

Dispense di
Metodi Numerici
per le Equazioni Differenziali

Dott. Marco Caliarì

a.a. 2012/13

La versione più aggiornata (al 24 settembre 2012) si trova
all'indirizzo

http://profs.sci.univr.it/~caliari/aa1213/equazioni_differenziali/dispense.pdf

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.

Indice

0	Preliminari	7
1	Interpolazione polinomiale a tratti	8
1.1	Interpolazione lineare a tratti	8
1.1.1	Errore di interpolazione	9
2	Formule di quadratura gaussiana	10
2.1	Quadratura gaussiana di Chebyshev(-Lobatto)	11
3	Metodi iterativi per sistemi lineari	13
3.1	Metodi di Richardson	13
3.1.1	Metodo del gradiente preconditionato	14
3.1.2	Metodo del gradiente coniugato preconditionato	15
3.1.3	Test d'arresto	16
4	Memorizzazione di matrici sparse	17
4.1	Alcuni comandi per matrici sparse	18
5	Sistemi tridiagonali	19
6	Metodo di Newton	21
6.1	Metodo di Newton inesatto	22
7	Esponenziale di matrice	23
7.1	Formula delle <i>variazioni delle costanti</i>	23
7.2	Calcolo di $\exp(A)$	24
7.2.1	Matrici piene, di modeste dimensioni	24
7.2.2	Matrici sparse, di grandi dimensioni	26
8	Esercizi	28

1	BVPs	30
9	Introduzione	31
10	Differenze finite	32
10.1	Differenze finite centrate del secondo ordine	32
10.2	Convergenza per un problema modello	35
10.2.1	Unicità	35
10.2.2	Esistenza	36
10.2.3	Regolarità	36
10.2.4	Esistenza ed unicità per il problema discretizzato . . .	37
10.2.5	Consistenza	38
10.2.6	Proprietà di A	38
10.2.7	Stabilità	38
10.2.8	Convergenza	39
10.3	Altre differenze finite	40
10.3.1	Su nodi non equispaziati	40
10.3.2	Non centrate	41
10.3.3	Di ordine più elevato	41
10.4	Condizioni al bordo	41
10.4.1	Condizioni di Dirichlet	41
10.4.2	Condizioni di Neumann	42
10.4.3	Importanza delle condizioni al bordo	43
10.5	Un esempio: l'equazione della catenaria	44
10.5.1	Iterazioni di punto fisso	44
10.5.2	Metodo di Newton	45
10.6	Norme ed errori	45
11	Metodo di shooting	48
11.1	Metodo di bisezione	48
11.2	Metodo di Newton	49
11.3	Problema ai limiti con frontiera libera	50
12	Equazione di Poisson	52
12.1	Equazione di Poisson bidimensionale	52
12.1.1	Condizioni al bordo di Dirichlet	52
12.1.2	Condizioni al bordo miste	54
13	Metodi variazionali	56
13.1	Un problema modello	56
13.1.1	Metodo di approssimazione variazionale	58

<i>INDICE</i>	5
13.1.2 Estensione al caso bidimensionale	65
13.2 Metodi spettrali	66
13.2.1 Trasformata di Fourier	68
13.2.2 Trasformata di Fourier discreta	69
13.3 Metodi di collocazione	76
13.3.1 Condizioni al bordo	76
14 Esercizi	80
2 ODEs	82
15 Introduzione	83
15.1 Riduzione in forma autonoma	84
15.2 Equazioni di ordine superiore al primo	84
16 Metodi ad un passo	85
16.1 Metodo di Eulero	85
16.2 Metodo dei trapezi	87
16.3 theta-metodo	89
16.3.1 Risoluzione di un metodo implicito	91
16.3.2 Newton inesatto e passo variabile	93
16.3.3 Caso lineare	94
16.3.4 Metodi semiimpliciti	94
16.4 Verifica dell'implementazione	95
17 Metodi multistep	96
17.1 Metodi di Adams-Bashforth	96
17.2 Metodi lineari multistep	98
17.2.1 Metodi BDF	100
17.3 Consistenza e stabilità	102
17.4 Influenza degli errori di arrotondamento	106
18 Metodi di Runge-Kutta	107
18.1 Un esempio di metodo Runge-Kutta implicito	112
18.2 Metodi di Runge-Kutta embedded	114
19 A-stabilità	118
19.1 A-stabilità dei metodi di Runge-Kutta espliciti	120
19.2 A-stabilità dei metodi lineari multistep	122
19.3 Equazioni stiff	122
19.3.1 Risoluzione di un metodo implicito per problemi stiff	124

20	Integratori esponenziali	126
21	Esercizi	129
3	PDEs	132
22	Equazione del calore	133
22.1	Equazione del calore	133
22.1.1	Esistenza di una soluzione	133
22.1.2	Unicità della soluzione	136
22.2	Metodo di Fourier	137
22.3	Metodo delle linee	138
22.3.1	Differenze finite	138
22.3.2	Condizioni al bordo di Dirichlet	139
22.3.3	Condizioni al bordo di Neumann (costanti)	140
22.3.4	Equazione di trasporto-diffusione	141
22.3.5	Elementi finiti	145
22.3.6	Errori spaziali e temporali	147
22.4	Esercizi	147
4	Temî d'esame	149
5	Bibliografia	161
	Bibliografia	162

Parte 0
Preliminari

Capitolo 1

Interpolazione polinomiale a tratti

Data una funzione $f: [a, b] \rightarrow \mathbb{R}$ 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 $k - 1$

$$a_{i,1}(x - x_i)^{k-1} + a_{i,2}(x - x_i)^{k-2} + \dots + a_{i,k-1}(x - x_i) + a_{i,k}. \quad (1.1)$$

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,P)`, ove \mathbf{x} è il vettore di nodi e \mathbf{P} è la matrice, con riferimento a (1.1),

$$P_{ij} = a_{i,j}.$$

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

1.1 Interpolazione lineare a tratti

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

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

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)])
```

1.1.1 Errore di interpolazione

Il risultato fondamentale sull'errore di interpolazione è

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

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

$$f(x) - p_1^c f(x) = \frac{f''(\xi)}{2}(x - x_i)(x - x_{i+1}) \quad (1.2)$$

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

$$\|f - p_1^c f\|_\infty \leq \frac{h^2}{8} \|f''\|_\infty$$

Derivando rispetto a x l'equazione (1.2), 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 2

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

$$(f, g) = \int_a^b f(x)g(x)w(x)dx$$

con l'ipotesi

$$\int_a^b |x|^k w(x)dx < \infty, \quad k \geq 0$$

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

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

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

$$w_n = \int_a^b L_n(x)w(x)dx, \quad 1 \leq n \leq m$$

ove $L_n(x)$ è il polinomio di Lagrange 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

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

In particolare

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

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,

$$\delta_{ij} = \int_a^b p_j(x)p_i(x)w(x)dx = \sum_{n=1}^m p_j(\bar{x}_n)p_i(\bar{x}_n)\bar{w}_n, \quad \begin{array}{l} 0 \leq i \leq m-3 \\ 0 \leq j \leq m-1 \end{array}$$

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

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

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

2.1 Quadratura gaussiana di Chebyshev e di Chebyshev-Lobatto

Per integrali del tipo

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

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

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

Gli zeri del polinomio di grado m soddisfano

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

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

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

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

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

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

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

e i corrispondenti pesi

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

Capitolo 3

Metodi iterativi per sistemi di equazioni lineari

I metodi iterativi per la soluzione del sistema lineare

$$Ax = b \quad (3.1)$$

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

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

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

$$Px^{(l+1)} = Mx^{(l)} + b, \quad l \geq 1 \quad (3.2)$$

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*.

Lemma 1. *Si ha $\lim_{l \rightarrow \infty} e^{(l)} = 0$ per ogni $e^{(1)}$ se e solo se $\lim_{l \rightarrow \infty} B^l = 0$, cioè se e solo se $\rho(B) < 1$.*

3.1 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 (3.2) può essere riscritto come

$$P(x^{(l+1)} - x^{(l)}) = r^{(l)}. \quad (3.3)$$

In questo contesto, P viene chiamata *matrice di preconditionamento* 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 (3.3) sia “facile”.

Una generalizzazione dello schema (3.3) 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

$$\begin{aligned} Pz^{(l)} &= r^{(l)} \\ x^{(l+1)} &= x^{(l)} + \alpha z^{(l)} \\ r^{(l+1)} &= r^{(l)} - \alpha Az^{(l)} \end{aligned} \quad (3.4)$$

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 $r^{(l+1)} = r^{(l)} - \alpha Az^{(l)}$ (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 .

3.1.1 Metodo del gradiente preconditionato

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

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

sia minima. Si ha

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

e dunque il minimo è dato dalla scelta

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

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

$$\begin{aligned} Pz^{(l)} &= r^{(l)} \\ \alpha_l &= \frac{z^{(l)\text{T}} r^{(l)}}{z^{(l)\text{T}} A z^{(l)}} \\ x^{(l+1)} &= x^{(l)} + \alpha_l z^{(l)} \\ r^{(l+1)} &= r^{(l)} - \alpha_l A z^{(l)} \end{aligned} \tag{3.5}$$

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

3.1.2 Metodo del gradiente coniugato preconditionato

Siano A e P simmetriche e definite positive. Il *metodo del gradiente coniugato preconditionato* 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

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

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

$$\begin{aligned} \alpha_l &= \frac{z^{(l)\text{T}} r^{(l)}}{p^{(l)\text{T}} A p^{(l)}} \\ x^{(l+1)} &= x^{(l)} + \alpha_l p^{(l)} \\ r^{(l+1)} &= r^{(l)} - \alpha_l A p^{(l)} \\ Pz^{(l+1)} &= r^{(l+1)} \\ \beta_{l+1} &= \frac{z^{(l+1)\text{T}} r^{(l+1)}}{z^{(l)\text{T}} r^{(l)}} \\ p^{(l+1)} &= z^{(l+1)} + \beta_{l+1} p^{(l)} \end{aligned} \tag{3.6}$$

Teorema 1. *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 $p^{(1)}, \dots, p^{(N)}$ sono vettori linearmente indipendenti e non ce ne possono essere più di N . \square

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 $\|e^{(l)}\|_A \leq 0$.

3.1.3 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

$$\|r^{(l)}\| \leq \text{tol} \cdot \|r^{(1)}\| \tag{3.7}$$

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

Capitolo 4

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

```
data = [1, 4, 2, 3, 5, 6, 7]
ridx = [1, 3, 2, 2, 3, 3, 4]
cidx = [1, 3, 4, 6, 8]
```

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.

4.1 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 5

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 6

Metodo di Newton per sistemi di equazioni non lineari

Consideriamo il sistema di equazioni non lineari

$$\begin{cases} f_1(x_1, x_2, \dots, x_N) = 0 \\ f_2(x_1, x_2, \dots, x_N) = 0 \\ \vdots \\ f_N(x_1, x_2, \dots, x_N) = 0 \end{cases}$$

che può essere riscritto, in forma compatta,

$$f(x) = 0 .$$

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

$$\begin{aligned} J^{(l)} \delta x^{(l)} &= -f(x^{(l)}) \\ x^{(l+1)} &= x^{(l)} + \delta x^{(l)} \end{aligned} \tag{6.1}$$

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

$$J_{ij}^{(l)} = \frac{\partial f_i(x^{(l)})}{\partial x_j^{(l)}} . \tag{6.2}$$

Il criterio d'arresto solitamente usato è

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

6.1 Metodo di Newton inesatto

Il metodo di Newton (6.1) richiede il calcolo della matrice Jacobiana e la sua “inversione” ad ogni passo k . Questo potrebbe essere troppo oneroso. 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. In tal modo, per esempio, è possibile usare la stessa fattorizzazione $L^{(l)}U^{(l)}$ per più iterazioni successive.

Capitolo 7

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

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

7.1 Formula delle *variazioni delle costanti*

Data l'equazione differenziale

$$\begin{cases} y'(t) = ay(t) + b(t, y(t)), & t > 0 \\ y(t_0) = y_0 \end{cases} \quad (7.1)$$

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

$$y(t) = e^{(t-t_0)a}y_0 + \int_{t_0}^t e^{(t-\tau)a}b(\tau, y(\tau))d\tau \quad (7.2)$$

Infatti, si ha

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

Si osservi che

$$\begin{aligned} \int_{t_0}^t e^{(t-\tau)a} d\tau &= -\frac{1}{a} \int_{t_0}^t -ae^{(t-\tau)a} d\tau = -\frac{1}{a} e^{(t-\tau)a} \Big|_{t_0}^t = \\ &= -\frac{1}{a} (1 - e^{(t-t_0)a}) = (t-t_0) \frac{e^{(t-t_0)a} - 1}{(t-t_0)a} = \\ &= (t-t_0) \varphi_1((t-t_0)a), \end{aligned}$$

ove

$$\varphi_1(z) = \frac{e^z - 1}{z} = \sum_{j=0}^{\infty} \frac{z^j}{(j+1)!} \quad (7.3)$$

e, analogamente,

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

ove

$$\varphi_2(z) = \frac{e^z - 1 - z}{z^2} = \sum_{j=0}^{\infty} \frac{z^j}{(j+2)!} \quad (7.4)$$

Consideriamo ora un sistema differenziale

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

Ancora, la soluzione esplicita può essere scritta come

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

7.2 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.

7.2.1 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 $e^{d_1}, e^{d_2}, \dots, e^{d_N}$. 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

$$e^z \approx \frac{a_1 z^{p-1} + a_2 z^{p-2} + \dots + a_p}{b_1 z^{q-1} + b_2 z^{q-2} + \dots + b_q}, \quad (7.5)$$

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

$$\begin{aligned} \left(1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \dots\right) (b_1 z + 1) &= a_1 z + a_2 \\ b_1 z + 1 + b_1 z^2 + z + \frac{z^2}{2} + o(z^2) &= a_1 z + a_2 \end{aligned}$$

da cui

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

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 $(b_1 A + I)B = a_1 A + a_2 I$.

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

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

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`.

7.2.2 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 $V_m^T A V_m = H_m$, ove $V_m \in \mathbb{R}^{n \times m}$, $V_m^T V_m = I_n$, $V_m e_1 = v$ 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 $\xi_0, \xi_1, \dots, \xi_m$, 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

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

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 8

Esercizi

1. Implementare 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 $A^T x$.
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 $[1, 1, \dots, 1]^T$ usando il comando `\` di GNU Octave, il metodo del gradiente preconditionato e il metodo del gradiente coniugato preconditionato. Per questi ultimi due, si usi una tolleranza pari a 10^{-6} , un numero massimo di iterazioni pari a 2000, il preconditionatore 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 (6.1). Si usi una tolleranza pari a 10^{-6} , un numero massimo di iterazioni pari a 150 e un vettore iniziale $x^{(1)} = [1, 1]^T$. 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 coef-

ficienti dell'approssimazione razionale di Padé (7.5) (con $p = q$) per la funzione esponenziale.

Parte 1
BVPs
(Problemi ai limiti)

Capitolo 9

Introduzione

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

$$\begin{cases} u''(x) = f(x, u(x), u'(x)), & x \in (a, b) \\ u(a) = u_a \\ u(b) = u_b \end{cases} \quad (9.1)$$

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

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

Quando i valori prescritti ai bordi sono nulli, si parla di condizioni *omogenee*.

Capitolo 10

Differenze finite

10.1 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

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

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

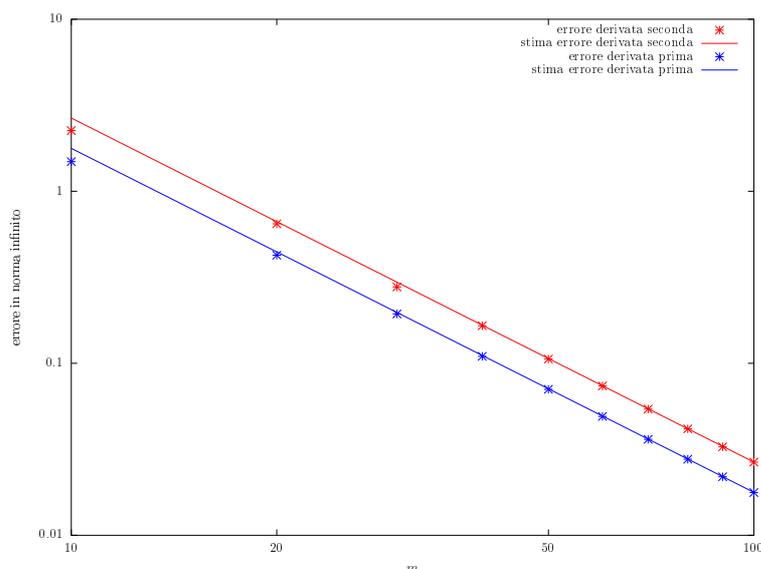
$$\begin{aligned}u(x_{i+1}) &= u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{6}u^{(3)}(x_i) + \frac{h^4}{24}u^{(4)}(\hat{x}_i) \\u(x_{i-1}) &= u(x_i) - hu'(x_i) + \frac{h^2}{2}u''(x_i) - \frac{h^3}{6}u^{(3)}(x_i) + \frac{h^4}{24}u^{(4)}(\tilde{x}_i)\end{aligned}$$

da cui

$$u''(x_i) = \Delta^2 u(x_i) - \tau_i^{(2)} = \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1}))}{h^2} - \tau_i^{(2)} \quad (10.1)$$

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.



diff12.m

Figura 10.1: Errori nell'approssimazione delle derivate prima e seconda per la funzione $u(x) = \sin(3x)$.

In Figura 10.1 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 10.2 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 (9.1) si risolve il problema discretizzato

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

diff12ns.m

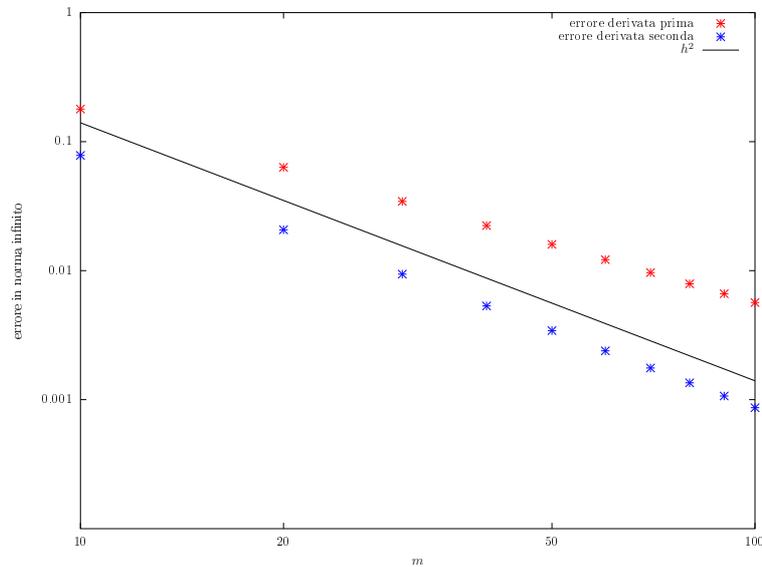


Figura 10.2: Errori nell'approssimazione delle derivate prima e seconda per la funzione $u(x) = |x|^{7/2}$.

nell'incognita $\mathbf{u} = [u_1, u_2, \dots, u_{m-1}, u_m]^T$, 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,

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

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.

10.2 Convergenza per un problema modello

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

$$\begin{cases} -u''(x) + q(x)u(x) = g(x), & x \in (a, b) \\ u(a) = u_a \\ u(b) = u_b \end{cases} \quad (10.2)$$

con $q, g \in C^0([a, b])$, $q(x) \geq 0$ per $x \in [a, b]$. La funzione $q(x)$ dipende dal materiale di cui è fatta la trave e $g(x)$ è la densità di carico trasversale. La soluzione $u(x)$ rappresenta il momento flettente. Vogliamo studiare l'esistenza, l'unicità e la regolarità della soluzione analitica.

10.2.1 Unicità

Se $u_1(x)$ e $u_2(x)$ sono due soluzioni di (10.2), allora $z(x) = u_1(x) - u_2(x)$ soddisfa il problema *omogeneo*

$$\begin{cases} -z''(x) + q(x)z(x) = 0, & x \in (a, b) \\ z(a) = 0 \\ z(b) = 0 \end{cases} \quad (10.3)$$

Proposizione 1. *Se $z(x)$ è soluzione di (10.3), allora $z(x) \equiv 0$.*

Dimostrazione (metodo dell'energia). Moltiplicando l'equazione per $z(x)$ ed integrando si ha

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

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

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

10.2.2 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 (10.3) (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 (10.2) si ottiene imponendo

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

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

10.2.3 Regolarità

Proposizione 2. *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 $u \in \mathcal{C}^{k+3}([a, b])$. \square

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

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

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

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

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

10.2.4 Esistenza ed unicità per il problema discretizzato

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

$$\frac{1}{h^2} \begin{bmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ -1 & 2 + q_2 h^2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 + q_3 h^2 & -1 & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -1 & 2 + q_{m-1} h^2 & -1 \\ 0 & \dots & \dots & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{m-1} \\ u_m \end{bmatrix} = \begin{bmatrix} u_a/h^2 \\ g_2 \\ g_3 \\ \vdots \\ g_{m-1} \\ u_b/h^2 \end{bmatrix}$$

e può essere semplificato in

$$\frac{1}{h^2} \begin{bmatrix} 2 + q_2 h^2 & -1 & 0 & \dots & \dots & 0 \\ -1 & 2 + q_3 h^2 & -1 & 0 & \dots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \dots & 0 & -1 & 2 + q_{m-2} h^2 & -1 \\ 0 & \dots & \dots & 0 & -1 & 2 + q_{m-1} h^2 \end{bmatrix} \begin{bmatrix} u_2 \\ u_3 \\ \vdots \\ \vdots \\ u_{m-2} \\ u_{m-1} \end{bmatrix} = \begin{bmatrix} g_2 + u_a/h^2 \\ g_3 \\ \vdots \\ \vdots \\ g_{m-2} \\ g_{m-1} + u_b/h^2 \end{bmatrix}$$

cioè

$$A\mathbf{u} = \mathbf{g} \tag{10.4}$$

ove adesso $\mathbf{u} = [u_2, \dots, u_{m-1}]^T$.

Proposizione 3. *Il sistema lineare (10.4) è non singolare e dunque ammette un'unica soluzione.*

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

$$\begin{aligned} \mathbf{z}^T A \mathbf{z} &= \frac{1}{h^2} [(2 + q_2 h^2) z_2^2 - z_2 z_3 - z_3 z_2 + (2 + q_3 h^2) z_3^2 - z_3 z_4 + \dots + \\ &\quad + \dots - z_{m-1} z_{m-2} + (2 + q_{m-1} h^2) z_{m-1}^2] = \\ &= \frac{1}{h^2} [z_2^2 + (z_2 - z_3)^2 + (z_3 - z_4)^2 + \dots + (z_{m-2} - z_{m-1})^2 + z_{m-1}^2] + \\ &\quad + \sum_{i=2}^{m-1} q_i z_i^2 \geq 0 \end{aligned}$$

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 \mathbf{z} nullo. Dunque la matrice A è definita positiva e quindi non singolare. \square

10.2.5 Consistenza

Se si sostituisce u_i con la soluzione analitica $u(x_i)$, da (10.1) si ottiene

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

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

10.2.6 Proprietà di A

A è una matrice simmetrica 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.

10.2.7 Stabilità

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

$$\begin{aligned} A\tilde{\mathbf{u}} &= \tilde{\mathbf{g}} \\ A\bar{\mathbf{u}} &= \bar{\mathbf{g}} \end{aligned}$$

da cui

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

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 $h \rightarrow 0$. Consideriamo la matrice $A_{q=0}$ corrispondente alla stessa discretizzazione nel caso $q(x) \equiv 0$. Si ha $A - A_{q=0} = \text{diag}(q_2, \dots, q_{m-1}) \geq 0$. 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 $A^{-1} \leq A_{q=0}^{-1}$. Osserviamo poi che $\mathbf{v} = A_{q=0}^{-1}[1, \dots, 1]^T$ 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 $v^{(3)}(x) \equiv 0$, così è per $v^{(4)}$ e dunque l'errore locale, per questo problema, è nullo. Dunque

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

e poiché $\|A^{-1}\|_{\infty} \leq \|A_{q=0}^{-1}\|_{\infty}$, si ha la maggiorazione richiesta.

10.2.8 Convergenza

Definiamo $\mathbf{e}_h = [e_{2,h}, \dots, e_{m-1,h}]^T = [u_2 - u(x_2), \dots, u_{m-1} - u(x_{m-1})]^T$, $h = (b-a)/(m-1)$. Poiché

$$\begin{aligned} A[u_2, \dots, u_{m-1}]^T &= \mathbf{g} \\ A[u(x_2), \dots, u(x_{m-1})]^T &= \mathbf{g} - \boldsymbol{\tau}_h^{(2)} \end{aligned}$$

ove $\boldsymbol{\tau}_h^{(2)} = [\tau_{2,h}^{(2)}, \dots, \tau_{m-1,h}^{(2)}]^T$, si deduce $\mathbf{e}_h = A^{-1}\boldsymbol{\tau}_h^{(2)}$. Combinando i risultati di *consistenza* e *stabilità*, si ottiene, per il problema (10.2) discretizzato mediante differenze finite centrate del secondo ordine,

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

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

10.3 Altre differenze finite

10.3.1 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

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

da cui

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

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 conto 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 \mathbf{x} (vettore colonna di lunghezza m), è possibile costruire il vettore $[h_1, h_2, \dots, h_{m-1}]^T$ 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.

10.3.2 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,

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

Ciò può risultare utile per l'approssimazione ai bordi.

10.3.3 Di ordine più elevato

Si possono per esempio costruire differenze finite di ordine quattro centrate

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

10.4 Condizioni al bordo

10.4.1 Condizioni di Dirichlet

Se vengono prescritti i valori $u(a) = u_a$ o $u(b) = u_b$, 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

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

Poi, si correggono le equazioni relative ai nodi al bordo

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

In questo modo, però, la simmetria della matrice 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

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

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

10.4.2 Condizioni di Neumann

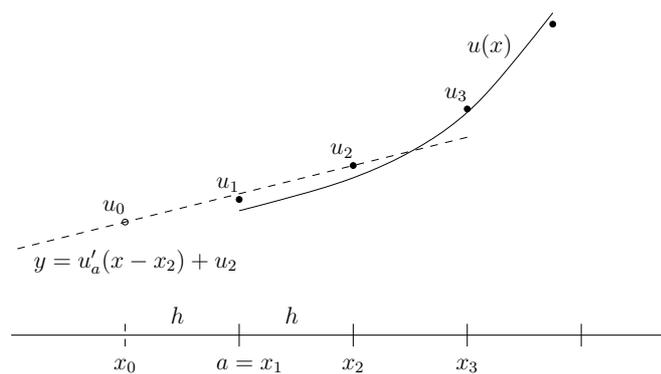


Figura 10.3: Imposizione di una condizione di Neumann sull'estremo sinistro.

Un modo per imporre le condizioni di Neumann, per esempio $u'(a) = u'_a$, e conservare l'ordine di accuratezza due (e che si può usare anche con i problemi

parabolici, vedi paragrafo 22.3.3) è quella di introdurre una variabile fittizia $u_0 \approx u(a - h)$, approssimare la derivata $u'(a)$ con

$$u'(a) \approx \frac{u_2 - u_0}{2h}$$

e porla uguale a u'_a . In tal modo, risulta

$$u_0 = u_2 - 2hu'_a$$

Tale espressione per u_0 va usata poi in qualunque stencil di discretizzazione. Per esempio, la discretizzazione dell'equazione

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

in $x = a$ sarà

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

da cui

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

10.4.3 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

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

munita delle condizioni al bordo

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

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

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

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

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

(nessuna soluzione) oppure

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

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

10.5 Un esempio: l'equazione della catenaria

Consideriamo l'equazione della *catenaria*

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

La discretizzazione mediante differenze finite centrate del secondo ordine è

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

Si tratta dunque di risolvere il sistema non lineare

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

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 $\mathbf{b} = [1/h^2 - \alpha, 0, \dots, 0, 1/h^2 - \alpha]^T$.

10.5.1 Iterazioni di punto fisso

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

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

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.

10.5.2 Metodo di Newton

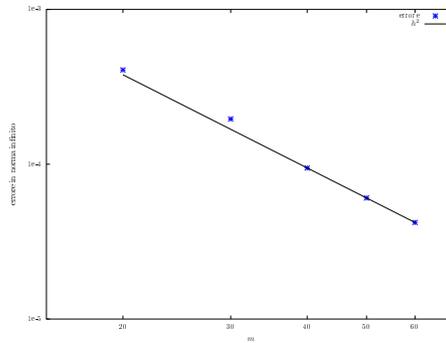
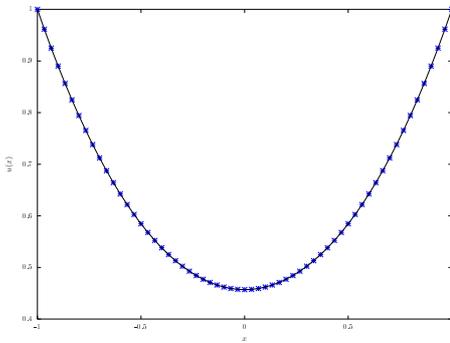
Volendo applicare il metodo di Newton (a convergenza quadratica) è necessario calcolare lo jacobiano di $F(\mathbf{u})$, che risulta essere

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

ove

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

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



catenaria.m

Figura 10.4: Soluzione dell'equazione della catenaria (sinistra, $m = 60$) e ordine di convergenza (destra).

del sistema lineare

$$J_F(\mathbf{u}^{(r)}) (\mathbf{u}^{(r+1)} - \mathbf{u}^{(r)}) = -F(\mathbf{u}^{(r)})$$

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

$$J_F(\mathbf{u}^{(r)})\mathbf{v} = \frac{dF}{d\mathbf{u}}(\mathbf{u}^{(r)})\mathbf{v} = \lim_{\varepsilon \rightarrow 0} \frac{F(\mathbf{u}^{(r)} + \varepsilon\mathbf{v}) - F(\mathbf{u}^{(r)})}{\varepsilon}$$

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

10.6 Norme ed errori

Data una funzione $u(x)$ e due diverse discretizzazioni su nodi equispaziati $[\tilde{u}_1, \dots, \tilde{u}_m] \approx [u(\tilde{x}_1), \dots, u(\tilde{x}_m)]$ e $[\hat{u}_1, \dots, \hat{u}_l] \approx [u(\hat{x}_1), \dots, u(\hat{x}_l)]$, $\{\tilde{x}_i\}_i \subset$

$[a, b]$, $\{\hat{x}_i\}_i \subset [a, b]$, non ha molto senso confrontare gli errori $\| [u(\hat{x}_1) - \tilde{u}_1, u(\hat{x}_2) - \tilde{u}_2, \dots, u(\hat{x}_m) - \tilde{u}_m] \|_2$ e $\| [u(\hat{x}_1) - \hat{u}_1, y(\hat{x}_2) - \hat{u}_2, \dots, u(\hat{x}_i) - \hat{u}_i] \|_2$. 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 $u(x)$.

fsc.m

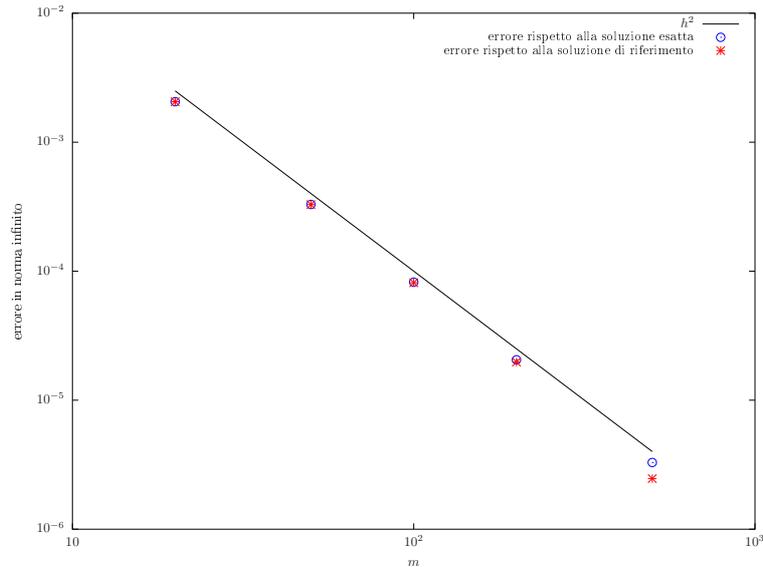


Figura 10.5: Convergenza e *falsa superconvergenza* per la risoluzione di $u''(x) = -\sin(x)$, $u(0) = u(\pi) = 0$.

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 10.5). Se si calcola una soluzione di riferimento con \bar{m} punti di discretizzazione, si ha

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

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 $\|u_m - u_{\bar{m}}\|_\infty \lesssim 2\varepsilon$. 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 $\mathbf{e}_{\tilde{m}}$ e $\mathbf{e}_{\hat{m}}$ gli errori corrispondenti a due discretizzazioni con \tilde{m} e \hat{m} punti. Supponiamo che

$$\|\mathbf{e}_{\tilde{m}}\|_{\infty} = \frac{C}{(\tilde{m} - 1)^p}$$
$$\|\mathbf{e}_{\hat{m}}\|_{\infty} = \frac{C}{(\hat{m} - 1)^p}$$

Si ricava

$$\log\|\mathbf{e}_{\tilde{m}}\|_{\infty} - \log\|\mathbf{e}_{\hat{m}}\|_{\infty} = -p(\log(\tilde{m} - 1) - \log(\hat{m} - 1))$$

Dunque, in un grafico *logaritmico-logaritmico*, l'errore $\|\mathbf{e}_m\|_{\infty}$ si dispone su una retta di pendenza $-p$ (cioè parallelo alla "retta" $(m - 1)^{-p}$) rispetto a $m - 1$.

Capitolo 11

Metodo di shooting

È possibile trasformare il problema (9.1) in un sistema differenziale del primo ordine

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

tramite il cambiamento di variabili $t = x$, $y_1(t) = u(x)$, $y_2(t) = u'(x)$, $\mathbf{f}(t, \mathbf{y}(t)) = [y_2(t), f(t, y_1(t), y_2(t))]^T$. 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

$$\begin{cases} \mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)), & t \in (a, b] \\ y_1(a) = u_a \\ y_2(a) = s \end{cases} \quad (11.1)$$

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) $F(s) = 0$.

11.1 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 (11.1) 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.

11.2 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

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

(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

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

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

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

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.

11.3 Problema ai limiti con frontiera libera

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

$$\begin{cases} u''(x) = f(x, u(x), u'(x)), & x \in (s, b) \\ u(s) = \alpha \\ u'(s) = \beta \\ u(b) = u_b \end{cases} \quad (11.2)$$

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

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

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

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

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

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

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

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

ove il valore $u''(s)$ si ricava dal problema (11.2) e vale $f(s, \alpha, \beta)$.

Capitolo 12

Equazione di Poisson

Di particolare interesse è l'equazione di Poisson

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

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.

12.1 Equazione di Poisson bidimensionale

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

12.1.1 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 $x_i = a + (i - 1)h_x$, $i = 1, 2, \dots, m_x$, $h_x = (b - a)/(m_x - 1)$

e $y_j = c + (j - 1)h_y$, $j = 1, 2, \dots, m_y$, $h_y = (d - c)/(m_y - 1)$. 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 12.1. La matrice di discretizzazio-

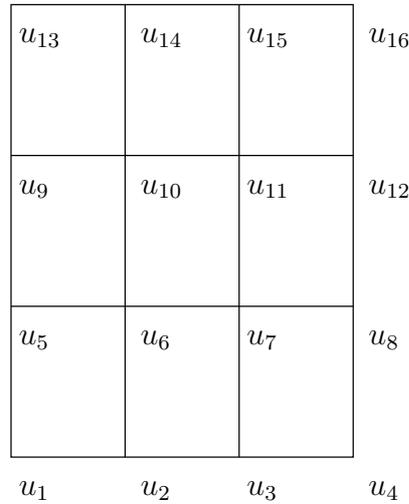


Figura 12.1: Numerazione di una griglia bidimensionale

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

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

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 $[b_1, b_2, \dots, b_{m_x m_y}]^T$, 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`.

12.1.2 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

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

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 $[b_1, b_2, \dots, b_{m_x m_y}]^T$, 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 13

Metodi variazionali

13.1 Formulazione variazionale di un problema modello

Consideriamo il seguente problema ai limiti (equazione di Poisson)

$$\begin{cases} -u''(x) = g(x), & x \in (0, 1) \\ u(0) = u(1) = 0 \end{cases} \quad (13.1)$$

con $g \in C^0([0, 1])$. Introduciamo il seguente spazio lineare:

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

e il prodotto scalare su V

$$(v, w) = \int_0^1 v(x)w(x)dx$$

Teorema 2 (Formulazione variazionale). *Se $u(x)$ è la soluzione del problema (13.1), allora $u \in V$ e*

$$(u', v') = (g, v), \quad \forall v \in V \quad (13.2)$$

Dimostrazione. Sia u soluzione di (13.1). Allora, per ogni $v \in V$,

$$\int_0^1 -u''(x)v(x)dx = \int_0^1 g(x)v(x)dx = (g, v)$$

Integrando per parti,

$$\int_0^1 -u''(x)v(x)dx = -u'(x)v(x)\Big|_0^1 + \int_0^1 u'(x)v'(x)dx = (u', v')$$

poiché $v(0) = v(1) = 0$. □

Per quanto visto per il problema modello (10.2), la soluzione di (13.1) esiste ed è unica. Per quanto appena dimostrato, essa è soluzione anche di (13.2).

A cosa serve dunque la formulazione variazionale? Consideriamo invece il seguente problema

$$\begin{cases} -u''(x) = -g_\varepsilon(x), & x \in (0, 1) \\ u(0) = u(1) = 0 \end{cases} \quad (13.3)$$

ove

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

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

$$\int_0^1 g_\varepsilon(x) dx = 1$$

La “soluzione” di (13.3) è

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

In che senso soluzione? Chiaramente $u_\varepsilon''(1/2 \pm \varepsilon)$ non esiste e quindi non è vero che $-u_\varepsilon''(x) = g_\varepsilon(x)$, $x \in (0, 1)$. Ma è vero che $u_\varepsilon \in V$ e

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

Dunque u_ε è soluzione della formulazione variazionale.

Ritornando al problema modello (13.1), ove però non facciamo ipotesi su g , la soluzione classica, quella di classe (almeno) \mathcal{C}^2 , si chiama *soluzione forte* del problema (13.1), mentre la soluzione di (13.2) si chiama *soluzione*

debole del problema (13.1). 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). 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 (13.2) e $u \in \mathcal{C}^2([0, 1])$ (da notare che $\mathcal{C}^2([0, 1]) \subset V$), allora $0 = (u' - g, v) = (-u'' - g, v)$ 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 (13.2) del problema (13.1) è in realtà la 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 (13.2), allora

$$(u'_1 - u'_2, v') = 0, \quad \forall v \in V$$

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

$$\int_0^1 (u'_1(x) - u'_2(x))^2 dx = 0$$

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 $u_1(x) - u_2(x) = 0$.

13.1.1 Metodo di approssimazione variazionale

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

$$(\hat{u}, v)' = (g, v), \quad \forall v \in V_m \tag{13.4}$$

(*metodo di Galerkin*) ove, per brevità, $(\hat{u}, v)' = (\hat{u}', v')$, o, equivalentemente

$$J(\hat{u}) = \inf_{v \in V_m} J(v)$$

(*metodo di Ritz*).

Teorema 3. *Il problema (13.4) ha un'unica soluzione.*

Dimostrazione. Sia $\{\phi_j\}_{j=1}^m$ una base di V_m . Allora

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

e il problema (13.4) si riscrive, per $i = 1, 2, \dots, m$,

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

ove $A = (a_{ij}) = (\phi_j, \phi_i)'$ e $\mathbf{u} = [\hat{u}_1, \dots, \hat{u}_m]^T$. Calcoliamo ora $\mathbf{w}^T A \mathbf{w}$ per $\mathbf{w} = [w_1, \dots, w_m]^T$. Si ha

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

da cui, per la linearità del prodotto scalare,

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

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. \square

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

Teorema 4. *Se u è soluzione di (13.2) e \hat{u} di (13.4), allora*

$$\|u - \hat{u}\|' \leq \inf_{v \in V_m} \|u - v\|' \quad (13.5)$$

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

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. \square

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

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 rigidità e di risolvere il sistema lineare.

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 (13.4), allora

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

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

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

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, *supponendo* 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)

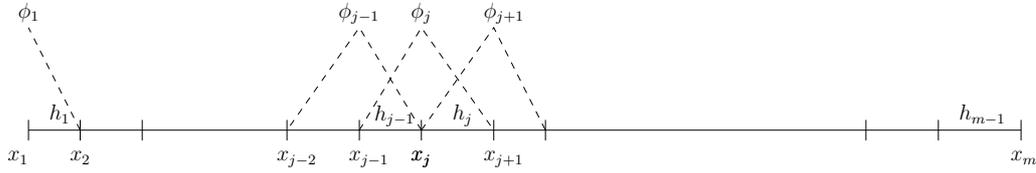


Figura 13.1: Hat functions

Introduciamo una discretizzazione dell'intervallo $[0, 1]$ a passo *variabile*, come in Figura 13.1. Lo spazio V_m è generato dalle funzioni di base $\{\phi_j\}_{j=2}^{m-1}$, le quali sono definite da

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

e

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

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

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

e

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

e

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

e

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

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 (13.4) si riscrive

$$\begin{aligned} \int_0^1 \hat{u}'(x) \phi'_i(x) dx &= \sum_{j=1}^m \hat{u}_j \int_0^1 \phi'_j(x) \phi'_i(x) dx = \sum_{j=1}^m \hat{u}_j \int_{x_i-h_{i-1}}^{x_i+h_i} \phi'_j(x) \phi'_i(x) dx = \\ &= \sum_{j=1}^m \hat{u}_j a_{ij} = \int_{x_i-h_{i-1}}^{x_i+h_i} g(x) \phi_i(x) dx \end{aligned}$$

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 $1 < i < m$,

$$\begin{aligned} a_{ii} &= (\phi_i, \phi_i)' = \int_{x_i-h_{i-1}}^{x_i} \left(\frac{1}{h_{i-1}}\right)^2 dx + \int_{x_i}^{x_i+h_i} \left(-\frac{1}{h_i}\right)^2 dx = \frac{1}{h_{i-1}} + \frac{1}{h_i} \\ a_{i,i-1} &= (\phi_{i-1}, \phi_i)' = \int_{x_i-h_{i-1}}^{x_i} -\frac{1}{h_{i-1}} \cdot \frac{1}{h_{i-1}} dx = -\frac{1}{h_{i-1}} = a_{i-1,i} \end{aligned}$$

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

$$\begin{aligned} a_{11} &= \int_{x_1}^{x_1+h_1} \left(-\frac{1}{h_1}\right)^2 dx = \frac{1}{h_1} \\ a_{21} &= \int_{x_2-h_1}^{x_2} -\frac{1}{h_1} \cdot \frac{1}{h_1} dx = -\frac{1}{h_1} = a_{12} \\ a_{mm-1} &= \int_{x_m-h_{m-1}}^{x_m} -\frac{1}{h_{m-1}} \cdot \frac{1}{h_{m-1}} dx = -\frac{1}{h_{m-1}} = a_{m-1,m} \\ a_{mm} &= \int_{x_m-h_{m-1}}^{x_m} \left(-\frac{1}{h_{m-1}}\right)^2 dx = \frac{1}{h_{m-1}} \end{aligned}$$

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

$$\begin{aligned} g_i &= (g, \phi_i) = \int_{x_{i-1}}^{x_i} g(x) \frac{x-x_{i-1}}{h_{i-1}} dx + \int_{x_i}^{x_{i+1}} g(x) \frac{x_{i+1}-x}{h_i} dx \approx \\ &\approx g\left(\frac{x_{i-1}+x_i}{2}\right) \frac{h_{i-1}}{2} + g\left(\frac{x_i+x_{i+1}}{2}\right) \frac{h_i}{2} \end{aligned}$$

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

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

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

L'approssimazione di

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

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,

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

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

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

per $1 < i < m$ e da

$$g_1 = (g, \phi_1) = \bar{g}_1 \int_{x_1}^{x_1+h_1} \phi_1(x) dx = \bar{g}_1 \frac{h_1}{2}$$

$$g_m = (g, \phi_m) = \bar{g}_{m-1} \int_{x_m-h_{m-1}}^{x_m} \phi_m(x) dx = \bar{g}_{m-1} \frac{h_{m-1}}{2}$$

La riga i -esima del sistema lineare risulta dunque essere

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

e dunque *molto simile* (il termine noto è diverso, anche se dello stesso ordine) a quella della discretizzazione con differenze finite del secondo ordine. 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). Facilmente, usando la disuguaglianza (13.5) e i risultati della Sezione 1.1.1, si può dimostrare

$$\|u - \hat{u}\|' \leq \|u - p_1^c u\|' = \sqrt{\int_0^1 [u'(x) - (p_1^c u)'(x)]^2 dx} \leq h \|u''\|_\infty$$

E quindi, ancora, ci si aspetta che se l'errore sulle derivate va a 0 come h , allora l'errore sulla funzione vada a 0 come h^2 . 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.

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

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

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

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

Dunque, la prima riga del sistema lineare da risolvere è

$$\int_0^1 \hat{u}'(x)\phi_1'(x)dx = -u'_0 + \int_0^1 g(x)\phi_1(x)dx$$

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

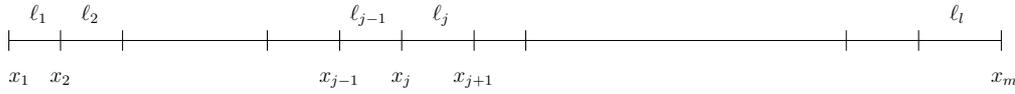


Figura 13.2: Nodi (numerati in basso) ed elementi (numerati in alto).

riferimento alla Figura 13.2, ove $m = l + 1$, si ha

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

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

$$\begin{aligned} \phi_{\ell_j, k \ell_j, k} &= \int_{\ell_j} \phi'_{\ell_j, k}(x) \phi'_{\ell_j, k}(x) dx = \frac{1}{h_j}, \quad k = 1, 2 \\ \phi_{\ell_j, k \ell_j, 3-k} &= \int_{\ell_j} \phi'_{\ell_j, k}(x) \phi'_{\ell_j, 3-k}(x) dx = -\frac{1}{h_j}, \quad k = 1, 2 \end{aligned}$$

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

- $a_{ij} = 0, 1 \leq i, j \leq m, g_i = 0, 1 \leq i \leq m$
- FOR $j = 1, \dots, l$

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

13.1.2 Estensione al caso bidimensionale

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

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

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

13.2 Metodi spettrali

Sia

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

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

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

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 un sistema $\{\phi_j\}_j$ ortonormale rispetto al prodotto scalare

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

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

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

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

$$\sum_{j=1}^m \hat{u}_j \left(\sum_{n=1}^m L\phi_j(x_n) \phi_i(x_n) w_n \right) = \sum_{n=1}^m g(x_n) \phi_i(x_n) w_n, \quad 1 \leq i \leq m \quad (13.6)$$

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

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

i coefficienti \hat{u}_j dati da

$$\hat{u}_j = \sum_{n=1}^m u(x_n) \phi_j(x_n) w_n \approx \int_a^b u(x) \phi_j(x) w(x) dx = u_j, \quad 1 \leq j \leq m$$

si calcolano mediante il prodotto matrice-vettore

$$\begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_m \end{bmatrix} = F \begin{bmatrix} u(x_1)w_1 \\ u(x_2)w_2 \\ \vdots \\ u(x_m)w_m \end{bmatrix}$$

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

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

Solitamente le funzioni $\{\phi_j(x)\}_j$ sono polinomi ortonormali rispetto alla funzione peso $w(x)$. 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

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

Ad ogni modo, la differenza sostanziale 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.

13.2.1 Trasformata di Fourier

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

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

Allora,

$$\int_a^b \phi_j(x) \overline{\phi_k(x)} dx = \delta_{jk} \quad (13.7)$$

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

$$\phi_j(x) \overline{\phi_k(x)} = \frac{e^{i2\pi(j-k)(x-a)/(b-a)}}{b-a}$$

e quindi

$$\int_a^b \phi_j(x) \overline{\phi_k(x)} dx = \int_0^1 \frac{e^{i2\pi(j-k)y}}{b-a} (b-a) dy = 0,$$

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

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

Un risultato utile è il seguente

$$\sum_{n=1}^m e^{i(n-1)2\pi(j-k)/m} = m\delta_{jk}, \quad 1 \leq j, k \leq m \quad (13.8)$$

È ovvio per $j = k$; altrimenti

$$\begin{aligned} \sum_{n=1}^m e^{i(n-1)2\pi(j-k)/m} &= \sum_{n=0}^{m-1} (e^{i2\pi(j-k)/m})^n = \\ &= \frac{1 - e^{i2\pi(j-k)}}{1 - e^{i2\pi(j-k)/m}} = \frac{1 - \cos(2\pi(j-k))}{1 - e^{i2\pi(j-k)/m}} = 0 \end{aligned}$$

poiché $-m+1 \leq j-k \leq m-1$.

13.2.2 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

$$u(x) = \sum_{j=-\infty}^{+\infty} u_j \phi_j(x), \quad u_j \in \mathbb{C} \quad (13.9)$$

Fissato $k \in \mathbb{Z}$, moltiplicando entrambi i membri per $\overline{\phi_k(x)}$ e integrando nell'intervallo $[a, b]$, usando (13.7) si ottiene

$$\begin{aligned} \int_a^b u(x) \overline{\phi_k(x)} dx &= \int_a^b \left(\sum_{j=-\infty}^{+\infty} u_j \phi_j(x) \overline{\phi_k(x)} \right) dx = \\ &= \sum_{j=-\infty}^{\infty} u_j \int_a^b \phi_j(x) \overline{\phi_k(x)} dx = u_k \end{aligned} \quad (13.10)$$

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

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

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

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

Stimiamo adesso u_j : posto $\lambda_j = i(j-1-m/2)2\pi/(b-a)$ si ha, per funzioni di classe $u \in C^1$ (e u' derivabile), integrando per parti

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

Se anche $u'(a) = u'(b)$ e $u' \in \mathcal{C}^1$ (e u'' derivabile), allora, integrando ancora per parti, si ottiene $u_j = \mathcal{O}((j-1-m/2)^{-3})$ e così via. Se dunque u è infinitamente derivabile e *periodica* (cioè tutte le derivate sono periodiche), allora u_j decade più velocemente di ogni potenza negativa di j .

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 $y_n = (n-1)/m$, $n = 1, \dots, m+1$ per funzioni periodiche:

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

Usando la (13.8), abbiamo

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

cioè la famiglia $\{\phi_j\}_j$, $1 \leq j \leq m$, è ortonormale anche rispetto al prodotto scalare *discreto*

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

Questo significa che la formula di quadratura trapezoidale a m punti è esatta per le funzioni $\{\phi_j\}_{j=2-m/2}^{m+m/2}$. Applicando la formula di quadratura ai coefficienti (13.10) si ottiene

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

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

La funzione (*serie troncata di Fourier*)

$$\begin{aligned}\hat{u}(x) &= \sum_{j=1}^m \hat{u}_j \phi_j(x) = \sum_{k=-m/2}^{m/2-1} \hat{u}_{k+1+m/2} \phi_{k+1+m/2}(x) = \\ &= \sum_{k=-m/2}^{m/2-1} \hat{u}_{k+1+m/2} \frac{e^{ik2\pi(x-a)/(b-a)}}{\sqrt{b-a}}\end{aligned}$$

è un polinomio trigonometrico che approssima $u(x)$ ed è *interpolante* nei nodi x_n . Infatti, usando (13.8),

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

Si può far vedere poi che

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

La trasformazione

$$[u(x_1), u(x_2), \dots, u(x_m)]^T \rightarrow [\hat{u}_1, \hat{u}_2, \dots, \hat{u}_m]^T$$

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

$$F = (f_{jk}), \quad f_{jk} = e^{-i(j-1)2\pi y_k} = \overline{\phi_{j+m/2}\left(\frac{x_k-a}{b-a}\right)} \sqrt{b-a}.$$

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

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

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

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

$$\begin{aligned} \hat{\hat{v}}_n &= \sum_{j=1}^m \hat{v}_j \phi_j(x_n) = \sum_{j=1}^m \hat{v}_j \frac{e^{i(j-1-m/2)2\pi(x_n-a)/(b-a)}}{\sqrt{b-a}} = \\ &= \frac{m}{\sqrt{b-a}} \frac{1}{m} \left(\sum_{j=1}^m \hat{v}_j e^{i(j-1)2\pi y_n} \right) e^{-im\pi y_n}. \end{aligned}$$

La trasformazione

$$[\hat{v}_1, \hat{v}_2, \dots, \hat{v}_m]^T \rightarrow [\hat{\hat{v}}_1, \hat{\hat{v}}_2, \dots, \hat{\hat{v}}_m]^T$$

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

Applicazione ad un problema modello

Consideriamo la soluzione del problema

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

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

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

fourier.m

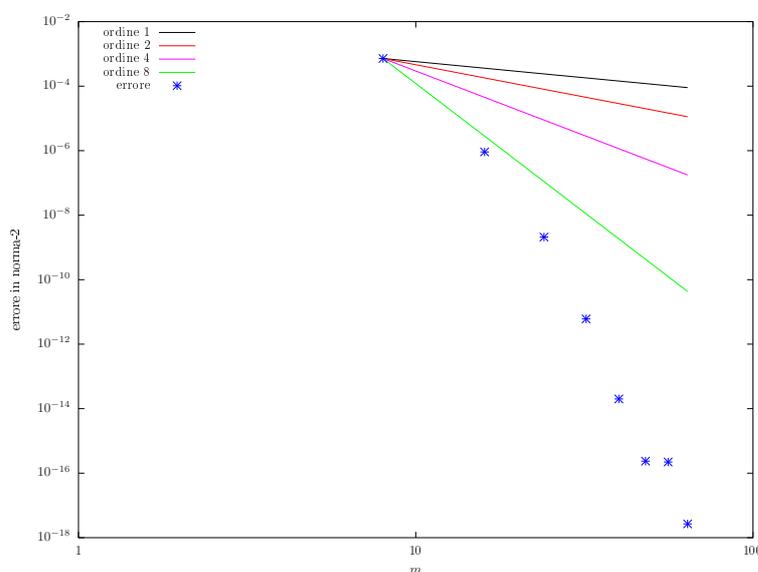


Figura 13.3: Convergenza spettrale di Fourier.

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

$$\begin{aligned} \int_{-\pi}^{\pi} \left(-\sum_{j=1}^m u_j \phi_j(x) \right)'' \overline{\phi_k(x)} dx + \int_{-\pi}^{\pi} \left(\sum_{j=1}^m u_j \phi_j(x) \right) \overline{\phi_k(x)} dx &= \\ &= \int_{-\pi}^{\pi} g(x) \overline{\phi_k(x)} dx \end{aligned}$$

da cui

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

Poiché

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

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 13.3 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). 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), 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

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

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 $\hat{u}_{m/2+2}$ 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 Cesàro*. Infatti, si sostituisce la serie troncata di Fourier

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

con la media delle troncate

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

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 , $j = 1, \dots, m$ 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, $x_k = (k-1)/n$, $1 \leq k \leq n$, $n > m$, n pari. Si possono introdurre dei coefficienti fittizi \hat{U}_k

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

Si avrà

$$\begin{aligned} \hat{u}(x_k) &= \sum_{j=1}^m \hat{u}_j \phi_j(x_k) = \sum_{j=1}^n \hat{U}_j \frac{e^{i(j-1-n/2)2\pi(x_k-a)/(b-a)}}{\sqrt{b-a}} = \\ &= \frac{n}{\sqrt{b-a}} \left[\frac{1}{n} \left(\sum_{j=1}^n \hat{U}_j e^{i(j-1)2\pi y_k} \right) \right] e^{-in\pi y_k} \end{aligned}$$

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.

13.3 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 5. *La soluzione del sistema lineare*

$$\sum_{j=1}^m \hat{u}_j L\phi_j(x_n) = g(x_n), \quad 1 \leq n \leq m \quad (13.11)$$

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

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

quando si approssimino gli integrali con le formule gaussiane.

Dimostrazione. Per ogni i , $1 \leq i \leq m$, da (13.11), si ha

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

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

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

che è precisamente la formulazione di Galerkin pseudospettrale (13.6). \square

13.3.1 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

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

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

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

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

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

e soddisfano

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

(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 $x_n = \cos((n-1)\pi/(m-1))$, $1 \leq n \leq m$. Possiamo allora definire la seguente famiglia di funzioni ortonormali

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

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

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

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

$$\begin{aligned}\phi'_1(x) &= 0, & \phi'_2(x) &= \sqrt{\frac{2}{\pi}}, & \phi'_3(x) &= 2\phi_2(x) + 2x\phi'_2(x), \\ \phi'_{j+1}(x) &= 2\phi_j(x) + 2x\phi'_j(x) - \phi'_{j-1}(x), & & & & 3 \leq j \leq m-1 \\ \phi''_1(x) &= 0, & \phi''_2(x) &= 0, & \phi''_3(x) &= 4\phi'_2(x), \\ \phi''_{j+1}(x) &= 4\phi'_j(x) + 2x\phi''_j(x) - \phi''_{j-1}(x), & & & & 3 \leq j \leq m-1\end{aligned}$$

Conviene calcolare le matrici

$$\begin{aligned}\mathbb{T}_0 &= \begin{bmatrix} \phi_1(x_1) & \phi_1(x_2) & \dots & \phi_1(x_m) \\ \phi_2(x_1) & \phi_2(x_2) & \dots & \phi_2(x_m) \\ \vdots & \vdots & \dots & \vdots \\ \phi_m(x_1) & \phi_m(x_2) & \dots & \phi_m(x_m) \end{bmatrix} \\ \mathbb{T}_1 &= \begin{bmatrix} \phi'_1(x_1) & \phi'_1(x_2) & \dots & \phi'_1(x_m) \\ \phi'_2(x_1) & \phi'_2(x_2) & \dots & \phi'_2(x_m) \\ \vdots & \vdots & \dots & \vdots \\ \phi'_m(x_1) & \phi'_m(x_2) & \dots & \phi'_m(x_m) \end{bmatrix} \\ \mathbb{T}_2 &= \begin{bmatrix} \phi''_1(x_1) & \phi''_1(x_2) & \dots & \phi''_1(x_m) \\ \phi''_2(x_1) & \phi''_2(x_2) & \dots & \phi''_2(x_m) \\ \vdots & \vdots & \dots & \vdots \\ \phi''_m(x_1) & \phi''_m(x_2) & \dots & \phi''_m(x_m) \end{bmatrix}\end{aligned}$$

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

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

Per imporre le condizioni al bordo, si sostituisce la prima riga della matrice con la prima riga di \mathbb{T}_0^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^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

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

Capitolo 14

Esercizi

1. Si risolva il problema ai limiti

$$\begin{cases} u''(x) = u(x) + x, & x \in (0, 1) \\ u(0) = 0 \\ u(1) = 0 \end{cases} \quad (14.1)$$

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

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

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 (14.1) 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 15

Introduzione

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

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

con dato iniziale

$$\begin{cases} y_1(t_0) = y_{10} \\ y_2(t_0) = y_{20} \\ \vdots \\ y_d(t_0) = y_{d0} \end{cases}$$

che può essere riscritto, in forma compatta,

$$\begin{cases} \mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)), & t > t_0 \\ \mathbf{y}(t_0) = \mathbf{y}_0 \end{cases} \quad (15.1)$$

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

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

Allora il sistema (15.1) ha un'unica soluzione.

15.1 Riduzione in forma autonoma

Un sistema in forma *non autonoma*

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

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

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

Si giunge a

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

15.2 Equazioni di ordine superiore al primo

Le equazioni differenziali di ordine d del tipo

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

(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

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

ove

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

Capitolo 16

Metodi ad un passo

16.1 Metodo di Eulero

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

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

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

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

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

$$\begin{aligned} \mathbf{y}_{n+1} &= \mathbf{y}_n + k\mathbf{f}(t_n, \mathbf{y}_n), \quad n \geq 0, \\ \mathbf{y}_0 &= \mathbf{y}(t_0) \end{aligned} \tag{16.1}$$

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

$$\begin{cases} \mathbf{y}^{*'}(t) = \mathbf{f}(t, \mathbf{y}^*(t)) \\ \mathbf{y}^*(t_n) = \mathbf{y}_n \end{cases}$$

Se consideriamo l'intervallo temporale $[t_0, t_0 + t^*]$, indichiamo con $\mathbf{y}_{n,m}$ (oppure $\mathbf{y}_{n,k}$), $n \leq m$, la soluzione approssimata al tempo t_n mediante un generico metodo $\mathbf{y}_{n+1,k} = \mathcal{Y}_n(\mathbf{f}, k, \mathbf{y}_0, \mathbf{y}_{1,k}, \dots, \mathbf{y}_{n,k})$ per la soluzione del sistema differenziale (15.1), ove il passo temporale è $k = t^*/m$.

Definizione 1. La quantità $\mathbf{y}(t_{n+1}) - \mathcal{Y}_n(\mathbf{f}, k, \mathbf{y}_0, \mathbf{y}(t_1), \dots, \mathbf{y}(t_n))$ si chiama errore locale del metodo.

Se consideriamo il problema differenziale

$$\begin{cases} \mathbf{y}^*(t) = \mathbf{f}(t, \mathbf{y}^*(t)) \\ \mathbf{y}^*(t_n) = \mathbf{y}(t_n) \end{cases}$$

si vede che l'errore locale coincide con $\mathbf{y}(t_{n+1}) - \mathbf{y}_{n+1}^*$, cioè con 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 . È chiaro allora che ad ogni passo si commette un nuovo errore che si accumula con l'errore prodotto ai passi precedenti. Per il metodo di Eulero, si ha

$$\begin{aligned} \mathbf{y}(t_{n+1}) - \mathbf{y}(t_n) - k\mathbf{f}(t_n, \mathbf{y}(t_n)) &= \\ &= \mathbf{y}(t_n) + k\mathbf{y}'(t_n) + \frac{k^2}{2}\mathbf{y}''(t_n) + \mathcal{O}(k^3) - \mathbf{y}(t_n) - k\mathbf{y}'(t_n) = \mathcal{O}(k^2) \end{aligned} \quad (16.2)$$

(supponendo $\mathbf{y} \in \mathcal{C}^2$). Poiché ad ogni passo si commette un errore di ordine $\mathcal{O}(k^2)$ e i passi sono $m = t^*/k$, se tutto va bene, anche quando $m \rightarrow \infty$ alla fine si commette un errore di ordine $\mathcal{O}(h^2 \cdot t^*/k) = \mathcal{O}(k)$. È giustificata allora la seguente

Definizione 2. Un metodo $\mathbf{y}_{n+1} = \mathcal{Y}_n(\mathbf{f}, k, \mathbf{y}_0, \dots, \mathbf{y}_n)$ è di ordine p se $\mathbf{y}(t_{n+1}) - \mathcal{Y}_n(\mathbf{f}, k, \mathbf{y}(t_0), \dots, \mathbf{y}(t_n)) = \mathcal{O}(k^{p+1})$, per $k \rightarrow 0$, per qualunque \mathbf{f} analitica e $0 \leq n \leq m - 1$.

La definizione sopra dice in verità che il metodo è *almeno* di ordine p . Un metodo di ordine $p \geq 1$ si dice *consistente di ordine p* , o semplicemente *consistente*. Se $\mathbf{y}(t_{n+1}) - \mathcal{Y}_n(\mathbf{f}, k, \mathbf{y}(t_0), \dots, \mathbf{y}(t_n)) = 0$ 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*.

Definizione 3. Il metodo $\mathbf{y}_{n+1,k} = \mathcal{Y}_n(\mathbf{f}, k, \mathbf{y}_0, \mathbf{y}_{1,k}, \dots, \mathbf{y}_{n,k})$ è convergente se

$$\lim_{k \rightarrow 0^+} \max_{0 \leq n \leq m} \|\mathbf{e}_{n,k}\| = 0$$

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

Teorema 6. Il metodo di Eulero è convergente.

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

$$\begin{aligned}\mathbf{y}_{n+1,k} &= \mathbf{y}_{n,k} + k\mathbf{f}(t_n, \mathbf{y}_{n,k}) && \text{(definizione del metodo)} \\ \mathbf{y}(t_{n+1}) &= \mathbf{y}(t_n) + k\mathbf{f}(t_n, \mathbf{y}(t_n)) + \mathcal{O}(k^2) && \text{(errore locale (16.2))}\end{aligned}$$

si ricava

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

da cui, siccome \mathbf{f} è lipschitziana,

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

Allora

$$\|\mathbf{e}_{n,k}\| \leq \frac{c}{\lambda}k[(1 + k\lambda)^n - 1], \quad 0 \leq n \leq m$$

(si dimostra per induzione). Poiché $1 + k\lambda < e^{k\lambda}$, $(1 + k\lambda)^n < e^{nk\lambda} \leq e^{mk\lambda} = e^{t^*\lambda}$. Dunque

$$\|\mathbf{e}_{n,k}\| \leq \frac{c}{\lambda}k(e^{t^*\lambda} - 1), \quad 0 \leq n \leq m$$

da cui

$$\lim_{k \rightarrow 0^+} \max_{0 \leq n \leq m} \|\mathbf{e}_{n,k}\| = 0$$

□

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).

16.2 Metodo dei trapezi

Il *metodo dei trapezi* (o metodo di *Crank–Nicolson*) si basa sull'approssimazione

$$\begin{aligned}\mathbf{y}'(t) &\approx \frac{\mathbf{y}(t) - \mathbf{y}(t_0)}{t - t_0} \\ \mathbf{f}(t, \mathbf{y}(t)) &\approx \frac{1}{2}(\mathbf{f}(t_0, \mathbf{y}(t_0)) + \mathbf{f}(t, \mathbf{y}(t)))\end{aligned}$$

Può essere visto anche come applicazione della formula di quadratura del trapezio alla soluzione analitica

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

Data la sequenza $t_0, t_1 = t_0 + k, t_2 = t_0 + 2k, \dots, t_n = t_0 + nk, \dots$, lo schema numerico che ne risulta è

$$\begin{aligned} \mathbf{y}_{n+1} &= \mathbf{y}_n + \frac{k}{2}(\mathbf{f}(t_n, \mathbf{y}_n) + \mathbf{f}(t_{n+1}, \mathbf{y}_{n+1})), \quad n \geq 0, \\ \mathbf{y}_0 &= \mathbf{y}(t_0) \end{aligned} \quad (16.3)$$

ove $\mathbf{y}_n \approx \mathbf{y}(t_n)$. Dato lo schema

$$\mathbf{y}_{n+1} - \mathbf{y}_n - \frac{k}{2}(\mathbf{f}(t_n, \mathbf{y}_n) + \mathbf{f}(t_{n+1}, \mathbf{y}_{n+1})) = 0$$

sostituendo \mathbf{y}_n con $\mathbf{y}(t_n)$ e \mathbf{y}_{n+1} con $\mathbf{y}(t_{n+1})$ si ottiene

$$\begin{aligned} &\mathbf{y}(t_{n+1}) - \mathbf{y}(t_n) - \frac{k}{2}(\mathbf{f}(t_n, \mathbf{y}(t_n)) + \mathbf{f}(t_{n+1}, \mathbf{y}(t_{n+1}))) = \\ &= \mathbf{y}(t_n) + k\mathbf{y}'(t_n) + \frac{k^2}{2}\mathbf{y}''(t_n) + \frac{k^3}{6}\mathbf{y}'''(t_n) + \mathcal{O}(k^4) - \mathbf{y}(t_n) - \frac{k}{2}(\mathbf{y}'(t_n) + \mathbf{y}'(t_{n+1})) = \\ &\quad k\mathbf{y}'(t_n) + \frac{k^2}{2}\mathbf{y}''(t_n) + \frac{k^3}{6}\mathbf{y}'''(t_n) + \mathcal{O}(k^4) + \\ &\quad - \frac{k}{2} \left(\mathbf{y}'(t_n) + \mathbf{y}'(t_n) + k\mathbf{y}''(t_n) + \frac{k^2}{2}\mathbf{y}'''(t_n) + \mathcal{O}(k^3) \right) = \mathcal{O}(k^3) \end{aligned}$$

(supponendo $\mathbf{y} \in \mathcal{C}^3$). Dunque il metodo dei trapezi è di ordine 2.

Teorema 7. *Il metodo dei trapezi è convergente.*

Dimostrazione. Si ha

$$\begin{aligned} \mathbf{e}_{n+1,k} &= \mathbf{e}_{n,k} + \frac{k}{2}(\mathbf{f}(t_n, \mathbf{y}_n) - \mathbf{f}(t_n, \mathbf{y}(t_n))) + \\ &\quad + \frac{k}{2}(\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}) - \mathbf{f}(t_{n+1}, \mathbf{y}(t_{n+1}))) + \mathcal{O}(k^3) \end{aligned}$$

da cui

$$\|\mathbf{e}_{n+1,k}\| \leq \|\mathbf{e}_{n,k}\| + \frac{k\lambda}{2}(\|\mathbf{e}_{n,k}\| + \|\mathbf{e}_{n+1,k}\|) + ck^3, \quad c > 0$$

Se $k < 2/\lambda$,

$$\|\mathbf{e}_{n+1,k}\| \leq \left(\frac{1 + \frac{k\lambda}{2}}{1 - \frac{k\lambda}{2}} \right) \|\mathbf{e}_{n,k}\| + \left(\frac{c}{1 - \frac{k\lambda}{2}} \right) k^3$$

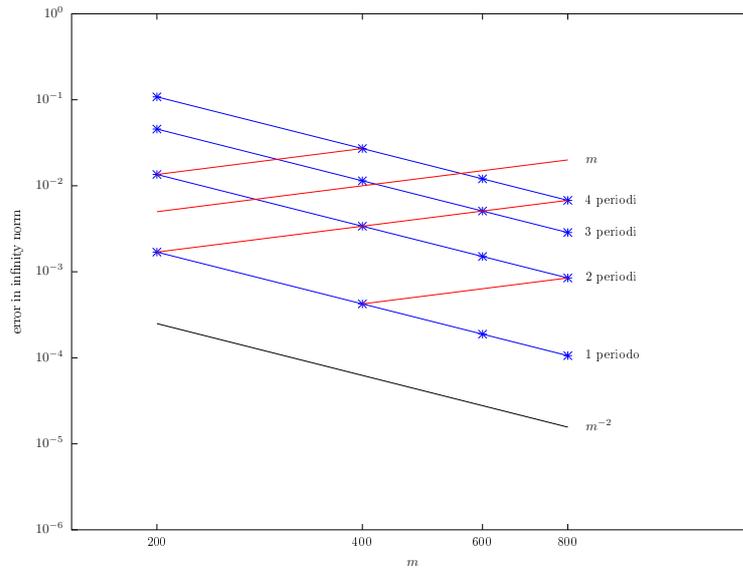
Allora

$$\|\mathbf{e}_{n,k}\| \leq \frac{ck^2}{\lambda} \left[\left(\frac{1 + \frac{k\lambda}{2}}{1 - \frac{k\lambda}{2}} \right)^n - 1 \right], \quad 0 \leq n \leq m$$

(si dimostra per induzione). Si conclude osservando che

$$\left(\frac{1 + \frac{k\lambda}{2}}{1 - \frac{k\lambda}{2}}\right)^n = \left(1 + \frac{k\lambda}{1 - \frac{k\lambda}{2}}\right)^n \leq \exp\left(\frac{nk\lambda}{1 - k\lambda/2}\right) \leq \exp\left(\frac{t^*\lambda}{1 - k\lambda/2}\right)$$

□



pendolo period

Figura 16.1: Errori del metodo dei trapezi per l'equazione del pendolo linearizzata.

In Figura 16.1 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 \mathbf{y}_{n+1} dipende esplicitamente solo da \mathbf{y}_n). Il metodo dei trapezi è però *implicito*, cioè la soluzione \mathbf{y}_{n+1} è implicitamente definita dall'equazione (in generale non lineare)

$$F_n(\mathbf{y}_{n+1}) = \mathbf{y}_{n+1} - \frac{k}{2}\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}) - \mathbf{y}_n - \frac{k}{2}\mathbf{f}(t_n, \mathbf{y}_n) = 0$$

16.3 θ -metodo

Il θ -metodo è una generalizzazione dei metodi precedenti e si scrive

$$\begin{aligned} \mathbf{y}_{n+1} &= \mathbf{y}_n + k[(1 - \theta)\mathbf{f}(t_n, \mathbf{y}_n) + \theta\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1})], \quad n \geq 0 \\ \mathbf{y}_0 &= \mathbf{y}(t_0) \end{aligned} \quad (16.4)$$

È facile verificare che

$$\begin{aligned} \mathbf{y}(t_{n+1}) - \mathbf{y}(t_n) - k[(1 - \theta)\mathbf{f}(t_n, \mathbf{y}(t_n)) + \theta\mathbf{f}(t_{n+1}, \mathbf{y}(t_{n+1}))] = \\ = \left(\frac{1}{2} - \theta\right)k^2\mathbf{y}''(t_n) + \left(\frac{1}{6} - \frac{\theta}{2}\right)k^3\mathbf{y}'''(t_n) + \mathcal{O}(k^4) \end{aligned} \quad (16.5)$$

e dunque il metodo ha ordine due se $\theta = \frac{1}{2}$, e ordine uno altrimenti. In particolare, se $\mathbf{y}''(t)$ è nulla, tale è l'errore locale per il θ -metodo. E se $\theta = \frac{1}{2}$ e $\mathbf{y}'''(t)$ è nulla, tale è l'errore locale. Poiché però ad ogni passo si commettono errori di arrotondamento (nel caso esplicito) o di arrotondamento e approssimazione (nel caso implicito, in cui c'è da risolvere un sistema non lineare), per avere la convergenza è comunque necessario che questi non si accumulino in maniera distruttiva.

Possiamo considerare anche in questo caso il problema differenziale

$$\begin{cases} \mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}^*(t)) \\ \mathbf{y}^*(t_n) = \mathbf{y}(t_n) \end{cases}$$

e applicare un passo del θ -metodo. Si ha

$$\mathbf{y}_{n+1}^* = \mathbf{y}(t_n) + k[(1 - \theta)\mathbf{f}(t_n, \mathbf{y}(t_n)) + \theta\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}^*)]$$

da cui

$$\begin{aligned} \mathbf{y}(t_{n+1}) - \mathbf{y}_{n+1}^* = k\theta[\mathbf{f}(t_{n+1}, \mathbf{y}(t_{n+1})) - \mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}^*)] + \\ + \left(\frac{1}{2} - \theta\right)k^2\mathbf{y}''(t_n) + \left(\frac{1}{6} - \frac{\theta}{2}\right)k^3\mathbf{y}'''(t_n) + \mathcal{O}(k^4) \end{aligned}$$

Posto $\mathbf{e}_{n+1}^* = \mathbf{y}(t_{n+1}) - \mathbf{y}_{n+1}^*$, si ha

$$\|\mathbf{e}_{n+1}^*\| \leq k\theta\lambda\|\mathbf{e}_{n+1}^*\| + \left(\frac{1}{2} - \theta\right)\mathcal{O}(k^2) + \left(\frac{1}{6} - \frac{\theta}{2}\right)\mathcal{O}(k^3) + \mathcal{O}(k^4)$$

Dunque, se k è sufficientemente piccolo,

$$\|\mathbf{e}_{n+1}^*\| \leq \left(\frac{1}{2} - \theta\right)\mathcal{O}(k^2) + \left(\frac{1}{6} - \frac{\theta}{2}\right)\mathcal{O}(k^3) + \mathcal{O}(k^4)$$

Quindi, anche per uno schema implicito, l'errore di troncamento 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 .

Teorema 8. *Il θ -metodo, per $\theta \in [0, 1]$, è convergente.*

Osserviamo che:

- il metodo è esplicito per $\theta = 0$ (e in tal caso si riduce al metodo di Eulero esplicito);
- il metodo è di ordine due per $\theta = \frac{1}{2}$ (e in tal caso si riduce al metodo dei trapezi);
- 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.

16.3.1 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 $F_n(\mathbf{x}) = 0$, $\mathbf{x} = \mathbf{y}_{n+1}$, ove

$$F_n(\mathbf{x}) = \mathbf{x} - k\theta \mathbf{f}(t_{n+1}, \mathbf{x}) - \mathbf{y}_n - k(1 - \theta) \mathbf{f}(t_n, \mathbf{y}_n)$$

La prima idea potrebbe essere quella di applicare il metodo di punto fisso:

$$\mathbf{x}^{(r+1)} = \mathbf{g}(\mathbf{x}^{(k)}) = k\theta \mathbf{f}(t_{n+1}, \mathbf{x}^{(r)}) + \mathbf{y}_n + k(1 - \theta) \mathbf{f}(t_n, \mathbf{y}_n)$$

La funzione \mathbf{g} soddisfa

$$\|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})\| = \|k\theta \mathbf{f}(t_{n+1}, \mathbf{x}) - k\theta \mathbf{f}(t_{n+1}, \mathbf{y})\| \leq k\theta\lambda \|\mathbf{x} - \mathbf{y}\|$$

ed è una contrazione se

$$k\theta\lambda < 1 \Rightarrow k < \frac{1}{\theta\lambda} \quad (16.6)$$

Dunque, a patto di prendere k sufficientemente piccolo, si può ricavare la soluzione $\mathbf{x} = \mathbf{y}_{n+1}$ con il metodo del punto fisso, partendo da $\mathbf{x}^{(0)} = \mathbf{y}_n$. 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

$$J_n(\mathbf{x}) = I - k\theta \left(\frac{\partial f_i(t_{n+1}, \mathbf{x})}{\partial x_j} \right)_{ij}$$

Il vettore iniziale per il calcolo di \mathbf{y}_{n+1} è di solito la soluzione al passo precedente \mathbf{y}_n . Il calcolo di \mathbf{y}_{n+1} a partire da \mathbf{y}_n con il metodo di Newton avviene dunque secondo il seguente algoritmo:

- $r = 0$
- $\mathbf{x}^{(r)} = \mathbf{y}_n$
- $J_n(\mathbf{x}^{(r)})\boldsymbol{\delta}^{(r)} = -F_n(\mathbf{x}^{(r)})$
- WHILE $\|\boldsymbol{\delta}^{(r)}\| > \text{Newt_tol}$

$$\mathbf{x}^{(r+1)} = \mathbf{x}^{(r)} + \boldsymbol{\delta}^{(r)}$$

$$r = r + 1$$

$$J_n(\mathbf{x}^{(r)})\boldsymbol{\delta}^{(r)} = -F_n(\mathbf{x}^{(r)})$$

END

- $\mathbf{y}_{n+1} = \mathbf{x}^{(r)}$

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 $\mathbf{x}^{(0)}$, invece di \mathbf{y}_n si può prendere $\mathbf{y}_n + k\mathbf{f}(t_n, \mathbf{y}_n)$, che corrisponde a $\mathbf{y}(t_{n+1})$ approssimato con il metodo di Eulero e dunque dovrebbe essere un'approssimazione migliore di \mathbf{y}_{n+1} .

16.3.2 Newton inesatto e passo variabile

Nel caso in cui il calcolo e/o la risoluzione dei sistemi lineari con $J_n(\mathbf{x}^{(r)})$ risulti particolarmente oneroso, si può ricorrere al metodo di Newton inesatto, considerando ad ogni passo la matrice jacobiana costante $J_n(\mathbf{x}^{(r)}) \equiv J_n(\mathbf{x}^{(0)})$, 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. Il numero di iterazioni necessarie alla convergenza dipende (anche) dalla vicinanza della soluzione iniziale $\mathbf{x}^{(0)} = \mathbf{y}_n$ a quella finale \mathbf{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 ($\theta \neq 0$):

- $r = 0$
- $\mathbf{x}^{(r)} = \mathbf{y}_n$
- $F_n(\mathbf{x}^{(r)}) = \mathbf{x}^{(r)} - k_{n+1}\theta \mathbf{f}(t_{n+1}, \mathbf{x}^{(r)}) - \mathbf{y}_n - k_{n+1}(1 - \theta)\mathbf{f}(t_n, \mathbf{y}_n)$
- $J_n(\mathbf{x}^{(r)}) = I - k_{n+1}\theta \left(\frac{\partial f_i(t_{n+1}, \mathbf{x}^{(r)})}{\partial x_j} \right)_{ij}$
- $P_n J_n(\mathbf{x}^{(r)}) = L_n U_n$
- $L_n U_n \boldsymbol{\delta}^{(r)} = -P_n F_n(\mathbf{x}^{(r)})$
- WHILE $\|\boldsymbol{\delta}^{(r)}\| > \text{Newt_tol}$
 - $\mathbf{x}^{(r+1)} = \mathbf{x}^{(r)} + \boldsymbol{\delta}^{(r)}$
 - $r = r + 1$
 - $L_n U_n \boldsymbol{\delta}^{(r)} = -P_n F_n(\mathbf{x}^{(r)})$
- END
- $\mathbf{y}_{n+1} = \mathbf{x}^{(r)}$
- 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 \mathbf{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.

16.3.3 Caso lineare

Un caso molto frequente è quello lineare autonomo a coefficienti costanti

$$\begin{cases} \mathbf{y}'(t) = A\mathbf{y}(t) + \mathbf{b} \\ \mathbf{y}(t_0) = \mathbf{y}_0 \end{cases}$$

con passo di integrazione k costante. In tal caso, il metodo si scrive

$$(I - k\theta A)\mathbf{y}_{n+1} = (I + k(1 - \theta)A)\mathbf{y}_n + k\mathbf{b}$$

Nel caso implicito, si tratta dunque di risolvere un sistema *lineare* di matrice $I - k\theta A$ ad ogni passo. Pertanto, per problemi di piccola dimensione, è [thetametodolin.m](#) conveniente precalcolare la fattorizzazione LU della matrice. Altrimenti, si può considerare un metodo iterativo, ove si scelga come vettore iniziale per il calcolo di \mathbf{y}_{n+1} la soluzione al passo precedente \mathbf{y}_n . Se il termine \mathbf{b} dipende invece dal tempo, allora $k\mathbf{b}$ va sostituito, ovviamente, con $k(1 - \theta)\mathbf{b}(t_n) + k\theta\mathbf{b}(t_{n+1})$.

16.3.4 Metodi semiimpliciti

Da scrivere

16.4 Verifica della correttezza dell'implementazione

Supponiamo di aver implementato un metodo di ordine p per la soluzione del sistema differenziale

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

e di volerne testare la corretta implementazione. L'idea è quella di creare una soluzione artificiale $\mathbf{x}(t)$, inserirla nell'equazione e calcolarne il residuo

$$\mathbf{x}'(t) - \mathbf{f}(t, \mathbf{x}(t)) = \mathbf{g}(t)$$

A questo punto, si risolve il sistema differenziale

$$\begin{cases} \mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)) + \mathbf{g}(t) = \hat{\mathbf{f}}(t, \mathbf{y}(t)) \\ \mathbf{y}(t_0) = \mathbf{x}(t_0) \end{cases}$$

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 $\mathbf{e}_{m_1, k_1} = \|\mathbf{y}_{m_1, k_1} - \mathbf{x}(t_0 + t^*)\| = Ck_1^p$ e $\mathbf{e}_{m_2, k_2} = \|\mathbf{y}_{m_2, k_2} - \mathbf{x}(t_0 + t^*)\| = Ck_2^p$. Si ha dunque

$$\frac{\mathbf{e}_{m_2, k_2}}{\mathbf{e}_{m_1, k_1}} = \left(\frac{k_2}{k_1}\right)^p,$$

da cui

$$\log \mathbf{e}_{m_2, k_2} - \log \mathbf{e}_{m_1, k_1} = p(\log k_2 - \log k_1) = -p(\log m_2 - \log m_1).$$

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 $\mathbf{f}(t, \mathbf{y}(t))$ sia particolarmente complicato, invece di calcolare il residuo, si può calcolare una *soluzione di riferimento* $\mathbf{y}_{\bar{m}, \bar{k}}$ e poi confrontare con essa le soluzioni \mathbf{y}_{m_1, k_1} e \mathbf{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 17

Metodi multistep

17.1 Metodi di Adams–Bashforth

Invece di costruire la soluzione \mathbf{y}_{n+1} a partire dalla sola soluzione al passo precedente \mathbf{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

$$\mathbf{y}(t_{n+s}) = \mathbf{y}(t_{n+s-1}) + \int_{t_{n+s-1}}^{t_{n+s}} \mathbf{f}(\tau, \mathbf{y}(\tau)) d\tau \quad (17.1)$$

l'idea è quella di sostituire la funzione integranda in (17.1) con il “suo” polinomio interpolatore sui nodi equispaziati $t_n, t_{n+1}, \dots, t_{n+s-1}$ ($t_{n+j} = t_n + jk$)

$$\mathbf{p}(\tau) = \sum_{j=0}^{s-1} L_j(\tau) \mathbf{f}(t_{n+j}, \mathbf{y}_{n+j})$$

ove $L_j(t)$ è il polinomio elementare di Lagrange di grado $s-1$ definito da $L_j(t_{n+i}) = \delta_{ij}$. Poiché $\mathbf{p}(t_{n+j}) = \mathbf{f}(t_{n+j}, \mathbf{y}_{n+j})$, $0 \leq j \leq s-1$ (e non, ovviamente, $\mathbf{p}(t_{n+j}) = \mathbf{f}(t_{n+j}, \mathbf{y}(t_{n+j}))$), dobbiamo supporre di avere già a disposizione i valori iniziali $\mathbf{y}_{n+j} \approx \mathbf{y}(t_{n+j})$, $0 \leq j \leq s-1$. Si ha dunque

$$\int_{t_{n+s-1}}^{t_{n+s}} \mathbf{p}(\tau) d\tau = \sum_{j=0}^{s-1} \left(\int_{t_{n+s-1}}^{t_{n+s}} L_j(\tau) d\tau \right) \mathbf{f}(t_{n+j}, \mathbf{y}_{n+j}) = k \sum_{j=0}^{s-1} b_j \mathbf{f}(t_{n+j}, \mathbf{y}_{n+j})$$

da cui il metodo esplicito *multistep Adams–Bashforth*

$$\mathbf{y}_{n+s} = \mathbf{y}_{n+s-1} + k \sum_{j=0}^{s-1} b_j \mathbf{f}(t_{n+j}, \mathbf{y}_{n+j}) \quad (17.2)$$

I coefficienti b_j non dipendono da n e neanche da k : infatti

$$\begin{aligned} b_j &= \frac{1}{k} \int_{t_{n+s-1}}^{t_{n+s}} \prod_{\substack{i=0 \\ i \neq j}}^{s-1} \frac{\tau - t_{n+i}}{t_{n+j} - t_{n+i}} d\tau = \int_0^1 \prod_{\substack{i=0 \\ i \neq j}}^{s-1} \frac{t_{n+s-1} + rk - t_{n+i}}{t_{n+j} - t_{n+i}} dr = \\ &= \int_0^1 \prod_{\substack{i=0 \\ i \neq j}}^{s-1} \frac{((s-1-i)k + rk)}{(j-i)k} dr = \int_0^1 \prod_{\substack{i=0 \\ i \neq j}}^{s-1} \frac{((s-1-i) + r)}{(j-i)} dr \end{aligned}$$

Dunque possono essere calcolati una volta per tutte. Calcoliamo l'ordine di tale metodo: come al solito, dobbiamo valutare l'espressione

$$\mathbf{y}(t_{n+s}) - \mathbf{y}(t_{n+s-1}) - k \sum_{j=0}^{s-1} b_j \mathbf{f}(t_{n+j}, \mathbf{y}(t_{n+j}))$$

L'ultimo termine è l'integrale del polinomio $\mathbf{q}(\tau)$ di grado $s-1$ che interpola $\mathbf{f}(t_{n+j}, \mathbf{y}(t_{n+j}))$, $0 \leq j \leq s-1$. Dunque, per $t_{n+s-1} \leq \tau \leq t_{n+s}$,

$$\begin{aligned} \|\mathbf{q}(\tau) - \mathbf{f}(\tau, \mathbf{y}(\tau))\| &\leq \frac{\|\mathbf{y}^{(s+1)}(\bar{\tau})\|}{s!} \underbrace{|(\tau - t_n) \cdots (\tau - t_{n+s-1})|}_{s \text{ termini}} \leq \\ &\leq \frac{\|\mathbf{y}^{(s+1)}(\bar{\tau})\|}{s!} s! k^s = \mathcal{O}(k^s), \quad t_{n+s-1} \leq \tau \leq t_{n+s} \end{aligned}$$

e quindi

$$\begin{aligned} \mathbf{y}(t_{n+s}) - \mathbf{y}(t_{n+s-1}) - k \sum_{j=0}^{s-1} b_j \mathbf{f}(t_{n+j}, \mathbf{y}(t_{n+j})) &= \\ - k \sum_{j=0}^{s-1} b_j \mathbf{f}(t_{n+j}, \mathbf{y}(t_{n+j})) + \int_{t_{n+s-1}}^{t_{n+s}} \mathbf{f}(\tau, \mathbf{y}(\tau)) d\tau &= \\ - \int_{t_{n+s-1}}^{t_{n+s}} \mathbf{q}(\tau) d\tau + \int_{t_{n+s-1}}^{t_{n+s}} \mathbf{f}(\tau, \mathbf{y}(\tau)) d\tau &= \mathcal{O}(k^{s+1}) \end{aligned}$$

perché un ulteriore fattore k deriva dal fatto che si integra in un intervallo di ampiezza k . Quindi, se anche $\mathbf{y}_n = \mathbf{y}(t_n) + \mathcal{O}(k^s)$, $1 \leq n \leq s-1$ (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 $\mathbf{f}(t_n, \mathbf{y}_n)$. È ovviamente $\mathbf{p}(\tau) \equiv \mathbf{f}(t_n, \mathbf{y}_n)$ e $b_0 = 1$, quindi

$$\mathbf{y}_{n+1} = \mathbf{y}_n + k \mathbf{f}(t_n, \mathbf{y}_n)$$

e si ritrova il metodo di Eulero. Nel caso $s = 2$, il polinomio interpolatore è

$$\mathbf{p}(\tau) = \frac{t_{n+1} - \tau}{k} \mathbf{f}(t_n, \mathbf{y}_n) + \frac{\tau - t_n}{k} \mathbf{f}(t_{n+1}, \mathbf{y}_{n+1})$$

e dunque

$$b_0 = \frac{1}{k} \int_{t_{n+1}}^{t_{n+2}} \frac{t_{n+1} - \tau}{k} d\tau = -\frac{1}{2}, \quad b_1 = \frac{1}{k} \int_{t_{n+1}}^{t_{n+2}} \frac{\tau - t_n}{k} d\tau = \frac{3}{2}$$

da cui

$$\mathbf{y}_{n+2} = \mathbf{y}_{n+1} - \frac{k}{2} \mathbf{f}(t_n, \mathbf{y}_n) + \frac{3k}{2} \mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}) \quad (17.3)$$

Il valore \mathbf{y}_1 può essere ricavato, per esempio, anche con il metodo di Eulero, poiché si ha, in tal caso, $\mathbf{y}_1 = \mathbf{y}(t_1) + \mathcal{O}(k^2)$.

17.2 Metodi lineari multistep

Una semplice generalizzazione del metodo di Adams–Bashforth è permettere che il metodo sia implicito:

$$\mathbf{y}_{n+s} = \mathbf{y}_{n+s-1} + k \sum_{j=0}^s b_j \mathbf{f}(t_{n+j}, \mathbf{y}_{n+j}) \quad (17.4)$$

Questa famiglia di metodi prende il nome di *Adams–Moulton* e vi rientra anche il θ -metodo ($b_0 = (1 - \theta)$, $b_1 = \theta$). Ancora più in generale,

$$\sum_{j=0}^s a_j \mathbf{y}_{n+j} = k \sum_{j=0}^s b_j \mathbf{f}(t_{n+j}, \mathbf{y}_{n+j}) \quad (17.5)$$

con la normalizzazione $a_s = 1$. Il metodo è di ordine p se, come al solito,

$$\sum_{j=0}^s a_j \mathbf{y}(t_{n+j}) - k \sum_{j=0}^s b_j \mathbf{f}(t_{n+j}, \mathbf{y}(t_{n+j})) = \mathcal{O}(k^{p+1})$$

per ogni funzione \mathbf{f} analitica e $0 \leq n \leq m - s$. Siccome la verifica può risultare molto tediosa, risulta utile il seguente

Teorema 9. *Dato un metodo multistep (17.5), 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

$$\rho(1 + \xi) - \sigma(1 + \xi) \cdot \ln(1 + \xi) = c\xi^{p+1} + \mathcal{O}(\xi^{p+2}) \quad \text{per } \xi \rightarrow 0$$

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 (17.3) 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 9. Si ha

$$\begin{aligned} & \sum_{j=0}^s a_j \mathbf{y}(t_{n+j}) - k \sum_{j=0}^s b_j \mathbf{f}(t_{n+j}, \mathbf{y}(t_{n+j})) = \\ & \quad \sum_{j=0}^s a_j \mathbf{y}(t_n + jk) - k \sum_{j=0}^s b_j \mathbf{y}'(t_n + jk) = \\ & = \sum_{j=0}^s a_j \sum_{i=0}^{\infty} \mathbf{y}^{(i)}(t_n) \frac{j^i k^i}{i!} - k \sum_{j=0}^s b_j \sum_{i=0}^{\infty} \mathbf{y}^{(i+1)}(t_n) \frac{j^i k^i}{i!} = \\ & = \sum_{j=0}^s a_j \left(\mathbf{y}(t_n) + \sum_{i=1}^{\infty} \mathbf{y}^{(i)}(t_n) \frac{j^i k^i}{i!} \right) - k \sum_{j=0}^s b_j \sum_{i=1}^{\infty} \mathbf{y}^{(i)}(t_n) \frac{j^{i-1} k^{i-1}}{(i-1)!} = \\ & = \left(\sum_{j=0}^s a_j \right) \mathbf{y}(t_n) + \sum_{i=1}^{\infty} \frac{1}{i!} \left(\sum_{j=0}^s j^i a_j - i \sum_{j=0}^s j^{i-1} b_j \right) k^i \mathbf{y}^{(i)}(t_n) \end{aligned}$$

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

$$\sum_{j=0}^s j^i a_j = i \sum_{j=0}^s j^{i-1} b_j, \quad i = 0, 1, \dots, p \quad (17.6)$$

Per finire la dimostrazione, si calcola lo sviluppo in serie di Taylor di $\rho(e^z) - \sigma(e^z)z$ per $z \rightarrow 0$ e si osserva che esso è $\mathcal{O}(z^{p+1})$ se e solo se vale (17.6). Posto $w = e^z$, l'ordine è p se e solo se lo sviluppo di $\rho(w) - \sigma(w) \cdot \ln w$ per $w \rightarrow 1$ vale

$$-c(\ln w)^{p+1} + \mathcal{O}((\ln w)^{p+2}) = c(w-1)^{p+1} + \mathcal{O}((w-1)^{p+2})$$

A questo punto, si pone $\xi = w - 1$. □

Non bisogna dimenticare che l'ordine è p se $\mathbf{y}(t)$ 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 (17.6) sono molto utili per determinare l'ordine di un metodo multistep: per esempio, per il metodo lineare multistep (implicito) a due passi

$$\begin{aligned} \mathbf{y}_{n+2} - 3\mathbf{y}_{n+1} + 2\mathbf{y}_n &= \\ &= \frac{k}{12} [-5\mathbf{f}(t_n, \mathbf{y}_n) - 20\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}) + 13\mathbf{f}(t_{n+2}, \mathbf{y}_{n+2})] \end{aligned} \quad (17.7)$$

si ha

$$\begin{aligned} a_0 + a_1 + a_2 &= 0 \\ a_1 + 2a_2 &= b_0 + b_1 + b_2 \Rightarrow \text{ordine (almeno) 1} \\ a_1 + 4a_2 &= 2(b_1 + 2b_2) \Rightarrow \text{ordine (almeno) 2} \\ a_1 + 8a_2 &\neq 3(b_1 + 4b_2) \Rightarrow \text{ordine 2} \end{aligned}$$

17.2.1 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

$$\sum_{j=0}^s a_j \mathbf{y}_{n+j} = kb_s \mathbf{f}(t_{n+s}, \mathbf{y}_{n+s})$$

e $kb_s \mathbf{f}(t_{n+s}, \mathbf{y}_{n+s}) \approx kb_s \mathbf{y}'(t_n + sk)$, conviene cercare una combinazione lineare di \mathbf{y}_{n+j} , $0 \leq j \leq s$ che approssimi $kb_s \mathbf{y}'(t_{n+s})$. Si procede dunque con lo sviluppo in serie di Taylor di $\mathbf{y}(t_n + jk)$, $0 \leq j \leq s$, centrato in $\mathbf{y}(t_n + sk)$. Per esempio, per $s = 1$,

$$\begin{cases} \mathbf{y}(t_n) = \mathbf{y}(t_n + k) - k\mathbf{y}'(t_n + k) + \mathcal{O}(k^2) \\ \mathbf{y}(t_n + k) = \mathbf{y}(t_n + k) \end{cases}$$

da cui $a_1 = 1$ e $a_0 = -1$. Dunque, il metodo BDF di ordine 1 è il metodo di Eulero implicito (*backward Euler*) ed è di ordine 1, come noto. Per $s = 2$

$$\begin{cases} \mathbf{y}(t_n) = \mathbf{y}(t_n + 2k) - 2k\mathbf{y}'(t_n + 2k) + \frac{4k^2}{2}\mathbf{y}''(t_n + 2k) + \mathcal{O}(k^3) \\ \mathbf{y}(t_n + k) = \mathbf{y}(t_n + 2k) - k\mathbf{y}'(t_n + 2k) + \frac{k^2}{2}\mathbf{y}''(t_n + 2k) + \mathcal{O}(k^3) \\ \mathbf{y}(t_n + 2k) = \mathbf{y}(t_n + 2k) \end{cases}$$

da cui

$$\begin{cases} a_2 = 1 \\ a_0 + a_1 + a_2 = 0 \\ -2a_0 - a_1 = b_2 \\ 2a_0 + \frac{a_1}{2} = 0 \end{cases} \Rightarrow \begin{cases} a_0 = \frac{1}{3} \\ a_1 = -\frac{4}{3} \\ a_2 = 1 \\ b_2 = \frac{2}{3} \end{cases} \quad (17.8)$$

e il metodo è di ordine 2.

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

$$\begin{cases} \mathbf{y}(t_n) = \mathbf{y}(t_n + k_{n+1} + k_{n+2}) - (k_{n+1} + k_{n+2})\mathbf{y}'(t_n + k_{n+1} + k_{n+2}) + \\ + \frac{(k_{n+1} + k_{n+2})^2}{2}\mathbf{y}''(t_n + k_{n+1} + k_{n+2}) + \dots \\ \mathbf{y}(t_n + k_{n+1}) = \mathbf{y}(t_n + k_{n+1} + k_{n+2}) - k_{n+2}\mathbf{y}'(t_n + k_{n+1} + k_{n+2}) + \\ + \frac{k_{n+2}^2}{2}\mathbf{y}''(t_n + k_{n+1} + k_{n+2}) + \dots \\ \mathbf{y}(t_n + k_{n+1} + k_{n+2}) = \mathbf{y}(t_n + k_{n+1} + k_{n+2}) \end{cases}$$

da cui i coefficienti

$$\begin{cases} a_2 = 1 \\ a_0 + a_1 + a_2 = 0 \\ -a_0(k_{n+1} + k_{n+2}) - a_1k_{n+2} = b_2k_{n+2} \\ \frac{a_0}{2}(k_{n+1} + k_{n+2})^2 + \frac{a_1}{2}k_{n+2}^2 = 0 \end{cases} \Rightarrow \begin{cases} a_0 = \frac{k_{n+2}^2}{(k_{n+1} + 2k_{n+2})k_{n+1}} \\ a_1 = -\frac{(k_{n+1} + k_{n+2})^2}{(k_{n+1} + 2k_{n+2})k_{n+1}} \\ a_2 = 1 \\ b_2 = \frac{(k_{n+1} + k_{n+2})}{(k_{n+1} + 2k_{n+2})} \end{cases}$$

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 16.3.2). Questi metodi risultano particolarmente vantaggiosi quando la valutazione della funzione \mathbf{f} è onerosa, poiché permettono di raggiungere un ordine elevato con una sola valutazione (nel caso lineare, altrimenti è necessario valutare \mathbf{f} in un ciclo di Newton ad ogni passo temporale).

17.3 Consistenza e stabilità

Dalle condizioni d'ordine (17.6), si vede che un metodo lineare multistep è consistente se

$$\rho(1) = 0, \quad \rho'(1) = \sigma(1)$$

La consistenza, però, non è sufficiente ad assicurare la convergenza di un metodo. Consideriamo l'applicazione del metodo BDF2 (17.8) e del metodo a due passi del secondo ordine (17.7) al semplice problema differenziale

$$\begin{cases} y'(t) = 0, & t > 0 \\ y(0) = e \end{cases} \quad (17.9)$$

la cui soluzione è evidentemente $y(t) \equiv e$. L'applicazione del secondo metodo

step	BDF2 (17.8)	metodo (17.7)
y_2	2.7182818284590451e+00	2.7182818284590455e+00
y_3	2.7182818284590451e+00	2.7182818284590473e+00
y_{10}	2.7182818284590451e+00	2.7182818284593857e+00
y_{20}	2.7182818284590451e+00	2.7182818288082906e+00
y_{30}	2.7182818284590451e+00	2.7182821860869133e+00
y_{40}	2.7182818284590451e+00	2.7186480393965446e+00
y_{50}	2.7182818284590451e+00	3.0932818284590446e+00
y_{60}	2.7182818284590451e+00	3.8671828182845888e+02
y_{70}	2.7182818284590451e+00	3.9321871828182810e+05
y_{80}	2.7182818284590451e+00	4.0265318671828151e+08
y_{90}	2.7182818284590451e+00	4.1231686041871790e+11
y_{100}	2.7182818284590451e+00	4.2221246506598656e+14

Figura 17.1: Primi passi di risoluzione del problema test (17.9).

porge

$$y_{n+2} - 3y_{n+1} + 2y_n = 0$$

Dunque $y_0 = e$ e prendiamo pure $y_1 = e$. Allora, $y_2 = 3e - 2e$ che, a causa degli errori di arrotondamento, vale $e + \varepsilon$, $\varepsilon < 9 \cdot 10^{-16}$. Poi, $y_3 = 3y_2 - 2y_1 = 3(e + \varepsilon) - 2e$. Anche senza introdurre ulteriori errori di arrotondamento (che, in realtà, ci sono), si ha $y_3 = e + 3\varepsilon$. In generale, si ha

$$y_n = e + (2^{n-1} - 1)\varepsilon$$

Dunque, se il numero di passi è $m = t^*/k$, si ha

$$\lim_{k \rightarrow 0} y_m = \lim_{m \rightarrow \infty} y_m = \infty$$

come ben si vede dalla Tabella 17.1. 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 che si accumula in maniera distruttiva. Infatti, posto, come al solito,

$$\mathbf{y}_{n+s}^* + \sum_{j=0}^{s-1} a_j \mathbf{y}(t_{n+j}) = k \sum_{j=0}^{s-1} b_j \mathbf{f}(t_{n+j}, \mathbf{y}(t_{n+j})) + kb_s \mathbf{f}(t_{n+s}, \mathbf{y}_{n+s}^*)$$

l'errore al passo $n + s$ può essere espresso come

$$\mathbf{e}_{n+s} = \mathbf{y}_{n+s} - \mathbf{y}(t_{n+s}) = (\mathbf{y}_{n+s} - \mathbf{y}_{n+s}^*) + (\mathbf{y}_{n+s}^* - \mathbf{y}(t_{n+s})) \quad (17.10)$$

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. È giustificata allora (in analogia con quanto visto al paragrafo 10.2.7) la seguente

Definizione 4. Dato un metodo lineare multistep (17.5), siano \mathbf{z}_n^i , $i = 1, 2$, due perturbazioni della soluzione definite da

$$\begin{aligned} \mathbf{z}_j^i &= \mathbf{y}_j + \boldsymbol{\delta}_j^i, & 0 \leq j \leq s-1 \\ \sum_{j=0}^s a_j \mathbf{z}_{n+j}^i &= k \sum_{j=0}^s b_j \mathbf{f}(t_{n+j}, \mathbf{z}_{n+j}^i) + \boldsymbol{\delta}_{n+s}^i, & 0 \leq n \leq m-s \end{aligned}$$

Se da $\max_n \|\boldsymbol{\delta}_n^1 - \boldsymbol{\delta}_n^2\| \leq \varepsilon$ segue $\max_n \|\mathbf{z}_n^1 - \mathbf{z}_n^2\| \leq C\varepsilon$ per k piccolo a piacere, allora il metodo (17.5) si dice (zero-)stabile.

Guardando la rappresentazione dell'errore (17.10), si vede che il primo termine $(\mathbf{y}_{n+s} - \mathbf{y}_{n+s}^*)$ è la differenza tra due particolari soluzioni perturbate \mathbf{z}_n^1 e \mathbf{z}_n^2 corrispondenti a

$$\boldsymbol{\delta}_j^1 = 0, \quad 0 \leq j \leq m$$

e

$$\begin{aligned} \boldsymbol{\delta}_j^2 &= \mathbf{y}(t_j) - \mathbf{y}_j, & 0 \leq j \leq s-1 \\ \boldsymbol{\delta}_{n+s}^2 &= -k \sum_{j=0}^s b_j \mathbf{f}(t_{n+j}, \mathbf{z}_{n+j}^2) + \\ &+ k \sum_{j=0}^{s-1} b_j \mathbf{f}(t_{n+j}, \mathbf{y}(t_{n+j})) + kb_s \mathbf{f}(t_{n+s}, \mathbf{y}_{n+s}^*) + \\ &+ \sum_{j=0}^s a_j \mathbf{z}_{n+j}^2 - \mathbf{y}_{n+s}^* - \sum_{j=0}^{s-1} a_j \mathbf{y}(t_{n+j}), & 0 \leq n \leq m-s \end{aligned}$$

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 arrotondamento e/o approssimazione) rimangano limitate (cioè non fa esplodere la prima parte di (17.10)) e che l'errore locale tenda a zero con k (cioè fa andare a zero la seconda parte di (17.10)). Cerchiamo un criterio semplice per determinare se un metodo è stabile. Consideriamo ancora il problema visto all'inizio di questo paragrafo. Supponiamo che θ sia una radice del polinomio $\rho(w)$ (che d'ora in poi chiameremo *caratteristico*) di un metodo multistep almeno consistente. Allora $z_n = k^p \theta^n + e$ ($p \geq 1$) è una soluzione numerica (in un metodo multistep a s stadi vi è una certa libertà di costruire i primi z_n , $0 \leq n \leq s-1$, da cui la *non* unicità di soluzione). Infatti

$$\sum_{j=0}^s a_j z_{n+j} = k^p \theta^n \sum_{j=0}^s a_j \theta^j + e \sum_{j=0}^s a_j = k^p \theta^n \rho(\theta) + 0 = 0$$

Essa è generata dai valori iniziali $y_j = e$, $0 \leq j \leq s-1$, *perturbati* dalla quantità $\delta_j = k^p \theta^j$, perfettamente lecita dal punto di vista della consistenza, e supponendo che non intervengano ulteriori perturbazioni. Se vogliamo che z_m converga alla soluzione analitica $y(t) \equiv e$ per $k = t^*/m \rightarrow 0$ (o, equivalentemente, per $m \rightarrow \infty$), deve essere $|\theta| \leq 1$. Se θ è radice multipla di $\rho(w)$, allora è radice anche di $\rho'(w)$ e pertanto soddisfa

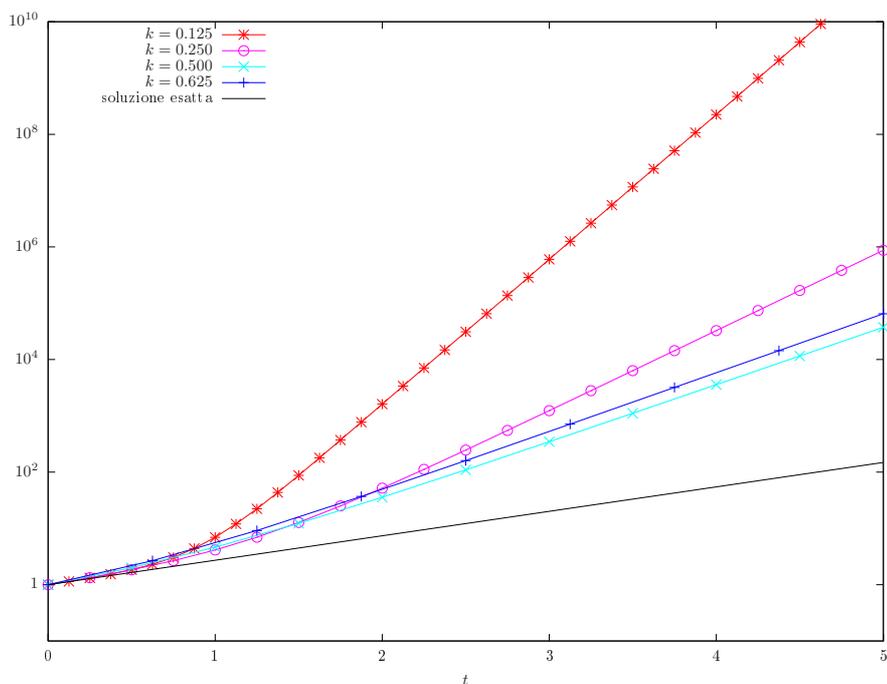
$$\rho'(\theta) = \sum_{j=0}^s a_j j \theta^{j-1} = 0$$

Allora anche $z_n = k^p n \theta^n + e$ è una soluzione numerica. Infatti

$$\begin{aligned} \sum_{j=0}^s a_j z_{n+j} &= k^p \left(\theta^n \sum_{j=0}^s a_j n \theta^j + \theta^{n+1} \sum_{j=0}^s a_j j \theta^{j-1} \right) + \sum_{j=0}^s a_j = \\ &= k^p \theta^n n \rho(\theta) + k^p \theta^{n+1} \rho'(\theta) + 0 = 0 \end{aligned}$$

ed è generata dai valori iniziali $y_j = e$ perturbati da $\delta_j = k^p j \theta^j$, $0 \leq j \leq s-1$. Se vogliamo che z_m converga alla soluzione analitica, 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*. Da notare che la condizione delle radici risulta necessaria per la stabilità anche per perturbazioni che siano $\mathcal{O}(k^p)$ e non solo $\mathcal{O}(1)$ (come per esempio gli errori di arrotondamento). La condizione delle radici risulta anche sufficiente per avere la stabilità e, assieme alla consistenza, per la convergenza del metodo lineare multistep. Si ha infatti il seguente teorema fondamentale:

Teorema 10 (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).*



msinstabile.m

Figura 17.2: Applicazione del metodo (17.7) al problema differenziale $y'(t) = y(t)$, $y(0) = 1$.

Ritornando al metodo (17.7), si ha che $\theta = 2$ è radice del polinomio caratteristico e infatti $y_n = k^2 2^n + e$ è una soluzione del problema differenziale che abbiamo usato come modello. Pertanto, il metodo non è stabile, come si vede anche in Figura 17.2 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 11 (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 $1 \leq s \leq 6$.

17.4 Influenza degli errori di arrotondamento

Abbiamo visto che la stabilità assicura che perturbazioni consistenti di ordine $\mathcal{O}(k^p)$ non impediscono ad un metodo consistente di convergere. Tali perturbazioni si generano, 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

$$\tilde{\mathbf{y}}_{n+1} = \tilde{\mathbf{y}}_n + k\mathbf{f}(t_n, \tilde{\mathbf{y}}_n) + \varepsilon_{n+1}$$

e dunque

$$\|\mathbf{e}_n\| \leq \frac{ck^2 + \varepsilon}{k\lambda} [(1 + k\lambda)^n - 1] + (1 + k\lambda)^n \|\mathbf{e}_0\|, \quad 0 \leq n \leq m$$

ove $\varepsilon = \max_{0 \leq n \leq m} \varepsilon_n$ e quindi

$$\|\mathbf{e}_n\| \leq e^{t^*\lambda} \|\mathbf{e}_0\| + (e^{t^*\lambda} - 1) \left(\frac{ck}{\lambda} + \frac{\varepsilon}{k\lambda} \right), \quad 0 \leq n \leq m$$

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.

Capitolo 18

Metodi di Runge–Kutta

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

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{k}{2}(\mathbf{f}(t_n, \mathbf{y}_n) + \mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}))$$

per renderlo esplicito è sostituire \mathbf{y}_{n+1} con $\mathbf{y}_n + k\mathbf{f}(t_n, \mathbf{y}_n)$ così da avere

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{k}{2}(\mathbf{f}(t_n, \mathbf{y}_n) + \mathbf{f}(t_{n+1}, \mathbf{y}_n + k\mathbf{f}(t_n, \mathbf{y}_n))) \quad (18.1)$$

Da un punto di vista “logico”, esso può essere definito come

$$\begin{aligned} \boldsymbol{\xi}_1 &= \mathbf{y}_n \approx \mathbf{y}(t_n) \\ \boldsymbol{\xi}_2 &= \mathbf{y}_n + k\mathbf{f}(t_n, \boldsymbol{\xi}_1) \approx \mathbf{y}(t_{n+1}) \\ \mathbf{y}_{n+1} &= \mathbf{y}_n + \frac{k}{2}(\mathbf{f}(t_n, \boldsymbol{\xi}_1) + \mathbf{f}(t_{n+1}, \boldsymbol{\xi}_2)) \approx \mathbf{y}(t_{n+1}) \end{aligned}$$

Un altro modo di rendere esplicito il metodo dei trapezi è sostituire la media delle funzioni \mathbf{f} con la funzione \mathbf{f} valutata nel “punto medio”

$$\mathbf{y}_{n+1} = \mathbf{y}_n + k\mathbf{f}\left(t_n + \frac{k}{2}, \mathbf{y}_n + \frac{k}{2}\mathbf{f}(t_n, \mathbf{y}_n)\right) \quad (18.2)$$

cioè

$$\begin{aligned}\boldsymbol{\xi}_1 &= \mathbf{y}_n \approx \mathbf{y}(t_n) \\ \boldsymbol{\xi}_2 &= \mathbf{y}_n + \frac{k}{2} \mathbf{f}(t_n, \boldsymbol{\xi}_1) \approx \mathbf{y}\left(t_n + \frac{k}{2}\right) \\ \mathbf{y}_{n+1} &= \mathbf{y}_n + k \mathbf{f}\left(t_n + \frac{k}{2}, \boldsymbol{\xi}_2\right) \approx \mathbf{y}(t_{n+1})\end{aligned}$$

L'idea generale dei metodi espliciti di *Runge-Kutta* è quella, come al solito, di sostituire l'integrale nella formula risolutiva

$$\mathbf{y}(t_{n+1}) = \mathbf{y}(t_n) + \int_{t_n}^{t_{n+1}} \mathbf{f}(\tau, \mathbf{y}(\tau)) d\tau$$

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

$$\mathbf{y}(t_{n+1}) \approx \mathbf{y}(t_n) + k \sum_{j=1}^{\nu} b_j \mathbf{f}(t_n + c_j k, \mathbf{y}(t_n + c_j k))$$

Si tratta ora di trovare delle approssimazioni $\boldsymbol{\xi}_j$ di $\mathbf{y}(t_n + c_j k)$. Si procede iterativamente in questo modo

$$\left\{ \begin{array}{l} \mathbf{y}(t_n) \approx \mathbf{y}_n = \boldsymbol{\xi}_1 \quad (\Rightarrow c_1 = 0) \\ \mathbf{y}(t_n + c_2 k) \approx \mathbf{y}_n + c_2 k \mathbf{f}(t_n, \mathbf{y}_n) = \mathbf{y}_n + a_{2,1} k \mathbf{f}(t_n, \boldsymbol{\xi}_1) = \boldsymbol{\xi}_2 \\ \vdots \\ \mathbf{y}(t_n + c_i k) \approx \mathbf{y}_n + k \sum_{j=1}^{i-1} a_{i,j} \mathbf{f}(t_n + c_j k, \boldsymbol{\xi}_j) = \boldsymbol{\xi}_i \\ \vdots \\ \mathbf{y}(t_n + c_\nu k) \approx \mathbf{y}_n + k \sum_{j=1}^{\nu-1} a_{\nu,j} \mathbf{f}(t_n + c_j k, \boldsymbol{\xi}_j) = \boldsymbol{\xi}_\nu \\ \mathbf{y}_{n+1} = \mathbf{y}_n + k \sum_{j=1}^{\nu} b_j \mathbf{f}(t_n + c_j k, \boldsymbol{\xi}_j) \end{array} \right. \quad (18.3)$$

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

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_ν	$a_{\nu,1}$	$a_{\nu,2}$	\dots	$a_{\nu,\nu-2}$	$a_{\nu,\nu-1}$	
	b_1	b_2	\dots	$b_{\nu-2}$	$b_{\nu-1}$	b_ν

Tabella 18.1: Tableau di Butcher per metodi di Runge–Kutta espliciti.

metodo di Eulero. Per $\nu = 2$, l'ordine si ricava al solito modo

$$\begin{aligned}
& \mathbf{y}(t_{n+1}) - \mathbf{y}(t_n) - kb_1 \mathbf{f}(t_n, \mathbf{y}(t_n)) - kb_2 \mathbf{f}(t_n + c_2 k, \mathbf{y}(t_n) + a_{2,1} k \mathbf{f}(t_n, \mathbf{y}(t_n))) = \\
& = \mathbf{y}(t_n) + k \mathbf{y}'(t_n) + \frac{k^2}{2} \mathbf{y}''(t_n) + \mathcal{O}(k^3) - \mathbf{y}(t_n) - kb_1 \mathbf{y}'(t_n) + \\
& - kb_2 \left[\mathbf{f}(t_n, \mathbf{y}(t_n)) + \frac{\partial \mathbf{f}}{\partial t}(t_n, \mathbf{y}(t_n)) c_2 k + \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(t_n, \mathbf{y}(t_n)) a_{2,1} k \mathbf{y}'(t_n) + \mathcal{O}(k^2) \right] = \\
& = k \mathbf{y}'(t_n) + \frac{k^2}{2} \mathbf{y}''(t_n) + \mathcal{O}(k^3) - kb_1 \mathbf{y}'(t_n) - kb_2 \mathbf{y}'(t_n) + \\
& - kb_2 \frac{\partial \mathbf{f}}{\partial t}(t_n, \mathbf{y}(t_n)) c_2 k - k^2 a_{2,1} b_2 \left[\frac{\partial \mathbf{f}}{\partial \mathbf{y}}(t_n, \mathbf{y}(t_n)) \mathbf{y}'(t_n) \right] = \\
& = k \mathbf{y}'(t_n) + \frac{k^2}{2} \mathbf{y}''(t_n) + \mathcal{O}(k^3) - kb_1 \mathbf{y}'(t_n) - kb_2 \mathbf{y}'(t_n) + \\
& - kb_2 \frac{\partial \mathbf{f}}{\partial t}(t_n, \mathbf{y}(t_n)) c_2 k - k^2 a_{2,1} b_2 \left[\mathbf{y}''(t_n) - \frac{\partial \mathbf{f}}{\partial t}(t_n, \mathbf{y}(t_n)) \right] = \\
& = k(1 - b_1 - b_2) \mathbf{y}'(t_n) + k^2 \left(\frac{1}{2} - a_{2,1} b_2 \right) \mathbf{y}''(t_n) + \\
& - k^2 (b_2 c_2 - a_{2,1} b_2) \frac{\partial \mathbf{f}}{\partial t}(t_n, \mathbf{y}(t_n)) + \mathcal{O}(k^3)
\end{aligned}$$

Dunque l'ordine è due se

0					
1	1				
	$\frac{1}{2}$	$\frac{1}{2}$			

0				
$\frac{1}{2}$	$\frac{1}{2}$			
	0	1		

0			
$\frac{2}{3}$	$\frac{2}{3}$		
	$\frac{1}{4}$	$\frac{3}{4}$	

Tabella 18.2: Metodi Runge–Kutta espliciti di ordine 2.

$$\begin{cases} b_1 + b_2 = 1 \\ a_{2,1}b_2 = \frac{1}{2} \\ b_2c_2 = b_2a_{2,1} \end{cases} \quad (18.4)$$

da cui, per esempio, i metodi di ordine due riportati in Tabella 18.2. 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). Il punto cruciale di questo sviluppo in serie di Taylor è l'uguaglianza tra

$$\begin{aligned} \mathbf{y}'(t_n) + k\mathbf{y}''(t_n) &= \mathbf{y}'(t_n) + k \frac{d}{dt} \mathbf{f}(t_n, \mathbf{y}(t_n)) = \\ &= \mathbf{y}'(t_n) + k \frac{\partial \mathbf{f}}{\partial t}(t_n, \mathbf{y}(t_n)) + k \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(t_n, \mathbf{y}(t_n)) \mathbf{y}'(t_n) \end{aligned}$$

e

$$\begin{aligned} \mathbf{f}(t_n + k, \mathbf{y}(t_n) + k\mathbf{f}(t_n, \mathbf{y}(t_n))) &= \mathbf{f}(t_n, \mathbf{y}(t_n)) + \\ &+ \frac{\partial \mathbf{f}}{\partial t}(t_n, \mathbf{y}(t_n))k + \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(t_n, \mathbf{y}(t_n))k\mathbf{f}(t_n, \mathbf{y}(t_n)) + \mathcal{O}(k^2) = \\ &= \mathbf{y}'(t_n) + k \frac{\partial \mathbf{f}}{\partial t}(t_n, \mathbf{y}(t_n)) + k \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(t_n, \mathbf{y}(t_n)) \mathbf{y}'(t_n) + \mathcal{O}(k^2) \end{aligned}$$

(a meno di $\mathcal{O}(k^2)$) in cui le derivate di ordine superiore di \mathbf{y} (e quindi di \mathbf{f}) sono sostituite da funzioni di funzioni \mathbf{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

$$\sum_{j=1}^{i-1} a_{i,j} = c_i, \quad 2 \leq i \leq \nu$$

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 18.3.

numero stadi ν	1	2	3	4	5	6	7	8
massimo ordine p	1	2	3	4	4	5	6	6

Tabella 18.3: Massimo ordine dato il numero di stadi.

Il numero di stadi equivale al numero di valutazioni della funzione \mathbf{f} (e dunque al costo del metodo). Ai fini dell'implementazione, per evitare di calcolare più volte la funzione \mathbf{f} negli stessi punti, si usa lo schema

$$\begin{cases} \mathbf{f}_i = \mathbf{f} \left(t_n + c_i k, \mathbf{y}_n + k \sum_{j=1}^{i-1} a_{i,j} \mathbf{f}_j \right), & i = 1, \dots, \nu \\ \mathbf{y}_{n+1} = \mathbf{y}_n + k \sum_{j=1}^{\nu} b_j \mathbf{f}_j \end{cases}$$

Si ricava da (18.3) ponendo $\mathbf{f}_j = \mathbf{f}(t_n + c_j k, \boldsymbol{\xi}_j)$ e osservando che $\mathbf{f}_i = \mathbf{f}(t_n + c_i k, \boldsymbol{\xi}_i) = \mathbf{f}(t_n + c_i k, \mathbf{y}_n + k \sum_{j=1}^{i-1} a_{i,j} \mathbf{f}_j)$, $i = 1, \dots, \nu$.

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_{ν}	$a_{\nu,1}$	$a_{\nu,2}$	\dots	$a_{\nu,\nu-2}$	$a_{\nu,\nu-1}$	$a_{\nu,\nu}$
	b_1	b_2	\dots	$b_{\nu-2}$	$b_{\nu-1}$	b_{ν}
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_{ν}	$a_{\nu,1}$	$a_{\nu,2}$	\dots	\dots	$a_{\nu,\nu-1}$	$a_{\nu,\nu}$
	b_1	b_2	\dots	\dots	$b_{\nu-1}$	b_{ν}

Tabella 18.4: Tableaux di Butcher per i metodi di Runge–Kutta semiimpliciti (sopra) e impliciti (sotto).

È possibile generalizzare i metodi espliciti di Runge–Kutta per ottenere metodi *semiimpliciti* e impliciti, i cui tableaux sono riportati in Tabella 18.4. Per tali metodi, l'ordine massimo raggiungibile dato il numero di stadi ν è

$p = 2\nu$. Anche per essi valgono le condizioni

$$\sum_{j=1}^{\nu} a_{i,j} = c_i, \quad 1 \leq i \leq \nu$$

e

$$\sum_{j=1}^{\nu} b_j = 1$$

18.1 Un esempio di metodo Runge–Kutta implicito

Consideriamo il seguente metodo di Runge–Kutta implicito

$$\boldsymbol{\xi}_1 = \mathbf{y}_n + \frac{k}{2} \mathbf{f} \left(t_n + \frac{k}{2}, \boldsymbol{\xi}_1 \right) \quad (18.5a)$$

$$\mathbf{y}_{n+1} = \mathbf{y}_n + k \mathbf{f} \left(t_n + \frac{k}{2}, \boldsymbol{\xi}_1 \right) \quad (18.5b)$$

di tableau

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

Esso equivale al metodo *punto medio implicito*

$$\mathbf{y}_{n+1} = \mathbf{y}_n + k \mathbf{f} \left(t_n + \frac{k}{2}, \frac{\mathbf{y}_n + \mathbf{y}_{n+1}}{2} \right) \quad (18.6)$$

infatti, per quest'ultimo vale

$$\frac{\mathbf{y}_n + \mathbf{y}_{n+1}}{2} = \mathbf{y}_n + \frac{k}{2} \mathbf{f} \left(t_n + \frac{k}{2}, \frac{\mathbf{y}_n + \mathbf{y}_{n+1}}{2} \right)$$

da cui si deduce che $(\mathbf{y}_n + \mathbf{y}_{n+1})/2 = \boldsymbol{\xi}_1$ e dunque (18.5b) coincide con (18.6). È un metodo di ordine 2 (lo si dimostra a partire da (18.6) con conti 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

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t))$$

$\mathbf{y}(t)^T \mathbf{y}(t)$ è costante, cioè $\mathbf{y}(t)^T \mathbf{f}(t, \mathbf{y}(t)) = 0$, allora per (18.6) vale $\mathbf{y}_{n+1}^T \mathbf{y}_{n+1} = \mathbf{y}_n^T \mathbf{y}_n$. Prima di dimostrarlo, osserviamo che la proprietà è interessante quando $\mathbf{y}(t)$ è un vettore di dimensione maggiore di uno, altrimenti deve essere

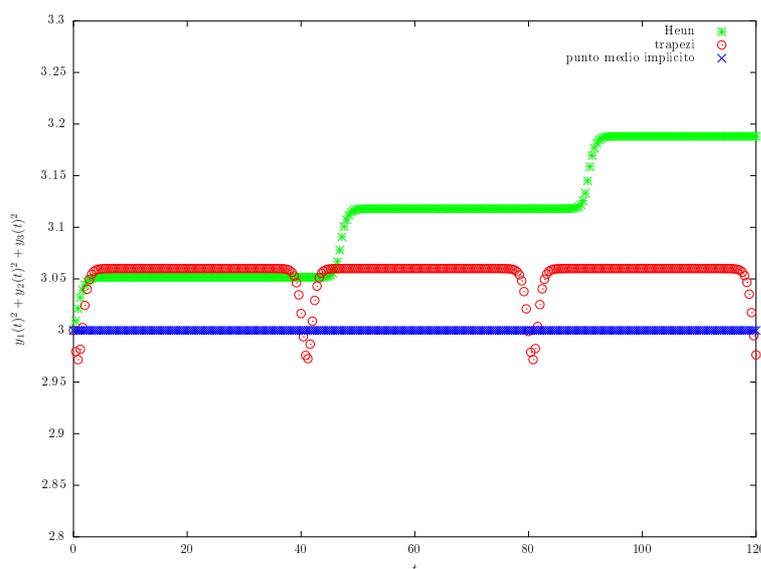
$\mathbf{f} = 0$ e dunque banalmente $\mathbf{y}_{n+1} = \mathbf{y}_n$. Poiché $\mathbf{y}(t)^\top \mathbf{f}(t, \mathbf{y}(t)) = 0$, allora anche $\mathbf{x}^\top \mathbf{f}(t_n + k/2, \mathbf{x}) = 0$, \mathbf{x} generico. Infatti, esiste sicuramente $\mathbf{y}(t_0)$ tale che per $t = t_n + k/2$ si ha $\mathbf{y}(t) = \mathbf{x}$ (per trovarlo, basta partire da $\mathbf{y}(t_n + k/2) = \mathbf{x}$ e risolvere all'indietro il sistema fino al tempo t_0). Osserviamo poi che

$$\mathbf{y}_{n+1} - \mathbf{y}_n = k\mathbf{f}\left(t_n + \frac{k}{2}, \frac{\mathbf{y}_n + \mathbf{y}_{n+1}}{2}\right)$$

e quindi

$$0 = \left(\frac{\mathbf{y}_n + \mathbf{y}_{n+1}}{2}\right)^\top k\mathbf{f}\left(t_n + \frac{k}{2}, \frac{\mathbf{y}_n + \mathbf{y}_{n+1}}{2}\right) = k\left(\frac{\mathbf{y}_n + \mathbf{y}_{n+1}}{2}\right)^\top (\mathbf{y}_{n+1} - \mathbf{y}_n)$$

da cui la tesi.



symplectic.m

Figura 18.1: Evoluzione di $y_1(t)^2 + y_2(t)^2 + y_3(t)^2$ per il problema 18.7.

Consideriamo il seguente problema (equazioni di Eulero per il corpo rigido libero):

$$\begin{cases} y_1'(t) = \frac{I_2 - I_3}{I_2 I_3} y_2(t) y_3(t) \\ y_2'(t) = \frac{I_3 - I_1}{I_3 I_1} y_3(t) y_1(t) \\ y_3'(t) = \frac{I_1 - I_2}{I_1 I_2} y_1(t) y_2(t) \end{cases} \quad (18.7)$$

\mathbf{y} rappresenta il momento angolare e I_1 , I_2 e I_3 sono i momenti principali d'inerzia. Evidentemente $|\mathbf{y}(t)|^2 = y_1(t)^2 + y_2(t)^2 + y_3(t)^2$ 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 ($I_1 = 2$, $I_2 = 1$, $I_3 = 2/3$, $\mathbf{y}(0) = [1, 1, 1]^T$) produce il grafico in Figura 18.1, in cui si vede che il metodo punto medio implicito conserva "esattamente" la quantità $|\mathbf{y}(t)|^2$, 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 $\mathbf{y}_{n+1}^T \mathbf{y}_{n+1} = \mathbf{y}_n^T \mathbf{y}_n$ anche per il metodo punto medio implicito. I tre metodi testati sono dello stesso ordine, ma uno produce soluzioni "qualitativamente" migliori.

È buona norma, dopo aver definito la matrice \mathbf{A} e i vettori \mathbf{c} e \mathbf{b} del tableau, controllare che queste condizioni siano soddisfatte, confrontando $\text{sum}(\mathbf{A}, 2)$ con \mathbf{c} e $\text{sum}(\mathbf{b})$ con 1, al fine di evitare banali errori. Ovviamente, si tratta di condizioni necessarie ma *non sufficienti* per garantire la corretta implementazione.

Anche il θ -metodo può essere fatto rientrare nella classe dei metodi di Runge–Kutta semiimpliciti:

$$\begin{cases} \boldsymbol{\xi}_1 = \mathbf{y}_n \\ \boldsymbol{\xi}_2 = \mathbf{y}_n + k(1 - \theta)\mathbf{f}(t_n, \boldsymbol{\xi}_1) + k\theta\mathbf{f}(t_n + k, \boldsymbol{\xi}_2) \\ \mathbf{y}_{n+1} = \mathbf{y}_n + k(1 - \theta)\mathbf{f}(t_n, \boldsymbol{\xi}_1) + k\theta\mathbf{f}(t_n + k, \boldsymbol{\xi}_2) \end{cases}$$

o, in forma *implementativa* (anche se non si usa in pratica),

$$\begin{cases} \mathbf{f}_1 = \mathbf{f}(t_n, \mathbf{y}_n) \\ \mathbf{f}_2 = \mathbf{f}(t_n + k, \mathbf{y}_n + k(1 - \theta)\mathbf{f}_1 + k\theta\mathbf{f}_2) \\ \mathbf{y}_{n+1} = \mathbf{y}_n + k(1 - \theta)\mathbf{f}_1 + k\theta\mathbf{f}_2 \end{cases}$$

Dunque, abbiamo risposto anche alla prima domanda all'inizio di questo capitolo. Vediamo ora come rispondere alla terza domanda.

18.2 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'aver 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

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	$a_{\nu-1,\nu-2}$	
	b_1	b_2	\dots	$b_{\nu-2}$	$b_{\nu-1}$
0	$a_{2,1}$				
c_2	$a_{3,1}$	$a_{3,2}$			
c_3	$a_{3,1}$	$a_{3,2}$	\dots	$a_{\nu-1,\nu-2}$	
\vdots	\vdots	\vdots	\ddots		
$c_{\nu-1}$	$a_{\nu-1,1}$	$a_{\nu-1,2}$	\dots	$a_{\nu-1,\nu-2}$	
c_ν	$a_{\nu,1}$	$a_{\nu,2}$	\dots	$a_{\nu,\nu-2}$	$a_{\nu,\nu-1}$
	\hat{b}_1	\hat{b}_2	\dots	$\hat{b}_{\nu-2}$	$\hat{b}_{\nu-1}$
					\hat{b}_ν

Tabella 18.5: Metodi di Runge–Kutta di ordine $p - 1$ e p .

rispettivamente, i cui tableaux sono riportati in Tabella 18.5. È chiaro che, dopo aver costruito il primo metodo, con una sola nuova valutazione della funzione \mathbf{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 18.6. 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.

0					
c_2	$a_{2,1}$				
c_3	$a_{3,1}$	$a_{3,2}$			
\vdots	\vdots	\vdots	\ddots		
c_ν	$a_{\nu,1}$	$a_{\nu,2}$	\dots	$a_{\nu,\nu-1}$	
	b_1	b_2	\dots	$b_{\nu-1}$	
	\hat{b}_1	\hat{b}_2	\dots	$\hat{b}_{\nu-1}$	\hat{b}_ν

Tabella 18.6: Metodi di Runge–Kutta embedded di ordine $p - 1$ e p .

Consideriamo il sistema differenziale

$$\begin{cases} \tilde{\mathbf{y}}'(t) = \mathbf{f}(t, \tilde{\mathbf{y}}(t)) \\ \tilde{\mathbf{y}}(t_n) = \mathbf{y}_n^{(p)} \end{cases}$$

ove $\mathbf{y}_n^{(p)}$ è l'approssimazione di $\mathbf{y}(t_n)$ ottenuta con il metodo Runge–Kutta di

ordine p . Si ha allora

$$\|\mathbf{y}_{n+1}^{(p)} - \mathbf{y}_{n+1}^{(p-1)}\| = \|\mathbf{y}_{n+1}^{(p)} - \tilde{\mathbf{y}}(t_{n+1}) + \tilde{\mathbf{y}}(t_{n+1}) - \mathbf{y}_{n+1}^{(p-1)}\| = Ck_{n+1}^p, \quad (18.8)$$

per un opportuno $C > 0$, ove $k_{n+1} = t_{n+1} - t_n$ è il passo di integrazione e Ck_{n+1}^p è l'errore locale del metodo di ordine $p - 1$. Se si vuole controllare tale errore si può allora imporre, ad ogni passo, che

$$\|\mathbf{y}_{n+1}^{(p)} - \mathbf{y}_{n+1}^{(p-1)}\| \leq \text{tol}_a + \|\mathbf{y}_{n+1}^{(p-1)}\| \cdot \text{tol}_r \quad (18.9)$$

rifutando $\mathbf{y}_{n+1}^{(p)}$ (e $\mathbf{y}_{n+1}^{(p-1)}$) nel caso la disuguaglianza non sia soddisfatta e calcolando un nuovo passo di integrazione minore del precedente. Per calcolare il successivo passo di integrazione k , sia nel caso che la disuguaglianza sia stata soddisfatta o meno, imponiamo che valga

$$Ck^p = \text{tol}_a + \|\mathbf{y}_{n+1}^{(p-1)}\| \cdot \text{tol}_r$$

e, ricavando $1/C$ da (18.8), si ha

$$k = \left(\frac{\text{tol}_a + \|\mathbf{y}_{n+1}^{(p-1)}\| \cdot \text{tol}_r}{\|\mathbf{y}_{n+1}^{(p)} - \mathbf{y}_{n+1}^{(p-1)}\|} \right)^{1/p} \cdot k_{n+1}$$

Tale espressione indica il nuovo passo temporale $k_{n+2} = k$ nel caso la disuguaglianza (18.9) sia stata soddisfatta oppure la nuova misura del vecchio $k_{n+1} = k$ nel caso contrario. Per evitare che il passo di integrazione cambi troppo bruscamente, si può adottare una correzione del tipo

$$k = \min \left(2, \max \left(0.6, 0.9 \cdot \left(\frac{\text{tol}_a + \|\mathbf{y}_{n+1}^{(p-1)}\| \cdot \text{tol}_r}{\|\mathbf{y}_{n+1}^{(p)} - \mathbf{y}_{n+1}^{(p-1)}\|} \right)^{1/p} \right) \right) \cdot k_{n+1}$$

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 $\mathbf{f}_1 = \mathbf{f}(t_n, \mathbf{y}_n)$ 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

$$\begin{array}{c|cc} 0 & & \\ c_2 & a_{2,1} & \\ \hline & \hat{b}_1 & \hat{b}_2 \end{array}$$

può essere implementato a passo variabile secondo lo schema (da t_n a t_{n+1})

- $\mathbf{f}_1 = \mathbf{f}(t_n, \mathbf{y}_n)$
- $\mathbf{y}_{n+1}^{(1)} = \mathbf{y}_n^{(2)} + k_{n+1} \mathbf{f}_1$ (metodo di Eulero)
- $\mathbf{f}_2 = \mathbf{f}(t_n + c_2 k_{n+1}, \mathbf{y}_n + a_{2,1} k_{n+1} \mathbf{f}_1)$
- $\mathbf{e}_{n+1}^{(2-1)} = k_{n+1} [(\hat{b}_1 - 1) \mathbf{f}_1 + \hat{b}_2 \mathbf{f}_2]$
- IF $\|\mathbf{e}_{n+1}^{(2-1)}\| > \text{tol}_a + \|\mathbf{y}_{n+1}^{(p-1)}\| \text{tol}_r$
 $n = n - 1$ (time step rifiutato)
- ELSE
- $$\mathbf{y}_{n+1}^{(2)} = \mathbf{y}_{n+1}^{(1)} + \mathbf{e}_{n+1}^{(2-1)}$$
- END
- $k_{n+2} = \left[(\text{tol}_a + \|\mathbf{y}_{n+1}^{(p-1)}\| \cdot \text{tol}_r) / \|\mathbf{e}_{n+1}^{(2-1)}\| \right]^{1/2} \cdot k_{n+1}$
- $n = n + 1$

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	$\frac{6656}{12825}$	$\frac{28561}{56430}$	$-\frac{9}{50} \quad \frac{2}{55}$

Tabella 18.7: 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 18.7.

Capitolo 19

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

$$\begin{cases} y'(t) = \lambda y(t) & t > t_0 \\ y(t_0) = y_0 \end{cases} \quad (19.1)$$

la cui soluzione esatta $y(t) = e^{\lambda(t-t_0)}y_0$ tende a zero, per $t \rightarrow +\infty$, se $\Re(\lambda) < 0$. Analizziamo il comportamento del metodo di Eulero per questo problema, supponendo di avere fissato il passo temporale k : 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

$$\lim_{n \rightarrow \infty} y_n = 0 \Leftrightarrow |1 + k\lambda| < 1 \Leftrightarrow 1 + k^2\Re(\lambda)^2 + 2k\Re(\lambda) + k^2\Im(\lambda)^2 < 1$$

da cui

$$\lim_{n \rightarrow \infty} y_n = 0 \Leftrightarrow k < -\frac{2\Re(\lambda)}{|\lambda|^2} \quad (19.2)$$

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 ($\lim_{n \rightarrow \infty} |y_n| = |y_0|$ o $\lim_{n \rightarrow \infty} y_n = \infty$). Nel caso di Eulero implicito, invece, si ha

$$y_n = \left(\frac{1}{1 - k\lambda} \right)^n y_0$$

da cui

$$\lim_{n \rightarrow \infty} y_n = 0 \Leftrightarrow |1 - k\lambda| > 1 \Leftrightarrow |1 - k\Re(\lambda) - ki\Im(\lambda)| > 1$$

disuguaglianza sempre soddisfatta, poiché $\Re(\lambda) < 0$. Anche per il metodo dei trapezi la soluzione numerica tende a 0 per $n \rightarrow \infty$. 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

$$\begin{aligned} \lim_{n \rightarrow \infty} y_n = 0 &\Leftrightarrow \left| \frac{1 + (1 - \theta)k\lambda}{1 - \theta k\lambda} \right| < 1 \Leftrightarrow |1 + (1 - \theta)k\lambda| < |1 - \theta k\lambda| \Leftrightarrow \\ 0 &< (\theta^2 - (1 - \theta)^2)k^2\Re(\lambda)^2 - (2\theta + 2(1 - \theta))k\Re(\lambda) + (\theta^2 - (1 - \theta)^2)k^2\Im(\lambda)^2 \end{aligned}$$

da cui

$$\lim_{n \rightarrow \infty} y_n = 0 \Leftrightarrow 0 < (2\theta - 1)k^2|\lambda|^2 - 2k\Re(\lambda)$$

Se $2\theta - 1 \geq 0$, certamente la disequazione è soddisfatta. Altrimenti,

$$\lim_{n \rightarrow \infty} y_n = 0 \Leftrightarrow k < \frac{2\Re(\lambda)}{(2\theta - 1)|\lambda|^2}, \quad (2\theta - 1 < 0) \quad (19.3)$$

Definizione 5. Dato un metodo numerico $\mathbf{y}_{n+1} = \mathcal{Y}_n(\mathbf{f}, k, \mathbf{y}_0, \dots, \mathbf{y}_n)$, la regione di assoluta stabilità (o linear stability domain) è l'insieme dei numeri $z = k\lambda$ per cui la soluzione di (19.1) soddisfa $\lim_{n \rightarrow \infty} \mathbf{y}_n = 0$.

Con riferimento al θ -metodo, la regione di assoluta stabilità del metodo di Eulero è $\{z \in \mathbb{C}: |1+z| < 1\}$, per Eulero implicito è $\{z \in \mathbb{C}: |1-z| > 1\}$ e per il metodo dei trapezi è $\{z \in \mathbb{C}: \Re(z) < 0\}$. Diremo che un metodo è *A-stabile* se la sua regione di assoluta stabilità contiene $\mathbb{C}^- = \{z \in \mathbb{C}: \Re(z) < 0\}$, cioè se riproduce correttamente il comportamento della soluzione analitica di (19.1) quando $\Re(\lambda) < 0$. Da notare che, indicato con $r(k\lambda)$ il termine (che dovrebbe essere in modulo minore di 1)

$$r(k\lambda) = \left[\frac{1 + (1 - \theta)k\lambda}{1 - \theta k\lambda} \right]$$

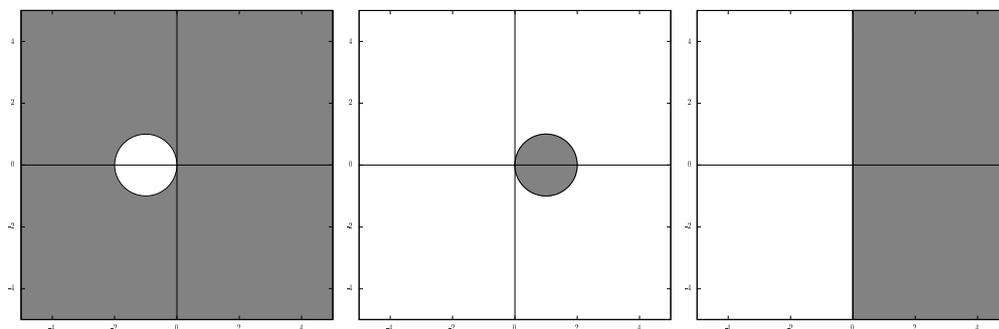


Figura 19.1: Regioni di assoluta stabilità (bianche) per i metodi di Eulero, Eulero implicito e trapezi.

si ha

$$\lim_{k\Re(\lambda) \rightarrow -\infty} |r(k\lambda)| = \left| \frac{\theta - 1}{\theta} \right|$$

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 ($\theta = 0$) anche se il metodo con $\theta = 2/3$, pur del primo ordine, ha una costante d'errore minore (vedi (16.5)). Per inciso, se λ è puramente immaginario $\lambda = \delta i$, $|y_n| \rightarrow +\infty$ per Eulero, $|y_n| \rightarrow 0$ per Eulero implicito e $|y_n| = 1 = |y(t_n)|$ per il metodo dei trapezi.

19.1 A-stabilità dei metodi di Runge–Kutta espliciti

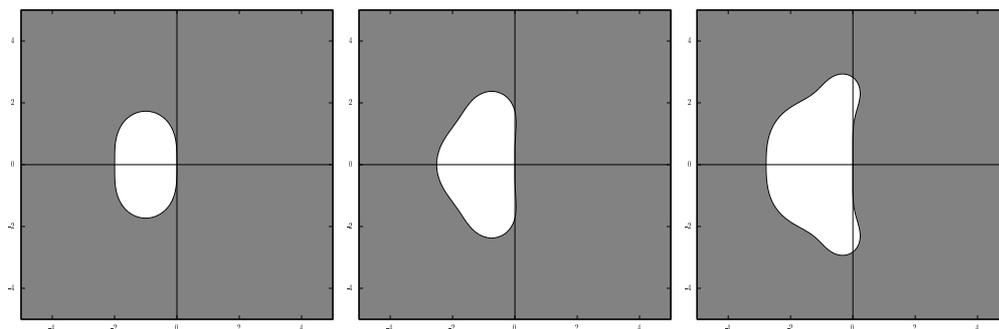


Figura 19.2: Regioni di assoluta stabilità (bianche) per i metodi di Runge–Kutta di ordine 2, 3 e 4.

Teorema 12. 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

$$r(z) = 1 + z + \frac{z^2}{2!} + \dots + \frac{z^\nu}{\nu!}$$

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 $j = 2, 3, \dots, \nu-1$: 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 $y(t_0 + k) = e^{k\lambda}y_0$. Quindi $r(k\lambda) - e^{k\lambda} = \mathcal{O}(k^{p+1})$ e dunque

$$r(k\lambda) = r(z) = \left(1 + z + \frac{z^2}{2!} + \dots + \frac{z^p}{p!}\right) + \mathcal{O}(z^{p+1})$$

da cui la tesi. □

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 Runge–Kutta esplicito $r(z)$ è un polinomio di grado ν (e dunque di grado maggiore di 0).

Teorema 13. Nessun metodo di Runge–Kutta esplicito è A-stabile.

Dimostrazione. Si ha

$$\lim_{n \rightarrow \infty} y_n = 0 \Leftrightarrow |r(z)| < 1, \quad z = k\lambda$$

ma $r(z)$ è un polinomio di grado maggiore di 0. Dunque, $\lim_{x \rightarrow -\infty} r(x) = \infty$, 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}^- . □

19.2 A-stabilità dei metodi lineari multistep

Ci limitiamo a riportare alcuni risultati.

Teorema 14. *Nessun metodo esplicito multistep è A-stabile.*

Teorema 15. *I metodi BDF ad un passo (Eulero implicito) e a due passi sono A-stabili.*

Teorema 16 (Seconda barriera di Dahlquist). *L'ordine più alto che un metodo multistep A-stabile può raggiungere è due.*

19.3 Equazioni stiff

Se consideriamo il problema

$$\begin{cases} y'(t) = -100y(t), & t > 0 \\ y(0) = 1 \end{cases}$$

la condizione (19.3) 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). Dunque, con poco più di 20 passi il metodo di Eulero arriva a calcolare la soluzione sino a t^* .

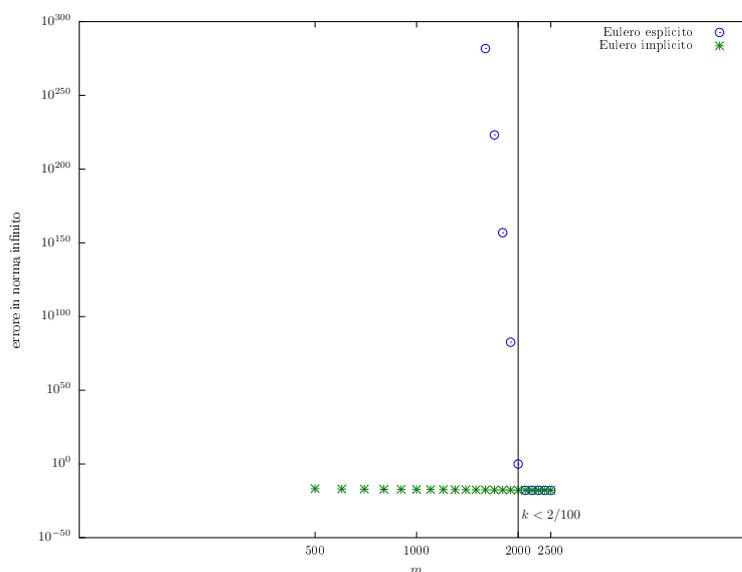
Qual è dunque il problema? Ecco:

$$\begin{cases} \mathbf{y}'(t) = \begin{bmatrix} -100 & 0 \\ 0 & -1 \end{bmatrix} \mathbf{y}(t), & t > 0 \\ \mathbf{y}(0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{cases} \quad (19.4)$$

La soluzione analitica è

$$\mathbf{y}(t) = \begin{bmatrix} e^{-100t} \\ e^{-t} \end{bmatrix}$$

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 19.3), 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



stiff.m

Figura 19.3: Eulero esplicito e Eulero implicito per la soluzione di (19.4) fino al tempo $t^* = 40$.

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, ma ci si può sempre ridurre, eventualmente in maniera approssimata, ad uno disaccoppiato e ragionare per componenti. Infatti, se A è una matrice diagonalizzabile,

$$\mathbf{y}'(t) = A\mathbf{y}(t) \Leftrightarrow \mathbf{z}'(t) = D\mathbf{z}(t) \Leftrightarrow \mathbf{z}(t) = \exp(tD)\mathbf{z}_0$$

ove $AV = VD$, $D = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_d\}$, e $\mathbf{y}(t) = V\mathbf{z}(t)$. Poi

$$\begin{aligned} \mathbf{y}'(t) = A\mathbf{y}(t) + \mathbf{b} &\Leftrightarrow \mathbf{z}'(t) = D\mathbf{z}(t) + V^{-1}\mathbf{b} \Leftrightarrow \\ &\Leftrightarrow \mathbf{z}(t) = \mathbf{z}_0 + t\varphi_1(tD)(D\mathbf{z}_0 + V^{-1}\mathbf{b}) \end{aligned}$$

ove

$$\varphi_1(\lambda) = \begin{cases} \frac{e^\lambda - 1}{\lambda} & \text{se } \lambda \neq 0 \\ 1 & \text{se } \lambda = 0 \end{cases}$$

Infine (considerando un problema autonomo per semplicità e sviluppando in serie di Taylor)

$$\mathbf{y}'(t) = \mathbf{f}(\mathbf{y}(t)) \Leftrightarrow \mathbf{y}'(t) \approx \mathbf{f}(t_n, \mathbf{y}_n) + J_n(\mathbf{y}(t) - \mathbf{y}_n)$$

ove J_n è la matrice jacobiana

$$J_n = \frac{\partial f_i}{\partial y_j}(\mathbf{y}_n)$$

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 (15.1) 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. Il metodo implicito richiede la soluzione ad ognuno dei “pochi” passi di un sistema, in generale, non lineare.

19.3.1 Risoluzione di un metodo implicito per problemi stiff

Consideriamo, per semplicità, il problema

$$\mathbf{y}'(t) = A\mathbf{y}(t)$$

con A stiff e *simmetrica*. La restrizione sul passo per il metodo di Eulero esplicito è

$$k < \frac{2}{\rho_{\max}}$$

ove ρ_{\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

$$\|A\mathbf{x} - A\mathbf{y}\|_2 \leq \|A\|_2 \|\mathbf{x} - \mathbf{y}\|_2 = \rho_{\max} \|\mathbf{x} - \mathbf{y}\|_2$$

si avrebbe la restrizione (vedi (16.6))

$$k < \frac{1}{\rho_{\max}}$$

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 20

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

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

La soluzione analitica è

$$\mathbf{y}(t) = \exp((t-t_0)A)\mathbf{y}_0 + (t-t_0)\varphi_1((t-t_0)A)\mathbf{b} = \mathbf{y}_0 + t\varphi_1((t-t_0)A)(A\mathbf{y}_0 + \mathbf{b})$$

Infatti $\mathbf{y}(t_0) = \mathbf{y}_0$ e

$$\begin{aligned} \mathbf{y}'(t) &= A \exp((t-t_0)A)\mathbf{y}_0 + \exp((t-t_0)A)\mathbf{b} = \\ &= A(\exp((t-t_0)A)\mathbf{y}_0 + (t-t_0)((t-t_0)A)^{-1} \exp((t-t_0)A)\mathbf{b} + \\ &\quad - (t-t_0)((t-t_0)A)^{-1}\mathbf{b} + A^{-1}\mathbf{b}) = \\ &= A(\exp((t-t_0)A)\mathbf{y}_0 + (t-t_0)\varphi_1((t-t_0)A)\mathbf{b}) + \mathbf{b} = A\mathbf{y}(t) + \mathbf{b} \end{aligned}$$

Le funzioni \exp e φ_1 di matrice possono essere approssimate come visto al paragrafo 7. Da questa osservazione, per un problema

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

il metodo *Eulero esponenziale* è

$$\mathbf{y}_{n+1} = \exp(kA)\mathbf{y}_n + k\varphi_1(kA)\mathbf{b}(t_n, \mathbf{y}_n)$$

Proposizione 4. *Il metodo di Eulero esponenziale è esatto se $\mathbf{b}(\mathbf{y}(t)) = \mathbf{b}(\mathbf{y}_0) \equiv \mathbf{b}$ e di ordine uno altrimenti.*

Dimostrazione. Si ha

$$\mathbf{y}_{n+1} = \exp(kA)\mathbf{y}_n + \int_{t_n}^{t_{n+1}} \exp((t_{n+1} - \tau)A)\mathbf{g}(t_n)d\tau$$

ove si è posto $\mathbf{g}(t) = \mathbf{b}(t, \mathbf{y}(t))$. Per la formula di variazioni delle costanti (7.2)

$$\begin{aligned} \mathbf{y}(t_{n+1}) - \exp(kA)\mathbf{y}(t_n) - \int_{t_n}^{t_{n+1}} \exp((t_{n+1} - \tau)A)\mathbf{g}(t_n)d\tau &= \\ &= \exp(kA)\mathbf{y}(t_n) + \int_{t_n}^{t_{n+1}} \exp((t_{n+1} - \tau)A)\mathbf{g}(\tau)d\tau + \\ &- \exp(kA)\mathbf{y}(t_n) - \int_{t_n}^{t_{n+1}} \exp((t_{n+1} - \tau)A)\mathbf{g}(t_n)d\tau = \\ &= \int_{t_n}^{t_{n+1}} \exp((t_{n+1} - \tau)A)(\mathbf{g}(t_n) + \mathbf{g}'(\tau_n)(\tau - t_n) - \mathbf{g}(t_n))d\tau = \\ &= k^2\varphi_2(kA)\mathbf{g}'(\tau_n) = \mathcal{O}(k^2) \end{aligned}$$

□

Si può inoltre dimostrare che il metodo converge (cioè è stabile). Poiché risolve esattamente i problemi lineari, il metodo è A-stabile.

Proposizione 5. *Per un problema lineare, non autonomo*

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

il metodo esponenziale—punto medio

$$\mathbf{y}_{n+1} = \exp(kA)\mathbf{y}_n + k\varphi_1(kA)\mathbf{b}(t_n + k/2)$$

è esatto se $\mathbf{b}(t) \equiv \mathbf{b}$ e di ordine 2 altrimenti.

Dimostrazione. Procedendo come sopra, si arriva a

$$\begin{aligned} \mathbf{y}(t_{n+1}) - \exp(kA)\mathbf{y}(t_n) - \int_{t_n}^{t_{n+1}} \exp((t_{n+1} - \tau)A)\mathbf{b}(t_n + k/2)d\tau &= \\ &= \int_{t_n}^{t_{n+1}} \exp((t_{n+1} - \tau)A)\mathbf{b}'(\tau_n + k/2)(\tau - (t_n + k/2))d\tau = \\ &= \int_{t_n}^{t_{n+1}} \exp((t_{n+1} - \tau)A)\mathbf{b}'(\tau_n + k/2)(\tau - t_n - k/2)d\tau = \\ &= (k^2\varphi_2(kA) - k^2/2\varphi_1(kA))\mathbf{b}'(\tau_n + k/2) = \\ &= \left(\frac{k^2I}{2} + \frac{k^3A}{6} + \mathcal{O}(k^4) - \frac{k^2I}{2} - \frac{k^3A}{2} + \mathcal{O}(k^4) \right) \mathbf{b}'(\tau_n + k/2) = \\ &= \mathcal{O}(k^3) \end{aligned}$$

□

Anche in questo caso si può dimostrare che il metodo converge e che è A-stabile. Dato un problema differenziale in forma autonoma

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

si può pensare di linearizzarlo ad ogni passo

$$\mathbf{y}'(t) = J_n \mathbf{y}(t) + \mathbf{b}_n(\mathbf{y}(t))$$

ove

$$J_n = \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(\mathbf{y}_n), \quad \mathbf{b}_n(\mathbf{y}(t)) = \mathbf{f}(\mathbf{y}(t)) - J_n \mathbf{y}(t)$$

e applicarvi il metodo di Eulero esponenziale. Si arriva così al metodo di Eulero–Rosenbrock esponenziale

$$\mathbf{y}_{n+1} = \exp(kJ_n) \mathbf{y}_n + k\varphi_1(kJ_n) \mathbf{b}_n(\mathbf{y}_n) = \mathbf{y}_n + k\varphi_1(kJ_n) \mathbf{f}(\mathbf{y}_n)$$

Il metodo è di ordine 2 e convergente.

Gli integratori esponenziali sono particolarmente utili per la risoluzione di problemi stiff (essendo A-stabili). Convieni 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 preconditionatori efficaci. Inoltre, per i metodi visti, si può risparmiare il calcolo dell'esponenziale di matrice notando che

$$\exp(kA) \mathbf{y}_n + k\varphi_1(kA) \mathbf{b} = \mathbf{y}_n + k\varphi_1(kA) (A\mathbf{y}_n + \mathbf{b})$$

Capitolo 21

Esercizi

1. Si consideri il seguente problema differenziale del secondo ordine *ai limiti*

$$\begin{cases} u''(x) - 3 \cos(u(x)) = 0, & x \in (0, 1) \\ u(0) = 0 \\ u(1) = 1 \end{cases}$$

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 21.1, 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 $\mathbf{y}(1)$, ove $\mathbf{y}'(t) = A\mathbf{y}(t)$, $\mathbf{y}(0) = [1, \dots, 1]^T$, con A data da $A = 100 \cdot \text{toeplitz}(\text{sparse}([1, 1], [1, 2], [-2, 1], 1, 10))$, usando il θ -metodo con $\theta = 0, 1/2, 1$ e diversi passi temporali $k = 2^{-3}, 2^{-4}, \dots, 2^{-8}$. 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.

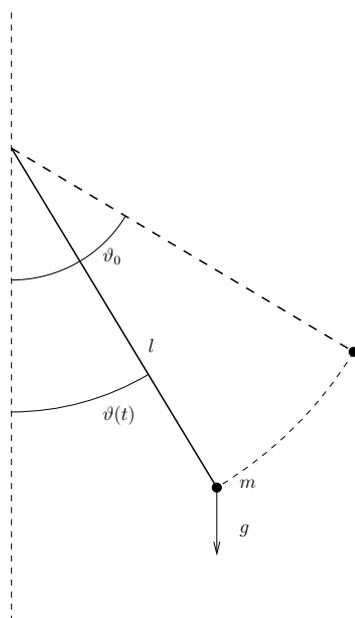


Figura 21.1: Pendolo

4. Si risolva il sistema di ODEs

$$\begin{cases} A'(t) = -2a(t)A(t) \\ a'(t) = A(t)^2 + \Omega(t)^2 - a(t)^2 - 1 \\ \Omega'(t) = -2(a(t) + A(t))\Omega(t) \end{cases} \quad (21.1)$$

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 18.2) e lo si testi per il sistema differenziale (21.1), producendo il grafico della quantità $E(t)$.
6. Si implementino gli altri due metodi di ordine 2 in Tabella 18.2, li si testi per il sistema differenziale (21.1), 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 21.1: Metodo di Runge–Kutta a 4 stadi.

7. Si implementi il metodo di Runge–Kutta di tableau in Tabella 21.1, determinandone numericamente l'ordine.
8. Si implementi la function relativa ad un generico metodo di Runge–Kutta con tableaux dato da

$$\begin{array}{c|c} \mathbf{c} & A \\ \hline & \mathbf{b}^T \end{array}$$

ove \mathbf{c} , A e \mathbf{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 21.1.

10. Si implementi il metodo di Runge–Kutta–Fehlberg il cui tableau è riportato nella Tabella 18.7, e se ne mostri l'ordine. Lo si testi sul sistema differenziale (21.1).

Parte 3

PDEs

**(Equazioni alle derivate
parziali)**

Capitolo 22

Equazione del calore

22.1 Equazione del calore con dati iniziali e condizioni ai bordi

Consideriamo la seguente equazione alle derivate parziali

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) = \frac{\partial^2 u}{\partial x^2}(t, x), & t > 0, x \in (0, L) \\ u(t, 0) = u(t, L) = 0, & t > 0 \text{ (condizioni ai bordi)} \\ u(0, x) = u_0(x), & x \in (0, L) \text{ (condizioni iniziali)} \end{cases} \quad (22.1)$$

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 $u_0(x)$.

22.1.1 Esistenza di una soluzione

Cerchiamo una soluzione *a variabili separabili*

$$u(t, x) = \psi(t)\phi(x)$$

Inserendo tale rappresentazione in (22.1), si deduce

$$\psi'(t)\phi(x) = \psi(t)\phi''(x), \quad t > 0, x \in (0, L)$$

da cui

$$\frac{\psi'(t)}{\psi(t)} = -K \text{ (costante)} \Rightarrow \psi(t) = Ae^{-Kt}$$

Per quanto riguarda $\phi(x)$, la soluzione generale è

$$\phi(x) = Be^{\sqrt{-K}x} + Ce^{-\sqrt{-K}x}$$

Imponendo le condizioni al bordo

$$0 = \phi(0) = B + C$$

$$0 = \phi(L) = Be^{\sqrt{-K}L} + Ce^{-\sqrt{-K}L} = B \left(e^{\sqrt{-K}L} - e^{-\sqrt{-K}L} \right)$$

Se $K \leq 0$, allora $e^{\sqrt{-K}L} + e^{-\sqrt{-K}L} > 0$ 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 $K = \lambda^2 > 0$, $\lambda > 0$, allora

$$\phi(x) = B (e^{i\lambda x} - e^{-i\lambda x}) = 2Bi \sin(\lambda x) = B \sin(\lambda x)$$

(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

$$u_0(x) = u(0, x) = \sum_{j=1}^{\infty} c_j \sin\left(\frac{j\pi}{L}x\right) \quad (22.2)$$

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 13.2.1,

$$\begin{aligned} u_{0m/2+1+j} &= \int_{-L}^L \bar{u}_0(x) \overline{\phi_{m/2+1+j}(x)} dx = \frac{-i}{\sqrt{2L}} \int_{-L}^L \bar{u}_0(x) \sin\left(\frac{2\pi j(x+L)}{2L}\right) dx = \\ &= \frac{-i\sqrt{2}}{\sqrt{L}} \int_0^L \bar{u}_0(x) \sin\left(\frac{j\pi x}{L} + j\pi\right) dx = \\ &= \frac{-i\sqrt{2}}{\sqrt{L}} \int_0^L u_0(x) \sin\left(\frac{j\pi x}{L} + j\pi\right) dx \end{aligned}$$

e

$$\begin{aligned} u_{0_{m/2+1-j}} &= \int_{-L}^L \bar{u}_0(x) \overline{\phi_{m/2+1-j}(x)} dx = \frac{-i}{\sqrt{2L}} \int_{-L}^L \bar{u}_0(x) \sin\left(\frac{-2\pi j(x+L)}{2L}\right) dx = \\ &= \frac{i\sqrt{2}}{\sqrt{L}} \int_0^L \bar{u}_0(x) \sin(j\pi x/L + j\pi) dx = -u_{0_{m/2+1+j}} \end{aligned}$$

da cui

$$\begin{aligned} \bar{u}_0(x) &= \sum_{j=-\infty}^{+\infty} u_{0_{m/2+1+j}} \phi_{m/2+1+j}(x) = \\ &= \sum_{j=-\infty}^{+\infty} u_{0_{m/2+1+j}} \frac{\cos(j\pi x/L + j\pi) + i \sin(j\pi x/L + j\pi)}{\sqrt{2L}} = \\ &= \sum_{j=-\infty}^{-1} u_{0_{m/2+1+j}} \frac{\cos(j\pi x/L + j\pi) + i \sin(j\pi x/L + j\pi)}{\sqrt{2L}} + \\ &+ \sum_{j=1}^{+\infty} u_{0_{m/2+1+j}} \frac{\cos(j\pi x/L + j\pi) + i \sin(j\pi x/L + j\pi)}{\sqrt{2L}} = \\ &= \sum_{j=1}^{+\infty} -u_{0_{m/2+1+j}} \frac{\cos(j\pi x/L + j\pi) - i \sin(j\pi x/L + j\pi)}{\sqrt{2L}} + \\ &+ \sum_{j=1}^{+\infty} u_{0_{m/2+1+j}} \frac{\cos(j\pi x/L + j\pi) + i \sin(j\pi x/L + j\pi)}{\sqrt{2L}} = \\ &= \sum_{j=1}^{\infty} u_{0_{m/2+1+j}} \frac{\sqrt{2}}{\sqrt{L}} i \sin(j\pi x/L + j\pi) = \\ &= \sum_{j=1}^{\infty} \left[\frac{2}{L} \int_0^L u_0(x) \sin\left(\frac{j\pi}{L}x\right) dx \right] \sin\left(\frac{j\pi}{L}x\right) \end{aligned}$$

Confrontando quest'ultima espressione con (22.2), si deduce

$$c_j = \left[\frac{2}{L} \int_0^L u_0(x) \sin\left(\frac{j\pi}{L}x\right) dx \right]$$

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 (22.1) (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 $t \rightarrow +\infty$ (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.

22.1.2 Unicità della soluzione

Introduciamo la seguente quantità (*energia*)

$$E(t) = \int_0^L \frac{1}{2} u^2(t, x) dx$$

Si ha

$$\frac{dE}{dt} = \int_0^L \frac{\partial}{\partial t} \left[\frac{1}{2} u^2(t, x) \right] dx = \int_0^L u \frac{\partial u}{\partial t} dx = \int_0^L u \frac{\partial^2 u}{\partial x^2} dx$$

Integrando per parti e tenendo conto delle condizioni ai bordi, si ha

$$\frac{dE}{dt} = - \int_0^L \left(\frac{\partial u}{\partial x} \right)^2 dx \leq 0$$

Per dimostrare l'unicità, consideriamo come al solito il problema omogeneo (corrispondente a (22.1) 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 (22.1), 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 $\varepsilon \rightarrow 0$, si ottiene

$$\min_{\Gamma} u \leq \min u$$

Poiché ovviamente vale anche la disuguaglianza opposta,

$$\min u = \min_{\Gamma} u = \min\{\min u_0, 0\} = 0$$

22.2 Metodo di Fourier

Per quanto visto, il metodo spettrale basato su approssimazione in serie di Fourier (vedi paragrafo 13.2.1) 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 $\{\phi_j\}_j$,

$$\begin{cases} \hat{u}'_k(t) = \lambda_k^2 \hat{u}_k(t), & 1 \leq k \leq m \\ \hat{u}_k(0) = \hat{u}_{0k} \end{cases}$$

ove $\lambda_k = i(k-1-m/2)2\pi/(2L)$ (si deve lavorare infatti del dominio $[-L, L]$, dove si è prolungata per simmetria la funzione $u_0(x)$) e \hat{u}_{0k} sono i coefficienti di Fourier discreti di u_0 prolungata (per inciso, si ha $i\sqrt{2}\sqrt{L}\hat{u}_{0m/2+1+k} = (-1)^k c_k$). Si trova, dunque,

$$\hat{u}_k(t) = e^{-(k-1-m/2)^2\pi^2 t/L^2} \hat{u}_{0k}, \quad 1 \leq k \leq m$$

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

$$\begin{cases} \frac{\partial \bar{v}}{\partial t}(t, x) = -\frac{\partial \bar{u}}{\partial t}(t, -x) \\ \frac{\partial^2 \bar{v}}{\partial x^2}(t, x) = (-\cdot -) \cdot -\frac{\partial^2 \bar{u}}{\partial x^2}(t, -x) \end{cases}$$

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.

22.3 Metodo delle linee

Il *metodo delle linee* per la risoluzione di problemi del tipo

$$\left\{ \begin{array}{l} \frac{\partial u}{\partial t}(t, x) = \frac{\partial^2 u}{\partial x^2}(t, x) + g(u(t, x)) + s(t, x), \quad t > 0, x \in (a, b) \\ + \text{condizioni ai bordi} \\ + \text{condizioni iniziali} \end{array} \right. \quad (22.3)$$

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. Vediamo qualche esempio.

22.3.1 Differenze finite

Trascurando per il momento le condizioni ai bordi e usando differenze finite centrate del secondo ordine a passo costante h

$$\begin{bmatrix} y_1'(t) \\ y_2'(t) \\ \vdots \\ y_{m-1}'(t) \\ y_m'(t) \end{bmatrix} = \frac{1}{h^2} \begin{bmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \cdots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 1 & -2 & 1 \\ 0 & \cdots & 0 & 1 & -2 \end{bmatrix} \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_{m-1}(t) \\ y_m(t) \end{bmatrix} + \begin{bmatrix} b(t_1, y_1(t)) \\ b(t_2, y_2(t)) \\ \vdots \\ b(t_{m-1}, y_{m-1}(t)) \\ b(t_m, y_m(t)) \end{bmatrix}$$

ove $y_j(t) \approx y(t, x_j)$ o, in maniera compatta,

$$\mathbf{y}'(t) = \mathbf{A}\mathbf{y}(t) + \mathbf{b}(t, \mathbf{y}(t)) = \mathbf{f}(t, \mathbf{y}(t)) \quad (22.4)$$

(con l'ovvia definizione dei simboli). A questo punto, si sceglie il metodo di integrazione temporale (θ -metodo, Runge–Kutta, multistep, esponenziale). Si tenga presente che il problema (22.4), 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

$$\lambda_j = -4 \sin^2 \left(\frac{j}{2m+1} \pi \right), \quad 1 \leq j \leq m$$

che vengono poi amplificati dal coefficiente $1/h^2$. Dunque, con riferimento alla condizione (19.2) 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 $\mathcal{O}(k + h^2) = \mathcal{O}(h^2)$). 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 19.2) 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 $\mathcal{O}(k^2 + h^2) = \mathcal{O}(h^4 + h^2) = \mathcal{O}(h^2)$.

22.3.2 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

$$\mathbf{y}'(t) = A\mathbf{y}(t) + \mathbf{b}(t, \mathbf{y}(t)) = \mathbf{f}(t, \mathbf{y}(t))$$

Costanti nel tempo

Si deve modificare la prima riga di $\mathbf{f}(t, \mathbf{y}(t))$ 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 \mathbf{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 $\mathbf{x} = \mathbf{y}_{n+s}$ tale che

$$F_{n+s-1}(\mathbf{x}) = 0$$

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

$$F_n(\mathbf{y}_{n+1}) = (I - kA)\mathbf{y}_{n+1} - k\mathbf{b}(t_{n+1}, \mathbf{y}_{n+1}) - \mathbf{y}_n = 0$$

e l'imposizione della condizione al bordo avviene, per esempio, ponendo a zero la prima riga di A e la prima componente di \mathbf{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 \mathbf{y}_n .

- **metodi esponenziali:** per i metodi esponenziali visti si ha

$$\mathbf{y}_{n+1} = \exp(kA)\mathbf{y}_n + k\varphi_1(kA)\mathbf{b}_n$$

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 \mathbf{b}_n uguale a $(u_a(t_{n+1}) - u_a(t_n))/k$.

22.3.3 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 [10.4.2](#)).

22.3.4 Equazione di trasporto-diffusione

Consideriamo l'equazione del *trasporto* (in un dominio non limitato)

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) + c \frac{\partial u}{\partial x}(t, x) = 0, & t > 0, x \in \mathbb{R} \\ u(0, x) = u_0(x) \end{cases} \quad (22.5)$$

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

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) + c \frac{\partial u}{\partial x}(t, x) = 0, & t > 0, x \in (a, b), c > 0 \\ u(t, a) = 0, & t > 0 \\ u(0, x) = u_0(x), & x \in (a, b) \end{cases} \quad (22.6)$$

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*

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) + c \frac{\partial u}{\partial x}(t, x) = d \frac{\partial^2 u}{\partial x^2} & t > 0, x \in (a, b) \\ u(0, x) = u_0(x) \\ u(t, a) = u(t, b) = 0 \end{cases}$$

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 17. *Dato*

$$\begin{cases} \mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)), & t > 0 \\ \mathbf{y}(0) = \mathbf{y}_0 \end{cases}$$

sono equivalenti le seguenti proprietà:

- se $\mathbf{y}_0 \geq 0$, allora $\mathbf{y}(t) \geq 0$ per ogni t (il sistema si dice positivo)
- dato \mathbf{x} , con $x_i = 0$, $x_j \geq 0$, $j \neq i$, allora $f_i(t, \mathbf{x}) \geq 0$

Da questo teorema segue, come corollario, il seguente, che può però essere dimostrato in maniera diretta.

Teorema 18. Un sistema lineare $\mathbf{y}'(t) = A\mathbf{y}(t)$ 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 $\mathbf{y}_0 \geq 0$, si ha $\mathbf{y}(\tau) \geq 0$. Ma

$$\mathbf{y}(\tau) = \exp(\tau A)\mathbf{y}_0 = (I + \tau A)\mathbf{y}_0 + \mathcal{O}(\tau^2)$$

se τ è sufficientemente piccolo. Se, per assurdo, $a_{\bar{i}\bar{j}} < 0$, $\bar{j} \neq \bar{i}$, allora, preso $\mathbf{y}_0 = \mathbf{e}_{\bar{j}}$,

$$(I + \tau A)\mathbf{e}_{\bar{j}} = \begin{bmatrix} * \\ \vdots \\ * \\ \tau a_{\bar{i}\bar{j}} \\ * \\ \vdots \\ * \end{bmatrix} \leftarrow \text{riga } \bar{i}$$

e dunque la componente \bar{i} -esima di $\exp(\tau A)\mathbf{e}_{\bar{j}}$ sarebbe negativa, assurdo.

Se invece $a_{ij} \geq 0$, $j \neq i$, allora

$$\exp(tA) = \lim_{n \rightarrow \infty} \left(I + \frac{t}{n} A \right)^n \geq 0$$

da cui la positività. □

Tornando all'equazione (22.5), 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 $Pe = |c|h/(2d)$ 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 $\mathbf{y}'(t) = A\mathbf{y}(t)$ non garantisce che qualunque metodo numerico per ODEs la preservi (di certo lo garantisce il metodo esponenziale, poiché esatto). Pertanto, la condizione sul numero di Péclet di griglia è solo necessaria per avere una soluzione numerica positiva.

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.

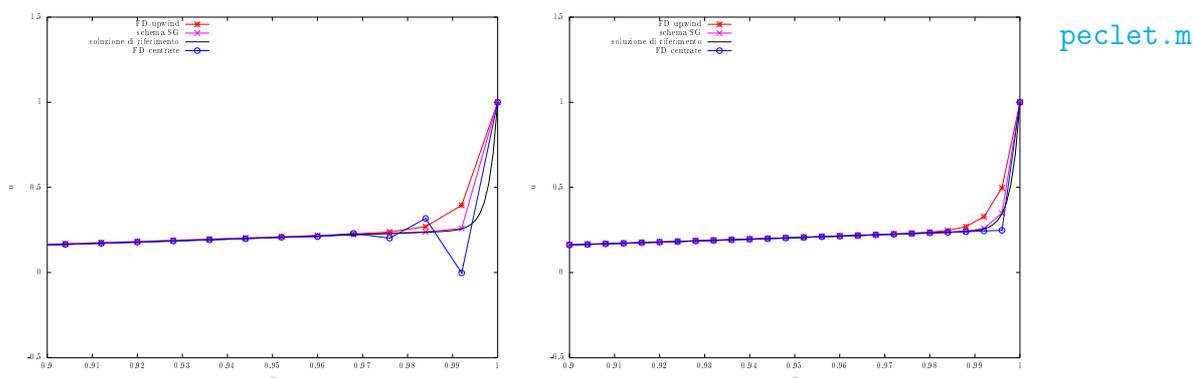


Figura 22.1: Soluzione di (22.7) con diversi schemi di differenze finite, $h = 1/125$ (sinistra), $h = 1/250$ (destra) e Eulero esponenziale al tempo $t^* = 0.05$. La soluzione di riferimento è stata ottenuta con differenze finite centrate e $h = 1/1000$.

Consideriamo, per esempio, l'equazione di trasporto-diffusione

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) + c \frac{\partial u}{\partial x}(t, x) = d \frac{\partial^2 u}{\partial x^2} & t > 0, x \in (0, 1) \\ u(0, x) = x^2 \\ u(t, 0) = 0 \\ u(t, 1) = 1 \end{cases} \quad (22.7)$$

con $c = 10$ e $d = 0.02$. Il numero di Péclet di griglia, con $h = 1/100$, vale 2.5. La risoluzione mediante differenze finite centrate e Eulero esponenziale (esatto nel tempo) produce il grafico blu a sinistra in Figura 22.1. 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

$$\frac{ch}{2d \left(1 + \frac{ch}{2d}\right)} = \frac{ch}{2d(1 + \text{Pe})} = \frac{\text{Pe}}{1 + \text{Pe}} < 1, \quad \forall h$$

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 $d(1 + \phi(\text{Pe}))$ in modo che il nuovo numero di Péclet valga

$$\frac{ch}{2d(1 + \phi(\text{Pe}))}$$

La funzione $\phi(\text{Pe})$ dovrà soddisfare:

- $\phi(\text{Pe}) \geq \text{Pe} - 1$ (così il nuovo numero di Péclet sarà minore o uguale a 1), ma non troppo grande (altrimenti si introduce troppa diffusione artificiale)
- $\phi(ch/(2d)) = \mathcal{O}(h^2)$, $h \rightarrow 0$ (così la discretizzazione a differenze finite centrare sarà di ordine 2)

Una scelta possibile è

$$\phi(t) = t - 1 + e^{-t}, \quad \phi(t) = \frac{t^2}{2} + \mathcal{O}(t^3), \quad t \rightarrow 0$$

Una scelta migliore è

$$\phi(t) = t - 1 + \frac{2t}{e^{2t} - 1}, \quad \phi(t) = \frac{t^2}{3} + \mathcal{O}(t^4), \quad t \rightarrow 0$$

(da notare che lo schema upwind corrisponde a $\phi(t) = t$). Il risultato corrisponde al grafico magenta in Figura 22.1 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 $\text{Pe} = |c|h/(2d)$, in modo da *aggiungere* diffusione artificiale e non togliere. Per esempio, nel caso upwind con $c < 0$, $\phi(\text{Pe}) = \text{Pe} = -ch/(2d)$ e pertanto la discretizzazione al primo ordine che ne risulta è

$$\begin{aligned} -c \frac{\partial u}{\partial x}(t, x_i) + d \frac{\partial^2 u}{\partial x^2}(t, x_i) &\approx -c \frac{u_{i+1} - u_i}{h} + d \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} = \\ &= -c \frac{u_{i+1} - u_{i-1}}{2h} + d \left(1 - \frac{ch}{2d}\right) \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} \end{aligned}$$

22.3.5 Elementi finiti

Nel caso di discretizzazione spaziale con elementi finiti lineari, la discretizzazione del problema (22.3) porta al sistema di ODEs

$$P\mathbf{y}'(t) = A\mathbf{y}(t) + \mathbf{g}(\mathbf{y}(t)) + \mathbf{s}(t) \quad (22.8)$$

ove A è (l'opposta de) la *stiffness matrix* e P la *mass matrix*, definita da,

$$\begin{aligned} p_{jj} &= \int_{x_{j-1}}^{x_{j+1}} \phi_j(x) \phi_j(x) dx = \frac{h_{j-1} + h_j}{3} \\ p_{j,j+1} &= p_{j+1,j} = \int_{x_j}^{x_{j+1}} \phi_j(x) \phi_{j+1}(x) dx = \frac{h_j}{6} \end{aligned} \quad (22.9a)$$

mentre, per $j = 1$ e $j = m$,

$$\begin{aligned}
 p_{11} &= \int_{x_1}^{x_2} \phi_1(x)\phi_1(x)dx = \frac{h_1}{3} \\
 p_{12} &= \int_{x_1}^{x_2} \phi_1(x)\phi_2(x)dx = \frac{h_1}{6} \\
 p_{m-1m} &= p_{mm-1} = \int_{x_{m-1}}^{x_m} \phi_m(x)\phi_{m-1}(x)dx = \frac{h_{m-1}}{6} \\
 p_{mm} &= \int_{x_{m-1}}^{x_m} \phi_m(x)\phi_m(x)dx = \frac{h_{m-1}}{3}
 \end{aligned} \tag{22.9b}$$

Poi, per $1 < i < m$,

$$\begin{aligned}
 g_i &= \int_{x_{i-1}}^{x_{i+1}} g \left(\sum_{j=1}^m u_j \phi_j(x) \right) \phi_i(x) dx = \\
 &= \int_{x_{i-1}}^{x_i} g \left(\sum_{j=1}^m u_j \phi_j(x) \right) \phi_i(x) dx + \int_{x_i}^{x_{i+1}} g \left(\sum_{j=1}^m u_j \phi_j(x) \right) \phi_i(x) dx \approx \\
 &\frac{g(y_{i-1}) + g(y_i)}{2} \frac{h_{i-1}}{2} + \frac{g(y_i) + g(y_{i+1})}{2} \frac{h_i}{2} \\
 s_i &= \int_{x_{i-1}}^{x_{i+1}} s(t, x) \phi_i(x) dx \approx \frac{s(t, x_{i-1}) + s(t, x_i)}{2} \frac{h_{i-1}}{2} + \frac{s(t, x_i) + s(t, x_{i+1})}{2} \frac{h_i}{2}
 \end{aligned}$$

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 (22.8), 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 (22.9) mediante la formula dei trapezi e dunque non riduce l'accuratezza del metodo. Infatti, la matrice $P_L^{(-1)}A$ (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 (22.8), 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.

22.3.6 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](#)

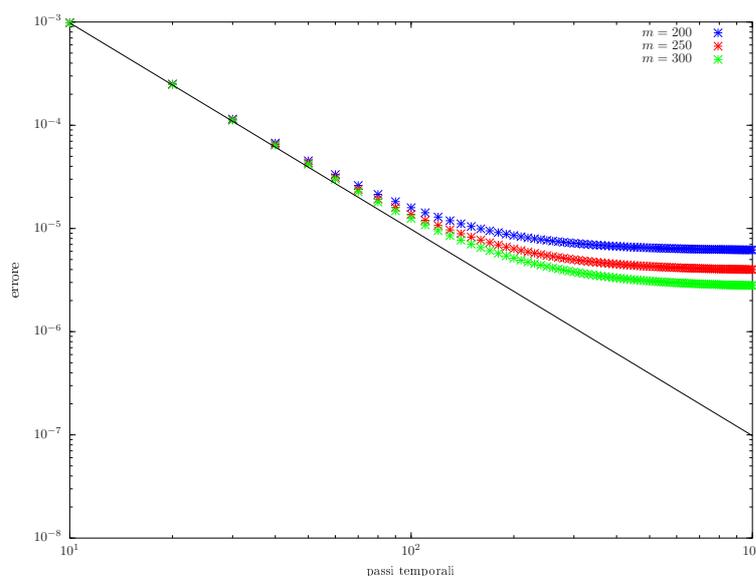


Figura 22.2: Errore temporale per un numero m di passi spaziali.

in Figura 22.2, 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.

22.4 Esercizi

1. Si calcoli la soluzione analitica dell'equazione del calore con sorgente

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) = \frac{\partial^2 u}{\partial x^2}(t, x) + 2e^t \sin(x), & t > 0, x \in (0, \pi/2) \\ u(t, 0) = 0, & t > 0 \\ \frac{\partial u}{\partial x}(t, \pi/2) = 0, & t > 0 \\ u(0, x) = \sin(x), & x \in (0, \pi/2) \end{cases}$$

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 21.1
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*

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) + c \frac{\partial u}{\partial x}(t, x) = d \frac{\partial^2 u}{\partial x^2}(t, x) + \rho u(t, x)(u(t, x) - 1/2)(1 - u(t, x)), & t > 0, x \in \\ u(t, 0) = 0, & t > 0 \\ \frac{\partial u}{\partial x}(t, 1) = 0, & t > 0 \\ u(0, x) = 5x(1 - x)^2, & x \in (0, 1) \end{cases}$$

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
Temi d'esame

29-06-2010

1. Si implementi il metodo di Runge–Kutta definito dal tableau

0					
$\frac{1}{2}$	$\frac{1}{2}$				
$\frac{1}{2}$	0	$\frac{1}{2}$			
1	0	0	1		
$\frac{3}{4}$	$\frac{5}{32}$	$\frac{7}{32}$	$\frac{13}{32}$	$-\frac{1}{32}$	
	$-\frac{1}{2}$	$\frac{7}{3}$	$\frac{7}{3}$	$\frac{13}{6}$	$-\frac{16}{3}$

e se ne determini numericamente l'ordine.

2. Si usi il metodo di Runge–Kutta implementato sopra per determinare numericamente il periodo del pendolo

$$\begin{cases} \theta''(t) = -g \sin(\theta(t)), & t > 0 \\ \theta(0) = \frac{\pi}{3} \\ \theta'(0) = 0 \end{cases}$$

e si dica se è minore, maggiore o uguale al periodo del pendolo *linearizzato*

$$\begin{cases} \theta_1''(t) = -g\theta_1(t), & t > 0 \\ \theta_1(0) = \frac{\pi}{3} \\ \theta_1'(0) = 0 \end{cases}$$

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

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) = \frac{1}{100} \frac{\partial^2 u}{\partial x^2}(t, x) + 5u(t, x) \left(u(t, x) - \frac{1}{2} \right) (1 - u(t, x)), & t > 0, x \in (0, 1) \\ u(0, x) = 4x(1 - x), & x \in (0, 1) \\ u(t, 0) = u(t, 1) = 0, & t > 0 \end{cases}$$

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

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}\left(t_n + \frac{h}{2}, \frac{1}{2}(\mathbf{y}_n + \mathbf{y}_{n+1})\right), \quad n \geq 0$$

per la soluzione di un problema ai valori iniziali

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

Lo si implementi e lo si applichi al problema autonomo

$$\begin{cases} y'(t) = y(t)(1 - y(t)), & t \geq 0 \\ y(0) = \frac{1}{2} \end{cases}$$

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

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) = \frac{\partial^2 u}{\partial x^2}(t, x) - (\sin t)(-x^2 + 2x) + 2 \cos t, & t \geq 0, x \in (0, 1) \\ u(0, x) = -x^2 + 2x, & x \in (0, 1) \\ u(t, 0) = 0, & t \geq 0 \\ \frac{\partial u}{\partial x}(t, 1) = 0, & t \geq 0 \end{cases}$$

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

$$\begin{cases} \mathbf{f}_i = \mathbf{f}\left(t_n + c_i k, \mathbf{y}_n + k \sum_{j=1}^i a_{i,j} \mathbf{f}_j\right), & i = 1, 2 \\ \mathbf{y}_{n+1} = \mathbf{y}_n + k \sum_{j=1}^2 b_j \mathbf{f}_j \end{cases}$$

di tableau

$$\begin{array}{c|cc|c|cc} c_1 & a_{1,1} & a_{1,2} & \frac{3+\sqrt{3}}{6} & \frac{3+\sqrt{3}}{6} & 0 \\ c_2 & a_{2,1} & a_{2,2} & \frac{3-\sqrt{3}}{6} & -\frac{\sqrt{3}}{3} & \frac{3+\sqrt{3}}{6} \\ \hline & b_1 & b_2 & & \frac{1}{2} & \frac{1}{2} \end{array} =$$

per la soluzione di un problema ai valori iniziali

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

Lo si implementi per il problema autonomo

$$\begin{cases} y'(t) = y(t)(1 - y(t)), & t \geq 0 \\ y(0) = \frac{1}{2} \end{cases}$$

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

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) = \frac{\partial^2 u}{\partial x^2}(t, x) + (2x - x^2) \cos t + 2 \sin t, & t \geq 0, x \in (0, 1) \\ u(0, x) = 0, & x \in (0, 1) \\ u(t, 0) = 0, & t \geq 0 \\ u(t, 1) = \sin t, & t \geq 0 \end{cases}$$

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

$$\begin{cases} u''(x) = \frac{1}{8}(32 + 2x^3 - u(x)u'(x)), & x \in (1, 3) \\ u(1) = 17 \\ u(3) = \frac{43}{3} \end{cases}$$

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

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) = \frac{\partial^2 u}{\partial x^2}(t, x) + 2e^t \sin x, & t > 0, x \in (0, \pi/2) \\ u(0, x) = \sin x, & x \in (0, \pi/2) \\ u(t, 0) = 0, & t \geq 0 \\ \frac{\partial u}{\partial x}(t, \pi/2) = 0, & t \geq 0 \end{cases}$$

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 $(t, x) \in [0, 1] \times [0, 1]$

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) = \varepsilon \frac{\partial^2 u}{\partial x^2}(t, x) + \frac{\partial u}{\partial x}(t, x) + \rho u(t, x)(u(t, x) - 1/2)(1 - u(t, x)) \\ u(0, x) = 10x^2(1 - x)^2 + 1/2 \\ \frac{\partial u}{\partial x}(t, 0) = \frac{\partial u}{\partial x}(t, 1) = 0 \end{cases}$$

(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

$$\begin{cases} u''(x) = \frac{1}{8}(32 + 2x^3 - u(x)u'(x)), & x \in (1, 3) \\ u(1) = 17 \\ u(3) = \frac{43}{3} \end{cases}$$

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

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) = \frac{\partial^2 u}{\partial x^2}(t, x) + \frac{5}{4}e^t \sin \frac{x}{2}, & t > 0, x \in (0, \pi) \\ u(0, x) = \sin \frac{x}{2}, & x \in (0, \pi) \\ u(t, 0) = 0, & t \geq 0 \\ \frac{\partial u}{\partial x}(t, \pi) = 0, & t \geq 0 \end{cases}$$

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

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) = \frac{1}{100} \frac{\partial^2 u}{\partial x^2}(t, x) + \sin(u(t, x)), & t > 0, x \in (0, 1) \\ u(0, x) = 10x(1 - x)^2, & x \in [0, 1] \\ u(t, 0) = 0, & t \geq 0 \\ \frac{\partial u}{\partial x}(t, 1) = 0, & t \geq 0 \end{cases}$$

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

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) = \frac{1}{50} \frac{\partial^2 u}{\partial x^2}(t, x) + \cos(u(t, x)), & t > 0, x \in (0, 1) \\ u(0, x) = 10x^2(1 - x) + 1, & x \in [0, 1] \\ \frac{\partial u}{\partial x}(t, 0) = 0, & t \geq 0 \\ u(t, 1) = 1, & t \geq 0 \end{cases}$$

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

$$\begin{cases} x''(t) = -Bx'(t) \\ x(0) = 0 \\ x'(0) = v_0 \cos(\alpha) \end{cases} \quad \begin{cases} y''(t) = -By'(t) - g \\ y(0) = 0 \\ y'(0) = v_0 \sin(\alpha) \end{cases}$$

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

$$\begin{cases} \frac{\partial u}{\partial t} = \frac{1}{100} \frac{\partial^2 u}{\partial x^2} + u(1-u)(u-1/2), & t > 0, x \in (0, 1) \\ u(0, x) = 10x^2(1-x)^2, & x \in [0, 1] \\ \frac{\partial u}{\partial x}(t, 0) = 0, & t \geq 0 \\ \frac{\partial u}{\partial x}(t, 1) = 0, & t \geq 0 \end{cases}$$

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

$$\begin{cases} y''(x) = 2y^3 - 6y - 2x^3, & x \in (1, 2) \\ y(1) = 2 \\ y(2) = \frac{5}{2} \end{cases}$$

mostrando il corretto ordine di convergenza del metodo scelto.

24. Si applichi il metodo delle linee al problema di diffusione non omogeneo

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) = \frac{\partial^2 u}{\partial x^2}(t, x) + \frac{5}{4}e^t \cos \frac{x}{2}, & t > 0, x \in (\pi, 2\pi) \\ u(0, x) = \cos \frac{x}{2}, & x \in (\pi, 2\pi) \\ u(t, \pi) = 0, & t \geq 0 \\ \frac{\partial u}{\partial x}(t, 2\pi) = 0, & t \geq 0 \end{cases}$$

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 $t^* = 1$.

03-02-2011

25. Si risolva il seguente problema differenziale

$$\begin{cases} y''(x) + \frac{y'(x)}{x} = \cos(y(x)) & x \in (0, 1] \\ y(0) = 1 \\ y'(0) = 0 \end{cases}$$

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

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) = \frac{\partial^2 u}{\partial x^2}(t, x) + \frac{3}{2}e^{t/2} \sin x, & t > 0, x \in (-\pi, \pi/2) \\ u(0, x) = \sin x, & x \in (-\pi, \pi/2) \\ u(t, -\pi) = 0, & t \geq 0 \\ \frac{\partial u}{\partial x}(t, \pi/2) = 0, & t \geq 0 \end{cases}$$

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

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) = \frac{x - x^2}{2} \cdot \frac{\partial^2 u}{\partial x^2}(t, x), & t > 0, x \in (0, 1) \\ u(0, x) = x^2 - x, & x \in (0, 1) \\ u(t, 0) = 0, & t \geq 0 \\ u(t, 1) = 0, & t \geq 0 \end{cases}$$

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

$$\begin{aligned} \mathbf{y}_{n+3} - \frac{18}{11}\mathbf{y}_{n+2} + \frac{9}{11}\mathbf{y}_{n+1} - \frac{2}{11}\mathbf{y}_n &= k \frac{6}{11} \mathbf{f}(t_{n+3}, \mathbf{y}_{n+3}) \\ \mathbf{y}_{n+3} - \frac{18}{11}\mathbf{y}_{n+2} + \frac{9}{11}\mathbf{y}_{n+1} - \frac{3}{11}\mathbf{y}_n &= k \frac{6}{11} \mathbf{f}(t_{n+3}, \mathbf{y}_{n+3}) \end{aligned}$$

si dica quale è consistente e perché e se ne determini numericamente l'ordine.

30. Si risolva il problema differenziale di diffusione-trasporto-reazione

$$\begin{cases} \frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = d \frac{\partial^2 u}{\partial x^2} + \rho u(u - 1/2)(1 - u), & t > 0, x \in (0, 1) \\ u(t, 0) = 1, & t > 0 \\ \frac{\partial u}{\partial x}(t, 1) = 0, & t > 0 \\ u(0, x) = (x - 1)^2, & x \in (0, 1) \end{cases}$$

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

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) = 4 \frac{\partial^2 u}{\partial x^2}(t, x) + 3e^{2t} \cos \frac{x}{2}, & t > 0, x \in (0, \pi) \\ u(0, x) = \cos \frac{x}{2}, & x \in (0, \pi) \\ \frac{\partial u}{\partial x}(t, 0) = 0, & t \geq 0 \\ u(t, \pi) = 0, & t \geq 0 \end{cases}$$

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

$$\begin{cases} y_1'(t) = y_2(t)y_3(t) \sin t - y_1(t)y_2(t)y_3(t) \\ y_2'(t) = -y_1(t)y_3(t) \sin t + \frac{1}{20}y_1(t)y_3(t) \\ y_3'(t) = y_1^2(t)y_2(t) - \frac{1}{20}y_1(t)y_2(t) \end{cases}$$

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

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) = 4\frac{\partial^2 u}{\partial x^2}(t, x) - 2\frac{\partial u}{\partial x}(t, x), & t \geq 0, x \in (0, 1) \\ u(0, x) = x^2, & x \in (0, 1) \\ \frac{\partial u}{\partial x}(t, 0) = 0, & t \geq 0 \\ u(t, 1) = 1, & t \geq 0 \end{cases}$$

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.

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.