Analisi Numerica

Professoressa Angeles Martinez Calomardo

Riassunto da: Cortinovis Nicola

Laurea Triennale in Intelligenza Artificiale e Data Analytics 2021/2022
Università di Trieste II semestre II anno

Contents

1 Note sul riassunto

In questo riassunto c'è tutto quello che è stato visto a lezione eccezion fatta per le parti riguardanti i laboratori su matlab (è già tutto scritto nei *lucidi*) e per gli esercizi del tutorato.

2 Aritmetica di macchina

2.1 Introduzione

Sappiamo che la soluzione di un calcolatore ad un problema può avere degli errori, questi errori sono di vario tipo:

- **Errori dovuti alla modellazione matematica del problema reale ed errori presenti nei dati sperimentali**. Questo è un errore "a monte", cioè un errore antecedente all'attività del calcolatore. Tra questi rientrano anche gli errori commessi con gli strumenti di misurazione.
- **Errori di troncamento** legati alla trasformazione di un problema matematico di dimensione **infinita** in uno di dimensione **finita**. Per un calcolatore ovviamente il concetto di infinito non esiste, ogni operazione "all'infinito" va quindi ridotta ad un calcolo finito introducendo tassativamente un errore. Un esempio può essere la somma di Riemann per il calcolo integrale: teoricamente sommiamo infiniti rettangolini, nella pratica no.
- **Errori di arrotondamento** dovuti al fatto che sul calcolatore si può rappresentare soltanto un **sottoinsieme finito** di $\mathbb{R}$. Ad esempio per 0.1 il passaggio da base 10 a base 2 (binario) è problematico.

2.2 Sistema posizionale e conversione di base

Definizione 1: Rappresentazione posizionale

Fissati un numero $B \in \mathbb{N}$ con $B > 1$ che chiamiamo **base** ed un numero $x \in \mathbb{R}$ con cifre finite d_k $k = -m, -m+1, \dots, -1, 0, 1, \dots n-1, n$ si definisce x_B la **rappresentazione posizionale** di x in base B

$$x_B = (-1)^S(d_n d_{n-1} \dots d_1 d_0 d_{-1} \dots d_m) = (-1)^S\left(\sum_{k=-m}^{n} d_k B^k\right) \quad m \neq 0$$

Il simbolo S viene utilizzato per specificare il **segno**: se $S = 0$ allora il numero è positivo, altrimenti se $S = 1$ è negativo.

Esempio:

$$(141.2123)_{10} = (-1)^0 \cdot (1 \cdot 10^2 + 4 \cdot 10^1 + 1 \cdot 10^0 + 2 \cdot 10^{-1} + 1 \cdot 10^{-2} + 2 \cdot 10^{-3} + 3 \cdot 10^{-4})$$
$$(-10010.001)_2 = (-1)^1 \cdot (1 \cdot 2^4 + 1 \cdot 2^1 + 1 \cdot 2^{-3})$$

In generale è più comodo dividere le sommatorie che rappresentano la parte intera e quella frazionaria:

$$x_B = (-1)^s \underbrace{\left(\sum_{k=0}^{n} d_k B^K\right)}_{Parte\ intera} + \underbrace{\left(\sum_{k=1}^{\infty} d_{-k} B^{-k}\right)}_{Parte\ frazionaria}$$

La serie che rappresenta la parte frazionaria converge perché è una serie maggiorata da una geometrica con modulo della ragione < 1. In maniera rigorosa se noi sostituiamo a tutti i d_{-k}

il valore massimo che può assumere, cioè $B-1$ otteniamo:

$$\sum_{k=1}^{\infty}(B-1)B^{-k} = (B-1)\sum_{k=1}^{\infty}B^{-k}$$

Chiaramente vale che

$$\left(\sum_{k=1}^{\infty}d_{-k}B^{-k}\right) \leq (B-1)\sum_{k=1}^{\infty}B^{-k} = \frac{B^{-1}}{1-B^{-1}}$$

Un numero irrazionale ha la parte frazionaria infinità **in tutte le basi** ($\pi, \sqrt{2}\ldots$). Per i numeri razionali almeno una base deve avere parte frazionaria finita, ma potrebbero esserci basi dove il numero ha parte frazionaria infinita.

$$x = \frac{1}{3} \quad x_{10} = 0.\overline{3} \quad x_3 = 0.1$$
$$x = 0.1 \quad x_{10} = 0.1 \quad x_2 = 0.0\overline{0011}$$

La conversione da base 2 a base 10 è molto semplice, basta esprimere il numero binario in notazione posizionale

Esempio

$$11101.01 = (-1)^0 \cdot (1\cdot 2^4 + 1\cdot 2^3 + 1\cdot 2^2 + 1\cdot 2^0 + 1\cdot 2^{-2}) = 16+8+4+1+0.25 = 29.25$$

La conversione da base 10 a base 2 è lievemente più complicata perché prevede due procedimenti diversi per la parte frazionaria e per quella intera, vediamo un esempio:

Esempio Vogliamo convertire 389.1_{10} in base 2:

Parte intera

	quoziente	resto	
$389 \div 2$	194	1	
$194 \div 2$	97	0	
$97 \div 2$	48	1	
$48 \div 2$	24	0	
$24 \div 2$	12	0	
$12 \div 2$	6	0	
$6 \div 2$	3	0	
$3 \div 2$	1	1	
$1 \div 2$	0	1	$\uparrow$

Parte frazionaria

$.1 \times 2 = 0.2$	$\to 0$	$\downarrow$
$.2 \times 2 = 0.4$	$\to 0$	
$.4 \times 2 = 0.8$	$\to 0$	
$.8 \times 2 = 1.6$	$\to 1$	
$.6 \times 2 = 1.2$	$\to 1$	
$.2 \times 2 = 0.4$	$\to 0$	$.2 \times 2$ è già apparso
$.4 \times 2 = 0.8$	$\to 0$	
$.8 \times 2 = 1.6$	$\to 1$	
$.6 \times 2 = 1.2$	$\to 1$	

In questo caso otteniamo che $389.1_{10} = 110000101.0\overline{0011}_2$

2.3 Rappresentazione dei numeri in virgola mobile normalizzata

Data una base B ogni numero reale o intero diverso da 0 si può scrivere in virgola mobile normalizzata come:

$$x = (-1)^s B^e \left(\sum_{k=1}^{\infty} d_k B^{-k} \right) \text{ con } \begin{cases} d_1 > 0 \\ 0 \leq d_k \leq B-1 \\ e \in \mathbb{Z} \end{cases}$$

In maniera compatta possiamo riscrivere.

$$x = \pm p \cdot B^e, \text{ dove } B^{-1} \leq p \leq 1$$

La differenza importante tra virgola mobile e virgola mobile normalizzata è che la prima cifra dopo il . è **sempre diversa da 0**. Indichiamo con p la **mantissa** e con e **l'esponente**

Esempio

$$x = 0.00745_{10} \quad \text{diventa} \quad 0.745 \cdot 10^{-2}$$
$$x = 70408.102_{10} \quad \text{diventa} \quad 0.70408102 \cdot 10^{5}$$
$$x = 11001.111_{2} \quad \text{diventa} \quad 0.11001111 \cdot 2^{5}$$

La rappresentazione in virgola mobile normalizzata **garantisce l'unicità della rappresentazione** per tutti i numeri.

2.4 Rappresentazione dei numeri nel calcolatore

Definizione 2: Numeri macchina

I numeri che riusciamo a rappresentare nel calcolatore sono detti **numeri macchina**. Dedichiamo un numero finito t di cifre per la mantissa ed un numero finito e di cifre per codificare l'esponente ($L \leq e \leq U$). Data una base B, fissato t e fissati $L < 0$ e $U > 0$ definiamo l'**insieme di numeri macchina** $\mathbb{F}(B, t, L, U)$:

$$\mathbb{F}(B,t,L,U) = \{x \mid x = (-1)^S B^e \left(\sum_{k=1}^{t} d_k B^{-k} \right)\} \cup \{0\} \quad \text{con} \quad \begin{cases} d_1 > 0 \\ 0 \leq d_k \leq B-1 \\ L \leq e \leq U \end{cases}$$

Nota che:

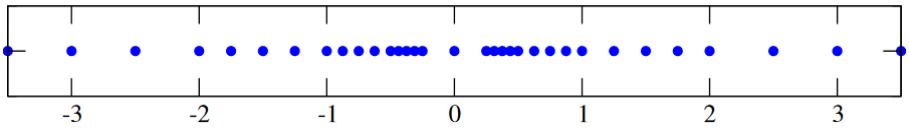
- L sta per **lower bound**, U sta per **upper bound** e rappresenta l'intervallo di valori interi che e può assumere
- L'insieme $\mathbb{F}$ (sta per floating) definisce l'**aritmetica di macchina**;
- Non c'è lo 0 perché la mantissa deve avere avere la prima cifra dopo la parte decimale diversa da 0, non è quindi un numero rappresentabile in virgola mobile normalizzata. Deve essere quindi aggiunto manualmente.
- La condizione $d_1 > 0$ è quella che garantisce l'**unicità** della rappresentazione.

Un altro modo più compatto per esprimere un numero macchina è:

$$x = (-1)^s \cdot (0.d_1 d_2 \dots d_t) \cdot B^e = (-1)^s \cdot \sum_{k=1}^{t} (d_k \cdot B^{-k}) \cdot B^e$$

Esempio $\mathbb{F}(2, 3, -1, 2)$ Qui $\mathbb{F}$ è un insieme di numeri macchina in base 2 con 3 cifre per mantissa, le possibili mantisse sono quindi in totale 4: 0.100, 0.101, 0.110, 0.111. Ciò è dovuto dalla prima cifra dopo il . che è sempre diversa da 0, e quindi in base 2 è forzatamente 1.

Per ciascuna mantissa è possibile abbinare uno tra $U - L + 1 = 2 - (-1) + 1 = 4$ esponenti, cioè gli interi nel range di $L = -1$ e $U = 2$: $[-1, 2] = (-1, 0, 1, 2)$. La distanza ogni volta che cambia l'esponente raddoppia, infatti notiamo che i numeri sono più addensati quanto più sono piccoli e la loro distanza aumenta man mano che aumenta il loro valore assoluto. In totale contiamo 16 valori positivi diversi, 16 valori negativi opposti a quelli positivi e lo 0 per un totale di 33 numeri diversi.



I puntini blu rappresentano i numeri macchina possibili dato l'insieme $\mathbb{F}(2, 3, -1, 2)$

2.5 Standard IEEE-754r

Per codificare un numero macchina abbiamo bisogno di 3 informazioni:

- il segno, (**1 bit**)
- le cifre della mantissa (**t bit**)
- l'esponente

Ogni numero macchina occupa uno spazio di memoria di 32 bit in **singola precisione** o 2 spazi consecutivi di 32 bit (64 bit) in **doppia precisione**. Esiste anche la half precision (16 bit). Nella realtà la più usata è la doppia tuttavia se la precisione non è fondamentale conviene usare la singola o la half diminuendo la dimensione dei dati per guadagnare in termini di spazio.

- **Singola precisione**: dedichiamo 32 bit per ciascun numero macchina distribuendoli in questo modo:

1	8	23
s	esponente	mantissa

  - L'insieme dei numeri macchina a singola precisione è $\mathbb{F}(2, 24, -126, 127)$.
  - Dato che $L = -126$ e $U = 127$ abbiamo un totale di $2^8 = 256$ esponenti possibili, 2 esponenti sono riservati ad uso speciale (NaN ed ∞).
  - Con i 23 bit disponibili si codificano 24 cifre della mantissa, un bit è nascosto (vedremo poi)
  - I numeri rappresentabili in singola precisione sono:

$$\underbrace{2}_{\text{segno}} \cdot \underbrace{(U - L + 1)}_{\text{Esponenti possibili}} \cdot \underbrace{(B-1) \cdot B^{t-1}}_{\text{Mantisse possibili}} + \underbrace{1}_{\text{lo zero}} = 2 \cdot 2541 \cdot 2^{23} \approx 4.2789 \cdot 10^9$$

    Nota: **Non** è il valore massimo esprimibile in singola precisione, ma bensì la **cardinalità** dell'insieme $\mathbb{F}(2, 23, -126, 127)$

- **Doppia precisione**: dedichiamo 64 bit per ciascun numero macchina distribuendoli in questo modo:

1	11	52
s	esponente	mantissa

  - L'insieme dei numeri macchina a doppia precisione $\mathbb{F}(2, 53, -1022, 1023)$
  - Con i 52 bit disponibili si codificano 53 cifre della mantissa, un bit è nascosto
  - Dato che $L = -1022$ e $U = 1023$ abbiamo un totale di 2^{11} esponenti possibili, 2 riservati per uso speciale (NaN e ∞).
  - I numeri rappresentabili in doppia precisione (cardinalità di $\mathbb{F}(2, 53, -1022, 1023)$) sono:

$$\underbrace{2}_{\text{segno}} \cdot \underbrace{(U - L + 1)}_{\text{Esponenti possibili}} \cdot \underbrace{(B-1) \cdot B^{t-1}}_{\text{Mantisse possibili}} + \underbrace{1}_{\text{lo zero}} = 2 \cdot 20461 \cdot 2^{52} \approx 1.8438 \cdot 10^1 9$$

In realtà i bit della mantissa sono rispettivamente 24 e 53, infatti si aggiunge un bit a sinistra che si assume essere sempre 1 (se la base è 2), nel caso di $\mathbb{F}(2, 53, -1022, 1023)$ un numero non nullo normalizzato si scrive come:

$$x = (-1)^s \cdot (1 + f) \cdot 2^{e-bias}$$

La mantissa è ora $1.d_1, d_2, d_\tau$, $t = 1 + \tau$ con $\tau = 52$. Anche l'esponente ha una rappresentazione diversa, infatti in genere si preferisce usare una rappresentazione in **eccesso a N** dove $N = U$

$$e^* = e + N$$

Questa scelta ci permette di effettuare un confronto tra due numeri in questo formato in maniera più semplice ed efficace confrontando prima i due esponenti e trattandoli come numeri interi.

Esempio Sia $x = 126$ il numero che vogliamo rappresentare secondo lo standard IEEE (singola precisione) allora dobbiamo:

- Passare in formato binario $126_{10} = (1111110)_2$
- Passare in virgola mobile normalizzata in standard IEEE $(1111110)_2 = 1.111110 \cdot 2^6$
- Calcolare il valore in eccesso $N = U = 127$ dell'esponente, $e^* = 6 + 127 = 133$ e passare in formato binario $133_{10} = 10000101_2$
- Codificarlo come stringa binaria:

$$\underbrace{0}_{s} \quad \underbrace{10000101}_{esponente} \quad \underbrace{111110}_{valore} \quad \underbrace{00000000000000000}_{17\ zeri}$$

Possiamo determinare quali sono il **massimo** ed il **minimo** numero rappresentabile dato un insieme macchina $\mathbb{F}(B, t, L, U)$:

- Il massimo si calcola ponendo tutti i bit della mantissa a $B - 1$ e scegliendo $e = U$. In doppia precisione $\mathbb{F}(2, 53, -1022, 1023)$ il massimo numero rappresentabile è:

$$(2 - 2^{-52}) \cdot 2^{1023} = 1.7977 \cdot 10^{308} \quad \text{("va saputo")}$$

- Il minimo si calcola ponendo tutti i bit della mantissa a 0 tranne il primo e scegliendo $e = L$. In doppia precisione $\mathbb{F}(2, 53, -1022, 1023)$ il minimo numero rappresentabile è:

$$1 \cdot 2^{-1022} = 2.2251 \cdot 10^{-308} \quad \text{("va saputo")}$$

2.6 Errore assoluto ed errore relativo di arrotondamento

Dato un $\mathbb{F}(B,t,L,U)$ la rappresentazione di un numero reale risulta problematica quando $x = pB^e$ (ricordando che p è la mantissa) ha più di t cifre nella mantissa, in questi casi il numero viene **approssimato** con un numero macchina che chiamiamo **fl(x)**. Esistono 2 metodi principali per approssimare:

- **Troncamento**, cancella nella mantissa p la parte che eccede la t-esima cifra;
- **Arrotondamento**, alla mantissa p aggiunge $\frac{B}{2}B^{-(t+1)}$ e poi si tronca alla t-esima cifra.

Esempio Supponiamo che $t = 6$ e $x = 0.745645897$

$$\begin{cases} \text{Troncamento: } fl(x) = tr(0.745645897) = 0.745645 \\ \text{Arrotondamento: } fl(x) = tr(0.745645897 + 0.0000005 = tr(0.745646397) = 0.745646 \end{cases}$$

Dove $\frac{B}{2}B^{-(t+1)} = \frac{10}{2} * 10^{-7}$.

In soldoni arrotondare equivale ad aggiungere 1 alla t-esima cifra della mantissa (d_t) se la cifra successiva $(d_{t+1}) \geq \frac{B}{2}$, altrimenti t rimane invariata e si tronca e basta.

Definizione 3: Overflow e Underflow

In $\mathbb{F}(B,t,L,U)$ nell'approssimare x con $fl(x)$ definiamo:

- Un errore di **Overflow** se l'esponente $e > U$ ed il numero viene rappresentato come Inf;
- Un errore di **Underflow** se l'esponente $e < L$ ed il numero viene rappresentato come 0.

Definizione 4: Errore assoluto e relativo

Sia x un numero reale e x^* una sua approssimazione allora definiamo:

- **Errore assoluto** la quantità $|x - x^*|$. E' la differenza tra il valore vero ed il valore approssimato, dà un'informazione molto vaga sulla dimensione dell'errore.
- **Errore relativo** la quantità $\frac{|x-x^*|}{|x|}$ con $x \neq 0$. Questa quantità dà un'informazione molto più dettagliata sulla dimensione dell'errore

Concettualmente la definizione rimane la stessa anche per enti diversi dai $\mathbb{R}$ come vettori, funzioni etc... basta sostituire i valori assoluti con la appropriata norma (vedrai dopo).

Esempio Supponiamo $x = 0.456789 \cdot 10^{-30}$ e $x^* = 0.6 \cdot 10^{-30}$, calcoliamo errore assoluto e relativo:

$$\begin{cases} \text{Errore assoluto } |x - x^*| = 0.143211 \cdot 10^{-30} \\ \text{Errore relativo } \frac{|x-x^*|}{|x|} = 0.313517 \end{cases}$$

Inizialmente l'errore assoluto potrebbe sembrarci molto piccolo, ma quando calcoliamo quello relativo notiamo che è maggiore del 31% quindi in realtà abbiamo commesso un errore grande.

2.6.1 Errori di rappresentazione

In generale vale che per l'**errore assoluto**:

- Ogni errore assoluto di **troncamento** è sicuramente limitato superiormente da:

$$|x - fl(x)| = |pB^e - \overline{p}B^e| = |p - \overline{p}|B^e \leq B^{-t}B^e$$

Dimostrazione 1: Massimo errore assoluto nel troncamento

Vale che se $x = pB^e$ dove ricordiamo che $p = 0.d_1 \dots d_t d_{t+1} d_{t+2} \dots \infty$ allora sicuramente

$$\sum_{k=t+1}^{\infty} d_k \cdot B^{-k} \leq (B-1) \sum_{k=t+1}^{\infty} B^{-k} = (B-1) \cdot \frac{B^{-(t+1)} - B^{-\infty}}{1 - B^{-1}} = (B-1) \cdot \frac{B^{-(t+1)})}{1 - \frac{1}{B}} = B^{-t}$$

Dove in sostanza abbiamo detto che la parte che rimane facendo la differenza tra p e $\overline{p}$ è $d_{t+1} \dots$ che sicuramente è minore od uguale al caso "massimo" dove tutte le d_k con $k \geq t+1$ sono uguali a $B-1$

- Ogni errore assoluto di **arrotondamento** è sicurmanete limitato superiormente da:

$$|x - fl(x)| = |pB^e - \overline{p}B^e| = |p - \overline{p}|B^e \leq \frac{B}{2} B^{-(t+1)} B^e$$

Dimostrazione 2: Massimo errore assoluto nell'arrotondamento

Si distinguono 2 casi, quando la $d_{t+1} < \frac{B}{2}$ e quando la $d_{t+1} \geq \frac{B}{2}$. Se $d_{t+1} < \frac{B}{2}$ allora il caso peggiore lo si ha quando $d_{t+1} = \frac{B}{2} - 1$ e tutti gli altri sono $(B-1)$:

$$(\frac{B}{2} - 1)B^{-(t+1)} + (B-1) \sum_{k=t+2}^{\infty} B^{-k}$$

Se $d_{t+1} \geq \frac{B}{2}$ il caso peggiore lo si ha quando $d_{t+1} = \frac{B}{2}$ e tutte le altre successive uguali a 0 cioè si pone esattamente a metà

Similmente per gli errori **relativi** vale che:

- Ogni errore relativo di **troncamento** è limitato superiormente da:

$$\frac{|x - fl(x)|}{|x|} = \frac{|pB^e - \overline{p}B^e|}{pB^e} = \frac{|p - \overline{p}|}{p} \leq \frac{B^{-t}}{B^{-1}} = B^{1-t}$$

  In $\mathbb{F}(2, 53, -1022, 1023)$ questo numero equivale a 2^{-52}. Dove per ottenere l'errore massimo abbiamo preso la mantissa a denominatore più piccola possibile in virgola mobile normalizzata

- Ogni errore relativo di **arrotondamento** è limitato superiormente da:

$$\frac{|x - fl(x)|}{|x|} = \frac{|pB^e - \overline{p}B^e}{pB^e|} = \frac{|p - \overline{p}|}{p} \leq \frac{\frac{B}{2}B^{-(t+1)}}{B^{-1}} = \frac{1}{2}B^{1-t}$$

  In $\mathbb{F}(2, 53, -1022, 1023)$ questo numero equivale a 2^{-53}. Per la scelta della mantissa vale come sopra.

Generalmente la maggior parte dei sistemi implementa la tecnica di arrotondamento perché tende a produrre errori più piccoli.

2.7 Distanza tra numeri macchina e precisione di macchina

Definizione 5: Precisione macchina

Definiamo **precisione di macchina** il valore:

$$u = \frac{1}{2} \cdot B^{1-t} \qquad u \text{ sta per unit-roundoff}$$

La precisione di macchina rappresenta il massimo errore relativo ottenibile approssimando un $x \in \mathbb{R}$ ad un $fl(x)$ per arrotondamento (forse domanda esame). Vale che:

- Per la singola precisione $\mathbb{F}(2, 24, -126, 127)$ si ha che $u = 2^{-24} \approx 5.96 \cdot 10^{-8}$
- Per la doppia precisione $\mathbb{F}(2, 53, -1022, 1023)$ si ha che $u = 2^{-53} \approx 1.11 \cdot 10^{-16}$

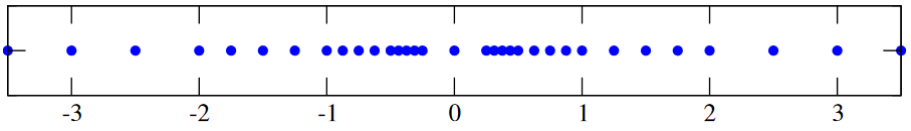
Definizione 6: Distanza assoluta

Dato un insieme macchina $\mathbb{F}(B, t, L, U)$ la distanza tra un numero $x \in \mathbb{F}$ ed il suo successivo x_+ **non** è una quantità infinitamente piccola, ma un valore **ben determinato**. La **distanza assoluta** Δx vale:

$$\begin{aligned} x &= -(1)^s \cdot (1 + 0.d_1 d_2 \dots d_\tau) \cdot B^e \\ x_+ &= -(1)^s \cdot (1 + 0.d_1 d_2 \dots d_\tau + 1) \cdot B^e \\ \Delta x &= |x_+ - x| = B^{-\tau} \cdot B^e = B^{e-\tau} \end{aligned}$$

Questa distanza gode di alcune proprietà:

- E' **uguale** per tutti i numeri macchina aventi lo stesso esponente
- Incrementando/decrementando di un'unità l'esponente incrementiamo/decrementiamo di un fattore pari alla base della distanza assoluta tra due numeri macchina consecutivi. Ad esempio nel binario moltiplico o divido per 2.



Definizione 7: Distanza relativa

Possiamo calcolare la **distanza relativa** dividendo quella assoluta per il valore assoluto di x:

$$\frac{|x_+ - x|}{|x|} = \frac{B^{e-\tau}}{p \cdot B^e} = \frac{B^{-\tau}}{p}$$

Vale che la distanza relativa tra due numeri macchina consecutivi ha un andamento periodico (?).

La distanza relativa massima tra due numeri macchina la otteniamo quando la mantissa $p = 1.000\ldots0$ e la definiamo come:

$$\varepsilon_M = B^{-\tau}$$

In IEEE-754r $\varepsilon_M = 2^{-52}$, notiamo che $\varepsilon_M = 2u$ cioè il doppio della precisione di macchina. In generale abbiamo che nello standard IEEE-754r se usiamo la doppia precisione avremo una precisione di macchina $u = 2^{-53} \approx 1.11 \cdot 10^{-16}$ che implica un errore nella 16° cifra (praticamente sempre).

2.8 Aritmetica di macchina e propagazione degli errori

Gli errori di rappresentazione si propagano quando svogliamo delle operazioni aritmetiche tra numeri macchina, definiamo le solite operazioni ma svolte tra numeri macchina con questa notazione.

Definizione 8: Aritmetica di macchina

Dati $x, y \in \mathbb{R}$

- Somma $\oplus$ e vale che $x \oplus y = fl(fl(x) + fl(y))$
- Sottrazione $\ominus$ e vale che $x \ominus y = fl(fl(x) - fl(y))$
- Prodotto $\otimes$ e vale che $x \otimes y = fl(fl(x) \cdot fl(y))$
- Divisione $\oslash$ e vale che $x \oslash y = fl(fl(x) \backslash fl(y))$

Cioè arrotondiamo x ed y, svogliamo l'operazione ed arrotondiamo il risultato.

Definizione 9: Errore relativo delle operazioni macchina

Definiamo l'**errore relativo** introdotto dalle operazioni macchina:

$$\varepsilon^{\oplus}_{x,y} \frac{|(x+y) - (x \oplus y)|}{|x+y|}$$

Ovviamente è definito in maniera analoga anche per le altre operazioni

Possiamo dare dei limiti superiori per gli errori relativi di tutte le operazioni (considerando somma e sottrazione come una singola operazione).

$$\begin{aligned}
\varepsilon^{\oplus}_{x,y} &\leq \left|\frac{x}{x+y}\right|\varepsilon_x + \left|\frac{y}{x+y}\right|\varepsilon_y \\
\varepsilon^{\otimes}_{x,y} &\lessapprox \varepsilon_x + \varepsilon_y \\
\varepsilon^{\oslash}_{x,y} &\leq |\varepsilon_x - \varepsilon_y|
\end{aligned}$$

Posto che $\varepsilon_x, \varepsilon_y < \mathbf{u}$ (precisione di macchina).

Dimostrazione 3

Dimostriamo che

$$\varepsilon_{x,y}^{\oplus} \leq |\frac{x}{x+y}|\varepsilon_x + |\frac{y}{x+y}|\varepsilon_y$$

Assumiamo che $x \neq 0$ e $y \neq 0$ e che $fl(fl(x) + fl(y)) = fl(x) + fl(y)$ cioè che la somma di due numeri arrotondati è già "perfetta" e non serve ri-arrotondarla (questa condizione rende solo la dimostrazione più semplice, non è necessaria). Vale che:

$$\begin{aligned}
\varepsilon_{x,y}^{\oplus} &= \frac{|(x+y) - (x \oplus y)|}{|x+y|} = \frac{|(x+y) - fl(fl(x) + fl(y))}{|x+y|} \\
&=_1 \frac{|(x+y) - (fl(x) + fl(y))|}{|x+y|} \leq_2 \frac{|(x - fl(x)|}{|x+y|} + \frac{|(y - fl(y)|}{|x+y|} \\
&= \frac{|(x - fl(x)||x|}{|x+y||x|} + \frac{|(y - fl(y)||y|}{|x+y||y|} \\
&= \underbrace{\frac{|(x - fl(x)|}{|x|}}_{\text{Errore relativo x}} \frac{|x|}{|x+y|} + \underbrace{\frac{|(y - fl(y)|}{|y|}}_{\text{Errore relativo y}} \frac{|y|}{|x+y|} = \varepsilon_x \frac{|x|}{|x+y|} + \varepsilon_y \frac{|y|}{|x+y|}
\end{aligned}$$

Dove per $=_1$ abbiamo usato l'assunzione e per $\leq_2$ la diseguaglianza triangolare.

Dimostrazione 4

Dimostriamo che

$$\varepsilon_{x,y}^{\otimes} \lessapprox \varepsilon_x + \varepsilon_y$$

Come sopra assumiamo che $x \neq 0$ e $y \neq 0$ e che $fl(fl(x) + fl(y)) = fl(x) + fl(y)$ (di nuovo questa assunzione è solo per semplificare la dimostrazione)

$$\begin{aligned}
\varepsilon_{x,y}^{\otimes} &= \frac{|(x \cdot y) - (x \otimes y)|}{|x \cdot y|} = \frac{|(x \cdot y) - fl(fl(x) \cdot fl(y))|}{|x \cdot y|} = \frac{|(x \cdot y) - fl(x) \cdot fl(y)|}{|x \cdot y|} \\
&= \frac{|x \cdot y - x \cdot fl(y) + x \cdot fl(y) - fl(x) \cdot fl(y)|}{|x \cdot y|} = \frac{|x \cdot (y - fl(y)) + (x - fl(x)) \cdot fl(y)|}{|x \cdot y|} \\
&\leq_1 \frac{|\not{x} \cdot (y - fl(y))|}{|\not{x} \cdot y|} + \frac{|(x - fl(x)) \cdot \cancel{fl(y)}|}{|x \cdot \not{y}|} \text{ assumo che } \frac{fl(y)}{y} \approx 1 \\
&= \underbrace{\frac{|y - fl(y)}{|y|}}_{\text{errore relativo x}} + \underbrace{\frac{|x - fl(x)|}{|x|}}_{\text{errore relativo y}} = \varepsilon_x + \varepsilon_y
\end{aligned}$$

Dove per $\leq_1$ abbiamo usato la diseguaglianza triangolare

Vediamo come alcune proprietà delle operazioni con i numeri $\mathbb{R}$ non valgono in $\mathbb{F}$:

- Non vale la proprietà associativa $(a \oplus b) \oplus c \neq a \oplus (b \oplus c)$. Questa proprietà viene infranta quando svolgiamo operazioni tra numeri che distano molto tra loro oppure per problemi

di overflow o underflow.

$$(1 \oplus 10^{-15}) \ominus 1 = 1.11 \cdot 10^{-15} \quad (1 \ominus 1) \oplus 10^{-15} = 10^{-15}$$
$$a = 1.0 \cdot 10^{308} \; b = 1.1 \cdot 10^{308} \; c = -1.001 \cdot 10^{308}$$
$$a \oplus (b \oplus c) = 1.0 \cdot 10^{308} \oplus (0.99 \cdot 10^{307}) = 1.099 \cdot 10^{308}$$
$$(a \oplus b) \oplus c = Inf \oplus c = Inf$$

- Non esiste un **unico** elemento neutro per la somma

2.9 Cancellazione numerica, Stabilità di un algoritmo e Condizionamento di un problema

Notiamo come le operazioni di prodotto e divisione introducano un errore nell'ordine di **u**, mentre potenzialmente la somma (sottrazione) possono causare errori molto più grandi, in particolare ciò si verifica quando $x \approx -y$, questo fenomeno prende il nome di cancellazione numerica. Formalmente:

Definizione 10: Cancellazione numerica

Definiamo **cancellazione numerica** la perdita di cifre significative dovute alla sottrazione di due numeri quasi uguali.

Esempio Supponiamo di lavorare con $\mathbb{F}(10, 5, *, *)$ e con $a = 0.73415507$ e $b = 0.73415448$ segue che $fl(a) = 0.73416$ e $fl(b) = 0.73415$. La differenza tra i valori esatti è $a-b = 0.59 \cdot 10^{-6}$, tuttavia nell'aritmetica di macchina abbiamo $fl(fl(x) - fl(y)) = 0.00001 = 10^{-5}$ potrebbe sembrare piccolo tuttavia studiando l'errore relativo notiamo che:

$$\frac{|(a-b) - (a \ominus b)|}{|a-b|} = \frac{|0.59 \cdot 10^{-6} - 10^{-5}|}{0.59 \cdot 10^{-6}} = \frac{0.941}{0.059} = 15.949 = 1595\%$$

Cioè una schifezza, questa aritmetica ha una precisione di macchina **u**$= 5 \cdot 10^{-5}$ ed in linea di massima consideriamo accettabili errori vicini alla precisione di macchina.

Definizione 11: Algoritmo stabile

Un metodo numerico (algoritmo, formula etc...) si dice **stabile** se non propaga errori inevitabili dovuti alla rappresentazione dei numeri nel calcolatore. Altrimenti si dice **instabile**.

La cancellazione numerica è una delle cause principali dell'instabilità degli algoritmi, per evitarla si trasformano formule in altre più stabili.

Esempio Prendiamo la formula $\sqrt{x+\delta} - \sqrt{x}$ abbiamo un problema di cancellazione numerica per $\delta \to 0$, tuttavia possiamo aggirarlo razionalizzando:

$$\sqrt{x+\delta} - \sqrt{x} = \frac{\sqrt{x+\delta} + \sqrt{x}}{\sqrt{x+\delta} + \sqrt{x}} = \frac{\delta}{\sqrt{x+\delta} + \sqrt{x}}$$

Per rendercene conto vediamo che in doppia precisione per $x = 1$ e $\delta = 10^{-14}$ abbiamo:

$$\sqrt{1+10^{-14}} - \sqrt{1} = 4.884998130835069 \cdot 10^{-15} \quad \varepsilon = 0.023 = 2.3\%$$
$$\frac{10^{-14}}{\sqrt{1+10^{-14}} + \sqrt{1}} = 4.99999999999999 \cdot 10^{-15} \quad \varepsilon = 1.58 \cdot 10^{-16}$$

Esempio Supponiamo di voler risolvere l'equazione

$$ax^2 + \frac{b}{a}x + \frac{c}{a} = 0 \quad \text{equivalente a} \quad x^2 + 2px - q \text{ con } \begin{cases} \frac{b}{a} = 2p \\ \frac{c}{a} = -q \end{cases}$$

Sappiamo che la formula risolutiva di un'equazione di secondo grado è:

$$x_{1,2} = \frac{-2p \pm \sqrt{4p^2 + 4q}}{2} = -p \pm \sqrt{p^2 + q}$$

La cancellazione numerica si verifica quando $q << p$ cioè q è tanto più piccolo di p (tanto più piccolo significa che la somma tra p^2 e q si posiziona nell'intervallo tra p^2 ed il successivo numero macchina portando per forza ad una perdita di cifre). Quando poi faccio la differenza tra i due valori abbiamo la cancellazione numerica. Esattamente come fatto sopra una soluzione al problema è razionalizzare:

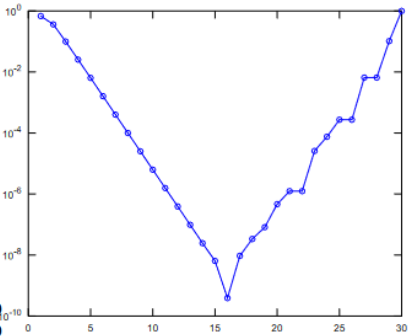
$$x_1 = -p + \sqrt{p^2 + q} = \frac{(-p + \sqrt{p^2 + q})(p + \sqrt{p^2 + q})}{(p + \sqrt{p^2 + q})} = \frac{q}{(p + \sqrt{p^2 + q})}$$

Esempio Vogliamo approssimare il valore di π con questa formula ricorsiva:

$$z_n = \begin{cases} z_n = 2 \quad \text{se } n = 2 \\ z_{n+1} = 2^{n-0.5}\sqrt{1 - \sqrt{1 - 4^{1-n}\ z_n^2}} \end{cases}$$

Detta $r = 4^{1-n} z_n^2$ l'applicazione della formula ricorsiva ci porta ad ottenere i seguenti valori:

$n+1$	r	$1-\sqrt{1-r}$	z_{n+1}	$\frac{\lvert z_{n+1} - \pi \rvert}{\pi}$
...	...	...	...	...
10	**1.505e-04**	**7.529e-05**	**3.14157294036**	**6.27e-06**
11	**3.764e-05**	**1.882e-05**	**3.14158772527**	**1.57e-06**
12	**9.412e-06**	**4.706e-06**	**3.14159142150**	**3.92e-07**
13	**2.353e-06**	**1.176e-06**	**3.14159234561**	**9.80e-08**
14	**5.882e-07**	**2.941e-07**	**3.14159257654**	**2.45e-08**
15	**1.470e-07**	**7.353e-08**	**3.14159263346**	**6.41e-09**
16	**3.676e-08**	**1.838e-08**	**3.14159265480**	**3.88e-10**
17	**9.191e-09**	**4.595e-09**	**3.14159264532**	**2.63e-09**
18	**2.297e-09**	**1.148e-09**	**3.14159260737**	**1.47e-08**
19	**5.744e-10**	**2.872e-10**	**3.14159291093**	**8.19e-08**
...	...	...	...	...
28	**2.220e-15**	**1.110e-15**	**3.16227766016**	**6.58e-03**
29	**5.551e-16**	**3.330e-16**	**3.46410161513**	**1.03e-01**
30	**1.665e-16**	**1.110e-16**	**4.00000000000**	**2.73e-01**
31	**5.551e-17**	**0.000e+00**	**0.00000000000**	**1.00e+00**
32	**0.000e+00**	**0.000e+00**	**0.00000000000**	**1.00e+00**



Abbiamo un episodio di cancellazione numerica perché per n sempre più grandi $r \to 0$ (4^{1-n} tende a 0), quando è abbastanza piccolo svolgiamo la differenza tra 1 ed un numero molto vicino ad 1 e quindi si verifica la cancellazione numerica.

Esempio Supponiamo di voler calcolare la seguente successione di integrali definiti:

$$\begin{cases} I_0 = \frac{1}{e}\int_0^1 e^x dx = 1 - \frac{1}{e} \approx 0.632121 \\ \\ I_n = \frac{1}{e}\int_0^1 x^n e^x dx \quad (n \geq 0) \end{cases}$$

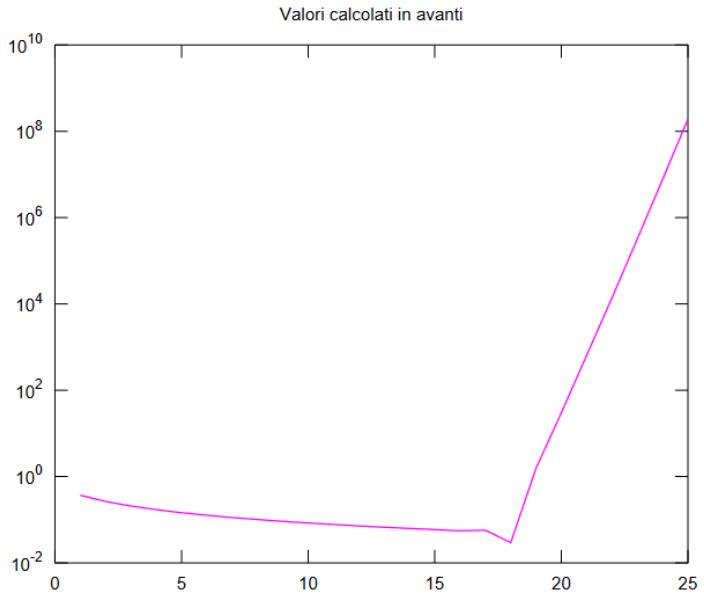
Posto $n \geq 1$ ed integrando per parti otteniamo la seguente formula ricorsiva:

$$I_n = \frac{1}{e}\left\{ \left[x^n e^x \right]_0^1 - n\int_0^1 x^{n-1} e^x dx \right\} = 1 - nI_{n-1}$$

Che possiamo riscrivere come

$$\begin{cases} I_0 \approx 0.632121 \text{ se n} = 0 \\ I_n = 1 - nI_{n-1} \text{ se } n \geq 1 \end{cases}$$

Notiamo come $0 < I_n < 1$, vediamo come il comportamento generi degli errori enormi superata una certa soglia, tuttavia questo è un esempio di algoritmo **instabile** ma **non** per colpa della cancellazione numerica.



La formula $I_n = 1 - nI_{n-1}$ è instabile perché già al primo passo svogliamo ovviamente un'approssimazione $I_0 = \frac{1}{e}$ che avrà un suo errore, per ciascun passo successivo questo errore viene propagato e moltiplicato per n volte. Infatti se riscriviamo la formula "vera" in quella che gestisce il calcolatore abbiamo che:

$$(I_n + \varepsilon_n) = 1 - n(I_{n-1} + \varepsilon_{n-1})$$

A questo punto possiamo sottrarre la formula vera a quella del calcolatore per quantificare l'errore:

$$\varepsilon_n = -n\varepsilon_{n-1} \quad \text{per induzione otteniamo} \quad |\varepsilon_n| = n!|(\varepsilon_0)|$$

Già per I_{20} c'è un errore $\varepsilon_{20} = 20!\varepsilon_0 = 2.7 \cdot 10^2$

Questo tipo di approccio al problema è detto "in avanti", nel senso che viene costruito dal basso $I_n = 1 - nI_{n-1}$. Un approccio migliore è quello "all'indietro" che possiamo ricavarci da quella in avanti ottenendo:

$$I_{n-1} = \frac{1 - I_n}{n}$$

L'idea qui è di partire abbastanza lontani dal valore che vogliamo calcolare e tornare indietro fino all'I_{n-1} che ci interessa. (Es vogliamo I_{50} allora partiamo da I_{90} e scendiamo). Questa nuova formula è **stabile** perché invece di aumentare l'errore lo diminuisce, infatti con un procedimento analogo a quello fatto precedentemente per isolare l'errore notiamo che:

$$\varepsilon_{n-1} = \frac{-1}{n}\varepsilon_n$$

Scelto un m di partenza che riteniamo appropriato scendiamo fino a quando $m - k$ non è l'iterazione che ci interessa. L'errore non si propaga ma bensì si riduce:

$$|\varepsilon_{m-1}| = \frac{|\varepsilon_m|}{m} \quad |\varepsilon_{m-2}| = \frac{|\varepsilon_m|}{m(m-1)} \quad \dots \quad |\varepsilon_{m-k}| = \frac{|\varepsilon_m|}{m(m-1)\dots(m-k+1)}$$

Per rendere l'idea: se vogliamo calcolarci I_{25} e partiamo da $I_{40} = 0.5$ l'errore iniziale $\varepsilon_0 < 0.5$ viene abbattuto di un fattore pari a $40 \cdot 39 \cdot \ \dots \ \cdot 27 \cdot 26 \approx 5.26 \cdot 10^{22}$

Definizione 12: Condizionamento di un problema

Un problema si dice **mal condizionato** se a **piccole variazioni** dei dati corrispondono **grandi variazioni** nei risultati

Notiamo immediatamente che qui non parliamo più dell'algoritmo per risolvere un problema ma del problema stesso. Non esiste algoritmo, per quanto stabile, che possa risolvere correttamente il problema. Vale che un problema si dice **ben condizionato** se non è mal condizionato. Una delle aree matematiche dove questo problema è più marcato è la risoluzione di sistemi.

Esempio

$$\begin{cases} x + y = 2 \\ 1001x + 1000y = 2001 \end{cases}$$

Questo semplice sistema ha come soluzioni $x = 1$ e $y = 1$, tuttavia se perturbiamo il coefficiente x della prima equazione di una quantità relativamente piccola come 0.01 otteniamo un nuovo sistema:

$$\begin{cases} 1.01x + y = 2 \\ 1001.01x + 1000y = 2001 \end{cases}$$

Ora questo sistema ha come soluzioni $x = -0.11111111$ e $y = 2.11222222$ con un errore relativo pari a :

$$err_x = \frac{1 + 0.11111111}{1} = 1.111111111 \qquad err_y = \frac{|1 - 2.11222222|}{1} = 1.11222222$$

Entrambi gli errori sono maggiori del 100%. Una piccola variazione ha causato un errore enorme e quindi il problema è mal condizionato.

Definizione 13: Numero di condizionamento

Possiamo quantificare quanto un problema è suscettibile alle piccole variazione dei dati con il **numero di condizionamento**.

Consideriamo il problema di valutare una funzione in un punto x e assumiamo che questa sia derivabile in un intorno del punto x. Se il dato x subisce una **perturbazione** Δx per il Teorema di Lagrange vale che:

$$f(x + \Delta x) - f(x) = \Delta x f'(\xi)$$

Indicando con $\Delta y = f(x + \Delta x) - f(x)$ l'**errore assoluto** sul valore della funzione possiamo definire quello **relativo** in questo modo:

$$\left|\frac{\Delta y}{y}\right| = \left|\frac{\Delta x f'(\xi)}{y}\right| = \frac{|\Delta x|}{|x|}\frac{|x \cdot f'(\xi)|}{|y|}$$

Vale che per $\Delta x \to 0$ ho $\xi \to x$. Possiamo riscrivere il tutto in questo modo:

$$\left|\frac{\Delta y}{y}\right| = K(f, x)\frac{|\Delta x|}{|x|}$$

dove con $K(f, x)$ indichiamo $\frac{|x \cdot f'(\xi)|}{|y|}$ cioè il **numero di condizionamento**

Esempio Supponiamo di avere la funzione:

$$f(x) = \sqrt{1 - x^2}$$

Di questa volgiamo calcolarci il numero di condizionamento, possiamo procedere analiticamente applicando la formula:

$$K(f, x) = \frac{|x \cdot f'(\xi)|}{|y|} = \frac{\left|x \cdot \frac{-2x}{2\sqrt{1-x^2}}\right|}{|\sqrt{1 - x^2}|} = \frac{x^2}{1 - x^2}$$

Chiaramente il numero di condizionamento esplode quando $x \to 1$

3 Equazioni non lineari

3.1 Risoluzione di equazioni non lineari

La ricerca degli 0 di funzioni è un problema ricorrente in moltissime discipline scientifiche, in particolare nel machine e nel deep learning, infatti l'allenamento di una rete neurale (training) avviene tramite la risoluzione di un problema di minimo: viene cercato il minimo di una funzione che misura la differenza del valore predetto dalla rete neurale rispetto al valore corretto.

Definizione 14: Risoluzione di equazioni non lineari

L'operazione di risoluzione di equazioni non lineari data una funzione $f : \mathbb{R} \to \mathbb{R}$ con $(f : I \subset R \to R)$ consiste nel trovare $\alpha \in I$ tale che $f(\alpha) = 0$.

Chiameremo il valore $\alpha \mid f(\alpha) = 0$ uno **zero di f**

La ricerca degli 0 di una funzione è un problema strettamente collegato alla ricerca dei **punti di minimo** di una funzione. Ricordiamo infatti che data una funzione $f(x)$ i punti in cui la sua derivata $f'(x)$ si annulla sono i punti stazionari e potenzialmente punti di minimo locali o globali.

3.1.1 Categorizzazione degli zeri di una funzione

Data una funzione $f : \mathbb{R} \to \mathbb{R}$ esistono diverse **tipologie di zeri**

- α è uno **zero semplice** di f se $f'(\alpha) \neq 0$
- α è uno **zero multiplo** di f se $f'(\alpha) = 0$. In questo caso diciamo che lo zero ha **molteplicità** che corrisponde all'ordine della prima derivata che non si annulla in α.

Esempio

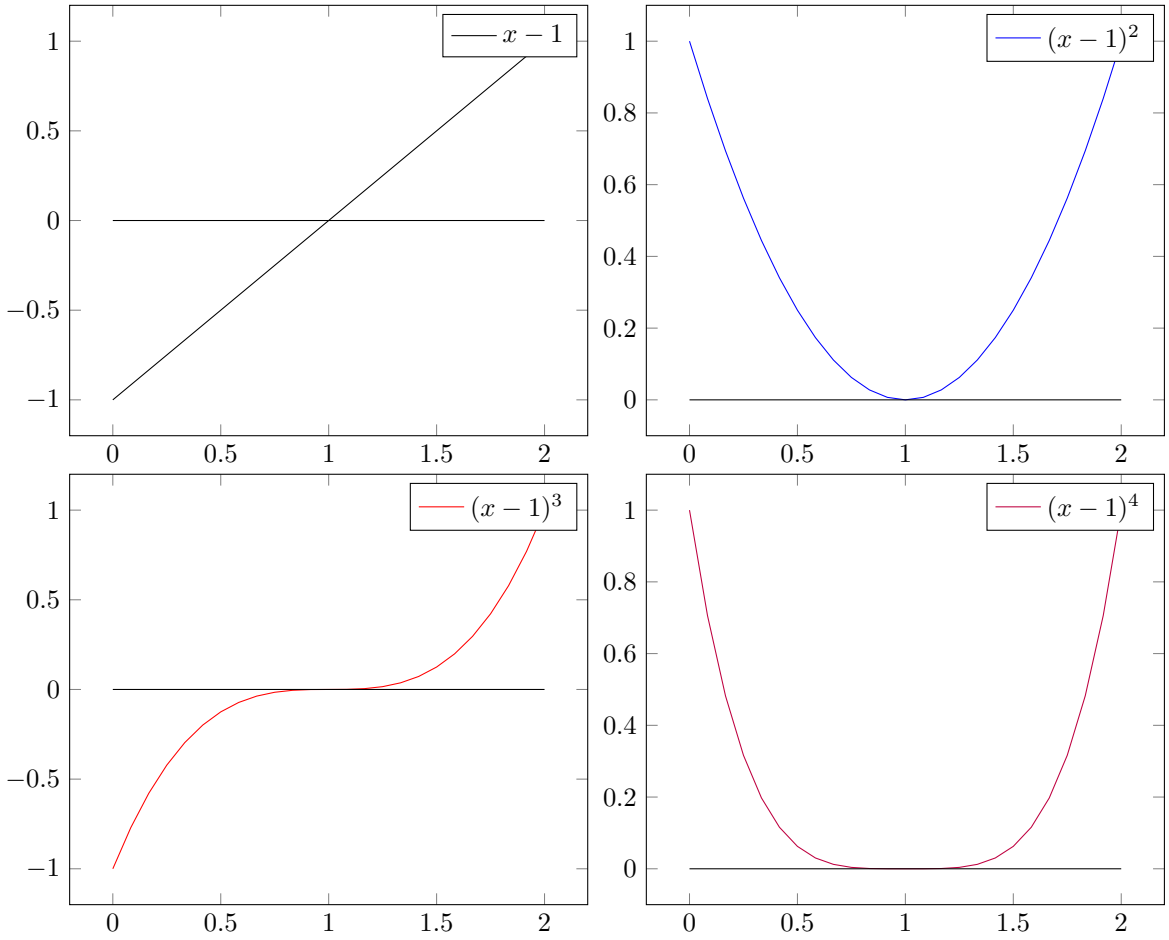
$$f(x) = \cos(x) - 1 + \frac{x^2}{2} + \frac{x^5}{5}$$

Immediatamente uno zero di $f(x)$ è $\alpha = 0$. A questo punto passiamo ad analizzare la derivata d'ordine primo per vedere se lo zero è semplice o multiplo.

$$f'(x) = -\sin x + x + x^4 \qquad f'(\alpha) = 0$$

La prima derivata si annulla in α quindi lo zero è **multiplo**, per scoprire la sua molteplicità continuiamo a derivare

$$\begin{aligned} f^{(2)}(x) &= -\cos x + 1 + 4x^3 & f^{(2)}(\alpha) &= 0 \\ f^{(3)}(x) &= \sin(x) + 12x^2 & f^{(3)}(\alpha) &= 0 \\ f^{(4)}(x) &= \cos(x) + 24x & f^{(4)}(\alpha) &= 1 \end{aligned}$$



Questi sono 4 esempi di funzioni aventi rispettivamente zero semplice e di molteplicità 2, 3 e 4.

Nota bene che la molteplicità degli zeri non ha nulla a che vedere con il numero di zeri presenti in un determinato intervallo, ad esempio

$$f(x) = x^2 - 2 \qquad \alpha_{1,2} = \pm\sqrt{2} \qquad f'(x) = 2x \qquad f'(\alpha_1) \neq 0 \quad \text{e} \quad f'(\alpha_2) \neq 0$$

Qui abbiamo 2 zeri in $\alpha_{1,2} = \pm\sqrt{2}$ ma sono entrambi zeri semplici.

3.1.2 Esistenza ed unicità degli zeri

Ci chiediamo data una funzione $f(x)$ come possiamo determinare se questa ha o meno degli zeri. Dobbiamo determinare se c'è un intervallo nel dominio della funzione che rispetta le richieste del teorema di Bolzano.

Teorema 1: Teorema di Bolzano

Sia f continua in un intervallo chiuso limitato $[a, b]$ e tale che $f(a) \cdot f(b) < 0$ allora $\exists \alpha \in [a, b] \mid f(\alpha) = 0$.

Questo teorema è il nostro strumento per verificare l'**esistenza degli zeri**.

Da solo il teorema di Bolzano non ci garantisce che lo zero sia unico, per fare ciò dobbiamo anche richiedere che la funzione sia **monotona** nell'intervallo $[a, b]$. Quindi noi siamo interessati ad intervalli del dominio dove la funzione è monotona e rispetta il teorema di Bolzano.

Esempio Supponiamo di avere

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

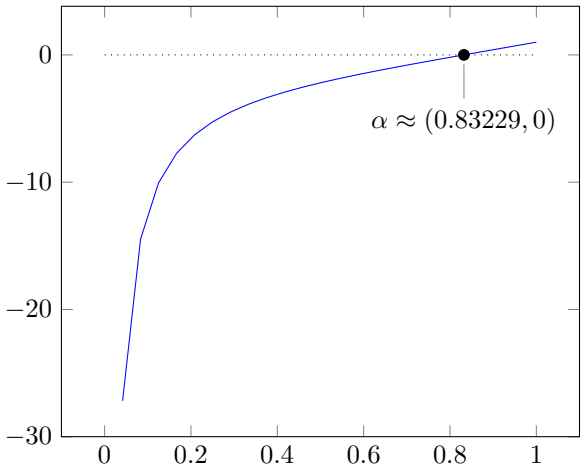
Esiste un α in $[0, 1] = I \mid f(\alpha) = 0$. Vale che:

$$f(0.1) = -12.28 < 0 \qquad f(1) = 1 > 0$$

Quindi esiste $\alpha[0, 1, 1] \mid f(\alpha) = 0$. Studiamo il segno della funzione nell'intervallo che ci interessa:

$$f'(x) = 4x + \frac{1}{x} + \frac{1}{x^2} > 0 \quad \forall x \in [0.1, 1]$$

Dato che la derivata è > 0 su I ciò implica che f è **monotona crescente** nell'intervallo I e quindi che lo zero è unico.



3.2 Metodi iterativi

Definizione 15: Metodi iterativi

Un metodo iterativo è una procedura che a partire da un valore iniziale x_0 genera una **successione di valori** $\{x_k\}_{k\geq 0}$ dove ogni termine x_k viene ottenuto a partire da uno o più termini precedenti.

Strettamente collegata a questa definizione vi è quella di convergenza.

Definizione 16: Convergenza

Un metodo iterativo (o la successione che genera) è **convergente** ad α se

$$\lim_{k\to\infty} x_k = \alpha \equiv \lim_{k\to\infty} \underbrace{|\alpha - x_k|}_{\text{errore } e_k} = 0$$

Chiaramente nella pratica non mettiamo $k \to \infty$, ci fermiamo prima

Definizione 17: Ordine di convergenza

Un metodo convergente (tale che $x_k \to \alpha$ dove $f(\alpha) = 0$) ha **ordine di convergenza** pari a **p** se esiste $c > 0$ tale che

$$\lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|^p} = c \quad \text{cioè} \quad \lim_{k\to\infty} \frac{|\alpha - x_{k+1}|}{|\alpha - x_k|^p} = c$$

Dove $e_k = \alpha - x_k$ è l'errore al passo k. L'ordine di convergenza descrive la **velocità di un metodo iterativo**.

La successioni di valori x_k, se tutto va bene, genererà un errore sempre più piccolo. Il metodo è convergente di ordine p se questo limite tra i rapporti degli errori consecutivi esiste ed è una **costante**.

- Se $p = 1$ il metodo è detto **a convergenza lineare**
- Se $p = 2$ il metodo è detto **a convergenza quadratica**
- Se $p > 1$ e $\lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|^p} = 0$ il metodo è detto a **a convergenza superlineare**

Chiamiamo **profilo di convergenza** il grafico che misura la dimensione dell'errore in base al numero d'iterazioni svolte.

$$p = 2 \qquad \lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|^p} = c \equiv |e_{k+1}| \approx c \cdot |e_k|^p$$

Cioè l'errore all'iterazione successiva è il quadrato di quella precedente.

$$p = 1 \qquad \lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|^p} = c \equiv |e_{k+1}| \approx c \cdot |e_k|^p$$

Notiamo che per $p = 1$ perché l'errore diminuisca c deve essere per forza minore di 1, cosa non necessaria per $p = 2$. Chiamiamo c **fattore di convergenza**.

Noi studieremo questi metodi di convergenza:

- Bisezione: ha $p = 1$ e lo studiamo perché è un metodo che richiede alla funzione solamente di essere continua. Non è generalizzabile per funzioni a più variabili.
- Newton: ha $p = 2$ e nelle ipotesi normali di convergenza è il metodo il migliore. Richiede che la funzione sia derivabile. (In più variabili calcolare derivate diventa costoso e a volte impossibile).
- Punto fisso: ha $p = 1$ e $x_{k+1} = g(x_k)$ cioè applica una certa funzione al valore precedente della successione per ottenere quello successivo.

3.2.1 Metodo di bisezione

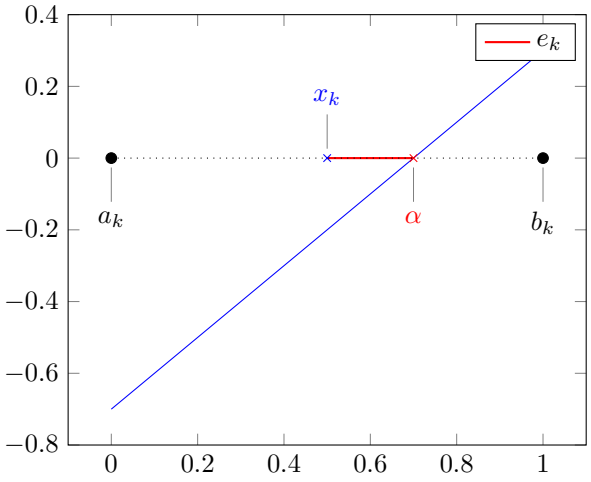
Definizione 18: Metodo di bisezione o dicotomico

E' il metodo più semplice dato che richiede solo che la funzione sia **continua** nell'intervallo $I = [a, b]$ e chiaramente che abbia radice $\alpha \in [a, b]$. Nota che continua non implica derivabile.

Data f continua in I e tale che dentro I via sia una sua radice α la bisezione costruisce una **successione di intervalli chiusi** $I_k = [a_k, b_k]$ gli uni dentro gli altri, ovvero vale che $I_k \subset I_{k-1}$, dove per ogni passo l' **ampiezza viene dimezzata** ma la radice $\alpha \in I_k \ \forall k$, ovvero tali che $f(a_k) \cdot f(b_k) < 0 \ \forall k$

Posto l'intervallo iniziale $[a, b]$ tale che $f(a) \cdot f(b) < 0$ e $a_0 = a, b_0 = b$ per ogni iterazione $k = 0, 1 \dots$ si calcola il **punto medio** $x_k = \frac{a_k + b_k}{2}$ e il valore della funzione nel punto medio $f(x_k)$. Per determinare in quale dei due intervalli è contenuta la radice dobbiamo effettuare dei controlli:

- Se $f(x_k) = 0$ allora abbiamo terminato, il punto medio è proprio la radice
- Se $f(a_k) \cdot f(x_k) < 0$ allora $b_k = x_k$, la radice sta a sinistra del punto medio
- Se $f(x_k) \cdot f(b_k) < 0$ allora $a_k = x_k$, la radice sta a destra del punto medio



Ad ogni iterazione la dimensione dell'intervallo dimezza, in realtà non si pone come condizione di fine $f(x_k) = 0$ ma bensì un valore "abbastanza" vicino allo zero. Esistono altre quantità che mi danno informazioni sull'errore, ad esempio la **semi-lunghezza dell'intervallo** che al passo k-esimo ci permette di dire che $|e_k| \leq \frac{1}{2}(b_k - a_k)$ e questo permette di provare che l'errore converge.

Corollario 1

La bisezione è un metodo a **convergenza globale**

Dimostrazione 5

Se abbiamo che $|e_k| \leq \frac{1}{2}(b_k - a_k)$ allora:

$$\frac{1}{2}(b_k - a_k) \to \frac{1}{2}\frac{1}{2}(b_{k-1} - a_{k-1}) \to \frac{1}{2}\frac{(b-a)}{2^k}$$

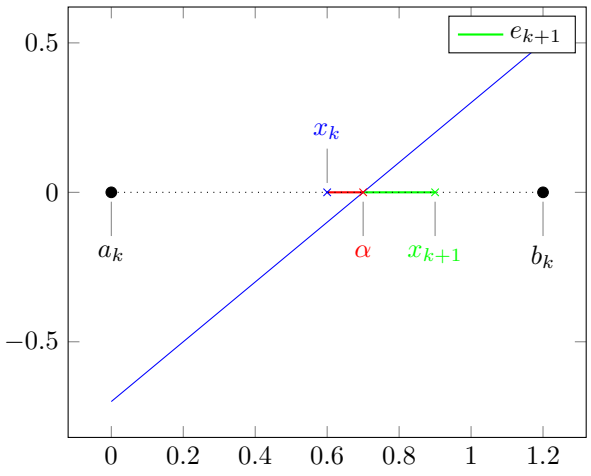
Cioè la dimensione della semi-lunghezza considerata al passo k-esimo è sempre uguale alla metà della precedente semi-lunghezza.

$$0 \leq |e_k| \leq \frac{b-a}{2^{k+1}} \quad \text{svolgendo il limite ho che} \quad \lim_{k\to\infty} \frac{b-a}{2^{k+1}} = 0$$

Ma allora per il **teorema dei carabinieri**:

$$\lim_{k\to\infty} |e_k| = 0$$

Perché l'ordine di convergenza di un limite sia $p = 1$ allora $\lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|} = c$. Non è detto che ad un passo successivo l'iterazione successiva diminuisca l'errore. Per rendere l'idea pensa se alla k-esima iterazione trovi un punto molto vicino ad α ma comunque diverso da 0 (o in realtà diverso dal valore che riteniamo accettabile), all'iterazione k+1 il valore di x_{k+1} sarà per forza più distante di x_k generando un errore più grande rispetto al precedente passo. Vale però che in media il fattore di convergenza $c = \frac{1}{2}$, tuttavia il limite formalmente non esiste.



Definizione 19: Criterio d'arresto e test sul residuo

L'idea è di arrestare il criterio quando la semi-lunghezza è più piccola di una determinata **tolleranza** che impostiamo noi e che chiamiamo tol (dipende da quanto preciso lo vogliamo, ad esempio se voglio la precisione di macchina sceglierò $tol = 10^{-15}$).

$|e_k| < tol \quad \varepsilon = 10^{-t} \quad \text{semilunghezza } = \frac{1}{2}(b_k - a_k) < tol \quad f(x_k) < tol$ (test sul residuo)

Il **test sul residuo** non è utilizzato perché non è affidabile, intuitivamente se la funzione è **molto ripida** vicino ad α potrei avere un residuo molto alto nonostante la vicinanza alla radice.

Esempio $f(x) = x \cdot 10^{50}$, per $x = 10^{-30}$ ho un residuo di molto grande di 10^{20} nonostante io sia abbastanza vicino a $\alpha = 0$

Un discorso simile vale anche per funzioni **molto piatte**.

Esempio $f(x) = x \cdot 10^{-50}$, per $x = 10^{30}$ ho un residuo piccolissimo di 10^{-20} ma sono decisamente lontano dallo zero $\alpha = 0$

Definizione 20: Test sul residuo pesato

Viene pesato il residuo in base a quanto velocemente cambia la funzione (quindi in base alla derivata). Una buona approssimazione dell'errore è:

$$\frac{f(x_k)}{f'(x_k)}$$

E' un approssimazione al primo ordine dell'errore. Abbiamo definito l'errore $e_k = \alpha - x_k$ segue che la radice vale $\alpha = x_k + e_k$.

$$0 = f(\alpha) = f(x_k + e_k) = f(x_k) + f'(x_k)e_k + \frac{1}{2}f''(x_k)e_k^2 + O(e_k^2)$$

Ricordando che lo sviluppo di Taylor è:

$$0 = f(x_k) + f'(x_k)e_k + O(e_k^2)$$

Dato che è un'approssimazione al primo ordine ci fermiamo alla derivata prima. Assumendo che $f'(k) \neq 0$ otteniamo che

$$\frac{-f(x_k)}{f'(x_k)} = e_k + O(e_k^2)$$

(Importante). Ci ricordiamo però che per il metodo della bisezione abbiamo richiesto una funzione **continua** e non necessariamente derivabile! Per ovviare a questo problema approssimiamo la derivata con il rapporto incrementale:

$$f'(x_k) = \frac{f(b_k) - f(a_k)}{b_k - a_k}$$

Vale quindi che:

$$e_k + O(e_k^2) = -f(x_k) \cdot \frac{b_k - a_k}{f(b_k) - f(a_k)}$$

E' possibile determinare il numero d'iterazioni necessarie al metodo di bisezione per raggiungere una certa accuratezza cioè possiamo trovare qual è il minimo numero d'iterazioni del metodo per avere un errore più piccolo di una certa tolleranza.

Supponiamo di volere un errore più piccolo di 10^{-t} $|e_k| < \varepsilon = 10^{-t}$, vale che

$$|e_k| \leq \frac{b-a}{2^{k+1}} < \varepsilon$$

Portando i termini diseguaglianza sui logaritmi ed isolando k ottengo:

$$\begin{aligned}
&\log(b-a) - (k+1)\log 2 < \log \varepsilon \ ; \\
&\frac{\log(b-a) - \log(\varepsilon)}{\log 2} < k+1 \ ; \\
&k > \left\lceil \frac{\log(\frac{b-a}{\varepsilon})}{\log 2} \right\rceil - 1
\end{aligned}$$

Per diminuire l'errore di un ordine di grandezza il metodo di bisezione deve svolgere tra le 3 e le 4 iterazioni. (Per rendere l'idea $100 \to 50 \to 25 \to 12.5 \to 6.75$).

Concludiamo la panoramica sul metodo di bisezione elencandone i pregi ed i difetti:

- Converge sempre, quindi la convergenza locale è assicurata
- Richiede solamente che la funzione sia continua, un vincolo "abbastanza" leggero sopratutto se paragonato con il vincolo di derivabilità
- E' lento, infatti ha un ordine di convergenza $p = 1$ ed un fattore di convergenza $c = M = 0.5$, (la costante asintotica si può chiamare M o c)
- Non è generalizzabile a funzioni in più variabili, cioè non si può usare per funzioni tipo $f(X) : \mathbb{R}^n \to \mathbb{R}^n$

3.3 Metodo di Newton

Parentesi storica Isaac Newton lo ha usato per risolvere questa equazione nel 1660:

$$x^3 - 2x - 5 = 0$$

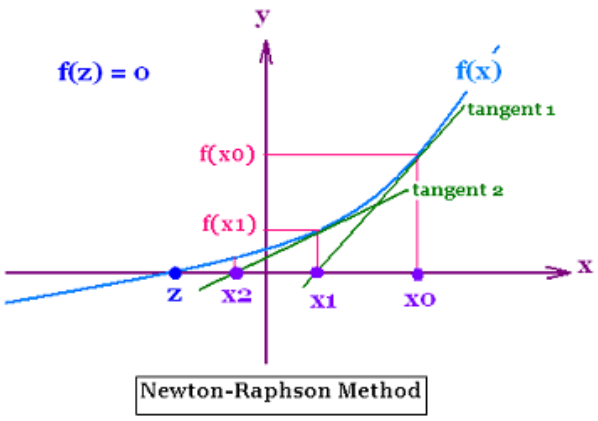
Successivamente Joseph Raphson ha pubblicato l'algoritmo, quindi questo metodo prende anche il nome di metodo di Newton Raphson.

Questo metodo ha tanti nomi:

- Metodo di Newton
- Metodo di Newton - Raphson
- Metodo who was - Raphson
- Metodo delle tangenti/ Metodo delle tangenti variabile

Diamo una panoramica generale del metodo di Newton:

- Applicabilità: richiede che la funzione f sia **derivabile** in un intorno dello 0 (cioè di α).
- Funzionamento: dato x_0 si calcola la retta tangente alla curva $y = f(x)$ nel punto $(x_0, f(x_0))$ e si interseca la tangente con l'asse x, l'ascissa del punto d'intersezione è x_1 e si ripete la procedura.
- Geometricamente:



Esempio, dove chiaramente la z è la nostra α

In soldoni il metodo di Newton riceve in input un punto $(x_k, f(x_k))$, ne calcola la retta tangente e la interseca con l'asse delle ascisse. La successiva iterazione del metodo prenderà il punto ottenuto dall'intersezione ed il valore della funzione in quel punto $(x_{k+1}, f(x_{k+1}))$ e rifarà i passaggi del calcolo della tangente e dell'intersezione.

Ricavare il punto d'intersezione tra la retta tangente nel punto $(x_k, f(x_k))$ e l'asse delle ascisse consiste nel risolvere un sistema

$$\begin{cases} y = f(x) + f'(x)(x - x_k) \\ y = 0 \end{cases} \quad \begin{cases} 0 = f(x_K) + f'(x_k)(x - x_k) \\ y = 0 \end{cases} \quad \begin{cases} x = x_k - \frac{f(x_k)}{f'(x_k)} \\ y = 0 \end{cases}$$

Questo metodo risulta ben posto se e solo se $f'(x_k) \neq 0$, cioè se la funzione ha uno zero semplice.

Lo **schema iterativo** è che:

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

La retta tangente è un'approssimazione del primo ordine della funzione, cioè equivale a fare un'approssimazione con Taylor fermandosi al primo ordine. Vediamo infatti:

$$\text{Con Taylor:} \qquad x_{k+1} = x_k + s \qquad f(x_{k+1}) = f(x_k + s) = f(x_k) + f'(x_k) \cdot s = 0$$

$$\text{Ricavando } s \text{ ottengo} \qquad s = \frac{-f(x_k)}{f'(x_k)} \qquad x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

Definizione 21: Convergenza locale

Il metodo di Newton **non converge globalmente**, ma si dimostra che il metodo di Newton **converge localmente**.

Diciamo che un metodo iterativo **converge localmente** se esiste un "raggio" δ tale che

$$\forall \delta > 0 \mid \forall x_0 \in B(\alpha, \delta) \iff x_0 \in [\alpha - \delta, \alpha + \delta]$$

Allora si ha che $\{x_k\} \to \alpha$.

[Nota: In sostanza esiste un intervallo attorno alla radice dove se prendo un punto vicino a lì la successione convergerà?]

Teorema 2: Convergenza locale del metodo di Newton

Sia f una funzione di classe 2 $f \in C^2(a, b)$ (cioè derivabile 2 volte) con la radice $\alpha \in (a, b) \mid f(\alpha = 0)$. Indichiamo con I_α l'intervallo $[a, b]$

Supponiamo che $f'(x) \neq 0 \ \forall x \in I_\alpha$ (cioè che la funzione abbia uno zero semplice). In queste condizioni esiste $\delta > 0$ tale che se prendiamo $x \in B(\alpha, \delta)$ allora la successione **generata a partire dal punto** x_0 $\{x_k\}_{k \geq 0}$ è ben definita e $\{x_k\} \subset B(\alpha, \delta)$, inoltre il rapporto $\frac{e_{k+1}}{e_k^2} \leq M$ e quindi $\{x_k\}_{k \geq 0} \to \alpha$ (converge ad α).

Inoltre vale anche che

$$\lim_{k \to \infty} \frac{|e_{k+1}|}{|e_k|^2} = \frac{1}{2} \cdot \frac{|f''(\alpha)|}{|f'(\alpha)|} \Rightarrow f''(\alpha) \neq 0 \to p = 2$$

Se invece $f''(\alpha)$ è uguale a 0 allora l'ordine di convergenza è almeno 3

$$\lim_{k \to \infty} \frac{e_{k+1}}{e_k} = 0 \quad \text{implica} \quad p > 2$$

Dimostrazione 6

Sappiamo che $e_k = \alpha - x_k$ e che allora $\alpha = x_k + e_k$. Approssimando la funzione con Taylor ottengo che

$$0 = f(x_k + e_k) = f(x_k) + f'(x_k)e_k + \frac{1}{2}f''(\xi_k)e_k^2$$

Dove $\xi_k \in (x_k, \alpha)$.

Dividendo per $f'(x_k) \neq 0$ ottengo:

$$\frac{-f(x_k)}{f'(x_k)} = e_k + \frac{1}{2} \cdot \frac{f''(\xi_k)}{f'(x_k)}e_k^2$$

Ricordando che $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$ decidiamo di chiamare **scarto** $s_{k+1} = x_{k+1} - x_k$

Segue che:

$$x_{k+1} - x_k = e_k + \frac{1}{2} \cdot \frac{f''(\xi_k)}{f'(x_k)}e_k^2 \quad x_{k+1} - \alpha + \alpha - x_k = e_k + \frac{1}{2} \cdot \frac{f''(\xi_k)}{f'(x_k)}e_k^2$$

Ricordando che $e_k = \alpha - x_k$

$$-e_{k+1} + e_k = e_k + \frac{1}{2} \cdot \frac{f''(\xi_k)}{f'(x_k)}e_k^2 \quad |e_{k+1}| = \frac{1}{2}\frac{|f''(\xi_k)|}{|f'(x_k)|}e_k^2 \quad |e_{k+1}| \leq \underbrace{\frac{1}{2}\frac{max|f''(x)|}{min|f'(x)|}}_{M} e_k^2$$

Dove max e min sono presi nell'intervallo $x \in [a, b]$.

Dimostriamo adesso che il rapporto è più piccolo di M e che quindi $\{x_k\}_{k\geq 0} \to \alpha$

$$\frac{|e_{k+1}|}{e_k^2} \leq M \qquad |e_{k+1}| \leq Me_k^2 \qquad |e_k| \leq M(e_{k-1})^2$$

Capito questo possiamo estendere fino a e_0

$$M|e_k| \leq (M|e_{k-1}|)^2 \leq (M|e_{k-2}|)^{2^2} \leq \cdots \leq (M|e_0|)^{2^k}$$

Per dimostrare la convergenza applico il **teorema dei carabinieri**:

$$0 \leq M|e_k| \leq (M|e_0|)^{2^k} \qquad \lim_{k\to\infty} (M|e_0|)^{2^k} = 0 \quad \text{per } M|e_0| < 1$$

Dimostrato che tende a 0 vale che

$$M|\alpha - x_0| < 1 \iff |\alpha - x_0| < \frac{1}{M} \iff -\frac{1}{M} < \alpha - x_0 < \frac{1}{M}$$

che è vera se e solo se

$$x_0 \in [\alpha - \underbrace{\frac{1}{M}}_{-\delta}, \alpha + \underbrace{\frac{1}{M}}_{\delta}] \ \delta \text{ era il "raggio"}$$

Dimostrazione 7

Dalla dimostrazione precedente vale che

$$|e_{k+1}| = \frac{1}{2}\frac{f''(\xi_k)}{f'(x_k)}e_k^2 \quad \text{con } \xi_k \in (x_k, \alpha)$$

Segue:

$$\lim_{k\to\infty}\frac{|e_{k+1}|}{e_k^2} = \lim_{k\to\infty}\frac{1}{2}\cdot\frac{f''(\xi_k)}{f'(x_k)} = \underbrace{\frac{1}{2}\frac{|f''(\alpha)|}{|f'(\alpha)|}}_{\text{Cost asintotica c}}$$

Dato che $x_k \to \alpha$

Funzioni crescenti convesse, concave decrescenti hanno convergenza monotona.

Definizione 22: Criterio d'arresto per il metodo di Newton

Consideriamo il rapporto tra l'errore e lo scarto, l'errore al passo k è $e_k = \alpha - x_k$ mentre lo scarto al passo k è $s_k = x_k - x_{k-1}$. Dato che per il metodo di Newton abbiamo richiesto la derivabilità della funzione ho che:

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

Ricordando che $0 = f(\alpha) = f(x_k + e_k) = f(x_k) + f'(x_k)e_k + O(e_k^2)$.

$$\frac{-f(x_k)}{f'(x_k)} = e_k + O(e_k^2) \iff |s_{k+1}| \approx |e_k| + O(e_k)^2$$

L'errore al passo k ha **lo stesso ordine di grandezza** dello scarto all'iterata successiva. Quindi noi arresteremo il metodo di Newton quando lo scarto sarà minore di un determinato valore che scegliamo.

Esempio Se lo scarto è 10^{-8} alla 3° iterazione, allora già alla 2° iterazione l'errore era di circa 10^{-8}.

Supponiamo $s_3 = 10^{-8}$ allora $e_2 = 10^{-8}$ allora per $k = 3$ quanto vale e_3 ? L'errore al passo $k = 3$ è circa $|e_{k+1}| \approx c \cdot |e_k|^2 \approx c \cdot s_{k+1}^2$. Segue che non ha senso scegliere scarti troppo grandi dato che l'errore viene elevato al quadrato rispetto al precedente, per rendere l'idea già con uno scarto di 10^{-8} l'errore sarà 10^{-16} e quindi siamo già in precisione macchina e non ha senso andare avanti.

- $f'(\alpha) \neq 0$ zero semplice
- Se $f'(\alpha) \neq 0$ e $f''(\alpha) \neq 0$ allora p = 2
- Se $f''(\alpha) = 0$ allora vale che $p \geq 3$
- $f'(\alpha) = 0$ zero multiplo e la convergenza del metodo di newton è lineare

Metodo di Newton per radici multiple, quando $p = 1$ la costante asintotica è $c = 1 - \frac{1}{r}$ dove r è la molteplicità della radice. Cioè abbiamo che:

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

Definizione 23: Metodo di Newton modificato

Per ripristinare la convergenza quadratica del metodo di Newton per radici multiple basta porre basta porre:

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

Dove $k \geq 0$

Il metodo di Newton risulta essere particolarmente pesante ed a volte non applicabile per funzioni a più variabili, ciò deriva dal fatto che ad ogni iterazione dobbiamo calcolare delle derivate parziali, un'operazione che diventa molto **costosa**. Esistono tuttavia delle varianti del metodo di Newton che aggirano questo problema, pagando però un "prezzo" sull'ordine di convergenza che non sarà più quadratico.

3.3.1 Metodo della tangente fissa

L'idea di questo metodo è di tenere **fissa** la retta tangente con cui approssimiamo la funzione localmente in ogni punto. In pratica ci calcoliamo solo la tangente del punto di partenza e questa verrà usata da tutti i calcoli successivi. Il vantaggio principale è che evitiamo di calcolare tante derivate, tuttavia l'ordine di convergenza è $p = 1$ cioè **lineare**. Il fatto che sia l'ordine di convergenza sia lineare non implica un comportamento "lento" perché ricordiamo che la velocità in questo caso dipende fortemente dal fattore di convergenza c o M (che però è forzatamente più piccolo di 1).

Il metodo non di discosta troppo da quello Newtoniano quando trattiamo funzioni dove la derivata non varia molto nell'intervallo che analizziamo. Spesso questo metodo è utilizzato in modo alterno con altri.

Possiamo definire questo schema iterativo

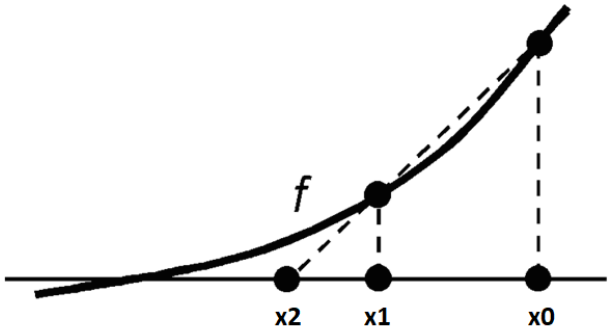
$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_0)} \quad k \geq 0$$

E queste sono le caratteristiche principali del metodo:

$$p = 1 \qquad c = \left| 1 - \frac{f'(\alpha)}{f'(x_0)} \right|$$

3.3.2 Metodo delle secanti

Detto anche **metodo della secante variabile**. L'idea è di sfruttare due punti x_1 e x_0, si calcola la secante ed il punto dove la secante interseca l'asse delle ascisse sarà x_2 che sfrutterò per calcolare la nuova secante e così via.



Esempio per rendere l'idea geometricamente

La pendenza della secante è il rapporto incrementale quindi lo schema iterativo del metodo è:

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

O se vogliamo scrivere una formula più leggibile:

$$\begin{cases} x_{k+1} = x_k - \frac{f(x_k)}{a_k} & k \geq 1 \\ a_k = \frac{f(x_k)-f(x_{k-1})}{x_k - x_{k-1}} \end{cases}$$

Questo metodo ha un ordine di convergenza tra i più veloci dopo Newton, tantè che in più variabili si chiama **quasi-newton**. Vale che

- $p = 1.618 = \frac{1+\sqrt{5}}{2}$ quindi è un metodo con ordine di convergenza **superlineare**
- La costante asintotica o fattore di convergenza è

$$c = \left| \frac{\frac{1}{2}f''(\alpha)}{f'(\alpha)} \right|^{0.618}$$

3.3.3 Metodo del punto fisso

Detto anche **metodo d'iterazione funzionale** è un metodo molto usato composto da una famiglia di metodi iterativi che calcolano zeri di funzioni basati sulla riscrittura dell'equazione originale $f(x) = 0$ nella forma

$$x = g(x)$$

Il metodo del punto fisso ha ordine di convergenza $p = 1$ (quando converge). Per trovare non un α tale che f(x) = 0 ma bensì un α tale che $\alpha = g(\alpha)$. Dove ricordiamo che per **punto fisso** intendiamo un punto che la funzione mappa in sé stesso. La riscrittura della funzione può essere fatta in molti modi.

Esempio Prendiamo questa funzione:

$$f(x) = e^{-x} - \cos(x) + 3x^2 = 0$$

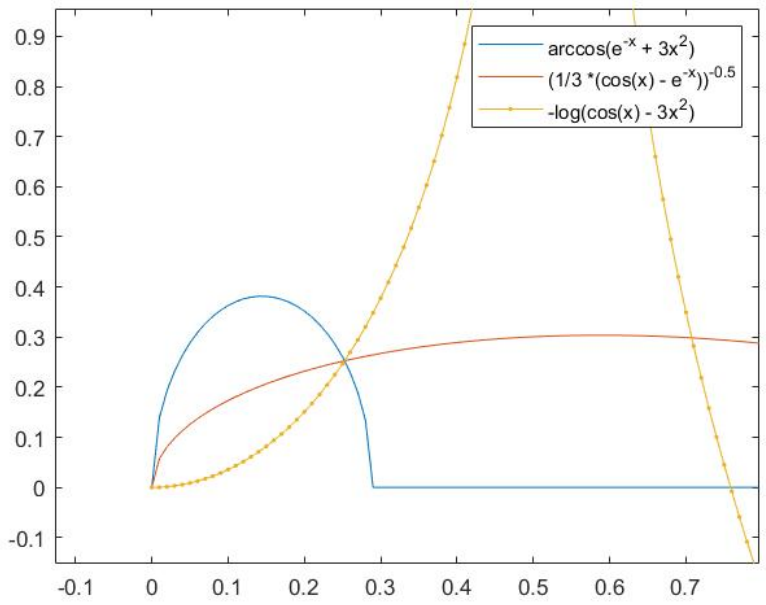
Posso riscriverla in "forma punto fisso" in tanti modi:

$$e^{-x} = \cos(x) - 3x^2 \quad -x = \log(\cos(x) - 3x^2) \quad \underbrace{x = -\log(\cos(x) - 3x^2)}_{g_1(x)}$$

$$3x^2 = \cos(x) - e^{-x} \quad \underbrace{x = \sqrt{\frac{1}{3}(\cos(x) - e^{-x})}}_{g_2(x)}$$

Vale che $\alpha_1 = 0$ e $\alpha_2 = 0.25266781$. Un'altra forma è:

$$\underbrace{x = \arccos(e^{-x} + 3x^2)}_{g_3(x)}$$



E chiaramente c'è la banale riscrittura $x + f(x) = x$. I metodi iterativi del punto fisso si basano sulla riscrittura della formula, non tutte convergeranno. Vediamo lo schema iterativo:

$$x = g(x) \quad \begin{cases} x_0 \in \mathbb{R} \\ x_{k+1} = g(x_k) \end{cases} \tag{1}$$

Lemma 1

Se $g(x)$ è **continua** lo schema iterativo (1) converge ad α e allora α è **punto fisso di** $g(x)$. Chiamiamo $g(x)$ **funzione d'iterazione**.

La successione generata da questo metodo iterativo è convergente significa che

$$\alpha = \lim_{k\to\infty} x_{k+1} = \lim_{k\to\infty} g(x_k) = g\Big(\underbrace{\lim_{k\to\infty} x_k}_{\alpha}\Big) = g(\alpha)$$

Esempio 7

$$\begin{cases} x_0 = 1 \\ x_{k+1} = \cos(x_k) \end{cases} \qquad \begin{cases} x = \cos(x) \\ f(x) = x - \cos(x) = 0 \end{cases} \qquad \begin{cases} x_1 = \cos(1) = 0.540302 \\ x_{20} = \cos(19) = 0.73918... \end{cases}$$

Quando arrestiamo lo schema iterativo? Per newton scarto e residuo pesato erano la stessa cosa. Per questi metodi utilizzeremo esclusivamente lo scarto dato che non lavoriamo con derivate.

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

Esempio

$$f(x) = e^{-2x}(x-1) = 0 \quad x = \underbrace{e^{-2x}(x-1) + x}_{g(x)} \quad \alpha = 1$$

Vediamo che un metodo del punto fisso per una $g(x)$ la quale radice α è banale non converge partendo da $x_0 = 0.99$

$$\begin{cases} x_0 = 0.99 \\ x_1 = e^{-2x_0}(x_0 - 1) + x_0 = 0.9886 \\ x_2 = g(x_1) = 0.9870 \\ \vdots \\ x_{27} = g(x_{26}) = 0.1665 \\ \vdots \end{cases}$$

Teorema 3: Teorema di convergenza globale

Sia $g(x)$ continua su $[a,b]$ e derivabile con continuità (cioè che la derivata sia continua) in (a,b) e tale che $g(x) \in [a,b]$ $\forall x \in [a,b]$, cioè che $g([a,b]) \subset [a,b]$ ovvero che g è una **contrazione** di $[a,b]$.

Allora esiste almeno un punto fisso α di g, se inoltre $\exists m < 1$ tale che

$$|g'(x)| \leq m < 1 \quad \forall (x) \in [a,b]$$

Allora il punto fisso è unico e la successione $\{x_k\}_{k\geq 0}$ generata dallo schema iterativo 1 è convergente ad α per tutti gli $x_0 \in [a,b]$

Riassumendo questo teorema dice che:

- il punto fisso esiste quando g è una contrazione (1)
- se in un intorno di α cioè $[a,b]$ riesci a maggiorare la derivata prima con m e questa $m < 1$ allora possiamo provare unicità (2) e convergenza (3)

Dimostrazione 8

Proviamo l'esistenza (1), la nostra ipotesi è che $g([a,b]) \subseteq [a,b]$. Definiamo $x = g(x)$ e $\phi(x) = g(x) - x$, ora applichiamo il teorema degli zeri su $\phi(x)$.

$$\phi(a) = g(a) - a \text{ siamo certi data la contrazione che } g(a) \geq a$$

Segue che

$$\phi(a) = g(a) - a > 0 \quad \phi(b) = g(b) - b < 0$$

Esiste α tale che $\phi(\alpha) = 0$ quindi esiste α tale che $g(\alpha) = \alpha$

Dimostrazione 9

Dimostriamo l'unicità (2), vale che se $|g'(x)| \leq m < 1 \forall x \in [a,b]$ è vero allora devo provare che il punto fisso è unico. Per assurdo supponiamo esistano due punti fissi α_1 e α_2 con $\alpha_1 \neq \alpha_2$. Proviamo a calcolare quanto distano.

$$|\alpha_1 - \alpha_2| = |g(\alpha_1) - g(\alpha_2)| = g'|(\xi)||\alpha_1 - \alpha_2| \text{ per il teorema del valore medio}$$

Segue che $\xi \in (\alpha_1, \alpha_2)$ e quindi $\xi \in [a,b]$. Vale che $|g'(\xi)| < 1$.

Ma allora $|\alpha_1 - \alpha_2| = g'|(\xi)||\alpha_1 - \alpha_2|$ impossibile dato che $g'|(\xi)| < 1$.

Dimostrazione 10

Proviamo adesso la convergenza (3), partiamo dal fatto che

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

Applichiamo di nuovo il teorema del valore medio:

$$|g(x_k) - g(\alpha)| = |g'(\xi)| \cdot |x_k - \alpha|$$

Dato che $\xi \in (x_k, \alpha)$ allora $\xi \in [a,b]$ quindi $|g'(\xi)| \leq m < 1$.

$$|g'(\xi)| \cdot |x_k - \alpha| \leq m \cdot |e_k| \leq m^2 |e_{k-1}| \leq \cdots \leq m^{k+1} |e_0|$$

A questo punto applico il teorema dei carabinieri:

$$0 \leq |e_{k+1}| \leq m^{k+1} \cdot |e_0| \quad \lim_{k \to \infty} m^{k+1} \cdot |e_0| = 0 \ \forall e_0 \ \forall x_0$$

Quindi vale che $\lim_{k \to \infty} |\{e_{k+1}\}| = 0$

Teorema 4: Teorema di convergenza locale

Sia α un punto fisso di g funzione continua e derivabile in $[a, b]$ (derivabile in (a,b)) con $\alpha \in [a, b]$. Se **esattamente** in α vale che:

$$|g'(\alpha)| < 1 \rightarrow \exists \delta > 0 \text{ tale che } \forall x_0 \in [\alpha - \delta, \alpha + \delta]$$

Il metodo 1 converge ad α e si ha che

$$\lim_{k \to \infty} \frac{e_{k+1}}{e_k} = g'(\alpha)$$

Cioè se in α la derivata è minore di 1 allora sicuramente c'è convergenza partendo vicino ad α e la convergenza del metodo è lineare e converge ad una costante $g'(\alpha)$ che è la costante asintotica di questo metodo.

Dobbiamo dimostrare due cose:

1. Se α è un punto fisso di una funzione g continua e derivabile in $[a, b] = I$ con $\alpha \in I$. Se $|g'(\alpha)| < 1$ allora esiste un raggio $\delta > 0$ tale che $\forall x_0 \in [\alpha - \delta, \alpha + \delta]$ il metodo del punto fisso converge ad α.
2. $$\lim_{k \to \infty} = \frac{e_{k+1}}{e_k} = g'(\alpha)$$

Dimostrazione 11: 1.

Vogliamo riportarci alle condizioni del teorema di convergenza globale, dobbiamo quindi dimostrare che la derivata sia minore di 1 e che la funzione sia una contrazione. Per la definizione di funzione continua sappiamo che se in α $g'(x) < 1$ allora sicuramente esiste un intervallo $(\alpha - \delta, \alpha + \delta) = I$ tale per cui $|g'(x)| \leq m < 1$. Dobbiamo dimostrare che $g(x)$ è una contrazione nell'intervallo $\forall x \in I_\alpha$, quindi che $|g(x) - g(\alpha)| < \delta$

$$|x - \alpha| < \delta \quad \text{per la definizione di punto fisso} \quad |g(x) - \alpha| = |g(x) - g(\alpha)|$$

Applico il teorema del valore medio e trovo che:

$$|g'(\xi)| \cdot |x - \alpha| = |g(x) - g(\alpha)|$$

$\xi \in (x, \alpha)$ ma $g'(\xi) < 1$ e $|x - \alpha| < \delta$ quindi $|g(x) - g(\alpha)| < \delta$.

Siccome esiste I dove $|g'(x)| \leq m < 1$ per continuità e abbiamo dimostrato che $g(I) \subset I$ per il teorema precedente il metodo converge ad α $\forall x \in I$.

Dimostrazione 12: 2.

Per definizione so che $e_{k+1} = \alpha - x_{k+1} = g(\alpha) - g(x_k)$ ($x_{k+1} = g(x_k)$ per lo schema iterativo del metodo del punto fisso). Applico di nuovo il teorema del valore medio:

$$e_{k+1} = g(\alpha) - g(x_k) = g'(\xi_k) \cdot (\alpha - x_k) = g'(\xi_k) \cdot e_k \quad \text{passo ai limiti}$$

$$\lim_{k\to\infty} \frac{e_{k+1}}{e_k} = \lim_{k\to\infty} g'(\xi_k) = g'(\alpha)$$

Con $\xi_k \in (x_k, \alpha)$

Prendendo l'ultimo risultato in valore assoluto ottengo esattamente la definizione di **ordine di convergenza**

$$\lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|} = |g'(\alpha)|$$

Per il metodo del punto fisso, assunto che $g'(\alpha) \neq 0$, ho che:

- p = 1
- c = $|g'(\alpha)|$ con $0 < c < 1$

Per k sufficientemente grande vale che $e_{k+1} \approx g'(\alpha) \cdot e_k$. Se si ha che $g'(\alpha) = 0$ allora significa che l'ordine di convergenza che abbiamo supposto è sbagliato, in realtà è più grande di 1.

Teorema 5

Se $g(x)$ è continua e derivabile due volte con continuità su $[a, b]$ e $\alpha \in [a, b]$ punto fisso di $g(x)$ con $g'(\alpha) = 0$ e $g''(\alpha) \neq 0$ e si ha che:

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

Segue che la costante asintotica è $\left|\frac{g''(\alpha)}{2}\right|$ e $p = 2$

Dimostrazione 13

Sfruttiamo che $-e_{k+1} = x_{k+1} - \alpha = g(x_k) - g(\alpha)$. Qui non ha senso usare il teorema del valore medio perché non compare la derivata seconda, uso quindi Taylor centrato in α. Vale che $x_k = \alpha + (x_k - \alpha) = \alpha - e_k$

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

(Metto ξ_k perché stiamo facendo Taylor con il resto di Lagrange) Vale che:

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

E quindi ottengo che:

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

Passando ai limiti:

$$\lim_{k\to\infty} \frac{x_{k+1} - \alpha}{(x_k - \alpha)^2} = \lim_{k\to\infty} \frac{1}{2}g''(\xi_k) = \frac{1}{2}g''(\alpha)$$ perché sto ultimo passaggio?

In generale si ha che per $g(x) \in C^p([a,b])$ (continua e derivabile p volte su $[a,b]$) allora:

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

Allora l'ordine di convergenza del metodo del punto fisso è p e la costante asintotica è:

$$c = \left|\frac{g^{(p)}(\alpha)}{p!}\right| \qquad \text{dove p ovviamente è l'ordine di derivata}$$

Definizione 24: Test d'arresto per il metodo del punto fisso

La relazione che sussiste tra scarto ed errore nel punto fisso è la seguente:

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

$$(1 - g'(\xi_k))e_{k+1} = g'(\xi_k)s_{k+1} \qquad e_{k+1} = \frac{g'(\xi_k)}{1 - g'(\xi_k)}s_{k+1} \quad \xi_k \in (x_k, \alpha)$$

Possiamo approssimare a:

$$e_{k+1} = \frac{g'(\alpha)}{1 - g'(\alpha)}s_{k+1}$$

Per valori di $g'(\alpha) \approx 1$ ho un denominatore molto piccolo ed ottengo un errore sovrastimato.

Definizione 25: Metodo di Newton come metodo di punto fisso

Dato che $x_{k+1} = g(x_k)$ è il metodo del punto fisso, ma il metodo di newton si scrive così:

$$x_{k+1} = \underbrace{x_k - \frac{f(x_k)}{f'(x_{k+1})}}_{g_N(x_k)}$$

Dal metodo di Newton sappiamo che

$$c = \frac{1}{2}\frac{f''(\alpha)}{f'(\alpha)} \quad \text{p} = 2 \text{ se } f'(\alpha) \neq 0 \wedge f''(\alpha) \neq 0$$

Dimostrazione 14

Visto che $g_N(x) = x - \frac{f(x)}{f'(x)}$ segue che

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

Valutiamo ora in α e otteniamo:

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

Se $f'(\alpha) \neq 0$ allora $g'_N(\alpha) = 0$ e quindi $p > 1$.
Se $g''_N(\alpha) \neq 0$ allora il metodo converge con ordine 2 e la costante asintotica è:

$$c = \left| \frac{g''_N(\alpha)}{2} \right|$$

4 Approssimazione di funzioni

Le funzioni possono essere approssimate con un'altra funzione scritta come combinazione lineare all'interno di uno spazio vettoriale di funzioni. Generalmente vogliamo riscrivere le funzioni come polinomi perché i polinomi si possono valutare con le operazioni elementari (moltiplicazione ed addizione), inoltre è facile derivare ed integrare polinomi. Il concetto cardine è di sostituire una funzione perché l'originale è troppo complicata oppure perché l'originale non la conosciamo ma sappiamo che valori assume in certi punti del dominio. Noi applicheremo le tecniche d'approssimazione a funzioni da $\mathbb{R} \to \mathbb{R}$, ma i concetti sono applicabili anche in più dimensioni. Studieremo due metodi d'approssimazione:

- L'**approssimazione ai minimi quadrati** che come metrica usa $\| \ \|_2$ (norma 2)
- L'**interpolazione polinomiale** che come metrica usa $\| \ \|_\infty$ (norma massima)

4.1 Interpolazione polinomiale

Spazio polinomiale $\mathbb{P}_n(x)$ (o anche $\pi_n(x)$) è lo **spazio dei polinomi di grado fino ad** n

$$p_n(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \qquad p_n(x) \in \mathbb{P}_n$$

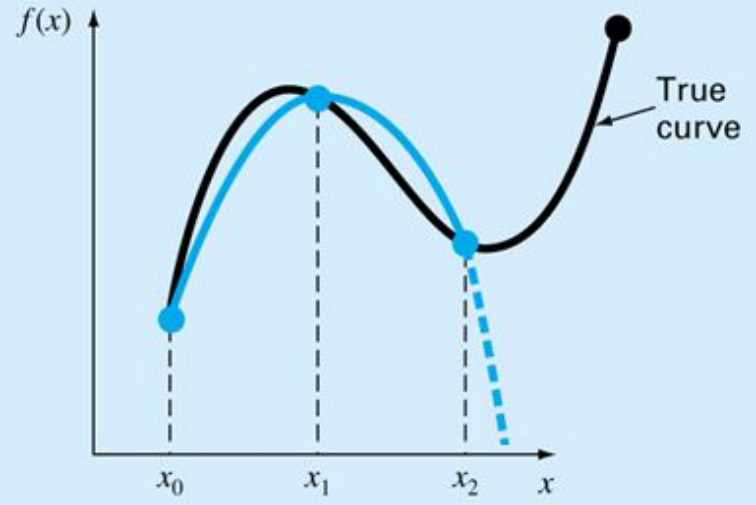
Teorema 6: Esistenza ed unicità del polinomio d'interpolazione

Siano dati $n+1$ coppie di numeri (x_i, y_i) con $i = 0, 1, \ldots, n$ dove tutte le ascisse sono diverse $x_i \neq x_j$, **esiste ed è unico** il polinomio $p_n(x)$ di grado n $p_n(x) \in \mathbb{P}_n(x)$ tale che:

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

Queste sono dette **condizioni di interpolazione** ed esprimono il valore che la funzione assume in certi punti x_i detti **nodi d'interpolazione**. Quindi la condizione che chiediamo alla funzione approssimata è che assuma gli stessi valori dei nodi x_i che noi vogliamo approssimare.

$$y_i = f(x_i)$$



In azzurro il polinomio ed in nero la funzione "vera", x_0, x_1 e x_2 sono i nodi

Dimostrazione 15: Dimostrazione dell'esistenza e dell'unicità "algebrica"

Le condizioni di interpolazione sono che $p_n(x_i) = y_i$ nei nodi che cerchiamo per le $n+1$ coppie (x_i, y_i). Il polinomio che cerchiamo è quindi di questa forma:

$$p_n(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

Applichiamo le condizioni di interpolazione al polinomio p_n:

$$\begin{aligned} p_n(x_0) &= a_0 + a_1 x_0 + a_2 x_0^2 + \cdots + a_n x_0^n \\ p_n(x_1) &= a_0 + a_1 x_1 + a_2 x_1^2 + \cdots + a_n x_1^n \\ &\vdots \\ p_n(x_n) &= a_0 + a_1 x_n + a_2 x_n^2 + \cdots + a_n x_n^n \end{aligned}$$

Sistema di equazioni lineari in $n+1$ incognite con $n+1$ soluzioni. Possiamo riscrivere il tutto come una matrice:

$$\underbrace{\begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{bmatrix}}_{V} \cdot \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix}}_{a} = \underbrace{\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}}_{y} \qquad Va = y \to a = V^{-1}y$$

Devo dimostrare che le soluzioni esistono e che siano uniche, per fare ciò basta dimostrare che il rango della matrice, che chiamiamo **matrice di Vandermonde**, sia uguale alla dimensione. Per fare ciò calcoliamo il **determinante** e vediamo che è diverso da 0.

$$det(V) = \prod (x_i - x_j) \neq 0 \text{ se } x_i \neq x_j \qquad i \neq j \text{ con } 0 \leq j < i \leq n$$

Esempio

$$V = \begin{bmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \end{bmatrix} \qquad det(V) = \prod_{0 \le j < i \le n} (x_i - x_j) = (x_2 - x_0)(x_2 - x_1)(x_1 - x_0)$$

Dimostrazione 16: Dimostrazione dell'unicità "costruttiva"

Cominciamo con dimostrare l'unicità per assurdo. Supponiamo esistano due polinomi $p_n(x)$ e $q_n(x)$ entrambi $\in \mathbb{P}_n(x)$ tali che:

$$\begin{cases} p_n(x_i) = y_i \\ q_n(x_i) = y_i \end{cases}$$

Vale che:

$$(p_n - q_n)(x_i) = p_n(x_i) - q_n(x_i) = 0 \qquad i = 0, 1, \dots, n$$

Questo significa che $p_n - q_n = 0$ se e solo se $p_n = q_n$, ma per il teorema fondamentale dell'algebra sappiamo che un polinomio di grado n ha al massimo n radici, non può annullarsi in più di n radici, altrimenti è il polinomio nullo. Segue che i due polinomi sono lo stesso polinomio.

Definizione 26: Polinomi elementari di Lagrange

Per dimostrare l'esistenza dobbiamo prima introdurre il **polinomi elementari di Lagrange**:

$$L_i(x) = \prod_{j=0, j \neq i}^{n} \frac{(x - x_j)}{(x_i - x_j)} = \frac{(x - x_0)(x - x_1) \dots (x - x_{i-1})(x - x_{i+1}) \dots (x - x_n)}{(x_i - x_0)(x_i - x_1) \dots (x_i - x_{i-1})(x_i - x_{i+1}) \dots (x_i - x_n)}$$

$$L_i(x_j) = \delta_{i,j} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

Per $n = 2$ avrò questi nodi $\{x_0, x_1, x_2\}$ e quindi ho che

$$L_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} \qquad L_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} \qquad L_2(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}$$

$$L_0(x_0) = \frac{(x_0 - x_1)(x_0 - x_2)}{(x_0 - x_1)(x_0 - x_2)} = 1 \qquad L_0(x_1) = 0 \qquad L_0(x_2) = 0$$

Discorso analogo per $L_1(x_0) = 0, L_1(x_1) = 1 \dots$. Questo è il polinomio d'interpolazione in forma di Lagrange:

$$P_n(x) = \sum_{i=0}^{n} y_i L_i(x)$$

Notiamo come la base canonica dei polinomi di grado n $\{1, x, x^2, x^3, \dots, x^n\}$ e la base dei polinomi elementari di Lagrange $\{L_0, L_1, \dots, L_n\}$ siano basi dello stesso spazio vettoriale. Ogni polinomio di grado n è esprimibile come combinazione lineare di polinomi elementari di Lagrange.

Teorema 7: Costruzione dell'interpolante in forma di Lagrange

Dati $n+1$ punti distinti $x_0, x_1 \dots x_n$ ed i valori $y_0, y_1 \dots y_n$ (i valori che assume la funzione nei nodi), il polinomio di grado al più n così definito:

$$p_n(x) = \sum_{i=0}^{n} y_i L_i(x)$$

è tale che $p_n(x) = y_i$ per $i = 0, 1 \dots n$ (è interpolante)

Il teorema dice due cose:

- Il polinomio è di grado al più n, ovvio dato che la somma di polinomi di Lagrange (che al più hanno grado n) risulta in un polinomio ancora di grado n.
- $p_n(x_i) = y_i$ di cui vediamo la dimostrazione

Dimostrazione 17

$$p_n(x) = \sum_{i=0}^{n} y_i L_i(x_i) = y_0 L_0(x_i) + \dots + y_i L_i(x_i) + \dots + y_n L_n(x_i)$$

Ma sappiamo che i polinomi di Lagrange valgono 0 se L_i e x_i non coincidono e quindi abbiamo dimostrato che è interpolante.

Esempio supponiamo di avere questi punti:

$$\begin{cases} x_0 = -2 & x_1 = 1 & x_2 = 3 \\ y_0 = -2 & y_1 = 11 & y_2 = 17 \end{cases}$$

Ho 3 punti quindi $P_2(x)$ un polinomio di interpolazione di grado 2 quindi:

$$L_0(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{(x-1)(x-3)}{(-3)(-5)} = \frac{1}{15}(x-1)(x-3)$$

$$L_1(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{(x+2)(x-3)}{3(-2)} = \frac{-1}{6}(x+2)(x-3)$$

$$L_2(x) = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{(x+2)(x-1)}{5 \cdot 2} = \frac{1}{10}(x+2)(x-1)$$

Quindi ottengo che

$$P_2 = -2L_0(x) + 11L_1(x) + 17L_2(x) = -\frac{2}{15}(x-1)(x-3) - \frac{11}{6}(x+1)(x-3) + \frac{17}{10}(x+2)(x-1)$$

Allora ho

$$P_2(x_0) = -2 \qquad P_2(x_1) = 11 \qquad P_2(x_2) = 17$$

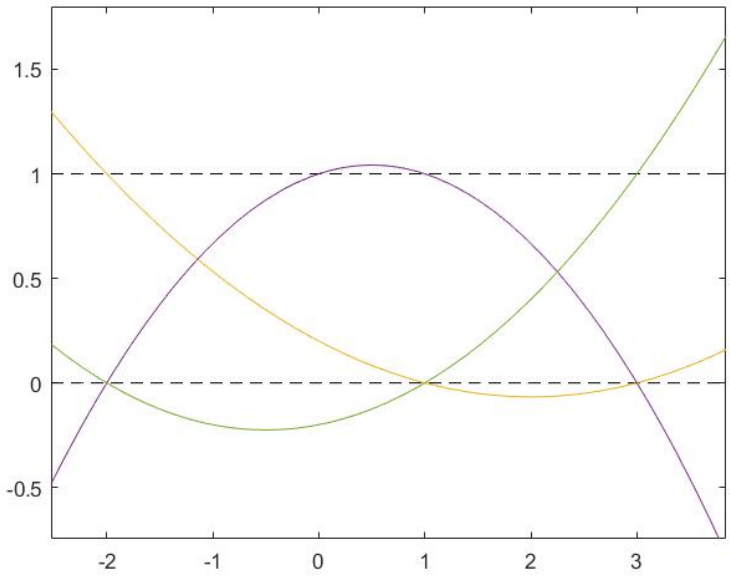


Grafico dei 3 polinomi di Lagrange, notiamo ad esempio come L0 (giallo) si annulla in $x_1 = 1$ e $x_2 = 3$ mentre assume valore 1 in $x_0 = -2$

Definizione 27: Errori di interpolazione

Definiamo **resto di interpolazione** o anche **errore di interpolazione**

$$r_n(x) = f(x) - p_n(x)$$

Dove $a \leq x_0 < x_1 < x_2 < \cdots < x_n \leq b$ con $x \in [a, b]$

Teorema 8: Formula dell'errore puntuale o Resto d'interpolazione

Sia $f \in C^{n+1}([a, b])$ e $p(x_i)$ il polinomio di interpolazione di f nei nodi $a \leq x_0 < x_1 < x_2 < \cdots < x_n \leq b$. Per ogni $x \in [a, b]$ esiste $\xi \in (a, b)$ tale che:

$$E_n(x) = r_n(x) = f(x) - p_n(x) = \frac{f^{(n+1)}(\xi) \cdot \omega_{n+1}(x)}{(n+1)!}$$

Il polinomio $\omega_{n+1}(x)$ prende il nome di **polinomio nodale**

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

Dal punto di vista **pratico**, oltre a ragionare per valore assoluto, non conosciamo ξ quindi sostituiamo con il massimo valore che $f^{(n+1)}$ assume in $[a, b]$

Teorema 9: Teorema di Rolle

Per la dimostrazione che segue ci servirà il **Teorema di Rolle**: data f una funzione continua e derivabile in $[a,b]$ e tale che $f(a)=f(b)$ allora esiste almeno un punto $c\in[a,b]$ tale per cui $f'(c)=0$

Dimostrazione 18

Se x è un nodo d'interpolazione il polinomio nodale vale 0 e quindi l'errore chiaramente è 0. Se x è diverso da tutti i nodi d'interpolazione x_i allora il polinomio nodale è diverso da 0, $\omega_{n+1}(x)\neq 0$. Definiamo questa funzione:

$$g(y)=r_n(y)-\omega_{n+1}(y)\cdot\frac{r_n(x)}{\omega_{n+1}(x)}$$

Con $y\in[a,b]$, volendo possiamo riscrivere $g(y)$:

$$g(y)=\underbrace{f(y)}_{\in C^{n+1}([a,b])}-\underbrace{p_n(y)}_{\in C^\infty}-\underbrace{\omega_{n+1}(y)}_{\in C^\infty}\frac{r_n(x)}{\omega_{n+1}(x)}$$

Quindi il tutto è derivabile $n+1$ volte. Si annulla nei nodi x_i con $i=0,\dots,n$ e quindi:

$$g(x_i)=r_n(x_i)-\omega_{n+1}(x_i)\cdot\frac{r_n(x_i)}{\omega_{n+1}(x_i)}=0$$

Si annulla anche in x perché:

$$g(x)=r_n(x)-\omega_{n+1}(x)\cdot\frac{r_n(x)}{\omega_{n+1}(x)}=0$$

Dato che $g\in C^{n+1}([a,b])$ e si annulla in $n+2$ punti, usando il Teorema di Rolle possiamo concludere che la funzione $g'(x)$ si annulla in $n+1$ punti di $[a,b]$. Ora possiamo considerare $g'(x)$ che si annulla in $n+1$ punti ma allora di nuovo per il Teorema di Rolle $g''(x)$ si annulla in n punti. Procediamo in questo modo fino ad arrivare a $g^{(n+1)}$ che si annulla in un solo punto, questo punto è proprio ξ.

$$g^{(n+1)}(y)=r_n^{(n+1)}(y)-(n+1)!\cdot\frac{r_n(x)}{\omega_{n+1}(x)}$$

Dove $(n+1)!$ sbuca perché derivando $n+1$ volte $\omega_{n+1}(y)$ ottengo $(n+1)!$

$$0=g^{n+1}(\xi)=r_n^{(n+1)}(\xi)-(n+1)!\cdot\frac{r_n(x)}{\omega_{n+1}(x)}$$

Isolando $r_n(x)$ ottengo:

$$r_n(x)=\frac{\omega_{n+1}(x)\cdot r_n^{(n+1)}(\xi)}{(n+1)!}=\frac{f^{n+1}(\xi)\omega_{n+1}(x)}{(n+1)!}$$

Dato che $r_n^{(n+1)}(\xi)=f^{(n+1)}(\xi)-0=f^{(n+1)}(\xi)$

Esempio Supponiamo di voler calcolare $|r_n(x)|$ per la funzione $f(x) = e^x$ sull'intervallo $[-1, 1]$ avendo a disposizione 5 nodi equi-spaziati $x_i = -1 + i \cdot h$ con $h = \frac{1}{2}$ e $i = 0, \ldots, 4$. Cioè di fatto consideriamo:

$$x_0 = -1 \quad x_1 = \frac{-1}{2} \quad x_2 = 0 \quad x_3 = \frac{1}{2} \quad x_4 = 1$$

Vale che:

$$|r_n(x)| = \frac{|f^{(5)}(\xi)| \cdot |\omega_5(x)|}{5!}$$

Possiamo calcolare il polinomio nodale:

$$|\omega_5(x)| = (x+1)(x+\frac{1}{2})(x-0)(x-\frac{1}{2})(x-1) \leq 0.12$$

$f^{(5)}(x) = e^x$ e dato che non conosciamo ξ:

$$max(|f^{(5)}(x)|) = e \qquad \text{con } x \in [-1, 1]$$

E allora concludiamo che:

$$|r_n(x)| \leq \frac{0.12}{120} \approx 0.0027 = 2.7 \cdot 10^{-3}$$

Definizione 28: Convergenza dell'interpolante polinomiale

Per ogni n abbiamo un polinomio diverso:

$$\begin{aligned} n = 0 \quad & p_0 \text{ (una costante)} \\ n = 1 \quad & p_1 \\ n = 2 \quad & p_2 \\ & \vdots \\ n \to \infty \quad & p_n \end{aligned}$$

Ci chiediamo se questa successione di funzioni $\{p_n\}_{n\in\mathbb{N}}$ si comporti esattamente come $f(x)$ per n sufficientemente grandi. In altre parole ci chiediamo se converge ad f

$$\lim_{n\to\infty} \{p_n\}_{n\in\mathbb{N}} \to f$$

Esistono due tipologie di convergenze:

- **Convergenza puntuale**, analizziamo che succede nei punti dell'intervallo:

$$\lim_{n\to\infty} |f(x) - p_n(x)| = 0 \ \forall x \in I = [x_0, x_n]$$

- **Convergenza uniforme** (quella che ci interessa). Si misura sfruttando la norma del massimo:

$$\lim_{n\to\infty} ||f - p_n||_\infty = 0$$

  Norma infinito equivale al massimo valore che la funzione assume in quell'intervallo infatti possiamo riscrivere la condizione di convergenza uniforme così:

$$\lim_{n\to\infty} \max_{x\in I} |\underbrace{f(x) - p_n(x)}_{E_n(x)}| = 0$$

In generale la convergenza uniforme non è sempre verificata.

Se vogliamo ottenere una maggiorazione di $||E_n||_\infty$ possiamo svolgere:

$$||f - p_n||_\infty \leq \max_{x\in[a,b]} |f^{(n+1)}(x)| \frac{(b-a)^{n+1}}{(n+1)!}$$

Definizione 29: Formula dell'errore per nodi equispaziati

I nodi equispaziati sono i nodi così definiti:

$$x_i = x_0 + i \cdot h \qquad i = 0, 1, \dots, n \quad h > 0 \qquad x_{i+1} = x_i + h$$

Se $x_0 = a$ e $x_n = b$, che equivale a dire $[a, b]$, allora $h = \frac{b-a}{n}$. Per i nodi equispaziati si può dimostrare che $\forall x \in I$ con $I = [x_0, x_n]$ il polinomio nodale in valore assoluto assume questi valori:

$$|\omega_{n+1}(x)| = \left| \prod_{i=0}^{n} (x - x_i) \right| \leq \frac{h^{n+1}}{4} \cdot n!$$

Quando sostituiamo questa espressione nel calcolo dell'errore otteniamo a sua volta una maggiorazione per l'errore sui nodi equispaziati.

$$|E_n(x)| \leq \frac{\max_{x \in I} |f^{(n+1)}(x)| \cdot |\omega_{n+1}(x)|}{(n+1)!} \leq \frac{\max_{x \in I} |f^{(n+1)}(x)|}{(n+1)n!} \frac{h^{n+1} n!}{4}$$

Il tutto allora diventa

$$|E_n(x)| \leq \max_{x \in I} |f^{(n+1)}(x)| \cdot \frac{h^{n+1}}{4(n+1)}$$

Abbiamo che la frazione di destra tende a 0 per $n \to \infty$ (ricorda che $h = \frac{b-a}{n}$). Tuttavia questa espressione non mi permette di dedurre che l'errore tende a 0 perché esistono delle funzioni dove il massimo valore assunto dalla derivata $(n+1)$-esima tende ad ∞

Definizione 30: Funzioni di Runge e Fenomeno di Runge

Esistono funzioni dove $|E_n(x)| \to \infty$, tra queste vi è la **funzione di Runge**:

$$f(x) = \frac{1}{1 + x^2} \qquad I = [-5, 5]$$

Se la interpoliamo su questo intervallo e su **nodi equispaziati** otteniamo $\lim_{n \to \infty} |E_n(x)| = \infty$. Questo perché l'ordine di infinito del termine $\max_{x \in I} |f^{(n+1)}(x)|$ supera l'ordine di infinitesimo di $\frac{h^{n+1}}{4(n+1)}$.

Fenomeno di Runge: aumentando il grado del polinomio d'interpolazione otteniamo una rappresentazione della funzione peggiore, cioè l'errore aumenta.

Definizione 31: Nodi di Chebyshev

Esistono dei nodi ottimali per svolgere l'interpolazione polinomiale tra questi vi sono i **nodi di Chebyshev** che corrispondono alle radici dei polinomi di Chebyshev. Fissato un grado n abbiamo:

- Nodi di Gauss-Chebyshev scalati su $[a, b]$:

$$x_k = \frac{a+b}{2} + \frac{b-a}{2} t_k \quad k = 0, \dots, n$$

$$t_k = \cos\left(\frac{2k+1}{2n+2}\right) \quad k = 0, \dots, n$$

- Nodi di Chebyshev-Lobatto scalati su $[a, b]$:

$$x_k = \frac{a+b}{2} + \frac{b-a}{2} t_k \quad k = 0, \dots, n$$

$$t_k = \cos\left(\frac{k\pi}{n}\right) \quad k = 0, \dots, n$$

Polinomio di Chebyshev:

$$T_k(x) = \cos(k \arccos(x))$$

Per l'interpolante polinomiale su nodi equispaziati non c'è nemmeno la **convergenza puntuale** ovvero esistono punti per cui:

$$\lim_{n\to\infty} |f(x) - p_n(x)| \neq 0$$

Per i nodi di Chebyshev si dimostra la **convergenza uniforme** dell'interpolante alla funzione di Runge. Il motivo è che in questo caso per i nodi di Chebyshev, dato che sono ammassati agli estremi, la maggiorazione che si ha per il polinomio nodale è tale per cui:

$$|\omega_{n+1}(x)| \to 0$$

Quindi quando calcoliamo la formula dell'errore abbiamo:

$$|E_n(x)| \leq \frac{\overbrace{\max_{x\in I} |f^{(n+1)}(x)|}^{\approx (n+1)!} \cdot |\omega_{n+1}(x)|}{(n+1)!}$$

Teorema 10: Teorema di Faber

Per ogni distribuzione di nodi esiste almeno una funzione $f \in C([a, b])$ con $-\infty < a < b < +\infty$ tale che l'errore d'interpolazione $||E_n(f)||_\infty$ non converge a 0 per $n \to \infty$. In soldoni i nodi di Chebyshev non garantiscono la convergenza uniforme dell'interpolante per tutte le funzioni

Teorema 11

Per ogni funzione $f \in C([a,b])$ con $-\infty < a < b < +\infty$ esiste almeno una distribuzione di nodi tale che $||E_n(f)||_\infty \to 0$ per $n \to \infty$. In maniera complementare a quanto detto nel teorema sopra esistono sempre dei nodi che garantiscono la convergenza uniforme.

Teorema 12: Teorema di Bernstein

Per ogni funzione $f \in C^1([a,b])$ con $-\infty < a < b < +\infty$ se p_n è l'interpolante di f in $n+1$ nodi di Chebyshev allora $||E_n[f]||_\infty \to 0$ per $n \to +\infty$. Cioè abbiamo definito la classe di funzioni per le quali i nodi di Chebyshev garantiscono la convergenza uniforme del polinomio interpolante.

4.1.1 Formula di Newton del polinomio d'interpolazione

I polinomi elementari che costituiscono la base di Lagrange devono essere ricostruiti da 0 se decidiamo di aumentare di un solo nodo l'intervallo. In generale cambiamenti anche di un solo nodo implica il ricalcolare tutto quanto. Introduciamo quindi un nuovo metodo per calcolare il polinomio d'interpolazione più efficiente dal punto di vista computazionale e più "flessibile"

Definizione 32: Differenze divise

Dati $n+1$ punti (x_i, y_i) con $i = 0, \dots, n$ e $x_i \neq x_j$ e $i \neq j$ presi due punti x_0 e x_1 qualsiasi si chiama **differenza divisa prima** o **differenza divisa di ordine uno** di $f(x)$ la funzione:

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} \quad \text{(E' il rapporto incrementale)}$$

Per il teorema del valor medio questo rapporto incrementale coincide con il valore della derivata prima di f in un certo punto ξ, $f[x_0, x_i] = f'(\xi)$ per $\xi \in (x_0, x_1)$. Notiamo anche che l'ordine dei punti non cambia il valore della differenza divisa:

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f(x_0) - f(x_1)}{x_0 - x_1} = f[x_1, x_0]$$

Dati 3 punti x_0, x_1, x_2 si chiama **differenza divisa d'ordine due** la funzione:

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$$

In generale dati $n+1$ punti la **differenza divisa di ordine n** della funzione f è:

$$f[x_0, x_1, \dots, x_n] = \frac{f[x_1, \dots, x_n] - f[x_0, x_1, \dots, x_{n-1}]}{x_n - x_0}$$

Esempio Supponiamo di avere questi nodi d'interpolazione:

x_i	-1	0	1
y_1	2	1	3

$$\begin{array}{ll} -1 & 2 \\ & \frac{1-2}{0-(-1)} = -1 \\ 0 & 1 \qquad\qquad \frac{2-(-1)}{1-(-1)} = \frac{3}{2} \\ & \frac{3-1}{1-0} = 2 \\ 1 & 3 \end{array}$$

$$P_2(x) = 2 - 1(x+1) + \frac{3}{2}(x+1)x = 2 - x - 1 + \frac{3}{2}x^2 + \frac{3}{2}x = 1 + \frac{1}{2}x + \frac{3}{2}x^2$$

Teorema 13: Polinomi di interpolazione in forma di Newton

Data $f(x)$ in I con $x_i \neq x_j$ e $0 \leq i \leq n$ si ha che:

$$p_n(x) \begin{cases} f(x) = f(x_0) + f[x_0, x_1](x - x_0) \\ \qquad + f[x_0, x_1, x_2](x - x_0)(x - x_1) \\ \qquad + f[x_0, x_1, x_2, x_3](x - x_0)(x - x_1)(x - x_2) \\ \qquad \vdots \\ \qquad + f[x_0, x_1, x_2, x_3, \ldots, x_n](x - x_0)\ldots(x - x_{n-1}) \end{cases}$$

$$E_n(x) \begin{cases} \qquad + f[x_0, x_1, \ldots, x_n, x](x - x_0)(x - x_1)\ldots(x - x_n) \end{cases}$$

Osserviamo come:

$$E_n(x) = f(x) - p_n(x) = f[x_0, x_1, \ldots, x_n, x] \cdot \omega_{n+1}(x) = \frac{f^{(n+1)}(\xi) \cdot \omega_{n+1}(x)}{(n+1)!}$$

E quindi:

$$f[x_0, x_1, \ldots, x_n, x] = \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

Corollario 2

Se $f \in C^k([x_0, x_k])$ allora esiste almeno un punto $\xi \in (x_o, x_k)$ tale che

$$f[x_0, x_1, \ldots, x_k] = \frac{f^{(k)}(\xi)}{k!}$$

Definizione 33: Differenze divise per punti coincidenti

Fino ad adesso abbiamo sempre assunto che i nodi $x_0, x_1, \ldots, x_n$ fossero tutti diversi, vediamo come calcolare le differenze divise per punti coincidenti. Ricordiamo la definizione di differenza divisa

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} \qquad \text{Per calcolarla necessariamente } x_1 \neq x_0$$

Per trattare il caso dove $x_0 = x_1$ dobbiamo **passare ai limiti**:

$$f[x_0, x_1] = \lim_{x_1 \to x_0} \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} = f'(x_0)$$

Quindi per calcolare la differenza divisa in punti coincidenti dobbiamo calcolare la **derivata prima**.

Analizziamo ora il caso dove abbiamo 3 punti uguali, $x_0 = x_1 = x_2$:

$$f[x_0, x_1, x_2] = f[x_0, x_0, x_0] = \lim_{\substack{x_1 \to x_0 \\ x_2 \to x_0}} f[x_0, x_1, x_2] = \lim_{\substack{x_1 \to x_0 \\ x_2 \to x_0}} \frac{f^{(2)}(\xi)}{2}$$

Stiamo dicendo che esiste un punto che rende vera l'uguaglianza (Ricordiamo che ξ dipende dal punto x dove stiamo calcolando l'errore). Più sono i punti uguali più diminuisce l'intervallo dove si trova ξ ed in generale tende a x_0

$$\lim_{\substack{x_1 \to x_0 \\ x_2 \to x_0}} \frac{f^{(2)}(\xi)}{2} = \frac{f^{(2)}(x_0)}{2}$$

Nel caso generale di $k+1$ punti coincidenti otteniamo:

$$f[\underbrace{x_0, x_0, \ldots, x_0}_{k+1}] = \frac{f^{(k)}(x_0)}{k!}$$

Esempio Siano assegnati due punti $x_0 = 0$ e $x_1 = 2$ e sia $f(x_0) = 1$ e $f(x_1) = 7$. Sia inoltre $f'(x_0) = 4$ e $f^{(2)}(x_0) = 8$. Scrivere il polinomio d'interpolazione in forma di Newton. Ho 3 condizioni per x_0 ed una condizione per x_1 quindi il polinomio sarà di grado 3: $p_3(x)$

$$\begin{array}{cc|cc} x_0 & f(x_0) & 0 & 1 \\ & & & f[x_0, x_0] = f'(x_0) = 4 \\ x_0 & f(x_0) & 0 & 1 \\ & & & f[x_0, x_0] = f'(x_0) = 4 \\ x_0 & f(x_0) & 0 & 1 \\ & & & f[x_0, x_1] = (7-1)/2 \\ x_1 & f(x_1) & 2 & 7 \end{array}$$

$$f[x_0, x_0, x_0] = f^{(2)}(x_0)/2 = 4 \qquad (-0.5 - 4)/2 = -9/4$$

$$f[x_0, x_0, x_1] = (3-4)/2$$

$$p_3(x) = 1 + 4(x - x_0) + 4(x - x_0)(x - x_0) - \frac{9}{4}(x - x_0)(x - x_0)(x - x_0) = 1 + 4x + 4x^2 - \frac{9}{4}x^3$$

Notiamo che se tutti i punti fossero coincidenti la forma del polinomio d'interpolazione sarebbe:

$$f(x_0) + f'(x_0)(x - x_0) + \frac{f^{(2)}(x_0)(x - x_0)^2}{2!} + \frac{f^{(3)}(x_0)(x - x_0)^3}{3!} + \ldots$$

Quindi nel caso in cui tutti i punti sono coincidenti la formula di Newton equivale alla formula di Taylor.

4.2 Approssimazione di funzioni ai minimi quadrati discreti

L'idea è che date delle determinate misurazioni della funzione f vogliamo trovare una funzione che mi permetta di predire il valore della funzione in altri punti. Vogliamo cercare una funzione semplice (modello), per noi sarà un polinomio, che si discosti dalla funzione f (incognita) il meno possibile, questo "meno possibile" implica rendere minima una norma. L'approssimazione ai minimi quadrati si basa sul rendere minima la **norma euclidea vettoriale**.

Definizione 34: Norma euclidea vettoriale

Dato x un vettore definiamo **norma euclidea vettoriale** la radice della somma di tutti i valori di un vettore elevati al quadrato:

$$||x||_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$$

Definizione 35: Approssimazione ai minimi quadrati discreti

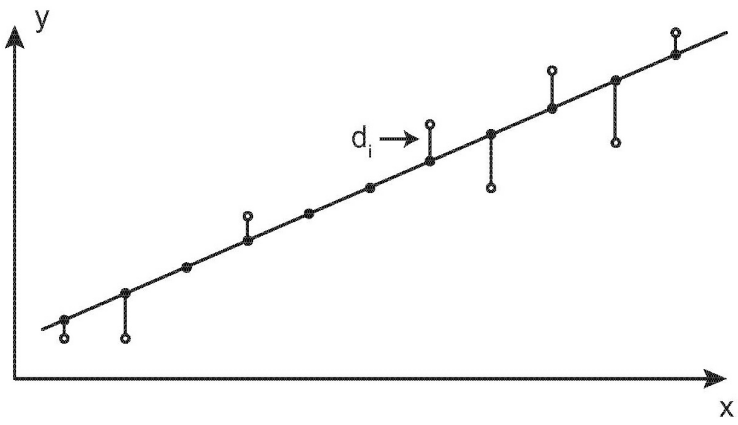
Dati una serie di dati sperimentali $\{(x_i, \overbrace{y_i}^{f(x_i)}), i = 0, \dots, n\}$ e data una base nello spazio dei polinomi di grado $\leq n$ che denotiamo in questo modo $\mathbb{P}_n(x)$ (ad esempio $\{1, x, x^2, \dots, x^n\}$) cerco il polinomio p_m con $m << n$

$$p_m = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m = \sum_{i=0}^{m} a_i \cdot x^i$$

tale per cui la quantità:

$$\sum_{i=1}^{n} (y_i - p_m(x_i))^2$$

sia **minima**. In altre parole cerco il p_m che renda minima la norma 2 del vettore differenza tra i valori $f(x_i)$ ed i valori $p_m(x_i)$ (L'aggettivo discreto distingue la norma 2 che utilizziamo, cioè quella nello spazio $\mathbb{R}^n$, dalla norma 2 nello spazio di funzioni continue). Questo processo viene anche chiamato **minimizzazione degli scarti verticali**.



Supponiamo $m = 1$ cerchiamo quel polinomio (in questo caso una retta) che minimizzi la norma 2 tra i punti della funzione "vera" e quelli del polinomio.

La ricerca di questo polinomio può essere svolta con due approcci diversi, l'approccio dell'analisi matematica e quello dell'algebra.

4.2.1 Ricerca del polinomio minimo tramite l'analisi matematica

Definizione 36: Polinomio di miglior approssimazione ai minimi quadrati

Definiamo $p_m(x)$ come il **polinomio di miglior approssimazione ai minimi quadrati** di una serie di dati $\{(x_i, y_i), i = 1, \dots, n\}$ se:

$$\Phi(\underbrace{b_0, b_1, \dots, b_m}_{\text{coeff del polinomio}}) = \sum_{i=1}^{n} \Big[y_i - p_m(x_i) \Big]^2 \leq \sum_{i=1}^{n} \Big[y_i - q_m(x_i) \Big]^2 \qquad \forall\ q_m \in \mathbb{P}_m(x)$$

Il nostro obiettivo è quindi:

$$\min \Phi(b_0, b_1, \dots, b_m) \qquad \Phi(b_0, b_1, \dots, b_m) \in \mathbb{R}^{m+1}$$

Vediamo come risolvere il problema di minimo nel caso in cui $m = 1$, quindi il nostro modello è una retta

$$y = a_0 + a_1 x = p_1(x)$$

La quantità che dobbiamo minimizzare è:

$$\Phi(b_0, b_1) = \sum_{i=1}^{n} \Big[y_i - p_1(x) \Big]^2 = \sum_{i=1}^{n} \Big[y_i - (b_0 + b_1 x_i) \Big]^2 = \sum_{i=1}^{n} y_i^2 + (b_0 + b_1 x_i)^2 - 2y_i(b_0 + b_1 x_i)$$

Ovviamente i punti di min/max delle funzioni derivabili sono punti **stazionari**, ovvero punti

dove la derivata si annulla. Quando le variabili sono multiple parliamo di **gradienti**, ovvero il vettore delle **derivate parziali**.

$$\nabla\Phi = \begin{pmatrix} \frac{\delta\Phi}{\delta b_0} \\ \\ \frac{\delta\Phi}{\delta b_1} \end{pmatrix}$$

In sintesi ci calcoliamo le derivate parziali, imponiamo le componenti uguali a 0 e ci troviamo b_0 e b_1. Nel caso di $p_m = 1$ abbiamo:

$$\begin{cases} \frac{\delta\Phi}{\delta b_0}(b_0, b_1) = \sum_{i=1}^{n} 2(b_0 + b_1 x_i) - 2y_i = 0 \\ \\ \frac{\delta\Phi}{\delta b_1}(b_0, b_1) = \sum_{i=1}^{n} 2(b_0 + b_1 x_i)x_1 - 2y_i x_i = 0 \end{cases}$$

Possiamo riscrivere il tutto così:

$$\begin{cases} \sum_{i=1}^{n}(b_0 + b_1 x_i) = \sum_{i=1}^{n} y_i \\ \\ \sum_{i=1}^{n}(b_0 x_i + b_1 x_i^2) = \sum_{i=1}^{n} x_i y_i \end{cases}$$

Ed in forma matriciale otteniamo:

$$\begin{pmatrix} \frac{\delta\Phi}{\delta b_0}(b_0, b_1) \\ \\ \frac{\delta\Phi}{\delta b_1}(b_0, b_1) \end{pmatrix} = \begin{pmatrix} 0 \\ \\ 0 \end{pmatrix}$$

Che diventa:

$$\begin{pmatrix} \sum_{i=1}^{n} 1 & \sum_{i=1}^{n} x_i \\ \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} b_0 \\ \\ b_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

Questo rappresenta il sistema da risolvere per trovare la retta, ovvero il nostro polinomio di miglior approssimazione ai minimi quadrati quando $m = 1$. Questa tecnica di ricerca della retta è detta **regressione lineare**.

4.2.2 Ricerca del polinomio minimo tramite l'algebra lineare

Fissiamo un polinomio di un generico grado m

$$p_m = b_0 + b_1 x + \cdots + b_m x^m$$

Quando cercavamo il polinomio interpolante abbiamo imposto la condizione di passaggio, creando un sistema di equazioni dove la matrice dei coefficienti (Vandermonde) aveva questa forma:

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^m \\ \\ 1 & x_1 & x_1^2 & \dots & x_1^m \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_m & x_m^2 & \dots & x_m^m \end{bmatrix}$$

Analizzando il caso della retta $p_1(x) = b_0 + b_1 x$ con n dati sperimentali se imponiamo il passaggio per i punti otteniamo:

$$\begin{cases} b_0 + b_1 x_0 = y_0 \\ b_0 + b_1 x_0 = y_1 \\ b_0 + b_1 x_0 = y_2 \\ \vdots \\ b_0 + b_1 x_0 = y_n \end{cases}$$

Ovvero un **sistema sovradeterminato** che non ha soluzioni.

$$\begin{pmatrix} 1 & x_0 \\ 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \cdot \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \longrightarrow \qquad Vb = y$$

Per il metodo dei minimi quadrati cerchiamo il vettore b che rende minimo la differenza:

$$\|Vb - y\|_2$$

Quindi se imponiamo il passaggio per i punti (dati sperimentali) $\{(x_i, y_i) i = 1 \dots n\}$ al polinomio $p_m(x) \in \mathbb{P}_m(x)$ con $m << n$ cioè di grado basso abbiamo rispetto ai dati sperimentali otteniamo:

$$p_n(x) = a_0 + a_1 x + \cdots + a_n x^m \qquad p_1(x) a_0 + a_1 x$$

$$\begin{cases} x_1 : a_0 + a_1 x_1 + \cdots + a_n x_1^m &= y_1 \\ x_2 : a_0 + a_1 x_2 + \cdots + a_n x_2^m &= y_2 \\ & \vdots \\ x_n : a_0 + a_1 x_n + \cdots + a_n x_n^m &= y_n \end{cases} \quad \rightarrow y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad a = \begin{bmatrix} a_0 \\ \vdots \\ a_m \end{bmatrix} \rightarrow Va = y$$

Quello che otteniamo in forma matriciale è $V \in \mathbb{R}^{n \times (m+1)}$ rettangolare $n >> m$ quindi moltissime righe rispetto alle colonne, quindi un sistema sovrastimato.

Risolvere il sistema significa calcolare l'anti-immagine di un vettore per una trasformazione lineare. Nel nostro caso abbiamo una matrice $A \in R^{n \times m}$ una trasformazione lineare fissata tra due spazi vettoriali: $\mathbb{R}^m \rightarrow \mathbb{R}^n$. Porre $Ax = b$ significa chiedersi se esista un vettore in R^m tale che la sua immagine è $b \in \mathbb{R}^n$. Ci sono due sottospazi fondamentali che caratterizzano l'applicazione lineare:

- Il **nucleo** $ker(A) \in \mathbb{R}^m$, quando $ker(A) = \{0\}$ allora la soluzione è unica.
- L'**immagine** $im(A) \in \mathbb{R}^n$, tutti i vettori che sono immagini di vettori dello spazio di partenza ($\mathbb{R}^m$)

Per dimensione di spazio vettoriale intendiamo il numero di vettori di una sua base, per fare un esempio nello spazio vettoriale dei polinomi di grado due $\mathbb{P}_2(x)$ una sua base è $\{1, x, x^2\}$ e quindi la dimensione di $dim(\mathbb{P}_2) = 3$.

Teorema 14: Teorema delle dimensioni

Data un'applicazione lineare $A \in \mathbb{R}^{n\times m}$: $\mathbb{R}^m \to \mathbb{R}^n$ la dimensione dello spazio d'uscita m è uguale alla dimensione del nucleo più la dimensione dell'immagine:

$$dim(\mathbb{R}^n) = m = dim(ker(A)) + dim(Im(A))$$

Noi abbiamo:

$$Va = y \quad \text{con } V \in \mathbb{R}^{n\times m+1} \quad : \ \mathbb{R}^{m+1} \to \mathbb{R}^n$$

E quindi dal teorema:

$$m+1 = dim(ker(V)) + \underbrace{dim(Im(V))}_{rank(V)}$$

Se V è di **rango pieno** (massimo), ricordando che il rango equivale al numero di colonne o di righe linearmente indipendenti, ho che:

$$3 = \underbrace{dim(ker(V))}_{0} + \underbrace{dim(Im(V))}_{3}$$

Risolvere $Va = y$ nel senso dei minimi quadrati significa trovare il vettore $\overline{a}$ che minimizza la norma 2 della differenza, $||Va - y||_2^2$. Risolvere o trovare il vettore $\overline{a} \in \mathbb{R}^{m+1}$ che minimizza la distanza $||Va - y||_2^2$ è equivalente a risolvere il sistema lineare

$$V^T V a = V^T y \qquad \text{Equazioni normali}$$

$Va = y$ con $y \notin Im(V)$. Imponiamo che $y - V\overline{a}$ sia ortogonale ad una base di V, ad esempio alle colonne di V. Quindi il prodotto scalare tra $y - V\overline{a}$ e le colonne di V sia 0. Con $< V^0, y - V\overline{a} >$ indichiamo il prodotto scalare tra la colonna 0 di V e $y - V\overline{a}$, quindi deve valere che:

$$\begin{cases} < V^0, y - V\overline{a} >= 0 \\ < V^1, y - V\overline{a} >= 0 \\ \qquad \vdots \\ < V^m, y - V\overline{a} >= 0 \end{cases} \qquad \text{Che equivale a} \quad V^T(y - V\overline{a}) = 0$$

Se V ha rango pieno $rank(V) = n+1$ allora $V^T V$ è invertibile e la soluzione è unica. Nel caso della retta ho:

$$p_1(x) = a_0 + a_1 x$$

$$V = \begin{pmatrix} 1 & x_0 \\ 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad a = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} \quad y = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Quindi

$$V^T V \overline{a} = V^T y$$

Diventa

$$\begin{pmatrix} 1 & 1 & \dots & 1 \\ x_0 & x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_0 \\ 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_0 & x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Che in forma compatta diventa

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} {}_y i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

Che è esattamente quello che avevamo ottenuto approcciando il problema tramite l'analisi matematica.

Esempio Dati i dati sperimentali della seguente tabella:

1	2	4	5
1.8364	2.2232	1.4025	0.3985

Calcoliamo la retta di regressione lineare, cioè il polinomio di grado 1 che meglio approssima i dati rispetto ai minimi quadrati.

$$\begin{pmatrix} 4 & 12 \\ 12 & 46 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 5.8579 \\ 13.872 \end{pmatrix}$$

Svogliamo un passo di Gauss

$$\begin{pmatrix} 4 & 12 \\ 0 & 10 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 5.8579 \\ -3.7021 \end{pmatrix} \Rightarrow \begin{cases} a_0 = 2.5751 \\ a_1 = -0.37021 \end{cases}$$

Il che vuol dire che la retta che meglio approssima rispetto ai minimi quadrati è:

$$P_1(x) = 2.5751 - 0.37021x$$

Esempio Dati i dati sperimentali della seguente tabella:

1	1.5	2	2.5	3
0	0.10820	0.38629	0.79073	1.2983

Trovare la retta $y = a_0 + a_1 x$ che meglio approssima i dati nel senso dei minimi quadrati:

$$\begin{pmatrix} 1 & 1 \\ 1 & 1.5 \\ 1 & 2 \\ 1 & 2.5 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 0.0 \\ 0.10820 \\ 0.38629 \\ 0.79073 \\ 1.2983 \end{pmatrix}$$

$$Va = y \iff V^T V a = V^T y$$

$$\begin{pmatrix} 5 & 10 \\ 10 & 29.5 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 2.581 \\ 6.7991 \end{pmatrix}$$

$$\begin{pmatrix} 5 & 10 \\ 0 & 2.5 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 2.581 \\ 1.6371 \end{pmatrix}$$

$$a_1 = 0.6548 \qquad a_0 = -0.7934$$

E quindi la retta che meglio approssima i dati è $y = -0.7934 + 0.6548x$

Esercizio proposto Calcolare il miglior polinomio d'approssimazione ai minimi quadrati di grado 2 considerando i seguenti dati sperimentali:

0	0.7	1.5	2.5	3.4
-12	-3.0	2.8	3.2	-4

5 Integrazione Numerica

La **integrazione numerica** o anche detta **quadratura numerica** si occupa di trovare valori approssimati per integrali definiti. Si vuole approssimare l'integrale definito di una funzione $f \in C([a,b])$ con $[a,b]$ intervallo limitato. L'idea è di sostituire

$$I(f) = \int_a^b f(x)dx$$

Con una formula detta **di quadratura** della forma

$$I_n(f) = \sum_{i=1}^{n} w_i f(x_i)$$

I termini w_i e $x_i \in [a,b]$ sono detti rispettivamente **pesi** e **nodi** della formula. Nota che **n è il numero di nodi di quadratura**. Questa sommatoria del prodotto di $w_i f(x_i)$ chiaramente equivale al prodotto scalare dei vettori:

$$w = \begin{pmatrix} w_1 & w_2 & \dots & w_n \end{pmatrix} \qquad f = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{pmatrix} \quad \rightarrow \quad < w, f >$$

Teorema 15: Stabilità d'integrazione

L'operazione funzionale di integrazione nel continuo è stabile, ovvero se $\tilde{f}$ approssima una funzione $f \in C([a,b])$ con $[a,b]$ intervallo limitato.

$$\left| \int_a^b f(x)dx - \int_a^b \tilde{f}(x)dx \right| \leq (b-a) \underbrace{\max_{x\in[a,b]} |f(x) - \tilde{f}(x)|}_{\|f-\tilde{f}\|_\infty}$$

Cioè se prendiamo $\tilde{f}$ che è vicina ad f relativamente ad una distanza che abbiamo definito così:

$$dist(f, \tilde{f}) = \max_{x\in[a,b]} |f(x) - \tilde{f}(x)|$$

Allora $\int_a^b \tilde{f}(x)dx$ non può essere arbitrariamente distante da $\int_a^b f(x)dx$ concetto descritto come **stabilità del funzionale d'integrazione**.

Dimostrazione 19

Il teorema segue dal fatto che:

$$\left| \int_a^b f(x)dx - \int_a^b \tilde{f}(x)dx \right| \leq \max_{x\in[a,b]} |f(x) - \tilde{f}(x)| \int_a^b 1 \cdot dx = (b-a) \max_{x\in[a,b]} |f(x) - \tilde{f}(x)|$$

Corollario 3

Sia $\{f_n\}$ una successione di funzioni continue che converge uniformemente a $f \in C([a,b])$ con $[a,b]$ intervallo limitato ovvero:

$$\lim_n dist(f_n, f) = \lim_n \max_{x\in[a,b]} |f_n(x) - f(x)| = 0$$

allora

$$\lim_n \left| \int_a^b f_n(x)dx - \int_a^b f(x)dx \right| = 0$$

Cioè

$$\int_a^b f_n(x)dx \to \int_a^b f(x)dx$$

L'idea è di ottenere la formula di quadratura sostituendo ad f un suo polinomio interpolante p_n e integrare quest'ultimo. Quindi sia p_n l'unico polinomio che interpola f relativamente ai nodi $\{x_i\}_{i=0,\dots,n}$ e sia L_k il k-esimo polinomio di Lagrange:

$$L_k(x) = \frac{(x-x_o)\dots(x-x_{k-1})(x-x_{k+1})\dots(x-x_n)}{(x_k-x_0)\dots(x_k-x_{k-1})(x_k-x_{k+1})\dots(x_k-x_n)}$$

Sapendo che:

$$p_n(x) = \sum_{k=0}^{n} f(x_k)L_k(x)$$

Possiamo scrivere

$$\int_a^b f(x)dx \approx \int_a^b p_n(x)dx = \int_a^b \sum_{k=0}^{n} f(x_k)L_k(x)dx = \sum_{k=0}^{n} f(x_k)\int_a^b L_k(x)dx = \sum_{k=0}^{n} w_k f(x_k)$$

Quindi i **pesi sono gli integrali definiti nell'intervallo** $[a,b]$ **dei polinomi dei Lagrange**.

Definizione 37: Grado di precisione

Diciamo che una formula :

$$\int_a^b f(x)w(x)dx \approx \sum_{i=1}^{M} w_i f(x_i)$$

- ha **grado di precisione almeno** N se e solo se è esatta per tutti i polinomi f di grado inferiore o uguale a N (esatta $\forall f \in \mathbb{P}_N$).
- ha **grado di precisione esattamente** N se e solo se ha grado di precisione almeno n ed esiste un polinomio di grado $N+1$ per cui non lo sia.

Non è quindi un concetto che ha a vedere con l'accuratezza (errore), ma bensì indica che la formula è esatta (errore nullo) per alcuni tipi di funzione.

Teorema 16

Una formula a n nodi è **interpolatoria** se e solo se ha grado di precisione almeno $n-1$

Definizione 38: Formula del rettangolo

Se $f \in C([a,b])$ con $-\infty < a < b < +\infty$ e $x_0 \in [a,b]$ ricaviamo $L_0(x) = 1$ (quindi il numero di nodi che utilizza la formula è $n = 1$) ed essendo:

$$\int_a^b L_0(x)dx = b - a$$

Deduciamo dalla regola del rettangolo

$$\int_a^b f(x) \approx \sum_{k=0}^{0} w_k f(x_k) = (b-a)f(x_0)$$

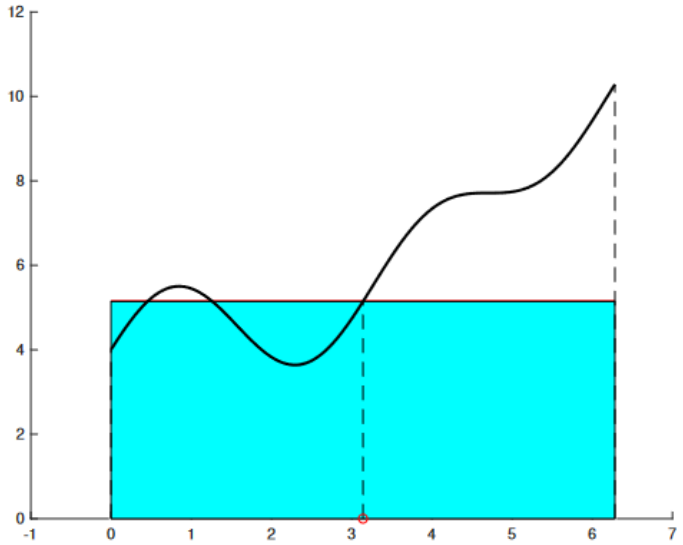
Per costruzione, se :

- f è un polinomio di grado 0
- p_0 è il polinomio che interpola il dato (x_0, y_0)

Per l'unicità del polinomio interpolatore abbiamo $f = p_0$ e quindi:

$$\int_a^b f(x)dx = \int_a^b p_0(x)dx = (b-a)f(x_0)$$

Di conseguenza il grado di precisione è **almeno** 0, quindi potrebbe essere anche maggiore di 0.



Formula del rettangolo con nodo $x_0 = \frac{a+b}{2}$ per il calcolo di $\int_0^{2\pi} 3x + \sin(2x) + \cos(x) + x dx$

Definizione 39: Formula del punto medio

Un caso particolare della formula del rettangolo la otteniamo se scegliamo come punto d'interpolazione $x_0 = \frac{a+b}{2}$ questa formula è detta **formula del punto medio** che denoteremo con I_0. Nonostante sfrutti solo un nodo è una formula accurata, in effetti se $f \in C^{(2)}(a,b)$ l'errore risulta:

$$E_0(f) = I(f) - I_0(f) = \frac{-(b-a)^3}{24} f^{(2)}(\xi) \quad \xi \in (a,b)$$

Come al solito l'errore è definito dalla differenza del valore vero ed il valore approssimato. Visto che un polinomio $q_1 \in \mathbb{P}_1$ ha derivata seconda nulla, deduciamo che $E_0(q_1) = 0$ ovvero che il grado di precisione è almeno 1. Per il polinomio $x^2 \in \mathbb{P}_2$ l'errore risulta uguale a:

$$\frac{-(b-a)^3}{24} f^{(2)}(\xi) \cdot 2 = \frac{-(b-a)^3}{12} \neq 0$$

E quindi il grado di precisione della formula del punto medio è esattamente 1.

Teorema 17: Teorema della media integrale generalizzata

Sia $f(x)$ e $g(x)$ due funzioni $\in C([a,b])$ con $[a,b]$ limitato e sia $g(x) \geq 0$ (o equivalentemente $g(x) \leq 0$) $\forall x \in [a,b]$. Allora $\xi \in (a,b)$ tale che:

$$\int_a^b f(x) \cdot g(x) dx = f(\xi) \cdot \int_a^b g(x) dx$$

Dimostrazione 20: Errore della formula del punto medio

Dato il punto medio $c = \frac{a+b}{2}$ definiamo l'errore in questo modo:

$$E = I(f) - I_0(f) = I(f) - (b-a)f(c) = \int_a^b f(x)dx - \int_a^b 1 \cdot f(c)dx = \int_a^b f(x) - f(c)dx$$

Considerando $x = c + (x - c)$ e sfruttando Taylor attorno ad x (ordine 2) ottengo:

$$f(x) = f(c) + f'(c)(x-c) + \frac{1}{2}f^{(2)}(\xi)(x-c)^2 \quad \xi \in (a,b)$$

E quindi

$$\int_a^b f(x) - f(c)dx = \int_a^b f'(c)(x-c)dx + \int_a^b \frac{1}{2}f^{(2)}(\xi)(x-c)^2dx$$

$$f'(c)\int_a^b (x-c)dx + \frac{1}{2}\int_a^b f^{(2)}(\xi)(x-c)^2dx \qquad f'(c)\int_a^b (x-c)dx = 0$$

Dato che $(x-c)^2 \geq 0 \ \forall x \in [a,b]$ per il teorema della media integrale generalizzata otteniamo:

$$\frac{1}{2}f^{(2)}(\xi) \cdot \int_a^b (x-c)^2dx = \frac{1}{2}f^{(2)}(\xi)\frac{(x-c)^3}{3}\Big|_a^b = \cdots = \frac{-(b-a)^3}{24}f^{(2)}(\xi)$$

Definizione 40: Formule di Newton-Cotes chiuse

Sia $[a,b]$ un intervallo compatto di $\mathbb{R}$ una formula:

$$I_n(f) = \sum_{i=1}^n w_i f(x_i) \approx \int_a^b f(x)dx$$

si dice di tipo **Newton-Cotes chiusa** se:

- I nodi sono equispaziati sono compresi gli estremi, cioè

$$x_i = a + \frac{(i-1)(b-a)}{(n-1)} \quad i = 1, \ldots, n$$

- I pesi sono:

$$w_i = \int_a^b L_i(x)dx \quad i = 1, \ldots, n \qquad L_i(x) = \prod_{j=1, j\neq i}^n \frac{(x - x_i)}{x_j - x_i}$$

Questa formula è interpolatoria ed ha grado di precisione almeno $n-1$. Gli estremi a, b sono **nodi di quadratura**, esistono formule di Newton-Cotes aperte in cui i nodi sono equidistanti ma gli estremi non sono compresi (es: formula del rettangolo).

Definizione 41: Formula del trapezio

Una delle formule di Newton-Cotes è la formula del trapezio. Se $f \in C([a,b])$ con $-\infty < a < b < +\infty$ e $x_0 = a$ e $x_1 = b$ ricaviamo:

$$L_0(x) = \frac{x - x_1}{x_0 - x_1} = \frac{b - x}{b - a} \qquad L_1(x) = \frac{x - x_0}{x_1 - x_0} = \frac{x - a}{b - a}$$

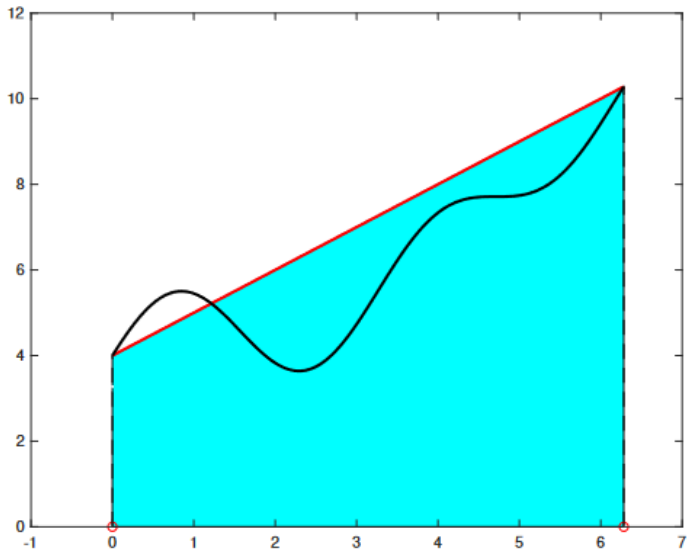
Calcoliamo i pesi:

$$\begin{aligned}\int_a^b L_0(x)dx &= \int_a^b \frac{b-x}{b-a}dx = \frac{1}{b-a}\int_a^b (b-x)dx \\ &= \frac{1}{b-a}\cdot\frac{-1}{2}((b-b)^2 - (b-a)^2) = \frac{b-a}{2}\end{aligned}$$

$$\begin{aligned}\int_a^b L_1(x)dx &= \int_a^b \frac{x-a}{b-a}dx = \frac{1}{b-a}\int_a^b (x-a)dx \\ &= \frac{1}{b-a}\cdot\frac{1}{2}((b-a)^2 - (a-a)^2) = \frac{b-a}{2}\end{aligned}$$

Deduciamo quindi la **formula del trapezio**

$$\int_a^b f(x) \approx I_T(f) = \frac{b-a}{2}\cdot(f(a) + f(b))$$



Formula del trapezio per il calcolo di $\int_0^{2\pi} 3x + \sin(2x) + \cos(x) + xdx$

Corollario 4

Nelle formule di Newton-Cotes si ha che la sommatoria dei pesi vale sempre:

$$\sum_{i=1}^{n} w_i = b - a$$

Definizione 42: Grado di precisione ed errore della formula del trapezio

Si dimostra che se $f \in C^2([a,b])$ allora l'errore compiuto è:

$$E_1(f) = I(f) - I_T(f) = \frac{-(b-a)^3}{12} f^{(2)}(\xi) \qquad \xi \in (a,b)$$

Incipit per la dimostrazione:

- Visto che un polinomio p_1 di grado ≤ 1 ha derivata seconda nulla deduciamo che $E_1(p_1) = 0$, ovvero che il grado di precisione è almeno 1.
- Per $f(x) = x^2$ cioè $f(x) \in \mathbb{P}_2$ l'errore risulta uguale a $\frac{a-(b-a)^3}{6}$ da cui abbiamo una dimostrazione alternativa che il grado di precisione è almeno 1

Dimostrazione 21: Calcolo della formula dell'errore della formula del trapezio

Sappiamo che l'errore come sempre equivale alla differenza tra il valore vero ed il valore ottenuto dalla formula:

$$E_1(f) = I(f) - I_T(f) = \int_a^b f(x)dx - \int_a^b p_1(x)dx = \int_a^b \underbrace{f(x) - p_1(x)}_{\text{Errore interpolazione } E_1(x)} \quad dx$$

Ma sappiamo che

$$E_1(x) = \frac{f^{(2)}(\xi(x)) \cdot \omega_2(x)}{2!} \qquad \omega_2(x) = (x-a)(x-b)$$

Ed allora:

$$\frac{1}{2}\int_a^b f^{(2)}(\xi_x) \cdot (x-a)(x-b)dx$$

Siccome $(x-a)(x-b) \geq 0$ posso applicare il teorema della media generalizzata e quindi ottengo:

$$\frac{1}{2}f^{(2)}(\xi) \cdot \int_a^b (x-a)(x-b)dx \quad \xi \in (a,b)$$

Svolgo un cambio di variabile $x = a + s \cdot h$ con $h = b-a$, $s \in [0,1]$, $dx = h \cdot ds$ ottengo:

$$\int_0^1 s \cdot h \cdot (s-1)h \cdot hds = h^3 \int_0^1 s^2 - sds = h^3 \left[\frac{s^3}{3} - \frac{s^2}{2}\right]_0^1 = h^3(\frac{1}{3} - \frac{1}{2}) = \frac{-h^3}{6}$$

Ma $h = b-a$ e quindi si conclude che:

$$\frac{1}{2}f^{(2)}(\xi) \cdot \int_a^b (x-a)(x-b)dx = \frac{-1}{12}f^{(2)}(\xi)(b-a)^3$$

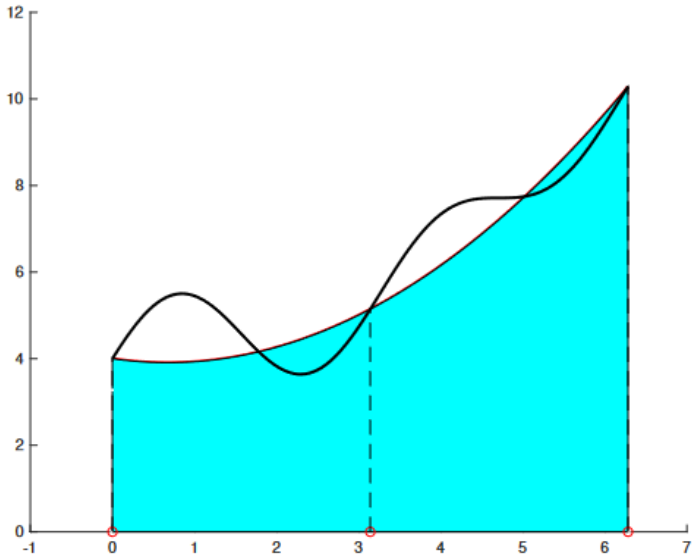
Definizione 43: Formula di Cavalieri-Simpson

Un'altra formula di Newton-Cotes chiusa è quella di **Cavalieri-Simpson** che otteniamo per $n = 3$. Qui i nodi di quadratura sono $x_0 = a$, $x_1 = (b+a)/2$ e $x_2 = b$. Integrando il polinomio p_2 di grado 2 che interpola i dati $(x_k, y_k)_{k=0,1,2}$ e calcolando i pesi otteniamo la **formula di Cavalieri- Simpson**:

$$I_{CS}(f) = \frac{b-a}{6} \cdot f(a) + \frac{2(b-a)}{3} \cdot f\left(\frac{a+b}{2}\right) + \frac{b-a}{6} \cdot f(b)$$

Oppure equivalentemente:

$$I_{CS}(f) = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]$$



Formula di Cavalieri-Simpson per il calcolo di $\int_0^{2\pi} 3x + \sin(2x) + \cos(x) + x dx$

Definizione 44: Errore e grado di precisione della formula di Cavalieri-Simpson

La formula è ottenuta integrando i polinomi di grado 2, quindi il grado di precisione è almeno 2. In realtà si dimostra che il grado di precisione è esattamente 3, infatti se $f \in C^4([a,b])$ si dimostra che l'errore compiuto è:

$$E_2(f) = I(f) - I_2(f) = \frac{-h^5 f^{(4)}(\xi)}{90} \qquad h = \frac{b-a}{2} \quad \xi \in (a,b)$$

Dato che un polinomio p_3 di grado ≤ 3 ha sicuramente derivata quarta nulla deduciamo che $E_2(p_3) = 0$ dimostrando che il grado di precisione è almeno 3. Inoltre per $f(x) = x^4 \in \mathbb{P}_4$ l'errore risulta essere uguale a:

$$\frac{-h^5}{90} 4! \neq 0$$

Dal quale si deduce che il grado di precisione per la formula di Cavalieri Simpson è esattamente 3

Esempio Consideriamo il seguente integrale semplice:

$$\int_1^3 \frac{1}{x} dx = \log_x \Big|_1^3 = \log_3 = 1.0986$$

Conosciamo il valore "vero". Vediamo cosa otteniamo con le formule:

- Formula dei trapezi, utilizzando gli estremi dell'intervallo $a = 1$ e $b = 3$ ed ottengo:

$$I_1 = I_T = \frac{b-a}{2}\left[f(a) + f(b)\right] = \left[1 + \frac{1}{3}\right] = 1.\overline{3}$$

L'errore che stiamo commettendo con la formula del trapezio è:

$$E_t = I - I_T = -0.235$$

- Formula Cavalieri-Simpson con $a = 1$, $\frac{a+b}{2} = 2$ e $b = 3$ ottengo:

$$I_2 = I_{CS} = \frac{b-a}{6}\left[f(a) + 4f(\frac{b-a}{2}) + f(b)\right] = \frac{1}{3}\left[1 + 4 \cdot \frac{1}{2} + \frac{1}{3}\right] = 1.\bar{1}$$

L'errore che stiamo commettendo con la formula di Cavalieri-Simpson è:

$$E_{CS} = I - I_{CS} = 1.0986 - 1.\bar{1} = -0.0125$$

5.1 Formule di Newton-Cotes composte

Vogliamo far tendere $b - a$ a 0, chiaramente non possiamo scegliere direttamente l'ampiezza dell'intervallo, ma possiamo suddividere l'intervallo originale in tanti subintervalli d'ampiezza "vicina" a 0. Suddividendo l'intervallo d'integrazione $[a, b]$ in N **subintervalli** che indichiamo con $T_k = [x_{k-1}, x_k]$ $k = 1 \dots N$ definiti dagli $N+1$ punti equidistanti $x_i = a + ih$ con $h = \frac{(b-a)}{N}$ e $i = 0 \dots N$ dove $a = x_0$ e $b = x_n$ possiamo calcolare l'integrale su $[a, b]$ come **somma degli integrali sui subintervalli**

$$\int_{x_0}^{x}{}_N f(x)dx = \int_a^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \cdots + \int_{x_{N-1}}^{b} f(x)dx$$

Su ciascun intervallo $[x_{k-1}, x_k]$ per $k = 1, \dots, N$ approssimiamo l'integrale della f mediante la formula di quadratura semplice (noi vedremo Trapezi o Simpson).Vale che:

- Per la formula del Trapezio composta utilizzeremo $N + 1$ punti
- Per la formula di Cavalieri-Simpson composta $2N + 1$ punti
- Per la formula del Rettangolo composta N punti

Definizione 45: Formula dei Trapezi composta ed errore

La **Formula dei Trapezi composta** è definita da:

$$I_T^{(c)}(f, a, b, N) = \frac{b-a}{N}\left[\frac{f(x_0)}{2} + f(x_1) + \cdots + f(x_{N-1}) + \frac{f(x_n)}{2}\right]$$

Ottenuta semplicemente concatenando l'applicazione della formula del trapezio ai singoli subintervalli. Si dimostra che l'errore compiuto per un certo $\xi \in (a, b)$ è:

$$E_1^{(c)} = I(f) - I_T^{(c)}(f, a, b, N) = \frac{-(b-a)}{12}h^2 f^{(2)}(\xi) \quad h = \frac{(b-a)}{N}$$

Il grado di precisione è 1 come per la formula del trapezio, questa formula è indicata per integrare funzioni periodiche con derivate anch'esse periodiche (tipo le trigonometriche).

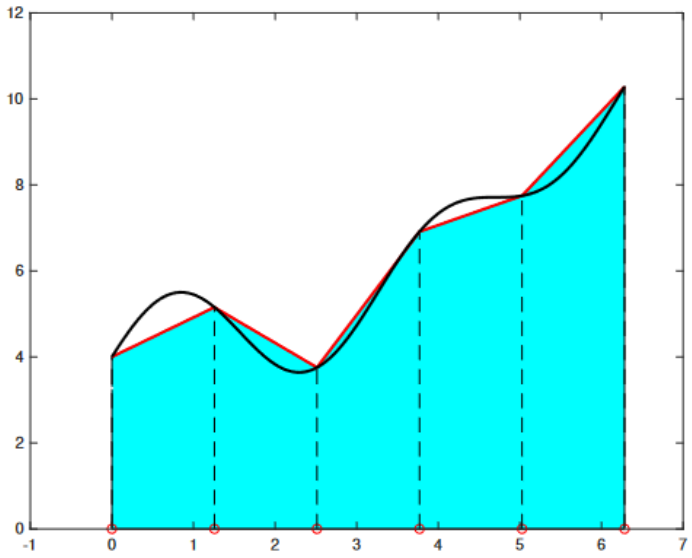
Dimostrazione 22: Errore della formula composta dei Trapezi

Quando svolgiamo una formula composta de facto stiamo applicando la formula semplice N volte. L'errore della formula composta allora sarà:

$$E_T^C(f) = I - I_T^C = \sum_{i=1}^{N} \frac{-f^{(2)}(\xi_i)h^3}{12} = \frac{-h^3}{12}\sum_{i=1}^{N} f^{(2)}(\xi_i) \qquad h = \frac{b-a}{N}$$

$$\frac{-(b-a)^3}{12N^2} \cdot \underbrace{\sum_{i=1}^{N} \frac{1}{N} \cdot f^{(2)}(\xi_i)}_{f^{(2)}(\xi) \quad \xi \in [a,b]} = \frac{-(b-a)^3 f^{(2)}(\xi)}{12N^2} = \frac{-(b-a)}{12} h^2 f^{(2)}(\xi)$$

Dove $\sum_{i=1}^{N} \frac{1}{N} \cdot f^{(2)}(\xi_i) = f^{(2)}(\xi)$ segue dal teorema dei valori intermedi



Formula dei trapezi composta per il calcolo di $\int_0^{2\pi} 3x + \sin(2x) + \cos(x) + x dx$

Definizione 46: Formula composta di Cavalieri-Simpson ed errore

a **formula composta di Cavalieri-Simpson** è definita da:

$$I_{CS}^{(c)}(f,a,b,N) = \frac{h}{6}\left[f(x_0) + 2\sum_{r=1}^{N-1} f(x_{2r}) + 4\sum_{s=0}^{N-1} f(x_{2s+1}) + (x_{2N})\right]$$

Se devo calcolare il valore dell'integrale con Cavalieri-Simpson composta devo calcolare:

$$\begin{aligned} I_{CS}^{(c)} = &\frac{h}{6}\left[f(x_0) + 4f(\frac{x_0+x_1}{2}] + f(x_1)\right] + \\ &+ \frac{h}{6}\left[f(x_1) + 4f(\frac{x_1+x_2}{2}] + f(x_2)\right] + \\ &+ \frac{h}{6}\left[f(x_2) + 4f(\frac{x_2+x_3}{2}] + f(x_3)\right] + \\ &\vdots \\ &\frac{h}{6}\left[f(x_{N-1}) + 4f(\frac{x_{N-1}+x_N}{2}] + f(x_N)\right] \end{aligned}$$

Si mostra che l'errore compiuto per un certo $\xi \in (a,b)$:

$$E_2^{(c)} = I(f) - I_{CS}^{(c)}(f,a,b,N) = \frac{-(b-a)}{180}\left(\frac{h}{2}\right)^4 f^{(4)}(\xi) = \frac{-(b-a)}{2880} f^{(4)}(\xi) h^4$$

Il grado di precisione è 3, come per la formula standard.

Definizione 47: Ordine di accuratezza

Una formula di quadratura composta $I_n^{(c)}(f,a,b,N)$ ha **ordine di accuratezza p** se l'errore compiuto tende a zero per $h \to 0$ come h^p ovvero:

$$|E_n^{(c)}(f)| = |I(f) - I_n^{(c)}(f,a,b,N)| = O(h^p) \quad \text{(O grande)}$$

L'ordine di accuratezza della formula dei trapezi composta è 2 (c'è un h^2)

$$E_1^{(c)}(f) = I(f) - I_T^{(c)}(f,a,b,N) = \frac{-(b-a)}{12} f^{(2)(\xi)} h^2$$

L'ordine di accuratezza della formula di Cavalieri-Simpson composta è 4 (c'è un h^4)

$$E_2^{(c)}(f) = I(f) - I_{CS}^{(c)}(f,a,b,N) = \frac{-(b-a)}{180} f^{(4)}(\xi)\left(\frac{h}{2}\right)^4 = \frac{-(b-a)}{2880} f^{(4)}(\xi) h^4$$

Esempio Applichiamo le formule composte per approssimare l'integrale definito così:

$$I = \int_0^{\pi} exp(x)cos(x)dx = -(exp(\pi)+1)/2$$

Con $N = 2^i$ con $i = 1, 2, \ldots, 9$ suddivisioni dell'intervallo $[0, \pi]$ otteniamo:

N	$E_0^{(c)}(f)$	$E_1^{(c)}(f)$	$E_2^{(c)}(f)$	$\#_N^{\mathsf{R}}$	$\#_N^{\mathsf{T}}$	$\#_N^{\mathsf{CS}}$
1	$1.2e+01$	$2.3e+01$	$4.8e-01$	1	2	3
2	$2.8e+00$	$5.3e+00$	$8.5e-02$	2	3	5
4	$6.4e-01$	$1.3e+00$	$6.1e-03$	4	5	9
8	$1.6e-01$	$3.1e-01$	$3.9e-04$	8	9	17
16	$3.9e-02$	$7.8e-02$	$2.5e-05$	16	17	33
32	$9.7e-03$	$1.9e-02$	$1.6e-06$	32	33	65
64	$2.4e-03$	$4.8e-03$	$9.7e-08$	64	65	129
128	$6.1e-04$	$1.2e-03$	$6.1e-09$	128	129	257
256	$1.5e-04$	$3.0e-04$	$3.8e-10$	256	257	513
512	$3.8e-05$	$7.6e-05$	$2.4e-11$	512	513	1025

Definizione 48: Rapporto tra errori

Calcoliamo il rapporto tra 2 errori successivi per ogni formula, ovvero, se $(E_n^{(c)}(f))_N$ con $n = 0, 1, 2$ è l'errore compiuto dalla formula $I_n^{(c)}$ relativamente al calcolo di $\int_a^b f(x)dx$ utilizzando N suddivisioni mostriamo:

$$r_n^{(c)}(f))_N = \frac{(E_n^{(c)}(f))_N}{(E_n^{(c)}(f))_{2N}}$$

Per $n = 0, 1, 2$ ossia per formule composte del punto medio, dei trapezi e di Cavalieri-Simpson.

N	$(r_0^{(c)}(f))_N$	$(r_1^{(c)}(f))_N$	$(r_3^{(c)}(f))_N$
1	4.33	4.27	5.59
2	4.34	4.20	13.92
4	4.10	4.06	15.54
8	4.03	4.02	15.89
16	4.01	4.00	15.97
32	4.00	4.00	15.99
64	4.00	4.00	16.00
128	4.00	4.00	16.00
256	4.00	4.00	16.00

Dimostrazione 23: Rapporto tra errori consecutivi

- Formula dei Trapezi, l'errore nei due casi vale:

$$N : \frac{-(b-a)^3 f^{(2)}(\xi_N)}{12N^2} \qquad 2N : \frac{-(b-a)^3 f^{(2)}(\xi_{2N})}{12(2N)^2}$$

Calcolando il rapporto:

$$\frac{E_N}{E_{2N}} = \frac{\frac{-(b-a)^3 f^{(2)}(\xi_N)}{12N^2}}{\frac{-(b-a)^3 f^{(2)}(\xi_{2N})}{12(2N)^2}} = \frac{f^{(2)}(\xi_N)}{f^{(2)}(\xi_{2N})} \cdot 4$$

Quindi tutto dipende dalla derivata della funzione.

- Formula di Cavalieri-Simpson, l'errore nei due casi vale:

$$N : \frac{-(b-a)^5 f^{(4)}(\xi_N)}{2880N^4} \qquad 2N : \frac{-(b-a)^5 f^{(4)}(\xi_{2N})}{2880(2N)^4}$$

$$\frac{\frac{-(b-a)^5 f^{(4)}(\xi_N)}{2880N^4}}{\frac{-(b-a)^5 f^{(4)}(\xi_{2N})}{2880(2N)^4}} = \frac{f^{(4)}(\xi_N)}{f^{(4)}(\xi_{2N})} \cdot 16$$

Definizione 49: Estrapolazione di Richardson

Procedura che consiste nel combinare opportunamente 2 approssimazione di una certa quantità in modo da trovare una terza approssimazione più accurata. Nell'integrazione numerica, la procedura si basa sui rapporti tra errori ottenuti raddoppiando il numero N di suddivisioni dell'intervallo d'integrazione (che sarà 4 per Trapezi e 16 per Cavalieri-Simpson sotto certe ipotesi sulle derivate della funzione integranda.

Per Trapezi, assumendo che $E_N = 4E_{2N}$:

$$I = I_N + E_N = I_N - 4E_{2N} \qquad I = I_{2N} + E_{2N}$$

Da cui sottraendo otteniamo:

$$0 = I_N - I_{2N} + 3E_{2N} \Rightarrow E_{2N} = \frac{I_{2N} - I_N}{3}$$

Ottengo finalmente:

$$I = I_{2N} + E_{2N} = I_{2N} + \frac{I_{2N} - I_N}{3} = \frac{4I_{2N} - I_N}{3}$$

Per Cavalieri-Simspon ottengo che

$$E_n = 16E_{2N} \qquad I_R = \frac{16I_{2N} - I_N}{15}$$

Esercizi proposti

- Sia data:
$$\int_1^3 \frac{1}{x}dx \qquad N = 2 \qquad I_{vero} = 1.0986$$
  - Applichiamo la formula del trapezio con $N = 2$, vale che $h = \frac{3-1}{2} = 1$ quindi:
  $$\int_1^3 \frac{1}{x}dx \approx \frac{h}{2}\Big[f(1)+f(2)\Big]+\frac{h}{2}\Big[f(2)+f(3)\Big] = \frac{1}{2}\Big[f(1)+2f(2)+f(3)\Big] = \frac{1}{2}\Big[1+1+\frac{1}{3}\Big] = \frac{7}{6}$$
  L'errore utilizzando questa formula è:
  $$E_T^c = 1.0986 - \frac{7}{6} = -0.068$$
  - Applichiamo la formula di Cavalieri-Simpson, ho bisogno dei punti medi dei due intervalli, quindi 1.5 e 2.5 ora applicando ho:
  $$I_{cs}^c = \frac{1}{6}\Big[f(1)+4f(1.5)+2f(2)+4f(2.5)+f(3)\Big] = \frac{1}{6}\Big[1+4\frac{1}{1.5}+2\frac{1}{2}+4\frac{1}{2.5}+\frac{1}{3}\Big] = 1.1$$
  L'errore che stiamo commettendo con questa formula è quindi:
  $$E_{cs}^c = 1.0986 - 1.1 = -0.00139 = -1.39 \cdot 10^{-3}$$

- Sia data:
$$\int_1^2 \log_x dx = 0.386294$$
Determinare il numero di intervalli necessari per ottenere un errore $|E_N| < 10^{-3}$ per la formula dei trapezi e di Cavalieri-Simpson composta.
$$E_{cs}^c = \frac{-(b-a)^5 \cdot f^{(4)}(\xi)}{2880 \cdot N^4} =$$
Per poter maggiorare l'errore lavoriamo con il valore assoluto, quindi:
$$|E_{cs}^c| \leq \frac{-(b-a)^5 \cdot \max_{x\in[1,2]} |f^{(4)}(x)|}{2880 \cdot N^4} < 10^{-3}$$
$$f(x) = \log(x) \quad f'(x) = \frac{1}{x} \quad f^{(2)}(x) = \frac{-1}{x^2} \quad f^{(3)}(x) = \frac{2}{x^3} \quad f^{(4)}(x) = \frac{-6}{x^4}$$
Quindi:
$$\max_{x\in[1,2]} |f^{(4)}(x)| = \max_{x\in[1,2]} |\frac{6}{x^4}| = 6$$
$$\frac{6}{2880 \cdot N^4} < 10^{-3} \qquad N^4 > \frac{6000}{2880} \to N > \sqrt{2.1} \approx 1.2$$
Siccome N deve essere un intero naturale prendiamo la parte intera superiore e quindi $N = \lceil 1.2 \rceil = 2$. Calcoliamo il valore con la formula di Cavalieri-Simpson utilizzando due suddivisioni e quindi 5 punti, $h = \frac{2-1}{2} = \frac{1}{2}$:
$$I_{cs}^c = \frac{1}{12}\Big[f(1) + 4f(1.25) + 2f(1.5) + 4f(1.75) + f(2)\Big] = 0.386260$$
$$E_{cs}^c = I_{vero} - I_{cs}^c = 0.386294 - 0.386260 = 3.4 \cdot 10^{-5} < 10^{-3}$$

- Si consideri la seguente formula di quadratura per l'approssimazione di questo integrale generico:
$$\int_{-2}^{2} f(x)dx \approx \alpha_1 f(0) + \frac{2}{3}\alpha_2 f(-c) + f(c)$$
Dove α_i sono i pesi. Questa è una formula di quadratura perché rientra nella definizione di approssimazione di integrale definito come somma finita di pesi $\times$ valori funzione. Determinare α_1, α_2 e c sapendo che $c > 0$ in modo che la formula data abbia grado di precisione almeno due. Ricordiamo che avere grado di precisione almeno N implica che tutti i polinomi fino al grado N devono essere integrati dalla formula. Devo quindi imporre che sia integrabile una base dei polinomi di grado 2, come base di $\mathbb{P}_2$ prendiamo $\{1, x, x^2\}$. Quindi:
$$\int_{-2}^{2} f(x)dx = \alpha_1 f(0) + \frac{2}{3}\alpha_2 f(-c) + f(c)$$
Imponiamo l'uguaglianza per $f = 1$, $f(x) = x$ e $f(x) = x^2$:
$$\begin{aligned}
f(x) = 1 \quad & \int_{-2}^{2} 1dx = 4 = \alpha_1 1 + \frac{2}{3}\alpha_2 \cdot 1 + 1 \\
f(x) = x \quad & \int_{-2}^{2} xdx = \frac{x^2}{2}\Big|_{-2}^{2} = 0 = \alpha_1 \cdot 0 + \frac{2}{3}\alpha_2(-c) + c \\
f(x) = x^2 \quad & \int_{-2}^{2} x^2 dx = \frac{x^3}{3}\Big|_{-2}^{2} = \frac{16}{3} = \alpha_1 \cdot 0 + \frac{2}{3}\alpha_2 c^2 + c^2
\end{aligned}$$
Dalla seconda equazione ricaviamo:
$$(1 - \frac{2}{3}\alpha_2)c = 0 \rightarrow c = 0 \vee \alpha_2 = \frac{3}{2}$$
Siccome $c > 0$ vale che $\alpha_2 = \frac{3}{2}$. Dalla terza equazione ricaviamo:
$$\frac{16}{3} = \frac{2}{3}\frac{3}{2}c^2 + c^2 \rightarrow \frac{16}{3} = 2c^2 \rightarrow c = \sqrt{\frac{8}{3}}$$

- Si consideri la formula di quadratura:
$$\int_{-1}^{1} f(x)dx \approx \alpha_1 f(-1) + \alpha_2 f(0) + \alpha_3 f(\frac{1}{2})$$
Determinare i pesi α_i con $i = 1, 2, 3$ in modo che la formula di quadratura abbia grado di precisione almeno 2.

Esercizio Sia nota la funzione
$$f(x) = x^4 - 5x + 2\sqrt{x}$$
nei cinque nodi $x_i = 1 + ih$ con $h = 0.2$ e $i = 0, 1, 2, 3, 4$.

- Utilizzando i valori delle funzione nei punti dati si approssimi il valore dell'integrale:
$$I = \int_{x_0}^{x_4} f(x)dx \qquad I_{vero} = -0.134259$$

mediante la formula di Cavalieri-Simpson composta. Chiamata Q_1 questa approssimazione si dia una maggiorazione dell'errore commesso e la si confronti con l'errore vero. Dovendo utilizzare C-S sui 5 nodi ho solo un modo di procedere, nota che nella formula $h = 0.4$ (è un altro h diverso dalla consegna ovviamente)

$$Q_1 = I_{cs}^c = \frac{0.4}{6}\Big[f(1) + 4f(1.2) + 2f(1.4) + 4f(1.6) + f(1.8)\Big] = -0.1340937$$

$$E_{cs} = |I_{vero} - Q_1| = 1.65 \cdot 10^{-4}$$

- Utilizzando i valori in x_0, x_2, x_4 utilizzare C-S per ottenere una seconda approssimazione dell'integrale che chiamiamo Q_2.

$$Q_2 = \frac{0.8}{6}\Big[f(1) + 4f(1.4) + f(1.8)\Big] = -0.131599$$

- A partire da Q_1 e Q_2 si fornisca una terza Q_3 e più accurata approssimazione dell'integrale mediante l'estrapolazione di Richardson, calcolare infine l'approssimazione. Sappiamo che per la formula di Cavalieri-Simpson l'estrapolazione di Richardson ci dà:

$$I_r = \frac{16I_{2N} - I_N}{15} = \frac{16Q_1 - Q_2}{15} = -0.134260 = Q_3 \qquad |E_3| = |Q_3 - I_{vero}| = 1.005 \cdot 10^{-6}$$

6 Algebra lineare numerica

Di tutta l'algebra lineare numerica che si occupa principalmente di risolvere sistemi lineari e calcolare autovalori ci interessiamo alla parte che studia **le fattorizzazioni di matrici** (rientra nel primo gruppo). Noi tratteremo due sistemi d'equazione

- Sistemi dove la matrice è **quadrata** e quindi la soluzione è unica

$$Ax = b \qquad A \in \mathbb{R}^{n \times n} \qquad x, b \in \mathbb{R}^n$$

- Sistemi dove la matrice è **rettangolare**

$$Ax = b \qquad A \in \mathbb{R}^{m \times n} \qquad x \in \mathbb{R}^n \quad b \in \mathbb{R}^m$$

In realtà le fattorizzazioni che vedremo non servono solo per risolvere sistemi, ma sono alla base di moltissime scienze (data mining, information retrival ...).

6.1 Richiami di algebra lineare

In questa parte ripassiamo i concetti di algebra lineare indispensabili per la fattorizzazione di matrici.

Definizione 50: Matrice quadrata

Diciamo che una matrice $A \in \mathbb{R}^{n \times n}$ è quadrata se ha lo stesso numero di righe e di colonne. La matrice quadrata per antonomasia è la matrice identità.

$$\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

Una matrice è invertibile se e solo se

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

La **trasposta** di una matrice è così definita:

$$A^T \in \mathbb{R}^{n \times n} \qquad A \in \mathbb{R}^{m \times n} \to A^T \in \mathbb{R}^{n \times m}$$

Una matrice è simmetrica se

$$A^T ij = Aij \qquad A = A^T$$

Proprietà della matrice trasposta :

- $(A^T)^T = \mathrm{A}$
- $(A + B)^T = A^T + A^B$
- $(A \cdot B)^T = B^T \cdot A^T$
- Se A^{-1} allora $(A^{-1})^T = (A^T)^{-1}$

Definizione 51: Matrice diagonale

Una matrice si dice diagonale quando tutti gli elementi fuori dalla diagonale principale sono nulli.

$$\begin{bmatrix} 4 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 2 \end{bmatrix}$$

Ad esempio questa è diagonale. Formalmente $d_{ij} = 0 \quad \forall i \neq j$

Definizione 52: Matrice triangolare

Ne esistono di due tipi:

- Matrice triangolare superiore, hanno tutti 0 sotto la diagonale principale, cioè $a_{ij} = 0 \quad \forall i > j$

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

- Matrice triangolare inferiore, hanno tutti 0 sopra la diagonale principale, cioè $a_{ij} = 0 \quad \forall i < j$

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

Definizione 53: Matrici tridiagonali

Sono matrici che hanno tutte le componenti uguali a 0 eccetto le "3 diagonali principali" (diagonale principale, quella sotto e quella sopra).

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

Definizione 54: Autovalori e Autovettori

Considerando una matrice quadrata $A \in \mathbb{R}^{n\times n}$ definiamo

$$\lambda \in \mathbb{C} \text{ autovalore di A se } \exists x \in \mathbb{R}^n \backslash \{0\} \ (x \neq 0) \text{ tale che } Ax = \lambda x$$

x si dice **autovettore** di A relativo a λ, nota che gli autovettori non possono essere nulli.

Polinomio caratteristico : $\det(A - \lambda I)$ di grado pari al rango della matrice (n).

Equazione caratteristica : $\det(A - \lambda I) = 0$.

Definizione 55: Spettro

Definiamo **spettro** $\lambda \in \sigma(A)$ l'insieme di tutti gli autovalori di una matrice. Vale che per ogni $\lambda \in \sigma(A)$ ho $\det(A - \lambda I) = 0$. Il determinante di una matrice è il prodotto di tutti i suoi autovalori, mentre la traccia di una matrice è la somma di tutti i suoi valori sulla diagonale principale:

$$det(A) = \prod_{i=1}^{n} \lambda_i \qquad tr(A) = \sum_{i=1}^{n} a_{ii} = \sum_{i=1}^{n} \lambda_i$$

. Se $A - \lambda I$ non è invertibile per $\lambda \in \sigma(A)$ ciò implica che il suo rango non è completo e quindi dal teorema delle dimensioni:

$$n = \dim \underbrace{(ker(A - \lambda I)}_{>0} + \underbrace{Rg(A - \lambda I)}_{<n}$$

Gli autovettori di una matrice stanno in $E_\lambda = \ker(A - \lambda I)$

Definizione 56: Raggio spettrale di una matrice

Il massimo dei moduli degli autovalori

$$\rho(A) = \max_{i\in\{1,\dots,n\}} |\lambda_i| \qquad \lambda_i \in \sigma(A)$$

(Il fatto che chiediamo che i moduli degli autovalori dipende dall'alta probabilità che questi assumano valori complessi)

Definizione 57: Prodotto scalare

Dato uno spazio vettoriale qualsiasi, a noi interessa $W \in \mathbb{R}^n$, la funzione $< \cdot, \cdot >$: $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ che verifica le 4 proprietà seguenti:

1. $< x, x > \geq 0$ e $< x, x >= 0 \leftrightarrow x = 0$
2. $< x, y >=< y, x > \quad \forall x, y \in \mathbb{R}^n$
3. $< x, y + z >=< x, y > + < x, z > \quad \forall x, y, z \in \mathbb{R}^n$
4. $< \alpha x, y >= \alpha < x, y > \quad \alpha \in \mathbb{R}$

allora è detto **prodotto scalare** su $\mathbb{R}^n$. Definiamo **prodotto scalare canonico**:

$$\mathbb{R}^n :< x, y >= x^T y = \sum_{i=1}^{n} x_i y_i$$

Due vettori sono ortogonali se $< x, y >= 0$.

Definizione 58: Norma

Dato uno spazio vettoriale W generico la funzione $|| \cdot || : \mathbb{R}^n \to \mathbb{R}$ che verifica le 3 proprietà seguenti:

1. $||x|| \geq 0$ e $||x|| = 0 \leftrightarrow x = 0 \in \mathbb{R}^n$ (vettore nullo)
2. $||\alpha x|| = |\alpha| \cdot ||x|| \quad \forall x \in \mathbb{R}^n \quad \forall \alpha \in \mathbb{R}$
3. $||x + y|| \leq ||x|| + ||y|| \quad \forall x, y \in \mathbb{R}^n$

è una norma vettoriale su $\mathbb{R}^n$. La norma dedotta da un prodotto scalare è:

$$||x|| = \sqrt{< x, x >}$$

La norma dedotta dal prodotto scalare canonico su $\mathbb{R}^n$ è la **norma euclidea** o anche detta norma 2:

$$||x||_2 = \sqrt{x^T x} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

Principali norme di vettore:

- Norma 1, $||x||$ definita come la somma delle componenti in valore assoluto:

$$||x|| = \sum_{i=1}^{n} |x_i|$$

- Norma 2, $||x||_2$ definita come la radice della somma delle componenti in valore assoluto elevato al quadrato

$$||x||_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2}$$

- Norma massima, $||x||_\infty$ definita:

$$\max_{i\in\{1,\dots,n\}} |x_i| = \lim_{p\to\infty} ||x||_p$$

Sono tutte e 3 facenti parte della famiglia delle norme p:

$$||x||_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{(1/p)} \qquad p \in [1, +\infty)$$

Definizione 59: Norma di matrice

Definiamo **norma di matrice** su $\mathbb{R}^{n\times n}$ (non è necessario che sia quadrato) come la funzione $||\cdot|| : \mathbb{R}^{n\times n} \to \mathbb{R}$ che verifica:

- Proprietà di positività: $||A|| \geq 0$ e $||A|| = 0$ se e solo se $A = 0$ (matrice nulla)
- Proprietà di omogeneità: $||\alpha A|| = |\alpha| \cdot ||A|| \quad \forall \alpha \in \mathbb{R}\ \forall A \in \mathbb{R}^{n\times n}$
- Diseguaglianza triangolare: $||A + B|| \leq ||A|| + ||B|| \quad \forall A, B \in \mathbb{R}^{n\times n}$
- Proprietà di molteplicità: $||A \cdot B|| \leq ||A|| \cdot ||B|| \quad \forall A, B \in \mathbb{R}^{n\times n}$

è una norma su $\mathbb{R}^{n\times n}$

Definizione 60: Compatibilità con le norme vettoriali

Una norma di matrice $||\cdot||_M$ è **compatibile** con una norma vettoriale $||\cdot||_V$ se :

$$||A \cdot x||_V \leq ||A||_M \cdot ||x||_v \quad \forall A \in \mathbb{R}^{n\times n}\ \forall x \in \mathbb{R}^n$$

Definizione 61: Norma matriciale indotta da una norma vettoriale

La funzione $||\cdot|| : \mathbb{R}^{n\times n} \to \mathbb{R}$ data da:

$$||A|| := \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{||Ax||}{||x||}$$

è una norma di matrice su $\mathbb{R}^{n\times n}$ chiamata **norma indotta dalla norma vettoriale** (anche detta norma d'operatore)

Principali norme indotte:

- Norma 1: $||A||_1$, corrisponde alla norma indotta da $||x||_1$

$$\max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|$$

- Norma ∞: $||A||_\infty$, corrisponde alla norma indotta da $||x||_\infty$

$$\max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|$$

- Norma 2: $||A||_2$, corrisponde alla norma indotta da $||x||_2$

$$||A||_2 = \sqrt{\rho(A^T A)} = \sqrt{\rho(AA^T)}$$

Esempio Consideriamo questa matrice:

$$A = \begin{pmatrix} 4 & -1 & 1 \\ 1 & 3 & -1 \\ 0 & 1 & 1 \end{pmatrix} \quad ||A||_1 = \max\{5,5,3\} = 5 \quad ||A||_\infty \max\{6,5,2\} = 6$$

$$||A||_2 = \sqrt{\rho(A^T A)} = \sqrt{18} = 4.2426$$

Definizione 62: Norma di Frobenius

Definiamo **norma di Frobenius** $||A||_F$ in questo modo:

$$||A||_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2} = \sqrt{tr(A^T A)}$$

In pratica è la norma due di una matrice le quali righe sono tutte concatenate in modo da ottenere una matrice con tante colonne ed una sola riga.

La norma di Frobenius **non è una norma indotta da nessuna norma vettoriale**, ciò è un problema perché le norme indotta hanno due proprietà:

1. Per ogni norma indotta si ha che ogni norma indotta è compatibile con la norma da cui deriva.

   $$||Ax|| \leq ||A|| \cdot ||x||$$

   Si dimostra notando che:

   $$\frac{||Ax||}{||x||} \leq \sup_{z \in \mathbb{R}^n \setminus \{0\}} \frac{||Az||}{||z||} = ||A||$$

2. Per ogni norma indotta

   $$||I|| = \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{||Ix||}{||x||} = \sup \frac{||x||}{||x||} = 1$$

   $$||I_n||_F = \sqrt{\sum_{i=1}^{n} 1} = \sqrt{n} \neq 1$$

Per ogni norma matriciale indotta si ha

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

Dimostrazione 24

Prendiamo $\lambda \in \sigma(A)$ allora $\exists x \in \mathbb{R}^n \backslash \{0\}$ tale che $Ax = \lambda x$

$$||\lambda x|| = |\lambda| \cdot ||x|| = ||Ax|| \leq ||A|| \cdot ||x||$$

Posso dividere per $||x|| \neq 0$ ed ottengo:

$$|\lambda| \leq ||A||$$

Definizione 63: Matrice ortogonale

Definiamo **matrice ortogonale** $Q \in \mathbb{R}^{n \times n}$ (in campo complesso le equivalenti sono le matrice unitarie) se le sue colonne $\{q^{(1)}, q^{(2)}, \ldots, q^{(n)}\}$ sono una **base ortonormale di** $\mathbb{R}^n$ (ortogonali e di norma 1).

$$Q = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad Q^T = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Alcune proprietà che mi permettono di vedere facilmente se una matrice è ortogonale o meno:

- Q è ortogonale se e soltanto se $Q^T Q = I_N$
- Se Q è ortogonale allora automaticamente è **invertibile**, cioè esiste $Q^{-1} = Q^T$
- Q è ortogonale se e solo se $QQ^T = I_N$
- Q è una matrice ortogonale se $\forall x, y \in \mathbb{R}^n$ si ha che il prodotto scalare

  $$< Qx, Qy >=< x, y >$$

  Piccola dimostrazione banale:

  $$< Qx, Qy >= (Qx)^T Qy = x^T \underbrace{Q^T Q}_{I_n} y = x^T y =< x, y >$$

  Quindi sono trasformazioni che **non alterano la lunghezza dei vettori**:

  $$||Qx||_2 = \sqrt{< Qx, Qy >} = \sqrt{< x, x >} = ||x||_2$$

Definizione 64: Determinante di una matrice ortogonale

Il **determinante** di una matrice ortogonale $Q \in \mathbb{R}^{n \times n}$ assume valore $\det(Q) = \pm 1$, in generale vale che:

$$|\det(Q)| = 1$$

Dimostrazione 25

$$I = Q^T Q \quad \det(I) = 1 = \det(Q^T Q) = \det(Q)\det(Q^T) = [\det(Q)]^2$$

Definizione 65: Autovalore di una matrice ortogonale

Ogni autovalore di una matrice ortogonale $\lambda \in \sigma(Q)$ assume valore $|\lambda| = 1$ (modulo, quindi possono essere reali o complessi).

Dimostrazione 26

Dal fatto che le matrici ortogonali conservano la norma 2 ho che se $\lambda \in \sigma(Q)$ allora esisterà $x \in \mathbb{R}^n \backslash \{0\}$ tale che $Qx = \lambda x$

$$\begin{cases} ||Qx||_2 = ||\lambda x|| = |\lambda| \cdot ||x||_2 \\ ||Qx||_2 = ||x||_2 \end{cases} \qquad |\lambda| = 1$$

Definizione 66: Matrice simmetrica

Una matrice si dice **simmetrica** quando **coincide con la sua trasposta**, (quindi già questo ci dice che per essere simmetrica una matrice è quadrata). Se $A = A^T$ allora:

- Se B è simmetrica allora $A + B$ è simmetrica
- αA è ancora simmetrica
- Se A è simmetrica non singolare (regolare) allora anche A^{-1} è simmetrica
- $A, B \in \mathbb{R}^{n \times n}$ simmetriche allora $A \cdot B$ è simmetrica se e soltanto se A e B commutano
$$(AB)^T = B^T A^T = BA = AB$$
- A simmetrica è sempre **diagonalizzabile**
- A simmetrica è sempre **diagonalizzabile** con una matrice di trasformazione ortogonale.
- Per un A simmetrica esiste sempre una base **ortonormale**

Definizione 67: Matrici simili

Due matrici A e B entrambe quadrate $\in \mathbb{R}^{n \times n}$ sono simili se esiste una matrice invertibile V tale per cui:

$$AV = VB \quad \equiv \quad A = VBV^{-1} \quad \equiv \quad B = V^{-1}AV$$

Se A e B sono simili allora hanno lo stesso **spettro**, $\sigma(A) = \sigma(B)$ (il viceversa non è vero, cioè spettro uguale non implica similitudine). (La dimostrazione si basa sul dimostrare che hanno lo stesso polinomio caratteristico).

Definizione 68: Matrice diagonalizzabile

Diciamo che una matrice A è **diagonalizzabile** se è simile ad una matrice diagonale, cioè esiste una V invertibile tale che:

$$AV = VD \quad A = VDV^{-1}$$

Se la matrice è quadrata ed è diagonalizzabile allora esiste sempre una base di $\mathbb{R}^n$ formata da autovettori di A. Questi autovettori compongono le colonne di V, vale che:

$$AV = V \cdot D \quad \text{con } D = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

Per A simmetriche vale che:

$$A = QDQ^T \quad \equiv \quad D = A^T AQ$$

Esempio

$$A = \begin{pmatrix} 1 & 9 \\ 4 & 1 \end{pmatrix} \quad \det(A - \lambda I) = \begin{vmatrix} 1-\lambda & 9 \\ 4 & 1-\lambda \end{vmatrix} = (\lambda - 7)(\lambda + 5)$$

$$\sigma(A) = \{7, 5\} \quad \dim E\lambda = 7 \quad \dim \ker(A - 7I) = 1 \quad \dim E\lambda = 5 \quad \dim \ker(A + 5I) = 1$$

Esempio Dato che la matrice è triangolare superiore non dobbiamo calcolare nessun polinomio caratteristico

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \sigma(A) = \{1\} \quad P_A(\lambda) = (\lambda - 1)^2 \quad (2 \text{ è la molteplicità algebrica})$$

$$A - I = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad Rg(A - I) = 1 \Rightarrow \ \dim \ker(A - I) = 1$$

$$\lambda \in \sigma(A) \Rightarrow \lambda^n \in \sigma^n(A^n)$$

$$\lambda \in \sigma(A) \Rightarrow \exists x \neq 0 | Ax = \lambda x \quad A^2 x = \lambda A x = \lambda \cdot \lambda x = \lambda^2 x$$

$$A^{n-1} x = \lambda^{n-1} x$$

Moltpeliticità e simmetria per A:

$$A^n x = \lambda^{n-1} A x = \lambda^{n-1} \cdot \lambda x = \lambda^n x$$

Se A invertibile è vero anche per $n < 0$:

$$\lambda \in \sigma(A) \Rightarrow \frac{1}{\lambda} \sigma(A^{-1})$$

$$\lambda \in \sigma(A) \Rightarrow \frac{1}{\lambda^2} \sigma(A^{-1})$$

$$\sigma(A) = \sigma(A^T)$$

Dimostrazione 27

$$P_A(\lambda) = \det(A - \lambda I) = \det((A - \lambda I)^T) = \det(A^T - \lambda I) = P_{A^T}(\lambda)$$

Definizione 69

Se A è simmetrica (Hermitiana nel campo complesso) allora il suo spettro è **puramente reale**:

$$A = A^T \Rightarrow \forall \lambda \in \sigma(A), \quad \lambda \in \mathbb{R}$$

Definizione 70

Partendo dall'assunzione che

$$\forall x, y \in \mathbb{R}^n \backslash \{0\} \text{ autovettori di } A \text{ con } A = A^T \text{ relativi a } \lambda, \mu \text{ con } \lambda \neq \mu$$

Allora vale che

$$< x, y >= 0$$

Dimostrazione 28

Dato che $Ax = \lambda x$ e $Ay = \mu y$ (dalla definizione di autovettore), per le proprietà del prodotto scalare:

$$\begin{cases} < Ax, y >=< \lambda x, y >= \lambda < x, y > \\ < Ax, y >=< x, A^T y >=< x, Ay >=< x, \mu y >= \mu < x, y > \end{cases}$$

Applicando il metodo di riduzione ottengo:

$$0 = \lambda < x, y > -\mu < x, y >= \underbrace{(\lambda - \mu)}_{\neq 0} < x, y >\neq 0$$

Concludo che

$$< x, y >= 0$$

Teorema 18: Teorema spettrale

Se $A \in \mathbb{R}^{n\times n}$ è simmetrica allora $Q \in \mathbb{R}^{n\times n}$ ortogonale tale che $AQ = QD$ con $D = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Cioè esiste una base ortonormale di $\mathbb{R}^N$ formata da autovettori di A

Esempio Calcoliamo la matrice simile diagonale di una matrice simmetrica A

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \quad P_A(\lambda) = \det(A - \lambda I) = \begin{vmatrix} 1-\lambda & 2 \\ 2 & 1-\lambda \end{vmatrix} = \begin{vmatrix} 3-\lambda & 2 \\ 3-\lambda & 1-\lambda \end{vmatrix}$$

$$(3-\lambda)\begin{vmatrix} 1 & 2 \\ 1 & 1-\lambda \end{vmatrix} = (3-\lambda)(\lambda+1) = (\lambda-3)(\lambda+1) \quad \sigma(A) = \{3, -1\}$$

$$\lambda = 3 \to \ker(A - 3I) =< (1,1) > \to u^{(1)} = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix}$$

$$\lambda = 1 \to \ker(A + I) =< (-1,1) > \to u^{(2)} = \begin{pmatrix} -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix}$$

Dove abbiamo ricavato $u^{(1)}$ e $u^{(2)}$ dividendo le basi dei nuclei per la loro norma, cioè li abbiamo normalizzati (?)

$$A\begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}\begin{pmatrix} 3 & 0 \\ 0 & -1 \end{pmatrix}$$

Definizione 71: Gran-Schmidt

E' una procedura per ottenere a partire da una base di un sottospazio vettoriale una **base ortogonale**. Supponiamo che il sistema originale sia tipo: $\{u^{(1)}, u^{(2)}, \ldots, u^{(n)}\}$ allora Gran-Schmidt produce $\{w^{(1)}, w^{(2)}, \ldots, w^{(n)}\}$ secondo queste regole:

$$\begin{aligned} w^{(1)} &= u^{(1)} \\ w^{(2)} &= u^{(2)} - \frac{< u^{(2)}, w^{(1)} >}{||w^{(1)}||_2^2} w^{(1)} \\ w^{(3)} &= u^{(3)} - \frac{< u^{(3)}, w^{(1)} >}{||w^{(1)}||_2^2} w^{(1)} - \frac{< u^{(3)}, w^{(2)} >}{||w^{(2)}||_2^2} w^{(2)} \\ &\vdots \\ w^{(k)} &= u^{(k)} - \sum_{i=1}^{k-1} \frac{< u^{(k)}, w^{(i)} >}{||w^{(i)}||_2^2} w^{(i)} \end{aligned}$$

$$< Ax, x > =< x, Ax > = x^T Ax \in \mathbb{R} \text{ se } A = A^T$$

Definizione 72: Quoziente di Rayleigh

Data A simmetrica si definisce **quoziente di Rayleigh** il rapporto:

$$q(A,x) = \frac{< x, Ax >}{< x, x >} = \frac{x^T Ax}{x^T x} \quad x \neq 0$$

Si può dimostrare che il minimo autovalore di una matrice simmetrica è

$$\lambda \min(A) \leq q(A,x) \leq \lambda \max(A)$$

Quando $x = \lambda$ otteniamo:

$$\dots$$

Per x autovettore di A relativo a λ otteniamo

$$q(x) = \frac{x^T Ax}{x^T x} = \frac{x^T(\lambda x)}{x^T x} = \frac{\lambda(x^T x)}{(x^T x)} = \lambda$$

Esempio Determinare la matrice ortogonale U che diagonalizza la matrice simmetrica A

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

Cominciamo con il calcolare gli autovalori per ottenere il polinomio caratteristico.

$$P(\lambda) = \det(A - \lambda I) = \begin{vmatrix} 5-\lambda & 2 & 1 \\ 2 & 2-\lambda & -2 \\ 1 & -2 & 5-\lambda \end{vmatrix}$$

Come primo passo svogliamo $C1 \leftarrow C1 + C3$:

$$\begin{vmatrix} 5-\lambda & 2 & 1 \\ 2 & 2-\lambda & -2 \\ 1 & -2 & 5-\lambda \end{vmatrix} = \begin{vmatrix} 6-\lambda & 2 & 1 \\ 0 & 2-\lambda & -2 \\ 6-\lambda & -2 & 5-\lambda \end{vmatrix} = (6-\lambda) \begin{vmatrix} 1 & 2 & 1 \\ 0 & 2-\lambda & -2 \\ 1 & -2 & 5-\lambda \end{vmatrix}$$

Come secondo passo svolgiamo $R3 \leftarrow R3 - R1$ e otteniamo:

$$(6-\lambda) \begin{vmatrix} 1 & 2 & 1 \\ 0 & 2-\lambda & -2 \\ 0 & -4 & 4-\lambda \end{vmatrix} \quad \text{Applicando la formula di Laplace} \quad (6-\lambda) \begin{vmatrix} 2-\lambda & -2 \\ -4 & 4-\lambda \end{vmatrix}$$

Come terzo passo svogliamo $C1 \leftarrow C1 + C2$:

$$(6-\lambda) \begin{vmatrix} -\lambda & -2 \\ -\lambda & 4-\lambda \end{vmatrix} = -\lambda(6-\lambda) \begin{vmatrix} 1 & -2 \\ 1 & 4-\lambda \end{vmatrix} = -\lambda(6-\lambda)[4-\lambda+2] = -\lambda(\lambda-6)^2$$

Gli autovalori sono : $\sigma(A) = \{0, 6\}$ $\lambda = 6$ ha molteplicità algebrica 2

L'autospazio relativo a $\lambda = 0$ avrà dimensione 1, quello relativo a $\lambda = 6$ dimensione 2

$$E_{\lambda=0} = \ker(A) \to \dim = 1 \qquad E_{\lambda=6} = \ker(A - 6I) \to \dim = 2$$

$$E_{\lambda=6} \qquad A-6I = \begin{pmatrix} -1 & 2 & 1 \\ 2 & -4 & -2 \\ 1 & -2 & -1 \end{pmatrix} \quad rank(A-6I) = 1$$

Per trovare il nucleo devo risolvere:

$$\{(x_1,x_2,x_3) \in \mathbb{R}^3 | \begin{pmatrix} -1 & 2 & 1 \\ 2 & -4 & -2 \\ 1 & -2 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Dalla prima equazione ottengo:

$$x_1 = 2x_2 + x_3 \qquad \{(2x_2+x_3, x_2, x_3) | x_2, x_3 \in \mathbb{R}\} = < (2,1,0)(1,0,1) >$$

Nota che $< (2,1,0),(1,0,1) >$ indica lo spazio generatori e non il prodotto scalare nonostante la notazione sia la stessa

$$< (2,1,0),(1,0,1) >= 2+0+0 = 2 \neq 0$$

Applichiamo la procedura di Gram-Schmidt che ci consente dato un sistema di vettori linearmente indipendenti di trovare un altro sistema di vettori linearmente indipendenti equivalente ma **ortogonali**.

$$w^{(1)} = (2,1,0)$$
$$w^{(2)} = (1,0,1) - \frac{< (1,0,1), w^1 >}{||w^{(1)}||_2^2} w^{(1)} = (1,0,1) - \frac{2}{5}(2,1,0) = \left(\frac{1}{5}, \frac{-2}{5}, 1\right)$$

Nota che la norma $||w^{(1)}||_2 = \sqrt{4+1}$, $||w^{(2)}||_2 = \sqrt{6/5}$
Verifichiamo che sia ortogonale calcolando il prodotto scalare del vettore con $w^{(1)}$:

$$< (\frac{1}{5}, \frac{-2}{5}, 1), (2,1,0) >= \frac{2}{5} - \frac{2}{5} + 0 = 0$$

Non ci bastano i vettori ortogonali, li vogliamo di norma 1. Prendiamo quindi come colonne della matrice U i vettori $w^{(1)}$ e $w^{(2)}$ divisi per la loro norma 2.

$$u^{(1)} = \frac{1}{\sqrt{5}}(2,1,0)^T = \left(\frac{2\sqrt{5}}{5}, \frac{\sqrt{5}}{5}, 0\right)^T$$
$$u^{(2)} = \left(\frac{1}{\sqrt{30}}, \frac{-2}{30}, \frac{\sqrt{5}}{\sqrt{6}}\right)$$

$$E_\lambda = 0 = \ker(A) = \{x \in \mathbb{R}^3 | Ax = 0\} =< (-1,2,1) >$$

Possiamo verificare che $(-1,2,1)$ è ortogonale a $u^{(1)}$ e $u^{(2)}$. La terza colonna di U sarà data da questo vettore diviso la sua norma 2.

$$||(-1,2,1||_2 = \sqrt{6} \Rightarrow U = \begin{pmatrix} \frac{2\sqrt{5}}{5} & \frac{1}{\sqrt{30}} & \frac{\sqrt{-6}}{6} \\ \frac{\sqrt{5}}{5} & \frac{-2}{\sqrt{30}} & \frac{\sqrt{6}}{3} \\ 0 & \frac{\sqrt{5}}{\sqrt{6}} & \frac{\sqrt{6}}{6} \end{pmatrix}$$

Esercizio Determinare la matrice ortogonale Q che diagonalizza la matrice simmetrica A:

$$A = \begin{pmatrix} 2 & 0 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & 0 \end{pmatrix} \qquad \text{Una possibile soluzione è: } Q = \begin{pmatrix} \frac{-\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} \\ 0 & 1 & 0 \\ \frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} \end{pmatrix}$$

Definizione 73: Matrice definite positive, negative e indefinite

Sia $A \in \mathbb{R}^{n \times n}$ simmetrica, diciamo che

- A è **definita positiva** se
$$\forall x \in \mathbb{R}^n \backslash \{0\} \quad x^T A x > 0$$
- A è **semidefinita positiva** se
$$\forall x \in \mathbb{R}^n \quad x^T A x \geq 0$$
- A è **definita negativa** se
$$\forall x \in \mathbb{R}^n \backslash \{0\} \quad x^T A x < 0$$
- A è **semidefinita negativa** se
$$\forall x \in \mathbb{R}^n \quad x^T A x > 0$$

Vale che:

$$A \in \mathbb{R}^{m \times n} \Rightarrow A^T A \text{ e } A A^T \text{ sono sempre simmetriche semidefinite positive}$$

$$x^T A^T A x = (Ax)^T \cdot (Ax) = ||Ax||_2^2 \geq 0$$

La matrice $A^T A$ è definita positiva $\leftrightarrow rank(A) = n$ (rango pieno)

$$x^T A^T A x = (Ax)^T \cdot (Ax) = ||Ax||_2^2 = 0 \iff Ax = 0 \iff x$$
$$x = 0 \iff ker(A) = \{0\} \iff rank(A) = n$$

$A^T A$ è sempre semidefinita positiva e è definita positiva se rank(A) = n (pieno per colonne)

AA^T è sempre semidefinita positiva e AA^T è definita positiva se rank(A) = m (pieno per righe)

Teorema 19

Sia $A \in \mathbb{R}^{n \times n}$ simmetrica A

- E' definita positiva se e solo se $\lambda_i > 0 \ \forall \lambda_i \in \sigma(A)$
- E' definita semipositiva se e solo se $\lambda_i \geq 0 \ \forall \lambda_i \in \sigma(A)$ (quindi non è invertibile)
- E' definita negativa se e solo se $\lambda_i < 0 \ \forall \lambda_i \in \sigma(A)$
- E' definita seminegativa se e solo se $\lambda_i \leq 0 \ \forall \lambda_i \in \sigma(A)$ (quindi non è invertibile)

Definizione 74: Seconda caratterizzazione della matrice simmetrica definita positiva

$A \in \mathbb{R}^{n\times n}$ è definita positiva se e solo se tutti i minori principali di testa sono positivi:

$$A = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 9 & -1 \\ 0 & -1 & 6 \end{pmatrix} \qquad A_1 = |4| = 4 \; A_2 = \begin{vmatrix} 4 & 1 \\ 1 & 9 \end{vmatrix} = 36 \; A_3 = \det(A) = 206$$

Teorema 20: 1° Teorema di Gershgorin, localizzazione degli autovalori

Data una matrice $A \in \mathbb{C}^{n\times n}$ definiamo per $k = 1, 2, \ldots, n$ i **dischi di Gershgorin** gli insiemi:

$$R_k = \{z \in \mathbb{C} \;\Big|\; |z - a_{kk}| \leq \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{kj}|\}$$

Un cerchio di centro a_{kk} e di raggio pari alla somma in valore assoluto dei valori al di fuori della diagonale principale (per quella riga). Possiamo affermare che lo spettro della matrice è incluso nell'unione di tutti questi dischi.

$$\sigma(A) \subseteq \bigcup_{i=1}^{n} Ri \text{ righe i-esime}$$

Dato che $\sigma(A) = \sigma(A^T)$ se consideriamo la trasposta A^T e calcoliamo i suoi dischi di Gershgorin, ciò equivale a calcolare i dischi di Gershgorin per le colonne di A. Quindi possiamo affermare che:

$$k = 1, 2, \ldots, n \qquad C_k = \{z \in \mathbb{C} \;\Big|\; |z - a_{kk}| \leq \sum_{\substack{i=1 \\ i\neq k}} |a_{ik}|\}$$

Allora:

$$\sigma(A) \subseteq \bigcup_{i=1}^{n} Ci \text{ colonne i-esime}$$

Segue quindi che lo spettro di A sarà compreso nell'intersezione dei dischi di Gershgorin calcolati per le righe e per le colonne:

$$\sigma(A) \subseteq \left(\bigcup_{i=1}^{n} R_i \cap \bigcup_{i=1}^{n} C_i\right)$$

Dimostrazione 29: 1° Teorema di Gershgorin

Dobbiamo dimostrare che $\sigma(A) \subseteq \bigcup_{i=1}^{n} Ri$ con

$$R_k = \{z \in \mathbb{C} \;\Big|\; |z - a_{kk}| \leq \sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{kj}|\}$$

Che equivale a dimostrare che ogni $\lambda \in \sigma(A)$ è contenuto in R_k. Sia $\lambda \in \sigma(A)$ allora $\exists u \in \mathbb{R}^n \backslash \{0\}$ tale che $Au = \lambda u$

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{21n} \\ \vdots & \vdots & \dots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} = \lambda \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

Riscrivendo il tutto come sistema otteniamo:

$$\begin{cases} a_{11}u_1 + a_{12}u_2 + \dots + a_{1n}u_n = \lambda u_1 \\ a_{21}u_1 + a_{22}u_2 + \dots + a_{2n}u_n = \lambda u_2 \\ \vdots \\ a_{k1}u_1 + a_{k2}u_2 + \dots + a_{kn}u_n = \lambda u_k \\ \vdots \\ a_{n1}u_1 + a_{n2}u_2 + \dots + a_{nn}u_n = \lambda u_n \end{cases}$$

Prendiamo come esempio la riga k-esima e portiamo a membro destro l'elemento corrispondente a quello sulla diagonale:

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

Seleziono l'indice k tale che $|u_k| = ||u||_\infty$ (selezioniamo la più grande componente in modulo del vettore u, in questo caso supponiamo sia la k-esima).

$$|\lambda - a_{kk}| \cdot |u_k| = \left| \sum_{\substack{j=1 \\ j \neq k}}^{n} a_{kj}u_j \right| \leq \sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{kj}u_j|$$

Dividendo per $|u_k|$:

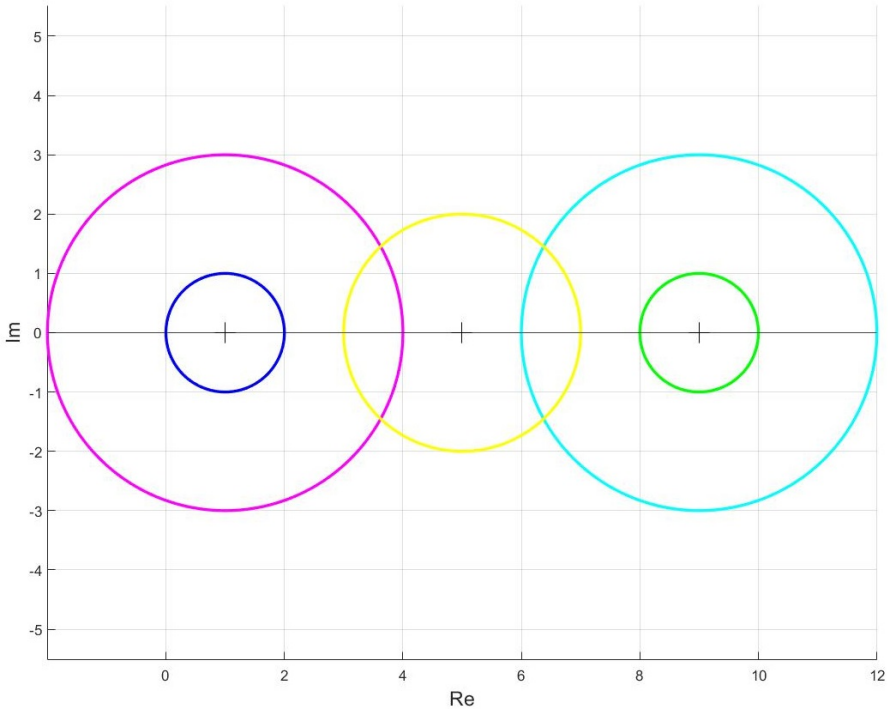
$$|\lambda - a_{kk}| \leq \sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{kj}| \cdot \left| \frac{u_j}{u_k} \right| \leq \sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{kj}|$$

Esempio Data la matrice:

$$\begin{pmatrix} 1 & -1 & 0 \\ 1 & 5 & 1 \\ -2 & -1 & 9 \end{pmatrix}$$

1. Questa matrice può avere autovalori reali? Notiamo che se esiste un autovalore complesso sicuramente anche il suo coniugato sarà un autovalore. Dato che la matrice è 3×3 possiamo affermare con certezza che almeno un autovalore sarà reale.
2. Usando Gershgorin disegnare la regione piana contenente gli autovalori di A

$$R_1 = \{z \in \mathbb{C} \mid |z-1| \leq 1\}$$
$$R_2 = \{z \in \mathbb{C} \mid |z-5| \leq 2\}$$
$$R_3 = \{z \in \mathbb{C} \mid |z-9| \leq 3\}$$
$$C_1 = \{z \in \mathbb{C} \mid |z-1| \leq 3\}$$
$$C_2 = \{z \in \mathbb{C} \mid |z-5| \leq 2\}$$
$$C_3 = \{z \in \mathbb{C} \mid |z-9| \leq 1\}$$



Gli autovalori si trovano nell'intersezione tra i dischi riga e quelli colonna

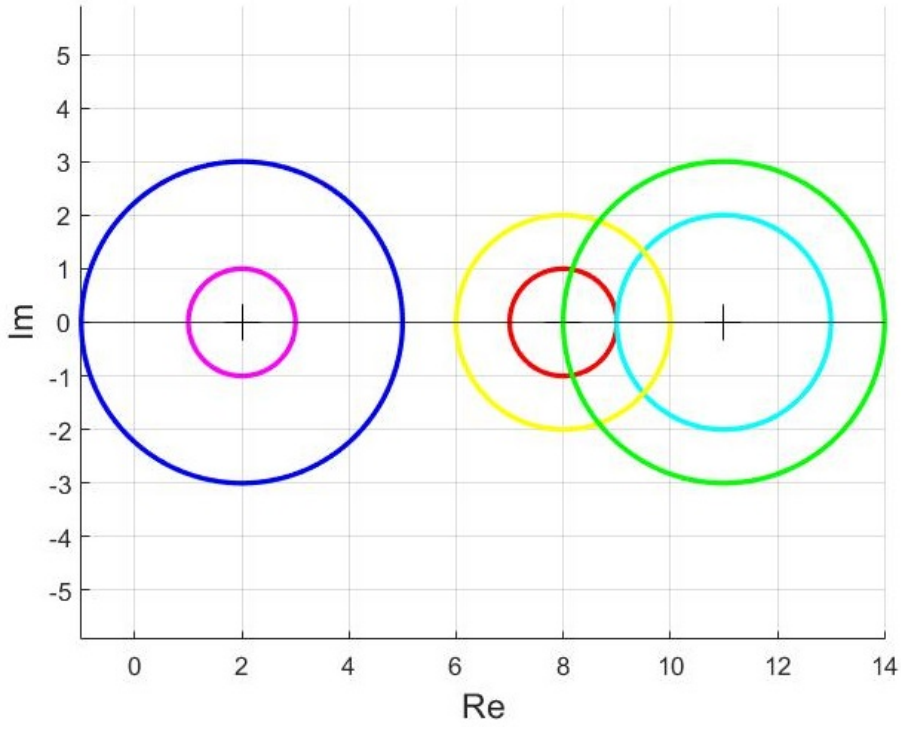
3. Dire se c'è un autovalore di A con parte reale negativa, no dal disegno
4. Può un autovalore di A avere parte reale uguale a 2.5, no dal disegno.

Dal 2° teorema di Gershgorin sappiamo che in un cerchio disgiunto c'è necessariamente un autovalore, possiamo quindi affermare che gli autovalori di questa matrice sono tutti reali.

Esempio Dimostrare se $(3+i) \in \sigma(A)$

$$\begin{pmatrix} 2 & 1 & 2 \\ 0 & 8 & 1 \\ 1 & 1 & 11 \end{pmatrix}$$

$$R_1 = \{z \in \mathbb{C} \mid |z-2| \leq 3\}$$
$$R_2 = \{z \in \mathbb{C} \mid |z-8| \leq 1\}$$
$$R_3 = \{z \in \mathbb{C} \mid |z-11| \leq 2\}$$
$$C_1 = \{z \in \mathbb{C} \mid |z-2| \leq 1\}$$
$$C_2 = \{z \in \mathbb{C} \mid |z-8| \leq 2\}$$
$$C_3 = \{z \in \mathbb{C} \mid |z-11| \leq 3\}$$



3 cerchi disgiunti, quindi per il 2° Teorema di Gershgorin ho un autovalore per disco. Concludiamo quindi che $(3+i) \notin \sigma(A)$. Anche dal disegno si poteva intuire che non era un autovalore.

Definizione 75: Dominanza diagonale

Una matrice quadrata $A \in \mathbb{R}^{n \times n}$ è **diagonalmente dominante per riga** se per ogni riga $\forall i = 1 \dots n$ vale che

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

Cioè l'elemento diagonale della riga $i-$esima è maggiore o uguale alla somma degli altri elementi della riga. Questa definizione renderebbe anche la matrice nulla una matrice diagonalmente dominante; per escluderla richiediamo che almeno una diseguaglianza sia stretta, cioè

$$\exists i \mid |a_{ii}| > \sum_{j=1, j \neq i}^{n} |a_{ij}|$$

Esiste anche una definizione in senso stretto, una matrice quadrata $A \in \mathbb{R}^{n \times n}$ è **diagonalmente dominante in senso stretto per riga** (DDSS) se $\forall i = 1 \dots n$ vale che:

$$|a_{ii}| > \sum_{j=1,\ j \neq i}^{n} |a_{ij}|$$

Le matrici diagonalmente dominanti per colonne sono definite in maniera simile.

Vale che una matrice DDSS per righe o per colonne è **regolare** (invertibile). La dimostrazione si ottiene applicando il Teorema di Gershgorin (esercizio). Se inoltre è simmetrica e ha elementi diagonali positivi $a_{ii} > 0$ per $i = 1 \dots n$ allora la matrice è anche definita positiva.

$$\begin{cases} A = A^T \\ a_{ii} > 0 \\ DDSS \end{cases} \implies A \text{ è definita positiva}$$

6.2 Risoluzione numerica di sistemi di equazioni lineari

Esistono **metodi diretti** e **metodi iterativi**. Quelli diretti sono i prediletti per matrici dense, fattorizzano la matrice A quindi da $Ax = b$ si calcola una fattorizzazione di A e si risolve il sistema. Quelli iterativi sono preferibili per matrici sparse (numero di elementi diversi da 0 nell'ordine di n). Ci occupiamo principalmente di sistemi in questa forma:

$$Ax = b \qquad A \in \mathbb{R}^{n \times n} \text{ invertibile} \qquad x = A^{-1}b$$

Definizione 76: Errore e Residuo

Il calcolatore non genererà gli stessi valori del vettore x "vero", se siamo fortunati i valori del vettore generato dal colatore disteranno ad una distanza uguale alla precisione di macchina dai valori reali. Quindi come al solito definiamo **errore** la differenza tra il valore vero ed il valore ottenuto dal calcolatore, in questo caso i valori fanno riferimento alle soluzioni del sistema.

$$e = \delta x = x - \overline{x}$$

Un ragionamento analogo si applica alle trasformazioni delle matrici, definiamo **residuo**

$$r = b - A\overline{x} = Ax - A\overline{x}$$

Definizione 77: Sistemi triangolari

La soluzione di sistemi lineari in cui la matrice è triangolare risulta essere efficiente in quanto è possibile disaccoppiare le equazioni ovvero ricavare da ciascuna equazione una sola incognita alla volta. Dividiamo i sistemi a seconda della forma della matrice in:

- **Sistemi triangolari superiori**: $Ux = b$

$$\begin{pmatrix} u_{11} & u_{12} & u_{13} & \dots & u_{1n} \\ & u_{22} & u_{23} & \dots & u_{2n} \\ & & u_{33} & \dots & u_{3n} \\ & & & & \dots \\ & & & & u_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

  Per cui è definito questo algoritmo di **sostituzione all'indietro** con un costo computazionale di $O(n^2/2)$

  $x_n = \frac{b_n}{u_{nn}}$
  `for i = n - 1 : -1 : 1 do`
  $x_i = b_i - \frac{1}{u_{ii}} \left(\sum_{k=i+1}^{n} u_{ik} x_k \right)$;
  `end for`

- **Sistemi triangolari inferiori**: $Lx = b$

$$\begin{pmatrix} l_{11} & & & & \\ l_{21} & l_{22} & & & \\ l_{31} & l_{32} & l_{33} & & \\ \dots & \dots & \dots & \dots & \\ l_{n1} & l_{n2} & l_{n3} & \dots & l_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

  Per cui è definito questo algoritmo di **sostituzione in avanti** con un costo computazionale di $O(n^2/2)$

  $x_1 = \frac{b_1}{l_{11}}$
  `for i = 2 : n do`
  $x_i = b_i - \frac{1}{l_{ii}} \left(\sum_{k=1}^{i-1} l_{ik} x_k \right)$;
  `end for`

Definizione 78: Condizionamento dei sistemi lineari

Consideriamo un sistema di equazioni lineari $Ax = b$ con $A \in \mathbb{R}^{n \times n}$ e $b \in \mathbb{R}^n$. Sia $\delta x = e = \overline{x} - x$ l'errore sul risultato in seguito ad una perturbazione δb. In generale l'analisi diventa molto più complessa se consideriamo entrambi δA e δb diversi da 0, tuttavia per semplicità ora consideriamo $\delta A = 0 \in \mathbb{R}^n$, cioè accumuliamo l'errore solo sul termine noto. Il sistema da risolvere diventa:

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

Poiché $Ax = b$:

$$A\delta x = \delta b \qquad \delta x = A^{-1}\delta b$$

Rispetto ad una qualsiasi norma matriciale indotta da quella vettoriale seguono le maggiorazioni:

$$||\delta x| = ||A^{-1}\delta b|| \leq ||A^{-1}||\ ||\delta b||$$

$$||b|| = ||Ax|| \leq ||A||\ ||x|| \Longrightarrow \frac{1}{||x||} \leq \frac{||A||}{||b||}$$

Moltiplicando la prima per la seconda ottengo una correlazione tra l'errore nella soluzione e l'errore nei dati.

$$\frac{||\delta x||}{||x||} \leq ||A^{-1}||\ ||A||\ \frac{||\delta b||}{||b||}$$

$$\frac{||\delta x||}{||x||} = \text{ Errore relativo nella soluzione}$$

$$\frac{||\delta b||}{||b||} = \text{ Errore relativo nei dati}$$

$$\kappa(A) = ||A^{-1}||\ ||A|| = \text{ Numero o indice di condizionamento di A}$$

Vale che il numero di condizionamento è sempre ≥ 1 in quanto:

$$1 = ||I|| = ||AA^{-1}|| \leq ||A||\ ||A^{-1}|| = \kappa(A)$$

L'errore nei dati è quindi **amplificato** dal numero di condizionamento.

Teorema 21

Se si assume che sulla matrice A vi sua una perturbazione ($\delta A \neq 0$)

$$\det(A) \neq 0 \qquad ||A^{-1}|| \cdot ||\delta A|| < 1$$

Si ottiene:

$$\frac{||\delta x||}{||x||} \leq \frac{\kappa(A)}{1 - \frac{||\delta A||}{||A||}\kappa(A)} \cdot \left(\frac{||\delta b||}{||b||} + \frac{||\delta A||}{||A||} \right)$$

Su ipotesi che $\frac{||\delta A||}{||A||} << \kappa(A)$ allora

$$\frac{\kappa(A)}{1 - \frac{||\delta A||}{||A||}\kappa(A)} \approx \kappa(A)$$

Definizione 79: Malcondizionamento

Generalmente non conosciamo l'errore relativo sui dati, considerando che:

$$A(\underbrace{x + \delta x}_{\overline{x}}) = b + \delta b$$

si ha che:

$$r = b - A\overline{x} = b - A(x + \delta x) = Ax - Ax - A\delta x = -\delta b$$

Deduciamo che la norma dell'errore relativo e quella del residuo relativo sono legate mediante la relazione:

$$\frac{||\delta x||}{||x||} \leq \kappa(A) \frac{||r||}{||b||}$$

Per valori molto grandi di $\kappa(A)$ ($\kappa(A) > 10^3$) l'errore relativo sulla soluzione può essere molto grande anche se è piccolo l'errore relativo sui dati. Ciò significa che ad un residuo piccolo non necessariamente corrisponde un errore piccolo, in questi casi vi è un **malcondizionamento del sistema** (o della matrice).

Definizione 80: Numero di condizionamento e indice di condizionamento spettrale

Precedentemente abbiamo definito il **numero di condizionamento** come

$$\kappa(A) = ||A^{-1}|| \cdot ||A||$$

Vale che il numero di condizionamento cambia a seconda della norma matriciale, per esempio potremmo calcolarlo anche così:

$$\kappa(A) = ||A||_\infty \cdot ||A^{-1}||_\infty$$

Notiamo che poiché gli autovalori dell'inversa A^{-1} sono uguali a $[_i(A)]^{-1}$ si ha l'indice:

$$\kappa_2(A) = \left(\frac{\lambda_{max}(A^T A)}{\lambda_{min}(A^T A)}\right)^{\frac{1}{2}}$$

Se la matrice è simmetrica (hermitiana) allora l'indice $\kappa_2(A)$ è detto **indice di condizionamento spettrale**:

$$\kappa_2(A) = \frac{max|\lambda(A)|}{min(\lambda(A)|} = \rho(A)\rho(A^{-1})$$

Se infine A è anche definita positiva allora $\kappa_2(A)$ diventa:

$$\kappa_2(A) = \frac{max(A)}{min(A)}$$

6.2.1 Metodi diretti

I **metodi diretti** sono i metodi basati sul metodo di eliminazione di Gauss (MEG) che trasforma il sistema originale $Ax = b$ scrivendola in forma completa $(A|b)$ (aggiungiamo come ultima colonna il termine noto) e tramite una serie di trasformazioni otteniamo una matrice triangolare superiore ed applicando il metodo di sostituzione all'indietro si risolve il sistema $Ux = \tilde{b}$ (sistema trasformato). I passaggi si articolano in $n-1$ passi, perché azzeriamo una "colonna inferiore alla diagonale" alla volta fino ad arrivare alla penultima colonna. Le operazioni ammesse sono di 3 tipi:

- Scambi di righe, $R_i \leftrightarrow R_j$
- Moltiplicare una riga per un numero $\neq 0$ $R_i \leftarrow \alpha \cdot R_i$ $\alpha \neq 0$
- Combinazione lineare di 2 righe: $R_i \leftarrow R_i - m_{ik} \cdot R_k$

Esempio

$$A = \begin{pmatrix} 2 & 1 & 3 \\ 4 & 3 & 10 \\ 2 & 4 & 17 \end{pmatrix} \qquad b = \begin{pmatrix} 11 \\ 28 \\ 31 \end{pmatrix} \qquad (A|b) = \begin{pmatrix} 2 & 1 & 3 & 11 \\ 4 & 3 & 10 & 28 \\ 2 & 4 & 17 & 31 \end{pmatrix}$$

Calcoliamo i moltiplicatori:

- riga 2 $l_{21} = \frac{4}{2} = 2$ $R_2 \leftarrow R_2 - 2R_1$

- riga 3 $l_{31} = \frac{2}{2} = 1$ $R_3 \leftarrow R_3 - 1R_1$

$$(A|b) = \begin{pmatrix} 2 & 1 & 3 & 11 \\ 4 & 3 & 10 & 28 \\ 2 & 4 & 17 & 31 \end{pmatrix} \rightarrow \begin{pmatrix} 2 & 1 & 3 & 11 \\ 0 & 1 & 4 & 6 \\ 0 & 3 & 14 & 20 \end{pmatrix}$$

- riga 3 $l_{32} = \frac{3}{1} = 3$ $R_3 \leftarrow R_3 - 3R_2$

$$\begin{pmatrix} 2 & 1 & 3 & 11 \\ 0 & 1 & 4 & 6 \\ 0 & 3 & 14 & 20 \end{pmatrix} \rightarrow \begin{pmatrix} 2 & 1 & 3 & 11 \\ 0 & 1 & 4 & 6 \\ 0 & 0 & 2 & 2 \end{pmatrix}$$

Risolviamo il sistema triangolare superiore tramite sostituzione all'indietro:

$$\begin{cases} 2x_3 = 2 \quad x_3 = 1 \\ x_2 + 4x_3 = 6 \quad x_2 = 6 - 4 = 2 \\ x_1 = \frac{11-2-3}{2} = 3 \end{cases} \qquad x = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}$$

Verificare sempre se si ottiene b svolgendo Ax (non avrà pietà nel compito).

Se si porta a termine (cioè non otteniamo un pivot uguale a 0) dal metodo di eliminazione di Gauss possiamo ricavare $A = L \cdot U$ (matrice triangolare inferiore). Con MEG abbiamo già la matrice U, per ottenere L dobbiamo:

- Porre la diagonale uguale ad 1 e tutti gli elementi superiori uguali a 0
- Porre gli elementi della colonna $i-$esima inferiori alla diagonali uguali ai moltiplicatori ricavati nel passaggio $i-$esimo.

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 3 & 1 \end{pmatrix}; \quad U = \begin{pmatrix} 2 & 1 & 3 \\ 0 & 1 & 4 \\ 0 & 0 & 2 \end{pmatrix}; \quad L \cdot U = \begin{pmatrix} 2 & 1 & 3 \\ 4 & 3 & 10 \\ 2 & 4 & 17 \end{pmatrix} = A$$

Definizione 81: Algoritmo di eliminazione di Gauss

```
for k = 1,...,n-1 do
    for i = k +1,..., do
        l_ik = a_ik^(k) / a_kk^(k)
        for j = k,...,n do
            a_ij^(k) = a_ij^(k) - l_ik a_kj^(k)
        end for
    end for
end for
```

Il costo computazionale è pari a $O(\frac{2}{3}n^3)$

Se dobbiamo risolvere una sequenza n sistemi tutti con la stessa matrice A calcolo una volta $A = L \cdot U$ e costa $O(\frac{2}{3}n^3)$, poi ogni sistema lo risolvo mediante sostituzione in avanti e sostituzione all'indietro dal costo $O(\frac{n^2}{2})$. Ciò ci è permesso dalla fattorizzazione LU infatti:

$$Ax = b \text{ ma } A = L \cdot U \text{ quindi } L \cdot \underbrace{U \cdot x}_{y} = b$$

$$\begin{cases} Ly = b \\ Ux = y \end{cases}$$

Quindi se io conosco la fattorizzazione LU di una matrice A la risoluzione di un qualsiasi sistema che ha come matrice dei coefficienti A si traduce nella risoluzione di due sistemi triangolari dal costo computazionale quadratico.

Dimostrazione 30: Numero di condizionamento spettrale

Dimostriamo che per una matrice simmetrica (hermitiana)

$$k_2(A) = ||A||_2 \cdot ||A^{-1}||_2 = \frac{\max|\lambda(A)|}{\min|\lambda(A)|}$$

Partiamo dalla definizione di norma 2:

$$||A||_2 = \sqrt{\rho(A^T A)}$$

Allora supponendo $A = A^T$ ho che $A^T A = A^2$ possiamo scrivere che la norma 2 di A così:

$$||A||_2 = \sqrt{\rho(A^T A)} = \sqrt{\rho(A^2)} = \sqrt{[\rho(A)]^2} = |\rho(A)| = \rho(A) = \max|\lambda(A)|$$

La norma 2 dell'inversa è:

$$||A^{-1}||_2 = \max|\lambda(A^{-1}| = \frac{1}{\min|\lambda(A)|}$$

E quindi:

$$\kappa_2 = ||A||_2 \cdot ||A^{-1}||_2 = \max|\lambda(A)| \cdot \frac{1}{\min|\lambda(A)|}$$

La dimostrazione è identica se A è anche definita positiva, spariscono soltanto i valori assoluti.

Esempio Matrice malcondizionata. Supponiamo di avere questa matrice:

$$A = A^T = \begin{pmatrix} 1 & 10 \\ 10 & 101 \end{pmatrix} \qquad A^{-1} = \begin{pmatrix} 101 & -10 \\ -10 & 1 \end{pmatrix}$$

Notiamo che le due righe che la compongono sono quasi linearmente dipendenti, in genere questo è un indice che la matrice è malcondizionata. Troviamo che:

$$\sigma(A) = \{0.9804864 \times 10^{-2},\ 0.1019902 \times 10^{3}\}$$

Quindi

$$\kappa_\infty = \kappa_1 = 12321;\ \kappa_2 = 10402$$

Notiamo che:

- Se $b^T = (11, 111)$ il sistema $Ax = b$ ha per soluzione $x^T = (1, 1)$.
- Se $b^T = (11.11, 112.11)$ il sistema ha per soluzione $x^T = (1.01, 1.01)$ quindi otteniamo un errore relativo sui dati ed un errore relativo sui risultati:

  $$\frac{||\delta b||}{||b||} = 10^{-2} \qquad \frac{||\delta x||}{||x||} = 10^{-2}$$

  Quindi qui non c'è stata un'amplificazione dovuta all'errore di condizionamento
- Se $b^T = (11.1, 111)$ il sistema ha per soluzione $x^T = (11.1, 0)$ ed i seguenti errori relativi sui dati e sui risultati:

  $$\frac{||\delta b||}{||b||} = 0.9 \times 10^{-3} \qquad \frac{||\delta x||}{||x||} = 10.1$$

  Quindi c'è stata una grande amplificazione dovuta all'errore di condizionamento nonostante una perturbazione sui dati in teoria minore rispetto al caso precedente.

Teorema 22: Teorema di esistenza ed unicità della fattorizzazione LU

Sia $A \in \mathbb{R}^{n \times n}$, se tutti i minori principali di testa sono diversi da 0 cioè $|A_i| \neq 0,\ i = 1, 2, \ldots, n$ (con A_i indichiamo la sottomatrice di dimensione $i \times i$) allora esiste una matrice triangolare inferiore L ed una matrice triangolare superiore U tali che $A = L \cdot U$. Se A è invertibile allora la fattorizzazione è unica se imponiamo che la matrice L abbia diagonale unitaria ($l_{ii} = 1,\ i = 1, \ldots, n$) (in realtà basta che L o U abbiano diagonale unitaria, ma per noi sarà sempre L).

Dimostrazione 31: Esistenza della fattorizzazione LU (informale)

Consideriamo la **matrice elementare di Gauss**

$$L_k = I - l_k \cdot e_k^T$$

Dove l_k è il vettore dei moltiplicatori:

$$l_k = (\underbrace{0, 0, \ldots, 0}_{k}, l_{k+1,k}, l_{k+2,k}, \ldots, l_{n,k})$$

$$L_k = \begin{bmatrix} 1 & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & 1 & \ldots & \ldots & \ldots & \ldots & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \ldots & \ldots & \ldots & 1 & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & -l_{k+1},k & 1 & \ldots & \ldots \\ \ldots & \ldots & \ldots & -l_{k+1},k & \ldots & 1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \ldots & \ldots & \ldots & -l_n,k & \ldots & \ldots & 1 \end{bmatrix}$$

Questa è la matrice che viene moltiplicata alle matrici che otteniamo durante MEG fino ad arrivare alla matrice U. Per esempio al primo passo avremo:

$$L_1 = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ -l_{2,1} & 1 & \ldots & 0 \\ -l_{3,1} & \ldots & \ldots & \ldots \\ \vdots & \vdots & \vdots & \vdots \\ -l_{n,1} & \ldots & \ldots & 1 \end{pmatrix}$$

In pratica durante MEG svolgiamo questo prodotto:

$$L_{n-1} \ldots L_2 L_1 A = U$$

Il determinante di ciascuna di queste matrici L_k è 1 quindi sono invertibili e allora anche il loro prodotto sarà invertibile. Concludiamo quindi che

$$A = (L_{n-1} L_{n-2} \ldots L_2 L_1)^{-1} U \text{ ma quindi } (L_{n-1} L_{n-2} \ldots L_2 L_1)^{-1} = L$$

Dimostrazione 32: Unicità della fattorizzazione LU

Supponiamo per assurdo che data una A esistano due fattorizzazioni diverse:

$$A = L_1 \cdot U_1 = L_2 \cdot U_2$$

Possiamo premoltiplicare per l'inversa di L_2 e postmoltiplicare per l'inversa di U_1 :

$$L_2^{-1} L_1 \underbrace{U_1 U_1^{-1}}_{I} = \underbrace{L_2^{-1} L_2}_{I} U_2 U_1^{-1} \rightarrow L_2^{-1} L_1 = U_2 U_1^{-1} = D$$

A sinistra abbiamo un prodotto di triangolari inferiori che genera ancora una triangolare inferiore, a destra un prodotto di triangolari superiori che genera ancora una triangolare superiore. Dato che sono uguali devono per forza avere tutti gli elementi uguali a 0 tranne quelli sulla diagonale che saranno uguali tra di loro, quindi sono matrici diagonali.

$$L_2^{-1} L_1 = D \rightarrow L_1 = L_2 D$$

Cioè stiamo svolendo una scalatura sulle colonne di L_2

$$\underbrace{\begin{pmatrix} 1 & \dots & \dots & \dots \\ \dots & 1 & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots \\ \dots & \dots & \dots & 1 \end{pmatrix}}_{L_1} = \underbrace{\begin{pmatrix} 1 & \dots & \dots & \dots \\ \dots & 1 & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots \\ \dots & \dots & \dots & 1 \end{pmatrix}}_{L_2} \underbrace{\begin{pmatrix} d_{11} & \dots & \dots & \dots \\ \dots & d_{22} & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots \\ \dots & \dots & \dots & d_{nn} \end{pmatrix}}_{D}$$

Allora obbligatoriamente si ha che $D = I$, ma quindi:

$$L_1^{-1} = L_2^{-1} \iff L_1 = L_2 \text{ analogamente } U_1 = U_2$$

Non tutte le matrici hanno tutti i minori principali di testa diversi da 0, per esempio questa matrice ha soluzione unica ($det(A) \neq 0$) eppure ha un minore principale di testa nullo.

$$A = \begin{pmatrix} 0 & 2 & 1 \\ 1 & 0 & 0 \\ 3 & 0 & 1 \end{pmatrix} \quad b = \begin{pmatrix} 5 \\ -1 \\ -2 \end{pmatrix} \qquad \det(A) = -2 \neq 0$$

Definizione 82: Pivoting

Definiamo **pivoting** come la scelta opportuna dell'elemento **pivot** ad ogni passo del MEG, cioè chi va al denominatore quando si calcola il moltiplicatore. Gli obiettivi della tecnica del pivoting sono 2:

1. Permettere la **conclusione** del MEG anche in quei casi in cui non si potrebbe applicare, cioè quando esiste un minore principale uguale a 0. Quindi di conseguenza ci permette di calcolare la fattorizzazione LU
2. Conferire **stabilità numerica** al MEG, senza pivoting MEG è instabile cioè l'errore si propaga.

Esistono due tecniche per selezionare il miglior pivot al passo $k-$esimo

Definizione 83: Pivoting parziale

Al passo k viene selezionato come elemento pivot l'elemento più grande (in valore assoluto) di tutta la colonna. Scelto il pivot ogni riga "sotto" il pivot subisce questa trasformazione:

$$R_i \leftarrow R_i - m_{ik} R_k \qquad m_{ik} = \frac{a_{ik}}{a_{kk}}$$

Scegliamo come pivot il numero più grande della colonna per generare un moltiplicatore il più piccolo possibile (a_{kk} è il pivot). Ricordiamo che quando sommiamo numeri aventi ordini di grandezza molto diversi c'è una perdita d'informazione sulle cifre più significative del più piccolo: ciò è fonte di errore. Applicando la tecnica del pivot parziale siamo certi che:

$$|l_{ik}| \leq 1$$

Quindi abbiamo eliminato il problema dei "numeri grandi". Formalmente al passo $k-$esimo la procedure è:

1. Si seleziona l'indice s tale che

$$|a_{sk}| \geq |a_{ik}| \quad i = k, \ldots, n$$

2. Si scambiano le righe s e k

$$R_s \longleftrightarrow R_k$$

3. Si ottiene ad ogni passo che

$$|l_{ik}| \leq 1 \quad \text{per } i = k+1, \ldots, n$$

Potenzialmente in questo algoritmo si può ancora verificare una crescita esponenziale dell'errore ed è quindi ancora instabile per alcune matrici, però è estremamente raro.

Definizione 84: Pivoting totale

Al passo $k-$esimo cerco il più grande elemento in tutta la sottomatrice (per sottomatrice intendiamo la matrice che va da k fino ad n sulla quale dobbiamo ancora lavorare). Supponiamo si trovi in a_{sr}, per portarlo in $[k, 1]$ bisogna svolgere scambi di righe e scambi di colonne.

$$|a_{sr}| \geq a_{ij} \text{ per } i, j = k, \ldots,$$

1. Scambio delle righe

$$R_s \longleftrightarrow R_k$$

2. Scambio delle colonne

$$C_r \longleftrightarrow C_k$$

Il costo delle operazioni è $O(n^2)$ e va fatto un totale di n volte, costa molto e quindi non viene quasi mai applicato. Anche la migliore stabilità che fornisce nella pratica trova pochissime applicazioni.

Definizione 85: Matrice di permutazione

Per effettuare lo scambio di due righe q e u dobbiamo pre-moltiplicare la matrice A per una **matrice di permutazione** P. Una matrice di permutazione è una matrice ortogonale che si ottiene dalla matrice I, basta scambiare la posizione righe q e u della matrice I. Fatto ciò ci basta moltiplicare A e P.

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad P \cdot \begin{pmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 3 & 3 & 3 \end{pmatrix} = \begin{pmatrix} 2 & 2 & 2 \\ 1 & 1 & 1 \\ 3 & 3 & 3 \end{pmatrix}$$

Nel pivoting totale per effettuare lo scambio di due colonne dobbiamopost-moltiplicare la matrice A, ottenendo questa equazione:

$$PAQQ^T x = Pb \qquad QQ^T = I$$

Abbiamo aggiunto Q^T per evitare di modificare la soluzione del sistema, tuttavia ciò implica che quando risolviamo il sistema otteniamo

$$LU\ (Q^T x) = Pb \qquad Q^T x = z \Longleftrightarrow x = Qz$$

Cioè per ottenere la soluzione vera del sistema dobbiamo effettuare una trasformazione su z

Dobbiamo quindi aggiungere un passaggio nell'algoritmo di fattorizzazione, perché per ogni L_k dovrà essere moltiplicato per un appropriata P_k.

$$L_{n-1}P_{n-1} \dots L_4P_4\ L_3P_3\ L_2P_2\ L_1P_1\ A = U$$

Quello che si ottiene è

$$(L)^{-1}PA = U \longrightarrow P \cdot A = L \cdot U$$

Dove P è la matrice risultante dal prodotto al contrario (nel senso da $n-1$ a 1)di tutte le matrici di permutazione.

Esempio Calcolare la fattorizzazione LU sfruttando la tecnica del pivoting parziale. Risolvere poi il sistema usando la fattorizzazione trovata.

$$A = \begin{pmatrix} 0.5 & -1.5 & 1 \\ 2 & 1 & -0.5 \\ 1 & -3 & 1.5 \end{pmatrix} \qquad b = \begin{pmatrix} 0 \\ 10 \\ -2 \end{pmatrix}$$

Per la prima colonna l'elemento più grande è 2, dobbiamo scambiare $R_2 \leftrightarrow R_1$ quindi:

$$P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}; \quad P_1 \cdot A = \begin{pmatrix} 2 & 1 & -0.5 \\ 0.5 & -1.5 & 1 \\ 1 & -3 & 1.5 \end{pmatrix} \qquad l_{21} = \frac{0.5}{2} = 0.25 \quad l_{31} = \frac{1}{2} = 0.5$$

Applichiamo le trasformazioni alla matrice A: $R_2 \leftarrow R_2 - 0.25R_1$ e $R_3 \leftarrow R_3 - 0.5R_1$

$$\begin{pmatrix} 2 & 1 & -0.5 \\ 0 & -1.75 & 1.125 \\ 0 & -3.5 & 1.75 \end{pmatrix} \qquad L = \begin{pmatrix} & & \\ 0.25 & & \\ 0.5 & & \end{pmatrix}$$

Per la seconda colonna l'elemento più grande è $|-3.5|$ quindi devo scambiare R_2 con R_3. Ricordiamo che si deve applicare la permutazione anche alla matrice L che abbiamo ottenuto al passaggio precedente

$$P_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}; \quad P_2 \cdot A = \begin{pmatrix} 2 & 1 & -0.5 \\ 0 & -3.5 & 1.75 \\ 0 & -1.75 & 1.125 \end{pmatrix} \quad P_2 \cdot L = \begin{pmatrix} & & \\ 0.5 & & \\ 0.25 & & \end{pmatrix}$$

Dobbiamo calcolare il moltiplicatore e applicare la trasformazione, in questo caso l_{32} e $R_3 \leftarrow R_3 - l_{32}R_2$

$$l_{32} = \frac{-1.75}{-3.5} = 0.5 \qquad \underbrace{\begin{pmatrix} 2 & 1 & -0.5 \\ 0 & -3.5 & 1.75 \\ 0 & 0 & 0.25 \end{pmatrix}}_{U}$$

Ora dobbiamo solo ricavarci la L e la P:

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 0.25 & 0.5 & 1 \end{pmatrix} \qquad P = P_2 \cdot P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Verifichiamo se $P \cdot A = L \cdot U$:

$$P \cdot A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0.5 & -1.5 & 1 \\ 2 & 1 & -0.5 \\ 1 & -3 & 1.5 \end{pmatrix} = \begin{pmatrix} 2 & 1 & -0.5 \\ 1 & -3 & 1.5 \\ 0.5 & -1.5 & 1 \end{pmatrix}$$

$$L \cdot U = \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 0.25 & 0.5 & 1 \end{pmatrix} \cdot \begin{pmatrix} 2 & 1 & -0.5 \\ 0 & -3.5 & 1.75 \\ 0 & 0 & 0.25 \end{pmatrix} = \begin{pmatrix} 2 & 1 & -0.5 \\ 1 & -3 & 1.5 \\ 0.5 & -1.5 & 1 \end{pmatrix}$$

Ora dobbiamo risolvere il sistema, siamo partiti con $Ax = b$, abbiamo permutato A ma quindi dobbiamo permutare anche b per ottenere $PAx = Pb$. Dato che $PA = LU$ lavoriamo con $LUx = Pb$ ma quindi

$$\begin{cases} Ly = Pb \\ Ux = y \end{cases}$$

$$Ly = Pb; \quad \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 0.25 & 0.5 & 1 \end{pmatrix} \cdot Pb = \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 0.25 & 0.5 & 1 \end{pmatrix} \begin{pmatrix} 10 \\ -2 \\ 0 \end{pmatrix} \quad \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 10 \\ -7 \\ 2 \end{pmatrix}$$

Nota: conviene sempre controllare che il vettore y risolva $Ly = Pb$, se è sbagliato e si va avanti si sbaglia anche il vettore x

$$Ux = y; \quad \begin{pmatrix} 2 & 1 & -0.5 \\ 0 & -3.5 & 1.75 \\ 0 & 0 & 0.25 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 10 \\ -7 \\ -1 \end{pmatrix} \rightarrow x = \begin{pmatrix} 4 \\ 4 \\ 4 \end{pmatrix}$$

Quando applichiamo pivoting cambia anche il calcolo del determinante:

$$\det(P) \cdot \det(A) = \det(L) \cdot \det(U) \quad \det(A) = \frac{\det(U)}{\det(P)} = \pm \det(U)$$

Vale che $\det(P) = +1$ se il numero di scambi di righe è pari, altrimenti se è dispari $\det(P) = -1$

6.2.2 Fattorizzazione LDU

Teorema 23: Teorema (LDU)

Sia A una matrice di ordine n con tutti i minori principali $\neq 0$ allora esiste ed è unica la fattorizzazione :

$$A = LDU$$

con L triangolare inferiore e U triangolare superiore e con $l_{ii} = u_{ii} = 1$ e D matrice diagonale. In soldoni dividiamo U per la sua diagonale (che contiene gli elementi utilizzati come pivot) in modo da ottenere 1 sulla diagonale.

Considerando $A = L\overline{U}$ e posto

$$D = (\overline{U}) = (\overline{u}_{11}, \overline{u}_{22}, \dots \overline{u}_{nn})$$

Poiché D è invertibile possiamo scrivere:

$$L\overline{U} = LD(D^{-1}\overline{U}) = LDU \quad (u_{ii} = 1)$$

$$D = \begin{pmatrix} 7 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix} \qquad \underbrace{\begin{pmatrix} 1/7 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/2 \end{pmatrix}}_{D^{-1}} \underbrace{\begin{pmatrix} 7 & 1 & 3 \\ 0 & 3 & -1 \\ 0 & 0 & 2 \end{pmatrix}}_{\overline{U}} = \underbrace{\begin{pmatrix} 1 & 1/7 & 3/7 \\ 0 & 1 & -1/3 \\ 0 & 0 & 1 \end{pmatrix}}_{U}$$

Teorema 24: LDL^T

Sia $A = A^T$ una matrice reale di ordine n con tutti i minori principali $\neq 0$ allora esiste ed è unica la fattorizzazione

$$A = LDL^T$$

Dove L è triangolare inferiore con $l_{ii} = 1$ e D diagonale (chiaramente $U = L^T$ data la simmetria). Il vantaggio principale di LDL^T rispetto alla fattorizzazione LU è che non abbiamo una L ed una U diverse, questo si riflette sul costo computazionale che diventa la metà di quello LU (rimane comunque cubico)

Dimostrazione 33: LDL^T

Dall'ipotesi abbiamo che $A = LDU$ ma per simmetrica si ha anche

$$A = LDU = (LDU)^T = U^T D L^T$$

Ma allora per unicità della fattorizzazione segue che $L = U^T$ e $U = L^T$

Teorema 25: Fattorizzazione di Cholesky

Sia A una matrice reale simmetrica e definita positiva (e quindi con i minori di testa diversi da 0) allora a è fattorizzabile nella forma:

$$A = RR^T$$

Dove R è una matrice triangolare inferiore reale non singolare

Dimostrazione 34: Fattorizzazione di Cholesky

L'idea è che vogliamo eliminare la D della fattorizzazione LDL^T "mettendola" metà in L e metà in L^T. Essendo A simmetrica essa ammette la fattorizzazione

$$A = LDL^T$$

Dalla definizione di matrice definita positiva si ha che per ogni vettore $x \neq 0$

$$0 < x^T Ax = \underbrace{x^T L}_{y^T} D \underbrace{L^T x}_{y} = y^T Dy \qquad (L^T x)^T = x^T L$$

con $y \neq 0$ perché L è invertibile e quindi D è definita positiva.

Gli elementi diagonali della matrice D sono positivi in quanto se scegliamo $y_i = e_i$ l'$i-$esimo vettore della base canonica si ha che $0 < e_i^T De_i = d_{ii}$. Possiamo quindi dedurre che la matrice D ammette una radice quadrata reale:

$$\sqrt{D} = (\sqrt{d_{11}}, \ldots, \sqrt{d_{nn}})$$

E quindi

$$A = LDL^T = \underbrace{L\sqrt{D}}_{R} \underbrace{\sqrt{D}L^T}_{R^T} = RR^T$$

Nota: probabile domanda d'esame

Esempio Supponiamo di avere questa matrice A

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

Notiamo immediatamente che è simmetrica e definita positiva, possiamo quindi procedere con la fattorizzazione di Cholesky, imponiamo $RR^T = A$:

$$\begin{pmatrix} r_{11} & 0 & 0 \\ r_{21} & r_{22} & 0 \\ r_{31} & r_{32} & r_{33} \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

Il metodo più efficiente per procedere è di calcolare R per le colonne e quindi R^T per righe:

$$\text{Prima colonna di R: } \begin{cases} r_{11}^2 = a_{11} = 2 \\ r_{21}r_{11} = a_{21} = 1 \\ r_{31}r_{11} = a_{31} = 1 \end{cases} \Rightarrow \begin{cases} r_{11} = \sqrt{2} \\ r_{21} = 1\sqrt{2} \\ r_{31} = 1\sqrt{2} \end{cases}$$

$$\text{Seconda colonna di R: } \begin{cases} r_{21}^2 r_{22}^2 = a_{22} = 2 \\ r_{21}r_{31} + r_{22}r_{r2} = a_{32} = 1 \end{cases} \Rightarrow \begin{cases} r_{22} = \sqrt{3/2} \\ r_{32} = 1\sqrt{6} \end{cases}$$

$$\text{Terza colonna di R: } \begin{cases} r_{31}^2 + r_{32}^2 + r_{33}^2 = a_{33} = 2 \end{cases} \Rightarrow \begin{cases} r_{33} = \sqrt{4/3} \end{cases}$$

Abbiamo trovato R

$$R = \begin{pmatrix} \sqrt{2} & 0 & 0 \\ 1\sqrt{2} & \sqrt{3/2} & 0 \\ 1/\sqrt{2} & 1/\sqrt{6} & \sqrt{4/3} \end{pmatrix}$$

Una volta in possesso della fattorizzazione possiamo risolvere il sistema lineare $Ax = b$ con $b^T = (1, 2, 3)$ otteniamo:

$$Ax = RR^T x = b \Leftrightarrow \begin{cases} Ry = b \\ R^T x = y \end{cases} \qquad y = \begin{pmatrix} 1/\sqrt{2} \\ \sqrt{3/2} \\ \sqrt{3} \end{pmatrix} \quad x = \begin{pmatrix} -1/2 \\ 1/2 \\ 3/2 \end{pmatrix}$$

6.2.3 Risoluzone di sistemi sovradeterminati

Sia $A \in \mathbb{R}^{m \times n}$ con $m \geq n$ e $b \in \mathbb{R}^m$ il sistema lineare $Ax = b$ ammette soluzione se e solo se $b \in Im(A)$, inoltre la soluzione è unica se $rank(A) = n$. Se il sistema **non ammette soluzione** (ad esempio nel caso $A \in \mathbb{R}^{m \times n}$ con $m >> n$ generalmente $b \notin Im(A)$) il problema dei minimi quadrati diventa:

$$\min_{x \in \mathbb{R}^n} ||b - Ax||_2$$

ovvero trovare $x \in \mathbb{R}^n$ tale che

$$||b - Ax||_2 \leq ||b - Ay||_2 \quad \forall y \in \mathbb{R}^n$$

Teorema 26: Teorema

Il vettore x è soluzione del sistema delle equazioni normali

$$A^T Ax = A^T b \qquad A^T (b - Ax) = 0$$

Se e solo se x è soluzione di

$$||b - Ax||_2 \leq ||b - Ay||_2 \quad \forall y \in \mathbb{R}^n$$

Dimostrazione 35

- $\Rightarrow$, Sia $x \in \mathbb{R}^n$ la soluzione di $A^T(b - Ax) = 0$, per ogni $y \in \mathbb{R}^n$ si ha:

$$b - Ay = b - Ax + Ax - Ay = b - Ax + A(x - y)$$

Da cui prendendo le norme al quadrato otteniamo

$$\begin{aligned} ||b - Ay||_2^2 &= ||b - Ax + A(x - y)||_2^2 \\ &= (b - Ax + A(x - y))^T (b - Ax + A(x - y)) = \\ &= ||b - Ax||_2^2 + ||A(x - y)||_2^2 + 2(x - y)^T \underbrace{A^T(b - Ax)}_{0} = \\ &= ||b - Ax||_2^2 + ||A(x - y)||_2^2 \geq ||b - Ax||_2^2 \end{aligned}$$

- $\Leftarrow$, Dimostrare la tesi equivale a dimostrare che se x soddisfa

$$||b - Ax||_2 \leq ||b - Ay||_2 \quad \forall y \in \mathbb{R}^n$$

allora $b - Ax$ è ortogonale a $Im(A)$, scrivendo $b = b_1 + b_2$ con $b_1 \in Im(A)$ e $b_2 \in Im(A)^\perp$ si ottiene

$$b - Ax = b_1 + b_2 - Ax = b_1 - Ax + b_2$$

Passando alle norme al quadrato

$$\begin{aligned} ||b - A||_2^2 &= ||(b_1 - Ax) + b_2|| =< (b_1 - Ax + b_2, (b_1 - Ax) + b_2 > \\ &= ||b_1 - Ax||_2^2 + ||b_2||_2^2 + 2 < \underbrace{(b_1 - Ax)}_{\in Im(A)}, \underbrace{b_2}_{\in Im(A)^\perp} > \\ &= ||b_1 - Ax||_2^2 + ||b_2||_2^2 \end{aligned}$$

Quindi la soluzione di $||b - Ax||_2 \leq ||b - Ay||_2 \quad \forall y \in \mathbb{R}^n$ si ottiene per $b_1 - Ax = 0$, pertanto

$$b - Ax = b_1 + b_2 - Ax = b_2 \in Im(A)^\perp$$

Nella pratica il sistema delle equazioni normali ha soluzione **unica** se la matrice A^TA è **non singolare**. Da $A^TAx = A^Tb$ ricaviamo x:

$$x = (A^TA)^{-1}A^Tb = A^+b$$

Dove con A^+ si definisce la **pseudoinversa** di A. In pratica invece di formare la matrice A^TA e risolvere il sistema delle equazioni normali si risolve direttamente il sistema sovradeterminato $Ax = b$ mediante la fattorizzazione QR. Non risolviamo il sistema

$$A^T(Ax - b) = 0 \longrightarrow A^TAx = A^Tb$$

Perché il condizionamento della A (che è una Vandermonde) è altissimo così come il condizionamento di A^TA è un quadrato del condizionamento di A (quindi ancora più alto).

Teorema 27: Fattorizzazione QR

Data una matrice $A \in \mathbb{R}^{m\times n}$ esiste una matrice ortogonale $Q \in \mathbb{R}^{m\times m}$ tale che:

$$A = QR = Q \begin{pmatrix} \tilde{R} \\ \bigcirc \end{pmatrix}$$

Dove $R \in \mathbb{R}^{m\times n}$ con $(r_{ij}) = 0$ con $i > j$ ($\tilde{R} \in \mathbb{R}^{n\times n}$ è triangolare superiore)

Nota: $\bigcirc$ si intende tante righe di zeri

Osservando che le ultime $m - n$ righe di R sono nulle possiamo scrivere

$$A = QR = \begin{pmatrix} \tilde{Q} & Q_2 \end{pmatrix} \begin{pmatrix} \tilde{R} \\ \bigcirc \end{pmatrix} = \tilde{Q}\tilde{R}$$

Dove $\tilde{Q} = \mathbb{R}^{m\times n}$, questa fattorizzazione ridotta è detta *skinny* o *light* QR. Dunque abbiamo:

$$Ax = b \iff \tilde{Q}\tilde{R}x = b \iff \underbrace{\tilde{Q}^T\tilde{Q}}_{I_n}\tilde{R}x = \tilde{Q}^T b \iff \tilde{R}x = \tilde{Q}^T b$$

Dobbiamo quindi risolvere un sistema triangolare superiore, oppure possiamo usare la **decomposizione ai valori singolari** (SVD = singular value decomposition).

6.3 Singular value decomposition

Teorema 28: Teorema

Rappresenta una generalizzazione della decomposizione spettrale a matrici generali (non necessariamente quadrate). $\forall A \in \mathbb{R}^{m\times n}$ esistono due matrici ortogonali $V \in \mathbb{R}^{m\times m}$ e $U \in \mathbb{R}^{n\times n}$ ed una matrice $\Sigma \in \mathbb{R}^{m\times n}$ con la forma seguente:

$$\Sigma = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \quad \Sigma_r = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_r \end{bmatrix}$$

Nota: Σ_r non è ovviamente quadrata anche se qua sopra sembra esserlo

Dove $\sigma_1 \geq \sigma_2 \cdots \geq \sigma_r > 0$ con $r \leq \min\{m, n\}$ tali che

$$A = V\Sigma U^T$$

Gli scalari σ_i sono detti **valori singolari** di A e le colonne di V e U si dicono **vettori singolari sinistri e destri** rispettivamente di A.La matrice U è la matrice ortogonale che diagonalizza A^TA cioè

$$U^TA^TAU = D = diag(\lambda_1, \ldots, \lambda_r, \lambda_{r+1}, \ldots, \lambda_n) \quad \lambda_1 \geq \cdots \geq \lambda_r > \lambda_{r+1} = \cdots = \lambda_n = 0$$

La matrice V è la matrice ortogonale che diagonalizza AA^T cioè

$$V^TAA^TV = diag(\mu_1, \ldots, \mu_s, \mu_{s+1}, \ldots, \mu_n) \quad \mu_1 \geq \cdots \geq \mu_s > \mu_{s+1} = \cdots = \mu_n = 0$$

Poiché A^TA e AA^T hanno gli **stessi autovalori non nulli** con le stesse molteplicità geometriche si ha che $r = s$ ed inoltre

$$\begin{matrix} \lambda_1 = \mu_1 \\ \lambda_2 = \mu_2 \\ \vdots \\ \lambda_r = \mu_r \end{matrix} \qquad \sigma_i = \sqrt{\lambda_i} \quad i = 1, \ldots, r$$

Siccome A^TA e AA^T sono semidefinite positive simmetriche possiamo trovare una base ortonormale e quindi sia A^TA che AA^T sono diagonalizzabili con una matrice di trasformazione ortogonale. Questa matrice è la U per A^TA mentre è la V per AA^T cioè U contiene gli autovalori per A^TA e V per AA^T

La soluzione formale del sistema delle equazioni normali è

$$x = A^+b$$

Possiamo calcolare la SVD di A

$$A = V\sigma U^T$$

e definiamo

$$A^+ = U\Sigma^+V^T \qquad \Sigma^+ : \sigma^+_{ii} = \begin{cases} \frac{1}{\sigma_i} & i = 1, \ldots, r \\ 0 & i = r+1, \ldots, \min\{r, n\} \end{cases}$$

Teorema 29

Sia $B \in \mathbb{R}^{p\times q}$:

- Sia $\lambda \in \mathbb{R}$ un autovalore non nullo di B^TB con autovettore relativo $x \in \mathbb{R}^q$ allora λ è autovalore di BB^T con autovettore corrispondente Bx
- Inoltre se $x^{(1)} \dots x^{(I)}$ sono autovettori indipendenti di B^TB relativi all'autovalore λ allora anche gli autovettori $Bx^{(1)} \dots Bx^{(I)}$ sono linearmente indipendenti. Segue che la molteplicità geometrica di λ come autovalore di B^TB è minore o uguale della molteplicità geometrica di λ come autovalore di BB^T

Cioè in soldoni se λ è u autovalore di A^TA con molteplicità geometrica l allora λ come autovalore di AA^T ha molteplicità geometrica $\geq l$

Dimostrazione 36: Bozza della dim

- Sia $\lambda \neq 0$ con $B^TBx = \lambda x$. Posto $y = Bx$ ($y \neq 0$ altrimenti λ sarebbe 0) si ha
$$BB^Ty = B(B^TB)x = \lambda Bx = \lambda y$$
- Da
$$0 = \alpha_1 Bx^{(1)} + \dots + \alpha_I Bx^{(I)}$$
premoltiplicando per B^T otteniamo
$$0 = \alpha_1 B^TBx^{(1)} + \dots + \alpha_I B^TBx^{(I)} = \lambda(\alpha x^{(1)} + \dots + \alpha_I x^{(I)})$$
Essendo che $\lambda \neq 0$ deve essere $\alpha_1 x^{(1)} + \dots + \alpha_I x^{(I)} = 0$ da cui possiamo dedurre che $\alpha_1 = \dots = \alpha_I$ perché $x^{(1)} \dots x^{(I)}$ sono linermente indipendenti.

Nell'atto pratico quando vogliamo calcolarci questa decomposizione $A = V\Sigma U$ dove ricordiamo che V è $m \times m$, Σ $m \times n$ e U è $n \times n$ dobbiamo:

1. Calcolare i valori singolari di A
$$\sigma_i = \sqrt{}_i \quad i = 1 \dots r \quad \lambda_i \neq 0 \wedge \lambda_i \in \sigma(A^TA) \wedge \lambda_i \in \sigma(AA^T)$$
Cioè ci prendiamo i $_i$ da $\sigma(AA^T)$ o $\sigma(A^TA)$ che tanto quelli diversi da 0 sono gli stessi. Così facendo otteniamo Σ
2. Dobbiamo calcolare U quindi gli autovettori ortogonali e unitari (ortonormali) di A^TA
3. Dobbiamo calcolare V, se U^i (la colonna i-esima) è autovettore di A^TA allora AU^i è autovettore di AA^T
$$< AU^{(i)}, AU^{(j)} >=< U^{(i)}, A^TAU^{(j)} >=< U^{(i)}, \lambda_j U^{(j)} >= \lambda_j < U^{(i)}, U^{(j)} >= \begin{cases} \lambda_j & i = j \\ 0 & i \neq j \end{cases}$$
$$||AU^{(i)}||_2^2 =< AU^{(i)}, AU^{(j)} >= \lambda_i \neq 1 \quad V^i = \frac{1}{\sqrt{\lambda_i}} AU^{(i)}$$

A partire dalle colonne della matrice ortogonale U che diagonalizza A^TA è possibile ottenere le colonne della matrice V per $i = 1, \dots, r$ (quando finisco gli autovalori positivi ovviamente mi devo fermare)

$$V^{(i)} = \frac{1}{\sqrt{\lambda_i}} AU^{(i)}$$

Siccome

$$V = \left[V^{(1)} \dots V^{(r)} \;\middle|\; V^{(r+1)} \dots V^{(m)} \right] \in \mathbb{R}^{m \times m}$$

Come ottengo le altre colonne (da $r+1$ a m)? Consideriamo che $V^{(r+1)} \dots V^{(m)}$ sono autovettori di AA^T relativi a $\lambda = 0$ vanno quindi calcolati nel **nucleo** di AA^T.

$$\ker(AA^T) = span\{\tilde{V}^{(r+1)}, \dots, \tilde{V}^{(n)}\}$$

Potremmo dover usare Gram-Schmidt per ortogonalizzare i vettori

Esempio Determinare una SVD della matrice A

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \in \mathbb{R}^{3\times 2} \qquad A = V\Sigma U^T$$

In questo caso $V \in \mathbb{R}^{3\times 3}$ ortogonale, $U \in \mathbb{R}^{2\times 2}$ ortogonale e $\Sigma \in \mathbb{R}^{3\times 2}$. Per prima cosa ci calcoliamo i **valori singolari**

$$A^TA = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 3 \\ 3 & 3 \end{pmatrix}$$

Calcoliamo gli autovalori di questa matrice

$$P(\lambda) = \det(A^TA - I) = \begin{vmatrix} 3-\lambda & 3 \\ 3 & 3-\lambda \end{vmatrix} = \begin{vmatrix} 6-\lambda & 3 \\ 6-\lambda & 3-\lambda \end{vmatrix} = (6-\lambda) \begin{vmatrix} 1 & 3 \\ 1 & 3-\lambda \end{vmatrix} = (6-\lambda)(-) = \lambda(\lambda-6)$$

Quindi

$$\sigma(A^TA) = \{6, 0\} \text{ Ordinati in senso decrescente } \; \sigma_1 = \sqrt{6} \; \sigma_2 = 0 \to \Sigma = \begin{pmatrix} \sqrt{6} & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

Ora dobbiamo calcolare la matrice U che richiede il calcolo degli autospazi relativi a tutti gli autovalori diversi da 0 quindi di A^TA.

$$\lambda = 6 \to E_{\lambda=6} = \ker(A^TA - 6I) = \ker \begin{pmatrix} -3 & 3 \\ 3 & -3 \end{pmatrix} = \{(x_1, x_2) \in \mathbb{R}^2 \mid -3x_1 + 3x_2 = 0\} = < (1,1) >$$

Non possiamo ancora usarlo come colonna della matrice U perché non ha norma 1, quindi dobbiamo dividerlo per la sua norma:

$$U^{(1)} = \begin{pmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix}$$

$$\lambda = 0 \to E_{\lambda=0} = \ker(A^TA) = < (-1,1) >$$

$$U^{(2)} = \begin{pmatrix} -\sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix} \longrightarrow U = \begin{pmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{pmatrix}$$

Nota: potevamo scegliere un'altra base per $U^{(2)}$ perché la fattorizzazione SVD **non** è unica

Ora dobbiamo calcolare la V, abbiamo un singolo valore λ non nullo e quindi $r = 1$, possiamo calcolarci solo una colonna con la formula:

$$v^{(1)} = \frac{1}{\sqrt{\lambda_1}} A U^{(1)} = \frac{1}{\sigma_1} A U^{(1)} = \frac{1}{\sqrt{6}} \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix} = \begin{pmatrix} \sqrt{3}/3 \\ \sqrt{3}/3 \\ \sqrt{3}/3 \end{pmatrix}$$

Gli altri dobbiamo cercarli nel nucleo di AA^T, quindi:

$$AA^T = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{pmatrix}$$

Tutto torna, ci servono 2 colonne e per il teorema delle dimensione i nuclei sono 2

$$rank(AA^T) = 1 \longrightarrow ker(AA^T) = 2$$

$$\begin{pmatrix} 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \qquad x_1+x_2+x_3 = 0 \quad \ker(AA^T) = \{(-x_2-x_3, x_2, x_3) \big| x_2, x_3 \in \mathbb{R}\}$$

$$\ker(AA^T) = \{x_2(-1,1,0) + x_3(-1,0,1), x_1, x_2 \in \mathbb{R}\} = span\{(-1,1,0)(-1,0,1)\}$$

Sono già ortogonali? No, quindi applico Gram-Schmidt

$$V^{(2)} = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -\sqrt{2}/2 \\ \sqrt{2}/2 \\ 0 \end{pmatrix} \qquad \tilde{V}^{(3)} = (-1,0,1) - \frac{< (-1,0,1), (-1,1,0) >}{2}(-1,1,0) = \left(-\frac{1}{2}, \frac{-1}{2}, 1\right)$$

Ora abbiamo che è ortogonale all'altro vettore, però non ha ancora norma 1 quindi lo dividiamo per la sua norma

$$||(\tilde{V}^{(3)}||_2 = ||\big(-1/2, -1/2, 1\big)|| = \sqrt{\frac{1}{4} + \frac{1}{4} + 1} = \sqrt{\frac{3}{2}} \qquad V^{(3)} = \frac{1}{\sqrt{3/2}} \begin{pmatrix} -1/2 \\ -1/2 \\ 1 \end{pmatrix} = \begin{pmatrix} -\sqrt{6}/6 \\ -\sqrt{6}/6 \\ -\sqrt{6}/3 \end{pmatrix}$$

$$V = \begin{pmatrix} \sqrt{3}/3 & -\sqrt{2}/2 & -\sqrt{6}/6 \\ \sqrt{3}/3 & \sqrt{2}/2 & -\sqrt{6}/6 \\ \sqrt{3}/3 & 0 & \sqrt{6}/3 \end{pmatrix}$$

Verificare sempre che $V\Sigma U^T = A$

Poiché V ed U sono matrici ortogonali la SVD di A si può scrivere equivalentemente moltiplicando a destra per U:

$$AU = V\Sigma \quad A \begin{bmatrix} u_1 & \dots & u_n \end{bmatrix} = \begin{bmatrix} v_1 & \dots & v_m \end{bmatrix} \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right]$$

Per colonne $U = [u_1 \dots u_2]$ e $V = [v_1 \dots v_m]$ si ha che:

$$\begin{cases} Au_1 = \sigma_1 v_1 \\ Au_2 = \sigma_2 v_2 \\ \quad \vdots \\ Au_r = \sigma_r v_r \\ Au_{r+1} = 0 \\ \quad \vdots \\ Au_n = 0 \end{cases}$$

Questo significa che i vettori $\{Au_{r+1} = 0 \dots Au_n = 0\} \in \ker(A)$ (C'è un altro risultato che afferma che sono esattamente questi i vettori del nucleo). Notiamo anche che

$$v^{(i)} = \frac{1}{\sigma_i} AU^{(i)} \in Im(A)$$

Le prime r colonne stanno nell'immagine di A mentre le restanti $n - r$ stanno nel nucleo di A. Vale che

$$\ker(A) = span\{u^{(r+1)}, u^{(r+2)}, \dots, u^{(n)}\} \quad Im(A) = span\{v^{(1)}, v^{(2)}, \dots, v^{(r)}\}$$

Quindi $rank(A) = r$

Esercizio Trovare $\ker(A)$, $Im(A)$ e $rank(A)$ della matrice A dell'esercizio precedente

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

6.3.1 SVD leggera

Definizione 86: SVD leggera

Sia $A = V\Sigma U^T$ una SVD di $A \in \mathbb{R}^{m \times m}$ con $\sigma_1, \sigma_2, \dots, \sigma_r$ i valori singolari non nulli. Siano poi $u^{(1)}, \dots, u^{(r)}$ e $v^{(1)}, \dots, v^{(r)}$ i corrispondenti vettori singolari destri e sinistri rispettivamente.

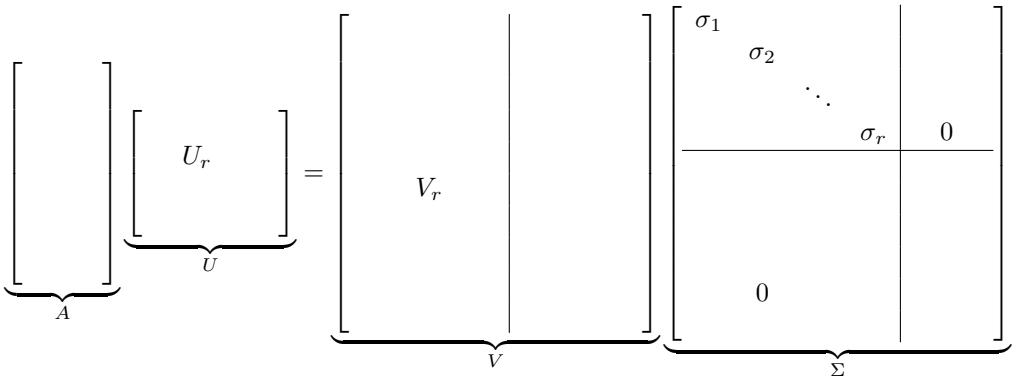
$$A = V_r \Sigma_r U_r^T = \sum_{i=1}^{r} \sigma_i \underbrace{v^{(i)} u^{(i)T}}_{E_i}$$

Con E_i matrici di rango 1. Questa scomposizione è detta **SVD leggera** di A in quanto contiene solo le colonne delle matrici U e V corrispondenti ai valori singolari non nulli. Queste sono le informazioni **indispensabili** per la ricostruzione della matrice A. Notiamo che:

$$V_r = [v^{(1)} \dots v^{(r)}] \in \mathbb{R}^{m \times r} \quad U_r = [u^{(1)} \dots u^{(r)}] \in \mathbb{R}^{n \times r} \quad \Sigma_r \in \mathbb{R}^{m \times r}$$

Le matrici U_r e V_r sono **rettangolari** e non più quadrate, esse soddisfano $U_r^T U_r = I$ e $V_r^T V_r = I$ ma nel caso generale in cui $r < \min\{m, n\}$ $U_r^T U_r \neq I$ e $V_r^T V_r \neq I$.

Per rendere l'idea di quanto spazio si guadagna, della U e della V ci teniamo le prime r colonne e della Σ ci teniamo soltanto la sottomatrice $r \times r$



Dimostrazione 37: SVD leggera

Dimostriamo che

$$V\Sigma U^T = V_r \Sigma_r U_r^T$$

Si ha che

$$\mathrm{V}\Sigma U^T = \left[\begin{array}{c|c} V_r & Z \end{array}\right] \left[\begin{array}{cc} \Sigma_r & 0 \\ 0 & 0 \end{array}\right] \left[\begin{array}{c} U_r^T \\ \hline W^T \end{array}\right] = \left[\begin{array}{c|c} V_r & Z \end{array}\right] \left[\begin{array}{c} \Sigma_r U_r^T \\ 0 \end{array}\right] = V_r \Sigma_r U_r^T$$

Inoltre

$$\underbrace{\left[\begin{array}{ccc} v^{(1)} & \dots & v^{(r)} \end{array}\right]}_{V_r} \underbrace{\left[\begin{array}{ccc} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{array}\right]}_{\Sigma_r} \underbrace{\left[\begin{array}{c} {u^{(1)}}^T \\ \dots \\ {u^{(r)}}^T \end{array}\right]}_{U_r^T} = \left[\begin{array}{ccc} v^{(1)} & \dots & v^{(r)} \end{array}\right] \left[\begin{array}{c} \sigma_1 {u^{(1)}}^T \\ \dots \\ \sigma_r {u^{(r)}}^T \end{array}\right] =$$

$$= \sigma_1 v^{(1)} {u^{(1)}}^T + \dots + \sigma_r v^{(r)} {u^{(r)}}^T$$

$$= \sum_{i=1}^{r} \sigma_i v^{(i)} {u^{(i)}}^T$$

Quindi A è una sommatoria di matrici di rango 1

Teorema 30: Norma 2 e Norma di Frobenius per valori singolari

ia $A \in \mathbb{R}^{m \times n}$ e siano $\sigma_1, \sigma_2, \dots, \sigma_p$ tutti i valori singolari di A. Vale che:

$$||A||_2 = \sigma_1 \qquad ||A||_F = \sqrt{\sigma_1^2 + \dots + \sigma_p^2}$$

Nota che consideriamo tutti i valori singolari, anche quelli nulli. In questo modo possiamo considerare anche il caso dove tutti i valori singolari son nulli ($rank(A) = 0$)

Dimostrazione 38

Sia $A = V\Sigma U^T$ una SVD di A. Dato che la norma 2 e la norma di Frobenius sono ortogonalmente invarianti si ha che:

$$||A||_2 = ||V\Sigma U^T||_2 = ||\Sigma||_2 \qquad ||A||_F = ||V\Sigma U^T||_F = ||\Sigma||_F$$

Calcoliamo queste norme per

$$\Sigma = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix}$$

Dalla definizione di norma 2

$$||A||_2 = \sqrt{\rho(A^t A)}$$

Segue che

$$||\Sigma||_2 = \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{||\Sigma x||_2}{||x||_2} = \sigma_1$$

Infatti ($p = \min\{m, n\}$)

$$||\Sigma x||_2 = \sqrt{\sum_{i=1}^{p} (\sigma_i x_i)^2} = \sqrt{\sum_{i=1}^{p} \sigma_i^2 x_i^2} \leq \sqrt{\sum_{i=1}^{p} \sigma_1^2 x_i^2} = \sigma_1 \sqrt{\sum_{i=1}^{p} x_i^2} \leq \sigma_1 \sqrt{\sum_{i=1}^{n} x_i^2} = \sigma_1 ||x||_2$$

Segue che

$$||\Sigma||_2 = \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{||\Sigma x||_2}{||x||} \leq \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\sigma_1 ||x||_2}{||x||_2} = \sigma_1$$

Vale l'uguaglianza $||\Sigma||_2 = \sigma_1$ poiché c'è almeno un vettore per cui questo limite superiore viene raggiunto. Per esempio scegliendo $x = e^{(1)}$ si ha $||x||_2 = 1$ e $||\Sigma x||_2 = ||\Sigma(:, 1)||_2 = \sqrt{\sigma_1^2} = \sigma_1$. ($\Sigma(:, 1)$ signifca la prima colonna di Σ). Risulta poi per la norma di Frobenius

$$||\Sigma||_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} \Sigma_{ij}^2} = \sqrt{\sigma_1^2 + \cdots + \sigma_p^2}$$

Possiamo ottenere una SVD di una matrice simmetrica facilmente a partire dalla diagonalizzazione (mediante trasformazione ortogonale) della matrice:ù

$$A \in \mathbb{R}^{n \times n} \text{ simmetrica } \Rightarrow U^T A U = (\lambda_1, \ldots, \lambda_n) \text{ con } U = [u^{(1)}, \ldots, u^{(n)}] \text{ ortogonale}$$

Ordiniamo autovalori e corrispondenti autovettori in modo che:

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n| \text{ con } |\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_r| > 0$$

Si noti che $A = UDU^T$ **non** è una SVD di A perché gli elementi diagonali di D non sono (in generale) tutti ≥ 0. Scriviamo pertanto:

$$\begin{aligned} A &= U diag(\lambda_1, \ldots, \lambda_2) U^T \\ &= U diag(sgn(\lambda_1), \ldots, sgn(\lambda_r), 1, \ldots, 1) \cdot diag(|\lambda_1|, \ldots, |\lambda_r|, 0, \ldots, 0) U^T \\ &= \underbrace{\left[sgn(\lambda_1) u^{(1)}, \ldots, sgn(\lambda_r) u^{(r)}, u^{(r+1)}, \ldots, u^{(n)}\right]}_{V} |D| U^T \end{aligned}$$

Esercizio Determinare una SVD della matrice simmetrica A

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

Diagonalizziamo A con una trasformazione ortogonale:

$$P_A(\lambda) = \begin{vmatrix} 1-\lambda & 2 \\ 2 & 1-\lambda \end{vmatrix} = (1-\lambda)^2 - 4 = (-1-\lambda)(3-) = (\lambda+1)(\lambda-3)$$

Quindi $D = diag(3, 1-)$ autovalori ordinati per modulo decrescente $|3| > |-1|$. L'autospazio relativo a $\lambda)3$:

$$\ker(A-3I) = \ker \begin{pmatrix} -2 & 2 \\ 2 & -2 \end{pmatrix} = \{-x_1+x_2 = 0, (x_1, x_2) \in \mathbb{R}^2\} =< (1,1) > \rightarrow u^{(1)} = \begin{pmatrix} \sqrt{2}/2, \sqrt{2}/2 \end{pmatrix}^T$$

L'altro autovettore è ortogonale e vale pertanto $u^{(2)} = (-\sqrt{2}/2, \sqrt{2}/2)^T$. Abbiamo quindi:

$$U = \begin{bmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix} \qquad V = [u^{(1)} - u^{(2)}] = \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ \sqrt{2}/2 & -\sqrt{2}/2 \end{bmatrix}$$

Infine

$$A = V|D|U^T = \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ \sqrt{2}/2 & -\sqrt{2}/2 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix}$$

Come le altre fattorizzazioni (LU, QR) anche SVD esprime la matrice come somma di matrici di rango 1, tuttavia nella SVD gli addendi sono **in ordine di importanza** infatti:

- $A_1 = \sigma_1 v_1 u_1^T$ è la matrice di rango 1 più vicina ad A ...
- $A_k = \sigma_1 v_1 u_1^T + \cdots + \sigma_k v_k u_k^T = \sum_{i=1}^{k} \sigma_i v_i u_i^T$ è la matrice di rango k più vicina ad A

Teorema 31: Teorema di Eckart-Young

Se B ha rango k allora

$$||A - A_k|| \leq ||A - B|$$

per ogni norma matriciale unitariamente invariante (come la norma 2 o quella di Frobenius). In soldoni A_k è la migliore approssimazione in assoluto di A avente rango k

Definizione 87: SVD Troncata

Se invece di usare tutti gli r addendi della SVD leggera come:

$$A = \sum_{i=1}^{r} \sigma_i v^{(1)} u^{(i)^T}$$

Si decide di usarne solo alcuni ($l \leq r9$ si ottiene una ricostruzione approssimata di A che chiameremo A_I dove

$$A_I = \sum_{i=1}^{l} \sigma_i v_{(i)} u^{(i)^T} \qquad 0 \leq l \leq r$$

Si ha che

$$A_l = \begin{pmatrix} v^{(1)} \dots v^{(l)} \end{pmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_l \end{bmatrix} \begin{bmatrix} u^{(1)^T} \\ \vdots \\ u^{(l)^T} \end{bmatrix}$$

E' un'approssimazione di A di rango l. Più addendi si aggiungo più A_l si avvicina ad A

Dal teorema di Eckart-Young sappiamo che troncando la decomposizione ai valori singolari al $k-$esimo termine per $k << \min\{m, n\}$ si ottiene la migliore approssimazione di rango k di A rispetto ad una qualsiasi norma di matrice unitariamente invariante. La migliore approssimazione di rango k di A, A_k si ottiene dunque azzerando i valori singolari dal $k+1-$esimo in poi

L'errore commesso approssimando A con A_k può essere quantificato esplicitamente:

$$\min_{rank(B) \leq k} ||A - B||_2 = ||A - A_k||_2 = \sigma_{k+1}$$

$$\min_{rank(B) \leq k} ||A - B||_F = ||A - A_k||_F = \sqrt{\sigma_{k+1}^2 + \dots + \sigma_r^2}$$

Esempio Calcolare una SVD della matrice A, una volta ottenuta la SVD trova un'approssimazione di rango 1.

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

Per prima cosa dobbiamo calcolare i valori singolari, quindi gli autovalori di $A^T A$

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

Ora lo spettro di $A^T A$ è:

$$\sigma(A^T A) = \{2, 1\} \to \sigma_1 = \sqrt{2} \quad \sigma_2 = 1 \quad \Sigma = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$$

Per calcolare U dobbiamo calcolare gli autovettori della matrice A^TA:

$$\lambda = 2 \quad \ker(A^TA - 2I) = \ker \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix} \rightarrow \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow x_2 = 0 \quad span\{(1,0)\}$$

E' già di norma 1 quindi non dobbiamo fare nulla. La seconda colonna sarà il vettore ortogonale ad u_1

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

Ora dobbiamo calcolare la matrice V, in questo esercizio $r = 2$ (numero di valori singolari divesi da 0), quindi per 2 colonne possiamo usare la formula:

$$v^{(i)} = \frac{1}{\sigma_i} A \cdot U^{(i)}$$

Quindi:

$$v^{(1)} = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -\sqrt{2}/2 \\ 0 \\ \sqrt{2}/2 \end{pmatrix}$$

$$v^{(2)} = \frac{1}{1} \begin{pmatrix} -1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

Per la terza colonna dobbiamo calcolarci AA^T e da questa ricavarci il nucleo che conterrà $v^{(3)}$

$$v^{(3)} \in \ker(AA^T) = \{(x_1, 0, x_1) | x_1 \in \mathbb{R}\} = span\{(1,0,1)\}$$

Il nucleo da solo non è la colonna perché non ha norma 1, lo dividiamo quindi per la sua norma

$$v^{(3)} = \begin{pmatrix} \sqrt{2}/2 \\ 0 \\ \sqrt{2}/2 \end{pmatrix} \qquad V = \begin{pmatrix} -\sqrt{2}/2 & 0 & \sqrt{2}/2 \\ 0 & 1 & 0 \\ \sqrt{2}/2 & 0 & \sqrt{2}/2 \end{pmatrix}$$

E' sempre una buona idea svolgere la verifica, tuttavia può darsi che anche con una SVD non corretta io ottenga la A. Ottenuta la SVD dobbiamo trovare un'approssimazione di rango 1, noi sappiamo che:

$$A = \sigma_1 v^{(1)} {u^{(1)}}^T + \sigma_2 v^{(2)} {u^{(2)}}^T$$

Per ottenere la migliore approssimazione avente rango 1 di A tronchiamo al primo addendo:

$$A_1 = \sigma_1 v^{(1)} {u^{(1)}}^T = \sqrt{2} \begin{pmatrix} -\sqrt{2}/2 \\ 0 \\ \sqrt{2}/2 \end{pmatrix} \begin{pmatrix} 1 & 0 \end{pmatrix} = \sqrt{2} \begin{pmatrix} -\sqrt{2}/2 & 0 \\ 0 & 0 \\ \sqrt{2}/2 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 0 \\ 1 & 0 \end{pmatrix}$$

L'errore assoluto di questa approssimazione equivale al primo valore singolo che abbiamo trascurato, cioè σ_2

$$||A - A_1||_2 = \sigma_2 = 1$$

Se volessimo calcolare l'errore relativo ci basta calcolare l'errore assoluto fratto il valore vero

$$\frac{||A - A_1||_2}{||A||_2} = \frac{\sigma_2}{\sigma_2} = \frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2} \approx 0.71$$