

Contents

1 Aritmetica computazionale	1
1.1 Rappresentazione dei numeri reali	1
1.2 Errori di rappresentazione	6
1.3 Aritmetica finita	7
1.3.1 Caratterizzazione di u	8
1.4 Analisi degli errori	9
1.4.1 Analisi in avanti degli errori nelle operazioni di moltiplicazione e addizione	10
1.4.2 Condizionamento di un problema e stabilità di un algoritmo	11
1.4.3 Numero di condizione	13
2 Funzioni polinomiali	18
2.1 Valutazione di un polinomio	18
2.1.1 Valutazione numerica della derivata	20
2.2 Polinomi nella base di Bernstein	22
2.2.1 Cambio di variabile	24
2.2.2 Proprietà dei polinomi di Bernstein	25
2.2.3 Valutazione di un polinomio nella base di Bernstein	26
3 Interpolazione polinomiale	31
3.1 Esistenza e unicità dell'interpolazione polinomiale	31
3.2 Metodi di costruzione	33
3.2.1 Base di Newton	33
3.2.2 Base di Bernstein	35
3.2.3 Base di Lagrange	36
3.3 Errore di interpolazione (interpolazione di funzioni)	39
3.3.1 Punti equispaziati e di Chebyshev	40
3.4 Interpolazione polinomiale a tratti	41
3.5 Condizionamento dell'interpolazione (interpolazione di funzioni)	42
4 Integrazione numerica	44
4.1 Formule di quadratura di Newton-Cotes	44
4.1.1 Formula (dei Trapezi) per $n = 1$	45
4.1.2 Formula (di Simpson) per $n = 2$	46
4.1.3 Errore di integrazione	47
4.2 Formule di quadratura composite	48
4.3 Metodi adattivi	50
4.3.1 Estrapolazione di Richardson	50
5 Zeri di funzioni non lineari	53
5.1 Metodo di bisezione	53
5.1.1 Metodo della falsa posizione	54
5.2 Metodo di Newton	55
5.2.1 Metodo delle secanti	60
6 Algebra lineare numerica	61
6.1 Fattorizzazione LU	61
6.1.1 Sostituzione in avanti	62
6.1.2 Sostituzione all'indietro	62
6.1.3 Metodo di Gauss	62
6.1.4 Fattorizzazione LU con scambio delle righe e perno massimo	65
6.2 Condizionamento del problema $Ax = b$	67
6.3 Fattorizzazione QR	69
6.3.1 Matrici elementari di Householder	70
6.3.2 Metodo di Householder	72
6.4 Metodi iterativi	74
6.4.1 Metodi di Jacobi e Gauss-Seidel	75

1 Aritmetica computazionale

1.1 Rappresentazione dei numeri reali

I **numeri finiti** sono utilizzati dai calcolatori per rappresentare i numeri reali poiché questi ultimi possono avere un numero infinito di cifre, che i calcolatori, avendo una memoria limitata, non sono in grado di rappresentare.

Teorema (Rappresentazione in base). *Sia α un numero reale non nullo. Possiamo rappresentare tale numero con una base $\beta \geq 2$, un numero intero scelto da noi, nel seguente modo:*

$$\begin{aligned}\alpha &= \pm(\alpha_1\beta^{-1} + \alpha_2\beta^{-2} + \dots)\beta^p \\ \alpha &= \pm\left(\sum_{i=1}^{\infty} \alpha_i\beta^{-i}\right)\beta^p\end{aligned}\tag{1.1}$$

I vari termini della 1.1 vengono detti:

β	base
p	esponente
α_i	cifre del numero
$\sum_{i=1}^{\infty} \alpha_i\beta^{-i}$	mantissa

Ogni cifra α_i è un numero intero che varia tra 0 e $\beta - 1$. Ad esempio, se lavoriamo in base 10, le cifre saranno numeri interi compresi tra 0 e 9.

Per garantire l'unicità della rappresentazione, è necessario che $\alpha_1 \neq 0$. Se così non fosse, il numero 13 potrebbe essere rappresentato come 13, 013, 0013, eccetera, il che va contro l'unicità della rappresentazione.

Possiamo scrivere un numero $\alpha \in \mathbb{R}$ con $\alpha \neq 0$ in due modi:

1. **forma mista.**

$$\alpha = \begin{cases} \pm(0.000\alpha_1\alpha_2\dots)\beta & p \leq 0 \\ \pm(\alpha_1\alpha_2\dots)\beta & p > 0 \end{cases}$$

2. **forma scientifica.** L'idea è quella di spostare il punto decimale al primo numero $\neq 0$ e poi moltiplicare il tutto per β^p per riportare il numero al suo valore originale.

$$\alpha = \pm 0.\alpha_1\alpha_2\dots \cdot \beta^p$$

Esempio:

$$\begin{aligned}\alpha &= (12.37)_{10} & \alpha &= 0.12237 \cdot 10^2 \\ \alpha &= (0.0045)_{10} & \alpha &= 0.45 \cdot 10^{-2} \\ & & &= (4 \cdot 10^{-1} + 5 \cdot 10^{-2}) \cdot 10^{-2}\end{aligned}$$

Definizione (Numeri finiti). L'insieme \mathbb{F} dei numeri finiti è definito come l'insieme dei numeri espressi in base β (dove $\beta \geq 2$), utilizzando t cifre (con $t \geq 1$). Poiché anche l'esponente p potrebbe essere così grande da non poter essere rappresentato, è necessario limitare l'intervallo degli esponenti rappresentabili. Qui, λ indica il più piccolo esponente che può essere rappresentato e ω il più grande esponente rappresentabile.

$$\begin{aligned}\mathbb{F}(\beta, t, \lambda, \omega) &= \{0\} \cup \{\alpha \in \mathbb{R} : \alpha = \pm 0.\alpha_1\alpha_2\dots\alpha_t \cdot \beta^p, \\ &= \{0\} \cup \{\alpha \in \mathbb{R} : \alpha = \pm\left(\sum_{i=1}^t \alpha_i\beta^{-i}\right)\beta^p, \\ &\quad \text{con } 0 \geq \alpha_i < \beta, \text{ per } i = 1, 2, \dots, t, \alpha_1 \neq 0, \lambda \leq p \leq \omega\}\end{aligned}$$

\mathbb{F} è un sottoinsieme che rappresenta una discretizzazione di \mathbb{R} . In altre parole, \mathbb{F} è un insieme discreto di numeri presi da \mathbb{R} , dove ciascun numero può essere espresso al più in t cifre. Questo significa che gli elementi di \mathbb{F} sono una selezione discreta di numeri reali con una precisione limitata a t cifre decimali.

Per convenzione, utilizzeremo α per scrivere i numeri reali e $\tilde{\alpha}$ per scrivere i numeri finiti.

Esempio:

Determinare e posizionare sull'asse reale gli elementi di $\mathbb{F}(2, 3, -1, 2)$.

I numeri rappresentabili possono essere espressi come:

$$\tilde{\alpha} = \pm 0.\alpha_1\alpha_2\alpha_3 \cdot 2^p$$

$$\tilde{\alpha} = \pm \left(\sum_{i=1}^3 \alpha_i \cdot 2^{-i} \right) \cdot 2^p$$

con $\tilde{\alpha} \in \mathbb{F}$, $-1 \leq p < 3$ e $\alpha_1 \neq 0$

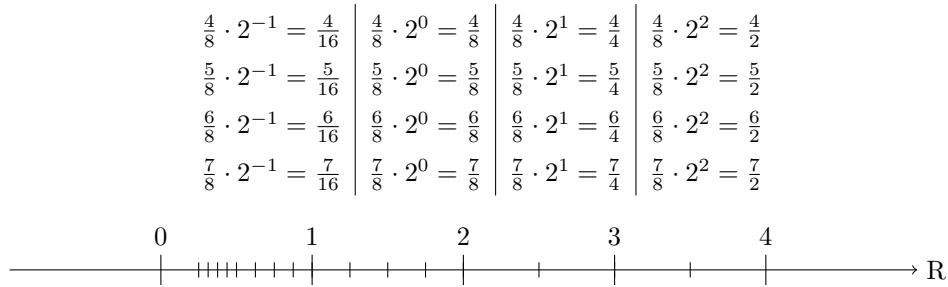
L'insieme delle possibili mantisse m_3 è dato da:

$$m_3 = \{0.100, \\ 0.101, \\ 0.110, \\ 0.111\} \times \{2^{-1}, 2^0, 2^1, 2^2\}$$

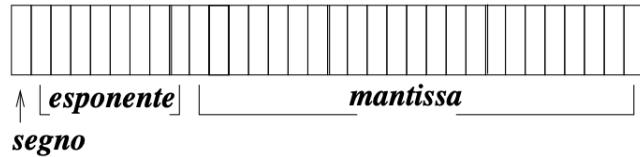
Pertanto, l'insieme degli elementi di $\mathbb{F}(2, 3, -1, 2)$ è composto da 33 elementi. Di questi, 16 sono positivi, 16 sono negativi e uno è lo zero.

Per capire come questi elementi sono posizionati sull'asse reale, li portiamo in base 10.

$$0.100 = 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 0 \cdot 2^{-3} = \frac{1}{2} = \frac{4}{8} \\ 0.101 = 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} = \frac{5}{8} \\ 0.110 = 1 \cdot 2^{-1} + 1 \cdot 2^{-2} + 0 \cdot 2^{-3} = \frac{3}{4} = \frac{6}{8} \\ 0.111 = 1 \cdot 2^{-1} + 1 \cdot 2^{-2} + 1 \cdot 2^{-3} = \frac{7}{8}$$



Basic precision single	$\mathbb{F}(2, 24, -127, 128)$	32 bit
Basic precision double	$\mathbb{F}(2, 53, -1023, 1024)$	64 bit



In precisione singola vengono destinati 24 bit alla mantissa (in realtà solo $2^{3^1} = 23$ ¹) e 8 all'esponente ($2^8 = 256 = \omega - \lambda + 1$, con $\lambda = -127$ e $\omega = 128$), mentre in precisione doppia le cifre della mantissa sono 53 (memorizzati 52 bit) e dell'esponente 11 ($2^{11} = 2048 = \omega - \lambda + 1$, con $\lambda = -1023$ e $\omega = 1024$).

Si osservi che l'esponente è memorizzato per traslazione (*exponent biased*) e che la costante di traslazione (*bias*) è $-\lambda$. Quindi, se p è l'esponente del numero e \tilde{p} è l'esponente memorizzato, possiamo trovare l'esponente memorizzato a partire dall'esponente originale utilizzando la seguente relazione:

$$\tilde{p} = p - \lambda$$

Dato un numero reale non nullo, α , per associare un numero finito ad esso, procediamo come segue:

1. **Rappresentazione esatta.** Se α è scritto nella forma $\alpha = \pm(\alpha_1\alpha_2\dots) \times \beta^p$ tale che $\lambda \leq p \leq \omega$, $\alpha_i = 0$ per $i > t$, allora è rappresentabile esattamente come un numero finito t di cifre e $\alpha \in \mathbb{F}(\beta, t, \lambda, \omega)$.
2. **Rappresentazione approssimata.** Altrimenti $\alpha \notin \mathbb{F}(\beta, t, \lambda, \omega)$ e quindi bisogna associargli un numero approssimato $\tilde{\alpha}$ che indicheremo con $fl(\alpha)$. Si hanno i seguenti casi:

- $p \notin [\lambda, \omega]$, viene segnalata una condizione d'errore:

$$\begin{array}{ll} p < \lambda & underflow \\ p > \omega & overflow \end{array}$$

- $p \in [\lambda, \omega]$, ma le cifre a_i con $i > t$ non sono tutte nulle, allora viene assegnato un numero finito $fl(\alpha)$ seguendo due possibili criteri:

- **Troncamento** di α alla t -esima cifra

$$fl_T(\alpha) = \pm \left(\sum_{i=1}^t \alpha_i \beta^{-i} \right) \beta^p$$

- **Arrotondamento** di α alla t -esima cifra

$$fl_A(\alpha) = \pm fl_T \left(\left(\sum_{i=1}^{t+1} \alpha_i \beta^{-i} + \frac{\beta}{2} \beta^{-(t+1)} \right) \beta^p \right)$$

Esempio:

Il numero $\alpha = (0.11011)_2$ ha una mantissa di lunghezza 5, che è più lunga delle 3 cifre consentite in $\mathbb{F}(2, 3, -1, 2)$. Quindi, procediamo con l'operazione di arrotondamento:

$$\begin{array}{r} fl_A(\alpha) = 0.11011 + \\ 0.00010 = \\ \hline 0.11100 \end{array}$$

¹Essendo sempre $\alpha_1=1$ per la rappresentazione binaria, la prima cifra può essere sottintesa senza mai essere fisicamente memorizzata.

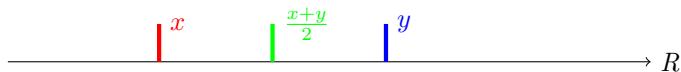
Esempio:

Consideriamo l'insieme dei numeri finiti $\mathbb{F}(10, 5, -50, 49)$. Per rappresentare un numero finito in questo insieme in memoria, dobbiamo definire il numero di posizioni necessarie. Nello specifico:

- **Segno:** una posizione è riservata per il segno. Se il numero è positivo si usa 0; se è negativo, si usa $\beta - 1$.
- **Esponente:** due posizioni sono destinate all'esponente. Usando la tecnica di memorizzazione per traslazione ($p - \lambda = \tilde{p}$), possiamo rappresentare gli esponenti da -50 a 49 attraverso valori memorizzati da 00 a 99.
- **Mantissa:** cinque posizioni sono dedicate alla mantissa.

$$\begin{aligned}\alpha &= 0.0532 = 0.532 \cdot 10^{-1} & fl(\alpha) &= 04953200 \\ \alpha &= -237141 = -0.237141 \cdot 10^6 & fl(\alpha) &= 95623714\end{aligned}$$

Osservazione. Siano x ed y due numeri $\in \mathbb{F}$ consecutivi positivi. Sia $\alpha \in \mathbb{R}$ tale che $x \leq \alpha < y$.



Allora possiamo affermare che α non appartiene all'insieme \mathbb{F} perché, per ipotesi, x e y sono consecutivi e non ci può essere un altro numero tra loro. Tuttavia, la rappresentazione approssimata $fl(\alpha)$ risulta essere:

$$fl_T(\alpha) = x \quad fl_A(\alpha) = \begin{cases} x & \text{se } \alpha < \frac{x+y}{2} \\ y & \text{se } \alpha \geq \frac{x+y}{2} \end{cases}$$

L'errore commesso nel troncamento sarà sempre maggiore o uguale dell'errore commesso nell'arrotondamento. Questo è il motivo per cui, con una base numerica pari, si preferisce utilizzare l'arrotondamento, poiché fornirà una migliore approssimazione del numero reale rispetto al troncamento.

La modalità di arrotondamento dello standard ANSI/IEEE-754 coincide con quella precedentemente descritta, con la particolarità dell'**arrotondamento ai pari**. Questa particolarità si applica quando un numero reale α è esattamente equidistante dai numeri finiti consecutivi x ed y , in altre parole, quando $\alpha = \frac{x+y}{2}$. In questa situazione l'arrotondamento funziona nel seguente modo:

$$fl_{AP}(\alpha) = \begin{cases} x & \text{se } x \text{ è pari} \\ y & \text{se } y \text{ è pari} \end{cases}$$

Sempre parlando dello standard ANSI/IEEE-754, per gestire risultati non rappresentabili, vengono utilizzati due valori speciali:

- **NaN**
- **Inf**

Invece, di avere un buco vicino allo zero dove i numeri molto piccoli verrebbero immediatamente arrotondati a zero, vengono inseriti dei numeri ulteriori per riempire questo vuoto e permettere ai valori di avvicinarsi progressivamente a zero. Questo meccanismo è chiamato **gradual underflow**.

Per rappresentare questi numeri estremamente piccoli, si fa uso della rappresentazione **denormalizzata**. In questa rappresentazione, la mantissa non inizia con il solito bit隐含的 di 1, ma con una serie di 0.

Esempio:

Si eseguano i passi necessari per rappresentare il numero reale $(-13.9)_{10}$ in un'area di memoria di 8 bit (1 per il segno, 3 per l'exponent biased e 4 per la mantissa), che permettono di memorizzare $\mathbb{F}(2, 5, -3, 4)$ per troncamento e arrotondamento.

1. Conversione in binario, prima la parte intera, quindi la parte decimale:

- *Parte intera:*

- Dividi il numero per 2.
- Registra il resto della divisione (sarà 0 o 1).
- Usa il quoziente ottenuto come nuovo numero e ripeti la divisione per 2.
- Continua il processo fino a quando il quoziente diventa 0.
- Leggi i resti della divisione in ordine inverso: questo sarà il numero in base 2 della parte intera.

$$(13)_{10} = (1101)_2$$

- *Parte decimale:*

- Moltiplica la parte decimale per 2.
- Registra la parte intera del risultato (sarà 0 o 1).
- Usa la parte decimale del risultato come nuovo numero e ripeti la moltiplicazione per 2.
- Continua questo processo finché non ottieni una parte decimale di 0 o si arriva al limite di precisione della mantissa.
- Leggi i numeri interi in ordine di apparizione: questo sarà il numero in base 2 della parte decimale.

$$0.9 \times 2 = \underline{1}.8$$

$$0.8 \times 2 = \underline{1}.6$$

$$0.6 \times 2 = \underline{1}.2$$

$$0.2 \times 2 = \underline{0}.4$$

$$0.4 \times 2 = \underline{0}.8$$

$$(0.9)_{10} = (11100\dots)_2$$

da cui

$$(-13.9)_{10} = (-1101.11100\dots)_2$$

2. Normalizzazione: nello standard IEEE-754, la rappresentazione normalizzata dei numeri in virgola mobile prevede che la parte intera sia sempre 1.

$$(-1101.11100\dots)_2 = (-1.10111100\dots)_2 \times 2^3$$

3. Calcolo dell'esponente biased:

$$p - \lambda = \tilde{p} \rightarrow 3 - (-3) = 6$$

$$(-1.10111100\dots)_2 \times 2^3 = (-1.10111100\dots)_2 \times 2^{(110)_2}$$

4. Rappresentazione della mantissa:

arrotondamento	troncamento
$1.10111 + 0.00001 = 1.1100$	1.1011

5. Rappresentazione in memoria: nello standard IEEE-754, con una mantissa di 5 bit, solo 4 bit vengono effettivamente memorizzati in memoria.

1	1	1	0	1	1	0	0
1	1	1	0	1	0	1	1

1.2 Errori di rappresentazione

Definizione. Consideriamo un valore $\alpha \in \mathbb{R}$. Se $\alpha \notin \mathbb{F}(\beta, t, \lambda, \omega)$, allora la sua migliore approssimazione all'interno di questo insieme è data da $\tilde{\alpha} \in \mathbb{F}(\beta, t, \lambda, \omega)$. L'approssimazione di α con $\tilde{\alpha}$ introduce un **errore di rappresentazione**. Per quantificare tale errore, definiamo le seguenti metriche:

$$E_{abs} = |\alpha - fl(\alpha)| \quad \text{errore assoluto}$$

$$E_{rel} = \left| \frac{\alpha - fl(\alpha)}{\alpha} \right| \text{ se } \alpha \neq 0 \quad \text{errore relativo}$$

Nel calcolo scientifico, l'errore relativo è preferito poiché fornisce una misura dell'errore “normalizzata”, che non dipende dalla grandezza dei numeri confrontati.

Esempio:

Si converta quanto rappresentato nell'esempio precedente nuovamente in base 10. Successivamente, si valuti l'errore assoluto e l'errore relativo della rappresentazione.

6. **Decodifica:** per riconvertire il numero floating point appena determinato, faremo:

arrotondamento	troncamento
$(-1.1100)_2 \times 2^{(110)_2}$	$(-1.1011)_2 \times 2^{(110)_2}$
$(-1.1100)_2 \times 2^{(3)_{10}}$	$(-1.1011)_2 \times 2^{(3)_{10}}$
$(-1110.0)_2$	$(-1101.1)_2$
$-(8 + 4 + 2 + 0 + 0)_{10}$	$-(8 + 4 + 0 + 1 + 0.5)_{10}$
$-(14)_{10}$	$-(13.5)_{10}$

	arrotondamento	troncamento
errore assoluto	$-13.9 - (-14) = 0.1$	$-13.9 - (-13.5) = -0.4$
errore relativo	$\frac{0.1}{-13.9} = -0.0072$	$\frac{-0.4}{-13.9} = 0.0288$

Definizione. Dato l'insieme dei numeri finiti $\mathbb{F}(\beta, t, \lambda, \omega)$, si dice **unità di arrotondamento** e la si indica con u , la quantità:

$$u = \begin{cases} \beta^{1-t} & \text{per troncamento} \\ \frac{1}{2}\beta^{1-t} & \text{per arrotondamento} \end{cases}$$

Teorema. Per ogni $\alpha \in \mathbb{R}$ e $\alpha \neq 0$ vale

$$\left| \frac{\alpha - fl(\alpha)}{\alpha} \right| < u$$

Il teorema afferma che u , l'unità di arrotondamento, rappresenta il limite superiore dell'errore relativo quando si rappresenta un numero reale in un formato numerico finito.

Esempio:

Consideriamo l'insieme dei numeri finiti $\mathbb{F}(2, 5, -3, 4)$. Calcolare l'unità di arrotondamento u sia nel caso di troncamento che di arrotondamento.

$$u = \begin{cases} 2^{1-5} = \frac{1}{16} = 0.0625 & \text{per troncamento} \\ \frac{1}{2} \cdot 2^{1-5} = \frac{1}{32} = 0.0325 & \text{per arrotondamento} \end{cases}$$

Indicheremo con ϵ l'errore relativo.

Corollario. Per ogni $\alpha \in \mathbb{R}$ e $\alpha \neq 0$ vale

$$fl(\alpha) = \alpha(1 \pm \epsilon), \quad \text{con } |\epsilon| < u$$

Dimostrazione.

Banalmente dato $\epsilon = \frac{\alpha - fl(\alpha)}{\alpha}$, per il Teorema si ha che $|\epsilon| < u$ e $fl(\alpha) = \alpha\epsilon + \alpha = \alpha(1 + \epsilon)$.

■

Precisione desiderata in base 10. La questione chiave è: quante cifre in base 10 sono necessarie per rappresentare con precisione ciò che è memorizzato in base 2?

Supponiamo di avere un numero rappresentato con t cifre in base 2. Vogliamo sapere a quante cifre, s , in base 10 questo corrisponde.

Partendo dall'equazione:

$$2^{-t} = 10^{-s}$$

e applicando il logaritmo in base 10 ad entrambi i lati:

$$-t \times \log_{10}(2) = -s$$

da qui possiamo isolare s :

$$s = t \times \log_{10}(2)$$

usando un'approssimazione per il logaritmo:

$$s \approx t \times 0.30103$$

Per esempio:

- Nella precisione ‘basic single’, con $t = 24$ cifre in base 2 per la mantissa, abbiamo:

$$s \approx 24 \times 0.30103 \approx 7.224$$

- Nella precisione ‘basic double’, con $t = 53$ cifre in base 2 per la mantissa, abbiamo:

$$s \approx 53 \times 0.30103 \approx 15.95459$$

Questo indica che ci servono circa 7-8 cifre in ‘basic single’ o circa 16 cifre in ‘basic double’ in base 10 per rappresentare con precisione ciò che è memorizzato in base 2. Utilizzando meno cifre, stiamo arrotondando e potremmo perdere informazioni.

Possiamo calcolare l'unità di arrotondamento u per ‘basic single’ e ‘basic double’:

- $u_{single} = \frac{1}{2} \times 2^{1-24} = 2^{-24} \approx 5.96 \times 10^{-8}$
- $u_{double} = \frac{1}{2} \times 2^{1-53} = 2^{-53} \approx 1.116 \times 10^{-16}$

Ora, se confrontiamo questi valori con le cifre necessarie in base 10 per una rappresentazione accurata, notiamo una relazione. Le cifre necessarie sono legate all'ordine di grandezza dell'unità di arrotondamento.

Questi valori forniscono un indicatore sull'ordine di grandezza minimo dei numeri che possono essere rappresentati accuratamente e sul numero massimo di cifre che possiamo stampare senza perdere informazioni.

1.3 Aritmetica finita

Dati due numeri a e b appartenenti a $\mathbb{F}(\beta, t, \lambda, \omega)$, l'operazione $a \ op \ b$ potrebbe produrre un risultato che non è contenuto in $\mathbb{F}(\beta, t, \lambda, \omega)$.

Esempio:

Siano $a = (0.34)_{10} \times 10^0$ e $b = (0.12)_{10} \times 10^{-2} \in \mathbb{F}(10, 2, \lambda, \omega)$. Eseguendo la somma si ha:

$$0.34 + 0.0012 = 0.3412$$

ma, $0.3412 \notin \mathbb{F}(10, 2, \lambda, \omega)$.

Per eseguire le operazioni in questo dominio, vengono definiti degli operatori specifici, che indichiamo con \tilde{op} (ad esempio, $\tilde{+}$, $\tilde{-}$, $\tilde{\times}$, $\tilde{/}$).

Definizione. L'operatore \tilde{op} tra due numeri $a, b \in \mathbb{F}$ è definito nel modo seguente:

$$a \ \tilde{op} \ b = fl(a \ op \ b)$$

Questo significa che viene prima eseguita l'operazione in aritmetica esatta, e il risultato viene arrotondato per rientrare nell'insieme di numeri finiti \mathbb{F} .

Per soddisfare tali requisiti, si utilizzano dei registri posizionati vicino al processore. Questi registri, dotati di bit aggiuntivi, rispetto a quelli della mantissa, permettono di eseguire operazioni con una precisione superiore rispetto a quella raggiungibile con solo t bit. Tale maggiore precisione assicura che, arrotondando a t cifre, il risultato ottenuto aderisce alla definizione delineata sopra. Idealmente, un registro di lunghezza $t + 1$ bit, superiore a quella della mantissa stessa, garantirebbe la conformità a questa definizione.

Errore in aritmetica finita. Qual'è l'errore massimo che possiamo commettere durante un'operazione con numeri finiti? Consideriamo l'errore relativo tra il risultato ottenuto in aritmetica finita e quello in aritmetica esatta, si può notare un interessante comportamento. Notiamo che, il risultato esatto di $(a \text{ op } b)$ è un numero $\alpha \in \mathbb{R}$. Per definizione, in aritmetica finita, $(a \tilde{\text{op}} b)$ è l'approssimazione floating-point di α . Allora, per il teorema sopra menzionato, possiamo dedurre che l'errore relativo massimo tra α e $fl(\alpha)$ è minore dell'unità di arrotondamento u . Questo significa che u rappresenta l'errore relativo massimo che possiamo aspettarci in una singola operazione in aritmetica finita. Estendendo questa logica, se effettuiamo una serie di n operazioni, l'errore totale potrebbe essere al più $n \cdot u$.

$$\left| \frac{\overbrace{(a \tilde{\text{op}} b) - (a \text{ op } b)}^{fl(\alpha)} - \overbrace{\alpha}^{\alpha}}{\underbrace{a \text{ op } b}_{\alpha}} \right| < u$$

Proprietà associativa. La proprietà associativa afferma che l'ordine in cui si raggruppano i termini durante un'operazione non modifica il risultato. Tuttavia, essa, così come le altre proprietà, non vale nell'aritmetica finita.

Esempio:

Considerati $a = 0.11 \times 10^0, b = 0.13 \times 10^{-1}, c = 0.14 \times 10^{-1} \in \mathbb{F}(10, 2, \lambda, \omega)$. Verificare se la proprietà associativa $(a \tilde{+} b) \tilde{+} c = a \tilde{+} (b \tilde{+} c)$ è valida.

$$\begin{aligned} (0.11 \tilde{+} 0.013) \tilde{+} 0.014 &= 0.11 \tilde{+} (0.013 \tilde{+} 0.014) \\ fl(0.123) \tilde{+} 0.014 &= 0.11 \tilde{+} fl(0.027) \\ 0.12 \tilde{+} 0.014 &= 0.11 \tilde{+} 0.03 \\ fl(0.134) &= 0.14 \\ 0.13 \times 10^0 &\neq 0.14 \times 10^0 \end{aligned}$$

Concludendo, la proprietà associativa non è valida nell'ambito dell'aritmetica finita.

Questo significa che la stessa istruzione o espressione, se scritta in modi diversi, può produrre risultati differenti. Di conseguenza, è importante comprendere come evitare di scrivere operazioni che potrebbero causare errori più grandi.

1.3.1 Caratterizzazione di u

L'unità di arrotondamento u ha un'importanza numerica sia in relazione alla precisione di rappresentazione (Teorema) che in termini di precisione di calcolo (risultato precedente). La sua importanza numerica è ulteriormente sottolineata dalla seguente caratterizzazione:

u è il più piccolo numero finito positivo tale che, se sommato a 1, viene “sentito” e risulta essere $>$ di 1.

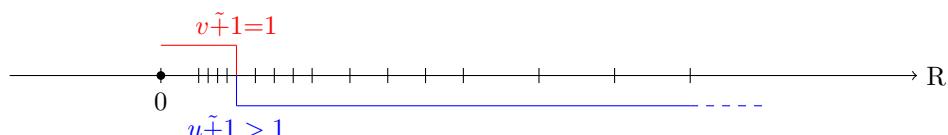
$$u \tilde{+} 1 > 1$$

Questo implica che per ogni numero finito $v < u$ sarà $v \tilde{+} 1 = 1$.

Infatti, se sommiamo un numero v a 1:

$$\underbrace{1.0 \dots 0}_{t \text{ cifre}} + \underbrace{0.0 \dots 0}_1$$

Tuttavia, il valore 1 rimane fuori dalle t cifre e nella somma viene arrotondato, e non viene “sentito” poiché il risultato sarà comunque 1.



Invece, se sommiamo un numero u a 1:

$$\begin{aligned}
 u\tilde{+}1 &= \frac{1}{2}\beta^{1-t} + 1 \\
 &= \frac{\beta}{2}\beta^{-t} + 1 \\
 &= 0.0\dots 0\frac{\beta}{2} + \underbrace{1.0\dots 0}_{t \text{ cifre}} \\
 &= 1.0\dots 0\frac{\beta}{2} + 0.0 + 0.0\dots 0\frac{\beta}{2} \text{ per arrotondamento} \\
 &= \underbrace{1.0\dots 1}_{t \text{ cifre}} 0 \text{ per troncamento} \\
 1.0\dots 1 &> 1
 \end{aligned}$$

Nello standard IEEE-754, se consideriamo l'arrotondamento ai pari è dato che $1.0\dots 0\frac{\beta}{2} = \frac{x+y}{2}$, dove $x = 1$ e $y = 1.0\dots 1$, il valore pari più vicino è 1. Pertanto, 1 verrà scelto come risultato dell'arrotondamento. Questo cambia la caratterizzazione di u :

u è il più grande numero finito positivo tale che, se sommato a 1, risulta essere = 1.

$$u\tilde{+}1 = 1$$



Ma cosa ci serve tutto a questo? Per determinare proceduralmente u :

1. Esaminare le potenze negative di 2.
2. Continuare finché non si individua una potenza 2^{-t} che, quando sommata a 1, produce come risultato esattamente 1.
3. Tale potenza 2^{-t} rappresenta l'unità di arrotondamento u .

```

u=1
t=0
while (u+1>1)
    u=u/2
    t=t+1
end
stampa(u,t)

```

1.4 Analisi degli errori

Nella risoluzione di problemi in aritmetica finita su un calcolatore, dobbiamo prima chiederci cosa sia un “problema ben posto”. Per definire un problema come **ben posto**, è necessario che soddisfi due requisiti:

- Il problema deve ammettere una e una sola soluzione.
- La soluzione deve dipendere con continuità dai dati in ingresso.

Un buon modo per modelizzare il nostro problema è tramite una funzione $f : \mathbb{R} \rightarrow \mathbb{R}$ che a partire da un dato x produce un risultato $f(x)$. Affinché il problema sia ben posto, questa funzione deve essere continua. E per la sua stessa definizione, una funzione restituisce sempre un unico output per ogni input.

Ora, consideriamo un tipico flusso di lavoro legato alla risoluzione di un problema su un calcolatore:

- **Input.** Il dato viene letto e tradotto in una rappresentazione in aritmetica finita.
- **Elaborazione.** Si applica un algoritmo, anch'esso in aritmetica finita, per elaborare il dato.
- **Valutazione del risultato.** Si esamina il risultato, l'errore associato e si decide se il risultato è accettabile oppure no.

Per valutare l'entità dell'errore, possiamo utilizzare:

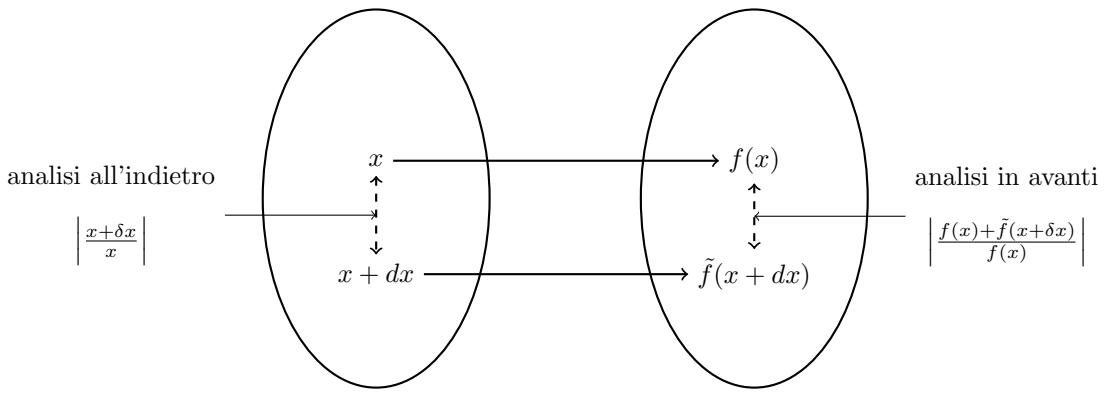


Figure 1: Analisi dell'errore

- **Analisi in avanti:** si calcola l'errore relativo sul risultato finale in termini degli errori introdotti dalle singole operazioni, trascurando i termini in cui compaiono prodotti di errori (analisi del 1° ordine).
- **Analisi all'indietro:** approccio opposto al precedente consiste nel considerare il risultato $\tilde{f}(x+\delta x)$ come risultato esatto derivato da dati iniziali perturbati rispetto a quelli reali. La valutazione è data quindi da un fattore δx sul dato iniziale x .

L'immagine 1 mostra che se la distanza tra x e $x+\delta x$ è piccola e, analogamente, la distanza tra $f(x)$ e $\tilde{f}(x+\delta x)$ è anch'essa piccola, allora possiamo considerare il risultato come "buono".

1.4.1 Analisi in avanti degli errori nelle operazioni di moltiplicazione e addizione

Applicando l'analisi in avanti alle operazioni di moltiplicazione e addizione si ottengono alcuni importanti risultati:

- **Moltiplicazione:** Siano $x, y \in \mathbb{R}$. Consideriamo la funzione $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ definita come $f(x, y) = x \cdot y$, che moltiplica due numeri reali e restituisce il risultato.

Applichiamo l'analisi in avanti, tenendo conto sia dell'errore sui dati che dell'errore di calcolo:

$$\begin{aligned} x &\rightarrow fl(x) = x(1 + \epsilon_1) & |\epsilon_1| < u \\ y &\rightarrow fl(y) = y(1 + \epsilon_2) & |\epsilon_2| < u \\ fl(x) \cdot fl(y) &= fl(fl(x)fl(y)) \\ &= fl(x)fl(y)(1 + \epsilon_3) & |\epsilon_3| < u \\ &= x(1 + \epsilon_1)y(1 + \epsilon_2)(1 + \epsilon_3) \end{aligned}$$

Calcoliamo ora l'errore relativo:

$$\begin{aligned} \left| \frac{fl(fl(x)fl(y)) - f(x, y)}{f(x, y)} \right| &= \left| \frac{x(1 + \epsilon_1)y(1 + \epsilon_2)(1 + \epsilon_3) - xy}{xy} \right| \\ &= \left| \frac{(xy)((1 + \epsilon_1)(1 + \epsilon_2)(1 + \epsilon_3) - 1)}{xy} \right| \\ &= \left| (1 + \epsilon_1)(1 + \epsilon_2)(1 + \epsilon_3) - 1 \right| \\ &= \left| 1 + \epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_1\epsilon_2 + \epsilon_1\epsilon_3 + \epsilon_2\epsilon_3 + \epsilon_1\epsilon_2\epsilon_3 - 1 \right| \end{aligned}$$

Trascuriamo il prodotto di errori, poiché numericamente irrilevanti (analisi di 1° ordine):

$$\approx |\epsilon_1 + \epsilon_2 + \epsilon_3| \leq |\epsilon_1| + |\epsilon_2| + |\epsilon_3| < 3u$$

Abbiamo quantificato un limite superiore per l'errore sul risultato finale. Dato che ci sono 3 operazioni e l'errore finale massimo è di $3u$, il risultato è sia accettabile che aspettato.

- **Addizione:** Siano dati $x, y \in \mathbb{R}$. Consideriamo la funzione $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ definita come $f(x, y) = x + y$, che somma due numeri reali e restituisce il risultato.

Applichiamo l'analisi in avanti, tenendo conto sia dell'errore sui dati che dell'errore di calcolo:

$$\begin{aligned} x \rightarrow fl(x) &= x(1 + \epsilon_1) & |\epsilon_1| < u \\ y \rightarrow fl(y) &= y(1 + \epsilon_2) & |\epsilon_2| < u \\ fl(x) + fl(y) &= fl(fl(x) + fl(y)) \\ &= (fl(x)fl(y))(1 + \epsilon_3) & |\epsilon_3| < u \\ &= (x(1 + \epsilon_1) + y(1 + \epsilon_2))(1 + \epsilon_3) \end{aligned}$$

Calcoliamo ora l'errore relativo:

$$\begin{aligned} \left| \frac{fl(fl(x) + fl(y)) - f(x, y)}{f(x, y)} \right| &= \left| \frac{(x(1 + \epsilon_1) + y(1 + \epsilon_2))(1 + \epsilon_3) - (x + y)}{x + y} \right| \\ &= \left| \frac{(x + y + x\epsilon_1 + y\epsilon_2)(1 + \epsilon_3) - (x + y)}{x + y} \right| \\ &= \left| \frac{(x + y) + x\epsilon_1 + y\epsilon_2 + x\epsilon_3 + y\epsilon_3 + x\epsilon_1\epsilon_3 + y\epsilon_2\epsilon_3 - (x + y)}{x + y} \right| \end{aligned}$$

Trascuriamo il prodotto di errori, poiché numericamente irrilevanti (analisi del 1° ordine):

$$\approx \left| \frac{x}{x+y}\epsilon_1 + \frac{y}{x+y}\epsilon_2 + \frac{x+y}{x+y}\epsilon_3 \right| \leq \left| \frac{x}{x+y} \right| |\epsilon_1| + \left| \frac{y}{x+y} \right| |\epsilon_2| + |\epsilon_3| \not< 3u$$

Che cosa è successo? Non possiamo dire che l'errore sia sempre $< 3u$. Infatti, i fattori $\frac{x}{x+y}$ e $\frac{y}{x+y}$ agiscono come amplificatori degli errori ϵ_1 e ϵ_2 . Ma quando questi fattori diventano grandi? Quando $x, y \in \mathbb{R}$ sono di segno opposto e con valori quasi uguali.

Questo fenomeno è conosciuto come **errore di cancellazione numerica**. Si tratta di una perdita di precisione durante operazioni di addizione o sottrazione. Esso si verifica quando:

- x e y sono di segno opposto e con valori quasi uguali;
- vi è un errore ϵ_1 nella rappresentazione di x o un errore ϵ_2 in quella di y .

Anche se questi fattori fossero grandi, in assenza degli errori ϵ_1 o ϵ_2 , non avremmo un amplificazione dell'errore. Tuttavia, è proprio la presenza di errori nella rappresentazione, unita ai fattori di amplificazione, che può rendere il risultato finale meno preciso di quanto ci si potrebbe aspettare.

Esempio:

Sia $\mathbb{F}(10, 6, \lambda, \omega)$ con rappresentazione per arrotondamento e siano dati i numeri reali $\alpha = 0.147554326$ e $b = -0.147251742$.

La loro approssimazione nell'insieme \mathbb{F} sarà: $fl(a) = 0.1475543 + 0.0000005 = 0.147554$ e $fl(b) = 0.1472517 + 0.0000005 = 0.147252$.

L'addizione esatta darà: $a + b = 0.302584 \times 10^{-3}$, mentre in aritmetica finita darà: $fl(fl(a) + fl(b)) = fl(0.147554 - 0.147252) = 0.000302 = 0.302000 \times 10^{-3}$.

L'errore relativo commesso sarà: $\frac{0.302000 \times 10^{-3} - 0.302584 \times 10^{-3}}{0.302584 \times 10^{-3}} \approx 0.2 \times 10^{-2}$, mentre $u = \frac{1}{2}10^{1-6} = 0.5 \times 10^{-5}$ comportando un grave errore.

Nel cancellare delle cifre a causa dell'arrotondamento, ho perso delle informazioni successive. Questa perdita si traduce in un errore che è maggiorato di ben tre ordini di grandezza.

1.4.2 Condizionamento di un problema e stabilità di un algoritmo

Ci interessa comprendere dell'errore totale, quanto di quest'ultimo sia risultato dall'approssimazione dei dati e quanto invece sia dovuto all'algoritmo che stiamo utilizzando.

Iniziamo dividendo l'errore totale in due componenti:

- **Condizionamento del problema:** Questo rappresenta l'errore che è intrinsecamente associato ai dati di input del problema. In altre parole, è l'errore che non possiamo eliminare in quanto è legato alla qualità dei dati stessi. Possiamo chiamarlo **errore inherente**. Per quantificarlo, prendiamo il dato reale, lo approssimiamo e poi valutiamo l'errore relativo tra il risultato ottenuto in aritmetica esatta utilizzando il dato approssimato e quello che avremmo ottenuto utilizzando il dato vero.

$$E_{in} = \left| \frac{f(\tilde{x}) - f(x)}{f(x)} \right|$$

- **Stabilità dell'algoritmo:** Questo rappresenta l'errore introdotto dall'algoritmo che stiamo utilizzando per risolvere il problema. È il contributo dell'algoritmo nell'amplificare gli errori presenti nei dati. Possiamo chiamarlo **errore algoritmico**. Per valutarlo, confrontiamo il risultato finale ottenuto utilizzando l'algoritmo in aritmetica finita con il risultato teorico che l'algoritmo avrebbe fornito operando in aritmetica esatta, considerando che dati iniziali siano già stati approssimati.

$$E_{alg} = \left| \frac{\tilde{f}(\tilde{x}) - f(\tilde{x})}{f(\tilde{x})} \right|$$

Teorema. Siano x e \tilde{x} tale che $f(x) \neq 0$ ed $f(\tilde{x}) \neq 0$. Indicati con $E_{tot} = \left| \frac{\tilde{f}(\tilde{x}) - f(x)}{f(x)} \right|$ l'errore relativo nell'analisi in avanti e con E_{in} ed E_{alg} gli errori inerente e algoritmo, si ha

$$E_{tot} = E_{alg}(1 + E_{in}) + E_{in} \quad (1.2)$$

Trascurando il prodotto di errori, la 1.2 risulta così semplificata:

$$E_{tot} \approx E_{alg} + E_{in}$$

Un problema è definito come mal condizionato quando presenta un elevato errore inerente. Al contrario, quando l'errore inerente è ridotto, il problema è definito come ben condizionato. D'altro canto, se un algoritmo produce un ampio errore algoritmico, viene definito instabile. Se, invece, l'errore algoritmico è minimo, l'algoritmo è definito come numericamente stabile.

Esempio:

Calcolare gli errori inerente e algoritmico associati all'addizione di $x + y$, dove $x, y \in \mathbb{R}$.

$$\begin{aligned} \tilde{x} &= fl(x) = x(1 + \epsilon_1) \quad |\epsilon_1| < u \\ \tilde{y} &= fl(y) = y(1 + \epsilon_2) \quad |\epsilon_2| < u \\ E_{in} &= \left| \frac{f(\tilde{x}, \tilde{y}) - f(x, y)}{f(x, y)} \right| \\ &= \left| \frac{x(1 + \epsilon_1) + y(1 + \epsilon_2) - (x + y)}{x + y} \right| \\ &= \left| \frac{(x + y) + x\epsilon_1 + y\epsilon_2 - (x + y)}{x + y} \right| \\ &= \left| \frac{x}{x + y} \epsilon_1 + \frac{y}{x + y} \epsilon_2 \right| \\ &\leq \left| \frac{x}{x + y} \right| |\epsilon_1| + \left| \frac{y}{x + y} \right| |\epsilon_2| \\ E_{alg} &= \left| \frac{\tilde{f}(\tilde{x}, \tilde{y}) - f(\tilde{x}, \tilde{y})}{f(\tilde{x}, \tilde{y})} \right| \\ &= \left| \frac{(\tilde{x} + \tilde{y})(1 + \epsilon_3) - (\tilde{x} + \tilde{y})}{\tilde{x} + \tilde{y}} \right| \\ &= |\epsilon_3| \end{aligned}$$

Esempio:

Si analizzi il condizionamento e la stabilità dell'algoritmo utilizzato calcolare la funzione $f : \mathbb{R} \rightarrow \mathbb{R}$ definita come $f(x) = \frac{(1+x)-1}{x}$ e dove $x \in \mathbb{F} \subset \mathbb{R}$ e $x \neq 0$.

$$\begin{aligned}\tilde{x} &= fl(x) = x(1 + \epsilon_1) & |\epsilon_1| < u \\ E_{in} &= \left| \frac{f(\tilde{x}) - f(x)}{f(x)} \right| \\ &= \left| \frac{(1 + x(1 + \epsilon_1)) - 1}{x(1 + \epsilon_1)} - 1 \right| \\ &= 0\end{aligned}$$

Le operazioni principali e i loro errori associati sono:

- un'addizione con errore ϵ_1 ,
- una sottrazione con errore ϵ_2 , e
- una divisione con errore ϵ_3 .

$$\begin{aligned}E_{alg} &= \left| \frac{\tilde{f}(\tilde{x}) - f(\tilde{x})}{f(\tilde{x})} \right| \\ &= \left| \frac{\frac{(1+x)(1+\epsilon_1)-1)(1+\epsilon_2)}{x}(1+\epsilon_3) - 1}{1} \right| \\ &= \left| \frac{1+x+(1+x)\epsilon_1-\cancel{1}}{x}(1+\epsilon_2)(1+\epsilon_3) - 1 \right| \\ &= \left| \frac{x(1+\epsilon_2)(1+\epsilon_3)+(1+x)\epsilon_1(1+\epsilon_2)(1+\epsilon_3)}{x} - 1 \right| \\ &= \left| 1 + \epsilon_2 + \epsilon_3 + \epsilon_2\epsilon_3 + \frac{1+x}{x}\epsilon_1(1+\epsilon_2+\epsilon_3+\epsilon_2\epsilon_3) - 1 \right|\end{aligned}$$

Trascurando il prodotto di errori, poiché numericamente irrilevanti (analisi di 1° ordine):

$$\begin{aligned}&\approx \frac{1+x}{x}\epsilon_1 + \epsilon_2 + \epsilon_3 \\ \frac{1}{x}\epsilon_1 + \epsilon_1 + \epsilon_2 + \epsilon_3 &\leq \left| \frac{1}{x} \right| |\epsilon_1| + \underbrace{|\epsilon_1| + |\epsilon_2| + |\epsilon_3|}_{3u}\end{aligned}$$

Come interpretare questo risultato finale? Su quest'ultimo può esserci un errore $> 3u$. Se l'errore ϵ_1 nella prima addizione è $\neq 0$ e x è piccolo, il termine $\frac{1}{x}$ diventa grande, amplificando così l'errore.

1.4.3 Numero di condizione

Funzioni $f : \mathbb{R} \rightarrow \mathbb{R}$. Consideriamo una funzione $f : \mathbb{R} \rightarrow \mathbb{R}$ che sia differenziabile. Supponendo che vogliamo valutare questa funzione in un punto vicino a x_0 , possiamo usare lo sviluppo di Taylor centrato in x_0 :

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + o(h)$$

dove $h = x - x_0$.

Se vogliamo approssimare f in un punto vicino x_0 , diciamo \tilde{x}_0 abbiamo:

$$f(\tilde{x}_0) = f(x_0) + f'(x_0)(\tilde{x}_0 - x_0) + o(h)$$

Calcoliamo l'errore inerente:

$$E_{in} = \left| \frac{f(\tilde{x}_0) - f(x_0)}{f(x_0)} \right|$$

Sostituendo lo sviluppo di Taylor per $f(\tilde{x}_0)$, possiamo riscrivere:

$$\begin{aligned} &\approx \left| \frac{f(x_0) + f'(x_0)(\tilde{x}_0 - x_0) - f(x_0)}{f(x_0)} \right| \\ &= \left| \frac{f'(x_0)(\tilde{x}_0 - x_0)}{f(x_0)} \cdot \frac{x_0}{x_0} \right| \quad \text{per ipotesi } f(x_0) \neq 0 \text{ e } x_0 \neq 0 \\ &= \left| \frac{f'(x_0)x_0}{f(x_0)} \frac{\tilde{x}_0 - x_0}{x_0} \right| \end{aligned}$$

Se si considera $\frac{\tilde{x}_0 - x_0}{x_0}$ come l'errore sui dati, possiamo riscrivere:

$$= \left| \frac{f'(x_0)x_0}{f(x_0)} \right| |\epsilon_{x_0}|$$

Ciò che emerge è che l'errore inerente non dipende soltanto dall'errore sui dati, ma anche da una quantità $\left| \frac{f'(x_0)x_0}{f(x_0)} \right|$ che amplifica tale errore. Questa quantità è nota come **numero di condizione** e lo denotiamo con $C(f, x_0)$.

Un valore elevato di $C(f, x_0)$ indica che il problema è mal condizionato in x_0 , cioè piccoli errori nei dati possono portare a grandi errori nella soluzione.

Pertanto, per determinare se un problema differenziabile è mal condizionato in un dato punto, si può guardare il suo numero di condizione.

Generalizzazione per funzioni $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Dati una funzione $f : \mathbb{R}^n \rightarrow \mathbb{R}$ e due vettori $x = (x_1, x_2, \dots, x_n)$ e $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n)$.

L'espansione di Taylor di funzione in più variabili può essere espressa come:

$$f(\tilde{x}) = f(x) + \sum_{i=1}^n (\tilde{x}_i - x_i) \frac{\delta f}{\delta x_i} + o(h)$$

dove $\frac{\delta f}{\delta x_i}$ rappresenta la derivata parziale di f rispetto alla i -esima componente e $h = \sum_{i=1}^n (\tilde{x}_i - x_i)$.

Calcoliamo l'errore inerente:

$$E_{in} = \left| \frac{f(\tilde{x}) - f(x)}{f(x)} \right|$$

Sostituendo lo sviluppo di Taylor per $f(\tilde{x})$, possiamo riscrivere:

$$\begin{aligned} &\approx \left| \frac{f(x) + \sum_{i=1}^n (\tilde{x}_i - x_i) \frac{\delta f}{\delta x_i} - f(x)}{f(x)} \right| \\ &\leq \sum_{i=1}^n \left| \frac{(\tilde{x}_i - x_i) \frac{\delta f}{\delta x_i}}{f(x)} \right| \quad \text{per ipotesi } x_i \neq 0 \\ &= \sum_{i=1}^n \left| \frac{(\tilde{x}_i - x_i) \frac{\delta f}{\delta x_i}}{f(x)} \frac{x_i}{x_i} \right| \end{aligned}$$

Se si considera $\epsilon_i = \frac{\tilde{x}_i - x_i}{x_i}$ come gli errori sui dati e le quantità $c_i = \frac{\frac{\delta f}{\delta x_i} x_i}{f(x)}$ i numeri di condizione, possiamo riscrivere:

$$= \sum_{i=1}^n |c_i \epsilon_i|$$

Ricapitolando:

$$\left| \frac{f(\tilde{x}) - f(x)}{f(x)} \right| \leq \sum_{i=1}^n |c_i \epsilon_i| \quad (1.3)$$

dove

$$c_i = \frac{\frac{\delta f}{\delta x_i} x_i}{f(x)} \quad (1.4)$$

Se anche uno solo di questi numeri di condizione è grande, l'errore inerente sarà significativo. Per avere un errore inerente piccolo, è necessario che tutti i numeri di condizione siano piccoli.

Inoltre, grazie a quanto abbiamo visto, se la funzione è differenziabile, saremo in grado di stimare l'errore inerente più velocemente.

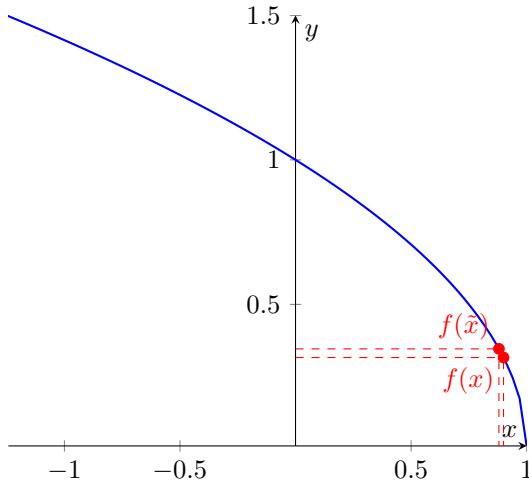
Esempio:

Consideriamo la funzione $f : \mathbb{R} \rightarrow \mathbb{R}$ definita come $f(x) = \sqrt{1-x}$, con $x \in \mathbb{R}$ e $x < 1$. La sua derivata è data da $f'(x) = -\frac{1}{2\sqrt{1-x}}$.

Calcoliamo il numero di condizione:

$$\left| \frac{x f'(x)}{f(x)} \right| = \left| \frac{x}{\sqrt{1-x}} \left(-\frac{1}{2\sqrt{1-x}} \right) \right| = \left| \frac{x}{2(1-x)} \right|$$

Guardando questa espressione, notiamo che quando x si avvicina a 1, il numero di condizione aumenta rapidamente, suggerendo che il problema è mal condizionato in prossimità di $x = 1$.



Dal grafico, possiamo vedere che quando x si avvicina molto a 1, anche piccole variazioni in x possono causare variazioni significative in $f(x)$. Ciò significa che errori piccoli in x possono portare a errori grandi in $f(x)$.

Per avere un'idea numerica di questo comportamento, consideriamo l'insieme $\mathbb{F}(10, 4, \lambda, \omega)$.

Siano:

$$x_0 = 0.99984$$

$$\tilde{x}_0 = 0.9998$$

Calcoliamo l'errore sui dati:

$$\left| \frac{\tilde{x}_0 - x_0}{x_0} \right| = \left| \frac{0.9998 - 0.99984}{0.99984} \right| \approx 4 \times 10^{-5}$$

L'errore inerente è dato da:

$$E_{in} = \left| \frac{f(\tilde{x}_0) - f(x_0)}{f(x_0)} \right| = \left| \frac{0.014142 - 0.0126491}{0.0126491} \right| \approx 0.1180$$

Il numero di condizione in x_0 è:

$$C(f, x_0) = C(\sqrt{1-x}, 0.99984) = \left| \frac{0.99984}{2(1-0.99984)} \right| \approx 0.312 \times 10^4$$

Osserviamo che l'errore sui dati si amplifica di 4 ordini di grandezza, a causa del numero di condizione, causando un grave errore sul risultato.

Esempio:

Si stimi l'errore inerente applicando 1.3 nei casi già visti di moltiplicazione e addizione fra numeri reali:

- **moltiplicazione** $f(x_1, x_2) = x_1 \cdot x_2$; applicando la 1.4 si ha

$$c_1 = \frac{x_1}{x_1 x_2} = 1 \quad c_2 = \frac{x_2}{x_1 x_2} x_1 = 1$$

da cui si deduce che il problema in oggetto è ben condizionato e risulta

$$E_{in} = \left| \frac{f(\tilde{x}) - f(x)}{f(x)} \right| \leq \sum_{i=1}^2 |c_i \epsilon_i| = |c_1 \epsilon_1| + |c_2 \epsilon_2| = 1 |\epsilon_1| + 1 |\epsilon_2|$$

- **addizione** $f(x_1, x_2) = x_1 + x_2$; applicando la 1.4 si ha

$$c_1 = \frac{x_1}{x_1 + x_2} 1 \quad c_2 = \frac{x_2}{x_1 + x_2} 1$$

da cui si deduce che il problema è mal condizionato per $x_1 + x_2 \rightarrow 0$ e risulta

$$E_{in} = \dots = \left| \frac{x_1}{x_1 + x_2} \right| |\epsilon_1| + \left| \frac{x_2}{x_1 + x_2} \right| |\epsilon_2|$$

Errore analitico. Supponiamo di avere una funzione $g : \mathbb{R} \rightarrow \mathbb{R}$ o $g : \mathbb{R}^n \rightarrow \mathbb{R}$ che non può essere risolta in aritmetica reale. Un esempio tipico potrebbe essere una funzione che calcola il seno di un numero, per la quale un calcolatore potrebbe non avere una definizione diretta.

Per trattare questo problema computazionalmente, possiamo sostituire $g(x)$ con una sua approssimazione $f(x)$ che, al contrario, è direttamente calcolabile. Un modo comune per approssimare questa funzione è utilizzare lo sviluppo di Taylor.

L'errore introdotto nell'usare $f(x)$ al posto di $g(x)$ è ciò che chiamiamo **errore analitico** ed è definito come:

$$E_{an} = \left| \frac{f(x) - g(x)}{g(x)} \right|$$

L'errore analitico si verifica quando si approssima un problema *irrazionale* (cioè, un problema non risolvibile in aritmetica reale) con un problema *razionale* o direttamente calcolabile.

Esempio:

Si esamina il problema della valutazione della seguente funzione:

$$f(x_1, x_2) = \sqrt{x_1 + x_2} - \sqrt{x_1} \quad \text{con } x_1, x_1 + x_2 \geq 0$$

Esistono delle condizioni dei dati in ingresso che possono causare errori gravi e che possono invalidare il risultato?

1. Se x_2 è estremamente piccolo in confronto a x_1 , si incorre in un errore di cancellazione numerica.
2. Se x_1 e x_2 sono di segno opposto e con valori quasi uguali, si incorre in un errore di cancellazione numerica.

Alla luce di ciò, si cerca un algoritmo differente che eviti il problema. Razionalizzando, otteniamo:

$$\begin{aligned} f(x_1, x_2) &= (\sqrt{x_1 + x_2} - \sqrt{x_1}) \frac{\sqrt{x_1 + x_2} + \sqrt{x_1}}{\sqrt{x_1 + x_2} + \sqrt{x_1}} \\ &= \frac{x_2}{\sqrt{x_1 + x_2} + \sqrt{x_1}} \end{aligned}$$

Ora, a denominatore si effettua un'addizione fra quantità non negative. Questo elimina il rischio di cancellazione per la 1° condizione, ma non per la 2° condizione.

Per convincerci di ciò, analizziamo l'errore inherente. Utilizziamo la stima 1.3: nel caso specifico sarà $E_{in} = c_1 \epsilon_1 + c_2 \epsilon_2$ con c_1 e c_2 calcolati tramite la 1.4; facendo i conti si ottiene:

$$c_1 = -\frac{1}{2} \sqrt{\frac{x_1}{x_1 + x_2}} \quad c_2 = \frac{1}{2} \left(1 + \sqrt{\frac{x_1}{x_1 + x_2}}\right) = \frac{1}{2} - c_1$$

e il problema risulta mal condizionato proprio quando $x_1 + x_2 \rightarrow 0$.

Illustriamo la situazione con un esempio numerico, dove con f denotiamo il valore esatto, mentre con \tilde{f} e \hat{f} , rispettivamente i risultati dei due algoritmi lavorando in $\mathbb{F}(10, 7, \lambda, \omega)$.

- Siano dati $x_1 = 0.1 \times 10^1$ e $x_2 = 0.1 \times 10^{-3}$, allora $c_2 \approx 1$ che indica un buon condizionamento in questo caso. I risultati esatto e calcolati con i due algoritmi sono:

$$f(x_1, x_2) = 0.4999875 \times 10^{-6}$$

$$\tilde{f}(x_1, x_2) = 0.4994869 \times 10^{-6}$$

$$\hat{f}(x_1, x_2) = 0.4999875 \times 10^{-6}$$

che indica l'instabilità del primo algoritmo nel caso di $|x_2| \ll |x_1|$.

- Siano dati $x_1 = 1$ e $x_2 = -1 + 10^{-6} = 0.999999$, allora $c_2 \approx 0.5 \times 10^{-3}$ che indica un cattivo condizionamento, infatti $x_1 + x_2 \rightarrow 0$. I risultati esatto e calcolati con i due algoritmi sono:

$$f(x_1, x_2) = -0.999$$

$$\tilde{f}(x_1, x_2) = \hat{f}(x_1, x_2) = -0.9985858$$

che mostra come entrambi gli algoritmi ci restituiscono risultati scadenti.

- Siano dati $x_1 = 0.1$ e $x_2 = 0.2$, allora $c_2 \approx 0.85$ che indica un buon condizionamento anche in questo caso. I risultati esatto e calcolati con i due algoritmi sono:

$$f(x_1, x_2) = \tilde{f}(x_1, x_2) = \hat{f}(x_1, x_2) = 0.1309858$$

2 Funzioni polinomiali

Definizione. Una funzione $p : \mathbb{R} \rightarrow \mathbb{R}$ definita da

$$p(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n = \sum_{i=0}^n a_i x^i \quad (2.1)$$

è detta funzione polinomiale, dove n è un intero non negativo detto grado e a_0, a_1, \dots, a_n sono numeri reali fissati detti coefficienti. Inoltre,

- se $a_n \neq 0$, si dice che $p(x)$ ha grado n ;
- se tutti i coefficienti a_i , sono nulli, allora $p(x)$ è detto polinomio nullo.

Con \mathbf{P}_n si denota l'insieme di tutte le funzioni polinomiali con grado $\leq n$, insieme al polinomio nullo. \mathbf{P}_n è uno spazio vettoriale di dimensione $n+1$. In uno spazio vettoriale di dimensione $n+1$, è necessario identificare $n+1$ funzioni polinomiali linearmente indipendenti per rappresentare univocamente tutte le altre funzioni all'interno di tale spazio come combinazione lineare di queste $n+1$ funzioni base.

È importante sottolineare che esistono infinite possibili basi per uno spazio vettoriale come \mathbf{P}_n .

Nel contesto del calcolo numerico, la scelta della base è fondamentale. Basi diverse corrispondono a coefficienti diversi, che possono influenzare l'errore inerente durante i calcoli.

Osservazione. Un polinomio è una funzione razionale, direttamente calcolabile su un calcolatore.

Teorema (fondamentale dell'algebra). Sia $p(x)$ un polinomio di grado $n \geq 1$. Allora, $p(x)$ ha esattamente n radici reali o complesse, ciascuna contata con la sua molteplicità. Ciò significa che $p(x) = 0$ può essere riscritto come:

$$p(x) = a_n(x - \alpha_1)^{m_1}(x - \alpha_2)^{m_2} \dots (x - \alpha_k)^{m_k} \quad (2.2)$$

dove:

- α_i (per $i = 1, \dots, k$) sono le radici distinte del polinomio,
- m_i (per $i = 1, \dots, k$) rappresenta la molteplicità della radice α_i ,
- e la somma totale delle molteplicità è $m_1 + m_2 + \dots + m_k = n$.

Il teorema fondamentale dell'algebra ci fornisce un altro modo per esprimere il polinomio.

Teorema. Siano $a(x)$ e $b(x)$ polinomi (dove $b(x)$ non è il polinomio nullo); allora è sempre possibile dividere $a(x)$ per $b(x)$ in modo da ottenere un quoziente $q(x)$ ed un resto $r(x)$. In altre parole, ogni polinomio $a(x)$ può essere espresso nella forma:

$$a(x) = q(x)b(x) + r(x)$$

con $r(x) = 0$ o $r(x)$ con grado minore di quello di $b(x)$.

2.1 Valutazione di un polinomio

Il primo problema che affronteremo riguarda la valutazione di un polinomio. Ciò significa determinare il valore del polinomio per un assegnato valore \bar{x} .

Formalmente, data la funzione:

$$f(a_0, a_1, \dots, a_n, \bar{x}) = a_0 + a_1\bar{x} + \dots + a_n\bar{x}^n$$

dove $f : \mathbb{R}^{n+2} \rightarrow \mathbb{R}$, vogliamo trovare il risultato di:

$$p(\bar{x}) = f(a_0, a_1, \dots, a_n, \bar{x})$$

In altre parole, inserendo i dati a_0, a_1, \dots, a_n e un valore specifico \bar{x} , vogliamo calcolare il valore del polinomio per quel valore.

Considerando una funzione polinomiale di grado n nella sua rappresentazione canonica 2.1, il metodo più immediato per la sua valutazione in corrispondenza di un assegnato valore \bar{x} può essere come segue:

```
p=a[0]
s=1
for k=1..n
    s=s*x_bar
    p=p+a[k]*s
```

Il calcolo di $p(\bar{x})$ richiede $2n$ moltiplicazioni ed n addizioni, con un costo asintotico $\mathcal{O}(n)$. Possiamo fare meglio? Se si scrive il polinomio 2.1 nella seguente forma:

$$\begin{aligned}
 p(x) &= a_0 + x(a_1 + a_2x + \dots + a_{n-1}x^{n-2} + a_nx^{n-1}) \\
 &= a_0 + x(a_1 + x(a_2 + a_3x + \dots + a_{n-1}x^{n-3} + a_nx^{n-2})) \\
 &\vdots \\
 &= a_0 + x(a_1 + x(a_2 + \dots + x(a_{n-1} + xa_n) \cdots))
 \end{aligned} \tag{2.3}$$

si ricava un differente metodo dovuto ad Horner:

```

p=a[n]
for k=n-1..0
    p=a[k]+x_bar*p

```

L'algoritmo di Horner richiede n moltiplicazioni ed n addizioni, con un costo asintotico $\mathcal{O}(n)$. Ancora, possiamo fare meglio?

Richiamando il teorema 2.2 sulla divisione di due polinomi, nel caso particolare in cui il polinomio divisore sia il binomio $(x - \bar{x})$ per un assegnato valore reale \bar{x} , si ha che esistono i polinomi unici $q(x)$ ed $r(x)$ per cui

$$p(x) = q(x)(x - \bar{x}) + r(x)$$

e poiché $r(x)$ deve essere di grado inferiore a quello di $(x - \bar{x})$ sarà una costante che indicheremo con r .

Teorema. *Il resto della divisione del polinomio $p(x)$ per $(x - \bar{x})$ è $p(\bar{x})$.*

Dal teorema appena enunciato si deduce che un metodo per valutare un polinomio $p(x)$ in un punto \bar{x} consiste nel determinare il resto della divisione fra $p(x)$ e il binomio $(x - \bar{x})$. A tal fine è ben nota la regola di Ruffini, la quale viene applicata seguendo il seguente schema di calcolo:

\bar{x}	a_n	a_{nb-1}	\dots	\dots	a_2	a_1	a_0
		$\bar{x}b_n$	$\bar{x}b_{n-1}$	\dots	$\bar{x}b_3$	$\bar{x}b_2$	$\bar{x}b_1$
	b_n	b_{n-1}	\dots	\dots	b_2	b_1	$r = p(\bar{x})$

- **Preparazione:**

- Scrivi i coefficienti del polinomio $p(x)$ in ordine decrescente di grado sulla prima riga. Se manca un termine di un certo grado, includi un coefficiente 0 per quel grado.
- Senza effettuare alcuna operazione, porta giù il primo coefficiente.

- **Procedimento:**

- Procedi con la compilazione della seconda riga e della terza riga. Moltiplica l'elemento appena riportato nella terza riga per il valore assegnato \bar{x} . Scrivi il risultato nella seconda riga, nella colonna successiva.
- Somma il coefficiente della prima riga con l'elemento presente nella seconda riga della colonna corrispondente e riporta il risultato nella terza riga della stessa colonna.

Reiterando il procedimento arriviamo all'ultimo elemento a destra sulla terza riga, che rappresenta il resto. Le operazioni fatte sono:

$$\begin{aligned}
 b_n &= a_n \\
 b_{n-1} &= a_{n-1} + \bar{x}b_n \\
 b_{n-2} &= a_{n-2} + \bar{x}b_{n-1} \\
 &\vdots \\
 r &= a_0 + \bar{x}b_1
 \end{aligned}$$

Come si può osservare, questo schema si traduce esattamente nello pseudocodice di Horner. Tuttavia, mentre Horner serve solamente a valutare il polinomio, Ruffini può essere anche utilizzato anche per valutare la derivata del polinomio.

2.1.1 Valutazione numerica della derivata

Se siamo interessati solo al valore della derivata in \bar{x} , allora, in modo più economico si può procedere nel seguente modo: per quanto detto nella sezione precedente, possiamo riscrivere $p(x)$ come

$$p(x) = q(x)(x - \bar{x}) + p(\bar{x})$$

derivando tale espressione si ha

$$p'(x) = q'(x)(x - \bar{x}) + q(x)$$

e valutandola in \bar{x}

$$\begin{aligned} p'(\bar{x}) &= q'(\bar{x}) \underbrace{(\bar{x} - \bar{x})}_0 + q(\bar{x}) \\ &= q(\bar{x}) \end{aligned}$$

dove $q(x)$ ed i suoi coefficienti sono quelli che si ottengono applicando l'algoritmo di Ruffini per valutare $p(x)$ in \bar{x} .

```
p=a[n]
p1=0
for k=n-1..0
    p1=p+x_bar*p1
    p=a[k]+x_bar*p
```

analogamente si possono calcolare anche le derivate di ordine superiore. Derivando l'espressione ottenuta per $p'(x)$ si ottiene:

$$p''(x) = q''(x)(x - \bar{x}) + 2q'(x)$$

e valutando questa espressione in \bar{x}

$$\begin{aligned} p''(\bar{x}) &= q''(\bar{x}) \underbrace{(\bar{x} - \bar{x})}_0 + 2q'(\bar{x}) \\ &= 2q'(\bar{x}) \end{aligned}$$

Allora per ottenere $p''(\bar{x})$ è sufficiente calcolare $q'(\bar{x})$ che possiamo ottenere da Ruffini utilizzando per valutare $p'(x)$ in \bar{x} .

Esempio:

Dato il polinomio $p(x) = 1 + x - 2x^2 + 3x^4$, vogliamo calcolare $p(2)$, $p'(2)$ e $p''(2)$. Per fare ciò, possiamo utilizzare la regola di Ruffini:

$$\begin{array}{c} \begin{array}{r|rrrr} & 3 & 0 & -2 & 1 \\ 2 & \hline & 6 & 12 & 20 \\ \hline & 3 & 6 & 10 & 21 \end{array} & 43 = p(2) \\[10pt] \begin{array}{r|rrr} & 3 & 6 & 10 \\ 2 & \hline & 6 & 24 \\ \hline & 3 & 12 & 34 \end{array} & 21 \\[10pt] \begin{array}{r|rr} & 3 & 12 \\ 2 & \hline & 6 \\ \hline & 3 & 18 \end{array} & 36 \\[10pt] 70 \rightarrow 2 * 70 = 140 = p''(2) \end{array}$$

Stima dell'errore algoritmico del metodo di Horner/Ruffini. Procediamo con un'analisi in avanti per determinare l'errore algoritmico associato al metodo di Horner/Ruffini. Questo metodo è utilizzato per calcolare il valore di un polinomio di grado n nella base canonica in un punto \bar{x} , rappresentato come $f(a_1, a_2, \dots, a_n, x) = p(x)$, dove $f : \mathbb{R}^2 \rightarrow \mathbb{R}$. Si ottiene:

$$E_{alg} = \left| \frac{\tilde{f}(\tilde{a}_0, \tilde{a}_1, \dots, \tilde{a}_n, \tilde{x}) - f(\tilde{a}_0, \tilde{a}_1, \dots, \tilde{a}_n, \tilde{x})}{f(\tilde{a}_0, \tilde{a}_1, \dots, \tilde{a}_n, \tilde{x})} \right| \leq \frac{\gamma 2n}{|p(\tilde{x})|} \sum_{i=0}^n |a_i \tilde{x}^i|$$

con

$$\gamma 2n \leq 2.01nu$$

Se ci limitassimo a considerare soltanto quest'ultimo, potremmo considerare l'algoritmo come stabile, dato che cresce linearmente con il grado del polinomio. Però, se i coefficienti o i valori di x fossero particolarmente grandi, ciò potrebbe causare un incremento significativo nell'errore algoritmico.

Stima dell'errore inherente del metodo di Horner/Ruffini. Assegnato un polinomio ed un punto in cui valutarlo, l'errore inherente misura a piccole variazioni sui coefficienti (dati), come varia, in senso relativo, il valore del polinomio (risultato).

Esempio:

Sia assegnato il polinomio

$$p(x) = a_0 + a_1 x = 100 - x$$

e lo si voglia valutare in punti $\bar{x} \in [100, 101]$. Si perturba il coefficiente a_1 dell'1%, cioè $\left| \frac{\tilde{a}_1 - a_1}{a_1} \right| = \frac{1}{100}$; segue che $\tilde{a}_1 = a_1 \pm \frac{1}{100}a_1$, per cui il polinomio perturbato risulta:

$$\tilde{p}(x) = 100 - (1 - \frac{1}{100})x = 100 - \frac{99}{100}x$$

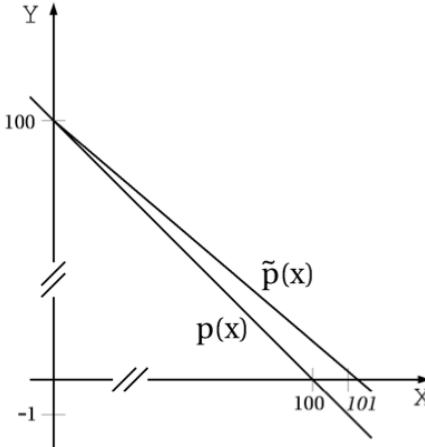
Valutandoli in $x = 101$ si ha:

$$p(101) = -1 \quad \tilde{p}(101) = 0.01$$

commettendo un errore relativo dato da:

$$\left| \frac{\tilde{p}(101) - p(101)}{p(101)} \right| = 101 \frac{1}{100}$$

Risulta, quindi che una perturbazione iniziale dell'1% porta una variazione sul risultato del 101%.



La causa di questa amplificazione dell'errore è evidente nel grafico sopra riportato. In pratica il coefficiente a_1 rappresenta l'inclinazione della retta la quale, anche se alterata minimamente, comporta grossi errori per punti lontani dall'origine.

Lo stesso comportamento si ottiene perturbando anche solo a_0 , entrambi i coefficienti o valutando in altri punti dell'intervallo.

Procediamo, per questo esempio, all'analisi dell'errore inherente mediante la stima 1.3:

$$\begin{aligned} E_{in} &= \left| \frac{f(\tilde{a}_0, \tilde{a}_1, \tilde{x}) - f(a_0, a_1, x)}{f(a_0, a_1, x)} \right| \leq |c_0 \epsilon_0| + |c_1 \epsilon_1| + |c_x \epsilon_x| \\ &= \left| \frac{a_0}{p(x)} \epsilon_0 \right| + \left| \frac{a_1 x}{p(x)} \epsilon_1 \right| + \left| \frac{x \alpha_1}{p(x)} \epsilon_x \right| \end{aligned}$$

e nel caso $a_0 = 100$, $a_1 = -1$ e $x \in [100, 101]$ con, per esempio, $x = 101$ si ha:

$$= \left| \frac{100}{-1} \epsilon_0 \right| + \left| \frac{-101}{-1} \epsilon_1 \right| + \left| \frac{-101}{-1} \epsilon_x \right|$$

da qui si vede come il problema è mal condizionato. Infatti una minima perturbazione su uno dei coefficienti, per esempio, a_1 dell'1% ($\epsilon_0 = 0, \epsilon_1 = 1/100, \epsilon_x = 0$), fa sì che l'errore relativo assuma il valore $101 \frac{1}{100}$ e cioè 101 volte maggiore di quello sul dato iniziale.

Dall'analisi fatta si evince che, per qualunque punto $x \in [100, 101]$, questo comportamento sarà inevitabile in quanto una lieve modifica dei coefficienti (ϵ_0 o $\epsilon_1 \neq 0$) viene grandemente amplificata (coefficienti c_0 e c_1).

Generalizzando l'analisi fatta in questo esempio alla valutazione di un generico polinomio $p(x)$ nella base canonica l'errore inerente si può rappresentare come la somma di due componenti:

$$\begin{aligned} E_{in} &\leq \left| \frac{a_0}{p(x)} \epsilon_0 \right| + \left| \frac{a_1 x}{p(x)} \epsilon_1 \right| + \left| \frac{a_2 x^2}{p(x)} \epsilon_2 \right| + \dots + \left| \frac{a_n x^n}{p(x)} \epsilon_n \right| + \left| \frac{x p'(x)}{p(x)} \epsilon_x \right| \\ &= E_{in1} + E_{in2} \end{aligned}$$

Si può desumere che:

- E_{in1} dipende da $p(x)$ e dai valori $a_i x^i$. Questi termini dipendono dalla base di rappresentazione;
- E_{in2} dipende dal valore di x , dal polinomio in esame $p(x)$ e dalla sua derivata, per cui è un errore che non dipende dalla base di rappresentazione.

Pertanto, l'espressione di un polinomio può incidere sull'errore inerente associato alla sua valutazione. Nel seguito vogliamo esaminare se è possibile cambiare la base di rappresentazione al fine di ridurre l'errore inerente. Consideriamo una base di \mathbb{P}_n costituita da $n+1$ polinomi linearmente indipendenti come $\{\phi_0(x), \phi_1(x), \dots, \phi_n(x)\}$; allora un generico polinomio $p(x)$ di grado al massimo n può essere espresso come una combinazione lineare di questi polinomi:

$$p(x) = \sum_{i=0}^n b_i \phi_i$$

Se ripetiamo l'analisi fatta per determinare E_{in} nel caso che $p(x)$ sia rappresentato nella base $\{\phi_i\}$, avremo:

$$E_{in} \leq \sum_{i=0}^n |c_i \epsilon_i| + |c_x \epsilon_x|$$

con

$$C_i = \frac{b_i \phi_i(x)}{p(x)} \quad C_x = \frac{x p'(x)}{p(x)}$$

L'idea è che, identificando una base di rappresentazione con valori "piccoli", potremmo ottenere un miglioramento sull'errore inerente E_{in1} .

Esempio:

Un primo esempio è rappresentato dalla base con centro $\{1, (x-c), (x-c)^2, \dots, (x-c)^n\}$ con $c \in [a, b]$.

2.2 Polinomi nella base di Bernstein

Si potrebbe chiedersi se esista una base che, indipendentemente dal polinomio considerato, il problema sia sempre ben condizionato. La risposta, in generale, è no: spesso la scelta della base ottimale dipende dallo specifico problema. Tuttavia, esiste una base, la base di Bernstein, che si distingue come particolarmente efficace e versatile rispetto ad altre.

Definizione. Con funzione polinomiale in forma di Bernstein si intende un'espressione del tipo

$$p(x) = \sum_{i=0}^n b_i B_{i,n}(x) \quad \text{con } x \in [a, b]$$

dove i $B_{i,n}(x)$ sono i polinomi base di Bernstein definiti sull'intervallo $[a, b]$ da

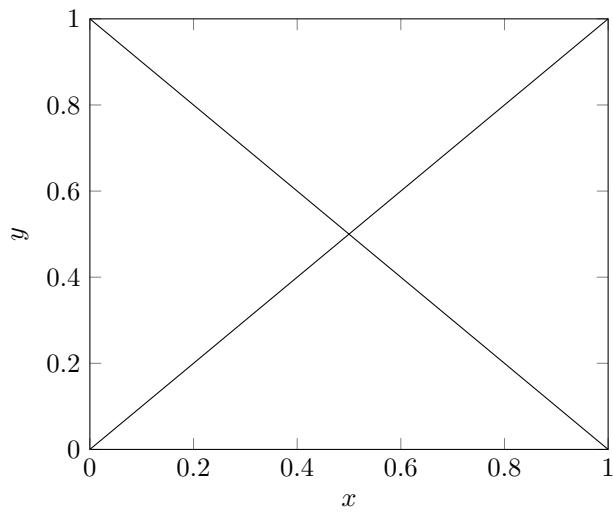
$$B_{i,n}(x) = \binom{n}{i} \frac{(b-x)^{n-i}(x-a)^i}{(b-a)^n} \quad (2.4)$$

e $b_0, b_1, \dots, b_n \in \mathbb{R}$ sono i coefficienti nella base di Bernstein.

con $\binom{n}{i} = \frac{n!}{i!(n-i)!}$.

- **n=1**

$$\begin{aligned} B_{0,1}(x) &= \binom{1}{0} \frac{(b-x)^1(x-a)^0}{b-a} = \frac{b-x}{b-a} \\ B_{1,1}(x) &= \binom{1}{0} \frac{(b-x)^0(x-a)^1}{b-a} = \frac{x-a}{b-a} \end{aligned}$$

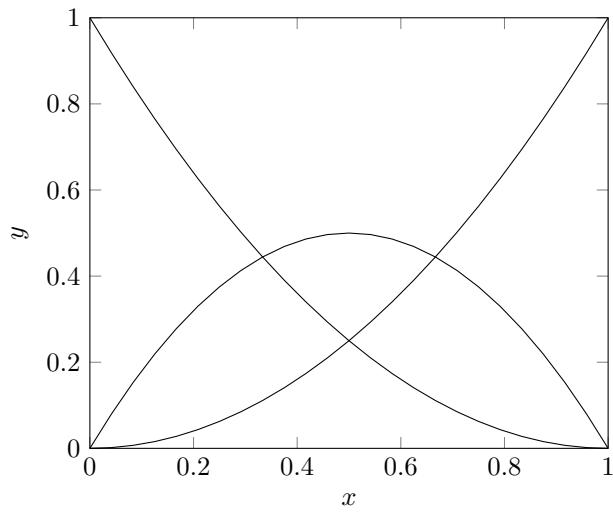


- n=2

$$B_{0,2}(x) = \binom{2}{0} \frac{(b-x)^2(x-a)^0}{(b-a)^2} = \frac{(b-x)^2}{(b-a)^2}$$

$$B_{1,2}(x) = \binom{2}{1} \frac{(b-x)^1(x-a)^1}{(b-a)^2} = \frac{(b-x)(x-a)}{(b-a)^2}$$

$$B_{2,2}(x) = \binom{2}{2} \frac{(b-x)^0(x-a)^2}{(b-a)^2} = \frac{(x-a)^2}{(b-a)^2}$$



Esempio:

Riprendiamo il polinomio $p(x) = 100 - x$ e lo riscriviamo nella base di Bernstein:

$$B_{0,1}(x) = \binom{1}{0} \frac{(101-x)1}{1} = 101 - x \quad B_{1,0}(x) = \binom{1}{1} \frac{1(x-100)}{1} = x - 100$$

da cui

$$\begin{aligned} p(x) &= 100 - x = b_0 B_{0,1}(x) + b_1 B_{1,0}(x) \\ &= b_0(101 - x) + b_1(x - 100) \end{aligned}$$

Ossia $b_0 = 0$ e $b_1 = -1$.

Rieseguendo l'analisi sull'errore inerente si ha:

$$\begin{aligned} E_{in} &\leq |c_0\epsilon_0| + |c_1\epsilon_1| + |c_x\epsilon_x| \\ &= \left| \frac{b_0(101-x)}{p(x)}\epsilon_0 \right| + \left| \frac{b_1(x-100)}{p(x)}\epsilon_1 \right| + \left| \frac{xp'(x)}{p(x)}\epsilon_x \right| \end{aligned}$$

valutandoli in $x = 101$ si ha:

$$\begin{aligned} &= \left| \frac{0(101-101)}{-1}\epsilon_0 \right| + \left| \frac{-1(1)}{-1}\epsilon_1 \right| + \left| \frac{101(-1)}{-1}\epsilon_x \right| \\ &= |\epsilon_1| + |101\epsilon_x| \end{aligned}$$

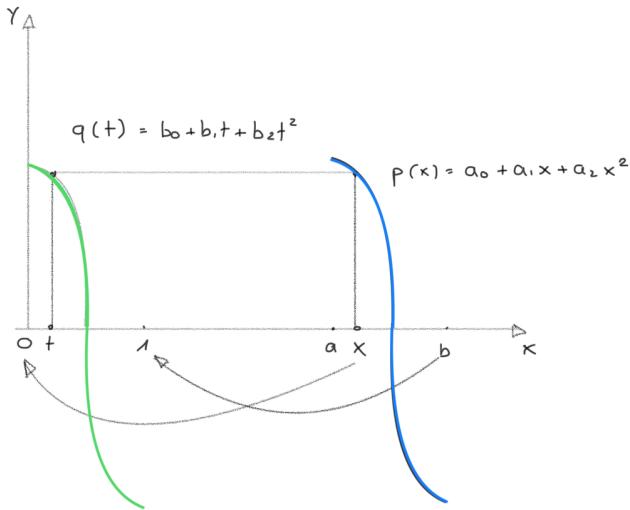
Siamo riusciti a rendere "piccoli" i numeri di condizioni che dipendono dalla base di rappresentazione. Tuttavia, il termine c_x , essendo indipendente dalla base di rappresentazione scelta, potrebbe causare un incremento significativo dell'errore inerente.

Considerando il problema evidenziato nell'esempio, si può fare qualcosa?

2.2.1 Cambio di variabile

I polinomi godono della proprietà di essere **invarianti per traslazione e scala dell'intervallo di definizione o cambio di variabile**. Questo significa che, per qualsiasi dato polinomio $p(x)$, definito in un intervallo $[a, b]$, è sempre possibile individuare un altro polinomio che assume gli stessi valori all'interno di un intervallo che è stato traslato o scalato. Questo permette di definire un'applicazione (mapping) tra i due intervalli $[a, b]$ e, per esempio, $[0, 1]$

$$\begin{aligned} x \in [a, b] &\rightarrow t \in [0, 1] \\ t = \frac{x-a}{b-a} \text{ o viceversa } x &= a + t(b-a) \end{aligned} \tag{2.5}$$



Per effettuare un cambio di variabile, possiamo sostituire, all'interno del polinomio, x con $a+t(b-a)$. Otteniamo un nuovo polinomio espresso in termini di t piuttosto che di x .

Consideriamo un polinomio $p(x)$ espresso nella base di Bernstein:

$$p(x) = \sum_{i=0}^n b_i B_{i,n}(x) \quad \text{con } x \in [a, b]$$

Vogliamo effettuare un cambio di variabile da x a t , con $t \in [0, 1]$. Riscriviamo la 2.4 nel seguente modo:

$$= \binom{n}{i} \left(\frac{b-x}{b-a} \right)^{n-i} \left(\frac{x-a}{b-a} \right)^i$$

Sostituendo x con $a + t(b - a)$, otteniamo:

$$\begin{aligned} &= \binom{n}{i} \left(\frac{(b-a)(1-t)}{b-a} \right)^{n-i} \left(\frac{a+t(b-a)-a}{b-a} \right)^i \\ &= \binom{n}{i} (1-t)^{n-i} t^i = B_{i,n}(t) \end{aligned}$$

Il cambio di variabile ha trasformato la base di Bernstein in funzione di t , mantenendo inalterati i coefficienti b_i . Pertanto, il polinomio dopo il cambio di variabile è:

$$p(t) = \sum_{i=0}^n b_i B_{i,n}(t) \quad \text{con } t \in [0, 1]$$

Quindi un cambio di variabile, per un polinomio nella base di Bernstein, non comporta alcun errore o costo computazionale sui coefficienti in quanto restano uguali.

Ma come ci aiuta esattamente il cambio di variabile a ridurre il numero di condizione c_x ?

Consideriamo la trasformazione:

$$\frac{xp'(x)}{p(x)} \rightarrow \frac{tp'(t)}{p(t)}$$

Se effettuiamo un cambio di variabile da x a t , dove t è definito nell'intervallo $[0, 1]$, stiamo effettivamente ridimensionando la variabile x . Questa operazione ha un impatto anche sulla sua derivata. Infatti, il ridimensionamento di questo intervallo, per effetto della scala, potrebbe attenuare o "rilassare" la derivata del polinomio.

Esempio:

Riprendiamo il polinomio $p(x) = 100 - x$. Questo può essere espresso nella base di Bernstein come $p(x) = -1 \underbrace{(x-100)}_{B_{1,1}(x)}$ con $x \in [100, 101]$. Ora, effettuiamo il cambio di variabile in $t \in [0, 1]$.

I polinomi di Bernstein in termini di t sono:

$$B_{0,1}(t) = 1 - t \quad B_{1,1}(t) = t$$

In termini di t , il polinomio $p(x)$ diventa:

$$q(t) = -1 \underbrace{\frac{(t)}{B_{1,1}(t)}}$$

Se valutiamo in $x = 101$, questo corrisponde a $t = 1$. Calcoliamo ora, c_t :

$$c_t = \frac{tq'(t)}{q(t)} = \frac{1}{-1}(-1) = 1$$

Grazie al cambio di variabile, c_t è stato ridotto a un valore che non amplifica l'errore.

2.2.2 Proprietà dei polinomi di Bernstein

Da ora in poi useremo sempre i polinomi di Bernstein nell'intervallo $[0, 1]$.

Definizione. Il polinomio di Bernstein definito sull'intervallo $[0, 1]$ è dato da:

$$B_{in} = \binom{n}{i} (1-x)^{n-i} \cdot x^i \quad \text{con } x \in [0, 1]$$

Inoltre, i polinomi della base di Bernstein $B_{i,n}(x)$ godono delle seguenti proprietà:

1. $B_{i,n} \geq 0 \quad i = 0, \dots, n \quad \forall x \in [0, 1];$

- Il valore del polinomio è sempre non negativo;

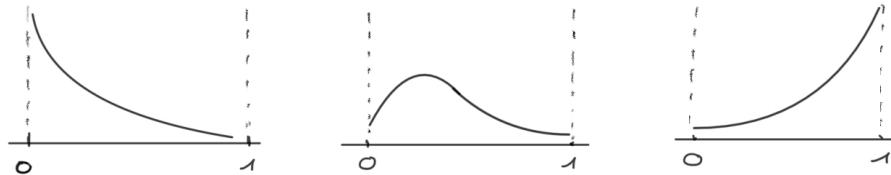
- $B_{i,n} = 0$ agli estremi dell'intervallo.

$$B_{0,n}(x) = (1-x)^n \quad \text{tutti gli zero in 1 e nessuno zero in 0}$$

$$B_{1,n}(x) = \binom{n}{1}(1-x)^{n-1}x^1 \quad \text{uno zero in 0 ed } n \text{ zeri in 1}$$

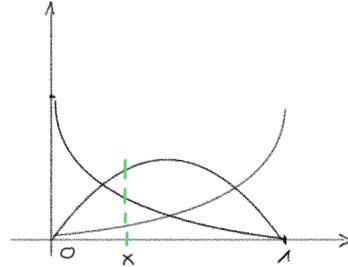
⋮

$$B_{n,n}(x) = x^n \quad \text{tutti gli zero in 0 e nessuno zero in 1}$$



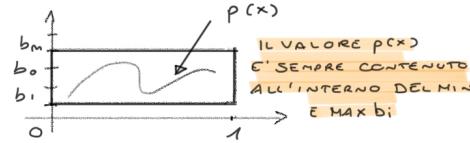
2. **Partizione dell'unità:** per ogni x nell'intervallo $[0, 1]$, la somma dei valori dei polinomi di Bernstein è sempre uguale a 1.

$$\sum_{i=0}^n B_{i,n}(x) = 1 \quad \forall x \in [0, 1]$$



3. Per le proprietà precedenti, $p(x)$ espresso nella base di Bernstein è una combinazione convessa dei b_i , da cui segue

$$\min_i \{b_i\} \leq p(x) \leq \max_i \{b_i\}$$



Ciò suggerisce che affinché il polinomio $p(x)$ possa avere zeri reali, i coefficienti b_i devono essere sia positivi che negativi.

4. Valutando il polinomio in $x = 0$, solo $B_{0,n}(0)$ ha valore 1, mentre tutti gli altri polinomi di Bernstein sono nulli. Di conseguenza, $p(0)$ corrisponde al coefficiente b_0 . Analogamente, in $x = 1$, $p(1)$ coincide con b_n dato che solo $B_{n,n}(1)$ è non nullo e vale 1.

$$p(0) = b_0 \quad p(1) = b_n$$

Ma se volessimo valutare un polinomio nella base di Bernstein in un punto interno, anziché agli estremi? Poiché gli algoritmi di valutazione polinomiale che abbiamo visto finora sono pensati per la base canonica, è necessario introdurre due nuovi algoritmi specificamente pensati per la base di Bernstein.

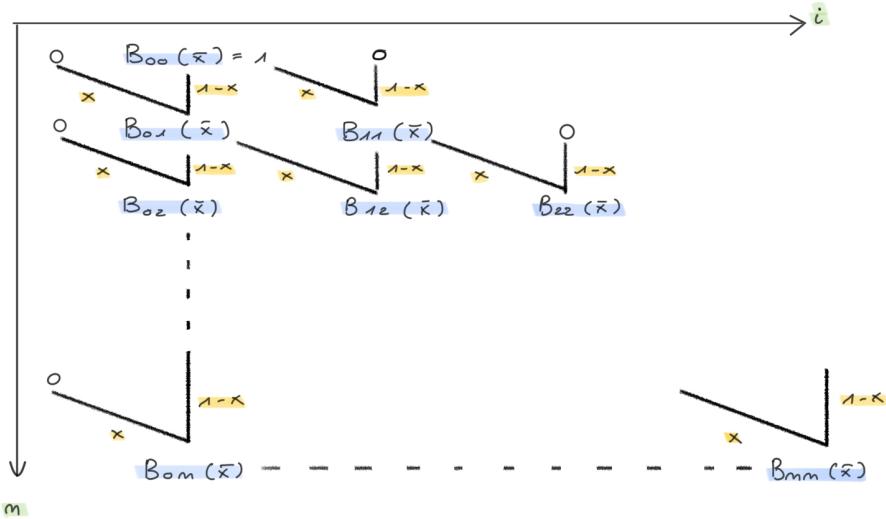
2.2.3 Valutazione di un polinomio nella base di Bernstein

Formula ricorrente base di Bernstein. I polinomi base di Bernstein di grado n , sono determinabili dai polinomi di Bernstein di grado $n - 1$, dalla seguente formula ricorrente

$$B_{i,n}(x) = xB_{i-1,n-1}(x) + (1-x)B_{i,n-1}(x)$$

con $B_{0,0}(x) = 1$ e $B_{i,n}(x) = 0$ per ogni $i \notin \{0, n\}$.

- Passi:
- Inizializzazione:** si crea un vettore B di lunghezza $n + 1$, inizializzato con tutti zeri tranne il primo elemento ($B_{0,0}$) che sarà 1. Questo vettore conterrà i valori dei polinomi base di Bernstein.
 - Valutazione delle funzioni base di Bernstein:** si calcolano iterativamente i valori dei polinomi base di Bernstein di grado n usando la formula ricorrente.



- Calcolo di $p(x) = \sum_{i=0}^n b_i B_{in}(\bar{x})$:** si esegue il prodotto scalare del vettore riga b e il vettore colonna B^T per ottenere il valore del polinomio in \bar{x} .

```

1 def bernstein_evaluation(b, x_bar):
2     n = len(b)
3     B = np.zeros((n,n))
4     B[0][0] = 1.
5
6     for i in range(1,n):
7         for j in range(0,i+1):
8             if j == 0:
9                 B[i][j] = (1-x_bar) * B[i-1][j]
10            elif j == i:
11                B[i][j] = x_bar * B[i-1][j-1]
12            else:
13                B[i][j] = x_bar * B[i-1][j-1] + (1-x_bar)*B[i-1][j]
14    return sum([b[i] * B[n-1][i] for i in range(n)])

```

Quando si utilizza l'algoritmo per valutare un polinomio nella base di Bernstein:

- Il calcolo del valore dei polinomi base $B_{i,n}(\bar{x})$ per $i = 0, \dots, n$ necessita di $\frac{n(n+1)}{2}$ addizioni e $n(n + 1)$ moltiplicazioni.
- Per ottenere il valore finale del polinomio in \bar{x} attraverso la combinazione lineare dei polinomi base e dei coefficienti del polinomio, sono necessarie ulteriori n addizioni e n moltiplicazioni

In totale, l'algoritmo ha un costo computazionale di $\mathcal{O}(n^2)$. Anche se questo costo potrebbe sembrare alto, va sottolineato che l'algoritmo è numericamente stabile, specialmente, nel calcolo dei polinomi base.

Algoritmo di de Casteljau. Un’alternativa all’algoritmo proposto è l’algoritmo di de Casteljau, che deriva dall’applicare ripetutamente la formula ricorrente all’espressione del polinomio. Vediamolo:

$$\begin{aligned} p(x) &= \sum_{i=0}^n b_i B_{i,n}(x) \stackrel{\text{def}}{=} \sum_{i=0}^n b_i (xB_{i-1,n-1}(x) + (1-x)B_{i,n-1}(x)) \\ &= \sum_{i=0}^n b_i x B_{i-1,n-1}(x) + \sum_{i=0}^n b_i (1-x) B_{i,n-1}(x) \end{aligned}$$

A questo punto, dobbiamo notare che il termine $B_{-1,n-1}$ non esiste. Inoltre, dato che la sommatoria effettivamente va solo fino a $n-1$, il termine $B_{n,n-1}=0$. Dunque, regolando gli indici, abbiamo:

$$= \sum_{i=0}^{n-1} b_{i+1} x B_{i,n+1}(x) + \sum_{i=0}^{n-1} b_i (1-x) B_{i,n-1}(x)$$

Raccogliendo:

$$\begin{aligned} &= \sum_{i=0}^{n-1} [b_{i+1}x + b_i(1-x)] B_{i,n-1}(x) \\ &= \sum_{i=0}^{n-1} b_i^{[1]} B_{i,n-1}(x) \end{aligned}$$

Riapplicando la formula ricorrente più volte, alla fine si ottiene:

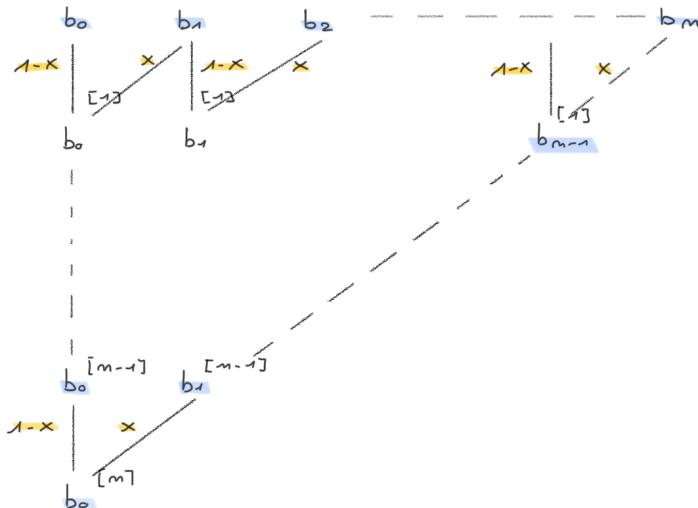
$$= \sum_{i=0}^0 b_0^{[n]} \overbrace{B_{0,0}(x)}^1 = b_0^{[n]}$$

Si ha quindi il seguente algoritmo:

$$b_i^{[j]} = xb_{i+1}^{[j-1]} + (1-x)b_i^{[j-1]} \quad (2.6)$$

con $j = 0, \dots, n$ e $i = 0, \dots, n-j$.

Si inizia con i coefficienti iniziali del polinomio e si applica iterativamente l’equazione 2.6 fino a quando non si ottiene $b_0^{[n]}$. Quest’ultimo valore corrisponde al valore del polinomio quando viene valutato in x .



```

1 def de_casteljau(b, x_bar):
2     if len(b) == 1:
3         return b[0]
4
5     b_new = []
6     for i in range(len(b)-1):
7         b_new.append(x_bar * b[i+1] + (1-x_bar) * b[i])
8
9     return de_casteljau(b_new, x_bar)

```

Quando si utilizza l’algoritmo di Casteljau per valutare un polinomio nella base di Bernstein:

- Il calcolo del coefficiente $b_0^{[n]}$ necessita di $\frac{n(n+1)}{2}$ addizioni e $n(n+1)$

Il costo totale dell'algoritmo è di $\mathcal{O}(n^2)$. Tuttavia, è meno costoso rispetto al precedente algoritmo. Ciò è dovuto al fatto che l'algoritmo di Casteljau elimina la necessità di calcolare esplicitamente la combinazione lineare per determinare il valore di $p(x)$.

Esempio:

Si consideri il polinomio nella base di Bernstein:

$$p(x) = 2 \overbrace{B_{0,3}(x)}^{b_0} + 2 \overbrace{B_{2,3}(x)}^{b_2} + 0 \overbrace{B_{a1,3}(x)}^{b_1} + 0 \overbrace{B_{3,3}(x)}^{b_3}$$

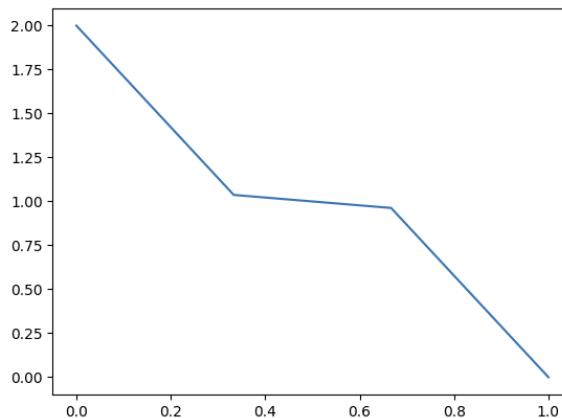
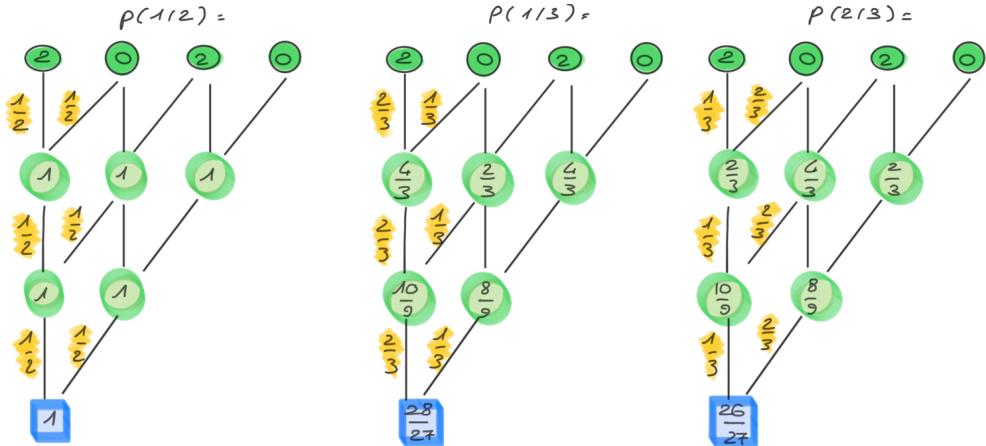
Applicare l'algoritmo di de Casteljau ai seguenti punti:

$$x = \left[0, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 1\right]$$

Infine, disegnare un grafico per i punti $p(x)$.

Eseguendo l'algoritmo, otteniamo:

- $p(0) = 2, p(1) = 0$ dati dalle proprietà dei polinomi di Bernstein;
- $p(\frac{1}{2}) = 1, p(\frac{1}{3}) = \frac{28}{27}, p(\frac{2}{3}) = \frac{26}{27}$.



I polinomi base di Bernstein rappresentano una delle basi preferite per la rappresentazione di curve, in particolare per le curve di Bézier. Una curva 2D è definita come segue:

$$C(t) = (f(t), g(t)), \quad \text{dove } t \in [0, 1].$$

dove $t \in [0, 1]$. Ogni t individua un punto sulla curva.

Utilizzando i polinomi di Bernstein come base, possiamo esprimere le funzioni $f(t)$ e $g(t)$ come segue:

$$f(t) = \sum_{i=0}^n x_i B_{in}(t) \quad g(t) = \sum_{i=0}^n y_i B_{in}(t)$$

In termini vettoriali, la curva può essere descritta come una funzione vettoriale nella base di Bernstein:

$$C(t) = \sum_{i=0}^n P_i B_{in}(t)$$

dove $P_i = (x_i, y_i)$ rappresentano i punti di controllo.

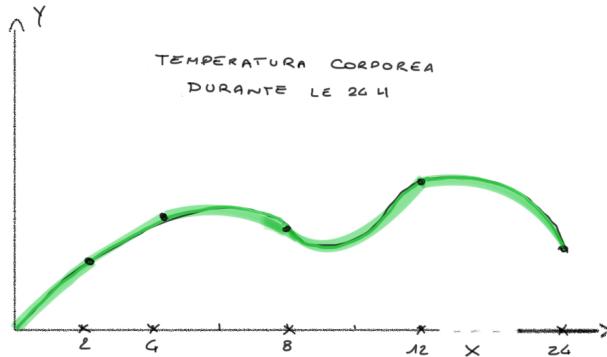
Le curve di Bézier sono utilizzate nel disegno su calcolatore e nei font vettoriali. In questi font, un carattere è definito da curve. A differenza delle immagini raster, i disegni vettoriali mantengono nitidezza a qualsiasi livello di ingrandimento. Così, zoomando su un carattere o un disegno vettoriale, non si perde qualità.

3 Interpolazione polinomiale

Interpolazione di dati. Il problema dell'interpolazione di dati consiste nel trovare un polinomio che passi per un insieme di punti (x_i, y_i) dati. Più formalmente, si cerca un polinomio $p(x)$ appartenente all'insieme dei polinomi \mathbb{P}_n tale che $p(x_i) = y_i$ per ogni $i = 0, \dots, n$.

In questo problema, abbiamo a disposizione gli y_i - i valori che desideriamo che il nostro polinomio assuma - e gli x_i - i punti corrispondenti a questi valori. Ciò che manca sono i coefficienti del polinomio, indicati con a . Potremmo considerare l'interpolazione come il problema inverso della valutazione di un polinomio. Nella valutazione, abbiamo i coefficienti del polinomio e gli x_i dove desideriamo valutare il polinomio. Invece, nell'interpolazione, abbiamo i valori y_i del polinomio e gli x_i , ma non conosciamo i coefficienti.

In pratica, i dati potrebbero provenire, ad esempio, da misurazioni sperimentali di un fenomeno e vogliamo rappresentarli con una funzione. Una volta identificato un polinomio interpolante adatto, questo può essere utilizzato per prevedere il comportamento della funzione all'interno dell'intervallo di interesse, incluso in punti che non sono stati originariamente misurati.



Interpolazione di funzioni. Il problema dell'interpolazione di funzioni mira a trovare un polinomio che approssima una data funzione $f(x)$ in un intervallo $x \in [a, b]$. Questo comporta:

1. Selezione dei punti dell'interpolazione:

- A differenza dell'interpolazione di dati dove gli x_i sono dati, qui scegliamo autonomamente i punti x_i nell'intervallo $[a, b]$, con $i = 0, \dots, n$.
- La scelta di questi punti è fondamentale: una selezione appropriata può portare a una rappresentazione accurata della funzione originale, mentre una scelta inadeguata può portare un'approssimazione imprecisa.

2. Costruzione del polinomio:

- Si determina un polinomio $p(x) \in \mathbb{P}_n$ tale che $p(x_i)$ coincida con $f(x_i)$ per ogni $i = 0, \dots, n$.

3. Valutazione dell'errore:

- Una volta costruito il polinomio, si valuta l'errore $|f(x) - p(x)|$ per ogni $x \in [a, b]$ per misurare l'accuratezza dell'approssimazione.

3.1 Esistenza e unicità dell'interpolazione polinomiale

Teorema. *Dati $n+1$ punti (x_i, y_i) , $i = 0, \dots, n$ con x_i distinti, allora esiste ed è unico il polinomio $p \in \mathbb{P}_n$ che verifica le condizioni*

$$p(x_i) = y_i \quad i = 0, \dots, n$$

Dimostrazione.

Si consideri $p(x)$ in forma canonica

$$p(x) = a_0 + a_1 x + \dots + a_n x^n$$

Imponendo le $n+1$ condizioni di interpolazione, cioè che $p(x_i) = y_i$ con $i = 0, \dots, n$, otteniamo:

$$\begin{cases} a_0 + a_1 x_0 + a_2 x_0^2 + \dots + a_n x_0^n &= y_0 \\ a_0 + a_1 x_1 + a_2 x_1^2 + \dots + a_n x_1^n &= y_1 \\ \vdots \\ a_0 + a_1 x_n + a_2 x_n^2 + \dots + a_n x_n^n &= y_n \end{cases}$$

Cioè un sistema lineare di $n+1$ equazioni in $n+1$ incognite; se tale sistema ammette una ed una sola soluzione, allora tale sarà il polinomio. In forma matriciale il sistema si presenterà come:

$$Va = y \quad (3.1)$$

con

$$V = \begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \dots & & & & \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{bmatrix} \quad a = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} \quad y = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}$$

il sistema $Va = y$ ammette una e una sola soluzione se e solo se la matrice V , conosciuta come matrice di Vandermonde, è non singolare, il che si verifica se e solo se il $\det(V) \neq 0$. Il determinante della matrice di Vandermonde può essere espresso come:

$$\det(V) = \prod_{i,j=0, j>i}^n (x_j - x_i)$$

che nelle ipotesi che i punti x_i siano distinti risulta sempre non nullo. Ciò dimostra che il sistema lineare ha un'unica soluzione e quindi $p(x)$ esiste ed è unico. ■

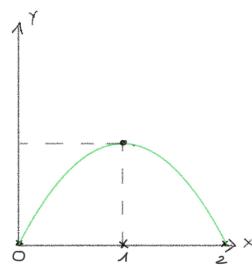
La dimostrazione del Teorema fornisce anche un metodo per procedere alla determinazione del polinomio interpolante risolvendo il sistema lineare.

Esempio:

Siano dati il set di punti $x = [0, 1, 2]$, $y = [0, 1, 0]$. Si determini il polinomio che interpola i dati. In base al teorema visto, il polinomio che interpola questi dati esiste ed è unico, e può essere trovato in \mathbb{P}_2 , ovvero $p(x) = a_0 + a_1x + a_2x^2$.

$$\begin{aligned} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} &= \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\ &= \left[\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 0 \end{array} \right] \\ &= \left[\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ \frac{3}{4} & \frac{1}{2} & 0 & 1 \\ 1 & 2 & 4 & 0 \end{array} \right] \\ &= \left[\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 1 \\ 0 & 2 & 4 & 0 \end{array} \right] \\ &= \left[\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 2 \\ 0 & 0 & 4 & -4 \end{array} \right] \\ &= \left[\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & -1 \end{array} \right] \\ p(x) &= 0 + 2x - x^2 \end{aligned}$$

Si osserva inoltre che il polinomio trovato è il polinomio di grado MINIMO che interpola i dati in \mathbb{P}_n .



3.2 Metodi di costruzione

Ora esamineremo vari metodi di interpolazione polinomiale, ciascuno basato su una diversa base polinomiale, allo scopo di identificare quale tra queste sia la migliore.

3.2.1 Base di Newton

Siano dati $n + 1$ punti distinti (x_i, y_i) con $i = 0, \dots, n$. Il polinomio interpolante nella forma di Newton è costruito su una particolare base chiamata base di Newton, che viene costruita sui punti x_i , $i = 0, \dots, n$.

La base di Newton è costituita da $n + 1$ funzioni base linearmente indipendenti

$$1, (x - x_0), (x - x_0)(x - x_1), \dots, (x - x_0)(x - x_1) \cdots (x - x_{n-1})$$

e forma una base polinomiale per lo spazio \mathbb{P}_n .

Il polinomio interpolante nella base di Newton può essere scritto come combinazione lineare di queste funzioni base:

$$p(x) = b_0 + b_1(x - x_0) + b_2(x - x_0)(x - x_1) + \dots + b_n(x - x_0)(x - x_1) \cdots (x - x_n)$$

Imponendo le condizioni di interpolazione a questo polinomio si ottiene un sistema di $n + 1$ equazioni in $n + 1$ incognite:

$$\begin{aligned} p(x_0) = y_0 &\Rightarrow b_0 + b_1 \overbrace{(x_0 - x_0)}^0 + b_2 \overbrace{(x_0 - x_0)}^0 (x_0 - x_1) + \dots + b_n \overbrace{(x_0 - x_0)}^0 (x_0 - x_1) \cdots (x_0 - x_{n-1}) = y_0 \\ &\qquad\qquad\qquad b_0 = y_0 \\ p(x_1) = y_0 &\Rightarrow b_0 + b_1(x_1 - x_0) + b_2(x_1 - x_0) \overbrace{(x_1 - x_1)}^0 + \dots + b_n(x_1 - x_0) \overbrace{(x_1 - x_1)}^0 \cdots (x_1 - x_{n-1}) = y_0 \\ &\qquad\qquad\qquad b_0 + b_1(x_1 - x_0) = y_0 \\ &\vdots && \vdots \\ p(x_n) = y_n &\Rightarrow b_0 + b_1(x_n - x_0) + b_2(x_n - x_0)(x_n - x_1) + \dots + b_n(x_n - x_0)(x_n - x_1) \cdots (x_n - x_{n-1}) = y_n \end{aligned}$$

Questo sistema può essere espresso in forma matriciale come:

$$Nb = y$$

con

$$N = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & (x_1 - x_0) & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 1 & (x_n - x_0) & (x_n - x_0)(x_n - x_1) & \dots & (x_n - x_0) \cdots (x_n - x_{n-1}) \end{bmatrix} \quad c = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix} \quad y = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}$$

Poiché la matrice N è triangolare inferiore, il sistema è facilmente risolvibile per sostituzione in avanti. Ciò fornisce i coefficienti b_i , che rappresentano il polinomio che intercala i dati.

```

1 def calculate_newton_base(x_bar, x):
2     return np.prod([x_bar - x_i for x_i in x])
3
4 def newton_interpolation(x, y):
5     n = len(x)
6     L = np.zeros((n,n))
7
8     L[:,0] = 1
9     for i in range(1,n):
10        for j in range(1, i+1):
11            L[i][j] = calculate_newton_base(x[i], x[:j])
12
13    return forward_substitution(L, y)

```

Esempio:

Siano dati nuovamente il set di punti $x = [0, 1, 2]$, $y = [0, 1, 0]$. Vogliamo determinare il polinomio che interpola questi dati utilizzando la base di Newton.

Il polinomio interpolante si trova in \mathbb{P}_2 ed ha la seguente forma:

$$\begin{aligned} p(x) &= b_0 + b_1(x - x_0) + b_2(x - x_0)(x - x_1) \\ &= b_0 + b_1(x - 0) + b_2(x - 0)(x - 1) \\ &= b_0 + b_1x + b_2x(x - 1) \end{aligned}$$

Possiamo calcolare i coefficienti risolvendo il seguente sistema lineare:

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$\begin{array}{c|cc|c} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 2 & 2 & 0 \end{array}$$

$$\begin{bmatrix} 1 & 0 & 0 & | & 0 \\ 0 & 1 & 0 & | & 1 \\ 0 & 0 & 1 & | & -1 \end{bmatrix}$$

$$p(x) = 0 + x - x(x - 1)$$

Se cambiato l'ordine dei punti in $x = [0, 2, 1]$, $y = [0, 0, 1]$, il polinomio risultante rimane lo stesso. Il teorema, infatti, ci garantisce l'esistenza e unicità di un polinomio che passa attraverso quei punti specifici, indipendentemente dall'ordine. Quello che cambia è la base di rappresentazione utilizzata per esprimere il polinomio. Anche se il polinomio rimane invariato, i coefficienti possono cambiare a seconda dell'ordine dei punti. Quindi analiticamente, il polinomio non cambia, ma numericamente potrebbe cambiare.

$$\begin{aligned} p(x) &= b_0 + b_1(x - 0) + b_2(x - 0)(x - 2) \\ &\quad \begin{bmatrix} 1 & 0 & 0 & | & 0 \\ 1 & 2 & 0 & | & 0 \\ 1 & 1 & -1 & | & 1 \end{bmatrix} \\ &\quad \begin{array}{c|cc|c} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 1 & -1 & 1 \end{array} \\ &\quad \begin{bmatrix} 1 & 0 & 0 & | & 0 \\ 0 & 1 & 0 & | & 0 \\ 0 & 0 & 1 & | & -1 \end{bmatrix} \\ &\quad p(x) = -x(x - 2) \end{aligned}$$

Soluzione di un sistema lineare con matrice triangolare inferiore. Dato un sistema lineare con una matrice triangolare inferiore L , possiamo risolvere il sistema seguendo questi passaggi:

1. Dalla prima equazione, calcoliamo l'incognita $b_1 = \frac{y_1}{l_{11}}$.
2. Sostituiamo b_1 nelle equazioni successive e aggiorniamo i termini noti come segue:

$$y_i = y_i - l_{i1} \cdot b_1, \text{ per } i = 2, \dots, n$$

Questo aggiornamento “sposta” il valore che abbiamo sostituito a destra nel vettore dei termini noti, riducendo la matrice del sistema e ottenendo un nuovo termine noto per tutte le equazioni successive.

3. Ripetiamo questo processo per tutte le incognite sconosciute, calcolando ogni volta l'incognita e sostituendo.

$$\begin{bmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & l_{22} & \dots & 0 \\ \vdots & \ddots & & \vdots \\ l_{n1} & l_{n2} & \dots & l_{nm} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\begin{bmatrix} l_{22} & 0 & \dots & 0 \\ l_{32} & l_{33} & \dots & 0 \\ \vdots & \ddots & & \vdots \\ l_{n2} & l_{n3} & \dots & l_{nm} \end{bmatrix} \begin{bmatrix} b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix} = \begin{bmatrix} y_2 - l_{21} * \frac{y_1}{l_{11}} \\ y_3 - l_{31} * \frac{y_1}{l_{11}} \\ \vdots \\ y_n - l_{n1} * \frac{y_1}{l_{11}} \end{bmatrix}$$

Il costo computazionale di questo algoritmo è dell'ordine di $\mathcal{O}(m^2)$. In particolare, si effettuano $\frac{m(m+1)}{2}$ calcoli, ciascuno dei quali comprende un'addizione e una moltiplicazione, più m divisioni.

```

1 def forward_substitution(L, y):
2     n = len(y)
3     b = np.zeros(n)
4     b[0] = y[0] / L[0][0]
5     for k in range(0, n-1):
6         for i in range(k+1, n):
7             y[i] = y[i] - L[i][k] * b[k]
8         b[k+1] = y[k+1] / L[k+1][k+1]
9     return b.tolist()

```

Valutazione di un polinomio nella base di Newton. Il polinomio nella base di Newton è dato da:

$$p(x) = b_0 + b_1(x - x_0) + b_2(x - x_0)(x - x_1) + \dots + b_n(x - x_0) \cdots (x - x_{n-1})$$

Possiamo esprimere il polinomio in una forma che facilita la sua valutazione, similmente all'algoritmo di Horner, raccogliendo i termini simili:

$$\begin{aligned} p(x) &= b_0 + b_1(x - x_0) + b_2(x - x_0)(x - x_1) + \dots + b_n(x - x_0) \dots (x - x_{n-1}) \\ p(x) &= b_0 + (x - x_0)(b_1 + b_2(x - x_1) + \dots + b_n(x - x_1) \dots (x - x_{n-1})) \\ p(x) &= b_0 + (x - x_0)(b_1 + (x - x_1)(b_2 + \dots + b_n(x - x_2) \dots (x - x_{n-1}))) \\ &\vdots \\ p(x) &= b_0 + (x - x_0)(b_1 + (x - x_1)(b_2 + \dots + (x - x_{n-2})(b_{n-1} + b_n(x - x_{n-1})))) \dots \end{aligned}$$

Questa riscrittura del polinomio rende possibile la sua valutazione seguendo una procedura analoga a quella dell'algoritmo di Horner. Tuttavia, a differenza dell'algoritmo di Horner, dove \bar{x} è un valore fisso, qui esso è un vettore contenente tutte le differenze tra \bar{x} e i punti della base di Newton. La valutazione viene effettuata partendo dalle parentesi interne e procedendo verso l'esterno.

```

1 def newton_evaluation(b, x, x_bar):
2     p = b[-1]
3     for k in range(len(b)-2, -1, -1):
4         p = b[k] + (x_bar - x[k]) * p
5     return p

```

3.2.2 Base di Bernstein

Siano dati $n+1$ punti (x_i, y_i) , $i = 0, \dots, n$ con x_i distinti e si vuole determinare $p(x) \in \mathbb{P}_n$ nella base di Bernstein:

$$p(x) = \sum_{i=0}^n c_i B_{i,n} \quad \text{con } x \in [a, b]$$

tale che

$$p(x_i) = y_i \quad i = 0, \dots, n$$

La base di Bernstein è costruita su un intervallo $[a, b]$, dove $a = \min\{x_i\}$ e $b = \max\{x_i\}$, che sarà il più piccolo intervallo che contenga i punti. In questo intervallo viene definito il polinomio nella base di Bernstein per l'interpolazione.

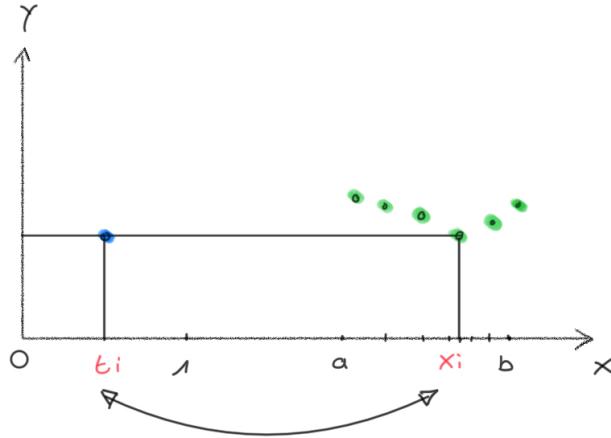
Si costruisce un sistema lineare imponendo le condizioni di esistenza:

$$\begin{aligned} p(0) = y_0 &\Rightarrow \sum_{i=0}^n c_i B_{i,n}(x_0) = y_0 \\ p(1) = y_1 &\Rightarrow \sum_{i=0}^n c_i B_{i,n}(x_1) = y_1 \\ &\vdots & &\vdots \\ p(n) = y_n &\Rightarrow \sum_{i=0}^n c_i B_{i,n}(x_n) = y_n \end{aligned}$$

Questo sistema può essere espresso in forma matriciale:

$$\begin{bmatrix} B_{0,n}(x_0) & B_{1,n}(x_0) & \dots & B_{n,n}(x_0) \\ B_{0,n}(x_1) & B_{1,n}(x_1) & \dots & B_{n,n}(x_1) \\ \vdots & \ddots & & \vdots \\ B_{0,n}(x_n) & B_{1,n}(x_n) & \dots & B_{n,n}(x_n) \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}$$

La matrice del sistema lineare è composta dai valori delle funzioni base di Bernstein valutati nei punti x_i . Tuttavia, per migliorare la precisione numerica, è conveniente traslare e scalare i punti di interpolazione nell'intervallo $[0, 1]$ mediante la 2.5.



Ora il problema consiste nell'interpolare i punti $(t_i, y_i), i = 0, \dots, n$ con un polinomio definito nella base di Bernstein:

$$p(t) = \sum_{i=0}^n c_i B_{i,n}(t) \quad \text{con } t \in [0, 1]$$

Imponendo le condizioni di esistenza, otteniamo un sistema lineare con la stessa matrice, ma valutata nei punti $t_i = 0, \dots, n$ invece che x_i . Questa rappresentazione consente di ottenere gli stessi coefficienti c_i che si otterrebbero lavorando in $[a, b]$, ma con una maggiore precisione, in quanto meno suscettibile a problemi numerici.

Per la soluzione del sistema lineare si può applicare un algoritmo ad hoc perché la matrice è totalmente positiva, e perciò qualunque sottomatrice presa quadrata che noi ritagliamo da questa avrà il determinante > 0 . Per matrici fatte così esistono algoritmi efficienti per trovare una soluzione con complessità $\mathcal{O}(n^2)$.

3.2.3 Base di Lagrange

Sia il metodo di Newton che il metodo di Bernstein, l'interpolazione richiede la soluzione di un sistema lineare. Esiste però un metodo di interpolazione che elimini completamente la necessità di risolvere un sistema lineare? Supponiamo di avere un insieme di punti (x_i, y_i) e vogliamo trovare un polinomio $p(x) \in \mathbb{P}_n$ che li interpoli. Anziché risolvere il problema di interpolazione su tutti i dati, possiamo semplificarlo costruendo un polinomio che ha zeri in tutti i punti x_i , tranne in uno specifico punto x_j , dove vogliamo che il polinomio valga 1. Per il Teorema 2.1, i polinomi $L_{i,n}(x)$ avendo come radici i punti x_j con $j = 0, \dots, n$ e $j \neq i$ saranno nella forma

$$L_{i,n}(x) = d_i(x - x_0)(x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)$$

con d_i una costante da determinare in modo che $L_{i,n}(x_i) = 1$. Troviamo d_i come:

$$d_i(x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n) = 1$$

$$d_i = \frac{1}{(x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)}$$

da cui in definitiva

$$L_{i,n}(x) = \frac{(x - x_0)(x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)}$$

$$= \frac{\prod_{j=0, j \neq i}^n (x - x_j)}{\prod_{j=0, j \neq i} (x_i - x_j)} \quad i = 0, \dots, n$$

Sono $n + 1$ funzioni che formano una base polinomiale per lo spazio \mathbb{P}_n detta base di Lagrange.

$$\{L_{0,n}(x), L_{1,n}(x), \dots, L_{n,n}(x)\} \in \mathbb{P}_n$$

Ora che abbiamo questi polinomi base di Lagrange, il problema è banalmente risolto da

$$p(x) = \sum_{i=0}^n y_i L_{i,n}(x) \tag{3.2}$$

infatti per le condizioni $L_{i,j}(x_j) = \begin{cases} 1 & \text{se } i = j \\ 0 & \text{altrimenti} \end{cases}$ che definiscono i polinomi $L_{i,n}(x)$, banalmente in ogni punto x_j l'espressione polinomiale 3.2 varrà y_j per ogni $j = 0, \dots, n$.

Dal punto di vista computazionale determinare tale polinomio non costa nulla, essendo i coefficienti del polinomio in tale base proprio i valori y_i assegnati. Al contrario la valutazione del polinomio interpolante comporta la determinazione e la valutazione dei polinomi $L_{i,n}(x)$.

Valutazione di un polinomio nella base di Lagrange. Calcolare i polinomi base di Lagrange può essere costoso, ma possiamo semplificarlo utilizzando una formula alternativa.

I forma baricentrica:

Iniziamo con il polinomio di Lagrange:

$$L_{i,n}(x) = d_i \prod_{j=0, j \neq i}^n (x - x_j)$$

$$\text{dove } d_i = \frac{1}{\prod_{j=0, j \neq i} (x_i - x_j)}$$

Possiamo calcolare una volta sola prima di valutare il polinomio:

$$p(\bar{x}) = \sum_{i=0}^n y_i L_{i,n}(\bar{x})$$

Successivamente, determiniamo:

$$l(x) = (x - x_0)(x - x_1) \cdots (x - x_n)$$

cioè, il prodotto di tutti i termini senza saltarne nessuno e riscriviamo il polinomio di Lagrange come:

$$L_{i,n}(x) = d_i \frac{l(x)}{x - x_i}$$

Da cui segue:

$$p(x) = \sum_{i=0}^n y_i \frac{w_i l(x)}{x - x_i}$$

Poiché $l(x)$ non dipende più dall'indice di sommatoria, possiamo portarlo fuori:

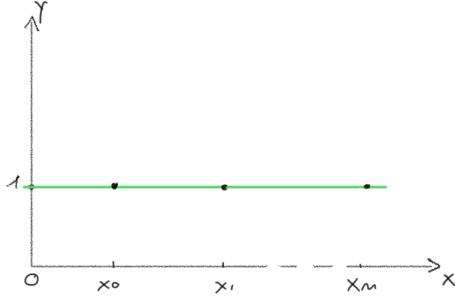
$$l(x) \sum_{i=0}^n y_i \frac{w_i}{x - x_i}$$

dove w_i e $l(x)$ vengono calcolati una sola volta con n moltiplicazioni, e all'interno della sommatoria vengono effettuate 2 moltiplicazioni per n volte. Il costo computazionale totale per la valutazione è quindi $\mathcal{O}(n)$.

II forma baricentrica:

Consideriamo il polinomio nella forma di Lagrange che interpola $(x_i, y_i = 1)$ per $i = 0, \dots, n$. Essendo la funzione costante 1 l'unica che interpola i nostri dati in \mathbb{P}_n , allora il polinomio nella I forma di Lagrange corrispondente a 1, è dato da:

$$p(x) = l(x) \sum_{i=0}^n \frac{d_i}{x - x_i} 1 = 1 \quad (3.3)$$



Ora, dividendo la I forma di Lagrange per 1, ma scrivendo 1 come in 3.3, otteniamo:

$$p(x) = \frac{l(x) \sum_{i=0}^n y_i \frac{d_i}{x - x_i}}{l(x) \sum_{i=0}^n \frac{d_i}{x - x_i}}$$

Il vantaggio di questa rappresentazione deriva dal fatto che, avendo semplificato analiticamente $l(x)$, non è più necessario calcolarlo. Il costo computazionale rimane $\mathcal{O}(n)$, ma con meno operazioni da fare, e si svolge come segue:

- Nel ciclo, si eseguono una divisione e una moltiplicazione al numeratore.
- Dopo il ciclo, si effettua un'ultima divisione.

Questo rende l'algoritmo più efficiente e altamente stabile numericamente.

```

1 def calculate_di(x, i):
2     return 1/np.prod([x[i] - x[j] for j in range(len(x)) if j != i])
3
4 def lagrange_evaluation(x, y, x_bar):
5     n = len(x)
6     d = [calculate_di(x, i) for i in range(n)]
7
8     num = 0
9     den = 0
10    for i in range(n):
11        div = d[i]/(x_bar-x[i])
12        num += y[i] * div
13        den += div
14    return num/den

```

3.3 Errore di interpolazione (interpolazione di funzioni)

Recap:

I termini C^k dove k è un intero non negativo, sono utilizzati per classificare le funzioni in base alla loro regolarità. Nello specifico:

- C^0 : La funzione è continua.
- C^1 : La funzione ha una derivata prima continua.
- C^2 : La funzione ha derivate prima e seconda continue.
- \vdots
- C^∞ : La funzione è infinitamente differenziabile e tutte le sue derivate sono continue.

Queste notazioni sono utilizzate per indicare quanto sia “liscia” una funzione, cioè quante derivate continue possiede.

Sia $f(x)$ la funzione assegnata nell’intervallo $[a, b]$ e sia $p(x)$ il polinomio interpolante nei punti $(x_i, f(x_i))$, $i = 0, \dots, n$; ha senso chiedersi quanto sia grande l’errore di interpolazione

$$R(x) = |f(x) - p(x)|$$

che si commette in un punto $\bar{x} \in [a, b]$ diverso dai punti di interpolazione x_i .

Teorema. Sia $f(x) \in C_{[a,b]}^{(n+1)}$; siano $x_0, x_1, \dots, x_n \in [a, b]$ punti di interpolazione distinti allora esiste un punto ξ che dipende da \bar{x} (punto di valutazione) per cui

$$f(\bar{x}) = p(\bar{x}) + \frac{w(\bar{x})}{(n+1)!} f^{(n+1)}(\xi) \quad (3.4)$$

con $w(\bar{x}) = (\bar{x} - x_0)(\bar{x} - x_1) \cdots (\bar{x} - x_n)$.

La parte evidenziata in rosso rappresenta l’errore di interpolazione. Questa quantità ci mostra come l’errore dipende dalla regolarità della funzione (($n+1$)-esima derivata), dai punti interpolanti scelti, e dal punto di valutazione \bar{x} specifico.

Esempio:

Sia $f(x) = e^x$ con $x \in [a, b]$. Si osserva che:

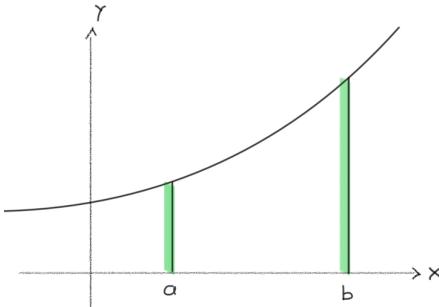
- $|f^{(n+1)}(\xi)| \leq e^b$
- $w(\bar{x}) = (\bar{x} - x_0)(\bar{x} - x_1) \cdots (\bar{x} - x_n) \leq (b-a)(b-a) \cdots (b-a) = (b-a)^{n+1}$

da cui

$$R(x) \leq \frac{(b-a)^{n+1}}{(n+1)!} e^b \xrightarrow{n \rightarrow \infty} 0$$

Notiamo che il fattore $(n+1)!$ cresce più velocemente di $(b-a)^{n+1}$, e quindi l’intero fattore tende a zero quando n tende all’infinito.

Questo dimostra che il polinomio interpolante converge alla funzione all’aumentare del numero dei punti di interpolazione.



3.3.1 Punti equispaziati e di Chebyshev

Consideriamo la funzione $f(x) = \frac{1}{1+x^2}$, definita sull'intervallo $x \in [-5, 5]$, e il polinomio di interpolazione costruito utilizzando $n+1$ punti equispaziati $x_i = \frac{10}{n}i - 5 \quad i = 0, \dots, n$.

Il matematico Runge dimostrò un caso in cui l'errore di interpolazione non migliora, ma peggiora all'aumentare dei punti. Invece di convergere alla funzione, il polinomio inizia ad oscillare vicino agli estremi dell'intervallo, e queste oscillazioni si intensificano all'aumentare del numero di punti in cui vado ad interpolare e del grado del polinomio che vado ad utilizzare.

La causa di questo comportamento risiede proprio nella scelta dei punti equispaziati.

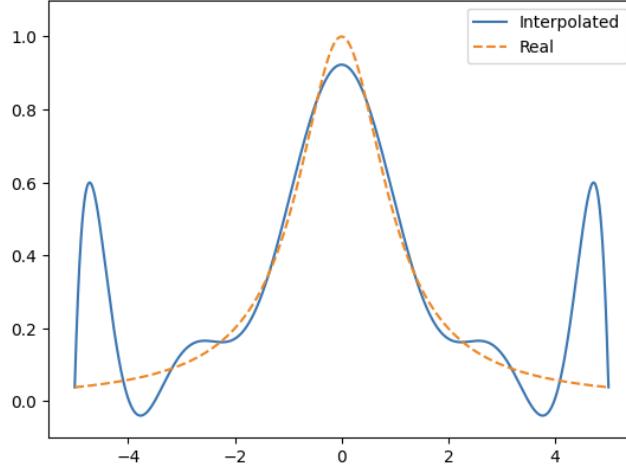


Figure 2: Interpolazione polinomiale della funzione di Runge su 11 punti equispaziati

Teorema. Sia $I \in [-1, 1]$ e i punti di interpolazione siano gli zeri del polinomio di Chebyshev di grado $n+1$. Se $f(x) \in C_I^0$ e soddisfa la condizione di Lipschitz²

$$|f(\bar{x}) - f(\tilde{x})| < K |\bar{x} - \tilde{x}|$$

allora il polinomio interpolante converge uniformemente a $f(x)$ su I per $n \rightarrow \infty$ e $|f(x) - p_n(x)| \xrightarrow{n \rightarrow \infty} 0$

Quindi se i punti di interpolazione vengono scelti opportunamente (come per esempio gli zeri del polinomio di Chebyshev di grado $n+1$ ³) e la funzione $f(x)$ è Lipschitziana si ha la convergenza dell'interpolante polinomiale. Più la f è regolare e più veloce la convergenza.

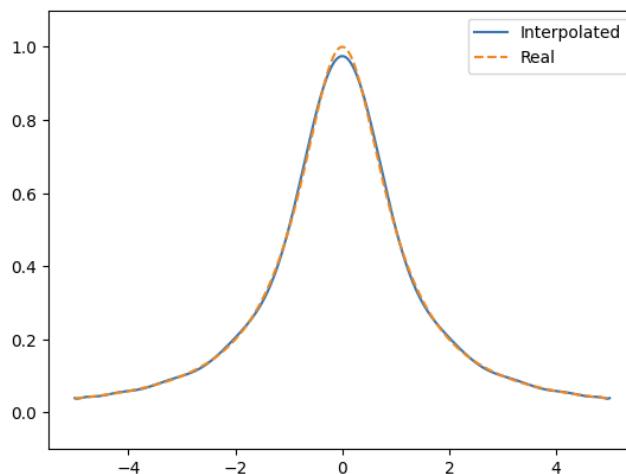


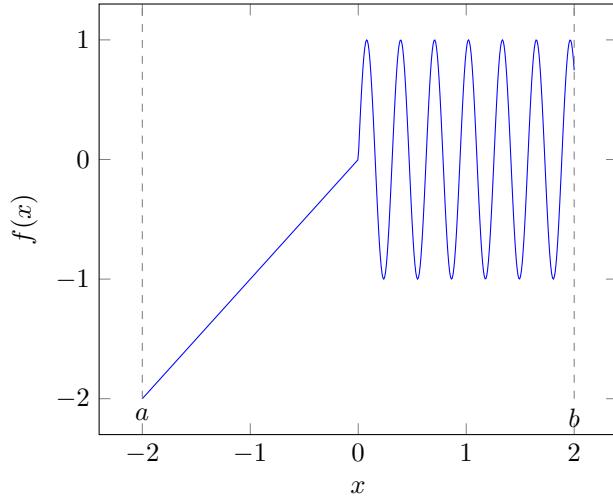
Figure 3: Interpolazione polinomiale della funzione di Runge su 21 punti zeri di Chebyshev

²La condizione di Lipschitz richiede che la funzione non abbia variazioni troppo brusche tra due punti vicini, aiutando a garantire la regolarità della funzione.

³ $x_i = \cos\left(\frac{(2i+1)\pi}{2(n+1)}\right)$, $i = 0, \dots, n$ sono gli $n+1$ zeri reali del polinomio di Chebyshev di prima specie di grado $n+1$ definito in $[-1, 1]$; se i punti di interpolazione sono in $[a, b]$ si applichi un mapping degli zeri da $[-1, 1]$ in $[a, b]$

3.4 Interpolazione polinomiale a tratti

I polinomi non sono sufficientemente flessibili per approssimare funzioni che hanno differenti andamenti in differenti parti dell'intervallo di definizione. Se si usa un polinomio di grado basso, questo approssima bene le parti lisce della funzione ma fallisce nel rappresentare le parti più variabili. Al contrario, aumentando il grado del polinomio, si ottiene una buona approssimazione delle parti più variabili, ma ciò potrebbe causare un comportamento oscillante nelle regioni in cui la funzione è più liscia.



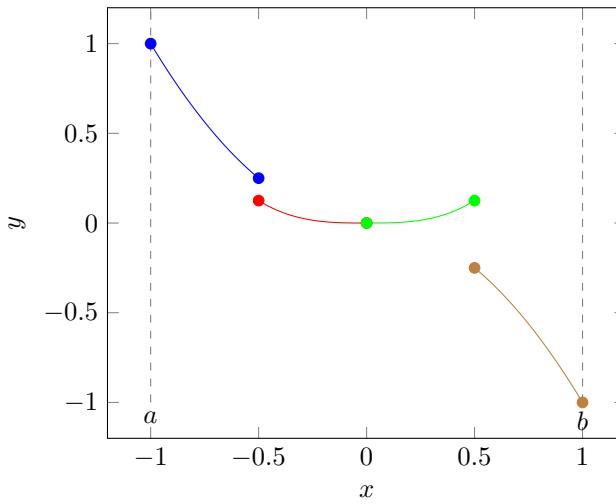
Una soluzione a questo problema è utilizzare la classe dei polinomi a tratti.

Definizione. Sia $[a, b]$ un intervallo e siano $\{x_i\}_{i=1, \dots, m-1}$ punti distinti e crescenti dell'intervallo $[a, b]$, cioè

$$a = x_0 < x_1 < x_2 < \dots < x_{m-1} < x_m = b$$

Questa sequenza di punti forma una partizione dell'intervallo $[a, b]$ in m sottointervalli. Allora si definisce un polinomio a tratti la funzione:

$$pp(x) = \begin{cases} p_0(x) & x \in [x_0, x_1] \\ p_1(x) & x \in [x_1, x_2] \\ \vdots \\ p_{m-1}(x) & x \in [x_{m-1}, x_m] \end{cases}$$



Questa classe di polinomi risolve il problema dell'inflessibilità nella rappresentazione delle funzioni. Consente di partizionare un intervallo in sottointervalli e di utilizzare polinomi distinti per ciascun sottointervallo. In questo modo, si può adattare la rappresentazione ai vari comportamenti della funzione all'interno dell'intervallo complessivo.

Tuttavia, questa soluzione introduce un problema di discontinuità nei nodi, cioè nei punti in cui i sottointervalli si congiungono. Questa discontinuità può essere problematica quando si vuole ricostruire una funzione che dovrebbe essere regolare.

Per risolvere questo problema, imponiamo delle CONDIZIONI DI CONTINUITÀ nei nodi. In altre parole, si può richiedere che i polinomi nei tratti interni si accordino di valore

$$p_{i-1}(x) \equiv p_i(x) \quad i = i, \dots, m-1$$

Oltre a ciò, possiamo imporre che le derivate ℓ -esime si accordino di valore nei nodi. In questo modo, si ottiene un certo ordine di continuità C^ℓ , che garantisce maggiore regolarità.

$$p_{i-1}^{(\ell)}(x) \equiv p_i^{(\ell)} \quad i = i, \dots, m-1 \text{ e } \ell = 1, \dots, n-1$$

Quando $\ell = n-1$ si ha la massima regolarità e la funzione polinomiale a tratti viene chiamata funzione **spline**.

Con interpolazione polinomiale a tratti si intende l'interpolazione di un set di dati in un intervallo $[a, b]$ con una funzione polinomiale a tratti. In particolare, siano assegnati $m+1$ punti distinti e ordinati, cioè siano assegnati (x_i, y_i) $i = 0, \dots, m$ con $x_i < x_{i+1} \forall i$. L'intervallo $[a, b]$ è il più piccolo intervallo che contiene questi punti, quindi avremo $a = x_0$ e $b = x_m$.

Utilizziamo i punti di interpolazione x_i per partizionare l'intervallo $[a, b]$ in sottointervalli:



Definiamo un polinomio a tratti come segue:

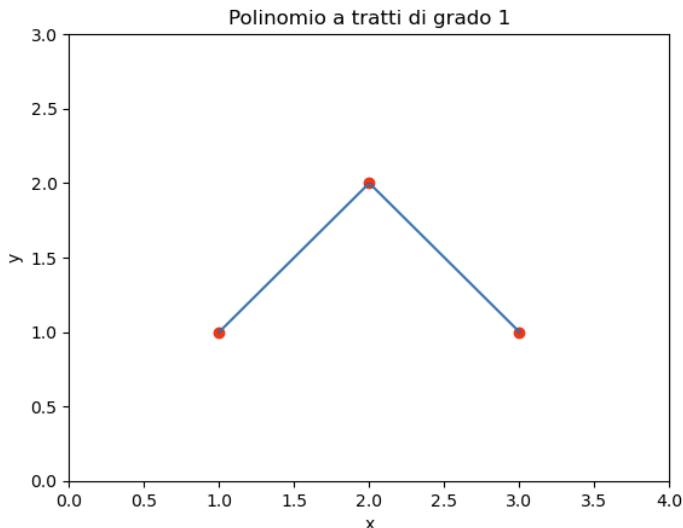
$$pp(x) = \begin{cases} p_0(x) & x \in [x_0, x_1] \\ \vdots \\ p_{m-1}(x) & x \in [x_{m-1}, x_m] \end{cases}$$

Successivamente, imponiamo le condizioni di interpolazione $pp(x_i) = y_i, i = 0 \dots, m$:

$$p_0(x_0) = y_0, \dots, p_{m-1}(x_m) = y_m$$

Esempio:

Siano dati $(x_i, y_i)_{i=0, \dots, 2} = \{(1, 1), (2, 2), (3, 1)\}$. Scegliamo $p(x) \in \mathbb{P}_1$, cioè un polinomio di grado 1 per ciascuno degli intervalli $[x_0, x_1]$ e $[x_1, x_2]$. Per farlo, utilizziamo la formula per un polinomio lineare tra due punti.



3.5 Condizionamento dell'interpolazione (interpolazione di funzioni)

L'uso della forma di Lagrange per il polinomio interpolante è molto comoda per analizzare l'errore inerente associato a questo problema.

Consideriamo il polinomio interpolante nella forma di Lagrange:

$$p(x) = \sum_{i=0}^n f(x_i) L_{i,n}(x)$$

Nell'analisi del condizionamento di questo problema, dobbiamo identificare i dati coinvolti. Poiché gli x_i sono numeri finiti scelti da noi, non sono soggetti a errori di approssimazione. L'errore potrebbe verificarsi nel valore della funzione f , dato che potrebbe essere rappresentato in modo approssimato.

Supponiamo ora che gli $f(x_i)$ siano perturbati da una quantità $\delta f(x_i)$, e denotiamo le valutazioni della funzione perturbate come:

$$\hat{f}(x_i) = f(x_i) + \delta f(x_i)$$

Di conseguenza, il polinomio perturbato sarà:

$$\hat{p}(x) = \sum_{i=0}^n \hat{f} L_{i,n}(x)$$

L'errore inerente in questo caso consiste nel valutare la differenza tra $p(x)$ e $\hat{p}(x)$, ovvero tra il polinomio calcolato con i dati esatti e il polinomio calcolato con i dati perturbati.

$$\begin{aligned} |\hat{p}(x) - p(x)| &= \left| \sum_{i=0}^n (f(x_i) + \delta f(x_i)) L_{i,n}(x) - \sum_{i=0}^n f(x_i) L_{i,n}(x) \right| \\ &= \left| \sum_{i=0}^n \delta f(x_i) L_{i,n}(x) \right| \\ &\leq \sum_{i=0}^n |\delta f(x_i)| |L_{i,n}(x)| \\ &\leq \max_{0 \leq i \leq n} |\delta f(x_i)| \sum_{i=0}^n |L_{i,n}(x)| \end{aligned}$$

Definiamo il numero di condizione del problema di interpolazione nel punto x come:

$$C_{\text{Int}}(p(x)) = \sum_{i=0}^n |L_{i,n}(x)|$$

Il massimo numero di condizione su tutto il dominio X sarà quindi:

$$\Lambda_n(X) = \max_{x \in X} C_{\text{Int}}(p(x)) \quad \text{con } X = \{x_0, x_1, \dots, x_n\}$$

Questa è chiamata **costante di Lesbegue**. Se la costante è dell'ordine di $\log(n)$, allora il problema è ben condizionato; altrimenti, è mal condizionato.

4 Integrazione numerica

Teorema (Teorema fondamentale del calcolo integrale). Sia $f : [a, b] \rightarrow \mathbb{R}$ una funzione integrabile. Supponiamo che f abbia una primitiva $F : [a, b] \rightarrow \mathbb{R}$. Allora

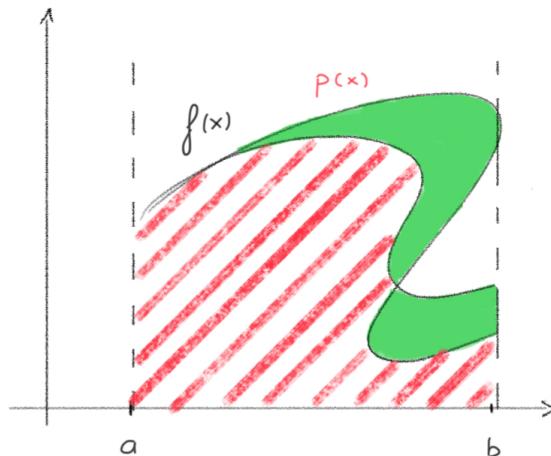
$$\int_a^b f(x) dx = F(b) - F(a) = F(x)|_{x=a}^b$$

Questo teorema afferma che il calcolo di un integrale può essere semplificato trovando una primitiva della funzione e valutandola agli estremi dell'intervallo. È importante notare, però, che non tutte le funzioni possiedono una primitiva.

Esempio:

$$\begin{aligned}\int_0^1 x dx &= \frac{1}{2}x^2\Big|_{x=0}^1 = \frac{1}{2} \\ \int_0^1 e^x dx &= e^x\Big|_{x=0}^1 = e - 1\end{aligned}$$

Se la funzione da integrare è difficile da risolvere analiticamente, non possiede una primitiva o è irrazionale, l'approccio rimanente è quello di procedere numericamente calcolando un valore approssimato del suo integrale. L'idea dietro all'integrazione numerica consiste nell'approssimare la funzione $f(x)$ con un polinomio ottenuto tramite l'interpolazione polinomiale. Questo polinomio interpolante viene poi integrato nell'intervallo $[a, b]$ per ottenere un'approssimazione del valore dell'integrale.



Poiché l'integrazione viene effettuata su un'approssimazione polinomiale, si introduce inevitabilmente un errore. Questo errore si manifesta come aree in eccesso o in difetto, che contribuiscono all'errore rispetto al risultato finale.

$$\int_a^b f(x) dx = \int_a^b p(x) dx + \text{errore} \approx \int_a^b p(x) dx$$

4.1 Formule di quadratura di Newton-Cotes

Scegliamo $n + 1$ punti di interpolazione x_i equispaziati nell'intervallo $[a, b]$. Questi punti sono dati da:

$$x_i = a + ih \quad i = 0, \dots, n \quad \text{con} \quad h = \frac{b-a}{n} \quad \text{per} \quad n > 0$$

Successivamente, costruiamo un polinomio interpolante $p \in \mathbb{P}_n$ sui punti $(x_i, f(x_i))_{i=0, \dots, n}$. Questo polinomio è espresso nella forma di Lagrange:

$$p(x) = \sum_{i=0}^n f(x_i) L_{i,n}(x)$$

Per approssimare l'integrale di $f(x)$ sull'intervallo $[a, b]$, possiamo sostituire la funzione integranda $f(x)$ con il polinomio interpolante $p(x)$:

$$\int_a^b f(x) dx \approx \int_a^b p(x) dx = \int_a^b \sum_{i=0}^n f(x_i) L_{i,n}(x) dx$$

Applicando la proprietà che l'integrale della somma è la somma degli integrali e notando che $f(x_i)$ è una costante rispetto all'integrazione, otteniamo:

$$= \sum_{i=0}^n f(x_i) \int_a^b L_{i,n}(x) dx$$

Definiamo ora gli integrali dei polinomi di Lagrange su $[a, b]$ come i coefficienti w_i :

$$= \sum_{i=0}^n f(x_i) w_i$$

Questa è la **formula di quadratura**, una combinazione lineare della funzione $f(x)$ pesata dai coefficienti w_i . Procediamo alla loro integrazione.

$$\int_a^b L_{i,n}(x) dx = \int_a^b \prod_{j=0, j \neq i}^n \frac{x - x_j}{x_i - x_j} dx$$

Effettuiamo un cambio di variabile di integrazione da $x \in [a, b]$ a $t \in [0, n]$ mediante la $x = a + ht$ con $h = \frac{b-a}{n}$

$$\begin{aligned} &= \int_0^n \prod_{j=0, j \neq i}^n \frac{\cancel{a} + \cancel{h}t - \cancel{a} - \cancel{h}j}{\cancel{a} + \cancel{h}i - \cancel{a} - \cancel{h}j} h dt \\ &= h \int_0^n \prod_{j=0, j \neq i}^n \frac{t - j}{i - j} dt \\ &= h W_i \end{aligned}$$

Notiamo che i coefficienti W_i , dipendono soltanto dal grado n del polinomio interpolante e non dalla specifica funzione $f(x)$ da integrare o dall'intervallo di integrazione. Quindi, una volta calcolati, possono essere utilizzati per approssimare l'integrale di qualsiasi funzione $f(x)$ senza doverli ricalcolare.

4.1.1 Formula (dei Trapezi) per $n = 1$

Si procede al calcolo dei W_i per $i = 0, 1$ (caso $n = 1$).

$$W_0 = \int_0^1 \frac{t - 1}{0 - 1} dt = \int_0^1 1 - t dt = \left(-\frac{t^2}{2} + t \right) \Big|_{x=0}^1 = \frac{1}{2}$$

$$W_1 = \int_0^1 \frac{t - 0}{1 - 0} dt = \int_0^1 t dt = \frac{t^2}{2} \Big|_{x=0}^1 = \frac{1}{2}$$

e quindi

$$\int_a^b f(x) dx \approx \int_a^b p_1(x) dx = h \sum_{i=0}^1 f(x_i) W_i = \frac{h}{2} (f(x_0) + f(x_1))$$

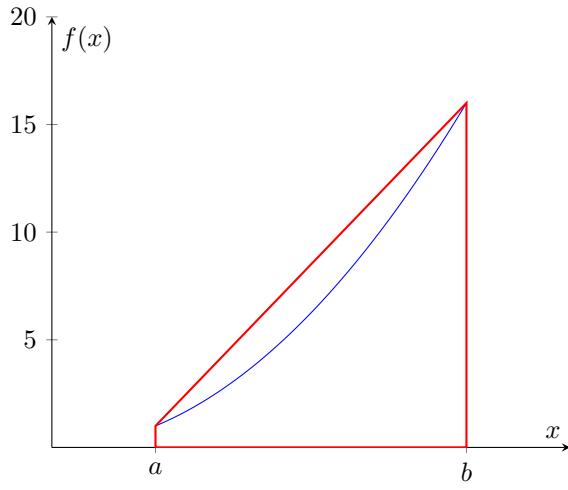


Figure 4: Interpolazione lineare per formula dei trapezi

4.1.2 Formula (di Simpson) per $n = 2$

Si procede al calcolo dei W_i per $i = 0, 1, 2$ (caso $n = 2$).

$$W_0 = \int_0^2 \frac{t-1}{0-1} \frac{t-2}{0-2} dt = \frac{1}{2} \int_0^2 t^2 - 3t + 2 dt = \frac{1}{2} \left(\frac{t^3}{3} - \frac{t^2}{2} + 2t \right) \Big|_{x=0}^2 = \frac{1}{3}$$

$$W_1 = \int_0^2 \frac{t-0}{1-0} \frac{t-2}{1-2} dt = - \int_0^2 t^2 - 2t dt = - \left(\frac{t^3}{3} - 2\frac{t^2}{2} \right) \Big|_{x=0}^2 = \frac{4}{3}$$

$$W_2 = \int_0^2 \frac{t-0}{2-0} \frac{t-1}{2-1} dt = \frac{1}{2} \int_0^2 t^2 - t dt = \frac{1}{2} \left(\frac{t^3}{3} - 2\frac{t^2}{2} \right) \Big|_{x=0}^2 = \frac{1}{3}$$

e quindi

$$\int_a^b f(x) dx \approx \int_a^b p_2(x) dx = h \sum_{i=0}^2 f(x_i) W_i = \frac{h}{3} (f(x_0) + 4f(x_1) + f(x_2))$$

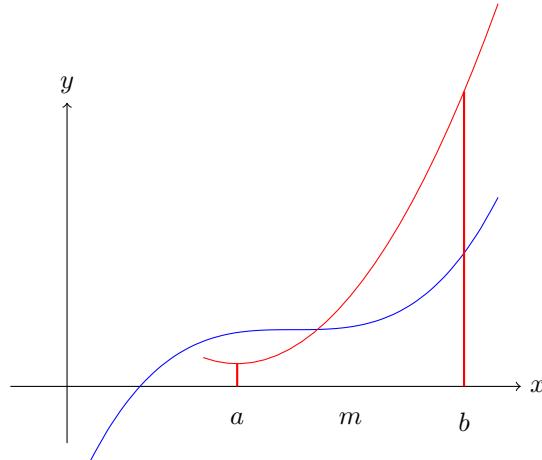


Figure 5: Interpolazione quadratica per formula di Simpson

Esempio:

Applichiamo le formule dei Trapezi e di Simpson per il calcolo di $\int_0^1 e^{-x^2} dx$, che è un esempio di integrale non risolubile tramite il calcolo della funzione primitiva. Il valore esatto alle prime 6 cifre significative è 0.746824.

Applichiamo la formula dei Trapezi:

$$\int_0^1 e^{-x^2} dx \approx \frac{h}{2}(e^0 + e^{-1}) = 0.683939.$$

Applichiamo la formula di Simpson:

$$\int_0^1 e^{-x^2} dx \approx \frac{h}{3}(e^0) + 4e^{-1/4} + e^{-1} = 0.747180.$$

4.1.3 Errore di integrazione

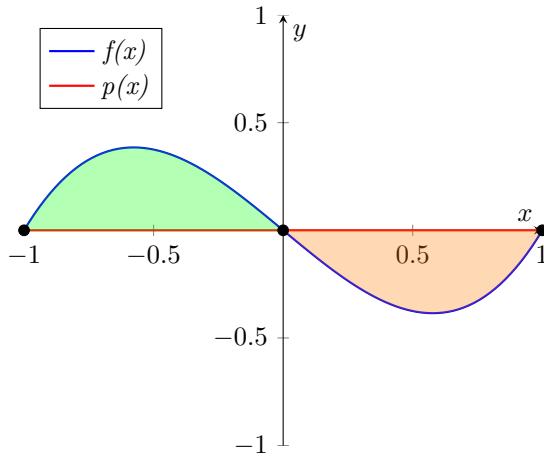
Table 1: Errore nell'Integrazione Numerica

n	σ_i	s	R
1	1 1	2	$-\frac{1}{12}h^3 f^{(2)}(\eta)$
2	1 4 1	3	$-\frac{1}{90}h^5 f^{(4)}(\eta)$
3	1 3 3 1	$\frac{8}{3}$	$-\frac{3}{80}h^5 f^{(4)}(\eta)$
4	7 32 12 32 7	$\frac{45}{9}$	$-\frac{8}{945}h^7 f^{(6)}(\eta)$

Definizione (Grado di precisione di una formula di quadratura). Una formula di quadratura si dice che ha grado di precisione n se fornisce il valore esatto dell'integrale di grado $\leq n$.

Per la formula dei Trapezi, il grado di precisione è 1, mentre per la formula di Simpson il grado di precisione è 3.

Osservazione. Si osserva che, mentre la formula dei trapezi, come ci si potrebbe aspettare è esatta per polinomi lineari di grado 1, la formula di Simpson rimane esatta non solo per polinomi di grado 2, come ci si aspetterebbe, ma anche per polinomi di grado 3. Questo avviene nonostante il fatto che un polinomio interpolante di grado 2 non dovrebbe, in teoria, essere in grado di rappresentare esattamente il polinomio di grado 3. Tuttavia, gli errori in eccesso e in difetto che si verificano si compensano esattamente, rendendo la formula esatta. La chiave sta nelle derivate: la derivata seconda di una funzione lineare è zero, e la derivata quarta di un polinomio di grado 2 o 3 è anche zero, annullando quindi l'errore.



Questo comportamento si ha in tutte le formule con n dispari ed n pari rispettivamente.

Teorema. Il grado di precisione delle formule di Newton-Cotes è n se n è dispari, è $n+1$ se n è pari.

Per $n \geq 8$, le formule diventano instabili a causa dei coefficienti di segno diverso e dell'aumento del loro valore assoluto che portano a problemi di cancellazione numerica e perdita di precisione. Di conseguenza, si rende necessaria la ricerca di formule più stabili e precise.

4.2 Formule di quadratura composite

Le formule composte consistono nel considerare dei polinomi a tratti come funzioni approssimanti della funzione integranda nell'intervallo $[a, b]$.

$$\int_a^b f(x) dx \approx \int_a^b pp_1(x) dx$$

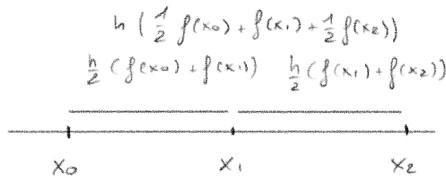
In pratica tale intervallo viene suddiviso in sottointervalli di uguale ampiezza ($h = \frac{b-a}{m}$), $[x_i, x_{i+1}]$ $i = 0, \dots, m-1$, e su ciascuno di essi si applica una formula di quadratura. La formula di quadratura composita è data dalla somma degli integrali calcolati su tutti gli intervalli.

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

Al variare di n abbiamo le varie formule composite:

- **Formula dei Trapezi composita:**

$$\sum_{i=0}^{m-1} \int_{x_i}^{x_{i+1}} p_1(x) dx = h \left(\frac{1}{2} f(x_0) + f(x_1) + \dots + f(x_{m-1}) + \frac{1}{2} f(x_m) \right) + R_{TC}$$



Dove R_{TC} è definito come segue:

$$R_{TC} = -\frac{h^3}{12} \sum_{i=0}^{m-1} f^{(2)}(\eta_i) \quad \eta_i \in (x_i, x_{i+1})$$

Sostituendo $h = \frac{b-a}{m}$, l'errore diventa:

$$= -\frac{h^2}{12} \left(\frac{b-a}{m} \right) \sum_{i=0}^{m-1} f^{(2)}(\eta_i)$$

La parte in rosso rappresenta una media dei valori delle derivate seconde da 0 a $m-1$. Utilizzando il TEOREMA DEL VALOR MEDIO, questo termine è uguale alla derivata seconda valutata in un opportuno punto η nell'intervallo (a, b) . Pertanto, possiamo riscrivere R_{TC} come segue:

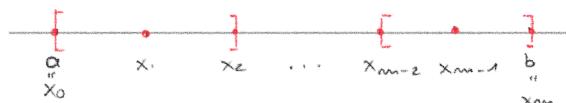
$$= -\frac{b-a}{12} h^2 f^{(2)}(\eta) \quad \eta \in (a, b) \tag{4.1}$$

Poiché $h = \frac{b-a}{m}$, man mano che m aumenta, h diventa più piccolo, migliorando la precisione dell'integrale e riducendo l'errore.

$$R_{TC} \xrightarrow[m \rightarrow \infty]{} 0$$

Un vantaggio chiave dell'uso delle formule di quadratura composte, è proprio questo incremento di precisione quando si aumenta il numero di sottointervalli m .

- **Formula di Simpson composita:** se prendiamo $m = 2k$ l'intervallo $[a, b]$ viene diviso in un numero pari di sottointervalli. Ogni sottointervallo è delimitato dai punti $[x_{2i}, x_{2i+2}]$ e include tre punti distinti. In questo modo, possiamo utilizzare la formula di Simpson per calcolare l'integrale in ogni sottointervallo.



$$\sum_{i=0}^{k-1} \int_{x_{2i}}^{x_{2i+1}} p_2(x) dx = \frac{h}{3} (f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + \dots + f(x_m)) + R_{SC}$$

Dove R_{SC} è definito come segue:

$$R_{SC} = -\frac{h^5}{90} \sum_{i=0}^{k-1} f^{(4)}(\eta_i) \quad \eta_i(x_{2i}, x_{2i+1})$$

Per le stesse argomentazione usate per la formula dei trapezi, questo si può scrivere:

$$\begin{aligned} &= -\frac{h^4}{90} \frac{b-a}{2k} \sum_{i=0}^{k-1} f^{(4)}(\eta_i) \\ &= -\frac{b-a}{180} h^4 f^{(4)}(\eta) \quad \eta \in (a, b) \end{aligned} \tag{4.2}$$

In questo caso, il valore di h converge più rapidamente rispetto al metodi dei Trapezi composito quando si aumenta m .

Esempio:

Dato l'integrale

$$\int_0^1 \frac{1}{1+x} dx$$

il cui valore esatto è 0.693147.

Calcolare il valore approssimato dell'integrale e l'errore R utilizzando sia la formula dei Trapezi composita che la formula di Simpson composita.

h	$T(h)$	R_{TC}	$S(h)$	R_{SC}
$h=1$	0.75	0.0568	-	-
$h/2=0.5$	0.708	0.01518	0.6944	0.00129
$h/4=0.25$	0.697	0.0038	0.6932	0.000107
$h/8=0.125$	0.694	0.0097	0.69315	0.00009

Come visto precedentemente, l'errore R_{TC} nel metodo dei Trapezi composito è proporzionale a h^2 , mentre l'errore R_{SC} nel metodo di Simpson composito è proporzionale a h^4 . Effettivamente, nell'esempio questo è confermato: il rapporto di riduzione dell'errore nel metodo dei Trapezi è di un fattore 4, mentre nel metodo di Simpson è di un fattore 16 quando h viene dimezzato.

Esempio:

Determinare il passo h da utilizzare nella formula dei Trapezi composita e nella formula di Simpson composita, affinché l'integrale $\int_0^1 \frac{1}{1+x} dx$ sia approssimato alla tolleranza 0.5×10^{-3} .

Si considera la formula 4.1 per l'errore di integrazione dei Trapezi composita e si cerca un limite superiore per la $f^{(2)}$ in $[a, b]$; nel caso specifico sarà:

$$f'(x) = (1+x)^{-1} \quad f''(x) = 2(1+x)^{-3}$$

Allora

$$\max_{0 \leq x \leq 1} f''(x) = f''(0) = 2$$

e ricordando che $h = \frac{b-a}{2k}$ si ha:

$$\left| \frac{b-a}{12} h^2 f''(\eta) \right| \leq \frac{1}{12} \cdot \frac{1}{6} h^2 \cdot 2 = \frac{1}{6} \cdot \frac{1}{4k^2} = \frac{1}{24k^2} \leq 0.5 \times 10^{-3}$$

da cui

$$k^2 \geq \frac{2}{24} \times 10^3 \quad \text{e quindi } k \geq 9.12$$

Segue che per approssimare l'integrale dato alla tolleranza fissata con il metodo dei Trapezi composito è sufficiente usare $k = 10$, cioè il più piccolo intero che soddisfa $k \geq 9.12$.

Si considera la formula 4.2 per l'errore di integrazione di Simpson composita e si cerca un limite superiore per la $f^{(4)}$ in $[a, b]$; nel caso specifico sarà:

$$f^{(3)}(x) = 6(1+x)^{-4} \quad f^{(4)}(x) = 24(1+x)^{-5}$$

Allora

$$\max_{0 \leq x \leq 1} f^{(4)}(x) = f^{(4)}(0) = 24$$

si ha:

$$\left| \frac{b-a}{180} h^4 f^{(4)}(\eta) \right| \leq \frac{1}{180} \cdot \frac{1}{15} h^4 \cdot 24 = \frac{2}{15} \cdot \frac{1}{16k^4} = \frac{1}{120k^4} \leq 0.5 \times 10^{-3}$$

da cui

$$k^4 \geq \frac{2}{120} \times 10^3 \quad \text{e quindi } k \geq 2.0205$$

Segue che per approssimare l'integrale dato alla tolleranza fissata con il metodo di Simpson composita è sufficiente usare $k = 3$, cioè il più piccolo intero che soddisfa $k \geq 2.0205$.

4.3 Metodi adattivi

Si desidera che la differenza tra l'integrale esatto $\int_a^b f(x) dx$ e la sua approssimazione I_A sia contenuta entro una certa tolleranza tol .

$$\left| \int_a^b f(x) dx - I_A \right| \leq tol$$

Purtroppo, quando non si conosce il valore esatto dell'integrale, quanto visto fino ad ora non permette di soddisfare questa richiesta.

Nella prossima sezione vedremo una tecnica nota come estrapolazione di Richardson che permette di stimare numericamente l'errore di integrazione.

4.3.1 Estrapolazione di Richardson

Con **estrapolazione di Richardson** ci si riferisce ad una tecnica che permette di ottenere dall'applicazione di due formule di integrazione composte con passi rispettivamente h ed $\frac{h}{2}$ un valore di approssimazione per l'integrale più preciso dei due precedenti. Vediamo questa tecnica nel caso delle formule composite:

Caso Trapezi.

Corollario. Assumendo che la $f(x)$ sia derivabile almeno 4 volte su $[a, b]$, si ha:

$$\int_a^b f(x) dx - T(h/2) \approx \frac{T(h/2) - T(h)}{3} + \mathcal{O}(h^4)$$

da cui

$$\int_a^b f(x) dx \approx \frac{4T(h/2) - T(h)}{3} + \mathcal{O}(h^4)$$

Osservazione.

- * **Miglioramento della stima.** Questo metodo consente di eliminare la potenza più bassa (2) combinando opportunamente $T(h)$ e $T(h/2)$, al fine di ottenere un'approssimazione più precisa dell'integrale.
- * **Stima dell'errore.** Inoltre, il corollario fornisce una stima del termine principale dell'errore e quindi fornisce un metodo pratico per raggiungere l'obiettivo che ci si è posti. Si può progettare un metodo iterativo in cui si dimezza il passo fino a che

$$\left| \frac{T(h/2) - T(h)}{3} \right| \leq tol \quad (4.3)$$

quando questo si verifica, per il corollario, sarà anche

$$\left| \int_a^b f(x) dx - T(h/2) \right| \leq tol$$

e avremo una buona integrazione con $T(h/2)$. Questo metodo iterativo può essere applicato in modo adattivo.

In un approccio adattivo il numero di punti viene scelto in base alla regolarità della funzione. Si suddivide l'intervallo di integrazione, solo dove serve, in sottointervalli e si applica ricorsivamente a questi una formula di quadratura sfruttando la stima dell'errore di integrazione 4.3 per il test di arresto. La funzione integranda viene così valutata in pochi punti nei sottointervalli in cui ha un andamento regolare e in molti punti negli intervalli dove sono presenti irregolarità.

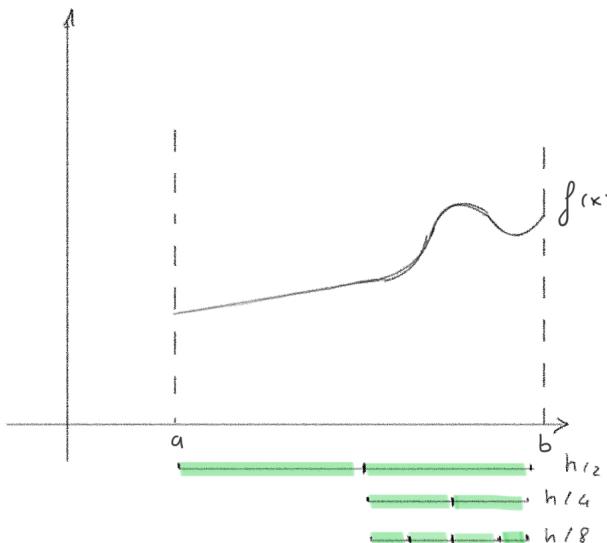


Figure 6: Metodo adattivo con controllo dell'errore che ci permette di calcolare il valore approssimato dell'integrale della funzione alla tolleranza fissata suddividendo solo dove serve

```

1 def trap_adapt(fn, a, b, fa, fb, tol, tab):
2     m = 0.5 * (a + b)
3     fm = fn(m)
4
5     tam = trap_sing(a, m, fa, fm)
6     tmb = trap_sing(m, b, fm, fb)
7
8     if math.fabs(tab - tam - tmb) / 3 < tol:
9         return (4 * (tam + tmb) - tab) / 3 # Estrapolazione di Richardson
10    else:
11        return (trap_adapt(fn, a, m, fa, fm, 0.5 * tol, tam) +
12                trap_adapt(fn, m, b, fm, fb, 0.5 * tol, tmb))

```

Caso Simpson. Calcolando un Simpson con passo $h = \frac{b-a}{2}$ ($S(h)$) e un Simpson con passo $h/2 = \frac{b-a}{4}$ ($S(h/2)$); avremo la seguente formula che fornisce un'approssimazione dell'integrale d'ordine superiore:

$$\left| \int_a^b f(x) - S(h/2) dx \right| \approx \left| \frac{1}{15} [S(h/2) - S(h)] \right|$$

da questa relazione, possiamo ottenere un'approssimazione migliore dell'integrale come segue:

$$\int_a^b f(x) dx \approx \frac{1}{15} [16 \cdot S(h/2) - S(h)]$$

Utilizzando la [prima relazione](#) possiamo implementare un procedimento iterativo in cui il passo viene dimezzato fino a quando la seguente condizione è soddisfatta:

$$\left| \frac{1}{15} [S(h/2) - S(h)] \leq tol \right|$$

Una volta soddisfatta questa condizione, avremo che:

$$\left| \int_a^b f(x) dx - S(h/2) \right| \leq tol$$

Di seguito il metodo adattivo per Simpson:

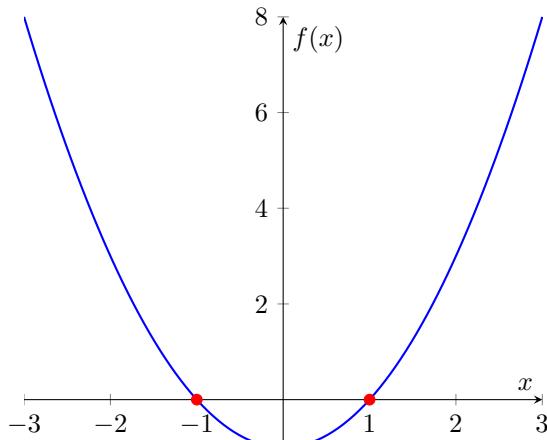
```

1 def simp_adapt(fn, a, m, b, fa, fm, fb, tol, sab):
2     m1 = 0.5*(a+m)
3     m2 = 0.5*(m+b)
4     fm1 = fn(m1)
5     fm2 = fn(m2)
6     sam = simp_sing(a, m, fa, fm1, fm)
7     smb = simp_sing(m, b, fm, fm2, fb)
8
9     if math.fabs(sab - sam - smb) / 15 < tol:
10        return (16*(sam+smb) - sab)/15 # Estrapolazione di Richardson
11    else:
12        return (simp_adapt(fn, a, m1, m, fa, fm1, fm, 0.5*tol, sam) +
13                simp_adapt(fn, m, m2, b, fm, fm2, fb, 0.5*tol, smb))

```

5 Zeri di funzioni non lineari

Si considera il problema di trovare le radici di una funzione non lineare, ovvero risolvere $f(x) = 0$. Geometricamente, questo equivale a trovare i punti in cui la funzione $f(x)$ interseca l'asse delle ascisse (gli zero della funzione).



Per equazioni con grado $n \leq 4$, esistono soluzioni analitiche che permettono di determinare le radici in modo esatto. Tuttavia, per equazioni di grado superiore, non esistono soluzioni analitiche esplicite.

In questi casi, si ricorre a metodi numerici per approssimare le radici della funzione.

5.1 Metodo di bisezione

Per poter applicare questo metodo la funzione $f(x)$ con $x \in [a, b]$ deve rispettare le seguenti condizioni:

1. **Continuità della funzione:** la funzione $f(x)$ deve essere continua nell'intervallo.

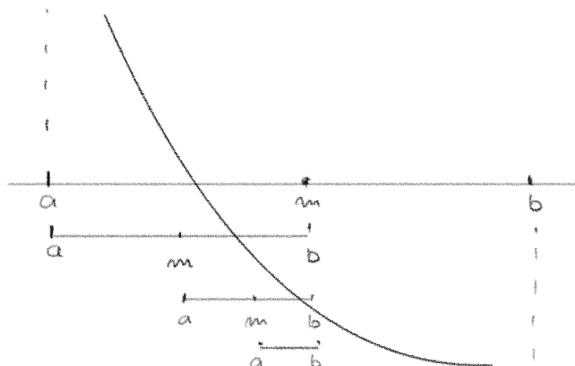
$$f \in C[a, b]$$

2. **Segno opposto agli estremi:** la funzione deve avere segni opposti agli estremi dell'intervallo a e b .

$$f(a) \cdot f(b) < 0$$

Se entrambe queste condizioni sono soddisfatte, allora esiste almeno un valore z nell'intervallo (a, b) per cui $f(z) = 0$.

Il metodo procede suddividendo ripetutamente l'intervallo $[a, b]$ a metà, calcolando il punto medio m . Successivamente, si verifica quale dei due nuovi intervalli $[a, m]$ o $[m, b]$ soddisfa la condizione per contenere x^* . In particolare, se $f(a)f(m) < 0$ allora $x^* \in [a, m]$; altrimenti $x^* \in [m, b]$. L'intervallo scelto diventa il nuovo $[a, b]$, creando una successione di intervalli $[a_k, b_k]$ per $k = 1, \dots, m$. Il procedimento viene iterato sull'intervallo che è risultato contenere x^* e viene arrestato quando l'ampiezza dell'intervallo in esame risulterà minore di una prefissata tolleranza tol .



Osservazione.

Il valore di tol non può essere scelto arbitrariamente piccolo, poiché lavorando con numeri finiti la condizione di arresto

$$\frac{|b_k - a_k|}{\min(|a_k|, |b_k|)} < tol$$

potrebbe non essere mai soddisfatta. Poiché a_k e b_k sono numeri finiti, la loro distanza non sarà mai esattamente zero. Se la tolleranza tol viene impostata a un valore minore della distanza tra a_k e b_k il metodo potrebbe non convergere.

Si può dimostrare che la distanza relativa tra due numeri finiti consecutivi X e Y è data da

$$\frac{Y - X}{X} = 2u$$

Ne consegue che tol deve essere maggiore di $2u$.

Per evitare la divisione per zero quando a_k o b_k sono uguali a zero, il test di arresto può essere riformulato come segue:

$$|b_k - a_k| < 2u \cdot \min(|a_k|, |b_k|) + tol$$

dove ora tol può essere scelto arbitrariamente e il problema della divisione per zero è evitato.

Osservazione. Se $x^* \equiv 0$ sarà sempre $a_k < 0$ e $b_k > 0$ per cui

$$\frac{|b_k - a_k|}{\min(|a_k|, |b_k|)} > 1$$

per evitare questo caso basta controllare se $f(0) \equiv 0$ quando $a < 0 < b$.

```

1 def bisez(fn, a, b, tol):
2     if a<0 and b>0 and fn(0) == 0:
3         return 0
4     while abs(b-a) > 2*u*(min(abs(a), abs(b)))+tol:
5         m = (a+b) / 2
6         fm = fn(m)
7         if fn(a)*fm < 0:
8             b=m
9         else:
10            a=m
11    return (a+b)/2

```

Tuttavia, il metodo di bisezione presenta lo svantaggio di essere lento. In particolare, ad ogni iterazione, l'approssimazione della radice migliora soltanto di un bit. Considerando una precisione di tipo double, che dispone di una mantissa di 53 cifre, sarebbero necessarie 53 iterazioni per raggiungere l'approssimazione desiderata della radice.

Il problema del metodo di bisezione è che utilizza solo una frazione delle informazioni disponibili. Nello specifico, si limita a considerare solamente il segno della funzione agli estremi dell'intervallo in esame. Fortunatamente, esistono metodi che superano questa limitazione, sfruttando informazioni addizionali come il valore effettivo della funzione agli estremi, per accelerare il processo di convergenza.

5.1.1 Metodo della falsa posizione

Nel metodo della falsa posizione, si interpola una retta tra i punti $(a, f(a))$ e $(b, f(b))$ e si trova il punto y in cui questa retta interseca l'asse x tramite la formula:

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

Successivamente, si verifica quale dei due nuovi intervalli, $[a, y]$ o $[y, b]$, soddisfa la condizione per contenere x^* . Si sostituisce uno degli estremi con y e si continua iterativamente.

A differenza del metodo di bisezione, non stiamo rimpicciolendo gli intervalli ma abbiamo che uno dei due estremi si avvicina sempre di più alla soluzione x^* . Formalmente, avremo una successione x_k con $k = 0, \dots, n$ di estremi tale che:

$$x_k \xrightarrow{k \rightarrow \infty} x^*$$

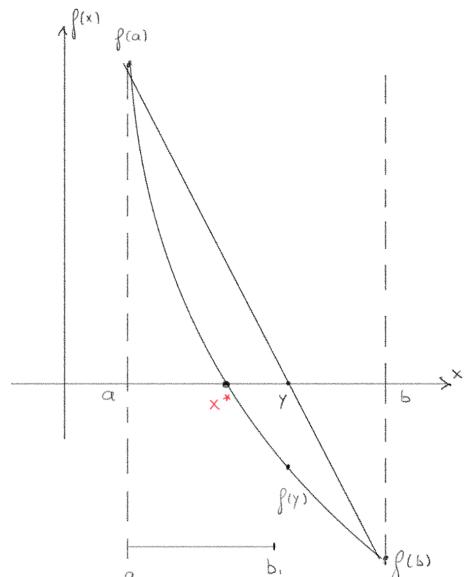
Il test d'arresto delle falsi posizioni è diverso da quello del metodo della bisezione. Si calcola la differenza relativa tra due iterative successive x_k e x_{k-1} e confrontarla con la tolleranza desiderata:

$$\frac{|x_k - x_{k-1}|}{\min(|x_{k-1}|, |x_k|)} < tol$$

e, come prima:

$$|x_k - x_{k-1}| < 2u \cdot \min(|x_k|, |x_{k-1}|) + tol$$

Quindi, il metodo converge in molti meno passi rispetto al metodo di bisezione.



```

1 def false_position(fn, a, b, tol):
2     prev_y = None
3     fa = fn(a)
4     fb = fn(b)
5     while True:
6         y = b - (fb * (b - a)) / (fb - fa)
7         if prev_y is not None and abs(y - prev_y) < 2*u*(min(abs(y), abs(prev_y)))+tol:
8             return y
9         fy = fn(y)
10        if fa*fy < 0:
11            b=y
12            fb=fy
13        else:
14            a=y
15            fa=fy
16        prev_y = y

```

5.2 Metodo di Newton

Supponiamo di avere una funzione $f \in C^2[a, b]$ e di conoscere un punto $\bar{x} \in [a, b]$ che sia un'approssimazione della soluzione x^* tale che $f'(\bar{x}) \neq 0$ e che $|\bar{x} - x^*|$ sia piccolo.

Partiamo da uno sviluppo di Taylor centrato in \bar{x} :

$$f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + f''(\bar{x}) \left(\frac{x - \bar{x}}{2} \right)^2 + \dots$$

Valutiamo x^* , usando lo sviluppo di Taylor:

$$\underbrace{f(x^*)}_0 = f(\bar{x}) + f'(\bar{x})(x^* - \bar{x}) + f''(\bar{x}) \left(\frac{x^* - \bar{x}}{2} \right)$$

Poiché $|\bar{x} - x^*|$ si è assunto piccolo, $(x^* - \bar{x})^2$ sarà ancora più piccolo e ancora di più i termini successivi; trascurando allora i termini non lineari abbiamo:

$$0 = f(\bar{x}) + f'(\bar{x})(x^* - \bar{x})$$

Risolvendo per x^* otteniamo:

$$x^* = \frac{f(\bar{x})}{f'(\bar{x})}$$

Questa relazione fornisce l'idea per il **metodo di Newton**, che consiste nel generare una successione $\{x_k\}$ definita da:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \quad \text{con } k \geq 0 \text{ e } f'(x_k) \neq 0 \forall k. \quad (5.1)$$

Geometricamente, indica che ogni nuovo iterato x_{k+1} è dato dall'intersezione fra la retta tangente $f'(x_k)$, alla $y = f(x_k)$ nel punto x_k e l'asse x .

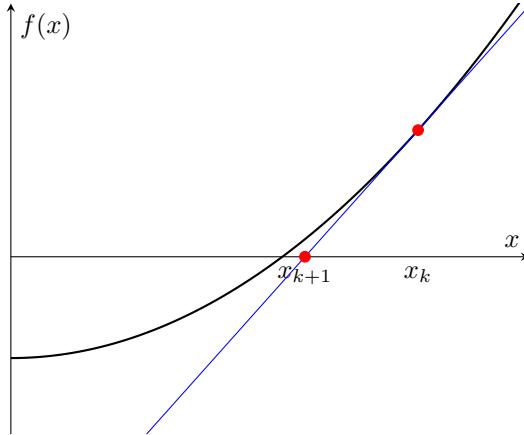


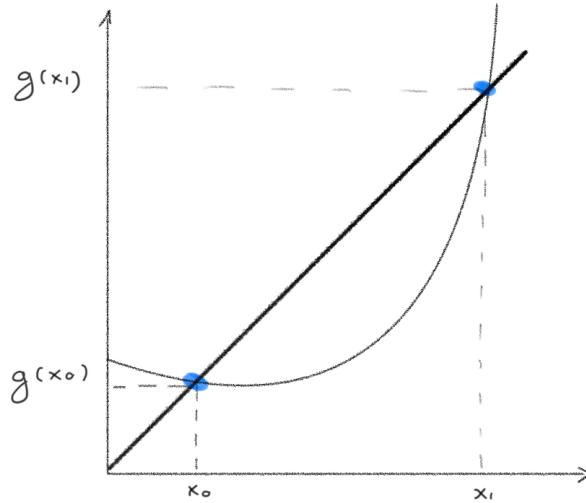
Figure 7: Iterazione del metodo di Newton

Per verificare la convergenza della successione $\{x_k\}_{k=0,1,\dots}$, consideriamo una famiglia più generale di metodi, di cui il metodo di Newton è un caso particolare. Questa famiglia di metodi è nota come metodi di iterazione funzionale e sono descritti dalla seguente formula

$$x_{k+1} = g(x_k) \quad k = 0, 1, \dots \quad (5.2)$$

Tali metodi risolvono un problema diverso: trovare le soluzioni dell'equazione

$$x = g(x)$$



Ciò significa trovare i punti di intersezione tra la retta bisettrice $y = x$ e la curva $y = g(x)$. Questi sono detti **punti fissi** di $g(x)$, cioè i punti in cui la funzione assume lo stesso valore dell'argomento.

Se $g(x) = x - \frac{f(x)}{h(x)}$ con $h(x) \neq 0$, allora i due problemi sono equivalenti e le radici di $f(x) = 0$ sono anche i punti fissi di $g(x)$ e viceversa.

Dimostrazione.

- Sia $f(x^*) = 0$. Allora abbiamo

$$\begin{aligned} g(x^*) &= x^* - \overbrace{\frac{f(x^*)}{h(x^*)}}^0 \\ &= x^* \end{aligned}$$

il che implica che x^* è un punto fisso di $g(x)$.

- Viceversa, sia x^* un punto fisso di $g(x)$. In tal caso, si ha

$$g(x^*) = x^* - \frac{f(x^*)}{h(x^*)}$$

che semplificato diventa $0 = f(x^*)$.

■

Dopo aver stabilito l'equivalenza tra i due problemi, procediamo a dimostrare la convergenza di un metodo iterativo funzionale.

Teorema. Se $g(x)$ possiede un punto fisso x^* e se $g(x)$ è continua e derivabile in $[x^* - \rho, x^* + \rho]$ con $\rho > 0$ e soddisfa la condizione

$$|g'(x)| \leq \lambda < 1 \quad \text{con } x \in [x^* - \rho, x^* + \rho]$$

Allora per ogni $x_0 \in [x^* - \rho, x^* + \rho]$, tutti i successivi x_k generati dalla 5.2 appartengono a questo intervallo e la successione converge a x^* .

In altre parole, il teorema ci dice che se partiamo da un punto iniziale x_0 che appartiene all'intervallo che contiene x^* , allora tutti i punti x_k iterati successivi rimarranno all'interno di questo intervallo e convergeranno al punto x^* .

Dimostrazione.

$$|x_{k-1} - x^*| < |x_k - x^*|$$

L'obiettivo è dimostrare che se la successione soddisfa questa condizione per ogni k , allora necessariamente converge a x^* .

Per farlo, riscriviamo nel seguente modo:

$$|x_{k+1} - x^*| = |g(x_k) - g(x^*)|$$

Dove abbiamo utilizzato il fatto che $x_{k+1} = g(x_k)$ e $x^* = g(x^*)$.
Applicando il teorema del valor medio, otteniamo:

$$= |g'(\xi_k)(x_k - x^*)|, \quad \text{dove } \xi_k \in (x^*, x_k)$$

Dato che per ipotesi $|g'(x)| < \lambda$ su un intervallo che contiene x^* , abbiamo:

$$\leq \lambda |x_k - x^*|, \quad \text{dove } \lambda < 1$$

Iterando questa diseguaglianza, otteniamo:

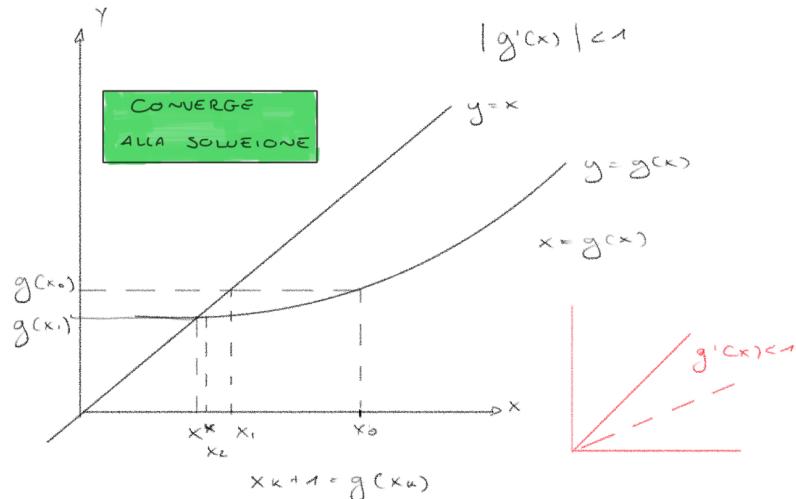
$$\leq \lambda^k |x_0 - x^*|$$

Poiché $\lambda < 1$, possiamo concludere che:

$$\lim_{k \rightarrow \infty} |x_{k+1} - x^*| \leq \lim_{k \rightarrow \infty} \lambda^k |x_0 - x^*| = 0$$

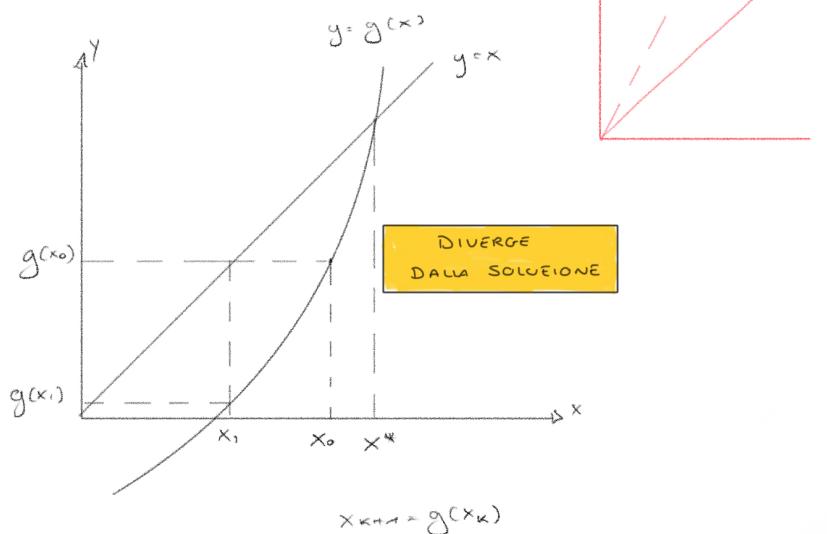
Pertanto, la successione converge a x^* . ■

ESEMPIO 1 $|g'(x)| < 1$



SU QUESTO ESEMPIO GRAFICO VOGLIANO PRENDERE
UN PUNTO INIZIALE x_0 VICINO A x^* E VEDERE LA CONVERGENZA A x^*

ESEMPIO 2 $|g'(x)| > 1$



Si vuole sfruttare il Teorema precedente per dimostrare il seguente risultato:

Teorema (Convergenza del metodo di Newton). Se $f(x) \in C^2[a, b]$, $f(x^*) = 0$ e $f'(x^*) \neq 0$, allora esiste un intervallo $I = [x^* - \rho, x^* + \rho]$ contenente x^* tale che se $x_0 \in I$ il metodo di Newton converge a x^* .

Dimostrazione. Se scriviamo $g(x) = x - \frac{f(x)}{h(x)}$, con $h(x) \neq 0$, le soluzioni di un problema sono le soluzioni dell'altro. Se prendiamo $h(x) = f'(x)$, otteniamo il metodo di Newton:

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

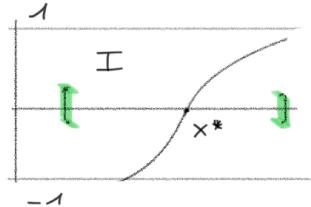
Il Teorema precedente afferma che la successione $x_{k+1} = g(x_k)$ converge a x^* se $|g'(x)| < 1$ per ogni $x \in I$. Per verificare questa condizione, deriviamo $g(x)$ e otteniamo:

$$\begin{aligned} g'(x) &= 1 - \frac{[f'(x)]^2}{[f'(x)]^2} + \frac{f''(x)f(x)}{[f'(x)]^2} \\ &= \frac{f''(x)f(x)}{[f'(x)]^2} \end{aligned}$$

Se valutiamo questa espressione in $x = x^*$, otteniamo:

$$\frac{f''(x^*)f(x^*)}{[f'(x)]^2}$$

Per ipotesi $f'(x^*) \neq 0$ e $f(x^*) = 0$ quindi l'espressione si annulla in x^* . Di conseguenza esiste un intorno di x^* in cui $|g'(x)| < 1$.



■

Da ciò, possiamo concludere che la convergenza è garantita se, e solo se, si seleziona un intervallo opportuno in cui $g'(x) < 1$. Sotto queste condizioni, il metodo convergerà sempre. Il passo successivo è quindi determinare un intervallo I attorno a x^* che soddisfi questa condizione, in modo da innescare con successo il metodo di Newton.

Test di arresto. Nell'implementare il metodo di Newton, è bene prevedere almeno tre test al verificarsi di uno dei quali ci si deve arrestare:

1. $|x_k - x_{k-1}| < 2u \cdot \min(|x_{k-1}|, |x_k|) + tol$. Questo è analogo al test di arresto utilizzato nel metodo della falsa posizione.
2. $k > max_iter$. Se il valore iniziale x_0 non appartiene all'intervallo corretto, il metodo iterativo divergerà. In altre parole, gli iterati successivi non soddisfaranno mai il test di arresto, risultando in un ciclo infinito. Per prevenire questo problema, inseriamo un numero massimo di iterazioni, oltre al quale il metodo si arresterà.
3. $|f(x_k)| \leq tol_2$. L'errore inerente nel calcolo di uno zero x^* di una funzione $f(x)$ è dato dalla formula:

$$|\hat{x}^* - x^*| = \frac{1}{f'(\xi)} E(\hat{x}^*)$$

Questa equazione mostra che l'errore assoluto nel risultato dipende sia dall'errore nei dati $E(\hat{x}^*)$ e sia dal fattore che potrebbe amplificarlo, ovvero l'inverso della derivata prima $f'(\xi)$ in un punto ξ nell'intervallo (x^*, \hat{x}^*) .

Osserviamo che se la derivata $f'(\xi)$ è molto piccola, il suo inverso sarà molto grande, e quindi anche un piccolo errore nei dati iniziali può essere amplificato in modo significativo. In tal caso, il problema è mal condizionato, e il metodo dovrebbe essere interrotto per evitare risultati inaccurati.

```

1 def newton(fn, fn_p, x0, tol):
2     n=0
3     max_iter=50
4     small_real = 100*real_min
5
6     x_curr = x0
7     x_prev = x0 + 1
8     while abs(x_curr - x_prev) > 2*u*min(abs(x_curr), abs(x_prev))+tol and n < max_iter:
9         x_prev = x_curr
10        fx_prev = fn(x_prev)
11        if abs(fx_prev)>=small_real:
12            x_curr = x_prev - (fx_prev / fn_p(x_prev))
13            n=n+1
14    return x_curr

```

Ordine di convergenza

Definizione. Sia $\{x_k\}$ una successione convergente ad x^* e sia $x_k \neq x^* \forall k$. Se esiste un numero $p \geq 1$, tale che

$$\lim_{k \rightarrow \infty} \frac{|e_{k+1}|}{|e_k|^p} = \gamma \quad \text{dove } e_k = x_k - x^*$$

con

$$\begin{cases} 0 < \gamma \leq 1 & \text{se } p = 1 \\ \gamma > 0 & \text{se } p > 1 \end{cases}$$

si dice che la successione ha **ordine di convergenza** p . La costante γ è detta **fattore di convergenza**.

- se $p = 1$ e $0 < \gamma < 1$ si dice che la convergenza è **lineare**;
- se $p = 1$ e $\gamma = 1$ si dice che la convergenza è **sublineare**;
- se $p > 1$ e $\gamma > 0$ si dice che la convergenza è **superlineare**.

Osservazione. Dalla definizione segue che esiste una costante c tale che, da un certo k in poi, si ha:

$$|e_{k+1}| \leq c |e_k|^p$$

Questa diseguaglianza misura la riduzione dell'errore assoluto ad ogni iterazione.

Esempio:

Il metodo di bisezione

$$|e_{k+1}| = \frac{1}{2} |e_k|^1$$

ha un ordine di convergenza lineare.

Teorema. Sia $\{x_k\}$ una successione generata da 5.2 convergente a x^* e sia $g(x)$ sufficientemente regolare in un intorno di x^* . Allora la successione ha ordine di convergenza p se e solo se

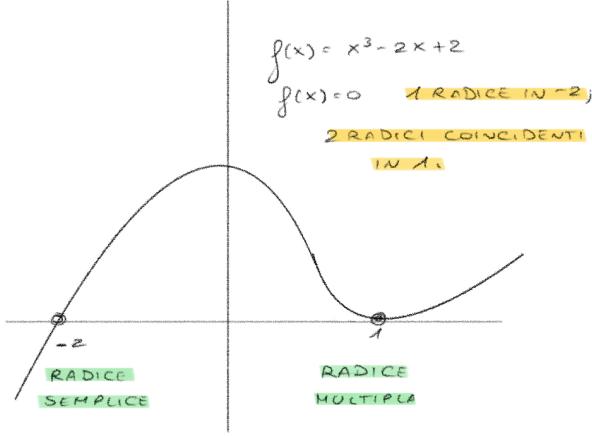
$$g'(x^*) = g''(x^*) = \dots = g^{(p-1)}(x^*) = 0, \quad g^{(p)}(x^*) \neq 0$$

Per determinare l'ordine di convergenza di un metodo di iterazione funzionale, si valuta x^* in $g(x)$ e tutte le sue derivate successive. L'ordine di convergenza p è determinato da tutte le derivate nulle che precedono l'ultima derivata non nulla $g^{(p)}(x^*)$.

Osservazione.

Poiché nel metodo di Newton abbiamo $g'(x^*) = 0$, il metodo ha una convergenza pari a 2.

Questo vale solo nell'ipotesi che $f'(x^*) \neq 0$, cioè se x^* sia una radice semplice di $f(x)$. Se invece la radice x^* ha una molteplicità $m > 1$, l'ordine di convergenza non sarà più 2. Il metodo rimane tuttavia convergente, ma con un ordine di convergenza ridotto a 1.



5.2.1 Metodo delle secanti

Il metodo di Newton è veloce ma richiede di conoscere la derivata $f'(x)$ della funzione $f(x)$ che si sta esaminando. Quando la derivata è difficile da calcolare o la sua valutazione è costosa, si può utilizzare il metodo delle secanti come alternativa.

Nel metodo delle secanti, invece di utilizzare la retta tangente alla curva della funzione in un punto x_k , si utilizza la retta secante che passa attraverso due punti $(x_{k-1}, f(x_{k-1}))$ e $(x_k, f(x_k))$ sulla curva. L'intersezione di questa retta con l'asse x fornisce il prossimo iterato x_{k+1} .

La formula per calcolare x_{k+1} nel metodo delle secanti è:

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

Questa si può anche vedere considerando il metodo di Newton in cui alla $f'(x_k)$ si sostituisce il rapporto incrementale.

Teorema. Se $f(x) \in C^2[a, b]$, $f(x^*) = 0$, $f'(x^*) \neq 0$ e $f''(x^*) \neq 0$, allora esiste un intervallo $I = [x^* - \rho, x^* + \rho]$ tale che $x_0, x_1 \in I$ ($x_0 \neq x_1$) allora la successione $\{x_k\}$ (generata dal metodo delle secanti) converge a x^* per $k \rightarrow \infty$.

Inoltre, il

$$\lim_{k \rightarrow \infty} \frac{|e_{k+1}|}{|e_k|^p} = \gamma \neq 0$$

dove $p = \frac{1}{2}(\sqrt{5} + 1) \approx 1.618\dots$

Il metodo delle secanti ha un ordine di convergenza che si colloca tra 1 e 2. Questo lo rende più efficiente di un algoritmo con convergenza lineare, ma meno efficiente del metodo di Newton in termini di velocità di convergenza.

Tuttavia, c'è un aspetto da considerare: nel metodo delle secanti, il valore x_{k-1} è già stato calcolato nell'iterazione precedente, quindi non è necessario rivalutarlo. Al contrario, il metodo di Newton richiede la valutazione sia della funzione che della sua derivata ad ogni iterazione.

In altre parole, sebbene il metodo di Newton possa convergere più rapidamente, ogni sua iterazione è computazionalmente più costosa rispetto al metodo delle secanti. Questo diventa particolarmente rilevante se la derivata della funzione è onerosa da calcolare. In tali casi, il costo computazionale associato al calcolo della derivata nel metodo di Newton potrebbe annullare i vantaggi della sua più rapida convergenza.

```

1 def secant(fn, x0, x1, tol):
2     n=0
3     max_iter=50
4     small_real=100*real_min
5
6     x_next = x1
7     x_curr = x0
8     fx_curr = fn(x0)
9     while abs(x_next - x_curr) > 2*u*min(abs(x_next), abs(x_curr)) and n < max_iter:
10        x_prev = x_curr
11        fx_prev = fx_curr
12        x_curr = x_next
13        fx_curr = fn(x_curr)
14        if (abs(fx_curr)>=small_real):
15            x_next = x_curr - fx_curr * (x_curr-x_prev) / (fx_curr-fx_prev)
16            n=n+1
17    return x_next

```

6 Algebra lineare numerica

Uno dei problemi più frequenti del calcolo scientifico è la soluzione di un sistema lineare. In forma matriciale può essere scritto come

$$Ax = b$$

dove A è una data matrice di ordine $n \times n$, b è un dato vettore colonna con n elementi ed x è il vettore delle incognite.

In algebra lineare si studiano metodi per risolvere sistemi lineari non singolari. Un metodo noto è quello di Cramer nel quale ogni componente della soluzione è espressa come:

$$x_i = \frac{\det \begin{bmatrix} a_1, \dots, \overset{i}{b}, \dots, a_n \end{bmatrix}}{\det [a_1, \dots, a_i, \dots, a_n]}$$

dove il numeratore si ottiene dal denominatore sostituendo alla i -ma colonna a_i la colonna b .

Esempio:

Se si cerca di risolvere un sistema di 20 equazioni con la regola di Cramer sarebbe necessario calcolare 21 determinanti di ordine 20. Per calcolare il determinante di una matrice $n \times n$, con lo sviluppo di Laplace, servono $(n - 1)n!$ moltiplicazioni; nel nostro esempio sarà $19 \cdot 20!$ e quindi l'intero sistema comporterà $19 \cdot 20! \cdot 21$ moltiplicazioni, più un ugual numero di addizioni. Su un Macbook M1, oggi si possono fare 2.6×10^{12} moltiplicazioni al secondo (2600 GFLOPs), così che, solo le moltiplicazioni richiederanno un tempo di calcolo di circa 11 anni.

Nell'esempio precedente abbiamo visto che l'uso della regola di Cramer per risolvere il sistema lineare $Ax = b$ è impraticabile dal punto di vista computazionale. Pertanto, in questo capitolo introduciamo due metodi alternativi per affrontare questo tipo di problemi.

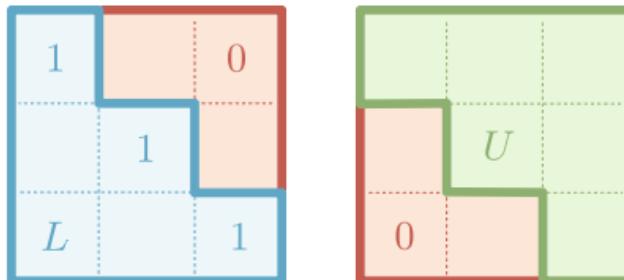
6.1 Fattorizzazione LU

L'obiettivo è fattorizzare una matrice quadrata A nel prodotto di una matrice triangolare inferiore L per una matrice triangolare superiore U :

$$A_{n \times n} = L_{n \times n} \cdot U_{n \times n}$$

con

$$L = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ l_{21} & 1 & 0 & \cdots & 0 \\ l_{31} & l_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \cdots & 1 \end{bmatrix} \quad U = \begin{bmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & u_{2n} \\ 0 & 0 & u_{33} & \cdots & u_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & u_{nn} \end{bmatrix}$$



Una volta ottenute le due matrici, è possibile riformulare il sistema lineare $Ax = b$ come:

$$L \cdot \underbrace{U \cdot \underbrace{x}_{y}}_{n \times 1} = b$$

Definiamo $y = U \times b$, e risolviamo il sistema in due passaggi:

1. Risolviamo $Ly = b$ utilizzando la sostituzione in avanti per trovare y .
2. Una volta trovato y , risolviamo $Ux = y$ utilizzando la sostituzione all'indietro, per ottenere x .

In questo modo, la soluzione del sistema originale $Ax = b$ viene ricondotta alla soluzione di due sistemi lineari più semplici e computazionalmente meno costosi.

6.1.1 Sostituzione in avanti

Per risolvere il primo sistema $Ly = b$, dove L è una matrice triangolare inferiore, si può fare riferimento al metodo discusso nel Paragrafo 3.2.1.

6.1.2 Sostituzione all'indietro

La sostituzione all'indietro è essenzialmente analogo alla sostituzione in avanti, ma si parte dall'ultima equazione e si procede all'indietro. Entrambi i metodi hanno un costo computazionale di $\mathcal{O}(n^2)$.

Consideriamo il seguente sistema lineare triangolare superiore:

$$\begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

Dopo aver calcolato x_3 , aggiorniamo il vettore y e risolviamo il sistema ridotto:

$$\begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} y_1 - u_{13}x_3 \\ y_2 - u_{23}x_3 \end{bmatrix}$$

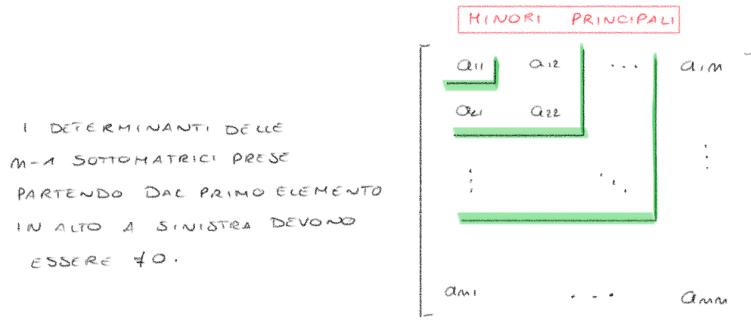
E così via, fino a quando tutte le incognite x sono state calcolate.

```

1 fn backward_substitution<const N: usize>(u: &SquareMatrix<N>, y: &Vector<N>) -> Vector<N> {
2     let mut x = zeros_vector::<N>();
3     let mut y = y.clone();
4
5     for k in (0..N).rev() {
6         x[k] = y[k] / u[k][k];
7         for i in 0..k {
8             y[i] = y[i] - u[i][k] * x[k];
9         }
10    }
11    x
12 }
```

6.1.3 Metodo di Gauss

Teorema. Se i minori principali di ordine di k di A per $k = 1, \dots, n-1$ sono diversi da zero, allora esiste una e una sola fattorizzazione LU di A .



Per illustrare il metodo di Gauss per la fattorizzazione LU , consideriamo una matrice $A_{3 \times 3}$:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Affinché la fattorizzazione LU sia possibile, è necessario che siano soddisfatte le seguenti condizioni:

$$a_{11} \neq 0 \quad \text{e} \quad \det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \neq 0$$

Il procedimento di fattorizzazione si svolge in $n-1$ passi, dove n è la dimensione della matrice. Per una matrice 3×3 , ci sono quindi 2 passi.

Prima di procedere, è utile notare come sono strutturate le matrici L_k che useremo. Ogni matrice L_k è una matrice identità modificata, in cui gli elementi l_{ik} nella k -esima colonna sotto la diagonale sono calcolati cambiando segno all'elemento a_{ik} della matrice A e dividendo per l'elemento diagonale corrispondente a_{kk} , cioè $l_{ik} = -\frac{a_{ik}}{a_{kk}}$. Questi elementi sono costruiti per annullare gli elementi sotto la diagonale nella k -esima colonna della matrice A corrente quando L_k viene moltiplicata per A .

1. **Primo passo:** costruiamo una matrice L_1 come segue:

$$L_1 = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{a_{21}}{a_{11}} & 1 & 0 \\ -\frac{a_{31}}{a_{11}} & 0 & 1 \end{bmatrix}$$

Notiamo che è essenziale $a_{11} \neq 0$ per evitare la divisione per zero.

Moltiplicando L_1 per A , otteniamo una nuova matrice A_1 , in cui tutti gli elementi nella prima colonna sotto a_{11} sono zero:

$$L_1 A = A_1 = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & \overset{(1)}{a_{22}} & \overset{(1)}{a_{23}} \\ 0 & \overset{(1)}{a_{32}} & \overset{(1)}{a_{33}} \end{bmatrix}$$

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

2. **Secondo passo:** costruiamo una matrice L_2 siffatta:

$$L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{\overset{(1)}{a_{32}}}{\overset{(1)}{a_{22}}} & 1 \end{bmatrix}$$

Moltiplicando L_2 per A_1 , otteniamo una nuova matrice A_2 , in cui tutti gli elementi nella seconda colonna sotto a_{22} sono zero:

$$L_2 A_1 = A_2 = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & \overset{(1)}{a_{22}} & \overset{(1)}{a_{23}} \\ 0 & 0 & \overset{(2)}{a_{33}} \end{bmatrix}$$

Dove $\overset{(2)}{a_{33}} = \overset{(1)}{a_{33}} - \frac{\overset{(1)}{a_{32}}}{\overset{(1)}{a_{22}}} \overset{(1)}{a_{23}}$.

$$L_2(L_1 A) = U$$

Dopo aver completato i passi della fattorizzazione, otteniamo U come A_2 . Osserviamo che le matrici L_1 e L_2 sono non singolari. Questo è garantito dal fatto che il determinante di ciascuna matrice, calcolato come il prodotto degli elementi della sua diagonale, è diverso da zero. Pertanto, entrambe le matrici sono invertibili.

$$A = L_1^{-1} L_2^{-1} U$$

Per ottenere le matrici L_1^{-1} e L_2^{-1} basta invertire il segno degli elementi al di fuori della diagonale nelle matrici originali L_1 e L_2 .

Infine, invece di eseguire un costoso prodotto matriciale, la matrice composta $L_1^{-1} L_2^{-1}$ può essere “assemblata” prendendo gli elementi sotto la diagonale da ciascuna delle matrici inverse e inserirli nella posizione corrispondente nella matrice L finale.

Esempio:

Consideriamo la matrice A :

$$A = \begin{bmatrix} 2 & 1 & 0 \\ 4 & 5 & 2 \\ 6 & 15 & 12 \end{bmatrix}$$

Verifica condizioni:

$$a_{11} = 2 \neq 0 \quad \text{e} \quad \det \begin{bmatrix} 2 & 1 \\ 4 & 5 \end{bmatrix} = 10 - 4 = 6 \neq 0$$

1° passo: L_1 deve essere triangolare inferiore e tale da rendere nulli gli elementi della prima colonna
 A sotto l'elemento $a_{11} = 2$

$$L_1 = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -3 & 0 & 1 \end{bmatrix} \quad L_1 A = A_1 = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 3 & 2 \\ 0 & 12 & 12 \end{bmatrix}$$

2° passo: L_2 deve essere triangolare inferiore e tale da rendere nulli gli elementi della seconda colonna
 A_1 sotto l'elemento $\overset{(1)}{a_{22}} = 3$.

$$L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -4 & 1 \end{bmatrix} \quad L_2 A_1 = A_2 = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 3 & 2 \\ 0 & 0 & 4 \end{bmatrix}$$

$$\textcolor{red}{U} = A_2 \quad \textcolor{red}{L} = L_1^{-1} L_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{bmatrix}$$

Se ora volessimo risolvere un sistema lineare $Ax = b$, dove b è un vettore arbitrario, potremmo farlo risolvendo i seguenti sistemi:

$$Ly = b \quad \text{e} \quad Ux = y$$

Nella implementazione effettiva si utilizza un'unica matrice U che all'inizio della k -ma interazione contiene A^k e viene quindi sovrascritta con $A^{(k+1)}$. Se la procedura viene portata a termine con successo, alla fine U conterrà esattamente la matrice U della fattorizzazione LU . Per quanto riguarda la matrice L , essa viene inizializzata come $L = I$ ed alla iterazione k -ma gli elementi sotto la diagonale della k -esima colonna di L vengono aggiornati.

$$l_{ij}^{(k)} = \begin{cases} 1 & \text{per } i = j \\ -\frac{a_{ij}^{(k-1)}}{a_{jj}^{(k-1)}} & \text{per } i > k \text{ e } j = k \\ 0 & \text{altrimenti} \end{cases} \quad a_{ij}^{(k)} = \begin{cases} a_{ij}^{(k-1)} & \text{per } i \leq k-1 \\ 0 & \text{per } i \geq k \text{ e } j \leq k-1 \\ a_{ij}^{(k-1)} + l_{ik}^{(k)} a_{kj}^{(k-1)} & \text{per } i \geq k \text{ e } j \geq k \end{cases}$$

$L_k = \begin{bmatrix} x & & & \\ x & x & & \\ & x & x & \\ & & x & x \end{bmatrix}$

$A_k = \begin{bmatrix} x & \dots & \dots & x \\ 0 & x & \dots & x \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \# & \# \end{bmatrix}$

K-ESIMA COLONNA

```

1 fn lu_decomposition<const N: usize>(a: &SquareMatrix<N>) -> (SquareMatrix<N>, SquareMatrix<N>)
2 {
3     let mut u = a.clone();
4     let mut l = identity_matrix();
5
6     for k in 0..N - 1 {
7         for i in k + 1..N {
8             l[i][k] = u[i][k] / u[k][k];
9             for j in 0..=k {
10                 u[i][j] = 0.;
11             }
12             for j in k + 1..N {
13                 u[i][j] -= l[i][k] * u[k][j];
14             }
15         }
16     }
17 }
```

```

12         }           u[i][j] -= l[i][k] * u[k][j];
13     }
14 }
15 (l, u)
16 }
17 }
```

Per determinare la suddetta fattorizzazione, sono necessarie $\frac{n(n-1)}{2}$ divisioni per calcolare gli elementi di L , e $(n-1)^2 + (n-2)^2 + \dots + 1 = \frac{(n-1)n(2n-1)}{6}$ addizioni e moltiplicazioni per gli elementi di U . In totale, il costo computazionale è dell'ordine di $\mathcal{O}(n^3)$.

Calcolo di A^{-1} . Consideriamo una matrice A che sia invertibile. Per calcolare A^{-1} in modo efficiente possiamo sfruttare la fattorizzazione LU . L'obiettivo è di risolvere il sistema lineare $AX = I$, dove I è la matrice identità. Per ogni colonna e della matrice identità I , eseguiamo i seguenti passaggi:

$$\begin{cases} Ly_i = e_i \\ Ux_i = y_i \end{cases}$$

La soluzione x sarà la i -esima colonna di A^{-1} . Combinando tutte le colonne x trovate, otteniamo A^{-1} .

La fattorizzazione LU ha un costo di $\mathcal{O}(n^3)$. Ogni sistema lineare da risolvere ha un costo di $\mathcal{O}(n^2)$, e dato che ci sono n sistemi da risolvere, il costo totale per questa parte è anche $\mathcal{O}(n^3)$. Sommando i due, il costo totale per calcolare A^{-1} è $\mathcal{O}(n^3)$.

Calcolo del $\det(A)$. Consideriamo una matrice A e procediamo con la sua fattorizzazione LU . Utilizzando il teorema di Binet, possiamo calcolare il determinante di A come segue:

$$\det(A) = \det(L) \times \det(U)$$

Dato che L è una matrice triangolare inferiore con tutti gli elementi della diagonale uguale a 1, il suo determinante sarà $\det(L) = 1$. Inoltre, il determinante della matrice triangolare U è dato dal prodotto degli elementi sulla diagonale. Di conseguenza:

$$\det(A) = \det(U) = \prod_{i=1}^n u_{i,i}$$

In questo modo, il calcolo del determinante di A ha una complessità computazionale di $\mathcal{O}(n^3)$.

6.1.4 Fattorizzazione LU con scambio delle righe e perno massimo

Consideriamo la matrice A :

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

In questo caso, $a_{11} = 0$, il che ci rende impossibile applicare una fattorizzazione LU . Tuttavia, effettuando uno scambio di righe, possiamo riscrivere la matrice come segue:

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

Ora, con $a_{11} \neq 0$, siamo di nuovo nelle condizioni di poter procedere con la fattorizzazione. Arricchiamo pertanto il metodo visto in precedenza con questa nuova informazione.

Teorema. Per ogni matrice $A_{n \times n}$ è sempre possibile trovare una matrice di permutazione P per cui PA sia fattorizzabile in LU , cioè

$$PA = LU$$

La matrice di permutazione P è una matrice di comodo ottenuta dalla matrice identità scambiando opportunamente alcune righe. Moltiplicare P per un'altra matrice A ha l'effetto di scambiare le righe di A .

Esempio:

Usiamo la matrice di permutazione P per scambiare la prima e la terza riga di A :

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}, \quad P = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

Moltiplicando P per A , otteniamo:

$$PA = \begin{bmatrix} 7 & 8 & 9 \\ 4 & 5 & 6 \\ 1 & 2 & 3 \end{bmatrix}$$

Se al passo k il perno a_{kk} è zero, cerchiamo un nuovo perno in una riga h successiva alla riga k e scambiamo le righe in modo che a_{hk} prenda il posto di a_{kk} . Tuttavia, se $a_{kk} = 0$ e tutti i possibili nuovi perni nelle righe successive sono anch'essi nulli, allora il procedimento si arresta. In questo caso, quanto trovato fino a quel momento rappresenta la fattorizzazione cercata.

Prima di formalizzare l'algoritmo di fattorizzazione LU con scambio delle righe, applichiamo l'analisi all'indietro per stimare la stabilità dell'algoritmo di fattorizzazione descritto.

Stabilità della fattorizzazione LU . Poiché le operazioni nell'algoritmo di fattorizzazione LU sono effettuate in aritmetica finita, i fattori L e U generati non sono esatti. In altre parole, otteniamo $\tilde{L} = L + \delta L$ e $\tilde{U} = U + \delta U$. Una volta ottenute le matrici \tilde{L} e \tilde{U} , possiamo considerarle come la fattorizzazione di una matrice A perturbata, $A + \delta A$, ottenuta attraverso un algoritmo esatto. Quindi abbiamo:

$$\begin{aligned} A + \delta A &= \tilde{L}\tilde{U} = (L + \delta L)(U + \delta U) \\ &= LU + L\delta U + \delta LU + \delta L\delta U \end{aligned}$$

Da questa relazione, notiamo che A e LU sono identici, quindi rimane:

$$\delta A = L\delta U + \delta LU + \delta L\delta U$$

Trascurando il prodotto di errori, poiché numericamente irrilevanti, otteniamo:

$$\delta A \approx L\delta U + \delta LU$$

da cui segue che se gli elementi di L e U sono grandi, essi amplificano l'errore introdotto in δA (gli errori di arrotondamento).

Perciò diremo che la fattorizzazione $A = LU$ è numericamente **stabile** se gli elementi di L ed U non sono troppo grandi rispetto agli elementi di A .

Definizione. Se esistono delle costanti a e b indipendenti dagli elementi di A tali che gli elementi della matrice L sono $\leq a$ e gli elementi della matrice U sono $\leq b$, allora la fattorizzazione LU è detta **stabile in senso forte**; se le costanti a e b dipendono dall'ordine di A , allora la fattorizzazione LU è detta **stabile in senso debole**.

Sebbene si sia visto che per ogni matrice, a patto di un'opportuna permutazione, esiste sempre la fattorizzazione LU , in generale questa potrebbe non essere stabile. Infatti gli elementi

$$l_{ik} = \frac{\underset{(k-1)}{a_{ik}}}{\underset{(k-1)}{a_{kk}}}$$

possono assumere valori molto grandi e quindi gli elementi di L possono crescere oltre ogni limite.

Quindi questo porta a modificare l'algoritmo di Gauss introducendo lo scambio delle righe non solo per individuare un pivot non nullo, ma quello massimo. Questo si realizza al passo k -esimo scegliendo come pivot fra gli $\underset{(k-1)}{a_{ik}}$ con $i = k, \dots, n$ il più grande in valore assoluto, così che a scambio delle righe effettuato risultì

$$\left| \underset{(k-1)}{a_{kk}} \right| \geq \left\{ \left| \underset{(k-1)}{a_{ik}} \right| \right\}_{i=k+1, \dots, n}$$

$$A_K = \left[\begin{array}{cccc} & \cdots & & \times \\ & \cdots & \cdots & \times \\ \times & \cdots & \cdots & \times \\ & \vdots & \ddots & \vdots \\ & \times & \cdots & \times \end{array} \right]$$

Applicando l'algoritmo di Gauss di scambio delle righe e perno massimo avremo che tutti gli elementi di L sono ≤ 1 , mentre non riesce a limitare gli elementi di U che possono crescere esponenzialmente con l'ordine di A . Infatti si ha che:

$$\max |u_{ij}| \leq 2^{n-1} \max |a_{ij}|$$

Perciò l'algoritmo genera una fattorizzazione **stabile in senso debole**.

Esempio:

$$A_0 = \begin{bmatrix} 2 & 3 & 0 \\ 4 & 1 & 4 \\ 6 & 3 & 3 \end{bmatrix}$$

$$I_{2,0}A_0 = \begin{bmatrix} 6 & 3 & 3 \\ 4 & 1 & 4 \\ 2 & 3 & 0 \end{bmatrix}, \quad L_1 = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{2}{3} & 1 & 0 \\ -\frac{1}{3} & 0 & 1 \end{bmatrix}, \quad L_1(I_{2,0}A_0) = \begin{bmatrix} 6 & 3 & 3 \\ 0 & -1 & 2 \\ 0 & 2 & -1 \end{bmatrix}$$

$$I_{2,1}A_1 = \begin{bmatrix} 6 & 3 & 3 \\ 0 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}, \quad L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{2} & 1 \end{bmatrix}, \quad L_2(I_{2,1}A_1) = \begin{bmatrix} 6 & 3 & 3 \\ 0 & 3 & 6 \\ 0 & 0 & \frac{3}{2} \end{bmatrix}$$

$$L_2 I_{2,1} L_1 I_{2,0} A = U$$

$$\underbrace{L_2}_{\tilde{L}_1} \underbrace{P_2 L_1 P_2}_{P} \underbrace{P_2 P_1}_{P_1} A = U$$

6.2 Condizionamento del problema $Ax = b$

In questa sezione si vuole esaminare come, perturbazioni sugli elementi della matrice A e sugli elementi del termine noto b influenzano la soluzione x del sistema lineare. Queste perturbazioni sono tipicamente dovute agli errori di approssimazione quando la matrice A ed il termine noto b vengono rappresentati con numeri finiti. Per essere in grado di stimare gli errori, dobbiamo introdurre una misura della *distanza fra vettori*.

Definizione. Siano $x, y \in \mathbb{R}^n$ e $\alpha \in \mathbb{R}$. Una **norma vettoriale** $\|\cdot\|$ è una funzione $\mathbb{R}^n \rightarrow \mathbb{R}$ che associa ad un vettore di \mathbb{R} un valore reale (la lunghezza del vettore). Tale funzione per essere una norma vettoriale deve soddisfare le tre seguenti proprietà:

1. $\|x\| \geq 0$ e $\|x\| = 0$ se e solo se $x = 0$
2. $\|x + y\| \leq \|x\| + \|y\|$ (diseguaglianza triangolare)
3. $\|\alpha x\| = |\alpha| \|x\|$

Le norme vettoriali più importanti sono:

- $\|x\|_1 = \sum_{i=1}^n |x_i|$
- $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2} = (x^T x)^{\frac{1}{2}}$
- $\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$

La $\|\cdot\|_2$ è una generalizzazione ad \mathbb{R}^n dell'usuale distanza in \mathbb{R}^2 ed è detta norma Euclidea.

Definizione. Una **norma matriciale** $\|\cdot\|$ è una funzione $\mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ che associa ad una matrice di $\mathbb{R}^{m \times n}$ un valore reale. Tale funzione, per essere una norma matriciale deve soddisfare le seguenti proprietà:

1. $\|A\| \geq 0$ e $\|A\| = 0$ se e solo se $A = 0$
2. $\|A + B\| \leq \|A\| + \|B\|$
3. $\|\alpha A\| = |\alpha| \|A\|$

Definizione (Norma matriciale indotta da una norma vettoriale). In alternativa, è possibile definire la norma matriciale a partire dalla norma vettoriale come

$$\|A\| = \max_{x \neq 0} \frac{\left\| \begin{matrix} A & x \\ (m \times n) & (n \times 1) \end{matrix} \right\|}{\|x\|}$$

Le norme indotte sono particolarmente utili perché godono delle seguenti proprietà:

- $\|Ax\| \leq \|A\| \|x\|$
- $\|AB\| \leq \|A\| \|B\|$

Le norme matriciali più importanti sono:

- $\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}|$
- $\|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}|$
- $\|A\|_2 = (\lambda_{\max}(A^T A))^{\frac{1}{2}}$ (norma spettrale)

Con le norme, possiamo introdurre i concetti di distanza in \mathbb{R}^n . Sia \tilde{x} un vettore soluzione calcolata e x un vettore soluzione esatta. Per una data norma vettoriale $\|\cdot\|$, si definisce

- **Errore assoluto**

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

- **Errore relativo**

$$\frac{\|\tilde{x} - x\|}{\|x\|}$$

Errore inerente. Affrontiamo ora lo studio dell'errore inerente del problema $Ax = b$ considerando separatamente eventuali perturbazioni sulla matrice A e sul vettore dei termini noti b .

- Introduciamo un vettore di perturbazione $\delta b \in \mathbb{R}^n$ sul termine noto; cerchiamo $x + \delta x \in \mathbb{R}^n$ soluzione del sistema perturbato

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

poiché $Ax = b$ risulterà

$$A\delta x = \delta b$$

da cui

$$\delta x = A^{-1} \delta b$$

Passando alle norme

$$\|\delta x\| = \|A^{-1} \delta b\| \leq \|A^{-1}\| \|\delta b\|$$

inoltre vale

$$\|b\| = \|Ax\| \leq \|A\| \|x\|$$

Allora

$$\begin{aligned} \frac{\|\delta x\|}{\|x\|} &\leq \frac{\|A^{-1}\| \|\delta b\|}{\|x\|} \\ &\leq \color{red} \|A^{-1}\| \|A\| \frac{\|\delta b\|}{\|b\|} \end{aligned}$$

Dunque il condizionamento di $Ax = b$ dipende dalla costante

$$K = \|A^{-1}\| \|A\|$$

detto numero di condizione di A .

- Analogamente, per una perturbazione δA su A .

Teorema. Sia A non singolare e sia $r = \|A^{-1}\| \|\delta A\| < 1$; allora la matrice $A + \delta A$ è non singolare, e

$$\|(A + \delta A)^{-1}\| \leq \frac{\|A^{-1}\|}{1 - r}$$

La soluzione del sistema perturbato

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

soddisfa

$$\leq \frac{K(A)}{1 - r} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right)$$

Se stiamo cercando di risolvere un sistema lineare $Ax = b$, è importante considerare che l'inserimento della matrice A in un calcolatore introduce delle perturbazioni. Queste potrebbero alterare la matrice, rendendola singolare e quindi non invertibile. La soluzione del sistema perturbato è garantita se la condizione $r < 1$ è soddisfatta.

L'errore relativo nella soluzione x è influenzato sia dagli errori nei dati di input A e b , sia da un fattore di amplificazione $\frac{K(A)}{1-r}$. Qui, $K(A)$ è il numero di condizione di A , una quantità che dovrebbe essere calcolata prima di risolvere il sistema, per valutare quanto il problema sia ben condizionato e decidere se procedere.

```

1 fn cond<const N: usize>(a: &SquareMatrix<N>, norm_type: NormType) -> f64 {
2     norm(&a.inverse().expect("The given matrix is not invertible"), norm_type) * norm(a,
3         norm_type)

```

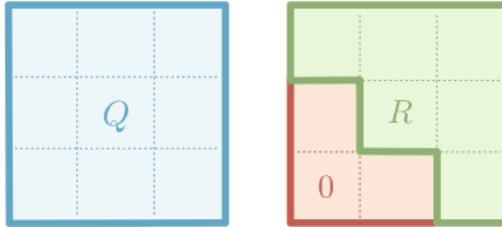
6.3 Fattorizzazione QR

L'obiettivo è fattorizzare una matrice quadrata A nel prodotto di una matrice ortogonale Q per una matrice triangolare superiore R :

$$A_{n \times n} = Q_{n \times n} \cdot R_{n \times n}$$

con

$$Q^T Q = I \quad \text{od anche} \quad Q^T = Q^{-1} \quad R = \begin{bmatrix} r_{11} & r_{12} & r_{13} & \cdots & r_{1n} \\ 0 & r_{22} & r_{23} & \cdots & r_{2n} \\ 0 & 0 & r_{33} & \cdots & r_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & r_{nn} \end{bmatrix}$$



Osservazione.

Sia Q ortogonale ed a un vettore $n \times 1$, allora $\|Qa\|_2 = \|a\|_2$.

In altre parole, applicare una matrice ortogonale Q a un vettore a non cambia la sua norma euclidea, ossia la sua distanza dall'origine.

Dimostrazione.

$$\begin{aligned} \|Qa\|_2^2 &= (Qa)^T (Qa) = (a^T Q^T)(Qa) = \\ &= a^T (Q^T Q)a = a^T Ia = a^T a = \|a\|_2^2 \end{aligned}$$

■

Una volta ottenute le due matrici, è possibile riformulare il sistema lineare $Ax = b$ come:

$$Q \cdot \underbrace{R \cdot \underbrace{x}_{y}}_{n \times n \times 1} = b$$

Definiamo $Rx = y$, e risolviamo il sistema in due passaggi:

1. Risolviamo il sistema $Qy = b$ sfruttando la definizione di Q ortogonale cioè $y = Q^T b$.
2. Una volta trovato y , risolviamo $Rx = y$ utilizzando la sostituzione all'indietro per ottenere x .

6.3.1 Matrici elementari di Householder

Nella fattorizzazione LU , facevamo uso di matrici elementari di Gauss per azzerare elementi e ottenere una forma triangolare. In modo analogo, nella fattorizzazione QR , utilizzeremo matrici elementari conosciute come matrici di Householder per raggiungere un obiettivo simile.

Per una matrice elementare di Householder si intende una matrice H tale che

$$Ha = \pm \|a\|_2 e_1 = \begin{bmatrix} \pm \|a\|_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Le matrici elementari di Householder sono sia simmetriche ($H = H^T$) che ortogonali.

Dato un vettore a , possiamo costruire una trasformazione di Householder nel modo seguente: definiamo il vettore di Householder v come

$$v = \begin{bmatrix} a_1 \pm \|a\|_2^2 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

Quindi, la matrice H è

$$H = I - \beta vv^T \quad \text{dove} \quad \beta = \frac{2}{\|v\|_2^2}$$

H è simmetrica per costruzione, essendo $I - \beta vv^T$.

Dimostriamo ora che H è ortogonale:

Dimostrazione.

$$\begin{aligned} H^T H &= (I - \beta vv^T)(I - \beta vv^T) \\ &= I - \beta vv^T - \beta vv^T + \beta vv^T \beta vv^T \\ &= I - 2\beta vv^T + \beta \underbrace{\frac{2}{\|v\|_2^2} vv^T}_{\|v\|_2^2} vv^T \\ &= I - 2\beta vv^T + 2\beta vv^T \\ &= I \end{aligned}$$

■

Infine, verifichiamo che, quando applicata al vettore a , la matrice H azzeri tutte le componenti tranne la prima:

$$\begin{aligned} Ha &= (I - \beta vv^T)a \\ &= a - \beta vv^T a \quad \text{dico che} \quad bv^T a = 1 \\ &= a - v \\ &= \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} - \begin{bmatrix} a_1 \pm \|a\|_2 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \pm \|a\|_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \pm \|a\|_2 e_1 \end{aligned}$$

verifichiamo $\beta v^T a = 1$

Dimostrazione.

$$\begin{aligned} v^T a &= [a_1 \pm \|a\|_2 \ a_2 \ \dots \ a_n] \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \\ &= a_1^2 \pm a_1 \|a\|_2 + a_2^2 + \dots + a_n^2 \\ &= \|a\|_2^2 \pm a_1 \|a\|_2 \\ &= \|a\|_2 (\|a\|_2 \pm a_1) \end{aligned}$$

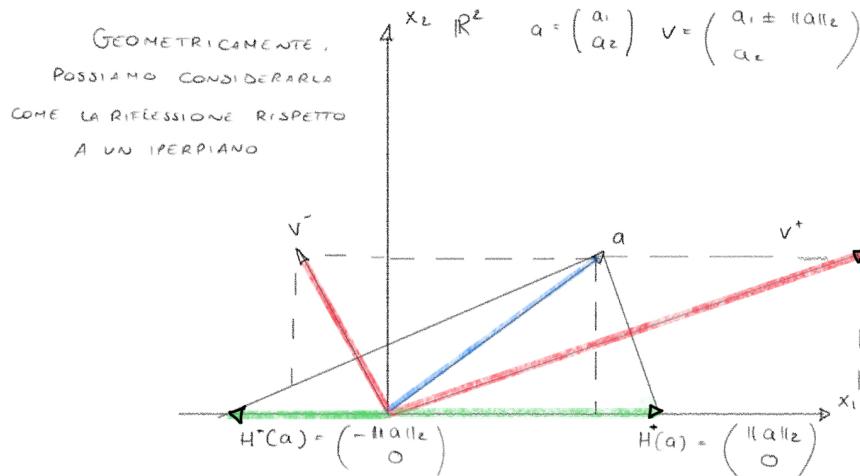
inoltre $\beta = \frac{2}{\|v\|_2^2}$ e

$$\begin{aligned}\|v\|_2^2 &= v^T v = [a_1 \pm \|a\|_2 \quad a_2 \quad \dots \quad a_n] \begin{bmatrix} a_1 \pm \|a\|_2 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \\ &= (a_1 \pm \|a\|_2)^2 = a_1^2 + \dots + a_n^2 \\ &= a_1^2 + 2a_1\|a\|_2 + \|a\|_2^2 + a_2^2 + \dots + a_n^2 \\ &= 2\|a\|_2^2 \pm 2a_1\|a\|_2 \\ &= 2\|a\|_2(\|a\|_2 \pm a_1)\end{aligned}$$

e quindi

$$\beta v^T a = \frac{2\|a\|_2(\|a\|_2 \pm a_1)}{2\|a\|_2(\|a\|_2 \pm a_1)} = 1$$

■



Esempio:

Dato il vettore $a = \begin{bmatrix} 72 \\ -144 \\ -144 \end{bmatrix}$, si determini la matrice H che azzera la seconda e terza componente.

1.

$$\|a\|_2 = \sqrt{72^2 + 144^2 + 144^2} = 216, \quad v = \begin{bmatrix} 72 + 216 \\ -144 \\ -144 \end{bmatrix} = \begin{bmatrix} 288 \\ -144 \\ -144 \end{bmatrix}$$

2.

$$\beta = \frac{2}{288^2 + 144^2 + 144^2} = \frac{1}{62808}$$

3.

$$\begin{aligned}H &= I - \beta vv^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \frac{1}{62208} \begin{bmatrix} 288 \\ -144 \\ -144 \end{bmatrix} \begin{bmatrix} 288 & -144 & -144 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} \frac{4}{3} & -\frac{2}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{1}{3} & \frac{1}{3} \\ -\frac{2}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \\ &= \begin{bmatrix} -\frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{bmatrix}\end{aligned}$$

4.

$$Ha = \begin{bmatrix} -\frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{bmatrix} \begin{bmatrix} 72 \\ -144 \\ -144 \end{bmatrix} = \begin{bmatrix} -216 \\ 0 \\ 0 \end{bmatrix}$$

6.3.2 Metodo di Householder

Procediamo ad illustrare come funziona l'algoritmo per fattorizzare QR una matrice $A_{n \times n}$. Si pone $A_0 = A$ e si costruisce una matrice H_1 tale che

$$H_1 A_0 = \begin{bmatrix} \|a_1^{(0)}\| & \times & \dots & \times \\ 0 & \vdots & & \vdots \\ \vdots & & & \\ 0 & \times & \dots & \times \end{bmatrix} = A_1$$

quindi una H_2 tale che

$$H_2 A_1 = \begin{bmatrix} \|a_1^{(0)}\| & \times & \times & \dots & \times \\ 0 & \|a_2^{(1)}\| & \times & \dots & \times \\ \vdots & 0 & \vdots & & \\ \vdots & & & & \\ 0 & 0 & \times & \dots & \times \end{bmatrix} = A_2$$

e così via fino ad una H_{n-1} tale che

$$H_{n-1} A_{n-2} = \begin{bmatrix} \|a_1^{(0)}\| & \times & \times & \dots & \times \\ 0 & \|a_2^{(1)}\| & \times & \dots & \times \\ \vdots & 0 & \ddots & & \vdots \\ \vdots & & \ddots & & \times \\ 0 & 0 & \dots & & \|a_n^{(n-2)}\| \end{bmatrix} = R$$

così che

$$H_{n-1} H_{n-2} \dots H_2 H_1 A = R$$

Ma le H_k matrici sono tutte ortogonali ed il prodotto di matrici ortogonali è ancora una matrice ortogonale, da cui

$$A = \underbrace{H_1^T H_2^T \dots H_{n-2}^T H_{n-1}^T}_Q R$$

$$A = QR$$

$$A_0 = A$$

per $k = 1, \dots, m-1$

$$v_k = a_k \pm \|a_k\|_2$$

$$b_k = \frac{2}{v_k^T v_k}$$

$$h_k = I - b_k v_k v_k^T$$

$$A_k = h_k A_{k-1}$$

$$R = A_{m-1}$$

$$Q = H_1^T H_2^T \dots H_{m-1}^T$$

$$\left\{ \begin{array}{l} + \text{ se } a_{kk}^{(k-1)} \geq 0 \\ - \text{ se } a_{kk}^{(k-1)} < 0 \end{array} \right.$$

IN MODO DA FARE SOMME TRA DUE QUANTITA' DI SEGNIO CONCORDE ED EVITARE ERRORI DI CANCELLAZIONE NUMERICA.

$$v_k = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \begin{matrix} (k-1) & (k-1) \\ a_{kk} + \|a_k\|_2 & a_{k+1,k} \\ (k-1) & (k+1,k) \\ \vdots & \vdots \\ (k-1) & a_{mk} \end{matrix} \end{bmatrix} \} \text{ k-1 POSIZIONI}$$

Esempio:

Si calcoli la fattorizzazione QR della matrice

$$A_0 = A = \begin{bmatrix} 72 & -144 & -144 \\ -144 & -36 & -360 \\ -144 & -360 & 450 \end{bmatrix}$$

1.

$$\beta_1 = \frac{1}{62208}, \quad v_1 = \begin{bmatrix} 288 \\ -144 \\ -144 \end{bmatrix}$$

$$H_1 = I - \beta_1 v_1^T v_1 = \frac{1}{6} \begin{bmatrix} -2 & 4 & 4 \\ 4 & 4 & -2 \\ 4 & -2 & 4 \end{bmatrix}, \quad H_1 A_0 = \begin{bmatrix} -216 & -216 & 108 \\ 0 & 0 & 486 \\ 0 & -324 & 324 \end{bmatrix} = A_1$$

2.

$$\beta_2 = \frac{1}{104976}, \quad v_2 = \begin{bmatrix} 0 \\ 324 \\ -324 \end{bmatrix}$$

$$H_2 = I - \beta_2 v_2^T v_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad H_2 A_1 = \begin{bmatrix} -216 & -216 & 108 \\ 0 & -324 & 324 \\ 0 & 0 & -486 \end{bmatrix} = A_2$$

per cui $R = A_2$ e

$$Q = H_1 H_2 = \frac{1}{6} \begin{bmatrix} -2 & 4 & 4 \\ 4 & -2 & 4 \\ 4 & 4 & -2 \end{bmatrix}$$

Il metodo di Householder, per risolvere il sistema lineare $Ax = b$, può essere implementato senza calcolare effettivamente le matrici H_k . La procedura è la seguente:

Iniziamo considerando la matrice T_0 , che è composta dalla matrice A e dal vettore b come colonna aggiuntiva:

$$T_0 = [A_0 | b_0] = [A | b]$$

A ogni passo k , costruiamo i valori β_k , v_k , e un vettore y_k^T calcolato come $y_k^T = v_k^T T_k$. La matrice T_{k+1} , viene quindi aggiornata come segue:

$$T_{k+1} = T_k - \beta_k v_k y_k^T$$

Dopo $n - 1$ iterazioni, otteniamo la matrice T_{n-1} , che è nella forma:

$$T_n = [A_n | b_n] = [R | b_n]$$

Con questa matrice, possiamo risolvere il sistema $Rx = b_n$, che è equivalente al sistema iniziale $Ax = b$.

Costo computazionale e stabilità della fattorizzazione QR . La complessità computazionale della fattorizzazione QR è di $\mathcal{O}(n^3)$, o più specificamente di $\frac{2}{3}n^3$, se non si tiene conto del calcolo delle matrici H_k e dei loro prodotti. Questo è circa il doppio del costo associato alla fattorizzazione LU . Tuttavia, un vantaggio significativo della fattorizzazione QR è la sua maggiore stabilità numerica.

Procedendo con un'analisi all'indietro, esattamente come già fatto nel caso della fattorizzazione LU si ha che

$$A + \delta A = (Q + \delta Q)(R + \delta R)$$

$$A + \delta A = QR + Q\delta R + \delta QR + \delta Q\delta R$$

Trascurando il prodotto di errori, poiché numericamente irrilevanti, otteniamo:

$$\delta A = Q\delta R + \delta QR$$

Gli elementi della matrice Q sono limitati dalla sua norma, che è 1 poiché Q è una matrice ortogonale. Gli elementi della matrice R , invece sono limitati da $\sqrt{n} \max_{i,j} |a_{i,j}|$.

Da cui risulta che l'algoritmo di fattorizzazione QR è stabile in senso debole con estremo \sqrt{n} , che risulta comunque più stabile dell'algoritmo LU essendo $\sqrt{N} \ll 2^{n-1}$.

6.4 Metodi iterativi

Per risolvere un sistema lineare $Ax = b$, oltre ai metodi diretti, si possono utilizzare anche i metodi iterativi, che risultano essere particolarmente convenienti se la matrice è molto grande o sparsa, ovvero se il numero degli elementi non nulli di A è dell'ordine della dimensione della matrice. Questo perché i metodi diretti possono generare matrici intermedie con elementi non nulli dove nella matrice originale erano presenti elementi non nulli. Di conseguenza, la matrice perde la sua sparsità durante le operazioni intermedie.

Idea. Sia $A_{n \times n}$ una matrice e si consideri la decomposizione di A nella forma

$$A = M - N$$

dove M è una matrice non singolare e facilmente invertibile. Sostituendo tale decomposizione nel sistema $Ax = b$ si ha

$$(M - N)x = b$$

$$Mx = Nx + b$$

ed essendo M non singolare

$$x = M^{-1}Nx + M^{-1}b$$

Posto $P = M^{-1}N$ e $q = M^{-1}b$ si ottiene il seguente sistema

$$x = Px + q$$

equivalente ad $Ax = b$.

Questo ci consente di introdurre un metodo iterativo come segue:

$$x^{(k)} = Px^{(k-1)} + q \quad k = 1, 2, \dots \quad (6.1)$$

con x_0 assegnato.

L'obiettivo è che la successione di vettori $\{x^{(k)}\}$ converga alla soluzione x^* del sistema lineare, ovvero

$$\lim_{k \rightarrow \infty} x^k = x^*$$

Definizione (Convergenza di una successione di vettori). Una successione $\{x^{(k)}\}$ di vettori di \mathbb{R}^n si dice convergente al vettore $x^* \in \mathbb{R}^n$ se esiste una norma vettoriale per cui

$$\lim_{k \rightarrow \infty} \|x^{(k)} - x^*\| = 0 \quad \forall x^{(0)} \text{ iniziale assegnato.}$$

In questo caso si scrive $\lim_{k \rightarrow \infty} x^{(k)} = x^*$

Teorema. Il metodo iterativo 6.1 è convergente se e solo se il **raggio spettrale** $\rho(P)$, definito come l'autovalore massimo in modulo della matrice P è minore di 1.

Questo teorema fornisce una condizione necessaria e sufficiente che permette di determinare a priori la convergenza del metodo iterativo.

È utile ricordare che gli autovalori di A sono quei valori λ che soddisfano $Ax = \lambda x$. Per trovare gli autovalori, costruiamo il polinomio caratteristico di grado n in λ . Gli autovalori sono le radici di questo polinomio e il raggio spettrale è il modulo del più grande di questi autovalori. Tuttavia, calcolare il raggio spettrale per una grande matrice è oneroso, quindi si cercano condizioni più facilmente verificabili.

Teorema. Se esiste una norma matriciale indotta $\|\cdot\|$, per cui $\|P\| < 1$ allora il metodo iterativo 6.1 è convergente.

Questo secondo teorema offre un criterio alternativo basato sulla norma della matrice, che è più facile da calcolare. Va notato, tuttavia, che questa è una condizione solo sufficiente, non necessaria. Vale a dire che se $\rho(P) < 1 < \|P\|$, il secondo teorema non fornisce alcuna informazione sulla convergenza, anche se il metodo può comunque convergere. Ma è certamente più pratico del calcolo di tutti gli autovalori e del raggio spettrale.

Costo computazionale. In un metodo iterativo, ad ogni iterazione, il costo computazionale è dominato dalle operazioni di moltiplicazione della matrice P per un vettore, che ha una complessità di $\mathcal{O}(n^2)$. Tuttavia, se la matrice P è sparsa, le operazioni sono limitate agli elementi non nulli della matrice. Di conseguenza, la complessità diventa dell'ordine di $\mathcal{O}(\text{numero di elementi non nulli}^2)$. Quindi, in questi casi, i metodi iterativi possono risultare competitivi con quelli diretti.

Test di arresto. Per interrompere il metodo iterativo è necessario stabilire un test di arresto appropriato. In maniera analoga a quanto fatto per le equazioni non lineari, possiamo considerare due differenti misure per valutare la convergenza del metodo:

$$\|x^{(k)} - x^{(k-1)}\| \leq tol \quad \text{e} \quad \frac{\|x^{(k)} - x^{(k-1)}\|}{\min(\|x^{(k)}\|, \|x^{(k-1)}\|)} \leq tol$$

Va notato che queste due condizioni non assicurano che la soluzione ottenuta sia approssimata con una precisione pari a tol .

Condizionamento e stabilità. Nel contesto dei metodi iterativi, il condizionamento della soluzione di un sistema lineare è influenzato non solo dal numero di condizionamento $K(A)$ della matrice originale A , ma anche dal raggio spettrale $\rho(P)$ della matrice iterativa P . Più $\rho(P)$ si avvicina a 1, più il problema è mal condizionato. Per quanto riguarda l'errore algoritmico, i metodi iterativi hanno il vantaggio di essere generalmente meno sensibili alla propagazione degli errori rispetto ai metodi diretti. Questo accade perché ogni vettore $x^{(k)}$ nell'iterazione può essere visto come il risultato di una singola iterazione a partire dal vettore precedente $x^{(k-1)}$. Di conseguenza, $x^{(k)}$ è influenzato principalmente dagli errori di arrotondamento generati nell'ultima iterazione, non da un accumulo di errori attraverso iterazioni multiple come nei metodi diretti.

6.4.1 Metodi di Jacobi e Gauss-Seidel

Si consideri la decomposizione della matrice A

$$A = D - L - U$$

dove

$$D = \begin{bmatrix} a_{11} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{nn} \end{bmatrix} \quad L = \begin{bmatrix} 0 & \dots & 0 \\ -a_{21} & \ddots & \vdots \\ \vdots & \ddots & \\ -a_{n1} & \dots & -a_{nn-1} & 0 \end{bmatrix} \quad U = \begin{bmatrix} 0 & -a_{12} & \dots & -a_{1n} \\ \vdots & \ddots & \ddots & \\ 0 & \dots & & 0 \end{bmatrix}$$

Scegliendo

$$M = D \quad N = L + U$$

si ottiene il metodo di Jacobi, mentre scegliendo

$$M = D - L \quad N = U$$

si ottiene il metodo di Gauss-Seidel.

Nel metodo di Jacobi, la matrice M è la diagonale D di A . Nel metodo di Gauss-Seidel $M = D - L$ è la matrice triangolare inferiore di A . In entrambi i casi, affinché i metodi siano applicabili, è necessario che tutti gli elementi della diagonale della matrice A siano diversi da zero. Questo assicura che M sia non singolare e quindi invertibile.

- **Jacobi:** Indicando con J la matrice di iterazione del metodo di Jacobi, dalla $P = M^{-1}N$ si ha

$$J = D^{-1}(L + U)$$

per cui la 6.1 diviene

$$x^{(k)} = Jx^{(k-1)} + q = D^{-1}(L + U)x^{(k-1)} + b$$

con $q = D^{-1}b$.

$$= \begin{bmatrix} \frac{1}{a_{11}} & & & & \\ & \frac{1}{a_{22}} & & & \\ & & \ddots & & \\ & & & \frac{1}{a_{nn}} & \end{bmatrix} \left(\begin{bmatrix} 0 & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & 0 & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & 0 & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & 0 \end{bmatrix} \begin{bmatrix} x_1^{(k-1)} \\ \vdots \\ x_n^{(k-1)} \end{bmatrix} + \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} \right)$$

- **Gauss-Seidel.**