Konvergenz des Jacobi- und Gauß-Seidel-Verfahrens

Bachelor-Arbeit

im 1-Fach Bachelorstudiengang Mathematik
der Mathematisch-Naturwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel

vorgelegt von
Alina Sophie Wrage

Erstgutachter: Prof. Dr. Steffen Börm
Zweitgutachter: Prof. Dr. Malte Braack

Kiel im August 2016

Inhaltsverzeichnis

1 Einleitung **3**

2 Grundlagen **5**

2.1 Eigenwerte und Spektralradius 5

2.2 Irreduzibilität und Diagonaldominanz 7

3 Jacobi- und Gauß-Seidel-Verfahren **9**

3.1 Lineare Iterationsverfahren 9

3.2 Jacobi-Verfahren 11

3.3 Gauß-Seidel-Verfahren 11

4 Konvergenzanalyse **13**

4.1 Konsistenz 13

4.2 Konvergenz 13

4.3 Konvergenzgeschwindigkeit 19

5 Relaxationsverfahren **23**

5.1 Jacobi-Relaxationsverfahren 24

5.2 Konvergenz Jacobi-Relaxationsverfahren 24

5.3 Gauß-Seidel-Relaxationsverfahren 26

5.4 Konvergenz Gauß-Seidel-Relaxationsverfahren 27

6 Beispiele und Fazit **41**

6.1 Vergleichende Beispiele 41

6.2 Fazit 47

1 Einleitung

In vielen Anwendungen in der Naturwissenschaft ergeben sich Probleme, die der Lösung eines linearen Gleichungssystems

$$Ax = b \tag{1.1}$$

für eine invertierbare Matrix $A \in \mathbb{R}^{n\times n}$ und einen Vektor $b \in \mathbb{R}^n$ für ein $n \in \mathbb{N}$ bedürfen. Oft treten dabei sehr große Systeme auf, bei denen die Matrix A allerdings nur schwach besetzt ist. Dann ist es ratsam, anstelle von direkten Verfahren wie der Gauß-Elimination, der QR- oder der Cholesky-Zerlegung auf Iterationsverfahren zurückzugreifen. Diese können zwar nur noch eine approximative Lösung bestimmen, lassen sich allerdings im Gegensatz zu den eben genannten direkten Verfahren gut anpassen an schwache Besetzungsstrukturen der zu betrachtenden Matrix und sind dann mit einem vergleichbar sehr viel geringeren Rechenaufwand realisierbar.

Die vorliegende Bachelor-Arbeit beschäftigt sich mit dem Jacobi-Verfahren und dem Gauß-Seidel-Verfahren sowie einer Verallgemeinerung dieser, den sogenannten Relaxationsverfahren, insbesondere im Hinblick auf die Konvergenz. Hierzu lassen sich verschiedene Zusammenhänge mit gewissen Eigenschaften der Ausgangsmatrix A zeigen. So bieten sich für die beiden herkömmlichen Verfahren diagonaldominante Matrizen an, wie in Kapitel 4 gezeigt wird. Anschließend werden wir in Kapitel 5 die Konvergenz der zugehörigen Relaxationsverfahren im Zusammenhang mit Eigenschaften von A wie Symmetrie, positiver Definitheit und sogenannter konsistenter Ordnung untersuchen.

In vielen praktischen Anwendungen ergeben sich Gleichungssysteme, die durch eine Matrix mit Blockstruktur gegeben sind. Hierfür existieren vom Jacobi- und Gauß-Seidel-Verfahren abgewandelte Blockversionen. Auf diese und deren Konvergenzverhalten wird in dieser Arbeit nicht eingegangen, dazu finden sich verschiedene Ergebnisse in [3], [5] und [6], von denen die meisten sich von den hier gezeigten Resultaten für die ursprünglichen Verfahren ableiten lassen.

Schließlich werden die verschiedenen Verfahren anhand zweier konkreter Beispiele hinsichtlich ihrer Konvergenz verglichen. Die dabei angeführten Grafiken wurden mithilfe des Programms *gnuplot* erstellt und beruhen auf Ergebnissen einer eigenen Imple-

mentierung der Verfahren in zwei Versionen. Für das Minimalbeispiel, auf das in den Beispielen 4.9 und 6.1 eingegangen wird, wurde eine Implementierung für vollbesetzte Matrizen verwendet, für Beispiel 6.2 eine speziell an Tridiagonalmatrizen angepasste Implementierung.

2 Grundlagen

In diesem Kapitel werden einige Definitionen und grundlegende Resultate aus der linearen Algebra wiederholt, die wir für spätere Resultate benötigen werden. Die Definitionen orientieren sich dabei an [1], die Sätze wie in den Beweisen gekennzeichnet an [7] und [4].

Im Folgenden sei stets $n \in \mathbb{N}$. Ferner bezeichne I_m für alle $m \in \mathbb{N}$ die $(m \times m)$-Einheitsmatrix.

2.1 Eigenwerte und Spektralradius

Der sogenannte Spektralradius einer Matrix wird in den kommenden Betrachtungen eine zentrale Rolle spielen, weshalb wir diesen Begriff und den zugrundeliegenden Begriff des Eigenwerts zunächst noch einmal einführen werden.

Definition 2.1. Sei $A \in \mathbb{C}^{n\times n}$. $\lambda \in \mathbb{C}$ heißt *Eigenwert* von A, wenn ein Vektor $e \in \mathbb{C}^n$ mit $Ae = \lambda e$ existiert. Der Vektor e heißt dann *Eigenvektor* von A bezüglich des Eigenwerts λ. Ferner bezeichne

$$\sigma(A) \ := \ \{\lambda \in \mathbb{C} \ : \ \lambda \text{ Eigenwert von } A\}$$

das *Spektrum* von A. Das Maximum der Beträge der Eigenwerte

$$\rho(A) \ := \ \max \{|\lambda| \ : \ \lambda \in \sigma(A)\}$$

wird *Spektralradius* von A genannt.

Bemerkung 2.2. Für alle $\lambda \in \mathbb{C}$ und $A \in \mathbb{C}^{n\times n}$ gilt

$$\begin{aligned} \lambda I_n - A \text{ ist nicht invertierbar} \quad &\Leftrightarrow \quad \ker(\lambda I_n - A) \neq \{0\} \\ &\Leftrightarrow \quad \exists x \neq 0 : (\lambda I_n - A)x = 0 \\ &\Leftrightarrow \quad \exists x \neq 0 : \lambda x = Ax \\ &\Leftrightarrow \quad \lambda \text{ ist Eigenwert von } A. \end{aligned}$$

Insbesondere ist A invertierbar genau dann, wenn 0 kein Eigenwert von A ist.

Der folgende Satz wird für den späteren Konvergenzbeweis relevant sein.

Satz 2.3. *Für $A \in \mathbb{C}^{n\times n}$ gilt $A^k \xrightarrow{k\to\infty} 0$ genau dann, wenn $\rho(A) < 1$ gilt.*

Beweis. Der Beweis beruht auf dem Beweis von Theorem 1.10 in [7].
Sei $A \in \mathbb{C}^{n\times n}$.
Gelte zunächst $A^k \xrightarrow{k\to\infty} 0$. Sei λ Eigenwert von A mit maximalem Betrag und x zugehöriger Eigenvektor mit $\|x\|_2 = 1$. Da Vielfache des Eigenvektors x ebenfalls Eigenvektoren zum Eigenwert λ sind, folgt induktiv für $k \in \mathbb{N}$

$$A^k x = A^{k-1}(Ax) = A^{k-1}(\lambda x) = \ldots = \lambda^k x.$$

Nehmen wir von beiden Seiten die euklidische Norm, so erhalten wir wegen $\|x\|_2 = 1$

$$|\lambda|^k = |\lambda^k| = |\lambda^k|\|x\|_2 = \|\lambda^k x\|_2 = \|A^k x\|_2 \quad \text{für alle } k \in \mathbb{N}.$$

Da A^k für $k \to \infty$ gegen die Nullmatrix konvergiert, konvergiert auch die Norm auf der rechten Seite gegen Null und somit $|\lambda|^k \xrightarrow{k\to\infty} 0$. Es muss also $\rho(A) = |\lambda| < 1$ gelten.

Für die Rückrichtung gelte nun $\rho(A) < 1$. Betrachte eine zu A ähnliche Matrix J in Jordan-Normalform, es gilt also $A = XJX^{-1}$ für eine invertierbare Matrix X. Für $k \in \mathbb{N}$ folgt dann

$$A^k = (XJX^{-1})^k = XJ^kX^{-1}.$$

Um zu zeigen, dass A^k für $k \to \infty$ gegen die Nullmatrix konvergiert, reicht es somit, zu zeigen, dass J^k für $k \to \infty$ gegen die Nullmatrix konvergiert. Nach Definition besteht J aus Jordan-Blöcken J_i für $i \in \{1, \ldots, m\}$, die diagonal angeordnet sind, und ansonsten nur aus Nullblöcken, die Matrix ist also von der Form

$$J = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_m \end{pmatrix}.$$

Es lässt sich einsehen, dass diese Blockgestalt beim Potenzieren erhalten bleibt, für alle $k \in \mathbb{N}$ gilt nämlich

$$J^k = \begin{pmatrix} J_1^k & & \\ & \ddots & \\ & & J_m^k \end{pmatrix},$$

wie per Induktion nach k gezeigt werden kann.
Für jedes $i \in \{1, \ldots, m\}$ ist der i-te Jordan-Block von der Form

$$J_i = \lambda_i I_{p_i} + E_i \in \mathbb{C}^{p_i \times p_i} \quad \text{für ein } p_i \in \mathbb{N},$$

wobei λ_i der zugehörige Eigenwert von A ist und E_i eine nilpotente Matrix vom Grad p_i, d.h. es gilt $E_i^{p_i} = 0$. Mit dem Binomischen Lehrsatz folgt für alle $i \in \{1, \dots, m\}$ und $k \geq p_i$

$$J_i^k = \sum_{j=0}^{p_i - 1} \frac{k!}{j!(k-j)!} \lambda_i^{k-j} E_i^j.$$

Die Dreiecksungleichung impliziert für jede beliebige Norm $\|\cdot\|$

$$\|J_i^k\| \leq \sum_{j=0}^{p_i - 1} \frac{k!}{j!(k-j)!} |\lambda_i|^{k-j} \|E_i^j\|.$$

Die Norm $\|E_i^j\|$ lässt sich für alle $j \in \{0, \dots, p_i - 1\}$ beschränken durch das Maximum $\max \{\|E_i^j\| : 0 \leq j \leq p_i - 1\}$. Weil zudem für alle $j \in \{0, \dots, p_i - 1\}$ der Faktor

$$\frac{k!}{j!(k-j)!} = \frac{1}{j!} \underbrace{k \cdot (k-1) \cdot (k-2) \cdots (k-j+1)}_{< k^j}$$

nur polynomielles Wachstum bewirkt, konvergiert für $k \to \infty$ wegen $|\lambda_i| < 1$ jeder Summand gegen Null. Da die Summe endlich ist, gilt somit $J_i^k \xrightarrow{k \to \infty} 0$ für alle $i \in \{1, \dots, m\}$ und folglich $J^k \xrightarrow{k \to \infty} 0$. □

Wir können außerdem zeigen, dass der Spektralradius einer Matrix sich durch jede beliebige induzierte Matrixnorm beschränken lässt, wie der folgende Satz zeigt.

Satz 2.4. *Sei $A \in \mathbb{C}^{n \times n}$ und $\|\cdot\|$ eine beliebige Norm. Dann gilt für die davon induzierte Matrixnorm*

$$\rho(A) \ \leq \ \|A\|.$$

Beweis. Der Beweis ist angelehnt an den Beweis von Satz 2.33 in [4].
Sei $\lambda \in \mathbb{C}$ ein betraglich größter Eigenwert von A, d.h. es gilt $|\lambda| = \rho(A)$, und u ein zugehöriger Eigenvektor mit $\|u\| = 1$. Dann gilt

$$\|A\| = \sup_{\|x\|=1} \|Ax\| \overset{\|u\|=1}{\geq} \|Au\| = \|\lambda u\| = |\lambda| \|u\| \overset{\|u\|=1}{=} |\lambda| = \rho(A).$$

□

2.2 Irreduzibilität und Diagonaldominanz

Es wird sich zeigen, dass es für die Konvergenz der Verfahren von Jacobi und Gauß-Seidel vorteilhaft ist, wenn die Ausgangsmatrix bestimmte Eigenschaften erfüllt. In Kapitel 4 werden wir dies vor allem für Matrizen zeigen, deren Diagonale gewissermaßen betraglich dominiert. Daher führen wir die nun folgenden Begriffe ein.

Definition 2.5. Eine Matrix $A \in \mathbb{C}^{n\times n}$ heißt *irreduzibel*, wenn für alle $i, j \in \{1, \ldots, n\}$ eine *Verbindung* von i nach j existiert, d.h. ein Tupel $(i_\ell)_{\ell=0}^{m}$ mit $m \in \mathbb{N}_0$ derart, dass

$$i = i_0, \quad j = i_m \quad \text{und} \quad a_{i_\ell, i_{\ell+1}} \neq 0 \ \text{ für alle } \ell \in \{0, \ldots, m-1\}.$$

Definition 2.6. Eine Matrix $A \in \mathbb{C}^{n\times n}$ heißt

- *(schwach) diagonaldominant*, falls

$$|a_{ii}| \ \geq \ \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \quad \text{für alle } i \in \{1, \ldots, n\}.$$

- *streng/strikt diagonaldominant*, falls

$$|a_{ii}| \ > \ \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \quad \text{für alle } i \in \{1, \ldots, n\}.$$

- *irreduzibel diagonaldominant*, falls A irreduzibel und schwach diagonaldominant ist und (mindestens) ein $i \in \{1, \ldots, n\}$ existiert mit

$$|a_{ii}| \ > \ \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|.$$

3 Jacobi- und Gauß-Seidel-Verfahren

Dieses Kapitel beruht größtenteils auf dem entsprechenden Abschnitt von Yousef Saad in [7], nur Definition 3.1 orientiert sich an [2].

3.1 Lineare Iterationsverfahren

Bevor wir uns den speziellen Iterationsverfahren von Jacobi und Gauß-Seidel zuwenden, werden zunächst lineare Iterationsverfahren im Allgemeinen definieren.

Definition 3.1. Ein *Iterationsverfahren* für das Gleichungssystem (1.1) ist gegeben durch eine Abbildung

$$\Phi : \mathbb{R}^n \to \mathbb{R}^n.$$

Es heißt *linear*, falls eine invertierbare Matrix $M \in \mathbb{R}^{n\times n}$ derart existiert, dass

$$\Phi(x) = x + M^{-1}(b - Ax) \qquad \text{für alle } x \in \mathbb{R}^n.$$

Für jeden Anfangsvektor $x^{(0)} \in \mathbb{R}^n$ wird eine Folge von Iterierten $\left(x^{(k)}\right)_{k=0}^{\infty}$ definiert durch

$$x^{(m+1)} := \Phi\left(x^{(m)}\right) \qquad \text{für alle } m \in \mathbb{N}_0.$$

Durch Umformen erhalten wir die Iterationsvorschrift

$$x^{(m+1)} = (I_n - M^{-1}A)x^{(m)} + M^{-1}b \qquad \text{für alle } m \in \mathbb{N}_0.$$

Die Matrix $G := I_n - M^{-1}A$ bezeichnen wir als zugehörige *Iterationsmatrix*.

Die Iterationsvorschrift

$$x^{(m+1)} = Gx^{(m)} + M^{-1}b \qquad \text{für alle } m \in \mathbb{N}_0$$

lässt sich mit $N = M - A$ und folglich $A = M - N$ wegen

$$M^{-1}N = M^{-1}(M - A) = I_n - M^{-1}A = G$$

auch schreiben in der Form

$$x^{(m+1)} = M^{-1}Nx^{(m)} + M^{-1}b \qquad \text{für alle } m \in \mathbb{N}_0,$$

bzw. mit $f = M^{-1}b$

$$x^{(m+1)} = Gx^{(m)} + f \qquad \text{für alle } m \in \mathbb{N}_0. \tag{3.1}$$

Schreiben wir das Gleichungssystem $Ax = b$ als einzelne Gleichungen, erhalten wir

$$\begin{array}{rcccccl} a_{11}x_1 & + & \dots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & \dots & + & a_{2n}x_n & = & b_2 \\ & & & & & \vdots & \\ a_{n1}x_1 & + & \dots & + & a_{nn}x_n & = & b_n. \end{array}$$

Sowohl das Jacobi- als auch das Gauß-Seidel-Verfahren beruhen auf der Auflösung der jeweils $i-$ten Gleichung für $i \in \{1, \dots, n\}$ nach x_i, die

$$x_i \;=\; \frac{1}{a_{ii}}\left(b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j\right)$$

liefert. Voraussetzung für die Wohldefiniertheit beider Verfahren ist also, dass die Diagonaleinträge von $A = (a_{ij})_{i,j=1,\dots,n}$ alle von Null verschieden sind, d.h.

$$a_{ii} \neq 0 \quad \text{für alle } i \in \{1, \dots, n\}. \tag{3.2}$$

Daraus folgt, dass die Diagonalmatrix $D := \operatorname{diag}\{a_{11}, a_{22}, \dots, a_{nn}\}$ von A ebenfalls invertierbar ist, da $\det(D) = \prod_{i=1}^{n} a_{ii} \neq 0$ gilt.
Zusätzlich zu D definieren wir die strikte untere Dreiecksmatrix $L = (\ell_{ij})_{i,j=1,\dots,n}$ und die strikte obere Dreiecksmatrix $R = (r_{ij})_{i,j=1,\dots,n}$ durch

$$\ell_{ij} \;=\; \begin{cases} -a_{ij}, & \text{falls } i > j \\ 0, & \text{sonst} \end{cases}, \quad r_{ij} \;=\; \begin{cases} -a_{ij}, & \text{falls } i < j \\ 0, & \text{sonst} \end{cases} \qquad \text{für alle } i, j \in \{1, \dots, n\}.$$

Wir erhalten somit eine additive Zerlegung von A gegeben durch

$$A = D - L - R.$$

3.2 Jacobi-Verfahren

Ausgehend von einem Startvektor $x^{(0)} \in \mathbb{R}^n$ ist die Jacobi-Iteration definiert durch

$$x_i^{(m+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij} x_j^{(m)}\right) \quad \text{für alle } i \in \{1,\dots,n\},\ m \in \mathbb{N}_0.$$

Mit unserer Zerlegung $A = D - L - R$ können wir das Verfahren in Matrixschreibweise auch beschreiben durch

$$x^{(m+1)} = D^{-1}(L+R)x^{(m)} + D^{-1}b \quad \text{für alle } m \in \mathbb{N}_0.$$

Mit $M_J := D$ und der Iterationsmatrix

$$G_J := D^{-1}(L+R) = D^{-1}(D-A) = I_n - D^{-1}A$$

ist das Jacobi-Verfahren somit gegeben durch

$$x^{(m+1)} = G_J x^{(m)} + {M_J}^{-1} b \qquad \text{für alle } m \in \mathbb{N}_0,$$

es handelt sich also um ein lineares Iterationsverfahren gemäß Definition 3.1.

3.3 Gauß-Seidel-Verfahren

Bei der Gauß-Seidel-Iteration wird im Gegensatz zur Jacobi-Iteration nach jeder Korrektur eines Eintrags die aktuelle Näherung angepasst, sodass wir für einen gegebenen Startvektor $x^{(0)} \in \mathbb{R}^n$ ein Iterationsverfahren gegeben durch die folgende Vorschrift erhalten:

$$x_i^{(m+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(m+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(m)}\right) \quad \text{für alle } i \in \{1,\dots,n\},\ m \in \mathbb{N}_0. \tag{3.3}$$

Mithilfe unserer Matrixdarstellung können wir dies auch beschreiben durch

$$x^{(m+1)} = D^{-1}\left(b + Lx^{(m+1)} + Rx^{(m)}\right) \qquad \text{für alle } m \in \mathbb{N}_0,$$

woraus

$$Dx^{(m+1)} = b + Lx^{(m+1)} + Rx^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0$$

und somit

$$(D-L)x^{(m+1)} = Rx^{(m)} + b \qquad \text{für alle } m \in \mathbb{N}_0$$

folgt. Da $D - L$ der untere Triangularteil von A ist, ist wegen unserer Voraussetzung (3.2) die Matrix $D - L$ ebenfalls invertierbar, wir erhalten also

$$x^{(m+1)} = (D - L)^{-1} R x^{(m)} + (D - L)^{-1} b \qquad \text{für alle } m \in \mathbb{N}_0.$$

Mit den Matrizen $M_{GS} := D - L$ und

$$G_{GS} := (D - L)^{-1} R = (D - L)^{-1}((D - L) - A) = I_n - (D - L)^{-1} A$$

lautet die Iterationsvorschrift folglich

$$x^{(m+1)} = G_{GS}\ x^{(m)} + {M_{GS}}^{-1} b \qquad \text{für alle } m \in \mathbb{N}_0.$$

Somit ist auch das Gauß-Seidel-Verfahren ein lineares Iterationsverfahren mit der Iterationsmatrix G_{GS}.

4 Konvergenzanalyse

Nachdem wir nun die Verfahren von Jacobi und Gauß-Seidel eingeführt haben, möchten wir deren Konvergenz analysieren. Das folgende Kapitel ist ebenfalls an Ergebnisse von Yousef Saad in [7] angelehnt, lediglich Bemerkung 4.8 verweist auf Resultate aus [4].

4.1 Konsistenz

Bevor wir darauf eingehen, unter welchen Voraussetzungen wir eine Konvergenz erwarten können, beschäftigen wir uns mit der Konsistenz der allgemeinen linearen Iteration (3.1). Wir sind also daran interessiert, ob im Falle der Konvergenz der Grenzwert auch tatsächlich eine Lösung unseres Ausgangssystems darstellt.

Dies lässt sich durch einige wenige Umformungen der Gleichung einsehen. Wenn die Iteration (3.1) konvergiert, ist der Grenzwert x^* Fixpunkt der Iteration, es gilt also

$$x^* = Gx^* + f,$$

bzw. wegen $G = M^{-1}N$ und $f = M^{-1}b$

$$Mx^* = Nx^* + b$$

und somit

$$\underbrace{(M - N)}_{=A} x^* = b.$$

Der Grenzwert ist also tatsächlich Lösung des linearen Gleichungssystems, von dem wir ausgegangen sind.

4.2 Konvergenz

Nun können wir uns der Frage widmen, wann die Verfahren konvergieren. Auch diesbezüglich werden wir uns zunächst mit allgemeinen linearen Iterationsverfahren befassen. Für deren Konvergenz kann ein enger Zusammenhang mit den Eigenwerten der Iterationsmatrix hergestellt werden.

Satz 4.1. *Sei G eine quadratische Matrix mit $\rho(G) < 1$. Dann ist $I_n - G$ invertierbar und die Iteration* (3.1) *konvergiert für jedes f und $x^{(0)}$. Umgekehrt folgt aus der Konvergenz von* (3.1) *für jedes f und $x^{(0)}$, dass $\rho(G) < 1$ gilt.*

Beweis. Der folgende Beweis ist angelehnt an die Herleitung von Theorem 4.1 in [7]. Gelte zunächst $\rho(G) < 1$. Da somit $1 \notin \sigma(G)$ gilt, ist nach Bemerkung 2.2 die Matrix $I_n - G$ invertierbar. Es existiert also eine Lösung x^* von

$$(I_n - G)x = f,$$

was äquivalent ist zu $x = Gx + f$. Somit ist x^* Fixpunkt von (3.1), wir erhalten also für alle $k \in \mathbb{N}_0$

$$x^{(k+1)} - x^* = G(x^{(k)} - x^*) = \ldots = G^{k+1}(x^{(0)} - x^*). \tag{4.1}$$

Mit Satz 2.3 folgt, dass $x^{(k)} - x^*$ für $k \to \infty$ gegen Null konvergiert.

Konvergiert umgekehrt die Iteration (3.1) für jedes f und $x^{(0)}$, so zeigt die Gleichung

$$x^{(k+1)} - x^{(k)} = G(x^{(k)} - x^{(k-1)}) = \ldots = G^k(x^{(1)} - x^{(0)}) = G^k(f - (I_n - G)x^{(0)}),$$

dass $G^k v \xrightarrow{k\to\infty} 0$ für jeden Vektor v gilt und somit $G^k \xrightarrow{k\to\infty} 0$. Folglich muss, erneut nach Satz 2.3, $\rho(G) < 1$ gelten. □

Mit Satz 2.4 erhalten wir direkt das folgende Korollar.

Korollar 4.2. *Sei G eine quadratische Matrix derart, dass $\|G\| < 1$ für eine induzierte Matrixnorm $\|\cdot\|$ gilt. Dann ist $I_n - G$ invertierbar und die Iteration* (3.1) *konvergiert für jeden Anfangsvektor $x^{(0)}$.*

Der nun folgende Satz von Gerschgorin bietet die Möglichkeit, die Position der Eigenwerte einer Matrix in der komplexen Ebene einzugrenzen.

Satz 4.3 (**Gerschgorin**). *Für jeden Eigenwert $\lambda \in \mathbb{C}$ einer Matrix $A \in \mathbb{C}^{n\times n}$ existiert ein $i \in \{1, \ldots, n\}$ derart, dass*

$$|\lambda - a_{ii}| \;\leq\; \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| =: \rho_i.$$

Die Scheiben $D(a_{ii}, \rho_i)$ in der komplexen Ebene um die Mittelpunkte a_{ii} mit Radius ρ_i für $i \in \{1, \ldots, n\}$ nennen wir die n Gerschgorin-Scheiben.

Beweis. Der Beweis orientiert sich am Beweis von Theorem 4.6 in [7].
Sei λ ein Eigenwert von A und x zugehöriger Eigenvektor. Sei $m \in \{1, \ldots, n\}$ derart, dass $|x_m| \geq |x_i|$ für alle $i \in \{1, \ldots, n\}$ gilt. Ohne Beschränkung der Allgemeinheit sei

x so skaliert, dass $|x_m| = 1$ und $|x_i| \leq 1$ für $i \neq m$ gilt.
Da x Eigenvektor zum Eigenwert λ ist, gilt insbesondere

$$\lambda x_m = (\lambda x)_m = (Ax)_m = \sum_{j=1}^{n} a_{mj} x_j$$

und somit

$$(\lambda - a_{mm}) x_m = \sum_{\substack{j=1 \\ j \neq m}}^{n} a_{mj} x_j.$$

Es folgt mit $|x_m| = 1$, der Dreiecksungleichung und $|x_j| \leq 1$ für $j \neq m$

$$|\lambda - a_{mm}| \ \leq \ \sum_{\substack{j=1 \\ j \neq m}}^{n} |a_{mj}||x_j| \ \leq \ \sum_{\substack{j=1 \\ j \neq m}}^{n} |a_{mj}| = \rho_m. \tag{4.2}$$

□

Für irreduzible Matrizen kann nun mithilfe der Gerschgorin-Scheiben folgendes Resultat gezeigt werden, welches wir für den Beweis des anschließend folgenden Satzes zur Invertierbarkeit irreduzibel diagonaldominanter Matrizen benötigen werden.

Satz 4.4. *Sei A eine irreduzible Matrix und λ ein Eigenwert von A, der auf dem Rand der Vereinigung der n Gerschgorin-Scheiben liegt. Dann liegt λ auf dem Rand jeder einzelnen Gerschgorin-Scheibe.*

Beweis. Der Beweis beruht auf dem Beweis zu Theorem 4.7 in [7].
Sei x wieder Eigenvektor zum Eigenwert λ mit $|x_m| = 1$ und $|x_i| \leq 1$ für $i \neq m$. Wie im Beweis des Satzes von Gershgorin erhalten wir die Ungleichung (4.2).
Da λ auf dem Rand der Vereinigung aller Scheiben liegt, kann λ kein innerer Punkt der Scheibe $D(a_{mm}, \rho_m)$ sein, folglich gilt $|\lambda - a_{mm}| = \rho_m$. Die beiden Ungleichheiten in (4.2) müssen in diesem Fall also Identitäten sein, es folgt

$$|\lambda - a_{mm}| \ = \ \sum_{\substack{j=1 \\ j \neq m}}^{n} |a_{mj}||x_j| \ = \ \sum_{\substack{j=1 \\ j \neq m}}^{n} |a_{mj}| = \rho_m.$$

Daraus folgt $|x_j| = 1$ für alle $j \neq m$ mit $a_{mj} \neq 0$, da wir bereits wissen, dass $|x_j| \leq 1$ für alle $j \neq m$ gilt.
Sei nun $j \in \{1, \ldots, n\}$, $j \neq m$. Wir wollen zeigen, dass λ auf dem Rand der j-ten Gershgorin-Scheibe liegt, also $|\lambda - a_{jj}| = \rho_j$. Da A irreduzibel ist, existiert eine Verbindung $(m_\ell)_{\ell=0}^{k}$, $k \in \mathbb{N}$, von m nach j, es gilt also

$$m_0 = m, \quad m_k = j \quad \text{und} \quad a_{m_\ell, m_{\ell+1}} \neq 0 \ \text{ für alle } \ell \in \{0, \ldots, k-1\}.$$

Insbesondere gilt $a_{m,m_1} \neq 0$ und somit mit obiger Beobachtung $|x_{m_1}| = 1$.
Wendet man das gleiche Argument wie oben auf m_1 an statt auf m, so erhält man die Gleichung

$$|\lambda - a_{m_1,m_1}| = \sum_{\substack{j=1 \\ j\neq m_1}}^{n} |a_{m_1,j}||x_j| = \sum_{\substack{j=1 \\ j\neq m_1}}^{n} |a_{m_1,j}| = \rho_{m_1}.$$

und induktiv $|\lambda - a_{m_i,m_i}| = \rho_{m_i}$ für alle $i \in \{1,\dots,k\}$. Schließlich folgt aus dieser Eigenschaft für $i = k$ wegen $m_k = j$, dass λ auf dem Rand der j-ten Gerschgorin-Scheibe liegt. Da $j \neq m$ beliebig gewählt war und λ wie zu Beginn erläutert auch auf dem Rand der m-ten Scheibe liegt, folgt die Behauptung. □

Korollar 4.5. *Jede streng diagonaldominante oder irreduzibel diagonaldominante Matrix A ist invertierbar.*

Beweis. Der Beweis ist angelehnt an den Beweis von Korollar 4.8 in [7].
Sei A zunächst streng diagonaldominant. Dann kann $\lambda = 0$ kein Eigenwert sein, da sonst nach dem Satz von Gerschgorin gelten würde

$$|a_{mm}| \ \leq \ \sum_{\substack{j=1 \\ j\neq m}}^{n} |a_{mj}| \quad \text{für ein } m \in \{1,\dots,n\}.$$

Dies stellt einen Widerspruch zur strengen Diagonaldominanz dar. Mit Bemerkung 2.2 ist A also invertierbar.

Betrachten wir nun den zweiten Fall, A sei also irreduzibel diagonaldominant. Angenommen, A ist nicht invertierbar. Dann ist nach Bemerkung 2.2 $\lambda = 0$ Eigenwert von A, liegt also in einer Gerschgorin-Scheibe $D(a_{kk}, \rho_k)$ und somit in der Vereinigung aller Scheiben. Da aufgrund der schwachen Diagonaldominanz insbesondere

$$|a_{ii}| \ \geq \ \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|.$$

für alle $i \in \{1,\dots,n\}$ gilt, kann der Eigenwert 0 außerdem nicht im Inneren einer der Scheiben liegen und muss somit auf dem Rand der Vereinigung der Gerschgorin-Scheiben liegen. Nach Satz 4.4 liegt er damit auf dem Rand aller Scheiben, es gilt also

$$|a_{ii}| \ = \ \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \quad \text{für alle } i \in \{1,...,n\},$$

was ein Widerspruch zur irreduziblen Diagonaldominanz ist. □

Damit können wir uns schließlich der Konvergenz von Jacobi- und Gauß-Seidel-Verfahren widmen, welche wir im Folgenden im Zusammenhang mit Diagonaldominanz untersuchen werden.

Bemerkung 4.6. Ist A streng diagonaldominant, so sind insbesondere alle Diagonaleinträge von A von Null verschieden, die beiden Verfahren sind also wohldefiniert.
Im Falle der irreduziblen Diagonaldominanz von A ist diese Voraussetzung nicht so direkt zu erkennen, sie ist aber ebenso gegeben. Da hier nur schwache Diagonaldominanz vorliegt, erhalten wir zunächst nur, dass $a_{ii} \neq 0$ für $i \in \{1, \dots, n\}$ gilt, falls die i-te Zeile von A nicht nur aus Nullen besteht. Aufgrund der Irreduzibilität von A kann es jedoch keine reinen Nullzeilen geben, daher sind auch in diesem Fall beide Verfahren wohldefiniert.

Satz 4.7. *Ist A eine streng diagonaldominante oder irreduzibel diagonaldominante Matrix, dann konvergieren das assoziierte Jacobi- und Gauß-Seidel-Verfahren für jeden Startvektor $x^{(0)}$.*

Beweis. Der Satz entspricht Theorem 4.9 in [7], der Beweis wurde jedoch abgewandelt. Wir führen den Beweis per Widerspruch. Sei A streng diagonaldominant oder irreduzibel diagonaldominant und λ sei Eigenwert der jeweiligen Iterationsmatrix G mit $|\lambda| = \rho(G)$. Ferner sei x zugehöriger Eigenvektor mit $|x_m| = 1$ und $|x_i| \leq 1$ für $i \neq m$. Wir nehmen an, dass $|\lambda| \geq 1$ gilt.

Im Falle des Jacobi-Verfahrens haben wir somit

$$D^{-1}(L + R)x = \lambda x,$$

woraus wir

$$(\lambda D - L - R)x = 0 \tag{4.3}$$

erhalten. Für die Einträge der Matrix $\lambda D - L - R =: A' = (a'_{ij})_{i,j=1,\dots,n}$ gilt

$$a'_{ij} = \begin{cases} \lambda a_{ii}, & \text{falls } i = j \\ a_{ij}, & \text{falls } i \neq j \end{cases} \quad \text{für } i, j \in \{1, \dots, n\}.$$

Es gilt also $a'_{ij} \neq 0 \Leftrightarrow a_{ij} \neq 0$ für alle $i, j \in \{1, \dots, n\}$, da wegen $|\lambda| \geq 1$ insbesondere $\lambda \neq 0$ gilt, und für die Beträge der Einträge von A' gilt

$$|a'_{ii}| \geq |a_{ii}| \text{ für alle } i \in \{1, \dots, n\} \quad \text{sowie} \quad |a'_{ij}| = |a_{ij}| \text{ für alle } i, j \in \{1, \dots, n\}, i \neq j.$$

Folglich überträgt sich nach Definition die strikte bzw. irreduzible Diagonaldominanz von A auf die Matrix A'.

Im Falle des Gauß-Seidel-Verfahrens gilt

$$(D - L)^{-1}Rx = \lambda x,$$

woraus sich

$$(\lambda D - \lambda L - R)x = 0 \tag{4.4}$$

ergibt. Betrachten wir die Einträge der Matrix $\lambda D - \lambda L - R =: A'' = (a''_{ij})_{i,j=1,\dots,n}$, so erhalten wir

$$a''_{ij} = \begin{cases} \lambda a_{ij}, & \text{falls } i \geq j \\ a_{ij}, & \text{falls } i < j \end{cases} \quad \text{für } i,j \in \{1,\dots,n\}.$$

Hier gilt folglich im Falle der strengen Diagonaldominanz von A

$$|a''_{ii}| \;=\; |\lambda||a_{ii}| \;>\; |\lambda| \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \;=\; \sum_{\substack{j=1 \\ j\neq i}}^{n} |\lambda||a_{ij}| \;=\; \sum_{j=1}^{i-1} \underbrace{|\lambda||a_{ij}|}_{\geq |a_{ij}|=|a''_{ij}|} \;+\; \sum_{j=i+1}^{n} \underbrace{|\lambda a_{ij}|}_{=|a''_{ij}|} \;\geq\; \sum_{\substack{j=1 \\ j\neq i}}^{n} |a''_{ij}|$$

für alle $i \in \{1,\dots,n\}$, die Matrix A'' ist dann also ebenfalls streng diagonaldominant. Ist A irreduzibel diagonaldominant, so erhalten wir analog dazu

$$|a''_{ij}| \;\geq\; \sum_{\substack{j=1 \\ j\neq i}}^{n} |a''_{ij}| \quad \text{für alle } i \in \{1,\dots,n\}$$

sowie

$$|a''_{kj}| \;>\; \sum_{\substack{j=1 \\ j\neq k}}^{n} |a''_{kj}| \quad \text{für ein } k \in \{1,\dots,n\}.$$

Ferner gilt wegen $\lambda \neq 0$ erneut $a''_{ij} = 0 \;\Leftrightarrow\; a_{ij} = 0$ für alle $i,j \in \{1,\dots,n\}$, die Matrix A'' ist in dem Fall somit ebenfalls irreduzibel diagonaldominant.

Nach Korollar 4.5 sind A' und A'' also in allen Fällen invertierbar. Damit ergibt sich allerdings ein Widerspruch zu den Gleichungen (4.3) und (4.4), da der Vektor x als Eigenvektor $x \neq 0$ erfüllt. Der Widerspruch impliziert, dass die Annahme falsch war und somit in allen Fällen $\rho(G) = |\lambda| < 1$ für die Iterationsmatrix G gilt, nach Satz 4.1 konvergieren somit die jeweiligen Verfahren.

□

Bemerkung 4.8. (vgl. Satz 4.11 in [4])
Die Konvergenz des Jacobi-Verfahrens im Falle der strikten Diagonaldominanz von A lässt sich auch sehr kurz mithilfe von Korollar 4.2 über die Zeilensummennorm $\|\cdot\|_\infty$ zeigen, denn es gilt

$$\|G_J\|_\infty \;=\; \max_{i\in\{1,\dots,n\}} \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1.$$

Ferner kann analog auch die Konvergenz des Jacobi-Verfahrens unter Voraussetzung des *starken Spaltensummenkriteriums*

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

gezeigt werden, da für die sogenannte Spaltensummennorm $\|\cdot\|_1$ gilt, dass $\|G_J\|_1 = q$.

4.3 Konvergenzgeschwindigkeit

Neben der reinen Konvergenz sind wir vor allem im Hinblick auf die praktische Anwendung der Verfahren auch daran interessiert, wie schnell diese konvergieren. Wir werden sehen, dass die Konvergenzgeschwindigkeit in direktem Zusammenhang mit dem Spektralradius der Iterationsmatrix steht. Die folgenden Betrachtungen beruhen auf dem entsprechenden Abschnitt in [7], der anschließend an das dortige Korollar 4.2 zu finden ist.

Sei im Folgenden x^* die Lösung der Iteration (3.1) und für alle $k \in \mathbb{N}_0$

$$d_k := x^{(k)} - x^*$$

der Fehler in Stufe k. Dann gilt nach Gleichung (4.1) aus dem Beweis von Satz 4.1

$$d_k = G^k d_0 \qquad \text{für alle } k \in \mathbb{N}_0.$$

Um die Konvergenzgeschwindigkeit zu untersuchen, betrachten wir wieder eine zu G ähnliche Jordan-Normalform J, d.h. es gilt $G = XJX^{-1}$ für eine invertierbare Matrix X.

Der Einfachheit halber nehmen wir an, dass G nur einen einzigen Eigenwert λ mit maximalem Betrag $\rho(G)$ hat, und damit J ebenso, da aufgrund der Ähnlichkeit zu G die Eigenwerte der beiden Matrizen identisch sind. Wir erhalten

$$d_k \;=\; \lambda^k \left(\frac{1}{\lambda} G\right)^k d_0 \;=\; \lambda^k \left(X \left(\frac{1}{\lambda} J\right) X^{-1}\right)^k d_0 \;=\; \lambda^k X \left(\frac{1}{\lambda} J\right)^k X^{-1} d_0. \qquad (4.5)$$

Für die Potenzen der Jordanmatrix $\frac{1}{\lambda}J$ gilt, dass alle ihre Blöcke bis auf den zum Eigenwert λ gehörenden Block für $k \to \infty$ gegen die Nullmatrix der jeweiligen Größe konvergieren, da diese Jordan-Blöcke mit Diagonaleinträgen betraglich echt kleiner als 1 sind. Sei der zum Eigenwert λ gehörende Jordan-Block von Größe p und von der Form $J_\lambda = \lambda I_p + E$ für eine nilpotente Matrix E vom Grad p. Dann gilt $E^p = 0$, also gilt für $k \geq p$ mit dem Binomischen Lehrsatz

$$J_\lambda^k \;=\; (\lambda I_p + E)^k \;=\; \lambda^k (I_p + \lambda^{-1} E)^k \;=\; \lambda^k \left(\sum_{i=0}^{p-1} \lambda^{-i} \binom{k}{i} E^i\right).$$

Da E nilpotent vom Grad p ist, gilt auch $E^\ell \neq 0$ für alle $\ell < p$ und somit insbesondere $E^{p-1} \neq 0$. Ist k groß genug, ist wegen des Faktors $\binom{k}{i}$ der dominante Term dieser Summe der letzte, also lässt sich

$$J_\lambda^k \;\approx\; \lambda^{k-p+1} \binom{k}{p-1} E^{p-1},$$

approximieren, da E nilpotent vom Grad p ist und somit insbesondere $E^{p-1} \neq 0$ gilt. Mit (4.5) folgt für die Norm des Fehlers

$$\|d_k\| \;=\; \|XJ^kX^{-1}d_0\| \;\approx\; C|\lambda^{k-p+1}|\binom{k}{p-1} \qquad \text{für eine Konstante } C.$$

Für den sogenannten *Konvergenzfaktor*

$$\rho \;:=\; \lim_{k\to\infty}\left(\frac{\|d_k\|}{\|d_0\|}\right)^{1/k}$$

gilt wegen $\alpha^{1/k} \xrightarrow{k\to\infty} 1$ für $\alpha \in \mathbb{R}_{>0}$

$$\rho = \lim_{k\to\infty}\left(\left(\|d_k\|\right)^{1/k}\left(\frac{1}{\|d_0\|}\right)^{1/k}\right) \;=\; \lim_{k\to\infty}\left(\|d_k\|\right)^{1/k},$$

mit obiger Approximation erhalten wir also

$$\rho \;\approx\; \lim_{k\to\infty}\left(C^{1/k}|\lambda|^{\frac{k-p+1}{k}}\binom{k}{p-1}^{1/k}\right).$$

Da für großes k gilt, dass $\binom{k}{p-1} \approx \binom{k}{1} = k$ ist, folgt wegen $k^{1/k} \xrightarrow{k\to\infty} 1$ und wegen $C > 0$ mit obigem Argument

$$\rho \;\approx\; \lim_{k\to\infty}|\lambda|^{\frac{k-p+1}{k}} \;\overset{\text{Stetigkeit}}{=}\; |\lambda|^{\lim_{k\to\infty}\frac{k-p+1}{k}} \;=\; |\lambda|^1 \;=\; |\lambda| \;=\; \rho(G).$$

Beispiel 4.9. Als Minimalbeispiel betrachten wir das einfache lineare Gleichungssystem $Ax = b$ mit

$$A = \begin{pmatrix} 8 & 5 & 0 \\ 3 & 6 & 2 \\ 0 & 4 & 5 \end{pmatrix} \in \mathbb{R}^{3\times 3}, \quad b = \begin{pmatrix} 2 \\ 11 \\ 13 \end{pmatrix} \in \mathbb{R}^3.$$

Die Lösung ist $x = (-1, 2, 1)^T$ und die Iterationsmatrizen der beiden Verfahren lauten

$$G_J = I_3 - \begin{pmatrix} 8 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 5 \end{pmatrix}^{-1}\begin{pmatrix} 8 & 5 & 0 \\ 3 & 6 & 2 \\ 0 & 4 & 5 \end{pmatrix} = \begin{pmatrix} 0 & -5/8 & 0 \\ -1/2 & 0 & -1/3 \\ 0 & -4/5 & 0 \end{pmatrix},$$

$$G_{GS} = I_3 - \begin{pmatrix} 8 & 0 & 0 \\ 3 & 6 & 0 \\ 0 & 4 & 5 \end{pmatrix}^{-1}\begin{pmatrix} 8 & 5 & 0 \\ 3 & 6 & 2 \\ 0 & 4 & 5 \end{pmatrix} = \begin{pmatrix} 0 & -5/8 & 0 \\ 0 & 5/16 & -1/3 \\ 0 & -1/4 & 4/15 \end{pmatrix}.$$

Daraus ergeben sich die Spektren

$$\sigma(G_J) = \{-\alpha,\ 0,\ \alpha\} \quad \text{mit } \alpha = \sqrt{\frac{139}{240}} = \frac{1}{4}\sqrt{\frac{139}{15}} \quad \text{und}$$
$$\sigma(G_{GS}) = \left\{0,\ \frac{139}{240}\right\},$$

und somit die Spektralradien

$$\rho(G_J) = \alpha \approx 0.76103, \qquad \rho(G_{GS}) = \frac{139}{240} \approx 0.579167.$$

Bei der Durchführung des Jacobi- und Gauß-Seidel-Verfahrens für dieses Beispiel mit dem Anfangsvektor $x^{(0)} = (10, 10, 10)^T$ ergeben sich die in Abbildung 4.1 erkennbaren Ergebnisse für die Norm $e_k := \|d_k\|_2$ des Fehlers d_k nach $k = 0, \ldots, 50$ Schritten.

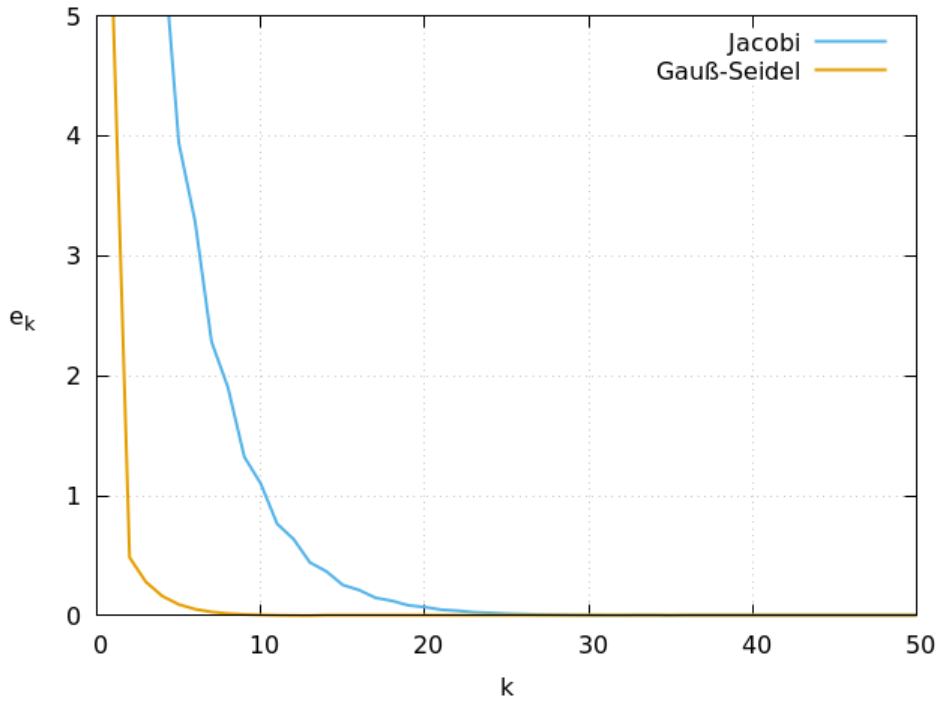

Abbildung 4.1: Fehler bei Verwendung des Jacobi- und Gauß-Seidel-Verfahrens

Betrachten wir in jedem Schritt der Iteration außerdem jeweils als Annäherung an den Konvergenzfaktor den Wert

$$\rho_k := \left(\frac{\|d_k\|}{\|d_0\|}\right)^{1/k},$$

so erhalten wir im Vergleich der beiden Verfahren die Grafik in Abbildung 4.2.

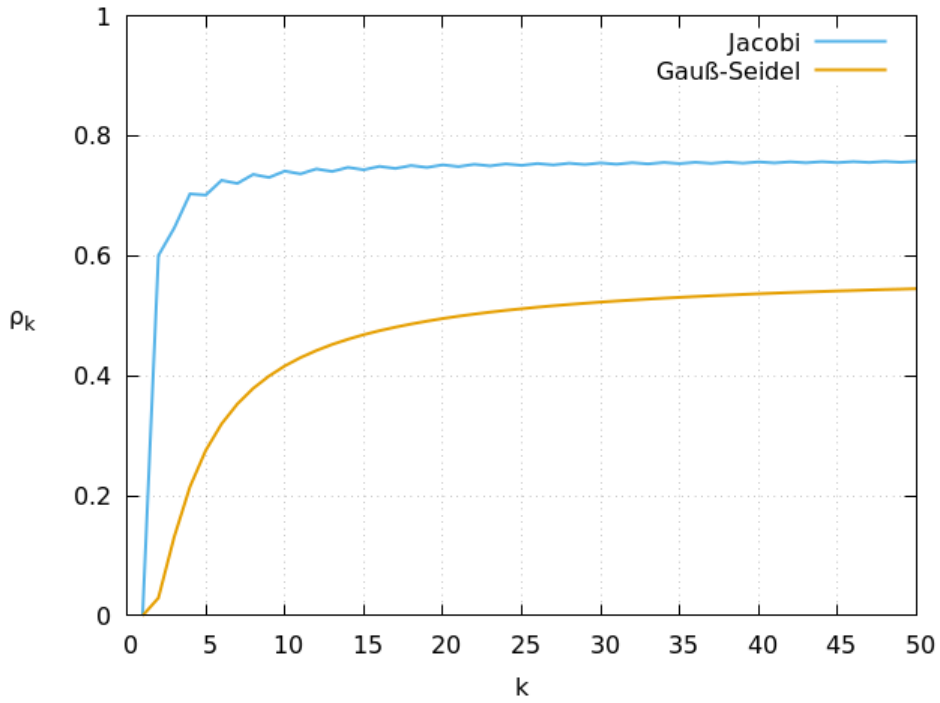

Abbildung 4.2: Annäherung an den Konvergenzfaktor

Es ist also tatsächlich deutlich zu erkennen, dass sich dieser Wert dem jeweiligen Spektralradius annähert.

5 Relaxationsverfahren

Wir werden nun noch eine Verallgemeinerung des Jacobi- und des Gauß-Seidel-Verfahrens kennenlernen, die sogenannten Relaxationsverfahren, und auch deren Konvergenz untersuchen. Der Aufbau und die Resultate dieses Kapitels beruhen auf dem entsprechenden Abschnitt in [4] von Andreas Meister.

Wieder betrachten wir zunächst eine allgemeine lineare Iteration. Schreiben wir das zu betrachtende lineare Iterationsverfahren wie in Definition 3.1 in der Form

$$x^{(m+1)} = x^{(m)} + M^{-1}\left(b - Ax^{(m)}\right) \qquad \text{für alle } m \in \mathbb{N}_0, \tag{5.1}$$

so wird ersichtlich, dass für alle $m \in \mathbb{N}_0$ der $(m+1)$-te Iterationsschritt eine Korrektur der alten Näherung $x^{(m)}$ durch den Vektor $r^{(m)} := M^{-1}\left(b - Ax^{(m)}\right)$ ist.
Ziel der von Jacobi- und Gauß-Seidel-Verfahren abgeleiteten Relaxationsverfahren ist es, die Konvergenzgeschwindigkeit durch Gewichtung des Vektors $r^{(m)}$ zu verbessern, bzw. überhaupt Konvergenz herzustellen, falls das jeweilige Ausgangsverfahren nicht konvergiert. Im Falle von Gesamtschrittverfahren wie dem Jacobi-Verfahren wird ein Relaxationsparameter $\omega \in \mathbb{R}_{>0}$ hinzugenommen, der die Gleichung (5.1) abwandelt zu

$$x^{(m+1)} = x^{(m)} + \omega M^{-1}(b - Ax^{(m)}) \qquad \text{für alle } m \in \mathbb{N}_0. \tag{5.2}$$

Äquivalent dazu erhalten wir

$$x^{(m+1)} = (I_n - \omega M^{-1}A)x^{(m)} + \omega M^{-1}b \qquad \text{für alle } m \in \mathbb{N}_0,$$

die Iterationsmatrix des Relaxationsverfahrens ist also gegeben durch

$$G(\omega) = I_n - \omega M^{-1}A.$$

Gesucht wird dann für alle $m \in \mathbb{N}_0$ eine optimale neue Näherung $x^{(m+1)}$ ausgehend von $x^{(m)}$ in Richtung $r^{(m)}$, also ein $\omega \in \mathbb{R}_{>0}$, für das der Spektralradius der neuen Iterationsmatrix $G(\omega)$ minimal ist.

Gilt $\omega < 1$, so spricht man auch von *Unterrelaxation*, bei $\omega > 1$ von *Überrelaxation*.

5.1 Jacobi-Relaxationsverfahren

Gemäß (5.1) und (5.2) erhalten wir aus der Iterationsvorschrift des Jacobi-Verfahrens in der Schreibweise

$$x^{(m+1)} = x^{(m)} + D^{-1}(b - Ax^{(m)}) \qquad \text{für alle } m \in \mathbb{N}_0$$

das *Jacobi-Relaxationsverfahren* für $\omega \in \mathbb{R}_{>0}$ und einen Startvektor $x_0 \in \mathbb{R}^n$, das gegeben ist durch

$$\begin{aligned} x^{(m+1)} &= x^{(m)} + \omega D^{-1}(b - Ax^{(m)}) \\ &= (I_n - \omega D^{-1}A)x^{(m)} + \omega D^{-1}b \qquad \text{für alle } m \in \mathbb{N}_0. \end{aligned}$$

Mit der Matrix $M_J(\omega) := \frac{1}{\omega}D$ und der Iterationsmatrix

$$G_J(\omega) := I_n - \omega D^{-1}A$$

ergibt sich also wiederum ein lineares Iterationsverfahren. Damit die Diagonalmatrix D invertierbar ist, benötigen wir hier weiterhin (3.2) als Voraussetzung.

Offensichtlich ist das herkömmliche Jacobi-Verfahren der Spezialfall dieses Relaxationsverfahrens für den Relaxationsparameter $\omega = 1$.

5.2 Konvergenz Jacobi-Relaxationsverfahren

Für das Jacobi-Relaxationsverfahren können wir folgende Aussage über die Eigenwerte der Iterationsmatrix und den optimalen Relaxationsparameter machen.

Satz 5.1. *Die Iterationsmatrix G_J des Jacobi-Verfahrens habe nur reelle Eigenwerte $\lambda_1 \leq \cdots \leq \lambda_n$ mit den zugehörigen linear unabhängigen Eigenvektoren $u_1, \ldots, u_n$, und für ihren Spektralradius gelte $\rho(G_J) < 1$.*
Dann besitzt die Iterationsmatrix $G_J(\omega)$ des Jacobi-Relaxationsverfahrens die Eigenwerte

$$\mu_i = 1 - \omega - \omega\lambda_i \qquad \textit{für } i \in \{1, \ldots, n\},$$

und für den optimalen Relaxationsparameter ω_{opt}, der einen minimalen Spektralradius von $G_J(\omega)$ garantiert, gilt

$$\omega_{opt} = \frac{2}{2 - \lambda_1 - \lambda_n}.$$

Beweis. Der folgende Beweis orientiert sich am Beweis von Satz 4.21 in [4].
Da $G_J = D^{-1}(L + R)$ gilt, erhalten wir

$$D^{-1}(L + R)u_i = \lambda_i u_i \qquad \text{für alle } i \in \{1, \ldots, n\}.$$

Es folgt für alle $i \in \{1, \ldots, n\}$ mit der Zerlegung $A = D - L - R$

$$\begin{aligned} G_J(\omega)u_i &= (I_n - \omega D^{-1}A)u_i \\ &= \big(I_n - \omega D^{-1}(D - L - R)\big)u_i \\ &= \big((1-\omega)I_n + \omega D^{-1}(L+R)\big)u_i \\ &= (1-\omega)u_i + \omega \underbrace{D^{-1}(L+R)u_i}_{=\lambda_i u_i} \\ &= (1 - \omega + \omega\lambda_i)u_i. \end{aligned}$$

Somit ist $\mu_i(\omega) = (1 - \omega + \omega\lambda_i)$ für alle $i \in \{1, \ldots, n\}$ ein Eigenwert von $G_J(\omega)$ zum Eigenvektor u_i, und da die Vektoren $u_1, \ldots, u_n$ linear unabhängig sind, existieren keine weiteren Eigenwerte. Wegen $\omega > 0$ und der Voraussetzung $\lambda_1 \leq \ldots \leq \lambda_n$ gilt weiterhin

$$\mu_1(\omega) \leq \ldots \leq \mu_n(\omega).$$

Sei nun $\omega^* \in \mathbb{R}_{>0}$ derart, dass die Bedingung

$$\mu_n(\omega^*) = -\mu_1(\omega^*) \tag{5.3}$$

erfüllt ist. Dann gilt insbesondere $\mu_1(\omega^*) < 0$ und $\mu_n(\omega^*) > 0$ (vorausgesetzt $n > 1$), da aufgrund der linearen Unabhängigkeit von $u_1, \ldots, u_n$ nicht alle Eigenwerte gleich Null sein können, sowie $\rho\big(G_J(\omega^*)\big) = |\mu_1(\omega^*)| = |\mu_n(\omega^*)|$.
Für alle ω mit $\omega > \omega^*$ folgt

$$\mu_1(\omega) = 1 - \omega + \omega\lambda_1 = 1 - \omega(1 - \lambda_1) < 1 - \omega^*(1 - \lambda_1) = 1 - \omega^* + \omega^*\lambda_1 = \mu_1(\omega^*) < 0,$$

da wegen $\rho(G_J) < 1$ insbesondere $1 - \lambda_1 > 0$ gilt. Dies impliziert

$$\rho\big(G_J(\omega)\big) \geq |\mu_1(\omega)| > |\mu_1(\omega^*)| = \rho\big(G_J(\omega^*)\big).$$

Analog erhalten wir für alle ω mit $0 < \omega < \omega^*$ wegen $1 - \lambda_n > 0$

$$\mu_n(\omega) = 1 - \omega(1 - \lambda_n) > 1 - \omega^*(1 - \lambda_n) = \mu_n(\omega^*)$$

und somit

$$\rho\big(G_J(\omega)\big) \geq |\mu_n(\omega)| > |\mu_n(\omega^*)| = \rho\big(G_J(\omega^*)\big).$$

Der Spektralradius von $G_J(\omega^*)$ ist also minimal, es gilt folglich $\omega_{opt} = \omega^*$. Aus (5.3) folgt für ω^* außerdem

$$1 - \omega^* + \omega^*\lambda_n = \mu_n(\omega^*) = -\mu_1(\omega^*) = -(1 - \omega^* + \omega^*\lambda_1)$$

und somit

$$2 = \omega^*(1 - \lambda_1 + 1 - \lambda_n) = \omega^*(2 - \lambda_1 - \lambda_n).$$

Wegen $\rho(G_J) < 1$ gilt zudem $2 - \lambda_1 - \lambda_n > 0$, wir erhalten also insgesamt

$$\omega_{opt} = \omega^* = \frac{2}{2 - \lambda_1 - \lambda_n}.$$

□

5.3 Gauß-Seidel-Relaxationsverfahren

Bei sogenannten Einzelschrittverfahren wie dem Gauß-Seidel-Verfahren ist es sinnvoll, die Relaxation bei jedem Einzelschritt zu berücksichtigen. Dafür betrachten wir die Gauß-Seidel-Iteration wieder in Komponentenschreibweise und erhalten aus (3.3) die äquivalente Formulierung

$$x_i^{(m+1)} = x_i^{(m)} + \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(m+1)} - \sum_{j=i}^{n} a_{ij}x_j^{(m)}\right) \quad \text{für alle } i \in \{1,\dots,n\},\ m \in \mathbb{N}_0.$$

Hiervon leiten wir für einen Relaxationsparameter $\omega \in \mathbb{R}_{>0}$ folgende abgewandelte Iterationsvorschrift für alle $i \in \{1,\dots,n\}$ und $m \in \mathbb{N}_0$ ab:

$$\begin{aligned} x_i^{(m+1)} &= x_i^{(m)} + \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(m+1)} - \sum_{j=i}^{n} a_{ij}x_j^{(m)}\right) \\ &= (1-\omega)x_i^{(m)} + \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(m+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(m)}\right). \end{aligned}$$

Diese lässt sich in Matrixschreibweise darstellen durch

$$x^{(m+1)} = (1-\omega)x^{(m)} + \omega D^{-1}b + \omega D^{-1}Lx^{(m+1)} + \omega D^{-1}Rx^{(m)} \quad \text{für alle } m \in \mathbb{N}_0,$$

was sich zusammenfassen lässt zu

$$(I_n - \omega D^{-1}L)x^{(m+1)} = \big((1-\omega)I_n + \omega D^{-1}R\big)x^{(m)} + \omega D^{-1}b \quad \text{für alle } m \in \mathbb{N}_0.$$

Multiplikation von links mit D liefert

$$(D - \omega L)x^{(m+1)} = \big((1-\omega)D + \omega R\big)x^{(m)} + \omega b \quad \text{für alle } m \in \mathbb{N}_0.$$

Die Matrix $D - \omega L$ ist wiederum eine untere Dreicksmatrix mit der gleichen Hauptdiagonalen wie A und somit unter der Voraussetzung (3.2) invertierbar, diese muss also auch beim Gauß-Seidel-Relaxationsverfahren erfüllt sein. Damit erhalten wir die

Iterationsvorschrift

$$x^{(m+1)} = (D - \omega L)^{-1}\big((1-\omega)D + \omega R\big)x^{(m)} + \omega(D - \omega L)^{-1}b \quad \text{für alle } m \in \mathbb{N}_0,$$

durch welche das *Gauß-Seidel-Relaxationsverfahren* für den Relaxationsparameter $\omega \in \mathbb{R}_{>0}$ definiert wird. Dieses lineare Iterationsverfahren ist somit gegeben durch die Iterationsmatrix

$$G_{GS}(\omega) := (D - \omega L)^{-1}\big((1-\omega)D + \omega R\big)$$

und die Matrix $M_{GS}(\omega) := \frac{1}{\omega}(D - \omega L)$.

Auch hier ist wie bereits beim Jacobi-Relaxationsverfahren zu erkennen, dass sich im Spezialfall $\omega = 1$ das ursprüngliche Gauß-Seidel-Verfahren ergibt.

5.4 Konvergenz Gauß-Seidel-Relaxationsverfahren

Für das Gauß-Seidel-Relaxationsverfahren lässt sich folgende allgemeine Abschätzung für den Spektralradius der Iterationsmatrix $G_{GS}(\omega)$ zeigen.

Satz 5.2. *Sind alle Diagonaleinträge von A von Null verschieden, so gilt für alle $\omega \in \mathbb{R}_{>0}$ die Ungleichung*

$$\rho\big(G_{GS}(\omega)\big) \geq |\omega - 1|.$$

Beweis. Dieser Beweis beruht auf dem Beweis von Satz 4.22 sowie den Erläuterungen zu Definition 2.30 in [4].
Seien $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ die Eigenwerte der Iterationsmatrix $G_{GS}(\omega)$, das charakteristische Polynom $\chi_{G_{GS}(\omega)}(\lambda) = \det\big(G_{GS}(\omega) - \lambda I_n\big)$ zerfällt also insbesondere über $\mathbb{C}$ in Linearfaktoren und ist von der Form

$$\chi_{G_{GS}(\omega)}(\lambda) = \prod_{i=1}^{n}(\lambda_i - \lambda),$$

da es sich bei den Eigenwerten $\lambda_1, \ldots, \lambda_n$ um dessen Nullstellen handelt. Wegen

$$\det G_{GS}(\omega) = \chi_{G_{GS}(\omega)}(0) = \prod_{i=1}^{n}\lambda_i$$

erhalten wir unter Verwendung des Determinanten-Produktsatzes

$$\begin{aligned}\prod_{i=1}^{n} \lambda_i &= \det G_{GS}(\omega) \\ &= \det\big((D-\omega L)^{-1}\big) \det\big((1-\omega)D + \omega R\big) \\ &\overset{(*)}{=} \det(D^{-1}) \det\big((1-\omega)D\big) \\ &= (\det D)^{-1}(1-\omega)^n \det D \\ &= (1-\omega)^n,\end{aligned}$$

wobei bei $(*)$ benutzt wurde, dass L und R strikte untere bzw. obere Dreiecksmatrizen sind sowie D eine Diagonalmatrix. Es folgt, dass $|\lambda_j| \geq |1-\omega|$ für mindestens ein $j \in \{1, \ldots, n\}$ gelten muss und somit

$$\rho\big(G_{GS}(\omega)\big) = \max_{i \in \{1,\ldots,n\}} |\lambda_i| \;\geq\; |\lambda_j| \;\geq\; |1-\omega| = |\omega - 1|.$$

□

Aus dem Satz folgt also, dass der Spektralradius der Iterationsmatrix $G_{GS}(\omega)$ nur dann echt kleiner als 1 sein kann, wenn $\omega \in (0, 2)$ gilt. Mit Satz 4.1 erhalten wir das folgende Korollar.

Korollar 5.3. *Wenn das Gauß-Seidel-Relaxationsverfahren konvergiert, so folgt für den Relaxationsparameter $\omega \in (0, 2)$.*

Fügen wir weitere Voraussetzungen an die Matrix A hinzu, so können wir Äquivalenz erhalten.

Satz 5.4. *Sei A symmetrisch und positiv definit. Dann konvergiert das Gauß-Seidel-Relaxationsverfahren genau dann, wenn $\omega \in (0, 2)$ gilt.*

Beweis. Der Beweis ist angelehnt an den Beweis von Satz 4.24 aus [4], wurde allerdings angepasst auf reelle Matrizen.
Aus der positiven Definitheit von A folgt für die Diagonaleinträge

$$a_{ii} > 0 \qquad \text{für alle } i \in \{1, \ldots, n\}, \tag{5.4}$$

da für die Vektoren $e_1, \ldots, e_n$ der Standardbasis gegeben durch

$$(e_i)_j = \begin{cases} 1, & \text{falls } j = i \\ 0, & \text{ansonsten} \end{cases}$$

für alle $i \in \{1, \ldots, n\}$ gilt

$$a_{ii} = {e_i}^T A e_i \overset{A \text{ positiv definit}}{>} 0.$$

Das Gauß-Seidel-Relaxationsverfahren ist demnach wohldefiniert.

Ist das Gauß-Seidel-Relaxationsverfahren konvergent, so folgt mit Korollar 5.3, dass $\omega \in (0,2)$ gilt. Gelte also $\omega \in (0,2)$. Sei λ ein beliebiger Eigenwert von $G_{GS}(\omega)$ mit zugehörigem Eigenvektor x. Da A symmetrisch ist, gilt $R = L^T$ und somit

$$G_{GS}(\omega) = (D - \omega L)^{-1}\big((1-\omega)D + \omega L^T\big).$$

Da λ Eigenwert dieser Matrix zum Eigenvektor x ist, gilt also

$$\big((1-\omega)D + \omega L^T\big)x = \lambda(D - \omega L)x. \tag{5.5}$$

Ferner gilt

$$\begin{aligned} 2\big((1-\omega)D + \omega L^T\big) &= (2-2\omega)D + 2\omega L^T \\ &= (2-\omega)D + \omega(-D + 2L^T) \\ &= (2-\omega)D + \omega(\underbrace{-D + L^T + L}_{=-A} + L^T - L) \\ &= (2-\omega)D - \omega A + \omega(L^T - L) \end{aligned}$$

und

$$\begin{aligned} 2(D - \omega L) &= (2-\omega)D + \omega(D - 2L) \\ &= (2-\omega)D + \omega(\underbrace{D - L - L^T}_{=A} - L + L^T) \\ &= (2-\omega)D + \omega A + \omega(L^T - L). \end{aligned}$$

Mit (5.5) folgt durch Multiplikation von links mit $2x^T$

$$x^T\big((2-\omega)D - \omega A + \omega(L^T - L)\big)x = \lambda x^T\big((2-\omega)D + \omega A + \omega(L^T - L)\big)x. \tag{5.6}$$

Wegen

$$\begin{aligned} x^T(L^T - L)x &= x^T L^T x - x^T L x \\ &= \sum_{j=1}^{n} x_j \left(\sum_{k=1}^{n} \ell_{jk} x_k \right) - \sum_{k=1}^{n} x_k \left(\sum_{j=1}^{n} \ell_{jk} x_j \right) \\ &= \sum_{j,k=1}^{n} x_j \ell_{jk} x_k - \sum_{k,j=1}^{n} x_k \ell_{jk} x_j \;=\; 0 \end{aligned} \tag{5.7}$$

erhalten wir

$$(2-\omega)x^T D x - \omega x^T A x = \lambda\big((2-\omega)x^T D x + \omega x^T A x\big).$$

Aufgrund von (5.4) ist neben A auch D positiv definit, es gilt also

$$d := x^T Dx > 0 \quad \text{und} \quad a := x^T Ax > 0,$$

da x als Eigenvektor nicht der Nullvektor sein kann. Unsere Gleichung ist damit von der Form

$$(2-\omega)d - \omega a = \lambda\big((2-\omega)d + \omega a\big),$$

Division durch ω liefert somit

$$\frac{2-\omega}{\omega}d - a = \lambda\left(\frac{2-\omega}{\omega}d + a\right).$$

Aufgrund der Voraussetzung $\omega \in (0,2)$ gilt $\mu := \frac{2-\omega}{\omega} > 0$ und wegen $d, a > 0$ folgt $\mu d + a > 0$, also

$$\lambda = \frac{\mu d - a}{\mu d + a}. \tag{5.8}$$

Wiederum wegen $d, a, \mu > 0$ gilt $|\mu d - a| < |\mu d + a|$ und somit

$$|\lambda| = \left|\frac{\mu d - a}{\mu d + a}\right| < 1.$$

Da λ als beliebiger Eigenwert der Matrix $G_{GS}(\omega)$ gewählt worden ist, folgt

$$\rho\big(G_{GS}(\omega)\big) < 1,$$

das Gauß-Seidel-Relaxationsverfahren konvergiert folglich nach Satz 4.1. □

Bemerkung 5.5. (vgl. Beweis von Satz 4.24 in [4])
Satz 5.4 lässt sich ebenfalls zeigen für eine Matrix $A \in \mathbb{C}^{n\times n}$ mit komplexen Einträgen, die hermitesch statt symmetrisch ist. Statt der transponierten Matrizen und Vektoren werden dann im Beweis die Adjungierten betrachtet, und bei der zu (5.7) analogen Umformung ergibt sich, dass

$$x^*(L^* - L)x = s \cdot i$$

für ein $s \in \mathbb{R}$ ist, wobei i die imaginäre Einheit bezeichnet. Im Bruch in (5.8) muss dann sowohl im Zähler als auch im Nenner der Summand $s \cdot i$ ergänzt werden. Dieser ändert allerdings nichts an der folgenden Abschätzung des Betrags (im Komplexen), da der Imaginärteil von Zähler und Nenner folglich identisch ist, der Realteil sich hingegen ebenso wie im betrachteten reellen Fall unterscheidet.

Der vorangegangene Satz garantiert uns also Konvergenz für alle $\omega \in (0,2)$, wenn A symmetrisch und positiv definit ist. Wie wir allerdings schon im vorigen Kapitel gesehen haben, interessieren wir uns neben der Konvergenz an sich auch für die Konvergenzgeschwindigkeit, also den Spektralradius der Iterationsmatrix. Für das Jacobi-

Relaxationsverfahren haben wir mit Satz 5.1 bereits Informationen über die Eigenwerte der Iterationsmatrix und schließlich über einen optimalen Relaxationsparameter erhalten. Derartige Ergebnisse möchten wir nun auch für das Gauß-Seidel-Relaxationsverfahren erlangen und müssen dafür andere Voraussetzungen an die Matrix A stellen. Beispielsweise ist die Eigenschaft einer konsistenten Ordnung wünschenswert, die wir zunächst definieren werden.

Definition 5.6. Sei $A = D - L - R$ eine additive Zerlegung der Matrix A wie bisher, insbesondere sei die Diagonalmatrix D invertierbar. A heißt *konsistent geordnet*, falls die Eigenwerte der Matrix

$$C(\alpha) = \alpha D^{-1}L + \alpha^{-1}D^{-1}R, \qquad \alpha \in \mathbb{R} \setminus \{0\},$$

unabhängig von α sind.

Beispiel 5.7. (vgl. Beispiel 4.27 in [4])
Jede Tridiagonalmatrix

$$A = \begin{pmatrix} a_1 & b_1 & & & \\ c_2 & a_2 & b_2 & & \\ & \ddots & \ddots & \ddots & \\ & & c_{n-1} & a_{n-1} & b_{n-1} \\ & & & c_n & a_n \end{pmatrix} \in \mathbb{R}^{n\times n}$$

mit $a_i \neq 0$ für alle $i \in \{1, \dots, n\}$ ist konsistent geordnet.

Beweis. Es gilt

$$C(1) = D^{-1}(L+R) = \begin{pmatrix} 0 & d_1 & & \\ e_2 & \ddots & \ddots & \\ & \ddots & \ddots & d_{n-1} \\ & & e_n & 0 \end{pmatrix}$$

mit $d_i = {}^{b_i}\!/_{a_i}$ für $i \in \{1, \dots, n-1\}$ und $e_i = {}^{c_i}\!/_{a_i}$ für $i \in \{2, \dots, n\}$. Ferner gilt für alle $\alpha \in \mathbb{R} \setminus \{0\}$

$$\begin{aligned} C(\alpha) &= \alpha D^{-1}L + \alpha^{-1}D^{-1}R \\ &= \alpha \begin{pmatrix} 0 & 0 & & \\ e_2 & \ddots & \ddots & \\ & \ddots & \ddots & 0 \\ & & e_n & 0 \end{pmatrix} + \alpha^{-1} \begin{pmatrix} 0 & d_1 & & \\ 0 & \ddots & \ddots & \\ & \ddots & \ddots & d_{n-1} \\ & & 0 & 0 \end{pmatrix}. \end{aligned}$$

Mit der invertierbaren Diagonalmatrix

$$S(\alpha) = \text{diag}\,\{1, \alpha, \alpha^2, \ldots, \alpha^{n-1}\} \in \mathbb{R}^{n\times n},$$

folgt für alle $\alpha \in \mathbb{R} \setminus \{0\}$ für die Einträge der Matrix

$$S(\alpha)C(1)\big(S(\alpha)\big)^{-1} = (\beta_{\alpha,ij})_{i,j=1,\ldots,n}$$

wegen

$$\big(S(\alpha)\big)^{-1} = \text{diag}\,\{1, \alpha^{-1}, \alpha^{-2}, \ldots, \alpha^{-(n-1)}\}$$

mit $C(1) = (\gamma_{ij})_{i,j=1,\ldots,n}$ für alle $i, j \in \{1, \ldots, n\}$

$$\beta_{\alpha,ij} = \alpha^{i-1}\alpha^{-(j-1)}\gamma_{ij} = \alpha^{i-j}\gamma_{ij} = \begin{cases} \alpha^{-1}d_i, & \text{falls } i = j+1, \\ \alpha e_i, & \text{falls } i = j-1, \\ 0, & \text{sonst.} \end{cases}$$

Es gilt somit $S(\alpha)C(1)\big(S(\alpha)\big)^{-1} = C(\alpha)$, die Matrix $C(\alpha)$ ist also ähnlich zur Matrix $C(1)$ für alle $\alpha \in \mathbb{R} \setminus \{0\}$. Demnach sind alle $C(\alpha),\ \alpha \in \mathbb{R} \setminus \{0\}$, zueinander ähnlich und haben folglich die gleichen Eigenwerte. □

Satz 5.8. *Sei A konsistent geordnet und $\omega \in (0, 2)$. Dann ist $\mu \in \mathbb{C} \setminus \{0\}$ genau dann Eigenwert von $G_{GS}(\omega)$, wenn*

$$\lambda = \frac{\mu + \omega - 1}{\omega\sqrt{\mu}}$$

ein Eigenwert der Iterationsmatrix G_J des Jacobi-Verfahrens ist.

Beweis. Der Beweis des Satzes beruht auf dem Beweis von Satz 4.28 in [4].
Sei $\mu \in \mathbb{C} \setminus \{0\}$, dann gilt

$$\begin{aligned}
&(I_n - \omega D^{-1}L)\big(\mu I_n - G_{GS}(\omega)\big) \\
&\qquad = \mu(I_n - \omega D^{-1}L) - (I_n - \omega D^{-1}L)\big(G_{GS}(\omega)\big) \\
&\qquad = \mu(I_n - \omega D^{-1}L) - \big(D^{-1}(D - \omega L)\big)\Big((D - \omega L)^{-1}\big((1-\omega)D + \omega R\big)\Big) \\
&\qquad = \mu(I_n - \omega D^{-1}L) - D^{-1}\big((1-\omega)D + \omega R\big) \\
&\qquad = \big(\mu - (1-\omega)\big)I_n - \omega D^{-1}(\mu L + R) \\
&\qquad = \big(\mu - (1-\omega)\big)I_n - \omega\sqrt{\mu}D^{-1}\left(\sqrt{\mu}L + \frac{1}{\sqrt{\mu}}R\right) \qquad (5.9)
\end{aligned}$$

Da $(I_n - \omega D^{-1}L)$ eine untere Dreiecksmatrix ist, deren Hauptdiagonale nur aus Einsen besteht, gilt $\det(I_n - \omega D^{-1}L) = 1$. Aufgrund der Äquivalenz

$$\det\big(\mu I_n - G_{GS}(\omega)\big) = 0 \quad \Leftrightarrow \quad \mu \text{ ist Eigenvektor von } G_{GS}(\omega)$$

ist mit (5.9) nach dem Determinanten-Produktsatz μ genau dann Eigenvektor von $G_{GS}(\omega)$, wenn

$$\det\left(\big(\mu-(1-\omega)\big)I_n-\omega\sqrt{\mu}D^{-1}\left(\sqrt{\mu}L+\frac{1}{\sqrt{\mu}}R\right)\right)=0$$

gilt. Dies ist äquivalent dazu, dass

$$\det\left(\frac{\mu-(1-\omega)}{\omega\sqrt{\mu}}I_n-D^{-1}\left(\sqrt{\mu}L+\frac{1}{\sqrt{\mu}}R\right)\right)=0$$

gilt, also dass

$$\lambda:=\frac{\mu-(1-\omega)}{\omega\sqrt{\mu}}$$

Eigenwert der Matrix $D^{-1}\left(\sqrt{\mu}L+\frac{1}{\sqrt{\mu}}R\right)$ ist.
Da A konsistent geordnet ist, stimmen die Eigenwerte der beiden Matrizen

$$D^{-1}\left(\sqrt{\mu}L+\frac{1}{\sqrt{\mu}}R\right)\quad\text{und}\quad D^{-1}(L+R)=G_J$$

überein. Somit ist μ genau dann Eigenwert von $G_{GS}(\omega)$, wenn λ Eigenwert von G_J ist. □

Da das Gauß-Seidel-Verfahren dem Gauß-Seidel-Relaxationsverfahren mit Relaxationsparameter $\omega = 1$ entspricht, erhalten wir aus Satz 5.8 insbesondere, dass für konsistent geordnete Ausgangsmatrizen A

$$\mu\in\mathbb{C}\setminus\{0\}\text{ ist Eigenwert von }G_{GS}\quad\Leftrightarrow\quad\frac{\mu}{\sqrt{\mu}}=\sqrt{\mu}\text{ ist Eigenwert von }G_J$$

und somit

$$\rho\big(G_{GS}\big)=\rho\big(G_J\big)^2 \tag{5.10}$$

gilt. In dem Fall benötigt das Gauß-Seidel-Verfahren für dieselbe Genauigkeit also nur etwa halb so viele Iterationen wie das Jacobi-Verfahren, sofern Letzteres konvergiert.

Wie bereits beim Jacobi-Relaxationsverfahren können wir uns nun der Frage widmen, welcher Relaxationsparameter für das Verfahren optimal ist.

Satz 5.9. *Sei A konsistent geordnet. Die Eigenwerte der Iterationsmatrix G_J des Jacobi-Verfahrens seien reell und es gelte*

$$\rho:=\rho(G_J)<1.$$

Dann gilt

(a) Das Gauß-Seidel-Relaxationsverfahren konvergiert für alle $\omega\in(0,2)$.

(b) Der Relaxationsparameter

$$\omega_{opt} = \frac{2}{1+\sqrt{1-\rho^2}}$$

liefert einen minimalen Spektralradius der Iterationsmatrix des Gauß-Seidel-Relaxationsverfahrens, für den

$$\rho\big(G_{GS}(\omega_{opt})\big) = \omega_{opt} - 1 = \frac{1-\sqrt{1-\rho^2}}{1+\sqrt{1-\rho^2}}$$

gilt.

Beweis. Der folgende Beweis beruht auf dem Beweis von Satz 4.29 aus [4].
Seien $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ die Eigenwerte von G_J, dann ist nach Satz 5.8 μ genau dann Eigenwert von $G_{GS}(\omega)$, wenn

$$\lambda = \frac{\mu+\omega-1}{\omega\sqrt{\mu}} \in \sigma(G_J) = \{\lambda_1, \dots, \lambda_n\} \tag{5.11}$$

gilt. Da A konsistent geordnet ist, sind die Eigenwerte der beiden Matrizen

$$D^{-1}(L+R) = G_J \quad \text{und} \quad D^{-1}(-L-R) = -D^{-1}(L+R) = -G_J$$

identisch, es ist also $\lambda \in \mathbb{R}$ Eigenwert von G_J genau dann, wenn $-\lambda$ Eigenwert von G_J ist. Wir können daher das Vorzeichen in (5.11) außer Acht lassen. Somit können wir ohne Beschränkung der Allgemeinheit $\lambda \geq 0$ annehmen und die Gleichung (5.11) quadrieren, sodass wir

$$\lambda^2\omega^2\mu = (\mu+\omega-1)^2 \tag{5.12}$$

erhalten. Da nach Voraussetzung $\rho(G_J) < 1$ gilt, haben wir nun $\lambda \in [0,1)$. Falls $\lambda = 0$, so erhalten wir aus (5.12), dass $\mu+\omega-1=0$ und somit $\mu = 1-\omega$ gilt. Die folgende Analyse befasst sich daher mit dem Fall, dass $\lambda \in (0,1)$ gilt.
Aus Korollar 5.3 wissen wir, dass $\omega \in (0,2)$ eine notwendige Bedingung für die Konvergenz ist. Ferner gilt

$$\begin{aligned}
\lambda^2\omega^2\mu &= (\mu+\omega-1)^2 = \mu^2+\omega^2+1+2\mu\omega-2\mu-2\omega \\
\Leftrightarrow \qquad 0 &= \mu^2 - \lambda^2\omega^2\mu + 2\mu\omega - 2\mu + \omega^2 + 1 - 2\omega \\
&= \mu^2 - (\lambda^2\omega^2 - 2\omega + 2)\mu + (\omega-1)^2 \\
&\overset{(*)}{=} \Big(\mu - \Big(\frac{1}{2}\lambda^2\omega^2 - \omega + 1\Big)\Big)^2 - \underbrace{\Big(\frac{1}{2}\lambda^2\omega^2 - \omega + 1\Big)^2}_{=\frac{1}{4}\lambda^4\omega^4 - \lambda^2\omega^2(\omega-1) + (\omega-1)^2} + (\omega-1)^2 \\
&= \Big(\mu - \Big(\frac{1}{2}\lambda^2\omega^2 - (\omega-1)\Big)\Big)^2 - \Big(\lambda^2\omega^2\Big(\frac{1}{4}\lambda^2\omega^2 - (\omega-1)\Big)\Big),
\end{aligned}$$

wobei bei $(*)$ quadratische Ergänzung angewandt wurde, wir erhalten somit für alle $\omega \in (0,2)$ und $\lambda \in \{\lambda_1, \dots, \lambda_n\} \cap (0,1)$ die beiden Eigenwerte

$$\mu^+ = \mu^+(\omega,\lambda) = \frac{1}{2}\lambda^2\omega^2 - (\omega-1) + \lambda\omega\sqrt{\frac{1}{4}\lambda^2\omega^2 - (\omega-1)} \quad \text{und}$$
$$\mu^- = \mu^-(\omega,\lambda) = \frac{1}{2}\lambda^2\omega^2 - (\omega-1) - \lambda\omega\sqrt{\frac{1}{4}\lambda^2\omega^2 - (\omega-1)}.$$

Definiere $g(\omega,\lambda) := \frac{1}{4}\lambda^2\omega^2 - (\omega-1)$ für alle $\omega \in (0,2)$ und $\lambda \in (0,1)$, es gilt also

$$\mu^\pm(\omega,\lambda) = \frac{1}{2}\lambda^2\omega^2 - (\omega-1) \pm \lambda\omega\sqrt{g(\omega,\lambda)}.$$

Damit liegt es nahe, die Nullstellen von g bezüglich Minimalität der Eigenwerte zu überprüfen. Da $g(\omega,\lambda) = 0$ genau dann gilt, wenn

$$0 = \omega^2 - \frac{4}{\lambda^2}\omega + \frac{4}{\lambda^2} = \Big(\omega - \frac{2}{\lambda^2}\Big)^2 - \frac{4}{\lambda^4} + \frac{4}{\lambda^2} = \Big(\omega - \frac{2}{\lambda^2}\Big)^2 - \frac{4}{\lambda^4}\big(1-\lambda^2\big),$$

hat g für alle $\lambda \in (0,1)$ Nullstellen bei

$$\begin{aligned}
\omega^+ = \omega^+(\lambda) &= \frac{2}{\lambda^2} - \frac{2}{\lambda^2}\sqrt{1-\lambda^2} \\
&= \frac{2(1-\sqrt{1-\lambda^2})}{\lambda^2} \\
&= \frac{2(1-(1-\lambda^2))}{\lambda^2(1+\sqrt{1-\lambda^2})} \\
&= \frac{2\lambda^2}{\lambda^2(1+\sqrt{1-\lambda^2})} \\
&= \frac{2}{1+\sqrt{1-\lambda^2}}
\end{aligned} \tag{5.13}$$

sowie (mit analoger Rechnung) bei

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

Wegen $\lambda \in (0,1)$ gilt $\sqrt{1-\lambda^2} \in (0,1)$ und somit

$$1+\sqrt{1-\lambda^2} \in (1,2) \quad \text{und} \quad 1-\sqrt{1-\lambda^2} \in (0,1).$$

Es folgt für alle $\lambda \in (0,1)$

$$\omega^+(\lambda) = \frac{2}{1+\sqrt{1-\lambda^2}} \in (1,2), \qquad \omega^-(\lambda) = \frac{2}{1-\sqrt{1-\lambda^2}} > 2. \tag{5.14}$$

Da wir nur $\omega \in (0,2)$ betrachten, können wir somit ω^- vernachlässigen und fahren nur

mit ω^+ fort.
Weiterhin gilt für alle $\lambda \in (0,1)$ und $\omega \in (0,2)$

$$\frac{\partial g}{\partial \omega}(\omega,\lambda) = \frac{1}{2}\lambda^2\omega - 1 < 0,$$

g ist also streng monoton fallend bezüglich ω auf $(0,2)$. Da für ein festes λ nach vorangegangenen Beobachtungen $\omega^+(\lambda)$ die einzige Nullstelle von g für $\omega \in (0,2)$ ist, folgt also für $\omega \in (0,2)$ für ein festes λ

$$g(\omega,\lambda) \begin{cases} > 0, & \text{falls } 0 < \omega < \omega^+(\lambda), \\ = 0 & \text{falls } \omega = \omega^+(\lambda), \\ < 0, & \text{falls } \omega^+(\lambda) < \omega < 2. \end{cases}$$

Für jedes λ ergeben sich somit für $\omega \in (0,2)$ die folgenden drei Fälle:

1. $\omega^+(\lambda) < \omega < 2$:
   Dann gilt $g(\omega,\lambda) < 0$ und somit

   $$\sqrt{g(\omega,\lambda)} = \underbrace{\sqrt{-g(\omega,\lambda)}}_{\in\mathbb{R}}\, i,$$

   die beiden Eigenwerte $\mu^+(\omega,\lambda)$ und $\mu^-(\omega,\lambda)$ sind also komplex mit

   $$\mu^{\pm}(\omega,\lambda) = \underbrace{\frac{1}{2}\lambda^2\omega^2 + (\omega-1)}_{\text{Realteil}} + \underbrace{\big(\pm\lambda\omega\sqrt{-g(\omega,\lambda)}\big)}_{\text{Imaginärteil}}\, i.$$

   Folglich gilt nach Definition des Betrags im Komplexen

   $$\begin{aligned} |\mu^+(\omega,\lambda)| = |\mu^-(\omega,\lambda)| &= \sqrt{\Big(\frac{1}{2}\lambda^2\omega^2 + (\omega-1)\Big)^2 + (\lambda\omega\sqrt{-g(\omega,\lambda)})^2} \\ &= \sqrt{\frac{1}{4}\lambda^4\omega^4 - \lambda^2\omega^2(\omega-1) + (\omega-1)^2 + \lambda^2\omega^2\Big(-\frac{1}{4}\lambda^2\omega^2 + (\omega-1)\Big)} \\ &= \sqrt{\frac{1}{4}\lambda^4\omega^4 - \lambda^2\omega^2(\omega-1) + (\omega-1)^2 - \frac{1}{4}\lambda^4\omega^4 + \lambda^2\omega^2(\omega-1)} \\ &= \sqrt{(\omega-1)^2} \\ &= |\omega-1| \\ &\overset{(*)}{=} \omega - 1, \end{aligned}$$

   wobei $(*)$ wegen $\omega > \omega^+(\lambda) \overset{(5.14)}{>} 1$ gilt.

2. $\omega = \omega^+(\lambda)$:
   Es folgt also wegen

   $$g(\omega,\lambda) = \frac{1}{4}\lambda^2\omega^2 - (\omega-1) = 0, \tag{5.15}$$

dass $\lambda^2\omega^2 = 4(\omega - 1)$ gilt und somit

$$\lambda^2 = \frac{4}{\omega} - \frac{4}{\omega^2}. \tag{5.16}$$

Wir erhalten wiederum mit (5.15)

$$\begin{aligned} |\mu^+(\omega,\lambda)| = |\mu^-(\omega,\lambda)| &= \left|\tfrac{1}{2}\lambda^2\omega^2 - (\omega - 1)\right| \\ &\overset{(5.16)}{=} |2\omega - 2 - \omega + 1| \\ &= |\omega - 1| \\ &= \omega - 1. \end{aligned}$$

Die letzte Gleichheit gilt dabei wegen $\omega = \omega^+(\lambda) \overset{(5.14)}{>} 1$.

3. $0 < \omega < \omega^+(\lambda)$: In diesem Fall gilt $g(\omega,\lambda) > 0$, also folgt wegen

$$\mu^\pm = \underbrace{\frac{1}{4}\lambda^2\omega^2 + g(\omega,\lambda)}_{>0} \pm \underbrace{\lambda\omega\sqrt{g(\omega,\lambda)}}_{>0},$$

dass $\mu^+(\omega,\lambda) > 0$ sowie $\mu^+(\omega,\lambda) = |\mu^+(\omega,\lambda)| > |\mu^-(\omega,\lambda)|$ gilt.

In jedem der drei Fälle gilt also

$$|\mu^+(\omega,\lambda)| \geq |\mu^-(\omega,\lambda)|,$$

zur Bestimmung des Spektralradius $\rho\big(G_{GS}(\omega)\big)$ ist für uns somit jeweils nur $\mu^+(\omega,\lambda)$ relevant. Für $\lambda \in (0,1)$ und $\omega \in (0,2)$ betrachten wir daher

$$\mu(\omega,\lambda) := |\mu^+(\omega,\lambda)| = \begin{cases} \mu^+(\omega,\lambda) & \text{für } 0 < \omega < \omega^+(\lambda), \\ \omega - 1 & \text{für } \omega^+(\lambda) \leq \omega < 2. \end{cases} \tag{5.17}$$

Offenbar ist μ für $\omega \in [\omega^+(\lambda), 2)$ für ein fest gewähltes λ streng monoton wachsend. Um zu zeigen, dass $\omega^+(\lambda)$ ein Minimum im Intervall $(0,2)$ darstellt, bleibt also noch zu zeigen, dass μ für ein festes λ für $\omega \in (0, \omega^+(\lambda))$ strikt monoton fällt. Es gilt für alle $\lambda \in (0,1)$ und $\omega \in (0, \omega^+(\lambda))$

$$\begin{aligned} \left(\frac{1}{2}\lambda\omega + \sqrt{\frac{1}{4}\lambda^2\omega^2 - (\omega - 1)}\right)^2 &= \frac{1}{4}\lambda^2\omega^2 + \lambda\omega\sqrt{\frac{1}{4}\lambda^2\omega^2 - (\omega - 1)} + \frac{1}{4}\lambda^2\omega^2 - (\omega - 1) \\ &= \frac{1}{2}\lambda^2\omega^2 - (\omega - 1) + \lambda\omega\sqrt{\frac{1}{4}\lambda^2\omega^2 - (\omega - 1)} \\ &= \mu(\omega,\lambda) \end{aligned}$$

und somit nach der Kettenregel

$$\frac{\partial \mu}{\partial \omega}(\omega,\lambda) = \underbrace{2\left(\frac{1}{2}\lambda\omega + \sqrt{\frac{1}{4}\lambda^2\omega^2 - (\omega-1)}\right)}_{>0} \underbrace{\left(\frac{\lambda}{2} + \frac{1}{2}\frac{\frac{1}{2}\lambda^2\omega - 1}{\sqrt{\frac{1}{4}\lambda^2\omega^2 - (\omega-1)}}\right)}_{=:q(\omega,\lambda)}. \tag{5.18}$$

Indem wir den Faktor $\frac{1}{2\sqrt{g(\omega,\lambda)}}$ ausklammern, erhalten wir

$$q(\omega,\lambda) = \frac{1}{2\sqrt{\frac{1}{4}\lambda^2\omega^2 - (\omega-1)}}\left(\lambda\sqrt{\frac{1}{4}\lambda^2\omega^2 - (\omega-1)} + \frac{1}{2}\lambda^2\omega - 1\right),$$

mit

$$q_1(\omega,\lambda) := \lambda\sqrt{\frac{1}{4}\lambda^2\omega^2 - (\omega-1)} \quad \text{und} \quad q_2(\omega,\lambda) := \frac{1}{2}\lambda^2\omega - 1$$

für alle $\lambda \in (0,1)$ und $\omega \in (0,\omega^+(\lambda))$ erhalten wir also

$$q(\omega,\lambda) = \frac{1}{2\sqrt{g(\omega,\lambda)}}\left(q_1(\omega,\lambda) + q_2(\omega,\lambda)\right).$$

Für alle $\lambda \in (0,1)$ und $\omega \in (0,\omega^+(\lambda))$ gilt wegen $\frac{\omega^+(\lambda)}{2} < 1$

$$q_1(\omega,\lambda) = \lambda\sqrt{\frac{1}{4}\lambda^2\omega^2 - (\omega-1)} > 0 \quad \text{und} \quad q_2(\omega,\lambda) = \frac{1}{2}\underbrace{\lambda^2\omega}_{\in(0,\omega^+(\lambda))} - 1 < 0$$

sowie wegen $\lambda^2 < 1$

$$\big(q_1(\omega,\lambda)\big)^2 = \frac{1}{4}\lambda^4\omega^2 - \lambda^2\omega + \lambda^2 < \frac{1}{4}\lambda^4\omega^2 - \lambda^2\omega + 1 = \big(q_2(\omega,\lambda)\big)^2.$$

Wir erhalten also

$$|q_1(\omega,\lambda)| < |q_2(\omega,\lambda)|$$

und somit

$$q(\omega,\lambda) = \underbrace{\frac{1}{2\sqrt{g(\omega,\lambda)}}}_{>0}\underbrace{\left(q_1(\omega,\lambda) + q_2(\omega,\lambda)\right)}_{<0} < 0.$$

Insgesamt ergibt sich also mit (5.18) die Ungleichung

$$\frac{\partial \mu}{\partial \omega}(\omega,\lambda) < 0 \quad \text{für alle } \lambda \in (0,1),\ \omega \in (0,\omega^+(\lambda)).$$

Außerdem erhalten wir aus (5.17)

$$\lim_{\omega \searrow 0} \mu(\omega, \lambda) = \lim_{\omega \searrow 0} \mu^+(\omega, \lambda) = 1 = \lim_{\omega \nearrow 2}(\omega - 1) = \lim_{\omega \nearrow 2} \mu(\omega, \lambda),$$

folglich gilt $\mu(\omega, \lambda) < 1$ für alle $\lambda \in (0, 1)$ und $\omega \in (0, 2)$. Da nach Definition von μ gilt, dass $\mu(\omega, \lambda) \geq 0$ für alle $\lambda \in (0, 1)$ und $\omega \in (0, 2)$, folgt mit den vorangegangenen Überlegungen $\rho\big(G_{GS}(\omega)\big) < 1$. Nach Satz 4.1 gilt somit Behauptung (a).
Ferner wird wie bereits gezeigt für jeden Eigenwert λ der Wert $|\mu(\omega, \lambda)| = \mu(\omega, \lambda)$ minimal für $\omega_{opt} = \omega^+(\lambda)$.
Da außerdem für $0 < \omega < \omega^+(\lambda)$ und $\lambda \in (0, 1)$ mit Produkt- und Kettenregel

$$\begin{aligned}\frac{\partial \mu}{\partial \lambda}(\omega, \lambda) &= \underbrace{\lambda\omega^2}_{>0} + \omega\sqrt{\frac{1}{4}\lambda^2\omega^2 - (\omega - 1)} + \lambda\omega\left(\frac{1}{2}\lambda\omega^2 \cdot \frac{1}{2}\Big(\frac{1}{4}\lambda^2\omega^2 - (\omega - 1)\Big)^{-1/2}\right) \\ &= \underbrace{\lambda\omega^2}_{>0} + \underbrace{\omega\sqrt{\frac{1}{4}\lambda^2\omega^2 - (\omega - 1)}}_{>0} + \underbrace{\frac{\lambda^2\omega^3}{4\sqrt{\frac{1}{4}\lambda^2\omega^2 - (\omega - 1)}}}_{>0} > 0 \qquad (5.19)\end{aligned}$$

gilt, wird μ für ein festes ω maximal für den betraglich größten Eigenwert λ_{max} mit $\lambda_{max} \overset{\lambda \in [0,1)}{=} |\lambda_{max}| = \rho(G_J) = \rho$. Wir erhalten also insgesamt

$$\begin{aligned}\rho\big(G_{GS}(\omega_{opt})\big) &= \big|\mu\big(\omega_{opt}, \rho(G_J)\big)\big| \\ &\overset{(5.17)}{=} \omega^+(\rho) - 1 \\ &\overset{(5.13)}{=} \frac{2}{1 + \sqrt{1 - \rho^2}} - 1 \\ &= \frac{2 - (1 + \sqrt{1 - \rho^2})}{1 + \sqrt{1 - \rho^2}} \\ &= \frac{1 - \sqrt{1 - \rho^2}}{1 + \sqrt{1 - \rho^2}}.\end{aligned}$$

□

Wie im Beweis an (5.14) zu sehen ist, erhalten wir unter der Voraussetzung der konsistenten Ordnung immer einen optimalen Relaxationsparameter $\omega_{opt} \in (1, 2)$. In dem Fall kommt es also zu einer Überrelaxation, das Verfahren ist daher auch bekannt unter der Bezeichnung *SOR-Methode (successive overrelaxation method).*

Bemerkung 5.10. (vgl. Bemerkung nach Satz 4.28 in [4])
Für alle $\lambda \in (0, 1)$ gilt

$$\lim_{\omega \searrow \omega_{opt}} \frac{\partial \mu}{\partial \omega}(\omega, \lambda) = \lim_{\omega \searrow \omega_{opt}} \frac{\partial}{\partial \omega}(\omega - 1) = 1$$

und mit (5.18)

$$\begin{aligned}\lim_{\omega\nearrow\omega_{opt}}\frac{\partial\mu}{\partial\omega}(\omega,\lambda) &= \lim_{\omega\nearrow\omega_{opt}}\left[\left(\frac{\omega\lambda}{2}+\sqrt{\frac{1}{4}\lambda^2\omega^2-(\omega-1)}\right)\left(\lambda+\frac{\frac{1}{2}\lambda^2\omega-1}{\sqrt{\frac{1}{4}\lambda^2\omega^2-(\omega-1)}}\right)\right]\\ &= \lim_{\omega\nearrow\omega_{opt}}\left[\frac{\omega\lambda^2}{2}+\frac{\omega\lambda\left(\frac{1}{2}\lambda^2\omega-1\right)}{2\sqrt{g(\omega,\lambda)}}+\lambda\sqrt{g(\omega,\lambda)}+\left(\frac{1}{2}\lambda^2\omega-1\right)\right]\\ &= \lim_{\omega\nearrow\omega_{opt}}\left[\omega\lambda^2+\frac{\omega\lambda\left(\frac{1}{2}\lambda^2\omega-1\right)}{2\sqrt{g(\omega,\lambda)}}+\lambda\sqrt{g(\omega,\lambda)}-1\right].\end{aligned}$$

Da zudem

$$\begin{aligned}\lim_{\omega\nearrow\omega_{opt}}\omega\lambda^2 &= \omega_{opt}\lambda^2=\frac{2\lambda^2}{1+\sqrt{1-\rho^2}} \in (0,2),\\ \lim_{\omega\nearrow\omega_{opt}}\lambda\sqrt{g(\omega,\lambda)} &= \lambda\sqrt{g(\omega_{opt},\lambda)}=0,\\ \lim_{\omega\nearrow\omega_{opt}}\frac{\omega\lambda\left(\frac{1}{2}\lambda^2\omega-1\right)}{2} &= \frac{\omega_{opt}\lambda\left(\frac{1}{2}\lambda^2\omega_{opt}-1\right)}{2}\\ &= \frac{\lambda}{1+\sqrt{1-\rho^2}}\left(\frac{\lambda^2}{1+\sqrt{1-\rho^2}}-1\right)\\ &= \frac{\lambda^3}{\left(1+\sqrt{1-\rho^2}\right)^2}-\frac{\lambda}{1+\sqrt{1-\rho^2}}\\ &= \frac{\lambda^3-\lambda\left(1+\sqrt{1-\rho^2}\right)}{\left(1+\sqrt{1-\rho^2}\right)^2}\\ &= \frac{\lambda^3-\lambda-\lambda\sqrt{1-\rho^2}}{\left(1+\sqrt{1-\rho^2}\right)^2} \overset{\lambda\in(0,1)}{\in} (-\infty,0)\end{aligned}$$

gilt, folgt wegen $g(\omega_{opt},\lambda)=0$

$$\lim_{\omega\nearrow\omega_{opt}}\frac{\partial\mu}{\partial\omega}(\omega,\lambda)=-\infty.$$

Es sollte also im Zweifelsfall eher $\omega>\omega_{opt}$ als $\omega<\omega_{opt}$ gewählt werden.

6 Beispiele und Fazit

6.1 Vergleichende Beispiele

Zum Abschluss möchten wir noch an zwei Beispielen die verschiedenen betrachteten Verfahren bezüglich ihrer Konvergenz vergleichen. Zunächst kommen wir dabei auf unser anfängliches Minimalbeispiel aus Kapitel 4 zurück und ergänzen dies mithilfe der soeben erhaltenen Ergebnisse für die Relaxationsverfahren.

Beispiel 6.1. Betrachten wir erneut das lineare Gleichungssystem aus Beispiel 4.9, so erhalten wir aus den reellen Eigenwerten der Iterationsmatrix G_J des Jacobi-Verfahrens

$$\lambda_1 = -\frac{1}{4}\sqrt{\frac{139}{15}},\ \lambda_2 = 0,\ \lambda_3 = \frac{1}{4}\sqrt{\frac{139}{15}} \quad \text{mit } \lambda_1 \leq \lambda_2 \leq \lambda_3,$$

dass nach Satz 5.1 der optimale Relaxationsparameter für das Jacobi-Relaxationsverfahren

$$\omega_{opt,J} = \frac{2}{2-\lambda_1-\lambda_3} = \frac{2}{2-\left(-\frac{1}{4}\sqrt{\frac{139}{15}}\right)-\left(\frac{1}{4}\sqrt{\frac{139}{15}}\right)} = \frac{2}{2} = 1$$

ist, das Jacobi-Verfahren ist also bereits optimal gewichtet.
Da es sich bei der Matrix A um eine Tridiagonalmatrix handelt, ist diese nach Beispiel 5.7 konsistent geordnet. Mit Satz 5.9 folgt also, dass der optimale Relaxationsparameter für die SOR-Methode durch

$$\omega_{opt,SOR} = \frac{2}{1+\sqrt{1-\rho(G_J)^2}} = \frac{2}{1+\sqrt{1-\frac{1}{16}\cdot\frac{139}{15}}} = \frac{2}{1+\frac{1}{4}\sqrt{\frac{101}{15}}} \approx 1.213065$$

gegeben ist. Wegen Bemerkung 5.10 wählen wir $\omega = 1.22$ und erhalten den in Abbildung 6.1 abgebildeten Fehlerverlauf vom Gauß-Seidel- und SOR-Verfahren im Vergleich, wieder vom Startvektor $x^{(0)} = (10, 10, 10)^T$ ausgehend.

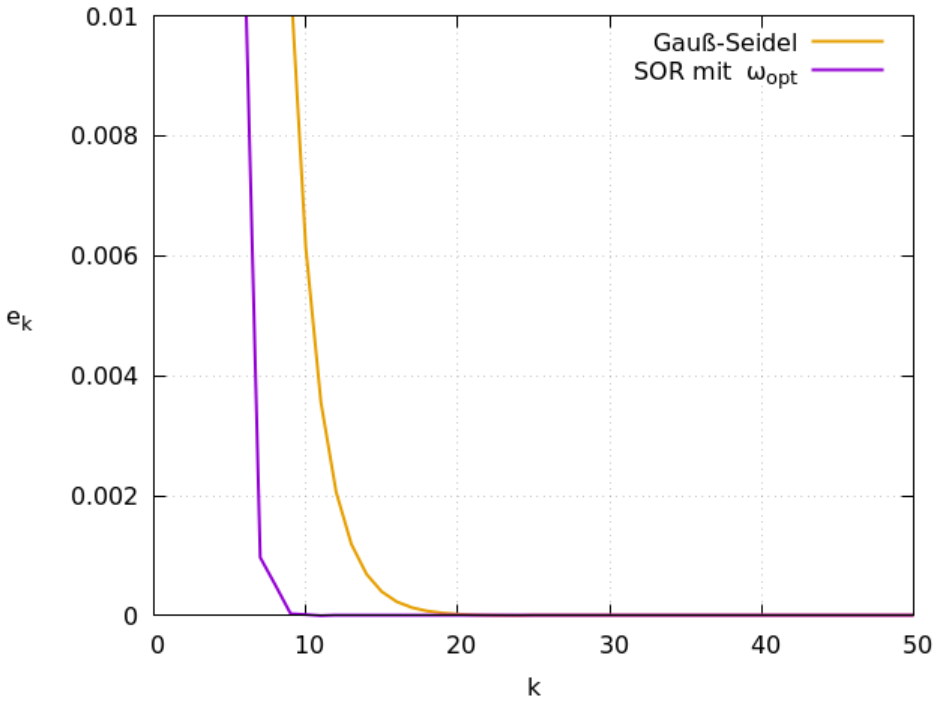

Abbildung 6.1: Fehler von Gauß-Seidel-Verfahren und SOR-Methode

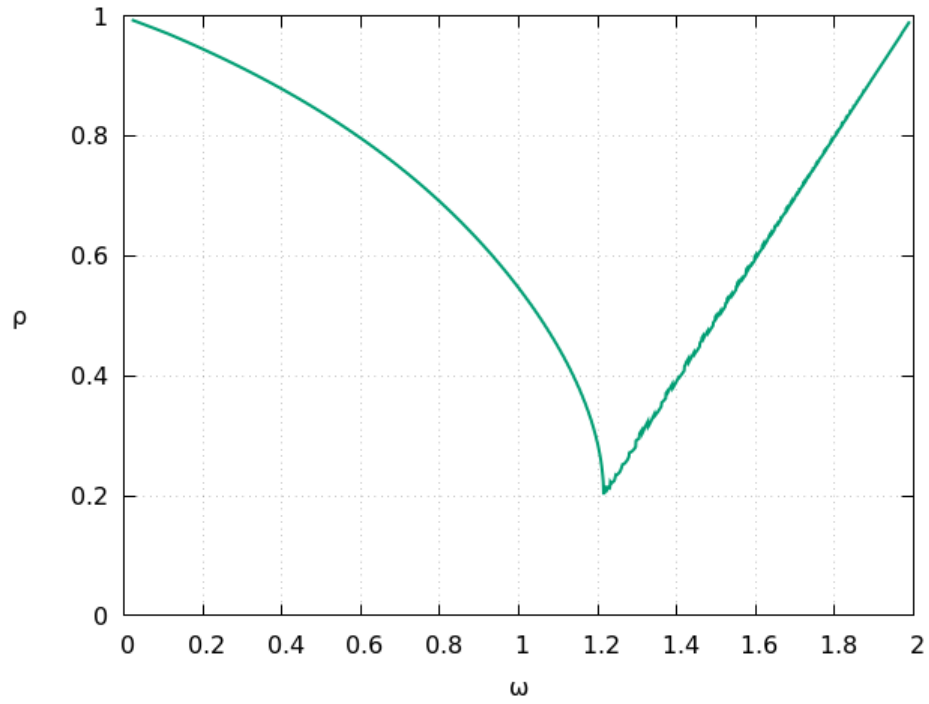

Abbildung 6.2: Abhängigkeit des Konvergenzfaktors ρ vom Relaxationsparameter ω

Abbildung 6.2 zeigt außerdem den Zusammenhang vom Relaxationsparameter ω mit dem zugehörigen Konvergenzfaktor ρ des Gauß-Seidel-Relaxationsverfahrens. Da-

bei wurde für ρ als Näherung

$$\rho = \lim_{k\to\infty} \left(\frac{\|d_k\|_2}{\|d_0\|_2} \right)^{1/k} \approx \left(\frac{\|d_m\|_2}{\|d_0\|_2} \right)^{1/m}$$

verwendet, wobei m die jeweilige Anzahl an Iterationsschritten bezeichnet, die benötigt wurde, um die Fehlerschranke von 10^{-12} zu unterschreiten.

Das bisher betrachtete Beispiel entspricht offenbar nicht dem Muster einer typischen Anwendung der betrachteten Iterationsverfahren, da diese sich, wie bereits in der Einleitung erwähnt, hauptsächlich für sehr große und schwach besetzte Systeme eignen. Ein solches werden wir uns daher nun auch noch ansehen.

Beispiel 6.2. Für alle $n \in \mathbb{N}$ und $h := \frac{1}{n+1}$ betrachten wir die Matrix

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

die durch Diskretisierung der eindimensionalen *Potentialgleichung*, auch bekannt als *Poisson-Gleichung*, durch ein Gitter mit n inneren Punkten entsteht. Die aus L hergeleitete Iterationsmatrix des Jacobi-Verfahrens lautet also

$$G_J = I - \frac{h^2}{2} I \cdot L = \begin{pmatrix} 0 & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{2} & 0 & \frac{1}{2} \\ & & & \frac{1}{2} & 0 \end{pmatrix}. \tag{6.1}$$

Diese hat die Eigenwerte

$$\lambda_{h,\nu} = \cos(\pi\nu h), \qquad \nu = 1, \ldots, n$$

mit den zugehörigen Eigenvektoren

$$e_{h,\nu} = \big(\sin(\pi\nu h), \sin(2\pi\nu h), \sin(3\pi\nu h), \ldots, \sin(n\pi\nu h)\big)^T, \quad \nu = 1, \ldots, n,$$

denn mit der aus der Analysis bekannten trigonometrischen Gleichung

$$\sin(\alpha+\beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta) \quad \text{für alle } \alpha, \beta \in \mathbb{R} \tag{6.2}$$

erhalten wir für alle $x \in \mathbb{R}$ und $k \in \mathbb{N}$

$$\begin{aligned}
&\sin\big((k-1)x\big) + \sin\big((k+1)x\big) \\
&\quad = \sin(kx - x) + \sin(kx + x) \\
&\quad \overset{(6.2)}{=} \sin(kx)\underbrace{\cos(-x)}_{=\cos(x)} + \cos(kx)\underbrace{\sin(-x)}_{=-\sin(x)} + \sin(kx)\cos(x) + \cos(kx)\sin(x) \\
&\quad = 2\sin(kx)\cos(x) \qquad (6.3)
\end{aligned}$$

und somit wegen

$$\sin(0 \cdot \pi\nu h) = \sin(0) = 0 = \sin(\pi\nu) = \sin\big((n+1)\pi\nu h\big) \qquad (6.4)$$

für alle $\nu \in \{1, \ldots, n\}$ mit (6.1)

$$\begin{aligned}
\left(G_J \cdot e_{h,\nu}\right)_k &\overset{(6.4)}{=} \frac{1}{2}\big(\sin((k-1)\pi\nu h) + \sin((k+1)\pi\nu h)\big) \\
&\overset{(6.3)}{=} \sin(k\pi\nu h)\cos(\pi\nu h) \\
&= \lambda_{h,\nu}\left(e_{h,\nu}\right)_k \qquad \text{für alle } k \in \{1, \ldots, n\}.
\end{aligned}$$

Da der Kosinus im Intervall $(0, 1)$ monoton fällt, ergeben sich die maximalen und minimalen Eigenwerte durch

$$\begin{aligned}
\lambda_{h,\max} = \lambda_{h,1} = \cos(h\pi) = \cos\left(\frac{\pi}{n+1}\right) \quad \text{und} \\
\lambda_{h,\min} = \lambda_{h,n} = \cos(nh\pi) = \cos\left(\frac{n\pi}{n+1}\right) = -\lambda_{h,\max}. \qquad (6.5)
\end{aligned}$$

Wir erhalten den Spektralradius $\rho(G_J) = |\lambda_{h,\min}| = |\lambda_{h,\max}| = \lambda_{h,\max} = \cos\left(\frac{\pi}{n+1}\right)$ und somit

$$\rho(G_J) \xrightarrow{n\to\infty} 1.$$

Je feiner das Gitter gewählt wird, desto mehr stagniert somit die Iteration. Da wir es erneut mit einer Tridiagonalmatrix und somit konsistent geordneten Matrix zu tun haben, erhalten wir mit (5.10) durch ersatzweise Anwendung des Gauß-Seidel-Verfahrens immerhin eine Quadrierung des Spektralradius.
Es bleibt also zu hoffen, dass wir mithilfe der Relaxationsverfahren eine bessere Konvergenz erzielen können. Aus (6.5) erhalten wir wie im vorangegangenen Beispiel, dass das Jacobi-Verfahren bereits optimal gewichtet ist und wir durch Relaxation keine Besserung erzielen können. Für das SOR-Verfahren ergibt sich mit Satz 5.9 der optimale Relaxationsparameter

$$\omega_{opt,SOR} = \frac{2}{1 + \sqrt{1 - \cos^2\left(\frac{\pi}{n+1}\right)}},$$

der einen Spektralradius von $\rho\big(G_{GS}(\omega_{opt,SOR})\big) = \omega_{opt,SOR} - 1$ bewirkt. In Abbildung 6.3 ist zu sehen, dass sich $\omega_{opt,SOR}$ für $n \to \infty$ dem Wert 2 annähert, der zugehörige Spektralradius tendiert folglich gegen 1.

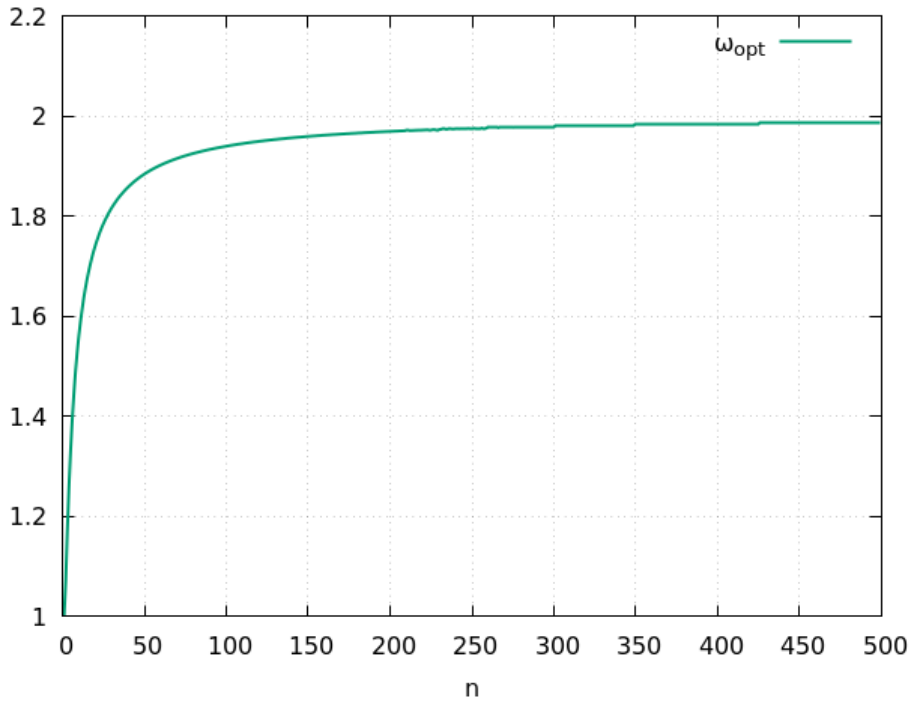

Abbildung 6.3: $\omega_{opt,SOR}$ in Abhängigkeit von n

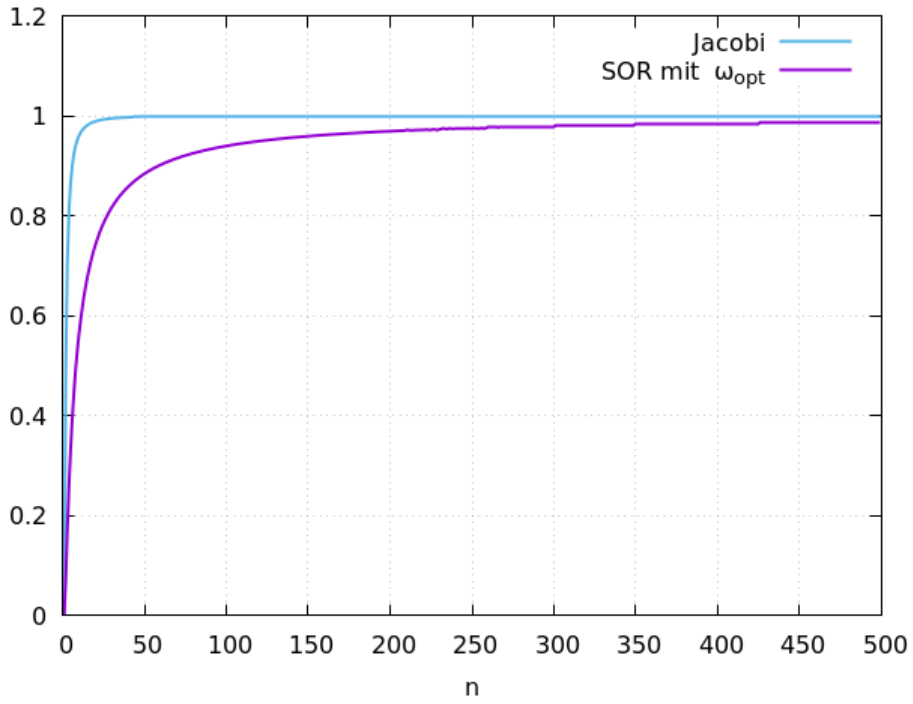

Abbildung 6.4: Spektralradien vom Jacobi-Verfahren und der SOR-Methode im Vergleich

Dass dies jedoch deutlich langsamer als beim Spektralradius

$$\rho(G_J) = \cos\left(\frac{\pi}{n+1}\right)$$

der Iterationsmatrix des Jacobi-Verfahrens geschieht, ist Abbildung 6.4 zu entnehmen, die die Spektralradien der beiden Verfahren im Zusammenhang mit der Anzahl n an inneren Punkten miteinander vergleicht.

Als konkretes Beispiel wollen wir nun den (immer noch recht überschaubaren) Fall $n = 50$ betrachten. Dann haben wir

$$\omega_{opt,SOR} = \frac{2}{1 + \sqrt{1 - \cos^2\left(\frac{\pi}{51}\right)}} \approx 1.884018$$

und somit einen minimalen Spektralradius von näherungsweise 0.884018. Wir wählen als rechte Seite $b = (\frac{1}{h^2}, 0, \ldots, 0, \frac{1}{h^2})^T$, die Lösung ist somit $x = (1, \ldots, 1)^T$.
Wenn wir auch hier beim SOR-Verfahren für die verschiedenen ω-Werte wie beim vorigen Beispiel eine Näherung des Konvergenzfaktors berechnen, erhalten wir den in Abbildung 6.5 gezeigten Verlauf.
Der Fehlerverlauf für die Größe $n = 50$ der drei Verfahren im Vergleich ist der Abbildung 6.6 zu entnehmen, diesmal mit dem Startvektor $x^{(0)} = (10, \ldots, 10)^T$.

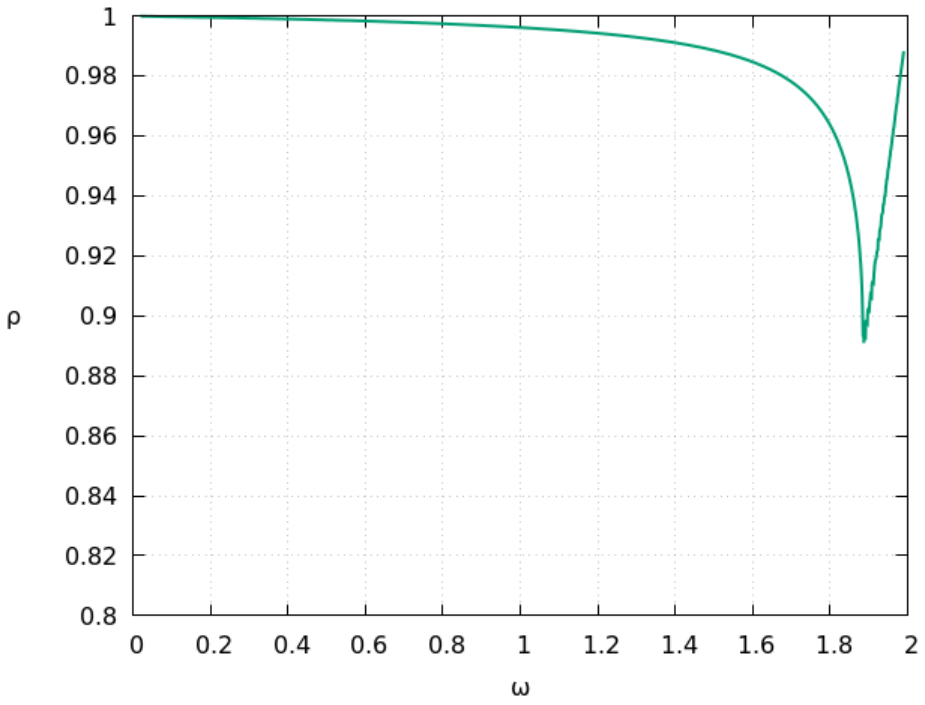

Abbildung 6.5: Abhängigkeit des Konvergenzfaktors ρ vom Relaxationsparameter ω

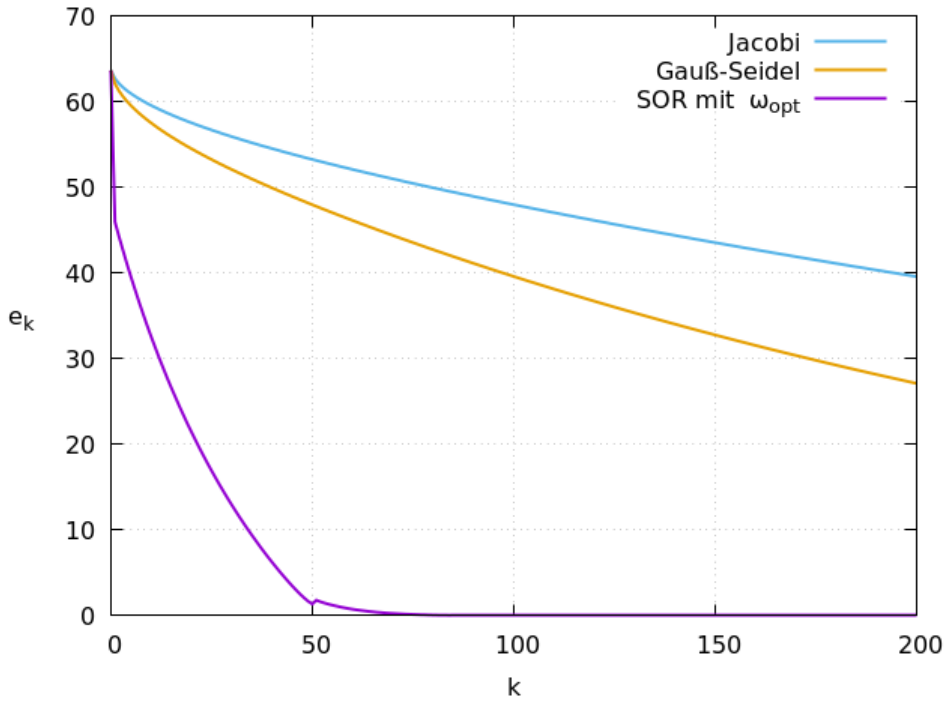

Abbildung 6.6: Fehler des Jacobi- und Gauß-Seidel-Verfahrens im Vergleich mit der SOR-Methode

Hierbei ist also schon ein deutlicher Unterschied zwischen den beiden herkömmlichen Verfahren und dem optimierten SOR-Verfahren erkennbar, der vor allem bei einem wie im Beispiel recht großen Anfangsfehler zum Tragen kommt.

6.2 Fazit

Wir haben nun anhand von unterschiedlichen theoretischen Resultaten sowie einigen Beispielen gesehen, dass die Konvergenz des Jacobi- und Gauß-Seidel-Verfahrens sowie der zugehörigen Relaxationsverfahren stark abhängig von der Ausgangsmatrix A und den daraus resultierenden Iterationsmatrizen ist. Je nach betrachtetem Gleichungssystem können die verschiedenen Verfahren divergieren, sehr langsam konvergieren oder aber sogar sehr schnell konvergieren. Ob eines der betrachteten Verfahren für eine bestimmte Anwendung infrage kommt, ist also immer vorher anhand des Spektralradius der Iterationsmatrix zu überprüfen, eventuell mithilfe von einem der hier vorgestellten Resultate.

Im Falle von strikt diagonaldominanten oder irreduzibel diagonaldominanten Ausgangsmatrizen haben wir gezeigt, dass sowohl das Jacobi-Verfahren als auch das Gauß-Seidel-Verfahren immer konvergieren. Es kann allerdings vorkommen, dass die Konvergenz dabei extrem langsam vonstatten geht und daher die Verfahren keinen wirklichen praktischen Nutzen haben, wie wir in Beispiel 6.2 gesehen haben. In diesem Fall können

die allgemeineren Relaxationsverfahren eine deutliche Verbesserung erwirken, was mit Sicherheit allerdings wieder nur unter bestimmten Voraussetzungen gezeigt werden kann, wie bei den hier gezeigten Ergebnissen für konsistent geordnete Matrizen.

Ähnliche Resultate zur SOR-Methode wie die hier gezeigten können auch für abweichende Voraussetzungen gezeigt werden, zum Beispiel für Matrizen, die der sogenannten A-Eigenschaft genügen und für die die Eigenwerte der Iterationsmatrix G_J des Jacobi-Verfahrens reell sind, wie in [6] (Eigenschaft 4.4) nachzulesen ist. In [5] lassen sich außerdem noch einige Resultate für die verschiedenen betrachteten Verfahren für sogenannte H-Matrizen finden.

Literaturverzeichnis

[1] Börm, Steffen: *Numerical Methods for Partial Differential Equations.* Vorlesungsskript, 2016.

[2] Börm, Steffen: *Wissenschaftliches Rechnen.* Vorlesungsskript, 2016.

[3] Hackbusch, Wolfgang: *Iterative Lösung großer schwachbesetzter Gleichungssysteme.* B.G. Teubner Stuttgart, 1993.

[4] Meister, Andreas: *Numerik linearer Gleichungssysteme.* 5. Auflage. Springer Spektrum, 2015.

[5] Meurant, Gerard: *Computer Solution of Large Linear Systems.* Elsevier Science B.V., 1999.

[6] Quarteroni, Alfio / Sacco, Riccarco / Saleri, Fausto: *Numerische Mathematik 1.* Springer, 2002.

[7] Saad, Yousef: *Iterative Methods for Sparse Linear Systems.* Second Edition. SIAM, 2003.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe angefertigt und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Weiterhin versichere ich, dass diese Arbeit noch nicht als Abschlussarbeit an anderer Stelle vorgelegen hat.

30. August 2016, _____
Alina Sophie Wrage