Kapitel 6

Lösung linearer Gleichungssysteme II: Iterative Verfahren

Wir kehren zurück zur Lösung linearer Gleichungssysteme

$$\boxed{Ax = b}$$

mit regulärer Matrix $A \in \mathbb{R}^{n\times n}$ und rechter Seite $b \in \mathbb{R}^n$. Bei Gleichungssystemen mit sehr großem n ist der Aufwand direkter Lösungsverfahren oft zu groß: $n^3/3$ bei vollbesetzter Matrix A, immer noch $nm^2/3$ bei Bandmatrizen der Bandbreite m. Dies ergibt für $n = 10^6$, $m = 200$ bereits über 10^{10} Rechenoperationen.

Bei vielen Aufgaben für zwei- und dreidimensionale Modelle in der Mechanik treten Matrizen mit sehr vielen Nullen auf, die jedoch keine starke Bandstruktur aufweisen.

- Wir haben dies bereits in Beispiel 2.3.9 bei der Diskretisierung der zweidimensionalen Poisson-Gleichung beobachtet. Die Matrix A der Dimension $n \times n$ hat die Bandbreite $N = \sqrt{n}$, aber höchstens 5 von Null verschiedene Einträge in jeder Zeile.

- Der Aufwand für ein direktes Verfahren ist also circa $n^2/3$, weil die Eliminationsverfahren zum "Auffüllen" der Nullen zwischen der Diagonalen und den äußeren "Bändern" der Matrix A führt.

- **Iterationsverfahren** bestimmen die Lösung $x^* \in \mathbb{R}^n$ näherungsweise und nutzen dabei die Struktur der Matrix A aus, indem sie nur auf die von Null verschiedenen Einträge zugreifen. Dies ermöglicht

  1. eine **kompakte Speicherung** dünn besetzter Matrizen (matlab "sparse"),

2. sehr effiziente Lösungsverfahren, bei denen nur eine kleine Anzahl von Matrix-Vektor Produkten gebildet wird. Jedes Matrix-Vektor Produkt kostet im angegbenen Beispiel nur $5n$ Rechenoperationen.

Die Verfahren in den ersten beiden Abschnitten fallen unter die Kategorie der Fixpunktverfahren im $\mathbb{R}^n$, die bereits in Kapitel 5.1 behandelt wurden. Im ersten Abschnitt wird das allgemeine Prinzip zur Lösung linearer Gleichungssysteme inclusive der Konvergenzanalyse behandelt. Im zweiten Teil werden dann mehrere Verfahren betrachtet.

6.1 Fixpunktverfahren allgemein

Die allgemeine Form der Fixpunkt-Iteration zur Lösung von $Ax = b$ lautet folgendermaßen.

6.1.1 Allgemeine Form der Fixpunkt-Iteration für LGS'e

Mit einer beliebigen invertierbaren Matrix $C \in \mathbb{R}^{n\times n}$ gilt die Äquivalenz

$$\boxed{Ax = b \iff x = x - C^{-1}(Ax - b) \iff x = (I - C^{-1}A)x + C^{-1}b.}$$

Man versucht C so zu wählen, dass

- zu gegebenem Vektor $x^{(k)}$ der neue Vektor

  $$x^{(k+1)} = x^{(k)} - C^{-1}(Ax^{(k)} - b)$$

  mit wenig Aufwand zu berechnen ist,

- die Iterationsfunktion

  $$\phi : \mathbb{R}^n \to \mathbb{R}^n, \qquad \phi(x) = (I - C^{-1}A)x + C^{-1}b$$

  kontrahierend ist und eine möglichst kleine Kontraktionszahl besitzt.

Die Kontraktionszahl L hängt (wie alle Lipschitz-Konstanten) von der gewählten Norm im $\mathbb{R}^n$ ab. Das Hauptergebnis dieses Abschnitts stellt jedoch fest, dass für die Konvergenz der Fixpunktiteration nicht eine Matrixnorm, sondern der Spektralradius der Iterationsmatrix

$$B = I - C^{-1}A \tag{6.1}$$

verantwortlich ist.

6.1.2 Satz: Globale Konvergenz der Fixpunktiteration
Die Matrizen A und C seien regulär, der Vektor x^* sei die eindeutige Lösung des linearen Gleichungssystems $Ax = b$, und $B := I - C^{-1}A$ bezeichne die Iterationsmatrix in (6.1).

Dann sind die folgenden Aussagen äquivalent:

(i) Die Fixpunktiteration $x^{(k+1)} = Bx^{(k)} + C^{-1}b$ konvergiert bei beliebigem Startvektor $x^{(0)} \in \mathbb{R}^n$ gegen x^*.

(ii) Der **Spektralradius** von B ist kleiner als 1, d.h.

$$\rho(B) = \max\{|\lambda|;\ \lambda \text{ Eigenwert von } B\} < 1.$$

(iii) Es gibt eine **natürliche** Matrixnorm $\|.\|$ auf $\mathbb{R}^{n\times n}$ mit $\|B\| < 1$.

In den Anwendungen wird oft Teil (iii) verwendet: falls für eine bestimmte Matrixnorm, z.B. die Zeilensummennorm $\|B\|_\infty < 1$ gilt, so ist die globale Konvergenz der Fixpunktiteration nachgewiesen. Dies finden wir in den Aussagen zu diagonaldominanten Matrizen.

Wir wollen kurz den Begriff der natürlichen Matrixnorm aus Definition 2.1.3 in Erinnerung rufen.

6.1.3 Natürliche Matrixnormen
Zu einer gegebenen (Vektor-)Norm $\|\cdot\|$ auf $\mathbb{R}^n$ definieren wir

$$\|A\|_{\text{op}} := \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Ax\|}{\|x\|} = \sup_{x\in\mathbb{R}^n,\ \|x\|=1} \|Ax\|.$$

Die Norm $\|\cdot\|_{\text{op}}$ heißt die *natürliche Matrixnorm* (oder *Operatornorm*) zur gegebenen Vektor-Norm $\|\cdot\|$ auf $\mathbb{R}^n$.

Beispiele natürlicher Matrixnormen sind die Zeilensummennorm $\|A\|_\infty$, Spaltensummennorm $\|A\|_1$ und Spektralnorm $\|A\|_2$. Viele weitere Beispiele natürlicher Matrixnormen erhält man wie in der Übungsaufgabe 1 von Blatt 3: Mit einer invertierbaren Matrix $S \in \mathbb{R}^{n\times n}$ wird durch

$$\|A\|_{\text{op}} := \|S^{-1}AS\|_\infty$$

eine natürliche Matrixnorm definiert. Dies zeigt bereits einen engen Zusammenhang zum Spektralradius: falls A diagonalisierbar ist, gilt mit einem regulären S, das als Spalten eine Basis aus Eigenvektoren von A besitzt,

$$S^{-1}AS = \text{diag}(\lambda_1, \ldots, \lambda_n),$$

also

$$\|S^{-1}AS\|_\infty = \rho(A).$$

Dieser Zusammenhang wurde in Übungsaufgabe 2 von Blatt 3 dargestellt. Ganz allgemein kann man sogar die folgende wichtige Aussage zeigen.

6.1.4 Satz zum Spektralradius von Matrizen

Gegeben sei die Matrix $B \in \mathbb{R}^{n\times n}$ mit Spektralradius $\rho(B)$.

a) Für jede natürliche Matrixnorm $q : \mathbb{R}^{n\times n} \to \mathbb{R}$ gilt

$$\rho(B) \leq q(B).$$

b) Zu beliebigem $\epsilon > 0$ existiert eine natürliche Matrixnorm $q : \mathbb{R}^{n\times n} \to \mathbb{R}$ mit

$$q(B) \leq \rho(B) + \epsilon.$$

Insgesamt gilt also

$$\rho(B) = \inf\{q(B) : q \text{ ist natürliche Matrixnorm auf } \mathbb{R}^{n\times n}\}.$$

Außerdem gilt für **jede** beliebige Matrixnorm auf $\mathbb{R}^{n\times n}$

$$\rho(B) = \lim_{k\to\infty} \left(\|B^k\|\right)^{1/k}.$$

Beweis: (a) Es sei $\lambda \in \mathbb{C}$ ein betragsgrößter Eigenwert und $x \in \mathbb{C}^n \setminus \{0\}$ ein zugehöriger Eigenvektor. Mit der zugehörigen Vektornorm $\|.\|$ auf $\mathbb{R}^n$ ist dann $\|Bx\| = \|\lambda x\| = |\lambda|\ \|x\|$, also nach Definition der Operatornorm $q(B) = \|B\|_{\text{op}} \geq |\lambda|$.

In Teil (b) wird eine reguläre Matrix ST zur Definition einer natürlichen Matrixnorm konstruiert. Hierbei dient S zur Erstellung der Jordan-Normalform von B

$$S^{-1}BS = \begin{bmatrix} \lambda_1 & d_1 & & \\ & \ddots & \ddots & \\ & & \lambda_{n-1} & d_{n-1} \\ & & & \lambda_n \end{bmatrix},$$

wobei $\lambda_1, \ldots, \lambda_n$ die Eigenwerte von B und $d_1, \ldots, d_{n-1}$ entweder 0 oder 1 sind. Die Matrix

$$T = \text{diag}(1, \epsilon, \ldots, \epsilon^{n-1})$$

liefert eine Skalierung der Nicht-Diagonalelemente d_j mit dem Faktor ϵ,

$$T^{-1}S^{-1}BST = \begin{bmatrix} \lambda_1 & \epsilon d_1 & & \\ & \ddots & \ddots & \\ & & \lambda_{n-1} & \epsilon d_{n-1} \\ & & & \lambda_n \end{bmatrix}.$$

Die natürliche Matrixnorm $q : \mathbb{R}^{n\times n} \to \mathbb{R}$, $q(B) = \|T^{-1}S^{-1}BST\|_\infty$, hat den Wert

$$q(B) = \max\{|\lambda_1| + \epsilon d_1, \ldots, |\lambda_{n-1}| + \epsilon d_{n-1}, |\lambda_n|\} \leq \rho(B) + \epsilon.$$

Damit ist auch die erste Darstellung des Spektralradius als Infimum gezeigt.

Die zweite Darstellung folgt aus der Äquivalenz der Normen: jede beliebige Matrixnorm ist äquivalent zur konstruierten Norm q zu dem gegebenen $\epsilon > 0$. Mit Konstanten $\alpha, \beta > 0$ gilt also

$$\alpha q(B^k) \leq \|B^k\| \leq \beta q(B^k).$$

Die Eigenwerte von B^k sind die Potenzen der Eigenwerte von B (Lineare Algebra), also gilt $\rho(B^k) = (\rho(B))^k$. Mit (a) und (b) und der Submultiplikativität der Norm q folgt weiter

$$(\rho(B))^k = \rho(B^k) \leq q(B^k) \leq q(B)^k \leq (\rho(B) + \epsilon)^k .$$

Die k-te Wurzel ist monoton, erhält also die Ungleichungen. Wir erhalten somit

$$\alpha^{1/k}\rho(B) \leq \left(\|B^k\|\right)^{1/k} \leq \beta^{1/k}\rho(B) + \epsilon.$$

Die Folgen $\alpha^{1/k}$ und $\beta^{1/k}$ konvergieren gegen 1 für $k \to \infty$. Also ist die Folge der $\left(\|B^k\|\right)^{1/k}$ beschränkt und alle Häufungspunkte liegen im Intervall $[\rho(B), \rho(B)+\epsilon]$. Nun folgt die Zusatzaussage mit dem üblichen Trick der Analysis: weil $\epsilon > 0$ beliebig war, ist $\rho(B)$ der einzige Häufungspunkt. Daher konvergiert diese Folge gegen $\rho(B)$.

6.1.5 Korollar

Es sei $B \in \mathbb{R}^{n\times n}$. Die folgenden Aussagen sind äquivalent:

(i) Die Folge $(B^k)_{k\geq 0}$ der Matrixpotenzen konvergiert in jeder beliebigen Matrixnorm gegen die Nullmatrix.

(ii) $\rho(B) < 1$.

Beweis: Wir benutzen die übliche 2-Norm $\|.\|_2$ im $\mathbb{C}^n$.

Falls $\rho(B) \geq 1$ gilt, gilt für einen Eigenvektor $x \in \mathbb{C}^n \setminus \{0\}$ zum betragsgrößten Eigenwert λ wieder $\|B^k x\| = |\lambda|^k \ \|x\| \not\to 0$, also konvergiert B^k nicht gegen die Nullmatrix.

Falls $\rho(B) < 1$ gilt, folgt mit $c = (1 + \rho(B))/2 < 1$ aus der Zusatzaussage in Satz 6.1.4 $\|B^k\| \leq c^k$ für alle $k \geq k_0$, also folgt $\|B^k\| \to 0$ und damit auch $B^k \to 0$.

Nun ist es leicht, die Konvergenz der Fixpunktiteration zu beweisen.

Beweis von 6.1.2: x^* sei die Lösung von $Ax = b$, also gilt $x^* = Bx^* + c$. Wir definieren den Fehlervektor $e^{(k)} := x^{(k)} - x^*$ und erhalten

$$e^{(k+1)} = x^{(k+1)} - x^* = Bx^{(k)} + c - Bx^* - c = Be^{(k)}, \qquad \text{also } \ e^{(k)} = B^k e^{(0)}.$$

Beliebige Wahl des Startvektors $x^{(0)}$ entspricht beliebigem Fehlervektor $e^{(0)} \in \mathbb{R}^n$. Daher ist (i) in Satz 6.1.2 äquivalent zur Beziehung

$$\lim_{k\to\infty} B^k y = 0 \qquad \text{für alle } \ y \in \mathbb{R}^n.$$

Dies ist äquivalent zur Konvergenz der Matrixfolge $(B^k)_{k\geq 0}$ gegen die Nullmatrix und wegen Korollar 6.1.5 äquivalent zu $\rho(B) < 1$.

Die Äquivalenz von (ii) und (iii) wurde in Satz 6.1.4 (a) und (b) bewiesen.

6.1.6 Bemerkung: Man beobachtet in der Praxis meist die lineare Konvergenz

$$\lim_{k\to\infty} \frac{\|x^{(k)} - x^*\|}{\|x^{(k-1)} - x^*\|} = \rho(B) < 1.$$

D.h. die numerisch beobachtete Kontraktionszahl ist $\rho(B)$. Genau untersuchen wir dies erst in einem späteren Kapitel bei der Behandlung von Eigenwert-Verfahren (vgl. Potenzmethode). Hier genügt eine kürzere Begründung.

a) Falls $\|B\|_{\text{op}} < 1$ gilt, so ist die Iterationsfunktion $\phi(x) = Bx + C^{-1}b$ sogar kontrahierend bzgl. der zugehörigen Vektornorm auf $\mathbb{R}^n$, die Kontraktionskonstante ist $L = \|B\|_{\text{op}}$:

$$\|\phi(x) - \phi(y)\| = \|B(x-y)\| \leq \|B\|_{\text{op}}\|x-y\|.$$

Deshalb gelten in diesem Fall die a-priori und a-posteriori Fehlerabschätzungen des Banachschen Fixpunktsatzes 5.1.6 wörtlich. Diese sollten als **Abbruchkriterien** verwendet werden. Wegen $L \geq \rho(B)$ sind sie etwas zu pessimistisch.

b) Falls $\|B\|_{\text{op}} \geq 1$, aber $\rho(B) < 1$ gilt, so verwenden wir die Zusatzaussage in Satz 6.1.4 in der folgenden Form: zu $\epsilon > 0$ mit $\tau := \rho(B) + \epsilon < 1$ gibt es ein $l \in \mathbb{N}$ so, dass

$$\|B^l\|_{\text{op}} \leq \tau^l$$

gilt. Dann betrachtet man statt der einzelnen Iterationsschritte die Teilfolgen

$$x^{(j)},\ x^{(j+l)},\ x^{(j+2l)},\ \ldots$$

für $j = 0, \ldots, l-1$ und erhält

$$\|x^{(j+kl)} - x^*\| \leq \tau^l \|x^{(j+(k-1)l)} - x^*\|, \qquad k \in \mathbb{N}.$$

Damit lassen sich dann die Fehlerabschätzungen im Banachschen Fixpunktsatz für diese l Teilfolgen verwenden, die Kontraktionszahl jedes "Päckchens" von l Schritten ist $L = \tau^l$. Grob gesagt liegt für Einzelschritte die "Kontraktion im Mittel" mit dem Faktor $\tau \approx \rho(B)$ vor.

6.2 Die praktischen Verfahren: Jacobi, Gauß-Seidel und SOR

Wir kommen nun zum praktischen Teil und behandeln mehrere spezielle Fixpunkt-Verfahren zur Lösung des linearen Gleichungssystems $Ax = b$. Für die praktische Durchführung soll vorab gesagt werden, dass die Iterationsmatrizen $B = I - C^{-1}A$ nie explizit aufgestellt werden, sie dienen allein der Konvergenzanalyse in diesem Abschnitt. Die praktische Rechnung hingegen arbeitet nur mit den Matrixelementen von A.

Weiterhin sollte man Satz 6.1.2 als zentrale Aussage für die Konvergenz der Verfahren im Kopf behalten, die ja die globale Konvergenz vollständig charakterisiert. Wir werden weitere **hinreichende** Bedingungen für die Konvergenz einzelner Verfahren erhalten. Diese sind als Ergänzung anzusehen, die das praktische Überprüfen der Konvergenz erleichtern.

Zuerst betrachten wir die zwei bekanntesten Verfahren, die wenig Aufwand zur Berechnung von $x^{(k+1)}$ erfordern.

6.2.1 Gesamtschrittverfahren (Jacobi-Verfahren)

Die Matrix A habe Diagonalelemente $a_{ii} \neq 0$.

1. Wähle beliebigen Startvektor $x^{(0)} \in \mathbb{R}^n$, z.B. $x_i^{(0)} = \frac{b_i}{a_{ii}}$.
2. Für $k = 0, 1, 2, \ldots$

   für $i = 1, 2, \ldots, n$: ("löse die i-te Gleichung nach x_i")

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right).$$

6.2.2 Einzelschrittverfahren (Gauß-Seidel-Verfahren)

Die Matrix A habe Diagonalelemente $a_{ii} \neq 0$.

1. Wähle beliebigen Startvektor $x^{(0)} \in \mathbb{R}^n$, z.B. $x_i^{(0)} = \frac{b_i}{a_{ii}}$.
2. Für $k = 0, 1, 2, \ldots$

   für $i = 1, 2, \ldots, n$: ("löse die i-te Gleichung nach x_i")

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(\mathbf{k+1})} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right).$$

6.2.3 Beispiel: eindimensionales Modellproblem (analytische Lösung)

Auf einen an beiden Enden $x = 0$ und $x = l$ aufgelegten Balken wirke eine Kraft F in der Mitte $x = l/2$. Die Biegelinie $u(x)$ lässt sich unter weiteren Annahmen (kleine Auslenkung aus der horizontalen) näherungsweise beschreiben durch die lineare Differentialgleichung 2. Ordnung

$$-u''(x) = \frac{M(x)}{C}, \qquad x \in (0, l),$$

mit den Randwerten $u(0) = u(l) = 0$. Hierbei ist C die Biegesteifigkeit und $M(x) = \frac{Fx}{2}$ für $0 \leq x \leq \frac{l}{2}$, $M(x) = \frac{F(l-x)}{2}$ für $\frac{l}{2} < x < l$ das Biegemoment. Zweimalige Integration und Berücksichtigung der Randwerte ergibt die Lösung

$$u(x) = \frac{F}{4C} \left(\frac{l^2 x}{4} - \frac{x^3}{3} \right) \quad \text{für} \quad x \in [0, l/2],$$

und $u(x) = u(l - x)$ für $x \in [l/2, l]$. Die maximale Auslenkung in der Balkenmitte ist $u(l/2) = \frac{Fl^3}{48C}$.

6.2.4 Beispiel: eindimensionales Modellproblem (numerische Lösung)

Wir betrachten das obige Modell mit beliebiger rechter Seite $f \in C[0, l]$,

$$-u''(x) = f(x), \quad x \in (0, l), \qquad u(0) = u(l) = 0.$$

Zu $n \in \mathbb{N}$ werden die Schrittweite $h = \frac{l}{n+1}$ und die Stellen $x_k = kh$, $0 \leq k \leq n + 1$, eingeführt. Dann lautet die übliche Diskretisierung der 2. Ableitung

$$-u''(x_k) \approx \frac{2u(x_k) - u(x_{k-1}) - u(x_{k+1})}{h^2}, \qquad 1 \leq k \leq n.$$

$x^{(0)}$	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	$x^{(6)}$	$x^{(7)}$	$\cdots$	$x^{(20)}$
0.5000	1.0000	1.2500	1.5000	1.6250	1.7500	1.8125	1.8750		1.9985
1.0000	1.5000	2.0000	2.2500	2.5000	2.6250	2.7500	2.8125		2.9980
0.5000	1.0000	1.2500	1.5000	1.6250	1.7500	1.8125	1.8750		1.9985

Tabelle 6.1: Werte $4\widehat{u}_k$ beim Jacobi-Verfahren zum 1-dim. Modellproblem, $h = \frac{1}{4}$

$x^{(0)}$	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	$x^{(6)}$	$x^{(7)}$	$\cdots$	$x^{(20)}$
0.5000	1.0000	1.3750	1.6875	1.8438	1.9219	1.9609	1.9805		2.0000
1.0000	1.7500	2.3750	2.6875	2.8438	2.9219	2.9609	2.9805		3.0000
0.5000	1.3750	1.6875	1.8438	1.9219	1.9609	1.9805	1.9902		2.0000

Tabelle 6.2: Werte $4\widehat{u}_k$ beim Gauß-Seidel Verfahren zum 1-dim. Modellproblem, $h = \frac{1}{4}$

Die Näherungswerte $\widehat{u}_k \approx u(x_k)$, $1 \le k \le n$, erhält man aus dem linearen Gleichungssystem $A\widehat{u} = b$ mit der Tridiagonalmatrix

$$A = \begin{bmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{n\times n}$$

zur rechten Seite $b_k = h^2 f(x_k)$, $1 \le k \le n$. Für spätere Zwecke halten wir fest: die Eigenwerte von A sind

$$\lambda_k = 2\left(1 - \cos\frac{k\pi}{n+1}\right), \qquad 1 \le k \le n,$$

und zugehörige Eigenvektoren sind

$$v_k = \left(\sin\frac{jk\pi}{n+1}\right)_{1\le j\le n}.$$

Konkretes Beispiel: Zu $l = 2$, $F/C = 4$ erhalten wir für $n = 3$ die rechte Seite $b = \frac{1}{4}(1,2,1)^T$ und die Lösung $\widehat{u} = \frac{1}{4}(2,3,2)^T$. Gesamt- und Einzelschrittverfahren liefern die Werte der Tabellen 6.1 und 6.2. Zum Vergleich: die Werte der analytischen Lösung $u(x) = x - x^3/3$ sind $u(1/2) = 11/24 = u(3/2)$, $u(1/2) = 2/3$.

Zur Konvergenzanalyse beider Verfahren stellen wir die Iterationsmatrizen B der Verfahren auf (vgl. (6.1)). Dazu wird die Iterationsvorschrift in 6.2.1 und 6.2.2 von der komponentenweisen Form in die Vektorform übertragen. Dazu teilen wir die Matrix A auf in strikte untere Hälfte, Diagonale und strikte obere Hälfte, also $A = L + D + R$ mit

$$L = \begin{bmatrix} 0 & \cdots & \cdots & 0 \\ a_{21} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ a_{n1} & \cdots & a_{n,n-1} & 0 \end{bmatrix}, \quad D = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{nn} \end{bmatrix}, \quad R = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & a_{n-1,n} \\ 0 & \cdots & \cdots & 0 \end{bmatrix}.$$

Beim Gesamtschrittverfahren ist der Iterationsschritt sofort in Vektorform anzugeben als

$$x^{(k+1)} = D^{-1}\left(b - (L+R)x^{(k)}\right) = -D^{-1}(L+R)x^{(k)} + D^{-1}b.$$

Das GSV wird daher sehr oft für parallele Verarbeitung eingesetzt: die Komponenten von $x^{(k+1)}$ können unabhängig voneinander zeitlich berechnet werden.

Beim Einzelschrittverfahren bringt man zunächst alle Komponenten von $x^{(k+1)}$ auf die linke Seite und multipliziert mit den Diagonalelementen, dies ergibt

$$(D + L)x^{(k+1)} = b - Rx^{(k)}.$$

Man formt weiter um zu

$$x^{(k+1)} = -(D + L)^{-1}Rx^{(k)} + (D + L)^{-1}b.$$

6.2.5 Proposition
Die Iterationsfunktion des Gesamtschrittverfahrens lautet

$$\phi_{\text{GSV}}(x) = J_1x + D^{-1}b \quad \text{mit} \quad J_1 = -D^{-1}(L + R),$$

und die des Einzelschrittverfahrens lautet

$$\phi_{\text{ESV}}(x) = H_1x + (L + D)^{-1}b \quad \text{mit} \quad H_1 = -(L + D)^{-1}R.$$

Die Bezeichnungen der Iterationsmatrizen mit J_1 und H_1 werden später deutlich. Dann werden noch weitere Verfahren (JOR und SOR) zu Matrizen J_ω, H_ω mit $\omega > 0$ definiert.

Ein nützliches Konvergenzkriterium für beide Verfahren ergibt sich aus der folgenden Eigenschaft. Dieses Kriterium ist anhand der Matrix-Einträge sehr leicht zu überprüfen.

6.2.6 Satz: GSV und ESV für stark diagonaldominante Matrizen
Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt *stark diagonaldominant*, falls

$$|a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \qquad \text{für alle } 1 \leq i \leq n.$$

Falls A stark diagonaldominant ist, so gilt für die Iterationsmatrizen J_1 und H_1

$$\rho(J_1) \leq \|J_1\|_\infty < 1, \qquad \rho(H_1) \leq \|H_1\|_\infty \leq \|J_1\|_\infty < 1.$$

Insbesondere sind GSV und ESV global konvergent, $L = \|J_1\|_\infty$ ist eine Kontraktionskonstante bzgl. der Zeilensummennorm für beide Verfahren.

Beweis: Die Voraussetzung liefert $a_{ii} \neq 0$ für alle i, dies muss also nicht separat verlangt werden. Also sind J_1 und H_1 wohldefiniert.

Die Abschätzung für J_1 folgt aus Satz 6.1.4(a) und der starken Diagonaldominanz, denn $\rho(J_1) \leq \|J_1\|_\infty$ und

$$\|J_1\|_\infty = \|D^{-1}(L + R)\|_\infty = \max_{1\leq i\leq n} \frac{1}{|a_{ii}|} \sum_{j\neq i} |a_{ij}| < 1. \tag{6.2}$$

Für die Operatornorm $\|H_1\|_\infty$ müssen wir für beliebiges $v \in \mathbb{R}^n$ mit $\|v\|_\infty = 1$ die Maximumsnorm des Vektors $w = H_1 v$ nach oben abschätzen. Dazu betrachten wir die rekursiv definierten Zahlen

$$s_i = \frac{1}{|a_{ii}|}\Big(\sum_{j=1}^{i-1} |a_{ij}| s_j + \sum_{j=i+1}^{n} |a_{ij}|\Big), \qquad 1 \le i \le n. \tag{6.3}$$

(Für $i = 1$ ist die erste Summe leer (also Null), für $i = n$ die letzte.)

Die Komponenten w_i von $w = H_1 v$ sind wie in 6.2.2 gegeben durch

$$w_i = -\frac{1}{a_{ii}}\Big(\sum_{j=1}^{i-1} a_{ij} w_j + \sum_{j=i+1}^{n} a_{ij} v_j\Big).$$

Speziell für $i = 1$ folgt aus $|v_j| \le 1$

$$|w_1| \le \frac{1}{|a_{11}|} \sum_{j=2}^{n} |a_{ij}|\ |v_j| \le s_1.$$

Rekursiv ergibt sich weiter für $i = 2, \ldots, n$

$$|w_i| \le \frac{1}{|a_{ii}|}\Big(\sum_{j=1}^{i-1} |a_{ij}| \underbrace{|w_j|}_{\le s_j} + \sum_{j=i+1}^{n} |a_{ij}| \underbrace{|v_j|}_{\le 1}\Big) \le s_i.$$

Also ist

$$\|H_1\|_\infty = \max_{\|v\|_\infty = 1} \|H_1 v\|_\infty \le \max_{1 \le i \le n} s_i =: s.$$

Aus der Diagonaldominanz von A erhalten wir mit (6.2)

$$s_1 = \frac{1}{|a_{11}|} \sum_{j=2}^{n} |a_{ij}| \le \|J_1\|_\infty < 1,$$

und weiter per Rekursion für $i = 2, \ldots, n$

$$s_i = \frac{1}{|a_{ii}|}\Big(\sum_{j=1}^{i-1} |a_{ij}| \underbrace{s_j}_{<1} + \sum_{j=i+1}^{n} |a_{ij}|\Big) \le \|J_1\|_\infty < 1.$$

Damit ist $\|H_1\|_\infty \le s \le \|J_1\|_\infty < 1$ bewiesen.

Der obige Beweis zum ESV liefert sogar die Konvergenz für eine etwas größere Klasse von Matrizen.

6.2.7 Korollar: Kriterium von Sassenfeld

Die Matrix $A \in \mathbb{R}^{n \times n}$ habe nicht-verschwindende Diagonalelemente. Das ESV konvergiert, falls die rekursiv definierten Zahlen s_i in (6.3) kleiner als 1 sind; in diesem Fall gilt

$$\rho(H_1) \le \|H_1\|_\infty \le \max_{1 \le i \le n} s_i =: s < 1.$$

Die **starke** Diagonaldominanz in Satz 6.2.6 ist in vielen praktischen Beispielen verletzt, vgl. das Modellproblem in Beispiel 6.2.4. Deshalb lohnt es sich, die Voraussetzungen noch etwas abzuschwächen. Dabei können wir aber das "$>$" nicht einfach durch ein "$\ge$" ersetzen, sondern müssen noch etwas mehr verlangen. Dies wird in den folgenden zwei Definitionen erklärt.

6.2.8 Definition: Schwach diagonaldominante Matrix
Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt *schwach diagonaldominant*, wenn

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

gilt **und** wenn mindestens ein $1 \leq s \leq n$ existiert mit

$$|a_{ss}| > \sum_{\substack{j=1\\ j\neq s}}^{n} |a_{sj}|.$$

6.2.9 Definition: Unzerlegbare Matrix
Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt *zerlegbar*, wenn es zwei nichtleere Indexmengen $I, J \subset \{1, 2, \ldots, n\}$ gibt mit

$$I \cup J = \{1, 2, \ldots, n\}, \qquad I \cap J = \emptyset,$$

und

$$a_{i,j} = 0 \qquad \text{für alle} \quad i \in I,\ j \in J.$$

Mit anderen Worten: A ist zerlegbar, wenn es eine Permutationsmatrix P gibt, so dass PAP^T die Block-Gestalt

$$PAP^T = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}$$

besitzt, bei der A_{11} eine quadratische Matrix mit $1 \leq k \leq n-1$ Zeilen ist und entsprechend A_{22} eine quadratische Matrix mit $n-k$ Zeilen ist.

Falls keine solche Zerlegung existiert, heißt *A unzerlegbar* **oder auch** *irreduzibel***.**

Bemerkung: Die Irreduzibilität lässt sich mit einem diskreten Graphen erklären: die Knoten des Graphen sind die Zahlen 1 bis n, eine gerichtete Kante vom Knoten j zum Knoten k liegt genau dann vor, wenn $a_{j,k} \neq 0$ gilt. Die Matrix A ist genau dann unzerlegbar, wenn es von jedem Knoten r mindestens einen Weg entlang der gerichteten Kanten zu jedem anderen Knoten s gibt. (Man sagt dann, dass der Graph zusammenhängend ist.)

Im folgenden Beweis verwenden wir diese Eigenschaft in folgender Form: Zu beliebigen Indizes $1 \leq r, s \leq n$ gibt es eine "Kette" weiterer Indizes $k_1, \ldots, k_m$ mit

$$a_{r,k_1} \neq 0, \quad a_{k_1,k_2} \neq 0, \qquad a_{k_2,k_3} \neq 0 \quad \ldots \qquad a_{k_m,s} \neq 0. \tag{6.4}$$

6.2.10 Satz: GSV und ESV für schwach diagonaldominante unzerlegbare Matrizen

Die Matrix $A \in \mathbb{R}^{n\times n}$ sei schwach diagonaldominant und unzerlegbar.

Dann gilt:

a) A ist regulär,

b) alle Diagonalelemente a_{ii} sind ungleich 0,

c) sowohl das Gesamtschrittverfahren als auch das Einzelschrittverfahren konvergieren für jeden Startvektor $x^{(0)}$.

Beweis: 1. Da A unzerlegbar ist, enthält A keine Nullzeile. Aus der schwachen Diagonaldominanz folgt $|a_{ii}| > 0$ für alle i, und die Matrizen J_1 und H_1 sind wohldefiniert.
2. Genau wie in Satz 6.2.6 folgt aus der schwachen Diagonaldominanz bereits

$$\rho(J_1) \le \|J_1\|_\infty \le 1, \qquad \rho(H_1) \le \|H_1\|_\infty \le \|J_1\|_\infty \le 1.$$

Wir müssen also nur ausschließen, dass Eigenwerte vom Betrag 1 auftreten. Wir betrachten zuerst J_1.
3. Angenommen, es existiert ein Eigenwert $\lambda \in \mathbb{C}$ von J_1 mit $|\lambda| = 1$. Sei $v \in \mathbb{C}^n$ zugehöriger Eigenvektor mit $\|v\|_\infty = 1$. Der Index r wird gewählt mit $|v_r| = 1$. Aus $|\lambda| = 1$ und $\lambda v = J_1 v$ folgt (in jeder Zeile i)

$$|v_i| = |\lambda v_i| \le \frac{1}{|a_{ii}|} \sum_{j\neq i} |a_{ij}|\ |v_j|, \qquad 1 \le i \le n. \tag{6.5}$$

Für mindestens einen Index s liefert die "starke" Bedingung in Definition 6.2.8 sowie $|v_j| \le 1$

$$|v_s| \le \frac{1}{|a_{ss}|} \sum_{j\neq s} |a_{sj}| < 1$$

und wegen $|v_r| = 1$ ist $r \neq s$. Aufgrund der Unzerlegbarkeit können wir nun die strikte Ungleichung auch für $|v_r|$ erzielen, das ergibt dann den Widerspruch. Dazu wählen wir eine Kette aus Indizes $r, k_1, \ldots, k_m, s$ wie in (6.4) und erhalten aus (6.5) Schritt für Schritt

$$\begin{aligned}
|v_{k_m}| &\le \frac{1}{|a_{k_m,k_m}|}\left(\sum_{j\neq k_m,s} |a_{k_m,j}|\ |v_j| + \underbrace{|a_{k_m,s}|}_{>0}\ \underbrace{|v_s|}_{<1}\right) < \frac{1}{|a_{k_m,k_m}|}\sum_{j\neq k_m} |a_{k_m,j}| \le 1,\\
&\vdots\\
|v_r| &\le \frac{1}{|a_{rr}|}\left(\sum_{j\neq r,k_1} |a_{r,j}|\ |v_j| + \underbrace{|a_{r,k_1}|}_{>0}\ \underbrace{|v_{k_1}|}_{<1}\right) < \frac{1}{|a_{rr}|}\sum_{j\neq r} |a_{rj}| \le 1.
\end{aligned}$$

Damit ist der Widerspruch zu $|v_r| = 1$ bzw. $\rho(J_1) = 1$ gezeigt.
4. Der Widerspruch zu $\rho(H_1) = 1$ folgt ähnlich. Man geht aus von $\lambda v = -(D+L)^{-1}Rv$ mit $|\lambda| = 1$ und formt dies um zu

$$\lambda(D+L)v = -Rv \implies \lambda Dv = -(\lambda L + R)v \implies \lambda v = -D^{-1}(\lambda L + R)v.$$

Daraus ergibt sich die gleiche Ungleichung (6.5) für $1 \le i \le n$, die genau so zum Widerspruch geführt wird.

6.2.11 Bemerkung: Die bisherigen Aussagen gelten für Klassen von Matrizen, bei denen sowohl das GSV als auch das ESV global konvergieren. Satz 6.2.6 legt sogar die

Vermutung nahe, dass das ESV immer schneller konvergiert als das GSV. Das ist in der Praxis häufig der Fall, für $n = 2$ kann man sogar leicht beweisen, dass $\rho(H_1) = \rho(J)^2$ gilt, also im Falle der Konvergenz beider Verfahren (d.h. $\rho(J_1) < 1$) der Spektralradius von H_1 kleiner ist als der von J_1. Aber:

- Es gibt diagonaldominante Matrizen mit $1 > \rho(H_1) > \rho(J_1)$. Beispiel

$$A = \begin{pmatrix} 270 & 260 & -1 \\ 1 & 20 & -19 \\ 7 & 2 & 9 \end{pmatrix}$$

  mit $\rho(J_1) \approx 0.9254$, $\rho(H_1) \approx 0.9305$. Dies ist kein Widerspruch zu den bisherigen Aussagen: es gilt $\|H_1\|_\infty = 0.998 < \|J_1\|_\infty = 1$. Bei der Iteration wird das Jacobi-Verfahren tatsächlich etwas schneller konvergieren, siehe Bemerkung 6.1.6.

- Ohne die Eigenschaft der Diagonaldominanz können sogar Unterschiede im Konvergenzverhalten auftreten: Es gibt Matrizen mit $\rho(J_1) < 1 < \rho(H_1)$ und auch solche mit $\rho(H_1) < 1 < \rho(J_1)$, vgl. Übungsblatt 14.

Man muss also bei den Konvergenzaussagen zu beiden Verfahren sehr sorgfältig sein.

Für das Gauß-Seidel Verfahren gibt es eine weitere hinreichende Bedingung der Konvergenz, die eine große Klasse von Matrizen umfasst.

6.2.12 Satz: Konvergenz des ESV für positiv-definite Matrizen
Falls $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit ist, so konvergiert das Einzelschrittverfahren.

Beweis: 1. Die Diagonalelemente von A erfüllen $a_{ii} > 0$, also ist H_1 wohldefiniert.
2. Wir definieren die Vektornorm

$$\|x\|_A := \sqrt{x^T A x}.$$

Dies ist eine Norm, weil A symmetrisch und positiv definit ist. Wir zeigen für die zugehörige Operatornorm

$$\|H_1\|_A = \max\{\|H_1 x\|_A;\ \|x\|_A = 1\} < 1.$$

Dann folgt aus Satz 6.1.4(a) $\rho(H_1) \le \|H_1\|_A < 1$, also die Konvergenz des ESV.
3. Die Definition der Operatornorm ergibt die Äquivalenz

$$\begin{aligned} \|H_1\|_A < 1 &\iff \|H_1 x\|_A < \|x\|_A \text{ für alle } x \neq 0 \\ &\iff x^T H_1^T A H_1 x < x^T A x \text{ für alle } x \neq 0 \\ &\iff x^T (A - H_1^T A H_1) x > 0 \text{ für alle } x \neq 0. \end{aligned}$$

Zu zeigen ist also, dass die Matrix $A - H_1^T A H_1$ positiv definit ist. Die Symmetrie von A ergibt die Zerlegung $A = L + D + L^T$ mit strikter unterer Dreiecksmatrix L und Diagonalmatrix D. Wir setzen $X := -(D + L)^{-1}$ und erhalten

$$H_1 = XL^T = X(A - L - D) = X(A + X^{-1}) = I + XA,$$

und weiter wegen $A = A^T$

$$\begin{aligned} A - H_1^T A H_1 &= A - (I + XA)^T A (I + XA) = A - (I + AX^T) A (I + XA) \\ &= A - A - AX^T A - AXA - AX^T AXA = -(AX^T A + AXA + AX^T AXA) \\ &= -AX^T \underbrace{(X^{-1} + (X^T)^{-1} + A)}_{=-D} XA \\ &= AX^T DXA = (XA)^T D (XA). \end{aligned}$$

Tabelle 6.3: Werte $4\widehat{u}_k$ beim SOR zum 1-dim. Modellproblem aus 6.2.4, $h = \frac{1}{4}$, $\omega = 1.2$

$x^{(0)}$	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	$x^{(6)}$	$x^{(7)}$	$x^{(8)}$
0.5000	1.1000	1.5560	1.9371	1.9890	2.0007	2.0003	2.0001	2.0000
1.0000	1.9600	2.7472	2.9607	2.9975	3.0008	3.0003	3.0001	3.0000
0.5000	1.6760	1.9131	1.9938	1.9998	2.0005	2.0001	2.0000	2.0000

Die Diagonalmatrix D ist positiv definit (wegen $a_{ii} > 0$ wie oben), X und A sind regulär, also ist $(XA)^T D(XA)$ ebenfalls positiv definit (Sylvesterscher Trägheitssatz). q.e.d.

Man beobachtet beim ESV häufig, dass die Schritte sich "aus einer Richtung" an die Lösung x^* annähern. Man könnte daher versuchen, die Konvergenz zu beschleunigen, indem in jedem Berechnungsschritt (innere Schleife des ESV) eine Verlängerung der Korrektur um einen festen Faktor $\omega > 1$ vorgenommen wird.

6.2.13 Das SOR (=successive over-relaxation)-Verfahren

Die Matrix A habe Diagonalelemente $a_{ii} \neq 0$. Weiter sei $\omega \in \mathbb{R}$. (ω wird *Relaxationsparameter* genannt.)

1. Wähle beliebigen Startvektor $x^{(0)} \in \mathbb{R}^n$, z.B. $x_i^{(0)} = \frac{b_i}{a_{ii}}$.
2. Für $k = 0, 1, 2, \ldots$

   für $i = 1, 2, \ldots, n$

$$\widetilde{x}_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right)$$

$$x_i^{(k+1)} = \omega \widetilde{x}_i^{(k+1)} + (1 - \omega) x_i^{(k)}.$$

Bemerkung: Eine entsprechende Änderung beim Gesamtschrittverfahren liefert das JOR-Verfahren; dies wird hier nicht diskutiert.

6.2.14 Beispiel: Das eindimensionale Modellproblem aus 6.2.4 wird in den Tabellen 6.3 und 6.4 mit den Parametern $\omega = 1.2$ bzw. $\omega = 1.3$ im SOR-Verfahren gerechnet. Zur Wahl des "optimalen" Relaxationsparameters vgl. Satz 6.2.17.

Die Konvergenzanalyse erfolgt wieder über die Betrachtung der Iterationsmatrix des SOR-Verfahrens. Dazu verwenden wir die Zerlegung $A = L + D + R$ wie vor Proposition 6.2.5.

6.2.15 Proposition: Iterationsfunktion des SOR-Verfahrens
Mit $A = L + D + R$ wie in 6.2.5 lautet die Iterationsfunktion des SOR-Verfahrens

$$\phi_{\mathrm{SOR}}(x) = H_\omega x + \omega(D + \omega L)^{-1} b$$

mit der Iterationsmatrix $H_\omega = (D + \omega L)^{-1}((1-\omega)D - \omega R)$.

Herleitung: Umstellung der Gleichungen in 6.2.13 gibt

$$a_{ii} x_i^{(k+1)} + \omega \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} = \omega b_i - \omega \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} + (1-\omega) a_{ii} x_i^{(k)}.$$

Dies lautet

$$(D + \omega L) x^{(k+1)} = \omega b + ((1-\omega)D - \omega R) x^{(k)}.$$

Bemerkung: Für $\omega = 1$ ist $\phi_{\mathrm{SOR}} = \phi_{\mathrm{ESV}}$ in 6.2.5, daher rührt auch die Bezeichnung H_1 beim ESV.

6.2.16 Satz: Konvergenzanalyse

a) *von Kahan:* Für beliebiges A mit Diagonalelementen $a_{ii} \neq 0$ und beliebiges $\omega \in \mathbb{R}$ ist

$$\rho(H_\omega) \geq |\omega - 1|.$$

Die Konvergenz des SOR-Verfahrens kann also höchstens für $\omega \in (0, 2)$ eintreten.

b) *von Reich und Ostrowski:* Für jede positiv definite Matrix A und jedes $\omega \in (0, 2)$ gilt $\rho(H_\omega) < 1$, also ist das SOR-Verfahren konvergent.

Beweis:

a) Mit $\widetilde{L} = D^{-1}L$ und $\widetilde{R} = D^{-1}R$ schreibt man H_ω um in

$$H_\omega = (D + \omega L)^{-1}((1-\omega)D - \omega R) = (I + \omega\widetilde{L})^{-1} \underbrace{D^{-1}D}_{=I} ((1-\omega)I - \omega\widetilde{R}).$$

Weil $\widetilde{L}$ und $\widetilde{R}$ strikte untere bzw. obere Dreiecksmatrizen sind, folgt

$$\det H_\omega = \det(I + \omega\widetilde{L})^{-1} \ \det((1-\omega)I - \omega\widetilde{R}) = (1-\omega)^n.$$

Weil $\det H_\omega$ gleich dem Produkt aller Eigenwerte von H_ω ist, folgt $\rho(H_\omega) \geq |1 - \omega|$.

b) Wir zeigen $\|H_\omega\|_A < 1$ wie in 6.2.12. Dies ist wieder äquivalent zur positiven Definitheit der Matrix $A - H_\omega^T A H_\omega$.

Mit $X := -(D + \omega L)^{-1}$ und $R = A - D - L$ ist

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

Weiter ergibt sich wie in 6.2.12

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

mit der positiv definiten Matrix AX^TDXA aus 6.2.12. Weil $\omega(2-\omega) > 0$ für alle $\omega \in (0,2)$ gilt, ist alles gezeigt.

6.2.17 Beispiel: Die Modellmatrix $A = \texttt{tridiag}(-1, 2, -1) \in \mathbb{R}^{n\times n}$ in Beispiel 6.2.4 besitzt die Eigenwerte

$$\lambda_k = 2\left(1 - \cos\frac{k\pi}{n+1}\right), \qquad 1 \le k \le n.$$

Einfache Rechnung ergibt für $J = -D^{-1}(L+R) = \frac{1}{2}(2I - A)$ die Eigenwerte

$$\mu_k = \cos\frac{k\pi}{n+1}.$$

Mit $h := \frac{1}{n+1}$ ist

$$\begin{aligned} \rho(J_1) &= \cos\pi h = 1 - \frac{\pi^2 h^2}{2} + \mathcal{O}(h^4), \\ \rho(H_1) &= \rho(J_1)^2 = \cos^2\pi h = 1 - \pi^2 h^2 + \mathcal{O}(h^4), \\ \rho(H_\omega) &= \begin{cases} \omega - 1 \quad \text{für } \omega \ge \omega^*, \\ \frac{1}{4}\left(\omega\rho(J_1) + \sqrt{\omega^2\rho(J_1)^2 - 4(\omega-1)}\right)^2 \quad \text{für } \omega < \omega^*. \end{cases} \end{aligned}$$

Der Wert $\omega = \omega^* = \dfrac{2}{1 + \sqrt{1-\rho(J)^2}} \in (1,2)$ liefert das Minimum

$$\rho(H_{\omega^*}) = \omega^* - 1 = \frac{1 - \sqrt{1-\rho(J)^2}}{1 + \sqrt{1-\rho(J)^2}} = \frac{1 - \sin\pi h}{1 + \sin\pi h} = 1 - 2\pi h + \mathcal{O}(h^2).$$

Dies ist eine deutliche Verbesserung der Rate der linearen Konvergenz.

6.2.18 Diskussion: Die im Beispiel 6.2.17 gemachte Beobachtung trifft bei Matrizen eines *speziellen Typs* zu, die bei der Diskretisierung partieller Dgl'n auftreten. In diesen Spezialfällen ($\to$ M-Matrizen) gilt

$$\rho(H_1) = \rho(J)^2 < 1$$

und $\rho(H_\omega)$ hat die in 6.2.17 angegebene Form. Das absolute Minimum liegt beim optimalen Relaxationsparameter

$$1 < \omega^* = \frac{2}{1 + \sqrt{1-\rho(J)^2}} < 2,$$

es hat den Wert

$$\rho(H_{\omega^*}) = \omega^* - 1 = \frac{1 - \sqrt{1-\rho(J)^2}}{1 + \sqrt{1-\rho(J)^2}} < 1.$$

Zahlenwerte für zwei verschiedene M-Matrizen mit $\rho(J_1) = 0.809$ bzw. $\rho(J_1) = 0.99$ sind in Tabelle 6.5 enthalten. Die Graphen der Funktionen $g(\omega) = \rho(H_\omega)$ sind in Abbildung 6.1 nebeneinander dargestellt.

Tabelle 6.4: Vergleich der Werte der Spektralradien für GSV, ESV und SOR mit optimalem Relaxationsparameter

	$\rho(J_1)$	$\rho(H_1)$	$\rho(H_{\omega^*})$	ω^*
linkes Bild, s. Beispiel 6.2.14	0.809	0.6545	0.2596	1.2596
rechtes Bild	0.99	0.9801	0.7527	1.7527

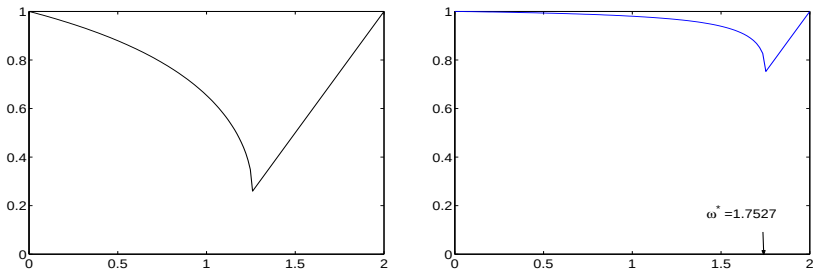

Abbildung 6.1: Graph der Funktion $g(\omega) = \rho(H_\omega)$ für zwei verschiedene M-Matrizen

6.3 Methode der konjugierten Gradienten

Wir stellen kurz ein weiteres Verfahren vor, das in der Praxis sehr häufig verwendet wird. Es handelt sich eigentlich um ein direktes Lösungsverfahren für das lineare Gleichungssystem $Ax = b$ ist. **Hierbei ist A symmetrisch und positiv definit!**

Das CG-Verfahren beruht nicht auf der Elimination, sondern auf der sukzessiven Annäherung einer Vektorfolge $x^{(k)}$, $k \geq 0$, an die Lösung x^*, die nach spätestens n Schritten die exakte Lösung ergibt. Weil man meist bereits nach $r < n$ Schritten eine gute Genauigkeit erzielt, wird hier auch von einem Iterationsverfahren gesprochen. Das Prinzip des Verfahrens beruht auf folgendem Resultat.

6.3.1 Satz: Minimierungsfunktional
Die eindeutige Lösung x^* von $Ax = b$ mit positiv definitem $A \in \mathbb{R}^{n\times n}$ ist auch das eindeutige Minimum der Funktion

$$Q : \mathbb{R}^n \to \mathbb{R}, \qquad Q(x) = \frac{1}{2}x^T Ax - x^T b. \tag{6.6}$$

Beweis: Q ist ein quadratisches Polynom, also beliebig oft differenzierbar. Der Gradient ist (wegen $A = A^T$)

$$\nabla Q(x) = Ax - b, \tag{6.7}$$

die einzige Lösung von $\nabla Q(x) = 0$ ist die Lösung x^* von $Ax = b$. Die Hessematrix der 2. partiellen Ableitungen im Punkt x^* ist $H(x^*) = A$, alle höheren partiellen Ableitungen sind Null. Also lautet die Taylor-Entwicklung im Punkt x^* für jedes $x \neq x^*$

$$Q(x) = Q(x^*) + \frac{1}{2}(x - x^*)^T A(x - x^*) > Q(x^*).$$

Ausgehend von einem (beliebigen) Startvektor $x^{(0)}$ bestimmt das CG-Verfahren linear unabhängige "Suchrichtungen" $d^{(k)}$ und Näherungen $x^{(k)}$, $k \geq 0$, so dass

$$Q(x^{(k)}) = \min\{Q(x) | x \in x^{(0)} + \mathrm{Span}(d^{(0)}, \ldots, d^{(k-1)})\} \tag{6.8}$$

gilt. Das bedeutet, dass in Schritt k das Minimum von Q über einen affinen Teilraum der Dimension k gebildet wird, der mit Aufpunkt $x^{(0)}$ und den aufspannenden Vektoren der ersten k Suchrichtungen definiert ist. Weil der Funktionswert von Q dabei in jedem Schritt verringert wird, spricht man von einem "Abstiegsverfahren". Nach höchstens n Schritten (also wenn der Teilraum ganz $\mathbb{R}^n$ ist) liegt die Lösung x^* in diesem affinen Raum.

Bild 6.3.1 veranschaulicht dies im Fall $n = 2$. Das Bild zeigt die Höhenlinien von Q: dies sind die Lösungsmengen von $0 = Q(x) - c = \frac{1}{2}x^T Ax - b^T x - c$, also Ellipsen im $\mathbb{R}^2$. Ausgehend von $x^{(0)}$ wird entlang der Richtung

$$d^{(0)} = -\nabla Q(x^{(0)}) = b - Ax^{(0)} \qquad \text{vgl. (6.7)}$$

das Minimum von Q bestimmt. Der Punkt $x^{(1)}$ liegt dann auf der Höhenlinie, die von dem Strahl $x^{(0)} + td^{(0)}$, $t > 0$, tangiert wird. Die neue Suchrichtung $d^{(1)}$ wird in der Ebene von $d^{(0)}$ und $-\nabla Q(x^{(1)})$ so gewählt, dass die Minimalbeziehung in (6.8) erreicht werden kann. Man sieht am Bild, dass im $\mathbb{R}^2$ die Suche entlang $d^{(1)}$ bereits das Minimum x^* von Q liefert.

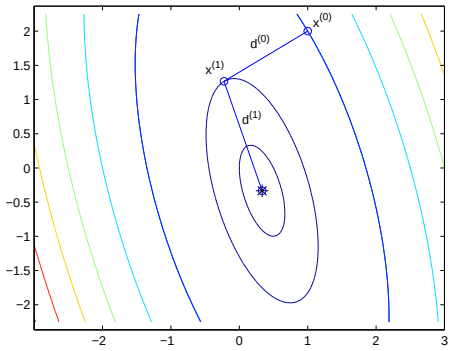

Abbildung 6.2: CG-Verfahren im $\mathbb{R}^2$

6.3.2 Algorithmus: CG-Verfahren

Gegeben: $A \in \mathbb{R}^{n\times n}$ positiv definit, $b \in \mathbb{R}^n$

Startwert: $x^{(0)} \in \mathbb{R}^n$

Initialisierung: $g^{(0)} = Ax^{(0)} - b; \quad d^{(0)} = -g^{(0)}$

Für $k = 0, 1, 2, \ldots$

$$x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)} \qquad \text{mit} \quad \alpha_k = \frac{(g^{(k)})^T g^{(k)}}{(d^{(k)})^T A d^{(k)}}$$
$$g^{(k+1)} = g^{(k)} + \alpha_k A d^{(k)}$$
$$d^{(k+1)} = -g^{(k+1)} + \beta_k d^{(k)} \qquad \text{mit} \quad \beta_k = \frac{(g^{(k+1)})^T g^{(k+1)}}{(g^{(k)})^T g^{(k)}}$$

Bemerkung: Das CG-Verfahren ist sehr effizient: die Berechnung der neuen Suchrichtung $d^{(k+1)}$ erfordert nur ein Matrix-Vektor-Produkt $Ad^{(k)}$ und zwei Skalarprodukte. Dies lässt sich genau wie beim Jacobi- und Gauss-Seidel-Verfahren ausnutzen, falls A sehr viele Nullen enthält.

Die Besonderheit der Suchrichtungen $d^{(k)}$ beim CG-Verfahren wird deutlich, wenn wir das Skalarprodukt

$$\langle x, y\rangle_A = x^T A y, \qquad x, y \in \mathbb{R}^n, \tag{6.9}$$

einführen. Dass hiermit ein Skalarprodukt definiert ist, folgt sofort aus der positiven Definitheit von A. Die zugehörige Norm $\|x\|_A = \sqrt{x^T Ax}$ haben wir schon im Beweis von Satz 6.2.12 verwendet. Das wichtigste Resultat lautet:

Solange $d^{(k)} \neq 0$ gilt, sind $d^{(0)}, \ldots, d^{(k)}$ paarweise "A-konjugiert", d.h. sie bilden ein Orthogonalsystem bzgl. des neuen Skalarprodukts:

$$\langle d^{(i)}, d^{(j)}\rangle_A = (d^{(i)})^T A d^{(j)} = 0 \quad \text{für} \quad i \neq j.$$

Die A-Konjugiertheit wird durch die spezielle Wahl von β_k erreicht. Der Beweis per Induktion ist etwas komplizierter, siehe z.B. Schaback/Wendland, Satz 16.5.

Die A-Konjugiertheit ist einerseits die Begründung für die lineare Unabhängigkeit der $d^{(j)}$: paarweise Orthogonalität bezüglich dieses Skalarprodukts ergibt

$$\left\| \sum_{j=0}^{k} c_j d^{(j)} \right\|_A^2 = \sum_{j=0}^{k} c_j^2 \left\| d^{(j)} \right\|_A^2 > 0$$

für jede nicht-triviale Linearkombination von $d^{(0)}, \ldots, d^{(k)}$.

Andererseits ist die A-Konjugiertheit genau der Grund dafür, dass die Minimierung (6.8) eintritt. Dies wollen wir kurz überlegen:

(i) Die Formel $x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$ mit dem angegebenen α_k liefert zunächst nur das Minimum von Q auf dem Strahl $x^{(k)} + td^{(k)}$, $t > 0$.

Begründung: Der Wert von α_k ergibt sich dabei aus der eindimensionalen Minimierungsaufgabe für $h(t) = Q(x^{(k)} + td^{(k)})$. Mit der Kettenregel und (6.7) erhalten wir

$$h'(t) = \nabla Q(x^{(k)} + td^{(k)})^T d^{(k)} = (Ax^{(k)} + tAd^{(k)} - b)^T d^{(k)}.$$

Berücksichtigt man $g^{(k)} = Ax^{(k)} - b$ im Algorithmus, so ist

$$h'(t) = (g^{(k)} + tAd^{(k)})^T d^{(k)}.$$

Aus $h'(t) = 0$ ergibt sich der Wert $t = \frac{-(g^{(k)})^T d^{(k)}}{(d^{(k)})^T A d^{(k)}}$. Der Zähler lässt sich umformen zu $(g^{(k)})^T g^{(k)}$, weil die vorherige Suchrichtung $d^{(k-1)}$ orthogonal zu $g^{(k)} = \nabla Q(x^{(k)})$ ist.

(ii) Die mehrdimensionale Minimierungsaufgabe für

$$F(t_0, \ldots, t_{k-1}) = Q(x^{(0)} + t_0 d^{(0)} + \cdots + t_{k-1} d^{(k-1)})$$

führt auf die Bedingung $\nabla F = 0$, was komponentenweise die Gleichungen

$$(\nabla Q(x^{(0)} + t_0 d^{(0)} + \cdots + t_{k-1} d^{(k-1)}))^T d^{(j)} = 0, \qquad j = 0, \ldots, k-1,$$

ergibt. Einsetzen von $\nabla Q(x) = Ax - b$ ergibt

$$(Ax^{(0)} - b + t_0 Ad^{(0)} + t_1 Ad^{(1)} + \cdots + t_{k-1} Ad^{(k-1)})^T d^{(j)} = 0, \qquad j = 0, \ldots, k-1. \quad (6.10)$$

Die A-Konjugiertheit ergibt für $j = 0$ die Beziehung $(Ax^{(0)} - b + t_0 Ad^{(0)})^T d^{(0)} = 0$, also $t_0 = \alpha_0$ wie in (i). Wegen $x^{(1)} = x^{(0)} + \alpha_0 d^{(0)}$ verkürzt sich (6.10) zu

$$(Ax^{(1)} - b + t_1 Ad^{(1)} + \cdots + t_{k-1} Ad^{(k-1)})^T d^{(j)} = 0, \qquad j = 1, \ldots, k-1.$$

Für $j = 1$ ergibt sich analog $t_1 = \alpha_1$, usw.

6.3.3 Bemerkung:

a) Die **Residuen** $g^{(k)} = Ax^{(k)} - b$ sind paarweise orthogonal, ihre Norm $\|g^{(k)}\|_2$ fällt monoton. Hieraus folgt ebenfalls $g^{(n)} = 0$, also $x^{(n)} = x^*$.

b) Die **Fehlervektoren** $e^{(k)} = x^{(k)} - x^*$ besitzen monoton fallende Norm $\|e^{(k)}\|_A$.

c) Mit $\kappa = \text{cond}_2(A)$ erhält man die a-priori Fehlerabschätzung

$$\|x^{(k)} - x^*\|_A \leq \left(\frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}}\right)^k \|x^{(0)} - x^*\|_A.$$

(d) Zur Reduzierung von $\kappa = \text{cond}_2(A)$ werden Verfahren der *Vorkonditionierung* verwendet, die das LGS in $C^{-1}Ax = C^{-1}b$ umwandeln.

e) Die Matlab-Routine `pcg`=**P**reconditioned **C**onjugate **G**radient steht sowohl für das gewöhnliche CG-Verfahren als auch für die vorkonditionierte Methode, siehe Matlab-Help.

6.4 Anwendungsbeispiel

Wir wollen die Leistungsfähigkeit der neuen Verfahren dieses Kapitels mit den direkten Methoden aus Kapitel 2 vergleichen. Dazu betrachten wir das Modellproblem zur Poisson-Gleichung aus Beispiel 2.3.9

$$\begin{array}{rcll} -\frac{\partial^2 u}{\partial x^2}(x,y) - \frac{\partial^2 u}{\partial y^2}(x,y) & = & f(x,y) & \text{für } (x,y) \in \Omega \\ u(x,y) & = & 0 & \text{für } (x,y) \in \partial\Omega \end{array} \tag{6.11}$$

auf dem Einheitsquadrat $\Omega = (0,1) \times (0,1) \subset \mathbb{R}^2$. Eine Lösung $u(x,y)$ ist i.allg. nicht geschlossen angebbar, so daß man sich numerisch eine Näherungslösung verschafft. Dazu wird zunächst das Gebiet Ω mit einem Quadratgitter überdeckt. Bild 6.4.0 zeigt ein solches Gitter zur Schrittweite $h = \frac{1}{m+1}$ in beide Koordinatenrichtungen, hier ist $m = 4$.

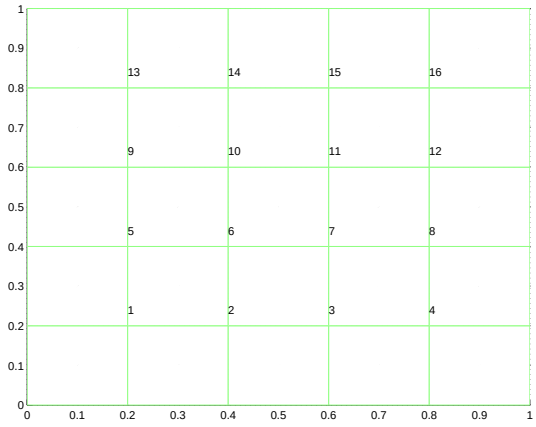

Abbildung 6.3: Diskretisierung der Poisson-Gleichung im Einheitsquadrat

Die inneren Gitterpunkte werden zeilenweise durchnumeriert. Ersetzt man dann in der obigen Differentialgleichung die 2. Ableitungen durch die entsprechenden zentralen Differenzenquotienten 2. Ordnung, erhält man

$$-h^{-2}\{u(x+h,y) - 2u(x,y) + u(x-h,y) + u(x,y+h) - 2u(x,y) + u(x,y-h)\} \cong f(x,y).$$

Dies wird nur für die inneren Gitterpunkte gefordert. Setzt man die Randwerte 0 direkt ein, so erhält man das lineare Gleichungssystem

$$Ax = b \tag{6.12}$$

für den Vektor $x \in \mathbb{R}^n$ der unbekannten Knotenwerte

$$x_i \sim u(P_i)\,, \quad P_i \quad \text{Gitterpunkt}\,.$$

Die Matrix hat die schon bekannte Gestalt

$$A = \left.\begin{bmatrix} B & -I & & \\ -I & B & -I & \\ & -I & B & \ddots \\ & & \ddots & \ddots \end{bmatrix}\right\} n \qquad B = \left.\begin{bmatrix} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & \ddots \\ & & \ddots & \ddots \end{bmatrix}\right\} m$$

mit der $m \times m$-Einheitsmatrix I. Die rechte Seite ist

$$b = h^2 (f(P_1), \ldots, f(P_n))^T .$$

Die Matrix A ist

a) eine dünn besetzte Bandmatrix mit der Bandbreite $2m + 1$;

b) symmetrisch, irreduzibel;

c) schwach diagonal dominant und positiv definit;

d) eine M-Matrix, d.h. die Theorie zur Optimierung des Relaxationsparameters ω für das SOR-Verfahren ist anwendbar.

Die Eigenwerte und zugehörigen Eigenvektoren von A lassen sich explizit angeben. Für $k, l = 1, \ldots, m$ ergibt sich:

$$\begin{aligned} \lambda_{kl} &= 4 - 2\,(\cos(kh\pi) + \cos(lh\pi)) \\ w^{kl} &= (\sin(ikh\pi)\ \sin(jlh\pi))_{i,j=1,\ldots,m} . \end{aligned}$$

Also ist (für $h \ll 1$)

$$\begin{aligned} \Lambda &:= \lambda_{\max} = 4 - 4\,\cos\,(1-h)\pi \ \approx\ 8 \\ \lambda &:= \lambda_{\min} = 4 - 4\,\cos\,(h\pi) = 4 - 4\,(1 - \frac{\pi^2}{2}\,h^2 + \mathcal{O}(h^4)) \ \approx\ 2\pi^2 h^2 \end{aligned}$$

und somit

$$\kappa := \mathrm{cond}_2(A) \ \approx\ \frac{4}{\pi^2 h^2} \tag{6.13}$$

Die Eigenwerte der Jacobimatrix $J_1 = -D^{-1}(L + R)$ sind

$$\mu_{kl} = \tfrac{1}{2}\,(\cos(kh\pi) + \cos(lh\pi)) \quad (k, l = 1, \ldots, m)$$

Folglich wird

$$\mu_{\max} = \cos(h\pi) = 1 - \frac{\pi^2}{2}\,h^2 + \mathcal{O}(h^4)\,,$$

bzw.

$$\rho(J_1) = \mu_{\max} \ \approx\ 1 - \frac{\pi^2}{2}\,h^2 . \tag{6.14}$$

Für die Iterationsmatrizen H_1 und $H_{\omega_{\mathrm{opt}}}$ des Gauß-Seidel-Verfahrens und des optimalen SOR-Verfahrens gilt dann

$$\rho\,(H_1) = \rho\,(J_1)^2 = 1 - \pi^2 h^2 + \mathcal{O}(h^4)$$

$$\rho\,(H_{\omega_{\mathrm{opt}}}) = \frac{1 - \sqrt{1 - \rho(J_1)^2}}{1 + \sqrt{1 - \rho(J_1)^2}} = \frac{1 - \pi h + \mathcal{O}(h^2)}{1 + \pi h + \mathcal{O}(h^2)} = 1 - 2\,\pi h + \mathcal{O}(h^2)\,.$$

Um den Anfangsfehler $\|x^{(0)} - x\|_2$ durch Anwendung der Iterationsverfahren um den Faktor $\epsilon \ll 1$ zu reduzieren, sind etwa

$$T(\epsilon) \approx \frac{\ln(\epsilon)}{\ln \rho(H)}, \quad H = I - B^{-1}A \quad \text{Iterationsmatrix},$$

Iterationsschritte erforderlich. Für die einzelnen Verfahren ergibt sich somit

$$\begin{aligned}
T_{GSV}(\epsilon) &\sim -\frac{\ln(1/\epsilon)}{\ln(1-\frac{\pi^2}{2}h^2)} \sim 2\frac{\ln(1/\epsilon)}{\pi^2 h^2} = \frac{2}{\pi^2} n \ \ln(1/\epsilon) \\
T_{ESV}(\epsilon) &\sim -\frac{\ln(1/\epsilon)}{\ln(1-\pi^2 h^2)} \sim \frac{\ln(1/\epsilon)}{\pi^2 h^2} = \frac{1}{\pi^2} n \ \ln(1/\epsilon) \\
T_{SOR}(\epsilon) &\sim -\frac{\ln(1/\epsilon)}{\ln(1-2\pi h)} \sim \frac{\ln(1/\epsilon)}{2\pi h} = \frac{1}{2\pi}\sqrt{n} \ \ln(1/\epsilon).
\end{aligned}$$

Das CG-Verfahren benötigt zur Reduzierung des Anfangsfehlers $\|x^{(0)} - x\|_2$ um den Faktor $\epsilon \ll 1/\sqrt{\kappa(A)} \sim \sqrt{2/n}$ etwa

$$T_{CG}(\epsilon) = \frac{\ln(2/\epsilon)}{\ln(\frac{1-1/\sqrt{\kappa}}{1+1/\sqrt{\kappa}})} \sim -\frac{\ln(2/\epsilon)}{\ln(1-\pi h)} \sim \frac{1}{2}\sqrt{\kappa} \ln(\frac{2}{\epsilon}) \sim \frac{1}{\pi h} \ln(\frac{2}{\epsilon})$$

Wir sehen, daß das Jacobi-Verfahren (GSV) am langsamsten ist. Das CG-Verfahren ist zwar nur halb so schnell wie das optimale SOR-Verfahren, erfordert aber nicht die Bestimmung eines Iterationsparameters ω.

Für die spezielle rechte Seite $f(x,y) = 2\pi^2 \sin(\pi x)\sin(\pi x)$ ist die exakte Lösung der obigen Randwertaufgabe gerade

$$u(x,y) = \sin(\pi x)\sin(\pi y). \tag{6.15}$$

Für den Diskretisierungsfehler der Differenzenapproximation läßt sich folgende Darstellung zeigen

$$\max_{P_i} |u(P_i) - x_i| = \frac{\pi^2}{12} h^2 + \mathcal{O}(h^4). \tag{6.16}$$

Zur Erzielung einer (absoluten) Genauigkeit von $\epsilon = 10^{-4}$ ist also die Gitterweite erforderlich:

$$h \sim \frac{\sqrt{12}}{\pi} 10^{-2} \sim 10^{-2}.$$

Die Anzahl von Unbekannten ist dann $n \sim 10^4$. Für die Spektralradien bzw. Konditionszahlen der betrachteten Iterationsverfahren und für die Anzahl der Iterationsschritte, die zur Erzielung einer Fehlergröße von etwa 10^{-4} erforderlich sind, ergibt sich in diesem Fall ($\ln(1/\epsilon) \sim 10$):

$$\begin{aligned}
\rho(J_1) &\sim 0,9995, & T_J(\epsilon) &\sim 20.000, \\
\rho(H_1) &\sim 0.999 \ , & T_{GS}(\epsilon) &\sim 10.000, \\
\rho(H_{\omega*}) &\sim 0,995 \ , & T_{SOR}(\epsilon) &\sim \ \ 170, \\
\kappa(A) &\sim 5.000 \ , & T_{CG}(\epsilon) &\sim \ \ 340.
\end{aligned}$$

Zum Vergleich der Effizienz der Iterationsverfahren muß natürlich auch der Aufwand pro Iterationsschritt berücksichtigt werden. Für die Anzahl OP der arithmetischen Operationen (1 Multiplikation + 1 Addition) pro Iterationsschritt gilt

$$\begin{aligned} &\mathrm{OP}_{J_1} \approx \mathrm{OP}_{H_1} \approx \mathrm{OP}_{H_\omega} \approx 6\,n\,, \\ &\mathrm{OP}_{CG} \approx 10\,n\,. \end{aligned}$$

Als Endresultat finden wir, daß zur Bestimmung der Lösung des durch Diskretisierung der Randwertaufgabe (6.11) entstehenden $(n \times n)$-Gleichungssystems $Ax = b$ das Jacobi-Verfahren und das Gauß-Seidel-Verfahren $\mathcal{O}(n^2)$ Operationen benötigen. Zur Lösung des Gleichungssystems $Ax = b$ mit einem direkten Verfahren würde man das Cholesky-Verfahren verwenden. Bei Berüchsichtigung der speziellen Struktur der Modellmatrix erfordert dies $\mathcal{O}(n^2) = \mathcal{O}(m^2 n)$ Operationen zur Berechnung der Zerlegung $A = LL^T$ und weitere $\mathcal{O}(n^{3/2}) = \mathcal{O}(mn)$ Operationen für das Vorwärts- und Rückwärtseinsetzen. Damit scheint das direkte Verfahren dem Gauß-Seidel-Verfahren z.B. überlegen zu sein. Es ist jedoch zu berücksichtigen, daß letzteres nur $\mathcal{O}(n)$ Speicherplätze benötigt im Gegensatz zu den $\mathcal{O}(n^{3/2}) = \mathcal{O}(mn)$ für das Cholesky-Verfahren. In den letzten Jahren wurden sehr effiziente Verfahren zur Lösung von Problemen des obigen Typs entwickelt, die im wesentlichen die n Unbekannten mit $\mathcal{O}(n)$ Operationen berechnen.