Capitolo 5

METODI ITERATIVI PER LA RISOLUZIONE DI SISTEMI DI EQUAZIONI LINEARI

1. Successioni di vettori e di matrici

Per risolvere un sistema lineare $A\mathbf{x} = \mathbf{b}$, oltre ai metodi diretti, si possono utilizzare anche i metodi iterativi, che risultano particolarmente convenienti se la matrice A è sparsa, cioè se il numero degli elementi non nulli di A è dell'ordine della dimensione della matrice. Infatti quando si utilizza un metodo di risoluzione diretto, ad esempio il metodo di Gauss, può accadere che nelle matrici intermedie vengano generati molti elementi diversi da zero in corrispondenza ad elementi nulli della matrice iniziale (questo fenomeno si chiama *fill-in*). Poiché i metodi diretti non sfruttano adeguatamente la sparsità della matrice, per questo tipo di problemi, soprattutto se A è di grandi dimensioni, può essere più conveniente utilizzare un metodo iterativo. Esistono però dei casi nei quali la matrice A è sparsa, ma è comunque conveniente applicare dei metodi diretti che sfruttano specifiche proprietà di struttura della matrice.

5.1 Definizione. Una successione $\{\mathbf{x}^{(k)}\}$ di vettori di $\mathbf{C}^n$ si dice *convergente* al vettore $\mathbf{x}^*$ di $\mathbf{C}^n$ se esiste una norma per cui risulta

$$\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}^*\| = 0; \tag{1}$$

in tal caso si pone

$$\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x}^*. \qquad \blacksquare$$

Per il teorema 3.4 di equivalenza delle norme su $\mathbf{C}^n$, la definizione 5.1 non dipende dalla particolare norma considerata. La condizione di convergenza data dalla (1) si traduce in una condizione di convergenza delle successioni formate dalle singole componenti. Infatti, considerando la norma ∞, poiché è

$$|x_i^{(k)} - x_i^*| \le \|\mathbf{x}^{(k)} - \mathbf{x}^*\|_\infty, \quad i = 1, \ldots, n,$$

dalla (1) si ha

$$\lim_{k\to\infty} |x_i^{(k)} - x_i^*| = 0,$$

e quindi

$$\lim_{k\to\infty} x_i^{(k)} = x_i^*, \quad i = 1, ..., n; \tag{2}$$

viceversa, se vale la (2), è ovviamente verificata la condizione (1), per la norma ∞.

Per le successioni di matrici $\{A^{(k)}\}$ si può dare una definizione di convergenza analoga alla 5.1.

Il seguente teorema è di fondamentale importanza nello studio della convergenza dei metodi iterativi per la risoluzione dei sistemi lineari.

5.2 Teorema. *Sia $A \in \mathbf{C}^{n\times n}$, allora*

$$\lim_{k\to\infty} A^k = O \quad \textit{se e solo se} \quad \rho(A) < 1.$$

Dim. Per il teorema 2.18 esiste una matrice non singolare $T \in \mathbf{C}^{n\times n}$, tale che $A = TJT^{-1}$, dove J è la forma normale di Jordan di A; allora risulta

$$A^k = TJ^kT^{-1}. \tag{3}$$

Usando la notazione del teorema 2.18, risulta

$$J^k = \begin{bmatrix} J_1^k & & & \\ & J_2^k & & \\ & & \ddots & \\ & & & J_p^k \end{bmatrix},$$

dove

$$J_i^k = \begin{bmatrix} [C_i^{(1)}]^k & & & \\ & [C_i^{(2)}]^k & & \\ & & \ddots & \\ & & & [C_i^{(\tau(\lambda_i))}]^k \end{bmatrix},$$

per $i = 1, \ldots, p$, e i blocchi $C_i^{(j)} \in \mathbf{C}^{\nu_i(j)\times\nu_i(j)}$ per $j = 1, \ldots, \tau(\lambda_i)$ sono della forma

$$C_i^{(j)} = \lambda_i I + U,$$

in cui

$$U = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ & & & 0 \end{bmatrix}.$$

Per ogni i e j, $1 \le i \le p$, $1 \le j \le \tau(\lambda_i)$, risulta

$$[C_i^{(j)}]^k = (\lambda_i I + U)^k = \sum_{r=0}^{k} \binom{k}{r} \lambda_i^{k-r} U^r,$$

assumendo $U^0 = I$. Posto $s = \nu_i^{(j)}$, per $r \ge s$ risulta $U^r = O$, e quindi per $k \ge s$ è

$$[C_i^{(j)}]^k = \sum_{r=0}^{s-1} \binom{k}{r} \lambda_i^{k-r} U^r = \begin{bmatrix} \lambda_i^k & \binom{k}{1}\lambda_i^{k-1} & \cdots & \binom{k}{s-1}\lambda_i^{k-s+1} \\ & \lambda_i^k & \ddots & \\ & & \ddots & \binom{k}{1}\lambda_i^{k-1} \\ & & & \lambda_i^k \end{bmatrix}. \qquad (4)$$

Ne segue che condizione necessaria e sufficiente affinché λ_i^k e $\binom{k}{r}\lambda_i^{k-r}$ tendano a zero per $k \to \infty$ è che sia $|\lambda_i| < 1$ per $i = 1, \ldots, n$, cioè $\rho(A) < 1$.

∎

5.3 Esempio. La matrice

$$E = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

ha autovalori $\lambda_1 = \lambda_2 = 0$, $\lambda_3 = 3$. Quindi risulta

$$\rho(E) = 3 \text{ e } \lim_{k\to\infty} E^k \ne O.$$

Si osservi infatti che per $k \ge 1$ è

$$E^k = 3^{k-1} E.$$

La matrice $F = \frac{1}{4} E$ ha autovalori $\lambda_1 = \lambda_2 = 0,\ \lambda_3 = \frac{3}{4}$. Quindi risulta

$$\rho(F) = \frac{3}{4} < 1 \text{ e } \lim_{k\to\infty} F^k = O.$$

Si osservi infatti che per $k \ge 1$ è

$$F^k = \left(\frac{3}{4}\right)^{k-1} F.$$

∎

5.4 Teorema. *Sia $A \in \mathbf{C}^{n\times n}$. Allora*

$$\det(I-A)\neq 0 \ e \ \lim_{k\to\infty}\sum_{i=0}^{k} A^i = (I-A)^{-1} \quad \textit{se e solo se} \quad \rho(A)<1.$$

Dim. Sia $\rho(A)<1$, allora gli autovalori di A hanno tutti modulo minore di 1, quindi la matrice $I-A$ non ha autovalori nulli e risulta non singolare. Inoltre, poiché

$$(I-A)\sum_{i=0}^{k} A^i = I - A^{k+1},$$

si ha

$$\sum_{i=0}^{k} A^i = (I-A)^{-1}(I-A^{k+1}),$$

e quindi

$$\begin{aligned}\lim_{k\to\infty}\sum_{i=0}^{k} A^i &= \lim_{k\to\infty}(I-A)^{-1}(I-A^{k+1})\\ &= (I-A)^{-1}\lim_{k\to\infty}(I-A^{k+1}) = (I-A)^{-1},\end{aligned}$$

in quanto per il teorema 5.2 si ha $\lim_{k\to\infty} A^{k+1} = O$. Viceversa, sia $I-A$ non singolare e

$$\lim_{k\to\infty}\sum_{i=0}^{k} A^i = (I-A)^{-1}.$$

Indicato con λ un autovalore di A tale che $|\lambda| = \rho(A)$ e con $\mathbf{x}$ un autovettore corrispondente a λ, è $\lambda \neq 1$ perché $I-A$ è non singolare, ed inoltre vale

$$\lim_{k\to\infty}\sum_{i=0}^{k} A^i\mathbf{x} = (I-A)^{-1}\mathbf{x}$$

e quindi

$$\lim_{k\to\infty}\Big(\sum_{i=0}^{k}\lambda^i\Big)\mathbf{x} = \frac{1}{1-\lambda}\ \mathbf{x}.$$

Ne segue la convergenza della serie numerica

$$\sum_{i=0}^{\infty}\lambda^i = \frac{1}{1-\lambda},$$

per cui $|\lambda| < 1$. ∎

Come per le serie numeriche, si usa scrivere

$$\sum_{i=0}^{\infty} A^i = (I - A)^{-1}.$$

2. Generalità sui metodi iterativi

Sia $A \in \mathbf{C}^{n\times n}$ una matrice non singolare e si consideri la decomposizione di A nella forma

$$A = M - N, \tag{5}$$

dove M è una matrice non singolare. Dalla (5), sostituendo nel sistema lineare

$$A\mathbf{x} = \mathbf{b}, \tag{6}$$

risulta

$$M\mathbf{x} - N\mathbf{x} = \mathbf{b},$$

cioè

$$\mathbf{x} = M^{-1}N\mathbf{x} + M^{-1}\mathbf{b}.$$

Posto

$$P = M^{-1}N \quad \text{e} \quad \mathbf{q} = M^{-1}\mathbf{b}, \tag{7}$$

si ottiene il seguente sistema

$$\mathbf{x} = P\mathbf{x} + \mathbf{q}, \tag{8}$$

equivalente al sistema (6).

Dato un vettore iniziale $\mathbf{x}^{(0)}$, si considera la successione $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$, così definita

$$\mathbf{x}^{(k)} = P\mathbf{x}^{(k-1)} + \mathbf{q}, \quad k = 1, 2, \ldots\ . \tag{9}$$

Se la successione $\mathbf{x}^{(k)}$ è convergente e si indica con

$$\mathbf{x}^* = \lim_{k\to\infty} \mathbf{x}^{(k)},$$

allora passando al limite nella (9) risulta

$$\mathbf{x}^* = P\mathbf{x}^* + \mathbf{q}, \tag{10}$$

cioè $\mathbf{x}^*$ è la soluzione del sistema (8) e quindi del sistema (6).

La relazione (9) individua un *metodo iterativo* in cui, partendo da un vettore iniziale $\mathbf{x}^{(0)}$, la soluzione viene approssimata utilizzando una successione $\{\mathbf{x}^{(k)}\}$ di vettori. La matrice P si dice *matrice di iterazione del metodo*.

Al variare del vettore iniziale $\mathbf{x}^{(0)}$ si ottengono dalla (9) diverse successioni $\{\mathbf{x}^{(k)}\}$, alcune delle quali possono essere convergenti ed altre no. Un metodo iterativo è detto *convergente* se, qualunque sia il vettore iniziale $\mathbf{x}^{(0)}$, la successione $\{\mathbf{x}^{(k)}\}$ è convergente.

5.5 Esempio. Si consideri il sistema (8) in cui

$$P = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 2 \end{bmatrix}, \quad \mathbf{q} = \mathbf{0}, \quad \text{e quindi} \quad \mathbf{x}^* = \mathbf{0}.$$

Allora

$$P^k = \begin{bmatrix} \left(\frac{1}{2}\right)^k & 0 & 0 \\ 0 & \left(\frac{1}{2}\right)^k & 0 \\ 0 & 0 & 2^k \end{bmatrix}.$$

Se $\mathbf{x}^{(0)} = [1, 0, 0]^T$, si ottiene la successione

$$\mathbf{x}^{(k)} = [\left(\tfrac{1}{2}\right)^k, 0, 0]^T, \ k = 1, 2, \ldots,$$

che converge alla soluzione del sistema. Se invece $\mathbf{x}^{(0)} = [0, 1, 1]^T$, si ottiene la successione

$$\mathbf{x}^{(k)} = [0, \left(\tfrac{1}{2}\right)^k, 2^k]^T, \ k = 1, 2, \ldots,$$

che non converge. Questo è un esempio di metodo non convergente. ∎

5.6 Teorema. *Il metodo iterativo (9) è convergente se e solo se* $\rho(P) < 1$.

Dim. Sia $\mathbf{x}^*$ la soluzione del sistema (6), che soddisfa quindi la (10). Sottraendo membro a membro la (9) dalla (10) risulta

$$\mathbf{x}^* - \mathbf{x}^{(k)} = P(\mathbf{x}^* - \mathbf{x}^{(k-1)}), \quad k = 1, 2, \ldots \ . \tag{11}$$

Indicato con

$$\mathbf{e}^{(k)} = \mathbf{x}^* - \mathbf{x}^{(k)},$$

il vettore *errore* alla k-esima iterazione, si ha dalla (11)

$$\mathbf{e}^{(k)} = P\mathbf{e}^{(k-1)}, \quad k = 1, 2, \ldots \tag{12}$$

e quindi

$$\mathbf{e}^{(k)} = P\mathbf{e}^{(k-1)} = P^2\mathbf{e}^{(k-2)} = \ldots = P^k\mathbf{e}^{(0)}. \tag{13}$$

Se $\rho(P) < 1$, per il teorema 5.2 risulta

$$\lim_{k\to\infty} P^k = O,$$

e dalla (13), per ogni vettore $\mathbf{e}^{(0)}$, segue che

$$\lim_{k\to\infty} \mathbf{e}^{(k)} = \mathbf{0}. \tag{14}$$

Viceversa, se il metodo è convergente, la (14) vale per ogni $\mathbf{x}^{(0)}$, e in particolare deve valere se $\mathbf{x}^{(0)}$ è tale che il vettore $\mathbf{e}^{(0)} = \mathbf{x}^* - \mathbf{x}^{(0)}$ è un autovettore di P corrispondente ad un autovalore λ di modulo massimo, cioè $|\lambda| = \rho(P)$. In questo caso risulta

$$P\mathbf{e}^{(0)} = \lambda\mathbf{e}^{(0)}$$

e quindi

$$\mathbf{e}^{(k)} = P^k\mathbf{e}^{(0)} = \lambda^k\mathbf{e}^{(0)}.$$

Ne segue che

$$\lim_{k\to\infty} [\rho(P)]^k = 0$$

e quindi $\rho(P) < 1$. ∎

La condizione $\rho(P) < 1$, necessaria e sufficiente per la convergenza del metodo (9), non è in generale di agevole verifica. Conviene allora utilizzare, quando è possibile, delle condizioni sufficienti di convergenza di più facile verifica. Una tale condizione è data nel seguente teorema.

5.7 Teorema. *Se esiste una norma matriciale indotta* $\|\ .\ \|$ *per cui* $\|P\| < 1$, *il metodo iterativo (9) è convergente.*

Dim. La tesi segue dal teorema 5.6 e dalla proprietà

$$\rho(P) \le \|P\|,$$

dimostrata nel teorema 3.10. ∎

Poiché il determinante di una matrice è uguale al prodotto degli autovalori, se $|\det P| \ge 1$, almeno uno degli autovalori di P è in modulo maggiore o uguale a 1 e quindi il metodo (9) non è convergente. Poiché la traccia di una matrice è uguale alla somma degli autovalori, se $|\text{tr } P| \ge n$, almeno uno degli autovalori di P è in modulo maggiore o uguale a 1 e quindi il metodo (9) non è convergente. Quindi le condizioni $|\det P| < 1$ e $|\text{tr } P| < n$ sono necessarie affinché il metodo iterativo (9) sia convergente.

3. Controllo della convergenza

Fissata una norma vettoriale $\| \cdot \|$ e la corrispondente norma matriciale indotta, dalla (13) si ottiene la seguente maggiorazione della norma dell'errore da cui è affetto $\mathbf{x}^{(k)}$ rispetto alla soluzione del sistema $\mathbf{x}^*$:

$$\|\mathbf{e}^{(k)}\| \leq \|P^k\| \, \|\mathbf{e}^{(0)}\|, \tag{15}$$

dove il segno di uguaglianza vale per particolari vettori $\mathbf{e}^{(0)}$, perché la norma matriciale considerata è indotta. Quindi $\|P^k\|$ esprime la riduzione, rispetto all'errore iniziale, dell'errore al k-esimo passo. Questa misura risulta però inadatta per una valutazione della velocità di convergenza di un metodo, che sia indipendente dal numero delle iterazioni. Infatti, se P e Q sono due matrici di iterazione associate a due diversi metodi, può accadere che per una particolare norma $\| \cdot \|$ esistano due interi j e k, con $k \neq j$, tali che

$$\|P^k\| < \|Q^k\| \quad \text{e} \quad \|P^j\| > \|Q^j\|.$$

5.8 Esempio. Siano

$$P = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.6 \end{bmatrix}, \quad Q = \begin{bmatrix} 0.5 & 0.25 \\ 0 & 0.5 \end{bmatrix}.$$

Si ha

$$P^k = \begin{bmatrix} 0.5^k & 0 \\ 0 & 0.6^k \end{bmatrix}, \quad Q^k = \begin{bmatrix} 0.5^k & k\ 0.5^{k+1} \\ 0 & 0.5^k \end{bmatrix}.$$

Utilizzando la norma ∞ risulta

$$\|P^k\|_\infty = 0.6^k \quad \text{e} \quad \|Q^k\|_\infty = (2+k)0.5^{k+1}.$$

Per $k = 1, \ldots, 15$ si ottengono i valori

k	$\|P^k\|_\infty$	$\|Q^k\|_\infty$
1	0.6000000	0.7500000
2	0.3600000	0.5000000
3	0.2160000	0.3125000
.	.	.
.	.	.
8	$0.1679614\ 10^{-1}$	$0.1953125\ 10^{-1}$
9	$0.1007769\ 10^{-1}$	$0.1074219\ 10^{-1}$
10	$0.6046608\ 10^{-2}$	$0.5859375\ 10^{-2}$
11	$0.3627964\ 10^{-2}$	$0.3173828\ 10^{-2}$
.	.	.
.	.	.
15	$0.4701840\ 10^{-3}$	$0.2593994\ 10^{-3}$

Si noti che $\|P^k\|_\infty < \|Q^k\|_\infty$ per $k \leq 9$, e $\|P^k\|_\infty > \|Q^k\|_\infty$ per $k \geq 10$. Utilizzando la norma 2, per $k = 1, \ldots, 15$, si ottengono i valori

k	$\|P^k\|_2$	$\|Q^k\|_2$
1	0.6000000	0.6403882
2	0.3600000	0.4045085
3	0.2160000	0.2500000
.	.	.
.	.	.
6	$0.4665595\ 10^{-1}$	$0.5160587\ 10^{-1}$
7	$0.2799357\ 10^{-1}$	$0.2941847\ 10^{-1}$
8	$0.1679614\ 10^{-1}$	$0.1654713\ 10^{-1}$
9	$0.1007769\ 10^{-1}$	$0.9203542\ 10^{-2}$
.	.	.
.	.	.
15	$0.4701840\ 10^{-3}$	$0.2328810\ 10^{-3}$

e quindi $\|P^k\|_2 < \|Q^k\|_2$ per $k \leq 7$, e $\|P^k\|_2 > \|Q^k\|_2$ per $k \geq 8$. ∎

Se $\mathbf{e}^{(k-1)} \neq \mathbf{0}$, la quantità $\|\mathbf{e}^{(k)}\|/\|\mathbf{e}^{(k-1)}\|$ esprime la riduzione dell'errore al k-esimo passo e la media geometrica delle riduzioni dell'errore sui primi k passi:

$$\sigma_k = \sqrt[k]{\frac{\|\mathbf{e}^{(1)}\|}{\|\mathbf{e}^{(0)}\|}\ \frac{\|\mathbf{e}^{(2)}\|}{\|\mathbf{e}^{(1)}\|}\ \cdots\ \frac{\|\mathbf{e}^{(k)}\|}{\|\mathbf{e}^{(k-1)}\|}} = \sqrt[k]{\frac{\|\mathbf{e}^{(k)}\|}{\|\mathbf{e}^{(0)}\|}}$$

esprime la *riduzione media per passo* dell'errore relativo ai primi k passi. Dalla (15) risulta

$$\sigma_k \leq \sqrt[k]{\|P^k\|},$$

dove il segno di uguaglianza vale per particolari vettori $\mathbf{e}^{(0)}$. La quantità che si ottiene facendo tendere k all'infinito esprime la *riduzione asintotica media per passo* e, come risulta dal seguente teorema, è indipendente dalla particolare norma utilizzata.

5.9 Teorema. *Sia $A \in \mathbf{C}^{n\times n}$ e sia $\|\ .\ \|$ una qualunque norma indotta. Allora*

$$\lim_{k\to\infty} \sqrt[k]{\|A^k\|} = \rho(A).$$

Dim. Si dimostra prima che il limite, se esiste, non dipende dalla particolare norma usata. Per l'equivalenza delle norme, se $\|\ .\ \|'$ e $\|\ .\ \|''$ sono due norme matriciali indotte, esistono due costanti α e β positive, tali che

$$\alpha\ \|A^k\|'' \leq \|A^k\|' \leq \beta\ \|A^k\|'',$$

per cui

$$\sqrt[k]{\alpha}\,\sqrt[k]{\|A^k\|''} \leq \sqrt[k]{\|A^k\|'} \leq \sqrt[k]{\beta}\,\sqrt[k]{\|A^k\|''}.$$

Poiché

$$\lim_{k\to\infty} \sqrt[k]{\alpha} = \lim_{k\to\infty} \sqrt[k]{\beta} = 1,$$

dalla relazione precedente segue che se esiste

$$\lim_{k\to\infty} \sqrt[k]{\|A^k\|''},$$

allora esiste anche

$$\lim_{k\to\infty} \sqrt[k]{\|A^k\|'},$$

e tali limiti coincidono. Si dimostra adesso che il limite esiste per un'opportuna norma indotta. Dalla (3) si ha $A^k = TJ^kT^{-1}$, dove J^k è una matrice diagonale formata dai blocchi $[C_i^{(j)}]^k$, $i = 1, \ldots, p$, $j = 1, \ldots, \tau(\lambda_i)$, in cui λ_i, $i = 1, \ldots, p$, sono gli autovalori distinti di A e i blocchi $[C_i^{(j)}]^k$ sono quelli riportati nella (4). Per il teorema 3.11 l'applicazione

$$A \to \|T^{-1}AT\|_\infty$$

è una norma indotta di A e, indicando con $\|\ .\ \|$ tale norma, risulta $\|A^k\| = \|J^k\|_\infty$. Se λ_1 è l'autovalore di A per cui $|\lambda_1| = \rho(A)$, e, fra tutti i blocchi relativi a $\lambda_1, C_1^{(1)}$ è quello di ordine s massimo, allora esiste un intero k_0 tale che per ogni $k \geq k_0$ si ha

$$\|A^k\| = \|[C_1^{(1)}]^k\|_\infty = \sum_{r=0}^{s-1} \binom{k}{r} |\lambda_1|^{k-r} = [\rho(A)]^k \sum_{r=0}^{s-1} \binom{k}{r} [\rho(A)]^{-r}.$$

La quantità

$$p(k) = \sum_{r=0}^{s-1} \binom{k}{r} [\rho(A)]^{-r}$$

è un polinomio in k di grado $s-1$, e quindi

$$\lim_{k\to\infty} \sqrt[k]{p(k)} = 1.$$

Ne segue che il

$$\lim_{k\to\infty} \sqrt[k]{\|A^k\|}$$

esiste e vale $\rho(A)$. ∎

La quantità $\rho(P)$, indipendente dalla norma utilizzata e dall'indice di iterazione k, viene quindi assunta come misura della velocità di convergenza del metodo (9). Il numero k di iterazioni richieste per ridurre l'errore di $1/10$ (cioè, approssimativamente, per ottenere una cifra decimale in più) è tale che

$$[\rho(P)]^k \approx \frac{1}{10}, \quad \text{da cui} \quad k \approx -1/\log_{10}\rho(P).$$

5.10 Definizione. Si definisce *tasso asintotico di convergenza* del metodo iterativo (9) la costante $R = -\log_{10}\rho(P)$. ∎

Poiché con un metodo iterativo non è ovviamente possibile calcolare in generale la soluzione con un numero finito di iterazioni, occorre individuare dei criteri per l'arresto del procedimento. I criteri più comunemente usati, fissata una tolleranza ϵ, che tiene conto anche della precisione utilizzata nei calcoli, sono i seguenti:

$$\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq \epsilon, \tag{16}$$

oppure, se $\mathbf{x}^{(k)} \neq \mathbf{0}$,

$$\frac{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|}{\|\mathbf{x}^{(k)}\|} \leq \epsilon. \tag{17}$$

Si noti però che le condizioni (16) e (17) non garantiscono che la soluzione sia stata approssimata con la precisione ϵ. Infatti per la (12) è:

$$\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)} = [\mathbf{x}^* - \mathbf{x}^{(k-1)}] - [\mathbf{x}^* - \mathbf{x}^{(k)}] = \mathbf{e}^{(k-1)} - \mathbf{e}^{(k)} = (I - P)\mathbf{e}^{(k-1)}$$

e, passando alle norme, se $\|P\| < 1$, per il teorema 3.13 si ha:

$$\|\mathbf{e}^{(k-1)}\| \leq \|(I-P)^{-1}\|\ \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq \frac{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|}{1 - \|P\|},$$

per cui può accadere che $\|\mathbf{e}^{(k-1)}\|$ sia elevata anche se la condizione (16) è verificata.

In un programma che implementa un metodo iterativo deve essere comunque previsto un controllo per interrompere l'esecuzione quando il numero delle iterazioni diventa troppo elevato. Può anche accadere che un metodo iterativo la cui matrice di iterazione P è tale che $\rho(P) < 1$, per gli effetti indotti dagli errori di arrotondamento non converga in pratica, e questo accade, in particolare, quando la matrice A è fortemente mal condizionata e $\rho(P)$ è molto vicino ad 1.

È opportuno rilevare che un metodo iterativo rispetto ad un metodo diretto è in generale meno sensibile alla propagazione degli errori. Infatti

il vettore $\mathbf{x}^{(k)}$ può essere considerato come il vettore generato con una sola iterazione a partire dal vettore iniziale $\mathbf{x}^{(k-1)}$, e quindi risulta affetto dagli errori di arrotondamento generati dalla sola ultima iterazione.

In un metodo iterativo ad ogni iterazione il costo computazionale è principalmente determinato dalla operazione di moltiplicazione della matrice P per un vettore, che richiede n^2 operazioni moltiplicative se la matrice A non ha specifiche proprietà. Se invece A è sparsa, cioè ha un numero di elementi non nulli dell'ordine di n, la moltiplicazione di P per un vettore richiede un numero di operazioni moltiplicative dell'ordine di n. In questo caso i metodi iterativi possono risultare competitivi con quelli diretti. Particolarmente interessante è il caso in cui la matrice, oltre a essere sparsa, ha specifiche proprietà di struttura, che possono essere convenientemente sfruttate anche per ridurre l'ingombro di memoria richiesto.

4. Metodi iterativi di Jacobi e Gauss-Seidel

Fra i metodi iterativi individuati da una particolare scelta della decomposizione (5) sono particolarmente importanti il metodo di Jacobi e il metodo di Gauss-Seidel, per i quali è possibile dare delle condizioni sufficienti di convergenza verificate da molte delle matrici che si ottengono risolvendo problemi differenziali.

Si consideri la decomposizione della matrice A

$$A = D - B - C$$

dove

$$d_{ij} = \begin{cases} a_{ij} & \text{se } i = j \\ 0 & \text{se } i \neq j, \end{cases} \qquad b_{ij} = \begin{cases} \text{-}a_{ij} & \text{se } i > j \\ 0 & \text{se } i \leq j, \end{cases} \qquad c_{ij} = \begin{cases} 0 & \text{se } i \geq j \\ \text{-}a_{ij} & \text{se } i < j. \end{cases}$$

Scegliendo $M = D, \quad N = B + C$, si ottiene il *metodo di Jacobi*.
Scegliendo $M = D - B, \quad N = C$, si ottiene il *metodo di Gauss-Seidel*.

Per queste decomposizioni risulta $\det M \neq 0$ se e solo se tutti gli elementi principali di A sono non nulli.

Indicando con J la matrice di iterazione del metodo di Jacobi, dalla (7) si ha

$$J = D^{-1}(B + C),$$

per cui la (9) diviene:

$$\mathbf{x}^{(k)} = J\mathbf{x}^{(k-1)} + D^{-1}\mathbf{b}$$

e, in termini di componenti :

$$x_i^{(k)} = \frac{1}{a_{ii}} \left[b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k-1)} \right], \quad i = 1, 2, \ldots, n. \tag{18}$$

Il metodo di Jacobi è detto anche *metodo degli spostamenti simultanei*, in quanto le componenti del vettore $\mathbf{x}^{(k)}$ sostituiscono simultaneamente al termine dell'iterazione le componenti di $\mathbf{x}^{(k-1)}$.

Indicando con G la matrice di iterazione del metodo di Gauss-Seidel, dalla (7) si ha

$$G = (D - B)^{-1} C,$$

per cui la (9) diviene:

$$\mathbf{x}^{(k)} = G\mathbf{x}^{(k-1)} + (D - B)^{-1}\mathbf{b}. \tag{19}$$

Per descrivere la (19) in termini di componenti, conviene prima trasformarla nel modo seguente:

$$(D - B)\mathbf{x}^{(k)} = C\mathbf{x}^{(k-1)} + \mathbf{b}$$

$$D\mathbf{x}^{(k)} = B\mathbf{x}^{(k)} + C\mathbf{x}^{(k-1)} + \mathbf{b}$$

$$\mathbf{x}^{(k)} = D^{-1}B\mathbf{x}^{(k)} + D^{-1}C\mathbf{x}^{(k-1)} + D^{-1}\mathbf{b}, \tag{20}$$

ottenendo quindi:

$$x_i^{(k)} = \frac{1}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} \right], \quad i = 1, 2, \ldots, n. \tag{21}$$

Confrontando la (21) con la (18), risulta che nel metodo di Gauss-Seidel per calcolare le componenti del vettore $\mathbf{x}^{(k)}$ (contrariamente a quanto accade nel metodo di Jacobi) sono utilizzate componenti già calcolate dello stesso vettore. Per questo motivo il metodo prende anche il nome di *metodo degli spostamenti successivi*. Quindi nella implementazione del metodo di Jacobi è necessario disporre, contemporaneamente, di entrambi i vettori $\mathbf{x}^{(k)}$ e $\mathbf{x}^{(k-1)}$, mentre per il metodo di Gauss-Seidel è sufficiente disporre di un solo vettore.

In molte applicazioni il metodo di Gauss-Seidel, che utilizza immediatamente i valori calcolati nella iterazione corrente, risulta più veloce del metodo di Jacobi. Però esistono casi in cui risulta non solo che il metodo di Jacobi sia più veloce del metodo di Gauss-Seidel, ma anche che il metodo di Jacobi sia convergente e quello di Gauss-Seidel no.

5.11 Esempi. Si esamina la convergenza dei metodi di Jacobi e di Gauss-Seidel applicati al sistema $A\mathbf{x} = \mathbf{b}$, per diverse matrici $A \in \mathbf{R}^{3\times 3}$. Il vettore $\mathbf{b}$ è sempre scelto in modo che la soluzione sia $\mathbf{x}^* = [1, 1, 1]^T$. Il criterio di arresto utilizzato è quello espresso dalla (16), in norma ∞, con $\epsilon = 10^{-5}$. Si noti che, poiché $\|\mathbf{x}^*\|_\infty = 1$, in questo caso i criteri di arresto espressi dalla (16) e dalla (17) da un certo valore di k in poi sono equivalenti.

Nelle figure sono riportati i grafici delle norme degli errori assoluti $\|\mathbf{e}^{(k)}\|_\infty$ delle successioni ottenute a partire dal vettore iniziale $\mathbf{x}^{(0)} = \mathbf{0}$. Con i quadratini vuoti sono indicati gli errori generati dal metodo di Jacobi, con i quadratini pieni gli errori generati dal metodo di Gauss-Seidel.

a) Nel caso

$$A = \begin{bmatrix} 3 & 0 & 4 \\ 7 & 4 & 2 \\ -1 & -1 & -2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 7 \\ 13 \\ -4 \end{bmatrix} \tag{22}$$

risulta

$$J = \begin{bmatrix} 0 & 0 & -\frac{4}{3} \\ -\frac{7}{4} & 0 & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & 0 \end{bmatrix} \quad G = \begin{bmatrix} 0 & 0 & -\frac{4}{3} \\ 0 & 0 & \frac{11}{6} \\ 0 & 0 & -\frac{1}{4} \end{bmatrix}$$

e $\rho(J) = 1.337510$, $\rho(G) = 0.25$. Quindi il metodo di Gauss-Seidel è convergente mentre il metodo di Jacobi non lo è. La successione ottenuta con il metodo di Gauss-Seidel si arresta alla 11-esima iterazione. I grafici degli errori sono riportati nella figura 5.1.

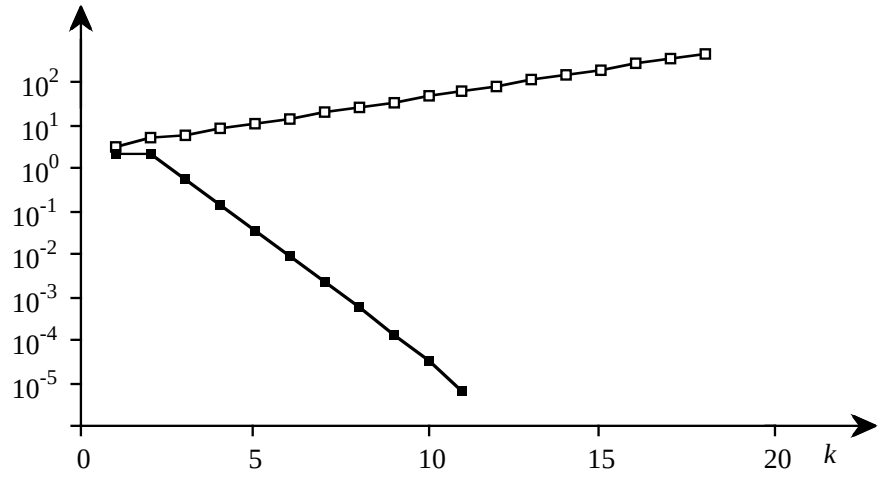

Fig. 5.1 - Grafici degli errori dei metodi di Jacobi e di Gauss-Seidel per il problema (22).

b) Nel caso

$$A = \begin{bmatrix} -3 & 3 & -6 \\ -4 & 7 & -8 \\ 5 & 7 & -9 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} -6 \\ -5 \\ 3 \end{bmatrix} \tag{23}$$

risulta

$$J = \begin{bmatrix} 0 & 1 & -2 \\ \frac{4}{7} & 0 & \frac{8}{7} \\ \frac{5}{9} & \frac{7}{9} & 0 \end{bmatrix} \quad G = \begin{bmatrix} 0 & 1 & -2 \\ 0 & \frac{4}{7} & 0 \\ 0 & 1 & -\frac{10}{9} \end{bmatrix}$$

e $\rho(J) = 0.8133091$, $\rho(G) = 1.111111$. Quindi il metodo di Jacobi è convergente e il metodo di Gauss-Seidel non lo è. La successione ottenuta con il metodo di Jacobi si arresta alla 49-esima iterazione. I grafici degli errori sono riportati nella figura 5.2.

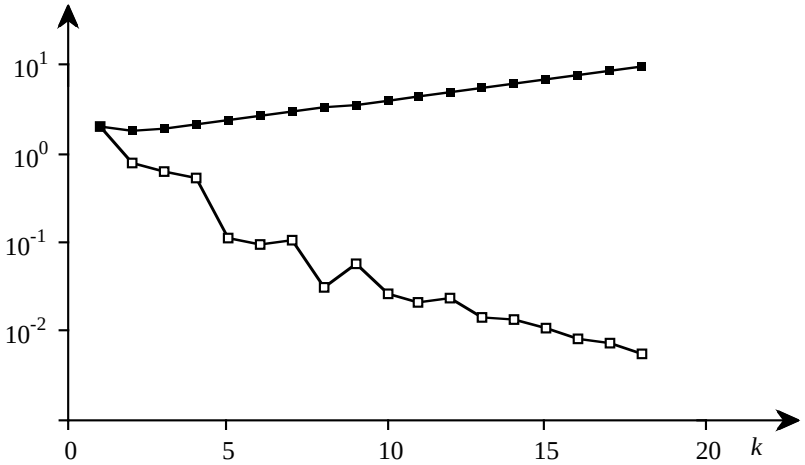

Fig. 5.2 - Grafici degli errori dei metodi di Jacobi e di Gauss-Seidel per il problema (23).

c) Nel caso

$$A = \begin{bmatrix} 4 & 1 & 1 \\ 2 & -9 & 0 \\ 0 & -8 & -6 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 6 \\ -7 \\ -14 \end{bmatrix} \tag{24}$$

risulta

$$J = \begin{bmatrix} 0 & -\frac{1}{4} & -\frac{1}{4} \\ \frac{2}{9} & 0 & 0 \\ 0 & -\frac{4}{3} & 0 \end{bmatrix} \quad G = \begin{bmatrix} 0 & -\frac{1}{4} & -\frac{1}{4} \\ 0 & -\frac{1}{18} & -\frac{1}{18} \\ 0 & \frac{2}{27} & \frac{2}{27} \end{bmatrix}$$

e $\rho(J) = 0.4438188, \quad \rho(G) = 0.01851852$. Quindi entrambi i metodi sono convergenti e la successione generata dal metodo di Gauss-Seidel, che si arresta alla quinta iterazione, converge più rapidamente di quella generata dal metodo di Jacobi, che si arresta alla 16-esima iterazione. I grafici degli errori sono riportati nella figura 5.3.

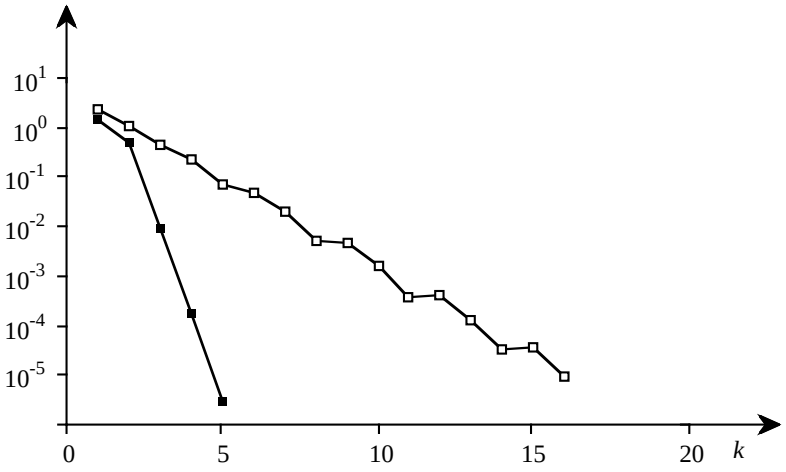

Fig. 5.3 - Grafici degli errori dei metodi di Jacobi e di Gauss-Seidel per il problema (24).

d) Nel caso

$$A = \begin{bmatrix} 7 & 6 & 9 \\ 4 & 5 & -4 \\ -7 & -3 & 8 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 22 \\ 5 \\ -2 \end{bmatrix} \tag{25}$$

risulta

$$J = \begin{bmatrix} 0 & -\frac{6}{7} & -\frac{9}{7} \\ -\frac{4}{5} & 0 & \frac{4}{5} \\ \frac{7}{8} & \frac{3}{8} & 0 \end{bmatrix} \qquad G = \begin{bmatrix} 0 & -\frac{6}{7} & -\frac{9}{7} \\ 0 & \frac{24}{35} & \frac{64}{35} \\ 0 & -\frac{69}{140} & -\frac{123}{280} \end{bmatrix}$$

e $\rho(J) = 0.6411328, \quad \rho(G) = 0.7745967$. Quindi entrambi i metodi sono convergenti e la successione generata dal metodo di Jacobi, che si arresta alla 30-esima iterazione, converge più rapidamente di quella generata dal metodo di Gauss-Seidel, che si arresta alla 48-esima iterazione. I grafici degli errori sono riportati nella figura 5.4. ∎

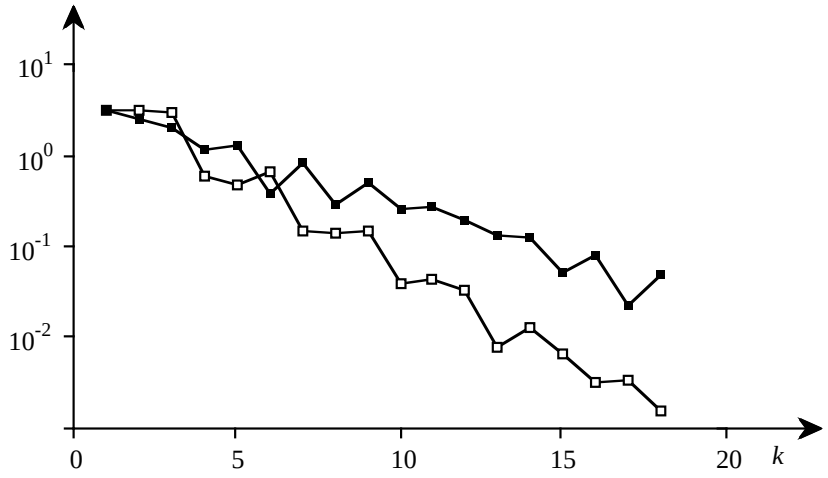

Fig. 5.4 - Grafici degli errori dei metodi di Jacobi e di Gauss-Seidel per il problema (25).

Dai teoremi 5.6 e 5.7 si possono ricavare delle condizioni di convergenza per i metodi di Jacobi e di Gauss-Seidel, applicati alla risoluzione del sistema lineare $A\mathbf{x} = \mathbf{b}$, $A \in \mathbf{C}^{n\times n}$. Particolarmente importanti e di facile verifica sono le condizioni basate sulla proprietà di predominanza della matrice A (si vedano le definizioni 2.40).

5.12 Teorema. *Sia $A = M - N$ la decomposizione della matrice A corrispondente al metodo di Jacobi (cioè $M = D$ e $N = B + C$) o al metodo di Gauss-Seidel (cioè $M = D - B$ e $N = C$). Se vale una delle seguenti ipotesi:*

a) la matrice A è a predominanza diagonale in senso stretto,

b) la matrice A è a predominanza diagonale ed è irriducibile,

c) la matrice A è a predominanza diagonale in senso stretto per colonne,

d) la matrice A è a predominanza diagonale per colonne ed è irriducibile,

allora $\rho(M^{-1}N) < 1$ e quindi il metodo di Jacobi e il metodo di Gauss-Seidel sono convergenti.

Dim. Nelle ipotesi fatte, gli elementi principali di A sono non nulli e quindi la matrice M è non singolare. Un numero complesso λ è autovalore di $M^{-1}N$ se e solo se

$$\det(M^{-1}N - \lambda I) = 0, \tag{26}$$

ed essendo $M^{-1}N - \lambda I = -M^{-1}(\lambda M - N)$, per la regola di Binet dalla (26) segue che λ è autovalore di $M^{-1}N$ se e solo se

$$\det(\lambda M - N) = 0. \tag{27}$$

La matrice $H = \lambda M - N$ ha gli elementi

$$h_{ij} = \begin{cases} \lambda a_{ij} & \text{se } i = j \\ a_{ij} & \text{se } i \neq j \end{cases} \quad \text{per il metodo di Jacobi,}$$

$$h_{ij} = \begin{cases} \lambda a_{ij} & \text{se } i \geq j \\ a_{ij} & \text{se } i < j \end{cases} \quad \text{per il metodo di Gauss-Seidel.}$$

Se $|\lambda| \geq 1$ si ha

$$|h_{ii}| = |\lambda|\ |a_{ii}| \quad \text{e} \quad |h_{ij}| \leq |\lambda|\ |a_{ij}| \quad \text{per} \quad i \neq j,$$

e quindi la matrice H ha le proprietà *a), b) c)* o *d)* della matrice A. In tal caso, per il teorema 2.41 la matrice H è non singolare e quindi un numero λ, tale che $|\lambda| \geq 1$, non può verificare la (27), cioè non può essere autovalore di $M^{-1}N$. Ne segue che gli autovalori di $M^{-1}N$ hanno modulo minore di 1, e per il teorema 5.6 i metodi di Jacobi e di Gauss-Seidel sono convergenti. ∎

5.13 Esempi. a) La matrice

$$A = \begin{bmatrix} \text{-}4 & \text{-}1 & 1 & 1 \\ 0 & \text{-}4 & \text{-}1 & 1 \\ \text{-}1 & \text{-}1 & 4 & 1 \\ 1 & \text{-}1 & 0 & 4 \end{bmatrix}$$

ha predominanza diagonale in senso stretto, sia per righe che per colonne. Per il teorema 5.12 le matrici di iterazione di Jacobi e di Gauss-Seidel hanno entrambe raggio spettrale minore di 1. È infatti

$$J = \frac{1}{4} \begin{bmatrix} 0 & \text{-}1 & 1 & 1 \\ 0 & 0 & \text{-}1 & 1 \\ 1 & 1 & 0 & \text{-}1 \\ \text{-}1 & 1 & 0 & 0 \end{bmatrix}, \qquad G = \frac{1}{16} \begin{bmatrix} 0 & \text{-}4 & 4 & 4 \\ 0 & 0 & \text{-}4 & 4 \\ 0 & \text{-}1 & 0 & \text{-}2 \\ 0 & 1 & \text{-}2 & 0 \end{bmatrix}.$$

Lo spettro degli autovalori di J è dato dall'unione degli spettri delle matrici (si veda l'esercizio 2.26)

$$\frac{1}{4}\left(\begin{bmatrix} 0 & \text{-}1 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ \text{-}1 & 1 \end{bmatrix}\right) \quad \text{e} \quad \frac{1}{4}\left(\begin{bmatrix} 0 & \text{-}1 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ \text{-}1 & 1 \end{bmatrix}\right).$$

Gli autovalori di J risultano

$$\lambda_1 = \lambda_2 = \frac{1}{4}, \quad \lambda_3 = \frac{-1 + \mathbf{i}\sqrt{2}}{4}, \quad \lambda_4 = \frac{-1 - \mathbf{i}\sqrt{2}}{4},$$

quindi $\rho(J) = \dfrac{\sqrt{3}}{4}$. Il polinomio caratteristico della G è

$$p(\lambda) = \lambda^4 - \frac{3}{64}\lambda^2 - \frac{\lambda}{256} = \lambda\Big(\lambda - \frac{1}{4}\Big)\Big(\lambda + \frac{1}{8}\Big)^2,$$

per cui gli autovalori di G sono

$$\lambda_1 = 0, \quad \lambda_2 = \frac{1}{4}, \quad \lambda_3 = \lambda_4 = -\frac{1}{8}$$

e il raggio spettrale è $\rho(G) = \dfrac{1}{4}$.

b) La matrice

$$A = \begin{bmatrix} \text{-4} & \text{-1} & 1 & 1 \\ 0 & \text{-4} & \text{-1} & \text{-3} \\ \text{-1} & \text{-1} & 4 & 1 \\ 1 & 3 & 0 & 4 \end{bmatrix}$$

ha predominanza diagonale ed è irriducibile. Per il teorema 5.12 le matrici di iterazione di Jacobi e di Gauss-Seidel hanno entrambe raggio spettrale minore di 1. È infatti

$$J = \frac{1}{4}\begin{bmatrix} 0 & \text{-1} & 1 & 1 \\ 0 & 0 & \text{-1} & \text{-3} \\ 1 & 1 & 0 & \text{-1} \\ \text{-1} & \text{-3} & 0 & 0 \end{bmatrix}, \qquad G = \frac{1}{16}\begin{bmatrix} 0 & \text{-4} & 4 & 4 \\ 0 & 0 & \text{-4} & \text{-12} \\ 0 & \text{-1} & 0 & \text{-6} \\ 0 & 1 & 2 & 8 \end{bmatrix}.$$

Procedendo come nel caso a), si trova che gli autovalori di J risultano

$$\lambda_1 = \frac{1}{4}, \quad \lambda_2 = -\frac{3}{4}, \quad \lambda_3 = \frac{1+\sqrt{2}}{4}, \quad \lambda_4 = \frac{1-\sqrt{2}}{4},$$

da cui $\rho(J) = \dfrac{3}{4}$, e che gli autovalori di G sono

$$\lambda_1 = 0, \quad \lambda_2 = \frac{1}{4}, \quad \lambda_3 = \lambda_4 = \frac{1}{8},$$

da cui $\rho(G) = \dfrac{1}{4}$.

c) Si noti che la sola condizione di predominanza diagonale non è sufficiente per la convergenza. Si consideri infatti la matrice

$$A = \begin{bmatrix} \text{-4} & \text{-1} & 1 & 1 \\ 0 & \text{-4} & 0 & \text{-4} \\ 1 & 1 & 4 & 1 \\ 0 & \text{-4} & 0 & 4 \end{bmatrix}$$

che ha predominanza diagonale ma è riducibile. Le due matrici di iterazione sono

$$J = \frac{1}{4}\begin{bmatrix} 0 & -1 & 1 & 1 \\ 0 & 0 & 0 & -4 \\ -1 & -1 & 0 & -1 \\ 0 & 4 & 0 & 0 \end{bmatrix}, \qquad G = \frac{1}{16}\begin{bmatrix} 0 & -4 & 4 & 4 \\ 0 & 0 & 0 & -16 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & -16 \end{bmatrix}.$$

Il polinomio caratteristico della J è

$$p(\lambda) = \lambda^4 + \frac{17}{16}\lambda^2 + \frac{1}{16} = (\lambda^2 + 1)\left(\lambda^2 + \frac{1}{16}\right),$$

per cui gli autovalori di J sono $\lambda_1 = \mathbf{i}$, $\lambda_2 = -\mathbf{i}$, $\lambda_3 = \dfrac{\mathbf{i}}{4}$, $\lambda_4 = -\dfrac{\mathbf{i}}{4}$ e quindi $\rho(J) = 1$. Il polinomio caratteristico della G è

$$p(\lambda) = \lambda^4 + \frac{17}{16}\lambda^3 + \frac{\lambda^2}{16} = \lambda^2\left(\lambda^2 + \frac{17}{16}\lambda + \frac{1}{16}\right),$$

per cui gli autovalori di G sono $\lambda_1 = \lambda_2 = 0$, $\lambda_3 = -\dfrac{1}{16}$, $\lambda_4 = -1$, e quindi $\rho(G) = 1$. In questo caso né il metodo di Jacobi né quello di Gauss-Seidel convergono. ∎

5.14 Teorema. *Sia A una matrice hermitiana non singolare con elementi principali reali e positivi. Allora il metodo di Gauss-Seidel è convergente se e solo se A è definita positiva.*

Dim. Essendo la matrice A hermitiana, è $C = B^H$ e quindi

$$A = D - B - B^H,$$

e la matrice di iterazione del metodo di Gauss-Seidel risulta

$$G = (D - B)^{-1}B^H = I - (D - B)^{-1}A. \tag{28}$$

Per dimostrare che il metodo è convergente, conviene prima dimostrare che la matrice $A - G^H AG$ è definita positiva. Posto per semplicità

$$F = (D - B)^{-1}A,$$

dalla (28) si ha $G = I - F$ e

$$\begin{aligned} A - G^H AG &= A - (I - F)^H A(I - F) = A - A + F^H A + AF - F^H AF \\ &= F^H(AF^{-1} + F^{-H}A - A)F = F^H(D - B + D - B^H - A)F \\ &= F^H DF. \end{aligned}$$

La matrice F è non singolare perché tali sono le due matrici $(D-B)^{-1}$ e A, ed essendo gli elementi di D positivi, la matrice $A - G^H AG$ risulta definita positiva. Infatti per ogni $\mathbf{x} \neq \mathbf{0}$ risulta

$$\mathbf{x}^H A\mathbf{x} - \mathbf{x}^H G^H AG\mathbf{x} = \mathbf{x}^H F^H DF\mathbf{x} > 0. \tag{29}$$

Si supponga ora che la matrice A sia definita positiva e si consideri un autovalore λ di G e un corrispondente autovettore $\mathbf{x}$. Dalla (29) si ha:

$$\mathbf{x}^H A\mathbf{x} - \lambda\overline{\lambda}\,\mathbf{x}^H A\mathbf{x} > 0,$$

e cioè

$$(1 - |\lambda|^2)\,\mathbf{x}^H A\mathbf{x} > 0. \tag{30}$$

Essendo A definita positiva, dalla (30) risulta $|\lambda| < 1$ e quindi, per il teorema 5.6, il metodo di Gauss-Seidel è convergente.

Viceversa, si supponga che il metodo sia convergente e si consideri il vettore $\mathbf{e}^{(k)} = \mathbf{x}^* - \mathbf{x}^{(k)}$. Per la (12) si ha che

$$\mathbf{e}^{(k)} = G\mathbf{e}^{(k-1)},$$

e, sostituendo nella (30) $\mathbf{e}^{(k-1)}$ al posto di $\mathbf{x}$, poiché la matrice $A - G^H AG$ è definita positiva, risulta:

$$[\mathbf{e}^{(k-1)}]^H A\mathbf{e}^{(k-1)} > [\mathbf{e}^{(k)}]^H A\mathbf{e}^{(k)}. \tag{31}$$

Se A non fosse definita positiva, allora esisterebbe un vettore $\mathbf{e}^{(0)} \neq \mathbf{0}$ per cui $[\mathbf{e}^{(0)}]^H A\mathbf{e}^{(0)} \leq 0$ e quindi la successione $[\mathbf{e}^{(k)}]^H A\mathbf{e}^{(k)}$, che per la (31) è monotona decrescente, non potrebbe convergere a zero, ciò che è assurdo perché il metodo di Gauss-Seidel è convergente, cioè

$$\lim_{k\to\infty} \mathbf{e}^{(k)} = \mathbf{0}. \qquad \blacksquare$$

Nella figura 5.5 sono sinteticamente rappresentate le classi delle matrici hermitiane, delle matrici definite positive e delle matrici con predominanza diagonale in senso stretto e la classe della matrici per cui il metodo di Gauss-Seidel è convergente.

Si può dimostrare (si veda l'esercizio 5.15) che per le matrici a predominanza diagonale in senso stretto vale la relazione $\|G\|_\infty \leq \|J\|_\infty < 1$. Però, anche se $\rho(G) \leq \|G\|_\infty$ e $\rho(J) \leq \|J\|_\infty$, non sempre ne segue che $\rho(G) \leq \rho(J)$, cioè non sempre per le matrici a predominanza diagonale in senso stretto il metodo di Gauss-Seidel è asintoticamente più veloce del metodo di Jacobi.

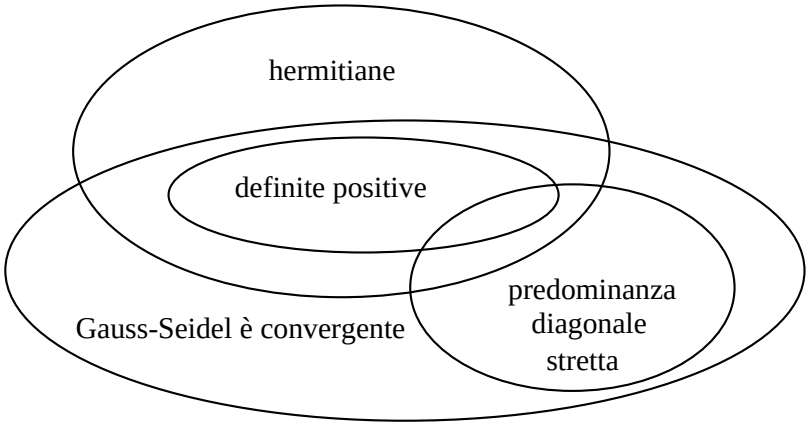

Fig. 5.5 - Classi di matrici per cui il metodo di Gauss-Seidel è convergente.

5.15 Esempio. Per il sistema $A\mathbf{x} = \mathbf{b}$, dove

$$A = \begin{bmatrix} 11 & -5 & -5 \\ 5 & 12 & 6 \\ 6 & -4 & 11 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 23 \\ 13 \end{bmatrix},$$

che ha la soluzione $\mathbf{x}^* = [1,\ 1,\ 1]^T$, risulta

$$J = \begin{bmatrix} 0 & \frac{5}{11} & \frac{5}{11} \\ -\frac{5}{12} & 0 & -\frac{1}{2} \\ -\frac{6}{11} & \frac{4}{11} & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 0 & \frac{5}{11} & \frac{5}{11} \\ 0 & -\frac{25}{132} & -\frac{91}{132} \\ 0 & -\frac{115}{363} & -\frac{181}{363} \end{bmatrix}$$

e $\rho(J) = 0.7917518$, $\rho(G) = 0.8362568$, $\|J\|_\infty = 0.9166667$, $\|G\|_\infty = 0.9090909$. Quindi $\rho(J) < \rho(G)$, mentre $\|G\|_\infty \leq \|J\|_\infty$. Il tasso asintotico di convergenza del metodo di Jacobi è maggiore di quello del metodo di Gauss-Seidel. Assumendo $\mathbf{x}^{(0)} = \mathbf{0}$, e usando il criterio di arresto espresso dalla (16) in norma ∞ con $\epsilon = 10^{-5}$, la successione ottenuta con il metodo di Jacobi si arresta alla 52-esima iterazione, mentre la successione ottenuta con il metodo di Gauss-Seidel si arresta alla 68-esima iterazione. ∎

Il seguente teorema individua un'ampia classe di matrici per cui è possibile stabilire una relazione più precisa fra le velocità di convergenza dei metodi di Gauss-Seidel e di Jacobi.

5.16 Teorema (di Stein-Rosenberg). *Sia $A \in \mathbf{R}^{n\times n}$. Se gli elementi principali di A sono non nulli e gli elementi della matrice di iterazione di Jacobi J sono non negativi, allora vale una e una sola delle seguenti relazioni:*

a) $\rho(G) = \rho(J) = 0$;
b) $\rho(G) < \rho(J) < 1$;
c) $\rho(G) = \rho(J) = 1$;
d) $\rho(G) > \rho(J) > 1$;

(Per la dimostrazione si veda [10]) . ∎

5.17 Esempi. a) Per la matrice

$$A = \begin{bmatrix} 6 & 0 & 0 \\ -7 & 9 & 0 \\ -4 & -1 & 8 \end{bmatrix},$$

si ha

$$J = \begin{bmatrix} 0 & 0 & 0 \\ \frac{7}{9} & 0 & 0 \\ \frac{1}{2} & \frac{1}{8} & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

e $\rho(J) = \rho(G) = 0$.

b) Per la matrice

$$A = \begin{bmatrix} 9 & -3 & -1 \\ -2 & 9 & 0 \\ -2 & 0 & 9 \end{bmatrix},$$

si ha

$$J = \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{9} \\ \frac{2}{9} & 0 & 0 \\ \frac{2}{9} & 0 & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{9} \\ 0 & \frac{2}{27} & \frac{2}{81} \\ 0 & \frac{2}{27} & \frac{2}{81} \end{bmatrix},$$

e $\rho(J) = 0.3142697,\ \rho(G) = 0.09876543$.

c) Per la matrice

$$A = \begin{bmatrix} 1 & 0 & 0 \\ -8 & 1 & -2 \\ -6 & -3 & 6 \end{bmatrix},$$

si ha

$$J = \begin{bmatrix} 0 & 0 & 0 \\ 8 & 0 & 2 \\ 1 & \frac{1}{2} & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 1 \end{bmatrix},$$

e $\rho(J) = \rho(G) = 1$.

d) Per la matrice

$$A = \begin{bmatrix} 8 & -6 & -8 \\ -6 & 7 & 0 \\ 0 & -8 & 7 \end{bmatrix},$$

si ha

$$J = \begin{bmatrix} 0 & \frac{3}{4} & 1 \\ \frac{6}{7} & 0 & 0 \\ 0 & \frac{8}{7} & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 0 & \frac{3}{4} & 1 \\ 0 & \frac{9}{14} & \frac{6}{7} \\ 0 & \frac{36}{49} & \frac{48}{49} \end{bmatrix},$$

e $\rho(J) = 1.206222,\ \rho(G) = 1.6224490$. ∎

Molte delle matrici che si ottengono risolvendo numericamente problemi differenziali di tipo ellittico hanno predominanza diagonale e soddisfano alle condizioni del teorema di Stein-Rosenberg: in tal caso è conveniente usare il metodo di Gauss-Seidel. Nel caso delle matrici tridiagonali è possibile stabilire esattamente di quanto il metodo di Gauss-Seidel è più veloce del metodo di Jacobi.

5.18 Teorema. *Sia $A \in \mathbf{C}^{n\times n}$ la matrice tridiagonale*

$$A = \begin{bmatrix} a_1 & c_1 & & & \\ b_1 & a_2 & c_2 & & \\ & b_2 & a_3 & \ddots & \\ & & \ddots & \ddots & c_{n-1} \\ & & & b_{n-1} & a_n \end{bmatrix},$$

in cui $a_i \neq 0$ per $i = 1, \ldots, n$. Valgono le seguenti relazioni:

a) se μ è autovalore di J, allora μ^2 è autovalore di G;

b) se λ è autovalore non nullo di G, allora le radici quadrate di λ sono autovalori di J.

Dim. Sia $S \in \mathbf{C}^{n\times n}$ la matrice diagonale

$$S = \begin{bmatrix} 1 & & & & \\ & \alpha & & & \\ & & \alpha^2 & & \\ & & & \ddots & \\ & & & & \alpha^{n-1} \end{bmatrix},$$

in cui $\alpha \in \mathbf{C}$ è una costante non nulla. Si ha

$$SJS^{-1} = \begin{bmatrix} 0 & -\frac{c_1}{\alpha a_1} & & & \\ -\alpha\,\frac{b_1}{a_2} & 0 & -\frac{c_2}{\alpha a_2} & & \\ & -\alpha\,\frac{b_2}{a_3} & 0 & \ddots & \\ & & \ddots & \ddots & -\frac{c_{n-1}}{\alpha a_{n-1}} \\ & & & -\alpha\,\frac{b_{n-1}}{a_n} & 0 \end{bmatrix}$$

$$= \alpha\, D^{-1}B + \frac{1}{\alpha}\, D^{-1}C,$$

e quindi le matrici J e $\alpha\ D^{-1}B + \frac{1}{\alpha}\ D^{-1}C$ hanno lo stesso polinomio caratteristico qualunque sia $\alpha \neq 0$, cioè se μ è autovalore di J allora

$$\det(\alpha^2\ D^{-1}B + D^{-1}C - \alpha\mu I) = 0, \tag{32}$$

per ogni $\alpha \neq 0$ e viceversa, se esiste un $\alpha \neq 0$ per cui μ soddisfa la (32), allora μ è autovalore di J. Si ha

$$\begin{aligned} G - \lambda I &= (D-B)^{-1}C - \lambda I = (D-B)^{-1}[C - \lambda(D-B)] \\ &= (I - D^{-1}B)^{-1}(\lambda D^{-1}B + D^{-1}C - \lambda I), \end{aligned}$$

e quindi, se μ è autovalore di J, posto $\lambda = \alpha\mu$ e $\alpha^2 = \lambda$, dalla (32) segue che $\det(G - \lambda I) = 0$, per cui i λ tali che $\mu^2 = \lambda$ sono autovalori di G. Viceversa, se $\lambda \neq 0$ è un autovalore di G, siano $\alpha \neq 0$ e μ tali che $\alpha^2 = \lambda$ e $\alpha\mu = \lambda$. Allora

$$0 = \det(\lambda D^{-1}B + D^{-1}C - \lambda I) = \det(\alpha^2 D^{-1}B + D^{-1}C - \alpha\mu I),$$

e per la (32) μ è autovalore di J. ∎

Quindi per le matrici tridiagonali il metodo di Gauss-Seidel è convergente se e solo se lo è il metodo di Jacobi e vale

$$\rho(G) = \rho^2(J).$$

Perciò il tasso asintotico di convergenza del metodo di Gauss-Seidel è doppio di quello del metodo di Jacobi e, asintoticamente, sono necessarie metà iterazioni del metodo di Gauss-Seidel per ottenere la stessa precisione che con il metodo di Jacobi.

5.19 Esempio. Sia $A \in \mathbf{R}^{6\times 6}$, la matrice tridiagonale

$$a_{ij} = \begin{cases} 2 & \text{per } i = j, \\ \text{-1} & \text{per } |i-j| = 1, \\ 0 & \text{altrimenti.} \end{cases}$$

Essendo A simmetrica, si ha

$$A = 2I - U - U^T,$$

dove U è la matrice

$$u_{ij} = \begin{cases} 1 & \text{per } j = i-1, \\ 0 & \text{altrimenti,} \end{cases}$$

e quindi

$$J = \frac{1}{2}(U + U^T) \quad \text{e} \quad G = (2I - U)^{-1}U^T = \left[\sum_{i=0}^{5} \left(\frac{1}{2}\right)^{i+1} U^i\right] U^T$$

(si veda l'esercizio 1.52). I rispettivi polinomi caratteristici sono dati da

$$p_J(\mu) = \mu^6 - \frac{5}{4}\mu^4 + \frac{3}{8}\mu^2 - \frac{1}{64}$$
$$p_G(\lambda) = \lambda^3 \left(\lambda^3 - \frac{5}{4}\lambda^2 + \frac{3}{8}\lambda - \frac{1}{64}\right),$$

da cui si ricavano gli autovalori (dall'esercizio 2.40 si ha che gli autovalori di J sono dati da $\pm\cos\frac{\pi}{7}$, $\pm\cos\frac{2\pi}{7}$, $\pm\cos\frac{3\pi}{7}$)

$$\mu_1 = -\mu_6 = 0.9009688,\ \mu_2 = -\mu_5 = 0.6234898,\ \mu_3 = -\mu_4 = 0.2225209,$$
$$\lambda_1 = \lambda_2 = \lambda_3 = 0,$$
$$\lambda_4 = \mu_1^2 = 0.8117447,\ \lambda_5 = \mu_2^2 = 0.3887395,\ \lambda_6 = \mu_3^2 = 0.04951555.$$

Risulta pertanto che

$$\rho(J) = 0.9009688 \quad \text{e} \quad \rho(G) = \rho^2(J) = 0.8117447. \qquad \blacksquare$$

Nei casi particolari $n = 5, 10, 20$ risulta

	$n = 5$	$n = 10$	$n = 20$
$\rho(J)$	0.8660254	0.9594930	0.9888308
$\rho(G)$	0.7500000	0.9206268	0.9777864
$\rho[H(\omega_o)]$	0.3333333	0.5603879	0.7405800
$\rho(J_B)$	0.7637079	0.9221398	0.9779084
$\rho(G_B)$	0.5832498	0.8503418	0.9563048
$\rho[H_B(\omega_o)]$	0.2153903	0.4421100	0.6542134

Per n grande si possono dare le seguenti valutazioni approssimate

$$\rho(J) \approx 1 - \frac{\pi^2}{2(n+1)^2}, \quad \rho(G) \approx 1 - \frac{\pi^2}{(n+1)^2}, \quad \rho[H(\omega_o)] \approx 1 - 2\frac{\pi}{n+1},$$

$$\rho(J_B) \approx 1 - \frac{\pi^2}{(n+1)^2}, \quad \rho(G_B) \approx 1 - \frac{2\pi^2}{(n+1)^2}, \quad \rho[H_B(\omega_o)] \approx 1 - 2\sqrt{2}\frac{\pi}{n+1}.$$

Perciò asintoticamente per ottenere un risultato con la stessa precisione il metodo di Jacobi richiede un numero di iterazioni doppio di quello richiesto dai metodi di Jacobi a blocchi e di Gauss-Seidel e pari a quattro volte il numero di iterazioni richiesto dal metodo di Gauss-Seidel a blocchi. Applicando il metodo di rilassamento la riduzione del numero di iterazioni rispetto al metodo di Jacobi è proporzionale ad n. ∎

7. Metodo del gradiente coniugato

Se la matrice A è reale e definita positiva, il sistema lineare $A\mathbf{x} = \mathbf{b}$ può essere risolto con il metodo del *gradiente coniugato*. Questo metodo, anche se in teoria è un metodo diretto, in quanto viene costruita una successione $\{\mathbf{x}^{(k)}\}_{k=0,1,\ldots}$ di vettori tali che $\mathbf{x}^{(m)} = \mathbf{x}^* = A^{-1}\mathbf{b}$, per un qualche indice $m \leq n$, in pratica però, per la presenza degli errori di arrotondamento, non termina all'm-esimo passo e viene utilizzato come metodo iterativo. In molti casi significativi il numero di iterazioni che occorrono per raggiungere la precisione richiesta è di gran lunga inferiore alla dimensione del sistema, e ciò rende il metodo molto conveniente per trattare problemi di grosse dimensioni, anche rispetto al metodo di rilassamento, poiché non richiede, fra l'altro, la determinazione preliminare di alcun parametro. Del metodo del gradiente coniugato esistono diverse varianti, che all'atto pratico generano successioni confrontabili. L'algoritmo che viene descritto in questo paragrafo è quello originario dovuto a Hestenes e Stiefel [6].

Sia $A \in \mathbf{R}^{n\times n}$, definita positiva, e $\mathbf{b} \in \mathbf{R}^n$, e si consideri il problema di minimizzare su $\mathbf{R}^n$ il funzionale

$$\Phi(\mathbf{x}) = \tfrac{1}{2}\ \mathbf{x}^T A\mathbf{x} - \mathbf{b}^T\mathbf{x}. \tag{44}$$

Tale problema ha una e una sola soluzione che, come si vedrà nel capitolo 7, è data da $\mathbf{x}^* = A^{-1}\mathbf{b}$. Quindi per calcolare la soluzione del sistema lineare $A\mathbf{x} = \mathbf{b}$ possono essere utilizzati dei metodi che minimizzano il funzionale (44). I metodi del *gradiente* sono metodi iterativi che minimizzano il funzionale (44) sfruttando il gradiente negativo di $\Phi(\mathbf{x})$, cioè il vettore

$$-\nabla\Phi(\mathbf{x}) = -\Big[\frac{\partial\Phi}{\partial x_1}\ (\mathbf{x}), \frac{\partial\Phi}{\partial x_2}\ (\mathbf{x}), \ldots, \frac{\partial\Phi}{\partial x_n}\ (\mathbf{x})\Big]^T = \mathbf{b} - A\mathbf{x} = \mathbf{r}(\mathbf{x}).$$

Il vettore $\mathbf{r} = \mathbf{r}(\mathbf{x})$ viene detto *residuo* del sistema $A\mathbf{x} = \mathbf{b}$.

Un metodo del gradiente procede nel modo seguente (per semplificare le notazioni, si scriverà in basso l'indice di iterazione): sia al k-esimo passo $\mathbf{x}_k \neq \mathbf{x}^*$, scelto un vettore direzione $\mathbf{p}_k \neq \mathbf{0}$ di decrescita per $\Phi(\mathbf{x})$, cioè tale che $\mathbf{p}_k^T\nabla\Phi(\mathbf{x}_k) < 0$, si determina il punto $\mathbf{x}_{k+1}$ di minimo del funzionale (44) sulla retta passante per $\mathbf{x}_k$ e di direzione $\mathbf{p}_k$:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{p}_k, \tag{45}$$

dove $\alpha_k \in \mathbf{R}$ è tale che

$$\Phi(\mathbf{x}_{k+1}) = \min_{\alpha\in\mathbf{R}} \Phi(\mathbf{x}_k + \alpha\mathbf{p}_k).$$

Derivando la funzione $\Phi(\mathbf{x}_k + \alpha\mathbf{p}_k)$ rispetto ad α si ottiene

$$\frac{\partial\Phi}{\partial\alpha} = (\mathbf{x}_k + \alpha\mathbf{p}_k)^T A\mathbf{p}_k - \mathbf{b}^T\mathbf{p}_k,$$

da cui, imponendo che $\dfrac{\partial\Phi}{\partial\alpha} = 0$, si ricava

$$\alpha_k = \frac{(\mathbf{b} - A\mathbf{x}_k)^T\mathbf{p}_k}{\mathbf{p}_k^T A\mathbf{p}_k} = \frac{\mathbf{r}_k^T\mathbf{p}_k}{\mathbf{p}_k^T A\mathbf{p}_k}, \tag{46}$$

in cui $\mathbf{r}_k = \mathbf{r}(\mathbf{x}_k)$ è il residuo in $\mathbf{x}_k$. Poiché

$$\mathbf{r}_k^T\mathbf{p}_k = -\mathbf{p}_k^T\nabla\Phi(\mathbf{x}_k) > 0,$$

ne segue che $\alpha_k > 0$. Così procedendo si ottiene una successione $\{\mathbf{x}_k\}$ che converge al punto $\mathbf{x}^*$.

Dalla (45) segue per $k = 0, 1, \ldots$

$$\mathbf{b} - A\mathbf{x}_{k+1} = \mathbf{b} - A\mathbf{x}_k - \alpha_k A\mathbf{p}_k,$$

e quindi

$$\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k A\mathbf{p}_k, \tag{47}$$

da cui per la (46)

$$\mathbf{r}_{k+1}^T \mathbf{p}_k = (\mathbf{r}_k - \alpha_k A\mathbf{p}_k)^T \mathbf{p}_k = \mathbf{r}_k^T \mathbf{p}_k - \alpha_k \mathbf{p}_k^T A\mathbf{p}_k = 0, \tag{48}$$

e quindi ad ogni passo il residuo $\mathbf{r}_{k+1}$ è ortogonale al vettore direzione $\mathbf{p}_k$ del passo precedente.

I diversi metodi del gradiente si distinguono per la diversa scelta del vettore $\mathbf{p}_k$: un metodo classico è quello dello *steepest descent*, in cui si sceglie $\mathbf{p}_k = \mathbf{r}_k = -\nabla\Phi(\mathbf{x}_k)$, cioè ad ogni passo il vettore $\mathbf{p}_k$ coincide con la direzione di massima pendenza per $\Phi(\mathbf{x})$. Questa strategia, anche se in ciascun punto $\mathbf{x}_k$ sfrutta la direzione della massima pendenza, può non essere la migliore, in particolare quando la matrice A è mal condizionata.

Nella figura 5.11, ad esempio, è illustrato il comportamento del metodo dello steepest descent nel caso di una matrice A di ordine 2. In ogni punto $\mathbf{x}_k$ si individua nel piano $\mathbf{R}^2$ la direzione $\mathbf{p}_k$, lungo la quale il funzionale $\Phi(\mathbf{x})$ decresce con la massima pendenza: il punto $\mathbf{x}_{k+1}$ è quello in cui il funzionale $\Phi(\mathbf{x})$ ha il valore minimo e in $\mathbf{x}_{k+1}$ la direzione $\mathbf{p}_k$ è tangente alla curva di livello $\Phi(\mathbf{x}) = \Phi(\mathbf{x}_{k+1})$. Il nuovo vettore direzione $\mathbf{p}_{k+1}$ è ortogonale al precedente vettore $\mathbf{p}_k$: infatti dalla (48) risulta

$$\mathbf{p}_{k+1}^T \mathbf{p}_k = 0, \quad \text{per} \quad k = 0, 1, \ldots \ .$$

La velocità di convergenza di questo procedimento dipende dalla eccentricità delle ellissi che rappresentano le curve di livello $\Phi(\mathbf{x}) = c$, dove c è una costante. L'eccentricità è tanto maggiore quanto maggiore è il rapporto λ_1/λ_2 degli autovalori λ_1 e λ_2, con $\lambda_1 > \lambda_2$, della matrice A, e quindi quanto più la matrice A è mal condizionata.

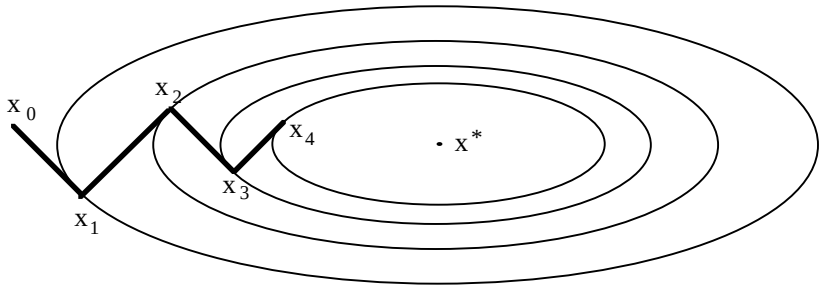

Fig. 5.11 - Il metodo dello steepest descent.

In generale se $\lambda_{\max}$ e $\lambda_{\min}$ sono il massimo e il minimo autovalore della matrice A di ordine n, si può dimostrare [3] che, indicato con $\mathbf{e}_k = \mathbf{x}^* - \mathbf{x}_k$ l'errore al k-esimo passo risulta (si veda l'esercizio 5.28)

$$\mathbf{e}_{k+1}^T A \mathbf{e}_{k+1} \leq \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}\right)^2 \mathbf{e}_k^T A \mathbf{e}_k.$$

Introducendo la norma vettoriale (si veda l'esercizio 3.6)

$$\|\mathbf{x}\|_A = \sqrt{\mathbf{x}^T A \mathbf{x}}, \tag{49}$$

e notando che $\mu_2(A) = \dfrac{\lambda_{\max}}{\lambda_{\min}}$, si ottiene la limitazione dell'errore al k-esimo passo per il metodo dello steepest descent

$$\|\mathbf{e}_k\|_A \leq \left(\frac{\mu_2(A) - 1}{\mu_2(A) + 1}\right)^k \|\mathbf{e}_0\|_A. \tag{50}$$

È possibile ottenere una migliore strategia per la minimizzazione del funzionale $\Phi(\mathbf{x})$, con una scelta di $\mathbf{p}_k$ che tiene conto anche delle direzioni $\mathbf{p}_j,\ j = 1, 2, \ldots, k-1$, calcolate ai passi precedenti. Un metodo che utilizza questa strategia è il metodo del *gradiente coniugato*, in cui il vettore direzione $\mathbf{p}_k$ viene scelto nel modo seguente

$$\mathbf{p}_k = \begin{cases} \mathbf{r}_0 & \text{se } k = 0, \\ \\ \mathbf{r}_k + \beta_k \mathbf{p}_{k-1} & \text{se } k \geq 1, \end{cases} \tag{51}$$

dove β_k è tale che

$$\mathbf{p}_k^T A \mathbf{p}_{k-1} = 0. \tag{52}$$

Il vettore $\mathbf{p}_k$ viene detto A-*coniugato* con il vettore $\mathbf{p}_{k-1}$. Sostituendo nella (52) l'espressione di $\mathbf{p}_k$ data dalla (51), si ricava

$$\beta_k = -\frac{\mathbf{r}_k^T A \mathbf{p}_{k-1}}{\mathbf{p}_{k-1}^T A \mathbf{p}_{k-1}}, \quad k \geq 1. \tag{53}$$

La direzione $\mathbf{p}_k$ così scelta è una direzione di decrescita del funzionale $\Phi(\mathbf{x})$. Si ha infatti da (51) e (48)

$$-\mathbf{p}_k^T \nabla \Phi(\mathbf{x}_k) = \mathbf{p}_k^T \mathbf{r}_k = \mathbf{r}_k^T \mathbf{r}_k + \beta_k \mathbf{p}_{k-1}^T \mathbf{r}_k = \mathbf{r}_k^T \mathbf{r}_k > 0 \tag{54}$$

se $\mathbf{r}_k \neq \mathbf{0}$, cioè $\mathbf{x}_k \neq \mathbf{x}^*$.

Sostituendo la (54) nella (46) si ha che per il metodo del gradiente coniugato

$$\alpha_k = \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{p}_k^T A \mathbf{p}_k} \ . \tag{55}$$

Da (51) e (48) si ha

$$\mathbf{r}_k^T \mathbf{r}_{k-1} = \mathbf{r}_k^T \mathbf{p}_{k-1} - \beta_{k-1} \mathbf{r}_k^T \mathbf{p}_{k-2} = -\beta_{k-1} \mathbf{r}_k^T \mathbf{p}_{k-2},$$

e poiché per la (47), la (48) e la (52) è

$$\mathbf{r}_k^T \mathbf{p}_{k-2} = \mathbf{r}_{k-1}^T \mathbf{p}_{k-2} - \alpha_{k-1} \mathbf{p}_{k-1}^T A \mathbf{p}_{k-2} = 0,$$

ne segue che

$$\mathbf{r}_k^T \mathbf{r}_{k-1} = 0, \tag{56}$$

cioè ogni residuo è ortogonale al precedente. Inoltre da (51), (56) e (54) si ha

$$\mathbf{p}_k^T \mathbf{r}_{k-1} = \mathbf{r}_k^T \mathbf{r}_{k-1} + \beta_k \mathbf{p}_{k-1}^T \mathbf{r}_{k-1} = \beta_k \mathbf{p}_{k-1}^T \mathbf{r}_{k-1} = \beta_k \mathbf{r}_{k-1}^T \mathbf{r}_{k-1},$$

e da (47), (52) e (54)

$$\mathbf{p}_k^T \mathbf{r}_{k-1} = \mathbf{p}_k^T \mathbf{r}_k + \alpha_{k-1} \mathbf{p}_k^T A \mathbf{p}_{k-1} = \mathbf{p}_k^T \mathbf{r}_k = \mathbf{r}_k^T \mathbf{r}_k,$$

da cui si ottiene un'altra relazione per β_k

$$\beta_k = \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{r}_{k-1}^T \mathbf{r}_{k-1}} \ . \tag{57}$$

Indicato con S_k lo spazio generato dai k vettori $\mathbf{p}_0, \ldots, \mathbf{p}_{k-1}$, si può dimostrare (si veda l'esercizio 5.33) che le direzioni $\mathbf{p}_i$, $i = 0, \ldots, k-1$, ottenute con la (51) sono tali che

$$\Phi(\mathbf{x}_k) = \min_{\mathbf{x} \in S_k} \Phi(\mathbf{x}),$$

e quindi il metodo del gradiente coniugato determina la soluzione $\mathbf{x}^*$ in al più n passi, cioè esiste un $m \leq n$ tale che

$$\mathbf{r}_m = \mathbf{0}. \tag{58}$$

Questa proprietà si può ricavare direttamente dal seguente teorema.

5.32 Teorema. *Siano* $\mathbf{r}_0 \neq \mathbf{0}$ *ed* $h \geq 1$ *tale che* $\mathbf{r}_k \neq \mathbf{0}$ *per ogni* $k \leq h$. *Allora*

$$\left.\begin{aligned} \mathbf{r}_k^T \mathbf{r}_j &= 0 \\ \mathbf{p}_k^T A \mathbf{p}_j &= 0 \end{aligned}\right\} \quad \text{per } k \neq j,\ k, j = 0, \ldots, h, \tag{59}$$

cioè i primi h *residui costituiscono un insieme di vettori ortogonali e i vettori* $\mathbf{p}_k$ *costituiscono un insieme di vettori* A*-coniugati.*

Dim. Si procede per induzione su h.
Per $h = 1$, la (59) vale per $k = 1$ e $j = 0$, essendo da (56) e (52)

$$\mathbf{r}_1^T \mathbf{r}_0 = 0, \quad \text{e} \quad \mathbf{p}_1^T A \mathbf{p}_0 = 0.$$

Per $h > 1$, si suppone che valgano le (59) e si dimostra che

$$\left.\begin{aligned} \mathbf{r}_{h+1}^T \mathbf{r}_j &= 0 \\ \mathbf{p}_{h+1}^T A \mathbf{p}_j &= 0 \end{aligned}\right\} \quad \text{per} \quad j = 0, \ldots, h-1,$$

in quanto per $j = h$ l'ortogonalità dei residui è già stata dimostrata con la (56) e $\mathbf{p}_{h+1}$ è A-coniugato con $\mathbf{p}_h$ per la (52). Si ha dalla (47) per $j = 0, \ldots, h-1$

$$\mathbf{r}_{h+1}^T \mathbf{r}_j = \mathbf{r}_h^T \mathbf{r}_j - \alpha_h \mathbf{p}_h^T A \mathbf{r}_j = -\alpha_h \mathbf{p}_h^T A \mathbf{r}_j$$

per l'ipotesi induttiva, e per (51) è

$$\mathbf{p}_h^T A \mathbf{r}_j = \mathbf{p}_h^T A \mathbf{p}_j - \beta_j \mathbf{p}_h^T A \mathbf{p}_{j-1} = 0$$

per l'ipotesi induttiva. Quindi

$$\mathbf{r}_{h+1}^T \mathbf{r}_j = 0. \tag{60}$$

Inoltre per $j = 0, \ldots, h-1$ dalla (47) è

$$A \mathbf{p}_j = \frac{1}{\alpha_j}(\mathbf{r}_j - \mathbf{r}_{j+1}), \tag{61}$$

e quindi dalla (51) e dall'ipotesi induttiva

$$\mathbf{p}_{h+1}^T A \mathbf{p}_j = \mathbf{r}_{h+1}^T A \mathbf{p}_j + \beta_{h+1} \mathbf{p}_h^T A \mathbf{p}_j = \mathbf{r}_{h+1}^T A \mathbf{p}_j,$$

e da (61)

$$\mathbf{p}_{h+1}^T A \mathbf{p}_j = \frac{1}{\alpha_j}(\mathbf{r}_{h+1}^T \mathbf{r}_j - \mathbf{r}_{h+1}^T \mathbf{r}_{j+1}) = -\frac{1}{\alpha_j} \mathbf{r}_{h+1}^T \mathbf{r}_{j+1} = 0$$

per la (60) se $j = 0, \ldots, h-2$, e per la (56) se $j = h-1$. ∎

Dal teorema 5.32 segue che, poiché l'insieme dei primi h residui è formato da vettori ortogonali, non vi possono essere più di n vettori $\mathbf{r}_k \neq 0$ e quindi esiste un $m \leq n$ tale che vale la (58). Inoltre $\mathbf{r}_k$ appartiene al sottospazio generato dai vettori $\mathbf{r}_0, A\mathbf{r}_0, A^2\mathbf{r}_0, \ldots, A^k\mathbf{r}_0$, come si può vedere per induzione utilizzando la (47) e la (51). Se la matrice A ha al più s autovalori distinti, $s \leq n$, allora $\mathbf{r}_m = 0$ per qualche $m \leq s$. Infatti lo spazio generato da $\mathbf{r}_0, A\mathbf{r}_0, \ldots, A^{n-1}\mathbf{r}_0$, ha dimensione al più s (si veda l'esercizio 2.21) e quindi in esso non può esistere un vettore $\mathbf{r}_s$ non nullo che sia ortogonale a $\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_{s-1}$. Ne segue che $\mathbf{r}_s = 0$.

Riassumendo, il metodo del gradiente coniugato può essere così descritto:

1. $k = 0$, $\mathbf{x}_0$ arbitrario, $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$
2. se $\mathbf{r}_k = \mathbf{0}$, stop
3. altrimenti si calcoli

$$\beta_k = \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{r}_{k-1}^T \mathbf{r}_{k-1}} \quad (\beta_0 = 0 \text{ per } k = 0),$$

$$\mathbf{p}_k = \mathbf{r}_k + \beta_k \mathbf{p}_{k-1} \quad (\mathbf{p}_0 = \mathbf{r}_0 \text{ per } k = 0),$$

$$\alpha_k = \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{p}_k^T A \mathbf{p}_k},$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k,$$

$$\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k A \mathbf{p}_k,$$

$k = k + 1$ e si vada al punto 2.

In questo procedimento per α_k e β_k si usano la (55) e la (57) che hanno un minor costo computazionale rispetto alle (46) e (53).

Per gli errori di arrotondamento il metodo può non terminare in n passi e viene di solito usato come metodo iterativo. Il calcolo si arresta quando il residuo $\mathbf{r}_k$ diventa sufficientemente piccolo. Poiché la quantità $\mathbf{r}_k^T \mathbf{r}_k$ viene già calcolata nel corso dell'algoritmo, conviene usare la seguente condizione di arresto:

$$\|\mathbf{r}_k\|_2 < \epsilon \, \|\mathbf{b}\|_2. \tag{62}$$

Un aspetto delicato dell'algoritmo dal punto di vista della stabilità è il calcolo del residuo $\mathbf{r}_{k+1}$ con la relazione ricorrente (47): allo scopo di contenere gli errori che si accumulano nel calcolo di $\mathbf{r}_{k+1}$, è opportuno dopo un certo numero di passi calcolare il residuo con la relazione $\mathbf{r}_{k+1} = \mathbf{b} - A\mathbf{x}_{k+1}$. In [11] si suggerisce di fare questa correzione ogni m passi, dove m è dell'ordine di $\sqrt{n}$.

L'operazione che presenta in generale un maggior costo computazionale è quella relativa alla moltiplicazione della matrice A per il vettore $\mathbf{p}_k$. Quindi la complessità del metodo, se questo richiedesse un numero di passi dell'ordine di n e se la matrice A non fosse sparsa, sarebbe dell'ordine di n^3, superiore a quella del metodo di Cholesky. Però nella risoluzione di sistemi di equazioni lineari che scaturiscono dalla discretizzazione di problemi differenziali, il numero di iterazioni richieste è di solito molto inferiore alla dimensione della matrice e la matrice A è sparsa (si veda a questo proposito l'esempio 5.36).

5.33 Esempio. Si applichi il metodo del gradiente coniugato al sistema lineare $A\mathbf{x} = \mathbf{b}$, dove

$$A = \begin{bmatrix} 7 & 4 & -7 \\ 4 & 5 & -3 \\ -7 & -3 & 8 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 4 \\ 6 \\ -2 \end{bmatrix},$$

la cui soluzione è $\mathbf{x}^* = [1, 1, 1]^T$. Tale sistema è stato già studiato nell'esempio 5.27. Con il metodo del gradiente coniugato, partendo dal vettore iniziale $\mathbf{x}_0 = \mathbf{0}$, si ottiene una successione dei residui $\mathbf{r}_k$ le cui norme sono

$$\|\mathbf{r}_0\|_2 = 7.483314$$
$$\|\mathbf{r}_1\|_2 = 3.712894$$
$$\|\mathbf{r}_2\|_2 = 0.2249498$$
$$\|\mathbf{r}_3\|_2 = 0.1500715\ 10^{-3}$$
$$\|\mathbf{r}_4\|_2 = 0.6938835\ 10^{-5}.$$

Se $\epsilon = 10^{-5}$, la condizione di arresto espressa dalla (62) risulta verificata dopo 4 iterazioni. ∎

Una limitazione dell'errore al k-esimo passo del metodo del gradiente coniugato, è data in [4] per la norma definita in (49)

$$\|\mathbf{e}_k\|_A \leq 2 \left(\frac{\sqrt{\mu_2(A)} - 1}{\sqrt{\mu_2(A)} + 1} \right)^k \|\mathbf{e}_0\|_A. \tag{63}$$

Si confronti questa relazione con la (50) relativa al metodo dello steepest descent.

Dalla relazione (63) segue che se $\mu_2(A)$ è un numero prossimo ad 1, sono sufficienti pochi passi per ottenere una buona riduzione dell'errore iniziale, mentre se $\mu_2(A)$ è elevato, è possibile che siano necessari fino ad n

passi per ottenere una approssimazione accettabile della soluzione, e se n è molto grande è possibile che non si riesca ad ottenere una approssimazione accettabile per la presenza degli errori di arrotondamento.

5.34 Esempio. Sia $n = 1000$ e siano

$$A = \begin{bmatrix} \alpha & 1 & & \\ 1 & \alpha & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & \alpha \end{bmatrix} \in \mathbf{R}^{n\times n},\ \alpha \geq 2, \quad \text{e} \quad \mathbf{b} = \begin{bmatrix} \alpha+1 \\ \alpha+2 \\ \vdots \\ \alpha+2 \\ \alpha+1 \end{bmatrix}.$$

La soluzione del sistema lineare $A\mathbf{x} = \mathbf{b}$ è data da $\mathbf{x}^* = [1, 1, \ldots, 1]^T$. Se $\mathbf{x}_0 = \mathbf{0}$ risulta

$$\|\mathbf{e}_0\|_A = \sqrt{\mathbf{x}^{*T} A \mathbf{x}^*} = \sqrt{n(\alpha+2) - 2}.$$

Gli autovalori della matrice A sono

$$\lambda_i = \alpha + 2\cos\frac{i\pi}{n+1}, \quad i = 1, 2, \ldots, n,$$

(si veda l'esercizio 2.40), per cui

$$\mu_2(A) = \frac{\alpha + 2\cos\dfrac{\pi}{n+1}}{\alpha - 2\cos\dfrac{\pi}{n+1}}.$$

Se $\alpha = 20$, allora $\mu_2(A) = 1.105541$, $\|\mathbf{e}_0\|_A = 148.3037$ e dalla (63) si ha

$$\|\mathbf{e}_k\|_A = \sigma^k \|\mathbf{e}_0\|_A, \quad \text{dove} \quad \sigma = 0.05012535,$$

e bastano 7 iterazioni del metodo del gradiente coniugato per ridurre l'errore di un fattore dell'ordine di 10^{-6}. Se invece $\alpha = 2.5$, allora $\mu_2(A) = 2.999968$, $\|\mathbf{e}_0\|_A = 67.03731$ e dalla (63) si ha

$$\|\mathbf{e}_k\|_A = \sigma^k \|\mathbf{e}_0\|_A, \quad \text{dove} \quad \sigma = 0.4999960,$$

e bastano 26 iterazioni del metodo del gradiente coniugato per ridurre l'errore di un fattore dell'ordine di 10^{-6}. In pratica, a causa degli errori di arrotondamento, non è possibile ottenere una soluzione affetta da un errore dell'ordine di 10^{-6} se si usa la relazione ricorrente (47) per il calcolo del residuo $\mathbf{r}_{k+1}$: infatti per $\alpha = 20$ dopo 6 iterazioni l'errore è dell'ordine di $\frac{1}{2}10^{-5}$ e non diminuisce più, mentre per $\alpha = 2.5$ dopo 19 iterazioni l'errore

è dell'ordine di $\frac{1}{2}10^{-4}$ e non diminuisce più. Se invece il residuo viene calcolato con la relazione $\mathbf{r}_{k+1} = \mathbf{b} - A\mathbf{x}_{k+1}$, per $\alpha = 20$ l'errore si riduce dell'ordine di $\frac{1}{2}10^{-6}$ in 5 passi per $\alpha = 20$ e in 18 passi per $\alpha = 2.5$, come è illustrato nella figura 5.12, in cui con i quadratini vuoti sono indicati gli errori generati nel caso $\alpha = 20$ e con i quadratini pieni gli errori per $\alpha = 2.5$.

■

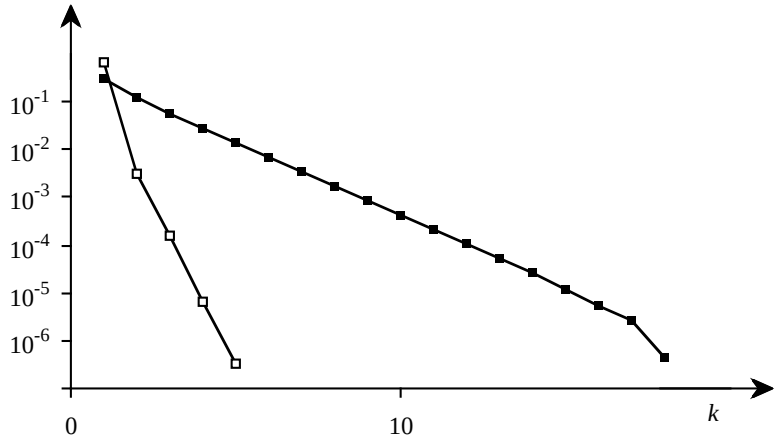

Fig. 5.12 - Errori generati dal metodo del gradiente coniugato applicato al sistema dell'esempio 5.34.

In alcuni casi è possibile ottenere una migliore convergenza utilizzando tecniche di *precondizionamento*, che trasformano il problema originale in un problema equivalente meglio condizionato.

Le tecniche di precondizionamento, che hanno avuto recentemente consistenti sviluppi, consistono essenzialmente nell'individuare una matrice C reale e non singolare, in modo che la matrice

$$B = C^{-1}AC^{-T}$$

sia tale che $\mu_2(B) < \mu_2(A)$. Il sistema a cui il metodo viene applicato è

$$B\mathbf{y} = \mathbf{c},$$

dove $\mathbf{y} = C^T\mathbf{x}$ e $\mathbf{c} = C^{-1}\mathbf{b}$. Naturalmente, per non aumentare eccessivamente il costo computazionale, la matrice C deve essere scelta di forma opportuna. La matrice

$$M = CC^T,$$

reale e definita positiva, è detta *precondizionatore* ed è quella che interviene nel seguente algoritmo del metodo del *gradiente coniugato con precondizionamento*. Infatti, poiché il residuo $\mathbf{s}_k$ del punto $\mathbf{y}_k$ nel metodo con

precondizionamento è legato al residuo $\mathbf{r}_k$ del punto $\mathbf{x}_k$ nel metodo senza precondizionamento dalla relazione

$$\mathbf{s}_k = C^{-1}\mathbf{r}_k,$$

allora

$$\mathbf{s}_k^T\mathbf{s}_k = \mathbf{r}_k^T M^{-1}\mathbf{r}_k.$$

Definito il vettore $\mathbf{z}_k$ tale che

$$M\mathbf{z}_k = \mathbf{r}_k,$$

si ha

$$\mathbf{s}_k^T\mathbf{s}_k = \mathbf{z}_k^T\mathbf{r}_k.$$

L'algoritmo del metodo del gradiente coniugato con precondizionamento risulta allora il seguente:

1. $k = 0,\ \mathbf{x}_0$ arbitrario, $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$
2. se $\mathbf{r}_k = \mathbf{0}$, stop
3. altrimenti si calcoli

   $\mathbf{z}_k$ tale che $M\mathbf{z}_k = \mathbf{r}_k$,

   $\beta_k = \dfrac{\mathbf{z}_k^T\mathbf{r}_k}{\mathbf{z}_{k-1}^T\mathbf{r}_{k-1}} \quad (\beta_0 = 0)$,

   $\mathbf{p}_k = \mathbf{z}_k + \beta_k\mathbf{p}_{k-1} \quad (\mathbf{p}_0 = \mathbf{z}_0)$,

   $\alpha_k = \dfrac{\mathbf{z}_k^T\mathbf{r}_k}{\mathbf{p}_k^T A\mathbf{p}_k}$,

   $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{p}_k$,

   $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k A\mathbf{p}_k$,

   $k = k + 1$ e si vada al punto 2.

Per la (63) la velocità di convergenza del metodo con precondizionamento può risultare tanto maggiore quanto più la matrice B è "vicina" alla matrice identica, cioè quanto più la matrice M è "vicina" alla matrice A. Varie sono le tecniche di precondizionamento: di notevole interesse fra di esse quelle che si applicano a matrici con strutture particolari, come ad esempio le tecniche che si applicano alle matrici tridiagonali a blocchi trattate in [2].

Semplici tecniche di precondizionamento utilizzano la matrice

$$M = \begin{bmatrix} a_{11} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{nn} \end{bmatrix},$$

che ha come elementi principali quelli di A o la matrice

$$M = \begin{bmatrix} A_{11} & & & \\ & A_{22} & & \\ & & \ddots & \\ & & & A_{nn} \end{bmatrix},$$

diagonale a blocchi, dove i blocchi diagonali quadrati A_{ii} sono i corrispondenti blocchi di A. Una tecnica più raffinata, frequentemente utilizzata e che è particolarmente conveniente se A è sparsa, è quella basata sulla *fattorizzazione incompleta di Cholesky*, in cui $M = LL^T$, dove L è una matrice triangolare inferiore, ottenuta nel modo seguente

$$l_{ii} = \sqrt{a_{ii} - \sum_{r=1}^{i-1} l_{ir}^2}, \quad i = 1, \ldots, n,$$

$$l_{ij} = \left\{ \begin{array}{ll} 0 & \text{se } a_{ij} = 0, \\ \dfrac{1}{l_{jj}} \left[a_{ij} - \displaystyle\sum_{r=1}^{j-1} l_{ir} l_{jr} \right] & \text{se } a_{ij} \neq 0, \end{array} \right\}, \; j = 1, \ldots, i-1, \; i = 2, \ldots, n.$$

La matrice L così ottenuta non è la matrice che si ottiene con fattorizzazione di Cholesky di A, in quanto in essa vengono posti a zero gli elementi che corrispondono a elementi nulli di A. In questo modo, se la matrice A è sparsa, anche L risulta una matrice sparsa e la risoluzione del sistema $M\mathbf{z}_k = \mathbf{r}_k$ viene ricondotta alla risoluzione di due sistemi con matrice triangolare sparsa.

5.35 Esempio Applicando il metodo del gradiente coniugato con il precondizionamento a blocchi di ordine 2 al sistema $A\mathbf{x} = \mathbf{b}$ dell'esempio 5.34 e calcolando il residuo esatto ogni 3 iterazioni, la soluzione viene calcolata con un errore dell'ordine di 10^{-6} con 3 iterazioni se $\alpha = 20$ e con 15 iterazioni se $\alpha = 2.5$. ∎

5.36 Esempio (problema di Dirichlet). Siano Ω il quadrato

$$\Omega = \{ \, (x, y) \in \mathbf{R}^2, 0 < x, y < 1 \, \}$$

e $\partial\Omega$ la sua frontiera e siano $f(x,y)$ e $g(x,y)$ due funzioni assegnate, definite rispettivamente su Ω e su $\partial\Omega$, tali che il problema

$$\begin{cases} \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = f(x,y) & \text{per } (x,y) \in \Omega, \\ u(x,y) = g(x,y) & \text{per } (x,y) \in \partial\Omega, \end{cases} \tag{64}$$

abbia una e una sola soluzione sufficientemente regolare $u(x,y)$. La soluzione $u(x,y)$ può essere approssimata nel modo seguente: si costruisce un reticolo formato di linee parallele agli assi, di solito ugualmente distanti fra di loro, e si considera un problema discreto che approssima il problema (64) nei nodi del reticolo. Più esattamente, fissato un intero n si considerano i punti

$$(x_i, y_j), \quad \text{tali che } x_i = ih,\ y_j = jh,\ h = \frac{1}{n+1},\ i,j = 0,\ldots,n+1.$$

Il reticolo corrispondente è illustrato nella figura 5.13.

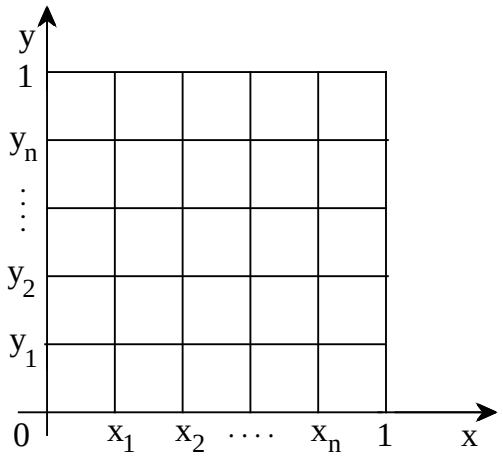

Fig. 5.13 - Reticolo per l'approssimazione della soluzione del problema di Dirichlet.

Per calcolare un'approssimazione della derivata seconda di una funzione $F(x)$ che si suppone derivabile 4 volte con continuità, si utilizzano le due espressioni ottenute con la formula di Taylor

$$F(x_0 - h) = F(x_0) - hF'(x_0) + \frac{h^2}{2}F''(x_0) - \frac{h^3}{3!}F'''(x_0) + \frac{h^4}{4!}F^{(4)}(\xi_1),$$

$$F(x_0 + h) = F(x_0) + hF'(x_0) + \frac{h^2}{2}F''(x_0) + \frac{h^3}{3!}F'''(x_0) + \frac{h^4}{4!}F^{(4)}(\xi_2),$$

da cui si ricava che

$$F''(x_0) = \frac{1}{h^2}[F(x_0 - h) - 2F(x_0) + F(x_0 + h)] + O(h^2).$$

Utilizzando questa relazione, la restrizione della prima equazione del problema (64) ai nodi del reticolo è data da

$$u(x_{i-1}, y_j) - 2u(x_i, y_j) + u(x_{i+1}, y_j) + u(x_i, y_{j-1}) - 2u(x_i, y_j) + u(x_i, y_{j+1})$$

$$= h^2[f(x_i, y_j) + O(h^2)], \qquad i,\ j = 1, \ldots, n. \tag{65}$$

La restrizione della seconda equazione ai nodi del reticolo è data da

$$\begin{aligned} u(x_i, y_j) = g(x_i, y_j), \quad &\text{per } i = 0 \text{ e } i = n+1,\ j = 1, \ldots, n, \\ &\text{e per } i = 1, \ldots, n,\ j = 0 \text{ e } j = n+1. \end{aligned} \tag{66}$$

Trascurando i termini in $O(h^2)$, le relazioni (65) e (66) si riducono ad un sistema di equazioni lineari che consente di calcolare le approssimazioni $u_{i,j}$ della funzione $u(x, y)$ nei punti (x_i, y_j), $i,\ j = 1, \ldots, n$. Il sistema che si ottiene è il seguente

$$-u_{i,j-1} - u_{i-1,j} + 4u_{i,j} - u_{i+1,j} - u_{i,j+1} = -h^2 f(x_i, y_j), \quad i,\ j = 1, \ldots, n,$$

$$\begin{aligned} u_{i,j} = g(x_i, y_j), \quad &\text{per } i = 0 \text{ e } i = n+1,\ j = 1, \ldots, n, \\ &\text{e per } i = 1, \ldots, n,\ j = 0 \text{ e } j = n+1. \end{aligned}$$

Per semplicità si studia il caso in cui $f(x, y) = 0$ per $(x, y) \in \Omega$ e $g(x, y) = 1$ per $(x, y) \in \partial\Omega$, la cui soluzione è $u(x, y) = 1$. Il sistema lineare $A\mathbf{u} = \mathbf{b}$ ottenuto è

$$\begin{bmatrix} B & -I & & \\ -I & B & \ddots & \\ & \ddots & \ddots & -I \\ & & -I & B \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_n \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_n \end{bmatrix}, \tag{67}$$

dove

$$B = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{bmatrix} \in \mathbf{R}^{n \times n},\ \mathbf{u}_j = \begin{bmatrix} u_{1,j} \\ u_{2,j} \\ \vdots \\ u_{n,j} \end{bmatrix} \in \mathbf{R}^n,\ \text{per } j = 1, \ldots, n,$$

$$\mathbf{b}_1 = \mathbf{b}_n = \begin{bmatrix} 2 \\ 1 \\ \vdots \\ 1 \\ 2 \end{bmatrix} \in \mathbf{R}^n, \quad \mathbf{b}_j = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \in \mathbf{R}^n, \text{ per } j = 2, \ldots, n-1.$$

La matrice A ha predominanza diagonale ed è irriducibile: per il teorema 2.41 A è non singolare e quindi il sistema (67) ha una e una sola soluzione $u_{i,j},\ i, j = 1, \ldots, m$, che si assume come approssimazione della soluzione del problema (64). In generale è possibile dimostrare [7] che, sotto opportune ipotesi di regolarità della funzione $u(x, y)$, per $h \to 0$, cioè al tendere a zero dell'ampiezza delle maglie del reticolo, la soluzione del sistema (67) tende alla soluzione $u(x, y)$ del problema (64), nel senso che

$$\lim_{h \to 0} \max_{i,j=1,\ldots,n} |u(x_i, y_j) - u_{ij}| = 0.$$

I metodi esposti in questo testo per la risoluzione dei sistemi lineari sono stati utilizzati per risolvere il sistema (67) nei casi $n = 5$, 10 e 20, cioè per sistemi di ordine rispettivamente 25, 100 e 400. È opportuno rilevare che i sistemi lineari che scaturiscono da problemi reali hanno dimensioni molto maggiori, dell'ordine di decine di migliaia di equazioni, ed è possibile risolverli efficientemente solo perché la matrice ha specifiche proprietà di struttura, quale quella di essere sparsa. Il problema dell'esempio, pur con le sue ridotte dimensioni, è comunque significativo, per confrontare e mettere in evidenza le caratteristiche dei metodi proposti.

Si sono utilizzati i seguenti metodi

G	metodo di Gauss	metodi diretti
GB	metodo di Gauss a blocchi	
C	metodo di Cholesky	
H	metodo di Householder	

J	metodo di Jacobi	metodi iterativi
JB	metodo di Jacobi a blocchi	
GS	metodo di Gauss-Seidel	
GSB	metodo di Gauss-Seidel a blocchi	
S	metodo di rilassamento	
SB	metodo di rilassamento a blocchi	

GC	metodo del gradiente coniugato	metodi del gradiente coniugato
GCB	metodo del gradiente coniugato con il precondizionamento a blocchi	
GCC	metodo del gradiente coniugato con la fattor. incompleta di Cholesky	

Nella tabella 5.1 sono riportati i risultati ottenuti utilizzando i metodi diretti. t indica il tempo impiegato, misurato in millesimi di secondo, ed $\|\mathbf{e}\|_2$ indica la norma 2 dell'errore assoluto della soluzione calcolata.

Nella tabella 5.2 sono riportati i risultati ottenuti utilizzando i metodi iterativi. È stata usata la condizione di arresto (16) $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty < \epsilon$ con la tolleranza $\epsilon = 10^{-5}$ e k indica il numero delle iterazioni richieste. Come si nota, i risultati ottenuti sono del tutto in accordo con i risultati dell'esempio 5.31, e in particolare per il caso n=20 con le stime asintotiche della velocità di convergenza.

Nella tabella 5.3 sono riportati i risultati ottenuti utilizzando il metodo del gradiente coniugato: si è usata la condizione di arresto (62) $\|\mathbf{r}_k\|_2 < \epsilon\|\mathbf{b}\|_2$, con la tolleranza $\epsilon = 10^{-5}$. Come si nota, il metodo del gradiente coniugato consente di approssimare la soluzione con una precisione maggiore e in un tempo minore rispetto agli altri metodi e la precisione richiesta viene raggiunta con un numero di iterazioni assai inferiore a n^2, dimensione della matrice A.

Per implementare su calcolatore i vari metodi sono state sfruttate le specifiche proprietà della matrice. In particolare sia il metodo di Gauss che quello di Cholesky, utilizzando il fatto che la matrice è definita positiva e a banda di ampiezza n richiedono un numero di operazioni dell'ordine $n^4/2$ (si veda l'esercizio 4.41). Il metodo di Householder genera una matrice a banda superiore di ampiezza $2n$, e richiede $4n^4$ operazioni (si veda l'esercizio 4.41). I tempi di calcolo sono legati, oltre che al numero delle operazioni, anche ad altri fattori (per esempio l'individuazione e il trasferimento dei dati), che dipendono anche dalle tecniche di programmazione. In particolare per valori piccoli di n, al numero delle operazioni, va aggiunto il numero delle radici quadrate per il metodo di Cholesky e di Householder e per quello del gradiente coniugato con la fattorizzazione incompleta di Cholesky.

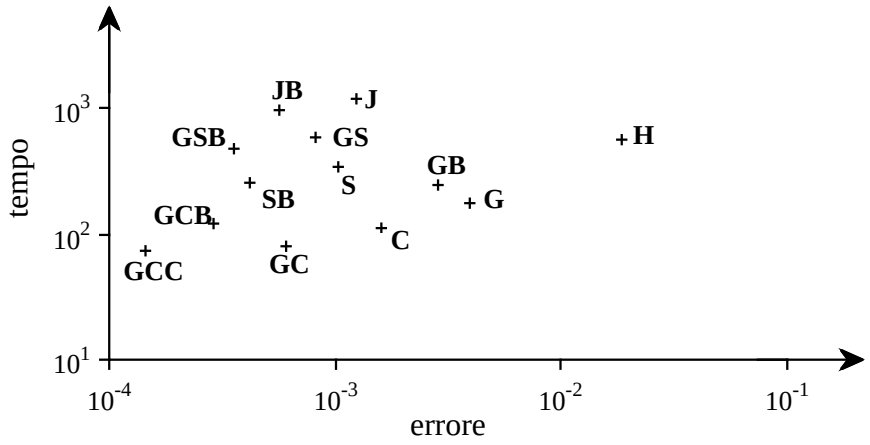

Fig. 5.14 - Tempo (in millisec.) ed errore dei metodi usati per risolvere il sistema (67).

Per il caso $n = 20$ (ordine del sistema 400) i risultati delle tabelle sono sintetizzati nella figura 5.14, in cui in ascissa sono riportati gli errori del risultato generato da ciascun metodo e in ordinata i tempi di esecuzione in millesimi di secondo. ∎

Esercizi proposti

5.1 Sia $A \in \mathbf{C}^{n\times n}$. Senza utilizzare il teorema 5.2,

a) si dimostri per ogni norma matriciale indotta $\| \, . \, \|$ che se

$$\lim_{k\to\infty} \|A^k\| = 0 \quad \text{allora} \quad \rho(A) < 1.$$

Posto $A = UTU^H$, U unitaria e T triangolare superiore, se $\rho(A) < 1$, si dimostri che

b) se T ha n autovettori linearmente indipendenti, allora $\lim_{k\to\infty} T^k = O$;

c) se T non ha n autovettori linearmente indipendenti, si può costruire una matrice $S \in \mathbf{R}^{n\times n}$ tale che $S > |T|$, $\rho(S) < 1$ e S ha n autovettori linearmente indipendenti, e quindi

$$\lim_{k\to\infty} S^k = O;$$

d) se $\rho(A) < 1$, esiste una norma matriciale indotta $\| \, . \, \|$ per cui

$$\lim_{k\to\infty} \|A^k\| = 0.$$

Sfruttando i risultati dei punti a), b) e c), si dia una dimostrazione del teorema 5.2, in cui si faccia riferimento alla forma normale di Schur, anziché alla forma normale di Jordan; sfruttando i risultati dei punti a) e d), si dia un'altra dimostrazione del teorema 5.2.

(Traccia: a) per il teorema 3.10 è $\|A^k\| \geq [\rho(A)]^k$; b) sia D diagonale tale che $T = CDC^{-1}$, allora $T^k = CD^kC^{-1}$, inoltre $\rho(D) = \rho(T) = \rho(A) < 1$ e quindi $\lim_{k\to\infty} D^k = O$; c) poiché $\rho(T) < 1$, gli elementi principali di T sono in modulo minori di 1, basta porre S la matrice i cui elementi sono

$$s_{ij} = \begin{cases} \alpha_i & \text{per } j = i, \\ \beta & \text{per } j > i, \\ 0 & \text{per } j < i, \end{cases} \quad \text{dove} \quad \begin{cases} \alpha_i > |t_{ii}| & \text{per } i = 1, \ldots, n, \\ \beta > |t_{ij}| & \text{per } i, j = 1, \ldots, n, \end{cases}$$