

# Capitolo 3

## Sistemi di equazioni lineari

In questo capitolo si studiano due tipi di metodi risolutivi per sistemi di equazioni lineari, solitamente detti *metodi diretti* e *metodi iterativi*. Nei metodi diretti si giunge alla soluzione esatta (a meno degli errori di arrotondamento) con un numero finito di operazioni sui dati; nei metodi iterativi la soluzione viene invece approssimata dai termini di una successione di cui la soluzione cercata è il limite. La convenienza dell'uno o dell'altro tipo di metodo dipende da particolari proprietà del sistema che si vuole risolvere.

Per semplicità si fa riferimento al caso di sistemi reali, notando che l'estensione degli algoritmi al campo complesso non presenta particolari difficoltà.

### 3.1 Algoritmo base del metodo di Gauss

Dato il sistema di  $n$  equazioni lineari

$$\begin{array}{ccccccc} a_{11}x_1 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ \cdots & & \cdots & & \cdots & & \cdots \\ a_{n1}x_1 & + & \cdots & + & a_{nn}x_n & = & b_n \end{array} \quad (3.1)$$

dove i *coefficienti*  $a_{ij}$  e i *termini noti*  $b_i$  sono numeri reali, si cerca un vettore  $x^T = (x_1, x_2, \dots, x_n)$  che verifichi le (3.1). Introdotta la matrice dei coefficienti  $A$  e il vettore dei termini noti  $b$ , il sistema si può scrivere nella forma

$$Ax = b, \quad (3.2)$$

dove si suppone  $A$  non singolare, per garantire l'esistenza e l'unicità della soluzione.

Il *metodo di Gauss* o di *eliminazione* consiste nel trasformare il sistema (3.2) in un altro equivalente

$$Rx = c, \quad (3.3)$$

dove  $R$  è una matrice triangolare superiore con  $r_{ii} \neq 0$ ,  $i = 1, 2, \dots, n$ .

Il sistema (3.3) è quindi della forma

$$\begin{array}{ccccccc} r_{11}x_1 & + & r_{12}x_2 & + & \cdots & + & r_{1n}x_n & = & c_1 \\ & & r_{22}x_2 & + & \cdots & + & r_{2n}x_n & = & c_2 \\ & & & & \cdots & & \cdots & & \cdots \\ & & & & & & r_{nn}x_n & = & c_n \end{array} \quad (3.4)$$

e si risolve immediatamente con le formule

$$x_n = c_n / r_{nn}$$

$$x_i = (c_i - \sum_{j=i+1}^n r_{ij}x_j) / r_{ii}, \quad i = n-1, \dots, 1.$$

Per passare dal sistema (3.1) ad uno equivalente della forma (3.4) occorre *eliminare* dalla  $i$ -esima equazione le incognite con indice minore di  $i$ , per  $i = 2, 3, \dots, n$ . Ciò si effettua utilizzando la proprietà che la soluzione non cambia se si sostituisce all'equazione  $i$ -esima una sua combinazione lineare con un'altra equazione del sistema. Pertanto, se  $a_{11} \neq 0$ , si elimina  $x_1$  da tutte le equazioni che seguono la prima, sottraendo membro a membro dalla  $i$ -esima equazione,  $i = 2, 3, \dots, n$ , la prima equazione i cui membri siano stati moltiplicati per il coefficiente, detto appunto *moltiplicatore*,

$$l_{i1} = \frac{a_{i1}}{a_{11}}. \quad (3.5)$$

Ponendo per ragioni formali  $a_{ij}^{(1)} := a_{ij}$ ,  $i, j = 1, 2, \dots, n$ , il sistema, dopo la prima eliminazione, assume la forma:

$$\begin{array}{ccccccc} a_{11}^{(1)}x_1 & + & a_{12}^{(1)}x_2 & + & a_{13}^{(1)}x_3 & + & \cdots & + & a_{1n}^{(1)}x_n & = & b_1^{(1)} \\ & & a_{22}^{(2)}x_2 & + & a_{23}^{(2)}x_3 & + & \cdots & + & a_{2n}^{(2)}x_n & = & b_2^{(2)} \\ & & \cdots & & \cdots & & \cdots & & \cdots & & \cdots \\ & & a_{n2}^{(2)}x_2 & + & a_{n3}^{(2)}x_3 & + & \cdots & + & a_{nn}^{(2)}x_n & = & b_n^{(2)} \end{array} \quad (3.6)$$

dove

$$a_{ij}^{(2)} = a_{ij}^{(1)} - l_{i1}a_{1j}^{(1)}, \quad b_i^{(2)} = b_i^{(1)} - l_{i1}b_1^{(1)}, \quad i, j = 2, 3, \dots, n.$$

Questo sistema si può sostituire al sistema (3.1) senza cambiarne la soluzione.

Se nel sistema (3.6) risulta  $a_{22}^{(2)} \neq 0$ , si può eliminare  $x_2$  da tutte le equazioni che seguono la seconda, utilizzando ora i moltiplicatori  $l_{i2} = \frac{a_{i2}^{(2)}}{a_{22}^{(2)}}$ ,  $i = 3, 4, \dots, n$ , e così via. Supposto che tale procedimento possa ripetersi  $n - 1$  volte, si giunge al sistema

$$\begin{array}{ccccccc} a_{11}^{(1)}x_1 & + & a_{12}^{(1)}x_2 & + & \cdots & + & a_{1n}^{(1)}x_n & = & b_1^{(1)} \\ & & a_{22}^{(2)}x_2 & + & \cdots & + & a_{2n}^{(2)}x_n & = & b_2^{(2)} \\ & & & & \cdots & & \cdots & & \cdots \\ & & & & & & a_{nn}^{(n)}x_n & = & b_n^{(n)} \end{array} \quad (3.7)$$

che è della forma (3.4) ed è equivalente a (3.1).

Le condizioni perché l'algoritmo possa giungere al termine come descritto, sono

$$a_{11}^{(1)} \neq 0, \quad a_{22}^{(2)} \neq 0, \quad \dots, \quad a_{nn}^{(n)} \neq 0. \quad (3.8)$$

In mancanza di una di queste condizioni l'algoritmo si interrompe.

Le (3.8) equivalgono, com'è facile verificare, alla proprietà che la matrice  $A$  abbia i minori principali di testa diversi da zero, cioè

$$a_{11} \neq 0, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{vmatrix} \neq 0, \dots, \det(A) \neq 0. \quad (3.9)$$

In realtà poche matrici godono di questa proprietà; fra queste si trovano le matrici simmetriche e definite, che ricorrono spesso nelle applicazioni.

Nella pratica del calcolo l'algoritmo di base viene modificato sia per garantirne la completa esecuzione, sia per ridurre la propagazione degli errori di arrotondamento, anche quando le condizioni (3.9) fossero verificate.

Le modifiche apportate in questo senso non alterano comunque il numero di operazioni essenziali (moltiplicazioni e divisioni) che, per un sistema di  $n$  equazioni in  $n$  incognite, si può verificare che ammonta a  $\frac{n^3}{3} + n^2 - \frac{n}{3}$ .

Si noti che lo stesso sistema, risolto con la regola di Cramer, che è pure un metodo diretto, richiede circa  $(n - 1)(n + 1)!$  operazioni.

L'uso della tecnica di eliminazione, evitando il calcolo dei determinanti, riduce notevolmente anche il numero delle operazioni necessarie per il calcolo

della matrice inversa di  $A$ . La matrice  $A^{-1}$  è infatti la soluzione del sistema matriciale

$$AX = I,$$

che equivale ad  $n$  sistemi lineari della forma

$$Ax^{(i)} = e^{(i)}, \quad i = 1, 2, \dots, n,$$

dove si è posto  $X = [x^{(1)} \mid x^{(2)} \mid \dots \mid x^{(n)}]$  e  $I = [e^{(1)} \mid e^{(2)} \mid \dots \mid e^{(n)}]$ .

## 3.2 Tecniche di pivoting

I coefficienti  $a_{11}^{(1)}, a_{22}^{(2)}, a_{33}^{(3)}, \dots$  del sistema (3.7) si dicono *elementi pivotali* (dal francese **pivot**=perno). Le modifiche dell'algoritmo che si introducono per i motivi detti nel paragrafo precedente, consistono nello stabilire a priori un criterio di scelta dell'elemento pivotale per ciascuna eliminazione.

Un primo criterio, detto del *pivoting parziale*, è il seguente: si supponga che nel sistema (3.1) sia

$$\max_{1 \leq i \leq n} |a_{i1}^{(1)}| = |a_{r1}^{(1)}|; \quad (3.10)$$

allora, se  $r \neq 1$ , si scambiano di posto la prima e l' $r$ -esima equazione e quindi si considera un sistema in cui i coefficienti  $a_{1j}^{(1)}, j = 1, 2, \dots, n$ , sono i coefficienti  $a_{rj}^{(1)}, j = 1, 2, \dots, n$ , del sistema di partenza e viceversa. Effettuata la prima eliminazione, si supponga che nel sistema (3.6) si abbia

$$\max_{2 \leq i \leq n} |a_{i2}^{(2)}| = |a_{s2}^{(2)}|; \quad (3.11)$$

allora, se  $s \neq 2$ , si scambiano di posto l'equazione di indice  $s$  con quella di indice 2, quindi si procede alla seconda eliminazione e così via.

Un'altra strategia è quella del *pivoting totale* in cui il pivot è ancora l'elemento di massimo modulo, ma scelto ogni volta sull'intera matrice del sistema parziale da trasformare anziché su una sola colonna come nelle (3.10) e (3.11). È chiaro che in questo caso per portare il pivot selezionato nella posizione di testa può essere necessario un riordinamento delle equazioni e delle incognite.

Nel caso di sistemi con equazioni *sbilanciate*, cioè con coefficienti di una stessa equazione molto diversi nell'ordine di grandezza, il criterio del pivoting

parziale può risultare inefficace ai fini della riduzione degli errori di arrotondamento. In questi casi conviene ricorrere al pivoting totale oppure al così detto *pivoting parziale bilanciato* che consiste nello scegliere come elementi pivotali gli elementi  $a_{r1}^{(1)}, a_{s2}^{(2)}, \dots$ , tali che si abbia

$$\frac{|a_{r1}^{(1)}|}{m_r^{(1)}} = \max_{1 \leq i \leq n} \frac{|a_{i1}^{(1)}|}{m_i^{(1)}}, \quad \frac{|a_{s2}^{(2)}|}{m_s^{(2)}} = \max_{2 \leq i \leq n} \frac{|a_{i2}^{(2)}|}{m_i^{(2)}}, \dots, \quad (3.12)$$

dove i numeri  $m_i^{(1)} = \max_{1 \leq j \leq n} |a_{ij}^{(1)}|$ ,  $i = 1, 2, \dots, n$ , vanno calcolati all'inizio, sulla matrice  $A$  del sistema di partenza, i numeri  $m_i^{(2)} = \max_{2 \leq j \leq n} |a_{ij}^{(2)}|$ ,  $i = 2, 3, \dots, n$ , si calcolano sulla matrice del sistema (3.6) etc..

### 3.3 Fattorizzazione LR

L'algoritmo di eliminazione può essere considerato come un procedimento che trasforma una data matrice  $A$  in una matrice triangolare  $R$ .

Per vedere in quale relazione sono le matrici  $A$  ed  $R$  si supponga che la matrice  $A$  verifichi le condizioni (3.9) e quindi che si possa applicare l'algoritmo di eliminazione senza effettuare scambi tra le righe.

**Teorema 3.3.1** *Nell'ipotesi che valgano le condizioni (3.9) l'algoritmo di eliminazione produce la fattorizzazione*

$$A = LR, \quad (3.13)$$

dove  $R$  è la matrice triangolare superiore data dai coefficienti del sistema (3.7) ed  $L$  ha la forma

$$L = \begin{pmatrix} 1 & & & & & \\ l_{21} & 1 & & & & \\ l_{31} & l_{32} & 1 & & & \\ \vdots & \vdots & \ddots & \ddots & & \\ \vdots & \vdots & & \ddots & 1 & \\ l_{n1} & l_{n2} & \cdots & \cdots & l_{n,n-1} & 1 \end{pmatrix}$$

in cui gli elementi al disotto della diagonale principale coincidono con i moltiplicatori dell'algoritmo di eliminazione.

**DIMOSTRAZIONE.** Siano  $A_1, A_2, \dots, A_{n-1} = R$  le matrici dei successivi sistemi equivalenti a (3.1) che si ottengono dopo ciascuna eliminazione.

Si constata che

$$A_1 = H_1 A, A_2 = H_2 A_1, \dots, A_{n-1} = H_{n-1} A_{n-2} = R \quad (3.14)$$

con

$$H_i = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & -l_{i+1,i} & \ddots & & \\ & & \vdots & & \ddots & \\ & & -l_{n,i} & & & 1 \end{pmatrix}.$$

Posto  $L_i := H_i^{-1}$  e tenuto conto che

$$H_i^{-1} = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & l_{i+1,i} & \ddots & & \\ & & \vdots & & \ddots & \\ & & l_{n,i} & & & 1 \end{pmatrix}$$

e che  $L_1 L_2 \dots L_{n-1} = L$ , dalle (3.14) segue

$$H_{n-1} H_{n-2} \dots H_1 A = R,$$

da cui

$$A = L_1 L_2 \dots L_{n-1} R = LR.$$

□

Nel caso di una matrice  $A$  qualunque si può dimostrare che l'algoritmo di Gauss con l'eventuale uso del pivoting parziale conduce ancora ad una fattorizzazione della forma

$$PA = L_p R_p \quad (3.15)$$

dove  $P$  è una matrice di permutazione definita dagli scambi di righe richiesti dall'algoritmo,  $R_p$  è triangolare superiore ed  $L_p$  è triangolare inferiore con elementi diagonali unitari.

Una conseguenza delle decomposizioni (3.13) e (3.15) è data rispettivamente dalle uguaglianze

$$\det(A) = \det(R), \quad \det(A) = (-1)^s \det(R_p)$$

dove  $s$  è il numero degli scambi di righe dovuti all'uso del pivoting, mentre i determinanti di  $R$  ed  $R_p$  sono dati dal prodotto dei termini diagonali. Si osservi che il costo computazionale di  $\det(A)$  mediante la definizione è di circa  $n!$  operazioni mentre il numero di operazioni per costruire  $R$  ed  $R_p$  con l'eliminazione gaussiana è di circa  $n^3/3$ .

### 3.4 Metodi di fattorizzazione

La conoscenza di una fattorizzazione della matrice  $A$  può essere utile ai fini della risoluzione del sistema (3.1), infatti se ad esempio si conosce a priori la decomposizione (3.13), il sistema si può scrivere

$$LRx = b,$$

e la sua risoluzione si riconduce a quella immediata dei due sistemi triangolari

$$Lc = b, \quad Rx = c. \quad (3.16)$$

Nell'ipotesi che valgano le condizioni (3.9) l'eliminazione gaussiana produce le due matrici  $R$  ed  $L$ , quest'ultima essendo fornita dai moltiplicatori  $l_{ij}$  che si possono convenientemente memorizzare durante l'esecuzione dell'algoritmo.

Tuttavia se lo scopo della fattorizzazione è la risoluzione dei sistemi (3.16) si preferisce costruire direttamente le matrici  $L$  ed  $R$  sulla base della definizione di prodotto fra matrici, ferma restando l'ipotesi (3.9).

Si hanno così i *metodi di fattorizzazione diretta* che si fondano sulla (3.13) pensata come un sistema di  $n^2$  equazioni

$$a_{ij} = \sum_{h=1}^{\min(i,j)} l_{ih}r_{hj}, \quad i, j = 1, 2, \dots, n. \quad (3.17)$$

Per ricavare gli elementi di  $L$  ed  $R$  dalla (3.17) si possono seguire diversi schemi di calcolo.

Nel *metodo di Doolittle* si pone nelle (3.17)  $l_{ii} = 1$ ,  $i = 1, 2, \dots, n$ , sicché le  $n^2$  incognite sono gli  $n(n+1)/2$  elementi  $r_{ij}$  di  $R$  con  $j \geq i$  e gli  $(n-1)n/2$  elementi  $l_{ij}$  di  $L$  al disotto della diagonale principale. L'ordine che si segue nella risoluzione delle (3.17) è il seguente:

1. si pone  $i = 1$  e si ricavano le  $r_{1j}$  per la prima riga di  $R$  dalle  $n$  equazioni

$$a_{1j} = l_{11}r_{1j}, \quad j = 1, 2, \dots, n;$$

2. si pone  $j = 1$  e si ricavano le  $l_{i1}$  per la prima colonna di  $L$  dalle  $n-1$  equazioni

$$a_{i1} = l_{i1}r_{11}, \quad i = 2, 3, \dots, n;$$

3. si pone  $i = 2$  e si ricavano le  $r_{2j}$  per la seconda riga di  $R$  da

$$a_{2j} = \sum_{h=1}^2 l_{2h}r_{hj}, \quad j = 2, 3, \dots, n.$$

Così proseguendo, si costruiscono alternativamente una riga completa di  $R$  e una colonna senza l'elemento diagonale di  $L$ , seguendo l'ordine rappresentato in Fig. 3.1.

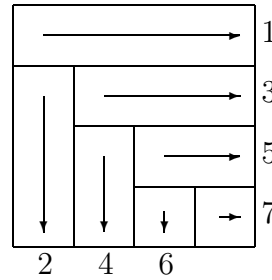


Figura 3.1: Metodo di Doolittle.

Se nelle (3.17) si pone  $r_{ii} = 1$ ,  $i = 1, 2, \dots, n$ , si ottiene il *metodo di Crout* in cui si costruiscono alternativamente una colonna completa di  $L$  ed una riga senza l'elemento diagonale di  $R$  secondo un ordinamento che è il trasposto di quello della Fig. 3.1 (cfr. Fig. 3.2).

Lo schema di calcolo per usare le (3.17) è quindi il seguente:

1. si pone  $j = 1$  e si ricavano le  $l_{i1}$  da

$$a_{i1} = l_{i1}r_{11}, \quad i = 1, 2, \dots, n;$$



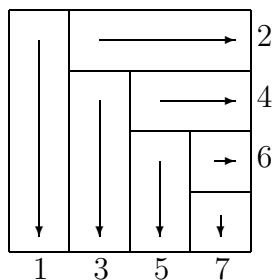


Figura 3.2: Metodo di Crout.

2. si pone  $i = 1$  e si ricavano le  $r_{1j}$  da

$$a_{1j} = l_{11}r_{1j}, \quad j = 2, 3, \dots, n;$$

3. si pone  $j = 2$  e si ricavano le  $l_{i2}$  da

$$a_{i2} = \sum_{h=1}^2 l_{ih}r_{h2}, \quad i = 2, 3, \dots, n.$$

E così via.

Con la scelta  $l_{ii} = 1$  associata alla costruzione per righe alternate di  $R$  e di  $L$  si ha il *metodo di Banachiewicz*.

Tutti questi metodi sono applicabili solo quando siano verificate le condizioni (3.9) ed hanno, rispetto alla eliminazione gaussiana, il solo vantaggio di una esecuzione più compatta che non richiede la memorizzazione di stadi intermedi.

Nel caso speciale dei sistemi lineari con matrice simmetrica definita positiva esiste la possibilità di ridurre anche il numero di operazioni essenziali, quasi dimezzandole, rispetto al metodo di eliminazione. Ciò si ottiene ricorrendo alla fattorizzazione

$$A = LL^T \quad (3.18)$$

valida per ogni matrice  $A$  simmetrica e definita positiva, con  $L$  matrice triangolare inferiore ed elementi diagonali positivi ma non necessariamente uguali ad 1.

Sulla (3.18) si fonda il *metodo di Cholesky* in cui ci si limita a costruire solo la matrice  $L$  procedendo per colonne. Posto nella (3.17)  $r_{hj} := l_{jh}$  si ha per  $i \geq j$ ,

$$a_{ij} = l_{i1}l_{j1} + l_{i2}l_{j2} + \dots + l_{ij}l_{jj}, \quad j = 1, 2, \dots, n, \quad (3.19)$$

dando a  $i$  tutti i valori da  $j$  ad  $n$ , dalla (3.19) si ricavano gli elementi  $l_{ij}$  della colonna  $j$ -esima di  $L$ . Si noti che per  $i = j$  la (3.19) diventa

$$a_{jj} = l_{j1}^2 + l_{j2}^2 + \cdots + l_{jj}^2$$

da cui

$$l_{jj} = \sqrt{a_{jj} - \sum_{h=1}^{j-1} l_{jh}^2}$$

dove per  $j = 1$  la sommatoria è nulla.

In generale, quando si conosce una fattorizzazione  $A = LR$  si ha formalmente  $A^{-1} = R^{-1}L^{-1}$ , perciò per avere l'inversa di  $A$  basta risolvere i due sistemi matriciali triangolari

$$RX = I, \quad LY = I,$$

che forniscono rispettivamente  $R^{-1}$  ed  $L^{-1}$  e poi eseguire il prodotto  $R^{-1}L^{-1}$ . In particolare se  $A$  è simmetrica e definita positiva basta risolvere soltanto il secondo sistema, avendosi

$$A^{-1} = (L^T)^{-1}L^{-1} = (L^{-1})^T L^{-1}.$$

### 3.5 Errori, stabilità e condizionamento

Qualunque metodo per la risoluzione di un sistema lineare produce una soluzione approssimata a causa degli errori di arrotondamento introdotti nel corso dei calcoli. Tali errori vengono amplificati e trasmessi alla soluzione attraverso un meccanismo che dipende sia dall'algoritmo che dal sistema stesso.

Sia  $x$  la soluzione esatta del sistema  $Ax = b$ , al quale si supponga di applicare un qualunque metodo diretto e sia  $x + \delta x$  la soluzione approssimata che si ottiene. Si usa ammettere che l'influenza dell'algoritmo equivalga ad una certa perturbazione  $\delta A$ ,  $\delta b$  dei dati iniziali, per cui la soluzione numerica  $x + \delta x$  si può pensare come la soluzione esatta del sistema perturbato

$$(A + \delta A)(x + \delta x) = b + \delta b.$$

Un algoritmo che produce forti perturbazioni si dice *instabile*, mentre si dice *stabile* se le perturbazioni prodotte sono modeste.

L'entità dell'errore relativo  $\frac{\|\delta x\|}{\|x\|}$  dipende dalla sensibilità della soluzione alle perturbazioni dei dati  $A$  e  $b$  o, come si dice, dal *condizionamento* del sistema, termine col quale si designa, più in generale, l'attitudine che ha un dato problema a trasmettere, più o meno amplificate, le perturbazioni dei dati alla soluzione. Precisamente vale il teorema seguente.

**Teorema 3.5.1** *Nell'ipotesi che la matrice  $A + \delta A$  sia non singolare e che, rispetto ad una data norma, sia  $\|\delta A\| < 1/\|A^{-1}\|$ , vale la relazione:*

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\mu(A)}{1 - \mu(A) \frac{\|\delta A\|}{\|A\|}} \left( \frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right), \quad (3.20)$$

dove

$$\mu(A) = \|A\| \|A^{-1}\|. \quad (3.21)$$

Si osservi che quando il numero  $\mu(A)$  definito dalla (3.21) è "molto grande" si ha una forte amplificazione del membro destro della (3.20) e l'errore relativo della soluzione può essere molto grande. Per questo si suole assumere  $\mu(A)$  come misura del condizionamento del sistema o della matrice  $A$  e si dice appunto *numero di condizionamento* rispetto alla norma considerata.

Il numero  $\mu(A)$  è non inferiore all'unità, avendosi, per ogni norma,

$$\mu(A) = \|A\| \|A^{-1}\| \geq \|AA^{-1}\| = \|I\| \geq 1.$$

In generale  $A$  si dice *malcondizionata* se  $\mu(A) \gg 1$  e *bencondizionata* se  $\mu(A)$  non è "molto grande", ma è chiaro che, tranne casi estremi, l'adozione di questi termini dipende da criteri contingenti.

Si osservi che se nella (3.20) si pone  $\|\delta A\| = 0$ , l'errore relativo può crescere al più linearmente al crescere di  $\|\delta b\|$  mentre al crescere di  $\|\delta A\|$  l'errore potrebbe subire aumenti assai più forti in quanto nel membro destro della (3.20) cresce anche il fattore  $\mu(A) / (1 - \mu(A) \frac{\|\delta A\|}{\|A\|})$ .

Per questo motivo per misurare la stabilità o meno dell'algoritmo usato, si cerca di risalire alla perturbazione  $\|\delta A\|$ , partendo dagli errori di arrotondamento introdotti dall'algoritmo di fattorizzazione della matrice  $A$ . Questa tecnica, detta di *analisi dell'errore all'indietro*, viene usata in generale anche per altri algoritmi.

Nota una stima della perturbazione sui dati corrispondente ad un certo algoritmo diretto, la (3.20) fornisce una maggiorazione a priori dell'errore relativo della soluzione. Di validità più generale è invece una maggiorazione

a posteriori che si ricava come segue: sia  $\tilde{x}$  la soluzione ottenuta per il sistema  $Ax = b$  risolto con un qualunque metodo e si abbia

$$b - A\tilde{x} = r$$

dove  $r$  è il *vettore residuo*. In corrispondenza alla soluzione esatta,  $r$  risulta nullo, cioè si ha  $b - Ax = 0$ ; ne segue

$$A(\tilde{x} - x) = -r, \quad (\tilde{x} - x) = -A^{-1}r$$

e, per una qualunque norma naturale:

$$\|\tilde{x} - x\| \leq \|A^{-1}\| \|r\|$$

d'altra parte da  $Ax = b$  si ha

$$\|x\| \geq \frac{\|b\|}{\|A\|}$$

e dalle ultime due relazioni segue la detta maggiorazione

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq \mu(A) \frac{\|r\|}{\|b\|}. \quad (3.22)$$

La (3.22) mostra che la dipendenza dell'errore finale da  $\mu(A)$  è un fatto generale e mette in guardia dal ritenere buona un'approssimazione  $\tilde{x}$  quando il corrispondente residuo sia "piccolo".

In generale non si conosce la matrice inversa di  $A$  e quindi  $\mu(A)$ ; tuttavia la (3.22) può essere effettivamente utilizzata ricorrendo ad appositi procedimenti per il calcolo approssimato di  $\mu(A)$ .

### 3.6 Metodi iterativi in generale

Molti problemi conducono alla risoluzione di un sistema  $Ax = b$  di dimensioni molto grandi con matrice  $A$  *sparsa*, cioè con pochi elementi non nulli. Se a un tale sistema si applica un metodo diretto, le matrici dei sistemi intermedi o di arrivo possono diventare matrici *dense*, cioè con un elevato numero di elementi non nulli.

Sorgono così seri problemi di costo computazionale e di ingombro di memoria. In questi casi può giovare il ricorso ai metodi iterativi, in cui ogni iterazione richiede il prodotto di una matrice  $H$  per un vettore.

Poiché la densità di  $H$  è paragonabile a quella di  $A$ , se questa è una matrice sparsa, ogni iterazione comporta una mole di calcoli relativamente modesta ed un ingombro di memoria limitato.

Il procedimento generale per costruire un metodo iterativo è il seguente.

Dato il sistema

$$Ax - b = 0, \quad \text{con } A \in \mathbb{R}^{n \times n}, \quad b \in \mathbb{R}^n, \quad \det(A) \neq 0,$$

si trasforma il sistema dato in un altro equivalente della forma

$$x = Hx + c. \quad (3.23)$$

Ciò può farsi in molti modi; per esempio, con  $G$  matrice non singolare qualsiasi, si può scrivere

$$x = x - G(Ax - b)$$

da cui

$$x = (I - GA)x + Gb$$

che è della forma voluta.

La (3.23) suggerisce il processo iterativo

$$x^{(k+1)} = Hx^{(k)} + c, \quad k = 0, 1, \dots, \quad (3.24)$$

dove  $x^{(0)}$  è una approssimazione iniziale della soluzione.

La matrice  $H$  è detta *matrice di iterazione* e definisce il metodo. Un metodo iterativo si dice convergente se la successione  $\{x^{(k)}\}$  converge alla soluzione del sistema dato.

La convergenza o meno della successione  $\{x^{(k)}\}$  generata da un metodo iterativo dipende dalla sua matrice di iterazione  $H$  in base al seguente teorema.

**Teorema 3.6.1** *Condizione necessaria e sufficiente affinché un metodo iterativo della forma (3.24) sia convergente per qualunque vettore iniziale  $x^{(0)}$ , è che la sua matrice di iterazione  $H$  sia convergente.*

**DIMOSTRAZIONE.** Sia  $a$  la soluzione esatta del sistema  $Ax = b$  e si voglia usare un metodo iterativo del tipo (3.24).

Essendo  $x = Hx + c$  equivalente al sistema dato, vale l'identità

$$a = Ha + c;$$

sottraendo membro a membro questa dalla (3.24) e indicando con  $e^{(k)} = x^{(k)} - a$  l'errore associato a  $x^{(k)}$ , si ha

$$e^{(k+1)} = He^{(k)}, \quad k = 0, 1, \dots,$$

da cui

$$e^{(k)} = H^k e^{(0)}; \quad (3.25)$$

perciò, per un arbitrario  $e^{(0)}$ , si avrà  $\lim_{k \rightarrow \infty} e^{(k)} = 0$  allora e solo che sia  $\lim_{k \rightarrow \infty} H^k = \mathbf{O}$ .  $\square$

Da note proprietà delle matrici e delle loro norme (cfr. Teorema 2.7.1 e Teorema 2.10.3), si ottengono i seguenti corollari.

**Corollario 3.6.1** *Per la convergenza del metodo (3.24) è necessario e sufficiente che sia*

$$\rho(H) < 1. \quad (3.26)$$

**Corollario 3.6.2** *Condizione sufficiente per la convergenza del metodo (3.24) è l'esistenza di una norma naturale per cui si abbia*

$$\|H\| < 1.$$

La (3.25) consente di studiare la riduzione dell'errore nel corso delle iterazioni. Infatti si dimostra che per una qualunque norma naturale si ha

$$\lim_{k \rightarrow \infty} \sqrt[k]{\|H^k\|} = \rho(H);$$

quindi asintoticamente, cioè per  $k$  abbastanza grande, si ha

$$\sqrt[k]{\|H^k\|} \simeq \rho(H); \quad (3.27)$$

da questa e dalla (3.25) segue, se  $\|e^{(0)}\| \neq 0$ ,

$$\frac{\|e^{(k)}\|}{\|e^{(0)}\|} \leq \|H^k\| \simeq \rho^k(H). \quad (3.28)$$

Perciò, in un metodo convergente,  $\|e^{(k)}\|$  si riduce almeno a  $\|e^{(0)}\| \times 10^{-m}$  dopo un numero  $k$  di iterazioni tale che  $\rho^k(H) \leq 10^{-m}$  ossia se

$$\frac{k}{m} \geq -\frac{1}{\text{Log } \rho(H)} \quad (3.29)$$

(si ricordi che per la (3.26) è  $\text{Log } \rho(H) < 0$ ).

Dalla (3.29) si vede che, nell'ambito dei metodi convergenti, la convergenza risulta tanto più rapida quanto più grande è il numero  $-\text{Log } \rho(H)$ .

Poiché la (3.29) è stata dedotta dalle relazioni asintotiche (3.27) e (3.28), al numero

$$V = \frac{m}{k} = -\text{Log } \rho(H) \quad (3.30)$$

si dà il nome di *velocità asintotica di convergenza* del metodo avente matrice di iterazione  $H$ .

In base alla (3.30) se due metodi hanno matrici di iterazione con diverso raggio spettrale è più veloce quello che corrisponde al raggio spettrale minore.

Sottraendo ad entrambi i membri della (3.24) il vettore  $x^{(k)}$  e tenendo conto che  $c = -(H - I)a$  si perviene alla

$$\|e^{(k)}\| \leq \|(H - I)^{-1}\| \|x^{(k+1)} - x^{(k)}\|. \quad (3.31)$$

L'uso di un metodo iterativo comporta il ricorso a qualche criterio di arresto. Se  $\epsilon$  è una tolleranza d'errore prestabilita, un criterio spesso seguito è il seguente

$$\|x^{(k+1)} - x^{(k)}\| \leq \epsilon \quad (3.32)$$

che si basa sulla maggiorazione (3.31). Tale criterio è chiaramente inefficiente se il numero  $\|(H - I)^{-1}\|$  è molto grande.

Un altro criterio si fonda sulla (3.22) che, ponendo  $r^{(k)} = b - Ax^{(k)}$ , può scriversi

$$\frac{\|e^{(k)}\|}{\|a\|} \leq \mu(A) \frac{\|r^{(k)}\|}{\|b\|}$$

e suggerisce il criterio di arresto:

$$\frac{\|r^{(k)}\|}{\|b\|} \leq \epsilon. \quad (3.33)$$

La (3.33) comporta che l'errore relativo di  $x^{(k)}$  non superi in norma il numero  $\mu(A)\epsilon$ . Anche questo criterio è poco affidabile se  $A$  è molto malcondizionata. Comunque per garantire che l'algoritmo termini dopo un numero massimo  $N$  di iterazioni, si affianca ai criteri (3.32) o (3.33) l'ulteriore condizione che il calcolo si arresti allorché sia

$$k \geq N.$$

Si osservi che il teorema di convergenza 3.6.1 non tiene conto degli errori di arrotondamento, cioè vale nell'ipotesi ideale che le iterate siano esattamente quelle definite dalla (3.24). In realtà, indicando con  $\delta_k$  l'errore di arrotondamento che si commette ad ogni passo nel calcolo della funzione  $Hx^{(k)} + c$ , in luogo della (3.24) si dovrebbe scrivere

$$\tilde{x}^{(k+1)} = H\tilde{x}^{(k)} + c + \delta_k, \quad k = 0, 1, \dots$$

dove  $\{\tilde{x}^{(k)}\}$  è la successione effettivamente calcolata a partire da un arbitrario  $x^{(0)}$ .

Di conseguenza si può vedere che, in presenza di errori di arrotondamento, la convergenza del metodo nel senso del Teorema 3.6.1 non garantisce che l'errore effettivo tenda al vettore nullo. Tuttavia si può dire che in un metodo convergente l'effetto degli errori di arrotondamento sia abbastanza contenuto.

Questo giustifica l'uso del criterio di arresto  $\|\tilde{x}^{(k+1)} - \tilde{x}^{(k)}\| \leq \epsilon$  che in pratica si sostituisce alla (3.32).

### 3.7 Metodi di Jacobi e di Gauss-Seidel

Per definire due classici metodi iterativi si scomponga  $A$  nella forma

$$A = D - E - F \tag{3.34}$$

dove  $D = \text{diag}(a_{11}, a_{22}, \dots, a_{nn})$  mentre  $-E$  e  $-F$  sono matrici triangolari, rispettivamente inferiore e superiore, con la diagonale nulla.

Il sistema  $Ax - b = 0$  si può quindi scrivere

$$Dx = (E + F)x + b,$$

da cui, se  $a_{ii} \neq 0$ ,  $i = 1, 2, \dots, n$ , si ottiene il *metodo di Jacobi*

$$x^{(k+1)} = D^{-1}(E + F)x^{(k)} + D^{-1}b, \quad k = 0, 1, \dots, \tag{3.35}$$

la cui matrice di iterazione è  $H_J = D^{-1}(E + F)$ .

Le equazioni del sistema (3.35) sono date da

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left( b_i - \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij} x_j^{(k)} \right), \quad i = 1, 2, \dots, n, \quad k = 0, 1, \dots \tag{3.36}$$



Il vettore  $x^{(k+1)}$  ottenuto con l'algoritmo (3.36) viene prima memorizzato in una posizione distinta da quella occupata da  $x^{(k)}$  poi le  $n$  componenti  $x_i^{(k+1)}$  vengono trasferite simultaneamente nelle posizioni prima occupate dalle  $x_i^{(k)}$ . Per questo motivo il metodo è detto anche metodo delle *sostituzioni simultanee*.

Se si scrive il sistema dato nella forma equivalente

$$(D - E)x = Fx + b$$

e si suppone ancora che sia  $a_{ii} \neq 0$ ,  $i = 1, 2, \dots, n$ , si ottiene il *metodo di Gauss-Seidel*

$$x^{(k+1)} = (D - E)^{-1}Fx^{(k)} + (D - E)^{-1}b, \quad k = 0, 1, \dots, \quad (3.37)$$

dove la matrice di iterazione è data da

$$H_G = (D - E)^{-1}F.$$

Nel calcolo pratico si fa uso di una formulazione equivalente alla (3.37) e cioè

$$x^{(k+1)} = D^{-1}Ex^{(k+1)} + D^{-1}Fx^{(k)} + D^{-1}b, \quad k = 0, 1, \dots, \quad (3.38)$$

dove le singole equazioni sono

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij}x_j^{(k)} \right), \quad i = 1, 2, \dots, n, \quad k = 0, 1, \dots \quad (3.39)$$

L'algoritmo (3.39) consente una maggiore economia di memoria rispetto a quello di Jacobi, in quanto ogni singola componente  $x_i^{(k+1)}$  appena calcolata può essere subito memorizzata nella posizione prima occupata dalla vecchia componente  $x_i^{(k)}$ . Ciò giustifica la denominazione di *metodo delle sostituzioni successive* spesso usata per il processo (3.39).

Si osservi che in entrambi i metodi sopra definiti è necessaria la condizione  $a_{ii} \neq 0$ ,  $i = 1, 2, \dots, n$ . Se tale condizione non fosse verificata è sempre possibile ottenerla riordinando le equazioni e, eventualmente, anche le incognite, purché la matrice  $A$  non sia singolare. In generale, a diversi ordinamenti soddisfacenti la condizione  $a_{ii} \neq 0$ , corrispondono diverse matrici di iterazione.

Ammessi che sia verificata la condizione ora detta, un esame della matrice  $A$  del sistema può dire subito se vi sono condizioni sufficienti per la convergenza. Valgono infatti i seguenti teoremi.

**Teorema 3.7.1** *Se  $A$  è una matrice a predominanza diagonale forte, allora il metodo di Jacobi e quello di Gauss-Seidel sono convergenti (cfr. Esempio 3.10.5).*

**Teorema 3.7.2** *Se  $A$  è una matrice irriducibile e a predominanza diagonale debole, allora il metodo di Jacobi e quello di Gauss-Seidel sono convergenti (cfr. Esempio 3.10.6).*

In generale la convergenza di uno dei due metodi non implica quella dell'altro.

I metodi di Jacobi e di Gauss-Seidel possono essere usati anche nella versione a blocchi, con riferimento ad una partizione di  $A$  in sottomatrici  $A_{ij}$ ,  $i, j = 1, 2, \dots, m$ ,  $m < n$ , con le  $A_{ii}$  matrici quadrate non singolari (non necessariamente dello stesso ordine), cui deve corrispondere una ripartizione in  $m$  blocchi anche per i vettori  $x$  e  $b$ .

Il sistema  $Ax = b$  rappresentato a blocchi assume quindi la forma:

$$\begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \cdots & \cdots & \cdots \\ A_{m1} & \cdots & A_{mm} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix},$$

dove ora  $x_i$  e  $b_i$  sono vettori con un numero di componenti pari all'ordine di  $A_{ii}$ .

Alle equazioni a elementi (3.36) e (3.39) corrispondono rispettivamente le seguenti equazioni a blocchi:

$$x_i^{(k+1)} = A_{ii}^{-1} \left( b_i - \sum_{\substack{j=1 \\ j \neq i}}^m A_{ij} x_j^{(k)} \right), \quad i = 1, 2, \dots, m, \quad (3.40)$$

$$x_i^{(k+1)} = A_{ii}^{-1} \left( b_i - \sum_{j=1}^{i-1} A_{ij} x_j^{(k+1)} - \sum_{j=i+1}^n A_{ij} x_j^{(k)} \right), \quad i = 1, 2, \dots, m. \quad (3.41)$$

La convergenza dei metodi (3.40) e (3.41) è ora legata ai raggi spettrali delle corrispondenti matrici di iterazione che sono rispettivamente

$$\tilde{H}_J = \tilde{D}^{-1}(\tilde{E} + \tilde{F}),$$

$$\tilde{H}_G = (\tilde{D} - \tilde{E})^{-1} \tilde{F},$$

dove  $\tilde{D} = \text{diag}(A_{11}, A_{22}, \dots, A_{mm})$ , mentre  $-\tilde{E}$  e  $-\tilde{F}$  sono triangolari a blocchi con i blocchi diagonali nulli, seguendo la scomposizione (3.34).

Generalmente i metodi a blocchi vengono usati per sistemi con matrici di tipo speciale. Un caso del genere è contemplato nel seguente teorema.

**Teorema 3.7.3** *Se  $A$  è una matrice tridiagonale a blocchi, cioè con  $A_{ij} = \mathbf{O}$  per  $|i - j| > 1$ , e se si ha  $\det(A_{ii}) \neq 0$ ,  $i = 1, 2, \dots, m$ , i metodi a blocchi di Jacobi e di Gauss-Seidel convergono o divergono insieme, avendosi*

$$\rho(\tilde{H}_G) = \rho^2(\tilde{H}_J) .$$

In generale la convergenza di un metodo per punti, per un dato sistema, non implica quella dello stesso metodo usato a blocchi.

## 3.8 Metodi di rilassamento

Si consideri il metodo di Gauss-Seidel nella forma (3.38)

$$x^{(k+1)} = D^{-1}(Ex^{(k+1)} + Fx^{(k)} + b) ;$$

scrivendo

$$x^{(k+1)} = x^{(k)} + D^{-1}(Ex^{(k+1)} + Fx^{(k)} + b) - x^{(k)} ;$$

e ponendo

$$c^{(k)} := D^{-1}(Ex^{(k+1)} + Fx^{(k)} + b) - x^{(k)} \quad (3.42)$$

si ha

$$x^{(k+1)} = x^{(k)} + c^{(k)} ;$$

quindi ogni iterazione del metodo di Gauss-Seidel può pensarsi come una correzione del vettore  $x^{(k)}$  mediante un altro vettore  $c^{(k)}$  per ottenere  $x^{(k+1)}$ .

Questa interpretazione suggerisce di introdurre una correzione del tipo  $\omega c^{(k)}$  dove  $\omega$  è un parametro reale, che, opportunamente scelto, può servire ad accelerare la convergenza del metodo. Si ha così il *metodo di rilassamento* definito da

$$x^{(k+1)} = x^{(k)} + \omega c^{(k)} , \quad k = 0, 1, \dots ,$$

ossia, tenendo conto della (3.42),

$$x^{(k+1)} = (1 - \omega)x^{(k)} + \omega D^{-1}(Ex^{(k+1)} + Fx^{(k)} + b) , \quad k = 0, 1, \dots , \quad (3.43)$$

dove le equazioni a elementi sono

$$x_i^{(k+1)} = (1 - \omega)x_i^{(k)} + \frac{\omega}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij}x_j^{(k)} \right), \\ i = 1, 2, \dots, n, \quad k = 0, 1, \dots$$

La matrice di iterazione  $H_\omega$  del metodo di rilassamento si ottiene subito scrivendo la (3.43) nella forma

$$x^{(k+1)} = (D - \omega E)^{-1}[(1 - \omega)D + \omega F]x^{(k)} + \omega(D - \omega E)^{-1}b$$

da cui

$$H_\omega = (D - \omega E)^{-1}[(1 - \omega)D + \omega F].$$

Si dimostra che per la convergenza del metodo di rilassamento è necessario scegliere  $\omega$  in modo che sia

$$0 < \omega < 2. \quad (3.44)$$

Nel caso speciale di matrice  $A$  hermitiana definita positiva, si può dimostrare che la (3.44) è anche condizione sufficiente per la convergenza del metodo.

Naturalmente per  $\omega = 1$  si ottiene il metodo di Gauss-Seidel. Anche il metodo di rilassamento può essere impiegato nella versione a blocchi, che consente, in qualche caso particolare, una scelta ottimale di  $\omega$  (cfr. 3.10.6).

### 3.9 Metodo del gradiente coniugato

Sia  $A \in \mathbb{R}^{n \times n}$  simmetrica e definita positiva ed  $a$  la soluzione del sistema lineare

$$Ax = b; \quad (3.45)$$

se si considera il funzionale

$$\varphi(x) = \frac{1}{2}(b - Ax)^T A^{-1}(b - Ax) \quad (3.46)$$

si ha evidentemente  $\varphi(a) = 0$  mentre, essendo  $A^{-1}$  simmetrica e definita positiva, risulta  $\varphi(x) > 0$  per ogni vettore reale  $x \neq a$ .

La risoluzione del sistema (3.45) è quindi un problema equivalente a quello della ricerca del punto di minimo in  $\mathbb{R}^n$  per il funzionale  $\varphi(x)$ .

Dalla (3.46) si ricava

$$\varphi(x) = \frac{1}{2}x^T Ax - x^T b + \frac{1}{2}b^T A^{-1}b;$$

quindi minimizzare  $\varphi(x)$  equivale a minimizzare

$$F(x) = \frac{1}{2}x^T Ax - x^T b,$$

che differisce da  $\varphi(x)$  per una costante.

Si noti la relazione

$$\text{grad } \varphi(x) = \text{grad } F(x) = Ax - b = -r(x),$$

dove il vettore  $r(x)$  è il residuo del sistema (3.45) (cfr. 3.5).

Vari metodi numerici per il calcolo di  $a$  consistono nel costruire una successione  $\{x^{(k)}\}$  a cui corrisponda una successione  $\{F(x^{(k)})\}$  che sia decrescente. Il più semplice di questi metodi, ideato da Cauchy, è quello della *discesa più ripida*, che, partendo da  $x^{(0)}$  arbitrario, produce la successione

$$x^{(k+1)} = x^{(k)} + \lambda_k d^{(k)} \quad (3.47)$$

dove  $d^{(k)}$  è un vettore orientato nel senso di massima decrescenza di  $F(x)$  in  $x_k$  e  $\lambda_k$  è un valore reale che minimizza la funzione  $F(x^{(k)} + \lambda d^{(k)})$  della sola variabile  $\lambda$ . In pratica si pone

$$d^{(k)} = -\text{grad } F(x)_{x=x^{(k)}} = r(x^{(k)}). \quad (3.48)$$

La (3.47) si interpreta geometricamente come il passaggio dal punto  $x^{(k)}$  (tale che  $r(x^{(k)}) \neq 0$ ) al punto  $x^{(k+1)}$  lungo la retta passante per  $x^{(k)}$  e parallela a  $d^{(k)}$ . La convergenza della successione (3.47) alla soluzione  $a$  può essere molto lenta se  $A$  è malcondizionata.

Una variante di notevole importanza prende il nome di *metodo del gradiente coniugato*.

Anche questo metodo è espresso formalmente dalla (3.47) ma la scelta del vettore  $d^{(k)}$  è diversa dalla (3.48) e consente di migliorare la convergenza rispetto al metodo della discesa più ripida. Per  $x^{(0)}$  arbitrario si pone ora

$$\begin{aligned} d^{(0)} &= r^{(0)} = b - Ax^{(0)} \\ d^{(k)} &= r^{(k)} + \rho_k d^{(k-1)}, \quad r^{(k)} = r(x^{(k)}), \quad k \geq 1, \end{aligned} \quad (3.49)$$

dove il numero  $\rho_k$  si calcola in modo che il vettore  $d^{(k)}$  risulti *coniugato di*  $d^{(k-1)}$  *rispetto ad*  $A$  cioè sia

$$\left(d^{(k)}\right)^T A d^{(k-1)} = 0 ;$$

questa condizione permette di ricavare per  $\rho_k$  la seguente espressione

$$\rho_k = \frac{\left(r^{(k)}\right)^T r^{(k)}}{\left(r^{(k-1)}\right)^T r^{(k-1)}} , \quad k \geq 1 . \quad (3.50)$$

Calcolato  $d^{(k)}$  in base alle (3.49), (3.50),  $F(x^{(k)} + \lambda d^{(k)})$  risulta minima per  $\lambda = \lambda_k$ , dove

$$\lambda_k = \frac{\left(r^{(k)}\right)^T r^{(k)}}{\left(d^{(k)}\right)^T A d^{(k)}} . \quad (3.51)$$

In questo metodo il verso del vettore  $d^{(k)}$  coincide con quello di massima decrescenza di  $F(x)$  solo per  $k = 0$ , tuttavia si dimostra che anche per  $k \geq 1$  gli spostamenti di  $x^{(k)}$  avvengono lungo rette orientate nel senso in cui  $F(x)$  decresce e che si raggiunge la soluzione  $a$  in un numero  $p \leq n$  di passi. In realtà la presenza degli errori di arrotondamento consente solo di approssimare  $a$ , cioè il metodo del gradiente coniugato finisce per assumere un comportamento simile a quello di un metodo iterativo.

Il metodo può essere usato anche per un sistema con matrice dei coefficienti  $A \in \mathbb{R}^{n \times n}$  non simmetrica applicandolo al sistema equivalente  $A^T A x = A^T b$  la cui matrice dei coefficienti  $A^T A$  risulta simmetrica e definita positiva. Va detto che il sistema così trasformato ha in genere un condizionamento peggiore di quello iniziale. È possibile però effettuare ulteriori trasformazioni in modo da ridurre il malcondizionamento del sistema. Le tecniche numeriche ideate a questo scopo, sulle quali non ci si sofferma, si dicono *metodi di preconditionamento* e vengono spesso associate al metodo del gradiente coniugato.

## 3.10 Complementi ed esempi

### 3.10.1 Il metodo di Gauss-Jordan

Una variante del metodo di Gauss è il *metodo di Gauss-Jordan*. Esso consiste nell'operare sulla matrice dei coefficienti del sistema (3.1) delle com-

binazioni tra le righe in modo da ottenere un sistema lineare equivalente la cui matrice dei coefficienti sia diagonale.

Per fare ciò, basta effettuare, dal secondo passo in poi, le combinazioni lineari opportune anche con le righe che precedono la riga a cui appartiene l'elemento pivotale. In altre parole, al passo  $i$ -esimo del metodo di Gauss si elimina l'incognita  $x_i$  da tutte le equazioni esclusa l' $i$ -esima.

Il risultato finale è un sistema del tipo

$$Dx = b'$$

dove  $D$  è una matrice diagonale.

Come per il metodo base di Gauss, è possibile che uno o più elementi pivotali risultino nulli. Non si presenta questo caso se e solo se valgono le condizioni (3.9) che assicurano l'applicabilità del metodo senza dover ricorrere a scambi di righe.

Per ridurre la propagazione degli errori di arrotondamento si ricorre ai criteri esposti in 3.2.

L'applicazione del metodo di Gauss-Jordan a un sistema di ordine  $n$  comporta un costo computazionale di  $\frac{n^3}{2} + n^2 - \frac{n}{2}$  operazioni, cioè superiore a quello del metodo di Gauss.

### 3.10.2 Calcolo della matrice inversa

Come accennato in 3.1, data una matrice  $A$  di ordine  $n$  non singolare, la sua matrice inversa  $A^{-1}$  è la soluzione del sistema matriciale

$$AX = I. \quad (3.52)$$

Si tratta quindi di risolvere  $n$  sistemi lineari  $Ax^{(i)} = e^{(i)}$ ,  $i = 1, 2, \dots, n$ , dove  $x^{(i)}$  e  $e^{(i)}$  sono la  $i$ -esima colonna, rispettivamente, della matrice  $X$  e della matrice  $I$ . Tali sistemi vengono risolti simultaneamente considerando la matrice completa  $(A \mid I)$  ed effettuando su di essa le operazioni di eliminazione gaussiana.

Gli eventuali scambi di righe dovute ad un eventuale pivoting parziale, non comportano variazioni nella soluzione  $X$  del sistema lineare (3.52) che rimane la matrice inversa di  $A$ . Infatti il sistema effettivamente risolto in tal caso è della forma  $PAX = PI$ , da cui segue  $X = A^{-1}$ . Se invece si effettua uno scambio di colonne sulla matrice di (3.52), ciò equivale a risolvere un sistema della forma  $APX = I$ , da cui si ha  $X = P^{-1}A^{-1}$  e infine  $A^{-1} = PX$ .

**Osservazione 3.10.1** Le considerazioni fatte nel caso della risoluzione del sistema (3.52) permettono la risoluzione simultanea di più sistemi lineari di uguale matrice  $A$  come un solo sistema matriciale del tipo  $AX = B$ , dove le colonne di  $B$  sono i vettori dei termini noti di tutti i sistemi dati.

**Osservazione 3.10.2** Analogamente ai metodi iterativi, anche i metodi diretti di Gauss e Gauss-Jordan possono essere usati nella versione a blocchi (cfr. Esempio 3.10.1). In questo caso, la forma dei moltiplicatori usati per le combinazioni lineari tra le righe differisce dalla (3.5) in quanto si opera tra sottomatrici e, per esempio al primo passo, si ha

$$L_{i1} = A_{i1}A_{11}^{-1}.$$

Nella versione a blocchi, i moltiplicatori devono essere usati come *premultiplcatori* perché così facendo si operano combinazioni lineari tra le equazioni del sistema senza alterarne la soluzione, mentre la postmoltiplicazione comporterebbe una combinazione tra le colonne e quindi una alterazione della soluzione.

**Esempio 3.10.1** Sia  $A \in \mathbb{R}^{n \times n}$  non singolare e partizionata a blocchi nel seguente modo:

$$A = \begin{pmatrix} B & u \\ u^T & c \end{pmatrix}, \quad B \in \mathbb{R}^{(n-1) \times (n-1)}, \quad u \in \mathbb{R}^{n-1}, \quad c \in \mathbb{R}.$$

Nell'ipotesi che  $B$  sia invertibile, si può esprimere l'inversa di  $A$  operando sui blocchi anziché sugli elementi. Applicando il metodo di Gauss-Jordan a blocchi al sistema  $AX = I$ , si considera la matrice completa  $(A \mid I)$ , con  $I$  partizionata a blocchi coerentemente con  $A$ :

$$(A \mid I) = \left( \begin{array}{cc|cc} B & u & I_{n-1} & 0 \\ u^T & c & 0 & 1 \end{array} \right).$$

Premoltiplicando la prima riga per  $B^{-1}$  si ottiene la matrice

$$\left( \begin{array}{cc|cc} I_{n-1} & B^{-1}u & B^{-1} & 0 \\ u^T & c & 0 & 1 \end{array} \right).$$



Premoltiplicando la prima riga per  $u^T$  e sottraendola dalla seconda si ha

$$\left( \begin{array}{cc|cc} I_{n-1} & B^{-1}u & B^{-1} & 0 \\ 0 & c - u^T B^{-1}u & -u^T B^{-1} & 1 \end{array} \right).$$

Posto  $\gamma = (c - u^T B^{-1}u)^{-1}$ , si ottiene

$$\left( \begin{array}{cc|cc} I_{n-1} & B^{-1}u & B^{-1} & 0 \\ 0 & 1 & -\gamma u^T B^{-1} & \gamma \end{array} \right),$$

da cui, infine, sottraendo dalla prima riga la seconda premoltiplicata per  $B^{-1}u$ , si ha

$$\left( \begin{array}{cc|cc} I_{n-1} & 0 & B^{-1} + \gamma B^{-1}uu^T B^{-1} & -\gamma B^{-1}u \\ 0 & 1 & -\gamma u^T B^{-1} & \gamma \end{array} \right).$$

Il sistema  $AX = I$  è così diventato della forma

$$IX = S, \quad \text{dove} \quad S = \left( \begin{array}{cc} B^{-1} + \gamma B^{-1}uu^T B^{-1} & -\gamma B^{-1}u \\ -\gamma u^T B^{-1} & \gamma \end{array} \right).$$

Si deduce quindi  $S = A^{-1}$ . □

### 3.10.3 Fattorizzazione $LL^T$

Una matrice  $A$  reale e definita positiva è fattorizzabile nella forma  $LL^T$  con  $L$  matrice triangolare inferiore (cfr. 3.4).

**Esempio 3.10.2** Sia

$$A = \begin{pmatrix} 1 & t & 0 \\ t & 1 & t \\ 0 & t & 1 \end{pmatrix}, \quad t \in \mathbb{R}.$$

Gli autovalori della matrice  $A$  sono

$$\lambda_1 = 1, \quad \lambda_2 = 1 + t\sqrt{2}, \quad \lambda_3 = 1 - t\sqrt{2}.$$

I tre autovalori risultano positivi se  $|t| < \frac{\sqrt{2}}{2}$  e quindi per questi valori di  $t$ , per il Teorema 2.11.2, la matrice  $A$  è definita positiva.

Applicando il metodo di Cholesky si ottiene:

$$a_{11} = 1, \quad a_{22} = 1, \quad a_{33} = 1,$$

da cui

$$\begin{aligned} l_{11}^2 &= a_{11} = 1, \\ l_{22}^2 &= a_{22} - l_{21}^2 = 1 - t^2, \\ l_{33}^2 &= a_{33} - l_{31}^2 - l_{32}^2 = \frac{1-2t^2}{1-t^2}. \end{aligned}$$

Risulta infine

$$L = \begin{pmatrix} 1 & 0 & 0 \\ t & \sqrt{1-t^2} & 0 \\ 0 & \frac{t}{\sqrt{1-t^2}} & \sqrt{\frac{1-2t^2}{1-t^2}} \end{pmatrix}.$$

□

### 3.10.4 Sistemi malcondizionati

**Esempio 3.10.3** Si consideri il sistema  $Ax = b$  con

$$A = \begin{pmatrix} 1 & 1 \\ 0.999 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 2 \\ 1.999 \end{pmatrix}.$$

La soluzione è  $x^T = (1, 1)$  e si verifica che risulta, in norma infinito,  $\mu(A) = 4 \times 10^3$ .

Si supponga di perturbare la matrice  $A$  con la matrice

$$\delta A = 0.00024 \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}.$$

Ne risulta evidentemente  $\|\delta A\|/\|A\| = 24 \times 10^{-5}$ . In questo caso il fattore di amplificazione nella (3.20) vale

$$\frac{\mu(A)}{1 - \mu(A) \frac{\|\delta A\|}{\|A\|}} = 10^5,$$

da cui la limitazione

$$\frac{\|\delta x\|}{\|x\|} \leq 24$$

per l'errore relativo della soluzione perturbata. In effetti, risolvendo il sistema

$$(A + \delta A)(x + \delta x) = b$$

si trova

$$x + \delta x = \begin{pmatrix} 0.04023 \dots \\ 1.9593 \dots \end{pmatrix}$$

a cui corrisponde  $\|\delta x\|/\|x\| \simeq 0.96$ .

Si noti, quindi, che ad una variazione  $\|\delta A\|$  pari al 0.024% di  $\|A\|$  corrisponde una variazione della soluzione pari a circa il 96%.  $\square$

**Esempio 3.10.4** Si consideri la *matrice di Hilbert* del quarto ordine

$$A = \begin{pmatrix} 1 & 1/2 & 1/3 & 1/4 \\ 1/2 & 1/3 & 1/4 & 1/5 \\ 1/3 & 1/4 & 1/5 & 1/6 \\ 1/4 & 1/5 & 1/6 & 1/7 \end{pmatrix}$$

e il vettore  $b^T = (1, 1, 1, 1)$ . Il sistema  $Ax = b$  ha soluzione  $x^T = (-4, 60, -180, 140)$ . Si può verificare che il numero di condizione di  $A$ , in norma euclidea, risulta

$$\mu(A) \simeq 1.55 \times 10^4.$$

Si perturbi il vettore  $b$  con

$$\delta b = (\epsilon, -\epsilon, \epsilon, -\epsilon)^T$$

essendo  $\epsilon$  un numero positivo arbitrario.

Per la (3.20) il fattore di amplificazione dell'errore relativo coincide con  $\mu(A)$ .

Risolvendo il sistema perturbato

$$A(x + \delta x) = b + \delta b$$

si ottiene

$$(x + \delta x)^T = (-4 + 516\epsilon, 60 - 5700\epsilon, -180 + 13620\epsilon, 140 - 8820\epsilon),$$

da cui  $\|\delta x\|/\|x\| \simeq 73\epsilon$ .

Pertanto, se  $\epsilon = 0.01$ , si può affermare che una perturbazione pari a 1% del vettore  $b$  induce una variazione superiore al 70% del vettore soluzione.  $\square$

### 3.10.5 Metodi iterativi

Gli esempi che seguono sono una applicazione dei Teoremi 3.7.1 e 3.7.2.

**Esempio 3.10.5** È dato il sistema lineare  $Ax = b$  con

$$A = \begin{pmatrix} 4 & 1 & 1 & 1 \\ 1 & 4 & 1 & 1 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 4 \end{pmatrix}, \quad b = \begin{pmatrix} 6 \\ 6 \\ 8 \\ -2 \end{pmatrix}.$$

La predominanza diagonale forte di  $A$  garantisce la convergenza dei metodi iterativi di Jacobi e di Gauss-Seidel. Infatti, si ottengono, rispettivamente, le matrici di iterazione

$$H_J = -\frac{1}{4} \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

$$H_G = -\frac{1}{256} \begin{pmatrix} 0 & 64 & 64 & 64 \\ 0 & -16 & 48 & 48 \\ 0 & 4 & -12 & 52 \\ 0 & -1 & 3 & -13 \end{pmatrix}.$$

Si verifica immediatamente che  $\|H_J\|_\infty = 0.75$  e  $\|H_G\|_\infty = 0.75$ , per cui, per il Corollario 3.6.2, i metodi risultano convergenti.  $\square$

In generale si può dimostrare che per ogni matrice con predominanza diagonale forte le corrispondenti matrici di iterazione  $H_J$  e  $H_G$  sono tali che  $\|H_J\|_\infty < 1$  e  $\|H_G\|_\infty < 1$ . In ciò consiste la dimostrazione del Teorema 3.7.1.

**Esempio 3.10.6** È dato il sistema lineare  $Ax = b$  con

$$A = \begin{pmatrix} 3 & 2 & 1 \\ 1 & 1 & 0 \\ -1 & 1 & 3 \end{pmatrix}, \quad b = \begin{pmatrix} -3 \\ -2 \\ 6 \end{pmatrix}.$$

Si osservi che  $A$  è a predominanza diagonale debole e irriducibile.

Per i metodi iterativi di Jacobi e di Gauss-Seidel si ottengono, rispettivamente, le matrici di iterazione

$$H_J = -\frac{1}{3} \begin{pmatrix} 0 & 2 & 1 \\ 3 & 0 & 0 \\ -1 & 1 & 0 \end{pmatrix},$$

$$H_G = -\frac{1}{9} \begin{pmatrix} 0 & 6 & 3 \\ 0 & -6 & -3 \\ 0 & 4 & 2 \end{pmatrix}.$$

Tali matrici risultano anch'esse irriducibili. Analizzando i rispettivi cerchi di Gershgorin, si osserva che, per il Teorema 2.8.3, gli autovalori di  $H_J$  e  $H_G$  hanno modulo minore di 1, e quindi i due metodi sono convergenti.  $\square$

Il ragionamento qui seguito serve in generale a dimostrare il Teorema 3.7.2.

Si mostra, ora, come, in generale, i metodi di Jacobi e Gauss-Seidel possano non convergere contemporaneamente.

**Esempio 3.10.7** Sia dato il sistema lineare  $Ax = b$  con

$$A = \begin{pmatrix} 1 & 1 & -2 \\ 2 & 1 & 2 \\ 2 & 1 & 1 \end{pmatrix}.$$

Le matrici di iterazione di Jacobi e di Gauss-Seidel sono

$$H_J = \begin{pmatrix} 0 & -1 & 2 \\ -2 & 0 & -2 \\ -2 & -1 & 0 \end{pmatrix}, \quad H_G = \begin{pmatrix} 0 & -1 & 2 \\ 0 & 2 & -6 \\ 0 & 0 & 2 \end{pmatrix}.$$

Si verifica che  $\rho(H_J) = 0$  per cui il metodo di Jacobi è convergente, mentre  $\rho(H_G) = 2$  per cui il metodo di Gauss-Seidel risulta divergente.

Per contro, dato il sistema lineare  $Ax = b$  con

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix},$$

per le matrici di iterazione di Jacobi e di Gauss-Seidel

$$H_J = \begin{pmatrix} 0 & 0 & -1 \\ -1 & 0 & 0 \\ -1 & -1 & 0 \end{pmatrix}, \quad H_G = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix},$$

si verifica che  $\rho(H_J) > 1$  e  $\rho(H_G) = 0$ . □

**Esempio 3.10.8** Come esempio illustrativo del Teorema 3.7.3, si consideri la matrice

$$A = \begin{pmatrix} I & B \\ B & I \end{pmatrix}$$

dove  $B \in \mathbb{R}^{n \times n}$  e  $I$  è la matrice identica di ordine  $n$ .

Le matrici di iterazione di Jacobi e di Gauss-Seidel sono

$$H_J = \begin{pmatrix} \mathbf{O} & -B \\ -B & \mathbf{O} \end{pmatrix}, \quad H_G = \begin{pmatrix} \mathbf{O} & -B \\ \mathbf{O} & B^2 \end{pmatrix}.$$

Sia  $\lambda$  un autovalore di  $H_J$  e  $y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$  un autovettore ad esso associato. Dall'equazione  $H_J y = \lambda y$  si ha

$$\begin{cases} -By_2 &= \lambda y_1 \\ -By_1 &= \lambda y_2 \end{cases} \quad (3.53)$$

che può scriversi anche

$$\begin{cases} -By_2 &= -\lambda(-y_1) \\ -B(-y_1) &= -\lambda y_2 \end{cases}.$$

Si deduce che  $-\lambda$  è autovalore della matrice  $H_J$  con autovettore  $y = \begin{pmatrix} -y_1 \\ y_2 \end{pmatrix}$ .

Sia  $\mu$  un autovalore non nullo della matrice  $H_G$  e  $z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$  un autovettore associato a  $\mu$ . Dalla relazione  $H_G z = \mu z$  si ha

$$\begin{cases} -Bz_2 &= \mu z_1 \\ B^2 z_2 &= \mu z_2 \end{cases} \quad \text{ovvero} \quad \begin{cases} -Bz_2 &= \mu z_1 \\ -B(\mu z_1) &= \mu z_2 \end{cases}$$

da cui

$$\begin{cases} -Bz_2 &= \sqrt{\mu}(\sqrt{\mu}z_1) \\ -B(\sqrt{\mu}z_1) &= \sqrt{\mu}z_2. \end{cases}$$

Dal confronto con (3.53) si evidenzia come  $\sqrt{\mu}$  sia autovalore della matrice  $H_J$  con autovettore  $\begin{pmatrix} \sqrt{\mu}z_1 \\ z_2 \end{pmatrix}$ . Segue quindi  $\mu = \lambda^2$  conforme alla tesi del teorema citato.  $\square$

Nei metodi iterativi, la scelta del vettore iniziale  $x^{(0)}$  non è soggetta a particolari condizioni. Ciò non esclude che una buona scelta per  $x^{(0)}$  riduca il numero delle iterazioni necessarie per ottenere una data accuratezza. Ad esempio, se la matrice  $A$  ha una qualunque predominanza diagonale una buona scelta è  $x_i^{(0)} = b_i/a_{ii}$ ,  $i = 1, 2, \dots, n$ .

### 3.10.6 Scelta ottimale del parametro $\omega$

Dalla scelta del parametro  $\omega$  dipendono la convergenza e la velocità asintotica di convergenza del metodo di rilassamento.

Per fare un esempio si considera il caso particolare di un sistema con matrice reale e definita positiva.

**Esempio 3.10.9** Sia dato il sistema lineare  $Ax = b$  con

$$A = \begin{pmatrix} 2 & -2 & 0 \\ -2 & 3 & -1 \\ 0 & -1 & 4 \end{pmatrix}. \quad (3.54)$$

Nella Fig. 3.3 è riportata la variazione del raggio spettrale della matrice di iterazione  $H_\omega = (D - \omega E)^{-1}[(1 - \omega)D + \omega F]$  in funzione del parametro  $\omega$  nell'intervallo  $]0, 2[$ .

Il minimo  $\rho(H_{\omega^*}) \simeq 0.3334$  si ottiene per  $\omega^* \simeq 1.3334$ . Con tale valore di  $\omega$ , il metodo di rilassamento applicato al sistema dato ha velocità di convergenza massima.  $\square$

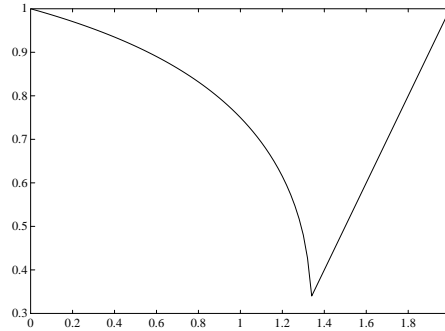


Figura 3.3: Grafico di  $\rho(H_\omega)$  per la (3.54).

Se non si dispone del valore esatto di  $\omega^*$  è evidente dalla Fig. 3.3 che conviene preferire una sua approssimazione per eccesso ad una per difetto.

Nel caso di matrici tridiagonali e tridiagonali a blocchi non è pratico costruire caso per caso la funzione  $\rho(H_\omega)$  come nell'esempio sopra considerato; in effetti, in questi casi, esistono formule che danno una stima immediata del valore ottimale di  $\omega$ .

### 3.10.7 L'algoritmo del metodo del gradiente coniugato

Dato un sistema lineare  $Ax = b$  con  $A$  matrice reale e definita positiva, scelto il vettore iniziale  $x^{(0)}$ , l'algoritmo del metodo del gradiente coniugato si può descrivere come segue:

1. Calcolo del vettore residuo  $r^{(k)} = b - Ax^{(k)}$  per  $k = 0$ ;
2. se  $r^{(k)} = 0$  si arresta il calcolo;
3. si calcola il numero reale  $\rho_k$  dato dalla (3.50) (per  $k = 0$  si pone  $\rho_0 = 0$ );
4. si calcola il vettore  $d^{(k)} = r^{(k)} + \rho_k d^{(k-1)}$ ;
5. si calcola il numero reale  $\lambda_k$  dato dalla (3.51);
6. si calcolano i vettori  $x^{(k+1)} = x^{(k)} + \lambda_k d^{(k)}$  e  $r^{(k+1)} = b - Ax^{(k+1)}$ ;
7. si pone  $k := k + 1$  e si riparte dal punto 2.



In pratica, il calcolo si arresta, per esempio, quando una norma del vettore  $r^{(k)}$  risulta minore di un valore prefissato. Poiché al punto 3 si calcola il prodotto scalare  $(r^{(k)})^T r^{(k)}$ , è conveniente usare il seguente criterio di arresto

$$\|r^{(k)}\|_2 < \epsilon \|b\|_2. \quad (3.55)$$

**Esempio 3.10.10** Si consideri il sistema lineare  $Ax = b$  dove

$$A = \begin{pmatrix} 4 & -1 & -1 & 0 & 0 & 0 \\ -1 & 4 & -1 & -1 & 0 & 0 \\ -1 & -1 & 4 & -1 & -1 & 0 \\ 0 & -1 & -1 & 4 & -1 & -1 \\ 0 & 0 & -1 & -1 & 4 & -1 \\ 0 & 0 & 0 & -1 & -1 & 4 \end{pmatrix}, \quad b = \begin{pmatrix} 19/6 \\ 5/12 \\ -37/60 \\ -1/5 \\ 1/20 \\ 13/60 \end{pmatrix},$$

la cui soluzione è  $a^T = (1, 1/2, 1/3, 1/4, 1/5, 1/6)$ .

Applicando il metodo del gradiente coniugato con vettore iniziale  $x^{(0)}$  di componenti  $x_i^{(0)} = b_i/a_{ii}$  e usando il criterio di arresto (3.55) con  $\epsilon = 10^{-8}$ , il processo iterativo si arresta alla quinta iterazione e si ha

$$x^{(5)} = \begin{pmatrix} 1.0000046 \dots \\ 0.4999859 \dots \\ 0.3333464 \dots \\ 0.2499855 \dots \\ 0.2000196 \dots \\ 0.1666564 \dots \end{pmatrix}.$$

□

**Bibliografia:** [2], [4], [14], [28], [31]

