Dispense Analisi Numerica

Federico De Sisti

2026-01-26

1 Norme Matriciali

Definizione 1 (Norma Matriciale)
Una norma matriciale è un applicazione $\|\cdot\| : \mathbb{C}^{m\times n} \to \mathbb{R}$ *tale che*

1. $\|A\| \geq 0 \quad \forall A \in \mathbb{C}^{m\times n}$ *e* $\|A\| = 0$ *se e solo se* $A = 0$
2. $\|\alpha A\| = |\alpha|\|A\| \forall \alpha \in C, \forall A \in \mathbb{C}^{m\times n}$ *(omogeneità)*
3. $\|A+B\| \leq \|A\|+\|B\| \quad \forall A, B \in \mathbb{C}^{m\times n}$ *(disuguaglianza triangolare)*

Definizione 2 (Norma compatibile)
Diciamo che una norma matriciale $\|\cdot\|$ *è compatibile o consistente con una norma vettoriale* $\|\cdot\|$ *se*

$$\|Ax\| \leq \|A\|\|x\|. \quad \forall x \in \mathbb{C}^n.$$

Definizione 3 (Matrice coniugata trasposta (aggiunta))
Sia $A \in \mathbb{C}^{m\times n}$*; la matrice* $B = A^H \in \mathbb{C}^{m\times n}$ *è detta matrice coniugata trasposta (o aggiunta) di* A *se* $b_{ij} = \overline{a_{ji}}$ *essendo* $\overline{a_{ji}}$ *il numero complesso coniugato di* a_{ji}

Definizione 4 (Norma di Frobenius)

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

Teorema 1
Sia $\|\cdot\|$ *una norma vettoriale su* $\mathbb{C}^n$*. La funzione*

$$\|A\| = \sup_{x\neq 0}\frac{\|Ax\|}{\|x\|}.$$

è una norma matriciale su $\mathbb{C}^{m\times n}$, che viene detta norma matriciale indotta (o subordinata) o norma matriciale naturale.

La dimostrazione di questo teorema non è richiesta

Teorema 2
Se $|||\cdot|||$ è una norma matriciale naturale indotta da $\|\cdot\|$, allora

1. $\|Ax\| \leq |||A|||\|x\|$, *ossia è una norma compatibile*
2. $|||I||| = 1$
3. $|||AB||| \leq |||A|||||B|||$, *ossia è sub moltiplicativa*

Dimostrazione
guarda libro □

Definizione 5 (Raggio spettrale)
Chiamiamo $\rho(A)$ il raggio spettrale della matrice $A \in \mathbb{C}^{n\times n}$

$$\rho(A) = \max_{i=1,\dots,n} |\lambda_i|.$$

Teorema 3
Sia $\|\cdot\|$ una norma matriciale compatibile con una norma vettoriale che indichiamo con lo stesso simbolo, allora

$$\rho(A) \leq \|A\| \quad \forall A \in \mathbb{C}^{n\times n}.$$

La norma spettrale è la norma indotta dalla norma euclidea.

Teorema 4 (di Ostrowski)
Sia $A \in \mathbb{C}^{n\times n}$ e $\varepsilon > 0$. Allora esiste una norma matriciale naturale $\|\cdot\|_{\rho,\varepsilon}$ (dipendente da ε tale che

$$\|A\|_{\rho,\varepsilon} \leq \rho(A) + \varepsilon.$$

Di conseguenza fissata una tolleranza piccola a piacere, esiste sempre una norma matriciale che è arbitrariamente vicina la raggio spettrale di A,

ossia

$$\rho(A) = \inf_{\|\cdot\|} \|A\|.$$

Teorema 5 (Successione di matrici e raggio spettrale)
Sia A una matrice quadrata e sia $\|\cdot\|$ una norma naturale allora

$$\lim_{m\to+\infty} \|A^m\|^{1/m} = \rho(A)$$

Definizione 6 (Convergenza di una successione di matrici)
Una successione di matrici $\{A^{(k)} \in \mathbb{R}^{n\times m}$ è detta convergente ad una matrice $A \in \mathbb{R}^{n\times n}$ se

$$\lim_{k\to+\infty} \|A^{(k)} - A\| = 0.$$

La scelta della norma è ininfluente in quanto in $\mathbb{R}^{n\times n}$ le norme sono equivalenti.

Definizione 7 (Matrice Convergente)
Una matrice A si dice convergente se

$$\lim_{k\to+\infty} A^k = 0.$$

Teorema 6
Sia A una matrice quadrata. Allora:

$$\lim_{k\to+\infty} A^k = 0 \Leftrightarrow \rho(A) < 1.$$

Inoltre se $\rho(A) < 1$ allora la matrice $I - A$ è invertibile.
La serie geometrica $\sum_{k=0}^{+\infty} A^k$ è convergente se e solo se $\rho(A) < 1$ e, in tal caso

$$\sum_{k=0}^{+\infty} A^k = (I - A)^{-1}.$$

Infine, sia $\|\cdot\|$ una norma matriciale naturale tale che $\|A\| < 1$. Allora,

se $I - A$ è invertibile, valgono le seguenti disuguaglianze

$$\frac{1}{1+\|A\|} \leq \|(I-A)^{-1}\| \leq \frac{1}{1-\|A\|}.$$

Dimostrazione
guarda il libro ed è da fare, non il 1.24, mi sa manco l'enunciato. □

2 Sistemi Lineari

Definizione 8 (Numero di condizionamento)
Si definisce numero di condizionamento di una matrice $A \in \mathbb{C}^{n\times n}$ la quantità

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

Essendo $\|\cdot\|$ una norma matriciale indotta. In generale $K(A)$ dipende dalla norma scelta.

Definizione 9
Una martice si dice singolare quando non è invertibile, ovvero quando il suo determinante è nullo

Osservazione
In generale, se definiamo la distanza relatia di $A \in \mathbb{C}^{n\times n}$ dall'insieme delle matrici singolari rispetto alla norma p come

$$dist_p(A) = \min\{\frac{\|\delta A\|_p}{\|A\|_p}; A + \delta A \text{ è singolare }\}.$$

e si può dimostrare che

$$dist_p(A) = \frac{1}{K_p(A)}.$$

Ciò significa che una matrice con un numero di condizionamento elevato potrebbe comportarsi come una matrice singolare della forma $A + \delta A$. In altre parole, in tal caso, a perturbazioni nulle del termine noto potrebbero non corrispondere perturbazioni nulle sulla soluzione.

Teorema 7
Siano $A \in \mathbb{R}^{n\times n}$ una matrice non singolare e $\delta A \in \mathbb{R}^{n\times n}$ tali che sia soddisfatta:

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

per una generica norma matriciale indotta $\|\cdot\|$.
Allora se $x \in \mathbb{R}^n$ è soluzione di $Ax = b$ con $b \in \mathbb{R}^n$ $(b \neq 0)$ e $\delta x \in \mathbb{R}^n$ verifica

$$(A + \delta A)(x + \delta x) = b + \delta b.$$

si ha che

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{K(A)}{1 - K(A)\|\delta A\|/\|A\|}\left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|}\right).$$

Dimostrazione
guarda il libro □

Corollario 1
Si suppongano valide le ipotesi del teorema 7, sia $\delta A = 0$ Allora

$$\frac{1}{K(A)}\frac{\|\delta b\|}{\|b\|} \leq \frac{\|\delta x\|}{\|x\|} \leq K(A)\frac{\|\delta b\|}{\|b\|}.$$

Per poter impiegare queste due disuguaglianze nell'analisi della propagazione degli errori di arrotondamento per i metodi diretti, $\|\delta A\|$ e $\delta b\|$ dovranno essere stimati in funzione della dimensione del sistema e delle caratteristiche dell'aritmetica floating-point usata.
È infatti ragionevole aspettarsi che le perturbazioni indotte da un metodo per la risoluzione di un sistema lineare siano tali che $\|\delta A\| \leq \gamma\|A\|$ e $\|\delta b\| \leq \gamma\|b\|$, essendo γ un numero positivo che dipende da u, l'unità di roundoff (ad esempio si porrà in seguito $\gamma = \beta^{1-t}$, essendo β la base e t il numero di cifre della mantissa del sistema $\mathbb{F}$ scelto).

2.1 Metodo di eliminazione gaussiana (MEG) e fattorizzazione LU

Il metodo di eliminazione gaussiana si basa sul ridurre il sistema $Ax = b$ ad un nuovo sistema equivalente $Ux = \hat{b}$ dove U è triangolare superiore e $\hat{b}$ è un

nuovo termine noto.
Il MEG equivale a fattorizzare la matrice di partenza nel prodotto di due matrici $A = LU$ con $U = A^{(n)}$.
Fattorizzare la matrice A in questo modo è utile perché non dipende dal termine noto, e quindi posso risolvere più sistemi lineari con matrice dei coefficienti uguale e termine noto diverso.
Sostianzialmente possiamo prendere la matrice di trasformazione gaussiana del k-esimo passo M_k e ricavare che

$$A^{k+1} = M_k(A^{(k)}.$$

possiamo quindi scrivere

$$M_{n-1}M_{n-2}\dots M_1 A = A^{(n)} = U.$$

Le matrici M_k sono matrici triangolari inferiori con elementi diangolai pari ad uno e con inversa data da

$$M_k^{-1} = 2I_n - M_k = I_n + m_k e_k^T.$$

essendo $(m_i e_i^T)(m_j e_j^T)$ uguale alla matrice nulla se $i \leq j$ di conseguenza

$$\begin{aligned} A &= M_1^{-1}\dots M_{n-1}^{-1}U \\ &= (I_n + m_1 e_1^T)\dots(I_n + m_{n-1}e_{n-1}^T U \\ &= \left(I_n + \sum_{i=1}^{n-1} m_i e_i^T\right) U \\ &= \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ m_{21} & 1 & & & \vdots \\ \vdots & m_{32} & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ m_{n_1} & m_{n_2} & \dots & m_{n,n-1} & 1 \end{pmatrix} U \end{aligned}$$

Definiamo allora $L = M_{n-1}\dots M_1)^{-1} = M_1^{-1}\dots M_{n-1}^{-1}$ sia

$$A = LU.$$

La risoluzione del sistema lineare diventa quindi la soluzione, in sequenza, di questi due sistemi lineari

$$Ly = b \tag{1}$$
$$Ux = y \tag{2}$$

Il programma numero 3 a pagina 78 è lukji della fattorizzazione LU.
Studio fowardrow e backwardrow.

Definizione 10 (Pivoting Parziale)
Il pivoting parziale è l'applicazione del metodo MEG su una matrice con la scelta dell'elemento di modulo massimo come PIVOT per ridurre l'errore

Esistenza e unicità della fattorizzazione LU
Sia $A \in R^{n\times n}$. La fattorizzazione LU di A con $l_i i = 1$ per $i = 1, \dots, n$ esiste ed è unica se e solo se le sottomatrici principali A_i di A di ordine $i = 1, \dots, n-1$ sono non singolari.

Definizione 11
*Una matrice quadrata $A = (a_{ij}) \in \mathbb{C}^{n\times n}$ è **a diagonale dominante per righe** se:*

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

*e **strettamente dominante diagonale** se:*

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

Matrici a dominanza diagonale e fattorizzazione LU
Se A è una matrice a dominanza diagonale stretta per righe o per colonne, allora esiste ed è unica la fattorizzazione LU di A con elementi diagonali di L tutti pari ad 1. Lo stesso risultato vale se A è a dominanza diagonale nel caso sia non singolare.

Teorema 8 (Fattorizzazione di Cholesky)
Sia $A \in \mathbb{R}^{n\times n}$ una matrice simmetrica e definita positiva; allora esiste un'unica matrice triangolare superiore H con elementi diagonali positivi

tale che

$$A = H^T H.$$

Questa è la fattorizzazioen di Cholesky, Gli elementi h_{ij} di H sono dati dalle fomrule seguenti $h_{11} = \sqrt{a_{11}}$, per $j = 2, \ldots, n$

$$h_{ij} = \left(a_{ij} - \sum_{k=1}^{i-1} j_{ki} h_{kj} \right) / h_i i, i = 1, \ldots, j-1.$$

$$h_j j = \left(a_j j - \sum_{k=1}^{j-1} h_{kj}^2 \right)^{1/2}.$$

Il teorema si generalizza alle matrici ermitiane definite positive

Teorema 9 (Fattorizzazione di Cholesky complessa)
Una matrice quadrata $A \in C^{n \times n}$ è hermitiana definita positiva se:

- $A = A^H$ *(proprietà hermitiana)*
- $x^H A x > 0$ *per ogni* $x \neq 0$

In tal caso, esiste una matrice triangolare inferiore $L \in C^{n \times n}$ con elementi diagonali reali e positivi tali che

$$A = L L^H.$$

dove L^H è la trasposta coniugata di L
Per $k = 1, \ldots, n$:

$$l_{kk} = \sqrt{a_{kk} - \sum_{j=1}^{k-1} |l_{kj}|^2} \tag{3}$$

$$l_{ik} = \frac{1}{l_{kk}} \left(a_{ik} - \sum_{j=1}^{k-1} l_{ij} \overline{l_{kj}} \right), \quad i = k+1, \ldots, n \tag{4}$$

Qui c'è il codice della fattorizzazione di chol2, il programma 7 a pagina 84

2.2 Matrici tridiagonali e algoritmo di Thomas

Definizione 12
Una matrice tridiagonale è una matrice nulla se non per la diagonale, i coefficienti streattamente al di sopra e strettamente al di sotto

In caso di matrici tridiagonali le matrici LU sono del tipo:
L ha sulla diagonale 1 e sulla diagonale subito sotto i coefficienti β_k con $k = 2, \dots, n$
U ha sulla diagonale i coefficienti α_k e sulla diagonale subito sopra i coefficienti c_k con $k = 1, \dots, n-1, n$
i coefficienti c_k sono gli stessi della matrice di partenza e per calcolare gli α e i β si usa la formula

$$\alpha_1 = a_1, \beta_i = \frac{b_i}{\alpha_{i-1}}, \alpha_i = a_i - \beta_i c_{i-1}, i = 2, \dots, n.$$

Questo algoritmo prende il nome di algoritmo di Thomas.
Questo algoritmo può essere utilizzato per risolvere il sistema tridiagonale traimte

$$(Ly = f)\ y_1 = f_1, y_i = f_i - \beta_i y_{i-1}, i = 2, \dots, n.$$
$$(Ux = y)\ x_n = \frac{y_n}{\alpha_n},\ x_i = (y_i - c_i x_{i+1})/\alpha_i, i = n-1, \dots, 1.$$

2.3 Metodi iterativi

I metodi iterativi sono tutti quelli che ci portano alla soluzione tramite un numero di iterazioni determinato dalla precisione scelta.
Splitting additivo
Una strategia generale per costruire metodi iterativi lineari consiste in una decomposizione additiva della matrice A, della forma $A = P - N$ dove P e N sono due matrici opportune e P è non singolare, P è detta matrice di precondizionamento o precondizionaore.
Precisamnte, assegnare x^0, si ottiene x^{k+1} per $k \geq 0$ risolvendo i nuovi sistemi

$$Px^{(k+1)} = Nx^{(k)} + b \quad k \geq 0.$$

Più in generale per i metodi iterativi avremo $x^{(0)}$ dato e

$$x^{(k+1)} = Bx^{(k)} + f \quad k \geq 0.$$

avendo indicato con B una matrice quadrata $n \times n$ detta matrice di iterazione e con f un vettore che si ottiene a partire dal termine noto b.

Definizione 13 (Condizione di consistenza)
Un metodo iterativo della forma appena descritta si dirà consistente con il sistema $Ax = b$ se e solo se f e B sono tali che $x = Bx + f$. Equivalentemente

$$f = (I - B)A^{-1}b.$$

Indicato con

$$e^{(k)} = x^{(k)} - x.$$

L'errore al passo k, la condizione di convergenza è quivalente a richiedere che $\lim_{k\to+\infty} e^{(k)} = 0$ per ogni scelta del vettore iniziale $x^{(0)}$

Teorema 10 (Convergenza e Raggio spettrale)
Se il metodo è consistente, esso converge alla soluzione del sistema per ogni scelta del vettore iniziale $x^{(0)}$ se e solo se $\rho(B) < 1$

Dimostrazione
Questa te la fai sul libro a pagina 116 □

2.3.1 Metodo di Jacobi

Nel metodo di Jacobi, scelto un dato iniziale $x^{(0)}$, si calcola $x^{(k+1)}$ attraverso le formule

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left[b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij} x_j^{(k)} \right], \quad i = 1, \ldots, n.$$

Ciò equivale allo splitting $A = P - N$ con

$$P = D, \quad N = D - A = E + F.$$

Dove D è la matrice diagonale rappresentata dagli elementi diagonali di A, E è la matrice triangolare inferiore di coefficienti $e_{ij} = -a_{ij}$ se $i > j, e_i j = 0$ se $i \leq j$, mentre F è la matrice triangolare superiore di coefficienti $d_{ij} = -a_{ij}$ se $j > i, f_{ij} = 0$ se $j \leq i$. Di conseguenza, $A = D - (E + F)$
La matrice di iterazione corrispondente è

$$B_J = D^{-1}(E + F) = I - D^{-1}A.$$

2.3.2 Metodo di Gauss-Seidel

Il metodo di Gauss-Seidel si differenzia dal metodo di Jacobi per il fatto che al passo $k+1$ si utilizzano i valori di $x_i^{(k+1)}$ qualora sianodisponibili

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_k^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right], \quad i = 1, \ldots, n.$$

Questo metodo equivale ad aver utilizzato lo splitting $A = P - N$ dove

$$P = D - E, \quad N = F.$$

La corrispondente matrice di iterazione è data da

$$B_{GS} = (D - E)^{-1} F.$$

Teorema 11 (Convergenza di Jacobi e Gauss Seidel)
Se A è una matrice a dominanza diagonale stretta per righe, i metodi di Jacobi e Gauss-Seidel sono convergenti.

Dimostrazione
questa va fatta sul libro ma è una mezza cazzata (pag 121) □

Teorema 12 (Convergenza tridiagonali)
Nel caso in cui A sia una matrice tridiagonale (per punti o per blocchi), si può dimostrare che

$$\rho(B_{GS}) = \rho^2(B_J).$$

Da tale relazione si conclude che due metodi convergono o divergono contemporaneamente. Nel caso in cui convergano, il metodo di Gauss-Seidel converge più rapidamente

2.3.3 Metodo del rilassamento successivo (SOR)

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

per $i = 1, \ldots, n$
La matrice di iterazione è per tanto

$$B(\omega) = (1 - \omega D^{-1}E)^{-1}[(1-\omega)I + \omega D^{-1}F].$$

Si puo trovare la formula seguente per il metodo SOR

$$x^{(k+1)} = x^{(k)} + \left(\frac{1}{\omega}D = E\right)^{-1} r^{(k)}.$$

Esso risulta consistente per ogni $\omega \neq 0$ e per $\omega 1$ concide con il metodo di Gauss-Seidel. In particolare, se $\omega \in (0,1)$ il metodo si dice sottorilassamento, mentre se $\omega > 1$ si dice sovrarilassamento.

Teorema 13 (Convergenza SOR (Teorema di Kahan))
Per ogni $\omega \in \mathbb{R}$ *si ha* $\rho(B(\omega)) \geq |\omega - 1|$*, pertanto il metodo SOR diverge se* $\omega \leq 0$ *o se* $\omega \geq 2$.

Proprietà 4.3 (Ostrowski)
Se A è una matrice simmetrica definita positiva, il metodo SOR converge se e solo se $0 < \omega < 2$. Inoltre, la convergenza è monotona rispetto alla norma $\|\cdot\|_A$
Infine, se A è a dominanza diagonale stretta per righe, il metodo SOR converge se $0 < \omega \leq 1$

Proprietà 4.4
Sia A simmetrica definita positiva e tridiagonale. Allora il metodo SOR converge per ogni valore iniziale $x^{(0)}$ se $0 < \omega < 2$. In tal caso

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - \rho^2(B_J)}}.$$

ed è in corrispondenza di tale valore il fattore asintotico di convergenza è pari a

$$\rho(B(\omega_{opt})) = \frac{1 - \sqrt{1 - \rho^2(B_J)}}{1 + \sqrt{1 - \rho^2(B_J)}}.$$

2.3.4 Un criterio basato sul controllo dell'incremento

Dalla relazione ricorsiva sull'errore $e^{(k+1)} = Be^{(k)}$, otteniamo

$$\|e^{(k+1)}\| \leq \|B\|\|e^{(k)}\|.$$

Usando la disuguaglianza triangolare si trova

$$\|e^{(k+1)}\| \leq \|B\| \left(\|e^{(k+1)}\| + \|x^{(k_1)} - x^{(k)}\| \right).$$

dunque se $\|B\| > 1$

$$\|x - x^{(k+1)}\| \leq \frac{\|B\|}{1 - \|B\|} \|x^{(k+1)} - x^{(k)}\|.$$

In particolare prendendo $k = 0$ e applicando ricorsivamente la formula precedente si trova

$$\|x - x^{(k+1)}\| \leq \frac{\|B\|^{k+1}}{1 - \|B\|} \|x^{(1)} - x^{(0)}.$$

che può essere usata per sistemare il numero di iterazioni necessario per soddisfare la condizione $\|e^{(k+1)}\| \leq \varepsilon$ per unna data tolleranza ε
In pratica, $\|B\|$ può essere stimata nel modo seguente avendosi

$$x^{(k+1)} - x^{(k)} = -(x - x^{(k+1)}) + (x - x^{(k)}) = B(x^{(k)} - x^{(k+1)}).$$

2.3.5 Un criterio basato sul controllo del residuo

Un criterio d'arresto più pratico dei precedenti è quello in cui ci si ferma quando $\|r^{(k)}| \leq \varepsilon$, essendo ε una tolleranza fissata. In tal caso si ricava

$$\|x - x^{(k)}\| = \|A^{-1}b - x^{(k)}\| = \|A^{-1}r^{(k)} \leq \|A^{-1}\|\varepsilon.$$

e di conseguenza, se vogliamo che l'errore sia minore di δ, dovremo scegliere $\varepsilon \leq \delta / \|A^{-1}\|$. In generale conviene tuttavia effettuare un test d'arresto sul residuo normalizzato: ci si ferma dunque non appena $\frac{\|r^{(k)}\|}{\|r^0\|} \leq \varepsilon$ oppure quando risulta $\frac{\|r^{(k)}\|}{\|b\|} \leq \varepsilon$ (che corrisponde ad aver scelto $x^{(0)} = 0$). In quest ultimo caso si ha il seguente controllo sull'errore relativo commesso.

$$\frac{\|x - x^{(k)}\|}{\|x\|} \leq \frac{\|A^{-1}\|\|r^{(k)}\|}{\|x\|} \leq K(A) \frac{\|r^{(k)}\|}{\|b\|} \leq \varepsilon K(A).$$

Ne caso di metodi precondizionati, al residuo si sotituisce il residuo precondizionato e quindi il criterio precedente diventa

$$\frac{\|P^{-1}\|r^{(k)}\|}{\|P^{-1}r^{(0)}\|} \leq \varepsilon.$$

essendo P la matrice di precondizionamento.

3 Sistemi non lineari

Affrontiamo in questa sezione la risoluzione numerica di sistemi di equazioni non lineari della forma:

$$\text{data } F : \mathbb{R}^n \to \mathbb{R}^n, \quad \text{trovare } x^* \in \mathbb{R}^n \quad \text{tale che } F(x^*) = 0.$$

Notazione 1
Chiameremo $J_F(x)$ la matrice jacobiana associata a F e valutata in $x = (x_1, \ldots, x_n)^T$

3.1 Il metodo di Newton

Una immediata estensione al caso vettoriale del metodo di Newton si formula nel modo seguente:
dato $x^{(0)} \in \mathbb{R}^n$, per $k = 0, 1, \ldots,$ fino a convergenza

$$\text{risolvere} \quad J_F(x^{(k)})\delta x^{(k)} = -F(x^{(k)}).$$

$$\text{porre} \quad x^{(k+1)} = x^{(k)} + \delta x^{(k)}.$$

dove $\delta x^{(k)}$ è la correzione rispetto al termine precedente

Teorema 14 (Convergenza metodo di Newton)
Sia $F : \mathbb{R}^n \to \mathbb{R}^n$ una funzione vettoriale di classe C^1 in un aperto convesso D di $\mathbb{R}^n$ contenente x^*. Supponiamo che $J_F^{-1}(x^*)$ esista e che esistano delle costanti positive, R, C ed L tali che $\|J_F^{-1}(x^*)\| \leq C$ e

$$\|J_F(x) - J_F(y)\| \leq L\|x - y\| \quad \forall x, y \in B(x^*; R).$$

Avendo indicato con lo stesso simbolo $\|\cdot\|$ una norma vettoriale ed una norma matriciale consistenti. Esiste allora $r > 0$ tale che , per ogni $x^{(0)} \in B(x^*; r)$ la successione è univocamente definita, converge a x^* e

$$\|x^{(k+1)} - x^*\| \leq CL\|x^{(k)} - x^*\|^2.$$

Dimostrazione
sul libro a pagina 251 □

4 Interpolaione polinomiale

In questo capito illustreremo i principali metodi per l'apprissimazione di funzioni attraverso i loro valori nodali.

4.1 Interpolazione polinomiale di Lagrange

Consideriamo $n+1$ coppie di valori (x_i, y_i). Cerchiamo un polinomio $\Pi_n \in \mathbb{P}_m$ detto polinomio interpolatore tale che

$$\Pi_m(x_i) = a_m x_i^m + \ldots + a_1 x_i + a_0 = y_0, \quad i = 0, \ldots, n.$$

Teorema 15
Dati $n+1$ punti distinti $x_0, \ldots, x_n$ e $n+1$ corrispondenti $y_0, \ldots, y_n$ esiste un unico polinomio $\Pi_n \in \mathbb{P}_n$ tale che $\Pi_n(x_i) = y_i$ per $i = 0, \ldots, n$

La dimostrazione non è necessaria ma costruisce così il polinomio

$$\Pi_n(x_i) = \sum_{j=0}^{n} b_j l_j(x_i) = y_i, \quad i = 0, \ldots, n$$

$$l_i \in \mathbb{P}_n : \quad l_i = \prod_{\substack{i=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} \qquad i = 0, \ldots, n$$

Questa forma è detta forma di Lagrange del polinomio interpolatore.

4.2 Forma di Newton

Introduciamo una forma aternativa dal costo computazionale inferiore. Poniamo il seguente obbiettivo:
date $n+1$ coppie $\{x_i, y_i\}$ si può rappresentare Π_n come la soma di Π_{n-1} e di un polinomio di grado n dipendente dai nodi x_i e da un solo coefficiente incognito. Precisamente poniamo

$$\Pi_n(x) = \Pi_{n-1}(x) + q_n(x).$$

doev $q_n \in \mathbb{P}_n$. Poiché $q_n(x_i) = \Pi_n - \Pi_{n-1}(x_i) = 0$ per $i = 0, \ldots, n-1$, dovrà necessariamente essere

$$q_n(x) = a_n(x - x_0) \ldots (x - x_{n-1}) = a_n \omega_n(x).$$

Per determinare il coefficiente incognito a_n, supponiamo che $y_i = f(x_i)$, $i = 0, \ldots, n$ dove f è una funzione opportuna, non necessariamente nota in forma esplicita. Siccome $\Pi_n f(x_n) = f(x_n)$ si avrà

$$a_n = \frac{f(x_n) - \Pi_{n-1} f(x_n)}{\omega_n(x_n)}.$$

Il coefficiente a_n è detto n-esima differenza divisa di Newton e viene generalmente indicato con

$$a_n = f[x_0, \ldots, x_n].$$

per $n \geq 1$. Di conseguenza la prima formula diventa

$$\Pi_n f(x) = \Pi_{n-1} f(x) + \omega_n(x) f[x_0, \ldots, x_n].$$

Se poniamo $y_0 = f(x_0) = f[x_0]$ e $\omega_0 = 1$, per ricorsione su n possiamo ottenere la formula seguente

$$\Pi_n f(x) = \sum_{k=0}^{n} \omega_k(x) f[x_0, \ldots, x_k].$$

4.2.1 Formula ricorsiva per il calcolo delle differenze divise

Rielaborando algebricamente la formula si giunge alla seguente formula ricorsiva per il calcolo delle differenze divise:

$$f[x_0, \ldots, x_n] = \frac{f[x_1, \ldots, x_n] - f[x_0, \ldots, x_{n-1}]}{x_n - x_0}, \quad n \geq 1.$$

Il programma di interpolazione è il numero 57 a pagina 273

4.2.2 Stabilità della procedura

Il valore assunto dalla differenza divisa è invariante rispetto ad una permutazione degli indici dei nodi. Questo fatto puo essere sfruttato opportunamente quando problemi di stabilità suggeriscano l'uso di una opportuna permutazione degli indici (ad esempio se x è il punto in cui si vuole valuare il polinomio è conveniente utilizzare una permutazione degli indici in modo tale che $|x - x_k| \leq |x - x_{k-1}|$ con $k = 1, \ldots, n$)

4.2.3 L'errore di interpolazione

Teorema 16
Siano $x_0, \ldots, x_n$ $n+1$ nodi distinti e x un punto appartenente al dominio di una data funzione f. Supponiamo che $f \in C^{n+1}(I_x)$ essendo I_x il pi u piccolo intervallo contenente i nodi $x_0, \ldots, x_n$ ed è il punto x. Allora l'errore di interpolazione nel generico punto x è dato da

$$E_n(x) = f(x) - \Pi_n f(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}\omega_{n+1}(x).$$

con $\xi(x) \in I_x$ e dove ω_{n+1} è il polinomio nodale di grado $n+1$

4.2.4 L'errore di interpolazione usando le differenze divise

Consideriamo i nodi $x_0, \ldots, x_n$ e sia $\Pi_n f$ il polinomio interpolatore di f su tali nodi. Sia ora x un nodo distinto dai precedenti; posto $x_{n+1} = x$, denotiamo con $\Pi_{n+1} f$ il polinomio interpolatore sui nodi $x_k, k = 0, \ldots, n+1$. Usando la formula delle differenze divide di Newton si trova

$$\Pi_{n+1} f(t) = \Pi_n f(t) + (t - x_0) \ldots (t - x_n) f[x_0, \ldots x_n, t].$$

Essendo $\Pi_{n+1} f(x) = f(x)$, si ricava la seguente formula dell'errore di interpolazione in $t = x$

$$\begin{aligned} E_n(x) &= f(x) - \Pi_n f(x) = \Pi_{n+1} f(x) - \Pi_n f(x) \\ &= (x - x_0) \ldots (x - x_n) f[x_0, \ldots, x_n, x] \\ &= \omega_{n+1}(x) f[0, \ldots, x_n, x] \end{aligned}$$

Supponendo $f \in C^{(n+1)}(I_x)$ otteniamo

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

$\xi(x) \in I_x$. Per la somiglianza della formula con il resto dello svilupop in serie di Taylor, la formula di Newton del polinomio interpolatore viene paragonata ad uno sviluppo troncato intorno ad x_0 a patto che $|x_n - x_0|$ non sia troppo grande.

4.2.5 Limiti dell'interpolazione polinomiale su nodi equidistanziati

Per analizzare il comportamento dell'errore di interpolazione quando n tende all'infinito, definiamo per ogni funzione $f \in C^0([a,b])$ la norma infinito come

$$\|f\|_\infty = \max_{x\in[a,b]} |f(x)|.$$

Introduciamo una matrice triangolare inferiore X di dimensione infinita, detta matrice di interpolazione su $[a,b]$ i cui elementi $x_{ij}, i,j = 0,1,\ldots$ rappresentano i punti di $[a,b]$ con l'assunzione che in ogni riga gli elementi siano tutti distinti
In tal caso, per ogni $n \geq 0$, la $n+1$-esima riga di X contiene $n+1$ valori che possiamo identificare come nodi di inteprolazione. Per una data funzione f, si puo allora definire in modo univoco un polinomio interpolatore $\Pi_n f$ di grado n rispetto a detti nodi. Naturalmente, $\Pi_n f$ dipenderà solamente da X, oltre che da f.

Fissata f e la matrice di interpolazione X, definiamo l'errore di interpolazione

$$E_{n,\infty}(X) = \|f - \Pi_n f\|_\infty, \quad n = 0,1,\ldots.$$

Indichiamo con $p^* \in \mathbb{P}_n$ il polinomio di miglior approssimazione uniforma (detto in inglese polinomio di best approximation), in corrispondenza del quale si ha

$$E_n^* = \|f - p_n^*\|_\infty \leq \|f - q_n\|_\infty \quad \forall q \in \mathbb{P}_n.$$

Vale il seguente risultato di confronto

Proprietà 7.1

Sia $f \in X^0([a,b])$ e X una matrice di interpolazione su $[a,b]$. Allora

$$E_{n,\infty}(X) \leq (1 + \Lambda_n(X))E_n^*, \quad n = 0,1,\ldots.$$

dove $\Lambda_n(X)$ denota la costante di Lebesgue di X definita come

$$\Lambda_n(X) = \|\sum_{j=0}^{n} |l_j^{(n)}|\|_\infty.$$

essendo $l_j^{(n)} \in \mathbb{P}_n$ il j-esimo polinomio caratteristico associato alla $n+1$-esima riga di X, ovvero tale che $l_j^{(n)}(x_{nk}) = \delta_{jk}, \quad j,k = 0,1,\ldots$

Essendo E_n^* indipendente da X, tutte le informazioni relative agli effetti di X su $E_{n,\infty}(X)$ andranno ricercate in $\Lambda_n(X)$. Pur esistendo una matrice

di interpolazione X^* in corrispondenza della quale $\Lambda_n(X)$ è minimo, non è possibile, se non in rari casi, determiunare esplicitamente gli elementi.

D'altro canto per ogni possibile scelta di X, esiste una costante $C > 0$ tale che

$$\Lambda_n(X) > \frac{2}{\pi} \log(n+1) - C.$$

Questa proprietà mostra che $\Lambda_n(X) \to +\infty$ per $n \to +\infty$. In particolare si può dedurre che per ogni matrice di interpolazione X su un intervallo $[a, b]$ esiste sempre una funzione f, continua in $[a, b]$ tale che $\Pi_n f$ non converge uniformemente (nella norma massimo) ad f. Non è quindi possibile approssimare tramite l'interpolazione polinomiale tutte le funzioni continue, il seguente ne è un esempio.

Controesempio di Runge

Si voglia approssimare la seguente funzione

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

con l'interpolazione di Lagrange su nodi equispaziati. Si verifica che esistono dei punti x interni all'intervallo tali che

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

In particolare si ha divergenza per $|x| > 3.63\ldots$. Questo fenomeno è particolarmente marcato agli estremi dell'intervallo di interpolazione, come mostrato nella figura (pag 270) ed è legato alla scelta di nodi equispaziati.

4.3 Stabilità dell'interpolazione polinomiale

Supponiamo di considerare una approssimazione, $\tilde{f}(x_i)$, di un insieme di dati $f(x_i)$ relativi ai nodi x_i, con $i = 0, \ldots, n$ in un intervallo $[a, b]$. La perturbazione $f(x_i) - \tilde{f}(x_i)$ potrà essere dovuta ad esempio all'effetto degli errori di arrotondamento, oppure causata da un errore nella misurazione dei dati stessi.
Indicando con $\Pi_n \tilde{f}$ il polinomio interpolatore corrispondente ai valori $\tilde{f}(x_i)$ si ha

$$\begin{aligned} \|\Pi_n f - \Pi_n \tilde{f}\|_\infty &= \max_{a \le x \le b} \left| \sum_{j=1}^{n} (f(x_j) - \tilde{f}(x_j)) l_j(x) \right| \\ &\le \Lambda_n(X) \max_{i=0,\ldots,n} |f(x_i) - \tilde{f}(x_i)| \end{aligned}$$

Di conseguenza, a piccole perturbazione sui dati corrisponderanno piccole variazioni sul polinomio interpolatore puchè la costante di Lebesgue sia piccola. Quest'ultima assume il significato di numero di condizionamento del problema dell'interpolazione. Come abbiamo già osservato, Λ_n cresce per $n \to +\infty$ ed in particolare, nel caso dell'interpolazione polinomiale di Lagrange su nodi equispaziati, si trova

$$\Lambda_n(X) \simeq \frac{2^{n+1}}{en(\log n + \gamma)}.$$

dove e è il numero di nepero e $\gamma \simeq_0 .547721$ rappresenta la costante di Eulero. Di conseguenza per n grande questo tipo di interpolazione può essere instabile. Facciamo notare come siano stati del tutto trascurati gli errori generati dal processo di interpolazione nella costruzione di Π_n. Tuttavia, l'effetto di questi errori è in generale trascurabile

Esempio
Sull'intervallo $[-1, 1]$ interpoliamo la funzione $f(x) = \sin(2\pi x)$ con 22 nodi equidistanti x_i, Generiamo un insieme perturbato di valori $\tilde{f}(x_i)$ in modo che

$$\max_{i=0,\ldots,21} |f(x_i) - \tilde{f}(x_i) \simeq 9.5 \cdot 10^{-4}.$$

4.4 Interpolazione composita di Lagrange

In generale, per distribuzioni equispaziate dei nodi di interpolazione, non si ha convergenza uniforme di $\Pi_n f$ a f per $n \to \infty$. D'altra parte, da un

lato l'equispaziatura dei nodi presenta considerevoli vantaggi ocmputazionali, dall'altro l'interpolazione di Lagrange risulta ragionveolmente accurata per gradi bassi, a patto di interpolare su intervalli sufficientemente piccoli.
È per tanto naturale introdurre una partizione $\mathcal{T}_h$ di $[a,b]$ in M sottointervalli $I_j = [x_j, x_{j+1}]$ con $j = 0, \ldots, M$ di lunghezza h_j, con $h = \max_{0 \le j \le M-1} h_j$, tali che $[a,b] = \mathbb{U}_{j=0}^{M-1} I_j$ ed usare poi l'interpolazione di Lagrange su ciascun intervallo I_j con k piccolo e su $k+1$ nodi equidistanziati $\{x_J^{(i)}, 0 \le i \le k\}$.

Per $k \ge 1$, introduciamo su $\mathcal{T}$ lo spazio dei polinomi compositi

$$X_h^k = \{v \in C^0([a,b]) \ : \ v|_{I_j} \in \mathbb{P}_k(I_j) \forall I_j \in \mathcal{T}_h\}.$$

definito come lo spazio delle funzioni continue su $[a,b]$ le cui restrizioni a ciascun I_j sono polinomi di grado $\le k$. Allora, per ogni funzione f continua su $[a,b]$ il polinomio inteprolatore composito, $\Pi_h^k f$, coincide su I_j con il polinomio interpolatore di $f_{|I_j}$ sui $k+1$ nodi $\{x_j^{(i)}, 0 \le i \le k\}$. Di conseguenza, se $f \in C^{k+1}([a,b])$, su ricava la seguente stima dell'errore

$$\|f - \Pi_h^k f\|_\infty \le C h^{k+1} \|f^{(k+1)}\|_\infty.$$

4.5 Funzioni spline

Introduciamo in questa sezione le funzioni splines, che consentono di effetuare l'interpolazione di una funzione attraverso polinomi compositi non solo continui, ma anche derivabili su tutto l'intervallo $[a,b]$

Definizione 14
Siano $x_0, \ldots, x_n, n+1$ nodi distinti e ordinati sull'intervallo $[a,b]$ con $a = x_0 < x_1 < \ldots < x_n = b$. Una funzione $s_j(x)$ sull'intervallo $[a,b]$ è detta spline di grado k $(k \ge 1)$ relativa ai nodi x_j se

$$s_{k|[x_j, x_{j+1}]} \in \mathbb{P}_k, \quad j = 0, 1 \ldots, n-1.$$

$$s_k \in C^{k-1}[a,b].$$

4.5.1 Spline cubiche interpolatorie

Le spline cubiche di grfado 3 di tipo interpolatorio hanno un rilievo particolare, in quanto

1. sono le spline di grado minimo che consentano di ottenere approssimazioni almeno di classe C^2

2. sono sufficientemente regolari in presenza di piccole curvature.

Consideriamo dunque in $[a, b], n+1$ nodi ordinati $a = x_0 < x_1 < \ldots < x_n = b$ e le corrispondenti valutazioni f_i. Si vuole fornire una procedura efficiente per la costruzione della spline cubica interpolante tali valori. Essendo la spline di grado 3, essa dovrà presentare derivate continue fino al second'ordine. Introduciamo le seguenti notazioni

$$f_i = s_3(x_i), \quad m_i = s_3'(x_i), \quad M_i = s_3''(x_i), \quad i = 0, \ldots, n.$$

Avendosi $s_{3,i-1}(x) \in \mathbb{P}_3, s_{3,i-1}''$ sarà lineare e

$$s_{3,i-1}'' = M_{i-1}\frac{x_i - x}{h_i} + M_i\frac{x - x_{i-1}}{h_i} \quad \text{per } x \in [x_{i-1}, x_i]$$

essendo $h_i = x - x_i, i = 1, \ldots, n$. Integrando l'uguaglianza precedente due volte otteniamo

$$s_{3,i-1} = M_{i-1}\frac{(x_i - x)^3}{6h_i} + M_i\frac{(x - x_i)^3}{6h_i} + C_{i-1}(x - x_{i-1} + \tilde{C}_{i-1}.$$

e le costanti C_{i-1} e $\tilde{C}_{i-1}$ vengono deliminate imponendo che $s_3(x_{i-1} = f_{i-1}$ e $s_3(x_i) = f_i$. Si ricava quindi che, per $i = 1, \ldots, n-1$

$$\tilde{C}_{i-1} = f_{i-1} - M_{i-1}\frac{h_i^2}{6}, \quad C_{i-1} = \frac{f_i - f_{i-1}}{h_i} = \frac{h_i}{6}(M_i - M_{i-1}).$$

Imponiamo ora la continuità delle derivate prime in x_i; avremo:

$$s_{3,i-1}'(x_i) = \frac{h_i}{6}M_{i-1} + \frac{h_i}{3}M_i + \frac{f_i - f_{i-1}}{h_i} \tag{5}$$

$$= -\frac{h_{i+1}}{3}M_i - \frac{h_{i+1}}{6}M_{i+1} + \frac{f_{i+1} - f_i}{h_{i+1}} = s_{3,i}'(x_i) \tag{6}$$

Si giunge così al seguente sistema lineare (detto di M-continuità)

$$\mu_i M_{i-1} + 2M_i + \lambda_i M_{i+1} = d_i \quad i = 1, \ldots, n-1.$$

avendo posto

$$\mu_i = \frac{h_i}{h_i + h_{i+1}}, \qquad \lambda_i = \frac{h_{i+1}}{h_i + h_{i+1}}.$$

$$d_i = \frac{6}{h_i + h_{i+1}} \left(\frac{f_{i+1} - f_i}{h_{i+1}} - \frac{f_i - f_{i-1}}{h_i} \right), \quad i = 1, \ldots, n-1.$$

Il sistema ha $n+1$ incognite e $n-1$ equazioni: servono ancora $2(= k-1)$ condizioni. In generale, queste condizioni potranno essere della formula

$$2M_0 + \lambda_0 M_1 = d_0, \quad \mu_n M_{n-1} + 2M_n = d_n.$$

con $0 \leq \lambda_0, \mu_n \leq 1$ e d_0, d_n valorai assegnati. Ad esempio nel caso in cui si vogliano ottenere spline naturali (soddisfacenti $s_3''(a) = s_3''(b) = 0$), si prenderanno tali coefficienti tutti nulli.

Proprietà 7.2

Sia $f \in C^2([a,b])$ e sia s_3 la spline cubica natruale interpolante di f. Allora

$$\int_a^b [s_3''(x)]^2 dx \leq \int_a^b [f''(x)]^2 dx.$$

valendo l'uguaglianza se e solo se $f = s_3$

Proprietà 7.3

Sia $f \in C^4([a,b])$ e si consideri una partizione di $[a,b]$ in sottointervalli di ampiezza h_i, sia s_3 la spline cubica interpolante f. Allora

$$\|f^{(r)} - s_3^{(r)}\|_\infty \leq C_r h^{4-r} \|f^{(4)}\|_\infty \quad r = 0, 1, 2, 3.$$

dove $h = \max_i h_i, C_0 5/384, C_1 = 1/24, C_2 = 3/8, C_3 = (\beta + \beta)^{-1}/2$ e $\beta = h/\min_i h_i$

5 Integrazione numerica

In questo capitolo vengono illustrati i metodi più comunemente usati per il calcolo numerico degli integrali.

Sia f una funzione reale integrabile sull'intervallo $[a,b]$. Il suo integrale definito $I(f) = \int_a^b f(x)dx$ può non essere sempre valutabile in forma esplicita

o comunque può risultare difficile da calcolare. Una qualunque formula esplicita che consenta di approssimare $I(f)$ viene detta formula di quadratura o formula di integrazione numerica.

Un esempio può essere ottenuto sostituendo ad f una sua approssimazione f_n, dipendente dall'intero $n \geq 0$, e calcolare $I(f_n)$ in luogo di $I(f)$. Ponendo $I_n(f) = I(f_n)$ si ha dunque

$$I_n(f; a, b) = \int_a^b f_n(x)dx \quad n \geq 0.$$

La dipendenza dagli estremi di integrazione a, b sarà omessa e scriveremo semplicemente $I_n(f)$.

Se $fC^0([a, b])$, l'errore di quadratura $E_n(f) = I(f) - I_n(f)$ soddisfa

$$|E_n(f)| \leq \int_a^b |f(x) - f_n(x)|dx \leq (b - a)\|f - f_n\|_\infty.$$

e dunque, se per qualche n, $\|f - f_n\|_\infty < \varepsilon$, si avrà $|E_n(f)| \leq \varepsilon(b - a)$.

Ovviamente, sarà opportuno scegliere f_n in modo tale che il suo integrale sia facile da calcolare. Un modo naturale di proecdere consiste nello scegliere $f_n = \Pi_n f$ il polinomio interpolatore di Lagrange di f su $n + 1$ nodi distinti, in tal caso si ha

$$I_n(f) = \sum_{i=0}^{n} f(x_i) \int_a^b l_i(x)dx.$$

dove l_i è il polinomio caratteristico di Lagrange di grado n relativo al nodo x_i, in generale è un caso particolare della seguente formula

$$I_n(f) = \sum_{i=0}^{n} \alpha_i f(x_i).$$

nel caso in cui i coefficienti α_i della combinazione lineare siano dati da $\int_a^b l_i(x)dx$

5.1 Formule di quadratura interpolatorie

Vediamo tre esempi significativi della formula precedente,corrispondenti a $n = 0, 1$ e 2

5.1.1 Formula punto medio o del rettangolo

Sostituiamo a f su $[a,b]$ la funzione costante pari al valore assunto da f nel punto medio di $[a,b]$. Avremo

$$I_0(f) = (b-a)f\left(\frac{a+b}{2}\right).$$

con peso $\alpha_0 = b-a$ e nodo $x_0 = (a+b)/2$. Se $f \in C^2([a,b])$, l'errore di quadratura è dato da

$$E_0(f) = \frac{h^3}{3}f''(\xi),\ h = \frac{b-a}{2}.$$

essendo ξ un punto interno all'intervallo di integrazione (a,b).
Infatti, sviluppando la funzione f in serie di Taylod nell'intorno di $c = (a+b)/2$ ed arrestandosi al secondo termine, si ottiene

$$f(x) = f(c) + f'(c)(x-c) + f''(\eta(x))(x-c)^2/2.$$

da cui integrando e usando i lteorema del valor medio integrale si ottiene la formula precedente.
Introducendo i nodi di quadratura $x_k = a+(2k+1)H/2$, per $k = 0,\ldots,m-1$, si ottiene la formula del punto medio composita

$$I_{0,m}(f) = H\sum_{k=0}^{m-1} f(x_k), \quad m \geq 1.$$

L'errore di quadratura $E_{o,m}(f) = I(f) - I_{0,m}(f)$ è dato da

$$E_{0,m}(f) = \frac{b-a}{24}H^2 f''(\xi), \quad H = \frac{b-a}{m}.$$

5.1.2 La formula del trapezio

Questa formula si ottiene sostituendo a f il suo polinomio interpolatore di Lagrange di grado 1, $\Pi_1 f$, relativo ai nodi $x_0 = a$ e $x_1 = b$. L'espressione della formula di quadratura, che ha nodi $x_0 = a, x_1 = b$ e pesi $\alpha_0 = \alpha_1 = (b-a)/2$ risulta essere

$$I_1(f) = \frac{b-a}{2}[f(a) - f(b)].$$

Se $f \in C^2([a,b])$, l'errore di quadratura è dato da

$$E_1(f) = -\frac{h^3}{12} f''(\xi), \quad h = b - a.$$

essendo ξ un punto interno all'intervallo di integrazione. Per ottenere la formula composita si procede come nel caso $n = 0$

$$I_{1,m} = \frac{H}{2} \sum_{k=0}^{m-1} (f(x_k) + f(x_{k+1})) \quad m \geq 1.$$

Possiamo dunque scrivere

$$I_{1,m}(f) = H[\frac{1}{2} f(x_0) + f(x_1) + \ldots + f(x_{m-1}) = \frac{1}{2} f(x_m)].$$

Si dimostra che l'errore di quadratura associato è dato da

$$E_{1,m}(f) = -\frac{b-a}{12} H^2 f''(\xi).$$

purché $f \in C^2([a,b])$ ed essendo $\xi \in (a,b)$. Il grado di esattezza è ancora pari a 1.

5.1.3 Formula della parabola

La formula si ottiene integrando sull'intervallo $[a,b]$ anziché f, il suo polinomio interpolatore di grado 2 nei nodi $x_0 = a, x_1 = (a+b)/2$ e $x_2 = b$. I pesi risultano dati pertanto da $\alpha_0 = \alpha_2 = (b-a)/6$ e $\alpha_1 = 4(b-a)/6$, e la formula risulta essere

$$I_2(f) = \frac{b-a}{6} \left[f(a) + 4f(\frac{a+b}{2}) + f(b) \right].$$

Si può dimostrare che l'errore di quadratura è dato da

$$E_2(f) = -\frac{h^5}{90} f^{(4)}(\xi), \; h = \frac{b-a}{2}.$$

purché $f \in C^4([a,b])$, ed essendo ξ un punto interno all'intervallo (a,b). Si deduce che la prima formula ha grado di esattezza 3.

Sostituendo ad f il polinomio inteprolatore composito di grado 2 su $[a,b]$ si perviene dalla formula composita. Introduciamo i nodi di quadratura $x_k = a + kH/2,$, per $k = 0, \ldots, 2m$ e $H = (b-a)/m$ con $, \geq 1$ si ha

$$I_{2,m} = \frac{H}{6} \left[f(x_0) + 2 \sum_{r=1}^{m-1} f(x_{2r}) + 4 \sum_{s=0}^{m-1} f(x_{2s+1}) + f(x_{2m}) \right].$$

5.2 Formula di Newton-Cotes

Queste formule sono basate sul metodo di interpolazione di Lagrange con nodi equidistanziati in $[a, b]$. Per $n \geq 0$ fissato, indichiamo i nodi di quadratura con $x_k = x_0 + kh, k = 0, \ldots, n$. Le formule del punto medio, del trapezio sono esempi di formule di Newton-Cotes, dove, rispettivamente, $n = 0, n = 1$. Nel caso generale si definiscono:

- **formule chiuse**, quelle in cui $x_0 = a, x_n = b$ e $h = \frac{b-a}{n}$ $(n \geq 1)$
- **formule aperte**, quelle in cui $x_0 = a + h, x_n = b - h$ e $h = \frac{b-a}{n-2}$ $(n \geq 0)$

Una proprietà interessante delle formule di Newton-Cotes è quella di avere pesi di quadratura α_i che dipendono solo da n e da h, ma non dall'intervallo di integrazione $[a, b]$.
Si può ricavare la formula dei pesi di quadratura

$$\alpha_i = \int_a^b l_i(x)dx = \int_0^n \varphi_i(t)hdt = h\int_0^n \varphi_i(t)dt.$$

con

$$l_i(x) = \prod_{\substack{k=0 \\ k\neq i}}^{n} \frac{t-k}{i-k} = \varphi_i(t) \quad 0 \leq i \leq n.$$

da cui si ottiene ela formula di quadratura

$$I_n(f) = h\sum_{i=0}^{n} w_i f(x_i), \quad w_i = \int_0^n \varphi_i(t)dt.$$

In modo analogo lavorando si possono reinterpretare le formule aperte

$$I_n(f) = h\sum_{i=0}^{n} w_i f(x_i), \quad w_i = \int_{-1}^{n+1} \varphi_i(t)dt.$$

Teorema 17
Data una formula di Newton-Cotes con n pari, aperta o chiusa, vale la seguente rappresentazione dell'errore

$$E_n(f) = \frac{M_n}{(n_2)!}h^{n+3}f^{(n+2)}(\bar{\xi}).$$

purché $f \in C^{n+2}$, *dove* $\bar{\xi} \in (a,b)$ *e*

$$M_n = \begin{cases} \int_0^n t\pi_{n+1}(t)dt < 0 & \text{per formule chiuse,} \\ \int_{-1}^{n+1} t\pi_{n+1}(t)dt > 0 & \text{per formule aperte} \end{cases}.$$

avendo definito $\pi_{n+1}(t) = \prod_{i=0}^{n}(t-i)$. *Si deduce che il grado di esattezza è pari a* $n+1$ *e che l'errore è proporzionale a* h^{n+3}.
Similmente per n *dispari vale la seguente rappresentazione dell'errore*

$$E_n(f) = \frac{K_n}{(n+1)!} h^{n+2} f^{(n+1)}(\bar{\eta}).$$

purché $f \in C^{n+1}([a,b])$, *dove* $\bar{\eta} \in (a,b)$ *e*

$$K_n = \begin{cases} \int_0^n \pi_{n+1}(t)dt < 0 & \text{per formule chiuse,} \\ \int_{-1}^{n+1} \pi_{n+1}(t)dt > 0 & \text{per formule aperte} \end{cases}.$$

Dunque il grado di esattezza è pari a n *e l'errore è proporzionale a* h^{n+2}

5.3 Formule di Newton-Cotes composite

Si può ottenere la formula di quadratura interpolatoria composita sostituendo $I(f)$ con

$$I_{n,m}(f) = \sum_{j=0}^{m-1} \sum_{k=0}^{n} \alpha_k^{(j)} f(x_k^{(j)}).$$

L'errore che si genera sarà $E_{n,m}(f) = I(f) - I_{n,m}(f)$

Teorema 18
Data una formula di Newton-Cotes composita, aperta o chiusa su ogni sottointervallo e con n *pari, se* $f \in C^{n+2}([a,b])$ *si ha*

$$E_{n,m}(f) = \frac{b-a}{(n_2)!} \frac{M_n}{\gamma_n^{n+3}} H^{n+2} f^{(n+2)}(\xi).$$

per un opportuno $\xi \in (a,b)$. Pertanto, la formula di quadratura ha ordine di infinitesimo $n+2$ rispetto ad H e di grado di esattezza pari a $n+1$.

Nel caso in cui n sia dispari, se $f \in C^{n+1}([a,b])$ si ha

$$E_{n,m}(f) = \frac{b-a}{(n+1)!}\frac{K_n}{\gamma_m^{n+2}}H^{n+1}f^{(n+1)}(\eta).$$

per un opportuno $\eta \in (a,b)$. Pertanto, la formula di quadratura ha ordine di infinitesimo $n+1$ rispetto ad H e grado di esattezza pari a n.
nelle due formule precedenti si ha $\gamma_n = (n+2)$ se la formula è aperta e $\gamma_n = n$ se la formula è chiusa.

5.4 L'estrapolazione di Richardson

Il metodo di estrapolazione di Richardson è una procedura che combina opoprtunamente varie approssimazioni di una certa quantità α_0 in modo da trovare una stima più accurata di α_0 con una quantità $A(h)$ che sia calcolabile per ogni valore del parametro $h \neq 0$. Assumiamo inoltre che per $A(h)$ valga uno sviluppo del tipo

$$A(h) = \alpha_0 + \alpha_1 h + \ldots + \alpha_k h^k + R_{k+1}(h)..$$

per un opportuno $k \geq 0$, dove $|R_{k+1}| \leq C_{k+1}h^{K+1}$. La costante C_{k+1} e i coefficienti α_i sono indipendenti da h. Quindi $\alpha_0 = \lim_{h\to 0} A(h)$.

Scrivendo la precedente formula con δh invece di h con $0 < \delta < 1$, si ottiene

$$A(\delta h) = \alpha_0 + \alpha_1(\delta h) + \ldots + \alpha_k(\delta h)^k + R_{k+1}(\delta h).$$

Possiamo calcolare quindi

$$B(h) = \frac{A(\delta h) - \delta A(h)}{1-\delta} = \alpha_0 + \tilde{\alpha}_2 h^2 + \ldots \tilde{\alpha}_k h^k + \tilde{R}_{k+1}(h).$$

avendo definito, per $k \geq 2, \tilde{\alpha}_i = \alpha_i(\delta^i - \delta)/(1-\delta)$, per $i = 2, \ldots, k$ e $\tilde{R}_{k+1}(h) = [R_{k+1}(\delta h) - \delta R_{k+1}(h)]/(1-\delta)$

Si noti che $\tilde{\alpha}_i \neq 0$ se e solo se $\alpha_i \neq 0$; dunque, se in particolare risulta $\alpha_1 \neq 0$, allora $A(h)$ è un'approssimazione al prim'ordine per α_0, mentre $B(h)$ lo è almeno al second'ordine. Più in generale, se $A(h)$ è un'approssimazione di α_0 all'ordine p, allora la quantità $B(h) = [A(\delta h) - \delta^p A(h)]/(1-\delta^p)$ è un'approssimazione di $\alpha 0$ almeno all'ordine $p+1$.

Procedendo per induzione si genera il seguente algoritmo di estrapolazione di Richardson, fissati $n \geq 0, h > 0$ e $\delta \in (0,1)$ si costruiscono le successioni

$$A_{m,0} = A(\delta^m h) \qquad m = 0, \ldots, n \tag{7}$$

$$A_{m,q+1} = \frac{A_{m,q} - \delta^{q+1} A_{m-1,q}}{1 - \delta^{q+1}} \quad \substack{q=0,\ldots,n-1 \\ m=q+1,\ldots n} \tag{8}$$

6 Risoluzione numerica di equazioni differenziali ordinarie

In questo capitolo affrontiamo la risoluzione numerica del problema di Cauchy per le equazioni differenziali.

Definizione 15 (Problema di Cauchy (ODE))
trovare $y \in C^1(I)$ a valori reali tale che

$$\begin{cases} y'(t) = f(t, y(t)) & t \in I \\ t(t_0) = y_0 \end{cases} .$$

dove $f(t,y)$ è una funzione assegnata nella striscia $S = I \times (-\infty, +\infty)$ a valori reali e continua rispetto ad entrambe le variabili.

Definizione 16 (Problema di Cauchy vettoriale)

$$\begin{cases} y' = F(t, y) \\ y(t_0) = y_0 \end{cases} .$$

dove y è il vettore soluzione e $F : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ unna funzione assegnata.

Proprietà (esistenza ed unicità in grande Sia $F : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ una funzione continua su $D = [t_0, T] \times \mathbb{R}^n$ con t_0 e T finiti. Allora, se esiste una costante positiva L tale che la disuguaglianza

$$\|F(t,y) - F(t,\bar{y})\| \leq L\|y - \bar{y}\|.$$

valga per ogni $(t,y), (t,\bar{y}) \in D$, allora per ogni $y_0 \in \mathbb{R}^n$ esiste un'unica y continua e differenziabile per ogni $(t,y) \in D$, soluzione del problema di Cauchy

6.1 Metodi numerici ad un passo

Consideriamo l approssimazione numerica del problema di Cauchy, fissato $0 < T < +\infty$, sia $I = (t_0, t_0 + T)$ l'intervallo di integrazione e, in corrispondenza di $h > 0$, sia $t_n = t_0 + nh$, con $n = 0, 1, 2, \ldots, N_h$, la successione dei nodi di discretizzazione di I in sottointervalli $I_n = [t_n, t_{n+1}]$ con $n = 0, 1, \ldots, N_h - 1$. L'ampiezza h di tali intervalli verrà detta passo di discretizzazione. Si noti che N_h è il massimo intero per il quale risulti $t_{N_h} \leq t_0 + T$. Indichiamo con u_j l'approssimazione del nodo t_j della soluzione esatta $y(t_j)$; quest'ultima verrà denotata per comodità come y_j. Analogamente, f_j indicherà il valore di $f(t_j, u_j)$. Ovviamente in generale si porrà $u_0 = y_0$

Definizione 17 (Metodo numerico ad un passo)
Un metodo numerico per l'approssimazione del problema ODE si dice ad un passo se $\forall n \geq 0, u_{n+1}$ dipende solo da u_n. In caso contrario si dirà a più passi o multistep.

Definizione 18 (Metodi espliciti ed impliciti)
Un metodo si dice esplicito se u_{n+1} si ricava direttamente in funzione dei valori nei soli punti precedenti. Un metodo è implicito se u_{n+1} dipende implicitamente da se stessa attraverso f.

Alcuni di questi metodi sono

1. **Il metodo di Eulero in avanti (o di Eulero Esplicito**

$$u_{n+1} = u_n + hf_n.$$

2. **Il metodo di Eulero all'indietro (o di Eulero implicito)**

$$u_{n+1} = u_n + hf_{n+1}.$$

3. **Il metodo del trapezio (o di Cranck-Nicolson)**

$$u_{n+1} = u_n + \frac{h}{2}[f_n + f_{n+1}].$$

4. **Il metodo di Heun**

$$u_{n+1} = u_n + \frac{h}{2}[f_n + f(t_{n+1}, u_n + hf_n)].$$

5. **Eulero modificato**

$$u_{n+1} = u_n + hf(t_n + \frac{h}{2}, u_n + \frac{h}{2}f_n), \quad n \geq 0.$$

6.2 Analisi dei metodi ad un passo

Ogni metodo esplicito ad un passo si può scrivere nella forma compatta

$$u_{n+1} = u_n + h\Phi(t_n, u_n, f_n : h), \quad 0 \leq n \leq N_h - 1, \quad u_0 = y_0.$$

dove $\Phi(\cdot\cdot, \cdot; \cdot)$ è detta funzione di incremento. Ponendo come al solito $y_n = y_n(t_n)$, in analogia alla formula precedente possiamo scrivere

$$y_{n+1} = y_n = h\Phi(t_n, y_n, f(t_n, y_n); h) + \varepsilon_{n+1}, \quad 0 \leq n \leq N_h - 1.$$

dove ε_{n+1} è il residuo che si genera nel punto t_{n+1} avendo preteso che la soluzione esatta soddisfi lo schema numerico. Riscriviamo il residuo nella forma seguente

$$\varepsilon_{n+1} = h\tau_{n+1}(h).$$

La quantità $\tau_{n+1}(h)$ è detta errore di troncamento locale nel nodo t_{n+1}. Definiamo allora errore di troncamento globale la quantità

$$\tau(h) = \max_{0 \leq N_h - 1} |\tau_{n+1}(h)|.$$

Si noti che $\tau(h)$ dipende dalla funzione y, soluzione del problema di Chauchy. Il metodo di Eulero in avanti è un caso particolare della prima formula ove si ponga

$$\Phi(t_n, u_n, f_n; h) = f_n.$$

mentre per ritrovare il metodo di Heun si deve porre

$$\Phi(t_n, u_n, f_n; h) = \frac{1}{2}[f_n + f(t_n + h, u_n + hf_n).$$

Uno schema esplicito ad un passo è completamente caratterizzato dalla sua funzione di incremento Φ. Quest'ultima, in tuttii casi sin qui considerati è tale che

$$\lim_{h \to 0} \Phi(t_n, u_n, f_n; h) = f(t_n, y_n), \quad \forall t_n \geq t_0.$$

La proprietà precedente, unita all'ovvia proprietà che $y_{n+1} - y_n = hy'(t_n) + O(h^2) \ \ \forall n \geq 0$, assicura che segua $\lim_{h\to 0} \tau_{n+1}(h) = 0, \ 0 \leq n \leq N_h - 1$. A sua volta questa condizione garantisce che

$$\lim_{h\to 0} \tau(h) = 0.$$

Proprietà che esprime la consistenza del metodo numerico ad inizio pagina con il problema di Cauchy. In generale un metodo si dirà consistente quando il suo LTE (local truncation error) è infinitesimo rispetto ad h. Inoltre uno schema ha ordine p se, $\forall t \in I$, la soluzione $y(t)$ del problema di Cauchy soddisfa la condizione

$$\tau(h) = O(h^p) \ \text{ per } h \to 0.$$

Usando gli sviluppi di Taylor si può stabilire che il metodo di Eulero in avanti ha ordine uno, mentre il metodo di Crank-Nicolson e quello di Heun hanno ordine due.

6.3 Metodo di Runge-Kutta a 3 stadi

$$u_{n+1} = u_n + \frac{h}{6}(K_1 + 2K_2 + 2K_3 + K_4)$$

$K_1 = f_n$
$K_2 = f(t_n + \frac{h}{2}, u_n + \frac{h}{2}K_1)$
$K_3 = f(t_n + \frac{h}{2}, u_n + \frac{h}{2}K_2)$
$K_4 = f(t_{n+1}, u_n + hK_3)$

6.4 Zero stabilità

Il metodo numerico ad un passo per la risoluzione del problema di Cauchy è zero-stabile se $\exists h_0 > 0, \exists C > 0$ ed $\exists \varepsilon_0 > 0$ tali che $\forall h \in (0, h_0]$ e $\forall \varepsilon \in (0, \varepsilon_0]$, se $|\delta_n| \leq \varepsilon$, $0 \leq n \leq N_h$ allora

$$|z_n^{(h)} - u_n^{(h)} \leq C\varepsilon, \ \ 0 \leq n \leq N_h.$$

dove $z_n^{(h)}$ e $u_n^{(h)}$ sono rispettivamente le soluzioni dei problemi

$$\begin{cases} z_{n+1}^{(h)} = z_n^{(h)} + h[\Phi(t_n, z_n^{(h)}, f(t_n, z_n^{(h)}); h) + \delta_{n+1}], n = 0, \ldots, N_h - 1 \\ z_0^{(h)} = y_0 + \delta_0 \end{cases}.$$

$$\begin{cases} u_{n+1}^{(h)} = u_n^{(h)} + h\Phi(t_n, u_n^{(h)}, f(t_n, u_n^{(h)}); h), n = 0, 1, \ldots, N_h - 1 \\ u_0^{(h)} = y_0 \end{cases}.$$

6.5 Prima barriera di Dahlquist per i metodi ad un passo lineari

Caso Esplicito: La barriera dice $p \leq k$. Con $k = 1$, otteniamo $p \leq 1$. Esempio: Il metodo di Eulero in avanti è esplicito a un passo ed ha ordine 1. Non è possibile costruire un metodo lineare esplicito a un passo con ordine > 1 (senza aggiungere stadi interni, vedi sotto).

Caso Implicito : La barriera dice $p \leq k + 1$ (per k dispari). Con $k = 1$, otteniamo $p \leq 2$. Esempio: Il metodo di Crank-Nicolson (o dei Trapezi) è implicito a un passo ed ha ordine 2. Questo è il massimo ordine raggiungibile per questa struttura.

Conclusione: Se restiamo nella struttura lineare classica $y_{n+1} = y_n + h(\ldots)$, non possiamo superare l'ordine 2.