6 Lineare Gleichungssysteme II (Iterative Verfahren)

Für sehr große Gleichungssysteme mit $n \gg 1.000$ ist die Gauß-Elimination nur sehr schwer zu realisieren, da sie zu viel Speicherplatz erfordert. Für eine $n \times n$-Matrix A mit $n = 10^6$ und Bandbreite $m = 10^2$ sind dies bereits 10^8 Speicherplätze, was die Arbeitsspeicherkapazität der meisten zur Zeit im Einsatz befindlichen Rechenanlagen übersteigt. Zur Durchführung der Elimination müsste man in diesem Fall also mit Hintergrundspeichern arbeiten, was wegen des erforderlichen Datentransfers die Rechenzeit stark verlängert. Bei vielen in der Praxis auftretenden großen Gleichungssystemen hat man es jedoch mit sehr dünn besetzten Bandmatrizen mit nur $5 - 25$ von Null verschiedenen Elementen pro Zeile zu tun. Die im Folgenden betrachteten „iterativen Verfahren" benötigen zur näherungsweisen Lösung des Gleichungssystems $Ax = b$ nicht viel mehr Speicherplatz, als zur Speicherung von A selbst erforderlich ist.

Als erstes betrachten wir Fixpunktiterationen zur Lösung des Systems $Ax = b$ mit einer regulären $n \times n$-Matrix A und einem n-Vektor b. Zur Konstruktion solcher Iterationsvorschriften geht man etwa wie folgt vor:

Das Gleichungssystem $Ax = b$ lautet ausgeschrieben

$$a_{jj}x_j + \sum_{\substack{k=1 \\ k \neq j}}^{n} a_{jk}x_k = b_j, \quad j = 1, \ldots, n.$$

Im Falle $a_{jj} \neq 0$ ist dies äquivalent zu

$$x_j = \frac{1}{a_{jj}} \Big\{ b_j - \sum_{\substack{k=1 \\ k \neq j}}^{n} a_{jk}x_k \Big\}, \quad j = 1, \ldots, n.$$

Das sog. „Gesamtschritt-Verfahren" (oder auch „Jacobi-Verfahren") erzeugt Iterierte $x^t \in \mathbb{R}^n$, $t = 1, 2, \ldots$, durch die Iterationsvorschrift

$$x_j^t = \frac{1}{a_{jj}} \Big\{ b_j - \sum_{\substack{k=1 \\ k \neq j}}^{n} a_{jk}x_k^{t-1} \Big\}, \quad j = 1, \ldots, n. \tag{6.0.1}$$

Zum Zeitpunkt der Berechnung von x_j^t sind die vorausgehenden neuen Komponenten x_r^t, $r < j$, bereits berechnet. Zur Beschleunigung der Konvergenz liegt es also nahe, diese Zusatzinformation schon zur Berechnung von x_j^t auszunutzen. Diese Idee ist die Grundlage des sog. „Einzelschritt-Schritt-Verfahrens" (oder auch „Gauß-Seidel[1]-Verfahrens"):

$$x_j^t = \frac{1}{a_{jj}} \Big\{ b_j - \sum_{k<j} a_{jk}x_k^t - \sum_{k>j} a_{jk}x_k^{t-1} \Big\}, \quad j = 1, \ldots, n. \tag{6.0.2}$$

[1]Philipp Ludwig von Seidel (1821–1896): Deutscher Mathematiker; Professor in München; Beiträge zur Analysis (u. a. Methode der kleinsten Fehlerquadrate) owie Himmelsmechanik und Astronomie.

6.1 Fixpunktiterationen

Zur kompakteren Schreibweise der betrachteten Iterationsverfahren führen wir die Aufspaltung $A = D + L + R$ ein, wobei

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

Damit schreibt sich das Jacobi-Verfahren in der Form

$$x^t = D^{-1}\{b - (L+R)x^{t-1}\} = \underbrace{-D^{-1}(L+R)}_{J}\, x^{t-1} + D^{-1}b,$$

mit der sog. "Jacobi-Matrix" J und das Gauß-Seidel-Verfahren in der Form

$$x^t = D^{-1}\{b - Lx^t - Rx^{t-1}\} = \underbrace{-(D+L)^{-1}R}_{H_1} x^{t-1} + (D+L)^{-1}b.$$

mit der sog. „Gauß-Seidel-Matrix" H_1 (Die Notation H_1 für die Gauß-Seidel-Matrix wird später klar werden.). Beide Verfahren besitzen also die Gestalt

$$x^t = Bx^{t-1} + c \tag{6.1.3}$$

mit einer sog. „Iterationsmatrix" B. Konvergiert nun die Folge der Iterierten $(x^{(t)})_{t\in\mathbb{N}}$ gegen einen Vektor $x \in \mathbb{R}^n$, so gilt für diesen offenbar

$$x = Bx + c, \tag{6.1.4}$$

d. h. er ist ein „Fixpunkt" der Abbildung $g : x \to Bx + c$; daher auch die Bezeichnung „Fixpunktiteration". Ein sinnvolles iteratives Verfahren dieser Art muss also so gebaut sein, dass die Fixpunkte von g automatisch Lösungen des ursprünglichen Gleichungssystems $Ax = b$ sind. Dies ist beim Jacobi- und beim Gauß-Seidel-Verfahren aufgrund ihrer Konstruktion der Fall. Zur Konstruktion allgemeinerer iterativer Verfahren dieses Typs wählt man etwa eine reguläre $n\times n$-Matrix C und iteriert ausgehend von der Beziehung

$$Ax = b \quad \leftrightarrow \quad Cx = Cx - Ax + b \quad \leftrightarrow \quad x = x + C^{-1}(b - Ax)$$

in der Form

$$x^t = \underbrace{(I - C^{-1}A)}_{=:\,B}\, x^{t-1} + \underbrace{C^{-1}b}_{=:\,c}. \tag{6.1.5}$$

Dies wird in der Praxis auf dem Rechner als sog. „Defektkorrekturiteration2' realisiert, bei der in jedem Schritt im wesentlichen ein lineares Gleichungssystem mit der gewählten

Matrix C gelöst werden muss:

$$d^{t-1} = Ax^{t-1} - b\,, \qquad C\delta x^t = d^{t-1}\,, \qquad x^t = x^{t-1} - \delta x^t.$$

Ein hinreichendes Kriterium für die Konvergenz der Iteration (6.1.3) ist nach dem Banachschen Fixpunktsatz, dass

$$\|B\| < 1$$

für irgendeine Matrizennorm $\|\cdot\|$ auf $\mathbb{R}^{n\times n}$. Die Gültigkeit dieser Beziehung kann aber für eine konkrete Matrix sehr wohl von der speziellen Wahl der Norm abhängen. Daher verwendet man zur Charakterisierung der Fixpunktiteration besser den sog. „Spektralradius“ der Iterationsmatrix:

$$\mathrm{spr(B)} := \max\{\,|\lambda| : \lambda \in \sigma(\mathrm{B})\,\}.$$

Hierbei bezeichnet $\sigma(B) \subset \mathbb{C}$ das „Spektrum“ der Matrix B, d. h.: die Menge ihrer Eigenwerte. Offenbar ist spr(B) der Radius der kleinsten Kreisscheibe um den Nullpunkt in der komplexen Zahlenebene, die alle Eigenwerte von B enthält. Mit einer beliebigen natürlichen Matrizennorm gilt

$$\mathrm{spr(B)} \le \|\mathrm{B}\|. \tag{6.1.6}$$

Für symmetrisches B ist sogar

$$\mathrm{spr(B)} = \|\mathrm{B}\|_2 = \sup_{\mathrm{x}\in\mathbb{R}^n\setminus\{0\}} \frac{\|\mathrm{Bx}\|_2}{\|\mathrm{x}\|_2}; \tag{6.1.7}$$

jedoch ist $\mathrm{spr}(\cdot)$ keine Norm auf $\mathbb{R}^{n\times n}$, da i. Allg. die Dreiecksungleichung nicht gilt.

Satz 6.1 (Fixpunktiteration): *Die durch*

$$x^t = Bx^{t-1} + c \tag{6.1.8}$$

erzeugten Iterierten $x^t \in \mathbb{R}^n$, $t = 1, 2, \ldots$, konvergieren genau dann für jeden Startwert $x^0 \in \mathbb{R}^n$ gegen die Lösung $x \in \mathbb{R}^n$ der Fixpunktgleichung $x = Bx + c$, wenn

$$\mathrm{spr(B)} < 1. \tag{6.1.9}$$

Im Falle der Konvergenz ist das asymptotische Konvergenzverhalten bzgl. einer beliebigen Vektornorm $\|\cdot\|$ charakterisiert durch

$$\sup_{x^0\in\mathbb{R}^n} \limsup_{t\to\infty} \Big(\frac{\|x^t - x\|}{\|x^0 - x\|}\Big)^{1/t} = \mathrm{spr}(B). \tag{6.1.10}$$

Beweis: Wir führen die Fehlervektoren $e^t := x^t - x$ ein und finden (wegen $x = Bx + c$)

$$e^t = x^t - x = Bx^{t-1} + c - \underbrace{(Bx + c)}_{=\,x} = Be^{t-1},$$

d. h. $e^t = B^t e^0$, $t \in \mathbb{N}$.

(i) Im Fall $\mathrm{spr(B)} < 1$ existiert gemäß Hilfssatz 6.1 eine natürliche Matrizennorm $\|\cdot\|_\varepsilon$, so dass

$$\|B\|_\varepsilon \leq \mathrm{spr(B)} + \varepsilon < 1$$

für ein $\varepsilon < 1 - \mathrm{spr(B)}$. Folglich konvergiert in der zugehörigen Vektornorm $\|\cdot\|_\varepsilon$:

$$\|e^t\|_\varepsilon = \|B^t e^0\|_\varepsilon \leq \|B^t\|_\varepsilon \, \|e^0\|_\varepsilon \leq \|B\|_\varepsilon^t \, \|e^0\|_\varepsilon \to 0, \quad (t \to \infty).$$

Aufgrund der Äquivalenz aller Normen auf $\mathbb{R}^n$ konvergiert also $x^t \to x \, (t \to \infty)$.

(ii) Aus der Konvergenz der Iteration (für jeden Startwert) folgt bei Wahl von $x^0 = w + x$ mit einem Eigenvektor $w \in \mathbb{R}^n \setminus \{0\}$ zum betragsgrößten Eigenwert λ von B:

$$\lambda^t w = B^t w = B^t e^0 = e^t \to 0 \quad (t \to \infty).$$

Dies impliziert notwendig $|\lambda| < 1$ für $\lambda \in \sigma(B)$, d. h. $\mathrm{spr(B)} < 1$. Als Nebenprodukt erhalten wir noch die Beziehung

$$\Big(\frac{\|e^t\|}{\|e^0\|}\Big)^{1/t} = |\lambda|.$$

(iii) Zu beliebig kleinen $\varepsilon > 0$ sei wieder $\|\cdot\|_\varepsilon$ eine natürliche Matrizennorm mit $\|B\|_\varepsilon \leq \mathrm{spr(B)} + \varepsilon$. Dann existieren Zahlen $m, M > 0$, so dass für die gegebene (beliebige) Vektornorm $\|\cdot\|$ gilt:

$$m\|x\| \leq \|x\|_\varepsilon \leq M\|x\|, \quad x \in \mathbb{R}^n,$$

Damit erhalten wir

$$\begin{aligned} \|e^t\| &\leq \frac{1}{m}\|e^t\|_\varepsilon = \frac{1}{m}\|B^t e^0\|_\varepsilon \leq \frac{1}{m}\|B\|_\varepsilon^t \|e^0\|_\varepsilon \\ &\leq \frac{M}{m}(\mathrm{spr(B)} + \varepsilon)^{\mathrm{t}}\|\mathrm{e}^0\|, \end{aligned}$$

bzw. wegen $\big(\frac{M}{m}\big)^{1/t} \to 1 \, (t \to \infty)$:

$$\limsup_{t\to\infty} \Big(\frac{\|e^t\|}{\|e^0\|}\Big)^{1/t} \leq \mathrm{spr(B)} + \varepsilon.$$

Da $\varepsilon > 0$ beliebig klein gewählt werden kann, ergibt sich die Behauptung. Q.E.D.

Wir tragen noch den im obigen Beweis verwendeten Hilfssatz nach:

Hilfssatz 6.1 (Spektralradius): *Für jede Matrix $B \in \mathbb{R}^{n\times n}$ gibt es zu jedem beliebig kleinen $\varepsilon > 0$ eine natürliche Matrizennorm $\|\cdot\|_\varepsilon$, so dass*

$$\mathrm{spr}(B) \leq \|B\|_\varepsilon \leq \mathrm{spr}(B) + \varepsilon. \tag{6.1.11}$$

Beweis: Die Matrix B ist ähnlich zu einer Dreiecksmatrix

$$B = T^{-1}RT\,, \quad R = \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{bmatrix},$$

mit den Eigenwerten von B auf der Hauptdiagonalen, d. h.:

$$\mathrm{spr(B)} = \max_{1\le \mathrm{i} \le \mathrm{n}} |\mathrm{r_{ii}}|.$$

Für ein beliebiges $\delta \in (0,1]$ setzen wir

$$S_\delta = \begin{bmatrix} 1 & & & 0 \\ & \delta & & \\ & & \ddots & \\ 0 & & & \delta^{n-1} \end{bmatrix}, \qquad R_0 = \begin{bmatrix} r_{11} & & 0 \\ & \ddots & \\ 0 & & r_{nn} \end{bmatrix},$$

$$Q_\delta = \begin{bmatrix} 0 & r_{12} & \delta r_{13} & \cdots & \delta^{n-2} r_{1n} \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & \delta r_{n-2,n} \\ & & & \ddots & r_{n-1,n} \\ & & & & 0 \end{bmatrix},$$

und haben damit

$$R_\delta := S_\delta^{-1} R S_\delta = \begin{bmatrix} r_{11} & \delta r_{12} & \cdots & \delta^{n-1} r_{1n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \delta r_{n-1,n} \\ 0 & & & r_{nn} \end{bmatrix} = R_0 + \delta Q_\delta.$$

Wegen der Regularität von $S_\delta^{-1}T$ wird durch

$$\|x\|_\delta := \|S_\delta^{-1} T x\|_2\,, \quad x \in \mathbb{R}^n,$$

eine Vektornorm erklärt. Dann ist wegen $R = S_\delta R_\delta S_\delta^{-1}$:

$$B = T^{-1}RT = T^{-1} S_\delta R_\delta S_\delta^{-1} T$$

für alle $x \in \mathbb{R}^n$ und $y = S_\delta^{-1} T x$:

$$\begin{aligned} \|Bx\|_\delta &= \|T^{-1}S_\delta R_\delta S_\delta^{-1}Tx\|_\delta = \|R_\delta y\|_2 \\ &\le \|R_0 y\|_2 + \delta\|Q_\delta y\|_2 \le \{\max_{1\le i\le n}|r_{ii}| + \delta\mu\}\,\|y\|_2 \\ &\le \{\mathrm{spr(B)} + \delta\mu\}\|\mathrm{x}\|_\delta \end{aligned}$$

mit der Konstante

$$\mu = \Big(\sum_{i,j=1}^{n} |r_{ij}|^2\Big)^{1/2}.$$

Also ist

$$\sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Bx\|_\delta}{\|x\|_\delta} \le \mathrm{spr(B)} + \mu\delta,$$

und die Behauptung folgt mit $\delta = \varepsilon/\mu$. Q.E.D.

Der Spektralradius der Iterationsmatrix B bestimmt also das asymptotische Konvergenzverhalten der Iterierten x^t bzgl. jeder Vektornorm. Zu jedem $\varepsilon > 0$ existiert ein $t_\varepsilon \in \mathbb{N}$, so dass

$$\|x^t - x\| \le (\mathrm{spr(B)} + \varepsilon)^{\mathrm{t}}\|\mathrm{x}^0 - \mathrm{x}\| \quad (\mathrm{t} \ge \mathrm{t}_\varepsilon).$$

Dies lässt sich wie folgt interpretieren: Ist etwa $\mathrm{spr(B)} \le \rho < 1$, so erhält man nach t_0 Schritten die zur weiteren Reduktion des Fehlers $\|x^{t_0} - x\|$ um den Faktor 10^{-1} (d. h. zur Gewinnung einer Dezimalstelle Genauigkeit) erforderliche Anzahl von Iterationsschritten aus der Beziehung $\rho^t \le 10^{-1}$ zu

$$t \sim -\frac{1}{\log_{10}\rho} = -\frac{\ln(10)}{\ln(\rho)}. \tag{6.1.12}$$

In ungünstigsten Fällen mit z. B. $\mathrm{spr(B)} \sim 0.99$ ist $t_1 \sim 230$. Für Gleichungssysteme der Größenordnung $n > 1000$ bedeutet dies einen beträchtlichen Rechenaufwand zur Erlangung einer akzeptablen Genauigkeit.

Abbruchkriterien

Bei iterativen Verfahren ist es erforderlich, ein Abbruchkriterium anzugeben. Zunächst erhalten wir durch Anwendung des Banachschen Fixpunktsatzes die Fehlerabschätzung

$$\|x^t - z\| \le \frac{q}{1-q}\,\|x^t - x^{t-1}\|, \tag{6.1.13}$$

mit der „Kontraktionskonstante“ $q = \|B\| < 1$. Bei vorgegebener Fehlertoleranz $\varepsilon > 0$ könnte man das Verfahren dann abbrechen, sobald für die relative Änderung gilt:

$$\frac{\|\delta^t\|}{\|x^t\|} \le \frac{\|B\|}{1-\|B\|}\,\varepsilon. \tag{6.1.14}$$

Zur Realisierung dieser Strategie wird aber eine Schätzung für die Norm $\|B\|$ bzw. für $\mathrm{spr(B)}$ benötigt. Diese muss indirekt aus den berechneten Iterierten x^t, d. h. *a posteriori*

im Verlauf der Rechnung, gewonnen werden. In der Regel kann die Iterationsmatrix $B = I - C^{-1}A$ mit vertretbarem Aufwand gar nicht explizit gebildet werden. Methoden zur Bestimmung von spr(B) werden im Kapitel über Eigenwertaufgaben diskutiert.

Alternativ könnte man auch das Residuum $\|Ax^t - b\|$ abfragen. Über die Argumentation

$$e^t = x^t - x = A^{-1}(Ax^t - b)\,, \quad b = Ax$$
$$\|e^t\| \le \|A^{-1}\|\,\|Ax^t - b\|\,, \quad \frac{1}{\|b\|} \ge \frac{1}{\|A\|\,\|x\|}$$

erhält man

$$\frac{\|e^t\|}{\|x\|} \le \mathrm{cond(A)}\,\frac{\|\mathrm{Ax^t - b}\|}{\|\mathrm{b}\|}.$$

Dies hat allerdings den Nachteil, daß dazu extra Ax^t berechnet werden müßte, und daß im Falle $\mathrm{cond(A)} \gg 1$ eine starke Unterschätzung des tatsächlichen Fehlers erfolgt. Zudem ist cond(A) selbst natürlich wieder nur schwer schätzbar (noch schwieriger als spr(B)). Wir verweisen hierfür auch auf das Kapitel über Eigenwertaufgaben.

Konstruktion von Iterationsverfahren

Bei der Konstruktion der Iterationsverfahren etwa auf dem ersten der angegebenem Wege, d. h. bei der Wahl der Matrix C, müssen zwei wesentliche Ziele berücksichtigt werden:

- $\mathrm{spr(I - C^{-1}A)}$ soll möglichst klein sein.
- Die Gleichungssysteme $Cx^t = (C - A)x^{t-1} + b$ sollen möglichst leicht (und mit wenig zusätzlichem Speicherplatzbedarf!) lösbar sein.

Leider widersprechen sich diese beiden Prämissen; die extremen Lösungen sind:

$$C = A \quad \Rightarrow \quad \mathrm{spr(I - C^{-1}A)} = 0$$
$$C = D \quad \Rightarrow \quad \mathrm{spr(I - D^{-1}A)} \sim 1.$$

Man wird also einen gewissen Kompromiss eingehen. Inwieweit dies beim Jacobi- und beim Gauß-Seidel-Verfahren gelungen ist, wollen wir jetzt untersuchen. Zunächst ist festzustellen, dass Punkt (ii) in beiden Fällen gut erfüllt ist, denn in jedem Iterationsschritt ist beim Jacobi-Verfahren nur ein Diagonalsystem und beim Gauß-Seidel-Verfahren ein unteres Dreieckssystem zu lösen. Es wird außerdem nicht mehr Speicherplatz benötigt, als zur Speicherung der Matrix A erforderlich ist. Dies lässt vermuten, dass der Spektralradius von $I - C^{-1}A$ nicht besonders klein sein wird. Trotzdem lässt sich für eine große Klasse von Matrizen wenigstens die Konvergenz der Verfahren garantieren, wenn diese auch oft sehr langsam ist.

6.1.1 Jacobi- und Gauß-Seidel-Verfahren

Satz 6.2 (Starkes Zeilensummenkriterium): *Genügen die Zeilensummen der Matrix* $A \in \mathbb{R}^{n\times n}$ *der Bedingung (strikte Diagonaldominanz)*

$$\sum_{k=1,k\neq j}^{n} |a_{jk}| < |a_{jj}|\,, \quad j = 1,\ldots,n, \tag{6.1.15}$$

so ist $\mathrm{spr}(J) < 1$ *und* $\mathrm{spr}(H_1) < 1$*, d. h. Jacobi- und Gauß-Seidel-Verfahren konvergieren.*

Beweis: Seien $\lambda \in \sigma(J)$ bzw. $\mu \in \sigma(H_1)$ und v bzw. w zugehörige Eigenvektoren (beachte $a_{jj} \neq 0$), d. h.:

$$\lambda v = Jv = -D^{-1}(L+R)v$$

bzw.

$$\mu w = H_1 w = -(D+L)^{-1}Rw \iff \mu w = -D^{-1}(\mu L+R)w.$$

Hieraus folgt zunächst im Falle $\|v\|_\infty = \|w\|_\infty = 1$

$$|\lambda| \le \|D^{-1}(L+R)\|_\infty = \max_{j=1,\ldots,n} \Big\{ \frac{1}{|a_{jj}|} \sum_{k=1,k\neq j}^{n} |a_{jk}| \Big\} < 1.$$

Ferner ist

$$|\mu| \le \|D^{-1}(\mu L+R)\|_\infty \le \max_{1\le j\le n} \Big\{ \frac{1}{|a_{jj}|} \Big[\sum_{k<j} |\mu|\,|a_{jk}| + \sum_{k>j} |a_{jk}| \Big] \Big\}.$$

Im Falle $|\mu| \ge 1$ ergäbe sich der Widerspruch

$$|\mu| \le |\mu|\,\|D^{-1}(L+R)\|_\infty < |\mu|,$$

so dass auch $|\mu| < 1$ sein muss. Q.E.D.

Matrizen mit der Eigenschaft aus Satz 6.2 heißen „strikt diagonaldominant“. Für die Bedürfnisse der Praxis ist die Bedingung zu einschränkend; die einfache Modellmatrix aus Abschnitt 4.3

$$A = \left[\begin{array}{cccc} B & -I_4 & & \\ -I_4 & B & -I_4 & \\ & -I_4 & B & -I_4 \\ & & -I_4 & B \end{array}\right] \Bigg\} 16\,, \qquad B = \left[\begin{array}{cccc} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & -1 \\ & & -1 & 4 \end{array}\right] \Bigg\} 4$$

ist z. B. zwar diagonaldominant, aber nicht *strikt* diagonaldominant. Sie ist jedoch in einigen Zeilen (z. B. der ersten) strikt diagonaldominant. Dieser Umstand kann nun zum Nachweis der Konvergenz des Jacobi- und des Gauß-Seidel-Verfahrens verwendet werden.

Definition 6.1: *Eine Matrix* $A \in \mathbb{R}^{n\times n}$ *heißt „irreduzibel“, wenn es keine Permutationsmatrix* P *gibt, so dass*

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

(simultane Zeilen- und Spaltenvertauschung) mit Matrizen $\tilde{A}_{11} \in \mathbb{R}^{p\times p}$, $\tilde{A}_{22} \in \mathbb{R}^{q\times q}$, $\tilde{A}_{21} \in \mathbb{R}^{q\times p}$, $p, q > 0$, $p + q = n$.

Hilfssatz 6.2 (Irreduzibilität): *Eine Matrix* $A \in \mathbb{R}^{n\times n}$ *ist genau dann irreduzibel, wenn der zugehörige gerichtete Graph*

$$G(A) := \left\{ \textit{Knoten } P_1, ..., P_n\,,\ \textit{Kanten } \overline{P_j P_k} \Leftrightarrow a_{jk} \neq 0,\ j, k = 1, ..., n \right\}$$

zusammenhängend ist, d. h.: wenn zu jedem Knotenpaar $\{P_j, P_k\}$ *eine gerichtete Kantenverbindung zwischen* P_j *und* P_k *existiert.*

Beweis: Die Reduzibilität von A lässt sich auch wie folgt formulieren: Es existiert eine (nicht–triviale) Zerlegung $N_n = J \cup K$ der Indexmenge $N_n = \{1, ..., n\}$, $J, K \neq \emptyset$, $J \cap K = \emptyset$, so dass $a_{jk} = 0$ für alle Paare $\{j, k\} \in J \times K$. Der Zusammenhang des Graphen $G(A)$ bedeutet nun, dass es zu je zwei Indizes j, k stets eine Kette von Indizes $i_1, \ldots, i_m \in \{1, \ldots, n\}$ gibt, so dass

$$a_{ji_1} \neq 0\,,\ a_{i_1 i_2} \neq 0\,, \ldots\,, a_{i_{m-1} i_m} \neq 0\,,\ a_{i_m k} \neq 0.$$

Hieraus liest man direkt die behauptete Charakterisierung ab (Übungsaufgabe). Q.E.D.

Für irreduzible Matrizen kann die Bedingung des starken Zeilensummenkriteriums deutlich abgemildert werden.

Satz 6.3 (Schwaches Zeilensummenkriterium): *Genügen die Zeilensummen einer irreduziblen Matrix* $A \in \mathbb{R}^{n\times n}$ *den Bedingungen*

$$\sum_{k=1, k\neq j}^{n} |a_{jk}| \leq |a_{jj}| \quad \textit{für} \quad j = 1, \ldots, n, \tag{6.1.16}$$

$$\sum_{k=1, k\neq r}^{n} |a_{rk}| < |a_{rr}| \quad \textit{für ein} \quad r \in \{1, \ldots, n\}, \tag{6.1.17}$$

so ist A *regulär und* $\mathrm{spr}(J) < 1$ *sowie* $\mathrm{spr}(H_1) < 1$*, d. h. Jacobi- und Gauß-Seidel-Verfahren konvergieren.*

Beweis: Wegen der Irreduzibilität von A ist notwendig

$$\sum_{k=1}^{n} |a_{jk}| > 0\,, \quad j = 1, \ldots, n,$$

und wegen der Diagonaldominanz folglich $a_{jj} \neq 0$, $j = 1, \ldots, n$. Jacobi- und Gauß-Seidel-Verfahren sind also durchführbar. Mit Hilfe der Diagonaldominanz erschließt man analog zum Beweis von Satz 6.2:

$$\mathrm{spr}(\mathrm{J}) \leq 1\,, \quad \mathrm{spr}(\mathrm{H}_1) \leq 1.$$

Angenommen, es gibt einen Eigenwert $\lambda \in \sigma(J)$ mit $|\lambda| = 1$. Sei $v \in \mathbb{C}^n$ ein zugehöriger normierter Eigenvektor, so dass $|v_s| = \|v\|_\infty = 1$. Es gilt dann

$$|\lambda|\,|v_i| \leq \frac{1}{|a_{ii}|} \sum_{k \neq i} |a_{ik}|\,|v_k|\,, \quad i = 1, \ldots, n. \tag{6.1.18}$$

Aufgrund der Irreduzibilität von A gibt es nun Indizes $i_1, \ldots, i_m$, so dass $a_{si_1} \neq 0\,, \ldots,$ $a_{i_m r} \neq 0$. Durch mehrfache Anwendung von (6.1.18) folgt so der Widerspruch ($|\lambda| = 1$)

$$\begin{aligned}
|v_r| &= |\lambda v_r| \leq \tfrac{1}{|a_{rr}|} \sum_{k \neq r} |a_{rk}|\,\|v\|_\infty && < \|v\|_\infty \\
|v_{i_m}| &= |\lambda v_{i_m}| \leq \tfrac{1}{|a_{i_m i_m}|} \Big\{ \sum_{k \neq i_m, r} |a_{i_m k}|\,\|v\|_\infty + \underbrace{|a_{i_m r}|}_{\neq 0}\,|v_r| \Big\} && < \|v\|_\infty \\
&\vdots \\
|v_{i_1}| &= |\lambda v_{i_1}| \leq \tfrac{1}{|a_{i_1 i_1}|} \Big\{ \sum_{k \neq i_1, i_2} |a_{i_1 k}|\,\|v\|_\infty + \underbrace{|a_{i_1 i_2}|}_{\neq 0}\,|v_{i_2}| \Big\} && < \|v\|_\infty \\
\|v\|_\infty &= |\lambda v_s| \leq \tfrac{1}{|a_{ss}|} \Big\{ \sum_{k \neq s, i_1} |a_{sk}|\,\|v\|_\infty + \underbrace{|a_{si_1}|}_{\neq 0}\,|v_{i_1}| \Big\} && < \|v\|_\infty .
\end{aligned}$$

Also muss $\mathrm{spr}(\mathrm{J}) < 1$ sein. Analog erschließt man unter Verwendung von (6.1.18) auch $\mathrm{spr}(\mathrm{H}_1) < 1$. Wegen $A = D(I - J)$ muss A dann regulär sein. Q.E.D.

6.1.2 SOR-Verfahren

Für die in der Praxis auftretenden großen, aber dünn besetzten Matrizen ist $\mathrm{spr}(\mathrm{J})$ bzw. $\mathrm{spr}(\mathrm{H}_1)$ nahe bei Eins, so dass Jacobi- und Gauß-Seidel-Verfahren viel zu langsam konvergieren. Man versucht daher, die Konvergenz durch Einführung eines (oder mehrerer) sog. „Relaxationsparameter“ zu beschleunigen. Beim „SOR-Verfahren“ (**S**uccessive **O**ver**R**elaxation method) berechnet man im t-ten Iterationsschritt ausgehend von dem Gauß-Seidel-Zwischenwert

$$\tilde{x}_j^t = \frac{1}{a_{jj}} \Big\{ b_j - \sum_{k<j} x_k^t - \sum_{k>j} x_k^{t-1} \Big\}$$

die nächste Iterierte x_j^t als Linearkombination

$$x_j^t = \omega \tilde{x}_j^t + (1 - \omega)\, x_j^{t-1}$$

mit einem Parameter $\omega \geq 1$. Im Fall $\omega = 1$ ist dies gerade das Gauß-Seidel-Verfahren. Im Falle $\omega < 1$ spricht man von „Unterrelaxation". Die Iterationsmatrix des Relaxationsverfahrens erhält man aus den Beziehungen

$$x^t = \omega D^{-1} \{ b - Lx^t - Rx^{t-1} \} + (1-\omega)\, x^{t-1}$$

als

$$H_\omega = (D + \omega L)^{-1} [\, (1-\omega)\, D - \omega R \,],$$

d. h.: Der Iterationsschritt lautet

$$x^t = H_\omega x^{t-1} + \omega\, (D + \omega L)^{-1} b. \tag{6.1.19}$$

Der folgende Hilfssatz zeigt, dass man sich beim Relaxationsverfahren auf den Parameterbereich $0 < \omega < 2$ beschränken muss.

Hilfssatz 6.3 (Relaxation): *Für eine beliebige Matrix $A \in \mathbb{R}^{n\times n}$ mit regulärem Diagonalanteil D gilt*

$$\operatorname{spr}(H_\omega) \geq |\omega - 1|\,, \quad \omega \in \mathbb{R}. \tag{6.1.20}$$

Beweis: Der Beweis ist überraschend einfach. Wir formen wie folgt um:

$$H_\omega = (D+\omega L)^{-1}[\,(1-\omega)\, D - \omega R\,] = (I + \omega \underbrace{D^{-1}L}_{=:\, L'})^{-1} \underbrace{D^{-1}D}_{=\, I} [\,(1-\omega)\, I - \omega \underbrace{D^{-1}R}_{=:\, R'}]$$

Dann gilt

$$\det(H_\omega) = \underbrace{\det(I+\omega L')}_{=\,1})^{-1} \cdot \underbrace{\det((1-\omega)\, I - \omega R')}_{=\,(1-\omega)^n} = (1-\omega)^n.$$

Wegen $\det(H_\omega) = \prod_{i=1}^n \lambda_i$ ($\lambda_i \in H_\omega$) folgt notwendigerweise

$$\operatorname{spr}(H_\omega) = \max_{1\leq i\leq n} |\lambda_i| \geq \Big(\prod_{i=1}^n |\lambda_i| \Big)^{1/n} = |1-\omega|.$$

Q.E.D.

Für positiv definite Matrizen lässt sich die Aussage von Hilfssatz 6.3 in gewisser Weise umkehren.

Satz 6.4 (SOR-Verfahren): *Für eine positiv definite Matrix $A \in \mathbb{R}^{n\times n}$ gilt*

$$\operatorname{spr}(H_\omega) < 1\,, \quad \textit{für} \quad 0 < \omega < 2\,; \tag{6.1.21}$$

insbesondere ist das Gauß-Seidel-Verfahren konvergent.

Beweis: Wegen der Symmetrie von A ist $R = L^T$, d. h. $A = L+D+L^T$. Sei $\lambda \in \sigma(H_\omega)$ beliebig für $0 < \omega < 2$, mit einem Eigenvektor $v \in \mathbb{R}^n$, d. h. $H_\omega v = \lambda v$. Es gilt also

$$\left((1-\omega)\, D - \omega L^T\right) v = \lambda\, (D+\omega L)\, v$$

bzw.

$$\omega\, (D+L^T)\, v = (1-\lambda)\, Dv - \lambda\omega L v.$$

Hiermit erschließt man

$$\begin{aligned} \omega A v &= \omega\, (D+L^T)\, v + \omega L v \\ &= (1-\lambda)\, Dv - \lambda\omega L v + \omega L v \\ &= (1-\lambda)\, Dv + \omega\, (1-\lambda)\, Lv\,, \end{aligned}$$

und

$$\begin{aligned} \lambda\omega A v &= \lambda\omega\, (D+L^T)\, v + \lambda\omega L v \\ &= \lambda\omega\, (D+L^T)\, v + (1-\lambda)\, Dv - \omega\, (D+L^T)\, v \\ &= (\lambda-1)\omega (D+L^T)\, v + (1-\lambda)\, Dv \\ &= (1-\lambda)(1-\omega)\, Dv - (1-\lambda)\, \omega L^T v. \end{aligned}$$

Wegen $v^T L v = v^T L^T v$ folgt

$$\begin{aligned} \omega v^T A v &= (1-\lambda)\, v^T D v + \omega\, (1-\lambda)\, v^T L v \\ \lambda\omega v^T A v &= (1-\lambda)(1-\omega)\, v^T D v - (1-\lambda)\, \omega v^T L v, \end{aligned}$$

und hieraus durch Addition

$$\omega\, (1+\lambda)\, v^T A v = (1-\lambda)\, (2-\omega)\, v^T D v.$$

Da mit A auch D positiv definit ist, gilt

$$v^T A v > 0, \quad v^T D v > 0.$$

Folglich ist unter Beachtung von $0<\omega<2$ notwendig $\lambda \neq \pm 1$, und es gilt

$$\mu := \frac{1+\lambda}{1-\lambda} = \frac{2-\omega}{\omega}\, \frac{v^T D v}{v^T A v} > 0.$$

Durch Auflösen nach λ erhalten wir schließlich

$$|\lambda| = \left| \frac{\mu - 1}{\mu + 1} \right| < 1,$$

was zu zeigen war. Q.E.D.

Definition 6.2: *Die qualitative Konvergenzaussagen der letzten Sätze lassen sich für eine gewisse Klasse von Matrizen wesentlich verschärfen. Man nennt die Matrix* $A \in \mathbb{R}^{n\times n}$ *mit der additiven Aufspaltung* $A = L+D+R$ *„konsistent geordnet", wenn die Eigenwerte der Matrizen*

$$J(\alpha) = -D^{-1}\{\alpha L + \alpha^{-1}R\}, \quad \alpha \in \mathbb{C},$$

unabhängig vom Parameter α *also stets gleich denen der Jacobi–Matrix* $J = J(1)$ *sind.*

Man kann zeigen, dass neben anderen die oben eingeführte Modellmatrix „konsistent geordnet" ist. Die Bedeutung dieser Eigenschaft besteht darin, dass man in diesem Fall explizit angeben kann, wie die Eigenwerte von J mit denen von H_ω zusammenhängen.

Satz 6.5 (Optimales SOR-Verfahren): *Die Matrix* $A \in \mathbb{R}^{n\times n}$ *sei konsistent geordnet und* $0 \le \omega \le 2$. *Dann besteht zwischen den Eigenwerten* $\mu \in \sigma(J)$ *und* $\lambda \in \sigma(H_\omega)$ *die Beziehung*

$$\lambda^{1/2}\omega\mu = \lambda + \omega - 1. \tag{6.1.22}$$

Beweis: Seien $\lambda, \mu \in \mathbb{C}$ zwei Zahlen, welche der Gleichung (6.1.22) genügen. Im Falle $0 \neq \lambda \in \sigma(H_\omega)$ ist dann $H_\omega v = \lambda v$ äquivalent zu

$$\left((1-\omega)I - \omega D^{-1}R\right) v = \lambda(I + \omega D^{-1}L)v$$

bzw.

$$(\lambda + \omega - 1)v = -\lambda^{1/2}\omega\left(\lambda^{1/2}D^{-1}L + \lambda^{-1/2}D^{-1}R\right) v = \lambda^{1/2}\omega J(\lambda^{1/2})\, v.$$

Also ist v Eigenvektor von $J(\lambda^{1/2})$ zum Eigenwert

$$\mu = \frac{\lambda + \omega - 1}{\lambda^{1/2}\omega}.$$

Mit der Voraussetzung an A folgt auch $\mu \in \sigma(J)$. Umgekehrt erhält man für $\mu \in \sigma(J)$ auf diese Weise auch $\lambda \in \sigma(H_\omega)$. Q.E.D.

Als direkte Folgerung aus diesem Resultat erhalten wir für konsistent geordnete Matrizen für das Gauß-Seidel-Verfahren (Fall $\omega = 1$) alternativ $\mathrm{spr}(\mathrm{H}_1) = 0$ oder die Beziehung

$$\mathrm{spr}(\mathrm{H}_1) = \mathrm{spr}(\mathrm{J})^2. \tag{6.1.23}$$

Im Falle $\mathrm{spr}(\mathrm{J}) < 1$ konvergiert das Jacobi-Verfahren. Zur Reduzierung des Fehlers um den Faktor 10^{-1} sind dann mit dem Gauß-Seidel-Verfahren nur halb so viele Iterationen erforderlich. Im allgemeinen ist das Gauß-Seidel-Verfahren dem Jacobi-Verfahren vorzuziehen. (Dies darf jedoch nicht generalisiert werden, da man Beispiele konstruieren kann, bei denen jeweils das eine, aber nicht das andere Verfahren konvergiert.)

Für konsistent geordnete Matrizen lässt sich aus der Identität (6.1.22) der „optimale“ Relaxationsparameter $\omega_{\text{opt}} \in (0,2)$ mit

$$\text{spr}(H_{\omega_{\text{opt}}}) \le \text{spr}(H_\omega)\,, \quad \omega \in (0,2),$$

explizit ableiten. Im Falle $\rho := \text{spr}(J) < 1$ gilt für $0 < \omega < 2$:

$$\text{spr}(H_\omega) = \begin{cases} \omega - 1 & , \quad \omega_{\text{opt}} \le \omega \\ \frac{1}{4}\left(\rho\,\omega + \sqrt{\rho^2\omega^2 - 4(\omega-1)}\right)^2 & , \quad \omega \le \omega_{\text{opt}}\,, \end{cases}$$

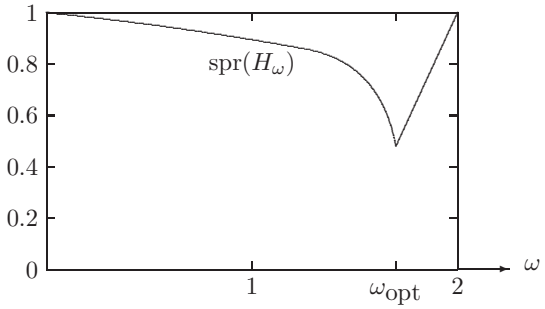


Abbildung 6.1: *Spektralradius der SOR-Matrix als Funktion von* ω

Dann ist

$$\omega_{\text{opt}} = \frac{2}{1+\sqrt{1-\rho^2}}, \qquad \text{spr}(H_{\omega_{\text{opt}}}) = \omega_{\text{opt}} - 1 = \frac{1-\sqrt{1-\rho^2}}{1+\sqrt{1-\rho^2}} < 1. \tag{6.1.24}$$

I. Allg, ist der genaue Wert für $\text{spr}(J)$ nicht bekannt. Da die linksseitige Ableitung der Funktion $f(\omega) = \text{spr}(H_\omega)$ für $\omega \to \omega_{\text{opt}}$ singulär wird, ist es bei Schätzungen von ω_{opt} besser, einen etwas zu großen als zu kleinen Wert zu nehmen. Mit Hilfe von Einschließungssätzen oder auch nur der Schranke $\rho \le \|J\|_\infty$ erhält man obere Schätzungen $\bar\rho \ge \rho$. Im Falle $\bar\rho < 1$ erhält man damit durch

$$\bar\omega \;:=\; \frac{2}{1+\sqrt{1-\bar\rho^2}} \ge \frac{2}{1+\sqrt{1-\rho^2}} = \omega_{\text{opt}}$$

eine obere Schätzung $\bar\omega \ge \omega_{\text{opt}}$ mit

$$\text{spr}(H_{\bar\omega}) = \bar\omega - 1 = \frac{1-\sqrt{1-\bar\rho^2}}{1+\sqrt{1-\bar\rho^2}} < 1\,. \tag{6.1.25}$$

Dies setzt natürlich voraus, dass die Formel (6.1.24) überhaupt anwendbar ist.

Beispiel 6.1: Konvergenzverbesserung durch Überrelaxation:

$$\text{spr}(H_1) = \text{spr}\,(J)^2 = \begin{cases} 0.81 \\ 0.99 \end{cases} \implies \text{spr}(H_{\omega_{\text{opt}}}) = \begin{cases} 0.39 \\ 0.8 \end{cases}\,.$$

6.2 Abstiegsverfahren

Im Folgenden betrachten wir eine Klasse von Verfahren zur Lösung linearer Gleichungssysteme, die primär auf symmetrische, positiv definite Koeffizientenmatrizen zugeschnitten sind.

Sei $A \in \mathbb{R}^{n\times n}$ eine (symmetrische) positiv definite Matrix, d. h.:

$$\begin{aligned} (Ax, y) &= (x, Ay)\,, \quad \forall\, x\,,\, y \in \mathbb{R}^n \\ (Ax, x) &> 0\,, \quad \forall\, x \in \mathbb{R}^{n\times n} \setminus \{0\}. \end{aligned} \tag{6.2.26}$$

Es bezeichnet wieder $(\cdot,\cdot)$ das euklidische Skalarprodukt auf $\mathbb{R}^n$ und $\|\cdot\|$ die eukidische Vektornorm. Zugehörig zur Matrix A werden das sog. „A-Skalarprodukt“ und die zugehörige „A-Norm“ definiert:

$$(x, y)_A := (Ax, y)\,, \quad \|x\|_A := (Ax, x)^{1/2}. \tag{6.2.27}$$

Wir haben früher schon einige wichtige Eigenschaften symmetrischer, positiv definiter Matrizen kennengelernt: Ihre Eigenwerte sind reell und positiv, $\lambda := \lambda_1 \le \ldots \le \lambda_n =: \Lambda$ und es existiert eine zugehörige Orthonormalbasis $\{w_1, \ldots, w_n\}$ von Eigenvektoren. Für den Spektralradius und die Spektralkonditonszahl gilt

$$\mathrm{spr}(A) = \Lambda\,, \quad \mathrm{cond_2(A)} = \frac{\Lambda}{\lambda}. \tag{6.2.28}$$

Grundlegend für das Folgende ist die Charakterisierung der Lösung $x \in \mathbb{R}^n$ des Gleichungssystems $Ax = b$ als Minimum eines quadratischen Funktionals auf $\mathbb{R}^n$.

Satz 6.6 (Minimierungseigenschaft): *Die Matrix A sei (symmetrisch) positiv definit. Die eindeutige Lösung des Gleichungssystems $Ax = b$ ist charakterisiert durch die Eigenschaft*

$$Q(x) < Q(y) \qquad \forall\, y \in \mathbb{R}^n\,,\, y \neq x, \tag{6.2.29}$$

mit dem quadratischen Funktional

$$Q(y) := \tfrac{1}{2}\,(Ay, y) - (b, y), \quad y \in \mathbb{R}^n. \tag{6.2.30}$$

Beweis: Sei zunächst $Ax = b$. Für $y \neq x$ ist dann

$$\begin{aligned} Q(y) - Q(x) &= \tfrac{1}{2}\,\{\,(Ay, y) - 2(b, y) - (Ax, x) + 2(b, x)\,\} \\ &= \tfrac{1}{2}\,\{\,(Ay, y) - 2(Ax, y) + (Ax, x)\,\} \\ &= \tfrac{1}{2}\,(A[x - y], x - y) > 0, \end{aligned}$$

wegen der Definitheit von A. Ist umgekehrt $Q(x) < Q(y)$, für $x \neq y$, d. h. ist x ein Minimum der Funktion Q auf $\mathbb{R}^n$, so muss notwendig $\mathrm{gradQ(x)} = 0$ sein. Dies bedeutet

gerade, dass gilt:

$$\frac{\partial Q}{\partial x_i}(x) = \frac{1}{2}\frac{\partial}{\partial x_i}\sum_{j,k=1}^{n} a_{jk}x_j x_k - \frac{\partial}{\partial x_i}\sum_{k=1}^{n} b_k x_k = \sum_{k=1}^{n} a_{ik}x_k - b_i = 0\,,$$

für $i = 1, \ldots, n$; man beachte $a_{jk} = a_{kj}$. Also ist $Ax = b$. Q.E.D.

Wir halten fest, dass der Gradient von Q in einem Punkt $y \in \mathbb{R}^n$ gegeben ist durch

$$\operatorname{grad} Q(y) = \tfrac{1}{2}(A + A^T)y - b = Ay - b. \tag{6.2.31}$$

(Dies ist gerade der „Defekt“ im Punkt y.) Die sog. „Abstiegsverfahren“ bestimmen nun ausgehend von einem geeigneten Startvektor $x^{(0)} \in \mathbb{R}^n$ eine Folge von Iterierten x^t, $t \in \mathbb{N}$, durch

$$x^{t+1} = x^t + \alpha_t r^t. \tag{6.2.32}$$

Dabei sind die r^t vorgegebene oder auch erst im Verlauf der Iteration berechnete “Abstiegsrichtungen”, und die “Schrittweiten” $\alpha_t \in \mathbb{R}$ sind durch die Vorschrift bestimmt (sog. “line search”):

$$Q(x^{t+1}) = \min_{\alpha\in\mathbb{R}} Q(x^t + \alpha r^t). \tag{6.2.33}$$

Wegen

$$\frac{d}{d\alpha}Q(x^t + \alpha r^t) = \operatorname{grad}Q(x^t + \alpha r^t)\cdot r^t = (Ax^t - b, r^t) + \alpha(Ar^t, r^t)$$

ist notwendig

$$\alpha_t = -\frac{(g^t, r^t)}{(Ar^t, r^t)}\,, \quad g^t := Ax^t - b = \operatorname{gradQ}(\mathrm{x^t})\,.$$

Definition 6.3: *Das allgemeine Abstiegsverfahren bestimmt ausgehend von einem Startwert $x^0 \in \mathbb{R}^n$ eine Folge von Iterierten $x^t \in \mathbb{R}^n$, $t = 1, 2, \ldots$ nach der Vorschrift:*

$$\begin{aligned} &\textit{Gradient } g^t = Ax^t - b, \quad \textit{Abstiegsrichtung } r^t, \\ &\alpha_t = -\frac{(g^t, r^t)}{(Ar^t, r^t)}, \quad x^{t+1} = x^t + \alpha_t r^t. \end{aligned}$$

Praktisch günstiger ist die folgende Schreibweise, bei der man eine Matrix-Vektor-Multiplikation spart, wenn man den Vektor Ar^t abspeichert:

$$\begin{aligned} &\text{Startwerte:} && x^0 \in \mathbb{R}^n, \quad g^0 := Ax^0 - b \\ &\text{für } t \geq 0: && g^t = Ax^t - b, \quad \text{Abstiegsrichtung} \quad r^t \\ & && \alpha_t = -\tfrac{(g^t, r^t)}{(Ar^t, r^t)}, \quad x^{t+1} = x^t + \alpha_t r^t, \quad g^{t+1} = g^t + \alpha_t Ar^t. \end{aligned}$$

Unter Verwendung der Notation $\|y\|_B := (By, y)^{1/2}$ gilt

$$2Q(y) = \|Ay - b\|^2_{A^{-1}} - \|b\|^2_{A^{-1}} = \|y - x\|^2_A - \|x\|^2_A, \tag{6.2.34}$$

d. h.: Die Minimierung des Funktionals $Q(\cdot)$ ist äquivalent zur Minimierung der Defektnorm $\|Ay - b\|_{A^{-1}}$ bzw. der Fehlernorm $\|y - x\|_A$.

6.2.1 Gradienten-Verfahren

Die verschiedenen Abstiegsverfahren unterscheiden sich im Wesentlichen durch die jeweilige Wahl der Abstiegsrichtungen r^t. Die einfachste Möglichkeit wäre, die Richtungen r^t zyklisch die kartesischen Einheitsvektoren $\{e^1, \ldots, e^n\}$ durchlaufen zu lassen. Die so erhaltene iterative Methode wird „Koordinatenrelaxation“ genannt; sie ist im Wesenlichen äquivalent zum Gauß-Seidel-Verfahren (Übungsaufgabe). Naheliegender ist die Wahl der Richtung des stärksten Abfalls von $Q(\cdot)$ im Punkt x^t als Suchrichtung:

$$r^t = -\mathrm{gradQ(x^t)} = -\mathrm{g^t}. \tag{6.2.35}$$

Definition 6.4: *Das „Gradientenverfahren“ bestimmt eine Folge von Iterierten $x^t \in \mathbb{R}^n$ gemäß der Vorschrift:*

$$\begin{array}{ll} \textit{Startwerte:} & x^0 \in \mathbb{R}^n\,, \quad g^0 := Ax^0 - b \\ \textit{für } t \geq 0\textit{:} & \alpha_t = \frac{\|g^t\|^2}{(Ag^t, g^t)} \\ & x^{t+1} = x^t - \alpha_t g^t, \quad g^{t+1} = g^t - \alpha_t A g^t. \end{array}$$

Im Falle $(Ag^t, g^t) = 0$ ist notwendig auch $g^t = 0$, d. h.: Die Iteration kann nur mit $Ax^t = b$ abbrechen, d. h.: x^t ist dann Lösung des Gleichungssystems.

Satz 6.7 (Gradientenverfahren): *Ist die Matrix $A \in \mathbb{R}^{n\times n}$ positiv definit, so konvergiert das Gradientenverfahren für jeden Startvektor $x^0 \in \mathbb{R}^n$ gegen die Lösung des Gleichungssystems $Ax = b$.*

Beweis: Wir führen das sog. „Fehlerfunktional“ ein

$$E(y) \;:=\; \|y - x\|^2_A = (y - x, A[y - x])\,, \quad y \in \mathbb{R}^n,$$

und setzen zur Abkürzung $e^t := x^t - x$. Mit diesen Bezeichnungen gilt dann

$$\begin{aligned} \frac{E(x^t) - E(x^{t+1})}{E(x^t)} &= \frac{(e^t, Ae^t) - (e^{t+1}, Ae^{t+1})}{(e^t, Ae^t)} \\ &= \frac{(e^t, Ae^t) - (e^t - \alpha_t g^t, A[e^t - \alpha_t g^t])}{(e^t, Ae^t)} \\ &= \frac{2\alpha_t (e^t, Ag^t) - \alpha_t^2 (g^t, Ag^t)}{(e^t, Ae^t)} \end{aligned}$$

und folglich, wegen $Ae^t = Ax^t - Ax = Ax^t - b = g^t$,

$$\frac{E(x^t) - E(x^{t+1})}{E(x^t)} = \frac{2\alpha_t \|g^t\|^2 - \alpha_t^2 (g^t, Ag^t)}{(g^t, A^{-1}g^t)} = \frac{\|g^t\|^4}{(g^t, Ag^t)(g^t, A^{-1}g^t)}.$$

Für die positiv definite Matrix A gilt

$$\lambda \|y\|^2 \le (y, Ay) \le \Lambda \|y\|^2, \quad \Lambda^{-1}\|y\|^2 \le (y, A^{-1}y) \le \lambda^{-1}\|y\|^2,$$

mit $\lambda = \lambda_{\min}(A)$ und $\Lambda = \lambda_{\max}(A)$. Im Falle $x^t \neq x$, d. h. $E(x^t) \neq 0$ und $g^t \neq 0$, erschließt man damit

$$\frac{\|g^t\|^4}{(g^t, Ag^t)(g^t, A^{-1}g^t)} \ge \frac{\|g^t\|^4}{\Lambda\|g^t\|^2\, \lambda^{-1}\|g^t\|^2} = \frac{\lambda}{\Lambda},$$

bzw.

$$E(x^{t+1}) \le \{\, 1 - \kappa^{-1} \,\} E(x^t), \quad \kappa := \mathrm{cond}_{\mathrm{nat}}(\mathrm{A}).$$

Wegen $0 < 1 - 1/\kappa < 1$ konvergiert somit für jedes $x^0 \in \mathbb{R}^n$ das Fehlerfunktional $E(x^t) \to 0$ $(t \to \infty)$, d. h.: $x^t \to x$ $(t \to \infty)$. Q.E.D.

Für eine verschärfte Abschätzung der Konvergenzgeschwindigkeit des Gradientenverfahrens benötigen wir das folgende Resultat von Kantorowitsch:

Hilfssatz 6.4 (Lemma von Kantorowitsch): *Für die positiv definite Matrix $A \in \mathbb{R}^n$ mit kleinstem Eigenwert λ und größtem Eigenwert Λ gilt*

$$4\,\frac{\lambda\Lambda}{(\lambda - \Lambda)^2} \le \frac{\|y\|^4}{(y, Ay)(y, A^{-1}y)}, \quad \forall\, y \in \mathbb{R}^n. \tag{6.2.36}$$

Beweis: Seien $\lambda = \lambda_1 \le \ldots \le \lambda_n = \Lambda$ die Eigenwerte von A und $\{w_1, \ldots, w_n\}$ eine zugehörige Orthonormalbasis von Eigenvektoren. Ein beliebiger Vektor $y \in \mathbb{R}^n$ gestattet die Entwicklung $y = \sum_{i=1}^n y_i w_i$ mit den Koeffizienten $y_i = (y, w_i)$. Dann gilt

$$\frac{\|y\|^4}{(y, Ay)(y, A^{-1}y)} = \frac{(\sum_{i=1}^n y_i^2)^2}{(\sum_{i=1}^n \lambda_i y_i^2)\,(\sum_{i=1}^n \lambda_i^{-1} y_i^2)} = \frac{1}{(\sum_{i=1}^n \lambda_i \zeta_i)\,(\sum_{i=1}^n \lambda_i^{-1}\zeta_i)} = \frac{\varphi(\zeta)}{\psi(\zeta)}$$

mit den Bezeichnungen

$$\zeta = (\zeta_i)_{i=1,\ldots,n}, \quad \zeta_i = y_i^2 (\sum_{i=1}^n y_i^2)^{-1},$$

$$\psi(\zeta) = \sum_{i=1}^n \lambda_i^{-1}\zeta_i, \quad \varphi(\zeta) = (\sum_{i=1}^n \lambda_i \zeta_i)^{-1}.$$

Da die Funktion $f(\lambda) = \lambda^{-1}$ konvex ist, folgt aus $0 \le \zeta_i \le 1$ und $\sum_{i=1}^n \zeta_i = 1$, dass gilt:

$$\sum_{i=1}^{n} \lambda_i^{-1}\zeta_i \ge (\sum_{i=1}^{n} \lambda_i\zeta_i)^{-1}.$$

Wir setzen $g(\lambda) := (\lambda_1 + \lambda_n - \lambda)/(\lambda_1\lambda_n)$.

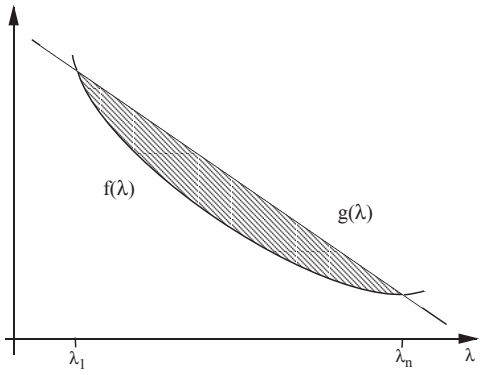


Abbildung 6.2: *Skizze zu Beweis des Lemma von Kantorowitsch*

Offenbar liegt $\varphi(\zeta)$ für alle Argumente ζ stets auf der Kurve $f(\lambda)$, und $\psi(\zeta)$ liegt stets zwischen den Kurven $f(\lambda)$ und $g(\lambda)$ (schraffierter Bereich). Folglich gilt

$$\frac{\varphi(\zeta)}{\psi(\zeta)} \ge \min_{\lambda_1 \le \lambda \le \lambda_n} \frac{f(\lambda)}{g(\lambda)} = \frac{f([\lambda_1 + \lambda_n]/2)}{g([\lambda_1 + \lambda_n]/2)} = \frac{4\lambda_1\lambda_n}{(\lambda_1 + \lambda_n)^2}.$$

Q.E.D.

Satz 6.8 (Fehlerabschätzung): *Für das Gradientenverfahren gilt die Fehlerabschätzung*

$$\|x^t - z\|_A \le \Big(\frac{1 - 1/\kappa}{1 + 1/\kappa}\Big)^t \|x^0 - z\|_A\,, \quad t \in \mathbb{N}, \tag{6.2.37}$$

mit der Spektralkonditionszahl $\kappa = \mathrm{cond}_2(A) = \Lambda/\lambda$ *von* A.

Beweis: Im Beweis von Satz 6.7 wurde die folgende Identität gezeigt:

$$E(x^{t+1}) = \Big\{\, 1 - \frac{\|g^t\|^4}{(g^t, Ag^t)\,(g^t, A^{-1}g^t)} \,\Big\}\, E(x^t).$$

Diese ergibt mit der Ungleichung von Kantorowitsch

$$E(x^{t+1}) \le \Big\{\, 1 - 4\frac{\lambda\Lambda}{(\lambda + \Lambda)^2} \,\Big\}\, E(x^t) = \Big(\frac{\lambda - \Lambda}{\lambda + \Lambda}\Big)^2 E(x^t).$$

Daraus folgt dann durch sukzessive Anwendung

$$\|x^t - x\|_A^2 \leq \Big(\frac{\lambda - \Lambda}{\lambda + \lambda}\Big)^{2t} \|x^0 - x\|_A^2\,, \quad t \in \mathbb{N}.$$

Q.E.D.

Der Beziehung

$$\begin{aligned}(g^{t+1}, g^t) &= (g^{(t)} - \alpha_t A g^t, g^t) \\ &= \|g^t\|^2 - \alpha_t (A g^t, g^t) = 0.\end{aligned} \tag{6.2.38}$$

entnehmen wir, dass die Abstiegsrichtungen $r^t = -g^t$ des Gradientenverfahrens in jeweils direkt aufeinanderfolgenden Schritten orthogonal sind. Dagegen kann g^{t+2} weit von Orthogonalität zu g^t abweichen. Dies führt zu einem stark oszillatorischen Konvergenzverhalten des Gradientenverfahrens besonders bei Matrizen A mit weit auseinander liegenden Eigenwerten. Dies bedeutet etwa in Fall $n = 2$, dass das Funktional $Q(\cdot)$ stark exzentrische Niveaulinien hat, und sich die Iterierten in einem Zickzackkurs der Lösung annähern (siehe Abb. 6.2.1).

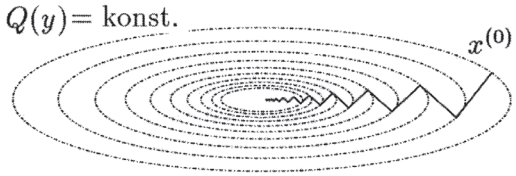


Abbildung 6.3: *Oszillierende Konvergenz des Gradientenverfahrens*

6.2.2 CG-Verfahren

Das Gradientenverfahren nutzt die Struktur des quadratischen Funktionals $Q(\cdot)$, d. h. die Verteilung der Eigenwerte der Matrix A, nur lokal von einem Schritt zum nächsten aus. Es wäre günstiger, wenn bei der Wahl der Abstiegsrichtungen auch die bereits gewonnenen Informationen über die globale Struktur von $Q(\cdot)$ berücksichtigt würde, d. h. wenn etwa die Abstiegsrichtungen paarweise orthogonal wären. Dies ist die Grundidee des „Verfahrens der konjugierten Richtungen“ („conjugate gradient method“; kurz „CG-Verfahren“) nach Hestenes[2] und Stiefel[3] (1992), welches sukzessive eine Folge von Abstiegsrichtungen $d^{(t)}$ erzeugt, die bzgl. des Skalarprodukts $(\cdot,\cdot)_A$ orthogonal sind („A-orthogonal“).

[2]Magnus R. Hestenes (1906–1991): US-Amerikanischer Mathematiker; arbeitete am National Bureau of Standards (NBS)und an der University of California at Los Angeles (UCLA); Beiträge zur Optimierung und Kontrolltheorie und zur Numerischen Linearen Algebra.

[3]Eduard Stiefel (1909–1978): Schweizer Mathematiker; seit 1943 Professor für Angewandte Mathematik an der ETH Zürich; wichtige Beiträge zu Topologie, Gruppentheorie, Numerische Lineare Algebra (CG-verfahren) und Approximationstheorie sowie zur Himmelsmechanik.

Zur Herleitung des CG-Verfahrens verwenden wir den Ansatz

$$B_t := \mathrm{span}\{d^0, \cdots, d^{t-1}\} \tag{6.2.39}$$

mit einem noch zu bestimmenden linear unabhängigen Satz von Vektoren d^i und suchen die Iterierten in der Form

$$x^t = x^0 + \sum_{i=0}^{t-1} \alpha_i d^i \ \in \ x^0 + B_t \tag{6.2.40}$$

so zu bestimmen, dass

$$Q(x^t) = \min_{y \in x^0 + B_t} Q(y) \quad \leftrightarrow \quad \|Ax^t - b\|_{A^{-1}} = \min_{y \in x^0 + B_t} \|Ay - b\|_{A^{-1}}. \tag{6.2.41}$$

Durch Nullsetzten der Ableitungen von $Q(\cdot)$ nach den α_i sehen wir, dass dies äquivalent ist zu den sog. „Galerkin[4]-Gleichungen“:

$$(Ax^t - b, d^i) = 0\,, \quad i = 0, \ldots, t-1. \tag{6.2.42}$$

bzw. in Kurzschreibweise $Ax^t - b = g^t \perp B_t$.

Bemerkung 6.1: Es sei bemerkt, dass (6.2.42) nicht von der Symmetrie der Matrix A abhängt. Ausgehend von dieser Beziehung lassen sich auch CG-artige Verfahren für unsymmetrische und sogar indefinite Gleichungssysteme ableiten. Diese werden allgemein „Projektionsverfahren“ genannt.

Setzt man den obigen Ansatz für x^t in die Orthogonalitätsbedingung (6.2.42) ein, so erhält man ein reguläres Gleichungssystem für die Koeffizienten $\alpha_i\,(i = 0, \ldots, t-1)$. Es sei nochmals daran erinnert, dass die Galerkin-Gleichungen (6.2.42) äquivalent sind zur Minimierung der Defektnorm $\|Ax^t - b\|_{A^{-1}}$ oder der Fehlernorm $\|x^t - x\|_A$ über $x^0 + B_t$. Eine natürliche Wahl für B_t sind die sog. Krylow[5]-Räume

$$B_t = K_t(d^0; A) := \mathrm{span}\{d^0, Ad^0, \ldots, A^{t-1}d^0\}, \tag{6.2.43}$$

mit einem Vektor d^0, etwa $d^0 := b - Ax^0$ zu irgend einem Startvektor x^0. Dies ist motiviert durch die Beobachtung, dass aus $A^t d^0 \in K_t(d^0; A)$ notwendig $-g^t = b - Ax^t = d^0 + A(x^0 - x^t) \in d^0 + AK_t(d^0; A) \in K_t(d^0; A)$ folgt. Wegen $g^t \perp K_t(d^0; A)$ impliziert dies dann $g^t = 0$ gemäß Konstruktion.

Das CG-Verfahren erzeugt nun Abstiegsrichtungen, die eine A-orthogonale Basis des Krylow-Raumes $K_t(d^0; A)$ bilden. Dazu wird induktiv vorgegangen: Ausgehend von ei-

[4]Boris Grigorievich Galerkin (1871–1945): Russischer Bauingenieur und Mathematiker; Professor in St. Petersburg; Beiträge zur Struktur-Mechanik, insbesondere zur Plattentheorie.

[5]Aleksei Nikolaevich Krylov (1863–1945): Russischer Mathematiker; Professor an der Sov. Akademie der Wissensch. in St. Petersburg; Beiträge zu Fourier-Analysis und Differentialgleichungen, Anwendungen in der Schiffstechnik.

nem Startpunkt x^0 mit „Residuum“ (negative „Defekt“) $d^0 = b - Ax^0$ seien Iterierte x^i und zugehörige Abstiegsrichtungen $d^i (i = 0, ..., t-1)$ bereits bestimmt, so dass $\{d^0, ..., d^{t-1}\}$ eine A-orthogonale Basis von $K_t(d^0; A)$ ist. Zur Konstruktion des nächsten $d^t \in K_{t+1}(d^0; A)$ mit der Eigenschaft $d^t \perp_A K_t(d^0; A)$ machen wir den Ansatz

$$d^t = -g^t + \sum_{j=0}^{t-1} \beta_j^{t-1} d^j \ \in \ K_{t+1}(d^0; A) \tag{6.2.44}$$

Dabei wird o.B.d.A. angenommen, dass $g^t = Ax^t - b \notin K_t(d^0; A)$ ist, da andernfalls $g^t = 0$ bzw. $x^t = x$ wäre. Dann gilt für $i = 0, ..., t-1$:

$$(d^t, Ad^i) = (-g^t, Ad^i) + \sum_{j=0}^{t-1} \beta_j^{t-1} (d^j, Ad^i) = (-g^t + \beta_i^{t-1} d^i, Ad^i) . \tag{6.2.45}$$

Für $i < t-1$ ist $(g^t, Ad^i) = 0$ wegen $Ad^i \in K_t(d^0; A)$ und demnach $\beta_i^{t-1} = 0$. Für $i = t-1$ führt die Bedingung

$$0 = (-g^t, Ad^{t-1}) + \beta_{t-1}^{t-1} (d^{t-1}, Ad^{t-1}) \tag{6.2.46}$$

zu den Formeln

$$\beta_{t-1} := \beta_{t-1}^{t-1} = \frac{(g^t, Ad^{t-1})}{(d^{t-1}, Ad^{t-1})} , \quad d^t = -g^t + \beta_{t-1} d^{t-1}. \tag{6.2.47}$$

Die nächsten Iterierten x^{t+1} und $g^{t+1} = Ax^{t+1} - b$ sind dann bestimmt durch

$$\alpha_t = -\frac{(g^t, d^t)}{(d^t, Ad^t)} , \quad x^{t+1} = x^t + \alpha_t d^t , \quad g^{t+1} = g^t + \alpha_t Ad^t . \tag{6.2.48}$$

Dies sind die Rekursionsformeln des klassischen CG-Verfahrens. Nach Konstruktion gilt

$$(d^t, Ad^i) = (g^t, d^i) = 0, \quad i \leq t-1 , \quad (g^t, g^{t-1}) = 0. \tag{6.2.49}$$

Damit folgern wir, dass

$$\|g^t\|^2 = (d^t - \beta_{t-1} d^{t-1}, -g^{t+1} + \alpha_t Ad^t) = \alpha_t (d^t, Ad^t), \tag{6.2.50}$$

$$\|g^{t+1}\|^2 = (g^t + \alpha_t Ad^t, g^{t+1}) = \alpha_t (Ad^t, g^{t+1}). \tag{6.2.51}$$

Dies gestattet einige Vereinfachungen in den Formeln, nämlich

$$\alpha_t = \frac{\|g^t\|^2}{(d^t, Ad^t)}, \quad \beta_t = \frac{\|g^{t+1}\|^2}{\|g^t\|^2}, \tag{6.2.52}$$

solange die Iteration nicht mit $g^t = 0$ abbricht.

Definition 6.5: *Das CG-Verfahren bestimmt eine Folge von Iterierten $x^t \in \mathbb{R}^n$ gemäß der Vorschrift:*

$$\begin{array}{ll} \textit{Startwerte:} & x^0 \in \mathbb{R}^n\,, \quad d^0 = -g^0 = b - Ax^0, \\ \textit{für } t \geq 0\textit{:} & \alpha_t = \frac{\|g^t\|^2}{(d^t, Ad^t)}, \quad x^{t+1} = x^t + \alpha_t d^t, \quad g^{t+1} = g^{(t)} + \alpha_t A d^t, \\ & \beta_t = \frac{\|g^{(t+1)}\|^2}{\|g^{(t)}\|^2}, \quad d^{t+1} = -g^{t+1} + \beta_t d^t. \end{array}$$

Nach Konstruktion erzeugt das CG-Verfahren eine Folge von Abstiegsrichtungen d^t, welche automatisch paarweise A-orthogonal sind. Dies impliziert, dass die Vektoren $d^0, \ldots, d^t$ jeweils linear unabhängig sind, und dass folglich gilt:

$$\operatorname{span}\{d^0, \ldots, d^{n-1}\} = \mathbb{R}^n. \tag{6.2.53}$$

Wir fassen die bisher abgeleiteten Eigenschaften des CG-Verfahrens in dem folgendem Satz zusammen:

Satz 6.9 (CG-Verfahren): *Das CG-Verfahren bricht für jeden Startvektor $x^0 \in \mathbb{R}^n$ (bei rundungsfreier Rechnung) nach spätestens n Schritten mit $x^{t_n} = x$ ab. Dabei gilt in jedem Schritt*

$$Q(x^t) = \min_{\alpha \in \mathbb{R}} Q(x^{t-1} + \alpha d^{t-1}) = \min_{y \in x^0 + K_t} Q(y) \tag{6.2.54}$$

mit $K_t := \operatorname{span}\{d^0, \ldots, d^{t-1}\}$.

Das CG-Verfagren gehört also im Gegensatz zum Gradientenverfahren eigentlich zur Klasse der „direkten“ Verfahren. In der Praxis wird sie jedoch wie ein iteratives Verfahren angewendet, da

1. aufgrund von Rundungsfehlern die Richtungen d^t nicht wirklich A-orthogonal sind, und die Iteration nicht abbricht;
2. bei großen Matrizen auch mit deutlich weniger als n Iterationsschritten schon brauchbare Näherungen erzielbar sind.

Zur Vorbereitung des Hauptsatzes über die Konvergenzgeschwindigkeit des CG-Verfahrens beweisen wir zunächst den folgenden Hilfssatz.

Hilfssatz 6.5 (Polynomiale Normschranke): *Für ein Polynom $p \in P_t$, $p(0) = 1$, gelte auf einer Menge $S \subset \mathbb{R}$, welche alle Eigenwerte von A enthält,*

$$\sup_{\mu \in S} |p(\mu)| \leq M. \tag{6.2.55}$$

Dann gilt

$$\|x^t - x\|_A \leq M \|x^0 - x\|_A. \tag{6.2.56}$$

Beweis: Unter Beachtung der Beziehung

$$\|x^t - x\|_A^{\|} = \min_{y \in x^0 + K_t} \|y - x\|_A,$$

$$K_t = \operatorname{span}\{d^0, \ldots, d^{t-1}\} = \operatorname{span}\{A^0 g^{(0)}, \ldots, A^{t-1} g^0\}$$

finden wir

$$\|x^t - x\|_A = \min_{p \in P_{t-1}} \|x^0 - x + p(A) g^0\|_A.$$

Wegen $g^0 = Ax^0 - b = A(x^0 - x)$ folgt weiter

$$\begin{aligned} \|x^t - x\|_A &= \min_{p \in P_{t-1}} \|[I + Ap(A)](x^0 - x)\|_A \\ &\leq \min_{p \in P_{t-1}} \|I + Ap(A)\|_A \, \|x^0 - x\|_A \leq \min_{p \in P_t,\, p(0)=1} \|p(A)\|_A \, \|x^0 - x\|_A, \end{aligned}$$

mit der von $\|\cdot\|_A$ erzeugten natürlichen Matrizennorm $\|\cdot\|_A$. Für beliebiges $y \in \mathbb{R}^n$ gilt mit einer Orthonormalbasis $\{w_1, \ldots, w_n\}$ aus Eigenvektoren von A:

$$y = \sum_{i=1}^{n} \gamma_i w_i\,, \quad \gamma_i = (y, w_i),$$

und folglich

$$\|p(A)y\|_A^2 = \sum_{i=1}^{n} \lambda_i p(\lambda_i)^2 \gamma_i^2 \leq M^2 \sum_{i=1}^{n} \lambda_i \gamma_i^2 = M^2 \|y\|_A^2.$$

Dies impliziert

$$\|p(A)\|_A = \sup_{y \in \mathbb{R}^n,\, y \neq 0} \frac{\|p(A)y\|_A}{\|y\|_A} \leq M$$

und damit die Behauptung. Q.E.D.

Satz 6.10 (CG-Konvergenz): *Für das CG-Verfahren gilt die Fehlerabschätzung*

$$\|x^t - x\|_A \leq 2 \left(\frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}} \right)^t \|x^0 - x\|_A\,, \quad t \in \mathbb{N}\,, \tag{6.2.57}$$

mit der Spektralkonditionszahl $\kappa = \operatorname{cond}_2(A) = \Lambda/\lambda$ *von* A*. Zur Reduzierung des Anfangsfehlers um den Faktor* ε *sind höchstens*

$$t(\varepsilon) \leq \tfrac{1}{2} \sqrt{\kappa}\, \ln(2/\varepsilon) + 1 \tag{6.2.58}$$

Iterationsschritte erforderlich.

Beweis: Setzt man $S := [\lambda, \Lambda]$ in Hilfssatz 6.5, so folgt

$$\|x^t - x\|_A \leq \min_{p \in P_t,\, p(0)=1} \Big\{ \sup_{\lambda \leq \mu \leq \Lambda} |p(\mu)| \Big\} \|x^0 - x\|_A\,.$$

Dies ergibt die Behauptung, wenn wir zeigen können, dass

$$\min_{p\in P_t,\, p(0)=1} \Big\{ \sup_{\lambda\le\mu\le\Lambda} |p(\mu)| \Big\} \le 2\Big(\frac{1-\sqrt{\lambda/\Lambda}}{1+\sqrt{\lambda/\Lambda}}\Big)^t.$$

Dabei handelt es sich um ein Problem der Bestapproximation mit Polynomen bzgl. der Maximumnorm (Tschebyscheff-Approximation). Die Lösung $\bar p$ ist gegeben durch

$$\bar p(\mu) = T_t\Big(\frac{\Lambda+\lambda-2\mu}{\Lambda-\lambda}\Big)\; T_t\Big(\frac{\Lambda+\lambda}{\Lambda-\lambda}\Big)^{-1},$$

mit dem t-ten Tschebyscheff-Polynom T_t auf $[-1,1]$. Dabei ist

$$\sup_{\lambda\le\mu\le\Lambda} \bar p(\mu) = T_t\Big(\frac{\Lambda+\lambda}{\Lambda-\lambda}\Big)^{-1}.$$

Aus der Darstellung

$$T_t(\mu) = \tfrac12\left[\big(\mu+\sqrt{\mu^2-1}\big)^t + \big(\mu-\sqrt{\mu^2-1}\big)^t\right], \quad \mu\in[-1,1],$$

für die Tschebyscheff-Polynome folgt über die Identität

$$\frac{\kappa+1}{\kappa-1} + \sqrt{\Big(\frac{\kappa+1}{\kappa-1}\Big)^2-1} = \frac{\kappa+1}{\kappa-1} + \frac{2\sqrt\kappa}{\kappa-1} = \frac{(\sqrt\kappa+1)^2}{\kappa-1} = \frac{\sqrt\kappa+1}{\sqrt\kappa-1}$$

die Abschätzung nach unten

$$T_t\Big(\frac{\Lambda+\lambda}{\Lambda-\lambda}\Big) = T_t\Big(\frac{\kappa+1}{\kappa-1}\Big) = \frac12\left[\Big(\frac{\sqrt\kappa+1}{\sqrt\kappa-1}\Big)^t + \Big(\frac{\sqrt\kappa-1}{\sqrt\kappa+1}\Big)^t\right] \ge \frac12\Big(\frac{\sqrt\kappa+1}{\sqrt\kappa-1}\Big)^t.$$

Also wird

$$\sup_{\lambda\le\mu\le\Lambda} \bar p(\mu) \le 2\Big(\frac{\sqrt\kappa-1}{\sqrt\kappa+1}\Big)^t,$$

was (6.2.57) impliziert. Zur Herleitung von (6.2.58) fordern wir

$$2\Big(\frac{\sqrt\kappa-1}{\sqrt\kappa+1}\Big)^{t(\varepsilon)} < \varepsilon$$

bzw.

$$t(\varepsilon) > \ln\Big(\frac{2}{\varepsilon}\Big)\ln\Big(\frac{\sqrt\kappa+1}{\sqrt\kappa-1}\Big)^{-1}.$$

Wegen

$$\ln\frac{x+1}{x-1} = 2\left\{\frac1x + \frac13\frac{1}{x^3} + \frac15\frac{1}{x^5} + \dots\right\} \ge \frac2x$$

ist dies erfüllt für $t(\varepsilon) \ge \frac12\sqrt\kappa\,\ln(2/\varepsilon)$. Q.E.D.

Wegen $\kappa = \mathrm{cond}_{\mathrm{nat}}(\mathrm{A}) > 1$ ist $\sqrt\kappa < \kappa$. Da die Funktion $f(\lambda) = (1-\lambda^{-1})\,(1+\lambda^{-1})^{-1}$ für $\lambda > 0$ streng monoton wachsend ist $(f'(\lambda) > 0)$, folgt:

$$\frac{1-1/\sqrt{\kappa}}{1+1/\sqrt{\kappa}} < \frac{1-1/\kappa}{1+1/\kappa},$$

d. h.: Die Methode der konjugierten Richtungen sollte schneller konvergieren als das Gradientenverfahren. Dies ist auch praktisch der Fall. Beide Verfahren konvergieren offenbar umso schneller, je näher die Kondition $\mathrm{cond}_{\mathrm{nat}}(\mathrm{A})$ bei 1 liegt. Ist aber $\Lambda \gg \lambda$, was in der Praxis leider häufig der Fall ist, konvergiert auch die Methode der konjugierten Richtungen nur sehr langsam. Eine Beschleunigung der Konvergenz kann durch sog. „Vorkonditionierung“ erreicht werden, die wir im Folgenden beschreiben werden.

6.2.3 Allgemeinere CG-Verfahren und Vorkonditionierung

Unsymmetrische Systeme

Zur Lösung allgemeiner Gleichungssysteme $Ax = b$ mit einer regulären, aber nicht notwendig positiv definiten Matrix $A \in \mathbb{R}^n$ mit Hilfe des CG-Verfahrens kann man etwa zu dem äquivalenten System

$$A^T A x = A^T b \tag{6.2.59}$$

mit der positiv definiten Matrix $A^T A$ übergehen. Hierauf angewendet, schreibt sich das CG-Verfahren wie folgt:

$$\begin{array}{ll} \text{Startwerte:} & x^0 \in \mathbb{R}^n\,, \quad d^0 = A^T(b - Ax^0) \\ \text{für } t \geq 0\text{:} & \alpha_t = \frac{\|g^t\|^2}{\|Ad^t\|^2}\,, \quad (g^t = A^T Ax - A^T b) \\ & x^{t+1} = x^t + \alpha_t d^t, \quad g^{t+1} = g^t + \alpha_t A^T A d^t \\ & \beta_t = \frac{\|g^{t+1}\|^2}{\|g^t\|^2}, \quad d^{t+1} = -g^{t+1} + \beta_t d^t\,. \end{array}$$

Die Konvergenzgeschwindigkeit ist dabei charakterisiert durch $\kappa(A^T A)$. Das ganze Verfahren beruht offenbar auf der Minimierung des Funktionals

$$Q(y) := \tfrac{1}{2}\,(A^T A y, y) - (A^T b, y) = \|Ay - b\|^2 - \tfrac{1}{2}\,\|b\|^2, \tag{6.2.60}$$

d. h. der Minimierung der Residuumsnorm $\|Ay - b\|$, daher der Name „MINRES-Verfahren”. Da $\kappa(A^T A) \sim \kappa(A)^2$ ist, muss man mit einer recht langsamen Konvergenz des CG-Verfahrens für nicht symmetrische Systeme rechnen. Andere CG-artige Verfahren for unsymmetrische Matrizen mit i. Allg. besseren Konvergenzeigenschaften sind das „GMRES-“ und das „BiCGstab-Verfahren“ (s. die Literatur über Numerische Lineare Algebra).

Vorkonditionierung, PCG-Verfahren

Offensichtlich funktioniert das CG-Verfahren um so besser, je näher die Kondition der Matrix A bei Eins liegt:

$$\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \geq 1. \tag{6.2.61}$$

Daher wird eine „Vorkonditionierung“ vorgenommen, d. h.: Das System $Ax = b$ wird in ein äquivalentes umgeformt, $\tilde{A}\tilde{x} = \tilde{b}$, dessen Matrix $\tilde{A}$ besser konditioniert ist. Sei C eine symmetrische, positiv definite Matrix, welche sich als Produkt darstellen lässt,

$$C = KK^T \tag{6.2.62}$$

mit einer regulären Matrix K. Das System $Ax = b$ wird dann in der folgenden äquivalenten Form geschrieben:

$$\underbrace{K^{-1}A\,(K^T)^{-1}}_{\tilde{A}}\ \underbrace{K^T x}_{\tilde{x}} = \underbrace{K^{-1}b}_{\tilde{b}}. \tag{6.2.63}$$

Das CG-Verfahren wird nun auf das System $\tilde{A}\tilde{x} = \tilde{b}$ angewendet. Die Idee dabei ist, die Matrix C so zu wählen, dass $\kappa(\tilde{A}) \ll \kappa(A)$ wird. Die Beziehung

$$(K^T)^{-1}\tilde{A}K^T = (K^T)^{-1}K^{-1}A(K^T)^{-1}K^T = C^{-1}A \tag{6.2.64}$$

zeigt, dass bei der Wahl $C = uivA$ die Matrix $\tilde{A}$ ähnlich zu I, d. h. $\kappa(\tilde{A}) = \kappa(I) = I$ wäre. Folglich wird man C als möglichst gute Approximation von A wählen, wobei natürlich die Zerlegung $C = KK^T$ mit möglichst einfach zu invertierendem Faktor K bekannt sein muss. Das CG-Verfahren sieht dann wie folgt aus:

Startwerte:	$\tilde{x}^{(0)} \in \mathbb{R}^n\,, \quad \tilde{d}^0 = -\tilde{g}^0 = \tilde{b} - \tilde{A}\tilde{x}^0$
für $t \geq 0$:	$\alpha_t = \frac{\|\tilde{g}^t\|^2}{(\tilde{d}^t, \tilde{A}\tilde{d}^t)}$
	$\tilde{x}^{t+1} = \tilde{x}^t + \alpha_t\tilde{d}^t, \quad \tilde{g}^{t+1} = \tilde{g}^t + \alpha_t\tilde{A}\tilde{d}^t$
	$\beta_t = \frac{\|\tilde{g}^{t+1}\|^2}{\|\tilde{g}^t\|^2}, \quad \tilde{d}^{t+1} = -\tilde{g}^{t+1} + \beta_t\tilde{d}^t.$

Diesen Algorithmus schreibt man üblicherweise bezogen auf die ursprüngliche Matrix A und erhält so das sog. „PCG-Verfahren“ („Preconditioned PC method“).

Definition 6.6: *Das PCG-Verfahren mit (regulärer) Vorkonditionierungsmatrix $C = KK^T$ bestimmt eine Folge von Iterierten $x^t \in \mathbb{R}^n$ gemäß der Vorschrift:*

Startwerte:	$x^0 \in \mathbb{R}^n\,, \quad g^0 = Ax^0 - b\,, \quad C\rho^0 = g^0\,, \quad d^0 = -\rho^0$
für $t \geq 0$:	$\alpha_t = \frac{(g^t, \rho^t)}{(d^t, Ad^t)}, \quad x^{t+1} = x^t + \alpha_t d^t, \quad g^{t+1} = g^t + \alpha_t Ad^{(t)}$
	$C\rho^{t+1} = g^{t+1}$
	$\beta_t = \frac{(g^{t+1}, \rho^{t+1})}{(g^t, \rho^t)}, \quad d^{t+1} = -\rho^{t+1} + \beta_t d^t.$

Beim PCG-Verfahren ist in jedem Iterationsschritt ist also ein lineares Gleichungssystem mit der Koeffizientenmatrix $C = KK^T$ zu lösen. Dies bedingt, dass K etwa eine Dreiecksmatrix oder ähnliches sein sollte, so dass die Lösung von $C\rho^t = g^t$ durch einfaches Vorwärts- und Rückwärtseinsetzen erfolgen kann.

Beispiel 6.2: Wir listen einige einfache in der Praxis gebräuchliche Vorkonditionierungen für das CG-Verfahren.

a) Skalierung: Mit der üblichen Zerlegung $A = D + L + R,\ R = L^T$ wird gesetzt:

$$C = D\,,\ K = D^{1/2} \quad \text{Skalierungsmatrix}$$
$$\tilde{A} = D^{-1/2} A D^{-1/2} \Rightarrow \tilde{a}_{ii} = 1 \quad (1 \le i \le n).$$

Die Skalierung bewirkt, dass die Elemente von A auf etwa gleiche Größenordnung gebracht werden. Dies reduziert die Kondition, denn es gilt folgender Satz: *Der kleinste (größte) Eigenwert einer symmetrischen, positiv definiten Matrix ist höchstens (mindestens) so groß wie das kleinste (größte) Diagonalelement, und die Kondition der Matrix ist mindestens so groß wie der Quotient aus dem größten und dem kleinsten Diagonalelement.*

b) SSOR-Vorkonditionierung: Mit einem Parameter ω wird gesetzt

$$C = (D + \omega L) D^{-1} (D + \omega R) = (\underbrace{D^{1/2} + \omega L D^{-1/2}}_{K})(\underbrace{D^{1/2} + \omega D^{-1/2} R}_{K^T}).$$

Offenbar besitzt die Dreiecksmatrix K dieselbe schwache Besetzung wie A. Pro Iterationsschritt erfordert das so vorkonditionierte Verfahren etwa doppelt so viel Aufwand wie das einfache CG-Verfahren. Dagegen gilt für die Modellmatrix (vgl. Abschnitt 6.3) bei optimaler Wahl des Parameters ω (i. a. nicht leicht zu bestimmen!)

$$\kappa(\tilde{A}) = \sqrt{\kappa(A)}.$$

c) ICCG-Vorkonditionierung (**I**ncomplete **C**holesky **C**onjugate **G**radient): Die symmetrische, positiv definite Matrix A besitzt eine Cholesky-Zerlegung $A = LL^T$ mit einer unteren Dreiecksmatrix

$$L = \begin{bmatrix} l_{11} & & 0 \\ \vdots & \ddots & \\ l_{n1} & \cdots & l_{nn} \end{bmatrix}.$$

Die Elemente von L sind bestimmt durch die Rekursionsformeln

$$l_{ii} = [\, a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2 \,]^{1/2}\,, \quad i = 1, \ldots, n\,,$$
$$l_{ji} = [a_{ji} - \sum_{k=1}^{i-1} l_{jk} l_{ik}] / l_{ii}\,, \quad j = i+1, \ldots, n.$$

Die Matrix L hat i. Allg. innerhalb der Hülle von A von Null verschiedene Elemente, erfordert also in der Regel weit mehr Speicherplatz als A selbst. Dies wird jedoch dadurch ausgeglichen, dass man nur eine „unvollständige Cholesky-Zerlegung“ vornimmt, d. h.: Im

Cholesky-Algorithmus werden einige der l_{ji} willkürlich Null gesetzt, z. B.:

$$\tilde{l}_{ji} = 0\,, \quad \text{wenn} \quad a_{ji} = 0. \tag{6.2.65}$$

Dies ergibt dann eine Zerlegung

$$A = \tilde{L}\tilde{L}^T + E \tag{6.2.66}$$

mit einer unteren Dreiecksmatrix $\tilde{L} = (\tilde{l}_{ij})_{i,j=1,\ldots,n}$, welche eine ähnliche (dünne) Besetzungsstruktur wie A besitzt. Man spricht vom „ICCG(0)-Verfahren“, wenn (6.2.65) gefordert wird. Werden im Fall einer Bandmatrix A weitere p Nebendiagonalen mit von Null verschiedenen Elementen in $\tilde{L}$ hinzugefügt bzw. weggestrichen, so nennt man dies „ICCG(+p)- bzw. ICCG(-p)-Verfahren“. Zur Vorkonditionierung verwendet das ICCG-Verfahren die Matrix

$$C = KK^T = \tilde{L}\tilde{L}^T. \tag{6.2.67}$$

Obwohl keine strenge theoretische Begründung für den Erfolg dieses Ansatzes vorliegt, so zeigen doch numerische Tests an Modellproblemen, welchen Einfluß diese Konditionierung auf die Verteilung der Eigenwerte der Matrix $\tilde{A}$ hat. Zwar wird die Konditionszahl $\kappa(\tilde{A})$ nicht deutlich kleiner als $\kappa(A)$, doch die Eigenwerte von $\tilde{A}$ häufen sich im Gegensatz zu denen von A stark bei $\lambda = 1$. Dies bewirkt, wie eine feinere Analyse in (6.2.55) zeigt, eine deutliche Beschleunigung der Konvergenz.

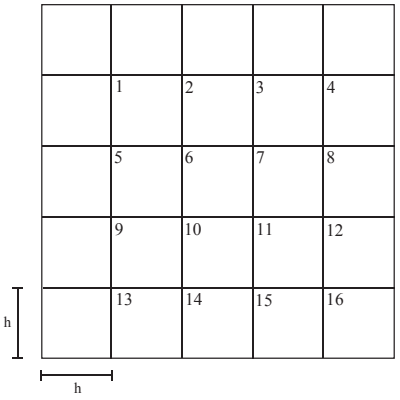
6.3 Ein Modellproblem

Wir wollen im Folgenden die Leistungsfähigkeit der bisher untersuchten Verfahren zur approximativen Lösung großer linearer Gleichungssysteme anhand eines Modellproblems vergleichen. Dazu betrachten wir zunächst das sog. „1. Randwertproblem des Laplace[6]-Operators“

$$\begin{aligned} -\frac{\partial^2 u}{\partial x^2}(x,y) - \frac{\partial^2 u}{\partial y^2}(x,y) &= f(x,y) \quad \text{für } (x,y) \in \Omega \\ u(x,y) &= 0 \quad \text{für } (x,y) \in \partial\Omega, \end{aligned} \tag{6.3.68}$$

auf dem Einheitsquadrat $\Omega = (0,1) \times (0,1) \subset \mathbb{R}^2$. Die Lösung $u = u(x,y)$ beschreibt u. a. die Auslenkung einer (idealisierten) elastischen Membran, die über dem Bereich $\bar{\Omega}$ horizontal gespannt und mit einer Kraftdichte f vertikal belastet wird. Eine Lösung ist i. Allg. nicht geschlossen angebbar, so dass man sich numerisch eine Näherungslösung verschafft. Dazu wird zunächst das Gebiet Ω mit einem Quadratgitter überdeckt.

[6]Pierre Simon Marquis de Laplace (1749–1827): Französischer Mathematiker und Astronom; Professor in Paris; begründete u.a. die Wahrscheinlichkeitsrechnung.



$h = \frac{1}{m+1}$ Gitterweite

$n = m^2$ Anzahl der gesuchten Knotenwerte

Abbildung 6.4: *Zur „5-Punkte-Diskretisierung“ des Modellproblems*

Die „inneren“ Gitterpunkte werden zeilenweise durchnumeriert. Ersetzt man dann in der obigen Differentialgleichung die 2. Ableitungen durch die entsprechenden zentralen Differenzenquotienten 2. Ordnung (sog. „5-Punkte-Differenzendiskretisierung“) und fordert die Gleichung nur in den inneren Gitterpunkten, so erhält man die Beziehungen

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

Durch Berücksichtigung der Randbedingung $u(x,y) = 0$ für $(x,y) \in \partial\Omega$ wird dies äquivalent zu dem linearen Gleichungssystem

$$Ax = b \tag{6.3.69}$$

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 uns schon bekannte Gestalt (Modellmatrix)

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

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

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

– symmetrisch, irreduzibel;

– schwach diagonal dominant und positiv definit;

– derartig, dass die Theorie für das SOR-Verfahren anwendbar ist.

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

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

Also ist (für $h \ll 1$)

$$\Lambda := \lambda_{\max} = 4 - 4\,\cos\,(1-h)\pi \;\approx\; 8$$
$$\lambda := \lambda_{\min} = 4 - 4\,\cos\,(h\pi) = 4 - 4\left(1 - \tfrac{1}{2}\pi^2h^2 + O(h^4)\right) \;\approx\; 2\pi^2h^2$$

und somit

$$\kappa \;:=\; \mathrm{cond}_{\mathrm{nat}}(\mathrm{A}) \;\approx\; \frac{4}{\pi^2\mathrm{h}^2} \tag{6.3.70}$$

Die Eigenwerte der Jacobi-Matrix $J = -D^{-1}(L+R)$ sind

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

Folglich wird

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

bzw.

$$\rho := \mathrm{spr(J)} = \mu_{\max} \;\approx\; 1 - \tfrac{1}{2}\pi^2\mathrm{h}^2. \tag{6.3.71}$$

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

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

$$\mathrm{spr}(\mathrm{H}_{\omega_{\mathrm{opt}}}) = \frac{1-\sqrt{1-\rho^2}}{1+\sqrt{1-\rho^2}} = \frac{1-\pi\mathrm{h}+\mathrm{O(h^2)}}{1+\pi\mathrm{h}+\mathrm{O(h^2)}} = 1 - 2\,\pi\mathrm{h} + \mathrm{O(h^2)}\,.$$

Wir kommen nun zum Leistungsvergleich der verschiedenen Verfahren. Um den Anfangsfehler $\|x^0 - x\|_2$ durch Anwendung der Fixpunktiterationen um den Faktor $\varepsilon \ll 1$ zu reduzieren, sind etwa

$$T(\varepsilon) \;\approx\; \frac{\ln(\varepsilon)}{\ln(\mathrm{spr(B)})}\,, \quad B = I - C^{-1}A \quad \text{Iterationsmatrix},$$

Iterationsschritte erforderlich. Es ergibt sich somit

$$T_J(\varepsilon) \sim \frac{\ln(\varepsilon)}{\ln(1-\frac{1}{2}\pi^2h^2)} \sim 2\frac{\ln(1/\varepsilon)}{\pi^2h^2} = \frac{2}{\pi^2}n\ \ln(1/\varepsilon),$$
$$T_{GS}(\varepsilon) \sim \frac{\ln(\varepsilon)}{\ln(1-\pi^2h^2)} \sim \frac{\ln(1/\varepsilon)}{\pi^2h^2} = \frac{1}{\pi^2}\,n\ \ln(1/\varepsilon),$$
$$T_{SOR}(\varepsilon) \sim \frac{\ln(\varepsilon)}{\ln(1-2\pi h)} \sim \frac{\ln(1/\varepsilon)}{2\pi h} = \frac{1}{2\pi}\sqrt{n}\ \ln(1/\varepsilon).$$

Das Gradientenverfahren und das CG-Verfahren benötigen zur Reduzierung des Anfangsfehlers $\|x^0 - x\|_2$ um den Faktor $\varepsilon \ll 1$ die folgenden Iterationszahlen:

$$T_G(\varepsilon) = \frac{1}{2}\,\kappa\ \ln(2/\varepsilon) \sim \frac{2}{\pi^2h^2}\ \ln(1/\varepsilon) \sim \frac{2}{\pi^2}n\ \ln(1/\varepsilon),$$
$$T_{CG}(\varepsilon) = \frac{1}{2}\,\sqrt{\kappa}\ \ln(2/\varepsilon) \sim \frac{1}{\pi h}\ \ln(2/\varepsilon) \sim \frac{1}{\pi}\,\sqrt{n}\ \ln(2/\varepsilon).$$

Wir sehen, dass das Jacobi-Verfahren und das Gradientenverfahren ungefähr gleich schnell sind. Das CG-Verfahren ist zwar nur halb so schnell wie das „optimale“ SOR-Verfahren, erfordert aber nicht die Bestimmung eines optimalen Iterationsparameters ω_{opt}.

Für die spezielle rechte Seite $f(x,y) = 2\pi^2\sin(\pi x)\sin(\pi y)$ ist die exakte Lösung der obigen Randwertaufgabe gerade $u(x,y) = \sin(\pi x)\sin(\pi y)$ und für den Diskretisierungsfehler der Differenzenapproximation gilt:

$$\max_{P_i} |u(P_i) - x_i| = \frac{\pi^2}{12}\,h^2 + O(h^4). \tag{6.3.72}$$

Zur Erzielung einer Genauigkeit von $\varepsilon = 10^{-4}$ (vier Stellen) ist also die Gitterweite

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

erforderlich. 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 $(\ln(1/\varepsilon) \sim 10)$:

$\text{spr}(\text{J})$	$\sim 0,9995$	$T_{\text{J}}(\varepsilon)$	~ 20.000	
$\text{spr}(\text{H}_1)$	$\sim 0,999$	$T_{\text{GS}}(\varepsilon)$	~ 10.000	
$\text{spr}(\text{H}_{\omega*})$	$\sim 0,9372$	$T_{\text{SOR}}(\varepsilon)$	~ 160	
$\text{cond}_{\text{nat}}(\text{A})$	~ 5.000	$T_G(\varepsilon)$	$\sim 20.000,$	$T_{\text{CG}}(\varepsilon) \sim 340$

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

$$\mathrm{OP_J} \approx \mathrm{OP_{H_1}} \approx \mathrm{OP_{H_\omega}} \approx 6\,\mathrm{n}\,,$$
$$\mathrm{OP_G} \approx \mathrm{OP_{CG}} \approx 10\,\mathrm{n}\,.$$

Als Endresultat finden wir, dass zur Bestimmung der Lösung des durch Diskretisierung der Randwertaufgabe (6.3.68) entstehenden $(n \times n)$-Gleichungssystems $Ax = b$ das Jacobi-Verfahren, das Gauß-Seidel-Verfahren und das Gradientenverfahren $O(n^2)$ a. Op. benötigen. Zur Lösung des Gleichungssystems $Ax = b$ mit einem direkten Verfahren würde man das Cholesky-Verfahren verwenden. Bei Berücksichtigung der speziellen Struktur der Modellmatrix erfordert dies $O(n^2) = O(m^2 n)$ a. Op. zur Berechnung der Zerlegung $A = LL^T$ und weitere $O(n^{3/2}) = O(mn)$ a. Op. für Vorwärts- und Rückwärtseinsetzen. Damit scheint das direkte Verfahren z. B. dem Gauß-Seidel-Verfahren überlegen zu sein. Es ist jedoch zu berücksichtigen, dass letzteres nur $O(n)$ Speicherplätze benötigt im Gegensatz zu den $O(n^{3/2}) = O(mn)$ für das Cholesky-Verfahren. In den letzten Jahren wurden sehr effiziente Verfahren zur Lösung von Problemen des obigen Typs entwickelt, welche „komplexitäts-optimal“ sind, d. h: die n Unbekannten mit $O(n)$ a. Op. berechnen.

6.4 Übungsaufgaben

Übung 6.1: Für die Matrizen

$$A_1 = \begin{bmatrix} 2 & -1 & 2 \\ 1 & 2 & -2 \\ 2 & 2 & 2 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 5 & 5 & 0 \\ -1 & 5 & 4 \\ 2 & 3 & 8 \end{bmatrix},$$

untersuche man, ob das Jacobi- und das Gauß-Seidel-Verfahren für die Gleichungssysteme $A_i x = b$ $(i = 1, 2)$ konvergiert. (Hinweis: Man wende die Konvergenzkriterien von oben an bzw. schätze den Spektralradius ab.)

Übung 6.2: Das Gleichungssystem

$$\begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

soll mit dem Jacobi- und dem Gauß-Seidel-Verfahren gelöst werden. Wieviele Iterationen sind jeweils ungefähr erforderlich, um den Iterationsfehler $\|x^t - x\|_2$ um den Faktor 10^{-6} zu reduzieren? (Hinweis: Man verwende die Fehlerabschätzung von oben.)

Übung 6.3: Zur Lösung des linearen (2×2)-Gleichungssystems

$$\begin{bmatrix} 1 & -a \\ -a & 1 \end{bmatrix} x = b, \quad x, b \in \mathbb{R}^2,$$

sei das folgende Iterationsverfahren angesetzt

$$\begin{bmatrix} 1 & 0 \\ -\omega a & 1 \end{bmatrix} x^t = \begin{bmatrix} 1-\omega & \omega a \\ 0 & 1-\omega \end{bmatrix} x^{t-1} + \omega b, \quad \omega \in \mathbb{R}.$$

a) Für welche $a \in \mathbb{R}$ ist diese Methode mit $\omega = 1$ konvergent?

b) Man bestimme für $a = 0.5$ den Wert

$$\omega \in \{0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4\},$$

für den der Spektralradius der Iterationsmatrix B_ω minimal wird und skizziere den Graphen der Funktion $f(\omega) = \mathrm{spr}(B_\omega)$.

Übung 6.4: Man zeige, dass die oben angegebenen beiden Definitionen der "Irreduzibilität" einer Matrix $A \in \mathbb{R}^{n\times n}$ äquivalent sind.
(Hinweis: Die Definition der „Nichtzerlegbarkeit" des Gleichungssystems in einer Form

$$PAP^T = \tilde{A} = \begin{bmatrix} \tilde{A}_{11} & 0 \\ \tilde{A}_{12} & \tilde{A}_{22} \end{bmatrix}, \quad \tilde{A}_{11} \in \mathbb{R}^{p\times p}, \quad \tilde{A}_{22} \in \mathbb{R}^{q\times q}, \quad n = p+q,$$

lässt sich auch wie folgt ausdrücken: Es gibt keine (nicht-triviale) Partitionierung $\{J, K\}$ von $N_n = \{1, \ldots, n\}$, $J \cup K = N_n$, $J \cap K = \emptyset$, so dass $a_{jk} = 0$ für $j \in J$, $k \in K$.)

Übung 6.5 (Praktische Aufgabe): Man betrachte das lineare Gleichungssystem $A_n x = b$ mit der $(n \times n)$-Blockmatrix $(n = m^2,\ h := (m+1)^{-1})$

$$A_n = \begin{bmatrix} B_m & -I_m & & \\ -I_m & B_m & \ddots & \\ & \ddots & \ddots & -I_m \\ & & -I_m & B_m \end{bmatrix} \in \mathbb{R}^{n\times n}, \qquad B_m = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{bmatrix} \in \mathbb{R}^{m\times m}$$

mit der Einheitsmatrix $I_m \in \mathbb{R}^{m\times m}$ und dem Vektor $b = h^2(1, \ldots, 1)^T \in \mathbb{R}^n$. Man schreibe ein Programm zur Lösung dieses Gleichungssystems mit Hilfe

a) des Jacobi-Verfahrens;

b) des Gauß-Seidel-Verfahrens;

c) des SOR-Verfahrens mit „optimalem" Relaxationsparameter ω_{opt} gemäß der oben angegebenen Theorie:

$$1 < \omega_{\mathrm{opt}} = \frac{2}{1 + \sqrt{1 - \mathrm{spr}(J)^2}} < 2, \qquad \mathrm{spr}(J) = \cos(h\pi) < 1.$$

Als Startvektoren verwende man jeweils $x^0 = 0$ und als Abbruchkriterium

$$\frac{\|Ax^t - b\|_\infty}{\|x^t\|_\infty} \leq 10^{-8} \qquad \text{oder} \qquad t_{\max} \leq 20000.$$

Für $m = 2^k$, $k = 1, \ldots, 6$, vergleiche man das Konvergenzverhalten dieser Verfahren. (Das Programm soll möglichst sparsam im Speicherverbrauch sein!)