Dispense di

Metodi Numerici per le Equazioni Differenziali

Dott. Marco Caliari

a.a. 2015/16

Questi appunti non hanno nessuna pretesa di completezza. Sono solo alcune note ed esercizi che affiancano l'insegnamento di Metodi Numerici per le Equazioni Differenziali. Sono inoltre da considerarsi in perenne "under revision" e pertanto possono contenere discrepanze, inesattezze o errori.

Questa è la versione del 24 settembre 2015. La versione più aggiornata si trova all'indirizzo

http://profs.scienze.univr.it/caliari/aa1516/equazioni_differenziali/dispense.pdf

Tutti i grafici riportano a fianco il link al codice usato per ottenerli. Nel caso in cui il link non funzionasse, i codici si trovano all'indirizzo

http://profs.scienze.univr.it/caliari/aa1516/equazioni_differenziali/

Indice

Parte 0

Preliminari

Capitolo 1

Matrici e autovalori

Una matrice $A \in \mathbb{C}^{n\times n}$ è detta *normale* se $AA^* = A^*A$, ove A^* denota la trasposta coniugata. Esempi di matrici normali sono le hermitiane (simmetriche se reali) $A^* = A$, le antihermitiane (antisimmetriche se reali) $A^* = -A$ e le unitarie (ortogonali se reali) $A^* = A^{-1}$. Vale il seguente

Teorema 1 (Teorema spettrale)**.** *Una matrice A è normale se e solo se si può decomporre come*

$$A = UDU^*$$

ove U è una matrice unitaria e D la matrice diagonale degli autovalori di A.

In particolare, gli autovettori di una matrice normale (le colonne di U) relativi ad autovalori distinti sono tra loro ortogonali e dunque formano una base per $\mathbb{C}^n$.

Il *campo dei valori (field of values o numerical range)* $W(A)$ di A è il sottoinsieme di $\mathbb{C}$ dei numeri della forma

$$W(A) = \{z \in \mathbb{C}\colon z = x^*Ax,\ x \in \mathbb{C}^n,\ x^*x = 1\}$$

Gli autovalori di A stanno nel campo dei valori: infatti, se $Ax = \lambda x$, con x di norma unitaria, allora $\lambda = x^*Ax$. Il campo dei valori gode della proprietà subadditiva: $W(A+B) \subseteq W(A) + W(B)$. Per le matrici normali, il campo dei valori è l'involucro convesso dello spettro. Infatti, se $\lambda = x^*Ax$ e $x = \sum_{i=1}^n c_i x_i$, ove gli $\{x_i\}$ sono gli autovettori (di norma unitaria) di A, allora

$$\lambda = x^*Ax = \sum_{i,j=1}^n \overline{c}_i c_j x_i^* A x_j = \sum_{i,j=1}^n \overline{c}_i c_j \lambda_j x_i^* x_j = \sum_{j=1}^n |c_j|^2 \lambda_j$$

e $\sum_j |c_j|^2 = 1$ perché x di norma unitaria.

Teorema 2 (Dischi di Gerschgorin)**.** *Sia $A = (a_{ij})$ una matrice quadrata di dimensione n. Allora gli autovalori sono compresi nella regione*

$$\left(\bigcup_{i=1}^{n} R_i\right) \cap \left(\bigcup_{j=1}^{n} C_j\right)$$

ove

$$R_i = \left\{ z \in \mathbb{C} \colon |z - a_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \right\}, \quad C_j = \left\{ z \in \mathbb{C} \colon |z - a_{jj}| \leq \sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{ij}| \right\}$$

Dimostrazione. Dimostriamo che gli autovalori (destri) di A stanno in $\cup_i R_i$: seguirà che gli autovalori di A^T stanno in $\cup_j C_j$ e poiché i due insiemi di autovalori coincidono, staranno nell'intersezione. Sia λ un autovalore e v l'autovettore associato, normalizzato in modo che

$$\max_k |v_k| = |v_i| = 1$$

per un qualche $1 \leq i \leq n$. Allora

$$\sum_{j=1}^{n} a_{ij} v_j - \lambda v_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} u_j + a_{ii} v_i - \lambda v_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} v_j + (a_{ii} - \lambda) v_i = 0$$

Passando ai moduli

$$|a_{ii} - \lambda| = |\sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} v_j| \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}||v_j| \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|$$

e dunque $\lambda \in R_i$. □

Capitolo 2

Interpolazione polinomiale a tratti

Data una funzione $f\colon [a,b] \to \mathbb{R}$ e un'insieme $\{x_i\}_{i=1}^m \subset [a,b]$ di nodi ordinati ($x_{i-1} < x_i$), consideriamo l'interpolante polinomiale a tratti $p_{k-1}^c f$ di grado $k-1$. Su ogni intervallo $[x_i, x_{i+1}]$ di lunghezza $h_i = x_{i+1} - x_i$ essa è il polinomio di grado $k-1$

$$a_{i,1}(x-x_i)^{k-1} + a_{i,2}(x-x_i)^{k-2} + \ldots + a_{i,k-1}(x-x_i) + a_{i,k}\ . \tag{2.1}$$

Dunque, l'interpolante polinomiale a tratti è completamente nota una volta noti i nodi e i coefficienti di ogni polinomio.

In GNU Octave, l'interpolante polinomiale a tratti è definita mediante una struttura solitamente chiamata `pp` *(piecewise polynomial)* che si costruisce con il comando `mkpp(x,A)`, ove `x` è il vettore di nodi e `A` è la matrice, con riferimento a (2.1),

$$A_{ij} = a_{i,j}\ .$$

Nota una struttura `pp`, è possibile valutare il valore dell'interpolante in un generico target (o vettore di target) $\bar{x}$ con il comando `ppval(pp,xbar)`.

2.1 Interpolazione lineare a tratti

Dati i vettori $[x_1, \ldots, x_m]^{\mathrm{T}}$ e $[f_1, \ldots, f_m]^{\mathrm{T}}$, nell'intervallo $[x_i, x_{i+1}]$ l'interpolante lineare a tratti coincide con il polinomio di grado uno

$$\frac{f_{i+1} - f_i}{h_i}(x - x_i) + f_i$$

ove $h_i = x_{i+1} - x_i$. Pertanto, si costruisce la corrispondente struttura `pp` con il comando

```
pp = mkpp(x,[(f(2:m)-f(1:m-1))./h,f(1:m-1)])
```

ove `h` è il vettore delle distanze tra nodi consecutivi (e si può ottenere con il comando `diff(x)`).

2.1.1 Errore di interpolazione

Il risultato fondamentale sull'errore di interpolazione è

$$f(x) - p_{m-1}f(x) = \frac{f^{(m)}(\xi)}{n!}(x - x_1) \cdot (x - x_2) \cdot \ldots \cdot (x - x_m)$$

ove $p_{m-1}f$ è il polinomio di grado $m-1$ interpolatore di f sui nodi $\{x_i\}_{i=1}^m$ e ξ un opportuno punto nell'involucro convesso di $x \cup \{x_i\}_{i=1}^m$. Per l'interpolante lineare a tratti $p_1^c f$ su nodi equispaziati in un intervallo $[a, b]$, si ha dunque per $x \in [x_i, x_{i+1}]$

$$f(x) - p_1^c f(x) = \frac{f''(\xi)}{2}(x - x_i)(x - x_{i+1}) \tag{2.2}$$

e pertanto

$$|f(x) - p_1^c f(x)| \leq \max_{\xi \in [x_i, x_{i+1}]} \frac{f''(\xi)}{2} \frac{h^2}{4}, \quad x \in [x_i, x_{i+1}]$$

$((x - x_i)(x_{i+1} - x)$ è una parabola rivolta verso il basso di vertice $(x_i + h/2, h^2/4))$ da cui

$$\max_{[a,b]} |f(x) - p_1^c f(x)| = \|f - p_1^c f\|_\infty \leq \frac{h^2}{8}\|f''\|_\infty$$

Derivando rispetto a x l'equazione (2.2), si ottiene

$$f'(x) - (p_1^c f)'(x) = \frac{f''(\xi)}{2}[(x - x_{i+1}) + (x - x_i)]$$

e pertanto

$$|f'(x) - (p_1^c f)'(x)| \leq \max_{\xi \in [x_i, x_{i+1}]} \frac{f''(\xi)}{2} 2h, \quad x \in [x_i, x_{i+1}]$$

da cui

$$\|f' - (p_1^c f)'\|_\infty \leq h\|f''\|_\infty$$

Capitolo 3

Formule di quadratura gaussiana

Dato un intervallo (a, b) (eventualmente anche non limitato) e una funzione peso $w(x)$ non negativa su (a, b), si considera il prodotto scalare

$$(f, g) = \int_a^b f(x)g(x)w(x)\mathrm{d}x$$

con l'ipotesi

$$\int_a^b |x|^k w(x)\mathrm{d}x < \infty, \quad k \geq 0$$

Allora, esiste un'unica famiglia $\{p_j(x)\}_j$, $p_j(x)$ polinomio di grado j, *ortonormale* rispetto al prodotto scalare

$$\int_a^b p_j(x)p_i(x)w(x)\mathrm{d}x = \delta_{ij}$$

Gli zeri $\{x_n\}_{n=1}^m$ del polinomio $p_m(x)$ sono interni all'intervallo (a, b) e assieme ai pesi

$$w_n = \int_a^b L_n(x)w(x)\mathrm{d}x, \quad 1 \leq n \leq m$$

ove $L_n(x)$ è il polinomio di Lagrange di grado $m-1$ che vale 1 in x_n e zero in tutti gli altri nodi, costituiscono una formula di quadratura *gaussiana* esatta fino al grado polinomiale $2m-1$, cioè

$$\int_a^b q_j(x)w(x)\mathrm{d}x = \sum_{n=1}^m q_j(x_n)w_n, \quad 0 \leq j \leq 2m-1$$

ove $q_j(x)$ è un qualunque polinomio di grado $2m-1$. In particolare, per la famiglia $\{p_j\}_j$, vale

$$\delta_{ij} = \int_a^b p_j(x)p_i(x)w(x)\mathrm{d}x = \sum_{n=1}^{m} p_j(x_n)p_i(x_n)w_n, \quad 0 \le i,j \le m-1 \quad (3.1)$$

(perché $p_j(x)p_i(x)$ è un polinomio di grado $i+j \le 2m-2 < 2m-1$). Nel caso in cui (a,b) sia limitato, esiste un'unica formula di quadratura esatta fino al grado polinomiale $2m-3$ che usa come nodi $\bar{x}_1 = a$, $\bar{x}_m = b$ e gli zeri $\{\bar{x}_n\}_{n=2}^{m-1}$ del polinomio di grado $m-2$ della famiglia di polinomi ortogonali rispetto alla funzione peso $w(x)(x-a)(b-x)$. In questo caso si ha, in particolare,

$$\delta_{ij} = \int_a^b p_j(x)p_i(x)w(x)\mathrm{d}x = \sum_{n=1}^{m} p_j(\bar{x}_n)p_i(\bar{x}_n)\bar{w}_n, \quad \begin{matrix} 0 \le i \le m-3 \\ 0 \le j \le m-1 \end{matrix}$$

La famiglia $\{\phi_j(x)\}_{j=1}^m$, ove $\phi_j(x) = p_{j-1}(x)\sqrt{w(x)}$ è ovviamente ortonormale rispetto al prodotto scalare

$$(f,g) = \int_a^b f(x)g(x)\mathrm{d}x$$

e per essa valgono le osservazioni fatte sopra riguardo al calcolo degli integrali.

3.1 Quadratura gaussiana di Chebyshev e di Chebyshev–Lobatto

Per integrali del tipo

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\mathrm{d}x$$

i polinomi ortogonali da considerare sono quelli di Chebyshev

$$p_j(x) = T_j(x) = \cos(j \arccos(x))$$

che soddisfano la relazione di ricorrenza

$$\begin{aligned} T_0(x) &= 1, \quad T_1(x) = 1 \\ T_{j+1}(x) &= 2xT_j(x) - T_{j-1}(x), \quad j \ge 1 \end{aligned}$$

Gli zeri del polinomio di grado m soddisfano

$$m \arccos(x) = \frac{\pi}{2} + (n-1)\pi, \quad 1 \le n \le m$$

(gli angoli devono essere compresi tra 0 e $m\pi$) da cui

$$x_n = \cos\left(\frac{\frac{\pi}{2} + (n-1)\pi}{m}\right) = \sin\left(\frac{\pi}{2} - \frac{(2n-1)\pi}{2m}\right), \quad 1 \le n \le m$$

(la seconda formula produce nodi *anche numericamente* simmetrici) e i corrispondenti pesi di quadratura sono costanti e valgono

$$w_n = \frac{\pi}{m}, \quad 1 \le n \le m$$

I nodi di (Gauss–)Chebyshev–Lobatto sono invece

$$\bar{x}_n = \cos\left(\frac{(n-1)\pi}{m-1}\right) = \sin\left(\frac{\pi}{2} - \frac{(n-1)\pi}{m-1}\right), \quad 1 \le n \le m$$

e i corrispondenti pesi

$$\bar{w}_n = \begin{cases} \dfrac{\pi}{2(m-1)} & \text{per } n = 1 \text{ o } n = m \\ \dfrac{\pi}{m-1} & \text{per } 2 \le n \le m-1 \end{cases}$$

Capitolo 4

Metodi iterativi per sistemi di equazioni lineari

I metodi iterativi per la soluzione del sistema lineare

$$Ax = b \tag{4.1}$$

si basano sull'idea di calcolare la soluzione come limite di una successione di vettori

$$x = \lim_{l\to\infty} x^{(l)} .$$

Una strategia generale per costruire la successione $\{x^{(l)}\}_l$ è basata sullo splitting $A = P - M$, ove P è non singolare. Assegnato $x^{(1)}$, il termine $x^{(l+1)}$ è calcolato ricorsivamente come

$$Px^{(l+1)} = Mx^{(l)} + b, \quad l \geq 1 \tag{4.2}$$

Posto $e^{(l)} = x - x^{(l)}$, si ha

$$e^{(l)} = Be^{l-1}, \quad B = P^{-1}M = I - P^{-1}A ,$$

ove B è chiamata *matrice di iterazione*. È molto importante questo risultato

Proposizione 1. *Sia B una matrice quadrata. Sono equivalenti*

1. $\lim_{l\to\infty} B^l = 0$ *(ove 0 indica la matrice nulla)*
2. $\lim_{l\to\infty} \|B^l\| = 0$, *per una qualunque norma*
3. $\rho(B) < 1$, *ove* $\rho(B) = \max_\lambda |\lambda|$, $\lambda \in \sigma(B)$ *è il raggio spettrale di* B

Dimostrazione. Poiché la funzione norma $\|\cdot\|$ è continua, 1. implica 2. Per l'equivalenza delle norme

$$\|B^l\|_\infty \leq M\|B^l\|$$

e per definizione di $\|\cdot\|_\infty$, se $\|B^l\|$ tende a zero, B^l tende alla matrice nulla. Pertanto 2. implica 1. Per una qualunque norma naturale[1]

$$\|B\| \geq \rho(B)$$

(infatti $\|B\| = \max_{\|x\|=1}\|Bx\| \geq \|Bu\| = |\lambda|$ se λ e u con norma unitaria sono una coppia autovalore-autovettore) e quindi

$$\|B^l\| \geq \rho(B^l) = \rho(B)^l$$

e dunque 2. implica 3. Il raggio spettrale di una matrice gode della seguente proprietà (qui non dimostrata): per ogni ε esiste una norma naturale $\|\cdot\|$ tale che

$$\|B\| \leq \rho(B) + \varepsilon$$

Pertanto, se $\rho(B) < 1$ allora esiste ε e una norma naturale $\|\cdot\|$ tali che

$$\|B\| \leq \rho(B) + \varepsilon = \mu < 1$$

e dunque (per la disuguaglianza moltiplicativa valida per le norme naturali)

$$\|B^l\| \leq \|B\|^l \leq \mu^l$$

e quindi $\lim_{l\to\infty}\|B^l\| = 0$ e così per ogni altra norma, per l'equivalenza delle norme. Dunque, 3. implica 2. □

Da questa proposizione discende immediatamente la seguente

Proposizione 2. *Se $\rho(B) < 1$ (o se $\|B\| < 1$ per una qualunque norma $\|\cdot\|$), allora*

$$(I - B)^{-1} = \sum_{l=0}^{\infty} B^l$$

Dimostrazione. Basta considerare

$$(I-B)(I+B+B^2+\ldots+B^{L-1}) = I-B^L = (I+B+B^2+\ldots+B^{L-1})(I-B)$$

e passare al limite $L \to \infty$. □

Tornando ai metodi iterativi, si ha

$$e^{(l)} = B^l e^{(0)}$$

e dunque $e^{(l)}$ tende a zero se e solo se $\rho(B) < 1$.

[1] Indotta da una norma sui vettori

4.1 Metodi di Richardson

Indicato con $r^{(l)}$ il *residuo*

$$r^{(l)} = b - Ax^{(l)} = Ax - Ax^{(l)} = A(x - x^{(l)}) = Ae^{(l)} ,$$

il metodo (4.2) può essere riscritto come

$$P(x^{(l+1)} - x^{(l)}) = r^{(l)} . \tag{4.3}$$

In questo contesto, P viene chiamata *matrice di precondizionamento* o *precondizionatore* di A e viene scelta in modo che la matrice di iterazione $B = I - P^{-1}A$ abbia un raggio spettrale minore di 1 e la risoluzione di (4.3) sia "facile".

Una generalizzazione dello schema (4.3) è il *metodo di Richardson*: dato $x^{(1)}$, $x^{(l+1)}$ è calcolato ricorsivamente come

$$P(x^{(l+1)} - x^{(l)}) = \alpha r^{(l)} ,$$

ove α è un opportuno parametro di accelerazione. Dati $x^{(1)}$ e $r^{(1)} = b - Ax^{(1)}$, l'algoritmo per calcolare $x^{(l+1)}$ è

$$\begin{aligned} Pz^{(l)} &= r^{(l)} \\ x^{(l+1)} &= x^{(l)} + \alpha z^{(l)} \\ r^{(l+1)} &= r^{(l)} - \alpha A z^{(l)} \end{aligned} \tag{4.4}$$

Il costo di un'iterazione è dato essenzialmente dalla risoluzione di un sistema lineare $Pz^{(l)} = r^{(l)}$ facile e dal prodotto matrice-vettore $Az^{(l)}$. Tali metodi risulteranno particolarmente vantaggiosi per matrici *sparse*, in cui il numero di elementi diversi da zero è $\mathcal{O}(N)$ piuttosto che $\mathcal{O}(N^2)$ (e dunque il costo di un prodotto matrice-vettore è $\mathcal{O}(N)$), se l'ordine della matrice è N.

Il calcolo del residuo $r^{(l+1)} = r^{(l)} - \alpha A z^{(l)}$ (invece di $r^{(l+1)} = b - Ax^{(l+1)}$) permette di ridurre la propagazione, attraverso il prodotto matrice-vettore, degli errori, in quanto il vettore $z^{(l)}$, contrariamente a $x^{(l+1)}$, diminuisce in modulo al crescere di l.

4.1.1 Metodo del gradiente precondizionato

Siano A e P simmetriche e definite positive. Il metodo di Richardson può essere generalizzato con una scelta dinamica del parametro di accelerazione, prendendo $\alpha = \alpha_l$ in modo tale che

$$\|x - x^{(l+1)}\|_A, \quad \|y\|_A = \sqrt{y^{\mathrm{T}} A y}$$

sia minima. Si ha

$$\|x - x^{(l+1)}\|_A^2 = (x - x^{(l)} - \alpha_l z^{(l)})^{\mathrm{T}} A(x - x^{(l)} - \alpha_l z^{(l)}) =$$
$$= \alpha_l^2 {z^{(l)}}^{\mathrm{T}} A z^{(l)} - 2\alpha_l {z^{(l)}}^{\mathrm{T}} A(x - x^{(l)}) + (x - x^{(l)})^{\mathrm{T}} A(x - x^{(l)})$$

e dunque il minimo è dato dalla scelta

$$\alpha_l = \frac{{z^{(l)}}^{\mathrm{T}} r^{(l)}}{{z^{(l)}}^{\mathrm{T}} A z^{(l)}} .$$

Il metodo ottenuto si chiama *metodo del gradiente precondizionato*. Dati $x^{(1)}$ e $r^{(1)}$, l'algoritmo per calcolare $x^{(l+1)}$ è

$$\begin{aligned} Pz^{(l)} &= r^{(l)} \\ \alpha_l &= \frac{{z^{(l)}}^{\mathrm{T}} r^{(l)}}{{z^{(l)}}^{\mathrm{T}} A z^{(l)}} \\ x^{(l+1)} &= x^{(l)} + \alpha_l z^{(l)} \\ r^{(l+1)} &= r^{(l)} - \alpha_l A z^{(l)} \end{aligned} \tag{4.5}$$

Nel caso si scelga $P = I$, si ottiene il *metodo del gradiente* (noto anche come *steepest descent*).

4.1.2 Metodo del gradiente coniugato precondizionato

Siano A e P simmetriche e definite positive. Il *metodo del gradiente coniugato precondizionato* è una generalizzazione del metodo di Richardson in cui

$$x^{(l+1)} = x^{(l)} + \alpha_l p^{(l)}$$

ove i $\{p^{(l)}\}_l$ sono *coniugati*, cioè soddisfano

$$ {p^{(i)}}^{\mathrm{T}} A p^{(j)} = 0, \quad i \neq j$$

Per soddisfare questa proprietà è necessaria l'introduzione di un ulteriore parametro β_l. Dati $x^{(1)}$, $r^{(1)}$, $Pz^{(1)} = r^{(1)}$ e $p^{(1)} = z^{(1)}$, l'algoritmo per calcolare $x^{(l+1)}$ è

$$\begin{aligned} \alpha_l &= \frac{{z^{(l)}}^{\mathrm{T}} r^{(l)}}{{p^{(l)}}^{\mathrm{T}} A p^{(l)}} \\ x^{(l+1)} &= x^{(l)} + \alpha_l p^{(l)} \\ r^{(l+1)} &= r^{(l)} - \alpha_l A p^{(l)} \\ Pz^{(l+1)} &= r^{(l+1)} \\ \beta_{l+1} &= \frac{{z^{(l+1)}}^{\mathrm{T}} r^{(l+1)}}{{z^{(l)}}^{\mathrm{T}} r^{(l)}} \\ p^{(l+1)} &= z^{(l+1)} + \beta_{l+1} p^{(l)} \end{aligned} \tag{4.6}$$

Teorema 3. *Il metodo del gradiente coniugato applicato ad una matrice di ordine N converge in al più N iterazioni (in aritmetica esatta).*

Dimostrazione. La dimostrazione (omessa) si basa essenzialmente sul fatto che $p^{(1)}, \ldots, p^{(N)}$ sono vettori linearmente indipendenti e non ce ne possono essere più di N. □

Per questo motivo, tale metodo è detto *semiiterativo*.

Stima dell'errore

Vale la seguente stima dell'errore:

$$\|e^{(l)}\|_A \leq 2 \left(\frac{\sqrt{\text{cond}_2(P^{-1}A)} - 1}{\sqrt{\text{cond}_2(P^{-1}A)} + 1} \right)^{l-1} \|e^{(1)}\|_A$$

dalle quale si osserva che

- la stima d'errore decresce in ogni caso, poiché il numeratore è più piccolo del denominatore;
- in particolare, nel caso $P = I$;
- tanto più è piccolo il numero di condizionamento di $P^{-1}A$, tanto più il metodo ha convergenza veloce;
- nel caso limite di $P = A$, si ha $\|e^{(l)}\|_A \leq 0$.

4.1.3 Test d'arresto

Un primo stimatore è costituito dal residuo: si arresta cioè il metodo iterativo quando

$$\|r^{(l)}\| \leq \text{tol} \cdot \|b\|$$

Infatti, dalla precedente si ricava

$$\frac{\|e^{(l)}\|}{\|x\|} \leq \text{tol} \cdot \text{cond}(A)$$

Una modifica consiste in

$$\|r^{(l)}\| \leq \text{tol} \cdot \|r^{(1)}\| \tag{4.7}$$

che coincide con il precedente nel caso in cui come $x^{(1)}$ venga scelto il vettore di zeri.

Capitolo 5

Memorizzazione di matrici sparse

Sia A una matrice sparsa di ordine N con m elementi diversi da zero. Esistono molti formati di memorizzazione di matrici sparse. Quello usato da GNU Octave è il Compressed Column Storage (CCS). Consiste di tre array: un primo, `data`, di lunghezza m contenente gli elementi diversi da zero della matrice, ordinati prima per colonna e poi per riga; un secondo, `ridx`, di lunghezza m contenente gli indici di riga degli elementi di `data`; ed un terzo, `cidx`, di lunghezza $N + 1$, il cui elemento i-esimo ($i < N + 1$) è la posizione dentro `data` del primo elemento della colonna i e l'elemento $(N+1)$-esimo è il numero totale di elementi diversi da zero incrementato di uno. Per esempio, alla matrice

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

corrispondono i vettori

$$\begin{aligned} \texttt{data} &= [1, 4, 2, 3, 5, 6, 7] \\ \texttt{ridx} &= [1, 3, 2, 2, 3, 3, 4] \\ \texttt{cidx} &= [1, 3, 4, 6, 8] \end{aligned}$$

Talvolta, soprattutto in linguaggi di calcolo con array che iniziano dall'indice 0, gli array `ridx` e `cidx` hanno elementi decrementati di uno.

In GNU Octave, il formato CCS e l'implementazione del prodotto matrice-vettore sono automaticamente usati dalla function `sparse` e dall'operatore `*`, rispettivamente.

5.1 Alcuni comandi per matrici sparse

- Il comando `speye(N)` genera la matrice identità di ordine N.
- Il comando `spdiags(v,0,N,N)`, ove v è un vettore colonna, genera la matrice diagonale di ordine n avente v in diagonale. Se la dimensione di v è minore di n, la diagonale viene riempita con zeri posti dopo il vettore v. Se invece la dimensione di v è maggiore di N, vengono usate solo le prime N componenti di v.

  Sia V la matrice

$$V = \begin{pmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ \vdots & \vdots & \vdots \\ v_{N1} & v_{N2} & v_{N3} \end{pmatrix}$$

  Il comando `spdiags(V,-1:1,N,N)` genera la matrice

$$\begin{pmatrix} v_{12} & v_{23} & 0 & 0 & \dots & 0 \\ v_{11} & v_{22} & v_{33} & 0 & \dots & 0 \\ 0 & v_{21} & v_{32} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & v_{N-21} & v_{N-12} & v_{N3} \\ 0 & \dots & \dots & 0 & v_{N-11} & v_{N2} \end{pmatrix}$$

Capitolo 6

Sistemi tridiagonali

La risoluzione di sistemi tridiagonali

$$Ax = b$$

con

$$A = \begin{bmatrix} a_1 & c_1 & 0 & \dots & \dots & 0 \\ b_1 & a_2 & c_2 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & b_{n-2} & a_{n-1} & c_{n-1} \\ 0 & \dots & \dots & 0 & b_{n-1} & a_n \end{bmatrix}$$

risulta particolarmente economica. Infatti, nel caso non sia necessario il pivoting, si ha $A = LU$, ove

$$L = \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ \beta_1 & 1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \beta_{n-1} & 1 \end{bmatrix}, \quad U = \begin{bmatrix} \alpha_1 & c_1 & 0 & \dots & 0 \\ 0 & \alpha_2 & c_2 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \alpha_{n-1} & c_{n-1} \\ 0 & \dots & \dots & 0 & \alpha_n \end{bmatrix}$$

con

$$\begin{cases} \alpha_1 = a_1 \\ \beta_{k-1} = b_{k-1}/\alpha_{k-1}, \quad \alpha_k = a_k - \beta_{k-1}c_{k-1}, \quad k = 2, 3, \dots, n \end{cases}$$

e dunque la fattorizzazione LU costa $\mathcal{O}(2n)$ flops. A questo punto si risolvono i due sistemi $Ly = b$ e $Ux = y$, mediante

$$\begin{cases} y_1 = b_1 \\ y_k = b_k - \beta_{k-1}y_{k-1}, \quad k = 2, 3, \dots n \end{cases}$$

e

$$\begin{cases} x_n = y_n/\alpha_n \\ x_k = (y_k - c_k x_{k+1})/\alpha_k, \quad k = n-1, n-2, \dots 1 \end{cases}$$

con un ulteriore costo $\mathcal{O}(2n)$ flops. GNU Octave usa automaticamente questo algoritmo per le matrici tridiagonali.

Capitolo 7

Metodo di Newton per sistemi di equazioni non lineari

Consideriamo il sistema di equazioni non lineari

$$\begin{cases} f_1(x_1, x_2, \ldots, x_m) = 0 \\ f_2(x_1, x_2, \ldots, x_m) = 0 \\ \vdots \\ f_m(x_1, x_2, \ldots, x_m) = 0 \end{cases}$$

che può essere riscritto, in forma compatta,

$$f(x) = 0 \, .$$

Dato $x^{(1)}$, il metodo di Newton per calcolare $x^{(r+1)}$ è

$$\begin{aligned} J^{(r)} \delta x^{(r)} &= -f(x^{(r)}) \\ x^{(r+1)} &= x^{(r)} + \delta x^{(r)} \end{aligned} \tag{7.1}$$

ove $J^{(r)}$ è la matrice Jacobiana, definita da

$$J_{ij}^{(r)} = \frac{\partial f_i(x^{(r)})}{\partial x_j^{(r)}} \, . \tag{7.2}$$

Il criterio d'arresto solitamente usato è

$$\|\delta x^{(r)}\| \leq \text{tol}$$

In Matlab/Octave l'implementazione potrebbe essere:

```
f = @(x) ... % f
J = @(x) ... % jacobiano di f
x0 = ... % guess iniziale
x = x0;
errest = -J(x) \ f(x);
while (norm(errest,inf) > Newt_tol)
  x = x + errest;
  errest = -J(x) \ f(x);
end
```

7.1 Metodo di Newton inesatto

Il metodo di Newton (7.1) richiede il calcolo della matrice Jacobiana e la sua "inversione" ad ogni passo k. Questo potrebbe essere troppo oneroso ($\mathcal{O}(N^3)$ per un metodo diretto). Una strategia per ridurre il costo computazionale è usare sempre la stessa matrice Jacobiana $J^{(1)}$, oppure aggiornarla solo dopo un certo numero di iterazioni, oppure ancora usarne una sua approssimazione costante. In tal modo, per esempio, è possibile usare la stessa fattorizzazione $L^{(r)}U^{(r)}$ per più iterazioni successive e usarla per risolvere i sistemi lineari (di costo $\mathcal{O}(N^2)$).

7.2 Calcolo della matrice Jacobiana

Riportiamo qui le definizioni e le regole di base per il calcolo della matrice Jacobiana. Le esatte condizioni di esistenza si trovano su un qualunque libro di analisi.

Consideriamo $f\colon \mathbb{R}^m \to \mathbb{R}^m$. La sua derivata rispetto al vettore $v \in \mathbb{R}^m$ calcolata in $u \in \mathbb{R}^m$ è, se esiste,

$$\partial_v f(u) = \lim_{\varepsilon \to 0} \frac{f(u+\varepsilon v) - f(u)}{\varepsilon} \in \mathbb{R}^m$$

Una funzione f è differenziabile in u se esiste $T_u\colon \mathbb{R}^m \to \mathbb{R}^m$ lineare e continua tale che

$$\lim_{v \to u} \frac{\|f(v) - f(u) - T_u(v-u)\|}{\|v-u\|} = 0$$

Se f è differenziabile in u, allora $\partial_v f(u)$ esiste per ogni $v \in \mathbb{R}^m$ e si ha

$$\partial_v f(u) = T_u(v)$$

Essendo T_u un operatore lineare, scelta la base canonica per $\mathbb{R}^m$, esso può essere rappresentato dalla *matrice Jacobiana*

$$T_u = J_u = \begin{bmatrix} \partial_1 f_1(u) & \partial_2 f_1(u) & \dots & \partial_m f_1(u) \\ \partial_1 f_2(u) & \partial_2 f_2(u) & \dots & \partial_m f_2(u) \\ \vdots & \vdots & & \vdots \\ \partial_1 f_m(u) & \partial_2 f_m(u) & \dots & \partial_m f_m(u) \end{bmatrix}$$

ove $\partial_i f_j(u) = \partial_{e_i} f_j(u)$.

Indichiamo ora con $f'(u)$ il differenziale di f calcolato in u. Valgono le usuali regole di derivazione: (da scrivee usando la matrice jacobiana)

- $(\alpha f + \beta g)'(u) = \alpha f'(u) + \beta g'(u)$
- $g \circ f)'(u) = g'(f(u)) \circ f'(u)$
- $(f(u).g(u))' =$

Capitolo 8

Esponenziale di matrice

Data una matrice quadrata $A \in \mathbb{R}^{N\times N}$, si definisce

$$\exp(A) = \sum_{j=0}^{\infty} \frac{A^j}{j!}$$

Tale serie converge per qualunque matrice A, essendo A un operatore lineare tra spazi di Banach e avendo la serie esponenziale raggio di convergenza ∞. Se A e B sono *permutabili* (cioè $AB = BA$), allora

$$\exp(A + B) = \exp(A)\exp(B)$$

8.1 Formula delle *variazioni delle costanti*

Data l'equazione differenziale

$$\begin{cases} y'(t) = ay(t) + b(t, y(t)), \quad t > 0 \\ y(t_0) = y_0 \end{cases} \tag{8.1}$$

$y(t) \in \mathbb{R}$, la soluzione può essere scritta analiticamente mediante la formula delle *variazioni delle costanti*

$$y(t) = e^{(t-t_0)a}y_0 + \int_{t_0}^{t} e^{(t-\tau)a}b(\tau, y(\tau))\mathrm{d}\tau \tag{8.2}$$

Infatti, si ha

$$y'(t) = ae^{(t-t_0)a}y_0 + a\int_{t_0}^{t} e^{(t-\tau)a}b(\tau, y(\tau))\mathrm{d}\tau + e^{(t-t)a}b(t, y(t)) = ay(t) + b(t, y(t))$$

Si osservi che

$$\int_{t_0}^{t} e^{(t-\tau)a}\mathrm{d}\tau = -\frac{1}{a}\int_{t_0}^{t} -ae^{(t-\tau)a}\mathrm{d}\tau = -\frac{1}{a}e^{(t-\tau)a}\Big|_{t_0}^{t} =$$
$$= -\frac{1}{a}\left(1-e^{(t-t_0)a}\right) = (t-t_0)\frac{e^{(t-t_0)a}-1}{(t-t_0)a} =$$
$$= (t-t_0)\varphi_1((t-t_0)a)\ ,$$

ove

$$\varphi_1(z) = \frac{e^z-1}{z} = \sum_{j=0}^{\infty}\frac{z^j}{(j+1)!} \tag{8.3}$$

e, analogamente,

$$\int_{t_0}^{t} e^{(t-\tau)a}(\tau-t_0)\mathrm{d}\tau = (t-t_0)^2\varphi_2((t-t_0)a)$$

ove

$$\varphi_2(z) = \frac{e^z-1-z}{z^2} = \sum_{j=0}^{\infty}\frac{z^j}{(j+2)!} \tag{8.4}$$

Consideriamo ora un sistema differenziale

$$\begin{cases} \boldsymbol{y}'(t) = A\boldsymbol{y}(t) + \boldsymbol{b}(t,\boldsymbol{y}(t)), \quad t>0 \\ \boldsymbol{y}(t_0) = \boldsymbol{y}_0 \end{cases}$$

Ancora, la soluzione esplicita può essere scritta come

$$\boldsymbol{y}(t) = \exp((t-t_0)A)\boldsymbol{y}_0 + \int_{t_0}^{t}\exp((t-\tau)A)\boldsymbol{b}(\tau,\boldsymbol{y}(\tau))\mathrm{d}\tau$$

8.2 Calcolo di $\exp(A)$

Come per la risoluzione di sistemi lineari, non esiste *il* modo per calcolare $\exp(A)$, ma diversi modi, ognuno adatto a particolari situazioni.

8.2.1 Matrici piene, di modeste dimensioni

Questi metodi si applicano, in pratica, a quelle matrici per le quali si usano i metodi diretti per la risoluzione di sistemi lineari.

Decomposizione spettrale Se la matrice è diagonalizzabile, cioè $A = VDV^{-1}$, allora $\exp(A) = V\exp(D)V^{-1}$, ove $\exp(D)$ è la matrice diagonale con elementi $e^{d_1}, e^{d_2}, \ldots, e^{d_N}$. Basta infatti osservare che

$$A^2 = (VDV^{-1})^2 = (VDV^{-1})(VDV^{-1}) = VD^2V^{-1}$$

e scrivere $\exp(A)$ come serie di Taylor. La decomposizione spettrale di una matrice costa, in generale, $\mathcal{O}(N^3)$. Si ottiene in GNU Octave con il comando `eig`.

Approssimazione razionale di Padé Si considera un'approssimazione razionale della funzione esponenziale

$$e^z \approx \frac{a_1 z^{p-1} + a_2 z^{p-2} + \ldots + a_p}{b_1 z^{q-1} + b_2 z^{q-2} + \ldots + b_q} , \tag{8.5}$$

ove $b_q = 1$ per convenzione. Essa è chiamata *diagonale* quando $p = q$. Si può dimostrare che le approssimazioni diagonali sono le più efficienti. Fissato il grado di approssimazione, si sviluppa in serie di Taylor la funzione esponenziale e si fanno coincidere quanti più coefficienti possibile. Per esempio, fissiamo $p = q = 2$. Si ha allora

$$\left(1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \ldots\right)(b_1 z + 1) \approx a_1 z + a_2$$

$$b_1 z + 1 + b_1 z^2 + z + \frac{z^2}{2} + o(z^2) \approx a_1 z + a_2$$

da cui

$$\begin{cases} 1 = a_2 \\ b_1 + 1 = a_1 \\ b_1 + \dfrac{1}{2} = 0 \end{cases}$$

L'approssimazione di Padé si estende banalmente al caso matriciale. Considerando sempre il caso $p = q = 2$, si ha

$$\exp(A) \approx B = (b_1 A + I)^{-1}(a_1 A + a_2 I) ,$$

cioè B è soluzione del sistema lineare $(b_1 A + I)B = a_1 A + a_2 I$. In questo caso, l'approssimazione della soluzione di un problema differenziale $\boldsymbol{y}'(t) = A\boldsymbol{y}(t)$ diventa

$$\boldsymbol{y}(t) = \exp(tA)\boldsymbol{y}_0 \approx (I - tA/2)^{-1}(I + tA/2)\boldsymbol{y}_0$$

e dunque coincide con la risoluzione del problema mediante il metodo dei trapezi (vedi § 17.2).

L'approssimazione di Padé è accurata solo quando $|z| < 1/2$ (o, nel caso matriciale, $\|A\|_2 < 1/2$). Per la funzione esponenziale esiste una tecnica, chiamata *scaling and squaring* che permette di aggirare il problema. Si usa infatti la proprietà

$$e^z = \left(e^{z/2}\right)^2 = \left(e^{z/2^j}\right)^{2^j}$$

Se $|z| > 1/2$, allora $|z|/2^j < 1/2$ per $j > \log_2(|z|) + 1$. Si calcola dunque l'approssimazione di Padé di $e^{z/2^j}$ e poi si eleva al quadrato j volte. Per la funzione φ_1 vale

$$\varphi_1(z) = \frac{1}{2}(e^{z/2} + 1)\varphi_1\left(\frac{z}{2}\right)$$

Anche l'approssimazione di Padé matriciale ha costo $\mathcal{O}(N^3)$. In GNU Octave si usa una variante di questa tecnica nel comando `expm`.

8.2.2 Matrici sparse, di grandi dimensioni

I metodi visti nel paragrafo precedente ignorano l'eventuale sparsità delle matrici. Inoltre, negli integratori esponenziali, non è mai richiesto di calcolare esplicitamente funzioni di matrice, ma solo funzioni di matrice applicate a vettori, cioè $\exp(A)v$ (è l'analoga differenza tra calcolare A^{-1} e $A^{-1}v$). Si possono allora usare dei metodi *iterativi*.

Metodo di Krylov Mediante la *tecnica di Arnoldi* è possibile, tramite prodotti matrice-vettore, decomporre A in $V_m^{\mathrm{T}} A V_m = H_m$, ove $V_m \in \mathbb{R}^{n\times m}$, $V_m^{\mathrm{T}} V_m = I_n$, $V_m e_1 = v$ e H_m è matrice di Hessenberg di ordine m (con $m \ll n$). Allora $AV_m \approx V_m H_m$ e quindi

$$\exp(A)V_m \approx V_m \exp(H_m) \Rightarrow \exp(A)v \approx V_m \exp(H_m)e_1$$

Il calcolo di $\exp(H_m)$ è fatto mediante l'approssimazione di Padé. Il costo della tecnica di Arnoldi è $\mathcal{O}(nm^2)$ se A è matrice sparsa. È necessario inoltre memorizzare la matrice V_m.

Interpolazione su nodi di Leja Se il polinomio $p_m(z)$ interpola e^z nei nodi $\xi_0, \xi_1, \ldots, \xi_m$, allora $p_m(A)v$ è una approssimazione di $\exp(A)v$. È una *buona* approssimazione se i nodi sono buoni (*non* equispaziati, per esempio) e se sono contenuti nell'involucro convesso dello spettro di A. È difficile stimare a priori il grado di interpolazione m necessario. È conveniente usare la formula di interpolazione di Newton

$$p_{m-1}(z) = d_1 + d_2(z-\xi_1) + d_3(z-\xi_1)(z-\xi_2) + \cdot\ldots\cdot + d_m(z-\xi_1)\cdots(z-\xi_{m-1})$$

ove $\{d_i\}_i$ sono le differenze divise. Tale formula si può scrivere, nel caso matriciale,

$$p_{m-1}(A)v = p_{m-2}v + d_m w_m, \quad w_m = \left(\prod_{i=1}^{m-1}(A - \xi_i I)\right) v = (A - \xi_{m-1})w_{m-1}$$

Dunque, la complessità è $\mathcal{O}(Nm)$ è richiesta la memorizzazione di un solo vettore w.

Quali nodi usare? I nodi di Chebyshev, molto buoni per l'interpolazione, non possono essere usati, in quanto non permettono un uso efficiente della formula di interpolazione di Newton (cambiano tutti al cambiare del grado). I *nodi di Leja* sono distribuiti asintoticamente come i nodi di Chebyshev e, dati i primi $m-1$, ξ_m è il nodo per cui

$$\prod_{i=1}^{m-1} |\xi_m - \xi_i| = \max_{\xi \in [a,b]} \prod_{i=1}^{m-1} |\xi - \xi_i| \ ,$$

ove l'intervallo $[a, b]$ è in relazione con lo spettro di A, per esempio $[a, b] = \sigma(A) \cap \{y = 0\}$. Il primo nodo coincide, solitamente, con l'estremo dell'intervallo $[a, b]$ di modulo massimo. È chiaro che l'insieme dei primi m nodi di Leja coincide con l'unione di $\{\xi_m\}$ con l'insieme dei primi $m-1$ nodi di Leja.

Capitolo 9

Esercizi

1. Implemetare le functions `[data,ridx,cidx] = full2ccs(A)` e `[A] = ccs2full(data,ridx,cidx)` e le functions che, dati `data`, `ridx` e `cidx`, implementano i prodotti matrice vettore Ax e $A^{\mathrm{T}}x$.

2. Si risolvano 6 sistemi lineari con le matrici di Hilbert di ordine $N = 4, 6, 8, 10, 12, 14$ (`hilb(N)`) e termine noto scelto in modo che la soluzione esatta sia il vettore $[1, 1, \ldots, 1]^{\mathrm{T}}$ usando il comando \ di GNU Octave, il metodo del gradiente precondizionato e il metodo del gradiente coniugato precondizionato. Per questi ultimi due, si usi una tolleranza pari a 10^{-6}, un numero massimo di iterazioni pari a 2000, il precondizionatore diagonale e un vettore iniziale $x^{(1)}$ di zeri. Si riporti, per ogni N, il numero di condizionamento della matrice, l'errore in norma infinito rispetto alla soluzione esatta e il numero di iterazioni dei metodi iterativi.

3. Risolvere il sistema non lineare

$$\begin{cases} f_1(x_1, x_2) = x_1^2 + x_2^2 - 1 = 0 \\ f_2(x_1, x_2) = \sin(\pi x_1/2) + x_2^3 = 0 \end{cases}$$

   con il metodo di Newton (7.1). Si usi una tolleranza pari a 10^{-6}, un numero massimo di iterazioni pari a 150 e un vettore iniziale $x^{(1)} = [1, 1]^{\mathrm{T}}$. Si risolva lo stesso sistema non lineare usando sempre la matrice Jacobiana relativa al primo passo e aggiornando la matrice Jacobiana ogni r iterazioni, ove r è il più piccolo numero di iterazioni che permette di ottenere la stessa soluzione con la tolleranza richiesta calcolando solo due volte la matrice Jacobiana.

4. Si implementi una function `[a,b] = padeexp(p)` che restituisce i coefficienti dell'approssimazione razionale di Padé (8.5) (con $p = q$) per la funzione esponenziale.

Parte 1

BVPs
(Problemi ai limiti)

Capitolo 10

Introduzione

Consideriamo il seguente *problema ai limiti (boundary value problem)*

$$\begin{cases} u''(x) = f(x, u(x), u'(x)), & x \in (a, b) \\ u(a) = u_a \\ u(b) = u_b \end{cases} \tag{10.1}$$

ove $u(x) \in \mathbb{R}$. Le condizioni ai bordi sono di *Dirichlet* quando viene prescritto il valore della soluzione $u(x)$ e di *Neumann* quando viene prescritto il valore della derivata della soluzione $u'(x)$. Si possono avere anche condizioni *miste*, ad esempio

$$\begin{cases} u''(x) = f(x, u(x), u'(x)), & x \in (a, b) \\ u(a) = u_a \\ u'(b) = u'_b \end{cases}$$

Con un'unica notazione si può scrivere

$$\begin{cases} u''(x) = f(x, u(x), u'(x)), & x \in (a, b) \\ \alpha_a u(a) + \beta_a u'(a) = \gamma_a \\ \alpha_b u(b) + \beta_b u'(b) = \gamma_b \end{cases}$$

Se $\alpha_a \neq 0$ e $\beta_a \neq 0$ (oppure $\alpha_b \neq 0$ e $\beta_b \neq 0$) si parla di condizioni di *Robin*. Quando i valori prescritti ai bordi sono nulli, si parla di condizioni *omogenee*.

Capitolo 11

Differenze finite

11.1 Differenze finite centrate del secondo ordine

Sia $u \in \mathcal{C}^3([a,b])$ e $x_i = a + (i-1)h$, $1 \leq i \leq m$, $h = (b-a)/(m-1)$. Sviluppando in serie di Taylor (resto di Lagrange), si ha

$$u(x_{i+1}) = u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{6}u^{(3)}(\hat{x}_i)$$
$$u(x_{i-1}) = u(x_i) - hu'(x_i) + \frac{h^2}{2}u''(x_i) - \frac{h^3}{6}u^{(3)}(\tilde{x}_i)$$

da cui

$$u'(x_i) = \Delta u(x_i) - \tau_i^{(1)} = \frac{u(x_{i+1}) - u(x_{i-1})}{2h} - \tau_i^{(1)}$$

ove $\tau_i^{(1)} = \frac{h^2}{6}u^{(3)}(\bar{x}_i)$ è l'*errore locale* ($u^{(3)}(\hat{x}_i) + u^{(3)}(\tilde{x}_i) = 2u^{(3)}(\bar{x}_i)$, per un opportuno $\bar{x}_i$, per il teorema dei valori intermedi). Analogamente, sia $u \in \mathcal{C}^4([a,b])$. Si ha

$$u(x_{i+1}) = u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{6}u^{(3)}(x_i) + \frac{h^4}{24}u^{(4)}(\hat{x}_i)$$
$$u(x_{i-1}) = u(x_i) - hu'(x_i) + \frac{h^2}{2}u''(x_i) - \frac{h^3}{6}u^{(3)}(x_i) + \frac{h^4}{24}u^{(4)}(\tilde{x}_i)$$

da cui

$$u''(x_i) = \Delta^2 u(x_i) - \tau_i^{(2)} = \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} - \tau_i^{(2)} \qquad (11.1)$$

ove $\tau_i^{(2)} = \frac{h^2}{12}u^{(4)}(\bar{x}_i)$. Queste approssimazioni della derivata prima e seconda di chiamano *differenze finite centrate del secondo ordine*. Il termine "centrate" si riferisce al fatto che i punti x_i sono equispaziati e si usano i valori

della funzione $u(x)$ in uno stesso numero di punti a sinistra e a destra di x_i per ricavare un'approssimazione delle derivate. Il termine "secondo ordine" si riferisce al fatto che l'errore locale è proporzionale alla seconda potenza del *passo di discretizzazione* h. Ovviamente sono possibili altri tipi di approssimazione, basati su nodi non equispaziati, non centrate e di ordine diverso (vedi § 11.3).

`diff12.m`

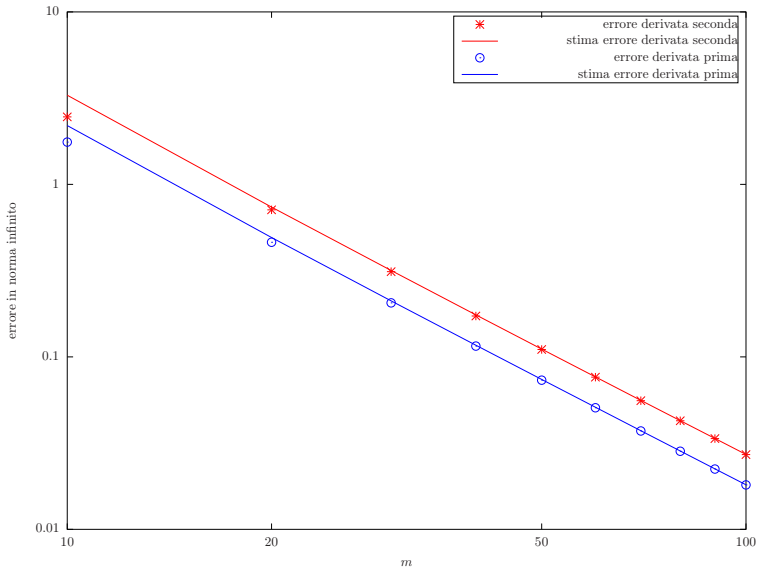
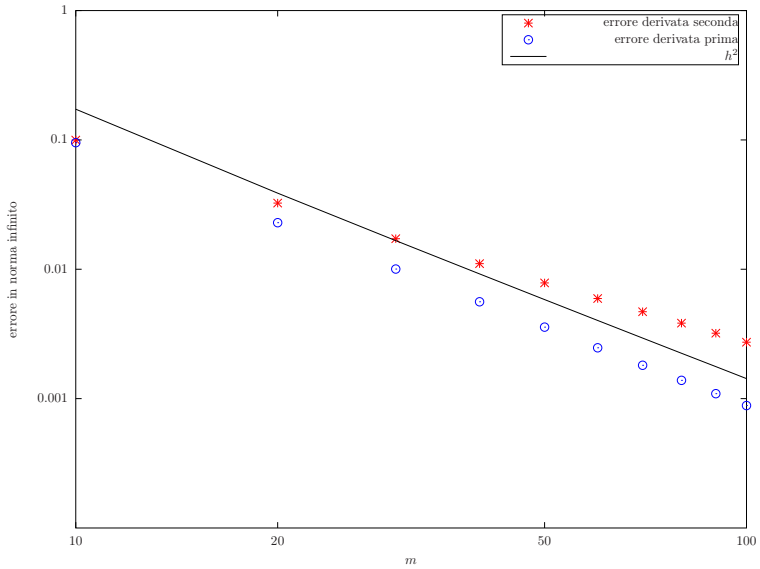


Figura 11.1: Errori nell'approssimazione delle derivate prima e seconda per la funzione $u(x) = \sin(3x)$.

In Figura 11.1 si vedono gli errori (in norma infinito) tra la derivata prima e seconda della funzione $u(x) = \sin(3x)$ e la relativa approssimazione mediante differenze finite centrate del secondo ordine (asterischi) e le stime $h^2/6 \cdot \|u^{(3)}\|_\infty$ e $h^2/12 \cdot \|u^{(4)}\|_\infty$ (linea continua), rispettivamente, ove $h = 2\pi/(m-1)$. In Figura 11.2 si vede invece che per la funzione $u(x) = |x|^{7/2}$, l'approssimazione della derivata prima mediante differenze finite centrate ha effettivamente ordine due, mentre quella della derivata seconda no, in quanto non esiste la derivata quarta di $u(x)$ ($h = 2/(m-1)$).

Una volta scelto il tipo di discretizzazione, invece del problema originale (10.1) si risolve il problema discretizzato

$$\begin{cases} \Delta^2 u_i = f(x_i, u_i, \Delta u_i), \quad 2 \le i \le m-1 \\ u_1 = u_a \\ u_m = u_b \end{cases}$$



diff12ns.m

Figura 11.2: Errori nell'approssimazione delle derivate prima e seconda per la funzione $u(x) = |x|^{7/2}$.

nell'incognita $\boldsymbol{u} = [u_1, u_2, \ldots, u_{m-1}, u_m]^{\mathrm{T}}$, ove

$$\Delta u_i = \frac{u_{i+1} - u_{i-1}}{2h}$$
$$\Delta^2 u_i = \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}$$

Si tratta dunque di risolvere un sistema di m equazioni (in generale) non lineari nelle incognite u_i, $1 \leq i \leq m$.

In forma matriciale,

$$\begin{bmatrix} \Delta u_1 \\ \Delta u_2 \\ \Delta u_3 \\ \vdots \\ \Delta u_{m-1} \\ \Delta u_m \end{bmatrix} = \frac{1}{2h} \begin{bmatrix} * & * & * & * & * & * \\ -1 & 0 & 1 & 0 & \dots & 0 \\ 0 & -1 & 0 & 1 & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -1 & 0 & 1 \\ * & * & * & * & * & * \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_{m-1} \\ u_m \end{bmatrix}$$

$$\begin{bmatrix} \Delta^2 u_1 \\ \Delta^2 u_2 \\ \Delta^2 u_3 \\ \vdots \\ \Delta^2 u_{m-1} \\ \Delta^2 u_m \end{bmatrix} = \frac{1}{h^2} \begin{bmatrix} * & * & * & * & * & * \\ 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 & -2 & 1 \\ * & * & * & * & * & * \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_{m-1} \\ u_m \end{bmatrix}$$

ove la prima e l'ultima riga devono essere trattate a parte, solitamente per includere le condizioni al bordo. Le matrici relative alle approssimazione della derivata prima e seconda possono essere costruite con i comandi

```
> toeplitz(sparse(1,2,-1/(2*h),1,m),sparse(1,2,1/(2*h),1,m));
```

e

```
> toeplitz(sparse([1,1],[1,2],[-2/h^2,1/h^2],1,m));
```

rispettivamente.

11.2 Convergenza per un problema modello

Consideriamo il seguente problema modello *(elasticità della trave)*

$$\begin{cases} -u''(x) + q(x)u(x) = g(x), \quad x \in (a,b) \\ u(a) = u_a \\ u(b) = u_b \end{cases} \tag{11.2}$$

con $q, g \in \mathcal{C}^0([a,b])$, $q(x) \geq 0$ per $x \in [a,b]$. La funzione $q(x)$ è definita dal rapporto $P/(E \cdot I(x))$ (P la tensione orizzontale, E il modulo di Young e $I(x)$ il momento principale di inerzia) e $g(x)$ è la densità di carico verticale (tipicamente positiva). La soluzione $u(x)$ rappresenta il momento flettente. Vogliamo studiare l'esistenza, l'unicità e la regolarità della soluzione analitica.

11.2.1 Unicità

Se $u_1(x)$ e $u_2(x)$ sono due soluzioni di (11.2), allora $z(x) = u_1(x) - u_2(x)$ soddisfa il problema *omogeneo*

$$\begin{cases} -z''(x) + q(x)z(x) = 0, \quad x \in (a,b) \\ z(a) = 0 \\ z(b) = 0 \end{cases} \tag{11.3}$$

Proposizione 3. *Se $z(x)$ è soluzione di (11.3), allora $z(x) \equiv 0$.*

Dimostrazione (metodo dell'energia). $z(x) \equiv 0$ è certamente una soluzione. Supponiamo, per assurdo, che esista anche $z(x) \neq 0$ soluzione. Moltiplicando l'equazione per $z(x)$ ed integrando si ha

$$\begin{aligned} 0 &= \int_a^b -z''(x)z(x)\mathrm{d}x + \int_a^b q(x)z(x)^2\mathrm{d}x = \\ &= [-z'(x)z(x)]_a^b + \int_a^b z'(x)^2\mathrm{d}x + \int_a^b q(x)z(x)^2\mathrm{d}x = \\ &= \int_a^b z'(x)^2\mathrm{d}x + \int_a^b q(x)z(x)^2\mathrm{d}x \end{aligned}$$

Poiché le funzioni integrande sono non negative, si ha che deve essere necessariamente $z(x) \equiv 0$, quindi assurdo. □

Dunque, $u_1(x) \equiv u_2(x)$.

11.2.2 Esistenza

Sia $z(x) = c_1 z_1(x) + c_2 z_2(x)$ la soluzione generale di $-z''(x) + q(x)z(x) = 0$, con $z_1(x)$ e $z_2(x)$ indipendenti (lo spazio delle soluzioni dell'equazione lineare omogenea ha proprio dimensione due). La soluzione di (11.3) (che corrisponde a $c_1 = c_2 = 0$) si ottiene imponendo

$$\begin{cases} c_1 z_1(a) + c_2 z_2(a) = 0 \\ c_1 z_1(b) + c_2 z_2(b) = 0 \end{cases}$$

Poiché sappiamo che $z(x) \equiv 0$ è l'unica soluzione, si ha che la matrice

$$\begin{bmatrix} z_1(a) & z_2(a) \\ z_1(b) & z_2(b) \end{bmatrix}$$

è non singolare.

La soluzione generale di $-u''(x) + q(x)u(x) = g(x)$ è $u(x) = c_1 z_1(x) + c_2 z_2(x) + s(x)$ ($s(x)$ soluzione particolare che si ottiene dalla tecnica delle variazioni delle costanti, cioè supponendo $s(x) = c_1(x)z_1(x) + c_2(x)z_2(x)$, $c_1(x)$ e $c_2(x)$ da ricavare). La soluzione di (11.2) si ottiene imponendo le condizioni al bordo

$$\begin{cases} c_1 z_1(a) + c_2 z_2(a) = u_a - s(a) \\ c_1 z_1(b) + c_2 z_2(b) = u_b - s(b) \end{cases}$$

cioè risolvendo un sistema lineare non singolare che ammette dunque (unica) soluzione.

11.2.3 Regolarità

Proposizione 4. *Se $q, g \in \mathcal{C}^k([a,b])$, allora $u \in \mathcal{C}^{k+2}([a,b])$.*

Dimostrazione. Se $q, g \in \mathcal{C}^0([a,b])$, poiché la soluzione u esiste, u'' è definita in ogni punto $x \in [a,b]$, e dunque u' esiste (ed è derivabile). Quindi $u \in \mathcal{C}^0([a,b])$ e quindi $u'' \in \mathcal{C}^0([a,b])$. Dunque $u \in \mathcal{C}^2([a,b])$. Sia vero adesso l'enunciato per k e siano $q, g \in \mathcal{C}^{k+1}([a,b])$: poiché anche $u \in \mathcal{C}^{k+1}([a,b])$, si ha $u'' \in \mathcal{C}^{k+1}([a,b])$ da cui $u \in \mathcal{C}^{k+3}([a,b])$. □

Si è costretti a ridursi ad un problema modello perché problemi ai limiti anche molto semplici possono non avere soluzione: si consideri, per esempio,

$$\begin{cases} u''(x) + u(x) = 0, \quad x \in (0,\pi) \\ u(0) = 0 \\ u(\pi) = 1 \end{cases}$$

La soluzione generale è $c_1 \cos(x) + c_2 \sin(x)$, ma non è possibile imporre le condizioni al bordo.

Ci occupiamo adesso di analizzare la convergenza del problema modello discretizzato mediante differenze finite centrate del secondo ordine, che si scrive

$$\begin{cases} -\dfrac{u_{i+1} - 2u_i + u_{i-1}}{h^2} + q_i u_i = g_i, \quad 2 \le i \le m-1 \\ u_1 = u_a \\ u_m = u_b \end{cases}$$

ove $q_i = q(x_i)$ e $g_i = g(x_i)$.

11.2.4 Esistenza ed unicità per il problema discretizzato

Il sistema lineare da risolvere per trovare $\boldsymbol{u} = [u_1, u_2, \ldots, u_{m-1}, u_m]^{\mathrm{T}}$ è

$$\frac{1}{h^2}\begin{bmatrix} 1 & 0 & \ldots & \ldots & \ldots & 0 \\ -1 & 2+q_2h^2 & -1 & 0 & \ldots & 0 \\ 0 & -1 & 2+q_3h^2 & -1 & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & -1 & 2+q_{m-1}h^2 & -1 \\ 0 & \ldots & \ldots & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{m-1} \\ u_m \end{bmatrix} = \begin{bmatrix} u_a/h^2 \\ g_2 \\ g_3 \\ \vdots \\ g_{m-1} \\ u_b/h^2 \end{bmatrix}$$

e può essere semplificato in

$$\frac{1}{h^2}\begin{bmatrix} 2+q_2h^2 & -1 & 0 & \ldots & \ldots & 0 \\ -1 & 2+q_3h^2 & -1 & 0 & \ldots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ldots & 0 & -1 & 2+q_{m-2}h^2 & -1 \\ 0 & \ldots & \ldots & 0 & -1 & 2+q_{m-1}h^2 \end{bmatrix} \begin{bmatrix} u_2 \\ u_3 \\ \vdots \\ \vdots \\ u_{m-2} \\ u_{m-1} \end{bmatrix} = \begin{bmatrix} g_2 + u_a/h^2 \\ g_3 \\ \vdots \\ \vdots \\ g_{m-2} \\ g_{m-1} + u_b/h^2 \end{bmatrix}$$

cioè

$$A\boldsymbol{u} = \boldsymbol{g} \tag{11.4}$$

ove adesso $\boldsymbol{u} = [u_2, \ldots, u_{m-1}]^{\mathrm{T}}$ (parliamo in questo caso di discretizzazione *a nodi interni*).

Proposizione 5. *Il sistema lineare* (11.4) *è non singolare e dunque ammette un'unica soluzione.*

Dimostrazione (metodo dell'energia discreto). Dato $\boldsymbol{z} = [z_2, z_3, \ldots, z_{m-1}]^{\mathrm{T}}$, consideriamo il prodotto $\boldsymbol{z}^{\mathrm{T}}A\boldsymbol{z}$. Si ha

$$\begin{aligned} z^{\mathrm{T}}Az = \frac{1}{h^2}[&(2+q_2h^2)z_2^2 - z_2z_3 - z_3z_2 + (2+q_3h^2)z_3^2 - z_3z_4 + \ldots + \\ &+ \ldots - z_{m-1}z_{m-2} + (2+q_{m-1}h^2)z_{m-1}^2] = \\ &= \frac{1}{h^2}[z_2^2 + (z_2-z_3)^2 + (z_3-z_4)^2 + \ldots + (z_{m-2}-z_{m-1})^2 + z_{m-1}^2] + \\ &+ \sum_{i=2}^{m-1} q_i z_i^2 \geq 0 \end{aligned}$$

Poiché si ha una somma di elementi non negativi, l'uguaglianza a 0 si può avere solo quando tutti gli elementi sono nulli e quindi per solo per $\boldsymbol{z}$ nullo. Dunque la matrice A è definita positiva e quindi non singolare. □

11.2.5 Proprietà di A

A è una matrice simmetrica, definita positiva e diagonalmente dominante. È possibile usare i metodi iterativi, semi-iterativi e diretti *senza* pivoting per la soluzione del sistema lineare. Inoltre, è una M-matrice, cioè i suoi elementi extra-diagonali sono non positivi e la sua inversa ha elementi non negativi (vedi § A.1).

11.2.6 Consistenza

Se si sostituisce u_i con la soluzione analitica $u(x_i)$, da (11.1) si ottiene

$$\begin{cases} -\dfrac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} + q(x_i)u(x_i) - g(x_i) = -\tau_i^{(2)},\ 2 \le i \le m-1 \\ u(x_1) = u_a \\ u(x_m) = u_b \end{cases}$$

da cui si deduce che il metodo numerico è *consistente* di ordine 2.

Definiamo l'errore $\boldsymbol{e}_h = [e_{2,h}, \ldots, e_{m-1,h}]^{\mathrm{T}} = [u_2 - u(x_2), \ldots, u_{m-1} - u(x_{m-1})]^{\mathrm{T}}$, $h = (b-a)/(m-1)$. Poiché

$$A[u_2, \ldots, u_{m-1}]^{\mathrm{T}} = \boldsymbol{g}$$
$$A[u(x_2), \ldots, u(x_{m-1})]^{\mathrm{T}} = \boldsymbol{g} - \boldsymbol{\tau}_h^{(2)}$$

ove $\boldsymbol{\tau}_h^{(2)} = [\tau_{2,h}^{(2)}, \ldots, \tau_{m-1,h}^{(2)}]^{\mathrm{T}}$, si deduce $\boldsymbol{e}_h = A^{-1}\boldsymbol{\tau}_h$. Si può concludere che l'errore tende a zero come $\boldsymbol{\tau}_h$ quando $h \to$? No, perché non abbiamo informazioni sul comportamento di $A_h^{-1} = A^{-1}$ per $h \to 0$.

11.2.7 Stabilità

Consideriamo due soluzioni relative a dati perturbati $\tilde{\boldsymbol{g}}$ e $\bar{\boldsymbol{g}}$. Si ha

$$A\tilde{\boldsymbol{u}} = \tilde{\boldsymbol{g}}$$
$$A\bar{\boldsymbol{u}} = \bar{\boldsymbol{g}}$$

da cui

$$(\tilde{\boldsymbol{u}} - \bar{\boldsymbol{u}}) = A^{-1}(\tilde{\boldsymbol{g}} - \bar{\boldsymbol{g}})$$

Se si vuole che le perturbazioni sui dati non si ripercuotano in maniera distruttiva sulle soluzioni, occorre che la matrice A^{-1} sia limitata in norma *indipendentemente* da h, in particolare per $h \to 0$. Consideriamo la matrice

$A_{q=0}$ corrispondente alla stessa discretizzazione nel caso $q(x) \equiv 0$. Si ha $A - A_{q=0} = \text{diag}(q_2, \ldots, q_{m-1}) \geq 0$. Allora

$$A_{q=0}^{-1} - A^{-1} = A_{q=0}^{-1}(A - A_{q=0})A^{-1} \geq 0$$

perché $A_{q=0}$ e A sono M-matrici. Allora $A^{-1} \leq A_{q=0}^{-1}$ e quindi $\|A^{-1}\|_\infty \leq \|A_{q=0}^{-1}\|_\infty$. Osserviamo poi che $\|A_{q=0}^{-1}\|_\infty = \|A_{q=0}^{-1}[1, \ldots, 1]^{\mathrm{T}}\|_\infty$ e che $\boldsymbol{v} = A_{q=0}^{-1}[1, \ldots, 1]^{\mathrm{T}}$ è la soluzione discreta (approssimata) di

$$\begin{cases} -v''(x) = 1 \\ v(a) = 0 \\ v(b) = 0 \end{cases}$$

la cui soluzione analitica è $v(x) = (x-a)(b-x)/2$. Poiché $v^{(4)}(x) \equiv 0$ l'errore locale $\tau_i^{(2)}$, per questo problema, è nullo e la soluzione discreta coincide, nei nodi, con la soluzione analitica. Dunque

$$\begin{aligned} \|A_{q=0}^{-1}\|_\infty = \|A_{q=0}^{-1}[1, \ldots, 1]^{\mathrm{T}}\|_\infty = \max_{2 \leq i \leq m-1} v_i = \\ = \max_{2 \leq i \leq m-1} v(x_i) \leq \max_{x \in [a,b]} v(x) \leq \frac{(b-a)^2}{8} \end{aligned}$$

e poiché $\|A^{-1}\|_\infty \leq \|A_{q=0}^{-1}\|_\infty$, si ha la maggiorazione richiesta. Abbiamo dunque mostrato che se si hanno due dati iniziali perturbati, le rispettive soluzioni saranno *diverse*, ma di poco distanti tra loro, qualunque sia il passo di discretizzazione h.

11.2.8 Convergenza

Combinando i risultati di *consistenza* e *stabilità*, si ottiene, per il problema (11.2) discretizzato mediante differenze finite centrate del secondo ordine,

$$\|\boldsymbol{e}_h\|_\infty \leq \frac{(b-a)^2}{8} \frac{h^2}{12} \|u^{(4)}\|_\infty$$

e dunque l'errore è proporzionale a h^2, posto che $u \in \mathcal{C}^4([a, b])$.

11.3 Altre differenze finite

11.3.1 Su nodi non equispaziati

Dati tre nodi x_{i-1}, x_i, x_{i+1}, con $h_{i-1} = x_i - x_{i-1}$ e $h_i = x_{i+1} - x_i$, si ha

$$u(x_{i+1}) = u(x_i) + h_i u'(x_i) + \frac{h_i^2}{2} u''(x_i) + \frac{h_i^3}{6} u^{(3)}(x_i) + \mathcal{O}(h_i^4)$$
$$u(x_{i-1}) = u(x_i) - h_{i-1} u'(x_i) + \frac{h_{i-1}^2}{2} u''(x_i) - \frac{h_{i-1}^3}{6} u^{(3)}(x_i) + \mathcal{O}(h_{i-1}^4)$$

da cui

$$u'(x_i) = \frac{u(x_{i+1}) - u(x_{i-1})}{h_{i-1} + h_i} - \frac{1}{2} \frac{h_i^2 - h_{i-1}^2}{h_{i-1} + h_i} u''(x_i) - \frac{1}{6} \frac{h_{i-1}^3 + h_i^3}{h_{i-1} + h_i} u^{(3)}(x_i) + \\ + \mathcal{O}(\max\{h_{i-1}^4, h_i^4\})$$

Se h_{i-1} e h_i non differiscono troppo (precisamente, se la loro differenza è $\mathcal{O}(\max\{h_{i-1}^2, h_i^2\})$), allora l'approssimazione con il rapporto incrementale centrato è di ordine $\mathcal{O}(\max\{h_{i-1}^2, h_i^2\})$. Analogamente, si può costruire un'approssimazione della derivata seconda

$$u''(x_i) \approx \frac{\frac{u(x_{i+1}) - u(x_i)}{h_i} - \frac{u(x_i) - u(x_{i-1})}{h_{i-1}}}{\frac{h_{i-1} + h_i}{2}}$$

La matrice corrispondente all'approssimazione mediante differenze finite di ordine due della derivata prima con griglia *non* equispaziata è (senza tener condo delle condizioni ai bordi)

$$\begin{bmatrix} u'(x_1) \\ u'(x_2) \\ u'(x_3) \\ \vdots \\ u'(x_{m-1}) \\ u'(x_m) \end{bmatrix} \approx \begin{bmatrix} * & * & * & * & * & * \\ \frac{-1}{h_1+h_2} & 0 & \frac{1}{h_1+h_2} & 0 & \dots & 0 \\ 0 & \frac{-1}{h_2+h_3} & 0 & \frac{1}{h_2+h_3} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \frac{-1}{h_{m-2}+h_{m-1}} & 0 & \frac{1}{h_{m-2}+h_{m-1}} \\ * & * & * & * & * & * \end{bmatrix} \begin{bmatrix} u(x_1) \\ u(x_2) \\ u(x_3) \\ \vdots \\ u(x_{m-1}) \\ u(x_m) \end{bmatrix}$$

Dati i nodi `x` (vettore colonna di lunghezza m), è possibile costruire il vettore $[h_1, h_2, \dots, h_{m-1}]^{\mathrm{T}}$ con il comando `h=diff(x)`. Allora la matrice, a meno della prima e dell'ultima riga, può essere costruita, direttamente in formato sparso, con i comandi

```
> d = 1./(h(1:m-2)+h(2:m-1));
> spdiags([[-d;0;0],[0;0;d]],[-1,1],m,m)
```

La costruzione della matrice relativa alla derivata seconda è analoga.

11.3.2 Non centrate

È possibile approssimare la derivata prima e seconda usando i nodi che stanno solo a sinistra (o a destra) del nodo corrente. Per esempio,

$$u'(x_1) = \frac{u(x_2) - u(x_1)}{h} + \mathcal{O}(h) = \frac{-3u(x_1) + 4u(x_2) - u(x_3)}{2h} + \mathcal{O}(h^2)$$
$$u''(x_1) = \frac{u(x_1) - 2u(x_2) + u(x_3)}{h^2} + \mathcal{O}(h) =$$
$$= \frac{2u(x_1) - 5u(x_2) + 4u(x_3) - u(x_4)}{h^2} + \mathcal{O}(h^2)$$

Ciò può risultare utile per l'approssimazione ai bordi.

11.3.3 Di ordine più elevato

Si possono per esempio costruire differenze finite di ordine quattro centrate

$$u'(x_i) = \frac{u(x_{i-2}) - 8u(x_{i-1}) + 8u(x_{i+1}) - u(x_{i+2})}{12h} + \mathcal{O}(h^4)$$
$$u''(x_i) = \frac{-u(x_{i-2}) + 16u(x_{i-1}) - 30u(x_i) + 16u(x_{i+1}) - u(x_{i+2})}{12h^2} + \mathcal{O}(h^4)$$

11.4 Condizioni al bordo

L'applicazione delle approssimazioni introdotte porta alla trasformazione del problema differenziale in un sistema di equazioni in generale non lineari

$$\tilde{F}(\boldsymbol{u}) = 0$$

Tale sistema deve essere opportunamente modificato (o completato) in modo da esprimere le condizioni al bordo.

11.4.1 Condizioni di Robin

Le più generali condizioni al bordo che consideriamo sono quelle di Robin, del tipo (per esempio per $x = a$)

$$\alpha u(a) + \beta u'(a) = \gamma \tag{11.5}$$

ove α, β e γ sono noti. Si può procedere in questo modo: si suppone l'esistenza di un nodo x_0 definito come $a - h$ ($a - h_1$ in generale) al quale corrisponde il valore $u(a - h)$ "fantasma" della soluzione approssimato da u_0. A questo

punto si discretizza l'equazione (11.5) con differenze finite centrate di ordine due

$$\alpha u_1 + \beta \frac{u_2 - u_0}{2h} = \gamma$$

Se $\beta = 0$ (e $\alpha \neq 0$), allora le condizioni al bordo sono di Dirichlet e sono usualmente date nella forma

$$u(a) = u_a = \gamma/\alpha$$

da cui si deve imporre

$$u_1 = u_a \tag{11.6}$$

Se invece $\beta \neq 0$, si ricava u_0

$$u_0 = \frac{2h}{\beta}(\alpha u_1 - \gamma) + u_2 \tag{11.7}$$

e questo valore va usato nell'approssimare $u'(x)$ e $u''(x)$, per $x = a$, nell'equazione differenziale. Se $\alpha = 0$, si parla di condizioni di Neumann e sono usualmente date nella forma

$$u'(a) = u'_a$$

da cui si deve imporre

$$u_0 = -2hu'_a + u_2 \tag{11.8}$$

Condizioni di Dirichlet

Conviene discretizzare, in un *primo momento*, il problema ai limiti senza tener conto delle condizioni al bordo. Per esempio, la discretizzazione del problema ai limiti

$$\begin{cases} u''(x) = 1, & x \in (a, b) \\ u(a) = u_a \\ u(b) = u_b \end{cases}$$

senza tener conto delle condizioni al bordo diventa

$$\tilde{F}(\boldsymbol{u}) = \tilde{A}\boldsymbol{u} - \tilde{\boldsymbol{b}} = \frac{1}{h^2} \begin{bmatrix} -2 & 1 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -2 & 1 \\ 0 & \cdots & 0 & 0 & 1 & -2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_{m-1} \\ u_m \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ \vdots \\ \vdots \\ 1 \\ 1 \end{bmatrix} = 0$$

Poi, si correggono le equazioni relative ai nodi al bordo

$$F(\boldsymbol{u}) = A\boldsymbol{u} - \boldsymbol{b} = \frac{1}{h^2}\begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -2 & 1 \\ 0 & \cdots & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_{m-1} \\ u_m \end{bmatrix} - \begin{bmatrix} u_a/h^2 \\ 1 \\ \vdots \\ \vdots \\ 1 \\ u_b/h^2 \end{bmatrix} = 0$$

Si vede che la prima riga del sistema $F(\boldsymbol{u}) = 0$ corrisponde all'equazione (11.6). A questo punto si risolve il sistema (in questo esempio lineare) di equazioni

$$F(\boldsymbol{u}) = A\boldsymbol{u} - \boldsymbol{b} = 0$$

In questo modo, però, la simmetria della matrice $\tilde{A}$ viene persa. Pertanto, non è più possibile applicare gli appositi metodi per la risoluzione di sistemi lineari simmetrici. Un metodo *numericamente* equivalente, detto metodo di penalizzazione, è quello di modificare i soli elementi diagonali della prima e dell'ultima riga inserendo un numero molto grande M

$$\frac{1}{h^2}\begin{bmatrix} M & 1 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -2 & 1 \\ 0 & \cdots & 0 & 0 & 1 & M \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_{m-1} \\ u_m \end{bmatrix} - \begin{bmatrix} Mu_a/h^2 \\ 1 \\ \vdots \\ \vdots \\ 1 \\ Mu_b/h^2 \end{bmatrix} = 0$$

Per poter usare l'algoritmo di Choleski, è inoltre necessario che la matrice sia definita positiva. Conviene allora considerare il problema $-u''(x) = -1$.

Condizioni di Neumann

L'espressione trovata per u_0 nel caso generale delle condizioni di Robin (11.7) va usata poi in qualunque stencil di discretizzazione. Per esempio, la discretizzazione del problema ai limiti

$$\begin{cases} u''(x) - u'(x) = 1 \\ u'(a) = u'_a \\ u(b) = u_b \end{cases}$$

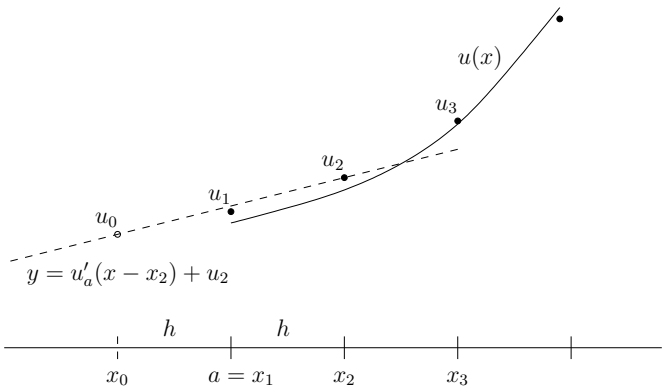


Figura 11.3: Imposizione di una condizione di Neumann sull'estremo sinistro.

in $x = a$ sarà

$$\frac{u_0 - 2u_1 + u_2}{h^2} - \frac{u_2 - u_0}{2h} = \frac{(u_2 - 2hu'_a) - 2u_1 + u_2}{h^2} - \frac{u_2 - (u_2 - 2hu'_a)}{2h} = \\ = \frac{2u_2 - 2u_1 - 2hu'_a}{h^2} - u'_a = 1$$

da cui

$$\frac{2u_2 - 2u_1}{h^2} = 1 + \frac{2u'_a}{h} + u'_a$$

e dunque

$$F(\boldsymbol{u}) = A\boldsymbol{u} - \boldsymbol{b} = \frac{1}{h^2} \begin{bmatrix} -2 & 2 & 0 & 0 & \cdots & 0 \\ 1 + \frac{h}{2} & -2 & 1 - \frac{h}{2} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 + \frac{h}{2} & -2 & 1 - \frac{h}{2} \\ 0 & \cdots & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_{m-1} \\ u_m \end{bmatrix} - \begin{bmatrix} 1 + \frac{2u'_a}{h} + u'_a \\ 1 \\ \vdots \\ \vdots \\ 1 \\ u_b/h^2 \end{bmatrix} = 0$$

Anche in questo caso, è possibile recuperare l'eventuale simmetria di partenza riscalando opportunamente la prima equazione.

11.4.2 Importanza delle condizioni al bordo

Spesso si trascura l'importanza di una corretta imposizione delle condizioni al bordo e si pensa che l'influenza delle condizioni riguardi solamente un intorno del bordo. Ovviamente non è così: basti pensare all'equazione differenziale

$$-u''(x) = 2, \quad x \in (-1, 1)$$

munita delle condizioni al bordo

$$u(-1) = u(1) = 0$$

(la cui soluzione è $u(x) = -x^2 + 1$) oppure

$$\begin{cases} u(-1) = 0 \\ u'(1) = 0 \end{cases}$$

(la cui soluzione è $u(x) = -x^2 + 2x + 3$) oppure

$$u'(-1) = u'(1) = 0$$

(nessuna soluzione) oppure

$$\begin{cases} u'(-1) = 2 \\ u'(1) = -2 \end{cases}$$

(infinite soluzioni $u(x) = -x^2 + k$).

11.5 Un esempio: l'equazione della catenaria

Consideriamo l'equazione della *catenaria* (corda flessibile inestensibile appesa agli estremi)

$$\begin{cases} u''(x) = \alpha\sqrt{1 + u'(x)^2}, \quad x \in (-1, 1) \\ u(-1) = 1 \\ u(1) = 1 \end{cases} \tag{11.9}$$

la cui soluzione analitica è

$$u(x) = \frac{\cosh(ax)}{a} - \frac{\cosh a}{a} + 1$$

e il parametro a dipende dalla lunghezza della corda (vedi Appendice A.5). La discretizzazione mediante differenze finite centrate del secondo ordine è

$$A \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_{m-1} \\ u_m \end{bmatrix} - \alpha \begin{bmatrix} 1 \\ \sqrt{1 + \left(\frac{u_3 - u_1}{2h}\right)^2} \\ \vdots \\ \sqrt{1 + \left(\frac{u_m - u_{m-2}}{2h}\right)^2} \\ 1 \end{bmatrix} = \boldsymbol{b}$$

Si tratta dunque di risolvere il sistema non lineare nell'incognita $\boldsymbol{u}$

$$F(\boldsymbol{u}) = A\boldsymbol{u} - \alpha\sqrt{1 + (B\boldsymbol{u})^2} - \boldsymbol{b} = 0$$

ove

$$A = \frac{1}{h^2}\begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -2 & 1 \\ 0 & \cdots & 0 & 0 & 0 & 1 \end{bmatrix}, \quad B = \frac{1}{2h}\begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ -1 & 0 & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1 & 0 & 1 \\ 0 & \cdots & 0 & 0 & 0 & 0 \end{bmatrix}$$

e $\boldsymbol{b} = [1/h^2 - \alpha, 0, \ldots, 0, 1/h^2 - \alpha]^{\mathrm{T}}$.

11.5.1 Iterazioni di punto fisso

Si può tentare di risolvere il sistema di equazioni $F(\boldsymbol{u}) = 0$ mediante iterazioni di punto fisso, che consistono nel risolvere il sistema lineare

$$A\boldsymbol{u}^{(r+1)} = \alpha\sqrt{1 + (B\boldsymbol{u}^{(r)})^2} + \boldsymbol{b}$$

L'applicazione del metodo risulta molto semplice: si può decomporre A nei fattori LU una sola volta e risolvere due sistemi lineari triangolari ad ogni iterazione. La funzione G deve essere una contrazione e ciò può essere difficile da verificare. Inoltre, la convergenza risulta essere lineare. Per l'esempio della catenaria, comunque, il metodo delle iterazioni di punto fisso converge adeguatamente.

11.5.2 Metodo di Newton

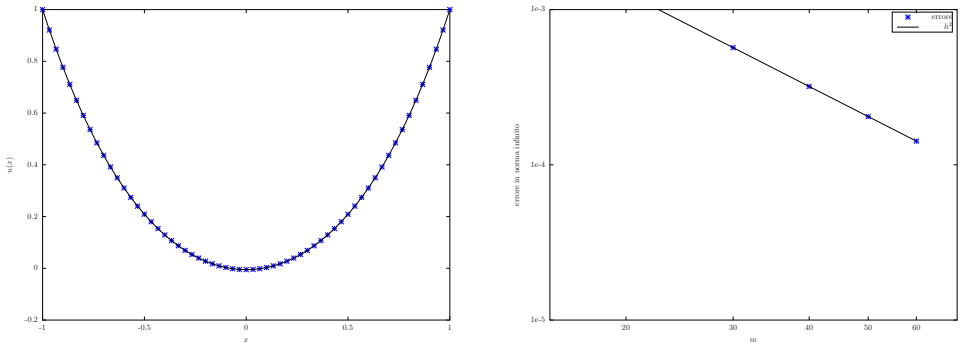
Volendo applicare il metodo di Newton (a convergenza quadratica) è necessario calcolare lo jacobiano di $F(\boldsymbol{u})$, che risulta essere

$$J_F(\boldsymbol{u}^{(r)}) = A - \alpha D(\boldsymbol{u}^{(r)})B$$

ove

$$D = (d_{ij}(\boldsymbol{u}^{(r)})), \quad d_{ij}(\boldsymbol{u}^{(r)}) = \begin{cases} \dfrac{(B\boldsymbol{u}^{(r)})_i}{\sqrt{1 + (B\boldsymbol{u}^{(r)})_i^2}}, & i = j \\ 0, & i \neq j \end{cases}$$

A questo punto, l'iterazione del metodo di Newton consiste nella risoluzione



catenaria.m

Figura 11.4: Soluzione dell'equazione della catenaria (sinistra, $m = 60$) e ordine di convergenza (destra).

del sistema lineare

$$J_F(\boldsymbol{u}^{(r)})\left(\boldsymbol{u}^{(r+1)} - \boldsymbol{u}^{(r)}\right) = -F(\boldsymbol{u}^{(r)})$$

In generale, lo jacobiano di un funzionale $F(\boldsymbol{u})$ calcolato in $\boldsymbol{u}^{(r)}$ e applicato a $\boldsymbol{v}$ è

$$J_F(\boldsymbol{u}^{(r)})\boldsymbol{v} = \frac{\mathrm{d}F}{\mathrm{d}\boldsymbol{u}}(\boldsymbol{u}^{(r)})\boldsymbol{v} = \lim_{\varepsilon\to 0}\frac{F(\boldsymbol{u}^{(r)} + \varepsilon\boldsymbol{v}) - F(\boldsymbol{u}^{(r)})}{\varepsilon}$$

Come soluzione iniziale si prende solitamente una funzione semplice che soddisfi le condizioni al bordo.

11.6 Norme ed errori

Data una funzione $u(x)$ e due diverse discretizzazioni su nodi equispaziati $[\tilde{u}_1, \ldots, \tilde{u}_m] \approx [u(\tilde{x}_1), \ldots, u(\tilde{x}_m)]$ e $[\hat{u}_1, \ldots, \hat{u}_l] \approx [u(\hat{x}_1), \ldots, u(\hat{x}_l)]$, $\{\tilde{x}_i\}_i \subset [a, b]$, $\{\hat{x}_i\}_i \subset [a, b]$, non ha molto senso confrontare gli errori $\|[u(\tilde{x}_1) - \tilde{u}_1, u(\tilde{x}_2) - \tilde{u}_2, \ldots, u(\tilde{x}_m) - \tilde{u}_m]\|_2$ e $\|[u(\hat{x}_1) - \hat{u}_1, y(\hat{x}_2) - \hat{u}_2, \ldots, u(\hat{x}_l) - \hat{u}_l]\|_2$. Si preferisce usare la norma infinito, oppure la norma $\|u\|_2\sqrt{\frac{b-a}{m}}$, che risulta essere una approssimazione mediante quadratura con formula dei rettangoli della norma in L^2 di $u(x)$.

Se si devono invece confrontare tra loro le due discretizzazioni, occorre che i nodi siano "intercalati" e bisogna fare attenzione alla *falsa superconvergenza* (vedi Figura 11.5). Se si calcola una soluzione di riferimento con $\bar{m}$ punti di discretizzazione, si ha

$$\left|\|u_m - u\|_\infty - \|u - u_{\bar{m}}\|_\infty\right| \leq \|u_m - u_{\bar{m}}\|_\infty \leq \|u_m - u\|_\infty + \|u - u_{\bar{m}}\|_\infty$$

fsc.m

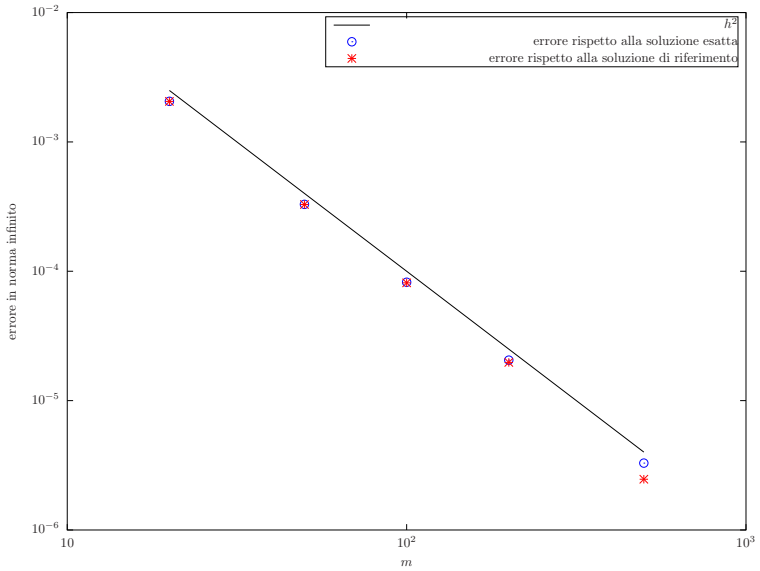


Figura 11.5: Convergenza e *falsa superconvergenza* per la risoluzione di $u''(x) = -\sin(x)$, $u(0) = u(\pi) = 0$.

da cui

$$\|u_m - u\|_\infty - \varepsilon \leq \|u_m - u_{\bar{m}}\|_\infty \leq \|u_m - u\|_\infty + \varepsilon$$

se $\|u - u_{\bar{m}}\|_\infty = \varepsilon < \|u_m - u\|_\infty$. Ciò significa che si può stimare l'errore di u_m usando una soluzione di riferimento $u_{\bar{m}}$ solo se questa dista poco dalla soluzione analitica e se $m \ll \bar{m}$, altrimenti la stima dice solo che $\|u_m - u_{\bar{m}}\|_\infty \lessapprox 2\varepsilon$. Si ha cioè l'impressione che la soluzione numerica sia più vicina alla soluzione analitica di quello che dovrebbe, invece è *solo* molto vicina a quella di riferimento (per assurdo, se $m = \bar{m}$, $\|u_m - u_{\bar{m}}\|_\infty = 0 \neq \|u_m - u\|_\infty$).

Una maniera molto comoda per verificare l'ordine di un metodo si basa sulla seguente osservazione. Siano $\boldsymbol{e}_{\tilde{m}}$ e $\boldsymbol{e}_{\hat{m}}$ gli errori corrispondenti a due discretizzazioni con $\tilde{m}+1$ e $\hat{m}+1$ punti. Supponiamo che

$$\|\boldsymbol{e}_{\tilde{m}}\|_\infty = \frac{C}{\tilde{m}^p}$$
$$\|\boldsymbol{e}_{\hat{m}}\|_\infty = \frac{C}{\hat{m}^p}$$

Si ricava

$$\log\|\boldsymbol{e}_{\tilde{m}}\|_\infty - \log\|\boldsymbol{e}_{\hat{m}}\|_\infty = -p(\log\tilde{m} - \log\hat{m})$$

Dunque, in un grafico *logaritmico-logaritmico*, l'errore $\|\boldsymbol{e}_m\|_\infty$ si dispone su una retta di pendenza $-p$ (cioè parallelo alla "retta" m^{-p}) rispetto a m.

11.7 Derivate ed equazioni differenziali

Il calcolo esplicito di derivate e la risoluzione di equazioni differenziali sono problemi ben distinti. Per esempio, se è vero che il calcolo della derivata di $\exp(x)$ tramite rapporto incrementale centrato

$$\exp'(x) \approx \frac{\exp(x+h) - \exp(x-h)}{2h}$$

soffre di severi errori di cancellazione per h piccolo, meno problematica è la risoluzione del problema differenziale

$$\begin{cases} u'(x) = \exp(x), & x \in (0,1) \\ u(0) = 1 \\ u(1) = \mathrm{e} \end{cases}$$

tramite differenze finite centrate del secondo ordine, in cui è il numero di condizionamento della matrice ad influire sulla corretta risoluzione del sistema lineare.

derequadiff.m

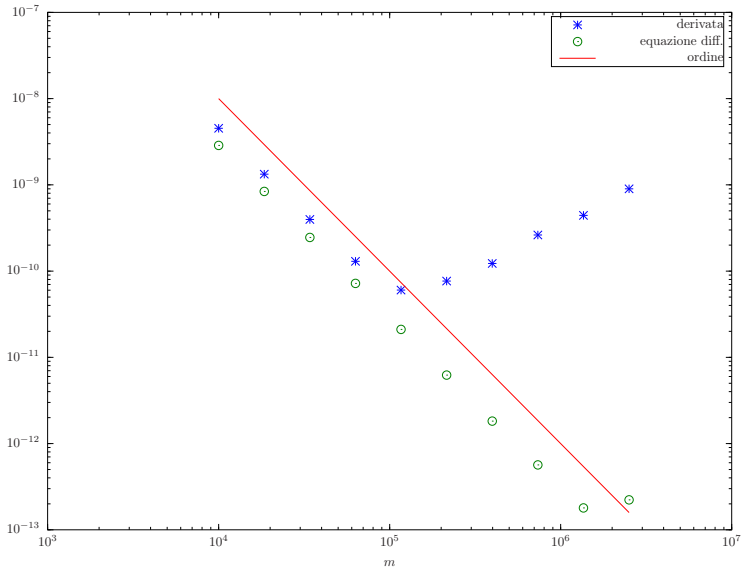


Figura 11.6: Andamento degli errori nel calcolo della derivata di $\exp(x)$.

In Figura 11.6 è riportato l'andamento di $\max_{2\le i\le m}\{|(\exp(x_i + 1/m) - \exp(x_i - 1/m))/(2/m) - \exp(x_i)|\}$ e $\max_{2\le i\le m}\{|u_i - \exp(x_i)|\}$, $x_i = (i-1)/m$, $i = 1, 2, \ldots, m+1$ per valori di m da 10000 a più di 2500000. L'andamento del numero di condizionamento della matrice si comporta come m.

Capitolo 12

Metodo di shooting

È possibile trasformare il problema (10.1) in un sistema differenziale del primo ordine

$$\boldsymbol{y}'(t) = \boldsymbol{f}(t, \boldsymbol{y}(t)), \quad t \in (a, b]$$

tramite il cambiamento di variabili $t = x$, $y_1(t) = u(x)$, $y_2(t) = u'(x)$, $\boldsymbol{f}(t, \boldsymbol{y}(t)) = [y_2(t), f(t, y_1(t), y_2(t)]^{\mathrm{T}}$. Per quanto riguarda le condizioni iniziali, mentre quella per $y_1(t)$ è $y_1(a) = u_a$, quella per $y_2(t)$ non è definita. Si può allora introdurre un parametro $s \in \mathbb{R}$ e considerare la seguente famiglia di problemi ai valori iniziali

$$\begin{cases} \boldsymbol{y}'(t) = \boldsymbol{f}(t, \boldsymbol{y}(t)), \quad t \in (a, b] \\ y_1(a) = u_a \\ y_2(a) = s \end{cases} \tag{12.1}$$

Dato s, il sistema sopra può essere risolto con un opportuno metodo per problemi ai valori iniziali. Poiché s è il valore della derivata prima di $u(x)$, tale metodo di risoluzione prende il nome di *shooting*. Chiamiamo $y_1(t \mid y_2(a) = s)$ (da leggersi "valore di y_1 in t dato che y_2 in a vale s") la prima componente della soluzione. Si dovrà ovviamente trovare $\bar{s}$ tale che $y_1(t \mid y_2(a) = \bar{s}) = u(x)$, $t = x \in [a, b]$. In particolare, dovrà essere $y_1(b \mid y_2(a) = \bar{s}) = u_b$. Introduciamo allora la funzione

$$F(s) = y_1(b \mid y_2(a) = s) - u_b$$

Si tratta di risolvere l'equazione (in generale non lineare) $F(s) = 0$.

12.1 Metodo di bisezione

Dati due valori s_1 e s_2 per cui $F(s_1)F(s_2) < 0$, è possibile applicare il metodo di bisezione per trovare lo zero di $F(s)$. Poiché la soluzione di (12.1) è

approssimata a meno di un errore dipendente dal passo di discretizzazione temporale, la tolleranza richiesta per il metodo di bisezione dovrà essere (leggermente) inferiore a tale errore.

12.2 Metodo di Newton

Per applicare il metodo di Newton, è necessario calcolare $F'(s)$. Definiamo a tal scopo

$$v(x) = \frac{\partial}{\partial s} u(x \mid u'(a) = s) = \frac{\partial}{\partial s} y_1(t \mid y_2(a) = s)$$

Derivando rispetto a s nel problema ai limiti

$$\begin{cases} u''(x) = f(x, u(x), u'(x)), \quad x \in (a, b) \\ u(a) = u_a \\ u'(a) = s \end{cases}$$

(la cui incognita $u(x)$ è proprio $u(x \mid u'(a) = s)$) si ha

$$\frac{\partial}{\partial s} u''(x) = \frac{\partial}{\partial s} f(x, u(x), u'(x))$$

da cui, scambiando l'ordine di derivazione

$$v''(x) = f_u(x, u(x), u'(x))v(x) + f_{u'}(x, u(x), u'(x))v'(x), \quad x \in (a, b)$$

Per quanto riguarda le condizioni iniziali per $v(x)$, si ha

$$v(a) = \frac{\partial}{\partial s} u(a \mid u'(a) = s) = 0$$
$$v'(a) = \frac{\partial}{\partial s} u'(a \mid u'(a) = s) = 1$$

Dunque, per calcolare $F'(s) = v(b)$ occorre risolvere il *sistema variazionale* (lineare in $v(x)$)

$$\begin{cases} v''(x) = f_u(x, u(x), u'(x))v(x) + f_{u'}(x, u(x), u'(x))v'(x), \quad x \in (a, b) \\ v(a) = 0 \\ v'(a) = 1 \end{cases}$$

In conclusione, per calcolare la coppia $F(s)$ e $F'(s)$ in un generico punto s, occorre risolvere il sistema differenziale del primo ordine ai dati iniziali

$$\begin{cases} y_1'(t) = y_2(t) \\ y_2'(t) = f(t, y_1(t), y_2(t)) \\ y_3'(t) = y_4(t) \\ y_4'(t) = f_{y_1}(t, y_1(t), y_2(t))y_3(t) + f_{y_2}(t, y_1(t), y_2(t))y_4(t) \\ y_1(a) = u_a \\ y_2(a) = s \\ y_3(a) = 0 \\ y_4(a) = 1 \end{cases}$$

fino al tempo $t = b$. Quindi $F(s) = y_1(b)$ e $F'(s) = y_3(b)$. Poiché le equazioni per $y_1'(t)$ e $y_2'(t)$ non dipendono da $y_3(t)$ e $y_4(t)$, è possibile disaccoppiare le prime due componenti dalle seconde due.

Una semplificazione del metodo di Newton che non richiede il calcolo di $F'(s)$ è il metodo delle secanti.

12.3 Problema ai limiti con frontiera libera

Un caso particolarmente interessante per l'applicazione del metodo di shooting è quello a frontiera libera *(free boundary)*

$$\begin{cases} u''(x) = f(x, u(x), u'(x)), \quad x \in (s, b) \\ u(s) = \alpha \\ u'(s) = \beta \\ u(b) = u_b \end{cases} \tag{12.2}$$

ove i valori di u e di u' sono assegnati in un punto incognito s, $s < b$. La funzione di cui si deve trovare lo zero è, in questo caso,

$$F(s) = u(b \mid u(s) = \alpha, u'(s) = \beta) - u_b$$

(scriveremo $F(s) = u(b \mid s) - u_b$ per brevità). Dati due punti s_1 e s_2 tali che $F(s_1)F(s_2) < 0$, l'applicazione del metodo di bisezione non presenta difficoltà. Per quanto riguarda il metodo di Newton, il sistema variazionale per

$$v(x) = \frac{\partial}{\partial s} u(x \mid s) = \lim_{h \to 0} \frac{u(x \mid s + h) - u(x \mid s)}{h}$$

è analogo al caso precedente. L'unica diversità è data dalle condizioni iniziali (in s). Si ha

$$v(s) = \lim_{h\to 0} \frac{u(s \mid s+h) - u(s \mid s)}{h}$$

Ora, $u(s \mid s) = \alpha$. Poi

$$u(s \mid s+h) = u(s+h \mid s+h) - hu'(s+h \mid s+h) + \mathcal{O}(h^2) = \alpha - h\beta + \mathcal{O}(h^2)$$

Dunque, $v(s) = -\beta$. In maniera analoga

$$v'(s) = \lim_{h\to 0} \frac{u'(s \mid s+h) - u'(s \mid s)}{h} = -u''(s)$$

ove il valore $u''(s)$ si ricava dal problema (12.2) e vale $f(s, \alpha, \beta)$.

Capitolo 13

Equazione di Poisson

Di particolare interesse è l'equazione di Poisson

$$-\nabla^2 u(\boldsymbol{x}) = f(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega \subset \mathbb{R}^d$$

ove ∇^2 è l'operatore *laplaciano* definito da

$$\nabla^2 = \sum_{k=1}^{d} \frac{\partial^2}{\partial x_k^2}$$

L'equazione è solitamente accompagnata da condizioni al bordo di Dirichlet o di Neumann.

13.1 Equazione di Poisson bidimensionale

Analizziamo numericamente in dettaglio il caso $d = 2$ ($\boldsymbol{x} = (x, y)$) e $\Omega = [a, b] \times [c, d]$.

13.1.1 Condizioni al bordo di Dirichlet

Consideriamo dapprima il caso con condizioni al bordo di Dirichlet. Dunque

$$\begin{cases} -\nabla^2 u(x, y) = f(x, y), & (x, y) \in [a, b] \times [c, d] \subset \mathbb{R}^2 \\ u(a, y) = D_a(y) \\ u(b, y) = D_b(y) \\ u(x, c) = D_c(x) \\ u(x, d) = D_d(x) \end{cases}$$

con le necessarie condizioni di compatibilità ai vertici. Introduciamo una discretizzazione $x_i = a + (i - 1)h_x$, $i = 1, 2, \ldots, m_x$, $h_x = (b - a)/(m_x - 1)$

e $y_j = c + (j-1)h_y$, $j = 1, 2, \ldots, m_y$, $h_y = (d-c)/(m_y - 1)$. Introduciamo infine la discretizzazione di $u(x, y)$ definita da

$$u_k \approx u(x_i, y_j), \quad k = (j-1)m_x + i$$

di cui si vede un esempio in Figura 13.1. La matrice di discretizzazio-

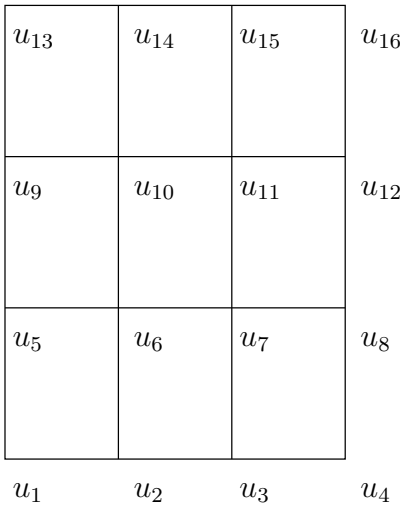


Figura 13.1: Numerazione di una griglia bidimensionale

ne alle differenze finite centrate del secondo ordine, *senza* tener conto delle condizioni al bordo, è data da

$$A = I_{m_y} \otimes A_x + A_y \otimes I_{m_x}$$

ove $\otimes$ indica il prodotto di Kronecker e

$$A_x = \frac{1}{h_x^2}\begin{bmatrix} 2 & -1 & 0 & \ldots & \ldots & 0 \\ -1 & 2 & -1 & 0 & \ldots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ldots & 0 & -1 & 2 & -1 \\ 0 & \ldots & \ldots & 0 & -1 & 2 \end{bmatrix}, \quad A_y = \frac{1}{h_y^2}\begin{bmatrix} 2 & -1 & 0 & \ldots & \ldots & 0 \\ -1 & 2 & -1 & 0 & \ldots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ldots & 0 & -1 & 2 & -1 \\ 0 & \ldots & \ldots & 0 & -1 & 2 \end{bmatrix}$$

ove $A_x \in \mathbb{R}^{m_x \times m_x}$ e $A_y \in \mathbb{R}^{m_y \times m_y}$. Poi, le righe di indice, diciamo k, corrispondente ad un nodo al bordo vanno sostituite con il vettore della base

canonica e_k, diviso per $h_x^2 + h_y^2$. Il termine noto è $[b_1, b_2, \ldots, b_{m_x m_y}]^{\mathrm{T}}$, ove

$$b_k = \begin{cases} f(x_i, y_j) & \text{se } (x_i, y_j) \text{ è un nodo interno, } k = (j-1)m_x + i \\ \frac{D_a(y_j)}{h_x^2+h_y^2} & \text{se } x_i = a,\ k = (j-1)m_x + i \\ \frac{D_b(y_j)}{h_x^2+h_y^2} & \text{se } x_i = b,\ k = (j-1)m_x + i \\ \frac{D_c(x_i)}{h_x^2+h_y^2} & \text{se } y_j = c,\ k = (j-1)m_x + i \\ \frac{D_d(x_i)}{h_x^2+h_y^2} & \text{se } y_j = d,\ k = (j-1)m_x + i \end{cases}$$

Alternativamente, si può sostituire il solo termine diagonale delle righe corrispondenti ad un nodo al bordo con un coefficiente $M/(h_x^2 + h_y^2)$, $M \gg 1$ e moltiplicare per M il corrispondente elemento nel termine noto. Questa procedura permette di assegnare, di fatto, le condizioni al bordo di Dirichlet, mantenendo la matrice A *simmetrica*.

In GNU Octave, la corretta numerazione dei nodi avviene con i comandi

```
> x = linspace(a,b,mx);
> y = linspace(c,d,my);
> [X,Y] = ndgrid(x,y);
```

e la costruzione della matrice A tramite il comando `kron`.

13.1.2 Condizioni al bordo miste

L'equazione di Poisson non può essere accompagnata solo da condizioni al bordo di Neumann, altrimenti la soluzione è indeterminata. Consideriamo allora il seguente problema con condizioni al bordo miste

$$\begin{cases} -\nabla^2 u(x,y) = f(x,y), \quad (x,y) \in [a,b] \times [c,d] \subset \mathbb{R}^2 \\ u(b,y) = D_b(y) \\ u(x,c) = D_c(x), \quad D_c(b) = D_b(c) \\ -\dfrac{\partial u}{\partial x}(x,y) = N_a(y), \quad x = a,\ c < y < d \\ \dfrac{\partial u}{\partial y}(x,y) = N_d(x), \quad y = d,\ x < b \end{cases}$$

La matrice di discretizzazione alle differenze finite centrate del secondo ordine è data da

$$A = I_{m_y} \otimes A_x + A_y \otimes I_{m_x}$$

ove

$$A_x = \frac{1}{h_x^2}\begin{bmatrix} 2 & -2 & 0 & \dots & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \dots & 0 & -1 & 2 & -1 \\ 0 & \dots & \dots & 0 & -1 & 2 \end{bmatrix}, \quad A_y = \frac{1}{h_y^2}\begin{bmatrix} 2 & -1 & 0 & \dots & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \dots & 0 & -1 & 2 & -1 \\ 0 & \dots & \dots & 0 & -2 & 2 \end{bmatrix}$$

Poi, le righe di indice, diciamo k, corrispondente ad un nodo al bordo su cui sono prescritte condizioni di Dirichlet vanno sostituite con il vettore della base canonica e_k, diviso per $h_x^2 + h_y^2$. La riga di indice m_y, corrispondente al nodo di bordo (a, c), va sostituita con

$$[0, \dots, 0, 1] \otimes \frac{1}{h_x^2}[-2, 5, -4, 1, 0, \dots, 0] + \frac{1}{h_y^2}[0, \dots, 0, -2, 2] \otimes [1, 0, \dots, 0]$$

(si può verificare che lo stencil $[2, -5, 4, -1]/h_x^2$ è un'approssimazione al secondo ordine della derivata seconda). Il termine noto è $[b_1, b_2, \dots, b_{m_x m_y}]^{\mathrm{T}}$, ove

$$b_k = \begin{cases} f(x_i, y_j) & \text{se } (x_i, y_j) \text{ è un nodo interno, } k = (j-1)m_x + i \\ \frac{D_b(y_j)}{h_x^2+h_y^2} & \text{se } x_i = b,\ k = (j-1)m_x + i \\ \frac{D_c(x_i)}{h_x^2+h_y^2} & \text{se } y_j = c,\ k = (j-1)m_x + i \\ f(x_i, y_j) + \frac{2N_a(y_i)}{h_x} & \text{se } x_i = a,\ k = (j-1)m_x + i,\ j \neq 1,\ j \neq m_y \\ f(x_i, y_j) + \frac{2N_d(x_i)}{h_y} & \text{se } y_j = d,\ k = (j-1)m_y + i,\ i \neq m_x \end{cases}$$

Capitolo 14

Metodi variazionali

14.1 Formulazione variazionale di un problema modello

Un filo elastico (con tensione unitaria) sottoposto ad un carico soddisfa, sotto opportune ipotesi di regolarità e nel caso di piccole deformazioni, l'equazione

$$\begin{cases} -u''(x) = g(x), & x \in (0,1) \\ u(0) = u(1) = 0 \end{cases} \tag{14.1}$$

ove $u(x)$ rappresenta lo scostamento dalla posizione di riposo orizzontale. Supponiamo che il carico sia $g(x) = g_\varepsilon(x)$

$$\begin{cases} -u''(x) = g_\varepsilon(x), & x \in (0,1) \\ u(0) = u(1) = 0 \end{cases} \tag{14.2}$$

ove

$$g_\varepsilon(x) = \begin{cases} 0 & 0 \le x < \frac{1}{2} - \varepsilon \\ -\frac{1}{2\varepsilon} & \frac{1}{2} - \varepsilon \le x \le \frac{1}{2} + \varepsilon \\ 0 & \frac{1}{2} + \varepsilon < x \le 1 \end{cases}$$

La funzione discontinua g_ε rappresenta la densità di carico e il carico totale vale

$$\int_0^1 g_\varepsilon(x)\mathrm{d}x = -1$$

La "soluzione" di (14.2) è

$$u_\varepsilon(x) = \begin{cases} -\dfrac{1}{2}x & 0 \le x \le \dfrac{1}{2} - \varepsilon \\ \dfrac{1}{4\varepsilon}\left(x - \dfrac{1}{2}\right)^2 + \dfrac{\varepsilon - 1}{4} & \dfrac{1}{2} - \varepsilon \le x \le \dfrac{1}{2} + \varepsilon \\ -\dfrac{1}{2}(1 - x) & \dfrac{1}{2} + \varepsilon \le x \le 1 \end{cases}$$

In che senso soluzione? Chiaramente $u''_\varepsilon(1/2 \pm \varepsilon)$ non esiste e quindi non è vero che $-u''_\varepsilon(x) = g_\varepsilon(x)$, $x \in (0,1)$. Cerchiamo dunque una formulazione alternativa che renda sensato il modello per un problema così semplice e fisico come (14.2).

Introduciamo il seguente spazio lineare:

$$V = \{v\colon\ v \in \mathcal{C}^0([0,1]),\ v' \text{ continua a tratti e limitata},\ v(0) = v(1) = 0\}$$

e il prodotto scalare su V

$$(v, w) = \int_0^1 v(x)w(x)\mathrm{d}x$$

Con il termine "v' continua a tratti e limitata" intendiamo che v è derivabile ovunque eccetto alpiù un insieme finito di punti e dove è derivabile v' è continua e limitata.

Teorema 4 (Formulazione variazionale)**.** *Se $u(x)$ è la soluzione del problema (14.1), allora $u \in V$ e*

$$(u', v') = (g, v), \quad \forall v \in V \tag{14.3}$$

Dimostrazione. Sia u soluzione di (14.1). Allora, per ogni $v \in V$,

$$\int_0^1 -u''(x)v(x)\mathrm{d}x = \int_0^1 g(x)v(x)\mathrm{d}x = (g, v)$$

Integrando per parti,

$$\int_0^1 -u''(x)v(x)\mathrm{d}x = -u'(x)v(x)\Big|_0^1 + \int_0^1 u'(x)v'(x)\mathrm{d}x = (u', v')$$

poiché $v(0) = v(1) = 0$. □

Per quanto visto per il problema modello (11.2), se $g \in \mathcal{C}^0(0,1)$, la soluzione di (14.1) esiste ed è unica. Per quanto appena dimostrato, essa è soluzione anche di (14.3). Ritornando al problema modello (14.1), ove però non facciamo ipotesi su g, la soluzione classica, quella di classe (almeno) $\mathcal{C}^2$, si chiama *soluzione forte* del problema (14.1), mentre la soluzione di (14.3) si chiama *soluzione debole* del problema (14.1). Con il teorema e l'esempio precedenti abbiamo dimostrato che se esiste la soluzione forte, essa è anche soluzione debole, ma non è sempre vero il contrario (cioè può esistere la sola soluzione debole). Questo è il caso di (14.2), per cui la soluzione debole è $u_\varepsilon \in V$, visto che

$$\begin{aligned}\int_0^1 u_\varepsilon'(x)v'(x)\mathrm{d}x &= \int_0^{\frac{1}{2}-\varepsilon} u_\varepsilon'(x)v'(x)\mathrm{d}x + \int_{\frac{1}{2}-\varepsilon}^{\frac{1}{2}+\varepsilon} u_\varepsilon'(x)v'(x)\mathrm{d}x + \int_{\frac{1}{2}+\varepsilon}^1 u_\varepsilon'(x)v'(x)\mathrm{d}x = \\ &= -\int_0^{\frac{1}{2}-\varepsilon} u_\varepsilon''(x)v(x)\mathrm{d}x - \int_{\frac{1}{2}-\varepsilon}^{\frac{1}{2}+\varepsilon} u_\varepsilon''(x)v(x)\mathrm{d}x - \int_{\frac{1}{2}+\varepsilon}^1 u_\varepsilon''(x)v(x)\mathrm{d}x = \\ &= -\int_{\frac{1}{2}-\varepsilon}^{\frac{1}{2}+\varepsilon} \frac{1}{2\varepsilon}v(x)\mathrm{d}x = \int_{\frac{1}{2}-\varepsilon}^{\frac{1}{2}+\varepsilon} g_\varepsilon(x)v(x)\mathrm{d}x = \int_0^1 g_\varepsilon(x)v(x)\mathrm{d}x\end{aligned}$$

Qualora si trovi una soluzione debole, ha però senso verificare se per caso non sia anche forte. Infatti, se $u \in V$ è soluzione di (14.3) *e* $u \in \mathcal{C}^2([0,1])$ *e* g è continua allora $0 = (u', v') - (g, v) = (-u'', v) - (g, v) = -(u'' + g, v)$ per ogni $v \in V$. Poiché $u'' + g$ è continua, si deduce $-u''(x) = g(x)$ per $0 < x < 1$.

Per quanto visto, la formulazione variazionale (14.3) del problema (14.1) è in realtà le più "fisica": pensando al problema della trave, essa permette di descrivere, per esempio, anche il caso in cui la densità di carico $g(x)$ non sia continuo. Basta infatti che sia possibile calcolare (g, v), $v \in V$ e dunque basta, per esempio, che g sia continua a tratti. Quindi, in generale, è possibile come modello per un fenomeno fisico la sola formulazione debole. La soluzione debole, se esiste, è unica: infatti, se u_1 e u_2 sono due soluzioni di (14.3), allora

$$(u_1' - u_2', v') = 0, \quad \forall v \in V$$

e in particolare per $v = u_1 - u_2$. Dunque

$$\int_0^1 (u_1'(x) - u_2'(x))^2 \mathrm{d}x = 0$$

e quindi $u_1'(x) - u_2'(x) = (u_1(x) - u_2(x))' = 0$. Pertanto $u_1 - u_2$ è costante e siccome $u_1(0) - u_2(0) = 0$, allora $u_1(x) - u_2(x) = 0$.

14.1.1 Metodo di approssimazione variazionale

Prendiamo un sottospazio V_m di V di dimensione finita. Si cerca allora $\hat{u} \in V_m$ tale che

$$(\hat{u}', v') = (g, v), \quad \forall v \in V_m \tag{14.4}$$

(metodo di Galerkin).

Teorema 5. *Il problema (14.4) ha un'unica soluzione.*

Dimostrazione. Sia $\{\phi_j\}_{j=1}^m$ una base di V_m. Allora, se esiste $\hat{u}(x) \in V_m$ soluzione è

$$\hat{u}(x) = \sum_{j=1}^{m} \hat{u}_j \phi_j(x)$$

e il problema (14.4) si riscrive, per $i = 1, 2, \ldots, m$,

$$\int_0^1 \hat{u}'(x)\phi_i'(x)\mathrm{d}x = \left(\left(\sum_{j=1}^{m} \hat{u}_j \phi_j \right)', \phi_i' \right) = \sum_{j=1}^{m} (\phi_j', \phi_i')\hat{u}_j = A\boldsymbol{u} = (g, \phi_i)$$

ove $A = (a_{ij}) = (\phi_j', \phi_i')$ e $\boldsymbol{u} = [\hat{u}_1, \ldots, \hat{u}_m]^{\mathrm{T}}$. Quindi, l'esistenza ed unicità di $\hat{u}(x)$ equivale all'esistenza ed unicità della soluzione di un sistema lineare di matrice A. Calcoliamo ora $\boldsymbol{w}^{\mathrm{T}} A \boldsymbol{w}$ per $\boldsymbol{w} = [w_1, \ldots, w_m]^{\mathrm{T}}$. Si ha

$$\boldsymbol{w}^{\mathrm{T}} A \boldsymbol{w} = \sum_{i=1}^{m} w_i \left(\sum_{j=1}^{m} (\phi_i', \phi_j') w_j \right)$$

da cui, per la linearità del prodotto scalare,

$$\boldsymbol{w}^{\mathrm{T}} A \boldsymbol{w} = \left(\left(\sum_{i=1}^{m} w_i \phi_i(x) \right)', \left(\sum_{j=1}^{m} w_j \phi_j(x) \right)' \right) = \int_0^1 \left(\sum_{j=1}^{m} w_j \phi_j'(x) \right)^2 \mathrm{d}x \geq 0$$

e l'unica possibilità per avere 0 è che $\sum w_j \phi_j(x)$ sia costante e dunque nullo (poiché nullo ai bordi). Dunque, A è definita positiva. □

La matrice A, che risulta essere simmetrica e definita positiva, si chiama matrice di rigidezza *(stiffness matrix)* e il vettore (g, ϕ_i) vettore di carico *(load vector)*. Vale poi il seguente risultato:

Teorema 6. *Se u è soluzione di (14.3) e $\hat{u}$ di (14.4), allora*

$$\|u - \hat{u}\| \leq \inf_{v \in V_m} \|u - v\| \tag{14.5}$$

ove $\|v\| = \sqrt{(v', v')}$.

Dimostrazione. Dalle uguaglianze

$$\begin{aligned}(u', v') &= (g, v) \quad \forall v \in V \text{ e, dunque, } \forall v \in V_m \\ (\hat{u}', v') &= (g, v) \quad \forall v \in V_m\end{aligned}$$

si ricava $((u - \hat{u})', v') = 0$ per ogni $v \in V_m$. Dunque, se $v \in V_m$, allora $v - \hat{u} \in V_m$ e quindi

$$\begin{aligned}((u - \hat{u})', (u - \hat{u})') = ((u - \hat{u})', (u - v + v - \hat{u})') = ((u - \hat{u}), (u - v)') \leq \\ \leq \|u - \hat{u}\| \|u - v\|\end{aligned}$$

(per la disuguaglianza di Cauchy–Schwartz) da cui

$$\|u - \hat{u}\| \leq \|u - v\|, \quad \forall v \in V_m$$

e quindi la tesi. □

Per definizione, $\hat{u}$ è allora la proiezione ortogonale della soluzione esatta u sul sottospazio V_m, tramite il prodotto scalare $\langle u, v \rangle = (u', v')$.

Stabilità e consistenza

La consistenza del metodo di Galerkin discende dal fatto che se $u \in V$, allora

$$(u', v') = (g, v), \quad \forall v \in V_m$$

(il metodo si dice *fortemente* consistente). Per quanto riguarda la stabilità, cominciamo ad osservare che se $\hat{u}$ soddisfa (14.4), allora

$$\left| \int_0^1 2x\hat{u}(x)\hat{u}'(x)\mathrm{d}x \right| \leq 2 \left| \int_0^1 \hat{u}(x)\hat{u}'(x)\mathrm{d}x \right| \leq 2\sqrt{(\hat{u}, \hat{u})}\sqrt{(\hat{u}', \hat{u}')}$$

per la monotonia degli integrali ($x \leq 1$ in $[0, 1]$) e la disuguaglianza di Cauchy–Schwartz e

$$\int_0^1 2x\hat{u}(x)\hat{u}'(x)\mathrm{d}x = \int_0^1 x\hat{u}^2(x)'\mathrm{d}x = \hat{u}^2(x)x\Big|_0^1 - \int_0^1 \hat{u}^2(x)\mathrm{d}x$$

da cui

$$(\hat{u}, \hat{u}) \leq 2\sqrt{(\hat{u}, \hat{u})}\sqrt{(\hat{u}', \hat{u}')} = 2\sqrt{(\hat{u}, \hat{u})}\sqrt{(\hat{u}', \hat{u}')}$$

cioè

$$\sqrt{(\hat{u}, \hat{u})} \leq 2\|\hat{u}\|$$

Poiché $\hat{u}$ soddisfa, in particolare,

$$(\hat{u}', \hat{u}') = (g, \hat{u})$$

si ricava, *suppondendo* g a quadrato sommabile,

$$\|\hat{u}\|^2 \leq \sqrt{(g,g)}\sqrt{(\hat{u},\hat{u})} \leq 2\sqrt{(g,g)}\|\hat{u}\|$$

da cui

$$\|\hat{u}\| \leq 2\sqrt{(g,g)}$$

Si conclude osservando che date due perturbazioni della soluzione $\tilde{u}$ e $\bar{u}$ corrispondenti rispettivamente a $\tilde{g}$ e $\bar{g}$, allora

$$((\tilde{u} - \bar{u})', v') = (\tilde{g} - \bar{g}, v), \quad \forall v \in V_m$$

e pertanto

$$\|\tilde{u} - \bar{u}\| \leq 2\sqrt{(\tilde{g} - \bar{g}, \tilde{g} - \bar{g})}$$

e cioè che piccole variazioni sui dati producono piccole variazioni sulle soluzioni.

Metodo degli elementi finiti (FEM)

La scelta di V_m caratterizza il metodo. Da un lato bisogna considerare la regolarità della soluzione richiesta. Dall'altro la difficoltà di *assemblare* la matrice di rigidezza e di risolvere il sistema lineare. Vediamo un esempio. Introduciamo una discretizzazione dell'intervallo $[0, 1]$ a passo *variabile*, come

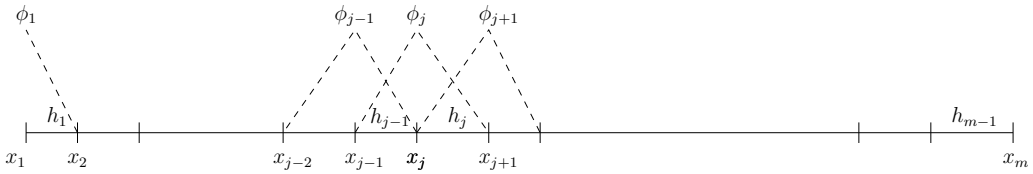


Figura 14.1: Hat functions

in Figura 14.1. Lo spazio V_m è generato dalle funzioni di base $\{\phi_j\}_{j=2}^{m-1}$, le quali sono definite da

$$\phi_j(x) = \begin{cases} \dfrac{x - x_{j-1}}{h_{j-1}}, & x_{j-1} \leq x \leq x_j \\ \dfrac{x_{j+1} - x}{h_j}, & x_j \leq x \leq x_{j+1} \\ 0, & \text{altrimenti} \end{cases}$$

e

$$\phi_j'(x) = \begin{cases} \dfrac{1}{h_{j-1}}, & x_{j-1} < x < x_j \\ -\dfrac{1}{h_j}, & x_j < x < x_{j+1} \\ 0, & \text{altrimenti} \end{cases}$$

Tuttavia, per permettere la trattazione di problemi con differenti condizioni al bordo, consideriamo anche

$$\phi_1(x) = \begin{cases} \dfrac{x_2 - x}{h_1}, & x_1 \le x \le x_2 \\ 0, & \text{altrimenti} \end{cases}$$

e

$$\phi_1'(x) = \begin{cases} -\dfrac{1}{h_1}, & x_1 < x < x_2 \\ 0, & \text{altrimenti} \end{cases}$$

e

$$\phi_m(x) = \begin{cases} \dfrac{x - x_{m-1}}{h_{m-1}}, & x_{m-1} \le x \le x_m \\ 0, & \text{altrimenti} \end{cases}$$

e

$$\phi_m'(x) = \begin{cases} \dfrac{1}{h_{m-1}}, & x_{m-1} < x < x_m \\ 0, & \text{altrimenti} \end{cases}$$

Dunque, nell'approssimazione

$$\hat{u}(x) = \sum_{j=1}^{m} \hat{u}_j \phi_j(x)$$

i coefficienti $\hat{u}_j$ sono i valori di $\hat{u}$ nei nodi x_j. Il problema (14.4) si riscrive

$$\int_0^1 \hat{u}'(x)\phi_i'(x)\mathrm{d}x = \sum_{j=1}^{m} \hat{u}_j \int_0^1 \phi_j'(x)\phi_i'(x)\mathrm{d}x = \sum_{j=1}^{m} \hat{u}_j \int_{x_i - h_{i-1}}^{x_i + h_i} \phi_j'(x)\phi_i'(x)\mathrm{d}x =$$
$$= \sum_{j=1}^{m} \hat{u}_j a_{ij} = \int_{x_i - h_{i-1}}^{x_i + h_i} g(x)\phi_i(x)\mathrm{d}x$$

Siccome il supporto di $\phi_j(x)$ è $[x_{j-1}, x_{j+1}]$, gli unici elementi non nulli a_{ij} sono a_{ii}, $a_{i\,i-1}$ e $a_{i+1\,i} = a_{i\,i+1}$. Per $1 < i < m$,

$$a_{ii} = (\phi_i', \phi_i') = \int_{x_i - h_{i-1}}^{x_i} \left(\frac{1}{h_{i-1}}\right)^2 \mathrm{d}x + \int_{x_i}^{x_i + h_i} \left(-\frac{1}{h_i}\right)^2 \mathrm{d}x = \frac{1}{h_{i-1}} + \frac{1}{h_i}$$

$$a_{i\,i-1} = (\phi_{i-1}', \phi_i') = \int_{x_i - h_{i-1}}^{x_i} -\frac{1}{h_{i-1}} \cdot \frac{1}{h_{i-1}} \mathrm{d}x = -\frac{1}{h_{i-1}} = a_{i-1\,i}$$

Per $i = 1$ e $i = m$, si ha invece

$$a_{11} = \int_{x_1}^{x_1 + h_1} \left(-\frac{1}{h_1}\right)^2 \mathrm{d}x = \frac{1}{h_1}$$

$$a_{21} = \int_{x_2 - h_1}^{x_2} -\frac{1}{h_1} \cdot \frac{1}{h_1} \mathrm{d}x = -\frac{1}{h_1} = a_{12}$$

$$a_{m\,m-1} = \int_{x_m - h_{m-1}}^{x_m} -\frac{1}{h_{m-1}} \cdot \frac{1}{h_{m-1}} \mathrm{d}x = -\frac{1}{h_{m-1}} = a_{m-1\,m}$$

$$a_{mm} = \int_{x_m - h_{m-1}}^{x_m} \left(-\frac{1}{h_{m-1}}\right)^2 \mathrm{d}x = \frac{1}{h_{m-1}}$$

Per quanto riguarda il calcolo di (g, ϕ_i) si può ricorrere alla formula del punto medio: per $1 < i < m$ è

$$g_i = (g, \phi_i) = \int_{x_{i-1}}^{x_i} g(x) \frac{x - x_{i-1}}{h_{i-1}} \mathrm{d}x + \int_{x_i}^{x_{i+1}} g(x) \frac{x_{i+1} - x}{h_i} \mathrm{d}x \approx$$

$$\approx g\left(\frac{x_{i-1} + x_i}{2}\right) \frac{h_{i-1}}{2} + g\left(\frac{x_i + x_{i+1}}{2}\right) \frac{h_i}{2}$$

Per $i = 1$ e $i = m$ si ha invece

$$g_1 = (g, \phi_1) = \int_{x_1}^{x_2} g(x) \frac{x_2 - x}{h_1} \mathrm{d}x \approx g\left(\frac{x_1 + x_2}{2}\right) \frac{h_1}{2}$$

$$g_m = (g, \phi_m) = \int_{x_{m-1}}^{x_m} g(x) \frac{x - x_{m-1}}{h_{m-1}} \mathrm{d}x \approx g\left(\frac{x_{m-1} + x_m}{2}\right) \frac{h_{m-1}}{2}$$

L'approssimazione di

$$\int_{x_{i-1}}^{x_i} g(x) \phi_i(x) \mathrm{d}x = \int_{x_{i-1}}^{x_i} g(x) \frac{x - x_{i-1}}{h_{i-1}} \mathrm{d}x$$

mediante la formula del punto medio produce un errore

$$\left| \frac{h_{i-1}^3}{24} \left(g''(\xi_{i-1}) \frac{\xi_{i-1} - x_{i-1}}{h_{i-1}} + \frac{2g'(\xi_{i-1})}{h_{i-1}} \right) \right| = \mathcal{O}(h_{i-1}^2), \quad \xi_{i-1} \in (x_{i-1}, x_i)$$

(occorre infatti valutare la derivata seconda di $g(x)\phi_i(x)$ in un opportuno punto ξ_{i-1}). Siccome

$$g\left(\frac{x_{i-1}+x_i}{2}\right) = \frac{g(x_{i-1})+g(x_i)}{2} + \mathcal{O}(h_{i-1}^2) = \bar{g}_{i-1} + \mathcal{O}(h_{i-1}^2)$$

e, essendo la formula del punto medio esatta sulle funzioni lineari,

$$\int_{x_{i-1}}^{x_i} \phi_i(x)\mathrm{d}x = \phi_i\left(\frac{x_{i-1}+x_i}{2}\right) h_{i-1} = \frac{h_{i-1}}{2}$$

la formula del punto medio viene di solito sostituita dalla formula equivalente (nel senso dell'ordine di approssimazione)

$$g_i = (g,\phi_i) \approx \bar{g}_{i-1}\int_{x_i-h_{i-1}}^{x_i} \phi_i(x)\mathrm{d}x + \bar{g}_i\int_{x_i}^{x_i+h_i} \phi_i(x)\mathrm{d}x = \bar{g}_{i-1}\frac{h_{i-1}}{2} + \bar{g}_i\frac{h_i}{2}$$

per $1 < i < m$ e da

$$g_1 = (g,\phi_1) = \bar{g}_1\int_{x_1}^{x_1+h_1} \phi_1(x)\mathrm{d}x = \bar{g}_1\frac{h_1}{2}$$
$$g_m = (g,\phi_m) = \bar{g}_{m-1}\int_{x_m-h_{m-1}}^{x_m} \phi_m(x)\mathrm{d}x = \bar{g}_{m-1}\frac{h_{m-1}}{2}$$

La riga i-esima del sistema lineare risulta dunque essere

$$\begin{bmatrix} 0 & \dots & 0 & -\frac{1}{h_{i-1}} & \left(\frac{1}{h_{i-1}}+\frac{1}{h_i}\right) & -\frac{1}{h_i} & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \vdots \\ \hat{u}_{i-1} \\ \hat{u}_i \\ \hat{u}_{i+1} \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \frac{\bar{g}_{i-1}h_{i-1}+\bar{g}_i h_i}{2} \\ \vdots \end{bmatrix}$$

e dunque *molto simile* (il termine noto è diverso, anche se dello stesso ordine) a quella della discretizzazione con differenze finite del secondo ordine. È importante sottolineare che la similitudine con le differenze finite si ha *solo* per questo semplice problema modello, per la scelta delle funzioni di base e per la scelta della formula di quadratura. Pertanto, è naturale aspettarsi, sotto opportune ipotesi di regolarità, che l'errore, nella norma indotta dal prodotto scalare, rispetto alla soluzione analitica tenda a zero come h^2, $h = \max_j h_j$ (e ciò giustifica, a posteriori, la scelta della formula di quadratura). Si ha infatti, nell'intervallo $[x_j, x_{j+1}]$

$$|u'(x) - (p_1^{\mathrm{c}}u)'(x)| = \left|\int_{z_j}^{x} (u''(s) - (p_1^{\mathrm{c}}u)''(s))\mathrm{d}s\right| = \left|\int_{z_j}^{x} u''(s)\mathrm{d}s\right| \leq h_j\|u''\|_{[x_j,x_{j+1}]}$$

(per il teorema di Rolle) e quindi

$$\|u - p_1^c u\|^2 = \int_0^1 |u'(x) - (p_1^c u)'(x)|^2 \mathrm{d}x = \sum_{j=1}^{m-1} \int_{x_j}^{x_{j+1}} |u'(x) - (p_1^c u)'(x)|^2 \mathrm{d}x \leq$$
$$\leq \sum_{j=1}^{m-1} h_j (h_j \|u''\|_{[x_j, x_{j+1}]})^2 \leq \sum_{j=1}^{m-1} h_j (h^2 \max_k \|u''\|^2_{[x_k, x_{k+1}]}) =$$
$$= h^2 \max_k \|u''\|^2_{[x_k, x_{k+1}]}$$

Dunque, usando (14.5), si ha

$$\|u - \hat{u}\| \leq h \max_j \|u''\|_{x_j, x_{j+1}}$$

Questa è una stima che coinvolge la derivata prima della funzione. Se consideriamo ancora l'intervallo $[x_j, x_{j+1}]$, possiamo scrivere

$$|u(x) - p_1^c u(x)| = \left| \int_{x_j}^{x} (u'(s) - (p_1^c u)'(s)) \mathrm{d}s \right| \leq h_j \cdot h_j \|u''\|_{[x_j, x_{j+1}]}$$

e quindi

$$\|u - p_1^c u\|^2_{L^2} = \int_0^1 |u(x) - p_1^c u(x)|^2 \mathrm{d}x = \sum_{j=1}^{m-1} \int_{x_j}^{x_{j+1}} |u(x) - p_1^c u(x)|^2 \mathrm{d}x \leq$$
$$\leq \sum_{j=1}^{m-1} h_j (h^4 \|u''\|^2_{[x_j, x_{j+1}]}) = h^4 \max_k \|u''\|^2_{[x_k, x_{k+1}]}$$

da cui

$$\|u - \hat{u}\|_{L^2} \leq h^2 \max_j \|u''\|^2_{[x_k, x_{k+1}]}$$

Si noti come nelle stime sopra sia sufficiente che la derivata seconda della soluzione esista in ogni intervallo, ma non sia necessario che esista dappertutto.

A questo punto si risolve il sistema lineare, dopo aver opportunamente modificato la matrice e il termine noto per imporre le condizioni al bordo di Dirichlet.

Nel caso di condizioni di Neumann (per esempio in $u'(0) = u'_0$), la forma debole del problema è

$$-\hat{u}'(x)\phi_i(x)\Big|_0^1 + \int_0^1 \hat{u}'(x)\phi_i'(x)\mathrm{d}x = \int_0^1 g(x)\phi_i(x)\mathrm{d}x, \quad 1 \leq i \leq m$$

Per $i = 1$, che è il caso di interesse, si ha

$$\hat{u}'(0) + \int_0^1 \hat{u}'(x)\phi_1'(x)\mathrm{d}x = \int_0^1 g(x)\phi_1(x)\mathrm{d}x$$

Dunque, la prima riga del sistema lineare da risolvere è

$$\int_0^1 \hat{u}'(x)\phi_1'(x)\mathrm{d}x = -u_0' + \int_0^1 g(x)\phi_1(x)\mathrm{d}x$$

Da notare che il problema con due condizioni di Neumann non è ben definito, in quanto se $u(x)$ è soluzione, allora lo è anche $u(x) + k$.

Ovviamente, lo spazio V_m può essere costituito da funzioni molto più regolari (per esempio polinomi di grado superiore).

Vediamo un approccio più generale all'implementazione del metodo degli elementi finiti. Supponiamo di avere l elementi $\{\ell_j\}_{j=1}^{l}$ (nel caso unidimensionale, sono gli intervalli) ad ognuno dei quali sono associati due nodi. Con

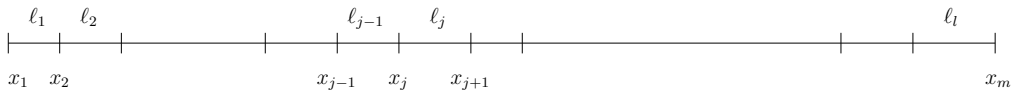


Figura 14.2: Nodi (numerati in basso) ed elementi (numerati in alto).

riferimento alla Figura 14.2, ove $m = l + 1$, si ha

$$\ell_{j,1} = j, \ \ell_{j,2} = j + 1, \quad 1 \le j \le l$$

che significa che l'elemento ℓ_j ha associati i nodi x_j e x_{j+1}. Ogni elemento contribuisce all'*assemblaggio* della matrice di stiffness e del termine noto. Per quanto riguarda la matrice di stiffness, ad ogni elemento ℓ_j si associa

$$\phi_{\ell_{j,k}\ell_{j,k}} = \int_{\ell_j} \phi_{\ell j,k}'(x)\phi_{\ell j,k}'(x)\mathrm{d}x = \frac{1}{h_j}, \quad k = 1, 2$$

$$\phi_{\ell_{j,k}\ell_{j,3-k}} = \int_{\ell_j} \phi_{\ell j,k}'(x)\phi_{\ell j,3-k}'(x)\mathrm{d}x = -\frac{1}{h_j}, \quad k = 1, 2$$

Per quanto riguarda il termine noto, ad ogni elemento ℓ_j si associa

$$g_{\ell_j} = \bar{g}_j = \frac{g(x_{\ell_{j,1}}) + g(x_{\ell_{j,2}})}{2}$$

Pertanto si ha

- $a_{ij} = 0$, $1 \le i, j \le m$, $g_i = 0$, $1 \le i \le m$

- FOR $j = 1, \ldots, l$

  FOR $k = 1, 2$

  $a_{\ell_{j,k}\ell_{j,k}} = a_{\ell_{j,k}\ell_{j,k}} + \phi_{\ell_{j,k}\ell_{j,k}}$, $g_{\ell_{j,k}} = g_{\ell_{j,k}} + g_{\ell_j}\frac{h_j}{2}$

  $a_{\ell_{j,k}\ell_{j,3-k}} = a_{\ell_{j,k}\ell_{j,3-k}} + \phi_{\ell_{j,k}\ell_{j,3-k}}$

  END

END

14.1.2 Estensione al caso bidimensionale

Tutto quanto detto si estende, in particolare, al caso bidimensionale. Si deve usare la formula di Green

$$\int_\Omega \nabla^2 u(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = -\int_\Omega \nabla u(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} v(s)\nabla u(s)\cdot\nu(s)\mathrm{d}s$$

ove $\nu(s)$ è il versore esterno normale a $\partial\Omega$.

14.2 Metodi spettrali

Sia

$$u(x) = \sum_j u_j\phi_j(x)$$

L'*indice algebrico* di convergenza è il più grande k tale che

$$\lim_{j\to\infty} |u_j| j^k < +\infty$$

Se tale limite è finito per ogni k, allora la serie si dice convergere *esponenzialmente* oppure *spettralmente*. Significa che $|u_j|$ decade più velocemente di ogni potenza negativa di j. Parleremo di *metodi spettrali* quando useremo un'approssimazione di una serie convergente spettralmente

$$\hat{u}(x) = \sum_{j=1}^{m} \hat{u}_j\phi_j(x)$$

al posto di $u(x)$. Quindi, useremo i metodi spettrali quando ci aspettiamo soluzioni molto regolari.

Consideriamo una famiglia di funzioni $\{\phi_j\}_j$ ortonormali rispetto al prodotto scalare

$$\int_a^b \phi_j(x)\phi_i(x)w(x)\mathrm{d}x = \delta_{ji}$$

La formulazione di Galerkin di un problema ai limiti $Lu = g$, L operatore differenziale *lineare*, diventa

$$\sum_{j=1}^{m} \hat{u}_j \int_a^b L\phi_j(x)\phi_i(x)w(x)\mathrm{d}x = \int_a^b g(x)\phi_i(x)w(x)\mathrm{d}x, \quad 1 \leq i \leq m$$

Nel caso non si possano calcolare analiticamente o con formule di quadratura esatte gli integrali, si ricorre alle formule di quadratura gaussiana (relative alle funzioni $\phi_j(x)$) a m punti, dando origine al sistema lineare

$$\sum_{j=1}^{m} \hat{u}_j \left(\sum_{n=1}^{m} L\phi_j(x_n)\phi_i(x_n)w_n \right) = \sum_{n=1}^{m} g(x_n)\phi_i(x_n)w_n = \hat{g}_i, \quad 1 \leq i \leq m \tag{14.6}$$

In tal caso si parla di metodi *pseudospettrali*. I coefficienti $\hat{u}_j$ che si trovano risolvendo il sistema lineare si chiamano solitamente soluzione *nello spazio spettrale*. Dati i coefficienti, si ricostruisce la soluzione *nello spazio fisico* $\sum_j \hat{u}_j\phi_j(x)$. Detta F la matrice

$$F = (f_{ij}) = \phi_i(x_j)\sqrt{w_j}$$

e dati i coefficienti $\hat{u}_j$, si calcola la soluzione fisica $\sum_j \hat{u}_j\phi_j(x)$ sui nodi di quadratura mediante

$$\begin{bmatrix} \sqrt{w_1}\hat{u}(x_1) \\ \sqrt{w_2}\hat{u}(x_2) \\ \vdots \\ \sqrt{w_n}\hat{u}(x_n) \end{bmatrix} = F^{\mathrm{T}} \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_m \end{bmatrix}$$

Per quanto riguarda invece i coefficienti g_i, essi sono dati da

$$\begin{bmatrix} \hat{g}_1 \\ \hat{g}_2 \\ \vdots \\ \hat{g}_n \end{bmatrix} = F \begin{bmatrix} g(x_1)\sqrt{w_1} \\ g(x_2)\sqrt{w_2} \\ \vdots \\ g(x_n)\sqrt{w_n} \end{bmatrix}$$

Solitamente le funzioni $\{\phi_j(x)\}_j$ sono polinomi ortonormali rispetto alla funzione peso $w(x)$ e in tal caso, ma non solo in questo, $F^{\mathrm{T}} = F^{-1}$ (lo si dimostra calcolando FF^{T} e usando l'equazione (3.1) con $p_j(x) = \phi_{j+1}(x)$). La soluzione $u(x)$, però, potrebbe non essere efficacemente approssimata da polinomi, per esempio se deve soddisfare particolari condizioni al contorno (tipo *vanishing boundary conditions, condizioni al bordo periodiche* o altro). Può essere

utile allora la decomposizione

$$u(x) \approx \sum_{j=1}^{m} \hat{u}_j \phi_j(x) \sqrt{w(x)} = \sum_{j=1}^{m} \hat{u}_j \varphi_j(x)$$

La formulazione di Galerkin di $Lu = g$ diventa allora

$$\sum_{j=1}^{m} \hat{u}_j \int_a^b L(\phi_j(x)\sqrt{w(x)})\phi_i(x)\sqrt{w(x)}\mathrm{d}x = \int_a^b g(x)\phi_i(x)\sqrt{w(x)}\mathrm{d}x,\ 1 \le i \le m$$

Ad ogni modo, la differenza sostanzionale con il metodo degli elementi finiti è che le funzioni di base sono regolari e a supporto globale.

Consideriamo ora un caso particolare di fondamentale importanza (per l'analisi numerica in generale). Molte proprietà risultano comuni anche agli altri metodi pseudospettrali.

14.2.1 Trasformata di Fourier

Sia $[a, b]$ un intervallo di $\mathbb{R}$, $m > 0$ pari e fissato. Consideriamo, per ogni $j \in \mathbb{Z}$,

$$\phi_j(x) = \frac{\mathrm{e}^{\mathrm{i}(j-1-m/2)2\pi(x-a)/(b-a)}}{\sqrt{b-a}}$$

Allora,

$$\int_a^b \phi_j(x)\overline{\phi_k(x)}\mathrm{d}x = \delta_{jk} \tag{14.7}$$

Infatti, se $j = k$ allora $\phi_j(x)\overline{\phi_k(x)} = 1/(b-a)$, altrimenti

$$\phi_j(x)\overline{\phi_k(x)} = \frac{\mathrm{e}^{\mathrm{i}2\pi(j-k)(x-a)/(b-a)}}{b-a}$$

e quindi

$$\int_a^b \phi_j(x)\overline{\phi_k(x)}\mathrm{d}x = \int_0^1 \frac{\mathrm{e}^{\mathrm{i}2\pi(j-k)y}}{b-a}(b-a)\mathrm{d}y = 0\ ,$$

poiché l'integrale delle funzioni sin e cos in un intervallo multiplo del loro periodo è nullo. La famiglia di funzioni $\{\phi_j\}_j$ si dice *ortonormale* nell'intervallo $[a, b]$ rispetto al prodotto scalare

$$(\phi_j, \phi_k) = \int_a^b \phi_j(x)\overline{\phi_k(x)}\mathrm{d}x$$

Un risultato utile è il seguente

$$\sum_{n=1}^{m} \mathrm{e}^{\mathrm{i}(n-1)2\pi(j-k)/m} = m\delta_{jk}, \quad 1 \le j,k \le m \tag{14.8}$$

È ovvio per $j = k$; altrimenti

$$\begin{aligned}\sum_{n=1}^{m} \mathrm{e}^{\mathrm{i}(n-1)2\pi(j-k)/m} &= \sum_{n=0}^{m-1} \left(\mathrm{e}^{\mathrm{i}2\pi(j-k)/m}\right)^n = \\ &= \frac{1-\mathrm{e}^{\mathrm{i}2\pi(j-k)}}{1-\mathrm{e}^{\mathrm{i}2\pi(j-k)/m}} = \frac{1-\cos(2\pi(j-k))}{1-\mathrm{e}^{\mathrm{i}2\pi(j-k)/m}} = 0\end{aligned}$$

poiché $-m+1 \le j-k \le m-1$.

14.2.2 Trasformata di Fourier discreta

Sia u una funzione da $[a,b]$ a $\mathbb{C}$ tale che $u(a) = u(b)$. Supponiamo che u si possa scrivere (ciò è vero, per esempio, per funzioni di classe C^1) come

$$u(x) = \sum_{j=-\infty}^{+\infty} u_j \phi_j(x), \quad u_j \in \mathbb{C} \tag{14.9}$$

Fissato $k \in \mathbb{Z}$, moltiplicando entrambi i membri per $\overline{\phi_k(x)}$ e integrando nell'intervallo $[a,b]$, usando (14.7) si ottiene

$$\begin{aligned}\int_a^b u(x)\overline{\phi_k(x)}\mathrm{d}x &= \int_a^b \left(\sum_{j=-\infty}^{+\infty} u_j \phi_j(x)\overline{\phi_k(x)}\right)\mathrm{d}x = \\ &= \sum_{j=-\infty}^{\infty} u_j \int_a^b \phi_j(x)\overline{\phi_k(x)}\mathrm{d}x = u_k\end{aligned} \tag{14.10}$$

Dunque, abbiamo un'espressione esplicita per u_j. Analogamente si vede che

$$\int_a^b |u(x)|^2 \mathrm{d}x = \sum_{j=-\infty}^{+\infty} |u_j|^2 \quad \text{(identità di Parseval)}$$

La prima approssimazione da fare consiste nel troncare la serie infinita. Osserviamo che, definito $J = \mathbb{Z} \setminus \{1, 2, \dots, m\}$,

$$\begin{aligned}\int_a^b \left| u(x) - \sum_{j=1}^{m} u_j \phi_j(x) \right|^2 \mathrm{d}x &= \int_a^b \left| \sum_{j \in J} u_j \phi_j(x) \right|^2 \mathrm{d}x = \\ &= \int_a^b \left(\sum_{j \in J} u_j \phi_j(x) \right) \left(\sum_{k \in J} \overline{u_k} \overline{\phi_k(x)} \right) \mathrm{d}x = \\ &= \sum_{j \in J} |u_j|^2\end{aligned}$$

Stimiamo adesso u_j: posto $\lambda_j = \mathrm{i}(j - 1 - m/2)2\pi/(b-a)$ si ha, per funzioni di classe $u \in \mathcal{C}^2$, integrando per parti e tenendo conto che $\overline{\phi_j(a)} = \overline{\phi_j(b)} = 1/\sqrt{b-a}$ e che $\phi_j'(x) = \lambda_j \phi(x)$

$$\begin{aligned}u_j &= \int_a^b u(x)\overline{\phi_j(x)}\mathrm{d}x = -\frac{1}{\lambda_j \sqrt{b-a}} \left(u(b) - u(a)\right) + \frac{1}{\lambda_j} \int_a^b u'(x)\overline{\phi_j(x)}\mathrm{d}x = \\ &= -\frac{1}{\lambda_j \sqrt{b-a}} \left(u(b) - u(a)\right) - \frac{1}{\lambda_j^2 \sqrt{b-a}} \left(u'(b) - u'(a)\right) + \frac{1}{\lambda_j^2} \int_a^b u''(x)\overline{\phi_j(x)}\mathrm{d}x \\ &= \mathcal{O}\left((j - 1 - m/2)^{-2}\right)\end{aligned}$$

Se anche $u'(a) = u'(b)$ e $u \in \mathcal{C}^3$, allora, integrando ancora per parti, si ottiene $|u_j| = \mathcal{O}\left(|j - 1 - m/2|^{-3}\right)$ e così via: in generale, se $u \in \mathcal{C}^k$ e $u^{(l)}$ periodica per $l = 1, 2, \dots, k-2$, allora

$$|u_j| = \mathcal{O}\left(|j - 1 - m/2|^{-k}\right), \quad j \to \infty$$

Si ha dunque

$$\begin{aligned}\sum_{j \in J} |u_j|^2 &= |u_{m+1}|^2 + \left(|u_0|^2 + |u_{m+2}|^2\right) + \left(|u_{-1}|^2 + |u_{m+3}|^2\right) + \dots = \\ &= \mathcal{O}\left((m/2)^{-2k}\right) + \mathcal{O}\left((1 + m/2)^{-2k}\right) + \mathcal{O}\left((2 + m/2)^{-2k}\right) + \dots\end{aligned}$$

con $|u_{m+1}|$ termine dominante.

Se poi u è infinitamente derivabile e *periodica* (cioè tutte le derivate sono periodiche), allora, fissato k,

$$0 \le \lim_{j \to \infty} |u_j| j^k \le \lim_{j \to \infty} M \left(|j - 1 - m/2|^{-(k+1)}\right) j^k = 0$$

La seconda approssimazione da fare è utilizzare una formula di quadratura per il calcolo di u_j. Riportiamo per comodità la formula di quadratura

trapezoidale a $m+1$ nodi equispaziati $x_n = (b-a)y_n + a$, ove $y_n = (n-1)/m$, $n = 1, \ldots, m+1$ per funzioni periodiche:

$$\int_a^b g(x)\mathrm{d}x \approx \frac{b-a}{2m}\left(g(x_1) + 2\sum_{n=2}^{m} g(x_n) + g(x_{m+1})\right) = h\sum_{n=1}^{m} g(x_n)$$

ove $h = (b-a)/m$. Usando la (14.8), abbiamo

$$\begin{aligned} m\delta_{jk} &= \sum_{n=1}^{m} \mathrm{e}^{\mathrm{i}(n-1)2\pi(j-k)/m} = \sum_{n=1}^{m} \mathrm{e}^{\mathrm{i}(j-k)2\pi y_n} = \sum_{n=1}^{m} \mathrm{e}^{\mathrm{i}(j-k)2\pi(x_n-a)/(b-a)} = \\ &= (b-a)\sum_{n=1}^{m} \frac{\mathrm{e}^{\mathrm{i}(j-1-m/2)2\pi(x_n-a)/(b-a)}}{\sqrt{b-a}} \frac{\mathrm{e}^{-\mathrm{i}(k-1-m/2)2\pi(x_n-a)/(b-a)}}{\sqrt{b-a}} = \\ &= (b-a)\sum_{n=1}^{m} \phi_j(x_n)\overline{\phi_k(x_n)} = m\int_a^b \phi_j(x)\overline{\phi_k(x)}\mathrm{d}x \end{aligned}$$

cioè la famiglia $\{\phi_j\}_{j=1}^m$ è ortonormale anche rispetto al prodotto scalare *discreto*

$$(\phi_j, \phi_k)_{\mathrm{d}} = \frac{b-a}{m}\sum_{n=1}^{m} \phi_j(x_n)\overline{\phi_k(x_n)}$$

Applicando la formula di quadratura ai coefficienti (14.10) si ottiene

$$\begin{aligned} u_j &= \int_a^b u(x)\frac{\mathrm{e}^{-\mathrm{i}(j-1-m/2)2\pi(x-a)/(b-a)}}{\sqrt{b-a}}\mathrm{d}x = \\ &= \sqrt{b-a}\int_0^1 u((b-a)y+a)\mathrm{e}^{-\mathrm{i}(j-1)2\pi y}\mathrm{e}^{\mathrm{i}m\pi y}\mathrm{d}y \approx \\ &\approx \frac{\sqrt{b-a}}{m}\boxed{\sum_{n=1}^{m}\left(u(x_n)\mathrm{e}^{\mathrm{i}m\pi y_n}\right)\mathrm{e}^{-\mathrm{i}(j-1)2\pi y_n}} = \hat{u}_j \end{aligned}$$

ove $x = (b-a)y + a$.

La funzione *(serie troncata di Fourier)*

$$\begin{aligned} \hat{u}(x) &= \sum_{j=1}^{m} \hat{u}_j\phi_j(x) = \sum_{k=-m/2}^{m/2-1} \hat{u}_{k+1+m/2}\phi_{k+1+m/2}(x) = \\ &= \sum_{k=-m/2}^{m/2-1} \hat{u}_{k+1+m/2}\frac{\mathrm{e}^{\mathrm{i}k2\pi(x-a)/(b-a)}}{\sqrt{b-a}} \end{aligned}$$

è un polinomio trigonometrico che approssima $u(x)$ ed è *interpolante* nei nodi x_n. Infatti, usando (14.8),

$$\begin{aligned}\hat{u}(x_n) &= \sum_{j=1}^{m} \hat{u}_j \phi_j(x_n) = \\ &= \sum_{j=1}^{m} \left(\frac{\sqrt{b-a}}{m} \sum_{k=1}^{m} \left(u(x_k) \mathrm{e}^{\mathrm{i}m\pi y_k} \right) \mathrm{e}^{-\mathrm{i}(j-1)2\pi y_k} \right) \frac{\mathrm{e}^{\mathrm{i}(j-1-m/2)2\pi(x_n-a)/(b-a)}}{\sqrt{b-a}} = \\ &= \frac{1}{m} \sum_{k=1}^{m} u(x_k) \mathrm{e}^{\mathrm{i}m\pi(k-1)/m} \mathrm{e}^{-\mathrm{i}m\pi(n-1)/m} \sum_{j=1}^{m} \mathrm{e}^{-\mathrm{i}(j-1)2\pi(k-1)/m} \mathrm{e}^{\mathrm{i}(j-1)2\pi(n-1)/m} = \\ &= \frac{1}{m} \sum_{k=1}^{m} u(x_k) \mathrm{e}^{\mathrm{i}(k-n)\pi} \sum_{j=1}^{m} \mathrm{e}^{\mathrm{i}(j-1)2\pi(n-k)/m} = \frac{1}{m} u(x_n) m = u(x_n) \, .\end{aligned}$$

Si può far vedere poi che

$$\int_a^b \left| u(x) - \sum_{j=1}^{m} \hat{u}_j \phi_j(x) \right|^2 \mathrm{d}x \leq 2 \sum_{j \in J} |u_j|^2$$

Si usa spesso considerare allora il doppio del primo termine trascurato (cioè u_{m+1}), oppure della somma dei primi termini trascurati come stima di errore.

La trasformazione

$$[u(x_1), u(x_2), \ldots, u(x_m)]^{\mathrm{T}} \to [\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_m]^{\mathrm{T}}$$

si chiama *trasformata di Fourier discreta* di u e $\hat{u}_1, \ldots, \hat{u}_m$ *coefficienti di Fourier* di u. Il vettore $m \cdot [\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_m]^{\mathrm{T}} / \sqrt{b-a}$ può essere scritto come prodotto matrice-vettore $F\sqrt{m}[u(x_1)\mathrm{e}^{\mathrm{i}m\pi y_1}, u(x_2)\mathrm{e}^{\mathrm{i}m\pi y_2}, \ldots, u(x_m)\mathrm{e}^{\mathrm{i}m\pi y_m}]^{\mathrm{T}}$, ove

$$F = (f_{jn}), \quad f_{jn} = \frac{\mathrm{e}^{-\mathrm{i}(j-1)2\pi y_n}}{\sqrt{m}} = \overline{\phi_{j+m/2}\left(\frac{x_n - a}{b-a} \right)} \sqrt{h} \, .$$

Alternativamente, si può usare la Fast Fourier Transform (FFT). Il comando `fft` applicato al vettore $[u(x_1)\mathrm{e}^{\mathrm{i}m\pi y_1}, u(x_2)\mathrm{e}^{\mathrm{i}m\pi y_2}, \ldots, u(x_m)\mathrm{e}^{\mathrm{i}m\pi y_m}]^{\mathrm{T}}$ produce il vettore $m \cdot [\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_m]^{\mathrm{T}} / \sqrt{b-a}$, così come il comando `fftshift` applicato al risultato del comando `fft` applicato a $[u(x_1), u(x_2), \ldots, u(x_m)]$.

Dati dei coefficienti $\hat{v}_j$, $j = 1, \ldots, m$, si può considerare la funzione (periodica)

$$\sum_{j=1}^{m} \hat{v}_j \phi_j(x)$$

La valutazione nei nodi x_n, $1 \le n \le m$, porge

$$\hat{\hat{v}}_n = \sum_{j=1}^{m} \hat{v}_j \phi_j(x_n) = \sum_{j=1}^{m} \hat{v}_j \frac{\mathrm{e}^{\mathrm{i}(j-1-m/2)2\pi(x_n-a)/(b-a)}}{\sqrt{b-a}} =$$

$$= \frac{m}{\sqrt{b-a}} \boxed{\frac{1}{m}\left(\sum_{j=1}^{m} \hat{v}_j \mathrm{e}^{\mathrm{i}(j-1)2\pi y_n}\right)} \mathrm{e}^{-\mathrm{i}m\pi y_n} .$$

La trasformazione

$$[\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_m]^{\mathrm{T}} \to [\hat{\hat{v}}_1, \hat{\hat{v}}_2, \ldots, \hat{\hat{v}}_m]^{\mathrm{T}}$$

si chiama *anti-trasformata di Fourier discreta*. Se i $\hat{v}_j$ sono i coefficienti di Fourier di una funzione $v(x)$, la proprietà di interpolazione comporta $\hat{\hat{v}}_n = v(x_n)$. Ma, in generale, non è vero che

$$v(x) = \sum_{j=1}^{m} \hat{v}_j \phi_j(x)$$

Il vettore $\sqrt{b-a} \cdot [\hat{\hat{v}}_1 \mathrm{e}^{\mathrm{i}m\pi y_1}, \hat{\hat{v}}_2 \mathrm{e}^{\mathrm{i}m\pi y_2}, \ldots, \hat{\hat{v}}_m \mathrm{e}^{\mathrm{i}m\pi y_m}]^{\mathrm{T}}/m$ può essere scritto come prodotto matrice-vettore $F'[\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_m]^{\mathrm{T}}/\sqrt{m}$, ove F' denota, come in GNU Octave, la trasposta coniugata di F. Alternativamente, il comando `ifft` applicato al vettore $[\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_m]$ produce il vettore $\sqrt{b-a} \cdot [\hat{\hat{v}}_1 \mathrm{e}^{\mathrm{i}m\pi y_1}, \hat{\hat{v}}_2 \mathrm{e}^{\mathrm{i}m\pi y_2}, \ldots, \hat{\hat{v}}_m \mathrm{e}^{\mathrm{i}m\pi y_m}]/m$, mentre, se applicato al risultato del comando `ifftshift` applicato al vettore $[\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_m]$, produce il vettore $\sqrt{b-a} \cdot [\hat{\hat{v}}_1, \hat{\hat{v}}_2, \ldots, \hat{\hat{v}}_m]/m$.

Da notare che $F' = F^{-1}$ e che, detta $D = \sqrt{h} \cdot \mathrm{diag}(\mathrm{e}^{\mathrm{i}m\pi y_1}, \mathrm{e}^{\mathrm{i}m\pi y_2}, \ldots, \mathrm{e}^{\mathrm{i}m\pi y_m})$, si ha $[\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_m]^{\mathrm{T}} = FD[u(x_1), u(x_2), \ldots, u(x_m)]^{\mathrm{T}}$ e $(FD)^{-1} = D^{-1}F^{-1} = (1/h)D'F' = (1/h)(FD)'$.

Sul range delle frequenze

Con le notazioni attuali la serie troncata di Fourier è definita come

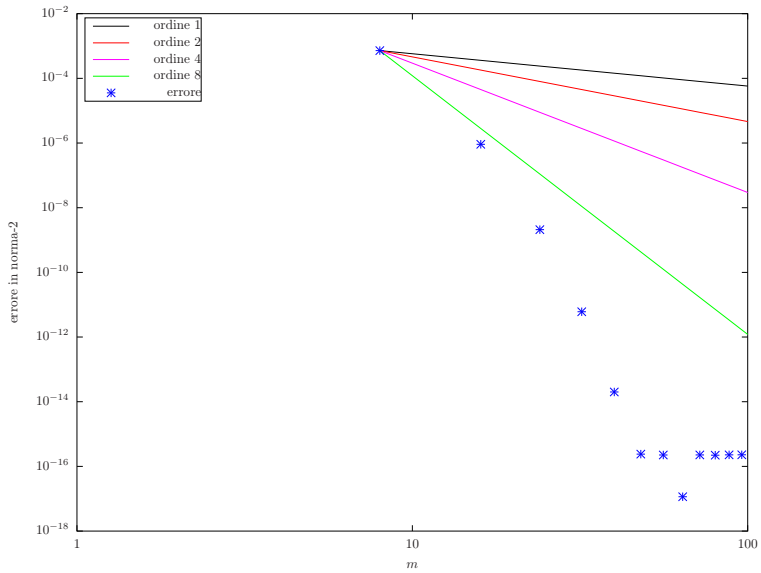
$$\hat{u}(x) = \sum_{k=-m/2}^{m/2-1} \ldots$$

mentre molte volte si vuole calcolare

$$\sum_{k=-m/2}^{m/2} \ldots$$

È sufficiente considerare allora la famiglia di funzioni $\{\phi_j\}_{j=1}^{m+1}$ con nodi $x_n = (b-a)y_n + a$, $y_n = (n-1)/(m+1)$, $n = 1, \ldots, m+2$.

Applicazione ad un problema modello



fourier.m

Figura 14.3: Convergenza spettrale di Fourier per il problema (14.11) rispetto ad una soluzione di riferimento ottenuta con $m = 64$.

Consideriamo la soluzione del problema

$$\begin{cases} -u''(x) + u(x) = \dfrac{1}{\sin x + 2}, & x \in (-\pi, \pi) \\ u(-\pi) = u(\pi) \end{cases} \tag{14.11}$$

mediante decomposizione in funzioni di Fourier. Posto $a = -\pi$, $b = \pi$, $g(x) = 1/(\sin x + 2)$, si ha

$$\phi_j(x) = \frac{\mathrm{e}^{\mathrm{i}(j-1-m/2)2\pi(x-a)/(b-a)}}{\sqrt{b-a}}$$

ove m è pari e fissato e, per $1 \le k \le m$,

$$\int_{-\pi}^{\pi} \left(-\sum_{j=1}^{m} u_j \phi_j(x) \right)'' \overline{\phi_k(x)} \mathrm{d}x + \int_{-\pi}^{\pi} \left(\sum_{j=1}^{m} u_j \phi_j(x) \right) \overline{\phi_k(x)} \mathrm{d}x = \\ = \int_{-\pi}^{\pi} g(x) \overline{\phi_k(x)} \mathrm{d}x$$

da cui

$$-\sum_{j=1}^{m} u_j \int_{-\pi}^{\pi} \phi_j''(x) \overline{\phi_k(x)} \mathrm{d}x + \sum_{j=1}^{m} u_j \int_{-\pi}^{\pi} \phi_j(x) \overline{\phi_k(x)} \mathrm{d}x = \int_{-\pi}^{\pi} g(x) \overline{\phi_k(x)} \mathrm{d}x$$

Poiché

$$\phi_j''(x) = \left(\frac{\mathrm{i}(j-1-m/2)2\pi}{b-a}\right)^2 \phi_j(x) = \lambda_j^2 \phi_j(x)$$

usando l'ortonormalità delle funzioni di Fourier e calcolando i coefficienti di Fourier di $g(x)$, si ha

$$-\lambda_k^2 \hat{u}_k + \hat{u}_k = \hat{g}_k, \quad 1 \leq k \leq m$$

da cui

$$\hat{u}_k = \frac{\hat{g}_k}{1-\lambda_k^2}, \quad 1 \leq k \leq m$$

e quindi

$$u(x) \approx \sum_{j=1}^{m} \hat{u}_j \phi_j(x)$$

In Figura 14.3 si vede la convergenza spettrale del metodo.

Da notare che le condizioni al bordo devono essere di tipo periodico: condizioni come

$$\begin{cases} u''(x) = f(x, u(x), u'(x)), & x \in (a,b) \\ u(a) = 0 \\ u(b) = 0 \end{cases}$$

sono invece di Dirichlet omogenee. Inoltre, la soluzione del problema *deve* poter essere periodica (o, almeno, avere più derivate possibili periodiche): per esempio, non possono esserci termini non omogenei *non periodici*.

Costi computazionali e stabilità

La Fast Fourier Transform di un vettore di lunghezza m ha costo $\mathcal{O}(m \log m)$, mentre il prodotto matrice-vettore $\mathcal{O}(m^2)$. Tali costi sono però asintotici e nascondono i fattori costanti. Inoltre, GNU Octave può far uso di implementazioni ottimizzate di algebra lineare (come, ad esempio, le librerie ATLAS[1] o OpenBLAS[2]). In pratica, dunque, esiste un m_0 sotto il quale conviene, dal punto di vista del costo computazionale, usare il prodotto matrice-vettore e sopra il quale la FFT.

Per quanto riguarda l'accuratezza, in generale la FFT è più precisa del prodotto matrice vettore. Poiché la trasformata di Fourier discreta comporta l'uso di aritmetica complessa (anche se la funzione da trasformare è reale),

[1] http://math-atlas.sourceforge.net/
[2] http://xianyi.github.com/OpenBLAS/

la sequenza trasformata/anti-trasformata potrebbe introdurre una quantità immaginaria spuria che può essere eliminata con il comando `real`.

Anche per la trasformata di Fourier vi possono essere problemi di stabilità simili al fenomeno di Runge (qui chiamato *fenomeno di Gibbs*). Una tecnica per "smussare" (in inglese "to smooth") eventuali oscillazioni, consiste nel moltiplicare opportunamente i coefficienti di Fourier $\hat{u}_j$ per opportuni termini σ_j che decadono in j, per esempio

$$\sigma_j = \frac{\frac{m}{2} + 1 - |\frac{m}{2} + 1 - j|}{\frac{m}{2} + 1}, \quad 1 \le j \le m$$

Il risultato è che il coefficiente $\hat{u}_{m/2+1}$ è pesato da $\sigma_{m/2+1} = 1$, i coefficienti $\hat{u}_{m/2}$ e $u_{m/2+2}$ sono pesati da $m/(m+2)$ e così via fino al coefficiente $\hat{u}_1$ pesato da $2/(m+2)$. Questa scelta corrisponde alle *medie di Cesáro*. Infatti, si sostituisce la serie troncata di Fourier

$$\sum_{j=1}^{m} \hat{u}_j \phi_j(x)$$

con la media delle troncate

$$\frac{\displaystyle\sum_{k=0}^{\frac{m}{2}} \sum_{j=\frac{m}{2}+1-k}^{\max\{\frac{m}{2}+1+k,m\}} \hat{u}_j \phi_j(x)}{\frac{m}{2} + 1}$$

Si ricorda che se una serie è convergente, allora il limite delle medie delle sue troncate è la somma della serie.

Valutazione di un polinomio trigonometrico

Supponiamo di conoscere i coefficienti $\hat{u}_j$, $j = 1, \ldots, m$ e di voler valutare la funzione

$$\hat{u}(x) = \sum_{j=1}^{m} \hat{u}_j \phi_j(x)$$

su un insieme di nodi target x_k equispaziati, $x_k = (k-1)/n$, $1 \le k \le n$, $n > m$, n pari. Si possono introdurre dei coefficienti fittizi $\hat{U}_k$

$$\begin{aligned}
&\hat{U}_k = 0 && 1 \le k \le \frac{n-m}{2} \\
&\hat{U}_k = \hat{u}_{k-\frac{n-m}{2}} && \frac{n-m}{2} + 1 \le k \le m - \frac{n-m}{2} \\
&\hat{U}_k = 0 && m - \frac{n-m}{2} + 1 \le k \le n
\end{aligned}$$

Si avrà

$$\hat{u}(x_k) = \sum_{j=1}^{m} \hat{u}_j \phi_j(x_k) = \sum_{j=1}^{n} \hat{U}_j \frac{\mathrm{e}^{\mathrm{i}(j-1-n/2)2\pi(x_k-a)/(b-a)}}{\sqrt{b-a}} =$$

$$= \frac{n}{\sqrt{b-a}} \boxed{\frac{1}{n}\left(\sum_{j=1}^{n} \hat{U}_j \mathrm{e}^{\mathrm{i}(j-1)2\pi y_k}\right)} \mathrm{e}^{-\mathrm{i}n\pi y_k}$$

Oppure si può costruire la matrice F relativa ai nodi (ciò funziona anche per nodi non equispaziati). Infine, si può usare la trasformata di Fourier non equispaziata NFFT.

14.3 Metodi di collocazione

Si assume comunque

$$\hat{u}(x) = \sum_{j=1}^{m} \hat{u}_j \phi_j(x)$$

ove $\{\phi_j\}$ è un sistema ortonormale rispetto ad un prodotto scalare, ma si impone poi che l'equazione differenziale $Lu = g$ sia soddisfatta in certi nodi x_n. Si ha il seguente risultato interessante:

Teorema 7. *La soluzione del sistema lineare*

$$\sum_{j=1}^{m} \hat{u}_j L\phi_j(x_n) = g(x_n), \quad 1 \le n \le m \tag{14.12}$$

ove gli $\{x_n\}$ sono i nodi della quadratura gaussiana relativa alla famiglia $\{\phi_j\}$ è la stessa del problema di Galerkin

$$\sum_{j=1}^{m} \hat{u}_j \int_a^b L\phi_j(x)\phi_i(x)w(x)\mathrm{d}x = \int_a^b g(x)\phi_i(x)w(x)\mathrm{d}x$$

quando si approssimino gli integrali con le formule gaussiane.

Dimostrazione. Per ogni i, $1 \le i \le m$, da (14.12), si ha

$$\sum_{j=1}^{m} \hat{u}_j L\phi_j(x_n)\phi_i(x_n)w_n = g(x_n)\phi_i(x_n)w_n, \quad 1 \le n \le m$$

ove i $\{w_n\}_n$ sono i pesi di quadratura gaussiana, da cui, sommando su n,

$$\sum_{n=1}^{m}\left(\sum_{j=1}^{m}\hat{u}_j L\phi_j(x_n)\phi_i(x_n)w_n\right)=\sum_{j=1}^{m}\hat{u}_j\left(\sum_{n=1}^{m}L\phi_j(x_n)\phi_i(x_n)w_n\right)=$$
$$=\sum_{n=1}^{m}g(x_n)\phi_i(x_n)w_n,\quad 1\le i\le m$$

che è precisamente la formulazione di Galerkin pseudospettrale (14.6). □

14.3.1 Condizioni al bordo

Consideriamo il problema

$$\begin{cases}Lu(x)=g(x)\\ u(a)=u_a\\ u'(b)=u'_b\end{cases}$$

Con il metodo di collocazione, si ha

$$\begin{cases}\displaystyle\sum_{j=1}^{m}\hat{u}_j L\phi_j(x_n)=g(x_n),\quad 1\le n\le m-2\\ \displaystyle\sum_{j=1}^{m}\hat{u}_j\phi_j(a)=u_a\\ \displaystyle\sum_{j=1}^{m}\hat{u}_j\phi'_j(b)=u'_b\end{cases}$$

Anche in questo caso il metodo di collocazione può essere visto come un metodo di Galerkin pseudospettrale: basta prendere come nodi di collocazione gli $m-2$ nodi di quadratura gaussiana. Si ha poi

$$\begin{cases}\displaystyle\sum_{j=1}^{m}\hat{u}_j\left(\sum_{n=1}^{m-2}L\phi_j(x_n)\phi_i(x_n)w_n\right)=\sum_{n=1}^{m-2}g(x_n)\phi_i(x_n)w_n,\quad 1\le i\le m-2\\ \displaystyle\sum_{j=1}^{m}\hat{u}_j\phi_j(a)=u_a\\ \displaystyle\sum_{j=1}^{m}\hat{u}_j\phi'_j(b)=u'_b\end{cases}$$

Alternativamente, si possono usare, come nodi di collocazione, quelli delle formule di quadratura di Gauss–Lobatto (che contengono i nodi al bordo).

Collocazione Gauss–Lobatto–Chebyshev

I polinomi di Chebyshev sono definiti da

$$T_j(x) = \cos(j \arccos(x)), \quad -1 \le x \le 1$$

e soddisfano

$$\int_{-1}^{1} \frac{T_j(x) T_i(x)}{\sqrt{1-x^2}} \mathrm{d}x = \begin{cases} \pi & i = j = 0 \\ \frac{\pi}{2} & i = j \neq 0 \\ 0 & i \neq j \end{cases}$$

(lo si vede con il cambio di variabile $x = \cos\theta$ e applicando le formule di Werner, oppure integrando per parti due volte). I nodi di (Gauss–)Chebyshev–Lobatto sono $x_n = \cos((n-1)\pi/(m-1))$, $1 \le n \le m$. Possiamo allora definire la seguente famiglia di funzioni ortonormali

$$\phi_1(x) = \sqrt{\frac{1}{\pi}} T_0(x), \quad \phi_j(x) = \sqrt{\frac{2}{\pi}} T_{j-1}(x), \quad 2 \le j \le m$$

Ricordando la formula di ricorrenza tra polinomi di Chebyshev, possiamo scrivere

$$\begin{aligned} &\phi_1(x) = \sqrt{\frac{1}{\pi}}, \quad \phi_2(x) = \sqrt{\frac{2}{\pi}} x, \quad \phi_3(x) = 2x\phi_2(x) - \sqrt{2}\phi_1(x), \\ &\phi_{j+1}(x) = 2x\phi_j(x) - \phi_{j-1}(x), \quad 3 \le j \le m-1 \end{aligned}$$

Da qui, possiamo calcolare anche la derivata prima e seconda delle funzioni:

$$\begin{aligned} &\phi_1'(x) = 0, \quad \phi_2'(x) = \sqrt{\frac{2}{\pi}}, \quad \phi_3'(x) = 2\phi_2(x) + 2x\phi_2'(x), \\ &\phi_{j+1}'(x) = 2\phi_j(x) + 2x\phi_j'(x) - \phi_{j-1}'(x), \quad 3 \le j \le m-1 \end{aligned}$$

$$\begin{aligned} &\phi_1''(x) = 0, \quad \phi_2''(x) = 0, \quad \phi_3''(x) = 4\phi_2'(x), \\ &\phi_{j+1}''(x) = 4\phi_j'(x) + 2x\phi_j''(x) - \phi_{j-1}''(x), \quad 3 \le j \le m-1 \end{aligned}$$

Conviene calcolare le matrici

$$\mathbb{T}_0 = \begin{bmatrix} \phi_1(x_1) & \phi_1(x_2) & \dots & \phi_1(x_m) \\ \phi_2(x_1) & \phi_2(x_2) & \dots & \phi_2(x_m) \\ \vdots & \vdots & \dots & \vdots \\ \phi_m(x_1) & \phi_m(x_2) & \dots & \phi_m(x_m) \end{bmatrix}$$

$$\mathbb{T}_1 = \begin{bmatrix} \phi_1'(x_1) & \phi_1'(x_2) & \dots & \phi_1'(x_m) \\ \phi_2'(x_1) & \phi_2'(x_2) & \dots & \phi_2'(x_m) \\ \vdots & \vdots & \dots & \vdots \\ \phi_m'(x_1) & \phi_m'(x_2) & \dots & \phi_m'(x_m) \end{bmatrix}$$

$$\mathbb{T}_2 = \begin{bmatrix} \phi_1''(x_1) & \phi_1''(x_2) & \dots & \phi_1''(x_m) \\ \phi_2''(x_1) & \phi_2''(x_2) & \dots & \phi_2''(x_m) \\ \vdots & \vdots & \dots & \vdots \\ \phi_m''(x_1) & \phi_m''(x_2) & \dots & \phi_m''(x_m) \end{bmatrix}$$

Consideriamo, a titolo di esempio, il seguente problema modello

$$\begin{cases} -u''(x) + q(x)u(x) = g(x) \\ u(-1) = u_a \\ u'(1) = u_b' \end{cases}$$

Il sistema lineare risultante da risolvere per il metodo di collocazione Gauss–Chebyshev–Lobatto (per il momento *senza* tener conto delle condizioni al bordo) è

$$\left(-\mathbb{T}_2^{\mathrm{T}} + \begin{bmatrix} q(x_1) & & & \\ & q(x_2) & & \\ & & \ddots & \\ & & & q(x_m) \end{bmatrix} \mathbb{T}_0^{\mathrm{T}} \right) \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{bmatrix} = \begin{bmatrix} g(x_1) \\ g(x_2) \\ \vdots \\ g(x_m) \end{bmatrix}$$

Per imporre le condizioni al bordo, si sostituisce la prima riga della matrice con la prima riga di $\mathbb{T}_0^{\mathrm{T}}$ e il primo elemento del termine noto con u_a. Poi, l'ultima riga della matrice con l'ultima riga di $\mathbb{T}_1^{\mathrm{T}}$ e l'ultimo elemento del termine noto con u_b'. Una volta noti i coefficienti u_j, si ricostruisce la soluzione nello spazio fisico tramite

$$\begin{bmatrix} u(x_1) \\ u(x_2) \\ \vdots \\ u(x_m) \end{bmatrix} = \mathbb{T}_0^{\mathrm{T}} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{bmatrix}$$

Capitolo 15

Esercizi

1. Si risolva il problema ai limiti

$$\begin{cases} u''(x) = u(x) + x, & x \in (0,1) \\ u(0) = 0 \\ u(1) = 0 \end{cases} \tag{15.1}$$

usando il metodo delle differenze finite del secondo ordine. Sapendo che la soluzione esatta è $u(x) = (e^x - e^{-x})/(e - e^{-1}) - x$, si mostri inoltre l'ordine del metodo mediante un grafico logaritmico-logaritmico dell'errore in norma infinito.

2. Si risolva il problema ai limiti

$$\begin{cases} u''(x) + u'(x) + u(x) - \cos(x) = 0, & x \in (0, \pi/2) \\ u(0) = 0 \\ u(\pi/2) = 1 \end{cases}$$

usando il metodo delle differenze finite del secondo ordine. Si mostri inoltre l'ordine del metodo mediante un grafico logaritmico-logaritmico dell'errore in norma infinito rispetto ad una soluzione di riferimento.

3. Si risolva il problema ai limiti

$$\begin{cases} u''(x) + u'(x) + u(x) - \cos(x) = 0, & x \in (0, \pi/2) \\ u'(0) = 1 \\ u(\pi/2) = 1 \end{cases}$$

usando il metodo delle differenze finite del secondo ordine. Si mostri inoltre l'ordine del metodo mediante un grafico logaritmico-logaritmico dell'errore in norma infinito rispetto ad una soluzione di riferimento.

4. Si risolva il problema ai limiti

$$\begin{cases} u''(x) = \cos(u(x)), & x \in (0,1) \\ u(0) = 0 \\ u(1) = 1 \end{cases}$$

usando il metodo delle differenze finite del secondo ordine.

5. Si risolva il problema ai limiti

$$\begin{cases} -\dfrac{\mathrm{d}}{\mathrm{d}x}\left((1+x)\dfrac{\mathrm{d}}{\mathrm{d}x}u(x)\right) = 1, & x \in (0,1) \\ u(0) = 0 \\ u(1) = 0 \end{cases}$$

Si mostri inoltre l'ordine del metodo mediante un grafico logaritmico-logaritmico dell'errore in norma infinito rispetto alla soluzione esatta.

6. Si risolva il problema ai limiti

$$\begin{cases} u''(x) = 20u'(x) + u(x), & x \in (0,1) \\ u(0) = 0 \\ u(1) = 1 \end{cases}$$

Visto l'andamento della soluzione, si implementi uno schema di differenze finite su nodi non equispaziati secondo una distribuzione di tipo coseno. Si confrontino gli errori rispetto alla soluzione analitica.

7. Si ricavi la relazione di ricorrenza dei polinomi ortonormali nell'intervallo $[-\infty, \infty]$ rispetto alla funzione peso $w(x) = e^{-\alpha^2 x^2}$

8. Noti gli zeri dei polinomi di Legendre e i pesi di quadratura della rispettiva formula gaussiana, si ricavino i nodi e i pesi di una formula gaussiana nell'intervallo $[a, b]$ rispetto al peso $w(x) = 1$.

9. Si risolva il problema ai limiti (15.1) usando il metodo di collocazione con polinomi di Legendre. Gli N nodi di collocazione in $[a, b]$ e la valutazione dei polinomi di Legendre e delle loro derivate sono dati dalla function `[L,x,L1,L2] = legendrepolynomials(N,a,b)`. Si mostri inoltre l'ordine del metodo mediante un grafico logaritmico-logaritmico dell'errore in norma infinito.

Parte 2

ODEs (Equazioni differenziali ordinarie)

Capitolo 16

Introduzione

Consideriamo il sistema di equazioni differenziali ordinarie (ODEs) *ai valori iniziali (initial value problem)*

$$\begin{cases} y_1'(t) = f_1(t, y_1(t), y_2(t), \dots, y_d(t)) \\ y_2'(t) = f_2(t, y_1(t), y_2(t), \dots, y_d(t)) \\ \quad \vdots \\ y_d'(t) = f_d(t, y_1(t), y_2(t), \dots, y_d(t)) \end{cases}$$

con dato iniziale

$$\begin{cases} y_1(t_0) = y_{1_0} \\ y_2(t_0) = y_{2_0} \\ \quad \vdots \\ y_d(t_0) = y_{d_0} \end{cases}$$

che può essere riscritto, in forma compatta,

$$\begin{cases} \boldsymbol{y}'(t) = \boldsymbol{f}(t, \boldsymbol{y}(t)), \quad t > t_0 \\ \boldsymbol{y}(t_0) = \boldsymbol{y}_0 \end{cases} \tag{16.1}$$

Assumiamo $\boldsymbol{y}_0 \in \mathbb{R}^d$ e $\boldsymbol{f} \colon [t_0, +\infty) \times \mathbb{R}^d \to \mathbb{R}^d$ globalmente lipschitziana nel secondo argomento

$$\|\boldsymbol{f}(t, \boldsymbol{x}) - \boldsymbol{f}(t, \boldsymbol{y})\| \leq \lambda \|\boldsymbol{x} - \boldsymbol{y}\| \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$$

Allora il sistema (16.1) ha un'unica soluzione.

16.1 Riduzione in forma autonoma

Un sistema in forma *non autonoma*

$$\begin{cases} \boldsymbol{y}'(t) = \boldsymbol{f}(t, \boldsymbol{y}(t)), & t > t_0 \\ \boldsymbol{y}(t_0) = \boldsymbol{y}_0 \end{cases}$$

può essere ricondotto in forma autonoma mediante l'introduzione della variabile

$$y_{d+1}(t) = t$$

Si giunge a

$$\begin{cases} \boldsymbol{y}'(t) = \boldsymbol{f}(y_{d+1}(t), \boldsymbol{y}(t)), & t > t_0 \\ y'_{d+1}(t) = 1, & t > t_0 \\ \boldsymbol{y}(t_0) = \boldsymbol{y}_0 \\ y_{d+1}(t_0) = t_0 \end{cases}$$

16.2 Equazioni di ordine superiore al primo

Le equazioni differenziali di ordine d del tipo

$$\begin{cases} y^{(d)}(t) = f(t, y(t), y'(t), \ldots, y^{(d-1)}(t)), & t > t_0 \\ y(t_0) = y_{0,0} \\ y'(t_0) = y_{0,1} \\ \vdots \\ y^{(d-1)}(t_0) = y_{0,d-1} \end{cases}$$

(cioè in cui vengono prescritti i valori iniziali della funzione e delle derivate) si possono ricondurre ad un sistema di ODEs di ordine uno, mediante la sostituzione

$$\begin{cases} y_1(t) = y(t) \\ y_2(t) = y'(t) \\ \vdots \\ y_d(t) = y^{(d-1)}(t) \end{cases}$$

dando così luogo a

$$\begin{cases} \boldsymbol{y}'(t) = \boldsymbol{f}(t, \boldsymbol{y}(t)) \\ \boldsymbol{y}(t_0) = [y_{0,0}, y_{0,1}, \ldots, y_{0,d-1}]^{\mathrm{T}} \end{cases}$$

ove

$$\boldsymbol{f}(t, \boldsymbol{y}(t)) = [y_2(t), y_3(t), \ldots, y_d(t), f(t, y_1(t), y_2(t), \ldots, y_{d-1}(t))]^{\mathrm{T}}$$

Capitolo 17

Metodi ad un passo

17.1 Metodo di Eulero

Il *metodo di Eulero* (o *Eulero esplicito*, o *forward Euler*) si basa sull'approssimazione

$$\boldsymbol{y}'(t) \approx \frac{\boldsymbol{y}(t) - \boldsymbol{y}(t_0)}{t - t_0}$$
$$\boldsymbol{f}(t, \boldsymbol{y}(t)) \approx \boldsymbol{f}(t_0, \boldsymbol{y}(t_0))$$

per cui $\boldsymbol{y}(t) \approx \boldsymbol{y}(t_0) + (t - t_0)\boldsymbol{f}(t_0, \boldsymbol{y}(t_0))$. Pertanto l'approssimazione di $\boldsymbol{y}(t)$ è ottenuta per interpolazione lineare a partire da $(t_0, \boldsymbol{f}(t_0, \boldsymbol{y}(t_0)))$, con pendenza $\boldsymbol{f}(t_0, \boldsymbol{y}(t_0))$. Può essere visto anche come applicazione della formula di quadratura del rettangolo (estremo sinistro) alla soluzione analitica

$$\boldsymbol{y}(t) = \boldsymbol{y}(t_0) + \int_{t_0}^{t} \boldsymbol{f}(\tau, \boldsymbol{y}(\tau))\mathrm{d}\tau$$

Data la sequenza $t_0, t_1 = t_0 + k, t_2 = t_0 + 2k, \ldots, t_n = t_0 + nk, \ldots$, ove k è il *passo temporale* (o *time step*), lo schema numerico che ne risulta è

$$\begin{aligned} \boldsymbol{y}_{n+1} &= \boldsymbol{y}_n + k\boldsymbol{f}(t_n, \boldsymbol{y}_n), \quad n \geq 0, \\ \boldsymbol{y}_0 &= \boldsymbol{y}(t_0) \end{aligned} \tag{17.1}$$

ove $\boldsymbol{y}_n \approx \boldsymbol{y}(t_n)$. In pratica, $\boldsymbol{y}_{n+1}$ è la soluzione approssimata al tempo $t_n + k$, mediante un passo del metodo di Eulero, del sistema

$$\begin{cases} \tilde{\boldsymbol{y}}'(t) = \boldsymbol{f}(t, \tilde{\boldsymbol{y}}(t)), \quad t > t_n \\ \tilde{\boldsymbol{y}}(t_n) = \boldsymbol{y}_n \end{cases}$$

Se consideriamo l'intervallo temporale $[t_0, t_0 + t^*]$ e abbiamo fissato m, indichiamo con $\boldsymbol{y}_n$, $n \leq m$, la soluzione approssimata al tempo t_n mediante un generico metodo $\boldsymbol{y}_{n+1} = \boldsymbol{\mathcal{Y}}_n(\boldsymbol{f}, k, \boldsymbol{y}_n)$ per la soluzione del sistema differenziale (16.1), ove il passo temporale è $k = t^*/m$.

Definizione 1. *La quantità* $\boldsymbol{y}(t_{n+1}) - \boldsymbol{\mathcal{Y}}_n(\boldsymbol{f}, k, \boldsymbol{y}(t_n))$ *si chiama* errore locale *del metodo.*

Detto in altre parole, esso è la differenza tra la soluzione esatta al tempo t_{n+1} e la soluzione approssimata al tempo t_{n+1} che si *otterrebbe* applicando il metodo numerico al problema differenziale e supponendo esatta la soluzione al tempo t_n. Per il metodo di Eulero, si ha

$$\begin{aligned}\boldsymbol{y}(t_{n+1}) - \boldsymbol{\mathcal{Y}}_n(\boldsymbol{f}, k, \boldsymbol{y}(t_n)) &= \boldsymbol{y}(t_{n+1}) - \boldsymbol{y}(t_n) - k\boldsymbol{f}(t_n, \boldsymbol{y}(t_n)) = \\ &= \boldsymbol{y}(t_n) + k\boldsymbol{y}'(t_n) + \mathcal{O}(k^2) - \boldsymbol{y}(t_n) - k\boldsymbol{y}'(t_n) = \mathcal{O}(k^2) \quad (17.2)\end{aligned}$$

(supponendo $\boldsymbol{y} \in \mathcal{C}^2$). Poiché ad ogni passo si commette un errore di ordine $\mathcal{O}(k^2)$ (che si accumula agli errori prodotti ai passi precendeti) e i passi sono $m = t^*/k$, *se tutto va bene, anche quando* $m \to \infty$ *(cioè* $k \to 0^+$*)* si commette un errore massimo di ordine $t^*/k \cdot \mathcal{O}(k^2) = \mathcal{O}(k)$. È giustificata allora la seguente

Definizione 2. *Un metodo* $\boldsymbol{y}_{n+1} = \boldsymbol{\mathcal{Y}}_n(\boldsymbol{f}, k, \boldsymbol{y}_n)$ *è di ordine* p *se* $\boldsymbol{y}(t_{n+1}) - \boldsymbol{\mathcal{Y}}_n(\boldsymbol{f}, k, \boldsymbol{y}(t_n)) = \mathcal{O}(k^{p+1})$, $k = t^*/m$, *per* $m \to \infty$, *per qualunque* $\boldsymbol{f}$ *analitica e* $0 \le n \le m-1$.

Un metodo di ordine $p \ge 1$ si dice *consistente di ordine* p, o semplicemente *consistente*. Se $\boldsymbol{y}(t_{n+1}) - \boldsymbol{\mathcal{Y}}_n(\boldsymbol{f}, k, \boldsymbol{y}(t_n)) = 0$ il metodo si dice *fortemente consistente*. Dunque il metodo di Eulero è consistente di ordine 1. Si tratta ora di dimostrare che *tutto va bene*.

Nella definizione sopra, al variare di m il significato di t_n (e di $\boldsymbol{y}_n$) cambia, nel senso che $\boldsymbol{y}_n \approx \boldsymbol{y}(t_n) = \boldsymbol{y}(t_0 + nt^*/m)$. Per essere più chiari (ma anche più pesanti) dobbiamo introdurre la notazione $\boldsymbol{y}_{n,m}$ per indicare l'approssimazione di $\boldsymbol{y}(t_{n,m}) = \boldsymbol{y}(t_0 + nt^*/m)$, osservando che $\boldsymbol{y}(t_{n,m_1}) \neq \boldsymbol{y}(t_{n,m_2})$ se $m_1 \neq m_2$.

Definizione 3. *Il metodo* $\boldsymbol{y}_{n+1,m} = \boldsymbol{\mathcal{Y}}_n(\boldsymbol{f}, k, \boldsymbol{y}_{n,m})$ *è* convergente *se*

$$\lim_{m\to\infty} \max_{0\le n\le m} \|\boldsymbol{e}_{n,m}\| = 0$$

ove $\boldsymbol{e}_{n,m} = \boldsymbol{y}_{n,m} - \boldsymbol{y}(t_n)$. *La quantità* $\max_n \|\boldsymbol{e}_{n,m}\|$ *si chiama* errore globale.

Di solito si parla di convergenza, e anche di ordine, per $k \to 0^+$. È la stessa cosa, ma se si fissa k, non è detto che t^*/k sia un numero intero. Si può allora considerare il limite

$$\lim_{k_n\to 0} \max_{0\le n\le m} \|\boldsymbol{e}_{n,m}\|$$

ove $k_m = t^*/m$. Nel seguito assumeremo comunque sempre che valga la relazione $k = t^*/m$. Con questa assunzione, useremo anche le notazioni $t_{n,k} = t_{n,m}$, $\boldsymbol{y}_{n,k} = \boldsymbol{y}_{n,m}$ e $\boldsymbol{e}_{n,k} = \boldsymbol{e}_{n,m}$.

Teorema 8. *Il metodo di Eulero è convergente.*

Dimostrazione. Assumiamo $\boldsymbol{f}$ (e dunque $\boldsymbol{y}$) analitica. Dalle uguaglianze

$$\begin{aligned} \boldsymbol{y}_{n+1,k} &= \boldsymbol{y}_{n,k} + k\boldsymbol{f}(t_{n,k}, \boldsymbol{y}_{n,k}) && \text{(definizione del metodo)} \\ \boldsymbol{y}(t_{n+1,k}) &= \boldsymbol{y}(t_n) + k\boldsymbol{f}(t_{n,k}, \boldsymbol{y}(t_n)) + \mathcal{O}(k^2) && \text{(errore locale (17.2))} \end{aligned}$$

si ricava

$$\boldsymbol{e}_{n+1,k} = \underbrace{\boldsymbol{e}_{n,k} + k[\boldsymbol{f}(t_{n,k}, \boldsymbol{y}_n) - \boldsymbol{f}(t_{n,k}, \boldsymbol{y}(t_{n,k})]}_{\boldsymbol{y}_{n+1,k} - \boldsymbol{\mathcal{Y}}_n(\boldsymbol{f},k,\boldsymbol{y}(t_n))} + \underbrace{\mathcal{O}(k^2)}_{\boldsymbol{\mathcal{Y}}_n(\boldsymbol{f},k,\boldsymbol{y}(t_{n,k})) - \boldsymbol{y}(t_{n+1,k})}$$

da cui, siccome $\boldsymbol{f}$ è lipschitziana,

$$\begin{aligned} \|\boldsymbol{e}_{n+1,k}\| &\le \|\boldsymbol{e}_{n,k}\| + k\|\boldsymbol{f}(t_{n,k}, \boldsymbol{y}_{n,k}) - \boldsymbol{f}(t_{n,k}, \boldsymbol{y}(t_{n,k})\| + ck^2 \le \\ &\le (1 + k\lambda)\|\boldsymbol{e}_{n,k}\| + ck^2, \quad c > 0 \end{aligned}$$

Applicando l'analoga disuguaglianza a $\boldsymbol{e}_{n,k}$ si trova

$$\|\boldsymbol{e}_{n+1,k}\| \le (1 + k\lambda)^2\|\boldsymbol{e}_{n-1,k}\| + ((1 + k\lambda) + 1)ck^2$$

e dunque

$$\begin{aligned} \|\boldsymbol{e}_{n,k}\| &\le (1 + k\lambda)^n\|\boldsymbol{e}_{0,k}\| + ((1 + k\lambda)^{n-1} + \ldots + (1 + k\lambda) + 1)ck^2 = \\ &= (1 + k\lambda)^n\|\boldsymbol{e}_{0,k}\| + \frac{(1 + k\lambda)^n - 1}{(1 + k\lambda) - 1}ck^2 \end{aligned}$$

Poiché $1 + k\lambda < \mathrm{e}^{k\lambda}$, $(1 + k\lambda)^n < \mathrm{e}^{nk\lambda} \le \mathrm{e}^{mk\lambda} = \mathrm{e}^{t^*\lambda}$. Dunque

$$\|\boldsymbol{e}_{n,k}\| \le \mathrm{e}^{t^*\lambda}\|\boldsymbol{e}_{0,k}\| + \frac{\mathrm{e}^{t^*\lambda} - 1}{\lambda}ck, \quad 1 \le n \le m$$

Si assume generalmente che $\boldsymbol{y}_{0,k} = \boldsymbol{y}(t_0)$ e dunque

$$\|\boldsymbol{e}_{n,k}\| \le \frac{\mathrm{e}^{t^*\lambda} - 1}{\lambda}ck, \quad 0 \le n \le m$$

da cui

$$\lim_{k \to 0^+} \max_{0 \le n \le m} \|\boldsymbol{e}_{n,k}\| = 0$$

□

In particolare, l'errore globale tende a 0 come k, come ci si aspettava. Da notare che l'errore globale dipende anche dall'intervallo di tempo t^* (anche se la stima ottenuta è molto pessimistica, generalmente l'errore globale cresce linearmente con l'ampiezza dell'intervallo di tempo).

17.2 Metodo dei trapezi

Il *metodo dei trapezi* (o metodo di *Crank–Nicolson*) si basa sull'approssimazione

$$\boldsymbol{y}'(t) \approx \frac{\boldsymbol{y}(t) - \boldsymbol{y}(t_0)}{t - t_0}$$
$$\boldsymbol{f}(t, \boldsymbol{y}(t)) \approx \frac{1}{2}(\boldsymbol{f}(t_0, \boldsymbol{y}(t_0)) + \boldsymbol{f}(t, \boldsymbol{y}(t)))$$

Può essere visto anche come applicazione della formula di quadratura del trapezio alla soluzione analitica

$$\boldsymbol{y}(t) = \boldsymbol{y}(t_0) + \int_{t_0}^{t} \boldsymbol{f}(\tau, \boldsymbol{y}(\tau))\mathrm{d}\tau$$

Data la sequenza $t_0, t_1 = t_0 + k, t_2 = t_0 + 2k, \ldots, t_n = t_0 + nk, \ldots$, lo schema numerico che ne risulta è

$$\begin{aligned} \boldsymbol{y}_{n+1} &= \boldsymbol{y}_n + \frac{k}{2}(\boldsymbol{f}(t_n, \boldsymbol{y}_n) + \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1})), \quad n \geq 0, \\ \boldsymbol{y}_0 &= \boldsymbol{y}(t_0) \end{aligned} \tag{17.3}$$

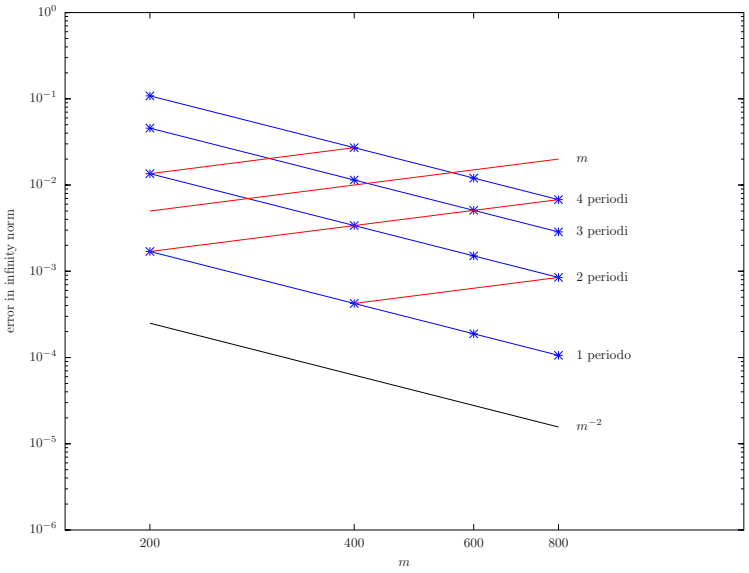
ove $\boldsymbol{y}_n \approx \boldsymbol{y}(t_n)$. Indicato con $\boldsymbol{y}_{n+1} = \boldsymbol{\mathcal{Y}}_n(\boldsymbol{f}, k, \boldsymbol{y}_n, \boldsymbol{y}_{n+1})$ un generico metodo di questo tipo, l'errore locale si ottiene calcolando

$$\begin{aligned} &\boldsymbol{y}(t_{n+1}) - \boldsymbol{\mathcal{Y}}_n(\boldsymbol{f}, k, \boldsymbol{y}(t_n), \boldsymbol{y}(t_{n+1})) = \\ &\quad = \boldsymbol{y}(t_{n+1}) - \boldsymbol{y}(t_n) - \frac{k}{2}(\boldsymbol{f}(t_n, \boldsymbol{y}(t_n)) + \boldsymbol{f}(t_{n+1}, \boldsymbol{y}(t_{n+1}))) = \\ &= \boldsymbol{y}(t_n) + k\boldsymbol{y}'(t_n) + \frac{k^2}{2}\boldsymbol{y}''(t_n) + \frac{k^3}{6}\boldsymbol{y}'''(t_n) + \mathcal{O}(k^4) - \boldsymbol{y}(t_n) - \frac{k}{2}(\boldsymbol{y}'(t_n) + \boldsymbol{y}'(t_{n+1})) = \\ &\quad = k\boldsymbol{y}'(t_n) + \frac{k^2}{2}\boldsymbol{y}''(t_n) + \frac{k^3}{6}\boldsymbol{y}'''(t_n) + \mathcal{O}(k^4) + \\ &\quad - \frac{k}{2}\left(\boldsymbol{y}'(t_n) + \boldsymbol{y}'(t_n) + k\boldsymbol{y}''(t_n) + \frac{k^2}{2}\boldsymbol{y}'''(t_n) + \mathcal{O}(k^3)\right) = \mathcal{O}(k^3) \end{aligned}$$

(supponendo $\boldsymbol{y} \in \mathcal{C}^3$). Dunque il metodo dei trapezi è di ordine 2. Rimandiamo la dimostrazione della convergenza del metodo al paragrafo successivo.

In Figura 17.1 si vede chiaramente che l'errore (per l'equazione del pendolo linearizzata) decade come $k^2 \propto m^{-2}$ e che lo stesso cresce linearmente con l'intervallo di tempo. Entrambi i metodi descritti sono *ad un passo* (cioè la soluzione $\boldsymbol{y}_{n+1}$ dipende esplicitamente solo da $\boldsymbol{y}_n$). Il metodo dei trapezi è però *implicito,* cioè la soluzione $\boldsymbol{y}_{n+1}$ è implicitamente definita dall'equazione (in generale non lineare)

$$F_n(\boldsymbol{y}_{n+1}) = \boldsymbol{y}_{n+1} - \frac{k}{2}\boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}) - \boldsymbol{y}_n - \frac{k}{2}\boldsymbol{f}(t_n, \boldsymbol{y}_n) = 0$$



pendoloperiod.m

Figura 17.1: Errori del metodo dei trapezi per l'equazione del pendolo linearizzata.

17.3 θ-metodo

Il θ-metodo è una generalizzazione dei metodi precedenti e si scrive

$$\begin{aligned} \boldsymbol{y}_{n+1} &= \boldsymbol{y}_n + k[(1-\theta)\boldsymbol{f}(t_n, \boldsymbol{y}_n) + \theta \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1})], \quad n \geq 0 \\ \boldsymbol{y}_0 &= \boldsymbol{y}(t_0) \end{aligned} \tag{17.4}$$

È facile verificare che

$$\begin{aligned} &\boldsymbol{y}(t_{n+1}) - \boldsymbol{\mathcal{Y}}_n(\boldsymbol{f}, k, \boldsymbol{y}(t_n), \boldsymbol{y}(t_{n+1})) = \\ &\quad = \boldsymbol{y}(t_{n+1}) - \boldsymbol{y}(t_n) - k[(1-\theta)\boldsymbol{f}(t_n, \boldsymbol{y}(t_n)) + \theta \boldsymbol{f}(t_{n+1}, \boldsymbol{y}(t_{n+1}))] = \\ &\quad\quad = \left(\frac{1}{2} - \theta\right) k^2 \boldsymbol{y}''(t_n) + \left(\frac{1}{6} - \frac{\theta}{2}\right) k^3 \boldsymbol{y}'''(t_n) + \mathcal{O}(k^4) \end{aligned} \tag{17.5}$$

e dunque il metodo ha ordine due se $\theta = \frac{1}{2}$, e ordine uno altrimenti. In particolare, se $\boldsymbol{y}''(t)$ è nulla, tale è l'errore locale per il θ-metodo. E se $\theta = \frac{1}{2}$ e $\boldsymbol{y}'''(t)$ è nulla, tale è l'errore locale. Si assume $\theta \in [0, 1]$.

Qual è il significato dell'errore locale per un metodo implicito? Possiamo considerare il problema differenziale

$$\begin{cases} \boldsymbol{y}^{*\prime}(t) = \boldsymbol{f}(t, \boldsymbol{y}^*(t)), \quad t > t_n \\ \boldsymbol{y}^*(t_n) = \boldsymbol{y}(t_n) \end{cases}$$

ed approssimare $\boldsymbol{y}^*(t_{n+1})$ mediante un passo del θ-metodo. Si ha

$$\boldsymbol{y}^*_{n+1} = \boldsymbol{\mathcal{Y}}_n(\boldsymbol{f}, k, \boldsymbol{y}(t_n), \boldsymbol{y}^*_{n+1}) = \boldsymbol{y}(t_n) + k[(1-\theta)\boldsymbol{f}(t_n, \boldsymbol{y}(t_n)) + \theta \boldsymbol{f}(t_{n+1}, \boldsymbol{y}^*_{n+1})]$$

da cui (usando (17.5))

$$\begin{aligned}\boldsymbol{y}(t_{n+1}) - \boldsymbol{y}^*_{n+1} = k\theta[\boldsymbol{f}(t_{n+1}, \boldsymbol{y}(t_{n+1})) - \boldsymbol{f}(t_{n+1}, \boldsymbol{y}^*_{n+1})] + \\ + \left(\frac{1}{2} - \theta\right) k^2 \boldsymbol{y}''(t_n) + \left(\frac{1}{6} - \frac{\theta}{2}\right) k^3 \boldsymbol{y}'''(t_n) + \mathcal{O}(k^4)\end{aligned}$$

Per il metodo di Eulero ($\theta = 0$), $\boldsymbol{y}^*_{n+1} = \boldsymbol{\mathcal{Y}}_n(\boldsymbol{f}, k, \boldsymbol{y}(t_n))$ e dunque $\boldsymbol{y}(t_{n+1}) - \boldsymbol{y}^*_{n+1}$ è l'errore locale per definizione. Altrimenti, posto $\boldsymbol{e}^*_{n+1} = \boldsymbol{y}(t_{n+1}) - \boldsymbol{y}^*_{n+1}$, si ha

$$\|\boldsymbol{e}^*_{n+1}\| \leq k\theta\lambda\|\boldsymbol{e}^*_{n+1}\| + \left(\frac{1}{2} - \theta\right) ck^2 + \left(\frac{1}{6} - \frac{\theta}{2}\right) dk^3$$

Dunque, se $0 < \varepsilon < 1 - k\theta\lambda$ (cioè k è sufficientemente piccolo), a meno di rinominare le costanti c e d

$$\begin{aligned}\|\boldsymbol{e}^*_{n+1}\| \leq \left(\frac{1}{2} - \theta\right) ck^2 + \left(\frac{1}{6} - \frac{\theta}{2}\right) dk^3 = \\ = \boldsymbol{y}(t_{n+1}) - \boldsymbol{\mathcal{Y}}_n(\boldsymbol{f}, k, \boldsymbol{y}(t_n), \boldsymbol{y}(t_{n+1}))\end{aligned}$$

Quindi, anche per uno schema implicito, l'errore locale è *dello stesso ordine* della differenza tra la soluzione esatta al tempo t_{n+1} e la soluzione che si otterrebbe applicando il metodo numerico e supponendo esatta la soluzione al tempo t_n. Per questo motivo, con un abuso di linguaggio, spesso si identificano le due cose.

Per dimostrare che il θ-metodo converge, introduciamo la seguente funzione (definita implicitamente per $\theta \neq 0$)

$$\psi(t, \boldsymbol{y}) = (1-\theta)\boldsymbol{f}(t, \boldsymbol{y}) + \theta \boldsymbol{f}(t + k, \boldsymbol{y} + k\psi(t, \boldsymbol{y}))$$

Osserviamo che

$$\underbrace{\boldsymbol{y}_n + k\psi(t_n, \boldsymbol{y}_n)}_{} = \boldsymbol{y}_n + k[(1-\theta)\boldsymbol{f}(t_n, \boldsymbol{y}_n) + \theta \boldsymbol{f}(t_n + k, \underbrace{\boldsymbol{y}_n + k\psi(t_n, \boldsymbol{y}_n)}_{})]$$

da cui

$$\boldsymbol{y}_n + k\psi(t_n, \boldsymbol{y}_n) = \boldsymbol{y}_{n+1}$$

Analogamente,

$$\boldsymbol{y}(t_n) + k\psi(t_n, \boldsymbol{y}(t_n)) = \boldsymbol{y}^*_{n+1}$$

Osserviamo infine che

$$\begin{aligned}\|\psi(t,\boldsymbol{x})-\psi(t,\boldsymbol{y})\| &= \|(1-\theta)\boldsymbol{f}(t,\boldsymbol{x})+\theta\boldsymbol{f}(t+k,\boldsymbol{x}+k\psi(t,\boldsymbol{x}))+\\ &\quad -(1-\theta)\boldsymbol{f}(t,\boldsymbol{y})-\theta\boldsymbol{f}(t+k,\boldsymbol{y}+k\psi(t,\boldsymbol{y}))\| \leq\\ &\quad \leq (1-\theta)\lambda\|\boldsymbol{x}-\boldsymbol{y}\|+\theta\lambda\|\boldsymbol{x}+k\psi(t,\boldsymbol{x})-\boldsymbol{y}-k\psi(t,\boldsymbol{y})\| \leq\\ &\quad \leq (1-\theta)\lambda\|\boldsymbol{x}-\boldsymbol{y}\|+\theta\lambda\|\boldsymbol{x}-\boldsymbol{y}\|+k\theta\lambda\|\psi(t,\boldsymbol{x})-\psi(t,\boldsymbol{y})\|\end{aligned}$$

da cui, se $0<1-k\theta\lambda$ (cioè k sufficientemente piccolo),

$$\|\psi(t,\boldsymbol{x})-\psi(t,\boldsymbol{y})\| \leq \frac{\lambda}{1-k\theta\lambda}\|\boldsymbol{x}-\boldsymbol{y}\| = \Lambda\|\boldsymbol{x}-\boldsymbol{y}\|$$

Quindi $\psi(t,\boldsymbol{y})$ è uniformemente lipshitziana nel secondo argomento.

Teorema 9. *Il θ-metodo è convergente.*

Dimostrazione. Si ha

$$\begin{aligned}\|\boldsymbol{e}_{n+1}\| &= \|\boldsymbol{y}_{n+1}-\boldsymbol{y}(t_{n+1})\| = \|(\boldsymbol{y}_{n+1}-\boldsymbol{y}^*_{n+1})+(\boldsymbol{y}^*_{n+1}-\boldsymbol{y}(t_{n+1}))\| \leq\\ &\leq \|\boldsymbol{y}_{n+1}-\boldsymbol{y}^*_{n+1}\|+\|\boldsymbol{y}^*_{n+1}-\boldsymbol{y}(t_{n+1})\| \leq\\ &\leq \|\boldsymbol{e}_n\|+k\|\psi(t_n,\boldsymbol{y}_n)-\psi(t_n,\boldsymbol{y}(t_n))\|+\left(\frac{1}{2}-\theta\right)ck^2+\left(\frac{1}{6}-\frac{\theta}{2}\right)dk^3 \leq\\ &\leq (1+k\Lambda)\|\boldsymbol{e}_n\|+\left(\frac{1}{2}-\theta\right)ck^2+\left(\frac{1}{6}-\frac{\theta}{2}\right)dk^3\end{aligned}$$

Si conclude adesso esattamente come nel caso del metodo di Eulero. Quindi il θ-metodo è convergente con ordine uno se $\theta \neq \frac{1}{2}$ e ordine due altrimenti. □

Osserviamo che:

- il metodo per $\theta = 1$ si chiama *Eulero implicito (backward Euler)*;
- per $\theta = \frac{1}{3}$ il metodo è di ordine uno, ma il termine contenente la derivata terza della soluzione è annullato.

17.3.1 Risoluzione di un metodo implicito

Nel caso implicito ($\theta \neq 0$), ad ogni passo n si deve risolvere un sistema di equazioni in generale non lineari $F_n(\boldsymbol{x}) = 0$, $\boldsymbol{x} = \boldsymbol{y}_{n+1}$, ove

$$F_n(\boldsymbol{x}) = \boldsymbol{x} - k\theta\boldsymbol{f}(t_{n+1},\boldsymbol{x}) - \boldsymbol{y}_n - k(1-\theta)\boldsymbol{f}(t_n,\boldsymbol{y}_n)$$

La prima idea potrebbe essere quella di applicare il metodo di punto fisso:

$$\boldsymbol{x}^{(r+1)} = \boldsymbol{g}(\boldsymbol{x}^{(r)}) = k\theta\boldsymbol{f}(t_{n+1},\boldsymbol{x}^{(r)}) + \boldsymbol{y}_n + k(1-\theta)\boldsymbol{f}(t_n,\boldsymbol{y}_n)$$

La funzione $\boldsymbol{g}$ soddisfa

$$\|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{y})\| = \|k\theta\boldsymbol{f}(t_{n+1}, \boldsymbol{x}) - k\theta\boldsymbol{f}(t_{n+1}, \boldsymbol{y})\| \leq k\theta\lambda\|\boldsymbol{x} - \boldsymbol{y}\|$$

ed è una contrazione se

$$k\theta\lambda < 1 \Rightarrow k < \frac{1}{\theta\lambda} \tag{17.6}$$

Dunque, a patto di prendere k sufficientemente piccolo, si può ricavare la soluzione $\boldsymbol{x} = \boldsymbol{y}_{n+1}$ con il metodo del punto fisso, partendo, per esempio, da $\boldsymbol{x}^{(0)} = \boldsymbol{y}_n$. Il metodo del punto fisso, facile da applicare, ha convergenza lineare ed è inoltre soggetto ad una restrizione sul passo k che potrebbe essere eccessiva. Per questi motivi è utile considerare anche il metodo di Newton per la risoluzione del sistema non lineare. La matrice jacobiana associata è

$$J_n(\boldsymbol{x}) = I - k\theta \left(\frac{\partial f_i(t_{n+1}, \boldsymbol{x})}{\partial x_j} \right)_{ij}$$

Il calcolo di $\boldsymbol{y}_{n+1}$, con $\boldsymbol{y}_n$ come valore iniziale, con il metodo di Newton avviene dunque secondo il seguente algoritmo:

- $r = 0$
- $\boldsymbol{x}^{(r)} = \boldsymbol{y}_n$
- $J_n(\boldsymbol{x}^{(r)})\boldsymbol{\delta}^{(r)} = -F_n(\boldsymbol{x}^{(r)})$
- WHILE $\|\boldsymbol{\delta}^{(r)}\| >$ Newt_tol

  $\boldsymbol{x}^{(r+1)} = \boldsymbol{x}^{(r)} + \boldsymbol{\delta}^{(r)}$

  $r = r + 1$

  $J_n(\boldsymbol{x}^{(r)})\boldsymbol{\delta}^{(r)} = -F_n(\boldsymbol{x}^{(r)})$

  END
- $\boldsymbol{y}_{n+1} = \boldsymbol{x}^{(r)}$

In Matlab/Octave l'implementazione potrebbe essere:

```
f = @(t,y) ... % rhs dell'equazione
j = @(t,y) ... % jacobiano di f
F = @(tn,yn,t,y) y - k * theta * f(t,y) - yn - k * (1 - theta) * f(tn,yn)
J = @(t,y) eye(length(y)) - k * theta * j(t,y)
y(:,1) = y0;
t = 0;
```

```
for n = 1:ts
  y(:,n+1) = y(:,n);
  errest = -J(t + k,y(:,n + 1)) \ F(t,y(:,n),t + k,y(:,n + 1));
  while (norm(errest,inf) > Newt_tol)
    y(:,n + 1) = y(:,n + 1) + errest;
    errest = -J(t + k,y(:,n + 1)) \ F(t,y(:,n),t + k,y(:,n + 1));
  end
  t = t + k;
end
```

La tolleranza Newt_tol va presa tenendo conto che si sta comunque commettendo un errore proporzionale a k^2 (trapezi) o addirittura k. Come valore iniziale $\boldsymbol{x}^{(0)}$, invece di $\boldsymbol{y}_n$ si può prendere $\boldsymbol{y}_n + k\boldsymbol{f}(t_n, \boldsymbol{y}_n)$, che corrisponde a $\boldsymbol{y}(t_{n+1})$ approssimato con il metodo di Eulero e dunque dovrebbe essere un'approssimazione migliore di $\boldsymbol{y}_{n+1}$ (ma si veda il capitolo 20). Anche altre scelte possono andare bene, a patto di avere convergenza del metodo.

17.3.2 Newton inesatto e passo variabile

Nel caso in cui il calcolo e/o la risoluzione dei sistemi lineari con $J_n(\boldsymbol{x}^{(r)})$ risulti particolarmente oneroso, si può ricorrere al metodo di Newton inesatto, considerando ad ogni passo la matrice jacobiana costante $J_n(\boldsymbol{x}^{(r)}) \equiv J_n(\boldsymbol{x}^{(0)})$, o, in generale, qualunque modifica della matrice jacobiana. A questo punto, la si può fattorizzare in LU una sola volta per passo temporale e poi risolvere i sistemi con le corrispondenti matrici triangolari. Il metodo di Newton inesatto converge generalmente in maniera lineare e dunque serviranno più iterazioni (ma meno costose) rispetto al metodo di Newton esatto. Per inciso, questo è il sintomo principale di una matrice jacobiana non corretta. Il numero di iterazioni necessarie alla convergenza dipende (anche) dalla vicinanza della soluzione iniziale $\boldsymbol{x}^{(0)} = \boldsymbol{y}_n$ a quella finale $\boldsymbol{y}_{n+1}$. Tanto più sono vicine, tante meno iterazioni serviranno, e viceversa. Se in un certo intervallo di tempo la soluzione non varia molto, allora è plausibile pensare di prendere i successivi passi temporali più grandi. Viceversa, se varia molto, può essere necessario prendere i successivi passi temporali più piccoli. La velocità di convergenza del metodo di Newton inesatto è un indicatore della variazione della soluzione. Il metodo di Newton inesatto a passo variabile potrebbe essere implementato nel seguente modo ($\theta \neq 0$):

- $r = 0$
- $\boldsymbol{x}^{(r)} = \boldsymbol{y}_n$
- $F_n(\boldsymbol{x}^{(r)}) = \boldsymbol{x}^{(r)} - k_{n+1}\theta\boldsymbol{f}(t_{n+1}, \boldsymbol{x}^{(r)}) - \boldsymbol{y}_n - k_{n+1}(1-\theta)\boldsymbol{f}(t_n, \boldsymbol{y}_n)$

- $J_n(\boldsymbol{x}^{(r)}) = I - k_{n+1}\theta \left(\frac{\partial f_i(t_{n+1}, \boldsymbol{x}^{(r)})}{\partial x_j}\right)_{ij}$
- $P_n J_n(\boldsymbol{x}^{(r)}) = L_n U_n$
- $L_n U_n \boldsymbol{\delta}^{(r)} = -P_n F_n(\boldsymbol{x}^{(r)})$
- WHILE $\|\boldsymbol{\delta}^{(r)}\| >$ Newt_tol

  $\boldsymbol{x}^{(r+1)} = \boldsymbol{x}^{(r)} + \boldsymbol{\delta}^{(r)}$

  $r = r + 1$

  $L_n U_n \boldsymbol{\delta}^{(r)} = -P_n F_n(\boldsymbol{x}^{(r)})$

  END
- $\boldsymbol{y}_{n+1} = \boldsymbol{x}^{(r)}$
- IF $r < r_{\min}$ THEN $k_{n+2} = k_{n+1}\rho$

  ELSE IF $r > r_{\max}$ THEN $k_{n+2} = k_{n+1}/\rho$

Dunque, dato il passo temporale $k_{n+1} = t_{n+1} - t_n$, il successivo passo temporale k_{n+2} è uguale a k_{n+1} se il metodo di Newton inesatto ha raggiunto la convergenza in un numero di iterazioni r compreso tra $r_{\min}$ e $r_{\max}$, è amplificato da un fattore $\rho > 1$ se il numero di iterazioni è stato più piccolo di $r_{\min}$ ed è ridotto dello stesso fattore se il numero di iterazioni è stato più grande di $r_{\max}$ (ove $r_{\max} > r_{\min}$). Ovviamente i valori di ρ, $r_{\min}$ e $r_{\max}$ dipendono dal problema. Bisognerebbe prevedere anche un numero massimo di iterazioni R dentro il ciclo WHILE del metodo di Newton: raggiunto tale numero, il passo corrente k_{n+1} andrebbe ridotto (per esempio a k_{n+1}/δ e si dovrebbe procedere nuovamente al calcolo di $\boldsymbol{y}_{n+1}$.

Nel caso in cui non sia necessario applicare il metodo di Newton (perché il problema è lineare), il numero di iterazioni da considerare potrebbe essere quello necessario ad un metodo iterativo per risolvere i sistemi lineari.

17.3.3 Caso lineare

Un caso molto frequente è quello lineare autonomo a coefficienti costanti

$$\begin{cases} \boldsymbol{y}'(t) = A\boldsymbol{y}(t) + \boldsymbol{b} \\ \boldsymbol{y}(t_0) = \boldsymbol{y}_0 \end{cases}$$

con passo di integrazione k costante. In tal caso, il metodo si scrive

$$(I - k\theta A)\boldsymbol{y}_{n+1} = (I + k(1-\theta)A)\boldsymbol{y}_n + k\boldsymbol{b}$$

Nel caso implicito, si tratta dunque di risolvere un sistema *lineare* (non è necessario quindi il metodo di Newton) di matrice $I - k\theta A$ ad ogni passo. Pertanto, per problemi di piccola dimensione, è conveniente precalcolare la fattorizzazione LU della matrice. Altrimenti, si può considerare un metodo iterativo, ove si scelga come vettore iniziale per il calcolo di $\boldsymbol{y}_{n+1}$ la soluzione al passo precedente $\boldsymbol{y}_n$. Se il termine $\boldsymbol{b}$ dipende invece dal tempo, allora $k\boldsymbol{b}$ va sostituito, ovviamente, con $k(1-\theta)\boldsymbol{b}(t_n) + k\theta\boldsymbol{b}(t_{n+1})$. thetamlin.m

17.4 Verifica della correttezza dell'implementazione

Supponiamo di aver implementato un metodo di ordine p per la soluzione del sistema differenziale

$$\begin{cases} \boldsymbol{y}'(t) = \boldsymbol{f}(t, \boldsymbol{y}(t)) \\ \boldsymbol{y}(t_0) = \boldsymbol{y}_0 \end{cases}$$

e di volerne testare la corretta implementazione. L'idea è quella di creare una soluzione artificiale $\boldsymbol{x}(t)$, inserirla nell'equazione e calcolarne il residuo

$$\boldsymbol{x}'(t) - \boldsymbol{f}(t, \boldsymbol{x}(t)) = \boldsymbol{g}(t)$$

A questo punto, si risolve il sistema differenziale

$$\begin{cases} \boldsymbol{y}'(t) = \boldsymbol{f}(t, \boldsymbol{y}(t)) + \boldsymbol{g}(t) = \hat{\boldsymbol{f}}(t, \boldsymbol{y}(t)) \\ \boldsymbol{y}(t_0) = \boldsymbol{x}(t_0) \end{cases}$$

fino ad un tempo $t_0 + t^*$ fissato, con due discretizzazioni di passo costante $k_1 = t^*/m_1$ e $k_2 = t^*/m_2$, rispettivamente. Si avranno errori finali $\boldsymbol{e}_{m_1,k_1} = \|\boldsymbol{y}_{m_1,k_1} - \boldsymbol{x}(t_0 + t^*)\| = Ck_1^p$ e $\boldsymbol{e}_{m_2,k_2} = \|\boldsymbol{y}_{m_2,k_2} - \boldsymbol{x}(t_0 + t^*)\| = Ck_2^p$. Si ha dunque

$$\frac{\boldsymbol{e}_{m_2,k_2}}{\boldsymbol{e}_{m_1,k_1}} = \left(\frac{k_2}{k_1}\right)^p ,$$

da cui

$$\log \boldsymbol{e}_{m_2,k_2} - \log \boldsymbol{e}_{m_1,k_1} = p(\log k_2 - \log k_1) = -p(\log m_2 - \log m_1).$$

Dunque, rappresentando in un grafico logaritmico-logaritmico l'errore in dipendenza dal numero di passi, la pendenza della retta corrisponde all'ordine del metodo, cambiato di segno. Tale verifica è valida anche nel caso di passi non costanti.

Nel caso $\boldsymbol{f}(t, \boldsymbol{y}(t))$ sia particolarmente complicato, invece di calcolare il residuo, si può calcolare una *soluzione di riferimento* $\boldsymbol{y}_{\bar{m},\bar{k}}$ e poi confrontare con essa le soluzioni $\boldsymbol{y}_{m_1,k_1}$ e $\boldsymbol{y}_{m_2,k_2}$, ove $m_1, m_2 \ll \bar{m}$. In questo caso, però, si può mostrare solo che il metodo converge con l'ordine giusto ad *una* soluzione, non necessariamente quella giusta.

Capitolo 18

Metodi multistep

18.1 Metodi di Adams–Bashforth

Invece di costruire la soluzione $\boldsymbol{y}_{n+1}$ a partire dalla sola soluzione al passo precedente $\boldsymbol{y}_n$, si può pensare di usare le soluzioni di più passi precedenti. Fissato s numero naturale maggiore di 0 e una discretizzazione dell'intervallo $[t_0, t_0 + t^*]$ in m passi di ampiezza costante k, data la formula di risoluzione

$$\boldsymbol{y}(t_{n+s}) = \boldsymbol{y}(t_{n+s-1}) + \int_{t_{n+s-1}}^{t_{n+s}} \boldsymbol{f}(\tau, \boldsymbol{y}(\tau))\mathrm{d}\tau \tag{18.1}$$

l'idea è quella di sostituire la funzione integranda in (18.1) con il "suo" polinomio interpolatore sui nodi equispaziati $t_n, t_{n+1}, \ldots, t_{n+s-1}$ ($t_{n+j} = t_n + jk$)

$$\boldsymbol{p}(\tau) = \sum_{j=0}^{s-1} L_j(\tau)\boldsymbol{f}(t_{n+j}, \boldsymbol{y}_{n+j})$$

ove $L_j(t)$ è il polinomio elementare di Lagrange di grado $s-1$ definito da $L_j(t_{n+i}) = \delta_{ij}$. Poiché $\boldsymbol{p}(t_{n+j}) = \boldsymbol{f}(t_{n+j}, \boldsymbol{y}_{n+j})$, $0 \le j \le s-1$ (e non, ovviamente, $\boldsymbol{p}(t_{n+j}) = \boldsymbol{f}(t_{n+j}, \boldsymbol{y}(t_{n+j}))$), dobbiamo supporre di avere già a disposizione i valori iniziali $\boldsymbol{y}_{n+j} \approx \boldsymbol{y}(t_{n+j})$, $0 \le j \le s-1$. Si ha dunque

$$\int_{t_{n+s-1}}^{t_{n+s}} \boldsymbol{p}(\tau)\mathrm{d}\tau = \sum_{j=0}^{s-1} \left(\int_{t_{n+s-1}}^{t_{n+s}} L_j(\tau)\mathrm{d}\tau \right) \boldsymbol{f}(t_{n+j}, \boldsymbol{y}_{n+j}) = k \sum_{j=0}^{s-1} b_j \boldsymbol{f}(t_{n+j}, \boldsymbol{y}_{n+j})$$

da cui il metodo esplicito *multistep Adams–Bashforth*

$$\boldsymbol{y}_{n+s} = \boldsymbol{y}_{n+s-1} + k \sum_{j=0}^{s-1} b_j \boldsymbol{f}(t_{n+j}, \boldsymbol{y}_{n+j}) \tag{18.2}$$

I coefficienti b_j non dipendono da n e neanche da k: infatti

$$b_j = \frac{1}{k}\int_{t_{n+s-1}}^{t_{n+s}} \prod_{\substack{i=0\\ i\neq j}}^{s-1} \frac{\tau - t_{n+i}}{t_{n+j} - t_{n+i}} \mathrm{d}\tau = \int_0^1 \prod_{\substack{i=0\\ i\neq j}}^{s-1} \frac{t_{n+s-1} + rk - t_{n+i}}{t_{n+j} - t_{n+i}} \mathrm{d}r =$$

$$= \int_0^1 \prod_{\substack{i=0\\ i\neq j}}^{s-1} \frac{((s-1-i)k + rk)}{(j-i)k} \mathrm{d}r = \int_0^1 \prod_{\substack{i=0\\ i\neq j}}^{s-1} \frac{(s-1-i+r)}{(j-i)} \mathrm{d}r$$

Dunque possono essere calcolati una volta per tutte. Calcoliamo l'ordine di tale metodo: come al solito, dobbiamo valutare l'espressione

$$\boldsymbol{y}(t_{n+s}) - \boldsymbol{y}(t_{n+s-1}) - k\sum_{j=0}^{s-1} b_j \boldsymbol{f}(t_{n+j}, \boldsymbol{y}(t_{n+j}))$$

L'ultimo termine è l'integrale del polinomio $\boldsymbol{q}(\tau)$ di grado $s-1$ che interpola $\boldsymbol{f}(t_{n+j}, \boldsymbol{y}(t_{n+j}))$, $0 \le j \le s-1$. Dunque, per $t_{n+s-1} \le \tau \le t_{n+s}$,

$$\|\boldsymbol{f}(\tau, \boldsymbol{y}(\tau)) - \boldsymbol{q}(\tau)\| \le \frac{\|\boldsymbol{y}^{(s+1)}(\bar{\tau})\|}{s!} \underbrace{|(\tau - t_n)\cdot \ldots \cdot (\tau - t_{n+s-1})|}_{s \text{ termini}} \le$$

$$\le \frac{\|\boldsymbol{y}^{(s+1)}(\bar{\tau})\|}{s!} s! k^s = \mathcal{O}(k^s), \quad t_{n+s-1} \le \tau \le t_{n+s}$$

e quindi

$$\boldsymbol{y}(t_{n+s}) - \boldsymbol{y}(t_{n+s-1}) - k\sum_{j=0}^{s-1} b_j \boldsymbol{f}(t_{n+j}, \boldsymbol{y}(t_{n+j})) =$$

$$\int_{t_{n+s-1}}^{t_{n+s}} \boldsymbol{f}(\tau, \boldsymbol{y}(\tau))\mathrm{d}\tau - k\sum_{j=0}^{s-1} b_j \boldsymbol{f}(t_{n+j}, \boldsymbol{y}(t_{n+j})) =$$

$$\int_{t_{n+s-1}}^{t_{n+s}} \boldsymbol{f}(\tau, \boldsymbol{y}(\tau))\mathrm{d}\tau - \int_{t_{n+s-1}}^{t_{n+s}} \boldsymbol{q}(\tau)\mathrm{d}\tau = \mathcal{O}(k^{s+1})$$

perché un ulteriore fattore k deriva dal fatto che si integra in un intervallo di ampiezza k. Quindi, se anche $\boldsymbol{y}_n = \boldsymbol{y}(t_n) + \mathcal{O}(k^s)$, $0 \le n \le s-1$ (queste approssimazioni *non* possono essere ottenute con il metodo stesso), il metodo è di ordine s. Calcoliamo esplicitamente i metodi che corrispondono a $s = 1$ e $s = 2$. Se $s = 1$, dobbiamo cercare il polinomio di grado 0 che interpola $\boldsymbol{f}(t_n, \boldsymbol{y}_n)$. È ovviamente $\boldsymbol{p}(\tau) \equiv \boldsymbol{f}(t_n, \boldsymbol{y}_n)$ e $b_0 = 1$, quindi

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + k\boldsymbol{f}(t_n, \boldsymbol{y}_n)$$

e si ritrova il metodo di Eulero. Nel caso $s = 2$, il polinomio interpolatore è

$$\boldsymbol{p}(\tau) = \frac{t_{n+1} - \tau}{k}\boldsymbol{f}(t_n, \boldsymbol{y}_n) + \frac{\tau - t_n}{k}\boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1})$$

e dunque

$$b_0 = \frac{1}{k}\int_{t_{n+1}}^{t_{n+2}} \frac{t_{n+1} - \tau}{k}\mathrm{d}\tau = -\frac{1}{2}, \quad b_1 = \frac{1}{k}\int_{t_{n+1}}^{t_{n+2}} \frac{\tau - t_n}{k}\mathrm{d}\tau = \frac{3}{2}$$

da cui

$$\boldsymbol{y}_{n+2} = \boldsymbol{y}_{n+1} - \frac{k}{2}\boldsymbol{f}(t_n, \boldsymbol{y}_n) + \frac{3k}{2}\boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}) \tag{18.3}$$

Il valore $\boldsymbol{y}_1$ può essere ricavato, per esempio, anche con il metodo di Eulero, poiché si ha, in tal caso, $\boldsymbol{y}_1 = \boldsymbol{y}(t_1) + \mathcal{O}(k^2)$.

18.2 Metodi lineari multistep

Una semplice generalizzazione del metodo di Adams–Bashforth è permettere che il metodo sia implicito:

$$\boldsymbol{y}_{n+s} = \boldsymbol{y}_{n+s-1} + k\sum_{j=0}^{s} b_j \boldsymbol{f}(t_{n+j}, \boldsymbol{y}_{n+j}) \tag{18.4}$$

Il caso banale $s = 0$ dà

$$\boldsymbol{y}_n = \boldsymbol{y}_{n-1} + kb_0\boldsymbol{f}(t_n, \boldsymbol{y}_n)$$

da cui, per $b_0 = 1$, si ottiene una riscrittura del metodo di Eulero implicito. Per $s = 1$ abbiamo il θ-metodo ($b_0 = (1 - \theta)$, $b_1 = \theta$). Il metodo (18.4) ha ordine massimo $s + 1$. In tal caso si chiama metodo *Adams–Moulton*. Una generalizzazione ulteriore è

$$\sum_{j=0}^{s} a_j \boldsymbol{y}_{n+j} = k\sum_{j=0}^{s} b_j \boldsymbol{f}(t_{n+j}, \boldsymbol{y}_{n+j}) \tag{18.5}$$

con la normalizzazione $a_s = 1$. Il metodo è di ordine p se, come al solito, $\boldsymbol{y}_n = \boldsymbol{y}(t_n) + \mathcal{O}(k^p)$ e

$$\sum_{j=0}^{s} a_j \boldsymbol{y}(t_{n+j}) - k\sum_{j=0}^{s} b_j \boldsymbol{f}(t_{n+j}, \boldsymbol{y}(t_{n+j})) = \mathcal{O}(k^{p+1})$$

per ogni funzione $\boldsymbol{f}$ analitica e $0 \le n \le m - s$. Siccome la verifica può risultare molto tediosa, risulta utile il seguente

Teorema 10. *Dato un metodo multistep (18.5), definiamo i due polinomi*

$$\rho(w) = \sum_{j=0}^{s} a_j w^j, \quad \sigma(w) = \sum_{j=0}^{s} b_j w^j$$

Allora il metodo è di ordine p se e solo se esiste $c \neq 0$ tale che

$$\rho(1+\xi) - \sigma(1+\xi) \cdot \ln(1+\xi) = c\xi^{p+1} + \mathcal{O}\left(\xi^{p+2}\right) \text{ per } \xi \to 0$$

Prima di vedere la traccia della dimostrazione, proviamo ad applicare il teorema a qualche caso noto. Per il metodo di Eulero si ha $\rho(1+\xi) = \xi$ e $\sigma(1+\xi) = 1$. Si ha

$$\xi - 1 \cdot \left(\xi - \frac{\xi^2}{2} + \mathcal{O}(\xi^3)\right) = \frac{\xi^2}{2} + \mathcal{O}(\xi^3)$$

e dunque il metodo è di ordine 1, come noto. Per il metodo di Adams–Bashforth (18.3) si ha $\rho(1+\xi) = (1+\xi)^2 - (1+\xi) = \xi^2 + \xi$ e $\sigma(1+\xi) = 3(1+\xi)/2 - 1/2$. Si ha

$$(\xi^2 + \xi) - \left(\frac{3}{2}(1+\xi) - \frac{1}{2}\right) \cdot \left(\xi - \frac{\xi^2}{2} + \frac{\xi^3}{3} - \mathcal{O}(\xi^4)\right) = \frac{5}{12}\xi^3 + \mathcal{O}(\xi^4)$$

e dunque il metodo è di ordine 2, come noto.

Traccia della dimostrazione del Teorema 10. Si ha

$$\begin{aligned}
&\sum_{j=0}^{s} a_j \boldsymbol{y}(t_{n+j}) - k \sum_{j=0}^{s} b_j \boldsymbol{f}(t_{n+j}, \boldsymbol{y}(t_{n+j})) = \\
&\qquad \sum_{j=0}^{s} a_j \boldsymbol{y}(t_n + jk) - k \sum_{j=0}^{s} b_j \boldsymbol{y}'(t_n + jk) = \\
&\quad = \sum_{j=0}^{s} a_j \sum_{i=0}^{\infty} \boldsymbol{y}^{(i)}(t_n) \frac{j^i k^i}{i!} - k \sum_{j=0}^{s} b_j \sum_{i=0}^{\infty} \boldsymbol{y}^{(i+1)}(t_n) \frac{j^i k^i}{i!} = \\
&= \sum_{j=0}^{s} a_j \left(\boldsymbol{y}(t_n) + \sum_{i=1}^{\infty} \boldsymbol{y}^{(i)}(t_n) \frac{j^i k^i}{i!}\right) - k \sum_{j=0}^{s} b_j \sum_{i=1}^{\infty} \boldsymbol{y}^{(i)}(t_n) \frac{j^{i-1} k^{i-1}}{(i-1)!} = \\
&\quad = \left(\sum_{j=0}^{s} a_j\right) \boldsymbol{y}(t_n) + \sum_{i=1}^{\infty} \frac{1}{i!} \left(\sum_{j=0}^{s} j^i a_j - i \sum_{j=0}^{s} j^{i-1} b_j\right) k^i \boldsymbol{y}^{(i)}(t_n)
\end{aligned}$$

Dunque l'ordine è p se e solo se i coefficienti delle potenze fino a p di k sono nulli, e cioè se le $2s+1$ incognite $\{a_j\}_{j=0}^{s-1}$ e $\{b_j\}_{j=0}^{s}$ soddisfano il sistema lineare

$$\sum_{j=0}^{s} j^i a_j = i \sum_{j=0}^{s} j^{i-1} b_j, \quad i = 0, 1, \ldots, p \tag{18.6}$$

Da notare che, per $i = j = 0$, si intende $j^i = 1$. Per finire la dimostrazione, si calcola lo sviluppo in serie di Taylor di $\rho(\mathrm{e}^z) - \sigma(\mathrm{e}^z)z$ per $z \to 0$ e si osserva che esso è $\mathcal{O}(z^{p+1})$ se e solo se vale (18.6). Posto $w = \mathrm{e}^z$, l'ordine è p se e solo se lo sviluppo di $\rho(w) - \sigma(w) \cdot \ln w$ per $w \to 1$ vale

$$-c(\ln w)^{p+1} + \mathcal{O}\left((\ln w)^{p+2}\right) = c(w-1)^{p+1} + \mathcal{O}\left((w-1)^{p+2}\right)$$

A questo punto, si pone $\xi = w - 1$. □

Non bisogna dimenticare che l'ordine è p se $\boldsymbol{y}(t)$ è almeno di classe $\mathcal{C}^{p+1}$. Pertanto, qualora ci si aspetti una soluzione non regolare, potrebbe non essere adatto un metodo di ordine alto. In realtà, anche le condizioni (18.6) sono molto utili per determinare l'ordine di un metodo multistep: per esempio, per il metodo lineare multistep (implicito) a due passi

$$\begin{aligned} &\boldsymbol{y}_{n+2} - 3\boldsymbol{y}_{n+1} + 2\boldsymbol{y}_n = \\ &\qquad = \frac{k}{12}\left[-5\boldsymbol{f}(t_n, \boldsymbol{y}_n) - 20\boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}) + 13\boldsymbol{f}(t_{n+2}, \boldsymbol{y}_{n+2})\right] \end{aligned} \tag{18.7}$$

si ha

$$\begin{aligned} &a_0 + a_1 + a_2 = 0 \\ &\quad a_1 + 2a_2 = b_0 + b_1 + b_2 \quad \Rightarrow \text{ordine (almeno) } 1 \\ &\quad a_1 + 4a_2 = 2(b_1 + 2b_2) \quad \Rightarrow \text{ordine (almeno) } 2 \\ &\quad a_1 + 8a_2 \neq 3(b_1 + 4b_2) \quad \Rightarrow \text{ordine } 2 \end{aligned}$$

18.2.1 Implementazione dei metodi multistep

Dati iniziali, scrivere come $y_{n+1} =$

18.2.2 Metodi BDF

I metodi BDF *(Backward Differentiation Formulas)* sono metodi multistep impliciti a s passi, di ordine s e con $\sigma(w) = \beta w^s$. Dato $\sigma(w)$ e le condizioni d'ordine, si può costruire $\rho(w)$. Poiché però tali metodi sono della forma

$$\sum_{j=0}^{s} a_j \boldsymbol{y}_{n+j} = k b_s \boldsymbol{f}(t_{n+s}, \boldsymbol{y}_{n+s})$$

e $kb_s \boldsymbol{f}(t_{n+s}, \boldsymbol{y}_{n+s}) \approx kb_s \boldsymbol{y}'(t_n + sk)$, conviene cercare una combinazione lineare di $\boldsymbol{y}_{n+j}$, $0 \le j \le s$ che approssimi $kb_s \boldsymbol{y}'(t_{n+s})$. Si procede dunque con lo sviluppo in serie di Taylor di $\boldsymbol{y}(t_n + jk)$, $0 \le j \le s$, centrato in $\boldsymbol{y}(t_n + sk)$. Per esempio, per $s = 1$,

$$\begin{cases} \boldsymbol{y}(t_n) = \boldsymbol{y}(t_n + k) - k\boldsymbol{y}'(t_n + k) + \mathcal{O}(k^2) \\ \boldsymbol{y}(t_n + k) = \boldsymbol{y}(t_n + k) \end{cases}$$

da cui $a_1 = 1$ e $a_0 = -1$. Dunque, il metodo BDF di ordine 1 è il metodo di Eulero implicito *(backward Euler)*. Per $s = 2$

$$\begin{cases} \boldsymbol{y}(t_n) = \boldsymbol{y}(t_n + 2k) - 2k\boldsymbol{y}'(t_n + 2k) + \dfrac{4k^2}{2}\boldsymbol{y}''(t_n + 2k) + \mathcal{O}(k^3) \\ \boldsymbol{y}(t_n + k) = \boldsymbol{y}(t_n + 2k) - k\boldsymbol{y}'(t_n + 2k) + \dfrac{k^2}{2}\boldsymbol{y}''(t_n + 2k) + \mathcal{O}(k^3) \\ \boldsymbol{y}(t_n + 2k) = \boldsymbol{y}(t_n + 2k) \end{cases}$$

da cui

$$\begin{cases} a_2 = 1 \\ a_0 + a_1 + a_2 = 0 \\ -2a_0 - a_1 = b_2 \\ 2a_0 + \dfrac{a_1}{2} = 0 \end{cases} \Rightarrow \begin{cases} a_0 = \dfrac{1}{3} \\ a_1 = -\dfrac{4}{3} \\ a_2 = 1 \\ b_2 = \dfrac{2}{3} \end{cases} \tag{18.8}$$

e il metodo è di ordine 2.

In generale, un metodo BDF a s passi può essere scritto nella forma

$$\sum_{j=1}^{s} \frac{1}{j} \nabla^j \boldsymbol{y}_{n+s} = k\boldsymbol{f}(t_{n+s}, \boldsymbol{y}_{n+s})$$

ove $\nabla^0 \boldsymbol{y}_{n+s} = \boldsymbol{y}_{n+s}$ e $\nabla^j \boldsymbol{y}_{n+s} = \nabla^{j-1}\boldsymbol{y}_{n+s} - \nabla^{j-1}\boldsymbol{y}_{n+s-1}$. In questo modo però $a_s \neq 1$ per $s > 1$.

I metodi BDF sono gli unici metodi multistep in cui non è difficile calcolare i coefficienti anche nel caso di passi temporali variabili. Sempre per $s = 2$, se $t_{n+1} = t_n + k_{n+1}$ e $t_{n+2} = t_{n+1} + k_{n+2}$, allora

$$\begin{cases} \boldsymbol{y}(t_n) = \boldsymbol{y}(t_n + k_{n+1} + k_{n+2}) - (k_{n+1} + k_{n+2})\boldsymbol{y}'(t_n + k_{n+1} + k_{n+2}) + \\ \quad + \dfrac{(k_{n+1} + k_{n+2})^2}{2}\boldsymbol{y}''(t_n + k_{n+1} + k_{n+2}) + \dots \\ \boldsymbol{y}(t_n + k_{n+1}) = \boldsymbol{y}(t_n + k_{n+1} + k_{n+2}) - k_{n+2}\boldsymbol{y}'(t_n + k_{n+1} + k_{n+2}) + \\ \quad + \dfrac{k_{n+2}^2}{2}\boldsymbol{y}''(t_n + k_{n+1} + k_{n+2}) + \dots \\ \boldsymbol{y}(t_n + k_{n+1} + k_{n+2}) = \boldsymbol{y}(t_n + k_{n+1} + k_{n+2}) \end{cases}$$

da cui i coefficienti

$$\begin{cases} a_2 = 1 \\ a_0 + a_1 + a_2 = 0 \\ -a_0(k_{n+1} + k_{n+2}) - a_1 k_{n+2} = b_2 k_{n+2} \\ \frac{a_0}{2}(k_{n+1} + k_{n+2})^2 + \frac{a_1}{2} k_{n+2}^2 = 0 \end{cases} \Rightarrow \begin{cases} a_0 = \dfrac{k_{n+2}^2}{(k_{n+1} + 2k_{n+2})k_{n+1}} \\ a_1 = -\dfrac{(k_{n+1} + k_{n+2})^2}{(k_{n+1} + 2k_{n+2})k_{n+1}} \\ a_2 = 1 \\ b_2 = \dfrac{(k_{n+1} + k_{n+2})}{(k_{n+1} + 2k_{n+2})} \end{cases}$$

Va notato però che il metodo che ne risulta in generale *non* converge se $k_{n+2}/k_{n+1} \geq 1 + \sqrt{2}$. E rimane aperto poi il problema di scegliere come cambiare il passo (vedi però il paragrafo 17.3.2). Questi metodi risultano particolarmente vantaggiosi quando la valutazione della funzione $\boldsymbol{f}$ è onerosa, poiché permettono di raggiungere un ordine elevato con una sola valutazione (nel caso lineare, altrimenti è necessario valutare $\boldsymbol{f}$ in un ciclo di Newton ad ogni passo temporale).

18.3 Consistenza e stabilità

Dalle condizioni d'ordine (18.6), si vede che un metodo lineare multistep è consistente se

$$\rho(1) = 0, \quad \rho'(1) = \sigma(1)$$

La consistenza è una condizione necessaria per la convergenza: infatti, se consideriamo il semplice problema differenziale

$$\begin{cases} y'(t) = 0, \quad t \in (0, t^*] \\ y(0) = y_0 \end{cases} \tag{18.9}$$

una soluzione numerica al passo $n = m - s$, $k = t^*/m$ soddisfa

$$a_0 y_{m-s,m} + a_1 y_{m-s+1,m} + \ldots + a_s y_{m,m} = 0$$

Se il metodo è convergente, $y_{m-s+j,m} \to y(t^*)$, $0 \leq j \leq s$, per $m \to \infty$ e dunque

$$0 = a_0 y_{m-s,m} + a_1 y_{m-s+1,m} + \ldots + a_s y_{m,m} \to (a_0 + a_1 + \ldots + a_s) y(t^*) = \rho(1) y(t^*)$$

da cui $\rho(1) = 0$. Lo stesso metodo applicato invece a

$$\begin{cases} y'(t) = 1, \quad t \in (0, t^*] \\ y(0) = 0 \end{cases}$$

ammette una soluzione numerica del tipo $y_{n,m} = nk\alpha = t_{n,m}\alpha$. Infatti, da

$$\alpha(a_0 nk + a_1(n+1)k + \ldots + a_s(n+s)k) = k(b_0 + b_1 + \ldots + b_s)$$

si ricava

$$\alpha[n(a_0 + a_1 + \ldots + a_s) + (a_1 + 2a_2 + \ldots + sa_s)] = \alpha[n\rho(1) + \rho'(1)] = \sigma(1)$$

Se il metodo è convergente, allora $\rho(1) = 0$ e dunque $\alpha = \sigma(1)/\rho'(1)$. Siccome la soluzione analitica è $y(t) = t$, $y_{m,m} \to t_{m,m} = t^*$, per $m \to \infty$ e dunque $\alpha = 1$.

La consistenza, però, non è sufficiente ad assicurare la convergenza di un metodo. Consideriamo l'applicazione del metodo a due passi del secondo ordine (18.7) al problema differenziale (18.9) la cui soluzione è evidentemente $y(t) \equiv y_0$. Si ha

$$y_{n+2} - 3y_{n+1} + 2y_n = 0$$

Prendiamo $y_1 = y_0 + \mathcal{O}(k^2) = y_0 + ck^2$ (valore iniziale lecito per un metodo multistep a due passi di ordine due). Allora, $y_2 = 3y_0 + 3ck^2 - 2y_0 = y_0 + c3k^2$. Si ha poi $y_3 = y_0 + c7k^2$. In generale, si ha

$$y_n = y_0 + c(2^n - 1)k^2$$

Dunque, se il numero di passi è $m = t^*/k$, si ha

$$y(t^*) \approx y_m = y_0 + c(2^m - 1)\left(\frac{t^*}{m}\right)^2$$

e facendo tendere $m \to \infty$ (o, equivalentemente, $k \to 0$), si ha $y_m \to \infty$. Abbiamo quindi un metodo la cui soluzione numerica diverge facendo tendere il passo temporale a 0 (cioè proprio l'opposto di quanto dovrebbe succedere). È proprio un piccolo errore commesso ad un passo (in questo caso al solo passo y_1) che si accumula in maniera distruttiva. Per il problema (18.9) consideriamo un metodo ad s passi di ordine p i cui primi valori siano

$$z_j = y_0 + c\theta^j k^p, \quad 0 \le j \le s-1$$

ove θ è una radice del polinomio $\rho(w)$ (che d'ora in poi chiameremo *caratteristico*). Sono valori accettabili, anche per $j = 0$, in quanto distano $\mathcal{O}(k^p)$ dalla soluzione analitica. Ora z_s si trova risolvendo

$$\sum_{j=0}^{s} a_j z_j = 0$$

da cui

$$z_s = -\sum_{j=0}^{s-1} a_j z_j = -y_0 \sum_{j=0}^{s-1} a_j - ck^p \sum_{j=0}^{s-1} a_j \theta^j = y_0 + c\theta^s k^p$$

ove si è usato $\sum_j a_j = 0$ e $\rho(\theta) = 0$. Per induzione si arriva a provare che

$$z_{n,k} = z_n = y_0 + c\theta^n k^p$$

è una soluzione di (18.9), con $z_{n,k} - y(t_n) = z_{n,k} - y_0 = c\theta^n k^p$. Dunque il metodo è consistente ma, se vogliamo che si abbia $\lim_{m\to\infty} z_{m,k} = \lim_{m\to\infty} z_{m,t^*/m} = y(t^*) = y_0$ occorre che $|\theta| \leq 1$. Se θ è radice multipla di $\rho(w)$, allora è radice anche di $\rho'(w)$ e pertanto soddisfa anche

$$\rho'(\theta) = \sum_{j=0}^{s} a_j j\theta^{j-1} = 0$$

Allora anche $z_{n,k} = z_n = y_0 + cn\theta^n k^p$ è una soluzione numerica. Infatti

$$\sum_{j=0}^{s} a_j z_{n+j} = y_0 \sum_{j=0}^{s} a_j + ck^p \left(\theta^n \sum_{j=0}^{s} a_j n\theta^j + \theta^{n+1} \sum_{j=0}^{s} a_j j\theta^{j-1} \right) =$$
$$= 0 + ck^p\theta^n n\rho(\theta) + ck^p\theta^{n+1}\rho'(\theta) = 0$$

ed è generata dai valori iniziali $z_j = y_0 + cj\theta^j k^p$, $0 \leq j \leq s-1$. Se vogliamo che $z_{n,k}$ converga alla soluzione analitica per $m \to \infty$, deve essere $|\theta| < 1$. Se il polinomio caratteristico ha radici semplici θ tali che $|\theta| \leq 1$ e radici multiple θ tali che $|\theta| < 1$, diremo che il polinomio soddisfa la *condizione delle radici*. Consideriamo adesso $y_n = y_0$ e $z_n = y_0 + c\theta^n k^p$, θ radice del polinomio caratteristico. Sono entrambe soluzioni consistenti del problema e le seconde sono state generate ammettendo delle perturbazioni $\delta_j = c\theta^j k^p$ per i primi s termini. Dunque

$$|z_{j,k} - y_{j,k}| = |\delta_j| \leq \max_{0\leq j\leq s-1} c|\theta|^j k^p$$

mentre le soluzioni distano

$$|z_{n,k} - y_{n,k}| = c|\theta|^n k^p = |\theta||z_{n-1,k} - y_{n-1,k}|$$

e se $|\theta| > 1$, tali differenze non sono limitate. Dunque, un insieme di perturbazioni limitate può produrre un insieme di soluzioni non limitate. Allo

stesso modo, se si considerano le soluzioni $y_n = y_0$ e $z_n = y_0 + cn\theta^n k^p$, θ radice multipla del polinomio caratteristico, si ha

$$|z_{j,k} - y_{j,k}| = |\delta_j| \leq \max_{0 \leq j \leq s-1} cj|\theta|^j k^p$$

mentre

$$|z_{n,k} - y_{n,k}| = cn|\theta|^n k^p = |\theta| \frac{n}{n-1} |z_{n-1,k} - y_{n-1,k}|$$

e quindi le soluzioni divergono se $|\theta| \geq 1$. È giustificata allora (in analogia con quanto visto al paragrafo 11.2.7) la seguente

Definizione 4. *Dato un metodo lineare multistep* (18.5)*, siano* $\boldsymbol{z}_n^i$*,* $i = 1, 2$*, due perturbazioni della soluzione definite da*

$$\begin{aligned} &\boldsymbol{z}_j^i = \boldsymbol{y}_j + \boldsymbol{\delta}_j^i, && 0 \leq j \leq s-1 \\ &\sum_{j=0}^{s} a_j \boldsymbol{z}_{n+j}^i = k \sum_{j=0}^{s} b_j \boldsymbol{f}(t_{n+j}, \boldsymbol{z}_{n+j}^i) + k\boldsymbol{\delta}_{n+s}^i, && 0 \leq n \leq m-s \end{aligned}$$

Se da $\max_n \|\boldsymbol{\delta}_n^1 - \boldsymbol{\delta}_n^2\| \leq \varepsilon$ *segue* $\max_n \|\boldsymbol{z}_n^1 - \boldsymbol{z}_n^2\| \leq C\varepsilon$ *per* k *piccolo a piacere, allora il metodo* (18.5) *si dice* (zero-)stabile.

Oltre alla consistenza, l'altro ingrediente per avere la convergenza di un metodo è proprio la stabilità. Infatti, posto, come al solito,

$$\boldsymbol{y}_{n+s}^* + \sum_{j=0}^{s-1} a_j \boldsymbol{y}(t_{n+j}) = k \sum_{j=0}^{s-1} b_j \boldsymbol{f}(t_{n+j}, \boldsymbol{y}(t_{n+j})) + kb_s \boldsymbol{f}(t_{n+s}, \boldsymbol{y}_{n+s}^*)$$

(e $\boldsymbol{y}_j^* = \boldsymbol{y}(t_j)$, $0 \leq j \leq s-1$) l'errore al passo $n+s$ può essere espresso come

$$\boldsymbol{e}_{n+s} = \boldsymbol{y}_{n+s} - \boldsymbol{y}(t_{n+s}) = (\boldsymbol{y}_{n+s} - \boldsymbol{y}_{n+s}^*) + (\boldsymbol{y}_{n+s}^* - \boldsymbol{y}(t_{n+s})) \tag{18.10}$$

ove il secondo termine è dell'ordine dell'errore locale e il primo termine tiene conto dell'accumulazione degli errori ai passi precedenti, cioè delle *perturbazioni* tra la soluzione esatta e la soluzione numerica ai passi precedenti. Guardando la rappresentazione dell'errore (18.10), si vede che il primo termine $(\boldsymbol{y}_{n+s} - \boldsymbol{y}_{n+s}^*)$ è la differenza tra due particolari soluzioni perturbate $\boldsymbol{z}_n^1$ e $\boldsymbol{z}_n^2$ corrispondenti a

$$\boldsymbol{\delta}_j^1 = 0, \quad 0 \leq j \leq m$$

e

$$\begin{aligned}
\boldsymbol{\delta}_j^2 &= \boldsymbol{y}(t_j) - \boldsymbol{y}_j, && 0 \le j \le s-1 \\
\boldsymbol{\delta}_{n+s}^2 &= -\sum_{j=s-n}^{s-1} b_j \boldsymbol{f}(t_{j+n}, \boldsymbol{z}_{j+n}^2) + \sum_{j=s-n}^{s-1} b_j \boldsymbol{f}(t_{j+n}, \boldsymbol{y}(t_{j+n})) + \\
&\quad + \frac{1}{k} \sum_{j=s-n}^{s-1} a_j \left(\boldsymbol{z}_{j+n}^2 - \boldsymbol{y}(t_{j+n}) \right), && 0 \le n \le s \\
\boldsymbol{\delta}_{n+s}^2 &= -\sum_{j=0}^{s-1} b_j \boldsymbol{f}(t_{j+n}, \boldsymbol{z}_{j+n}^2) + \sum_{j=0}^{s-1} b_j \boldsymbol{f}(t_{j+n}, \boldsymbol{y}(t_{j+n})) + \\
&\quad + \frac{1}{k} \sum_{j=0}^{s-1} a_j \left(\boldsymbol{z}_{j+n}^2 - \boldsymbol{y}(t_{j+n}) \right), && s+1 \le n \le m-s
\end{aligned}$$

Dunque, perché un metodo sia convergente (cioè l'errore tenda a zero con k), occorre che le perturbazioni della soluzione introdotte ad ogni passo da errori di approssimazione del metodo stesso rimangano limitate (ciò non fa esplodere la prima parte di (18.10)) e che l'errore locale tenda a zero con k (ciò fa andare a zero la seconda parte di (18.10)).

Abbiamo visto che la condizione delle radici è necessaria affinché perturbazioni limitate generino soluzioni limitate. In realtà essa è anche sufficiente. Si ha infatti il seguente teorema fondamentale:

Teorema 11 (Equivalenza di Dahlquist)**.** *Un metodo lineare multistep con valori iniziali consistenti è convergente se e solo se è consistente e stabile (cioè il suo polinomio caratteristico soddisfa la condizione delle radici).*

La grande portata di questo teorema è che il risultato è valido non solo per il problema modello (18.9). Inoltre, se i valori iniziali approssimano con ordine p le soluzioni analitiche, allora il metodo è convergente con ordine p. La dimostrazione della sufficienza della consistenza e della condizione delle radici passa attraverso la riscrittura di un metodo multistep come un metodo ad un passo in uno spazio di dimensione maggiore. A quel punto, la dimostrazione procede essenzialmente come nel caso del θ-metodo, introducento una opportuna funzione $\psi(t, \boldsymbol{y})$, lipschitziana con costante Λ. Si ottiene quindi

$$\|\boldsymbol{E}_{n+1}\| \le (1 + k\Lambda)\|\boldsymbol{E}_n\| + ck^{p+1}$$

in cui il valore 1 deriva dal fatto che il metodo è stabile. Dalla disuguaglianza sopra si ricava

$$\|\boldsymbol{E}_n\| \le (1 + k\Lambda)^n \|\boldsymbol{E}_0\| + \frac{(1 + k\Lambda)^n - 1}{(1 + k\Lambda) - 1} ck^{p+1}$$

msinstabile.m

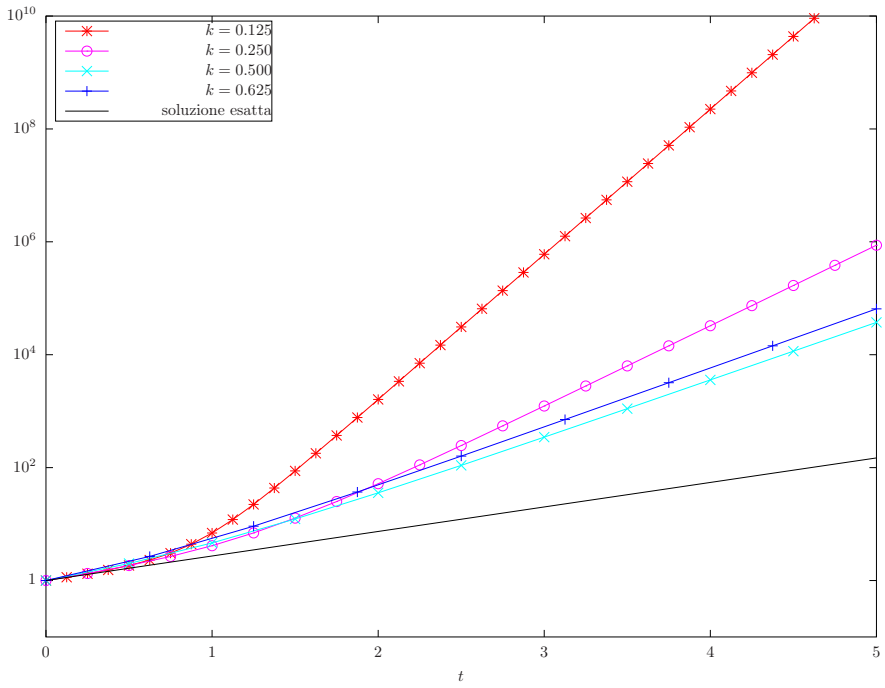


Figura 18.1: Applicazione del metodo (18.7) al problema differenziale $y'(t) = y(t)$, $y(0) = 1$.

Il vettore $\boldsymbol{E}_0$ contiene tutte le differenze dei valori iniziali $\boldsymbol{y}_j - \boldsymbol{y}(t_j)$, $j = 0, 1, \ldots, s-1$ che, anche se non nulle, sono di ordine k^p. L'importante è che $(1+k\Lambda)^n \leq \mathrm{e}^{t^*\Lambda}$ (mentre invece, per esempio, $(2+k\Lambda)^m = 2^m(1+k\Lambda/2)^m = 2^m(1+t^*\Lambda/(2m))^m = 2^m[(1+1/(2m/(t^*\Lambda)))^{2m/(t^*\Lambda)}]^{t^*\Lambda/2} \to \infty$ per $m \to \infty$, cioè $k \to 0$). Ritornando al metodo (18.7), si ha che $\theta = 2$ è radice del polinomio caratteristico e pertanto, il metodo non è stabile, come si vede anche in Figura 18.1 ove il metodo è stato applicato al problema differenziale

$$\begin{cases} y'(t) = y(t), & t \in (0, 5] \\ y(0) = 1 \end{cases}$$

Come corollario al teorema precedente, abbiamo che ogni metodo ad un passo è stabile (perché $\rho(w) = w - 1$) e che i metodi di Adams–Bashforth e Adams–Moulton sono stabili (perché $\rho(w) = w^s - w^{s-1}$). Esiste un limite superiore per l'ordine di un metodo a s passi, dato dal seguente

Teorema 12 (Prima barriera di Dahlquist)**.** *Il massimo ordine per un metodo a s passi convergente è $2\lfloor (s+2)/2 \rfloor$ se implicito e s se esplicito.*

Per quanto riguarda i metodi BDF (speciali metodi impliciti) si ha che sono convergenti (cioè sono stabili) solo per $1 \leq s \leq 6$.

18.4 Influenza degli errori di arrotondamento

Abbiamo visto che la stabilità (assieme alla consistenza) assicura che perturbazioni del tipo $k\boldsymbol{\delta}_{n+s}$, $n \geq 0$, permettono ad un metodo di convergere. Tali perturbazioni si generano per approssimazione dei passi precedenti e, per esempio, quando si risolve un sistema (non)lineare per un metodo implicito e si chiede una tolleranza proporzionale a k^p se il metodo è di ordine p. Gli errori di arrotondamento sono però di ordine $\mathcal{O}(1)$ ed è del tutto evidente che in aritmetica di macchina non si può in generale pretendere di trovare la soluzione esatta. Vediamo esplicitamente il caso del metodo di Eulero: si avrà

$$\tilde{\boldsymbol{y}}_{n+1} = \tilde{\boldsymbol{y}}_n + k\boldsymbol{f}(t_n, \tilde{\boldsymbol{y}}_n) + \boldsymbol{\varepsilon}_{n+1}$$

e dunque, definito

$$\tilde{\boldsymbol{e}}_n = \tilde{\boldsymbol{y}}_n - \boldsymbol{y}(t_n)$$

si ottiene

$$\|\tilde{\boldsymbol{e}}_{n+1}\| \leq (1 + k\lambda)\|\tilde{\boldsymbol{e}}_n\| + ck^2 + \|\boldsymbol{\varepsilon}_{n+1}\|, \quad c > 0$$

da cui, per induzione,

$$\|\tilde{\boldsymbol{e}}_n\| \leq \frac{ck^2 + \varepsilon}{k\lambda}[(1 + k\lambda)^n - 1] + (1 + k\lambda)^n\|\tilde{\boldsymbol{e}}_0\|, \quad 0 \leq n \leq m$$

ove $\varepsilon = \max_{0\leq n\leq m}\|\boldsymbol{\varepsilon}_n\|$ e quindi

$$\|\tilde{\boldsymbol{e}}_n\| \leq (\mathrm{e}^{t^*\lambda} - 1)\left(\frac{ck}{\lambda} + \frac{\varepsilon}{k\lambda}\right) + \mathrm{e}^{t^*\lambda}\|\tilde{\boldsymbol{e}}_0\|, \quad 0 \leq n \leq m$$

Anche se non è detto che gli errori di arrotondamento si sommino ad ogni passo (potrebbero anche compensarsi in un certo qual modo), si vede che al tendere di k a 0 l'errore potrebbe esplodere. Pertanto esisterebbe un k ottimale (non eccessivamente piccolo) che minimizza l'errore globale del metodo quando implementato in aritmetica di macchina.

Nel caso si usasse un metodo non stabile la situazione sarebbe ancora peggiore, come si vede in Tabella 18.1, ove il metodo (18.7) è stato applicato al problema (18.9) avendo preso $y_0 = y_1 = 1/3$. Il problema nasce dal fatto che $y_2 = 3y_1 - 2y_0$, a causa degli errori di arrotondamento, non è uguale a y_1 e y_0.

y_2	3.333333333333334e-01
y_3	3.333333333333333e-01
y_{10}	3.333333333333051e-01
y_{20}	3.333333333042297e-01
y_{30}	3.333333035310111e-01
y_{40}	3.333028157552085e-01
y_{50}	3.020833333333335e-01
y_{60}	-3.166666666666666e+01
y_{70}	-3.276766666666667e+04
y_{80}	-3.355443166666666e+07
y_{90}	-3.435973836766667e+10
y_{100}	-3.518437208883166e+13

Tabella 18.1: Primi passi di risoluzione del problema test (18.9) con il metodo (18.7).

Capitolo 19

Metodi di Runge–Kutta

19.1 Metodi di Runge–Kutta espliciti

I metodi lineari multistep lasciano aperti alcuni problemi. Come calcolare i valori iniziali per i metodi di ordine elevato? Abbiamo visto che il massimo ordine per un metodo ad un passo convergente è 2 se implicito (lo raggiunge il solo metodo dei trapezi). È possibile modificarlo e renderlo esplicito (e dunque di più facile applicazione)? Si possono costruire metodi di ordine elevato e che permettano un passo temporale "adattabile" all'andamento della soluzione? Cominciamo a rispondere alla seconda domanda: una modifica abbastanza ovvia al metodo dei trapezi

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + \frac{k}{2}(\boldsymbol{f}(t_n, \boldsymbol{y}_n) + \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}))$$

per renderlo esplicito è sostituire $\boldsymbol{y}_{n+1}$ con $\boldsymbol{y}_n + k\boldsymbol{f}(t_n, \boldsymbol{y}_n)$ così da avere

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + \frac{k}{2}(\boldsymbol{f}(t_n, \boldsymbol{y}_n) + \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_n + k\boldsymbol{f}(t_n, \boldsymbol{y}_n)) \tag{19.1}$$

Da un punto di vista "logico", esso può essere definito come

$$\begin{aligned}
\boldsymbol{\xi}_1 &= \boldsymbol{y}_n \approx \boldsymbol{y}(t_n) \\
\boldsymbol{\xi}_2 &= \boldsymbol{y}_n + k\boldsymbol{f}(t_n, \boldsymbol{\xi}_1) \approx \boldsymbol{y}(t_{n+1}) \\
\boldsymbol{y}_{n+1} &= \boldsymbol{y}_n + \frac{k}{2}(\boldsymbol{f}(t_n, \boldsymbol{\xi}_1) + \boldsymbol{f}(t_{n+1}, \boldsymbol{\xi}_2)) \approx \boldsymbol{y}(t_{n+1})
\end{aligned}$$

Un altro modo di rendere esplicito il metodo dei trapezi è sostituire la media delle funzioni $\boldsymbol{f}$ con la funzione $\boldsymbol{f}$ valutata nel "punto medio"

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + k\boldsymbol{f}\left(t_n + \frac{k}{2}, \boldsymbol{y}_n + \frac{k}{2}\boldsymbol{f}(t_n, \boldsymbol{y}_n)\right) \tag{19.2}$$

cioè

$$\begin{aligned}\boldsymbol{\xi}_1 &= \boldsymbol{y}_n \approx \boldsymbol{y}(t_n)\\ \boldsymbol{\xi}_2 &= \boldsymbol{y}_n + \frac{k}{2}\boldsymbol{f}(t_n, \boldsymbol{\xi}_1) \approx \boldsymbol{y}\left(t_n + \frac{k}{2}\right)\\ \boldsymbol{y}_{n+1} &= \boldsymbol{y}_n + k\boldsymbol{f}\left(t_n + \frac{k}{2}, \boldsymbol{\xi}_2\right) \approx \boldsymbol{y}(t_{n+1})\end{aligned}$$

L'idea generale dei metodi espliciti di *Runge–Kutta* è quella, come al solito, di sostituire l'integrale nella formula risolutiva

$$\boldsymbol{y}(t_{n+1}) = \boldsymbol{y}(t_n) + \int_{t_n}^{t_{n+1}} \boldsymbol{f}(\tau, \boldsymbol{y}(\tau))\mathrm{d}\tau$$

con una formula di quadratura su nodi $t_n + c_j k$, $1 \leq j \leq \nu$ nell'intervallo $[t_n, t_{n+1}]$. Si giunge quindi a

$$\boldsymbol{y}(t_{n+1}) \approx \boldsymbol{y}(t_n) + k\sum_{j=1}^{\nu} b_j \boldsymbol{f}(t_n + c_j k, \boldsymbol{y}(t_n + c_j k))$$

Si tratta ora di trovare delle approssimazioni $\boldsymbol{\xi}_j$ di $\boldsymbol{y}(t_n + c_j k)$. Si procede iterativamente in questo modo

$$\begin{cases} \boldsymbol{y}(t_n) \approx \boldsymbol{y}_n = \boldsymbol{\xi}_1 \quad (\Rightarrow\ c_1 = 0)\\ \boldsymbol{y}(t_n + kc_2) \approx \boldsymbol{y}_n + kc_2\boldsymbol{f}(t_n, \boldsymbol{y}_n) = \boldsymbol{y}_n + a_{2,1}k\boldsymbol{f}(t_n, \boldsymbol{\xi}_1) = \boldsymbol{\xi}_2\\ \quad\vdots\\ \boldsymbol{y}(t_n + kc_i) \approx \boldsymbol{y}_n + k\sum_{j=1}^{i-1} a_{i,j}\boldsymbol{f}(t_n + kc_j, \boldsymbol{\xi}_j) = \boldsymbol{\xi}_i\\ \quad\vdots\\ \boldsymbol{y}(t_n + kc_\nu) \approx \boldsymbol{y}_n + k\sum_{j=1}^{\nu-1} a_{\nu,j}\boldsymbol{f}(t_n + kc_j, \boldsymbol{\xi}_j) = \boldsymbol{\xi}_\nu\\ \boldsymbol{y}_{n+1} = \boldsymbol{y}_n + k\sum_{j=1}^{\nu} b_j\boldsymbol{f}(t_n + kc_j, \boldsymbol{\xi}_j) \end{cases} \tag{19.3}$$

ove i parametri c_j, b_j e $a_{i,j}$ sono da determinare in modo da ottenere l'ordine desiderato. Il numero ν indica il numero di *stadi*. I parametri c_j, b_j e $a_{i,j}$ si racchiudono di solito nel *tableau di Butcher*. Se $\nu = 1$, ci si riconduce al

$$
\begin{array}{c|cccccc}
0 & (0) & & & & & \\
c_2 & a_{2,1} & & & & & \\
c_3 & a_{3,1} & a_{3,2} & & & & \\
\vdots & \vdots & \vdots & \ddots & & & \\
c_{\nu-1} & a_{\nu-1,1} & a_{\nu-1,2} & \dots & a_{\nu-1,\nu-2} & & \\
c_\nu & a_{\nu,1} & a_{\nu,2} & \dots & a_{\nu,\nu-2} & a_{\nu,\nu-1} & \\
\hline
 & b_1 & b_2 & \dots & b_{\nu-2} & b_{\nu-1} & b_\nu
\end{array}
$$

Tabella 19.1: Tableau di Butcher per metodi di Runge–Kutta espliciti.

metodo di Eulero. Per $\nu = 2$, l'ordine si ricava al solito modo

$$
\begin{aligned}
\boldsymbol{y}(t_{n+1}) - \boldsymbol{y}(t_n) - kb_1\boldsymbol{f}(t_n, \boldsymbol{y}(t_n)) - kb_2\boldsymbol{f}(t_n + c_2k, \boldsymbol{y}(t_n) + a_{2,1}k\boldsymbol{f}(t_n, \boldsymbol{y}(t_n))) = \\
= \boldsymbol{y}(t_n) + k\boldsymbol{y}'(t_n) + \frac{k^2}{2}\boldsymbol{y}''(t_n) + \mathcal{O}(k^3) - \boldsymbol{y}(t_n) - kb_1\boldsymbol{y}'(t_n) + \\
-kb_2\left[\boldsymbol{f}(t_n, \boldsymbol{y}(t_n)) + \frac{\partial \boldsymbol{f}}{\partial t}(t_n, \boldsymbol{y}(t_n))c_2k + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}}(t_n, \boldsymbol{y}(t_n))a_{2,1}k\boldsymbol{y}'(t_n) + \mathcal{O}(k^2)\right] = \\
= k\boldsymbol{y}'(t_n) + \frac{k^2}{2}\boldsymbol{y}''(t_n) + \mathcal{O}(k^3) - kb_1\boldsymbol{y}'(t_n) - kb_2\boldsymbol{y}'(t_n) + \\
- kb_2\frac{\partial \boldsymbol{f}}{\partial t}(t_n, \boldsymbol{y}(t_n))c_2k - k^2a_{2,1}b_2\left[\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}}(t_n, \boldsymbol{y}(t_n))\boldsymbol{y}'(t_n)\right] = \\
= k\boldsymbol{y}'(t_n) + \frac{k^2}{2}\boldsymbol{y}''(t_n) + \mathcal{O}(k^3) - kb_1\boldsymbol{y}'(t_n) - kb_2\boldsymbol{y}'(t_n) + \\
- kb_2\frac{\partial \boldsymbol{f}}{\partial t}(t_n, \boldsymbol{y}(t_n))c_2k - k^2a_{2,1}b_2\left[\boldsymbol{y}''(t_n) - \frac{\partial \boldsymbol{f}}{\partial t}(t_n, \boldsymbol{y}(t_n))\right] = \\
= k(1 - b_1 - b_2)\boldsymbol{y}'(t_n) + k^2\left(\frac{1}{2} - a_{2,1}b_2\right)\boldsymbol{y}''(t_n) + \\
- k^2(b_2c_2 - a_{2,1}b_2)\frac{\partial \boldsymbol{f}}{\partial t}(t_n, \boldsymbol{y}(t_n)) + \mathcal{O}(k^3)
\end{aligned}
$$

Dunque l'ordine è due se

$$
\begin{array}{c|cc}
0 & & \\
1 & 1 & \\
\hline
 & \frac{1}{2} & \frac{1}{2}
\end{array}
\qquad
\begin{array}{c|cc}
0 & & \\
\frac{1}{2} & \frac{1}{2} & \\
\hline
 & 0 & 1
\end{array}
\qquad
\begin{array}{c|cc}
0 & & \\
\frac{2}{3} & \frac{2}{3} & \\
\hline
 & \frac{1}{4} & \frac{3}{4}
\end{array}
$$

Tabella 19.2: Metodi di Runge–Kutta espliciti di ordine 2.

$$\begin{cases} b_1 + b_2 = 1 \\ a_{2,1} b_2 = \dfrac{1}{2} \\ b_2 c_2 = b_2 a_{2,1} \end{cases} \tag{19.4}$$

da cui, per esempio, i metodi di ordine due riportati in Tabella 19.2. I primi due corrispondono ai due metodi visti all'inizio del capitolo e si chiamano, rispettivamente, *metodo di Heun* e *metodo di Eulero modificato*. Da notare come non esista, tra le infinite soluzioni, una soluzione che annulla il termine $\mathcal{O}(k^3)$ (bisognerebbe scriverlo esplicitamente per verificarlo). Lo si può verificare indirettamente, considerando l'equazione

$$\begin{cases} y'(t) = y(t) \\ y(0) = 1 \end{cases}$$

L'applicazione del metodo sopra porge

$$\begin{cases} \xi_1 = y_n \\ \xi_2 = y_n + k a_{2,1} y_n \\ y_{n+1} = y_n + k b_1 \xi_1 + k b_2 \xi_2 = y_n + k b_1 y_n + k b_2 y_n + k^2 b_2 a_{2,1} y_n \end{cases}$$

da cui, usando le condizioni d'ordine,

$$y_{n+1} = y_n + k y_n + \frac{k^2}{2} y_n$$

Sostituendo la soluzione analitica in questo schema, si ottiene

$$\begin{aligned} y(t_{n+1}) - y(t_n) - k y(t_n) - \frac{k^2}{2} y(t_n) &= e^k y(t_n) - y(t_n) - k y(t_n) - \frac{k^2}{2} y(t_n) = \\ &= \frac{k^3}{6} y(t_n) + \mathcal{O}(k^4) \end{aligned}$$

e pertanto l'ordine è 2.

Il punto cruciale del calcolo dell'ordine per $\nu = 2$ è l'uguaglianza tra

$$\begin{aligned} \boldsymbol{y}'(t_n) + k \boldsymbol{y}''(t_n) &= \boldsymbol{y}'(t_n) + k \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{f}(t_n, \boldsymbol{y}(t_n)) = \\ &= \boldsymbol{y}'(t_n) + k \frac{\partial \boldsymbol{f}}{\partial t}(t_n, \boldsymbol{y}(t_n)) + k \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}}(t_n, \boldsymbol{y}(t_n)) \boldsymbol{y}'(t_n) \end{aligned}$$

e

$$\begin{aligned} \boldsymbol{f}(t_n + k, \boldsymbol{y}(t_n) + k \boldsymbol{f}(t_n, \boldsymbol{y}(t_n))) &= \boldsymbol{f}(t_n, \boldsymbol{y}(t_n)) + \\ &+ \frac{\partial \boldsymbol{f}}{\partial t}(t_n, \boldsymbol{y}(t_n)) k + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}}(t_n, \boldsymbol{y}(t_n)) k \boldsymbol{f}(t_n, \boldsymbol{y}(t_n)) + \mathcal{O}(k^2) = \\ &= \boldsymbol{y}'(t_n) + k \frac{\partial \boldsymbol{f}}{\partial t}(t_n, \boldsymbol{y}(t_n)) + k \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}}(t_n, \boldsymbol{y}(t_n)) \boldsymbol{y}'(t_n) + \mathcal{O}(k^2) \end{aligned}$$

(a meno di $\mathcal{O}(k^2)$) in cui le derivate di ordine superiore di $\boldsymbol{y}$ (e quindi di $\boldsymbol{f}$) sono sostituite da funzioni di funzioni $\boldsymbol{f}$.

Per ogni $\nu > 1$, il corrispondente sistema *non* lineare che si ottiene per la determinazione dell'ordine può avere infinite soluzioni. Solitamente si impone l'ulteriore vincolo (non strettamente necessario)

$$\sum_{j=1}^{i-1} a_{i,j} = c_i, \quad 2 \leq i \leq \nu$$

che forza $\boldsymbol{\xi}_i$ ad essere un'approssimazione almeno del primo ordine di $\boldsymbol{y}_n$. Da notare che la condizione

$$\sum_{j=1}^{\nu} b_j = 1$$

è necessaria per avere almeno ordine 1 (cioè la consistenza). Per quanto riguarda la stabilità, si può ripetere tutto il ragionamento fatto per il caso dei metodi multistep: si arriva ad osservare che il polinomio caratteristico è $\rho(w) = w - 1$ e pertanto tutti i metodi di Runge–Kutta espliciti sono stabili. Ne discende la convergenza. Per quanto riguarda il massimo ordine che si può raggiungere dato il numero di stadi ν, si ha quanto riportato in Tabella 19.3.

numero stadi ν	1	2	3	4	5	6	7	8
massimo ordine p	1	2	3	4	4	5	6	6

Tabella 19.3: Massimo ordine dei metodi di Runge–Kutta espliciti dato il numero di stadi.

Il numero di stadi equivale al numero di valutazioni della funzione $\boldsymbol{f}$ (e dunque al costo del metodo).

È possibile generalizzare i metodi espliciti di Runge–Kutta per ottenere metodi *semiimpliciti* e impliciti, i cui tableaux sono riportati in Tabella 19.4. Per tali metodi, l'ordine massimo raggiungibile dato il numero di stadi ν è $p = 2\nu$. Anche per essi valgolo le condizioni

$$\sum_{j=1}^{\nu} a_{i,j} = c_i, \quad 1 \leq i \leq \nu$$

e

$$\sum_{j=1}^{\nu} b_j = 1$$

$$
\begin{array}{c|ccccccc}
c_1 & a_{1,1} & & & & & & \\
c_2 & a_{2,1} & a_{2,2} & & & & & \\
c_3 & a_{3,1} & a_{3,2} & a_{3,3} & & & & \\
\vdots & \vdots & \vdots & \ddots & \ddots & & & \\
c_{\nu-1} & a_{\nu-1,1} & a_{\nu-1,2} & \dots & a_{\nu-1,\nu-2} & a_{\nu-1,\nu-1} & & \\
c_\nu & a_{\nu,1} & a_{\nu,2} & \dots & a_{\nu,\nu-2} & a_{\nu,\nu-1} & a_{\nu,\nu} \\
\hline
 & b_1 & b_2 & \dots & b_{\nu-2} & b_{\nu-1} & b_\nu
\end{array}
$$

$$
\begin{array}{c|cccccc}
c_1 & a_{1,1} & a_{1,2} & \dots & \dots & a_{1,\nu-1} & a_{1,\nu} \\
c_2 & a_{2,1} & a_{2,2} & \dots & \dots & a_{2,\nu-1} & a_{2,\nu} \\
c_3 & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\
c_{\nu-1} & a_{\nu-1,1} & a_{\nu-1,2} & \dots & \dots & a_{\nu-1,\nu-1} & a_{\nu-1,\nu} \\
c_\nu & a_{\nu,1} & a_{\nu,2} & \dots & \dots & a_{\nu,\nu-1} & a_{\nu,\nu} \\
\hline
 & b_1 & b_2 & \dots & \dots & b_{\nu-1} & b_\nu
\end{array}
$$

Tabella 19.4: Tableaux di Butcher per i metodi di Runge–Kutta semiimpliciti (sopra) e impliciti (sotto).

Lo schema generale di un metodo di Runge–Kutta si scrive dunque

$$
\begin{cases}
\boldsymbol{\xi}_i = \boldsymbol{y}_n + k \displaystyle\sum_{j=1}^{j^*} a_{i,j} \boldsymbol{f}(t_n + k c_j, \boldsymbol{\xi}_j), \quad i = 1, \dots, \nu \\
\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + k \displaystyle\sum_{j=1}^{\nu} b_j \boldsymbol{f}(t_n + k c_j, \boldsymbol{\xi}_j)
\end{cases}
\tag{19.5}
$$

ove $j^* = i - 1$ per gli schemi espliciti, $j^* = i$ per gli schemi semiimpliciti e $j^* = \nu$ per gli schemi impliciti. Il guess iniziale per il calcolo di $\boldsymbol{\xi}_i$ è generalmente $\boldsymbol{y}_n$.

È buona norma, dopo aver definito la matrice `A` e i vettori `c` e `b` del tableau, controllare che queste condizioni siano soddisfatte, confrontando `sum(A,2)` con `c` e `sum(b)` con 1, al fine di evitare banali errori. Ovviamente, si tratta di condizioni necessarie ma *non sufficienti* per garantire la corretta implementazione. Ai fini dell'implementazione dei metodi di Runge–Kutta, per evitare di calcolare più volte la funzione $\boldsymbol{f}$ negli stessi punti, si può usare

lo schema

$$\begin{cases} \boldsymbol{f}_i = \boldsymbol{f}\left(t_n + kc_i, \boldsymbol{y}_n + k\sum_{j=1}^{j^*} a_{i,j}\boldsymbol{f}_j\right), \quad i = 1, \dots, \nu \\ \boldsymbol{y}_{n+1} = \boldsymbol{y}_n + k\sum_{j=1}^{\nu} b_j \boldsymbol{f}_j \end{cases}$$

Si ricava da (19.5) ponendo $\boldsymbol{f}_j = \boldsymbol{f}(t_n + c_j k, \boldsymbol{\xi}_j)$ e osservando che $\boldsymbol{f}_i = \boldsymbol{f}(t_n + c_i k, \boldsymbol{\xi}_i) = \boldsymbol{f}(t_n + c_i k, \boldsymbol{y}_n + k\sum_{j=1}^{j^*} a_{i,j}\boldsymbol{f}_j)$, $i = 1, \dots, \nu$. In questo caso, il guess iniziale per il calcolo di $\boldsymbol{f}_i$ è $\boldsymbol{f}(t_n, \boldsymbol{y}_n)$.

19.2 Due esempi di metodi di Runge–Kutta semiimpliciti

Consideriamo il seguente metodo di Runge–Kutta (semi)implicito

$$\boldsymbol{\xi}_1 = \boldsymbol{y}_n + \frac{k}{2}\boldsymbol{f}\left(t_n + \frac{k}{2}, \boldsymbol{\xi}_1\right) \tag{19.6a}$$

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + k\boldsymbol{f}\left(t_n + \frac{k}{2}, \boldsymbol{\xi}_1\right) \tag{19.6b}$$

di tableau

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

Esso equivale al metodo *punto medio implicito*

$$\bar{\boldsymbol{y}}_{n+1} = \boldsymbol{y}_n + k\boldsymbol{f}\left(t_n + \frac{k}{2}, \frac{\boldsymbol{y}_n + \bar{\boldsymbol{y}}_{n+1}}{2}\right) \tag{19.7}$$

infatti, per quest'ultimo vale

$$\frac{\boldsymbol{y}_n + \bar{\boldsymbol{y}}_{n+1}}{2} = \boldsymbol{y}_n + \frac{k}{2}\boldsymbol{f}\left(t_n + \frac{k}{2}, \frac{\boldsymbol{y}_n + \bar{\boldsymbol{y}}_{n+1}}{2}\right)$$

da cui si deduce che $(\boldsymbol{y}_n + \bar{\boldsymbol{y}}_{n+1})/2 = \boldsymbol{\xi}_1$ e dunque (19.6b) coincide con (19.7). È un metodo di ordine 2 (lo si dimostra a partire da (19.7) con passaggi molto simili a quelli fatti per i metodi di Runge–Kutta espliciti di ordine 2) e in qualche modo simile al metodo dei trapezi. Gode della seguente importante proprietà: se per

$$\boldsymbol{y}'(t) = \boldsymbol{f}(t, \boldsymbol{y}(t))$$

$\boldsymbol{y}(t)^{\mathrm{T}}\boldsymbol{y}(t)$ è costante per ogni t_0 e $\boldsymbol{y}_0$ allora per (19.7) vale $\boldsymbol{y}_{n+1}^{\mathrm{T}}\boldsymbol{y}_{n+1} = \boldsymbol{y}_n^{\mathrm{T}}\boldsymbol{y}_n$. Prima di dimostrarlo, osserviamo che la condizione sopra significa $\boldsymbol{y}(t)^{\mathrm{T}}\boldsymbol{f}(t,\boldsymbol{y}(t)) = 0 = \boldsymbol{y}_0^{\mathrm{T}}\boldsymbol{f}(t_0,\boldsymbol{y}_0)$ cioè $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{f}(t,\boldsymbol{x}) = 0$ per ogni t e ogni $\boldsymbol{x}$. Osserviamo infine che la proprietà è interessante quando $\boldsymbol{y}(t)$ è un vettore di dimensione maggiore di uno, altrimenti deve essere $\boldsymbol{f} = 0$ e dunque banalmente $\boldsymbol{y}_{n+1} = \boldsymbol{y}_n$. Osserviamo poi che

$$\boldsymbol{y}_{n+1} - \boldsymbol{y}_n = k\boldsymbol{f}\left(t_n + \frac{k}{2}, \frac{\boldsymbol{y}_n + \boldsymbol{y}_{n+1}}{2}\right)$$

e quindi, posto $\boldsymbol{x} = (\boldsymbol{y}_n + \boldsymbol{y}_{n+1})/2$,

$$0 = \left(\frac{\boldsymbol{y}_n + \boldsymbol{y}_{n+1}}{2}\right)^{\mathrm{T}} k\boldsymbol{f}\left(t_n + \frac{k}{2}, \frac{\boldsymbol{y}_n + \boldsymbol{y}_{n+1}}{2}\right) = \left(\frac{\boldsymbol{y}_n + \boldsymbol{y}_{n+1}}{2}\right)^{\mathrm{T}} (\boldsymbol{y}_{n+1} - \boldsymbol{y}_n)$$

da cui la tesi.

symplectic.m

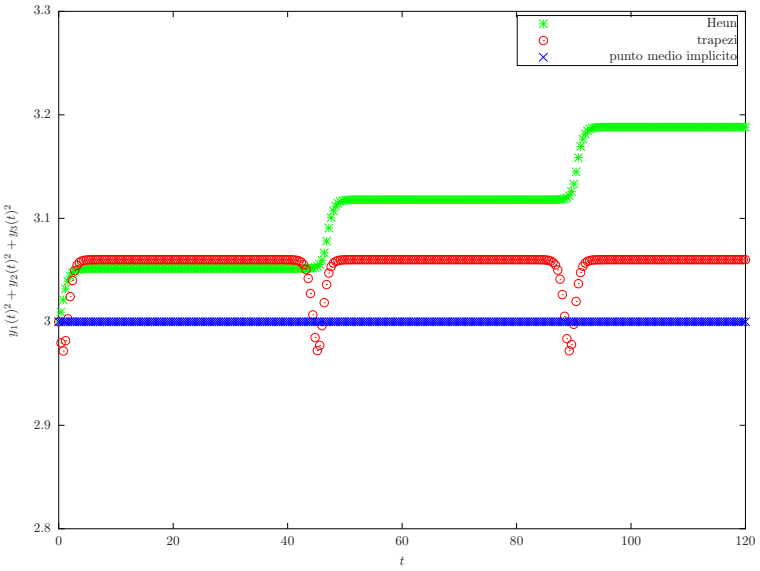


Figura 19.1: Evoluzione di $y_1(t)^2 + y_2(t)^2 + y_3(t)^2$ per il problema (19.8).

Consideriamo il seguente problema (equazioni di Eulero per il corpo rigido libero):

$$\begin{cases} y_1'(t) = \dfrac{I_2 - I_3}{I_2 I_3} y_2(t) y_3(t) \\ y_2'(t) = \dfrac{I_3 - I_1}{I_3 I_1} y_3(t) y_1(t) \\ y_3'(t) = \dfrac{I_1 - I_2}{I_1 I_2} y_1(t) y_2(t) \end{cases} \tag{19.8}$$

$\boldsymbol{y}$ rappresenta il momento angolare e I_1, I_2 e I_3 sono i momenti principali d'inerzia. Evidentemente $|\boldsymbol{y}(t)|^2 = y_1(t)^2 + y_2(t)^2 + y_3(t)^2$ è costante. Proviamo a risolvere il sistema con i metodi punto medio implicito, trapezi e Heun (tutti di ordine 2): l'integrazione fino al tempo $t^* = 120$ con 300 passi temporali ($I_1 = 2$, $I_2 = 1$, $I_3 = 2/3$, $\boldsymbol{y}(0) = [1, 1, 1]^{\mathrm{T}}$) produce il grafico in Figura 19.1, in cui si vede che il metodo punto medio implicito conserva "esattamente" la quantità $|\boldsymbol{y}(t)|^2$, a differenza dei metodi dei trapezi e di Heun. Ovviamente bisogna tenere in conto che gli errori di approssimazione (risoluzione del sistema non lineare) non garantiscono l'esatta uguaglianza $\boldsymbol{y}_{n+1}^{\mathrm{T}}\boldsymbol{y}_{n+1} = \boldsymbol{y}_n^{\mathrm{T}}\boldsymbol{y}_n$ anche per il metodo punto medio implicito. I tre metodi testati sono dello stesso ordine, ma uno produce soluzioni "qualitativamente" migliori.

Un metodo di Runge–Kutta di ordine tre è quello di tableau

$$\begin{array}{c|cc} \frac{3+\sqrt{3}}{6} & \frac{3+\sqrt{3}}{6} & 0 \\ \frac{3-\sqrt{3}}{6} & -\frac{\sqrt{3}}{3} & \frac{3+\sqrt{3}}{6} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \tag{19.9}$$

È interessante perché $\boldsymbol{\xi}_1$ e $\boldsymbol{\xi}_2$ soddisfano

$$\begin{aligned} \boldsymbol{\xi}_1 &= \boldsymbol{y}_n + a_{1,1}k\boldsymbol{f}(t_n + c_1 k, \boldsymbol{\xi}_1) \\ \boldsymbol{\xi}_2 &= \boldsymbol{y}_n + a_{2,1}k\boldsymbol{f}(t_n + c_1 k, \boldsymbol{\xi}_1) + a_{2,2}k\boldsymbol{f}(t_n + c_2 k, \boldsymbol{\xi}_2) \end{aligned}$$

con $a_{1,1} = a_{2,2}$. Dunque, gli jacobiani dei due sistemi da risolvere sono

$$\begin{aligned} J_n^1(\boldsymbol{x}) &= \left(I - ka_{1,1}\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}}(t_n + c_1 k, \boldsymbol{x})\right) \\ J_n^2(\boldsymbol{x}) &= \left(I - ka_{1,1}\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}}(t_n + c_2 k, \boldsymbol{x})\right) \end{aligned}$$

Spesso si usa l'approssimazione

$$J_n^1(\boldsymbol{x}) \approx J_n^2(\boldsymbol{x}) \approx J_n(\boldsymbol{y}_n) = \left(I - ka_{1,1}\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}}(t_n + (c_1 + c_2)k/2, \boldsymbol{y}_n)\right)$$

e quindi si risolvono entrambi i sistemi non lineari con un metodo di Newton modificato con jacobiano costante e fattorizzato una volta e per tutte. I metodi di Runge–Kutta semiimpliciti in cui il tableau ha diagonale costante si dicono *singolarmente (semi)impliciti (singly diagonally implicit)*.

Anche il θ-metodo può essere fatto rientrare nella classe dei metodi di Runge–Kutta semiimpliciti:

$$\begin{cases} \boldsymbol{\xi}_1 = \boldsymbol{y}_n \\ \boldsymbol{\xi}_2 = \boldsymbol{y}_n + k(1-\theta)\boldsymbol{f}(t_n, \boldsymbol{\xi}_1) + k\theta\boldsymbol{f}(t_n + k, \boldsymbol{\xi}_2) \\ \boldsymbol{y}_{n+1} = \boldsymbol{y}_n + k(1-\theta)\boldsymbol{f}(t_n, \boldsymbol{\xi}_1) + k\theta\boldsymbol{f}(t_n + k, \boldsymbol{\xi}_2) \end{cases}$$

o, in forma *implementativa* (anche se non si usa in pratica),

$$\begin{cases} \boldsymbol{f}_1 = \boldsymbol{f}(t_n, \boldsymbol{y}_n) \\ \boldsymbol{f}_2 = \boldsymbol{f}(t_n + k, \boldsymbol{y}_n + k(1-\theta)\boldsymbol{f}_1 + k\theta\boldsymbol{f}_2) \\ \boldsymbol{y}_{n+1} = \boldsymbol{y}_n + k(1-\theta)\boldsymbol{f}_1 + k\theta\boldsymbol{f}_2 \end{cases}$$

Dunque, abbiamo risposto anche alla prima domanda all'inizio di questo capitolo. Vediamo ora come rispondere alla terza domanda.

19.3 Metodi di Runge–Kutta *embedded*

Per i metodi ad un passo risulta alquanto facile adottare un passo temporale k_n variabile nel tempo (non così con i multistep, in cui i parametri dipendono dall'avere assunto i passi temporali costanti). In generale, più l'equazione ha un comportamento "lineare", più i passi possono essere presi grandi. Ma come adattare automaticamente il passo all'andamento della soluzione? Supponiamo di avere due metodi di Runge–Kutta espliciti di ordine $p-1$ e p rispettivamente, i cui tableaux sono riportati in Tabella 19.5. È chiaro che, dopo aver costruito il primo metodo, con una sola nuova valutazione della funzione $\boldsymbol{f}$ si può costruire il secondo metodo. Una tale coppia di metodi si dice *embedded* e si scrive di solito un unico tableau, come nella Tabella 19.6. Il fatto che per trovare metodi di Runge–Kutta sia necessario risolvere sistemi non lineari per i coefficienti, rende difficile *ma non impossibile* trovare coppie di metodi con tali caratteristiche.

Consideriamo il sistema differenziale

$$\begin{cases} \tilde{\boldsymbol{y}}'(t) = \boldsymbol{f}(t, \tilde{\boldsymbol{y}}(t)) \\ \tilde{\boldsymbol{y}}(t_n) = \boldsymbol{y}_n^{(p)} \end{cases}$$

ove $\boldsymbol{y}_n^{(p)}$ è l'approssimazione di $\boldsymbol{y}(t_n)$ ottenuta con il metodo di Runge–Kutta di ordine p. Si ha allora

$$\|\boldsymbol{y}_{n+1}^{(p)} - \boldsymbol{y}_{n+1}^{(p-1)}\| = \|\boldsymbol{y}_{n+1}^{(p)} - \tilde{\boldsymbol{y}}(t_{n+1}) + \tilde{\boldsymbol{y}}(t_{n+1}) - \boldsymbol{y}_{n+1}^{(p-1)}\| \leq C_{n+1}k_{n+1}^p, \quad (19.10)$$

$$
\begin{array}{c|ccccc}
0 & & & & & \\
c_2 & a_{2,1} & & & & \\
c_3 & a_{3,1} & a_{3,2} & & & \\
\vdots & \vdots & \vdots & \ddots & & \\
c_{\nu-1} & a_{\nu-1,1} & a_{\nu-1,2} & \dots & a_{\nu-1,\nu-2} & \\
\hline
 & b_1 & b_2 & \dots & b_{\nu-2} & b_{\nu-1}
\end{array}
$$

$$
\begin{array}{c|cccccc}
0 & & & & & & \\
c_2 & a_{2,1} & & & & & \\
c_3 & a_{3,1} & a_{3,2} & & & & \\
\vdots & \vdots & \vdots & \ddots & & & \\
c_{\nu-1} & a_{\nu-1,1} & a_{\nu-1,2} & \dots & a_{\nu-1,\nu-2} & & \\
c_\nu & a_{\nu,1} & a_{\nu,2} & \dots & a_{\nu,\nu-2} & a_{\nu,\nu-1} & \\
\hline
 & \hat{b}_1 & \hat{b}_2 & \dots & \hat{b}_{\nu-2} & \hat{b}_{\nu-1} & \hat{b}_\nu
\end{array}
$$

Tabella 19.5: Metodi di Runge–Kutta di ordine $p-1$ e p.

$$
\begin{array}{c|ccccc}
0 & & & & & \\
c_2 & a_{2,1} & & & & \\
c_3 & a_{3,1} & a_{32} & & & \\
\vdots & \vdots & \vdots & \ddots & & \\
c_\nu & a_{\nu,1} & a_{\nu,2} & \dots & a_{\nu,\nu-1} & \\
\hline
 & b_1 & b_2 & \dots & b_{\nu-1} & \\
\hline
 & \hat{b}_1 & \hat{b}_2 & \dots & \hat{b}_{\nu-1} & \hat{b}_\nu
\end{array}
$$

Tabella 19.6: Metodi di Runge–Kutta embedded di ordine $p-1$ e p.

per un opportuno $C_{n+1} > 0$, ove $k_{n+1} = t_{n+1} - t_n$ è il passo di integrazione e $\mathcal{O}(k_{n+1}^p)$ è l'errore locale del metodo di ordine $p-1$. Se si vuole controllare tale errore si può allora richiedere, ad ogni passo, che

$$\|\boldsymbol{y}_{n+1}^{(p)} - \boldsymbol{y}_{n+1}^{(p-1)}\| \leq \mathrm{tol_a} + \|\boldsymbol{y}_{n+1}^{(p-1)}\| \cdot \mathrm{tol_r} \tag{19.11}$$

Se la disuguaglianza non è soddisfatta, si rifiuta $\boldsymbol{y}_{n+1}^{(p)}$ (e $\boldsymbol{y}_{n+1}^{(p-1)}$) e si calcola un nuovo passo di integrazione $\tilde{k}_{n+1}$ minore di k_{n+1}. Per fare questo, si suppone che valga

$$\begin{aligned}
\|\boldsymbol{y}_{n+1,k_{n+1}}^{(p)} - \boldsymbol{y}_{n+1,k_{n+1}}^{(p-1)}\| &= C_{n+1} k_{n+1}^p \\
\|\boldsymbol{y}_{n+1,\tilde{k}_{n+1}}^{(p)} - \boldsymbol{y}_{n+1,\tilde{k}_{n+1}}^{(p-1)}\| &= \tilde{C}_{n+1} \tilde{k}_{n+1}^p = C_{n+1} \tilde{k}_{n+1}^p
\end{aligned}$$

(cioè $\tilde{C}_{n+1} = C_{n+1}$) e si impone che l'errore $\tilde{C}_{n+1}\tilde{k}_{n+1}^p$ valga proprio quanto

la tolleranza richiesta, ricavando

$$\tilde{k}_{n+1} = \left(\frac{\mathrm{tol_a} + \|\boldsymbol{y}^{(p-1)}_{n+1,k_{n+1}}\| \cdot \mathrm{tol_r}}{\tilde{C}_{n+1}} \right)^{1/p} = \left(\frac{\mathrm{tol_a} + \|\boldsymbol{y}^{(p-1)}_{n+1,k_{n+1}}\| \cdot \mathrm{tol_r}}{\|\boldsymbol{y}^{(p)}_{n+1,k_{n+1}} - \boldsymbol{y}^{(p-1)}_{n+1,k_{n+1}}\|} \right)^{1/p} \cdot k_{n+1}$$

Se invece la disuguaglianza (19.11) è soddisfatta, si accetta il passo corrente e per il successivo si suppone che valga

$$\|\boldsymbol{y}^{(p)}_{n+1,k_{n+1}} - \boldsymbol{y}^{(p-1)}_{n+1,k_{n+1}}\| = C_{n+1} k^p_{n+1}$$
$$\|\boldsymbol{y}^{(p)}_{n+2,k_{n+2}} - \boldsymbol{y}^{(p-1)}_{n+2,k_{n+2}}\| = C_{n+2} k^p_{n+2} = C_{n+1} k^p_{n+2}$$

(cioè $C_{n+2} = C_{n+1}$) e, di nuovo, si impone che l'errore successivo $C_{n+2} k^p_{n+2}$ valga quanto la tolleranza richiesta, ricavando

$$k_{n+2} = \left(\frac{\mathrm{tol_a} + \|\boldsymbol{y}^{(p-1)}_{n+1,k_{n+1}}\| \cdot \mathrm{tol_r}}{C_{n+2}} \right)^{1/p} = \left(\frac{\mathrm{tol_a} + \|\boldsymbol{y}^{(p-1)}_{n+1,k_{n+1}}\| \cdot \mathrm{tol_r}}{\|\boldsymbol{y}^{(p)}_{n+1,k_{n+1}} - \boldsymbol{y}^{(p-1)}_{n+1,k_{n+1}}\|} \right)^{1/p} \cdot k_{n+1}$$

Evidentemente si è supposto anche che, se i passi non sono molto diversi tra loro, $\|\boldsymbol{y}^{(p-1)}_{n+1,k_{n+1}}\| = \|\boldsymbol{y}^{(p-1)}_{n+1,\tilde{k}_{n+1}}\| = \|\boldsymbol{y}^{(p-1)}_{n+2,k_{n+2}}\|$. L'espressione trovata è la stessa. Per evitare che il passo di integrazione cambi troppo bruscamente, si può adottare una correzione del tipo

$$\min\left(2, \max\left(0.6, 0.9 \cdot \left(\frac{\mathrm{tol_a} + \|\boldsymbol{y}^{(p-1)}_{n+1}\| \cdot \mathrm{tol_r}}{\|\boldsymbol{y}^{(p)}_{n+1} - \boldsymbol{y}^{(p-1)}_{n+1}\|} \right)^{1/p}\right)\right) \cdot k_{n+1}$$

Vediamo un esempio facile di costruzione di metodi di Runge–Kutta embedded. Innanzitutto, osserviamo che qualunque metodo di Runge–Kutta (in, particolare, quelli di ordine due) richiede la valutazione di $\boldsymbol{f}_1 = \boldsymbol{f}(t_n, \boldsymbol{y}_n)$ che è praticamente tutto ciò che serve per il metodo di Runge–Kutta di ordine uno, cioè il metodo di Eulero. Quindi, qualunque metodo di Runge–Kutta di ordine due di tableau

0		
c_2	$a_{2,1}$	
	$\hat{b}_1$	$\hat{b}_2$

può essere implementato a passo variabile secondo lo schema (da t_n a t_{n+1})

- $\boldsymbol{f}_1 = \boldsymbol{f}(t_n, \boldsymbol{y}_n)$
- $\boldsymbol{y}^{(1)}_{n+1} = \boldsymbol{y}^{(2)}_n + k_{n+1} \boldsymbol{f}_1$ (metodo di Eulero)

- $\boldsymbol{f}_2 = \boldsymbol{f}(t_n + c_2 k_{n+1}, \boldsymbol{y}_n + a_{2,1} k_{n+1} \boldsymbol{f}_1)$
- $\boldsymbol{e}_{n+1}^{(2-1)} = k_{n+1}[(\hat{b}_1 - 1)\boldsymbol{f}_1 + \hat{b}_2 \boldsymbol{f}_2]$
- IF $\|\boldsymbol{e}_{n+1}^{(2-1)}\| > \mathrm{tol_a} + \|\boldsymbol{y}_{n+1}^{(1)}\| \mathrm{tol_r}$

  $n = n - 1$ (time step rifiutato)

  ELSE

  $\boldsymbol{y}_{n+1}^{(2)} = \boldsymbol{y}_{n+1}^{(1)} + \boldsymbol{e}_{n+1}^{(2-1)}$

  END
- $k_{n+2} = \left[(\mathrm{tol_a} + \|\boldsymbol{y}_{n+1}^{(1)}\| \cdot \mathrm{tol_r})/\|\boldsymbol{e}_{n+1}^{(2-1)}\|\right]^{1/2} \cdot k_{n+1}$
- $n = n + 1$

0						
$\frac{1}{4}$	$\frac{1}{4}$					
$\frac{3}{8}$	$\frac{3}{32}$	$\frac{9}{32}$				
$\frac{12}{13}$	$\frac{1932}{2197}$	$-\frac{7200}{2197}$	$\frac{7296}{2197}$			
1	$\frac{439}{216}$	-8	$\frac{3680}{513}$	$-\frac{845}{4104}$		
$\frac{1}{2}$	$-\frac{8}{27}$	2	$-\frac{3544}{2565}$	$\frac{1859}{4104}$	$-\frac{11}{40}$	
	$\frac{25}{216}$	0	$\frac{1408}{2565}$	$\frac{2197}{4104}$	$-\frac{1}{5}$	
	$\frac{16}{135}$	0	$\frac{6656}{12825}$	$\frac{28561}{56430}$	$-\frac{9}{50}$	$\frac{2}{55}$

Tabella 19.7: Metodo di Runge–Kutta–Fehlberg.

Forse il più importante metodo di Runge–Kutta embedded è il Runge–Kutta–Fehlberg, di ordine (4)5, il cui tableau è riportato in Tabella 19.7.

Capitolo 20

A-stabilità

Purtroppo la consistenza e la stabilità di un metodo non sono sufficienti per avere un *buon* solutore di qualunque equazione differenziale ordinaria. Consideriamo infatti il seguente problema lineare

$$\begin{cases} y'(t) = \lambda y(t) & t > t_0 \\ y(t_0) = y_0 \end{cases} \tag{20.1}$$

La soluzione esatta $y(t) = \mathrm{e}^{\lambda(t-t_0)} y_0$ tende a zero per $t \to +\infty$ quando $\Re(\lambda) < 0$. Analizziamo il comportamento del metodo di Eulero per questo problema, supponendo di avere fissato il passo temporale k (e dunque $t_n = t_0 + nk$): si ha

$$y_{n+1} = y_n + k\lambda y_n = (1 + k\lambda) y_n$$

da cui

$$y_n = (1 + k\lambda)^n y_0$$

Si ha

$$\lim_{n\to\infty} y_n = 0 \Leftrightarrow |1 + k\lambda| < 1 \Leftrightarrow 1 + k^2\Re(\lambda)^2 + 2k\Re(\lambda) + k^2\Im(\lambda)^2 < 1$$

da cui

$$\lim_{n\to\infty} y_n = 0 \Leftrightarrow k < -\frac{2\Re(\lambda)}{|\lambda|^2} \tag{20.2}$$

Dunque, la soluzione numerica ottenuta con il metodo di Eulero ha lo stesso comportamento della soluzione analitica solo se il passo temporale è sufficientemente piccolo. Altrimenti, la soluzione può essere completamente diversa ($\lim_{n\to\infty} |y_n| = |y_0|$ o $\lim_{n\to\infty} y_n = \infty$). Nel caso di Eulero implicito, invece, si ha

$$y_n = \left(\frac{1}{1 - k\lambda}\right)^n y_0$$

da cui

$$\lim_{n\to\infty} y_n = 0 \Leftrightarrow |1 - k\lambda| > 1 \Leftrightarrow |1 - k\Re(\lambda) - ki\Im(\lambda)| > 1$$

disuguaglianza sempre soddisfatta, poiché $\Re(\lambda) < 0$. Anche per il metodo dei trapezi la soluzione numerica tende a 0 per $n \to \infty$. Ma non è vero, in generale, per qualunque metodo implicito. Analizziamo infatti il comportamento generale del θ-metodo per questo problema: si ha

$$y_{n+1} = y_n + (1-\theta)k\lambda y_n + \theta k\lambda y_{n+1}$$

da cui

$$y_n = \left[\frac{1 + (1-\theta)k\lambda}{1 - \theta k\lambda}\right]^n y_0$$

Si ha

$$\lim_{n\to\infty} y_n = 0 \Leftrightarrow \left|\frac{1 + (1-\theta)k\lambda}{1 - \theta k\lambda}\right| < 1 \Leftrightarrow |1 + (1-\theta)k\lambda| < |1 - \theta k\lambda| \Leftrightarrow$$
$$0 < (\theta^2 - (1-\theta)^2)k^2\Re(\lambda)^2 - (2\theta + 2(1-\theta))k\Re(\lambda) + (\theta^2 - (1-\theta)^2)k^2\Im(\lambda)^2$$

da cui

$$\lim_{n\to\infty} y_n = 0 \Leftrightarrow 0 < (2\theta - 1)k^2|\lambda|^2 - 2k\Re(\lambda)$$

Se $2\theta - 1 \geq 0$, certamente la disequazione è soddisfatta. Altrimenti,

$$\lim_{n\to\infty} y_n = 0 \Leftrightarrow k < \frac{2\Re(\lambda)}{(2\theta - 1)|\lambda|^2}, \quad (2\theta - 1 < 0) \tag{20.3}$$

Definizione 5. *Dato un metodo numerico* $\boldsymbol{y}_{n+1} = \boldsymbol{\mathcal{Y}}_n(\boldsymbol{f}, k, \boldsymbol{y}_0, \ldots, \boldsymbol{y}_n)$*, la* regione di assoluta stabilità *(o* linear stability domain*) è l'insieme dei numeri* $z = k\lambda$ *per cui la soluzione di* (20.1) *soddisfa* $\lim_{n\to\infty} \boldsymbol{y}_n = 0$.

Con riferimento al θ-metodo, la regione di assoluta stabilità del metodo di Eulero è $\{z \in \mathbb{C}\colon |1+z| < 1\}$, per Eulero implicito è $\{z \in \mathbb{C}\colon |1-z| > 1\}$ e per il metodo dei trapezi è $\{z \in \mathbb{C}\colon \Re(z) < 0\}$. Diremo che un metodo è *A-stabile* se la sua regione di assoluta stabilità contiene $\mathbb{C}^- = \{z \in \mathbb{C}\colon \Re(z) < 0\}$, cioè se riproduce correttamente il comportamento della soluzione analitica di (20.1) quando $\Re(\lambda) < 0$. Da notare che, indicato con $r(k\lambda)$ il termine (che dovrebbe essere in modulo minore di 1)

$$r(k\lambda) = \left[\frac{1 + (1-\theta)k\lambda}{1 - \theta k\lambda}\right]$$

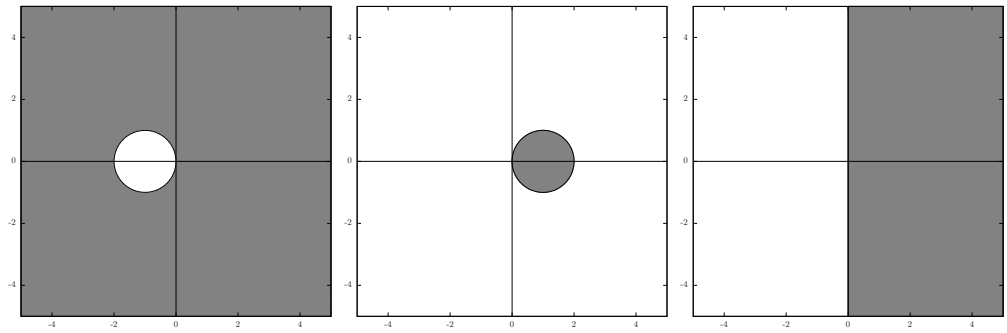


Figura 20.1: Regioni di assoluta stabilità (bianche) per i metodi di Eulero, Eulero implicito e trapezi.

si ha

$$\lim_{k\Re(\lambda)\to-\infty} |r(k\lambda)| = \left|\frac{\theta-1}{\theta}\right|$$

Tale limite vale proprio 1 per $\theta = 1/2$. Significa che se $\Re(\lambda) \ll 0$ oppure $k \gg 0$ il metodo dei trapezi potrebbe mostrare qualche problema di instabilità. In tal caso, il metodo migliore, da questo punto di vista, è il metodo di Eulero implicito ($\theta = 1$). In Figura 20.2 vediamo l'applicazione dei due metodi al problema

$$\begin{cases} y'(t) = -2000(y - \cos t), & t \leq 1.5 \\ y(0) = 0 \end{cases} \tag{20.4}$$

Se

$$\lim_{k\Re(\lambda)\to-\infty} |r(k\lambda)| = 0$$

diremo che il metodo è *L-stabile*.

Per inciso, se λ è puramente immaginario $\lambda = \delta$i, $|y_n| \to +\infty$ per Eulero, $|y_n| \to 0$ per Eulero implicito e $|y_n| = 1 = |y(t_n)|$ per il metodo dei trapezi.

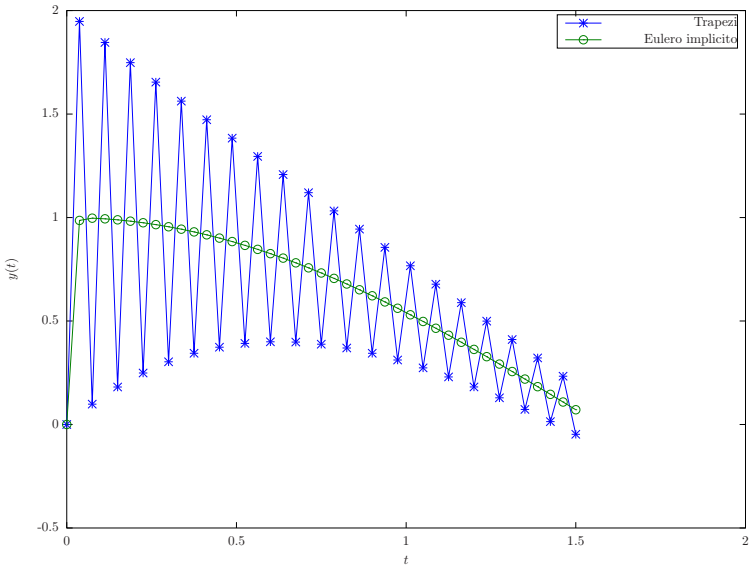
20.1 A-stabilità dei metodi di Runge–Kutta espliciti

Teorema 13. *Per un metodo di Runge–Kutta esplicito a ν stadi, si ha*

$$y_{n+1} = r(k\lambda)y_n \Rightarrow y_n = r(k\lambda)^n y_0$$

con $r(k\lambda)$ polinomio di grado ν in $z = k\lambda$. Inoltre, se l'ordine p è uguale al numero di stadi ν, si ha

$$r(z) = 1 + z + \frac{z^2}{2!} + \ldots + \frac{z^p}{p!}$$



Lstability.m

Figura 20.2: Metodi dei trapezi e di Eulero implicito per la soluzione di (20.4) con $k = 1.5/40$.

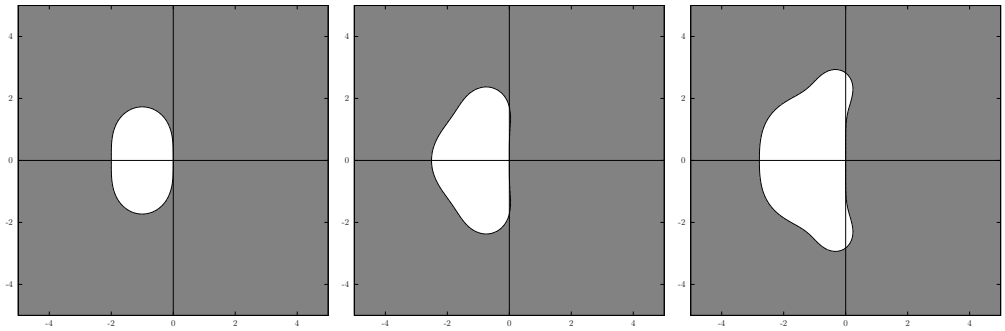


Figura 20.3: Regioni di assoluta stabilità (bianche) per i metodi di Runge–Kutta di ordine $p = 2$ ($\nu = 2$), $p = 3$ ($\nu = 3$) e $p = 4$ ($\nu = 4$).

Dimostrazione. Si ha che $\xi_1 = y_n$ è un polinomio di grado 0 in z. Supponiamo che ξ_j sia un polinomio $p_{j-1}(z)y_n$ di grado $j-1$ in $z = k\lambda$ per $j = 2, 3, \ldots, \nu-1$: allora

$$\xi_\nu = y_n + k\sum_{j=1}^{\nu-1} a_{\nu,j}\lambda\xi_j = y_n + k\lambda\sum_{j=1}^{\nu-1} a_{\nu,j}\xi_j = p_{\nu-1}(z)y_n$$

è un polinomio di grado $\nu - 1$ in z. Quindi

$$y_{n+1} = y_n + k\lambda\sum_{j=1}^{\nu} b_j\xi_j = r(k\lambda)y_n$$

e dunque $r(k\lambda)$ è un polinomio di grado ν in $z = k\lambda$. Poi, se l'ordine del metodo è p, significa che

$$y_1 - y(t_0 + k) = r(k\lambda)y_0 - y(t_0 + k) = \mathcal{O}(k^{p+1})$$

Ma $y(t_0 + k) = \mathrm{e}^{k\lambda}y_0$. Quindi $r(k\lambda) - \mathrm{e}^{k\lambda} = \mathcal{O}(k^{p+1})$ e dunque, l'unica possibilità per il polinomio di grado p $r(k\lambda)$ è

$$r(k\lambda) = r(z) = \left(1 + z + \frac{z^2}{2!} + \ldots + \frac{z^p}{p!}\right)$$

□

Dunque, i metodi di Runge–Kutta di ordine p uguale al numero di stadi ν hanno tutti la stessa regione di stabilità. In ogni caso, la dimostrazione qui sopra mostra che per un metodo di Runge–Kutta esplicito $r(z)$ è un polinomio di grado ν (e dunque di grado maggiore di 0).

Teorema 14. *Nessun metodo di Runge–Kutta esplicito è A-stabile.*

Dimostrazione. Si ha

$$\lim_{n\to\infty} y_n = 0 \Leftrightarrow |r(z)| < 1, \quad z = k\lambda$$

ma $r(z)$ è un polinomio di grado maggiore di 0. Dunque, $\lim_{x\to-\infty} r(x) = \infty$, x reale. Quindi, certamente esiste $z \in \mathbb{C}^- \cap \mathbb{R}$ tale che $|r(z)| > 1$ e dunque la regione di assoluta stabilità non contiene $\mathbb{C}^-$. □

Per quanto riguarda la regione di assoluta stabilità dei due metodi di Runge–Kutta implicito che conosciamo, cioè il metodo del punto medio implicito (19.7) e il metodo singolarmente semiimplicito (19.9), osserviamo che il primo coincide, per problemi lineari, con il metodo dei trapezi. Pertanto, ha la stessa regione di assoluta stabilità. Il secondo è anch'esso A-stabile (vedi Appendice A.6).

20.2 A-stabilità dei metodi lineari multistep

Ci limitiamo a riportare alcuni risultati.

Teorema 15. *Nessun metodo esplicito multistep è A-stabile.*

Teorema 16. *I metodi BDF ad un passo (Eulero implicito) e a due passi sono A-stabili.*

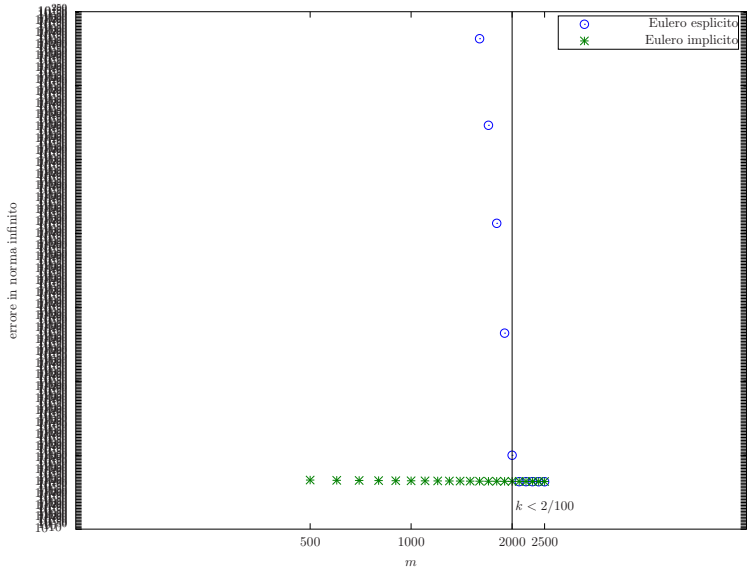
Teorema 17 (Seconda barriera di Dahlquist)**.** *L'ordine più alto che un metodo multistep A-stabile può raggiungere è due.*

20.3 Equazioni stiff

Se consideriamo il problema

$$\begin{cases} y'(t) = -100y(t), & t > 0 \\ y(0) = 1 \end{cases}$$

la condizione (20.3) per il metodo di Eulero impone $k < 1/50 = 0.02$. D'altra parte, la soluzione analitica del problema per $t^* = 0.4$ è minore di 10^{-17} (e dunque, trascurabile, nel senso che $y(0) - y(t^*) = y(0)$, in precisione doppia). Dunque, con poco più di 20 passi il metodo di Eulero arriva a calcolare adeguatamente la soluzione sino a t^*.



stiff.m

Figura 20.4: Eulero esplicito e Eulero implicito per la soluzione di (20.5) fino al tempo $t^* = 40$.

Qual è dunque il problema? Eccolo:

$$\begin{cases} \boldsymbol{y}'(t) = \begin{bmatrix} -100 & 0 \\ 0 & -1 \end{bmatrix} \boldsymbol{y}(t), & t > 0 \\ \boldsymbol{y}(0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{cases} \tag{20.5}$$

La soluzione analitica è

$$\boldsymbol{y}(t) = \begin{bmatrix} \mathrm{e}^{-100t} \\ \mathrm{e}^{-t} \end{bmatrix}$$

e la sua norma infinito è minore di 10^{-17} per $t^* = 40$. Poiché però per poter calcolare la prima componente serve un passo temporale $k < 0.02$, sono necessari più di 2000 passi (vedi Figura 20.4), anche se la prima componente diventa trascurabile dopo pochi passi e la seconda non richiederebbe un così elevato numero di passi. Dunque, anche se il metodo è convergente e il passo, per esempio, $k = 0.1$ garantisce un errore locale proporzionale a $k^2 = 0.01$, il metodo di Eulero non può essere usato con tale passo. Usando il metodo di Eulero implicito sarebbe possibile invece usare un passo piccolo all'inizio e poi, quando ormai la prima componente è trascurabile, si potrebbe incrementare il passo, senza pericolo di esplosione della soluzione. Per questo semplice problema, sarebbe possibile calcolare le due componenti separatamente. Nel caso generale, però, il sistema non è disaccoppiato. Per l'analisi, ci si può ricondurre, eventualmente in maniera approssimata, ad uno disaccoppiato e ragionare per componenti. Infatti, se A è una matrice diagonalizzabile,

$$\boldsymbol{y}'(t) = A\boldsymbol{y}(t) \Leftrightarrow \boldsymbol{z}'(t) = D\boldsymbol{z}(t) \Leftrightarrow \boldsymbol{z}(t) = \exp(tD)\boldsymbol{z}_0$$

ove $AV = VD$, $D = \operatorname{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_d\}$, e $\boldsymbol{y}(t) = V\boldsymbol{z}(t)$. Poi

$$\begin{aligned}\boldsymbol{y}'(t) = A\boldsymbol{y}(t) + \boldsymbol{b} \Leftrightarrow \boldsymbol{z}'(t) = D\boldsymbol{z}(t) + V^{-1}\boldsymbol{b} \Leftrightarrow \\ \Leftrightarrow \boldsymbol{z}(t) = \boldsymbol{z}_0 + t\varphi_1(tD)(D\boldsymbol{z}_0 + V^{-1}\boldsymbol{b})\end{aligned}$$

ove

$$\varphi_1(\lambda) = \begin{cases} \dfrac{\mathrm{e}^\lambda - 1}{\lambda} & \text{se } \lambda \neq 0 \\ 1 & \text{se } \lambda = 0 \end{cases}$$

Infine (considerando un problema autonomo per semplicità e sviluppando in serie di Taylor)

$$\boldsymbol{y}'(t) = \boldsymbol{f}(\boldsymbol{y}(t)) \Leftrightarrow \boldsymbol{y}'(t) \approx \boldsymbol{f}(\boldsymbol{y}_n) + J_n(\boldsymbol{y}_n)(\boldsymbol{y}(t) - \boldsymbol{y}_n)$$

ove J_n è la matrice jacobiana

$$J_n(\boldsymbol{y}_n) = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}}(\boldsymbol{y}_n)$$

e, se J_n è diagonalizzabile, ci si riconduce al caso precedente. Dunque, si ha sempre a che fare con gli autovalori di J_n (nel caso J_n non sia diagonalizzabile, si ragiona in maniera equivalente con blocchi di Jordan) e il più piccolo di questi è quello che determina la restrizione massima sul passo temporale.

Definizione 6. *Un sistema di ODEs* (16.1) *si dice* stiff *in un intorno di* t_n *se esiste almeno una coppia di autovalori* λ_1, λ_2 *della matrice jacobiana* J_n *tali che*

- $\Re(\lambda_1) < 0$, $\Re(\lambda_2) < 0$
- $\Re(\lambda_1) \ll \Re(\lambda_2)$

In pratica, può essere molto difficile capire se un sistema non lineare presenta regioni di *stiffness* o meno. Altrettanto difficile è rispondere alla domanda: per un problema stiff, conviene usare un metodo esplicito con passo piccolo o un metodo implicito? È chiaro che il metodo esplicito è di facile implementazione e applicazione, ma richiede molti passi temporali (e vedi 18.4). Il metodo implicito richiede la soluzione ad ognuno dei "pochi" passi di un sistema, in generale, non lineare.

20.3.1 Risoluzione di un metodo implicito per problemi stiff

Consideriamo, per semplicità, il problema

$$\boldsymbol{y}'(t) = A\boldsymbol{y}(t)$$

con A stiff e *simmetrica*. La restrizione sul passo per il metodo di Eulero esplicito è

$$k < \frac{2}{\rho_{\max}}$$

ove $\rho_{\max}$ è il raggio spettrale di A. Applicando il metodo di Eulero implicito e le iterazioni di punto fisso per risolvere l'equazione (per assurdo, poiché l'equazione da risolvere è lineare), siccome

$$\|A\boldsymbol{x} - A\boldsymbol{y}\|_2 \leq \|A\|_2\|\boldsymbol{x} - \boldsymbol{y}\|_2 = \rho_{\max}\|\boldsymbol{x} - \boldsymbol{y}\|_2$$

si avrebbe la restrizione (vedi (17.6))

$$k < \frac{1}{\rho_{\max}}$$

dunque una restrizione ancora più severa.

Da questo esempio si deduce che i metodi impliciti per problemi stiff vanno risolti con il metodo di Newton (eventualmente modificato).

Capitolo 21

Integratori esponenziali

I problemi di assoluta stabilità per semplici problemi lineari visti nel capitolo precedente, portano alla ricerca di nuovi metodi. Consideriamo il sistema differenziale

$$\begin{cases} \boldsymbol{y}'(t) = A\boldsymbol{y}(t) + \boldsymbol{b}, & t > 0 \\ \boldsymbol{y}(t_0) = \boldsymbol{y}_0 \end{cases}$$

La soluzione analitica è

$$\begin{aligned} \boldsymbol{y}(t) &= \exp((t-t_0)A)\boldsymbol{y}_0 + (t-t_0)\varphi_1((t-t_0)A)\boldsymbol{b} = \\ &= \boldsymbol{y}_0 + (t-t_0)\varphi_1((t-t_0)A)(A\boldsymbol{y}_0 + \boldsymbol{b}) \end{aligned}$$

Infatti $\boldsymbol{y}(t_0) = \boldsymbol{y}_0$ e, osservando che

$$\frac{\mathrm{d}}{\mathrm{d}t}[(t-t_0)\varphi_1((t-t_0)A)\boldsymbol{b}] = \exp((t-t_0)A)\boldsymbol{b} = (t-t_0)A\varphi_1((t-t_0)A)\boldsymbol{b} + \boldsymbol{b}$$

si ha

$$\begin{aligned} \boldsymbol{y}'(t) &= A\exp((t-t_0)A)\boldsymbol{y}_0 + \exp((t-t_0)A)\boldsymbol{b} = \\ &= A[\exp((t-t_0)A)\boldsymbol{y}_0 + (t-t_0)\varphi_1((t-t_0)A)\boldsymbol{b}] + \boldsymbol{b} \\ &= A\boldsymbol{y}(t) + \boldsymbol{b} \end{aligned}$$

Le funzioni exp e φ_1 di matrice possono essere approssimate come visto al paragrafo 8. Da questa osservazione, per un problema

$$\begin{cases} \boldsymbol{y}'(t) = A\boldsymbol{y}(t) + \boldsymbol{b}(t, \boldsymbol{y}(t)), & t > 0 \\ \boldsymbol{y}(t_0) = \boldsymbol{y}_0 \end{cases}$$

il metodo *Eulero esponenziale* è

$$\boldsymbol{y}_{n+1} = \exp(kA)\boldsymbol{y}_n + k\varphi_1(kA)\boldsymbol{b}(t_n, \boldsymbol{y}_n) = \boldsymbol{y}_n + k\varphi_1(kA)(A\boldsymbol{y}_n + \boldsymbol{b}(t_n, \boldsymbol{y}_n))$$

Proposizione 6. *Il metodo di Eulero esponenziale è* esatto *se* $\boldsymbol{b}(\boldsymbol{y}(t)) = \boldsymbol{b}(\boldsymbol{y}_0) \equiv \boldsymbol{b}$ *e di ordine uno altrimenti.*

Dimostrazione. Si ha

$$\boldsymbol{y}_{n+1} = \exp(kA)\boldsymbol{y}_n + \int_{t_n}^{t_{n+1}} \exp((t_{n+1} - \tau)A)\boldsymbol{b}(t_n, \boldsymbol{y}_n)\mathrm{d}\tau$$

Come al solito, inseriamo la soluzione analitica nello schema al posto della soluzione numerica. Ponendo $\boldsymbol{g}(t) = \boldsymbol{b}(t, \boldsymbol{y}(t))$ e usando la formula di variazioni delle costanti (8.2)

$$\begin{aligned}
&\boldsymbol{y}(t_{n+1}) - \exp(kA)\boldsymbol{y}(t_n) - \int_{t_n}^{t_{n+1}} \exp((t_{n+1} - \tau)A)\boldsymbol{g}(t_n)\mathrm{d}\tau = \\
&\quad = \exp(kA)\boldsymbol{y}(t_n) + \int_{t_n}^{t_{n+1}} \exp((t_{n+1} - \tau)A)\boldsymbol{g}(\tau)\mathrm{d}\tau + \\
&\quad - \exp(kA)\boldsymbol{y}(t_n) - \int_{t_n}^{t_{n+1}} \exp((t_{n+1} - \tau)A)\boldsymbol{g}(t_n)\mathrm{d}\tau = \\
&\quad = \int_{t_n}^{t_{n+1}} \exp((t_{n+1} - \tau)A)(\boldsymbol{g}(t_n) + \boldsymbol{g}'(\tau_n)(\tau - t_n) - \boldsymbol{g}(t_n))\mathrm{d}\tau = \\
&\quad = k^2\varphi_2(kA)\boldsymbol{g}'(\tau_n) = \mathcal{O}(k^2)
\end{aligned}$$

□

Si può inoltre dimostrare che il metodo converge (cioè è stabile). Poiché risolve esattamente i problemi lineari, il metodo è A-stabile e la sua regione di assoluta stabilità è

$$|r(z)| = |\mathrm{e}^z| < 1$$

e dunque $\mathbb{C}^-$.

Proposizione 7. *Per un problema* lineare, non autonomo

$$\begin{cases} \boldsymbol{y}'(t) = A\boldsymbol{y}(t) + \boldsymbol{b}(t), \quad t > 0 \\ \boldsymbol{y}(t_0) = \boldsymbol{y}_0 \end{cases}$$

il metodo esponenziale—punto medio

$$\boldsymbol{y}_{n+1} = \exp(kA)\boldsymbol{y}_n + k\varphi_1(kA)\boldsymbol{b}(t_n + k/2) = \boldsymbol{y}_n + k\varphi_1(kA)(A\boldsymbol{y}_n + \boldsymbol{b}(t_n + k/2))$$

è esatto se $\boldsymbol{b}(t) \equiv \boldsymbol{b}$ *e di ordine 2 altrimenti.*

Dimostrazione. Procedendo come sopra, si arriva a

$$\begin{aligned}
&\boldsymbol{y}(t_{n+1}) - \exp(kA)\boldsymbol{y}(t_n) - \int_{t_n}^{t_{n+1}} \exp((t_{n+1}-\tau)A)\boldsymbol{b}(t_n + k/2)\mathrm{d}\tau = \\
&\quad = \int_{t_n}^{t_{n+1}} \exp((t_{n+1}-\tau)A)\boldsymbol{b}'(\tau_n + k/2)(\tau - (t_n + k/2))\mathrm{d}\tau = \\
&\quad = \int_{t_n}^{t_{n+1}} \exp((t_{n+1}-\tau)A)\boldsymbol{b}'(\tau_n + k/2)(\tau - t_n - k/2)\mathrm{d}\tau = \\
&\quad = (k^2\varphi_2(kA) - k^2/2\varphi_1(kA))\boldsymbol{b}'(\tau_n + k/2) = \\
&\quad = \left(\frac{k^2 I}{2} + \frac{k^3 A}{6} + \mathcal{O}(k^4) - \frac{k^2 I}{2} - \frac{k^3 A}{2} + \mathcal{O}(k^4)\right)\boldsymbol{b}'(\tau_n + k/2) = \\
&\quad = \mathcal{O}(k^3)
\end{aligned}$$

□

Anche in questo caso si può dimostrare che il metodo converge e che è A-stabile. Dato un problema differenziale in forma autonoma

$$\begin{cases} \boldsymbol{y}'(t) = \boldsymbol{f}(\boldsymbol{y}(t)), & t > t_0 \\ \boldsymbol{y}(t_0) = \boldsymbol{y}_0 \end{cases}$$

si può pensare di linearizzarlo ad ogni passo

$$\boldsymbol{y}'(t) = J_n\boldsymbol{y}(t) + \boldsymbol{b}_n(\boldsymbol{y}(t))$$

ove

$$J_n = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}}(\boldsymbol{y}_n), \quad \boldsymbol{b}_n(\boldsymbol{y}(t)) = \boldsymbol{f}(\boldsymbol{y}(t)) - J_n\boldsymbol{y}(t)$$

e applicarvi il metodo di Eulero esponenziale. Si arriva così al metodo di Eulero–Rosenbrock esponenziale

$$\boldsymbol{y}_{n+1} = \exp(kJ_n)\boldsymbol{y}_n + k\varphi_1(kJ_n)\boldsymbol{b}_n(\boldsymbol{y}_n) = \boldsymbol{y}_n + k\varphi_1(kJ_n)\boldsymbol{f}(\boldsymbol{y}_n)$$

Il metodo è di ordine 2 e convergente. Esso richiede di valutare la funzione di matrice $\varphi_1(kJ_n)$ ad ogni passo temporale.

Gli integratori esponenziali sono particolarmente utili per la risoluzione di problemi stiff (essendo A-stabili). Conviene usare un metodo implicito o un metodo esponenziale? Nel primo caso, è necessario risolvere sistemi lineari, nel secondo calcolare funzioni di matrici. Per problemi di grosse dimensioni, non è per niente ovvio quale sia la strategia migliore. In generale, per matrici sparse senza struttura è più semplice calcolare funzioni di matrici, che non richiedono l'uso di precondizionatori efficaci.

Capitolo 22

Esercizi

1. Si consideri il seguente problema differenziale del secondo ordine *ai limiti*
$$\begin{cases} u''(x) - 3\cos(u(x)) = 0, \quad x \in (0,1) \\ u(0) = 0 \\ u(1) = 1 \end{cases}$$
Lo si trasformi in un sistema del primo ordine ($t = x$, $y_1(t) = u(x)$, $y_2(t) = u'(x)$) da risolvere con il metodo di Eulero esplicito e si determini, con una opportuna strategia, quale dovrebbe essere il valore iniziale $y_2(0)$ affinché $y_1(t) = u(x)$ sia soluzione del problema originale.

2. Con riferimento alla Figura 22.1, l'equazione del pendolo è
$$\begin{cases} l\vartheta''(t) = -g \sin\vartheta(t) \\ \vartheta(0) = \vartheta_0 \\ \vartheta'(0) = 0 \end{cases}$$
La si risolva con il metodo dei trapezi fino al tempo $t^* = \pi\sqrt{l/g}$ (assumendo $l = 1$, $\vartheta_0 = \pi/4$). Si confronti la traiettoria con quella del pendolo *linearizzato* ($\sin\vartheta(t) \approx \vartheta(t)$). Di quest'ultimo, si trovi il numero minimo di passi temporali affinché il metodo di Eulero esplicito produca una soluzione al tempo t^* che dista da $\vartheta(t^*)$ meno di 10^{-2}.

3. Si calcoli $\boldsymbol{y}(1)$, ove $\boldsymbol{y}'(t) = A\boldsymbol{y}(t)$, $\boldsymbol{y}(0) = [1, \ldots, 1]^{\mathrm{T}}$, con A data da `A = 100*toeplitz(sparse([1,1],[1,2],[-2,1],1,10))`, usando il θ-metodo con $\theta = 0, 1/2, 1$ e diversi passi temporali $k = 2^{-3}, 2^{-4}, \ldots, 2^{-8}$. Si confrontino i risultati con la *soluzione di riferimento* ottenuta usando $\theta = 1/2$ e $k = 2^{-10}$, mettendo in evidenza l'ordine del metodo usato. Si provi anche il valore $\theta = 1/3$, discutendo i risultati ottenuti.

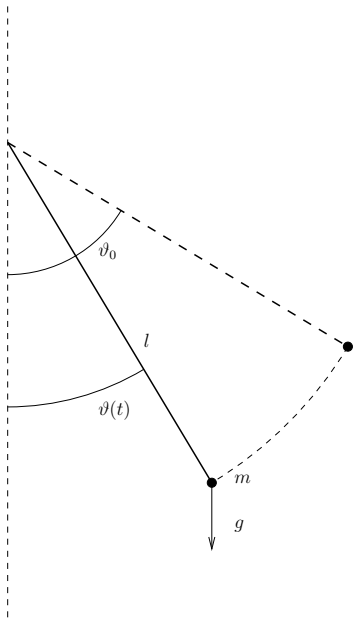


Figura 22.1: Pendolo

4. Si risolva il sistema di ODEs

$$\begin{cases} A'(t) = -2a(t)A(t) \\ a'(t) = A(t)^2 + \Omega(t)^2 - a(t)^2 - 1 \\ \Omega'(t) = -2(a(t) + A(t))\Omega(t) \end{cases} \tag{22.1}$$

con dato iniziale

$$\begin{cases} A(0) = 0.5 \\ a(0) = 2 \\ \Omega(0) = 10 \end{cases}$$

con il metodo di Eulero implicito fino ad un tempo finale $t^* = 15$, producendo un grafico della quantità $E(t) = (A(t)^2 + a(t)^2 + \Omega(t)^2 + 1)/(2A(t))$. Si confrontino le soluzioni ottenute usando 300 o 900 timesteps.

5. Si implementi il metodo di Eulero modificato (secondo tableau in Tabella 19.2) e lo si testi per il sistema differenziale (22.1), producendo il grafico della quantità $E(t)$.

6. Si implementino gli altri due metodi di ordine 2 in Tabella 19.2, li si testi per il sistema differenziale (22.1), mettendone in evidenza l'ordine.

0				
$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{1}{2}$	0	$\frac{1}{2}$		
1	0	0	1	
	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

Tabella 22.1: Metodo di Runge–Kutta a 4 stadi.

7. Si implementi il metodo di Runge–Kutta di tableau in Tabella 22.1, determinandone numericamente l'ordine.

8. Si implementi la function relativa ad un generico metodo di Runge–Kutta esplicito con tableau dato da

$\boldsymbol{c}$	A
	$\boldsymbol{b}^{\mathrm{T}}$

   ove $\boldsymbol{c}$, A e $\boldsymbol{b}$ sono dati.

9. Si implementi il metodo di Runge–Kutta (embedded) di tableau

0			
$\frac{1}{2}$	$\frac{1}{2}$		
1	-1	2	
	0	1	
	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

   e lo si applichi al problema differenziale 22.1.

10. Si implementi il metodo di Runge–Kutta–Fehlberg il cui tableau è riportato nella Tabella 19.7, e se ne mostri l'ordine. Lo si testi sul sistema differenziale (22.1).

11. Si consideri il problema differenziale

$$\begin{cases} \boldsymbol{y}'(t) = A\boldsymbol{y}(t) + \boldsymbol{y}(t)^2, & t > 0 \\ \boldsymbol{y}(0) = [\sqrt{(}2)/2, 1, \sqrt{(}2)/2]^{\mathrm{T}} \end{cases}$$

   ove

$$A = 16 \begin{bmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{bmatrix}$$

e lo si risolva fino al tempo $t^* = 1$ con i metodi di Eulero esplicito, Eulero implicito ed Eulero esponenziale, con diversi passi temporali.

Parte 3

PDEs (Equazioni alle derivate parziali)

Capitolo 23

Equazioni di trasporto-diffusione-reazione

Ci occuperemo in questo capitolo delle equazioni di trasporto-diffusione-reazione *(advection-diffusion-reaction, ADR)*.

23.1 Equazione del calore con dati iniziali e condizioni ai bordi

Consideriamo la seguente equazione alle derivate parziali

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \dfrac{\partial^2 u}{\partial x^2}(t,x), & t>0,\ x\in(0,L) \\ u(t,0)=u(t,L)=0, & t>0 \text{ } \textit{(condizioni ai bordi)} \\ u(0,x)=u_0(x), & x\in(0,L) \text{ } \textit{(condizioni iniziali)} \end{cases} \tag{23.1}$$

Supponiamo che $u_0(x)$ verifichi le *condizioni di compatibilità* $u_0(a)=u_0(b)=0$. Tale equazione rappresenta, per esempio, l'andamento della temperatura u su una barra di lunghezza L, i cui estremi sono tenuti a temperatura zero, e con una distribuzione iniziale di temperatura $u_0(x)$.

23.1.1 Esistenza di una soluzione

Cerchiamo una soluzione *a variabili separabili*

$$u(t,x)=\psi(t)\phi(x)$$

Inserendo tale rappresentazione in (23.1), si deduce

$$\psi'(t)\phi(x)=\psi(t)\phi''(x), \quad t>0,\ x\in(0,L)$$

da cui

$$\frac{\psi'(t)}{\psi(t)} = -K \text{ (costante)} \Rightarrow \psi(t) = A\mathrm{e}^{-Kt}$$

Per quanto riguarda $\phi(x)$, la soluzione generale è

$$\phi(x) = B\mathrm{e}^{\sqrt{-K}x} + C\mathrm{e}^{-\sqrt{-K}x}$$

Imponendo le condizioni al bordo

$$\begin{aligned} 0 &= \phi(0) = B + C \\ 0 &= \phi(L) = B\mathrm{e}^{\sqrt{-K}L} + C\mathrm{e}^{-\sqrt{-K}L} = B\left(\mathrm{e}^{\sqrt{-K}L} - \mathrm{e}^{-\sqrt{-K}L}\right) \end{aligned}$$

Se $K < 0$, allora $\mathrm{e}^{\sqrt{-K}L} - \mathrm{e}^{-\sqrt{-K}L} > 0$ e dunque $B = 0$ (e anche C). Quindi $\phi(x) = 0$, ma in tal caso $\psi(0)\phi(x) \neq u_0(x)$. Se invece $K = 0$, allora ancora $\phi(x) = B + C = 0$. Se invece $K = \lambda^2 > 0$, $\lambda > 0$, allora

$$\phi(x) = B\left(\mathrm{e}^{\mathrm{i}\lambda x} - \mathrm{e}^{-\mathrm{i}\lambda x}\right) = 2B\mathrm{i}\sin(\lambda x) = B\sin(\lambda x)$$

(avendo ridefinito B) e poiché $\phi(L) = 0$, l'unica possibilità non banale è $\lambda = j\pi/L$, j numero naturale non nullo. Pertanto, la funzione

$$u_j(t,x) = \exp\left(-\frac{j^2\pi^2}{L^2}t\right)\sin\left(\frac{j\pi}{L}x\right)$$

è soluzione dell'equazione del calore (e soddisfa le condizioni ai bordi) per ogni j. Quindi, la seguente serie

$$u(t,x) = \sum_{j=1}^{\infty} c_j u_j(t,x)$$

è soluzione *formale* dell'equazione del calore. Per quanto riguarda la condizione iniziale, si deve imporre

$$u_0(x) = u(0,x) = \sum_{j=1}^{\infty} c_j \sin\left(\frac{j\pi}{L}x\right) \tag{23.2}$$

Poiché $u_0(x)$ è nulla agli estremi, la possiamo prolungare per *antisimmetria* all'intervallo $[-L, L]$. Sotto opportune ipotesi, la sua serie di Fourier

$$\bar{u}_0(x) = \sum_{j=-\infty}^{+\infty} u_{0j}\phi_j(x)$$

converge in $[-L, L]$. Poiché $\bar{u}_0(x)$ è dispari, con riferimento al paragrafo 14.2.1,

$$u_{0m/2+1+j} = \int_{-L}^{L} \bar{u}_0(x)\overline{\phi_{m/2+1+j}(x)}\mathrm{d}x = \frac{-\mathrm{i}}{\sqrt{2L}} \int_{-L}^{L} \bar{u}_0(x) \sin\left(\frac{2\pi j(x+L)}{2L}\right) \mathrm{d}x =$$
$$= \frac{-\mathrm{i}\sqrt{2}}{\sqrt{L}} \int_0^L \bar{u}_0(x) \sin\left(\frac{j\pi x}{L} + j\pi\right) \mathrm{d}x =$$
$$= \frac{-\mathrm{i}\sqrt{2}}{\sqrt{L}} \int_0^L u_0(x) \sin\left(\frac{j\pi x}{L} + j\pi\right) \mathrm{d}x$$

e

$$u_{0m/2+1-j} = \int_{-L}^{L} \bar{u}_0(x)\overline{\phi_{m/2+1-j}(x)}\mathrm{d}x = \frac{-\mathrm{i}}{\sqrt{2L}} \int_{-L}^{L} \bar{u}_0(x) \sin\left(\frac{-2\pi j(x+L)}{2L}\right) \mathrm{d}x =$$
$$= \frac{\mathrm{i}\sqrt{2}}{\sqrt{L}} \int_0^L \bar{u}_0(x) \sin(j\pi x/L + j\pi)\mathrm{d}x = -u_{0m/2+1+j}$$

da cui

$$\bar{u}_0(x) = \sum_{j=-\infty}^{+\infty} u_{0m/2+1+j}\phi_{m/2+1+j}(x) =$$
$$= \sum_{j=-\infty}^{+\infty} u_{0m/2+1+j} \frac{\cos(j\pi x/L + j\pi) + \mathrm{i}\sin(j\pi x/L + j\pi)}{\sqrt{2L}} =$$
$$= \sum_{j=-\infty}^{-1} u_{0m/2+1+j} \frac{\cos(j\pi x/L + j\pi) + \mathrm{i}\sin(j\pi x/L + j\pi)}{\sqrt{2L}} +$$
$$+ \sum_{j=1}^{+\infty} u_{0m/2+1+j} \frac{\cos(j\pi x/L + j\pi) + \mathrm{i}\sin(j\pi x/L + j\pi)}{\sqrt{2L}} =$$
$$= \sum_{j=1}^{+\infty} -u_{0m/2+1+j} \frac{\cos(j\pi x/L + j\pi) - \mathrm{i}\sin(j\pi x/L + j\pi)}{\sqrt{2L}} +$$
$$+ \sum_{j=1}^{+\infty} u_{0m/2+1+j} \frac{\cos(j\pi x/L + j\pi) + \mathrm{i}\sin(j\pi x/L + j\pi)}{\sqrt{2L}} =$$
$$= \sum_{j=1}^{\infty} u_{0m/2+1+j} \frac{\sqrt{2}}{\sqrt{L}} \mathrm{i}\sin(j\pi x/L + j\pi) =$$
$$= \sum_{j=1}^{\infty} \left[\frac{2}{L}\int_0^L u_0(x) \sin\left(\frac{j\pi}{L}x\right) \mathrm{d}x\right] \sin\left(\frac{j\pi}{L}x\right)$$

Confrontanto quest'ultima espressione con (23.2), si deduce

$$c_j = \left[\frac{2}{L}\int_0^L u_0(x)\sin\left(\frac{j\pi}{L}x\right)\mathrm{d}x\right]$$

Si potrebbe mostrare adesso che

$$u(t,x) = \sum_{j=1}^{\infty} c_j \exp\left(-\frac{j^2\pi^2}{L^2}t\right)\sin\left(\frac{j\pi}{L}x\right)$$

è soluzione di (23.1) (bisogna poter derivare sotto il segno di serie). Dalla presenza del termine esponenziale negativo nel tempo per ogni componente $u_j(t,x)$, si deduce ogni componente tende a zero per $t \to +\infty$ (e dunque anche la soluzione), ma con *diverse* velocità dipendenti da un fattore proporzionale a j^2. L'equazione del calore rappresenta il modello dei fenomeni di *diffusione*. La diffusione è il processo mediante il quale la materia (o l'energia) è trasportata da una parte di un sistema ad un'altra come risultato di moti molecolari random.

23.1.2 Unicità della soluzione

Introduciamo la seguente quantità *(energia)*

$$E(t) = \int_0^L \frac{1}{2}u^2(t,x)\mathrm{d}x$$

Si ha

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \int_0^L \frac{\partial}{\partial t}\left[\frac{1}{2}u^2(t,x)\right]\mathrm{d}x = \int_0^L u\frac{\partial u}{\partial t}\mathrm{d}x = \int_0^L u\frac{\partial^2 u}{\partial x^2}\mathrm{d}x$$

Integrando per parti e tenendo conto delle condizioni ai bordi, si ha

$$\frac{\mathrm{d}E}{\mathrm{d}t} = -\int_0^L \left(\frac{\partial u}{\partial x}\right)^2 \mathrm{d}x \leq 0$$

Per dimostrare l'unicità, consideriamo come al solito il problema omogeneo (corrispondente a (23.1) con $u_0 \equiv 0$. Per tale problema $E_0(0) = 0$ e quindi $0 \leq E_0(t) \leq E_0(0)$ da cui $E_0(t) = 0$ per ogni t. Quindi $u(t,x) \equiv 0$ è l'unica soluzione del problema omogeneo. Dunque, se $u_1(t,x)$ e $u_2(t,x)$ fossero due soluzioni del problema (23.1), allora $u_1(t,x) - u_2(t,x)$ sarebbe soluzione del problema omogeneo e quindi $u_1(t,x) \equiv u_2(t,x)$.

Se $u_0(x) \geq 0$, si può dimostrare *(principio del massimo debole)* che la soluzione rimane non negativa per ogni t (dall'interpretazione fisica, è ovvio). Infatti, dato $\varepsilon > 0$, si ponga $v(t,x) = u(t,x) - \varepsilon x^2$. Allora $\partial_t v - \partial_{xx} v = 2\varepsilon > 0$. Se il minimo di $v(t,x)$ stesse in $(\bar{t}, \bar{x})$, $0 < \bar{t}$, $0 < \bar{x} < L$, allora $\partial_t v(\bar{t}, \bar{x}) = 0$ (punto critico) e $\partial_{xx} v(\bar{t}, \bar{x}) \geq 0$ (punto di minimo). Dunque

$$\partial_t v(\bar{t}, \bar{x}) - \partial_{xx} v(\bar{t}, \bar{x}) \leq 0$$

assurdo. Quindi, il punto di minimo per $v(t,x)$ sta in $\Gamma = \{0\} \times [0, L] \cup [0, +\infty) \times \{0, L\}$. Dunque

$$\min_\Gamma u - \varepsilon L^2 \leq \min_\Gamma v = \min v \leq \min u$$

e facendo tendere $\varepsilon \to 0$, si ottiene

$$\min_\Gamma u \leq \min u$$

Poiché ovviamente vale anche la disuguaglianza opposta,

$$\min u = \min_\Gamma u = \min\{\min u_0, 0\} = 0$$

23.2 Metodo di Fourier

Per quanto visto, il metodo spettrale basato su approssimazione in serie di Fourier (vedi paragrafo 14.2.1) dovrebbe essere particolarmente adatto alla risoluzione. Detta

$$\hat{u}(t,x) = \sum_{j=1}^{m} \hat{u}_j(t) \phi_j(x)$$

la soluzione approssimata, si ha

$$\sum_{j=1}^{m} \hat{u}_j'(t) \phi_j(x) = \sum_{j=1}^{m} \hat{u}_j(t) \lambda_j^2 \phi_j(x)$$

da cui, per l'ortonormalità della famiglia $\{\phi_j\}_j$,

$$\begin{cases} \hat{u}_k'(t) = \lambda_k^2 \hat{u}_k(t), & 1 \leq k \leq m \\ \hat{u}_k(0) = \hat{u_0}_k \end{cases}$$

ove $\lambda_k = \mathrm{i}(k - 1 - m/2)2\pi/(2L)$ (si deve lavorare infatti del dominio $[-L, L]$, dove si è prolungata per antisimmetria la funzione $u_0(x)$) e $\hat{u_0}_k$ sono i coefficienti di Fourier discreti di u_0 prolungata (si ha $\mathrm{i}\sqrt{2}\sqrt{L}\hat{u_0}_{m/2+1+k} = (-1)^k c_k$). Si trova, dunque,

$$\hat{u}_k(t) = \mathrm{e}^{-(k-1-m/2)^2 \pi^2 t / L^2} \hat{u_0}_k, \quad 1 \leq k \leq m$$

da cui poi si ricostruisce $\hat{u}(t,x)$. Avevamo visto che la decomposizione di Fourier si usa in caso di condizioni al bordo periodiche, mentre per l'equazione del calore sono di Dirichlet nulle. Poiché però il dato iniziale è la funzione dispari $\bar{u}_0(x)$, allora la soluzione $\bar{u}(t,x)$ dell'equazione del calore nell'intervallo $[-L,L]$ è pure dispari. Infatti, posto $\bar{v}(t,x) = -\bar{u}(t,-x)$, si ha

$$\begin{cases} \dfrac{\partial \bar{v}}{\partial t}(t,x) = -\dfrac{\partial \bar{u}}{\partial t}(t,-x) \\ \dfrac{\partial^2 \bar{v}}{\partial x^2}(t,x) = (-\cdot -)\cdot -\dfrac{\partial^2 \bar{u}}{\partial x^2}(t,-x) \end{cases}$$

inoltre, $\bar{v}(t,-L) = \bar{v}(t,L) = 0$ e $\bar{v}(0,x) = -\bar{u}(0,-x) = \bar{u}_0(x)$. Dunque, pure $\bar{v}(t,x)$ soddisfa l'equazione del calore. Ma questa è unica, quindi $\bar{v}(t,x) = -\bar{u}(t,-x) = \bar{u}(t,x)$, cioè $\bar{u}(t,x)$ è dispari. Quindi, $\partial_x \bar{u}(t,-x) = \partial_x \bar{u}(t,x)$ e, in particolare, $\partial_x \bar{u}(t,-L) = \partial_x \bar{u}(t,L)$. Per quanto visto, la serie di Fourier di $\bar{u}(t,x)$ converge (i coefficienti $u_j(t)$ decadono a zero almeno come j^2) e, poiché ogni troncata della serie è dispari e periodica, essa vale zero ai bordi $x = -L$ ed $x = L$ (e, di conseguenza, anche in $x = 0$: questo fatto non è vero per l'equazione originaria nel dominio $[0,L]$, poiché lì la soluzione non è dispari). Dunque, si può usare il metodo di Fourier. Se però $\bar{u}_0(x)$ non è periodica (nel senso che non lo sono le derivate di ordine superiore al primo), allora tale sarà la soluzione analitica e il metodo di Fourier *non* sarà spettralmente convergente.

23.3 Metodo delle linee

Il *metodo delle linee* per la risoluzione di problemi del tipo

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \dfrac{\partial^2 u}{\partial x^2}(t,x) + g(u(t,x)) + s(t,x), \quad t > 0,\ x \in (a,b) \\ + \textit{ condizioni ai bordi} \\ + \textit{ condizione iniziale} \end{cases} \tag{23.3}$$

ove il termine $g(u(t,x))$ si chiama *reazione* e il termine $s(t,x)$ *sorgente*, prevede di discretizzare gli operatori differenziali spaziali con uno dei metodi visti per i problemi con valori ai bordi e poi risolvere il sistema di ODEs che ne risulta con un metodo per problemi ai valori iniziali visti. Assumeremo sempre che la condizione iniziale soddisfi le condizioni ai bordi. Vediamo qualche esempio.

23.3.1 Differenze finite

Trascurando per il momento le condizioni ai bordi e usando differenze finite centrate del secondo ordine a passo costante h

$$\begin{bmatrix} y_1'(t) \\ y_2'(t) \\ \vdots \\ y_{m-1}'(t) \\ y_m'(t) \end{bmatrix} = \frac{1}{h^2} \begin{bmatrix} -2 & 1 & 0 & \dots & 0 \\ 1 & -2 & 1 & \dots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 1 & -2 & 1 \\ 0 & \dots & 0 & 1 & -2 \end{bmatrix} \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_{m-1}(t) \\ y_m(t) \end{bmatrix} + \begin{bmatrix} b_1(t, y_1(t)) \\ b_2(t, y_2(t)) \\ \vdots \\ b_{m-1}(t, y_{m-1}(t)) \\ b_m(t, y_m(t)) \end{bmatrix}$$

ove $y_j(t) \approx u(t, x_j)$ o, in maniera compatta,

$$\boldsymbol{y}'(t) = A\boldsymbol{y}(t) + \boldsymbol{b}(t, \boldsymbol{y}(t)) = \boldsymbol{f}(t, \boldsymbol{y}(t)) \tag{23.4}$$

(con l'ovvia definizione dei simboli). A questo punto, si sceglie il metodo di integrazione temporale (θ-metodo, multistep, Runge–Kutta, esponenziale). Si tenga presente che il problema (23.4), che si dice *semidiscretizzato*, è solitamente un problema stiff. Infatti, la matrice

$$\begin{bmatrix} -2 & 1 & 0 & \dots & 0 \\ 1 & -2 & 1 & \dots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 1 & -2 & 1 \\ 0 & \dots & 0 & 1 & -2 \end{bmatrix} \in \mathbb{R}^{m \times m}$$

ha autovalori reali negativi

$$\lambda_j = -4 \sin^2 \left(\frac{j}{2} \frac{\pi}{m+1} \right), \quad 1 \le j \le m$$

che vengono poi amplificati dal coefficiente $1/h^2$. Dunque, con riferimento alla condizione (20.2) per il metodo di Eulero, volendo usare questo metodo per l'integrazione temporale occorrerebbe un passo temporale k minore di (circa) $h^2/2$. Siccome il metodo di Eulero è del primo ordine, volendo che l'integrazione temporale non sia meno accurata dell'approssimazione spaziale, è giusto che il passo temporale sia proporzionale a h^2 (così che l'errore globale sia $\mathcal{O}(k) + \mathcal{O}(h^2) = \mathcal{O}(h^2)$). Per ridurre il numero di time steps, si può usare un metodo di ordine più alto, per esempio un metodo di Runge–Kutta esplicito di ordine 2. La restrizione sul time step è però la stessa (vedi la regione di assoluta stabilità del metodo in Figura 20.3) del metodo di Eulero. Dunque, ancora k dovrebbe essere proporzionale a $h^2/2$ (quindi il numero di time steps non diminuisce) e l'errore globale è ancora $\mathcal{O}(k^2) + \mathcal{O}(h^2) = \mathcal{O}(h^4) + \mathcal{O}(h^2) = \mathcal{O}(h^2)$.

23.3.2 Condizioni al bordo di Dirichlet

Vediamo come imporre una condizione di Dirichlet in $x_1 = a$ (eventualmente dipendente dal tempo $u(t, a) = y_1(t) = u_a(t)$) per il problema

$$\boldsymbol{y}'(t) = A\boldsymbol{y}(t) + \boldsymbol{b}(t, \boldsymbol{y}(t)) = \boldsymbol{f}(t, \boldsymbol{y}(t))$$

Costanti nel tempo

Si deve modificare la prima riga di $\boldsymbol{f}(t, \boldsymbol{y}(t))$ e porla uguale a zero. In tal modo, la prima riga del sistema differenziale risulta essere

$$y_1'(t) = 0 \Rightarrow y_1(t) = \text{costante} = y_1(t_0) = u_a$$

Poiché il dato iniziale soddisfa sempre le condizioni di compatibilità, la prima componente della soluzione assumerà sempre il valore u_a.

Variabili nel tempo

- **metodi espliciti:** basta calcolare $\boldsymbol{y}_{n+s}$ e poi modificarne la prima componente, ponendola uguale a $u_a(t_{n+s})$. Poiché però il problema è stiff, difficilmente i metodi espliciti sono efficaci, a causa della restrizione sul passo temporale.

- **metodi impliciti:** si deve trovare $\boldsymbol{x} = \boldsymbol{y}_{n+s}$ tale che

  $$F_{n+s-1}(\boldsymbol{x}) = 0$$

  Pertanto, si deve modificare la prima riga di questo sistema in modo che esprima l'uguaglianza $x_1 - u_a(t_{n+s}) = 0$. Per esempio, avendo scelto il metodo di Eulero implicito, si ha

  $$F_n(\boldsymbol{y}_{n+1}) = (I - kA)\boldsymbol{y}_{n+1} - k\boldsymbol{b}(t_{n+1}, \boldsymbol{y}_{n+1}) - \boldsymbol{y}_n = 0$$

  e l'imposizione della condizione al bordo avviene, per esempio, ponendo a zero la prima riga di A e la prima componente di $\boldsymbol{b}$ (ciò può essere fatto una volta per tutte, assieme alle necessarie modifiche allo Jacobiano) e ponendo uguale a $u_a(t_{n+1})$ la prima componente di $\boldsymbol{y}_n$.

- **metodi esponenziali:** per i metodi esponenziali visti si ha

  $$\boldsymbol{y}_{n+1} = \exp(kA)\boldsymbol{y}_n + k\varphi_1(kA)\boldsymbol{b}_n$$

  Se la prima riga di A viene messa a zero, la prima riga di $\exp(kA)$ e $\varphi_1(kA)$ è il primo vettore della base canonica e dunque basta porre il primo elemento di $\boldsymbol{b}_n$ uguale a $(u_a(t_{n+1}) - u_a(t_n))/k$.

23.3.3 Condizioni al bordo di Neumann (costanti)

Per quanto riguarda una condizione di Neumann omogenea, per esempio in $x = b$, si può pensare di introdurre la variabile fittizia $y_{m+1}(t) \approx u(t, x_{m+1})$, $x_{m+1} = b + h$ e imporre che $y_{m+1}(t) = y_{m-1}(t)$. L'approssimazione da usare per $\frac{\partial^2 u}{\partial x^2}(t, b)$ diventa dunque

$$\begin{aligned}\frac{\partial^2 u}{\partial x^2}(t, b) &\approx \frac{u(t, x_{m+1}) - 2u(t, x_m) + u(t, x_{m-1})}{h^2} = \\ &= \frac{y_{m+1}(t) - 2y_m(t) + y_{m-1}(t)}{h^2} = \frac{2y_{m-1}(t) - 2y_m(t)}{h^2}\end{aligned}$$

In maniera analoga si possono trattare condizioni di Neumann non omogenee (vedi paragrafo 11.4.1).

23.4 Equazione di trasporto-diffusione

Consideriamo l'equazione del *trasporto* (in un dominio non limitato)

$$\begin{cases} \dfrac{\partial u}{\partial t}(t, x) + c\dfrac{\partial u}{\partial x}(t, x) = 0, & t > 0,\ x \in \mathbb{R} \\ u(0, x) = u_0(x) \end{cases} \tag{23.5}$$

È facile verificare che la soluzione analitica è $u(t, x) = u_0(x - ct)$, da cui il nome dell'equazione. È ovviamente più *fisico* considerare un dominio limitato $x \in (a, b)$. Nel caso in cui $c > 0$, ha senso (ed è necessario) prescrivere un'unica condizione al bordo in $x = a$. Tale punto si chiama punto di *inflow* mentre il punto $x = b$ è detto di *outflow*. L'equazione di trasporto su un dominio limitato si scrive allora

$$\begin{cases} \dfrac{\partial u}{\partial t}(t, x) + c\dfrac{\partial u}{\partial x}(t, x) = 0, & t > 0,\ x \in (a, b),\ c > 0 \\ u(t, a) = 0, & t > 0 \\ u(0, x) = u_0(x), & x \in (a, b) \end{cases} \tag{23.6}$$

con $u_0(a) = 0$. La soluzione analitica è $u(t, x) = \tilde{u}_0(x - ct)$, ove

$$\tilde{u}_0(x) = \begin{cases} u_0(x) & x \in [a, b] \\ 0 & x < a \end{cases}$$

Nel caso in cui $c < 0$, il punto di inflow è $x = b$. Se consideriamo, più in generale, l'equazione di *trasporto-diffusione*

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) + c\dfrac{\partial u}{\partial x}(t,x) = d\dfrac{\partial^2 u}{\partial x^2} & t > 0,\ x \in (a,b) \\ + \text{ condizioni ai bordi} \\ + \text{ condizione iniziale} \end{cases} \tag{23.7}$$

ove $d > 0$, è lecito aspettarsi che entrambi i fenomeni di diffusione e trasporto si manifestino. Ancora, se $u_0(x) \geq 0$, tale rimane la soluzione per ogni t. Ma ciò è vero dopo aver discretizzato con il metodo delle linee? Abbiamo i due risultati seguenti.

Teorema 18. *Dato*

$$\begin{cases} \boldsymbol{y}'(t) = \boldsymbol{f}(t, \boldsymbol{y}(t)), & t > 0 \\ \boldsymbol{y}(0) = \boldsymbol{y}_0 \end{cases}$$

sono equivalenti le seguenti proprietà:

- *se* $\boldsymbol{y}_0 \geq 0$, *allora* $\boldsymbol{y}(t) \geq 0$ *per ogni* t *(il sistema si dice* positivo*)*
- *dato* $\boldsymbol{x}$, *con* $x_i = 0$, $x_j \geq 0$, $j \neq i$, *allora* $f_i(t, \boldsymbol{x}) \geq 0$

Da questo teorema segue, come corollario, il seguente, che può però essere dimostrato in maniera diretta.

Teorema 19. *Un sistema lineare* $\boldsymbol{y}'(t) = A\boldsymbol{y}(t)$ *è positivo se e solo se*

$$a_{ij} \geq 0 \text{ per ogni } j \neq i$$

ove $A = (a_{ij})$.

Dimostrazione. Supponiamo che il sistema sia positivo. Allora, se $\boldsymbol{y}_0 \geq 0$, si ha $\boldsymbol{y}(\tau) \geq 0$. Ma

$$\boldsymbol{y}(\tau) = \exp(\tau A)\boldsymbol{y}_0 = (I + \tau A)\boldsymbol{y}_0 + \mathcal{O}(\tau^2)$$

se τ è sufficientemente piccolo. Se, per assurdo, $a_{\bar{i}\bar{j}} < 0$, $\bar{j} \neq \bar{i}$, allora, preso $\boldsymbol{y}_0 = \boldsymbol{e}_{\bar{j}}$,

$$(I + \tau A)\boldsymbol{e}_{\bar{j}} = \begin{bmatrix} * \\ \vdots \\ * \\ \tau a_{\bar{i}\bar{j}} \\ * \\ \vdots \\ * \end{bmatrix} \leftarrow \text{riga } \bar{i}$$

e dunque la componente $\bar{i}$-esima di $\exp(\tau A)\boldsymbol{e}_{\bar{j}}$ sarebbe negativa, assurdo.

Se invece $a_{ij} \geq 0$, $j \neq i$, allora

$$\exp(tA) = \lim_{n\to\infty} \left(I + \frac{t}{n}A\right)^n \geq 0$$

da cui la positività. □

Tornando all'equazione (23.7), la discretizzazione mediante differenze finite centrate del secondo ordine porge, nei nodi interni,

$$y_i'(t) + c\frac{y_{i+1}(t) - y_{i-1}(t)}{2h} = d\frac{y_{i+1}(t) - 2y_i(t) + y_{i-1}(t)}{h^2}$$

I termini extradiagonali della matrice che ne deriva sono

$$\frac{c}{2h} + \frac{d}{h^2} \text{ e } -\frac{c}{2h} + \frac{d}{h^2}$$

che, per avere la positività, devono essere entrambi non negativi, da cui

$$\frac{|c|h}{2d} \leq 1$$

La quantità $\mathrm{Pe} = |c|h/(2d)$ si chiama *numero di Péclet di griglia*. La perdita di positività è solo uno degli effetti del numero di Péclet di griglia troppo elevato: si possono avere anche oscillazioni spurie, pertanto si chiederà sempre che il numero di Péclet di griglia sia minore o uguale a 1.

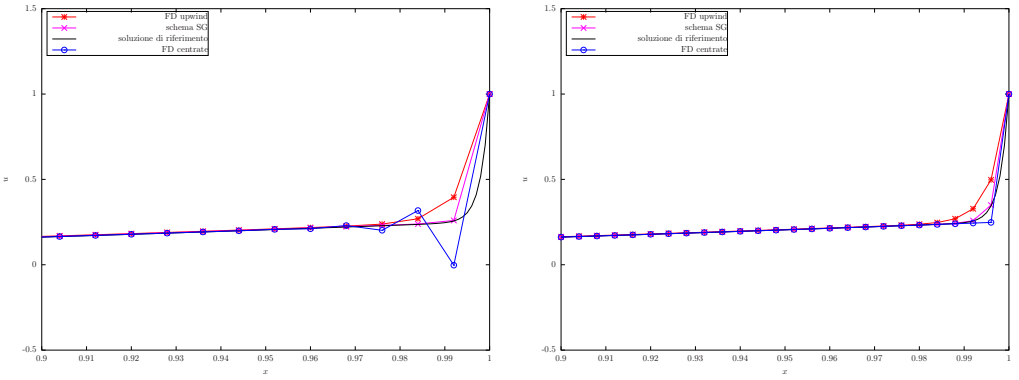
Da notare che la positività del sistema $\boldsymbol{y}'(t) = A\boldsymbol{y}(t)$ *non* garantisce che qualunque metodo numerico per ODEs la preservi (di certo lo garantisce il metodo esponenziale poiché esatto e, per il problema di diffusione e trasporto, anche Eulero implicito, vedi A.2). Pertanto, la condizione sul numero di Péclet di griglia è solo necessaria per avere una soluzione numerica positiva.

23.4.1 Stabilizzazione mediante diffusione artificiale

La restrizione sul passo di discretizzazione data dal numero di Péclet di griglia potrebbe essere irrealizzabile. Vediamo di stabilizzare lo schema delle differenze finite.

Consideriamo, per esempio, l'equazione di trasporto-diffusione

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) + c\dfrac{\partial u}{\partial x}(t,x) = d\dfrac{\partial^2 u}{\partial x^2} & t > 0,\ x \in (0,1) \\ u(0,x) = x^2 \\ u(t,0) = 0 \\ u(t,1) = 1 \end{cases} \tag{23.8}$$



peclet.m

Figura 23.1: Soluzione di (23.8) con diversi schemi di differenze finite, $h = 1/125$ (sinistra), $h = 1/250$ (destra) e Eulero esponenziale al tempo $t^* = 0.05$. La soluzione di riferimento è stata ottenuta con differenze finite centrate e $h = 1/1000$.

con $c = 10$ e $d = 0.02$. Il numero di Péclet di griglia, con $h = 1/125$, vale 2. La risoluzione mediante differenze finite centrate e Eulero esponenziale (esatto nel tempo) produce il grafico blu a sinistra in Figura 23.1. Se consideriamo invece la discretizzazione del primo ordine della derivata prima

$$\frac{\partial u}{\partial x}(t, x_i) \approx \frac{u_i - u_{i-1}}{h}$$

(in tale contesto si chiama discretizzazione *upwind*), otteniamo il grafico rosso, piuttosto lontano dalla soluzione esatta, ma privo di oscillazioni. Si può infatti vedere che i termini extradiagonali della matrice di discretizzazione sono non negativi. Per tentare di generalizzare (e migliorare l'ordine di accuratezza) questo approccio, scriviamo

$$\frac{u_i - u_{i-1}}{h} = \frac{u_{i+1} - u_{i-1}}{2h} - \frac{h}{2}\frac{u_{i-1} - 2u_i + u_{i+1}}{h^2}$$

e quindi approssimare al primo ordine

$$\begin{aligned} -c\frac{\partial u}{\partial x}(t, x_i) + d\frac{\partial^2 u}{\partial x^2}(t, x_i) &\approx -c\frac{u_i - u_{i-1}}{h} + d\frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} = \\ &= -c\frac{u_{i+1} - u_{i-1}}{2h} + d\left(1 + \frac{ch}{2d}\right)\frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} \end{aligned}$$

significa approssimare al secondo ordine (dunque, meglio)

$$-c\frac{\partial u}{\partial x}(t, x_i) + d\left(1 + \frac{ch}{2d}\right)\frac{\partial^2 u}{\partial x^2}(t, x_i)$$

cioè un'equazione con una diffusione *artificiale* (aggiuntiva) di coefficiente $ch/2$. Per questa equazione il numero di Péclet vale

$$\frac{ch}{2d\left(1+\frac{ch}{2d}\right)} = \frac{ch}{2d(1+\mathrm{Pe})} = \frac{\mathrm{Pe}}{1+\mathrm{Pe}} < 1, \quad \forall h$$

e ciò spiega l'assenza di oscillazioni. Vorremmo trovare una diffusione artificiale che stabilizzi lo schema e preservi l'ordine due delle differenze finite centrate. Cercheremo dunque una funzione ϕ del numero di Péclet e sostituiremo d con $d(1+\phi(\mathrm{Pe}))$ in modo che il nuovo numero di Péclet valga

$$\frac{ch}{2d(1+\phi(\mathrm{Pe}))}$$

La funzione $\phi(\mathrm{Pe})$ dovrà soddisfare:

- $\phi(\mathrm{Pe}) \gtrapprox \mathrm{Pe}-1$ (così il nuovo numero di Péclet sarà minore o uguale a 1), ma non troppo grande (altrimenti si introduce troppa diffusione artificiale)
- $\phi(ch/(2d)) = \mathcal{O}(h^2)$, $h \to 0$ (così la discretizzazione a differenze finite centrate sarà di ordine 2)

Una scelta possibile è

$$\phi(z) = z - 1 + \mathrm{e}^{-z}, \quad \phi(z) = \frac{z^2}{2} + \mathcal{O}(z^3), \quad z \to 0$$

Una scelta migliore è

$$\phi(z) = z - 1 + \frac{2z}{\mathrm{e}^{2z}-1}, \quad \phi(z) = \frac{z^2}{3} + \mathcal{O}(z^4), \quad z \to 0$$

(da notare che lo schema upwind corrisponde a $\phi(z) = z$). Il risultato corrisponde al grafico magenta in Figura 23.1 e lo schema si chiama di Scharfetter e Gummel.

L'esempio usato aveva il coefficiente c positivo: la funzione ϕ da usare deve essere funzione del numero di Péclet $\mathrm{Pe} = |c|h/(2d)$, in modo da *aggiungere* diffusione artificiale e non togliere. Per esempio, nel caso upwind con $c < 0$, $\phi(\mathrm{Pe}) = \mathrm{Pe} = -ch/(2d)$ e pertanto la discretizzazione al primo ordine che ne risulta è

$$\begin{aligned} -c\frac{\partial u}{\partial x}(t,x_i) + d\frac{\partial^2 u}{\partial x^2}(t,x_i) &\approx -c\frac{u_{i+1}-u_i}{h} + d\frac{u_{i-1}-2u_i+u_{i+1}}{h^2} = \\ &= -c\frac{u_{i+1}-u_{i-1}}{2h} + d\left(1-\frac{ch}{2d}\right)\frac{u_{i-1}-2u_i+u_{i+1}}{h^2} \end{aligned}$$

23.4.2 Elementi finiti

Nel caso di discretizzazione spaziale con elementi finiti lineari, la discretizzazione del problema (23.3) porta al sistema di ODEs

$$P\boldsymbol{y}'(t) = A\boldsymbol{y}(t) + \boldsymbol{g}(\boldsymbol{y}(t)) + \boldsymbol{s}(t) \tag{23.9}$$

ove A è (l'opposta de) la *stiffness matrix* e P la *mass matrix*, definita da,

$$\begin{aligned} p_{jj} &= \int_{x_{j-1}}^{x_{j+1}} \phi_j(x)\phi_j(x)\mathrm{d}x = \frac{h_{j-1}+h_j}{3} \\ p_{j\,j+1} = p_{j+1\,j} &= \int_{x_j}^{x_{j+1}} \phi_j(x)\phi_{j+1}(x)\mathrm{d}x = \frac{h_j}{6} \end{aligned} \tag{23.10a}$$

mentre, per $j = 1$ e $j = m$,

$$\begin{aligned} p_{11} &= \int_{x_1}^{x_2} \phi_1(x)\phi_1(x)\mathrm{d}x = \frac{h_1}{3} \\ p_{12} &= \int_{x_1}^{x_2} \phi_1(x)\phi_2(x)\mathrm{d}x = \frac{h_1}{6} \\ p_{m-1\,m} = p_{m\,m-1} &= \int_{x_{m-1}}^{x_m} \phi_m(x)\phi_{m-1}(x)\mathrm{d}x = \frac{h_{m-1}}{6} \\ p_{mm} &= \int_{x_{m-1}}^{x_m} \phi_m(x)\phi_m(x)\mathrm{d}x = \frac{h_{m-1}}{3} \end{aligned} \tag{23.10b}$$

Poi, per $1 < i < m$,

$$\begin{aligned} g_i &= \int_{x_{i-1}}^{x_{i+1}} g\left(\sum_{j=1}^{m} u_j\phi_j(x)\right)\phi_i(x)\mathrm{d}x = \\ &= \int_{x_{i-1}}^{x_i} g\left(\sum_{j=1}^{m} u_j\phi_j(x)\right)\phi_i(x)\mathrm{d}x + \int_{x_i}^{x_{i+1}} g\left(\sum_{j=1}^{m} u_j\phi_j(x)\right)\phi_i(x)\mathrm{d}x \approx \\ &\frac{g(y_{i-1})+g(y_i)}{2}\frac{h_{i-1}}{2} + \frac{g(y_i)+g(y_{i+1})}{2}\frac{h_i}{2} \\ s_i &= \int_{x_{i-1}}^{x_{i+1}} s(t,x)\phi_i(x)\mathrm{d}x \approx \frac{s(t,x_{i-1})+s(t,x_i)}{2}\frac{h_{i-1}}{2} + \frac{s(t,x_i)+s(t,x_{i+1})}{2}\frac{h_i}{2} \end{aligned}$$

mentre per $i = 1$ e $i = m$

$$\begin{aligned} g_1 &= \frac{g(y_1)+g(y_2)}{2}\frac{h_1}{2}, & g_m &= \frac{g(y_{m-1})+g(y_m)}{2}\frac{h_{m-1}}{2} \\ s_1 &= \frac{s(t,x_1)+s(t,x_2)}{2}\frac{h_1}{2}, & s_m &= \frac{s(t,x_{m-1})+s(t,x_m)}{2}\frac{h_{m-1}}{2} \end{aligned}$$

Usando un metodo esplicito per la risoluzione del sistema differenziale (23.9), è necessaria l'inversione della matrice di massa. Per tale motivo, si può ricorrere alla tecnica del *mass lumping* che consiste nel rendere diagonale la matrice P sostituendo ogni sua riga con una riga di zeri e la somma degli elementi originali in diagonale. Tale modifica è equivalente all'approssimazione degli integrali in (23.10) mediante la formula dei trapezi e dunque non riduce l'accuratezza del metodo. Infatti, la matrice $P_{\mathrm{L}}^{(-1)}A$ (P_{L} la matrice di massa con lumping) risulta uguale alla matrice che si ottiene discretizzando con differenze finite centrate del secondo ordine.

Usando invece un metodo implicito per la risoluzione del sistema differenziale (23.9), non è necessaria la tecnica del mass lumping: semplicemente, si devono risolvere sistemi lineari in cui la matrice identità è sostituita dalla matrice di massa.

23.4.3 Errori spaziali e temporali

Con il metodo delle linee, è facile capire cosa contribuisce all'errore spaziale e cosa all'errore temporale. Per esempio, se si usano differenze finite centrate di ordine due, si commette un errore Ch^2, $h \propto \frac{1}{m-1}$. Questo significa che, qualunque metodo si usi per l'integrazione temporale e con qualunque passo temporale, non è possibile scendere sotto tale errore. Ciò è esemplificato

erroretemporale.m

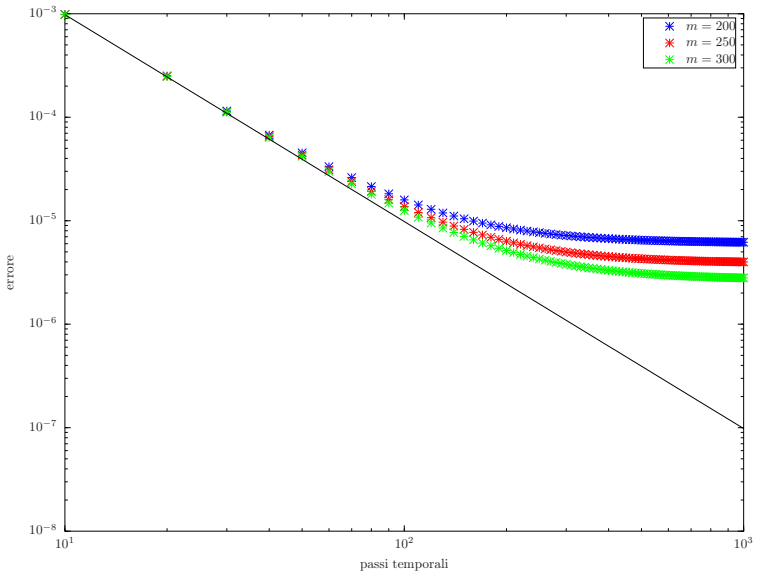


Figura 23.2: Errore temporale per un numero m di passi spaziali.

in Figura 23.2, ove si è risolto il problema dell'Esercizio 1 con un numero

m diverso di passi spaziali e si è misurato l'errore rispetto alla soluzione analitica. Vale ovviamente anche l'inverso: il metodo scelto per l'integrazione temporale e il numero di passi temporali pone un limite inferiore all'errore rispetto alla soluzione analitica.

23.5 Esercizi

1. Si calcoli la soluzione analitica dell'equazione del calore con sorgente

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \dfrac{\partial^2 u}{\partial x^2}(t,x) + 2\mathrm{e}^t \sin(x), & t > 0,\ x \in (0, \pi/2) \\ u(t,0) = 0, & t > 0 \\ \dfrac{\partial u}{\partial x}(t, \pi/2) = 0, & t > 0 \\ u(0,x) = \sin(x), & x \in (0, \pi/2) \end{cases}$$

   usando differenze finite del secondo ordine nello spazio e il metodo dei trapezi nel tempo. Si mostrino gli ordini spaziali e temporali della convergenza alla soluzione analitica al tempo $t^* = 1$.

2. Per l'esercizio sopra, discretizzato nello spazio tramite differenze finite centrate del secondo ordine con $m = 100$ nodi, si determini il numero minimo di passi temporali per avere un errore al tempo $t^* = 1$ rispetto alla soluzione analitica inferiore a 10^{-3}, avendo usato nel tempo

   - il metodo di Eulero
   - il metodo di Eulero implicito
   - il metodo dei trapezi
   - il metodo di Heun
   - il metodo Runge–Kutta di tableau in Tabella 22.1

3. Si ripeta l'esercizio 1. usando Eulero esponenziale e esponenziale—punto medio nel tempo.

4. Si studi l'andamento della soluzione del problema di *trasporto-diffusione-reazione*

$$\begin{cases} \dfrac{\partial u}{\partial t} + c\dfrac{\partial u}{\partial x} = d\dfrac{\partial^2 u}{\partial x^2}(t,x) + \rho u(u-1/2)(1-u), & t>0,\ x\in(0,1) \\ u(t,0) = 0, & t>0 \\ \dfrac{\partial u}{\partial x}(t,1) = 0, & t>0 \\ u(0,x) = 5x(1-x)^2, & x\in(0,1) \end{cases}$$

al variare dei coefficienti c, d e ρ (partendo da $c = 0.8$, $d = 0.01$, $\rho = 50$). Si usi un metodo implicito nel tempo. Si testi anche il caso di condizioni di Dirichlet omogenee per entrambi i bordi.

Parte 4

Appendici

Appendice A

Alcune dimostrazioni

A.1 M-matrici

Per dimostrare che la matrice A in (11.4) è una M-matrice, si procede così. Dato $\sigma = 1/\max_{2\le i\le m-1}\{2/h^2 + q_i\}$, definiamo

$$V = I - \sigma A$$

Innanzitutto $V \ge 0$. Ogni autovalore λ di A è reale positivo e dunque ogni autovalore $1-\sigma\lambda$ di V è minore di 1. Essendo V una matrice non negativa, per il teorema di Perron–Frobenius il raggio spettrale $\rho(V)$ (che è non negativo) di V è un autovalore di V. Pertanto, anch'esso è minore di 1 (in pratica significa che tutti gli autovalori sono in modulo minori di 1). Allora V è una matrice convergente (cioè $\lim_{m\to\infty} V^m = 0$ (vedi Proposizione 1)) e vale

$$\frac{1}{\sigma}A^{-1} = (\sigma A)^{-1} = (I - V)^{-1} = \sum_{m=0}^{\infty} V^m$$

Pertanto, A^{-1} è la somma di una serie di matrici non negative.

A.2 Positività di Eulero implicito per equazioni di trasporto-diffusione

Dato il problema di trasporto-diffusione (23.7) con condizioni al bordo di Dirichlet omogenee, si può considerare una discretizzazione spaziale a nodi interni tale che il problema semidiscretizzato sia

$$\begin{bmatrix} y_2'(t) \\ \vdots \\ y_{m-1}'(t) \end{bmatrix} = A \begin{bmatrix} y_2(t) \\ \vdots \\ y_{m-1}(t) \end{bmatrix}$$

con

$$A = -\frac{c}{2h}\begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ -1 & 0 & 1 & \dots & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & \dots & -1 & 0 & 1 \\ 0 & \dots & 0 & -1 & 0 \end{bmatrix} + \frac{d}{h^2}\begin{bmatrix} -2 & 1 & 0 & \dots & 0 \\ 1 & -2 & 1 & \dots & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & \dots & 1 & -2 & 1 \\ 0 & \dots & 0 & 1 & -2 \end{bmatrix}$$

Mediante il teorema dei dischi di Gershgorin possiamo dire che gli autovalori di A stanno nel disco di centro $-2d/h^2$ e passante per l'origine. Infatti, il raggio dei dischi vale

$$\left|\frac{c}{2h} + \frac{d}{h^2}\right| + \left|-\frac{c}{2h} + \frac{d}{h^2}\right|$$

Siccome, se il numero di Péclet di griglia è minore o uguale a 1, $|c|/(2h) \leq d/h^2$, allora tale raggio vale $2d/h^2$. La matrice "da invertire" per il metodo di Eulero implicito è $I - kA$, ove k è il passo temporale. Gli autovalori di tale matrice stanno nel disco di centro $1 + k2d/h^2$ e raggio $k2d/h^2$ e sono pertanto diversi da zero. Dunque, per qualunque k la matrice è invertibile. Prendiamo adesso $\sigma = 1/(1 + k2d/h^2)$ e definiamo

$$V = I - \sigma(I - kA)$$

Si osserva facilmente che $V \geq 0$. Poi, La matrice $-\sigma(I - kA)$ ha autovalori nel disco di centro -1 e raggio $\sigma k2d/h^2 < 1$. Pertanto V ha autovalori nel disco di centro 0 e $\rho(V) < 1$. E si conclude come sopra. Dunque la matrice $(I - kA)^{-1}$ ha elementi non negativi. Lo stesso per la matrice del metodo dei trapezi $(I - kA/2)^{-1}$. Ma mentre per Eulero implicito la matrice $(I - kA)^{-1}$ è applicata a $\boldsymbol{y}_n$ (che si suppone avere elementi non negativi), per il metodo dei trapezi la matrice $(I - k/2A)^{-1}$ si applica a $(I + kA/2)\boldsymbol{y}_n$. Se il numero di Péclet è minore o uguale a 1, i termini extradiagonali di $(I + kA/2)$ sono non negativi, mentre i termini diagonali lo sono solo per $1 - kd/h^2 \geq 0$, cioè $k \leq h^2/d$.

Per esempio, per il problema di diffusione

$$\begin{cases} u_t = u_{xx} & t > 0,\ x \in (0,1) \\ u(t,0) = u(t,1) = 0 & t > 0 \\ u(0,x) = \begin{cases} 0 & x \in [0,1/2) \\ 1 & x \in [1/2,1) \\ 0 & x = 1 \end{cases} \end{cases}$$

discretizzato con nodi interni e $h = 1/50$, si hanno i risultati in Figura A.1.

sispos.m

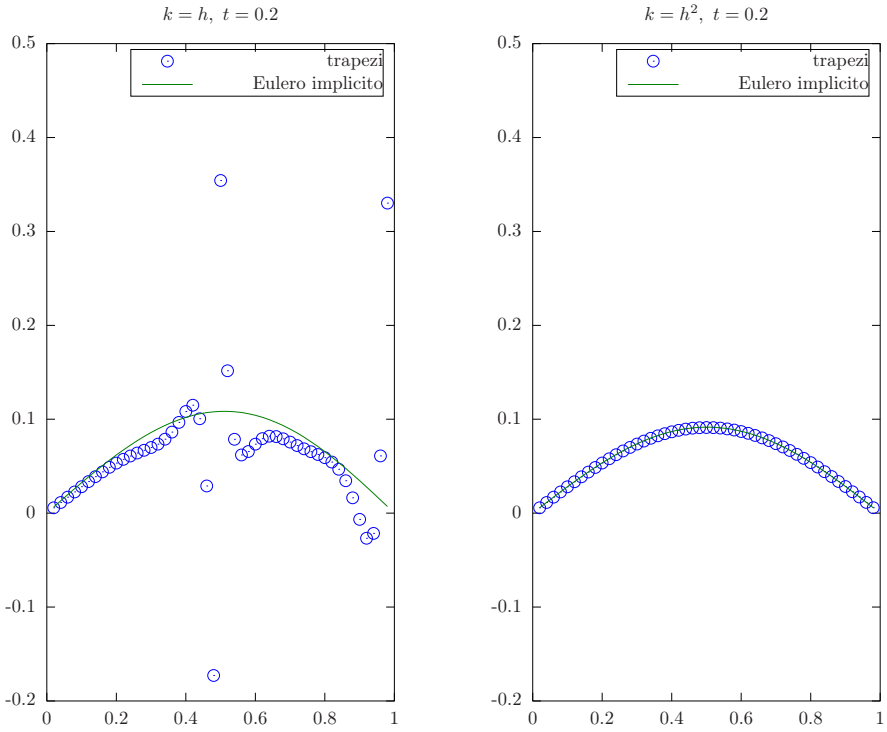


Figura A.1: Soluzione al tempo $t = 0.2$ con i metodi dei trapezi e di Eulero implicito, con $k = h$ (sinistra) e $k = h^2$ (destra).

Per le matrici di trasporto-diffusione è possibile dare un'altra stima per lo spettro. Se decomponiamo A nelle sue parti simmetrica A_S (diffusione) e antisimmetrica A_{AS} (trasporto) abbiamo che $W(A) \subseteq W(A_S) + W(A_{AS})$. Ma le parti simmetrica e antisimmetrica sono normali e dunque il campo dei valori è l'involucro convesso dello spettro. Per la parte simmetrica, usando i dischi di Gershgorin e il fatto che gli autovalori sono reali, abbiamo che $W(A_S) \subseteq [-4d/h^2, 0]$ (in realtà 0 si può escludere perché la matrice non è singolare). Per la parte antisimmetrica, usando i dischi di Gershgorin e il fatto che gli autovalori sono puramente immaginari, abbiamo che $W(A_{AS}) \subseteq \mathrm{i}[-|c|/(2h), -|c|/(2h)]$. Pertanto il campo dei valori $W(A)$ è contenuto nel rettangolo $[-4d/h^2, 0] + \mathrm{i}[-|c|/(2h), -|c|/(2h)]$ e a maggior ragione lo spettro. Se il numero di Péclet di griglia vale 1, allora tale rettangolo è il quadrato circoscritto al disco di centro $-2d/h^2$ e passante per l'origine (dunque non si guadagna informazione rispetto ai dischi di Gershgorin della matrice). Altrimenti, tale rettangolo ha intersezione non banale con il disco.

A.3 Equazione del filo elastico

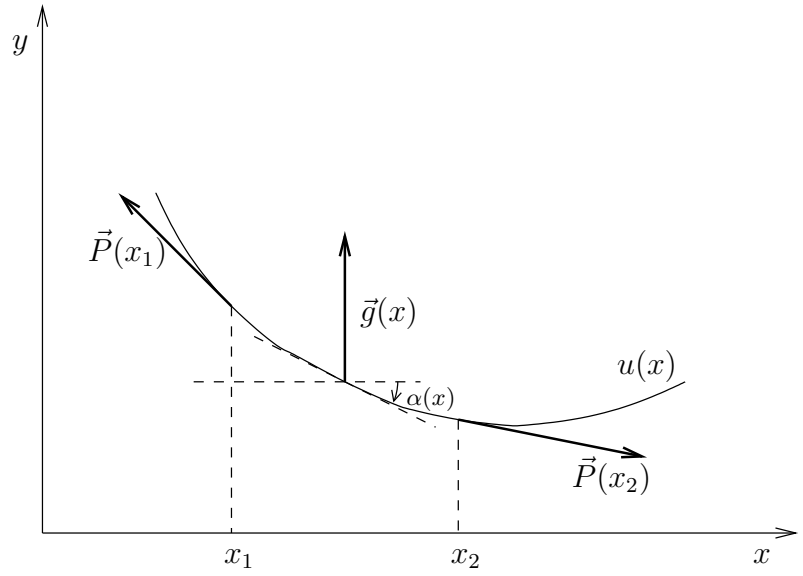


Figura A.2: Filo elastico sottoposto a carico $\vec{g}$.

L'equazione del filo elastico (14.1) si ricava nel seguente modo. Con riferimento alla Figura A.2, se il filo si trova a riposo significa che la risultante delle forze è nulla nell'intervallo $[x_1, x_2]$ e pertanto (trascurando il peso del filo)

$$\vec{P}(x_1) + \vec{P}(x_2) + \int_{x_1}^{x_2} \vec{g}(x)\mathrm{d}x = 0$$

In un generico punto x nell'intervallo $[x_1, x_2]$ si ha

$$\tan\alpha(x) = u'(x)$$

da cui

$$\sin\alpha(x) = \frac{u'(x)}{\sqrt{1+(u'(x))^2}} \approx u'(x)$$

se $\alpha(x)$ è piccolo e dunque $(u'(x))^2 \ll 1$. Proiettando sull'asse y (tenendo conto che $\alpha(x) < 0$) otteniamo

$$-P(x_1)\sin\alpha(x_1) + P(x_2)\sin\alpha(x_2) + \int_{x_1}^{x_2} g(x)\mathrm{d}x = 0$$

da cui

$$\int_{x_1}^{x_2} \frac{\mathrm{d}}{\mathrm{d}x}\left(P(x)u'(x)\right)\mathrm{d}x = -\int_{x_1}^{x_2} g(x)\mathrm{d}x$$

Proiettando sull'asse x, si ottiene invece

$$-P(x_1)\cos\alpha(x_1) + P(x_2)\cos\alpha(x_2) = 0$$

Ancora, se $\alpha(x)$ è piccolo, allora $\cos\alpha(x) \approx 1$ e dunque

$$P(x_1) = P(x_2) \equiv P$$

Pertanto

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

A.4 Equazione della trave

L'equazione della trave (11.2) sottoposta a carico e trazione, si ricava nel seguente modo. Con riferimento alla Figura A.3, nell'intervallo $[x_1, x_2]$ l'e-

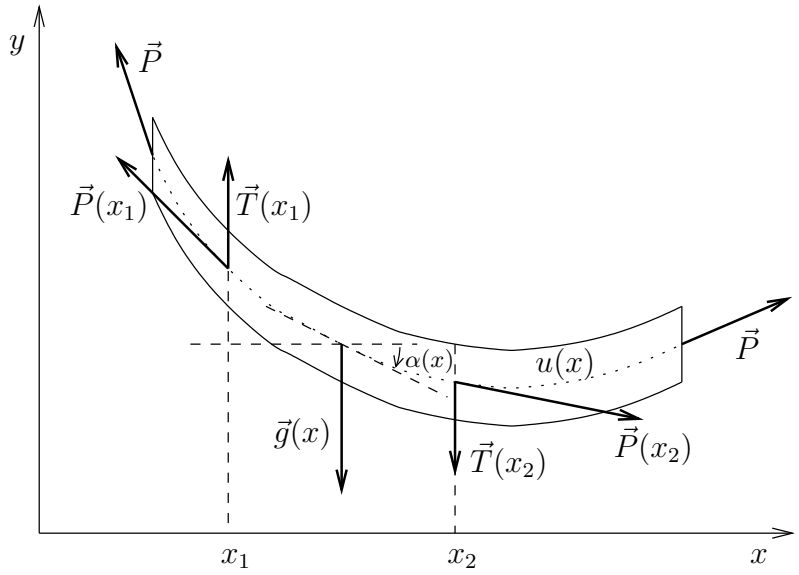


Figura A.3: Trave elastica sottoposta a carico $\vec{g}$ e trazione $\vec{P}$.

quilibrio delle forze si scrive

$$\vec{T}(x_1) + \int_{x_1}^{x_2} \vec{g}(x)\mathrm{d}x + \vec{T}(x_2) + \vec{P}(x_1) + \vec{P}(x_2) = 0$$

Le forze $\vec{T}(x)$ si chiamano *sforzi di taglio* e rappresenta la componente verticale delle forze di contatto a cui è soggetto un corpo. Lo sforzo di taglio $\vec{T}(x_2)$

ha orientamento opposto a $\vec{T}(x_1)$ in quanto inteso come reazione dell'azione di $-\vec{T}(x_2)$ in $x > x_2$. Proiettando sull'asse y, si ottiene

$$T(x_1) - \int_{x_1}^{x_2} g(x)\mathrm{d}x - T(x_2) - P(x_1)\sin\alpha(x_1) + P(x_2)\sin\alpha(x_2) = 0$$

Come nel caso del filo elastico, si suppone $\sin\alpha(x) \approx u'(x)$ e dunque

$$-\int_{x_1}^{x_2} \frac{\mathrm{d}}{\mathrm{d}x} T(x)\mathrm{d}x - \int_{x_1}^{x_2} g(x)\mathrm{d}x + \int_{x_1}^{x_2} \frac{\mathrm{d}}{\mathrm{d}x}\left(P(x)u'(x)\right)\mathrm{d}x$$

Proiettando sull'asse x, otteniamo, come nel caso del filo elastico,

$$P(x_1) = P(x_2) \equiv P$$

e pertanto

$$-T'(x) - g(x) + Pu''(x) = 0$$

Il *momento flettente* $\mu(x)$ è il momento risultante da tutte le forze verticali, e pertanto nell'intervallo $[x_1, x]$, $x \leq x_2$

$$\mu(x) - \mu(x_1) = \int_{x_1}^{x} T(t)\mathrm{d}t - \frac{x - x_1}{2}\int_{x_1}^{x} g(t)\mathrm{d}t$$

ove si è pensato il carico come una forza concentrata nel baricentro del segmento $[x_1, x]$. Dividento per $(x - x_1)$ e facendo tendere x_1 a x si ha

$$\mu'(x) = T(x)$$

Per finire, usando il *modello di elasticità lineare* che stabilisce

$$u''(x) = \frac{\mu(x)}{E \cdot I(x)}$$

si arriva a

$$-\mu''(x) + c(x)\mu(x) = g(x)$$

ove

$$c(x) = \frac{P}{E \cdot I(x)}$$

Se $P = 0$, l'equazione è uguale a quella del filo elastico, ma il significato è diverso. Per esempio, una trave appoggiata ai bordi (per esempio $x = 0$ e $x = 1$) e sotto un carico costante di densità 2, soddisfa l'equazione

$$\begin{cases} -\mu''(x) = 2, & x \in (0,1) \\ \mu(0) = \mu(1) = 0 \end{cases}$$

Infatti, per quanto visto, un carico diretto verso il basso ha intensità positiva (al contrario del caso del filo elastico) e l'appoggio ai bordi significa che lì non c'è momento flettente. La soluzione è

$$\mu(x) = x(1-x)$$

e dunque

$$u''(x) = 10x(1-x)$$

(supponendo $E \cdot I(x) \equiv 1/10$). La trave soddisfa condizioni di Dirichlet, per esempio $u(0) = u(1) = 0$. Quindi

$$u(x) = 10\left(-\frac{x^4}{12} + \frac{x^3}{6} - \frac{x}{12}\right)$$

Un filo elastico sottoposto ad un carico uniforme diretto verso il basso di intensità 2 (e tensione unitaria) soddisfa invece

$$\begin{cases} -u''(x) = -2, & x \in (0,1) \\ u(0) = u(1) = 0 \end{cases}$$

e dunque

$$u(x) = x(x-1)$$

A.4.1 Appoggi ottimali per una trave

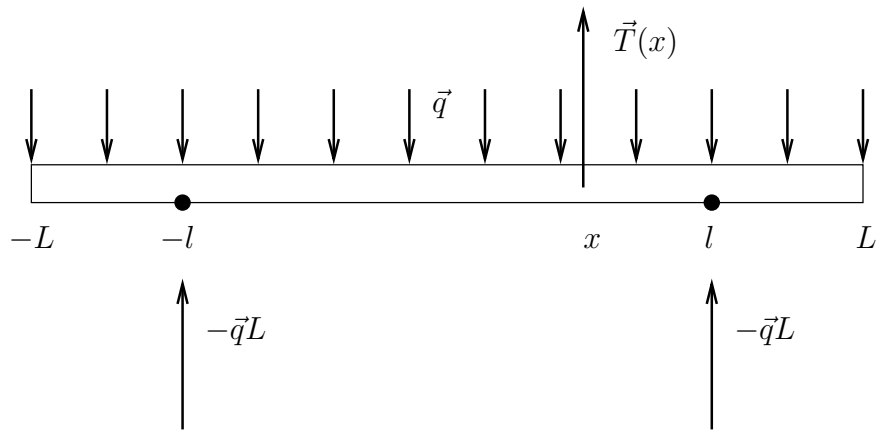


Figura A.4: Appoggi ottimali per una trave.

Si vuole determinare dove posizionare gli appoggi a sostegno di una trave omogenea, non in trazione, sottoposta a carico uniforme in modo da minimizzarne la deformazione. Con riferimento alla Figura A.4 e limitandosi a $x \geq 0$ per simmetria, lo sforzo di taglio $T(x)$ è

$$T(x) = \begin{cases} -qx, & 0 \leq x < l \\ qL - qx, & l \leq x \leq L \end{cases}$$

da cui il momento flettente

$$\mu(x) = \begin{cases} -\frac{1}{2}qx^2 + qLl - \frac{1}{2}qL^2, & 0 \leq x < l \\ qLx - \frac{1}{2}qx^2 - \frac{1}{2}qL^2, & l \leq x \leq L \end{cases}$$

in cui le costanti di integrazione sono state scelte per avere un momento continuo e $\mu(L) = 0$. Ora $u''(x) = \mu(x)$ (assumendo $E \cdot I = 1$) e si deve imporre la continuità di $u'(x)$ (due condizioni) e il passaggio di $u(x)$ da $(l, 0)$ (due condizioni). Sfruttando ancora la simmetria del problema, si trova

$$u(x) = \begin{cases} -\frac{q}{24}x^4 + \frac{q}{2}Llx^2 - \frac{q}{4}L^2x^2 + \frac{q}{24}l^4 - \frac{q}{2}Ll^3 + \frac{q}{4}qL^2l^2, & 0 \leq x < l \\ -\frac{q}{24}x^4 + \frac{q}{6}Lx^3 - \frac{q}{4}L^2x^2 + \frac{q}{2}Ll^2x + \frac{q}{24}l^4 - \frac{2q}{3}Ll^3 + \frac{q}{4}L^2l^2, & l \leq x \leq L \end{cases}$$

I punti critici di questa funzione sono

$$\begin{cases} x_1 = 0 & \\ x_2 = \sqrt{6Ll - 3L^2} & \text{se } l \geq L/2 \text{ e } x_2 < l, \text{ cioè } L/2 \leq l < (3 - \sqrt{6})L \\ x_3 = L - \sqrt[3]{L^3 - 3Ll^2} & \text{se } l \leq x_3 \leq L, \text{ cioè } (3 - \sqrt{6})L \leq l \leq \sqrt{3}L/3 \\ x_4 = L & \text{estremo} \end{cases}$$

(per ricavare x_3 si aggiunge e toglie $qL^3/6$ alla derivata del secondo pezzo di $u(x)$). Valutando $u(x)$ nei punti critici si ottengono quattro funzioni dipendenti da l: occorre stabilire per quale valore di l il massimo del valore assoluto di tali funzioni è minimo. Aiutandosi con un grafico, si scopre che tale l si ottiene dall'intersezione delle due funzioni corrispondenti a $x_1 = 0$ e $x_4 = L$, per cui l ottimale è soluzione di

$$\frac{q}{24}l^4 - \frac{q}{2}Ll^3 + \frac{q}{4}L^2l^2 = \\ - \frac{q}{24}L^4 + \frac{q}{6}L^4 - \frac{q}{4}L^4 + \frac{q}{2}L^2l^2 + \frac{q}{24}l^4 - \frac{2q}{3}Ll^3 + \frac{q}{4}L^2l^2$$

Si ottiene $l = (\sqrt{3}\sin(\arctan(\sqrt{39}/5)/3) - \cos(\arctan(\sqrt{39}/5)/3) + 1)L \approx 0.5537 \cdot L$.

Si osservi che la banale derivazione $u^{(4)}(x) = \mu''(x) = -q$ avrebbe portato al problema

$$\begin{cases} -u^{(4)}(x) = q, & x \in (-L, L) \\ u(-l) = u(l) = 0 \\ u''(-L) = u''(L) = 0 \end{cases}$$

la cui soluzione forte (cioè $\mathcal{C}^4$), del tutto diversa, è

$$u(x) = -\frac{q}{24}x^4 + \frac{q}{4}L^2x^2 + \frac{q}{24}l^4 - \frac{q}{4}L^2l^2$$

A.5 Lunghezza della catenaria

L'equazione della catenaria si riferisce alla curva assunta da una corda flessibile e inestensibile appesa agli estremi. Il parametro a presente nell'equazione (11.9) dipende allora solo dalla lunghezza l della corda, data da

$$\int_{-1}^{1} \sqrt{1 + (u'(x))^2}\mathrm{d}x = \frac{2\sinh a}{a} = l$$

Per risolvere l'equazione in a $f(a) = 2\sinh a - al = 0$, si può usare il metodo di Newton a partire dal punto a_0 soluzione di

$$2\left(a_0 + \frac{a_0^3}{6}\right) - a_0 l = 0$$

Infatti, $f(a)$ è convessa e $\sinh(a) = a + a^3/6 + a^5/5! + \mathcal{O}(a^7)$ e quindi $a < a_0$.

A.6 A-stabilità di un metodo di Runge–Kutta semiimplicito

Verifichiamo che il metodo di Runge–Kutta semiimplicito (19.9) è A-stabile.

Applicato al solito problema dà

$$\begin{cases} \xi_1 = y_n + ka_{1,1}\lambda\xi_1 \\ \xi_2 = y_n + ka_{2,1}\lambda\xi_1 + ka_{2,2}\lambda\xi_2 \\ y_{n+1} = y_n + kb_1\lambda\xi_1 + kb_2\lambda\xi_2 \end{cases}$$

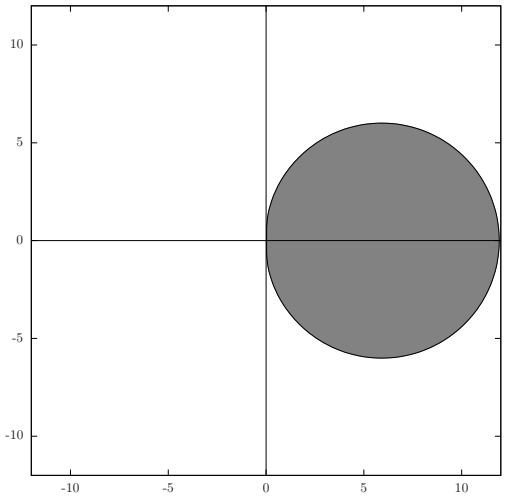


Figura A.5: Regione di assoluta stabilità (bianca) per il metodo di Runge–Kutta semiimplicito (19.9).

e dunque

$$r(k\lambda) = r(z) = 1 + b_1 z \frac{1}{1 - a_{1,1}z} + b_2 z \frac{1 + \frac{a_{2,1}z}{1-a_{1,1}z}}{1 - a_{2,2}z}$$

Quindi, chiedere $|r(z)| < 1$ è come chiedere (tenendo conto che $a_{1,1} = a_{2,2}$ e che $b_1 + b_2 = 1$)

$$|(1 - a_{1,1}z)^2 + z(1 - a_{1,1}z) + b_2 a_{2,1} z^2|^2 < |(1 - a_{1,1}z)^2|$$

da cui dividendo per $a_{1,1}$ e sostituendone il valore

$$\left| -\frac{\sqrt{3}}{3}z^2 + (1 - \sqrt{3})z + 3 - \sqrt{3} \right|^2 - \left| \frac{3 + \sqrt{3}}{6}z^2 - 2z + 3 - \sqrt{3} \right|^2 < 0$$

Scrivendo ora z come $x + \mathrm{i}y$, si trova

$$\sqrt{3}y^4 + 2\sqrt{3}x^2y^2 - 24xy^2 + \sqrt{3}x^4 - 24x^3 + 24\sqrt{3}x^2 + (72\sqrt{3} - 144)x > 0$$

da cui segue subito che la disuguaglianza è soddisfatta se $x = \Re(z) < 0$.

Appendice B

Estrapolazione di Richardson

Procedendo con gli sviluppi in serie di Taylor per le differenze finite, si trova

$$u'(x_i) = \Delta u(x_i) - \frac{h^2}{6} u^{(3)}(x_i) + \mathcal{O}(h^4)$$

e, analogamente,

$$u''(x_i) = \Delta^2 u(x_i) - \frac{h^2}{12} u^{(4)}(x_i) + \mathcal{O}(h^4)$$

Data dunque l'approssimazione u_i^h di un problema ai limiti mediante differenze finite centrate con passo h che soddisfa

$$\begin{aligned} u(x_i) &= u_i^h + c_i h^2 + \mathcal{O}(h^4) \\ u(x_i) &= u_i^{h/2} + c_i \left(\frac{h}{2}\right)^2 + \mathcal{O}(h^4) \end{aligned}$$

si ha

$$\frac{4u_i^{h/2} - u_i^h}{3} = u(x_i) + \mathcal{O}(h^4)$$

Appendice C

Temi d'esame

29-06-2010

1. Si implementi il metodo di Runge–Kutta definito dal tableau

$$\begin{array}{c|ccccc} 0 & & & & & \\ \frac{1}{2} & \frac{1}{2} & & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & & \\ 1 & 0 & 0 & 1 & & \\ \frac{3}{4} & \frac{5}{32} & \frac{7}{32} & \frac{13}{32} & -\frac{1}{32} & \\ \hline & -\frac{1}{2} & \frac{7}{3} & \frac{7}{3} & \frac{13}{6} & -\frac{16}{3} \end{array}$$

   e se ne determini numericamente l'ordine.

2. Si usi il metodo di Runge–Kutta implementato sopra per determinare numericamente il periodo del pendolo

$$\begin{cases} \theta''(t) = -g\sin(\theta(t)), \quad t > 0 \\ \theta(0) = \dfrac{\pi}{3} \\ \theta'(0) = 0 \end{cases}$$

   e si dica se è minore, maggiore o uguale al periodo del pendolo *linearizzato*

$$\begin{cases} \theta_1''(t) = -g\theta_1(t), \quad t > 0 \\ \theta_1(0) = \dfrac{\pi}{3} \\ \theta_1'(0) = 0 \end{cases}$$

3. Usando differenze finite del secondo ordine per lo spazio e il metodo di Eulero implicito per il tempo si discretizzi il problema di diffusione-reazione

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \dfrac{1}{100}\dfrac{\partial^2 u}{\partial x^2}(t,x) + 5u(t,x)\left(u(t,x) - \dfrac{1}{2}\right)(1-u(t,x)), & t>0,\ x\in(0,1) \\ u(0,x) = 4x(1-x), & x\in(0,1) \\ u(t,0) = u(t,1) = 0, & t>0 \end{cases}$$

4. Si mostri che il metodo usato per risolvere i sistemi non lineari e il metodo di Eulero implicito hanno il corretto ordine di convergenza.

20-07-2010

5. Si consideri il metodo implicito ad un passo

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h\boldsymbol{f}\left(t_n + \frac{h}{2}, \frac{1}{2}(\boldsymbol{y}_n + \boldsymbol{y}_{n+1})\right), \quad n \geq 0$$

per la soluzione di un problema ai valori iniziali

$$\begin{cases} \boldsymbol{y}'(t) = \boldsymbol{f}(t, \boldsymbol{y}(t)), & t > t_0 \\ \boldsymbol{y}(t_0) = \boldsymbol{y}_0 \end{cases}$$

Lo si implementi e lo si applichi al problema autonomo

$$\begin{cases} y'(t) = y(t)(1-y(t)), & t>0 \\ y(0) = \dfrac{1}{2} \end{cases}$$

per determinare numericamente l'ordine con cui viene approssimata la soluzione $y(t^*)$ al tempo $t^* = 1$.

6. Si calcoli la soluzione analitica del problema di diffusione non omogeneo

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \dfrac{\partial^2 u}{\partial x^2}(t,x) - (\sin t)(-x^2+2x) + 2\cos t, & t\geq 0,\ x\in(0,1) \\ u(0,x) = -x^2 + 2x, & x\in(0,1) \\ u(t,0) = 0, & t\geq 0 \\ \dfrac{\partial u}{\partial x}(t,1) = 0, & t\geq 0 \end{cases}$$

e si mostri l'ordine di convergenza temporale del metodo Eulero esponenziale per approssimare la soluzione $u(t^*, x)$ al tempo $t^* = 1$.

7. Per lo stesso problema, si mostri l'ordine di convergenza temporale del metodo esponenziale—punto medio.

15-09-2010

8. Si consideri il metodo di Runge–Kutta semi-implicito

$$\begin{cases} \boldsymbol{f}_i = \boldsymbol{f}\left(t_n + c_i k, \boldsymbol{y}_n + k\sum_{j=1}^{i} a_{i,j}\boldsymbol{f}_j\right), \quad i = 1,2 \\ \boldsymbol{y}_{n+1} = \boldsymbol{y}_n + k\sum_{j=1}^{2} b_j \boldsymbol{f}_j \end{cases}$$

di tableau

$$\begin{array}{c|cc} c_1 & a_{1,1} & a_{1,2} \\ c_2 & a_{2,1} & a_{2,2} \\ \hline & b_1 & b_2 \end{array} = \begin{array}{c|cc} \frac{3+\sqrt{3}}{6} & \frac{3+\sqrt{3}}{6} & 0 \\ \frac{3-\sqrt{3}}{6} & -\frac{\sqrt{3}}{3} & \frac{3+\sqrt{3}}{6} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

per la soluzione di un problema ai valori iniziali

$$\begin{cases} \boldsymbol{y}'(t) = \boldsymbol{f}(t, \boldsymbol{y}(t)), \quad t \geq t_0 \\ \boldsymbol{y}(t_0) = \boldsymbol{y}_0 \end{cases}$$

Lo si implementi per il problema autonomo

$$\begin{cases} y'(t) = y(t)(1 - y(t)), \quad t \geq 0 \\ y(0) = \dfrac{1}{2} \end{cases}$$

e si determini numericamente l'ordine con cui viene approssimata la soluzione $y(t^*)$ al tempo $t^* = 1$.

9. Si applichi il metodo delle linee al problema di diffusione non omogeneo

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \dfrac{\partial^2 u}{\partial x^2}(t,x) + (2x - x^2)\cos t + 2\sin t, & t \geq 0,\ x \in (0,1) \\ u(0,x) = 0, & x \in (0,1) \\ u(t,0) = 0, & t \geq 0 \\ u(t,1) = \sin t, & t \geq 0 \end{cases}$$

e si mostri l'ordine di convergenza temporale del metodo *Eulero implicito* con cui viene approssimata la soluzione $u(t^*, x)$ al tempo $t^* = 1$.

30-09-2010

10. Si risolva mediante il metodo delle differenze finite il problema ai limiti non lineare

$$\begin{cases} u''(x) = \dfrac{1}{8}(32 + 2x^3 - u(x)u'(x)), \quad x \in (1,3) \\ u(1) = 17 \\ u(3) = \dfrac{43}{3} \end{cases}$$

e si confronti il numero di iterazioni necessarie alla convergenza usando il metodo di Newton esatto e un metodo di Newton inesatto (la soluzione analitica è $u(x) = x^2 + 16/x$).

11. Si applichi il metodo delle linee al problema di diffusione non omogeneo

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \dfrac{\partial^2 u}{\partial x^2}(t,x) + 2e^t \sin x, & t > 0,\ x \in (0, \pi/2) \\ u(0,x) = \sin x, & x \in (0, \pi/2) \\ u(t,0) = 0, & t \geq 0 \\ \dfrac{\partial u}{\partial x}(t, \pi/2) = 0, & t \geq 0 \end{cases}$$

e si mostri l'ordine di convergenza temporale del metodo *Eulero esponenziale* con cui viene approssimata la soluzione analitica $u(t^*, x)$ al tempo $t^* = 1$.

01-02-2011

12. Si risolva il sistema di ODEs

$$\begin{cases} u'(t) = -2v(t)u(t) \\ v'(t) = u(t)^2 + z(t)^2 - v(t)^2 - 1 \\ z'(t) = -2(v(t) + u(t))z(t) \end{cases}$$

con dato iniziale

$$\begin{cases} u(0) = 1 \\ v(0) = 2 \\ z(0) = 15 \end{cases}$$

con il metodo di Crank–Nicolson fino ad un tempo finale $t^* = 1$. Presa come soluzione di riferimento quella ottenuta con 1000 passi temporali, si mostri l'ordine del metodo e la corretta convergenza del metodo di Newton.

13. Si applichi il metodo delle linee al problema di convezione-diffusione-reazione nel dominio $(t, x) \in [0, 1] \times [0, 1]$

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \varepsilon \dfrac{\partial^2 u}{\partial x^2}(t,x) + \dfrac{\partial u}{\partial x}(t,x) + \rho u(t,x)(u(t,x) - 1/2)(1 - u(t,x)) \\ u(0,x) = 10x^2(1-x)^2 + 1/2 \\ \dfrac{\partial u}{\partial x}(t,0) = \dfrac{\partial u}{\partial x}(t,1) = 0 \end{cases}$$

(ove $\varepsilon = 1/100$ e $\rho = 10$) e si mostri l'ordine di convergenza del metodo Eulero esponenziale, avendo preso come soluzione di riferimento quella ottenuta con un passo spaziale e un passo temporale entrambi pari a $1/100$.

24-02-2011

14. Si risolva mediante un metodo di shooting il problema ai limiti

$$\begin{cases} u''(x) = \dfrac{1}{8}(32 + 2x^3 - u(x)u'(x)), \quad x \in (1,3) \\ u(1) = 17 \\ u(3) = \dfrac{43}{3} \end{cases}$$

Sapendo che la soluzione analitica è $u(x) = x^2 + 16/x$, si determini sperimentalmente il numero minimo di passi temporali per avere un errore in norma infinito minore di 10^{-2}.

15. Si applichi il metodo delle linee al problema di diffusione non omogeneo

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \dfrac{\partial^2 u}{\partial x^2}(t,x) + \dfrac{5}{4}e^t \sin\dfrac{x}{2}, & t > 0,\ x \in (0,\pi) \\ u(0,x) = \sin\dfrac{x}{2}, & x \in (0,\pi) \\ u(t,0) = 0, & t \geq 0 \\ \dfrac{\partial u}{\partial x}(t,\pi) = 0, & t \geq 0 \end{cases}$$

e si mostri l'ordine di convergenza temporale del metodo *esponenziale punto medio* con cui viene approssimata la soluzione analitica $u(t^*, x)$ al tempo $t^* = 1$.

21-06-2011

16. Si risolva il seguente problema differenziale

$$\begin{cases} u''(x) + u(x) = 2\cos(x), & x \in (0,1] \\ u(0) = 0 \\ u'(0) = 0 \end{cases}$$

usando un metodo almeno del secondo ordine rispetto al passo di discretizzazione.

17. Si applichi il metodo delle linee al problema di diffusione-reazione

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \dfrac{1}{100}\dfrac{\partial^2 u}{\partial x^2}(t,x) + \sin(u(t,x)), & t > 0,\ x \in (0,1) \\ u(0,x) = 10x(1-x)^2, & x \in [0,1] \\ u(t,0) = 0, & t \geq 0 \\ \dfrac{\partial u}{\partial x}(t,1) = 0, & t \geq 0 \end{cases}$$

usando differenze finite centrate nello spazio e il metodo dei trapezi nel tempo. Si mostri il corretto ordine di convergenza del metodo dei trapezi per l'approssimazione della soluzione al tempo $t^* = 1$.

18. Per il problema sopra, si proponga un metodo di Newton modificato per la risoluzione dei sistemi non lineari. Si confrontino i tempi computazionali (con i comandi `tic`, `toc`) rispetto al caso di metodo di Newton esatto quando si usino un passo spaziale ed un passo temporale entrambi uguali a $\frac{1}{100}$.

11-07-2011

19. Si risolva il seguente problema ai limiti

$$\begin{cases} u''(x) = e^{u(x)} + 2 - e^{x^2}, & x \in (0,1) \\ u'(0) = 0 \\ u(1) = 1 \end{cases}$$

usando il metodo delle differenze finite di ordine due. È possibile verificare l'ordine di convergenza? Perché?

20. Si applichi il metodo delle linee al problema di diffusione-reazione

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \dfrac{1}{50}\dfrac{\partial^2 u}{\partial x^2}(t,x) + \cos(u(t,x)), & t > 0,\ x \in (0,1) \\ u(0,x) = 10x^2(1-x) + 1, & x \in [0,1] \\ \dfrac{\partial u}{\partial x}(t,0) = 0, & t \geq 0 \\ u(t,1) = 1, & t \geq 0 \end{cases}$$

usando differenze finite centrate nello spazio e il metodo Eulero implicito nel tempo. Si mostri il corretto ordine di convergenza del metodo di Eulero implicito per l'approssimazione della soluzione al tempo $t^* = 1$.

15-09-2011

21. La legge oraria *(lineare)* del moto di un proiettile sottoposto ad attrito viscoso in regime laminare è

$$\begin{cases} x''(t) = -Bx'(t) \\ x(0) = 0 \\ x'(0) = v_0\cos(\alpha) \end{cases} \qquad \begin{cases} y''(t) = -By'(t) - g \\ y(0) = 0 \\ y'(0) = v_0\sin(\alpha) \end{cases}$$

ove $B = b/m$, $m = 0.5$ la massa del proiettile e $b = 0.01$ il coefficiente d'attrito, $g = 9.81$ l'accelerazione di gravità, $v_0 = 200$ il modulo della velocità iniziale e $\alpha = \pi/3$ l'angolo di gittata. Determinare, mediante una opportuna strategia, il punto di atterraggio $x(T)$ del proiettile.

22. Si applichi il metodo delle linee al problema di diffusione-reazione

$$\begin{cases} \dfrac{\partial u}{\partial t} = \dfrac{1}{100}\dfrac{\partial^2 u}{\partial x^2} + u(1-u)(u-1/2), & t > 0,\ x \in (0,1) \\ u(0,x) = 10x^2(1-x)^2, & x \in [0,1] \\ \dfrac{\partial u}{\partial x}(t,0) = 0, & t \geq 0 \\ \dfrac{\partial u}{\partial x}(t,1) = 0, & t \geq 0 \end{cases}$$

usando differenze finite centrate nello spazio e il metodo dei trapezi nel tempo. Si mostri il corretto ordine di convergenza del metodo dei trapezi per l'approssimazione della soluzione al tempo $t^* = 1$.

29-09-2011

23. Si risolva il seguente problema differenziale

$$\begin{cases} y''(x) = 2y(x)^3 - 6y(x) - 2x^3, & x \in (1,2) \\ y(1) = 2 \\ y(2) = \dfrac{5}{2} \end{cases}$$

mostrando il corretto ordine di convergenza del metodo scelto.

24. Si applichi il metodo delle linee al problema di diffusione non omogeneo

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \dfrac{\partial^2 u}{\partial x^2}(t,x) + \dfrac{5}{4}e^t \cos\dfrac{x}{2}, & t > 0,\ x \in (\pi, 2\pi) \\ u(0,x) = \cos\dfrac{x}{2}, & x \in (\pi, 2\pi) \\ u(t,\pi) = 0, & t \geq 0 \\ \dfrac{\partial u}{\partial x}(t,2\pi) = 0, & t \geq 0 \end{cases}$$

e si mostri il corretto ordine di convergenza temporale del metodo *esponenziale punto medio* con cui viene approssimata la soluzione analitica $u(t^*, x)$ al tempo $t^* = 1$.

03-02-2012

25. Si risolva il seguente problema differenziale

$$\begin{cases} y''(x) + \dfrac{y'(x)}{x} = \cos(y(x)) & x \in (0,1] \\ y(0) = 1 \\ y'(0) = 0 \end{cases}$$

Si descriva esattamente quale metodo è stato usato e se ne mostri il corretto ordine di convergenza.

Si discuta inoltre il caso in cui $y(0) = \pi/2$.

26. Si applichi il metodo delle linee al problema di diffusione non omogeneo

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \dfrac{\partial^2 u}{\partial x^2}(t,x) + \dfrac{3}{2}e^{t/2} \sin x, & t > 0,\ x \in (-\pi, \pi/2) \\ u(0,x) = \sin x, & x \in (-\pi, \pi/2) \\ u(t,-\pi) = 0, & t \geq 0 \\ \dfrac{\partial u}{\partial x}(t,\pi/2) = 0, & t \geq 0 \end{cases}$$

usando differenze finite nello spazio e l'integratore esponenziale punto medio nel tempo. Si mostri il corretto ordine di convergenza spaziale con cui viene approssimata la soluzione $u(t^*, x)$ al tempo $t^* = 1$.

24-02-2012

27. Si risolva il seguente problema differenziale ai valori iniziali

$$\begin{cases} y_1'(t) = -2y_1(t)y_2(t) \\ y_2'(t) = y_1(t)^2 - y_2(t)^2 + y_3(t)^2 - 1 \\ y_3'(t) = -2(y_1(t) + y_2(t))y_3(t) \\ y_1(0) = y_2(0) = y_3(0) = 1 \end{cases}$$

fino al tempo $t^* = 1$ usando il metodo Runge–Kutta semiimplicito di tableau

$$\begin{array}{c|cc} 0 & 0 & \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

e se ne calcoli numericamente l'ordine.

28. Si applichi il metodo delle linee al problema di diffusione lineare

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \dfrac{x - x^2}{2} \cdot \dfrac{\partial^2 u}{\partial x^2}(t,x), & t > 0,\ x \in (0,1) \\ u(0,x) = x^2 - x, & x \in (0,1) \\ u(t,0) = 0, & t \geq 0 \\ u(t,1) = 0, & t \geq 0 \end{cases}$$

usando differenze finite nello spazio e il metodo di Eulero implicito nel tempo. Si mostri il corretto ordine di convergenza temporale con cui viene approssimata la soluzione al tempo $t^* = 1$. Qual è l'ordine di convergenza spaziale?

21-06-2012

29. Dei seguenti due metodi multistep

$$\boldsymbol{y}_{n+3} - \frac{18}{11}\boldsymbol{y}_{n+2} + \frac{9}{11}\boldsymbol{y}_{n+1} - \frac{2}{11}\boldsymbol{y}_n = k\frac{6}{11}\boldsymbol{f}(t_{n+3}, \boldsymbol{y}_{n+3})$$
$$\boldsymbol{y}_{n+3} - \frac{18}{11}\boldsymbol{y}_{n+2} + \frac{9}{11}\boldsymbol{y}_{n+1} - \frac{3}{11}\boldsymbol{y}_n = k\frac{6}{11}\boldsymbol{f}(t_{n+3}, \boldsymbol{y}_{n+3})$$

si dica quale è consistente e perché e se ne determini numericamente l'ordine.

30. Si risolva il problema differenziale di diffusione-trasporto-reazione

$$\begin{cases} \dfrac{\partial u}{\partial t} + c\dfrac{\partial u}{\partial x} = d\dfrac{\partial^2 u}{\partial x^2} + \rho u(u-1/2)(1-u), & t > 0,\ x \in (0,1) \\ u(t,0) = 1, & t > 0 \\ \dfrac{\partial u}{\partial x}(t,1) = 0, & t > 0 \\ u(0,x) = (x-1)^2, & x \in (0,1) \end{cases}$$

con $d = 0.01$, $c = 8$, $\rho = 50$, usando differenze finite nello spazio (con passo $1/100$) ed un opportuno metodo implicito nel tempo fino al tempo finale $t^* = 0.1$, usando il metodo di Newton per la risoluzione dei sistemi non lineari. A cosa sono dovute le oscillazioni vicino al bordo $x = 1$ al tempo finale? Come si possono eliminare?

05-07-2012

31. Si risolva il seguente problema ai limiti

$$\begin{cases} u''(x) - u'(x) + u^2(x) = e^{2x}, & x \in (0,1) \\ u'(0) = 1 \\ u(1) = e \end{cases}$$

con il metodo delle differenze finite. Si verichi il corretto ordine di approssimazione di $u'(0)$.

32. Si applichi il metodo delle linee al problema di diffusione non omogeneo

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = 4\dfrac{\partial^2 u}{\partial x^2}(t,x) + 3e^{2t}\cos\dfrac{x}{2}, & t > 0,\ x \in (0,\pi) \\ u(0,x) = \cos\dfrac{x}{2}, & x \in (0,\pi) \\ \dfrac{\partial u}{\partial x}(t,0) = 0, & t \geq 0 \\ u(t,\pi) = 0, & t \geq 0 \end{cases}$$

e si mostri il corretto ordine di convergenza temporale del metodo *esponenziale punto medio* con cui viene approssimata la soluzione analitica $u(t^*,x)$ al tempo $t^* = 1$.

10-09-2012

33. Dato il sistema differenziale del primo ordine

$$\begin{cases} y_1'(t) = y_2(t)y_3(t)\sin t - y_1(t)y_2(t)y_3(t) \\ y_2'(t) = -y_1(t)y_3(t)\sin t + \dfrac{1}{20}y_1(t)y_3(t) \\ y_3'(t) = y_1^2(t)y_2(t) - \dfrac{1}{20}y_1(t)y_2(t) \end{cases}$$

si dimostri che $m(t) = \sqrt{y_1(t)^2 + y_2(t)^2 + y_3(t)^2}$ rimane costante nel tempo. Si applichi il metodo di Eulero implicito, con dato iniziale $y_1(0) = y_2(0) = y_3(0) = \sqrt{3}/3$ fino al tempo $t^* = 1$ e si determini sperimentalmente il numero di passi temporali necessari perché $|\tilde{m}(0) - \tilde{m}(1)| \leq 2 \cdot 10^{-5}$, ove $\tilde{m}(t)$ è l'approssimazione di $m(t)$ calcolata sulla soluzione numerica.

34. Si risolva il seguente problema di convezione-diffusione

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = 4\dfrac{\partial^2 u}{\partial x^2}(t,x) - 2\dfrac{\partial u}{\partial x}(t,x), & t \geq 0,\ x \in (0,1) \\ u(0,x) = x^2, & x \in (0,1) \\ \dfrac{\partial u}{\partial x}(t,0) = 0, & t \geq 0 \\ u(t,1) = 1, & t \geq 0 \end{cases}$$

usando differenze finite del secondo ordine con passo $h = 1/50$ nello spazio e il metodo di Eulero esplicito nel tempo fino a $t^* = 1$. Perché serve un passo temporale minore di circa 1/20000 per avere convergenza? Si mostri infine il corretto ordine di convergenza temporale.

24-09-2012

35. Si trovi sperimentalmente l'ordine del seguente metodo di Runge–Kutta semiimplicito

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

e si dimostri che equivale al metodo *punto medio implicito*

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + k\boldsymbol{f}\left(t_n + \frac{k}{2}, \frac{\boldsymbol{y}_n + \boldsymbol{y}_{n+1}}{2}\right)$$

Sugg.: per quest'ultimo si calcoli $(\boldsymbol{y}_n + \boldsymbol{y}_{n+1})/2\ldots$

36. Si risolva il seguente problema di convezione-diffusione-reazione

$$\begin{cases} \dfrac{\partial u}{\partial t} + c\dfrac{\partial u}{\partial x} = d\dfrac{\partial^2 u}{\partial x^2}(t,x) + \rho u^2(1-u), & t \geq 0,\ x \in (0,1) \\ u(0,x) = -x^2 + 1, & x \in (0,1) \\ \dfrac{\partial u}{\partial x}(t,0) = 0, & t \geq 0 \\ u(t,1) = 0, & t \geq 0 \end{cases}$$

con $d = 0.1$, $c = 4$, $\rho = 50$, usando differenze finite nello spazio (con passo $1/100$) ed un metodo del secondo ordine implicito nel tempo fino al tempo finale $t^* = 0.1$, usando il metodo di Newton per la risoluzione dei sistemi non lineari. Si mostri il corretto ordine di convergenza temporale al tempo finale.

06-02-2013

37. Si risolva il seguente problema differenziale

$$\begin{cases} y'(x) = \sqrt{\dfrac{2 - y(x)}{y(x)}}, & x > 0 \\ y(0) = 0 \end{cases}$$

usando un opportuno metodo implicito. Si mostri il corretto ordine di convergenza con cui viene approssimata la soluzione $y(x^*)$ per $x^* = \pi/2$. *(Sugg.: $y'(x) > 0$ pertanto $y(x)$ è crescente e dunque $y(x) > 0$ per $x > 0$.)*

38. Si risolva il seguente problema di diffusione-trasporto

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \dfrac{1}{2}\dfrac{\partial^2 u}{\partial x^2}(t,x) - \dfrac{\partial u}{\partial x}(t,x) + e^{-t/2}\cos(x) & t > 0,\ x \in (0,\pi/2) \\ u(0,x) = \sin(x) & x \in (0,\pi/2) \\ u(t,0) = 0 & t > 0 \\ \dfrac{\partial u}{\partial x} u(t,\pi/2) = 0 & t > 0 \end{cases}$$

usando le differenze finite nello spazio e il metodo esponenziale—punto medio nel tempo. Si mostri il corretto ordine di convergenza temporale con cui viene approssimata la soluzione $u(t^*, x)$ al tempo finale $t^* = 1$.

20-02-2013

39. Si risolva il problema differenziale

$$\begin{cases} u''(x) = -3u'(x) + u(x) - 2\sin x + 3\cos x, & x \in (0, 2\pi) \\ u'(0) = 1 \\ u(2\pi) = 0 \end{cases}$$

mostrando il corretto ordine di convergenza del metodo scelto.

40. Si applichi il metodo delle linee al problema di diffusione-reazione

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \dfrac{1}{50}\dfrac{\partial^2 u}{\partial x^2}(t,x) + \sin(u(t,x)), & t > 0,\ x \in (0,1) \\ u(0,x) = 10x(1-x)^2, & x \in [0,1] \\ u(t,0) = 0, & t \geq 0 \\ \dfrac{\partial u}{\partial x}(t,1) = 0, & t \geq 0 \end{cases}$$

usando differenze finite centrate nello spazio con passo $h = 1/49$. Si mostrino i corretti ordini di convergenza dei metodi Eulero esponenziale e Eulero–Rosenbrock esponenziale con cui viene approssimata una soluzione di riferimento al tempo $t^* = 1$.

27-06-2013

41. Si determini analiticamente l'ordine del seguente metodo multistep

$$\boldsymbol{y}_{n+3} - \frac{18}{11}\boldsymbol{y}_{n+2} + \frac{9}{11}\boldsymbol{y}_{n+1} - \frac{2}{11}\boldsymbol{y}_n = k\frac{6}{11}\boldsymbol{f}(t_{n+3}, \boldsymbol{y}_{n+3})$$

Lo si applichi poi al problema

$$\begin{cases} y'(t) = -\dfrac{1}{2}y(t)\left(1 - \dfrac{y(t)}{3}\right), & t > 0 \\ y(0) = 1 \end{cases}$$

e si determini l'errore rispetto alla soluzione analitica

$$y(t) = \frac{3e^{-\frac{1}{2}t}}{2 + e^{-\frac{1}{2}t}}$$

al tempo $t^* = 10$, usando 10 passi temporali.
(Per i valori iniziali y_1 e y_2 si usi pure la soluzione analitica.)

42. Si applichi il metodo delle linee al problema di diffusione-trasporto

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \dfrac{1}{100}\dfrac{\partial^2 u}{\partial x^2}(t,x) + 6\dfrac{\partial u}{\partial x}(t,x) & t > 0,\ x \in (0,1) \\ u(0,x) = 5x^2\left(x - \dfrac{4}{5}\right), & x \in [0,1] \\ \dfrac{\partial u}{\partial x}(t,0) = 0, & t \geq 0 \\ u(t,1) = 1, & t \geq 0 \end{cases}$$

usando differenze finite centrate nello spazio con un opportuno passo di discretizzazione spaziale. Si mostri il corretto ordine di convergenza del metodo BDF di ordine 2 al tempo $t^* = 0.1$.

18-07-2013

43. Si consideri il seguente problema differenziale

$$\begin{cases} \dfrac{u''(x)}{10} + 900u'(x) + 500(\sin(u(x)) + u(x)) = 0, & x \in (0,10] \\ u(0) = \pi \\ u'(0) = 0 \end{cases}$$

Lo si risolva con un opportuno metodo esplicito, giustificando la scelta del passo di discretizzazione.

44. Si applichi il metodo delle linee al problema di diffusione-reazione

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \dfrac{1}{100}\dfrac{\partial^2 u}{\partial x^2}(t,x) + \sin(u(t,x)) & t > 0,\ x \in (0,1) \\ u(0,x) = 5x^2(x-1), & x \in [0,1] \\ \dfrac{\partial u}{\partial x}(t,0) = 0, & t \geq 0 \\ u(t,1) = 0, & t \geq 0 \end{cases}$$

usando differenze finite centrate nello spazio e il metodo punto medio implicito

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + k\boldsymbol{f}\left(t_n + \frac{k}{2}, \frac{\boldsymbol{y}_n + \boldsymbol{y}_{n+1}}{2}\right)$$

nel tempo. Si mostri il corretto ordine di convergenza temporale per la soluzione al tempo $t^* = 1$.

05-09-2013

45. Si risolva il seguente problema differenziale

$$\begin{cases} y'''(x) + \dfrac{1}{2}y(x)y''(x) = 0, & x \in (0,1) \\ y(0) = 0 \\ y'(0) = 0 \\ y''(0) = 10 \end{cases}$$

con un opportuno metodo del secondo ordine implicito. Che valore dovrebbe $y''(0)$ affinché risultasse $y'(1) = 11$?

46. Si applichi il metodo delle linee al problema di diffusione-reazione

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \dfrac{1}{20}\dfrac{\partial^2 u}{\partial x^2}(t,x) + \cos(u(t,x)), & t > 0,\ x \in (0,1) \\ u(0,x) = 10x^2(1-x), & x \in [0,1] \\ \dfrac{\partial u}{\partial x}(t,0) = 0, & t \geq 0 \\ u(t,1) = 0, & t \geq 0 \end{cases}$$

usando differenze finite centrate nello spazio con passo $h = 1/49$. Si mostrino i corretti ordini di convergenza dei metodi Eulero esponenziale e Eulero–Rosenbrock esponenziale con cui viene approssimata una soluzione di riferimento al tempo $t^* = 1$.

19-09-2013

47. Data la funzione $u(x)$ che soddisfa il seguente problema differenziale ($\log \equiv \log_e \equiv \ln$)

$$\begin{cases} u''(x) = \dfrac{2}{x}u'(x) - \dfrac{2}{x^2}u(x) + \sin(\log x), & x \in (1,2) \\ u(1) = 1 \\ u(2) = 2 \end{cases}$$

si calcoli quanto vale $u'(2)$ con almeno 4 cifre significative corrette.

48. Si applichi il metodo delle linee al problema di diffusione-reazione

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \dfrac{1}{4}\dfrac{\partial^2 u}{\partial x^2}(t,x) + u^2(t,x) - e^{-2t}\sin^2(2x), & t > 0,\ x \in \left(\dfrac{\pi}{4}, \dfrac{\pi}{2}\right) \\ u(0,x) = \sin(2x), & x \in \left[\dfrac{\pi}{4}, \dfrac{\pi}{2}\right] \\ \dfrac{\partial u}{\partial x}\left(t, \dfrac{\pi}{4}\right) = 0, & t \geq 0 \\ u\left(t, \dfrac{\pi}{2}\right) = 0, & t \geq 0 \end{cases}$$

usando differenze finite centrate nello spazio e il metodo di Eulero implicito nel tempo. Si mostri il corretto ordine di convergenza temporale con cui viene approssimata la soluzione analitica $u(t, x)$ al tempo $t^* = 1$.

11-02-2014

49. Si risolva il seguente problema differenziale

$$\begin{cases} y_1'(t) = -2y_1(t)y_2(t) \\ y_2'(t) = y_1(t)^2 - y_2(t)^2 + y_3(t) - 1 \\ y_3'(t) = -4(y_1(t) + y_2(t))y_3(t) \\ y_1(0) = 1/2,\ y_2(0) = 2,\ y_3(0) = \sqrt{10} \end{cases}$$

fino al tempo $t^* = 1$ usando il metodo Eulero–Rosenbrock esponenziale e se ne mostri l'ordine di convergenza per l'approssimazione di $\boldsymbol{y}(t_*)$.

50. Si consideri il seguente problema di diffusione-trasporto

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) + c\dfrac{\partial u}{\partial x}(t,x) = d\dfrac{\partial^2 u}{\partial x^2}(t,x) + c\dfrac{\pi}{2}e^{-t}\cos\left(\dfrac{\pi}{2}x\right) & t > 0,\ x \in (0,1) \\ u(t,0) = 0 & t > 0 \\ \dfrac{\partial u}{\partial x}(t,1) = 0 & t > 0 \\ u(0,x) = \sin\left(\dfrac{\pi}{2}x\right) & x \in (0,1) \end{cases}$$

ove $c = 150$ e $d = 4/\pi^2$ e se ne calcoli la soluzione analitica. Usando il metodo dei trapezi nel tempo con passo $k = 1/200$, si mostri il corretto ordine di convergenza spaziale con cui viene approssimata la soluzione analitica al tempo $t^* = 1$ usando la diffusione articiale $\phi_{\mathrm{U}}(\mathrm{Pe})$ corrispondente allo schema upwind e $\phi_{\mathrm{SG}}(\mathrm{Pe})$ corrispondente allo schema Scharfetter–Gummel, ove Pe è il numero di Péclet di griglia $ch/(2d)$.

25-02-2014

51. Si risolva il seguente problema differenziale

$$\begin{cases} u''(x) + u'(x) + e^{-x}u(x) = e^x, & x \in (0,1) \\ u(0) + u'(0) = 0 \\ u(1) = 1 \end{cases}$$

con un metodo del secondo ordine. Si verifichi che la condizione al bordo $x = 0$ è soddisfatta dalla soluzione numerica.

52. Si applichi il metodo delle linee al problema di diffusione-reazione

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \dfrac{1}{50}\dfrac{\partial^2 u}{\partial x^2}(t,x) + \cos(u(t,x)) & t > 0,\ x \in (0,1) \\ u(0,x) = x(x-1)^2, & x \in [0,1] \\ u(t,0) = 0, & t \geq 0 \\ \dfrac{\partial u}{\partial x}(t,1) = 0, & t \geq 0 \end{cases}$$

usando differenze finite centrate nello spazio e il metodo Runge–Kutta di tableau

$$\begin{array}{c|c} \frac{1}{2} & a \\ \hline & b \end{array}$$

(a e b da determinare). Si mostri il corretto ordine di convergenza temporale con cui viene approssimata la soluzione al tempo $t^* = 1$.

01-07-2014

53. Si consideri il problema differenziale

$$\begin{cases} y'(t) = \lambda(y(t) - \cos(t)), & t > 0 \\ y(0) = 0 \end{cases}$$

con $\lambda = -20$.

(a) Se ne determini la soluzione analitica. *(Sugg.: variazione delle costanti).*

(b) Si mostri il corretto ordine di convergenza del metodo Runge–Kutta semiimplicito di tableau

$$\begin{array}{c|cc} \frac{3+\sqrt{3}}{6} & \frac{3+\sqrt{3}}{6} & \\ \frac{3-\sqrt{3}}{6} & -\frac{\sqrt{3}}{3} & \frac{3+\sqrt{3}}{6} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

nell'approssimazione di $y(t^*)$, con $t^* = 1$.

54. Si consideri il seguente problema di diffusione-reazione

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \dfrac{1}{20}\dfrac{\partial^2 u}{\partial x^2}(t,x) + u^2(t,x) & t > 0,\ x \in (0,1) \\ \dfrac{\partial u}{\partial x}(t,0) = 0 & t > 0 \\ u(t,1) = 1 & t > 0 \\ u(0,x) = x^2 \end{cases}$$

Usando il metodo delle linee, si mostri il corretto ordine di convergenza del metodo Eulero–Rosenbrock esponenziale per l'approssimazione di $u(t^*, x)$, con $t^* = 1$.

29-09-2014

55. Si risolva il problema differenziale

$$\begin{cases} -\dfrac{\mathrm{d}^2}{\mathrm{d}x^2}u(x) + \dfrac{\mathrm{d}}{\mathrm{d}x}(1-u(x))^2 = 0 \\ u(1) = 0 \\ u(2) = \dfrac{1}{2} \end{cases}$$

e si mostri il corretto ordine di convergenza rispetto alla soluzione analitica $u(x) = 1 - 1/x$.

56. Si risolva il problema differenziale

$$\begin{cases} \dfrac{\partial}{\partial t}u(t,x) = \dfrac{\partial^2}{\partial x^2}u(t,x) - 3u(t,x) + x\sin(t), & t > 0,\ x \in (0,\pi) \\ u(0,x) = \sin\dfrac{x}{2}, & x \in (0,\pi) \\ u(t,0) = 0, & t \geq 0 \\ \dfrac{\partial u}{\partial x}(t,\pi) = 0, & t \geq 0 \end{cases}$$

con il metodo esponenziale punto medio e se ne mostri il corretto ordine di convergenza al tempo finale $t^* = 1$.

Parte 5

Bibliografia

Bibliografia

[1] J. P. Boyd, Chebyshev and Fourier Spectral Methods, DOVER Publications, Inc., 2000. http://www-personal.umich.edu/~jpboyd/BOOK_Spectral2000.html

[2] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, Spectral Methods in Fluid Dynamics, Springer-Verlag, 1986.

[3] V. Comincioli, Analisi numerica: metodi, modelli, applicazioni, McGraw–Hill, 1995.

[4] E. Hairer and G. Wanner, Solving Ordinary Differential Equations II, Stiff and Differential-Algebraic Problems, Springer, Second Revised Edition, 2002.

[5] E. Hairer, S. P. Nørsett, and G. Wanner, Solving Ordinary Differential Equations I, Nonstiff Problems, Springer, Second Revised Edition, 2000.

[6] W. Hundsdorfer, Numerical Solution of Advection-Diffusion-Reaction Equations, Lecture notes, Thomas Stieltjes Institute, 2000. http://homepages.cwi.nl/~willem/Coll_AdvDiffReac/notes.pdf

[7] A. Iserles, A First Course in the Numerical Analysis of Differential Equations, Cambridge Texts in Applied Mathematics, second ed., 2009.

[8] R. J. Leveque, Numerical Methods for Conservation Laws, Lectures in Mathematics, Birkhäuser, 1992.

[9] A. Quarteroni, Modellistica numerica per problemi differenziali, Springer, terza edizione, 2006.