



*centroappunti.it*

**CORSO LUIGI EINAUDI, 55/B - TORINO**

**Appunti universitari**

**Tesi di laurea**

**Cartoleria e cancelleria**

**Stampa file e fotocopie**

**Print on demand**

**Rilegature**

**NUMERO: 2574A**

**ANNO: 2024**

# **A P P U N T I**

**STUDENTE: Scaringi Nicolò**

**MATERIA: Metodi Numerici e Calcolo Scientifico - Prof.  
Pieraccini**

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 STUDENTI E NON SONO STATI VISIONATI DAL DOCENTE.  
IL NOME DEL PROFESSORE, SERVE SOLO PER IDENTIFICARE IL CORSO.



**Politecnico  
di Torino**

# **METODI NUMERICI E CALCOLO SCIENTIFICO**

Prof. ssa Sandra Pieraccini

1° anno

*Nicolò Scaringi*

# INDICE

|     |  |    |
|-----|--|----|
| 1.  | RAPPRESENTAZIONE DEI NUMERI                          | 1  |
| 1.1 | Sistema floating point                               | 1  |
| 1.2 | Stabilità e condizionamento                          | 2  |
| 2.  | SISTEMI LINEARI                                      | 3  |
| 2.1 | Algoritmi di algebra lineare                         | 5  |
| 2.2 | Metodi diretti                                       | 4  |
| 2.3 | Metodi iterativi                                     | 5  |
| 3.  | APPROSSIMAZIONI DI DATI E FUNZIONI                   | 11 |
| 3.1 | Interpolazione polinomiale                           | 11 |
| 3.2 | Interpolazione polinomiale a tratti                  | 12 |
| 3.3 | Metodo dei minimi quadrati                           | 13 |
| 3.4 | Derivazione numerica                                 | 13 |
| 4.  | QUADRATURA NUMERICA                                  | 15 |
| 4.1 | Formule interpolatorie                               | 15 |
| 4.2 | Quadratura composta                                  | 16 |
| 5.  | EQUAZIONI NON LINEARI                                | 17 |
| 5.1 | Metodo di bisezione                                  | 17 |
| 5.2 | Metodo di Newton e delle tangenti                    | 17 |
| 5.3 | Problema di punto fisso                              | 18 |
| 6.  | EQUAZIONI DIFFERENZIALI ORDINARIE (ODE)              | 20 |
| 6.1 | Metodi numerici                                      | 20 |
| 6.2 | Convergenza  | 20 |
| 6.3 | Stabilità  | 20 |
| 7.  | DISCRETIZZAZIONE DI EQUAZIONI ALLE DERIVATE PARZIALI | 22 |
| 7.1 | Modello di filo elastico e di sbarra termica         | 23 |
| 7.2 | Modello di membrana elastica e di piastra termica    | 27 |
| 7.3 | Modello di evoluzione temporale                      | 28 |
| 7.4 | Modello di convezione-diffusione e di trasporto      | 31 |
| 7.5 | Legge di conservazione                               | 32 |



# 1 RAPPRESENTAZIONI DI NUMERI

## 1.1 RAPPRESENTAZIONE DEI NUMERI SUL CALCOLATORE

### SISTEMA FLOATING POINT

fissata la base  $\beta$  di un sistema di numerazione, la rappresentazione floating point (virgola mobile) normalizzata di un numero reale consiste nell'usare una opportuna potenza di  $\beta$  in modo da non avere né parte intera e né zeri dopo la virgola.  $123,4567 \rightsquigarrow 0,1234567 \cdot 10^3$ . Su un calcolatore fissiamo:

- i)  $\beta \in \mathbb{N}^+, \beta \geq 2$ , la base del sistema di numerazione
- ii)  $t \in \mathbb{N}^+$ , numero di cifre di mantissa a disposizione sul calcolatore
- iii)  $L < 0$  e  $U > 0$ , rispettivamente limite minimo e massimo entro cui può variare la caratteristica.

L'insieme dei numeri rappresentabili su tale calcolatore è chiamato insieme dei numeri macchina (floating point):

$$F(\beta, t, L, U) = \left\{ 0 \right\} \cup \left\{ x \in \mathbb{R} : x = \pm d_1 d_2 \dots d_t \cdot \beta^e \right\} \quad \text{con } m = d_1 d_2 \dots d_t : \text{mantissa}$$

$$0 \leq d_s \leq \beta - 1 \text{ e } d_1 \neq 0 \quad \left\{ \begin{array}{l} s=0 \text{ per } + \\ s=1 \text{ per } - \end{array} \right.$$

$$L \leq e \leq U : \text{caratteristica}$$

### APPROSSIMAZIONE DI UN NUMERO REALE CON UN NUMERO DI MACCHINA

Non è possibile rappresentare alcun numero [a parte lo zero] minore in valore assoluto a  $x_{\min} = \beta^{L-1}$ . Per aggirare questa limitazione lo standard IEEE754 prevede una rappresentazione denormalizzata. Quando  $e = L$ , la condizione  $d_1 \neq 0$  può essere abbandonata e quindi vengono accettati  $\beta^L < m < \beta^{L-1} \cdot \beta^{-t}$

Un insieme di numeri reali è rappresentato da un numero di macchina, in base alla vicinanza.

Sea  $x$  un numero reale e  $\tilde{x}$  la sua approssimazione / modo macchina:  $x = (-1)^s m \cdot \beta^e, \tilde{x} = (-1)^s \tilde{m} \cdot \beta^e$

L'errore assoluto è  $E_a = |\tilde{x} - x|$ , l'errore relativo è  $E_r = \frac{|\tilde{x} - x|}{|x|}, x \neq 0$

Le tecniche di approssimazione (per lo standard IEEE754-2008) più utilizzate sono:

- a) arrotondamento  $\rightarrow$  approssima con modo macchina (m.m.) più vicino; se il numero cade esattamente a metà strada tra due m.m. consecutivi, approssima con quello con cifra meno significativa pari.
- b) troncamento  $\rightarrow$  taglia tutte le cifre della mantissa dopo la t-esima.

### ERRORI COMMESSI

Con troncamento tutte le mantisse  $m \in [m_1, m_1 + \beta^{-t}]$  vengono approssimate con  $\tilde{m} = m_1$ , l'errore commesso è  $m - m_1 < \beta^{-t}$  ed è sempre positivo. Con arrotondamento tutte le mantisse che cadono su  $(m_1 - \frac{1}{2} \beta^{-t}, m_1 + \frac{1}{2} \beta^{-t})$  vengono approssimate con  $m_1$ , l'errore commesso è  $|m - m_1| \leq \frac{1}{2} \beta^{-t}$  e può essere sia positivo che negativo.

L'arrotondamento è sicuramente più oneroso dal punto di vista di istruzioni del microprocessore, ma provoca un errore metà degli altri ed è sicuramente più conveniente dal punto di vista dell'accuratezza.

L'errore assoluto reale  $|x - \tilde{x}| = |m - \tilde{m}| \beta^e < \beta^{e-t}$  troncamento;  $|x - \tilde{x}| = |m - \tilde{m}| \beta^e \leq \frac{1}{2} \beta^{e-t}$  arrotondamento.

L'errore relativo è  $\frac{|x - \tilde{x}|}{|x|} < \frac{\beta^{e-t}}{\beta^{e-1}} < \beta^{1-t} =: E_m$  troncamento;  $\frac{|x - \tilde{x}|}{|x|} \leq \frac{\frac{1}{2} \beta^{e-t}}{\beta^{e-1}} < \frac{1}{2} \beta^{1-t} =: E_m$  arrotondamento

$\epsilon_{ps} = \beta^{1-t}$  è l'epsilon di macchina.  $E_m$  è la precisione di macchina, è una caratteristica dell'aritmetica di macchina e rappresenta la massima precisione relativa di calcolo raggiungibile sul calcolatore, due quantità la cui differenza relativa sia minore della precisione di macchina sono da considerarsi indistinguibili per il calcolatore.

Standard IEEE: MATLAB di default utilizza una doppia precisione: 64 bit così ripartiti: 1 bit per il segno + 11 bit per la caratteristica + 52 bit per la mantissa. In floating point  $0, d_1 d_2 \dots d_t \beta^e$ , con l'hidden bit  $d_1 = 1$  sempre per cui non spreca memoria per rappresentarlo, in questo modo ha 53 bit per la mantissa.  $E_m = \frac{1}{2} \beta^{1-t} = \frac{1}{2} 2^{1-53} = 2^{1-53-1} = 2^{-53} \approx 10^{-16}$

In MATLAB: format short stampa a schermo fino alla 4<sup>a</sup> cifra decimale, solo per fini grafici di visualizzazione format long stampa 16 cifre decimali, che è il massimo rappresentabile.

L'overflow è l'errore dovuto al tentativo di rappresentare numeri con  $e > U$ . L'underflow è l'errore dovuto al tentativo di

# 2 SISTEMI LINEARI

## 2.1 RICHAMI DI ALGEBRA LINEARE

Consideriamo sistemi lineari quadrati di  $n$  equazioni in  $n$  sconosciute e ben posti  $Ax=b$ , dove  $A \in \mathbb{R}^{n \times n}$  matrice di coef. fissati,  $b \in \mathbb{R}^n$  vettore [colonna] di termini noti,  $x \in \mathbb{R}^n$  vettore [colonna] soluzione. La soluzione è  $x=A^{-1}b$ ,  $A^{-1}$  è un formalismo -  $A^{-1} = \text{inv}(A)$  il comando inv si può usare solo se la matrice è diagonale. Se il sistema è ben posto  $\det(A) \neq 0$ ,  $A$  è quindi una matrice invertibile o non singolare [ $A$  singolare  $\Rightarrow \det(A)=0$ ]

### MATRICI CON STRUTTURE PARTICOLARI

Una matrice  $A$  si dice diagonale se  $a_{ij} = 0$  se  $i \neq j$

Una matrice  $A$  si dice triangolare superiore se  $a_{ij} = 0$  se  $i > j$ , triangolare inferiore se  $a_{ij} = 0$  se  $i < j$

Proprietà matrice diagonale: una matrice diagonale  $D$  è invertibile se  $d_{ss} \neq 0 \forall s$ , gli elementi di  $D^{-1}$  sono  $\frac{1}{d_{ss}}$ ; la moltiplicazione DA risale le righe, AD risale le colonne.

DEF  $A \in \mathbb{R}^{n \times n}$  si dice a predominanza diagonale / diagonale dominante per righe se  $\forall s=1, \dots, n \quad |a_{ss}| > \sum_{j \neq s} |a_{sj}|$   
 analogo definizione per predominanza diagonale per colonne.

DEF  $A \in \mathbb{R}^{n \times n}$  simmetrica si dice positiva se  $x^T A x > 0 \quad \forall x \neq 0$

PROP  $A \in \mathbb{R}^{n \times n}$  simmetrica è definita positiva  $\Leftrightarrow \lambda_s(A) > 0 \quad \forall s=1, \dots, n$

### NORME DI VETTORI E MATRICI

DEF La norma di vettore (matrice) è una funzione  $\|\cdot\|: V \rightarrow \mathbb{R}$  ( $V = \mathbb{R}^n$  per vettori,  $V = \mathbb{R}^{n \times n}$  per matrici) tale da:

i)  $\|x\| \geq 0 \quad \forall x \in V$

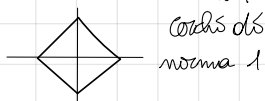
ii)  $\|\alpha x\| = |\alpha| \|x\| \quad \forall x \in V, \forall \alpha \in \mathbb{R}$

iii)  $\|x+y\| \leq \|x\| + \|y\| \quad \forall x, y \in \mathbb{R}^n$  (disuguaglianza triangolare)

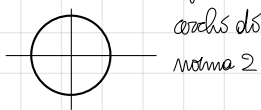
per le norme di matrici si)  $\|AB\| \leq \|A\| \|B\| \quad \forall A, B \in \mathbb{R}^{n \times n}$  (submoltiplicatività)

Norme di vettori più usate:

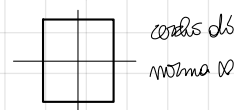
• norma 1  $\|x\|_1 = \sum_{i=1}^n |x_i|$



• norma 2/euclidea/spettale  $\|x\|_2 = \left(\sum_{i=1}^n x_i^2\right)^{1/2}$



• norma  $\infty$ /del massimo  $\|x\|_\infty = \max |x_i|$



Nello spazio di dimensione finita, le norme sono equivalenti

Norme di matrice più usate:

• norma 1  $\|A\|_1 = \max_j \sum_{i=1}^n |a_{ij}|$

norma verticale

• norma 2/euclidea/spettale  $\|A\|_2 = \sqrt{\rho(A^T A)}$  con  $\rho(B) = \max_s |\lambda_s(B)|$

• norma  $\infty$ /del massimo  $\|A\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|$

norma orizzontale

• norma di Frobenius  $\|A\|_F = \left(\sum_{i=1}^n \sum_{j=1}^n |a_{ij}|^2\right)^{1/2}$

Ogni norma di vettore  $\|\cdot\|_V$  induce una norma di matrice:  $\|A\|_M = \max_{\|x\|_V=1} \|Ax\|_V = \sup_{\|x\|_V=1} \frac{\|Ax\|_V}{\|x\|_V}$

$\|\cdot\|_M$  è detta norma normale/sudetta

$\|A\|_*$  è sudetta da  $\|\cdot\|_*$  per  $*$  = 1, 2,  $\infty$

$\|A\|_F$  non è normale, infatti  $\|I\| = 1$  per ogni norma ma  $\|I\|_F = \sqrt{n}$

La norma sudetta è il massimo allungamento che la matrice applica ad un vettore che ruota nello spazio.

PROP compatibilità:  $\|\cdot\|_M$  e  $\|\cdot\|_V$  sono dette compatibili se  $\|Ax\|_V \leq \|A\|_M \|x\|_V \quad \forall A, x$

Le norme 1, 2,  $\infty$  matrice e vettore sono compatibili.

### NUMERO DI CONDIZIONAMENTO

Se perturba  $b$ :  $b \rightarrow \delta b$ , se sistema  $Ax=b$  diventa  $Ax=b+\delta b$ . Sia  $\delta x = x - x_0$ , allora  $A(x+\delta x) = b+\delta b \Rightarrow Ax + A\delta x = b+\delta b \Rightarrow A\delta x = \delta b \Rightarrow \delta x = A^{-1}\delta b$ . Mancano due norme compatibili:

di le prime  $k$  righe e  $k$  colonne di  $A$ .

TH Sia  $A$  una matrice di ordine  $n$ . Se  $A_k$  è non singolare per  $k=1, \dots, n-1$  allora  $\exists!$  la fattorizzazione  $LU$  di  $A$ .  
 Se matrice a diagonale dominante per righe o per colonne e la matrice simmetrica definita positiva soddisfa il sistema del teorema precedente. Se anche  $A$  è singolare, ma  $\det(A_k) \neq 0, k=1, \dots, n-1$  esiste comunque la fattorizzazione  $LU$ , con la matrice  $U$  che avrà un elemento non nullo.

### PIVOTING

TH Sia  $A$  una matrice di ordine  $n$ , allora esiste una matrice di permutazione  $P$  per cui si può ottenere la fattorizzazione  $LU$  di  $A$ , cioè  $PA=LU$ .

Se durante l'eliminazione di Gauss al passo  $k$ -esimo risulta  $a_{kk}^{(k)} = 0$ , si scambiano due equazioni in modo che il nuovo elemento pivot  $a_{kk}^{(k)} \neq 0$  sia diverso da zero; la scelta è sempre tra le equazioni che seguono la  $k$ -esima.

Il  $k$ -esimo passo può essere rappresentato come  $A^{(k+1)} = M_k P_k A^{(k)}$ . Posto  $L = M^{-1}$  si ha  $U = MPA \Rightarrow LU = M^{-1}MPA \Rightarrow LU = PA$ , il sistema lineare  $Ax=b$  equivale a  $LUx=PB$ , posto  $y=Ux$  si ottiene  $Ly=PB$ . Si risolve  $Ly=PB$  per ricavare  $y$ , dato  $y$  si risolve  $Ux=y$  per ricavare  $x$ .

Comando MATLAB:  $[L,U,P] = lu(A)$  fattorizzazione  $PA=LU$

$x=A \setminus b$  metodo eliminazione di Gauss con pivoting parziale

Se  $A$  è simmetrica e definita positiva, essendo l'eliminazione gaussiana stabile senza pivoting, si può sfruttare la simmetria di  $A$  per ridurre al costo computazionale del calcolo della fattorizzazione usando la fattorizzazione di Cholesky anziché la fattorizzazione  $LU$ .

TH Fattorizzazione di Cholesky. Sia  $A$  una matrice di ordine  $n$  simmetrica e definita positiva, allora esiste un'unica matrice traug  $R$  con elementi diagonali positivi tale che  $A=R^T R$ .

Il costo computazionale della fattorizzazione di Cholesky è  $\mathcal{O}\left(\frac{n^3}{6}\right)$

Comando MATLAB  $R = chol(A)$ , il comando assume che  $A$  sia simmetrica definita positiva e usa solo la parte traug

Un'altra fattorizzazione è la fattorizzazione  $QR$  ( $qr(A)$ ):  $Q$  è ortogonale,  $R$  è traug; questa fattorizzazione è molto stabile ma anche più costosa della fattorizzazione  $PA=LU$ .

## 2.3

## METODI ITERATIVI

Per sistemi con matrici dense i metodi diretti sono generalmente preferibili (medie dimensioni  $\sim 10^3 - 10^5$ )

Per sistemi con matrici sparse e di ordine elevato, i metodi diretti diventano impraticabili a causa del fill in, per cui non può essere sfruttata la sparsità della matrice. Una matrice si dice sparsa se  $n^0 \text{ elementi} \neq 0 \ll n^2 \text{ elementi totali}$ .

Per matrici dense di ordine elevato si utilizzano metodi iterativi "ad hoc".

$\exists$  metodi iterativi basati su alterata la struttura di  $A$  e costruiscono una successione infinita di vettori  $\{x^{(k)}\}_{k \geq 0}$ , che sotto opportune ipotesi converge alla soluzione  $x$ . È importante considerare: i) condizioni di applicabilità, ii) costo computazionale, iii) problemi connessi alla convergenza (convergenza sì/no e velocità di convergenza).

DEF Una successione di vettori  $\{x^{(k)}\}_{k \geq 0}$  di  $\mathbb{R}^m$  si dice convergente ad un vettore  $x \in \mathbb{R}^m$  se esiste una norma per cui

$$\lim_{k \rightarrow \infty} \|x^{(k)} - x\| = 0 \quad [\text{indipendentemente dalla norma utilizzata}] \text{ e in tali casi si pone } \lim_{k \rightarrow \infty} x^{(k)} = x$$

grazie all'equivalenza delle norme su  $\mathbb{R}^m$  se una successione di vettori converge su una norma di  $\mathbb{R}^m$ , allora converge su tutte le norme di  $\mathbb{R}^m$ . Si ricorda che  $\{x^{(k)}\}_{k \geq 0}$  è una successione di vettori,  $\{\|x^{(k)} - x\|\}_{k \geq 0}$  è una successione numerica. la condizione di convergenza su una qualunque norma di  $\mathbb{R}^m$  si traduce in una condizione di convergenza per componenti:

$$\lim_{k \rightarrow \infty} x^{(k)} = x \Leftrightarrow \lim_{k \rightarrow \infty} x_s^{(k)} = x_s \quad \forall s = 1, \dots, m$$

TH Sia  $A \in \mathbb{R}^{n \times n}$ , si consideri la successione di matrici in cui al  $k$ -esimo termine è  $A^{(k)} := A^k$ . Svolgendo con  $\rho(A)$  il raggio spettrale di  $A$ , si ha  $\lim_{k \rightarrow \infty} A^k = 0 \Leftrightarrow \rho(A) < 1$ . Si ricorda che  $\rho(A) = \max_{1 \leq s \leq n} |\lambda_s(A)|$

PROP Per ogni norma di matrice compatibile con una norma di vettore si ha  $\rho(A) \leq \|A\|$ .

### METODI ITERATIVI STAZIONARI

Dato un sistema  $Ax=b$  con  $\det(A) \neq 0$ , data una stima iniziale  $x^{(0)}$  della soluzione  $x$  del sistema lineare, si costruisce una successione  $\{x^{(k)}\}_{k \geq 0}$  che - a riga - converge ad  $x$ , risolvendo ad iterazione dei sistemi lineari molto più semplici; ogni (5)



## TEST DI ARRESTO

Un metodo iterativo fornisce una successione infinita di stimate  $x^{(k)}$ ; è necessario stabilire un modo per interrompere il procedimento di generazione delle stimate, cioè stabilire un criterio di arresto della iterazione: se si potesse calcolare l'errore  $e^{(k)} = x^{(k)} - x$  ci si fermerebbe quando  $\|e^{(k)}\|$  è sufficientemente piccola, ma per farlo occorrerebbe conoscere la soluzione esatta che invece è l'incognita, si considera la distanza (assoluta o relativa) tra due iterazioni successive e si confronta con una tolleranza; si arresta il processo iterativo quando

$$a) \frac{\|x^{(k+1)} - x^{(k)}\|}{\|x^{(k+1)}\|} \leq tol_{rel}$$

$$b) \|x^{(k+1)} - x^{(k)}\| \leq tol_{abs}$$

Strada alternativa per definire un test di arresto consiste nel controllo del residuo dell'equazione: se  $x$  è la soluzione si ha  $b - Ax = 0$ , dato un qualunque vettore  $\hat{x}$  si definisce residuo del sistema lineare il vettore  $r = b - A\hat{x}$ ; quindi  $r = 0 \Leftrightarrow \hat{x}$  è la soluzione del sistema lineare

Si noti che residuo piccolo non garantisce errore piccolo  $\frac{\|e^{(k)}\|}{\|x\|} \leq K(A) \frac{\|r^{(k)}\|}{\|b\|}$ . In ogni caso qualsiasi test di arresto si mes è sempre indispensabile prevedere un n° massimo d'iterazioni.

## METODI DEL GRADIENTE

Se  $A$  è simmetrica e definita positiva, la soluzione del sistema lineare  $Ax = b$  può essere caratterizzata come l'unico punto di minimo del funzionale  $J: \mathbb{R}^n \rightarrow \mathbb{R}, J(x) = \frac{1}{2} x^T Ax - x^T b$ . Infatti  $\nabla J(x) = Ax - b \Rightarrow \nabla^2 J(x) = A$  (matrice Hessiana di  $J$ ), quindi i punti stazionari coincidono con eventuali soluzioni del sistema lineare.

ES 1  $J: \mathbb{R}^3 \rightarrow \mathbb{R} \quad J(x) = \frac{1}{2} [x_1 \ x_2 \ x_3] A \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} - [x_1 \ x_2 \ x_3] b$

$$\nabla J(x) = \begin{bmatrix} \frac{\partial J}{\partial x_1} \\ \frac{\partial J}{\partial x_2} \\ \frac{\partial J}{\partial x_3} \end{bmatrix} \Rightarrow \begin{cases} \frac{\partial J}{\partial x_1} = \frac{1}{2} [1 \ 0 \ 0] A \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \frac{1}{2} [x_1 \ x_2 \ x_3] A \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} - [1 \ 0 \ 0] b = \frac{1}{2} (A_{1,:})x + \frac{1}{2} x^T A(:,1) - b_1 = A(1,:)x - b_1 \\ \frac{\partial J}{\partial x_2} = A(2,:)x - b_2 \\ \frac{\partial J}{\partial x_3} = A(3,:)x - b_3 \end{cases} \Rightarrow \nabla J(x) = Ax - b$$

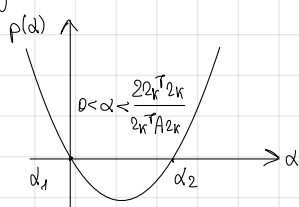
ES 2  $J: \mathbb{R} \rightarrow \mathbb{R} \quad J(x) = \frac{1}{2} ax^2 - bx \quad \text{con } a, b \in \mathbb{R} \quad \leadsto \text{parabola}$

ES 3  $J: \mathbb{R}^2 \rightarrow \mathbb{R} \quad J(x) = \frac{1}{2} [x \ y] A \begin{bmatrix} x \\ y \end{bmatrix} - [x \ y] b \quad \leadsto \text{paraboloide}$

$J$  è un funzionale quadratico, la direzione  $-\nabla J(x)$  è ortogonale alle linee di livello e rappresenta la direzione più rapida di discesa.

Si nota che  $-\nabla J(x) = b - Ax = r(x)$ . Ssa  $x_k \in \mathbb{R}^n$  fissato e ci si sposta da  $x_k$  lungo la direzione  $-\nabla J(x)$  di un passo  $\alpha_k$ :  $x_{k+1} = x_k + \alpha_k \nabla J(x) = x_k + \alpha_k r(x_k)$ . Per semplicità di notazione  $r(x_k) = r_k$ . Quindi  $J(x_k + \alpha_k r_k) = \frac{1}{2} (x_k + \alpha_k r_k)^T A (x_k + \alpha_k r_k) + (x_k + \alpha_k r_k)^T b = \frac{1}{2} x_k^T A x_k + \frac{1}{2} \alpha_k^2 r_k^T A r_k + \alpha_k r_k^T A x_k - \alpha_k r_k^T b = J(x_k) - \alpha_k r_k^T r_k + \frac{1}{2} \alpha_k^2 r_k^T A r_k$

Se  $\alpha_k$  è tale che  $\frac{1}{2} \alpha_k^2 r_k^T A r_k - \alpha_k r_k^T r_k$  sia negativo e più grande possibile in modulo, si avrebbe la maggior riduzione possibile di  $J$  muovendosi nella direzione  $r_k$ .



Il termine considerato rappresenta una parabola  $p = p(\alpha) = \frac{1}{2} \alpha^2 r_k^T A r_k - \alpha r_k^T r_k$  con concavità verso l'alto perché  $A$  è simmetrica definita positiva e quindi per definizione  $r_k^T A r_k > 0$ ,  $r_k$  è fissato  $r_k = b - Ax_k$ . Le soluzioni sono  $\alpha_1 = 0 \quad \alpha_2 = \frac{2 r_k^T r_k}{r_k^T A r_k} > 0$

Il valore ottimale per la minor riduzione possibile di  $J$  è il 1° membro  $\hat{\alpha} = \frac{r_k^T r_k}{r_k^T A r_k}$

3 passaggi da effettuare per il metodo del gradiente sono:

dato  $x_0$  per  $k \geq 0$

- 1) calcolare  $r_k = b - Ax_k$
- 2) calcolare  $\alpha_k = \frac{r_k^T r_k}{r_k^T A r_k}$
- 3) calcolare  $x_{k+1} = x_k + \alpha_k r_k$

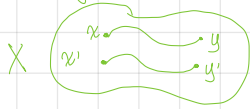
Per risolvere il modo prodotto matrice - vettore:

dato  $x_0, r_0 = b - Ax_0$ , per  $k \geq 0$

- 1) calcolare  $q_k = A r_k$
- 2) calcolare  $\alpha_k = \frac{r_k^T r_k}{r_k^T q_k}$
- 3) calcolare  $x_{k+1} = x_k + \alpha_k r_k$

gli autovalori di  $A$  appartengono all'unione dei cerchi di raggio  $r$  per colonna  $\lambda_s(A) \in \bigcup_{j=1}^m C_j =: C$ , e quindi appartengono all'insieme  $\mathbb{R} \cup C$ .  
 Cioè è possibile per cui  $A$  e  $A^T$  hanno lo stesso spettro.

**TA II** di Geršgorin - Ogni componente connessa di  $\mathbb{R}$  contiene tanti autovalori quanti sono i cerchi che la compongono.  
 Insieme formato da una sola componente connessa | Insieme formato da due componenti connesse



**DEF** Una matrice si dice **risolvibile** se esiste una matrice di permutazione  $P$  tale che  $PAP^T$  è triangolare a blocchi.  $A_{11}$  e  $A_{22}$  matrici quadrate.

$$PAP^T = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}$$

Si noti che la risolubilità di una matrice  $A$  può essere sfruttata nella fase di risoluzione di un sistema lineare  $Ax=b$  in cui essa sia matrice dei coefficienti

$$Ax=b \Rightarrow \begin{cases} A_{22}x_2 = b_2 \\ A_{11}x_1 = b_1 - A_{12}x_2 \end{cases}$$

**TA III** di Geršgorin - Se  $A$  è irrisolvibile ed  $\exists \lambda \in \mathbb{R} \Rightarrow \lambda \in \mathcal{D}R_s$   $\forall s$ , analogamente  $\exists \lambda \in \mathbb{C} \Rightarrow \lambda \in \mathcal{D}C_s$

## IN SINTESI

### RICHIAMI DI ALGEBRA LINEARE

- Norme di matrice più usate:
  - norma 1  $\|A\|_1 = \max_j \sum_{i=1}^n |a_{ij}|$  norma verticale
  - norma 2 / euclidea / spettrale  $\|A\|_2 = \sqrt{\rho(A^T A)}$  con  $\rho(B) = \max_s |\lambda_s(B)|$
  - norma  $\infty$  / del massimo  $\|A\|_\infty = \max_s \sum_{j=1}^n |a_{sj}|$  norma orizzontale
  - norma di Frobenius  $\|A\|_F = \left( \sum_{i=1}^n \sum_{j=1}^n |a_{ij}|^2 \right)^{1/2}$
- Numero di condizionamento  $K(A) = \|A\| \|A^{-1}\| (> 1 \text{ per norme naturali})$  Comando MATLAB cond(A)

### METODI DIRETTI

- **Eliminazione gaussiana** - Dato un sistema lineare  $Ax=b$  le seguenti operazioni conducono ad un sistema equivalente (con stessa soluzione): 1) scambiare due equazioni; 2) moltiplicare un'equazione per uno scalare; 3) sostituire un'equazione con una combinazione lineare della stessa e di un'altra equazione. Si passa da  $Ax=b$  a  $A^{(n)}x=b^{(n)}$  con  $A^{(n)}$  Triang. sup.
  - **Fattorizzazione LU** - Si fattorizza  $A=LU$  con  $L$ : triang. inf.,  $U$ : triang. sup. - Il sistema si risolve nel seguente modo:  $Ax=b \Rightarrow Ux=b$ , posto  $y=Ux$  si ottiene  $Ly=b$  - Si risolve  $Ly=b$  per ricavare  $y$ ; noto  $y$  si risolve  $Ux=y$  per ricavare  $x$ .
  - **Pivoting** - Si fattorizza  $PA=LU$ , posto  $L=M^{-1}$  si ha  $U=MPA \Rightarrow LU=M^{-1}MPA \Rightarrow U=PA$ , il sistema originale  $Ax=b$  è equivalente a  $LUx=PB$ , posto  $y=Ux$  si ottiene  $Ly=PB$  - Si risolve  $Ly=PB$  per ricavare  $y$ , noto  $y$  si risolve  $Ux=y$  per ricavare  $x$ .
- Comandi MATLAB:  $[L,U,P]=lu(A)$  fattorizzazione  $PA=LU$   
 $x=A \setminus b$  metodo eliminazione di Gauss con pivoting parziale

### METODI ITERATIVI

- **Metodi iterativi stazionari** -  $Ax=(M+N)x=b \Leftrightarrow Mx=-Nx+b$  - Dato  $x^{(k)}$ , si calcola  $x^{(k+1)}$  in modo da soddisfare  $x^{(k+1)} = -M^{-1}Nx^{(k)} + M^{-1}b$ , posto  $B=-M^{-1}N$  (matrice d'iterazione) e  $c=M^{-1}b$  si ha  $x^{(k+1)} = Bx^{(k)} + c$
- Nel metodo di Jacobi si considera il seguente splitting di  $A$ :  $A=E+D+F$ ,  $A = \begin{pmatrix} E & & F \\ & D & \\ & & F \end{pmatrix}$  e si prende  $M=D$  e  $N=E+F$ ,  
 $\Rightarrow B=-D^{-1}(E+F)$ , in forma compatta  $x^{(k+1)} = -M^{-1}Nx^{(k)} + M^{-1}b \Rightarrow x^{(k+1)} = D^{-1}(b - Ex^{(k)} - Fx^{(k)})$
- Implementazione in MATLAB dei comandi di Jacobi:  $A, b$  assegnati,  $x_0$  assegnato
- $$M = \text{diag}(\text{diag}(A));$$
- $$N = A - M;$$
- for  $k=1:m_{\text{MAX}}$
- $$x_{\text{new}} = M \setminus (b - N * x_0);$$
- <condizione per arrestare le iterazioni >
- $$x_0 = x_{\text{new}};$$

# 3 APPROSSIMAZIONE DI DATI E FUNZIONI

Si affida quando: a) si hanno  $(x_s, y_s)$  si vogliono approssimare con una funzione tale da  $y_s = f(x_s)$   
 b) data una funzione la si vuole approssimare con una più semplice, come un polinomio.

Data  $f \in F = C^k[a, b]$  (spazio di funzioni a dimensioni infinite) occorre:

- i) introdurre un sottospazio  $F_m$  di funzioni di dimensione finita, come: polinomi algebrici di grado  $m$   $P_m$  ( $\dim(P_m) = m+1$ ), funzioni polinomiali a tratti costruite da  $N$  tratti  $P_m^T$ , funzioni spline di ordine  $r$   $S_r$ , polinomi trigonometrici  $T_m$ .
  - ii) introdurre un criterio per determinare  $f_m \in F_m$ , come: interpolazione, minimi quadrati.
- Si noti che  $\dim F = n^o$  elementi della base di  $F_m = n^o$  parametri da introdurre  $f_m \in F = n^o$  condizioni da imporre. Nel caso dell'interpolazione, le condizioni di interpolazione sono tanto quanto i nodi  $(m+1)$ , e il grado massimo del polinomio che si può costruire è quindi  $m$ .  $\dim F = m+1$

## 3.1 INTERPOLAZIONE POLINOMIALE

La norma di funzione è  $\|f\|_\infty = \max_{x \in [a, b]} |f(x)|$  dove  $[a, b]$  è l'intervallo in cui si vuole approssimare  $f$ .  
 DEF Convergenza. Dati  $\{f_m\}_{m \geq 0}$  e  $\{g_m\} \in F$ , se  $\lim_{m \rightarrow \infty} \|f - f_m\|_\infty = 0$  si dice che si ha convergenza (uniforme) di  $f_m$  a  $f$ .

### BASE MONOMIALE

Sono assegnati  $s$  dati  $(x_s, y_s)$ ,  $s=1, \dots, m$ ;  $x_s$  sono detti nodi; si vuole costruire il polinomio  $p_m(x) \in P_m$  che interpola  $s$  dati, tale da  $p_m(x_s) = y_s$ ; si scrive  $p_m(x)$  nella forma  $p_m(x) = \sum_{k=0}^m c_k x^k$ ; questo corrisponde a considerare la base monomiale  $P_m = \text{span}\{1, x, x^2, \dots, x^m\} = \mathcal{L}(1, x, x^2, \dots, x^m) = \langle 1, x, x^2, \dots, x^m \rangle$ .  
 Impugnando il criterio di interpolazione si ha  $p_m(x_s) = y_s \Rightarrow \sum_{k=0}^m c_k x_s^k = y_s$ , si hanno  $m+1$  eq lineari nelle sconosciute  $c_k$ , che si traduce nel sistema  $AC=y$  con  $A \in \mathbb{R}^{(m+1) \times (m+1)}$ ,  $c \in \mathbb{R}^{m+1}$ ,  $y \in \mathbb{R}^{m+1}$ .

PROP Ssa  $A$  la matrice di Vandermonde (comando MATLAB `vander`), se  $s$  nodi  $x_s$  sono distinti, si può dimostrare che  $A$  è invertibile.  
 Se  $s$  nodi sono distinti, il polinomio interpolante esiste ed è unico, su linea di polinomi  $s$  coefficienti si possono determinare risolvendo il sistema lineare  $AC=y$   
 3! polinomio interpolante  $\Leftrightarrow$  invertibilità di  $A$

### BASE DI LAGRANGE

Il polinomio interpolante è nella forma  $p_m(x) = \sum_{k=0}^m c_k l_k(x)$  dove  $l_k(x_s) = \delta_{sk} = \begin{cases} 1 & \text{se } s=k \\ 0 & \text{se } s \neq k \end{cases}$ . Impugnando le condizioni d'interpolazione  $p_m(x_s) = y_s \Rightarrow \sum_{k=0}^m c_k l_k(x_s) = y_s \Rightarrow c_s l_s(x_s) = y_s \Rightarrow c_s = y_s$

Il polinomio  $l_k(x)$  deve essere 1 in tutti i nodi tranne  $x_k$ , quindi  $l_k(x) = \text{cost} \cdot \prod_{s \neq k} (x - x_s) = \text{cost} \cdot (x - x_0)(x - x_1) \dots (x - x_{k-1})(x - x_{k+1}) \dots (x - x_m)$  dove anche valore 1 nel nodo  $x_k$ , quindi  $l_k(x_k) = \text{cost} \cdot (x_k - x_0)(x_k - x_1) \dots (x_k - x_{k-1})(x_k - x_{k+1}) \dots (x_k - x_m)$ , per cui si ha

$$l_k(x) = \frac{\prod_{s \neq k} (x - x_s)}{\prod_{s \neq k} (x_k - x_s)} \quad \text{ES 3 nodi } (x_0, y_0); (x_1, y_1); (x_2, y_2)$$

$$l_0(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}, \quad l_1(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}, \quad l_2(x) = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} \Rightarrow p_2(x) = y_0 l_0(x) + y_1 l_1(x) + y_2 l_2(x)$$

### CONDIZIONAMENTO

Se sono  $y_s$   $s$  dati dell'interpolazione, e  $\tilde{y}_s = y_s + \delta_s$  dei dati perturbati; ssa  $p_m(x)$  il polinomio che interpola  $s$  dati  $y_s$ , e  $\tilde{p}_m(x)$  il polinomio che interpola  $s$  dati  $\tilde{y}_s$ . Si definisce costante di Lebesgue  $\Lambda_m = \max_{k=0, \dots, m} \|l_k(x)\|_\infty$ , dipende solo dai nodi.  
 $p_m(x) - \tilde{p}_m(x) = \sum_{k=0}^m (y_k - \tilde{y}_k) l_k(x) \Rightarrow \|p_m(x) - \tilde{p}_m(x)\|_\infty \leq \sum_{k=0}^m |\delta_k| \|l_k(x)\|_\infty \leq \max_{0 \leq k \leq m} |\delta_k| \sum_{k=0}^m \|l_k(x)\|_\infty \Rightarrow \|p_m - \tilde{p}_m\|_\infty \leq \|\delta\|_\infty \sum_{k=0}^m \|l_k(x)\|_\infty \Rightarrow \|p_m - \tilde{p}_m\|_\infty \leq \|\delta\|_\infty \Lambda_m$   
 Si ha  $\frac{\|p_m - \tilde{p}_m\|_\infty}{\|p_m\|_\infty} \leq \Lambda_m \frac{\|\delta\|_\infty}{\|y\|_\infty}$ , dove  $\frac{\|p_m - \tilde{p}_m\|_\infty}{\|p_m\|_\infty}$  è la perturbazione relativa sui polinomi,  $\frac{\|\delta\|_\infty}{\|y\|_\infty}$  è la perturbazione relativa sui dati, quindi  $\Lambda_m$  funge da modo condizionamento per il problema dell'interpolazione polinomiale

### CONVERGENZA

DEF Errore d'interpolazione:  $E_m(x) = f(x) - p(x)$ , dove  $f(x)$ : funzione da interpolare,  $p(x)$ : polinomio interpolante  
 TH Se  $f \in C[a, b]$  (funzione continua su  $a, b$ ) allora  $\|E_m\|_\infty \leq (1/\Lambda_m) \max_{q \in P_m} \|f - q\|_\infty$ , dove  $(1/\Lambda_m)$  dipende solo dai nodi, (11)



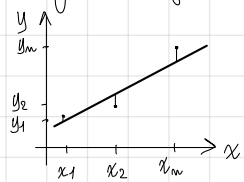
$\frac{1}{(1+f'(x)^2)^{3/2}}$ : curvatura di  $f$  nel pt  $x$ . Se  $|f'(x)|$  è sufficientemente piccolo  $C[f]$  è un' approssimazione della curvatura globale di  $f$  su  $[a, b]$ .

Comandi MATLAB: a)  $a = \text{polyfit}(x, y, m)$ ;  $yy = \text{polyval}(a, xx)$ ;  $x$ : vettore dei nodi,  $m$ : grado polinomio interpolante  
 b)  $yy = \text{spline}(x, y, xx)$ ;  $y$ : vettore delle ordinate,  $xx$ : vettore dei punti in cui calcolare la spline

Limiti dell'interpolazione sono: interpolazione polinomiale globale non adatta per  $m$  grandi; interpolazione in generale non adatta per dati sperimentali (molti dati e affetti da errore).

### 3.3 METODO DEI MINIMI QUADRATI

Si cerca a priori un modello per  $s$  dati in uno spazio di dimensione  $m$  (tipicamente  $m \ll s$ ): sono  $m$  funzioni linearmente indipendenti  $f_m(x) = c_1 \phi_1(x) + c_2 \phi_2(x) + \dots + c_m \phi_m(x)$



Si cerca la retta formata dai punti  $f_m(x)$  tale da far la distanza tra  $y$  e  $f_m$  sia la minima possibile (sommarievolmente distanza al più minore possibile)

DEF  $f_m(x)$  è la miglior approssimazione nel senso dei minimi quadrati

$$\min_{c \in \mathbb{R}^m} \sum_{s=1}^s (f_m(x_s) - y_s)^2 = \min_{c \in \mathbb{R}^m} \sum_{s=1}^s \left( \sum_{k=1}^m c_k \phi_k(x_s) - y_s \right)^2 = \min_{c \in \mathbb{R}^m} \|Ac - y\|_2^2$$

$\|Ac - y\|_2^2$  è il funzionale quadratico convesso, quindi il minimo del sistema è assunto in corrispondenza degli zeri del gradiente  $\min_{c \in \mathbb{R}^m} \|Ac - y\|_2^2 = (Ac - y)^T (Ac - y) = \frac{1}{2} c^T A^T A c + \frac{1}{2} y^T y - y^T A c \Rightarrow \nabla \|Ac - y\|_2^2 = A^T A c - A^T y = 0 \Rightarrow A^T A c = A^T y$  è detto sistema delle equazioni normali.

PROP Ssa  $A^T A$  simmetrica semidefinita positiva; se le colonne di  $A$  sono linearmente indipendenti,  $A^T A$  è simmetrica definita positiva. Se  $\phi_k(x)$  sono linearmente indipendenti e  $s$  nodi sono distinti, sicuramente le colonne di  $A$  sono linearmente indipendenti.

Gli vantaggi di questo approccio sono: è costoso costruire  $A^T A$ , tipicamente è molto mal condizionato. L'alternativa computazionale è l'uso della fattorizzazione QR (Q: ortogonale, R: triangolare).

Se si ignorassero le condizioni di interpolazione  $\sum_{k=1}^m c_k \phi_k(x_s) = y_s \forall s$ , si otterrebbe il sistema sovradimensionato  $Ac = y$  da non avrebbe soluzione in senso classico. Questo approccio corrisponde in generale ad un metodo per trovare una soluzione del sistema sovradeterminato nel senso dei minimi quadrati.

Il comando backslash di MATLAB risolve un sistema lineare sovradeterminato nel senso dei minimi quadrati; se  $A$  è a rango pieno, il comando  $c = A \backslash y$  fornisce la soluzione del problema.

Comandi MATLAB: a)  $\text{polyfit}$  se  $\phi_k(x)$  polinomiali, vuole in ingresso  $x$ : vettore dei nodi,  $y$ : vettore delle ordinate,  $m$  grado del polinomio  
 b)  $\text{fit}$

### 3.4 DERIVAZIONE NUMERICA

#### APPROSSIMAZIONE DERIVATA PRIMA

Approssimazione con polinomio di Taylor arrestato a I ordine con resto di Lagrange:

$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(\xi)(x - x_0)^2 \rightarrow$  resto di Lagrange, con  $x_0 \leq \xi \leq x_0$

Ssa  $x = x_0 + h \Rightarrow f(x_0 + h) = f(x_0) + f'(x_0)h + \frac{1}{2} f''(\xi_+) h^2 \Rightarrow f'(x_0) = \frac{f(x_0 + h) - f(x_0)}{h} - \frac{1}{2} f''(\xi_+) h \rightarrow$  errore =  $O(h)$  con  $\xi_+ \in [x_0, x_0 + h]$

Ssa  $x = x_0 - h \Rightarrow f(x_0 - h) = f(x_0) - f'(x_0)h + \frac{1}{2} f''(\xi_-) h^2 \Rightarrow f'(x_0) = \frac{f(x_0) - f(x_0 - h)}{h} + \frac{1}{2} f''(\xi_-) h \rightarrow$  errore =  $O(h)$  con  $\xi_- \in [x_0 - h, x_0]$

approssimazione con differenze in avanti / approssimazione con differenze all'indietro

Approssimazione con polinomio di Taylor arrestato al II ordine con resto di Lagrange:

$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2 + \frac{1}{6} f'''(\xi)(x - x_0)^3$

con  $x = x_0 + h \Rightarrow f(x_0 + h) = f(x_0) + f'(x_0)h + \frac{1}{2} f''(x_0)h^2 + \frac{1}{6} f'''(\xi_+) h^3$

con  $x = x_0 - h \Rightarrow f(x_0 - h) = f(x_0) - f'(x_0)h + \frac{1}{2} f''(x_0)h^2 - \frac{1}{6} f'''(\xi_-) h^3$

$f(x_0) = \frac{f(x_0 + h) - f(x_0 - h)}{2h} + \frac{1}{6} (f'''(\xi_+) + f'''(\xi_-)) h^2 \rightarrow$  errore =  $O(h^2)$

approssimazione con differenze centrate

Se  $h$  viene scelto troppo grande si hanno stime poco accurate, se troppo piccole si ha cancellazione numerica, si best trade-off è  $h \approx \sqrt{\epsilon_m / |f''|}$

# 4

# QUADRATURA NUMERICA

Se l'integrale definito  $\int_a^b f(x) dx$ , si ha difficoltà nel calcolarlo se:  $f(x)$  è nota solo per punti; per  $f(x)$  non è individuabile una primitiva elementare, ad es  $f(x) = \exp(-x^2)$ , da serve molto in calcolo della probabilità quando si gestiscono variabili aleatorie normali.

Le formule di quadratura numerica consentono di approssimare l'integrale definito usando  $n$  valori di  $f$  in punti assegnati  $\int_a^b f(x) dx \approx \sum_{s=1}^n w_s f(x_s)$  dove  $x_s$ : nodi,  $w_s$ : pesi.

Le formule d'interpolazione sono correlate all'interpolazione polinomiale, la quadratura composta all'interpolazione polinomiale a tratti.

## 4.1 FORMULE INTERPOLATORIE

Scelti  $x_s, s=1, \dots, n$  nodi distinti  $\in [a, b]$ , sia  $p_{n-1}(x) = \sum_{s=1}^n f(x_s) l_s(x)$  dove  $l_s(x)$ :  $s$ -esimo polinomio di Lagrange associato al nodo  $x_s$ ; il polinomio interpolante  $f$  nei nodi  $x_s$  è  $\int_a^b f(x) dx \approx \int_a^b p_{n-1}(x) dx = \sum_{s=1}^n f(x_s) \int_a^b l_s(x) dx$  con  $w_s = \int_a^b l_s(x) dx$ . Si parla di formule chiuse se  $a$  e  $b$  sono tra i nodi, aperte altrimenti.

### FORMULE DI NEWTON-COTES

Se  $n$  nodi sono equidistanti le formule interpolatorie sono dette di Newton-Cotes (quadratura interpolante su nodi equidistanti)

- $n=1$ , un solo nodo
  - rettangolo: il nodo è uno dei due estremi indifferente  $f(x) \approx p_0(x) = f(a) \Rightarrow \int_a^b f(x) dx \approx (b-a) f(a)$
  - punto medio: nodo dato dalla media degli estremi  $f(x) \approx p_0(x) = f(\frac{a+b}{2}) \Rightarrow \int_a^b f(x) dx \approx (b-a) f(\frac{a+b}{2})$
- $n=2$ , 2 nodi sono gli estremi
  - trapezio  $f(x) \approx p_1(x) = f(a) \frac{x-b}{a-b} + f(b) \frac{x-a}{b-a} \Rightarrow \int_a^b f(x) dx \approx \frac{b-a}{2} f(a) + \frac{b-a}{2} f(b) = \frac{b-a}{2} (f(a) + f(b))$
- $n=3$ , 3 nodi devono essere equidistanti, quindi si scelgono gli estremi e il punto medio  $c = \frac{a+b}{2}$ 
  - di Cavalieri-Simpson  $f(x) \approx p_2(x) = f(a) \frac{(x-c)(x-b)}{(a-c)(a-b)} + f(c) \frac{(x-a)(x-b)}{(c-a)(c-b)} + f(b) \frac{(x-a)(x-c)}{(b-a)(b-c)} \Rightarrow \int_a^b f(x) dx \approx \frac{b-a}{6} f(a) + 4 \frac{b-a}{6} f(c) + \frac{b-a}{6} f(b) = \frac{b-a}{6} (f(a) + 4f(c) + f(b))$

Non si ritiene di dover far crescere il grado del polinomio rispetto a 3 perché si avrebbe un'approssimazione errata, in quel caso si preferisce utilizzare una funzione polinomiale a tratti.

### ERRORE DI QUADRATURA

DEF Errore di quadratura. Dato una formula di quadratura l'errore è  $R_n = \int_a^b f(x) dx - \sum_{s=1}^n w_s f(x_s)$ , considerando la formula interpolatoria  $R_n = \int_a^b (f(x) - p_{n-1}(x)) dx = \int_a^b E_n(x) dx$ .

Se  $R_n = 0$  la formula di quadratura è dice esatta.

PROP Una formula di quadratura interpolatoria costruita su  $n$  nodi è esattamente esatta per tutti i polinomi di grado  $\leq n-1$  ( $E_n(x) \equiv 0 \forall x \in [a, b]$ )

Questa proprietà suggerisce un altro approccio per determinare i pesi di una formula interpolatoria, fissati  $n$  nodi. Si impone che la formula di quadratura  $\int_a^b f(x) dx \approx \sum_{s=1}^n w_s f(x_s)$  sia esatta su una base delle espressioni dei polinomi di grado  $n-1$ ,  $1, x, x^2, \dots, x^{n-1}$

$$\begin{cases}
 f(x) \equiv 1 & \left\{ \begin{array}{l} w_1 + w_2 + \dots + w_n = \int_a^b 1 dx = b-a \\
 f(x) \equiv x & \left\{ \begin{array}{l} w_1 x_1 + w_2 x_2 + \dots + w_n x_n = \int_a^b x dx = \frac{1}{2}(b^2 - a^2) \\
 \vdots \\
 f(x) \equiv x^{n-1} & \left\{ \begin{array}{l} w_1 x_1^{n-1} + w_2 x_2^{n-1} + \dots + w_n x_n^{n-1} = \int_a^b x^{n-1} dx = \frac{1}{n}(b^n - a^n)
 \end{array} \right.
 \end{array} \right.
 \end{array}
 \right.
 \end{cases}$$

Sistema lineare di  $n$  equazioni nelle  $n$  sconosciute  $w_s$  con  $s=1, \dots, n$

PROP Le formule di Newton-Cotes costruite su  $n$  nodi sono esatte per polinomi di grado fino a  $d = n-1$  se  $n$  è pari, e  $d = n$  se  $n$  è dispari.



# 5

# EQUAZIONI NON LINEARI

Sia  $f: \mathbb{R} \rightarrow \mathbb{R}$ , si vogliono trovare le soluzioni  $f(x) = 0$  dette zeri e radici. In generale non esiste una formula esatta, cioè non esiste un metodo diretto ma solo iterativi. Ci sono due aspetti da analizzare: convergenza e velocità di convergenza.

Il caso non lineare presenta diverse ostacoli rispetto al caso lineare: la convergenza dipende in modo critico dalla scelta di  $x_0$ , il metodo possono convergere a punti che non sono le soluzioni cercate. Qualunque metodo si intenda applicare, sarà sempre necessario effettuare uno studio preliminare della funzione in modo da localizzare le eventuali radici.

## 5.1

## METODO DI BISEZIONE

TH Esistenza degli zeri per funzioni continue - Se  $f \in C[a, b]$  e  $f(a)f(b) < 0 \Rightarrow \exists c \in [a, b]$  t.c.  $f(c) = 0$

Su questo th si basa il metodo di bisezione: sia  $[a, b]$  l'intervallo in cui è isolata la radice  $x^*$  e  $m = \frac{a+b}{2}$ , si hanno due possibilità: a) se  $f(m)$  è discorde con  $f(a)$ ,  $x^* \in [a, m]$

b) se  $f(m)$  è discorde con  $f(b)$ ,  $x^* \in [m, b]$

$x^*$  continua ad essere l'unica radice nel nuovo intervallo di ampiezza dimezzata; si procede iterativamente dimezzando ad ogni passo l'ampiezza dell'intervallo.

L'errore è  $e_k = |x_k - x^*| \leq \frac{b_k - a_k}{2} = \frac{b_0 - a_0}{2^{k+1}}$ , per il th del doppio confronto  $0 \leq e_k \leq \frac{b_0 - a_0}{2^{k+1}}$ , quindi  $\lim_{k \rightarrow \infty} e_k = 0$  indipendentemente dall'intervallo iniziale.

DEF Il metodo di bisezione è globalmente convergente, poiché la convergenza è sempre garantita, qualunque sia la stima iniziale della soluzione.

Si suppone di voler garantire un errore assoluto inferiore ad una certa tolleranza  $\epsilon$ . È possibile stabilire il n° massimo di iterazioni necessarie a soddisfare tale richiesta. Se si garantisce  $\frac{b_0 - a_0}{2^{k+1}} < \epsilon$  si garantisce  $e_k \leq \epsilon$ .

DEF Sia  $\{x_k\}$  una successione convergente a un limite  $x^*$ . Sia  $e_k = |x_k - x^*|$ , se esistono  $p > 1$  e una costante  $c > 0$  ( $c < 1$  se  $p = 1$ ) tale che  $\lim_{k \rightarrow \infty} \frac{e_{k+1}}{e_k^p} = c$  ( $\Rightarrow e_{k+1} = O(e_k^p)$ ) allora si dice che la successione converge con ordine  $p$ :

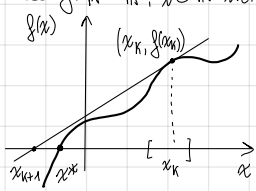
$p = 1$  convergenza lineare,  $p = 2$  convergenza quadratica,  $1 < p < 2$  convergenza superlineare.

Se c'è convergenza lineare,  $\lim_{k \rightarrow \infty} \frac{e_{k+1}}{e_k} < 1$ , quindi - per il th della permanenza del segno - asintoticamente l'errore deve diminuire monotonicamente ( $e_{k+1} < e_k$ ). Il metodo di bisezione non garantisce la decrescita monotona dell'errore, quindi la velocità di convergenza è sublineare.

## 5.2

## METODO DI NEWTON O DELLE TANGENTI

Sia  $f: \mathbb{R} \rightarrow \mathbb{R}$ ,  $x \in \mathbb{R}$  t.c.  $f(x) = 0$ . Il metodo di Newton sfrutta informazioni sia sul valore di  $f(x)$  che di  $f'(x)$ , sia  $x_k$  la stima corrente della soluzione di  $f(x) = 0$ , si approssima localmente  $f$  con il polinomio di Taylor di ordine 1, cioè con la retta tangente passante per il punto  $(x_k, f(x_k))$ :



$$f(x) \approx e_k(x) = f(x_k) + f'(x_k)(x - x_k); \text{ si cercano i punti in cui } e_k(x) = 0 : f(x_k) + f'(x_k)(x - x_k) = 0$$

$$\Rightarrow x = x_k - \frac{f(x_k)}{f'(x_k)} \text{ con } k=0, 1, \dots$$

TH Sia  $f \in C^2$  in un intorno della soluzione  $x^*$ , se la successione  $\{x_k\}$  generata dal metodo di Newton converge a  $x^*$  e  $f'(x^*) \neq 0$ , allora l'ordine di convergenza è (almeno)  $p = 2$ .

TH Sia  $f \in C^2[a, b]$  disuso e limitato, se: i)  $f(a)f(b) < 0$

ii)  $f'(x) > 0 \forall x \in [a, b]$  o  $f'(x) < 0 \forall x \in [a, b]$  (non si altera segno in  $[a, b]$ )

iii)  $f''(x) \geq 0 \forall x \in [a, b]$  o  $f''(x) \leq 0 \forall x \in [a, b]$  (non cambia segno in  $[a, b]$ )

iv)  $\left| \frac{f(a)}{f'(a)} \right| < b - a$  e  $\left| \frac{f(b)}{f'(b)} \right| < b - a$

allora  $f(x) = 0$  ha un'unica radice  $x^*$  in  $[a, b]$  e il metodo di Newton converge a  $x^* \forall x_0 \in [a, b]$ .

TH Sia  $f \in C^2[a, b]$  disuso e limitato, se: i)  $f(a)f(b) < 0$

ii)  $f'(x) > 0 \forall x \in [a, b]$  o  $f'(x) < 0 \forall x \in [a, b]$  (non si altera segno in  $[a, b]$ )

• Sia  $\{x_k\}$  una successione convergente a un limite  $x^*$ . Sia  $e_k = |x_k - x^*|$ , se esistono  $p > 1$  e una costante  $c > 0$  ( $c < 1$  se  $p=1$ ) tale che  $\lim_{k \rightarrow \infty} \frac{e_{k+1}}{e_k^p} = c$  ( $\Rightarrow e_{k+1} = O(e_k^p)$ ) allora si dice che la successione converge con ordine  $p$ .  
Non è garantita la decrescita monotona dell'errore, quindi la velocità di convergenza è sublineare.

### METODO DI NEWTON O DELLE TANGENTI

• Sia  $x_k$  la stima corrente della soluzione di  $f(x) = 0$ , si approssima localmente  $f$  con il polinomio di Taylor di ordine 1, cioè con la retta tangente passante per il punto  $(x_k, f(x_k))$ :  $f(x) \approx r_k(x) = f(x_k) + f'(x_k)(x - x_k)$ ; si cercano i punti in cui  $r_k(x) = 0$ :  $f(x_k) + f'(x_k)(x - x_k) = 0 \Rightarrow x = x_k - \frac{f(x_k)}{f'(x_k)}$  con  $k=0, 1, \dots$   
• Sia  $f \in C^2$  in un intorno della soluzione  $x^*$ , se la successione  $\{x_k\}$  generata dal metodo di Newton converge a  $x^*$  e  $f'(x^*) \neq 0$ , allora l'ordine di convergenza è (almeno)  $p=2$ .

Sia  $f \in C^2[a, b]$  discesa e limitato, se: i)  $f(a)f(b) < 0$ ; ii)  $f'(x) > 0 \forall x \in [a, b]$  o  $f'(x) < 0 \forall x \in [a, b]$  (non si altera mai in  $[a, b]$ ); iii)  $f''(x) \geq 0 \forall x \in [a, b]$  o  $f''(x) \leq 0 \forall x \in [a, b]$  (non cambia segno in  $[a, b]$ ); iv)  $|\frac{f(a)}{f'(a)}| < b-a$  e  $|\frac{f(b)}{f'(b)}| < b-a$ ; allora  $f(x) = 0$  ha

una unica radice  $x^*$  in  $[a, b]$  e il metodo di Newton converge a  $x^*$   $\forall x_0 \in [a, b]$ .

• Si cerca  $x_{k+1}$  tale che  $f(x_k) + f'(x_k)(x_{k+1} - x_k) = 0 \Rightarrow f'(x_k)x_{k+1} = -f(x_k) + f'(x_k)x_k \Rightarrow x_{k+1} = (f'(x_k))^{-1}(-f(x_k) + f'(x_k)x_k)$ . Per ottenere il prodotto matrice-vettore  $f'(x_k)x_k$ :  $f'(x_k) = -f(x_k)$  e  $x_{k+1} = x_k + S$   
Comandi MATLAB: a) `fzero` b) `fzero`

### PROBLEMA DI PUNTO FISSO

• Secondo il metodo delle stimate successive o delle stimate successive di punto fisso, dato  $x_0 \in \mathbb{R}$ ,  $x_{k+1} = \phi(x_k)$   
Se  $|\phi'(x^*)| < 1 \Rightarrow$  esiste un intorno  $I$  di  $x^*$  tale che la successione  $\{x_k\}$  converge se  $x_0 \in I$ . Se  $|\phi'(x^*)| > 1$  il metodo delle stimate successive non può convergere.  
Si suppone di trovare  $[a, b]$  tale che: i)  $\phi \in C^1[a, b]$ ; ii)  $\phi: [a, b] \rightarrow [a, b]$ ; iii)  $\exists k < 1$  tale che  $|\phi'(x)| \leq k \forall x \in [a, b]$ ; allora esiste uno e un solo punto fisso in  $[a, b]$  e il metodo delle stimate successive converge ad esso  $\forall x_0 \in [a, b]$ .  
• Ordine di convergenza  $\Leftrightarrow$  il primo ordine di derivata di  $\phi$  che non si annulla in  $x^*$ .