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 dati 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 posizione

in cui 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 spinge a considerare come primo problema quello della rappresentazione di numeri reali è che ovviamente si deve sapere il livello di affidabilità dei risultati forniti dall'elaboratore. Infatti bisogna osservare che i numeri reali sono infiniti mentre la memoria di un calcolatore ha una capacità finita che ne rende impossibile la rappresentazione esatta. Una seconda osservazione consiste nel fatto che un numero reale ammette molteplici modi di rappresentazione. Per esempio scrivere

$$x = 123.47$$

è la rappresentazione, in forma convenzionale, dell'espressione

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

da cui, mettendo in evidenza 10^2:

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

mentre, mettendo in evidenza 10^3 lo stesso numero viene scritto come

$$x = 10^3 \times \left(1 \times 10^{-1} + 2 \times 10^{-2} + 3 \times 10^{-3} + 4 \times \times 10^{-4} + 7 \times 10^{-5}\right)$$

deducendo che ogni numero, senza una necessaria rappresentazione convenzionale, può essere scritto in infiniti modi. Il seguente teorema è fondamentale proprio per definire la rappresentazione dei numeri reali in una determinata base β.

Teorema 1.2.1 *Sia $\beta \in \mathbb{N}$, $\beta \geq 2$, 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 la terza ipotesi è essenziale 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, senza la terza ipotesi del Teorema 1.2.1, al numero 1 corrisponderebbero due differenti rappresentazioni in base.
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 $0.d_1d_2 \dots d_k \dots$ si dice mantissa. Il numero x viene normalmente rappresentato con la cosiddetta notazione posizionale $x = segno(x)(.d_1d_2d_3 \dots) \times \beta^p$, 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) e comunque la base è sempre un numero pari. 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 $\beta, t, m, M \in \mathbb{N}$ si definisce insieme dei numeri di macchina con rappresentazione normalizzata in base β con t cifre significative

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

dove
1. $t \geq 1, \ \beta \geq 2, \ m, M > 0$;
2. $d_i \in \{0, 1 \dots, \beta - 1\}$;
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 non ammette rappresentazione in base normalizzata.
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),

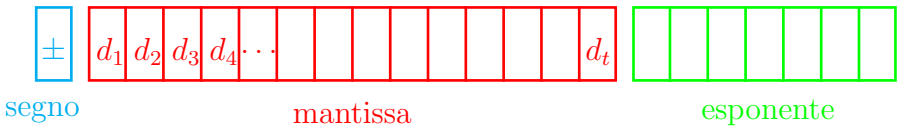


Figura 1.1: Locazione di memoria.

- un campo di memoria che può assumere $m + M + 1$ differenti configurazioni (e perciò può memorizzare i differenti valori p dell'esponente),
- 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

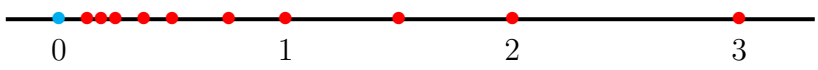


Figura 1.2: Elementi dell'insieme $\mathbb{F}(2,2,2,2)$.

esponente $-m$.

$$
\begin{aligned}
p=-2 \quad & x = 0.10 \times 2^{-2} = 2^{-1} \times 2^{-2} = 2^{-3} = 0.125; \\
& x = 0.11 \times 2^{-2} = (2^{-1} + 2^{-2}) \times 2^{-2} = 3/16 = 0.1875;
\end{aligned}
$$

$$
\begin{aligned}
p=-1 \quad & x = 0.10 \times 2^{-1} = 2^{-1} \times 2^{-1} = 2^{-2} = 0.25; \\
& x = 0.11 \times 2^{-1} = (2^{-1} + 2^{-2}) \times 2^{-1} = 3/8 = 0.375;
\end{aligned}
$$

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

$$
\begin{aligned}
p=1 \quad & x = 0.10 \times 2^{1} = 2^{-1} \times 2^{1} = 1; \\
& x = 0.11 \times 2^{1} = (2^{-1} + 2^{-2}) \times 2^{1} = 3/2 = 1.15;
\end{aligned}
$$

$$
\begin{aligned}
p=2 \quad & x = 0.10 \times 2^{2} = 2^{-1} \times 2^{2} = 2; \\
& x = 0.11 \times 2^{2} = (2^{-1} + 2^{-2}) \times 2^{2} = 3;
\end{aligned}
$$

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 l'esponente 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 $n > t$ 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

$$x = 0.ddddddddd \cdots \times \beta^M$$

con $d = \beta - 1$, 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 $\tilde{x} = \mathrm{tr}(x)$, con arrotondamento se $\tilde{x} = \mathrm{arr}(x)$. 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

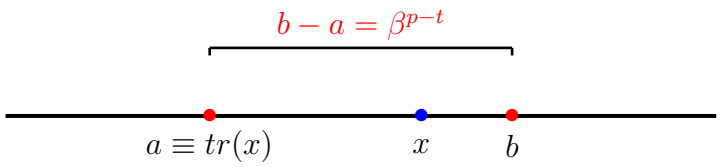


Figura 1.3: Stima dell'errore di rappresentazione nel caso di troncamento.

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

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

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:

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

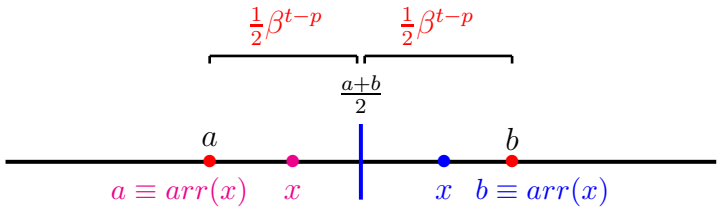


Figura 1.4: Stima dell'errore di rappresentazione nel caso di arrotondamento.

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

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*

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

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

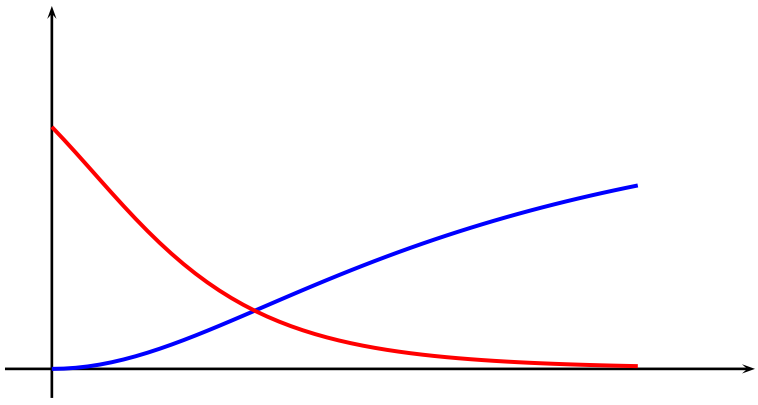
Nei successivi paragrafi saranno descritti alcuni metodi numerici per il calcolo approssimato delle radici di un'equazione non lineare. Tali metodi numerici sono di tipo iterativo, ovvero consistono nel definire una successione (o più successioni), che, a partire da un'assegnata approssimazione iniziale (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

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

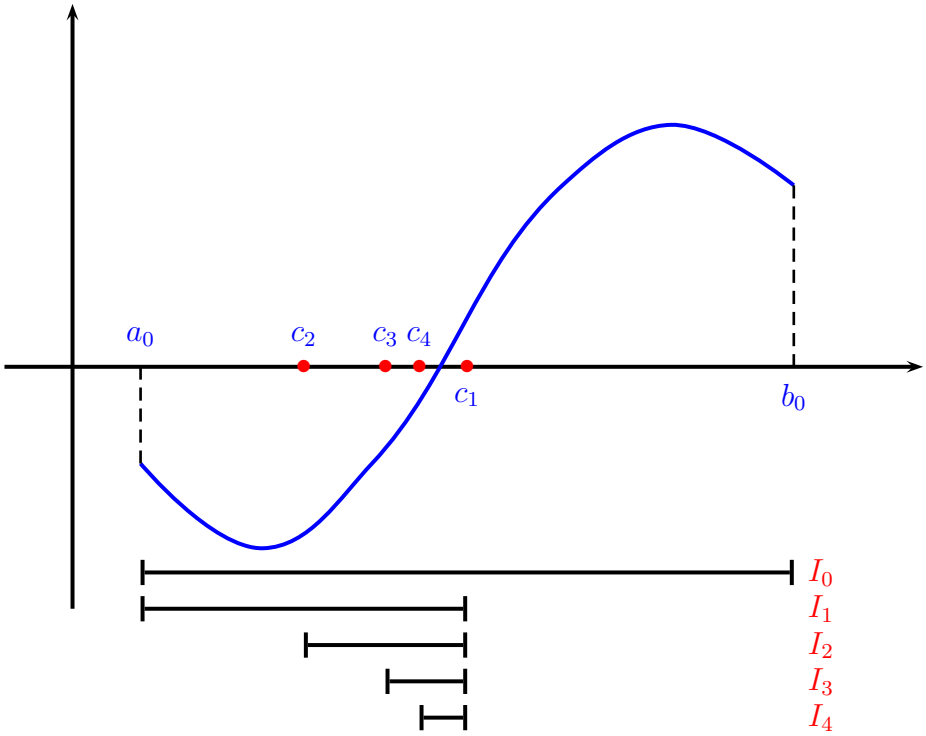
Se $f(c_{k+1}) = 0$ allora $\alpha = c_{k+1}$, altrimenti si pone:

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

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

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

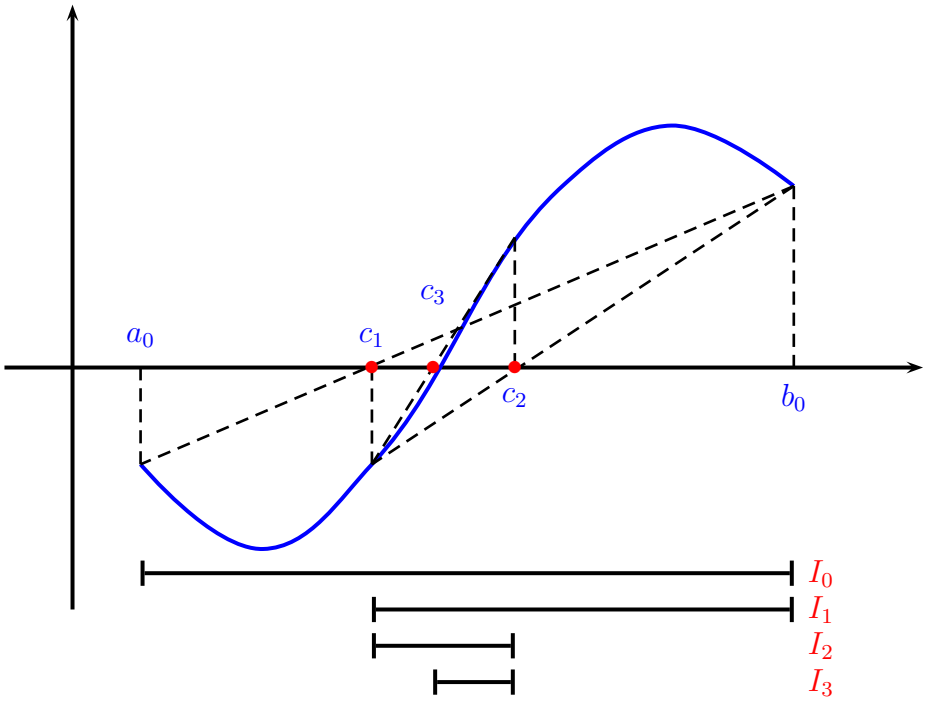
Ad un generico passo k si calcola

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

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.



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.

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

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

$$|g'(x)| < 1, \qquad \text{per ogni } x \in [\alpha-\rho, \alpha+\rho]$$

allora valgono le seguenti asserzioni:

1. *se $x_0 \in [\alpha-\rho, \alpha+\rho]$ allora anche $x_k \in [\alpha-\rho, \alpha+\rho]$ per ogni k;*
2. *la successione $\{x_k\}$ converge ad α;*
3. *α è l'unico punto fisso di $g(x)$ nell'intervallo $[\alpha-\rho, \alpha+\rho]$.*

Dimostrazione. Sia

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

Innanzitutto dimostriamo 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 $|\xi_k - \alpha| < |x_k - \alpha| \leq \rho$ e l'ultima uguaglianza segue dall'applicazione del teorema di Lagrange. 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:

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

Per dimostrare l'unicità del punto ragioniamo per assurdo che supponiamo ch i punti fissi sono due, $\alpha, \beta \in [\alpha-\rho, \alpha+\rho]$. Allora

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

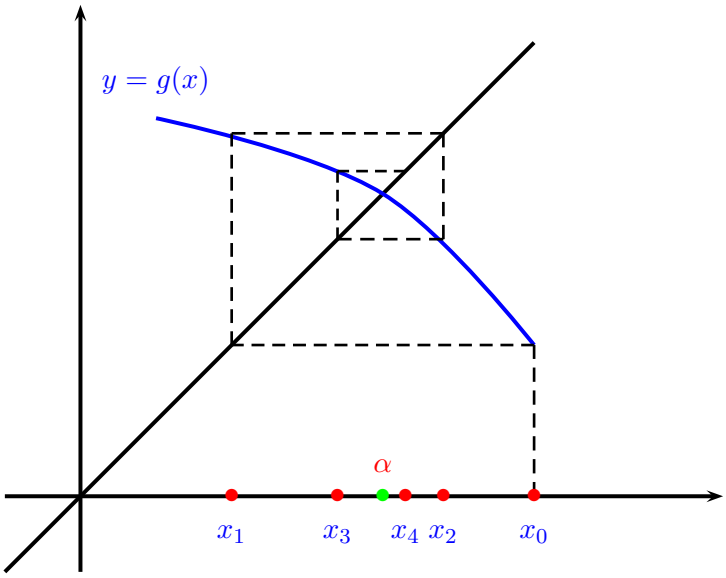


Figura 2.1: Interpretazione geometrica del processo $x_{k+1} = g(x_k)$, se $-1 < g'(\alpha) \leq 0$.

con $\xi \in [\alpha - \rho, \alpha + \rho]$. Poichè $|g'(\xi)| < 1$ si ha

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

e ciò è assurdo. □
Nelle figure 2.2 e 2.1 è 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.*

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

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

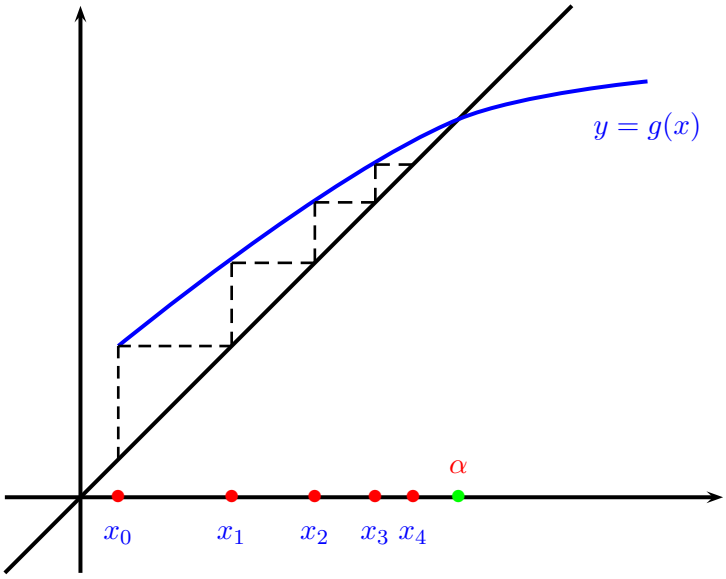


Figura 2.2: Interpretazione geometrica del processo $x_{k+1} = g(x_k)$, se $0 \leq g'(\alpha) < 1$.

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.
In generale l'ordine di convergenza è un numero reale maggiore o uguale a 1. Tuttavia per i metodi di iterazione funzionale di tipo (2.5) è un numero intero per il quale vale il seguente teorema.

Teorema 2.4.3 *Sia $\{x_k\}_{k=0}^{\infty}$ una successione generata dallo schema (2.5) convergente ad α, punto fisso di $g(x)$, funzione sufficientemente derivabile in un intorno di α. La successione ha ordine di convergenza $p \geq 1$ se e solo se*

$$g'(\alpha) = g''(\alpha) = \cdots = g^{(p-1)}(\alpha) = 0, \qquad g^{(p)}(\alpha) \neq 0. \tag{2.11}$$

Dimostrazione. Scriviamo lo sviluppo in serie di Taylor della funzione $g(x)$ in x_k prendendo come punto iniziale α:

$$\begin{aligned} g(x_k) &= g(\alpha) + g'(\alpha)(x_k - \alpha) + \frac{g''(\alpha)}{2!}(x_k - \alpha)^2 + \dots \\ &\quad \dots + \frac{g^{(p-1)}(\alpha)}{(p-1)!}(x_k - \alpha)^{p-1} + \frac{g^{(p)}(\xi_k)}{p!}(x_k - \alpha)^p. \end{aligned}$$

Sostituendo a $g(x_k)$ il valore x_{k+1} e sfruttando l'ipotesi che α è punto fisso di $g(x)$ risulta

$$\begin{aligned} x_{k+1} - \alpha &= g'(\alpha)(x_k - \alpha) + \frac{g''(\alpha)}{2!}(x_k - \alpha)^2 + \dots \\ &\quad \dots + \frac{g^{(p-1)}(\alpha)}{(p-1)!}(x_k - \alpha)^{p-1} + \frac{g^{(p)}(\xi_k)}{p!}(x_k - \alpha)^p \end{aligned}$$

dove ξ è compreso tra x_k e α. Quindi se vale l'ipotesi (2.11) e passando ai moduli risulta

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

e quindi

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

Viceversa supponiamo per ipotesi che la successione ha ordine di convergenza p e dimostriamo che

$$g'(\alpha) = g''(\alpha) = \dots = g^{(p-1)}(\alpha) = 0.$$

Ipotizziamo, per assurdo, che esista una derivata di ordine i, $i < p$, diversa da zero, ovvero

$$g^{(i)}(\alpha) \neq 0.$$

Scriviamo lo sviluppo in serie di Taylor di $x_{k+1} = g(x_k)$:

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

da cui

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

Passando ai moduli e calcolando il limite della successione si ottiene:

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

da cui segue che la successione ha ordine $i < p$ in contrasto con l'ipotesi fatta. □

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

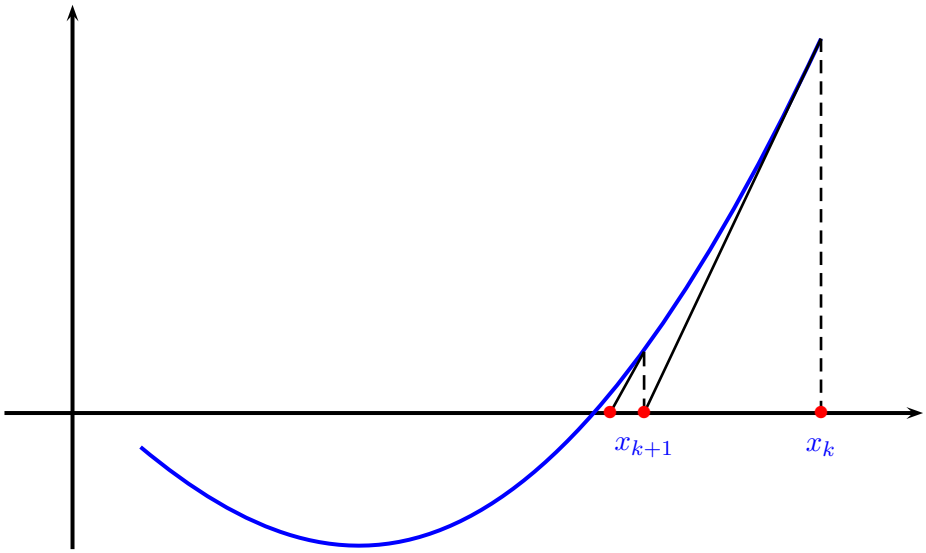
$$g'(\alpha) = g''(\alpha) = \cdots = g^{(q)}(\alpha) = 0,$$

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 la convergenza el'ordine del metodo di Newton-Raphson vale il seguente teorema.

Teorema 2.4.4 *Sia $f \in \mathcal{C}^3([a,b])$, tale che $f'(x) \neq 0$, per $x \in [a,b]$, dove $[a,b]$ è un opportuno intervallo contenente α, allora valgono le seguenti proposizioni:*

1. *esiste un intervallo $[\alpha - \rho, \alpha + \rho]$, tale che, scelto x_0 appartenente a tale intervallo, la successione definita dal metodo di Newton-Raphson è convergente ad α;*

2. *la convergenza è di ordine $p \geq 2$.*

Dimostrazione. Per valutare la convergenza del metodo calcoliamo la derivata prima della funzione iteratrice:

$$g'(x) = 1 - \frac{[f'(x)]^2 - f(x)f''(x)}{[f'(x)]^2} = \frac{f(x)f''(x)}{[f'(x)]^2}.$$

Poichè $f'(\alpha) \neq 0$ risulta:

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

quindi, fissato un numero positivo $\kappa < 1$, esiste $\rho > 0$ tale che per ogni $x \in [\alpha - \rho, \alpha + \rho]$ si ha $|g'(x)| < \kappa$ e quindi vale il teorema di convergenza 2.4.2.
Per dimostrare la seconda parte del teorema si deve calcolare la derivata seconda di $g(x)$:

$$g''(x) = \frac{[f'(x)f''(x) + f(x)f'''(x)][f'(x)]^2 - 2f(x)f'(x)[f''(x)]^2}{[f'(x)]^4}.$$

Calcolando la derivata seconda in $x = \alpha$ risulta

$$g''(\alpha) = \frac{f''(\alpha)}{f'(\alpha)} \tag{2.14}$$

ne segue che se $f''(\alpha) \neq 0$ allora anche $g''(\alpha) \neq 0$ e quindi, applicando il Teorema 2.4.3, l'ordine $p = 2$. Se invece $f''(\alpha) = 0$ allora l'ordine è almeno pari a 3. Dalla relazione 2.14 segue inoltre che la costante asintotica di convergenza vale

$$\gamma = \frac{1}{2}\left|\frac{f''(\alpha)}{f'(\alpha)}\right|. \ \square$$

Il Teorema 2.4.4 vale nell'ipotesi in cui $f'(\alpha) \neq 0$, cioè se α è una radice semplice di $f(x)$. Se invece la radice α ha molteplicità $r > 1$ l'ordine di convergenza del metodo non è più 2. In questo caso infatti si può porre

$$f(x) = q(x)(x - \alpha)^r, \qquad q(\alpha) \neq 0,$$

quindi riscrivendo la funzione iteratrice del metodo di Newton-Raphson risulta

$$g(x) = x - \frac{q(x)(x - \alpha)}{rq(x) + q'(x)(x - \alpha)},$$

da cui, dopo una serie di calcoli, risulta

$$g'(\alpha) = 1 - \frac{1}{r}. \tag{2.15}$$

Pertanto, poichè $r > 1$ risulta $|g'(x)| < 1$ e quindi per il Teorema 2.4.2 il metodo è ancora convergente ma, applicando il Teorema 2.4.3 l'ordine di convergenza è 1.
Se si conosce la molteplicità della radice si può modificare il metodo di Newton-Raphson ottenendo uno schema numerico con ordine 2. Ponendo

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

si definisce un metodo con la seguente funzione iteratrice

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

da cui segue, tenendo conto della (2.15), che

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

Esempio 2.4.1 *Approssimare il numero* $\alpha = \sqrt[m]{c}$ *con* $m \in \mathbb{R}$, $m \geq 2$, $c > 0$.

Il numero α cercato è lo zero della funzione

$$f(x) = x^m - c.$$

Poichè per $x > 0$ la funzione risulta essere monotona allora è sufficiente scegliere un qualsiasi $x_0 > 0$ per ottenere una successione convergente alla radice $m-$esima di c. Il metodo di Newton-Raphson fornisce la formula

$$x_{k+1} = x_k - \frac{x_k^m - c}{m x_k^{m-1}} = \frac{1}{m}\left[(m-1)x_k + c x_k^{1-m}\right], \quad k = 0, 1, 2, \ldots.$$

Per $m = 2$ lo schema diviene

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{c}{x_k}\right),$$

che è la cosiddetta formula di Erone per il calcolo della radice quadrata, nota già agli antichi Greci.
Considerando come esempio $m = 4$ e $c = 3$, 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, si ottiene il processo iterativo

$$x_{k+1} = x_k - \frac{1}{3}\left(2x_k + 3x_k^{-3}\right).$$

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

$$x_{k+1} = x_k - \frac{f(x_k)}{M} \tag{2.16}$$

anzichè la (2.12). La (2.16) 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

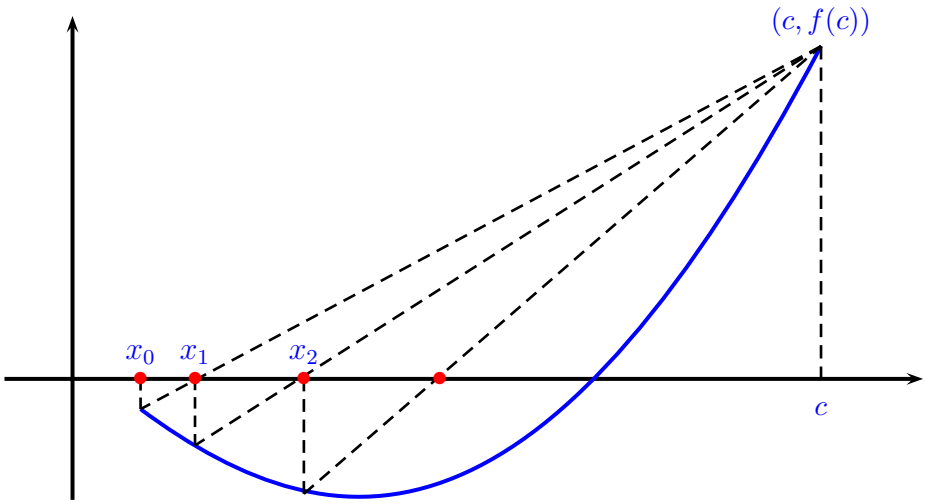
Il metodo della secante è definito dalla relazione

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

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 $f(x)$.

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:

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

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

$$x_i = \frac{\det A_i}{\det A} \qquad i = 1, \ldots, n, \tag{3.3}$$

dove A_i è la matrice ottenuta da A sostituendo la sua i-esima colonna con il termine noto $\mathbf{b}$. Dalla (3.3) è evidente che per ottenere tutte le componenti

del vettore soluzione è necessario il calcolo di $n+1$ determinanti di ordine n. Si può facilmente dedurre che il numero di operazioni necessarie per il calcolo del determinate di una matrice di ordine n applicando la regola di Laplace è circa $n!$, quindi questa strada non permette di poter determinare velocemente la soluzione del nostro sistema. Basti pensare che se $n = 100$ il il numero di operazioni per il calcolo di un solo determinante sarebbe all'incirca dell'ordine di 10^{157}.

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:

$$\begin{array}{ccccccccc} a_{11}x_1 & +a_{12}x_2 & \dots & +a_{1i}x_i & \dots & +a_{1n}x_n & = b_1 \\ & a_{22}x_2 & \dots & +a_{2i}x_i & \dots & +a_{2n}x_n & = b_2 \\ & & \ddots & \vdots & & \vdots & \vdots \\ & & & a_{ii}x_i & \dots & +a_{in}x_n & = b_i \\ & & & & \ddots & \vdots & \vdots \\ & & & & & a_{nn}x_n & = b_n. \end{array} \tag{3.4}$$

In questo caso la matrice A è detta triangolare superiore. Il determinante di una matrice di questo tipo è uguale al prodotto degli elementi diagonali pertanto la matrice è non singolare se risulta $a_{ii} \neq 0$ per ogni i. In questo caso, la soluzione è facilmente calcolabile infatti è sufficiente osservare che nell'ultima equazione compare solo un'incognita che può essere calcolata e che procedendo a ritroso da ogni equazione può essere ricavata un'incognita poichè le successive sono già state calcolate. Il metodo può essere riassunto nelle seguenti formule:

$$\begin{cases} x_n = \dfrac{b_n}{a_{nn}} \\ \\ x_i = \dfrac{b_i - \displaystyle\sum_{j=i+1}^{n} a_{ij}x_j}{a_{ii}} \qquad i = n-1, \dots, 1. \end{cases} \tag{3.5}$$

Il metodo (3.5) 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.

$$\begin{array}{lllllll} a_{11}x_1 & & & & & & = b_1 \\ a_{21}x_1 & +a_{22}x_2 & & & & & = b_2 \\ \vdots & \vdots & \ddots & & & & \vdots \\ a_{i1}x_1 & +a_{i2}x_2 & \dots & +a_{ii}x_i & & & = b_i \\ \vdots & \vdots & & & \ddots & & \vdots \\ a_{n1}x_1 & +a_{n2}x_2 & \dots & +a_{ni}x_i & \dots & +a_{nn}x_n & = b_n \end{array} \tag{3.6}$$

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 su dati reali. Considerano per esempio il metodo di sostituzione in avanti. Per calcolare x_1 è necessaria una sola operazione (una divisione), per calcolare x_2 le operazioni sono tre (un prodotto, una somma algebrica e una divisione), mentre il generico x_i richiede $2i-1$ operazioni ($i-1$ prodotti, $i-1$ somme algebriche 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

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

Il costo computazionale viene sempre valutato in funzione di un determinato parametro (il numero assoluto in sè non avrebbe alcun significato) che, in questo caso è la dimensione del sistema. In questo modo è possibile prevedere il tempo necessario per calcolare la soluzione del problema.

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[1] 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 la combinazione lineare di un'altra equazione 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.

[1] Due sistemi si dicono equivalenti se ammettono lo stesso insieme di soluzioni, quindi nel nostro caso la stessa soluzione. Osserviamo che se $\mathbf{x}^*$ è un vettore tale che $A\mathbf{x}^* = \mathbf{b}$ e B è una matrice non singolare allora $BA\mathbf{x}^* = B\mathbf{b}$; viceversa se $BA\mathbf{x}^* = B\mathbf{b}$ e B è non singolare allora $B^{-1}BA\mathbf{x}^* = B^{-1}B\mathbf{b}$ e quindi $A\mathbf{x}^* = \mathbf{b}$. Dunque se B è non singolare i sistemi $A\mathbf{x} = \mathbf{b}$ e $BA\mathbf{x} = B\mathbf{b}$ sono equivalenti.

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}:

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

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:

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

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:

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

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:

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

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:

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

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:

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

Cerchiamo ora di ricavare le formule di trasformazione del metodo di eliminazione di Gauss per rendere un generico sistema di ordine n in forma triangolare superiore.
Consideriamo il sistema di equazioni nella sua forma scalare (3.2):

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \qquad i = 1, \ldots, n. \tag{3.7}$$

Poichè il procedimento richiede un certo numero di passi indichiamo con $a_{ij}^{(1)}$ e $b_i^{(1)}$ gli elementi della matrice dei coefficienti e del vettore dei termini noti del sistema di partenza. Isoliamo in ogni equazione la componente x_1. Abbiamo:

$$a_{11}^{(1)} x_1 + \sum_{j=2}^{n} a_{1j}^{(1)} x_j = b_1^{(1)} \tag{3.8}$$

$$a_{i1}^{(1)} x_1 + \sum_{j=2}^{n} a_{ij}^{(1)} x_j = b_i^{(1)}, \qquad i = 2, \ldots, n. \tag{3.9}$$

Moltiplicando l'equazione (3.8) per $-a_{i1}^{(1)}/a_{11}^{(1)}, i = 2, \ldots, n$, si ottengono le seguenti $n-1$ equazioni:

$$-a_{i1}^{(1)} x_1 + \sum_{j=2}^{n} \left(-\frac{a_{i1}^{(1)}}{a_{11}^{(1)}} a_{1j}^{(1)} \right) x_j = -\frac{a_{i1}^{(1)}}{a_{11}^{(1)}} b_i^{(1)}, \qquad i = 2, \ldots, n. \tag{3.10}$$

Sommando alle equazioni (3.9) le (3.10) si ricavano $n-1$ nuove equazioni:

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

L'equazione (3.8) insieme alle (3.11) formano un nuovo sistema di equazioni, equivalente a quello originario, che possiamo scrivere nel seguente modo:

$$\begin{cases} a_{11}^{(1)}x_1+\displaystyle\sum_{j=2}^{n}a_{1j}^{(1)}x_j=b_1^{(1)} \\ \\ \displaystyle\sum_{j=2}^{n}a_{ij}^{(2)}x_j=b_i^{(2)} \qquad i=2,\ldots,n \end{cases} \tag{3.12}$$

dove

$$\begin{cases} a_{ij}^{(2)}= \quad a_{ij}^{(1)}-\dfrac{a_{i1}^{(1)}}{a_{11}^{(1)}}a_{1j}^{(1)} \qquad i,j=2,\ldots,n \\ \\ b_i^{(2)}= \quad b_i^{(1)}-\dfrac{a_{i1}^{(1)}}{a_{11}^{(1)}}b_1^{(1)} \qquad i=2,\ldots,n. \end{cases} \tag{3.13}$$

Osserviamo che la matrice dei coefficienti del sistema (3.12) è la seguente

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

Ora a partire dal sistema di equazioni

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

ripetiamo i passi fatti precedentemente:

$$a_{22}^{(2)}x_2+\sum_{j=3}^{n}a_{2j}^{(2)}x_j=b_2^{(2)} \tag{3.14}$$

$$a_{i2}^{(2)} x_2 + \sum_{j=3}^{n} a_{ij}^{(2)} x_j = b_i^{(2)}, \qquad i = 3, \ldots, n. \tag{3.15}$$

Moltiplicando l'equazione (3.14) per $-a_{i2}^{(2)}/a_{22}^{(2)}$, per $i = 3, \ldots, n$, si ottiene

$$a_{i2}^{(2)} x_2 + \sum_{j=3}^{n} \left(-\frac{a_{i2}^{(2)}}{a_{22}^{(2)}} a_{2j}^{(2)} \right) x_j = -\frac{a_{i2}^{(2)}}{a_{22}^{(2)}} b_2^{(2)}, \qquad i = 3, \ldots, n. \tag{3.16}$$

Sommando le equazioni (3.16) alle (3.15) si ottengono $n-2$ nuove equazioni:

$$\sum_{j=3}^{n} \left(a_{ij}^{(2)} - \frac{a_{i2}^{(2)}}{a_{22}^{(2)}} a_{2j}^{(2)} \right) x_j = b_i^{(2)} - \frac{a_{i2}^{(2)}}{a_{22}^{(2)}} b_2^{(2)}, \qquad i = 3, \ldots, n \tag{3.17}$$

che possiamo scrivere in forma più compatta:

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

dove

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

Abbiamo il nuovo sistema equivalente:

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

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:

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)} \qquad i,j = k+1,\dots,n \tag{3.18}$$

e

$$b_i^{(k+1)} = b_i^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} b_k^{(k)} \qquad i = k+1,\dots,n. \tag{3.19}$$

Soffermiamoci ora un momento sul primo passo del procedimento. Osserviamo che per ottenere il 1^o 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

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

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

$[4^a$ `riga di` $A^{(3)}$ `+` $]$	0	0	$1/3$	2	$+$
$[(-1/6)\times$ 3^a `riga di` $A^{(3)}]$	0	0	$-1/3$	$-1/3$	$=$
$[4^a$ `riga di` $A^{(4)}]$	0	0	0	$5/3$	

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

$$A\mathbf{x}_i = \mathbf{e}_i \tag{3.20}$$

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

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

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 (3.20), 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 (3.18) e (3.19)) che gli elementi pivotali $a_{kk}^{(k)}$ siano non nulli per ogni k. Tale situazione si verifica quando i minori principali di testa di ordine di A sono diversi da zero. Infatti vale il seguente risultato.

Teorema 3.3.1 *Se $A \in \mathbb{R}^{n\times n}$, indicata con A_k la matrice principale di testa di ordine k, risulta*

$$a_{kk}^{(k)} = \frac{\det A_k}{\det A_{k-1}}, \qquad k = 1, \ldots, n$$

avendo posto per convenzione $\det A_0 = 1$.

In pratica 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

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

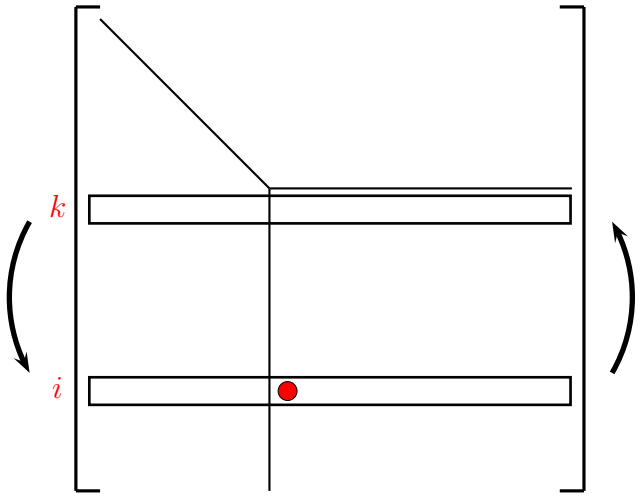


Figura 3.1: Strategia di pivoting parziale.

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 l'equazione 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 le equazioni del sistema di indice r e k.

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

1. determinare gli indici r, s tali che
$$|a_{rs}^{(k)}| = \max_{k \leq i,j \leq n} |a_{ij}^{(k)}|;$$

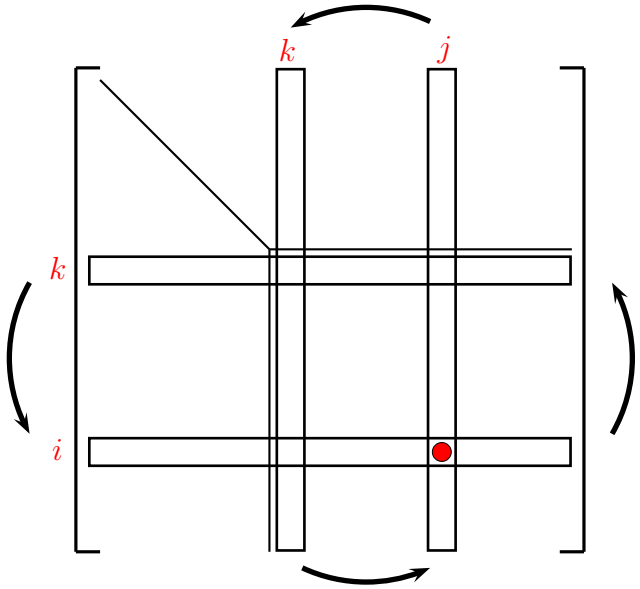


Figura 3.2: Strategia di pivoting totale.

2. effettuare lo scambio tra le equazioni del sistema di indice r e k.

3. effettuare lo scambio tra le colonne di indice s e k della matrice dei coefficienti.

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 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.

3.3.3 La Fattorizzazione *LU*

Introduzione

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

$$A^{(n)}\mathbf{x} = \mathbf{b}^{(n)}. \tag{3.21}$$

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

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

L'algoritmo che sarà descritto in questo paragrafo consentirà di evitare l'eventualità di dover rifare tutti i calcoli (o una parte di questi).

Fattorizzazione *LU*

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 quindi 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 quindi 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

$$a_{ij} = \sum_{k=1}^{n} l_{ik}u_{kj} = \sum_{k=1}^{\min(i,j)} l_{ik}u_{kj} \qquad i,j = 1,\dots,n. \tag{3.23}$$

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

$$a_{ij} = \sum_{k=1}^{i} l_{ik}u_{kj} \qquad j \geq i \tag{3.24}$$

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

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

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:

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

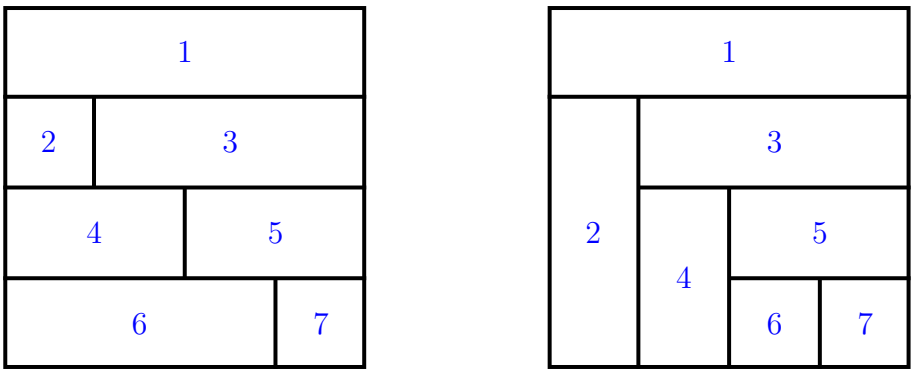
ovvero

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

da cui

$$l_{ij} = \frac{1}{u_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj} \right) \qquad i > j. \tag{3.27}$$

Si osservi che le formule (3.25) e (3.27) 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.

Esempio 3.3.3 *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 (3.25) 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 (3.27) 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 (3.25) 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 (3.27) 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 quindi

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

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

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

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.*

Capitolo 3

Interpolazione di Funzioni

3.1 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:

$$f(x) \simeq g(x; a_0, a_1, \ldots, a_n). \tag{3.1}$$

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 (3.1) 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

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

Vedremo nel successivo paragrafo di dare una risposta al nostro problema nel caso in cui si cerchino le funzioni $\Phi_i(x)$ di tipo polinomiale.

3.2 Il Polinomio Interpolante di Lagrange

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

$$L_n(x) = \sum_{k=0}^{n} l_{nk}(x) f(x_k) \tag{3.2}$$

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

$$l_{nk}(x_i) = \begin{cases} 0 & \text{se } k \neq i \\ 1 & \text{se } k = i \end{cases} \tag{3.3}$$

allora esse sono soddisfatte. In particolare la prima condizione di (3.3) indica che $l_{nk}(x)$ si annulla negli n nodi $x_0, x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n$ e quindi 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 (3.3)

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

si trova immediatamente:

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

In definitiva il polinomio interpolante ha la seguente forma:

$$L_n(x) = \sum_{k=0}^{n} \left(\prod_{i=0, i\neq k}^{n} \frac{x - x_i}{x_k - x_i} \right) f(x_k). \tag{3.4}$$

Il polinomio (3.4) 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.*

3.2.1 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 quindi

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

cioè

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

e in definitiva

$$e(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\omega_{n+1}(x). \tag{3.5}$$

Esempio 3.2.1 *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 $L_4(x)$:

$$L_4(x) = l_{4,0}(x)y_0 + l_{4,1}(x)y_1 + l_{4,2}(x)y_2 + l_{4,3}(x)y_3 + l_{4,4}(x)y_4 \tag{3.6}$$

e calcoliamo i 5 polinomi fondamentali di Lagrange:

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

Sostituendo in (3.6) 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

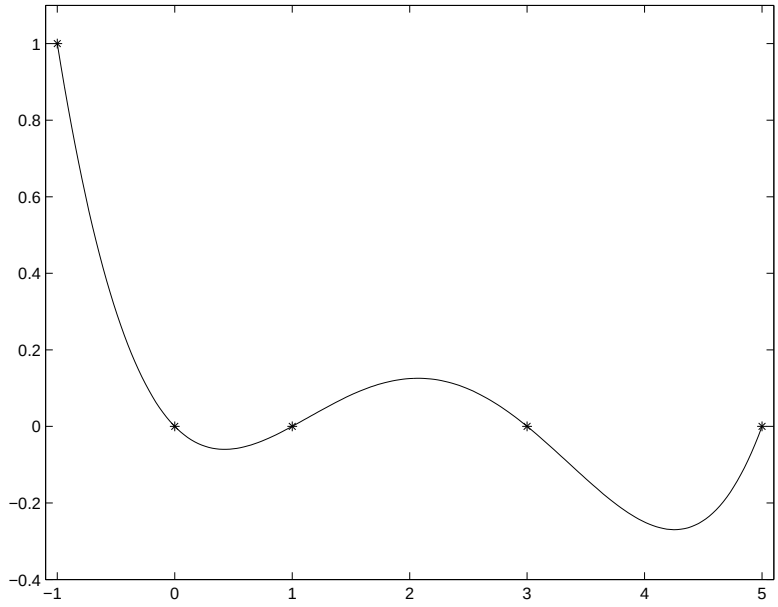


Figura 3.1: Grafico del polinomio $l_{40}(x)$.

polinomio interpolante $L_4(2)$.
Nelle figure 3.1-3.5 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 3.6 è tracciato il grafico del polinomio interpolante di Lagrange, i cerchi evidenziano ancora una volta i punti di interpolazione.

3.2.2 Il fenomeno di Runge

Nell'espressione dell'errore è presente, al denominatore, il fattore $(n+1)!$, che potrebbe indurre a ritenere che, interpolando la funzione con un elevato numero di nodi, l'errore tenda a zero e quindi che il polinomio interpolante tenda alla funzione $f(x)$. 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

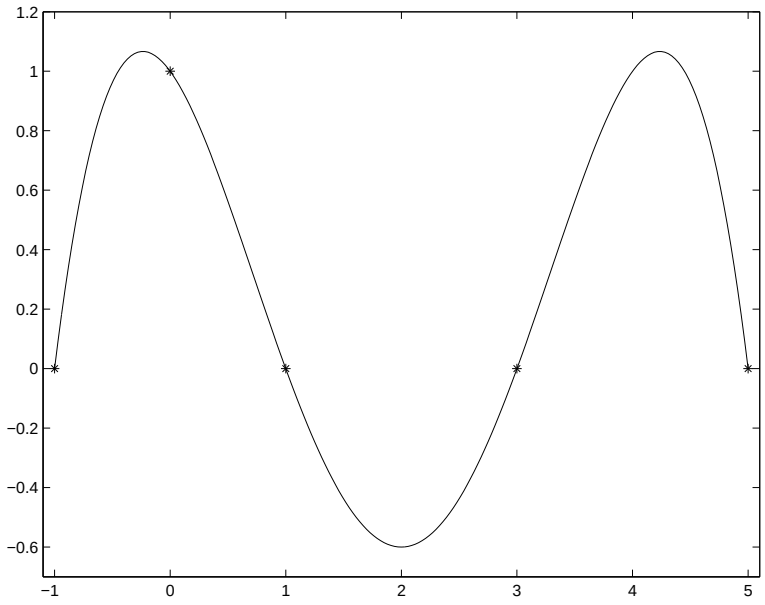


Figura 3.2: Grafico del polinomio $l_{41}(x)$.

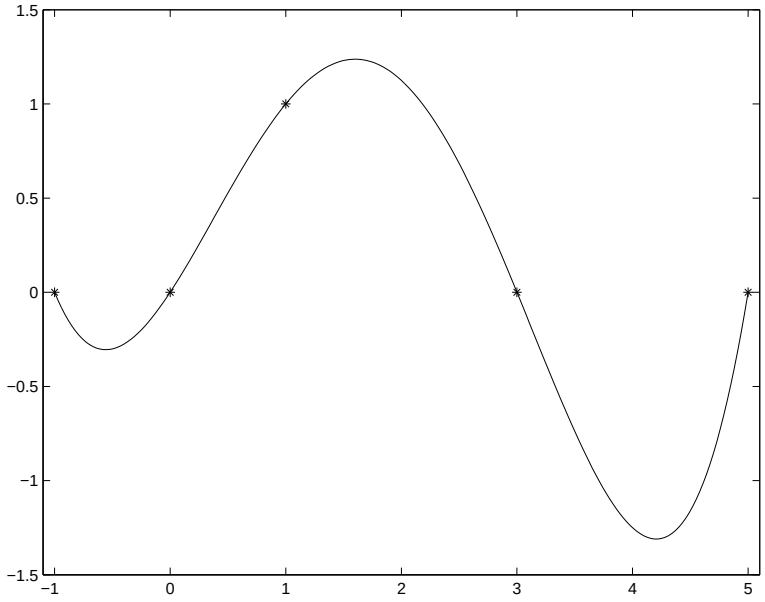


Figura 3.3: Grafico del polinomio $l_{42}(x)$.

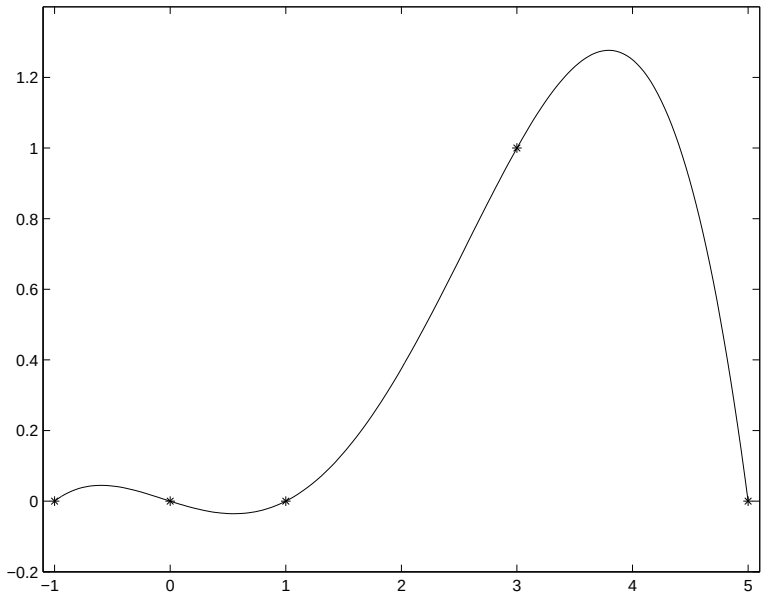


Figura 3.4: Grafico del polinomio $l_{43}(x)$.

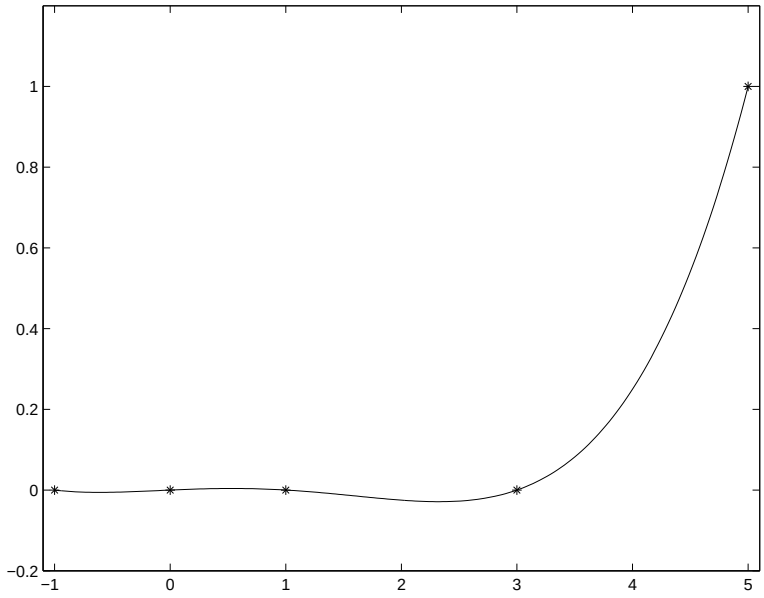


Figura 3.5: Grafico del polinomio $l_{44}(x)$.

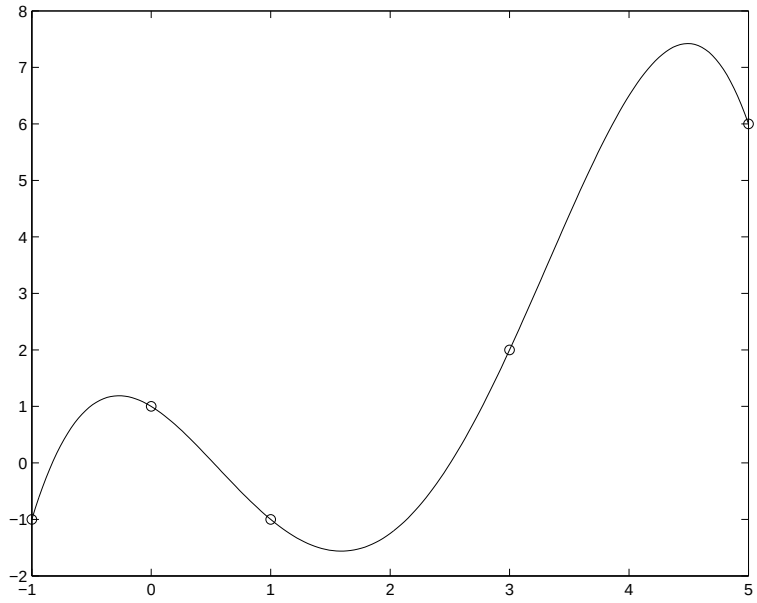


Figura 3.6: Grafico del polinomio interpolante di Lagrange $L_4(x)$.

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.

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).

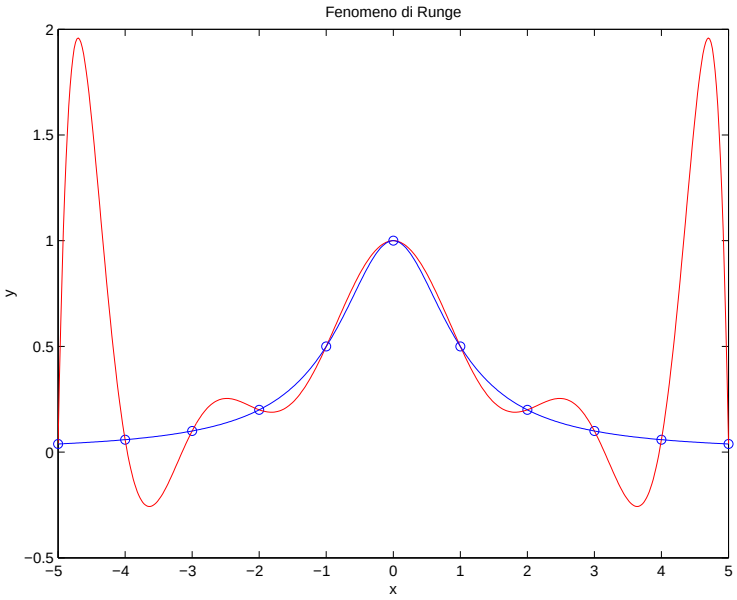


Figura 3.7: Il fenomeno di Runge.

Capitolo 4

Formule di Quadratura

4.1 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

$$\sum_{k=0}^{n} w_k f(x_k) \tag{4.1}$$

che approssimi realizzare $I(f)$.
Formule di tipo (4.1) 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 (4.1) 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

$$w_k = \int_a^b l_{nk}(x)dx \qquad k = 0, 1, \ldots, n \tag{4.2}$$

e

$$R_{n+1}(f) = \int_a^b e(x)dx \tag{4.3}$$

otteniamo

$$I(f) \simeq \sum_{k=0}^{n} w_k f(x_k)$$

con un errore stabilito dalla relazione (4.3). Le formule di quadratura con pesi definiti dalle formule (4.2) 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 4.1.1 *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

$$R_{n+1}(f) = \int_a^b \omega_{n+1}(x)\frac{p_n^{(n+1)}(x)}{(n+1)!}dx \equiv 0$$

e quindi la formula fornisce il risultato esatto dell'integrale, quindi $q \geq n$.

4.2 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.

4.2.1 Formula dei Trapezi

Siano $x_0 = a$, $x_1 = b$ e $h = b - a$.

$$T_2 = w_0 f(x_0) + w_1 f(x_1)$$

$$w_0 = \int_a^b l_{1,0}(x)dx = \int_a^b \frac{x - x_1}{x_0 - x_1}dx = \int_a^b \frac{x - b}{a - b}dx$$

$$= \frac{1}{a - b}\left[\frac{(x - b)^2}{2}\right]_{x=a}^{x=b} = \frac{h}{2}.$$

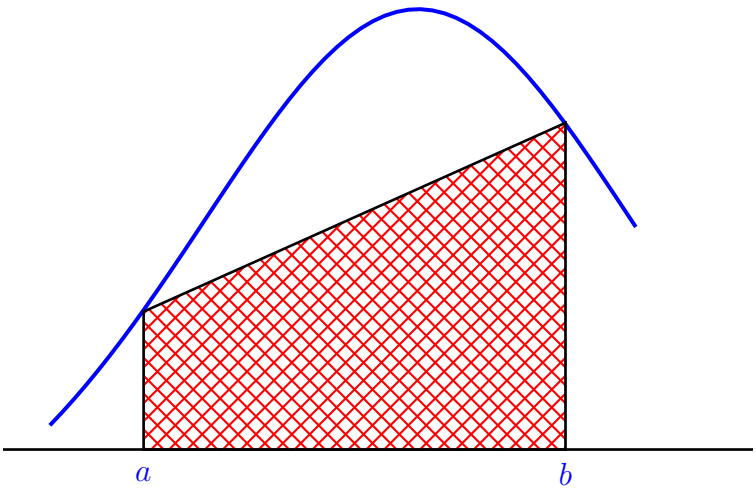
Poichè i nodi scelti sono simmetrici rispetto al punto medio $c = (a + b)/2$ è

$$w_1 = w_0 = \frac{h}{2}.$$

Otteniamo dunque la formula

$$T_2 = \frac{h}{2}\left[f(a) + f(b)\right]$$

che viene detta Formula dei Trapezi. 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 $[a, b]$).



Per quello che 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 4.2.1 *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

$$R_2(f) = \frac{1}{2}f''(\eta)h^3\int_0^1 t(t-1)dt = -\frac{1}{12}h^3 f''(\eta).$$

4.2.2 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

$$\begin{aligned} w_0 &= \int_a^b \frac{(x-c)(x-b)}{(a-c)(a-b)}dx = \int_{-1}^1 \frac{hth(t-1)}{(-h)(-2h)}h\ dt \\ &= \frac{h}{2}\int_{-1}^1 (t^2-t)dt = \frac{h}{2}\int_{-1}^1 t^2 dt = \frac{h}{2}\left[\frac{t^3}{3}\right]_{-1}^1 = \frac{h}{3}. \end{aligned}$$

Per la 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

$$R_3(f) = -h^5 \frac{f^{(4)}(\sigma)}{90} \qquad \eta, \sigma \in (a, b),$$

che assicura che la formula ha grado di precisione 3.

4.3 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,

$$h = \frac{b-a}{N} \tag{4.4}$$

sicchè

$$\int_a^b f(x)dx = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} f(x)dx$$

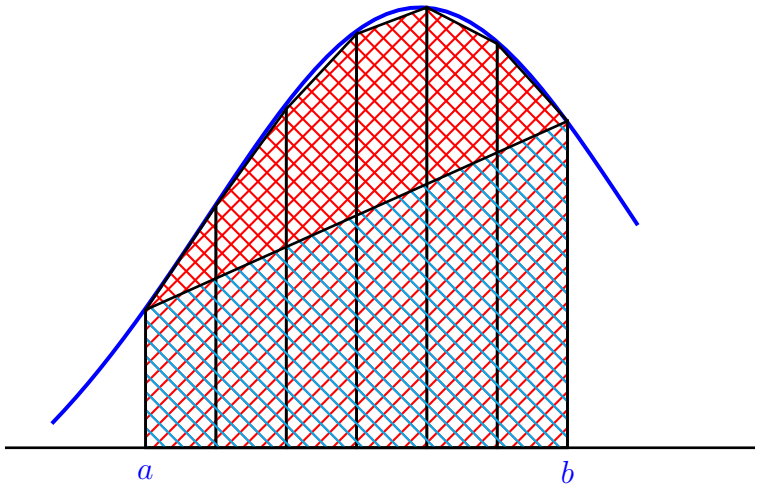
dove i punti x_i sono:

$$x_i \quad = a + ih \qquad i = 0, \ldots, N \tag{4.5}$$

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.

4.3.1 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 (4.4), e con i nodi x_i definiti in (4.5). 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 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

$$|R_T| \leq \frac{1}{12}\frac{(b-a)^3}{N^2}M, \qquad M = \max_{x\in[a,b]} |f''(x)|.$$

Imponendo che $|R_T| \leq \varepsilon$, precisione prefissata, segue

$$N_\varepsilon \geq \sqrt{\frac{(b-a)^3 M}{12\varepsilon}}. \tag{4.6}$$

Tuttavia questo numero spesso risulta una stima eccessiva a causa della maggiorazione della derivata seconda tramite M.

Esempio 4.3.1 *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 (4.6) segue che

$$N_\varepsilon \geq \sqrt{\frac{1}{12\varepsilon}} = 29.$$

4.3.2 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

$$N_\varepsilon \geq \sqrt[4]{\frac{(b-a)^5 M}{180\varepsilon}}. \tag{4.7}$$

Esempio 4.3.2 *Risolvere il problema descritto nell'esempio 4.3.1 applicando la formula di Simpson composta.*

La derivata quarta della funzione integranda è

$$f''(x) = -\frac{6}{x^4}$$

quindi è maggiorata da $M = 6$. Dalla relazione (4.7) segue che

$$N_\varepsilon \geq \sqrt[4]{\frac{6}{180\varepsilon}} > 4,$$

quindi $N_\varepsilon \geq 6$.

4.3.3 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 4.2.1. 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.

4.3.4 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

$$M_C(h) = 2h \sum_{i=0}^{\frac{N}{2}-1} f(x_{2i+1})$$

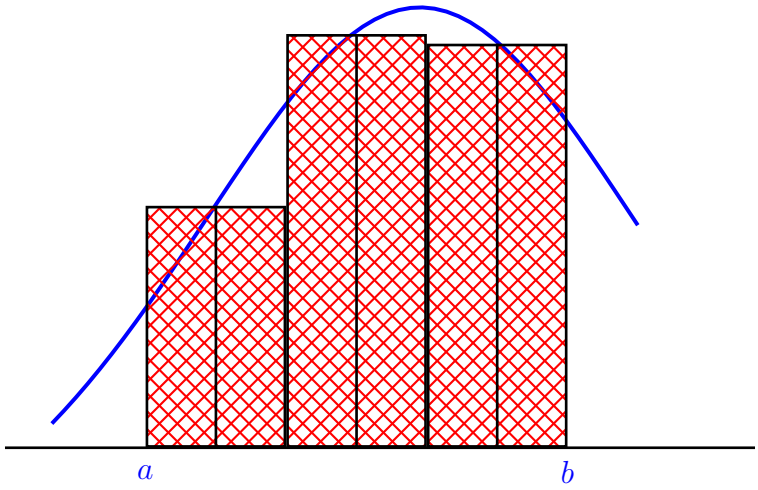


Figura 4.1: Formula del Punto di Mezzo Composta

mentre il resto è

$$R_M = \frac{(b-a)^3}{6N^2} f''(\eta). \tag{4.8}$$

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

$$N_\varepsilon \geq \sqrt{\frac{(b-a)^3 M}{6\varepsilon}}. \tag{4.9}$$

Nella Figura 4.1 sono evidenziate le aree che approssimano l'integrale utilizzando la formula del punto di mezzo composta.

Esempio 4.3.3 *Risolvere il problema descritto nell'esempio 4.3.1 applicando la formula di Simpson composta.*

La derivata seconda della funzione integranda è maggiorata da $M = 1$. Da (4.9) risulta

$$N_\varepsilon \geq \sqrt{\frac{1}{6\varepsilon}} > 40.$$