



Corso Luigi Einaudi, 55 - Torino

Appunti universitari

Tesi di laurea

Cartoleria e cancelleria

Stampa file e fotocopie

Print on demand

Rilegature

NUMERO: 1599A -

ANNO: 2015

A P P U N T I

STUDENTE: Aimar

MATERIA: Metodi Numerici e Statistici x l'Ing. + Eserc.
Prof.Falletta-Adami-Gasparini

Il presente lavoro nasce dall'impegno dell'autore ed è distribuito in accordo con il Centro Appunti.

Tutti i diritti sono riservati. È vietata qualsiasi riproduzione, copia totale o parziale, dei contenuti inseriti nel presente volume, ivi inclusa la memorizzazione, rielaborazione, diffusione o distribuzione dei contenuti stessi mediante qualunque supporto magnetico o cartaceo, piattaforma tecnologica o rete telematica, senza previa autorizzazione scritta dell'autore.

ATTENZIONE: QUESTI APPUNTI SONO FATTI DA STUDENTIE NON SONO STATI VISIONATI DAL DOCENTE.
IL NOME DEL PROFESSORE, SERVE SOLO PER IDENTIFICARE IL CORSO.

ARITHMETICA DEL CALCOLATORE

Rappresentazione Floating-point di un numero reale

Per poter parlare di questa rappresentazione, occorre fissare prima una BASE N generica (x noi $N=10$; x il calcolatore $N=2$).

Si definisce allora rappresentazione floating-point di un numero reale a (r.f-p) la seguente scrittura

$$a = (-1)^s p N^q$$

numero reale base

$$\begin{aligned} s &= 0; 1 \\ p &\in \mathbb{R} \\ q &\in \mathbb{Z} \end{aligned}$$

→ questa scrittura è detta r.f-p

Es Posto $N=10$, il numero reale vale

$$a = 0,015 \cdot 10^{-1} =$$

$$\begin{aligned} \rightarrow s &= 0 \text{ (dà il segno del numero reale)} \\ p &= 0,015 \text{ (} p \in \mathbb{R} \text{)} \\ N &= 10 \\ q &= -1 \end{aligned}$$

$$= 0,15 \cdot 10^{-2} =$$

→ muovendo la virgola a sinistra o destra, si hanno altri valori (si moltiplica / divide x 10)

$$= 0,0015 \cdot 10^0$$

Si nota che si hanno 3 r.f-p di uno stesso numero, in cui s e N sono fissi mentre varia la p .

Così esistono infinite r.f-p e il fatto che non sia unica può dare problemi. Pertanto si cerca la rappresentazione, cioè si fissa la r.f-p ponendola come univocamente determinata se si verifica la condizione sul numero reale p

$$N^{-1} \leq p < 1$$

Nell'esempio di prima, l'unica rappresentazione che soddisfa tale condizione è $0,15 \cdot 10^{-2}$ poiché

$$0,1 = 10^{-1} \leq p < 1$$

Più in generale, x $N=10$ occorre che il numero successivo alla virgola sia diversa da zero.

L'unica r.f-p che soddisfa tale condizione è detta rappresentazione floating-point normalizzata del numero reale (r.f-p n.) e corrisponde alla (è unica) r.f-p del numero.

Es Posto $N=10$, si ha

$$a = 0,0013 \cdot 10^{-4} =$$

Si normalizza, cioè si agisce in modo che la mantissa sia minore di 1 e abbia prima cifra decimale non nulla

$$= 1,3 \cdot 10^{-2}$$

$$\rightarrow \begin{aligned} p &= 0,13 \\ q &= 2 \end{aligned}$$

Visto che a un numero reale corrisponde un'unica r.f-p n., allora si può dire che questo numero reale è univocamente determinato dalla terna (che, del resto, definisce la sua r.f-p n.)

→ s = segno di a ($s = 0; 1$)

→ p = mantissa di a ($N^{-1} \leq p < 1$)

Il numero ottenuto in questo caso, in terminologia, non è un NUMERO MACCHINA

Numero macchina: è un numero che rispetta le esigenze del calcolatore, cioè che ha un numero di cifre minore di t e un esponente q compreso tra m e M .
Una volta fissate una base e l'aritmetica del calcolatore (cioè t e il range degli esponenti), si può sapere se un numero è di macchina.

Es $a = 0,0038245 =$
 $= 0,38245 \cdot 10^{-2}$

Si normalizza

→ è un numero macchina

Es $a = 12,29 \cdot 10^{128} =$
 $= 0,1229 \cdot 10^{130}$

→ non è un numero macchina

Esistono però infiniti numeri non rappresentabili e, ovviamente, non possono essere esclusi tutti. Si può però considerare un numero rappresentativo, cioè scegliere il numero macchina che sia il più vicino possibile al numero reale (che non è un numero macchina). Se, infatti, un numero non riesce a soddisfare le proprietà del calcolatore, bisogna rappresentarlo in qualche modo e pertanto si cerca il numero più vicino possibile che sia un numero macchina e si sostituisce questo numero reale.

Dato un generico numero reale a , si può determinare la sua approssimazione di macchina $\bar{a}$, se esiste, in due modi

→ TECNICA DI TRONCAMENTO

→ TECNICA DI ARROTONDAMENTO

Per avere una prima idea, si considera il numero

$$a = 0,1569 \cdot 10^1$$

Si vuole che la mantissa abbia solo 3 cifre e con il troncamento (si tolgono le ultime cifre), si ha

$$\bar{a} = 0,156 \cdot 10^1$$

Invece con l'arrotondamento (si approssima tagliando le ultime cifre e modificando opportunamente quelle rimanenti) si ha

$$\bar{a} = 0,157 \cdot 10^1$$

Ora, dato un numero reale in r.f.p.n.

$$a = (-1)^s p N^q$$

si suppone che il calcolatore fornisca t cifre x rappresentare la mantissa e un esponente compreso tra m e M .

Si suppone innanzitutto che il numero violi la prima proprietà, cioè p HA PIÙ DI t CIFRE e così il numero si esprime come

$$a = (-1)^s (0, a_1 a_2 a_3 \dots a_t a_{t+1} \dots a_n) N^q$$

Per risolvere il problema, esistono due tecniche

→ **tecnica di troncamento:** in essa si cancellano tutte le cifre dopo la t -esima. Di conseguenza, il numero macchina $\bar{a}$ si definisce come

$$\bar{a} = (-1)^s \bar{p} N^q$$

Per $t=2$, invece, si potrebbe avere

$$\bar{a}_1 = 0,10$$

$$\bar{a}_2 = 0,11$$

e la distanza vale

$$d = 0,01 = 10^{-2}$$

Procedendo a induzione, si può dire allora che con t cifre la distanza vale

$$d = N^{-t}$$

In questo intervallo si trovano infiniti numeri non rappresentabili perché non sono numeri macchina e allora si interviene con il troncamento o l'arrotondamento.

Con il TRONCAMENTO, poiché si trascurano le cifre dalla posizione $t+1$ in poi, ogni numero che cade in quell'intervallo viene approssimato con l'estremo di sinistra, cioè il valore CADE SEMPRE SULLO STESSO PUNTO (cioè l'estremo a sinistra).

Con l'ARROTONDAMENTO invece si fa una DISTINZIONE RISPETTO AL PUNTO MEDIO DELL'INTERVALLO. Nel primo sottointervallo, escluso il punto medio, tutti i numeri reali compresi in esso vengono approssimati con l'estremo a sinistra dell'intervallo (cioè $\bar{a}_1$). I numeri presenti nel secondo sottointervallo, invece, sono approssimati con l'estremo a destra (cioè $\bar{a}_2$). Questa tecnica di fatto divide a metà l'intervallo e fa un po' più di bilanciamento di rappresentazione. Si vedrà infatti che questa tecnica produce un errore più piccolo.

Es Per $t=5$, si consideri il numero

$$a = 0,158291 \cdot 10^1$$

Poiché a non è un numero macchina, si procede con

→ troncamento $\bar{a} = 0,15829 \cdot 10^1$

→ arrotondamento $\bar{a} = 0,15829 \cdot 10^1$

Anche il numero $a' = 0,158298 \cdot 10^1$ non è un numero macchina e dunque si procede con

→ troncamento $\bar{a}' = 0,15829 \cdot 10^1$

→ arrotondamento $\bar{a}' = 0,15830 \cdot 10^1$

Finora si è parlato di numeri non rappresentabili nel calcolatore perché la mantissa ha troppe cifre mentre l'esponente era corretto.

Ora si suppone che la mantissa soddisfi l'aritmetica del calcolatore, mentre l'ESPONENTE q NON RIENTRA NEL RANGE $[m; M]$ FISSATO.

In tal caso, si può verificare uno dei seguenti eventi:

→ $q < m$ Si dice che il numero a non è rappresentabile e si ha un problema di **underflow**.

→ $q > M$ Si dice che il numero a non è rappresentabile e si ha un problema di **overflow**.

In questi casi, non si può fare niente e il calcolatore segnala il problema. In particolare, se durante l'esecuzione di un programma si verifica

→ UNDERFLOW, cioè compare un numero troppo piccolo a essere rappresentabile, il computer lo APPROSSIMA A ZERO (lo trascura perché è troppo piccolo) e prosegue nell'esecuzione dell'algoritmo.

→ OVERFLOW, il computer INTERRUPE L'ESECUZIONE DEL PROGRAMMA e dà più problemi.

Dunque occorre evitare il problema di overflow perché non lascia terminare il programma.

Questa nuova espressione è equivalente alla precedente e inoltre può essere calcolata dall'elaboratore, cioè si è trovato un algoritmo che non dà problemi di overflow. Si suppone ora che nello stesso algoritmo si abbia

$$a_2 = 10^{-10}$$

Allora si ha che

$$\begin{aligned} \|v\| &= |a_1| \sqrt{1 + \left(\frac{a_2}{a_1}\right)^2} = \\ &= 10^{100} \cdot \sqrt{1 + (10^{-20})^2} = \end{aligned} \quad \rightarrow \text{compare } 10^{-180} \text{ che dà underflow,} \\ &= 10^{100} \quad \text{ma il calcolatore lo approssima a zero}$$

3 Errori nell'approssimazione:

ragionando in termini di mantissa, si è visto che, dato in generale un numero $a \in \mathbb{F}$, si può ottenere il numero macchina $\bar{a}$ con

→ troncamento

→ arrotondamento

Ovviamente, sostituendo ad a il valore $\bar{a}$ si fa un errore.

Come si misura tale errore?

Si potrebbe sottrarre i due valori e considerare il valore assoluto della quantità ottenuta. Si definisce così l'**errore assoluto** come

$$e_a = |a - \bar{a}|$$

→ è il valore assoluto della differenza tra valore reale e valore macchina

Si introduce anche il concetto di **errore relativo** e_r , definito come

$$e_r = \frac{|a - \bar{a}|}{|a|}$$

→ è l'errore assoluto diviso il modulo del valore reale

Esiste una differenza sostanziale tra questi due concetti. Posto infatti, ad es.,

$$x = 1000$$

se si vuole sostituirla con la quantità $\bar{x}$ ottenuta come

$$\bar{x} = x + 1$$

allora gli errori valgono

→ errore assoluto

$$e_a = |x - \bar{x}| = 1$$

→ errore relativo

$$e_r = \frac{|x - \bar{x}|}{|x|} = \frac{1}{1000} = 10^{-3}$$

Cioè mettendo $\bar{x}$ al posto di x , si ha un errore assoluto pari a 1 e un errore relativo pari a 10^{-3} .

Se ora si ripete la stessa operazione di sostituzione, cioè si fa

$$\bar{\bar{x}} = \bar{x} + 1$$

con $x = 10^{-4}$, si osserva che gli errori valgono

→ errore assoluto

$$e_a = |x - \bar{\bar{x}}| = 1$$

→ errore relativo

$$e_r = \frac{|x - \bar{\bar{x}}|}{|x|} = \frac{1}{10^{-4}} = 10^4$$

$$e_r \begin{cases} < N^{1-t} \\ \leq \frac{1}{2} N^{1-t} \end{cases}$$

$$e_r \begin{cases} < N^{1-t} & \text{troncamento} \\ \leq \frac{1}{2} N^{1-t} & \text{arrotondamento} \end{cases}$$

Si nota innanzitutto che l'ERRORE RELATIVO DIPENDE SOLO DALL'ARITMETICA DEL CALCOLATORE, cioè dalla base scelta N e dal numero di cifre t della mantissa.

Si nota inoltre che l'ERRORE RELATIVO È PIÙ PICCOLO IN CASO DI ARROTONDAMENTO. Dall'altra parte questa operazione è più costosa perché, prese due mantisse successive, occorre capire a quale intervallo la mantissa appartiene (c'è un controllo - è più accurata e più costosa).

In generale, comunque, l'errore relativo è molto piccolo e si dice essere dell'ordine della DICITURA DELLA MACCHINA. Più precisamente, si definisce **precisione di macchina** il valore

$$\text{eps} = \begin{cases} N^{1-t} & \text{troncamento} \\ \frac{1}{2} N^{1-t} & \text{arrotondamento} \end{cases}$$

→ come è definito, dipende solo dall'aritmetica del calcolatore

Così in generale vale la relazione

$$e_r = \frac{|a - \bar{a}|}{|a|} < \text{eps}$$

Se ora si denota con ε il valore

$$\varepsilon = \frac{\bar{a} - a}{a}$$

Si ha che

$$\begin{aligned} \bar{a} &= a + \varepsilon a = \\ &= a(1 + \varepsilon) \\ \bar{a} &= a(1 + \varepsilon) \end{aligned}$$

Così, dato un qualunque numero reale a , la sua r. f. p. ottenuta con troncamento/arrotondamento può essere vista come perturbazione di a , con la condizione che

$$|\varepsilon| \leq \text{eps}$$

cioè la differenza è di una quantità più piccola della precisione di macchina. La quantità eps vale

$$\text{eps} = 10^{-16}$$

ed è molto piccola, quindi quanto si approssima un numero reale con un numero macchina si commette un errore relativo molto piccolo (dell'ordine di 10^{-15}).

Visto che si stanno considerando non operazioni esatte (cioè quelle "classiche") ma operazioni di macchina, ci si chiede se, in generale, soddisfanno ancora le proprietà delle operazioni esatte

→ la PROPRIETÀ COMMUTATIVA vale sempre

→ in generale non valgono le proprietà associativa e distributiva.

Es (proprietà associativa)

Si supponga di lavorare in un sistema che impone $N=10$, $t=4$ e $q \in [-128; 127]$ e che operi solo la tecnica di troncamento.

Dati i numeri reali

$$a = 0,234567 \cdot 10^5$$

$$b = 0,967812 \cdot 10^1$$

$$c = 0,897834 \cdot 10^1$$

si deve verificare che non vale la proprietà associativa della somma intesa come operazione di macchina, cioè che

$$\bar{a} \oplus (\bar{b} \oplus \bar{c}) \neq (\bar{a} \oplus \bar{b}) \oplus \bar{c}$$

Prima si scrive il numero macchina corrispondente

$$\bar{a} = 0,2345 \cdot 10^5$$

$$\bar{b} = 0,9678 \cdot 10^1$$

$$\bar{c} = 0,8978 \cdot 10^1$$

Ora si calcola la somma

$$\bar{b} \oplus \bar{c} = \overline{\bar{b} + \bar{c}}$$

$$\begin{array}{r} \bar{b} + \bar{c} = 0,9678 \cdot 10^1 + \\ \quad 0,8978 \cdot 10^1 = \\ \hline 1,8656 \cdot 10^1 = 0,18656 \cdot 10^2 \end{array} \quad \rightarrow \text{è somma esatta}$$

Si nota che ~~la~~ la somma esatta non è un numero macchina ($5 > 4$ cifre), dunque

$$\bar{b} \oplus \bar{c} = 0,1865 \cdot 10^2$$

Si calcola la seconda parte della somma

$$\bar{a} \oplus (\bar{b} \oplus \bar{c}) = \overline{\bar{a} + (\bar{b} \oplus \bar{c})}$$

$$\begin{array}{r} \bar{a} + (\bar{b} \oplus \bar{c}) = 0,2345 \cdot 10^5 + 0,1865 \cdot 10^2 = 0,2345000 \cdot 10^5 + \\ \quad 0,0001865 \cdot 10^5 = \\ \hline 0,2346865 \cdot 10^5 \end{array} \quad \rightarrow \text{non è numero macchina}$$

Da qui si ha che

$$\bar{a} \oplus (\bar{b} \oplus \bar{c}) = 0,2346 \cdot 10^5$$

si vuole calcolare la differenza tra essi

$$a_1 \ominus a_2$$

Si nota che a_1 è quasi uguale a a_2 perché hanno delle cifre in comune (le prime tre dopo la virgola) e si esegue una sottrazione. Dunque si può parlare di cancellazione numerica perché si stanno sottraendo due numeri simili e poi questi numeri sono affetti da errore perché non sono numeri macchina (l'errore è dell'ordine della precisione di macchina).

Si individuano i numeri i macchina

$$\bar{a}_1 = 0,15782$$

$$\bar{a}_2 = 0,15735$$

→ 3 cifre in comune, mentre le altre sono diverse

Si calcola^{prima} la differenza esatta dei valori

$$\begin{array}{r} a_1 - a_2 = 0,157824831 - \\ \quad \quad \quad 0,157348212 = \\ \hline 0,000476619 = 0,476619 \cdot 10^{-3} \end{array}$$

Si esegue poi l'operazione nel calcolatore

$$\begin{array}{r} \bar{a}_1 \ominus \bar{a}_2 = 0,15782 - \\ \quad \quad \quad 0,15735 = \\ \hline 0,00047 = 0,470000 \cdot 10^{-3} \end{array}$$

Controllando i due risultati, si nota che le uniche cifre in comune dei risultati delle due operazioni sono le prime due, mentre quelle presenti nell'operazione esatta sui numeri reali esatti sono andate perse nell'operazione di macchina perché essa ha approssimato i numeri reali in numeri macchina. Così il risultato è un numero che ha perso cifre rispetto all'operazione giusta.

Questo fenomeno si dice CANCELLAZIONE NUMERICA.

Questo fatto è importante perché si sa che

$$e_{r, a_1} = \frac{|a_1 - \bar{a}_1|}{|a_1|} \leq \frac{1}{2} \cdot 10^{-4} \rightarrow e_r \leq \text{eps}$$

$$e_{r, a_2} = \frac{|a_2 - \bar{a}_2|}{|a_2|} \leq \frac{1}{2} \cdot 10^{-4}$$

→ c'è un errore piccolo sull'approssimazione dei numeri

Ora si valuta l'errore relativo nella differenza, cioè si confronta la quantità $a_1 - a_2$ con $\bar{a}_1 \ominus \bar{a}_2$

$$\frac{|(a_1 - a_2) - (\bar{a}_1 \ominus \bar{a}_2)|}{|a_1 - a_2|} = \frac{|0,476619 \cdot 10^{-3} - 0,47 \cdot 10^{-3}|}{|0,476619 \cdot 10^{-3}|} \approx 0,0139 = 1,39 \cdot 10^{-2}$$

L'errore è dell'ordine di 10^{-2} e dunque non è più piccola di 10^{-4} . Di conseguenza, anche se c'è un piccolo errore nell'approssimazione, con l'operazione l'errore sui dati aumenta.

Così tutte le operazioni aritmetiche di macchina possono in generale determinare errori sempre minori della precisione di macchina (cioè partendo da un numero macchina, con una qualunque operazione non si supera mai il valore eps). Se però i numeri di partenza sono già affetti da errori (cioè si hanno DATI PERTURBATI), allora l'operazione di sottrazione tra numeri quasi uguali può determinare una perdita di precisione, cioè di cifre significative.

Es (ELIMINAZIONE DI CANCELLAZIONE NUMERICA - III)

Si vuole calcolare

$$y = x - \sin x$$

I due elementi sono quasi uguali tra di loro quando $\sin x \sim x$, cioè $x \sim 0$.

Ora eliminare il fenomeno della cancellazione numerica è più difficile perché ci sono x e $\sin x$. Allora si può fare lo sviluppo di $\sin x$ nell'intorno dello zero troncato a un certo punto (perché non si può mettere tutto)

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots \Rightarrow y = \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!} + \dots$$

Si osserva però che, troncando più avanti lo sviluppo, più è preciso il risultato. Supponendo di voler calcolare la quantità y con $x \sim 10^{-3}$, si avrà che

$$y \sim 10^{-9} - 10^{-15} + 10^{-21}$$

Poiché in genere si ha $\text{eps} \sim 10^{-16}$, si può eliminare il termine x^7 perché nell'elaboratore darebbe zero.

Si osserva poi che qui compare una differenza tra numeri vicini allo zero ma non avviene cancellazione numerica perché si hanno potenze diverse di uno stesso numero e dunque non ci sono cifre in comune.

6 Condizionamento di un problema

Si suppone di avere dei dati x e un problema che, a partire da tali dati di input, restituisce dei dati di output y legati secondo la relazione

$$y = f(x)$$

f è il PROBLEMA NUMERICO che da x restituisce y . La relazione può essere scritta come esplicita ($y = f(x)$) o come implicita ($f(x; y) = 0$).

Es Sia dato il problema

$$Ax = b$$

In essa si riconoscono i dati di input A (matrice) e b mentre l'output è x . Da essa si può ricavare la relazione esplicita tra input e output

$$x = A^{-1}b = F(A; b)$$

La relazione implicita invece è

$$Ax - b = F(x; A; b) = 0$$

Quindi si suppone di partire da un dato noto x e, attraverso una legge, si vuole ottenere il dato d'output y . Quest'ultimo si ottiene tramite la relazione

$$x \xrightarrow{F} y$$

Cioè dal dato noto si ottiene la risposta y usando l'aritmetica esatta, con precisione infinita di calcolo.

Ora si suppone di perturbare il dato iniziale x in $\bar{x}$ (potrebbe essere il dato macchina) e si denota con $F(\bar{x})$ il valore d'output ottenuto con l'aritmetica infinita

$$\bar{x} \xrightarrow{F} f(\bar{x})$$

Si ricorda che, x ogni numero reale a , vale la relazione

$$\bar{a} = a(1 + \varepsilon), \quad |\varepsilon| \leq \text{eps}$$

Si può così scrivere che

$$\begin{aligned} \frac{|f(x_1, x_2) - f(\bar{x}_1, \bar{x}_2)|}{|f(x_1, x_2)|} &= \frac{|x_1 + x_2 - x_1(1 + \varepsilon_1) - x_2(1 + \varepsilon_2)|}{|x_1 + x_2|} = \\ &= \frac{|\varepsilon_1 x_1 + \varepsilon_2 x_2|}{|x_1 + x_2|} \leq \underbrace{\frac{|x_1|}{|x_1 + x_2|}}_{k_1} \varepsilon_1 + \underbrace{\frac{|x_2|}{|x_1 + x_2|}}_{k_2} \varepsilon_2 \end{aligned}$$

Ricordandosi che ε_1 e ε_2 sono gli errori dei dati di input, questi sono moltiplicati x due quantità dipendenti solo dai dati di input, cioè i numeri di condizionamento del problema k_1 e k_2 .

Si osserva che, $x_1 \approx -x_2$, queste quantità "esplodono" perché il denominatore tende a zero. Così a sinistra si ha qualcosa che è maggiorato da qualcosa che va all'infinito e il problema è malcondizionato. In questo caso, infatti, i numeri di condizionamento sono grandi. Il caso $x_1 \approx -x_2$ però fa sì che il problema diventi una somma di due numeri simili e opposti e ciò dà luogo a cancellazione numerica

$\Rightarrow$

dunque il fenomeno di CANCELLAZIONE NUMERICA DETERMINA UN PROBLEMA MALCONDIZIONATO. Se, invece, si avesse $x_1 \neq -x_2$, il problema è ben condizionato.

7 Stabilità di un algoritmo:

prima si è visto il concetto di condizionamento, che dipende solo dal problema e non dal tipo di algoritmo scelto.

Ora si vede il concetto di STABILITÀ, che dipende dall'algoritmo usato.

algoritmo: è una successione finita di operazioni aritmetiche e non.

Dato un problema f con input x e risposta y e denotati la perturbazione dell'input con $\bar{x}$ e l'output y^* , tramite il problema f della perturbazione con $f(\bar{x})$, queste operazioni permettano di introdurre un nuovo output y^*

$$\bar{x} \xrightarrow{\text{fin}} y^*$$

ottenuto a partire dal dato perturbato $\bar{x}$ con operazioni finite.

Così ora si hanno 3 possibili risposte

$$\begin{aligned} \rightarrow x &\xrightarrow{\infty} y \\ \rightarrow \bar{x} &\xrightarrow{\infty} f(\bar{x}) \\ \rightarrow \bar{x} &\xrightarrow{\text{fin}} y^* \end{aligned}$$

Per verificare la bontà dell'algoritmo, si confrontano $f(\bar{x})$ e y^* , cioè si confrontano i due output diversi (ottenuti con operazioni finite e non) presi dallo stesso input. Se l'ERRORE RELATIVO DELL'OUTPUT RISPETTO AL DATO ESATTO È DELL'ORDINE DELLA PRECISIONE DI MACCHINA eps, allora si parla di **algoritmo stabile**.

$$\frac{\|f(\bar{x}) - y^*\|}{\|f(\bar{x})\|} \approx \text{eps}$$

Dunque si guarda la risposta dell'algoritmo, che lavora con la matematica finita, e se non amplifica gli errori dovuti alla perturbazione iniziale, allora l'algoritmo è stabile.

RISOLUZIONE DI SISTEMI LINEARI

Premessa

1 **Norma di un vettore**: la norma è fondamentale perché interviene nel calcolo dell'errore sull'output dell'algoritmo.
Dato un vettore di $\mathbb{R}^n$ $\vec{x}$, definito come

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} \in \mathbb{R}^n$$

si definiscono

→ **norma euclidea**

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

→ è la somma dei quadrati delle componenti (in $\mathbb{R}^2$ è l'estensione del teorema di Pitagora) o prodotto del vettore riga x il vettore colonna (x la regola di moltiplicazione delle matrici)

→ **norma all'infinito**: è il massimo delle componenti in valore assoluto

$$\|\vec{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

→ **norma uno**: è la somma delle componenti del vettore $\vec{x}$ prese in valore assoluto

$$\|\vec{x}\|_1 = \sum_{i=1}^n |x_i|$$

Es Dato il vettore

$$\vec{x} = \begin{pmatrix} 1 \\ 0 \\ -3 \end{pmatrix}$$

si ha che

$$\rightarrow \|\vec{x}\|_2 = \sqrt{1^2 + 0^2 + (-3)^2} = \sqrt{10}$$

$$\rightarrow \|\vec{x}\|_\infty = 3$$

$$\rightarrow \|\vec{x}\|_1 = 1 + 0 + 3 = 4$$

La norma misura la "lunghezza" di un vettore (quella euclidea è proprio la lunghezza del vettore stesso) e queste quantità sono dette "norme" perché soddisfanno le proprietà delle norme, come

$$\|\vec{0}\| = 0$$

→ dato un vettore $\vec{x}$, si ha che

$$\|A\vec{x}\| \leq \|A\| \|\vec{x}\|$$

→ le norme una e all'infinito sono uguali in tutte le matrici simmetriche.

3 PARTICOLARI MATRICI

→ una matrice A si dice **triangolare superiore** se i suoi elementi sono tali che

$$a_{ij} = 0 \quad ; \quad i > j$$

$$A = \begin{bmatrix} \circ & \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ \end{bmatrix}$$

Cioè presenta zeri sotto la diagonale principale.

→ una matrice A si dice **triangolare inferiore** se ha elementi tali che

$$a_{ij} = 0 \quad , \quad i < j$$

$$A = \begin{bmatrix} \circ & \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ \end{bmatrix}$$

Cioè presenta zeri sopra la diagonale principale.

OSSERVAZIONE: si osserva che, ad es., la matrice A

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

è triangolare inferiore. Infatti ha almeno un elemento non nullo nel triangolo inferiore (bisogna solo controllare che gli zeri siano disposti opportunamente).

→ una matrice A si dice **diagonale** se ha elementi non nulli solo nella diagonale principale

$$a_{ij} = 0 \quad , \quad i \neq j$$

→ una matrice A si dice **tridiagonale** se si trovano solo più elementi nulli dalla codiagonale di ordine 2

$$A = \begin{bmatrix} \circ & \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ \end{bmatrix}$$

$$a_{ij} = 0 \quad , \quad |i-j| > 1$$

↓
è diverso da zero solo sulle 3 diagonali principali

→ una matrice A si dice **diagonale dominante per righe** se vale che

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}|, \quad \forall i=1, \dots, n$$

Cioè il valore assoluto degli elementi sulla diagonale, x ogni riga i , dev'essere maggiore della somma degli elementi su quella riga (eccetto l' i -esimo elemento)

Es. Si consideri la matrice A

$$A = \begin{bmatrix} 4 & 1 & -1 \\ 2 & -7 & 1 \\ 3 & -2 & 4 \end{bmatrix} \quad \begin{array}{l} \sim 4 > 2 \\ \sim 7 > 3 \\ \sim 4 > 5 \end{array} \quad \Rightarrow A \text{ è diagonale dominante x righe}$$

Sistemi lineari

1 Si vuole studiare la risoluzione di sistemi lineari con alto numero di equazioni e di incognite (perché in ingegneria si manipolano molti dati).

2 Si considera un generico **sistema quadrato** (dove il numero di equazioni è pari al numero di incognite)

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2 \\ (\dots) \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n = b_n \end{cases} \rightarrow \text{equazione del sistema lineare}$$

x_i = incognita
 b_i = termine noto dell'equazione
 a_{ij} = coefficiente (notazione con due indici perché esiste una notazione matriciale x i sistemi)

Il sistema può anche essere riscritto in forma matriciale come

$$\begin{bmatrix} a_{11} & a_{12} & (\dots) & a_{1n} \\ a_{21} & a_{22} & (\dots) & a_{2n} \\ (\dots) & & & \\ a_{n1} & a_{n2} & (\dots) & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

$\begin{bmatrix} a_{11} & a_{12} & (\dots) & a_{1n} \\ a_{21} & a_{22} & (\dots) & a_{2n} \\ (\dots) & & & \\ a_{n1} & a_{n2} & (\dots) & a_{nn} \end{bmatrix}$ $\parallel$ A (matrice dei coefficienti)
 $\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ $\parallel$ $\vec{x}$ (vettore colonna delle incognite)
 $\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$ $\parallel$ $\vec{b}$ (vettore colonna dei termini noti)

In forma più compatta si può scrivere come

$$A\vec{x} = \vec{b}$$

E se si consideri il sistema lineare

$$\begin{cases} x - y + z = 0 \\ 2x + 3z = 1 \\ 2x - 2y + z = 2 \end{cases}$$

allora a esso si associano

$$A = \begin{bmatrix} 1 & -1 & 1 \\ 2 & 0 & 3 \\ 2 & -2 & 1 \end{bmatrix} \rightarrow \text{nella prima riga ci sono i coefficienti della prima equazione ordinati secondo l'incognita considerata}$$

$$\vec{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \quad \vec{b} = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}$$

Si è interessati ai sistemi quadrati aventi un'unica soluzione, cioè quelli in cui la matrice A non è singolare. Pertanto, si suppone sempre che

$$\det A \neq 0$$

Così si sta cercando un'unica soluzione che si sa che esiste. La soluzione è quel vettore $\vec{x}$ che soddisfa la relazione

Gli errori relativi sui dati d'input valgono

$$\frac{\|A - \bar{A}\|}{\|A\|} = \frac{\|\delta A\|}{\|A\|} \quad \frac{\|\bar{b} - \tilde{b}\|}{\|\bar{b}\|} = \frac{\|\delta \bar{b}\|}{\|\bar{b}\|}$$

Dunque si vuole confrontare l'errore relativo dell'output con due errori relativi, riferiti all'input.
A tal proposito, si dimostra che, se la perturbazione δA , in qualunque norma considerata, è tale che

$$\|\delta A\| < \frac{1}{\|A^{-1}\|}$$

→ così automaticamente si nota che $A + \delta A = \bar{A}$ non è singolare

allora vale la relazione

$$\frac{\|\delta \tilde{x}\|}{\|\tilde{x}\|} \leq \frac{k(A)}{1 - k(A) \frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta \bar{b}\|}{\|\bar{b}\|} \right)$$

errore relativo dell'output
(si confronta il risultato da risolvere e il risultato con la perturbazione)

dove il parametro $k(A)$ vale

$$k(A) = \|A\| \|A^{-1}\|$$

→ esso è calcolabile perché la matrice A è nota

Se ora si suppone che

$$\|\delta A\| < \frac{1}{2\|A^{-1}\|}$$

cioè la scarto δA è più piccola rispetto all'ipotesi di prima, allora la relazione si semplifica

$$\frac{\|\delta \tilde{x}\|}{\|\tilde{x}\|} \leq 2k(A) \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta \bar{b}\|}{\|\bar{b}\|} \right)$$

Così l'errore relativo nell'output è controllato dal numero $k(A)$ e dalla somma degli errori relativi nell'input.

Ora, si sa che un problema è ben condizionato se a piccoli errori d'input corrispondono piccoli errori d'output. Supponendo che l'errore d'input sia dell'ordine di 10^{-6} , si ha che

→ se il numero $k(A)$ è 10^6 , l'errore con cui si controlla l'output è 2 e dunque il problema è mal condizionato

→ se il numero $k(A)$ è 2, l'errore è dell'ordine di 10^6 e dunque il problema è ben condizionato.

Dunque, per sapere il condizionamento del sistema, occorre studiare il parametro $k(A)$, detto **numero di condizionamento del sistema**, in quanto lo si vuole piccolo, in modo che a errori piccoli nell'input corrispondano errori piccoli nell'output. Si osserva, però, che tale numero non può andare sotto l'uno, cioè

$$k(A) \geq 1, \forall A$$

Dimostrazione

Per definizione, il numero di condizionamento $k(A)$ del sistema vale

$$k(A) = \|A\| \|A^{-1}\|$$

ES (MATRICE MALCONDIZIONATA - II)

Si consideri la **matrice di Vandermonde**. Essa si ottiene partendo da un vettore $\vec{x}$ di n componenti (supposti non ripetuti, cioè $x_i \neq x_j \times i \neq j$)

$$\vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

L'ultima colonna di tale matrice è tutta uno, la penultima è costituita dalle componenti del vettore $\vec{x}$, la terzultima è costituita dalle componenti di tale vettore ciascuna elevata al quadrato e così via, cioè ogni colonna ha le componenti del vettore $\vec{x}$ elevate a una certa potenza (l'ultima riga con esponente 0, la penultima con esponente 1 e così via).

Si ottiene così una matrice

$$V_n = \begin{bmatrix} x_1^{n-1} & x_1^{n-2} & (\dots) & x_1 & 1 \\ x_2^{n-1} & x_2^{n-2} & (\dots) & x_2 & 1 \\ \vdots & \vdots & (\dots) & \vdots & \vdots \\ x_n^{n-1} & x_n^{n-2} & (\dots) & x_n & 1 \end{bmatrix}$$

Si dimostra che anche essa è una matrice malcondizionata.

3 Risoluzione di un sistema lineare:

Per la risoluzione di un generico sistema quadrato si fa ricorso alla **regola di Cramer**

Es Sia dato il sistema

$$\begin{cases} x - y = 0 \\ x + y = 1 \end{cases}$$

Tale sistema, ovviamente, può essere scritto nella forma

$$A\vec{x} = \vec{b}$$

imponendo che

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

$$\vec{b} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Allora, la regola di Cramer dice che la soluzione vale

$$x = \frac{\det(b|A)}{\det A} = \frac{1}{2}$$

$$y = \frac{\det(A|\vec{b})}{\det A} = \frac{1}{2}$$

Cioè si ha che, sostituendo la colonna del termine noto nell' i -esima colonna della matrice dei coefficienti, calcolandone poi il determinante e poi dividendo il tutto x il determinante della matrice dei coefficienti, si ottiene la i -esima soluzione.

$$x_i = \frac{\det(A^1 A^2 \dots A^{i-1} \vec{b} A^{i+1} \dots A^n)}{\det A}$$

A^j = j -esima colonna di A

Il sistema successivo più facile da risolvere è il **sistema triangolare superiore**, in cui la matrice dei coefficienti è triangolare superiore.

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ & a_{22} & \dots & a_{2n} \\ & & \ddots & \vdots \\ & & & a_{nn} \end{bmatrix}$$

Dunque il sistema ha forma

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \dots + a_{1,n-1}x_{n-1} + a_{1n}x_n = b_1 \\ a_{22}x_2 + a_{23}x_3 + \dots + a_{2,n-1}x_{n-1} + a_{2n}x_n = b_2 \\ (-\dots-) \\ a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n = b_{n-1} \\ a_{nn}x_n = b_n \end{cases}$$

Perché il sistema abbia una sola soluzione, è necessario che

$$a_{ii} \neq 0, \forall i = 1, \dots, n \quad \rightarrow \text{così } \det A = a_{11}a_{22}\dots a_{nn} \neq 0$$

Per ricavare la soluzione, si considera l'ultima equazione, da cui si può ottenere direttamente l'ultima soluzione

$$x_n = \frac{b_n}{a_{nn}}$$

Ora si sostituisce il termine x_n nella penultima equazione, lo si porta a termine noto e si ricava la soluzione x_{n-1}

$$x_{n-1} = \frac{b_{n-1} - a_{n-1,n}x_n}{a_{n-1,n-1}}$$

Questo procedimento viene ripetuto fino alla prima equazione, dove si ha che

$$x_1 = \frac{b_1 - \sum_{j=2}^n a_{1j}x_j}{a_{11}} \quad \rightarrow \text{si porta a termine noto tutto ciò che si conosce}$$

Questa tecnica si dice **tecnica di backward substitution** (=sostituzione all'indietro) perché, durante la sostituzione, si va all'indietro e l'algoritmo in pseudolinguaggio si esprime nella forma più compatta come

$$\begin{cases} x_n = \frac{b_n}{a_{nn}} \\ x_i = \frac{b_i - \sum_{j=i+1}^n a_{ij}x_j}{a_{ii}}, \quad i = 1, \dots, n-1 \end{cases} \quad \rightarrow \text{dall'ultima equazione si esplicita } x_n \text{ e poi si va dritto.}$$

Si nota che il **costo computazionale c.c.**, cioè il numero di operazioni aritmetiche (somma algebrica, prodotto...) vale

$$c.c. = \frac{n^2}{2}$$

n = ordine della matrice

2 Metodo di eliminazione di Gauss:

esso è un metodo che consiste nel CAMBIARE IL SISTEMA LINEARE DI PARTENZA IN UN SISTEMA LINEARE TRIANGOLARE SUPERIORE IN UN NUMERO FINITO DI PASSI (di trasformazione), dove il numero di passi è

$$n_{\text{passi}} = n - 1$$

n = ordine della matrice
(matrice $2 \times 2 \Rightarrow$ basta 1 passo x trasformarla in matrice triangolare superiore)

Si ottiene così un **sistema lineare equivalente**, cioè che AMMETTE LA STESSA SOLUZIONE,

$$U\vec{x} = \vec{b}$$

dove però la matrice U è triangolare superiore e ciò consente di usare la tecnica di backward substitution x risolvere il sistema.

L'algoritmo di Gauss si basa su due proprietà fondamentali dei sistemi lineari

→ SCAMBIANDO DUE EQUAZIONI TRA DI LORO, LA SOLUZIONE DEL SISTEMA NON CAMBIA

→ SOSTITUENDO UN'EQUAZIONE DEL SISTEMA CON UNA COMBINAZIONE LINEARE DEL SISTEMA, LA SOLUZIONE DEL SISTEMA NON CAMBIA.

Es. Si consideri il sistema

$$\begin{cases} x + y = 1 \\ x - y = 0 \end{cases} \quad \left(x = \frac{1}{2}; y = \frac{1}{2} \right)$$

Si capisce chiaramente che tale sistema è equivalente al sistema

$$\begin{array}{l} \textcircled{\text{I}} \begin{cases} x - y = 0 \\ x + y = 1 \end{cases} \\ \textcircled{\text{II}} \end{array} \quad \rightarrow \text{si sono scambiate le equazioni}$$

Dall'altra parte, se ora si sommasse la prima equazione alla seconda, la soluzione è la stessa

$$\textcircled{\text{I}} \mapsto \textcircled{\text{I}} + \textcircled{\text{II}} \quad \begin{cases} x - y = 0 \\ 2x = 1 \end{cases} \quad \rightarrow \text{infatti si ha sempre } x = \frac{1}{2}; y = \frac{1}{2}$$

DESCRIZIONE DEL METODO DI GAUSS

Per descrivere tale metodo, si consideri x semplicità una generica matrice 4×4 (il ragionamento poi può essere esteso a un generico ordine n) e dunque si consideri il sistema 4×4

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 = b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 = b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 = b_3 \\ a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + a_{44}x_4 = b_4 \end{cases}$$

Esso può essere scritto in forma più compatta come

$$A\vec{x} = \vec{b}$$

in cui si ha che

Dopo di che si lascia invariata la prima riga della matrice (cioè la prima equazione del sistema) e si sostituisce la II equazione del sistema con

$$\text{II EQUAZIONE} + m_{21} \cdot \text{I EQUAZIONE}$$

La soluzione, infatti, non cambia perché si sta sostituendo un'equazione con una combinazione lineare delle altre equazioni del sistema. Si ha così che

$$\begin{array}{l} \text{II EQUAZIONE} + \\ m_{21} \cdot \text{I EQUAZIONE} \end{array} \quad \begin{array}{l} a_{21}^{(1)} x_1 + a_{22}^{(1)} x_2 + a_{23}^{(1)} x_3 + a_{24}^{(1)} x_4 = b_2^{(1)} \\ m_{21} (a_{11}^{(1)} x_1 + a_{12}^{(1)} x_2 + a_{13}^{(1)} x_3 + a_{14}^{(1)} x_4) = m_{21} b_1^{(1)} \end{array} \quad = \begin{array}{l} + \\ \sim \text{ occorre moltiplicare anche il} \\ \text{termine noto (se no cambia} \\ \text{l'equazione)} \end{array}$$

$$\underbrace{(a_{21}^{(1)} + m_{21} a_{11}^{(1)})}_{=0} x_1 + a_{22}^{(2)} x_2 + a_{23}^{(2)} x_3 + a_{24}^{(2)} x_4 = b_2^{(2)}$$

x comodità, il termine $a_{22}^{(1)} + m_{21} a_{12}^{(1)}$ lo si fa chiamare $a_{22}^{(2)}$ e lo stesso si fa x i termini successivi

Così nella II equazione si è eliminato il coefficiente nella posizione della prima incognita (cioè a_{21}).

Poi si prende la III equazione e la si sostituisce con

$$\text{III EQUAZIONE} + m_{31} \cdot \text{I EQUAZIONE}$$

e, come prima, si elimina il coefficiente a_{31} associato alla prima incognita. Infine, si prende la IV equazione e la si sostituisce con

$$\text{IV EQUAZIONE} + m_{41} \cdot \text{I EQUAZIONE}$$

e di nuovo si elimina il coefficiente a_{41} della prima incognita.

Così si ottiene una nuova matrice, contrassegnata con $A^{(2)}$, avente la stessa prima riga e parte della prima colonna nulla, grazie a come sono stati definiti i moltiplicatori.

$$A^{(2)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & a_{14}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & a_{24}^{(2)} \\ 0 & a_{32}^{(2)} & a_{33}^{(2)} & a_{34}^{(2)} \\ 0 & a_{42}^{(2)} & a_{43}^{(2)} & a_{44}^{(2)} \end{bmatrix}$$

Analogamente, poiché si stanno facendo operazioni anche sul termine noto, si ottiene che

$$b^{(2)} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(2)} \\ b_4^{(2)} \end{pmatrix} \quad \rightarrow \text{ la prima equazione non è stata modificata}$$

PASSO III)

Bisogna focalizzarsi sulla sottomatrice 2×2 (evidenziata in rosso) di $A^{(3)}$ e cercare di annullare l'elemento sotto $a_{33}^{(3)}$.
 Supponendo che $a_{33}^{(3)} \neq 0$ (altrimenti si scambiano le due righe della sottomatrice, in modo tale che al posto di tale elemento ne venga uno sicuramente diverso da zero), si azzerare l'elemento $a_{43}^{(3)}$ si introduce il moltiplicatore

$$m_{43} = -\frac{a_{43}^{(3)}}{a_{33}^{(3)}}$$

A questo punto si sostituisce la quarta equazione

$$\text{IV equazione} \rightarrow \text{IV EQUAZIONE} + m_{43} \cdot \text{III EQUAZIONE}$$

Si ottiene così la matrice

$$A^{(4)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & a_{14}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & a_{24}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & a_{34}^{(3)} \\ 0 & 0 & 0 & a_{44}^{(4)} \end{bmatrix}$$

Dunque si è partiti da un sistema del tipo

$$A^{(1)} \vec{x} = \vec{b}^{(1)}$$

Poi, nel passo I, si sono modificati matrice dei coefficienti e vettore dei termini noti in $A^{(2)}$ e $\vec{b}^{(2)}$ senza però modificare il sistema (sono equivalenti)

$$A^{(2)} \vec{x} = \vec{b}^{(2)}$$

Nel passo II si ottiene il sistema

$$A^{(3)} \vec{x} = \vec{b}^{(3)}$$

Nel passo IV si ottiene infine il sistema

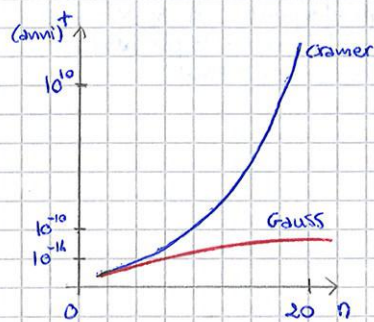
$$A^{(4)} \vec{x} = \vec{b}^{(4)}$$

dove

$$A^{(4)} = U$$

$$\rightarrow 3 \text{ passi } (3 = n - 1 = 4 - 1)$$

Si nota che il metodo di Gauss è un METODO DIRETTO perché richiede un numero finito di passi e non cambia la soluzione del sistema.
 Poi, a risolvere il sistema, si usa la tecnica di backward substitution (che richiede $n^2/2$ operazioni)



$$\begin{cases} x_n = \frac{b_n^{(n)}}{a_{nn}^{(n)}} \\ x_i = \frac{b_i^{(n)} - \sum_{j=i+1}^n a_{ij}^{(n)} x_j}{a_{ii}^{(n)}}, \quad \forall i=1; \dots; n-1 \end{cases}$$

Si dimostra che, x l'algoritmo di Gauss, il costo computazionale vale

$$c.c. = O\left(\frac{n^3}{3}\right)$$

Il termine $\frac{n^2}{2}$ dovuto alla backward substitution è trascurato perché, x alti n (sistemi con tante equazioni e incognite), asintoticamente domina il termine n^3

più rigorosi sarebbe $c.c. = O\left(\frac{n^3}{3}\right) + \frac{n^2}{2}$

Es (APPLICAZIONE DEL METODO DI ELIMINAZIONE DI GAUSS)

Sia dato il sistema

$$A\vec{x} = \vec{b}$$

dovente

$$A = \begin{bmatrix} 2 & -1 & 1 & -2 \\ 0 & 2 & 0 & -1 \\ 1 & 0 & -2 & 1 \\ 0 & 2 & 1 & 1 \end{bmatrix} \quad \vec{b} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$

Si osserva fin da subito che in realtà la soluzione è nota fin da subito perché, quando l' i -esimo elemento del vettore $\vec{b}$ è pari alla somma degli elementi sull' i -esima riga della matrice A , allora la soluzione è

$$\vec{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \quad \rightarrow \text{vettore uniforme}$$

La matrice $A = A^{(1)}$ è 4×4 , x cui x applicare il metodo di Gauss occorrono 4 passi

Passo (I) Si calcolano i moltiplicatori (infatti $a_{11}^{(1)} = 2 \neq 0$), ricordandosi che

$$m_{i,1} = -\frac{a_{i1}^{(1)}}{a_{11}^{(1)}}$$

$$\rightarrow m_{2,1} = 0$$

$$\rightarrow m_{3,1} = -1/2$$

$$\rightarrow m_{4,1} = 0$$

Dunque la matrice si trasforma come

$$\rightarrow \frac{29}{10} x_4 = \frac{29}{10} \cdot 1 \Rightarrow x_4 = 1$$

$$\rightarrow -\frac{5}{2} x_3 = -\frac{1}{4} - \frac{9}{4} \cdot 1 = -\frac{5}{2} \Rightarrow x_3 = 1$$

$$\rightarrow 2x_2 = 1 - 0 \cdot 1 + 1 \cdot 1 = 2 \Rightarrow x_2 = 1$$

$$\rightarrow 2x_1 = 0 + 1 \cdot 1 - 1 \cdot 1 + 2 \cdot 1 = 2 \Rightarrow x_1 = 1$$

Dunque, come è già stato accennato prima, la soluzione del sistema è

$$\vec{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

OSSERVAZIONE: nel metodo di eliminazione di Gauss una condizione necessaria è che

$$a_{kk}^{(k)} \neq 0$$

Tale condizione è garantita se la matrice dei coefficienti A soddisfa almeno una delle seguenti proprietà

~ si consideri la matrice A_k , formata dagli elementi a_{ij} fino alla riga e colonna k , cioè

$$a_{k,i,j} = a_{ij}, \quad 1 \leq i,j \leq k \quad k = 1, \dots, n$$

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1k} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2k} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \dots & a_{kk} & \dots & a_{kn} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & \dots & \dots & a_{nn} \end{bmatrix}$$

deve valere allora che

$$\det A_k \neq 0, \quad \forall k = 1, \dots, n$$

Cioè, considerando la sottomatrice fino a a_{kk} , se a ogni k si ha determinante non nullo, allora si è sicuri che in ogni passaggio del metodo di eliminazione di Gauss non si trovano mai elementi della diagonale nulli e si può procedere tranquillamente

- la matrice A è una **matrice diagonale dominante per righe**
- la matrice A è una **matrice diagonale dominante per colonne**
- la matrice A è una **matrice simmetrica definita positiva**

$$\frac{\|\tilde{x} - \bar{x}\|_{\infty}}{\|\tilde{x}\|_{\infty}} \sim 10^{-9} \rightarrow \text{è una buona approssimazione}$$

Applicando ora il pivoting, si costruisce un algoritmo più stabile in quanto si controllano anche i pivot in modo che si trovi l'elemento più grande e l'errore relativo vale

$$\frac{\|\tilde{x} - \bar{x}\|_{\infty}}{\|\tilde{x}\|_{\infty}} \sim 10^{-15} \rightarrow \text{è un algoritmo più costoso ma anche più accurato}$$

Dunque la tecnica di pivoting è preferibile, anche se in alcuni casi, in presenza di matrici particolari, essa può essere evitata in quanto con esse si è sicuri che l'elemento $a_{kk}^{(k)}$ è il più grande in valore assoluto nella colonna k . Queste matrici sono

- matrici diagonali dominanti per colonne
- matrici simmetriche a diagonale dominante
- matrici simmetriche definite positive

4. Fattorizzazione di una matrice:

essa è un'operazione che può rivelarsi utile alla risoluzione dei sistemi lineari. Per semplicità, si applica tale operazione a una matrice A 4×4 (il discorso può però essere esteso a ordini diversi), definita come

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

Poi, a livello di notazione, si definisce

$$A^{(1)} = A$$

Nel metodo di eliminazione di Gauss, o ora SENZA PIVOTING (cioè si suppone che, a ogni passo del metodo, l'elemento $a_{kk}^{(k)}$ sia non nullo in modo da non dover scambiare equazioni), il primo passo è di eliminare la prima colonna definendo i moltiplicatori

$$m_{i1} = -\frac{a_{i1}^{(1)}}{a_{11}^{(1)}}, \quad i = 2, 3, 4$$

Ora si definisce la matrice M_1 come

$$M_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ m_{21} & 1 & 0 & 0 \\ m_{31} & 0 & 1 & 0 \\ m_{41} & 0 & 0 & 1 \end{bmatrix}$$

→ è una matrice con diagonale uno, triangolare inferiore e avente i moltiplicatori sulla prima colonna

Moltiplicando la matrice $M_1 \times A^{(1)}$, si ottiene

Così, partendo da un sistema

$$A\vec{x} = \vec{b}$$

si sono costruite tre matrici M_1, M_2 e M_3 .

Ora le si moltiplicano a sinistra x A e, ovviamente, anche x $\vec{b}$ (x mantenere l'uguaglianza)

$$M_3 M_2 M_1 A \vec{x} = M_3 M_2 M_1 \vec{b}$$

$$\underbrace{M_3 M_2 M_1 A}_{A^{(2)}} \vec{x} = M_3 M_2 M_1 \vec{b}$$

(poiché $A = A^{(1)}$)

$$\underbrace{M_3 M_2 A^{(2)}}_{A^{(3)}} \vec{x} = M_3 M_2 M_1 \vec{b}$$

$$\underbrace{M_3 A^{(3)}}_U \vec{x} = \underbrace{M_3 M_2 M_1 \vec{b}}_{\vec{b}}$$

$$U \vec{x} = \vec{b}$$

Dunque si è giunti a un sistema

$$U \vec{x} = \vec{b}$$

dove

$$M_3 M_2 M_1 A = U$$

cioè

$$A = (M_3 M_2 M_1)^{-1} U$$

Però cos'è la matrice $(M_3 M_2 M_1)^{-1}$?

Innanzitutto, si osserva che essa è l'inverso del prodotto di 3 matrici triangolari inferiori invertibili (in quanto la loro diagonale non ha elementi nulli), x cui il loro prodotto è invertibile, cioè $(M_3 M_2 M_1)^{-1}$ esiste e vale

$$(M_3 M_2 M_1)^{-1} = M_1^{-1} M_2^{-1} M_3^{-1}$$

~ l'inverso del prodotto è il prodotto degli inversi con ordine invertito

Ora si osserva che i singoli termini valgono

$$M_1^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -m_{21} & 1 & 0 & 0 \\ -m_{31} & 0 & 1 & 0 \\ -m_{41} & 0 & 0 & 1 \end{bmatrix}$$

$$M_2^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -m_{32} & 1 & 0 \\ 0 & -m_{42} & 0 & 1 \end{bmatrix}$$

$$M_3^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -m_{43} & 1 \end{bmatrix}$$

Ossia gli inversi mantengono la stessa struttura, cioè sono triangolari inferiori, hanno una sulla diagonale principale e gli zeri e i moltiplicatori sono sulle stesse posizioni, ma questi ultimi sono cambiati di segno.

PROPRIETÀ DELLE MATRICI DI PERMUTAZIONE:

Quanto vale la matrice P^2 ?
Si sa che, x definizione, essa vale

$$P^2 = PP$$

Se però si moltiplica una qualunque matrice x P , si scambiano due righe all'interno di tale matrice. Questa matrice però è proprio pari a P , cioè è ottenuta scambiando le proprio quelle due righe nella matrice identica. Così, visto che di fatto si scambiano a due volte quelle righe, il risultato è che si torna alla matrice identica

$$P^2 = PP = I$$

Scrivere tale espressione, però, equivale a dire che

$$P^{-1} = P$$

Ora si torna alla tecnica di eliminazione di Gauss con pivoting, perché con tale tecnica si vuole ottenere una fattorizzazione della matrice A (sempre supposta essere $n \times n$).
Data un sistema lineare

$$A\tilde{x} = \tilde{b}$$

prima di definire i moltiplicatori, si guarda la I colonna e in essa, tra gli elementi a_{11}, a_{21}, a_{31} e a_{n1} , si cerca l'elemento a_{r1} con valore assoluto massimo. Dopodiché si permuta moltiplicando la matrice A x P_1 , ottenuta dalla matrice identica scambiando la I riga con la r -esima. Ovviamente, si moltiplica x P_1 anche il vettore $\tilde{b}$, perché occorre fare lo scambio di righe anche in esso.

$$P_1 A \tilde{x} = P_1 \tilde{b}$$

→ è possibile che $P_1 = I$, se in modulo l'elemento a_{11} è massimo

Ora si possono definire i moltiplicatori e definire la matrice M_1 , in modo da completare così il I passo di Gauss

$$\underbrace{M_1 P_1 A}_{A^{(2)}} \tilde{x} = M_1 P_1 \tilde{b}$$

Si è così definito il nuovo sistema

$$A^{(2)} \tilde{x} = M_1 P_1 \tilde{b}$$

Nel II passo, si guarda la II colonna della matrice $A^{(2)}$ e si cerca l'elemento più grande in valore assoluto e si scambia la II riga con la r' -esima, attraverso la matrice P_2 .
Si ha così che

$$P_2 M_1 P_1 A \tilde{x} = P_2 M_1 P_1 \tilde{b}$$

A questo punto si definisce la matrice dei moltiplicatori M_2

$$\underbrace{M_2 P_2 M_1 P_1 A}_{A^{(3)}} \tilde{x} = M_2 P_2 M_1 P_1 \tilde{b}$$

Si ottiene così la matrice $A^{(3)}$

Il III e ultima passo prevede di applicare ancora il pivoting introducendo la matrice P_3 , poi definire la matrice associata ai moltiplicatori M_3 e ottenere così che

$$\underbrace{M_3 P_3 M_2 P_2 M_1 P_1 A}_{A^{(4)} = U} \tilde{x} = M_3 P_3 M_2 P_2 M_1 P_1 \tilde{b}$$

→ se non ci sono scambi, poiché $P_i = I$, compaiono solo le matrici M_1, M_2 e M_3

$$\tilde{M} = P_2 M_1 P_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 3 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ 5 & 0 & 0 & 1 \end{bmatrix}$$

Se ora si porta la matrice $\tilde{M}$ dall'altro membro, cioè lo si inverte, si ottiene una matrice L , definita come

$$L = \tilde{M}^{-1}$$

Ricordandosi però ^{di} come è definita la matrice $\tilde{M}$, si ha che

$$L = \tilde{M}^{-1} = (M_3 M_2 M_1)^{-1} = M_1^{-1} M_2^{-1} M_3^{-1}$$

e, come nel caso della fattorizzazione $A = LU$, questa matrice L è una matrice triangolare inferiore con tutti i 1 sulla diagonale principale e contenente gli stessi elementi delle matrici M_1 , M_2 e M_3 ~~ma~~ cambiati però di segno. Si è però appena detto che queste matrici contengono i moltiplicatori scambiati di posizione.

Di conseguenza, la matrice L è una MATRICE TRIANGOLARE INFERIORE CON TUTTI I SULLA DIAGONALE PRINCIPALE E CONTENENTE NEL TRIANGOLO INFERIORE I MOLTIPLICATORI CAMBIATI DI SEGNO E ANCHE SCAMBIATI DI POSIZIONE.

Si ottiene così la fattorizzazione della matrice A

$$PA = LU$$

OSSERVAZIONE: si osserva che, se la matrice A è simmetrica definita positiva, diagonale dominante $\times$ righe oppure diagonale dominante $\times$ colonne, allora la matrice PA sarà semplicemente "definita" positiva $\times$ via della tecnica di pivoting parziale.

Es (CALCOLO DELLA FATTORIZZAZIONE)

Data la matrice

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

Ricavare la fattorizzazione $PA = LU$, cioè valutare le matrici P , L e U .

I passo) Si considera la I colonna e si valuta in essa l'elemento massimo in valore assoluto. In questo caso, occorre invertire la I riga con la III riga e, pertanto, si definisce la matrice P_1 come

$$P_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

$$M_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{1}{2} & 1 \end{bmatrix}$$

La matrice $A^{(3)}$ si può così calcolare come

$$A^{(3)} = M_2 P_2 A^{(2)} = M_2 P_2 M_1 P_1 A = \begin{bmatrix} -6 & 0 & 2 \\ 0 & 2 & 2 \\ 0 & 0 & -1 \end{bmatrix} = U$$

A questo punto, calcolata la matrice U , si può passare a calcolare le altre matrici della fattorizzazione

→ P

$$P = P_2 P_1 = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

→ L : Bisogna innanzitutto calcolare la matrice $\bar{M}$, che è pari in questo caso a

$$\bar{M} = M_2 \bar{M}_1 = M_2 P_2 M_1 P_2$$

→ infatti nel calcolo della fattorizzazione si fa l'operazione $M_2 P_2 M_1 P_1 A = M_2 P_2 M_1 P_2 P_1 A$

Da qui la matrice $\bar{M}$ vale

$$\bar{M} = M_2 \bar{M}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{2} & -\frac{1}{2} & 1 \end{bmatrix}$$

si può ricavare direttamente ricordando che i moltiplicatori associati alla matrice M_1 sono scambiati di posizione x effetto della matrice P_2

Di conseguenza, la matrice L vale

$$L = \bar{M}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix}$$

→ la si può ricavare direttamente da $\bar{M}$ senza calcoli semplicemente cambiando segno ai moltiplicatori

Il costo computazionale della fattorizzazione tiene conto del fatto che essa passa attraverso l'eliminazione di Gauss e, infatti, costa quanto questa operazione, cioè

$$C.C. = O(N^3)$$

→ calcolo del determinante:

data

data una matrice A , di cui è stata calcolata la fattorizzazione $PA = LU$, se ne vuole calcolare il determinante.

Si applica allora il determinante a entrambi i membri della fattorizzazione

$$\det(PA) = \det(LU)$$

Si sa che il determinante del prodotto è pari al prodotto dei determinanti

$$\det P \det A = \det L \det U$$

Da qui si ricava il determinante della matrice A

$$\det A = (\det P)^{-1} \det L \det U$$

Si ricorda che il reciproco del determinante corrisponde al determinante dell'inversa

$$(\det P)^{-1} = \det(P^{-1})$$

Da qui si ha che

$$\det A = \det(P^{-1}) \det L \det U$$

Si ricorda però che vale la relazione

$$P^{-1} = P$$

Pertanto si ottiene che

$$\det A = \det P \det L \det U$$

Ora si valutano i singoli fattori

→ $\det P$: si osserva che, se la matrice P è ottenuta scambiando 2 righe della matrice identica, il suo determinante è pari all'opposto del determinante della matrice identica, cioè vale -1 .
Se si scambiano altre ^{due} righe (cioè si fanno 2 scambi), il determinante diventa positivo e vale 1 .

Così si può dire che

$$\det P = (-1)^s$$

dove s è il numero di scambi di righe nel metodo di Gauss con pivoting parziale (si nota che, se $s=0$, cioè se non ci sono scambi, allora $\det P = 1$).

→ $\det L$: la matrice L è triangolare inferiore con tutti 1 sulla diagonale principale (e nel triangolo inferiore ha i moltiplicatori cambiati di segno ed eventualmente scambiati di posizione), a cui il determinante vale

$$\det L = 1 \quad \rightarrow \text{è il prodotto degli elementi sulla diagonale}$$

→ $\det U$: è una matrice triangolare superiore, a cui il determinante è pari al prodotto degli elementi u_{ii} ($i=1, \dots, n$) sulla diagonale

In questi sistemi lineari però la matrice dei coefficienti A è sempre la stessa. Pertanto si può CALCOLARE A MONTE LA FATTORIZZAZIONE $PA = LU$, operazione che ha costo $O(n^3/3)$ una volta e questa è fatta soltanto una volta.

Dopo di che si applica la fattorizzazione al I sistema, nello stesso modo di prima

$$PA\vec{x}_1 = P\vec{b}_1$$

Ciò equivale a scrivere

$$LU\vec{x}_1 = P\vec{b}_1$$

Ora si spezza tale sistema in 2 sistemi lineari

$$\begin{cases} L\vec{y}_1 = P\vec{b}_1 \\ U\vec{x}_1 = \vec{y}_1 \end{cases}$$

Si hanno così 2 sistemi, ciascuno con costo computazionale $O(n^2/2)$ perché triangolare, e cui il I sistema si hanno $O(n^2)$ operazioni.

Si passa ora al II sistema e anche lì si usa la fattorizzazione

$$PA\vec{x}_2 = P\vec{b}_2$$

Ciò equivale a scrivere che

$$LU\vec{x}_2 = P\vec{b}_2$$

Si spezza e si ottengono i due sistemi

$$\begin{cases} L\vec{y}_2 = P\vec{b}_2 \\ U\vec{x}_2 = \vec{y}_2 \end{cases}$$

Il costo di tale sistema è di nuovo $O(n^2)$.

Si ripete lo stesso ragionamento per gli altri sistemi lineari, in modo da spezzarli in una coppia di sistemi lineari del tipo

$$\begin{cases} L\vec{y}_i = P\vec{b}_i \\ U\vec{x}_i = \vec{y}_i \end{cases}, i = 1, \dots, j, n$$

e per ciascuno di essi il costo computazionale sarà $O(n^2)$

Così, in totale, il costo computazionale sarà dato dal contributo della fattorizzazione (fatta una volta sola) e dal contributo delle operazioni fatte per risolvere i singoli sistemi e dunque esso vale

$$c.c. = O\left(\frac{n^3}{3} + pn^2\right)$$

A differenza di prima, il termine p non moltiplica n^3 ma n^2 e dunque la spesa è minore (poiché n^2 è trascurabile rispetto a n^3).

5 Ora si introduce una particolare fattorizzazione, cioè la **Fattorizzazione di Cholesky** di una matrice:

si consideri una MATRICE A SIMMETRICA DEFINITA POSITIVA, in modo che il pivoting parziale sia superfluo in quanto si è sicuri che ogni elemento a_{kk} sia massimo in valore assoluto nella colonna di competenza. Così non occorrono scambi di righe (x cui $P = I$) e si può parlare di fattorizzazione

$$A = LU$$

Per ipotesi, si sa che

$$A = {}^t A \quad \rightarrow \text{x simmetria}$$

Di conseguenza, vale la relazione

$$LU = {}^t(LU) = {}^t U {}^t L$$

Così si ha la relazione

$$LU = {}^t U {}^t L$$

Si nota che la matrice ${}^t U$ è triangolare inferiore x cui ha la stessa struttura della matrice L ; viceversa, la matrice ${}^t L$ è triangolare superiore e ha così la stessa struttura della matrice U .

Ora, si può dimostrare che la **Fattorizzazione $PA = LU$ è unica**, cioè esistono un'unica matrice L e un'unica matrice U che realizzano la fattorizzazione $PA = LU$. ~~Già~~ Guardando però la relazione appena scritta, si osserva che esistono un'altra matrice triangolare inferiore (${}^t U$) e un'altra matrice triangolare superiore (${}^t L$) che realizzano questo prodotto. L'unica anomalia è che sulla diagonale principale della matrice ${}^t U$ non esistono solo elementi 1, così, a meno di elementi sulla diagonale principale, si può dire che ${}^t U$ coincide con L (in tal modo, la condizione di unicità è soddisfatta), cioè che

$${}^t U \sim L$$

Si dimostra però che, x una matrice A simmetrica definita positiva, si può ottenere una fattorizzazione di A pari a

$$A = L {}^t L$$

dove L non è la matrice L della fattorizzazione ma L è una matrice triangolare inferiore con elementi sulla diagonale principale tutti positivi.

Così si può ottenere una fattorizzazione con un'unica matrice L .

Si osserva, però, che a livello di notazione si tende a usare la matrice R definita come

$$R = {}^t L$$

In tal modo, si può rileggere la fattorizzazione della matrice A come

$$A = {}^t R R$$

Questa prende il nome di **Fattorizzazione di Cholesky**:

esiste una matrice R triangolare superiore con elementi sulla diagonale principale tutti positivi tale che la matrice A si presenti come il prodotto ${}^t R R$. In tal modo, si ottiene una FATTORIZZAZIONE CON UN'UNICA MATRICE R .

5 RIEPILOGO SUL METODO DI ELIMINAZIONE DI GAUSS:

esso è un metodo diretto a risolvere un sistema lineare.
Per definizione, un **metodo diretto** ha le seguenti proprietà:

- la soluzione $\vec{x}$ è determinata con un NUMERO FINITO DI PASSI. Per il metodo di Gauss, il numero di passi a ottenere una matrice triangolare superiore è pari a $n-1$, dove n è la dimensione della matrice dei coefficienti.
- SE SI LAVORASSE IN ARITMETICA ESATTA, LA SOLUZIONE $\vec{x}$ SAREBBE DETERMINATA IN MODO ESATTO. Usando invece le operazioni di macchina, la soluzione $\vec{x}$ è approssimata a seguito dell'introduzione di operazioni di macchina e l'errore è dato dalla precisione di macchina (parametro indipendente dall'utente).
- i metodi diretti MODIFICANO LA MATRICE DEI COEFFICIENTI e, a questo, tali tecniche sono utili a sistemi caratterizzati da:
 - **matrici dei coefficienti di piccole dimensioni**
 - **matrici dei coefficienti dense:**
queste sono le opposte delle matrici sparse, in cui sono presenti tanti elementi nulli (es matrici diagonali, triangolari, etc.)
Infatti, non è detto che il metodo di Gauss conservi gli zeri, cioè può aggiungere più elementi non nulli rispetto a quelli iniziali.
 - ↳ le matrici sparse si vuole mantenere la struttura iniziale, in modo da avere pochi dati da memorizzare

Metodi iterativi

Metodi iterativi: dati

dato un generico sistema lineare

$$A\vec{x} = \vec{b}$$

in cui si esplicitano tutte le equazioni:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2 \\ \dots \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n = b_n \end{cases}$$

Si suppone innanzitutto che gli elementi a_{ii} sulla diagonale principale di A siano non nulli, a ogni $i = 1, \dots, n$.
Allora si può prendere la I equazione e isolare da essa l'incognita x_1

$$x_1 = \frac{b_1 - \sum_{j=2}^n a_{1j}x_j}{a_{11}}$$

Tale ragionamento può essere anche fatto a la II equazione

→ metodo di Gauss-Seidel:

scelto un vettore iniziale $\vec{x}^{(0)}$, allora all'iterazione $k+1$ l' i -esima soluzione x_i vale

$$x_i^{(k+1)} = \frac{b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij} x_j^{(k)}}{a_{ii}}$$

Anche questa è un metodo iterativo in quanto, noto il vettore $\vec{x}^{(0)}$, con questa relazione si può descrivere il vettore $\vec{x}^{(1)}$ e poi il vettore $\vec{x}^{(2)}$ e così via.
Inoltre si osserva che, al passo 1, a costruire la soluzione $x_1^{(1)}$, si nota che il II addendo vale

$$\sum_{j=1}^0 a_{ij} x_j^{(k+1)} \rightarrow \text{poichè } i=1$$

Cioè tale termine scompare e si ricava subito $x_1^{(1)}$.
Passando invece all'incognita x_2 , il II addendo diventa

$$\sum_{j=1}^1 a_{ij} x_j^{(k+1)} \rightarrow \text{poichè } i=2$$

Cioè diventa pari al termine $x_1^{(1)}$, che si è appena ricavato.
Dunque il II addendo contiene i termini che sono già stati calcolati prima, in quanto, se la somma si arresta all'indice $i-1$, allora i termini $x_j^{(k+1)}$ si arrestano all'indice $i-1$ ma, fino a tale indice, questi termini sono già stati calcolati.
} infatti quello che si sta calcolando è proprio $x_i^{(k+1)}$

ES (APPLICAZIONE DEI METODI DI ITERAZIONE)

Sia dato il sistema lineare

$$\begin{cases} 3x_1 - x_2 = 2 \\ x_1 + 2x_2 = 3 \end{cases}$$

→ si nota che la soluzione è il vettore unitario

Lo si risolve ricorrendo ai metodi di iterazione.

Inanzitutto, si isolano le due incognite x_1 e x_2

$$\begin{cases} x_1 = \frac{2+x_2}{3} \\ x_2 = \frac{3-x_1}{2} \end{cases}$$

Si applica, a iniziare, il metodo di Jacobi, che, a sua definizione, fornisce il metodo iterativo

$$\begin{cases} x_1^{(k+1)} = \frac{2+x_2^{(k)}}{3} \\ x_2^{(k+1)} = \frac{3-x_1^{(k)}}{2} \end{cases}$$

Si suppone di scegliere come dato iniziale $\vec{x}^{(0)}$ il vettore

$$\vec{x}^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

A questo punto si applica la I iterazione e si ottiene che

Per ottenere un metodo iterativo, si parte come sempre da un dato iniziale $\vec{x}^{(0)}$ e lo schema iterativo si ottiene imponendo che, al passo k , tra i vettori $\vec{x}^{(k+1)}$ e $\vec{x}^{(k)}$ valga la relazione

$$D\vec{x}^{(k+1)} = \vec{b} - C\vec{x}^{(k)}$$

Così, assegnato il dato iniziale $\vec{x}^{(0)}$, a ricavare al I passo il vettore $\vec{x}^{(1)}$ si risolve il sistema lineare avente come matrice dei coefficienti D e termine noto $\vec{b} - C\vec{x}^{(0)}$ e in esso si esplicita la soluzione $\vec{x}^{(1)}$

$$\vec{x}_i^{(1)} = D^{-1}(\vec{b} - C\vec{x}^{(0)})$$

Una volta ricavato questo vettore, nel II passo si ripete un procedimento analogo in cui si risolve il sistema e si ricava il vettore $\vec{x}^{(2)}$, e così via.

Dunque si può dire che ogni passo dell'iterazione può essere interpretato come la risoluzione di un sistema lineare, avente come matrice dei coefficienti la matrice D e come vettore dei termini noti il vettore $\vec{b} - C\vec{x}^{(0)}$.

Si può ora notare che esistono infinite modalità con cui decomporre la matrice A , in quanto esistono infiniti modi a definire la matrice D (la matrice C viene di conseguenza, poiché $C = A - D$).

Pertanto si può scegliere la matrice D in modo che essa sia la più semplice possibile. In tal modo, a ogni passo dell'iterazione, si ha un sistema lineare da risolvere che è il più semplice possibile. A tal proposito, si sa che i sistemi più semplici sono quelli diagonali e poi quelli triangolari.

Adesso si valuta come si comportano i due metodi di iterazione

→ **metodo di Jacobi:**

esso è espresso dalla relazione

$$a_{ij}x_i^{(k+1)} = b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^n a_{ij}x_j^{(k)}, \quad \forall i=1, \dots, n$$

Si vuole vedere com'è strutturata la matrice D (e di conseguenza anche la matrice C) nel metodo di Jacobi.

A tal proposito, nell'espressione di Jacobi si osserva che, a ogni equazione, si riesce a isolare ogni variabile $x_i^{(k+1)}$ al passo $k+1$ e ciascuna di esse è moltiplicata x un coefficiente a_{ij} . Questa relazione può essere realizzata con una relazione $D\vec{x}^{(k+1)}$ se D è una matrice diagonale.

Dunque, **nel metodo di Jacobi la matrice D_j è una matrice diagonale**, contenente gli elementi a_{ii}

$$D_j = \begin{bmatrix} a_{11} & & & 0 \\ & a_{22} & & \\ & & \ddots & \\ 0 & & & a_{nn} \end{bmatrix}$$

Di conseguenza, la matrice C_j vale

$$C_j = A - D_j = \begin{bmatrix} 0 & & & a_{1n} \\ & 0 & & \\ & & \ddots & \\ a_{n1} & & & 0 \end{bmatrix}$$

→ si sottrae ad A la diagonale e rimangono solo i due triangoli

2 Convergenza dei metodi iterati:

occorre ancora vedere sotto quali ipotesi questi metodi garantiscono che la soluzione $\vec{x}^{(k)}$ converga alla soluzione $\vec{x}$ e con quale velocità, in quanto possono esistere diversi casi.

Innanzitutto, se si implementa un metodo nell'elaboratore, occorre definire, come x ogni metodo iterativo, quando fermarsi, cioè **occorre indicare all'elaboratore il numero di iterazioni**. Se, infatti, nel metodo di Gauss il numero di passi è predefinito (vale $n-1$), i metodi iterativi procedono finché ~~non~~ lo decide l'utente. Si ha così la necessità di fissare dei **criteri d'arresto del metodo d'iterazione**, che diano il numero di passi, e questi sono due e sono da applicare insieme.

→ si fissa innanzitutto un **tetto massimo per i passi** $k_{\max}$, che indica il numero di iterazioni concesse.

Così si definisce il metodo iterativo come

$$\vec{x}^{(k+1)} = \vec{b} - C\vec{x}^{(k)}, \quad k = 0; 1; \dots; k_{\max} \quad \rightarrow \text{non va più all'infinito ma a un valore } k_{\max} \text{ scelto}$$

→ considerando due iterate successive, un'iterata al passo k e una al passo $k+1$, si valuta la loro vicinanza ricorrendo a una generica norma vettoriale. Se la distanza è più piccola di una tolleranza $Toll$ fissata, s'interrompe il processo iterativo, altrimenti si prosegue finché non si raggiunge il passo $k_{\max}$. La tolleranza può essere riferita sia all'errore assoluto (e dunque a una distanza assoluta) sia all'errore relativo, calcolato rispetto all'iterata al passo $k+1$.

$$\|\vec{x}^{(k+1)} - \vec{x}^{(k)}\| \leq Toll$$

oppure

$$\frac{\|\vec{x}^{(k+1)} - \vec{x}^{(k)}\|}{\|\vec{x}^{(k+1)}\|} \leq Toll$$

In base al tipo di errore scelto (relativo o assoluto), si definisce un certo criterio d'arresto, cioè un certo valore della tolleranza $Toll$, e se le due iterate sono abbastanza vicine rispetto alla tolleranza, allora ci si ferma (altrimenti si continua).

Si osserva che, x tolleranze più piccole, si dovranno fare più iterate perché queste dovranno essere più vicine. Se, alla fine, il numero E di iterazioni fatte finché tale criterio non è soddisfatto è tale che

$$E < k_{\max}$$

significa che si è usciti dall'iterazione prima del tetto massimo definito e dunque si è usciti perché l'errore è diventato più piccolo della tolleranza.

Si osserva che i metodi di arresto, in generale, hanno senso quando il metodo iterativo converge rapidamente, cioè quando ci si aspetta che due vettori consecutivi siano sempre più vicini tra di loro e più la convergenza è veloce, più è basso il numero di iterazioni con cui si raggiunge.

In tal caso, si uscirà fuori dall'implementazione perché si è passati sotto la tolleranza $Toll$ e non perché si è raggiunto il limite $k_{\max}$.

In caso contrario, in genere si esce con un numero di passi $E = k_{\max}$ perché si è scelto un tetto massimo troppo basso oppure perché si ha una matrice poco convergente.

Tale matrice, durante l'iterazione, rimane inalterata (non cambia al cambiare dei passi) e può essere di Jacobi o di Gauss-Seidel in base al fatto che la matrice D sia diagonale o triangolare.
In tal modo, a ogni passo si lega l'errore al passo k con l'errore al passo $k-1$ attraverso la matrice d'iterazione

$$\vec{e}^{(k)} = B \vec{e}^{(k-1)}$$

Valutando però l'errore al passo $k-1$, si osserva che

$$\vec{e}^{(k-1)} = B \vec{e}^{(k-2)}$$

Di conseguenza si può dire che

$$\vec{e}^{(k)} = B \vec{e}^{(k-1)} = B(B \vec{e}^{(k-2)}) = B^2 \vec{e}^{(k-2)}$$

Questo ragionamento, ovviamente, può essere ripetuto k volte in modo da trovare una relazione tra l'errore al passo k e l'errore al passo 0

$$\vec{e}^{(k)} = B \vec{e}^{(k-1)} = B^2 \vec{e}^{(k-2)} = B^3 \vec{e}^{(k-3)} = \dots = B^k \vec{e}^{(0)}$$

L'errore al passo 0, però, è dettato dalla scelta del dato iniziale $\vec{x}^{(0)}$ e, pertanto, si può dire che l'errore al passo k è legato all'errore iniziale.
Poiché, in generale, la soluzione del sistema non è nota, l'errore iniziale $\vec{e}^{(0)}$ non è noto ma esso comunque è un valore fisso, cioè è semplicemente una costante.

Ora, si sa che, a definizione, una successione $\vec{x}^{(k)}$ converge a una soluzione $\vec{x}$ se

$$\lim_{k \rightarrow \infty} \vec{x}^{(k)} = \vec{x}$$

Ciò equivale a imporre la condizione

$$\lim_{k \rightarrow \infty} \vec{e}^{(k)} = 0$$

Quando il vettore $\vec{e}^{(k)}$ tende al vettore nullo?
Ricordandosi della relazione

$$\vec{e}^{(k)} = B^k \vec{e}^{(0)}$$

allora deve verificarsi che

$$B^k \vec{e}^{(0)} \rightarrow 0, \quad k \rightarrow \infty$$

Il vettore $\vec{e}^{(0)}$ però è un vettore costante indipendente da k e dunque è necessario che la matrice B^k tenda a zero, cioè alla matrice nulla.
Così si ha convergenza se vale la proprietà

$$\lim_{k \rightarrow \infty} B^k = [0]$$

Si dimostra però che condizione necessaria e sufficiente affinché valga tale proprietà è che il raggio spettrale della matrice B sia tale che

$$\rho(B) = \max_{i=1, \dots, n} \{ |\lambda_i| \} < 1 \quad \lambda_i = \text{autovalore di } B$$

In base a tutto questo, si può dire che **condizione necessaria e sufficiente affinché un metodo iterativo descritto in questo modo converga all'unica soluzione del sistema lineare (si suppone che la matrice dei coefficienti A sia invertibile) è che il raggio spettrale della matrice**

OSSERVAZIONE: si osserva che la convergenza di un metodo di iterazione non comporta la convergenza di un altro.
In genere, però, se entrambi convergono, il metodo di Gauss-Seidel converge più velocemente e la velocità si lega al raggio spettrale (in virtù della relazione $\rho(B) < 1$).

Es (APPLICAZIONE DEI METODI DI ITERAZIONE - II)

Sia dato il sistema lineare

$$A\vec{x} = \vec{b}$$

definito come

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

$$\vec{b} = \begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix}$$

→ si osserva che la soluzione esatta $\vec{x}$ è il vettore unitario

Il metodo di Gauss-Seidel converge?

Prima di usare la condizione necessaria e sufficiente, si guarda se si realizza una delle condizioni sufficienti: si osserva che la matrice dei coefficienti è diagonale dominante x righe, x cui il metodo di Gauss-Seidel converge.

Ora, a partire dal dato iniziale $\vec{x}^{(0)}$ coincidente con il vettore nullo, si calcolano le prime 3 iterate del metodo di Gauss-Seidel.

Per fare questo, si esplicitano dapprima le equazioni del sistema lineare

$$\begin{cases} 3x_1 - 2x_2 = 1 \\ x_1 + 2x_2 = 3 \\ x_3 = 1 \end{cases}$$

Ora si isolano le incognite x_i :

$$\begin{cases} x_1 = \frac{1+2x_2}{3} \\ x_2 = \frac{3-x_1}{2} \\ x_3 = 1 \end{cases}$$

A questo punto, si può definire il metodo iterativo

$$\begin{cases} x_1^{(k+1)} = \frac{1+2x_2^{(k)}}{3} \\ x_2^{(k+1)} = \frac{3-x_1^{(k+1)}}{2} \\ x_3^{(k+1)} = 1 \end{cases}$$

→ si mette l'indice $k+1$ alle incognite di indice minore di i (perché i coefficienti con $k+1$ stanno nel triangolo inferiore)
→ iterando, x x_3 si ottiene sempre 1

Tale metodo può essere anche scritto introducendo la matrice D_{GS} che, in questo caso, è la parte triangolare inferiore della matrice A

$$D_{GS} = \begin{bmatrix} 3 & 0 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

3 RIEPILOGO SUI METODI ITERATIVI:

in generale, a definizione, un **metodo iterativo** presenta le seguenti caratteristiche

→ la soluzione $\vec{x}$ (o meglio, la sua approssimazione) è determinata come limite di una successione di vettori $\{\vec{x}^{(n)}\}$ convergente

$$\vec{x} = \lim_{n \rightarrow \infty} \{\vec{x}^{(n)}\}$$

→ LAVORANDO IN ARITMETICA INFINITA, LA SOLUZIONE $\vec{x}$ È DETERMINATA IN MANIERA APPROSSIMATA, a differenza del metodo di eliminazione di Gauss dove, con un numero finito di operazioni, si ottiene la soluzione precisa.

Nell'aritmetica con precisione finita, viene determinata una successione decisa dall'utente (in quanto i metodi di iterazione produrrebbero infinite operazioni) e dunque la soluzione può essere determinata con una precisione stabilita dall'operatore stesso, in base al numero di operazioni ~~spese~~ fatte spendere (l'utente infatti può un arresto fissando il numero di iterazioni consentite o un limite di tolleranza)

↓
nei metodi diretti invece la precisione non può essere decisa

→ nei metodi iterativi LA MATRICE DEI COEFFICIENTI NON VIENE MODIFICATA in quanto questi sono definiti come

$$\vec{x}^{(k+1)} = D^{-1} (\vec{b} - C\vec{x}^{(k)})$$

In tale relazione, le matrici D e C non vengono mai alterate e pertanto i metodi iterati convergono in presenza di

→ **matrici di grandi dimensioni**, in quanto è l'utente che decide il numero di passi (nel metodo di eliminazione di Gauss, invece, il numero dipende dalla dimensione n)

→ **matrici sparse**, in quanto, non modificando gli elementi, si conserva la sparsità e dunque si risparmia nell'occupazione della memoria.

↓
così, se nel metodo di eliminazione di Gauss si cambia la matrice dei coefficienti ma la soluzione rimane invariata, nei metodi iterativi la matrice non viene cambiata ma cambia la soluzione, poiché cambia la soluzione di ogni iterata.

ES (APPLICAZIONE DEL CONDIZIONAMENTO DI UN SISTEMA)

Scrivere la relazione che lega l'errore relativo associato alla soluzione del sistema lineare

$$A\vec{x} = \vec{b}$$

con gli errori relativi ai dati del sistema A e $\vec{b}$.
Dire quanto il sistema si definisce ben condizionato.

Partendo dal suddetto sistema lineare, s'introducano la matrice $\bar{A}$ e il vettore $\bar{b}$ definiti come

$$\begin{aligned}\bar{A} &= A + \delta A \\ \bar{b} &= \vec{b} + \delta \vec{b}\end{aligned}$$

dove δA e $\delta \vec{b}$ rappresentano la perturbazione rispettivamente di A e $\vec{b}$.
Così l'errore relativo su A e $\vec{b}$ vale

APPROSSIMAZIONE DI DATI E DI FUNZIONI

1. Studio dell'approssimazione di funzioni

In genere esistono situazioni in cui, assegnata una funzione f , questa ha espressione complessa e non facile da studiare. Pertanto bisogna trovare una funzione $\tilde{f}$ che sia vicina a f e che sia più semplice e più facile da manipolare. In tal modo, può risolvere problemi come l'integrazione imponente che

$$f \sim \tilde{f} \Rightarrow \int_a^b f(x) dx \sim \int_a^b \tilde{f}(x) dx$$

Se, infatti, la funzione f non è integrabile elementarmente, non si potrebbe applicare il teorema fondamentale del calcolo integrale e dunque si bypassa il problema cercando, a livello numerico, una funzione integrabile elementarmente. Ades. graficando la funzione f in un intervallo, si potrebbe notare che essa è vicina alla retta $y=x$ e allora si può approssimare l'integrale come

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

Dall'altra parte, anziché conoscere in maniera analitica la funzione f , si può avere a disposizione una serie di dati ottenuti da misurazioni statistiche e analitiche e non si conosce l'andamento della legge che sta al di sotto di tali dati.

Allora, scopo dell'approssimazione è scoprire la legge, cioè estendere i dati in modo da conoscere la legge che sta sotto e ricavare anche i valori che non coincidano con i dati.

2. Nell'approssimazione di una funzione occorre

- I → INDIVIDUARE UNA CLASSE DI FUNZIONI APPROSSIMANTI tra cui scegliere la funzione $\tilde{f}$, cioè bisogna trovare una funzione vicina fatta in un certo modo.
- II → definita la classe, occorre ADOTTARE UN CRITERIO e DETERMINARE IN MANIERA UNIVOCAMENTE LA FUNZIONE APPROSSIMANTE $\tilde{f}$.
- III → VALUTARE LA BONTÀ DELLA FUNZIONE APPROSSIMANTE $\tilde{f}$ e, a fare ciò, occorre individuare un criterio di valutazione.

(I) Per ipotesi, si lavorerà sempre su funzioni regolari, cioè funzioni f tali che

$$f \in C^k([a; b]) \quad \rightarrow \text{qui } k = \infty$$

Allora si hanno diverse possibili classi di funzioni approssimanti

→ **polinomi algebrici**: la funzione approssimante si scrive come polinomio, cioè come combinazione lineare (c.l.) dei monomi x^k , $k = 0, \dots, n$

$$\tilde{f}(x) = \sum_{k=0}^n a_k x^k$$

o approssimare una funzione continua su intervalli chiusi e limitati

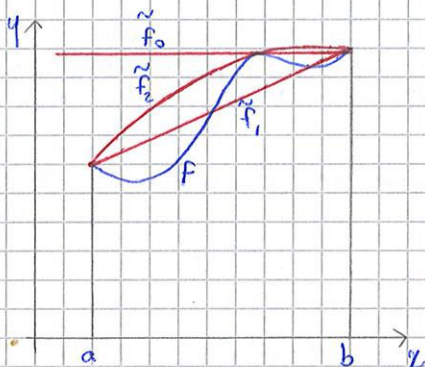
→ **somma di esponenziali di ordine n** : è c.l. a coefficienti a_k dei monomi e^{kx}

$$\tilde{f}(x) = \sum_{k=0}^n a_k e^{kx}$$

Essa si usa a funzioni con andamento di tipo esponenziale, a cui serve un'informazione di partenza sulla funzione. Se la funzione decresce, si riscrive tutto mettendo e^{-kx}

III Una volta trovata la funzione approssimante $\tilde{f}$, ci si chiede se è "buona", cioè se è abbastanza vicina alla funzione f . Per fare ciò, si ha bisogno di uno strumento e si usa la norma infinita.

$$\|f - \tilde{f}\|_{\infty} = \max_{a \leq x \leq b} |f(x) - \tilde{f}(x)|$$



Così, supponendo di avere una funzione f definita su un intervallo $[a, b]$, si può scegliere ad es. la classe dei polinomi x approssimarla. All'inizio si sceglie la costante passante x un nodo ma così si ha una bassa approssimazione, poiché la funzione f_0 è distante da f . Allora si rimane sempre nei polinomi ma si sceglie una funzione f_1 di grado 1 passante x gli estremi della funzione nell'intervallo. Così si ha un'approssimazione migliore ma non ottima. Pertanto si aumenta ancora il grado e, di conseguenza, si aumenta il numero di nodi di passaggio. In questo modo, con alti valori di grado n , si costruisce una successione di funzioni $\tilde{f}_n$ che si spera che converga a f , cioè tale che

$$\tilde{f}_n \rightarrow f$$

Così vuol dire però che la successione di funzioni $\tilde{f}_n$ converge alla funzione f ? Per definire la convergenza, infatti, bisogna definire la metrica e la convergenza. Visto che si è scelta come norma x misurare la distanza tra le funzioni f e $\tilde{f}$ la norma infinita, si può dire che la successione di funzioni approssimanti $\tilde{f}_n$ converge alla funzione f se la convergenza è data nella norma infinita, cioè se

$$\lim_{n \rightarrow \infty} \|f - \tilde{f}_n\|_{\infty} = 0$$

Criterio di interpolazione polinomiale

Un primo esempio di approssimazione di funzioni mediante polinomi è dato dallo sviluppo di TAYLOR.

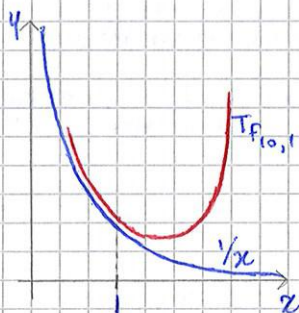
Si sa infatti che ogni funzione regolare può essere approssimata mediante uno sviluppo di Taylor, cioè un polinomio centrato in x_0 del tipo

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2 + \dots$$

Questo è un primo esempio di polinomio approssimante ma si ricorda che lo sviluppo di Taylor centrato in x_0 fornisce una buona approssimazione solo nell'intorno del punto x_0 .

Es. Si consideri la funzione

$$f(x) = \frac{1}{x}$$



e si calcoli lo sviluppo di Taylor centrato in $x_0 = 1$ e di ordine 10, indicato con $T_{f,10,1}(x)$.

Si osserva che il polinomio trovato approssima bene la funzione f nell'intorno di 1 (lo fa fino a 1,6).

Poi, man mano che ci si allontana, il discostamento si fa sempre più grande.

Per questo motivo, lo sviluppo di TAYLOR NON È un buon modo per approssimare una funzione ma comunque rappresenta un primo esempio di polinomio che approssima una funzione.

Dimostrazione

Questa dimostrazione è interessante perché, oltre a dimostrare l'esistenza e l'unicità del polinomio interpolante, introduce un primo metodo a calcolarlo e ricavarlo.

Siano dati n dati associati

$$(x_i; y_i), \quad i = 1; \dots; n$$

e si suppone che i nodi siano distinti, cioè che

$$x_i \neq x_j, \quad i \neq j$$

Ora si introduce il polinomio generico di grado $n-1$, scritto come

$$p_{n-1}(x) = c_n + c_{n-1}x + \dots + c_1x^{n-1}$$

Si osserva che tale scrittura prevede di aver assunto come base dello spazio dei polinomi P_{n-1} di grado $n-1$ la base canonica, cioè l'insieme di monomi

$$\{1; x; x^2; \dots; x^n\}$$

Tornando però al teorema, ora si impone il passaggio del polinomio x i nodi, ossia si impone l'interpolazione

$$p_{n-1}(x_i) = y_i, \quad i = 1; \dots; n$$

Ciò equivale a scrivere il sistema

$$\begin{cases} c_n + c_{n-1}x_1 + \dots + c_1x_1^{n-1} = y_1 & \rightarrow \text{condizione imposta x il nodo } i=1 \\ c_n + c_{n-1}x_2 + \dots + c_1x_2^{n-1} = y_2 & \rightarrow \text{condizione imposta x il nodo } i=2 \\ \dots & \\ c_n + c_{n-1}x_n + \dots + c_1x_n^{n-1} = y_n & \rightarrow \text{condizione imposta x il nodo } i=n \end{cases}$$

Si è così ottenuto un sistema quadrato lineare quadrato di ordine n , in cui le incognite sono i coefficienti c_i del polinomio, il termine noto è dato dal vettore delle ordinate y_i e la matrice associata al sistema è

$$A = \begin{bmatrix} x_1^{n-1} & x_1^{n-2} & \dots & x_1 & 1 \\ x_2^{n-1} & x_2^{n-2} & \dots & x_2 & 1 \\ \dots & \dots & \dots & \dots & \dots \\ x_n^{n-1} & x_n^{n-2} & \dots & x_n & 1 \end{bmatrix} \quad \rightarrow A\vec{c} = \vec{y}$$

L'esistenza e unicità del polinomio dipende dal comportamento di questo sistema e, in particolare, ciò è garantito se la matrice A è invertibile.

Si osserva però che la matrice A è la matrice di Vandermonde e, a essa, si dimostra a induzione che il determinante è pari a

$$\det A = \prod_{i>j} (x_i - x_j)$$

Poiché si è fatta l'ipotesi di nodi distinti, in cui

$$x_i \neq x_j, \quad i \neq j$$

Così si definisce il vettore equidistante z nell'intervallo d'interesse

$$z = \text{linspace}(\min(x); \max(x))$$

In tal modo si ha un vettore equidistante con spaziatura uniforme tra i due nodi estremi. Ora si valuta il valore del polinomio ricorrendo alla funzione **polyval**, con sintassi:

$$\text{polyval}(c; z)$$

c = vettore dei coefficienti che determina univocamente il polinomio

z = vettore dei punti di valutazione

In tal modo si valuta il polinomio nei vari punti (in una sola o in un vettore di 100 punti) indicati da z

Es (PROBLEMATICHE LEGATE ALLA MATRICE DI VANDERMONDE)

Si consideri la funzione analiticamente nota

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

Si vuole costruire un polinomio che interpola la funzione f in 13 punti equispaziati nell'intervallo $[0; 1]$.

Di conseguenza, si cerca di costruire 13 punti in tale intervallo con il comando

$$x = \text{linspace}(0; 1; 13)$$

Così si sono costruiti i punti $x_1, x_2, \dots, x_{13}$ e, x completare i dati, occorrono anche i valori $y_1, y_2, \dots, y_{13}$, che sono la valutazione della funzione f nei vari nodi x_i :

$$y_i = f(x_i)$$

Questi punti verranno memorizzati nel vettore y .

Ora si calcola il polinomio interpolante con il comando

$$c = \text{polyfit}(x; y; 12) \rightarrow n-1$$

A questa istruzione, MATLAB dà un avviso (un "warning") sul malcondizionamento del problema. Infatti in questo comando si vuole risolvere un sistema lineare con una matrice di Vandermonde e dunque si ha un sistema malcondizionato. Così MATLAB avvisa che il risultato può essere poco corretto, cioè a piccoli errori relativi d'input associa alti errori relativi d'output. Pertanto spesso non conviene calcolare questo polinomio con questo metodo perché, con numeri n di nodi maggiori, il numero di condizionamento $\kappa(A)$ è maggiore e magari il vettore restituito in output non ha valore corretto.

↓

dunque IL CALCOLO DEL POLINOMIO INTERPOLANTE ASSUMENDO COME BASE LA BASE CANONICA E RICORRENDO ALLA MATRICE DI VANDERMONDE È UN BUON METODO PER BASSI NUMERI n DI NODI

Si osserva che tale metodo, dal punto di vista analitico, è corretto, ma non lo è x il computer.

$$(x_2; y_2) \rightarrow p_2 \in \mathbb{P}_{n-1}$$

Tale polinomio è definita come

$$p_2(x) = \frac{(x-x_1)(x-x_3) \cdot \dots \cdot (x-x_n)}{(x_2-x_1)(x_2-x_3) \cdot \dots \cdot (x_2-x_n)} \rightarrow \text{si salta il fattore con } x_2$$

Valutando nuovamente il polinomio nei nodi, si ha che

$$x = x_1 \rightarrow p_2(x_1) = 0$$

$$x = x_2 \rightarrow p_2(x_2) = 1$$

$$x = x_3 \rightarrow p_2(x_3) = 0$$

(...)

$$x = x_n \rightarrow p_2(x_n) = 0$$

Tale polinomio, come prima, è sempre nulla tranne nel nodo x_2 a cui è associato.

Ora si prova a scrivere l'espressione generale x un polinomio p_i associato all'i-esimo nodo

$$(x_i; y_i) \rightarrow p_i \in \mathbb{P}_{n-1}$$

Tale polinomio si costruisce come

$$p_i(x) = \prod_{\substack{j=1 \\ j \neq i}}^n \frac{(x-x_j)}{(x_i-x_j)} \rightarrow \text{si toglie il fattore } x-x_i$$

Un polinomio così costruito si dice **polinomio fondamentale di Lagrange associato al nodo $(x_i; y_i)$** .
Cosa succede se si valuta questo polinomio nei vari nodi?
Si ha che

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

Dove la variabile δ_{ij} si dice **delta di Kronecker** (è unitaria quando gli indici i e j sono uguali e nulla altrimenti).

Ora, alla luce di questi nuovi concetti, si vuole scrivere l'equazione del polinomio interpolante e, x fare questo, basta sfruttare le proprietà dei polinomi fondamentali di Lagrange p_i .
Si consideri infatti il polinomio

$$p_{n-1}(x) = \sum_{i=1}^n y_i p_i(x)$$

Esso è un polinomio di grado $n-1$ ed è una c.c. dei polinomi fondamentali di Lagrange con grado $n-1$ (che sono assunti dunque come base di questo nuovo spazio dei polinomi), i cui coefficienti sono le ordinate dei nodi.

Adesso si prova a valutare tale polinomio nei vari nodi

$$x = x_1 \quad p_{n-1}(x_1) = y_1 \underbrace{p_1(x_1)}_1 + y_2 \underbrace{p_2(x_1)}_0 + \dots + y_n \underbrace{p_n(x_1)}_0 = y_1 \rightarrow \text{x costruzione (tutti i polinomi fondamentali si annullano eccetto che in } x_1)$$

$$x = x_2 \quad p_{n-1}(x_2) = y_1 p_1(x_2) + y_2 p_2(x_2) + \dots + y_n p_n(x_2) = y_2$$

Si osserva allora che LA BASE DI LAGRANGE DIPENDE DALLA SCELTA DEI NODI e così, ogni volta che si estende l'insieme dei nodi assegnati, occorre ricalcolare tutto il polinomio fondamentale di Lagrange e ciò è costoso dal punto di vista computazionale.

Alla luce di tutto questo, si introduce la **rappresentazione di Newton del polinomio d'interpolazione** (si ricorda che tale polinomio è unico ma viene costruito secondo basi diverse).

Per poi

Per poterla definire, si introducono delle quantità dette **differenze divise**.

Dati una funzione f e i nodi x_1, x_2 , si definisce la **differenza divisa di ordine 1** la quantità denotata con il simbolo

$$f[x_1; x_2]$$

e che vale

$$f[x_1; x_2] = \frac{f(x_2) - f(x_1)}{x_2 - x_1}$$

→ essa è un rapporto incrementale che però prende, in questo caso, il nome di differenza divisa

Tale definizione ha senso se $x_1 \neq x_2$.

Ora si definisce la **differenza divisa di ordine 2**, dipendente da 3 ascisse x_1, x_2 e x_3 , la quantità

$$f[x_1; x_2; x_3] = \frac{f[x_2; x_3] - f[x_1; x_2]}{x_3 - x_1}$$

→ è la differenza tra le differenze divise di ordine 1

Più in generale, la **differenza divisa di ordine $n-1$** , dipendente da n nodi x_i , vale

$$f[x_1; x_2; \dots; x_{n-1}; x_n] = \frac{f[x_2; x_3; \dots; x_{n-1}; x_n] - f[x_1; x_2; \dots; x_{n-2}; x_{n-1}]}{x_n - x_{n-1}}$$

Essa è definita usando le differenze divise di ordine $n-2$.

Ovviamente, è anche utile definire la **differenza divisa di ordine 0** il valore (le altre differenze divise sono rapporti incrementali)

$$f[x_i] = f(x_i)$$

PROPRIETÀ DELLE DIFFERENZE DIVISE:

esse sono invarianti rispetto alla permutazione degli argomenti, ad es. a ordine 1 si ha che

$$f[x_1; x_2] = f[x_2; x_1]$$

Infatti, cambia segno sia al numeratore che al denominatore e dunque il risultato non cambia. Si dimostra ricorsivamente che tale proprietà vale anche a le differenze divise di ordine maggiore di 1.

Ora si può introdurre la rappresentazione di Newton del polinomio d'interpolazione.

Scelti n nodi d'interpolazione distinti

$$(x_i; y_i), \quad x_i \neq x_j, \quad i \neq j, \quad i, j = 1, \dots, n$$

Si sa allora che esiste un unico polinomio $p_{n-1} \in \mathbb{P}_{n-1}$ (di grado $n-1$) che interpola i dati.

Ora si prova a calcolare tutte le differenze divise associate ai nodi dall'ordine 0 a $n-1$

$$f[x_i] = f(x_i)$$

Di conseguenza, attraverso le differenze divise, si è definito un polinomio

$$P_{n-1}(x) = F[x_1] + F[x_1; x_2](x-x_1) + F[x_1; x_2; x_3](x-x_1)(x-x_2) + \dots + F[x_1; \dots; x_n](x-x_1) \dots (x-x_{n-1})$$

che costituisce una nuova rappresentazione, detta **rappresentazione di Newton del polinomio d'interpolazione**.

Perché tutto questo?

Si osserva che, dati n nodi distinti, se si vuole aggiungere il nodo $n+1$ -esimo, x poterlo aggiungere bisogna costruire il polinomio di grado n passante x i n nodi distinti. Per fare ciò, però, è sufficiente aggiungere al polinomio interpolante la differenza divisa x i nodi $x_1; x_2; \dots; x_{n+1}$ moltiplicata x i fattori $(x-x_1), (x-x_2), \dots, (x-x_n)$

$$P_n(x) = F[x_1] + \dots + F[x_1; x_2; \dots; x_n; x_{n+1}](x-x_1)(x-x_2) \dots (x-x_n)$$

Cioè è sufficiente aggiungere un termine in più a questa sommatoria e tutto ciò che è stato calcolato prima non deve essere modificato. Dunque

Dunque, aggiungendo un nodo, occorre solo calcolare una differenza divisa e aggiungere solo un termine in più alla sommatoria.

Ciò ha conseguenze sull'efficienza computazionale perché, aggiungendo un nodo in più x aumentare la precisione d'interpolazione, non occorre ricalcolare tutto da capo ma basta solo aggiungere il calcolo di una differenza divisa e inserire un nuovo addendo alla sommatoria.

Si osserva poi che, a questo punto, si ha la rappresentazione del polinomio interpolante come combinazione lineare di alcuni elementi

- $F[x_1]$ - Il coefficiente $F[x_1]$ moltiplica 1
- $F[x_1; x_2](x-x_1)$ - La differenza divisa di ordine 1 moltiplica il polinomio $(x-x_1)$, di grado 1
- $F[x_1; x_2; x_3](x-x_1)(x-x_2)$ - La differenza divisa di ordine 2 moltiplica il polinomio $(x-x_1)(x-x_2)$, di grado 2
- $F(\dots)$
- $F[x_1; \dots; x_n](x-x_1) \dots (x-x_{n-1})$ - La differenza divisa di ordine $n-1$ moltiplica il polinomio $(x-x_1) \dots (x-x_{n-1})$, di grado $n-1$

Dunque l'insieme

$$\{1; x-x_1; x-x_2; \dots; x-x_{n-1}\}$$

costituisce la **base fondamentale di Newton** x rappresentare i polinomi di grado minore o uguale a $n-1$.

In tal modo, si è riscritto il polinomio cambiando nuovamente la base, passando dalla base canonica (o la rappresentazione di Vandermonde) alla base dei polinomi fondamentali di Lagrange alla base di Newton.

Si osserva, a tal proposito, che, al livello di coefficienti, la rappresentazione più semplice da calcolare è quella di Lagrange, mentre nella rappresentazione di Newton occorre calcolare le differenze divise.

Per questo motivo, si descrive la cosiddetta **tabella delle differenze divise**, che costituisce una tecnica mnemonica per il calcolo delle differenze finite. Essa si costruisce

Es (APPLICAZIONE DELLA TABELLA DELLE DIFFERENZE DIVISE)

Si vuole scrivere il polinomio d'interpolazione x i nodi

$$(0; 1) \quad (1; -1) \quad (2; 1) \quad (\frac{1}{2}; 2)$$

Si costruisce la tabella, ricordandosi che nella I colonna vanno le ascisse x_i e nella II vanno le ordinate y_i

x_i	y_i					
0	1					
1	-1	$\frac{-1-1}{1-0} = -2$				
2	1	$\frac{1-(-1)}{2-1} = 2$	$\frac{2-(-2)}{2-0} = 2$			
$\frac{1}{2}$	2	$\frac{2-1}{\frac{1}{2}-2} = -\frac{2}{3}$	$\frac{-\frac{2}{3}-2}{\frac{1}{2}-1} = \frac{16}{3}$	$\frac{\frac{16}{3}-2}{\frac{1}{2}-0} = \frac{20}{3}$		

Ora si considerano gli elementi sulla diagonale principale e si costruisce il polinomio interpolante, che ha grado 3 (poiché i nodi sono 4)

$$p_3(x) = 1 - 2(x-0) + 2(x-0)(x-1) + \frac{20}{3}(x-0)(x-1)(x-2)$$

RIEPILOGO SUI METODI D'INTERPOLAZIONE POLINOMIALE:

dati n nodi distinti, esiste un unico polinomio passante x essi e si hanno 3 possibili rappresentazioni

→ secondo BASE CANONICA, dove i polinomi fondamentali sono $1; x; x^2; \dots; x^{n-1}$ e il polinomio è ricostruito risolvendo il sistema di VanderMonde

→ secondo BASE DI LAGRANGE, in cui si costruiscono n polinomi e i coefficienti sono le ordinate y_i dei nodi.

Tale metodo si evita quando, dato un insieme di nodi, si vuole aggiungere dei nodi al set di interpolazione

→ secondo BASE DI NEWTON, dove i polinomi fondamentali sono $1; (x-x_1); \dots; (x-x_{n-1})$ e i coefficienti sono le differenze divise, calcolate costruendo una tabella.

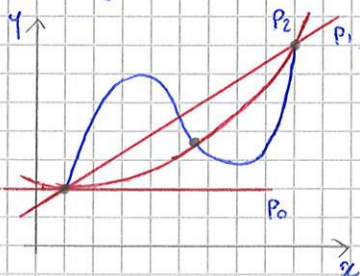
5 Convergenza dei metodi d'interpolazione polinomiale:

ora si studia il problema di convergenza che il polinomio ha rispetto a una funzione f .

Data una funzione f non nota (di cui ad es si vuole calcolare la derivata), si sceglie una funzione più semplice, cioè il polinomio interpolante, con cui la si vuole approssimare.

Si vuole poi migliorare l'approssimazione aumentando il numero di nodi (e dunque aumentando il grado del polinomio interpolante) e una così si aspetta che, aumentando i passaggi x punti, più punti si danno, più si avrà un polinomio che ben approssima la funzione.

Si immagina cioè che, dando più punti di passaggio (più nodi), l'approssimazione sia migliore.



Dando infatti un solo punto e costruendo il polinomio g_0 interpolante la funzione, si ha una brutta approssimazione della funzione. Si aggiunge allora un punto e si costruisce il polinomio interpolante g_1 e l'approssimazione è migliore.

Si danno poi 3 nodi e si costruisce il polinomio g_2 che approssima la funzione e così via.

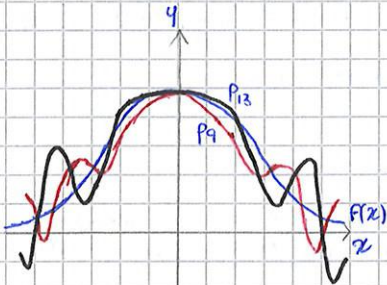
Dando ancora più punti di passaggio, ci si aspetta che il polinomio si avvicini sempre di più alla funzione.

Es (FENOMENO DI RUNGE)

Si consideri la cosiddetta **funzione di Runge**, definita come

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

Si vuole studiare il problema di interpolazione nell'intervallo $[-5; 5]$ su nodi equispaziati



La funzione f considerata è molto regolare (addirittura è analitica) e in sua grafica è nota.

Nonostante ciò, però, si osserva che la successione di polinomi interpolanti che si costruisce non converge alla funzione all'aumentare del numero di nodi d'interpolazione.

Tracciando infatti il polinomio di grado 9 e quello di grado 13 che interpolano la funzione (si passa da 10 a 14 nodi equispaziati), si nota che polinomio e funzione hanno sempre più punti in comune ma le oscillazioni aumentano.

Più in generale, pur aumentando i vincoli di passaggio, anziché avere un avvicinamento, si osserva che c'è un fenomeno di oscillazione agli estremi dell'intervallo di interpolazione $[-5; 5]$ e tali oscillazioni aumentano sempre di più all'aumentare del numero dei nodi d'interpolazione (cioè all'aumentare del grado del polinomio).

Così, anche se passa a più nodi, il polinomio interpolante non converge alla funzione stessa e, considerando il massimo dell'errore (con la norma infinito) si ha che

$$\lim_{n \rightarrow \infty} \|E_n(x)\|_{\infty} = +\infty$$

Questo fenomeno accade perché, anche se molto frequente, la scelta di nodi equispaziati non si è rivelata una buona scelta.

A tal proposito, si dimostra che, scegliendo nodi equidistanti, esiste sempre una funzione f per la quale l'errore d'interpolazione non tende a 0, per qualunque scelta dell'intervallo d'interpolazione.

Si osserva però che la scelta di nodi equispaziati non è pessima perché è buona se la funzione f rispetta certe proprietà, infatti, esiste un risultato importante:

se, scegliendo nodi d'interpolazione equispaziati in un intervallo $[a; b]$, in cui si vuole approssimare la funzione f , tale funzione f è regolare rispetto alla variabile complessa, cioè è analitica in una regione che contiene l'intervallo $[a; b]$, e se la distanza dei poli della funzione f dall'intervallo $[a; b]$ è maggiore dell'ampiezza dell'intervallo $[a; b]$ stesso, allora il polinomio interpolante $p_n(x)$ converge in $[a; b]$.

In parole povere, il polinomio interpolante converge se

→ la funzione è ANALITICA in una regione che contiene l'intervallo d'interpolazione $[a; b]$

→ la distanza dei poli (= punti dove si annulla il denominatore) è maggiore dell'ampiezza $b-a$ dell'intervallo

$$d_{\text{poli}} > b-a$$

Se tali condizioni sono soddisfatte, allora, a la particolare scelta dei nodi (nodi equispaziati), l'errore tende a 0.

Es (FENOMENO DI RUNGE)

Si consideri nuovamente la funzione di Runge

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

Si vuole approssimarla nell'intervallo $[1; 2]$ e, a tal proposito, si osserva che, disegnando

$$\lim_{n \rightarrow \infty} \|E_n(x)\|_{\infty} = 0 \quad \rightarrow \text{l'errore tende uniformemente a 0}$$

Il problema però è che tali nodi sono definiti solo nell'intervallo $(-1; 1)$ e solo lì si ha un risultato di convergenza, a cui si sarebbe vincolati a lavorare su tale intervallo a qualunque funzione f considerata.

In realtà non è così perché, scelto un qualunque intervallo $[a; b]$, si può definire la trasformazione tra intervalli

$$[-1; 1] \rightarrow [a; b]$$

In questa trasformazione, si associa a ogni valore $z \in [-1; 1]$ un valore x definito come

$$x = \left(\frac{b-a}{2}\right)z + \frac{b+a}{2}$$

Essa è una trasformazione lineare e simata che

$$\rightarrow xz = -1 \text{ si ha } x = a$$

$\rightarrow$ tale trasformazione manda l'intervallo $[-1; 1]$ nell'intervallo $[a; b]$

$$\rightarrow xz = 1 \text{ si ha } z = b$$

Di conseguenza, se si hanno i nodi di Chebyshev

$$z_i = -\cos\left(\frac{2i-1}{2(n+1)}\pi\right) \in (-1; 1)$$

allora i nodi x_i valgono

$$x_i = \left(\frac{b-a}{2}\right)z_i + \frac{b+a}{2} \in (a; b)$$

Questi sono nodi definiti nell'intervallo $(a; b)$ e sono i **nodi di Chebyshev nell'intervallo $(a; b)$** (sono traslati in tale intervallo).

Così, a ogni scelta dell'intervallo di approssimazione $[a; b]$, si possono definire tali nodi anche su tale intervallo.

Ovviamente, la conseguenza di tutto questo è che

se la funzione f è regolare nell'intervallo $[a; b]$, cioè tale che

$$f \in C'([a; b])$$

allora è garantita la convergenza uniforme del polinomio interpolante della funzione sui nodi x_i , derivanti dai nodi di Chebyshev per trasformazione lineare, all'aumentare del numero dei nodi (cioè all'aumentare del grado del polinomio interpolante).

Così, con una funzione regolare, si è sicuri che, con la scelta dei nodi di Chebyshev, il polinomio interpolante converge.

Se inoltre la funzione è ancora più regolare (cioè è analitica e non presenta poli), anche la scelta di nodi equispaziati si rivela conveniente.

ES

Tornando alla funzione di Runge, se nell'intervallo $[-5; 5]$ si considerano nodi non equispaziati (dove il polinomio interpolante non converge affatto) ma nodi di Chebyshev, allora si elimina il fenomeno di oscillazione agli estremi dell'intervallo ed è garantita la convergenza uniforme del polinomio interpolante alla funzione, poiché essa è di classe C' nell'intervallo $[-5; 5]$.

Così il polinomio interpolante vale

$$p_2(x) = y_2 l_2(x) = -\frac{4x(x-\pi)}{\pi^2}$$

ES (CASO PARTICOLARE)

Dati 5 nodi equispaziati nell'intervallo $[-1; 1]$, determinare il polinomio che interpola i punti

$$(x_i; x_i^2 - 1)$$

Poiché si hanno 5 nodi, bisogna determinare un polinomio p_4 di grado minore o uguale a 4. Si osserva, però, che le ordinate dei dati d'interpolazione sono la valutazione della funzione

$$f(x) = x^2 - 1$$

Si sa però che l'unico polinomio che interpola una funzione polinomiale su questi 5 nodi, a unicità, è il polinomio stesso, a cui

$$f(x) \equiv p_4(x) = x^2 - 1$$

Ora, però, in particolare ci si vuole occupare, in questo ambito, di approssimazione di dati o funzioni sempre x interpolazione (cioè imponendo il passaggio x punti) una scegliendo un polinomio di grado d su ogni sottointervallo determinata dalla partizione, in cui il GRADO d LOCALE È PICCOLO e, al massimo, si ha

$$d = 3$$

Dunque si useranno funzioni lineari, quadratiche e di grado 3 a tratti.

In particolare, ci si occuperà di una specifica categoria di funzioni cubiche dette **funzioni spline**:

sono funzioni di grado 3 a tratti che soddisfanno particolari condizioni di regolarità nei punti di raccordo, cioè sono globalmente di classe C^2 . Di conseguenza, le funzioni spline sono continue, derivabile con derivata seconda continua sui nodi di partizione.

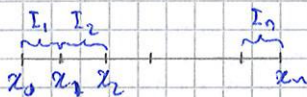
Si capisce allora perché non si tratteranno ad es. funzioni lineari a tratti perché non sono derivabili.

Si osserva, infatti, che con un'approssimazione mediante spline, si evitano fenomeni di oscillazione, presenti invece nel caso di approssimazione con polinomio globale. Essa ciò accade perché ci sono ottime proprietà di convergenza di questa funzione approssimante a una funzione assegnata e, infatti, questa metodo è molto usato nelle applicazioni grafiche, in cui le oscillazioni "sporcano" l'immagine.

2 Funzioni spline cubiche

Per costruirle, si suppone di partire da un insieme di dati assegnati, cioè un insieme di nodi d'interpolazione (x_i, y_i) , $i=0, \dots, n$ distinti, cioè tali che

$$x_i \neq x_j, \quad i \neq j$$



Così si hanno $n+1$ dati d'interpolazione. Questi nodi determinano degli intervalli chiamati come

$$I_i = [x_{i-1}, x_i], \quad i=1, \dots, n$$

Così, dati $n+1$ punti, questi determinano n sottointervalli dell'intervallo $[x_0, x_n]$.

Ora, su ciascuno di questi sottointervalli, si vuole definire il polinomio di grado 3 (quindi in tutto si avrà un polinomio che localmente ha grado 3) che interpoli i dati assegnati. Si denota così con il simbolo $S_3(x)$ la funzione approssimante, che prende il nome di **funzione spline cubica**.

Per costruirla, sulla base della definizione di funzione spline (che dev'essere soddisfatta -v.b.), si impongono 5 condizioni

① "polinomialità":

La funzione dev'essere un polinomio di grado 3 su OGNI SOTTOINTERVALLO. Così, a iniziare, si scrive $S_3(x)$ come un generico polinomio di grado 3, cioè come

$$S_3(x) = a + bx + cx^2 + dx^3$$

Si vuole però che la funzione $S_3(x)$ sia un polinomio di grado 3 non globalmente, ma ristretto a un intervallo generico, cioè si vuole che la funzione assuma questa espressione su ogni $x \in I_i$, $i=1, \dots, n$.

Di conseguenza, ai coefficienti del polinomio si mette un indice i , in quanto non è detto che questi siano gli stessi in tutti i sottointervalli.

Si hanno n intervalli, x cui in tutte le incognite sono $4n$.

Quante condizioni si impongono a determinare il polinomio?

Nell'interpolazione (condizione $\textcircled{V}$), s'impone una condizione a ogni nodo e da qui si hanno $n+1$ condizioni d'interpolazione.

Nella condizione di regolarità (condizioni $\textcircled{II}$, $\textcircled{III}$ e $\textcircled{IV}$), si chiede la continuità della funzione, della derivata prima e della derivata seconda nei nodi interni. I nodi interni sono $n-1$, x cui si hanno $n-1$ condizioni a la continuità della sola funzione, della derivata prima oppure della derivata seconda. Così, in totale, si hanno $3(n-1)$ condizioni a la regolarità.

Faccendo dunque un bilancio del problema, si hanno

→ $4n$ incognite

→ $n+1+3(n-1) = 4n-2$ condizioni imposte

Dunque il sistema tra condizioni e coefficienti non è bilanciato perché si hanno meno equazioni che incognite e, pertanto, bisogna aggiungere ancora 2 condizioni aggiuntive in modo da determinare univocamente la spline.

Queste 2 condizioni sono scelte in 3 modi diversi e, a seconda della scelta di tali condizioni fatta a definire questo polinomio interpolante localmente di grado 3 e di classe C^2 , si hanno diverse tipologie di spline.

→ se s'impongono come 2 ulteriori condizioni che la derivata seconda nel I e nell'ultimo nodo della partizione sia nulla, cioè che

$$S_3^{(2)}(x_0) = 0$$

$$S_3^{(2)}(x_n) = 0$$

allora la funzione spline ottenuta si dice **spline cubica naturale**.

Si osserva che il definire una spline cubica naturale prevede di dire tutte le condizioni dettate nei punti $\textcircled{I}$, $\textcircled{II}$, $\textcircled{III}$, $\textcircled{IV}$, $\textcircled{V}$ e $\textcircled{VI}$ e il nome "naturale" si lega alle due condizioni finali imposte.

→ le due condizioni finali consistono nell'imporre che la derivata terza della funzione e spline nel primo e nell'ultimo nodo interno (cioè nel $\textcircled{II}$ e nel penultimo nodo) della partizione sia continua, cioè che la funzione spline sia di classe C^3 nei nodi 1 e $n-1$.

$$S_3^{(3)}(x_1^+) = S_3^{(3)}(x_1^-)$$

$$S_3^{(3)}(x_{n-1}^+) = S_3^{(3)}(x_{n-1}^-)$$

In questo caso si parla di **spline cubica di tipo not a knot** (= "non un nodo").

→ le due ultime condizioni sono che la derivata prima della funzione spline calcolata nei nodi della partizione x_0 e x_n coincida con la derivata prima della funzione f nei nodi x_0 e x_n

$$S_3'(x_0) = f'(x_0)$$

$$S_3'(x_n) = f'(x_n)$$

S'impone dunque che la pendenza agli estremi dell'intervallo d'interpolazione della spline e della funzione che si vuole interpolare sia la stessa.

Si osserva che quest'ultima condizione ha senso solo se si ha una funzione da approssimare e dunque, se si ha un problema di approssimazione di dati, non la si sceglie.