) = \mathbf{x} - \theta \sum_{\kappa\in\mathcal{K}} \mathbf{P}_\kappa \mathbf{A}_\kappa^{-1} \mathbf{R}_\kappa (\mathbf{A}\mathbf{x} - \mathbf{b}) \qquad \textit{für alle } \mathbf{x},\mathbf{b} \in \mathbb{K}^{\mathcal{I}}$$

definierte lineare Iterationsverfahren bezeichnen wir als das (exakte) additive Unterraumverfahren *zu den Prolongationen $(\mathbf{P}_\kappa)_{\kappa\in\mathcal{K}}$ und dem Dämpfungsparameter θ.*

So wie das Gauß-Seidel-Verfahren als multiplikatives Unterraumverfahren mit eindimensionalen Unterräumen interpretiert werden kann lässt sich nun das Jacobi-Verfahren als additives Unterraumverfahren mit denselben Unterräumen behandeln.

Selbstverständlich gibt es noch weitere in der Praxis wichtige Spezialfälle von Unterraumverfahren. Ein Beispiel sind Block-Gauß-Seidel und Block-Jacobi-Verfahren, bei denen nicht einzelne Indizes, sondern vollständige Zeilen, Spalten oder kompliziertere Teilmengen eines Gitters auf einmal invertiert werden.

Im Fall des Modellproblems würde beispielsweise ein Zeilen-Block-Verfahren zu der Zerlegung

$$\mathcal{I}_\kappa := \{(i_x,\kappa) \ : \ i_x \in \{1,\dots,N\}\} \qquad \text{für alle } \kappa \in \mathcal{K} := \{1,\dots,N\}$$

gehören (vgl. Abbildung 4.10 links). Natürlich ist so ein Verfahren nur dann sinnvoll, wenn sich Gleichungssysteme mit den Matrizen $\mathbf{A}_\kappa$ effizient lösen lassen. Im Fall des

Zeilen-Block-Verfahrens lässt sich diese Aufgabe glücklicherweise relativ einfach handhaben: Infolge der speziellen Struktur des Modellproblems sind die Matrizen $\mathbf{A}_\kappa$ tridiagonal und positiv definit, so dass sie sich effizient mit Band-LR-Zerlegungen behandeln lassen.

Ein weiterer wichtiger Spezialfall sind die sogenannten *Gebietszerlegungsverfahren*, bei denen ebenfalls die Indexmenge $\mathcal{I}$ in eine Anzahl von Teilmengen zerlegt wird. Hier wird allerdings der Schwerpunkt nicht auf einfache Lösbarkeit der Teilprobleme, sondern auf möglichst kleine „Kontaktflächen" der Teilmengen gelegt, es soll also für jedes $\kappa \in \mathcal{K}$ die Menge

$$\partial\mathcal{I}_\kappa := \{j \in \mathcal{I} \setminus \mathcal{I}_\kappa \ : \ \text{es existiert ein } i \in \mathcal{I}_\kappa \text{ mit } A_{ij} \neq 0\}$$

möglichst klein werden (vgl. Abbildung 4.10 rechts, der „Rand" des oberen rechten Teilgebiets ist grau markiert). Die Idee bei diesen Verfahren besteht darin, die Behandlung der einzelnen Teilmengen $\mathcal{I}_\kappa$ verschiedenen Computern zu übertragen, die dann parallel arbeiten können. Auf dem für ein $\kappa \in \mathcal{K}$ zuständigen Computer werden die Komponenten x_i und b_i nur für $i \in \mathcal{I}_\kappa$ abgespeichert, so dass lediglich bei der Berechnung des Defekts

$$(d_\kappa)_i = \sum_{j\in\mathcal{I}} A_{ij}x_j - b_i = \sum_{j\in\mathcal{I}_\kappa} A_{ij}x_j + \sum_{j\in\partial\mathcal{I}_\kappa} A_{ij}x_j - b_i \qquad \text{für alle } i \in \mathcal{I}_\kappa, \kappa \in \mathcal{K}$$

eine Kommunikation mit den anderen Computern erforderlich ist, um den aktuellen Wert von $x_j \in \partial\mathcal{I}_\kappa$ zu erfahren. Ein additives Unterraumverfahren lässt sich mit diesem Ansatz sehr gut parallelisieren.

Schließlich lassen sich auch die verschiedenen von uns untersuchten Mehrgitterverfahren als Unterraumverfahren interpretieren: Wir nehmen an, dass $\mathbf{r}_\ell = \mathbf{p}_\ell^*$ gilt und dass die Hierarchie der Gleichungssysteme die Galerkin-Eigenschaft besitzt. Wenn die feinste Gitterstufe $L \in \mathbb{N}_0$ ist, setzen wir $\mathcal{I} := \mathcal{I}_L$, $\mathbf{A} = \mathbf{A}_L$, $\mathcal{K} := \{0, \dots, L\}$ und definieren die für das Unterraumverfahren benötigten Prolongationen $\mathbf{P}_\kappa \in \mathbb{K}^{\mathcal{I}\times\mathcal{I}_\kappa}$ durch

$$\mathbf{P}_\kappa := \begin{cases} \mathbf{I} & \text{falls } \kappa = L, \\ \mathbf{P}_{\kappa+1}\mathbf{p}_{\kappa+1} & \text{ansonsten} \end{cases} \qquad \text{für alle } \kappa \in \{0, \dots, L\}. \tag{4.31}$$

Dank der Galerkin-Eigenschaft erhalten wir so $\mathbf{A}_\kappa = \mathbf{A}_\ell$. Allerdings wurde im Mehrgitterverfahren nur auf dem gröbsten Gitter tatsächlich exakt gelöst, auf allen anderen Gitterstufen haben wir lediglich einige Glättungsschritte durchgeführt, also approximativ gelöst.

Um unsere Mehrgitterverfahren als Unterraumverfahren zu interpretieren, müssen wir also *approximative Unterraumkorrekturen*

$$\Phi_{\mathrm{AUK},\kappa}(\mathbf{x}, \mathbf{b}) := \mathbf{x} - \mathbf{P}_\kappa\mathbf{N}_\kappa\mathbf{R}_\kappa(\mathbf{A}\mathbf{x} - \mathbf{b}) \qquad \text{für alle } \mathbf{x}, \mathbf{b} \in \mathbb{K}^{\mathcal{I}},\ \kappa \in \mathcal{K}$$

zulassen, bei denen $\mathbf{N}_\kappa$ eine geeignete positiv definite Approximation von $\mathbf{A}_\kappa^{-1}$ ist. Die Iterationsmatrix von $\Phi_{\mathrm{AUK},\kappa}$ ist durch

$$\mathbf{M}_{\mathrm{AUK},\kappa} := \mathbf{I} - \mathbf{P}_\kappa\mathbf{N}_\kappa\mathbf{R}_\kappa\mathbf{A} \qquad \text{für alle } \kappa \in \mathcal{K}$$

gegeben. Betrachten wir das V-Zyklus-Mehrgitterverfahren mit einem Vorglättungsschritt. Die Glätter seien in der zweiten Normalform

$$\Phi_{\mathrm{Gl},\ell}(\mathbf{x}_\ell, \mathbf{b}_\ell) = \mathbf{x}_\ell - \mathbf{N}_{\mathrm{Gl},\ell}(\mathbf{A}_\ell \mathbf{x}_\ell - \mathbf{b}_\ell) \qquad \text{für alle } \mathbf{x}_\ell, \mathbf{b}_\ell \in \mathbb{K}^{\mathcal{I}_\ell},\ \ell \in \mathbb{N}$$

gegeben. Das durch (4.31) und

$$\mathbf{N}_\kappa = \begin{cases} \mathbf{A}_0^{-1} & \text{falls } \kappa = 0, \\ \mathbf{N}_{\mathrm{Gl},\kappa} & \text{ansonsten} \end{cases} \qquad \text{für alle } \kappa \in \mathcal{K} \tag{4.32}$$

gegebene approximative multiplikative Unterraumverfahren ist gerade das V-Zyklus-Mehrgitterverfahren:

Lemma 4.25 *Das V-Zyklus-Mehrgitterverfahren ist ein approximatives multiplikatives Unterraumverfahren mit den Prolongationen (4.31) und den approximativen Inversen (4.32).*

Beweis. Bekanntlich ist die Iterationsmatrix des Mehrgitterverfahrens durch

$$\mathbf{M}_\ell := \begin{cases} (\mathbf{I} - \mathbf{p}_\ell(\mathbf{I} - \mathbf{M}_{\ell-1})\mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\mathbf{A}_\ell)(\mathbf{I} - \mathbf{N}_\ell\mathbf{A}_\ell) & \text{falls } \ell > 0, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } \ell \in \{0, \dots, L\}$$

gegeben. Da es konsistent ist, können wir das Mehrgitterverfahren in der zweiten Normalform mit der Matrix

$$\widehat{\mathbf{N}}_\ell := (\mathbf{I} - \mathbf{M}_\ell)\mathbf{A}_\ell^{-1} \qquad \text{für alle } \ell \in \{0, \dots, L\}$$

darstellen, denn dann gilt $\mathbf{M}_\ell = \mathbf{I} - \widehat{\mathbf{N}}_\ell\mathbf{A}_\ell$. Die zu dieser Matrix gehörende approximative Unterraumkorrektur ist durch

$$\widehat{\mathbf{M}}_\ell := \mathbf{I} - \mathbf{P}_\ell\widehat{\mathbf{N}}_\ell\mathbf{R}_\ell\mathbf{A} \qquad \text{für alle } \ell \in \{0, \dots, L\}$$

gegeben und wird eine zentrale Rolle in unserem Beweis spielen: Wegen $\mathbf{P}_L = \mathbf{I}$ ist $\widehat{\mathbf{M}}_L = \mathbf{M}_L$ die Iterationsmatrix des Mehrgitterverfahrens, so dass wir lediglich

$$\widehat{\mathbf{M}}_\ell = \mathbf{M}_{\mathrm{AUK},0}\mathbf{M}_{\mathrm{AUK},1} \dots \mathbf{M}_{\mathrm{AUK},\ell} \qquad \text{für alle } \ell \in \{0, \dots, L\}$$

zu beweisen brauchen. Das tun wir natürlich per Induktion: Für $\ell = 0$ ist die Aussage wegen $\mathbf{M}_0 = \mathbf{0}$, also $\widehat{\mathbf{N}}_0 = \mathbf{A}_0^{-1} = \mathbf{N}_0$ bereits klar. Sei nun $\ell \in \{1, \dots, L\}$. Nach Definition gilt

$$\begin{aligned} \widehat{\mathbf{M}}_\ell &= \mathbf{I} - \mathbf{P}_\ell\widehat{\mathbf{N}}_\ell\mathbf{R}_\ell\mathbf{A} = \mathbf{I} - \mathbf{P}_\ell(\mathbf{I} - \mathbf{M}_\ell)\mathbf{A}_\ell^{-1}\mathbf{R}_\ell\mathbf{A} \\ &= \mathbf{I} - \mathbf{P}_\ell(\mathbf{p}_\ell(\mathbf{I} - \mathbf{M}_{\ell-1})\mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell + \mathbf{N}_\ell - \mathbf{p}_\ell(\mathbf{I} - \mathbf{M}_{\ell-1})\mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\mathbf{A}_\ell\mathbf{N}_\ell)\mathbf{R}_\ell\mathbf{A} \\ &= \mathbf{I} - \mathbf{P}_{\ell-1}(\mathbf{I} - \mathbf{M}_{\ell-1})\mathbf{A}_{\ell-1}^{-1}\mathbf{R}_{\ell-1}\mathbf{A} - \mathbf{P}_\ell\mathbf{N}_\ell\mathbf{R}_\ell\mathbf{A} \\ &\qquad + \mathbf{P}_{\ell-1}(\mathbf{I} - \mathbf{M}_{\ell-1})\mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\mathbf{A}_\ell\mathbf{N}_\ell\mathbf{R}_\ell\mathbf{A} \end{aligned}$$

$$
\begin{aligned}
&= \mathbf{I} - \mathbf{P}_{\ell-1}(\mathbf{I} - \mathbf{M}_{\ell-1})\mathbf{A}_{\ell-1}^{-1}\mathbf{R}_{\ell-1}\mathbf{A} - \mathbf{P}_\ell\mathbf{N}_\ell\mathbf{R}_\ell\mathbf{A} \\
&\qquad + \mathbf{P}_{\ell-1}(\mathbf{I} - \mathbf{M}_{\ell-1})\mathbf{A}_{\ell-1}^{-1}\mathbf{r}_\ell\mathbf{R}_\ell\mathbf{A}\ \mathbf{P}_\ell\mathbf{N}_\ell\mathbf{R}_\ell\mathbf{A} \\
&= \mathbf{I} - \mathbf{P}_{\ell-1}\widehat{\mathbf{N}}_{\ell-1}\mathbf{R}_{\ell-1}\mathbf{A} - \mathbf{P}_\ell\mathbf{N}_\ell\mathbf{R}_\ell\mathbf{A} + \mathbf{P}_{\ell-1}\widehat{\mathbf{N}}_{\ell-1}\mathbf{R}_{\ell-1}\mathbf{A}\ \mathbf{P}_\ell\mathbf{N}_\ell\mathbf{R}_\ell\mathbf{A} \\
&= (\mathbf{I} - \mathbf{P}_{\ell-1}\widehat{\mathbf{N}}_{\ell-1}\mathbf{R}_{\ell-1}\mathbf{A})(\mathbf{I} - \mathbf{P}_\ell\mathbf{N}_\ell\mathbf{R}_\ell\mathbf{A}) = \widehat{\mathbf{M}}_{\ell-1}\mathbf{M}_{\mathrm{AUK},\ell}.
\end{aligned}
$$

Damit ist die Induktion abgeschlossen. ■

In ähnlicher Weise können wir auch Mehrgitterverfahren mit mehreren Vor- und Nachglättungen als multiplikative Unterraumverfahren interpretieren, und auch der W-Zyklus stellt uns vor keine unüberwindbaren Schwierigkeiten.

Selbstverständlich lässt sich auch eine additive Unterraumkorrektur auf der Grundlage der für das Mehrgitterverfahren verwendeten Prolongationen und Restriktionen konstruieren. Für die praktische Implementierung bedeutet das lediglich, dass der Defekt *vor* dem Vorglättungsschritt auf das gröbere Gitter transferiert und die auf dem gröberen Gitter berechnete Korrektur erst *nach* dem Nachglättungsschritt von der Iterierten subtrahiert wird. Die Unterschiede in der Implementierung sind also minimal.

Wenden wir uns nun der Analyse des Konvergenzverhaltens zu. Wir untersuchen direkt die approximative Unterraumkorrektur, das exakte Verfahren lässt sich jederzeit durch $\mathbf{N}_\kappa := \mathbf{A}_\kappa^{-1}$ zurückgewinnen.

Zunächst untersuchen wir das additive Verfahren.

Definition 4.26 *Sei $\theta \in \mathbb{R}_{>0}$. Das durch*

$$\Phi_{\mathrm{AAdd},\theta}(\mathbf{x}, \mathbf{b}) = \mathbf{x} - \theta \sum_{\kappa\in\mathcal{K}} \mathbf{P}_\kappa\mathbf{N}_\kappa\mathbf{R}_\kappa(\mathbf{A}\mathbf{x} - \mathbf{b}) \qquad \textit{für alle } \mathbf{x}, \mathbf{b} \in \mathbb{K}^{\mathcal{I}}$$

definierte lineare Iterationsverfahren bezeichnen wir als das approximative additive Unterraumverfahren *zu den Prolongationen $(\mathbf{P}_\kappa)_{\kappa\in\mathcal{K}}$, den approximativen Inversen $(\mathbf{N}_\kappa)_{\kappa\in\mathcal{K}}$ und dem Dämpfungsparameter θ.*

In einem ersten Schritt beweisen wir, dass ein approximatives additives Unterraumverfahren konvergent sein kann. Dazu müssen wir nachweisen, dass die Matrix

$$\mathbf{N}_{\mathrm{AAdd},\theta} = \theta \sum_{\kappa\in\mathcal{K}} \mathbf{P}_\kappa\mathbf{N}_\kappa\mathbf{R}_\kappa$$

der zweiten Normalform positiv definit ist.

Lemma 4.27 *Die Matrix $\mathbf{N}_{\mathrm{AAdd},\theta}$ ist positiv definit.*

Beweis. Da die Matrizen $\mathbf{N}_\kappa$ positiv definit sind, muss $\mathbf{N}_{\mathrm{AAdd},\theta}$ zumindest positiv semidefinit sein. Sei nun $\mathbf{x} \in \mathbb{K}^{\mathcal{I}}$ mit $\langle \mathbf{N}_{\mathrm{AAdd},\theta}\mathbf{x}, \mathbf{x}\rangle_2 = 0$. Daraus folgt

$$0 = \langle \mathbf{N}_{\mathrm{AAdd},\theta}\mathbf{x}, \mathbf{x}\rangle_2 = \theta \sum_{\kappa\in\mathcal{K}} \langle \mathbf{P}_\kappa\mathbf{N}_\kappa\mathbf{R}_\kappa\mathbf{x}, \mathbf{x}\rangle_2 = \theta \sum_{\kappa\in\mathcal{K}} \langle \mathbf{N}_\kappa\mathbf{R}_\kappa\mathbf{x}, \mathbf{R}_\kappa\mathbf{x}\rangle_2.$$

Da die Matrix $\mathbf{N}_\kappa$ positiv definit sind, muss für alle $\kappa \in \mathcal{K}$ also $\mathbf{R}_\kappa \mathbf{x} = \mathbf{0}$ gelten. Für ein beliebiges $\kappa \in \mathcal{K}$ und $\mathbf{y}_\kappa \in \mathbb{K}^{\mathcal{I}_\kappa}$ erhalten wir

$$\langle \mathbf{R}_\kappa \mathbf{x}, \mathbf{y}_\kappa \rangle_2 = \langle \mathbf{x}, \mathbf{P}_\kappa \mathbf{y}_\kappa \rangle_2,$$

also steht $\mathbf{x}$ senkrecht auf $\operatorname{Bild} \mathbf{P}_\kappa$, und somit nach (4.30) auf $\mathbb{K}^{\mathcal{I}}$, muss also gleich Null sein. ■

Damit haben wir eine erste untere Schranke für das Spektrum von $\mathbf{N}_{\mathrm{AAdd},\theta}$. Um den Dämpfungsparameter θ korrekt wählen zu können, benötigen wir auch eine obere Schranke für das Spektrum.

Zur Herleitung dieser Schranke setzen wir $\mathbf{W}_\kappa := \mathbf{N}_\kappa^{-1}$ und nehmen an, dass es für alle $\kappa, \lambda \in \mathcal{K}$ eine Konstante $\epsilon_{\kappa\lambda} \in \mathbb{R}_{\geq 0}$ so gibt, dass

$$|\langle \mathbf{P}_\kappa \mathbf{x}_\kappa, \mathbf{P}_\lambda \mathbf{y}_\lambda \rangle_A| \leq \epsilon_{\kappa\lambda} \|\mathbf{x}_\kappa\|_{W_\kappa} \|\mathbf{y}_\lambda\|_{W_\lambda} \qquad \text{für alle } \mathbf{x}_\kappa \in \mathbb{K}^{\mathcal{I}_\kappa}, \mathbf{y}_\lambda \in \mathbb{K}^{\mathcal{I}_\lambda} \tag{4.33}$$

gilt. Ohne Beschränkung der Allgemeinheit können wir zusätzlich

$$\epsilon_{\kappa\lambda} = \epsilon_{\lambda\kappa} \qquad \text{für alle } \kappa, \lambda \in \mathcal{K}$$

voraussetzen (da (4.33) symmetrisch ist, können wir immer das Minimum von $\epsilon_{\kappa\lambda}$ und $\epsilon_{\lambda\kappa}$ verwenden). Falls alle Unterraumkorrekturen exakt durchgeführt werden, folgt diese Abschätzung beispielsweise aus der Cauchy-Schwarz-Ungleichung für das Energie-Skalarprodukt, allerdings werden sich in der Praxis häufig wesentlich bessere Abschätzungen finden lassen, indem man die Orthogonalität der einzelnen Unterräume ausnutzt. Deshalb wird die Annahme (4.33) im Allgemeinen als *verstärkte Cauchy-Schwarz-Ungleichung* bezeichnet.

Wir fassen die Koeffizienten $\epsilon_{\kappa\lambda}$ in der Matrix $\mathbf{E} \in \mathbb{K}^{\mathcal{K}\times\mathcal{K}}$ mit

$$E_{\kappa\lambda} = \epsilon_{\kappa\lambda} \qquad \text{für alle } \kappa, \lambda \in \mathcal{K}$$

zusammen und erhalten unsere erste Schranke für das Spektrum des Unterraumverfahrens:

Satz 4.28 (Obere Schranke) *Es gilt*

$$\mathbf{N}_{\mathrm{AAdd},\theta} \leq \theta \varrho(\mathbf{E}) \mathbf{A}^{-1}$$

Beweis. Wir setzen $\mathbf{N} := \mathbf{N}_{\mathrm{AAdd},1}$. Nach Definition gilt

$$\mathbf{N} = \sum_{\kappa\in\mathcal{K}} \mathbf{P}_\kappa \mathbf{N}_\kappa \mathbf{R}_\kappa,$$

also folgt für ein $\mathbf{x} \in \mathbb{K}^{\mathcal{I}}$ die Abschätzung

$$\|\mathbf{N}\mathbf{A}\mathbf{x}\|_A^2 = \sum_{\kappa\in\mathcal{K}} \sum_{\lambda\in\mathcal{K}} \langle \mathbf{P}_\kappa \mathbf{N}_\kappa \mathbf{R}_\kappa \mathbf{A}\mathbf{x}, \mathbf{P}_\lambda \mathbf{N}_\lambda \mathbf{R}_\lambda \mathbf{A}\mathbf{x} \rangle_A$$

$$
\begin{aligned}
&\leq \sum_{\kappa\in\mathcal{K}}\sum_{\lambda\in\mathcal{K}} |\langle \mathbf{P}_\kappa\mathbf{N}_\kappa\mathbf{R}_\kappa\mathbf{A}\mathbf{x}, \mathbf{P}_\lambda\mathbf{N}_\lambda\mathbf{R}_\lambda\mathbf{A}\mathbf{x}\rangle_A| \\
&\leq \sum_{\kappa\in\mathcal{K}}\sum_{\lambda\in\mathcal{K}} \epsilon_{\kappa\lambda}\|\mathbf{N}_\kappa\mathbf{R}_\kappa\mathbf{A}\mathbf{x}\|_{W_\kappa}\|\mathbf{N}_\lambda\mathbf{R}_\lambda\mathbf{A}\mathbf{x}\|_{W_\lambda}.
\end{aligned}
$$

Wir definieren den Hilfsvektor $\mathbf{z} \in \mathbb{R}^{\mathcal{K}}$ durch

$$
z_\kappa := \|\mathbf{N}_\kappa\mathbf{R}_\kappa\mathbf{A}\mathbf{x}\|_{W_\kappa} \qquad \text{für alle } \kappa \in \mathcal{K}
$$

ein und erhalten

$$
\begin{aligned}
\|\mathbf{N}\mathbf{A}\mathbf{x}\|_A^2 &\leq \sum_{\kappa\in\mathcal{K}}\sum_{\lambda\in\mathcal{K}} \epsilon_{\kappa\lambda} z_\kappa z_\lambda = \langle \mathbf{z}, \mathbf{E}\mathbf{z}\rangle_2 \leq \|\mathbf{E}\|_2\|\mathbf{z}\|_2^2 = \varrho(\mathbf{E})\sum_{\kappa\in\mathcal{K}}\|\mathbf{N}_\kappa\mathbf{R}_\kappa\mathbf{A}\mathbf{x}\|_{W_\kappa}^2 \\
&= \varrho(\mathbf{E})\sum_{\kappa\in\mathcal{K}}\langle \mathbf{W}_\kappa\mathbf{N}_\kappa\mathbf{R}_\kappa\mathbf{A}\mathbf{x}, \mathbf{N}_\kappa\mathbf{R}_\kappa\mathbf{A}\mathbf{x}\rangle_2 = \varrho(\mathbf{E})\sum_{\kappa\in\mathcal{K}}\langle \mathbf{R}_\kappa\mathbf{A}\mathbf{x}, \mathbf{N}_\kappa\mathbf{R}_\kappa\mathbf{A}\mathbf{x}\rangle_2 \\
&= \varrho(\mathbf{E})\sum_{\kappa\in\mathcal{K}}\langle \mathbf{A}\mathbf{x}, \mathbf{P}_\kappa\mathbf{N}_\kappa\mathbf{R}_\kappa\mathbf{A}\mathbf{x}\rangle_2 = \varrho(\mathbf{E})\sum_{\kappa\in\mathcal{K}}\langle \mathbf{x}, \mathbf{P}_\kappa\mathbf{N}_\kappa\mathbf{R}_\kappa\mathbf{A}\mathbf{x}\rangle_A \\
&= \varrho(\mathbf{E})\langle \mathbf{x}, \mathbf{N}\mathbf{A}\mathbf{x}\rangle_A \leq \varrho(\mathbf{E})\|\mathbf{x}\|_A\|\mathbf{N}\mathbf{A}\mathbf{x}\|_A.
\end{aligned}
$$

Aus dieser Ungleichung folgt

$$
\|\mathbf{N}\mathbf{A}\mathbf{x}\|_A \leq \varrho(\mathbf{E})\|\mathbf{x}\|_A \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^{\mathcal{I}},
$$

und mit Hilfe der Definition der Energienorm erhalten wir

$$
\begin{aligned}
\|\mathbf{A}^{1/2}\mathbf{N}\mathbf{A}^{1/2}\|_2 &= \sup\left\{\frac{\|\mathbf{A}^{1/2}\mathbf{N}\mathbf{A}^{1/2}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \ : \ \mathbf{x} \in \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\}\right\} \\
&= \sup\left\{\frac{\|\mathbf{A}^{1/2}\mathbf{N}\mathbf{A}\mathbf{y}\|_2}{\|\mathbf{A}^{1/2}\mathbf{y}\|_2} \ : \ \mathbf{y} \in \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\}\right\} \\
&= \sup\left\{\frac{\|\mathbf{N}\mathbf{A}\mathbf{y}\|_A}{\|\mathbf{y}\|_A} \ : \ \mathbf{y} \in \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\}\right\} \leq \varrho(\mathbf{E}),
\end{aligned}
$$

also insbesondere

$$
\begin{aligned}
\mathbf{A}^{1/2}\mathbf{N}\mathbf{A}^{1/2} &\leq \varrho(\mathbf{E})\mathbf{I}, \\
\mathbf{N} &\leq \varrho(\mathbf{E})\mathbf{A}^{-1}.
\end{aligned}
$$

Aus $\mathbf{N}_{\mathrm{AAdd},\theta} = \theta\mathbf{N}$ folgt die Behauptung. ■

Wir können die Schranke aus Satz 4.28 auch in der Form

$$
\mathbf{A} \leq \frac{1}{\theta\varrho(\mathbf{E})}\mathbf{W}_{\mathrm{AAdd},\theta}
$$

schreiben, um einen Bezug zu der für symmetrische Mehrgitterverfahren verwendeten Variante der Glättungseigenschaft herzustellen.

Indem wir Lemma 4.27 und Satz 4.28 kombinieren, erhalten wir für $\theta < 2/\varrho(\mathbf{E})$ die Abschätzung

$$\mathbf{I} > \mathbf{I} - \mathbf{A}^{1/2}\mathbf{N}_{\mathrm{AAdd},\theta}\mathbf{A}^{1/2} \geq \mathbf{I} - \theta\varrho(\mathbf{E})\mathbf{I} > \mathbf{I} - 2\mathbf{I} = -\mathbf{I},$$

also $\varrho(\mathbf{M}_{\mathrm{AAdd},\theta}) = \varrho(\mathbf{I} - \mathbf{A}^{1/2}\mathbf{N}_{\mathrm{AAdd},\theta}\mathbf{A}^{1/2}) < 1$, das approximative additive Unterraumverfahren ist also konvergent.

Um eine Schranke für die Konvergenzrate angeben zu können, benötigen wir auch eine untere Schranke für das Spektrum der Matrix $\mathbf{N}_{\mathrm{AAdd},\theta}$. In diesem Fall werden in der Regel *stabile Zerlegungen* verwendet:

Satz 4.29 (Untere Schranke) *Sei $C_S \in \mathbb{R}_{>0}$ so gegeben, dass für jedes $\mathbf{x} \in \mathbb{K}^{\mathcal{I}}$ eine Familie $(\mathbf{x}_\kappa)_{\kappa\in\mathcal{K}}$ mit*

$$\sum_{\kappa\in\mathcal{K}} \mathbf{P}_\kappa \mathbf{x}_\kappa = \mathbf{x}, \tag{4.34a}$$

$$\sum_{\kappa\in\mathcal{K}} \|\mathbf{x}_\kappa\|_{W_\kappa}^2 \leq C_S \|\mathbf{x}\|_A^2 \tag{4.34b}$$

existiert. Dann gilt die Abschätzung

$$\frac{\theta}{C_S}\mathbf{A}^{-1} \leq \mathbf{N}_{\mathrm{AAdd},\theta}.$$

Beweis. Wir setzen wieder $\mathbf{N} := \mathbf{N}_{\mathrm{AAdd},1}$. Sei $\mathbf{x} \in \mathbb{K}^{\mathcal{I}}$, und seien $\mathbf{x}_\kappa \in \mathbb{K}^{\mathcal{I}_\kappa}$ für alle $\kappa \in \mathcal{K}$ so gewählt, dass (4.34) gilt. Dann erhalten wir

$$\begin{aligned}
\|\mathbf{x}\|_A^2 &= \langle \mathbf{x}, \mathbf{x}\rangle_A = \sum_{\kappa\in\mathcal{K}} \langle \mathbf{x}, \mathbf{P}_\kappa \mathbf{x}_\kappa\rangle_A = \sum_{\kappa\in\mathcal{K}} \langle \mathbf{A}\mathbf{x}, \mathbf{P}_\kappa \mathbf{x}_\kappa\rangle_2 \\
&= \sum_{\kappa\in\mathcal{K}} \langle \mathbf{R}_\kappa \mathbf{A}\mathbf{x}, \mathbf{x}_\kappa\rangle_2 = \sum_{\kappa\in\mathcal{K}} \langle \mathbf{N}_\kappa^{1/2}\mathbf{R}_\kappa \mathbf{A}\mathbf{x}, \mathbf{W}^{1/2}\mathbf{x}_\kappa\rangle_2 \\
&\leq \sum_{\kappa\in\mathcal{K}} \|\mathbf{N}_\kappa^{1/2}\mathbf{R}_\kappa \mathbf{A}\mathbf{x}\|_2 \|\mathbf{W}^{1/2}\mathbf{x}_\kappa\|_2 = \sum_{\kappa\in\mathcal{K}} \|\mathbf{N}_\kappa^{1/2}\mathbf{R}_\kappa \mathbf{A}\mathbf{x}\|_2 \|\mathbf{x}_\kappa\|_{W_\kappa} \\
&\leq \left(\sum_{\kappa\in\mathcal{K}} \|\mathbf{N}_\kappa^{1/2}\mathbf{R}_\kappa \mathbf{A}\mathbf{x}\|_2^2\right)^{1/2} \left(\sum_{\kappa\in\mathcal{K}} \|\mathbf{x}\|_{W_\kappa}^2\right)^{1/2} \\
&\leq C_S^{1/2} \left(\sum_{\kappa\in\mathcal{K}} \|\mathbf{N}_\kappa^{1/2}\mathbf{R}_\kappa \mathbf{A}\mathbf{x}\|_2^2\right)^{1/2} \|\mathbf{x}\|_A.
\end{aligned}$$

Aus dieser Abschätzung folgt

$$\begin{aligned}
\|\mathbf{x}\|_A^2 &\leq C_S \sum_{\kappa\in\mathcal{K}} \|\mathbf{N}_\kappa^{1/2}\mathbf{R}_\kappa \mathbf{A}\mathbf{x}\|_2^2 = C_S \sum_{\kappa\in\mathcal{K}} \langle \mathbf{N}_\kappa^{1/2}\mathbf{R}_\kappa \mathbf{A}\mathbf{x}, \mathbf{N}_\kappa^{1/2}\mathbf{R}_\kappa \mathbf{A}\mathbf{x}\rangle_2 \\
&= C_S \sum_{\kappa\in\mathcal{K}} \langle \mathbf{A}\mathbf{P}_\kappa \mathbf{N}_\kappa \mathbf{R}_\kappa \mathbf{A}\mathbf{x}, \mathbf{x}\rangle_2 = C_S \langle \mathbf{A}\mathbf{N}\mathbf{A}\mathbf{x}, \mathbf{x}\rangle_2,
\end{aligned}$$

also die Ungleichungen

$$\mathbf{A} \leq C_S \mathbf{ANA}, \qquad \mathbf{A}^{-1} \leq C_S \mathbf{N}.$$

Dank $\mathbf{N}_{\mathrm{AAdd},\theta} = \theta \mathbf{N}$ folgt daraus unsere Behauptung. ■

Falls die Voraussetzungen der beiden Sätze 4.28 und 4.29 erfüllt sind, erhalten wir $\sigma(\mathbf{N}_{\mathrm{AAdd},\theta}\mathbf{A}) \subseteq [\theta/C_S, \theta\varrho(\mathbf{E})]$. Damit eignet sich das additive Unterraumverfahren gut als Vorkonditionierer für das cg-Verfahren: Falls C_S und $\varrho(\mathbf{E})$ unabhängig von der Anzahl der Unbekannten beschränkt sein sollten, erhalten wir ein Verfahren, dessen Konvergenzrate gleichmäßig beschränkt ist. Die explizite Kenntnis der Schranken ist, anders als im Fall des Richardson- oder Tschebyscheff-Verfahrens, nicht erforderlich.

Index