ACCURATEZZA DI UN RISULTATO

Errore assoluto : $E_a = |\text{risapprossimato} - \text{risesatto}|$

Errore relativo : $E_r = \frac{|E_a|}{|\text{risesatto}|}$ $\quad |\text{risesatto}| \neq 0$

Errore percentuale : $E_p = (E_r \cdot 100)\%$

RAPPRESENTAZIONE DEI NUMERI IN MEMORIA

Vengono rappresentati in forma binaria (base 2)

Fissato un numero $\beta \in \mathbb{N}$, con $\beta > 1$, rappresentiamo nella base β un numero reale qualunque $x \in \mathbb{R}$

$$x = \pm (d_0 d_1 d_2 \dots d_p . d_{p+1} \dots)_\beta = \sum_{k=0}^{p} d_k \beta^{p-k} + \sum_{k=1}^{\infty} d_{k+p} \beta^{-k}$$

NORMALIZZATA $\quad x = \pm (.d_1 d_2 \dots)\beta^p$

NUMERI INTERI

Rappresentazione in modulo e segno

Il modo più semplice per rappresentarli in memoria è quello di anteporre un bit da valore 0 se è positivo, 1 se negativo

Con N bit si possono rappresentare tutti i numeri da $[-(2^{N-1}-1), 2^{N-1}-1]$

NUMERI REALI

Si usa il metodo delle moltiplicazioni successive

$$\begin{cases} 0.2 \times 2 = 0.4 \\ 0.4 \times 2 = 0.8 \\ 0.8 \times 2 = 1.6 \\ 0.6 \times 2 = 1.2 \\ 0.2 \times 2 = 0.4 \end{cases} \qquad \begin{matrix} \text{parte intera} \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{matrix} \downarrow$$

$$(0.2)_{10} = (0.\overline{00110})_2$$

Rappresentazione scientifica normalizzata di un numero reale:

ogni $x \in \mathbb{R}$ può essere espresso come $\quad x = \pm (0.d_1 d_2 \dots)\beta^p = \pm \sum_{i=1}^{\infty} (d_i \beta^{-i})\, \beta^p$ $\quad \begin{cases} p \in \mathbb{N} \\ 0 \leq d_i \leq \beta - 1 \\ d_1 \neq 0 \end{cases}$

il numero $0.d_1 d_2 d_3 \dots$ viene detto mantissa di x, β^p è la parte esponente

p è detto caratteristica di x (o esponente), β è la base, d_i sono le cifre di rappresentazione

SISTEMA FLOATING POINT

Si definisce un insieme dei numeri macchina (floating-point) con t cifre significative, base β e range (L, U)

$$F(\beta, t, L, U) = \{0\} \cup \{x \in \mathbb{R} = \text{sign}(x)\, \beta^p \sum_{i=1}^{t} d_i \beta^{-i}\} \qquad \begin{bmatrix} t, \beta \in \mathbb{N} & \beta \geq 2 \\ 0 \leq d_i \leq \beta - 1 & d_1 \neq 0 \\ & L \leq p \leq U \end{bmatrix}$$

L ed U sono dello stesso ordine di grandezza.

t rappresenta il numero di cifre della mantissa (la precisione)

- $F(2, 24, -128, 127)$ precisione singola, 32 bit di cui 24 per la mantissa, 8 all'esponente
- $F(2, 53, -1024, 1023)$ precisione doppia, 64 bit di cui 53 per la mantissa, 11 per l'esponente

ERRORI DI RAPPRESENTAZIONE

Errore assoluto di arrotondamento: $|fl(x) - x| < \beta^{p-t}$

Errore relativo di arrotondamento: $\frac{|fl(x) - x|}{|x|} < \frac{1}{2}\beta^{1-t}$

$exp = \frac{1}{2}\beta^{1-t}$ è detta precisione macchina

exp è il più piccolo numero macchina positivo tale che $fl(1+exp) > 1$

Poiché $F \subseteq \mathbb{R}$, le operazioni sono definite anche per operandi in F, ma soltanto il risultato $\notin F$, quindi il risultato esatto è arrotondato ad un numero che sta in F

operazione aritmetica reale $\cdot : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$

operazione floating-point $\odot : F \times F \to F$ $\quad x \odot y = fl(x \cdot y)$

NB ogni operazione provoca un errore molto piccolo detto di arrotondamento $\left|\frac{(x \odot y) - (x \cdot y)}{x \cdot y}\right| < exp$

ERRORI NEL CALCOLO NUMERICO

di misura: dovuto alle imperfezioni dello strumento di misura dei dati del problema

di troncamento: quando un procedimento infinito viene realizzato come finito

algoritmico: dovuto al propagarsi di errori di arrotondamento nelle singole operazioni in un procedimento complesso

inerente: i dati del problema non sempre appartengono ad F e quindi vengono approssimati

RISOLUZIONE DI SISTEMI LINEARI

TEOREMA: Sia A una matrice $n \times n$, allora sono equivalenti:

1. $Ax = 0$ ha una sola soluzione: $x = 0$
2. $\forall$ vettore b, $Ax = b$ ha una sola soluzione
3. A è non singolare (determinante $\neq 0$)

Per risolvere un sistema lineare, possiamo adottare 2 tipi di metodi:

- DIRETTI: soluzione calcolata in un certo numero di passi finiti
- ITERATIVI: calcolo di una soluzione come approssimazione di una successione x_k (Adatti a sistemi grandi)

Sistema triangolare superiore

$$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ 0 & a_{2,2} & \cdots & \vdots \\ \vdots & & & \\ 0 & \cdots & 0 & a_{m,m} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$$

partendo dal basso

$$a_{m,m} \cdot x_m = b_m \rightarrow x_m = b_m / a_{m,m}$$

Sostituzioni successive verso l'alto

Complessità computazionale $O(m^2/2)$

Sistema triangolare inferiore

$$\begin{pmatrix} a_{1,1} & 0 & \cdots & 0 \\ \vdots & a_{2,2} & & \vdots \\ & \vdots & \ddots & 0 \\ a_{m,1} & \cdots & & a_{m,m} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$$

$$a_{1,1} \cdot x_1 = b_1 \rightarrow x_1 = b_1 / a_{1,1}$$

Sostituzioni successive verso il basso

Metodo di eliminazione di Gauss

Si eliminano le incognite in modo sistematico per trasformare il sistema lineare in uno equivalente a scala con forma triangolare superiore.

Costo computazionale $O(m^2/2)$

$$Ax = b \xrightarrow{\text{GAUSS}} Rx = y$$

dove R è triangolare superiore

Fattorizzazione LR

Operazioni necessarie a trasformare A nella matrice triangolare superiore R. $A = \begin{pmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{pmatrix}$

PASSO 1

Costruisco $L_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ -4 & 0 & 1 & 0 \\ -3 & 0 & 0 & 1 \end{pmatrix}$ tutti 1 sulla diagonale. Nella prima colonna c'è il risultato della divisione di tutti i termini della colonna con il primo termine, cambiato di segno

Si moltiplica L_1 per A

$$\begin{pmatrix} 1 & & & \\ -2 & 1 & & \\ -4 & & 1 & \\ -3 & & & 1 \end{pmatrix} \cdot \begin{pmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 3 & 5 & 5 \\ 0 & 4 & 6 & 8 \end{pmatrix}$$

$$L_1 \cdot A = A_2$$

PASSO 2

Costruisco L_2 nello stesso modo, partendo da A_2. $L_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -3 & 1 & 0 \\ 0 & -4 & 0 & 1 \end{pmatrix}$

Moltiplico L_2 per A_2 e ottengo $A_3 = \begin{pmatrix} 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 2 & 4 \end{pmatrix}$

PASSO 3

Costruisco L_3 nello stesso modo $L_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}$ e le moltiplico per A_3, ottenendo

$$R = \begin{pmatrix} 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 2 \end{pmatrix}$$

$L_3 \cdot L_2 \cdot L_1 \cdot A = R$ DI CONSEGUENZA: $A = LR$ con $L = L_1^{-1} \cdot L_2^{-1} \cdot L_3^{-1}$

calcolando le inverse di L_1, L_2, L_3 e moltiplicandole, ottengo $L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 3 & 1 & 0 \\ 3 & 4 & 1 & 1 \end{pmatrix}$

La complessità della fattorizzazione LR è $O(n^3/3)$

Non sempre la fattorizzazione LR è fattibile.

$A = \begin{pmatrix} 10^{-20} & 1 \\ 1 & 1 \end{pmatrix} = L \cdot R$ Con la fattorizzazione LR trovo

$L = \begin{pmatrix} 1 & 0 \\ 10^{20} & 1 \end{pmatrix}$ $R = \begin{pmatrix} 10^{-20} & 1 \\ 0 & 1-10^{20} \end{pmatrix}$ → Se sappiamo $F(2,53,-1024,1023)$ Con precisione 10^{-16}, perdo l'1

R diventa $\tilde{R} = \begin{pmatrix} 10^{-20} & 1 \\ 0 & -10^{20} \end{pmatrix}$ mentre $L = \tilde{L}$

$\tilde{L} \cdot \tilde{R} = \begin{pmatrix} 10^{-20} & 1 \\ 1 & 0 \end{pmatrix}$ quindi $\tilde{L}, \tilde{R}$ non sono la fattorizzazione LR di A

OSSERVAZIONI

In alcune matrici non si riesce a calcolare la fattorizzazione LR. Anche se $\tilde{L}, \tilde{R}$ sono vicine ad LR, i loro prodotti posso essere molto distanti

$\tilde{L} = L + \Delta L$ $\tilde{R} = R + \Delta R$ $\Delta L, \Delta R$ errori

$\tilde{L} \cdot \tilde{R} = (LR + L\Delta R + \Delta L R + (\Delta L \Delta R)) \neq A = A + \Delta A$

LR → A; $(\Delta L \Delta R)$ → Errore piccolo e trascurabile

La fattorizzazione LR è un algoritmo INSTABILE.

Si introduce quindi la fattorizzazione LR con PIVOT

pivot $\begin{pmatrix} x & x & x & \dots & x \\ a_{k,k} & x & \dots & x \\ x & x & \dots & x \\ \vdots & \vdots & & \vdots \\ x & x & \dots & x \end{pmatrix}$ Se $a_{k,k} = 0$ scelgo con $a_{j,k} \neq 0$ con $j > k$ e scambio le righe

PIVOTING PARZIALE

Nello scambiare le righe, scelgo come pivot nella colonna sottostante l'elemento max in valore assoluto → ho più stabilità

$$A = \begin{pmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{pmatrix}$$

$$\underbrace{\begin{pmatrix} & & 1 & \\ & 1 & & \\ 1 & & & \\ & & & 1 \end{pmatrix}}_{P_1} \cdot \underbrace{\begin{pmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{pmatrix}}_{A} = \underbrace{\begin{pmatrix} 8 & 7 & 9 & 5 \\ 4 & 3 & 3 & 1 \\ 2 & 1 & 1 & 0 \\ 6 & 7 & 9 & 8 \end{pmatrix}}_{A_1}$$

Si crea la matrice P che chiamata di permutazione. È la matrice identità con le righe che devo scambiare invertite

A questo punto creo L_1 nello stesso modo della fattorizzazione LR

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 & 0 \\ -\frac{1}{4} & 0 & 1 & 0 \\ -\frac{3}{4} & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 8 & 7 & 9 & 5 \\ 4 & 3 & 3 & 1 \\ 2 & 1 & 1 & 0 \\ 6 & 7 & 9 & 8 \end{pmatrix} = L_1 \cdot A_1 = A_2$$

Proseguo fino ad ottenere $L_3 \cdot A_3 = R$

Sistema lineare

$O\left(\frac{m^3}{3}\right)$

$Ax=b \quad \Rightarrow PAx = Pb$ con P matrice di permutazione

$$LRx = Pb \quad \Rightarrow \begin{cases} Ly = Pb \\ Rx = y \end{cases}$$

Qualunque matrice A non singolare ammette fattorizzazione LR con pivoting parziale

Caso generale $Ax=b$

Calcolo P, L, R tale che $PA = LR$

$PA = Pb$ quindi $z = Pb$

Sistema Triangolare inferiore $Ly = z$

Sistema Triangolare superiore $Rx = y$

NORME

Una norma è una funzione $\|\cdot\| : \mathbb{C}^m \to \mathbb{R}$ che assegna un valore reale ≥ 0 (lunghezza) a ogni vettore

$\forall x, y \in \mathbb{C}^m, \forall \alpha \in \mathbb{C}$

- $\|x\| \geq 0$ e $\|x\| = 0$ se $x = \vec{0}$
- $\|x+y\| \leq \|x\| + \|y\|$
- $\|\alpha x\| = |\alpha| \cdot \|x\|$
- $\|x\|_p = \left(\sum_{i=1}^{m} |x_i|^p\right)^{1/p}$ con $1 \leq p \leq \infty$ → NORMA EUCLIDEA: $\|x\|_2 = \left(\sum_{i=1}^{m} |x_i|^2\right)^{1/2}$
- $\|x\|_\infty = \max_{1 \leq i \leq m} |x_i|$
- $\|x\|_1 = \sum |x_i|$

Per i vettori $\|x\|_1 \geq \|x\|_2 \geq \|x\|_\infty$

NORMA DI MATRICI

- $\|AB\| \leq \|A\| \, \|B\|$
- $\|I\| = 1$
- $\|A\| \geq 0$, $\|A\| = 0 \Leftrightarrow A = 0$
- $\|A + B\| \leq \|A\| + \|B\|$
- $\|\alpha A\| = |\alpha| \cdot \|A\|$

- $\|A\|_1 = \max_{1 \leq j \leq m} \sum_{i=1}^{m} |a_{ij}|$

Sommo tutti gli elementi in val assoluto di una colonna → faccio la somma massima tra tutte le colonne $A = \begin{pmatrix} -1 & 2 \\ 3 & -4 \end{pmatrix}$ 1° colonna → 4 2° colonna → 6

- $\|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^{m} |a_{ij}|$

Come quella precedente, ma sommo gli elementi delle righe

$A = \begin{pmatrix} -1 & 2 \\ 3 & -4 \end{pmatrix}$ 1° riga → 3 2° riga → 7

- NORMA DI FROBENIUS

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{m} (a_{ij})^2}$$

Somma di tutti gli elementi al quadrato sotto radice

$A = \begin{pmatrix} -1 & 2 \\ 3 & -4 \end{pmatrix}$ $\|A\|_F = \sqrt{(-1)^2 + (2)^2 + (3)^2 + (-4)^2} = \sqrt{30}$

- $\|A\|_2 = \sqrt{\rho(A^T \cdot A)}$

dove ρ indica il raggio spettrale di una matrice = max autovalore in modulo

$\rho(A^T A)$ = max autovalore in modulo della matrice prodotto $A^T \cdot A$

PROBLEMA BEN CONDIZIONATO

errore dati → errore risultati

A piccole variazioni dei dati, corrispondono piccole variazioni dei risultati $\Delta r \approx \Delta w$

Sia x^* tale che $Ax^* = b$. Si studia la soluzione del sistema perturbato

$$(A + \Delta A)\,\tilde{x} = b + \Delta b \qquad \tilde{x} = x^* + \Delta x$$

Il condizionamento è legato all'errore INERENTE (causato dall'errore di rappresentazione o sui dati $\approx 10^{-16}$)

E_I : errore inerente , E_R : errore rappresentazione

Se $E_I >> E_R$ molto più grande in ordine di grandezza → Problema mal condizionato

Se $E_I > E_R$ è più o meno lo stesso ordine di grandezza → Problema ben condizionato

Dati $Ax = b$ e $(A + \Delta A)(x + \delta x) = b + \delta b$, per stimare l'errore inerente $E_r = \frac{\|\delta x\|}{\|x\|}$

Supponiamo di perturbare solo il termine noto

$\Delta A = 0$ e $\delta b \neq 0$

$$A x = b \quad (1) \qquad A(x+\delta x) = b + \delta b \quad (2)$$

$2 - 1 = A\delta x = \delta b$

$$\|A\| \, \|\delta x\| = \|\delta b\|$$

$$\|\delta x\| = \|A^{-1} \cdot \delta b\| \leq \|A^{-1}\| \cdot \|\delta b\|$$

unisco

D'altra parte $Ax = b \rightarrow \|b\| = \|Ax\| \leq \|A\| \, \|x\|$

$$\Rightarrow \frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}$$

e ottengo

$$\frac{\|\delta x\|}{\|x\|} \leq \|A^{-1}\| \, \|A\| \, \frac{\|\delta b\|}{\|b\|}$$

numero di condizionamento

$\frac{\|\delta b\|}{\|b\|} \rightarrow$ ordine $\approx 10^{-16}$

confronto tra errore inerente relativo $\frac{\|\delta x\|}{\|x\|}$ e l'errore inerente sul termine noto b $\frac{\|\delta b\|}{\|b\|}$

Perturbazione delle sola matrice ($\delta b = 0$) $\quad Ax = b \quad (A + \Delta A)(x + \delta x) = b$

$$Ax = (A+\Delta A)(x+\delta x) = A(x+\delta x) + \Delta A(x+\delta x) \rightarrow \delta x = A^{-1}\Delta A(x+\delta x)$$

$$\frac{\|\delta x\|}{\|x+\delta x\|} \leq \|A^{-1}\| \cdot \|A\| \cdot \frac{\|\Delta A\|}{\|A\|}$$

Perturbazione di matrice e termine noto

errore relativo sui dati che sul termine noto

$$\frac{\|\delta x\|}{\|x-y\|} \leq 2\delta \, \|A^{-1}\| \cdot \|A\| \, \frac{1}{(1-\alpha)}$$

legato all'errore sui dati $\delta > 0 \quad \frac{\|\Delta A\|}{\|A\|} \leq \delta \quad \frac{\|\delta b\|}{\|b\|} \leq \delta$

NUM DI CONDIZIONE

$K(A) = \|A\| \cdot \|A^{-1}\|$ → A deve essere quadrata e NON singolare

$$\frac{1}{K(A)} = \min_B \frac{\|A-B\|}{\|A\|}$$

distanza tra la matrice A e una matrice B singolare

distanza tra le 2 matrici

NB: quando K(A) alto, A "inizia a comportarsi" come una matrice singolare

RISOLUZIONE SISTEMI LINEARI : METODI ITERATIVI

$Ax = b$ costruisco una successione di vettori $\{\vec{x_1}, \vec{x_2}, \dots\}$ tale che $x_k \xrightarrow{k \to \infty} x^*$ soluzione del sistema

Schema algoritmo iterativo:
1. Dati: x_0
2. k=1
3. Ripeti finchè convergenza
   - 3.1 $x_k = G(x_{k-1})$
   - 3.2 $k = k + 1$

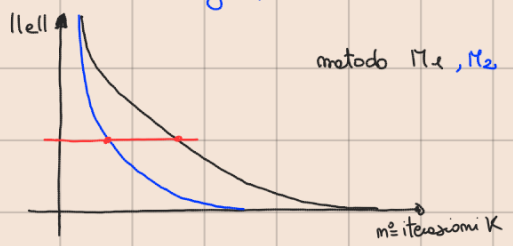
Cerco un K^* tale che $x_{K^*} \approx x^*$ → si genera un errore di troncamento

poiché un procedimento che dovrebbe essere infinito, viene troncato in passi finiti

errore troncamento = $\|x_K^* - x^*\|$

K viene solitamente definito ≤ maxit (num di iterazioni massime)

Velocità di convergenza



apparentemente M_2 è più veloce. M_1 ha T_{it1} come tempo per una singola iterazione. M_2 ha invece T_{it2}

Il tempo totale è dato da:

per M_1 → $T_{TOT1} = T_{it1} \cdot K_{it1}$

per M_2 → $T_{TOT2} = T_{it2} \cdot K_{it2}$

Convergenza metodi iterativi: la successione x_k si dice convergente ad un certo d con ordine $p \geq 1$ se

$$\exists C>0 : \frac{\|x_{k+1} - d\|}{\|x_k - d\|^p} \leq C \quad , \forall k : k \geq k_0 \qquad \{\text{dove } d = x^*\}$$

$$\|e_k\| \leq C \cdot \|e_{k-1}\| \quad \text{dove} \quad \|e_k\| = \|x_k - d\|$$

Nel caso $p=1$, per avere convergenza $C < 1$. In questo caso C = **fattore di convergenza**

Tra 2 metodi, a parità di ordine p, è migliore quello con la C più piccola, altrimenti, se due metodi hanno p diverse, è migliore quello con la p maggiore

Nei metodi iterativi il costo iterativo è quasi sempre un $O(n^2)$ → $T_{TOT} = T_{IT} \cdot N_{IT} = O(n^2) \cdot N_{IT}$

Se svolgo n iterazioni → ho un costo $O(n^3)$, mentre con fattorizzazione LR → $O(\frac{2}{3}n^3)$

NB i metodi iterativi si adattano bene alle matrici sparse (tanti elementi = 0)

$A \in \mathbb{R}^{n \times n}$ → A ha n^2 elementi di cui $m << n^2 \neq 0$

① Memorizzo di A solo gli m elementi e i loro indici (risparmio memoria)

② $T_{IT} = O(m)$ → $T_{TOT} = O(n^2 \cdot m)$

INTRODUZIONE AI METODI ITERATIVI STAZIONARI

$A \in \mathbb{R}^{n \times n}$ matrice non singolare → $A = M - N$ dove M (triang. superiore, triang. inferiore, diagonale ecc...) è non singolare e $Mz = h$ facilmente risolvibile

$$Ax = b \rightarrow Mx - Nx = b \rightarrow \underbrace{M^{-1}M}_{I}x - \underbrace{M^{-1}N}_{T}x = \underbrace{M^{-1} \cdot b}_{c}$$

$$x = T \cdot x + c \qquad x_k = T x_{k-1} + c \qquad k = 1, 2, \ldots$$

Devo dimostrare che converge alla soluzione

$$x_k \xrightarrow{k \to \infty} x^* \qquad x_{k-1} \xrightarrow{k \to \infty} x^* \quad \Rightarrow \quad x^* = T x^* + c$$

Un metodo è convergente se $\forall x_0$, la successione x_k converge

Es $T = \begin{pmatrix} 1/2 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$ $c = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$

$x_k = T \cdot x_0$

$$\begin{cases} x_{k+1} = Tx_k + c^{=0} \\ x_1 = Tx_0 \rightarrow x_2 = Tx_1 = T(Tx_0) = T^2(x_0) \\ x_3 = T(x_2) = T(Tx_1) = T(T(Tx_0)) = T^3 x_0 \end{cases}$$

$x_k \xrightarrow{k\to\infty}$ converge?

$$T^k = \begin{pmatrix} \frac{1}{2^k} & 0 & 0 \\ 0 & \frac{1}{2^k} & 0 \\ 0 & 0 & 2^k \end{pmatrix} \quad k \to \infty$$

Sia $x_0 = (1,0,0)$ $\rightarrow$ $x_k = T^k \cdot x_0 = T^k \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2^k} \\ 0 \\ 0 \end{pmatrix}$ converge per $k \to \infty$

Sia $x_0 = (0,0,1)$ $\rightarrow$ $x_k = T^k \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 2^k \end{pmatrix}$ diverge per $k \to \infty$

TEOREMA: il metodo iterativo è convergente se e solo se $\rho(T) < 1$ $\rightarrow$ il più grande autovalore di T in val assoluto

VELOCITÀ DI CONVERGENZA

Fissata una norma di vettori e la corrispondente norma di matrici indotta, si ha

errore al passo k $\quad \|e_k\| \leq \|T^k \cdot e_0\| \leq \|T^k\| \cdot \|e_0\|$ dove $e_k = \|x_k - x^*\|$ errore al passo k

Es

$$T = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.6 \end{pmatrix} \quad S = \begin{pmatrix} 0.5 & 0.25 \\ 0 & 0.5 \end{pmatrix} \rightarrow T^k = \begin{pmatrix} 0.5^k & 0 \\ 0 & 0.6^k \end{pmatrix} \quad S^k = \begin{pmatrix} 0.5^k & k \cdot 0.5^{k+1} \\ 0 & 0.5^k \end{pmatrix}$$

uso $\|\cdot\|_\infty$ $\quad \|T^k\|_\infty = 0.6^k \quad \|S^k\|_\infty = (2+k)\, 0.5^{k+1}$

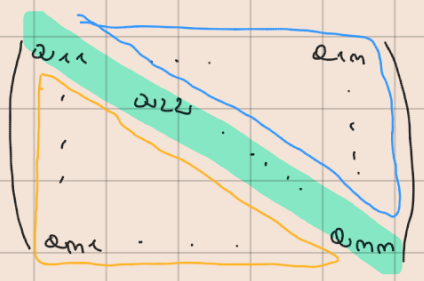
- per $k \leq 9$ ho $\|T^k\|_\infty < \|S^k\|_\infty$, mentre per $k \geq 10$ ho $\|T^k\|_\infty > \|S^k\|_\infty$

COSTRUZIONE DEI METODI ITERATIVI

$A = D - E - F$

dove $D = \{a_{11}, a_{22}, \dots, a_{mm}\}$ la diagonale di A

- E è la parte strettamente triangolare inferiore di A
- F è la parte strettamente triangolare superiore di A



Metodo di JACOBI (metodo delle sostituzioni simultanee)

$M = D$ diagonale $\quad N = E + F$ $\rightarrow$ E strettamente triangolare inferiore, F strett triangolare superiore

matrice di iterazione $J = D^{-1}(E+F) = I - D^{-1}A$ definito se D è non singolare (diagonale $\neq 0$)

$x_{k+1} = J \cdot x_k + D^{-1} b$ $\quad J x_k$

$$x_i^{(k)} = \left(b_i - \left[\sum_{j=1}^{i-1} a_{ij} \cdot x_j^{(k-1)} + \sum_{j=i+1}^{n} a_{ij} \cdot x_j^{(k-1)} \right] \right) / a_{ii}$$

componente i-esimo al passo k

Metodo di GAUSS-SEIDEL (metodo delle sostituzioni successive)

$M = D - E \qquad N = F$

matrice di iterazione $L_1 = (D-E)^{-1} F$ definito se $D-E$ è non singolare

$$x_{k+1} = L_1 x_k + (D-E)^{-1} \cdot b$$

Criteri d'arresto

$$\|x_k - x_{k-1}\| \leq \varepsilon \|x_k\|$$

dove ε è una quantità positiva prefissata

NOTA:

Gauss-Seidel, Jacobi convergono solo se $\rho(A) < 1$ (raggio spettrale < 1). Più $\rho(A)$ tende a 0 e più converge velocemente

Matrici particolari

- A è con diagonale dominante in senso stretto se
$$|a_{ii}| > \sum_{i \neq j} |e_{ij}| \quad \text{per ogni } i = 1, 2, \dots, m$$
- A è inriducibile se $\exists P$ di permutazione per cui
$$PAP^{-1} = \begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix} \quad \text{con } B_{11}, B_{22} \text{ matrici quadrate}$$
- A è inriducibile con diagonale dominante se A è inriducibile e
$$|a_{ii}| \geq \sum_{i \neq j} |e_{ij}| \quad , \; i = 1, 2, \dots, m \quad \text{con almeno un } i \text{ per cui vale in senso stretto}$$

JACOBI CONVERGENTE SE: A è con diagonale dominante in senso stretto o inriducibile con diagonale dominante

GAUSS-SEIDEL CONVERGENTE SE: A è con diagonale dominante in senso stretto o inriducibile con diagonale dominante

Sia A una matrice hermitiana (simmetrica) non singolare $\Rightarrow$ GAUSS SIEDEL converge se e solo se A è definita positiva

tutti autovalori: $\mathbb{R} \ni \lambda_i > 0 \;\; \forall i$

Nelle matrici tridiagonali

$$A = \begin{pmatrix} a_1 & c_1 & & & \\ b_1 & a_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{n-1} & 2 & c_{n-1} \\ & & & b_{n-1} & a_n \end{pmatrix}$$

Se μ autovalore di $J \Rightarrow \mu^2$ autoval di G

Se λ autoval non nullo di $G \Rightarrow \sqrt{(\lambda)}$ autoval di J

Quindi per queste matrici, Gauss-Seidel converge se e solo se Jacobi converge e vale

$$\rho(G) = \rho^2(J) \qquad (\Longrightarrow \rho(G) < \rho(J) \quad \text{perché entrambi} < 1)$$

immagine $\rho(J) = 0,4 \qquad \rho(G) = \rho(J)^2 = 0,16$ GS migliora

$$x_k = \mathcal{L} x_{k-1} + (D-E)^{-1} b$$

Metodi di rilassamento

Gauss-Seidel può essere scritto come:

$$x_k = x_{k-1} + r_k \qquad \text{dove } r_k = x_k - x_{k-1} = \overbrace{D^{-1}(Ex_k + Fx_{k-1} + b)}^{x_k} - x_{k-1}$$

Quindi il punto x_k si ottiene a partire da x_{k-1} effettuando un passo nella direzione r_k di lunghezza $\|r_k\|_2$.

Non sempre ciò provoca una convergenza veloce. Si modifica la lunghezza del passo introducendo ω

$$x_k = x_{k-1} + \omega r_k \qquad \begin{cases} \omega < 0 & \text{sottorilassamento} \\ \omega > 0 & \text{sovrarilassamento} \end{cases}$$

GAUSS-SEIDEL RILASSATO (SOR)

$$x_k = (D - \omega E)^{-1}((1-\omega)D + \omega F)x_{k-1} + \omega(D-\omega E)^{-1} \cdot b$$

la cui matrice di iterazione $\mathcal{L}_\omega = (D - \omega E)^{-1}((1-\omega)D + \omega F)$

Una condizione necessaria di convergenza per SOR è:

$$0 < \omega < 2$$

Teorema (Ostrowski - Reich): Se A è definita positiva e $0 < \omega < 2 \Rightarrow$ il metodo di rilassamento è convergente

FATTORIZZAZIONE DI CHOLESKY

IMPORTANTE

sdp: simmetrica definita positiva — metodo diretto

Ogni matrice A sdp può essere fattorizzata come

$$A = L \cdot L^T \qquad \text{con } L \text{ triang inf non singolare e } L^T \text{ triang sup non singolare}$$

↳ non necessariamente diagonale di $L = 1 \dots 1$

Se A è sdp $\Rightarrow$ tutte le sue sottomatrici sono sdp

→ Ad ogni passo j prendo la sottomatrice $A_j = L_j \cdot L_j^T$

- passo 1 $\quad A_1 = L_1 \cdot L_1^T \qquad A_1 = a_{11} > 0$

$$L_1 \cdot L_1^T = a_{11}^2 \longrightarrow L_{11} = \sqrt{a_{11}}$$

- passo 2 $\quad A_2 = L_2 \cdot L_2^T$

$$\begin{pmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{pmatrix} \cdot \begin{pmatrix} l_{11} & l_{21} \\ 0 & l_{22} \end{pmatrix} = \begin{pmatrix} l_{11}^2 & l_{11} \cdot l_{21} \\ l_{11} l_{21} & l_{22}^2 + l_{21}^2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

(sono noti)

ricavo $\quad l_{21} = \dfrac{a_{21}}{l_{11}} \qquad l_{22} = \sqrt{a_{22} - l_{21}^2}$

per induzione conosco $A_{k-1} = L_{k-1} \cdot L_{k-1}^T$

- al passo k conosco $A_{k-1} = L_{k-1} \cdot L_{k-1}^T$

$$A_k = \begin{pmatrix} A_{k-1} & a_k \\ a_k^T & a_{k,k} \end{pmatrix}$$

NOTAZIONE MATRICIALE

ad esempio: $A_3 = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} A_2 & a_3 \\ a_4 & a_{33} \end{pmatrix}$

A_2 sottomatrice chiusa; a_3; a_4

$$L_k = \begin{pmatrix} L_{k-1} & \vec{0} \\ l_k^T & \lambda \end{pmatrix}$$

$\vec{0}$: vettore nullo di $k-1$ elementi; L_{k-1}: $k-1$ colonne; l_k^T: $k-1$ elementi

$A_k = L_k L_k^T$

la complessità computazionale dell'algoritmo è $O(m^3/6)$ [la metà della fattorizzazione LR $\to O(\frac{m^3}{3})$]

l'algoritmo è stabile e non serve pivoting

$$Ax = b \to L \cdot L^T x = b \to \begin{cases} Ly = b \\ y = L^T \cdot x \end{cases}$$

METODO GRADIENTI CONIUGATI (GC)

metodo iterativo { dove $x_{k+1} = x_k + d_k p_k$ con $d_k \in \mathbb{R}$, $\vec{p_k}$ }

Vettore residuo: indica la "distanza" dalla soluzione esatta al passo k

$$r_k = b - A \cdot x_k$$

se x^* fosse la soluzione esatta $\Rightarrow Ax^* = b \Rightarrow r_k = b - A \cdot x^* = \vec{0}$

TEOREMA: Sia A sdp allora l'algoritmo dei gradienti coniugati calcola la soluzione in m passi: $x_m = x^*$

ma lavorando con i numeri finiti le iterazioni sono maggiori di m

Calcolo l'iterazione k-esima come: $x_{k+1} = x_k + d_k p_k$ $d_k \in \mathbb{R}$ detto passo e $\vec{p_k}$ è una direzione di discesa

N.B. in aritmetica finita le iterazioni possono essere anche infinite

Le iterazioni si fermano quando si raggiunge un criterio di convergenza

[ORDINE DI CONVERGENZA: $\exists c > 0$ t.c. $\|e_{k+1}\| \le c \|e_k\|^p < \dots < c^{k+1} \|e_0\|^p$ adottando ai metodi iterativi]

- Criterio di arresto: relazione di decrescita dell'errore in norma A

$$\|x_k - x^*\|_A \le 2 \|x_0 - x^*\|_A \left(\frac{K_2(A) - 1}{K_2(A) + 1} \right)^2$$

$$\|e_k\|_A \le 2 \|e_0\|_A \left(\frac{K_2(A) - 1}{K_2(A) + 1} \right)^2 \quad < 1$$

la norma A: $\|x\|_A = x^T \cdot A \cdot x > 0$ → norma energia sob se A simmetrica definita positiva

$K(A) = \|A\| \cdot \|A^{-1}\| \ge 1$ con $K_2(A)$ uso la norma 2 $\to K_2(A) = \|A\|_2 \cdot \|A^{-1}\|_2$

Tanto più $K(A)$ è vicino ad 1, tanto più il metodo decresce velocemente

- Criterio per terminare le iterazioni

$$\|b - A x_k\|_2 = \|r_k\|_2 \le \eta \|b\|_2$$ dove η è un valore fissato ($\approx 10^{-6}$)

se $r_0 \neq \vec{0}$ $\to \frac{\|r_k\|_2}{\|r_0\|_2} \le \eta$ criterio relativo al residuo cioè faccio $\frac{r_k}{r_0}$ supponendo $r_0 \neq 0$

AUTOVALORI

Date una matrice $A \in \mathbb{R}^{m \times m}$, $\lambda \in \mathbb{R}$, $x \in \mathbb{R}^m$, λ si definisce autovalore se

$$Ax = \lambda x$$

x è detto autovettore associato all'autovalore λ

Per calcolare gli autovalori, trovo gli zeri del polinomio caratteristico

$$p_A(\lambda) \to \det(A - \lambda I) = 0 \quad \text{radici}$$

Se $\lambda \in \mathbb{C}$ è autovalore $\Rightarrow \bar{\lambda}$ coniugato di λ è autovalore $(a+bi) \xrightarrow{\text{coniugato}} (a-bi)$

Per avere autovettori di norma 1, li divido per la loro lunghezza

$$\omega = \frac{v}{\|v\|_2} \to \|w\| = 1 \quad \text{normalizzato}$$

Calcolo autovettori:

METODO DELLE POTENZE

lo spettro di una matrice $A \to$ insieme di tutti i suoi autovalori. I più importanti sono i max e min in modulo

ipotesi: (1) $|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_m|$ NB λ_1 distinto dagli altri

(2) $x_1, \dots, x_m$ lin indipendenti

Supponiamo di voler calcolare l'autovettore di valore massimo di A (λ_1). Denotiamo con x_1 l'autovettore associato a λ_1 tale che $\|x_1\| = 1$. Se gli autovettore $x_1, \dots, x_m$ sono indipendenti, posso calcolare λ_1 e x_1 con il metodo delle potenze.

IDEA: Dato un vettore iniziale $x^{(0)} \in \mathbb{R}^m$ e posto $y^{(0)} = \frac{x^{(0)}}{\|x^{(0)}\|}$ (normalizzazione), si calcola per $k = 1, 2, \dots$

$$x^{(k)} = A \cdot y^{(k-1)}, \quad y^{(k)} = \frac{x^{(k)}}{\|x^{(k)}\|}, \quad \lambda^{(k)} = \left(y^{(k)}\right)^T A y^{(k)} \to \text{scalare}$$

$x_k \xrightarrow{k \to \infty} x_1$, $\lambda_k \xrightarrow{k \to \infty} \lambda_1$

$$y^{(1)} = \frac{x^{(1)}}{\|x^{(1)}\|}$$

$$y^{(2)} = \frac{x^{(2)}}{\|x^{(2)}\|} = \frac{A \cdot y^{(1)}}{\|A \cdot y^{(1)}\|} = \frac{A \frac{x^{(1)}}{\|x^{(1)}\|}}{\left\|A \frac{x^{(1)}}{\|x^{(1)}\|}\right\|} = \frac{A \frac{A y^{(0)}}{\|A y^{(0)}\|}}{\left\|A \frac{A y^{(0)}}{\|A y^{(0)}\|}\right\|} = \frac{A^2 y^{(0)}}{\left\|A^2 \frac{y^{(0)}}{\|A y^{(0)}\|}\right\|}$$

generalmente ho

$$y^{(k)} = \beta^{(k)} \cdot A^k \cdot y^{(0)} \quad \text{dove } \beta^{(k)} = \prod_{i=1}^{k} \left(\|x^{(i)}\|^{-1}\right)$$

raccogliendo

Analisi di convergenza

Poiché $x_1, \dots, x_m$ lin indipendenti, generano una base di $\mathbb{R}^m$. Di conseguenza, $x^{(0)}$ e $y^{(0)}$ possono essere scritte come combinazioni di $x_1, \dots, x_m$

$$x^{(0)} = \sum_{i=1}^{m} a_i x_i, \quad y^{(0)} = \beta^{(0)} \cdot \sum_{i=1}^{m} x_i a_i \quad \beta^{(0)} = \frac{1}{\|x^{(0)}\|}$$

$$x^{(1)} = A y^{(0)} = \beta^{(0)} \cdot A \cdot \sum_{i=1}^{m} x_i a_i$$

A non dipende da i → ob porto nella sommatoria

$$\to = \beta^{(0)} \cdot \sum_{i=1}^{m} A x_i a_i = \beta^{(0)} \cdot \sum_{i=1}^{m} \lambda_i x_i a_i$$

$Ax = \lambda x$ autovettore

$$y^{(k)} = \beta^{(k)} \cdot \sum_{i=1}^{m} a_i \cdot \lambda_i^k \cdot x_i \qquad \beta^{(k)} = \frac{1}{\|x^{(1)}\| \cdots \|x^{(k)}\|}$$

quindi $y^{(k)} = \lambda_1^k \cdot \beta^{(k)} \left(a_1 x_1 + \sum_{i=2}^{m} a_i \frac{\lambda_i^k}{\lambda_1^k} x_i \right)$ → tende a zero per $k \to \infty$ perché $\lambda_1 > \dots \lambda_i \dots$

Poiché $|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_m| \Rightarrow \left(\frac{\lambda_i}{\lambda_1}\right)^k < 1$ per $k \to \infty$ tende a zero

$y^{(k)} \xrightarrow{k \to \infty} \lambda_1^{(k)} \beta^{(k)} a_1 x_1$ quindi $y^{(k)}$ tende ad essere associato ad uno degli infiniti autovetti di λ_1

Gli errori e la velocità di convergenza sono proporzionali al oggetto $\left|\frac{\lambda_2}{\lambda_1}\right|^k$ $\left(\|e_k\| \leq c \|e_{k-1}\| \quad c = \left|\frac{\lambda_2}{\lambda_1}\right| < 1 \right)$

Tanto più c è piccolo, tanto più il metodo converge velocemente

METODO DELLE POTENZE INVERSE → metodo delle potenze applicato alla matrice inversa

Se $\lambda_1, \dots, \lambda_m$ autovalori di A, allora $\frac{1}{\lambda_1}, \dots, \frac{1}{\lambda_m}$ autovalori di A^{-1}

Per approssimare l'autovalore minimo di A, approssimo il max di A^{-1}, e poi faccio il reciproco

Dato $x^{(0)}$, $y^{(0)} = \frac{x^{(0)}}{\|x^{(0)}\|}$, per $k = 1, 2, \dots$

$x^{(k)} = A^{-1} \cdot y^{(k-1)}$, $y^{(k)} = \frac{x^{(k)}}{\|x^{(k)}\|}$, $\mu^{(k)} = (y^{(k)})^T A^{-1} \cdot y^{(k)}$

sono uguali

↳ $A x^{(k)} = A \cdot A^{-1} \cdot y^{(k-1)} \rightarrow A x^{(k)} = y^{(k-1)}$

NB In tutte le iterazioni A non cambia. Posso fattorizzarla LR (LU) [o cholesky se A simmetrica definita positiva] e ad ogni iterazione risolvo 2 problemi triangolari

Particolarità degli autovettori

$\|A\|_2 = \sqrt{\rho(A^T \cdot A)} = \sqrt{|\lambda_{max} A^T \cdot A|}$ autoval di modulo max di $A^T \cdot A$

① Se A è simmetrica → $A^T \cdot A = A \cdot A = A^2$ prodotto matriciale

② Se λ_i è autovalore di A allora $1/\lambda_i$ è autovalore di A^{-1}

③ Se A è simmetrica → A^{-1} è simmetrica $\Rightarrow (A^{-1})^T \cdot (A^{-1}) = A^{-1} \cdot A^{-1} = (A^{-1})^2$

allora gli autovalori di $(A^{-1})^2 = \left(\frac{1}{\lambda_i}\right)^2$ con λ_i autovalore di A

④ A^2 ha autovalori $(\lambda_i)^2$

NB: Se A è simmetrica $\Rightarrow K_2(A) = \|A\|_2 \cdot \|A^{-1}\|_2 = \sqrt{\lambda_{max} A^T \cdot A} \cdot \sqrt{\lambda_{max} (A^{-1})^T (A^{-1})} = \sqrt{\lambda_1^2} \cdot \sqrt{\frac{1}{\lambda_m^2}} =$

continua → $\sqrt{\frac{\lambda_1^2}{\lambda_m^2}} = \sqrt{\left(\frac{\lambda_1}{\lambda_m}\right)^2} = \left|\frac{\lambda_1}{\lambda_m}\right| \geq 1$

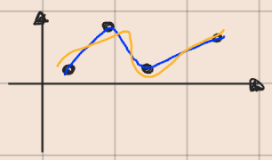
Se λ_1 è il max autovalore di A $\Rightarrow$ $\frac{1}{\lambda_m}$ è il max autovalore di A^{-1}

maggiore è $\left|\frac{\lambda_1}{\lambda_m}\right|$ e più la matrice è mal condizionata

Se A è sdp $\Rightarrow K_2(A) = \frac{\lambda_1}{\lambda_m}$ (autovalori tutti > 0)

$\lambda_1 > \lambda_2 \geq \dots \geq \lambda_m \rightarrow \frac{1}{\lambda_1} \leq \dots \leq \frac{1}{\lambda_{m-1}} < \frac{1}{\lambda_m}$

INTERPOLAZIONE



APPROSSIMAZIONE
INTERPOLAZIONE

coppie (x_i, y_i) $i=0,\dots,m$

$y_i = f(x_i)$

obiettivo: trovare una funzione che descrive l'andamento dei dati

Posso avere due tipi:

① Interpolazione → la funzione passa per i punti (se i dati non sono affetti da errori e sono pochi)

② Approssimazione → la funzione passa vicino ai punti

3 tipi di interpolazione: conosciamo $m+1$ coppie $\{x_i, y_i\}$ $i=0,\dots,m$ con gli x_i distinti tra loro (nodi)

① POLINOMIALE: $\tilde{f} = a_0 + a_1 x + a_2 x^2 + \dots + a_m x^m$

② TRIGONOMETRICA

③ RAZIONALE

interpolazione polinomiale di Lagrange

Per ogni coppia $\{x_i, y_i\}$ $i=0,\dots,m$ con x_i distinti tra loro, esiste un unico polinomio di grado $\leq m$ (indicato con Π_m) chiamato polinomio interpolatore tale che $\{\Pi_m$ è il polinomio$\}$

$\Pi_m(x_i) = y_i$ $i=0,\dots,m$ Se $y_i = f(x_i)$ → Π_m è detto interpolatore di f e si indica con $\Pi_m f$

- Π_m è unico:

Per assurdo, esiste Π_m^* polinomio interpolatore. ⇒ $\Pi_m - \Pi_m^*$ è un polinomio, ma $(\Pi_m - \Pi_m^*)(x_i) = 0$

quindi si annulla in $m+1$ punti → ha $m+1$ radici → grado $m+1$ ed è un assurdo ⇒ i polinomi sono uguali

- Π_m esiste sempre:

$$\Pi_m = a_0 + a_1 x + a_2 x^2 + \dots + a_m x^m$$

$$\left.\begin{array}{ll} i=0 \to & a_0 + a_1 x_0 + \dots + a_m x_0^m = y_0 \\ \vdots & \\ i=m \to & a_0 + a_1 x_m + \dots + a_m x_m^m = y_m \end{array}\right] \to \text{le incognite sono le } a_0, \dots, a_m$$

ho $m+1$ equazioni ed $m+1$ incognite → ho 1 soluzione

$$\vec{a} = \begin{pmatrix} a_0 \\ \vdots \\ a_m \end{pmatrix} \quad \vec{y} = \begin{pmatrix} y_0 \\ \vdots \\ y_m \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^m \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_m & x_m^2 & \dots & x_m^m \end{pmatrix}$$

X si chiama matrice di Vandermonde ed è malcondizionata → $X \cdot \vec{a} = \vec{y}$ non viene calcolato

Per calcolare le incognite $a_0, \dots, a_m$ possiamo usare l'algoritmo di Lagrange

DEF spazio funzionale: spazio vettoriale formato da funzioni. Ho una base di funzioni attraverso la quale posso esprimere tutte le funzioni come combinazione delle funzioni della base

base monomiale: $\begin{cases} \varphi_0(x) = 1 \\ \varphi_1(x) = x \\ \vdots \\ \varphi_m(x) = x^m \end{cases}$ → m+1 elementi φ_k corrisponde ad x^k nella base monomiale (sono i monomi)

$\Pi_m(x) = c_0 \varphi_0(x) + c_1 \cdot \varphi_1(x) + \dots + c_m \varphi_m(x)$ cambiando la base cambiano i componenti

NB Nella base di Lagrange, le φ non sono più i monomi e i coefficienti ci sono le ordinate dei dati

$c_0 = y_0$, $c_1 = y_1$, ..., $c_m = y_m$

Il polinomio nella forma di Lagrange sarà

$$\Pi_m(x) = y_0 \varphi_0(x) + y_1 \varphi_1(x) + \dots + y_m \varphi_m(x)$$

$\varphi_k = L_k$

La base di Lagrange è $\{L_0, L_1, \dots, L_m\}$ → $\Pi_m(x) = y_0 \cdot L_0(x) + y_1 L_1(x) + \dots + y_m L_m(x)$

- $L_k(x) = \varphi_k(x) = \prod_{\substack{j=0 \\ j \neq k}}^{m} \frac{x - x_j}{x_k - x_j}$ $k = 0, \dots, m$ prodotto per $j = 0 \dots m$ con $j \neq k$

$$L_1(x) = \varphi_1(x) = \left(\frac{x-x_0}{x_1-x_0}\right) \cdot \left(\frac{x-x_2}{x_1-x_2}\right) \cdot \dots \cdot \left(\frac{x-x_m}{x_1-x_m}\right)$$ ← Base di Lagrange di ordine 1

$$L_2(x) = \varphi_2(x) = \left(\frac{x-x_0}{x_2-x_0}\right) \cdot \left(\frac{x-x_1}{x_2-x_1}\right) \cdot \left(\frac{x-x_3}{x_2-x_3}\right) \cdot \dots \cdot \left(\frac{x-x_m}{x_2-x_m}\right)$$ ← indice 2

Calcolo L_1 in x_1:

$$L_1(x_1) = \left(\frac{x_1-x_0}{x_1-x_0}\right) \cdot \left(\frac{x_1-x_2}{x_1-x_2}\right) \cdots \left(\frac{x_1-x_m}{x_1-x_m}\right) = 1$$

Calcolo L_1 in x_2:

$$L_1(x_2) = \left(\frac{x_2-x_0}{x_1-x_0}\right) \cdot \left(\frac{x_2-x_2}{x_1-x_2}\right) \cdot \dots \cdot = 0$$

NB se calcolo $L_i(x_i) = 1$, invece $L_i(x_j) = 0$ se $j \neq i$

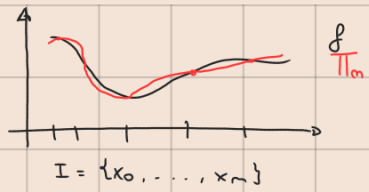
Il polinomio $\Pi_m(x_i) = y_i$ per $i = 0, \dots, m$

$$\Pi_m(x_0) = c_0 \underbrace{\varphi_0(x_0)}_{=1} + c_1 \underbrace{\varphi_1(x_1)}_{=0} + \underbrace{\dots}_{=0} + c_m \underbrace{\varphi_m(x_m)}_{=0} = y_0 \rightarrow c_0 \underbrace{\varphi_0(x_0)}_{=1} = y_0 \rightarrow c_0 = y_0$$

Il polinomio di interpolazione nella formula di Lagrange è quindi

$$\Pi_m(x) = \sum_{k=0}^{m} y_k \cdot \varphi_k(x)$$

Interpolazione di una funzione $(x_k, y_k) = (x_k, f(x_k))$



Voglio vedere l'errore commesso nei punti

Nei punti di interpolazione $|f(x_i) - \Pi_m(x_i)| = 0$, mentre negli altri punti

$$E_m f(x) = f(x) - \Pi_m f(x) = \frac{f^{(m+1)}(\xi)}{(m+1)!} \prod_{i=0}^{m} (x - x_i)$$

Se $m \to \infty$, cioè aumento i punti di interpolazione, allora $\lim_{m\to\infty} \max_{x \in I} |E_m f(x)| = \infty$

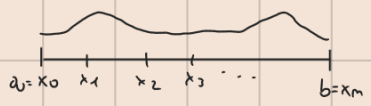
Usando i punti di Chebyshev nell'intervallo (a,b) → l'errore aumentando i punti tende a ZERO

$x_0 = a$, $x_m = b$, $x_i = \frac{a+b}{2} - \frac{b-a}{2} \cdot \cos\left(\frac{2i+1}{m+1} \cdot \frac{\pi}{2}\right)$ $\quad i = 0, \dots, m$

i punti sono più densi nei bordi dell'intervallo e meno densi al centro (sono simmetrici)

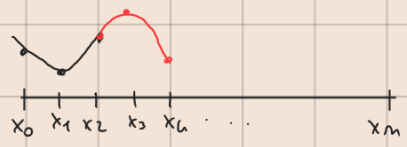
INTERPOLAZIONE POLINOMIALE A TRATTI

interpoliamo ogni sottointervallo $[x_i, x_{i+p}]$ con un polinomio di grado p ($p = 1, 2, 3$ solitamente)



x_i, $i = 0 \dots m$ equidistanti

y_i, $i = 0 \dots m$ misure

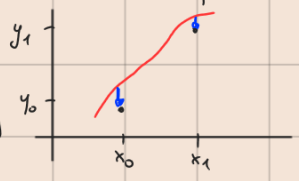


p = grado → $p=2$: parabole → ogni sottointervallo $[x_i, x_{i+2}]$ è una parabola

APPROSSIMAZIONE

Voglio trovare una funzione che approssima il fenomeno da cui ricaviamo i dati, e che non passa necessariamente per i punti

I punti (x_i, y_i) potrebbero essere affetti da errori.



Ho m punti e voglio calcolare una funzione (polinomio) di grado n ($n \ll m$ possibile)

in blu è la distanza tra punto e approssimazione $r_i = |p(x_i) - y_i|$

il vettore residui $\vec{r} = (r_0, \dots, r_m)$ → $\|r\|_2$ è la distanza tra polinomio e punti, cerco di minimizzarla (per semplicità calcolo $\min \|r\|_2^2$)

↳ la distanza tra il polinomio e i punti deve essere piccola

PROBLEMA AI MINIMI QUADRATI

sto nel determinare il vettore $x \in \mathbb{R}^n$ che minimizzi il vettore residui

$r = Ax - b$ $\qquad \min_{x \in \mathbb{R}^n} (\|Ax - b\|_2)^2 = \min_{x \in \mathbb{R}^n} (\|r\|_2)^2$

↳ $r = (r_0, r_1, \dots, r_m) = ((Ax)_0 - b_0, \dots, (Ax)_m - b_m)$ Calcolo Ax e prendo le componenti $0, 1, \dots, m$

Condizione di esistenza e unicità dei minimi quadrati

Data $A \in \mathbb{R}^{m \times n}$ con $m > n$, $b \in \mathbb{R}^m$, $x \in \mathbb{R}^n$

Sia $K = rg(A) \leq n$ → il problema dei minimi quadrati ammette almeno 1 soluzione sempre

$\boxed{K = n}$ → rango massimo: una sola soluzione unica

$\boxed{K < n}$ → soluzioni infinite che creano un sottospazio di $\mathbb{R}^n$ di dimensione $n - K$

① Caso $K=N$ — A ha rango max → A è simmetrica e definita positiva

$$\begin{aligned}||Ax-b||_2^2 &= (Ax-b)^T(Ax-b)\\ &= (-b^T+x^TA^T)(Ax-b)\\ &= -b^TAx+b^Tb+x^TA^TAx-x^TA^Tb\\ &\text{(sommando i termini simili } b^TAx = x^TA^Tb)\\ &= x^TA^TAx-2x^TAb+b^Tb\end{aligned}$$

Si pone:

$$f(x) = x^TA^TAx-2x^TAb+b^Tb$$

$\nabla f(x) = 2\cdot A^T\cdot A\cdot x - A^T\cdot b$

ponendo $\nabla f(x)=0$ (cerco punti di flesso) → $A^T\cdot A\cdot x = A^T\cdot b$

dove $A^T\cdot A$ è sdp → A ha rango massimo

si risolve il sistema $\begin{cases} Ly = A^Tb \\ y = L^T\cdot x \end{cases}$ con A fattorizzata con Cholesky $\quad K(A^T\cdot A) = K(A^2)$

② Caso $K\leq N$: decomposizione in valori singolari (SVD) — tra le infinite soluzioni vogliamo quella di norma minima

Sia $A\in\mathbb{R}^{m\times m}$, $m\geq m$, $K=rg(A)\leq m$ ALLORA ESISTONO:

- $U\in\mathbb{R}^{m\times m}$ ortogonale quadrata
- $V\in\mathbb{R}^{m\times m}$ ortogonale quadrata
- $\Sigma = diag(\sigma_1,\dots,\sigma_m)$ con $\sigma_1\geq\sigma_2\geq\dots\geq\sigma_m\geq 0$ detti valori singolari

$\Sigma\in\mathbb{R}^{m\times m}$: $\begin{pmatrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_m \\ 0 & \dots & 0 \\ 0 & \dots & 0\end{pmatrix}$ tutti zeri nelle m-n righe

ALLORA: $A = U\cdot\Sigma\cdot V^T$

NB: matrice ortogonale: $A\in\mathbb{R}^{m\times m}$ è ortogonale se, dette $a_1,\dots,a_m$ le colonne di A, si ha che

$\langle a_i, a_j\rangle = 0 \quad \forall i\neq j$ && $||a_i||_2 = 1$

Una matrice ortogonale ha alcune proprietà:

① $A^T = A^{-1}$ $\qquad A\cdot A^{-1} = A\cdot A^T = I$

② $\forall v\in\mathbb{R}^m$, $||v||_2 = ||A\cdot v||_2$ — proprietà di isometria — prendo un vettore in $\mathbb{R}^2$, lo moltiplico per A ortogonale ottengo un vettore lungo uguale (non stessa dir??)

③ A^T è ortogonale

Teorema: Sia $A\in\mathbb{R}^{m\times m}$, $rg(A)=K\leq m$, sia $A=U\cdot\Sigma\cdot V^T$ la sua decomposizione in valori singolari, allora il vettore

$$x^* = \sum_{i=1}^{K}\frac{u_i^T\cdot b}{\sigma_i}\cdot v_i$$

v_i è la colonna i-esima di V; dove u_i^T è la colonna i-esima di U, trasposta, $\sigma_i\in\mathbb{R}$

è la soluzione di minima norma del problema $\min_{x\in\mathbb{R}^m}(||Ax-b||_2)^2$

in corrispondenza ottengo: $(||r||_2)^2 = (||Ax^*-b||_2)^2 = \sum_{i=K+1}^{m}(u_i^Tb)^2$

DIM

posso farlo perché le norme restano uguali per l'isometria di U

$$||Ax-b||_2^2 = (\text{moltiplico per } U^T) = ||U^TAx-U^Tb||_2^2 = ||U^TA\underbrace{V\cdot V^T}_{I}x - U^Tb||_2^2$$

V ortogonale $\Rightarrow V^T = V^{-1} \Rightarrow V\cdot V^T = V\cdot V^{-1} = I$

posto $y = V^T\cdot x \in\mathbb{R}^m$, $\quad g = U^T\cdot b\in\mathbb{R}^m$

Se $A = U\cdot\Sigma\cdot V^T$ → per ortogonalità di U,V → $U^T\cdot A\cdot V = \Sigma$, allora

$$||Ax-b||_2^2 = ||\Sigma y - g||_2^2 = \sum_{i=1}^{K}(\sigma_i y_i - g_i)^2 + \sum_{i=K+1}^{m}(g_i)^2$$

NB i valori singolari di A: $\sigma_1, \dots, \sigma_m$ sono positivi in ordine decrescente e sono ≥ 0 fino a σ_k con $k = \text{rg}(A)$

$\sigma_1, \dots, \sigma_k > 0 \qquad \sigma_{k+1}, \dots, \sigma_m = 0$

$$\Sigma = \begin{pmatrix} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & \ddots & & \\ & & & \sigma_k & \\ 0 & \cdots & & & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} x \\ x \\ x \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{matrix} \} k \text{ elementi} \neq 0 \\ \} m-k \text{ elementi} = 0 \end{matrix}$$

$m \times m \qquad m \times 1 \qquad m \times 1$

Devo minimizzare rispetto ad x

con $y = V^T x$

$$\sum_{i=1}^{k} (\sigma_i y_i - g_i)^2 + \sum_{i=k+1}^{m} (g_i)^2$$

non dipende da x

è una somma di positivi $\to$ pongo $= 0$

$$\sigma_i \cdot y_i = g_i \longrightarrow y_i = \frac{g_i}{\sigma_i} = \frac{U^T b}{\sigma_i}$$

$$y_i = \frac{U^T b}{\sigma_i} \qquad \text{poiché } y = V^T \cdot x \longrightarrow Vy = V \cdot V^T \cdot x \longrightarrow x = Vy$$

$y = \frac{U^T b}{\sigma_i}$

ottengo così $\quad x_i^* = \dfrac{V \cdot U^T \cdot b}{\sigma_i} \iff x^* = \displaystyle\sum_{i=1}^{k} \frac{u_i^T \cdot b}{\sigma_i} \cdot v_i \qquad \text{con } x^* \in \mathbb{R}^m$

la norma del residuo, in corrispondenza di tale soluzione, è $\quad \|r\|_2^2 = \displaystyle\sum_{i=k+1}^{m} (g_i)^2 = \sum_{i=k+1}^{m} (u_i^T b)^2$

$(x_i, y_i) \quad i = 0, \dots, m \qquad p_m(x) = a_0 + a_1 x + a_2 x^2 + \dots + a_m x^m \qquad m > m$ $\begin{cases} m+1 \text{ punti} \\ m+1 \text{ coefficienti} \end{cases}$

$$r_i = p_m(x_i) - y_i = a_0 + a_1 x_i + a_2 x_i^2 + \dots + a_m x_i^m - y_i \qquad (i = 0, \dots, m)$$

scritto in forma matriciale ottengo

$\vec{r} = X \cdot \vec{a} - \vec{y}$ dove $\vec{a} = (a_0, \dots, a_m)^T \in \mathbb{R}^{m+1}$ incognita

$\vec{y} = (y_0, \dots, y_m)^T \in \mathbb{R}^{m+1}$

$$X = \begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^m \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_m & x_m^2 & & x_m^m \end{pmatrix}$$

rettangolare con $m+1$ righe e $m+1$ colonne

"matrice di Vandermonde rettangolare"

$\longrightarrow \quad \min \|r\|_2^2 = \min_{a \in \mathbb{R}^{m+1}} \|X\vec{a} - \vec{y}\|_2^2$ lo minimizzo rispetto ad a in $\mathbb{R}^{m+1}$

CAMBIO $\quad Ax - b \longrightarrow X\vec{a} - \vec{y}$

$$\begin{cases} A \longrightarrow X \\ \vec{x} \longrightarrow \vec{a} \\ \vec{b} \longrightarrow \vec{y} \end{cases}$$

Condizionamento del problema dei minimi quadrati

I valori singolari σ_i sono >0 fino a σ_k con $k = rg(A)$, inoltre $A v_i = \sigma_i u_i$ (vettori singolari destri v_i, vettori singolari sinistri u_i)

c'è una relazione tra i valori singolari di A e gli autovalori di $A^T \cdot A$ [Ricordiamo $(xy)^T = y^T \cdot x^T$]

$A = U \cdot \Sigma \cdot V^T$

$A^T A = (U \Sigma V^T)^T (U \Sigma V^T) = V \Sigma^T \overbrace{U^T U}^{I \text{ per ortogonalità}} \Sigma V^T = V \Sigma^T \Sigma V^T$ $\quad (m \times m)\cdot(m \times m) \to (m \times m)$

NB $(xy)^T = y^T x^T$

NB $\Sigma^T \Sigma$ è una matrice $m \times m$ diagonale, con i valori singolari al quadrato sulla diagonale

$$A^T A = V \Sigma^T \Sigma V^T = V \cdot \begin{bmatrix} \sigma_1^2 & & \\ & \sigma_2^2 & \\ & & \ddots \\ & & & \sigma_m^2 \end{bmatrix} \cdot V^T$$

V è la matrice degli autovettori di $A^T A$

un generico $\sigma_i = \sqrt{\lambda_i (A^T A)}$ $\quad i = 1 \dots n \longrightarrow (\sigma_i)^2 = \lambda_i (A^T A)$ autovalore i-esimo

in particolare, se $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_m$

- $\sigma_1 = \sqrt{\lambda_{max}(A^T A)} = \sqrt{\rho(A^T A)} = \|A\|_2$ Norma-2 di A
- $\sigma_m = \sqrt{\lambda_{min}(A^T A)}$
  - di conseguenza $\frac{1}{\sigma_m} = \|A^{-1}\|_2$

$$K_2(A) = \|A\|_2 \cdot \|A^{-1}\|_2 = \frac{\sigma_1}{\sigma_m}$$ condizionamento della matrice A

Approssimazione di una matrice tramite SVD → estrarre il contenuto informativo di una matrice

Data $A \in \mathbb{R}^{m \times n}$, $k = rg(A)$, $A = U \Sigma V^T$.

Sia $A_p = \sum_{i=1}^{p} \sigma_i v_i u_i^T$ $\quad \begin{cases} v_i = m \times 1 \\ u_i^T = 1 \times n \\ \sigma_i = \text{scalare} \end{cases} \to$ scalare · matrice $m \times n$

A_p sono matrici di rango 1 (DIADE: matrice di rango 1)

- Se $p = k$ allora $\sum_{i=1}^{k} \sigma_i v_i u_i^T = A$
- Se $p < k$ A_p è l'approssimazione di rango p di A $\quad (rg(A_p) = p)$

$\forall B \in \mathbb{R}^{m \times n}$, $rg(B) = p$, $\quad \|A - A_p\|_2 \leq \|A - B\|_2$

↳ tra tutte le matrici di rango p, A_p è l'approssimazione migliore di A

A può essere scritto come $\sum_{i=1}^{k} \sigma_i v_i \cdot u_i^T$, mentre $A_p = \sum_{i=1}^{p} \sigma_i v_i \cdot u_i^T$

di conseguenza $A - A_p = \sum_{i=p+1}^{k} \sigma_i v_i u_i^T$

$\|A - A_p\|_2 = \sigma_{p+1}$ se σ_{p+1} è piccolo, si ha una buona approssimazione di A

questa norma è il valore singolare massimo. La distanza è il primo valore singolare da "trascurare"

RISOLVERE FUNZIONI NON LINEARI

$f: I \subset \mathbb{R} \to \mathbb{R}$, vogliamo calcolare gli zeri di f, cioè $f(x)=0$

Weierstrass: se f è continua in $[a,b]$, con $f(a)\cdot f(b) < 0$ allora $\exists$ almeno uno zero di f in (a,b)

METODI ITERATIVI

$x_k = G(x_{k-1})$, questi iterati convergono a x^* $\left[x_k \xrightarrow{k\to\infty} x^*\right]$ tale che $f(x^*)=0$

x_k converge a x^* con un ordine $p \geq 1$ se $\dfrac{|x_{k+1} - x^*|}{|x_k - x^*|^p} = C \quad \forall k \geq k_0$

dove k_0 è un intero, e $C \in \mathbb{R}$ tale che $\begin{cases} 0 < C \leq 1 & \text{se } p=1 \\ C > 0 & \text{se } p>1 \end{cases}$

Metodo di bisezione:

Posto un intervallo (a,b) tale che $f(a)\cdot f(b) < 0$ $\quad [f(a)<0,\ f(b)>0]$

costruisco $I_1 = [a,b]$, $I_2 = [a_2, b_2]$, ..., $I_k = [a_k, b_k]$

tali che $x^* \in I_k \ \forall k$ $\quad I_k \subseteq I_{k+1} \subseteq \dots \subseteq I_1$ con $f(a_k)\cdot f(b_k) < 0$

calcolo $c_k = \dfrac{a_k + b_k}{2}$

se $f(c_k) = 0$ allora $x^* = c_k$

altrimenti $[a_{k+1}, b_{k+1}] = \begin{cases} [a_k, c_k] & \text{se } f(c_k) > 0 \\ [c_k, b_k] & \text{se } f(c_k) < 0 \end{cases}$

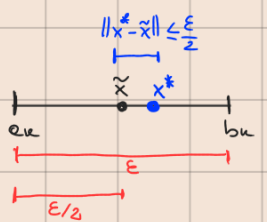
OSSERVAZIONI:

① $c = a + \dfrac{b-a}{2}$ per evitare che c_{k+1} cada fuori dall'intervallo $[a_k, b_k]$

② $|b_k - a_k| < \epsilon + \text{eps} \cdot \max(|b|, |a|)$ dove eps è la precisione macchina (per evitare under/over flow)

Iterando, mi fermo se $l_k = |b_k - a_k|$ (lunghezza intervallo) $\to l_k < \epsilon$

$\tilde{x} \approx c_{k+1} = \dfrac{b_k + a_k}{2}$ punto medio dell'intervallo $\to \|x^* - \tilde{x}\| \leq \dfrac{\epsilon}{2}$



Al passo K:

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

quindi $x^* = c_{k+1} \pm \epsilon_{k+1}$, dove

$$\epsilon_{k+1} \leq \frac{b-a}{2^{k+1}}$$ tolleranza

l'ordine di convergenza rappresenta quanto diminuisce l'errore $|e_k| < C|e_{k-1}|^p$

in questo caso $p=1$, $C=1/2$ $\Rightarrow$ il metodo è lento, di ordine convergente lineare

Metodo delle approssimazioni successive

Cercare $f(x)=0$, è comunque da cercare $g(x)=x$, cioè un punto fisso della funzione g, definita

$$g(x) = x - f(x)\cdot\phi(x)$$

dove $\phi(x)$ è limitata e $\neq 0$ $\quad 0 < |\phi(x)| < \infty \quad x \in [a,b]$

quindi risolvere $f(x)=0$ equivale a risolvere $g(x)=x$

$f(x) = 0 \iff g(x) = x$

$\Rightarrow$ $f(x^*)=0 \Rightarrow g(x^*) = x^* - \overbrace{f(x^*)}^{=0}\cdot\overbrace{\phi(x^*)}^{\neq 0} = x^*$

$\Leftarrow$ $g(x^*) = x^* \Rightarrow f(x^*)=0$

$x^* - f(x^*)\cdot\phi(x^*) = x^* \implies f(x^*)\cdot\overbrace{\phi(x^*)}^{\neq 0} = 0$ allora $f(x^*)=0$

Il punto fisso della funzione $g(x)$ è geometricamente l'intersezione delle curve

$y = x \qquad y = g(x)$

DEF: Convergenza globale = il metodo converge a x^* $\forall x_0$

Convergenza locale = il metodo converge a x^* solo per un $x_0 \in I_\rho$ (intorno centrato in x^* di raggio ρ) dove $I_\rho = [x^* - \rho, x^* + \rho]$

Teorema di esistenza e unicità del punto fisso nel modello continuo ($g(x) = x - f(x)\cdot\phi(x)$ ha il punto fisso?)

Sia $g(x)$ continua in $[a,b]$, tale che $g(x) \in [a,b]$. Sia L una costante $0 \leq L < 1$ tale che $\forall x, y \in [a$

$$|g(x) - g(y)| \leq L|x-y|$$

ossia g è una contrazione in $[a,b]$. Allora esiste un unico punto fisso x^* di $g(x)$ in $[a,b]$

NB accade che g sia continua, $g(x) \in [a,b]$, g non oscilla troppo

Data un'approssimazione iniziale x_0 di x^*, calcolo con una successione di iterati:

$$x_{k+1} = g(x_k)$$

Convergenza del metodo del tipo della funzione

Se $g(x)$ è continua e la successione $\{x_k\}$ converge per $k \to \infty$ ad un punto x^* allora x^* è punto fisso di $g(x)$

$$x^* = \lim_{k\to\infty} x_{k+1} = \lim_{k\to\infty} g(x_k) = g\left(\lim_{k\to\infty} x_k\right) = g(x^*)$$

TEOREMA DI CONVERGENZA GLOBALE del metodo delle approssimazioni successive

quando la successione $\{x_0, \dots x_k\}$ è convergente??

Sia $g(x)$ una funzione definita in $[a,b]$. Sia $g(x)$ continua in $[a,b]$, $g(x) \in [a,b]$, $g(x)$ contrazione in $[a,b]$

Allora $\forall\, x_0 \in [a,b]$, la successione degli iterati $x_k = g(x_{k-1})$ converge per $k \to \infty$ all'unico punto fisso x^* di $g(x)$

TEOREMA DI CONVERGENZA LOCALE

Ip centrato in x^*

Sia x^* punto fisso di $g(x)$; sia $g(x)$ continua e contrazione in $[a,b]$ $\forall\, x \in [x^* - \rho, x^* + \rho] = I_\rho$

Allora $\forall\, x_0 \in I_\rho$, la successione degli $\{x_k\} \in I_\rho$ e converge per $k \to \infty$ all'unico punto fisso x^* di $g(x)$

La condizione $g(x) \in [a,b]$ non è necessaria nella convergenza locale

CRITERI DI ARRESTO

x_k si ritiene approssimazione accettabile se valgono:

$$|f(x_k)| \le \epsilon_1 \quad \text{e} \quad |x_k - x_{k-1}| \le \epsilon_2$$

$|f(x_k)| \to |g(x_k) - x_k|$

Se f è "piatta", la sola condizione $|f(x_k)| \le \epsilon_1$ includerebbe un intervallo troppo grande

oppure se

$$\frac{|f(x_k)|}{f_{max}} \le \sigma_1 \quad \text{e} \quad \frac{|x_k - x_{k-1}|}{|x_k|} \le \sigma_2 \qquad \text{dove} \quad f_{max} = \max_{x \in I_\rho} |f(x)|$$

Metodo di Newton

particolare metodo di punto fisso

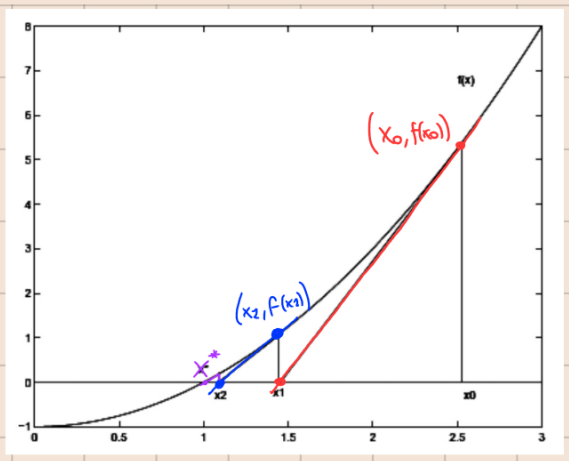
Un metodo iterativo, ha velocità di convergenza lineare se poniamo $\phi(x) \neq \frac{1}{f'(x)}$ con $f'(x) \neq 0$

Se invece poniamo $\phi(x^*) = \frac{1}{f'(x^*)}$ con $f'(x^*) \neq 0$, allora la velocità di convergenza è quadratica

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

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

Il metodo di Newton è detto anche metodo delle tangenti perché geometricamente x_{k+1} è il punto di intersezione tra l'asse delle x ($y=0$) e la retta tangente a $f(x)$ in $(x_k, f(x_k))$



x_1 è l'intersezione tra l'asse delle x, e $f'(x_0)$ (in rosso)

allo stesso modo, x_2 è l'intersezione tra $\begin{cases} y=0 \\ y=f'(x_1) \end{cases}$ (in blu)

SVANTAGGI → costo computazionale più alto (devo calcolare derivata prima)
→ deve essere derivate prime $\neq 0$ $\forall$ punto

Metodo di Newton non applicabile a tutte le g

Convergenza locale metodo di Newton (rispetto al precedente (espresso per g), questo lo esprimo per proprietà di f)

Sia x^* uno zero di $f(x)$. Sia f continua insieme alle sue derivate prime, seconde e terze

Sia $f'(x) \neq 0$ per $x \in I_\rho$ (dove $I_\rho = [x^* - \rho, x^* + \rho]$), e sia $f''(x^*) \neq 0$

Allora $\forall x_0 \in I_\rho$ la successione generata converge a x^* in modo quadratico

Convergenza globale del metodo di Newton

Sia $f \in C^2$ in $[a,b]$. Sia inoltre:

- $f(a) < 0 \quad f(b) > 0$
- $f'(x) \neq 0$
- $f''(x) \leq 0$
- $|f(b)| \leq (b-a)|f'(b)|$

$\Rightarrow$ allora il metodo di Newton genera una successione di iterati convergenti all'unica soluzione di $f(x)=0$ appartenente ad $[a,b]$ e partire da qualunque $x_0 \in [a,b]$

INTRODUZIONE PUNTI DI MINIMO

$f: \mathbb{R}^n \to \mathbb{R}$ → $\nabla f(x)=0$ NON PER FORZA x PUNTO DI MINIMO

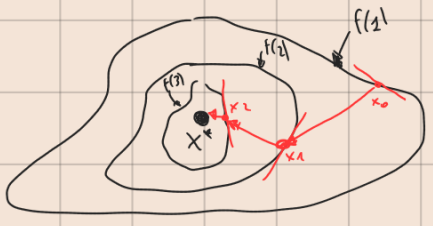
- Se $x^* \in \mathbb{R}^n$ è di minimo se $f \in C^1$ (esistono tutte le n derivate prime parziali e sono continue). allora $\nabla f(x^*) = 0$
  → è chiamato punto stazionario, cioè quando $\nabla f(x^*) = 0$

- Se $x^* \in \mathbb{R}$ è di minimo locale se $\nabla f(x^*) = 0$ e $\nabla^2 f(x^*)$ (cioè hessiana di f) è semidefinita positiva

METODI DI DISCESA

derivata prima = 0 non sufficiente per definire x punto di minimo → serve derivata seconda

Cerchiamo delle direzioni lungo le quali la funzione diminuisce il suo valore



$f(1)$, $f(2)$, $f(3)$ sono curve di livello in cui f decresce il suo valore

Le direzioni ci vengono date dai gradienti, perpendicolare alla tangente calcolata nel punto e curve di livello

Devo decidere la direzione lungo la quale la funzione diminuisce il suo valore, e la "step-length" (passo)

$x_0, x_1, \dots \quad x_k \xrightarrow{k\to\infty} x^*$

p_k è la direzione, d_k il passo

$x_{k+1} = x_k + d_k \cdot p_k$ ad ogni passo $f(x_{k+1}) < f(x_k)$ $\quad d_k \in \mathbb{R}^+$

DEF il vettore p è una direzione di discesa in f in x se esiste un $\bar{d} > 0$ tale che

$$f(x + d\,p) < f(x) \qquad \forall d \in]0, \bar{d}]$$

LEMMA Sia $f \in C^1$, il vettore p è di discesa in f in x se

$$p^T \nabla f(x) < 0$$

ESEMPIO

$f(x) = x_1^2 + 3x_2^2 \qquad P = (5,0) \qquad \bar{x} = (1,1)$ $\quad$ P è di discesa per f in $\bar{x}$?

$\nabla f(x) = (2x_1, \ 6x_2) \qquad \nabla f(\bar{x}) = (2, 6)$

$p^T \nabla f(\bar{x}) = \begin{pmatrix} 5 \\ 0 \end{pmatrix} (2 \ 6) = 10$

NO, perché $p^T \nabla f(x)$ non è < 0 → NO DISCESA

SCELTA DELLA DIREZIONE

Solitamente si sceglie $p_k = -\nabla f(x_k)$ che equivale alla max direzione di decrescita

$$p_k^T \nabla f(x_k) = -\nabla f(x_k)\, \nabla f(x_k) < 0 \quad \text{sempre}$$

NB $\quad v^T \cdot v = \sum_{i=1}^{m} v_i^T \cdot v_i = \sum_{i=1}^{m} v_i^2 \geq 0 \qquad \|v\|_2^2$

$$-\nabla f(x_k)\, \nabla f(x_k) = -\|\nabla f(x_k)\|_2^2 \leq 0$$

SCELTA DEL PASSO

Se scelgo il passo d_k in funzione della condizione $f(x_{k+1}) < f(x_k)$ non è sufficiente

Solitamente più il passo è piccolo, e più ho garanzia della decrescita della funzione. Ma passi troppo piccoli comportano una convergenza lenta. Il passo fisso non sempre garantisce la convergenza e la decrescita della funzione

Pertanto imporre ad ogni iterata $f(x_{k+1}) < f(x_k)$ è condizione sufficiente alla convergenza

Condizioni di Wolfe

I°: condizione di Wolfe - Armijo

condizione di decrescita sufficiente: limita inferiormente la decrescita ad ogni passo

$$f(x_k + \alpha_k p_k) \leq f(x_k) - \alpha_k \gamma$$

dove $\gamma = -\beta_1 \nabla f(x_k)^T p_k$ con $0 < \beta_1 < 1$

$$\rightarrow \quad f(x_{k+1} + \alpha_k p_k) \leq f(x_k) + \alpha_k \beta_1 \nabla f(x_k)^T p_k$$

II°: condizione di Wolfe

condizione di curvatura: Assicura che non vengano compiuti passi troppo corti

$$\frac{\nabla f(x_k + \alpha_k p_k)^T p_k}{\nabla f(x_k)^T p_k} \geq \beta_2 \qquad 0 < \beta_2 < 1$$

$$\nabla f(x_k + \alpha_k p_k)^T p_k \geqslant \beta_2 \nabla f(x_k)^T p_k$$

Uniamo le due condizioni

Posto $\alpha_0 = 1$, lo riduco fino a trovare un α_k che soddisfa la condizione di Armijo

Algoritmo di backtracking

① $\bar{\alpha} = 1$, $\rho > 0$, $c \in (0,1)$, $\alpha \leftarrow \bar{\alpha}$ solitamente $\rho = 1/2$

② while $f(x_k + \alpha p_k) \leq f(x_k) + \alpha \cdot c \cdot \nabla f(x_k)^T \cdot p_k$

do $\alpha = \alpha \cdot \rho$ endwhile

③ $\alpha_k = \alpha$

precisione macchina

Nel while si può aggiungere anche una condizione sulla lunghezza di α. Se $|\alpha| < \varepsilon$ → STOP

Teorema di convergenza:

Supponiamo f limitata inferiormente e differenziabile allora:

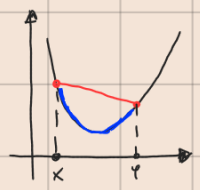
il metodo di discesa che calcola gli iterati $x_{k+1} = x_k + \alpha_k p_k$ con p_k direzione di discesa per f in x

e α_k calcolato con algoritmo di backTracking, converge a un punto di min o di sella di f ($\nabla f(x^*$

FUNZIONI CONVESSE

$f: \mathbb{R}^n \to \mathbb{R}$ è convessa se (strettamente convessa)

$$f(tx + (1-t)y) \underset{(<)}{\leq} t f(x) + (1-t) f(y) \qquad \forall t \in [0,1], \ \forall x, y \in \mathbb{R}^n$$



f converse tutti i punti compresi tra x, y, sono $\leq$ del segmento che unisce i due punti x, y

Lemma: Sia $f \in C^2(\mathbb{R}^n)$, $x \in \mathbb{R}^n$, allora vale

$$f \text{ convessa} \iff \nabla^2 f(x) \text{ è semidefinita positiva} \qquad \text{(Hessiana di } f\text{)}$$

Teorema Se f è convessa un punto di minimo locale è un punto di minimo globale

f convessa → ogni punto di minimo locale x^* è minimo globale di f

f strettamente convessa → esiste un unico punto di minimo locale (il metodo di discesa va verso il min glob

ES FUNZIONE CONVESSA

$f: \mathbb{R}^n \to \mathbb{R}$ $\qquad \min_{x \in \mathbb{R}^n} \underbrace{\|Ax - b\|_2^2}_{f}$

Se A è di rango massimo → f è strettamente convessa

$\nabla f(x) = A^T Ax - A^T b = 0 \longrightarrow A^T Ax = A^T b$

$\nabla^2 f(x) = A^T A$ se A è di rango massimo → $A^T A$ è definita positiva

Altri metodi di discesa

Nel metodo di Newton $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$ $\qquad \left(p_k = -\frac{1}{f'(x_k)}, \ d_k = f(x_k) \right)$

Metodi tipo Newton:

$$p_k = -\underbrace{\left(\nabla^2 f(x_k)\right)^{-1}}_{\text{hessiana}} \cdot \nabla f(x_k)$$

→ Non inverto la matrice, ma calcolo

calcolo la direzione di discesa risolvendo $\nabla^2 f(x_k) \cdot p_k = -\nabla f(x_k)$

Si dimostra che p_k è direzione di discesa per f in x_k se $\nabla^2 f(x_k)$ è definita positiva

ogni iterazione ha una complessità maggiore, ma servono meno iterazioni per la convergenza