CAPITOLO VIII

EQUAZIONI DIFFERENZIALI

1. IL PROBLEMA DI CAUCHY (PROBLEMA AI VALORI INIZIALI)

Consideriamo il seguente problema di Cauchy per i sistemi di equazioni differenziali del primo ordine :

$$y'(t) = f(t,y(t)) \qquad (8.1)$$
$$y(t_0)=y_0$$

dove f(t,y): $[t_0, t_f]\times R^m \to R^m$. La funzione f(t,y) è supposta continua rispetto a t , e lipschitziana rispetto ad y nella striscia illimitata $[t_0, t_f]\times R^m$,

$$||f(t,u)-f(t,v)||<L||u-v||, \quad \forall\, t\in[t_0, t_f] \text{ e } \forall\, u,v \in\times R^m.$$

E' noto che tali condizioni garantiscono l'esistenza e l'unicità della soluzione nell'intero intervallo di integrazione $[t_0, t_f]$. Inoltre, detta z(t) la soluzione dell'equazione (8.1) con valore iniziale $z(t_0)=z_0$, vale la seguente relazione che, tra l'altro, assicura la dipendenza continua delle soluzioni dai dati iniziali:

$$||y(t)-z(t)||<e^{L(t-t_0)}|| y_0- z_0||, \quad \forall\, t\in[t_0, t_f].$$

Per approssimare numericamente la soluzione del problema (8.1) fissiamo una discretizzazione dell'intervallo $[t_0, t_f]$ che, per semplicità di esposizione, supporremo uniforme:

$$t_0< t_1<...< t_N (=t_f) \qquad h=\frac{t_f - t_0}{N}.$$

e, per ogni intervallo $[t_n, t_{n+1}]$, consideriamo l'identità:

$$\int_{t_n}^{t_{n+1}} y'(t)dt = \int_{t_n}^{t_{n+1}} f(t,y(t))dt$$

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(t,y(t))dt\,. \qquad (8.2)$$

Ogni formula di quadratura che fa uso dei valori nodali di y(t) può essere utilizzata per creare una formula di integrazione numerica per il problema (8.1).

Metodi a un passo.

I metodi ad un passo sono quelli che, per approssimare l'integrale in (8.2), fanno uso di formule che richiedono soltanto valori della y relativi all'intervallo corrente $[t_n,\ t_{n+1}]$.

Limitiamoci, per ora, a considerare alcune formule, trattate nel capitolo precedente, che fanno uso di uno o di entrambi gli estremi dell'integrale. In particolare:

1) $\int_{t_n}^{t_{n+1}} f(t,y(t))dt = hf(t_n,y(t_n)) + \sigma_1(t_n,h)$

$$\sigma_1(t_n,h) = \tfrac{1}{2}\frac{\partial}{\partial t}f(\xi_n,y(\xi_n))h^2 = \tfrac{1}{2}y''(\xi_n)h^2 \qquad \xi_n \in (t_n,\ t_{n+1})$$

2) $\int_{t_n}^{t_{n+1}} f(t,y(t))dt = hf(t_{n+1},y(t_{n+1})) + \sigma_2(t_n,h)$

$$\sigma_2(t_n,h) = -\tfrac{1}{2}\frac{\partial}{\partial t}f(\xi_n,y(\xi_n))h^2 = -\tfrac{1}{2}y''(\xi_n)h^2 \qquad \xi_n \in (t_n,\ t_{n+1})$$

3) $\int_{t_n}^{t_{n+1}} f(t,y(t))dt = \tfrac{1}{2}h\big(f(t_n,y(t_n))+f(t_{n+1},y(t_{n+1}))\big) + \sigma_3(t_n,h)$

$$\sigma_3(t_n,h) = -\tfrac{1}{12}\frac{\partial^2}{\partial t^2}f(\xi_n,y(\xi_n))h^3 = -\tfrac{1}{12}y'''(\xi_n)h^3 \qquad \xi_n \in (t_n,\ t_{n+1})$$

Otteniamo così le relazioni:

1') $y(t_{n+1}) = y(t_n) + hf(t_n,y(t_n)) + \sigma_1(t_n,h)$

2') $y(t_{n+1}) = y(t_n) + hf(t_{n+1},y(t_{n+1})) + \sigma_2(t_n,h)$

3') $y(t_{n+1}) = y(t_n) + \tfrac{1}{2}h\big(f(t_n,y(t_n))+f(t_{n+1},y(t_{n+1}))\big) + \sigma_3(t_n,h)$.

Trascurando ad ogni passo l'errore $\sigma(t_n,h)$, detto **errore locale di troncamento**, si ottengono le formule ricorsive:

formula di **Eulero Esplicita** $\boxed{y_{n+1} = y_n + hf(t_n,y_n)}$

formula di **Eulero Implicita** $\boxed{y_{n+1} = y_n + hf(t_{n+1},y_{n+1})}$

formula dei **Trapezi** $\boxed{y_{n+1} = y_n + \tfrac{1}{2}h(f(t_n,y_n)+f(t_{n+1},y_{n+1}))}$

dove, per ogni n, y_n è l'approssimazione della soluzione nel punto $t_n = t_0+nh$ ed y_0 è il valore iniziale assegnato.

Si osservi che le formule di Eulero Implicita e dei trapezi presentano una maggiore complessità computazionale, rispetto alla formula di Eulero esplicita, poiche l'incognita y_{n+1} si presenta come la risoluzione di una equazione, in generale non lineare, in R^m,
Per comodità di trattazione, esprimiamo le precedenti formule nella forma generale:

$$y_{n+1} = y_n + h\Phi(t_n, y_n) \qquad (8.2)$$

dove la funzione $\Phi(t, y)$ è detta **funzione incrementale**.

Per ciacuna delle precedenti formule è immediato verificare la lipschitzianità della funzione incrementale rispetto a y, cioè l'esistenza di una costante M tale che

$$||\Phi(t,u)-\Phi(t,v)||<M||u-v||, \qquad \forall\, t\in[t_0, t_f] \text{ e } \forall\, u,v \in R^m,.$$

Per il metodo di Eulero esplicito, ciò si ricava immediatamente come conseguenza della lispchitzianità di f.
Per il metodo di Eulero implicito si osservi invece che $\Phi(t_n,y_n)=f(t_{n+1},\alpha)$, dove α è la soluzione del problema:

$$\alpha = y_n + hf(t_{n+1},\alpha) \qquad (8.3)$$

e $\Phi(t_n,z_n)=f(t_{n+1},\beta)$, dove β è la soluzione del problema:

$$\beta = z_n + hf(t_{n+1},\beta) \qquad (8.4)$$

Allora si ha:

$$||\Phi(t_n,y_n)-\Phi(t_n,z_n)||=||f(t_{n+1},\alpha)-f(t_{n+1},\beta)|| <L||\alpha-\beta||$$

dove $||\alpha-\beta||$ può essere ricavato dalle relazioni (8.3) ed (8.4). Si ottiene così, per h sufficientemente piccolo

$$||\alpha-\beta|| < \frac{1}{1-hL}||y_n-z_n||$$

e quindi

$$||\Phi(t_n,y_n)-\Phi(t_n,z_n)||<M||y_n-z_n||$$

dove la costante di Lipschitz per la Φ è data $M=\frac{L}{1-hL}$. (Si osservi che per hL<1/2 si ha M<2L)

Si verifichi che anche per il metodo dei trapezi la funzione di iterazione Φ è lipschitziana, e se ne valuti la costante.

Metodi di Runge-Kutta.

Altri metodi ad un passo si possono ottenere approssimando l'integrale in (8.2) con altre formule di quadratura che utilizzano punti t_n+c_ih i=1,...,s, anche diversi dagli estremi ma sempre inclusi nell'intervallo corrente $[t_n, t_{n+1}]$. Si ottengono così formule del tipo

$$y(t_{n+1})= y(t_n) + h\sum_{i=1}^{s} b_i f(t_n+c_ih , y(t_n+c_ih)) + \sigma(t_n,h)$$

per le quali dovrei conoscere i valori incogniti $y(t_n+c_ih)$. Ma questi possono a loro volta essere calcolati in modo approssimato non appena si osservi che, per un generico punto t dell'intervallo corrente, vale la formula:

$$y(t) = y(t_n) + \int_{t_n}^{t} f(s,y(s))ds.$$

e quindi, per ogni punto t_n+c_ih:

$$y(t_n+c_ih) = y(t_n) + \int_{t_n}^{t_n+c_ih} f(s,y(s))ds.$$

Consideriamo quindi, per ciascun integrale $\int_{t_n}^{t_n+c_ih} f(s,y(s))ds$, una formula di quadratura che faccia uso di tutti o alcuni dei nodi t_n+c_ih e dei corrispondenti valori incogniti $y(t_n+c_ih)$. Essa sarà del tipo

$$\int_{t_n}^{t_n+c_ih} f(s,y(s))ds = h\sum_{j=1}^{s} a_{i,j} f(t_n+c_jh , y(t_n+c_jh)) + \sigma_i(t_n,h)$$

Se imponiamo che, per ogni i, la formula di quadratura sia esatta almeno per le funzioni costanti, dobbiamo imporre ai pesi $a_{i,j}$ la condizione

$$c_i = \sum_{j=1}^{s} a_{i,j} \qquad i=1,..,s \qquad (8.5)$$

In conclusione si ottiene la relazione

$$y(t_n+c_ih)= y(t_n) + h\sum_{j=1}^{s} a_{i,j} f(t_n+c_jh , y(t_n+c_jh)) + \sigma_i(t_n,h)$$

dalla quale, indicando con Y_i le incognite $y(t_n+c_ih)$, ed ignorando come al solito gli errori di quadratura, si ricava la formula:

$$y_{n+1}= y_n+h\sum_{i=1}^{s} b_i f(t_n+c_ih,Y_i) \qquad (8.6)$$

$$Y_i = y_n + h\sum_{j=1}^{s} a_{i,j} f(t_n+c_jh , Y_j) \qquad i=1,...,s. \qquad (8.7)$$

Le formule (8.6)-(8.7) sono note come **formule di Runge-Kutta ad** s **livelli** e sono rappresentate, in forma sintetica, attraverso la seguente tabella

c_1	a_{11}	.	.	.	a_{1s}
c_2	.				
:	.				
c_s	a_{s1}	.	.	.	a_{ss}
	b_1	b_2	.	.	b_s

dove $A=(a_{i,j})$ è la **matrice dei coefficienti**, $b=(b_1,b_2,... b_s)^T$ è il vettore dei **pesi**, e $c=(c_1,c_2,... c_s)^T$ è il vettore delle **ascisse** per il quale deve valere la condizione (8.5).

Se la matrice A è triangolare inferiore il metodo di dirà **esplicito** e gli Y_i si calcolano direttamente "in avanti" a cominciare da Y_1 .Si osservi che in questo caso la condizione (8.5) impone $c_1=0$.

Se A è triangolare inferiore ed include anche la diagonale, il sistema si dice **semi-implicito**. In questo caso il problema si riduce alla risoluzione ricorsiva di s equazioni R^m (si ricordi che $Y_i \in R^m$). Infine se A è una matrice piena, il metodo si dice **implicito** e richiede la risoluzione di un sistema non lineare in $R^{m\times s}$.

In entrambi questi casi il sistema (8.7) si presenta come un problema di punto fisso nell'incognita

$$Y=(Y_1, Y_2,..., Y_s) \in R^{m\times s},$$

per il quale è facile verificare, attraverso il teorema di contrazione, l'esistenza ed unicità della soluzione per h sufficientemente piccolo. Inoltre la soluzione può essere trovata attraverso il metodo iterativo di Picard:

$$Y_i^{k+1} = y_n + h\sum_{j=1}^{s} a_{i,j}\, f(t_n + c_j h\,,\, Y_j^k) \qquad i=1,...,s.$$

dove i valori iniziali Y_j^0 sono assegnati (per esempio pari a y_n per ogni j).

Per i metodi Runge-Kutta (8.6)-(8.7), la funzione incrementale è del tipo:

$$\Phi(t,u) = u + h\sum_{i=1}^{s} b_i\, f(t + c_i h,\, Y_i)$$

dove gli Y_i dipendono anch'essi da u, e sono dati dalla soluzione di (8.7).
Si dimostra facilmente, come è stato fatto per il metodo di Eulero implicito, che la funzione $\Phi(t,u)$ è ancora lipschitziana. Ciò è lasciato al lettore come esercizio.

Esempi:

I metodi di Eulero e dei trapezi sono particolari metodi di Runge-Kutta. In particolare:

Il metodo di **Eulero esplicito** è un metodo di RK esplicito ad un livello con coefficienti:

0	0
	1

Il metodo di **Eulero implicito** è un metodo RK implicito a un livello con coefficienti:

1	1
	1

Il metodo dei **trapezi** è un metodo RK semi-implicito a due livelli con coefficienti:

0	0	0
1	1/2	1/2
	1/2	1/2

Infatti la formula $y_{n+1} = y_n+\frac{1}{2}h(f(t_n, y_n)+f(t_{n+1}, y_{n+1}))$ si può scrivere come:

$$y_{n+1} = y_n+\frac{1}{2}h(f(t_n, Y_1)+f(t_n+h, Y_2))$$

dove Y_1 e Y_2 sono dati dalla soluzione del sistema

$$Y_1 = y_n$$
$$Y_2 = y_n+\frac{1}{2}h(f(t_n, Y_1)+f(t_n+h, Y_2))$$

Analisi dell'errore e convergenza dei metodi di Runge-Kutta:

Detto $e_n := y_n - y(t_n)$ l'errore accumulato fino al passo n-esimo di integrazione, analizziamo come esso si propaga nel passo (n+1)-esimo ed ai successivi. A tale scopo estendiamo la nozione di errore locale di troncamento ad un generico metodo a un passo .

Definizione: *Dato il metodo* $y_{n+1}= y_n + h\Phi(t_n, y_n)$, *chiameremo* **errore locale di troncamento** *al passo n-esimo la quantità:*

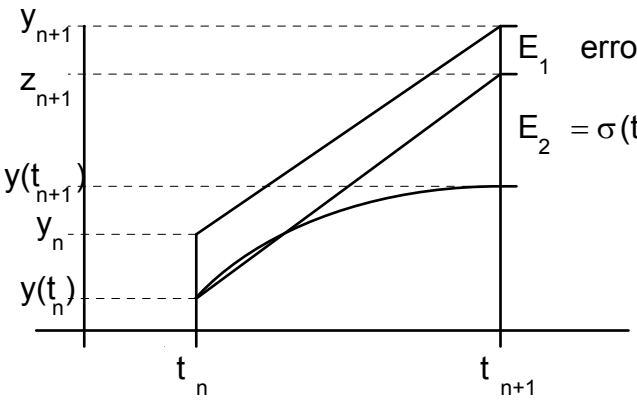
$$\sigma(t_n,h) := y(t_{n+1})-\big(y(t_n) + h\Phi(t_n, y(t_n))\big).$$

Osservato che il termine

$$z_{n+1}= y(t_n) + h\Phi(t_n, y(t_n))$$

è il valore fornito dalla formula qualora essa fosse applicata al punto $y(t_n)$ della traiettoria esatta, si ha

$$\sigma(t_n,h) := y(t_{n+1})- z_{n+1}$$



L'errore totale al passo n+1 è quindi dato, in relazione alla figura, dalla somma di due contributi: l'errore di propagazione E_1 e l'errore locale di troncamento E_2

$$e_{n+1} = y_{n+1} - y(t_{n+1}) = (y_{n+1} - z_{n+1}) + (z_{n+1} - y(t_{n+1})) = (y_{n+1} - z_{n+1}) + \sigma(t_n, h)$$

$$\| e_{n+1} \| \leq \| y_{n+1} - z_{n+1} \| + \| \sigma(t_n, h) \| = E_1 + E_2 \qquad (8.8)$$

Poichè

$$y_{n+1} - z_{n+1} = y_n + h\Phi(t_n, y_n) - (y(t_n) + h\Phi(t_n, y(t_n)))$$

$$y_{n+1} - z_{n+1} = e_n + h(\Phi(t_n, y_n) - \Phi(t_n, y(t_n)))$$

per la lipschitzianità di Φ, dimostrata nel paragrafo precedente, si ha

$$\| y_{n+1} - z_{n+1} \| \leq \| e_n \| + hM \| e_n \| = (1 + hM) \| e_n \|,$$

e tornando alla (8.8) si ottiene:

$$\| e_{n+1} \| < (1 + hM) \| e_n \| + \| \sigma(t_n, h) \|.$$

Maggiorando infine l'errore locale di troncamento $\| \sigma(t_n, h) \|$ in modo uniforme sull'intervallo di integrazione $[t_0, t_f]$

$$\sigma(h) := \max_{t \in [t_0, t_n]} \| \sigma(t, h) \|$$

si ottiene la seguente relazione ricorsiva per l'errore:

$$||e_{n+1}|| < (1+ hM)\,||e_n|| + \sigma(h), \qquad n=0,1,...,N-1. \qquad (8.9)$$

Lemma: *Se la successione* $\{a_n\}$, $a_n>0$, *soddisfa la relazione ricorsiva*

$$a_{n+1}<(1+hQ)a_n + c(h) \qquad n=0,1,2,...,N$$

con $(1+hQ)>0$, *allora vale la maggiorazione:*

$$a_m<(1+hQ)^N a_0 + c(h)\frac{(1+hQ)^N - 1}{hQ} \qquad \forall m\le N.$$

(La dimostrazione è lasciata come esercizio).

Applicando il lemma alla relazione ricorsiva (8.9), tenendo conto che $e_0=0$, si ottiene la maggiorazione uniforme per l'errore:

$$||e_m|| < \sigma(h)\frac{(1+hM)^N - 1}{hM} \qquad \forall m\le N$$

e, tenuto conto della disuguaglianza $(1+ hM)<e^{hM}$,

$$||e_m|| < \sigma(h)\frac{e^{MhN} - 1}{hM} \qquad \forall m\le N$$

Poichè il numero totale di passi N e l'ampiezza del passo h sono legati dalla relazione $Nh= (t_f - t_0)$, si ottiene infine

$$||e_m|| < \sigma(h)\frac{e^{M(t_f - t_0)} - 1}{hM} \qquad \forall m\le N. \qquad (8.10)$$

L'ultima relazione è fondamentale per l'analisi della convergenza del metodo.

Definizione: *Si dirà che il metodo* (8.2) *è* **convergente** *nell'intervallo d'integrazione* $[t_0, t_f]$, *se*

$$\max_{m\le N} ||e_m|| \to 0 \quad \text{per } N\to\infty \text{ e } h\to 0$$

ferma restando la relazione $Nh=(t_f - t_0)$. *Si dirà inoltre che il metodo ha* **ordine di convergenza** *uguale a* p *se il massimo errore sui nodi* $\max_{m\le N} ||e_m||$ *è infinitesimo di ordine* p.

Dalla relazione (8.10) si deduce immediatamente che la convergenza del metodo dipende dall'andamento del termine $\frac{\sigma(h)}{h}$. Vale quindi il seguente teorema di convergenza dei metodi Runge-Kutta e, più in generale, di ogni altro metodo $y_{n+1}= y_n + h\Phi(t_n, y_n)$ con funzione incrementale lipschitziana):

Teorema di convergenza. *Affinchè un metodo* $y_{n+1}= y_n + h\Phi(t_n, y_n)$ *sia convergente di ordine* p *nell'intervallo* $[t_0 , t_f]$ *è sufficiente che la funzione incrementale sia lipschitziana e che il rapporto* $\frac{\sigma(h)}{h}$ (detto **errore di discretizzazione**) *sia infinitesimo di ordine* p (*ossia che l'errore locale di troncamento sia infinitesimo di ordine* p+1 *uniformemente su tutto l'intervallo d'integrazione*).

Dalle espressioni dell'errore locale di troncamento si deduce che, su un intervallo chiuso e limitato $[t_0 , t_f]$, i metodi di Eulero esplicito ed implicito convergono con ordine p=1 per ogni f di classe $C^1(t_0 , t_f)$, mentre il metodo dei trapezi converge con ordine p=2 per ogni f di classe $C^2(t_0 , t_f)$.

<u>*Costruzione di metodi RK di ordine superiore.*</u>

Le formule di Runge-Kutta consentono di ottenere metodi di ordine superiore attraverso un opportuno numero s di livelli ed una opportuna scelta dei coefficienti. In base al *teorema di convergenza,* per ottenere un metodo convergente di ordine p è sufficiente che l'errore locale di troncamento $\sigma(t_n,h)$ sia uniformemente convergente a zero con ordine p+1.
Ciò può essere realizzato sviluppando l'errore locale di troncamento ed imponendo ai parametri in gioco di annullare tutti i termini che sono infinitesimi di ordine minore o uguale a p.

Illustriamo questa procedura attraverso un esempio scalare Vogliamo costruire un metodo di Runge-Kutta esplicito a due livelli di ordine p. In base alle considerazioni fatte in precedenza, la sua tabella sarà del tipo:

0	0	0
c	a	0
	b_1	b_2

con c=a

e la formula sarà, in forma compatta:

$$y_{n+1} = y_n + h\left(b_1 f(t_n, y_n) + b_2 f\left(t_n + ah, y_n + haf(t_n, y_n)\right)\right).$$

L'errore locale di discretizzazione è dato da

$$\sigma(t_n,h) = y(t_{n+1}) - y(t_n) - h\left(b_1 f(t_n, y(t_n)) + b_2 f\left(t_n + ah, y(t_n) + haf(t_n, y(t_n))\right)\right)$$

Sviluppando il termine $y(t_{n+1})$ in un intorno di t_n fino all'ordine 3, ed il termine $f\left(t_n + ah, y(t_n) + haf(t_n, y(t_n))\right)$ in un intorno di $(t_n, y(t_n))$ fino all'ordine 2 rispetto ad h, si trova (tenuto conto che $y''(t_n) = \frac{\partial}{\partial t} f(t_n, y(t_n)) = f_t(t_n, y(t_n)) + f_y(t_n, y(t_n)) y'(t_n)$) :

$$\sigma(t_n,h) = hy'(t_n) + \tfrac{1}{2}h^2\left(f_t(t_n, y(t_n)) + f_y(t_n, y(t_n))y'(t_n)\right) + O(h^3) -$$
$$-h\left[b_1 y'(t_n) + b_2\left(y'(t_n) + ahf_t(t_n, y(t_n)) + ahy'(t_n) f_y(t_n, y(t_n)) + O(h^2)\right)\right]$$

Uguagliando i termini simili in h e ponendoli uguali a zero si trovano le condizioni:

$$b_1 + b_2 = 1$$
$$a\, b_2 = \tfrac{1}{2}$$

per le quali l'errore di troncamento è infinitesimo di ordine 3.

Il precedente sistema ammette infinite soluzioni e quindi esistono infiniti metodi espliciti di Runge-Kutta a due livelli di ordine 2. Tra le possibili scelte troviamo la soluzione $b_1 = b_2 = \frac{1}{2}$ e a=1 che da luogo al **metodo di Heunn**

$$\boxed{y_{n+1} = y_n + h\tfrac{1}{2}\left(f(t_n, y_n) + f\left(t_n + h, y_n + hf(t_n, y_n)\right)\right).}$$

e la soluzione $b_1=0$, $b_2=1$ e $a=\frac{1}{2}$ che da luogo al metodo di **Eulero generalizzato**.

$$\boxed{y_{n+1} = y_n + hf\left(t_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}hf(t_n, y_n)\right).}$$

Si capisce che al crescere del numero dei livelli aumentano i termini dello sviluppo e quindi le condizioni da imporre ai parametri. Per i metodi Runge-Kutta è naturale chiedersi qual'è il massimo ordine che può avere un metodo ad s livelli, ovvero quanti livelli sono necessari per poter raggiungere l'ordine p. La relazione tra il numero di livelli s e il massimo ordine p(s) ottenibile con s livelli è data dalle seguenti tabelle:

Metodi espliciti

s	p(s)
1	1
2	2
3	3
4	4
5	4
6	5
7	5
8	6

Metodi impliciti

p(s)=2s

Alcuni esempi:

Metodo esplicito a 3 livelli di ordine 3;

0	0	0	0
1/3	1/3	0	0
2/3	0	2/3	0
	1/4	0	3/4

metodo esplicito a 4 livelli di ordine 4;.

0	0	0	0	0
1/2	1/2	0	0	0
1/2	0	1/2	0	0
1	0	0	1	0
	1/6	1/3	1/3	1/6

Metodo implicito ad 1 livello di ordine 2;

1/2	1/2
	1

Metodo implicito ad 2 livello di ordine 4;

$(3-\sqrt{3})/6$	1/4	$(3-2\sqrt{3})/12$
$(3+\sqrt{3})/6$	$(3+2\sqrt{3})/12$	1/4
	1/2	1/2

Propagazione dell'errore .

Abbiamo visto in precedenza per il problema iniziale (8.1), che ad ogni passo l'errore e_n si compone di due parti: l'errore di propagazione e l'errore di troncamento. Abbiamo altresì visto che gli errori si accumulano durante il processo di integrazione e la stima (8.10) ne rappresenta una limitazione uniforme su tutto l'intervallo $[t_0, t_f]$. Secondo la stima (8.10) l'errore potrebbe propagarsi lungo l'intervallo di integrazione in maniera drammatica, in dipendenza della costante di Lipschitz M e dell'ampiezza dell'intervallo di integrazione.

Se accade invece che, ad ogni passo, l'errore di propagazione $E_2 := ||y_{n+1} - z_{n+1}||$ risulta non superiore all'errore accumulato fino al passo precedente, cioè se:

$$||y_{n+1} - z_{n+1}|| \leq ||e_n||, \qquad (8.11)$$

allora non si ha propagazione dell'errore e si dice che il metodo è **stabile** (l'errore sul risultato è inferiore all'errore sul dato)

Per i metodi stabili la relazione (8.8) si può infatti sviluppare nel seguente modo:

$$||e_{n+1}|| \leq ||y_{n+1} - z_{n+1}|| + ||\sigma(t_n,h)|| \leq ||e_n|| + ||\sigma(t_n,h)||$$

$$\leq ||e_{n-1}|| + ||\sigma(t_{n-1},h)|| + ||\sigma(t_n,h)|| \leq ...$$

$$\leq ||e_0|| + ||\sigma(t_0,h)|| + ||\sigma(t_1,h)|| + ... + ||\sigma(t_n,h)||$$

$$= ||\sigma(t_0,h)|| + ||\sigma(t_1,h)|| + ... + ||\sigma(t_n,h)||.$$

e l'errore accumulato lungo i passi è dato (in realtà è maggiorato) dalla somma degli errori locali di troncamento.
Poichè, come abbiamo visto, $\sigma(t_k,h)<\sigma(h)$ per ogni k, allora si ottiene :

$$|| e_m || \leq m\, \sigma(h) \leq N\, \sigma(h) = \sigma(h)\, \frac{t_f - t_0}{h} \qquad \forall m \leq N.$$

Ciò significa che, per i metodi stabili, la crescita dell'errore è limitata in modo lineare rispetto all'intervallo $[t_0, t_f]$ anziche in modo esponenziale come indicato dalla (8.10) per un metodo qualunque. Vediamo, a questo proposito, un esempio numerico istruttivo:

Consideriamo il problema scalare:

$$y' = -100y + 100\, \mathrm{sen}(t)$$
$$y(0)=0$$

la cui soluzione esatta è:

$$y(t) = \frac{\mathrm{sen}(t) \; - \; 0.01\cos(t) \; + \; 0.01\, e^{-100t}}{1.001}$$

Supponiamo di integrare il problema nell'intervallo [0,3] con il metodo di RK esplicito di ordine 4. In corrispondenza a vari valori del passo h troviamo le seguenti approssimazioni nel punto finale t_f=3:

h	0.015	0.020	0.025	0.030
N	200	150	120	100
y(3)	0.151004	0.150996	0.150943	$6.7 \cdot 10^{11}$

Cosa è successo nel passare dal passo 0.025 al passo 0.030 ? Siamo passati da una situazione in cui la condizione (8.11) è verificata ad una in cui non lo è più. In altre parole siamo passati da una propagazione lineare ad una proagazione esponenziale dell'errore. Come vedremo tra poco, l'insorgenza del fenomeno di propagazione esponanziale dell'errore dipende sia dal problema trattato che dal metodo impiegato.
Naturalmente sarebbe preferibile utilizzare un metodo per il quale la condizione di stabilità (8.11) fosse verificata *per ogni scelta dal passo*.

In generale è difficile verificare la stabilità dei metodi per equazioni qualunque, e pertanto ci limiteremo a studiare la stabilità per una classe molto particolare di equazioni test.

Consideriamo dapprima la seguente equazione test scalare:

$$y'(t)=\lambda y(t)$$
$$y(0)=y_0 \qquad (8.12)$$

dove, per ragioni che vedremo tra poco, il coefficiente λ, e quindi la funzione y, sono *complessi*. E' noto che la soluzione è data della funzione $y(t)=y_0\, e^{\lambda t}$.

Detto $\lambda=\alpha+i\beta$, si ottiene:

$$y(t)=y_0\, e^{\lambda t}=y_0\, e^{(\alpha+i\beta)t}=y_0\, e^{\alpha t}(\cos\beta t + i\, \mathrm{sen}\beta t)$$

e per i moduli:

$$|y(t)|=|y_0\,|e^{\alpha t}$$

Per quanto riguarda la stabilità del problema (8.12) rispetto alle variazioni sul dato iniziale, sia z(t) la soluzione di (8.12) con dato iniziale z_0 . Per la linearità dell'equazione si ha

$$|y(t)-z(t)|=|y_0-z_0\,|e^{\alpha t}$$

e quindi la condizione $\alpha\leq 0$ è necessaria e sufficiente per avere

$$|y(t)-z(t)|\leq|y_0-z_0\,|e^{\alpha t} \qquad \text{per ogni } t > 0.$$

In questo caso diremo che (8.12) è un **problema stabile**.

Analizziamo ora la stabilità dei metodi numerici per il problema (8.12) nell'ipotesi che il problema stesso sia stabile, cioè che sia $\alpha= \mathrm{Re}(\lambda)<0$.

l metodo di Eulero esplicito, applicato all'equazione test, è:

$$y_{n+1}= y_n+h\lambda y_n= (1+h\lambda)y_n$$

ed il corrispondente valore z_{n+1} è dato da:

$$z_{n+1}= y(t_n)+h\lambda y(t_n)= (1+h\lambda)y(t_n).$$

Si ha quindi, per l'errore propagato:

$$y_{n+1} - z_{n+1} = (1+h\lambda)\, e_n.$$

$$| y_{n+1} - z_{n+1}| = |(1+h\lambda)|\, |e_n|.$$

In base alla definizione precedente, si osserva che il metodo è stabile per quei valori complessi del prodotto $h\lambda$ per i quali si ha:

$$|(1+h\lambda)| \leq 1.$$

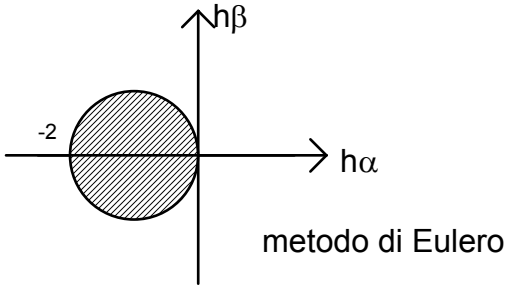
In generale per i metodi RK si ha:

$$| y_{n+1} - z_{n+1}| = |\varphi(h\lambda)|\, |e_n|.$$

La funzione $\varphi(h\lambda)$, detta **funzione di stabilità**, è un polinomio o una funzione razionale, e varia da metodo a metodo. La regione del piano complesso nella quale si ha:

$$|\varphi(h\lambda)| \leq 1$$

è detta **regione di assoluta stabilità** del metodo.
Per il metodo di Eulero la regione di assoluta stabilità è tratteggiata nella seguente figura:



metodo di Eulero

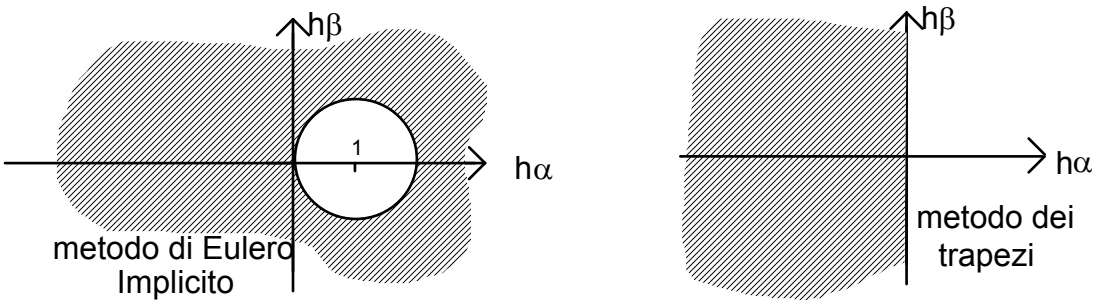
In maniera analoga si trovano le funzioni di stabilità:

$$\varphi(h\lambda) = \frac{1}{1-h\lambda}$$ per il metodo di Eulero implicito

e

$$\varphi(h\lambda) = \frac{1+\frac{h\lambda}{2}}{1-\frac{h\lambda}{2}}$$ per il metodo dei trapezi

alle quali corrispondono le seguenti regioni di assoluta stabilità:



Se la regione di assoluta stabilità include il semipiano negativo C^-, diremo che il metodo è **assolutamente stabile** o **incondizionatamente stabile** inquanto risulta stabile per tutte le equazioni (8.12) stabili e *per ogni passo* h.

Nell'esempio precedente (in cui si aveva $\lambda=-100$), usando un metodo RK esplicito di ordine 4, avevamo osservato una esplosione dell'errore nel passare dal passo h=0.025 al passo h=0.03, valori per i quali il termine $h\lambda$ passava da -2.5 a -3. Ciò è perfettamente in accordo col fatto che la regione di assoluta stabilità del metodo usato è tale che include il punto $h\lambda$ =-2.5 ed esclude il punto $h\lambda$ =-3.

Pur essendo la nostra equazione test di tipo molto particolare, essa può essere utile per analizzare, almeno *localmente*, equazioni più generali del tipo y'=f(t,y). Infatti basta osservare che, in un intorno di (t_n, y_n), essa può essere approssimata dall'equazione (linearizzata): $y'=\frac{\partial}{\partial y}f(t_n, y_n)y$ che rientra nella nostra classe con $\lambda=\frac{\partial}{\partial y}f(t_n, y_n)$.

Equazioni "stiff"

Consideriamo la seguente classe di equazioni differenziali:

$$y'=\lambda\,(y-F(t)) + F'(t)$$

con $\lambda<<-1$ (negativo e grande in modulo). Assegnato il valore iniziale $y(t_0)=y_0$, la soluzione è:

$$y(t) = (y_0 - F(t_0))e^{\lambda(t-t_0)} +F(t)$$

Per ogni $y_0 \neq F(t_0)$, la soluzione y(t) è una funzione che, quando t si allontana da t_0, precipita sulla funzione F(t). Finchè il termine $(y_0 - F(t_0))e^{\lambda(t-t_0)}$ non è trascurabile rispetto a F(t), si è nella **fase transitoria**, altrimenti si è nella **fase stazionaria**, nella quale la soluzione è praticamente uguale a F(t). Evidentemente la fase transitoria è tanto più breve quanto più grande è il modulo di λ. Si osservi però che, anche nella fase stazionaria, se

consideriamo un punto t_n ed un valore perturbato della soluzione y_n, la traiettoria uscente dal punto (t_n, y_n) è

$$x(t) = (y_n - F(t_n))e^{\lambda(t-t_n)} + F(t)$$

che a sua volta precipita sulla funzione F(t) ed è tale che la sua derivata in t_n si discosta sensibilmente da quella della soluzione esatta anche se siamo lontani da t_0. In altre parole, nella fase stazionaria le altre curve integrali sono sensibilmente diverse dalla soluzione esatta.
Come conseguenza, per un assegnato valore del passo h, l'errore di propagazione E_1 del metodo numerico può essere molto grande rispetto all'errore locale di troncamento. Più precisamente, come per l'equazione test (8.12), l'errore propagato dal metodo è:

$$| y_{n+1} - z_{n+1}| = |\varphi(h\lambda)|\, |e_n|.$$

Per i metodi assolutamente stabili si ha $|\varphi(h\lambda)| \leq 1$ per ogni $h\lambda < 0$ e quindi per ogni passo h. Viceversa per i metodi che hanno una regione di stabilità finita la condizione $|\varphi(h\lambda)| \leq 1$ è verificata solo per passi sufficientemente piccoli, tanto più piccoli quanto più grande è il modulo di λ.
Equazioni differenziali le cui soluzioni hanno un comportamento simile a questo sono dette **equazioni stiff** (rigide) e richiedono, per la loro integrazione numerica, l'impiego di metodi assolutamente stabili. In caso contrario esse possono comunque essere integrate con precisione assegnata, ma il passo richiesto diventa drammaticamente piccolo.

Stabilità dei sistemi di equazioni differenziali:

Per l'analisi della stabilità dei sistemi consideriamo ora, come equazione test, il sistema lineare:

$$y'(t) = Ay(t)$$
$$y(0) = u \qquad (8.13)$$

dove $A \in R^{m\times m}$ e $u=(1,1,...,1) \in R^m$. In questo caso il sistema è stabile se e solo se $\mathrm{Re}(\lambda) \leq 0$ per ogni autovalore λ di A

Il metodo di Eulero esplicito applicato al sistema (8.13) è :

$$y_{n+1} = y_n + hAy_n = (I+hA)y_n$$

mentre per il metodo di Eulero Implicito si ha:

$$y_{n+1} = y_n + hAy_{n+1}$$

e quindi

$$y_{n+1} = (I-hA)^{-1} y_n$$

Per il metodo dei trapezi si ha:

$$y_{n+1} = \left(I - \frac{hA}{2}\right)^{-1} \left(I + \frac{hA}{2}\right) y_n$$

e non è difficile vedere, più in generale, che per ogni metodo si ha:

$$y_{n+1} = \varphi(hA) y_n$$

dove la funzione φ, che ora trasforma matrici in matrici, è proprio la funzione di stabilità precedentemente definita per il caso scalare.
Come per il caso scalare, l'errore di propagazione è

$$||y_{n+1} - z_{n+1}|| \leq ||\varphi(hA)||\ ||e_n||$$

ed il metodo è stabile se $||\varphi(hA)|| \leq 1$, cioè se il raggio spettrale, e quindi il modulo di ogni autovalore della matrice $\varphi(hA)$ è ≤ 1.
Ricordando ora che se λ è autovalore di A allora $\varphi(h\lambda)$ è autovalore di $\varphi(hA)$, è sufficiente che sia $|\varphi(h\lambda)| \leq 1$ per ogni λ autovalore di A. Poichè λ è, in generale, un numero complesso, lo studio delle regioni di stabilità per l' equazione test scalare (8.12) è sufficiente anche per il caso vettoriale. Infatti un metodo risulta stabile se $h\lambda$ è incluso nella regione di assoluta stabilità per ogni λ autovalore di A.
In particolare se il metodo è assolutamente stabile, allora la condizione di stabilità

$$||y_{n+1} - z_{n+1}|| \leq ||e_n||$$

è verificata, *indipendentemente dal passo* h, per tutte le equazioni test stabili (cioè con autovalori di A a parte reale negativa).

Stima dell'errore locale ed algoritmi a passo variabile:

E' chiaro che la soluzione di una equazione differenziale puo' avere comportamenti qualitativi molto diversi lungo l'intervallo di integrazione. L'esempio piu' evidente e' quello delle equazioni stiff che, dopo un tratto transitorio nel quale la soluzione subisce una variazione molto rapida, passano al regime stazionario dove la soluzione e' liscia e potrebbe essere integrata con un passo molto piu' grande aumentando l'efficienza dell'algoritmo.

In questo paragrafo si propone una procedura ***empirica*** di integrazione a passo variabile che, ad ogni passo, adatta la lunghezza del passo stesso alle caratteristiche qualitative dell'equazione e della soluzione basandosi su una stima locale dell'errore.

Supponiamo di voler integrare l'equazione differenziale con un metodo di ordine locale p+1. Al passo n-esimo disponiamo del valore approssimato y_n e, utilizzando la formula approssimata con passo h_{n+1}, calcoliamo y_{n+1}.

Contrariamente a quanto fatto per l'analisi della convergenza, indichiamo con z_{n+1} la soluzione esatta uscente dal punto y_n nel punto t_{n+1} e indichiamo con

$$\sigma_{n+1} = ||y_{n+1} - z_{n+1}||$$

l'errore commesso (si noti che questo non e' l'errore locale di troncamento come e' stato definito in precedenza!). Di questo errore possiamo avere una stima utilizzando un metodo di ordine superiore, diciamo p+2, che fornisce il valore approssimato $\bar{y}_{n+1}$ da considerarsi "esatto" rispetto all'approssimazione fornita dal metodo di ordine p+1. Quindi possiamo concludere che, utilizzando il metodo di ordine p+1, si e' commesso un errore che, a meno di un infinitesimo di ordine p+2, vale

$$\sigma_{n+1} \approx ||y_{n+1} - \bar{y}_{n+1}|| \ .$$

A questo punto sottoponiamo l'errore σ_{n+1} cosi' stimato, al ***test di tolleranza***

$$\sigma_{n+1} \leq TOL \cdot h_{n+1}$$

dove TOL e' la ***tolleranza per unita' di passo***, cioe' il massimo errore che intendo accettare per un passo di integrazione di ampiezza $h_{n+1}=1$.

Passo rifiutato: Se il test non viene superato**,** il valore y_{n+1} viene *rifiutato* e la formula d'integrazione viene ricalcolata con un nuovo passo d'integrazione h_{new} , inferiore ad h_{n+1}.

La riduzione del passo non viene fatta in maniera arbitraria, per esempio dimezzando la lunghezza del passo rifiutato, ma viene valutata in maniera “ottimale” utilizzando i calcoli gia’ eseguiti. A tale scopo si osservi che l’errore σ_{n+1} ha la forma $K \cdot h^{p+1}$ per qualche valore di K che non conosco ma posso stimare dall’uguaglianza

$$\sigma_{n+1} = ||y_{n+1} - \bar{y}_{n+1}|| = k \cdot h^{p+1}.$$

Ottengo cosi’ una stima di K

$$K = \frac{\sigma_{n+1}}{h_{n+1}^{p+1}}$$

che ritengo valida anche per piccole variazioni del passo. A questo punto posso dire che, per il nuovo passo h_{new}, commettero’ un errore stimabile, a priori, in $K \cdot (h_{new})^{p+1}$.
Per passare il test di tolleranza con il nuovo passo, richiedero’ che tale errore soddisfi

$$K \cdot (h_{new})^{p+1} \leq TOL \cdot h_{new}.$$

Per evitare che il test fallisca a causa dei termini trascurati (che, sebbene di ordine superiore, possono compromettere la stima del nuovo passo) il nuovo passo viene calcolato sulla base della richiesta piu’ stringente

$$K \cdot (h_{new})^{p+1} = \tfrac{1}{2} TOL \cdot h_{new}. \qquad (1.14)$$

Da quest’ultima posso quindi ricavare una stima per h_{new}:

$$h_{new} = \sqrt[p]{\frac{TOL}{2K}} = \sqrt[p]{\frac{TOL \cdot h_{n+1}^{p+1}}{2\sigma_{n+1}}} = h_{n+1} \sqrt[p]{\frac{TOL \cdot h_{n+1}}{2\sigma_{n+1}}}$$

Poiche’ in prossimita’ di forti variazioni della soluzione il fattore di riduzione

$$R = \sqrt[p]{\frac{TOL \cdot h_{n+1}}{2\sigma_{n+1}}}$$

puo’ risultare molto piccolo, quando $\sigma_{n+1} >> 1$, allora, per evitare una riduzione eccessiva del passo , si definisce a priori una riduzione massima del passo, diciamo *non meno della meta’*, e quindi si definisce l’ampiezza del nuovo passo di tentativo

$$h_{new} = h_{n+1} \cdot \max\{½ , R\}$$

Con tale passo, rinominato h_{n+1} ($\leftarrow h_{new}$), si ripete l'intera procedura finche' il test di tolleranza viene superato. Quando cio' accade, il valore y_{n+1} viene accettato e si passa al passo successivo.

Passo accettato. Quando il valore y_{n+1} viene accettato e si passa al passo successivo, la procedura puo' essere ottimizzata utilizzando un passo di tentativo

$$h_{n+2} = R\, h_{n+1}$$

dove, per la (1,14), sara' R>1. Come nel caso della riduzione, anche nel caso di espansione del passo si pone una *protezione* del tipo

$$h_{n+2} = h_{n+1} \cdot \min\{2 , R\}$$

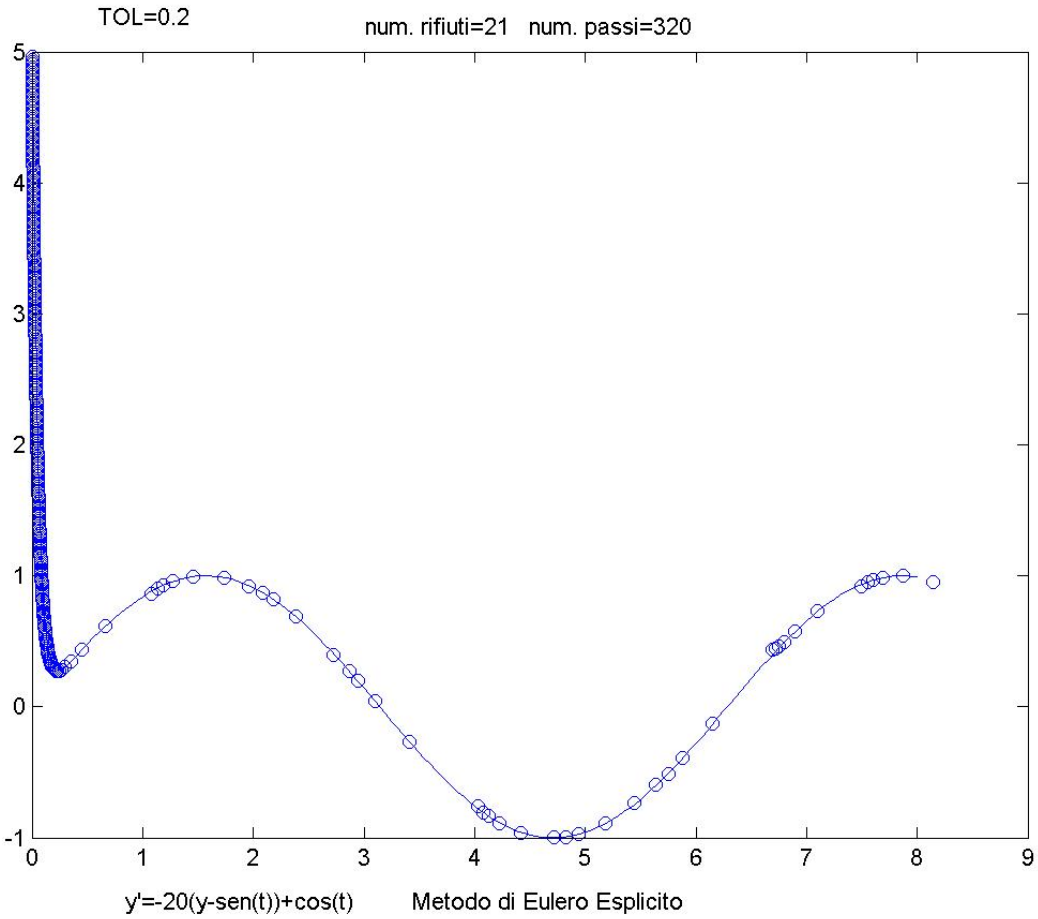
che ne limita l'allungamento.

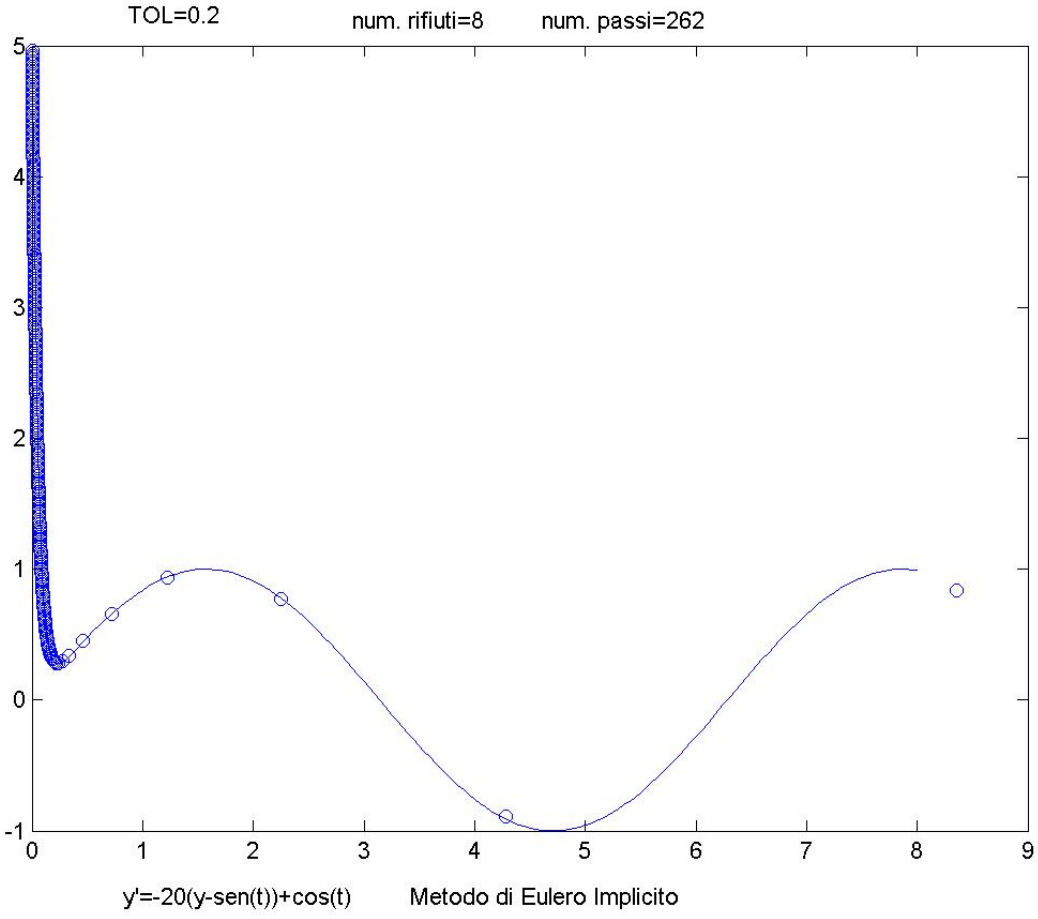
Con questo passo di tentativo, si applica la procedura descritta e si procede fino all'esaurimento dell'intervallo di integrazione.

Un'attenzione particolare va dedicata al primo passo per il quale si deve partire da un valore h_1 di tentativo e accorciarlo o allungarlo fino al primo passaggio o, rispettivamente, al primo rifiuto del test di tolleranza.

Per quanto riguarda la stima dell'errore, ci sono vari metodi. Negli esempi che seguono, si e' usato il metodo di estrapolazione di Richardson, del quale saltiamo la descrizione, ed i metodi di Runge-Kutta-Fehlberg che sono descritti nel paragrafo successivo.

Si noti che nell'esempio riportato i metodi di EE ed EI sono sostanzialmente equivalenti nella fase transitoria, mentre hanno un comportamento ben diverso nella fase stazionaria.


TOL=0.2
num. rifiuti=21 num. passi=320
y'=-20(y-sen(t))+cos(t) Metodo di Eulero Esplicito


TOL=0.2
num. rifiuti=8 num. passi=262
y'=-20(y-sen(t))+cos(t) Metodo di Eulero Implicito

Metodi di Runge-Kutta-Fehlberg:

I metodi di Runge-Kutta-Fehlberg (RKF) sono metodi progettati per fornire, in modo economico, coppie di metodi RK di ordine p e p+1 basati sull'utilizzo dello stesso insieme di livelli Y che vengono utilizzati con due insiemi diversi di pesi .

Metodo RKF23: E' un metodo esplicito a 3 livelli che fornisce la coppia di approssimazioni di ordine globale p=2 e p=3

0	0	0	0
1	1	0	0
1/2	1/4	1/4	0
	1/2	1/2	0
	1/6	1/6	2/3

I livelli Y_1, Y_2, Y_3 sono dati dalla soluzione del sistema:

$$Y_1 = y_n$$
$$Y_2 = y_n + hf(t_n , Y_1)$$
$$Y_3 = y_n + h(¼\, f(t_n , Y_1) + ¼\, f(t_n + h, Y_2))$$

mentre i 2 valori approssimati della soluzione $y(t_{n+1})$ sono dati da

$$y_{n+1} = y_n + h/2\left(f(t_n , Y_1) + f(t_n + h, Y_2)\right) \quad \text{metodo di ordine 2}$$

$$\bar{y}_{n+1} = y_n + h/6\left(f(t_n , Y_1) + f(t_n + h, Y_2) + 4\, f(t_n + h/2, Y_3)\right) \quad \text{metodo di ordine 3.}$$

Si osservi che il costo globale di RKF23 e' quello di un metodo a 3 livelli espliciti, cioe' 3 valutazioni della funzione f. Se avessi usato un metodo RK di ordine 2 ed un altro di ordine 3 avrei dovuto calcolare 5 valutazioni della f.

Metodo RKF45: E' un metodo esplicito a 6 livelli che fornisce la coppia di approssimazioni di ordine globale 4 e 5.

Lo schema dei coefficienti e':

0	0	0	0	0	0	0
2/9	2/9	0	0	0	0	0
1/3	1/12	1/4	0	0	0	0
3/4	69/128	-243/128	135/64	0	0	0
1	-17/12	27/4	-27/5	16/15	0	0
5/6	65/432	-5/16	13/16	4/27	5/144	0
	1/9	0	9/20	16/45	1/12	0
	47/450	0	12/25	32/225	1/30	6/25

Qui il risparmio computazionale e' superiore perche' sono sufficienti 6 valutazioni di f contro le 10 necessarie per implementare un RK di ordine 4 ed uno di ordine 5.

Analisi asintotica della soluzione

Per quanto riguarda l'andamento asintotico della soluzione di equazioni differenziali, riferiamoci ancora all'equazione test (8.12), ed osserviamo che, se $\alpha<0$, la soluzione

$$y(t)=y_0 e^{\alpha t}(\cos\beta t + i \operatorname{sen}\beta t)$$

tende a zero per t che tende a infinito. In altre parole la componente reale e immaginaria di y(t) tendono entrambe a zero. In questo caso si dice che la soluzione è **asintoticamente stabile**.
Se invece $\alpha=0$, allora la soluzione ha modulo costante uguale ad y_0, mentre le componenti oscillano periodicamente. Infine, se $\alpha>0$ la soluzione diverge.

Abbiamo visto che i vari metodi numerici, applicati all'equazione test, assumono la forma:

$$y_{n+1}= \varphi(h\lambda)y_n$$

per cui

$$|y_{n+1}| = |\varphi(h\lambda)|\,|y_n| = |\varphi(h\lambda)|^2\,|y_{n-1}| = \ldots = |\varphi(h\lambda)|^{n+1}\,|y_0|.$$

Tale relazione dice che la soluzione numerica ottenuta con passo h costante ha un comportamento asintotico che dipende da $|\varphi(h\lambda)|$ nel seguente modo:

$|\varphi(h\lambda)| < 1 \quad \Rightarrow \quad |y_n| \to 0$ per $n \to \infty$ (**metodo asintoticamente stabile**)

$|\varphi(h\lambda)| = 1 \quad \Rightarrow \quad |y_n| = |y_0| \quad \forall n$ (**metodo stazionario**)

$|\varphi(h\lambda)| > 1 \quad \Rightarrow \quad |y_n| \to \infty$ per $n \to \infty$.(**metodo instabile**)

Quindi sono asintoticamente stabili i metodi implementati con un passo tale che $h\lambda$ sia *interno* alla regione di assoluta stabilità; sono stazionari quelli per cui $h\lambda$ sta sul bordo della regione di assoluta stabilità e sono instabili quelli per cui $h\lambda$ è esterno alla regione.

Sono interessanti i metodi che risultano asintoticamente stabili per tutte le equazioni che hanno soluzioni asintoticamente stabili, cioè $\alpha<0$ Dalle considerazioni precedenti risulta che il metodo di Eulero esplicito è asintoticamente stabile solo per certi valori del passo h. Viceversa i metodi di Eulero implicito e dei trapezi risultano asintoticamente stabili per ogni valore del passo h, poichè le loro regioni di assoluta stabilità includono l'intero semipiano negativo.

Dunque i metodi assolutamente stabili sono anche asintoticamente stabili indipendentemente dal passo, sono cioè **incondizionatamente asintoticamente stabili**.

Si osservi infine che il metodo di Eulero implicito ha una regione di assoluta stabilità più ampia del semipiano negativo. Ciò causa, per certi valori del passo, un andamento asintoticamente stabile del metodo anche per equazioni che hanno $\alpha>0$, le cui soluzioni esatte divergono. Questa proprietà, nota come *smorzamento numerico* (numerical damping), è un aspetto negativo del metodo.

Un metodo perfetto, da questo punto di vista, è il metodo dei trapezi la cui soluzione ha, in ogni caso, lo stesso andamento qualitativo della soluzione esatta per ogni passo h.

Analisi asintotica dei sistemi:

Analogamente al caso scalare, l'analisi asintotica è fatta sull'equazione test (8.13):

$$y'(t) = Ay(t)$$
$$y(0) = u$$

per la quale è noto che la soluzione è asintoticamente stabile (tende a zero) se e solo se tutti gli autovalori di A hanno parte reale negativa.
Per ogni metodo numerico, dalla relazione

$$y_{n+1}=\varphi(hA)y_n$$

si ottiene:

$$||y_{n+1}||\leq||\varphi(hA)||\ ||y_n||\leq\ldots\leq||\varphi(hA)||^{n+1}\ ||y_0||$$

Come nel caso scalare, sono interessanti quei metodi numerici che risultano asintoticamente stabili quando si applicano ad una equazioni test (8.13) che sia asintoticamente stabile. Si vorrebbe cioè che sia $||\varphi(hA)||<1$, per ogni matrice A con autovalori a parte reale negativa.
Abbiamo già visto che ciò accade se $|\varphi(h\lambda)|<1$ per ogni $h\lambda$ con λ autovalore di A.

In particolare se il metodo è assolutamente stabile, allora la soluzione numerica tende a zero, *indipendentemente dal passo* h, per tutte le equazioni con soluzione asintoticamente stabile.

Un esempio di sistema stiff. Il metodo delle linee

Consideriamo l'equazione del calore in una dimensione spaziale:

$$\frac{\partial}{\partial t} y(t,x) = \frac{\partial^2}{\partial x^2} y(t,x) \qquad \text{per } t \in [0,T], \text{ ed } \quad x \in [0,1]$$

con le condizioni iniziali (rispetto a t) ed ai limiti (rispetto ad x):

$$y(0,x) = g(x)$$
$$y(t,0) = y(t,1) = 0$$

Discretizziamo il problema rispetto alla variabile spaziale x sui nodi x_i=ih, i=0,...,m con passo h=1/m.
Per ogni t e per ogni x_i, approssimiamo la derivata seconda rispetto ad x con la nota differenza centrale seconda:

$$\frac{\partial^2}{\partial x^2} y(t,x_i) = \frac{y(t,x_{i-1}) - 2y(t,x_i) + y(t,x_{i+1})}{h^2} \qquad i = 1,...,m-1$$

Otteniamo così, per ogni coordinata spaziale x_i, l'equazione differenziale:

$$\frac{\partial}{\partial t} y(t, x_i) = \frac{y(t,x_{i-1}) - 2y(t,x_i) + y(t,x_{i+1})}{h^2} \qquad i = 1,...,m$$

Con la notazione $y_i(t)=y(t, x_i)$ si ottiene, tenuto conto che y(t, 0)=y(t, 1)=0, il seguente sistema di equazioni differenziali:

$$\begin{bmatrix} y'_1(t) \\ y'_2(t) \\ \vdots \\ \vdots \\ y'_{m-1}(t) \end{bmatrix} = m^2 \begin{vmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & -2 \end{vmatrix} \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ \vdots \\ y_{m-1}(t) \end{bmatrix}$$

con le condizioni iniziali: $y_i(0) = y(0,x_i) = g(x_i)$ $\quad i = 1,...,m-1$.
Gli autovalori della matrice sono:

$$\lambda_i = m^2\left(-2 + 2\cos\left(\frac{i}{m}\pi\right)\right) \qquad i = 1,...,m-1$$

In particolare sono compresi tra $\lambda_{m-1} \cong -4m^2$ e $\lambda_1 \leq -\pi^2$. Il sistema è quindi stabile e, per m grande, richiede l'impiego di un metodo assolutamente stabile anche in fase stazionaria.

2. PROBLEMI AI LIMITI

Consideriamo alcuni problemi ai limiti e alcuni metodi numerici per la loro risoluzione approssimata.

Il metodo "shooting" per i problemi ai limiti

Un problema del primo ordine:

Consideriamo il seguente problema differenziale del primo ordine

$$y'(t)=f(t,y(t)) \qquad \text{per } t\in[a,b]$$

con la seguente condizione sui valori agli estremi y(a) ed y(b):

$$g(y(a),y(b))=0$$

e supponiamo che tale problema ammetta una soluzione. Una possibile condizione è, per esempio, quella di periodicità y(a)-y(b)=0.
Supponiamo inoltre che siano verificate le ipotesi per l'esistenza e l'unicità del problema di Cauchy associato all'equazione data. In questo caso supporremo che per ogni condizione iniziale $y(a)=\xi$ esista una ed una sola soluzione, che indicheremo con $y(t,\xi)$. Detta $y(b,\xi)$ tale soluzione nel punto b, è chiaro che se la coppia $(\xi, y(b,\xi))$ soddisfa la condizione ai limiti

$$g(\xi, y(b,\xi))=0$$

allora la funzione $y(t,\xi)$ è la soluzione del nostro problema. Ho così trasformato il problema ai limiti nella ricerca della radice dell'equazione

$$F(\xi):=g(\xi, y(b,\xi))=0 \qquad (8.14)$$

E' evidente che, a parte qualche caso eccezionale, non dispongo dell'espressione esplicita della funzione $F(\xi)$ in quanto essa dipende dal termine funzionale $y(b,\xi)$ il cui valore si ottiene attraverso la risoluzione di un problema di Cauchy.

Disponendo di metodi efficienti per la risoluzione del problema di Cauchy, posso però calcolare, in modo approssimato, il valore di $y(b,\xi)$ e quindi di $F(\xi)$ per ogni ξ.
Inoltre, disponendo di metodi veloci per la ricerca delle radici di equazioni non lineari che fanno uso solo di valutazioni puntuali della funzione F, posso ottenere una soluzione approssimata dell'equazione (8.14).
Il valore ξ^* così trovato è una approssimazione del valore iniziale y(a) (incognito) dal quale "sparare" un punto materiale la cui traiettoria soddisfa l'equazione differenziale e raggiunge nel punto b il valore y(b) richiesto affinchè la condizione ai limiti $g(y(a),y(b))=0$ sia soddisfatta. Da qui l'origine della denominazione di "metodo shooting"

Il metodo shooting, che si applica in modo sostanzialmente uguale anche in circostanze leggermente diverse, valorizza i metodi iterativi di Steffensen e delle secanti che non richiedono valutazioni puntuali della derivata della funzione $F(\xi)$.

Problemi del secondo ordine:

Consideriamo l'equazione differenziale del secondo ordine:

$$y''=f(t,y,y') \qquad \text{in } [a,b] \qquad (8.15)$$

con le condizioni ai limiti

$$g_1(y(a),y'(a),y(b),y'(b))=0$$
$$g_2(y(a),y'(a),y(b),y'(b))=0$$

per le quali supporremo l'esistenza della soluzione in [a,b]. Sono di particolare interesse le seguenti condizioni ai limiti:

$$y(a)=\alpha$$
$$y(b)=\beta \qquad \text{(problema classico dei due punti)}$$

$$y(a)-y(b)=0$$
$$y'(a)-y'(b)=0 \qquad \text{(problema periodico)}$$

$$g_1(y(a),y'(a))=0$$
$$g_2(y(b),y'(b))=0 \qquad \text{(condizioni di Sturm-Liouville)}$$

Anche qui supponiamo che il problema di Cauchy per l'equazione (8.15) ammetta una ed una sola soluzione. Una volta trasformata l'equazione (8.15) in un sistema del primo ordine nelle incognite u=y e v=y',

$$u'=v$$
$$v'=f(t,u,v) \qquad (8.16)$$

il problema ai limiti ammette per soluzione quella relativa ad opportuni valori iniziali u(a)=y(a) e v(a)=y'(a) da determinarsi.
Per esempio nel problema classico dei due punti, è noto u(a)=y(a)=α ma non e' noto v(a)=y'(a). Allora fissiamo v(a)=ξ ed indichiamo con u(t,ξ) la funzione u(t) soluzione di (8.16) con condizioni iniziali u(a)=α, v(a)=ξ. Naturalmente noi cerchiamo il valore di ξ per il quale si ha u(b,ξ)=β.
Come nel caso precedente, indichiamo con F(ξ) l'operatore che associa a ξ il valore u(b,ξ). L'equazione da risolvere è in questo caso è:

$$F(\xi)=\beta.$$

Leggermente più complesso è il caso delle condizioni periodiche nel quale non è noto né u(a) né v(a) che devono essere visti come incognite

$$u(a)=\xi \qquad v(a)=\eta \qquad (8.17)$$

Dette u(t,ξ,η) e v(t,ξ,η) le soluzioni dell'equazione (8.16) con condizioni iniziali (8.17), si tratta di determinare ξ e η in modo che nel punto finale b siano verificate le condizioni di periodicità

$$\xi=u(b,\xi,\eta) \quad \text{e} \quad \eta=v(b,\xi,\eta).$$

Definiamo l'operatore F(ξ,η) che associa alla coppia di valori iniziali x=(ξ , η), la coppia di valori finali (u(b,ξ,η), v(b,ξ,η)). Questa volta l'equazione da risolvere (in R^2 nel caso scalare, ed in R^{2m} nel caso dei sistemi) è

$$F(x)=x.$$

Il caso lineare

Consideriamo il caso di una equazione differenziale lineare, con condizioni ai limiti lineari:

$$y'=g(t)y + f(t) \qquad t\in[a,b]$$
$$\alpha y(a)+\beta y(b)=d.$$

Detto ξ il valore incognito di y(a), il metodo *shooting* consiste nella ricerca della radice dell'equazione $F(\xi)=0$ dove, con le solite notazioni, la funzione F è:

$$F(\xi)=\alpha\xi+\beta y(b,\xi) - d.$$

Osserviamo che la funzione F è una funzione affine in ξ. Infatti, detto u(t) l'integrale della equazione omogenea associata ed $\bar{y}(t)$ un integrale particolare della completa, l'integrale generale è $y(t)=cu(t)+\bar{y}(t)$ dove la costante c è determinata dalla condizione iniziale $y(a)=\xi$:

$$\xi=cu(a)+\bar{y}(a)$$
$$c=\frac{\xi-\bar{y}(a)}{u(a)}$$

La soluzione, relativa al dato iniziale ξ, è quindi $y(t,\xi)=\frac{\xi-\bar{y}(a)}{u(a)}u(t) + \bar{y}(t)$ che, calcolata in b, fornisce $y(b,\xi)=\frac{\xi-\bar{y}(a)}{u(a)}u(b) + \bar{y}(b)$. La funzione $F(\xi)$ assume quindi la forma affine:

$$F(\xi)=\alpha\xi+\beta\frac{\xi-\bar{y}(a)}{u(a)}u(b) + \beta\bar{y}(b) - d.$$

Di conseguenza i metodi iterativi delle secanti o di Steffensen raggiungono la soluzione esatta dopo una sola iterazione. (Si osservi che l'affinità di $F(\xi)$ è stata dimostrata nell'ipotesi che la soluzione $y(b,\xi)$ sia esatta. Il lettore dimostri che la proprietà rimane verificata anche quando $y(b,\xi)$ è ottenuto attraverso un metodo di RK con un numero indeterminato di passi).

Il metodo delle differenze

Abbiamo già visto all'inizio del secondo capitolo che il problema dei due punti:

$$y'' - g(t)y = f(t) \qquad t\in[a,b]$$
$$y(a)=\alpha, \qquad y(b)=\beta$$

con $g(t)\geq 0$, può essere risolto numericamente attraverso il *metodo delle differenze* approssimando la derivata seconda, in ogni punto della griglia, con una differenza centrale seconda:

$$y''(t) = \frac{y(t-h) - 2y(t) + y(t+h)}{h^2} + \sigma(t, h)$$

dove $\sigma(t, h)=O(h^2)$ uniformemente in [a,b]. Qui si vuole dimostrare che il procedimento è convergente.
Con le notazioni usate nel capitolo secondo, Il metodo si riduce alla risoluzione di un sistema lineare

$$AY = c \qquad (8.18)$$

nell'incognita $Y= (y_1,y_2,...,y_{N-1})^T$ dove la matrice A ha la forma:

$$A = 1/h^2 \begin{Vmatrix} 2+ h^2g_1 & -1 & & & & \\ -1 & 2+ h^2g_2 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -1 & 2 + h^2g_i & -1 & \\ & & & \ddots & \ddots & \ddots \\ & & & -1 & 2+ h^2g_{N-2} & -1 \\ & & & & -1 & 2+ h^2g_{N-1} \end{Vmatrix}$$

Viceversa, per il vettore $e= (y(t_1)-y_1, y(t_2)-y_2,..., y(t_{N-1})-y_{N-1})^T$ degli errori nei punti della griglia si ha:

$$Ae=\sigma \qquad (8.19)$$

dove σ è il vettore degli errori locali di troncamento:

$$\sigma= (\sigma(t_1,h),\sigma(t_2,h),,...,\sigma(t_{N-1},h))^T$$

Abbiamo già dimostrato nel capitolo secondo che, per ogni N, la matrice A è non singolare e definita positiva. Si ha quindi

$$||e|| \leq ||A^{-1}||\ ||\sigma||$$

dove $||\sigma||=O(h^2)$. Per ottenere la convergenza del metodo delle differenze per $N \longrightarrow +\infty$ (o se si preferisce per $h \longrightarrow 0$) è necessario avere la equilimitatezza di $||A^{-1}||$ rispetto ad N.

Poichè A è simmetrica e definita positiva, conviene considerare la norma 2 e osservare che $||A^{-1}||_2 = \rho(A^{-1}) = 1/\lambda_{min}(A)$
Quindi basta trovare una costante K tale che $1/\lambda_{min}(A) \leq K$.
Detto x l'autovettore corrispondente all'autovalore di modulo minimo di A, si ha

$$x^T Ax = \lambda_{min}(A) x^T x.$$

Separando A nella somma A=T+D (entrambe definite positive) si ha, ricordando che $\lambda_{min}(T) > \pi^2$ per ogni N e che $g(t) \geq 0$,

$$\lambda_{min}(A) x^T x = x^T Tx + x^T Dx \geq x^T Tx$$

Poiche' ogni x e' ottenibile come combinazione lineare di autovettori (ortogonali) di T: $x = \sum c_i y_i$, si ha $Tx = \sum c_i T y_i = \sum c_i \lambda_i(T) y_i$ e quest'ultima, moltiplicata a sinistra per $x^T = \sum c_i y_i^T$, fornisce≥

$$x^T Tx = \sum c_i^2 \lambda_i(T) \geq \lambda_{min}(T) \sum c_i^2 = \lambda_{min}(T) x^T x > \pi^2 x^T x.$$

In conclusione, $\lambda_{min}(A) > \pi^2$ e quindi:

$$||A^{-1}||_2 = 1/\lambda_{min} < 1/\pi^2$$

indipendentemente dalla dimensione di A.

Definizioni: La condizione $||\sigma|| = O(h^2)$ viene indicata come la **consistenza** dell'operatore di discretizzazione rispetto all'operatore differenziale e l'ordine di infinitesimo viene indicato come l'ordine di consistenza. La condizione dell'equilimitatezza dell'operatore A^{-1} viene invece indicata come la **stabilita'** dell'operatore stesso.

Abbiamo dimostrato che, per lo schema delle differenze centrali applicato al problema dei due punti, vale la relazione

consistenza + stabilita' = convergenza

Questa relazione, che si estende a molti schemi di approssimazione per problemi di equazioni differenziali, mette in evidenza che la sola consistenza non e', in generale, sufficiente a garantire la convergenza del metodo.

3. EQUAZIONI ALLE DERIVATE PARZIALI:

Consideriamo ancora l'equazione del calore monodimensionale

$$\frac{\partial}{\partial t} y(t,x) = a \frac{\partial^2}{\partial x^2} y(t,x) \qquad \text{per } t \in [0,T], \text{ ed } \quad x \in [0,1]$$

con le condizioni iniziali (rispetto a t) ed ai limiti (rispetto ad x):

$$y(0,x) = g(x)$$
$$y(t,0) = y(t,1) = 0$$

Abbiamo visto che discretizzando il problema rispetto alla variabile spaziale x sui nodi x_i=ih, i=0,...,m con passo h=1/m il problema si riduce ad un sistema di equazioni differenziali ordinarie

$$\frac{\partial}{\partial t} y(t, x_i) = \frac{y(t,x_{i-1}) - 2y(t,x_i) + y(t,x_{i+1})}{h^2} + \frac{h^2}{12} \frac{\partial^4}{\partial x^4} y(t,\eta_i) \qquad i = 1,...,m-1$$

con le condizioni iniziali: $y(0,x_i) = g(x_i)$ $\quad i = 1,...,m-1$.

Completiamo ora il processo di discretizzazione, discretizzando anche la derivata temporale con una differenza prima in avanti di passo k=T/N sui nodi t_n=nk, n=0,1,...,N

$$\frac{\partial}{\partial t} y(t_n, x_i) = \frac{y(t_{n+1},x_i) - y(t_n,x_i)}{k} + \frac{k}{2} \frac{\partial^2}{\partial t^2} y(\xi_n, x_i)$$

Si ottiene così l'uguaglianza:

$$\frac{y(t_{n+1},x_i) - y(t_n,x_i)}{k} = a \frac{y(t_n,x_{i-1}) - 2y(t_n,x_i) + y(t_n,x_{i+1})}{h^2} + E_{n,i} \qquad i = 1,...,m-1 \tag{8.20}$$

dove $E_{n,i}$ è la somma dei due errori di troncamento

$$E_{n,i} = \frac{k}{2} \frac{\partial^2}{\partial t^2} y(\xi_n, x_i) + a \frac{h^2}{12} \frac{\partial^4}{\partial x^4} y(t_n, \eta_i).$$

Sotto le solite ipotesi di regolarita' della soluzione, la discretizzazione globale che abbiamo applicato all'operatore differenziale $y_t = a y_{xx}$ e' consistente e l'errore di troncamento puo' essere maggiorato uniformemente con

$$E = \max_{n,i} ||E_{n,i}|| = O(k, h^2).$$

Trascurando l'errore di troncamento $E_{n,i}$, si ottiene, per ogni scelta della griglia spazio-temporale, il seguente **schema alle differenze in avanti**

$$\frac{y_{i,n+1} - y_{i,n}}{k} = a\frac{y_{i-1,n} - 2y_{i,n} + y_{i+1,n}}{h^2} \qquad i = 1,...,m\text{-}1 \qquad (8.21)$$

dove, per ogni i=1,...,m-1 ed n=0,1,...,N, $y_{i,n}$ è l'approssimazione numerica di $y(t_n, x_i)$. Le condizioni ai limiti sono, per ogni n,

$$y_{0,n} = y_{m,n} = 0,$$

e quelle iniziali sono, per ogni x_i

$$y_{i,0} = y(0, x_i) = g(x_i) \qquad i = 1,...,m\text{-}1.$$

Il metodo si esprime quindi con la seguente relazione ricorsiva nell' incognita $Y_{n+1}=(y_{1,n+1}, y_{2,n+1},..., y_{m-1,n+1})$:

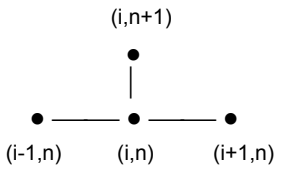
$$\begin{bmatrix} y_{1,n+1} \\ y_{2,n+1} \\ \vdots \\ \vdots \\ y_{m-1,n+1} \end{bmatrix} = \frac{ak}{h^2} \begin{vmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & -2 \end{vmatrix} \begin{bmatrix} y_{1,n} \\ y_{2,n} \\ \vdots \\ \vdots \\ y_{m-1,n} \end{bmatrix} + \begin{bmatrix} y_{1,n} \\ y_{2,n} \\ \vdots \\ \vdots \\ y_{m-1,n} \end{bmatrix}$$

ovvero:

$$Y_{n+1} = (I + \lambda T)\, Y_n$$

dove $\lambda = \frac{ak}{h^2}$.

Per ogni punto (t_n, x_i) della griglia di indici (i,n), l'equazione discretizzata coinvolge i quattro indici (i,n+1), (i-1,n), (i,n) e (i+1,n) secondo il seguente schema detto **FC** come acronimo per *differenze in avanti* (“Forward”) nel tempo e *differenze centrali* (“Centered”) nello spazio.



Sottranendo (8.21) da (8.20) si ottiene, per l'errore $e_{i,n}:=y(t_n,x_i)- y_{i,n}$ n=0,...,N, la relazione ricorsiva

$$\begin{bmatrix} e_{1,n+1} \\ e_{2,n+1} \\ \vdots \\ \vdots \\ e_{m-1,n+1} \end{bmatrix} = \frac{ak}{h^2} \begin{vmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & -2 & 1 \\ & & & 1 & -2 \end{vmatrix} \begin{bmatrix} e_{1,n} \\ e_{2,n} \\ \vdots \\ \vdots \\ e_{m-1,n} \end{bmatrix} + \begin{bmatrix} e_{1,n} \\ e_{2,n} \\ \vdots \\ \vdots \\ e_{m-1,n} \end{bmatrix} + kE_n.$$

In forma compatta si ha,

$$U_{n+1}=(I+\lambda T)U_n + kE_n \qquad \text{per n=1,2,...N-1}$$

dove $U_{n+1}=(e_{1,n+1}, e_{2,n+1},..., e_{m-1,n+1})^T$ e $E_n=(E_{1,n+1}, E_{2,n+1},..., E_{m-1,n+1})^T$.

Se la derivata temporale fosse stata approssimata con una differenza all'indietro

$$\frac{\partial}{\partial t}y(t_n, x_i)= \frac{y(t_n,x_i)-y(t_{n-1},x_i)}{k}+\frac{k}{2}\frac{\partial^2}{\partial t^2}y(\xi_n, x_i)$$

avrei ottenuto la relazione

$$\frac{y(t_n,x_i)-y(t_{n-1},x_i)}{k} = a\frac{y(t_n,x_{i-1})-2y(t_n,x_i)+y(t_n,x_{i+1})}{h^2} + E_{n,i} \qquad i = 1,...,m-1$$

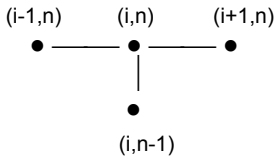
e quindi, trascurando l'errore, il seguente **schema alle differenze all'indietro**:

$$\begin{bmatrix} y_{1,n} \\ y_{2,n} \\ \vdots \\ \vdots \\ y_{m-1,n} \end{bmatrix} - \frac{ak}{h^2} \begin{vmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & -2 & 1 \\ & & & 1 & -2 \end{vmatrix} \begin{bmatrix} y_{1,n} \\ y_{2,n} \\ \vdots \\ \vdots \\ y_{m-1,n} \end{bmatrix} = \begin{bmatrix} y_{1,n-1} \\ y_{2,n-1} \\ \vdots \\ \vdots \\ y_{m-1,n-1} \end{bmatrix}$$

Come si può osservare, in questo caso la relazione per la soluzione numerica $Y_n=(y_{1,n}, y_{2,n},..., y_{m-1,n})$ è implicita e richiede, ad ogni passo temporale, la risoluzione di un sistema lineare

$$(I-\lambda T)\, Y_n= Y_{n-1}$$

Per ogni punto (t_n,x_i) della griglia di indici (i,n), l'equazione discretizzata coinvolge i quattro indici (i,n-1), (i-1,n), (i,n) e (i+1,n) secondo il seguente schema detto **BC**, dove B sta' per "Backward" (all'indietro)



La corrispondente relazione per l'errore è:

$$\begin{bmatrix} e_{1,n} \\ e_{2,n} \\ \vdots \\ \vdots \\ e_{m-1,n} \end{bmatrix} - \frac{ak}{h^2} \begin{vmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & -2 \end{vmatrix} \begin{bmatrix} e_{1,n} \\ e_{2,n} \\ \vdots \\ \vdots \\ e_{m-1,n} \end{bmatrix} = \begin{bmatrix} e_{1,n-1} \\ e_{2,n-1} \\ \vdots \\ \vdots \\ e_{m-1,n-1} \end{bmatrix} + kE_n$$

e quindi

$$U_n = (I-\lambda T)^{-1} U_{n-1} + k(I-\lambda T)^{-1} E_n$$

Per entrambi i metodi trattati, come per qualunque altro metodo alle differenze, l'errore soddisfa ad una relazione ricorsiva del tipo

$$BU_n = AU_{n-1} + kE_n$$

$$U_n = B^{-1}AU_{n-1} + kB^{-1}E_n$$

dove le matrici A e B, con $|B|\neq 0$, dipendono dai valori di discretizzazione k ed h, e dove la discretizzazione stessa e' supposta consistente col problema differenziale trattato, cioe'

$$E = \max_n ||E_n|| = O(k^p, h^q)$$

per opportuni valori di p e q (dipendenti dal tipo di differenza adottato).

Per quanto riguarda la **convergenza** di un siffatto schema per $h\to 0$, $k\to 0$, analizziamo l'errore e osserviamo che se $||B^{-1}A||\leq 1$ e $||B^{-1}||\leq M$ uniformemente rispetto alla dimensione di A (cioe' ad h) ed al valore di k, allora su ogni intervallo temporale finito [0,T] con T=Nk si ha

$$U_n=(B^{-1}A)^nU_0 + k((B^{-1}A)^{n-1}\, B^{-1}E_1 + (B^{-1}A)^{n-2}\, B^{-1}E_2+\ldots+B^{-1}E_n) \qquad \forall n\le N.$$

Poiche' l'errore iniziale U_0 e' nullo e vale l'ipotesi $||B^{-1}A||\le 1$,si ha

$$||U_n||\le k||B^{-1}||(||E_1 ||+ ||E_2||+\ldots+||E_n||)\le kMnE=TME = O(k^p ,h^q) \qquad \forall n\le N.$$

Avendo gia' verificato che entrambi gli schemi FC e BC sono consistenti (di ordine p=1 e q=2), per la convergenza ci rimane da verificare, per entrambi, le condizioni di **stabilita'** $||B^{-1}A||\le 1$ e $||B^{-1}||\le M$.

Per lo schema in FC si ha:

$$U_{n+1}=(I+\lambda T)U_n + kE_n \qquad \text{per } n=1,2,...N-1$$

quindi B=I e A= I+λT.

Si tratta dunque di verificare se $||(I+\lambda T)||\le 1$ in qualche norma.
Essendo A simmetrica, conviene usare la norma $||.||_2$,per la quale $\rho(A)=||A||_2$,e ricordare che gli autovalori di I+λT sono dati da $1+\lambda\sigma_i$ con

$$\sigma_i = \left(-2 + 2\cos\left(\frac{i}{m}\pi\right)\right) \qquad i = 1,...,m-1$$

autovalori di T.
Poichè gli autovalori di T sono negativi e quello di modulo massimo è maggiore di -4, la condizione $|1+\lambda\sigma_i|\le 1\ \forall i$ è soddisfatta se $1-4\lambda\ge -1$, cioè

$$\lambda=\frac{ak}{h^2}\le 1/2.$$

In conclusione lo schema FC è **condizionatamente stabile**, e quindi **condizionatamente convergente**, inquanto la convergenza e' garantita sotto la condizione che i passi h e k tendano a zero rispettando il vincolo $\frac{ak}{h^2}\le 1/2$. Nel caso di un passo spaziale h che sia già piccolo per esigenze di precisione, l'ampiezza del passo temporale k diventa così piccolo da rendere improponibile il metodo.

Per lo schema BC si ha invece

$$U_n=(I-\lambda T)^{-1}(U_{n-1}) + (I-\lambda T)^{-1}E'_n \qquad \text{per } n=1,2,...N$$

quindi $B=(I-\lambda T)^{-1}$ e $A=I$. In questo caso le condizioni per la convergenza si riducono a

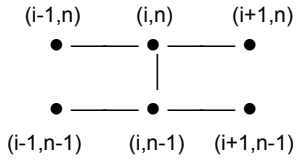
$$||(I-\lambda T)^{-1}||\leq 1 \qquad \text{uniformemente rispetto ad h e k.}$$

Gli autovalori di $I-\lambda T$ sono dati da $1/(1-\lambda\sigma_i)$. Poichè gli autovalori di T sono negativi, si ha $1/(1-\lambda\sigma_i) \leq 1$ per ogni λ.
Abbiamo quindi dimostrato che lo schema all'indietro è **incondizionatamente convergente**.

Si consideri infine il seguente schema di **Crank-Nicolson**, che consiste nell'approssimare la derivata seconda spaziale con la media di due differenze centrali ai tempi t_n e t_{n+1}. (Si lascia al lettore da verificare che tale schema fornisce un metodo consistente con ordine p=2, q=2)

$$\frac{y(t_n,x_i)-y(t_{n-1},x_i)}{k} = \frac{a}{2}\left(\frac{y(t_n,x_{i-1})-2y(t_n,x_i)+y(t_n,x_{i+1})}{h^2}+\frac{y(t_{n-1},x_{i-1})-2y(t_{n-1},x_i)+y(t_{n-1},x_{i+1})}{h^2}\right)+E_{n,i} \qquad i=1,...,m-1$$

Esso e' rappresentato dallo schema



La relazione ricorsiva per il vettore U_n degli errori assume la forma:

$$(I-\lambda T/2)^{-1}\, U_n = (I+\lambda T/2)\,(U_{n-1}) + kE_n$$

e quindi

$$U_n=(I-\lambda T/2)^{-1}(I+\lambda T/2)\,(U_{n-1}) + (I-\lambda T/2)^{-1}kE_n \qquad \text{per } n=1,2,...N.$$

Infine la condizione di stabilita'

$$||(I-\lambda T/2)^{-1}(I+\lambda T/2)||\leq 1$$

e' verificata **incondizionatamente** rispetto a k e ad h poiche' gli autovalori della matrice $(I-\lambda T/2)^{-1}(I+\lambda T/2)$ sono dati da $(1+½\lambda\sigma_i)/(1-½\lambda\sigma_i)$ che risultano ≤ 1 per ogni i.