Dispensa Analisi Numerica

"Numeri macchina e aritmetica floating point"

X

Sistema Posizionale

Sistema di Numerazione Posizionale

X

Definizione di sistema di numerazione posizionale. Notazione.

X

---

1. Sistema di Numerazione Posizionale

Q. Sia $x \in \mathbb{R}$ e supponiamo di avere $B \in \mathbb{N} \setminus \{0, 1\}$. Come posso rappresentare x con B "numeri"? Soppratutto, come possiamo rappresentarlo in un calcolatore?

Una prima risposta è fornita dal *sistema di numerazione posizionale*, dove ogni *cifra* ha un valore determinato dalla sua *posizione*

#Definizione

> Definizione (numero finito in base B).
>
> Sia $x \in \mathbb{R}$ con cifre finite da $-m$ a n, $B \in \mathbb{N} > 1$.
> Allora scriviamo la *rappresentazione di x in base B* con la seguente:
>
> $$x_B \sim (-1)^s \sum_{k=-m}^{n} d_k B^k$$
>
> Dove $(d_k)_{k \in [-m,n]} \supset \{1, \dots, B-1\}$, $d_n \neq 0$ e $s = 1 - \operatorname{sgn} x$.

Formalizzando su numeri con cifre infinite, ho la definizione ancora più generale:

#Definizione

> Definizione (numeri infinito in base B).
>
> Sia $x \in \mathbb{R}$ con cifre finite da $-m$ a n, $B \in \mathbb{N} > 1$.
> Allora scriviamo la *rappresentazione di x in base B* con la seguente:

$$x_B \sim (-1)^s \sum_{k=0}^{n} d_k B^k + \sum_{k=1}^{+\infty} d_{-k} B^{-k}$$

Dove $(d_k)_{k \in \mathbb{Z}}$ (limitata superiormente) tale che $d_n \neq 0$.

#Osservazione

Osservazione (la buona posizione della serie).

Notiamo che nella definizione ho la serie

$$\sum_{k=1}^{+\infty} d_{-k} B^{-k}$$

Questa converge? Sì, se consideriamo che $(d_k)_k \in l^\infty$ (infatti è sempre limitata da $B-1$), per cui ho

$$\sum_{k=1}^{+\infty} d_{-k} B^{-k} \leq \sum_{k=1}^{+\infty} (B-1) B^{-k} < +\infty$$

La seconda serie converge in quanto $B > 1$

#Osservazione

Osservazione (rappresentazione finita e infinita in basi diverse).

Notiamo che alcuni numeri $x \in \mathbb{R}$ hanno delle rappresentazioni finite e infinite in *basi diverse*.
Esempio: $x = 1/3$ ha rappresentazione infinita in x_{10}, ma finita in x_3 (infatti $x_3 = 0.1$)
Esempio: $x = 1/10$ ha rappresentazione finita in x_{10}, ma infinita in $x_2 = 0.0\overline{0011}$.

X

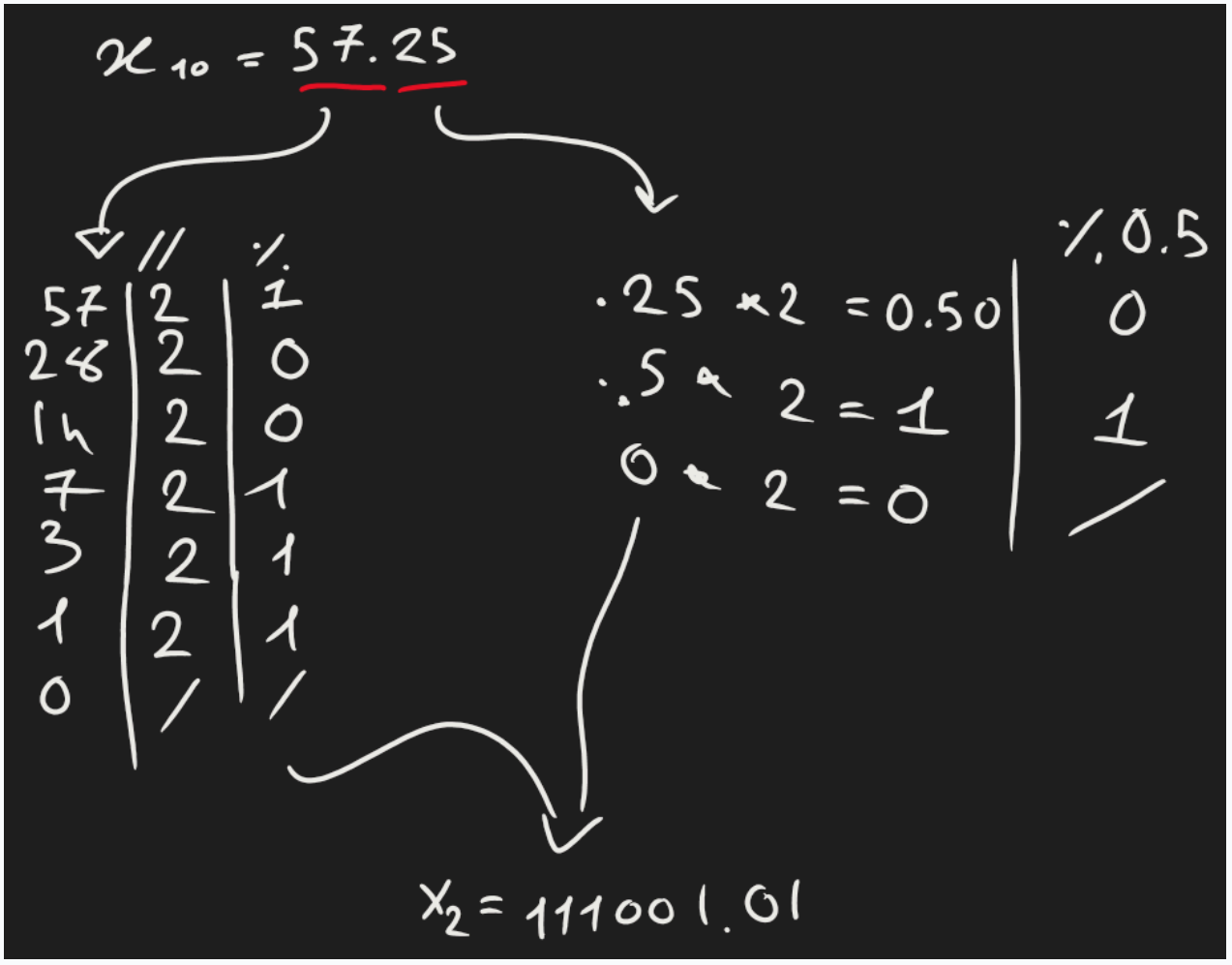
2. Conversione di Basi

Possiamo descrivere degli algoritmi per *convertire* $x_B \leftrightarrow x_{B'}$. Vediamo il caso $B = 2, B' = 10$.

$x_2 \to x_{10}$: Banale, basta esprimerlo come somme di 2^n.
$x_2 \leftarrow x_{10}$: La conversione si effettua in due step: parte intera e frazionaria

- Parte intera: Dividiamo per due la parte intera del numero, prendiamo il quoziente e lo dividiamo per due finchè il resto diventa nullo. La parte intera del numero binario sono i resti letti al contrario.

- Parte frazionaria: Stessa idea ma si moltiplica per due, e vogliamo tenere la parte intera da tenere per la conversione. Si osserva che qui è cruciale fare attenzione a casi in cui sono presenti cicli.



Rappresentazione in Virgola Mobile

Rappresentazione in Virgola Mobile Normalizzata

X

Rappresentazione dei numeri in virgola mobile

X

---

0. Voci correlate

- Sistema di Numerazione Posizionale

1. Rappresentazione in Virgola Mobile Normalizzata

Q. Sia $x \in \mathbb{R}$ e supponiamo di avere $B \in \mathbb{N} \setminus \{0, 1\}$. Come posso rappresentare x con B "numeri"? Soppratutto, come possiamo rappresentarlo in un calcolatore?

Come visto col sistema posizionale dei numeri, una rappresentazione in B di x ha la seguente forma:

$$x_B \sim (-1)^s \sum_{k=0}^{n} d_k B^k + \sum_{k=1}^{+\infty} d_{-k} B^{-k}$$

Tuttavia, questa forma è unica? Ovvero data x_B, abbiamo una sua unica forma? No, posso "spostare" la virgola a piacimento e dunque cambiare gli indici di $(d_k)_{k \leq n}$.
Esempio: sia $x = 2.312$ e $B = 10$. Ottengo le due seguenti possibili rappresentazioni:

$$x_B = 2.312 = 23.12 \cdot 10^{-1} = 0.2312 \cdot 10^{-1}$$

Definiamo dunque la *forma normalizzata della rappresentazione in virgola mobile* di un numero.

#Definizione

> Definizione (forma normalizzata di un numero in virgola mobile).
>
> Sia $x \neq 0$ e $B \geq 2$, possiamo scrivere in *virgola mobile normalizzata* x_B come
>
> $$x_B = (-1)^s B^e \underbrace{\left(\sum_{k \geq 1} d_k B^{-k} \right)}_{:= p} = \operatorname{sgn} x \cdot p B^e$$
>
> Dove $(d_k)_{k \geq 1} \supset \{0, \dots, B-1\}$ t.c. $d_1 \neq 0$ è una successione opportuna, e un coefficiente opportuno (intero).
> Definiamo p la *mantissa* di x, invece e l'*esponente*.

Esempi:

$$x_B = 0.2312 \cdot 10^{-1}$$

è in virgola mobile normalizzata, con mantissa 0.2312 e esponente -1.

#Osservazione

> Osservazione (l'unicità).
>
> Notiamo che la componente cruciale della definizione è $d_1 \neq 0$, che garantisce l'*unicità* della forma normalizzata.

Definizione dei Numeri Macchina

Definizione di Numeri Macchina

Definizione di insieme di numeri macchina. Esempio. Convenzioni standard per numeri macchina (FP32, FP64, FP16).

X

---

0. Voci Correlate

- Sistema di Numerazione Posizionale
- Rappresentazione in Virgola Mobile Normalizzata

1. Definizione di Numero Macchina

Q. Sia $x \in \mathbb{R}$ e supponiamo di avere $B \in \mathbb{N} \setminus \{0, 1\}$. Come posso rappresentare x con B "numeri"? Soppratutto, come possiamo rappresentarlo in un calcolatore?

Una risposta alla domanda finale è data dalla nozione di *numeri macchina*. Come abbiamo visto precedentemente, dato $x \in \mathbb{R}$ e una base $B \geq 2$, abbiamo la rappresentazione in virgola mobile normalizzata:

$$x_B = (-1)^s B^e \left(\sum_{k \geq 1} d_k B^{-k} \right)$$

Tuttavia abbiamo un paio di problemi principali:

- Questo presuppone una *successione* su $n = 1, 2, \dots, +\infty$; nei calcolatori abbiamo un numero limitato di *bits*.
- Inoltre $e \in \mathbb{Z}$, che presuppone una serie di possibilità "infinita"; pertanto devo trovare un range di *"troncamento"* per e, ossia dei "lower" o "upper" bound (lower per numeri "vicini" al zero, upper per numeri "lontani" da 0)
  Sia dunque $1 \leq t \leq +\infty$ e $L, U \in \mathbb{Z}$ con $L < 0, U > 0$. Andremo a definire lo spazio dei numeri macchina come segue:

#Definizione

> Definizione (numeri macchina di parametri B, t, L, U).
>
> Sia $\mathbb{N} \ni B \geq 2$ (generalmente $B = 2$) e $t, L, U \in \mathbb{Z}$ tali che $t \geq 1$, $L < 0$ e $U > 0$. Definiamo l'insieme dei *numeri macchina* $\mathbb{F}(B, t, L, U)$:
>
> $$\mathbb{F}(B, t, L, U) := \left\{ x : x = (-1)^s B^e \sum_{k \leq t} d_k B^{-k} \right\} \cup \{0\}$$
>
> dove $(d_k)_{k \leq t} \supset \{\mathbb{N}_0 < B\}$ è una sequenza finita tale che $d_1 \neq 0$ e l'esponente e è limitata da L, U (ossia $e \in [L, U]$).
>
> (Nota: $\mathbb{F}$ sta per *"floating"*)

Scriveremo $x \in \mathbb{F}(B,t,L,U)$ come

$$x = (-1)^s \cdot (0.d_1 d_2 \dots d_t) \cdot B^e$$

#Osservazione

Osservazione (l'elemento zero).

Notiamo che dobbiamo includere in una maniera arbitraria l'elemento 0. Infatti essa non ha nessuna rappresentazione in virgola mobile, in quanto una sua successione associata $(d_k)_k$ ha sempre valori nulli (e pertanto non può soddisfare $d_1 \neq 0$)

#Osservazione

Osservazione (proprietà dei numeri macchina).

Notiamo innanzitutto che $\mathbb{F}(B,t,L,U)$ è un insieme finito, con la seguente cardinalità:

$$\#(\mathbb{F}(B,t,L,U)) = \underbrace{2}_{s} \cdot \underbrace{(B-1)B^{t-1}}_{\text{combinazioni su } p} \cdot \underbrace{(|L|+|U|+1)}_{e} + \underbrace{1}_{\{0\}}$$

Osserviamo che data $\mathbb{F}(B,t,L,U)$ ha i massimi e minimi in modulo dati da:

$$\min |\mathbb{F}(B,t,L,U)| = \underbrace{B^{-1}}_{0.1} \cdot B^L = B^{L-1}$$

$$\max |\mathbb{F}(B,t,L,U)| = \underbrace{(1 - B^{-(t+1)})}_{0.(B-1)(B-1)\dots(B-1)} \; B^U$$

X

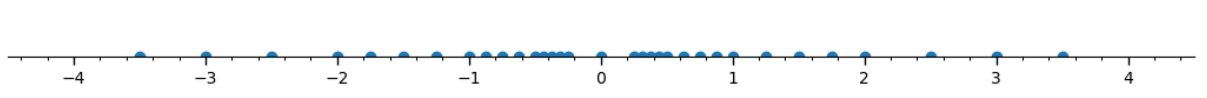
2. Esempi di Numeri Macchina

Esempio. Prendiamo $\mathbb{F}(2,3,-1,2)$. Ovvero ho l'insieme di numeri di base 2 con 3 cifre per la mantissa (normalizzata), con esponenti $e = -1, 0, 1, 2$.

- Mantisse possibili: ho $(2-1)\cdot(2^{3-1}) = 2^2 = 4$ combinazioni possibili (ho che la prima parte si "distacca" in quanto ho il vincolo $d_1 \neq 0$). Infatti ho le seguenti mantisse possibili: $(1,0,0)$; $(1,0,1)$; $(1,1,0)$; $(1,1,1)$.
- Ad ogni mantissa abbiniamo $|U| + |L| + 1$ esponenti possibili (aggiungiamo 1 per includere lo zero). In questo caso ho $|U| + |L| + 1 = 4$
- Ho dunque $4 \cdot 4 = 16$ numeri (senza segno e non nulli) possibili. Tenendo conto il segno ne ho 32
- Osservo che ho i seguenti massimi e minimi (in modulo, ossia ignorando i segni negativi):

$$\max |\mathbb{F}(2,3,-1,2)| = 2^3 \cdot 0.875 \wedge \min |\mathbb{F}(2,3,-1,2)| = 2^{-1} \cdot 0.5$$

Aggiungendo lo zero, concludo che $\#(\mathbb{F}(2,3,-1,2)) = 33$, con una grafica rappresentativa sottostante:



Osserviamo che i numeri sono più addensati quanto più piccoli sono, e la loro separazione aumenta con l'aumentare del modulo; questo è dovuto all'effetto "esponentizale" B^e.

Esempio. (Standard convenzionali dei calcolatori)
Supponiamo di avere un calcolatore con n bit. Per rappresentare $\mathbb{F}(2,t,L,U)$ i bit possono memorizzare il segno, le cifre della mantissa e l'esponente. Ci sono tre standard (di cui uno relativamente recente) per utilizzare questi bit per memorizzare $\mathbb{F}(2,t,L,U)$: FP32 (singola precisione), FP64 (doppia precisione), FP16 (mezza precisione).
Singola Precisione. Usiamo un bit per il segno s, otto per l'esponente e 23 per la mantissa. In questo modo possiamo codificare $\mathbb{F}(2,24,-126,127)$

- Notiamo che $126+127+1 = 254$ e non $256 = 2^8$ esponenti possibili! Infatti, riserviamo altri due bit per *usi speciali*, ovvero casi $\pm\infty$ e NaN
- Usiamo 23 bit per codificare 24 cifre della mantissa (un bit è infatti "nascosto" in quanto abbiamo sempre il vincolo $d_1 = 0$, pertanto viene omesso)

Doppia Precisione. Analogamente con 32 bit codifichiamo $\mathbb{F}(2,53,-1022,1023)$
Mezza Precisione. Analogamente con 16 bit codifichiamo $\mathbb{F}(2,11,L,U)$. Come esercizio calcoliamo L,U. Innanzittutto assumiamo che 1 bit viene usato per il segno, 10 per la mantissa e 5 per l'esponente.

- Pertanto ho $2^5 = 32$ esponenti possibili. Togliendo due "combinazioni" per rappresentare casi speciali, potrei rappresentare $L = -14, U = 15$.
- Codifichiamo 11 cifre della mantissa
- Pertanto abbiamo $2 \cdot 2^{10} \cdot (30) + 1 = 61441$ numeri possibili

Errore di Rappresentazione in Numero Macchina

Errore di Rappresentazione in Numeri Macchina

X

Approssimazione dei reali $\mathbb{R}$ in numeri macchina $\mathbb{F}$. Definizione di troncamento, arrotondamento. Casi speciali di rappresentazione in $\mathbb{F}$: Underflow e Overflow. Define di errore assoluto e relativo in rappresentazione dei reali in macchina. Maggiorazione dell'errore assoluto e relativo. Definizione di precisione di macchina.

X

0. Voci correlato

- Assiomi dei Numeri Reali
- Definizione di Numeri Macchina

1. Troncamento e Arrotondamento di un Reale in Macchina

Sia $x = pB^e \in \mathbb{R}$ tale che la sua mantissa ha più di $t > 0$ cifre. Come possiamo approssimare la sua rappresentazione in $\mathbb{F}(B, t, L, U)$, che chiameremo (con un leggero abuso di notazione) $\mathrm{fl}\,(x)$? Abbiamo due tecniche: troncamento e arrotondamento.

#Definizione

> **Definizione (troncamento di un numero).**
>
> Sia $x = pB^e \in \mathbb{R}$. Diciamo la troncatura $\mathrm{fl}\, x$ come il numero $\mathrm{fl}\, x \in \mathbb{F}(B, t, L, U)$ dove la nella mantissa p cancelliamo la parte che eccede la t-esima cifra. In altre parole, definiamo arbitrariamente la successione associata a $\mathrm{fl}\, x = B^e \sum_{t \in [1,t]} d_t B^{-t}$ come la stessa di p solo che *"ignoriamo"* tutte le cifre dopo la t-esima.
>
> Denotiamo questa operazione con $\mathrm{tr}_t\, x$.

#Definizione

> **Definizione (arrotondamento di un numero).**
>
> Sia $x = pB^e \in \mathbb{R}$. Diciamo l'arrotondamento alla cifra t-esima $\mathrm{fl}\, x$ come la troncatura della somma
>
> $$\mathrm{round}_t\, x := \mathrm{tr_t} + \frac{B}{2} B^{-(t+1)}$$

#Osservazione

> Osservazione (definizione equivalente dell'arrotondamento).
>
> Notiamo che arrotondare equivale incrementare la cifra t-esima della mantissa d_t se la sua cifra successiva è maggiore di $B/2$; altrimenti rimane invariata e si comporta come una troncatura.

Q. Se un numero non sta nel *"range"* di $\mathbb{F}$, cosa succede?
Vado a definire $\mathrm{fl}\, x$ come $+\infty$ (detto Inf) o 0

#Definizione

Definizione (underflow e overflow).

Sia $x = pB^e \in \mathbb{R}$ e prendiamo $\mathbb{F}(B, t, L, U)$

Se $e > U$, allora definiamo $\text{fl}\, x := (\text{sgn}\, x)(+\infty)$
Se $e < L$, allora definiamo $\text{fl}\, x := 0$

X

2. Maggiorazione del Troncamento e dell'Arrotondamento

Q. Abbiamo due modi diversi per rappresentare una mappa $\text{fl} : \mathbb{R} \longrightarrow \mathbb{F}$. Quale una delle due tecniche usare?
A. Vediamo che preferiremo usare l'*arrotondamento*, in quanto questa tecnica va a *"minimizzare"* degli errori. Quantifichiamo adesso il tutto.

#Definizione

Definizione (errore assoluto e relativo).

Sia $x \in \mathbb{R}$ e $x^* \in \mathbb{F}$ una *"approssimazione"* di x. Definiamo:

L'*errore assoluto* come la distanza $\|x - x^*\|$
L'*errore relativo* come la distanza normalizzata dal modulo di x:

$$\frac{\|x - x^*\|}{\|x\|}$$

(*nota: la nozione di errore relativo si applica per $x \neq 0$. Tuttavia questa "va bene" in quanto includiamo arbitrariamento lo zero nei numeri macchina*)

Preferiremo l'*errore relativo* come una metrica per quantificare l'errore, in quanto è in grado di darci un'*idea* della scala di *errore commesso*.

#Teorema

Teorema (maggiorazione degli errori assoluti per troncamento e arrotondamento).

Sia $x = pB^e \in \mathbb{R}$. Allora valgono che:

$$\|x - \text{tr}_t\,(x)\| \leq B^{-t}B^e$$

(*maggiorazione dell'errore assoluto per troncamento*)

$$\|x - \text{round}_t\,(x)\| \leq \frac{B}{2}B^{-(t+1)}B^e$$

(*maggiorazione dell'errore assoluto per arrotondamento*)

#Dimostrazione

DIMOSTRAZIONE del Teorema 6

Maggiorazione dell'errore assoluto per troncamento: Siano $p, \bar{p}$ le mantisse associate a $x, \operatorname{tr} x$. Calcolando la loro differenza $|p - \bar{p}|$ in termini di valori assoluti ottengo un numero del tipo

$$0.0 \dots 0 d_{t+1} d_{t+2} \dots$$

Certamente possiamo maggiorare $\forall t, d_t \leq B - 1$
Pertanto

$$0.0 \dots 0 d_{t+1} d_{t+2} \dots \leq 0. \dots 0 (B-1)(B-1) \dots$$

Che in sistema posizionale equivale alla serie

$$\sum_{n=t+1}^{+\infty} (B-1) B^{-n}$$

Questa certamente converge, infatti $B > 1$ e dunque viene maggiorata da una serie geometrica che è convergente sse $B > 1$. Calcoliamo la somma della serie, considerando la ridotta della serie geometrica (Esempi di Induzione > ^98ba76):

$$\begin{aligned}
\sum_{n=t+1}^{+\infty} (B-1) B^{-n} &= (B-1) \sum_{n=t+1}^{+\infty} \frac{1}{B^n} \\
&= (B-1) \left(\sum_{n} \frac{1}{B^n} - \sum_{n=0}^{t} \frac{1}{B^n} \right) \\
&= (B-1) \left(\frac{1}{1 - \frac{1}{B}} - \frac{1 - \frac{1}{B^{t+1}}}{1 - \frac{1}{B}} \right) \\
&= (B-1) \left(\frac{B^{-(t+1)}}{\frac{B-1}{B}} \right) \\
&= (B-1) \left(B^{-(t+1)} B (B-1)^{-1} \right) = B^{-t}
\end{aligned}$$

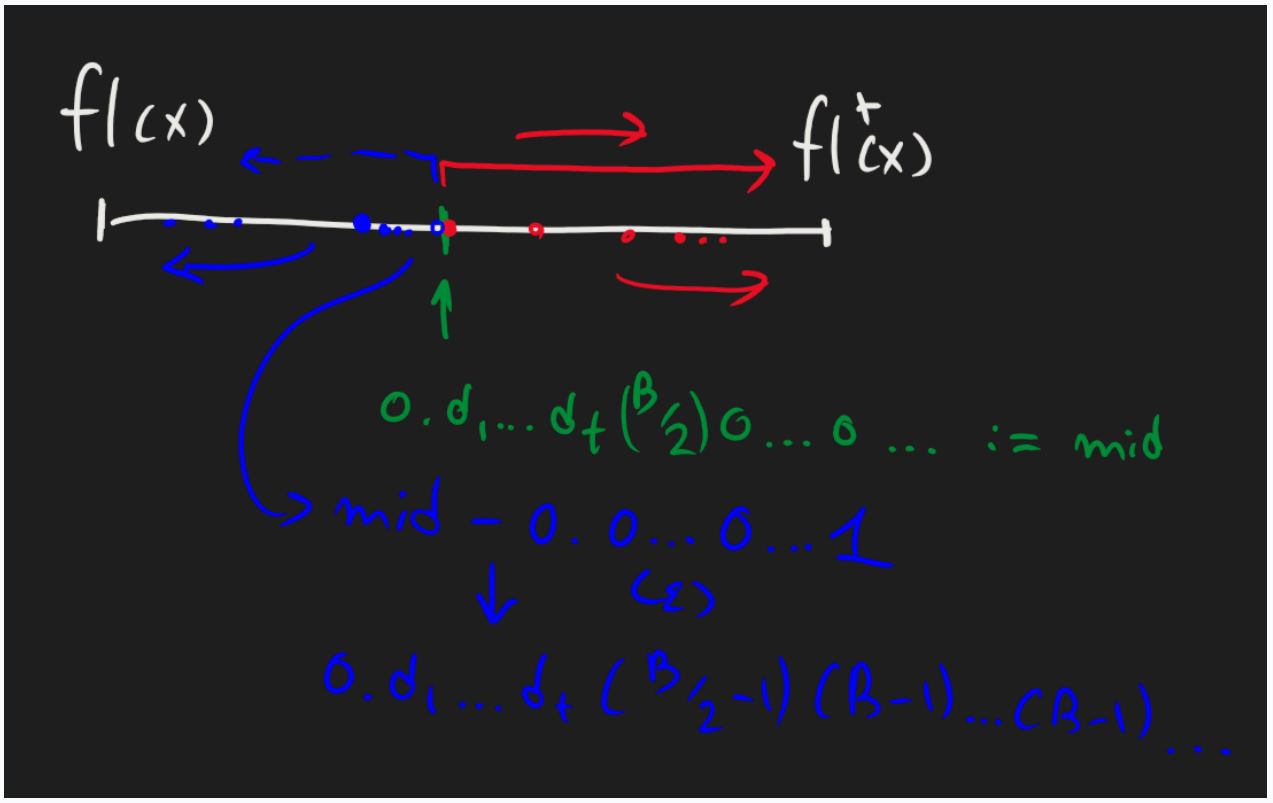
Considerando gli esponenti x, x^* sono uguali (ovvero hanno la stessa p), posso raccogliere per B^e e concludere.
Maggiorazione dell'errore assoluto per arrotondamento: Si tratta di fare un ragionamento analogo, ovvero considerando la differenza delle mantisse $|p - \bar{p}|$. Questa volta abbiamo che

$$\bar{p} = 0.d_1 d_2 \dots \tilde{d}_t$$

Se $d_{t+1} \geq \frac{B}{2}$ allora *"consideriamo"* il numero successivo nell'insieme dei numeri macchina. Pertanto l'errore massimo che posso commettere è quando la mantissa p è proprio

$$0.d_1 \dots d_t (B/2) 0 \dots$$

(per convincersi di questo ragionare che x è approssimato da fl, fl^+ e nel mezzo giace il numero in cui la mantissa è t.c. $d_t = B/2$)



Per cui la differenza delle mantisse è proprio

$$|p - \bar{p}| \leq \frac{B}{2} B^{-(t+1)}$$

Ricordiamo che se $d_{t+1} < \frac{B}{2}$ allora facciamo un semplice troncamento col fattore aggiuntivo. Posso stare in questo caso e commettere il maggior errore se ho una mantissa del tipo $p = 0.d_1 \dots d_t (B/2 - 1)(B-1) \dots$ (questo non è altro che $0.d_1 \dots d_t (B/2) 0 \dots$ sottratto per un numero "infinitesimo")

$$|p - \bar{p}| \leq 0.0 \dots 0 (B/2 - 1)(B-1)(B-1) \dots$$

Ovvero abbiamo di nuovo la serie geometrica

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

Come sempre possiamo verificare che questa converge con somma

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

Che conclude. ■

#Teorema

> Teorema (maggiorazione dell'errore relativo).

Sia $x = pB^e \in \mathbb{R}$. Allora valgono che:

$$\frac{\|x - \mathrm{tr}_t(x)\|}{\|x\|} \leq B^{1-t}$$

(*maggiorazione dell'errore relativo per troncamento*)

$$\frac{\|x - \mathrm{round}_t(x)\|}{\|x\|} \leq \frac{1}{2}B^{1-t}$$

(*maggiorazione dell'errore relativo per arrotondamento*)

#Dimostrazione

DIMOSTRAZIONE del Teorema 7

Sia $x = pB^e \in \mathbb{R}$. Consideriamo che la sua mantissa è una forma del tipo

$$p = 0.d_1 d_2 \dots d_t \dots$$

Questa sicuramente è più grande di $0.1 = B^{-1}$. Pertanto

$$\|x\| \geq B^{-1}B^e$$

Allora applichiamo il la *maggiorazione dell'errore assoluto* considerando questa minorazione (che diventa una maggiorazione in quanto si trova nel denominatore) e concludiamo. ■

#Osservazione

Osservazione (Conviene sempre l'arrotondamento).

Notiamo che l'errore relativo per arrtondamento è sempre più piccolo dell'errore relativo per troncamento. Pertanto definiamo canonicamente la seguente associazione

#Definizione

Definizione (rappresentatore in macchina).

Sia $x = (-1)^s pB^e \in \mathbb{R}$. Allora $\mathrm{fl}_{t,L,U} : \mathbb{R} \longrightarrow \mathbb{F}(B,t,L,U) \cup \{\infty\}$ come la seguente funzione:

$$\mathrm{fl}(x) := \begin{cases} \infty, e > U \\ 0, e < L \\ \mathrm{round}_t(x), \text{ alt.} \end{cases}$$

#Definizione

Definizione (precisione di macchina).

Per ogni insieme di numeri macchina $\mathbb{F}(B,t,L,U)$ definiamo la *precisione di macchina* (ovvero una metrica che esprime il *"massimo errore in fase di rappresentazione"*) come

$$\mathbf{u} := \frac{1}{2}B^{1-t}$$

($\mathbf{u}$ sta per *"unit roundoff"*)

#Esempio

Esempio (esempi di precisione di macchina).

FP16 (ossia $\mathbb{F}(2,24,-126,127)$) ha precisione di macchina
$\mathbf{u} = \frac{1}{2}2^{1-24} = 2^{-24} \approx 5.96 \cdot 10^{-8}$.
FP32 (ossia $\mathbb{F}(2,53,-1022,1023)$) ha precisione di macchina $\mathbf{u} = 2^{-53}$
La Half-precision (ossia $\mathbb{F}(2,11,-14,15)$) ha precisione di macchina $\mathbf{u} = 2^{-11}$

Lo standard ANSI IEEE-754r

Standard ANSI IEEE-754r

X

Lo standard ANSI IEEE-754r per rappresenatare i numeri macchina nei calcolatori. Mantissa col bit nascosto, esponente traslato.

X

0. Voci correlate

- Definizione di Numeri Macchina

1. Lo Standard ANSI IEEE-754r

Supponiamo di avere $x \in \mathbb{F}(2,t,L,U)$ e una macchina che contiene un numero finito di bit. Come possiamo codificare il numero x all'interno della macchina?

Lo standard ANSI IEEE-754r ci fornisce uno standard riconosciuto per rappresentare un numero macchina a due bit in un calcolatore

- Rappresentiamo 0 con la stringa vuota $0\ldots0$ (underflow)
- Sia $x \neq 0$: allora secondo lo standard lo scriviamo come

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

Notiamo che questa convenzione è diverso dal modello matematico. Infatti:

$$1 + f = 1.d_1d_2 \dots d_\tau = 0.d_1d_2 \dots d_td_{t+1}$$

Siccome stiamo lavorando in $B = 2$, lasciamo sott'inteso che $d_1 = 1$ e pertanto lo omettiamo, dandoci così un *"bit gratuito"* in più.
Inoltre codifichiamo l'esponente con $e^* - \beta$, dove β si dice il numero *"bias"* che determina la traslazione di e^*. Ricordandosi che serviamo due numeri dell'esponente per ∞ e Nan, abbiamo che $\beta = \frac{|L|+|U|+1}{2}$ e quindi ho $e^* \in (0, |L| + |U| + 2)$.

- In particolare:
  - Per $e^* = 0$ ho la codifica di 0 ed eventuali numeri denormalizzati
  - Per $e^* = |L| + |U| + 2$ (il massimo possibile togliendo l'ultimo bit) ho ∞ e Nan. Se la mantissa è nulla, allora è $+\infty$ altrimenti è NaN.

Esempio:
Vogliamo codificare $x = 10.25$. In binario questa è 1010.01, che in forma normalizzata diventa $0.101001 \cdot 10^4$. Supponiamo di essere in FP32, pertanto in ANSI IEEE-754r abbiamo:

- $s = 0$ (il numero è positivo)
- $f = 0.01001$ (la mantissa senza il bit nascosto)
- $e^* = e + (|L| + |U| + 2) = e + 255 = 258 \sim 100000010$ (l'esponente tenendo conto del bias)

Infine x è codificato dalla stringa

$$\underbrace{0}_{s}\ \underbrace{100000010}_{e^*}\ \underbrace{010010\overbrace{0\dots0}^{19\text{ zeri}}}_{f}$$

Distanza in Numeri Macchina

Distanza in Numeri Macchina

X

Distanza in numeri macchina. Definizione di distanza assoluta e relativa tra un numero f.p. e il suo successore.

X

Voci correlate

- Definizione di Numeri Macchina
- Standard ANSI IEEE-754r

1. Distanza tra Numeri Macchina

Q. Sia $|x| \neq \max \mathbb{F}(B,t,L,U)$ (ossia $\exists x_+$ successore di x). Come si comporta la distanza tra x, x_+?

#Definizione

> Definizione (distanza assoluta tra numero macchina e suo successore).
>
> Sia $|x| \neq \max \mathbb{F}(B,t,L,U)$ e x_+ il successore di x. Definiamo la *distanza assoluta* tra x, x_+ come:
>
> $$\Delta x := \|x - x_+\|_1$$
>
> Analogamente definiamo la *distanza relativa* come
>
> $$\frac{\Delta x}{x} := \frac{\|x - x_+\|}{\|x\|}$$

#Proposizione

> Proposizione (calcolo della distanza assoluta e relativa).
>
> In $\mathbb{F}(B,t,L,U)$ vale che $\forall x = (-1)^s(1+f)B^e$ s.t. $f \neq ((B-1)_t)_{t \leq \tau}$
>
> $$\Delta x = B^{e-\tau}$$
>
> dove τ è la cifra della mantissa secondo lo standard *IEEE 754-r* (Standard ANSI IEEE-754r).
>
> Analogamente la distanza relativa è
>
> $$\frac{\Delta x}{x} = \frac{B^{-\tau}}{p}$$

#Dimostrazione

DIMOSTRAZIONE della Proposizione 2

Si tratta banalmente di considerare che gli esponenti di x, x_+ sono uguali; inoltre le loro mantisse si differenziamo solo per una cifra, ossia se $p = 0.d_1 d_2 \dots d_\tau$, allora $p_+ = 0.d_1 d_2 \dots (d_\tau + 1)$. Pertanto la loro differenza risulta in $0.0 \dots 1 = B^{-\tau}$, concludendo. ■

X

---

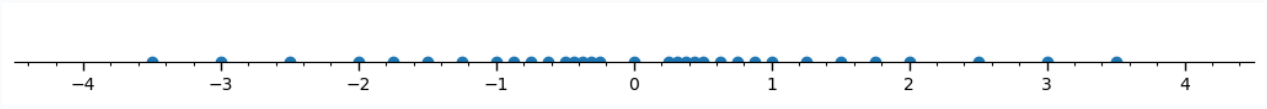
2. Conseguenze della Distanza tra Numeri Macchina

#Osservazione

Osservazione (i numeri macchina sono bucati).

Notiamo che una conseguenza di questa proprietà dei numeri macchina è che $\mathbb{F}(B,t,L,U)$ è un insieme *"bucato"* con dei *"gap"* che diventano più larghi con l'aumentare del numero (in modulo), in quanto la distanza assoluta è determinata dall'esponente $B^{e-\tau}$, dove e rappresentà la *"intensità"* di x.

In altre parole, al contrario di $\mathbb{R}$ che è un insieme denso, $\mathbb{F}(B,t,L,U)$ *non* è un insieme *denso*.



#Osservazione

Osservazione (i numeri macchina hanno distanza relativa periodica).

Inoltre, notiamo che la distanza relativa ha un *andamento periodico*; infatti p è un numero che assume un numero limitato di forme (ossia la combinazione delle sue cifre possibili).

Pertanto, imponendo $\sup p = 1$ otteniamo una maggiorazione della distanza relativa:

$$\frac{\Delta x}{x} = \frac{B^{-\tau}}{p} \leq B^{-\tau}$$

#Definizione

Definizione (massima distanza tra numeri macchina consecutivi).

Definiamo la *massima distanza relativa tra due numeri macchina consecutivi* con

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

#Osservazione

Osservazione (riformulazione alternativa dell'unit roundoff).

Notiamo che la *"unit roundoff"* può essere riformualta in termini di ε_M:

$$\mathbf{u} = \frac{\varepsilon_M}{2}$$

Aritmetica di Macchina

Aritmetica di Macchina

X

Operazioni Aritmetiche in numeri Macchina. Maggiorazione degli errori nelle operazioni macchina. Osservazione: cancellazione numerica. Osservazione: non vale la proprietà associativa.

X

0. Voci correlate

- Definizione di Numeri Macchina
- Errore di Rappresentazione in Numeri Macchina

1. Aritmetica di Macchina

Q. Sappiamo che già nella *rappresentazione* di un numero reale $x \mapsto \mathrm{fl}\, x$ si commette già un errore al massimo quantificabile come $\mathbf{u} = \varepsilon_M/2$. Se effettuiamo delle operazioni tra rappresentazioni di numeri reali in numeri macchina, come si propagheranno questi errori? Come sono confrontabili con gli errori commessi in fase di rappresentazione?

#Definizione

> Definizione (operatore in numeri macchina).
>
> Consideriamo un'operatore $\bullet : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$. Definiamo il suo *corrispondente operatore aritmetico in macchina* come
>
> $$\begin{aligned} \odot : \mathbb{R} \times \mathbb{R} &\longrightarrow \mathbb{F}(B,t,L,U) \\ (x,y) &\mapsto \mathrm{fl}\,(\mathrm{fl}\, x \bullet \mathrm{fl}\, y) \end{aligned}$$
>
> Definiamo canonicamente in questo modo le seguente operazioni: $\oplus, \ominus, \otimes, \oslash$.

Abbiamo che effettuiamo delle *analisi diverse* tra $\cdot$ e $\odot$. Calcoliamo infatti la loro maggiorazione dell'errore.

#Definizione

> Definizione (l'errore di un operatore).
>
> Sia $\odot : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{F}(B,t,L,U)$ un'operatore in numeri macchina. Associamo a loro l'*errore macchina* come la seguente quantità:

$$\varepsilon^{\odot}_{x,y} := \frac{\|(x \cdot y) - (x \odot y)\|}{\|x + y\|}$$

#Teorema

Teorema (maggiorazione degli operatori in numeri macchina).

Siano $x, y \in \mathbb{R}$ e consideriamo $\mathbb{F}(B, t, L, U)$. Valgono le seguenti maggiorazioni:

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

#Dimostrazione

DIMOSTRAZIONE del Teorema 3

Dimostreremo i risultati solo relativi agli operatori $\oplus$ e $\otimes$. Prima di cominciare le dimostrazioni, supponiamo per semplicità che $x, y, x \cdot y \neq 0$ e che $\mathrm{fl}(\mathrm{fl}\, x \cdot \mathrm{fl}\, y) = \mathrm{fl}\, x \cdot \mathrm{fl}\, y$ (ossia non abbiamo bisogno di ulteriori arrotondamenti quando abbiamo ottenuto il risultato teorico).

$\oplus$: Si tratta principalmente di scrivere la definizione per $\varepsilon^{\oplus}_{x,y}$ e di usare la disuguaglianza triangolare per maggiorare.

$$\begin{aligned} \varepsilon^{\oplus}_{x,y} :&= \frac{|(x+y) - (x \oplus y)|}{|x+y|} = \frac{|(x+y) - \cancel{\mathrm{fl}}\,(\mathrm{fl}\, x + \mathrm{fl}\, y)|}{|x+y|} \\ &= \frac{|x - \mathrm{fl}\, x + y - \mathrm{fl}\, y|}{|x+y|} \\ &\leq \frac{|x - \mathrm{fl}\, x|}{|x+y|} \cdot \frac{|x|}{|x|} + \frac{|y - \mathrm{fl}\, y|}{|x+y|} \cdot \frac{|y|}{|y|} =: \varepsilon_x \cdot \frac{|x|}{|x+y|} + \varepsilon_y \cdot \frac{|y|}{|x+y|} \end{aligned}$$

$\otimes$: La dimostrazione è analoga, solo che bisogna usare l'approssimazione $\mathrm{fl}\, y / y \approx 1$ (o WLOG $\mathrm{fl}\, x / x \approx 1$).

$$\begin{aligned} \varepsilon^{\otimes}_{x,y} :&= \frac{|(xy) - (\mathrm{fl}\, x\, \mathrm{fl}\, y)|}{|xy|} \\ &= \frac{|xy + x\,\mathrm{fl}\, y - x\,\mathrm{fl}\, y + \mathrm{fl}\, x\, \mathrm{fl}\, y|}{|xy|} \\ &= \frac{|x(y - \mathrm{fl}\, y) + \mathrm{fl}\, y(x - \mathrm{fl}\, x))|}{|xy|} \\ &\leq \frac{|x| \cdot |y - \mathrm{fl}\, y|}{|x| \cdot |y|} + \frac{|\,\mathrm{fl}\, y\,| \cdot |x - \mathrm{fl}\, x\,|}{|y| \cdot |x|} \\ &\approx \varepsilon_x + \varepsilon_y \end{aligned}$$

che conclude la dimostrazione. ■

#Osservazione

Osservazione (il prodotto e la divisione causano errori "accettabili").

Notiamo che $\varepsilon_{x,y}^{\otimes}$ e $\varepsilon_{x,y}^{\oslash}$ presentano una *combinazione lineare* di ε_x e ε_y, ossia $\mathbf{u}$. Ciò vuol dire che le operazioni del prodotto e della divisione in aritmetica dellam acchina introducono un errore dell'*ordine della precisione di macchina*.

In altre parole, *non* aplifico ulteriormente gli errori commessi.

#Osservazione

Osservazione (la cancellazione numerica).

Tuttavia il discorso cambia radicalmente quando consideriamo $\oplus$. Questa presenta invece il fattore

$$\frac{z \in \{x, y\}}{x + y}$$

Se consideriamo $x, y \in \mathbb{R}$ tali che $x + y \approx 0$ (in particolar modo quando $x \approx -y$) abbiamo che questo fattore viene amplificato in una maniera che non viene più controllata dalla precisione di macchina $\mathbf{u}$, generando delle potenziali instabilità.

Questo fenomeno si chiama *cancellazione numerica*, in quanto si tende a cancellare *cifre* del risultato.

#Esempio

Esempio (caso di cancellazione numerica).

Vediamo il seguente caso in cui succede la *cancellazione numerica*.

Consideriamo $\mathbb{F}(10, 5, L, U)$. Siano $a = 0.73415507$ e $b = 0.73415448$. Naturalmente si ha

$$\mathrm{fl}\, a = 0.73416, \mathrm{fl}\, b = 0.73415$$

Tuttavia $a - b = 0.59 \cdot 10^{-6}$ e $a \ominus b = 10^{-5}$, da cui

$$\varepsilon_{a,b}^{\oplus} = \frac{|0.59 \cdot 10^{-6} - 10^{-5}|}{0.59 \cdot 10^{-6}} = \frac{10^{-6}|0.59 - 10|}{0.59 \cdot 10^{-6}} = \frac{9.41}{0.59} \approx 15.949 = 1595\%$$

Ciò vuol dire che stiamo commettendo un errore in ordine di grandezza di 10^1, che è molto più grande di $\mathbf{u} = 5 \cdot 10^{-5}$! (circa sei volte in ordine di grandezza $10^\bullet$)

2. Non Associatività delle Operazioni

#Lemma

> Lemma (associatività dell'addizione nei reali).
>
> In $\mathbb{R}$ vale che $\forall x, y, z, (x + y) + z = x + (y + z)$

#Teorema

> Teorema (non associatività della somma in aritmetica della macchina).
>
> In $\oplus$ ovvero la somma sull'aritmetica della macchina $\mathbb{F}$ *non* vale la proprietà associativa. Ossia
>
> $$\exists x, y, z \in \mathbb{R} : (x \oplus y) \oplus z \neq x \oplus (y \oplus z)$$

#Dimostrazione

DIMOSTRAZIONE del Teorema 8

Si tratta di trovare dei controesempi che verifichino l'enunciato del teorema: useremo due strategie per trovare x, y, z. Da un lato si può sfruttare la *cancellazione numerica*, che va a causare un'approssimazione così grande da far cambiare i risultati. Dall'altro lato si può sfruttare anche i fenomeni di *overflow* (o *underflow*) nei numeri macchina.
Nel nostro particolare considereremo lo standard FP32, ossia $\mathbb{F}(2, 53, -1022, 1023)$ con $\mathbf{u} = 2^{-53} \approx 1.11 \cdot 10^{-16}$.
Controesempio 1: Siano $x, z = 1$ e $y = 10^{-15}$. Abbiamo che $(x \oplus y) \oplus -z = 1.11 \cdot 10^{-15}$. Questa è dovuta alla cancellazione numerica, infatti sommando $1 \oplus 10^{-15}$ vado a *"perdere"* delle cifre che vengono *"spostate"* a destra e quindi perse quando ci sottraggo $\text{//} \ominus 1$ (questa infatti è una conseguenza della *cancellazione numerica*).
Tuttavia banalmente

$$(1 \ominus 1) \oplus 10^{-15} = 0 \oplus 10^{-15} = 10^{-15}$$

Concludendo il controesempio.
Controesempio 2: Siano $x = 1.0 \cdot 10^{308}$ e $y = 1.1 \cdot 10^{308}$ e $z = -1.001 \cdot 10^{308}$. Ricordando che in FP32 si ha il massimo $1.111\ldots1 \cdot 2^{1023} \approx 2^{1024} \approx \cdot 1.79 \cdot 10^{308}$. Pertanto $x \oplus y = +\infty$. Allora abbiamo che

$$x \oplus (y \oplus z) = 10^{308} \oplus (0.099 \cdot 10^{308}) = 1.099 \cdot 10^{308}$$

Tuttavia

$$(x \oplus y) \oplus z = (+\infty) \oplus z = +\infty$$

Che conclude la dimostrazione. ■

Stabilità degli Algoritmi Numerici

Stabilità degli Algoritmi Numerici

X

Definizione di algoritmo numerico stabile e instabile. Proposizione: algoritmi che presentano la cancellazione numerica sono instabili. Esempi. Controesempio per il verso opposto: algoritmo numerico instabile non dovuto alla cancellazione numerica.

X

0. Voci correlate

- Errore di Rappresentazione in Numeri Macchina
- Aritmetica di Macchina

1. Definizione di Algoritmo Instabile

#Definizione

> Definizione (algoritmo numerico instabile).
>
> Diciamo che un *metodo numerico* (formula, algoritmo) è *stabile* sse questa formula non propaga gli errori dovuti alla rappresentazione dei numeri del calcolatori. Altrimenti è *instabile*.
>
> In altre parole, una formula $f : E \longrightarrow E'$ è *stabile* se vale che
>
> $$\varepsilon^{f}_{x \in E} := \frac{\|f(x) - \operatorname{fl} f(x)\|}{\|f(x)\|} \leq \mathbf{u}$$

X

2. Algoritmi Instabili per la Cancellazione Numerica

#Proposizione

> Proposizione (la cancellazione numerica causa l'instabilità degli algoritmi).
>
> Se un algoritmo presenta fenomeni di *cancellazione numerica*, allora è *instabile*.

Vediamo degli esempi immediati:

#Esempio

Esempio (formula instabile).

La formula

$$f_n(x) := \sqrt{x + \frac{1}{n}} - \sqrt{x}, n \to +\infty$$

è *instabile*. Infatti si ha che $x + \frac{1}{n} \approx x$ per cui $\sqrt{x + \frac{1}{n}} \oplus \sqrt{x}$ presenta un errore incontrollabile.

Infatti, fissato $x = 2$ e $n = 10^{18}$ ho fl $f_n(x) = 0$, presentando dunque l'errore del 100%.

#Osservazione

Osservazione (le formule instabili possono essere stabilizzate).

Notiamo che le formule instabili dovuti alla *cancellazione* possono essere manipolati in modo tale che matematicamente presentino la stessa formula, ma non sono più soggetti alla cancellazione numerica.

Il modo *"standard"* è quello di razionalizzare la formula, ossia f_n diventa

$$f_n^*(x) = \left(\sqrt{x + \frac{1}{n}} - \sqrt{x}\right) \cdot \frac{\sqrt{x + \frac{1}{n}} - \sqrt{x}}{\sqrt{x + \frac{1}{n}} - \sqrt{x}} = \frac{\frac{1}{n}}{\sqrt{x + \frac{1}{n}} + \sqrt{x}}$$

Ovviamente per $(x, n) \in (\mathbb{R} \times \mathbb{N}^{\not 0})$ si ha $f_n(x) = f_n^*(x)$.

#Esempio

Esempio (soluzioni alle equazioni di secondo grado).

Siano $a, b, c \in \mathbb{R} \setminus \{0\}$ tali che il discriminante $\Delta := b^2 - 4ac$ sia positivo ($\Delta \geq 0$).

Allora una sua soluzione è data dalla formula

$$(2a)x_1 = -b + \sqrt{b^2 - 4ac}$$

Questa è affetta da *cancellazione numerica* per $b^2 \gg 4ac$ (ossia $b \gg a, c$) per cui $\sqrt{b^2 - 4ac} \approx \sqrt{b^2} = |b|$.

La formula *"corretta"* è data da

$$(2a)x_1 = \frac{4ac}{b + \sqrt{b^2 - 4ac}}$$

#Esempio

Esempio (successione approssimante di π).

È noto che la seguente successione è un'approssimante di π:

$$(z_n)_{n\in\mathbb{N}} := \begin{cases} 2, n = 1 \\ 2^{n-0.5}\sqrt{1-\sqrt{1-4^{1-n}z_n^2}}, n > 1 \end{cases}$$

Ossia $\lim_n z_n = \pi$.

Questa è instabile per il fatto che $\lim_n 4^{1-n}z_n^2 = 0$ (siccome z_n si comporta come una costante) e dunque ho $1 - 4^{1-n}z_n^2 \approx 1$.

Come sempre, possiamo *"correggerla"* razionalizzando.

X

3. Altri Algoritmi Instabili

Q. Abbiamo visto che la cancellazione numerica implica algoritmo instabile. Ma vale il verso opposto, i.e. ogni algoritmo instabile è dovuto alla cancellazione numerica?
A. La risposta è negativa, fornendo il seguente controesempio

#Esempio

Esempio (algoritmo instabile non per cancellazione numerica).

Supponiamo di voler calcolare la seguente successione di R-integrali:

$$(I_n)_{n\in\mathbb{N}} := \frac{1}{e}\int_0^1 x^n e^x \, \mathrm{d}x$$

Notiamo che $I_0 = 1 - \frac{1}{e} \approx 0.632$.

Per $n \geq 1$ invece possiamo integrare per parti e ottenere la seguente relazione ricorsiva:

$$I_n = \frac{1}{e}\int_0^1 \underbrace{x^n}_{D}\underbrace{e^x}_{I} \, \mathrm{d}x = \frac{1}{e}\left(x^n e^x|_0^1 - n\int_0^1 x^{n-1}e^x \, \mathrm{d}x\right) =: 1 - nI_{n-1}$$

Notiamo che $(I_n)_n$ è una successione decrescente, infatti la sua derivata è sempre negativa. Pertante si ha che certamente $nI_{n-1} \neq 1$, per cui *non si ha la cancellazione numerica*.

Tuttavia, questa formula è comunque *instabile*. Supponiamo che ε_n sia un termine di errore associato all'integrale I_n. Pertanto per la formula iniziale ho

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

Ossia

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

Per induzione ho che

$$\varepsilon_n = -n(\varepsilon_{n-1}) \iff |\varepsilon_n| = n!|\varepsilon_0|$$

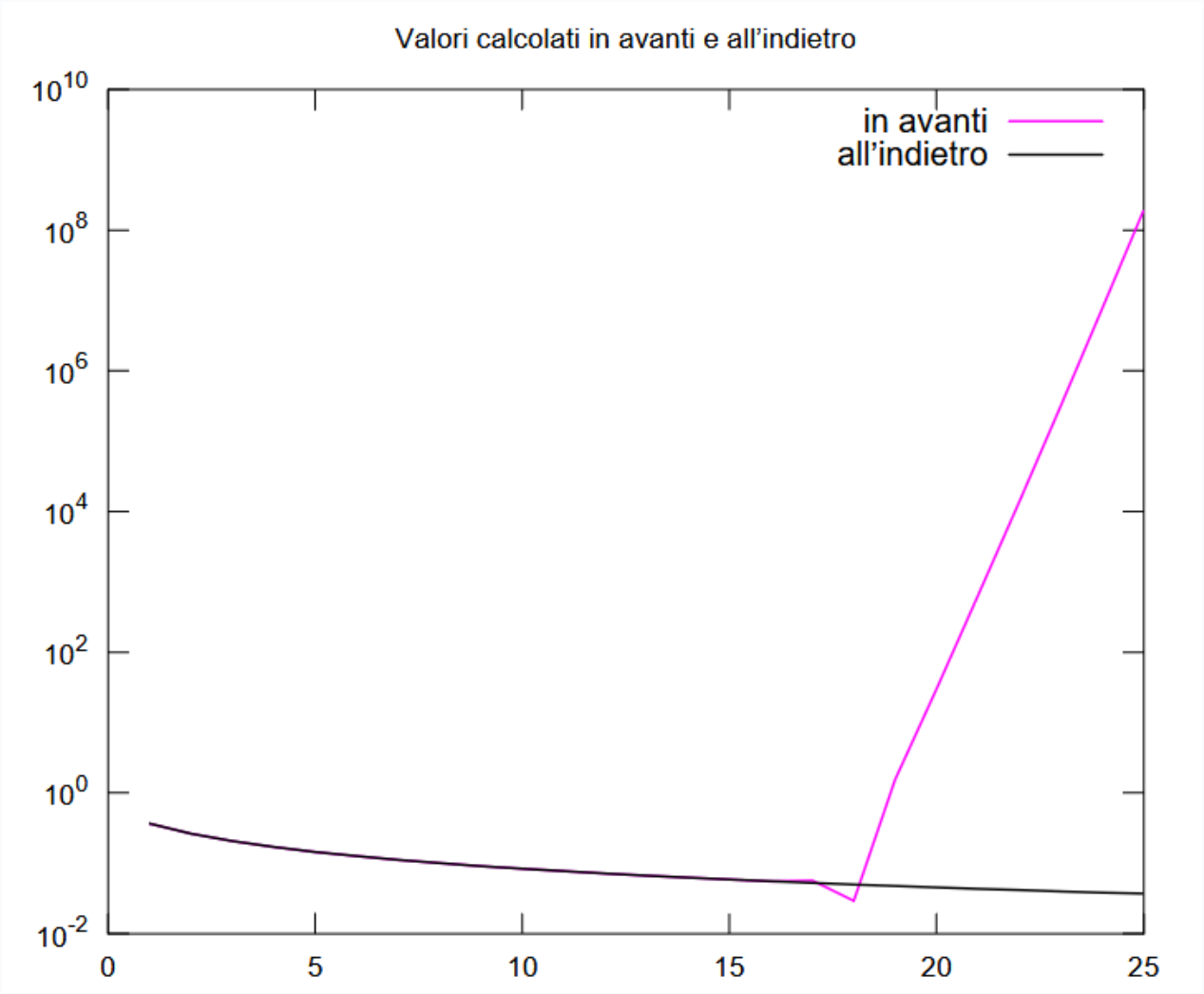
Pertanto ho l'amplificazione dell'errore, che aumenta quanto aumenta $O(n!)$.

Un modo per *"correggere"* la formula è di considerare la successione di ricorrenza *"al contrario"*, ossia calcolo

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

In questo caso vado pure a *"smorzare"* l'errore con l'aumentare di n. Tuttavia, per trovare I_n bisogna *"partire"* da un numero più largo $m \gg n$.

FIGURA 1. (*Grafico del calcolo di* I_n)



Condizionamento dei Problemi

Condizionamento dei Problemi

X

Definizione di condizionamento per problemi. Definizione di numero di condizionamento per equazioni.

X

0. Voci correlate

- Stabilità degli Algoritmi Numerici
- Teorema di Lagrange

1. Condizionamento di un Problema

Gli algoritmi possono essere *instabili*, ovvero risolvendo dei problemi possono propagare degli errori in una maniera incontrollabile.

Adesso vediamo un aspetto *intrinseco* dei problemi, ossia indipendente dall'algoritmo di risoluzione scelto, che risulta comunque *problematico*. Parleremo al *condizionamento di un problema*, ossia la *sensibilità* del problema alle piccole variazioni.

#Definizione

> Definizione (problema mal condizionato).
>
> Un problema si dice *mal condizionato* sse a piccole variazione nei dati corrispondono grandi variazioni nei risultati.
>
> In caso contrario, il problema si dice *ben condizionato*.

Un problema mal condizionato è profondamente problematico per gli *algoritmi numerici*, infatti per quanto un agoritmo numerico possa essere stabile, non potrà mai dare soluzioni corrette ai problemi mal condizionati.

#Esempio

> Esempio (sistema lineare).
>
> Supponiamo di voler risolvere il seguente sistema lineare
>
> $$\begin{pmatrix} 1 & 1 \\ 1001 & 1000 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2 \\ 2001 \end{pmatrix}$$

Certamente come soluzione ha $x, y = 1$. Tuttavia perturbando il coefficiente a_{11} di 0.01, otteniamo

$$\begin{pmatrix} 1.01 & 1 \\ 1001 & 1000 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2 \\ 2001 \end{pmatrix}$$

Che come soluzione ha $x = -0.1111\ldots$ e $y = 2.11222\ldots$ con gli errori relativi pari ha

$$\varepsilon_x = 111.11\%, \varepsilon_y = 111.22\%$$

Entrambi maggiori del 100%.

X

---

2. Numero di Condizionamento per Equazioni

Diremo che il *numero di condizionamento* di un problema è la quantità che misura il *grado di sensibilità* di un problema rispetto alle piccole variazione nei dati.

Prendiamo il caso particolare di valutare funzioni $y = f(x)$. Sia Δx la *pertubazione* in x, vogliamo calcolare Δy il cambiamento delle soluzioni (ovvero l'errore assoluto). Assumendo che $f \in \mathcal{C}^1(I := [x, x + \Delta x])$, per il *Teorema di Lagrange* ho

$$\exists \xi \in I : f'(\xi) = \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

Moltiplicando per $\Delta x \neq 0$ ho

$$\underbrace{f(x + \Delta x) - f(x)}_{\Delta y} = \Delta x f'(\xi)$$

Voglio ottenere l'*errore relativo*, per cui divido per $y = f(x)$:

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

Portando al limite $\Delta x \to 0$ ho che

$$\frac{|x f'(\xi)|}{|f(x)|} \to \frac{|x f'(x)|}{|x|}$$

Infatti l'intervallo ξ è compreso nell'intervallo $I = [x + \Delta x]$, per cui $\xi \to x$. Allora ho

$$\frac{\Delta y}{y} \xrightarrow{\Delta x \to 0} \frac{|x f'(x)|}{|f(x)|} \cdot \frac{|\Delta x|}{|x|}$$

Ossia l'errore relativo di $f(x)$ dipende da x parametrizzato da un *coefficiente*, che chiameremo *numero di condizionamento*.

#Definizione

Definizione (numero di condizionamento).

Sia $f : E \subseteq \mathbb{R} \longrightarrow \mathbb{R}$, sia $x_0 \in E$ tale che f è localmente derivabile in x_0. Diremo il *numero di condizionamento* di f in x_0 come il numero

$$K^f(x_0) := \frac{|x_0 f'(x_0)|}{|f(x_0)|}$$

#Esempio

Esempio (esempio del numero di condizionamento).

Sia $f(x) = (1-x^2)^{1/2}$. Siccome f è differenziabile L-q.o. in $[-1,1]$ possiamo calcolare analiticamente il suo *numero di condizionamento* per $\forall x \in [-1,1]$.

$$K^f(x) = \frac{|xf'(x)|}{|f(x)|} = \frac{x \cdot \frac{-x}{\sqrt{1-x^2}}}{|\sqrt{1-x^2}|} = \frac{x^2}{1-x^2}$$

Notiamo che per $x \to \pm 1$ ho che

$$\lim_{x \to \pm 1^{\mp}} K^f(x) = \lim_{x \to \pm 1^{\mp}} \frac{x^2}{1-x^2} = \lim_{''} \frac{1}{\frac{1}{x^2} - 1} = +\infty$$

Per cui il problema è mal condizionato per $x = -1, 1$. ■

X

"Calcolo dei Zeri delle Funzioni"

X

Calcolo dei Zeri delle Funzioni

Calcolo dei Zeri delle Funzioni

X

Calcolo dei Zeri delle Funzioni: problema, richiami preliminari di risultati di Analisi.

X

0. Voci correlate

- Teoremi sulle funzioni continue

1. Definizione di Zero di Una Funzione

PROBLEMA. Sia $f \in \mathcal{C}^0(I \Subset \mathbb{R}; \mathbb{R})$. Vogliamo trovare i valori $\alpha \in I$ tali che $f(\alpha) = 0$. Ossia, voglio trovare gli elementi dell'insieme

$$f^{\leftarrow}(\{0\})$$

#Definizione

Definizione (zero di una funzione).

Sia $f \in \mathcal{C}^0(I \Subset \mathbb{R}; \mathbb{R})$. Un *zero* (o *radice*) di f è un numero $\alpha \in I$ tale che $f(\alpha) = 0$.

#Esempio

Esempio (zeri di funzioni).

Notiamo subito che le *funzioni* possono avere *uno* o più zeri. Infatti ho che

$$f(x) = \cos(\log(1/x))$$

Ha infiniti zeri, tutti individuati i numeri x come

$$\log(1/x) = \mathbb{Z}\frac{\pi}{2}$$

Ovvero

$$x = e^{\frac{1}{2}\pi - \mathbb{Z}\pi}$$

#Definizione

Definizione (zero semplice).

Un zero α per una funzione $f \in \mathcal{C}^1$ si dice *semplice* sse vale che

$$f(\alpha) = 0 \wedge f'(\alpha) \neq 0$$

Possiamo estendere la nozione di *zero semplici* sulle derivate di ordine k-esimo, chiameremo questo concetto la *molteplicità* del zero α

#Definizione

Definizione (molteplicità di uno zero).

Sia $k \in \mathbb{N} \cup \{+\infty\}$ tale che $f \in C^{\geq k}$. Un zero α si dice di avere *molteplicità* $n \leq k$ sse n è l'*indice della prima derivata che non si annulla in* α, ossia

$$n = \min\left\{n \leq k : f^{(n)}(\alpha) \neq 0\right\}$$

Denotiamo la molteplicità del zero con r_α.

#Esempio

Esempio (molteplicità di uno zero).

Sia f definita come

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

Chiaramente $\alpha = 0$ è uno zero. Studiamo la sua molteplicità.

Vedo prima di tutto che $f \in \mathcal{C}^\infty$, che non ci dice molto se non il fatto che possiamo procedere a calcolare le derivate senzan essun problema.

$$\begin{aligned}
k = 1 &: f'(x) = -\sin x + x + x^4 \implies f'(\alpha) = 0 \\
k = 2 &: f''(x) = -\cos x + 1 + 4x^3 \implies f''(\alpha) = 0 \\
k = 3 &: f'''(x) = \sin x + 12x^2 \implies f'''(\alpha) = 0 \\
k = 4 &: f^{(\mathrm{iv})}(x) = \cos x + 24x \implies f^{(\mathrm{iv})}(\alpha) = 1 \neq 0
\end{aligned}$$

Pertanto $\alpha = 0$ ha molteplicità 4.

#Osservazione

Osservazione (significato della molteplicità).

La molteplicità di uno zero è importante per gli *algoritmi ricorsivi*.

#Osservazione

Osservazione (molteplicità infinita).

Notiamo che di solito, con funzioni *"decenti"*, si ha che ogni zero ha molteplicità finita.

Un caso dove si ha *zeri* con *molteplicità infinita* è la funzione banale $f(x) = 0$, oppure con la funzione $f(x) = xe^x$, che secondo Taylor non è altro che l'espansione della serie

$$f(x) = xe^x = x\sum_n \frac{x^n}{n!} = \sum_n \frac{x^{n+1}}{n!}$$

Da cui si ha che il zero $\alpha = 0$ ha molteplicità infinita, in quanto stiamo derivando un polinomio con termini infiniti. ■

2. Risultati Preliminari di ANALISI

Q. Data una funzione reale $f : I \Subset \mathbb{R} \longrightarrow \mathbb{R}$, quando vale che $\exists \alpha \in I : f(\alpha) = 0$? Quando vale invece l'unicità $\exists! \alpha \in I$?
A. Un paio di risultati classici di *Analisi* ci danno delle risposte

#Teorema

> **Teorema 6 (degli zeri).**
>
> Sia $f : [a, b] \longrightarrow \mathbb{R}$, f *continua* nel suo dominio. Sia $f(a) < 0$, $f(b) > 0$ oppure $f(a) > 0 \wedge f(b) < 0$, cioè sono di segni *discordi* (ovvero $f(a)f(b) < 0$).
> Allora
>
> $$\exists \xi \in]a, b[: f(\xi) = 0$$
>
> In parole deve esiste un valore ξ che *"taglia"* attraverso la linea orizzontale delle ascisse.

#Teorema

> **Teorema (l'esistenza unica dei zeri).**
>
> Se $f \in \mathcal{C}^0(I)$ ha *zeri* ed è *strettamente monotona* allora
>
> $$\exists! \alpha \in I : f(\alpha) = 0$$

#Esempio

> **Esempio (esempio di funzione con zero unico).**
>
> Sia f definita come
>
> $$f(x) = 2x^2 + 2 \ln x - \frac{1}{x}$$
>
> Sia $I = [0.1, 1]$. Notiamo che $f(0.1) = -12.28$ e $f(1) = 1$, pertanto per Bolzano $\exists \alpha \in [0.1, 1] : f(\alpha) = 0$.
> Dimostriamo che questa è anche unica.
>
> Calcoliamo f':
>
> $$f'(x) = 4x + \frac{2}{x} + \frac{1}{x^2}$$

> Chiaramente per $x > 0$ la derivata è sempre positiva, da cui f è *strettamente crescente*. Pertanto vale che α è l'*unico* zero di f in $[0.1, 1]$

Algoritmi Numerici Iterativi

Algoritmi Numerici Iterativi

X

Definizione di algoritmo numerico iterativo. Convergenza globale e locale di un algoritmo iterativo, ordine di convergenza e pseudo-ordine di convergenza.

X

1. Definizione di Algoritmo Numerico

Supponiamo di aver dimostrato che $f : I \longrightarrow \mathbb{R}$ ha un *zero unico*. Come facciamo a trovarlo? Un modo per farlo è approssimare α con un *algoritmo numerico*.

#Definizione

> Definizione (algoritmo numerico iterativo).
>
> Un *metodo iterativo* è una procedura che genera una soluzione a *partire da* ≥ 1 *valori iniziali* e da ≥ 1 *termini precedenti*. Ossia, abbiamo una successione $(x_k)_{k\geq 0}$ definita per induzione.
>
> Denotiamo un *algoritmo numerico* con la sua successione associata $(x_k)_k$.

X

2. Convergenza di Algoritmi Numerici

Quando un algoritmo numerico è *"buono"*?

#Definizione

> Definizione (algoritmo numerico convergente).
>
> Un *metodo iterativo* $(x_n)_n$ si dice *convergente ad* α sse vale il limite
>
> $$\lim_k x_k = \alpha$$
>
> Oppure se vale che lo scarto in L1 va a 0:

$$\lim_k |\alpha - x_k| := \lim_k \varepsilon_k = 0$$

#Definizione

Definizione (convergenza locale).

Un *metodo iterativo* $(x_k)_k$ converge *localmente ad* α sse esiste un intorno $U(\alpha)$ di α, tale che per $x_0 \in U(\alpha)$ allora $(x_k)_k$ converge ad α.

X

3. Convergenza dei Metodi Iterativi

#Definizione

Definizione (ordine di convergenza dei metodi iterativi).

Sia $(x_k)_k$ un metodo iterativo convergente ad α. Se $\exists C > 0$ tale che esiste il limite

$$\lim_k \frac{|\varepsilon_{k+1|}}{|\varepsilon_k|^p} = C$$

Allora $(x_k)_k$ *converge ad* α *con ordine di convergenza* p.

#Definizione

Definizione (convergenza lineare, quadratica e superlineare).

Sia $(x_k)_k$ un metodo iterativo convergente ad α.

Se converge ad α con ordine di convergenza $p = 1$, allora tale convergenza si dice *lineare*
Similmente se $p = 2$ è *quadratica*, e se $p > 1$ è *superlineare*.

Matematicamente questo è un buon concetto per definire la nozione di *"velocità"* della convergenza di un algoritmo. Tuttavia, praticamente è più utile approssimarlo togliendo il limite. Ovvero ho la seguente definizione di *pseudo convergenza*.

#Definizione

Definizione (pseudo ordine di convergenza dei metodi iterativi).

Sia $(x_k)_k$ un metodo iterativo convergente ad α. Se vale la stima per cui esiste $C > 0$ tale che

$$\frac{|\varepsilon_{k+1}|}{|\varepsilon_k|^p} \simeq C \iff |\varepsilon_{k+1}| \simeq C|\varepsilon_k|^p$$

Allora il metodo $(x_k)_k$ ha *pseudo convergenza di ordine* p.

#Osservazione

Osservazione (abbattimento dell'errore e pseudo ordine di convergenza).

Noto che se vogliamo assicurare che l'errore $|\varepsilon_k|_k$ si riduca allora:

- Per metodi iterativi *pseudo lineari*, vogliamo che la costante C sia minore di 1.
- Invece per metodi iterativi *pseudo superlineari*, non è necessaria nessuna condizione su C; è sufficiente che l'errore iniziale $|\varepsilon_0| < 1$, che è sempre certa essere sempre vera

Metodo della Bisezione

Metodo della Bisezione

X

Metodo della Bisezione per il calcolo dei zeri. Idea dell'algoritmo, definizione, pseudocodice e dimostrazione della convergenza globale. Pseudo-ordine di convergenza della bisezione. Test dell'arresto per il residuo (pesato).

X

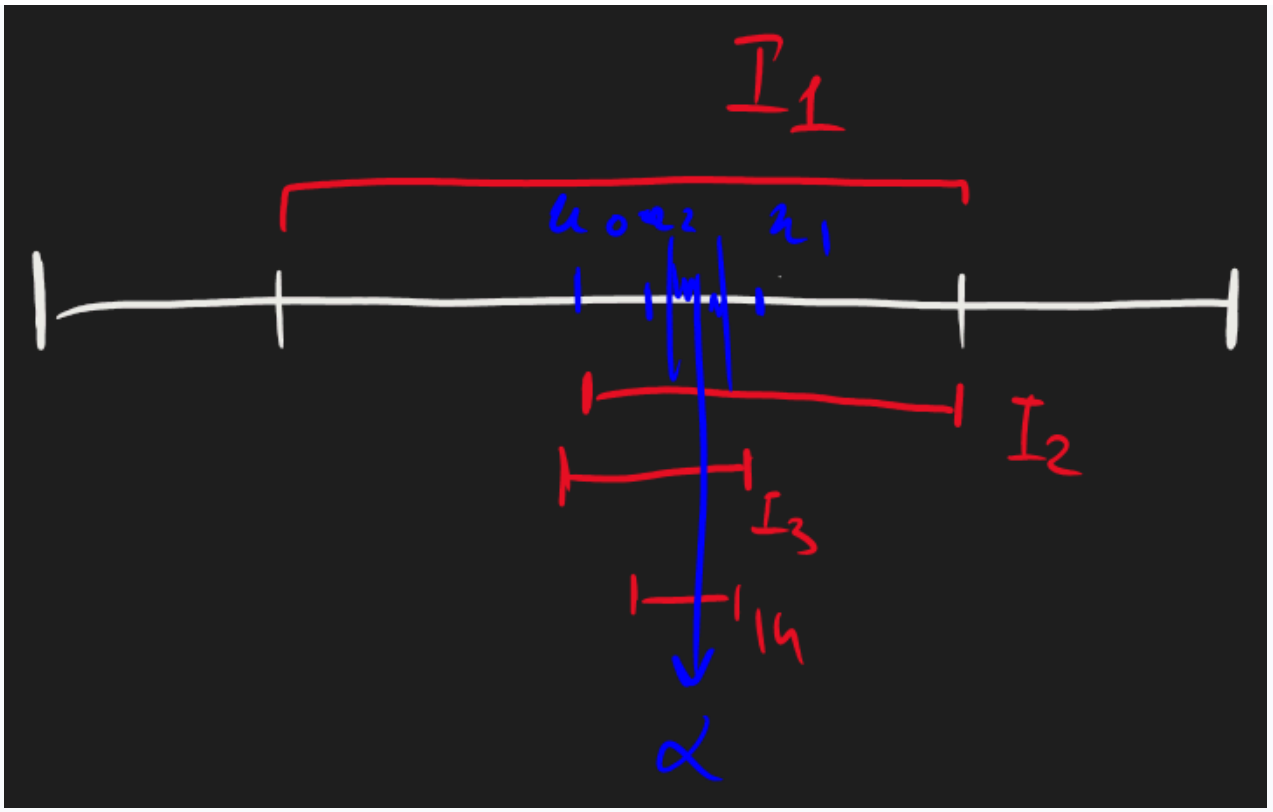
0. Voci correlate

- Calcolo dei Zeri delle Funzioni
- Algoritmi Numerici Iterativi

1. Idea del Metodo della Bisezione

OBBIETTIVO. Data una funzione $f \in \mathcal{C}^0([a,b])$ tale che $\exists! \alpha \in [a,b] : f(\alpha) = 0$, vogliamo trovare il valore α.

Un'idea *semplice* per trovare l'α è quello di *dividere* l'intervallo $I := [a,b]$ in due parti eque I_1, I_2. Siccome il valore α è unico, essa esiste solamente in una delle due parti. Pertanto, guardando i *segni* di f in $\partial I_1, \partial I_2$ (ovvero dove le frontiere sono valutate con segno discordi) per scegliere quale sotto-intervallo dividere nuovamente. Dopodiché, iteriamo il ragionamento finché per Cantor convergiamo a un valore ξ, che sarà proprio α.



X

2. Definizione del Metodo della Bisezione

Formalizziamo l'idea della *bisezione* in termini matematici precisi.

#Definizione

> Definizione (metodo della bisezione).
>
> Sia $f \in \mathcal{C}^0([a,b])$ tale che $\exists! \alpha \in [a,b] : f(\alpha) = 0$.
>
> Il *metodo della bisezione* è un *algoritmo numerico iterativo*, in cui definiamo la successione degli intervalli $(I_n)_{n \in \mathbb{N}}$ come
>
> $$I_n : \begin{cases} [a,b], n = 0 \\ \left[a_{n-1}, \dfrac{a_{n-1}+b_{n-1}}{2}\right], n > 0 \wedge \displaystyle\prod_{d \in \partial I_n} f(d) < 0 \\ \left[\dfrac{a_{n-1}+b_{n-1}}{2}, b_{n-1}\right], n > 0 \wedge \displaystyle\prod_{d \in \partial I_n} f(d) < 0 \end{cases}$$
>
> In particolare l'algoritmo della bisezione viene associata alla *successione dei punti medi*, posta come
>
> $$\bar{x}_k := \frac{a_k + b_k}{2}$$

Scriviamo lo pseudocodice per la bisezione.

```
LET [a,b], f

LET a(0) = a, b(0)=b
FOR  k=0, 1, ..., k
    LET x_k = (a(k)+b(k))/2
    IF f(x_k)=0: RETURN x_k
    ELSE:
        IF f(x_k)f(a(k)) < 0: b_k = x_k
        ELSE IF f(x_k)f(b(k)) < 0: a_k = x_k
RETURN x_k
```

X

3. Convergenza della Bisezione

Adesso dimostriamo che il metodo della bisezione effettivamente *funzioni*, studiando e dimostrando la sua convergenza alla soluzione α.

#Osservazione

> Osservazione (maggiorazione dell'errore).
>
> Notiamo che l'errore $|\varepsilon_k| := |\alpha - x_k|$ è al massimo l'ampiezza dell'intervallo k-esimo, ovvero ho la maggiorazione
>
> $$|\varepsilon_k| = |\alpha - x_k| \leq \frac{b_k - a_k}{2}$$
>
> Tuttavia non è vero che l'errore $|\varepsilon_k|$ diminuisca vero, infatti è vero solo che la sua maggiorazione decresce.
>
> Infatti è possibile che essa oscilli, presentando comunque un trend verso il basso.

#Teorema

> Teorema (convergenza della bisezione).
>
> Il *metodo della bisezione* converge globalmente.

#Dimostrazione

DIMOSTRAZIONE del Teorema 3

Voglio dimostrare il limite

$$\lim_k |\varepsilon_k| = 0$$

Usiamo la maggiorazione

$$0 \leq |\varepsilon_k| \leq \frac{b_k - a_k}{2}$$

Per induzione si dimostra che

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

Pertanto ho

$$0 \leq |\varepsilon_k| \leq \frac{b_0 - a_0}{2^{k+1}}$$

Ovviamente $\lim_k 0 = 0$ e $\lim_k (b_0 - a_0)2^{-(k+1)} = 0$, pertanto per il teorema dei due carabinieri ho il limite come si voleva all'inizio, concludendo. ■

Con quale ordine converge il metodo della bisezione? Purtroppo, non esiste il limite

$$\lim_k \frac{|\varepsilon_{k+1}|}{|\varepsilon_k|}$$

In quanto gli errori $|\varepsilon_k|$ possono oscillare. Tuttavia possiamo comunque *"togliere il limite"* e ottenere l'approssimazione

$$\frac{|\varepsilon_{k+1}|}{|\varepsilon_k|} \simeq \frac{1}{2}$$

Questo ha senso in quanto facciamo dimezzare il range, effettivamente

$$|\varepsilon_{k+1}| = \frac{1}{2}|\varepsilon_k|$$

#Osservazione

> Osservazione (ricavare il numero di iterazioni per ottenere una maggiorazione dell'errore).

Abbiamo

$$|\varepsilon_k| \leq \frac{b_0 - a_0}{2^{k+1}} \leq \texttt{tol} \sim 10^{-n}$$

Per ottenere esplicitamentei l valore k in funzione di n, effettuiamo dei calcoli

$$
\begin{aligned}
\frac{b_0 - a_0}{2^{k+1}} &\leq 10^{-n} \\
\log\left(\frac{b_0 - a_0}{2^{k+1}}\right) &\leq \log(10^{-n}) \\
\log(b_0 - a_0) - \log(2^{k+1}) &\leq -n\log(10) \\
(k+1)\log(2) &\geq \underbrace{n\log(10) + \log(b_0 - a_0)}_{\xi_n} \\
k &\geq \xi_n - 1 \\
k &\geq \lceil \xi_n \rceil - 1
\end{aligned}
$$

X

4. Test del Residuo Pesato

Q. Con un'implementazione *reale* di *Newton-Rhapson*, come determino il criterio d'arresto, dato un margine d'errore accettabile? Siccome non conosciamo il valore α, non possiamo conoscere l'errore k-esimo $|\varepsilon_k|$.
A. Una risposta parziale è fornita dal fatto che nello standard ANSI la precisione macchina è 2^{-53}, da cui per avere la *massima precisione* posso effettuare fino a 53 iterazioni. Tuttavia sono troppe, voglio trovare un modo più *"smart"* per trovare la condizione d'arresto.

#Definizione

Definizione (definizione di residuo).

Definiamo *residuo* di una funzione f come la quantità

$$|f(x_k)| := y_k$$

Possiamo dunque definire una tolleranza `tol` e dunque arrestare il programma quando k è tale che $y_k \leq$ `tol` (di solito `tol` $\approx 10^{-12} = \mathbf{u}$). Questo è noto come il *test di arresto del residuo*. Ma va bene? Adesso vediamo il problema.

#Osservazione

Osservazione (inaffidibilità del test del residuo).

Sia $f(x) = 10^{-50}x$ con $\alpha = 0$ una funzione *"piatta"*. Usando la *bisezione* col *test di residuo*, abbiamo che il programma termina per $\alpha \gg x_k$. Infatti per $x_k = 1$ ho $y_k = 10^{-50}$ per cui termina, ma ottengo un'errore $|\alpha - x_k| = 1 = 100\%$!

Similmente, definendo $f(x) = 10^{50}x$ una funzione *"ripida"* ho l'effetto opposto, ovvero il programma termina per $\alpha \simeq x_k$. Supponendo che $x_k = 10^{-50}$, ottengo $y_k = 1$, per cui continuo ancora a iterare fino quasi all'infinito.

Un modo per risolvere questo problema è di normalizzare il residuo con la sua derivata, ottenendo dunque

#Definizione

> **Definizione (residuo pesato).**
>
> Definiamo *residuo pesato* di una funzione f come la quantità
>
> $$\bar{y}_k := \frac{f(x_k)}{f'(x_k)}$$

#Proposizione

> **Proposizione (il residuo pesato è un'approssimante lineare dell'errore).**
>
> Si ha che il *residuo pesato* $\bar{y}_k$ è un'*approssimante lineare* dell'errore k-esimo ε_k, ovvero
>
> $$\bar{y}_k = \varepsilon_k + O(\varepsilon_k^2)$$

#Dimostrazione

DIMOSTRAZIONE della Proposizione 8

Osserviamo che

$$\varepsilon_k = \alpha - x_k \implies \alpha = x_k + \varepsilon_k$$

Da cui

$$f(\alpha) = 0 \iff f(x_k + \varepsilon_k) = 0$$

Per la formula di Taylor col resto di Peano centrato in x_k con ordine $n = 2$ otteniamo che

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

Dividendo per la derivata $f'(x_k)$ otteniamo

$$\varepsilon_k + O(\varepsilon_k^2) + \underbrace{\frac{f(x_k)}{f'(x_k)}}_{\bar{y}_k} = 0$$

Che è la tesi. ■

Osserviamo che affinché vale questo teorema, richiediamo che α è *localmente un zero semplice*. Ovvero esiste un intorno $U(\alpha)$ tale non si annulla la derivata.

Metodo di Newton-Rhapson

Metodo Di Newton-Raphson

X

Metodo di Newton-Rhapson. Cenni storici, idea dell'algoritmo, derivazione dello schema analitico. Teorema di convergenza locale di N-R, dimostrazione. Ordine di convergenza di N-R. Criterio d'arresto. Cenni al caso multidimensionale.

X

0. Voci correlate

- Calcolo dei Zeri delle Funzioni
- Algoritmi Numerici Iterativi
- Matrice Jacobiana di Funzioni in più Variabili

1. Cenni Storici

Isaac Newton e Joseph Raphson furono due matematici inglesi attivi intorno alla seconda metà del XVII secolo, precisamente attorno agli anni 1660-1690. In quel periodo, la comunità matematica era profondamente impegnata nella ricerca di metodi efficaci per risolvere equazioni, soprattutto quelle di grado superiore, che rappresentavano una sfida importante per studiosi e ricercatori.

Newton, nel corso dei suoi studi, elaborò un innovativo metodo, noto come "metodo delle tangenti", con il quale riusciva ad affrontare e risolvere equazioni di terzo grado sfruttando l'approssimazione delle soluzioni attraverso la geometria e le proprietà delle curve. Tuttavia, Newton non pubblicò mai esplicitamente i dettagli completi del suo procedimento, lasciando questo suo importante contributo in gran parte inedito.

Fu Joseph Raphson che, comprendendo il valore e l'importanza di questo metodo, decise di portarlo alla luce. Così, nel 1690, Raphson pubblicò definitivamente il metodo in forma analitica, utilizzando esplicitamente il concetto matematico di derivata, ponendo così le basi per quello che oggi è universalmente noto come il "metodo di Newton-Raphson". Questa tecnica rappresenta ancora oggi uno strumento fondamentale nell'ambito dell'analisi numerica, ampiamente utilizzato per approssimare le radici delle equazioni, e testimonia la rilevanza duratura dell'intuizione originale di Newton e dell'opera divulgativa e analitica di Raphson.

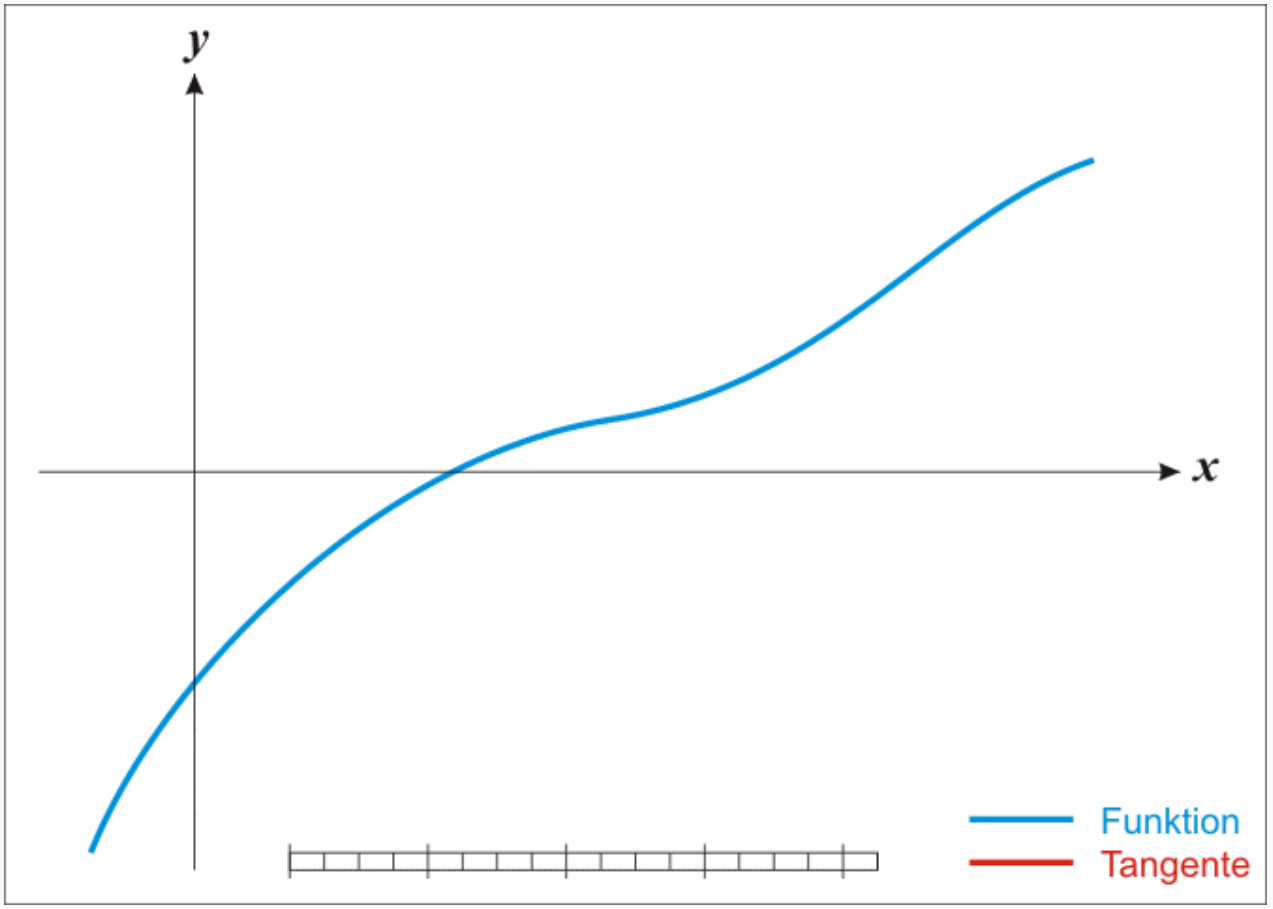
X

2. Intuizione del Metodo di Newton-Raphson

IDEA. Sia $f \in \mathcal{C}^1([a,b];\mathbb{R})$ con un zero in $[a,b]$ (detto α). Ad ogni punto x_k vogliamo approssimare la
funzione f col suo approssimante lineare $\bar{f}$, ovvero la sua *retta tangente* calcolata in quel punto.

Quindi, dato un punto iniziale $x_0 \in [a,b]$, calcolo l'approssimante lineare $\bar{f}(x_0)$ e trovo l'intersezione tra $\bar{f}(x_0)$ e la retta orizzontale. Poi si ripete

Geometricamente, faccio una *"discesa"* della funzione con la retta tangente.



Per ricavare analiticamente lo *schema iterativo* di N-R, abbiamo due modi:
Geometrico. Proseguiamo con l'intuizione geometrica di Newton, ponendod unque il sistema delle equazioni

$$\begin{cases} \bar{f}(x, x_k) = f(x_k) + f'(x_k)(x - x_k) \\ \bar{f}(x, x_k) = 0 \end{cases}$$

Da cui ottengo naturalmente

$$f(x_k) + f'(x_k)(x - x_k) = 0$$

L'idea è di isolare x e chiamarlo il nuovo valore x_{k+1}, ovvero

$$f(x_k) + f'(x_k)(x - x_k) = 0 \iff (x - x_k) = -\frac{f(x_k)}{f'(x_k)} \iff x = x_k - \frac{f(x_k)}{f'(x_k)}$$

Ottenendo in definitiva

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

Analitico. Possiamo definire S_k lo scarto k-esimo che è la misura della *"discesa"* della nostra successione $(x_k)_k$, ovvero

$$x_{k+1} = x_k + S_k$$

Come trovo S_k? Uso il fatto che sto effettuando una approssimazione lineare, quindi possiamo usare la formula di Taylor e dimenticare il termine approssimativo:

$$f(x_{k+1}) = f(x_k + S_k) = f(x_k) + f'(x_k)S_k + O(S_k^2) \simeq f(x_k) + f'(x_k)S_k \equiv 0$$

Lavorando sull'equazione $f(x_k) + f'(x_k)S_k \equiv 0$ isoliamo S_k ottenendo dunque

$$\boxed{S_k = -\frac{f(x_k)}{f'(x_k)}}$$

#Definizione

> **Definizione (metodo di Newton-Raphson).**
>
> Sia $f \in \mathcal{C}^1([a, b]; \mathbb{R})$ per cui $\exists! \alpha \in [a, b] : f(\alpha) = 0$. Allora, dato $x_0 \in [a, b]$, definiamo il *metodo di Newton-Raphson* come la successione
>
> $$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, k > 0$$
>
> Dove denotiamo il resto S_k come il termine
>
> $$S_{k+1} = -\frac{f(x_k)}{f'(x_k)}$$

3. Convergenza Locale di Newton-Raphson

Dimostriamo il punto cruciale di N-R, ovvero la sua *convergenza locale*.

#Teorema

> Teorema (convergenza locale di Newton-Raphson).
>
> Sia $f \in \mathcal{C}^2([a,b] : \mathbb{R})$ con $\alpha \in [a,b]$ tale che $f(\alpha) = 0$. Supponiamo che $\forall x \in [a,b], f'(x) \neq 0$, allora il metodo di *Newton-Raphson* è ben-definito, ossia esiste un intorno $U(\alpha)$ per cui $(x_n)_n \subset U(\alpha)$.
>
> Inoltre si ha che $\exists M > 0$ tale che
>
> $$\forall k, \frac{|\varepsilon_{k+1}|}{|\varepsilon_k|^2} \leq M$$
>
> Infine si ha il limite
>
> $$\lim_k \frac{|\varepsilon_{k+1}|}{|\varepsilon_k|^2} = \frac{1}{2}\frac{|f''(\alpha)|}{|f'(\alpha)|}$$

#Dimostrazione

DIMOSTRAZIONE del Teorema 2

Step 1. Dimostriamo innanzitutto la maggiorazione

$$\forall k, \frac{|\varepsilon_{k+1}|}{|\varepsilon_k|^2} < M$$

Ciò si fa trovando $M > 0$ valido per $\forall k$. Dalla definizione di scarto, ossia $\varepsilon_k := \alpha - x_k$, ho $\alpha = x_k + \varepsilon_k$. Allora per la la formula di Taylor col resto di Lagrange di $f(\alpha) = f(x_k + \varepsilon_k)$ centrata in x_k di ordine $n = 1$, ho che $\exists \xi_k \in \text{interval}(\alpha, x_k)$ tale che

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

Dividendo per $f'(x_k)$ ottengo lo scarto

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

Ricordandomi che in N-R ho la forma $S_k = x_{k+1} - x_k$, ottengo

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

Per relazionarmi con gli errori, aggiungo e sottraggo alpha nella RHS:

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

Dunque ho

$$-\varepsilon_{k+1} + \varepsilon_k = \varepsilon_k + \frac{1}{2}\frac{f''(\xi_k)}{f'(x_k)}\varepsilon_k^2$$

Cancellando ε_k e dividendo per ε_k^2 ottengo l'espressione

$$(*)\ \frac{|\varepsilon_{k+1}|}{|\varepsilon_k|^2} = \frac{1}{2}\frac{|f''(\xi_k)|}{|f'(x_k)|}$$

Notando che per ipotesi $f \in \mathcal{C}^2$ e dunque f', f'' sono limitate (per Weierstrass) e che $\xi_k, x_k \in [a,b]$ sicuramente, allora ho le maggiorazioni

$$f''(\xi_k) \leq \max_{x\in[a,b]} |f''(x)| := \mathtt{M}$$
$$f'(x_k) \geq \min_{x\in[a,b]} |f'(x)| := \mathtt{m}$$

Definendo $M := \mathtt{M}/\mathtt{m}$, ottengo

$$\frac{|\varepsilon_{k+1}|}{|\varepsilon_k|^2} \leq M$$

Step 2. Adesso dimostriamo che la successione $(x_n)_n$ converge ad α. Usiamo la disuguaglianza sopra, da cui

$$|\varepsilon_{k+1}| \leq M|\varepsilon_k|^2$$

Moltiplicando per M ottengo

$$M|\varepsilon_{k+1}| \leq M^2|\varepsilon_k|^2 = (M|\varepsilon_k|)^2$$

Allora queste disuguaglianze si ripetono fino a $k=0$, da cui

$$M|\varepsilon_{k+1}| \leq (M|\varepsilon_k|)^2 \leq (M|\varepsilon_{k-1}|)^{2^2} \leq \ldots \leq (M|\varepsilon_0|)^{2^k}$$

Per $M|\varepsilon_0| < 1$, ovvero per $x_0 \in (\alpha - 1/M, \alpha + 1/M)$, per due carabinieri ottengo il limite

$$0 \leq |\varepsilon_k| \leq \frac{1}{M}(M|\varepsilon_0|)^{2^k} \implies \lim_k |\varepsilon_k| = 0$$

Step 2.1. Notiamo che $(x_k)_k \subset B(\alpha, 1/M)$ è equivalente a dire che $(\varepsilon_k)_k \subset B(0, 1/M)$. Per $k=0$ è già vero per ipotesi del teorema, dimostriamo dunque per induzione per il caso $k > 0$. Come ipotesi induttiva ho

$$|\varepsilon_k| < \frac{1}{M}$$

Devo dimostrare che vale ugualmente la disuguaglianza $|\varepsilon_{k+1}| < \frac{1}{M}$. Mi ricordo che ho la disuguaglianza

$$|\varepsilon_{k+1}| \leq M|\varepsilon_k|^2$$

Da cui

$$|\varepsilon_{k+1}| \leq M|\varepsilon_k|^2 < M\frac{1}{M^2} = \frac{1}{M}$$

Dimostrando dunque il punto.

Step 3. Dimostriamo l'ultima parte, ovvero il limite

$$\lim_k \frac{|\varepsilon_{k+1}|}{|\varepsilon_k|^2} = \ ?$$

Usiamo la $(*)$:

$$(*)\ \frac{|\varepsilon_{k+1}|}{|\varepsilon_k|^2} = \frac{1}{2}\frac{|f''(\xi_k)|}{|f'(x_k)|}$$

Lo portiamo al limite

$$\lim_k \frac{|\varepsilon_{k+1}|}{|\varepsilon_k|^2} = \lim_k \frac{1}{2}\frac{|f''(\xi_k)|}{|f'(x_k)|}$$

Senza dimostrare le ipotesi necessarie, portiamo il limite dentro la frazione e le funzioni:

$$\lim_k \frac{1}{2}\frac{|f''(\xi_k)|}{|f'(x_k)|} = \frac{1}{2}\frac{|f''(\lim_k \xi_k)|}{|f'(\lim_k x_k)|}$$

Dimostriamo il limite nel numeratore:

$$\lim_k \xi_k = \alpha$$

Infatti sappiamo per Lagrange che $\xi_k \in \text{interval}(\alpha, x_k)$. Tuttavia, essendo che $x_k \to \alpha$, abbiamo che ξ_k va inevitabilmente a collassarsi su α.
Allora ho

$$\frac{1}{2}\frac{|f''(\lim_k \xi_k)|}{|f'(\lim_k x_k)|} = \frac{1}{2}\frac{|f''(\alpha)|}{|f'(\alpha)|}$$

che conclude la dimostrazione. ■

#Osservazione

> Osservazione (l'indebolimento delle condizioni).
>
> Possiamo indebolire la condizione per cui $f \in \mathcal{C}^2$, richiedendo invece che f' è *lipschitziana* (Definizione 2), ossia $\exists M > 0$ tale che
>
> $$\forall x, x' \in [a,b], \frac{|f'(x) - f'(x')|}{|x - x'|} \leq M$$

X

4. Ordine di Convergenza di Newton-Raphson

Adesso discutiamo, informalmente, l'ordine di convergenza di Newton-Raphson. Per il teorema di convergenza locale ho

$$\lim_k \frac{|\varepsilon_{k+1}|}{|\varepsilon_k|^2} = \frac{1}{2}\frac{|f''(\alpha)|}{|f'(\alpha)|}$$

Notiamo che in ogni caso $C = \frac{1}{2}\frac{|f''(\alpha)|}{|f'(\alpha)|} \geq 0$, per cui $p \geq 2$. Ciò vuol dire che Newton-Raphson è *al peggio quadratico*, che un risultato eccezionale. Tuttavia se $f''(\alpha) = 0$, ottengo $C = 0$ e ciò significa che $p > 2$, ossia N-R è sicuramente *superlineare*.

In particolare se ho α con molteplicità $r \geq 1$ (ovvero ho un zero multiplo e $f'(\alpha) = 0$), allora la sua costante asintotica è calcolata come

$$C = 1 - \frac{1}{r}$$

Noto che in tal caso per $r \to +\infty$ ho $C \to 1$, ossia l'algoritmo diventa man mano più lento.

Facciamo il seguente esercizio:

#Esercizio

> **Esercizio (calcolo della costante asintotica).**
>
> Sia $f \in C^3$ tale che $f(\alpha) = 0, f'(\alpha \neq 0), f''(\alpha) = 0$ ma $f'''(\alpha) \neq 0$. Calcolare la costante asintotica C per $p = 3$.

Per effettuare l'esercizio usiamo lo *sviluppo di Taylor col resto di lagrange* di $f(\alpha) = f(x_k + \varepsilon_k)$ centrato in x_k di ordine $n = 2$. Ovvero,

$$f(x_k + \varepsilon_k) = f(x_k) + f'(x_k)\varepsilon_k + \frac{1}{2}f''(x_k)\varepsilon_k^2 + \frac{1}{3!}f'''(\xi_k)\varepsilon_k^3 \equiv 0$$

Isoliamo $f(x_k)$ e dividiamo per $f'(x_k)$, ottenendo lo schema iterativo

$$-\frac{f(x_k)}{f'(x_k)} = x_{k+1} - x_k = \varepsilon_k + \frac{1}{2}\frac{f''(x_k)}{f'(x_k)}\varepsilon_k^2 + \frac{1}{6}\frac{f'''(\xi_k)}{f'(x_k)}\varepsilon_k^3$$

Riconoscendo che $x_{k+1} - x_k = x_{k+1} - \alpha - x_k + \alpha = \varepsilon_{k+1} + \varepsilon_k$, ho che

$$\varepsilon_{k+1} = \frac{1}{2}\frac{f''(x_k)}{f'(x_k)}\varepsilon_k^2 + \frac{1}{6}\frac{f'''(\xi_k)}{f'(x_k)}\varepsilon_k^3$$

Vogliamo *"incrementare"* il grado di ε_k^2 di uno, così possiamo dividere per ε_k^3. Per farlo effettuiamo un ulteriore sviluppo di Taylor per $f''(x_k) = f''(x_k + \varepsilon_k - \varepsilon_k) = f''(\alpha - \varepsilon_k)$ di ordine $n = 1$ e centrato in $x_0 = \alpha$. In effetti avrei

$$f''(x_k) = \underbrace{-f''(\alpha)}_{0} - f'''(\eta_k)(\varepsilon_k) = -f'''(\eta_k)(\varepsilon_k)$$

Sostituendo ottengo

$$\varepsilon_{k+1} = -\frac{1}{2}\frac{f'''(\eta_k)}{f'(x_k)}\varepsilon_k^3 + \frac{1}{6}\frac{f'''(\xi_k)}{f'(x_k)}\varepsilon_k^3$$

Dividendo per ε_k^3 ottengo

$$\frac{\varepsilon_{k+1}}{\varepsilon_k^3} = \frac{1}{6}\frac{f'''(\xi_k)}{f'(x_k)} - \frac{1}{2}\frac{f'''(\eta_k)}{f'(x_k)}$$

Dato che $\xi_k, \eta_k \to \alpha$ e $x_k \to \alpha$, otteniamo il risultato finale

$$\boxed{\frac{|\varepsilon_{k+1}|}{|\varepsilon_k^3|} = \frac{1}{3}\frac{|f'''(\alpha)|}{|f'(\alpha)|}}$$

Concludendo. ■

Generalizziamo il risultato con la seguente proposizione:

#Proposizione

> Proposizione (calcolo della coefficiente asintotica).
>
> Fissiamo $k > 0$. $f \in \mathcal{C}^{\geq k}$ tale che $\exists \alpha \in [a,b] : f(\alpha) = 0$ e $f'(\alpha) \neq 0$, $f''(\alpha) = \ldots = f^{(k-1)}(\alpha) = 0$ e $f^{(k)}(\alpha) \neq 0$. Allora $p = k$ e
>
> $$C = \lim_n \frac{|\varepsilon_{n+1}|}{|\varepsilon_n^p|} = \frac{p!}{(p-1)!}\frac{f^{(p)}(\alpha)}{f'(\alpha)}$$

Notiamo che la proposizione di cui sopra vale se $f'(\alpha) \neq 0$, ovvero la funzione presenta *radici semplici*. Nel caso in cui la radice fosse *multipla* con *molteplicità* r, allora *Newton-Raphson* potrebbe convergere lo stesso, con tuttavia una velocità *lineare*. Per *"recuperare"* la velocità quadratica, posso *"conoscere"* r ed effettuare la seguente modifica allo schema iterativo:

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

La dimostrazione è posticipata. ■

X

---

4. Criterio d'Arresto

Ricordiamo che $f' \subset \mathbb{R} \setminus \{0\}$., per cui ha senso definire lo scarto

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

Notiamo che questo non è altro che il *residuo pesato* $\bar{y}_k$, ho dunque che

$$S_{k+1} = x_{k+1} - x_k \simeq \varepsilon_k$$

Pertanto come test d'arresto si può imporre

$$S_k < \texttt{tol}$$

X

5. Cenni al Caso Multidimensionale

Supponiamo di avere il campo vettoriale $F : \mathbb{R}^{\texttt{d}} \longrightarrow \mathbb{R}^{\texttt{d}}$. La ricerca dei zeri corrisponde a risolvere il seguente sistema di equazioni:

$$F(\underline{x}) = \underline{0} \iff \begin{cases} f_1(\underline{x}) = 0 \\ \vdots \\ f_{\texttt{d}}(\underline{x}) = 0 \end{cases}$$

Non posso usare il *metodo della bisezione* in quanto la scelta del $\texttt{d}$-rettangolo diventa troppo *"sparsa"*. Tuttavia rimance sempre possibile usare il metodo di Newton-Raphson, solo dovendo apportare un paio di modifiche.
Ricordiamo che nel caso scalare ho

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

Posso estendere sul caso $n \in \mathbb{N}$ *"facendo finta"* di trasformare $f'(x_k)$ nella *matrice Jacobiana* $J_F(\underline{x_k})$ (Definizione 2). Siccome è una matrice quadrata $\mathbb{R}^{n \times n}$, possiamo definire l'inversa $J_F^{-1}(\underline{x_k})$. Dunque lo scarto nel caso n-dimensionale diventa

$$S_{k+1} = -(J_F^{-1}(\underline{x_k})) \cdot F(\underline{x_k})$$

Tuttavia, in termini pratici *non calcoleremo* mai esplicitamente l'inversa $J_F^{-1}(\underline{x_k})$, dato che presenta dei problemi computazionali. In alternativa si risolve il sistema lineare

$$J_F(\underline{x_k})S_{k+1} = -F(\underline{x_k})$$

Varianti del Metodo di Newton-Rhapson

Varianti di Newton-Raphson

X

Varianti del metodo di Newton-Raphson. Metodo della tangente fissa (vantaggi), metodo delle secanti (o della secante variabile o quasi-Newton).

X

0. Voci correlate

- Calcolo dei Zeri delle Funzioni
- Metodo Di Newton-Raphson

1. Metodo della Tangente Fissa

Notiamo che col metodo di Newton-Raphson, calcoliamo la successione degli scarti

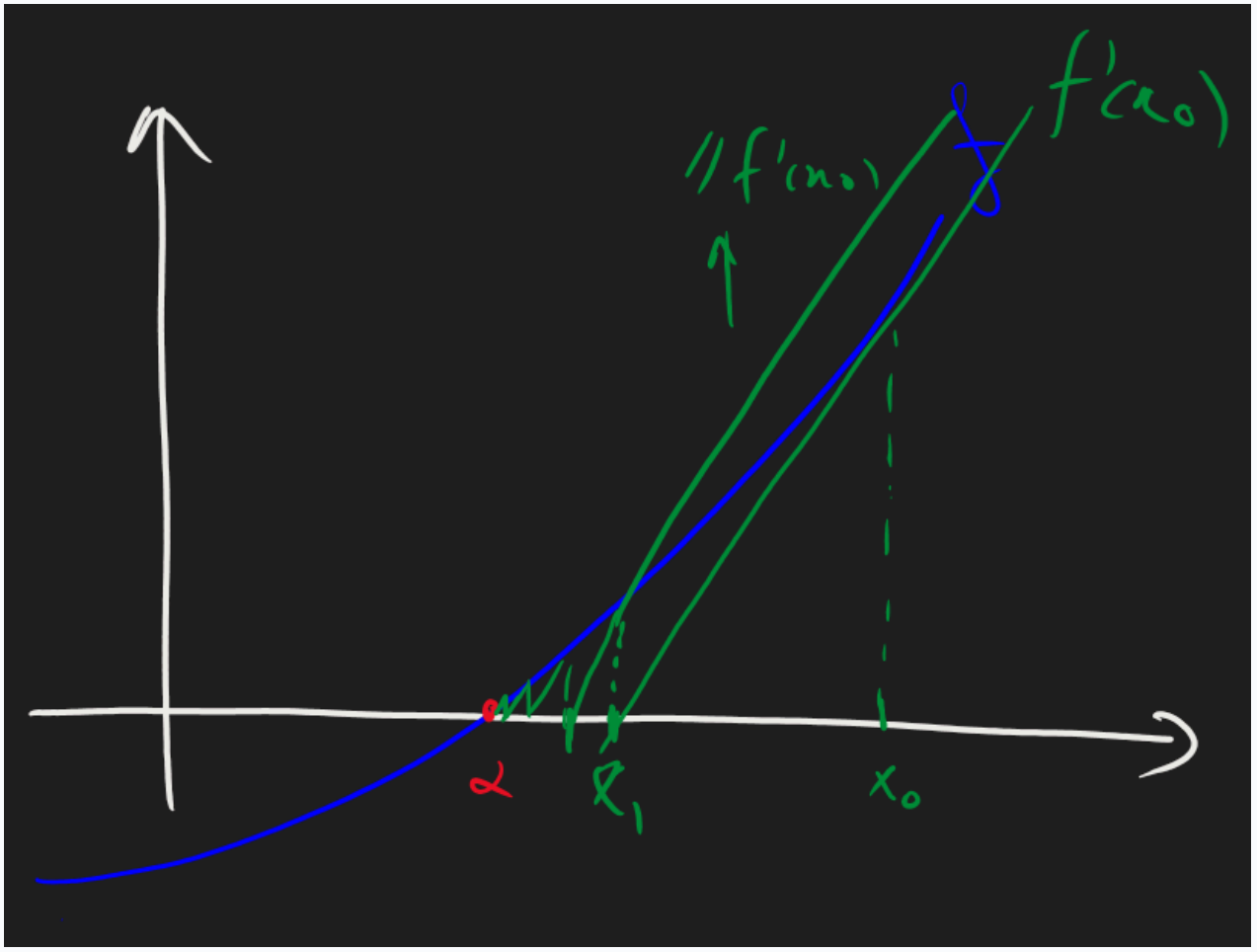
$$S_{k+1} = \frac{f(x_k)}{f'(x_k)}$$

In certi casi, specie quando si ha il caso `d`-dimensionale, il calcolo di questa successione diventa *computazionalmente costosa*. Un modo per alleggerire il costo computazionale è quello di calcolare *solamente* lo scarto S_1 e di mantenerla *"fissa"* per tutte le iterazioni.

Ossia, scelto un punto iniziale x_0, ho lo schema iterativo

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

Geometricamente vado ad effettuare la discesa con le *tangenti parallele alla tangente iniziale*.



Tuttavia il *minore costo computazionale* comporta un *trade-off*, ossia non ho più la *velocità quadratica* o *superlineare*; in effetti ho la velocità *lineare*, col limite

$$\lim_k \frac{|\varepsilon_{k+1}|}{|\varepsilon_k|} = 1 - \frac{f'(\alpha)}{f'(x_0)}$$

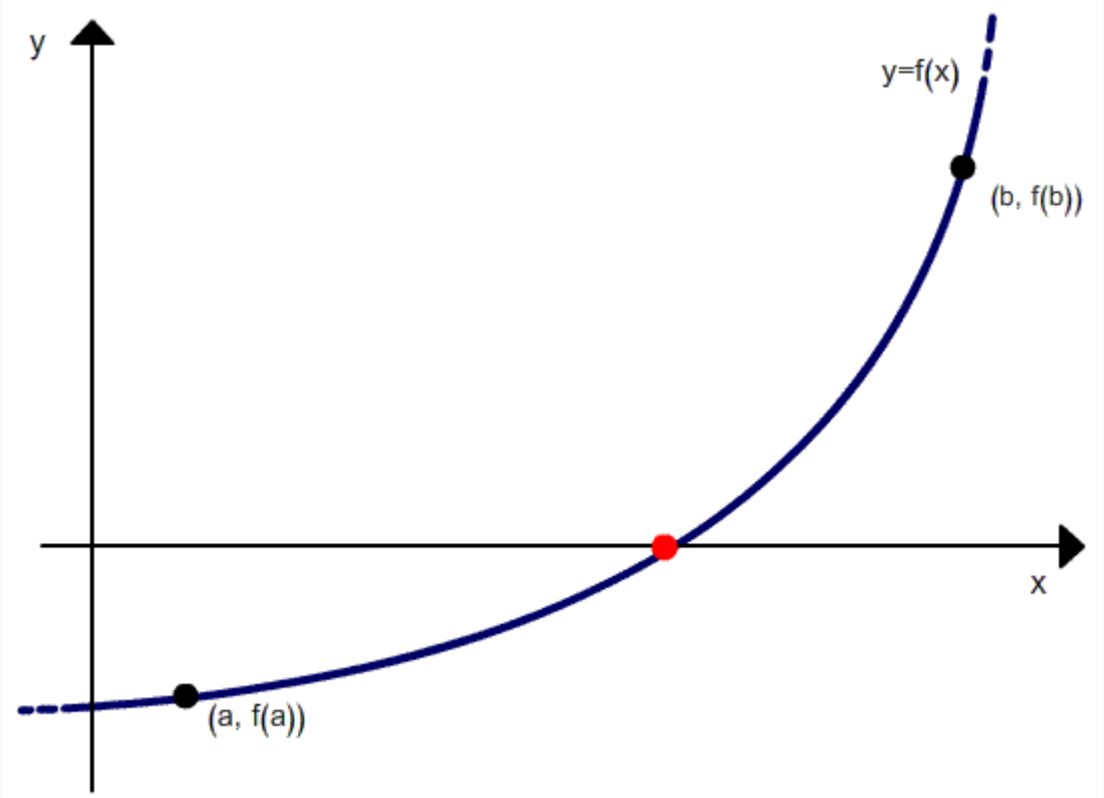
La dimostrazione di questo risultato è posticipata (vedremo col metodo del *punto fisso*). Tuttavia, osserviamo che abbiamo un altro trade off, per cui α dev'essere un *zero semplice*: infatti se $f'(\alpha) = 0$, allora si avrebbe che la costante asintotica è $C = 1$, per cui non converge il metodo. ■

X

2. Metodo delle Secanti

Idea. Uso le *secanti*, ovvero la retta che congiunge dei punti. Siano dunque $x_0 \neq x_1 \in I$, traccio la secante tra i punti iniziali x_0, x_1 e prendo la sua intersezione con l'asse x, che diventa il nuovo punto x_2. Dopodiché itero di nuovo con x_1, x_2.

[



La formulazione matematica è *quasi identica* a quella di Newton, solo che sostituiamo la *derivata* col *rapporto incrementale*.

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

Dove il rapporto incrementale $R^f_{\bar{x}}(x)$ è definita come (Definizione 1)

$$R^f_{\bar{x}}(x) = \frac{f(x) - f(\bar{x})}{x - \bar{x}}$$

In un certo senso, stiamo approssimando la derivata col rapporto incrementale. Infatti il metodo delle secante è *più lenta* di *Newton-Rhapson*, rimane tuttavia *superlineare*. Infatti, se $p > 1, f \in \mathcal{C}^2, f''(\alpha) \neq 0$ $(r > 2)$, allora vale che $p = \varphi^+$ dove φ^+ è la *sezione aurea positiva*, definita come

$$\varphi^+ := \frac{1+\sqrt{5}}{2} \approx 1.618$$

In particolare la costante asintotica vale

$$C = \frac{1}{2}\left|\frac{f''(\alpha)}{f'(\alpha)}\right|^{p-1}$$

Notiamo che se $f''(\alpha) = 0 \iff r = 2$, allora $p > \varphi^+$.

Questo metodo viene, in certi casi, riferito come un metodo *"quasi-Newton"*.

Metodo del Punto Fisso

Metodo del Punto Fisso

X

Metodo del punto fisso (o dell'iterazione funzionale). Definizione di punto fisso di una funzione, esempi. Schema del metodo del punto fisso, esempi di casi convergenti e divergenti.

X

---

0. Voci correlate

- Calcolo dei Zeri delle Funzioni
- Teorema di Lagrange

1. Fondamenti sui Punti Fissi

#Definizione

> Definizione (punto fisso di una funzione).
>
> Sia $g : [a,b] \longrightarrow \mathbb{R}$. $\alpha \in [a,b]$ si dice *punto fisso* di f sse vale che
>
> $$g(\alpha) = \alpha$$

Così, ci spostiamo al nostro problema del calcolo dei *zeri* in un problema del *calcolo dei punti fissi*. In particolare, data $f : [a,b] \longrightarrow \mathbb{R}$ per cui vogliamo trovare $f(\alpha) = 0$, vogliamo definire $g : [a,b] \longrightarrow \mathbb{R}$ per cui $g(\alpha) = \alpha$. Ossia, i valori di α coincidono.

Ci sono modi illimitati per effettuare la trasformazione $f \mapsto g$. Un metodo comune è quello di isolare un termine:

$$f(x) = e^{-x} - \cos x + 3x^2 \iff x = -\ln(\cos x - 3x^2) =: g(x)$$

Un altro modo (*non sempre convergente!*) per farlo è quello di sommare x da ambo i lati, ossia data una f qualunque questa diventa

$$f(x) = 0 \iff \underbrace{f(x) + x}_{:= g(x)} = x$$

X

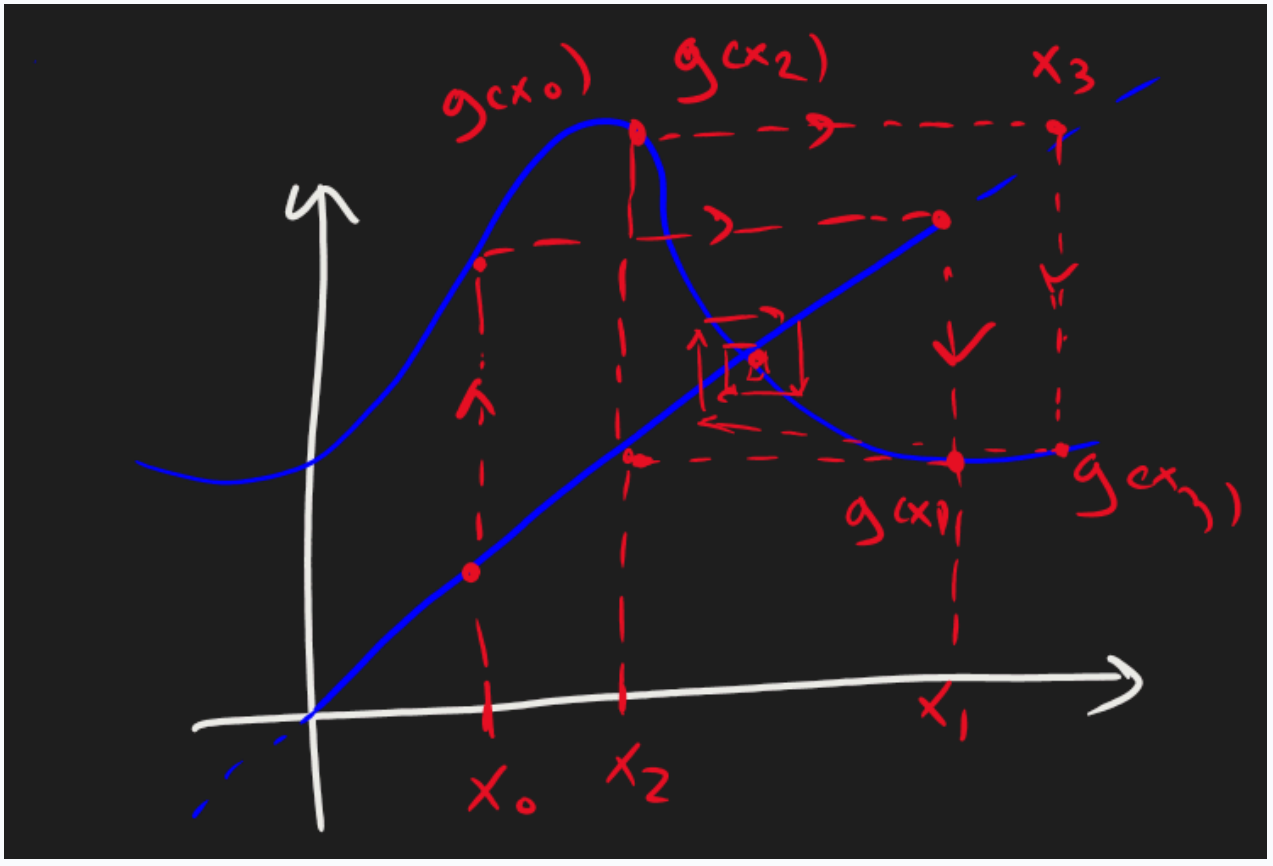
2. Schema del Punto Fisso

Definiamo lo schema del punto fisso come la seguente.

#Definizione

> **Definizione (metodo del punto fisso).**
>
> Si definisce il *metodo del punto fisso*, data una f per cui si cerca lo zero e g la sua forma in *"punto fisso"*, come la successione
>
> $$\forall k \geq 0, x_{k+1} = g(x_k)$$
>
> Notiamo che ovviamente x_0 è scelto manualmente.

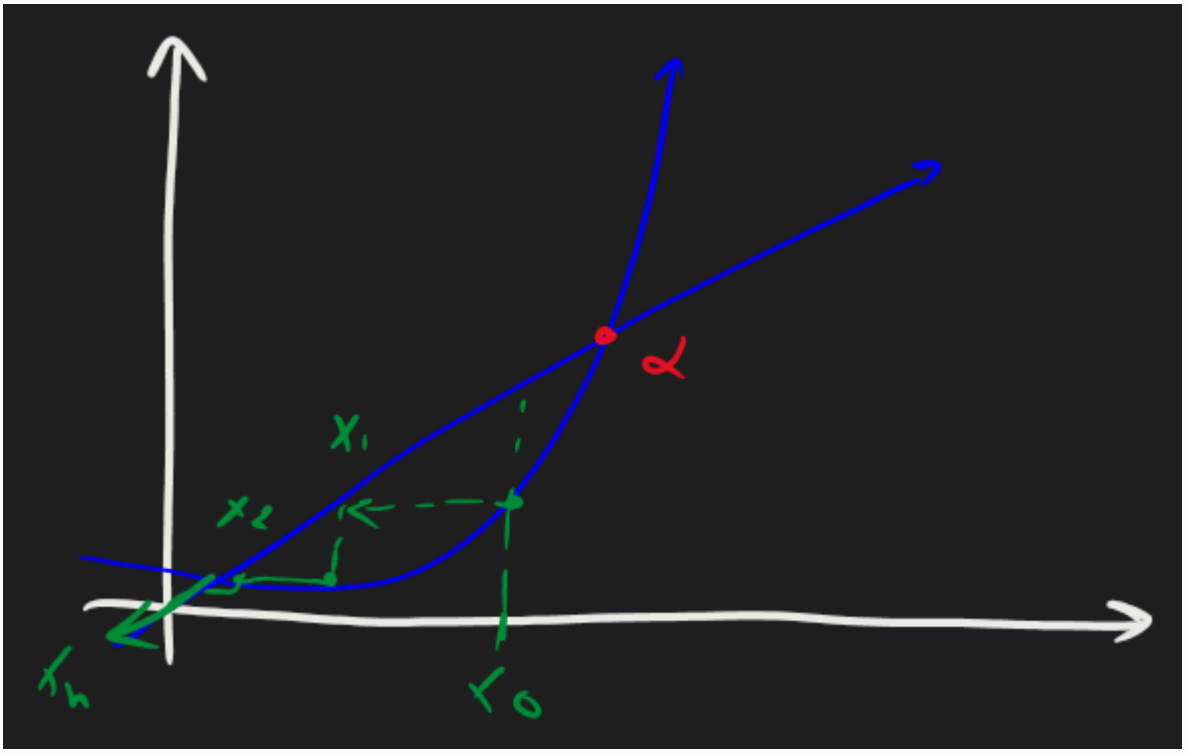
Geometricamente sto effettuando l'intersezione tra la *bisettrice* e la funzione g. e Per trovare il punto α effettuo una *"discesa"* sulla retta, trovando il valore $g(x_0)$ e definendo x_1 come $g(x_0)$, ossia la proiezione di $g(x_0)$ sulla bisettrice. In altre parole, mi sposto tra la funzione g e la bisettrice x.



X

3. Convergenza del Metodo del Punto Fisso

Osserviamo che la *convergenza* del metodo del punto fisso non è garantito. Infatti per funzioni che sono *"ripide"* attorno α, abbiamo la divergenza. Geometricamente, la ripidità va a causare l'allontamento della successione $(x_k)_k$.



Per un esempio concreto di convergenza invece (che al contrario, è *"piatta"* attorno ad α), prendiamo

$$f(x) = x - \cos x$$

Ossia per trovare il suo zero calcoliamo

$$x_{k+1} = \cos(x_k)$$

La successione certamente converge, infatti la successione è limitata ed eventualmente monotona.

Invece per un controesempio convergente, definiamo

$$f(x) = e^{-2x}(x-1), \alpha = 1$$

E definendo dunque

$$g(x) := e^{-2x}(x-1) + x$$

Abbiamo che, anche partendo *"abbastanza vicini"* ad $\alpha = 1$ (come $x_0 = 0.99$), lo schema iterativo diverge. Infatti per

$$x_{k+1} = e^{-2x_k}(x_k + 1) + x_k$$

Ottengo che $(x_k)_k$ non è limitata ed è crescente.

Enunciamo e dimostriamo *due teoremi* relativi alla convergenza del metodo del punto fisso.

3.1. Convergenza Globale

#Teorema

> Teorema (convergenza globale del punto fisso).
>
> Sia $g \in \mathcal{C}^1(I := [a,b]; \mathbb{R})$ tale che $g(I) \subseteq I$ (diremo che g è una *contrazione*). Allora *esiste* un punto fisso di g in I, ovvero $\exists \alpha \in I, g(\alpha) = \alpha$.
>
> Se inoltre vale che $\exists m \in (0,1) : \forall x \in I, |g'(x)| \leq m$, allora α è *unica* e il metodo converge, ovvero $\lim_n |\varepsilon_n| = 0$ indipendentemente da x_0 scelto.

#Dimostrazione

DIMOSTRAZIONE del Teorema 3

Step 1. $g(I) \subseteq I \implies \exists \alpha$:
Se vale la contrazione, allora graficamente vale che la *funzione* g è *"contenuta"* nel box $I \times I$. Matematicamente, ciò vuol dire che $g(a) \geq a$ e $g(b) \leq b$.

Definiamo la funzione ausiliaria $\varphi(x) = g(x) - x$: notiamo che φ ha una radice sse g ha un punto fisso. Pertanto dimostriamo che φ ha una radice.
Valutiamo φ negli estremi a, b:

$$\varphi(a) = g(a) - a \overset{g(a) \geq a}{\implies} \varphi(a) \geq 0$$

Analogamente $\varphi(b) \leq 0$. Allora per il *teorema degli zeri* abbiamo l'esistenza della radice, da cui l'esistenza del punto fisso.
Step 2. $|g'(x)| \leq m < 1 \implies \alpha!$:
D'ora in poi, supporremmo che

$$\exists m < 1 : \forall x \in I, |g'(x)| \leq m$$

Supponiamo per assurdo che $\exists \alpha, \alpha' \in [a,b] : g(\alpha) = \alpha, g(\alpha') = \alpha'$. Calcoliamo la differenza tra α, α':

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

Possiamo usare il *teorema di Lagrange* da cui $\exists \xi \in \texttt{Interval}(\alpha, \alpha') \subset [a,b]$ tale che

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

Per ipotesi $\forall x \in [a,b] |g'(x)| \leq m < 1$, da cui si ricava la maggiorazione stretta:

$$|g'(\xi)||\alpha - \alpha'| < |\alpha - \alpha'|$$

Che è assurdo, siccome $|g'(\xi)| \neq 1$ e dunque si avrebbe una forma del tipo $0.5x < x$.

Step 3. Convergenza globale

Dimostriamo infine il limite $|\varepsilon_k| \longrightarrow 0$. Ricordiamo che $\varepsilon_k := \alpha - x_k$, da cui

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

Usiamo nuovamente *Lagrange*, da cui $\exists \xi_k \in \texttt{Interval}(\alpha, x_k) \subset [a,b]$ tale che

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

Ovvero

$$|g'(\xi_k)||\varepsilon_{k-1}| = |\varepsilon_k|$$

Per ipotesi ho $\forall n |g(\xi_n)| \leq m < 1$, da cui

$$|\varepsilon_k| = |g(\xi_k)||\varepsilon_{k-1}| \leq m|\varepsilon_{k-1}|$$

Da cui ottengo la relazione ricorsiva

$$|\varepsilon_k| \leq m|\varepsilon_{k-1}| \leq m^2|\varepsilon_{k-2}| \leq \ldots \leq m^k|\varepsilon_0|$$

Quindi

$$0 \leq |\varepsilon_k| \leq m^k|\varepsilon_0|$$

Per due carabinieri ottengo la tesi. ■

3.2. Convergenza Locale

#Teorema

> **Teorema (convergenza locale del punto fisso).**
>
> Sia $g \in \mathcal{C}^1(I := [a,b])$ e sia $\alpha \in \mathbb{R}$ un punto fisso di g in $[a,b]$.
>
> Se $|g'(\alpha)| < 1$ allora esiste un intorno $U(\alpha)$ tale che $\forall x_0 \in U(\alpha), (x_n)_n \longrightarrow \alpha$. In particolare, $\exists \delta > 0$ per cui esista $U(\alpha) = B_\delta(\alpha) = (\alpha - \delta, \alpha + \delta)$.
>
> Inoltre si ha il seguente limite:
>
> $$\lim_k \frac{\varepsilon_{k+1}}{\varepsilon_k} = g'(\alpha)$$

Notiamo che questo corrisponde all'idea intuitiva di avere una certa *"piattezza"* vicino ad α.

#Dimostrazione

DIMOSTRAZIONE del Teorema 4

Ricordiamo che $g \in \mathcal{C}^1 \implies g' \in \mathcal{C}^0$, ossia per ipotesi vale che

$$\forall \varepsilon > 0, \exists \delta > 0 : \forall x \in I,$$
$$\underbrace{|x-\alpha|}_{x \in B_\delta(\alpha)} < \delta \implies |g'(x)| < \min\{1, \varepsilon\} \leq 1$$

Adesso, per dimostrare la *convergenza locale* usiamo la *convergenza globale* su $I_\delta = B_\delta(\alpha)$. Basta dimostrare la contrazione di g in I_δ. Ovvero,

$$(x \in I_\delta \implies g(x) \in I_\delta) \iff (|x-\alpha| < \delta \implies |g(x) - g(\alpha)| < \delta)$$

Per il teorema di Lagrange ho $\exists \xi \in \texttt{Interval}(x, \alpha) \subset I_\alpha$ per cui

$$|g(x) - g(\alpha)| = |g'(\xi)| \underbrace{|x-\alpha|}_{< \delta}$$

Tuttavia $|g'(\xi)| < 1$ per ipotesi (in quanto ξ sta in I_α), da cui $|g(x) - g(\alpha)| < \delta$ e dunque si ha la convergenza.
Dimostriamo il risultato finale, ovvero il limite degli scarti. Si tratta di usare Lagrange come prima, per cui

$$\varepsilon_{k+1} = g'(\xi_k)\varepsilon_k$$

da cui

$$\frac{\varepsilon_{k+1}}{\varepsilon_k} = g'(\xi_k)$$

Portanto al limite ho $\lim_k g'(\xi_k) \longrightarrow g'(\alpha)$ (continuità), da cui la tesi. ■

X

---

4. Ordine di Convergenza

Dimostriamo il seguente risultato relativo all'ordine di convergenza:

#Lemma

> Lemma (ordine di convergenza del punto fisso).
>
> Il metodo del punto fisso $(x_k)_k$ con g, nel caso convergente, ha:
>
> - Se $g'(\alpha) \neq 0$, allora $p = 1$ e $C < 1$
> - Se $g'(\alpha) = 0$ allora $p > 1$ (≥ 2) e in particolare
>
> $$p = \min\{r : f^{(r)}(\alpha) \neq 0\} \implies C = \frac{g^{(p)}(\alpha)}{p!}$$

#Dimostrazione

DIMOSTRAZIONE del Lemma 5

Dimostriamo il secondo punto per $p = 2$. Considero $-\varepsilon_{k+1}$ che è calcolato come

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

Uso Taylor su $g(x_k)$ centrato in $x_0 = \alpha$ all'ordine $n = 2$. Dunque $x - x_0 = -\varepsilon_k$ e ho

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

I $g(\alpha)$ si cancellano dato che $g'(\alpha) = 0$, da cui mi rimane solo

$$-\varepsilon_{k+1} = \frac{1}{2}g''(\xi_k)\varepsilon_k^2$$

Dividendo per ε_k^2 ottengo

$$\frac{\varepsilon_{k+1}}{\varepsilon_k^2} = -\frac{1}{2}g''(\xi_k) \overset{n\to+\infty}{\longrightarrow} -\frac{1}{2}g''(\alpha)$$

Portando al valore assoluto ho la tesi. ■

X

---

5. Scarto del Punto Fisso

Q. Come possiamo stimare l'errore al passo k-esimo, ε_k? Una buona *"approssimazione"* potrebbe essere, come negli altri metodi, lo scarto $s_k := x_k - x_{x-1}$.

Calcoliamo $\varepsilon_k = \alpha - x_k$:

$$\begin{aligned}\alpha - x_k &= g(\alpha) - g(x_{k-1}) \\ &= g'(\xi_k)(\alpha - x_{k-1}) \\ &= g'(\xi_k)(\alpha + x_k - x_{k-1} - x_k) \\ &= g'(\xi_k)(\varepsilon_k + s_k)\end{aligned}$$

Da cui

$$\varepsilon_k = (\varepsilon_k + s_k)g'(\xi_k) \implies \varepsilon_k = \frac{g'(\xi)}{1 - g'(\xi)}s_k \simeq \frac{|g'(\alpha)|}{1 - g'(\alpha)}s_k$$

Quindi per $g'(\alpha)$ *"lontana"* da 1 ho delle stime in ordine di grandezza per lo più pari (altrimenti ε_k viene sovrastimato da s_k). ■

Relazione scarto-errore per Metodi Superlineari

Scarto di Metodi Superlineari

X

Relazione tra scarto ed errore per metodi superlineari.

X

---

0. Voci correlate

- Algoritmi Numerici Iterativi

1. Scarto di Metodi Superlineari

Sia $(x_k)_k \longrightarrow \alpha$ un metodo *superlineare*, ovvero tale che

$$\frac{|\varepsilon_{k+1}|}{|\varepsilon_k|^{p>1}} = C > 0$$

Come possiamo *"approssimare"* l'errore al passo k-esimo, *conoscendo* solo lo scarto $S_{k+1} = x_{k+1} - x_k$ (di solito è così)? Per metodi più noti come *Newton-Raphson* usiamo semplicemente lo scarto stesso, in quanto possiamo *dimostrare* che sono degli approssimanti lineari di ε_k.

Adesso vediamo il *caso generale*, ossia *metodi supelineari*. Notiamo innanzitutto che la superlinearità comporta

$$|\varepsilon_{k+1}| = C|\varepsilon_k|^{p>1}$$

Quindi l'errore al passo $k+1$-esimo *decrementa* con un'ordine di grandezza più grande al passo k-esimo.

Calcoliamo dunque S_{k+1}:

$$S_{k+1} = x_{k+1} - x_k + \alpha - \alpha = \varepsilon_k - \varepsilon_{k+1}$$

Per l'osservazione fatta prima, abbiamo che $\varepsilon_k \ggg \varepsilon_{k+1}$, per cui possiamo *"far trascurare"* ε_{k+1} dunque ho

$$\boxed{S_{k+1} \approx \varepsilon_k}$$

Rivisitazione del Metodo di Newton-Raphson

Rivisitazione del Metodo di Newton-Raphson

X

Rivisitazione del Metodo di Newton-Raphson alla luce del metodo del punto fisso. Ordine di convergenza, costante asintotica

---

X

---

0. Voci correlate

- Metodo Di Newton-Raphson
- Varianti di Newton-Raphson
- Metodo del Punto Fisso

1. Rivisitazione di Newton-Raphson

#Osservazione

Osservazione. (*N-R è un metodo di punto fisso*)
Notiamo che lo schema iterativo di Newton-Raphson è

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

Definendo g come

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

Dunque $x_{k+1} = g(x_k)$ non è altro che *un metodo del punto fisso*. Questo discorso vale analogamente per il metodo della *tangente fissa*, ma non per la *secante variabile*.

X

---

2. Ordine di Convergenza di NR

#Osservazione

Osservazione. (*Ri-dimostrare l'ordine di convergenza e costante asintotica*)
Sapendo che NR non è altro che un *metodo del punto fisso*, sappiamo che la costante asintotica C dipende dalla derivata $|g'(\alpha)|$. Facciamo due conti:

$$\begin{aligned} g'(\alpha) &= 1 - \partial_\alpha \left(\frac{f(\alpha)}{f'(\alpha)} \right) \\ &= 1 - \frac{(f'(\alpha))^2 - f'(\alpha)f''(\alpha)}{(f'(\alpha))^2} \\ &= 1 - 1 + \frac{f(\alpha)f''(\alpha)}{(f'(\alpha))^2} = \frac{f(\alpha)f''(\alpha)}{(f'(\alpha))^2} \end{aligned}$$

Su questa facciamo un paio di osservazioni:

- Siccome $f(\alpha) = 0$, allora sicuramente la costante valutata in $p = 1$ è nulla e dunque $p > 1$. Per calcolare la sua vera costante asintotica, bisogna fare ulteriori calcoli su g (calcolando in particolare g'')
- Richiediamo che α sia una radice semplice, infatti altrimenti avrei un caso determinato del tipo $[0/0]$. Comunque si può applicare de l'Hopital ripetutamente:

  $$\frac{f(\alpha)f''(\alpha)}{(f'(\alpha))^2} \Longleftarrow \frac{f'(\alpha)f''(\alpha) + f(\alpha)f'''(\alpha)}{2f''(\alpha)f'(\alpha)} \Longleftarrow \frac{1}{2}\frac{(f''(\alpha))^2 + f'(\alpha)f'''(\alpha) + f'(\alpha)f'''(\alpha)}{f'''(\alpha)f'(\alpha) + f''(\alpha)f''}$$

  Se $r = 2$, allora $f''(\alpha) \neq 0$ da cui

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

  Pertanto il metodo converge linearmente.

#Osservazione

Osservazione. (*Stessa cosa ma tangente fissa*)
Sia

$$x_{k+1} := g_{\text{TF}}(x_k) = x_k - \frac{f(x_k)}{f'(x_0)}$$

Allora

$$g'_{\text{TF}}(x) = 1 - \frac{f'(x)}{f'(x_0)}$$

Pertanto per $f'(\alpha) \neq 0$ ho $g'_{\text{TF}}(\alpha) \neq 0$ per cui $C \neq 1$ e quindi converge linearmente. ■

X

"Approsimazione delle Funzioni e dei Dati"

X

Introduzione alla Teoria dell'Approssimazione

Introduzione alla Teoria dell'Approssimazione

X

Motivazioni per la teoria dell'approssimazione. Tecniche principali della teoria dell'approssimazione: interpolazione e approssimazione ai minimi quadrati (OLS). Spazi funzionali più comuni per l'approssimazione delle funzioni.

X

---

0. Voci correlate

- Analisi e Sintesi di Fourier

1. Teoria dell'Approssimazione delle Funzioni

PROBLEMA. Sia $f \in \mathcal{F}$ una funzione (di solito reale; $f : \mathbb{E} \subseteq R \longrightarrow \mathbb{R}$). Come possiamo trovare una funzione $\tilde{f}$ tale che $\tilde{f} \approx f$, anche se non conosciamo la forma analitica di f?

MOTIVAZIONI PER LA TEORIA DELL'APPROSSIMAZIONE

Ho *due motivazioni principali* per la formulazione di questo problema:

- Non conosco la forma analitica f, bensì solamente alcuni suoi *dati numerici* in forma $(x_n, f(x_n))_n$, dove $(x_n)_n \subset E$.

- Conosco f e voglio calcolare una sua *derivata* o *integrale*, ma diventa troppo *"difficile"* da calcolare (pensa ai casi in cui ho integrali non esprimibili in termini di funzioni elementari); quindi usiamo il *"surrogato"* $\tilde{f}$.

Il campo della matematica che fornisce la risposta al problema iniziale si chiama la *teoria dell'approssimazione*.

Vedremo principalmente due tecniche, tra cui l'*interpolazione polinomiale* e *approssimazione ai minimi quadrati*. Notiamo che un'altra possibile risposta è fornita dall'*analisi di Fourier*, dove le funzioni analitiche sono T-periodiche.

Per scegliere l'approssimante $\tilde{f}$, selezioniamo uno *spazio funzionale vettoriale* $\mathcal{V}$ con supporto $\mathbb{K}$ e esprimiamo $\tilde{f}$ come una combinazione lineare di $\mathcal{V}$:

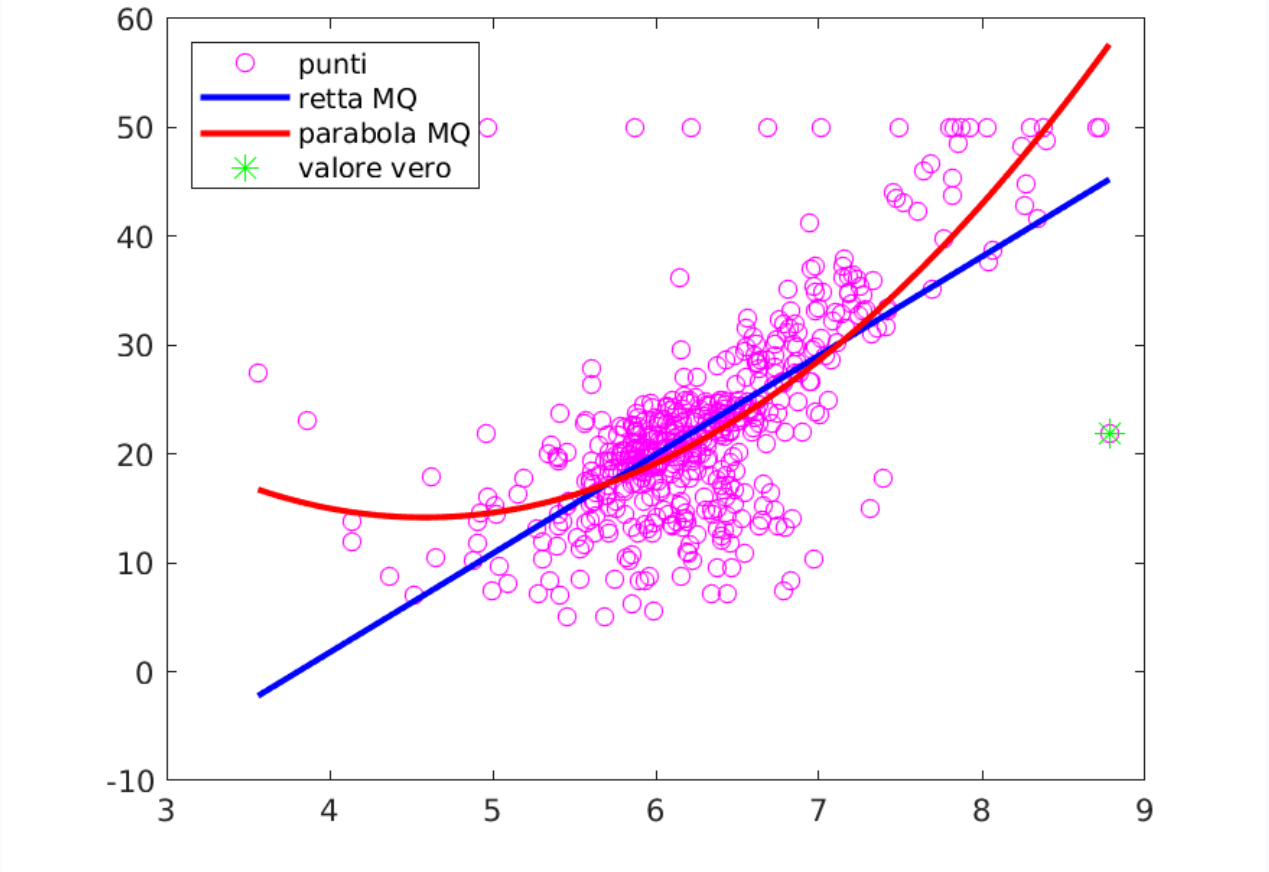
$$\tilde{f} = \langle \mathcal{V} \rangle = \sum_n \alpha_n \phi_n$$

dove $(\phi_n)_n \subset \mathcal{V}$ e $(\alpha_n)_n \subset \mathbb{K}$.

Quali sono i spazi $\mathcal{V}$ più comuni? Naturalmente lo *spazio* dei polinomi di grado $\mathtt{n}$, ossia

$$\mathcal{V} = \mathbb{P}_{\mathtt{n}} = \left\{ (a_n)_n \subset \mathbb{R} : \sum_{k \leq \mathtt{n}} a_k x^k \right\}$$

Dunque il problema di trovare la forma funzionale $\tilde{f}$ si riduce in trovare la successione $(a_n)_n$ opportuna.



Quando e se troveremo la funzione approssimata $\tilde{f}$, possiamo definire una sua *"misura d'errore"* rispetto a f come la seguente quantità:

$$r(x) := f(x) - \tilde{f}(x)$$

Questo si dice il *resto* dell'approssimazione.

Matrice di Vandermonde

Matrice di Vandermonde

X

Matrice di Vandermonde: definizione, rilevanza per l'interpolazione polinomiale, calcolo del determinante.

X

---

0. Voci correlate

- Interpolazione Polinomiale

1. Matrice di Vandermonde

Supponiamo di aver il set dei dati $(x_n, y_n)_{n\leq N}$ e di voler trovare un polinomio $p_N(x) \in \mathbb{P}_N$ tale che condizioni di interpolazione siano soddisfatte:

$$\forall i \leq N, p_N(x_i) = y_i$$

Scrivendo $p_N \sim (a_n)_{n\leq N}$ ossia $p_N(x) = \sum_{n\leq N} a_n x^n$, ottengo il seguente sistema lineare:

$$\begin{aligned}
p_N(x_0) = y_0 &\iff a_0 + a_1 x_0 + \ldots + a_N x_0^N = y_0 \\
p_N(x_1) = y_1 &\iff a_0 + a_1 x_1 + \ldots + a_N x_1^N = y_1 \\
\vdots &\iff \vdots \\
p_N(x_N) = y_N &\iff a_0 + a_1 x_N + \ldots + a_N x_N^N = y_N
\end{aligned}$$

Possiamo rappresentare la RHS in una forma più compatta usando le matrici: infatti notando che $p_N(x_i) = \langle \underline{a}, \underline{x_i} \rangle$ (dove $\underline{x_i} = (1, x_i, \ldots, x_i^N)^T$), possiamo vedere $(\underline{x_i})_i$ come i *"coefficienti"* e $\underline{a}$ come le *"soluzioni da trovare"*. Ossia ho la forma

$$\mathbf{V}\underline{a} = \underline{y}$$

dove $\underline{a} = (a_0, a_1, \ldots, a_N)^T$ e $\underline{y} = (y_0, y_1, \ldots, y_N)^T$ e $\mathbf{V}_{(i)} = (x_0^i, x_1^i, \ldots, x_N^i)^T$. Noto che $\mathbf{V} \in \mathbb{R}^{(N+1)\times(N+1)}$. Definisco la *matrice di Vandermonde* dunque come la seguente:

#Definizione

Definizione (matrice di Vandermonde).

Sia $N > 0$ e siano $(x_n)_{n \leq N} \subset \mathbb{R}$ i *punti nodali*. Allora la *matrice di Vandermonde* associata a $(x_n)_n$ è definita come la matrice $\mathbf{V} \in \mathbb{R}^{(N+1)\times(N+1)}$ tale che la sua riga i-esima è il vettore delle potenze di x_i da 0 fino a N, ovvero $\mathbf{V}_{(i)} = \mathbf{V}[:, i] = (1 \equiv x_i^0, x_i^1, \dots, x_i^N)^T$. La forma completa di $\mathbf{V}$ è

$$\mathbf{V} = \begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^N \\ 1 & x_1 & x_1^2 & \dots & x_1^N \\ \vdots & & & & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^N \end{pmatrix}$$

X

2. Determinante di Vandermonde

Calcoliamo $\mathbf{V}$. In particolare, dimostreremo induttivamente la seguente formula.

#Lemma

Lemma (determinante della matrice di Vandermonde).

Sia $\mathbf{V}$ la matrice di Vandermonde associata ai punti nodali $(x_n)_{n \leq N}$. Allora $\det \mathbf{V}$ viene calcolata come

$$\det \mathbf{V} = \prod_{0 \leq i < j \leq N} (x_j - x_i)$$

#Dimostrazione

DIMOSTRAZIONE del Lemma 2

Dimostriamo per *induzione*.

$N = 1$: Banalmente la matrice di Vandermonde per $N = 1$ è

$$\mathbf{V} = \begin{pmatrix} 1 & x_0 \\ 1 & x_1 \end{pmatrix}$$

Allora per definizione ho

$$\det \mathbf{V} = x_1 - x_0$$

Questo non è altro che la tesi per $N = 1$.

$N - 1 \implies N$: Sia $\mathbf{V}$ la matrice di Vandermonde *"completa"*, ossia

$$\mathbf{V} = \begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^N \\ 1 & x_1 & x_1^2 & \dots & x_1^N \\ \vdots & & & & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^N \end{pmatrix}$$

Scegliendo la matrice minore $\mathbf{V}_{N,1}$ ovvero togliendo l'ultima colonna e la prima riga otteniamo

$$\mathbf{V}_{N,1} = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{N-1} \\ \vdots & & & & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^{N-1} \end{pmatrix}$$

Come ipotesi induttiva supponiamo che

$$\det \mathbf{V}_{N,1} = \prod_{1 \le i < j \le N} (x_j - x_i)$$

L'idea cruciale della dimostrazione è quello di usare lo *sviluppo di Laplace del determinante* (Teorema 9) sulla prima riga, così in un modo ci *"riconduciamo"* all'ipotesi induttiva. Per farlo dobbiamo *"semplificare"* la matrice usando le *OE* (operazioni elementari, Definizione 2), così da dover *moltiplicare per* 1 la matrice minore e non effettuare altre somme strane.

Scegliamo, per $\forall i > 1, C_i \leftarrow C_i - x_0 \cdot C_{i-1}$, ovvero sottraiamo ogni colonna per la sua colonna precedente moltiplicata per x_0. In questo modo abbiamo la matrice equivalente

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^N \\ 1 & x_1 & x_1^2 & \dots & x_1^N \\ \vdots & & & & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^N \end{pmatrix} \leftarrow \begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & x_1 - x_0 & \dots & x_1^N - x_0 x_1^{N-1} \\ \vdots & & & \vdots \\ 1 & x_N - x_0 & \dots & x_N^N - x_0 x_N^{N-1} \end{pmatrix}$$

(Notiamo che effettuiamo tutte le OE *contemporaneamente*!; inoltre osserviamo che la scelta del punto nodale x_0 è arbitraria, la dimostrazione vale ugualmente anche se scegliamo un altro punto x_i)

Per le proprietà del determinante abbiamo che le determinanti delle due matrici trasformate in questo modo sono *uguali*. Allora sviluppando con Laplace la matrice RHS otteniamo

$$\det \mathbf{V} = 1 \cdot \det \mathbf{V}_{1,1} = \det \begin{pmatrix} x_1 - x_0 & \dots & x_1^N - x_0 x_1^{N-1} \\ \vdots & & \vdots \\ x_N - x_0 & \dots & x_N^N - x_0 x_N^{N-1} \end{pmatrix}$$

In ogni riga, per ogni elemento del tipo $x_i^j - x_0 x_i^{j-1}$ possiamo raccogliere per x_i, ottenendo dunque $(x_i^{j-1})(x_i - x_0)$. Per la *multilinearità* del determinante posso *"raccogliere fuori"* per ogni riga il termine $(x_i - x_0)$ moltiplicandolo per il determinante (Proposizione 1), ottenendo pertanto

$$\det\begin{pmatrix} x_1-x_0 & \dots & x_1^N-x_0x_1^{N-1} \\ \vdots & & \vdots \\ x_N-x_0 & \dots & x_N^N-x_0x_N^{N-1} \end{pmatrix} = (x_1-x_0)\dots(x_N-x_0)\det\begin{pmatrix} 1 & x_1 & \dots & x_1^{N-1} \\ 1 & x_2 & \dots & x_2^{N-1} \\ \vdots & & & \vdots \\ 1 & x_N & \dots & x_N^{N-1} \end{pmatrix}$$

Quest'ultima non è altro che la minore menzionata all'inizio, e dunque per ipotesi induttiva ho

$$(x_N-x_0)\dots(x_1-x_0)\det\begin{pmatrix} 1 & x_1 & \dots & x_1^{N-1} \\ 1 & x_2 & \dots & x_2^{N-1} \\ \vdots & & & \vdots \\ 1 & x_N & \dots & x_N^{N-1} \end{pmatrix} = (x_N-x_0)\dots(x_1-x_0)\prod_{1\le i<j\le N}(x_j-x_i)$$

Tuttavia $(x_N-x_0)\dots(x_1-x_0)=\prod_{0=i<j\le N}(x_j-x_i)$ e pertanto possiamo *"unire"* i prodotti, concludendo:

$$(x_N-x_0)\dots(x_1-x_0)\prod_{1\le i<j\le N}(x_j-x_i)=\prod_{0\le i<j\le N}(x_j-x_i)$$

CVD. ■

#Corollario

> Corollario (invertibilità della matrice di Vandermonde).
>
> Per ogni matrice di Vandermonde $\mathbf{V}$ associata ai dati $(x_n,y_n)_{n\le N}$ t.c. $\forall i\neq j, x_i\neq x_j$, vale che la matrice $\mathbf{V}$ è *invertibile* con inversa unica:
>
> $$\exists!\mathbf{V}^{-1}:\mathbf{V}\mathbf{V}^{-1}=\mathbf{V}^{-1}\mathbf{V}=\mathbb{1}$$

Polinomi di Lagrange

Polinomi di Lagrange

X

Polinomio di Lagrange. Definizione di polinomio di Lagrange. Base dello spazio dei N-polinomi. Lemma: calcolo dei polinomi di Lagrange nei punti nodali.

X

0. Voci correlate

- Interpolazione Polinomiale

1. Polinomi di Lagrange

Un modo per *"definire"* la base delle funzioni polinomiali $\mathbb{P}_N$ è semplicemente di usare il vettore di potenze $(x^0, x^1, \dots, x^N)$. Tuttavia si può definire una base *"migliore"* per $\mathbb{P}_N$, che rende più semplice *trovare* polinomi interpolatori per i dati.

#Definizione

> **Definizione (polinomi di Lagrange).**
>
> Sia $(x_n, y_n)_{n \leq N}$ una serie di dati. Supponiamo che $\forall i \neq j, x_i \neq x_j$. Allora definiamo il *polinomio elementare di Lagrange per il dato i-esimo* come
>
> $$L_i(x) := \prod_{j \neq i}^{N} \frac{x - x_j}{x_i - x_j} \in \mathbb{P}_N$$

Naturalmente L_i forniscono una base per $\mathbb{P}_N$ con $\mathcal{B}_{\mathcal{L}} = \{L_0, \dots, L_N\}$(ometteremo la dimostrazione). Ma qual è la peculiarità di questa base? Vedremo col seguente risultato

#Proposizione

> **Proposizione (i polinomi di Lagrange sono delle delta di Kronecker sui dati).**
>
> Siano $(L_i)_i$ dei polinomi di Lagrange sui dati $(x_n)_{n \leq N}$. Vale per $\forall i, j \leq N$ la seguente identità:
>
> $$L_j(x_i) = \delta_{ij} \ (1 \texttt{ if } i = j \texttt{ else } 0)$$

#Dimostrazione

DIMOSTRAZIONE della Proposizione 2

$i = j$: Il numeratore e il denominatore sono uguali, per cui si cancellano e abbiamo 1.

$i \neq j$: Uno dei fattori del numeratore si cancella (in particolare il j-esimo) e abbiamo dunque 0.

Interpolazione Polinomiale

Interpolazione Polinomiale

X

Formulazione del problema dell'interpolazione polinomiale. Esistenza unica del polinomio interpolatore. Dimostrazione con la matrice di Vandermonde, dimostrazione alternativa con i polinomi di Lagrange.

X

0. Voci correlate

- Matrice di Vandermonde
- Polinomi di Lagrange
- Introduzione alla Teoria dell'Approssimazione

1. Interpolazione Polinomiale

PROBLEMA. Sia $(x_n, y_n)_{n\leq N} \subset \mathbb{R}\times\mathbb{R}$ il *"dataset"*, ossia ho $N+1$ coppie di punti dove $(x_n)_n$ sono i nodi di interpolazione. Supponiamo che siano *"separabili"*, ovvero $\forall i \neq j, x_i \neq x_j$. Il problema dell'*interpolazione polinomiale* su $(x_n, y_n)_{n\leq N}$ si pone dunque come di trovare un polinomio $p_N(x)$ che *soddisfi tutte le condizioni di interpolazione*, ovvero

$$\forall i \leq N, p_N(x_i) = y_i$$

In tal caso diciamo che p_N è il *polinomio interpolante*.

---

2. Teorema dell'Interpolazione Polinomiale

Q. Possiamo garantirci l'esistenza del polinomio interpolante su un qualsiasi set di dati? Se sì, come possiamo costruire il *polinomio interpolante*?

Una risposta viene fornita dal seguente teorema:

#Teorema

> Teorema (dell'interpolazione polinomiale).
>
> Siano $(x_n, y_n)_{n\leq N} \subset \mathbb{R}^2$ t.c. $\forall i \neq j, x_i \neq x_j$. Allora $\exists! p_N \in \mathbb{P}_N$ tale che $\forall i \leq N, p_N(x_i) = y_i$

#Dimostrazione

DIMOSTRAZIONE del Teorema 1 (versione monomiale)
Le condizioni dell'interpolazione sono traducibili in termini del *sistema lineare* con la matrice di Vandermonde (Definizione 1), ovvero

$$p_N(\underline{x}) = \underline{y} \iff \mathbf{V}\underline{a} = \underline{y}$$

Ricordiamo che la sua *determinante* di $\mathbf{V}$ è calcolata come (Lemma 2)

$$\det \mathbf{V} = \prod_{0\leq i<j\leq N} (x_j - x_i)$$

Pertanto, per le nostre ipotesi il determinante non è nullo e dunque invertibile. Per *il teorema di Cramer* (Teorema 1) esiste ed è unica la soluzione $\underline{a}$, data da

$$\underline{a} = \mathbf{V}^{-1}\underline{y}$$

concludendo la dimostrazione. ■

#Osservazione

> Osservazione (svantaggi della versione monomiale).
>
> Notiamo che la *costruzione* data dalla prima versione della dimostrazione comporta il svantaggio di dover calcolare l'inversa della matrice $\mathbf{V}$, che è computazionalmente costosa. Inoltre, il problema dell'inversione di Vandermonde è un *problema mal-condizionato* (Definizione 1) per $n \geq 7$, per cui diventerebbe difficile calcolare il *polinomio giusto*.
>
> Dunque usiamo una versione più *"smart"* della dimostrazione, usando una base diversa per i polinomi $\mathbb{P}_N$.

#Dimostrazione

DIMOSTRAZIONE del Teorema 1 (versione Lagrange)
Questa versione della dimostrazione si articola in due fasi diverse.
!: Supponiamo che $\exists p_N \neq q_N$ t.c. entrambe verifichino le condizioni di interpolazione.
Definiamo il *polinomio differenza*

$$\Delta_N = p_N - q_N$$

Chiaramente $\Delta_N(x_i) = y_i - y_i = 0$, dunque ha $N+1$ radici. Tuttavia Δ_N è un polinomio di grado N, pertanto il fatto che ha $N+1$ radici ed è non-banale (ossia $\Delta_N \neq 0 \iff p_N \neq q_n$) deriva la contraddizione col *teorema fondamentale dell'algebra* (Teorema 3.2.).
$\exists$: Siano $(L_i(x))_i$ i *polinomi di Lagrange* calcolati su $(x_n)_n$. Notiamo che la condizione

$$p_N(x_i) = y_i$$

equivale a

$$y_i = \sum_n y_i \delta_{in}$$

Sapendo che $L_n(x_i) = \delta_{in}$ (Proposizione 2), ho che

$$p_N(x_i) = y_i = \sum_n y_n L_n(x_i)$$

Pertanto come soluzione ho

$$p_N(x) = \sum_{n \leq N} y_n L_n(x)$$

concludendo la dimostrazione. ■

Errore dell'Interpolazione Polinomiale

Errore dell'Interpolazione Polinomiale

X

Errore dell'interpolazione polinomiale. Lemma: regolarità del resto. Teorema dell'errore di interpolazione polinomiale. Osservazioni, dimostrazioni.

X

0. Voci correlate

- Interpolazione Polinomiale

1. Resto dell'Interpolazione Polinomaile

Sia $p_N(x)$ interpolante per i dati $(x_n, y_n)_{n\leq N}$ separabili t.c. $(x_n)_{n\leq N} \subset I \Subset \mathbb{R}$. Allora definiamo il resto N-esimo come la quantità

$$r_N(x) := f(x) - p_N(x)$$

Cosa possiamo dire su r_N per $x \in I$?

#Teorema

> Teorema (dell'errore di interpolazione polinomiale).
>
> Sia $N > 0$ fissato e $I = [a, b] \Subset \mathbb{R}$. Sia $f \in \mathcal{C}^{N+1}(I)$ e p_N il *polinomio di interpolazione* per $(x_n, f(x_n))_{n\leq N}$ separabile. In particolare assumiamo che $(x_n)_{n\leq N}$ sia *strettamente crescente*, ossia
>
> $$a \leq x_0 < x_1 < \ldots < x_N \leq b$$
>
> Allora per $\forall x \in [a, b], \exists \xi_x \in (a, b)$ t.c. valga l'identità
>
> $$r_N(x) = \frac{f^{(N+1)}(\xi_x)\omega_{N+1}(x)}{(N+1)!}$$
>
> Dove ω_N è il polinomio nodale definito come
>
> $$\omega_{N+1}(x) := \prod_{k\leq N}(x - x_k)$$

Il *polinomio nodale* ω_{N+1} è una *"misura"* della produttoria di tutte le distanze di un punto $x_0 \leq x \leq x_N$ da tutti i punti $(x_n)_n$.

Questo teorema ci dice che l'errore dipende dai seguenti fattori:

- Dalla derivata $N+1$-esima della funzione target f (primo termine sul numeratore, rosso)
- Il modo in cui scegliamo i nodi $(x_n)_n$ (secondo termine sul numeratore, verde)
- Il numero dei nodi che scegliamo (terzo termine sul numeratore, blu)

$$\frac{f^{(N+1)}(\xi_x) \cdot \omega_{N+1}(x)}{(N+1)!}$$

Purtroppo questo teorema *non ci garantisce* la convergenza uniforme, ossia

$$\|r_N\|_\infty \not\xrightarrow{N\to+\infty} 0$$

#Dimostrazione

DIMOSTRAZIONE del Teorema 1

Ricordiamo che dimostriamo la formula del resto *solamente per* $x \in [a,b]$.

$\exists i : x = x_i$: Ovviamente per le condizioni dell'interpolazione $r_N(x = x_i) = 0$. Ha senso? Sì, in quanto il polinomio nodale si annullerebbe (siccome uno dei membri della produttoria si cancellerebbe, ammazzando tutto).

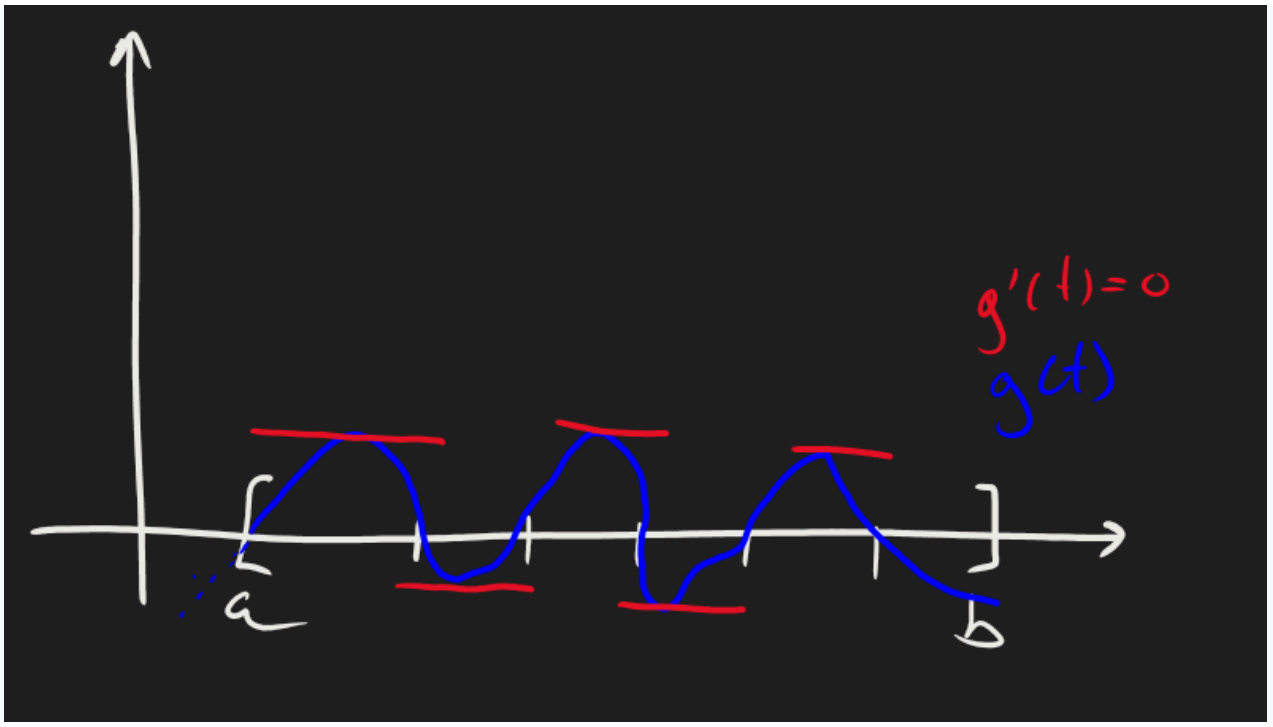
$\forall i, x \neq x_i$: In questo caso $\omega_{N+1}(x) \neq 0$. Fissiamo x e definiamo la seguente funzione ausiliaria:

$$g(t) := r_N(t) - \omega_{N+1}(t)\frac{r_N(x)}{\omega_{N+1}(x)}$$

Sicuramente $g \in \mathcal{C}^{N+1}(I)$, siccome per chiusura della continuità sulla somma si ha $r_N \in \mathcal{C}^{N+1}(I)$ (e banalmente ω_{N+1} è un polinomio $\mathcal{C}^\infty$). Allora si ricava che f, g sono *"ugualmente regolari"*.

Notiamo che per $t = x$ si ha $g(t) = g(x) = 0$. Poi ovviamente $\forall t = x_i, g(t) = 0$. Pertanto g si *annulla in* $N+2$ punti. Pertanto abbiamo $N+2$ punti valutati con lo stesso valore e per il *teorema di Rolle* (Teorema 1) esistono $N+1$ punti che annullano la derivata g'.

In altre parole, $N+2$ zeri di g inducono $N+1$ zeri. In una maniera più formale, si può dire che $x, x_0, \dots, x_N$ inducono la suddivisione Δ di $[a,b]$ dove ogni intervallo $\delta \in \Delta$ ha estremi valutati ugualmente.



Posso ripetere il ragionamento analogo per $N+1$ zeri di g' che inducono N zeri di g'', fino a chè raggiungo la $N+1$-esima derivata con *almeno un zero* di g^{N+1}. Pertanto

$$\exists \xi_x \in [a,b] : g^{(N+1)}(\xi_x) = 0$$

Come valutiamo la derivata $g^{(N+1)}$? Ricordando che *dipende* solo da t, abbiamo che

$$g^{(N+1)}(t) = r_N^{(N+1)}(t) - \omega_{N+1}^{(N+1)}(t)\frac{r_N(x)}{\omega_{N+1}(x)}$$

Tuttavia notiamo che ω_{N+1} è un *polinomio monico*, per cui la sua derivata $N+1$-esima è costante con valore $(N+1)!$. Pertanto, valutando $g^{(N+1)}(\xi_x)$ si ha

$$g^{(N+1)}(\xi_x) = 0 \iff r_N^{(N+1)}(\xi_x) - (N+1)!\frac{r_N(x)}{\omega_{N+1}(x)} = 0$$

Prima di isolare $r_N(x)$, osserviamo che $r_N^{(N+1)} = f^{(N+1)} - \underbrace{p_N^{(N+1)}}_{0} = f^{(N+1)}$ e pertanto concludiamo:

$$r_N(x) = \frac{f^{(N+1)}(\xi_x)\omega_{N+1}(x)}{(N+1)!}$$

CVD. ■

X

2. Maggiorazione del Resto

Q. Come conosciamo ξ_x dell'enunciato?
A. Non lo si conosce, infatto quello che faremo è quello di *maggiorare* il resto r_N per ottenere una stima. In particolare faremo

$$|r_N(x)| \leq \frac{\sup_{x\in[a,b]} |f^{(N+1)}(x)|}{(N+1)!} \sup_{x'\in[a,b]} |\omega_{N+1}(x')|$$

Per Weierstrass possiamo *"sostituire"* sup $\leftarrow$ max.

Scelta dei Punti Nodali

Scelta dei Punti Nodali

X

Problema della scelta dei punti nodali per l'interpolazione di una funzione conosciuta. Punti equidistanti. Metodo di Čebyšëv-Lobatto, di Čebyšëv. Intuizione geometrica di Čebyšëv-Lobatto.

X

---

0. Voci correlate

- Interpolazione Polinomiale

1. Punto Nodali Equidistanti

PROBLEMA. Sia $\mathcal{I} = [a, b] \Subset \mathbb{R}$ e $N > 0$ fissato. Supponiamo di conoscere $f : \mathcal{I} \longrightarrow \mathbb{R}$. Come possiamo *scegliere* i punti nodali $(x_n)_{n\leq N} \subset \mathcal{I}$?

Un primo esempio banale è quello di scegliere i *punti* che siano *equispaziati*. Ovvero, scegliamo x_n dato da

$$x_k = a + k\frac{b-a}{N} = a + t_k, t_k := k\frac{|\mathcal{I}|}{N}$$

X

---

2. Metodi di Čebyšëv

#Definizione

> Definizione (nodi di Čebyšëv-Lobatto).
>
> Sia $\mathcal{I} = [a, b]$ e $N > 0$. Allora si definiscono i *nodi di Čebyšëv-Lobatto* come quelli dati dall'equazione
>
> $$x_k = \frac{a+b}{2} + \frac{b-a}{2}t_k$$
>
> con

$$t_k := -\cos\left(\frac{k\pi}{N}\right)$$

#Definizione

Definizione (nodi di Čebyšëv).

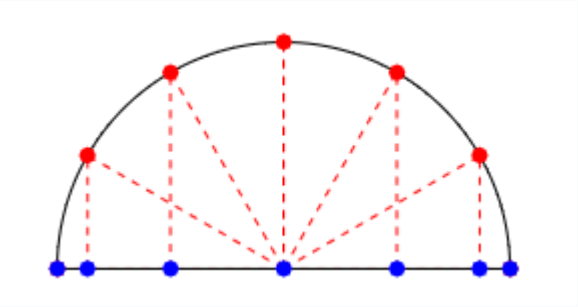
Sia $\mathcal{I} = [a, b]$ e $N > 0$. Allora si definiscono i *nodi di Čebyšëv* come quelli dati dall'equazione

$$x_k = \frac{a+b}{2} + \frac{b-a}{2} t_k$$

con

$$t_k := -\cos\left(\frac{(2k+1)\pi}{2(N+1)}\right)$$

Geometricamente i metodi di Čebyšëv corrispondono alla *proiezione di N nodi della semicirconferenza unitaria $\Omega^+ : \{\theta \in [0, \pi] : (\cos\theta, \sin\theta)\}$ sulla retta delle ascisse*.



Convergenza dell'Interpolazione Polinomiale

Convergenza dell'Interpolazione Polinomiale

X

Convergenza degli interpolanti polinomiali. Calcolo esplicito della norma infinita dello scarto. Controesempio: fenomeno di Runge.

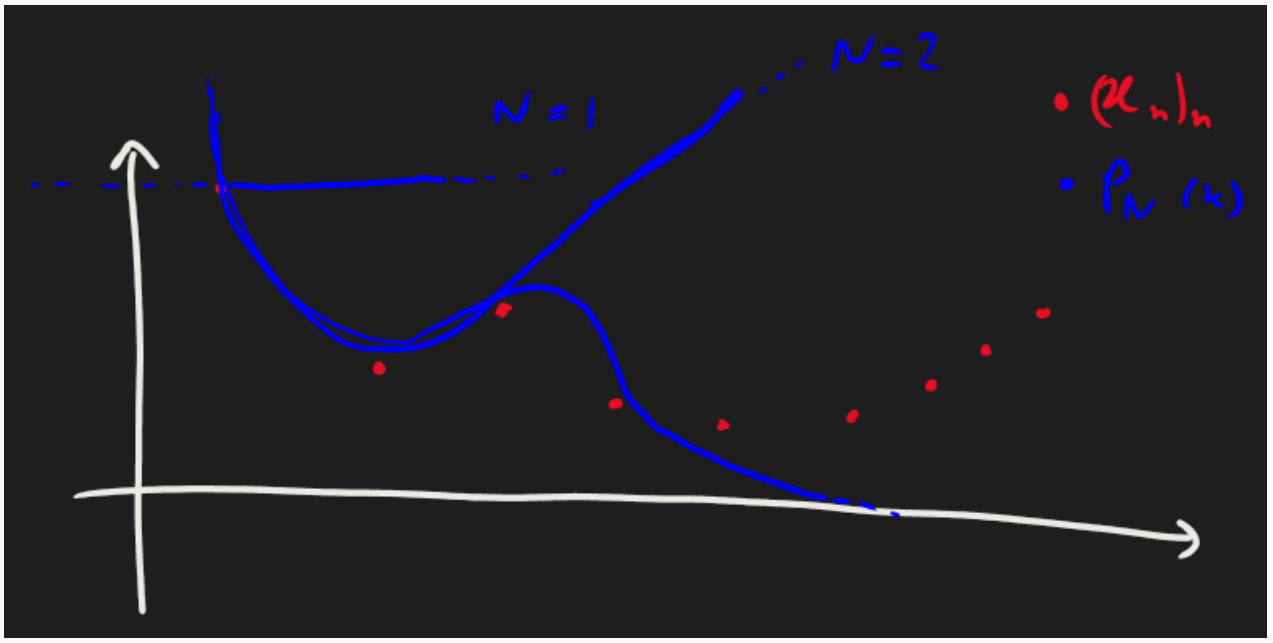
X

0. Voci correlate

- Interpolazione Polinomiale
- Errore dell'Interpolazione Polinomiale
- Scelta dei Punti Nodali
- Convergenza Puntuale e Uniforme per Successioni di Funzioni

1. Definizione di Convergenza

Sia $\mathcal{I} = [a, b]$ e $(x_n)_{n\in\mathbb{N}} \subset \mathcal{I}$. Costruisco, per ogni $N > 0$ un vettore di punti nodali $(x_n)_{n\leq N} \subset (x_n)_{n\in\mathbb{N}}$, un interpolante polinomiale $p_N \in \mathbb{P}_N$ di grado N.



Quando si ha che il *polinomio* p_N converge uniformemente alla funzione target f? In altre parole, quando lo scarto $r_N = f - p_N$ va a zero?

Richiamiamo la seguente definizione di convergenza uniforme:

#Definizione

> Definizione (convergenza uniforme di interpolanti polinomiali).
>
> Una *successione di interpolanti polinomiali* $(p_N)_{N\in\mathbb{N}}$ *converge uniformemente* ad f se e solo se vale che la norma infinita dello scarto tende a zero. Ossia $\lim_n \|f - p_N\|_\infty = 0$. Più esplicitamente, è
>
> $$\lim_n \sup_{x\in\mathcal{I}} |f(x) - p_N(x)| = 0$$

Osserviamo che se assumiamo che f è sufficientemente regolare, allora per Weierstrass possiamo usare max al posto di sup.

X

2. Fenomeno di Runge

Conoscendo la *formula* del resto r_N, enunciamo il seguente:

#Proposizione

> Proposizione (calcolo esplicito della norma infinita dell'errore).

Per una successione di interpolanti polinomiali p_N costruita su f si ha che

$$\|f - p_N\|_\infty = \frac{\|f^{(n+1)}\|_\infty \cdot \|\omega_{n+1}\|_\infty}{(n+1)!}$$

Nonostante il fatto che abbiamo un *fattoriale* sul denominatore, la convergenza uniforme non è garantita. Infatti le norme di $f^{(n+1)}$ e ω_{n+1} sono entrambi dipendenti da n, e pertanto se sono sufficientemente grandi posso portare alla divergenza.

Per un esempio tipico vedere il fenomeno di Runge.

X

---

3. Risultati Principali sulla Convergenza

Dimostriamo dei risultati principali relativi alla *convergenza* e *divergenza* di polinomi interpolatori

#Teorema

Teorema (Faber, 1914).

Per ogni distribuzione di nodi $(x_n)_n \subset \mathcal{I} \Subset \mathbb{R}$ esiste una funzione $f \in \mathcal{C}^0(\mathcal{I})$ tale che l'errore di interpolazione non converge a zero:

$$\lim_n \|r_n^f\|_\infty > 0$$

#Teorema

Teorema (esistenza dei nodi interpolanti convergenti).

Per ogni funzione f continua in $\mathcal{I} \Subset \mathbb{R}$ esiste una scelta di *distribuzione di nodi interpolanti* $(x_n)_n \subset \mathbb{R}$ tale che l'errore del polinomio interpolante tenda a 0:

$$\lim_n \|r_n^f\|_\infty = 0$$

#Teorema

Teorema (di Bernstein).

Sia $\mathcal{I} \Subset \mathbb{R}$. Per ogni funzione $f : \mathcal{I} \longrightarrow \mathbb{R} \in \mathcal{C}^1$, la scelta dei nodi interpolanti $(x_n)_n \subset \mathcal{I}$ fatta usando un metodo di Čebyšëv (Čebyšëv-Lobatto o Čebyšëv) porta alla convergenza del polinomio interpolante:

$$\lim_n \|r_n^f\|_\infty = 0$$

Tutti i teoremi sono omessi, lasciati da verificare sperimentalmente. ■

Fenomeno di Runge

Fenomeno di Runge

X

Study case di un caso divergente dell'interpolazione polinomiale.

X

---

0. Voci correlate

- Interpolazione Polinomiale
- Convergenza dell'Interpolazione Polinomiale

1. Fenomeno di Runge

Un esempio noto della divergenza di un'interpolazione polinomiale di una funzione (Interpolazione Polinomiale):

#Teorema

> Teorema (fenomeno di Runge).
>
> Esiste una funzione $f : \mathcal{I} \Subset \mathbb{R} \longrightarrow \mathbb{R}$ di classe $\mathcal{C}^\infty$ tale che il resto $r_n(x)$ sul polinomio interpolante $p_n(x)$ fittata una scelta di punti $(x_n)_n \subset \mathcal{I}$ diverge:
>
> $$\sup_{n\in\mathbb{Z}} \|r_n\|_\infty = +\infty$$

#Dimostrazione

DIMOSTERAZIONE del Teorema 1

La funzione che dimostra il teorema è la c.d. *funzione di Runge*:

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

La consideriamo sull'intervallo $\mathcal{I} = [-5, 5]$. Quindi il suo resto n-esimo è maggiorata, in modulo, come

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

Si dimostra (dim. omessa) che per *nodi equispaziati* con ampiezza $h_n = \frac{\mu(\mathcal{I})}{n}$ la maggiorazione del polinomio nodale è

$$\max_{x \in \mathcal{I}} |\omega_{n+1}(x)| \leq \frac{h_n^{n+1} n!}{4}$$

Pertanto rimpiazzandola in (1) otteniamo

$$\max_{x \in \mathcal{I}} |r_n(x)| \leq \max_{x \in \mathcal{I}} |f^{(n+1)}(x)| \cdot \frac{h_n^{(n+1)} \cdot \cancel{n!}}{4(n+1)\,\cancel{(n!)}}$$

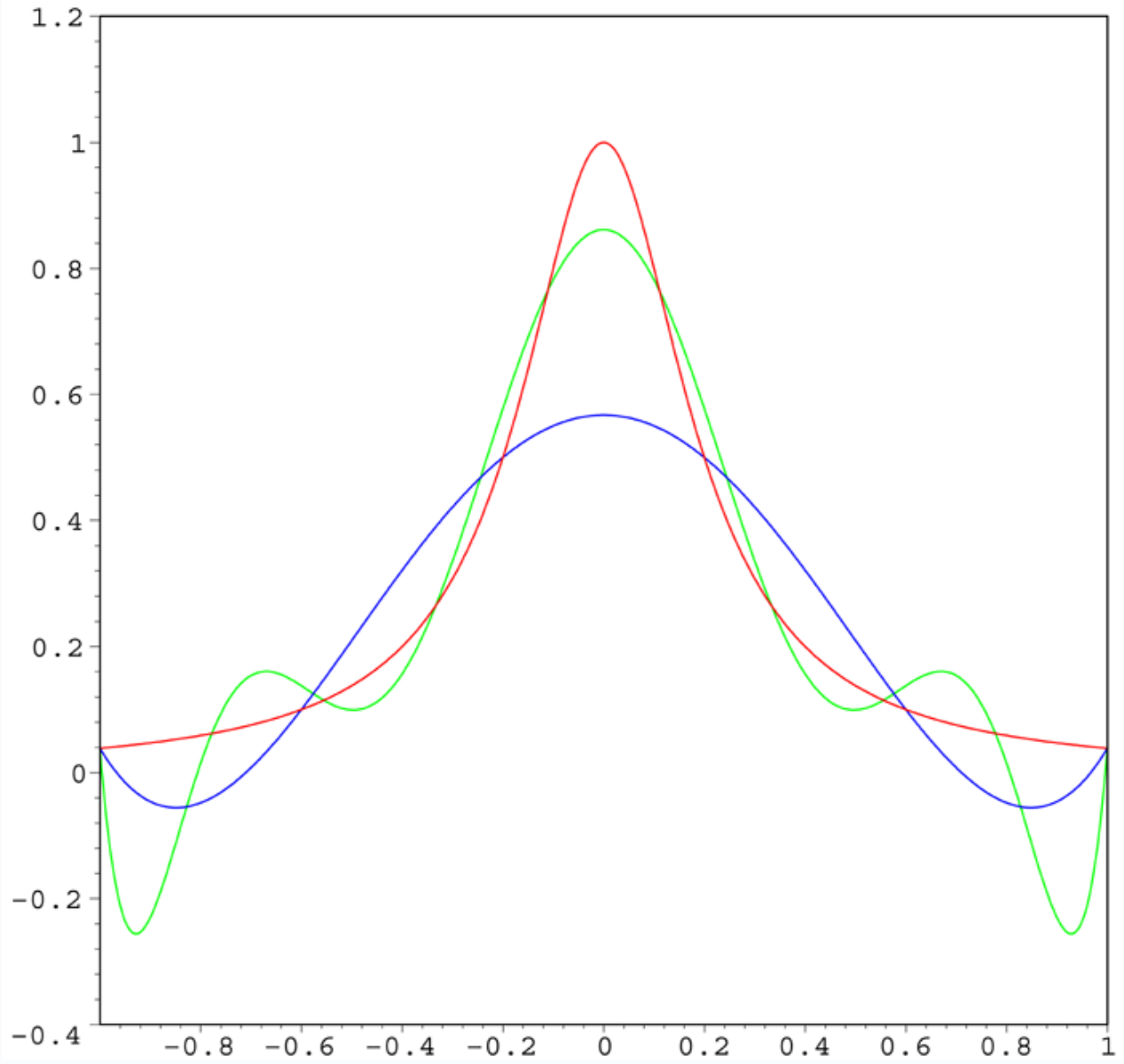
Chiaramente h_n è un infinitesimo, da cui

$$\lim_n \frac{h_n^{(n+1)}}{4(n+1)} = 0$$

Tuttavia si dimostra (dim. omessa) che la norma infinita della derivata $n+1$-esima di f si comporta asintoticamente come supera l'ordine di infinitesimo dell'infinitesimo in (2); pertanto il limite r_n diverge:

$$\lim_n |r_n(x)| = +\infty$$

Concludendo. ■



X

2. Nodi e Polionomio di Čebyšëv

Notiamo che questo problema è rimediabile per una scelta diversa di $(x_n)_n \subset \mathcal{I}$, minimizzando ulteriormente l'errore su ω_{n+1}. Andiamo a definire il *comportamento* su ω_{n+1} quando usiamo CGL (Definizione 1)

Definiamo il polinomio di Čebyšëv come segue:
#Definizione

> **Definizione (polinomio di Čebyšëv).**
>
> Si definisce un *polinomio di Čebyšëv* di grado n come la funzione $T_n : [-1, 1] \longrightarrow [-1, 1]$ posta come
>
> $$T_n(x) := \cos(n \arccos(x)) \in \mathbb{P}_n$$

Una formula per *calcolare* T_n è definita dalla seguente formula di ricorrenza:

$$(T_n)_n : \begin{cases} T_0(x) = 1 \\ T_1(x) = x \\ T_n(x) = 2x \cdot T_n(x) - T_{n-1}(x) \end{cases}$$

Notiamo che T_n è un polinomio del tipo

$$T_n(x) = 2^{n-1}x^n + \dots$$

Pertanto T_n *non* è monico. Lo normalizziamo e definiamo

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

Si dimostra il seguente risultato:
#Teorema

> Teorema (relazione tra Čebyšëv e polinomio nodale).
>
> Per $(x_n)_n$ distribuiti con un metodo di Čebyšëv si ha che il *polinomio nodale* $\omega_{n+1}(x)$ è equivalente a
>
> $$\omega_{n+1}(x) = \bar{T}_{n+1}(x)$$

La dimostrazione è omessa. Tuttavia notiamo che $\bar{T}$ è un coseno, pertanto sicuramente $\|T_n\|_\infty \leq 1$ e quindi

$$\|\bar{T}_n\|_\infty = \frac{1}{2^{n-1}}$$

(Notiamo che questa vale per $\mathcal{I} = [-1, 1]$; tuttavia si può generalizzare su $\mathcal{I} = [a, b]$ aggiungendo un termine dipendente da $\mu(\mathcal{I})$)

Però i nodi di Čebyšëv sono ottimali, in quanto si ha il seguente risultato:

$$\|\bar{T}_n\|_\infty \leq \|\Pi_n\|_\infty$$

Dove $\Pi_n(x)$ è un polinomio monico di grado n qualsiasi. Pertanto la scelta dei nodi di Čebyšëv minimizza il termine $\|\omega_{n+1}\|_\infty$ nella formula dell'errore (Proposizione 2). ■

Differenza Divisa

Differenza Divisa

X

Defininizione di una differenza divisa per una funzione relativa a dei dati. Metodo tabellare per calcolare la differenza divisa di una funzione.

X

0. Voci correlate

- Introduzione alla Teoria dell'Approssimazione
- Rapporto Incrementale

1. Differenza Divisa, Definizione

#Definizione

> Definizione (differenza divisa di una funzione relativa a dei punti).
>
> Sia $(x_n)_{n\leq N} \subset \mathcal{I} \Subset \mathbb{R}$ e $f : \mathcal{I} \longrightarrow \mathbb{R}$, tali che $\forall i \neq j, x_i \neq x_j$. Per convenzione poniamo $\mathcal{I} = [x_0, x_N]$.
>
> Definiamo *per induzione* la *differenza divisa* di ordine k-esimo di f come:
> $k = 0$: Semplicemente la funzione valutata in x_i, ossia
>
> $$f[x_i] = x_i$$
>
> $k \implies k+1$: Dati $k+1$ punti $x_0, \dots, x_k$, definiamo la sua differenza divisa come il *rapporto incrementale* tra i primi k punti e gli ultimi k punti, valutando le funzioni con la loro differenza divisa:
>
> $$f[x_0, \dots, x_k] = \frac{f[x_1, \dots, x_k] - f[x_0, \dots, x_{k-1}]}{x_k - x_0}$$
>
> Così via fino a $k = N$, se si sceglie di partire da x_0.

Osserviamo che per $k = 1$ (ordine primo), abbiamo che $\exists \xi : f'(\xi) = f[x_i, x_j] = R^f_{x_i}(x_j)$.

#Proposizione

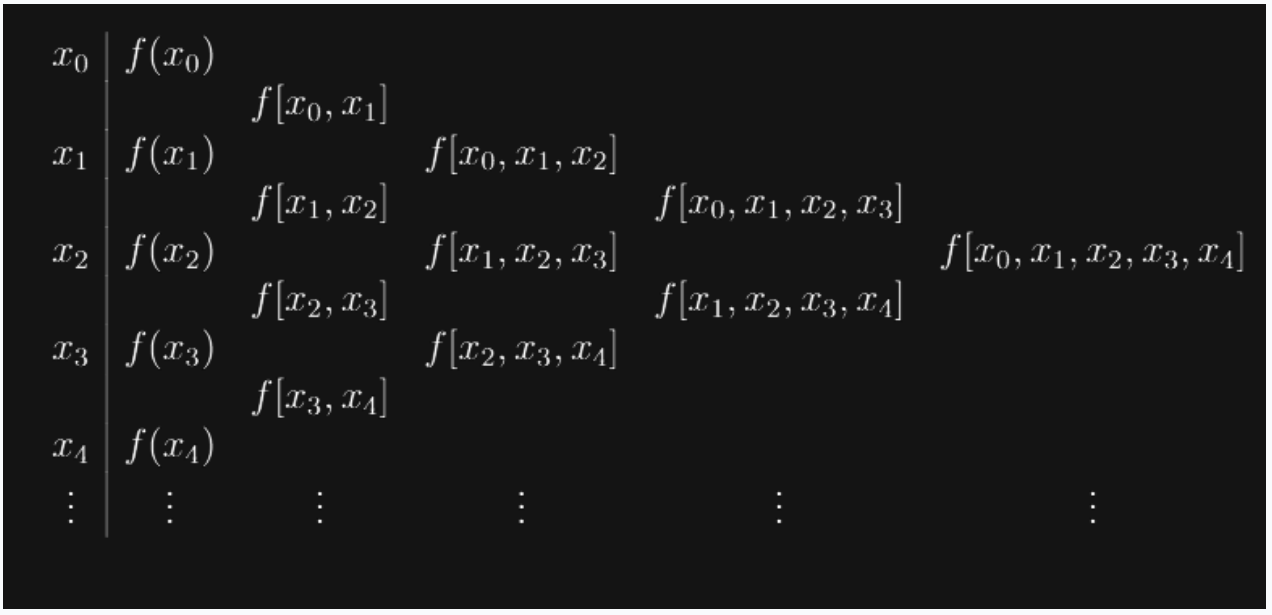
> Proposizione (commutabilità delle differenze divise).
>
> Sia $(\tilde{x}_n)_{n\leq N}$ una *permutazione fondamentale* di $(x_n)_{n\leq N}$, ossia dato $\pi : \mathbb{N} \longrightarrow \mathbb{N}$ *biiettiva* si ha $\tilde{x}_n = x_{\pi(n)}$. Allora si gode dell'uguaglianza, data f definita sull'immagine di $(x_n)_n$,
>
> $$f[x_0, \dots, x_N] = f[\tilde{x}_0, \dots, \tilde{x}_N]$$

Pertanto possiamo commutare $[x_0, \dots, x_N]$ come vogliamo e il risultato non cambia. Pertanto la Definizione 1 è ben posta senza perdere di generalità. La dimostrazione è omessa, ma è facilmente dimostrabile per induzione (partendo da $k = 1$).

PROBLEMA. Data f, $(x_0, \dots, x_n)$, come possiamo *calcolare* tutte le differenze divisi possibili in una maniera efficace?

Per farlo si può avvalersi della costruzione di una *tabella piramidale*, dove:

- Ogni strato rappresenta un ordine
- Il *"pre-strato"* è composto da tutti i punti $x_0, \dots, x_n$
- Il primo strato ($i = 0$) è dato da $f[x_k]$
- Ogni elemento dei strati successivi ($i > 0$) è dato dal *"merging"* di due elementi dello stato precedenti, dove si fa il semplice rapporto incrementale



Notiamo che aggiungento un punto x_{n+1}, basta effettuare un numero costante di calcoli per ricreare la tabella. Questa osservazione diventa fondamentale nel contesto della *interpolazione polinomiale con Newton*.

Come una *specie di spoiler*, anticiperemo che ai fini dell'interpolazione *useremo* solo il primo elemento di ogni strato. Ossia differenze divise dove compaiono x_0.

X

2. Differenza Divisa per Punti Coincidenti

Q. La Definizione 1 vale per *punti distinti* $x_i \neq x_j$. Possiamo trovare un *modo* per espandere tale nozione su punti coincidenti $x_i = x_j = \bar{x}$?

A. Una risposta generalizzante può essere data usando la nozione di *limite*, infatti

$$\lim_{x \to x_0} f[x, x_0] = \lim_{x \to x_0} R^f_{x_0}(x) = f'(x_0)$$

è ben definita per funzioni differenziabili in x_0. Pertanto definiamo

$$f[x_0, x_0] := f'(x_0)$$

Per trovare la definizione con tre punti coincidenti, troviamo (senza dimostrare)

$$f[x_0, x_0, x_0] := \lim_{(x_1, x_2) \to (x_0, x_0)} f[x_0, x_1, x_2] = \frac{f''(x_0)}{2!}$$

Pertanto generalizzeremo su $k \in \mathbb{N}$ con la seguente definizione:

#Definizione

> Definizione (differenza divisa per punti coincidenti).
>
> Definiamo la *differenza divisa* per k punti coincidenti $(x_0, \dots, x_0)$ come
>
> $$f[x_0 \dots, x_0] = \frac{f^{(k)}(x_0)}{k!}$$

Metodo delle Differenze Divise di Newton

Interpolazione Polinomiale secondo Newton

X

Metodo delle differenze divise di Newton. Definizione di base di Newton per lo spazio degli n-polinomi. Teorema: formula di Newton. Motivazione per la formula di Newton. Metodo delle differenze divise per punti coincidenti. Osservazione: formula di Taylor.

X

0. Voci Correlate

- Interpolazione Polinomiale
- Differenza Divisa
- Formula di Taylor

1. Motivazioni

Abbiamo dimostrato l'esistenza ed unicità del polinomio interpolatore su dati separabili (Teorema 1). Uno dei metodi più convenienti è quello di *Lagrange*, ossia usare dei polinomi che si comportino come delle delta di Kroenecker.

Tuttavia, questa scelta conviene sempre? Un problema legato all'interpolazione con Lagrange è il fatto che i polinomi vanno sempre *ricalcolati* se deciamo di aggiungere ulteriori punti sui dati $(x_n)_n$.

Il metodo di *Newton* va a sopperire questo problema, usando il concetto delle *differenze divise*.

X

2. Interpolazione con Newton

#Definizione

Definizione (base polinomiale di Newton).

La *base dei polinomi di Newton* di n-esimo grado, sui punti $x_0, \dots, x_n$ è dato da

$$\mathcal{B}_N = \left\{ 1, (x - x_0), (x - x_0)(x - x_1), \dots, \prod_{0 \le k \le n-1} (x - x_k) \right\}$$

Notiamo che $\prod_{0 \le k \le n-1}(x - x_k)$ è un *polinomio di grado* n, in quanto è incluso $(x - x_0)$.

#Teorema

Teorema (interpolazione con Newton).

Sia $(x_n)_{n \le N}$ una sequenza crescente, su cui definiamo $\mathcal{I} = [x_0, x_N]$. Sia $f : \mathcal{I} \longrightarrow \mathbb{R}$. Allora la *funzione* f è data dalla seguente formula:

$$\begin{aligned} f(x) &= f[x_0] + (x - x_0) f[x_0, x_1] + \dots + \prod_{0 \le k \le N-1} (x - x_k) \cdot f[x_0, \dots, x_N] \\ &\quad + f[x, x_0, \dots, x_N] \omega_{N+1}(x) \\ &:= p_N(x) + r_N^f(x) \end{aligned}$$

Definiamo p_N il polinomio di Newton, r_N^f il resto.

#Osservazione

Osservazione (forma unica del resto).

Osserviamo che, data l'unicità di p_N, otteniamo che il resto è unico per cui vale l'uguaglianza

$$f[x, x_0, \dots, x_N] \cdot \omega_{N+1}(x) = \frac{f^{(N+1)}(\xi_x)\omega_{N+1}(x)}{(N+1)!}$$

per il teorema dell'errore dell'interpolazione polinomiale (Teorema 1), e pertanto se f è $\mathcal{C}^{N+1}$ allora $\exists \xi_x$ tale che

$$f[x, x_0, \dots, x_N] = \frac{f^{(N+1)}(\xi_x)}{(N+1)!}$$

Notiamo che *"droppando"* l'ordine di uno rimuovendo x, otteniamo la formula

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

#Osservazione

> Osservazione (vantaggio di Newton).
>
> Il *vantaggio fondamentale* di questa formula è la seguente:
>
> $$p_{N+1}(x) = p_N(x) + (x - x_0)\dots(x - x_N)f[x_0, \dots, x_{N+1}]$$
>
> Ovvero, aggiungendo punti su punti dobbiamo solo aggiungere un piccolo termine alla formula.

#Dimostrazione

DIMOSTRAZIONE del Teorema 2

Nota: dimostrazione non svolta in classe

Diamo solo un'idea informale della dimostrazione, dimostrando la formula *per induzione*:
$k = 0$: Notiamo che dati dei punti x, x_0 abbiamo per definizione

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

da cui

$$f(x_0) - f(x) = (x_0 - x)f[x, x_0]$$

e quindi

$$f(x) = f(x_0) + f[x, x_0](x - x_0)$$

Questa verifica la formula del teorema. Infatti per $x = x_0$ abbiamo $f(x) = f(x_0)$, verificando la condizione interpolante.
$k = 1$: Analogamente prendiamo x, x_0, x_1 da cui

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

quindi effettuando una operazione analoga ottengo

$$f[x, x_0] = f[x_0, x_1] + (x - x_1)f[x, x_0, x_1]$$

Sostituendo in (1) ottengo

$$f(x) = f(x_0) + (x - x_0)f[x_0, x_1] + (x - x_1)(x - x_0)f[x, x_0, x_1]$$

Notiamo nuovamente che per $x = x_0$ ottengo $f(x) = f(x_0)$ e per $x = x_1$ ottengo

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

Quindi queste soddisfano le condizioni di interpolazione, e otteniamo anche il resto nella forma desiderata. Così iterando otteniamo la tesi. ■

Per una vera dimostrazione per induzione vedere il seguente paragrafo.

DIMOSTRAZIONE del Teorema 2

$k \implies k+1$: Supponiamo, per ipotesi induttiva, che

$$f(x) = f[x_0] + \ldots + (x - x_0)\ldots(x - x_k)f[x, x_0, \ldots, x_k]$$

Calcolando $f[x, x_0, \ldots, x_k]$ otteniamo

$$f[x, x_0, \ldots, x_{k+1}] = \frac{f[x_0, \ldots, x_{k+1}] - f[x, x_0, \ldots, x_k]}{x_{k+1} - x}$$

Isolando $f[x, x_0, \ldots, x_k]$ otteniamo

$$f[x, x_0, \ldots, x_k] = f[x_0, \ldots, x_{k+1}] + (x - x_{k+1})f[x_0, \ldots, x_{k+1}]$$

Sostituendo in (2) ottengo la tesi, i.e.

$$f(x) = f[x_0] + \ldots + (x - x_0)\ldots(x - x_k)f[x_0, \ldots, x_{k+1}] + (x - x_0)\ldots(x - x_{k+1})f[x, x_0, \ldots$$

Concludendo. ■

X

2. Fittare su Derivate

Notiamo che un altro *vantaggio* del metodo di Newton è che abbiamo un modo per *sfruttare* le derivate di $f(x)$. Infatti basta accettare la definizione

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

Da cui abbiamo un polinomio più opportuno. Notiamo che per x_0 tutti uguali otteniamo proprio la formula di Taylor:

$$f(x) = \sum_{k<N} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + f[x, x_0, \ldots, x_N]\omega_{N+1}(x)$$

Sostituendo il resto da Teorema 2 otteniamo proprio la formula di Taylor col resto di lagrange (Teorema 2.2.). ■

Approssimazione ai Minimi Quadrati

Approssimazione ai Minimi Quadrati

X

Approssimazione ai minimi quadrati (OLS) delle funzioni. Contesto, formulazione del problema. Proprietà del polinomio OLS. Modi per risolvere il problema OLS: metodo analitico, metodo geometrico. Intuizione geometrica del metodo geometrico.

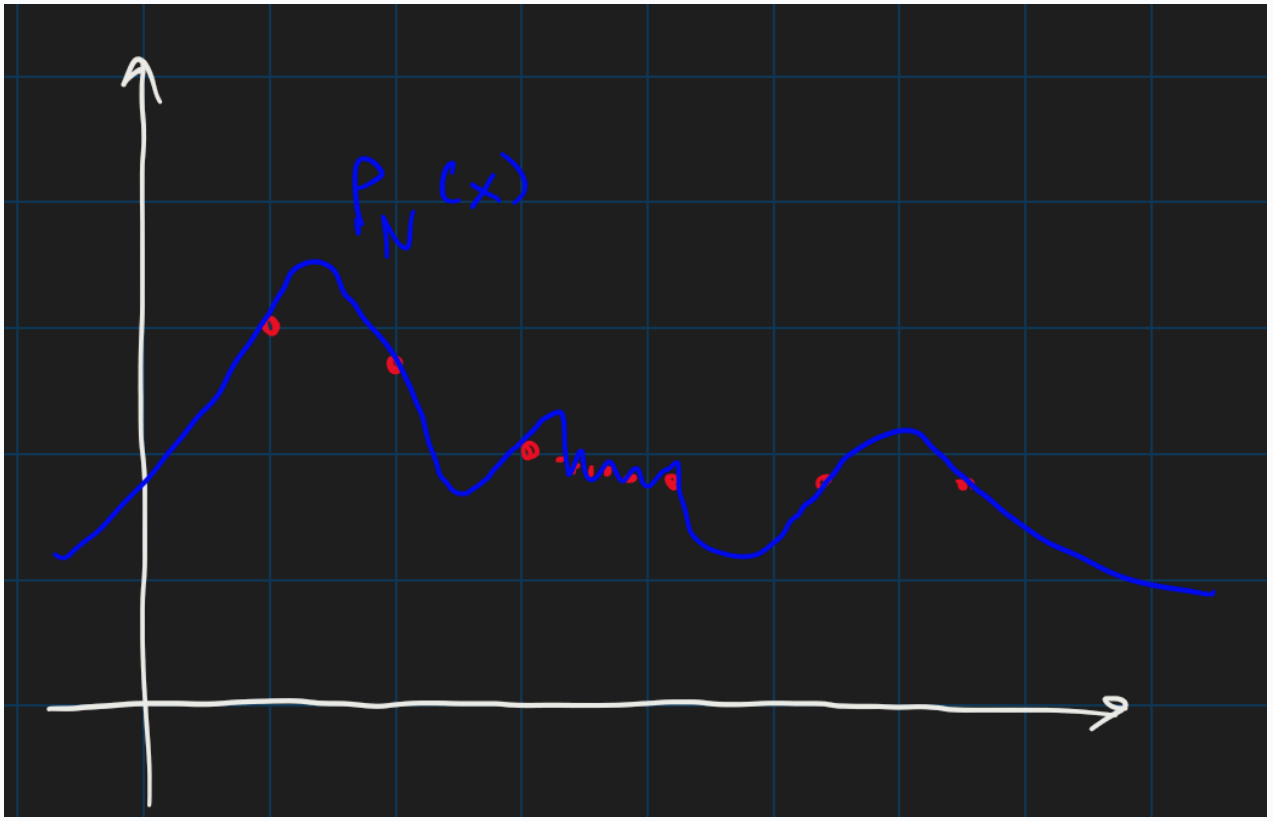
X

0. Voci correlate

- Introduzione alla Teoria dell'Approssimazione
- Test del Gradiente

1. Problema dell'Approssimazione OLS

CONTESTO. Sia $(x_n, y_n)_{n \leq N} \subset (\mathcal{I} \times f(\mathcal{I}))$, tale che $\forall i \neq j, x_i \neq x_j$ (ossia abbiamo un dataset separabile). Supponiamo di avere un numero N *"molto grande"*, ricavati tutti da dei dati sperimentali. Conviene usare l'approssimazione polinomiale? Ovviamente no, in quanto avrei un polinomio N-dimensionale, da cui avre un *"overfit"* drastico.



PROBLEMA. Pertanto l'obbiettivo dell'approssimazione OLS è quello di individuare uno spazio funzionale

$$\mathcal{F}_\varphi = \langle \varphi_0, \dots, \varphi_M \rangle$$

quindi

$$f \in \mathcal{F}_\varphi \iff f(x) = \sum_{m \leq M} \alpha_m \varphi_m(x)$$

in particolare $M \ll N$, usando *meno* funzioni di quanto ne avrei usati con l'interpolazione polinomiale. Per trovare $\varphi \in \mathcal{F}_\varphi$ vogliamo *minimizzare* la seguente *"misura d'errore"* (in Machine Learning viene riferita come *"loss"*):

$$\mathcal{L}(\varphi) = \|f - \varphi\|^2_{2,\texttt{discrete}} = \sum_{n \leq N} |f(x_n) - \varphi(x_n)|$$

quindi

$$\varphi^* = \arg\min_{\varphi \in \mathcal{F}_\varphi} \mathcal{L}(\varphi)$$

Fisseremo $\mathcal{F}_\varphi = \mathbb{P}_M$, ossia i *polinomi*, scegliendo la *base monomiale* $\mathcal{B} = \{x^0, \dots, x^M\}$. Pertanto una funzione generica diventa

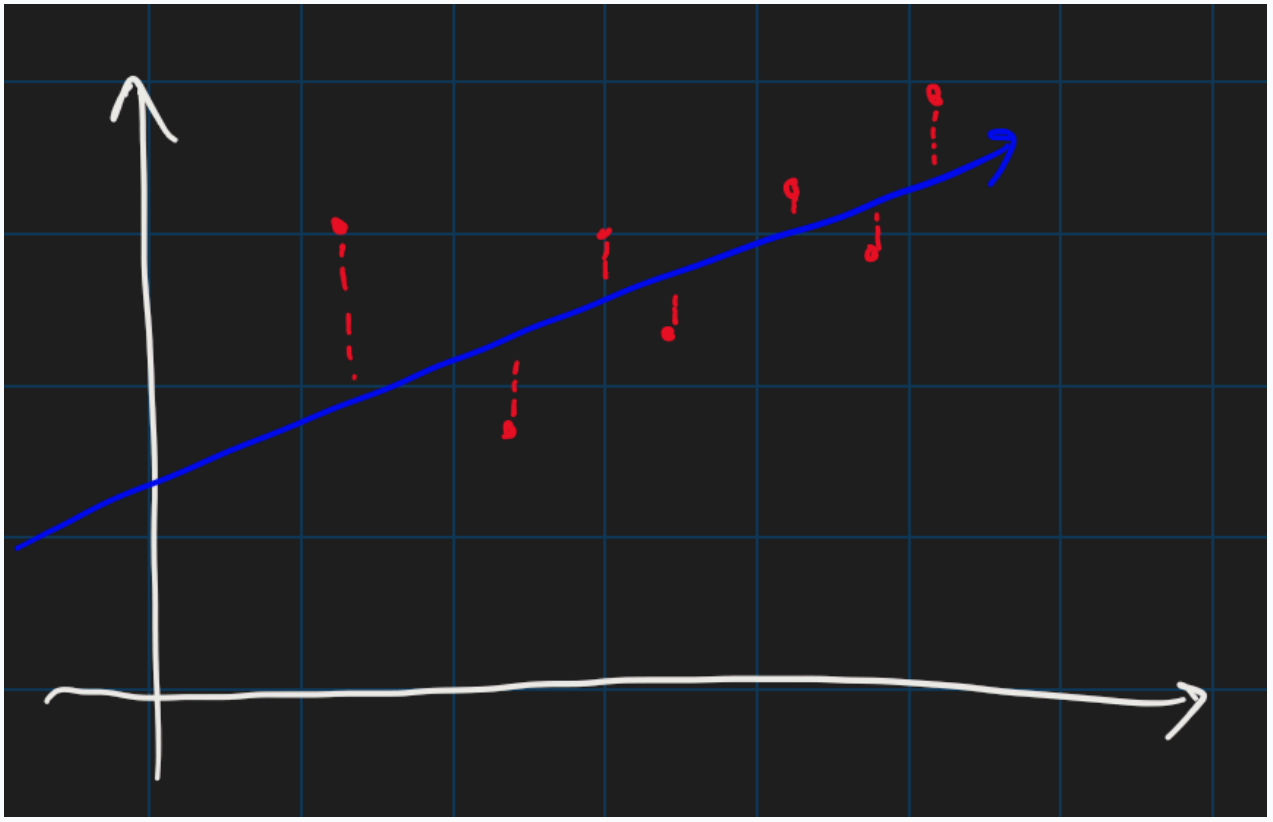
$$p_\alpha(x) = \sum_{0 \le m \le M} \alpha_m x^m$$

Denotiamo $\underline{\alpha} = (\alpha_0, \dots, \alpha_M)^T$. Vogliamo quindi ottimizzare

$$p_M^*(x) = \arg\min_{\underline{\alpha} \in \mathbb{R}^{M+1}} \left\| \underline{y} - p_\alpha(\underline{x}) \right\|_{2,\texttt{discrete}}$$

Dove $\underline{y} = (y_0, \dots, y_N)^T$ e $p_\alpha(\underline{x}) = (p_M(x_0), \dots, p_M(x_N))^T$.

INTUIZIONE GEOMETRICA. Fissato $M = 1$ (in tal caso abbiamo una *regressione lineare OLS*), l'intuizione geometrica consiste in *minimizzare* gli scarti verticali tra la funzione *"imparata"* $p^*(x)$ e i dati misurati. Notiamo che la funzione non deve necessariamente interpolare i nodi, ma deve *"avvicinarci"* il più possibile.



#Proposizione

> Proposizione (proprietà minimizzante della regressione OLS).
>
> Se p_M è la migliore approssimazione OLS dei dati, allora $\forall q_M \in \mathbb{P}_M$, vale la maggiorazione

$$0 \leq \|f - p_M\|^2_{2,\texttt{discrete}} \leq \|f - q_M\|^2_{2,\texttt{discrete}}$$

Q. Come facciamo a trovare i coefficienti ottimali $\underline{\alpha}$?

X

2. Metodo Analitico della Risoluzione

Fissiamo $M = 1$, ossia vogliamo trovare una funzione lineare del tipo

$$p(x) = \beta_0 + \beta_1 x$$

Scriviamo innanzitutto il *problema di minimo*, esprimento la loss $\mathcal{L}$ (forma funzionale) in termini di β_0, β_1:

$$\begin{aligned}\mathcal{L}(\beta_0, \beta_1) &= \sum_{n \leq N} (y_n - p(x_n))^2 \\ &= \sum_{n \leq N} (y_n - (\beta_0 + \beta_1 x))^2 \\ &= \sum_{n \leq N} y_n^2 - 2y_n(\beta_0 + \beta_1 x) + (\beta_0 + \beta_1 x)^2\end{aligned}$$

Prendiamo il *gradiente* di $\mathcal{L}$ rispetto a $\underline{\beta} = (\beta_0, \beta_1)^T$:

$$\nabla_{\underline{\beta}} \mathcal{L} = \begin{pmatrix} \partial_{\beta_0} \mathcal{L} \\ \partial_{\beta_1} \mathcal{L} \end{pmatrix}$$

$\partial_{\beta_0} \mathcal{L}$: Ammazziamo tutti i termini che *non* contengano β_0, e in particolare usiamo la *chain rule* per derivare $(\beta_0 + \beta_1 x)$. Ovviamente sfruttiamo la linearità della derivazione.

$$\partial_{\beta_0} \mathcal{L}(\beta_0, \beta_1) = \sum_{n \leq N} (-2y_n + 2(\beta_0 + \beta_1 x_n))$$

$\partial_{\beta_1} \mathcal{L}$: Analogamente abbiamo

$$\partial_{\beta_1} \mathcal{L}(\beta_0, \beta_1) = \sum_{n \leq N} (-2y_n x_n + 2x_n(\beta_0 + \beta_1 x_n))$$

Riordinando i termini abbiamo

$$\nabla_{\underline{\beta}} \mathcal{L} = \begin{pmatrix} \partial_{\beta_0} \mathcal{L} \\ \partial_{\beta_1} \mathcal{L} \end{pmatrix} = \begin{pmatrix} \sum_{n \leq N} [(2(\beta_0 + \beta_1 x_n) - 2y_n] \\ \sum_{n \leq N} [2x_n(\beta_0 + \beta_1 x_n) - 2y_n x_n] \end{pmatrix}$$

Per trovare il minimo *annulliamo il gradiente* (usiamo il test del gradiente, Teorema 1), ossia $\nabla_{\underline{\beta}} \mathcal{L} = \underline{0}$. Da questo ricaviamo il seguente sistema di equazioni:

$$\nabla_{\underline{\beta}} \mathcal{L} = 0 \iff \begin{cases} \sum_{n \leq N} [(2(\beta_0 + \beta_1 x_n) - 2y_n] = 0 \\ \sum_{n \leq N} [2x_n(\beta_0 + \beta_1 x_n) - 2y_n x_n] = 0 \end{cases}$$

Spezzando le sommatorie e ponendo i termini negativi al RHS e sviluppando dei termini, otteniamo

$$\begin{cases} 2\sum_{n\leq N}\beta_0 + 2\sum_{n\leq N}\beta_1 x_n = 2\sum_{n\leq N} y_n \\ 2\sum_{n\leq N}\beta_0 x_n + 2\sum_{n\leq N}\beta_1 x_n^2 = 2\sum_{n\leq N} y_n x_n \end{cases}$$

Notiamo che possiamo *"tirare fuori"* i coefficienti β_0, β_1 da ogni riga e dunque ri-ottenere il sistema lineare

$$\cancel{2}\begin{pmatrix} n & \sum_n x_n \\ \sum_n x_n & \sum_n x_n^2 \end{pmatrix}\begin{pmatrix}\beta_0 \\ \beta_1\end{pmatrix} = \cancel{2}\begin{pmatrix}\sum_n y_n \\ \sum_n y_n x_n\end{pmatrix}$$

Siccome la matrice dei coefficienti è *invertibile* (dimostrazione posticipata), la soluzione è unica e la ottengo invertendo il sistema lineare. ■

Notiamo che questa formula comporta il svantaggio di dover invertire matrici, un calcolo computazionalmente costoso e mal-condizionato.

X

---

3. Metodo Geometrico della Risoluzione

Come sempre, fissiamo $M = 1$ e vogliamo trovare

$$p_1(x) = \beta_0 + \beta_1 x$$

Quale sarebbe la retta che *azzera* la loss $\mathcal{L}$? Certamente il polinomio interpolatore. Pertanto *"facciamo finta"* di porre le condizioni di interpolazione:

$$\forall i \leq N, \beta_0 + \beta_1 x_i = y_i$$

Pertanto abbiamo un *sistema di equazioni* a forma *"rettangolare"*:

$$\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix}\begin{pmatrix}\beta_0 \\ \beta_1\end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

Ossia $V\beta = y$. Questo è un *sistema sovradeterminato*. Come lo risolviamo? Siccome ha forma rettangolare, non possiamo invertire la matrice V, in quanto tale nozione non è ben definita su matrici non quadre...

Non lo risolviamo nella maniera classica, bensì usiamo il fatto che V va a *rappresentare* una trasformazione lineare $V : \mathbb{R}^2 \longrightarrow \mathbb{R}^N$, dove $N \gg 2$. Il sistema ha soluzione sse $y \in \operatorname{im} V$, tuttavia ciò di solito non accade.

Infatti, Per il *teorema delle dimensioni delle applicazioni lineari* (Teorema 1), abbiamo

$$\dim V = \dim \operatorname{im} V + \dim \ker V = \operatorname{rg} V + \dim \ker V$$

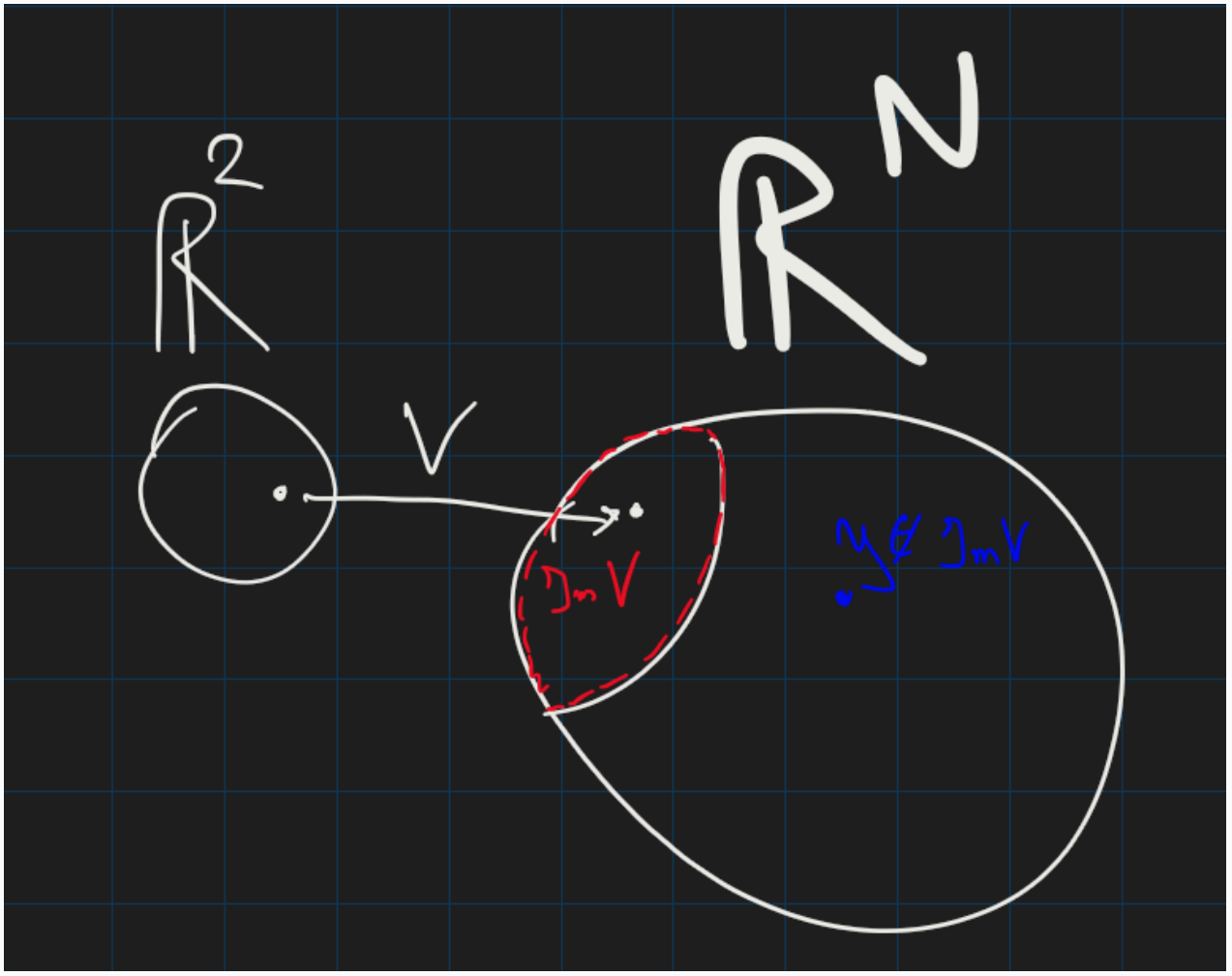
La dimensione di V è 2, inoltre V per assurdo ha rango pieno (infatti supponiamo che $x_i \neq i_j$ per tutti $i \neq j$), da cui

$$2 = 2 + \dim \ker V \implies \dim \ker V = 0$$

Pertanto abbiamo che il nucleo di V è nullo:

$$\ker V = \{\underline{0}\}$$

Da ciò consegue che la dimensione di $\operatorname{im} V$ sarà sempre $M + 1$, che è minore di N. Quindi la funzione V manda in uno *"sottospazio ristretto"* di $\mathbb{R}^N$.



Siccome in generale $y \notin \operatorname{im} V$, vogliamo trovare un *"surrogato"* $\tilde{\beta} \in \operatorname{im} V$ che minimizzi il residuo:

$$\min_{\tilde{\beta} \in \mathbb{R}^{M+1}} \|V\tilde{\beta} - y\|_{2,\ \texttt{discrete}}$$

Senza dimostrare, enunciamo che il *vettore* che minimizzi tale misura è data da

$$V^T V \tilde{\beta} = V^T y$$

Nel nostro caso si ha

$$\beta = (V^T V)^{-1} V^T y$$

Notiamo che i risultati sono consistenti con l'equazione (1), infatti:

$$V^TV = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_N \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix} = \begin{pmatrix} n & \sum_n x_n \\ \sum_n x_n & \sum_n x_n^2 \end{pmatrix}$$

$$V^ty = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_N \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} \sum_n y_n \\ \sum_n x_n y_n \end{pmatrix}$$

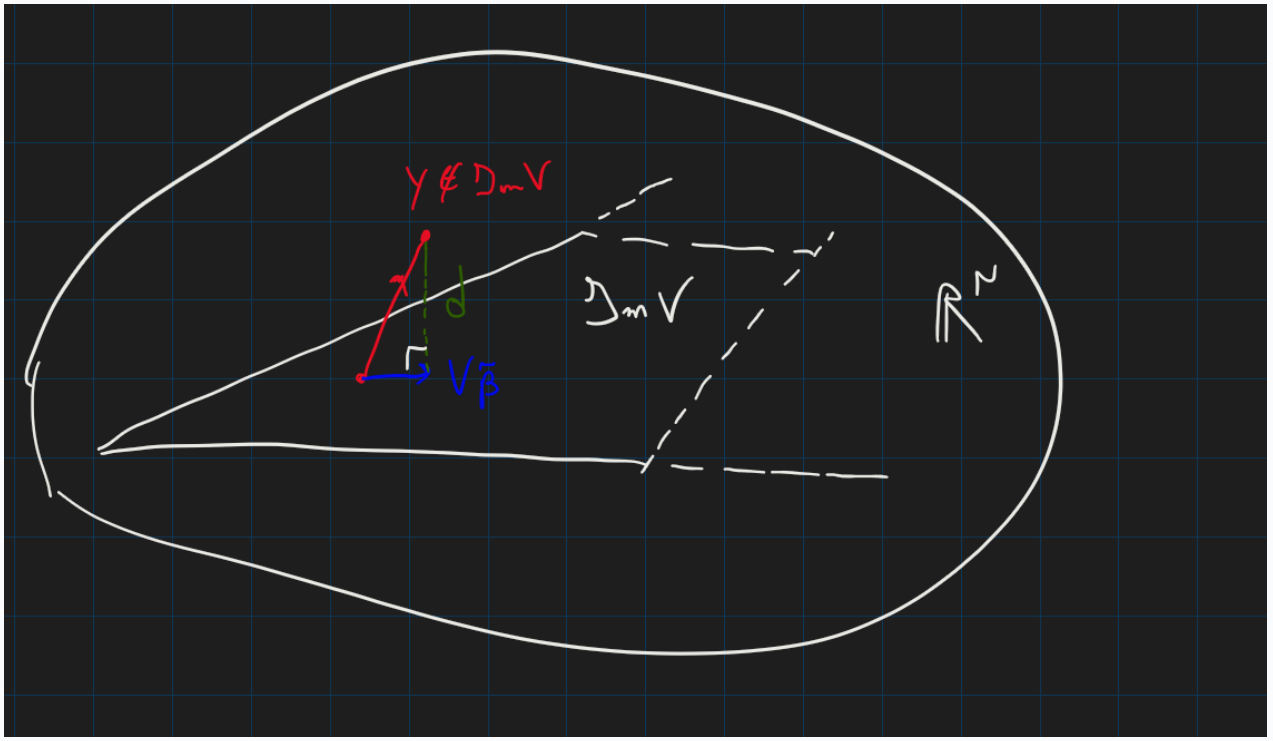
L'intuizione geometrica dell'equazione (2), ovvero il vettore che minimizza la norma è data da

$$V^TV\beta = V^Ty$$

consiste nell'ortogonalità del residuo. Immaginando $\operatorname{im} V$ come un *sottospazio affine* di $\mathbb{R}^N$, e supponendo che $y \notin \operatorname{im} V$, lo immaginamo come un *vettore* che sta *"fuori"* da questo sottospazio.

Adesso prendiamo $V\beta$, che è ovviamente nell'immagine di V, quindi sta *"dentro"* questo sottospazio. La distanza euclidea tra y e $V\beta$ è *minima* se e solo se è la *proiezione ortogonale* su $\operatorname{im} V$. Pertanto il residuo minimizzante appartiene all'ortogonale di $\operatorname{im} V$:

$$d := V\beta - y \in (\operatorname{im} V)^\perp$$



Per definizione questa vale se e solo se tutti i prodotti scalari tra d e le basi dell'immagine si annullano (Definizione 3). Siccome $\operatorname{im} V$ è proprio generata dalle colonne di V (vale in quanto ha rango pieno), abbiamo che deve valere la seguente identità:

$$\begin{aligned}\langle V, (V\beta - y)\rangle &= 0\\ V^T(V\beta - y) &= 0\\ V^TV\beta - V^Ty &= 0\\ V^TV\beta &= V^Ty\end{aligned}$$

concludendo. ■

X

"Algebra Lineare Numerica"

X

Richiami della Teoria Spettrale

Definizione di Autovalore, Autovettore, Autospazio

Nozioni basi della teoria spettrale. Definizione di autovalore, spettro di un'applicazione lineare; definizione di autovetture; definizione di autospazio.

X

---

1. Autovalore di un'applicazione

#Definizione

> Definizione (autovalore di un'applicazione lineare).
>
> Sia $f : V \longrightarrow V$ un'*applicazione lineare* (Definizione 1), con $\dim V = n$.
> Uno *scalare* $\lambda \in K$ si dice *autovalore* (in inglese *"Eigenvalue"* o in tedesco *"der Eigenwert"*) per l'applicazione f se si verifica il seguente
>
> $$\boxed{\exists v \in V \setminus \{0_V\} : f(v) = \lambda \cdot v}$$
>
> A parole, *"un scalare λ è autovalore di f se esiste un vettore di V (escluso il vettore nullo in quanto creerebbe dei problemi) tale che l'immagine di tale vettore è uguale al vettore scalato per il scalare scelto"*.

In termini di *matrici*, abbiamo che data una matrice $A \in \mathbb{K}^{n\times n}$, $\lambda \in \mathbb{K}$ è un *autovalore* sse $\exists v \in \mathbb{R}^n$ tale che

$$Av = \lambda v$$

#Osservazione

> Osservazione (l'esempio della riflessione rispetto alla retta l).

> Riprendiamo l'*esempio 1.2.* relativo alle considerazioni preliminari (Esempio 2): notiamo che $1, -1$ sono *autovalori* di ρ_l.
> Infatti,
>
> $$\rho_l(v_1) = 1 \cdot v_1; \rho_l(v_2) = -1 \cdot v_2; v_1, v_2 \neq 0_V$$

Osserviamo la seguente proprietà degli autovalori ed autovettori:

#Proposizione

> **Proposizione (esponenziazione dell'autovalore è chiusa).**
>
> Sia A una matrice associata ad un'applicazione lineare $f : V \longrightarrow V$. Allora vale la seguente proprietà induttiva:
>
> $$\lambda \in \sigma(A) \implies \lambda^2 \in \sigma(A^2) \implies \dots \implies \lambda^k \in \sigma(A^k) \implies \dots$$
>
> Inoltre - assumendo A invertibile - vale anche per $k = -1$, ossia
>
> $$\lambda \in \sigma(A) \implies \lambda^{-1} \in \sigma(A^{-1})$$

#Dimostrazione

DIMOSTRAZIONE della Proposizione 3
$k = 1$: Non c'è nulla da dimostrare
$k - 1 \implies k$: Come ipotesi induttiva assumiamo

$$\lambda^{k-1} \in \sigma(A^{k-1}) \iff \exists v : A^{k-1}v = \lambda^{k-1}v$$

(ovviamente con $k \geq 1$). Moltiplichiamo per A ottenendo

$$A^k v = \lambda^{k-1} Av = \lambda^{k-1}\lambda v = \lambda^k v$$

Concoludendo la dimostrazione. ■

Spettro di un'applicazione lineare

#Definizione

> **Definizione (spettro di un'applicazione lineare).**
>
> Data $f : V \longrightarrow V$, definiamo l'*insieme dei autovalori di* f come lo *spettro di* f e lo indichiamo con
>
> $$\text{Sp } f$$
>
> Oppure più convenzionalmente con $\sigma(f)$.

2. Autovettore di un'applicazione relativo ad un'autovalore

#Definizione

> Definizione (autovettore di un'applicazione lineare relativo ad un'autovalore).
>
> Sia $f : V \longrightarrow V$ un'*applicazione lineare*; sia $\lambda \in K$ un *autovalore* di f.
> Diciamo che il vettore $v \in V$ è *autovettore* (in inglese *"Eigenvector"* o in tedesco *"der Eigenvektor"*) se vale che
>
> $$f(v) = \lambda \cdot v$$

Notiamo che v è un autovalore di λ relativa ad una applicazione f associata alla matrice A, se e solo se

$$Av = \lambda v$$

Allora possiamo eseguire i seguenti calcoli:

$$\begin{aligned} Av - \lambda v &= 0 \\ Av - \mathbb{1}\lambda v &= 0 \\ v(A - \lambda\mathbb{1}) &= 0 \end{aligned}$$

Essendo $v \neq 0$, abbiamo che lo spazio degli autovettori è determinato dal nucleo di $A - \lambda\mathbb{1}$. Denotiamo tale spazio con

$$E_\lambda = \ker(A - \lambda\mathbb{1})$$

#Osservazione

> Osservazione (condizione necessaria per invertibilità).
>
> Notiamo che se $\lambda = 0$ è un *autovalore* di A (o f), allora ciò vuol dire che $\exists x : Ax = 0$.
> Pertanto $E_0 = \ker A$ *non* è il più piccolo possibile, e pertanto ha dimensione > 0.
> Ossia $\dim\ker A > 0$; dal teorema delle dimensioni (Teorema 1) abbiamo che
>
> $$n = \operatorname{rg} A + \dim\ker A \wedge \dim\ker A > 0 \implies \operatorname{rg} A < n$$
>
> Pertanto A *non è invertibile*.
>
> Prendendo la contronominale, abbiamo che
>
> $$\exists A^{-1} \implies 0 \neq \sigma(A)$$

3. Autospazio di un'autovalore

#Definizione

Definizione (autospazio di un autovalore).

Sia $\lambda \in K$ un *autovalore* per f.
Definiamo *l'autospazio* (in inglese *"Eigenspace"* o in tedesco *"der Eigenraum"*) di λ come *l'insieme di autovettori* di λ e lo denotiamo con

$$\operatorname{Aut} \lambda$$

#Osservazione

Osservazione (l'elemento nullo è elemento di qualsiasi autospazio).

Affinché lo scalare λ sia *autovalore*, per definizione, deve valore che

$$v \in V \setminus \{0_V\} : f(v) = \lambda \cdot v$$

Allora se λ è *autovalore*, consideriamo l'autovettore $w \in V$ relativa a λ:

$$f(w) = \lambda \cdot w$$

In particolare se $w = 0_V$, varrebbe che

$$f(0_V) = \lambda \cdot 0_V = 0_V$$

Dunque vale che il *vettore nullo* 0_V appartiene *sempre* all'autospazio di un qualunque autovalore.

$$\forall \lambda \text{ autovalore di } f, 0_V \in \operatorname{Aut} \lambda$$

- Ortonormalizzazione di Gram-Schmidt

Matrici Ortogonali e Simmetriche

Matrici Ortogonali e Simmetriche

X

Matrici ortogonali e simmetriche. Definizione e proprietà.

X

---

0. Voci correlate

- Matrice

1. Matrici Ortogonali

#Definizione

Definizione (matrice ortogonale).

Una matrice quadrata $Q \in \mathbb{R}^{n\times n}$ si dice *ortogonale* se le sue colonne formano una base ortonormale su $\mathbb{R}^n$ rispetto al prodotto scalare canonicamente indotto, i.e.

$$QQ^T = Q^TQ = \mathbb{1}$$

ossia

$$(Q^TQ)[i,j] = \langle Q[:,i], Q[:,j]\rangle = \delta_{ij}$$

Osserviamo che le matrici ortogonali sono automaticamente invertibili, con $Q^{-1} = Q^T$.

#Proposizione

Proposizione (proprietà delle matrici ortogonali).

Sia $Q \in \mathbb{R}^{n\times n}$ una matrice ortogonale. Allora si gode delle seguenti proprietà:
Conservazione del prodotto scalare: Si ha che con la trasformazione Q, il prodotto scalare tra due vettori rimane uguale, i.e. $\forall x, y \in \mathbb{R}^n, \langle Qx, Qy\rangle = \langle x, y\rangle$.
Determinante: Il determinante è ± 1
Autovalori: Tutti gli autovalori giaciono sul bordo della circonferenza unitaria nel piano complesso $\mathbb{S}$, ossia $\forall \lambda \in \sigma(Q), |\lambda| = 1$.

#Dimostrazione

DIMOSTRAZIONE della Proposizione 2
Conservazione del prodotto scalare: Trivialmente
$\langle Qx, Qy\rangle = (Qx)^T(Qy) = x^TQ^TQy = x^T\mathbb{1}y = x^Ty$, che è il prodotto scalare $\langle x, y\rangle$.
Determinante: Usiamo il teorema di Binet (Teorema 12) sfruttando l'identità $AA^T = \mathbb{1}$ e il fatto che il determinante rimane uguale sotto la trasposizione (Corollario 11):

$$\det(AA^T) = \det(A)\det(A^T) = (\det(A))^2 \equiv \det(\mathbb{1}) = 1$$

Pertanto ho un'equazione di secondo grado, risolto da $\det A = \pm 1$.
Autovalori: Omessa. ■

X

2. Matrici Simmetriche

#Definizione

Definizione (matrice simmetrica).

Una matrice $S \in \mathbb{R}^{n\times n}$ è *simmetrica* sse vale che $A = A^T$.

Nel caso complesso ho *matrici hermitiane*, ossia

$$H \in \mathbb{C}^{n\times n}, H^T = \overline{H}; H^\dagger := \overline{H}$$

Osserviamo che data una matrice quadrata A, sicuramente AA^T e A^TA sono simmetriche.

#Proposizione

> Proposizione (proprietà delle matrici simmetriche).
>
> Siano A, B due matrici simmetriche. Allora:
> *Chiusura sotto la somma e scalamento*: $A + B, \lambda S$ sono simmetriche
> *Commutazione del prodotto*: $(AB)^T = BA$

Matrici Definite con Segno

Matrici Definite con Segno

X

Nozione preliminare per l'ottimizzazione dei liberi: segno delle matrici.

X

---

0. Voci correlate

- Forme Lineari e Quadratiche
- Matrice
- Determinante

1. Definizione del Segno di una Matrice

Prima di enunciare un *criterio* per *distinguere i punti critici*, definiamo il segno di una matrice (nozioni che useremo poi sulla matrice hessiana).

#Definizione

> Definizione (segno di una matrice).
>
> Sia $A \in M_{n,n}(\mathbb{R})$ una matrice. Sia $Q : \mathbb{R}^n \longrightarrow \mathbb{R}$ la sua *forma quadratica* associata (1), ossia $Q(x) = x^TAx$
>
> Dato un qualsiasi $\underline{h} \neq \underline{0}$, si dice che Q è:
>
> - *Positiva* se $Q(\underline{h}) > 0$

    - *Semipositiva* se $Q(\underline{h}) \geq 0$
- *Negativa* se $Q(\underline{h}) < 0$
    - *Seminegativa* se $Q(\underline{h}) \leq 0$
- *Indefinita* se $\exists \underline{u}, \underline{v} \in \mathbb{R}^N$ tali che

$$Q(\underline{v}) < 0 < Q(\underline{u})$$

Si definisce il *segno della sua matrice* come il *segno della sua forma quadratica* Q.

Notiamo che A è positiva sse $-A$ è negativa.

#Proposizione

Proposizione (matrici simmetriche derivante da matrici arbitrarie).

Sia $A \in \mathbb{R}^{m \times n}$. Allora AA^T e A^TA sono *matrici simmetriche definite semi-positivamente*.

#Dimostrazione

DIMOSTRAZIONE della Proposizione 2
Calcoliamo

$$\langle x, A^TAx \rangle = x^TA^TAx = (Ax)^T(Ax) = \|Ax\|_2^2 \geq 0$$

Concludendo. ■

Osserviamo inoltre che sono *strettamente positive* sse A ha rango pieno. Infatti $\|Ax\|_2^2$ è 0 se e solo se $\det A = 0$ (escludiamo il caso $x = 0$ per definizione).

X

---

2. Caraterizzazione del Segno di una Matrice

Vediamo delle *condizioni equivalenti* per classificare la *positività* e la *negatività* della matrice. Vale a dire, delle *caratterizzazioni* delle matrici definite positivamente (WLOG).

2.1. Caratterizzazione con costante m (???)

#Proposizione

Proposizione (condizioni equivalenti per la positività e la negatività del segno).

Sia Q una *forma quadratica*. Si ha che

$$Q \text{ positiva} \iff Q(\underline{h}) \geq m\|\underline{h}\|^2, \forall \underline{h} \in \mathbb{R}^N$$

e

$$Q \text{ negativa} \iff Q(\underline{h}) \leq m\|\underline{h}\|^2, \forall \underline{h} \in \mathbb{R}^N$$

#Dimostrazione

DIMOSTRAZIONE della Proposizione 3.

Omessa. ■

2.2. Caratterizzazione in termini di teoria spettrale

#Teorema

> Teorema (caratterizzazione con la teoria spettral).
>
> Sia $A \in \mathbb{R}^{n\times n}$ una matrice simmetrica. Allora:
> A è definita *positivamente* sse tutti gli autovalori sono > 0 (sono in $\mathbb{R}$ per il teorema spettrale);
> A è definita *semipositivamente* sse tutti gli autovalori sono ≥ 0
>
> Il teorema vale simmetricamente per matrici definite *negativamente*.

Da questo abbiamo una proprietà peculiare:

#Proposizione

> Proposizione (matrici definite con segno sono invertibili e hanno segno positivo).
>
> Le matrici definite *positivamente* sono invertibili e hanno inversa *definita positivamente*

#Dimostrazione

DIMOSTRAZIONE della Proposizione 5

Discende dalla proprietà degli autovalori, per cui

$$\frac{1}{\lambda} \in \sigma(A^{-1}) \impliedby \lambda \in \sigma(A)$$

Concludendo. ■

Notiamo che se A è definita *positivamente* o *negativamente* nel senso stretto, allora essa è sicuramente invertibile. Infatto si ha per il Teorema 4 che $\sigma(A) \subset \mathbb{C} \setminus \{0 + 0i\}$. Dunque $\ker A$ è la *più piccola possibile* e quindi $\det A \neq 0$ e pertanto è invertibile.

2.3. Criterio di Sylvester

Vediamo il teorema più utile per poter classificare il segno della matrice.

#Teorema

Teorema (criterio di Sylvester).

Sia $Q : \mathbb{R}^N \longrightarrow \mathbb{R}$ una *forma quadratica* con $Q(\underline{h}) = \langle A \cdot \underline{h}, \underline{h} \rangle$. Sia A una matrice simmetrica ($A = {}^tA$).

Allora si ha che:

$$Q > 0 \iff \begin{cases} \det A_1 = a_{11} > 0 \\ \det A_2 = \det \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix} > 0 \\ \vdots \\ \det A_N = \det A > 0 \end{cases}$$

Ovvero prendendo *tutte le determinanti di ogni sottomatrice di* A ho solo numeri positivi

Inoltre ho che

$$Q < 0 \iff \begin{cases} \det A_1 = a_{11} > 0 \\ \det A_2 = \det \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix} < 0 \\ \vdots \\ (-1)^N \det A_N = (-1)^N \det A > 0 \end{cases}$$

Ovvero prendendo *tutte le determinante di ogni sottomatrice di* A ho un'oscillazione tra il negativo-positivo.

Se non vale nessuna delle condizioni equivalenti, si dice che il segno della Q è *indefinita*.

In particolare, definiamo $A_1, \dots, A_n$ i *minori principali in testa* di A

#Esempio

Esempio (caso $N = 2$).

Abbiamo che

$$\begin{aligned} Q > 0 &\iff a_{11} > 0 \wedge \det A > 0 \\ Q < 0 &\iff a_{11} < 0 \wedge \det A > 0 \\ Q \gtrless 0 \text{ (indeterminata)} &\iff a_{11} \in \mathbb{R} \wedge \det A < 0 \end{aligned}$$

Teorema di Gershgorin

Teorema di Gershgorin

X

Teorema di Gershgorin.

X

0. Voci correlate

- Definizione di Autovalore, Autovettore, Autospazio

1. Teorema di Gershgorin

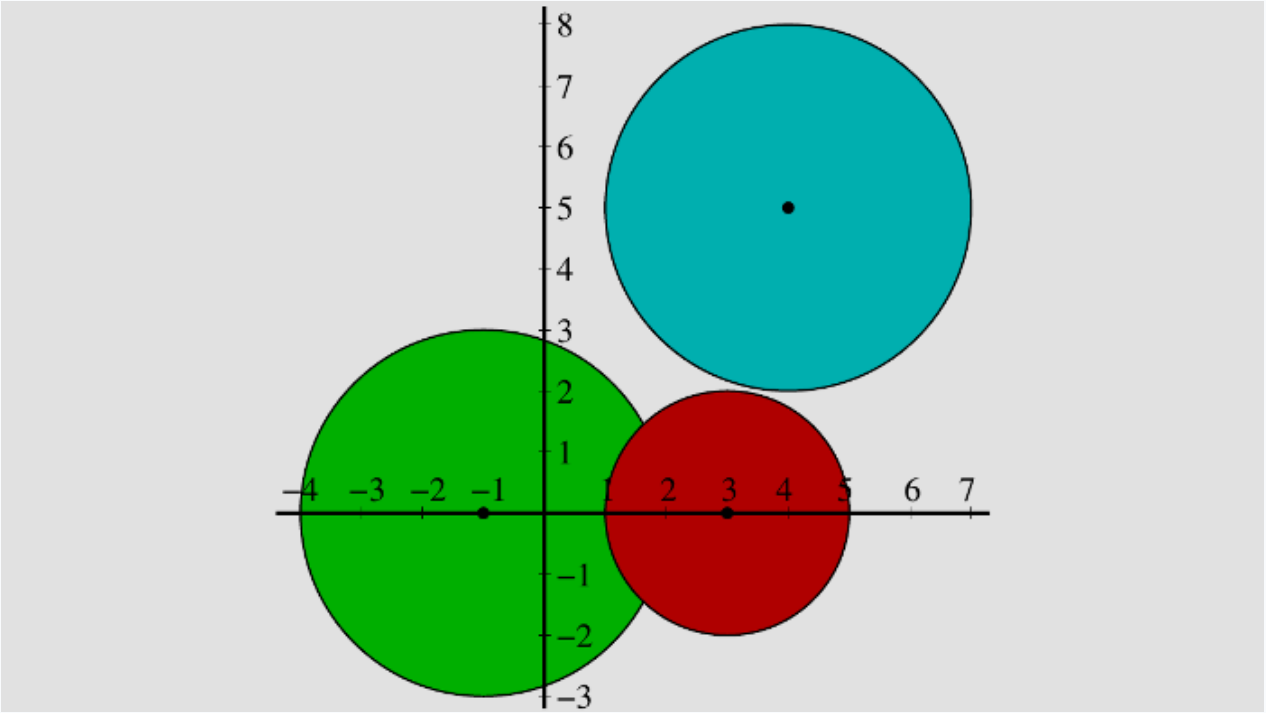
#Definizione

Definizione (disco di Gershgorin).

Sia $A \in \mathbb{C}^{n\times n}$. Per $i = 1, \dots, n$ definiamo il *disco i-esimo* R_i come

$$R_i := \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j\neq i} |a_{ij}| \right\}$$

Ossia il disco con raggio calcolato dalla somma della riga i-esima, con diagonale esclusa.



#Teorema

Teorema (teorema di Gershgorin).

Si ha che, per ogni matrice A, che il suo spettro è incluso nell'unione dei dischi di Gershgorin:

$$\sigma(A) \subseteq \bigcup_{i=1,\dots,n} R_i$$

#Definizione

Definizione (dischi colonna di Gershgorin).

Sia $A \in \mathbb{C}^{n\times n}$. Per $i = 1, \dots, n$ definiamo il *disco colonna* i*-esimo* C_i come

$$C_i := \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j\neq i} |a_{ji}| \right\}$$

Ossia il disco con raggio calcolato dalla somma della riga i-esima, con diagonale esclusa.

#Teorema

Teorema (Gershgorin 2).

Si ha che, per ogni matrice A, che il suo spettro è incluso nell'unione dei dischi-colonna di Gershgorin:

$$\sigma(A) \subseteq \bigcup_{i=1,\dots,n} C_i$$

Allora come corollario abbiamo che $\sigma(A)$ sta nell'intersezione dei dischi riga e dischi colonna di Gershgorin.

$$\sigma(A) \subseteq \left(\bigcup_n R_n\right) \cap \left(\bigcup_n C_n\right)$$

Da questo teorema possiamo dedurre la posizione degli autovalori, e in particolare se $0 \in \sigma(A)$ o meno (cruciale per capire se A *non* è invertibile!)

Un *secondo* teorema di Gershgorin enuncierebbe che *"se ho dei cerchi ben-separati allora sicuramente un autovalore sta proprio in uno dei cerchi separati"*.

Norme su Matrici

Norme Matriciali

X

Norme su matrici. Definizione di norma matriciale. Norma di Frobenius. Norme matriciali indotte da norme vettoriali. Regole di calcolo per norme matriciali indotte in $p = 1, +\infty$. Raggio spettrale, regola di calcolo per norma matriciale indotta in $p = 2$. Proprietà delle norme indotte: compatibilità, norma 1 su matrici unitarie.

X

0. Voci correlate

- Matrice
- Spazi Vettoriali Normati

1. Norma Matriciale

OSSERVAZIONE. $\mathbb{R}^{m\times n}$ forma un $\mathbb{R}$-spazio vettoriale. Infatti possiamo definire operazioni di *somma interna* e *scalamento esterno* per cui valgono gli assiomi vettoriali. Inoltre, sappiamo che sui spazi vettoriali possiamo definire una *norma*, ovvero una funzione $\|\bullet\| : V \longrightarrow \mathbb{R}^+$.

Q. Possiamo fare la stessa cosa su $\mathbb{R}^{m\times n}$?

Sì, diamo la seguente definizione ancora più *"ricca"*, che va anche comprendere l'operazione di moltiplicazione tra matrici.

#Definizione

> Definizione (norma matriciale).
>
> Definiamo la *norma matriciale* un'applicazione
>
> $$\|\bullet\| : \bigcup_{m,n\in\mathbb{N}^*} \mathbb{R}^{m\times n} \longrightarrow \mathbb{R}$$
>
> Tale che le seguenti proprietà vengono soddisfatte:
> *Non-degeneratezza*: $\|A\| \geq 0$ e $\|A\| = 0 \iff A = 0$ (matrice con tutti zeri)
> *Omogeneità*: $\|\alpha A\| = |\alpha| \cdot \|A\|$
> *Subadditività*: $\|A + B\| \leq \|A\| + \|B\|$
> *Submoltiplicatività*: $\|AB\| \leq \|A\| \cdot \|B\|$

La *"novità"* consiste nella submoltiplicatività. Vediamo un primo esempio di norma matriciale

X

2. Norma di Frobenius

#Definizione

> Definizione (norma di Frobenius).
>
> Sia $A \in \mathbb{R}^{m \times n}$. Si definisce la sua *norma di Frobenius* come la radice quadrata della somma quadratica dei suoi elementi:
>
> $$\|A\|_F = \left(\sum_{m,n} (A[m,n])^2 \right)^{1/2}$$

Senza dimostrare, enunciamo che la *norma di Frobenius* è una *norma matriciale*. Infatti, in un certo senso è *"indotta"* (MA NON LO E'!!!) dalla norma vettoriale $\mathbb{R}^\bullet$ (facendo, diciamo, un *"reshape"* della matrice). Tuttavia la norma di Frobenius è più *"povera"* rispetto alle norme che vedremo.

X

---

3. Norme Indotte da Norme Vettoriali

#Definizione

> Definizione (norma matriciale indotta da norma vettoriale).
>
> Fissate norme vettoriali su $\mathbb{R}^n$ e su $\mathbb{R}^m$ (li denotiamo con $\|\bullet\|_{\mathbb{R}^\bullet}$). Data matrice $A \in \mathbb{R}^{m \times n}$, definiamo la sua *norma matriciale indotta dalle norme vettoriali* come
>
> $$\|A\| = \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_{\mathbb{R}^m}}{\|x\|_{\mathbb{R}^n}}$$

In un certo senso, *"misuriamo"* la massima trasformazione relativa di A. Tuttavia, verifichiamo prima che la definizione sia ben posta:

#Lemma

> Lemma (lemma sulle norme matriciali indotte).
>
> Si ha che
>
> $$\sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_{\mathbb{R}^m}}{\|x\|_{\mathbb{R}^n}} < +\infty$$
>
> Ossia, la norma matriciale indotta è ben definita.

#Dimostrazione

DIMOSTRAZIONE del Lemma 4

Basta normalizzare x, i.e. effettuando il cambio di variabile $y = x/\|x\|_{\mathbb{R}^n}$, da cui restringiamo da $\mathbb{R}^n$ a $\mathbb{S}_1$ (la circonferenza in $\mathbb{R}^n$ di raggio 1)

$$\begin{aligned}\sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Ax\|_{\mathbb{R}^m}}{\|x\|_{\mathbb{R}^n}} &= \sup_{y\in\mathbb{S}_1} \frac{\|(A\|x\|_{\mathbb{R}^n}y)\|_{\mathbb{R}^m}}{\|x\|_{\mathbb{R}^n}} \\ &= \sup_{y\in\mathbb{S}_1} \frac{\|x\|_{\mathbb{R}^n}\|Ay\|_{\mathbb{R}^m}}{\|x\|_{\mathbb{R}^n}} \\ &= \sup_{y\in\mathbb{S}_1} \|Ay\|_{\mathbb{R}^m}\end{aligned}$$

Siccome $\mathbb{S}_1$ è un insieme chiuso e limitato (ossia compatto), sicuramente esiste l'estremo superiore ed è il massimo, concludendo. ■

#Definizione

> **Definizione (norma canonicamente indotta p).**
>
> Sia $p \in [1, +\infty]$. Definiamo la *p-norma matriciale canonicamente indotta* come la *norma canonicamente indotta* con norme p su vettori:
>
> $$\|A\|_p := \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Ax\|_p}{\|x\|_p}$$

Vediamo un paio di proprietà sulle norme matriciali

#Definizione

> **Definizione (compatibilità di norme matriciali con norme vettoriali).**
>
> Una *norma matriciale* si dice *compatibile con norma vettoriale* se vale che $\forall A \in \mathbb{R}^{m\times n}, x \in \mathbb{R}^m$, vale che
>
> $$\|Ax\|_{\mathbb{R}^m} \leq \|A\| \cdot \|x\|_{\mathbb{R}^n}$$

#Proposizione

> **Proposizione (compatibilità delle norme indotte).**
>
> Le *norme matriciali indotte* sono compatibili con le *norme vettoriali che lo inducono*.

#Dimostrazione

DIMOSTRAZIONE della Proposizione 7

Dividiamo la dimostrazione due casi: $x = 0$ e $x \neq 0$. Fissiamo A genericamente.

$x = 0$: Banalmente $\|Ax\| = 0$ e $\|x\| = 0$, per cui è sempre ovvia la tesi.
$x \neq 0$: Per definizione si ha che

$$\frac{\|Ax\|}{\|x\|} \leq \|A\|$$

Moltiplicando per $\|x\|$ otteniamo la tesi. ■

#Proposizione

> Proposizione (le norme indotte sono norme).
>
> La *norma matriciale indotta* è una norma

La dimostrazione è lasciata da fare per esercizio. Per la *submoltiplicatività* si suggerisce di usare la *compatibilità* (Proposizione 7).

#Proposizione

> Proposizione (le norme indotte sono 1 con matrici identità).
>
> Per ogni norma indotta si ha che $\|\mathbb{1}\| = 1$.

#Dimostrazione

DIMOSTRAZIONE della Proposizione 9
Notiamo che $\forall x, \mathbb{1}x = x$ (infatti è l'elemento neutro della moltiplicazione). Dalla definizione di norma indotta si ha la tesi. ■

#Osservazione

> Osservazione (la norma di frobenius non è indotta).
>
> Osserviamo che la *norma di Frobenius* non è una norma indotta. Infatti $\|\mathbb{1}_n\|_F = \sqrt{n}$.
>
> Tuttavia, la *norma di Frobenius* rimane comunque *compatibile* con norma vettoriale $p = 2$. ■

#Esercizio

> Esercizio (esercizi).
>
> Dimostrare che, data una norma matriciale indotta $\|\bullet\|$, che per ogni matrice $A \in \mathbb{R}^{n \times n}$ vale che:
> *Maggiorazione del raggio spettrale*:
>
> $$\|A\| \geq \rho(A)$$

Maggiorazione della norma della reciproca:

$$\|A\| \geq \frac{1}{\|A\|}$$

(*hint*: usare il fatto che $\mathbb{1} = AA^{-1}$ e sfruttare la submoltiplicatività)

X

---

4. Regole Pratiche delle Norme p

#Proposizione

Proposizione (regola pratica del calcolo norma $p = 1$).

Sia $A \in \mathbb{R}^{m \times n}$. Allora la norma $p = 1$ è calcolabile come il massimo della somme assolute delle sue colonne:

$$\|A\|_1 = \max_{i=1,\dots,n} \sum_{j \leq m} |A[i,j]| = \max_{i=1,\dots,n} \|A[:,i]\|_1$$

#Proposizione

Proposizione (regola pratica del calcolo norma $p = +\infty$).

Sia $A \in \mathbb{R}^{m \times n}$. Allora la norma $p = +\infty$ è calcolabile come il massimo della somme assolute delle sue righe:

$$\|A\|_{+\infty} = \max_{j=1,\dots,m} \sum_{i \leq m} |A[i,j]| = \max_{j=1,\dots,m} \|A[j,:]\|_1$$

Il calcolo della norma $p = 2$ sarà più difficile, ma comunque fondamentale per la *SVD*.

#Definizione

Definizione (raggio spettrale di matrici).

Sia $A \in \mathbb{R}^{n \times n}$. Si dice *raggio spettrale di* A, $\rho(A)$, come il *modulo massimo dei suoi autovalori*:

$$\rho(A) := \max_{\lambda_i \in \sigma(A)} |\lambda_i|$$

#Proposizione

Proposizione (regola pratica del calcolo della norma $p = 2$).

Sia $A \in \mathbb{R}^{m \times n}$. Allora la norma $p = 2$ è calcolabile come la radice quadrata del raggio spettrale di AA^T o A^TA:

$$\|A\|_2 = \sqrt{\rho(A^TA)} = \sqrt{\rho(AA^T)}$$

MEG

Algoritmo di Gauß

Definizioni preliminari per la descrizione dell'algoritmo di Gauß (Matrice completa e le operazioni elementari OE). Descrizione dell'algoritmo di Gauß per rendere un sistema lineare in un sistema lineare equivalente a scala come un programma.

X

---

1. Matrice completa di un sistema lineare

#Definizione

Definizione (matrice completa di un sistema lineare).

Consideriamo un sistema lineare di forma

$$A \cdot x = b$$

allora definiamo la *matrice* ottenuta aggiungendo alla matrice A la colonna data dai *termini noti* b come la *matrice completa* di questo sistema lineare. La denotiamo con

$$(A|b) := \left(\begin{array}{cccc|c} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ \vdots & & & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} & b_m \end{array}\right)$$

N. B. Il segno sbarra | per *"differenziare"* i termini noti dai coefficienti ha uno scopo puramente grafico.

2. Operazioni elementari OE

Ora definiamo una serie di *operazioni elementari* (OE) che sono in grado di trasformare un *sistema lineare* di forma $(A|B)$ in un altro *equivalente* (Definizione 9).

#Definizione

Definizione (le operazioni elementari).

OE1. *L'operazione scambia equazioni*
Dati due indici $i, j \in \{1, \dots, m\}$ scambiamo di posto l'equazione i-esima e j-esima. Questo corrisponde a *scambiare* la riga i-esima con la riga j-esima della matrice $(A|B)$.

OE2. *L'operazione scala equazioni*
Dato l'indice $i \in \{1, \dots, m\}$ e uno *scalare* $\lambda \in K$, moltiplichiamo l'i-esima equazione per λ. Precisamente questo corrisponde a *moltiplicare* per λ l'i-esima riga della matrice completa $(A|B)$.

OE3. *L'operazione somma equazioni*
Dati due indici $i, j \in \{1, \dots, m\}$ e uno scalare non nullo $\lambda \in K, \lambda \neq 0$, sommiamo alla i-esima equazione alla i-esima equazione la j-esima equazione dopo averla moltiplicata per λ.
Ovvero questo corrisponde a sommare alla riga i-esima della matrice completa $(A|B)$ λ volte la j-esima riga.

#Osservazione

Osservazione (Osservazione 2.1.).

Osserviamo che queste operazioni determinano dei sistemi lineari *equivalenti* in quanto queste operazioni sono *completamente invertibili*; infatti partendo da un sistema lineare *"trasformato"* mediante le **OE.**, possiamo tornare al sistema originario.

#Proposizione

Proposizione (le OE trasformano sistemi in sistemi equivalenti).

Se applico ad un sistema lineare qualsiasi una di queste operazione elementari, allora ottengo un sistema equivalente.

#Proposizione

Proposizione (con le OE posso portare un sistema a scala).

Dato un *qualsiasi sistema lineare arbitrario*, posso portarlo ad un *sistema a scala* con queste operazioni elementari **OE**. Infatti mostreremo un *algoritmo* (Nozioni Fondamentali di Programmazione) che è in grado di *"gradinizzare"* (ovvero portare a scala) una matrice completa $(A|B)$ qualsiasi.

3. L'algoritmo di Gauß

Premesse storiche

Riprendendo la *proposizione 2.2.* della sezione precedente, abbiamo appena enunciato che siamo in grado di portare un sistema lineare non a scala in un sistema lineare *a scala*; dimostreremo questa proposizione descrivendo uno degli algoritmi più noti dell'*Algebra Lineare*, ovvero *l'algoritmo di Gauß*.

NOTIZIE STORICHE. (*Trascrizione appunti + approfondimenti personali*)
Questo algoritmo è stato attribuito al noto matematico C. F. Gauß *(1777-1855)* in quanto fu proprio lui a formalizzare questo procedimento in latino; tuttavia ciò non significa che il matematico Gauß inventò questo algoritmo, in quanto ci sono evidenze storiche che prima esistevano già descrizioni su questo procedimento. Infatti, esiste un antico manoscritto cinese (*I Capitoli nove arte matematica*/*九章算術*, circa 179) che descrive un principio simile a quello che andremo a descrivere.
Per ulteriori approfondimenti consultare le seguenti pagine:
https://mathshistory.st-andrews.ac.uk/HistTopics/Matrices_and_determinants/
https://it.frwiki.wiki/wiki/Les_Neuf_Chapitres_sur_l%27art_math%C3%A9matique

---

Descrizione dell'algoritmo come programma

OBIETTIVO.
Come detto prima, il nostro *obiettivo* è quello di *"gradinizzare"* un sistema lineare qualsiasi che non sia a scala.

INPUT.
Quindi il nostro input è un sistema lineare qualsiasi del tipo

$$Ax = b$$

che lo *"condenseremo"* nella *matrice completa* $(A|B)$.

OUTPUT.
Vogliamo ottenere la matrice completa $(\tilde{A}\,|\tilde{B}\,)$ tale che

$$\tilde{A} \text{ è a scala e } \tilde{A}x = \tilde{B} \overset{\text{equiv.}}{\cong} Ax = b$$

ALGORITMO.
Il nostro procedimento si articola in una serie di *"istruzioni"* da eseguire per un certo numero di volte.

1. Determino il valore $\bar{j}$ come *l'indice di colonna minimo* per cui abbiamo una colonna *non nulla* di A. Ovvero

$$\bar{j} := \min\{j : A^j \neq 0\}$$

2. Determino l'indice $\bar{i}$ tale per cui abbiamo l'elemento $a_{\bar{i},\bar{j}} \neq 0$ (*l'esistenza di un tale* $\bar{i}$ *deriva dalla scelta di* $\bar{j}$)
3. Scambio le righe 1 con la $\bar{i}$-esima; in questo modo sarà possibile supporre che $a_{1\bar{j}} \neq 0$ (*OE1*)
4. Voglio assicurarmi che *non* ho altre colonne *nulle* in $A^{(\bar{j})}$ (eccetto ovviamente $A_{(1)}$).
    1. Moltiplico la riga $A_{(1)}$ per $\frac{1}{a_{1\bar{j}}}$ (*OE2*)
    2. Sommo alle altre righe $A_i, \forall i \in \{2, \dots, m\}$ un *multiplo opportuno* di $A_{(1)}$. Ovvero $\lambda = -a_{ij}$. (*OE3*)

$$A_{(i)} = A_{(i)} - a_{ij}A_{(1)}$$

5. Se la matrice ottenuta non è a scala, ripeto lo stesso procedimento a partire da *1.* sulla *sottomatrice* (ovvero una *"parte selezionata"* della matrice) con righe $\{2, \dots, m\}$ e colonne $\{\bar{j}+1, \dots, n\}$, del tipo

$$A' \in M_{m-1,n-\bar{j}-1}(K)$$

Queste operazioni corrispondono a:

$$0.\ \begin{pmatrix} 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \end{pmatrix}$$

$$1.\ \begin{pmatrix} 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \end{pmatrix}; \bar{j} = 3$$

$$2.\ \begin{pmatrix} 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \end{pmatrix}; \bar{i} = 1, 2, 3 \text{ (una di queste)}$$

$$3.\ \begin{pmatrix} 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \end{pmatrix} A_{(1)} \leftrightarrows A_{(1),(2),(3)} \text{ (una di queste)}$$

$$4.1.\ \begin{pmatrix} 0 & 0 & 1 & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \end{pmatrix}$$

$$4.2.\ \begin{pmatrix} 0 & 0 & 1 & * & * \\ 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & * & * \end{pmatrix} A_{(i)} = A_{(i)} - a_{ij}A_{(1)} \text{ per } i = 2, 3$$

$$5.\ \begin{pmatrix} 0 & 0 & 1 & * & * \\ 0 & 0 & 0 & + & + \\ 0 & 0 & 0 & + & + \end{pmatrix} \implies \begin{pmatrix} + & + \\ + & + \end{pmatrix} \text{ ripeto}$$

#Osservazione

> Osservazione (l'algoritmo è valido e ben posto?).
>
> Affinché questo algoritmo sia *valido* e *ben posto*, devo assicurarmi che:
>
> 1. Questo deve *eventualmente* terminare in un certo tempo *finito*; questo accade in quanto *prima o poi* le colonne e le righe delle *sottomatrici* della *5.* eventualmente si *"esauriranno"* e avremmo una matrice a scala.
> 2. Questo restituisce l'*output* corretto, come prescritto dalle specificazione. Anche questo si verifica in quanto ogni volta che raggiungo e svolgo il step *4.*, ho *"gradinizzato"* una scala.

COMPLESSITA'. Effettuiamo $n-1$ passaggi, dove ad ogni passaggio effettuiamo $n-i$ sottrazioni. Quindi la complessità dell'algoritmo è all'incirca $O(n^2)$. Tuttavia, questo algoritmo *presume* delle condizioni specifiche (vedremo bene con *pivoting*).

Esempio di applicazione.

Come un *programmatore* fa dei *"unit tests"* su un programma o algoritmo, tentiamo di applicare questo principio appena descritto ad un sistema lineare.

#Esempio

> Esempio (Esempio 3.1.).
>
> Consideriamo il sistema lineare dato da
>
> $$(A|B) = \begin{pmatrix} 0 & -1 & 2 & 1 & 3 \\ 2 & 4 & 8 & 6 & 2 \\ 3 & 1 & 5 & 3 & 1 \end{pmatrix}$$

Ora ci applichiamo *l'algoritmo di Gauß*.

0. $\begin{pmatrix} 0 & -1 & 2 & 1 & 3 \\ 2 & 4 & 8 & 6 & 2 \\ 3 & 1 & 5 & 3 & 1 \end{pmatrix}; j = 0, i = 2$
1. $\begin{pmatrix} 2 & 4 & 8 & 6 & 2 \\ 0 & -1 & 2 & 1 & 3 \\ 3 & 1 & 5 & 3 & 1 \end{pmatrix}; A_{(1)} \leftrightarrows A_{(2)}$
2. $\begin{pmatrix} 1 & 2 & 4 & 3 & 1 \\ 0 & -1 & 2 & 1 & 3 \\ 3 & 1 & 5 & 3 & 1 \end{pmatrix}; A_{(1)} = 0.5A_{(1)}$
3. $\begin{pmatrix} 1 & 2 & 4 & 3 & 1 \\ 0 & -1 & 2 & 1 & 3 \\ 0 & -5 & -7 & -6 & -2 \end{pmatrix}; A_{(2)} = A_{(2)} + 0A_{(1)}; A_{(3)} = A_{(3)} - 3A_{(1)}$
4. ripeto con $\begin{pmatrix} -1 & 2 & 1 & 3 \\ -5 & -7 & -6 & -2 \end{pmatrix}$
5. $\begin{pmatrix} 1 & -2 & -1 & -3 \\ -5 & -7 & -6 & -2 \end{pmatrix}; A_{(1)} = -A_{(1)}$
6. $\begin{pmatrix} 1 & -2 & -1 & -3 \\ 0 & -17 & -11 & -17 \end{pmatrix}; A_{(2)} = A_{(2)} - 5A_{(1)}$
7. la matrice in 6. è a scala; FINE

Dunque otteniamo la seguente matrice:

$$(\overline{A}\,|\overline{b}) = \begin{pmatrix} 1 & 2 & 4 & 3 & 1 \\ 0 & 1 & -2 & -1 & -3 \\ 0 & 0 & -17 & -11 & -17 \end{pmatrix}$$

che è *a scala*.

ESERCIZIO PERSONALE. Questo esercizio prevede un collegamento con *l'informatica*, in particolare con la *programmazione*.
A) Scrivere uno *pseudocodice* che *"emula"* questo principio
B) Implementare tale *pseudocodice* in *C/Python*
C) Calcolare la *"complessità"* di questo codice

Decomposizione LU

Decomposizione LU

X

Decomposizione LU delle matrici. Idea preliminare: usare informazioni ricavate dall'eliminazione di Gauss. Motivazioni. Teorema di esistenza e unicità della decomposizione LU delle matrici, dimostrazione. Calcolo del determinante mediante la decomposizione LU.

0. Voci correlate

X

- Metodo di Eliminazione di Gauss
- Algoritmo di Gauß

1. Idea della Decomposizione LU

Osserviamo che il *metodo di eliminazione di Gauss* è una procedura iterativa che va a trasformare una *matrice quadrata* in una *matrice triangolare superiore*. Possiamo sfruttare questa procedura? Sì, se vado anche a *memorizzare* i coefficienti moltiplicatori, ottengo una *matrice triangolare inferiore* per cui

$$A = L \cdot U$$

Esempio: data

$$A = \begin{pmatrix} 2 & 1 & 3 \\ 4 & 3 & 10 \\ 4 & 4 & 17 \end{pmatrix}$$

Abbiamo che col *metodo di eliminazione di Gauss* otteniamo

$$\tilde{A} = U = \begin{pmatrix} 2 & 1 & 3 \\ 0 & 1 & 4 \\ 0 & 0 & 2 \end{pmatrix}$$

E inoltre salvando anche i coefficienti abbiamo

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 3 & 1 \end{pmatrix}$$

Facendo dei calcoli otteniamo $A = LU$.

Q. Perché siamo interessati a questa decomposizione?

Una prima risposta è data da una *"maggiore semplicità"* nel risolvere sistemi lineari. Infatti dato $Ax = y$, otteniamo e $A = LU$, abbiamo

$$Ax = y \iff LUx = L(Ux) = y \implies \begin{cases} Lv = y \\ Ux = v \end{cases}$$

Ossia abbiamo un *sistema di equazioni lineari*, entrambi con *matrice di coefficienti a gradino* (quindi *"facilmente risolvibili"*). In questo modo, sostituendo una successione di sistemi lineari $(y_n)_n$, basta calcolare $Ux = v$ e poi risolvere $Lv = y$ di volta in volta, che è più semplice di applicare Gauss ad ogni sistema lineare.

Inoltre, un'altra motivazione particolare è quella di rendere più *veloce* il calcolo del determinante. Col metodo di Laplace si avrebbe un *costo fattoriale* $O(n!)$. Possiamo renderlo più semplice? Sì! Vediamo il seguente teorema.

#Osservazione

> Osservazione (calcolo del determinante di una decomposizione LU).

Si ha che, data una composizione LU di A:

$$A = LU$$

Allora per *Binet* il suo determinante è

$$\det A = \det L \det U$$

Siccome abbiamo entrambi matrici triangolari, il loro determinante si riduce alla produttoria delle diagonali:

$$\det A = \prod_n L[n,n] \prod_n U[n,n]$$

Tuttavia $L[n,n] = 1$ sempre, da cui ho che

$$\det A = \prod_n U[n,n]$$

X

2. Esistenza e Unicità della Decomposizione LU

#Teorema

> Teorema (di esistenza e di unicità della decomposizione LU).

Sia $A \in \mathbb{R}^{n \times n}$ con $n > 0$ fissato. Se *tutti i sottoinsiemi principali di testa* fino a ordine $n-1$ non sono singolari, i.e. $\det A_1, \det A_2, \dots, \det A_{n-1} \neq 0$, allora $\exists L, U$ matrici triangolari di cui L è inferiore e U è superiore per cui $A = LU$.

Se inoltre A è *invertibile* e L ha diagonale con tutti 1, allora L, U sono uniche.

#Dimostrazione

DIMOSTRAZIONE del Teorema 2

$\exists$: Diamo solamente dei *cenni* per questa parte della dimostrazione. L'idea è quella di *rappresentare* la trasformazione del passo k-esimo del MEG in una *forma matriciale*. In particolare, lo facciamo *definendo* L_k la *matrice elementare di Gauss*, data da

$$L_k := \mathbb{1} - l_k e_k^T$$

dove $e_k \in \mathcal{E}$ ed l_k è il vettore dei moltiplicatori dei moltiplicatori data da

$$l_k := (\underbrace{0, \dots, 0}_{k}, l_{k+1;k}, \dots, l_{k;k})$$

Naturalmente per $i > k$ ho $l_{i;k} = \frac{A[i,k]}{A[k,k]}$. Quindi L_k è una matrice identità dove alla k-esima colonna tutti gli elementi sotto la diagonale sono formati dai moltiplicatori, i.e.

$$L_k = \begin{pmatrix} 1 & \dots & \dots & \dots & 0 \\ 0 & \ddots & & & \vdots \\ \vdots & & 1 & & \vdots \\ \vdots & & -l_{k+1,k} & \ddots & \vdots \\ \vdots & & \vdots & & \vdots \\ 0 & & \underbrace{-l_{n,k}}_{k} & & 1 \end{pmatrix}$$

Notiamo che il *passo di eliminazione k-esimo* è proprio $A^{(k+1)} = L_k \cdot A^{(k)}$. Pertanto si ha complessivamente

$$L_{n-1}L_{n-2}\dots L_1 A = U$$

Notiamo che la *le matrici triangolari sono chiusi rispetto alla moltiplicazione*, e siccome le matrici *triangolari sono invertibili* (calcolare il loro determinante per credere!) posso pre-moltiplicare per la inversa e ottenere

$$A = (L_{n-1}L_{n-2}\dots L_1)^{-1}U = (L_{n-1}^{-1}\dots L_1^{-1})U$$

Chi è L_k^{-1}? So che $L_k = \mathbb{1} - l_i e^T$, allora chiaramente $L_k^{-1} = \mathbb{1} + l_i e^T$! (calcolare per crederci). Ponendo $L = (L_{n-1}^{-1}\dots L_1^{-1})$, abbiamo per la chiusura della *"triangolarità"* sotto il prodotto che L è triangolare inferiore. Inoltre, U è *triangolare superiore* in quanto *"rimuoviamo"* elementi del triangolo inferiore, facendo permanere dunque *solamente* gli elementi del triangolo superiore.

!: Supponiamo che L abbia diagonale unitaria (per ipotesi, non è una supposizione necessaria ma è stata attuata per semplificarci i conti). Notiamo inoltre che A invertibile dal fatto che è non singolare. Facciamo la dimostrazione di questa parte *per assurdo*, ossia suppongo che

$$\exists \tilde{L} \neq L, \tilde{U} \neq U : A = \tilde{L}\tilde{U}$$

Allora segue che

$$A = LU \implies \tilde{L}\tilde{U} = LU$$

Per la *formula di Binet* tutte le matrici triangolari L, L', U, U' sono *invertibili*; pertanto posso *"segregare"* tutte le matrici triangolari di tipi diversi su lati diversi, effettuando moltiplicazioni opportune. In particolare avrò

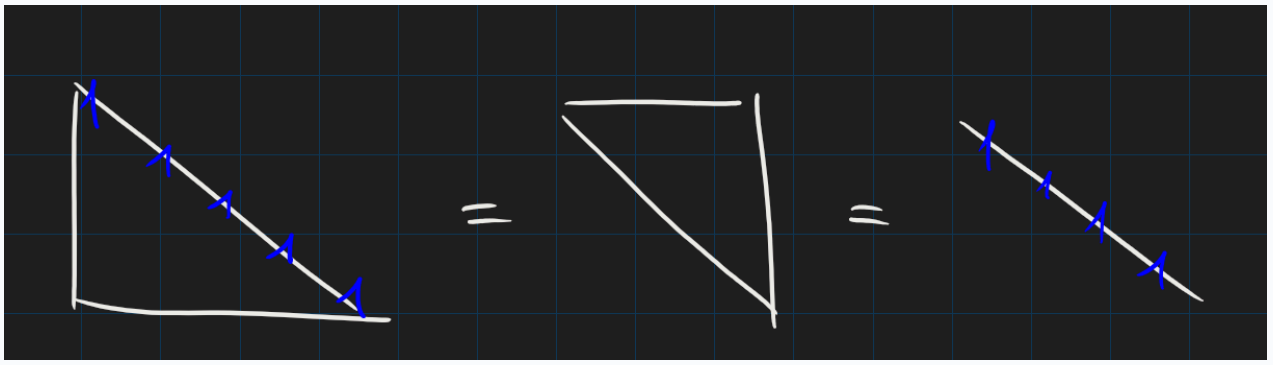
$$\tilde{L}(\tilde{U}\tilde{U}^{-1}) = L(U\tilde{U}^{-1}) \iff L^{-1}\tilde{L} = U\tilde{U}^{-1}$$

Per *chiusura* di triangolarità, deduco che $L^{-1}\tilde{L}$ è *triangolare inferiore* e $U\tilde{U}^{-1}$ è *triangolare superiore*. Notiamo che ho un'uguaglianza tra *matrici triangolari*! Dato che sicuramente

$L^{-1}\tilde{L}$ avrà sicuramente *diagonale unitaria*, segue che questa vale *se e solo se* i prodotti valgono $\mathbb{1}$. Ovvero

$$L^{-1}\tilde{L} = U\tilde{U}^{-1} = \mathbb{1}$$

Pertanto $L^{-1}\tilde{L} = \mathbb{1} \iff \tilde{L} = L$ e questo vale analogamente per $U, \tilde{U}$ facendo derivare l'assurdo con le ipotesi iniziali. ■



Matrici di Permutazione

Matrici di Permutazione

X

Matrici di permutazione. Definizione di matrice di permutazione, esempio. Osservazione: pre-multiplicazione e post-multiplicazione per una matrice di permutazione. Proprietà: ortogonalità delle matrici di permutazione, determinante delle matrici di permutazione.

X

0. Voci correlate

- Matrice
- Algoritmo di Gauß
- Tecniche di Pivoting per Eliminazione di Gauss

1. Definizione di Matrice di Permutazione

IDEA. Una delle *operazioni fondamentali* OE che trasforma matrici in matrici equivalenti è lo *scambio delle righe o colonne*. Come possiamo esprimerlo *sotto forma* di prodotto tra matrici? La nozione che permette di fare ciò sono le *matrici di permutazione*.

#Definizione

> Definizione (matrice di permutazione).

Una matrice $P \in \mathbb{R}^{n\times n}$ si dice *di permutazione* se è dato dallo scambio di ≥ 2 *righe o colonne* della matrice identità $\mathbb{1}$. Equivalentemente, è di permutazione sse ogni colonna e ogni riga ha somma 1

ESEMPIO. Siano date P, A date da

$$A = \begin{pmatrix} 1 & 1 & 3 \\ 2 & 2 & 20 \\ 3 & 6 & 4 \end{pmatrix}$$

Notiamo che questa matrice ha *sottomatrice di testa principale singolare*, i.e. $\det A_2 = 0$. Tuttavia, scambiando la prima con la terza riga non si ha più questo problema. Per farlo definiamo

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

Allora si ha che

$$PA = \begin{pmatrix} 3 & 6 & 4 \\ 2 & 2 & 20 \\ 3 & 6 & 4 \end{pmatrix}$$

#Osservazione

Osservazione (facciamo la pre-moltiplicazione).

Notiamo che il *passo cruciale* è che facciamo la *pre-moltiplicazione* invece della *post-moltiplicazione*. Possiamo dire che la *pre-moltiplicazione* va ad agire sulle *righe*, invece la *post-moltiplicazione* sulle colonne. Vedremo che ciò ha senso dimostrando che le matrici di permutazione sono effettivamente *ortonormali*.

X

2. Proprietà delle Matrici di Permutazioni

#Proposizione

Proposizione (ortogonalità).

Le matrici di permutazioni sono *ortogonali*.

#Dimostrazione

DIMOSTRAZIONE della Proposizione 3

Banale, le sue colonne (o righe) formano la base canonica $\mathcal{E}$ per $\mathbb{R}^n$. ■

Dunque si ha che le *matrici di permutazioni* sono pure *invertibili* (infatti $PP^T = P^TP = \mathbb{1} \implies P^{-1} = P^T$)

#Proposizione

> Proposizione (determinante della matrice di permutazione).
>
> Una matrice di permutazione P ha determinante $\det P = \pm 1$, ovvero $\det P \in \{-1, 1\}$.

#Dimostrazione

DIMOSTRAZIONE del Proposizione 4

Per definizione una matrice di permutazione è data dallo *scambio delle righe* di una matrice unità $\mathbb{1}$. Pertanto per la *multilinearità* del determinante (Proposizione 1), si ha che per ogni *scambio* riga/colonna il determinante cambia segno. In particolare, se ho effettuato k scambi allora il determinante è

$$\det P = (-1)^{k+1} \det \mathbb{1}$$

Essendo $\det \mathbb{1} = 1$, si ha la tesi. ■

Pivoting del Metodo di Eliminazione di Gauss

Tecniche di Pivoting per Eliminazione di Gauss

X

Tecnica di PIVOTING del metodo dell'eliminazione di Gauss. Pivoting parziale e completo. Stabilità di MEG con pivoting.

X

---

0. Voci correlate

- Decomposizione LU
- Algoritmo di Gauß
- Matrici di Permutazione

1. Idea di Pivoting

Q. Sia $A \in \mathbb{R}^{n \times n}$. Cosa succede se una *delle sottomatrici di testa principale* ha determinante nullo? Ovvero il *pivot* ad un passo k-esimo è nullo? Purtroppo, non possiamo usare MEG come lo conosciamo...

IDEA. Ad ogni passo della *MEG* vado a *selezionare* il *pivot "più opportuno"*, in particolare che *non sia nulla*. In particolare, andrò ad effettuare lo *scambio* tra righe. In questo modo:

- Permettiamo MEG anche nei casi in cui non è garantita l'esistenza della decomposizione LU; tuttavia la decomposizione è $PA = LU$.
- Rendiamo *stabile* l'algoritmo MEG. Vedremo di definire bene tale nozione dopo.

PIVOTING. Pivoting è la *selezione sistematica* del *pivot* ad ogni passo del metodo di eliminazione di Gauss. Viene fatta mediante lo *scambio* di righe (o colonne) della matrice dei coefficienti. Avremo due metodi.

X

---

2. Pivoting Parziale

PIVOTING PARZIALE. Al passo k-esimo del *metodo di eliminazione di Gauss* si va a selezionare come *elemento del pivot* l'elemento di modulo maggiore della *sottocolonna* k-esima ($A[k:n,k]$). Definiamo dunque l'indice del *pivot* come

$$s := \arg\max_{i=k,\dots,n} |A[i,k]|$$

Allora vogliamo *scambiare* $A[s,:]$ con $A[k,:]$. Viene fatta una matrice di permutazione P dove la riga s-esima è scambiata con la riga k-esima. Generalizzando su $k = 1, \dots, n-1$ abbiamo una *successione di permutazioni*:

$$A^{(k)} = P^{(k-1)} A^{(k-1)}$$

Ad esempio, con $k = 2$ abbiamo

$$L_2 P_2 L_1 P_1 A = U$$

Esisteranno $\tilde{L}_2, \tilde{L}_1$ tali che

$$\underbrace{\tilde{L}_2 \tilde{L}_1}_{L^{-1}} \cdot \underbrace{P_2 P_1}_{P} A = U$$

Quindi

$$PA = LU$$

STABILITA'. Per capire *come* il pivoting parziale va a sopperire (la maggior parte) dell'instabilità di MEG, capiamo prima *da dove* viene questa instabilità.

Una risposta sbagliata è quella di dire che l'instabilità è derivata dalla *sottrazione* per i moltiplicatori, ovvero

$$R_i \leftarrow R_i - l_{ik} R_k$$

La *cancellazione numerica* va ad avvenire *solo* per $R_i \approx l_{ik} R_k$, che è una cosa *altamente improbabile*.

Invece la risposta corretta è il fatto che abbiamo *operazioni* tra numeri *"troppo grandi"* e "*troppo piccoli*", causando al calcolatore di *"trascurare"* delle cifre (in certe istanze, la differenza è di oltre sedici ordini di grandezza!)

Il *pivoting parziale* va a sopperire questo problema *evitando* di creare *moltiplicatori grandi* (infatti dividiamo per il *massimo* tra i *"pivot candidati"*). Infatti, si ha che

$$l_{ik} = \frac{|a_{ik}|}{|a_{sk}|} = \max_{j=k,\dots,n} \frac{|a_{ik}|}{|a_{jk}|} \leq 1$$

Tuttavia, in certi casi *estremi* si ha comunque questo problema. Tuttavia, *di solito* questo è facilmente risolvibile portando in scala la matrice.

ESEMPIO. (*Instabilità di MEG*)
Sia $A \in \mathbb{R}^{3\times 3}$ data da

$$A = \begin{pmatrix} 1 & 1 & 3 \\ 2 & 2+10^{-10} & 20 \\ 3 & 6 & 4 \end{pmatrix} \rightsquigarrow \begin{pmatrix} 1 & 1 & 3 \\ 0 & 10^{-10} & 14 \\ 0 & 3 & -5 \end{pmatrix}$$

Al passo $k = 2$ in *MEG* abbiamo il pivot $\lambda = a_{22} = 10^{-10}$. Moltiplicandolo con a_{32} abbiamo

$$l_{3,2} = \frac{3}{10^{-10}} = 3 \cdot 10^{10}$$

Quindi naturalmente

$$\tilde{a}_{3,2} = a_{3,2} - l_{3,2} \cdot a_{2,2} = 0$$

Tuttavia, al prossimo elemento abbiamo il seguente problema:

$$\tilde{a}_{3,3} = -5 \ominus (3 \cdot 10^{10}) \approx -3 \cdot 10^{10}$$

Andiamo effettivamente a *"trascurare"* il termine -5, causando instabilità. ■

X

---

3. Pivoting Completo

PIVOTING COMPLETO (o TOTALE). Il *pivoting totale* è una variante del *pivoting parziale*, dove invece di cercare *solamente* sulla colonna dei *"pivot"* $A[k]$, andiamo a cercare sull'intera sottomatrice *"attiva"* del passo k-esimo:

$$(s, r) := \arg \max_{k \leq i,j \leq n} |A[i, j]|$$

Quindi scambieremo le righe k, s e le colonne k, r. ■

Condizionamento dei Sistemi Lineari

Condizionamento dei Sistemi Lineari

Condizionamento dei sistemi lineari, fattore di condizionamento su matrici, residui ed errori.

---

0. Voci correlate

- Condizionamento dei Problemi
- Norme Matriciali

1. Condizionamento dei Sistemi Lineari

INTRODUZIONE. Per la risoluzione dei sistemi lineari, abbiamo visto sia degli algoritmi *stabili* che *instabili*. Tuttavia, un problema ancora più intrinseca è il *condizionamento* dei sistemi lineari, ovvero la *"qualità di output cambiato per una perturbazione piccola dell'input"*. Ossia l'idea è date $A, \hat{A}$ e $b, \hat{b}$ leggermente *"diverse"*, le soluzioni date dai sistemi

$$Ax = b, \hat{A}\hat{x} = \hat{b}$$

Le soluzioni $x, \hat{x}$ potrebbero essere *"grandi"*!

Per fare l'analisi, faremo le seguenti assunzioni:

- $A \in \mathbb{R}^{n \times n}$ sarà *fissa* per rendere i calcoli semplici. In particolare sarà *invertibile* (non singolare)
- Invece defineremo le *perturbazioni* $\delta b, \delta x$ per cui $\hat{b} = b + \delta b$ e $\hat{x} = x + \delta x$. Inoltre assumeremo che $b \neq 0$, altrimenti si avrebbe la soluzione banale $x = 0$.
- Quindi l'errore accumulato sarà su δb

$$Ax = b \rightsquigarrow (A + \underbrace{\delta A}_{0})(x + \delta x) = (b + \delta b) \rightsquigarrow A(x + \delta x) = (b + \delta b)$$

#Proposizione

> Proposizione (relazione tra errore e residuo).
>
> Si ha che, dati A, x, b e le perturbazioni $\delta x, \delta b$ che
>
> $$\delta x = A^{-1}\delta b$$

#Dimostrazione

DIMOSTRAZIONE del Proposizione 1

Semplicemente isoliamo δx in (1), tenendo conto che $Ax = b$.

$$A(x + \delta x) = (b + \delta b) \iff \underbrace{Ax}_{b} + A\delta x = b + \delta b \iff \delta x = A^{-1}\delta b$$

Concludendo. ■

Vogliamo quantificare δx in un *numero scalare*. Come lo facciamo? Naturalmente con le morme!

$$\|\delta x\|_{\mathbb{R}^n} = \|A^{-1}\delta b\|_{\mathbb{R}^n}$$

Se usiamo una *norma matriciale compatibile* con la norma $\mathbb{R}^n$ (una qualsiasi va bene), come ad esempio la norma matriciale indotta, allora abbiamo la seguende disuguaglianza:

$$\|\delta x\|_{\mathbb{R}^n} \leq \|A^{-1}\| \cdot \|\delta b\|_{\mathbb{R}^n}$$

Volendo calcolare gli *errori relativi* vorremmo dividere tutto per la norma di b e x. In particolare ricaviamo quella per b come

$$\|b\|_{\mathbb{R}^n} = \|Ax\|_{\mathbb{R}^n} \leq \|A\| \cdot \|x\|_{\mathbb{R}^n} \implies \frac{1}{\|x\|_{\mathbb{R}^n}} \leq \frac{\|A\|}{\|b\|_{\mathbb{R}^n}}$$

Pertanto

$$\frac{\|\delta x\|_{\mathbb{R}^n}}{\|x\|_{\mathbb{R}^n}} \leq \frac{\|A\|}{\|b\|_{\mathbb{R}^n}}\|x\|_{\mathbb{R}^n} \leq \frac{\|\delta b\|_{\mathbb{R}^n}}{\|b\|_{\mathbb{R}^n}}\|A\| \cdot \|A^{-1}\|$$

Ossia il residuo su x dipende dal residuo su b scalato per $\|A\| \cdot \|A^{-1}\|$. Definiamo tale quantità come *numero di condizionamento* di questa matrice:

#Definizione

> **Definizione (fattore di condizionamento di una matrice).**
>
> Sia $A \in \mathbb{R}^{n \times n}$ invertibile. Si definisce il suo *numero di condizionamento* come
>
> $$K(A) := \|A\| \cdot \|A^{-1}\|$$
>
> dove la norma matriciale è una *compatibile* con la norma in $\mathbb{R}^n$.

Notiamo che usando la *norma indotta* si ha sicuramente che

$$1 = \|\mathbb{1}\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\|$$

Quindi sicuramente si *amplificherà* (o rimarrà uguale) l'errore causata dalla perturbazione in b.

In particolare, data una norma indotta in $p = [1, +\infty]$, parametrizzeremo il numero di condizionamento come $(K_p)_p$.

In particolare si ha che:

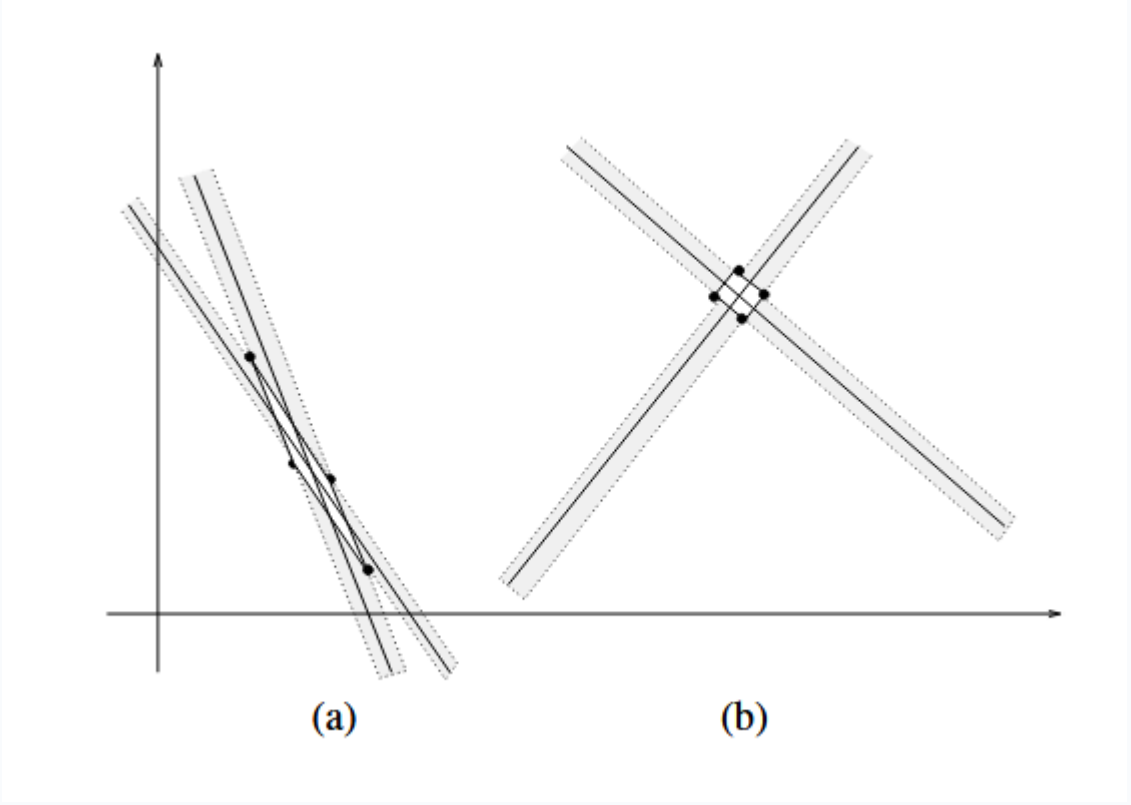
$$K_2(A) = \sqrt{\frac{\rho(A^T A)}{\lambda_{\min}(A^T A)}}$$

Se A simmetrica, allora $A^T A = A^2$ e dunque $\rho(A^2)$ è data da $\lambda^2_{\max}(A)$ e $\frac{1}{\lambda_{\min}(A^2)}$ è data da $\lambda^2_{\max}(A^{-1})$ e dunque

$$K_2(A) = \sqrt{\rho^2(A)\rho^2(A^{-1})} = \rho(A)\rho(A^{-1})$$

(non serve il modulo in quanto è già stato implicitamente posto nella definizione del raggio spettrale).

Geometricamente, il *numero di condizionamento* si può pensare a quanto *"siano paralleli"* le rette della matrice. Un buon condizionamento (i.e. 1) corrisponde a delle rette perfettamente ortogonali; altrimenti il *malcondizionamento* corrisponde a delle rette *quasi parallele*.



X

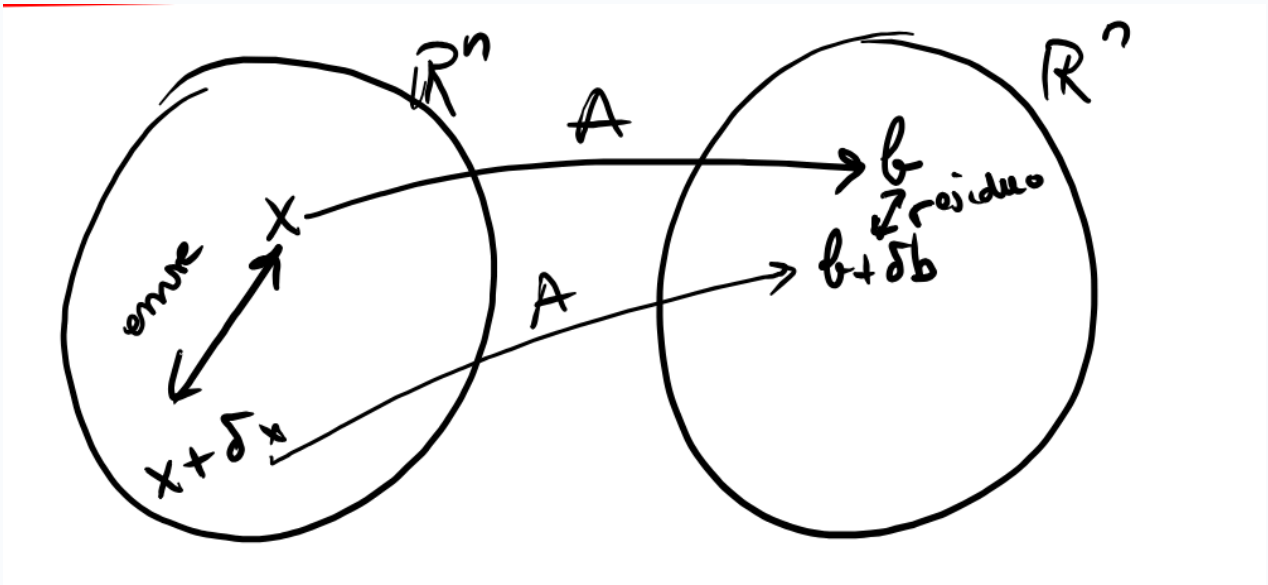
2. Errori e Residui

Q. Da un punto di vista pratico, come *possiamo* calcolare gli errori commessi?

Lo facciamo definendo il *residuo* r che va a *"misurare"* l'errore tra $b, b + \delta b$ e l'*errore* ε che misuri la distanza tra $x, x + \delta x$

$$\varepsilon = x - (x + \delta x)$$
$$r = b + \delta b = b + A(x + \delta x) = b - A\hat{x}$$

dove $\hat{x}$ è la soluzione calcolata.



Essendo che conosciamo $\hat{x}$, segue che di solito conosciamo r, ma non ε! Quindi vogliamo trovare un modo per relazionare r rispetto a ε. Tuttavia, notiamo che r è riconducibile proprio a δb, che a sua volta è riconducibile al fattore di condizionamento K. Infatti

$$r = b - Ax - A\delta x = -A\delta x$$

Sostituendo δx con Proposizione 1, otteniamo

$$-A\delta x = -AA^{-1}\delta b = -\delta b$$

Concludendo. ■

Pertanto si ha che

$$\boxed{\frac{\|\delta x\|_{\mathbb{R}^n}}{\|x\|_{\mathbb{R}^n}} \leq K(A)\frac{\|r\|_{\mathbb{R}^n}}{\|b\|_{\mathbb{R}^n}}}$$

Quindi il nostro obbiettivo sarà quello di *minimizzare* la norma di r il più possibile.

Decomposizione di Cholesky

Decomposizione di Cholesky

X

Decomposizione di Cholesky su matrici simmetriche e definite positive. Step preliminari: decomposizione LDM^T, LDL^T. Risoluzione di sistemi lineari con matrici decomposti secondo Cholesky.

X

0. Voci correlate

- Decomposizione LU
- Matrici Definite con Segno

1. Decomposizione di Cholesky

Vediamo un altro modo per decomporre matrici.

#Teorema

> Teorema (decomposizione di Cholesky).
>
> Sia $A \in \mathbb{R}^{n\times n}$ una *matrice simmetrica* e *definita positivamente* (dunque invertibile!). Allora $\exists R \in \mathbb{R}^{n\times n}$ matrice *triangolare inferiore* (e sopprattutto reale!) tale che
>
> $$A = RR^T$$

La dimostrazione della *decomposizione di Cholesky* richiede dei risultati preliminari, partendo dalla decomposizione LU.

#Lemma

> Lemma (decomposizione LDM^T).
>
> Sia $A \in \mathbb{R}^{n\times n}$ tale che esista una decomposizione LU. Allora esiste una matrice diagonale D e una matrice triangolare inferiore M tale che è composta da diagonale tutti uno e tale che
>
> $$A = LDM^T$$

#Dimostrazione

DIMOSTRAZIONE del Lemma 2

Notiamo che dalla *decomposizione LU* si ha

$$A = LU \iff A = L\mathbb{1}U = LDD^{-1}U$$

Dove D è una matrice diagonale. In particolare, selezioniamo $D = \texttt{diag}(U)$ (ossia è composta dalla diagonale di U). Allora si ha che $D^{-1}U$ è comunque una matrice triangolare superiore, in particolare di forma

$$D^{-1}U = \begin{pmatrix} u_{11}^{-1} & 0 & \dots & 0 \\ 0 & u_{22}^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & u_{nn}^{-1} \end{pmatrix} U$$

Quindi premoltiplicando andiamo a scalare la diagonale di U in un'unità, dandoci una forma del tipo

$$D^{-1}U = \begin{pmatrix} 1 & \tilde{u}_{12} & \dots & \tilde{u}_{1n} \\ & 1 & \dots & \tilde{u}_{2n} \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix}$$

Dove per $i > j$ definiamo $\tilde{u}_{ij} := \frac{u_{ij}}{u_{ii}}$. Definendo dunque $D^{-1}U =: M^T$, abbiamo la tesi. ■

#Lemma

> Lemma (decomposizione LDL^T).
>
> Sia A una *matricei invertibile e simmetrica* per cui esiste *unicamente* una decomposizione LU. Allora esiste una matrice triangolare inferiore L e diagonale D tali che
>
> $$A = LDL^T$$

#Dimostrazione

DIMOSTRAZIONE del Lemma 3

Per il Lemma 2 si ha che esistono L, D, M tali che $A = LDM^T$. Si tratta di verificare dunque che $M = L$. Notiamo che essendo A simmetrica si ha che

$$A = LDM^T \implies A^T = MDL^T$$

Pertanto

$$LDM^T = MDL^T$$

Essendo L, M *uniche* l'identità (1) vale se e solamente $M = L$, concludendo. ■

Adesso siamo pronti per dimostrare Cholesky.

#Dimostrazione

DIMOSTRAZIONE del Teorema 1

Osserviamo che A è *definita positivamente*, dunque vale la *decomposizione unica LU* da cui per il Lemma 3 abbiamo che $\exists L, D$ tali che

$$A = LDL^T$$

Infatti, A definita positivamente implica che è invertibile con tutti minori non-nulli (per criterio di Sylvester, Teorema 6) (*NOTA: QUESTO PASSAGGIO E' DA SCRIVERE E NON DA SALTARE!*)

Notiamo che inoltre pure D è definita positivamente, infatti effettuando un cambiamento di variabili nella definizione otteniamo

$$\forall x \neq 0, x^T Ax > 0 \iff x^T LDL^T x > 0 \iff y^T Dy > 0$$

Questo passaggio non è problematico, infatto la trasformazione $y := L^T x$ non comporta in alcun modo *"perdita"* di vettori, infatti $L^T x = 0$ vale sse $x = 0$ oppure L ha determinante

nullo. Chiaramente L non ha determinante nullo in quanto è una matrice triangolare con tutti uno; pertanto $\ker L = \{0\}$.

Questo significa che *ogni valore* della diagonale D è positiva, infatti per ogni elemento della base canonica $e_i \in \mathcal{E}$ si ha che $e_i^T D e_i > 0$ dove $e_i^T D e_i$ non è altro che $D[i,i]$ (infatti andiamo a *"selezionare"* la riga i-esima con la premoltiplicazione e la colonna i-esima con la postmoltiplicazione). Pertanto $\forall i, D[i,i] =: d_i > 0$.

Quindi definendo $\sqrt{D} := \texttt{diag}(\sqrt{d_1}, \dots, \sqrt{d_n})$ abbiamo che $D = \sqrt{D}\sqrt{D}$. Notiamo il fatto che R è una matrice reale è garantita dal fatto che $d_i > 0$, che a sua volta è assicurata dal fatto che A è definita positivamente, concludendo la dimostrazione. ■

X

---

2. Considerazioni Pratiche

Q. Perché?

La fattorizzazione di Cholesky ha le seguenti motivazioni:

- Il suo costo computazionale è minore, ma sopprattutto più *stabile* rispetto alla decomposizione LU (fatta con MEG)
- Ancora oggi è un aspetto di cui ricercare, sopprattutto nell'ambito ML e approssimazione dei dati
- La risoluzione dei sistemi lineari diventa più semplice

Q. Come calcolo una decomposizione di Cholesky?

Semplicemente, *DOPO AVER VERIFICATO LE IPOTESI*, vado ad imporre l'identità $RR^T = A$. Dopodiché lo risolvo colonna per colonna, ovvero ho una famiglia di sistemi lineari con coefficienti triangolari (dunque risolvibili in $O(n^2)$)

$$R(R^T)^{(i)} = A^{(i)}$$

In questo modo otterrò ogni colonna della matrice R, che va comporre la matrice finale. La complessità complessiva è $O(n^3)$.

Q. Sistemi lineari?

Banalmente

$$R\underbrace{R^T x}_{v} = b \iff \begin{cases} Rv = b \\ R^t x = v \end{cases}$$

Risoluzione dei Sistemi Sovradeterminati

Risoluzione dei Sistemi Sovradeterminati

X

Problema della risoluzione dei sistemi sovradeterminati. Caratterizzazione delle soluzioni OLS di sistemi sovradeterminato con al soluzione al sistema delle equazioni normali.

X

0. Voci correlate

- Approssimazione ai Minimi Quadrati

1. Sistemi Sovradeterminati

Sia $A \in \mathbb{R}^{m\times n}$ con $m \ggg n$. Allora data $b \in \mathbb{R}^n$, ho il seguente *sistema sovradeterminato*

$$Ax = b$$

Chhiamo che questo sistema è *risolvibile* (o compatibile) *se e solo se* vale che $b \in \text{im } A$, che è molto raro per $m \ggg n$! Pertanto per $b \notin \text{im } A$ definiamo la *"soluzione"* x come quella che minimizza lo scarto ai minimi quadrati:

$$x^* \approx x, x^* = \arg\min_{x\in\mathbb{R}^n} \|b - Ax\|_2$$

In particolare, essa si dice la *regressione OLS*.

Vediamo il seguente *teorema di caratterizzazione* della soluzione x^*:
#Teorema

> Teorema (caratterizzazione dei sistemi sovradeterminati).
>
> Si ha che x^* risolve un *sistema sovradeterminato* se e solo se essa risolve anche il *sistema delle equazioni normali* $(A^TA)x = A^Tb$. Ovvero
>
> $$x^* = \arg\min_{x\in\mathbb{R}^n} \|b - Ax\|_2 \iff x^* = (A^TA)^{-1}A^Tb$$

#Dimostrazione
DIMOSTRAZIONE del Teorema 1
" $\Longleftarrow$ ": Per ipotesi x^* va a risolvere il sistema delle equazioni normali, quindi $(A^TA)x^* = A^Tb$, ovvero $A^T(b - Ax^*) = 0$. Pertanto A e $b - Ax^*$ sono ortogonali. Allora calcolando $\|b - Ax\|_2$ abbiamo

$$\begin{aligned}\|b-Ay\|_2^2 &= \|b-Ax^*+Ax^*-Ay\|_2^2\\ &= \|(b-Ax)+A(x^*-y)\|_2^2\\ &= ((b-Ax^*)+A(x^*-y))^T((b-Ax)+A(x^*-y))\\ &= \|b-Ax^*\|_2^2+\|A(x^*-y)\|_2^2+2(A(x^*-y))^T(b-Ax)\\ &= \ldots+2(x^*-y)^T\underbrace{A^T(b-Ax)}_{0}\\ &= \|b-Ax^*\|_2^2+\|A(x^*-y)\|_2^2\end{aligned}$$

Siccome la *norma è sempre non-negativa*, sicuramente vale sempre che $\|b-Ay\|_2^2 \geq \|b-Ax^*\|_2^2$ e dunque x^* minimizza la distanza quadrata.
" $\implies$ ": Dimostrare questo equivale a dimostrare che se x^* minimizza la quantità $\|b-Ax\|_2^2$, allora il residuo $r=b-Ax^*$ è ortogonale a $\operatorname{im} A$. Ossia, $b-Ax^* \perp A$.
Partiamo osservando che se x^* minimizza la distanza euclidea allora $Ax^* \perp b-Ax^*$ (*teorema di Pitagora generalizzato su spazi vettoriali arbitrari*), quindi $\mathbb{R}^n$ è formata dalla *somma diretta* dall'immagine di A e la sua ortogonale:

$$\mathbb{R}^n = \operatorname{im} A \oplus (\operatorname{im} A)^\perp$$

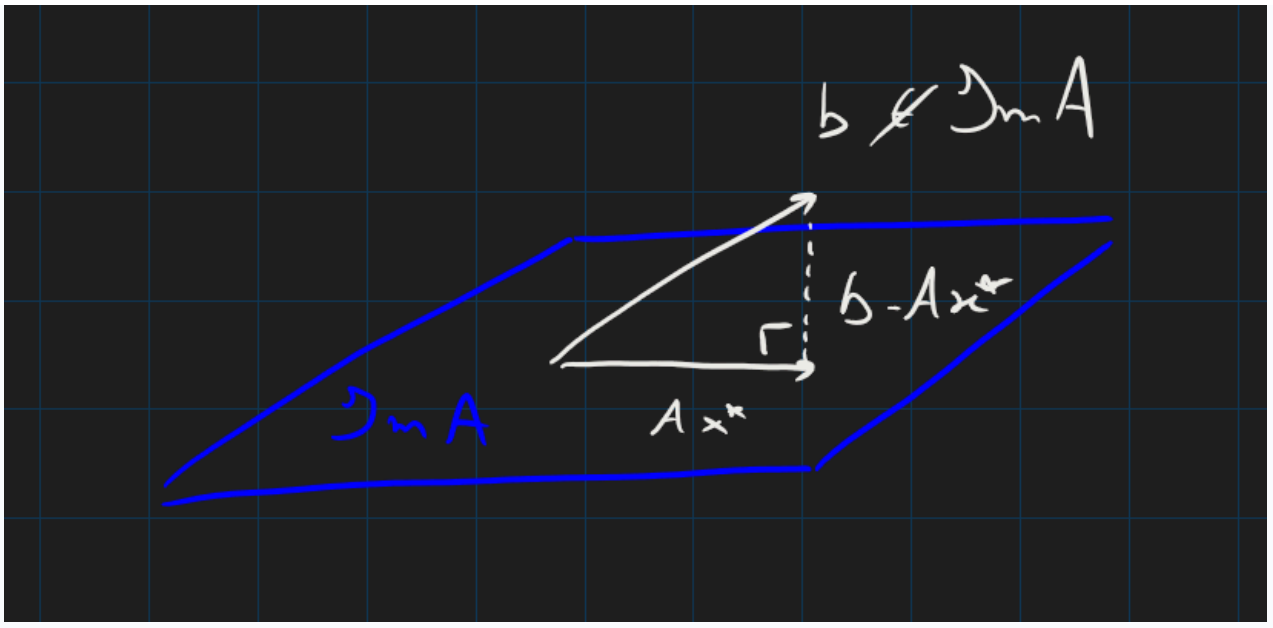
Pertanto dato un vettore qualsiasi $b \in \mathbb{R}^n$, sicuramente avrà una *componente* in $\operatorname{im} A$ e la sua perpendicolare. Denotiamo $b', b_\perp$ tali componenti: $b = b' + b_\perp$. Quindi calcolando la norma $\|b-Ax^*\|_2^2$ otteniamo

$$\begin{aligned}\|b-Ax^*\|_2^2 &= \|b'-Ax^*+b_\perp\|_2^2\\ &= \langle (b'-Ax^*)+b_\perp ; (b'-Ax^*)+b_\perp\rangle\\ &= \|b'-Ax^*\|_2^2+\|b_\perp\|_2^2+\underbrace{2(b'-Ax^*)^T(b_\perp)}_{\text{"}\perp\text{"} \implies 0}\\ &= \|b'-Ax^*\|_2^2+\|b_\perp\|_2^2\end{aligned}$$

Ricordandomi che x^* va a minimizzare la distanza euclidea tra Ax^* e b, otteniamo che $\|b'-Ax^*\|_2 = 0$. Pertanto per *non-degeneratezza* abbiamo che $b'-Ax=0$. Pertanto il residuo $r=b-Ax$ è data da

$$r = b-Ax = b'-Ax+b_\perp = b_\perp \in (\operatorname{im} A)^\perp$$

Concludendo. ■



Osserviamo che l'implicazione di verso " $\Longleftarrow$ " può essere interpretata come una sorta di *"bias implicito"* di scegliere x^* con quella formula (Proposizione 1), nell'ambito di *Machine Learning*.

CONCLUSIONE. Al posto risolvere $Ax = b$ risolveremo, per ottenere la *"migliore"* approssimazione il sistema

$$(A^T A)x = A^T b$$

Per risolverla, una soluzione *"naive"* sarebbe quella di osservare che $A^T A$ è simmetrica e (di solito!) definita positiva, dunque possiamo fattorizzarla con Cholesky. Tuttavia, non è una *buona* idea in quanto $A^T A$ è *mal condizionata* in quanto è una specie di *"matrice di Vandermonde"*. In effetti, si dimostra che

$$K(A^T A) = (K(A))^2$$

Come si può rimediare?

- Calcolare una decomposizione QR della matrice
- Calcolare una decomposizione SVD della matrice

Decomposizione QR

Decomposizione QR

X

Decomposizioen QR.

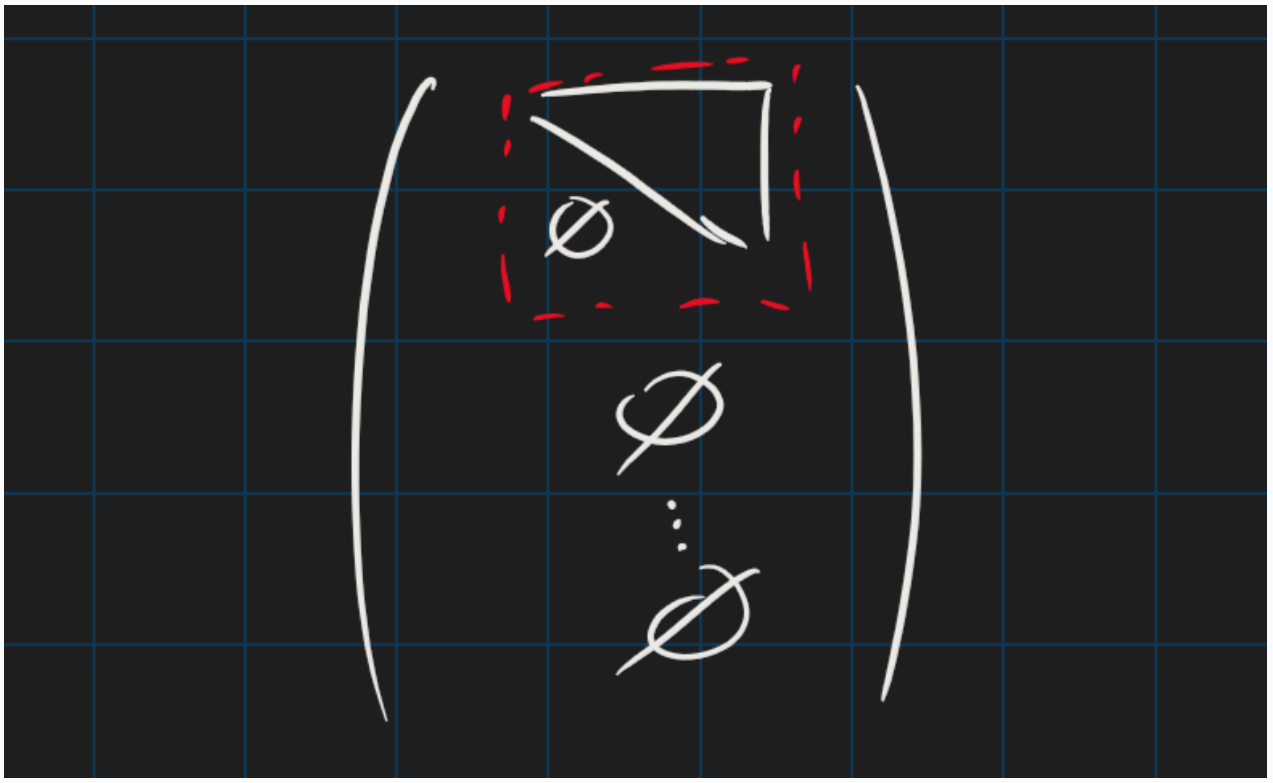
X

0. Voci correlate

- Risoluzione dei Sistemi Sovradeterminati

1. Decomposizione QR

Sia $A \in \mathbb{R}^{m \times n}$ con $m \gg n$. Una *decomposizione QR* di A è quando esistono una matrice $Q \in \mathbb{R}^{m \times m}$ ortogonale quadrata e R una matrice R *"pseudo-triangolare superiore"*, ovvero di forma



Quindi

$$A = QR$$

Notiamo che essendo Q ortogonale, essa preserva la norma dei vettori:

$$\|Qx\| = \|x\|$$

Tuttavia abbiamo un paio di problemi:

- Q abita in una *dimensione* m^2, che è alta! Dunque per la *curse of dimensionality* potrebbe essere una *matrice sparsa*.
- Inoltre *molte* informazioni in Q vengono perse, in quanto da $n+1$-esima riga di R in poi, azzeriamo tutto. In particolare, azzeriamo le colonne.

Quindi per *"risparmiare"* su questi costi, separeremo Q in due sottomatrici $(\tilde{Q}, Q)$ dove $\tilde{Q} \in \mathbb{R}^{m \times n}$ e $Q \in \mathbb{R}^{m \times m-n}$. Quidndi

$$A = \begin{pmatrix} \tilde{Q} & Q \end{pmatrix} \begin{pmatrix} \blacktriangledown \\ 0 \end{pmatrix}$$

Possiamo effettivamente *"buttare via"* Q, dandoci

$$A = \tilde{Q}\tilde{R}$$

Questa si dice *decomposizione QR "skinny"* (o *"light"*).

Relazione tra Autovettori di Matrici di Gram

Relazione tra Autovettori di Matrici di Gram

X

Lemma preliminare per la SVD: relazione tra autovalori e autovettori di Matrici di Gram.

X

---

0. Voci correlate

- Decomposizione ai Valori Singolari
- Teorema spettrale

1. Lemma Preliminare

Sia $A \in \mathbb{R}^{m \times n}$, con $m \gg n$. Richiamiamo che A^TA, AA^T (*le matrici di Gram* dei dati) sono *simmetriche e definite semipositive*, dunque per il *teorema spettrale* sono ortogonalizzabili con una base reale non-negativa e con cambiamenti di base ortogonali.

Notiamo che data U matrice che diagonalizza A^TA abbiamo

$$U^TA^TAU = D := \texttt{diag}(\lambda_1, \dots, \lambda_r, \lambda_{r+1}, \dots, \lambda_n)$$

Denotiamo r come l'*ultimo indice* per cui $\lambda_r > 0$; in particolare, senza perdere di generalità, richiediamo che $\lambda_\bullet$ siano ordinati nel senso decrescente, i.e.

$$\lambda_1 \geq \dots \geq \lambda_r > \lambda_{r+1} = \dots = \lambda_n = 0$$

Il discorso vale analogamente per AA^T con la matrice ortogonale di trasformazione V:

$$V^TAA^TV = \hat{D}$$

Dove $\hat{D} := \texttt{diag}(\mu_1, \dots, \mu_s, \mu_{s+1}, \dots, \mu_m)$ tali che

$$\mu_1 \geq \dots \geq \mu_s > \mu_{s+1} = \dots = \mu_m = 0$$

Q. Abbiamo un nesso tra gli *autovalori*, *autovettori* e *molteplicità* di A^TA e AA^T? Quale relazione intercorre tra r, s (che non sono altro che le somme delle *molteplicità geometriche* relativi agli autovalori non nulli per AA^T, A^TA!)

Questo risultato ci servira per giustificare bene la *decomposizione ai valori singolari* (SVD)

#Lemma

Lemma.

Sia $B \in \mathbb{R}^{m\times n}$. Allora valgono le seguenti:

a) Sia $\lambda \in \sigma(B^TB)\setminus\{0\}$, con autovettore relativo $x \in \mathbb{R}^n$. Allora λ è autovalore di BB^T ($\lambda \in \sigma(BB^T)$) con *autovettore corrispondente* Bx.
b) Inoltre, se $x_1, \dots, x_l$ sono *autovettori linearmente indipendenti relativi all'autovalore* $\lambda \neq 0$, allora lo saranno pure $Bx_1, \dots, Bx_l$.

Osserviamo che b) implica che la *molteplicità geometrica* di autovalori non nulli per B^TB saranno *minore o uguale* alla molteplicità geometrica di autovalori non nulli per BB^T:

$$\mu_G(\lambda \in \sigma(B^TB)\setminus\{0\}) \leq \mu_G(\lambda \in \sigma(BB^T))$$

Pertanto applicando il *lemma* su $B = A^TA$ e $B = AA^T$ si ottiene che tutte le molteplicità geometriche degli autovalori sono uguali, concludendo con $r = s$ in equazioni (1), (2).

Facciamo inoltre attenzione che *vale* solamnete per $\lambda \neq 0$! Infatti, vedremo che $n - r \neq m - s$ (le molteplicità geometriche dell'autovalore nullo in A^TA, AA^T).

#Dimostrazione

DIMOSTRAZIONE del Lemma 1

a) Banalmente per definizione ho che x è un *autovettore* relativo a $\lambda \neq 0$, i.e. $B^TBx = \lambda x$. Quindi calcolando $BB^T(Bx)$ otteniamo

$$BB^T(Bx) = B(B^TBx) = B\lambda x = \lambda(Bx)$$

Concludendo.
b) Come caratterizziamo l'*indipendenza lineare* tra vettori? Naturalmente esprimento una loro combinazione lineare, e imponendola nulla se e solo se i coefficienti sono nulli:

$$\sum_l \alpha_l x_l = 0 \iff \alpha = 0$$

Per dimostrare la tesi vogliamo mostrare che

$$\sum_l \alpha_l Bx_l = 0 \overset{?}{\iff} \alpha = 0$$

Per farlo premoltiplichiamo (3) per B^T; notiamo che questo passo è ben definito in quanto l'equazione rimane ugualmente vera ($B^T0 = 0$):

$$B^T\sum_l \alpha_l Bx_l = 0 \iff \sum_l \alpha_l B^TBx_l = 0$$

Sostituendo $B^TBx_l = \lambda x_l$ (per ipotesi, x_l è un autovettore relativo a $\lambda \neq 0$ in B^TB) in (4) otteniamo

$$\sum_l \lambda \alpha_l x_l = \lambda \left(\sum_l \alpha_l x_l \right) = 0$$

Quest'ultima equazione e vera se e solamente se si verificano uno dei due casi:
I) $\lambda = 0$
II) La sommatoria è nulla, che vale se e solo se $\alpha = 0$

Per assurdo la I) non è verificabile, lasciando dunque per forza vera II) e concludendo. ■

SVD

Decomposizione ai Valori Singolari

Decomposizione ai Valori Singolari (SVD). Definizione di una SVD per una matrice, teorema della SVD (idea della dimostrazione, costruzione). Applicazione delle SVD sulla risoluzione di sistemi lineari sovradeterminati.

---

0. Voci correlate

- Risoluzione dei Sistemi Sovradeterminati
- Relazione tra Autovettori di Matrici di Gram
- Analisi delle Componenti Principali
- Risoluzione dei Sistemi Sovradeterminati

1. Introduzione alla SVD

La *decomposizione ai valori singolari* è il miglior algoritmo di riduzione del rango delle matrici. In particolare, viene definita proprio la *"creme de la creme"* delle decomposizioni riducenti del rango (G.W. Stewart, 1998).

IDEA. Data una matrice *simmetrica* A, possiamo diagonalizzarla con degli *autovalori non-negativi* con la matrice di trasformazione P, data da

$$A = P \Lambda P^{-1}$$

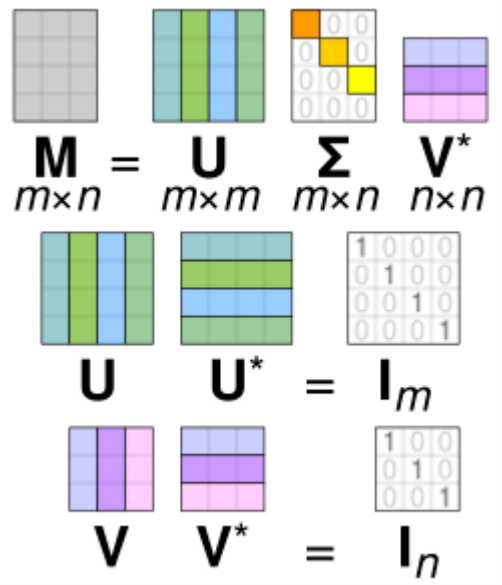
Dove Λ è la matrice diagonale degli autovalori. L'idea della *SVD* è quella di tentare di applicare lo stesso procedimento, usando però le sue *matrici di Gram* AA^T e A^TA. Questa è la *decomposizione spettrale di una matrice simmetrica*.
L'idea della *SVD* è quella di *generalizzare* la composizione spettrale su matrici non quadrate. In particolare, vorremmo usare le matrici di Gram A^TA, AA^T e applicare la decomposizione spettrale ad ognuno della matrice.
Per il lemma sulla relazione tra gli autovalori, le loro molteplicità, e gli autovettori di

A^TA, AA^T (Lemma 1) abbiamo che hanno gli stessi *autovalori* non nulli! Quindi li usiamo, definendo i *valori singolari* come la loro radice quadrata. Denotando U, V come le matrici ortogonali di cambiamento per A^TA, AA^T e denotando Σ come la *"matrice diagonale a blocchi dei valori singolari"*, i.e. $\Sigma = \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix}$ dove Σ_r è la matrice diagonale con valori singolari non nulli (notiamo che $r = \rho(A)$).
Allora otteniamo la decomposizione

$$A = V\Sigma U^T$$

Chiamiamo U, V come i *vettori singolari a destra/sinistra* di A.

Questo tipo di procedimento viene applicato nell'*analisi di dati* e *Machine Learning* (vedere *PCA*).



X

2. Teorema della SVD

#Teorema

> Teorema (della SVD).
>
> Sia $A \in \mathbb{R}^{m\times n}$. Allora sicuramente esistono $V \in \mathbb{R}^{m\times m}, U \in \mathbb{R}^{n\times n}$ ortogonali e $\Sigma \in \mathbb{R}^{m\times n}$ del tipo
>
> $$\Sigma = \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix}$$
>
> dove $\Sigma_r := \texttt{diag}(\sigma_1, \dots, \sigma_r) \in \mathbb{R}^{r\times r}$ per un $r \in \{0, 1, \dots, \rho(A)\}$ tali che
>
> $$A = V\Sigma U^T$$

#Dimostrazione

DIMOSTRAZIONE del Teorema 1

NOTA: IMPORTANTE PER GLI ESERCIZI

La dimostrale formale è omessa ed è oggetto di approfondimento personale (molto contosa...). Faremo la dimostrazione per *costruzione*, senza verificarne la correttezza. Supponiamo WLOG che $m > r$ (se $m = r$ allora abbiamo una semplice decomposizione spettrale).

U: Siccome A^TA è *simmetrica definita semipositiva*, allora sicuramente è diagonalizzabile con una matrice ortogonale. Dati $\lambda_1, \dots, \lambda_r \in \sigma(A) \setminus \{0\}$ ($r = \rho(A^TA)$), otteniamo gli *autovettori* $u_1, \dots, v_u$ tutti relativi ad autovalori diversi (a meno di autovalori con molteplicità multipla). Definiamo invece $u_{r+1}, \dots, u_n$ come gli *autovettori relativi* a 0 (se fa parte dello spettro); notiamo che ha senso che ci siano $n - r$ vettori! Infatti, per il teorema delle dimensioni

$$\dim A = \rho(A) + \dim \ker A \iff n - r = \dim \ker A$$

Definendo $\hat{u}_1, \dots, \hat{u}_n$ come i *vettori normalizzati* (basta dividere per la loro norma), otteniamo $U = (\hat{u}_1 \quad \dots \quad \hat{u}_n)$ e concludiamo.

Σ: La otteniamo gratis dagli autovalori $\lambda_1, \dots, \lambda_r$; basta prendere la loro radice quadrata.

V: Notiamo che dal *lemma di relazione tra autovettori* (Lemma 1) otteniamo che $AU[:, i]$ è *autovettore* di AA^T relativo all'autovalore $\lambda_i \neq 0$. Verifichiamo che tutti i vettori generati in quel modo siano *ortonormali*:

$$\begin{aligned}
\langle AU^{(i)}, AU^{(j)} \rangle &= (AU^{(i)})^T (AU^{(j)}) \\
&= U^{(i)} A^T A U^{(j)} \\
&= \langle U^{(i)}, A^T A U^{(j)} \rangle \\
&= \langle U^{(i)}, \lambda_j U^{(j)} \rangle \\
&= \lambda_j \langle U^{(i)}, U^{(j)} \rangle
\end{aligned}$$

Per costruzione $U^{(i)}, U^{(j)}$ sono già *ortonormali*, i.e. $\langle U^{(i)}, U^{(j)} \rangle = \delta_{i,j}$; pertanto otteniamo che

$$\langle AU^{(i)}, AU^{(j)} \rangle = \lambda_j \delta_{i,j}$$

Notiamo che non è garantita la *normalità*, dunque normalizzeremo tali vettori:

$$\boxed{\hat{v}_i := \frac{AU^{(i)}}{\|AU^{(i)}\|} = \frac{AU^{(i)}}{\sqrt{\lambda_i}}}$$

Per quanto riguarda i vettori restanti $\hat{v}_{r+1}, \dots, v_m$ basta calcolare una *base ortonormale* per l'autovalore nullo rispetto a AA^T. Notiamo che potrebbe essere necessario applicare ulteriori procedure di ortogonalizzazione, tra cui la *procedura di Gram Schmidt*. Definendo $V[:, i] := \hat{v}_i$, concludiamo. ■

Notiamo che non è garantita l'unicità del teorema. Infatti il calcolo delle basi del nucleo non restituisce risultati unici.

#Dimostrazione

DIMOSTRAZIONE del Teorema 1

NOTA: APPROFONDIMENTO SVOLTO IN VISTA PER L'ORALE

La dimostrazione si articolerà in tre fasi: prima la costruzione delle matrici U, V, Σ, poi un paio di osservazioni che vadano a legare U, V, Σ, A e infine dimostrare l'uguaglianza della tesi (i.e. $A = V\Sigma U^T$).

Step 1: Costruzione

U: Per il teorema spettrale esiste una matrice che diagonalizzi A^TA (essendo simmetrica e semidefinita positiva), dunque

$$A^TA = UDU^T$$

Ossia

$$U^TA^TAU = D$$

dove D è la matrice diagonale degli autovalori di A^TA, di dimensione $n \times n$. Inoltre si ha che

$$D = \begin{pmatrix} \Sigma_r^2 & 0 \\ 0 & 0 \end{pmatrix}$$

Per $r = \text{rank}\, A \leq \min\{m, n\}$. Definiamo inoltre $U = (U_r \quad W)$ dove W è la matrice delle colonne-vettori del nucleo id A^TA

Σ: Come già discusso prima, definiamo Σ_r come matrice dei valori singolari, i.e. $\Sigma_r = \texttt{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n})$.

V: Analogamente, $V^TAA^TV = D'$ dove D invece è di dimensione $m \times m$. Come prima, *"separiamo"* V in $V = (V_r \quad Z)$, dove le colonne di Z formano il nucleo di AA^T.

Step 2: Relazioni

Dinnanzi vediamo che $AW = 0$. Infatti si ha, per l'equazione (1):

$$U^TA^TAU = \begin{pmatrix} U_r^T \\ W^T \end{pmatrix} (A^TA) \begin{pmatrix} U_r & W \end{pmatrix} = \begin{pmatrix} U_r^TA^TAU_r & U_r^TA^TAW \\ W^TA^TAU_r & W^TA^TAW \end{pmatrix} = \begin{pmatrix} \Sigma_r^2 & 0 \\ 0 & 0 \end{pmatrix}$$

Dunque $W^TA^TAW = (AW)^TAW = 0$, da cui segue che la diagonale di AW è sicuramente nulla da cui $AW = 0$ (per esercizio si dimostra che per B con diagonale nulla si ha che $B^TB = 0$).

Inoltre, per il Lemma 1, si ha che

$$\forall i \leq r, V^{(i)} = \frac{1}{\sigma_i} AU^{(i)} \implies V^{(i)}\sigma_i = AU^{(i)}$$

Quindi facendo un paio di conti ottengo $AU_r = V\Sigma_r$.

Step 3: Uguaglianza
Adesso verifichiamo che $A = V\Sigma U^T$. Prendiamo nota che dimostrare quest'ultima identità equivale a dimostrare l'identità $V^T AU = \Sigma$. Facciamo i conti:

$$\begin{aligned} V^T AU &= \begin{pmatrix} V_r^T \\ Z^T \end{pmatrix} A \begin{pmatrix} U_r & W \end{pmatrix} \\ &= \begin{pmatrix} V_r^T AU_r & V_r^T AW \\ Z^T AU_r & Z^T AW \end{pmatrix} \end{aligned}$$

Per le osservazioni effettuate nello *Step 2*, notiamo che:

- $Z^T AW, V_r^T AW = 0$
- $Z^T AU_r = Z^T V_r \Sigma_r$. Essendo le colonne tra Z, V_r linearmente indipendenti, si ha che $Z^T V_r = 0$ e quindi annullando il prodotto.
- $V_r^T AU_r = V_r^T V_r^T \Sigma_r = \Sigma_r$, essendo le colonne di V tutte ortogonali tra di loro

Quindi concludiamo che

$$V^T AU = \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} = \Sigma$$

Come volevasi dimostrare. ■

X

---

3. Applicazioni delle SVD

La risoluzione dei *sistemi lineari sovradeterminati* può fare anche l'uso della SVD.

Sia dato il sistema lineare $Ax = b$, e data *una* SVD di $A = V\Sigma U^T$. Allora abbiamo

$$(V\Sigma U^T)x = b$$

L'idea sarebbe quella di *"invertire"* il termine $(V\Sigma U^T)$, premoltiplicarla e riportandoci ad una forma chiusa per x. Tuttavia, la matrice non è *quadrata*! Come facciamo?

Per questo defineremo la *pseudoinversa* di una matrice di Moore-Penrose:

#Definizione

> Definizione (pseudoinversa di M-P).
>
> Data $A \in \mathbb{R}^{m \times n}$, definiamo la sua *pseudoinversa di Moore-Penrose* come
>
> $$A^+ := (A^T A)^{-1} A^T$$

Notiamo che nel nostro caso

$$(V\Sigma U^T)^+ = (U\Sigma^+ V^T)$$

In particolare avremmo che $\Sigma^+ \in \mathbb{R}^{n\times m}$ dove

$$\Sigma^+ = \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$

Ovvero prendiamo i reciproci della diagonale e *"invertiamo le posizioni dei zeri"*. In questo modo otteniamo

$$x = (U\Sigma^+ V^T)b$$

Concludendo. ■

Dedurre Informazioni dalla SVD

Dedurre Informazioni sulle Matrici dalla SVD

X

Dedurre informazioni sulle matrici dalla SVD: rango, spazio dell'immagine e nucleo, norma euclidea e di Frobenius.

X

---

0. Voci correlate

- Decomposizione ai Valori Singolari
- Definizione di Nucleo e immagine
- Norme Matriciali

1. Rango, Immagine e Nucleo dalla SVD

OSSERVAZIONE. Data $A \in \mathbb{R}^{m\times n}$ e $V\Sigma U^T$ una sua *SVD*, possiamo ricavare facilmente il *rango* $\text{rg}(A)$. Infatti, essa è determinata dal parametro r per cui dal valore singolare σ_{r+1}-esimo sono tutti zeri.

Q. Possiamo ricavare invece ulteriori informazioni? In particolare, sappiamo già che la *dimensione* dell'immagine è proprio il rango $r = \text{rg}(A)$. Come conosciamo i *vettori* che compongono le basi per $\text{im}\, A$ e $\ker A$?

Osserviamo che

$$A = V\Sigma U^T \iff AU = V\Sigma$$

In quanto $U^TU = \mathbb{1}$; postmoltiplicando otteniamo il risultato voluto. Adesso *l'RHS dell'uguaglianza* (1) colonna per colonna:

$$AU = V\Sigma$$
$$A\begin{pmatrix} U^{(1)} & \dots & U^{(r)} \,\|\, U^{(r+1)} & \dots & U^{(n)} \end{pmatrix} = \begin{pmatrix} V^{(1)} & \dots & V^{(r)} \,\|\, V^{(r+1)} & \dots & V^{(m)} \end{pmatrix}\Sigma$$

Siccome Σ è una *matrice diagonale a blocchi*, abbiamo che $V^{(i)}\Sigma = \sigma_i V^{(i)}$. Allora otteniamo

$$\begin{cases} AU^{(1)} = \sigma_1 V^{(1)} \\ \quad \dots = \dots \\ AU^{(r)} = \sigma_r V^{(r)} \end{cases}$$

Possiamo omettere le equazioni da $r+1$ in poi in quanto sono tutti zero! Infatti per definizione $U^{(k>r)}$ sono *autovalori di* 0.

Allora dalle equazioni (2) otteniamo immediatamente che:

- $U^{(r+1)}, \dots, U^{(n)} \in \ker A$. Infatti $AU^{k>r} = \sigma_{k>r} V^{(k>r)} = 0$.
- Dividendo per le equazioni per lo scalare σ_i otteniamo $\left(A\left(\frac{1}{\sigma_i}U^{(i)}\right) = V^{(i)}\right)_{i=1,\dots,r}$; quindi certamente $V^{(1)}, \dots, V^{(r)} \in \operatorname{im} A$.

Inoltre i vettori di cui visti sopra formano pure *una base* per questi spazi; infatti per il *teorema delle dimensioni (Teorema 1) ho che*

$$\dim A = \dim \ker A + \dim \operatorname{im} A \implies n = r + n - r$$

X

---

2. Norme della Matrice

Andiamo a dimostrare una delle *proprietà* sui prodotti matriciali, in particolare la *regola pratica* per calcolare la norma due. Tuttavia enunciamo (senza dimostrare) prima una proprietà sulla *norma indotta* $p = 2$ e *norma di Frobenius*.

#Lemma

> Lemma (invarianza ortogonale delle norme matriciali).
>
> La norma matriciale euclidea e di Frobenius sono *ortogonalmente invarianti*, ossia date delle matrici ortogonali Q, R si ha che
>
> $$\|RAQ\|_{\{2,F\}} = \|A\|_{\{2,F\}}$$

Si può pensare questo lemma come un'*estensione* della conservazione del prodotto scalare sulle matrici ortogonali (Proposizione 2).

> Proposizione 15 (regola pratica del calcolo della norma $p = 2$).
>
> Sia $A \in \mathbb{R}^{m \times n}$. Allora la norma $p = 2$ è calcolabile come la radice quadrata del raggio spettrale di AA^T o A^TA:

$$\|A\|_2 = \sqrt{\rho(A^TA)} = \sqrt{\rho(AA^T)}$$

#Dimostrazione

DIMOSTRAZIONE della Proposizione 15.

Sia $A = V\Sigma U^T \in \mathbb{R}^{m\times n}$ *una* SVD. Allora per *invarianza ortogonale* ho che

$$\|V\Sigma U^T\|_2 = \|\Sigma\|_2$$

Calcoliamo la norma della matrice diagonale a blocchi dei valori singolari. Per definizione ho che

$$\|\Sigma\|_2 := \max_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|\Sigma x\|_2}{\|x\|_2}$$

Notiamo che il prodotto Σx non è altro che *"una specie di prodotto scalare"* tra i valori singolari σ_i e le componenti x_i (tutte le altre componenti sono nulle; in un certo senso, stiamo estendendo il vettore x da $\mathbb{R}^n$ in $\mathbb{R}^m$). Allora

$$\begin{aligned}
\|\Sigma x\|_2^2 &= \sum_{i=0}^{\min\{m,n\}} (\sigma_i x_i)^2 \\
&\leq \sum_{i=0}^{\min\{m,n\}} (\sigma_1 x_i)^2 \\
&= \sigma_1^2 \sum_{i=0}^{\min\{m,n\}} x_i^2 \\
&\leq \sigma_1^2 \sum_{i=1}^{n} x_i^2 \\
&\leq (\sigma_1 \|x\|_2)^2
\end{aligned}$$

(*FACCIAMO ATTENZIONE A QUESTI PASSAGGI, IN SPECIE DOVE MAGGIORIAMO!*)

Pertanto sostituendo in (3) ottengo

$$\|\Sigma\|_2 \leq \sigma_1 = \sqrt{\lambda_{\max}(A^TA)} = \sqrt{\lambda_{\max}(AA^T)}$$

Abbiamo concluso? No! Infatti, bisogna mostrare che vale *l'uguaglianza* invece della maggiorazione. Lo si fa considerando un *vettore* per cui la definizione di norma raggiunge effettivamente il suo limite superiore, in particolare sceglieremo il primo vettore della base canonica:

$$\|\Sigma e_1\|_2 = \frac{\|\Sigma e_1\|_2}{\|e_1\|_2}$$

- Notiamo che Σe_1 non è altro che un vettore che contiene solo σ_1 e sono tutti zeri altrove
- La norma di un elemento della base canonica è 1

Pertanto otteniamo che $\|\Sigma e_1\| = \sigma_1$, concludendo. ■

Possiamo fare dei conti analoghi per la norma di Frobenius, infatti otteniamo

$$\|A\|_F = \sqrt{\sum_{i=0}^{n} \sigma_r^2}$$

Notiamo che al posto di inserire r sul limite superiore della sommatoria, inseriamo n; lo facciamo per includere più matrici, infatti se fosse r non avremmo incluso matrici con tutti autovalori nulli (in quanto σ_1 non sarebbe neanche definito).

Varianti della SVD

Varianti della SVD

X

Varianti della SVD: SVD leggera, SVD troncata. Motivazioni, definizioni. Teorema di Eckert-Young.

X

0. Voci correlate

- Decomposizione ai Valori Singolari

1. SVD Leggera

Sia $A \in \mathbb{R}^{m \times n}$ e sia data *una* SVD come $AU = \Sigma V$.

OSSERVAZIONE. Osseriamo che nel *prodotto* ΣV, da una certa riga di Σ in poi abbiamo sempre zeri; quindi le righe corrispondenti del prodotto sono anche zeri, quindi in un certo senso *"inutili"*. Un discrso analogo si verifica per AU, in particolare per il fatto che U può contenere degli *autovettori* per l'autovalore nullo, ossia il *nucleo* di A.

IDEA. A $\longrightarrow$ SVD $\longrightarrow$ SVD leggera; in particolare andrò a risparmiare $n - r$ righe.

#Teorema

> Teorema (SVD leggera).
>
> Sia $A \in \mathbb{R}^{m \times n}$ con *una* SVD data da $A = V\Sigma U^T$. Denotando V_r, U_r come la matrice delle *colonne di* V, U che siano degli *autovettori* per autovalori non-nulli, abbiamo che
>
> $$A = V_r \Sigma_r U_r^T$$

#Dimostrazione

DIMOSTRAZIONE del Teorema 1

Denotiamo Z, W rispettivamente i *nuclei* di AA^T e A^TA. Quindi abbiamo che

$$V = (V_r \mid Z), U = (U_r \mid W)$$

Quindi calcolando $V\Sigma U^T$ otteniamo

$$\begin{aligned} V\Sigma U^T &= \begin{pmatrix} V_r & Z \end{pmatrix} \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} U_r^T \\ W^T \end{pmatrix} \\ &= \begin{pmatrix} V_r\Sigma_r & 0 \end{pmatrix} \begin{pmatrix} U_r^T \\ W^T \end{pmatrix} \\ &= V_r\Sigma_r U_r^T \end{aligned}$$

Concludendo. ■

(*Nota*: Questa non è più una *SVD*, in quanto non consideriamo più gli autovettori dell'autovalore nullo)

X

2. SVD Troncata

#Osservazione

OSSERVAZIONE. Data una *SVD* leggera $A = V_r\Sigma_r U_r^T$, notiamo subito che questa non è altro che il prodotto

$$V_r\Sigma_r U_r^T = \begin{pmatrix} V^{(1)} & \dots & V^{(r)} \end{pmatrix} \begin{pmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_r \end{pmatrix} \begin{pmatrix} U^{(1),T} \\ \vdots \\ U^{(r),T} \end{pmatrix}$$

Notiamo che premoltiplicando per Σ_r otteniamo una specie di prodotto scalare, i.e.

$$\begin{pmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_r \end{pmatrix} \begin{pmatrix} U^{(1),T} \\ \vdots \\ U^{(r),T} \end{pmatrix} = \begin{pmatrix} \sigma_1 U^{(1),T} \\ \vdots \\ \sigma_r U^{(r),T} \end{pmatrix}$$

Sostituendo in (1) otteniamo

$$V_r\Sigma_r U_r^T = \begin{pmatrix} V^{(1)} & \dots & V^{(r)} \end{pmatrix} \begin{pmatrix} \sigma_1 U^{(1),T} \\ \vdots \\ \sigma_r U^{(r),T} \end{pmatrix} = \sum_{i \leq r} \sigma_i V^{(i)} U^{(i),T}$$

Notiamo che $V^{(i)}U^{(i),T}$ ritorna una *matrice* in $\mathbb{R}^{m \times n}$. Qual è il rango di questa matrice?

- Naturalmente è $\neq 0$, infatti questa matrice è composta da *autovettori*, quindi per definizione $Ax = \lambda x$ ovvero è non-nulla

- Il *rango* preserva il prodotto tra moltiplicazioni di matrici, i.e. è sicuramente $\leq \min\{\dim V^{(i)}, \dim U^{(i)}\}$; tuttavia sono *vettori*, quindi è ≤ 1.

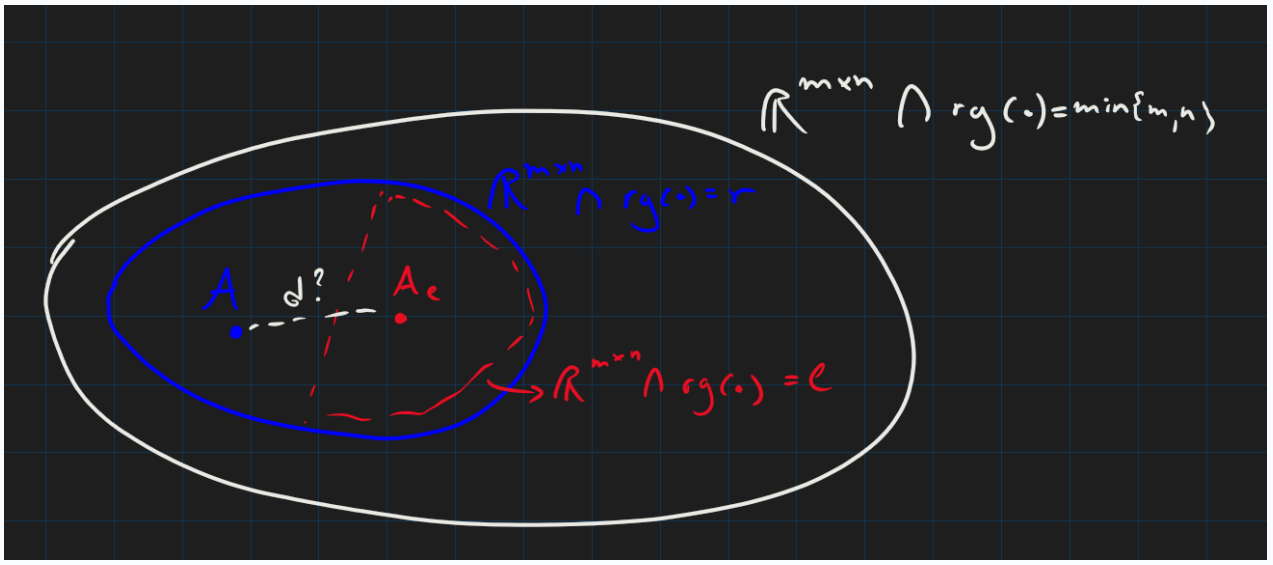
Interpretiamo questo risultato come *"A è un contributo di r matrici con rango uno"*.

Q. Cosa succede se, invece di considerare r somme, ne considero $l < r$?

In questo caso ho la definizione di una *SVD troncata* al rango l-esimo:

$$A_l := \sum_{i=0}^{l<r} \sigma_i V^{(i)} U^{(i),T}$$

L'intuizione consiste nel vedere A_l come una *"approssimazione low-rank"* della matrice A. Come sarà l'errore commesso nell'approssimazione? Vedremo che, in termini quantitativi, è la miglior possibile!



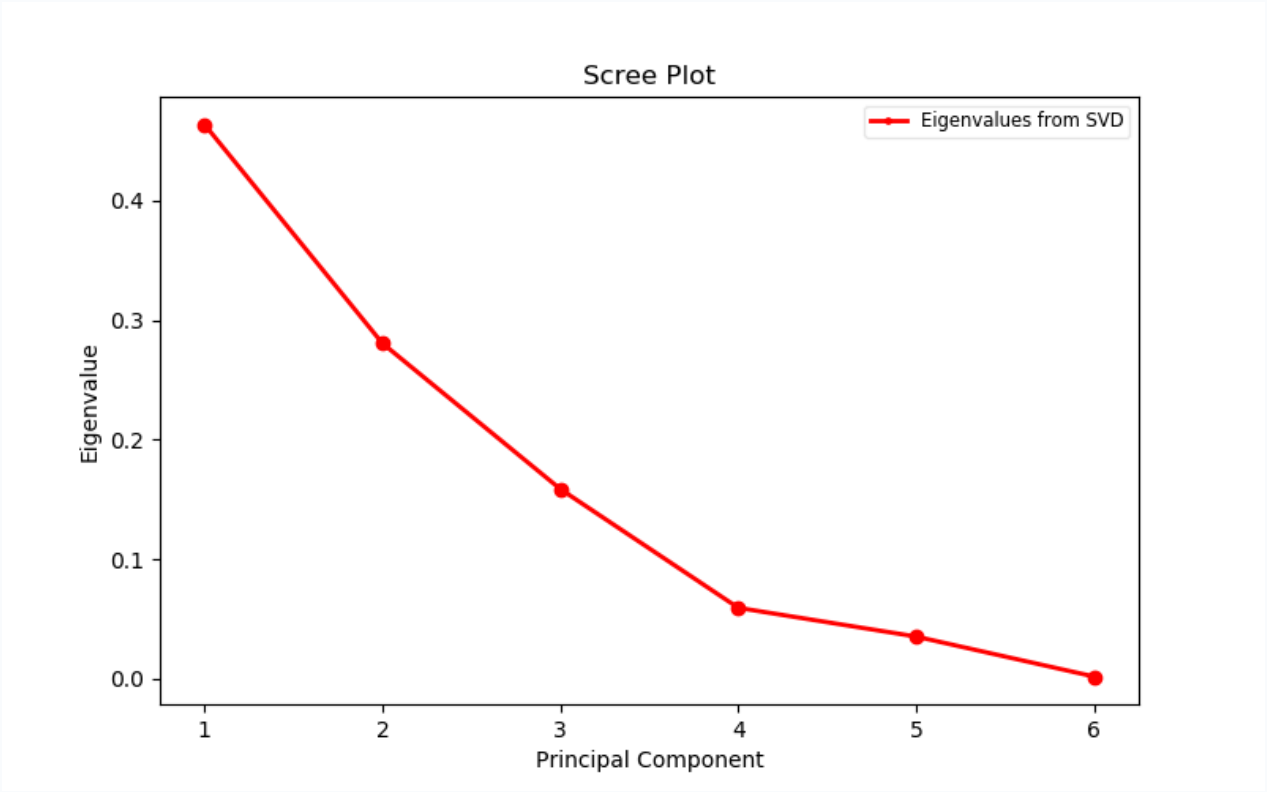
Mostriamo, senza dimostrare, il seguente risultato.

#Teorema

> **Teorema (di Young-Ecker).**
>
> La matrice di rango k che approssima una matrice A è la *SVD troncata al rango l-esimo*, in termini di *norma euclidea e/o di Frobenius*
>
> $$\arg \min_{\{B\in\mathbb{R}^{m\times n}:\ \mathrm{rg}\, B=k\}} \|A-B\|_{\{2,F\}} = A_k$$
>
> Inoltre possiamo quantificare l'errore commesso con i valori singolari della SVD:
>
> $$\|A-A_k\|_2 = \sigma_{k+1}$$
> $$\|A-A_k\|_F = \sqrt{\sum_{i<k\leq r} \sigma_i^2}$$

#Osservazione

OSSERVAZIONE. Osserviamo, su basi empiriche, che i valori singolari σ_i tendono a *"decadere velocemente"* da un certo indice in poi. Ovvero, per il teorema appena enunciato, ad un certo punto l'errore commesso diventa abbstanza per cui conviene *"scegliere"* solo quel valore (in quanto prima si commetterebbe un'errore troppo grande, o dopo andiamo a *"preservare"* troppe informazioni per salvare poco errore). Questo è noto come la *tecnica del gomito* (Scree plot)



X

"Integrazione Numerica"

X

Problema della Quadratura Numerica

Problema della Quadratura Numerica

X

Problema della quadratura (integrazione) numerica. Lemma: stabilità dell'integrazione, corollario.

X

0. Voci correlate

- Integrabilità secondo Riemann
- Tipologie di Funzioni Integrabili

1. Problema della Quadratura Numerica

PROBLEMA. Sia $f \in \mathcal{C}([a,b])$, ovvero sicuramente *Riemann-integrabile* in $[a,b] \Subset \mathbb{R}$ (Teorema 2). Definiamo l'integrale di f su $[a,b]$ come segue:

$$I(f) := \int_a^b f$$

Fissato $n > 0$, definiamo una *formula di quadratura* come una forma che *"sostituisce"* $I(f)$, passando dall'integrale ad una somma di Riemann:

$$I_n(f) := \sum_{k \leq n} w_k f(x_k)$$

Dove:

- $(w_i)_i$ sono i *"pesi"* della formula da *"imparare"*
- $(f(x_i))_i$ sono i *nodi* della formula
- $(x_i)_i$ sono i *nodi di quadratura*
- n è il *numero di nodi di quadratura*

Notiamo che, denotando $w = (w_1, \dots, w_n)$ e $f(x) = (f(x_1), \dots, f(x_n))$ abbiamo che la formula di quadratura non è altro che il seguente prodotto scalare:

$$I_n(f) = \langle w, f(x) \rangle$$

Q. Perché?

La motivazione di approssimare I con una somma discreta I_n potrebbe essere data da alcuni aspetti della funzione f, tra cui:

- Il fatto che f potrebbe non essere integrabile in termini di funzioni elementari (anzi, lo spazio delle funzioni che sono espressi medianti funzioni elementari hanno misura nulla!)
- f potrebbe essere non analitica, oppure sconosciuta o *"troppo difficile"* da calcolare

---

X

---

2. Lemma della Stabilità

Vediamo un lemma che ci garantisce che il *problema della quadratura numerica è ben posta*.

#Lemma

> Lemma (stabilità dell'integrazione).

L'operazione funzionale di integrazione nel continuo è stabile, ovvero se $\tilde{f}$ approssima una funzione $f \in \mathcal{C}[a,b]$ con $[a,b] \Subset \mathbb{R}$ si ha la seguente maggiorazione:

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

#Dimostrazione

DIMOSTRAZIONE del Lemma 1

Banale, si tratta di usare il fatto che $|\int \bullet| \leq \int |\bullet|$. Abbiamo dunque

$$\begin{aligned}\int_a^b f - \int_a^b \tilde{f} \ &= \int_a^b (f - \tilde{f}) \\ &\leq \int_a^b |f - \tilde{f}| \\ &\leq \max_{x\in[a,b]} \int_a^b |f(x) - \tilde{f}(x)| \, \mathrm{d}x \\ &= \max_{x\in[a,b]} |f(x) - \tilde{f}(x)| \underbrace{\int_a^b 1}_{\mu_{\mathcal{R}}([a,b])=b-a}\end{aligned}$$

Concludendo. ■

Dunque se $\tilde{f}$ è *"vicina"* a f, la differenza tra $I(f)$ e $I(\tilde{f})$ non possono essere toppo *"grandi"*. Questo risultato si applica al seguente risultato:

#Corollario

Corollario.

Sia $(f_n)_n$ una successione di funzioni che converge uniformemente a f (ovvero f_n converge a f nel senso della norma infinito; $\|f - f_n\|_\infty \to 0$) dove $f \in \mathcal{C}[a,b]$ con $[a,b] \Subset \mathbb{R}$. Allora si ottiene il seguente limite:

$$\lim_n |I(f_n) - I(f)| = 0$$

#Dimostrazione

DIMOSTRAZIONE del Corollario 2

Ancora più banale, basta applicare il Lemma 1 per ogni f_n per ottenere una maggiorazione e utilizzare dunque il teorema dei due carabinieri. ■

Quadratura Interpolatoria

Quadratura Interpolatoria

Formula di quadratura interpolatoria. Idea di base, definizione. Grado di precisione per una formula di quadratura.

X

---

0. Voci correlate

- Problema della Quadratura Numerica
- Interpolazione Polinomiale
- Polinomi di Lagrange

1. Idea di Base della Quadratura Interpolatoria

IDEA. Le formule di *quadratura interpolinomiale* consiste nel prendere l'approssimante $\tilde{f}$ come esattamente l'interpolante p_n ed integrare esattamente quest'ultimo.

Questa idea è ben posta per l'unicità dell'interpolante polinomiale.

Ricordando i polinomi di Lagrange:

> Definizione 1 (polinomi di Lagrange).
>
> Sia $(x_n, y_n)_{n\leq N}$ una serie di dati. Supponiamo che $\forall i \neq j, x_i \neq x_j$. Allora definiamo il *polinomio elementare di Lagrange per il dato i-esimo* come
>
> $$L_i(x) := \prod_{j\neq i}^{N} \frac{x - x_j}{x_i - x_j} \in \mathbb{P}_N$$

Abbiamo che

> Proposizione 2 (i polinomi di Lagrange sono delle delta di Kronecker sui dati).
>
> Siano $(L_i)_i$ dei polinomi di Lagrange sui dati $(x_n)_{n\leq N}$. Vale per $\forall i, j \leq N$ la seguente identità:
>
> $$L_j(x_i) = \delta_{ij} \ (1 \texttt{ if } i = j \texttt{ else } 0)$$

E dunque

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

Allora si ha che

$$\int_a^b p_n(x) = \int_a^b \sum_{k=0}^{n} f(x_k) L_k(x)\,\mathrm{d}x = \sum_{k=0}^{n} \int_a^b f(x_k) L_k(x)\,\mathrm{d}x = \sum_{k=0}^{n} \left[f(x_k) \underbrace{\int_a^b L_k(x)\,\mathrm{d}x}_{w_k} \right]$$

Ovvero una forma del tipo $\sum_k f(x_k) w_k$, che è proprio una formula di quadratura. ■

Inoltre, una quadratura interpolatoria si dice:

- *Chiusa* se i nodi comprendono anche gli estremi a, b
- *Aperta* se i nodi *comprendolo* solamente l'aperto $\mathcal{I}^\circ =]a, b[$.

X

2. Grado di Precisione di una Formula di Quadratura

#Definizione

> Definizione (grado di precisione per una formula di quadratura).
>
> Data una formula di quadratura generica
>
> $$I(f_M) = \sum_{i=1}^{M} w_i f(x_i)$$
>
> Si dice che ha *grado di precisione ALMENO* n sse è *esatta* per tutti i polinomi f di grado $\leq$ a n. Ovvero, $\forall f \in \mathbb{P}_{k \leq n}$ si ha che $I(f) = I(f_M)$
>
> Si dice che ha invece *grado di precisione ESATTAMENTE* n sse ha grado di precisione almeno n ed $\exists f \in \mathbb{P}_{n+1}$ tale che $I(f) \neq I(f_M)$.

#Teorema

> Teorema.
>
> Una formula a n nodi è *interpolatoria* sse ha grado di precisione *ALMENO* $n - 1$

Dimostrazione omessa, ma intuitiva. Infatti basta pensare che abbiamo un polinomio di grado $n - 1$ e dunque certamente può *"replicare"* un qualsiasi polinomio. ■

Formula del Rettangolo e del Punto Medio

Formula del Rettangolo e del Punto Medio

X

Primo esempio di quadratura interpolatoria: formula del rettangolo $N = 1$. Variante: formula del punto medio. Calcolo dell'errore commesso, grado delle quadrature interpolatorie.

---

0. Voci correlate

- Quadratura Interpolatoria

1. Formula del Rettangolo

#Definizione

> Definizione (formula del rettangolo).
>
> Sia $f \in \mathcal{C}[a,b]$, con $-\infty < a < b + \infty$ (ovvero $[a,b] \neq \{\xi\}$!), fissato $x_0 \in [a,b]$ otteniamo che $L_0(x) = 1$ (funzione costante) e dunque
>
> $$\int_a^b L_0 = b - a$$
>
> Da questa deduciamo la *regola del rettangolo*:
>
> $$\boxed{\int_a^b f \approx (b-a) f(x_0)}$$

Il grado di precisione della formula del rettangolo è *almeno* 0, infatti è in grado di interpolare una qualsiasi funzione costant.ma non è in grado di interpolare

---

2. Formula del Punto Medio

Per una *scelta più privilegiata* di x_0, si prende il *punto medio* dell'intervallo $[a,b]$, ovvero $\bar{x} := (a+b)/2$. Denotiamo tale scelta come I_0:

$$I_0(f) = (b-a) f\left(\frac{a+b}{2}\right)$$

In effetti, se $f \in \mathcal{C}^2[a,b]$ otteniamo la seguente formula dell'errore:

#Proposizione

> Proposizione (errore del punto medio).
>
> Data $f \in \mathcal{C}^2[a,b]$ abbiamo il seguente errore con la quadratura mediante il punto medio:

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

#Dimostrazione

DIMOSTRAZIONE del Proposizione 2

Si tratta di usare lo sviluppo di *Taylor* di f del secondo ordine centrato in $\bar{x}$.

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

Dove $\xi : x \mapsto \xi \in]a,b[$ è una forma funzionale dipendente da x. Calcolando lo scarto abbiamo dunque

$$\int_a^b f - \int_a^b f(\bar{x})\,\mathrm{d}x = \int_a^b f(x) - f(\bar{x})\,\mathrm{d}x$$

Sostituendo $f(x)$ da (1) in (2) otteniamo

$$\int_a^b f(x) - f(\bar{x})\,\mathrm{d}x = \int_a^b f'(\bar{x})(x-\bar{x}) + \frac{f''(\xi(x))}{2}(x-\bar{x})^2\,\mathrm{d}x$$

Per la linearità dell'operatore integrale

$$'' = f(\bar{x}) \underbrace{\int_a^b (x-\bar{x})\,\mathrm{d}x}_{(A)} + \underbrace{\int_a^b f''(\xi(x))\frac{(x-\bar{x})^2}{2}\,\mathrm{d}x}_{(B)}$$

Calcoliamo entrambi gli integrali:
(A): Disegnando un grafico di $x-\bar{x}$ si evince chiaramente che è una funzione dispari, dunque $(A) = 0$.
(B): Usiamo *il teorema dell'integrale media pesata (o generalizzata)* (Teorema della Media Integrale) con $g(x) = (x-\bar{x})^2/2$. Dunque $\exists \xi \in]a,b[$ tale che

$$(B) = f''(\xi) \int_a^b \frac{(x-\bar{x})^2}{2}\,\mathrm{d}x$$

Calcolando l'integrale deduco che

$$(B) = f''(\xi) \int_{a-\bar{x}}^{b-\bar{x}} \frac{u^2}{2}\,\mathrm{d}u = f''(\xi) \int_{-(b-a)/2}^{(b-a)/2} \frac{u^2}{2}\,\mathrm{d}u = \ldots = f''(\xi)\frac{(b-a)^3}{24}$$

Siccome $(A) = 0$, otteniamo che $E_0 = (B)$ ossia la tesi. ■

Formule di Newton-Cotes

Formule di Newton-Cotes

Distinzione tra formule di quadrature chiuse e aperte. Formule di Newton-Cotes. Esempi di Formule di Newton-Cotes: regola del trapezio, regola di Cavalieri-Simpson. Termine d'errore commesso.

---

0. Voci correlate

- Quadratura Interpolatoria

1. Formule (chiuse) di tipo Newton Cotes

Introdotte da Newton nel 1676 e perfezionate successivamente nel 1722.

#Definizione

> Definizione (formula di Newton-Cotes).
>
> Sia $[a,b] \Subset \mathbb{R}$. Una formula di quadratura interpolatoria
>
> $$I_n(f) = \sum_{i=1}^{n} w_n f(x_n)$$
>
> Si dice di *Newton-Cotes* se e solo se
> i. I nodi sono *equispaziati*, ossia
>
> $$x_i = a + \frac{(i-1)(b-a)}{n-1}, i = 1, \dots, n$$
>
> ii. I pesi sono calcolati con l'integrale del polinomio elementare di Lagrange:
>
> $$w_i = \int_a^b L_i, i = 1 \dots, n$$

Vediamo un paio di esempi.

---

2. Metodo del Trapezio SEMPLICE

#Definizione

> Definizione (regola del trapezio).
>
> La *regola del trapezio* è il caso più semplice di Newton-Cotes, ossia $n = 2$ e dunque per forza $x_1 = a$ e $x_2 = b$. Ovvero,

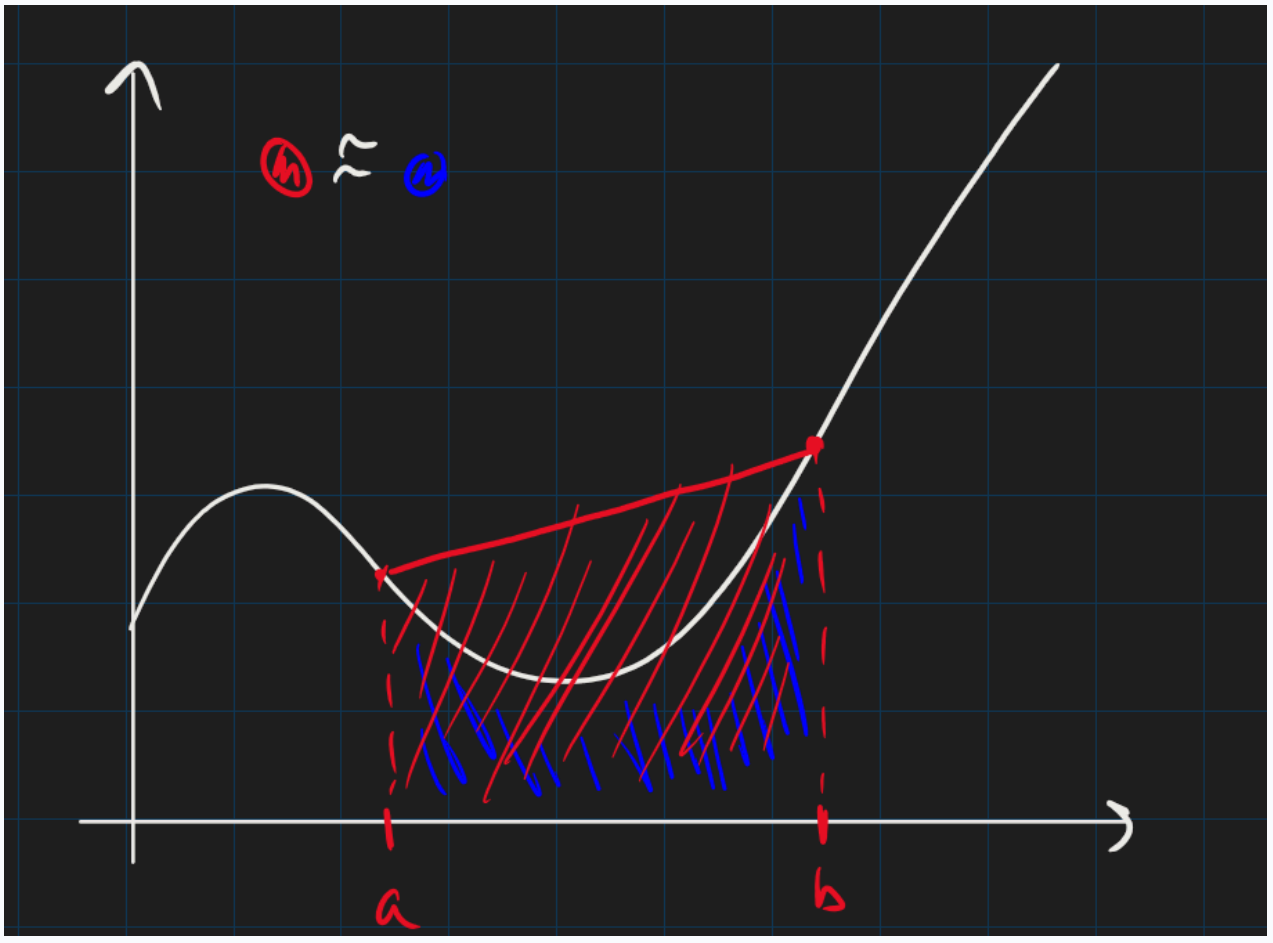
$$I_2(f) =: I_T(f) = \frac{b-a}{2}(f(a)+f(b))$$

In effetti, calcolando:

$$\begin{aligned} w_1 &= \int_a^b L_1(x) \\ &= \int_a^b \frac{b-x}{b-a}\,\mathrm{d}x \\ &= \frac{1}{b-a}\frac{-1}{2}((b-b)^2-(b-a)^2) = \frac{b-a}{2} \end{aligned}$$

Analogamente si ottiene che $w_2 = w_1$, da cui deduciamo la definizione Definizione 2. Geometricamente, calcoliamo l'area sotto la retta creata dai punti $(a, f(a))$ e $(b, f(b))$. In effetti, questo è un trapezio con basi a, $f(a)$ e b, $f(b)$.

Esercizio: si osservi che $w_1 + w_2 = (b-a)$. Dimostrare che per le quadrature interpolatorie si ha sempre che $\sum w = b - a$. Traccia: ricordare i gradi di precisione!!!



#Proposizione

Proposizione (grado di precisione ed errore di formula del trapezio).

Per $f \in \mathcal{C}^2[a,b]$, si dimostra che l'errore compiuto dalla regola del trapezio I_T è data da

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

Pertanto il grado di precisione è *almeno* 1 (infatti la derivata seconda di $a + bx$ è nulla), ed inoltre è anche *esattamente* 1 siccome x^2 non ha derivata nulla.

#Dimostrazione

DIMOSTRAZIONE della Proposizione 3

Analoga alla dimostrazione della regola del punto medio (Proposizione 2), pertanto omessa. ■

Traccia: Per una dimostrazione meno contosa tenere conto del teorema di errore di interpolazione nella forma di Lagrange:

> Teorema 1 (dell'errore di interpolazione polinomiale).
>
> Sia $N > 0$ fissato e $I = [a,b] \Subset \mathbb{R}$. Sia $f \in \mathcal{C}^{N+1}(I)$ e p_N il *polinomio di interpolazione* per $(x_n, f(x_n))_{n \leq N}$ separabile. In particolare assumiamo che $(x_n)_{n \leq N}$ sia *strettamente crescente*, ossia
>
> $$a \leq x_0 < x_1 < \ldots < x_N \leq b$$
>
> Allora per $\forall x \in [a,b], \exists \xi_x \in (a,b)$ t.c. valga l'identità
>
> $$r_N(x) = \frac{f^{(N+1)}(\xi_x)\omega_{N+1}(x)}{(N+1)!}$$
>
> Dove ω_N è il polinomio nodale definito come
>
> $$\omega_{N+1}(x) := \prod_{k \leq N} (x - x_k)$$

X

3. Formula di Cavalieri-Simpson

Sviluppate da Bonaventura Cavalieri (un milanese 🤢, 1635) e riscoperte da Thomas Simpson (un britannico 🤢, 1743)

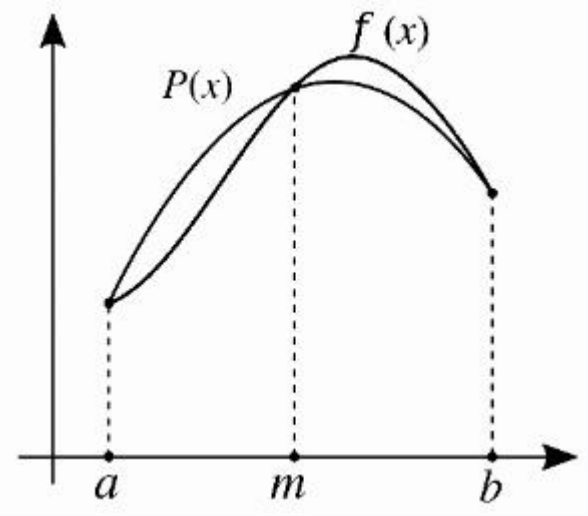
Possiamo *"arricchire"* la formula di Cavalieri-Simpson aggiungendo un *terzo nodo*. Essendo che i nodi devono essere equispaziati, l'unico nodo che possiamo aggiungere è la *media* $\bar{x} = (b+a)/2$.

#Definizione

> Definizione (regola di Cavalieri-Simpson).

La *regola di Cavalieri Simpson* è una formula di Newton-Cotes con $n = 3$, ovvero $x_1 = a, x_2 = \bar{x}, x_3 = b$. Ovvero

$$\boxed{I_{\text{CS}}(f) := \frac{b-a}{6}[f(a) + 4f(\bar{x}) + f(b)]}$$



#Proposizione

Proposizione (errore commesso da Cavalieri-Simpson).

Se $f \in \mathcal{C}^4[a, b]$ allora l'errore commesso da *Cavalieri-Simpson* è dato da

$$\boxed{E_{\text{CS}}(f) := E_2(f) = I(f) - I_{\text{CS}}(f) = -\frac{(b-a)^5}{2880} f^{(4)}(\xi)}$$

Pertanto il grado di precisione è *esattamente 3*, in quanto tutti i polinomi del tipo $a + bx + cx^2 + dx^3$ hanno derivata quarta nulla ma x^4 ha derivata quarta costante.

Formule Composte di Newton-Cotes

Formule Composte di Newton-Cotes

X

Formule di Newton-Cotes composte. Idea. Formula dei trapezi composta, formula di Cavalieri-Simpson composta.

X

0. Voci correlate

- Formule di Newton-Cotes

1. Formule di Newton-Cotes Composte

Q. Nelle *formule di Newton-Cotes* "*semplici*", il loro errore non presenta un termine che dipende da $N > 0$ numero dei nodi di quadratura. Tuttavia, abbiamo visto che per p_n che converge uniformemente a f allora si ha che

$$\lim_n I(f) - I(p_n) = \lim_n I(f) - I_n(f) = 0$$

Come possiamo garantirci che l'errore vada a 0?

L'idea di base è quello di *suddividere* l'intervallo di integrazione $[a, b] \Subset \mathbb{R}$ in una suddivisione di N subintervalli $T_k = [x_{k-1}, x_k)$ ossia t.c. $[a, b] = \sqcup_k T_k$. Dunque abbiamo che

$$\int_{x_0}^{x_N} f = \int_{x_0}^{x_1} f + \ldots + \int_{x_{N-1}}^{x_N} f = \sum_k \int_{T_k} f$$

Con un'abuso di notazione abbiamo denotato l'integrale in $[x_{k-1}, k)$ come il suo integrale generalizzato.

Denotiamo tale quadratura con

$$I_N^{(c)}(f)$$

In questa trattazione abbiamo gli $N > 0$ punti *EQUIDISTANTI*.

X

---

2. Esempi di Formule di Newton-Cotes Composte

FORMULA DEI TRAPEZI COMPOSTA. Non commenteremo la formula (in quanto è facilmente derivabile), tuttavia diciamo che il suo errore è dato da

$$\boxed{E_T^{(c)}(f, N) = -\frac{(b-a)^3}{12N^2} f''(\xi)}, \xi \in \ ]a, b[$$

Il grado di precisione è 1, come per trapezio semplice. Questa formula è particolarmente indicata per integrare funzioni *periodiche* con *derivate periodiche*.

FORMULA DI CAVALIERI SIMPSON COMPOSTA. Analogamente, denotiamo $I_{\text{CS}}^{(c)}(f, N)$ con la quadratura di *Cavalieri-Simpson* composta da N nodi. Il suo errore compiuto è dato da

$$\boxed{E_{\text{CS}}^{(c)}(f, N) = -\frac{(b-a)^5}{2880 \cdot N^4} f^{(\text{iv})}(\xi)}, \xi \in \ ]a, b[$$

Il grado di precisione è dunque 3. Osserviamo che il numero dei nodi è necessariamente dispari, in quanto per ogni suddivisione devo avere un nodo *"interpolatorio"* per calcolare la parabola di interpolazione (in particolare date N suddivisioni avrò $2N + 1$ nodi di quadratura).

X

3. Ordine di Accuratezza

Essendo la precisione delle *formule composte* effettivamente dipendenti dal parametro $N > 0$, possiamo definire la *"velocità"* in cui l'errore va a decadere.

#Definizione

> Definizione (ordine di accuratezza).
>
> Una *formula di quadratura composta* $I_n^{(c)}$ ha *ordine di accuratezza* $p > 0$ se il suo errore compiuto tende a zero per $h := (b-a)/n \longrightarrow 0$ con ordine $O(h^p)$.

Notiamo che:

- Trapezi composta ha ordine 2
- Cavalieri-Simpson ha ordine 4

Empiricamente, se confrontiamo l'approssimazione di un integrale mediante *Cavalieri-Simpson* e *metodo del trapezio* a pari costo computazionale (nodi di quadratura), vedremo che la precisione di *Cavalieri-Simpson* è doppia w.r.t. al metodo dei *trapezi*.

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

Estrapolazione di Richardson

Estrapolazione di Richardson

Osservazione: rapporto tra errori, ipotesi di Richardson. Estrapolazione di Richardson: idea, risultati empirici.

X

---

0. Voci correlate

- Formule Composte di Newton-Cotes

1. Rapporto tra Errori

IDEA. Data una formula di quadratura $I_n(f)$ e il suo errore associato E_n, vogliamo studiare *come* si comporta il rapporto tra E_n e un suo *"successore"*. Fissiamo $k > 0$, vogliamo calcolare

$$(r_n(f))_{N,k} = \frac{(E_n)_N}{(E_n)_{kN}}$$

Ad esempio, per $k = 2$ si ha

$$r_n(f) = \frac{E_n}{E_{2n}}$$

Sostituendo il calcolo dell'errore per I_n otteniamo una *formula chiusa* per $r_n(f)$ siccome i termini si cancellano, e otteniamo (in genere) una costante moltiplicata per un fattore che dipende da ξ_N, ξ_{kN}.

Ad esempio, per la regola dei trapezi si ha

$$r_n(f) = \frac{-(b-a)^3 f''(\xi_n)}{12n^2} \frac{12(2n)^2}{-(b-a^3)f''(\xi_{2n})} = 4\frac{f''(\xi_n)}{f''(\xi_{2n})}$$

IPOTESI. Possiamo, porre come *"ipotesi approssimativa"*, il rapporto $f''(\xi_n)/f''(\xi_{2n}) \simeq 1$ e dunque ottenere $r_n(f) = C > 1$. Intuitivamente, abbiamo una funzione *"sufficientemente liscia"* per cui la sua derivata non contiene dei picchi per cui il rapporto esplode. Diremo questa ipotesi come *ipotesi di Richardson*.

X

---

2. Estrapolazione di Richardson

Conoscendo il *rapporto degli errori* $r_n(f)$, come possiamo ottenere una *approssimazione migliore*?

Sappiamo che $r_n(f) = C > 0$, equivalentemente

$$E_{2n} = C^{-1}E_n$$

Per definizione di errore si ha che

$$I(f) = I_{2n}(f) + E_{2n} = I_n(f) + E_n$$

Sostituendo da (1) e isolando I_n otteniamo che

$$\begin{aligned} & I_{2n}(f) + C^{-1}E_n = I_n(f) + E_n \\ \implies & (1 - (1/C))E_n = I_{2n}(f) - I_n(f) \\ \implies & E_n = \frac{C}{C-1}(I_{2n}(f) - I_n(f)) \end{aligned}$$

Allora sostituendo E_n appena ottenuta in (2), all'interno di $I(f) = I_n(f) + E_n$ otteniamo una *"nuova*" formula per I, che è approssimata in quanto stiamo assumendo l'ipotesi di Richardson.

$$I(f) \simeq I_n(f) + \frac{C}{C-1}(I_{2n}(f) - I_n(f))$$

Facendo un paio di conti otteniamo la forma chiusa

$$I(f) \simeq I_R(f) := \frac{CI_{2n}(f) - I_n(f)}{C-1}$$

Notiamo che non abbiamo dimostrato in nessun modo che l'errore associato a I_R (diremo E_R) sia più piccolo di E_n o E_{kn}. Ricordiamo sempre l'intuizione dell'ipotesi di *Richardson*!