Capitolo 1

L'insieme dei numeri macchina

1.1 Introduzione al Calcolo Numerico

Il Calcolo Numerico è una disciplina che fa parte di un ampio settore della Matematica Applicata che prende il nome di Analisi Numerica. Si tratta di una materia che è al confine tra la Matematica e l'Informatica poichè cerca di risolvere i consueti problemi matematici utilizzando però una via algoritmica. In pratica i problemi vengono risolti indicando un processo che, in un numero finito di passi, fornisca una soluzione numerica e soprattutto che sia implementabile su un elaboratore. I problemi matematici che saranno affrontati nelle pagine seguenti sono problemi di base: risoluzione di sistemi lineari, approssimazione delle radici di funzioni non lineari, approssimazione di funzioni e dati sperimentali, calcolo di integrali definiti. Tali algoritmi di base molto spesso non sono altro se non un piccolo ingranaggio nella risoluzione di problemi ben più complessi.

1.2 Rappresentazione in base di un numero reale

Dovendo considerare problemi in cui l'elaboratore effettua computazioni esclusivamente su dari di tipo numerico risulta decisivo iniziare la trattazione degli argomenti partendo dalla rappresentazione di numeri. Innanzitutto è opportuno precisare che esistono due modi per rappresentare i numeri: la cosiddetta notazione posizionale, in cui il valore di una cifra dipende dalla

posizionein ci si trova all'interno del numero, da quella notazione non posizionale, in cui ogni numero è rappresentato da uno, o da un insieme di simboli (si pensi come esempio alla numerazione usata dai Romani). La motivazione che ci spinge a considerare come primo problema quello della rappresentazione di numeri reali è che ovviamente si deve sapere se i risultati forniti dall'elaboratore sono affidabili. Un primo problema che ci troviamo ad affrontare è il modo con cui i numeri reali sono rappresentati nella memoria di un elaboratore. Giusto per rendersi conto delle difficoltà che si incontrano va osservato che i numeri reali sono infiniti mentre la memoria di un calcolatore ha una capacità finita. Una seconda osservazione consiste nel fatto che un numero reale ammette molteplici rappresentazioni. Se consideriao il numero reale $x = 123.47$ in realtà questa simbologia èla rappresentazione, in forma convenzionale, dell'espressione matematica

$$x = 123.47 = 1 \times 10^2 + 2 \times 10^1 + 3 \times 10^0 + 4 \times \times 10^{-1} + 7 \times 10^{-2},$$

da cui, mettendo in evidenza 10^2:

$$x = 10^2 \times \left(1 \times 10^0 + 2 \times 10^{-1} + 3 \times 10^{-2} + 4 \times \times 10^{-3} + 7 \times 10^{-4}\right)$$

da cui si deduce che è necessario stabilire una rappresentazione convenzionale dei numeri reali, fornita dal seguente teorema.

Teorema 1.2.1 *Sia $\beta \in \mathbb{N}$, $\beta > 1$, allora ogni numero reale x, $x \neq 0$,, può essere rappresentato univocamente in base β nel seguente modo*

$$x = \pm\ \beta^p \sum_{i=1}^{\infty} d_i \beta^{-i}$$

dove $p \in \mathbb{Z}$, e i valori $d_i \in \mathbb{N}$ (detti cifre), verificano le seguenti proprietà:
1. $d_i \in \{1, 2, 3, \ldots, \beta - 1\}$;
2. $d_1 \neq 0$;
3. le cifre d_i non sono definivamente uguali a $\beta - 1$.

Evitiamo la dimostrazione del teorema 1.2.1 ma osserviamo che la richiesta che la terza ipotesi è importante per l'unicità della rappresentazione.

Consideriamo infatti il seguente esempio (in base $\beta = 10$).

$$\begin{aligned} x &= 0.999999999\ldots \\ &= 9 \times 10^{-1} + 9 \times 10^{-2} + 9 \times 10^{-3} + \ldots \\ &= \sum_{i=1}^{\infty} 9 \cdot 10^{-i} = 9 \sum_{i=1}^{\infty} \left(\frac{1}{10}\right)^i \\ &= 9 \left(\frac{1}{10}\right) \left(1 - \frac{1}{10}\right)^{-1} \\ &= 9 \left(\frac{1}{10}\right) \left(\frac{10}{9}\right) = 1. \end{aligned}$$

L'ultima uguaglianza deriva dalla covergenza della serie geometrica

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

quando $0 < q < 1$, da cui segue

$$1 + \sum_{i=1}^{\infty} q = \frac{1}{1-q}$$

e

$$\sum_{i=1}^{\infty} q = \frac{1}{1-q} - 1 = \frac{q}{1-q}.$$

In conclusione al numero 1 corrisponderebbero due differenti rappresentazioni. Considerato un numero reale $x \in \mathbb{R}$, $x \neq 0$, l'espressione

$$x = \pm\ \beta^p \times 0.d_1 d_2 \ldots d_k \ldots$$

prende il nome di rappresentazione in base β di x. Il numero p viene detto esponente (o caratteristica), i valori d_i sono le cifre della rappresentazione, mentre $0.d_1 d_2 \ldots d_k \ldots$ si dice mantissa. Il numero x viene normalmente rappresentato con la cosiddetta notazione posizionale $x = segno(x)(.d_1 d_2 d_3 \ldots) \times \beta^p$, che viene detta normalizzata. In alcuni casi à ammessa una rappresentazione in notazione posizionale tale che $d_1 = 0$, che viene detta denormalizzata.

La basi più utilizzate sono $\beta = 10$ (sistema decimale), $\beta = 2$ (sistema binario, che, per la sua semplicità, è quello utilizzato dagli elaboratori elettronici), e $\beta = 16$ (sistema esadecimale). Nel sistema esadecimale le cifre appartengono all'insieme

$$\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F\}.$$

Bisogna ternere presente che un qualunque numero reale $x \neq 0$ può essere rappresentato con infinite cifre nella mantissa e inoltre l'insieme dei numeri reali ha cardinalità infinita. Poichè un elaboratore è dotato di memoria finita non è possibile memorizzare:
a) gli infiniti numeri reali
b) le infinite (in generale) cifre di un numero reale.

1.3 L'insieme dei numeri macchina

Assegnati i numeri $\beta, t, m, M \in \mathbb{N}$ con $\beta \geq 2$, $t \geq 1$, $m, M > 0$, si dice insieme dei numeri di macchina con rappresentazione normalizzata in base β con t cifre significative l'insieme:

$$\mathbb{F}(\beta, t, m, M) = \left\{ x \in \mathbb{R} : \; x = \pm\beta^p \sum_{i=1}^{t} d_i \beta^{-i} \right\} \cup \{0\}$$

dove
1. $t \geq 0, \beta > 1$;
2. $d_i \in \{0, 1 \ldots, \beta - 1\}$;
3. $d_1 \neq 0$;
4. $p \in \mathbb{Z}$, $-m \leq p \leq M$.

È stato necessario aggiungere il numero zero all'insieme in quanto sfugge alla rappresentazione in base normalizzata viene assegnato per definizione all'insieme $\mathbb{F}(\beta, t, m, M)$.
Osserviamo che un elaboratore la cui memoria abbia le seguenti caratteristiche (riportate anche in figura 1.1:

- t campi di memoria per la mantissa, ciascuno dei quali può assumere β differenti configurazioni (e perciò può memorizzare una cifra d_i),
- un campo di memoria che può assumere $m + M + 1$ differenti configurazioni (e perciò può memorizzare i differenti valori p dell'esponente),

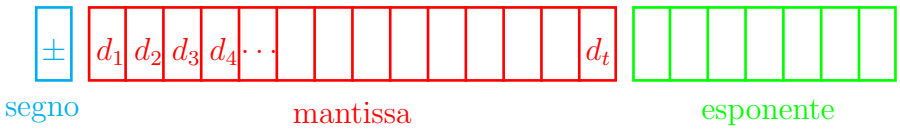


Figura 1.1: Locazione di memoria.

- un campo che può assumere due differenti configurazioni (e perciò può memorizzare il segno + o −),

è in grado di rappresentare tutti gli elementi dell'insieme $\mathbb{F}(\beta, t, m, M)$. In realtà poichè se $\beta = 2$ $d_1 = 1$, allora determinati standard non memorizzano la prima cifra della mantissa. Il più piccolo numero positivo appartenente all'insieme $\mathbb{F}(\beta, t, m, M)$ si ottiene prendendo la più piccola mantissa (ovvero 0.1) ed il più piccolo esponente

$$x = 0.1 \times \beta^{-m}$$

mentre il più grande ha tutte le cifre della mantissa uguali alla cifra più grande (ovvero $\beta - 1$) ed il massimo esponente

$$x = 0.\underbrace{dd\ldots dd}_{t}\ \beta^M, \qquad d = \beta - 1.$$

Consideriamo ora come esempio l'insieme $\mathbb{F}(2, 2, 2, 2)$, cioè i numeri binari con mantissa di due cifre ed esponente compreso tra -2 e 2. Enumeriamo gli elementi di questo insieme. Poichè il numero zero non appartiene all'insieme dei numeri macchina viene rappresentato solitamente con mantissa nulla ed

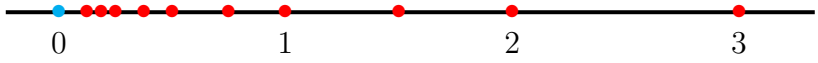


Figura 1.2: Elementi dell'insieme $\mathbb{F}(2,2,2,2)$.

esponente $-m$.

$$
\begin{aligned}
p=-2 \quad & x = 0.10 \times 2^{-2} = 2^{-1} \times 2^{-2} = 2^{-3} = 0.125; \\
& x = 0.11 \times 2^{-2} = (2^{-1} + 2^{-2}) \times 2^{-2} = 3/16 = 0.1875; \\
p=-1 \quad & x = 0.10 \times 2^{-1} = 2^{-1} \times 2^{-1} = 2^{-2} = 0.25; \\
& x = 0.11 \times 2^{-1} = (2^{-1} + 2^{-2}) \times 2^{-1} = 3/8 = 0.375; \\
p=0 \quad & x = 0.10 \times 2^{0} = 2^{-1} \times 2^{0} = 2^{-1} = 0.5; \\
& x = 0.11 \times 2^{0} = (2^{-1} + 2^{-2}) \times 2^{0} = 3/4 = 0.75; \\
p=1 \quad & x = 0.10 \times 2^{1} = 2^{-1} \times 2^{1} = 1; \\
& x = 0.11 \times 2^{1} = (2^{-1} + 2^{-2}) \times 2^{1} = 3/2 = 1.15; \\
p=2 \quad & x = 0.10 \times 2^{2} = 2^{-1} \times 2^{2} = 2; \\
& x = 0.11 \times 2^{2} = (2^{-1} + 2^{-2}) \times 2^{2} = 3;
\end{aligned}
$$

Nella Figura 1.2 è rappresentato l'insieme dei numeri macchina positivi appartenenti a $\mathbb{F}(2,2,2,2)$ (i numeri negativi sono esattamente simmetrici rispetto allo zero). Dalla rappresentazione dell'insieme dei numeri macchina si evincono le seguenti considerazioni:

1. L'insieme è discreto;
2. I numeri rappresentabili sono solo una piccola parte dell'insieme $\mathbb{R}$;
3. La distanza tra due numeri reali consecutivi è β^{p-t}, infatti, considerando per semplicità numeri positivi, sia

$$x = +\beta^{p} \times (0.d_1, d_2, \ldots, d_{t-1}, d_t)$$

il successivo numero macchina è

$$y = +\beta^p \times (0.d_1, d_2, \ldots, d_{t-1}, \tilde{d}_t)$$

dove

$$\tilde{d}_t = d_t + 1.$$

La differenza è pertanto

$$y - x = +\beta^p (0.\underbrace{00\ldots00}_{t-1}1) = \beta^{p-t}.$$

Nello standard IEEE (Institute of Electric and Electronic Engineers) singola precisione una voce di memoria ha 32 bit, dei quali 1 riservato al segno, 8 all'esponente e 23 alla mantissa. Allora $\beta = 2$, $t = 23, m = 127$ e $M = 128$. Per la doppia precisione si utilizzano 64 bit, di cui 1 per il segno, 11 per lesponente e 52 per la mantissa. Dunque $\beta = 2$, $t = 52, m = -1023$ e $M = 1024$. Dopo aver compreso la struttura dell'insieme $\mathbb{F}(\beta, t, m, M)$ resta da capire come, assegnato un numero reale x sia possibile rappresentarlo nell'insieme dei numeri macchina, ovvero quale elemento $\tilde{x} \in \mathbb{F}(\beta, t, m, M)$ possa essergli associato in modo da commettere il più piccolo errore di rappresentazione possibile. Supponiamo ora che la base β sia un numero pari. Possono presentarsi diversi casi:

- Sia
$$x = \pm\ \beta^p \sum_{i=1}^{n} d_i \beta^{-i}$$
con $d_1 \neq 0$, $n \leq t$, e $-m \leq p \leq M$. Allora è evidente che $x \in \mathbb{F}(\beta, t, m, M)$ e pertanto verrà rappresentato esattamente su un qualunque elaboratore che utilizzi $\mathbb{F}(\beta, t, m, M)$ come insieme dei numeri di macchina.

- Sia
$$x = \pm\ \beta^p \sum_{i=1}^{n} d_i \beta^{-i}$$
con $n \leq t$ ma supponiamo che $p \notin [-m, M]$. Se $p < -m$ allora x è più piccolo del più piccolo numero di macchina: in questo caso si dice che si è verificato un underflow (l'elaboratore interrompe la sequenza di calcoli e segnala con un messaggio l'underflow). Se $p > M$ allora

vuol dire che x è più grande del più grande numero di macchina e in questo caso si dice che si è verificato un overflow (anche in questo caso l'elaboratore si ferma e segnala l'overflow, anche se tale eccezione può anche essere gestita via software in modo tale che l'elaborazione continui).

- Sia
$$x = \pm\, \beta^p \sum_{i=1}^{n} d_i \beta^{-i}$$
con l'esponente $-m \leq p \leq M$ ma $t > n$ ed esiste un $k > t$ tale che $d_k \neq 0$. Anche in questo caso poichè x ha più di t cifre significative $x \notin \mathbb{F}(\beta, t, m, M)$. È però possibile rappresentare x mediante un numero in $\mathbb{F}$ con un'opportuna operazione di taglio delle cifre decimali che seguono la t-esima. Per questo si possono utilizzare due diverse tecniche di approssimazione:

  1. troncamento di x alla t-esima cifra significativa
  $$\tilde{x} = \mathrm{tr}(x) = \beta^p \times 0.d_1 d_2 \ldots d_t$$
  2. arrotondamento di x alla t-esima cifra significativa
  $$\tilde{x} = \mathrm{arr}(x) = \beta^p \times 0.d_1 d_2 \ldots \tilde{d}_t$$

dove
$$\tilde{d}_t = \begin{cases} d_t + 1 & \text{se } d_{t+1} \geq \beta/2 \\ d_t & \text{se } d_{t+1} < \beta/2. \end{cases}$$
Per esempio se $x = 0.654669235$ e $t = 5$ allora
$$tr(x) = 0.65466, \qquad arr(x) = 0.65467$$
In pratica quando il numero reale x non appartiene all'insieme $\mathbb{F}(\beta, t, m, M)$ esistono sicuramente due numeri $a, b \in \mathbb{F}(\beta, t, m, M)$, tali che
$$a < x < b. \tag{1.1}$$
Supponendo per semplicità $x > 0$ si ha che
$$tr(x) = a$$

mentre se $x \geq (a+b)/2$ allora

$$arr(x) = b$$

altrimenti

$$arr(x) = a.$$

L'arrotondamento è un'operazione che fornisce sicuramente un risultato più preciso (come risulterà evidente nel prossimo paragrafo), ma può dar luogo ad overflow. Infatti se $x = 0.9999999\cdots \times \beta^M$ allora

$$arr(x) = 1.0\beta^M = 0.1\beta^{M+1} \notin \mathbb{F}(\beta, t, m, M).$$

La rappresentazione di $x \in \mathbb{R}$ attraverso $\tilde{x} \in \mathbb{F}(\beta, t, m, M)$ si dice rappresentazione in virgola mobile di x o rappresentazione floating point, con troncamento se $\tilde{x} = \text{tr}(x)$, con arrotondamento se $\tilde{x} = \text{arr}(x)$. Talvolta il numero macchina che rappresenta $x \in \mathbb{R}$ viene indicato con $fl(x)$.

1.4 Errore Assoluto ed Errore Relativo

Una volta definite le modalità per associare ad un numero reale x la sua rappresentazione macchina $\tilde{x}$ si tratta di stabilire l'errore che si commette in questa operazione di approssimazione. Si possono definire due tipi di errori, l'errore assoluto e l'errore relativo.
Se $x \in \mathbb{R}$ ed $\tilde{x}$ è una sua approssmazione allora si definisce errore assoluto la quantità

$$E_a = |\tilde{x} - x|$$

mentre se $x \neq 0$ si definisce errore relativo la quantità

$$E_r = \frac{|\tilde{x} - x|}{|x|}.$$

Se $E_r \leq \beta^{-q}$ allora si dice che $\tilde{x}$ ha almeno q cifre significative corrette. Nel seguito assumeremo $x > 0$ e supporremo anche che la rappresentazione di x in $\mathbb{F}(\beta, t, m, M)$ non dia luogo ad underflow o overflow. Calcoliamo ora una maggiorazione per tali errori nel caso in cui $\tilde{x}$ sia il troncamento di $x > 0$. Nella Figura 1.3 a e b rappresentano i due numeri macchina tali che sia vera la relazione (1.1). È evidente ch risulta

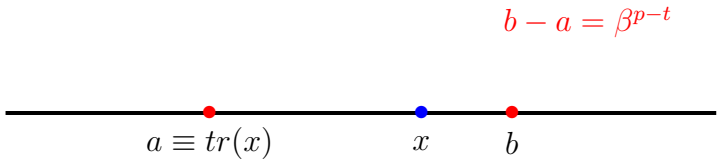


Figura 1.3: Stima dell'errore di rappresentazione nel caso di troncamento.

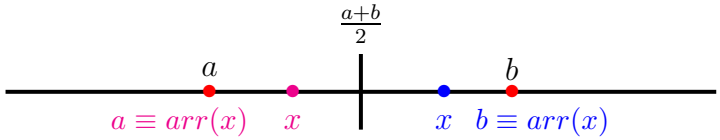


Figura 1.4: Stima dell'errore di rappresentazione nel caso di arrotondamento.

$$|tr(x) - x| < b - a = \beta^{p-t}.$$

Per maggiorare l'errore relativo osserviamo che

$$|x| = +\beta^p \times 0.d_1 d_2 d_3 \cdots \geq \beta^p \times 0.1 = \beta^{p-1}.$$

da cui

$$\frac{1}{|x|} \leq \beta^{1-p}$$

e quindi

$$\frac{|tr(x) - x|}{|x|} \leq \beta^{p-t} \times \beta^{1-p} = \beta^{1-t}. \tag{1.2}$$

Passiamo ora alla valutazione degli errori quando

$$\tilde{x} = arr(x).$$

Nella Figura 1.4 a e b rappresentano i due numeri macchina tali che sia vera la relazione (1.1). Se $x > 0$ si trova a sinistra del punto medio $(a + b)/2$ allora l'arrotondamento coincide con il valore a, se si trova nel punto medio

oppure alla sua destra allora coincide con b. È evidente che il massimo errore si ottiene quando x coincide con il punto medio tra a e b risulta

$$|arr(x) - x| \leq \frac{1}{2}(b-a) = \frac{1}{2}\beta^{p-t}.$$

Per maggiorare l'errore relativo procediamo come nel caso del troncamento di x:

$$\frac{|arr(x) - x|}{|x|} \leq \frac{1}{2}\beta^{p-t} \times \beta^{1-p} = \frac{1}{2}\beta^{1-t}. \tag{1.3}$$

Le quantità che compaiono a destra delle maggiorazioni (1.2) e (1.3), ovvero

$$u = \beta^{1-t}$$

oppure

$$u = \frac{1}{2}\beta^{1-t}$$

sono dette precisione di macchina o zero macchina per il troncamento (o per l'arrotondamento, in base alla tecnica in uso).
Posto

$$\varepsilon_x = \frac{\tilde{x} - x}{x}, \qquad |\varepsilon| \leq u$$

risulta

$$\tilde{x} = x(1 + \varepsilon_x) \tag{1.4}$$

che fornisce la relazione tra un numero $x \in \mathbb{R}$ e la sua rappresentazione macchina.

1.4.1 Operazioni Macchina

Se $x, y \in \mathbb{F}(\beta, t, m, M)$ è chiaro che il risultato di un'operazione aritmetica tra x e y non è detto che sia un numero macchina, inoltre è chiaro che quanto detto per la rappresentazione dei numeri reali sia valido anche per tale risultato. Se $\cdot$ è una delle quattro operazioni aritmetiche di base allora affinchè il risultato sia un numero macchina deve accadere che

$$x \cdot y = fl(x \cdot y). \tag{1.5}$$

L'operazione definita dalla relazione (1.5) è detta operazione macchina. L'operaziona macchina associata a $\cdot$ viene indicata con $\odot$ e deve soddisfare anch'essa la relazione (1.4), ovvero dev'essere:

$$x \odot y = (x \cdot y)(1 + \varepsilon), \qquad |\varepsilon| < u \tag{1.6}$$

per ogni $x, y \in \mathbb{F}(\beta, t, m, M)$ tali che $x \odot y$ non dia luogo ad overflow o underflow. Si può dimostrare che

$$x \odot y = \mathrm{tr}(x \cdot y)$$

e

$$x \odot y = \mathrm{arr}(x \cdot y)$$

soddisfano la (1.6) e dunque danno luogo ad operazioni di macchina. Le quattro operazioni così definite danno luogo alla aritmetica di macchina o aritmetica finita. La somma algebrica macchina (addizione e sottrazione) tra due numeri $x, y \in \mathbb{F}(\beta, t, m, M)$ richiede le seguenti fasi:

1. Si scala la mantissa del numero con l'esponente minore in modo tale che i due addendi abbiano lo stesso esponente (ovvero quello dell'esponente maggiore);

2. Si esegue la somma tra le mantisse;

3. Si normalizza il risultato aggiustando l'esponente in modo tale che la mantissa sia un numero minore di 1.

4. Si arrotonda (o si tronca) la mantissa alle prime t cifre;

Consideriamo per esempio i numeri $x, y \in \mathbb{F}(10, 5, m, M)$

$$x = 0.78546 \times 10^2, \qquad y = 0.61332 \times 10^{-1}$$

e calcoliamo il numero macchina $x \oplus y$.
1. Scaliamo il numero y fino ad ottenere esponente 2 (quindi si deve spostare il punto decimale di 3 posizioni), $y = 0.00061332 \times 10^2$;
2. Sommiamo le mantisse $0.78546 + 0.00061332 = 0.78607332$;
3. Questa fase non è necessaria perchè la mantissa è già minore di 1;
4. Si arrotonda alla quinta cifra decimale ottenendo

$$x \oplus y = 0.78607 \times 10^2.$$

Un fenomeno particolare, detto cancellazione di cifre significative, si verifica quando si effettua la sottrazione tra due numeri reali all'incirca uguali. Consideriamo per esempio la differenza tra i due numeri

$$x = 0.75868531 \times 10^2, \qquad y = 0.75868100 \times 10^2$$

nell'insieme $\mathbb{F}(10, 5, m, M)$. Risulta

$$fl(x) = 0.75869 \times 10^2, \qquad fl(y) = 0.75868 \times 10^2$$

e quindi

$$fl(fl(x) - fl(y)) = 0.1 \times 10^{-2}$$

mentre

$$x - y = 0.431 \times 10^{-3}$$

Calcolando l'errore relativo sul risultato dell'operazione si trova

$$E_r \simeq 1.32016$$

che è un valore piuttosto alto.
Il prodotto macchina tra due numeri $x, y \in \mathbb{F}(\beta, t, m, M)$ richiede le seguenti fasi:

1. Si esegue il prodotto tra le mantisse;
2. Si esegue l'arrotondamento (o il troncamento) alle prime t cifre;
3. Si sommano gli esponenti, normalizzando, se necessario, la mantissa ad un numero minore di 1.

Consideriamo per esempio il prodotto tra i due numeri

$$x = 0.11111 \times 10^3, \qquad y = 0.52521 \times 10^2$$

nell'insieme $\mathbb{F}(10, 5, m, M)$.
1. Il prodotto delle mantisse produce 0.05835608;
2. L'arrotondamento a 5 cifre produce 0.58356×10^{-1};
3. Somma degli esponenti $x * y = 0.58356 \times 10^4$.
La divisione macchina tra due numeri $x, y \in \mathbb{F}(\beta, t, m, M)$ richiede le seguenti fasi:

1. Si scala il dividendo x finchè la sua mantissa non risulti minore di quella del divisore y;
2. Si esegue la divisione tra le mantisse;
3. Si esegue l'arrotondamento (o il troncamento) alle prime t cifre;

4. Si sottraggono gli esponenti.

Consideriamo la divisione tra i due numeri

$$x = 0.12100 \times 10^5, \qquad y = 0.11000 \times 10^2$$

nell'insieme $\mathbb{F}(10, 5, m, M)$.
1. Scaliamo il dividendo di una cifra decimale 0.012100;
2. Dividiamo le mantisse $0.012100/0.11000 = 0.11000$;
3. Il troncamento fornisce lo stesso numero 0.11000;
4. Si sottraggono gli esponenti ottenendo il risultato

$$x \oslash y = 0.11000 \times 10^3.$$

Si può dimostrare che valgono le seguenti proprietà:

1. L'insieme $\mathbb{F}(\beta, t, m, M)$ non è chiuso rispetto alle operazioni macchina;
2. L'elemento neutro per la somma non è unico: infatti consideriamo i due numeri macchia

   $$x = 0.15678 \times 10^3, \qquad y = 0.25441 \times 10^{-2},$$

   appartenenti all'insieme $\mathbb{F}(10, 5, m, M)$, innanzitutto si scala y

   $$y = 0.0000025441 \times 10^3,$$

   sommando le mantisse si ottiene 0.1567825441 mentre l'arrotondamento fornisce il risultato finale

   $$x \oplus y = 0.15678 \times 10^3 = x.$$

3. L'elemento neutro per il prodotto non è unico;
4. Non vale la proprietà associativa di somma e prodotto;
5. Non vale la proprietà distributiva della somma rispetto al prodotto.

Capitolo 2

Equazioni non Lineari

2.1 Introduzione

Le radici di un'equazione non lineare $f(x) = 0$ non possono, in generale, essere espresse esplicitamente e anche se ciò è possibile spesso l'espressione si presenta in forma talmente complicata da essere praticamente inutilizzabile. Di conseguenza per poter risolvere equazioni di questo tipo siamo obbligati ad utilizzare metodi numerici che sono, in generale, di tipo iterativo, cioè partendo da una (o in alcuni casi più) approssimazioni della radice, producono una successione x_0, x_1, $x_2, \ldots$, convergente alla radice. Per alcuni di questi metodi per ottenere la convergenza è sufficiente la conoscenza di un intervallo $[a, b]$ che contiene la soluzione, altri metodi richiedono invece la conoscenza di una buona approssimazione iniziale. Talvolta è opportuno utilizzare in maniera combinata due metodi, uno del primo tipo e uno del secondo.
Prima di analizzare alcuni metodi per l'approssimazione delle radici dell'equazione $f(x) = 0$ diamo la definizione di molteplicità di una radice.

Definizione 2.1.1 *Sia $f \in \mathcal{C}^r([a, b])$ per un intero $r > 0$. Una radice α di $f(x)$ si dice di molteplicità r se*

$$\lim_{x \to \alpha} \frac{f(x)}{(x - \alpha)^r} = \gamma, \qquad \gamma \neq 0, c \neq \pm\infty. \tag{2.1}$$

Se α è una radice della funzione $f(x)$ di molteplicità r allora risulta

$$f(\alpha) = f'(\alpha) = \cdots = f^{(r-1)}(\alpha) = 0, \qquad f^{(r)}(\alpha) = \gamma \neq 0.$$

2.2 Localizzazione delle radici

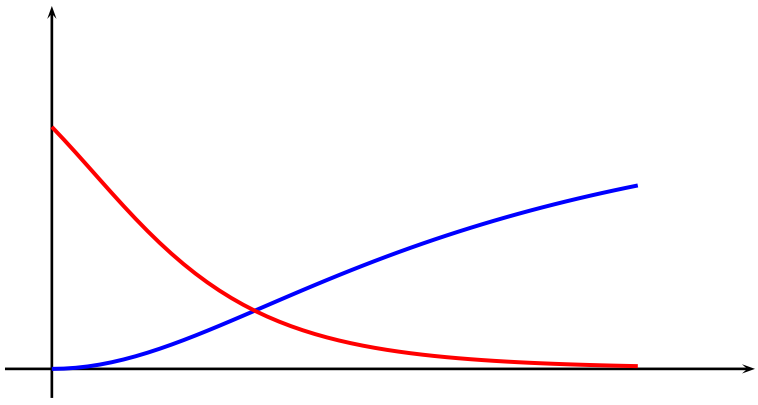
Nei successivi paragrafi saranno descritti alcuni metodi numerici per il calcolo approssimato delle radici di un'equazione non lineare. Tali metodi numerici sono di tipo iterativo, ovvero consistono nel definire una successione (o più successioni), che, a partire da un'assegnata approssimazione iniziale (nota), converga alla radice α in un processo al limite. Infatti poichè non esistono tecniche generali che consentano di trovare l'espressione esplicita di α in un numero finito di operazioni, allora questa può essere calcolata in modo approssimato solo in modo iterativo. Questa perculiarità tuttavia richiede che sia nota appunto un'approssimazione iniziale o, almeno, un intervallo di appartenenza. Il problema preliminare è quello di localizzare la radice di una funzione, problema che viene affrontato in modo grafico. Per esempio considerando la funzione

$$f(x) = \sin\big(\log(x^2+1)\big) - \frac{e^{-x}}{x^2+1}$$

risulta immediato verificare che il valore dell'ascissa in cui si annulla è quello in cui si intersecano i grafici delle funzioni

$$g(x) = \sin\big(\log(x^2+1)\big) \qquad\qquad h(x) = \frac{e^{-x}}{x^2+1}.$$

Un modo semplice per stimare tale valore è quello di tracciare i grafici delle due funzioni, come riportato nella seguente figura in cui il grafico di $h(x)$ è in rosso, mentre quello di $g(x)$ è blu, e l'intervallo di variabilità di x è $[0, 2.5]$.



Calcolando le funzioni in valori compresi in tale intervallo di variabilità si può restringere lo stesso intervallo, infatti risulta

$$g(0.5) = 0.2213 < h(0.5) = 0.48522$$

e

$$g(1) = 0.63896 > h(1) = 0.18394,$$

da cui si deduce che $\alpha \in]0.5, 1[$.

2.3 Il Metodo di Bisezione

Sia $f : [a, b] \to \mathbb{R}$, $f \in \mathcal{C}([a, b])$, e sia $f(a)f(b) < 0$. Sotto tali ipotesi esiste sicuramente almeno un punto nell'intervallo $[a, b]$ in cui la funzione si annulla. L'idea alla base del Metodo di Bisezione (o metodo delle bisezioni) consiste nel costruire una successione di intervalli $\{I_k\}_{k=0}^{\infty}$, con $I_0 = [a_0, b_0] \equiv [a, b]$, tali che:

1. $I_{k+1} \subset I_k$;
2. $\alpha \in I_k$, $\forall k \geq 0$;
3. l'ampiezza di I_k tende a zero per $k \to +\infty$.

La successione degli I_k viene costruita nel seguente modo. Innanzitutto si pone

$$I_0 = [a_0, b_0] = [a, b]$$

e si calcola il punto medio

$$c_1 = \frac{a_0 + b_0}{2}.$$

Se $f(c_1) = 0$ allora $\alpha = c_1$, altrimenti si pone:

$$I_1 = [a_1, b_1] \equiv \begin{cases} a_1 = a_0 \quad b_1 = c_1 & \text{se } f(a_0)f(c_1) < 0 \\ a_1 = c_1 \quad b_1 = b_0 & \text{se } f(a_0)f(c_1) > 0. \end{cases}$$

Ora, a partire da $I_1 = [a_1, b_1]$, si ripete la stessa procedura. In generale al passo k si calcola

$$c_{k+1} = \frac{a_k + b_k}{2}.$$

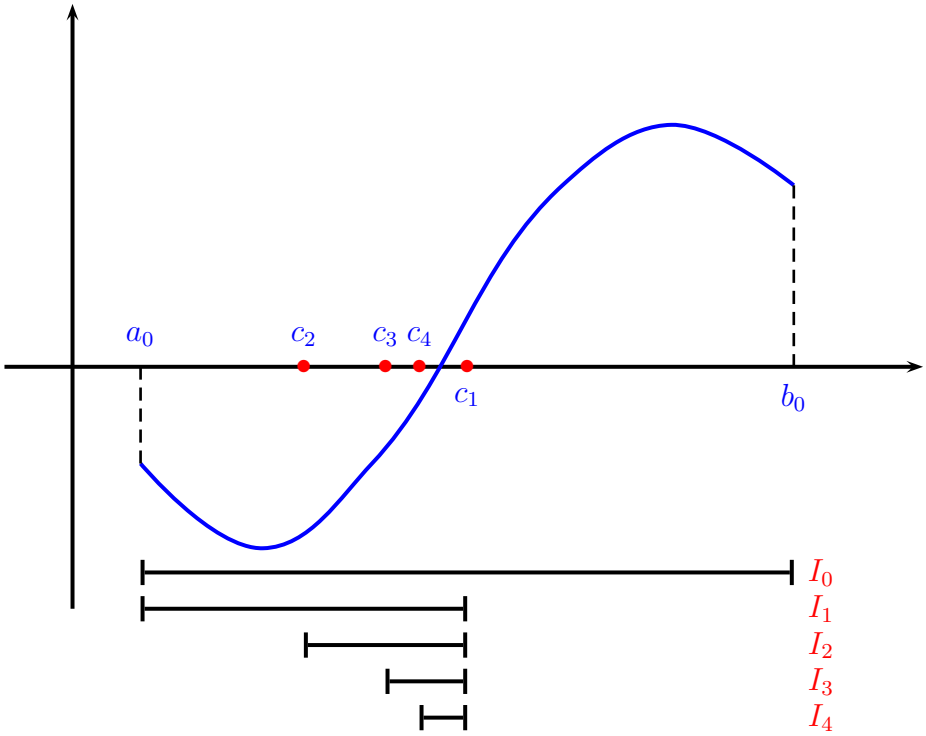
Se $f(c_{k+1}) = 0$ allora $\alpha = c_{k+1}$, altrimenti si pone:

$$I_{k+1} = [a_{k+1}, b_{k+1}] \equiv \begin{cases} a_{k+1} = a_k \quad b_{k+1} = c_k \quad \text{se } f(a_k)f(c_{k+1}) < 0 \\ a_{k+1} = c_{k+1} \quad b_{k+1} = b_k \quad \text{se } f(a_k)f(c_{k+1}) > 0. \end{cases}$$

La successione di intervalli I_k così costruita soddisfa automaticamente le condizioni 1) e 2). Per quanto riguarda la 3) abbiamo:

$$b_k - a_k = \frac{b_{k-1} - a_{k-1}}{2} = \frac{b_0 - a_0}{2^k}$$

e dunque l'ampiezza di I_k tende a zero quando $k \to +\infty$.



Generalmente costruendo le successioni $\{a_k\}$ e $\{b_k\}$ accade che la condizione $f(c_k) = 0$, per un certo valore k, non si verifica mai a causa degli errori di arrotondamento. Quindi è necessario stabilire un opportuno criterio di stop che ci permetta di fermare la procedura quando riteniamo di aver raggiunto una precisione soddisfacente. Per esempio si può imporre:

$$b_k - a_k \leq \varepsilon \tag{2.2}$$

dove ε è una prefissata tolleranza. La (2.2) determina anche un limite per il numero di iterate infatti:

$$\frac{b_0 - a_0}{2^k} \leq \varepsilon \qquad \Rightarrow \qquad k > \log_2 \left(\frac{b_0 - a_0}{\varepsilon} \right).$$

Poichè $b_k - \alpha \leq b_k - a_k$, il criterio (2.2) garantisce che α è approssimata da c_{k+1} con un errore assoluto minore di ε. Se $0 \notin [a, b]$ si può usare come criterio di stop

$$\frac{b_k - a_k}{\min(|a_k|, |b_k|)} \leq \varepsilon \tag{2.3}$$

che garantisce che α è approssimata da c_{k+1} con un errore relativo minore di ε. Un ulteriore criterio di stop è fornito dal test:

$$|f(c_k)| \leq \varepsilon. \tag{2.4}$$

È comunque buona norma utilizzare due criteri di stop insieme, per esempio (2.2) e (2.4) oppure (2.3) e (2.4).

2.3.1 Il metodo della falsa posizione

Una variante del metodo delle bisezioni è appunto il metodo della falsa posizione. Partendo sempre da una funzione $f(x)$ continua in un intervallo $[a, b]$ tale che $f(a)f(b) < 0$, in questo caso si approssima la radice considerando l'intersezione della retta passante per i punti $(a, f(a))$ e $(b.f(b))$ con l'asse x. L'equazione della retta è

$$y = f(a) + \frac{f(b) - f(a)}{b - a}(x - a)$$

pertanto il punto c_1, sua intersezione con l'asse x, è:

$$c_1 = a - f(a)\frac{b - a}{f(b) - f(a)}.$$

Si testa a questo punto l'appartenenza della radice α ad uno dei due intervalli $[a, c_1]$ e $[c_1, b]$ e si procede esattamente come nel caso del metodo delle bisezioni, ponendo

$$[a_1, b_1] \equiv \begin{cases} a_1 = a, \quad b_1 = c_1 & \text{se } f(a)f(c_1) < 0 \\ \\ a_1 = c_1, \quad b_1 = b & \text{se } f(a)f(c_1) > 0. \end{cases}$$

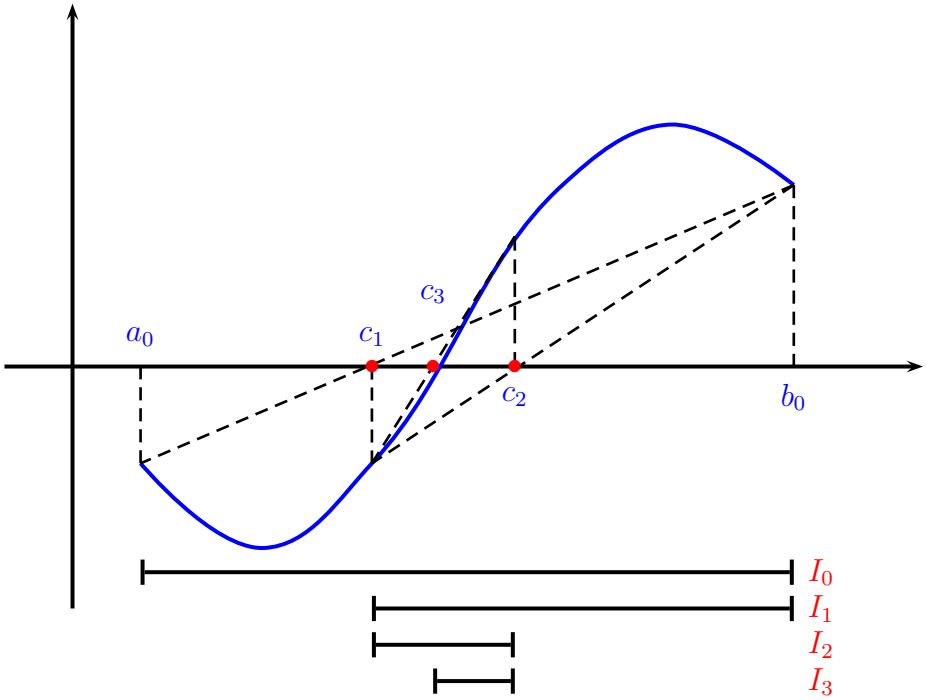
Ad un generico passo k si calcola

$$c_k = a_{k-1} - f(a_{k-1})\frac{b_{k-1} - a_{k-1}}{f(b_{k-1}) - f(a_{k-1})}$$

e si pone

$$[a_k, b_k] \equiv \begin{cases} a_k = a_{k-1} \quad b_k = c_k & \text{se } f(a_{k-1})f(c_k) < 0 \\ a_k = c_k \quad b_k = b_{k-1} & \text{se } f(a_{k-1})f(c_k) > 0. \end{cases}$$

Anche per questo metodo è possibile dimostrare la convergenza nella sola ipotesi di continuità della funzione $f(x)$. Nella seguente figura è rappresentato graficamente il metodo della falsa posizione.




```
function [alfa,k]=bisezione(f,a,b,tol)
%
% La funzione approssima la radice con il metodo di bisezione
%
% Parametri di input
```

```
% f = funzione della quale calcolare la radice
% a = estremo sinistro dell'intervallo
% b = estremo destro dell'intervallo
% tol = precisione fissata
%
% Parametri di output
% alfa = approssimazione della radice
% k = numero di iterazioni
%
if nargin==3
    tol = 1e-8; % Tolleranza di default
end
fa = feval(f,a);
fb = feval(f,b);
if fa*fb>0
    error('Il metodo non e'' applicabile')
end
c = (a+b)/2;
fc = feval(f,c);
k = 0;
while (b-a)>tol | abs(fc)>tol
    if fa*fc<0
        b = c;
        fb = fc;
    else
        a = c;
        fa = fc;
    end
    c = (a+b)/2;
    fc = feval(f,c);
    if nargout==2
        k = k+1;
    end
end
alfa = c;
return
```

2.4 Metodi di Iterazione Funzionale

Il metodo di bisezione può essere applicato ad una vastissima classe di funzioni, in quanto per poter essere applicato si richiede solo la continuità della funzione. Tuttavia ha lo svantaggio di risultare piuttosto lento, infatti ad ogni passo si guadagna in precisione una cifra binaria. Per ridurre l'errore di un decimo sono mediamente necessarie 3.3 iterazioni. Inoltre la velocità di convergenza non dipende dalla funzione $f(x)$ poichè il metodo utilizza esclusivamente il segno assunto dalla funzione in determinati punti e non il suo valore. Il metodo delle bisezioni può essere comunque utilizzato con profitto per determinare delle buone approssimazioni della radice α che possono essere utilizzate dai metodi iterativi che stiamo per descrivere.
Infatti richiedendo alla f supplementari condizioni di regolarità è possibile individuare una vasta classe di metodi che forniscono le stesse approssimazioni del metodo di bisezione utilizzando però un numero di iterate molto minore. In generale questi metodi sono del tipo:

$$x_{k+1} = g(x_k) \qquad k = 0, 1, 2, \ldots \tag{2.5}$$

dove x_0 è un'assegnato valore iniziale e forniscono un'approssimazione delle soluzioni dell'equazione

$$x = g(x). \tag{2.6}$$

Ogni punto α tale che $\alpha = g(\alpha)$ si dice punto fisso o punto unito di g.
Per poter applicare uno schema del tipo (2.5) all'equazione $f(x) = 0$, bisogna prima trasformare questa nella forma (2.6). Ad esempio se $[a, b]$ è l'intervallo di definizione di f ed $h(x)$ è una qualunque funzione tale che $h(x) \neq 0$, per ogni $x \in [a, b]$, si può porre:

$$g(x) = x - \frac{f(x)}{h(x)}. \tag{2.7}$$

Ovviamente ogni punto fisso di g è uno zero di f e viceversa.

Teorema 2.4.1 *Sia $g \in \mathcal{C}([a, b])$ e assumiamo che la successione $\{x_k\}$ generata da (2.5) sia contenuta in $[a, b]$. Allora se tale successione converge, il limite è il punto fisso di g.*

Dimostrazione.

$$\alpha = \lim_{k\to+\infty} x_{k+1} = \lim_{k\to+\infty} g(x_k) = g(\lim_{k\to+\infty} x_k) = g(\alpha). \ \square$$

Teorema 2.4.2 *Sia α punto fisso di g e $g \in \mathcal{C}^1([\alpha-\rho, \alpha+\rho])$, per qualche $\rho > 0$. Scelto x_0 tale che*

$$|x_0 - \alpha| \leq \rho$$

per la successione $\{x_k\}_{k=0}^{\infty}$ generata da (2.5) si ha che se $|g'(x)| < 1$, per $|x-\alpha| \leq \rho$, allora $|x_k - \alpha| \leq \rho$, per ogni k, e la successione $\{x_k\}$ converge a α.

Dimostrazione. Sia

$$\lambda = \max_{|x-\alpha|\leq\rho} |g'(x)| < 1.$$

Proviamo per induzione che tutti gli elementi della successione $\{x_k\}$ sono contenuti nell'intervallo di centro α e ampiezza 2ρ. Per $k = 0$ si ha banalmente $x_0 \in [\alpha-\rho, \alpha+\rho]$. Assumiamo che $|x_k - \alpha| \leq \rho$ e dimostriamolo per $k+1$.

$$|x_{k+1} - \alpha| = |g(x_k) - g(\alpha)| = |g'(\xi_k)||x_k - \alpha|$$

dove $|\xi_k - \alpha| < |x_k - \alpha| \leq \rho$. Pertanto

$$|x_{k+1} - \alpha| \leq \lambda|x_k - \alpha| < |x_k - \alpha| \leq \rho.$$

Proviamo ora che:

$$\lim_{k\to+\infty} x_k = \alpha.$$

Da $|x_{k+1} - \alpha| \leq \lambda|x_k - \alpha|$ segue

$$|x_{k+1} - \alpha| \leq \lambda^{k+1}|x_0 - \alpha|.$$

Conseguentemente qualunque sia x_0 si ha:

$$\lim_{k\to+\infty} |x_k - \alpha| = 0 \Leftrightarrow \lim_{k\to+\infty} x_k = \alpha. \ \square$$

Nella seguente figura viene rappresentata l'interpretazione geometrica di un metodo di iterazione funzionale in ipotesi di convergenza.

Definizione 2.4.1 *Un metodo iterativo del tipo (2.5) si dice localmente convergente ad una soluzione α del problema $f(x) = 0$ se esiste un intervallo $[a, b]$ contenente α tale che, per ogni $x_0 \in [a, b]$, la successione generata da (2.5) converge a α.*

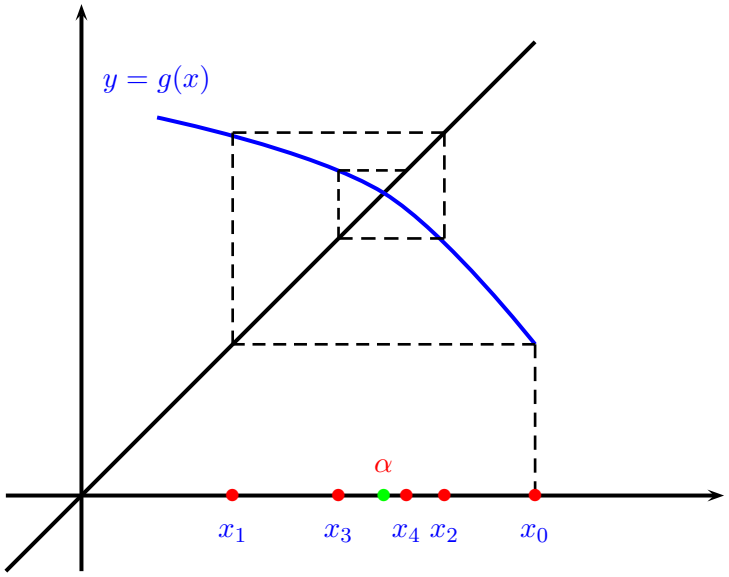


Figura 2.1: Interpretazione geometrica del processo $x_{k+1} = g(x_k)$, se $-1 < g'(\alpha) \leq 0$.

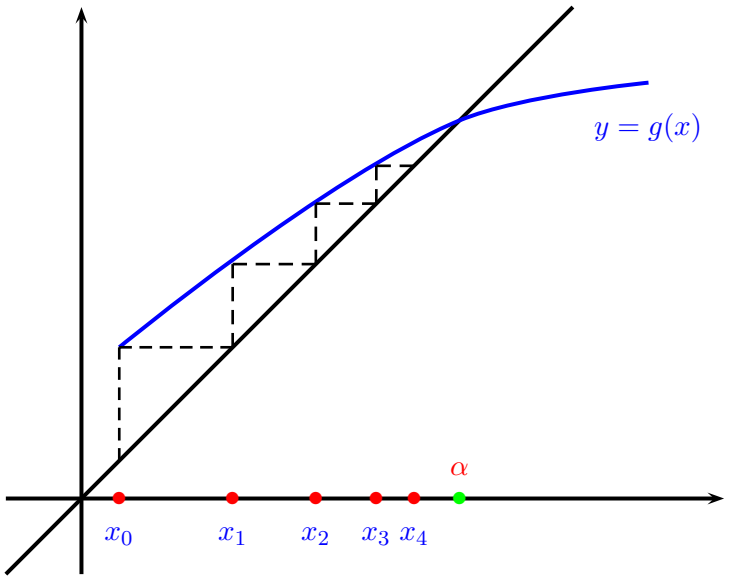


Figura 2.2: Interpretazione geometrica del processo $x_{k+1} = g(x_k)$, se $0 \leq g'(\alpha) < 1$.

Una volta determinata una condizione sufficiente per la convergenza della successione $\{x_k\}$ ad un punto fisso di $g(x)$ si deve essere sicuri che tale punto fisso è unico. Infatti se, altre ad α esistesse anche $\beta \in [a, b]$ tale che $\beta = g(\beta)$, con $\alpha \neq \beta$, allora

$$|\alpha - \beta| = |g(\alpha) - g(\beta)| = |g'(\xi)||\alpha - \beta|$$

con $\xi \in [a, b]$. Poichè $|g'(\xi)| < 1$ si ha:

$$|\alpha - \beta| < |\alpha - \beta|$$

e ciò è assurdo.
Come abbiamo già visto nel caso del metodo delle bisezioni anche per metodi di iterazione funzionale è necessario definire dei criteri di arresto per il calcolo delle iterazioni. Teoricamente, una volta stabilita la precisione voluta, ε, si dovrebbe arrestare il processo iterativo quando l'errore al passo k

$$e_k = |\alpha - x_k|$$

risulta minore della tolleranza prefissata ε. In pratica l'errore non può essere noto quindi è necessario utilizzare qualche stima. Per esempio si potrebbe considerare la differenza tra due iterate consecutive e fermare il calcolo degli elementi della successione quando

$$|x_{k+1} - x_k| \leq \varepsilon,$$

oppure

$$\frac{|x_{k+1} - x_k|}{\min(|x_{k+1}|, |x_k|)} \leq \varepsilon \qquad |x_{k+1}|, |x_k| \neq 0$$

se i valori hanno un ordine di grandezza particolarmente elevato. Una stima alternativa valuta il residuo della funzione rispetto al valore in α, cioè

$$|f(x_k)| \leq \varepsilon.$$

2.4.1 Ordine di Convergenza

Per confrontare differenti metodi iterativi che approssimano la stessa radice α di $f(x) = 0$, si può considerare la velocità con cui tali successioni convergono verso α. Lo studio della velocità di convergenza passa attraverso il concetto di ordine del metodo.

Definizione 2.4.2 *Sia $\{x_k\}_{k=0}^{\infty}$ una successione convergente ad α e tale che $x_k \neq \alpha$, per ogni k. Se esiste un numero reale $p \geq 1$ tale che*

$$\lim_{k\to+\infty} \frac{|x_{k+1}-\alpha|}{|x_k-\alpha|^p} = \gamma \qquad con \quad \begin{cases} 0 < \gamma \leq 1 & se\ p = 1 \\ \\ \gamma > 0 & se\ p > 1 \end{cases} \tag{2.8}$$

allora si dice che la successione ha ordine di convergenza p. La costante γ prende il nome di costante asintotica di convergenza.

In particolare se $p = 1$ e $0 < \gamma < 1$ allora la convergenza si dice *lineare*, se $p = 1$ e $\gamma = 1$ allora la convergenza si dice *sublineare*, mentre se $p > 1$ allora la convergenza si dice superlineare.
Osservazione. La relazione (2.8) implica che esiste una costante positiva β ($\beta \simeq \gamma$) tale che, per k sufficientemente grande:

$$|x_{k+1}-\alpha| \leq \beta |x_k-\alpha|^p \tag{2.9}$$

ed anche

$$\frac{|x_{k+1}-\alpha|}{|\alpha|} \leq \beta |\alpha|^{p-1} \left| \frac{x_k-\alpha}{\alpha} \right|^p . \tag{2.10}$$

Le (2.9) e (2.10) indicano che la riduzione di errore (assoluto o relativo) ad ogni passo è tanto maggiore quanto più alto è l'ordine di convergenza e, a parità di ordine, quanto più piccola è la costante asintotica di convergenza.

Teorema 2.4.3 *Sia $\{x_k\}_{k=0}^{\infty}$ una successione generata dallo schema (2.5) convergente ad α punto fisso di $g \in \mathcal{C}^1([\alpha-\rho, \alpha+\rho])$, $\rho > 0$. Se la convergenza della successione $\{x_k\}_{k=0}^{\infty}$ è lineare (risp. sublineare) allora:*

$$0 < |g'(\alpha)| < 1 \qquad (risp.\ |g'(\alpha)| = 1)$$

Dimostrazione.

$$x_{k+1} - \alpha = g(x_k) - g(\alpha)$$

da cui:

$$\frac{|x_{k+1}-\alpha|}{|x_k-\alpha|} = g'(\xi_k) \qquad \text{con } |\xi_k - \alpha| < |x_k - \alpha|.$$

Poichè

$$\lim_{k\to+\infty} |g'(\xi_k)| = |g'(\alpha)| = \gamma$$

la tesi segue direttamente dalla definizione di convergenza lineare (risp. sublineare). □

Teorema 2.4.4 *(Enunciato) Sia $\alpha \in [a,b]$ punto fisso di $g \in \mathcal{C}^1([a,b])$.*
1) Se $0 < |g'(\alpha)| < 1$ esiste $\rho > 0$ tale che per ogni x_0, $|x_0 - \alpha| < \rho$, la successione $\{x_k\}$ generata da (2.5) è convergente ed ha convergenza lineare.
2) Se $|g'(\alpha)| = 1$ ed esiste $\rho > 0$ tale che $0 < |g'(x)| < 1$, se $|x - \alpha| < \rho$, allora per ogni x_0, $|x_0 - \alpha| < \rho$, la successione $\{x_k\}$ generata da (2.5) è convergente ed ha convergenza sublineare. □

Teorema 2.4.5 *Sia α punto fisso di $g \in \mathcal{C}^p([a,b])$, $p \geq 2$, intero. Se per $x_0 \in [a,b]$ la successione $\{x_k\}$ generata dal metodo (2.5) converge a α con ordine p allora:*

$$g'(\alpha) = g''(\alpha) = \cdots = g^{(p-1)}(\alpha) = 0, \qquad g^{(p)}(\alpha) \neq 0.$$

Dimostrazione. Poichè la successione converge con ordine p, risulta:

$$\lim_{k \to +\infty} \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|^r} = 0 \qquad \text{se } r < p. \tag{2.11}$$

Dalla formula di Taylor

$$x_{k+1} - \alpha = g(x_k) - g(\alpha) = g'(\alpha)(x_k - \alpha) + \frac{1}{2} g''(\xi_k)(x_k - \alpha)^2$$

con $|\xi_k - \alpha| < |x_k - \alpha|$, ovvero

$$\frac{x_{k+1} - \alpha}{x_k - \alpha} = g'(\alpha) + \frac{1}{2} g''(\xi_k)(x_k - \alpha).$$

Passando al limite e tenendo conto della (2.11) e del fatto che $g''(x)$ è limitata segue

$$g'(\alpha) = 0.$$

Assumiamo ora che $g^{(i)}(\alpha) = 0$, per $i = 1, \ldots, r-1$, con $r < p$, e proviamolo per $i = r < p$. Dalla formula di Taylor

$$x_{k+1} - \alpha = g(x_k) - g(\alpha) = \sum_{i=1}^{r} \frac{g^{(i)}(\alpha)}{i!}(x_k - \alpha)^i + \frac{g^{(r+1)}(\sigma_k)}{(r+1)!}(x_k - \alpha)^{r+1}.$$

con $|\sigma_k - \alpha| < |x_k - \alpha|$. Per l'ipotesi induttiva:

$$x_{k+1} - \alpha = \frac{g^{(r)}(\alpha)}{r!}(x_k - \alpha)^r + \frac{g^{(r+1)}(\sigma_k)}{(r+1)!}(x_k - \alpha)^{r+1}.$$

In virtù della (2.11) e della limitatezza di $g^{(r+1)}(x)$ si ha

$$g^{(r)}(\alpha) = 0.$$

Infine, poichè:

$$\frac{x_{k+1} - \alpha}{(x_k - \alpha)^p} = \frac{1}{p!} g^{(p)}(\eta_k) \qquad |\eta_k - \alpha| < |x_k - \alpha|$$

si ha

$$\lim_{k \to +\infty} \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|^p} = \frac{1}{p!} |g^{(p)}(\alpha)| \neq 0. \ \square$$

Vale anche il viceversa.

Teorema 2.4.6 *Sia $\alpha \in [a, b]$ punto fisso di $g \in \mathcal{C}^p([a, b])$, $p \geq 2$, intero. Se*

$$g'(\alpha) = g''(\alpha) = \cdots = g^{(p-1)}(\alpha) = 0, \qquad g^{(p)}(\alpha) \neq 0$$

allora esiste $\rho > 0$ tale che per ogni $x \in [\alpha - \rho, \alpha + \rho]$ la successione $\{x_k\}$ generata da (2.5) è convergente con ordine di convergenza p.

Dimostrazione. Poichè $g'(\alpha) = 0$ esiste $\rho > 0$ tale che $|g'(x)| < 1$ per $|x - \alpha| < \rho$, e la convergenza segue dal teorema (1.2.1). Inoltre per ogni successione $\{x_k\}$ ottenuta da (2.5) si ha

$$\lim_{k \to +\infty} \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|^r} = \frac{1}{r!} |g^{(r)}(\alpha)| = 0 \qquad \text{per } r < p.$$

e

$$\gamma = \frac{1}{p!} |g^{(p)}(\alpha)| > 0,$$

e quindi la successione ha ordine di convergenza p. $\square$

Osservazione. L'ordine di convergenza p può essere anche un numero non intero. In questo caso, posto $q = [p]$, se $g \in \mathcal{C}^q([a, b])$ si ha anche

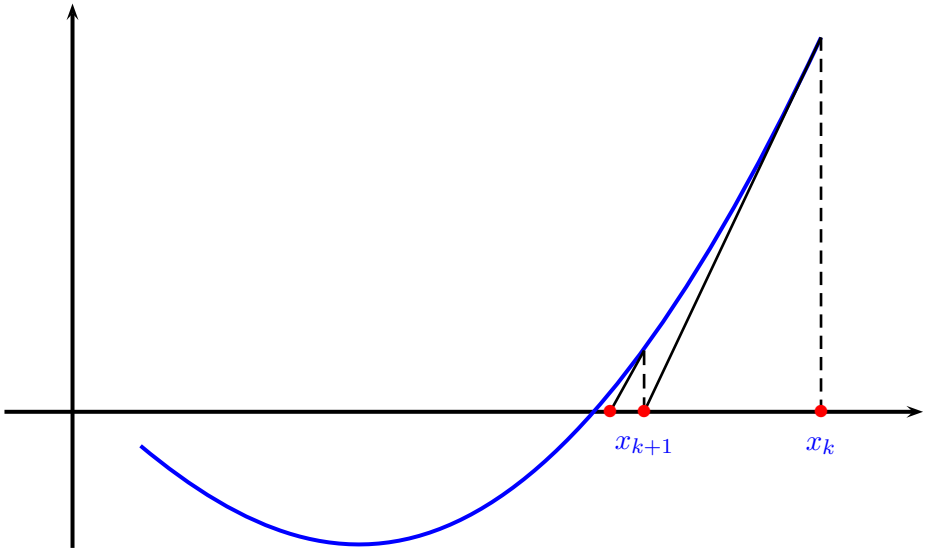
$$g'(\alpha) = g''(\alpha) = \cdots = g^{(q)}(\alpha) = 0,$$

e che g non ha derivata di ordine $q + 1$ altrimenti per il precedente teorema tutte le successioni ottenute da (2.5) a partire da $x_0 \in [\alpha - \rho, \alpha + \rho]$ avrebbero ordine almeno $q + 1$.

Definizione 2.4.3 *Un metodo iterativo convergente ad α si dice di ordine p (di ordine almeno p) se tutte le successioni ottenute al variare del punto iniziale in un opportuno intorno di α convergono con ordine di convergenza p (almeno p).*

2.4.2 Metodo di Newton-Raphson

Nell'ipotesi che f sia derivabile ed ammetta derivata prima continua allora un altro procedimento per l'approssimazione dello zero della funzione $f(x)$ è il metodo di Newton-Raphson, noto anche come metodo delle tangenti. Nella figura seguente è riportata l'interpretazione geometrica di tale metodo. A partire dall'approssimazione x_0 si considera la retta tangente la funzione f passante per il punto P_0 di coordinate $(x_0, f(x_0))$. Si calcola l'ascissa x_1 del punto di intersezione tra tale retta tangente e l'asse delle x e si ripete il procedimento a partire dal punto P_1 di coordinate $(x_1, f(x_1))$. Nella seguente figura è rappresentato graficamente il metodo di Newton-Raphson.



È facile vedere che il metodo definisce il seguente processo iterativo:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \qquad k = 0, 1, 2, \ldots \tag{2.12}$$

che equivale, scegliendo in (2.7) $h(x) = f'(x)$, al metodo di iterazione funzionale in cui la funzione $g(x)$ è

$$g(x) = x - \frac{f(x)}{f'(x)}. \tag{2.13}$$

Per esaminare la convergenza del metodo consideriamo che per ipotesi $f'(x) \neq 0$, per $x \in [a, b]$, dove $[a, b]$ è un opportuno intervallo contenente α. Calcolia-

mo quindi la derivata prima di $g(x)$:

$$g'(x) = \frac{f(x)f''(x)}{[f'(x)]^2}. \tag{2.14}$$

Poichè α è semplice risulta $f'(\alpha) \neq 0$ e quindi:

$$g'(\alpha) = \frac{f(\alpha)f''(\alpha)}{[f'(\alpha)]^2} = 0$$

esiste quindi un intorno di α nel quale $|g'(x)| < 1$, per ogni x, e per il teorema (2.4.2) comunque si sceglie un punto iniziale appartenente a tale intorno il metodo di Newton-Raphson risulta convergente. Se la radice α ha molteplicità $r > 1$ l'ordine di convergenza del metodo non è più 2. Se x_0 è sufficientemente vicino ad α è $|g'(x)| < 1$ e quindi per il teorema 2.4.2 il metodo è ancora convergente ma l'ordine di convergenza è 1.

```
function [alfa,k]=newton(f,f1,x0,tol,Nmax)
%
% La funzione calcolo un'approssimazione
% della radice con il metodo di Newton-Raphson
%
% Parametri di input
% f = funzione della quale calcolare la radice
% f1 = derivata prima della funzione f
% x0 = approssimazione iniziale della radice
% tol = precisione fissata
% Nmax = numero massimo di iterazioni fissate
%
% Parametri di output
% alfa = approssimazione della radice
% k = numero di iterazioni
%
if nargin==3
    tol=1e-8;
    Nmax=1000;
end
k=0;
x1=x0-feval(f,x0)/feval(f1,x0);
```

```
fx1 = feval(f,x1);
while abs(x1-x0)>tol | abs(fx1)>tol
    x0 = x1;
    x1 = x0-feval(f,x0)/feval(f1,x0);
    fx1 = feval(f,x1);
    k=k+1;
    if k>Nmax
        disp('Il metodo non converge');
        alfa = inf;
        break
    end
end
alfa=x1;
return
```

Esempio 2.4.1 *Approssimare il numero* $\alpha = \sqrt[4]{3}$.

Il numero α è lo zero della funzione

$$f(x) = x^4 - 3.$$

Poichè $f(0) < 0$ e $f(3) > 0$ allora si può applicare il metodo di bisezione ottenendo la seguente sucessione di intervalli:

Intervallo	Punto medio	Valore di f nel punto medio
$[0, 3]$	$c = 1.5$	$f(c) = 2.0625$
$[0, 1.5]$	$c = 0.75$	$f(c) = -2.6836$
$[0.75, 1.5]$	$c = 1.125$	$f(c) = -1.3982$
$[1.125, 1.5]$	$c = 1.3125$	$f(c) = -0.0325$
$\vdots$	$\vdots$	$\vdots$

Dopo 10 iterazioni $c = 1.3154$ mentre $\alpha = 1.3161$, e l'errore è pari circa a $6.4433 \cdot 10^{-4}$.
Applicando il metodo di Newton-Raphson,

$$f'(x) = 4x^3$$

si ottiene il processo iterativo

$$x_{k+1} = x_k - \frac{x_k^4 - 3}{4x_k^3}.$$

Poichè per $x > 0$ la funzione è monotona crescente allora si può scegliere $x_0 = 3$ come approssimazione iniziale, ottenendo la seguente successione:

$x_0 = 3$	$f(x_0) = 78$
$x_1 = 2.2778$	$f(x_1) = 23.9182$
$x_2 = 1.7718$	$f(x_2) = 6.8550$
$x_3 = 1.4637$	$f(x_3) = 1.5898$
$x_4 = 1.3369$	$f(x_4) = 0.1948$
$x_5 = 1.3166$	$f(x_5) = 0.0044$
$\vdots$	$\vdots$

Dopo 10 iterazioni l'approssimazione è esatta con un errore dell'ordine di 10^{-16}.

2.4.3 Il metodo della direzione costante

Se applicando ripetutamente la formula di Newton-Raphson accade che la derivata prima della funzione $f(x)$ si mantiene sensibilmente costante allora si può porre

$$M = f'(x)$$

e applicare la formula

$$x_{k+1} = x_k - \frac{f(x_k)}{M} \tag{2.15}$$

anzichè la (2.12). La (2.15) definisce un metodo che viene detto metodo di Newton semplificato oppure metodo della direzione costante in quanto geometricamente equivale all'applicazione del metodo di Newton in cui anzichè prendere la retta tangente la curva f si considera la retta avente coefficiente angolare uguale a M. La funzione iteratrice del metodo è

$$g(x) = x - \frac{f(x)}{M}$$

ed il metodo è convergente se

$$|g'(x)| = \left|1 - \frac{f'(x)}{M}\right| < 1$$

da cui si deduce che è necessario che $f'(x)$ ed M abbiano lo stesso segno.

2.4.4 Il Metodo della Secante

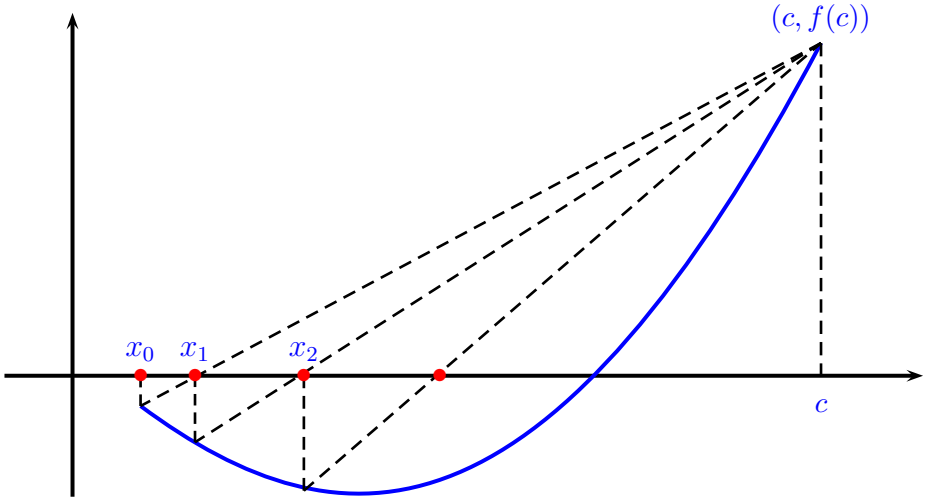
Il metodo della secante è definito dalla relazione

$$x_{k+1} = x_k - f(x_k)\frac{x_k - c}{f(x_k) - f(c)}$$

dove $c \in [a, b]$. Il significato geometrico di tale metodo è il seguente: ad un generico passo k si considera la retta congiungente i punti di coordinate $(x_k, f(x_k))$ e $(c, f(c))$ e si pone x_{k+1} pari al punto di intersezione di tale retta con l'asse x. Dalla formula si evince che la funzione iteratrice del metodo è

$$g(x) = x - f(x)\frac{x - c}{f(x) - f(c)}.$$

Il metodo è rappresentato graficamente nella seguente figura.



In base alla teoria vista nei paragrafi precedenti il metodo ha ordine di convergenze 1 se $g'(\alpha) \neq 0$. Può avere ordine di convergenza almeno 1 se $g'(\alpha) = 0$. Tale eventualità si verifica se la tangente alla curva in α ha lo stesso coefficiente angolare della retta congiungente i punti $(\alpha, 0)$ e $(c, f(c))$.
Poichè il metodo delle secanti ha lo svantaggio di avere, solitamente, convergenza lineare mentre il metodo di Newton-Raphson, pur avendo convergenza quadratica, ha lo svantaggio di richiedere, ad ogni passo, due valutazioni di funzioni: $f(x_k)$ ed $f'(x_k)$, quindi se il costo computazionale di $f'(x_k)$ è molto più elevato rispetto a quello di $f(x_k)$ può essere più conveniente l'uso di metodi che necessitano solo del calcolo del valore della funzione $f(x)$.

2.5 Il Metodo di Newton a Doppio Passo

Analizziamo in questa sezione una particolare applicazione del metodo di Newton per il calcolo degli zeri di un polinomio. In particolare supponiamo che il polinomio a coefficienti reali

$$p(x) = \sum_{i=0}^{n} a_i x^{n-i} \tag{2.16}$$

ammetta solo le radici reali e distinte

$$\alpha_1 > \alpha_2 > \alpha_3 > \cdots > \alpha_n.$$

In questo caso il metodo di Newton fornisce il seguente schema iterativo:

$$x_{k+1} = x_k - \frac{p(x_k)}{p'(x_k)} \qquad k = 0, 1, 2, \ldots \tag{2.17}$$

In questo caso vale il seguente teorema.

Teorema 2.5.1 *Sia $p(x)$ un polinomio reale di grado $n \geq 1$, che possiede solo radici reali α_i,*

$$\alpha_1 > \alpha_2 > \alpha_3 > \cdots > \alpha_n$$

allora il metodo di Newton, per ogni valore iniziale $x_0 > \alpha_1$ fornisce una successione monotona decrescente $\{x_k\}$ di approssimazioni, convergente verso α_1. Per $n \geq 2$ la successione delle x_k è strettamente monotona decrescente. □

Tuttavia nonostante la convergenza quadratica del metodo di Newton, se x_k non si trova abbastanza vicino ad uno zero semplice di $p(x)$, la successione delle x_k può avvicinarsi molto lentamente alla radice per piccoli valori di k; ciò accade quando, per esempio, è stato scelto un valore iniziale x_0 molto grande. Per x_k molto grande si ha infatti:

$$x_{k+1} = x_k - \frac{a_0 x_k^n + \ldots}{n a_0 x_k^{n-1} + \ldots} \simeq x_k \left(1 - \frac{1}{n}\right).$$

pertanto si preferisce applicare una variante del metodo di Newton, noto appunto come Metodo di Newton a Doppio Passo:

$$x_{k+1} = x_k - 2\frac{p(x_k)}{p'(x_k)} \qquad k = 0, 1, 2, \ldots. \tag{2.18}$$

Questo metodo non converge alla radice α_1 e pertanto viene utilizzato per far sì che la successione $\{x_k\}$ si avvicini più rapidamente ad α_1. Tuttavia per un certo indice m accade che:

$$\begin{array}{lll} p(x_j)p(x_{j-1}) & > 0 & j = 1, 2, \ldots, m-1 \\ p(x_m)p(x_{m-1}) & < 0 & \end{array}$$

e questo vuol dire:

$$x_m < \alpha_1 < x_{m-1} < x_{m-2} < \cdots < x_1 < x_0$$

cioè la successione delle x_k scavalca la radice α_1. A questo punto si può applicare il metodo di Newton semplice prendendo come punto iniziale proprio x_m. Infatti si può dimostrare che il punto x_m pur scavalcando la radice α_1 non scavalca il punto ξ_1, compreso tra α_2 e α_1, che annulla la derivata prima del polinomio $p(x)$. La successione $\{x_k\}$ così costruita converge alla radice α_1. Un altro problema da risolvere a questo punto è la scelta del valore iniziale x_0. Per quello possiamo sfruttare una delle maggiorazioni che vengono fornite dal seguente teorema.

Teorema 2.5.2 *Per ogni zero α_i del polinomio (2.16) risulta:*

$$|\alpha_i| \leq \max\left\{\left|\frac{a_n}{a_0}\right|, 1 + \left|\frac{a_{n-1}}{a_0}\right|, \ldots, 1 + \left|\frac{a_1}{a_0}\right|\right\}$$

$$|\alpha_i| \leq \max\left\{1, \sum_{i=1}^{n}\left|\frac{a_i}{a_0}\right|\right\}$$

$$|\alpha_i| \leq \max\left\{\left|\frac{a_n}{a_{n-1}}\right|, 2\left|\frac{a_{n-1}}{a_{n-2}}\right|, \ldots, 2\left|\frac{a_1}{a_0}\right|\right\}$$

$$|\alpha_i| \leq \sum_{i=0}^{n-1}\left|\frac{a_{i+1}}{a_i}\right|$$

$$|\alpha_i| \leq 2\max\left\{\left|\frac{a_1}{a_0}\right|, \sqrt{\left|\frac{a_2}{a_0}\right|}, \sqrt[3]{\left|\frac{a_3}{a_0}\right|}, \ldots, \sqrt[n]{\left|\frac{a_n}{a_0}\right|}\right\}.$$

Dopo aver determinato una buona approssimazione $\hat{\alpha}_1$ della radice più grande di $p(x)$ si pone il problema di determinare gli altri zeri. Un metodo può essere quello di calcolare, utilizzando per esempio la regola di Ruffini, il polinomio di grado $n-1$:

$$p_1(x) = \frac{p(x)}{x - \hat{\alpha}_1} \tag{2.19}$$

il cui massimo zero è ora α_2 e applicare il metodo di Newton. Tuttavia l'applicazione della (2.19) per il calcolo di $p_1(x)$ fa sì che l'errore con cui la radice α_1 è approssimata da $\hat{\alpha}_1$ si propaghi al calcolo dei coefficienti di $p_1(x)$ con l'effetto di ottenere un polinomio che sicuramente non ammette come zeri α_2, $\alpha_3, \ldots, \alpha_n$. Per evitare questo problema conviene non calcolare direttamente il polinomio $p_1(x)$ ma ricorrere alla cosiddetta Variante di Maehly. Infatti per il polinomio $p_1(x)$ abbiamo:

$$p_1'(x) = \frac{p'(x)}{x - \hat{\alpha}_1} - \frac{p(x)}{(x - \hat{\alpha}_1)^2}$$

e quindi il metodo di Newton a doppio passo è definito dalla formula:

$$x_{k+1} = x_k - 2\frac{p_1(x_k)}{p_1'(x_k)} = x_k - 2\frac{p(x_k)}{p'(x_k) - \dfrac{p(x_k)}{x_k - \hat{\alpha}_1}}. \tag{2.20}$$

Quindi per ottenere un'approssimazione di α_2 si considera nuovamente il metodo di Newton a doppio passo (2.20) prendendo come punto iniziale proprio il valore x_m trovato dopo lo scavalcamento di α_1. In generale per la derivata prima del polinomio

$$p_j(x) = \frac{p(x)}{(x - \hat{\alpha}_1)\ldots(x - \hat{\alpha}_j)}$$

vale la formula

$$p_j'(x) = \frac{p'(x)}{(x - \hat{\alpha}_1)\ldots(x - \hat{\alpha}_j)} - \frac{p(x)}{(x - \hat{\alpha}_1)\ldots(x - \hat{\alpha}_j)} \sum_{i=1}^{j} \frac{1}{x - \hat{\alpha}_i}.$$

Il metodo di Newton per la determinazione di α_{j+1} nella variante di Maehly assume la forma:

$$x_{k+1} = x_k - \frac{p(x_k)}{p'(x_k) - \displaystyle\sum_{i=1}^{j} \frac{p(x_k)}{x_k - \hat{\alpha}_i}}.$$

Il vantaggio di questa formula sta nel fatto che la successione delle x_k converge localmente in maniera quadratica verso la radice α_{j+1} indipendentemente dall'errore del quale i numeri $\hat{\alpha}_1, \ldots, \hat{\alpha}_j$ sono affetti in quanto approssimazioni delle radici del polinomio $p(x)$.

```
function [alfa, iter]=doppio(p,p1,epsilon)
%
% La funzione doppio(p,p1,epsilon) calcola gli zeri di un
% polinomio a coefficienti e radici reali utilizzando il
% metodo di Newton a doppio passo e la variante di Maehly.
%
% Parametri di input
%
% p = vettore dei coefficienti del polinomio
% p1 = vettore dei coefficienti della derivata
%      prima del polinomio
% epsilon = tolleranza prefissata
%
n=length(p)-1;
j=0;
alfa=[];
x0=inizio(p);
%
% La funzione inizio calcola l'approssimazione iniziale
%
while j<n
   x1=x0-2*newton(p,p1,x0,j,alfa);
   while polyval(p,x1)*polyval(p,x0)>0
        x0=x1;
        x1=x0-2*newton(p,p1,x0,j,alfa)
   end
   x2=x1;
%
% Scavalcamento avvenuto
% Si memorizza in x2 il valore assunto della successione
% dopo lo scavalcamento e che sara' il punto iniziale per
% il calcolo della successiva radice
%
```

```
   while abs(polyval(p,x1))>epsilon | abs(x1-x0)>epsilon
       x0=x1;
       x1=x0-newton(p,p1,x0,j,alfa)
   end
   alfa=[alfa x1];
   j=j+1;
   x0=x2;
end
return
```

La funzione `newton` calcola i rapporti $p(x_k)/p'(x_k)$ al primo passo e ai passi successivi quelli della variante di Maehly.

```
function y=newton(p,p1,x,j,alfa)
px=polyval(p,x);
p1x=polyval(p1,x);
somma=0;
for i=1:j
  somma=somma+1/(x-alfa(i));
end
y=px/(p1x-px*somma);
return
```

Per calcolare l'approssimazione iniziale si sfrutta il Teorema 2.5.2 calcolando la migliore delle 5 approssimazioni proposte.

```
function y=inizio(p)
%
% inizio(p) calcola l'approssimazione iniziale per il calcolo
% della radice piu' grande
%
n=length(p);
pp(1)=abs(p(n)/p(1));
for i=2:n
  pp(i)=1+abs(p(i)/p(1));
end
mx(1)=max(pp);
clear pp
pp(1)=1;
```

```
pp(2)=sum(abs(p(2:n)/p(1)));
mx(2)=max(pp);
clear pp
pp(1)=abs(p(n)/p(n-1));
for i=2:n
  pp(i)=2*abs(p(i)/p(i-1));
end
mx(3)=max(pp);
clear pp
mx(4)=0;
for i=1:n-1
  mx(4)=mx(4)+abs(p(i+1)/p(i));
end
pp(1)=abs(p(2)/p(1));
for i=2:n-1
  pp(i)=abs(p(i+1)/p(1))^(1/i);
end
mx(5)=2*max(pp);
y=min(mx);
return
```

Capitolo 3

Metodi diretti per sistemi lineari

3.1 Introduzione

Siano assegnati una matrice non singolare $A \in \mathbb{R}^{n\times n}$ ed un vettore $\mathbf{b} \in \mathbb{R}^n$. Risolvere un sistema lineare avente A come matrice dei coefficienti e $\mathbf{b}$ come vettore dei termini noti significa trovare un vettore $\mathbf{x} \in \mathbb{R}^n$ tale che

$$A\mathbf{x} = \mathbf{b}. \tag{3.1}$$

Esplicitare la relazione (3.1) significa imporre le uguaglianze tra le componenti dei vettori a primo e secondo membro:

$$\begin{array}{rcccl} a_{11}x_1+ & a_{12}x_2+ & \cdots+ & a_{1n}x_n = & b_1 \\ a_{21}x_1+ & a_{22}x_2+ & \cdots+ & a_{2n}x_n = & b_2 \\ \vdots & & & & \vdots \\ a_{n1}x_1+ & a_{n2}x_2+ & \cdots+ & a_{nn}x_n = & b_n. \end{array} \tag{3.2}$$

Le (3.2) definiscono un sistema di n equazioni algebriche lineari nelle n incognite $x_1, x_2, \ldots, x_n$. Il vettore $\mathbf{x}$ viene detto vettore soluzione. Prima di affrontare il problema della risoluzione numerica di sistemi lineari richiamiamo alcuni importanti concetti di algebra lineare.

Definizione 3.1.1 *Se $A \in \mathbb{R}^{n\times n}$ è una matrice di ordine 1, si definisce determinante di A il numero*

$$\det A = a_{11}.$$

Se la matrice A è quadrata di ordine n allora fissata una qualsiasi riga (colonna) di A, diciamo la i-esima (j-esima) allora applicando la cosiddetta regola di Laplace il determinante di A è:

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

dove A_{ij} è la matrice che si ottiene da A cancellando la i-esima riga e la j-esima colonna.

Il determinante è pure uguale a

$$\det A = \sum_{i=1}^{n} a_{ij}(-1)^{i+j} \det A_{ij},$$

cioè il determinante è indipendente dall'indice di riga (o di colonna) fissato. Se A è la matrice di ordine 2

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

allora

$$\det A = a_{11}a_{22} - a_{21}a_{12}.$$

Il determinante ha le seguenti proprietà:
1. Se A è una matrice triangolare o diagonale allora

$$\det A = \prod_{i=1}^{n} a_{ii};$$

2. $\det I = 1$;
3. $\det A^T = \det A$;
4. $\det AB = \det A \det B$ (Regola di Binet);
5. se $\alpha \in \mathbb{R}$ allora $\det \alpha A = \alpha^n \det A$;
6. $\det A = 0$ se una riga (o una colonna) è nulla, oppure una riga (o una colonna) è proporzionale ad un'altra riga (o colonna) oppure è combinazione lineare di due (o più) righe (o colonne) di A.
7. Se A è una matrice triangolare a blocchi

$$A = \begin{bmatrix} B & C \\ O & D \end{bmatrix}$$

con B e D matrici quadrate, allora

$$\det A = \det B \det D. \tag{3.3}$$

Una matrice A di ordine n si dice non singolare se il suo determinante è diverso da zero, in caso contrario viene detta *singolare*. Si definisce inversa di A la matrice A^{-1} tale che:

$$AA^{-1} = A^{-1}A = I_n$$

Per quello che riguarda il determinante della matrice inversa vale la seguente proprietà:

$$\det A^{-1} = \frac{1}{\det A}.$$

Un metodo universalmente noto per risolvere il problema (3.1) è l'applicazione della cosiddetta Regola di Cramer la quale fornisce:

$$x_i = \frac{\det A_i}{\det A} \qquad i = 1, \ldots, n, \tag{3.4}$$

dove A_i è la matrice ottenuta da A sostituendo la sua i-esima colonna con il termine noto $\mathbf{b}$. Dalla (3.4) è evidente che per ottenere tutte le componenti del vettore soluzione è necessario il calcolo di $n+1$ determinanti di ordine n. Calcoliamo ora il numero di operazioni aritmetiche necessario per calcolare une determinante con la regola di Laplace. Indichiamo con $f(n)$ il numero di operazioni aritmetiche su numeri reali necessario per calcolare un determinante di ordine n, ricordando che $f(2) = 3$. La regola di Laplace richiede il calcolo di n determinanti di matrici di ordine $n-1$ (il cui costo computazionale in termini di operazioni è $nf(n-1)$) inoltre n prodotti ed $n-1$ somme algebriche, ovvero

$$f(n) = nf(n-1) + 2n - 1.$$

Per semplicità tralasciamo gli ultimi addendi ottenendo il valore approssimato

$$f(n)\ simeqnf(n-1)$$

Applicando lo stesso ragionamento al numero $f(n-1) \simeq (n-1)f(n-2)$ e in modo iterativo si ottiene

$$f(n) \simeq n(n-1)(n-2)\ldots 3f(2) = \frac{3}{2}\, n!.$$

Se $n = 100$ si ha $100! \simeq 10^{157}$. Anche ipotizzando di poter risolvere il problema con un elaboratore in grado di eseguire miliardi di operazioni al secondo sarebbero necessari diversi anni di tempo per calcolare un singolo determinante. Questo esempio rende chiara la necessità di trovare metodi alternativi per risolvere sistemi lineari, in particolare quando le dimensioni sono particolarmente elevate.

3.2 Risoluzione di sistemi triangolari

Prima di affrontare la soluzione algoritmica di un sistema lineare vediamo qualche particolare sistema che può essere agevolmente risolto. Assumiamo che il sistema da risolvere abbia la seguente forma:

$$\begin{array}{ccccccccc} a_{11}x_1 & +a_{12}x_2 & \dots & +a_{1i}x_i & \dots & +a_{1n}x_n & = b_1 \\ & a_{22}x_2 & \dots & +a_{2i}x_i & \dots & +a_{2n}x_n & = b_2 \\ & & \ddots & \vdots & & \vdots & \vdots \\ & & & a_{ii}x_i & \dots & +a_{in}x_n & = b_i \\ & & & & \ddots & \vdots & \vdots \\ & & & & & a_{nn}x_n & = b_n \end{array} \tag{3.5}$$

con $a_{ii} \neq 0$ per ogni i. In questo caso la matrice A è detta *triangolare superiore*. È evidente che in questo caso, la soluzione è immediatamente calcolabile. Infatti:

$$\begin{cases} x_n = \dfrac{b_n}{a_{nn}} \\ \\ x_i = \dfrac{b_i - \displaystyle\sum_{j=i+1}^{n} a_{ij}x_j}{a_{ii}} \qquad i = n-1, \dots, 1. \end{cases} \tag{3.6}$$

Il metodo (3.6) prende il nome di metodo di sostituzione all'indietro, poichè il vettore $\mathbf{x}$ viene calcolato partendo dall'ultima componente. Anche per il

seguente sistema il vettore soluzione è calcolabile in modo analogo.

$$\begin{array}{lllllll} a_{11}x_1 & & & & & & = b_1 \\ a_{21}x_1 & +a_{22}x_2 & & & & & = b_2 \\ \vdots & \vdots & \ddots & & & & \vdots \\ a_{i1}x_1 & +a_{i2}x_2 & \dots & +a_{ii}x_i & & & = b_i \\ \vdots & \vdots & & & \ddots & & \vdots \\ a_{n1}x_1 & +a_{n2}x_2 & \dots & +a_{ni}x_i & \dots & +a_{nn}x_n & = b_n \end{array} \tag{3.7}$$

In questo caso la matrice dei coefficienti è triangolare inferiore e la soluzione viene calcolata con il metodo di sostituzione in avanti:

$$\begin{cases} x_1 = \dfrac{b_1}{a_{11}} \\ \\ x_i = \dfrac{b_i - \displaystyle\sum_{j=1}^{i-1} a_{ij}x_j}{a_{ii}} \qquad i = 2, \dots, n. \end{cases}$$

Concludiamo questo paragrafo facendo alcune considerazioni sul costo computazionale dei metodi di sostituzione. Per costo computazionale di un algoritmo si intende il numero di operazioni che esso richiede per fornire la soluzione di un determinato problema. Nel caso di algoritmi numerici le operazioni che si contano sono quelle aritmetiche che operano su dati reali. Considerano per esempio il metodo di sostituzione in avanti osserviamo che per calcolare x_1 è necessaria una sola operazione (una divisione), per calcolare x_2 le operazioni sono tre (un prodotto, una differenza e una divisione), mentre il generico x_i richiede $2i-1$ operazioni ($i-1$ prodotti, $i-1$ differenze e una divisione), indicato con $C(n)$ il numero totale di operazioni necessarie è:

$$C(n) = \sum_{i=1}^{n}(2i-1) = 2\sum_{i=1}^{n} i - \sum_{i=1}^{n} 1 = 2\frac{n(n+1)}{2} - n = n^2,$$

sfruttando la proprietà che

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}.$$

3.3 Metodo di Eliminazione di Gauss

L'idea di base del metodo di Gauss è appunto quella di operare delle opportune trasformazioni sul sistema originale $A\mathbf{x} = \mathbf{b}$, che non costino eccessivamente, in modo da ottenere un sistema equivalente avente come matrice dei coefficienti una matrice triangolare superiore.
Supponiamo di dover risolvere il sistema:

$$\begin{array}{rrrrcl} 2x_1 & +x_2 & +x_3 & & = & -1 \\ -6x_1 & -4x_2 & -5x_3 & +x_4 & = & 1 \\ -4x_1 & -6x_2 & -3x_3 & -x_4 & = & 2 \\ 2x_1 & -3x_2 & +7x_3 & -3x_4 & = & 0. \end{array}$$

Il vettore soluzione di un sistema lineare non cambia se ad un'equazione viene sommata un'altra equazione (eventualmente moltiplicata per una costante) oppure una combinazione lineare di alcune equazioni del sistema. L'idea alla base del metodo di Gauss è quella di ottenere un sistema lineare con matrice dei coefficienti triangolare superiore effettuando opportune combinazioni lineari tra le equazioni. Poniamo

$$A^{(1)} = \begin{bmatrix} 2 & 1 & 1 & 0 \\ -6 & -4 & -5 & 1 \\ -4 & -6 & -3 & -1 \\ 2 & -3 & 7 & -3 \end{bmatrix}, \qquad \mathbf{b}^{(1)} = \begin{bmatrix} -1 \\ 1 \\ 2 \\ 0 \end{bmatrix}$$

rispettivamente la matrice dei coefficienti e il vettore dei termini noti del sistema di partenza. Calcoliamo un sistema lineare equivalente a quello iniziale ma che abbia gli elementi sottodiagonali della prima colonna uguali a zero. Azzeriamo ora l'elemento $a_{21}(1)$. Lasciamo inalterata la prima equazione. Poniamo

$$l_{21} = -\frac{a_{21}}{a_{11}} = -\frac{-6}{2} = 3$$

e moltiplichiamo la prima equazione per l_{21} ottenendo:

$$6x_1 + 3x_2 + 3x_3 = -3.$$

La nuova seconda equazione sarà la somma tra la seconda equazione e la prima moltiplicata per l_{21}:

$$\begin{array}{rrrrcll} -6x_1 & -4x_2 & -5x_3 & +x_4 & = 1 & \\ 6x_1 & +3x_2 & +3x_3 & & = -3 & \\ \hline & -x_2 & -2x_3 & +x_4 & = -2 & [\texttt{Nuova seconda equazione}]. \end{array}$$

Prcediamo nello stesso modo per azzerare gli altri elementi della prima colonna. Poniamo

$$l_{31} = -\frac{a_{31}^{(1)}}{a_{11}^{(1)}} = -\frac{-4}{2} = 2$$

e moltiplichiamo la prima equazione per l_{31} ottenendo:

$$4x_1 + 2x_2 + 2x_3 = -2.$$

La nuova terza equazione sarà la somma tra la terza equazione e la prima moltiplicata per l_{31}:

$$\begin{array}{rrrrll} -4x_1 & -6x_2 & -3x_3 & -x_4 & = 2 & \\ 4x_1 & +2x_2 & +2x_3 & & = -2 & \\ \hline & -4x_2 & -x_3 & -x_4 & = 0 & [\texttt{Nuova terza equazione}]. \end{array}$$

Poniamo ora

$$l_{41} = -\frac{a_{41}^{(1)}}{a_{11}^{(1)}} = -\frac{2}{2} = -1$$

e moltiplichiamo la prima equazione per l_{41} ottenendo:

$$-2x_1 - x_2 - x_3 = 1.$$

La nuova quarta equazione sarà la somma tra la quarta equazione e la prima moltiplicata per l_{41}:

$$\begin{array}{rrrrll} 2x_1 & -3x_2 & +7x_3 & -3x_4 & = 0 & \\ -2x_1 & -x_2 & -x_3 & & = 1 & \\ \hline & -4x_2 & +6x_3 & -3x_4 & = 1 & [\texttt{Nuova quarta equazione}]. \end{array}$$

I numeri $l_{21}, l_{3,1}, \ldots$ sono detti moltiplicatori.
Al secondo passo il sistema lineare è diventato:

$$\begin{array}{rrrrl} 2x_1 & +x_2 & +x_3 & & = -1 \\ & -x_2 & -2x_3 & +x_4 & = -2 \\ & -4x_2 & -x_3 & -x_4 & = 0 \\ & -4x_2 & +6x_3 & -3x_4 & = 1. \end{array}$$

La matrice dei coefficienti e il vettore dei termini noti sono diventati:

$$A^{(2)} = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 0 & -1 & -2 & 1 \\ 0 & -4 & -1 & -1 \\ 0 & -4 & 6 & -3 \end{bmatrix}, \qquad \mathbf{b}^{(2)} = \begin{bmatrix} -1 \\ -2 \\ 0 \\ 1 \end{bmatrix}.$$

Cerchiamo ora di azzerare gli elementi sottodiagonali della seconda colonna, a partire da a_{32}, usando una tecnica simile. Innanzitutto osserviamo che non conviene prendere in considerazione una combinazione lineare che coinvolga la prima equazione perchè avendo questa un elemento in prima posizione diverso da zero quando sommata alla terza equazione cancellerà l'elemento uguale a zero in prima posizione. Lasciamo inalterate le prime due equazioni del sistema e prendiamo come equazione di riferimento la seconda. Poichè $a_{22}^{(2)} \neq 0$ poniamo

$$l_{32} = -\frac{a_{32}^{(2)}}{a_{22}^{(2)}} = -\frac{-4}{-1} = -4$$

e moltiplichiamo la seconda equazione per l_{32} ottenendo:

$$4x_2 + 8x_3 - 4x_4 = 8.$$

La nuova terza equazione sarà la somma tra la terza equazione e la seconda appena modificata:

$$\begin{array}{rrrll} -4x_2 & -x_3 & -x_4 & = 0 & \\ 4x_2 & +8x_3 & -4x_4 & = 8 & \\ \hline & 7x_3 & -5x_4 & = 8 & [\texttt{Nuova terza equazione}]. \end{array}$$

Poniamo

$$l_{42} = -\frac{a_{42}^{(2)}}{a_{22}^{(2)}} = -\frac{-4}{-1} = -4$$

e moltiplichiamo la seconda equazione per l_{42} ottenendo:

$$4x_2 + 8x_3 - 4x_4 = 8.$$

La nuova quarta equazione sarà la somma tra la quarta equazione e la seconda appena modificata:

$$\begin{array}{rrrll} -4x_2 & +6x_3 & -3x_4 & = 1 & \\ 4x_2 & +8x_3 & -4x_4 & = 8 & \\ \hline & 14x_3 & -7x_4 & = 9 & [\texttt{Nuova quarta equazione}]. \end{array}$$

Al terzo passo il sistema lineare è diventato:

$$\begin{array}{rrrrl} 2x_1 & +x_2 & +x_3 & & = -1 \\ & -x_2 & -2x_3 & +x_4 & = -2 \\ & & 7x_3 & -5x_4 & = 8 \\ & & 14x_3 & -7x_4 & = 9. \end{array}$$

La matrice dei coefficienti e il vettore dei termini noti sono quindi

$$A^{(3)} = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 0 & -1 & -2 & 1 \\ 0 & 0 & 7 & -5 \\ 0 & 0 & 14 & -7 \end{bmatrix}, \qquad \mathbf{b}^{(2)} = \begin{bmatrix} -1 \\ -2 \\ 8 \\ 9 \end{bmatrix}.$$

Resta da azzerare l'unico elemento sottodiagonali della terza colonna. Lasciamo inalterate le prime tre equazioni del sistema. Poniamo

$$l_{43} = -\frac{a_{43}^{(3)}}{a_{33}^{(3)}} = -\frac{14}{7} = -2$$

e moltiplichiamo la terza equazione per l_{43} ottenendo:

$$-14x_3 + 10x_4 = -16.$$

La nuova quarta equazione sarà la somma tra la quarta equazione e la terza appena modificata:

$$\begin{array}{rrll} 14x_3 & -7x_4 & = -16 & \\ -14x_3 & +10x_4 & = 9 & \\ \hline & 3x_4 & = -7 & [\texttt{Nuova quarta equazione}]. \end{array}$$

Abbiamo ottenuto un sistema triangolare superiore:

$$\begin{array}{rrrrl} 2x_1 & +x_2 & +x_3 & & = -1 \\ & -x_2 & -2x_3 & +x_4 & = 4 \\ & & 7x_3 & -5x_4 & = 8 \\ & & & 3x_4 & = -7. \end{array}$$

La matrice dei coefficienti e il vettore dei termini noti sono diventati:

$$A^{(4)} = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 0 & -1 & -2 & 1 \\ 0 & 0 & 7 & -5 \\ 0 & 0 & 0 & 3 \end{bmatrix}, \qquad \mathbf{b}^{(4)} = \begin{bmatrix} -1 \\ 4 \\ 8 \\ -7 \end{bmatrix}.$$

Vediamo come ciò sia possibile. Riconsideriamo il sistema di equazioni nella sua forma scalare (3.2):

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \qquad i = 1, \ldots, n. \tag{3.8}$$

Per motivi che risulteranno chiari tra poco poniamo $a_{ij}^{(1)} = a_{ij}$ e $b_i^{(1)} = b_i$. Isoliamo in ogni equazione la componente x_1. Abbiamo:

$$a_{11}^{(1)} x_1 + \sum_{j=2}^{n} a_{1j}^{(1)} x_j = b_1^{(1)} \tag{3.9}$$

$$a_{i1}^{(1)} x_1 + \sum_{j=2}^{n} a_{ij}^{(1)} x_j = b_i^{(1)}, \qquad i = 2, \ldots, n. \tag{3.10}$$

Dividiamo l'equazione (3.9) per $a_{11}^{(1)}$ e moltiplichiamo la stessa rispettivamente per $a_{21}^{(1)}, a_{31}^{(1)}, \ldots, a_{n1}^{(1)}$. In questo modo otteniamo $n-1$ nuove equazioni:

$$a_{i1}^{(1)} x_1 + \sum_{j=2}^{n} \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} a_{1j}^{(1)} x_j = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} b_i^{(1)}, \qquad i = 2, \ldots, n. \tag{3.11}$$

Sottraendo da (3.10) le equazioni (3.11) otteniamo:

$$\sum_{j=2}^{n} \left(a_{ij}^{(1)} - \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} a_{1j}^{(1)} \right) x_j = b_i^{(1)} - \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} b_i^{(1)}, \qquad i = 2, \ldots, n. \tag{3.12}$$

L'equazione (3.9) insieme alle (3.12) formano un nuovo sistema di equazioni, equivalente a quello originario, che possiamo scrivere così:

$$\begin{cases} a_{11}^{(1)} x_1 + \displaystyle\sum_{j=2}^{n} a_{1j}^{(1)} x_j = b_1^{(1)} \\ \\ \displaystyle\sum_{j=2}^{n} a_{ij}^{(2)} x_j = b_i^{(2)} \qquad i = 2, \ldots, n \end{cases} \tag{3.13}$$

dove

$$\begin{cases} a_{ij}^{(2)} = a_{ij}^{(1)} - \dfrac{a_{i1}^{(1)}}{a_{11}^{(1)}} a_{1j}^{(1)} \qquad i, j = 2, \ldots, n \\ \\ b_i^{(2)} = b_i^{(1)} - \dfrac{a_{i1}^{(1)}}{a_{11}^{(1)}} b_1^{(1)} \qquad i = 2, \ldots, n. \end{cases} \tag{3.14}$$

Osserviamo che la matrice dei coefficienti del sistema (3.13) è la seguente

$$A^{(2)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \dots & a_{nn}^{(2)} \end{bmatrix}.$$

Ora a partire dal sistema di equazioni

$$\sum_{j=2}^{n} a_{ij}^{(2)} x_j = b_i^{(2)} \qquad i = 2, \dots, n,$$

ripetiamo i passi fatti precedentemente:

$$a_{22}^{(2)} x_2 + \sum_{j=3}^{n} a_{2j}^{(2)} x_j = b_2^{(2)} \tag{3.15}$$

$$a_{i2}^{(2)} x_2 + \sum_{j=3}^{n} a_{ij}^{(2)} x_j = b_i^{(2)}, \qquad i = 3, \dots, n. \tag{3.16}$$

Moltiplicando l'equazione (3.15) per $\dfrac{a_{i2}^{(2)}}{a_{22}^{(2)}}$, per $i = 3, \dots, n$, otteniamo

$$a_{i2}^{(2)} x_2 + \sum_{j=3}^{n} a_{2j}^{(2)} \frac{a_{i2}^{(2)}}{a_{22}^{(2)}} x_j = \frac{a_{i2}^{(2)}}{a_{22}^{(2)}} b_2^{(2)}, \qquad i = 3, \dots, n. \tag{3.17}$$

Sottraendo a questo punto le equazioni (3.17) dalle (3.16) si ottiene:

$$\sum_{j=3}^{n} \left(a_{ij}^{(2)} - \frac{a_{i2}^{(2)}}{a_{22}^{(2)}} a_{2j}^{(2)} \right) x_j = b_i^{(2)} - \frac{a_{i2}^{(2)}}{a_{22}^{(2)}} b_2^{(2)}, \qquad i = 3, \dots, n \tag{3.18}$$

ovvero scritta in forma più compatta:

$$\sum_{j=3}^{n} a_{ij}^{(3)} x_j = b_i^{(3)} \qquad i = 3, \dots, n$$

dove

$$\begin{cases} a_{ij}^{(3)} = a_{ij}^{(2)} - \dfrac{a_{i2}^{(2)}}{a_{22}^{(2)}} a_{2j}^{(2)} & i,j = 3,\dots,n \\ \\ b_i^{(3)} = b_i^{(2)} - \dfrac{a_{i2}^{(2)}}{a_{22}^{(2)}} b_2^{(2)} & i = 3,\dots,n. \end{cases}$$

Abbiamo il nuovo sistema equivalente:

$$\begin{cases} \displaystyle\sum_{j=1}^{n} a_{1j}^{(1)} x_j = b_1^{(1)} \\ \\ \displaystyle\sum_{j=2}^{n} a_{2j}^{(2)} x_j = b_2^{(2)} \\ \\ \displaystyle\sum_{j=3}^{n} a_{ij}^{(3)} x_j = b_i^{(3)} \qquad i = 3,\dots,n. \end{cases}$$

Osserviamo che in questo caso la matrice dei coefficienti è

$$A^{(3)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \dots & a_{2n}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \dots & a_{2n}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & a_{n3}^{(3)} & \dots & a_{nn}^{(3)} \end{bmatrix}.$$

È evidente ora che dopo $n-1$ passi di questo tipo arriveremo ad un sistema equivalente a quello di partenza avente la forma:

$$\begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \dots & \dots & a_{2n}^{(2)} \\ 0 & 0 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & a_{n-1\ n-1}^{(n-1)} & a_{n-1\ n}^{(n-1)} \\ 0 & 0 & \dots & 0 & a_{nn}^{(n)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_{n-1}^{(n-1)} \\ b_n^{(n)} \end{bmatrix}$$

la cui soluzione, come abbiamo visto, si ottiene facilmente, e dove le formule di trasformazione al passo k sono:

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)} \qquad i, j = k+1, \ldots, n \tag{3.19}$$

e

$$b_{i}^{(k+1)} = b_{i}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} b_{k}^{(k)} \qquad i = k+1, \ldots, n. \tag{3.20}$$

Soffermiamoci ora un momento sul primo passo del procedimento. Osserviamo che per ottenere il 1^0 sistema equivalente abbiamo operato le seguenti fasi:

1. moltiplicazione della prima riga della matrice dei coefficienti (e del corrispondente elemento del termine noto) per un opportuno scalare;

2. sottrazione dalla riga i-esima di A della prima riga modificata dopo il passo 1.

Il valore di k varia da 1 (matrice dei coefficienti e vettori dei termini noti iniziali) fino a $n-1$, infatti la matrice $A^{(n)}$ avrà gli elementi sottodiagonali delle prime $n-1$ colonne uguali a zero.
Si può osservare che il metodo di eliminazione di Gauss ha successo se tutti gli elementi $a_{kk}^{(k)}$ sono diversi da zero, che sono detti elementi pivotali.
Un proprietà importante delle matrici $A^{(k)}$ è il fatto che le operazioni effettuate non alterano il determinante della matrice, quindi

$$\det A^{(k)} = \det A,$$

per ogni k. Poichè la matrice $A^{(n)}$ è triangolare superiore allora il suo determinante può essere calcolato esplicitamente

$$\det A^{(k)} = \prod_{k=1}^{n} a_{kk}^{(k)}.$$

Quello appena descritto è un modo, alternativo alla regola di Laplace per calcolare il determinante della matrice A.

Esempio 3.3.1 *Calcolare il determinante della matrice*

$$A = \begin{bmatrix} 3 & 3 & 5 & 0 \\ 3 & 2 & 6 & -1 \\ 0 & 2 & 0 & 4 \\ 1 & 3 & 0 & 4 \end{bmatrix}$$

utlizzando il metodo di eliminazione di Gauss.

Posto $A^{(1)} = A$, calcoliamo i tre moltiplicatori

$$l_{2,1} = -1, \qquad l_{3,1} = 0, \qquad l_{4,1} = -\frac{1}{3}.$$

Calcoliamo la seconda riga:

$[2^a$ `riga di` $A^{(1)}$ `+` $]$	3	2	6	-1	$+$
$[(-1)\times$ 1^a `riga di` $A^{(1)}]$	-3	-3	-5	0	$=$
$[2^a$ `riga di` $A^{(2)}]$	0	-1	1	-1	

La terza riga non cambia perchè il moltiplicatore è nullo, mentre la quarta riga è

$[4^a$ `riga di` $A^{(1)}$ `+` $]$	1	3	0	4	$+$
$[(-1/3)\times$ 1^a `riga di` $A^{(1)}]$	-1	-1	$-5/3$	0	$=$
$[4^a$ `riga di` $A^{(2)}]$	0	2	$-5/3$	4	

Abbiamo ottenuto la seguente matrice al passo 2:

$$A^{(2)} = \begin{bmatrix} 3 & 3 & 5 & 0 \\ 0 & -1 & 1 & -1 \\ 0 & 2 & 0 & 4 \\ 0 & 2 & -5/3 & 4 \end{bmatrix}.$$

Calcoliamo i due moltiplicatori

$$l_{3,2} = 2, \qquad l_{4,2} = 2.$$

Calcoliamo la terza riga:

$[3^a$ `riga di` $A^{(2)}$ `+` $]$	0	2	0	4	$+$
$[(2)\times$ 2^a `riga di` $A^{(2)}]$	0	-2	2	-2	$=$
$[3^a$ `riga di` $A^{(3)}]$	0	0	2	2	

La quarta riga è

$$\begin{array}{lrrrrl} [4^a \texttt{ riga di } A^{(2)} \texttt{ + }] & 0 & 2 & -5/3 & 4 & + \\ [(2)\times\, 2^a \texttt{ riga di } A^{(2)}] & 0 & -2 & 2 & -2 & = \\ \hline [4^a \texttt{ riga di } A^{(3)}] & 0 & 0 & 1/3 & 2 & \end{array}$$

Abbiamo ottenuto la seguente matrice al passo 3:

$$A^{(3)} = \begin{bmatrix} 3 & 3 & 5 & 0 \\ 0 & -1 & 1 & -1 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 1/3 & 2 \end{bmatrix}.$$

Calcoliamo l'unico moltiplicatore del terzo passo:

$$l_{4,3} = -\frac{1}{6}.$$

La quarta riga è

$$\begin{array}{lrrrrl} [4^a \texttt{ riga di } A^{(3)} \texttt{ + }] & 0 & 0 & 1/3 & 2 & + \\ [(-1/6)\times\, 3^a \texttt{ riga di } A^{(3)}] & 0 & 0 & -1/3 & -1/3 & = \\ \hline [4^a \texttt{ riga di } A^{(4)}] & 0 & 0 & 0 & 5/3 & \end{array}$$

La matrice triagolarizzata è

$$A^{(4)} = \begin{bmatrix} 3 & 3 & 5 & 0 \\ 0 & -1 & 1 & -1 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 5/3 \end{bmatrix}.$$

Il determinante della matrice è uguale al prodotto degli elementi diagonali della matrice triangolare, ovvero

$$\det A = -10.$$

Esempio 3.3.2 *Calcolare l'inversa della matrice*

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

utlizzando il metodo di eliminazione di Gauss.

L'inversa di A è la matrice X tale che

$$AX = I$$

ovvero, detta $\mathbf{x}_i$ la $i-$esima colonna di X, questo è soluzione del sistema lineare

$$A\mathbf{x}_i = \mathbf{e}_i \tag{3.21}$$

dove $\mathbf{e}_i$ è l'$i-$esimo versore della base canonica di $\mathbb{R}^n$. Posto $i = 1$ risolvendo il sistema

$$A\mathbf{x}_1 = \mathbf{e}_1, \qquad \begin{bmatrix} 2 & 1 & 0 & 1 \\ 1 & 1 & 2 & 0 \\ -1 & 0 & 3 & 1 \\ 1 & 1 & 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

si ottengono gli elementi della prima colonna di A^{-1}. Posto $A^{(1)} = A$ gli elementi della matrice al passo 2 sono calcolati applicando le formule

$$a_{ij}^{(2)} = a_{ij}^{(1)} - \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} a_{1j}^{(1)}, \qquad i, j = 2, 3, 4.$$

Tralasciando il dettaglio delle operazioni risulta

$$A^{(2)} = \begin{bmatrix} 2 & 1 & 0 & 1 \\ 0 & 1/2 & 2 & -1/2 \\ 0 & 1/2 & 3 & 3/2 \\ 0 & 1/2 & 2 & 3/2 \end{bmatrix}, \qquad \mathbf{e}_1^{(2)} = \begin{bmatrix} 1 \\ -1/2 \\ 1/2 \\ -1/2 \end{bmatrix}$$

Applicando le formula

$$a_{ij}^{(3)} = a_{ij}^{(2)} - \frac{a_{i2}^{(2)}}{a_{22}^{(2)}} a_{2j}^{(2)}, \qquad i, j = 3, 4.$$

si ottiene il sistema al terzo passo

$$A^{(3)} = \begin{bmatrix} 2 & 1 & 0 & 1 \\ 0 & 1/2 & 2 & -1/2 \\ 0 & 1/2 & 1 & 2 \\ 0 & 0 & 0 & 2 \end{bmatrix}, \qquad \mathbf{e}_1^{(3)} = \begin{bmatrix} 1 \\ -1/2 \\ 1 \\ 0 \end{bmatrix}.$$

In questo caso non è necessario applicare l'ultimo passo del metodo in quanto la matrice è già triangolare superiore e pertanto si può risolvere il sistema triangolare superiore ottenendo:

$$x_4 = 0, \quad x_3 = 1, \quad x_2 = -5, \quad x_1 = 3.$$

Cambiando i termini noti del sistema (3.21), ponendo $i = 2, 3, 4$ si ottengono le altre tre colonne della matrice inversa.

3.3.1 Costo Computazionale del Metodo di Eliminazione di Gauss

Cerchiamo ora di determinare il costo computazionale (cioè il numero di operazioni aritmetiche) richiesto dal metodo di eliminazione di Gauss per risolvere un sistema lineare di ordine n. Dalle relazioni

$$b_i^{(k+1)} = b_i^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} b_k^{(k)}, \qquad i = k+1, \ldots, n,$$

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)}, \qquad i, j = k+1, \ldots, n$$

è evidente che servono 3 operazioni aritmetiche per calcolare $b_i^{(k+1)}$ (noto $b_i^{(k)}$) mentre sono necessarie che solo 2 operazioni per calcolare $a_{ij}^{(k+1)}$ (noto $a_{ij}^{(k)}$), infatti il moltiplicatore viene calcolato solo una volta. Il numero di elementi del vettore dei termini noti che vengono modificati è pari ad $n - k$ mentre gli elementi della matrice cambiati sono $(n-k)^2$ quindi complessivamente il numero di operazioni per calcolare gli elementi al passo $k+1$ è:

$$2(n-k)^2 + 3(n-k)$$

Pertanto per trasformare A in $A^{(n)}$ e $\mathbf{b}$ in $\mathbf{b}^{(n)}$ è necessario un numero di operazioni pari ala somma, rispetto a k, di tale valore

$$f(n) = 2\sum_{k=1}^{n-1}(n-k)^2 + 3\sum_{k=1}^{n-1}(n-k).$$

Sapendo che

$$\sum_{k=1}^{n} n^2 = \frac{n(n+1)(2n+1)}{6}$$

ed effettuando un opportuno cambio di indice nelle sommatorie risulta

$$f(n) = 2\left[\frac{n(n-1)(2n-1)}{6}\right] + 3\frac{n(n-1)}{2} = \frac{2}{3}n^3 + \frac{n^2}{2} - \frac{7}{6}n.$$

A questo valore bisogna aggiungere le n^2 operazioni aritmetiche necessarie per risolvere il sistema triangolare superiore ottenendo

$$\frac{2}{3}n^3 + \frac{3}{2}n^2 - \frac{7}{6}n$$

che è un valore molto inferiore rispetto alle $n!$ operazioni richieste dalla regola di Cramer, applicata insieme alla regola di Laplace.

3.3.2 Strategie di Pivoting per il metodo di Gauss

Nell'eseguire il metodo di Gauss si è fatta l'implicita ipotesi (vedi formule (3.19) e (3.20)) che gli elementi pivotali $a_{kk}^{(k)}$ siano non nulli per ogni k. In vero questa non è un'ipotesi limitante in quanto la non singolarità di A permette, con un opportuno scambio di righe in $A^{(k)}$, di ricondursi a questo caso. Infatti scambiare due righe in $A^{(k)}$ significa sostanzialmente scambiare due equazioni nel sistema $A^{(k)}\mathbf{x} = \mathbf{b}^{(k)}$ e ciò non altera la natura del sistema stesso.
Consideriamo la matrice $A^{(k)}$ e supponiamo $a_{kk}^{(k)} = 0$. In questo caso possiamo scegliere un elemento sottodiagonale appartenente alla $k-$esima colonna diverso da zero, supponiamo $a_{ik}^{(k)}$, scambiare le equazioni di indice i e k e continuare il procedimento perchè in questo modo l'elemento pivotale è diverso da zero. In ipotesi di non singolarità della matrice A possiamo dimostrare tale elemento diverso da zero esiste sicuramente. Infatti supponendo che, oltra all'elemento pivotale, siano nulli tutti gli $a_{ik}^{(k)}$ per $i = k+1, \ldots, n$, allora $A^{(k)}$ ha la seguente struttura:

$$A^{(k)} = \begin{bmatrix} a_{11}^{(1)} & \ldots & a_{1,k-1}^{(1)} & a_{1k}^{(1)} & a_{1,k+1}^{(1)} & \ldots & a_{1n}^{(1)} \\ & \ddots & \vdots & \vdots & \vdots & & \vdots \\ & & a_{k-1,k-1}^{(k-1)} & a_{k-1,k}^{(k-1)} & a_{k-1,k+1}^{(k-1)} & \ldots & a_{k-1,n}^{(k-1)} \\ & & & 0 & a_{k,k+1}^{(k)} & & a_{kn}^{(k)} \\ & 0 & & \vdots & \vdots & & \vdots \\ & & & 0 & a_{n,k+1}^{(k)} & \ldots & a_{nn}^{(k)}. \end{bmatrix}$$

Se partizioniamo $A^{(k)}$ nel seguente modo

$$A^{(k)} = \begin{bmatrix} A_{11}^{(k)} & A_{12}^{(k)} \\ 0 & A_{22}^{(k)} \end{bmatrix}$$

con $A_{11}^{(k)} \in \mathbb{R}^{(k-1)\times(k-1)}$ allora il determinante di $A^{(k)}$ è

$$\det A^{(k)} = \det A_{11}^{(k)} \det A_{22}^{(k)} = 0$$

perchè la matrice $A_{22}^{(k)}$ ha una colonna nulla. Poichè tutte la matrici $A^{(k)}$ hanno lo stesso determinante di A, dovrebbe essere $\det A = 0$ e questo contrasta con l'ipotesi fatta. Possiamo concludere che se $a_{kk}^{(k)} = 0$ e $\det A \neq 0$ deve necessariamente esistere un elemento $a_{ik}^{(k)} \neq 0$, con $i \in \{k+1, k+2, \ldots, n\}$. Per evitare che un elemento pivotale possa essere uguale a zero si applica una delle cosiddette strategie di pivoting. La strategia di Pivoting parziale prevede che prima di fare ciò si ricerchi l'elemento di massimo modulo tra gli elementi $a_{kk}^{(k)}$, $a_{k+1,k}^{(k)}$, $\ldots$, $a_{nk}^{(k)}$ e si scambi la riga in cui si trova questo elemento con la k-esima qualora esso sia diverso da $a_{kk}^{(k)}$. In altri termini il pivoting parziale richiede le seguenti operazioni:

1. determinare l'elemento $a_{rk}^{(k)}$ tale che

$$|a_{rk}^{(k)}| = \max_{k\leq i\leq n} |a_{ik}^{(k)}|;$$

2. effettuare lo scambio tra la r-esima e la k-esima riga.

In alternativa si può adottare la strategia di Pivoting totale che è la seguente:

1. determinare gli indici r, s tali che

$$|a_{rs}^{(k)}| = \max_{k\leq i,j\leq n} |a_{ij}^{(k)}|;$$

2. effettuare lo scambio tra la r-esima e la k-esima riga e tra la s-esima e la k-esima colonna.

La strategia di pivoting totale è senz'altro migliore perchè garantisce maggiormente che un elemento pivotale non sia un numero piccolo (in questa eventualità potrebbe accadere che un moltiplicatore sia un numero molto

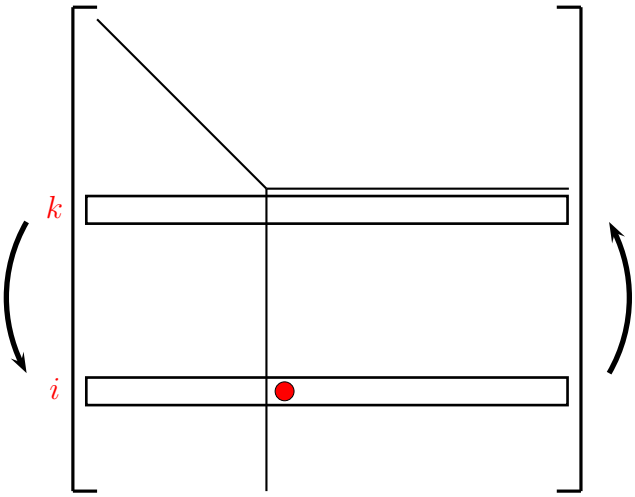


Figura 3.1: Strategia di pivoting parziale.

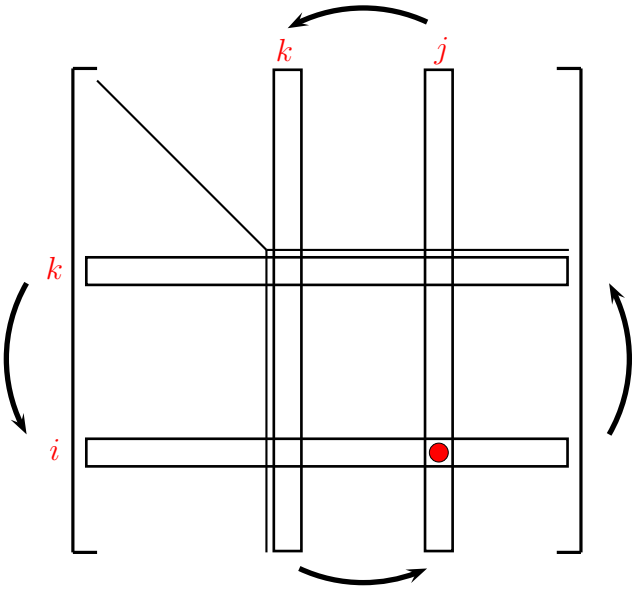


Figura 3.2: Strategia di pivoting totale.

grande) ma richiede che tutti gli eventuali scambi tra le colonne della matrice siano memorizzati. Infatti scambiare due colonne significa scambiare due incognite del vettore soluzione pertanto dopo la risoluzione del sistema triangolare per ottenere il vettore soluzione del sistema di partenza è opportuno permutare le componenti che sono state scambiate.

Esempio 3.3.3 *Risolvere il sistema lineare* $A\mathbf{x} = \mathbf{b}$ *dove*

$$A = \begin{bmatrix} 1 & 2 & -1 & 0 \\ 2 & -1 & -1 & 1 \\ 3 & 0 & -1 & 1 \\ 1 & -3 & 1 & 1 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 2 \\ 1 \\ 4 \\ 2 \end{bmatrix}$$

utlizzando il metodo di eliminazione di Gauss con strategia di pivoting parziale.

Posto $A^{(1)} = A$, osserviamo che l'elemento pivotale della prima colonna si trova sulla terza riga allora scambiamo per equazioni 1 e 3:

$$A^{(1)} = \begin{bmatrix} 3 & 0 & -1 & 1 \\ 2 & -1 & -1 & 1 \\ 1 & 2 & -1 & 0 \\ 1 & -3 & 1 & 1 \end{bmatrix}, \qquad \mathbf{b}^{(1)} = \begin{bmatrix} 4 \\ 1 \\ 2 \\ 2 \end{bmatrix}$$

calcoliamo i tre moltiplicatori

$$l_{2,1} = -\frac{2}{3}, \qquad l_{3,1} = -\frac{1}{3}, \qquad l_{4,1} = -\frac{1}{3}.$$

Calcoliamo la seconda riga:

[2^a `riga di` $A^{(1)}$ `+`]	2	-1	-1	1	1	+
[$(-2/3)\times$ 1^a `riga di` $A^{(1)}$]	-2	0	$2/3$	$-2/3$	$-8/3$	=
[2^a `riga di` $A^{(2)}$]	0	-1	$-1/3$	$1/3$	$-5/3$	

La terza riga è la seguente:

[3^a `riga di` $A^{(1)}$ `+`]	1	2	-1	0	2	+
[$(-1/3)\times$ 1^a `riga di` $A^{(1)}$]	-1	0	$1/3$	$-1/3$	$-4/3$	=
[3^a `riga di` $A^{(2)}$]	0	2	$-2/3$	$-1/3$	$2/3$	

mentre la quarta riga è

$[4^a$ `riga di` $A^{(1)}$ `+` $]$	1	-3	1	1	2	$+$
$[(-1/3)\times 1^a$ `riga di` $A^{(1)}]$	-1	0	$1/3$	$-1/3$	$-4/3$	$=$
$[4^a$ `riga di` $A^{(2)}]$	0	-3	$4/3$	$2/3$	$2/3$	

Abbiamo ottenuto la matrice ed il vettore al passo 2:

$$A^{(2)} = \begin{bmatrix} 3 & 0 & -1 & 1 \\ 0 & -1 & -1/3 & 1/3 \\ 0 & 2 & -2/3 & -1/3 \\ 0 & -3 & 4/3 & 2/3 \end{bmatrix}, \qquad \mathbf{b}^{(2)} = \begin{bmatrix} 4 \\ -5/3 \\ 2/3 \\ 2/3 \end{bmatrix}.$$

L'elemento pivotale della seconda colonna si trova sulla quarta riga quindi scambiamo le equazioni 2 e 4:

$$A^{(2)} = \begin{bmatrix} 3 & 0 & -1 & 1 \\ 0 & -3 & 4/3 & 2/3 \\ 0 & 2 & -2/3 & -1/3 \\ 0 & -1 & -1/3 & 1/3 \end{bmatrix}, \qquad \mathbf{b}^{(2)} = \begin{bmatrix} 4 \\ 2/3 \\ 2/3 \\ -5/3 \end{bmatrix}.$$

Calcoliamo i due moltiplicatori

$$l_{3,2} = \frac{2}{3}, \qquad l_{4,2} = -\frac{1}{3}.$$

La terza riga è la seguente:

$[3^a$ `riga di` $A^{(2)}$ `+` $]$	0	2	$-2/3$	$-1/3$	$2/3$	$+$
$[(2/3)\times 2^a$ `riga di` $A^{(2)}]$	0	-2	$8/9$	$4/9$	$4/9$	$=$
$[3^a$ `riga di` $A^{(3)}]$	0	0	$2/9$	$1/9$	$10/9$	

mentre la quarta riga è

$[4^a$ `riga di` $A^{(2)}$ `+` $]$	0	-1	$-1/3$	$1/3$	$-5/3$	$+$
$[(-1/3)\times 2^a$ `riga di` $A^{(2)}]$	0	1	$-4/9$	$-2/9$	$-2/9$	$=$
$[4^a$ `riga di` $A^{(3)}]$	0	0	$-7/9$	$1/9$	$-17/9$	

Abbiamo ottenuto la matrice ed il vettore al passo 3:

$$A^{(3)} = \begin{bmatrix} 3 & 0 & -1 & 1 \\ 0 & -3 & 4/3 & 2/3 \\ 0 & 0 & 2/9 & 1/9 \\ 0 & 0 & -7/9 & 1/9 \end{bmatrix}, \qquad \mathbf{b}^{(3)} = \begin{bmatrix} 4 \\ 2/3 \\ 10/9 \\ -17/9 \end{bmatrix}.$$

L'elemento pivotale della terza colonna si trova sulla quarta riga quindi scambiamo le equazioni 3 e 4:

$$A^{(3)} = \begin{bmatrix} 3 & 0 & -1 & 1 \\ 0 & -3 & 4/3 & 2/3 \\ 0 & 0 & -7/9 & 1/9 \\ 0 & 0 & 2/9 & 1/9 \end{bmatrix}, \qquad \mathbf{b}^{(3)} = \begin{bmatrix} 4 \\ 2/3 \\ -17/9 \\ 10/9 \end{bmatrix}.$$

Calcoliamo l'unico moltiplicatore del terzo passo:

$$l_{4,3} = \frac{2}{7}.$$

La quarta riga è

[4^a `riga di` $A^{(3)}$ `+`]	0	0	2/9	1/9	10/9	+
[(2/7)$\times$ 3^a `riga di` $A^{(3)}$]	0	0	−2/9	2/63	−34/63	=
[4^a `riga di` $A^{(4)}$]	0	0	0	1/7	4/7	

Il sistema triagolare superiore equivalente a quello iniziale ha come matrice dei coefficienti e come termine noto:

$$A^{(3)} = \begin{bmatrix} 3 & 0 & -1 & 1 \\ 0 & -3 & 4/3 & 2/3 \\ 0 & 0 & -7/9 & 1/9 \\ 0 & 0 & 0 & 1/7 \end{bmatrix}, \qquad \mathbf{b}^{(3)} = \begin{bmatrix} 4 \\ 2/3 \\ -17/9 \\ 4/7 \end{bmatrix}.$$

Risolvendo tale sistema triangolare superiore si ricava il vettore:

$$x_4 = 4, \qquad x_3 = 3, \qquad x_2 = 2, \qquad x_1 = 1.$$

Nelle pagine seguenti sono riportati i codici MatLab che implementano il metodo di Gauss con entrambe le strategie di pivoting descritte.

```
function x=Gauss(A,b)
%
% Metodo di eliminazione di Gauss
%
% Parametri di input:
% A = Matrice dei coefficienti del sistema
% b = Vettore dei termini noti del sistema
```

```
%
% Parametri di input:
% x = Vettore soluzione del sistema lineare
%
n = length(b);
x = zeros(n,1);
for k=1:n-1
  if abs(A(k,k))<eps
    error('Elemento pivotale nullo ')
  end
  for i=k+1:n
    A(i,k) = A(i,k)/A(k,k);
    b(i) = b(i)-A(i,k)*b(k);
    for j=k+1:n
      A(i,j) = A(i,j)-A(i,k)*A(k,j);
    end
 end
end
x(n) = b(n)/A(n,n);
for i=n-1:-1:1
  x(i) = (b(i)-A(i,i+1:n)*x(i+1:n))/A(i,i);
end
return

function x=Gauss_pp(A,b)
%
% Metodo di Gauss con pivot parziale
%
% Parametri di input:
% A = Matrice dei coefficienti del sistema
% b = Vettore dei termini noti del sistema
%
% Parametri di input:
% x = Vettore soluzione del sistema lineare
%
n = length(b);
x = zeros(n,1);
for k=1:n-1
```

```
  [a,i] = max(abs(A(k:n,k)));
  i = i+k-1;
  if i~=k
    A([i k],:) = A([k i],:);
    b([i k]) = b([k i]);
  end
  for i=k+1:n
    A(i,k) = A(i,k)/A(k,k);
    b(i) = b(i)-A(i,k)*b(k);
    for j=k+1:n
      A(i,j) = A(i,j)-A(i,k)*A(k,j);
    end
 end
end
x(n) = b(n)/A(n,n);
for i=n-1:-1:1
  x(i) = (b(i)-A(i,i+1:n)*x(i+1:n))/A(i,i);
end
return

function x=Gauss_pt(A,b)
%
% Metodo di Gauss con pivot totale
%
% Parametri di input:
% A = Matrice dei coefficienti del sistema
% b = Vettore dei termini noti del sistema
%
% Parametri di input:
% x = Vettore soluzione del sistema lineare
%
n = length(b);
x = zeros(n,1);
x1 = x;
indice = [1:n];
for k=1:n-1
  [a,riga] = max(abs(A(k:n,k:n)));
  [mass,col] = max(a);
```

```
  j = col+k-1;
  i = riga(col)+k-1;
  if i~=k
  A([i k],:) = A([k i],:);
  b([i k]) = b([k i]);
  end
  if j~=k
  A(:,[j k]) = A(:,[k j]);
    indice([j k]) = indice([k j]);
  end
  for i=k+1:n
    A(i,k) = A(i,k)/A(k,k);
    b(i) = b(i)-A(i,k)*b(k);
    for j=k+1:n
      A(i,j) = A(i,j)-A(i,k)*A(k,j);
    end
  end
end
%
% Risoluzione del sistema triangolare superiore
%
x1(n) = b(n)/A(n,n);
for i=n-1:-1:1
  x1(i) = (b(i)-A(i,i+1:n)*x1(i+1:n))/A(i,i);
end
%
% Ripermutazione del vettore
%
for i=1:n
  x(indice(i))=x1(i);
end
return
```

3.3.3 La Fattorizzazione *LU*

Supponiamo di dover risolvere un problema che richieda, ad un determinato passo, la risoluzione del sistema lineare $A\mathbf{x} = \mathbf{b}$ e di utilizzare il metodo di Gauss. La matrice viene resa triangolare superiore e viene risolto il sistema

triangolare

$$A^{(n)}\mathbf{x} = \mathbf{b}^{(n)}. \tag{3.22}$$

Ipotizziamo che, nell'ambito dello stesso problema, dopo un certo tempo sia necessario risolvere il sistema

$$A\mathbf{x} = \mathbf{c}$$

i cui la matrice dei coefficienti è la stessa mentre è cambiato il termine noto. Appare chiaro che non è possibile sfruttare i calcoli gia fatti in quanto il calcolo del vettore dei termini noti al passo n dipende dalle matrici ai passi precedenti all'ultimo, quindi la conoscenza della matrice $A^{(n)}$ è del tutto inutile. È necessario pertanto applicare nuovamente il metodo di Gauss e risolvere il sistema triangolare

$$A^{(n)}\mathbf{x} = \mathbf{c}^{(n)}. \tag{3.23}$$

L'algoritmo che sarà descritto in questo paragrafo consentirà di evitare l'eventualità di dover rifare tutti i calcoli (o una parte di questi). La Fattorizzazione *LU* di una matrice stabilisce, sotto determinate ipotesi, l'esistenza di una matrice L triangolare inferiore con elementi diagonali uguali a 1 e di una matrice triangolare superiore U tali che $A = LU$.
Vediamo ora di determinare le formule esplicite per gli elementi delle due matrici. Fissata la matrice A, quadrata di ordine n, imponiamo che risulti

$$A = LU.$$

Una volta note tali matrici il sistema di partenza $A\mathbf{x} = \mathbf{b}$ viene scritto come

$$LU\mathbf{x} = \mathbf{b}$$

e, posto $U\mathbf{x} = \mathbf{y}$, il vettore $\mathbf{x}$ viene trovato prima risolvendo il sistema triangolare inferiore

$$L\mathbf{y} = \mathbf{b}$$

e poi quello triangolare superiore

$$U\mathbf{x} = \mathbf{y}.$$

Imponiamo che la matrice A ammetta fattorizzazione LU:

$$\begin{bmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1n} \\ \vdots & & \vdots & & \vdots \\ a_{i1} & \dots & a_{ij} & \dots & a_{in} \\ \vdots & & \vdots & & \vdots \\ a_{n1} & \dots & a_{nj} & \dots & a_{nn} \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ l_{21} & 1 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & 0 & & \vdots \\ l_{i1} & \dots & l_{i,i-1} & 1 & \ddots & \vdots \\ \vdots & & \vdots & & \ddots & 0 \\ l_{n1} & \dots & l_{n,i-1} & l_{n,i} & \dots & 1 \end{bmatrix} \begin{bmatrix} u_{11} & \dots & \dots & u_{1j} & \dots & u_{1n} \\ 0 & u_{22} & \dots & u_{2j} & \dots & u_{2n} \\ \vdots & \ddots & \ddots & \vdots & & \vdots \\ \vdots & & \ddots & u_{jj} & \dots & u_{jn} \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & 0 & u_{nn} \end{bmatrix}.$$

Deve essere

$$a_{ij} = \sum_{k=1}^{n} l_{ik}u_{kj} = \sum_{k=1}^{\min(i,j)} l_{ik}u_{kj} \qquad i,j = 1,\dots,n. \tag{3.24}$$

Considerando prima il caso $i \leq j$, uguagliando la parte triangolare superiore delle matrici abbiamo

$$a_{ij} = \sum_{k=1}^{i} l_{ik}u_{kj} \qquad j \geq i \tag{3.25}$$

ovvero

$$a_{ij} = \sum_{k=1}^{i-1} l_{ik}u_{kj} + l_{ii}u_{ij} = \sum_{k=1}^{i-1} l_{ik}u_{kj} + u_{ij} \qquad j \geq i$$

infine risulta

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj} \qquad j \geq i \tag{3.26}$$

e ovviamente $u_{1j} = a_{1j}$, per $j = 1,\dots,n$. Considerando ora il caso $j < i$, uguagliando cioè le parti strettamente triangolari inferiori delle matrici risulta:

$$a_{ij} = \sum_{k=1}^{j} l_{ik}u_{kj} \qquad i > j \tag{3.27}$$

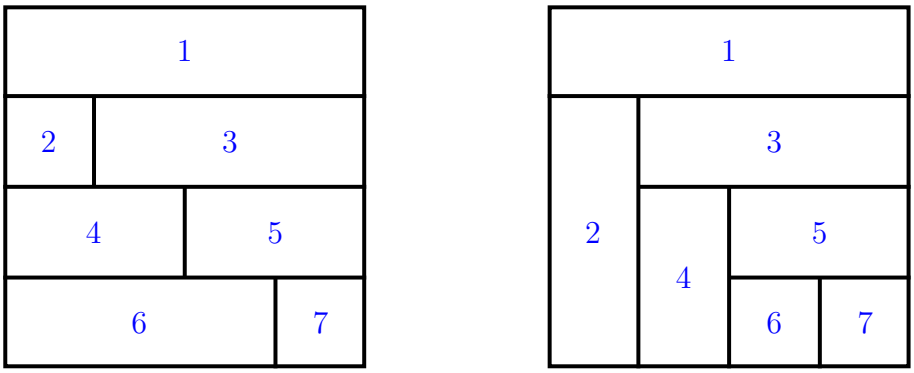
ovvero

$$a_{ij} = \sum_{k=1}^{j-1} l_{ik}u_{kj} + l_{ij}u_{jj} \qquad i > j$$

da cui

$$l_{ij} = \frac{1}{u_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj} \right) \qquad i > j. \tag{3.28}$$

Si osservi che le formule (3.26) e (3.28) vanno implementate secondo uno degli schemi riportati nella seguente figura.



Tecnica di Crout **Tecnica di Doolittle**

Ogni schema rappresenta in modo schematico una matrice la cui parte triangolare superiore indica la matrice U mentre quella triangolare inferiore la matrice L mentre i numeri indicano l'ordine con cui gli elementi saranno calcolati. Per esempio applicando la tecnica di Crout si segue il seguente ordine:

- 1^o Passo: Calcolo della prima riga di U;
- 2^o Passo: Calcolo della seconda riga di L;
- 3^o Passo: Calcolo della seconda riga di U;
- 4^o Passo: Calcolo della terza riga di L;
- 5^o Passo: Calcolo della terza riga di U;
- 6^o Passo: Calcolo della quarta riga di L;
- 7^o Passo: Calcolo della quarta riga di U;

e così via procedendo per righe in modo alternato. Nel caso della tecnica di Doolittle si seguono i seguenti passi:

- 1^o Passo: Calcolo della prima riga di U;

- 2^o Passo: Calcolo della prima colonna di L;
- 3^o Passo: Calcolo della seconda riga di U;
- 4^o Passo: Calcolo della seconda colonna di L;
- 5^o Passo: Calcolo della terza riga di U;
- 6^o Passo: Calcolo della terza colonna di L;
- 7^o Passo: Calcolo della quarta riga di U.

La fattorizzazione LU è un metodo sostanzialmente equivalente al metodo di Gauss, infatti la matrice U che viene calcolata coincide con la matrice $A^{(n)}$. Lo svantaggio del metodo di fattorizzazione diretto risiede essenzialmente nella maggiore difficoltà, rispetto al metodo di Gauss, di poter programmare una strategia di pivot. Infatti se un elemento diagonale della matrice U è uguale a zero non è possibile applicare l'algoritmo.

```
function [L,U]=crout(A);
%
% La funzione calcola la fattorizzazione LU della
% matrice A applicando la tecnica di Crout
%
% L = matrice triang. inferiore con elementi diagonali
%     uguali a 1
% U = matrice triangolare superiore
%
[m n] = size(A);
U = zeros(n);
L = eye(n);
U(1,:) = A(1,:);
for i=2:n
    for j=1:i-1
        L(i,j) = (A(i,j) - L(i,1:j-1)*U(1:j-1,j))/U(j,j);
    end
    for j=i:n
        U(i,j) = A(i,j) - L(i,1:i-1)*U(1:i-1,j);
    end
end
return
```

```
function [L,U]=doolittle(A);
%
% La funzione calcola la fattorizzazione LU della
% matrice A applicando la tecnica di Doolittle
%
% L = matrice triang. inferiore con elementi diagonali
%     uguali a 1
% U = matrice triangolare superiore
%
[m n] = size(A);
L = eye(n);
U = zeros(n);
U(1,:) = A(1,:);
for i=1:n-1
    for riga=i+1:n
        L(riga,i)=(A(riga,i)-L(riga,1:i-1)*U(1:i-1,i))/U(i,i);
    end
    for col=i+1:n
        U(i+1,col) = A(i+1,col)-L(i+1,1:i)*U(1:i,col);
    end
end
return
```

Esempio 3.3.4 *Applicare la tecnica di Doolittle per calcolare la fattorizzazione LU della matrice*

$$A = \begin{bmatrix} 1 & -1 & 3 & -4 \\ 2 & -3 & 9 & -9 \\ 3 & 1 & -1 & -10 \\ 1 & 2 & -4 & -1 \end{bmatrix}.$$

Gli elementi della prima riga di U vanno calcolati utilizzando la formula (3.26) con $i = 1$:

$$u_{1j} = a_{1j} - \sum_{k=1}^{0} l_{1k}u_{kj} = a_{1j}, \qquad j = 1, 2, 3, 4.$$

Quindi

$$U = \begin{bmatrix} 1 & -1 & 3 & -4 \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix}.$$

Gli elementi della prima colonna di L si ottengono applicando la formula (3.28) con $j = 1$:

$$l_{i1} = \frac{1}{u_{11}} \left(a_{i1} - \sum_{k=1}^{0} l_{ik} u_{k1} \right) = \frac{a_{i1}}{u_{11}}, \qquad i = 2, 3, 4,$$

da cui

$$l_{21} = \frac{a_{21}}{u_{11}} = 2; \quad l_{31} = \frac{a_{31}}{u_{11}} = 3; \quad l_{41} = \frac{a_{41}}{u_{11}} = 1.$$

La matrice L risulta essere, al momento, la seguente

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & l_{32} & 1 & 0 \\ 1 & l_{42} & l_{43} & 1 \end{bmatrix}.$$

Gli elementi della seconda riga di U vanno calcolati utilizzando la formula (3.26) con $i = 2$:

$$u_{2j} = a_{2j} - \sum_{k=1}^{1} l_{2k} u_{kj} = a_{2j} - l_{21} u_{1j}, \qquad j = 2, 3, 4,$$

quindi

$$\begin{aligned} u_{22} &= a_{22} - l_{21} u_{12} = -3 - 2 \cdot (-1) = -1; \\ u_{23} &= a_{23} - l_{21} u_{13} = 9 - 2 \cdot (3) = 3; \\ u_{24} &= a_{24} - l_{21} u_{14} = -9 - 2 \cdot (-4) = -1. \end{aligned}$$

$$U = \begin{bmatrix} 1 & -1 & 3 & -4 \\ 0 & -1 & 3 & -1 \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix}.$$

Gli elementi della seconda colonna di L si ottengono applicando la formula (3.28) con $j = 2$:

$$l_{i2} = \frac{1}{u_{22}} \left(a_{i2} - \sum_{k=1}^{1} l_{ik} u_{k2} \right) = \frac{a_{i2} - l_{i1} u_{12}}{u_{22}}, \qquad i = 3, 4,$$

e

$$l_{32} = \frac{a_{32} - l_{31}u_{12}}{u_{22}} = \frac{1 - 3 \cdot (-1)}{-1} = -4,$$

$$l_{42} = \frac{a_{42} - l_{41}u_{12}}{u_{22}} = \frac{2 - 1 \cdot (-1)}{-1} = -3.$$

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & -4 & 1 & 0 \\ 1 & -3 & l_{43} & 1 \end{bmatrix}.$$

Gli elementi della teza riga di U sono:

$$u_{3j} = a_{3j} - \sum_{k=1}^{2} l_{3k}u_{kj} = a_{3j} - l_{31}u_{1j} - l_{32}u_{2j}, \qquad j = 3, 4,$$

quindi

$$\begin{aligned} u_{33} &= a_{33} - l_{31}u_{13} - l_{32}u_{23} = -1 - 3 \cdot (3) - (-4) \cdot 3 = 2, \\ u_{34} &= a_{34} - l_{31}u_{14} - l_{32}u_{24} = -10 - 3 \cdot (-4) - (-4) \cdot (-1) = -2. \end{aligned}$$

Le matrici sono diventate

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & -4 & 1 & 0 \\ 1 & -3 & l_{43} & 1 \end{bmatrix}, \qquad U = \begin{bmatrix} 1 & -1 & 3 & -4 \\ 0 & -1 & 3 & -1 \\ 0 & 0 & 2 & -2 \\ 0 & 0 & 0 & u_{44} \end{bmatrix}.$$

L'unico elemento della terza colonna di L è:

$$l_{43} = \frac{1}{u_{33}} \left(a_{43} - \sum_{k=1}^{2} l_{4k}u_{k3} \right) =$$

ovvero

$$l_{43} = \frac{a_{43} - l_{41}u_{13} - l_{42}u_{23}}{u_{33}} = \frac{-4 - 1 \cdot 3 - (-3) \cdot 3}{2} = 1,$$

L'ultimo elemento da calcolare è:

$$\begin{aligned} u_{44} &= a_{44} - \sum_{k=1}^{3} l_{4k}u_{k4} \\ &= a_{44} - l_{41}u_{14} - l_{42}u_{24} - l_{43}u_{34} = -1 + 4 - 3 + 2 = 2. \end{aligned}$$

Le matrici L ed$U$ sono pertanto

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & -4 & 1 & 0 \\ 1 & -3 & 1 & 1 \end{bmatrix},$$

e

$$U = \begin{bmatrix} 1 & -1 & 3 & -4 \\ 0 & -1 & 3 & -1 \\ 0 & 0 & 2 & -2 \\ 0 & 0 & 0 & 2 \end{bmatrix}.$$

Esercizio 3.3.1 *Risolvere il problema descritto nell'esempio 3.3.2 calcolando la fattorizzazione LU della matrice A.*

3.4 Condizionamento di sistemi lineari

Nel Capitolo 1 è stato introdotto il concetto di rappresentazione in base ed è stata motivata la sostanziale inaffidabilità dei risultati dovuti ad elaborazioni numeriche, a causa dell'artimetica finita dell'elaboratore. Appare chiaro come la bassa precisione nel calcolo potrebbe fornire dei risultati numerici molto lontani da quelli reali. In alcuni casi tale proprietà è insita nel problema. Consideriamo il sistema lineare

$$A\mathbf{x} = \mathbf{b} \tag{3.29}$$

dove $A \in \mathbb{R}^{n\times n}$ è la cosiddetta matrice di Hilbert, i cui elementi sono

$$a_{ij} = \frac{1}{i+j-1}, \qquad i,j = 1,\ldots,n$$

mentre il vettore $\mathbf{b}$ è scelto in modo tale che il vettore soluzione abbia tutte componenti uguali a 1, cosicchè si possa conoscere con esattezza l'errore commesso nel suo calcolo. Risolvendo il sistema di ordine 20 con il metodo di Gauss senza pivoting si osserva che la soluzione è, in realtà, molto lontana da quella teorica. Questa situazione peggiora prendendo matrici di dimensioni crescenti ed è legata ad un fenomeno che viene detto malcondizionamento. Bisogna infatti ricordare che, a causa degli errori legati alla rappresentazione dei numeri reali, il sistema che l'elaboratore risolve non coincide con quello

teorico, poichè alla matrice A ed al vettore $\mathbf{b}$ e necessario aggiungere la matrice δA ed il vettore $\delta\mathbf{b}$ (che contengono le perturbazioni legate a tali errori), e che la soluzione ovviamente non è la stessa, pertanto la indichiamo con $\mathbf{x} + \delta\mathbf{x}$:

$$(A + \delta A)(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}. \tag{3.30}$$

Si può dimostrare che l'ordine di grandezza della perturbazione sulla soluzione è

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|A\|\|A^{-1}\| \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \right).$$

Il numero $K(A) = \|A\|\|A^{-1}\|$, detto indice di condizionamento del sistema, misura le amplificazioni degli errori sui dati del problema (ovvero la misura di quanto aumentano gli errori sulla soluzione). Il caso della matrice di Hilbert è appunto uno di quelli per cui l'indice di condizionamento assume valori molto grandi (di ordine esponenziale) all'aumentare della dimensione, si parla infatti di matrici malcondizionate. Quando ciò non accade si parla invece di matrici bencondizionate. A volte tale caratteristica può dipendere anche dalla scelta dell'algoritmo di risoluzione, ovvero vi sono algoritmi che forniscono risultati meno influenzati dal condizionamento dei dati (eseguendo il metodo di Gauss con pivoting parziale, per esempio, i risultati sono affetti comunque da errori, ma di meno rispetto al metodo di Gauss senza alcuna strategia di pivoting).

Capitolo 4

Interpolazione di dati e Funzioni

4.1 Introduzione

Nel campo del Calcolo Numerico si possono incontrare diversi casi nei quali è richiesta l'approssimazione di una funzione (o di una grandezza incognita):
1) non è nota l'espressione analitica della funzione $f(x)$ ma si conosce il valore che assume in un insieme finito di punti $x_1, x_2, \ldots, x_n$. Si potrebbe pensare anche che tali valori siano delle misure di una grandezza fisica incognita valutate in differenti istanti di tempo.
2) Si conosce l'espressione analitica della funzione $f(x)$ ma è così complicata dal punto di vista computazionale che è più conveniente cercare un'espressione semplice partendo dal valore che essa assume in un insieme finito di punti. In questo capitolo analizzeremo un particolare tipo di approssimazione di funzioni cioè la cosiddetta interpolazione che richiede che la funzione approssimante assume in determinate ascisse esattamente lo stesso valore di $f(x)$. In entrambi i casi appena citati è noto, date certe informazioni supplementari, che la funzione approssimante va ricercata della forma:

$$f(x) \simeq g(x; a_0, a_1, \ldots, a_n). \tag{4.1}$$

Se i parametri $a_0, a_1, \ldots, a_n$ sono definiti dalla condizione di coincidenza di f e g nei punti $x_0, x_1, \ldots, x_n$, allora tale procedimento di approssimazione si chiama appunto Interpolazione. Invece se $x \notin [\min_i x_i, \max_i x_i]$ allora si parla di *Estrapolazione*. Tra i procedimenti di interpolazione il più usato è

quello in cui si cerca la funzione g in (4.1) nella forma

$$g(x; a_0, a_1, \ldots, a_n) = \sum_{i=0}^{n} a_i \Phi_i(x)$$

dove $\Phi_i(x)$, per $i = 0, \ldots, n$, sono funzioni fissate e i valori di a_i, $i = 0, \ldots, n$, sono determinati in base alle condizioni di coincidenza di f con la funzione approssimante nei punti di interpolazione (detti anche nodi), x_j, cioè si pone

$$f(x_j) = \sum_{i=0}^{n} a_i \Phi_i(x_j) \qquad j = 0, \ldots, n. \tag{4.2}$$

Il processo di determinazione degli a_i attraverso la risoluzione del sistema (4.2) si chiama metodo dei coefficienti indeterminati. Il caso più studiato è quello dell'interpolazione polinomiale, in cui si pone:

$$\Phi_i(x) = x^i \qquad i = 0, \ldots, n$$

e perciò la funzione approssimante g assume la forma

$$\sum_{i=0}^{n} a_i x^i;$$

mentre le condizioni di coincidenza diventano

$$f(x_j) = \sum_{i=0}^{n} a_i x_j^i \qquad j = 0, \ldots, n. \tag{4.3}$$

Se i nodi x_j sono distinti allora la matrice dei coefficienti del sistema (4.3), detta matrice di Vandermonde, è non singolare e pertanto il problema dell'interpolazione ammette sempre un'unica soluzione. Descriviamo ora un modo alternativo di risolvere il problema di interpolazione in grado di fornire l'espressione esplicita del polinomio cercato.

4.2 Il Polinomio Interpolante di Lagrange

Al fine di dare una forma esplicita al polinomio interpolante, scriviamo il candidato polinomio nella seguente forma:

$$L_n(x) = \sum_{k=0}^{n} l_{nk}(x) f(x_k) \tag{4.4}$$

dove gli $l_{nk}(x)$ sono per il momento generici polinomi di grado n. Imponendo le condizioni di interpolazione

$$L_n(x_i) = f(x_i) \qquad i = 0, \ldots, n$$

deve essere, per ogni i:

$$L_n(x_i) = \sum_{k=0}^{n} l_{nk}(x_i) f(x_k) = f(x_i)$$

ed è evidente che se

$$l_{nk}(x_i) = \begin{cases} 0 & \text{se } k \neq i \\ \\ 1 & \text{se } k = i \end{cases} \tag{4.5}$$

allora esse sono soddisfatte. In particolare la prima condizione di (4.5) indica che $l_{nk}(x)$ si annulla negli n nodi $x_0, x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n$ pertanto deve avere la seguente struttura:

$$l_{nk}(x) = c_k \prod_{i=0, i\neq k}^{n} (x - x_i)$$

mentre impondendo la seconda condizione di (4.5)

$$l_{nk}(x_k) = c_k \prod_{i=0, i\neq k}^{n} (x_k - x_i) = 1$$

si trova immediatamente:

$$c_k = \frac{1}{\displaystyle\prod_{i=0, i\neq k}^{n} (x_k - x_i)}.$$

In definitiva il polinomio interpolante ha la seguente forma:

$$L_n(x) = \sum_{k=0}^{n} \left(\prod_{i=0, i\neq k}^{n} \frac{x - x_i}{x_k - x_i} \right) f(x_k). \tag{4.6}$$

Il polinomio (4.6) prende il nome di Polinomio di Lagrange mentre i polinomi:

$$l_{nk}(x) = \prod_{i=0, i\neq k}^{n} \frac{x - x_i}{x_k - x_i}; \qquad k = 0, 1, \ldots, n$$

si chiamano Polinomi Fondamentali di Lagrange.

4.2.1 Il Resto del Polinomio di Lagrange

Assumiamo che la funzione interpolata $f(x)$ sia di classe $\mathcal{C}^{n+1}([a,b])$ e valutiamo l'errore che si commette nel sostituire $f(x)$ con $L_n(x)$ in un punto $x \neq x_i$. Supponiamo che l'intervallo $[a,b]$ sia tale da contenere sia i nodi x_i che l'ulteriore punto x. Sia dunque

$$e(x) = f(x) - L_n(x)$$

l'errore (o resto) commesso nell'interpolazione della funzione $f(x)$. Poichè

$$e(x_i) = f(x_i) - L_n(x_i) = 0 \qquad i = 0, \ldots, n$$

è facile congetturare per $e(x)$ la seguente espressione:

$$e(x) = c(x)\omega_{n+1}(x)$$

dove

$$\omega_{n+1}(x) = \prod_{i=0}^{n}(x - x_i)$$

è il cosiddetto polinomio nodale mentre $c(x)$ è una funzione da determinare. Definiamo ora la funzione

$$\Phi(t;x) = f(t) - L_n(t) - c(x)\omega_{n+1}(t)$$

dove t è una variabile ed x è un valore fissato. Calcoliamo la funzione $\Phi(t;x)$ nei nodi x_i:

$$\Phi(x_i;x) = f(x_i) - L_n(x_i) - c(x)\omega_{n+1}(x_i) = 0$$

e anche nel punto x:

$$\Phi(x;x) = f(x) - L_n(x) - c(x)\omega_{n+1}(x) = e(x) - c(x)\omega_{n+1}(x) = 0$$

pertanto la funzione $\Phi(t;x)$ (che è derivabile con continuità $n+1$ volte poichè $f(x)$ è di classe $\mathcal{C}^{n+1}$) ammette almeno $n+2$ zeri distinti. Applicando il teorema di Rolle segue che $\Phi'(t;x)$ ammette almeno $n+1$ zeri distinti. Riapplicando lo stesso teorema segue che $\Phi''(t;x)$ ammette almeno n zeri distinti. Così proseguendo segue che

$$\exists \xi_x \in [a,b] \;\; \ni' \;\; \Phi^{(n+1)}(\xi_x;x) = 0.$$

Calcoliamo ora la derivata di ordine $n+1$ della funzione $\Phi(t;x)$, osservando innanzitutto che la derivata di tale ordine del polinomio $L_n(x)$ è identicamente nulla. Pertanto

$$\Phi^{(n+1)}(t;x) = f^{(n+1)}(t) - c(x)\frac{d^{n+1}}{dt^{n+1}}\omega_{n+1}(t).$$

Calcoliamo la derivata di ordine $n+1$ del polinomio nodale. Osserviamo innanzitutto che

$$\omega_{n+1}(t) = \prod_{i=0}^{n}(t - x_i) = t^{n+1} + p_n(t)$$

dove $p_n(t)$ è un polinomio di grado al più n. Quindi

$$\frac{d^{n+1}}{dt^{n+1}}\omega_{n+1}(t) = \frac{d^{n+1}}{dt^{n+1}}t^{n+1}.$$

Poichè

$$\frac{d}{dt}t^{n+1} = (n+1)t^n$$

e

$$\frac{d^2}{dt^2}t^{n+1} = (n+1)nt^{n-1}$$

è facile dedurre che

$$\frac{d^{n+1}}{dt^{n+1}}t^{n+1} = \frac{d^{n+1}}{dt^{n+1}}\omega_{n+1}(t) = (n+1)!.$$

Pertanto

$$\Phi^{(n+1)}(t;x) = f^{(n+1)}(t) - c(x)(n+1)!$$

e

$$\Phi^{(n+1)}(\xi_x;x) = f^{(n+1)}(\xi_x) - c(x)(n+1)! = 0$$

cioè

$$c(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}$$

e in definitiva

$$e(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\omega_{n+1}(x). \tag{4.7}$$

Esempio 4.2.1 *Supponiamo di voler calcolare il polinomio interpolante di Lagrange passante per i punti* $(-1,-1)$, $(0,1)$, $(1,-1)$, $(3,2)$ *e* $(5,6)$. *Il grado di tale polinomio è* 4, *quindi definiamo i nodi*

$$x_0 = -1, \quad x_1 = 0, \quad x_2 = 1, \quad x_3 = 3, \quad x_4 = 5,$$

cui corrispondono le ordinate che indichiamo con y_i, $i = 0, \ldots, 4$:

$$y_0 = -1, \quad y_1 = 1, \quad y_2 = -1, \quad y_3 = 2, \quad y_4 = 6.$$

Scriviamo ora l'espressione del polinomio $L_4(x)$:

$$L_4(x) = l_{4,0}(x)y_0 + l_{4,1}(x)y_1 + l_{4,2}(x)y_2 + l_{4,3}(x)y_3 + l_{4,4}(x)y_4 \tag{4.8}$$

e calcoliamo i 5 *polinomi fondamentali di Lagrange:*

$$\begin{aligned}
l_{4,0}(x) &= \frac{(x-0)(x-1)(x-3)(x-5)}{(-1-0)(-1-1)(-1-3)(-1-5)} = \\
&= \frac{1}{48}\, x(x-1)(x-3)(x-5) \\
l_{4,1}(x) &= \frac{(x+1)(x-1)(x-3)(x-5)}{(0+1)(0-1)(0-3)(0-5)} = \\
&= -\frac{1}{15}(x+1)(x-1)(x-3)(x-5) \\
l_{4,2}(x) &= \frac{(x+1)(x-0)(x-3)(x-5)}{(1+1)(1-0)(1-3)(1-5)} = \\
&= \frac{1}{16}x(x+1)(x-3)(x-5) \\
l_{4,3}(x) &= \frac{(x+1)(x-0)(x-1)(x-5)}{(3+1)(3-0)(3-1)(3-5)} = \\
&= -\frac{1}{48}x(x+1)(x-1)(x-5) \\
l_{4,4}(x) &= \frac{(x+1)(x-0)(x-1)(x-3)}{(5+1)(5-0)(5-1)(5-3)} = \\
&= \frac{1}{240}x(x+1)(x-1)(x-3)
\end{aligned}$$

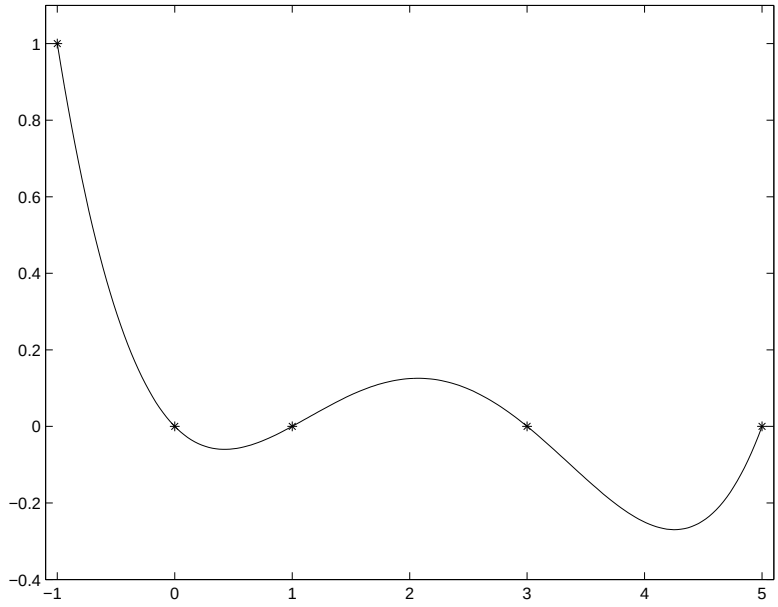


Figura 4.1: Grafico del polinomio $l_{40}(x)$.

Sostituendo in (4.8) il valore della funzione nei nodi si ottiene l'espressione finale del polinomio interpolante:

$$L_4(x) = -l_{4,0}(x) + l_{4,1}(x) - l_{4,2}(x) + 2l_{4,3}(x) + 6l_{4,4}(x).$$

Se vogliamo calcolare il valore approssimato della funzione $f(x)$ in un'ascissa diversa dai nodi, per esempio $x = 2$ allora dobbiamo calcolare il valore del polinomio interpolante $L_4(2)$.
Nelle figure 4.1-4.5 sono riportati i grafici dei cinque polinomi fondamentali di Lagrange: gli asterischi evidenziano il valore assunto da tali polinomi nei nodi di interpolazione. Nella figura 4.6 è tracciato il grafico del polinomio interpolante di Lagrange, i cerchi evidenziano ancora una volta i punti di interpolazione.

4.2.2 Il fenomeno di Runge

Nell'espressione dell'errore è presente, al denominatore, il fattore $(n + 1)!$, che potrebbe indurre a ritenere che, utilizzando un elevato numero di nodi, l'errore tenda a zero ed il polinomio interpolante converga alla funzione

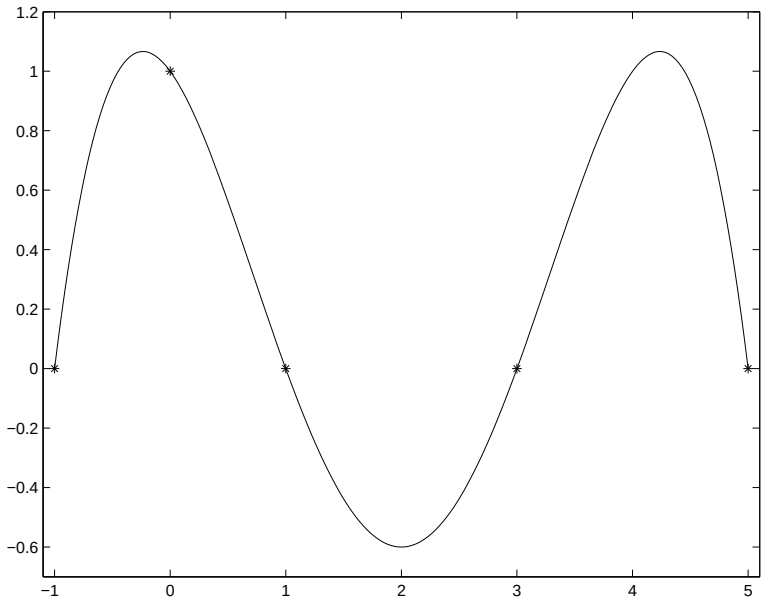


Figura 4.2: Grafico del polinomio $l_{41}(x)$.

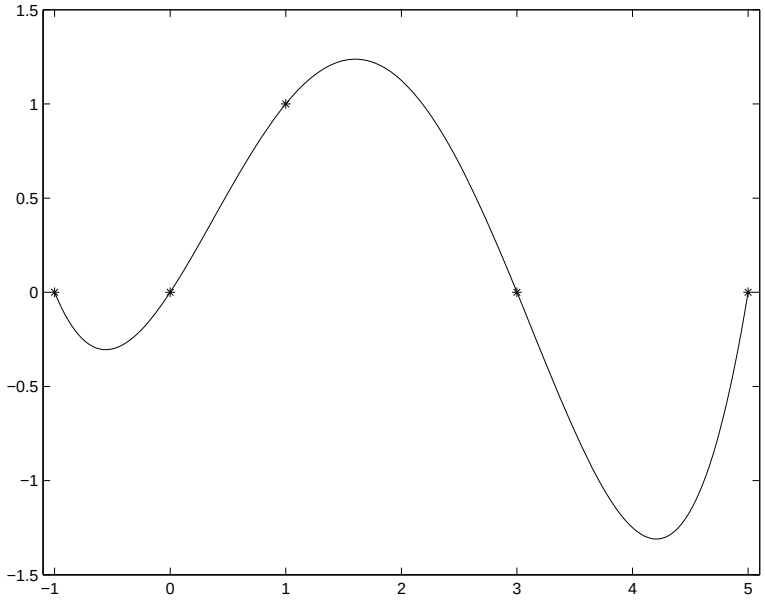


Figura 4.3: Grafico del polinomio $l_{42}(x)$.

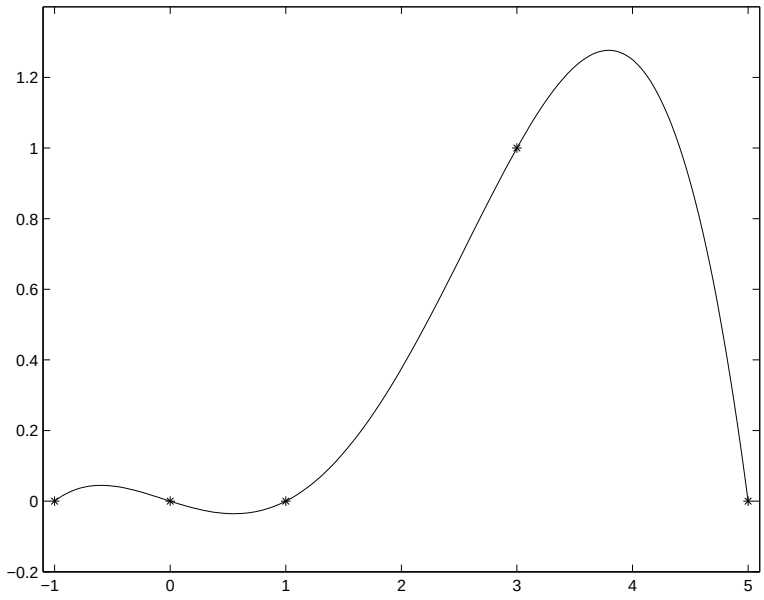


Figura 4.4: Grafico del polinomio $l_{43}(x)$.

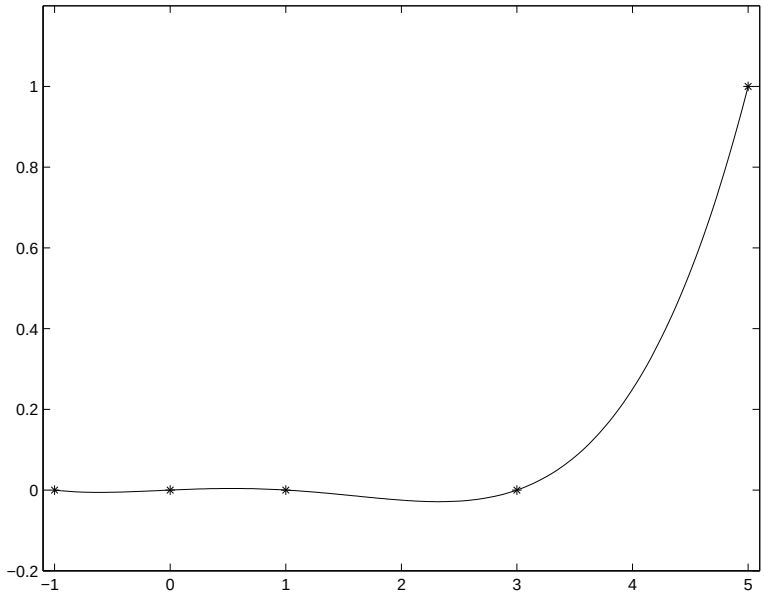


Figura 4.5: Grafico del polinomio $l_{44}(x)$.

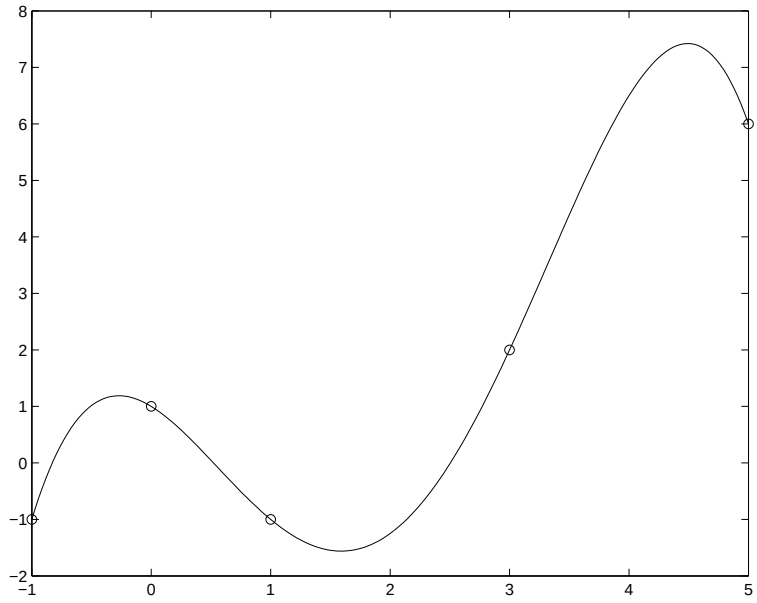


Figura 4.6: Grafico del polinomio interpolante di Lagrange $L_4(x)$.

$f(x)$. Questa ipotesi è confutata se si costruisce il polinomio che interpola la funzione

$$f(x) = \frac{1}{1+x^2}$$

nell'intervallo $[-5, 5]$ e prendendo 11 nodi equidistanti $-5, -4, -3, \ldots, 3, 4, 5$. Nella successiva figura viene appunto visualizzata la funzione (in blu) ed il relativo polinomio interpolante (in rosso).
Il polinomio interpolante presenta infatti notevoli oscillazioni, soprattutto verso gli estremi dell'intervallo di interpolazione, che diventano ancora più evidenti all'aumentare di n. Tale fenomeno, detto appunto fenomeno di Runge, è dovuto ad una serie di situazioni concomitanti:

1. il polinomio nodale, al crescere di n, assume un'andamento fortemente oscillante, soprattutto quando i nodi sono equidistanti;

2. alcune funzioni, come quella definita nell'esempio, hanno le derivate il cui valore tende a crescere con un ordine di grandezza talmente elevato da neutralizzare di fatto la presenza del fattoriale al denominatore dell'espressione dell'errore.

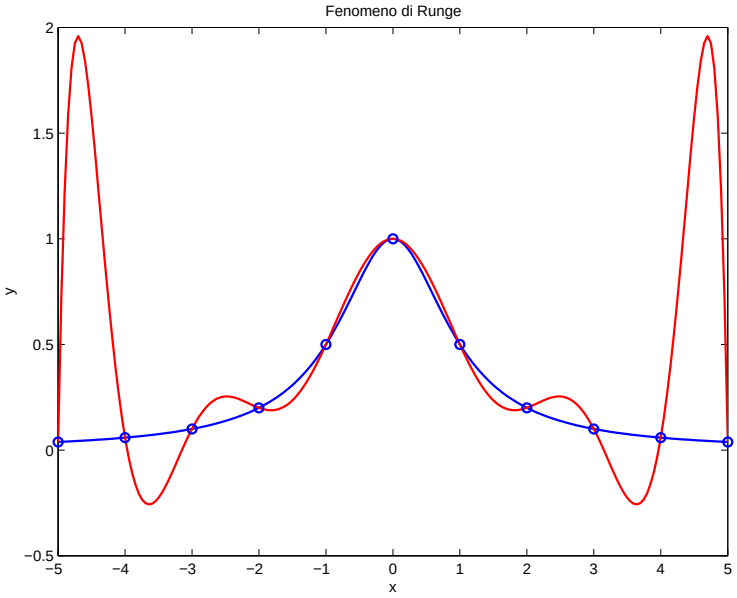


Figura 4.7: Il fenomeno di Runge.

Per ovviare al fenomeno di Runge si possono utilizzare insiemi di nodi non equidistanti oppure utilizzare funzioni interpolanti polinomiali a tratti (interpolando di fatto su intervalli più piccoli e imponendo le condizioni di continuità fino ad un ordine opportuno).

```
function  yy=lagrange(x,y,xx);
%
% La funzione calcola il polinomio interpolante di Lagrange
% in un vettore assegnato di ascisse
%
% Parametri di input
% x = vettore dei nodi
% y = vettore delle ordinate nei nodi
% xx = vettore delle ascisse in cui calcolare il polinomio
% Parametri di output
% yy = vettore delle ordinate del polinomio
%
n = length(x);
m = length(xx);
```

```
yy = zeros(size(xx));
for i=1:m
    yy(i)=0;
    for k=1:n
        yy(i)=yy(i)+prod((xx(i)-x([1:k-1,k+1:n]))./...
            (x(k)-x([1:k-1,k+1:n])))*y(k);
    end
end
return
```

4.3 Minimizzazione del Resto nel Problema di Interpolazione

Supponiamo che la funzione $f(x)$ sia approssimata su $[a,b]$ dal polinomio interpolante $L_n(x)$ e siano $x_0, x_1, \ldots, x_n$ i nodi di interpolazione. Come già sappiamo se $x \in [a,b]$ risulta

$$e(x) = f(x) - L_n(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\omega_{n+1}(x) \qquad \xi_x \in [a,b]$$

e dove

$$\omega_{n+1}(x) = \prod_{i=0}^{n}(x - x_i).$$

Si noti che variando i nodi x_i, $i = 0, \ldots, n$, cambia il polinomio $\omega_{n+1}(x)$ e di conseguenze cambia l'errore. Ha senso allora porsi il seguente problema: indicato con $\mathcal{P}_{n+1}$ l'insieme di tutti i polinomi di grado al più $n+1$ cerchiamo il polinomio $\tilde{p} \in \mathcal{P}_{n+1}$ tale che:

$$\max_{x\in[a,b]} |\tilde{p}(x)| = \min_{p\in\mathcal{P}_{n+1}} \max_{x\in[a,b]} |p(x)|. \tag{4.9}$$

Per dare una risposta a questo problema è essenziale introdurre i Polinomi di Chebyshev di 1^a Specie.

4.3.1 Polinomi di Chebyshev

I polinomi di Chebyshev $T_n(x)$, $n \geq 0$, sono così definiti:

$$T_n(x) = \cos(n \arccos x) \tag{4.10}$$

per $x \in [-1, 1]$. Per esempio:

$$\begin{aligned} T_0(x) &= \cos(0 \arccos x) = \cos 0 = 1 \\ T_1(x) &= \cos(1 \arccos x) = x \end{aligned}$$

e così via. È possibile ricavare una relazione di ricorrenza sui polinomi di Chebyshev che permette un più agevole calcolo. Infatti, posto

$$\arccos x = \theta \qquad (\text{ovvero } x = \cos\theta)$$

risulta

$$T_n(x) = \cos n\theta(x).$$

Considerando le relazioni

$$T_{n+1}(x) = \cos(n+1)\theta = \cos n\theta \ \cos\theta - \sin n\theta \ \sin\theta$$

$$T_{n-1}(x) = \cos(n-1)\theta = \cos n\theta \ \cos\theta + \sin n\theta \ \sin\theta$$

e sommandole membro a membro,

$$T_{n+1}(x) + T_{n-1}(x) = 2\cos\theta\cos n\theta = 2xT_n(x)$$

si ricava la seguente relazione di ricorrenza

$$T_{n+1}(x) = 2x \ T_n(x) - T_{n-1}(x), \qquad n \geq 1 \tag{4.11}$$

che, insieme all'espressione dei primi due polinomi,

$$T_0(x) = 1, \qquad T_1(x) = x.$$

consente di calcolare tutti i polinomi di Chebyshev.
L'espressione dei primi polinomi è la seguente

$$\begin{aligned} T_2(x) &= 2xT_1(x) - T_0(x) = 2x^2 - 1 \\ T_3(x) &= 2xT_2(x) - T_1(x) = 4x^3 - 3x \\ T_4(x) &= 2xT_3(x) - T_2(x) = 8x^4 - 8x^2 + 1 \\ T_5(x) &= 2xT_4(x) - T_3(x) = 16x^5 - 20x^3 + 5x \end{aligned}$$

Le seguenti proprietà dei polinomi di Chebyshev sono di facile dimostrazione:

1. $\max_{x\in[-1,1]} |T_n(x)| = 1$
2. $T_{2k}(x) = T_{2k}(-x)$
3. $T_{2k+1}(x) = -T_{2k+1}(-x)$
4. $T_n(x) = 2^{n-1}x^n + \ldots\ldots$
5. $T_n(x)$ assume complessivamente $n+1$ volte il valore $+1$ e -1 nei punti:

$$\begin{aligned} x_k &= \cos\frac{k\pi}{n} \qquad k = 0,\ldots,n; \\ T_n(x_k) &= (-1)^k \qquad k = 0,\ldots,n; \end{aligned}$$

6. $T_n(x)$ ha n zeri distinti nell'intervallo $]-1,1[$ dati da

$$t_k = \cos\frac{(2k+1)\pi}{2n} \qquad k = 0,\ldots,n-1.$$

Infatti è sufficiente porre

$$\cos n\theta = 0$$

da cui risulta

$$n\theta = \frac{\pi}{2} + k\pi = \frac{(2k+1)\pi}{2}, \qquad k = 0,\ldots,n-1.$$

Nella Figura 4.8 sono tracciati i grafici dei primi cinque polinomi di Chebyshev nell'intervallo $[-1,1]$. Ovviamente per calcolare il valore del polinomio $T_n(x)$ in un punto x fissato si usa la formula di ricorrenza (4.11), in quanto tale espressione è valida per ogni $x \in \mathbb{R}$.
Sia

$$\tilde{T}_n(x) = \frac{1}{2^{n-1}}T_n(x)$$

il polinomio di Chebyshev normalizzato in modo da risultare monico (ricordiamo che un polinomio di grado n è monico se il coefficiente del termine di grado massimo è 1). Vale allora la seguente proprietà di minimax.

Teorema 4.3.1 *(proprietà di minimax) Se $p_n(x)$ è un qualunque polinomio monico di grado n si ha:*

$$\frac{1}{2^{n-1}} = \max_{x\in[-1,1]} |\tilde{T}_n(x)| \le \max_{x\in[-1,1]} |p_n(x)|.$$

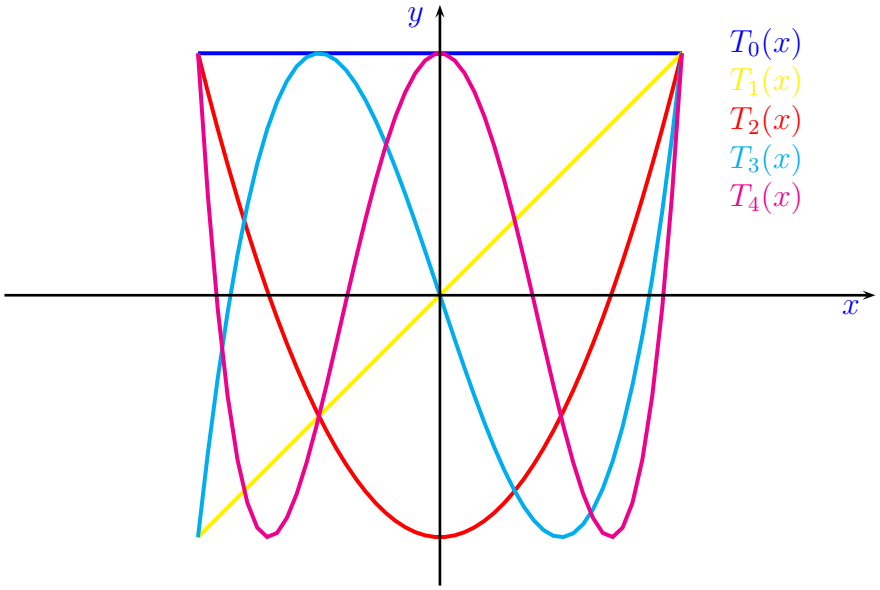


Figura 4.8: Grafico dei primi cinque polinomi di Chebyshev

Dimostrazione. Assumiamo per assurdo che sia

$$\max_{x\in[-1,1]} |p_n(x)| < \frac{1}{2^{n-1}}$$

e consideriamo il polinomio $d(x) = \tilde{T}_n(x) - p_n(x)$. Osserviamo subito che essendo sia $\tilde{T}_n(x)$ che $p_n(x)$ monici, $d(x)$ è un polinomio di grado al più $n-1$. Siano $t_0, t_1, \ldots, t_n$ i punti in cui T_n assume valore -1 e $+1$. Allora:

$$segn(d(t_k)) \quad = segn(\tilde{T}_n(t_k) - p_n(t_k)) = segn(\tilde{T}_n(t_k)).$$

Poichè $\tilde{T}_n(x)$ cambia segno n volte anche $d(x)$ cambia segno n volte e pertanto ammetterà n zeri, in contraddizione con il fatto che $d(x)$ è un polinomio di grado al più $n-1$. □

Osservazione. In verità vale un'affermazione più forte di quella del teorema, cioè se $p(x)$ è un polinomio monico di grado n diverso da $\tilde{T}_n(x)$ allora:

$$\max_{x\in[-1,1]} |p(x)| > \frac{1}{2^{n-1}}.$$

Il teorema di minimax stabilisce che, tra tutti i polinomi di grado n definiti nell'intervallo $[-1, 1]$, il polinomio di Chebyshev monico è quello che ha il massimo più piccolo. Supponendo che l'intervallo di interpolazione della funzione $f(x)$ sia appunto $[-1, 1]$ e scegliendo come nodi gi zeri del polinomio di Chebyshev risulta

$$\omega_{n+1}(x) = \tilde{T}_{n+1}(x)$$

pertanto

$$e(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\tilde{T}_{n+1}(x)$$

e, massimizzando tale errore, risulta

$$\begin{aligned}\max_{x\in[-1,1]} |e(x)| &\leq \max_{x\in[-1,1]} \left|\frac{f^{(n+1)}(\xi_x)}{(n+1)!}\right| \max_{x\in[-1,1]} |\omega_{n+1}(x)| \\ &= \frac{1}{2^n(n+1)!} \max_{x\in[-1,1]} |f^{(n+1)}(\xi_x)|.\end{aligned}$$

La crescita dell'errore può dipendere solo dalla derivata di ordine $n+1$ della funzione $f(x)$.
Se l'intervallo di interpolazione è $[a, b] \neq [-1, 1]$ allora il discorso può essere ripetuto egualmente effettuando una trasformazione lineare tra i due intervalli, nel modo riportato in Figura 4.9. Calcolando la retta nel piano (x, t) passante per i punti $(-1, a)$ e $(1, b)$:

$$t = \frac{b-a}{2}x + \frac{a+b}{2} \tag{4.12}$$

detti x_k gli zeri del polinomio di Chebyshev $T_{n+1}(x)$ allora si possono usare come nodi i valori

$$t_k = \frac{b-a}{2}x_k + \frac{a+b}{2}, \qquad k = 0, 1, \ldots, n,$$

ovvero

$$t_k = \frac{b-a}{2}\cos\frac{(2k+1)\pi}{2(n+1)} + \frac{a+b}{2} \qquad k = 0, 1, \ldots, n. \tag{4.13}$$

Il polinomio di Chebyshev, traslato nell'intervallo $[a, b]$, è

$$T_{n+1}^{[a,b]}(x) = T_{n+1}\left(\frac{2x-(b+a)}{b-a}\right),$$

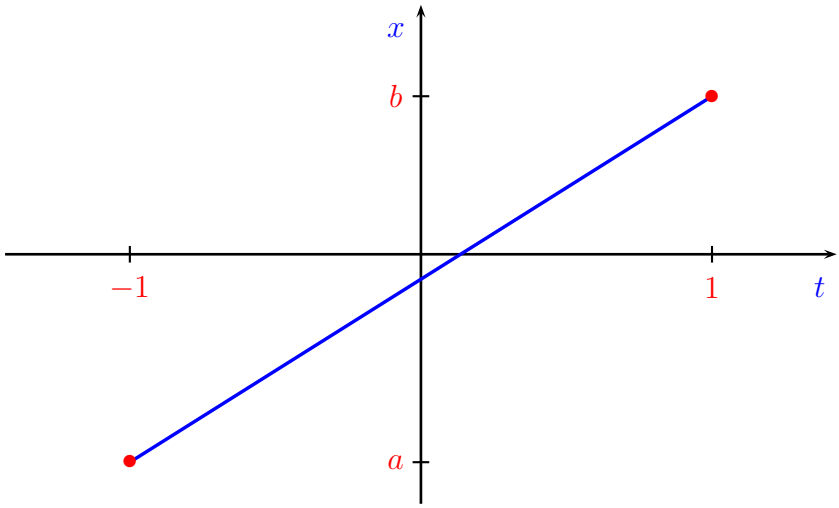


Figura 4.9: Trasformazione lineare tra gli intervalli $[-1,1]$ e $[a,b]$.

il cui coefficiente di grado massimo vale

$$2^n \frac{2^{n+1}}{(b-a)^{n+1}} = \frac{2^{2n+1}}{(b-a)^{n+1}}.$$

Se come nodi di interpolazione scegliamo i punti t_k dati da (4.13), cioè gli $n+1$ zeri del polinomio $\tilde{T}_{n+1}^{[a,b]}(x)$, allora abbiamo il polinomio monico è

$$\tilde{T}_{n+1}^{[a,b]}(x) = \frac{(b-a)^{n+1}}{2^{2n+1}} T_{n+1}\left(\frac{2x-(b+a)}{b-a}\right),$$

considerato che la trasformazione lineare inversa della (4.12) è

$$t = \frac{2x-(b+a)}{b-a}, \qquad x \in [a,b] \to t \in [-1,1]$$

quindi per l'errore dell'interpolazione vale la seguente maggiorazione:

$$\begin{aligned} \max_{x\in[a,b]} |e(x)| &\leq \max_{x\in[a,b]} \left|\frac{f^{(n+1)}(\xi_x)}{(n+1)!}\right| \max_{x\in[a,b]} |\tilde{T}_{n+1}^{[a,b]}(x)| \\ &= \max_{x\in[a,b]} \left|\frac{f^{(n+1)}(\xi_x)}{(n+1)!}\right| \frac{(b-a)^{n+1}}{2^{2n+1}}. \end{aligned}$$

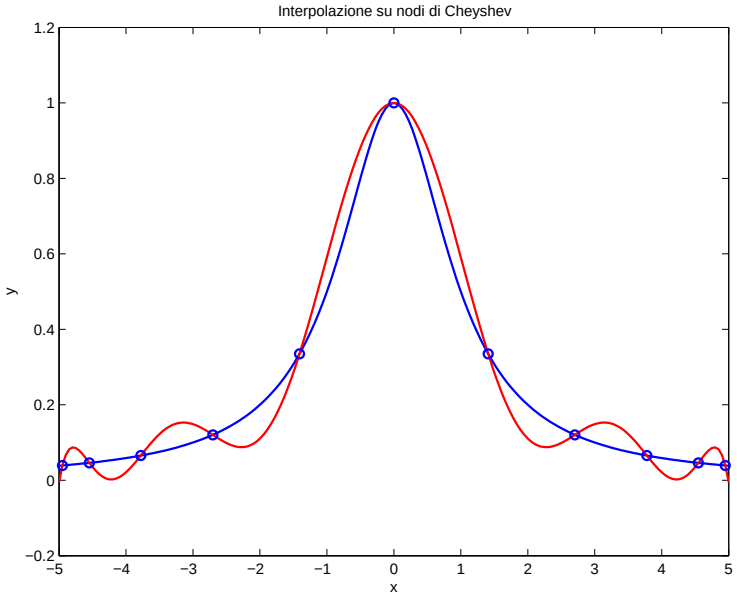


Figura 4.10: Interpolazione su nodi di Chebyshev.

Nella Figura 4.10 sono raffigurati la funzione di Runge ed il polinomio interpolante di Lagrange di grado 10 calcolato prendendo come nodi gli zeri del polinomio di Chebyshev di grado 11. Si può osservare la differenza con la Figura 4.7.

4.4 Interpolazione con Funzioni Polinomiali a Tratti

L'interpolazione polinomiale con un numero di nodi sufficientemente alto può dar luogo a polinomi interpolanti che mostrano un comportamento fortemente oscillatorio che può essere inaccettabile. In questo caso si preferisce usare una diversa strategia consistente nell'approssimare la funzione con polinomi di basso grado su sottointervalli dell'intervallo di definizione. Per esempio, supposto che l'intero n sia un multiplo di 3, denotiamo con $P_{3,j}(x)$ il polinomio di interpolazione di terzo grado associato ai nodi x_{3j-3}, x_{3j-2}, x_{3j-1}, x_{3j}, $j = 1, 2, \ldots, n/3$. Come funzione interpolante prendiamo poi la funzione:

$$I_n(x) = P_{3,j}(x) \qquad \text{in } [x_{3j-3}, x_{3j}]$$

che prende il nome di Funzione di tipo polinomiale a tratti. La tecnica esposta non è l'unica, anzi la più popolare è forse quella basata sull'uso delle cosiddette Funzioni Spline.

4.4.1 Interpolazione con Funzioni Spline

Con il termine spline si indica in lingua inglese un sottile righello usato nella progettazione degli scafi dagli ingegneri navali, per raccordare su un piano un insieme di punti (x_i, y_i), $i = 0, \ldots, n+1$.
Imponendo mediante opportune guide che il righello passi per i punti assegnati, si ottiene una curva che li interpola. Detta $y = f(x)$ l'equazione della curva definita dalla spline, sotto opportune condizioni $f(x)$ può essere approssimativamente descritta da pezzi di polinomi di terzo grado in modo che $f(x)$ e le sue prime due derivate risultino ovunque continue. La derivata terza può presentare discontinuità nei punti x_i. La spline può essere concettualmente rappresentata e generalizzata nel seguente modo.
Sia

$$\Delta =: a \equiv x_0 < x_1 < x_2 < \cdots < x_n < x_{n+1} \equiv b$$

una decomposizione dell'intervallo $[a, b]$.

Definizione 4.4.1 *Si dice funzione Spline di grado $m \geq 1$ relativa alla decomposizione Δ una funzione $s(x)$ soddisfacente le seguenti proprietà:*

1. *$s(x)$ ristretta a ciascun intervallo $[x_i, x_{i+1}]$, $i = 0, \ldots, n$, è un polinomio di grado al più m;*
2. *la derivata $s^{(k)}(x)$ è una funzione continua su $[a, b]$ per $k = 0, 1, \ldots, m-1$.*

Si verifica facilmente che l'insieme delle spline di grado assegnato è uno spazio vettoriale. In generale le spline vengono utilizzate in tutte quelle situazioni dove l'approssimazione polinomiale sull'intero intervallo non è soddisfacente. Per $m = 1$ si hanno le cosiddette spline lineari, mentre per $m = 3$ si hanno le spline cubiche.

4.4.2 Costruzione della Spline Cubica Interpolante con la Tecnica dei Momenti

Assegnata la decomposizione:

$$\Delta =: a \equiv x_0 < x_1 < x_2 < \cdots < x_n < x_{n+1} \equiv b$$

si vuole determinare una spline cubica $s(x)$ tale che

$$s(x_i) = y_i, \qquad i = 0, 1, \ldots, n+1. \tag{4.14}$$

dove y_i, $i = 0, \ldots, n+1$, sono $n+2$ assegnati valori.
Indichiamo con $s_i(x)$ la restrizione della spline nell'intervallo $[x_i, x_{i+1}]$, in cui $s_i''(x)$ è una funzione lineare mentre $s_i^{(3)}(x)$ è una costante, quindi

$$s_i''(x) = s_i''(x_i) + s_i^{(3)}(x_i)(x - x_i) \tag{4.15}$$

ovvero, posto

$$M_i = s_i''(x_i) \qquad\qquad c_i = s_i^{(3)}(x_i)$$

abbiamo

$$s_i''(x) = M_i + c_i(x - x_i). \tag{4.16}$$

Valutando (4.16) in x_{i+1} si ottiene

$$c_i = \frac{M_{i+1} - M_i}{h_i}, \qquad h_i = x_{i+1} - x_i. \tag{4.17}$$

Scriviamo lo sviluppo in serie di Taylor di $s_i(x)$ prendendo come punto iniziale x_i :

$$s_i(x) = s_i(x_i) + s_i'(x_i)(x - x_i) + s_i''(x_i)\frac{(x - x_i)^2}{2} + s_i^{(3)}(x_i)\frac{(x - x_i)^3}{6}, \tag{4.18}$$

sostituiamo i valori delle derivate seconda e terza, e calcoliamola in x_{i+1}

$$s_i(x_{i+1}) = s_i(x_i) + s_i'(x_i)(x_{i+1} - x_i) + M_i\frac{(x_{i+1} - x_i)^2}{2} + c_i\frac{(x_{i+1} - x_i)^3}{6}$$

e, imponendo le condizioni di interpolazione e sostituendo il valore dell'ampiezza dei sottointervalli, si ottiene

$$y_{i+1} = y_i + s_i'(x_i)h_i + M_i\frac{h_i^2}{2} + c_i\frac{h_i^3}{6}$$

da cui

$$\frac{y_{i+1} - y_i}{h_i} = s_i'(x_i) + M_i\frac{h_i}{2} + c_i\frac{h_i^2}{6}. \tag{4.19}$$

Scriviamo ora lo sviluppo in serie di Taylor di $s_{i-1}(x)$ prendendo come punto iniziale x_i :

$$s_{i-1}(x) = s_{i-1}(x_i) + s'_{i-1}(x_i)(x - x_i) + s''_{i-1}(x_i)\frac{(x-x_i)^2}{2} + s^{(3)}_{i-1}(x_{i-1})\frac{(x-x_i)^3}{6}$$

e sostituiamo il valori della derivate seconda (che à uguale a M_i in quanto è continua) e della derivata terza (che invece è uguala a c_{i-1} in quanto è discontinua), e poniamo $x = x_{i-1}$ e calcoliamola in x_{i-1},

$$s_{i-1}(x_{i-1}) = s_{i-1}(x_i) + s'_{i-1}(x_i)(x_{i-1} - x_i) + M_i\frac{(x_{i-1}-x_i)^2}{2} + c_{i-1}\frac{(x_{i-1}-x_i)^3}{6}.$$

Imponendo le condizioni di interpolazione anche sul nodo x_i in modo tale da assicurare la continuità della spline si ottiene

$$y_{i-1} = y_i - s'_{i-1}(x_i)h_{i-1} + M_i\frac{h^2_{i-1}}{2} - c_{i-1}\frac{h^3_{i-1}}{6}$$

$$\frac{y_{i-1} - y_i}{h_{i-1}} = -s'_{i-1}(x_i) + M_i\frac{h_{i-1}}{2} - c_{i-1}\frac{h^2_{i-1}}{6}. \tag{4.20}$$

Osserviamo dalla relazione (4.19) che $s'_i(x_i)$ può essere calcolata se sono noti i valori M_i. Di conseguenza la spline è completamente determinata se si conoscono i valori M_0, $M_1, \ldots,$ M_{n+1} (che sono detti momenti). A questo punto imponendo le condizioni di continuità della derivata prima, ovvero

$$s'_{i-1}(x_i) = s'_i(x_i)$$

sommando le equazioni (4.19) e (4.20) le derivate prime si semplificano ricavando l'equazione

$$M_i\frac{h_i + h_{i-1}}{2} - c_{i-1}\frac{h^2_{i-1}}{6} + c_i\frac{h^2_i}{6} = \frac{y_{i+1} - y_i}{h_i} + \frac{y_{i-1} - y_i}{h_{i-1}},$$

sostituendo l'espressione delle derivate terze nei due intervalli

$$c_{i-1} = \frac{M_i - M_{i-1}}{h_{i-1}}, \qquad c_i = \frac{M_{i+1} - M_i}{h_i}$$

$$3M_i(h_i + h_{i-1}) + (M_{i+1} - M_i)h_i - (M_i - M_{i-1})h_{i-1} = 6\left(\frac{y_{i+1} - y_i}{h_i} + \frac{y_{i-1} - y_i}{h_{i-1}}\right)$$

ottenendo, per $i = 1, \ldots, n$, le equazioni

$$h_{i-1}M_{i-1} + 2(h_i + h_{i-1})M_i + h_i M_{i+1} = 6\left(\frac{y_{i+1} - y_i}{h_i} + \frac{y_{i-1} - y_i}{h_{i-1}}\right) \quad (4.21)$$

che rappresentano un sistema lineare di n equazioni nelle $n+2$ incognite M_0, $M_1, \ldots, M_{n+1}$. Per risolvere il problema delle condizioni mancanti ci sono diverse possibili scelte:

1. Il modo piu semplice e quello di imporre che la spline abbia momenti nulli agli estremi dell'intervallo, ovvero

   $$M_0 = M_{n+1} = 0$$

   ottenendo la cosiddetta spline cubica naturale;

2. In alternativa si possono fissano i valori della derivata prima negli estremi, imponendo che sia

   $$\begin{aligned} s'(x_0) &= k_0 \\ s'(x_{n+1}) &= k_{n+1} \end{aligned}$$

   con k_0 e k_{n+1} valori assegnati: in tal caso la spline interpolante prende il nome di spline cubica completa;

3. Imponendo

   $$s'(x_0) = s'(x_{n+1}), \qquad s''(x_0) = s''(x_{n+1})$$

   si ottiene invece la cosiddetta spline cubica periodica.

Considerando la spline cubica naturale interpolante il sistema (4.21) si scrive:

$$A\mathbf{m} = \mathbf{b}$$

dove

$$A = \begin{bmatrix} 2(h_0+h_1) & h_1 & & & & \\ h_1 & 2(h_1+h_2) & h_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} & \\ & & & h_{n-1} & 2(h_{n-1}+h_n) & \end{bmatrix} \quad (4.22)$$

e

$$\mathbf{m} = \begin{bmatrix} M_1 \\ M_2 \\ M_3 \\ \vdots \\ M_{n-1} \\ M_n \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_{n-1} \\ b_n \end{bmatrix}$$

con

$$b_i = 6\left(\frac{y_{i+1} - y_i}{h_i} + \frac{y_{i-1} - y_i}{h_{i-1}}\right) \qquad i = 1, 2, \ldots, n.$$

4.4.3 Proprietà di Regolarità delle Spline Cubiche

Per ogni $f \in \mathcal{C}^2([a,b])$ definiamo

$$\sigma(f) = \int_{x_0}^{x_{n+1}} [f''(x)]^2 dx$$

che è, in prima approssimazione, una misura del grado di oscillazione di f. Infatti ricordando che:

$$f''(x)(1 + (f'(x))^2)^{-3/2}$$

definisce la curvatura della funzione f nel punto x, se $|f'(x)|$ è una quantità sufficientemente piccola rispetto a 1 allora la curvatura è definita approssimativamente da $f''(x)$. Conseguentemente

$$\int_a^b [f''(x)]^2 dx.$$

è una misura approssimata della curvatura totale di f su $[a,b]$.
Sia ora $s(x)$ la spline cubica naturale soddisfacente il problema di interpolazione (4.14) ed $f(x)$ una qualunque funzione con derivata seconda continua su $[a,b]$ soddisfacente anch'essa lo stesso problema di interpolazione. Assegnata cioè la decomposizione:

$$\Delta =: a \equiv x_0 < x_1 < x_2 < \cdots < x_n < x_{n+1} \equiv b$$

ed i valori $y_0, y_1, y_2, \ldots, y_n, y_{n+1}$ abbiamo

$$s(x_i) = y_i \qquad i = 0, 1, \ldots, n+1$$

$$f(x_i) = y_i \qquad i = 0, 1, \ldots, n+1.$$

Sia inoltre $e(x) = f(x) - s(x)$. Vale il seguente risultato.

Lemma 4.4.1

$$\int_{x_0}^{x_{n+1}} e''(x)s''(x)dx = e'(x_{n+1})s''(x_{n+1}) - e'(x_0)s''(x_0).$$

Dimostrazione. Osservato che:

$$e''(x)s''(x) = \frac{d}{dx}(e'(x)s''(x)) - e'(x)s^{(3)}(x),$$

si ha

$$\begin{aligned}\int_{x_0}^{x_{n+1}} e''(x)s''(x)dx &= \int_{x_0}^{x_{n+1}} \left[\frac{d}{dx}(e'(x)s''(x)) - e'(x)s^{(3)}(x)\right] dx \\ &= e'(x_{n+1})s''(x_{n+1}) - e'(x_0)s''(x_0) + \\ &\quad - \sum_{i=0}^{n} \int_{x_i}^{x_{i+1}} e'(x)s^{(3)}(x)dx.\end{aligned}$$

Poichè la derivata terza della spline è costante su ogni sottointervallo $[x_i, x_{i+1}]$, detta c_i tale costante, si può scrivere:

$$\int_{x_0}^{x_{n+1}} e''(x)s''(x)dx = e'(x_{n+1})s''(x_{n+1}) - e'(x_0)s''(x_0) - \sum_{i=0}^{n} c_i \int_{x_i}^{x_{i+1}} e'(x)dx.$$

La tesi segue poichè per ogni i risulta $e(x_i) = f(x_i) - s(x_i) = 0$. □

Teorema 4.4.1 *Se $s(x)$ è la spline naturale interpolante che soddisfa le condizioni (4.14) allora:*

$$\sigma(s) \leq \sigma(f)$$

qualunque sia f di classe $\mathcal{C}^2([a,b])$ interpolante gli stessi dati.

Dimostrazione.

$$\begin{aligned}\sigma(f) &= \int_{x_0}^{x_{n+1}} [f''(x)]^2 dx = \int_{x_0}^{x_{n+1}} [f''(x) - s''(x) + s''(x)]^2 dx \\ &= \int_{x_0}^{x_{n+1}} [e''(x) + s''(x)]^2 dx \\ &= \int_{x_0}^{x_{n+1}} [e''(x)]^2 dx + \int_{x_0}^{x_{n+1}} [s''(x)]^2 dx + 2\int_{x_0}^{x_{n+1}} e''(x)s''(x)dx.\end{aligned}$$

Poichè $s''(x)$ è una spline lineare possiamo applicare il Lemma 4.4.1 al terzo integrale a secondo membro, ottenendo

$$\begin{aligned}\sigma(f) &= \int_{x_0}^{x_{n+1}} [e''(x)]^2 dx + \int_{x_0}^{x_{n+1}} [s''(x)]^2 dx + \\ &\quad +2[e'(x_{n+1})s''(x_{n+1}) - e'(x_0)s''(x_0)].\end{aligned}$$

Poichè la spline in oggetto è naturale, segue:

$$\sigma(f) = \int_{x_0}^{x_{n+1}} [e''(x)]^2 dx + \sigma(s)$$

e dunque la tesi. □

Osservazione 1. Se $\sigma(s) = \sigma(f)$ allora $e''(x)$ è identicamente nulla. Pertanto $e(x)$ è un polinomio di primo grado e poichè esso si annulla in almeno due nodi è identicamente nullo. Di conseguenza $s \equiv f$.

Osservazione 2. La tesi del teorema (4.4.1) è verificata anche dalla spline cubica completa. Infatti in questo caso il termine

$$e'(x_{n+1})s''(x_{n+1}) - e'(x_0)s''(x_0) = 0$$

in virtù del fatto che $e'(x_{n+1}) = e'(x_0) = 0$. In definitiva abbiamo provato che la spline cubica naturale è l'unica funzione che risolve il problema di minimo:

$$\min_{f \in \mathcal{C}^2([a,b])} \int_a^b [f''(x)]^2 dx$$

$$f(x_i) = y_i \qquad i = 0, 1, \ldots, n+1$$

$$f''(a) = f''(b) = 0.$$

Analogamente la spline cubica completa risolve il problema di minimo:

$$\min_{f \in \mathcal{C}^2([a,b])} \int_a^b [f''(x)]^2 dx$$

$$f(x_i) = y_i \qquad i = 0, 1, \ldots, n+1$$

$$f'(a) = k_0, \qquad f'(b) = k_{n+1}.$$

4.5 Risoluzione di Sistemi Tridiagonali

In questo paragrafo descriviamo l'algoritmo per risolvere un sistema lineare con matrice dei coefficienti avente struttura come la (4.22) e che viene detta appunto matrice tridiagonale. Sia

$$A = \begin{bmatrix} a_1 & s_1 & & & \\ t_2 & a_2 & s_2 & & \\ & \ddots & \ddots & \ddots & \\ & & t_{n-1} & a_{n-1} & s_{n-1} \\ & & & t_n & a_n \end{bmatrix}, \qquad \det A \neq 0.$$

Assumiamo che A ammetta fattorizzazione LU.

$$L = \begin{bmatrix} 1 & & & & \\ \alpha_2 & 1 & & & \\ & \ddots & \ddots & & \\ & & \alpha_{n-1} & 1 & \\ & & & \alpha_n & 1 \end{bmatrix}, \quad U = \begin{bmatrix} \beta_1 & \lambda_1 & & & \\ & \beta_2 & \lambda_2 & & \\ & & \ddots & \ddots & \\ & & & \beta_{n-1} & \lambda_{n-1} \\ & & & & \beta_n \end{bmatrix}.$$

Tenendo presente che il prodotto di due matrici bidiagonali, una inferiore e l'altra superiore, è una matrice tridiagonale, per l'effettivo calcolo della matrici L ed U basterà imporre l'uguaglianza tra gli elementi non nulli di A e gli elementi non nulli di LU. Osserviamo innanzitutto che:

$$\beta_1 = a_1,$$

inoltre dal prodotto della i-esima riga di L per la colonna $i-1$ di U otteniamo:

$$t_i = (0, \ \ldots, \ 0, \ \underset{\substack{\uparrow \\ i-1}}{\alpha_i}, \ \underset{\substack{\uparrow \\ i}}{1}, \ 0, \ \ldots, \ 0) \cdot \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \lambda_{i-2} \\ \beta_{i-1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \\ \\ \\ \longleftarrow \text{riga } i-2 \\ \longleftarrow \text{riga } i-1 \\ \\ \\ \\ \end{matrix}$$

e quindi

$$t_i = \alpha_i \beta_{i-1} \Rightarrow \quad \alpha_i = \frac{t_i}{\beta_{i-1}} \qquad i = 2, \ldots, n.$$

Calcolando ora il prodotto della i-esima riga di L per la i-esima colonna di U abbiamo

$$a_i = (0, \ \ldots, \ 0, \ \underset{\substack{\uparrow \\ i-1}}{\alpha_i}, \ \underset{\substack{\uparrow \\ i}}{1}, \ 0, \ \ldots, \ 0) \cdot \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \lambda_{i-1} \\ \beta_i \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{array}{l} \\ \\ \\ \longleftarrow \text{riga } i-1 \\ \longleftarrow \text{riga } i \\ \\ \\ \\ \end{array}$$

cioè

$$a_i = \beta_i + \alpha_i \lambda_{i-1} \Rightarrow \quad \beta_i = a_i - \alpha \lambda_{i-1} \qquad i = 2, \ldots, n.$$

Calcoliamo ora il prodotto della i-esima riga di L per la $(i+1)$-esima colonna di U:

$$s_i = (0, \ \ldots, \ 0, \ \underset{\substack{\uparrow \\ i-1}}{\alpha_i}, \ \underset{\substack{\uparrow \\ i}}{1}, \ 0, \ \ldots, \ 0) \cdot \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \lambda_i \\ \beta_{i+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{array}{l} \\ \\ \\ \longleftarrow \text{riga } i \\ \longleftarrow \text{riga } i+1 \\ \\ \\ \\ \end{array}$$

quindi

$$s_i = \lambda_i \qquad i = 1, \ldots, n-1.$$

In definitiva la determinazione di L ed U si basa sulle seguenti formule ricorrenti:

$$\begin{aligned} \beta_1 &= a_1 \\ \beta_i &= a_i - \alpha_i s_{i-1} \quad i = 2, \ldots, n \\ \alpha_i &= \frac{t_i}{\beta_{i-1}} \qquad\quad i = 2, \ldots, n \end{aligned}$$

Risolvendo allora i sistemi $L\mathbf{y} = \mathbf{b}$ ed $U\mathbf{x} = \mathbf{y}$ si ottiene la soluzione che può essere calcolata attraverso le seguenti formule ricorrenti:

$$\begin{aligned} y_1 &= b_1 \\ y_i &= b_i - \alpha_i y_{i-1} \qquad i = 2, \ldots, n \end{aligned}$$

$$\begin{aligned} x_n &= \frac{y_n}{\beta_n} \\ x_i &= \frac{1}{\beta_i}(y_i - s_i x_{i+1}) \quad i = n-1, \ldots, 1. \end{aligned}$$

Capitolo 5

Formule di Quadratura

5.1 Formule di Quadratura di Tipo Interpolatorio

Siano assegnati due valori a, b, con $a < b$, ed una funzione f integrabile sull'intervallo (a, b). Il problema che ci poniamo è quello di costruire degli algoritmi numerici che ci permettano di valutare, con errore misurabile, il numero

$$I(f) = \int_a^b f(x)dx.$$

Diversi sono i motivi che possono portare alla richiesta di un algoritmo numerico per questi problemi.
Per esempio pur essendo in grado di calcolare una primitiva della funzione f, questa risulta così complicata da preferire un approccio di tipo numerico. Non è da trascurare poi il fatto che il coinvolgimento di funzioni, elementari e non, nella primitiva e la loro valutazione negli estremi a e b comporta comunque un'approssimazione dei risultati. Un'altra eventualità è che f sia nota solo in un numero finito di punti o comunque può essere valutata in ogni valore dell'argomento solo attraverso una routine. In questi casi l'approccio analitico non è neanche da prendere in considerazione.
Supponiamo dunque di conoscere la funzione $f(x)$ nei punti distinti $x_0, x_1, \ldots,$ x_n prefissati o scelti da noi, ed esaminiamo la costruzione di formule del tipo

$$\sum_{k=0}^{n} w_k f(x_k) \tag{5.1}$$

che approssimi realizzare $I(f)$.
Formule di tipo (5.1) si dicono di quadratura, i numeri reali $x_0, x_1, \ldots, x_n$ e $w_0, \ldots, w_n$ si chiamano rispettivamente nodi e pesi della formula di quadratura.
Il modo più semplice ed immediato per costruire formule di tipo (5.1) è quello di sostituire la funzione integranda $f(x)$ con il polinomio di Lagrange $L_n(x)$ interpolante $f(x)$ nei nodi x_i, $i = 0, \ldots, n$. Posto infatti

$$f(x) = L_n(x) + e(x)$$

dove $e(x)$ è la funzione errore, abbiamo:

$$\begin{aligned}\int_a^b f(x)dx &= \int_a^b [L_n(x) + e(x)]dx = \int_a^b L_n(x)dx + \int_a^b e(x)dx \\ &= \int_a^b \sum_{k=0}^{n} l_{nk}(x) f(x_k)dx + \int_a^b e(x)dx \\ &= \sum_{k=0}^{n} \left(\int_a^b l_{nk}(x)dx \right) f(x_k) + \int_a^b e(x)dx.\end{aligned}$$

Ponendo

$$w_k = \int_a^b l_{nk}(x)dx \qquad k = 0, 1, \ldots, n \tag{5.2}$$

e

$$R_{n+1}(f) = \int_a^b e(x)dx \tag{5.3}$$

otteniamo

$$I(f) \simeq \sum_{k=0}^{n} w_k f(x_k)$$

con un errore stabilito dalla relazione (5.3). Le formule di quadratura con pesi definiti dalle formule (5.2) si dicono interpolatorie. La quantità $R_{n+1}(f)$ prende il nome di Resto della formula di quadratura. Un utile concetto per misurare il grado di accuratezza con cui una formula di quadratura, interpolatoria o meno, approssima un integrale è il seguente.

Definizione 5.1.1 *Una formula di quadratura ha grado di precisione q se fornisce il valore esatto dell'integrale quando la funzione integranda è un*

qualunque polinomio di grado al più q ed inoltre esiste un polinomio di grado $q+1$ tale che l'errore è diverso da zero.

È evidente da questa definizione che ogni formula di tipo interpolatorio con nodi $x_0, x_1, \ldots, x_n$ ha grado di precisione almeno n. Infatti applicando una formula di quadratura costruita su $n+1$ nodi al polinomio $p_n(x)$, di grado n si ottiene:

$$\int_a^b p_n(x)dx = \sum_{i=0}^{n} w_i p_n(x_i) + R_{n+1}(f)$$

e

$$R_{n+1}(f) = \int_a^b \omega_{n+1}(x)\frac{p_n^{(n+1)}(x)}{(n+1)!}dx \equiv 0$$

ovvero la formula fornisce il risultato esatto dell'integrale, quindi $q \geq n$.

5.2 Formule di Newton-Cotes

Suddividiamo l'intervallo $[a, b]$ in n sottointervalli di ampiezza h, con

$$h = \frac{b-a}{n}$$

e definiamo i nodi

$$x_i = a + ih \qquad\qquad i = 0, 1, \ldots, n.$$

La formula di quadratura interpolatoria costruita su tali nodi, cioè

$$\int_a^b f(x)dx = \sum_{i=0}^{n} w_i f(x_i) + R_{n+1}(f)$$

è detta Formula di Newton-Cotes.
Una proprietà di cui godono i pesi delle formule di Newton-Cotes è la cosiddetta proprietà di simmetria. Infatti poichè i nodi sono a due a due simmetrici rispetto al punto medio c dell'intervallo $[a, b]$, cioè $c = (x_i + x_{n-i})/2$, per ogni i, tale proprietà si ripercuote sui pesi che infatti sono a due a due uguali, cioè $w_i = w_{n-i}$, per ogni i. Descriviamo ora due esempi di formule di Newton-Cotes.

5.2.1 Formula dei Trapezi

Siano $x_0 = a$, $x_1 = b$ e $h = b - a$.

$$T_2 = w_0 f(x_0) + w_1 f(x_1)$$

$$w_0 = \int_a^b l_{1,0}(x)dx = \int_a^b \frac{x - x_1}{x_0 - x_1}dx = \int_a^b \frac{x - b}{a - b}dx$$

$$= \frac{1}{a - b}\left[(x - b)^2\right]_{x=a}^{x=b} = \frac{h}{2}.$$

Poichè i nodi scelti sono simmetrici rispetto al punto medio $c = (a + b)/2$ è

$$w_1 = w_0 = \frac{h}{2}.$$

Otteniamo dunque la formula

$$T_2 = \frac{h}{2}\left[f(a) + f(b)\right].$$

che viene detta Formula dei Trapezi. Per quanto riguarda il resto abbiamo

$$R_2(f) = \frac{1}{2}\int_a^b (x - a)(x - b)f''(\xi_x)dx.$$

Prima di vedere come tale espressione può essere manipolata dimostriamo il seguente teorema che è noto come teorema della media generalizzato.

Teorema 5.2.1 *Siano $f, g : [a, b] \to \mathbb{R}$, funzioni continue con $g(x)$ a segno costante e $g(x) \neq 0$ per ogni $x \in]a, b[$. Allora*

$$\int_a^b f(x)g(x)dx = f(\xi)\int_a^b g(x)dx, \qquad \xi \in [a, b]. \ \square$$

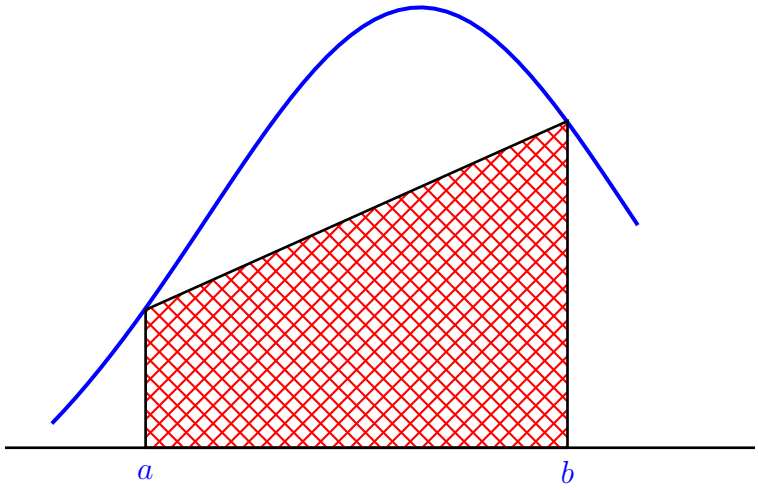
Poichè la funzione $(x - a)(x - b)$ è a segno costante segue:

$$R_2(f) = \frac{1}{2}f''(\eta)\int_a^b (x - a)(x - b)dx$$

posto $x = a + ht$ otteniamo

$$R_2(f) = \frac{1}{2}f''(\eta)h^3 \int_0^1 t(t-1)dt = -\frac{1}{12}h^3 f''(\eta).$$

L'interpretazione geometrica della formula del trapezio è riassunta nella seguente figura, l'area tratteggiata (ovvero l'integrale della funzione viene approssimato attraverso l'area del trapezio che ha come basi i valori della funzione in a e b e come altezza l'intervallo $[a, b]$).



5.2.2 Formula di Simpson

Siano $x_0 = a$, $x_2 = b$ mentre poniamo $x_1 = c$, punto medio dell'intervallo $[a, b]$. Allora

$$S_3 = w_0 f(a) + w_1 f(c) + w_2 f(b).$$

Posto

$$h = \frac{b-a}{2}$$

abbiamo

$$w_0 = \int_a^b l_{2,0}(x)dx = \int_a^b \frac{(x-c)(x-b)}{(a-c)(a-b)}dx.$$

Effettuando il cambio di variabile $x = c + ht$ è facile calcolare quest'ultimo integrale, infatti

$$x = a \;\Rightarrow\; a = c + ht \;\Rightarrow\; a - c = ht \;\Rightarrow\; -h = ht \;\Rightarrow\; t = -1$$

e

$$x = b \;\Rightarrow\; b = c + ht \;\Rightarrow\; b - c = ht \;\Rightarrow\; h = ht \;\Rightarrow\; t = 1.$$

Inoltre $a - c = -h$ e $a - b = -2h$ mentre

$$x - c = c + ht - c = ht, \qquad x - b = c + ht - b = c - b + ht = -h + ht = h(t-1),$$

ed il differenziale $dx = hdt$ cosicchè

$$\begin{aligned} w_0 &= \int_a^b \frac{(x-c)(x-b)}{(a-c)(a-b)}dx = \int_{-1}^1 \frac{hth(t-1)}{(-h)(-2h)}hdt \\ &= \frac{h}{2}\int_{-1}^1 (t^2 - t)dt = \frac{h}{2}\int_{-1}^1 t^2 dt = \frac{h}{2}\left[\frac{t^3}{3}\right]_{-1}^1 = \frac{h}{3}. \end{aligned}$$

Per la proprietà di simmetria è anche

$$w_2 = w_0 = \frac{h}{3}$$

mentre possiamo calcolare w_1 senza ricorrere alla definizione. Infatti possiamo notare che la formula deve fornire il valore esatto dell'integrale quando la funzione è costante nell'intervallo $[a, b]$, quindi possiamo imporre che, prendendo $f(x) = 1$ in $[a, b]$, sia

$$\int_a^b dx = b - a = \frac{h}{3}(f(a) + f(b)) + w_1 f(c)$$

da cui segue

$$w_1 = b - a - \frac{2}{3}h = 2h - \frac{2}{3}h = \frac{4}{3}h.$$

Dunque

$$S_3 = \frac{h}{3}\left[f(a) + 4f(c) + f(b)\right].$$

Questa formula prende il nome di Formula di Simpson. Per quanto riguarda l'errore si può dimostrare, e qui ne omettiamo la prova, che vale la seguente relazione

$$R_3(f) = -h^5 \frac{f^{(4)}(\sigma)}{90} \qquad \sigma \in (a, b),$$

che assicura che la formula ha grado di precisione 3.

5.3 Formule di Quadratura Composte

Come abbiamo già avuto modo di vedere le formule di quadratura interpolatorie vengono costruite approssimando su tutto l'intervallo di integrazione la funzione integranda con un unico polinomio, quello interpolante la funzione sui nodi scelti. Per formule convergenti la precisione desiderata si ottiene prendendo n sufficientemente grande. In tal modo comunque, per ogni fissato n, bisogna costruire la corrispondente formula di quadratura. Una strategia alternativa che ha il pregio di evitare la costruzione di una nuova formula di quadratura, e che spesso produce risultati più apprezzabili, è quella delle formule composte. Infatti scelta una formula di quadratura l'intervallo di integrazione (a, b) viene suddiviso in N sottointervalli di ampiezza h,

$$h = \frac{b-a}{N} \tag{5.4}$$

sicchè

$$\int_a^b f(x)dx = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} f(x)dx$$

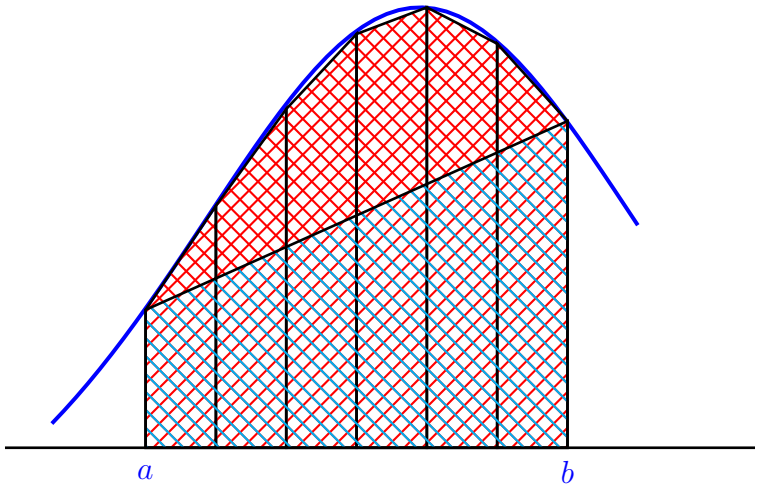
dove i punti x_i sono:

$$x_i = a + ih \qquad i = 0, \ldots, N \tag{5.5}$$

quindi la formula di quadratura viene applicata ad ognuno degli intervalli $[x_i, x_{i+1}]$. Il grado di precisione della formula di quadratura composta coincide con il grado di precisione della formula da cui deriva. Descriviamo ora la Formula dei Trapezi Composta.

5.3.1 Formula dei Trapezi Composta

Per quanto visto in precedenza suddividiamo l'intervallo $[a, b]$ in N sottointervalli, ognuno di ampiezza data da h, come in (5.4), e con i nodi x_i definiti in (5.5). Applichiamo quindi in ciascuno degli N intervalli $[x_i, x_{i+1}]$ la formula dei trapezi. Nella seguente figura sono evidenziate le aree che approssimano l'integrale utilizzando la formula dei trapezi semplice e quella composta.



Applicando la formula dei trapezi a ciascun sottointervallo si ottiene

$$\int_a^b f(x)dx = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} f(x)dx = \sum_{i=0}^{N-1} \left[\frac{h}{2} \left(f(x_i) + f(x_{i+1}) \right) - \frac{1}{12} h^3 f''(\eta_i) \right]$$

con $\eta_i \in (x_i, x_{i+1})$. Scrivendo diversamente la stessa espressione

$$\begin{aligned} \int_a^b f(x)dx &= \frac{h}{2} \left(f(x_0) + f(x_N) \right) + h \sum_{i=1}^{N-1} f(x_i) - \frac{1}{12} h^3 \sum_{i=0}^{N-1} f''(\eta_i) \\ &= \frac{h}{2} \left(f(x_0) + f(x_N) \right) + h \sum_{i=1}^{N-1} f(x_i) - \frac{1}{12} h^3 N f''(\eta) \end{aligned}$$

dove $\eta \in (a, b)$. L'esistenza di tale punto η è garantito dal cosiddetto Teorema della media nel discreto applicato a $f''(x)$, che stabilisce che se $g(x)$ è una funzione continua in un intervallo $[a, b]$ e $\eta_i \in [a, b]$ $i = 1, N$, sono N punti distinti, allora esiste un punto $\eta \in (a, b)$ tale che

$$\sum_{i=1}^{N} g(\eta_i) = N g(\eta).$$

Dunque la formula dei trapezi composta è data da:

$$T_C(h) = \frac{h}{2} \left(f(x_0) + f(x_N) \right) + h \sum_{i=1}^{N-1} f(x_i)$$

con resto

$$R_T = -\frac{1}{12}h^3 N f''(\eta) = -\frac{1}{12}\frac{(b-a)^3}{N^3}Nf''(\eta) = -\frac{1}{12}\frac{(b-a)^3}{N^2}f''(\eta).$$

Quest'ultima formula può essere utile per ottenere a priori una suddivisione dell'intervallo $[a, b]$ in un numero di intervalli che permetta un errore non superiore ad una prefissata tolleranza. Infatti

$$|R_T| \leq \frac{1}{12}\frac{(b-a)^3}{N^2}M, \qquad M = \max_{x\in[a,b]} |f''(x)|.$$

Imponendo che $|R_T| \leq \varepsilon$, precisione prefissata, segue

$$N_\varepsilon \geq \sqrt{\frac{(b-a)^3 M}{12\varepsilon}}. \tag{5.6}$$

Tuttavia questo numero spesso risulta una stima eccessiva a causa della maggiorazione della derivata seconda tramite M.

Esempio 5.3.1 *Determinare il numero di intervalli cui suddividere l'intervallo di integrazione per approssimare*

$$\int_1^2 \log x \; dx$$

con la formula dei trapezi composta con un errore inferiore a $\varepsilon = 10^{-4}$.

La derivata seconda della funzione integranda è

$$f''(x) = -\frac{1}{x^2}$$

quindi il valore di M è 1. Dalla relazione (5.6) segue che

$$N_\varepsilon \geq \sqrt{\frac{1}{12\varepsilon}} = 29.$$

5.3.2 Formula di Simpson Composta

Per ottenere la formula di Simpson composta, si procede esattamente come per la formula dei trapezi composta. Suddividiamo $[a, b]$ in N intervalli di ampiezza h, con N numero pari. Allora

$$\begin{aligned}\int_a^b f(x)dx &= \sum_{i=0}^{\frac{N}{2}-1} \int_{x_{2i}}^{x_{2i+2}} f(x)dx \\ &= \sum_{i=0}^{\frac{N}{2}-1} \left[\frac{h}{3}\left(f(x_{2i}) + 4f(x_{2i+1}) + f(x_{2i+2})\right) - \frac{h^5}{90} f^{(4)}(\eta_i)\right] \\ &= \frac{h}{3} \sum_{i=0}^{\frac{N}{2}-1} \left[f(x_{2i}) + 4f(x_{2i+1}) + f(x_{2i+2})\right] - \frac{h^5 N}{180} f^{(4)}(\eta)\end{aligned}$$

dove $\eta_i \in (x_i, x_{i+1})$ e $\eta \in (a, b)$.
La formula di Simpson composta è dunque

$$\begin{aligned}S_C(h) &= \frac{h}{3} \sum_{i=0}^{\frac{n}{2}-1} \left[f(x_{2i}) + 4f(x_{2i+1}) + f(x_{2i+2})\right] \\ &= \frac{h}{3} \left[f(x_0) + f(x_n) + 2\sum_{i=1}^{\frac{n}{2}-1} f(x_{2i}) + 4\sum_{i=1}^{\frac{n}{2}-1} f(x_{2i+1})\right]\end{aligned}$$

mentre la formula dell'errore è

$$R_S = -\frac{(b-a)^5}{180N^4} f^{(4)}(\eta)$$

Anche quest'ultima formula talvolta può essere utile per ottenere a priori una suddivisione dell'intervallo $[a, b]$ in un numero di intervalli che permetta un errore non superiore ad una prefissata tolleranza. Infatti

$$|R_S| \leq \frac{1}{180} \frac{(b-a)^5}{N^4} M, \qquad M = \max_{x \in [a,b]} |f^{(\mathrm{iv})}(x)|.$$

Imponendo che $|R_S| \leq \varepsilon$ segue

$$N_\varepsilon \geq \sqrt[4]{\frac{(b-a)^5 M}{180\varepsilon}}. \tag{5.7}$$

Esempio 5.3.2 *Risolvere il problema descritto nell'esempio 5.3.1 applicando la formula di Simpson composta.*

La derivata quarta della funzione integranda è

$$f^{\text{iv}}(x) = -\frac{6}{x^4}$$

quindi è maggiorata da $M = 6$. Dalla relazione (5.7) segue che

$$N_\varepsilon \geq \sqrt[4]{\frac{6}{180\varepsilon}} > 4,$$

quindi $N_\varepsilon \geq 6$.

5.3.3 La formula del punto di mezzo

Sia c il punto medio dell'intervallo $[a, b]$. Sviluppiamo $f(x)$ in serie di Taylor prendendo c come punto iniziale:

$$f(x) = f(c) + f'(c)(x - c) + \frac{f''(\xi_x)}{2}(x - c)^2, \qquad \xi_x \in [a, b].$$

Integrando membro a membro

$$\begin{aligned}\int_a^b f(x)dx &= \int_a^b f(c)dx + f'(c)\int_a^b (x - c)dx + \int_a^b \frac{f''(\xi_x)}{2}(x - c)^2 dx \\ &= (b - a)f(c) + \int_a^b \frac{f''(\xi_x)}{2}(x - c)^2 dx.\end{aligned}$$

Poichè la funzione $x - c$ è dispari rispetto a c il suo integrale nell'intervallo $[a, b]$ è nullo. La formula

$$\int_a^b f(x)dx \simeq (b - a)f(c)$$

prende appunto il nome di formula del punto di mezzo (o di midpoint).
Per quanto riguarda l'errore abbiamo

$$\begin{aligned}R(f) &= \int_a^b \frac{f''(\xi_x)}{2}(x - c)^2 dx \\ &= \frac{f''(\xi)}{2}\int_a^b (x - c)^2 dx.\end{aligned}$$

In questo caso la funzione $(x-c)^2$ è a segno costante quindi è stato possibile applicare il teorema 5.2.1. Calcoliamo ora l'integrale

$$\int_a^b (x-c)^2 dx = 2\int_c^b (x-c)^2 = \frac{2}{3}\left[(x-c)^3\right]_c^b = \frac{h^3}{12}$$

avendo posto $h = b-a$. L'espressione del resto di tale formula è quindi

$$R(f) = \frac{h^3}{24} f''(\xi).$$

Osserviamo che la formula ha grado di precisione 1, come quella dei trapezi, però richiede il calcolo della funzione solo nel punto medio dell'intervallo mentre la formula dei trapezi necessita di due valutazioni funzionali.

5.3.4 Formula del punto di mezzo composta

Anche in questo caso suddividiamo l'intervallo $[a,b]$ in N intervallini di ampiezza h, con N pari. Allora

$$\begin{aligned}
\int_a^b f(x)dx &= \sum_{i=0}^{\frac{N}{2}-1} \int_{x_{2i}}^{x_{2i+2}} f(x)dx \\
&= \sum_{i=0}^{\frac{N}{2}-1} \left[2hf(x_{2i+1}) + \frac{(2h)^3}{24} f''(\eta_i)\right] \\
&= 2h \sum_{i=0}^{\frac{N}{2}-1} f(x_{2i+1}) + \frac{Nh^3}{6} f''(\eta) \\
&= 2h \sum_{i=0}^{\frac{N}{2}-1} f(x_{2i+1}) + \frac{(b-a)^3}{6N^2} f''(\eta)
\end{aligned}$$

dove $\eta_i \in (x_{2i}, x_{2i+2})$ e $\eta \in (a,b)$. La formula del punto di mezzo composta è dunque

$$M_C(h) = 2h \sum_{i=0}^{\frac{N}{2}-1} f(x_{2i+1})$$

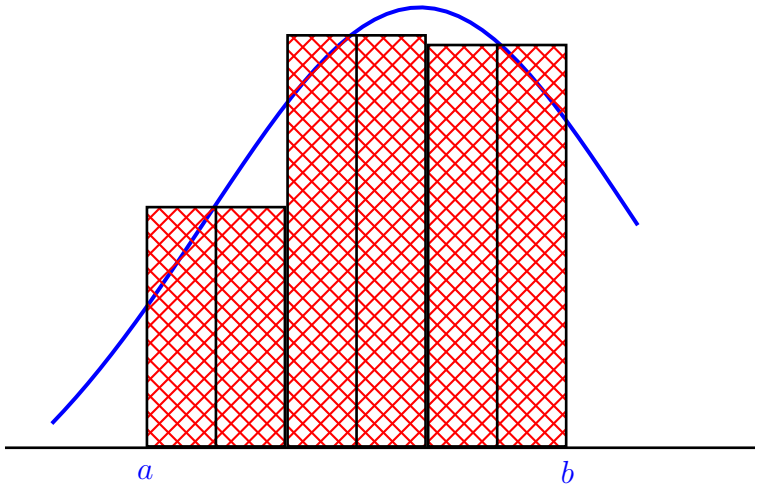


Figura 5.1: Formula del Punto di Mezzo Composta

mentre il resto è

$$R_M = \frac{(b-a)^3}{6N^2} f''(\eta). \tag{5.8}$$

Se ε è la tolleranza fissata risulta

$$|R_M| \leq \frac{1}{6}\frac{(b-a)^3}{N^2} M, \qquad M = \max_{x\in[a,b]} |f''(x)|.$$

Imponendo che $|R_T| \leq \varepsilon$, precisione prefissata, segue

$$N_\varepsilon \geq \sqrt{\frac{(b-a)^3 M}{6\varepsilon}}. \tag{5.9}$$

Nella Figura 5.1 sono evidenziate le aree che approssimano l'integrale utilizzando la formula del punto di mezzo composta.

Esempio 5.3.3 *Risolvere il problema descritto nell'esempio 5.3.1 applicando la formula di Simpson composta.*

La derivata seconda della funzione integranda è maggiorata da $M = 1$. Da (5.9) risulta

$$N_\varepsilon \geq \sqrt{\frac{1}{6\varepsilon}} > 40.$$

Capitolo 6

Metodi numerici per equazioni differenziali

6.1 Derivazione numerica

La risoluzione numerica di problemi differenziali (come equazioni differenziali ordinarie ed equazioni alle derivate parziali) è uno dei più importanti argomenti del Calcolo Numerico in quanto spesso non sono trattabili dal punto di vista analitico e contemporaneamente costituiscono lo strumento più efficace per la descrizione di problemi fisici, chimici e, in generale, delle scienze applicate. La risoluzione numerica di questi problemi passa anche attraverso il processo di discretizzazione delle derivate (totali o parziali), ovvero la loro approssimazione, che appunto è detta Derivazione numerica. Nei prossimi paragrafi capitolo affronteremo il problema relativo all'approssimazione delle derivate prima e seconda di una funzione in un punto del dominio utilizzando opportune combinazioni lineari tra i valori assunti dalla funzione in tale punto e in altri punti ad esso adiacenti. Tali approssimazioni saranno utilizzate anche per derivare semplici metodi per l'approssimazione della soluzione numerica di equazioni differenziali del primo ordine.

6.1.1 Approssimazione discreta delle derivate

Come detto in precedenza supponiamo che $f \in \mathcal{C}^k([a,b])$ e suddividiamo l'intervallo di variabilità di t in sottointervalli di ampiezza h. Consideriamo tre punti consecutivi appartenenti a tale reticolazione, rispettivamente t_{n-1},

t_n e t_{n+1} tali che

$$t_{n-1} = t_n - h, \qquad t_{n+1} = t_n + h.$$

Scriviamo lo sviluppo in serie di Taylor di $f(t_{n+1})$ prendendo come punto iniziale t_n:

$$f(t_{n+1}) = f(t_n) + hf'(t_n) + \frac{h^2}{2}f''(t_n) + \frac{h^3}{6}f'''(t_n) + \frac{h^4}{24}f^{\text{iv}}(\xi_n),\ \xi_n \in [t_n, t_{n+1}]$$

e procediamo in modo analogo per $f(t_{n-1})$:

$$f(t_{n-1}) = f(t_n) - hf'(t_n) + \frac{h^2}{2}f''(t_n) - \frac{h^3}{6}f'''(t_n) + \frac{h^4}{24}f^{\text{iv}}(\eta_n),\ \eta_n \in [t_{n-1}, t_n].$$

Sommiamo ora le due espressioni

$$f(t_{n+1}) + f(t_{n-1}) = 2f(t_n) + h^2 f''(t_n) + \frac{h^4}{24}\left[f^{\text{iv}}(\xi_n) + f^{\text{iv}}(\eta_n)\right]$$

ricavando

$$f''(t_n) = \frac{f(t_{n+1}) - 2f(t_n) + f(t_{n-1})}{h^2} - \frac{h^2}{24}\left[f^{\text{iv}}(\xi_n) + f^{\text{iv}}(\eta_n)\right]$$

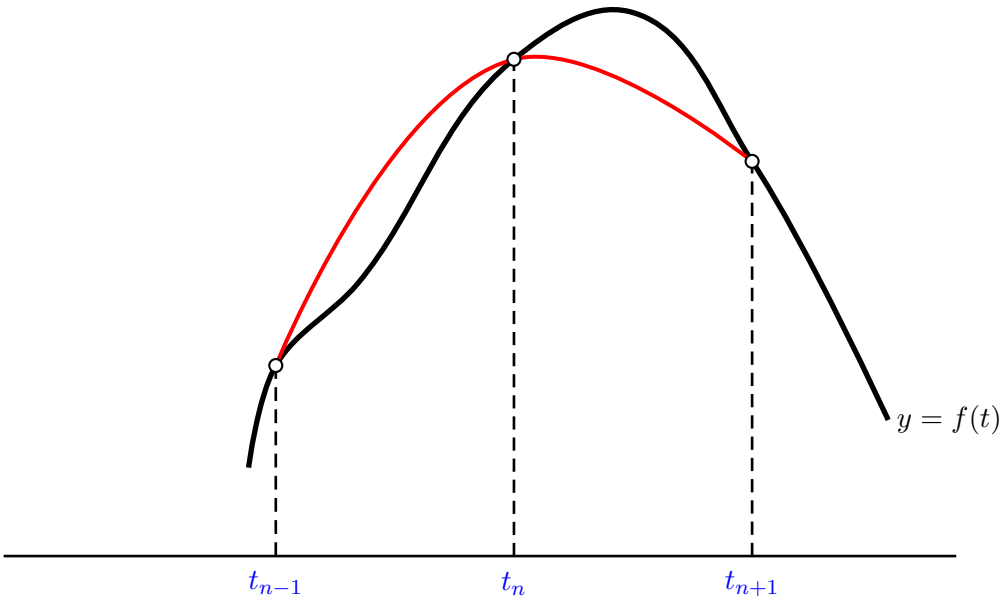
e, trascurando l'ultimo termine, l'approssimazione della derivata seconda è:

$$f''(t_n) \simeq \frac{f(t_{n+1}) - 2f(t_n) + f(t_{n-1})}{h^2} \tag{6.1}$$

mentre si può provare che l'errore vale:

$$E(f''(t_n)) = -\frac{h^2}{12}f^{\text{iv}}(\xi), \qquad \xi \in [t_{n-1}, t_{n+1}].$$

Nel seguente grafico viene evidenziata l'interpretazione geometrica della formula appena ricavata.



Infatti l'approssimazione appena trovata coincide con il valore della derivata seconda della parabola passante per i punti $(t_{n-1}, f(t_{n-1}))$, $(t_n, f(t_n))$ e $(t_{n+1}, f(t_{n+1}))$.
Infatti scrivendo l'equazione di tale parabola come:

$$p(t) = a(t - t_n)(t - t_{n-1}) + b(t - t_{n-1}) + c$$

risulta

$$\begin{aligned} c &= f(t_{n-1}) \\ b &= \frac{f(t_n) - f(t_{n-1})}{h} \\ a &= \frac{f(t_{n+1}) - 2f(t_n) + f(t_{n-1})}{2h^2} \end{aligned}$$

e la proprietà segue poichè:

$$p''(t) = 2a = \frac{f(t_{n+1}) - 2f(t_n) + f(t_{n-1})}{h^2}.$$

Poniamoci il problema di approssimare derivata prima e procediamo nello stesso modo cioè scrivendo le serie di Taylor per $f(t_{n+1})$ e $f(t_{n-1})$:

$$f(t_{n+1}) = f(t_n) + hf'(t_n) + \frac{h^2}{2}f''(t_n) + \frac{h^3}{6}f'''(\sigma_n),\ \sigma_n \in [t_n, t_{n+1}]$$

$$f(t_{n-1}) = f(t_n) - hf'(t_n) + \frac{h^2}{2}f''(t_n) - \frac{h^3}{6}f'''(\mu_n),\ \ \mu_n \in [t_{n-1}, t_n]$$

e questa volta sottraiamo la seconda dalla prima:

$$f(t_{n+1}) - f(t_{n-1}) = 2hf'(t_n) + \frac{h^3}{6}\left[f'''(\sigma_n) + f'''(\mu_n)\right]$$

ottenendo

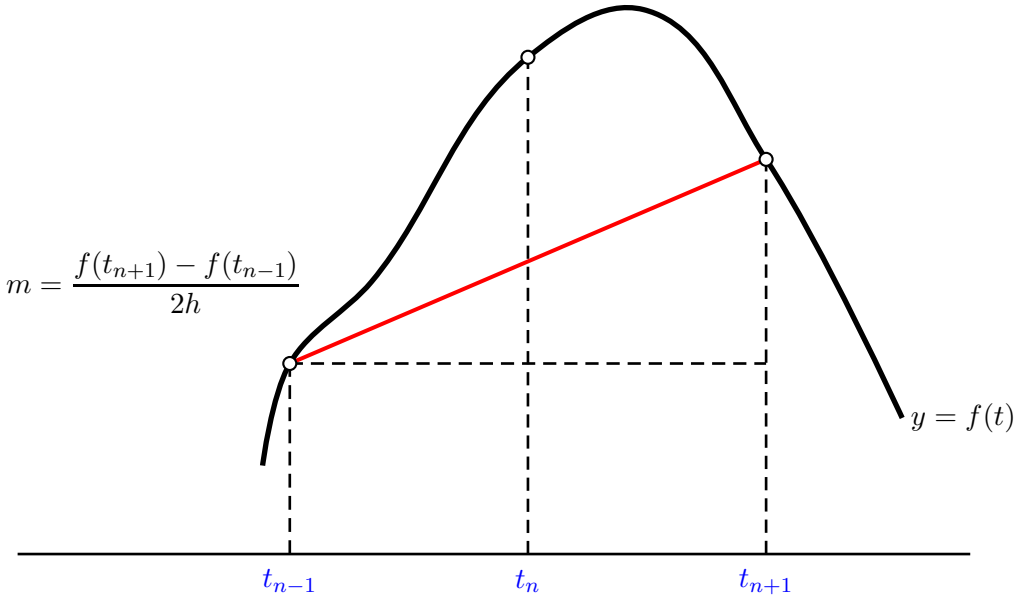
$$f'(t_n) = \frac{f(t_{n+1}) - f(t_{n-1})}{2h} - \frac{h^2}{12}\left[f'''(\sigma_n) + f'''(\mu_n)\right]$$

e, trascurando l'ultimo termine, l'approssimazione della derivata prima è:

$$f'(t_n) \simeq \frac{f(t_{n+1}) - f(t_{n-1})}{2h} \tag{6.2}$$

mentre si può provare che l'errore vale:

$$E(f'(t_n)) = -\frac{h^2}{6}f'''(\delta), \qquad \delta \in [t_{n-1}, t_{n+1}].$$



La formula (6.2) prende il nome di formula alle differenze centrali. Osserviamo che sia per questa che per l'approssimazione numerica per la derivata

seconda l'errore dipende da h^2, sono formule cioè *del secondo ordine*. Vediamo ora altre due approssimazioni per la derivata prima. Infatti possiamo anche scrivere:

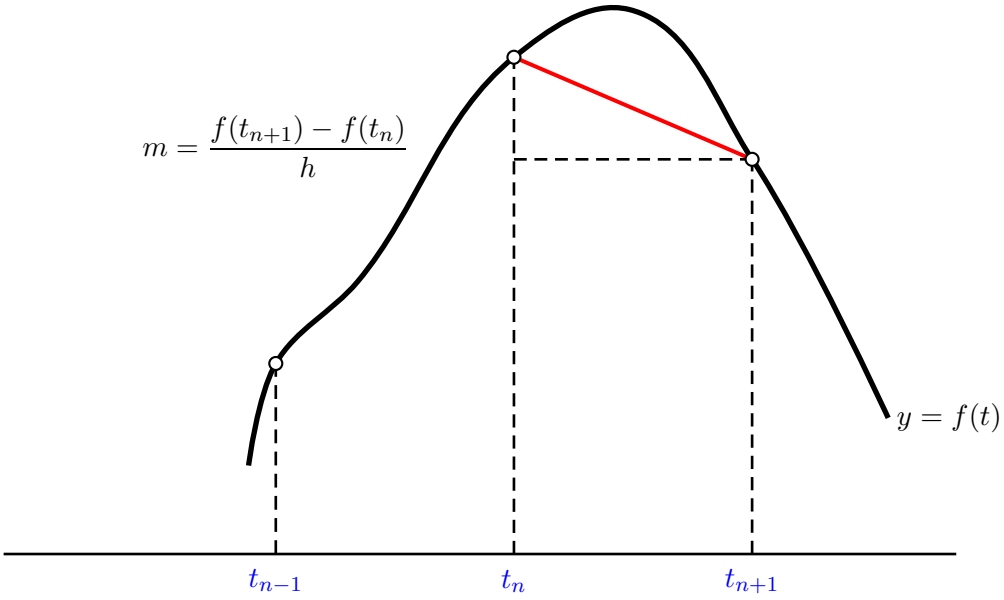
$$f(t_{n+1}) = f(t_n) + hf'(t_n) + \frac{h^2}{2}f''(\xi_n), \qquad \xi_n \in [t_n, t_{n+1}]$$

da cui si ricava immediatamente la formula alle differenze in avanti:

$$f'(t_n) \simeq \frac{f(t_{n+1}) - f(t_n)}{h} \tag{6.3}$$

con errore

$$E(f'(t_n)) = -\frac{h}{2}f''(\xi_n).$$



Analogamente si ricava

$$f(t_{n-1}) = f(t_n) - hf'(t_n) + \frac{h^2}{2}f''(\mu_n), \qquad \mu_n \in [t_{n-1}, t_n]$$

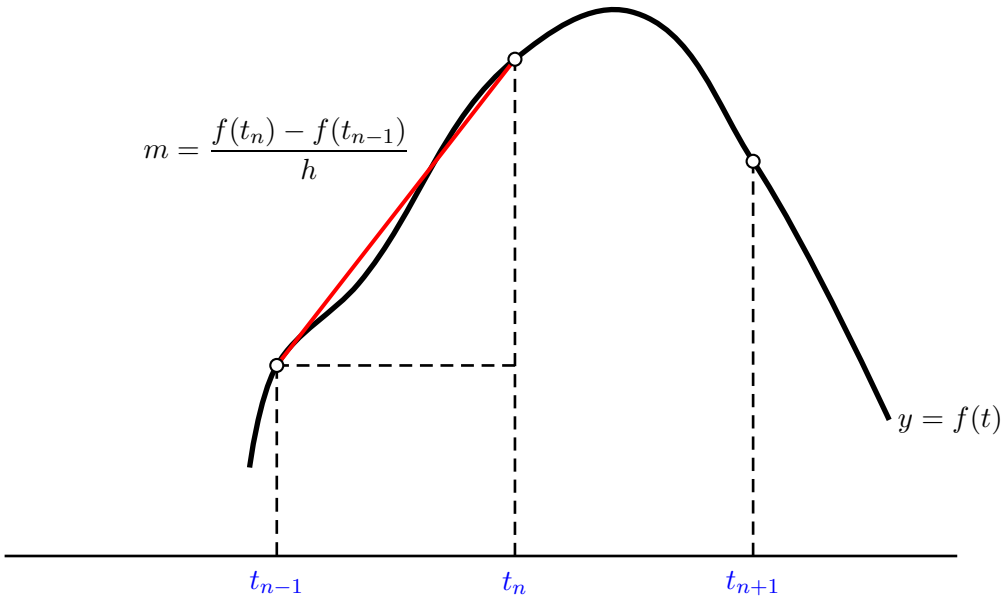
da cui si ricava immediatamente la formula alle differenze all'indietro:

$$f'(t_n) \simeq \frac{f(t_n) - f(t_{n-1})}{h} \tag{6.4}$$

con errore

$$E(f'(t_n)) = -\frac{h}{2}f''(\mu_n).$$

Queste due formule hanno ordine 1, quindi sono meno precise rispetto alla formula alle differenze centrali, tuttavia hanno il pregio di poter essere applicate quando la funzione è discontinua (oppure non è definita) a destra o a sinistra di t_n.



6.2 Metodi numerici per equazioni differenziali ordinarie

Supponiamo che sia assegnato il seguente problema differenziale:

$$\begin{cases} y'(t) = f(t, y(t)) \\ y(t_0) = y_0 \end{cases} \tag{6.5}$$

dove $f : [t_0, T] \times \mathbb{R} \to \mathbb{R}$ è una funzione continua rispetto a t e *Lipschitziana* rispetto a y, cioè esiste una costante positiva L tale che, per ogni $x, y \in \mathbb{R}$, risulta

$$|f(t, x) - f(t, y)| \leq L|x - y|, \qquad \forall t \in [t_0, T].$$

Il problema (6.5) prende il nome di *problema di Cauchy del primo ordine ai valori iniziali*. Risolvere (6.5) significa determinare una funzione $y(t)$ di classe $\mathcal{C}^1([t_0, T])$ la cui derivata prima soddisfi l'equazione assegnata e che passi per il punto (t_0, y_0). In base alle ipotesi fatte sulla funzione $f(t, y(t))$ il teorema di Cauchy assicura l'esistenza e l'unicità di tale funzione.

Teorema 6.2.1 (di Cauchy) *Sia $f(t, y) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, una funzione definita e continua per ogni (t, y) appartenente alla regione $[t_0, T] \times \mathbb{R}$, e sia inoltre Lipschitziana rispetto a y allora per ogni condizione iniziale esiste un'unica soluzione continua e differenziabile $y(t)$ del problema (6.5).*

L'equazione (6.5) dipende solo dalla derivata prima della soluzione, mentre si possono avere anche problemi di ordine superiore del tipo:

$$y^{(m)}(t) = f(t, y, y', y'', \ldots, y^{(m-1)}(t)).$$

È tuttavia possibile ricondursi ad un sistema differenziale del primo ordine con alcuni cambi di variabile, infatti, posto

$$\begin{cases} y_1 &= y \\ y_2 &= y' \\ y_3 &= y'' \\ \vdots & \\ y_m &= y^{(m-1)} \end{cases}$$

si ottiene il sistema differenziale equivalente:

$$\begin{cases} y_1' &= y_2 \\ y_2' &= y_3 \\ y_3' &= y_4 \\ \vdots & \\ y_{m-1}' &= y_m \\ y_m' &= f(t, y_1, y_2, \ldots, y_m) \end{cases}$$

da cui, ponendo

$$A = \begin{bmatrix} 0 & 1 & & & \\ 0 & 0 & 1 & & \\ \vdots & \ddots & \ddots & \ddots & \\ \vdots & & 0 & 0 & 1 \\ 0 & \ldots & \ldots & 0 & 0 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{m-1} \\ y_m \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ f(t, \mathbf{y}) \end{bmatrix}$$

si ricava il sistema differenziale in forma compatta:

$$\mathbf{y}' = A\mathbf{y} + \mathbf{f}.$$

Descriveremo nel seguito alcune classi di metodi per equazioni differenziali del primo ordine, considerando sempre che tali metodi possono essere applicati anche a sistemi. Tali metodi ovviamente non forniscono in forma chiusa l'espressione della soluzione $y(t)$ ma solo una sua approssimazione in un insieme discreto di punti. Se siamo interessati alla funzione $y(t)$ nell'intervallo $[t_0, T]$ lo suddividiamo in N sottointervalli ciascuno di ampiezza $h = (T - t_0)/N$ e definiamo i punti

$$t_n = t_{n-1} + h = t_0 + nh, \qquad n = 0, \ldots, N$$

dove la soluzione verrà approssimata.
Scriviamo l'equazione di Cauchy per $t = t_n$:

$$y'(t_n) = f(t_n, y(t_n))$$

e approssiamo la derivata prima con la formula alle differenze in avanti:

$$\frac{y(t_{n+1}) - y(t_n)}{h} \simeq f(t_n, y(t_n))$$

da cui, definendo le approssimazioni $y_n \simeq y(t_n)$ e $y_{n+1} \simeq y(t_{n+1})$ si ottiene la seguente uguaglianza tra quantità approssimate:

$$\frac{y_{n+1} - y_n}{h} = hf(t_n, y_n) \quad \Leftrightarrow \quad y_{n+1} = y_n + hf(t_n, y_n).$$

Tale metodo va sotto il nome di Metodo di Eulero Esplicito in quanto consente, noto y_n, di calcolare esplicitamente l'approssimazione nel punto successivo.
Scrivendo invece l'equazione di Cauchy per $t = t_{n+1}$:

$$y'(t_n) = f(t_{n+1}, y(t_{n+1}))$$

e approssimando la derivata prima con la formula alle differenze all'indietro:

$$\frac{y(t_{n+1}) - y(t_n)}{h} \simeq f(t_{n+1}, y(t_{n+1}))$$

si ottiene il cosiddetto Metodo di Eulero Implicito:

$$y_{n+1} = y_n + hf(t_{n+1}, y_{n+1}).$$

Esempio 6.2.1 *Applicare il metodo di Eulero esplicito per approssimare la soluzione del problema di Cauchy:*

$$y'(t) = e^{-y^2}, \qquad y(0) = 0$$

in $t = 0.1$.

Posto $h = 0.1$ si applica la formula con $n = 0$

$$y_1 = y_0 + hf(t_1, y_1), \qquad t_1 = t_0 + h = h$$

cosicchè risulti

$$y_1 \simeq y(0.1).$$

Sostituendo l'espressione della funzione l'approssimazione cercata è

$$y_1 = 0.1\ e^{-y_1^2}.$$

L'equazione, non lineare, può essere risolta solo utilizzando un metodo numerico, in quanto y_1 risulta essere lo zero della funzione

$$\varphi(x) = x - 0.1\ e^{-x^2}.$$

Osservando, per esempio, che risulta $\varphi(0) < 0$ e $\varphi(1) > 0$ si potrebbe applicare il metodo delle bisezioni, oppure un metodo iterativo di punto fisso

$$x_{k+1} = 0.1\ e^{-x_k^2}, \qquad x_0 = 0.$$

Un altro modo per derivare altri metodi numerici è quello di utilizzare le formule di quadratura descritte nel capitolo precedente. Infatti partendo dall'equazione differenziale

$$y'(t) = f(t, y(t)) \tag{6.6}$$

e supponendo di voler calcolare la funzione in t_{n+1} noto il suo valore in t_n, andiamo ad integrare membro a membro (6.6):

$$\int_{t_n}^{t_{n+1}} y'(t)dt = \int_{t_n}^{t_{n+1}} f(t, y(t))dt \tag{6.7}$$

cioè

$$y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} f(t, y(t))dt \tag{6.8}$$

quindi il problema equivale ad approssimare l'integrale a secondo membro. Un ulteriore formula per approssimare l'integrale in (6.7) è quello di usare l'area del trapezio avente come basi il valore della funzione $f(t, y)$ calcolato negli estremi dell'intervallo e come altezza lo stesso intervallo:

$$y(t_{n+1}) - y(t_n) \simeq \frac{h}{2} \left[f(t_{n+1}, y(t_{n+1})) + f(t_n, y(t_n)) \right]$$

che dà luogo al cosiddetto Metodo dei Trapezi:

$$y_{n+1} = y_n + \frac{h}{2} \left[f(t_{n+1}, y_{n+1}) + f(t_n, y_n) \right].$$

I due metodi appena descritti sono di tipo implicito, cioè l'approssimazione y_{n+1} dipende dal valore assunto dalla funzione $f(t, y)$ nell'incognita y_{n+1}. In questo caso è spesso necessario risolvere un'equazione non lineare (o un sistema di equazioni non lineari), che deve essere risolto numericamente.
I metodi descritti finora sono metodi ad un passo in quanto, per calcolare l'incognita y_{n+1} richiedono solo la conoscenza di y_n.
Un ulteriore metodo, applicabile all'intervallo $[t_n, t_{n+2}]$, consiste nell'approssimare l'integrale a secondo membro nell'equazione

$$y(t_{n+2}) - y(t_n) = \int_{t_n}^{t_{n+2}} f(t, y(t)) dt \tag{6.9}$$

con l'area del rettangolo avente come base l'intervallo $[t_n, t_{n+2}]$ e come altezza il valore assunto dalla funzione nel punto medio dello stesso intervallo:

$$y(t_{n+2}) - y(t_n) \simeq 2h f(t_{n+1}, y(t_{n+1}))$$

che fornisce il Metodo del Midpoint Esplicito:

$$y_{n+2} = y_n + 2h f(t_{n+1}, y_{n+1}).$$

Il metodo del midpoint esplicito è un metodo a due passi in quanto la soluzione nel punto t_{n+2}, che deve essere calcolata, dipende dalle approssimazioni in due punti precedenti, cioè y_n e y_{n+1}. È chiaro che in questo caso quando $n = 0$ si ottiene lo schema numerico

$$y_2 = y_0 + 2h f(t_1, y_1)$$

in cui il valore y_0 è noto in quanto coincide con la condizione iniziale, mentre il valore y_1 vien calcolato applicando un metodo ad un passo (metodi di

Eulero e dei Trapezi).
Per approssimare l'integrale (6.9) si potrebbe applicare anche la formula di Simpson

$$y(t_{n+2}) - y(t_n) \simeq \frac{h}{3}\left[f(t_n, y(t_n)) + 4f(t_{n+1}, y(t_{n+1})) + f(t_{n+2}, y(t_{n+2}))\right]$$

ottendendo appunto il cosiddetto metodo di Simpson:

$$y_{n+2} = y_n + \frac{h}{3}\left[f(t_n, y(t_n)) + 4f(t_{n+1}, y(t_{n+1})) + f(t_{n+2}, y(t_{n+2}))\right].$$

Anche il metodo di Simpson, come quello del Midpoint esplicito, e un metodo a due passi in quanto, per il calcolo di y_{n+2}, richiede la conoscenza delle approssimazioni y_n e y_{n+1}.

6.2.1 Accuratezza dei metodi numerici

Quando il passo di integrazione h tende a zero l'insieme di punti discreti $\{t_n\}$ diventa l'intero intervallo $[t_0, T]$. Una proprietà ovvia da richiedere ad un qualsiasi metodo numerico è che, quando $h \to 0$ la soluzione numerica y_n diventa la soluzione teorica $y(t)$, $t \in [t_0, T]$. Questa proprietà è detta Convergenza.

Definizione 6.2.1 *Un metodo numerico si dice convergente se, per ogni problemi ai valori iniziali soddisfacente le ipotesi si ha:*

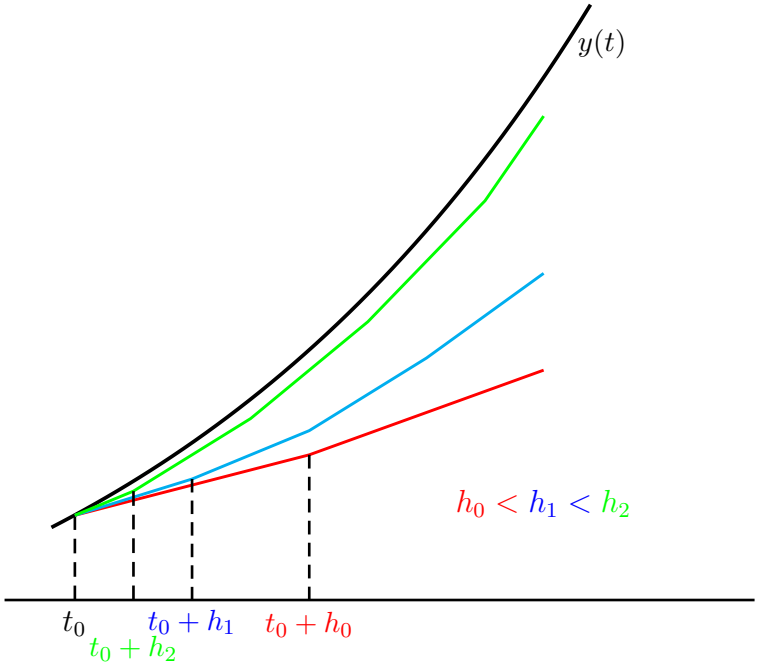
$$\lim_{\substack{h\to 0 \\ t=t_0+nh}} y_n = y(t)$$

per ogni $t \in [t_0, T]$. Un metodo che non è convergente si dice divergente.

Tale definizione necessita di alcuni chiarimenti. Consideriamo infatti un punto t della discretizzazione (cioè tale che $t = t_n = t_0 + nh$), un metodo convergente deve essere tale che la soluzione numerica y_n nel punto della discretizzazione $t = t_n$ tende a quella teorica $y(t)$ quando $h \to 0$. La definizione puntualizza l'esigenza che, anche se h tende a zero (e quindi $n \to \infty$), la quantità nh si mantiene costante all'ampiezza dell'intervallo $[t_0, t]$. Una definizione alternativa di convergenza richiede che

$$\lim_{h\to 0} \max_{0\le n\le N} |y(t_n) - y_n| = 0$$

quando il metodo numerico viene applicato ad un qualsiasi probema ai valori iniziali che soddisfa le ipotesi del Teorema 6.2.1. Nella seguente figura viene rappresentata tale proprietà.



Per i metodi che sono stati descritti nei paragrafi precedeti la convergenza è assicurata dal fatto che l'errore commesso nell'approssimazione (o della derivata prima della funzione o dell'integrale di $f(t, y)$) dipende da h (da potenze di h). Tuttavia la convergenza, da sola, non riesce a garantire, che, preso un valore del passo molto piccolo, la soluzione numerica sia molto vicina a quella teorica, a causa della presenza dei consuenti errori di rapprentazione dei dati e del condizionamento del problema differenziale.