Capitolo 1

Teoria degli errori

1.1 Rappresentazione dei numeri

Scelto un qualunque numero intero $\beta > 1$, ogni numero non nullo $x \in \mathbb{R}$ ammette una *rappresentazione in base* β

$$x = sign(x)\, \beta^b \sum_{i=1}^{\infty} \alpha_i \beta^{-i}\,,$$

dove $b \in \mathbb{Z}$ e ogni α_i è un intero tale che $0 \leq \alpha_i \leq \beta - 1$, $i = 1, 2, \ldots$, ($sign(x)$ è definita come una funzione che assume valore 1 se $x > 0$ e valore -1 se $x < 0$). Il numero β dicesi appunto la *base*, b l'*esponente* e i numeri α_i si dicono le *cifre* della rappresentazione.

Vale il seguente teorema.

Teorema 1.1.1 (di rappresentazione) *Data una base intera $\beta > 1$ e un qualunque numero reale x diverso da zero, esiste un'unica rappresentazione in base β tale che:*

1. *sia $\alpha_1 \neq 0$,*

2. *non vi sia un intero k per cui si abbia $\alpha_j = \beta - 1$, $\forall j > k$.*

La rappresentazione definita dal teorema precedente dicesi *rappresentazione in virgola mobile normalizzata* del numero reale x.

Fissato β, sono quindi univocamente determinati i numeri b e α_i, $i = 1, 2, \ldots$, della rappresentazione normalizzata e la serie $\sum_{i=1}^{\infty} \alpha_i \beta^{-i}$ è detta

mantissa del numero x; è immediato verificare che

$$\frac{1}{\beta} \leq \sum_{i=1}^{\infty} \alpha_i \beta^{-i} < 1.$$

All'interno di un calcolatore si possono rappresentare solo un certo numero m di cifre della mantissa di x. La rappresentazione sul calcolatore corrisponde perciò al numero $y = tr(x)$ oppure a $y = rd(x)$ dove $tr(x)$ ed $rd(x)$ sono approssimazioni di x, definite dai seguenti due criteri

$$y = tr(x) = sign(x)\, \beta^b \sum_{i=1}^{m} \alpha_i \beta^{-i}, \tag{1.1}$$

$$y = rd(x) = \begin{cases} tr(x) & \text{se } 0 \leq \alpha_{m+1} < \frac{\beta}{2} \\ sign(x)\, \beta^b \left[\sum_{i=1}^{m} \alpha_i \beta^{-i} + \beta^{-m}\right] & \text{se } \frac{\beta}{2} \leq \alpha_{m+1} < \beta \end{cases}. \tag{1.2}$$

Nel caso particolare in cui risulti $\alpha_1 = \alpha_2 = \cdots = \alpha_m = \beta - 1$ e $\alpha_{m+1} \geq \beta/2$, nella seconda delle (1.2) la normalizzazione porta ad aumentare di una unità l'esponente b. Per esempio, con $\beta = 10$ e $m = 3$, se $x = 0.9997 \times 10^5$ si ha $rd(x) = 0.100 \times 10^6$.

La (1.1) si chiama rappresentazione per *troncamento* del numero reale x mentre la (1.2) fornisce la rappresentazione per *arrotondamento*. Si può dimostrare che $\mid tr(x) - x \mid < \beta^{b-m}$ e $\mid rd(x) - x \mid \leq \frac{1}{2}\beta^{b-m}$, per cui, tra le due, la rappresentazione per arrotondamento è, in generale, una migliore approssimazione del numero reale x.

Si indichi con M l'insieme dei numeri z rappresentabili all'interno di un calcolatore, comunemente chiamati *numeri di macchina*.

M è un insieme finito; infatti, fissati β ed m e supposto $L \leq b \leq U$ $(L, U \in \mathbb{Z})$, la cardinalità di M risulta $2(\beta^m - \beta^{m-1})(U - L + 1) + 1$. Spesso l'insieme M viene indicato con il simbolo $F(\beta, m, L, U)$ per meglio evidenziare le caratteristiche della macchina.

Dato un qualunque numero reale $x \neq 0$, non è assicurata l'esistenza di $rd(x)$ fra i numeri di macchina.

Sia, per esempio, $F(10, 3, -99, 99)$ e $x = 0.9998 \times 10^{99}$; si ha $rd(x) = 0.1 \times 10^{100}$ che non rientra nell'insieme dei numeri di macchina considerato. In questo caso si ha una situazione di *overflow* e cioè il numero da rappresentare è "troppo grande" e non appartiene a M (tutti i calcolatori segnalano il presentarsi di un overflow, alcuni arrestano l'esecuzione del programma, altri proseguono con $rd(x) = sign(x) \max_{y \in M} \mid y \mid$).

Per contro, si abbia il numero $x = 0.01 \times 10^{-99}$; risulta $rd(x) = 0.1 \times 10^{-100}$ che non è un numero di macchina. In questo caso si ha una situazione di *underflow*, cioè il numero da rappresentare è "troppo piccolo" e non appartiene a M (non tutti i calcolatori segnalano questa situazione e nel caso in cui proseguano l'esecuzione del programma pongono $rd(x) = 0$).

Si dimostra che $rd(x)$ soddisfa la relazione

$$| rd(x) - x | \leq | z - x |, \quad \forall z \in M,$$

per cui, se $rd(x) \in M$, esso, in valore assoluto, differisce da x meno di qualunque altro numero di macchina.

Si definisce *errore assoluto* della rappresentazione del numero reale x il valore

$$\delta_x = rd(x) - x$$

ed *errore relativo* il valore

$$\epsilon_x = \frac{rd(x) - x}{x} = \frac{\delta_x}{x}.$$

Dalle precedenti considerazioni si ricavano le limitazioni

$$| \delta_x | \leq \frac{1}{2}\beta^{b-m},$$

$$| \epsilon_x | < \frac{1}{2}\beta^{1-m}.$$

Il numero $u = \frac{1}{2}\beta^{1-m}$ si dice *precisione di macchina*. In generale quando l'errore relativo di una approssimazione non supera $\frac{1}{2}\beta^{1-k}$ si dice che l'approssimazione è corretta almeno fino alla k-esima cifra significativa. Da quanto sopra segue quindi che $rd(x)$ approssima x almeno fino alla m-esima cifra significativa.

Se si assume la base $\beta = 2$ ogni cifra della rappresentazione di un numero ha valore 0 o 1 e si dice *bit* [1] (binary digit), mentre si dice *parola* di lunghezza t l'insieme dei t bit rappresentanti un numero. Per ragioni tecniche si usa spesso una base $\beta = 2^n$ $(n > 1)$ in cui ciascuna cifra, tradotta in rappresentazione binaria, richiede n bit.

Si distinguono la rappresentazione dei numeri interi da quella dei reali e l'aritmetica che opera solo con numeri interi da quella che opera indifferentemente con numeri interi e reali.

In 1.1.1, 1.1.2, 1.1.3, si farà riferimento al caso $t = 32$.

[1] Il termine bit è usato nel seguito in forma invariata anche al plurale.

1.1.1 Numeri interi

Per rappresentare un numero intero i 32 bit della parola vengono così utilizzati: un bit fornisce il segno del numero (per esempio 0 se il numero è positivo e 1 se è negativo), i restanti 31 bit servono a rappresentare in base 2 le cifre del numero dato. E' evidente che il massimo intero rappresentabile è, in valore assoluto, $2^{31} - 1 = 2147483647$.

1.1.2 Numeri reali (precisione semplice)

Nella rappresentazione dei numeri reali una situazione caratteristica è la seguente: la base è 16 (sistema esadecimale), un bit è riservato al segno del numero, sette bit sono riservati alla rappresentazione dell'esponente b, i restanti 24 bit sono utilizzati per le cifre α_i della mantissa che, in base 16, sono 6 in quanto occorrono 4 bit per rappresentare in binario una cifra esadecimale compresa tra 0 e 15. L'esponente b non è rappresentato in segno ma in traslazione rispetto a 64 e cioè viene rappresentato il numero $b^* = b+64$; ne segue che b è compreso tra -64 e 63.

Di conseguenza, indicando con F la cifra esadecimale che corrisponde al numero 15[2], il massimo numero rappresentabile (in precisione semplice) è $0.FFFFFF \times 16^{63} \simeq 7.23 \times 10^{75}$, il minimo numero non negativo rappresentabile risulta $0.1 \times 16^{-64} \simeq 5.39 \times 10^{-79}$ e la precisione di macchina è $u_s = \frac{1}{2}16^{-5} \simeq 4.75 \times 10^{-7}$.

1.1.3 Numeri reali (doppia precisione)

I numeri reali si dicono in doppia precisione quando sono rappresentati con una doppia parola e quindi con 64 bit. Rispetto alla precisione semplice cambia solo il numero delle cifre esadecimali della mantissa che da 6 divengono 14 in quanto essa viene ad includere i 32 bit aggiunti. L'insieme dei numeri rappresentabili non cambia molto mentre cambia la precisione di macchina che diviene $u_d = \frac{1}{2}16^{-13} \simeq 1.11 \times 10^{-16}$.

È possibile anche l'uso di precisioni multiple in cui si incrementa ulteriormente il numero delle cifre della mantissa, migliorando di conseguenza la precisione di macchina.

[2]Le cifre della rappresentazione in base 16 sono: $0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F$.

1.2 Le operazioni di macchina

Nell'insieme M non tutte le proprietà delle quattro operazioni elementari risultano verificate, in quanto il risultato di una operazione deve essere ricondotto ad un numero di macchina: perciò le operazioni elementari all'interno di una macchina sono diverse dalle corrispondenti operazioni ordinarie. Si indicano nel seguente modo le quattro *operazioni di macchina*:

$\oplus$ addizione
$\ominus$ sottrazione
$\otimes$ moltiplicazione
$\oslash$ divisione.

Un esempio in cui l'addizione ($\oplus$) non gode della proprietà associativa è il seguente.

Sia $F(10, 3, -99, 99)$ e sia $x = 0.135 \times 10^{-4}$, $y = 0.258 \times 10^{-2}$, $z = -0.251 \times 10^{-2}$; si ha

$$\begin{aligned} x \oplus (y \oplus z) &= 0.135 \times 10^{-4} \oplus \left(0.258 \times 10^{-2} \oplus -0.251 \times 10^{-2}\right) \\ &= 0.135 \times 10^{-4} \oplus 0.700 \times 10^{-4} \\ &= 0.835 \times 10^{-4}, \end{aligned}$$

mentre

$$\begin{aligned} (x \oplus y) \oplus z &= \left(0.135 \times 10^{-4} \oplus 0.258 \times 10^{-2}\right) \oplus -0.251 \times 10^{-2} \\ &= 0.259 \times 10^{-2} \oplus -0.251 \times 10^{-2} \\ &= 0.800 \times 10^{-4}. \end{aligned}$$

In modo analogo si possono dare esempi in cui l'operazione $\otimes$ non gode della proprietà distributiva rispetto alla addizione.

Quando si sottraggono due numeri di macchina dello stesso segno che hanno lo stesso esponente b e con le mantisse che differiscono di poco, si incorre in una perdita di cifre significative nel risultato. Tale fenomeno, detto *cancellazione*, produce, come si vedrà più avanti, una notevole amplificazione degli errori relativi.

1.3 Errore nel calcolo di una funzione

Una funzione non razionale φ viene sempre sostituita, all'interno di un calcolatore, da una funzione razionale f, il cui uso comporta un errore $f - \varphi$ che si dice *errore di troncamento*. Tale errore, all'occorrenza è facilmente maggiorabile.

Tuttavia anche nel calcolo di una funzione razionale $f(x_1, x_2, \ldots, x_n)$ in un punto assegnato $P_0 = (x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})$, in generale, non si ottiene il valore $f(P_0)$ cercato, a causa delle approssimazioni che si introducono.

Tali approssimazioni producono due tipi di errore.

Un primo errore nasce dal fatto che le operazioni aritmetiche che compaiono nella $f(P)$ devono essere sostituite con le corrispondenti operazioni di macchina e organizzate in un certo algoritmo e ciò equivale in definitiva alla sostituzione di $f(P)$ con un'altra funzione $f_a(P)$ che la approssimi.

Un secondo tipo di errore si presenta quando non è possibile rappresentare esattamente le coordinate del punto P_0 e quindi si devono approssimare tali coordinate con numeri di macchina (basti pensare, per esempio, al punto $P_0 = (\sqrt{2}, \pi)$).

1.3.1 Errore assoluto

Assegnato il punto $P_0 \in \mathbb{R}^n$ di coordinate $x_i^{(0)}$, $i = 1, 2, \ldots, n$, nel quale si vuole calcolare la funzione $f(P)$, si consideri l'insieme

$$D = \{P \in \mathbb{R}^n \mid a_i \leq x_i \leq b_i,\ i = 1, 2, \ldots, n\}$$

dove $a_i, b_i \in \mathbb{R}$ e si supponga che sia $a_i \leq x_i^{(0)} \leq b_i$, $i = 1, 2, \ldots, n$; D si dice *insieme di indeterminazione* del punto P_0.

In effetti il valore cercato $f(P_0)$ viene sostituito dal valore calcolato $f_a(P_1)$ dove P_1, di coordinate $x_i^{(1)}$, $i = 1, 2, \ldots, n$, appartiene a D e quindi l'errore commesso risulta $f_a(P_1) - f(P_0)$; di questo errore, detto *errore totale*, si può dare una stima.

Si pone

$$f_a(P_1) - f(P_0) = f_a(P_1) - f(P_1) + f(P_1) - f(P_0)$$

dove la differenza $f_a(P_1) - f(P_1)$ è detta *errore algoritmico* mentre la differenza $f(P_1) - f(P_0)$ è detta *errore trasmesso dai dati*.

L'errore algoritmico, una volta fissato l'algoritmo che fornisce $f_a(P)$, risulta definito e stimabile.

All'errore trasmesso, nell'ipotesi che $f(P) \in C^1(D)$, si può dare una rappresentazione generale: dalla formula di Taylor arrestata al primo termine e con punto iniziale P_0 si ottiene

$$f(P_1) - f(P_0) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_i^{(1)} - x_i^{(0)}) \tag{1.3}$$

dove le derivate parziali della funzione $f(P)$ sono calcolate in un punto opportuno. Indicando con

$$\begin{aligned} \delta_f &= f_a(P_1) - f(P_0) && \text{l'errore totale,} \\ \delta_a &= f_a(P_1) - f(P_1) && \text{l'errore algoritmico,} \\ \delta_d &= f(P_1) - f(P_0) && \text{l'errore trasmesso dai dati,} \end{aligned}$$

risulta

$$\delta_f = \delta_a + \delta_d.$$

Ponendo poi

$$\delta_{x_i} = x_i^{(1)} - x_i^{(0)}, \quad \rho_i = \frac{\partial f}{\partial x_i},$$

la (1.3) diventa

$$\delta_d = \sum_{i=1}^{n} \rho_i \delta_{x_i}.$$

I valori ρ_i sono detti *coefficienti di amplificazione* degli gli errori δ_{x_i}.

Una limitazione per il modulo dell'errore assoluto $f_a(P_1) - f(P_0)$ è

$$\mid f_a(P_1) - f(P_0) \mid \leq E_a + E_d$$

dove si è posto

$$\begin{aligned} E_a \;=\; & \text{massimo modulo dell}'\text{errore assoluto dovuto} \\ & \text{al particolare algoritmo usato,} \end{aligned}$$

$$E_d = \sum_{i=1}^{n} A_{x_i} \mid \delta_{x_i} \mid, \tag{1.4}$$

con

$$A_{x_i} \geq \sup_{x \in D} \left| \frac{\partial f}{\partial x_i} \right|, \quad i = 1, 2, \ldots, n. \tag{1.5}$$

Se si conosce una stima dell'errore di troncamento e degli errori δ_{x_i} nonché le A_{x_i}, si può stabilire a posteriori un confine superiore per l'errore assoluto con cui si è calcolata la funzione nel punto desiderato; questo problema è detto *problema diretto*.

Il *problema inverso* consiste nel richiedere a priori che il valore $f_a(P_1)$ sia tale che l'errore assoluto $\mid f_a(P_1) - f(P_0) \mid$ risulti minore di un valore prefissato, per cui si deve cercare sia un algoritmo $f_a(P)$ sia un opportuno punto P_1 che soddisfino la richiesta.

1.3.2 Errore relativo

L'*errore relativo* che si commette nel calcolo di una funzione $f(P)$ in un assegnato punto P_0 è definito da

$$\epsilon_f = \frac{f_a(P_1) - f(P_0)}{f(P_0)}.$$

Si verifica facilmente che

$$\epsilon_f = \frac{f(P_1) - f(P_0)}{f(P_0)} + \frac{f_a(P_1) - f(P_1)}{f(P_1)} \left(1 + \frac{f(P_1) - f(P_0)}{f(P_0)}\right)$$

per cui, indicando con

$$\epsilon_a = \frac{f_a(P_1) - f(P_1)}{f(P_1)}$$

l'*errore relativo algoritmico* e con

$$\epsilon_d = \frac{f(P_1) - f(P_0)}{f(P_0)}$$

l'*errore relativo trasmesso dai dati*, si ottiene

$$\epsilon_f = \epsilon_a + \epsilon_d + \epsilon_a \epsilon_d. \tag{1.6}$$

Nella (1.6) si trascura il termine $\epsilon_a \epsilon_d$ in quanto di ordine superiore.

Quanto all'errore relativo trasmesso dai dati, dalla relazione (1.3) si ricava

$$\epsilon_d = \frac{f(P_1) - f(P_0)}{f(P_0)} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{(x_i^{(1)} - x_i^{(0)})}{f(P_0)}$$

ed ancora, definendo $\epsilon_{x_i} = \frac{x_i^{(1)} - x_i^{(0)}}{x_i^{(0)}}$,

$$\epsilon_d = \sum_{i=1}^{n} \frac{x_i^{(0)}}{f(P_0)} \frac{\partial f}{\partial x_i} \epsilon_{x_i}.$$

Infine, ponendo $\gamma_i = \frac{x_i^{(0)}}{f(P_0)} \frac{\partial f}{\partial x_i}$, si ha

$$\epsilon_d = \sum_{i=1}^{n} \gamma_i \epsilon_{x_i} \ ,$$

dove i valori γ_i sono detti *coefficienti di amplificazione* degli errori relativi.

Se i coefficienti γ_i sono tali che l'errore ϵ_d risulta dello stesso ordine degli errori ϵ_{x_i} il problema del calcolo della funzione si dice *ben condizionato*; si dice invece *mal condizionato* se a piccoli errori relativi ϵ_{x_i} corrisponde un errore ϵ_d rilevante.

Dalla (1.6) si vede che all'errore ϵ_f concorre anche l'errore algoritmico ϵ_a: se l'algoritmo è tale da produrre errori accettabilmente limitati nella sua applicazione, si dice *stabile*, mentre è detto *instabile* nel caso contrario.

1.4 Gli errori nelle quattro operazioni

Analizzando le quattro operazioni fondamentali per quanto riguarda gli errori trasmessi dai dati si ottiene la seguente tabella:

operazione	δ_d	ϵ_d
$x \oplus y$	$\delta_x + \delta_y$	$\frac{x}{x+y}\epsilon_x + \frac{y}{x+y}\epsilon_y$
$x \ominus y$	$\delta_x - \delta_y$	$\frac{x}{x-y}\epsilon_x - \frac{y}{x-y}\epsilon_y$
$x \otimes y$	$y\delta_x + x\delta_y$	$\epsilon_x + \epsilon_y$
$x \oslash y$	$\frac{1}{y}\delta_x - \frac{x}{y^2}\delta_y$	$\epsilon_x - \epsilon_y$

Si deduce che le operazioni di addizione e sottrazione non danno problemi per quanto riguarda l'errore assoluto, mentre possono rendere grande l'errore relativo nel caso in cui i due termini dell'operazione siano molto vicini in valore assoluto, in quanto può accadere che i denominatori che compaiono nei coefficienti di amplificazione dell'errore relativo siano molto piccoli in valore assoluto (fenomeno della cancellazione già accennato in 1.2). La moltiplicazione non amplifica l'errore relativo e comporta un errore assoluto che

dipende dall'ordine di grandezza dei fattori; anche la divisione non produce amplificazione per quanto riguarda l'errore relativo, mentre l'errore assoluto diminuisce se aumenta (in valore assoluto) il divisore.

1.5 Complementi ed esempi

Si riportano due esempi sul calcolo degli errori assoluti e relativi.

Esempio 1.5.1 Si vuole calcolare la funzione $f(x_1, x_2) = x_1/x_2$ nel punto assegnato $P_0 = (\sqrt{5}, \pi)$.

Si assuma $D = \{(x_1, x_2) \in \mathbb{R}^2 | 2.23 < x_1 < 2.24,\ 3.14 < x_2 < 3.15\}$ come insieme di indeterminazione del punto P_0 (dalle (1.4) e (1.5) si ricava che più si riduce l'insieme D e più stringente è la maggiorazione dell'errore assoluto).

Problema diretto

Si calcola il rapporto x_1/x_2 arrotondando il risultato alla seconda cifra decimale, ottenendo quindi un errore $E_a \leq \frac{1}{2}10^{-2}$.
Assumendo $P_1 = (2.235, 3.145)$ si ha sicuramente $|\ \delta_{x_1}\ |, |\ \delta_{x_2}\ | \leq \frac{1}{2}10^{-2}$.

Dalla funzione e dall'insieme D si traggono le maggiorazioni

$$\sup_{x \in D} \left| \frac{\partial f}{\partial x_1} \right| \leq A_{x_1} = 0.32,$$

$$\sup_{x \in D} \left| \frac{\partial f}{\partial x_2} \right| \leq A_{x_2} = 0.23;$$

da cui si ha il massimo errore assoluto

$$E_a + E_d \leq \frac{1}{2}10^{-2} + (0.32 + 0.23)\frac{1}{2}10^{-2} = 0.775 \times 10^{-2}.$$

Problema inverso

Si vuole avere un valore che differisca dal valore esatto $f(P_0)$ meno di un prefissato errore $E = 10^{-2}$.

Si può imporre che E_a ed E_d siano, ciascuno, non superiore a metà dell'errore totale richiesto e quindi che sia $E_a \leq \frac{1}{2}10^{-2}$ e $E_d \leq \frac{1}{2}10^{-2}$.

Per l'errore di troncamento è sufficiente arrotondare il risultato della divisione alla seconda cifra decimale.

Poiché l'errore trasmesso dai dati è formato dalla somma di due addendi, si suddivide ancora il contributo a tale errore in due parti uguali per cui si ha

$$A_{x_1} \mid \delta_{x_1} \mid \leq \frac{1}{4}10^{-2},$$
$$A_{x_2} \mid \delta_{x_2} \mid \leq \frac{1}{4}10^{-2};$$

assumendo per A_{x_1} e A_{x_2} i valori calcolati in precedenza si ottiene

$$\mid \delta_{x_1} \mid \leq \frac{1}{4}10^{-2}\frac{1}{0.32} \simeq 0.0078,$$
$$\mid \delta_{x_2} \mid \leq \frac{1}{4}10^{-2}\frac{1}{0.23} \simeq 0.0109,$$

per cui si può porre $P_1 = (2.24, 3.14)$.

Si ha quindi $f_a(P_1) = 0.71$ che differisce da $f(P_0)$ meno di E. □

Esempio 1.5.2 Si vuole stimare l'errore relativo commesso nel calcolo della funzione $f(x, y, z, w) = x(y/z - w)$.

Si può ricorrere all'uso dei *grafi*; si calcola la funzione eseguendo una operazione dopo l'altra e stabilendo per ciascuna di esse l'entità degli errori relativi.

In questo esempio la sequenza delle operazioni è

$$r_1 = \frac{y}{z}, \qquad r_2 = r_1 - w, \qquad r_3 = xr_2\ .$$

Il grafo in Fig. 1.1 evidenzia per ogni operazione i coefficienti di amplificazione lungo i cammini orientati. Gli errori relativi dei dati sono $\epsilon_x, \epsilon_y, \epsilon_z, \epsilon_w$, mentre ϵ_{r_i} va inteso come l'errore relativo per la funzione r_i e si calcola dalla (1.6) senza l'ultimo termine.

Per stimare l'errore relativo totale si procede a ritroso seguendo il grafo. Indicando con ϵ_i l'errore algoritmico della i-esima operazione si ha

$$\begin{aligned}
\epsilon_f &= \epsilon_{r_3} \\
&= \epsilon_3 + \epsilon_x + \epsilon_{r_2} \\
&= \epsilon_3 + \epsilon_x + \epsilon_2 + \frac{y}{y - zw}\epsilon_{r_1} - \frac{zw}{y - zw}\epsilon_w \\
&= \epsilon_3 + \epsilon_x + \epsilon_2 + \frac{y}{y - zw}(\epsilon_1 + \epsilon_y - \epsilon_z) - \frac{zw}{y - zw}\epsilon_w \\
&= \epsilon_3 + \epsilon_2 + \frac{y}{y - zw}\epsilon_1 + \epsilon_x + \frac{y}{y - zw}(\epsilon_y - \epsilon_z) - \frac{zw}{y - zw}\epsilon_w\ .
\end{aligned}$$

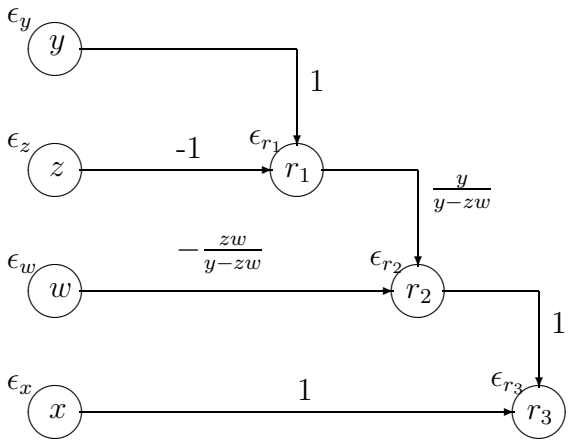

Figura 1.1: Esempio di grafo.

Se ne conclude che l'errore algoritmico è

$$\epsilon_a = \epsilon_3 + \epsilon_2 + \frac{y}{y - zw}\epsilon_1$$

e l'errore trasmesso è

$$\epsilon_d = \epsilon_x + \frac{y}{y - zw}(\epsilon_y - \epsilon_z) - \frac{zw}{y - zw}\epsilon_w \, .$$

□

Osservazione 1.5.1 L'errore relativo calcolato nell'Esempio 1.5.2 dipende dall'algoritmo seguito per il calcolo della funzione $f(x, y, z, w)$; seguendo un altro algoritmo si trova, in generale, un errore relativo diverso.

Bibliografia:

[1], [5], [28], [29].

Capitolo 2

Richiami di algebra lineare

2.1 Matrici e vettori

Con $A \in \mathbb{C}^{m\times n}$ si intende una *matrice* di m righe e n colonne formate da $m \times n$ numeri complessi a_{ij}, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$, detti *elementi* di A. Ogni elemento a_{ij} è individuato dall'indice di riga i e dall'indice di colonna j. Comunemente si scrive

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

Gli interi m ed n si dicono le *dimensioni* di A.

Se $m = n$ la matrice A si dice *quadrata di ordine* n; in tal caso gli elementi a_{ii}, $i = 1, 2, \ldots, n$, si dicono *elementi diagonali* o appartenenti alla *diagonale principale* di A.

Se $m \neq n$ la matrice A dicesi *rettangolare*.

Con $a \in \mathbb{C}^m$ si intende un *vettore* con m componenti complesse indicate come a_i, $i = 1, 2, \ldots, m$. I vettori sono particolari matrici con una sola colonna, potendosi identificare $\mathbb{C}^{m\times 1}$ con $\mathbb{C}^m$.

Se A ha tutti gli elementi reali è detta matrice reale e si suole scrivere $A \in \mathbb{R}^{m\times n}$; analogamente con $a \in \mathbb{R}^m$ si intende un vettore a componenti reali.

Data la matrice $A \in \mathbb{C}^{m\times n}$, la matrice $B \in \mathbb{C}^{n\times m}$ i cui elementi sono $b_{ij} = a_{ji}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$, si dice matrice *trasposta* della

matrice A e si indica con A^T mentre si dice matrice *trasposta coniugata* di A la matrice $B \in \mathbb{C}^{n\times m}$ i cui elementi sono $b_{ij} = \overline{a}_{ji}$ e si indica con A^H.

La matrice quadrata

$$I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

dicesi *matrice identica* o *matrice unità* .

Si dice *matrice diagonale* una matrice quadrata D con gli elementi non diagonali nulli, cioè della forma

$$D = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{pmatrix};$$

si scrive anche $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$.

Le matrici quadrate del tipo

$$L = \begin{pmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{pmatrix}, \quad R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{nn} \end{pmatrix},$$

si dicono *matrici triangolari* e precisamente L dicesi *triangolare inferiore* e R *triangolare superiore.*

2.2 Operazioni tra matrici

Date $A, B \in \mathbb{C}^{m\times n}$, la matrice somma, $C = A + B$, è definita ponendo

$$c_{ij} = a_{ij} + b_{ij}, \quad i = 1, 2, \ldots, m,\ j = 1, 2, \ldots, n.$$

Valgono, per l'addizione, le proprietà associativa e commutativa.

Se $A \in \mathbb{C}^{m\times k}$ e $B \in \mathbb{C}^{k\times n}$, la matrice prodotto $C = AB$, si definisce ponendo

$$c_{ij} = \sum_{l=1}^{k} a_{il} b_{lj}, \quad i = 1, 2, \ldots, m,\ j = 1, 2, \ldots, n,$$

e risulta quindi $C \in \mathbb{C}^{m\times n}$.

Per la moltiplicazione vale la proprietà associativa e la proprietà distributiva rispetto alla addizione.

È importante osservare che la moltiplicazione non gode della proprietà commutativa, né della legge di annullamento del prodotto; infatti in generale è $AB \neq BA$, mentre il prodotto AB può essere uguale alla matrice nulla senza che né A né B siano nulle (cfr. Esempio 2.11.1).

Per questo motivo, quando si moltiplica una matrice A per una matrice B occorre distinguere tra *premoltiplicazione* quando B moltiplica A a sinistra (BA) e *postmoltiplicazione* quando B moltiplica A a destra (AB).

Per la trasposta e la trasposta coniugata del prodotto di due matrici si ha

$$(AB)^T = B^T A^T \quad \text{e} \quad (AB)^H = B^H A^H.$$

Il prodotto $a^H b = \sum_{l=1}^{n} \overline{a}_l b_l$ di due vettori $a, b \in \mathbb{C}^n$ è detto *prodotto scalare* e se $a^H b = 0$ i due vettori si dicono *ortogonali.*

Infine la moltiplicazione di un numero α per una matrice (per un vettore) si esegue moltiplicando per α tutti gli elementi della matrice (del vettore).

Definizione 2.2.1 *k vettori $v^{(1)}, v^{(2)}, \ldots, v^{(k)} \in \mathbb{C}^m$ si dicono* linearmente indipendenti *se*

$$\alpha_1 v^{(1)} + \alpha_2 v^{(2)} + \cdots + \alpha_k v^{(k)} = 0\,,$$

con $\alpha_i \in \mathbb{C}$, $i = 1, 2, \ldots, k$, implica

$$\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0\,.$$

2.3 Determinante e inversa

Definizione 2.3.1 *Si dice* determinante *di una matrice $A \in \mathbb{C}^{n\times n}$ il numero*

$$\det(A) = \sum_{P} s(\pi) a_{1j_1} a_{2j_2} \cdots a_{nj_n}$$

dove π è la permutazione $(j_1, j_2, .., j_n)$ dei numeri $(1, 2, .., n)$, la funzione $s(\pi)$ vale 1 se la permutazione π è di classe pari o -1 se la permutazione π è di classe dispari e P è l'insieme delle $n!$ permutazioni π possibili.

Per il calcolo del determinante di una matrice A si può fare uso del *primo teorema di Laplace*

$$\det(A) = \begin{cases} a_{11} & \text{se } n = 1, \\ \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(A_{ij}) & \\ \quad = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det(A_{ij}) & \text{se } n > 1, \end{cases} \tag{2.1}$$

dove A_{ij} è la sottomatrice ottenuta da A eliminando la riga i-esima e la colonna j-esima.

Vale anche il *secondo teorema di Laplace*

$$\sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(A_{rj}) = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det(A_{is}) = 0$$

per $r \neq i$ e $s \neq j$.

Nel caso di una matrice diagonale o triangolare si ha immediatamente

$$\det(A) = \prod_{i=1}^{n} a_{ii}.$$

Il determinante della matrice A si indica anche con la notazione

$$\begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}.$$

Dalla (2.1) segue anche, per induzione, $\det(A) = \det(A^T)$.

Definizione 2.3.2 *Una matrice $A \in \mathbb{C}^{n \times n}$ si dice* singolare *(*non singolare*) se* $\det(A) = 0$ *(*$\det(A) \neq 0$*).*

Definizione 2.3.3 *Data una matrice $A \in \mathbb{C}^{m \times n}$ ed un intero $k \leq \min\{m, n\}$ si dice* minore di ordine k *il determinante di una matrice ottenuta da A prendendo tutti gli elementi sulla intersezione di k righe e k colonne comunque fissate.*

Definizione 2.3.4 *Data una matrice $A \in \mathbb{C}^{n \times n}$ si dicono* minori principali di ordine k *i determinanti delle sottomatrici di ordine k estratte da A e aventi diagonale principale composta da elementi della diagonale principale di A.*

Definizione 2.3.5 *Data una matrice $A \in \mathbb{C}^{n\times n}$ si dice* minore principale di testa di ordine *k il determinante della sottomatrice di ordine k formata dalle prime k righe e k colonne di A.*

Definizione 2.3.6 *Data una matrice $A \in \mathbb{C}^{m\times n}$ si dice* rango *o* caratteristica *della matrice A il numero $r(A)$ pari all'ordine più alto dei suoi minori diversi da zero.*

Teorema 2.3.1 (di Binet-Cauchy) *Date $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{n\times m}$, il determinante della matrice prodotto $C = AB \in \mathbb{C}^{m\times m}$ è nullo se $m > n$; altrimenti è dato dalla somma dei prodotti di tutti i possibili minori di ordine massimo di A per i corrispondenti minori di B.*[1]

Corollario 2.3.1 *Dal Teorema 2.3.1 segue, se $m = n$,*

$$\det(AB) = \det(A)\det(B)\,.$$

Definizione 2.3.7 *Data una matrice $A \in \mathbb{C}^{n\times n}$ non singolare si dice* matrice inversa *di A la matrice $B \in \mathbb{C}^{n\times n}$ tale che*

$$b_{ij} = (-1)^{i+j}\frac{\det(A_{ji})}{\det(A)}\,.$$

Da questa definizione e dai teoremi di Laplace segue

$$AB = BA = I.$$

In seguito, indicheremo con il simbolo A^{-1} la matrice inversa della matrice A.

Dal Corollario 2.3.1 si ha

$$\det(A)\det(A^{-1}) = \det(AA^{-1}) = \det(I) = 1\,,$$

per cui il determinante di una matrice A è il reciproco del determinante della sua inversa.

Dalla Definizione 2.3.7 segue che una matrice A singolare non ammette matrice inversa.

[1] Un minore di ordine massimo di A, nel caso $m \leq n$, è il determinante della matrice formata da m colonne di A di indici $k_1, k_2, \ldots, k_m$ comunque fissati; il corrispondente minore di B è il determinante della matrice formata da m righe di B di indici $k_1, k_2, \ldots, k_m$.

2.4 Matrici particolari

Una matrice $A \in \mathbb{C}^{n\times n}$ si dice

hermitiana	se	$A = A^H$;
antihermitiana	se	$A = -A^H$;
unitaria	se	$A^H A = AA^H = I$;
normale	se	$A^H A = AA^H$;
simmetrica	se	$A = A^T$;
antisimmetrica	se	$A = -A^T$.

In particolare, nel campo reale, una matrice hermitiana è anche simmetrica, mentre una matrice unitaria è detta anche *ortogonale*.

Le matrici unitarie hanno come inversa la loro trasposta coniugata e le matrici ortogonali hanno come inversa la loro trasposta.

Data una matrice hermitiana A ed un vettore $x \in \mathbb{C}^n$, lo scalare $x^H Ax$ è un numero reale in quanto

$$(x^H Ax)^H = x^H A^H x = x^H Ax;$$

la matrice A si dice *definita positiva (negativa)* se $x^H Ax > 0$ (< 0) per ogni $x \in \mathbb{C}^n$ con $x \neq 0$, mentre è detta *semidefinita positiva (negativa)* se $x^H Ax \geq 0$ (≤ 0).

Teorema 2.4.1 *Una matrice hermitiana è definita positiva se e solo se i suoi minori principali di testa sono tutti positivi.*

Definizione 2.4.1 *Una matrice $P \in \mathbb{R}^{n\times n}$ è detta matrice di* permutazione *se è ottenuta dalla matrice identica operando su di essa una permutazione delle colonne (delle righe).*

Una matrice di permutazione è una matrice ortogonale, avendosi $P^T P = PP^T = I$.

Il prodotto di una matrice A per una matrice di permutazione P produce su A una permutazione delle colonne o delle righe; precisamente:

AP presenta, rispetto ad A, la stessa permutazione di colonne che si è operata su I per ottenere P;

PA presenta, rispetto ad A, la stessa permutazione di righe che si è operata su I per ottenere P.

Si noti che se P si ottiene permutando le colonne di I, allora P^T si ottiene con la stessa permutazione delle righe di I.

Definizione 2.4.2 *Una matrice* $A \in \mathbb{C}^{n\times n}$ *si dice a* predominanza diagonale forte *se*

$$| a_{ii} | > \sum_{\substack{j=1 \\ j\neq i}}^{n} | a_{ij} |, \quad i = 1, 2, \ldots, n.$$

Definizione 2.4.3 *Una matrice* $A \in \mathbb{C}^{n\times n}$ *si dice a* predominanza diagonale debole *se*

$$| a_{ii} | \geq \sum_{\substack{j=1 \\ j\neq i}}^{n} | a_{ij} |, \quad i = 1, 2, \ldots, n,$$

e per almeno un indice r, $1 \leq r \leq n$, *si ha*

$$| a_{rr} | > \sum_{\substack{j=1 \\ j\neq r}}^{n} | a_{rj} | .$$

Definizione 2.4.4 *Una matrice* $A \in \mathbb{C}^{n\times n}$ *si dice* convergente *se*

$$\lim_{k\to\infty} A^k = \mathbf{O}$$

dove $\mathbf{O}$ *è la matrice nulla.*

2.5 Sistemi lineari

Dati una matrice $A \in \mathbb{C}^{n\times n}$ ed un vettore $b \in \mathbb{C}^n$, un sistema di n equazioni lineari in n incognite si può rappresentare nella forma (cfr. 3.1)

$$Ax = b \tag{2.2}$$

dove $x \in \mathbb{C}^n$ è il vettore delle incognite.

Sulla risolubilità del sistema (2.2) si ha il seguente teorema.

Teorema 2.5.1 (di Rouché - Capelli) *Il sistema lineare (2.2) ammette soluzione se e solo se*

$$r(A) = r(A \mid b),$$

dove con $(A \mid b)$ *è indicata la matrice completa del sistema costituita da* n *righe ed* $n+1$ *colonne. Se* $r(A) = n$ *la soluzione è unica mentre se* $r(A) < n$ *l'insieme delle soluzioni è un sottospazio di* $\mathbb{C}^n$ *di dimensione* $n - r(A)$.

Se $\det(A) \neq 0$, il sistema (2.2) si dice *normale*. In tal caso la soluzione è data formalmente da $x = A^{-1}b$ e può essere espressa mediante la *regola di Cramer*

$$x_i = \frac{\det(A_i)}{\det(A)}, \quad i = 1, 2, \ldots, m,$$

essendo A_i la matrice ottenuta da A sostituendo la i-esima colonna con il vettore b.

Nel caso particolare $b = 0$ il sistema (2.2) si dice *omogeneo* ed ha sicuramente soluzioni, avendosi $r(A) = r(A \mid b)$; se è anche $\det(A) \neq 0$ l'unica soluzione è $x = 0$. Ne segue che un sistema omogeneo ha soluzioni non nulle allora e solo che sia $\det(A) = 0$.

2.6 Matrici partizionate, matrici riducibili

Nelle applicazioni si incontrano spesso matrici A *partizionate a blocchi* e cioè matrici i cui elementi sono sottomatrici di A.

Una qualunque matrice può essere partizionata a blocchi in molti modi; è importante il caso in cui i blocchi diagonali sono quadrati [2].

Come per le matrici scritte ad elementi, anche per le matrici partizionate a blocchi si possono avere matrici *triangolari a blocchi*, cioè aventi una delle seguenti forme

$$\begin{pmatrix} A_{11} & \mathbf{O} & \cdots & \mathbf{O} \\ A_{21} & A_{22} & \cdots & \mathbf{O} \\ \vdots & \vdots & \ddots & \vdots \\ A_{k1} & A_{k2} & \cdots & A_{kk} \end{pmatrix}, \quad \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1k} \\ \mathbf{O} & A_{22} & \cdots & A_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{O} & \mathbf{O} & \cdots & A_{kk} \end{pmatrix},$$

o *diagonali a blocchi*

$$\begin{pmatrix} A_{11} & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & A_{22} & \cdots & \mathbf{O} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{O} & \mathbf{O} & \cdots & A_{kk} \end{pmatrix}.$$

In questi casi particolari è facile verificare che

$$\det(A) = \prod_{i=1}^{k} \det(A_{ii}).$$

[2] Ciò comporta che la matrice A sia quadrata.

Definizione 2.6.1 *Una matrice* $A \in \mathbb{C}^{n\times n}$ *si dice* riducibile *se esiste una matrice di permutazione* P *tale che la matrice* P^TAP *sia della forma*

$$B = P^TAP = \begin{pmatrix} A_{11} & \mathbf{O} \\ A_{21} & A_{22} \end{pmatrix}$$

con blocchi diagonali quadrati.

Nel caso in cui B contenga qualche blocco diagonale ancora riducibile si può operare una nuova trasformazione con un'altra matrice di permutazione e così via fino ad arrivare alla *forma ridotta* della matrice A in cui tutti i blocchi diagonali risultano non riducibili.

Una matrice che non sia riducibile è detta *irriducibile.*

Da quanto visto in 2.4, la matrice B si ottiene dalla matrice A operando sulle righe e sulle colonne la stessa permutazione operata sulle colonne di I per ottenere P.

Per stabilire se una matrice è riducibile o meno, cioè se esiste o no una matrice P che verifichi la Definizione 2.6.1, servono le definizioni ed il teorema seguenti.

Definizione 2.6.2 *Data una matrice* $A \in \mathbb{C}^{n\times n}$*, fissati* n *punti, detti* nodi*,* $N_1, N_2, \ldots, N_n$*, si dice* grafo orientato *associato ad* A*, il grafo che si ottiene congiungendo* N_i *a* N_j *con un cammino orientato da* N_i *a* N_j *per ogni* $a_{ij} \neq 0$*.*

Definizione 2.6.3 *Un grafo orientato si dice* fortemente connesso *se da ogni nodo* N_i*,* $i = 1, 2, \ldots, n$*, è possibile raggiungere un qualunque altro nodo* N_j*,* $j = 1, 2, \ldots, n$*, seguendo un cammino orientato eventualmente passante per altri nodi.*

Teorema 2.6.1 *Una matrice* $A \in \mathbb{C}^{n\times n}$ *è irriducibile se e solo se il grafo orientato ad essa associato risulta fortemente connesso.*

(cfr. Esempi 2.11.3 e 2.11.4)

Se A è riducibile, per costruire P si procede come nell'Esempio 2.11.5.

2.7 Autovalori e autovettori

Definizione 2.7.1 *Data una matrice* $A \in \mathbb{C}^{n\times n}$ *si dice* autovalore *di* A *ogni numero* $\lambda \in \mathbb{C}$ *tale che il sistema lineare*

$$Ax = \lambda x, \quad x \in \mathbb{C}^n, \tag{2.3}$$

abbia soluzioni $x \neq 0$; ogni tale vettore x è detto autovettore destro *associato all'autovalore λ, intendendo che x ed ogni vettore kx ($k \in \mathbb{C}$, $k \neq 0$) rappresentano lo stesso autovettore.*

In analogia è detto *autovettore sinistro* un vettore $y \in \mathbb{C}^n$ tale che

$$y^T A = \lambda y^T.$$

Trasponendo entrambi i membri della (2.3) si ha

$$x^T A^T = \lambda x^T$$

da cui risulta che gli autovettori destri di A sono gli autovettori sinistri di A^T associati allo stesso autovalore.

Come si è visto, dal teorema di Rouché -Capelli, un sistema lineare omogeneo ha soluzioni non nulle se e solo se la matrice dei coefficienti del sistema è singolare e cioè ha determinante nullo; poiché il sistema (2.3) è equivalente al sistema omogeneo

$$(A - \lambda I)x = 0\,,$$

ne segue che gli autovalori di A sono tutti e soli i numeri λ che soddisfano l'equazione

$$\det(A - \lambda I) = 0. \tag{2.4}$$

Essendo $\det(A - \lambda I) = \det[(A - \lambda I)^T] = \det(A^T - \lambda I)$ segue che A e A^T hanno gli stessi autovalori.

Dal calcolo di $\det(A - \lambda I)$, si ottiene

$$\begin{aligned} \det(A - \lambda I) \;\; = \;\; & (-1)^n \lambda^n + (-1)^{n-1} \sigma_1 \lambda^{n-1} + \\ & (-1)^{n-2} \sigma_2 \lambda^{n-2} + \cdots - \sigma_{n-1} \lambda + \sigma_n, \end{aligned} \tag{2.5}$$

dove i coefficienti σ_i, $i = 1, 2, \ldots, n$, sono, ciascuno, la somma dei minori principali di ordine i estratti dalla matrice A.

In particolare risulta $\sigma_1 = \sum_{j=1}^{n} a_{jj}$, e $\sigma_n = \det(A)$. σ_1 si dice *traccia* di A e si indica col simbolo $tr(A)$.

Il polinomio (2.5) è detto *polinomio caratteristico* della matrice A mentre l'equazione (2.4) prende il nome di *equazione caratteristica.*

Gli autovalori di una matrice $A \in \mathbb{C}^{n \times n}$ coincidono con le radici dell'equazione caratteristica, perciò sono n e verranno indicati con $\lambda_1, \lambda_2, \ldots, \lambda_n$.

Osservazione 2.7.1 Dalle (2.4) e (2.5) si deducono le proprietà:

$\lambda = 0$ è autovalore di $A \Longleftrightarrow \det(A) = 0$;

la somma degli autovalori coincide con σ_1 (cfr. le (4.50)) e quindi

$$\sum_{i=1}^{n} \lambda_i = tr(A)\,; \tag{2.6}$$

il prodotto degli autovalori coincide con σ_n (cfr. le (4.50)) e quindi

$$\prod_{i=1}^{n} \lambda_i = \det(A). \tag{2.7}$$

Definizione 2.7.2 *Si dice* raggio spettrale *della matrice A il numero reale non negativo*

$$\rho(A) = \max_{1 \le i \le n} \mid \lambda_i \mid \ .$$

Teorema 2.7.1 *Una matrice $A \in \mathcal{C}^{n \times n}$ è convergente se e solo se il suo raggio spettrale è minore di 1.*

Definizione 2.7.3 *Data una matrice A ed una matrice S non singolare, si dice* trasformata per similitudine *della matrice A, la matrice B tale che*

$$B = S^{-1}AS;$$

le matrici A e B si dicono simili.

Si verifica subito che la relazione di similitudine è riflessiva, simmetrica e transitiva.

Teorema 2.7.2 *Due matrici simili A e B hanno gli stessi autovalori. Inoltre, per ogni autovalore λ, se x è autovettore di A, allora $S^{-1}x$ è autovettore di B.*

DIMOSTRAZIONE. Per ottenere la prima parte della tesi basta verificare che due matrici simili A e B hanno lo stesso polinomio caratteristico.

Si ha infatti

$$\begin{aligned}\det(B-\lambda I) &= \det(S^{-1}AS-\lambda I)=\det(S^{-1}AS-\lambda S^{-1}S)\\ &= \det[S^{-1}(A-\lambda I)S]\\ &= \det(S^{-1})\det(A-\lambda I)\det(S)\\ &= \det(A-\lambda I).\end{aligned}$$

Per provare la seconda parte si osservi che da $Ax=\lambda x$ segue $(S^{-1}AS)S^{-1}x = \lambda S^{-1}x$. □

Teorema 2.7.3 *Se A possiede l'autovalore λ e $k \in \mathbb{N}$, A^k possiede l'autovalore λ^k e gli autovettori di A sono anche autovettori di A^k.*

Questo teorema, se A è non singolare, vale anche per $k \in \mathbb{Z}$.

Teorema 2.7.4 *Gli autovalori di una matrice hermitiana sono tutti reali.*

DIMOSTRAZIONE. Sia λ un autovalore di A. Dalla uguaglianza $Ax=\lambda x$, si ottiene, premoltiplicando per x^H, $x^HAx = \lambda x^Hx$ ed ancora, dividendo per il numero reale e positivo x^Hx (si ricordi che $x \neq 0$),

$$\lambda = \frac{x^HAx}{x^Hx}. \tag{2.8}$$

Nel secondo membro della (2.8) il numeratore è reale per quanto visto all'inizio di 2.4: ne segue la tesi. □

Definizione 2.7.4 *Dicesi* molteplicità algebrica *di un autovalore λ, la molteplicità di λ come radice dell'equazione caratteristica.*

La molteplicità algebrica di λ sarà indicata con $\alpha(\lambda)$.

Definizione 2.7.5 *Dicesi* molteplicità geometrica *di λ, e si indicherà con $\gamma(\lambda)$, la dimensione dello spazio delle soluzioni del sistema lineare omogeneo $(A-\lambda I)x=0$.*

Si osservi che la molteplicità geometrica così definita corrisponde al numero di autovettori, tra loro linearmente indipendenti, associati all'autovalore λ e si ha (cfr. Teorema 2.5.1) $\gamma(\lambda) = n - r(A-\lambda I)$.

Teorema 2.7.5 *Per ogni autovalore λ risulta*

$$1 \leq \gamma(\lambda) \leq \alpha(\lambda) \leq n \, .$$

Teorema 2.7.6 (traslazione di spettro) *Sia λ autovalore di A e $q \in \mathbb{C}$; allora $B = A + qI$ ha come autovalore $\lambda + q$ con molteplicità algebrica e geometrica pari a quelle di λ; inoltre B ha gli stessi autovettori di A.*

Teorema 2.7.7 *Se λ_i e λ_j sono autovalori distinti gli autovettori ad essi associati sono linearmente indipendenti.*

Ne segue che se $A \in \mathbb{C}^{n \times n}$ possiede n autovalori distinti, A ammette n autovettori linearmente indipendenti.

Definizione 2.7.6 *Una matrice A si dice* diagonalizzabile *se esiste una matrice X non singolare tale che*

$$X^{-1} A X = D \tag{2.9}$$

con $D = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$.

Una matrice X che verifica la (2.9) è tale che le sue colonne $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ sono autovettori linearmente indipendenti di A; infatti,

$$AX = XD$$

da cui segue $Ax^{(i)} = \lambda_i x^{(i)}$, $i = 1, 2, \ldots, n$.

Teorema 2.7.8 *Una matrice A è normale se e solo se esiste una matrice X unitaria per cui valga la (2.9).*

Teorema 2.7.9 (di Jordan) *Ogni matrice A, di ordine n, avente $k \leq n$ autovalori distinti, $\lambda_1, \lambda_2, \ldots, \lambda_k$, è simile ad una matrice diagonale a k blocchi, cioè esiste una matrice H non singolare tale che*

$$H^{-1} A H = J = diag(J^{(1)}, J^{(2)}, \ldots, J^{(k)}) \tag{2.10}$$

dove ogni blocco diagonale $J^{(i)}$ è di ordine $\alpha(\lambda_i)$, $i = 1, 2, \ldots, k$, ed è anch'esso di forma diagonale a blocchi, avendosi

$$J^{(i)} = diag(J_1^{(i)}, J_2^{(i)}, \ldots, J_{\gamma(\lambda_i)}^{(i)}) \, ;$$

ciascun blocco $J_l^{(i)}$*, detto anche* blocco di Jordan*, è della forma*

$$J_l^{(i)} = \begin{pmatrix} \lambda_i & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_i & 1 & \cdots & 0 & 0 \\ 0 & 0 & \lambda_i & \ddots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_i & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda_i \end{pmatrix}, \quad l = 1, 2, \ldots, \gamma(\lambda_i).$$

Si noti che, se $p_l^{(i)}$ è l'ordine di $J_l^{(i)}$, risulta

$$\sum_{l=1}^{\gamma(\lambda_i)} p_l^{(i)} = \alpha(\lambda_i)\,. \tag{2.11}$$

La matrice (2.10) si dice *forma canonica di Jordan* della matrice A. Il polinomio caratteristico del blocco di Jordan $J_l^{(i)}$ è $\det(J_l^{(i)} - \lambda I) = (\lambda_i - \lambda)^{p_l^{(i)}}$, e si dice un *divisore elementare* di A.

Si osservi che il numero dei blocchi di Jordan di una matrice A corrisponde al numero di autovettori linearmente indipendenti di A.

Le matrici diagonalizzabili costituiscono un caso particolare in cui i blocchi di Jordan sono tutti di ordine 1, cioè i divisori elementari sono tutti lineari: in tal caso le colonne di H sono gli autovettori di A.

Dal teorema di Jordan, e in particolare dalla (2.11), discende il seguente teorema.

Teorema 2.7.10 *Una matrice è diagonalizzabile se e solo se per ogni suo autovalore* λ *si ha* $\alpha(\lambda) = \gamma(\lambda)$.

2.8 Localizzazione degli autovalori

Teorema 2.8.1 (primo teorema di Gershgorin) *Sia* $A \in \mathcal{C}^{n\times n}$*, si indichino con* $\mathcal{F}_i$*,* $i = 1, 2, \ldots, n$*, gli insiemi*

$$\mathcal{F}_i = \{z \in \mathcal{C} \mid\mid z - a_{ii} \mid \leq \rho_i\}, \quad con \quad \rho_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} \mid a_{ij} \mid; \tag{2.12}$$

allora se λ *è autovalore di* A *si ha*

$$\lambda \in \mathcal{F} = \bigcup_{i=1}^{n} \mathcal{F}_i\,.$$

DIMOSTRAZIONE. Indicando con x un autovettore destro associato all'autovalore λ, sia x_k la sua componente di massimo modulo; dalla riga k-esima della (2.3) si ha

$$\sum_{j=1}^{n} a_{kj}x_j = \lambda x_k$$

da cui

$$(\lambda - a_{kk})x_k = \sum_{\substack{j=1 \\ j\neq k}}^{n} a_{kj}x_j \ .$$

Passando ai moduli dei due membri e maggiorando nel secondo membro ogni $\mid x_j \mid$ con $\mid x_k \mid$, si ottiene

$$\mid \lambda - a_{kk} \mid\mid x_k \mid \leq \sum_{\substack{j=1 \\ j\neq k}}^{n} \mid a_{kj} \mid\mid x_k \mid$$

ed essendo, per la definizione di autovettore, $x_k \neq 0$, si ha

$$\mid \lambda - a_{kk} \mid \leq \sum_{\substack{j=1 \\ j\neq k}}^{n} \mid a_{kj} \mid = \rho_k$$

che prova la tesi. □

Il Teorema 2.8.1 ha una evidente interpretazione grafica: ciascun insieme $\mathcal{F}_i$ di (2.12) è, nel piano complesso, un cerchio di centro il punto a_{ii} e raggio ρ_i; l'insieme $\mathcal{F}$ a cui appartengono tutti gli autovalori di A è quindi l'unione dei cerchi $\mathcal{F}_i$, detti anche *cerchi di Gershgorin*.

Teorema 2.8.2 (secondo teorema di Gershgorin) *Se M_1 è l'unione di k cerchi di Gershgorin e M_2 è l'unione dei rimanenti $n-k$ ed è $M_1 \bigcap M_2 = \emptyset$, allora k autovalori appartengono a M_1 e $n-k$ a M_2.*

Teorema 2.8.3 (terzo teorema di Gershgorin) *Sia A irriducibile; se un autovalore appartiene alla frontiera dell'unione dei cerchi di Gershgorin esso appartiene alla frontiera di tutti i cerchi costituenti l'insieme $\mathcal{F}$.*

Per la localizzazione degli autovalori di una matrice è utile anche il Corollario 2.10.2.

2.9 Valori singolari

Particolare importanza in alcune applicazioni (cfr. 6.8.5) ha il seguente teorema.

Teorema 2.9.1 *Qualunque sia $A \in \mathbb{C}^{m\times n}$, con $m \geq n$, esistono due matrici unitarie $U \in \mathbb{C}^{m\times m}$ e $V \in \mathbb{C}^{n\times n}$ tali che*

$$A = U\Sigma V^H \tag{2.13}$$

con $\Sigma = \begin{pmatrix} D \\ \mathbf{O} \end{pmatrix}$, $\mathbf{O}$ la matrice nulla, $D = diag(\sigma_1, \sigma_2, \ldots, \sigma_n)$, $\sigma_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$, e

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0.$$

La (2.13) è detta *decomposizione ai valori singolari* della matrice A, la matrice Σ è univocamente determinata (non così le matrici U e V) ed i numeri σ_i, $i = 1, 2, \ldots, n$, si dicono i *valori singolari* della matrice A. Si può verificare che il numero dei valori singolari non nulli è uguale a $r(A)$, le colonne di U sono autovettori di AA^H, mentre quelle di V sono autovettori di A^HA.

Teorema 2.9.2 *Sia $A \in \mathbb{C}^{m\times n}$; i quadrati dei valori singolari di A sono autovalori della matrice $A^HA \in \mathbb{C}^{n\times n}$.*

Teorema 2.9.3 *Se A è una matrice normale di ordine n si ha*

$$\sigma_i = \mid \lambda_i \mid, \quad i = 1, 2, \ldots, n,$$

dove i numeri λ_i sono gli autovalori della matrice A.

2.10 Norme

Spesso è necessario confrontare fra loro due vettori o due matrici; a tale scopo è utile il concetto di norma.

2.10.1 Norme vettoriali

Definizione 2.10.1 *Si dice* norma vettoriale*, e si indica con* $\|x\|$*, una funzione, definita nello spazio vettoriale* $\mathbb{C}^n$*, a valori reali non negativi, che verifica le seguenti condizioni:*

$$\begin{aligned}
&\|x\| = 0 \Longleftrightarrow x = 0\,; \\
&\|\alpha x\| = \mid \alpha \mid \|x\|\,, \qquad \forall x \in \mathbb{C}^n,\ \forall \alpha \in \mathbb{C}\ ; \\
&\|x + y\| \leq \|x\| + \|y\|\,, \qquad \forall x, y \in \mathbb{C}^n\,.
\end{aligned}$$

In $\mathbb{C}^n$ è possibile definire norme in modo arbitrario ma le più usate sono le seguenti:

$$\begin{aligned}
&\|x\|_1 = \textstyle\sum_{i=1}^n \mid x_i \mid, && \textit{norma 1};\\
&\|x\|_2 = \sqrt{\textstyle\sum_{i=1}^n \mid x_i \mid^2}, && \textit{norma 2} \text{ o } \textit{norma euclidea};\\
&\|x\|_\infty = \max_{1\leq i\leq n} \mid x_i \mid, && \textit{norma } \infty.
\end{aligned}$$

In Fig. 2.1 è riportato l'insieme

$$S = \{x \in I\!R^2 \mid \|x\| \leq 1\},$$

detto *sfera unitaria* di $I\!R^2$, per le tre norme definite.

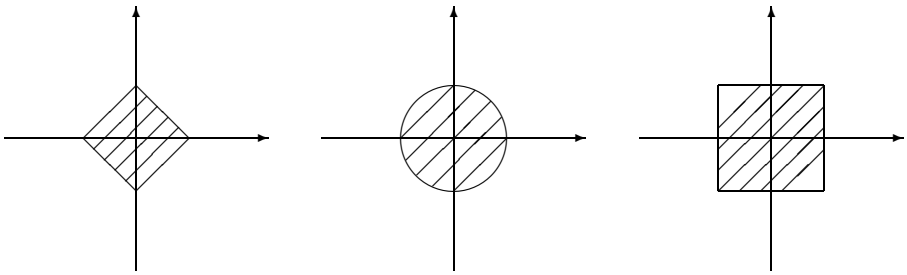

Figura 2.1: Sfere unitarie per le norme 1,2,∞.

Teorema 2.10.1 *Ogni norma vettoriale è uniformemente continua su* $\mathbb{C}^n$*.*

Teorema 2.10.2 (equivalenza tra norme) *Date due norme vettoriali* $\|x\|_p$ *e* $\|x\|_q$*, esistono due costanti reali e positive* α *e* β *tali che*

$$\alpha\|x\|_p \leq \|x\|_q \leq \beta\|x\|_p, \quad \forall x \in \mathcal{C}^n. \tag{2.14}$$

Il senso dell'equivalenza fra due norme appare evidente nello studio del comportamento di una successione di vettori: in virtù della (2.14) il carattere della successione è lo stesso in entrambe le norme.

Per le norme qui definite valgono le relazioni:

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2 \ ;$$
$$\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty \ ;$$
$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty \ .$$

2.10.2 Norme matriciali

Definizione 2.10.2 *Si dice* norma matriciale*, e si indica con* $\|A\|$*, una funzione, definita in* $\mathcal{C}^{n\times n}$*, a valori reali non negativi, che verifica le seguenti condizioni:*

$$\|A\| = 0 \Longleftrightarrow A = \mathbf{O}\,;$$
$$\|\alpha A\| = \mid \alpha \mid \|A\|\,, \qquad \forall A \in \mathcal{C}^{n\times n},\ \forall \alpha \in \mathcal{C}\ ;$$
$$\|A + B\| \leq \|A\| + \|B\|\,, \qquad \forall A, B \in \mathcal{C}^{n\times n}\ ;$$
$$\|AB\| \leq \|A\|\|B\|\,, \qquad \forall A, B \in \mathcal{C}^{n\times n}\ .$$

Si possono costruire norme matriciali facendo ricorso alle norme vettoriali definendo

$$\|A\| = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|};$$

in questo caso la norma matriciale si dice *naturale* o *indotta* dalla norma vettoriale considerata.

Una norma matriciale si dice *coerente* o *compatibile* con una norma vettoriale se si ha

$$\|Ax\| \leq \|A\|\|x\|.$$

Le norme naturali sono coerenti con le rispettive norme vettoriali.

Si può dimostrare che le norme matriciali indotte dalle tre norme vettoriali definite in 2.10.1 sono le seguenti:

$$\begin{aligned} &\|A\|_1 = \max_j \textstyle\sum_{i=1}^n \mid a_{ij} \mid, && \textit{norma 1};\\ &\|A\|_2 = \sqrt{\rho(A^H A)}, && \textit{norma 2 o norma euclidea};\\ &\|A\|_\infty = \max_i \textstyle\sum_{j=1}^n \mid a_{ij} \mid, && \textit{norma } \infty. \end{aligned}$$

Un esempio di norma matriciale non indotta è dato dalla *norma matriciale di Frobenius* definita come

$$\|A\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n \mid a_{ij} \mid^2};$$

essa non è indotta da alcuna norma vettoriale in quanto risulta $\|I\|_F = \sqrt{n}$ mentre è $\|I\| = \sup_{x \neq 0} \frac{\|Ix\|}{\|x\|} = 1$ qualunque sia la norma vettoriale considerata.

Teorema 2.10.3 *Sia $A \in \mathbb{C}^{n \times n}$; per ogni norma matriciale indotta vale la relazione*

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

DIMOSTRAZIONE. Sia λ un autovalore di A; quindi si ha $Ax = \lambda x$ con x autovettore destro associato a λ. Prendendo una qualunque norma dei due membri si ottiene

$$\mid \lambda \mid \|x\| = \|Ax\| ,$$

da cui, se si usa la norma matriciale indotta da quella vettoriale,

$$\mid \lambda \mid \|x\| \leq \|A\| \|x\| .$$

Essendo $x \neq 0$, dividendo per $\|x\|$, si ha

$$\mid \lambda \mid \leq \|A\| .$$

Poiché λ è un qualunque autovalore di A, la relazione precedente è valida anche per l'autovalore il cui modulo coincide con $\rho(A)$, da cui la tesi. □

Corollario 2.10.1 *Affinché una matrice sia convergente è sufficiente che una sua norma indotta risulti minore di 1.*

Un esempio in cui vale la relazione $\rho(A) = \|A\|$ è dato dalle matrici hermitiane; infatti si ha

$$\|A\|_2 = \sqrt{\rho(A^H A)} = \sqrt{\rho(A^2)} = \sqrt{\rho^2(A)} = \rho(A)\,.$$

Corollario 2.10.2 *Se $A \in \mathbb{C}^{n\times n}$, gli autovalori di A appartengono al cerchio*

$$\{z \in \mathbb{C} \mid |\, z \,| \leq \|A\|\}\,,$$

dove $\|\cdot\|$ è una qualunque norma indotta.

2.11 Complementi ed esempi

2.11.1 Prodotto di due matrici

Il seguente esempio mostra che l'operazione di moltiplicazione tra matrici non gode della proprietà commutativa e che non vale la legge di annullamento del prodotto.

Esempio 2.11.1 Sono date le matrici

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}, \qquad B = \begin{pmatrix} 2 & 1 \\ -1 & -\frac{1}{2} \end{pmatrix}.$$

Si ha

$$AB = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \qquad BA = \begin{pmatrix} 4 & 8 \\ -2 & -4 \end{pmatrix}.$$

□

2.11.2 Matrici definite e semidefinite

Teorema 2.11.1 *Le matrici hermitiane e definite positive (negative) hanno autovalori positivi (negativi) e viceversa.*

La dimostrazione della parte diretta del teorema risulta evidente ricorrendo alla (2.8).

Teorema 2.11.2 *Le matrici hermitiane e semidefinite positive (negative) hanno autovalori non negativi (non positivi) e viceversa.*

Esempio 2.11.2 Sia $A \in \mathbb{C}^{m\times n}$. La matrice $B = A^H A$ risulta hermitiana essendo

$$B^H = (A^H A)^H = A^H A = B\,.$$

Il numero $x^H Bx$, con $x \in \mathbb{C}^n$, $x \neq 0$, è dato da

$$x^H Bx = x^H A^H Ax = (Ax)^H (Ax)$$

per cui, ponendo $y = Ax$, si ha

$$x^H Bx = y^H y \geq 0\,.$$

La matrice B è quindi semidefinita positiva. Se la matrice A ha rango massimo, il vettore y non può essere nullo per cui $x^H Bx > 0$ e la matrice B è definita positiva. □

2.11.3 Matrici riducibili

Riportiamo due esempi ad illustrazione del Teorema 2.6.1.

Esempio 2.11.3 La matrice

$$A = \begin{pmatrix} 2 & 0 & 1 \\ -2 & 1 & 1 \\ 5 & 0 & 0 \end{pmatrix}$$

è riducibile infatti il suo grafo orientato non è fortemente connesso, non essendo il punto N_2 raggiungibile dagli altri due punti (Fig. 2.2).

Mediante la matrice $P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$ si ottiene

$$B = P^T AP = \left(\begin{array}{cc|c} 2 & 1 & 0 \\ 5 & 0 & 0 \\ \hline -2 & 1 & 1 \end{array}\right).$$

□

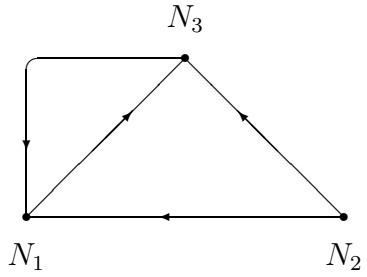

Figura 2.2: Grafo non fortemente connesso.

Esempio 2.11.4 La matrice

$$A = \begin{pmatrix} 1 & 0 & 1 \\ -1 & 2 & 5 \\ 1 & 1 & 0 \end{pmatrix}$$

è irriducibile in quanto il suo grafo orientato è fortemente connesso, come mostrato in Fig. 2.3. □

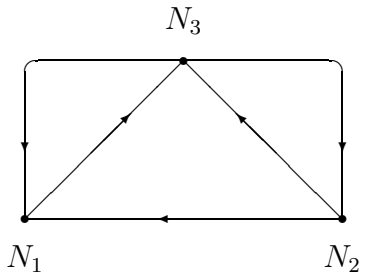

Figura 2.3: Grafo fortemente connesso.

Sia $A \in \mathbb{C}^{n\times n}$ una matrice riducibile e sia P la matrice di permutazione che la riduce. Poniamo

$$B = P^T AP = \begin{pmatrix} B_{11} & \mathbf{O} & \cdots & \mathbf{O} \\ B_{21} & B_{22} & \cdots & \mathbf{O} \\ \vdots & \vdots & \ddots & \vdots \\ B_{k1} & B_{k2} & \cdots & B_{kk} \end{pmatrix},$$

con i blocchi diagonali B_{ii}, $i = 1, 2, \ldots, k$, quadrati ed irriducibili; supponiamo che l'ordine dei blocchi B_{ii} sia p_i, $i = 1, 2, \ldots, k$, per cui $\sum_{i=1}^{k} p_i = n$.

La matrice A sia la matrice dei coefficienti di un sistema lineare $Ax = b$. Premoltiplicando per la matrice P^T si ha $P^T Ax = P^T b$ e $P^T APP^T x = P^T b$. Indicando con y il vettore $P^T x$ e con c il vettore $P^T b$, il sistema $Ax = b$ si trasforma nel sistema

$$By = c. \tag{2.15}$$

Partizionando i vettori y e c in blocchi di pari dimensione dei B_{ii}, $i = 1, 2, \ldots, k$, il sistema (2.15), scritto in forma esplicita, diviene

$$\begin{array}{lllllll} B_{11}y_1 & & & & & = & c_1 \\ B_{21}y_1 & + & B_{22}y_2 & & & = & c_2 \\ \vdots & & \vdots & \ddots & & \vdots & \vdots \\ B_{k1}y_1 & + & B_{k2}y_2 & + \cdots + & B_{kk}y_k & = & c_k\,. \end{array} \tag{2.16}$$

La prima equazione è un sistema lineare, con matrice dei coefficienti B_{11}, di ordine p_1, la cui incognita è il vettore y_1; si risolve tale sistema e si sostituisce il vettore y_1 nelle equazioni seguenti. La seconda equazione diviene un sistema lineare, quadrato di ordine p_2, da cui si ricava il vettore y_2 che può essere sostituito nelle equazioni seguenti. Procedendo in questo modo si ricavano tutti i blocchi y_i, $i = 1, 2, \ldots, k$, che costituiscono il vettore y. Una volta ottenuto l'intero vettore y si risale al vettore x tramite la relazione $x = Py$.

Si osservi che se la matrice A è non singolare tale è anche la matrice B in quanto ottenuta da A con una trasformazione per similitudine. La matrice B ha il determinante uguale al prodotto dei determinanti dei blocchi diagonali per cui se B è non singolare tali sono i blocchi B_{ii}, $i = 1, 2, \ldots, k$. Questo assicura l'esistenza e l'unicità della soluzione di tutti i sistemi lineari che via via si risolvono nella (2.16).

La sostituzione del sistema lineare $Ax = b$ con il sistema lineare (2.15) conduce alla risoluzione di k sistemi lineari tutti di ordine inferiore ad n al posto di un unico sistema lineare di ordine n. Il vantaggio dell'uso di questa trasformazione risulterà evidente nel Capitolo 3 quando saranno esposti i metodi numerici per la risoluzione dei sistemi lineari.

Si noti che le matrici A e B sono simili e quindi hanno gli stessi autovalori. Per il calcolo degli autovalori di B si può utilizzare la seguente osservazione.

Osservazione 2.11.1 Gli autovalori di una matrice triangolare o diagonale a blocchi, con blocchi diagonali quadrati, sono tutti e soli gli autovalori dei

blocchi diagonali.

Il seguente esempio mostra come si può costruire la matrice di permutazione P che porta in forma ridotta una data matrice A riducibile.

Esempio 2.11.5 Si consideri la matrice riducibile

$$A = \begin{pmatrix} 5 & 0 & 3 & 0 & 0 \\ 1 & 1 & 1 & 1 & 5 \\ 1 & 0 & 2 & 0 & 0 \\ 1 & 2 & 1 & 3 & 0 \\ 0 & 0 & 1 & 0 & 6 \end{pmatrix}.$$

Per trovare la matrice di permutazione che trasforma A nella sua forma ridotta, si costruisce il grafo orientato associato alla matrice A, avente i nodi N_1, N_2, N_3, N_4, N_5.

Da tale grafo si constata la seguente situazione

Nodo di partenza	Nodi raggiungibili	Nodi non raggiungibili
N_1	N_1, N_3	N_2, N_4, N_5;
N_2	tutti	$---$;
N_3	N_1, N_3	N_2, N_4, N_5;
N_4	tutti	$---$;
N_5	N_1, N_3, N_5	N_2, N_4.

Essendo N_1 il primo nodo a partire dal quale non è possibile raggiungere tutti gli altri nodi, si riordinano i nodi ponendo ai primi posti quei nodi che sono raggiungibili partendo da N_1 e di seguito gli altri. Si ha quindi un nuovo ordinamento dei nodi che è $Q_1 = N_1, Q_2 = N_3, Q_3 = N_2, Q_4 = N_4, Q_5 = N_5$. A questo punto si opera sulla matrice A una trasformazione per similitudine mediante la matrice di permutazione

$$P_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

ottenuta effettuando sulle colonne della matrice identica la stessa permutazione operata sui nodi N_i per ottenere i nodi Q_i.

Si ottiene la matrice

$$A_1 = P_1^T A P_1 = \begin{pmatrix} 5 & 3 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 5 \\ 1 & 1 & 2 & 3 & 0 \\ 0 & 1 & 0 & 0 & 6 \end{pmatrix}.$$

La matrice A_1 è triangolare a blocchi con blocchi diagonali

$$\begin{pmatrix} 5 & 3 \\ 1 & 2 \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 & 5 \\ 2 & 3 & 0 \\ 0 & 0 & 6 \end{pmatrix}.$$

Il secondo blocco diagonale risulta ancora riducibile in quanto nel suo grafo orientato esiste un nodo (quello di indice massimo) da cui non si possono raggiungere gli altri due.

Su questo blocco si opera analogamente a quanto fatto su A; ciò equivale ad effettuare una nuova trasformazione per similitudine sulla matrice A_1 mediante la matrice di permutazione

$$P_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

ottenendo la matrice

$$A_2 = P_2^T A_1 P_2 = \begin{pmatrix} 5 & 3 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 \\ 0 & 1 & 6 & 0 & 0 \\ 1 & 1 & 5 & 1 & 1 \\ 1 & 1 & 0 & 2 & 3 \end{pmatrix}.$$

I blocchi diagonali della matrice A_2 risultano irriducibili per cui essa è la forma ridotta di A.

Alla matrice A_2 si può giungere con un'unica trasformazione per similitudine mediante la matrice di permutazione

$$P = P_1 P_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

□

2.11.4 Autovalori e autovettori di matrici particolari

Esempio 2.11.6 Le matrici della forma

$$G = \begin{pmatrix} 1 & & & & & & & & & & \\ & \ddots & & & & & & & & & \\ & & 1 & & & & & & & & \\ & & & c & & & & -s & & & \leftarrow p \\ & & & & 1 & & & & & & \\ & & & & & \ddots & & & & & \\ & & & & & & 1 & & & & \\ & & & s & & & & c & & & \leftarrow q \\ & & & & & & & & 1 & & \\ & & & & & & & & & \ddots & \\ & & & & & & & & & & 1 \end{pmatrix},$$

dove $G \in \mathbb{R}^{n\times n}$ e $c^2 + s^2 = 1$, si dicono *matrici di rotazione*.

Si vogliono calcolare gli autovalori e gli autovettori di G nel caso $s \neq 0$.

Essendo G ortogonale, si ha $\det(GG^T) = 1$, per cui risulta anche $\det(G) = \lambda_1\lambda_2\cdots\lambda_n = \pm 1$.

Dal Teorema 2.10.3 si ottiene $|\ \lambda_i\ | \leq \|G\|_2 = \sqrt{\rho(GG^T)} = 1$; ne segue che può essere solo $|\ \lambda_i\ | = 1$, $i = 1, 2, \ldots, n$.

L'equazione caratteristica della matrice G è

$$(1-\lambda)^{n-2}(\lambda^2 - 2c\lambda + 1) = 0\,;$$

pertanto gli autovalori di G sono

$$\lambda_1 = \cdots = \lambda_{n-2} = 1, \quad \lambda_{n-1} = c - \sqrt{-s^2}, \quad \lambda_n = c + \sqrt{-s^2}\,.$$

Poiché in corrispondenza all'autovalore 1 risulta $r(G - I) = 2$, si ha $\alpha(1) = \gamma(1) = n - 2$; quindi la matrice G ammette n autovettori linearmente indipendenti che si ricavano risolvendo i sistemi $(G - \lambda_i I)x = 0$, $i = 1, 2, \ldots, n$.

Con semplici calcoli si ottiene la matrice degli autovettori

$$X = \begin{pmatrix}
1 & & & & & & & & & & \\
& \ddots & & & & & & & & & \\
& & 1 & & & & & & & & \\
& & & -\frac{1}{\sqrt{2}}\frac{\sqrt{-s^2}}{s} & & & & \frac{1}{\sqrt{2}} & & & \leftarrow p \\
& & & & 1 & & & & & & \\
& & & & & \ddots & & & & & \\
& & & & & & 1 & & & & \\
& & & \frac{1}{\sqrt{2}} & & & & \frac{1}{\sqrt{2}}\frac{\sqrt{-s^2}}{s} & & & \leftarrow q \\
& & & & & & & & 1 & & \\
& & & & & & & & & \ddots & \\
& & & & & & & & & & 1
\end{pmatrix}.$$

Si noti che, essendo G normale, X è unitaria (cfr. Teorema 2.7.8). □

Esempio 2.11.7 Si consideri la matrice $H = I - 2vv^H$ con $v \in \mathbb{C}^n$ e $v^H v = 1$; la matrice H è detta *matrice di Householder*.

Si vogliono calcolare gli autovalori e gli autovettori di H.

A tale scopo, si opera su H una traslazione di spettro e si calcolano gli autovalori della matrice $B = H - I = -2vv^H$. L'equazione caratteristica di B, tenuto conto che $v^H v = 1$, è

$$\mu^n + 2\mu^{n-1} = 0\,.$$

Si ricavano quindi gli autovalori di B che sono

$$\mu_1 = 0, \quad \mu_2 = -2, \quad \text{con} \quad \alpha(\mu_1) = n - 1, \quad \alpha(\mu_2) = 1\,.$$

Gli autovalori di H sono quindi

$$\lambda_1 = 1, \quad \lambda_2 = -1, \quad \text{con} \quad \alpha(\lambda_1) = n-1, \quad \alpha(\lambda_2) = 1\,.$$

Si verifica facilmente che $\gamma(\lambda_1) = \alpha(\lambda_1)$ e $\gamma(\lambda_2) = \alpha(\lambda_2)$, quindi gli autovettori linearmente indipendenti di H sono n. Gli autovettori associati a λ_1 sono

$$x_1 = \begin{pmatrix} \overline{v}_2 \\ -\overline{v}_1 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \end{pmatrix}, x_2 = \begin{pmatrix} \overline{v}_3 \\ 0 \\ -\overline{v}_1 \\ \vdots \\ 0 \\ 0 \\ 0 \end{pmatrix}, \ldots, x_{n-1} = \begin{pmatrix} \overline{v}_n \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ -\overline{v}_1 \end{pmatrix},$$

mentre l'autovettore associato a λ_2 è $x_n = v$.

Si osservi, infine, che la matrice H è una matrice hermitiana e unitaria. □

Esempio 2.11.8 Si consideri la matrice $J \in \mathbb{R}^{2n\times 2n}$ i cui elementi sono

$$(J)_{ik} = \begin{cases} 1 & \text{se } i+k = 2n+1, \\ 0 & \text{altrimenti.} \end{cases}$$

Essendo $J^2 = I$, indicato con λ un autovalore di J, si ha $\lambda^2 = 1$; per cui gli autovalori di J possono assumere i valori 1 o -1.

Osservando che $tr(J) = 0$ (cfr. (2.6)) si ricavano gli autovalori

$$\lambda_1 = 1, \quad \lambda_2 = -1 \quad \text{con} \quad \alpha(\lambda_1) = \alpha(\lambda_2) = n.$$

I due autovalori hanno molteplicità geometrica $\gamma(\lambda_1) = \gamma(\lambda_2) = n$. Gli autovettori associati a λ_1 sono

$$x_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 1 \end{pmatrix}, x_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \end{pmatrix}, \ldots, x_n = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

mentre gli autovettori associati a λ_2 sono

$$x_{n+1} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ -1 \end{pmatrix}, x_{n+2} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ -1 \\ 0 \end{pmatrix}, \ldots, x_{2n} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ -1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Si osservi che gli autovettori di J sono anche autovettori di J^2 mentre non è vero, in generale, il viceversa. Infatti, la matrice J^2 ammette come autovettore un qualunque vettore di $\mathbb{C}^n$. □

2.11.5 Matrici tridiagonali

Particolare interesse hanno le *matrici tridiagonali* cioè le matrici T i cui elementi t_{ij} sono nulli se $\mid i - j \mid > 1$, $i, j = 1, 2, \ldots, n$.

Pertanto possono scriversi nella seguente forma generale

$$T = \begin{pmatrix} a_1 & c_2 & & \\ b_2 & \ddots & \ddots & \\ & \ddots & \ddots & c_n \\ & & b_n & a_n \end{pmatrix}. \tag{2.17}$$

Per le matrici tridiagonali (2.17) il calcolo del determinante può effettuarsi mediante la relazione ricorrente

$$\begin{aligned} &D_0 = 1, \\ &D_1 = a_1, \\ &D_i = a_i D_{i-1} - b_i c_i D_{i-2}, \quad i = 2, 3, \ldots, n, \end{aligned}$$

il cui termine D_n è $\det(T)$.

In modo del tutto analogo si può calcolare $\det(T - \lambda I)$, cioè il polinomio caratteristico della matrice (2.17), mediante la successione di polinomi

$$
\begin{aligned}
&P_0(\lambda) = 1,\\
&P_1(\lambda) = a_1 - \lambda,\\
&P_i(\lambda) = (a_i - \lambda)P_{i-1}(\lambda) - b_i c_i P_{i-2}(\lambda), \quad i = 2, 3, \ldots, n,
\end{aligned}
$$

dove risulta $P_n(\lambda) = \det(T - \lambda I)$.

Esempio 2.11.9 Si calcola il determinante della matrice di ordine n

$$
T_n = \begin{pmatrix}
1 & 1 & & \\
-1 & \ddots & \ddots & \\
 & \ddots & \ddots & 1 \\
 & & -1 & 1
\end{pmatrix},
$$

caso particolare della (2.17).

Per $\det(T_n) = D_n$ risulta

$$
D_n = D_{n-1} + D_{n-2}\,. \tag{2.18}
$$

La (2.18), per $n = 2, 3, \ldots$, è una equazione lineare alle differenze (cfr. 8.3.1) la cui equazione caratteristica è $\mu^2 - \mu - 1 = 0$. La soluzione generale è $D_n = c_1\mu_1^n + c_2\mu_2^n$, con $\mu_1 = \frac{1+\sqrt{5}}{2}$ e $\mu_2 = \frac{1-\sqrt{5}}{2}$. Le costanti arbitrarie c_1 e c_2 possono essere determinate imponendo che per $n = 0$ ed $n = 1$, D_n assuma i valori dati per D_0 e D_1. Ponendo $D_0 = D_1 = 1$ e risolvendo quindi il sistema

$$
\begin{aligned}
c_1 \quad &+ \quad c_2 \quad &= \quad 1\\
c_1\mu_1 \quad &+ \quad c_2\mu_2 \quad &= \quad 1\,,
\end{aligned}
$$

si ricava $c_1 = \mu_1/\sqrt{5}$ e $c_2 = -\mu_2/\sqrt{5}$, da cui

$$
D_n = \frac{1}{\sqrt{5}}\mu_1^{n+1} - \frac{1}{\sqrt{5}}\mu_2^{n+1}\,.
$$

Quindi, ad esempio,

$$
\det(T_{100}) \simeq \frac{1}{\sqrt{5}}(1.618)^{101} \simeq 5.73 \times 10^{20}\,.
$$

La successione $\{D_n\}$ è nota con il nome di *successione di Fibonacci.* □

2.11.6 Conseguenze dei teoremi di Gershgorin

I teoremi di Gershgorin consentono di dedurre alcune proprietà notevoli.

Corollario 2.11.1 *Una matrice A a predominanza diagonale forte è non singolare.*

DIMOSTRAZIONE. Segue immediatamente dal primo teorema di Gershgorin tenendo presente che ogni cerchio di Gershgorin ha il centro a distanza $\mid a_{ii} \mid$ dall'origine degli assi ed il raggio $\rho_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} \mid a_{ij} \mid$; dall'ipotesi segue che nessuno dei detti cerchi contiene l'origine. Lo zero non è quindi autovalore della matrice ed il determinante, per la (2.7), non può essere nullo. □

Corollario 2.11.2 *Una matrice A a predominanza diagonale debole ed irriducibile è non singolare.*

DIMOSTRAZIONE. Ragionando come nel precedente corollario, si osserva che la predominanza diagonale debole consente ai cerchi di Gershgorin di avere la circonferenza passante per l'origine, eccettuato uno almeno di tali cerchi. D'altra parte se lo zero fosse autovalore della matrice irriducibile A, per il terzo teorema di Gershgorin esso dovrebbe appartenere alla frontiera di tutti i cerchi, contrariamente a quanto osservato. Dunque lo zero non è autovalore e perciò $\det(A) \neq 0$. □

Esempio 2.11.10 Sia

$$A = \begin{pmatrix} 2 & 2 & 0 \\ i & 2i & 1 \\ -1 & 0 & -2i \end{pmatrix}, \quad \text{dove} \quad i^2 = -1;$$

gli autovalori appartengono all'insieme rappresentato in Fig. 2.4 che è l'unione dei tre insiemi

$$\mathcal{F}_1 = \{z \in \mathbb{C} \mid\mid z - 2 \mid \leq 2\},$$
$$\mathcal{F}_2 = \{z \in \mathbb{C} \mid\mid z - 2i \mid \leq 2\},$$
$$\mathcal{F}_3 = \{z \in \mathbb{C} \mid\mid z + 2i \mid \leq 1\}.$$

La matrice A è a predominanza diagonale debole ed irriducibile; i suoi autovalori sono non nulli essendo $\det(A) = 2$. □

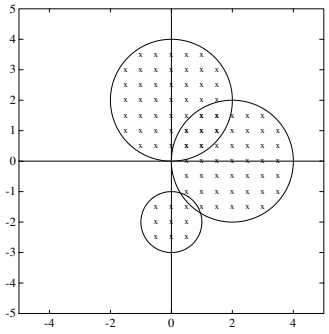

Figura 2.4: cerchi di Gershgorin.

Una interessante applicazione del Teorema 2.8.1 è la possibilità di localizzare le radici reali e complesse di una equazione algebrica.

Sia data l'equazione

$$x^k + a_{k-1}x^{k-1} + \cdots + a_1 x + a_0 = 0 \tag{2.19}$$

con $a_i \in \mathcal{C}$, $i = 0, 1, \ldots, k-1$.

Si consideri la matrice quadrata di ordine k

$$F = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{k-2} & -a_{k-1} \end{pmatrix}; \tag{2.20}$$

si verifica che la (2.19) è l'equazione caratteristica della matrice F: pertanto i suoi autovalori sono le radici dell'equazione assegnata. Quindi è possibile localizzare le radici dell'equazione (2.19) facendo uso del Teorema 2.8.1.

La matrice (2.20) è detta *matrice di Frobenius* o *matrice compagna* dell'equazione (2.19).

Esempio 2.11.11 Si consideri l'equazione algebrica

$$x^3 - 5x^2 + 3x - 1 = 0.$$

Le radici dell'equazione sono gli autovalori della matrice compagna

$$F = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & -3 & 5 \end{pmatrix}.$$

Applicando il primo teorema di Gershgorin alla matrice F, le radici dell'equazione appartengono all'insieme $\mathcal{F}$ unione dei cerchi

$$\mathcal{F}_1 = \mathcal{F}_2 = \{z \in \mathbb{C} \mid\mid z \mid \leq 1\}, \quad \mathcal{F}_3 = \{z \in \mathbb{C} \mid\mid z - 5 \mid \leq 4\}.$$

Se si considera la matrice F^T e si applica ad essa il teorema di Gershgorin (ciò equivale ad operare sulle colonne di F) si ottiene l'insieme $\mathcal{G}$ unione dei cerchi

$$\begin{aligned} \mathcal{G}_1 &= \{z \in \mathbb{C} \mid\mid z \mid \leq 1\}, \\ \mathcal{G}_2 &= \{z \in \mathbb{C} \mid\mid z \mid \leq 4\}, \\ \mathcal{G}_3 &= \{z \in \mathbb{C} \mid\mid z - 5 \mid \leq 1\}. \end{aligned}$$

Ricordando che gli autovalori delle matrici F e F^T sono gli stessi, si deduce che gli autovalori, quindi le radici dell'equazione proposta, appartengono all'insieme $\mathcal{F} \bigcap \mathcal{G}$, che è più ristretto sia di $\mathcal{F}$ che di $\mathcal{G}$, come mostra la Fig. 2.5. □

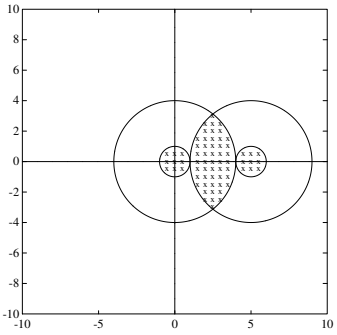

Figura 2.5: Intersezione $\mathcal{F} \bigcap \mathcal{G}$.

2.11.7 Polinomio minimo

Teorema 2.11.3 (di Cayley-Hamilton) *Una matrice $A \in \mathbb{C}^{n\times n}$ verifica la sua equazione caratteristica; si ha cioè*

$$P(A) = \mathbf{O}$$

essendo il secondo membro la matrice nulla e $P(\lambda)$ il polinomio caratteristico di A.

Si enuncia brevemente il Teorema 2.11.3 dicendo che A annulla $P(\lambda)$; è evidente come la matrice A annulli qualunque polinomio di cui $P(\lambda)$ è divisore; inoltre A può annullare anche polinomi di grado inferiore a n.

Definizione 2.11.1 *Data una matrice $A \in \mathbb{C}^{n\times n}$ si dice* polinomio minimo*, e si indicherà con $P_M(\lambda)$, il polinomio di grado più basso per il quale si ha*

$$P_M(A) = \mathbf{O}.$$

Si può dimostrare che $P_M(\lambda)$ è un divisore di $P(\lambda)$ e che è della forma

$$P_M(\lambda) = (\lambda - \lambda_1)^{p^{(1)}}(\lambda - \lambda_2)^{p^{(2)}} \cdots (\lambda - \lambda_k)^{p^{(k)}}$$

dove $p^{(i)}$ è l'ordine massimo dei blocchi di Jordan $J_l^{(i)}$ (cfr. Teorema 2.7.9), mentre $\lambda_1, \lambda_2, \ldots, \lambda_k$, sono gli autovalori distinti di A.

Esempio 2.11.12 Si calcolano i polinomi minimi di alcune matrici particolari.

1. Sia G una matrice di rotazione di ordine n con $s \neq 0$ (cfr. Esempio 2.11.6). Essa ha equazione caratteristica

   $$(\lambda - 1)^{n-2}(\lambda^2 - 2c\lambda + 1) = 0\,;$$

   si ha $\gamma(1) = n - 2$, quindi i blocchi di Jordan sono tutti di ordine 1 e il polinomio minimo di G risulta

   $$P_M(\lambda) = (\lambda - 1)(\lambda^2 - 2c\lambda + 1)\,.$$

2. Sia H una matrice di Householder (cfr. Esempio 2.11.7). Essendo H unitaria ed hermitiana, si ha $H^2 = I$ per cui $H^2 - I = \mathbf{O}$ ed il polinomio minimo è
$$P_M(\lambda) = \lambda^2 - 1$$
in quanto per nessun valore α può risultare $H + \alpha I = \mathbf{O}$.

3. Sia J la matrice definita nell'Esempio 2.11.8. Anche per J si ha $J^2 = I$ per cui il suo polinomio minimo, come per tutte le matrici ortogonali e simmetriche, risulta
$$P_M(\lambda) = \lambda^2 - 1\,.$$

□

La conoscenza di un polinomio che sia annullato da una matrice non singolare può essere sfruttata per il calcolo della matrice inversa come mostrato nel seguente esempio.

Esempio 2.11.13 Sia A la matrice reale di ordine n definita da
$$A = \begin{pmatrix} n & 1 & 1 & \cdots & 1 & 1 \\ 1 & n & 1 & \cdots & 1 & 1 \\ 1 & 1 & n & \cdots & 1 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & 1 & \cdots & n & 1 \\ 1 & 1 & 1 & \cdots & 1 & n \end{pmatrix}.$$

Conviene operare una traslazione di spettro (cfr. Teorema 2.7.6) e considerare la matrice $A - (n-1)I$. Per essa l'equazione caratteristica (2.4) è $\mu^n - n\mu^{n-1} = 0$ da cui gli autovalori $\mu_1 = \mu_2 = \cdots = \mu_{n-1} = 0$ e $\mu_n = n$. Gli autovalori di A sono perciò
$$\lambda_1 = \lambda_2 = \cdots = \lambda_{n-1} = n - 1, \quad \lambda_n = 2n - 1.$$

Poiché A è hermitiana, e quindi normale, per il Teorema 2.7.8, è diagonalizzabile, come è confermato dall'uguaglianza delle molteplicità algebrica e geometrica.

La forma canonica di Jordan della matrice A coincide con la matrice diagonale i cui elementi diagonali sono gli autovalori di A e quindi i blocchi

di Jordan sono tutti di ordine 1. Da quanto affermato sopra, il polinomio minimo della matrice A è il polinomio di secondo grado

$$P_M(\lambda) = (\lambda - (n-1))(\lambda - (2n-1)) = \lambda^2 + (2-3n)\lambda + 2n^2 - 3n + 1.$$

Poiché risulta $P_M(A) = \mathbf{O}$, premoltiplicando entrambi i membri per la matrice inversa A^{-1}, si ottiene

$$\begin{aligned} A^{-1} &= \frac{1}{2n^2-3n+1}((3n-2)I - A) \\ &= \frac{1}{2n^2-3n+1} \begin{pmatrix} 2n-2 & -1 & -1 & \cdots & -1 & -1 \\ -1 & 2n-2 & -1 & \cdots & -1 & -1 \\ -1 & -1 & 2n-2 & \cdots & -1 & -1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ -1 & -1 & -1 & \cdots & 2n-2 & -1 \\ -1 & -1 & -1 & \cdots & -1 & 2n-2 \end{pmatrix}. \end{aligned}$$

□

Bibliografia: [2], [5], [13], [15], [31].

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

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

dove i *coefficienti* a_{ij} e i *termini noti* b_i sono numeri reali, si cerca un vettore $x^T = (x_1, x_2, \ldots, x_n)$ 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

$$Ax = b\,, \tag{3.2}$$

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

$$Rx = c\,, \tag{3.3}$$

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

Il sistema (3.3) è quindi della forma

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

e si risolve immediatamente con le formule

$$\begin{aligned} x_n &= c_n/r_{nn} \\ x_i &= (c_i - \textstyle\sum_{j=i+1}^{n} r_{ij}x_j)/r_{ii},\ i = n-1, \ldots, 1. \end{aligned}$$

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 $i = 2, 3, \ldots, n$. 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, $i = 2, 3, \ldots, n$, la prima equazione i cui membri siano stati moltiplicati per il coefficiente, detto appunto *moltiplicatore*,

$$l_{i1} = \frac{a_{i1}}{a_{11}}\,. \tag{3.5}$$

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

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

dove

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

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 $l_{i2} = \frac{a_{i2}^{(2)}}{a_{22}^{(2)}}$, $i = 3,4,\ldots,n$, e così via. Supposto che tale procedimento possa ripetersi $n-1$ volte, si giunge al sistema

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

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

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

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

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

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

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 $a_{11}^{(1)}$, $a_{22}^{(2)}$, $a_{33}^{(3)}$,... 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

$$\max_{1 \le i \le n} \mid a_{i1}^{(1)} \mid = \mid a_{r1}^{(1)} \mid ; \tag{3.10}$$

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

$$\max_{2 \le i \le n} \mid a_{i2}^{(2)} \mid = \mid a_{s2}^{(2)} \mid ; \tag{3.11}$$

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 $a_{r1}^{(1)}$, $a_{s2}^{(2)}$, ..., tali che si abbia

$$\frac{\mid a_{r1}^{(1)} \mid}{m_r^{(1)}} = \max_{1\le i\le n} \frac{\mid a_{i1}^{(1)} \mid}{m_i^{(1)}}, \quad \frac{\mid a_{s2}^{(2)} \mid}{m_s^{(2)}} = \max_{2\le i\le n} \frac{\mid a_{i2}^{(2)} \mid}{m_i^{(2)}}, \ldots\ldots , \tag{3.12}$$

dove i numeri $m_i^{(1)} = \max_{1\le j\le n} \mid a_{ij}^{(1)} \mid$, $i = 1, 2, \ldots, n$, vanno calcolati sulla matrice A del sistema di partenza, i numeri $m_i^{(2)} = \max_{2\le j\le n} \mid a_{ij}^{(2)} \mid$, $i = 2, 3, \ldots, n$, 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*

$$A = LR\,, \tag{3.13}$$

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

Si constata che

$$A_1 = H_1 A,\ A_2 = H_2 A_1,\ \ldots,\ A_{n-1} = H_{n-1} A_{n-2} = R \tag{3.14}$$

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

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

da cui

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

□

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

$$PA = L_p R_p \tag{3.15}$$

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

$$Lc = b, \qquad Rx = c. \tag{3.16}$$

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

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

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) $l_{ii} = 1$, $i = 1, 2, \ldots, n$, 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
$$a_{1j} = l_{11} r_{1j}, \quad j = 1, 2, \ldots, n;$$
2. si pone $j = 1$ e si ricavano le l_{i1} per la prima colonna di L dalle $n - 1$ equazioni
$$a_{i1} = l_{i1} r_{11}, \quad i = 2, 3, \ldots n;$$
3. si pone $i = 2$ e si ricavano le r_{2j} per la seconda riga di R da
$$a_{2j} = \sum_{h=1}^{2} l_{2h} r_{hj}, \quad j = 2, 3, \ldots n.$$

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.

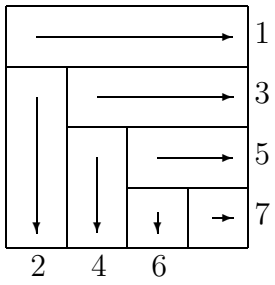

Figura 3.1: Metodo di Doolittle.

Se nelle (3.17) si pone $r_{ii} = 1$, $i = 1, 2, \ldots, n$, 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
$$a_{i1} = l_{i1} r_{11}, \quad i = 1, 2, \ldots, n;$$

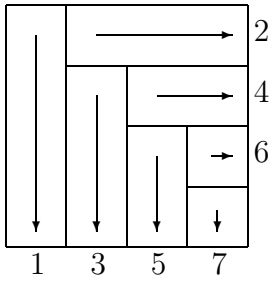

Figura 3.2: Metodo di Crout.

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

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

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

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

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

$$A = LL^T \tag{3.18}$$

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 $i \geq j$,

$$a_{ij} = l_{i1} l_{j1} + l_{i2} l_{j2} + \cdots + l_{ij} l_{jj}, \quad j = 1, 2, \ldots, n, \tag{3.19}$$

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

$$RX = I\,, \qquad LY = I\,,$$

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 δA, δb 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:*

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

dove

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

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 $\mu(A)/\left(1 - \mu(A)\frac{\|\delta A\|}{\|A\|}\right)$.

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

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

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

$$x = Hx + c\,. \tag{3.23}$$

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

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

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 $\left\{x^{(k)}\right\}$ converge alla soluzione del sistema dato.

La convergenza o meno della successione $\left\{x^{(k)}\right\}$ 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

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

da cui

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

perciò, per un arbitrario $e^{(0)}$, si avrà $\lim_{k\to\infty} e^{(k)} = 0$ se e solo se $\lim_{k\to\infty} H^k = \mathbf{O}$. □

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*

$$\rho(H) < 1\,. \tag{3.26}$$

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

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

quindi asintoticamente, cioè per k abbastanza grande, si ha

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

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

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

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

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

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

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

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

$$V = \frac{m}{k} = -\mathrm{Log}\,\rho(H) \tag{3.30}$$

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

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

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

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

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:

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

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 δ_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

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

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 $D = \mathrm{diag}(a_{11}, a_{22}, \ldots, a_{nn})$ 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 $a_{ii} \neq 0$, $i = 1, 2, \ldots, n$, si ottiene il *metodo di Jacobi*

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

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

Le equazioni del sistema (3.35) sono date da

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

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 $a_{ii} \neq 0$, $i = 1, 2, \ldots, n$, si ottiene il *metodo di Gauss-Seidel*

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

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

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

dove le singole equazioni sono

$$\begin{aligned} x_i^{(k+1)} = \tfrac{1}{a_{ii}} \left(b_i - \textstyle\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right), \\ i = 1, 2, \ldots, n, \quad k = 0, 1, \ldots. \end{aligned} \tag{3.39}$$

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 $a_{ii} \neq 0$, $i = 1, 2, \ldots, n$. 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.

Ammesso 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}, $i, j = 1, 2, \ldots, m$, $m < n$, 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:

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

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

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 $\tilde{D} = \mathrm{diag}(A_{11}, A_{22}, \ldots, A_{mm})$, 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 $\mid i - j \mid > 1$, e se si ha $\det(A_{ii}) \neq 0$, $i = 1, 2, \ldots, m$, 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

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

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 ω è un parametro reale, che, opportunamente scelto, può servire ad accelerare la convergenza del metodo. Si ha così il *metodo di rilassamento* definito da

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

ossia, tenendo conto della (3.42),

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

dove le equazioni a elementi sono

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

La matrice di iterazione H_ω 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 ω in modo che sia

$$0 < \omega < 2 . \tag{3.44}$$

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 ω (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

$$Ax = b ; \tag{3.45}$$

se si considera il funzionale

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

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 $\left\{x^{(k)}\right\}$ a cui corrisponda una successione $\left\{F(x^{(k)})\right\}$ 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

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

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

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

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

$$d^{(0)} = r^{(0)} = b - Ax^{(0)}$$

$$d^{(k)} = r^{(k)} + \rho_k d^{(k-1)}\,, \quad r^{(k)} = r(x^{(k)}), \quad k \geq 1\,, \tag{3.49}$$

dove il numero ρ_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 ρ_k la seguente espressione

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

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

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

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 non simmetrica applicandolo al sistema equivalente $A^T Ax = A^T b$ 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 precondizionamento* 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 combinazioni 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

$$AX = I\,. \tag{3.52}$$

Si tratta quindi di risolvere n sistemi lineari $Ax^{(i)} = e^{(i)}$, $i = 1, 2, \ldots, n$, 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 *premoltiplicatori* 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^Te sottraendola dalla seconda si ha

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

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

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

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

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

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

$$A = \left(\begin{array}{ccc} 1 & t & 0 \\ t & 1 & t \\ 0 & t & 1 \end{array}\right), \quad t \in \mathbb{R}.$$

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

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

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 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

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

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%. □

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 ϵ 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. □

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. □

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. □

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 λ 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

$$\begin{cases} -By_2 = \lambda y_1 \\ -By_1 = \lambda y_2 \end{cases} \tag{3.53}$$

che può scriversi anche

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

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

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

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

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. □

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 $x_i^{(0)} = b_i/a_{ii}$, $i = 1, 2, \ldots, n$.

3.10.6 Scelta ottimale del parametro ω

Dalla scelta del parametro ω 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

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

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 ω nell'intervallo $]0, 2[$.

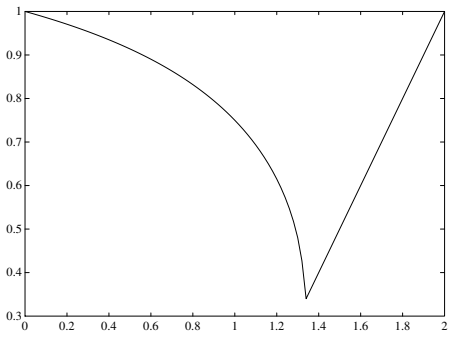

Figura 3.3: Grafico di $\rho(H_\omega)$ per la (3.54).

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

Se non si dispone del valore esatto di ω^* 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 ω.

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 ρ_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 λ_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

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

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

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

□

Bibliografia: [2], [5], [15], [29], [32]

Capitolo 4

Equazioni e sistemi non lineari

4.1 Introduzione

Sia $f(x)$:$\mathbb{R}\rightarrow\mathbb{R}$ una funzione continua almeno su un certo intervallo $\mathcal{I}$ e si supponga che $f(x)$ non sia della forma $f(x) = a_1x + a_0$ con a_1 e a_0 costanti; la relazione

$$f(x) = 0 \tag{4.1}$$

si dice *equazione non lineare* nell'incognita x.

Il problema di determinare (se esistono) gli zeri di $f(x)$, ossia di trovare le eventuali radici dell'equazione (4.1), raramente può essere risolto con un metodo diretto, cioè effettuando sulla (4.1) un insieme finito di operazioni che conducano ad una espressione esplicita di ciascuna radice. Come esempio basti considerare il caso di una equazione con $f(x)$ polinomio di grado maggiore di 1, cioè il caso in cui la (4.1) sia della forma

$$a_mx^m + a_{m-1}x^{m-1} + \cdots + a_1x + a_0 = 0, \quad (a_m \neq 0) \tag{4.2}$$

con m intero ≥ 2.

La (4.2), che prende il nome di *equazione algebrica di grado* m, com'è noto possiede m radici nel campo complesso, ma queste, salvo casi speciali, si possono trovare con un metodo diretto soltanto per $m \leq 4$. In generale per calcolare numericamente una radice α di una equazione non lineare della forma (4.1) si ricorre ad un metodo iterativo, cioè all'applicazione ripetuta di una formula del tipo

$$x_{n+1} = \phi_n(x_n, x_{n-1}, \ldots, x_{n-k+1}), \quad n = 0, 1, \ldots, \quad k \geq 1, \tag{4.3}$$

dove ϕ_n si dice la *funzione di iterazione* del metodo. Tale funzione dipende non solo dagli argomenti $x_n, x_{n-1}, \ldots, x_{n-k+1}$, ma anche dalla funzione $f(x)$ e la sua forma può variare al variare di n. Una opportuna scelta della funzione di iterazione e delle *approssimazioni iniziali* $x_0, x_{-1}, \ldots, x_{-k+1}$ può far sì che la successione $\{x_n\}$ converga alla radice α. Il calcolo viene arrestato al verificarsi di qualche criterio di accettabilità prestabilito. Ad eccezione dei k valori iniziali, ogni altro termine di $\{x_n\}$ viene calcolato in funzione di k termini già noti: per questo motivo la (4.3) viene detta *metodo iterativo a k punti*. Se la forma di ϕ_n non varia al variare di n il metodo si dice *stazionario*.

Definizione 4.1.1 *Data una successione* $\{x_n\}$ *convergente ad un limite* α, *si ponga* $e_n = x_n - \alpha$; *se esistono due numeri reali* $p \geq 1$ *e* $C \neq 0$ *tali che sia*

$$\lim_{n\to\infty} \frac{\mid e_{n+1} \mid}{\mid e_n \mid^p} = C, \tag{4.4}$$

si dice che la successione ha ordine di convergenza p *e* fattore di convergenza C.

Per $p = 1$ e $p = 2$ la convergenza si dice anche *lineare* e *quadratica* rispettivamente. Nel caso di $p = 1$ la convergenza ad α implica $C < 1$.

Dalla (4.4) segue che per n "abbastanza grande" si ha

$$\frac{\mid e_{n+1} \mid}{\mid e_n \mid^p} \simeq C\,. \tag{4.5}$$

Si dice che il metodo (4.3) è convergente di ordine p se tale è la successione da esso generata. Analogamente a quanto stabilito in 3.6 per i sistemi lineari, si può definire una velocità asintotica di convergenza, nel caso $p = 1$ (e $C < 1$), ponendo $V = -\mathrm{Log}C$; per $p > 1$ la velocità di convergenza ha una espressione più complicata che dipende dai valori di p e di C e cresce al crescere di p e al decrescere di C.

Nei paragrafi che seguono verranno esposti alcuni metodi iterativi a due e ad un punto.

4.2 Metodo di bisezione

Il *metodo di bisezione* è il più semplice metodo iterativo per approssimare gli zeri reali di una funzione $f(x)$. In questo metodo ad ogni passo si

costruisce un intervallo contenente uno zero di $f(x)$ e si assume come approssimazione di tale zero l'ascissa del punto medio del detto intervallo.

Sia $f(x)$ continua sull'intervallo $[a, b]$ e poniamo $x_0 = a$, $x_1 = b$. Supposto che si abbia $f(x_0)f(x_1) < 0$, per la continuità di $f(x)$ si avrà almeno uno zero in $]x_0, x_1[$. Per semplicità supporremo che $]x_0, x_1[$ contenga un solo zero di $f(x)$. Il numero

$$x_2 = \frac{x_1 + x_0}{2},$$

cioè l'ascissa del punto medio di $]x_0, x_1[$, sarà certamente una approssimazione di α migliore di almeno una delle precedenti x_0 e x_1. Se non si verifica $f(x_2) = 0$, si confronta il segno di $f(x_2)$ con quello di $f(x_1)$; se risulta $f(x_2)f(x_1) < 0$ allora $\alpha \in]x_2, x_1[$, nel caso contrario sarà $\alpha \in]x_0, x_2[$. Quindi la nuova approssimazione x_3 sarà data da

a) $x_3 = (x_2 + x_1)/2$, se $f(x_2)f(x_1) < 0$,

b) $x_3 = (x_2 + x_0)/2$, se $f(x_2)f(x_1) > 0$.

Indicando con $\hat{x}_2$ una variabile che può assumere i valori x_1 o x_0, possiamo unificare i due casi nella sola formula:

$$x_3 = \frac{x_2 + \hat{x}_2}{2} \quad \text{dove } \hat{x}_2 = \begin{cases} x_1 & \text{se } f(x_2)f(x_1) < 0\,, \\ x_0 & \text{altrimenti}\,. \end{cases}$$

Per esempio, nel caso della Fig. 4.1 si verifica il caso b).

Ripetendo il procedimento, si determinano $x_4, x_5, x_6, \ldots$ secondo la formula generale

$$x_{n+1} = \frac{x_n + \hat{x}_n}{2}, \quad n = 1, 2, \ldots; \tag{4.6}$$

dove per $n = 1$ è $\hat{x}_1 = x_0$ mentre per $n > 1$ si pone

$$\hat{x}_n = \begin{cases} x_{n-1} & \text{se } f(x_n)f(x_{n-1}) < 0\,, \\ \hat{x}_{n-1} & \text{altrimenti}\,. \end{cases}$$

Il metodo di bisezione è quindi un metodo iterativo a due punti non stazionario, infatti la funzione al secondo membro della (4.6) non è la stessa ad ogni passo.

Poiché ad ogni passo l'intervallo contenente α viene dimezzato, dopo n passi si ha una approssimazione x_{n+1}, tale che

$$| x_{n+1} - \alpha | \leq \frac{1}{2^n}(b - a), \tag{4.7}$$

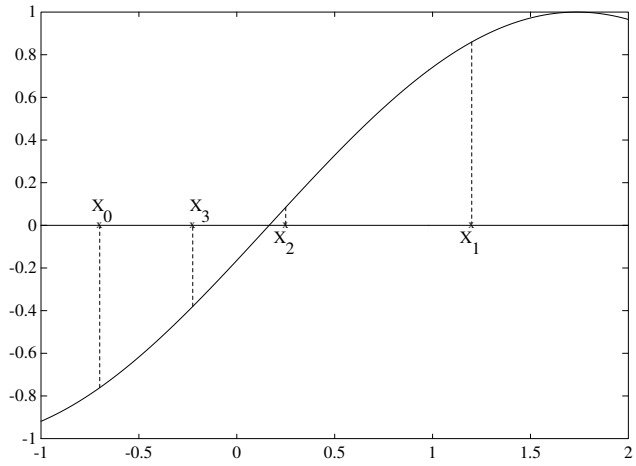

Figura 4.1: Metodo di bisezione.

da cui segue $\lim_{n\to\infty} \mid x_n - \alpha \mid = 0$, che prova la convergenza del metodo (4.6) alla radice α e fornisce anche una maggiorazione a priori dell'errore assoluto presente nell'iterata x_{n+1}. La (4.7) suggerisce come un possibile criterio di arresto la condizione $\frac{1}{2^n}(b-a) < \epsilon$, dove $\epsilon > 0$ è un numero opportunamente prefissato. La stessa condizione permette di conoscere a priori il numero n di iterazioni necessario per ridurre il modulo dell'errore assoluto al disotto di ϵ.

Il metodo (4.6) converge linearmente, infatti assumendo $\mid x_n - x_{n-1} \mid$ come stima di $\mid x_n - \alpha \mid = \mid e_n \mid$, si ha, per n abbastanza grande, l'uguaglianza approssimata della forma (4.5)

$$\frac{\mid e_{n+1} \mid}{\mid e_n \mid} \simeq \frac{\mid x_{n+1} - x_n \mid}{\mid x_n - x_{n-1} \mid} = \frac{1}{2} ;$$

poiché la convergenza è lenta, di solito questo metodo è usato per ottenere una prima approssimazione che consenta l'uso di altri metodi più efficienti.

4.3 Regula Falsi e metodo delle Secanti

Si abbiano le stesse condizioni iniziali del metodo di bisezione, ossia risulti $x_0 < \alpha < x_1$ e $f(x_0)f(x_1) < 0$. Il *metodo di falsa posizione*, noto anche col nome di *regula falsi*, è un altro metodo iterativo a due punti, in generale non

stazionario, nel quale, partendo da $]x_0, x_1[$, ad ogni passo si costruisce un nuovo intervallo di estremi x_n e $\hat{x}_n$ contenente uno zero di $f(x)$ e si assume come approssimazione x_{n+1} lo zero di una funzione lineare il cui grafico è la retta per i punti $A_n \equiv [x_n, f(x_n)]$ e $\hat{A}_n \equiv [\hat{x}_n, f(\hat{x}_n)]$. Si ha quindi lo schema

$$x_{n+1} = x_n - f(x_n)\frac{x_n - \hat{x}_n}{f(x_n) - f(\hat{x}_n)},\ n = 1, 2, \ldots, \tag{4.8}$$

dove per $n = 1$ è $\hat{x}_1 = x_0$ mentre per $n > 1$ si pone, come nel metodo di bisezione,

$$\hat{x}_n = \begin{cases} x_{n-1} & \text{se } f(x_n)f(x_{n-1}) < 0\,, \\ \hat{x}_{n-1} & \text{altrimenti}\,. \end{cases}$$

Si noti che anche nel metodo di bisezione il numero x_{n+1} fornito dalla (4.6) può interpretarsi come lo zero di una funzione lineare il cui grafico è la retta passante per i punti $A_n \equiv [x_n, sign_0(f(x_n))]$ e $\hat{A}_n \equiv [\hat{x}_n, sign_0(f(\hat{x}_n))]$, dove si è posto

$$sign_0(z) := \begin{cases} \frac{z}{|z|} & \text{per } z \neq 0\,, \\ 0 & \text{per } z = 0\,. \end{cases}$$

Il fatto che nella regula falsi si utilizzino i valori di $f(x)$ anziché i soli segni, spiega perché, quando il metodo (4.8) converge, la convergenza è più rapida che nel metodo (4.6). L'interpretazione geometrica del metodo è illustrata nella Fig. 4.2.

Teorema 4.3.1 *Se $f(x) \in C^2(\mathcal{I})$ e se il metodo regula falsi converge ad uno zero α di $f(x)$ con $\alpha \in \mathcal{I}$ tale che $f'(\alpha) \neq 0$, $f''(\alpha) \neq 0$, la convergenza è di ordine $p = 1$.*

Una variante importante della regula falsi è il *metodo delle secanti*, in cui sono richieste due approssimazioni iniziali senza alcun'altra condizione e senza la necessità di controllare il segno di $f(x)$ ad ogni passo.

In questo metodo il calcolo dell'approssimazione x_{n+1} utilizza le informazioni precedenti, x_n e x_{n-1}, secondo la formula

$$x_{n+1} = x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})},\ n = 1, 2 \ldots; \tag{4.9}$$

il numero x_{n+1} è individuato geometricamente (cfr. Fig. 4.3) dal punto in cui la secante passante per i punti $A_n \equiv [x_n, f(x_n)]$ e $A_{n-1} \equiv [x_{n-1}, f(x_{n-1})]$ incontra l'asse delle ascisse.

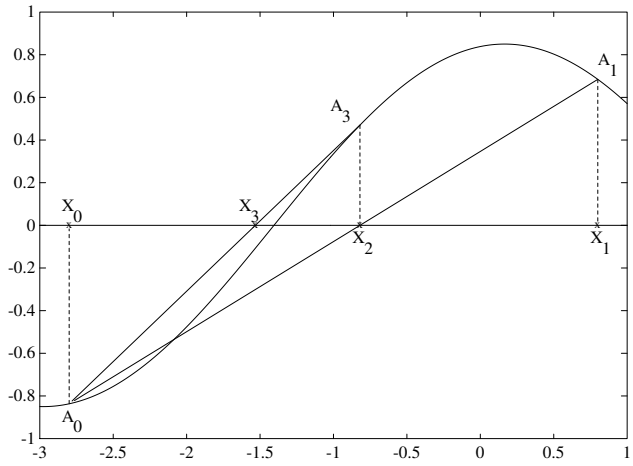

Figura 4.2: Regula falsi.

La (4.9) rappresenta un metodo iterativo stazionario a due punti, ma, per $n \geq 2$, ad ogni passo il calcolo di x_{n+1} richiede la sola valutazione di $f(x_n)$.

La convergenza del metodo è garantita se le approssimazioni x_0 e x_1 si scelgono "abbastanza vicine" alla radice α; vale il seguente teorema.

Teorema 4.3.2 *Se $f(x) \in C^2(\mathcal{I})$ e se il metodo delle secanti converge ad uno zero α di $f(x)$ con $\alpha \in \mathcal{I}$ e tale che $f'(\alpha) \neq 0$, $f''(\alpha) \neq 0$, allora l'ordine della convergenza è $p = (1 + \sqrt{5})/2 \simeq 1.618$.*

4.4 Metodi iterativi a un punto

In questo paragrafo si esporranno alcune proprietà generali dei metodi iterativi stazionari a un punto.

Data l'equazione $f(x) = 0$, si può sempre costruire una funzione $\phi(x)$ tale che l'equazione data sia equivalente alla seguente

$$x = \phi(x). \tag{4.10}$$

Basta infatti porre, ad esempio,

$$\phi(x) = x - g(x)f(x)\,,$$

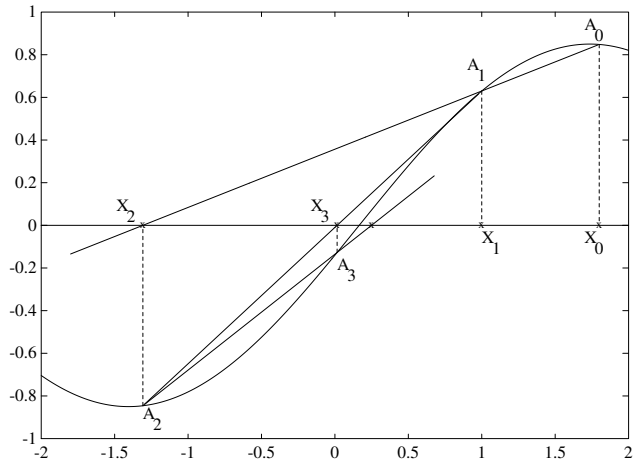

Figura 4.3: Metodo delle secanti.

dove $g(x)$ è un'arbitraria funzione continua che si suppone diversa da zero nei punti di una regione contenente gli zeri di $f(x)$. Se α è uno zero di $f(x)$ si ha anche $\alpha = \phi(\alpha)$ e viceversa; α si dice un *punto fisso* della trasformazione (4.10) o, semplicemente, di $\phi(x)$.

Per ogni scelta della funzione $\phi(x)$ si può considerare un metodo iterativo stazionario ad un punto della forma

$$x_{n+1} = \phi(x_n), \quad n = 0, 1, 2, \ldots, \tag{4.11}$$

e il problema di approssimare uno zero α di $f(x)$ si riduce a quello di costruire mediante la (4.11) una successione convergente ad α, punto fisso di $\phi(x)$.

Dal teorema seguente risulta una condizione sufficiente per la *convergenza locale* del metodo (4.11), cioè una convergenza assicurata dalla scelta di x_0 in un opportuno intorno di α.

Teorema 4.4.1 *Sia α un punto fisso di $\phi(x)$ interno ad un intervallo $\mathcal{I}$ sul quale $\phi(x)$ sia derivabile con continuità e si supponga che esistano due numeri positivi ρ e K con $K < 1$, tali che $\forall x \in [\alpha - \rho, \alpha + \rho] \subset \mathcal{I}$ si verifichi la condizione*

$$| \phi'(x) | \leq K; \tag{4.12}$$

allora per il metodo (4.11) valgono le seguenti proposizioni:

1. *se* $x_0 \in]\alpha-\rho, \alpha+\rho[$ *allora è anche* $x_n \in]\alpha-\rho, \alpha+\rho[$ *per* $n = 1, 2, \ldots$;
2. *per la successione* $\{x_n\}$, *con* $x_0 \in]\alpha-\rho, \alpha+\rho[$, *si ha* $\lim_{n\to\infty} x_n = \alpha$;
3. α *è l'unico punto fisso di* $\phi(x)$ *in* $[\alpha-\rho, \alpha+\rho]$.

DIMOSTRAZIONE. La proposizione 1 si dimostra per induzione, cioè, scelto un $x_0 \in]\alpha-\rho, \alpha+\rho[$, si ammette per ipotesi che sia, per un certo n, $x_n \in]\alpha-\rho, \alpha+\rho[$, ossia $\mid x_n - \alpha \mid < \rho$ e si deduce che deve essere $x_{n+1} \in]\alpha-\rho, \alpha+\rho[$, ovvero $\mid x_{n+1} - \alpha \mid < \rho$. Infatti, facendo uso della (4.11) e del teorema del valor medio, si ha

$$x_{n+1} - \alpha = \phi(x_n) - \phi(\alpha) = \phi'(\xi)(x_n - \alpha),$$

dove ξ è compreso tra x_n e α; dall'ipotesi fatta su x_n e da quelle del teorema segue poi

$$\mid x_{n+1} - \alpha \mid = \mid \phi'(\xi) \mid\mid x_n - \alpha \mid \leq K \mid x_n - \alpha \mid < \rho. \tag{4.13}$$

La proposizione 2 segue dall'ipotesi $0 < K < 1$ e dalla disuguaglianza

$$\mid x_{n+1} - \alpha \mid \leq K^{n+1} \mid x_0 - \alpha \mid$$

che si ottiene dalla (4.13).

Infine la proposizione 3 risulta per assurdo, infatti se in $]\alpha-\rho, \alpha+\rho[$ esistesse un altro punto fisso $\alpha' \neq \alpha$, si avrebbe

$$\mid \alpha - \alpha' \mid = \mid \phi(\alpha) - \phi(\alpha') \mid = \mid \phi'(\xi) \mid\mid \alpha - \alpha' \mid \leq K \mid \alpha - \alpha' \mid < \mid \alpha - \alpha' \mid .$$

□

Il Teorema 4.4.1 trova applicazione in molti casi concreti in cui qualche zero di $f(x)$ sia stato già localizzato (per esempio col metodo di bisezione) in un intervallo su cui valga la (4.12). Nelle Figure 4.4 e 4.5 sono riportati i grafici delle funzioni $y = x$ e $y = \phi(x)$ che si incontrano nel punto P di ascissa α. In esse è data l'interpretazione geometrica di un metodo della forma (4.11) con la condizione (4.12) verificata in due modi diversi.

In generale l'ordine di convergenza di un metodo iterativo è un numero reale ≥ 1 (vedi per esempio il metodo delle secanti). Per i metodi iterativi stazionari ad un punto vale però il teorema seguente.

Teorema 4.4.2 *Un metodo iterativo ad un punto, la cui funzione di iterazione* $\phi(x)$ *sia sufficientemente derivabile, ha ordine di convergenza uguale ad un numero intero positivo* p. *Precisamente se il metodo (4.11) converge ad* α, *la convergenza è di ordine* p *allora e solo che si abbia*

$$\phi(\alpha) = \alpha,\ \phi^{(i)}(\alpha) = 0 \ \ per\ \ 1 \leq i < p,\ \phi^{(p)}(\alpha) \neq 0. \tag{4.14}$$

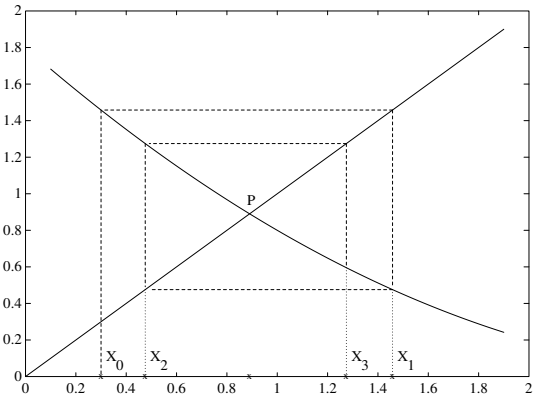

Figura 4.4: $-1 < \phi'(x) < 0$.

DIMOSTRAZIONE. Usando la (4.11) e la formula di Taylor si ha in generale

$$\begin{aligned} x_{n+1} - \alpha &= \phi(x_n) - \phi(\alpha) = \phi'(\alpha)(x_n - \alpha) + \phi''(\alpha)\frac{(x_n-\alpha)^2}{2!} + \cdots \\ &+ \ldots + \phi^{(p-1)}(\alpha)\frac{(x_n-\alpha)^{p-1}}{(p-1)!} + \phi^{(p)}(\xi)\frac{(x_n-\alpha)^p}{p!} \end{aligned}$$

dove ξ è compreso tra x_n e α; quindi, se valgono le (4.14), si ha

$$\lim_{n\to\infty} \frac{\mid x_{n+1} - \alpha \mid}{\mid x_n - \alpha \mid^p} = \frac{\mid \phi^{(p)}(\alpha) \mid}{p!} \neq 0, \tag{4.15}$$

cioè l'ordine di convergenza è p ed il fattore di convergenza è $C = \frac{|\phi^{(p)}(\alpha)|}{p!}$.

Viceversa se l'ordine è p, sia $\phi^{(i)}(\alpha)$ la prima derivata non nulla nel precedente sviluppo di Taylor intorno al punto α; se fosse $i \neq p$, per il ragionamento diretto anche l'ordine sarebbe $i \neq p$ contro l'ipotesi. Quindi deve essere $i = p$ cioè valgono le (4.14). □

È importante notare che i Teoremi 4.4.1 , 4.4.2 e quelli di 4.3 valgono per la successione teorica $\{x_n\}$ definita dal processo iterativo in assenza di errori di arrotondamento. Nella pratica però si introduce al passo n-esimo un errore δ_n dovuto agli arrotondamenti nel calcolo di ϕ e agli errori dei passi precedenti, per cui, nel caso specifico del processo (4.11), la successione che si ottiene è in realtà

$$\tilde{x}_{n+1} = \phi(\tilde{x}_n) + \delta_n, \quad n = 0, 1, \ldots; \tag{4.16}$$

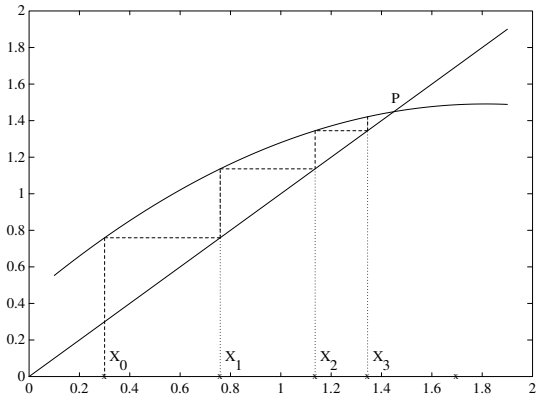

Figura 4.5: $0 < \phi'(x) < 1$.

per tale successione $\{\tilde{x}_n\}$, anche quando fossero verificate le ipotesi del Teorema 4.4.1, non è più garantita la convergenza ad α.

Tuttavia si può dimostrare che nelle ipotesi del teorema di convergenza e verificandosi la condizione $\mid \delta_n \mid \leq \delta$, $n = 0, 1, \ldots$, i termini della successione (4.16) finiscono col cadere in un intorno di α la cui ampiezza dipende linearmente da δ e inoltre, se K non è troppo vicino a 1 e se si opera con una precisione tale da rendere trascurabile il numero $\delta/(1-K)$, l'errore $\mid \tilde{x}_{n+1} - \alpha \mid$ è dell'ordine di $\mid \tilde{x}_{n+1} - \tilde{x}_n \mid$. Quindi come criterio di arresto dell'algoritmo iterativo si può assumere la condizione $\mid \tilde{x}_{n+1} - \tilde{x}_n \mid \leq \epsilon$ oppure $\left| \frac{\tilde{x}_{n+1} - \tilde{x}_n}{\min(|\tilde{x}_n|, |\tilde{x}_{n+1}|)} \right| \leq \epsilon$ a seconda che si voglia limitare l'errore assoluto o quello relativo.

Si abbia una successione $\{x_n\}$ convergente linearmente; in base alla (4.5) si ha quindi, per $p = 1$ e per n "sufficientemente" grande,

$$x_{n+1} - \alpha \simeq C(x_n - \alpha) \, ,$$

$$x_{n+2} - \alpha \simeq C(x_{n+1} - \alpha) \, ,$$

da cui

$$\alpha \simeq \frac{x_n x_{n+2} - x_{n+1}^2}{x_{n+2} - 2x_{n+1} + x_n}. \tag{4.17}$$

Si ha perciò una approssimazione di α costruita con tre termini successivi x_n, x_{n+1}, x_{n+2}. Il secondo membro della (4.17) si può considerare come il

termine z_n di una nuova successione, che, per evitare errori di cancellazione, si pone, di solito, nella forma

$$z_n = x_n - \frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n}. \tag{4.18}$$

Introducendo i simboli

$$\Delta_n := x_{n+1} - x_n, \quad \Delta_n^2 := x_{n+2} - 2x_{n+1} + x_n (\Delta_n^2 = \Delta_{n+1} - \Delta_n)$$

la successione $\{z_n\}$ può scriversi

$$z_n = x_n - \frac{(\Delta_n)^2}{\Delta_n^2},\ n = 0, 1, \ldots. \tag{4.19}$$

Lo schema di calcolo definito dalla (4.19) va sotto il nome di *processo* Δ^2 *di Aitken* ed ha lo scopo di accelerare la convergenza di una successione che converga linearmente. Vale infatti il seguente teorema.

Teorema 4.4.3 *Se una successione* $\{x_n\}$ *converge ad* α *linearmente, allora la successione* $\{z_n\}$ *converge allo stesso limite* α *più rapidamente di* $\{x_n\}$*, ossia si ha*

$$\lim_{n\to\infty} \frac{z_n - \alpha}{x_n - \alpha} = 0.$$

Nella Fig. 4.6 è data una interpretazione geometrica facilmente verificabile del processo di Aitken, nel caso di una successione generata con iterazioni del tipo $x_{n+1} = \phi(x_n)$: z_n è l'ascissa della intersezione della retta per i punti $P_n \equiv [x_n, \phi(x_n)]$ e $P_{n+1} \equiv [x_{n+1}, \phi(x_{n+1})]$ con la retta di equazione $y = x$.

Sulla base della (4.17), supponendola valida anche nel caso di una successione che converga non linearmente, è possibile, partendo da un metodo iterativo $x_{n+1} = \phi(x_n)$ di un dato ordine, costruirne un altro di ordine più elevato. Infatti, esprimendo il membro destro della (4.18) in funzione della sola x_n, si ha

$$F(x_n) = x_n - \frac{(\phi(x_n) - x_n)^2}{\phi(\phi(x_n)) - 2\phi(x_n) + x_n}.$$

In generale $F(x)$ ha gli stessi punti fissi di $\phi(x)$.

Il metodo iterativo

$$x_{n+1} = F(x_n), \quad n = 0, 1, \ldots, \tag{4.20}$$

è noto come *metodo di Steffensen* e per esso si può dimostrare il teorema seguente.

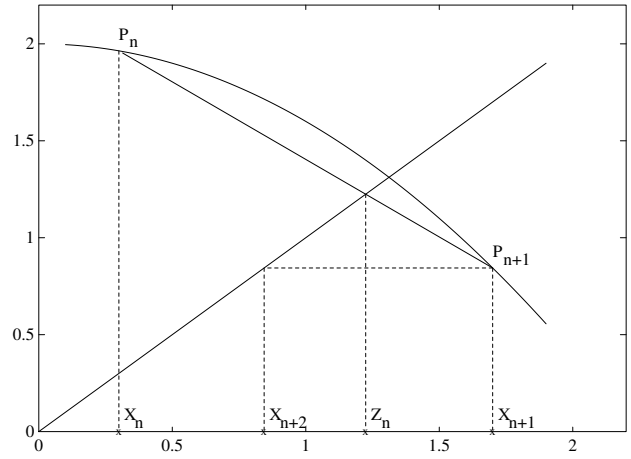

Figura 4.6: Processo di Aitken.

Teorema 4.4.4 *Sia $\phi(x)$ la funzione di iterazione di un metodo di ordine p per approssimare un suo punto fisso α. Per $p > 1$ il corrispondente metodo di Steffensen (4.20) per approssimare α ha ordine $2p - 1$, mentre per $p = 1$ e nella ipotesi $\phi'(\alpha) \neq 1$, il metodo (4.20) ha ordine almeno 2.*

4.5 Metodo di Newton

Il più importante fra i metodi ad un punto è il *metodo di Newton*. Tale metodo si può applicare per approssimare uno zero α di $f(x)$ se, in tutto un intorno di α, $f(x)$ è derivabile con continuitá. In tal caso, assumendo la funzione di iterazione della forma

$$\phi(x) = x - \frac{f(x)}{f'(x)} \tag{4.21}$$

si ha il metodo

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, 2, \ldots. \tag{4.22}$$

L'iterata x_{n+1} è individuata dal punto d'incontro dell'asse delle ascisse con la tangente alla curva $y = f(x)$ nel punto $A_n \equiv [x_n, f(x_n)]$ (cfr. Fig. 4.7); per questo si usa anche la denominazione di *metodo delle tangenti*.

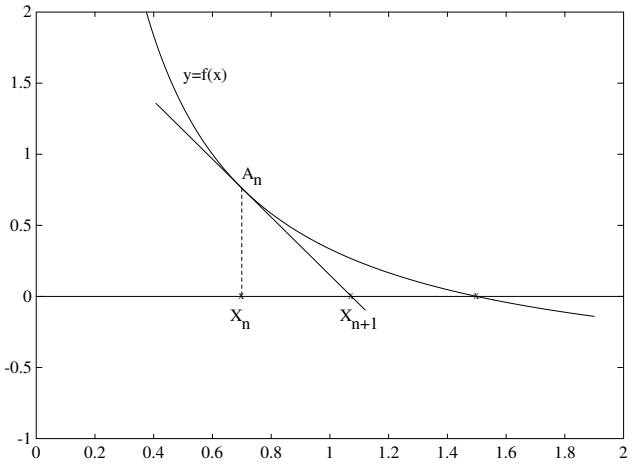

Figura 4.7: Metodo di Newton.

Sulla convergenza e l'ordine del metodo di Newton, vale il seguente teorema.

Teorema 4.5.1 *Sia* $f(x) \in C^3([a,b])$, $a < \alpha < b$, $f(\alpha) = 0$, $f'(\alpha) \neq 0$, *allora valgono le proposizioni:*

1. *esiste un numero* $\rho > 0$ *tale che per ogni* $x_0 \in [\alpha - \rho, \alpha + \rho]$ *il metodo (4.22) converge;*
2. *la convergenza è di ordine* $p \geq 2$*;*
3. *se* $p = 2$ *il fattore di convergenza è* $C = \frac{1}{2}\left|\frac{f''(\alpha)}{f'(\alpha)}\right|$.

Dimostrazione. Dalla (4.21) segue

$$\phi'(x) = \frac{f(x)f''(x)}{f'^2(x)} \tag{4.23}$$

da cui

$$\phi'(\alpha) = 0\,; \tag{4.24}$$

quindi, fissato un numero positivo $K < 1$, esiste un numero $\rho > 0$ tale che per ogni $x \in [\alpha-\rho, \alpha+\rho]$ si abbia $\mid \phi'(x) \mid \leq K$ e perciò vale il teorema di convergenza 4.4.1.

Per dimostrare l'asserto 2, dalla (4.23) si ricava

$$\phi''(\alpha) = \frac{f''(\alpha)}{f'(\alpha)};$$

ne segue che se $f''(\alpha) \neq 0$ è anche $\phi''(\alpha) \neq 0$; questa insieme alla (4.24), garantisce per il Teorema 4.4.2, che l'ordine è $p = 2$. Se invece si ha $f''(\alpha) = 0$ è anche $\phi''(\alpha) = 0$ e quindi si ha $p > 2$.

Se $p = 2$, l'asserto 3 segue dalla (4.15) essendo $\phi'(\alpha) = 0$ e $\phi''(\alpha) \neq 0$. □

Notiamo che la convergenza di cui si parla nella proposizione 1 del teorema precedente è di tipo locale, cioè si verifica se si sceglie x_0 in un intorno di α di raggio ρ abbastanza piccolo. A un tale intorno si può pervenire, ad esempio, col metodo di bisezione.

Vi sono però casi speciali in cui si può avere una *convergenza globale*, cioè che si verifica per qualunque scelta di x_0 in un dato intervallo limitato $[a, b]$. Ciò è asserito nel teorema seguente.

Teorema 4.5.2 *Sia $f(x) \in C^2([a, b])$ con $f(a)f(b) < 0$ e si supponga*

1. *$f'(x) \neq 0$, $\forall x \in [a, b]$,*
2. *$f''(x) \geq 0$, oppure $f''(x) \leq 0$, $\forall x \in [a, b]$,*
3. *$\left|\frac{f(a)}{f'(a)}\right| < b - a$, $\left|\frac{f(b)}{f'(b)}\right| < b - a$,*

allora $f(x)$ ha un solo zero $\alpha \in [a, b]$ ed il metodo di Newton converge ad α per ogni $x_0 \in [a, b]$.

Osservazione 4.5.1 Si noti che ponendo nella (4.22) $x_n = a$ oppure $x_n = b$ si ottiene rispettivamente

$$| x_{n+1} - a | = \left|\frac{f(a)}{f'(a)}\right| \quad \text{e} \quad | x_{n+1} - b | = \left|\frac{f(b)}{f'(b)}\right|,$$

per cui le condizioni dell'ipotesi 3 equivalgono a

$$| x_{n+1} - a | < b - a, \quad | x_{n+1} - b | < b - a,$$

cioè le tangenti negli estremi $A \equiv [a, f(a)]$ e $B \equiv [b, f(b)]$ intersecano l'asse delle ascisse all'interno di $[a, b]$.

Sotto le ipotesi 1 e 2 del Teorema 4.5.2 cadono in particolare le funzioni monotone convesse o concave su $[a, b]$. Una importante applicazione si ha nella speciale equazione

$$f(x) = x^m - c = 0, \quad m \in \mathbb{R},\ m \geq 2,\ c > 0, \tag{4.25}$$

che possiede la soluzione positiva $\alpha = \sqrt[m]{c}$, radice m-esima aritmetica di c.

Poiché per $x > 0$ si ha $f'(x) > 0$ e $f''(x) > 0$, per applicare il Teorema 4.5.2 basta constatare che si possono assegnare intervalli $[a, b]$ con $a < \alpha < b$, tali da soddisfare le condizioni 3. Se $a > 0$ è tale che sia $f(a) < 0$, si avrà $a < \alpha$. Ogni numero $b > a + \left|\frac{f(a)}{f'(a)}\right| = a + \frac{a^m - c}{ma^{m-1}}$ verifica la prima delle condizioni 3; inoltre, con questa scelta, b risulta a destra dell'intersezione della tangente in $A \equiv [a, f(a)]$ con l'asse delle ascisse, quindi è $b > \alpha$ e si ha

$$\begin{aligned} \left|\frac{f(b)}{f'(b)}\right| &= \frac{b^m - \alpha^m}{mb^{m-1}} \\ &= (b - \alpha)\frac{b^{m-1} + \alpha b^{m-2} + \cdots + \alpha^{m-1}}{mb^{m-1}} \\ &< b - \alpha < b - a. \end{aligned}$$

Pertanto un tale numero b verifica anche la seconda condizione 3. Ne segue che ogni numero positivo può considerarsi interno a un intervallo $[a, b]$ che verifichi le condizioni 3 e perciò il metodo di Newton applicato alla (4.25) converge ad α, comunque si scelga una approssimazione iniziale $x_0 > 0$. Il metodo fornisce la formula

$$x_{n+1} = \frac{1}{m}[(m-1)x_n + cx_n^{1-m}] \quad n = 0, 1, \ldots.$$

Per $m = 2$ si ha l'algoritmo $x_{n+1} = \frac{1}{2}\left(x_n + \frac{c}{x_n}\right)$ per estrarre la radice quadrata di c, notevole per il suo basso costo computazionale.

Il Teorema 4.5.1 vale nell'ipotesi $f'(\alpha) \neq 0$, cioè se α è uno zero semplice di $f(x)$. Supponiamo ora che α sia per $f(x)$ uno zero di molteplicità $s \geq 1$; in tal caso si può scrivere

$$f(x) = g(x)(x - \alpha)^s, \text{ con } g(x) = \frac{f(x)}{(x - \alpha)^s} \text{ e } g(\alpha) = \lim_{x \to \alpha} g(x) \neq 0.$$

Quindi per il metodo di Newton si ha

$$\phi(x) = x - \frac{f(x)}{f'(x)} = x - \frac{g(x)(x-\alpha)}{sg(x) + g'(x)(x-\alpha)}$$

da cui segue, con facili calcoli,

$$\phi'(\alpha) = 1 - \frac{1}{s}\,. \tag{4.26}$$

Perciò, per $s > 1$, il metodo convergere linearmente con fattore di convergenza $C = 1 - \frac{1}{s}$.

Se si sostituisce la funzione $f(x)$ con $F(x) = f(x)/f'(x)$ che ha gli stessi zeri di $f(x)$ ma tutti semplici, il metodo di Newton, applicato all'equazione $F(x) = 0$ fornisce ancora una convergenza almeno quadratica verso la radice α, essendo ora $F'(\alpha) \neq 0$. Il costo computazionale è però maggiore di quello richiesto dall'equazione $f(x) = 0$, poiché, in effetti, si usa la formula iterativa

$$x_{n+1} = x_n - \frac{f(x_n)f'(x_n)}{f'^2(x_n) - f(x_n)f''(x_n)}, \quad n = 0, 1, \ldots. \tag{4.27}$$

Quando si conosce la molteplicità della radice α si può usare una formula meno costosa della (4.27), ottenuta modificando il metodo di Newton in modo che la convergenza sia almeno quadratica anche per una radice di molteplicità $s > 1$. Scrivendo infatti

$$x_{n+1} = x_n - s\frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, \ldots, \tag{4.28}$$

si definisce un metodo per il quale si ha

$$\phi(x) = x - s\frac{f(x)}{f'(x)}\,,$$

da cui segue, tenuto conto della (4.26),

$$\phi'(\alpha) = 0\,.$$

Nell'uso del metodo (4.22) può accadere che i valori di $f'(x_n)$ varino molto lentamente al variare di n. In tal caso può essere utile una variante del metodo di Newton in cui lo stesso valore di $f'(x_n)$ viene utilizzato per un certo numero di passi successivi, dopo i quali viene ricalcolato; in altri termini

$f'(x_n)$ non si calcola ad ogni passo ma solo ogni volta che n coincide con un elemento di un prefissato sottoinsieme dei numeri naturali, lasciando $f'(x_n)$ uguale all'ultimo valore calcolato se n non appartiene a tale sottoinsieme. Si noti che se il valore di $f'(x_n)$ viene mantenuto costante per ogni n, la convergenza del metodo dipende dal valore di tale costante ed è comunque lineare. Anche il metodo delle secanti può considerarsi ottenuto dal metodo di Newton approssimando $f'(x_n)$ con il rapporto incrementale $\frac{f(x_n)-f(x_{n-1})}{x_n-x_{n-1}}$.

4.6 Metodi iterativi in $\mathbb{R}^n$

La teoria dei metodi iterativi precedentemente esposta può essere estesa al caso in cui, nella (4.1), sia $f(x):\mathbb{R}^n \to \mathbb{R}^n$, cioè al caso di un sistema di n equazioni (non lineari) in altrettante incognite

$$\begin{aligned} f_1(x_1,\ldots,x_n) &= 0 \\ f_2(x_1,\ldots,x_n) &= 0 \\ &\vdots \\ f_n(x_1,\ldots,x_n) &= 0\,. \end{aligned} \tag{4.29}$$

In analogia con quanto esposto in 4.4, il sistema (4.29) si può scrivere in una forma equivalente la quale consente di approssimare una soluzione $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)^T$ come punto fisso di una opportuna funzione di iterazione $\phi(x):\mathbb{R}^n \to \mathbb{R}^n$ data da

$$\phi(x) = x - G(x)f(x), \tag{4.30}$$

dove $\phi = (\phi_1, \phi_2, \ldots, \phi_n)^T$, $x = (x_1, x_2, \ldots, x_n)^T$, $f = (f_1, f_2, \ldots, f_n)^T$ e $G(x)$ è una matrice $n \times n$ non singolare in un dominio $D \subseteq \mathbb{R}^n$ contenente α.

Si considerano quindi metodi iterativi della forma

$$x^{(k+1)} = \phi(x^{(k)}), \quad k = 0, 1, \ldots, \tag{4.31}$$

per i quali si definisce l'ordine di convergenza come in 4.1 cambiando nella Definizione 4.1.1 i valori assoluti in norme di vettori.

Se esistono continue le derivate prime delle funzioni ϕ_i, introducendo la matrice jacobiana

$$\Phi(x) = \begin{pmatrix} \frac{\partial \phi_1}{\partial x_1} & \cdots & \frac{\partial \phi_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \phi_n}{\partial x_1} & \cdots & \frac{\partial \phi_n}{\partial x_n} \end{pmatrix}, \tag{4.32}$$

si può generalizzare il Teorema 4.4.1 e dimostrare il seguente teorema di convergenza locale.

Teorema 4.6.1 *Se α è un punto fisso di $\phi(x)$, condizione sufficiente per la convergenza ad α del metodo (4.31) è che esistano due numeri positivi K e ρ, con $K < 1$, tali che si abbia*

$$\| \Phi(x) \| \leq K, \quad \forall x \in D_\rho = \{x \mid \| x - \alpha \| \leq \rho\} ; \tag{4.33}$$

purché $x^{(0)}$ sia scelto in D_ρ; in tal caso α è l'unico punto fisso di ϕ in D_ρ.

Il Teorema 4.4.2 non può essere direttamente esteso in $\mathbb{R}^n$; tuttavia se esistono continue le derivate seconde delle ϕ_i, si può dimostrare che, se il metodo (4.31) converge ad un punto fisso α, condizione sufficiente perché converga linearmente è che sia

$$\Phi(\alpha) \neq \mathbf{O} .$$

Mentre, se $\Phi(\alpha) = \mathbf{O}$, la convergenza è almeno quadratica ed è esattamente quadratica se risulta non nulla almeno una delle matrici hessiane

$$H_i(\alpha) = \begin{pmatrix} \frac{\partial^2 \phi_i}{\partial x_1^2} & \frac{\partial^2 \phi_i}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 \phi_i}{\partial x_1 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 \phi_i}{\partial x_n \partial x_1} & \frac{\partial^2 \phi_i}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 \phi_i}{\partial x_n^2} \end{pmatrix}_{x=\alpha}, \ i = 1, \ldots, n.$$

Per estendere il metodo di Newton al sistema non lineare (4.29), supponiamo che le funzioni f_i siano derivabili con continuità rispetto a ciascuna variabile e che la matrice jacobiana

$$J(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}$$

sia non singolare in un dominio contenente nel suo interno una soluzione α del sistema (4.29).

Specializzando la funzione di iterazione (4.30) nella forma

$$\phi(x) = x - J^{-1}(x) f(x),$$

si ha il sistema $x = x - J^{-1}(x)f(x)$ equivalente a (4.29), da cui il metodo iterativo di Newton o di *Newton-Raphson*

$$x^{(k+1)} = x^{(k)} - J^{-1}(x^{(k)})f(x^{(k)}), \quad k = 0, 1, \ldots .$$

Nell'uso pratico del metodo l'iterata $x^{(k+1)}$ si ricava dalla soluzione del sistema lineare

$$J(x^{(k)})d^{(k)} = -f(x^{(k)}), \quad k = 0, 1, \ldots, \tag{4.34}$$

dove $d^{(k)} = x^{(k+1)} - x^{(k)}$.

Sull'ordine e sulla convergenza locale del metodo si ha il seguente teorema.

Teorema 4.6.2 *Sia $\alpha \in D$ soluzione di (4.29) e le funzioni f_i siano della classe $C^3(D)$ e tali che $J(x)$ sia non singolare in D; allora l'ordine di convergenza del metodo di Newton è almeno $p = 2$ e si ha convergenza per ogni scelta di $x^{(0)}$ in un opportuno dominio contenente α.*

DIMOSTRAZIONE. Tenuto conto della forma della funzione di iterazione ϕ, la colonna j-esima della matrice (4.32) risulta:

$$\begin{aligned}
\frac{\partial \phi}{\partial x_j} &= \frac{\partial x}{\partial x_j} - \frac{\partial}{\partial x_j}\left[J^{-1}(x)f(x)\right] \\
&= \frac{\partial x}{\partial x_j} - J^{-1}(x)\frac{\partial f(x)}{\partial x_j} - \frac{\partial J^{-1}(x)}{\partial x_j}f(x) \\
&= -\frac{\partial J^{-1}(x)}{\partial x_j}f(x)\,;
\end{aligned}$$

infatti i vettori $\frac{\partial x}{\partial x_j}$ e $J^{-1}(x)\frac{\partial f(x)}{\partial x_j}$ coincidono con la j-esima colonna della matrice identica. Ne segue che per $x = \alpha$ si ha

$$\left(\frac{\partial \phi}{\partial x_j}\right)_{x=\alpha} = -\left(\frac{\partial J^{-1}(x)}{\partial x_j}\right)_{x=\alpha} f(\alpha) = 0, \quad j = 1, 2, \ldots, n,$$

quindi è

$$\Phi(\alpha) = \mathbf{O}, \tag{4.35}$$

cioè l'ordine di convergenza del metodo è almeno $p = 2$.

Dalla continuità di $\Phi(x)$ e dalla (4.35) segue infine, per un assegnato $K < 1$, l'esistenza di un dominio $D_\rho = \{x \mid \| x - \alpha \| \leq \rho\} \subset D$ in cui vale la condizione (4.33) e quindi il metodo converge $\forall x^{(0)} \in D_\rho$. □

Nel caso di una lenta variazione della matrice jacobiana $J(x)$, si può ricorrere anche ora ad un *metodo di Newton semplificato* della forma

$$J(x^{(0)})d^{(k)} = -f(x^{(k)}), \quad k = 0, 1, \ldots, \tag{4.36}$$

dove la matrice jacobiana è valutata una sola volta in corrispondenza ad una buona approssimazione iniziale $x^{(0)}$. Se il metodo (4.36) converge, la convergenza è in generale lineare.

Un modo di evitare il calcolo dell'intera matrice $J(x)$ ad ogni passo, consiste nel considerare l'i-esima equazione del sistema (4.29) come una equazione nella sola incognita x_i ed applicare a ciascuna equazione del sistema il metodo di Newton per equazioni in una incognita. Supposto che l'ordinamento delle equazioni sia tale che, in un dominio contenente α, si abbia

$$\frac{\partial f_i(x_1, \ldots, x_n)}{\partial x_i} \neq 0, \quad i = 1, 2, \ldots, n,$$

si ottiene

$$x_i^{(k+1)} = x_i^{(k)} - \frac{f_i(x_1^{(k)}, \ldots, x_n^{(k)})}{\frac{\partial f_i(x_1^{(k)}, \ldots, x_n^{(k)})}{\partial x_i}}, \ i = 1, \ldots, n; \ k = 0, 1, \ldots . \tag{4.37}$$

Questo metodo, detto *metodo non lineare di Jacobi-Newton*, è un metodo iterativo ad un punto e può esprimersi nella forma matriciale

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

dove $D = \text{diag}(\frac{\partial f_1}{\partial x_1}, \frac{\partial f_2}{\partial x_2}, \ldots, \frac{\partial f_n}{\partial x_n})$; la convergenza, lineare, è garantita se vale la condizione (4.33), tenendo conto che nel metodo in esame si ha $\phi(x) = x - D^{-1}(x)f(x)$.

Una variante del metodo (4.37) si ottiene introducendo la tecnica iterativa di Gauss-Seidel, della forma

$$x_i^{(k+1)} = x_i^{(k)} - \frac{f_i(x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}, x_i^{(k)}, \ldots, x_n^{(k)})}{\frac{\partial f_i(x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}, x_i^{(k)}, \ldots, x_n^{(k)})}{\partial x_i}}, \tag{4.38}$$

$$i = 1, 2, \ldots, n, \ k = 0, 1, \ldots,$$

dove si sono utilizzate, per il calcolo di $x_i^{(k+1)}$, le prime $i-1$ componenti già note dell'iterata $x^{(k+1)}$ in corso di calcolo.

Il metodo (4.38) è noto come *metodo non lineare di Gauss-Seidel*.

Un modo di accelerare la convergenza lineare di questo metodo è quello di introdurre un fattore di rilassamento $\omega \in]0,2[$ ponendo

$$x_i^{(k+1)} = x_i^{(k)} - \omega F_i^{(k+1,k)},\ i = 1,2,\ldots,n,\ k = 0,1,\ldots, \tag{4.39}$$

dove $F_i^{(k+1,k)}$ indica la frazione che compare in (4.38). Alla (4.39), con $\omega > 1$ si dà il nome di *metodo di sovrarilassamento di Newton*.

Una classe di metodi, mutuati dal metodo di Newton-Raphson e che possono essere interpretati come una estensione del metodo delle secanti in $\mathbb{R}^n$, si ottiene sostituendo in (4.34) la matrice $J(x)$ con una sua approssimazione discreta; si ha così l'evidente vantaggio di non dovere calcolare le n^2 derivate parziali costituenti $J(x)$.

Per approssimare $J(x^{(k)})$ si può introdurre la matrice $\Delta(x^{(k)})$, reale di ordine n, i cui elementi sono dati da

$$\left(\Delta(x^{(k)})\right)_{ij} = \frac{f_i(x^{(k)}+h_j^{(k)}e^{(j)}) - f_i(x^{(k)})}{h_j^{(k)}},$$
$$i,j = 1,2,\ldots,n,\ k = 0,1,\ldots,$$

essendo $h_j^{(k)}$ opportuni scalari, infinitesimi al crescere di k; ad esempio si può assumere $h_j^{(k)} = f_j(x^{(k)})$ oppure, più semplicemente, $h_j^{(k)} = x_j^{(k-1)} - x_j^{(k)}$.
In generale questi metodi presentano un ordine di convergenza $p \leq 2$.

4.7 Zeri di polinomi

L'equazione (4.1) sia algebrica di grado $m \geq 2$, cioè della forma

$$P(x) = a_m x^m + a_{m-1}x^{m-1} + \cdots + a_1 x + a_0 = 0 \ \ (a_m \neq 0) \tag{4.40}$$

con i coefficienti a_i reali. All'equazione (4.40) sono applicabili tutti i metodi visti nei paragrafi precedenti, ma le particolari proprietà della classe dei polinomi forniscono ulteriori procedimenti per la localizzazione e l'approssimazione delle radici. Uno strumento utile a tale scopo è la *successione di Sturm*.

Definizione 4.7.1 *Si dice* successione di Sturm *per il polinomio $P(x)$ con $a_i \in \mathbb{R}$, una successione di polinomi reali*

$$P_0(x) := P(x),\ P_1(x),\ldots,\ P_k(x),\ \ (k \leq m) \tag{4.41}$$

per cui valgono le seguenti proprietà:

1. $P_k(x)$ *non ha zeri reali,*

2. $\alpha \in \mathbb{R}$, $P_r(\alpha) = 0$ *per* $1 \le r \le k-1$ *implica* $P_{r-1}(\alpha)P_{r+1}(\alpha) < 0$,

3. $\alpha \in \mathbb{R}$, $P_0(\alpha) = 0$ *implica* $P_0'(\alpha)P_1(\alpha) > 0$.

Dalla proprietà 3 della definizione segue che gli zeri reali di $P(x)$ sono tutti semplici.

Teorema 4.7.1 (di Sturm) *Se la (4.41) è una successione di Sturm il numero degli zeri del polinomio $P(x)$ nell'intervallo $a < x \le b$ è dato da $V(a) - V(b)$, dove $V(x)$ è il numero delle variazioni di segno presenti nella successione dei valori non nulli assunti dalla (4.41) nel punto x.*

DIMOSTRAZIONE. Al variare di x la funzione $V(x)$ può cambiare di valore soltanto se qualche polinomio $P_i(x)$ della successione (4.41) cambia di segno e pertanto si annulla in un punto $x = \alpha$. Per la proprietà 2 non possono annullarsi in $x = \alpha$ due polinomi consecutivi, inoltre, se $P_i(\alpha) = 0$, non può essere $i = k$ per la proprietà 1. Quindi deve essere $0 \le i \le k-1$.

Distinguiamo i due casi $i > 0$ e $i = 0$.

Nel primo caso, tenuto conto della proprietà 2 e della continuità , i polinomi $P_{i-1}(x)$, $P_i(x)$ e $P_{i+1}(x)$, calcolati nel punto $x = \alpha$ e in punti sufficientemente vicini $x = \alpha - \delta$ e $x = \alpha + \delta$ ($\delta > 0$), presentano una sola delle possibili distribuzioni di segni illustrate nei due quadri seguenti:

	$\alpha-\delta$	α	$\alpha+\delta$
P_{i-1}	$+$	$+$	$+$
P_i	$\pm$	0	$\pm$
P_{i+1}	$-$	$-$	$-$

$\alpha-\delta$	α	$\alpha+\delta$
$-$	$-$	$-$
$\pm$	0	$\pm$
$+$	$+$	$+$

.

Quindi si ha comunque $V(\alpha-\delta) = V(\alpha) = V(\alpha+\delta)$ e il valore di $V(x)$ non cambia quando x attraversa uno zero di $P_i(x)$.

Nel secondo caso, tenendo conto della proprietà 3, si hanno le due possibilità:

	$\alpha-\delta$	α	$\alpha+\delta$
P_0	$+$	0	$-$
P_1	$-$	$-$	$-$

$\alpha-\delta$	α	$\alpha+\delta$
$-$	0	$+$
$+$	$+$	$+$

.

Perciò se α è radice di $P(x) = 0$ nella successione (4.41) si ha

$$V(\alpha - \delta) - V(\alpha + \delta) = 1\,.$$

In particolare se $\alpha = a$, dalle ultime due colonne di ciascuno dei quadri relativi a P_0 e P_1, si ha $V(a) - V(a+\delta) = 0$, mentre se $\alpha = b$, dalle prime due colonne si ha $V(b-\delta) - V(b) = 1$.

Quindi il valore di $V(x)$ si abbassa di una unità ogni volta che x attraversa crescendo uno zero di $P(x)$, eccettuato il caso in cui tale zero coincida con $x = a$.

Da quanto sopra resta dimostrata la tesi. □

Corollario 4.7.1 *Se nella (4.41) si ha $k = m$ e se inoltre tutti i polinomi hanno i coefficienti dei termini di grado massimo dello stesso segno, allora l'equazione (4.40) ha m radici reali e distinte e viceversa.*

Per costruire effettivamente una successione di Sturm per il polinomio (4.40), si pone

$$P_0(x) := P(x)\,, \quad P_1(x) := P'(x)\,,$$
$$P_{r-1}(x) = P_r(x)Q_r(x) - P_{r+1}(x)\,, \;\; r = 1, 2, \ldots, \tag{4.42}$$

dove $Q_r(x)$ e $-P_{r+1}(x)$ sono rispettivamente quoziente e resto della divisione $P_{r-1}(x)/P_r(x)$, $r = 1, 2, \ldots$.

Il processo (4.42) ha termine, poiché il grado dei polinomi decresce al crescere dell'indice e perciò per un certo $k \leq m$ risulta

$$P_{k-1}(x) = P_k(x)Q_k(x).$$

Si riconosce nel processo (4.42) il noto *algoritmo di Euclide* che fornisce il massimo comune divisore di $P_0(x)$ e $P_1(x)$, cioè si ha

$$P_k(x) = \text{M.C.D. } \{P(x), P'(x)\}. \tag{4.43}$$

Dalle (4.42) e (4.43) segue che, nel caso in cui $P(x)$ e $P'(x)$ non abbiano zeri reali in comune, i polinomi $P_0(x), P_1(x), \ldots, P_k(x)$ formano una successione di Sturm per $P(x)$. Infatti, nell'ipotesi fatta e per la (4.43), $P_k(x)$ non ha zeri reali (proprietà 1).

Se $P_r(\alpha) = 0$ con $1 \leq r \leq k-1$, allora si ha $P_{r-1}(\alpha) = -P_{r+1}(\alpha) \neq 0$ come segue dalla (4.42) e quindi $P_{r-1}(\alpha)P_{r+1}(\alpha) < 0$ (proprietà 2).
Dalla definizione di $P_0(x)$ e di $P_1(x)$, se $P_0(\alpha) = 0$, segue $P_0'(\alpha)P_1(\alpha) = (P'(\alpha))^2 > 0$ (proprietà 3).

Se $P(x)$ e $P'(x)$ hanno zeri reali in comune, questi, per la (4.43), sono tutti zeri di $P_k(x)$ con la stessa molteplicità che hanno come zeri di $P'(x)$ e la (4.42) non fornisce una successione di Sturm.

In tal caso si può verificare che la successione

$$\frac{P_0(x)}{P_k(x)}, \quad \frac{P_1(x)}{P_k(x)}, \ldots, \frac{P_k(x)}{P_k(x)}, \tag{4.44}$$

è una successione di Sturm per il polinomio $P(x)/P_k(x)$ che ha tanti zeri semplici quanti sono gli zeri distinti di $P(x)$.

La differenza $V(a) - V(b)$ valutata per la successione (4.44) fornisce allora il numero delle radici reali e distinte (indipendentemente dalla loro molteplicità) dell'equazione $P(x) = 0$ sull'intervallo $a < x \leq b$.

Una successione di Sturm può essere usata per individuare un intervallo $[a, b]$ contenente una sola radice reale α di una equazione algebrica e quindi, con successivi dimezzamenti dell'intervallo come nel metodo di bisezione, si può approssimare α con qualsivoglia accuratezza anche se piuttosto lentamente.

L'approssimazione di uno zero di $P(x)$, che sia stato separato per esempio con una successione di Sturm, può essere fatta usando uno qualunque dei metodi esposti precedentemente. In particolare, si può utilizzare il metodo di Newton

$$x_{n+1} = x_n - \frac{P(x_n)}{P'(x_n)},$$

che richiede ad ogni passo il calcolo di $P(x_n)$ e $P'(x_n)$. Il calcolo di un polinomio in un punto $x = c$ può farsi agevolmente; infatti, posto

$$P(x) = (x - c)Q(x) + R\,, \tag{4.45}$$

si ha $P(c) = R$ e l'espressione di R insieme a quella dei coefficienti del quoziente $Q(x)$ si ottiene in funzione dei coefficienti a_i di $P(x)$ usando l'*algoritmo di Ruffini-Horner*, cioè ponendo

$$Q(x) = b_{m-1}x^{m-1} + b_{m-2}x^{m-2} + \cdots + b_0\,, \quad R = b_{-1}\,,$$

e calcolando successivamente (in base al principio di identità applicato alla (4.45))

$$\begin{aligned} b_{m-1} &= a_m\,, \\ b_{m-2} &= b_{m-1}c + a_{m-1}\,, \\ \cdots &\quad \cdots \\ b_i &= b_{i+1}c + a_{i+1}\,, \\ \cdots &\quad \cdots \\ b_{-1} &= b_0c + a_0 = R\,. \end{aligned} \tag{4.46}$$

L'algoritmo (4.46) ottimizza il calcolo di $P(c) = R$ riducendolo ad m moltiplicazioni ed altrettante addizioni ed equivale a calcolare per $x = c$ il polinomio $P(x)$ scritto nella forma

$$((\ldots((a_m x + a_{m-1})x + a_{m-2})x + a_{m-3})x + \cdots + a_1)x + a_0.$$

Con lo stesso algoritmo si può calcolare anche $P'(x)$ per $x = c$, avendosi dalla (4.45)

$$P'(x) = (x - c)Q'(x) + Q(x), \quad P'(c) = Q(c).$$

Un'operazione possibile nelle equazioni algebriche è la cosiddetta *deflazione* che consiste nell'utilizzare una radice α, già trovata, per abbassare il grado dell'equazione, in base al fatto che le rimanenti radici di $P(x) = 0$ sono quelle dell'equazione di grado $m - 1$

$$S_1(x) = \frac{P(x)}{x - \alpha} = 0\,.$$

Di questa equazione si può calcolare una nuova radice per poi ripetere una deflazione e così via, fino ad ottenere, almeno in linea di principio, tutte le radici.

Tale procedimento è però sconsigliabile in pratica, infatti quand'anche si conoscesse α esattamente, il calcolo dei coefficienti di $S_1(x)$ introduce inevitabilmente degli errori che possono privare di attendibilità i risultati successivi.

Per evitare queste difficoltà si può ricorrere alla cosddetta *deflazione implicita* in cui ogni radice già calcolata viene utilizzata per calcolarne un'altra, operando sempre sul polinomio originale $P(x)$, mediante una modifica del metodo di Newton.

Siano $\alpha_1, \alpha_2, \ldots, \alpha_m$ gli zeri di $P(x)$; partendo dalla nota decomposizione

$$P(x) = a_m(x - \alpha_1)(x - \alpha_2)\cdots(x - \alpha_m)\,,$$

e supposto di conoscere le radici $\alpha_1, \alpha_2, \ldots, \alpha_r$ $(1 \leq r < m)$, sia $S_r(x)$ il polinomio di grado $m - r$ che si otterrebbe operando senza errori r deflazioni; sarà allora

$$S_r(x) = \frac{P(x)}{(x - \alpha_1)(x - \alpha_2)\cdots(x - \alpha_r)}\,.$$

Da questa segue che si può applicare il metodo di Newton all'equazione $S_r(x) = 0$, senza conoscere esplicitamente il polinomio $S_r(x)$, ottenendo

$$x_{n+1} = x_n - \frac{1}{\frac{P'(x_n)}{P(x_n)} - \left(\frac{1}{x_n - \alpha_1} + \cdots + \frac{1}{x_n - \alpha_r}\right)} . \tag{4.47}$$

La (4.47) prende il nome di *metodo di Newton con deflazione implicita*.

Il metodo di Newton può essere usato anche nel campo complesso.

Esistono anche metodi efficienti che forniscono simultaneamente tutte le radici reali o complesse di una equazione algebrica.

Uno dei più usati consiste nel considerare le radici dell'equazione $P(x) = 0$ come autovalori della matrice di Frobenius associata a $P(x)$ (cfr. Capitolo 2) e nell'applicare per il calcolo degli autovalori il metodo QR descritto nel Capitolo 5.

Un altro metodo, che consente di approssimare tutte le radici di una equazione algebrica come soluzione di uno speciale sistema non lineare, è quello esposto nell'Esempio 6.8.3.

4.8 Complementi ed esempi

I metodi iterativi possono essere analizzati in base alla loro *efficienza*. Anche se questo concetto non ha una definizione formale unica, esso è legato essenzialmente all'ordine p del metodo (e quindi alla sua capacità di ridurre l'errore ad ogni iterazione) ed al numero v di valutazioni di funzione per ogni iterazione (e quindi alla sua complessità computazionale). Molto semplicemente, ad esempio, l'efficienza E può essere definita da

$$E = \frac{p}{v}.$$

Limitandosi al caso di radici semplici, per il metodo delle secanti (4.9) si ha $E \simeq 1.618$ avendosi $v = 1$ per ogni iterazione (tranne che per la prima), mentre risulta $E = 1$ sia per la regula falsi (4.8) sia per il metodo di Newton (4.22). Analogamente si ha $E = 1$ per il metodo di Steffensen

$$x_{n+1} = x_n - \frac{(f(x_n))^2}{f(x_n) - f(x_n - f(x_n))} \tag{4.48}$$

ottenuto dalla (4.20) ove si ponga $\phi(x) = x - f(x)$: infatti, come si verifica facilmente, per esso è $p = 2$ e $v = 2$.

Il vantaggio del metodo delle secanti rispetto ad altri metodi, ed in particolare rispetto al metodo di Newton, può risultare decisivo nel caso di funzioni complicate e di non immediata derivabilità.

Esempio 4.8.1 L'equazione

$$f(x) = x\left(\arccos\frac{1}{1+x}\right)^{\sin(\log(2x+1))} - \frac{1}{\pi} = 0$$

ha una radice in $]0,1[$.

Assumendo $x_0 = 0.6$ e $x_1 = 0.4$, il metodo delle secanti produce con 6 valutazioni di $f(x)$, $x_5 = 0.30212622627176\ldots$ dove le 14 cifre riportate sono esatte.

Il metodo di Newton, applicato con $x_0 = 0.4$, fornisce lo stesso risultato con l'iterazione x_4: sono però necessarie otto valutazioni di funzione di cui quattro di $f'(x)$, funzione, questa, dotata di una maggiore complessità computazionale. □

4.8.1 Radici multiple

Il metodo di Newton e le sue varianti (metodo delle secanti, metodo di Steffensen (4.48)) mostrano una riduzione dell'ordine di convergenza, e quindi dell'efficienza, in presenza di radici multiple.

Tuttavia, nota la molteplicità s di una radice α di $f(x) = 0$, è possibile, generalizzando la (4.28), costruire schemi iterativi aventi ordine di convergenza superiore al primo.

Esempio 4.8.2 Si consideri la funzione di iterazione

$$\phi(x) = x - \frac{2f^{(s-1)}(x)}{2f^{(s)}(x) + h(x)f^{(s+1)}(x)}$$

dove si intende $f^{(0)}(x) = f(x)$ e dove, per definizione di molteplicità di una radice, si ha $f^{(r)}(\alpha) = 0$, $r = 0, 1, \ldots, s-1$, $f^{(s)}(\alpha) \neq 0$. La funzione $h(x)$ può essere determinata in modo che il metodo $x_{n+1} = \phi(x_n)$, $n = 0, 1, \ldots$, abbia un ordine di convergenza superiore al primo.

Poiché risulta

$$\phi'(\alpha) = 1 - \frac{2f^{(s)}(\alpha)}{2f^{(s)}(\alpha) + h(\alpha)f^{(s+1)}(\alpha)}\,,$$

il metodo ha ordine di convergenza almeno due se $h(\alpha) = 0$.
Derivando ancora viene

$$\phi''(\alpha) = \frac{f^{(s)}(\alpha)\left[2f^{(s+1)}(\alpha) + h'(\alpha)f^{(s+1)}(\alpha)\right] - f^{(s+1)}(\alpha)f^{(s)}(\alpha)}{\left[f^{(s)}(\alpha)\right]^2}\,.$$

Pertanto l'ordine di convergenza è almeno tre se $\phi''(\alpha) = 0$, ossia se $h'(\alpha) = -1$.

In particolare se si assume $h(x) = -f^{(s-1)}(x)/f^{(s)}(x)$, le precedenti condizioni sono soddisfatte e poiché, come si constata facilmente, $\phi'''(\alpha) \neq 0$, il metodo ha ordine di convergenza esattamente $p = 3$.

Per esempio, applicando il metodo alla funzione $f(x) = (x^3 + x - 1)^2$ che ha uno zero di molteplicità due in $]0, 1[$, ponendo $s = 2$ e $x_0 = 1$ si ottiene:

$$x_1 = 0.73170731\ldots\,,$$
$$x_2 = 0.68269610\ldots\,,$$
$$x_3 = 0.68232780\ldots$$

(in x_3 le otto cifre decimali sono esatte). L'algoritmo di Newton (4.22), che in questo caso perde la convergenza quadratica, con $x_0 = 1$ fornisce otto cifre decimali esatte in x_{27}.

Applicando il metodo alla funzione $f(x) = x^3 + x - 1$ con $s = 1$ e $x_0 = 1$, si ottiene:

$$x_1 = 0.69230769\ldots\,,$$
$$x_2 = 0.68232811\ldots\,,$$
$$x_3 = 0.68232780\ldots\,.$$

In questo caso, per il Teorema 4.4.2, il fattore di convergenza è

$$C = \frac{1}{6} \mid \phi'''(\alpha) \mid = \frac{6\alpha^2 - 1}{3\alpha^2 + 1} \simeq 0.3122\,,$$

dove si è approssimato α con $0.68232780 \simeq x_3$.

Alternativamente C può essere stimato dalle iterazioni fatte, in base alla (4.5),

$$C \simeq \frac{\mid x_3 - x_2 \mid}{\mid x_3 - x_1 \mid^3} \simeq 0.3119.$$

Per questo metodo con $h(x) = -f^{(s-1)}(x)/f^{(s)}(x)$ risulta $E = 1$, indipendentemente da s. Tuttavia, a parte la necessaria conoscenza a priori di s, può costituire un inconveniente l'uso delle derivate successive di $f(x)$. □

4.8.2 Il caso delle equazioni algebriche

L'equazione algebrica (4.2) è dotata di importanti relazioni fra i coefficienti, reali o complessi, e le radici α_i, $i = 1, 2, \ldots, m$. Fra queste si riportano le due seguenti.

Posto

$$s_i = \sum_{1 \le k_1 < k_2 < \cdots < k_i \le m} \alpha_{k_1} \alpha_{k_2} \cdots \alpha_{k_i} , \quad i = 1, 2, \ldots, m,$$

dove in particolare è $s_1 = \alpha_1 + \alpha_2 + \cdots + \alpha_m$ e $s_m = \alpha_1 \alpha_2 \cdots \alpha_m$,

$$S_k = \sum_{i=1}^{m} \alpha_i^k, \quad k = 1, 2, \ldots,$$

si possono dimostrare le seguenti relazioni:

$$s_i = (-1)^i \frac{a_{m-i}}{a_m}, \quad i = 1, 2, \ldots, m; \tag{4.49}$$

$$\sum_{i=1}^{k} a_{m-i+1} S_{k-i+1} = -k a_{m-k}, \quad k = 1, 2, \ldots, m. \tag{4.50}$$

Le (4.50) sono dette *formule di Girard-Newton.*

Nel caso di equazioni a coefficienti interi, dalla (4.49) con $i = m$, segue facilmente che le eventuali radici razionali, della forma p/q, con $q \neq 0$, p e q interi e primi fra loro, sono da ricercarsi nell'insieme

$$\left\{ \frac{\pm \text{ divisori interi di } a_0}{\pm \text{ divisori interi di } a_m} \right\}.$$

Le (4.49) e (4.50) consentono, inoltre, di ricondurre la risoluzione di una particolare classe di sistemi algebrici non lineari alla risoluzione di una equazione algebrica.

Esempio 4.8.3 Si consideri il sistema

$$\sum_{i=0}^{2} b_i x_i^k = m_k , \quad k = 0, 1, 2, 3, 4, 5,$$

nelle incognite $b_0, b_1, b_2, x_0, x_1, x_2$. Tale sistema è formalmente analogo al sistema (7.4) con $n = 2$. Dalle prime tre equazioni si ricavano b_0, b_1, b_2, (implicate linearmente) in funzione di $x_0, x_1, x_2, m_0, m_1, m_2$, ottenendo

$$b_0 = \frac{m_0 x_1 x_2 - m_1(x_1 + x_2) + m_2}{(x_1 - x_0)(x_2 - x_0)},$$

$$b_1 = \frac{m_0 x_0 x_2 - m_1(x_0 + x_2) + m_2}{(x_0 - x_1)(x_2 - x_1)},$$

$$b_2 = \frac{m_0 x_0 x_1 - m_1(x_0 + x_1) + m_2}{(x_0 - x_2)(x_1 - x_2)}.$$

Sostituendo nelle ultime tre equazioni e ponendo $s_1 = x_0 + x_1 + x_2$, $s_2 = x_0x_1 + x_0x_2 + x_1x_2$, $s_3 = x_0x_1x_2$, si ottiene il sistema lineare in s_1, s_2, s_3,

$$\begin{pmatrix} m_2 & -m_1 & m_0 \\ m_3 & -m_2 & m_1 \\ m_4 & -m_3 & m_2 \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ s_3 \end{pmatrix} = \begin{pmatrix} m_3 \\ m_4 \\ m_5 \end{pmatrix}.$$

Ponendo $m_k = \frac{b^{k+1} - a^{k+1}}{k+1}$, si può verificare che la matrice dei coefficienti è non singolare, perció, risolvendo, si trova

$$s_1 = 3\left(\frac{b+a}{2}\right), \quad s_2 = 3\left(\frac{b+a}{2}\right)^2 - \frac{3}{5}\left(\frac{b-a}{2}\right)^2,$$

$$s_3 = \left[\left(\frac{b+a}{2}\right)^2 - \frac{3}{5}\left(\frac{b-a}{2}\right)^2\right]\frac{b+a}{2}.$$

Evidentemente, per le (4.49), i numeri x_0, x_1, x_2, sono le radici dell'equazione algebrica

$$x^3 - s_1x^2 + s_2x - s_3 = 0\,.$$

Cercando le soluzioni di questa equazione fra i divisori di s_3, si trova che $(b + a)/2$ è radice; ne segue quindi

$$x_0 = \frac{b+a}{2} - \frac{b-a}{2}\sqrt{\frac{3}{5}}, \quad x_1 = \frac{b+a}{2}, \quad x_2 = \frac{b+a}{2} + \frac{b-a}{2}\sqrt{\frac{3}{5}}.$$

Con tali valori si ricava, poi, facilmente

$$b_0 = b_2 = \frac{5}{18}(b-a), \quad b_1 = \frac{4}{9}(b-a).$$

Si noti che con $a = -1$ e $b = 1$ si ottengono i nodi e i pesi della formula di quadratura di Gauss-Legendre a tre punti (cfr. 7.1 e 7.4).

Un esempio di applicazione delle (4.50) si ha se b_0, b_1, b_2, sono arbitrariamente prefissati uguali. Dalla prima equazione del sistema si ottiene $b_0 = b_1 = b_2 = m_0/3$, mentre per il calcolo di x_0, x_1, x_2, possono usarsi le successive tre equazioni. Con $a = -1$ e $b = 1$ si ha:

$$\begin{aligned} x_0 + x_1 + x_2 &= 0 \\ x_0^2 + x_1^2 + x_2^2 &= 1 \\ x_0^3 + x_1^3 + x_2^3 &= 0\,. \end{aligned}$$

Questo sistema ha per soluzioni tutte le permutazioni dei tre numeri che sono radici del polinomio algebrico $a_3x^3 + a_2x^2 + a_1x + a_0$, i cui coefficienti possono ricavarsi con le formule di Girard-Newton. Ponendo nelle (4.50) $k = 3$, e, come suggerito dal sistema, $S_1 = 0$, $S_2 = 1$, $S_3 = 0$, si ottiene

$$\begin{aligned} &a_0 = 0 \\ &a_3 + 2a_1 = 0 \\ &a_2 + 3a_0 = 0\,. \end{aligned}$$

Fissando arbitrariamente uno dei coefficienti a_i, ad esempio $a_3 = 1$, si ha l'equazione

$$x^3 - \frac{1}{2}x = 0$$

e quindi $x_0 = -x_2 = \frac{\sqrt{2}}{2}$, $x_1 = 0$. □

4.8.3 Una particolare successione di Sturm

Si consideri la matrice tridiagonale hermitiana

$$T = \begin{pmatrix} a_1 & \overline{b}_2 & & \\ b_2 & \ddots & \ddots & \\ & \ddots & \ddots & \overline{b}_n \\ & & b_n & a_n \end{pmatrix}$$

con $b_i \neq 0$, $i = 2, 3, \ldots, n$. Il suo polinomio caratteristico $P(\lambda) = \det(A - \lambda I)$ può essere calcolato per ricorrenza come visto in 2.11.5 e si ha

$$P(\lambda) := P_n(\lambda) = (a_n - \lambda)P_{n-1}(\lambda) - \mid b_n \mid^2 P_{n-2}(\lambda)$$

con $P_0(\lambda) = 1$ e $P_1(\lambda) = a_1 - \lambda$.

Al riguardo vale il seguente teorema.

Teorema 4.8.1 *Nelle ipotesi fatte su T, la successione*

$$(-1)^n P_n(\lambda),\ (-1)^{n-1} P_{n-1}(\lambda), \ldots, -P_1(\lambda),\ P_0(\lambda), \tag{4.51}$$

è una successione di Sturm per $P(\lambda)$ e gli zeri di $P(\lambda)$ sono distinti.

Esempio 4.8.4 Si considera la matrice di ordine n

$$T = \begin{pmatrix} 0 & 1/2 & & \\ 1/2 & \ddots & \ddots & \\ & \ddots & \ddots & 1/2 \\ & & 1/2 & 0 \end{pmatrix}.$$

Si verifica facilmente che, per la successione (4.51), risulta $V(-1) - V(1) = n$. Inoltre se n è pari si ha $V(-1) - V(0) = V(0) - V(1) = n/2$, mentre se n è dispari si ha $V(-1) - V(0) = (n+1)/2$ e $V(0) - V(1) = (n-1)/2$: in questo secondo caso $\lambda = 0$ è uno zero di $P(\lambda)$. In entrambi i casi si verifica che $\lambda = 1$ non è uno zero di $P(\lambda)$; ne segue che $|\ \lambda_i\ | < 1$, $i = 1, 2, \ldots, n$. Questo risultato si trova anche applicando il Teorema 2.8.1 alla matrice T e osservando che $\lambda = 1$ e $\lambda = -1$ non possono essere autovalori in quanto le matrici $T - I$ e $T + I$ sono a predominanza diagonale debole con grafo orientato fortemente connesso e quindi non singolari (cfr. Corollario 2.11.2).

Si può poi dimostrare che gli autovalori di T sono

$$\lambda_i = \cos\left(\frac{i\pi}{n+1}\right), \quad i = 1, 2, \ldots, n.$$

□

4.8.4 Il teorema di Newton-Kantorovich

Per il metodo di Newton

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

sussiste il seguente teorema.

Teorema 4.8.2 (di Newton-Kantorovich) *Sia* $x^{(0)} \in \mathbb{R}^n$ *e* $f(x) \in C^2(D)$ *dove* $D = \left\{ x \mid \|x - x^{(0)}\| \leq 2b \right\} \subset \mathbb{R}^n$. *Se risulta:*

$$\|J^{-1}(x^{(0)})\| \leq a;$$
$$\|J^{-1}(x^{(0)}) f(x^{(0)})\| \leq b;$$
$$\sum_{r=1}^{n} \left| \frac{\partial^2 f_i}{\partial x_j \partial x_r} \right| \leq \frac{c}{n}, \quad \forall x \in D, \quad 1 \leq i, j \leq n;$$
$$abc \leq \frac{1}{2};$$

allora si ha:

$$x^{(k)} \in D,\ k = 1, 2, \ldots;$$
$$\lim_{k \to \infty} x^{(k)} = \alpha, \text{ essendo } \alpha \text{ radice di } f(x) = 0;$$
$$\alpha \text{ è unica in } D;$$
$$\|\alpha - x^{(k)}\| \leq \frac{b(2abc)^{2^k - 1}}{2^{k-1}}.$$

Per quanto non di agevole applicazione, il teorema è importante perché fornisce un legame tra la convergenza e la scelta della stima iniziale $x^{(0)}$ oltreché un risultato di esistenza e unicità della soluzione di $f(x) = 0$ (si noti che nel Teorema 4.6.2 l'esistenza della soluzione è ammessa per ipotesi).

Esempio 4.8.5 Si consideri in $\mathbb{R}^3$ il sistema

$$f(x) = \begin{pmatrix} 2x_1 + \cos x_2 + \cos x_3 - 1.9 \\ \cos x_1 + 2x_2 + \cos x_3 - 1.8 \\ \cos x_1 + \cos x_2 + 2x_3 - 1.7 \end{pmatrix} = 0.$$

La matrice jacobiana è

$$J(x) = \begin{pmatrix} 2 & -\sin x_2 & -\sin x_3 \\ -\sin x_1 & 2 & -\sin x_3 \\ -\sin x_1 & -\sin x_2 & 2 \end{pmatrix}.$$

Si assume come stima iniziale $x^{(0)} = 0$. Segue, con semplici calcoli,

$$\|J^{-1}(0)\| = \frac{1}{2} = a, \quad \|J^{-1}(0) f(0)\| = \frac{3}{20} = b\,,$$

dove si è usata la norma ∞.

Inoltre, per ogni x, si ha $\sum_{r=1}^{n} \mid \partial^2 f_i/\partial x_j \partial x_r \mid \leq 1$, $1 \leq i, j \leq 3$, da cui $c = 3$.

Poiché $abc = 9/40 < \frac{1}{2}$ il metodo di Newton è convergente.

Si noti che risulta $D = \left\{x \mid \|x\| \leq \frac{3}{10}\right\}$, per cui i punti di coordinate $\pm\frac{\pi}{2}$ non appartengono a D; ne viene che $J(x)$ è, per $x \in D$, diagonalmente dominante in senso forte e pertanto la soluzione del sistema (4.34) può essere approssimata, per ogni k, con il metodo iterativo di Jacobi o di Gauss-Seidel.

Il risultato

$$x_4 = \begin{pmatrix} -0.04236905027717\ldots \\ -0.09413400044134\ldots \\ -0.14733761588854\ldots \end{pmatrix}$$

è stato ottenuto con $x^{(0)} = 0$, usando il metodo di Jacobi. □

Bibliografia: [1], [5], [19], [25], [26], [29].

Capitolo 5

Calcolo di autovalori e autovettori

La conoscenza degli autovalori e degli autovettori di una matrice quadrata (cfr. Capitolo 2) è richiesta non solo nell'ambito di importanti teorie della matematica, ma anche in molte applicazioni, nelle quali si deve disporre di una loro buona approssimazione numerica.

Per stimare gli autovalori e gli autovettori di una matrice A sembrerebbe naturale ricorrere alla approssimazione delle radici dell'equazione caratteristica

$$\det(A - \lambda I) = 0, \tag{5.1}$$

usando i metodi studiati nel Capitolo 4 e successivamente, per ogni autovalore λ trovato, risolvere il sistema lineare omogeneo

$$(A - \lambda I)x = 0. \tag{5.2}$$

Tuttavia, tranne qualche caso speciale, (cfr. Complementi ed esempi del Capitolo 2 e Capitolo 4) è sconsigliabile seguire tale via, a causa degli inevitabili errori che si introducono nel calcolo dei coefficienti della (5.1). Infatti piccole variazioni nei coefficienti della (5.1) possono comportare forti variazioni delle radici, giungendo talvolta a mutare radici reali in complesse e viceversa. Inoltre, quand'anche si disponesse di un autovalore esatto λ, i metodi di ricerca degli autovettori associati a λ mediante la risoluzione del sistema (5.2) non sempre risultano di semplice applicazione.

Nei paragrafi che seguono sono esposti alcuni metodi iterativi più comunemente usati. Il primo di essi serve ad approssimare un autovalore di modulo

dominante ed un autovettore ad esso associato. Gli altri approssimano simultaneamente tutti gli autovalori, sfruttando la loro invarianza rispetto alle trasformazioni per similitudine; si deducono poi gli autovettori in base alla nota relazione tra autovettori di matrici simili (cfr. Teorema 2.7.2).

5.1 Metodo delle potenze

Il metodo delle potenze si fonda sul seguente teorema.

Teorema 5.1.1 *Sia $A \in \mathbb{C}^{n\times n}$ una matrice diagonalizzabile con autovalori soddisfacenti le condizioni*

$$| \lambda_1 | > | \lambda_2 | \geq \cdots \geq | \lambda_n | ; \tag{5.3}$$

sia $z^{(0)} \in \mathbb{C}^n$ un vettore arbitrario; allora il processo iterativo

$$\begin{aligned} y^{(0)} &= z^{(0)} \\ y^{(k)} &= Ay^{(k-1)}, \quad k = 1, 2, \ldots, \end{aligned} \tag{5.4}$$

è tale che

$$\lim_{k\to\infty} \frac{y^{(k)}}{y_j^{(k)}} = v, \quad \lim_{k\to\infty} \frac{y^{(k)^H} A y^{(k)}}{y^{(k)^H} y^{(k)}} = \lambda_1, \tag{5.5}$$

dove j è un indice per cui $y_j^{(k)} \neq 0$ e v è l'autovettore associato a λ_1.

Dimostrazione. La diagonalizzabilità di A implica l'esistenza di n autovettori $x^{(i)}$, $i = 1, 2, \ldots, n$, linearmente indipendenti e quindi che il vettore $z^{(0)}$ possa rappresentarsi nella forma $z^{(0)} = \sum_{i=1}^n c_i x^{(i)}$, dove è lecito supporre che sia $c_1 \neq 0$. Dalla (5.4) segue quindi

$$\begin{aligned} y^{(k)} &= A^k y^{(0)} = A^k (c_1 x^{(1)} + \cdots + c_n x^{(n)}) \\ &= c_1 A^k x^{(1)} + \cdots + c_n A^k x^{(n)} = c_1 \lambda_1^k x^{(1)} + \cdots + c_n \lambda_n^k x^{(n)} \\ &= \lambda_1^k \left(c_1 x^{(1)} + \sum_{i=2}^n c_i \left(\frac{\lambda_i}{\lambda_1} \right)^k x^{(i)} \right) \end{aligned} \tag{5.6}$$

e anche, per ogni indice j,

$$y_j^{(k)} = \lambda_1^k \left(c_1 x_j^{(1)} + \sum_{i=2}^n c_i \left(\frac{\lambda_i}{\lambda_1} \right)^k x_j^{(i)} \right). \tag{5.7}$$

In particolare, scegliendo $y_j^{(k)} \neq 0$, tenendo conto dell'ipotesi (5.3) e dividendo membro a membro la (5.6) per la (5.7), si ottiene

$$\lim_{k\to\infty} \frac{y^{(k)}}{y_j^{(k)}} = b_1 x^{(1)} = v \,, \tag{5.8}$$

dove $b_1 = 1/x_j^{(1)}$; perciò il vettore v è l'autovettore associato a λ_1, come afferma la tesi. Si ha quindi

$$Av = \lambda_1 v$$

e anche

$$v^H Av = \lambda_1 v^H v$$

da cui

$$\frac{v^H Av}{v^H v} = \lambda_1;$$

infine, tenendo conto della (5.8), si ha

$$\lim_{k\to\infty} \frac{y^{(k)^H} A y^{(k)}}{y^{(k)^H} y^{(k)}} = \lim_{k\to\infty} \frac{\left(y^{(k)}/y_j^{(k)}\right)^H A \left(y^{(k)}/y_j^{(k)}\right)}{\left(y^{(k)}/y_j^{(k)}\right)^H \left(y^{(k)}/y_j^{(k)}\right)} = \frac{v^H Av}{v^H v} = \lambda_1$$

per cui risulta dimostrata anche la seconda delle (5.5). □

Il rapporto

$$R(y^{(k)}) = \frac{y^{(k)^H} A y^{(k)}}{y^{(k)^H} y^{(k)}}$$

dicesi *quoziente di Rayleigh*.

Dalle relazioni (5.6) e (5.7) si deduce che gli errori $\frac{y^{(k)}}{y_j^{(k)}} - v$ ed $R(y^{(k)}) - \lambda_1$ tendono a zero come $\left(\frac{\lambda_2}{\lambda_1}\right)^k$.

Nel Teorema 5.1.1 l'ipotesi che A sia diagonalizzabile è, in generale, difficile da controllare; tuttavia tale ipotesi è certamente verificata per la vasta classe delle matrici normali, che sono riconoscibili in base alla proprietà $A^H A = AA^H$ (cfr. Capitolo 2). La convergenza del metodo delle potenze all'autovalore di modulo massimo e all'autovettore associato si può dimostrare anche se A non è diagonalizzabile, purché valga la condizione (5.3).

Un algoritmo basato sulla (5.4) può dar luogo a fenomeni di overflow o di underflow in quanto può produrre vettori con componenti di valore assoluto eccessivamente grande o eccessivamente piccolo. Pertanto si preferisce

ricorrere a qualche forma modificata delle iterazioni (5.4), introducendo una normalizzazione dei vettori. Un algoritmo che genera una successione di vettori $\left\{z^{(k)}\right\}$ normalizzati è il seguente,

$$\left.\begin{array}{rcl} y^{(k)} & = & Az^{(k-1)} \\ z^{(k)} & = & \frac{y^{(k)}}{\alpha_k} \end{array}\right\} k = 1, 2, \ldots, \tag{5.9}$$

dove è ancora $z^{(0)}$ arbitrario e α_k è una costante di normalizzazione opportuna. Se, ad esempio, α_k è una componente di massimo modulo di $y^{(k)}$, scelta, a partire da un certo k, sempre con lo stesso indice, risulta $\|z^{(k)}\|_\infty = 1$.[1]

Nelle ipotesi del Teorema 5.1.1, si dimostra che

$$\lim_{k\to\infty} z^{(k)} = w \quad \text{e} \quad \lim_{k\to\infty} R(z^{(k)}) = \lambda_1 \tag{5.10}$$

dove w è l'autovettore associato a λ_1.

In pratica si possono quindi assumere $R(z^{(k)})$ e $z^{(k)}$ come approssimazioni rispettivamente di λ_1 e dell'autovettore associato, adottando come criterio di arresto la condizione $\mid R(z^{(k)}) - R(z^{(k-1)}) \mid < \epsilon$ con $\epsilon > 0$ prefissato.

Il Teorema 5.1.1 si può estendere al caso più generale in cui λ_1 abbia molteplicità $r \geq 1$, modificando l'ipotesi (5.3) nella forma

$$\lambda_1 = \lambda_2 = \cdots = \lambda_r, \quad \mid \lambda_1 \mid > \mid \lambda_{r+1} \mid \geq \cdots \geq \mid \lambda_n \mid .$$

Si osservi che per $r = 1$ l'unico autovettore associato a λ_1 è approssimato da $z^{(k)}$; mentre, se $r > 1$, $z^{(k)}$ approssima un autovettore appartenente allo spazio generato dagli r autovettori associati a λ_1 e l'autovettore approssimato cambia in dipendenza dalla scelta del vettore iniziale $z^{(0)}$.

Nell'ipotesi $\mid \lambda_1 \mid > \mid \lambda_2 \mid > \mid \lambda_3 \mid \geq \cdots \geq \mid \lambda_n \mid$, il metodo delle potenze consente il calcolo dell'autovalore λ_2 e del corrispondente autovettore utilizzando la conoscenza di λ_1 e x_1. Per semplicità, ci si limita qui al caso leggermente più restrittivo delle matrici normali, le quali hanno autovettori due a due ortogonali; supponendoli normalizzati a lunghezza unitaria, per essi risulta

$$x^{(i)^H} x^{(j)} = \delta_{ij}, \quad 1 \leq i, j \leq n.$$

Applicando il metodo nella forma (5.9) con $\alpha_k = \|y^{(k)}\|_2$, e ottenuti i valori (approssimati) di $x^{(1)}$ e λ_1, si considera la matrice

$$A_1 = A - \lambda_1 x^{(1)} x^{(1)^H}.$$

[1] Per la (5.6) gli indici delle componenti di modulo massimo di $y^{(k)}$, per k sufficientemente grande, coincidono con gli indici delle componenti di massimo modulo di $x^{(1)}$.

Si constata subito che si ha

$$A_1 x^{(1)} = 0 \quad \text{e} \quad A_1 x^{(i)} = \lambda_i x^{(i)}, \quad i = 2, 3, \ldots, n;$$

la matrice A_1, quindi, ha autovalori $0, \lambda_2, \lambda_3, \ldots, \lambda_n$, e gli stessi autovettori di A. Il metodo, applicato ora a A_1, converge all'autovettore $x^{(2)}$ e all'autovalore λ_2. Questo procedimento, detto *deflazione*, può essere ripetuto, nell'ipotesi $| \lambda_1 | > | \lambda_2 | > \cdots > | \lambda_n |$, per calcolare tutti gli autovalori e tutti gli autovettori, considerando successivamente le matrici

$$A_i = A - \sum_{r=1}^{i} \lambda_r x^{(r)} x^{(r)H}, \quad i = 1, 2, \ldots, n-1.$$

5.2 Metodo delle potenze inverse

La matrice A di ordine n sia diagonalizzabile ed abbia un solo autovalore di modulo minimo, cioè sia

$$| \lambda_1 | \geq | \lambda_2 | \geq \cdots \geq | \lambda_{n-1} | > | \lambda_n | > 0;$$

la matrice inversa A^{-1} avrà quindi autovalori verificanti la condizione

$$\left| \frac{1}{\lambda_n} \right| > \left| \frac{1}{\lambda_{n-1}} \right| \geq \cdots \geq \left| \frac{1}{\lambda_1} \right|.$$

Ne segue che il metodo delle potenze, applicato alla matrice inversa A^{-1}, fornisce l'autovalore di massimo modulo $1/\lambda_n$ e, di conseguenza, anche l'autovalore di minimo modulo di A. Dalle (5.9) si ha, per $z^{(0)}$ arbitrario,

$$y^{(k)} = A^{-1} z^{(k-1)}, \quad z^{(k)} = \frac{y^{(k)}}{\alpha_k}, \quad k = 1, 2 \ldots .$$

In pratica si usa il seguente algoritmo, detto *metodo delle potenze inverse*,

$$A y^{(k)} = z^{(k-1)}, \quad z^{(k)} = \frac{y^{(k)}}{\alpha_k}, \quad k = 1, 2 \ldots, \tag{5.11}$$

nel quale ad ogni passo occorre risolvere un sistema lineare di matrice A per ottenere $y^{(k)}$. Questo metodo può essere vantaggiosamente usato per approssimare un qualunque autovalore λ_j di A quando se ne conosca già una

approssimazione iniziale $\tilde{\lambda}_j$. Partendo dall'osservazione che gli autovalori della matrice $A - \tilde{\lambda}_j I$ sono

$$\lambda_1 - \tilde{\lambda}_j, \ldots, \lambda_j - \tilde{\lambda}_j, \ldots, \lambda_n - \tilde{\lambda}_j,$$

mentre quelli della matrice $(A - \tilde{\lambda}_j I)^{-1}$ sono

$$\frac{1}{\lambda_1 - \tilde{\lambda}_j}, \ldots, \frac{1}{\lambda_j - \tilde{\lambda}_j}, \ldots, \frac{1}{\lambda_n - \tilde{\lambda}_j},$$

se $\tilde{\lambda}_j$ è abbastanza vicino a λ_j si può supporre che $\frac{1}{\lambda_j - \tilde{\lambda}_j}$ sia l'autovalore di massimo modulo per $(A - \tilde{\lambda}_j I)^{-1}$ e quindi si può applicare a questa matrice l'algoritmo (5.11) che ora assume la forma:

$$(A - \tilde{\lambda}_j I) y^{(k)} = z^{(k-1)}, \quad z^{(k)} = \frac{y^{(k)}}{\alpha_k}, \quad k = 1, 2 \ldots . \tag{5.12}$$

Arrestando, per esempio, le iterazioni per $k = m$, si ha

$$R(z^{(m)}) \simeq \frac{1}{\lambda_j - \tilde{\lambda}_j}$$

da cui si ottiene

$$\lambda_j \simeq \frac{1 + \tilde{\lambda}_j R(z^{(m)})}{R(z^{(m)})}.$$

Si noti che il vettore $z^{(m)}$ ottenuto da (5.12) approssima l'autovettore associato all'autovalore $\frac{1}{\lambda_j - \tilde{\lambda}_j}$ per la matrice $(A - \tilde{\lambda}_j I)^{-1}$, ma $z^{(m)}$ è anche un autovettore approssimato associato all'autovalore λ_j per la matrice A.

5.3 Metodo di Jacobi per matrici simmetriche

Il *metodo di Jacobi* permette di approssimare tutti gli autovalori di una matrice hermitiana. Per semplicità si considerano qui solo matrici A reali e simmetriche.

Il metodo consiste nell'operare successive trasformazioni per similitudine mediante matrici di rotazione G_{rs} della forma (cfr. Esempio 2.11.6)

$$G_{rs} = \begin{pmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ & & & \cos\varphi & & & -\sin\varphi & & & \leftarrow r \\ & & & & 1 & & & & & \\ & & & & & \ddots & & & & \\ & & & & & & 1 & & & \\ & & & \sin\varphi & & & \cos\varphi & & & \leftarrow s \\ & & & & & & & 1 & & \\ & & & & & & & & \ddots & \\ & & & & & & & & & 1 \end{pmatrix}.$$

Tali matrici sono determinate dagli interi r ed s, con $1 \leq r < s \leq n$, e dal parametro φ; inoltre sono ortogonali, cioè si ha

$$G_{rs}G_{rs}^T = G_{rs}^T G_{rs} = I \quad \text{e quindi} \quad G_{rs}^{-1} = G_{rs}^T.$$

Il metodo di Jacobi è un metodo iterativo in cui al passo k-esimo si trasforma una matrice A_k mediante una matrice ortogonale $G_{rs}^{(k)}$ secondo l'algoritmo seguente:

$$\begin{aligned} A_1 &:= A \\ A_{k+1} &= {G_{rs}^{(k)}}^T A_k G_{rs}^{(k)}, \quad k = 1, 2, \ldots, \end{aligned} \tag{5.13}$$

dove gli indici r ed s variano al variare di k.

Nella versione "classica" del metodo si scelgono gli indici di $G_{rs}^{(k)}$ coincidenti con quelli di un elemento non diagonale $a_{rs}^{(k)}$ di A_k avente modulo massimo e il valore di φ viene determinato in modo che nella matrice trasformata A_{k+1} risulti

$$a_{rs}^{(k+1)} = 0. \tag{5.14}$$

È facile constatare che la trasformazione (5.13) lascia inalterati tutti gli elementi di A_k non appartenenti alle righe e alle colonne di indici r ed s, cioè si ha

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} \quad \text{per} \quad i, j \neq r, s,$$

mentre vengono modificati solo gli elementi delle righe e colonne di indici r e s e inoltre conserva la simmetria iniziale.

Precisamente si ha:

$$a_{rs}^{(k+1)} = a_{sr}^{(k+1)} = -\left(a_{rr}^{(k)} - a_{ss}^{(k)}\right)\sin\varphi\cos\varphi + a_{rs}^{(k)}\left(\cos^2\varphi - \sin^2\varphi\right)\ , \quad (5.15)$$

$$\begin{aligned}
a_{rj}^{(k+1)} &= a_{jr}^{(k+1)} = a_{rj}^{(k)}\cos\varphi + a_{sj}^{(k)}\sin\varphi, \quad j \neq r, s,\\
a_{sj}^{(k+1)} &= a_{js}^{(k+1)} = -a_{rj}^{(k)}\sin\varphi + a_{sj}^{(k)}\cos\varphi, \quad j \neq r, s,\\
a_{rr}^{(k+1)} &= a_{rr}^{(k)}\cos^2\varphi + 2a_{rs}^{(k)}\sin\varphi\cos\varphi + a_{ss}^{(k)}\sin^2\varphi\ ,\\
a_{ss}^{(k+1)} &= a_{rr}^{(k)}\sin^2\varphi - 2a_{rs}^{(k)}\sin\varphi\cos\varphi + a_{ss}^{(k)}\cos^2\varphi\ .
\end{aligned} \quad (5.16)$$

Dalla (5.15) segue che la (5.14) è verificata per ogni φ soluzione dell'equazione

$$-\left(a_{rr}^{(k)} - a_{ss}^{(k)}\right)\sin\varphi\cos\varphi + a_{rs}^{(k)}\left(\cos^2\varphi - \sin^2\varphi\right) = 0.$$

In pratica si ricavano direttamente due valori, $\sin\varphi$ e $\cos\varphi$, per costruire la matrice $G_{rs}^{(k)}$, scrivendo la precedente equazione nella forma

$$t^2 + 2mt - 1 = 0, \quad (5.17)$$

dove si è posto $t := \tan\varphi$, $m := (a_{rr}^{(k)} - a_{ss}^{(k)})/2a_{rs}^{(k)}$.

Scegliendo fra le due radici della (5.17) quella di modulo minore, corrispondente a $|\ \varphi\ | \leq \pi/4$, si ha

$$t = \begin{cases} -m + \sqrt{1+m^2} & \text{se} \quad m > 0\ , \\ -m - \sqrt{1+m^2} & \text{se} \quad m < 0\ ; \end{cases}$$

per $m = 0$ si sceglie $t = 1$. In ogni caso si ottiene

$$\cos\varphi = \frac{1}{\sqrt{1+t^2}}, \quad \sin\varphi = t\cos\varphi,$$

con cui si costruisce la $G_{rs}^{(k)}$ voluta.

In questo modo ad ogni passo si annulla un elemento non diagonale (e il suo simmetrico); ciò non esclude che nei passi successivi un elemento non diagonale nullo possa essere modificato e assumere un valore non nullo; tuttavia vale il seguente teorema.

Teorema 5.3.1 *La successione* $\{A_k\}$ *generata dalle (5.13) col metodo di Jacobi classico converge alla matrice diagonale*

$$D = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n),$$

dove $\lambda_1, \ldots, \lambda_n$ *sono gli autovalori di* A.

In base a questo teorema, un criterio di arresto delle iterazioni (5.13) potrebbe essere il verificarsi della condizione

$$\max_{i>j} \mid a_{ij}^{(k+1)} \mid \leq \epsilon, \tag{5.18}$$

dove $\epsilon > 0$ è un numero prefissato.

Se il criterio (5.18) è verificato, allora la matrice A_{k+1} approssima la matrice diagonale D e quindi, indicando con $G^{(i)}$ la matrice di rotazione della i-esima iterazione, si ha

$$G^{(k)^T} G^{(k-1)^T} \cdots G^{(1)^T} A G^{(1)} \cdots G^{(k-1)} G^{(k)} = A_{k+1} \, ,$$

o anche, posto $H_k := G^{(1)} \cdots G^{(k)}$,

$$A H_k = H_k A_{k+1} \, .$$

Quindi gli elementi $a_{ii}^{(k+1)}$ di A_{k+1} sono approssimazioni degli autovalori di A mentre i corrispondenti autovettori sono approssimati dalle colonne della matrice H_k che si può ottenere mediante l'algoritmo

$$H_0 = I, \quad H_j = H_{j-1} G^{(j)}, \quad j = 1, 2, \ldots, k.$$

Una variante del metodo di Jacobi classico consiste nel sopprimere la ricerca dell'elemento di modulo massimo da annullare ad ogni passo, evitando così tutti i necessari confronti fra elementi. Ciò si ottiene annullando sistematicamente tutti gli elementi non nulli che si incontrano percorrendo per righe gli elementi al disopra della diagonale principale, cioè quelli di indici

$$(12), (13), \cdots, (1n); (23), \cdots, (2n); \cdots; (n-1, n).$$

L'operazione è ciclica nel senso che si ripete eseguendo gli annullamenti sempre nello stesso ordine. Per questo metodo, noto come *metodo di Jacobi ciclico*, si può dimostrare un teorema di convergenza analogo al Teorema 5.3.1.

5.4 Riduzione in forma tridiagonale e di Hessenberg

Se una matrice è hermitiana e tridiagonale, l'approssimazione dei suoi autovalori ed autovettori, sia col metodo di Jacobi che con altri metodi, risulta più agevole che per una matrice hermitiana qualunque non sparsa. Per questo motivo una generica matrice hermitiana A viene di solito trasformata in una matrice tridiagonale simile.

Fra i vari modi per effettuare la riduzione di A alla forma tridiagonale riportiamo qui il *metodo di Givens*, considerato, per semplicità, nel caso reale. In questo metodo si usano ancora le matrici di rotazione per ottenere termini nulli ma, a differenza di quanto si è visto nel metodo di Jacobi, un elemento che è stato annullato a un certo passo non viene più modificato nelle successive trasformazioni. Ciò si ottiene annullando ordinatamente i termini non nulli fra gli elementi a_{ij} con $i - j \geq 2$, considerati per colonne, nel seguente ordine

$$a_{31}, a_{41}, \ldots, a_{n1}; a_{42}, a_{52}, \ldots, a_{n2}; \ldots; a_{n,n-2},$$

usando, rispettivamente, le matrici di rotazione

$$G_{23}, G_{24}, \ldots, G_{2n}; G_{34}, G_{35}, \ldots, G_{3n}; \ldots; G_{n-1,n}. \tag{5.19}$$

Poiché ad ogni passo viene annullato un elemento e il suo simmetrico, bastano al più $\frac{(n-2)(n-1)}{2}$ rotazioni per trasformare A in una matrice simile A_1 di forma tridiagonale, conservando la simmetria.

In questo processo gli indici della matrice di rotazione e quelli dell'elemento da annullare non sono gli stessi come nel metodo di Jacobi, ma la matrice $G_{rs}^{(k)}$ viene costruita in modo che risulti

$$a_{s,r-1}^{(k+1)} = 0, \quad k = 1, 2, \ldots, \frac{(n-2)(n-1)}{2}; \tag{5.20}$$

se fosse già $a_{s,r-1}^{(k)} = 0$ si pone $G_{rs}^{(k)} = I$ (cioè $\varphi = 0$).

Utilizzando la (5.16) ove si ponga $j = r - 1$, la (5.20) fornisce

$$a_{s,r-1}^{(k+1)} = a_{s,r-1}^{(k)} \cos\varphi - a_{r,r-1}^{(k)} \sin\varphi = 0\,,$$

che è soddisfatta per

$$\cos\varphi = \frac{a_{r,r-1}^{(k)}}{\sqrt{\left(a_{s,r-1}^{(k)}\right)^2 + \left(a_{r,r-1}^{(k)}\right)^2}}, \quad \sin\varphi = \frac{a_{s,r-1}^{(k)}}{\sqrt{\left(a_{s,r-1}^{(k)}\right)^2 + \left(a_{r,r-1}^{(k)}\right)^2}}.$$

Formule numericamente più stabili si ottengono calcolando

$$t = \tan\varphi = \frac{a^{(k)}_{s,r-1}}{a^{(k)}_{r,r-1}}\,, \quad c = \cot\varphi = \frac{a^{(k)}_{r,r-1}}{a^{(k)}_{s,r-1}}\,,$$

e ponendo

$$\cos\varphi = \frac{1}{\sqrt{1+t^2}}\,, \qquad \sin\varphi = t\cos\varphi \quad \text{se} \quad |\,t\,| < 1\,,$$

$$\sin\varphi = \frac{1}{\sqrt{1+c^2}}\,, \qquad \cos\varphi = c\sin\varphi \quad \text{se} \quad |\,t\,| > 1\,.$$

La conoscenza delle matrici (5.19) consente di risalire dagli autovettori della matrice tridiagonale A_1 a quelli della matrice data. Infatti, sia y un autovettore di A_1 associato a λ e G la matrice prodotto di tutte le matrici (5.19) nell'ordine ivi considerato; G è ancora ortogonale e si ha $A_1 = G^T A G$ da cui, essendo $A_1 y = \lambda y$, segue $A_1 y = G^T A G y = \lambda y$ perciò $AGy = \lambda G y$ e $x = Gy$ è l'autovettore di A corrispondente a y.

Se il processo di Givens si applica ad una matrice A non simmetrica, la matrice H che si ottiene al termine delle $(n-2)(n-1)/2$ trasformazioni è della forma seguente

$$H = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ & \ddots & \ddots & \vdots \\ & & h_{n-1,n} & h_{nn} \end{pmatrix}.$$

H si dice *matrice di Hessenberg superiore*.

La riduzione di una matrice A non simmetrica alla forma di Hessenberg, al pari di quella delle matrici simmetriche alla forma tridiagonale, è utile ai fini dell'ulteriore approssimazione di autovalori ed autovettori.

Un metodo per ottenere queste approssimazioni e che risulta particolarmente efficiente quando opera su matrici delle due forme suddette è il *metodo* QR esposto nel paragrafo seguente.

5.5 Schema del metodo QR

L'approssimazione simultanea di tutti gli autovalori di una matrice qualunque, viene generalmente effettuata mediante il metodo QR, di cui ci si

limita ad accennare le linee fondamentali nel caso reale partendo dal seguente teorema.

Teorema 5.5.1 *Per ogni matrice $A \in \mathbb{R}^{n\times n}$ esiste una decomposizione nel prodotto di una matrice Q ortogonale per una matrice R triangolare superiore.*

La dimostrazione si ottiene costruendo effettivamente le matrici Q ed R. Uno dei vari modi per raggiungere lo scopo consiste nel premoltiplicare successivamente A con matrici di rotazione G_{rs} scelte in modo che ad ogni passo si ottenga una matrice prodotto in cui risulti nullo un elemento di indici (s, r) situato sotto la diagonale principale, ammesso che non fosse già nullo (in tal caso si pone $G_{rs} = I$). La strategia che si segue è quella di premoltiplicare A ordinatamente per le $n(n-1)/2$ matrici

$$G_{12}, G_{13}, \ldots, G_{1n}; G_{23}, \ldots, G_{2n}; \ldots; G_{n-1,n},$$

in modo che, nelle successive matrici prodotto, risultino nulli rispettivamente gli elementi di indici

$$(21), (31), \ldots, (n1); (32), \ldots, (n2); \ldots; (n, n-1).$$

In questo modo non vengono modificati gli elementi nulli ottenuti nei passi precedenti.

Poiché gli elementi annullati sono tutti situati al disotto della diagonale principale, la matrice R ottenuta dopo $n(n-1)/2$ prodotti è triangolare superiore e si ha

$$G_{n-1,n} \cdots G_{13} G_{12} A = R,$$

infine, essendo le G_{rs} ortogonali, si ha

$$A = G_{12}^T G_{13}^T \cdots G_{n-1,n}^T R = QR$$

che fornisce la cosiddetta *fattorizzazione* QR della matrice A, con la matrice $Q = G_{12}^T G_{13}^T \cdots G_{n-1,n}^T$ ortogonale.

In particolare, se la matrice A è della forma di Hessenberg superiore oppure tridiagonale, la fattorizzazione QR richiede al più $n-1$ premoltiplicazioni per matrici di rotazione, essendo al più $n-1$ gli elementi non nulli di A al disotto della diagonale principale.

Il Teorema 5.5.1 può essere generalizzato al caso di una matrice ad elementi complessi; in tal caso la matrice Q della fattorizzazione è una matrice unitaria.

L'algoritmo QR per la ricerca degli autovalori di $A \in \mathbb{R}^{n\times n}$ utilizza la fattorizzazione QR secondo lo schema

$$\begin{aligned} A_1 &:= A\,, \\ A_k &= Q_k R_k\,, \\ A_{k+1} &= R_k Q_k\,, \quad k = 1, 2, \ldots\,. \end{aligned} \tag{5.21}$$

Tutte le matrici della successione $\{A_k\}$ generata dall'algoritmo (5.21) sono simili ad A; infatti, per qualunque k, si ha $Q_k^T Q_k = I$ e anche

$$A_{k+1} = Q_k^T Q_k R_k Q_k = Q_k^T A_k Q_k.$$

Ogni matrice della successione $\{A_k\}$ ha quindi gli stessi autovalori di A; inoltre, se A è nella forma di Hessenberg, si può dimostrare che ogni matrice della successione $\{A_k\}$ si mantiene della stessa forma di A. In tal caso ad ogni passo la fattorizzazione richiede al più $(n-1)$ prodotti di matrici e il costo computazionale di una iterazione[2] è di circa $2n^2$ moltiplicazioni. Tale costo sale a n^3 moltiplicazioni se A è di forma qualsiasi.

La giustificazione teorica del metodo QR si completa con i due teoremi seguenti che ci si limita ad enunciare nel caso reale.

Teorema 5.5.2 (di Schur) *Per ogni matrice $A \in \mathbb{R}^{n\times n}$ esiste una matrice reale ortogonale B tale che la trasformata per similitudine $S = B^{-1}AB = B^T AB$ è una matrice triangolare a blocchi con i blocchi diagonali di ordine uno o due, della forma*

$$S = \begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1r} \\ & S_{22} & \cdots & S_{2r} \\ & & \ddots & \vdots \\ & & & S_{rr} \end{pmatrix}. \tag{5.22}$$

I blocchi diagonali di ordine 1 sono autovalori reali di A, mentre ogni blocco diagonale di ordine due ha come autovalori una coppia di autovalori complessi coniugati di A.

[2]Per "iterazione" si intende la fattorizzazione di A_k e il calcolo di A_{k+1} in base alle (5.21).

Teorema 5.5.3 *Se gli autovalori di $A \in \mathbb{R}^{n\times n}$ sono reali e distinti in modulo e quindi verificano la condizione*

$$| \lambda_1 | > | \lambda_2 | > \cdots > | \lambda_n | \tag{5.23}$$

e se gli autovettori di A formano una matrice X tale che X^{-1} sia fattorizzabile LR, allora le matrici A_k per $k \to \infty$ tendono ad una matrice triangolare superiore e gli elementi diagonali $a_{ii}^{(k)}$ di A_k tendono agli autovalori λ_i di A ordinati per modulo decrescente.

Se A possiede qualche coppia di autovalori complessi coniugati (e quindi la (5.23) non è verificata) ma i moduli di ciascuna coppia e quelli degli autovalori reali sono distinti e se vale l'ipotesi fatta su X^{-1}, allora le matrici A_k convergono ad una matrice triangolare a blocchi del tipo (5.22), in cui gli autovalori dei vari blocchi sono ancora ordinati per modulo decrescente.

Osservazione 5.5.1 Mancando l'ipotesi della fattorizzabilità LR di X^{-1}, si ha ancora la convergenza del metodo ma può venire meno l'ordinamento per modulo decrescente degli autovalori.

In pratica, partendo da una matrice A in forma di Hessenberg superiore le iterazioni si arrestano al raggiungimento di una matrice A_m che possiede qualche elemento della codiagonale principale talmente piccolo in valore assoluto da potersi considerare nullo. Si supponga che sia

$$A_m = \begin{pmatrix} A_{11} & A_{12} \\ \mathbf{O} & A_{22} \end{pmatrix}$$

con A_{11}, A_{22}, entrambe in forma di Hessenberg superiore.

Gli autovalori di A_m sono quindi approssimati dagli autovalori di A_{11} e A_{22}. Se per una o entrambe le matrici A_{11} e A_{22} non si è ancora raggiunta la forma triangolare o (5.22) si dovrà applicare ancora il metodo QR a una o a ciascuna delle due matrici e così via.

In linea di principio il metodo QR fornisce anche gli autovettori di una matrice A. Infatti si supponga che, dopo $m-1$ trasformazioni ortogonali, si sia ottenuta una matrice A_m (della forma (5.22) o triangolare superiore) e sia λ uno degli autovalori di A fornito da A_m; posto

$$Q = Q_1 Q_2 \cdots Q_{m-1} ,$$

Q è ancora una matrice ortogonale.

Ripetendo per Q, A_m ed A, il ragionamento fatto in 5.4 per le matrici G, A_1 ed A, segue che, se y è un autovettore di A_m associato a λ, $x = Qy$ è il corrispondente autovettore di A. Tuttavia il costo computazionale per il calcolo della matrice Q è elevato per cui, disponendo dell'approssimazione λ di un autovalore, conviene utilizzare il metodo delle potenze inverse descritto in 5.2 per ottenere un autovettore corrispondente di A.

Il metodo QR, qui esposto in una delle sue versioni più semplici, viene adoperato anche in forme modificate più efficienti; esso presenta comunque una notevole stabilità numerica e la proprietà di convergere sotto ipotesi anche più deboli di quelle del Teorema 5.5.3.

5.6 Complementi ed esempi

5.6.1 Precisazione sul metodo delle potenze

Nel metodo delle potenze, applicato nella forma (5.4), la scelta del vettore $z^{(0)}$ è arbitraria. La diagonalizzabilità di A, e quindi l'esistenza di n autovettori $x^{(i)}$ linearmente indipendenti, implica che, in ogni caso, risulti $z^{(0)} = \sum_{i=1}^{n} c_i x^{(i)}$.

Se per una data scelta di $z^{(0)}$ risulta $c_1 = 0$ e si suppone $| \lambda_2 | > | \lambda_3 |$, la successione dei vettori $y^{(k)}$ converge, come si riscontra dalla (5.6), ad un autovettore associato a λ_2 mentre $R(y^{(k)})$ tende a λ_2. Ciò accade effettivamente quando si ricorre a precisioni molto elevate. In caso contrario, gli errori di arrotondamento fanno sì che la successione $\{y^{(k)}\}$ finisca per convergere ancora, sia pure lentamente, ad un autovettore associato a λ_1 così come $R(y^{(k)})$ a λ_1. Infatti, in questo secondo caso, la (5.4) si scrive

$$y^{(k)} = Ay^{(k-1)} + \delta^{(k)}, \quad k = 1, 2, \ldots, \tag{5.24}$$

dove $\delta^{(k)}$ è il vettore di errore, anch'esso rappresentabile come una combinazione degli autovettori di A. Posto $\delta^{(1)} = d_1 x^{(1)} + \cdots + d_n x^{(n)}$, il vettore $y^{(1)}$, fornito dalla (5.24) per $k = 1$, si può esprimere nella forma

$$y^{(1)} = (c_1\lambda_1 + d_1)x^{(1)} + \cdots + (c_n\lambda_n + d_n)x^{(n)},$$

dove ora il coefficiente di $x^{(1)}$ è, in generale, diverso da zero e quindi, a partire da $y^{(1)}$, le ipotesi del Teorema 5.1.1 sono soddisfatte. Una situazione analoga si verifica per il metodo applicato nella forma normalizzata (5.9).

Esempio 5.6.1 Sia

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}$$

e si assuma $z^{(0)} = (1, 1, 1, 1)^T$: con questa scelta, nell'uguaglianza $z^{(0)} = \sum_{i=1}^{4} c_i x^{(i)}$ risulta $c_1 = 0$.

Si è applicato il metodo delle potenze in forma normalizzata con tre precisioni di macchina: $u_s \simeq 4.75 \times 10^{-7}$ (precisione semplice), $u_d \simeq 1.11 \times 10^{-16}$ (precisione doppia) e $u_q \simeq 2.58 \times 10^{-26}$ (precisione quadrupla). Si è adottato il criterio di arresto $\mid R(z^{(k)}) - R(z^{(k-1)}) \mid < \frac{1}{2}10^{-5}$. I risultati ottenuti sono nella tavola che segue.

	k	$R(z^{(k)})$
u_s	6	2.6180295...
u_d	6	2.6180339...
u_q	6	2.6180339...

Le otto cifre significative del valore riportato per la precisione doppia e la precisione quadrupla corrispondono a quelle dell'autovalore λ_2 e sono corrette. Tuttavia, proseguendo le iterazioni in precisione semplice e in precisione doppia, l'effetto degli errori di arrotondamento fa sì che, per $k \geq k^*$, $R(z^{(k)})$ si stabilizzi definitivamente su valori che approssimano λ_1. Mentre, la precisione quadrupla, essendo trascurabili gli errori di arrotondamento, continua a fornire l'approssimazione di λ_2.

Nella tavola che segue si danno i valori dell'indice k^* a partire dal quale le otto cifre significative riportate per $R(z^{(k)})$ rimangono fisse.

	k^*	$R(z^{(k)})$, $k \geq k^*$
u_s	79	3.6180315...
u_d	144	3.6180339...
u_q	6	2.6180339...

L'approssimazione di λ_1 ottenuta in precisione doppia è corretta nelle otto cifre significative riportate nella tavola. □

5.6.2 Accelerazione della convergenza

Si può dimostrare che per gli elementi della matrice A_k del metodo QR (5.21), nelle ipotesi del Teorema 5.5.3, risulta

$$a_{ij}^{(k)} = O\left(\left(\frac{\lambda_i}{\lambda_j}\right)^k\right), \quad i > j\,.$$

Ne segue che per questo metodo la convergenza può essere molto lenta se $\left|\frac{\lambda_i}{\lambda_j}\right| \simeq 1$; ad analoga conclusione si giunge, a causa della (5.6), per il metodo delle potenze se $\lambda_1 \simeq \lambda_2$.

Una tecnica che consente di ovviare a questo inconveniente consiste nell'effettuare una traslazione dello spettro di A (cfr. Teorema 2.7.6). Si considera, cioè, in luogo di A, la matrice $B = A + qI$, che ha gli stessi autovettori e i cui autovalori sono $\mu_i = \lambda_i + q$, $i = 1, 2, \ldots, n$, scegliendo il parametro q tale che sia ancora $|\ \mu_1\ | > |\ \mu_2\ | > \cdots > |\ \mu_n\ |$, ma risulti

$$\left|\frac{\mu_2}{\mu_1}\right| < \left|\frac{\lambda_2}{\lambda_1}\right|\,.$$

In tal caso il metodo delle potenze ed il metodo QR, applicati a B, convergono più rapidamente.

Esempio 5.6.2 Si consideri la matrice

$$A = \begin{pmatrix} 21 & -1 & 1 \\ -1 & 21 & 1 \\ -1 & 1 & 21 \end{pmatrix}.$$

Dal teorema di Gershgorin si ricava $19 \leq |\ \lambda_3\ | \leq |\ \lambda_2\ | \leq |\ \lambda_1\ | \leq 23$, per cui la convergenza sarà, nella migliore delle ipotesi, come quella del termine $(19/23)^k \simeq (0.826)^k$.

Effettuando una traslazione di spettro con $q = -19$, si ha la matrice B con $0 \leq |\ \mu_3\ | \leq |\ \mu_2\ | \leq |\ \mu_1\ | \leq 4$. Essendo $\mu_i = \lambda_i - q$, risulta $|\ \mu_2/\mu_1\ | \leq |\ \lambda_2/\lambda_1\ |$.

Si è applicato il metodo delle potenze in precisione doppia alle due matrici A e B, con il criterio di arresto $|\ R(z^{(k)}) - R(z^{(k-1)})\ | < 10^{-9}$ ottenendo i risultati riportati nella tavola che segue.

	k	$R(z^{(k)})$
matrice A	401	$22.0000000118\ldots$
matrice B	48	$3.0000000052\ldots$

Il metodo QR, applicato con i criteri di arresto $\max_{i>j} \mid a_{ij}^{(k)} \mid < 10^{-6}$ e $\max_{i>j} \mid b_{ij}^{(k)} \mid < 10^{-6}$, ha fornito i seguenti risultati:

$$\begin{aligned} a_{11}^{(286)} &= 22.00000166\ldots, & b_{11}^{(33)} &= 3.00000154\ldots, \\ a_{22}^{(286)} &= 19.99999944\ldots, & b_{22}^{(33)} &= 1.00000058\ldots, \\ a_{33}^{(286)} &= 20.99999888\ldots, & b_{33}^{(33)} &= 1.99999787\ldots. \end{aligned}$$

Gli autovalori di A sono $\lambda_1 = 22$, $\lambda_2 = 21$, $\lambda_3 = 20$.

Si noti inoltre che nelle matrici $A^{(k)}$ e $B^{(k)}$ gli autovalori compaiono sulla diagonale principale non ordinati per modulo decrescente: ciò è dovuto al fatto che la matrice X^{-1} di cui nel Teorema 5.5.3 non è fattorizzabile LR (cfr. Osservazione 5.5.1); si ha infatti

$$X^{-1} = \begin{pmatrix} -1 & 1 & 0 \\ 1 & -1 & 1 \\ 0 & 1 & -1 \end{pmatrix}.$$

□

Bibliografia: [2], [13], [15], [29], [31].

Capitolo 6

Interpolazione e approssimazione

In molti problemi si ha a che fare con una funzione $f(x)$ di forma non elementare, o addirittura sconosciuta, di cui si possiede solo una tabulazione in un numero finito di punti (sovente si tratta di misurazioni sperimentali).

In questi casi la stima di un valore di $f(x)$, in un punto diverso da quelli in cui è data, può essere fatta utilizzando i dati disponibili.

Questa operazione, detta *interpolazione*, di solito si effettua sostituendo a $f(x)$ una funzione che sia facilmente calcolabile come, per esempio, un polinomio. C'è quindi connessione fra il problema dell'interpolazione e quello più generale della *approssimazione* di una funzione $f(x)$, cioè della sostituzione di $f(x)$ con una funzione più semplice e che si discosti da $f(x)$ il meno possibile. Per misurare lo scostamento da $f(x)$ esistono vari criteri che danno luogo ad altrettanti metodi di approssimazione. In questo capitolo si descrivono alcune tecniche di interpolazione e l'approssimazione di una funzione col metodo dei minimi quadrati nel caso discreto.

6.1 Differenze divise

Assegnata una funzione $f(x) : \mathcal{I} \subseteq \mathbb{R} \rightarrow \mathbb{R}$ si ha la seguente definizione.

Definizione 6.1.1 *Siano* $x_0, x_1, \ldots, x_{k-1} \in \mathcal{I}$ *con* $x_i \neq x_j$ *se* $i \neq j$*; la funzione*

$$f[x_0, x_1, \ldots, x_{k-1}, x] = \frac{f[x_0, x_1, \ldots, x_{k-2}, x] - f[x_0, x_1, \ldots, x_{k-1}]}{x - x_{k-1}}, \qquad (6.1)$$

ove, per $k = 1$, $f[x_0, x] = \frac{f(x)-f(x_0)}{x-x_0}$, *è definita* $\forall x \in \mathcal{I}$, $x \neq x_i$, $i = 0, 1, \ldots, k-1$, *e si chiama* differenza divisa di ordine k.

Se $f(x)$ è derivabile su $\mathcal{I}$, la (6.1) si estende a tutto $\mathcal{I}$ (cfr. Proprietà 6.8.2).

Vale il seguente teorema di espansione, dimostrabile per induzione.

Teorema 6.1.1 *Sia* $f(x) : \mathcal{I} \subseteq \mathbb{R} \to \mathbb{R}$ *e siano* $x_0, x_1, \ldots, x_k \in \mathcal{I}$ *con* $x_i \neq x_j$ *se* $i \neq j$*; vale l'identità*

$$\begin{aligned} f(x) \;=\; & f(x_0) + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2] \\ & + \cdots + (x - x_0)(x - x_1) \cdots (x - x_{k-1}) f[x_0, x_1, \ldots, x_k] \\ & + (x - x_0)(x - x_1) \cdots (x - x_k) f[x_0, x_1, \ldots, x_k, x]. \end{aligned} \tag{6.2}$$

6.2 Interpolazione parabolica

Siano dati $k+1$ punti reali $x_0, x_1, \ldots, x_k \in \mathcal{I}$, due a due distinti, in corrispondenza dei quali siano noti i $k+1$ valori reali $f(x_0), f(x_1), \ldots\ldots, f(x_k)$. L'*interpolazione parabolica* consiste nel determinare un polinomio di grado al più k

$$P_k(x) = a_k x^k + a_{k-1} x^{k-1} + \cdots + a_1 x + a_0 \tag{6.3}$$

tale che

$$P_k(x_i) = f(x_i), \quad i = 0, 1, \ldots, k; \tag{6.4}$$

il polinomio $P_k(x)$ si chiama *polinomio di interpolazione*.

Nell'insieme dei polinomi del tipo (6.3) ne esiste uno ed uno solo che verifica le (6.4). Infatti, imponendo che il polinomio (6.3) verifichi le (6.4) si ottiene il sistema lineare di $k+1$ equazioni nelle $k+1$ incognite a_i, $i = 0, 1, \ldots, k$,

$$\begin{array}{ccccccccccc} a_0 & + & a_1 x_0 & + & \cdots & + & a_{k-1} x_0^{k-1} & + & a_k x_0^k & = & f(x_0) \\ \cdots & & \cdots & & \cdots & & \cdots & & \cdots & & \cdots \\ a_0 & + & a_1 x_k & + & \cdots & + & a_{k-1} x_k^{k-1} & + & a_k x_k^k & = & f(x_k)\,. \end{array} \tag{6.5}$$

Il sistema (6.5) ha la seguente matrice dei coefficienti

$$\begin{pmatrix} 1 & x_0 & \cdots & x_0^{k-1} & x_0^k \\ 1 & x_1 & \cdots & x_1^{k-1} & x_1^k \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_k & \cdots & x_k^{k-1} & x_k^k \end{pmatrix}$$

il cui determinante è $\prod_{0\le j<i\le k}(x_i - x_j)$ e risulta diverso da zero, essendo i punti x_i due a due distinti. Tale matrice è detta *matrice di Vandermonde*. Il sistema (6.5) ha quindi un'unica soluzione e perciò è unico il polinomio cercato.

Osservazione 6.2.1 Il polinomio di interpolazione è di grado minore di k se, nella soluzione del sistema (6.5), risulta $a_k = 0$.

Per la costruzione di $P_k(x)$ esistono procedimenti più pratici che non la risoluzione del sistema (6.5). Si possono, per esempio, utilizzare le differenze divise, in base al seguente teorema che si può dimostrare per induzione.

Teorema 6.2.1 *Il polinomio*

$$\begin{aligned} P_k(x) \;=\; & f(x_0) + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] \\ & + \cdots + (x - x_0)(x - x_1)\cdots(x - x_{k-1})f[x_0, x_1, \ldots, x_k] \end{aligned} \tag{6.6}$$

verifica le condizioni (6.4).

Il polinomio (6.6) è detto *polinomio di interpolazione di Newton.*

Una seconda forma del polinomio di interpolazione si può ottenere per mezzo delle funzioni polinomiali di grado k

$$l_r(x) = \frac{(x - x_0)\cdots(x - x_{r-1})(x - x_{r+1})\cdots(x - x_k)}{(x_r - x_0)\cdots(x_r - x_{r-1})(x_r - x_{r+1})\cdots(x_r - x_k)}\,, \tag{6.7}$$

$r = 0, 1, \ldots, k$. I polinomi (6.7) godono della proprietà

$$l_r(x_s) = \delta_{r,s}\,, \quad r, s = 0, 1, \ldots, k;$$

di conseguenza il polinomio

$$L_k(x) = \sum_{r=0}^{k} l_r(x) f(x_r)$$

verifica le condizioni (6.4). $L_k(x)$ si chiama *polinomio di interpolazione di Lagrange* e i polinomi (6.7) sono detti *polinomi fondamentali della interpolazione di Lagrange.*

L'errore che si commette se si sostituisce alla funzione $f(x)$ il polinomio di interpolazione $P_k(x)$ o $L_k(x)$ si ricava dalla (6.2) che può scriversi

$$f(x) = P_k(x) + (x - x_0) \cdots (x - x_k) f[x_0, x_1, \ldots, x_k, x];$$

si ha quindi

$$E_k(x) = f(x) - P_k(x) = \pi(x) f[x_0, x_1, \ldots, x_k, x] \,, \tag{6.8}$$

dove si è posto $\pi(x) = (x - x_0)(x - x_1) \cdots (x - x_k)$.

Teorema 6.2.2 *Se $f(x) \in C^{k+1}(\mathcal{I})$ si ha*

$$f(x) - P_k(x) = \pi(x) \frac{f^{(k+1)}(\xi)}{(k+1)!}$$

dove

$$\min\{x_0, x_1, \ldots, x_k, x\} < \xi < \max\{x_0, x_1, \ldots, x_k, x\} \,.$$

6.3 Interpolazione osculatoria di Hermite

Siano assegnati $k+1$ punti reali $x_0, x_1, \ldots, x_k \in \mathcal{I}$, due a due distinti, in corrispondenza dei quali siano noti $2k+2$ valori reali $f(x_0), f(x_1), \ldots, f(x_k)$, $f'(x_0), f'(x_1), \ldots, f'(x_k)$; *l'interpolazione osculatoria di Hermite* consiste nel determinare un polinomio $H(x)$ di grado al più $2k+1$ tale che

$$H(x_r) = f(x_r), \quad H'(x_r) = f'(x_r), \quad r = 0, 1, \ldots, k. \tag{6.9}$$

Introdotti i polinomi di grado $2k+1$

$$\begin{aligned} h_{0r}(x) &= [1 - 2l_r'(x_r)(x - x_r)] l_r^2(x), \quad r = 0, 1, \ldots, k, \\ h_{1r}(x) &= (x - x_r) l_r^2(x), \quad r = 0, 1, \ldots, k, \end{aligned} \tag{6.10}$$

è di semplice verifica il seguente teorema.

Teorema 6.3.1 *Il polinomio, di grado al più* $2k+1$*,*

$$H(x) = \sum_{r=0}^{k} h_{0r}(x) f(x_r) + \sum_{r=0}^{k} h_{1r}(x) f'(x_r) \tag{6.11}$$

soddisfa le condizioni (6.9).

Il polinomio (6.11) è detto *polinomio di interpolazione di Hermite.*

I polinomi $h_{0r}(x)$ e $h_{1r}(x)$ definiti dalle (6.10) sono detti, rispettivamente, *funzioni fondamentali di prima e di seconda specie dell'interpolazione di Hermite.*

Teorema 6.3.2 *Il polinomio* $H(x)$ *è unico.*

DIMOSTRAZIONE. Si supponga, per assurdo, che esista un secondo polinomio $S(x)$ di grado al più $2k+1$ che soddisfi le (6.9).

Il polinomio $R(x) = H(x) - S(x)$ e la sua derivata prima si annullano nei $k+1$ punti $x_0, x_1, \ldots, x_k$, per cui $R(x)$ risulta divisibile almeno per il polinomio $Q(x) = (x-x_0)^2 \cdots (x-x_k)^2$; ciò è assurdo essendo $Q(x)$ di grado $2k+2$ ed $R(x)$ di grado non superiore a $2k+1$. □

Una espressione dell'errore che si commette nel sostituire alla funzione $f(x)$ il polinomio $H(x)$ è data dal teorema seguente.

Teorema 6.3.3 *Se* $f(x) \in C^{2k+2}(\mathcal{I})$ *si ha*

$$f(x) - H(x) = (x-x_0)^2 \cdots (x-x_k)^2 \frac{f^{(2k+2)}(\xi)}{(2k+2)!}$$

dove

$$\min\{x_0, x_1, \ldots, x_k, x\} < \xi < \max\{x_0, x_1, \ldots, x_k, x\} .$$

6.4 Interpolazione con funzioni spline

I due tipi di interpolazione esposti nei precedenti paragrafi presentano lo svantaggio che, al crescere di k, aumenta, in generale, il grado del polinomio di interpolazione il quale finisce per assumere una forma poco maneggevole e può discostarsi anche sensibilmente dalla funzione $f(x)$ (cfr. Esempio 6.8.2).

Per ovviare a questo inconveniente si può ricorrere alla interpolazione mediante *funzioni spline* costituita da una *interpolazione polinomiale a tratti*

con polinomi di ugual grado su ciascun tratto, soddisfacenti certe condizioni di regolarità.

Siano dati $k+1$ punti reali $x_0 < x_1 < \cdots < x_k$ in corrispondenza dei quali siano noti $k+1$ valori reali $y_i = f(x_i)$, $i = 0, 1, \ldots, k$. Si ha la seguente definizione.

Definizione 6.4.1 *Dicesi* funzione spline di grado m, *relativa ai punti* $x_0, x_1, ..., x_k$, *una funzione* $S_m(x) : [x_0, x_k] \to \mathbb{R}$ *tale che:*

1. $S_m(x)$ *è un polinomio di grado non superiore a* m *in ogni intervallo* $[x_{i-1}, x_i]$, $i = 1, 2, \ldots, k$;
2. $S_m(x_i) = y_i, \quad i = 0, 1, \ldots, k$;
3. $S_m(x) \in C^{m-1}([x_0, x_k])$.

Nel seguito ci si limita ad analizzare il caso $m = 3$ che dà luogo alle cosiddette *spline cubiche*.

Una funzione spline cubica $S_3(x)$ è composta da k polinomi $p_i(x)$, $i = 1, 2, \ldots, k$, di grado al più 3; ciascun polinomio $p_i(x) : [x_{i-1}, x_i] \to \mathbb{R}$ è definito da quattro coefficienti. $S_3(x)$ risulterà quindi determinata dai $4k$ coefficienti dei polinomi che la compongono. Imponendo che siano verificate le proprietà 2 e 3, si ottengono le $4k-2$ condizioni

$$\begin{array}{ll} p_i(x_{i-1}) = y_{i-1}, & i = 1, 2, \ldots, k, \\ p_i(x_i) = y_i, & i = 1, 2, \ldots, k, \\ p_i'(x_i) = p_{i+1}'(x_i), & i = 1, 2, \ldots, k-1, \\ p_i''(x_i) = p_{i+1}''(x_i), & i = 1, 2, \ldots, k-1. \end{array} \tag{6.12}$$

Le due ulteriori condizioni si scelgono di solito fra le seguenti:

$$\begin{array}{ll} p_1''(x_0) = p_k''(x_k) = 0\,, & \textit{Spline naturale}; \\ p_1'(x_0) = p_k'(x_k)\,,\ p_1''(x_0) = p_k''(x_k)\,, & \textit{Spline periodica}; \\ p_1'(x_0) = y_0'\,,\ p_k'(x_k) = y_k'\,, & \textit{Spline vincolata}, \end{array}$$

se i valori $y_0' = f'(x_0)$ e $y_k' = f'(x_k)$ sono noti.

Teorema 6.4.1 *Se i punti* $x_0, \ldots, x_k$ *sono tali che* $x_i - x_{i-1} = h$, $i = 1, 2, \ldots, k$, *esiste una unica funzione spline cubica naturale* $S_3(x)$.

DIMOSTRAZIONE. Introdotti $k+1$ valori arbitrari m_i, si costruiscono i polinomi $p_i(x)$ nella forma di Hermite (6.11) a due punti con $f(x_r) = y_r$, $f'(x_r) = m_r$, $r = i-1, i$,

$$\begin{aligned} p_i(x) &= \left[1 + \frac{2}{h}(x - x_{i-1})\right] \left(\frac{x - x_i}{h}\right)^2 y_{i-1} \\ &\quad + \left[1 - \frac{2}{h}(x - x_i)\right] \left(\frac{x - x_{i-1}}{h}\right)^2 y_i \\ &\quad + (x - x_{i-1}) \left(\frac{x - x_i}{h}\right)^2 m_{i-1} + (x - x_i) \left(\frac{x - x_{i-1}}{h}\right)^2 m_i, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i = 1, 2, \ldots k. \end{aligned}$$

I polinomi $p_i(x)$ verificano, per costruzione, i primi tre gruppi di condizioni (6.12) qualunque siano i valori $m_0, m_1, \ldots, m_k$. Per determinare tali $k+1$ valori si utilizzano le ultime delle (6.12) e le due condizioni di spline naturale.

Essendo

$$p_i''(x_{i-1}) = \frac{2}{h^2}\left[3(y_i - y_{i-1}) - h(m_i + 2m_{i-1})\right],$$

$$p_i''(x_i) = \frac{2}{h^2}\left[3(y_{i-1} - y_i) + h(2m_i + m_{i-1})\right],$$

si perviene al sistema lineare

$$\begin{array}{ccccccccccc} 2m_0 & + & m_1 & & & & & & & = & \frac{3}{h}(y_1 - y_0) \\ m_0 & + & 4m_1 & + & m_2 & & & & & = & \frac{3}{h}(y_2 - y_0) \\ & & m_1 & + & 4m_2 & + & m_3 & & & = & \frac{3}{h}(y_3 - y_1) \\ & & & \ddots & & \ddots & & \ddots & & & \vdots \\ & & & & m_{k-2} & + & 4m_{k-1} & + & m_k & = & \frac{3}{h}(y_k - y_{k-2}) \\ & & & & & & m_{k-1} & + & 2m_k & = & \frac{3}{h}(y_k - y_{k-1}). \end{array} \tag{6.13}$$

La matrice dei coefficienti di questo sistema è a predominanza diagonale forte per cui è non degenere. Ne discende che la soluzione esiste ed è unica, quindi esistono, univocamente determinati, i polinomi $p_1(x), p_2(x), ..., p_k(x)$. □

Con considerazioni analoghe alle precedenti, l'esistenza e l'unicità si può dimostrare anche nel caso di partizione non uniforme, cioè con punti x_i non in progressione aritmetica, e per spline cubiche periodiche o vincolate (cfr. Esempio 6.8.9).

Nel caso delle spline cubiche vincolate e per partizioni uniformi si può dimostrare il seguente teorema di convergenza.

Teorema 6.4.2 *Se $f(x) \in C^4([x_0, x_k])$, allora esistono delle $K_p > 0$, $p = 0, 1, 2, 3$, tali che*

$$\max_{x_0 \le x \le x_k} \left| f^{(p)}(x) - S_3^{(p)}(x) \right| \le K_p M_4 h^{4-p}, \quad 0 \le p \le 2,$$

$$\max_{x_{i-1} \le x \le x_i} \left| f^{(3)}(x) - p_i^{(3)}(x) \right| \le K_3 M_4 h\,, \quad i = 1, 2 \ldots k,$$

dove $M_4 = \max_{x_0 \le x \le x_k} \left| f^{(4)}(x) \right|$.

Nelle precedenti maggiorazioni si può assumere $K_0 = \frac{7}{8}$, $K_1 = K_2 = \frac{7}{4}$, $K_3 = 2$.

6.5 Interpolazione con funzioni razionali

Un altro tipo di interpolazione si ottiene mediante l'uso di funzioni razionali.

Siano dati $n + m + 1$ punti $x_0, x_1, \ldots, x_{n+m}$, due a due distinti, in corrispondenza dei quali siano noti i valori reali $f(x_0), f(x_1), \ldots, f(x_{n+m})$.

Si consideri la funzione razionale a valori reali

$$R_m^n(x) = \frac{P_n(x)}{Q_m(x)} = \frac{a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0}{b_m x^m + b_{m-1} x^{m-1} + \cdots + b_1 x + b_0}.$$

L'*interpolazione con funzioni razionali* consiste nella determinazione degli $n + m + 2$ coefficienti a_i, $i = 0, 1, \ldots, n$, e b_i, $i = 0, 1, \ldots, m$, in modo che sia

$$R_m^n(x_i) = f(x_i), \quad i = 0, 1, \ldots, n + m. \tag{6.14}$$

Per questo è necessario che sia

$$P_n(x_i) - f(x_i) Q_m(x_i) = 0, \quad i = 0, 1, \ldots, n + m. \tag{6.15}$$

La determinazione di una funzione razionale che verifichi le (6.14) è quindi ricondotta alla risoluzione del sistema lineare omogeneo (6.15) di $n + m + 1$ equazioni nelle $n + m + 2$ incognite a_i, b_i. Il sistema (6.15) è risolubile e se la matrice dei coefficienti ha rango massimo, ha soluzioni non nulle che differiscono tra loro per una costante moltiplicativa.

Tali soluzioni non possono fornire coefficienti b_i tutti uguali a zero, cioè non possono essere tali da rendere $Q_m(x)$ identicamente nullo; infatti, se così fosse, $P_n(x)$ avrebbe $n + m + 1$ zeri quando il suo grado è al più n.

Se in corrispondenza alle soluzioni non nulle del sistema (6.15) si verifica

$$Q_m(x_r) = 0\,, \tag{6.16}$$

per qualche indice r compreso tra 0 e $n+m$, il problema dell'interpolazione mediante funzioni razionali per quei punti non ha soluzione ed i punti per i quali si verifica la (6.16) si dicono *punti inaccessibili*.

6.6 Approssimazione polinomiale

Non sempre i polinomi di interpolazione sono adatti per approssimare una funzione continua con una data accuratezza su tutto un intervallo. Infatti, per un noto *teorema di Bernstein*, dato un intervallo $[a,b]$ e fissati in esso $k+1$ punti, con $k = 1, 2, \ldots$, esiste certamente qualche funzione $f(x)$ continua su $[a,b]$, con la proprietà che la successione dei polinomi interpolanti $P_1(x), P_2(x), \ldots$, di grado pari all'indice, non converge uniformemente ad $f(x)$. Tuttavia, se, per una funzione continua $f(x)$, si cerca nella classe di tutti i polinomi, anziché fra quelli della particolare successione sopra definita, allora esiste qualche polinomio che approssima $f(x)$ quanto si vuole, uniformemente su $[a,b]$. Vale infatti il seguente fondamentale teorema.

Teorema 6.6.1 (di Weierstrass) *Sia $f(x) \in C^0([a,b])$; allora per ogni $\epsilon > 0$ esiste un intero n, dipendente da ϵ, ed un polinomio $p(x)$ di grado al più n, tale che sia*

$$|\, f(x) - p(x)\, | < \epsilon\,, \quad \forall x \in [a,b]. \tag{6.17}$$

Per esempio, Bernstein ha dimostrato che si possono costruire effettivamente polinomi $p(x)$ aventi la proprietà (6.17), in base al seguente teorema.

Teorema 6.6.2 *Data $f(x) \in C^0([0,1])$, si definisca il polinomio di grado n*

$$B_n(f;x) = \sum_{i=0}^{n} \frac{n!}{i!(n-i)!} x^i (1-x)^{n-i} f\left(\frac{i}{n}\right)\,, \tag{6.18}$$

allora la successione di polinomi $\{B_n\}$ converge uniformemente a $f(x)$ su $[0,1]$.

Si noti che, effettuando nella (6.18) la sostituzione $x = \frac{t-a}{b-a}$, il polinomio trasformato $B_n(f;t)$ risulta definito su $[a,b]$ e quindi può sostituirsi al polinomio $p(x)$ della (6.17).

L'approssimazione, polinomiale o no, di una funzione $f(x)$ può anche ottenersi richiedendo che siano verificate proprietà diverse dalla (6.17); per esempio, spesso nelle applicazioni si impone la proprietà che la funzione approssimante minimizzi una opportuna norma euclidea. Con questo criterio, detto dei *minimi quadrati*, si possono, per esempio, costruire approssimazioni che utilizzano un insieme discreto di punti dati $[x_i, f(x_i)]$, senza necessariamente imporre al grafico della funzione approssimante di passare per quei punti, come nel caso dell'interpolazione.

6.7 Metodo dei minimi quadrati nel discreto

Siano date $m+1$ funzioni $\phi_i(x)$, $i = 0, 1, \ldots, m$, $m \leq k$, continue almeno su un insieme $\mathcal{I}$ contenente $k+1$ punti x_j, $j = 0, 1, \ldots, k$, e si abbiano i valori $f(x_j)$ per ogni x_j. Si consideri la funzione combinazione lineare delle $\phi_i(x)$

$$\Phi(x) = c_0\phi_0(x) + c_1\phi_1(x) + \cdots + c_m\phi_m(x)$$

dove $c_i \in \mathbb{R}$, $i = 0, 1, \ldots, m$. Si vuole approssimare $f(x)$ con $\Phi(x)$.

È evidente che la funzione $\Phi(x)$ dipende dalla scelta dei coefficienti c_i. Il metodo dei minimi quadrati consiste nello scegliere i coefficienti c_i per i quali la funzione

$$\Psi(c_0, c_1, \ldots, c_m) = \sum_{j=0}^{k} \left[\sum_{i=0}^{m} c_i\phi_i(x_j) - f(x_j) \right]^2 \tag{6.19}$$

assume valore minimo. La funzione (6.19) rappresenta la somma degli *scarti quadratici* tra la funzione $\Phi(x)$ e la funzione $f(x)$ nei punti x_j e coincide con il quadrato della norma euclidea del vettore di componenti $\Phi(x_j) - f(x_j)$. Da qui la denominazione del metodo.

La (6.19) è una funzione continua di $m+1$ variabili, per cui il punto di minimo assoluto (se esiste) è da ricercarsi tra i punti di $\mathbb{R}^{m+1}$ per i quali sono nulle tutte le derivate parziali della Ψ rispetto a $c_0, c_1, \ldots, c_m$.

Si hanno quindi le relazioni

$$\frac{\partial \Psi}{\partial c_s} = 2 \sum_{j=0}^{k} \left[\sum_{i=0}^{m} c_i\phi_i(x_j) - f(x_j) \right] \phi_s(x_j) = 0, \quad s = 0, 1, \ldots, m, \tag{6.20}$$

che costituiscono un sistema lineare di $m+1$ equazioni nelle $m+1$ incognite $c_0, c_1, \ldots, c_m$.

Ponendo $c = (c_0, c_1, \ldots, c_m)^T$, $b = (f(x_0), f(x_1), \ldots, f(x_k))^T$,

$$A = \begin{pmatrix} \phi_0(x_0) & \phi_1(x_0) & \cdots & \phi_m(x_0) \\ \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_m(x_1) \\ \cdots & \cdots & \cdots & \cdots \\ \phi_0(x_k) & \phi_1(x_k) & \cdots & \phi_m(x_k) \end{pmatrix},$$

la (6.19) può scriversi

$$\Psi(c) = (Ac - b)^T(Ac - b) = \|Ac - b\|_2^2 \,, \tag{6.21}$$

mentre il sistema (6.20), detto sistema delle *equazioni normali*, può essere scritto nella forma

$$A^T Ac - A^T b = 0. \tag{6.22}$$

Questo sistema ha soluzione, in quanto si può dimostrare che $r(A^T A) = r(A^T A \mid A^T b)$.

Nel caso in cui la matrice $A \in \mathbb{R}^{(k+1)\times(m+1)}$ abbia rango uguale a $m+1$, per il Teorema 2.3.1, la matrice $A^T A \in \mathbb{R}^{(m+1)\times(m+1)}$ ha determinante diverso da zero; quindi il sistema (6.22) ha un'unica soluzione e la funzione $\Psi(c)$ ha un unico punto stazionario. Tale punto stazionario risulta il punto di minimo assoluto. Infatti, la matrice hessiana della (6.21) è la matrice $A^T A$ che, nel caso in esame, risulta definita positiva.

La scelta delle funzioni $\phi_i(x)$ può, in pratica, dipendere dalla distribuzione sul piano cartesiano dei punti di coordinate $[x_j, f(x_j)]$, $j = 0, 1, \ldots, k$. Le scelte più comuni sono:

$$\begin{aligned}
&\phi_0(x) = 1,\ \phi_1(x) = x,\ \phi_2(x) = x^2, \ldots; \\
&\phi_0(x) = \tfrac{1}{2}, \quad \phi_1(x) = \cos x, \quad \phi_2(x) = \sin x, \\
&\phi_3(x) = \cos(2x), \quad \phi_4(x) = \sin(2x), \ldots; \\
&\phi_0(x) = e^{\alpha_0 x}, \quad \phi_1(x) = e^{\alpha_1 x}, \quad \phi_2(x) = e^{\alpha_2 x}, \ldots, \\
&(\alpha_0, \alpha_1, \alpha_2 \ldots \in \mathbb{R}, \alpha_r \neq \alpha_s, \text{con } r \neq s) \,.
\end{aligned}$$

Si ha così la *migliore approssimazione* $\Phi(x)$ *nel senso dei minimi quadrati* rispettivamente di tipo polinomiale, trigonometrica ed esponenziale.

6.8 Complementi ed esempi

6.8.1 Proprietà ed applicazioni delle differenze divise

Proprietà 6.8.1 (di simmetria) *Se $i_0, i_1, \ldots, i_k$, è una qualunque permutazione dei numeri $0, 1, \ldots, k$, si ha*

$$f[x_{i_0}, x_{i_1}, \ldots, x_{i_k}] = f[x_0, x_1, \ldots, x_k].$$

DIMOSTRAZIONE. La differenza divisa $f[x_0, x_1, \ldots, x_k]$ è il coefficiente della potenza x^k del polinomio di interpolazione espresso dalla (6.6). Costruendo il polinomio di interpolazione relativo ai punti $x_{i_0}, x_{i_1}, \ldots, x_{i_k}$, ne viene che il coefficiente del termine in x^k è dato da $f[x_{i_0}, x_{i_1}, \ldots, x_{i_k}]$. Per l'unicità del polinomio di interpolazione si ha la tesi. □

Proprietà 6.8.2 *Se $f(x) \in C^1(\mathcal{I})$, la funzione $f[x_0, x_1, \ldots, x_{k-1}, x]$ è prolungabile per continuità su tutto $\mathcal{I}$.*

DIMOSTRAZIONE. Si procede per induzione.

Dalla Definizione 6.1.1 e dalla derivabilità di $f(x)$ segue

$$\lim_{x \to x_i} f[x_i, x] = \lim_{x \to x_i} \frac{f(x) - f(x_i)}{x - x_i} = f'(x_i)$$

per cui le differenze divise del primo ordine risultano prolungabili per continuità ponendo $f[x_i, x_i] = f'(x_i)$, $i = 0, 1, \ldots, k$.

Supposto che le differenze divise di ordine s siano continue su $\mathcal{I}$, per le differenze divise di ordine $s+1$, dalla definizione e dalla Proprietà 6.8.1, si ha

$$\begin{aligned} f[x_0, x_1, \ldots, x_s, x] &= f[x, x_0, x_1, \ldots, x_s] \\ &= \frac{f[x, x_0, x_1, \ldots, x_{s-2}, x_s] - f[x, x_0, x_1, \ldots, x_{s-1}]}{x_s - x_{s-1}}; \end{aligned}$$

dalla continuità del numeratore di quest'ultima frazione segue la continuità su $\mathcal{I}$ delle differenze divise di ordine $s+1$ se si pone

$$f[x_0, x_1, \ldots, x_s, x_i] = \frac{f[x_i, x_0, x_1, \ldots, x_{s-2}, x_s] - f[x_i, x_0, x_1, \ldots, x_{s-1}]}{x_s - x_{s-1}}, \quad i = 0, 1, \ldots, s\,.$$

La tesi risulta così provata. □

Osservazione 6.8.1 Come conseguenza della Proprietà 6.8.2 si ha che, nell'ipotesi $f(x) \in C^1(\mathcal{I})$, la (6.1) e la (6.2) restano valide anche quando $x_0, x_1, \ldots, x_k$ non sono tutti distinti.

Proprietà 6.8.3 *Se $f(x) \in C^k(\mathcal{I})$ esiste almeno un valore ξ, con la proprietà $\min_i x_i < \xi < \max_i x_i$, tale che*

$$f[x_0, x_1, \ldots, x_k] = \frac{f^{(k)}(\xi)}{k!}.$$

DIMOSTRAZIONE. La funzione $Q(x) = f(x) - P_k(x)$ ammette come zeri i valori $x_0, x_1, \ldots, x_k$, in quanto sono verificate le condizioni (6.4). Dal teorema di Rolle discende che la funzione $Q'(x)$ ha almeno k zeri distinti appartenenti all'intervallo $]\min_i x_i, \max_i x_i[$. Analogamente, la funzione $Q''(x)$ ammette almeno $k-1$ zeri distinti e così di seguito. Quindi la funzione $Q^{(k)}(x)$ ammette almeno uno zero ξ tale che $\min_i x_i < \xi < \max_i x_i$.

Calcolando la funzione $Q^{(k)}(x)$ nel punto ξ si ha

$$Q^{(k)}(\xi) = f^{(k)}(\xi) - P_k^{(k)}(\xi) = f^{(k)}(\xi) - k! f[x_0, x_1, \ldots, x_k] = 0$$

da cui la tesi. □

Come applicazione di questa proprietà si ottiene l'espressione data nel Teorema 6.2.2 per l'errore $E_k(x) = f(x) - P_k(x)$.

Per calcolare le differenze divise che intervengono nel polinomio di interpolazione di Newton (6.6) si può ricorrere alla costruzione di un quadro delle differenze divise come in Fig. 6.1 dove nella colonna DDr sono riportate opportune differenze divise di ordine r.

x	$f(x)$	DD1	DD2	DD3
x_0	$f(x_0)$			
x_1	$f(x_1)$	$f[x_0, x_1]$		
x_2	$f(x_2)$	$f[x_0, x_2]$	$f[x_0, x_1, x_2]$	
x_3	$f(x_3)$	$f[x_0, x_3]$	$f[x_0, x_1, x_3]$	$f[x_0, x_1, x_2, x_3]$

Figura 6.1: Quadro delle differenze divise.

Il quadro può essere esteso quanto si vuole, osservando che ogni differenza divisa si ottiene sottraendo al termine che si trova nella stessa riga e nella colonna immediatamente a sinistra il primo termine di questa colonna e dividendo per la differenza dei corrispondenti x_i.

Le differenze divise che figurano nel polinomio (6.6) sono date dal primo termine di ciascuna colonna.

Osservazione 6.8.2 Nel caso in cui gli elementi della colonna delle differenze di ordine r risultino tutti uguali fra loro, gli elementi delle colonne successive sono nulli, perciò il grado del polinomio di interpolazione è r.

Osservazione 6.8.3 Il valore della differenza divisa di ordine k si può anche ottenere direttamente dalla formula

$$\begin{aligned} f[x_0, x_1, \ldots, x_k] &= \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)\cdots(x_0 - x_k)} \\ &\quad + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)\cdots(x_1 - x_k)} + \cdots \\ &\quad \cdots + \frac{f(x_k)}{(x_k - x_0)(x_k - x_1)\cdots(x_k - x_{k-1})} \end{aligned}$$

che si dimostra per induzione.

Esempio 6.8.1 È data la seguente tabella di valori

x	0	-1	2	-2	3
y	5	3	3	-9	11

.

Si vuole il polinomio di interpolazione di Newton $P_4(x)$. Dai dati si ottiene il seguente quadro delle differenze divise

x	y	DD1	DD2	DD3
0	5			
-1	3	2		
2	3	-1	-1	
-2	-9	7	-5	1
3	11	2	0	1

Le differenze divise del terzo ordine risultano uguali tra loro per cui il polinomio di interpolazione è di grado 3. Usando la (6.6), si ha

$$\begin{aligned} P_4(x) &= 5 + 2x - x(x+1) + x(x+1)(x-2) \\ &= x^3 - 2x^2 - x + 5. \end{aligned}$$

□

A conferma di quanto accennato in 6.4 si riporta un esempio nel quale al crescere del grado del polinomio di interpolazione peggiora l'approssimazione della funzione $f(x)$.

Esempio 6.8.2 Si consideri la *funzione di Runge*, rappresentata in $[0, 1]$,

$$f(x) = \frac{1}{100x^2 - 100x + 26} ,$$

per la quale risulta $f(x) \in C^\infty([0, 1])$, e si scelgano i punti equidistanti $x_i = i/k$, $i = 0, 1, \ldots, k$.

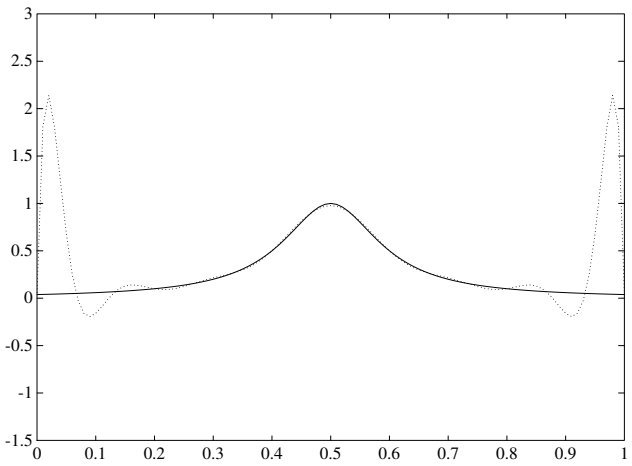

Figura 6.2: Polinomio di interpolazione $P_{15}(x)$ (linea tratteggiata) a confronto con la funzione di Runge (linea continua).

Si vede dalla Fig. 6.2 che l'errore $E_k(x)$ corrispondente al polinomio di interpolazione $P_k(x)$, in prossimità degli estremi dell'intervallo, assume valori elevati.

Si può anche dimostrare che

$$\lim_{k\to\infty} \max_{0\le x\le 1} \mid E_k(x) \mid = +\infty.$$

In generale, il comportamento dell'errore dipende dalla particolare scelta dei punti x_i, $i = 0, 1, \ldots, k$. Per esempio, se si scelgono come nodi gli zeri dei polinomi di Chebyshev di prima specie (cfr. 7.4) si dimostra che, per ogni $f(x) \in C^2([-1,1])$, $P_k(x)$ converge uniformemente ad $f(x)$ su $[-1,1]$ per $k \to \infty$ e si ha $\|f(x) - P_k(x)\|_\infty = O\left(\frac{1}{\sqrt{k}}\right)$

□

Una interessante applicazione delle differenze divise si ha nell'esempio che segue, in cui si espone il metodo di *Pasquini-Trigiante* che permette di approssimare simultaneamente tutte le radici di un'equazione algebrica.

Esempio 6.8.3 Si supponga che l'equazione

$$f(x) = x^m + a_{m-1}x^{m-1} + \cdots + a_0 = 0 \quad (a_i \in I\!R) \tag{6.23}$$

abbia m radici reali. Dalla (6.2) relativa ai punti $x_1, \ldots, x_m$, non necessariamente tutti distinti (cfr. Osservazione 6.8.1) si ha

$$\begin{aligned} f(x) \;=\; & f[x_1] + (x - x_1) f[x_1, x_2] + \cdots \\ & + (x - x_1) \cdots (x - x_{m-1}) f[x_1, \ldots, x_m] \\ & + (x - x_1) \cdots (x - x_m) f[x_1, \ldots, x_m, x] \end{aligned}$$

(dove si è posto $f[x_1] = f(x_1)$). Ne segue che se i numeri $x_1, \ldots, x_m$, verificano il sistema

$$\begin{aligned} f[x_1] &= 0 \\ f[x_1, x_2] &= 0 \\ f[x_1, x_2, x_3] &= 0 \\ \cdots \quad \cdots \quad \cdots &\quad \cdots \\ f[x_1, x_2, \ldots, x_m] &= 0\,, \end{aligned} \tag{6.24}$$

allora coincidono con le radici dell'equazione (6.23) e viceversa.

Si può dimostrare che il metodo di Newton-Raphson applicato al sistema (6.24) è convergente, comunque si scelga il vettore iniziale $x^{(0)} = (x_1^{(0)}, \ldots, x_m^{(0)})^T \in I\!R^m$, con l'esclusione al più di un insieme di punti di $I\!R^m$ di misura nulla.

Utilizzando note proprietà delle differenze divise e delle equazioni algebriche è possibile porre le iterazioni di Newton per il sistema (6.24) nella

forma

$$\begin{aligned} x_k^{(n+1)} &= x_k^{(n)} - \Delta_k, \quad k = 1, 2, \ldots, m-1, \\ x_m^{(n+1)} &= -a_{m-1} - \textstyle\sum_{k=1}^{m-1} x_k^{(n)} \, . \end{aligned}$$

Gli incrementi Δ_k si costruiscono col seguente algoritmo:

per $k = 1, 2, \ldots, m-1$
  $p_{0,k} := 1;$
  per $i = 1, 2, \ldots, m-k+1$
    $p_{i,0} := 0,$
    $p_{i,k} = p_{i,k-1} + p_{i-1,k} x_k^{(n)};$
  $P_k = \sum_{r=0}^{m-k+1} a_{r+k-1} p_{r,k};$
  per $j = 1, 2, \ldots, k$
    $q_{0,k}^{[j]} := 1,$
    per $h = 1, 2, \ldots, m-k$
      $q_{h,k}^{[j]} = p_{h,k} + q_{h-1,k}^{[j]} x_j^{(n)};$
    $Q_{j,k} = \sum_{r=0}^{m-k} a_{r+k} q_{r,k}^{[j]};$
  $\Delta_k = (P_k - \sum_{r=1}^{k-1} Q_{r,k}) / Q_{k,k}.$

Nell'algoritmo si è assunto $a_m = 1$ e $\sum_{r=1}^{0} = 0$.

Per esempio, si consideri l'equazione algebrica

$$x^6 - 4x^5 - 2x^4 + 32x^3 - 59x^2 + 44x - 12 = 0$$

le cui soluzioni sono $x_1 = x_2 = x_3 = 1$, $x_4 = x_5 = 2$, $x_6 = -3$.

Scegliendo $x^{(0)} = (0, 0, 0, 0, 0, 0)^T$, il metodo di Pasquini-Trigiante, programmato in precisione doppia, fornisce, dopo n iterazioni, i seguenti risultati che si riportano con 7 cifre decimali.

n	$x_1^{(n)}$	$x_2^{(n)}$	$x_3^{(n)}$	$x_4^{(n)}$	$x_5^{(n)}$	$x_6^{(n)}$
1	0.2727272	0.4730354	1.0979872	14.1562500	4.5000000	−16.5000000
5	0.8120109	1.1411024	0.9627325	3.5977632	0.7398782	−3.2534874
10	0.9720920	0.9717788	1.0563354	2.0696312	1.9301192	−2.9999568
20	0.9995045	0.9984509	1.0006503	2.0000913	1.9999086	−2.9999999
30	0.9999282	1.0000030	1.0000041	2.0000000	1.9999999	−3.0000000

□

6.8.2 Interpolazione inversa

In alcune applicazioni si presenta il problema di determinare per quali valori reali di x una funzione, nota per punti, assume un dato valore c.

Un modo di procedere consiste nel sostituire alla funzione $f(x)$ il polinomio di interpolazione $P_k(x)$ relativo ai punti dati e nel risolvere l'equazione algebrica

$$P_k(x) - c = 0 \tag{6.25}$$

ricorrendo, se necessario, alle tecniche esposte nel Capitolo 4.

Tuttavia, nel caso in cui la funzione $f(x)$ sia monotona, almeno su un intervallo contenente tutti i valori x_i, e quindi risulti invertibile su tale intervallo, si può evitare la risoluzione della (6.25).

Infatti, posto $y_i = f(x_i)$, $i = 0, 1, \ldots, k$, si costruisce il polinomio di interpolazione $Q_k(y)$ tale che $Q_k(y_i) = x_i$, $i = 0, 1, \ldots, k$, e si approssima il valore richiesto calcolando $Q_k(c)$.

Questa tecnica prende il nome di *interpolazione inversa*.

Esempio 6.8.4 È data la seguente tabella di valori

x	-1	0	1	3
$y(x)$	4	2	0	-2

.

Si vuole una stima del valore α per il quale $y(\alpha) = 0.5$.

Dai dati appare ragionevole supporre che la funzione $y(x)$ sia monotona al variare di x nell'intervallo contenente i quattro valori assegnati; si costruisce quindi il polinomio $Q_3(y)$ usando il seguente quadro delle differenze divise

$y(x)$	x	DD1	DD2	DD3
4	-1			
2	0	$-1/2$		
0	1	$-1/2$	0	
-2	3	$-2/3$	$1/24$	$-1/48$

Risulta

$$\begin{aligned} Q_3(y) &= -1 - \frac{1}{2}(y-4) - \frac{1}{48}y(y-4)(y-2) \\ &= -\frac{1}{48}(y^3 + 6y^2 - 32y + 48)\,. \end{aligned}$$

Una stima di α è data da

$$\alpha \simeq Q_3(0.5) = \frac{801}{1152} = 0.6953125.$$

Usando invece l'interpolazione diretta, si ha il polinomio $P_3(x) = \frac{1}{12}(x^3 - 25x + 24)$; si risolve quindi l'equazione $P_3(x) - 0.5 = 0$, ossia $x^3 - 25x + 18 = 0$, che ha un'unica soluzione nell'intervallo $[-1, 3]$. Applicando, ad esempio, il metodo di Newton si ottiene

$$\alpha \simeq 0.7359438\,.$$

□

6.8.3 Altri esempi di interpolazione

Esempio 6.8.5 È data la tabella di valori

x	-1	0	1	2	3
y	$1/3$	$1/2$	$3/5$	$2/3$	$5/7$

.

Si vogliono interpolare tali valori con la funzione

$$R_2^2(x) = \frac{a_2x^2 + a_1x + a_0}{b_2x^2 + b_1x + b_0}\,.$$

I coefficienti reali $a_0, a_1, a_2, b_0, b_1, b_2$, sono determinati dalle condizioni $R_2^2(x_i) = y_i$, $i = 0, 1, 2, 3, 4$, che forniscono il sistema lineare omogeneo di 5 equazioni in 6 incognite

$$\begin{array}{rcrcrcrcrcrcl}
a_2 & - & a_1 & + & a_0 & - & \frac{1}{3}b_2 & + & \frac{1}{3}b_1 & - & \frac{1}{3}b_0 & = & 0 \\
 & & & & a_0 & & & & & - & \frac{1}{2}b_0 & = & 0 \\
a_2 & + & a_1 & + & a_0 & - & \frac{3}{5}b_2 & - & \frac{3}{5}b_1 & - & \frac{3}{5}b_0 & = & 0 \\
4a_2 & + & 2a_1 & + & a_0 & - & \frac{8}{3}b_2 & - & \frac{4}{3}b_1 & - & \frac{2}{3}b_0 & = & 0 \\
9a_2 & + & 3a_1 & + & a_0 & - & \frac{45}{7}b_2 & - & \frac{15}{7}b_1 & - & \frac{5}{7}b_0 & = & 0\,.
\end{array}$$

Ovviamente esistono infinite soluzioni. Ponendo, per esempio, $b_2 = 1$, si ha

$$a_0 = -2,\ a_1 = 1,\ a_2 = 1,\ b_0 = -4,\ b_1 = 3,$$

per cui

$$R_2^2(x) = \frac{x^2 + x - 2}{x^2 + 3x - 4}\,.$$

Si noti che, in questo caso particolare, $R_2^2(x)$ si semplifica nella forma $\frac{x+2}{x+4}$. □

Esempio 6.8.6 Si vuole determinare la spline cubica naturale $S_3(x)$ che interpola i valori

x	0	1	2
y	a	$-a$	$2a$

dove $a \in \mathbb{R}$.

La spline $S_3(x)$ è composta da due polinomi $p_1(x), p_2(x)$ (cfr. Teorema 6.4.1), di grado al più tre, relativi agli intervalli $[0,1]$ e $[1,2]$.

Il sistema (6.13), in questo caso, diventa

$$\begin{aligned} 2m_0 &+ m_1 &+ &= -6a \\ m_0 &+ 4m_1 &+ m_2 &= 3a \\ &m_1 &+ 2m_2 &= 9a \end{aligned}$$

ed ha la soluzione

$$m_0 = -\frac{13}{4}a, \quad m_1 = \frac{1}{2}a, \quad m_2 = \frac{17}{4}a\,.$$

Si ha quindi la spline $S_3(x)$ formata dai polinomi

$$\begin{aligned} p_1(x) &= \tfrac{5}{4}ax^3 - \tfrac{13}{4}ax + a\,, && x \in [0,1]\,, \\ p_2(x) &= -\tfrac{5}{4}ax^3 + \tfrac{15}{2}ax^2 - \tfrac{43}{4}ax + \tfrac{7}{2}a\,, && x \in [1,2]\,. \end{aligned}$$

□

Esempio 6.8.7 Per la funzione $f(x) = \sin(2\pi x)$, sull'intervallo $[0,1]$, sono assegnate due tabelle di valori che si vogliono interpolare facendo ricorso a diversi tipi di interpolazione.

1. Sia data la tabella di valori

x	0	1/4	1/2	3/4	1
$f(x)$	0	1	0	−1	0

.

Si ottiene il polinomio di interpolazione di Lagrange

$$\begin{aligned} L_4(x) &= l_1(x) - l_3(x) \\ &= -\frac{128}{3}x(x-1/2)(x-3/4)(x-1) \\ &\quad +\frac{128}{3}x(x-1/4)(x-1/2)(x-1) \\ &= \frac{32}{3}(2x^3-3x^2+x)\,. \end{aligned}$$

2. Sia data la tabella di valori

x	0	1
$f(x)$	0	0
$f'(x)$	2π	2π

.

Il polinomio interpolante di Hermite risulta

$$\begin{aligned} H(x) &= h_{10}(x)2\pi + h_{11}(x)2\pi \\ &= x(x-1)^2 2\pi + (x-1)x^2 2\pi \\ &= 2\pi(2x^3-3x^2+x)\,. \end{aligned}$$

3. La spline cubica naturale $S_3(x)$ che interpola i valori della tabella del punto 1 si ottiene risolvendo il sistema lineare che risulta da (6.13) con $h = 1/4$, $k = 4$,

$$\begin{array}{rcrcrcrcrcr} 2m_0 & + & m_1 & & & & & & & = & 12 \\ m_0 & + & 4m_1 & + & m_2 & & & & & = & 0 \\ & & m_1 & + & 4m_2 & + & m_3 & & & = & -24 \\ & & & & m_2 & + & 4m_3 & + & m_4 & = & 0 \\ & & & & & & m_3 & + & 2m_4 & = & 12 \end{array}$$

la cui soluzione è

$$m_0 = 6, \quad m_1 = 0, \quad m_2 = -6, \quad m_3 = 0, \quad m_4 = 6.$$

Si ha, quindi, la spline $S_3(x)$

$$\begin{aligned} p_1(x) &= -32x^3 + 6x\,, & x \in [0, 1/4]\,, \\ p_2(x) &= 32x^3 - 48x^2 + 18x - 1\,, & x \in [1/4, 1/2]\,, \\ p_3(x) &= 32x^3 - 48x^2 + 18x - 1\,, & x \in [1/2, 3/4]\,, \\ p_4(x) &= -32x^3 + 96x^2 - 90x + 26\,, & x \in [3/4, 1]\,. \end{aligned}$$

□

6.8.4 Derivazione numerica

Si considera il problema di valutare le derivate di una funzione della quale sono assegnati i valori in alcuni punti di un intervallo dell'asse reale.

Siano $y_i = f(x_i)$, $i = 0, 1, \ldots, k$, con $x_0 < x_1 < \cdots < x_k$, i valori dati. Il problema può essere risolto approssimando $f'(x), f''(x), \ldots$, in un punto $x \in]x_0, x_k[$, con le corrispondenti derivate di un polinomio di interpolazione di $f(x)$.

Tuttavia, nell'interpolazione parabolica, polinomi di grado elevato possono scostarsi molto dai valori di $f(x)$ (cfr. Esempio 6.8.2). Inoltre, la proprietà $f(x_i) = P_k(x_i)$, $i = 1, 2, \ldots, k-1$, non implica che siano piccoli gli errori

$$| P_k'(x_i) - f'(x_i) |, | P_k''(x_i) - f''(x_i) |, \ldots, \; i = 1, 2, \ldots, k-1.$$

Per questo motivo, quando per approssimare una derivata si usa l'interpolazione parabolica, ci si limita a polinomi di grado non elevato e si stimano i valori di $f'(x), f''(x), \ldots$, utilizzando i dati solo in alcuni punti "vicini" al punto x.

In quello che segue, si danno le formule più comuni per la derivata prima e seconda.

Allo scopo si considerano i polinomi di secondo grado

$$q_i(x) = l_{i-1}(x)y_{i-1} + l_i(x)y_i + l_{i+1}(x)y_{i+1}, \quad i = 1, 2, \ldots, k-1,$$

dove

$$l_{i-1}(x) = \frac{(x - x_i)(x - x_{i+1})}{(x_{i-1} - x_i)(x_{i-1} - x_{i+1})}\,,$$

$$l_i(x) = \frac{(x - x_{i-1})(x - x_{i+1})}{(x_i - x_{i-1})(x_i - x_{i+1})} ,$$

$$l_{i+1}(x) = \frac{(x - x_{i-1})(x - x_i)}{(x_{i+1} - x_{i-1})(x_{i+1} - x_i)} .$$

Si assume quindi $f'(x) \simeq q_i'(x)$ e $f''(x) \simeq q_i''(x)$, $x \in]x_{i-1}, x_{i+1}[$. Se $x_i = x_0 + ih$, $i = 0, 1, \ldots, k$, $h > 0$, si hanno le cosiddette *formule centrali*

$$f'(x_i) \simeq q_i'(x_i) = \frac{y_{i+1} - y_{i-1}}{2h}, \quad i = 1, 2, \ldots, k-1, \tag{6.26}$$

$$f''(x_i) \simeq q_i''(x_i) = \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2}, \quad i = 1, 2, \ldots, k-1. \tag{6.27}$$

L'*errore di troncamento* della (6.26) è $E_1(x_i) = f'(x_i) - q_i'(x_i)$ e può essere determinato, nel caso che $f(x) \in C^3(]x_0, x_k[)$, considerando, per esempio, gli sviluppi di Taylor

$$y_{i+1} = y_i + f'(x_i)h + \frac{1}{2}f''(x_i)h^2 + \frac{1}{6}f^{(3)}(\theta_{i+1})h^3, \ \theta_{i+1} \in]x_i, x_{i+1}[,$$

$$y_{i-1} = y_i - f'(x_i)h + \frac{1}{2}f''(x_i)h^2 - \frac{1}{6}f^{(3)}(\theta_{i-1})h^3, \ \theta_{i-1} \in]x_{i-1}, x_i[.$$

Sottraendo membro a membro queste due relazioni si ottiene,

$$f'(x_i) = \frac{y_{i+1} - y_{i-1}}{2h} + E_1(x_i) ,$$

dove

$$E_1(x_i) = -\frac{1}{12}\left[f^{(3)}(\theta_{i+1}) + f^{(3)}(\theta_{i-1})\right] h^2 = \quad -\tfrac{1}{6}f^{(3)}(\xi_i)h^2,$$
$$\xi_i \in [\theta_{i-1}, \theta_{i+1}].$$

Analogamente, supposto $f(x) \in C^4(]x_0, x_k[)$, si ottiene l'errore di troncamento della (6.27). Dagli sviluppi di Taylor, arrestati ai termini del quarto ordine e sommati membro a membro, si ha

$$f''(x_i) = \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} + E_2(x_i) ,$$

dove

$$E_2(x_i) = -\frac{1}{24}\left[f^{(4)}(\tau_{i+1}) + f^{(4)}(\tau_{i-1})\right] h^2 = -\frac{1}{12}f^{(4)}(\eta_i)h^2 ,$$

con $\tau_{i+1} \in]x_i, x_{i+1}[$, $\tau_{i-1} \in]x_{i-1}, x_i[$ e $\eta_i \in [\tau_{i-1}, \tau_{i+1}]$.

Esempio 6.8.8 Si consideri il problema differenziale

$$\begin{aligned} &y''(x) = e^{y(x)} \,, \\ &y(0) = 0 \,, \\ &y(1) = 0 \,, \end{aligned}$$

la cui soluzione esatta è

$$y(x) = \log \left\{ \frac{1}{2} \left[\frac{c}{\cos\left(\frac{cx}{2} - \frac{c}{4}\right)} \right]^2 \right\} \,,$$

essendo la costante c, soluzione dell'equazione $c^2 - 2\cos^2(c/4) = 0$, approssimata da 1.3360556949061.

Fatta una partizione dell'intervallo $[0, 1]$ mediante i punti $x_i = ih$, $i = 1, 2, \ldots, k-1$, $h = 1/k$, si vogliono calcolare i valori y_i che approssimano, in tali punti, la soluzione $y(x)$ del problema dato.

Posto $y_0 = y(0) = 0$, $y_k = y(1) = 0$ e tenuto conto che per la (6.27) può scriversi

$$y''(x_i) \simeq \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2}, \quad i = 1, 2, \ldots, k-1,$$

i valori cercati sono la soluzione del sistema non lineare

$$\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} = e^{y_i}, \quad i = 1, 2, \ldots, k-1.$$

Posto $y^T = (y_1, \ldots, y_{k-1})$, tale sistema è della forma $g(y) = 0$ con $g^T(y) = (g_1(y), g_2(y), \ldots, g_{k-1}(y))$,

$$\begin{aligned} g_1(y) &= -2y_1 - h^2 e^{y_1} + y_2 \,, \\ g_i(y) &= y_{i-1} - 2y_i - h^2 e^{y_i} + y_{i+1} \,, \quad i = 2, \ldots, k-2 \,, \\ g_{k-1}(y) &= y_{k-2} - 2y_{k-1} - h^2 e^{y_{k-1}} \,. \end{aligned}$$

Applicando il metodo di Newton $y^{(r+1)} = y^{(r)} - J^{-1}(y^{(r)})g(y^{(r)})$, $r = 0, 1, \ldots$, con $k = 10$, $y^{(0)} = 0$ e arrestando le iterazioni al verificarsi di

$\|y^{(r+1)} - y^{(r)}\|_\infty < 10^{-6}$, si è ottenuto

$$y^{(3)} \simeq \begin{pmatrix} -0.041404 \\ -0.073214 \\ -0.095730 \\ -0.109159 \\ -0.113622 \\ -0.109159 \\ -0.095730 \\ -0.073214 \\ -0.041404 \end{pmatrix},$$

dove l'errore $\max_{1\le i\le k-1} \mid y_i^{(3)} - y(x_i) \mid$ è dell'ordine di 10^{-4}. □

Per la derivazione approssimata si possono anche utilizzare funzioni spline. Come è noto, esse hanno grado indipendente dal numero dei punti della partizione, e buone caratteristiche di convergenza (cfr. Teorema 6.4.2).

Se quindi si conosce la spline $S_3(x)$ che interpola i valori $y_i = f(x_i)$, $x_i = x_0 + ih$, $i = 0, 1, \ldots, k$, si assume, per ogni $x \in [x_0, x_k]$, $f'(x) \simeq S_3'(x)$ e $f''(x) \simeq S_3''(x)$. In particolare, dalla dimostrazione del Teorema 6.4.1, risulta

$$\begin{aligned} f'(x_i) &\simeq m_i\,, \quad i = 0, 1, \ldots, k, \\ f''(x_0) &\simeq \tfrac{2}{h^2}\left[3(y_1 - y_0) - h(m_1 + 2m_0)\right], \\ f''(x_i) &\simeq \tfrac{2}{h^2}\left[3(y_{i-1} - y_i) + h(2m_i + m_{i-1})\right], \quad i = 1, 2, \ldots, k. \end{aligned}$$

Esempio 6.8.9 È data la seguente tavola dei valori di $y = f(x)$

x	0	0.1	0.2	0.3
y	1	100/101	25/26	100/109

;

inoltre sono assegnati $y_0' = f'(0) = 0$ e $y_3' = f'(0.3) = -6000/11881$.

Si vogliono stimare $f'(0.1)$, $f''(0.1)$, e $f'(0.2)$, $f''(0.2)$.

È conveniente ricorrere ad una spline vincolata al fine di utilizzare i valori dati di $f'(x)$ negli estremi dell'intervallo. In casi di questo genere, poiché i momenti m_0 e m_k sono noti, nel sistema (6.13) la prima e l'ultima equazione vengono soppresse mentre nella seconda e nella penultima si portano m_0 e m_k a secondo membro. Il sistema quindi si riduce a sole $k - 1$ equazioni in altrettante incognite.

Nella fattispecie si ha

$$\begin{aligned} 4m_1 &+ m_2 &= \tfrac{3}{h}(y_2 - y_0) - y_0' \\ m_1 &+ 4m_2 &= \tfrac{3}{h}(y_3 - y_1) - y_3' \,. \end{aligned}$$

Posto $h = 0.1$, inserendo i dati e arrotondando i calcoli alla sesta cifra dopo la virgola, si ricava

$$m_1 = -0.196024\,, \quad m_2 = -0.369750\,,$$

che possono essere assunti come stime di $f'(0.1)$ e $f'(0.2)$ rispettivamente.

Le derivate seconde possono essere approssimate con (cfr. 6.4)

$$\begin{aligned} p_2''(x_1) &= \tfrac{2}{h^2}\left[3(y_2 - y_1) - h(m_2 + 2m_1)\right]\,, \\ p_2''(x_2) &= \tfrac{2}{h^2}\left[3(y_1 - y_2) + h(2m_2 + m_1)\right]\,. \end{aligned}$$

Sostituendo i valori trovati per m_1 ed m_2 ed arrotondando i calcoli come sopra si ottiene

$$p_2''(0.1) = -1.900369, \quad p_2''(0.2) = -1.574151,$$

che si assumono come stime di $f''(0.1)$ e $f''(0.2)$.

Al fine di evidenziare la bontà delle approssimazioni ottenute, si effettua il confronto con i valori esatti essendo i dati della tavola quelli della funzione $f(x) = \frac{1}{1+x^2}$. Si ha:

$$\begin{aligned} &| \, m_1 - f'(0.1) \, | \simeq 3.5 \times 10^{-5}, \quad | \, m_2 - f'(0.2) \, | \simeq 7.2 \times 10^{-5}, \\ &| \, p_2''(0.1) - f''(0.1) \, | \simeq 1.7 \times 10^{-2}, \quad | \, p_2''(0.2) - f''(0.2) \, | \simeq 1.3 \times 10^{-1}. \end{aligned}$$

Se, in luogo della spline vincolata, si usano le formule (6.26) e (6.27) si ottiene:

$$\begin{aligned} f'(0.1) &\simeq q_1'(0.1) &= -0.192307, \\ f'(0.2) &\simeq q_2'(0.2) &= -0.363339, \\ f''(0.1) &\simeq q_2''(0.1) &= -1.865956, \\ f''(0.2) &\simeq q_2''(0.2) &= -1.554672, \end{aligned}$$

da cui gli errori

$$\begin{aligned} &| \, q_1'(0.1) - f'(0.1) \, | \simeq 3.8 \times 10^{-3}, \quad | \, q_2'(0.2) - f'(0.2) \, | \simeq 6.5 \times 10^{-3}, \\ &| \, q_1''(0.1) - f''(0.1) \, | \simeq 1.7 \times 10^{-2}, \quad | \, q_2''(0.2) - f''(0.2) \, | \simeq 1.5 \times 10^{-1}. \end{aligned}$$

Si spiega la minore precisione raggiunta in questo secondo caso con il fatto che le formule (6.26) e (6.27) non tengono conto dei dati $f'(0)$ e $f'(0.3)$.

□

6.8.5 Sistemi lineari sovradeterminati

Siano k e m due numeri naturali con $k > m$ e inoltre si abbiano $A \in \mathbb{R}^{k\times m}$ e $b \in \mathbb{R}^k$. Il *sistema lineare sovradeterminato*

$$Ax = b \tag{6.28}$$

ha soluzione se e solo se $r(A) = r(A \mid b)$ (cfr. Teorema 2.5.1).

Nel caso $r(A) < r(A \mid b)$, il sistema (6.28) non ha soluzione e quindi per ogni vettore $x \in \mathbb{R}^m$ si ha

$$b - Ax = r$$

dove $r \in \mathbb{R}^k$, il vettore dei residui, risulta non nullo.

Si dice che si risolve il sistema (6.28) nel senso dei minimi quadrati se si determina un vettore $x \in \mathbb{R}^m$ che renda minimo il prodotto scalare

$$\Psi(x) = (b - Ax)^T(b - Ax) = r^T r = \sum_{i=1}^{k} r_i^2 = \|r\|_2^2. \tag{6.29}$$

La determinazione di un punto di minimo è ricondotta alla ricerca dei punti che annullano tutte le derivate parziali prime della $\Psi(x)$ ossia alla risoluzione del sistema lineare

$$\frac{\partial}{\partial x_i}\Big((b - Ax)^T(b - Ax)\Big) = 0, \quad i = 1, 2, \ldots, m,$$

che si può scrivere nella forma

$$A^T Ax = A^T b\,, \tag{6.30}$$

la cui risolubilità risulta da quanto si è detto per il sistema (6.22).

La risoluzione del sistema (6.28) nel senso dei minimi quadrati è semplificata se della matrice A si conosce una decomposizione ai valori singolari (cfr. 2.9) $A = U\Sigma V^T$ con $\Sigma \in \mathbb{R}^{k\times m}$, mentre $U \in \mathbb{R}^{k\times k}$ e $V \in \mathbb{R}^{m\times m}$ sono matrici unitarie.

Si può dimostrare che, indicando con u_i e v_i le colonne, rispettivamente, delle matrici U e V e con σ_i, $i = 1, 2, \ldots, s$ $(s = r(A))$, i valori singolari non nulli della matrice A, un vettore c che minimizza la $\Psi(x)$ è dato da

$$c = \sum_{i=1}^{s} \frac{u_i^T b}{\sigma_i} v_i\,. \tag{6.31}$$

Si può verificare che a tale vettore corrisponde lo scarto quadratico

$$(b - Ac)^T(b - Ac) = \sum_{i=s+1}^{k} \left| u_i^T b \right|^2 .$$

Se $s = m$, c è l'unica soluzione del sistema (6.30) mentre se $s < m$ nello spazio delle soluzioni del sistema (6.30) c è il vettore con norma euclidea minima.

Esempio 6.8.10 Si consideri il sistema lineare sovradeterminato

$$\begin{cases} 2x_1 - x_2 = 1 \\ -x_1 + x_2 = 0 \\ x_1 + 2x_2 = 1 . \end{cases}$$

La matrice del sistema è quindi

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

I valori singolari della matrice A, ricavati in base al Teorema 2.9.2, sono $\sigma_1 = \sqrt{7}$ e $\sigma_2 = \sqrt{5}$ ed una decomposizione ai valori singolari è $A = U\Sigma V^T$ con

$$U = \frac{\sqrt{70}}{70} \begin{pmatrix} 3\sqrt{5} & \sqrt{7} & 3\sqrt{2} \\ -2\sqrt{5} & 0 & 5\sqrt{2} \\ -\sqrt{5} & 3\sqrt{7} & -\sqrt{2} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sqrt{7} & 0 \\ 0 & \sqrt{5} \\ 0 & 0 \end{pmatrix},$$

$$V = \frac{\sqrt{2}}{2} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} .$$

La (6.31) fornisce la soluzione del sistema nel senso dei minimi quadrati

$$c = \left(\frac{19}{35}, \frac{9}{35} \right)^T .$$

Alla stessa soluzione si giunge risolvendo il sistema (6.30). □

Esempio 6.8.11 Si consideri il sistema lineare $Ax = b$ con $A \in \mathbb{R}^{n\times 2}$ e $b \in \mathbb{R}^n$ dati da

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 2 \\ \vdots \\ n-1 \\ n \end{pmatrix}.$$

La matrice A ha un unico valore singolare positivo $\sigma_1 = \sqrt{2n}$ e gli autovettori delle matrici AA^T e A^TA associati a σ_1^2 sono, rispettivamente,

$$u = \frac{\sqrt{n}}{n}\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \quad v = \frac{\sqrt{2}}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Un vettore che minimizza la $\Psi(x)$ è, per la (6.31),

$$c = \frac{u^T b v}{\sigma_1} = \frac{n+1}{4}\begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Poiché in questo caso è $r(A) = 1$ (quindi $s < m$), il vettore c ha norma euclidea minima tra tutte le soluzioni del sistema (6.30). Allo stesso risultato si giunge risolvendo direttamente il sistema (6.30) che assume ora la forma

$$\begin{pmatrix} n & n \\ n & n \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = n\frac{n+1}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

e le cui soluzioni sono

$$c = \begin{pmatrix} c_1 \\ \frac{n+1}{2} - c_1 \end{pmatrix}.$$

La norma euclidea

$$\|c\|_2 = \sqrt{c_1^2 + \left(\frac{n+1}{2} - c_1\right)^2}$$

è minima se $c_1 = (n+1)/4$. □

Bibliografia: [1], [19], [26] [29], [27].

Capitolo 7

Integrazione numerica

In questo capitolo si studiano alcuni metodi per il calcolo approssimato di integrali definiti.

Alcuni motivi che consigliano l'uso di metodi approssimati in luogo di metodi analitici ("esatti") sono i seguenti:

di molte funzioni integrabili non si conosce una funzione primitiva esprimibile con funzioni elementari (per esempio, $f(x) = e^{\cos x}$, $f(x) = x^x$);

in molti casi, funzioni semplici hanno funzioni primitive tanto complicate che spesso le formule approssimate sono più facilmente calcolabili di quelle esatte e con maggiore precisione (per esempio, una primitiva di $f(x) = \frac{x^2}{1+x^4}$ è $F(x) = \frac{\sqrt{2}}{4}\left[\frac{1}{2}\log\frac{x^2-\sqrt{2}x+1}{x^2+\sqrt{2}x+1} + \arctan\frac{\sqrt{2}x}{1-x^2}\right]$);

spesso della funzione integranda è nota solo una restrizione a un insieme discreto.

7.1 Grado di precisione ed errore

Sia $f(x)$ sufficientemente regolare sull'intervallo $[a, b]$ dell'asse reale. Sia $\rho(x)$ una *funzione peso* non negativa in $[a, b]$ e tale che esistano finiti i *momenti*

$$m_k = I(x^k\rho) = \int_a^b x^k\rho(x)dx, \quad k = 0, 1, \dots . \tag{7.1}$$

Si pone il problema di approssimare

$$I(\rho f) = \int_a^b \rho(x) f(x) dx$$

con una *formula di quadratura* della forma

$$J_n(f) = \sum_{i=0}^{n} a_i f(x_i) \,, \tag{7.2}$$

dove i numeri a_i, $i = 0, 1, \ldots, n$, detti *pesi* o *coefficienti*, sono reali. I punti x_i, $i = 0, 1, \ldots, n$, con $x_0 < x_1 < \cdots < x_n$ sono detti *nodi* e di solito appartengono all'intervallo $[a, b]$. L'errore è perciò formalmente dato da

$$E_n(f) = I(\rho f) - J_n(f) \,.$$

Considerata la base $1, x, x^2, \ldots, x^m, x^{m+1}$, dello spazio vettoriale dei polinomi algebrici di grado al piú $m + 1$, si puó dare la seguente definizione.

Definizione 7.1.1 *La formula (7.2) ha* grado di precisone *(algebrico)* $m \in \mathbb{N}$ *se si verifica*

$$E_n(1) = E_n(x) = \cdots = E_n(x^m) = 0, \quad E_n(x^{m+1}) \neq 0 \,. \tag{7.3}$$

Una formula del tipo (7.2) è individuata una volta che lo siano i nodi e i pesi. Per determinare gli $2n + 2$ parametri si impone che sia

$$E_n(1) = E_n(x) = \cdots = E_n(x^{2n+1}) = 0, \tag{7.4}$$

ovvero

$$\begin{array}{lcllcllcl} a_0 & + & a_1 & + & \cdots & + & a_n & = & m_0 \\ a_0 x_0 & + & a_1 x_1 & + & \cdots & + & a_n x_n & = & m_1 \\ \cdots & & \cdots & & \cdots & & \cdots & & \cdots \\ a_0 x_0^{2n+1} & + & a_1 x_1^{2n+1} & + & \cdots & + & a_n x_n^{2n+1} & = & m_{2n+1} \,, \end{array} \tag{7.5}$$

dove i termini noti m_k, $k = 0, 1, \ldots, 2n + 1$, sono dati dalla (7.1).

Si può dimostrare che la soluzione $(a_0, \ldots, a_n, x_0, \ldots, x_n)^T$ del sistema non lineare (7.5) è univocamente determinata (per il caso con $\rho(x) = 1$, cfr. Esempio 4.8.3). La formula (7.2) con coefficienti e nodi soddisfacenti

le (7.4) si dice che ha grado di precisione almeno $2n+1$; se inoltre risulta $E_n(x^{2n+2}) \neq 0$, il grado di precisione è esattamente $2n+1$.

Si noti che se $f(x) \in \Pi_{2n+1}$ allora dalle (7.4) segue $E_n(f(x)) = 0$.

L'errore (7.4) è suscettibile di una rappresentazione generale: sussiste infatti il seguente teorema, enunciato, per semplicità, nel caso di una formula (7.2) con $\rho(x) = 1$ e $a = x_0 < x_1 < \cdots < x_n = b$.

Teorema 7.1.1 (di Peano) *Sia $f(x) \in C^{m+1}([a,b])$ e $J_n(f)$ di grado di precisione m, allora risulta*

$$E_n(f) = \frac{1}{m!} \int_a^b f^{(m+1)}(t) G(t) dt \tag{7.6}$$

essendo $G(t) = E_n(s_m(x-t))$ e

$$s_m(x-t) = \begin{cases} (x-t)^m\,, & t < x, \\ 0\,, & t \geq x. \end{cases}$$

La funzione $G(t)$ dicesi *nucleo di Peano*.

Nel caso in cui $G(t)$ non cambi segno in $[a,b]$, usando la (7.6) e il teorema della media, si ottiene

$$E_n(f) = \frac{1}{m!} f^{(m+1)}(\theta) \int_a^b G(t) dt\,, \quad \theta \in]a,b[\,,$$

da cui si può ricavare $\int_a^b G(t)dt$ (che non dipende da f) ponendo, per esempio, $f(x) = x^{m+1}$; ne segue

$$E_n(f) = \frac{f^{(m+1)}(\theta)}{(m+1)!} E_n(x^{m+1})\,, \quad \theta \in]a,b[\,. \tag{7.7}$$

7.2 Le formule di Newton-Cotes

Se in (7.2) i nodi sono prefissati arbitrariamente (due a due distinti), i pesi sono determinati dalle prime $n+1$ equazioni di (7.5) che formano il sistema lineare

$$V\alpha = \mu \tag{7.8}$$

con

$$V = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_0 & x_1 & \cdots & x_n \\ \cdots & \cdots & \cdots & \cdots \\ x_0^n & x_1^n & \cdots & x_n^n \end{pmatrix},$$

$\alpha^T = (a_0, a_1, \ldots, a_n)$, $\mu^T = (m_0, m_1, \ldots, m_n)$. Poiché V^T è una matrice di Vandermonde, nelle ipotesi attuali è $\det(V) \neq 0$ e quindi α esiste ed è unico. Il grado di precisione è almeno n poiché $E_n(1) = E_n(x) = \cdots = E_n(x^n) = 0$.

Se $\rho(x) = 1$ e i nodi sono fissati in progressione aritmetica di ragione $h = (b-a)/n$, cioè $x_i = x_0 + ih$, $i = 0, 1, \ldots, n$, e quindi con $x_0 = a$ e $x_n = b$, si hanno le *formule di Newton-Cotes*; h si dice il *passo* della formula. I pesi sono definiti in base alla formulazione algebrica precedentemente data e quindi calcolabili risolvendo il sistema (7.8); tuttavia, per una loro rappresentazione esplicita conviene ricorrere ad un approccio interpolatorio.

Al riguardo si esprime $f(x)$ tramite il polinomio di interpolazione di Lagrange (cfr. Teorema 6.2.2), usando i nodi x_i. Cioè si pone

$$f(x) = \sum_{i=0}^{n} l_i(x) f(x_i) + \frac{1}{(n+1)!} f^{(n+1)}(\xi)\pi(x) \,,$$

essendo $l_i(x)$, $i = 0, 1, \ldots, n$, i polinomi fondamentali di Lagrange di grado n, $\xi \in]a, b[$ e $\pi(x) = (x - x_0)(x - x_1) \cdots (x - x_n)$. Ne viene:

$$I(f) = \sum_{i=0}^{n} I(l_i(x)) f(x_i) + \frac{1}{(n+1)!} I\left(f^{(n+1)}(\xi)\pi(x)\right) \,.$$

Si assume quindi $J_n(f) = \sum_{i=0}^{n} a_i f(x_i)$ con

$$a_i = I(l_i(x)) \tag{7.9}$$

e si ha

$$E_n(f) = \frac{1}{(n+1)!} I(f^{(n+1)}(\xi)\pi(x)) \,. \tag{7.10}$$

I pesi delle formule di Newton-Cotes sono stati calcolati per vari valori di n tramite le (7.9); fino a $n = 7$ (otto punti) i pesi sono positivi, mentre per $n > 7$ compaiono pesi negativi e le formule diventano numericamente instabili, cioè la loro capacità di amplificare gli errori di arrotondamento aumenta.

Una caratteristica importante delle formule di Newton-Cotes risiede nel fatto che per esse il nucleo di Peano $G(t)$ non cambia segno in $[a, b]$, per cui, per l'errore, può utilizzarsi la (7.7): m si determina in base alla definizione, mentre il termine $E_n(x^{m+1})$ si calcola facilmente una volta noti i pesi. In tal modo, indicando con $p = n + 1$ il numero dei nodi, l'errore assume la seguente forma caratteristica

$$E_n(f) = \begin{cases} c_p\, h^{p+1} f^{(p)}(\theta), & p \quad \text{pari}, \\ c_p\, h^{p+2} f^{(p+1)}(\theta), & p \quad \text{dispari}, \end{cases}$$

con c_p costante dipendente da p.

Si osservi come la (7.10), invece, non possa mettersi in una forma più semplice, tranne che nel caso $n = 1$, in cui $\pi(x)$ ha segno costante e quindi, applicando il teorema della media, la funzione $f^{(2)}(\xi)$ può essere portata fuori dal segno di integrale.

Notiamo infine che le formule di Newton-Cotes qui definite vengono dette *chiuse*, per distinguerle da formule analoghe dette *aperte* nelle quali si ha $a < x_0 < x_1 < \cdots < x_n < b$.

Di seguito si riportano le prime sette formule chiuse con i relativi errori.
Formula trapezoidale:

$$I(f) = \frac{b-a}{2}\,[f(x_0) + f(x_1)] - \frac{1}{12} h^3 f^{(2)}(\theta).$$

Formula di Simpson:

$$I(f) = \frac{b-a}{6}\,[f(x_0) + 4f(x_1) + f(x_2)] - \frac{1}{90} h^5 f^{(4)}(\theta).$$

Formula dei 3/8 o "pulcherrima":

$$I(f) = \frac{b-a}{8}\,[f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)] - \frac{3}{80} h^5 f^{(4)}(\theta).$$

Formula di Milne-Boole:

$$\begin{aligned} I(f) \;=\; & \frac{b-a}{90}\,[7f(x_0) + 32f(x_1) + 12f(x_2) + 32f(x_3) + 7f(x_4)] \\ & - \frac{8}{945} h^7 f^{(6)}(\theta). \end{aligned}$$

Formula dei sei punti:

$$I(f) = \frac{b-a}{288}\left[19f(x_0)+75f(x_1)+50f(x_2)+50f(x_3)+75f(x_4)\right.$$
$$\left.+19f(x_5)\right]-\frac{275}{12096}h^7 f^{(6)}(\theta).$$

Formula di Weddle:

$$I(f) = \frac{b-a}{840}\left[41f(x_0)+216f(x_1)+27f(x_2)+272f(x_3)+27f(x_4)\right.$$
$$\left.+216f(x_5)+41f(x_6)\right]-\frac{9}{1400}h^9 f^{(8)}(\theta).$$

Formula degli otto punti:

$$I(f) = \frac{b-a}{17280}\left[751f(x_0)+3577f(x_1)+1323f(x_2)+2989f(x_3)\right.$$
$$\left.+2989f(x_4)+1323f(x_5)+3577f(x_6)+751f(x_7)\right]$$
$$-\frac{8183}{518400}h^9 f^{(8)}(\theta).$$

Se, per una formula a $n+1$ punti, il passo di integrazione $h=(b-a)/n$ risulta troppo ampio, si divide $[a,b]$ in m parti uguali, mediante i punti $x_0=a<x_1<\cdots<x_m=b$, e si utilizza la proprietà

$$I(f)=\sum_{i=1}^{m}\int_{x_{i-1}}^{x_i} f(x)dx\,.$$

Si ottengono quindi le cosiddette *formule di Newton-Cotes generalizzate* applicando una stessa formula a $n+1$ punti per ognuno degli m integrali a secondo membro. Si hanno così $nm+1$ nodi con un passo $h=(b-a)/nm$.

Le formule più usate sono la formula trapezoidale ($n=1$) e quella di Simpson ($n=2$) le quali danno luogo alle corrispondenti formule generalizzate:

$$I(f)=\frac{b-a}{2m}\left[f(x_0)+2\sum_{i=1}^{m-1}f(x_i)+f(x_m)\right]-\frac{(b-a)^3}{12m^2}f^{(2)}(\tau)\,; \qquad (7.11)$$

$$I(f) = \frac{b-a}{6m}\left[f(x_0)+4\sum_{i=0}^{m-1}f\left(\frac{x_i+x_{i+1}}{2}\right)+2\sum_{i=1}^{m-1}f(x_i)+f(x_m)\right]$$
$$-\frac{(b-a)^5}{2880m^4}f^{(4)}(\tau).$$

Queste due formule generalizzate presentano il vantaggio di un errore che tende a zero al crescere di m. Si noti che da $h = (b-a)/nm$ segue che gli errori sono rispettivamente dell'ordine di h^2 e di h^4.

7.3 Applicazione della estrapolazione all'integrazione

Si può dimostrare che se $f(x)$ è sufficientemente regolare in $[a,b]$, l'errore della formula generalizzata (7.11) ammette uno sviluppo in serie di potenze pari di h. Più precisamente, scrivendo per semplicità I invece di $I(f)$ e ponendo

$$J_0^{(1)} = \frac{h}{2}\left[f(x_0) + 2\sum_{i=1}^{m-1} f(x_i) + f(x_m)\right], \tag{7.12}$$

si può dimostrare il seguente teorema.

Teorema 7.3.1 *Se $f(x) \in C^{2r+2}([a,b])$, la (7.11) può scriversi*

$$I = J_0^{(1)} + \alpha_1^{(1)}h^2 + \alpha_2^{(1)}h^4 + \cdots + \alpha_r^{(1)}h^{2r} + O(h^{2r+2}) \tag{7.13}$$

dove $\alpha_1^{(1)}, \alpha_2^{(1)}, \ldots, \alpha_r^{(1)}$, non dipendono da h.

La (7.13) oltre a mostrare, come già osservato, che l'errore di $J_0^{(1)}$ è dell'ordine di h^2, consente di ottenere, con un costo computazionale relativamente contenuto, stime di I migliori della (7.12) per mezzo della *tecnica di estrapolazione*.

Scelto l'intero $q > 1$, la formula trapezoidale generalizzata, con il passo $h/q = (b-a)/mq$, fornisce la stima di I,

$$J_1^{(1)} = \frac{h}{2q}\left[f(x_0) + 2\sum_{i=1}^{mq-1} f(x_i) + f(x_{mq})\right],$$

dove ora i nodi sono $x_i = a + i\frac{h}{q}$, $i = 0, 1, \ldots, mq$. D'altra parte, per il Teorema 7.3.1, si può scrivere

$$I = J_1^{(1)} + \alpha_1^{(1)}\left(\frac{h}{q}\right)^2 + \alpha_2^{(1)}\left(\frac{h}{q}\right)^4 + \cdots + \alpha_r^{(1)}\left(\frac{h}{q}\right)^{2r} + O(h^{2r+2}). \tag{7.14}$$

Eliminando il termine $\alpha_1^{(1)}h^2$ fra la (7.13) e la (7.14) si ottiene

$$I = J_0^{(2)} + \alpha_2^{(2)}h^4 + \cdots + \alpha_r^{(2)}h^{2r} + O(h^{2r+2}) \,, \tag{7.15}$$

dove si è posto

$$J_0^{(2)} = \frac{q^2 J_1^{(1)} - J_0^{(1)}}{q^2 - 1} \tag{7.16}$$

e dove i nuovi coefficienti, $\alpha_2^{(2)}, \ldots, \alpha_r^{(2)}$, non dipendono da h.

La (7.15) mostra che l'errore di $J_0^{(2)}$ è dell'ordine di h^4, cioè è una stima di I migliore di $J_0^{(1)}$ e $J_1^{(1)}$, mentre la (7.16) implica un costo computazionale di solo 3 operazioni essenziali; per contro, una stima di I più accurata di $J_1^{(1)}$, mediante la formula trapezoidale, implica necessariamente l'uso di un passo minore di h/q e quindi un costo computazionale più elevato di quello richiesto dalla (7.16).

Il procedimento di estrapolazione può essere ripetuto: ad ogni applicazione, l'ordine dell'errore della nuova stima di I aumenta di due. Infatti, con il passo h/q^2, dalla formula trapezoidale generalizzata, si ricava $J_2^{(1)}$ mentre il Teorema 7.3.1 fornisce

$$I = J_2^{(1)} + \alpha_1^{(1)} \left(\frac{h}{q^2}\right)^2 + \alpha_2^{(1)} \left(\frac{h}{q^2}\right)^4 + \cdots + \alpha_r^{(1)} \left(\frac{h}{q^2}\right)^{2r} + O(h^{2r+2});$$

eliminando $\alpha_1^{(1)}(h/q)^2$ fra questa relazione e la (7.14) si ha

$$I = J_1^{(2)} + \alpha_2^{(2)} \left(\frac{h}{q}\right)^4 + \cdots + \alpha_r^{(2)} \left(\frac{h}{q}\right)^{2r} + O(h^{2r+2}) \tag{7.17}$$

dove

$$J_1^{(2)} = \frac{q^2 J_2^{(1)} - J_1^{(1)}}{q^2 - 1} \,;$$

eliminando ora $\alpha_2^{(2)}h^4$ fra la (7.15) e la (7.17) si può scrivere

$$I = J_0^{(3)} + \alpha_3^{(3)}h^6 + \cdots + \alpha_r^{(3)}h^{2r} + O(h^{2r+2})$$

dove

$$J_0^{(3)} = \frac{q^4 J_1^{(2)} - J_0^{(2)}}{q^4 - 1} \,,$$

con $\alpha_3^{(3)}, \ldots, \alpha_r^{(3)}$ indipendenti da h.

Nella sua formulazione più generale, la tecnica di estrapolazione consiste, fissato l'intero $N < r$, nel calcolare, mediante la formula trapezoidale generalizzata, i valori

$$J_0^{(1)}, J_1^{(1)}, \ldots, J_N^{(1)}, \tag{7.18}$$

corrispondenti ai passi $h, h/q, \ldots, h/q^N$ (questa è la parte più costosa del metodo), quindi, a partire dai valori (7.18), si costruiscono nuove approssimazioni di I in base allo schema

$$J_i^{(k+1)} = \frac{q^{2k} J_{i+1}^{(k)} - J_i^{(k)}}{q^{2k} - 1}, \quad k = 1, 2, \ldots, N, \tag{7.19}$$

dove, per ciascun valore di k, l'indice i assume i valori $0, 1, \ldots, N - k$.

Dalle considerazioni precedenti si deduce che $J_i^{(k+1)}$ presenta un errore dell'ordine di $(h/q^i)^{2(k+1)}$.

È importante rilevare inoltre che la conoscenza di due delle approssimazioni (7.18), consente di stimare l'errore di entrambe, senza la necessità di conoscere le derivate di $f(x)$.

Per esempio, supposto di aver calcolato $J_0^{(1)}$ e $J_1^{(1)}$, dalle (7.13) e (7.14) si ha

$$\begin{aligned} I &\simeq J_0^{(1)} + \alpha_1^{(1)} h^2 \,, \\ I &\simeq J_1^{(1)} + \alpha_1^{(1)} (h/q)^2 \,. \end{aligned}$$

Evidentemente $\alpha_1^{(1)} h^2$ e $\alpha_1^{(1)} (h/q)^2$ forniscono stime dell'errore rispettivamente di $J_0^{(1)}$ e di $J_1^{(1)}$, non appena si calcoli il valore di α_1. A tale scopo, eliminando I fra le due relazioni, si ottiene

$$\alpha_1^{(1)} \simeq \frac{J_0^{(1)} - J_1^{(1)}}{\left(\frac{h}{q}\right)^2 - h^2} \,.$$

Comunemente vengono usati i valori $q = 2$ o $q = 4$ (cfr. Esempio 7.6.3).

Le approssimazioni di I ottenute col processo di estrapolazione, note come *formule di Romberg*, costituiscono una speciale applicazione di un procedimento generale, detto *estrapolazione di Richardson*. Tale procedimento può essere impiegato in tutti quei casi in cui una grandezza T si possa approssimare con un valore $\tau_0^{(1)}$ dipendente da un parametro h, e l'errore $T - \tau_0^{(1)}$ sia sviluppabile in serie di potenze di h, cioè si abbia

$$T = \tau_0^{(1)} + \beta_1^{(1)} h^{r_1} + \beta_2^{(1)} h^{r_2} + \cdots$$

con $\beta_1^{(1)}, \beta_2^{(1)}, \ldots$, non dipendenti da h e $0 < r_1 < r_2 < \cdots$ interi. La costruzione di approssimazioni di T più accurate di $\tau_0^{(1)}$ si ottiene con uno schema analogo a quello definito dalla (7.19).

7.4 Formule di quadratura di tipo gaussiano

Si premettono brevemente alcune definizioni e proprietà di una particolare classe di polinomi.

Sia $\rho(x)$ una funzione peso che verifichi le ipotesi fatte in 7.1.

Si indichi con Π lo spazio vettoriale dei polinomi algebrici a coefficienti reali e sia $[a, b]$ un intervallo, non necessariamente limitato. Per ogni coppia $r(x), s(x) \in \Pi$ si consideri il prodotto scalare

$$< r, s >=< s, r >= I(\rho rs) = \int_a^b \rho(x) r(x) s(x) dx. \tag{7.20}$$

Si definisce quindi la classe dei *polinomi ortogonali* Π^* come l'insieme dei polinomi algebrici, a coefficienti reali, ortogonali rispetto al prodotto scalare (7.20), cioè

$$\Pi^* = \{q_i(x) \mid \text{grado}(q_i) = i;\ < q_i, q_j >= h_i \delta_{ij}\ ; i, j = 0, 1, 2, ..\}\,.$$

I numeri positivi h_i sono le *costanti di normalizzazione.*
Dati $[a, b]$ e $\rho(x)$, gli elementi di Π^* restano definiti a meno di una costante moltiplicativa e costituiscono una base per Π; per essi valgono le seguenti proprietà:

$$< p, q_n >= 0 \quad \forall p \in \Pi_{n-1}; \tag{7.21}$$

$$q_n(x) \quad \text{ha } n \text{ zeri reali e distinti in }]a, b[. \tag{7.22}$$

Come osservato in 7.1, fissati $\rho(x)$, $[a, b]$ ed n, risulta univocamente determinata la formula di quadratura di grado di precisione almeno $2n + 1$

$$I(\rho f) = \int_a^b \rho(x) f(x) dx \simeq \sum_{i=0}^{n} a_i f(x_i) = J_n(f)\,,$$

che propriamente dicesi *formula di quadratura gaussiana.*

Teorema 7.4.1 *Dati $[a,b]$ e $\rho(x)$, sia $f(x) \in C^{2n+2}(]a,b[)$ e $J_n(f)$ una formula di quadratura gaussiana; allora i nodi $x_0, x_1, \ldots, x_n$ sono gli zeri di $q_{n+1}(x) \in \Pi^*$, i pesi sono positivi e dati da*

$$a_i = I(\rho l_i^2), \quad i = 0, 1, \ldots, n,$$

dove $l_i(x)$ sono i polinomi fondamentali di Lagrange relativi ai detti nodi e il grado di precisione è esattamente $2n+1$, essendo

$$E_n(f) = K_n \frac{f^{(2n+2)}(\theta)}{(2n+2)!}, \quad K_n > 0\,, \quad \theta \in]a,b[\,, \tag{7.23}$$

l'errore della formula.

DIMOSTRAZIONE. Siano $x_0, x_1, \ldots, x_n$, gli zeri di $q_{n+1}(x)$; relativamente a questi punti si considerano i polinomi fondamentali di Lagrange (cfr. 6.2) di grado n

$$l_i(x) = (x - x_0) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)/\alpha_i$$

con $\alpha_i = (x_i - x_0) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)$, $i = 0, 1, \ldots, n$, e le funzioni fondamentali dell'interpolazione di Hermite (cfr. 6.3) di grado $2n+1$ di prima e di seconda specie

$$h_{0i}(x) = [1 - 2l_i'(x_i)(x - x_i)]l_i^2(x), \quad i = 0, 1, \ldots, n,$$

$$h_{1i}(x) = (x - x_i)l_i^2(x), \quad i = 0, 1, \ldots, n.$$

Evidentemente $q_{n+1}(x)$ e $\pi(x) = (x - x_0) \cdots (x - x_n)$ differiscono per una costante moltiplicativa c_{n+1}, per cui si può scrivere

$$q_{n+1}(x) = c_{n+1}\pi(x) = c_{n+1}\alpha_i(x - x_i)l_i(x)\,.$$

Pertanto, per quanto riguarda le funzioni fondamentali di Hermite di prima specie, risulta

$$I(\rho h_{0i}) = I(\rho l_i^2) - \frac{2l_i'(x_i)}{c_{n+1}\alpha_i} I(\rho q_{n+1} l_i), \quad i = 0, 1, \ldots, n,$$

avendosi $\alpha_i \neq 0$ per la proprietà (7.22).
D'altra parte si ha $I(\rho q_{n+1} l_i) =< q_{n+1}, l_i >= 0$ per la (7.21), in quanto $l_i(x) \in \Pi_n$, per cui

$$I(\rho h_{0i}) = I(\rho l_i^2)\,, \quad i = 0, 1, \ldots, n. \tag{7.24}$$

Per le funzioni di seconda specie risulta

$$I(\rho h_{1i}) = \frac{1}{c_{n+1}\alpha_i} I(\rho q_{n+1} l_i) = 0\,, \quad i = 0, 1, \ldots, n. \tag{7.25}$$

Ricordando che (cfr. Teorema 6.3.3)

$$f(x) = \sum_{i=0}^{n} h_{0i}(x) f(x_i) + \sum_{i=0}^{n} h_{1i}(x) f'(x_i) + \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\pi^2(x),$$

con $\xi \in]a, b[$, e tenendo conto dei risultati (7.24) e (7.25) nonché della proprietà additiva degli integrali, si ha

$$I(\rho f) = \sum_{i=0}^{n} I(\rho l_i^2) f(x_i) + I\left(\frac{f^{(2n+2)}(\xi)}{(2n+2)!}\rho\pi^2\right).$$

Poiché $\rho(x)\pi^2(x)$ non cambia segno nell'intervallo di integrazione, nel secondo termine a secondo membro può applicarsi il teorema della media e scrivere

$$I\left(\frac{f^{(2n+2)}(\xi)}{(2n+2)!}\rho\pi^2\right) = \frac{f^{(2n+2)}(\theta)}{(2n+2)!} I(\rho\pi^2), \quad \theta \in]a, b[.$$

Ponendo quindi

$$a_i = I(\rho l_i^2), \quad i = 0, 1, \ldots, n,$$

$$K_n = I(\rho\pi^2), \quad n = 0, 1, \ldots,$$

$$E_n(f) = K_n \frac{f^{(2n+2)}(\theta)}{(2n+2)!},$$

si ha infine

$$I(\rho f) = \sum_{i=0}^{n} a_i f(x_i) + E_n(f)$$

dove $a_i > 0$, $i = 0, 1, \ldots, n$, $K_n > 0$, $n = 0, 1, \ldots$. □

Le formule aventi i pesi e i nodi definiti come nel Teorema 7.4.1 coincidono con quelle della formulazione algebrica data in 7.1. L'ipotesi di derivabilità, richiesta dal teorema, consente quindi una rappresentazione dell'errore nella forma (7.23) ma non è necessaria per la definizione della formula.

Nella tavola che segue si riportano le caratteristiche di alcuni polinomi ortogonali fra i più usati, che danno luogo alle corrispondenti formule di quadratura gaussiane da cui prendono il nome (nella tavola il simbolo π sta per 3.141592...).

Intervallo	Peso	Classificazione	K_n
$[-1,1]$	$1/\sqrt{1-x^2}$	Chebyshev 1^a specie	$\pi/2^{2n+1}$
$[-1,1]$	$\sqrt{1-x^2}$	Chebyshev 2^a specie	$\pi/2^{2n+3}$
$[-1,1]$	1	Legendre	$\frac{2^{2n+3}[(n+1)!]^4}{(2n+3)[(2n+2)!]^2}$
$[0,+\infty[$	e^{-x}	Laguerre	$[(n+1)!]^2$
$]-\infty,+\infty[$	e^{-x^2}	Hermite	$(n+1)!\sqrt{\pi}/2^{n+1}$

Osservazione 7.4.1 Nodi e pesi delle formule gaussiane sono numeri irrazionali e sono stati calcolati in precisione multipla per vari valori di n. Essi sono inseriti nei principali programmi di calcolo per l'integrazione approssimata. L'uso di tali formule per via automatica non presenta, in genere, particolari difficoltà.
Per completezza, nel paragrafo 7.6.3 sono riportate la formule di ricorrenza per la generazione dei polinomi ortogonali qui menzionati, nonché i corrispondenti nodi e pesi per alcuni valori di n.

Osservazione 7.4.2 Per ragioni di convenienza, per gli integrali estesi ad un intervallo limitato $[a,b]$, si usano polinomi ortogonali definiti in $[-1,1]$. In effetti ogni intervallo di integrazione $a \le t \le b$ può ricondursi all'intervallo $-1 \le x \le 1$ con la trasformazione $t = \frac{b-a}{2}x + \frac{b+a}{2}$ e la funzione $\rho(x)$ può essere comunque introdotta. Risulta infatti

$$\int_a^b g(t)dt = \int_{-1}^1 \rho(x)f(x)dx$$

ove si assuma $f(x) = \frac{b-a}{2\rho(x)} g\left(\frac{b-a}{2}x + \frac{b+a}{2}\right)$.
Si noti poi che l'integrazione gaussiana, basata sui polinomi di Laguerre e di Hermite, permette di approssimare integrali su intervalli non limitati (purché l'integrale esista finito).

Osservazione 7.4.3 La positività dei pesi consente di dimostrare, sotto ipotesi molto generali, la convergenza di $J_n(f)$.
Più precisamente, nel caso di intervalli limitati $[a,b]$ e per formule di grado $2n+1$, la semplice continuità di $f(x)$ è condizione sufficiente affinché

$$\lim_{n\to\infty} J_n(f) = I(\rho f)\,.$$

Pertanto l'errore $E_n(f)$ tende a zero per $n \to \infty$ anche nel caso che $f(x)$ non sia derivabile e quindi $E_n(f)$ non possa esprimersi nella forma (7.23).

Inoltre, nel caso di intervalli di integrazione non limitati, vale il seguente teorema relativo alle formula di Gauss-Laguerre.

Teorema 7.4.2 *Se esiste un numero reale $\alpha > 1$ ed un x^* tali che $\mid f(x) \mid \leq e^x/x^\alpha$ per $x > x^*$ e se $f(x) \in C^0([0, +\infty[)$, allora per la formula di quadratura di Gauss-Laguerre risulta*

$$\lim_{n\to\infty} J_n(f) = I(e^{-x} f) \, .$$

7.5 Integrazione in più dimensioni

L'approssimazione di integrali multipli, della forma cioè

$$I(\rho f) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} \cdots \int_{a_r}^{b_r} \rho(x_1, \ldots, x_r) f(x_1, \ldots, x_r) dx_1 \ldots dx_r$$

con $\rho(x_1, \ldots, x_r) \geq 0$ per $x_i \in [a_i, b_i]$, $i = 1, \ldots, r$, presenta difficoltà e costi computazionali assai più elevati rispetto al caso monodimensionale.

Ci si limita qui, per semplicità, ad accennare il problema nel caso bidimensionale con peso unitario e dominio di integrazione rettangolare.

Sia $f(x, y)$ integrabile su $[a, b] \times [c, d]$ con $x \in [a, b]$, $y \in [c, d]$.

Sia J_k una formula di quadratura monodimensionale con $k + 1$ nodi $x_i \in [a, b]$ e pesi a_i, $i = 0, 1, \ldots, k$, e J_n una con $n + 1$ nodi $y_j \in [c, d]$ e pesi b_j, $j = 0, 1, \ldots, n$.

In base al teorema di riduzione degli integrali doppi, risulta evidentemente

$$\begin{aligned} I(f) &= \int_a^b \int_c^d f(x, y) dx dy = \int_a^b \left[\int_c^d f(x, y) dy \right] dx \\ &\simeq \int_a^b \sum_{j=0}^{n} b_j f(x, y_j) dx \simeq \sum_{i=0}^{k} \sum_{j=0}^{n} a_i b_j f(x_i, y_j) \qquad (7.26) \\ &= J_k J_n(f). \end{aligned}$$

$J_k J_n(f)$ dicesi *formula di cubatura* e l'errore è

$$E_{k,n}(f) = I(f) - J_k J_n(f) \, .$$

Quindi, ad esempio, se J_2 è la formula di quadratura di Simpson si ha

$$\begin{aligned} J_2J_2(f) &= \frac{(b-a)(d-c)}{36}\,[f(x_0,y_0)+4f(x_0,y_1)+f(x_0,y_2) \\ &\quad +4f(x_1,y_0)+16f(x_1,y_1)+4f(x_1,y_2)+f(x_2,y_0) \\ &\quad +4f(x_2,y_1)+f(x_2,y_2)]\,. \end{aligned}$$

Se inoltre $f(x,y)$ è sufficientemente regolare si può dimostrare che l'errore è dato da

$$E_{2,2}(f) = -\frac{hk}{45}\left[h^4\frac{\partial^4 f(\bar{x},\bar{y})}{\partial x^4} + k^4\frac{\partial^4 f(\bar{\bar{x}},\bar{\bar{y}})}{\partial y^4}\right]$$

con $\bar{x},\bar{\bar{x}} \in]a,b[$, $\bar{y},\bar{\bar{y}} \in]c,d[$, e $h=(b-a)/2$, $k=(d-c)/2$.

Formule di cubatura possono essere costruite anche direttamente. Per esempio, considerati i punti $P_r \in [a,b]\times[c,d]$, $r=0,1,\ldots,N$, si cercano formule del tipo

$$I(f) \simeq J_N(f) = \sum_{r=0}^{N} c_r f(P_r). \tag{7.27}$$

I nodi P_r ed i pesi c_r sono determinati in modo che la (7.27) sia esatta quando $f(P)$ è un polinomio di grado non superiore a g nelle due variabili x e y, cioè in modo che sia

$$E_N(x^\alpha y^\beta) = 0, \quad 0 \le \alpha+\beta \le g\,,$$

dove $E_N(f) = I(f) - J_N(f)$.

Si osservi tuttavia che se J_n è una formula di quadratura con grado di precisione m, ossia se risulta

$$J_n(f) = \sum_{i=0}^{n} a_i x_i^\alpha = I(x^\alpha), \quad \alpha = 0,1,\ldots,m,$$

allora, facendo riferimento, per semplicità, ad un dominio di integrazione quadrato, si verifica facilmente che

$$J_nJ_n(x^\alpha y^\beta) = \sum_{i=0}^{n}\sum_{j=0}^{n} a_i a_j x_i^\alpha x_j^\beta = I(x^\alpha y^\beta), \quad 0 \le \alpha,\beta \le m. \tag{7.28}$$

La $J_nJ_n(f)$, detta *formula del prodotto cartesiano*, è della forma (7.27) ove si usino gli $N+1=(n+1)^2$ nodi $P_{ij}=(x_i,x_j)$ con i pesi a_ia_j e risulta esatta per polinomi fino al grado $2m$.

Ad esempio, nel caso $[a,b]=[-1,1]$ e $n=1$, la formula di quadratura di Gauss-Legendre J_1 ha nodi e pesi

$$-x_0 = x_1 = \frac{\sqrt{3}}{3}, \quad a_0 = a_1 = 1$$

e grado di precisione $m=3$.

Con $[c,d]=[-1,1]$ la formula di cubatura $J_1 J_1$ ha $N+1=2^2=4$ nodi dati da

$$\left(-\frac{\sqrt{3}}{3},-\frac{\sqrt{3}}{3}\right), \left(\frac{\sqrt{3}}{3},-\frac{\sqrt{3}}{3}\right), \left(-\frac{\sqrt{3}}{3},\frac{\sqrt{3}}{3}\right), \left(\frac{\sqrt{3}}{3},\frac{\sqrt{3}}{3}\right)$$

con coefficienti unitari; essa inoltre risulta esatta per ogni monomio del tipo $x^\alpha y^\beta$ con $0 \le \alpha, \beta \le 3$ ed è esplicitamente data da

$$\begin{aligned} I(f) &= \int_{-1}^{1}\int_{-1}^{1} f(x,y)dxdy \\ &\simeq f\left(-\frac{\sqrt{3}}{3},-\frac{\sqrt{3}}{3}\right) + f\left(\frac{\sqrt{3}}{3},-\frac{\sqrt{3}}{3}\right) + f\left(-\frac{\sqrt{3}}{3},\frac{\sqrt{3}}{3}\right) \\ &\quad + f\left(\frac{\sqrt{3}}{3},\frac{\sqrt{3}}{3}\right). \end{aligned}$$

7.6 Complementi ed esempi

7.6.1 Costruzione di formule e calcolo dell'errore

Un approccio algebrico analogo a quello che conduce al sistema (7.5) consente la costruzione di formule di quadratura più generali di quelle viste finora, potendosi utilizzare anche i valori di una o più derivate come nell'esempio seguente.

Esempio 7.6.1 Si vogliono calcolare i pesi e i nodi della formula

$$I(f) = \int_0^1 f(x)dx \simeq a_0 f(0) + \frac{1}{6}f'(x_1) + \frac{1}{6}f'(x_2) + a_1 f(1) = J(f)$$

in modo che sia massimo il grado di precisione.

Imponendo $E(x^r) = I(x^r) - J(x^r) = 0$, $r = 0, 1, 2, 3$, si trova subito $a_0 = 5/6$, $a_1 = 1/6$, ed inoltre si ha

$$\begin{aligned} x_1 &+ x_2 &= 1/2 \\ x_1^2 &+ x_2^2 &= 1/6 \end{aligned}$$

da cui $x_1 = (3 - \sqrt{3})/12$ e $x_2 = (3 + \sqrt{3})/12$.

Poiché risulta $E(x^4) = -1/120$, la formula ha grado di precisione $m = 3$. □

Il Teorema 7.1.1 indica una procedura generale per calcolare l'errore di una formula di quadratura quando $f(x)$ sia sufficientemente regolare.

Esempio 7.6.2 Si vuole calcolare l'errore $E_2(f) = I(f) - J_2(f)$ della formula di Simpson

$$I(f) = \int_{-1}^{1} f(x)dx \simeq \frac{1}{3}f(-1) + \frac{4}{3}f(0) + \frac{1}{3}f(1) = J_2(f)\,.$$

Ponendo $f(x) = x^r$, si trova $E_2(x^r) = 0$, $r = 0, 1, 2, 3$, ed $E_2(x^4) = -4/15$, per cui il grado di precisione è $m = 3$.

Il nucleo di Peano è dato da

$$\begin{aligned} G(t) &= \int_t^1 (x-t)^3 dx - \frac{4}{3}(0-t)^3 - \frac{1}{3}(1-t)^3 \\ &= \frac{1}{4}t^4 + \frac{2}{3}t^3 + \frac{1}{2}t^2 - \frac{1}{12} \qquad \text{per} \quad -1 \le t < 0; \end{aligned}$$

$$\begin{aligned} G(t) &= \int_t^1 (x-t)^3 dx - \frac{1}{3}(1-t)^3 \\ &= \frac{1}{4}t^4 - \frac{2}{3}t^3 + \frac{1}{2}t^2 - \frac{1}{12} \qquad \text{per} \;\; 0 \le t \le 1. \end{aligned}$$

Poiché $G(t)$ ha segno costante in $]-1, 1[$, dalla (7.7) si ricava

$$E_2(f) = -\frac{1}{90}f^{(4)}(\theta), \quad \theta \in]-1, 1[.$$

□

Esempio 7.6.3 Si consideri

$$I(f) = \int_0^1 \frac{dx}{1+x}\,,$$

il cui valore esatto è $I(f) = \log 2 = 0.69314718\ldots$.

Posto, nella (7.18), $N = 2$, si calcolano, $J_0^{(1)}$, $J_1^{(1)}$ e $J_2^{(1)}$ con $h = 1/2$ e $q = 2$. Applicando la (7.19), per $k = 1, 2$, si ottengono i seguenti valori estrapolati:

$$\begin{aligned}
&J_0^{(1)} = 0.70833333\ldots \\
&J_1^{(1)} = 0.69702380\ldots \quad J_0^{(2)} = 0.69325396\ldots \\
&J_2^{(1)} = 0.69412185\ldots \quad J_1^{(2)} = 0.69315453\ldots \quad J_0^{(3)} = 0.69314790\ldots\,.
\end{aligned}$$

Si noti che i valori $J_0^{(2)}$ e $J_0^{(3)}$ sono più accurati di quelli aventi l'indice superiore più piccolo (in particolare $J_0^{(2)}$ è una approssimazione migliore di $J_2^{(1)}$ che ha un costo computazionale maggiore).

Come osservato in 7.3, una stima dell'errore di $J_1^{(1)}$ basata sui valori $J_0^{(1)}$ e $J_1^{(1)}$ può ottenersi calcolando

$$\alpha_1^{(1)} \simeq \frac{J_1^{(1)} - J_0^{(1)}}{h^2 - h^2/4} \simeq -0.0603175$$

da cui

$$E(f) = I - J_1^{(1)} \simeq \alpha_1^{(1)} \frac{h^2}{4} \simeq -0.00377.$$

In modo analogo si può ottenere una stima dell'errore di $J_2^{(1)}$ utilizzando la coppia $J_2^{(1)}$, $J_0^{(1)}$ oppure $J_2^{(1)}$, $J_1^{(1)}$. □

Esempio 7.6.4 Si vuole calcolare il numero minimo $n + 1$ di nodi per approssimare, con una formula di quadratura di tipo gaussiano, l'integrale

$$I(f) = \int_0^1 \frac{\log(x+1)}{\sqrt{x(x+1)}} dx$$

in modo che risulti $| E_n | \leq 10^{-5}$.

Si trasforma l'intervallo $]0,1[$ nell'intervallo $]-1,1[$ con la sostituzione $t = 2x - 1$. Si ha quindi

$$\begin{aligned} I(f) &= \int_0^1 \frac{\log(x+1)}{\sqrt{x(x+1)}} dx \\ &= \int_{-1}^1 \log\left(\frac{t+3}{2}\right) \sqrt{\frac{1-t}{3+t}} \frac{1}{\sqrt{1-t^2}} dt \\ &= \sqrt{\frac{1-\tau}{3+\tau}} \int_{-1}^1 \log\left(\frac{t+3}{2}\right) \frac{1}{\sqrt{1-t^2}} dt \ , \ \tau \in]-1,1[\ , \end{aligned}$$

dove si è potuto applicare il teorema della media in quanto le funzioni $g(t) = \log\left(\frac{t+3}{2}\right)$ e $\rho(t) = \frac{1}{\sqrt{1-t^2}}$ non cambiano segno in $]-1,1[$. Posto $h(\tau) = \sqrt{\frac{1-\tau}{3+\tau}}$, si può considerare una formula di quadratura di Gauss-Chebyshev di prima specie e scrivere

$$I(f) = h(\tau) I(\rho g) = h(\tau) \left[J_n(g) + K_n \frac{g^{(2n+2)}(\theta)}{(2n+2)!} \right]$$

con $K_n = \pi / 2^{2n+1}$.

Pertanto $I(f) \simeq h(\tau) J_n(g)$ e inoltre, avendosi $0 \leq h(\tau) \leq 1$ e $\mid g^{(k)}(t) \mid \leq \frac{(k-1)!}{2^k}$, $k = 0, 1, \ldots$, si può verificare che, per l'errore, risulta

$$\mid E_n \mid = \left| K_n h(\tau) \frac{g^{(2n+2)}(\theta)}{(2n+2)!} \right| \leq 6 \times 10^{-6}$$

con $n = 3$, ossia con 4 nodi. □

Nell'integrazione gaussiana la scelta del peso, e quindi della formula, può risultare determinante ai fini dell'errore. Si consideri, ad esempio, l'integrale

$$I(f) = \int_{-1}^1 x^2 \sqrt{1-x^2} dx$$

il cui valore esatto è $\pi/8 = 0.39269908\ldots$.

Con una formula a due nodi di Gauss-Chebyshev di seconda specie ($\rho(x) = \sqrt{1-x^2}$), si ottiene

$$I(f) = I\left(\sqrt{1-x^2}\, x^2\right) = J_1(x^2) \, .$$

Utilizzando una formula a tre nodi di Gauss-Chebyshev di prima specie ($\rho(x) = 1/\sqrt{1-x^2}$), risulta

$$I(f) = I\left(\frac{1}{\sqrt{1-x^2}}\left(x^2[1-x^2]\right)\right) = J_2\left(x^2(1-x^2)\right) .$$

Quindi gli errori delle precedenti due formule sono nulli.

Per contro, con una formula di Gauss-Legendre con quattro nodi, si ha

$$I(f) \simeq J_3\left(x^2\sqrt{1-x^2}\right) = 0.40405278\ldots ,$$

con un errore $E_3(f) \simeq -1.1 \times 10^{-2}$.

Esempio 7.6.5 Si vogliono calcolare i pesi della formula

$$\begin{aligned}\int_a^b\int_a^b\int_a^b f(x,y,z)dxdydz \;&\simeq\; c_0 f(a,a,a) + c_1 f(b,0,0)\\ &\quad + c_2 f(0,b,0) + c_3 f(0,0,b) .\end{aligned}$$

Allo scopo si impone che la formula sia esatta per $f(x,y,z) = 1, x, y, z$. Ne viene il sistema

$$\begin{array}{lllllllll} c_0 & + & c_1 & + & c_2 & + & c_3 & = & (b-a)^3 \\ c_0 a & + & c_1 b & & & & & = & \frac{1}{2}(b-a)^3(b+a) \\ c_0 a & & & + & c_2 b & & & = & \frac{1}{2}(b-a)^3(b+a) \\ c_0 a & & & & & + & c_3 b & = & \frac{1}{2}(b-a)^3(b+a) \end{array}$$

che ha soluzione solo se $b \neq 3a$; in tal caso si ottiene

$$\begin{aligned} c_0 &= \tfrac{(b-a)^3(-b-3a)}{2(b-3a)} , \\ c_i &= \tfrac{(b-a)^4}{2(b-3a)} , \qquad i = 1,2,3 . \end{aligned}$$

□

7.6.2 Integrali impropri

Le formule viste per l'integrazione approssimata non sempre sono adatte per essere utilizzate direttamente per il calcolo di un integrale improprio.

Per esempio, nel caso di singolarità della funzione integranda, la presenza di punti singolari crea difficoltà nella scelta dei nodi e può compromettere la convergenza della formula.

Talvolta la singolarità è facilmente rimuovibile con una semplice sostituzione come nell'integrale

$$I(f) = \int_0^1 \frac{dx}{x^{1/\alpha}}\,, \quad \alpha \geq 2,$$

che, con la posizione $x = t^\alpha$, diviene

$$I(f) = \alpha \int_0^1 t^{\alpha-2} dt\,.$$

In altri casi si può ricorrere ad una formula gaussiana opportunamente pesata: l'integrale

$$I(f) = \int_{-1}^1 \frac{e^x}{\sqrt{1-x}} dx$$

può scriversi

$$I(f) = I\left(\frac{1}{\sqrt{1-x^2}} g(x)\right)$$

con $g(x) = e^x\sqrt{1+x}$ e quindi essere approssimato con una formula di quadratura di Gauss-Chebyshev di prima specie.

Altri due metodi sono proposti nei due esempi che seguono.

Esempio 7.6.6 Si considera l'integrale

$$I(f) = \int_0^1 \frac{e^{-x}}{\sqrt{1-x}} dx\,.$$

Con una integrazione per parti, scegliendo e^{-x} come fattore finito, si ottiene

$$I(f) = 2 - 2I(g) \tag{7.29}$$

dove si è posto $g(x) = e^{-x}\sqrt{1-x} \in C^0([0,1])$.

È da notare, tuttavia, che $g(x)$ non è derivabile per $x = 1$ per cui la stima dell'errore, per esempio mediante la (7.6) o la (7.23), può essere difficoltosa. Per ovviare a questo inconveniente si può procedere a due ulteriori integrazioni per parti, applicate a $I(g)$, portando la (7.29) nella forma

$$I(f) = \frac{18}{15} - \frac{8}{15} I(s)\,,$$

dove $s(x) = e^{-x}\sqrt{(1-x)^5} \in C^2([0,1])$.

L'approssimazione di $I(s)$ mediante la formula generalizzata (7.11), con $m = 8$, fornisce il seguente risultato

$$I(s) \simeq 0.231851$$

con un errore assoluto, in modulo, inferiore a 10^{-4}. □

Gli integrali della forma,

$$I(f) = \int_a^b \frac{g(x)}{(x-a)^\beta} dx$$

con $0 < \beta < 1$ e $g(x)$ regolare in $[a, b]$, possono essere affrontati sottraendo dal numeratore uno o più termini dello sviluppo di Taylor di $g(x)$ con punto iniziale $x = a$; questa tecnica va sotto il nome di *metodo della sottrazione della singolarità* .

Esempio 7.6.7 L'integrale

$$I(f) = \int_0^1 \frac{e^x}{x^\beta} dx$$

con $0 < \beta < 1$ può scriversi

$$I(f) = I(r) + I(s)$$

con $r(x) = (e^x - 1)/x^\beta$ e $s(x) = 1/x^\beta$.

Ponendo $r(0) = 0$, risulta, per prolungamento, $r(x) \in C^0([0,1])$ in quanto $\lim_{x\to 0} r(x) = 0$ e inoltre si ha $I(s) = 1/(1-\beta)$.

Si noti tuttavia che $r(x)$ non è della classe $C^1([0,1])$: se si richiede per $r(x)$ una maggiore regolarità conviene porre

$$r(x) = \frac{e^x - 1 - x}{x^\beta}, \quad s(x) = \frac{1}{x^\beta} + \frac{1}{x^{\beta-1}},$$

dove al numeratore di $r(x)$ si è sottratto un ulteriore termine dello sviluppo in serie di e^x.
In tal caso si ha $r(x) \in C^1([0,1])$ e $I(s) = (3-2\beta)/[(1-\beta)(2-\beta)]$. □

Un altro tipo di integrale improprio si ha quando l'intervallo di integrazione non è limitato. Possono allora risultare utili le formule gaussiane di Laguerre e di Hermite.

Esempio 7.6.8 Sia

$$I(f) = \int_0^\infty \sin(x^2)dx$$

il cui valore esatto è $\frac{1}{2}\sqrt{\frac{\pi}{2}} = 0.626657\ldots$.

Con evidenti passaggi può scriversi

$$I(f) = \frac{1}{2}\int_{-\infty}^{+\infty} e^{-x^2} g(x)dx \cong \frac{1}{2}J_n(g)$$

dove $g(x) = e^{x^2}\sin(x^2)$ e J_n è una formula di quadratura di Gauss-Hermite. Eseguendo i calcoli, con $n = 4$, si trova

$$\frac{1}{2}J_4(g) = 0.626868\ldots\ .$$

□

Alternativamente si può effettuare un troncamento dell'intervallo non limitato, avendo cura di stimare opportunamente la parte tralasciata, come nell'esempio che segue.

Esempio 7.6.9 Si consideri

$$I(f) = \int_0^\infty \frac{\cos x}{\cosh x}dx$$

il cui valore esatto è $\frac{\pi}{2}\mathrm{sech}\frac{\pi}{2} = 0.6260201\ldots$.

Può scriversi

$$I(f) = I^{(1)}(f) + I^{(2)}(f)$$

con

$$I^{(1)}(f) = \int_0^b \frac{\cos x}{\cosh x}dx, \quad I^{(2)}(f) = \int_b^\infty \frac{\cos x}{\cosh x}dx,$$

e si ha

$$\mid I^{(2)}(f) \mid \leq \int_b^\infty \frac{dx}{\cosh x} = 2\int_b^\infty \frac{e^x}{e^{2x}+1}dx \leq 2\int_b^\infty e^{-x}dx = 2e^{-b}.$$

Quindi, ad esempio, con $b = 10$, l'integrale proprio $I^{(1)}(f)$ approssima $I(f)$ con un errore assoluto non superiore a $2e^{-10} \simeq 10^{-4}$. □

Si osservi, come si constata facilmente, che per $x > 3$, risulta

$$\left| e^x \frac{\cos x}{\cosh x} \right| < \frac{e^x}{x^2} ;$$

quindi, per il Teorema 7.4.2 con $\alpha = 2$ e $x^* = 3$, la formula di Gauss-Laguerre $J_n(f)$ con $f(x) = e^x \frac{\cos x}{\cosh x}$ è convergente all'integrale dell'esempio precedente.

Avendosi $\lim_{n\to\infty}(J_{n+1}(f) - J_n(f)) = 0$, il numero $\mid J_{n+1}(f) - J_n(f) \mid$ può assumersi come una stima del modulo dell'errore assoluto che si commette approssimando $I(e^{-x}f)$ con $J_{n+1}(f)$.
Applicando la formula con $n = 4$ e $n = 5$, si trova:

$$J_5(f) = 0.6202592\ldots ,$$

$$\mid J_5(f) - J_4(f) \mid \simeq 10^{-2} .$$

L'errore effettivo risulta

$$\mid I(e^{-x}f) - J_6(f) \mid \simeq 5.8 \times 10^{-3} .$$

7.6.3 Polinomi ortogonali: formule di ricorrenza, nodi e pesi

I polinomi ortogonali possono essere costruiti con semplici formule di ricorrenza a tre termini. I loro zeri sono i nodi delle corrispondenti formule di quadratura. Su intervalli simmetrici rispetto all'origine i nodi sono simmetrici rispetto all'origine e a ogni coppia di nodi simmetrici compete lo stesso peso. Le costanti di normalizzazione h_i, $i \geq 0$, sono quelle date nella definizione di classe di polinomi ortogonali Π^* e riferite al prodotto scalare (7.20).

Polinomi di Chebyshev di 1^a *specie* ($h_0 = \pi,\ h_i = \pi/2,\ i > 0$).

$$T_{i+1}(x) = 2xT_i(x) - T_{i-1}(x),\ i = 1, 2, \ldots;\ T_0(x) = 1,\ T_1(x) = x.$$

I nodi e i pesi delle corrispondenti formule di quadratura sono esprimibili in forma chiusa. Poiché i pesi sono costanti per ogni valore di n, esse sono dette *formule a pesi uniformi*.

$$x_i = -\cos\frac{(2i+1)\pi}{2(n+1)}, \quad a_i = \frac{\pi}{n+1}, \quad i = 0, 1, \ldots, n.$$

Polinomi di Chebyshev di 2^a specie ($h_i = \pi/2$).

$$U_{i+1}(x) = 2xU_i(x) - U_{i-1}(x),\ i = 1, 2, \ldots;\ U_0(x) = 1,\ U_1(x) = 2x.$$

n	x_i	a_i
0	0.0000000000	1.5707963268
1	±0.5000000000	0.7853981634
2	±0.7071067812	0.3926990817
	0.0000000000	0.7853981634
3	±0.8090169944	0.2170787134
	±0.3090169944	0.5683194500
4	±0.8660254038	0.1308996939
	±0.5000000000	0.3926990817
	0.0000000000	0.5235987756
5	±0.9009688679	0.0844886909
	±0.6234898019	0.2743330561
	±0.2225209340	0.4265764164

Polinomi di Legendre ($h_i = 2/(2i+1)$).

$$\begin{aligned}(i+1)P_{i+1}(x) &= (2i+1)xP_i(x) - iP_{i-1}(x),\ i = 1, 2, \ldots;\\ P_0(x) &= 1,\ P_1(x) = x.\end{aligned}$$

n	x_i	a_i
0	0.0000000000	2.0000000000
1	±0.5773502692	1.0000000000
2	±0.7745966692	0.5555555555
	0.0000000000	0.8888888888
3	±0.8611363116	0.3478548451
	±0.3399810436	0.6521451549
4	±0.9061798459	0.2369268851
	±0.5384693101	0.4786286705
	0.0000000000	0.5688888888
5	±0.9324695142	0.1713244924
	±0.6612093865	0.3607615730
	±0.2386191861	0.4679139346

Polinomi di Laguerre ($h_i = 1$).

$$\begin{aligned}(i+1)L_{i+1}(x) &= (2i+1-x)L_i(x) - iL_{i-1}(x),\ i = 1, 2, \ldots; \\ L_0(x) &= 1,\ L_1(x) = -x + 1.\end{aligned}$$

n	x_i	a_i
0	1.0000000000	1.0000000000
1	0.5857864376	0.8535533906
	3.4142135624	0.1464466094
2	0.4157745568	0.7110930099
	2.2942803603	0.2785177336
	6.2899450829	0.0103892565
3	0.3225476896	0.6031541043
	1.7457611012	0.3574186924
	4.5366202969	0.0388879085
	9.3950709123	0.0005392947
4	0.2635603197	0.5217556106
	1.4134030591	0.3986668111
	3.5964257710	0.0759424497
	7.0858100059	0.0036117587
	12.640800844	0.0000233700
5	0.2228466042	0.4589646739
	1.1889321017	0.4170008308
	2.9927363261	0.1133733821
	5.7751435691	0.0103991975
	9.8374674184	0.0002610172
	15.982873981	0.0000008985

Polinomi di Hermite ($h_i = 2^i(i!)\sqrt{\pi}$).

$$H_{i+1}(x) = 2xH_i(x) - 2iH_{i-1}(x),\ i = 1, 2, \ldots;\ H_0(x) = 1,\ H_1(x) = 2x.$$

n	x_i	a_i
0	0.0000000000	1.7724538509
1	$\pm$0.7071067812	0.8862269255
2	$\pm$1.2247448714	0.2954089752
	0.0000000000	1.1816359006
3	$\pm$1.6506801239	0.0813128354
	$\pm$0.5246476233	0.8049140900
4	$\pm$2.0201828705	0.0199532421
	$\pm$0.9585724646	0.3936193232
	0.0000000000	0.9453087205
5	$\pm$2.3506049737	0.0045300099
	$\pm$1.3358490740	0.1570673203
	$\pm$0.4360774119	0.7246295952

Bibliografia: [7], [9], [30].

Capitolo 8

Metodi numerici per equazioni differenziali ordinarie

8.1 Introduzione

Il problema trattato nel presente capitolo riguarda l'approssimazione numerica di una funzione $y(t) : [a, b] \subseteq \mathbb{R} \rightarrow \mathbb{R}^m$, soluzione del seguente *problema di valori iniziali*, o *di Cauchy*, del primo ordine

$$\begin{aligned} y'(t) &= f(t, y(t)), \quad t \in [a, b], \\ y(t_0) &= y_0 \, , \end{aligned} \tag{8.1}$$

dove $f(t, y) : [a, b] \times \mathbb{R}^m \rightarrow \mathbb{R}^m$, $t_0 \in [a, b]$ e $y_0 \in \mathbb{R}^m$.

Prima di presentare alcune delle principali tecniche di approssimazione nel discreto, giova premettere qualche considerazione teorica sul problema continuo (8.1).

Teorema 8.1.1 *Sia $f(t, y)$ definita e continua nell'insieme*

$$D = \{(t, y) \mid -\infty < a \leq t \leq b < +\infty, \, \|y\| < +\infty\}$$

ed esista una costante L tale che

$$\|f(t, y) - f(t, y^*)\| \leq L\|y - y^*\| \tag{8.2}$$

per ogni $(t, y), (t, y^*) \in D$. Allora esiste un'unica soluzione $y(t) \in C^1([a, b])$ del problema (8.1), per ogni y_0 assegnato.

La costante L e la relazione (8.2) diconsi rispettivamente *costante* e *condizione di Lipschitz*.

Il precedente teorema vale sotto l'ipotesi, più restrittiva, che esista e sia continua la matrice jacobiana $J(t,y) : D \to \mathbb{R}^{m\times m}$ di f, definita da

$$J(t,y) = \frac{\partial f}{\partial y} \ :$$

in tal caso nella (8.2) può assumersi $L = \sup_{(t,y)\in D} \|\partial f/\partial y\|$.

Talvolta il problema di Cauchy è dato nella *forma autonoma*

$$\begin{aligned} y'(t) &= f(y), \\ y(t_0) &= y_0 \ . \end{aligned} \tag{8.3}$$

Ciò non lede la generalità in quanto ogni problema (8.1) può ricondursi alla forma (8.3) con la sostituzione $z_1(t) = y(t)$, $z_2(t) = t$, aggiungendo l'equazione $z_2'(t) = 1$ e completando le condizioni iniziali con $z_2(t_0) = t_0$; ponendo $z^T = (z_1^T, z_2)$, $g^T = (f(z)^T, 1)$, $z_0^T = (y_0^T, t_0)$ risulta infatti

$$\begin{aligned} z'(t) &= g(z) \ , \\ z(t_0) &= z_0 \ , \end{aligned}$$

che è, appunto, della forma (8.3).

Problemi differenziali di ordine superiore al primo possono essere trasformati in problemi equivalenti del primo ordine con una opportuna sostituzione. Si consideri, infatti, il problema di ordine $r > 1$

$$\begin{aligned} &y^{(r)} = f(t, y, y', \ldots, y^{(r-1)}) \ , \\ &y(t_0) = \eta_1 \ , \\ &y'(t_0) = \eta_2 \ , \\ &\qquad \vdots \\ &y^{(r-1)}(t_0) = \eta_r \ ; \end{aligned}$$

se si introduce il vettore ausiliario $z \in \mathbb{R}^{mr}$, $z^T = (z_1^T, z_2^T, \ldots, z_r^T)$, definito da

$$\begin{aligned} z_1 &= y \,, \\ z_2 &= z_1' = y' \,, \\ z_3 &= z_2' = y'' \,, \\ &\vdots \\ z_r &= z_{r-1}' = y^{(r-1)} \,, \end{aligned}$$

e si pone $z_r' = y^{(r)} = f(t, z_1, z_2, \ldots, z_r)$ e $z(t_0)^T = (\eta_1^T, \eta_2^T, \ldots, \eta_r^T) = z_0^T$, il precedente problema può scriversi come

$$\begin{aligned} z'(t) &= g(t, z) \,, \\ z(t_0) &= z_0 \,, \end{aligned}$$

con $g^T = (z_2^T, z_3^T, \ldots, z_r^T, f(t, z)^T)$, che è della forma (8.1).

Il problema (8.1) si dice *lineare* se è $f(t, y(t)) = K(t)y(t) + \alpha(t)$ con $K(t) : [a, b] \to \mathbb{R}^{m \times m}$ e $\alpha(t) : [a, b] \to \mathbb{R}^m$; si dice *lineare a coefficienti costanti* se K non dipende da t: con ciò il problema assume la forma

$$\begin{aligned} y'(t) &= Ky(t) + \alpha(t) \,, \\ y(t_0) &= y_0 \,. \end{aligned} \tag{8.4}$$

Nel caso importante che K sia diagonalizzabile, la soluzione generale di (8.4) è

$$y(t) = \sum_{i=1}^{m} d_i x^{(i)} e^{\lambda_i t} + \beta(t) \tag{8.5}$$

dove λ_i ed $x^{(i)}$, $i = 1, 2, \ldots, m$, sono rispettivamente gli autovalori e i corrispondenti autovettori di K, $\beta(t)$ è una soluzione particolare di (8.4) e le d_i, $i = 1, 2, \ldots, m$, sono costanti arbitrarie.

Il problema (8.4) si dice *omogeneo* se $\alpha(t) = 0$: nell'ipotesi fatta su K, la sua soluzione generale è la (8.5) con $\beta(t)$ identicamente nulla.

Nei metodi numerici considerati nel seguito, si farà riferimento ad un sottoinsieme discreto dell'intervallo $[a, b]$, $t_0 = a < t_1 < \cdots < t_N = b$, ottenuto con una progressione aritmetica di ragione $h > 0$, definita da $t_n = t_0 + nh$, $n = 0, 1, 2, \ldots, N$. La ragione h si dice *passo della discretizzazione.*

In corrispondenza ad una data discretizzazione, si indica con y_n una approssimazione di $y(t_n)$ ottenuta con un metodo numerico specifico in assenza di errori di arrotondamento, mentre, in presenza di tali errori (dovuti, in genere, ad una macchina da calcolo), l'approssimazione di $y(t_n)$ è indicata con $\tilde{y}_n$.

8.2 Metodi a un passo

8.2.1 Generalità

I *metodi a un passo* sono della forma generale

$$y_{n+1} = y_n + h\phi(h, t_n, y_n), \quad n = 0, 1, \ldots, \tag{8.6}$$

in cui la funzione ϕ dipende dalla funzione f del problema (8.1).

Posto $y_0 = y(t_0)$, la (8.6) serve a calcolare y_{n+1} conoscendo y_n.

Se ϕ dipende anche da y_{n+1} il metodo si dice *implicito* e, se ϕ non è lineare, y_{n+1} si calcola con una tecnica iterativa (cfr. (8.36)). Se ϕ non dipende da y_{n+1} il metodo è detto *esplicito* e la sua applicazione è immediata.

Due semplici metodi sono i seguenti:

la *formula trapezoidale*

$$y_{n+1} = y_n + \frac{h}{2}\left[f(t_n, y_n) + f(t_{n+1}, y_{n+1})\right] \quad \text{(implicito)};$$

il *metodo di Eulero*

$$y_{n+1} = y_n + hf(t_n, y_n) \quad \text{(esplicito)}.$$

Si introducono ora alcune definizioni.

Definizione 8.2.1 *Si dice* errore globale di discretizzazione *nel punto t_{n+1}, la differenza $e_{n+1} = y(t_{n+1}) - y_{n+1}$.*

Definizione 8.2.2 *Dicesi* errore locale di troncamento *la quantità τ_{n+1} definita da*

$$\tau_{n+1} = y(t_{n+1}) - u_{n+1} \tag{8.7}$$

dove

$$u_{n+1} = y(t_n) + h\phi(h, t_n, y(t_n))\,.$$

L'errore τ_{n+1} è quindi l'errore introdotto dal metodo al passo da t_n a t_{n+1} ed è uguale alla differenza fra il valore esatto $y(t_{n+1})$ e quello teorico u_{n+1} che si ottiene usando il metodo (8.6) col valore esatto $y(t_n)$ al posto di y_n.

Si consideri un passaggio al limite facendo tendere h a zero e n all'infinito in modo che $t_0 + nh$ resti fisso e si indichi tale operazione con $\lim_{\substack{h\to 0 \\ t=t_n}}$.

Definizione 8.2.3 *Un metodo (8.6) si dice* coerente *se vale la condizione*

$$\lim_{\substack{h\to 0 \\ t=t_{n+1}}} \frac{\tau_{n+1}}{h} = 0 \,.$$

Dalla (8.7) segue che la coerenza implica

$$\phi(0, t, y(t)) = f(t, y(t)) \,. \tag{8.8}$$

Si definisce *ordine del metodo* il più grande intero positivo p per cui risulta

$$\tau_{n+1} = O\left(h^{p+1}\right) \,.$$

Si noti che un metodo coerente ha ordine almeno $p = 1$ e che, in linea di principio, l'accuratezza del metodo cresce al crescere di p.

Definizione 8.2.4 *Un metodo si dice* convergente *se, applicato a un qualunque problema di Cauchy soddisfacente le ipotesi del Teorema 8.1.1, risulta, per ogni* $t \in [a, b]$,

$$\lim_{\substack{h\to 0 \\ t=t_{n+1}}} y_{n+1} = y(t_{n+1}) \,,$$

e quindi

$$\lim_{\substack{h\to 0 \\ t=t_{n+1}}} e_{n+1} = 0 \,.$$

Occorre considerare, infine, gli errori che nascono dagli arrotondamenti.

La differenza $\tilde{e}_{n+1} = y_{n+1} - \tilde{y}_{n+1}$ dicesi *errore globale di arrotondamento*. Tale errore è originato dagli errori di arrotondamento introdotti dalla macchina ad ogni passo. Questi si suppongono di solito indipendenti da h e quindi, per un fissato intervallo, il loro contributo cresce con il numero dei passi.

Si definisce *errore totale* $\hat{e}_{n+1} = y(t_{n+1}) - \tilde{y}_{n+1}$ l'errore che si accumula al passo t_{n+1} per l'effetto degli errori che si sono prodotti in ognuno dei passi precedenti.

Il seguente teorema stabilisce condizioni necessarie e sufficienti perché un metodo esplicito (8.6) sia convergente.

Teorema 8.2.1 *La funzione $\phi(h,t,y)$ sia continua nella regione*

$$\mathcal{D} = \{(h,t,y) \mid 0 < h \leq h_0,\ -\infty < a \leq t \leq b < +\infty\ ,\ \|y\| < +\infty\}$$

e inoltre soddisfi la seguente condizione di Lipschitz

$$\|\phi(h,t,y^*) - \phi(h,t,y)\| \leq M\|y^* - y\|$$

per ogni $(h,t,y^*), (h,t,y) \in \mathcal{D}$, allora il metodo (8.6) è convergente se e solo se è coerente.

Per i metodi considerati in 8.2.2 e 8.2.3 la funzione ϕ verifica le condizioni del Teorema 8.2.1 se la $f(t,y)$ soddisfa quelle del Teorema 8.1.1.

8.2.2 Metodi di Runge-Kutta

I *metodi di Runge-Kutta* costituiscono una importante classe di metodi della forma (8.6). La struttura generale di tali metodi è

$$y_{n+1} = y_n + h\sum_{i=1}^{s} b_i k_i \tag{8.9}$$

dove

$$k_i = f(t_n + c_i h, y_n + h\sum_{j=1}^{s} a_{ij}k_j), \quad i = 1,2,\ldots,s. \tag{8.10}$$

I parametri reali b_i, c_i, a_{ij} definiscono il metodo ed s è detto *numero di stadi*.

Pertanto un metodo di Runge-Kutta è un metodo a un passo in cui risulta

$$\phi(h,t_n,y_n) = \sum_{i=1}^{s} b_i k_i\ . \tag{8.11}$$

Si osservi che, dalla convergenza del metodo, segue

$$\lim_{\substack{h\to 0 \\ t=t_{n+1}}} k_i = f(t,y(t)), \quad i = 1,2,\ldots,s,$$

per cui, tenuto conto della (8.11), la condizione di coerenza (8.8) equivale a

$$\sum_{i=1}^{s} b_i = 1\ .$$

Per convenzione, comunemente adottata, si pone

$$c_i = \sum_{j=1}^{s} a_{ij}, \quad i = 1, 2, \ldots, s\,. \tag{8.12}$$

Si ottiene una utile rappresentazione compatta di un metodo di Runge-Kutta per mezzo della seguente *tavola di Butcher*

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array} = \begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \cdots & \cdots & \cdots & \cdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array}\,.$$

Si distinguono due classi di metodi, riconoscibili dalla forma della matrice A:

metodi espliciti se $a_{ij} = 0$, per ogni coppia i, j con $1 \le i \le j \le s$;

metodi impliciti se $a_{ij} \neq 0$ per qualche coppia i, j con $i \le j$.

Nel primo caso le (8.10) hanno la forma

$$k_i = f(t_n + c_i h, y_n + h \sum_{j=1}^{i-1} a_{ij} k_j), \quad i = 1, 2, \ldots, s,$$

e quindi ciascun vettore k_i si può calcolare esplicitamente in funzione dei precedenti k_j, $j = 1, 2, \ldots, i-1$.

Nel secondo caso, introducendo il vettore $k^T = (k_1^T, \ldots, k_s^T) \in \mathbb{R}^{ms}$, le (8.10) si possono scrivere in forma implicita

$$k = \varphi(k) \tag{8.13}$$

con $\varphi_i(k) = f(t_n + c_i h, y_n + h \sum_{j=1}^{s} a_{ij} k_j)$, $i = 1, 2, \ldots, s$.

Il sistema (8.13) è lineare o non lineare a seconda che lo sia $f(t, y)$ rispetto a y e può essere risolto con un algoritmo specifico.

Una sottoclasse dei metodi impliciti è costituita dai *metodi semi-impliciti* nei quali è $a_{ij} = 0$, $1 \le i < j \le s$, e $a_{ii} \neq 0$ per almeno un i. In tal caso si ha

$$k_i = f(t_n + c_i h, y_n + h \sum_{j=1}^{i-1} a_{ij} k_j + h a_{ii} k_i), \quad i = 1, 2, \ldots s,$$

e quindi le (8.10) possono essere risolte singolarmente con un costo computazionale più contenuto di quello richiesto per il sistema (8.13).

L'errore locale di troncamento di un metodo di Runge-Kutta di ordine p è della forma

$$\tau_{n+1} = h^{p+1}\psi$$

dove ψ è una funzione dipendente in modo non semplice da y_n, c, b e dagli elementi di A.

Di seguito si riportano alcuni metodi espliciti, con $s \leq 4$, scelti fra i più noti.

Metodo di Eulero, $p = 1$:

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}. \tag{8.14}$$

Metodo di Eulero modificato, $p = 2$:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1/2 & 1/2 \end{array}.$$

Metodo della poligonale, $p = 2$:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 0 & 1 \end{array}.$$

Formula di Heun, $p = 3$:

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 \\ 2/3 & 0 & 2/3 & 0 \\ \hline & 1/4 & 0 & 3/4 \end{array}.$$

Formula di Kutta, $p = 3$:

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 1 & -1 & 2 & 0 \\ \hline & 1/6 & 2/3 & 1/6 \end{array}.$$

Metodo di Runge-Kutta classico, $p = 4$:

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array} . \tag{8.15}$$

In generale i metodi impliciti sono classificati in base al tipo di formula di quadratura a cui danno luogo allorché vengono applicati al problema $y' = f(t)$. Si ha in questo caso

$$y_{n+1} - y_n = h \sum_{i=1}^{s} b_i f(t_n + c_i h) .$$

Il secondo membro può intendersi come una formula di quadratura, con pesi b_i e nodi c_i, che approssima l'integrale

$$\int_{t_n}^{t_{n+1}} f(t)dt ;$$

infatti si ha

$$y_{n+1} - y_n \simeq y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} f(t)dt = h \int_0^1 f(t_n + ch)dc .$$

Di seguito si riportano alcuni esempi con $s \leq 3$.

Metodi di Gauss-Legendre, $p = 2s$:

$$\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array} ,$$

$$\begin{array}{c|cc} (3-\sqrt{3})/6 & 1/4 & (3-2\sqrt{3})/12 \\ (3+\sqrt{3})/6 & (3+2\sqrt{3})/12 & 1/4 \\ \hline & 1/2 & 1/2 \end{array} .$$

Metodi di Radau IA, $p = 2s - 1$:

$$\begin{array}{c|c} 0 & 1 \\ \hline & 1 \end{array} , \tag{8.16}$$

$$\begin{array}{c|cc} 0 & 1/4 & -1/4 \\ 2/3 & 1/4 & 5/12 \\ \hline & 1/4 & 3/4 \end{array}.$$

Metodi di Radau IIA, $p = 2s - 1$:

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array},$$

$$\begin{array}{c|cc} 1/3 & 5/12 & -1/12 \\ 1 & 3/4 & 1/4 \\ \hline & 3/4 & 1/4 \end{array}.$$

Metodi di Lobatto IIIA, $p = 2s - 2$:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}, \tag{8.17}$$

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/2 & 5/24 & 1/3 & -1/24 \\ 1 & 1/6 & 2/3 & 1/6 \\ \hline & 1/6 & 2/3 & 1/6 \end{array}.$$

Metodi di Lobatto IIIB, $p = 2s - 2$:

$$\begin{array}{c|cc} 0 & 1/2 & 0 \\ 1 & 1/2 & 0 \\ \hline & 1/2 & 1/2 \end{array}, \tag{8.18}$$

$$\begin{array}{c|ccc} 0 & 1/6 & -1/6 & 0 \\ 1/2 & 1/6 & 1/3 & 0 \\ 1 & 1/6 & 5/6 & 0 \\ \hline & 1/6 & 2/3 & 1/6 \end{array}.$$

Metodi di Lobatto IIIC, $p = 2s - 2$:

$$\begin{array}{c|cc} 0 & 1/2 & -1/2 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array},$$

$$\begin{array}{c|ccc} 0 & 1/6 & -1/3 & 1/6 \\ 1/2 & 1/6 & 5/12 & -1/12 \\ 1 & 1/6 & 2/3 & 1/6 \\ \hline & 1/6 & 2/3 & 1/6 \end{array} .$$

Si osservi che i metodi (8.16) e (8.18) sono un esempio di eccezione alla condizione (8.12).

Esempi di metodi semi-impliciti sono il metodo (8.17), già riportato in 8.2.1 come "formula trapezoidale", e il metodo (8.18).

È da rilevare che i metodi impliciti hanno, in genere, a parità di numero di stadi, un ordine superiore a quello dei metodi espliciti. Inoltre per i metodi di Gauss-Legendre, Radau e Lobatto l'ordine p dipende in modo semplice e diretto da s. Un risultato di questo tipo non è disponibile per i metodi espliciti per i quali, invece, si hanno teoremi che stabiliscono per ogni ordine fissato p il minimo numero s degli stadi (o, per contro, per ogni s fissato, il massimo ordine ottenibile). Al riguardo la situazione desumibile dalla letteratura attuale è riassunta nella tavola seguente.

p	1	2	3	4	5	6	7	8	9	10
s minimo	1	2	3	4	6	7	9	11	$12 \leq s \leq 17$	$13 \leq s \leq 17$

Tavola 8.1: Corrispondenza tra ordine e numero minimo di stadi.

8.2.3 Stabilità dei metodi di Runge-Kutta

Si applichi un metodo di Runge-Kutta, nel caso scalare $m = 1$, al seguente *problema test*

$$y'(t) = \lambda y(t), \quad \lambda \in \mathbb{C}\,,\ Re(\lambda) < 0, \tag{8.19}$$

la cui soluzione è $y(t) = de^{\lambda t}$, con d costante arbitraria.

Poiché $\lim_{t\to\infty} y(t) = 0$, è naturale richiedere che la soluzione numerica abbia un comportamento analogo a quello della soluzione continua, cioè che sia, per ogni n,

$$| \, y_{n+1} \, | \leq c \, | \, y_n \, | \tag{8.20}$$

per qualche costante positiva $c < 1$.
Tale relazione implica che, per il passo h prescelto, gli errori di discretizzazione si mantengono limitati col procedere dei calcoli.

Questa proprietà riguarda la sensibilità agli errori di uno schema discreto, ovvero la sua *stabilità numerica*.

Per giungere ad una condizione che garantisca la (8.20) si faccia riferimento, ad esempio, al più elementare dei metodi di Runge-Kutta, il metodo di Eulero (8.14), il quale può scriversi

$$y_{n+1} = y_n + hf(t_n, y_n) \, . \tag{8.21}$$

Questo metodo, applicato al problema (8.19), fornisce

$$y_{n+1} = (q+1)y_n$$

dove si è posto

$$q = h\lambda \, .$$

Pertanto la (8.20) sarà verificata se e solo se

$$| \, q+1 \, | < 1 \, ,$$

ovvero se i valori del parametro q, nel piano complesso, sono interni al cerchio di centro $[-1, 0]$ e raggio unitario.

Generalizzando quanto sopra esposto per il metodo esplicito (8.21) al caso di un metodo di Runge-Kutta qualunque, cioè applicando un metodo (8.9)-(8.10) al problema (8.19), si ottiene un'equazione della forma

$$y_{n+1} = R(q)y_n \, . \tag{8.22}$$

$R(q)$ è detta *funzione di stabilità* ed è verificata la (8.20) se e solo se

$$| \, R(q) \, | < 1 \, . \tag{8.23}$$

Definizione 8.2.5 *Un metodo di Runge-Kutta si dice* assolutamente stabile*, per un dato q, se la sua funzione di stabilità soddisfa la condizione (8.23).*

Definizione 8.2.6 *L'insieme del piano complesso*

$$S_A = \{q \in \mathbb{C} \; | \; | \, R(q) \, | < 1\}$$

si chiama regione di assoluta stabilità *del metodo.*

Si può dimostrare che, introdotto il vettore $u = (1, 1, \ldots, 1)^T \in \mathbb{R}^s$, la funzione di stabilità di un metodo di Runge-Kutta è data da

$$R(q) = \frac{\det(I - qA + qub^T)}{\det(I - qA)} \,. \tag{8.24}$$

A titolo di verifica qui si ricavano la (8.22) e (8.24) per la formula trapezoidale

$$y_{n+1} = y_n + \frac{h}{2}\left[f(y_n) + f(y_{n+1})\right] \,:$$

applicandola al problema test (8.19) si ha

$$y_{n+1} = \frac{1 + q/2}{1 - q/2}\, y_n \,.$$

Facendo riferimento alla formula espressa nella forma (8.17), la stessa funzione di stabilità $R(q) = \frac{1+q/2}{1-q/2}$ si può ottenere direttamente dalla (8.24) ponendo

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ A = \begin{pmatrix} 0 & 0 \\ 1/2 & 1/2 \end{pmatrix},\ u = \begin{pmatrix} 1 \\ 1 \end{pmatrix},\ b = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix} \,.$$

Nella (8.24) $\det(I - qA + qub^T)$ e $\det(I - qA)$ sono polinomi in q, a coefficienti reali, di grado $\leq s$, quindi $R(q)$ è una funzione razionale (cfr. 6.5). Poiché allora $R(\bar{q}) = \overline{R(q)}$ ne viene che S_A è simmetrica rispetto all'asse reale.

Si vede subito che per la formula trapezoidale la regione S_A è l'intero semipiano $Re(q) < 0$.

Si noti che per la soluzione esatta del problema test si ha

$$y(t_{n+1}) = de^{\lambda t_{n+1}} = y(t_n)e^q \,,$$

mentre dalla (8.22), per definizione di errore locale di troncamento, si ottiene

$$y(t_{n+1}) = R(q)y(t_n) + \tau_{n+1} \,,$$

da cui, confrontando con la precedente e supponendo $\tau_{n+1} = O(h^{p+1})$, si ha

$$e^q - R(q) = O(h^{p+1}) \,. \tag{8.25}$$

Dalla (8.25) segue che, per $Re(q) > 0$ ed h sufficientemente piccolo, si ha $|\, R(q)\, | > 1$, cioè $q \notin S_A$ e quindi l'intersezione di S_A con l'asse reale è del tipo $]\alpha, 0[$ con $\alpha < 0$.

Definizione 8.2.7 *Un metodo si dice* A_0-stabile *se*

$$S_A \supseteq \{q \mid \ Re(q) < 0,\ Im(q) = 0\} \ ,$$

cioè se S_A *contiene l'intero semiasse reale negativo.*

Definizione 8.2.8 *Un metodo si dice* A-stabile *se*

$$S_A \supseteq \{q \mid \ Re(q) < 0\} \ .$$

In un metodo A-stabile, quindi, la condizione di stabilità $\mid R(q) \mid < 1$ è garantita indipendentemente dal passo h purché sia $Re(q) < 0$. Inoltre la A-stabilità implica la A_0-stabilità.

Definizione 8.2.9 *Un metodo si dice* L-stabile *se è A-stabile e se*

$$\lim_{Re(q)\to-\infty} \mid R(q) \mid = 0 \, .$$

Si noti, per esempio, che la formula trapezoidale è A-stabile, ma, essendo $\lim_{Re(q)\to-\infty} R(q) = -1$, non è L-stabile.

Poiché nei metodi espliciti la matrice A è triangolare inferiore con elementi diagonali nulli, risulta $\det(I - qA) = 1$ e pertanto $R(q)$ è un polinomio di grado compreso fra 1 e s; per questi metodi risulta $\lim_{Re(q)\to-\infty} \mid R(q) \mid = +\infty$. Resta quindi provato il seguente teorema.

Teorema 8.2.2 *Non esistono metodi di Runge-Kutta espliciti A-stabili.*

È da notare che le varie definizioni di stabilità date per il caso scalare $m = 1$, si estendono anche al caso $m > 1$ facendo riferimento al problema lineare a coefficienti costanti

$$y' = Ky \, , \tag{8.26}$$

dove si suppone che K abbia autovalori $\lambda_1, \lambda_2, \ldots, \lambda_m$ distinti e che sia

$$Re(\lambda_i) < 0 \, , \quad i = 1, 2, \ldots, m. \tag{8.27}$$

Posto

$$q_i = h\lambda_i \, , \quad i = 1, 2, \ldots, m,$$

l'equivalenza del sistema (8.26) ad un sistema di m equazioni indipendenti $z_i' = \lambda_i z_i$, $i = 1, 2, \ldots, m$, che si ottiene diagonalizzando la matrice K, giustifica la seguente definizione.

Definizione 8.2.10 *Un metodo di Runge-Kutta, applicato al problema (8.26), si dice assolutamente stabile per un insieme di valori* $q_i \in \mathcal{C}$, $i = 1, 2, \ldots, m$, *se*

$$q_i \in S_A \,, \quad i = 1, 2, \ldots, m, \tag{8.28}$$

dove S_A *è la regione di assoluta stabilità definita nel caso scalare (cfr. Definizione 8.2.6).*

Si può dimostare che nei metodi espliciti con $p = s$, cioè (cfr. Tavola 8.1) per $s = 1, 2, 3, 4$, si ha

$$R(q) = 1 + q + \frac{1}{2!}q^2 + \cdots + \frac{1}{s!}q^s \,. \tag{8.29}$$

Nella Tavola 8.2 si riportano il valore minimo di $Re(q)$ e il valore massimo di $\mid Im(q) \mid$ per le regioni di assoluta stabilità associate alla (8.29).

s	$Re(q)$	$\mid Im(q) \mid$
1	-2	1
2	-2	1.75
3	-2.5	2.4
4	-2.785	2.95

Tavola 8.2: Valori estremi di S_A per i metodi espliciti con $p = s$.

Le regioni S_A di assoluta stabilità per i metodi (8.29) con $s = 1, 2, 3, 4$ sono riportate in Fig. 8.1, e sono costituite dai punti interni alle porzioni di piano delimitate dalle curve date per ogni valore di s.

Si osservi ora che dalla (8.25) si ha $e^q \simeq R(q)$, e quindi $R(q)$ è una funzione razionale che approssima l'esponenziale e^q. Si comprende quindi come possa essere interessante confrontare le funzioni di stabilità $R(q)$ dei vari metodi di Runge-Kutta con le cosiddette *approssimazioni razionali di Padé* della funzione esponenziale. Esse sono date da

$$R_j^k(q) = \frac{P_k(q)}{Q_j(q)} \,,$$

dove

$$P_k(q) = \sum_{i=0}^{k} \frac{k!(j+k-i)!}{(k-i)!(j+k)!i!} q^i \,,$$

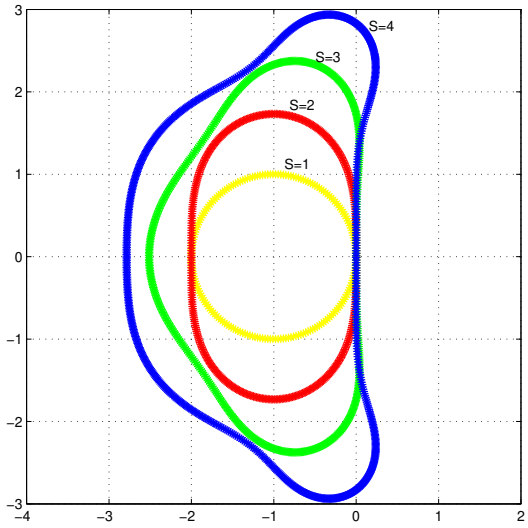

Figura 8.1: Regioni S_A dei metodi di Runge Kutta espliciti con $s = 1, 2, 3, 4$.

$$Q_j(q) = \sum_{i=0}^{j} \frac{j!(j+k-i)!}{(j-i)!(j+k)!i!}(-q)^i \, .$$

$R_j^k(q)$ è l'unica funzione razionale con numeratore di grado k e denominatore di grado j tale che

$$e^q = R_j^k(q) + O(q^{k+j+1}) \tag{8.30}$$

quando $q \to 0$. Si riportano nella Tavola 8.3 le espressioni di $R_j^k(q)$ con $0 \le k, j \le 3$.

	$k=0$	$k=1$	$k=2$	$k=3$
$j=0$	1	$1+q$	$1+q+\frac{1}{2}q^2$	$1+q+\frac{1}{2}q^2+\frac{1}{6}q^3$
$j=1$	$\frac{1}{1-q}$	$\frac{1+\frac{1}{2}q}{1-\frac{1}{2}q}$	$\frac{1+\frac{2}{3}q+\frac{1}{6}q^2}{1-\frac{1}{3}q}$	$\frac{1+\frac{3}{4}q+\frac{1}{4}q^2+\frac{1}{24}q^3}{1-\frac{1}{4}q}$
$j=2$	$\frac{1}{1-q+\frac{1}{2}q^2}$	$\frac{1+\frac{1}{3}q}{1-\frac{2}{3}q+\frac{1}{6}q^2}$	$\frac{1+\frac{1}{2}q+\frac{1}{12}q^2}{1-\frac{1}{2}q+\frac{1}{12}q^2}$	$\frac{1+\frac{3}{5}q+\frac{3}{20}q^2+\frac{1}{60}q^3}{1-\frac{2}{5}q+\frac{1}{20}q^2}$
$j=3$	$\frac{1}{1-q+\frac{1}{2}q^2-\frac{1}{6}q^3}$	$\frac{1+\frac{1}{4}q}{1-\frac{3}{4}q+\frac{1}{4}q^2-\frac{1}{24}q^3}$	$\frac{1+\frac{2}{5}q+\frac{1}{20}q^2}{1-\frac{3}{5}q+\frac{3}{20}q^2-\frac{1}{60}q^3}$	$\frac{1+\frac{1}{2}q+\frac{1}{10}q^2+\frac{1}{120}q^3}{1-\frac{1}{2}q+\frac{1}{10}q^2-\frac{1}{120}q^3}$

Tavola 8.3: Approssimazioni di Padé dell'esponenziale.

Si constata che molte approssimazioni di Padé coincidono con altrettante funzioni di stabiltà. Per esempio, dalla Tavola 8.3 si vede che le approssimazioni $R_0^k(q)$, $k = 1, 2, 3$, coincidono con le funzioni di stabilità dei metodi di Runge-Kutta espliciti con $p = s$.

Nella definizione seguente si caratterizzano le varie approssimazioni di Padé .

Definizione 8.2.11 *Una approssimazione $R_j^k(q)$ di Padé si dice:*
A_0-accettabile *se* $\mid R_j^k(q) \mid < 1$ *quando* $Re(q) < 0$ *e* $Im(q) = 0$*;*
A-accettabile *se* $\mid R_j^k(q) \mid < 1$ *quando* $Re(q) < 0$*;*
L-accettabile *se è A-accettabile e* $\lim_{Re(q)\to-\infty} \mid R_j^k(q) \mid = 0$.

Evidentemente $R_j^k(q)$ non può essere A-accettabile se $k > j$. Al riguardo si ha il seguente risultato che dimostra la cosiddetta *congettura di Ehle*.

Teorema 8.2.3 *Le approssimazioni di Padé $R_j^k(q)$ sono A-accettabili se e solo se $j - 2 \leq k \leq j$.*

Ne segue che se $j - 2 \leq k < j$, $R_j^k(q)$ sono anche L-accettabili.
Vale poi il seguente teorema.

Teorema 8.2.4 *$R_j^k(q)$ è A_0-accettabile se e solo se $k \leq j$.*

Se per un metodo risulta $R(q) = R_j^k(q)$, con $k \leq j$, allora esso sarà A_0-stabile, A-stabile oppure L-stabile a seconda che $R_j^k(q)$ sia A_0-accettabile, A-accettabile o L-accettabile.

Infine si possono dimostrare le proposizioni che seguono, le quali, per semplicità, si sono riunite in un unico teorema:

Teorema 8.2.5 *Risulta:*

$R(q) = R_s^s(q)$ per i metodi di Gauss-Legendre a s stadi;

$R(q) = R_s^{s-1}(q)$ per i metodi di Radau IA e Radau IIA a s stadi;

$R(q) = R_{s-1}^{s-1}(q)$ per i metodi di Lobatto IIIA e Lobatto IIIB a s stadi;

$R(q) = R_s^{s-2}(q)$ per i metodi di Lobatto IIIC a s stadi.

Quindi i metodi riportati da questo teorema sono A-stabili per ogni valore di s e, in particolare, i metodi di Radau IA, Radau IIA e di Lobatto IIIC sono anche L-stabili.

Una classe di problemi particolarmente importanti, per i quali sono utili i metodi A-stabili, è quella dei cosiddetti *problemi stiff*. Limitandosi, per semplicità, al caso di un problema lineare, siano λ^* e λ^{**} autovalori della matrice K di (8.4) tali che

$$| Re(\lambda^*) | = \min \{| Re(\lambda_1) |, | Re(\lambda_2) |, \ldots, | Re(\lambda_m) |\} ,$$

$$| Re(\lambda^{**}) | = \max \{| Re(\lambda_1) |, | Re(\lambda_2) |, \ldots, | Re(\lambda_m) |\} .$$

Il problema (8.4) si dice stiff se, per gli autovalori di K, vale l'ipotesi (8.27) e risulta anche:

$$| Re(\lambda^{**}) | \gg | Re(\lambda^*) | ,$$

$$| Re(\lambda^{**}) | \gg 1 .$$

In tal caso, nella soluzione (8.5), la parte $\sum_{i=1}^{m} d_i x^{(i)} e^{\lambda_i t}$, che prende il nome di *soluzione transitoria*, contiene la componente $e^{\lambda^{**} t}$ che tende a zero per $t \to \infty$, variando rapidamente in $T^{**} \equiv [t_0, t_0 + \mid \lambda^{**} \mid^{-1}]$.

È chiaro che per approssimare la componente $e^{\lambda^{**} t}$ sull'intervallo T^{**} è necessario un passo h^{**} dell'ordine di $\mid \lambda^{**} \mid^{-1}$. Un tale passo risulta troppo piccolo nei riguardi della componente $e^{\lambda^* t}$: infatti questa decresce lentamente in $T^* \equiv [t_0, t_0 + \mid \lambda^* \mid^{-1}]$ e risulta che l'ampiezza di T^* è molto maggiore di quella di T^{**}. Fuori dell'intervallo T^{**} sarebbe quindi opportuno aumentare la lunghezza del passo, compatibilmente con l'accuratezza che si vuole ottenere. Tuttavia, per i metodi con S_A limitata, il rispetto della condizione di stabilità (8.28) costringe ad usare il passo h^{**} anche su tutto l'intervallo T^*, facendo crescere in modo inaccettabile il numero dei passi necessari. Analoghe considerazioni possono farsi nei riguardi della componente $e^{\lambda^{**} t}$ confrontata con $\beta(t)$, cioè con quella parte della (8.5) che, nell'ipotesi (8.27), è detta *soluzione stazionaria*.

Al riguardo si consideri il seguente esempio.

Si vuole approssimare la soluzione del problema omogeneo (8.26) con un errore locale di troncamento dell'ordine di 10^{-4}; siano gli autovalori di K reali e $\lambda^{**} = -1000$, $\lambda^* = -0.1$; si applichi il metodo di Runge-Kutta classico (8.15), di ordine $p = 4$, il cui intervallo reale di stabilità è $]-2.785, 0[$. Il passo $h = 0.1$ è sufficiente per la precisione che si richiede, tuttavia le condizioni (8.28) sono soddisfatte solo se $q^{**} = h\lambda^{**} \in]-2.785, 0[$, ovvero se $h < 0.002785$. Assumendo $t_0 = 0$, il termine della soluzione relativo a

λ^{**} tende rapidamente a zero, mentre quello relativo a λ^* sarà prossimo a zero per un valore t_ν nettamente maggiore; sia esso $t_\nu \simeq \mid \lambda^* \mid^{-1} = 10$. Occorrono perciò $\nu = 10/h \simeq 3600$ passi per approssimare la soluzione $y(t)$ sull'intrevallo $[t_0, t_\nu]$. Poiché ad ogni passo si effettuano $s = 4$ valutazioni della funzione Ky, il costo computazionale risulta elevato e può divenire apprezzabile il fenomeno di accumulo degli errori di arrotondamento. Questi inconvenienti si possono evitare se si usa una tecnica di variazione del passo e se si utilizza un metodo A-stabile. In questo caso, poiché le (8.28) sono soddisfatte qualunque sia h, l'unico vincolo al passo è dato dal valore richiesto per l'errore locale di troncamento e a parità di ordine, con 100 passi si ottiene il risultato voluto.

8.3 Metodi a più passi

8.3.1 Equazioni alle differenze

Si consideri l'equazione

$$\sum_{j=0}^{k} \gamma_j y_{n+j} = b_n \,, \quad n = 0, 1, \ldots, \tag{8.31}$$

dove γ_j sono costanti scalari e b_n un vettore di $\mathbb{R}^m$ assegnato. La (8.31) prende il nome di *equazione lineare alle differenze a coefficienti costanti di ordine* k e la sua soluzione è una successione di vettori y_n di $\mathbb{R}^m$. Sia y_n^* una sua soluzione particolare e z_n la soluzione generale dell'equazione omogenea associata

$$\sum_{j=0}^{k} \gamma_j y_{n+j} = 0 \,, \quad n = 0, 1, \ldots, \tag{8.32}$$

allora la soluzione generale della (8.31) è data da

$$y_n = z_n + y_n^* \,.$$

Per sostituzione si trova che $z_n = d\mu^n$, $d \in \mathbb{R}^m$, è una soluzione della (8.32) se μ è radice del *polinomio caratteristico*

$$\pi(\mu) = \sum_{j=0}^{k} \gamma_j \mu^j \,.$$

Se $\pi(\mu)$ ha k radici distinte $\mu_1, \mu_2, \ldots, \mu_k$, l'insieme $\{\mu_1^n, \mu_2^n, \ldots, \mu_k^n\}$ forma un *sistema fondamentale di soluzioni* e la soluzione generale della (8.32) è

$$z_n = \sum_{i=1}^{k} d_i \mu_i^n \tag{8.33}$$

essendo d_i vettori arbitrari di $\mathbb{R}^m$. Se una radice, ad esempio μ_j, ha molteplicità ν e le rimanenti sono tutte distinte, l'insieme

$$\{\mu_1^n, \ldots, \mu_{j-1}^n, \mu_j^n, n\mu_j^n, n^2\mu_j^n, \ldots, n^{\nu-1}\mu_j^n, \mu_{j+1}^n, \ldots, \mu_{k-\nu+1}^n\}$$

forma ancora un sistema fondamentale di soluzioni e si ha

$$z_n = \sum_{i=1}^{j-1} d_i \mu_i^n + \sum_{i=j}^{j+\nu-1} d_i n^{i-j} \mu_j^n + \sum_{i=j+\nu}^{k} d_i \mu_{i-\nu+1}^n \, . \tag{8.34}$$

È immediata l'estensione al caso generale che $\pi(\mu)$ abbia r radici distinte μ_i, $i = 1, 2, \ldots, r$, ciascuna con molteplicità ν_i con $\sum_{i=1}^{r} \nu_i = k$.

Si indichi con E l'*operatore di avanzamento* definito da

$$E y_n = y_{n+1} \, ;$$

ammettendo che l'operatore E sia lineare, cioè risulti $E(\alpha y_n + \beta y_{n+1}) = \alpha E y_n + \beta E y_{n+1}$, e convenendo di porre $E^2 y_n = E(E y_n)$, la (8.31) può formalmente scriversi

$$\pi(E) y_n = b_n \, .$$

8.3.2 Metodi lineari

Un *metodo lineare a più passi*, o *a k passi*, con $k \geq 1$, per approssimare la soluzione del problema (8.1) ha la struttura seguente

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f(t_{n+j}, y_{n+j}) \, , \quad n = 0, 1, \ldots, \tag{8.35}$$

cioè ha la forma di una equazione alle differenze, lineare rispetto a y_{n+j} e $f(t_{n+j}, y_{n+j})$, $j = 0, 1, \ldots, k$.

I coefficienti α_j e β_j sono costanti reali e si ammette che sia

$$| \, \alpha_0 \, | + | \, \beta_0 \, | \neq 0 \, , \quad \alpha_k = 1 \, .$$

Per ogni n, la (8.35) fornisce il vettore y_{n+k} in funzione dei k vettori precedenti $y_{n+k-1}, y_{n+k-2}, \ldots, y_n$.
Si suppongono noti (dati o calcolati) i k *vettori iniziali* $y_0, y_1, \ldots, y_{k-1}$.

Se $\beta_k \neq 0$ il metodo si dice *implicito*: posto

$$w = -\sum_{j=0}^{k-1} \alpha_j y_{n+j} + h \sum_{j=0}^{k-1} \beta_j f(t_{n+j}, y_{n+j}) \,,$$

si calcola, per $n = 0, 1, \ldots$, una approssimazione z^* della soluzione dell'equazione

$$z = h\beta_k f(t_{n+k}, z) + w \tag{8.36}$$

e si assume quindi $y_{n+k} = z^*$. Se $f(t, y)$ è lineare rispetto a y la (8.36) si riduce ad un sistema lineare; in caso contrario si può utilizzare, ad esempio, il seguente procedimento iterativo

$$z^{(r+1)} = h\beta_k f(t_{n+k}, z^{(r)}) + w \,, \quad r = 0, 1, \ldots,$$

la cui convergenza è garantita se (cfr. Teorema 4.6.1)

$$h \mid \beta_k \mid L < 1 \,,$$

dove L è la costante di Lipschitz della funzione $f(t, z)$. Se esiste la matrice $\partial f/\partial z$ si può porre $L = \sup_{(t,z)\in D} \|\partial f/\partial z\|$.

Se $\beta_k = 0$ il metodo si dice *esplicito* e il calcolo di y_{n+k} è diretto.

Al metodo (8.35) sono associati i seguenti due polinomi, detti *primo* e *secondo polinomio caratteristico*,

$$\rho(\mu) = \sum_{j=0}^{k} \alpha_j \mu^j \,, \quad \sigma(\mu) = \sum_{j=0}^{k} \beta_j \mu^j \,.$$

Convenendo di porre

$$f_{n+j} = f(t_{n+j}, y_{n+j}) \,,$$

la (8.35) può formalmente scriversi

$$\rho(E) y_n = h\sigma(E) f_n \,.$$

Definizione 8.3.1 *Dicesi* errore locale di troncamento *di un metodo a più passi la quantità* τ_{n+k} *definita da*

$$\tau_{n+k} = \sum_{j=0}^{k} \alpha_j y(t_{n+j}) - h \sum_{j=0}^{k} \beta_j f(t_{n+j}, y(t_{n+j})) \,. \tag{8.37}$$

Questa definizione, nell'ambito dei metodi lineari, coincide con la Definizione 8.2.2 se il metodo è esplicito, mentre, se il metodo è implicito, le due definizioni forniscono errori che differiscono per termini dell'ordine di h. Analogamente si hanno poi le definizioni di coerenza, ordine e convergenza. Per un metodo a più passi la condizione di coerenza è data da

$$\lim_{\substack{h\to 0 \\ t=t_{n+k}}} \frac{\tau_{n+k}}{h} = 0 \,, \tag{8.38}$$

e l'ordine è il più grande intero p per cui risulta

$$\tau_{n+k} = O(h^{p+1}) \,.$$

Restano invariate, rispetto ai metodi a un passo, le definizioni di errore globale di discretizzazione, di errore globale di arrotondamento e di errore totale (riferite al calcolo di y_{n+k} nel punto t_{n+k}).

In generale i vettori $y_1, \ldots, y_{k-1}$ dipendono da h e si dice che formano un *insieme compatibile di vettori iniziali* se vale la proprietà

$$\lim_{h\to 0} y_i = y_0 \,, \quad i = 1, 2, \ldots, k-1.$$

Definizione 8.3.2 *Il metodo (8.35) si dice* convergente *se, applicato a un qualunque problema soddisfacente le ipotesi del Teorema 8.1.1, è tale che, per ogni* $t \in [a, b]$*, si abbia*

$$\lim_{\substack{h\to 0 \\ t=t_{n+k}}} y_{n+k} = y(t_{n+k}) \,,$$

per ogni insieme compatibile di vettori iniziali.

Come già osservato per i metodi ad un passo, anche per questi metodi, la riduzione di h su un dato intervallo può produrre un aumento dei contributi all'errore totale: questo fenomeno è dovuto all'accumulo degli errori locali di arrotondamento i quali possono ritenersi indipendenti da h.

Nell'ipotesi che $y(t) \in C^\infty(]a,b[)$, sviluppando il secondo membro della (8.37) si può scrivere formalmente

$$\tau_{n+k} = c_0 y(t_n) + c_1 y'(t_n)h + c_2 y''(t_n)h^2 + \cdots + c_r y^{(r)}(t_n)h^r + \cdots ,$$

dove le c_i, $i = 0, 1, \ldots$, sono costanti date da

$$\begin{aligned} c_0 &= \sum_{j=0}^{k} \alpha_j = \rho(1) , \\ c_1 &= \sum_{j=0}^{k} (j\alpha_j - \beta_j) = \rho'(1) - \sigma(1) , \\ c_r &= \sum_{j=0}^{k} \left(\frac{1}{r!} j^r \alpha_j - \frac{1}{(r-1)!} j^{r-1} \beta_j \right) , \ r = 2, 3, \ldots . \end{aligned} \tag{8.39}$$

Ne segue che per un metodo di ordine p deve essere $c_0 = c_1 = \cdots = c_p = 0$, $c_{p+1} \neq 0$; quindi risulta

$$\tau_{n+k} = c_{p+1} y^{(p+1)}(t_n) h^{p+1} + O(h^{p+2}) , \tag{8.40}$$

dove $c_{p+1} y^{(p+1)}(t_n) h^{p+1}$ si dice *parte principale* di τ_{n+k} mentre c_{p+1} prende il nome di *costante di errore del metodo*. Dalle (8.39) e (8.40) discende il seguente teorema.

Teorema 8.3.1 *Un metodo a più passi è coerente se e solo se*

$$\rho(1) = 0, \quad \rho'(1) = \sigma(1) . \tag{8.41}$$

L'applicazione della (8.35) al problema test

$$y' = 0 , \quad y(t_0) = y_0 , \tag{8.42}$$

la cui soluzione è $y(t) = y_0$, dà luogo all'equazione lineare omogenea alle differenze

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = 0 \tag{8.43}$$

il cui polinomio caratteristico coincide con il primo polinomio caratteristico $\rho(\mu)$ del metodo stesso. La soluzione generale y_n della (8.43) è della forma (8.33) o (8.34) a seconda che le radici $\mu_1, \mu_2, \ldots, \mu_k$ di $\rho(\mu) = 0$ siano

distinte o meno. Si può verificare che la soluzione numerica della (8.43), per un insieme compatibile di valori iniziali, converge alla soluzione del problema continuo (8.42) solo se vale la seguente *condizione delle radici*

$$\begin{gathered} |\,\mu_i\,| \leq 1\,, \quad i = 1, 2, \ldots, k, \\ \text{dove, se } |\,\mu_i\,| = 1, \text{ allora } \mu_i \text{ è semplice.} \end{gathered} \tag{8.44}$$

Si conclude che un metodo della forma (8.35) non è convergente se gli zeri del suo polinomio $\rho(\mu)$ non soddisfano la condizione (8.44).

Definizione 8.3.3 *Un metodo lineare (8.35) si dice* zero-stabile *se gli zeri del suo polinomio $\rho(\mu)$ soddisfano la condizione (8.44).*

Si può dimostrare il seguente teorema.

Teorema 8.3.2 *Un metodo lineare a più passi è convergente se e solo se è coerente e zero-stabile.*

In linea di principio è possibile costruire metodi lineari a k passi fino ad un ordine massimo $p = 2k$ determinando le $2k+1$ costanti $\alpha_0, \ldots, \alpha_{k-1}, \beta_0, \ldots, \beta_k$ in modo che, in base alle (8.39), risulti $c_0 = c_1 = \cdots = c_{2k} = 0$: tuttavia tali metodi possono risultare non zero-stabili. Sussiste infatti il seguente teorema detto *prima barriera di Dahlquist.*

Teorema 8.3.3 *Non esistono metodi lineari a k passi zero-stabili di ordine superiore a $k+1$ se k è dispari e a $k+2$ se k è pari.*

Si riportano qui alcuni dei più noti metodi lineari con relative costanti d'errore.

I *metodi di Adams* sono caratterizzati da $\rho(\mu) = \mu^k - \mu^{k-1}$. Se $\beta_k = 0$ si ha la sottoclasse dei *metodi* espliciti *di Adams-Bashforth*; per $k = 1$ si ha il metodo di Eulero (costante di errore $c_2 = 1/2$), mentre per $k = 2$ risulta

$$y_{n+2} - y_{n+1} = \frac{h}{2}(3f_{n+1} - f_n)\,, \quad c_3 = \frac{5}{12}\,.$$

Se $\beta_k \neq 0$ si ha la sottoclasse dei *metodi* impliciti *di Adams-Moulton*; per $k = 1$ si ha la formula trapezoidale (costante di errore $c_3 = -1/12$), per $k = 2$ si ottiene

$$y_{n+2} - y_{n+1} = \frac{h}{12}(5f_{n+2} + 8f_{n+1} - f_n)\,, \quad c_4 = -\frac{1}{24}\,.$$

I *metodi BDF* (Backward Differentiation Formulae) hanno $\sigma(\mu) = \beta_k \mu^k$, sono di ordine $p = k$ e zero-stabili solo per $k \leq 6$. Per esempio, per $k = 1, 2, 3$ si ha rispettivamente

$$y_{n+1} - y_n = h f_{n+1} \,, \quad c_2 = -\frac{1}{2} \,, \tag{8.45}$$

$$y_{n+2} - \frac{4}{3} y_{n+1} + \frac{1}{3} y_n = h \frac{2}{3} f_{n+2} \,, \quad c_3 = -\frac{2}{3} \,,$$

$$y_{n+3} - \frac{18}{11} y_{n+2} + \frac{9}{11} y_{n+1} - \frac{2}{11} y_n = h \frac{6}{11} f_{n+3}, c_4 = -\frac{3}{22}.$$

La (8.45) è nota come *formula di Eulero implicita*.

Nella classe dei metodi con $\rho(\mu) = \mu^k - \mu^{k-2}$ quelli aventi $\beta_k = 0$ si dicono *metodi di Nyström*; per $k = 2$ si ottiene la *formula del punto centrale*

$$y_{n+2} - y_n = 2h f_{n+1} \,, \quad c_3 = \frac{1}{3} \,. \tag{8.46}$$

Se invece $\beta_k \neq 0$ si hanno i *metodi generalizzati di Milne-Simpson*; in particolare per $k = 2$ si ha la *regola di Simpson*

$$y_{n+2} - y_n = \frac{h}{3}(f_{n+2} + 4 f_{n+1} + f_n) \,, \quad c_5 = -\frac{1}{90} \,. \tag{8.47}$$

I *metodi di Newton-Cotes* hanno $\rho(\mu) = \mu^k - 1$; con $k = 4$ e $\beta_k = 0$ si ha, per esempio,

$$y_{n+4} - y_n = \frac{4}{3} h (2 f_{n+3} - f_{n+2} + 2 f_{n+1}) \,, \quad c_5 = \frac{28}{90} \,. \tag{8.48}$$

8.3.3 Metodi a più passi: assoluta stabilità

Si è visto che la coerenza e la zero-stabilità garantiscono la convergenza di un metodo lineare (cfr. Teorema 8.3.2), ma, essendo la convergenza una proprietà limite per $h \to 0$, può accadere che un metodo convergente, usato con un fissato passo $h > 0$, anche se piccolo, produca errori globali relativamente grandi.

Ciò accade per esempio se si applica al problema test (8.19) la formula del punto centrale (8.46), che pure è coerente e zero-stabile.

Occorre pertanto definire un concetto di stabilità che garantisca non solo la convergenza del metodo per $h \to 0$ ma anche il contenimento degli errori per un dato h.

Si consideri il caso scalare $m = 1$.

Il metodo (8.35) applicato al problema test (8.19) fornisce

$$\sum_{j=0}^{k}(\alpha_j - q\beta_j)y_{n+j} = 0\,. \tag{8.49}$$

Alla (8.49) è associato il polinomio caratteristico, detto *polinomio di stabilità* ,

$$\pi(q,\mu) = \rho(\mu) - q\sigma(\mu)\,. \tag{8.50}$$

Se $\mu_1(q), \mu_2(q), \ldots, \mu_k(q)$ sono le radici di $\pi(q,\mu) = 0$, che si suppone si mantengano semplici, la soluzione generale della (8.49) è data da

$$y_n = \sum_{i=1}^{k} d_i \mu_i^n(q)\,, \quad n = 0, 1, \ldots, \tag{8.51}$$

dove d_i, $i = 1, 2, \ldots, k$, sono costanti arbitrarie.

D'altra parte, ponendo nella (8.49) il valore esatto $y(t_{n+j})$ al posto di y_{n+j}, per definizione di errore di troncamento locale (vedi la (8.37)) si ha

$$\sum_{j=0}^{k}(\alpha_j - q\beta_j)y(t_{n+j}) = \tau_{n+k}\,;$$

sottraendo membro a membro da questa equazione la (8.49) si ottiene

$$\sum_{j=0}^{k}(\alpha_j - q\beta_j)e_{n+j} = \tau_{n+k}$$

che può ritenersi un caso perturbato della (8.49). Ne segue che l'errore globale di discretizzazione $e_n = y(t_n) - y_n$ ha un andamento analogo a quello di y_n e si può quindi affermare che e_n non cresce, al crescere di n, per i valori di q per cui risulta

$$|\,\mu_i(q)\,| < 1\,, \quad i = 1, 2, \ldots, k. \tag{8.52}$$

Definizione 8.3.4 *Un metodo della forma (8.35) si dice* assolutamente stabile *per un dato $q \in \mathbb{C}$, se gli zeri del suo polinomio di stabilità $\pi(q,\mu)$ verificano la condizione (8.52).*

Definizione 8.3.5 *L'insieme del piano complesso*

$$S_A = \{q \in \mathbb{C} \;|\; |\,\mu_i(q)\,| < 1,\ i = 1, 2, \ldots, k\}$$

si chiama regione di assoluta stabilità *del metodo lineare a più passi.*

S_A è simmetrico rispetto all'asse reale: infatti se μ^* è tale che $\pi(q, \mu^*) = 0$, per la (8.50), risulta anche $\pi(\bar{q}, \bar{\mu}^*) = 0$.

Si osservi che si ha $\lim_{h\to 0} \pi(q, \mu) = \pi(0, \mu) = \rho(\mu)$; pertanto le radici caratteristiche $\mu_i(q)$, $i = 1, 2, \ldots, k$, per $h \to 0$ tendono alle radici di $\rho(\mu) = 0$. D'altra parte, una di queste radici, per le condizioni di coerenza (8.41), è uguale a 1. Quindi una delle radici $\mu_i(q)$, che si indicherà con $\mu_1(q)$, tende a 1 per $h \to 0$ e, a causa della condizione (8.44), essa è unica. A $\mu_1(q)$ si dà il nome di *radice principale* di $\pi(q, \mu) = 0$, perché è quella che nella (8.51) approssima la soluzione $y(t) = de^{\lambda t}$ del problema test.

Le altre radici $\mu_2(q), \ldots, \mu_k(q)$ si chiamano *radici spurie* o *parassite* perché nascono dalla sostituzione dell'equazione differenziale, del primo ordine, con l'equazione alle differenze (8.49), di ordine k, e il loro effetto è solo quello di accrescere l'errore.

Per la (8.37) ove si ponga $y(t) = e^{\lambda t}$ ed $f(t, y) = \lambda e^{\lambda t}$, si ottiene

$$\tau_{n+k} = e^{\lambda t_n} \pi(q, e^q) = O(h^{p+1}) .$$

D'altra parte vale l'identità

$$\pi(q, \mu) = (1 - q\beta_k)(\mu - \mu_1(q))(\mu - \mu_2(q)) \cdots (\mu - \mu_k(q));$$

ne segue quindi

$$\pi(q, e^q) = (1 - q\beta_k)(e^q - \mu_1(q))(e^q - \mu_2(q)) \cdots (e^q - \mu_k(q)) = O(h^{p+1}) .$$

Quando $q \to 0$ l'unico fattore di $\pi(q, e^q)$ che tende a zero, per quanto detto sopra, è $e^q - \mu_1(q)$: si ha pertanto

$$\mu_1(q) = e^q + O(h^{p+1}) . \tag{8.53}$$

Dalla (8.53) segue che se $Re(q) > 0$, ed h è sufficientemente piccolo, risulta $| \mu_1(q) | > 1$ e quindi $q \notin S_A$, cioè l'intersezione di S_A con l'asse reale è del tipo $]\alpha, 0[$ con $\alpha < 0$.

In Fig. 8.2 sono riportate le regioni S_A delle formule BDF con $k = 1, 2, 3$, e sono costituite dai punti esterni alle porzioni di piano delimitate dalle curve date per ogni valore di k.

Nel teorema seguente sono riuniti i risultati più significativi della teoria della stabilità per i metodi a più passi, avvertendo che anche per questi metodi le definizioni di A_0-stabilità e di A-stabilità sono le medesime date per i metodi di Runge-Kutta (cfr. Definizioni 8.2.7 e 8.2.8).

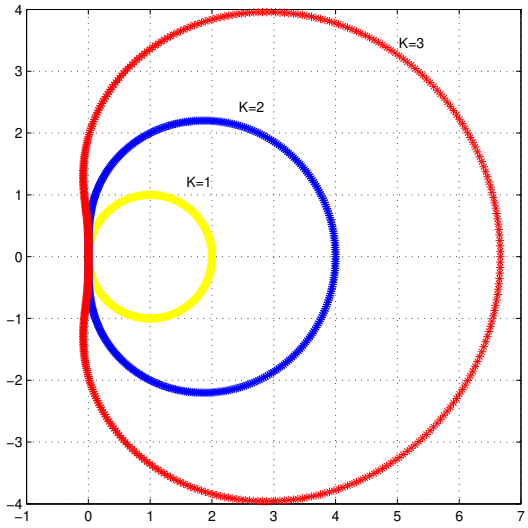

Figura 8.2: Regioni S_A delle prime tre fomule BDF.

Teorema 8.3.4 (di Dahlquist)

Non esistono metodi lineari a più passi espliciti A-stabili;

*l'ordine massimo di un metodo lineare a più passi A-stabile non può superare due (*seconda barriera di Dahlquist*);*

la formula trapezoidale è il metodo lineare A-stabile del secondo ordine con la costante di errore più piccola in modulo.

Successivamente la prima proposizione di questo teorema è stata ulteriormente precisata come segue.

Teorema 8.3.5 *Non esistono metodi lineari a più passi espliciti A_0-stabili.*

La definizione di assoluta stabilità data nel caso $m = 1$, si estende in modo naturale al caso $m > 1$ considerando il problema test (8.26): si dice ora che un metodo a più passi applicato al problema (8.26) è assolutamente stabile per un dato insieme di valori q_i, $i = 1, 2, \ldots, m$, se e solo se

$$q_i \in S_A\,, \quad i = 1, 2, \ldots, m,$$

dove S_A è la regione di assoluta stabilità introdotta con la Definizione 8.3.5.

8.3.4 Metodi di predizione e correzione

In 8.3.2 si è visto che l'uso di un metodo (8.35) di tipo implicito richiede, ad ogni passo, la risoluzione dell'equazione (8.36), che può essere non lineare. In tal caso l'incognita y_{n+k} viene approssimata con il processo iterativo

$$z^{(r+1)} = h\beta_k f(t_{n+k}, z^{(r)}) + w\ , \quad r = 0, 1, \dots\ .$$

È chiaro che, ammesso che vi sia convergenza, quanto più vicina alla soluzione si sceglie l'approssimazione iniziale $z^{(0)}$, tanto più piccolo è il numero delle iterazioni necessarie per ottenere una certa accuratezza.

Su quest'idea si basano i cosiddetti *metodi di predizione e correzione*, nei quali si sceglie $z^{(0)} = y^*_{n+k}$ dove y^*_{n+k} si calcola in precedenza con un metodo esplicito.

In questo contesto il metodo esplicito viene detto *predittore* mentre il metodo implicito prende il nome di *correttore*. In quello che segue si fa riferimento, per semplicità, al caso in cui predittore e correttore hanno lo stesso ordine; inoltre si suppone che il correttore venga usato una sola volta ad ogni passo.

In queste ipotesi, la struttura generale di un metodo di predizione e correzione è quindi data da

$$\begin{aligned} y^*_{n+k} &= -\sum_{j=0}^{k-1} \alpha^*_j y_{n+j} + h \sum_{j=0}^{k-1} \beta^*_j f_{n+j}\ , \\ y_{n+k} &= -\sum_{j=0}^{k-1} \alpha_j y_{n+j} + h \sum_{j=0}^{k-1} \beta_j f_{n+j} + h\beta_k f^*_{n+k}\ , \end{aligned} \tag{8.54}$$

dove si è posto $f^*_{n+k} = f(t_{n+k}, y^*_{n+k})$.

Si osservi che un metodo di predizione e correzione nella forma (8.54) ha un costo computazionale non superiore a quello del correttore usato da solo iterativamente.

L'algoritmo (8.54) viene designato con la sigla PEC per indicare le varie fasi che costituiscono un passo, cioè

P (Prediction): calcolo di y^*_{n+k} mediante il predittore,

E (Evaluation): valutazione del termine $f(t_{n+k}, y^*_{n+k})$,

C (Correction): calcolo di y_{n+k} mediante il correttore.

Al passo successivo si utilizza $f(t_{n+k}, y^*_{n+k})$ nel predittore per dare corso alla nuova fase P.

Una variante è costituita dall'algoritmo $PECE$ in cui compare l'ulteriore fase

E: valutazione del termine $f(t_{n+k}, y_{n+k})$;

in tal caso, nella fase P del passo successivo, il predittore utilizza la valutazione $f(t_{n+k}, y_{n+k})$ che, di solito, è più corretta della $f(t_{n+k}, y^*_{n+k})$.

Anche la stabilità di un metodo di predizione e correzione si studia mediante la sua applicazione al problema test (8.19) e, in generale, è diversa da quella del predittore e del correttore supposti usati singolarmente. Si consideri, ad esempio, il seguente metodo di predizione e correzione in cui il predittore è la formula del punto centrale (8.46) e il correttore è la formula trapezoidale

$$\begin{aligned} y^*_{n+2} &= y_n + 2hf_{n+1}\,, \\ y_{n+2} &= y_{n+1} + \frac{h}{2}(f_{n+1} + f^*_{n+2})\,. \end{aligned} \tag{8.55}$$

La formula del punto centrale non è assolutamente stabile mentre la formula trapezoidale è A-stabile.
Applicato al problema test, il metodo (8.55) diviene

$$\begin{aligned} y^*_{n+2} &= y_n + 2qy_{n+1}\,, \\ y_{n+2} &= y_{n+1} + \frac{1}{2}qy_{n+1} + \frac{1}{2}qy^*_{n+2}\,, \end{aligned}$$

da cui, eliminando y^*_{n+2}, si ottiene

$$y_{n+2} - \left(1 + \frac{1}{2}q + q^2\right) y_{n+1} - \frac{1}{2}qy_n = 0\,.$$

Il metodo (8.55) è dotato di una regione non vuota di assoluta stabilità come si verifica dalla sua regione S_A in Fig 8.3 costituita dai punti interni alla porzione di piano delimitata dalla curva data.

Particolare importanza hanno i metodi di predizione e correzione costruiti con due formule dello stesso ordine: in tal caso si può avere ad ogni passo una buona stima della parte principale dell'errore locale di troncamento sia del

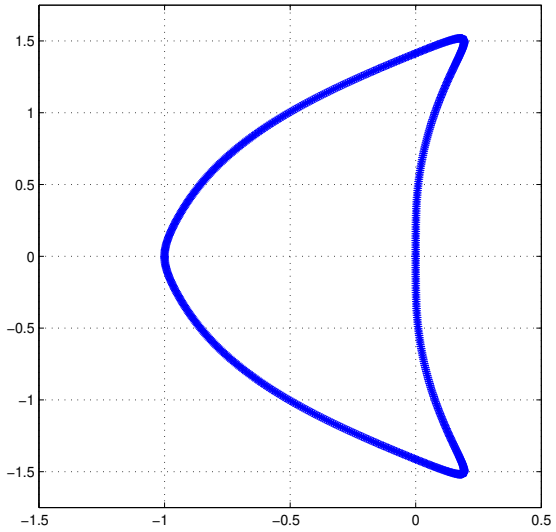

Figura 8.3: Regione S_A del metodo (8.55).

predittore che del correttore. Basandosi sull'assunto che nei membri destri delle (8.54) sia

$$y_{n+j} = y(t_{n+j}) \,, \quad j = 0, 1, \ldots, k-1,$$

tenuto conto della definizione (8.37) di errore locale di troncamento e della (8.40), può scriversi, per un predittore e un correttore entrambi di ordine p,

$$\begin{aligned} y(t_{n+k}) &= y^*_{n+k} + c^*_{p+1} y^{(p+1)}(t_n) h^{p+1} + O(h^{p+2}) \,, \\ y(t_{n+k}) &= y_{n+k} + c_{p+1} y^{(p+1)}(t_n) h^{p+1} + O(h^{p+2}) \,. \end{aligned} \tag{8.56}$$

Dalle precedenti uguaglianze si ricava la relazione

$$c_{p+1} y^{(p+1)}(t_n) h^{p+1} = \frac{c_{p+1}}{c^*_{p+1} - c_{p+1}} (y_{n+k} - y^*_{n+k}) + O(h^{p+2}) \tag{8.57}$$

che fornisce la detta stima per il correttore e richiede soltanto la conoscenza del valore predetto e del valore corretto.

Eliminando $y^{(p+1)}(t_n)$ tra la (8.57) e la (8.56) si ottiene una ulteriore approssimazione di $y(t_{n+k})$ data da

$$\hat{y}_{n+k} = y_{n+k} + \frac{c_{p+1}}{c^*_{p+1} - c_{p+1}} (y_{n+k} - y^*_{n+k}) \,, \tag{8.58}$$

e si ha $y(t_{n+k}) - \hat{y}_{n+k} = O(h^{p+2})$.

La (8.58), detta *correzione di Milne* o *estrapolazione locale*, può essere una ulteriore relazione in un metodo di predizione e correzione PEC o $PECE$: ad esempio, usando come predittore la (8.48) e come correttore la (8.47), si ottiene il seguente metodo

$$\begin{aligned} y^*_{n+4} &= y_n + \frac{4}{3}h(2f_{n+1} - f_{n+2} + 2f_{n+3}) \,, \\ y_{n+4} &= y_{n+2} + \frac{h}{3}(f_{n+2} + 4f_{n+3} + f^*_{n+4}) \,, \\ \hat{y}_{n+4} &= y_{n+4} - \frac{1}{29}(y_{n+4} - y^*_{n+4}) \,. \end{aligned} \tag{8.59}$$

La stima che si ottiene dalla (8.57) può servire anche per controllare ad ogni passo l'entità dell'errore locale, ai fini di una strategia di variazione del passo. In tal caso si riduce il passo se l'errore è troppo grande e si aumenta in caso contrario.

8.3.5 Metodi a più passi: stabilità relativa

Per tutti i metodi esposti in questo capitolo, l'assoluta stabilità viene definita per problemi con soluzioni $y(t)$ tali che $\lim_{t\to+\infty} y(t) = 0$, essendo questo tipo di problema molto frequente nelle applicazioni. Da qui l'uso di problemi test della forma

$$y' = \lambda y \,, \quad \text{o} \quad y' = Ky$$

con le rispettive condizioni

$$Re(\lambda) < 0 \,, \quad \text{e} \quad Re(\lambda_i) < 0 \,,\ i = 1, 2, \ldots, m.$$

Per i problemi con soluzione $y(t)$ tale che $\lim_{t\to+\infty} \|y(t)\| = +\infty$, si fa riferimento agli stessi problemi test, dove ora si assume rispettivamente

$$Re(\lambda) > 0 \,, \quad \text{e} \quad Re(\lambda_j) > 0 \,, \text{ per qualche } \ j \in \{1, 2, \ldots, m\}.$$

In tal caso si richiede, per la soluzione numerica, che sia $\lim_{n\to\infty} \|y_n\| = +\infty$ e a tale scopo si introduce il concetto di *stabilità relativa*, per significare che l'errore relativo $\frac{\|y_n - y(t_n)\|}{\|y(t_n)\|}$ si mantiene "accettabile" (cioè piccolo rispetto a 1) al crescere di n. Per esempio, nel caso dei metodi lineari e per $m = 1$, si propone, di solito, la seguente definizione.

Definizione 8.3.6 *Il metodo lineare (8.35) si dice* relativamente stabile *per un dato $q \in \mathbb{C}$, se, applicato al problema test $y' = \lambda y$, $Re(\lambda) > 0$, gli zeri del suo polinomio di stabilità $\pi(q, \mu)$ verificano le condizioni*

$$| \mu_i(q) | < | \mu_1(q) | , \quad i = 2, 3, \ldots, k,$$

dove $\mu_1(q)$ è la radice principale di $\pi(q, \mu) = 0$.

Definizione 8.3.7 *L'insieme del piano complesso*

$$S_R = \{q \in \mathbb{C} \ \ | \ | \mu_i(q) | < | \mu_1(q) |, \ i = 2, 3, \ldots, k\}$$

si dice regione di relativa stabilità *del metodo (8.35).*

Si consideri, ad esempio, la formula del punto centrale (8.46): il suo polinomio di stabilità è

$$\pi(q, \mu) = \mu^2 - 2q\mu - 1 ,$$

i cui zeri sono $\mu_1(q) = q + \sqrt{q^2 + 1}$ e $\mu_2 = q - \sqrt{q^2 + 1}$. Poiché $| \mu_1 \mu_2 | = 1$, la formula non ha una regione di assoluta stabilità. Tuttavia, limitandosi per semplicità al caso $q \in \mathbb{R}$, risulta $\mu_2(q) < \mu_1(q)$ se $q > 0$: essa, quindi, è dotata di un intervallo reale di relativa stabilità coincidente con il semiasse positivo.

Si conclude questa rassegna di metodi per problemi di valori iniziali osservando che le proprietà di stabilità qui riportate fanno parte della cosiddetta teoria della stabilità lineare, in cui si fa riferimento a problemi test lineari a coefficienti costanti, ma se ne estendono i risultati a problemi più generali.

Anche se la stabilità lineare è un requisito necessario, tuttavia essa può rivelarsi inadeguata quando un metodo, stabile in senso lineare, venga applicato ad un problema (8.1) con $f(t, y)$ non lineare, o anche lineare con la matrice K dipendente da t oppure non diagonalizzabile. In questi casi sarebbe più appropriata una forma di *stabilità non lineare*, la cui trattazione va oltre gli scopi del presente testo ed è reperibile in opere specializzate.

8.4 Problemi ai limiti e BV-metodi

8.4.1 Introduzione

Nell'Esempio 6.8.8 si è considerato un problema continuo per una equazione differenziale del secondo ordine, che differisce da un problema di Cauchy

per il fatto che la soluzione è assoggettata, invece che ad una condizione in un punto iniziale, a due condizioni poste agli estremi di un intervallo. Indipendentemente dall'ordine dell'equazione differenziale, che può sempre ricondursi ad un sistema del primo ordine (cfr. 8.1), i problemi continui con condizioni che coinvolgono i valori della soluzione in più punti si dicono *problemi ai limiti*.

La discretizzazione di questi problemi si può fare mediante un *problema ai limiti discreto*, cioè costituito da una equazione alle differenze di ordine k a cui sono associate condizioni su valori iniziali e valori finali della soluzione.

Più in generale, un problema ai limiti discreto si può sempre ottenere da un problema continuo, sia ai limiti che di valori iniziali, applicando una formula lineare a k passi del tipo (8.35)

$$\sum_{j=0}^{k} \alpha_j y_{n+j} - h \sum_{j=0}^{k} \beta_j f_{n+j} = 0 \,. \tag{8.60}$$

In tal caso la formula prende il nome di BV-*metodo* o, più brevemente, BVM (Boundary Value Method). Per contro quando la (8.60) è usata come nei paragrafi precedenti (cioè quando trasforma un problema continuo di valori iniziali in un problema discreto di valori iniziali), allora viene indicata con la sigla IVM (Initial Value Method). Nel caso dei problemi di valori iniziali, l'uso dei BVM presenta alcuni notevoli vantaggi rispetto ai tradizionali IVM. Per esempio il superamento della seconda barriera di Dahlquist (cfr. Teorema 8.3.4) e la maggiore facilità di stima dell'errore globale. I due usi della (8.60) come IVM e come BVM sono schematizzati nella Tavola 8.4. Si noti che esistono anche metodi che riconducono un problema continuo ai limiti ad un equivalente problema di valori iniziali che poi viene discretizzato con un IVM. Tali metodi, che qui non sono considerati, sono noti come *metodi shooting*.

Problema continuo	Metodo	Problema discreto
Valori iniziali	$\xrightarrow{\text{IVM}}$	Valori iniziali
	$\searrow^{\text{BVM}}$	
Valori ai limiti	$\xrightarrow{\text{BVM}}$	Valori ai limiti

Tavola 8.4: Caratterizzazione degli IVM e dei BVM.

8.4.2 Modo di impiego dei BVM

Di solito una formula del tipo (8.60) usata come BVM si dice *metodo base* e viene associata ad altre formule dello stesso tipo, generalmente implicite e con un numero di passi minore, che si dicono *metodi ausiliari*. Per delineare come si impiegano tali formule, si considererà il problema continuo di valori iniziali: il caso di un problema di valori ai limiti necessita solo di poche e semplici modifiche.
Si considera un passo di discretizzazione costante $h > 0$ sull'intervallo $[a, b]$. Se $t_0 = a$, $t_N = b$ e $t_n = t_0 + nh$, $n = 0, 1, \ldots, N$, si assume $h = (t_N - t_0)/N$. Applicando al problema il metodo base (8.60), per $n = 0, 1, \ldots, N - k$, si scrivono $N - k + 1$ relazioni indipendenti.

Queste $N - k + 1$ relazioni si possono considerare come un sistema nelle N incognite $y_1, \ldots, y_N$. Vi sono quindi ancora $k - 1$ incognite del problema discreto che devono essere determinate e ciò si può fare aggiungendo al sistema altrettanti metodi ausiliari. Se $f(t, y)$ è lineare rispetto a y tale è anche il sistema che così si ottiene: la soluzione esiste ed è unica per l'indipendenza lineare delle equazioni che lo costituiscono. Tale soluzione si assume come approssimazione discreta della soluzione del problema continuo (cfr. 8.5.5)

Se il problema continuo non è lineare, il procedimento ora descritto rimane valido, ma conduce ad un sistema discreto non lineare. In tal caso si può dimostrare, sotto opportune ipotesi, l'esistenza di una soluzione unica del problema discreto e la convergenza delle iterazioni di Newton (cfr. 4.6).

In particolare il sistema viene strutturato come segue.
Alle $N - k + 1$ equazioni ottenute dal metodo base vengono aggiunte $k - 1$ equazioni date da altrettanti metodi ausiliari. Tali metodi, per motivi che si chiariranno meglio nel paragrafo successivo, sono generalmente organizzati in due gruppi: $k_1 - 1$ metodi ausiliari "di testa" da porre come equazioni iniziali del sistema e k_2 metodi ausiliari "di coda" da porre come equazioni finali, con $k = k_1 + k_2$. Pertanto il sistema di N equazioni nelle N incognite $y_1, \ldots, y_N$ assume la forma seguente:

$$\begin{cases} \sum_{j=0}^{r} \alpha_{j\nu} y_j - h \sum_{j=0}^{r} \beta_{j\nu} f_j = 0, & \nu = 1, \ldots, k_1 - 1, \\ \sum_{j=0}^{k} \alpha_j y_{\nu+j-k_1} - h \sum_{j=0}^{k} \beta_j f_{\nu+j-k_1} = 0, & \nu = k_1, \ldots, N - k_2, \\ \sum_{j=0}^{s} \alpha_{j\nu} y_{N+j-s} - h \sum_{j=0}^{s} \beta_{j\nu} f_{N+j-s} = 0, & \nu = N - k_2 + 1, \ldots, N. \end{cases} \tag{8.61}$$

Si noti che nelle prime $k_1 - 1$ equazioni ausiliarie del sistema sono implicate le incognite $y_1, \ldots, y_r$, nelle equazioni del metodo di base le incognite $y_1, \ldots, y_N$, mentre nelle ultime k_2 equazioni ausiliarie le incognite $y_{N-s}, \ldots, y_N$. Tale sistema costituisce un *metodo BVM con* (k_1, k_2)*-condizioni al contorno* o, più semplicemente $\mathrm{BVM}_{k_1 k_2}$.
Poiché, in generale, è $N \gg k$, si può provare che le caratteristiche di convergenza e di stabilità del metodo (8.61) sono essenzialmente regolate dal metodo base il cui ordine p, per ovvi motivi di omogeneità, è il medesimo delle formule ausiliarie. Pertanto, in quello che segue, $\rho(\mu)$, $\sigma(\mu)$, $\pi(q, \mu)$ sono i polinomi del metodo base.

8.4.3 Stabilità e convergenza dei BVM

Le condizioni che si richiedono per la stabilità di un $\mathrm{BVM}_{k_1 k_2}$ sono alquanto diverse da quelle già viste per un IVM.

Si premettono alcune definizioni.

Definizione 8.4.1 *Sia* $k = k_1 + k_2 + k_3$ *con* k_1, k_2, k_3 *interi non negativi. Il polinomio a coefficienti reali*

$$p(\mu) = \sum_{j=0}^{k} a_j \mu^j$$

si dice del tipo $(\mathsf{k}_1,\mathsf{k}_2,\mathsf{k}_3)$ *se ha* k_1 *zeri in modulo minori di* 1, k_2 *zeri in modulo uguali a* 1 *e* k_3 *zeri in modulo maggiori di* 1.
Sono polinomi di Schur *quelli del tipo* $(k, 0, 0)$, *sono* polinomi di Von Neumann *quelli del tipo* $(k_1, k-k_1, 0)$ *con i* $k-k_1$ *zeri di modulo* 1 *tutti semplici, mentre si dicono* polinomi conservativi *quelli del tipo* $(0, k, 0)$.

Definizione 8.4.2 *Sia* $k = k_1 + k_2$ *con* k_1, k_2 *interi non negativi. Il polinomio a coefficienti reali*

$$p(\mu) = \sum_{j=0}^{k} a_j \mu^j$$

si dice un $\mathsf{S}_{k_1 k_2}$-polinomio *se per i suoi zeri risulta*

$$|\mu_1| \leq |\mu_2| \leq \ldots |\mu_{k_1}| < 1 < |\mu_{k_1+1}| \leq \ldots \leq |\mu_k|,$$

mentre si dice un $\mathsf{N}_{k_1 k_2}$-polinomio *se*

$$|\mu_1| \leq |\mu_2| \leq \ldots |\mu_{k_1}| \leq 1 < |\mu_{k_1+1}| \leq \ldots \leq |\mu_k|$$

e gli zeri di modulo unitario sono semplici.

Si osservi che un $S_{k_1k_2}$-polinomio è del tipo $(k_1, 0, k - k_1)$ e quindi un S_{k0}-polinomio è un polinomio di Schur, mentre un N_{k0}-polinomio è un polinomio di Von Neumann.
Di conseguenza un metodo lineare a k passi usato come IVM è zero-stabile se il suo polinomio caratteristico $\rho(\mu)$ è un polinomio di Von Neumann (ovvero un N_{k0}-polinomio) mentre è assolutamente stabile se $\pi(q, \mu)$ è un polinomio del tipo $(k, 0, 0)$ (ovvero un polinomio di Schur, ovvero ancora un S_{k0}-polinomio).

Per la zero-stabilità di un IVM si è fatto riferimento al problema test $y' = 0, \quad y(t_0) = y_0$. Ma lo stesso risultato si ottiene assumendo in un metodo a k passi $h = 0$. In questo senso la zero-stabilità è una stabilità asintotica per h che tende a 0. Un analogo concetto può essere dato nel caso di un $BVM_{k_1k_2}$. Precisamente vale il seguente risultato.

Teorema 8.4.1 *Un metodo $BVM_{k_1k_2}$ è convergente nel senso che*

$$\| y(t_n) - y_n \| = O(h^p)$$

se $\rho(1) = 0$, $\rho'(1) = \sigma(1)$ (coerenza) e se $\rho(\mu)$ è un $N_{k_1k_2}$-polinomio.

Ciò conduce alla seguente definizione.

Definizione 8.4.3 *Un $BVM_{k_1k_2}$ si dice* $0_{k_1k_2}$-stabile *se il corrispondente polinomio $\rho(z)$ è un $N_{k_1k_2}$-polinomio.*

Quindi coerenza e $0_{k_1k_2}$-stabilità sono condizioni sufficienti, ma non anche necessarie come nel caso di un IVM (cfr. Teorema 8.3.2), per la convergenza di un $BVM_{k_1k_2}$.

Si esamina ora la assoluta stabilità di un $BVM_{k_1k_2}$, cioè nel caso di h fisso.

Si consideri il problema test

$$y'(t) = \lambda y(t),\ y(t_0) = y_0,\ \lambda \in \mathbb{C},\ Re(\lambda) < 0$$

già utilizzato per studiare l'assoluta stabilità di un metodo di Runge-Kutta e di un IVM.
Il problema, quindi, è quello di approssimare con un $BVM_{k_1k_2}$ la soluzione $y(t_n) = y_0 e^{\lambda(nh)} = y_0(e^q)^n$.
Vale il seguente teorema.

Teorema 8.4.2 *Si supponga che per gli zeri del polinomio $\pi(q,\mu)$ si abbia*

$$\begin{aligned} |\mu_1(q)| \leq \ldots \leq |\mu_{k_1-1}(q)| &< |\mu_{k_1}(q)| < |\mu_{k_1+1}(q)| \leq \ldots \leq |\mu_k(q)| \\ |\mu_{k_1-1}(q)| &< 1 < |\mu_{k_1+1}(q)| \end{aligned}$$

con $\mu_{k_1}(0) = \mu_{k_1} = 1$, ovvero si ammetta che $\mu_{k_1}(q)$ sia radice principale, tale quindi da avere $\mu_{k_1}(q) = e^q + O(h^{p+1})$, allora si ha

$$y_n = \mu_{k_1}^n(q)(y_0 + O(h^p)) + O(h^p)(O(|\mu_{k_1-1}(q)|^n) + O(|\mu_{k_1+1}(q)|^{-(N-n)})).$$

Questo risultato conduce alla seguente definizione di *Assoluta stabilità* per un $\mathrm{BVM}_{k_1k_2}$

Definizione 8.4.4 *Un $BVM_{k_1k_2}$ si dice (k_1,k_2)-*Assolutamente stabile *per un dato q se il polinomio $\pi(q,\mu)$ è del tipo $(k_1,0,k_2)$, cioè un $S_{k_1k_2}$-polinomio.*

Definizione 8.4.5 *La regione del piano complesso*

$$S_{A_{k_1k_2}} = \{q \in \mathbb{C} \,|\pi(q,\mu) \text{è del tipo } (k_1,0,k_2)\}$$

*si dice regione di (k_1,k_2)-*Assoluta stabilità.

Si noti che se $k_2 = 0$ queste definizioni coincidono con quelle ordinarie per IVM.

Definizione 8.4.6 *Un $BVM_{k_1k_2}$ si dice* $\mathsf{A}_{k_1k_2}$-stabile *se*

$$S_{A_{k_1k_2}} \supseteq \{q \in \mathbb{C} \,|Re(q) < 0\}.$$

Per illustrare come una medesima formula possa avere comportamenti diversi se usata come IVM o $\mathrm{BVM}_{k_1k_2}$ si consideri la formula del punto centrale (8.46) per la quale risulta $\rho(\mu) = \mu^2 - 1$ e $\pi(q,\mu) = \mu^2(q) - 2q\mu(q) - 1$. Come già notato è zero-stabile essendo $\rho(\mu)$ un polinomio di Von Neumann, ma S_A è vuoto avendosi in ogni caso $|\mu_1(q)| > 1$ oppure $|\mu_2(q)| > 1$: tale formula non può essere usata da sola come IVM. Inoltre essa non è 0_{11}-stabile perché $\rho(\mu)$ non è un N_{11}-polinomio (dove è richiesto $|\mu_1| < |\mu_2|$). Tuttavia per $Re(q) < 0$ risulta $|\mu_1(q)| < 1 < |\mu_2(q)|$ e quindi $\pi(q,\mu)$ è un S_{11}-polinomio: per $Re(q) < 0$, quindi, usata con una equazione ausiliaria come BVM_{11} fornisce un metodo A_{11}-stabile (vedi paragrafo 8.5.5).

Come già accennato, uno degli aspetti più importanti dei BVM consiste nell'esistenza di metodi $\mathrm{A}_{k_1k_2}$-stabili di ordine $2k$, cioè del massimo ordine

possibile per formule lineari a k passi: ciò esclude di fatto ogni barriera d'ordine per metodi $\mathrm{BVM}_{k_1k_2}$ stabili.

Di seguito si forniscono tre esempi di metodi $\mathrm{A}_{k_1k_2}$-stabili: altri possono essere trovati nel testo *Brugnano-Trigiante* [4].

Metodo ETR (Extended Trapezoidal Rule),
p=4, $k_1 = 2, k_2 = 1$:

$$\begin{cases} y_1 - y_0 = \frac{h}{24}(f_3 - 5f_2 + 19f_1 + 9f_0), \\ \\ y_n - y_{n-1} = \frac{h}{24}(-f_{n+1} + 13f_n + 13f_{n-1} - f_{n-2}), \quad n = 2, \ldots, N-1, \\ \\ y_N - y_{N-1} = \frac{h}{24}(f_{N-3} - 5f_{N-2} + 19f_{N-1} + 9f_N). \end{cases}$$

Metodo GAM (Generalized Adams Method),
p=5, $k_1 = 2, k_2 = 2$:

$$\begin{cases} y_1 - y_0 = \frac{h}{720}(-19f_4 + 106f_3 - 264f_2 + 646f_1 + 251f_0), \\ \\ y_n - y_{n-1} = \frac{h}{720}(-19f_{n-2} + 346f_{n-1} + 456f_n - 74f_{n+1} + 11f_{n+2}), \\ \qquad\qquad n = 2, \ldots, N-1, \\ \\ y_N - y_{N-1} = \frac{h}{720}(-19f_{N+1} + 346f_N + 456f_{N-1} - 74f_{N-2} + 11f_{N-3}), \\ \\ y_{N+1} - y_N = \frac{h}{720}(-19f_{N-3} + 106f_{N-2} - 264f_{N-1} + 646f_N + 251f_{N+1}). \end{cases}$$

Metodo TOM (Top Order Method),
p=6, $k_1 = 2, k_2 = 1$:

$$\begin{cases} y_1 - y_0 = \frac{h}{1440}(27f_5 - 173f_4 + 482f_3 - 798f_2 + 1427f_1 + 475f_0), \\ \\ \frac{1}{60}(11y_{n+1} + 27y_n - 27y_{n-1} - 11y_{n-2}) = \\ \qquad\qquad \frac{h}{20}(f_{n+1} + 9f_n + 9f_{n-1} + f_{n-2}), \quad n = 2, \ldots, N-1, \\ \\ y_N - y_{N-1} = \\ \quad \frac{h}{1440}(27f_{N-5} - 173f_{N-4} + 482f_{N-3} - 798f_{N-2} + 1427f_{N-1} + 475f_N). \end{cases}$$

Si osservi che la formula trapezoidale

$$y_{n+1} - y_n = \frac{h}{2}(f_{n+1} + f_n),$$

che è A-stabile se usata come IVM, risulta essere A_{10}-stabile come BVM, cioè è un metodo BVM_{10}.
Infatti il suo polinomio di stabilità ha un unico zero $|\mu_1(q)| < 1$ per $Re(q) < 0$. Essa, pertanto, può essere usata da sola, cioè senza formule ausiliarie, con $n = 0, 1, \ldots, N-1$, avendosi $k_1 - 1 = 0$ e $k_2 = 0$.

8.5 Complementi ed esempi

8.5.1 I metodi di Runge-Kutta impliciti come metodi di collocazione

Il *metodo di collocazione* per un problema di valori iniziali della forma (8.1) consiste nel determinare un polinomio di grado s, con coefficienti in $\mathbb{R}^m$, che approssimi la soluzione sull'intervallo $[t_n, t_n + h]$.

Dati i numeri reali due a due distinti $c_1, c_2, \ldots, c_s \in [0, 1]$, il corrispondente *polinomio di collocazione* $u(t)$, di grado s, è definito univocamente dalle condizioni

$$u(t_n) = y_n \,, \tag{8.62}$$

$$u'(t_n + c_i h) = f(t_n + c_i h, u(t_n + c_i h)), i = 1, 2, \ldots, s \,. \tag{8.63}$$

Si assume come soluzione numerica della equazione differenziale nel punto t_{n+1} il valore

$$y_{n+1} = u(t_n + h) \,. \tag{8.64}$$

Il metodo di collocazione (8.62)-(8.63) è equivalente ad un metodo di Runge-Kutta implicito ad s stadi ove si ponga

$$a_{ij} = \int_0^{c_i} l_j(\tau) d\tau \,, \quad b_j = \int_0^1 l_j(\tau) d\tau \,, \quad i, j = 1, 2, \ldots, s, \tag{8.65}$$

essendo

$$l_j(\tau) = \frac{(\tau - c_1) \cdots (\tau - c_{j-1})(\tau - c_{j+1}) \cdots (\tau - c_s)}{(c_j - c_1) \cdots (c_j - c_{j-1})(c_j - c_{j+1}) \cdots (c_j - c_s)} \,, \; j = 1, 2, \ldots, s,$$

i polinomi fondamentali della interpolazione di Lagrange relativi ai punti $c_1, c_2, \ldots, c_s$. Infatti, poiché il polinomio $u'(t_n + \tau h)$ si può scrivere come polinomio di interpolazione relativo ai detti punti, posto

$$k_i = u'(t_n + c_i h) \,, \tag{8.66}$$

si ha (cfr. 6.2)

$$u'(t_n + \tau h) = \sum_{j=1}^{s} l_j(\tau) k_j \, ;$$

integrando entrambi i membri, rispetto a t, sugli intervalli $[t_n, t_n + c_i h]$, $i = 1, 2, \ldots, s$ e $[t_n, t_{n+1}]$, si ottiene rispettivamente

$$u(t_n + c_i h) = u(t_n) + h \sum_{j=1}^{s} \left(\int_0^{c_i} l_j(\tau) d\tau \right) k_j \, , \quad i = 1, 2, \ldots, s, \tag{8.67}$$

e

$$u(t_n + h) = u(t_n) + h \sum_{j=1}^{s} \left(\int_0^{1} l_j(\tau) d\tau \right) k_j \, . \tag{8.68}$$

La (8.68) si può anche scrivere, tenendo conto delle (8.64), (8.62) e (8.65),

$$y_{n+1} = y_n + h \sum_{j=1}^{s} b_j k_j \, ,$$

mentre l'equazione (8.63), utilizzando nel primo membro la (8.66) e nel secondo la (8.67), diventa

$$k_i = f(t_n + c_i h, y_n + h \sum_{j=1}^{s} a_{ij} k_j) \, , \quad i = 1, 2, \ldots, s.$$

8.5.2 Sulla stabilità e l'ordine dei metodi di Runge-Kutta

Esempio 8.5.1 Si vuole studiare la stabilità e calcolare l'ordine del seguente metodo di Runge-Kutta semi-implicito

$$A = \begin{pmatrix} 1/s & 0 & \cdots & 0 \\ 1/s & 1/s & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 1/s & 1/s & \cdots & 1/s \end{pmatrix}, \; b = \begin{pmatrix} 1/s \\ 1/s \\ \vdots \\ 1/s \end{pmatrix}, \; c \in I\!R^s \, .$$

Si verifica che

$$I - qA = \begin{pmatrix} 1 - q/s & 0 & \cdots & 0 \\ -q/s & 1 - q/s & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ -q/s & -q/s & \cdots & 1 - q/s \end{pmatrix}$$

e

$$I - qA + qub^T = \begin{pmatrix} 1 & q/s & \cdots & q/s \\ 0 & 1 & \cdots & q/s \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} .$$

Segue, dalla (8.24),

$$R(q) = \frac{1}{(1 - q/s)^s} . \tag{8.69}$$

Il metodo, quindi, è A-stabile.

Il calcolo dell'ordine può essere fatto in base alla (8.25). Sviluppando in serie di Taylor $R(q)$ si ottiene

$$R(q) = 1 + q + \frac{s+1}{2s} q^2 + O(q^3) .$$

Confrontando con la serie esponenziale si trova

$$e^q - R(q) = -\frac{1}{2s} q^2 + O(q^3) ,$$

da cui, essendo $q = h\lambda$, si ha $p = 1$.

Si osservi che, dalla (8.69), risulta

$$\lim_{s \to \infty} R(q) = e^q .$$

□

Esempio 8.5.2 Si vuole determinare l'ordine del metodo semi-implicito

0	0	0	0
1/2	1/4	1/4	0
1	0	1	0
	1/6	4/6	1/6

e calcolare un passo h idoneo ad approssimare, con tale metodo, la soluzione del problema

$$y' = Ky$$

dove

$$K = \begin{pmatrix} -4 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -4 \end{pmatrix} \in \mathbb{R}^{m\times m} .$$

Dalla (8.24) si ottiene

$$R(q) = \frac{1 + \frac{3}{4}q + \frac{1}{4}q^2 + \frac{1}{24}q^3}{1 - \frac{1}{4}q} .$$

Pertanto risulta (cfr. Tavola 8.3) $R(q) = R_1^3(q)$: da ciò e dalle (8.25) e (8.30) si ricava che l'ordine del metodo è $p = 4$.

Gli autovalori $\lambda_1, \lambda_2, \ldots, \lambda_m$ di K sono reali e, per il primo teorema di Gershgorin, si ha

$$\lambda_i \in]-6, -2[, \quad i = 1, 2, \ldots, m.$$

D'altro canto si verifica che $| \ R(q) \ | < 1$ se $q \in]-5.42, 0[$. Segue che la condizione (8.28) è soddisfatta se $h < \frac{-5.42}{-6} \simeq 0.9$.

Si osservi infine che, nell'applicazione del metodo al problema dato, k_1 si calcola direttamente e così k_3, una volta noto k_2. Il calcolo di k_2 richiede la risoluzione del sistema lineare

$$(I - \frac{1}{4}hK)k_2 = Ky_n + \frac{1}{4}hKk_1 .$$

Poiché gli autovalori di $I - \frac{1}{4}hK$ sono dati da $1 - \frac{1}{4}h\lambda_i$, $i = 1, 2, \ldots, m$, (cfr. Teorema 2.7.6) ed essendo $\lambda_i < 0$, $i = 1, 2, \ldots, m$, risulta sicuramente $\det(I - \frac{1}{4}hK) \neq 0$ per ogni $h > 0$ (cfr. Osservazione 2.7.1). □

8.5.3 Costruzione dei metodi a più passi ed esame della stabilità

Un metodo a più passi può essere costruito, a partire dalla sua forma generale (8.35), calcolandone i coefficienti mediante le condizioni di ordine (8.39) e imponendo che l'ordine p risulti massimo, compatibilmente con la condizione di zero-stabilità.

Esempio 8.5.3 Si vogliono determinare i coefficienti della formula BDF a due passi

$$\alpha_0 y_n + \alpha_1 y_{n+1} + \alpha_2 y_{n+2} = h\beta_2 f_{n+2} \, .$$

Per definizione $\alpha_2 = 1$. Per la coerenza deve aversi $\rho(1) = 0$ e $\rho'(1) - \sigma(1) = 0$ ovvero $\alpha_0 + \alpha_1 + 1 = 0$ e $2 + \alpha_1 - \beta_2 = 0$, da cui, esprimendo α_1 e β_2 in funzione di α_0, $\alpha_1 = -1 - \alpha_0$ e $\beta_2 = 1 - \alpha_0$.
Risulta quindi $\rho(\mu) = \mu^2 - (1+\alpha_0)\mu + \alpha_0 = (\mu - 1)(\mu - \alpha_0)$: la zero-stabilità è garantita se $\mid \alpha_0 \mid < 1$ oppure $\alpha_0 = -1$. L'ordine massimo si ottiene scegliendo α_0 in modo che risulti $c_2 = 0$. Dalle (8.39) con $r = 2$ si ha $c_2 = -\frac{1}{2} + \frac{3}{2}\alpha_0$ da cui $\alpha_0 = \frac{1}{3}$.
Allo stesso risultato si giunge, in virtù della (8.40), osservando che per un metodo di ordine p l'errore locale di troncamento è nullo se $y(t) = t^r$, $r = 0, 1, \ldots, p$, e quindi, utilizzando la (8.37), scrivendo $p + 1$ relazioni lineari nelle incognite α_j, β_j.

Il polinomio di stabilità del metodo è $\pi(q, \mu) = \left(1 - \frac{2}{3}q\right)\mu^2 - \frac{4}{3}\mu + \frac{1}{3}$ i cui zeri sono

$$\mu_1 = \frac{2 + \sqrt{1 + 2q}}{3 - 2q} \, , \quad \mu_2 = \frac{2 - \sqrt{1 + 2q}}{3 - 2q} \, .$$

Limitandosi al caso $q \in \mathbb{R}$, con semplici calcoli, si trova che il metodo è assolutamente stabile su tutto l'asse reale ad eccezione dell'intervallo $[0, 4]$ (il metodo, quindi, è A_0-stabile e si può, poi, dimostrare che è anche A-stabile), mentre è relativamente stabile per $q > -\frac{1}{2}$. □

8.5.4 Sulla stabilità dei metodi di predizione e correzione

Come si è visto in 8.3.4, le caratteristiche di stabilità di un metodo di predizione e correzione sono, in genere, diverse da quelle dei metodi che lo compongono. Per esempio, il metodo (8.55), pur essendo dotato di una regione di assoluta stabilità, perde la A-stabilità posseduta dal solo correttore. In altri casi, tuttavia, la stabilità è migliore di quella del predittore e di quella del correttore.

Esempio 8.5.4 Si consideri il metodo (8.59) senza la correzione di Milne.
Il polinomio di stabilità del predittore è

$$\pi^*(q, \mu) = \mu^4 - \frac{8}{3}q\mu^3 + \frac{4}{3}q\mu^2 - \frac{8}{3}q\mu - 1 \, .$$

Poiché gli zeri di $\pi^*(q,\mu)$ verificano la relazione $\mu_1\mu_2\mu_3\mu_4 = -1$ (cfr. la (4.50) con $i = m$), ne segue che almeno uno di essi è, in modulo, ≥ 1.

Il polinomio di stabilità del correttore è

$$\pi(q,\mu) = \left(1 - \frac{1}{3}q\right)\mu^2 - \frac{4}{3}q\mu - \left(1 + \frac{1}{3}q\right)$$

e si constata facilmente che per esso risulta $\mid \mu_1(q) \mid \geq 1$ se $Re(q) \geq 0$ e $\mid \mu_2(q) \mid > 1$ se $Re(q) < 0$. Se ne conclude che né il predittore né il correttore sono assolutamente stabili.

Il polinomio di stabilità del metodo di predizione e correzione definito dai due metodi è (cfr. la procedura già adottata per il metodo (8.55))

$$\hat{\pi}(q,\mu) = \mu^4 - \left(\frac{8}{9}q^2 + \frac{4}{3}q\right)\mu^3 + \left(\frac{4}{9}q^2 - \frac{1}{3}q - 1\right)\mu^2 - \frac{8}{9}q^2\mu - \frac{1}{3}q\,.$$

Limitandosi per semplicità al caso $q \in \mathbb{R}$, si verifica che il metodo è assolutamente stabile per $q = -\frac{1}{2}$. Posto $P_0(\mu) = \hat{\pi}\left(-\frac{1}{2},\mu\right)$, se ne consideri la successione di Sturm (cfr. Definizione e Teorema 4.7.1)

$$\begin{aligned}
P_0(\mu) &= \mu^4 + \tfrac{4}{9}\mu^3 - \tfrac{13}{18}\mu^2 - \tfrac{2}{9}\mu + \tfrac{1}{6},\\
P_1(\mu) &= \mu^3 + \tfrac{1}{3}\mu^2 - \tfrac{13}{36}\mu - \tfrac{1}{18},\\
P_2(\mu) &= \mu^2 + \tfrac{41}{129}\mu - \tfrac{56}{129},\\
P_3(\mu) &= \mu + \tfrac{1625}{164},\\
P_4(\mu) &= costante < 0.
\end{aligned}$$

Poiché risulta $V(-\infty) - V(0) = 0$ e $V(0) - V(+\infty) = 2$, il polinomio $\hat{\pi}\left(-\frac{1}{2},\mu\right)$ ha due zeri reali e positivi e due complessi coniugati. Per gli zeri reali si ha $0.6 < \mu_1 < 0.7$ e $0.4 < \mu_2 < 0.5$.
Indicato con ρ il modulo comune dei due zeri complessi coniugati, vale la relazione $\mu_1\mu_2\rho^2 = \frac{1}{6}$ (cfr. ancora la (4.50) con $i = m$) e quindi, poiché $0.24 < \mu_1\mu_2 < 0.35$, risulta $\rho^2 < 1$.

Uno studio più completo di $\hat{\pi}(q,\mu)$ mostra che il metodo in esame è dotato dell'intervallo reale di assoluta stabilità $]-0.8,-0.3[$. □

Un altro caso di miglioramento della stabilità è fornito dal *metodo di Hermite*

$$y_{n+2} = -4y_{n+1} + 5y_n + h(4f_{n+1} + 2f_n)\,,$$

addirittura non zero-stabile, e la regola di Simpson (8.47) che è zero-stabile, ma non assolutamente stabile. Usati insieme come predittore e correttore, danno luogo ad un metodo dotato di una regione di assoluta stabilità i cui estremi sono:

$$\text{minimo di } Re(q) = -1 \text{ e massimo di } \mid Im(q) \mid = 0.5.$$

8.5.5 Esempi di applicazione dei BVM

Esempio 8.5.5 Si abbia il problema di valori iniziali, scalare e lineare,

$$y'(t) = p(t)y(t) + q(t),\ p(t) < 0,\ a \le t \le b\,, \tag{8.70}$$

$$y(a) = y_0\,. \tag{8.71}$$

Volendo usare un $\mathrm{BVM}_{k_1k_2}$, seguendo quanto detto in 8.4.2, si ponga $t_0 = a$, $t_N = b$ ed $h = \frac{b-a}{N}$. L'intero N può essere fissato in base all'accuratezza che si vuole ottenere, tenendo conto che, se il metodo base prescelto è di ordine p, il $\mathrm{BVM}_{k_1k_2}$ produce errori dell'ordine di grandezza di $h^p = \left(\frac{b-a}{N}\right)^p$.

Essendo $p(t) < 0$ conviene usare un metodo base $\mathrm{A}_{k_1k_2}$-stabile. Scegliendo la formula del punto centrale (8.46), che è A_{11}-stabile, si ha

$$-y_n + y_{n+2} = 2h\left[p(t_{n+1})y_{n+1} + q(t_{n+1})\right]\ ,\ n = 0, 1, \ldots N-2. \tag{8.72}$$

Avendosi $k_1 = k_2 = 1$ occorre un solo metodo ausiliario finale, quale, ad esempio, la formula di Eulero implicita (8.45) con $n = N-1$:

$$-y_{N-1} + y_N = h\left[p(t_N)y_N + q(t_N)\right]\ . \tag{8.73}$$

Si ottiene perciò il sistema $Gy = c$ dove

$$G = \begin{pmatrix} -2hp(t_1) & 1 & & & & \\ -1 & -2hp(t_2) & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -1 & -2hp(t_{N-1}) & 1 & \\ & & & -1 & 1 - hp(t_N) & \end{pmatrix},$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N-1} \\ y_N \end{pmatrix}, \quad c = \begin{pmatrix} 2hq(t_1) + y_0 \\ 2hq(t_2) \\ \vdots \\ 2hq(t_{N-1}) \\ hq(t_N) \end{pmatrix}.$$

Invece di un problema di valori iniziali, si consideri ora il problema ai limiti continuo formato dalla (8.70) con la condizione ai limiti

$$\alpha y(a) + \beta y(b) = 0\,. \tag{8.74}$$

Il problema ai limiti discreto cui si giunge, usando le stesse formule, è costituito dalle (8.72) e (8.73). La corrispondente della (8.74)

$$\alpha y_0 + \beta y_N = 0$$

viene utilizzata ponendo $y_0 = -(\beta/\alpha)y_N$. Il sistema lineare che si ottiene in questo caso è $\tilde{G}y = \tilde{c}$ dove

$$\tilde{G} = \begin{pmatrix} -2hp(t_1) & 1 & & & \beta/\alpha \\ -1 & -2hp(t_2) & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & -2hp(t_{N-1}) & 1 \\ & & & -1 & 1 - hp(t_N) \end{pmatrix},$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N-1} \\ y_N \end{pmatrix}, \quad \tilde{c} = \begin{pmatrix} 2hq(t_1) \\ 2hq(t_2) \\ \vdots \\ 2hq(t_{N-1}) \\ hq(t_N) \end{pmatrix}.$$

□

Esempio 8.5.6 Si consideri la formula

$$y_{n+1} - y_n = h(\frac{5}{12}f_n + \frac{8}{12}f_{n+1} - \frac{1}{12}f_{n+2}).$$

Se ne studia la stabilità come BVM.
Si considera il problema test $y' = \lambda y$, per semplicità, solo nel caso reale, cioè $0 > \lambda \in$R. Il polinomio di stabilità della formula è

$$\pi(q, \mu) = \frac{1}{12}q\mu^2 + (1 - \frac{2}{3}q)\mu - (1 + \frac{5}{12}q).$$

Gli zeri sono quindi

$$\mu_1(q) = (-1 + \frac{2}{3}q + \sqrt{\Delta})/(\frac{1}{6}q), \quad \mu_2(q) = (-1 + \frac{2}{3}q - \sqrt{\Delta})/(\frac{1}{6}q),$$

avendo posto $\Delta = \frac{7}{12}q^2 - q + 1$. Si verifica facilmente che

$$|\mu_1(q)| < 1 < |\mu_2(q)|, \qquad \lim_{q\to 0^-} \mu_1(q) = 1,$$

per cui $\mu_1(q)$ è la radice principale, $k_1 = k_2 = 1$ e la formula usata come metodo base in un BVM è A_{11}-stabile.
Si trova poi $c_4 = \frac{1}{24}$ e quindi $p = 3$. Avendosi $k_1 - 1 = 0$ e $k_2 = 1$ occorre un metodo ausiliario "di coda". Allo scopo si può usare la formula del terzo ordine

$$y_{n+2} - y_{n+1} = h(-\frac{1}{12}f_n + \frac{8}{12}f_{n+1} + \frac{5}{12}f_{n+2}),$$

per la quale si ha $c_4 = -\frac{1}{24}$.
Il sistema, riferito al problema $y' = f(t, y), y(t_0) = y_0$, risulta quindi essere

$$\begin{cases} y_1 - \frac{8}{12}hf_1 + \frac{1}{12}hf_2 - y_0 - \frac{5}{12}hf_0 & = 0 \\ -y_1 - \frac{5}{12}hf_1 + y_2 - \frac{8}{12}hf_2 + \frac{1}{12}hf_3 & = 0 \\ -y_2 - \frac{5}{12}hf_2 + y_3 - \frac{8}{12}hf_3 + \frac{1}{12}hf_4 & = 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \ldots \\ -y_{N-2} - \frac{5}{12}hf_{N-2} + y_{N-1} - \frac{8}{12}hf_{N-1} + \frac{1}{12}hf_N & = 0 \\ \frac{1}{12}hf_{N-2} - y_{N-1} - \frac{8}{12}hf_{N-1} + y_N - \frac{5}{12}hf_N & = 0 \end{cases}.$$

La sua risoluzione numerica può essere effettuata con il metodo di Newton. Alternativamente si osserva che tale sistema è, in realtà, semilineare in quanto le incognite $y_1, y_2, \ldots, y_N$ si presentano sia in forma lineare sia coinvolte non linearmente come argomento della funzione f. Si verifica subito che il sistema si lascia scrivere nella forma

$$By = hF(y) + c,$$

dove:

$$B = \begin{pmatrix} 1 & & \\ -1 & 1 & \\ & \ddots & \ddots \\ & -1 & 1 \end{pmatrix},$$

il vettore delle incognite è

$$y = (y_1^T, y_2^T, \ldots, y_N^T)^T,$$

le componenti del vettore $F(y)$ sono

$$\begin{aligned}
&F_1(y) = \tfrac{8}{12}f_1 - \tfrac{1}{12}f_2,\\
&F_{i+1}(y) = \tfrac{5}{12}f_i + \tfrac{8}{12}f_{i+1} - \tfrac{1}{12}f_{i+2}, \quad i = 1, 2, \ldots, N-2,\\
&F_N(y) = -\tfrac{1}{12}f_{N-2} + \tfrac{8}{12}f_{N-1} + \tfrac{5}{12}f_N,
\end{aligned}$$

e il vettore dei dei termini noti è dato da

$$c = ((y_0 + h\frac{5}{12}f_0)^T, 0^T, \ldots, 0^T)^T.$$

Resta quindi definito il procedimento iterativo

$$y^{(s+1)} = hB^{-1}F(y^{(s)}) + B^{-1}c, \quad s = 0, 1, \ldots,$$

essendo, come si riscontra immediatamente, la matrice inversa di B triangolare inferiore data da

$$B^{-1} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 1 & 1 & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ 1 & 1 & \ldots & 1 \end{pmatrix}.$$

Per h sufficientemente piccolo, in virtù del Teorema 4.6.1, tale procedimento risulta convergente e pertanto può adottarsi il criterio di arrestare le iterazioni allorché $\|y^{(s+1)} - y^{(s)}\| \leq \epsilon$ con ϵ opportunamente prefissato. □

Esempio 8.5.7 La formula trapezoidale, $y_{n+1} - y_n = \frac{h}{2}(f_{n+1} + f_n)$, come si è osservato alla fine del paragrafo 8.4.3, può essere usata come BVM da sola. Facendo riferimento al problema di valori iniziali lineare

$$y' = K(t)y, \quad y(t_0) = y_0,$$

dove $K(t) \in \mathrm{R}^{m\times m}$, ponendo $K(t_n) = K_n$, il metodo si può scrivere

$$y_n + \frac{1}{2}hK_n y_n - y_{n+1} + \frac{1}{2}hK_{n+1}y_{n+1} = 0, \quad n = 0, 1, \ldots, N-1.$$

Introducendo il vettore delle incognite $y = (y_1^T, y_2^T, \ldots, y_N^T)^T$, il vettore dei termini noti $c = ((-y_0 - \frac{1}{2}hK_0y_0)^T, 0^T, \ldots, 0^T)^T$ e la matrice

$$G = \begin{pmatrix} -I + \frac{1}{2}hK_1 & & & \\ I + \frac{1}{2}hK_1 & -I + \frac{1}{2}hK_2 & & \\ & \ddots & \ddots & \\ & & I + \frac{1}{2}hK_{N-1} & -I + \frac{1}{2}hK_N \end{pmatrix},$$

si è condotti alla risoluzione del sistema lineare $Gy = c$.
Si osserva poi che

$$det(G) = det(-I + \frac{1}{2}hK_1) \cdots det(-I + \frac{1}{2}hK_N),$$

per cui $\lim_{h\to 0} |det(G)| = 1$. Ne segue che per h sufficientemente piccolo (e $K(t)$ sufficientemente regolare) il sistema ha un'unica soluzione.
Si supponga ora $K(t) = K =$ costante. Risulta

$$det(G) = [det(-I + \frac{1}{2}hK)]^N = [(-1 + \frac{1}{2}h\lambda_1) \cdots (-1 + \frac{1}{2}h\lambda_m)]^N,$$

essendo $\lambda_1, \ldots, \lambda_m$ gli autovalori di K. Pertanto se gli autovalori sono reali non positivi o complessi si ha $det(G) \neq 0$ per qualunque valore di h. Se vi sono autovalori reali e positivi si potrà assumere $h < 2/\rho(K)$. □

Bibliografia: [4], [3], [6], [8], [16], [17], [18], [20], [21], [22], [24].