EQUAZIONI DIFFERENZIALI 1

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

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

$$y'(t) = f(t,y(t)) \qquad (1.1)$$
$$y(t_0)=y_0$$

dove $y(t) : [t_0, t_f] \to R^m$ ed $f(t,y): [t_0, t_f]\times R^m \to R^m$. 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 (1.1) 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:

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

Nota : Si osservi che la Lipschitzianità della funzione $f(t,y)$ rispetto alla seconda variabile e' una condizione che garantisce la continuità ma non la derivabilità della f rispetto a tale variabile. In particolare essa afferma che il rapporto incrementale è limitato da L in ogni punto t e per ogni coppia di argomenti u e v. Dunque e' una condizione più forte della continuità ma più debole della derivabilità. Di conseguenza la funzione y'(t) risulta continua ma non derivabile. In seguito noi richiederemo delle condizioni più restrittive alla funzione f, precisamente la derivabilità fino ad un certo ordine p rispetto ad entrambi gli argomenti e quindi la derivabilità di ordine p+1 per la funzione incognita y(t). Ciò ci consentirà di ottenere metodi più veloci.

Per approssimare numericamente la soluzione del problema (1.1) fissiamo una discretizzazione dell'intervallo $[t_0, t_f]$ che, per semplicità di esposizione, supporremo uniforme:

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

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

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

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(t, y(t))dt. \qquad (1.2)$$

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 (1.1).

Metodi a un passo.

I metodi ad un passo sono quelli che, per approssimare l'integrale in (1.2), 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 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 di $y(t_n)$ 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:

$$y_{n+1}= y_n+h\Phi(t_n, y_n) \qquad (1.3)$$

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:

$$\alpha = y_n + hf(t_{n+1},\alpha) \qquad (1.4.a)$$

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

$$\beta = z_n + hf(t_{n+1},\beta) \qquad (1.4.b)$$

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 (1.4.a) ed (1.4.b). 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 (1.2) 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 $[t_n, t_{n+1}]$, quindi $0\leq c_i\leq 1$. 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

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

(1.5)

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:

$$\boxed{\begin{array}{lll} y_{n+1}= y_n+h\sum_{i=1}^{s} b_i\, f(\, t_n+c_ih,Y_i) & & (1.6) \\ Y_i = y_n + h\sum_{j=1}^{s} a_{i,j}\, f(\, t_n+c_jh\ ,\ Y_j) & i=1,...,s. & (1.7) \end{array}}$$

Le formule (1.6)-(1.7) sono note come **formule di Runge-Kutta ad** s **livelli** e sono rappresentate, in forma sintetica, attraverso la seguente tabella

$$\begin{array}{c|cccc} c_1 & a_{11} & \cdot & \cdot & \cdot & \cdot & a_{1s} \\ c_2 & \cdot & & & & & \\ \vdots & \cdot & & & & & \\ c_s & a_{s1} & \cdot & \cdot & \cdot & \cdot & a_{ss} \\ \hline & b_1 & b_2 & \cdot & \cdot & & b_s \end{array}$$

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 (1.5).

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 (1.5) 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 (1.7) si presenta come un problema di punto fisso Y=G(Y) 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:

$$Y^{k+1} = G(Y^k)$$

che, nel nostro caso, assume la forma:

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

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

Per i metodi Runge-Kutta (1.6)-(1.7), la funzione incrementale è del tipo:

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

dove gli Y_i dipendono anch'essi da u, e sono dati dalla soluzione di (1.7).
Si dimostra facilmente, come è stato fatto per il metodi 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 **<u>Eulero esplicito</u>** è un metodo di RK esplicito ad un livello con coefficienti:

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

Il metodo di **<u>Eulero implicito</u>** è un metodo RK implicito a un livello con coefficienti:

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

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

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}$$

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:*

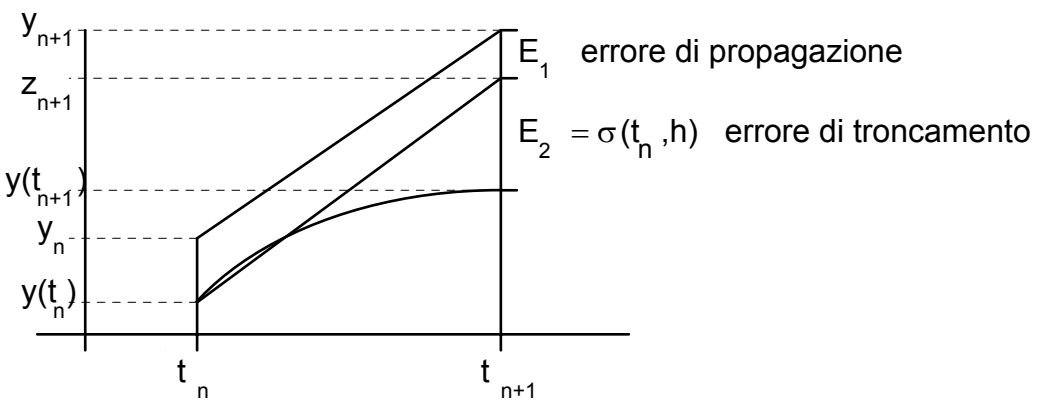
$$\sigma(t_n,h) := y(t_{n+1}) - \big(y(t_n) + h\Phi(t_n, y(t_n)) \big).$$

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

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

$$\| e_{n+1} \| \leq \| y_{n+1} - z_{n+1} \| + \| \sigma(t_n,h) \| = E_1 + E_2 \qquad (1.8)$$

Poichè

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

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

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 (1.8) 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:

$$\| e_{n+1} \| < (1 + hM) \| e_n \| + \sigma(h), \quad n = 0,1,\ldots,N-1. \qquad (1.9)$$

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:*

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

(La dimostrazione è lasciata come esercizio).

Applicando il lemma alla relazione ricorsiva (1.9), tenendo conto che $e_0=0$, si ottiene la maggiorazione uniforme per l'errore:

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

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

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

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

$$||e_m||< \sigma(h)\frac{e^{M(t_f -t_0)}-1}{hM} \qquad \forall m\leq N. \qquad (1.10)$$

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

Definizione: *Si dirà che il metodo* (1.3) *è* **convergente** *nell'intervallo d'integrazione* $[t_0, t_f]$, *se*

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

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 (1.10) 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:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ c & a & 0 \\ \hline & b_1 & b_2 \end{array} \qquad \text{con } c=a$$

e la formula sarà, in forma compatta:

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

L'errore locale di discretizzazione è dato da

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

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

$$\sigma(t_n,h) = hy'(t_n) + \tfrac{1}{2}h^2\big(f_t(t_n,y(t_n)) + f_y(t_n,y(t_n))y'(t_n)\big) + O(h^3) -$$

$$-h\Big[b_1 y'(t_n) + b_2\big(y'(t_n) + ahf_t(t_n,y(t_n)) + ahy'(t_n) f_y(t_n,y(t_n)) + O(h^2)\big)\Big]$$

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**

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

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

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

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 livelli 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 che, per il problema iniziale (1.1), 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 (1.10) ne rappresenta una limitazione uniforme su tutto l'intervallo $[t_0, t_f]$. Secondo la stima (1.10) 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 $E_1 := ||y_{n+1} - z_{n+1}||$ risulta non superiore all'errore accumulato fino al passo precedente, cioè se:

$$||y_{n+1} - z_{n+1}|| \leq ||e_n||, \qquad (1.11)$$

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

Per i metodi stabili la relazione (1.8) si può infatti sviluppare nel seguente modo:

$$\begin{aligned} || e_{n+1} || &\leq ||y_{n+1} - z_{n+1}|| + ||\sigma(t_n,h)|| \leq ||e_n|| + ||\sigma(t_n,h)|| \\ &\leq ||e_{n-1}|| + ||\sigma(t_{n-1},h)|| + ||\sigma(t_n,h)|| \leq ... \\ &\leq ||e_0|| + ||\sigma(t_0,h)|| + ||\sigma(t_1,h)|| + ... + ||\sigma(t_n,h)|| \\ &= ||\sigma(t_0,h)|| + ||\sigma(t_1,h)|| + ... + ||\sigma(t_n,h)||. \end{aligned}$$

e l'errore accumulato lungo i passi è dato (in realtà è maggiorato) dalla somma degli errori locali di troncamento.

Poichè, come abbiamo visto, $\sigma(t_k,h) \leq \sigma(h)$ 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 (1.10) 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 (1.11) è verificata, ad una in cui non lo è più. In altre parole siamo passati da una propagazione lineare ad una propagazione esponenziale dell'errore. Come vedremo tra poco, l'insorgenza del fenomeno di propagazione esponenziale dell'errore dipende sia dal problema trattato che dal metodo impiegato.
Naturalmente sarebbe preferibile utilizzare un metodo per il quale la condizione di stabilità (1.11) fosse verificata *per ogni scelta dal passo* che verra' quindi scelto solo in funzione dell'errore locale di troncamento.

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:

$$y'(t)=\lambda y(t)$$
$$y(0)=y_0 \qquad (1.12)$$

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 (1.12) rispetto alle variazioni sul dato iniziale, sia z(t) la soluzione di (1.12) 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

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

In questo caso diremo che (1.12) è un **problema stabile**.

Analizziamo ora la stabilità dei metodi numerici per il problema (1.12) nell'ipotesi che il problema stesso sia stabile, cioè che sia $\alpha=\mathrm{Re}(\lambda)<0$.

I 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:

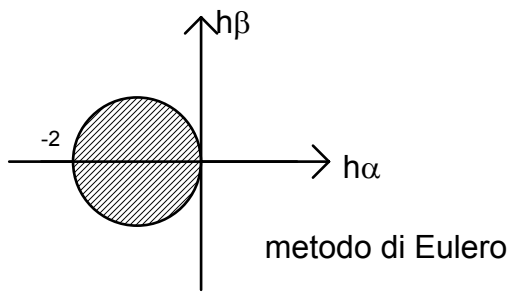
$$|\, y_{n+1} - z_{n+1}| = |\varphi(h\lambda)|\, |e_n|.$$

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

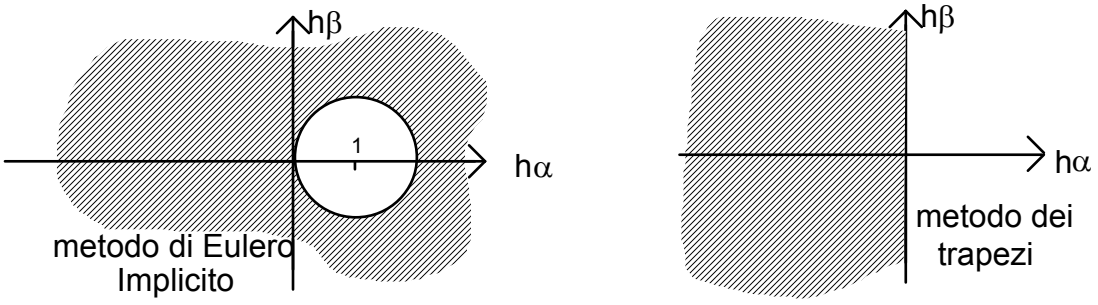
In maniera analoga si trovano le funzioni di stabilità:

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

e

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

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 (1.12) stabili e <u>*per ogni passo* h</u>.

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

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

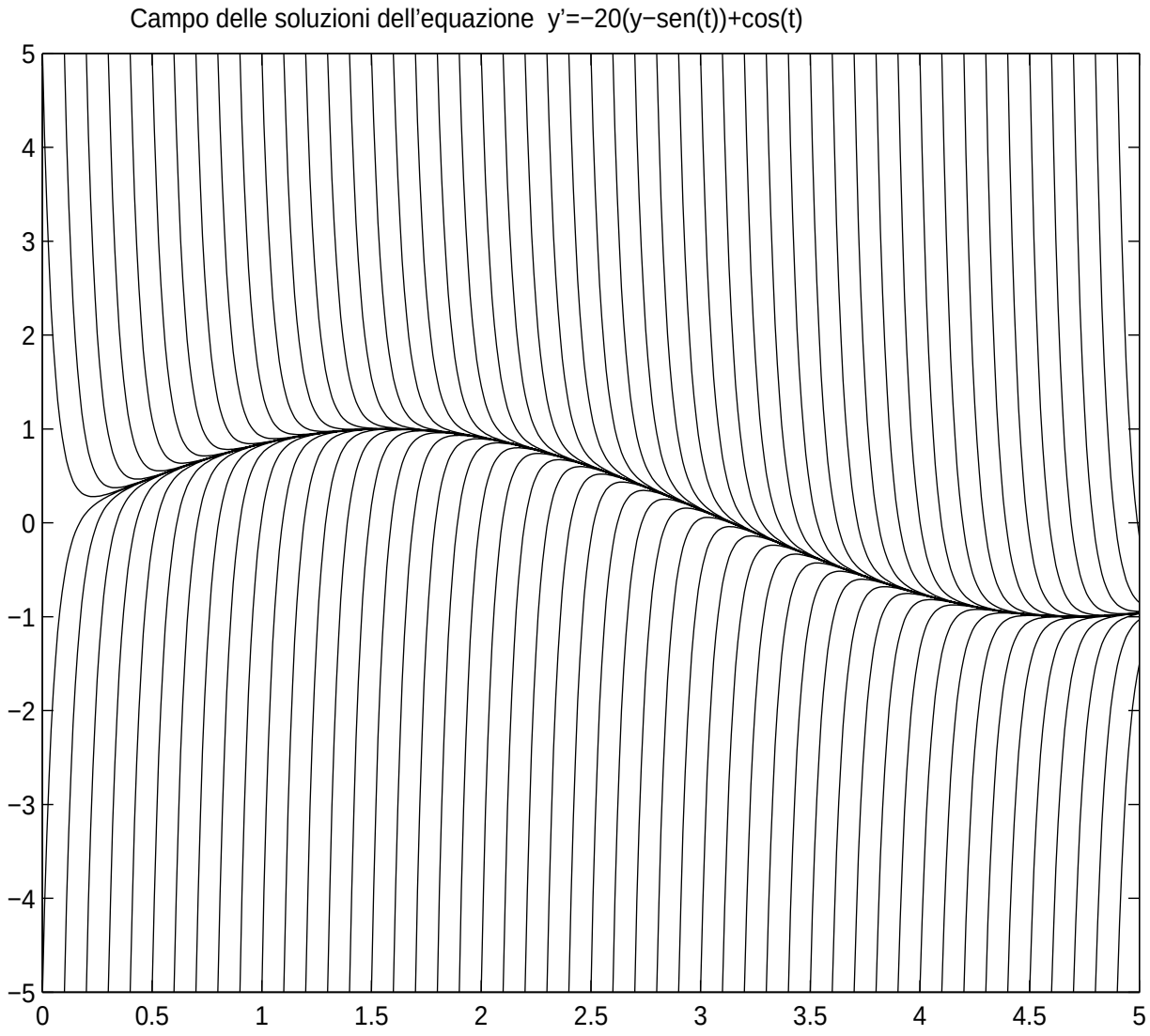
che a sua volta precipita sulla funzione F(t) ed è tale che la sua derivata in t_n si discosta sensibilmente da quella della soluzione esatta anche se siamo lontani da t_0. In altre parole, nella fase stazionaria le altre curve integrali sono sensibilmente diverse dalla soluzione esatta.

Nel seguente grafico si vede il campo delle soluzioni che escono da punti esterni alla soluzione esatta per il problema

$$y'=-20\,(y-\mathrm{sen}(t))+\cos(t)$$
$$y(0)=5$$

la cui soluzione e'

$$y(t) = 5\,e^{\lambda t} +\mathrm{sen}(t)$$

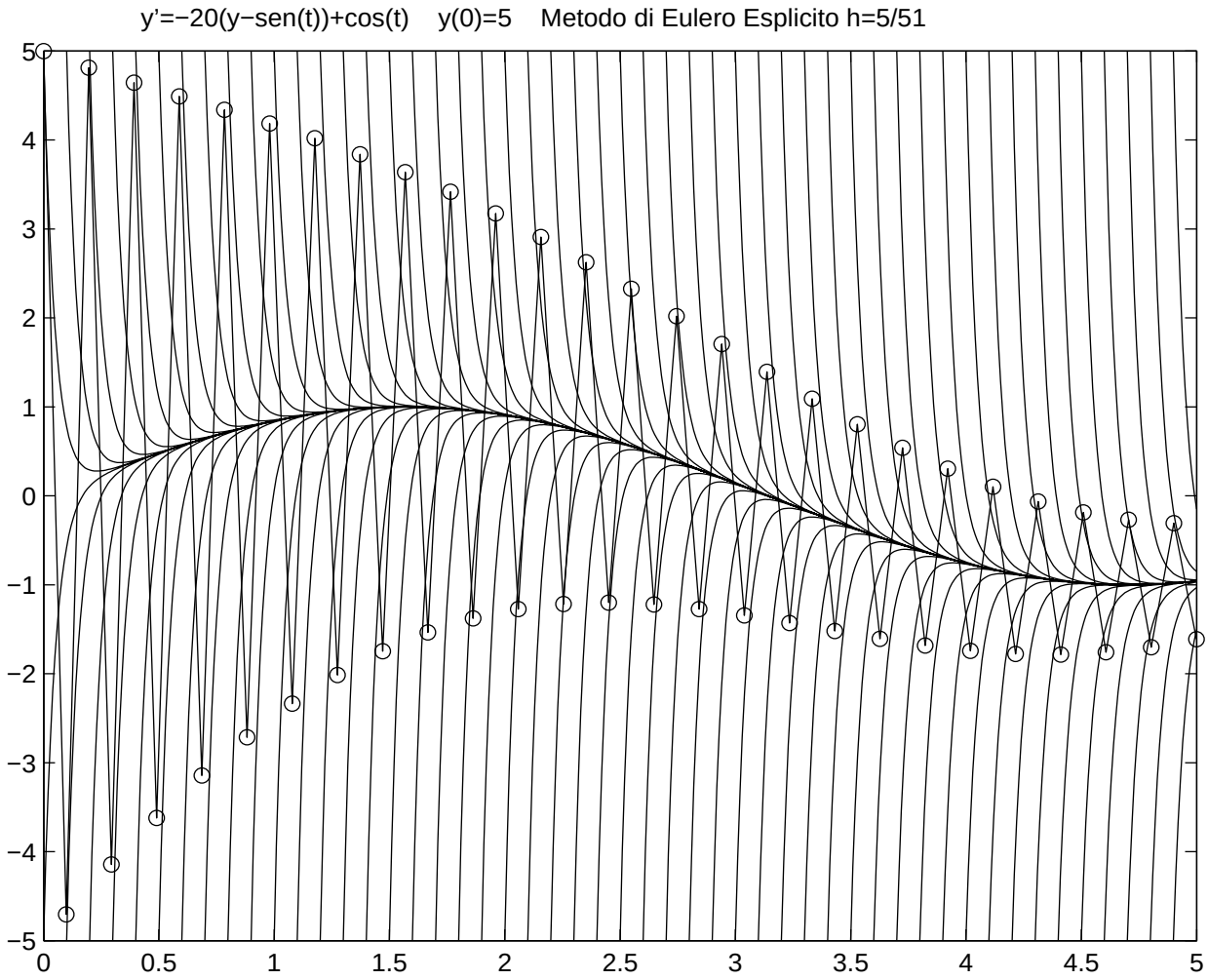


La figura successiva mostra l'approssimazione numerica ottenuta col metodo di Eulero Esplicito. Per il valore assegnato del passo d'integrazione h, l'errore di propagazione E_1 del metodo numerico e' molto grande rispetto all'errore locale di troncamento. Più precisamente, come per l'equazione test (1.12), l'errore propagato dal metodo è:

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

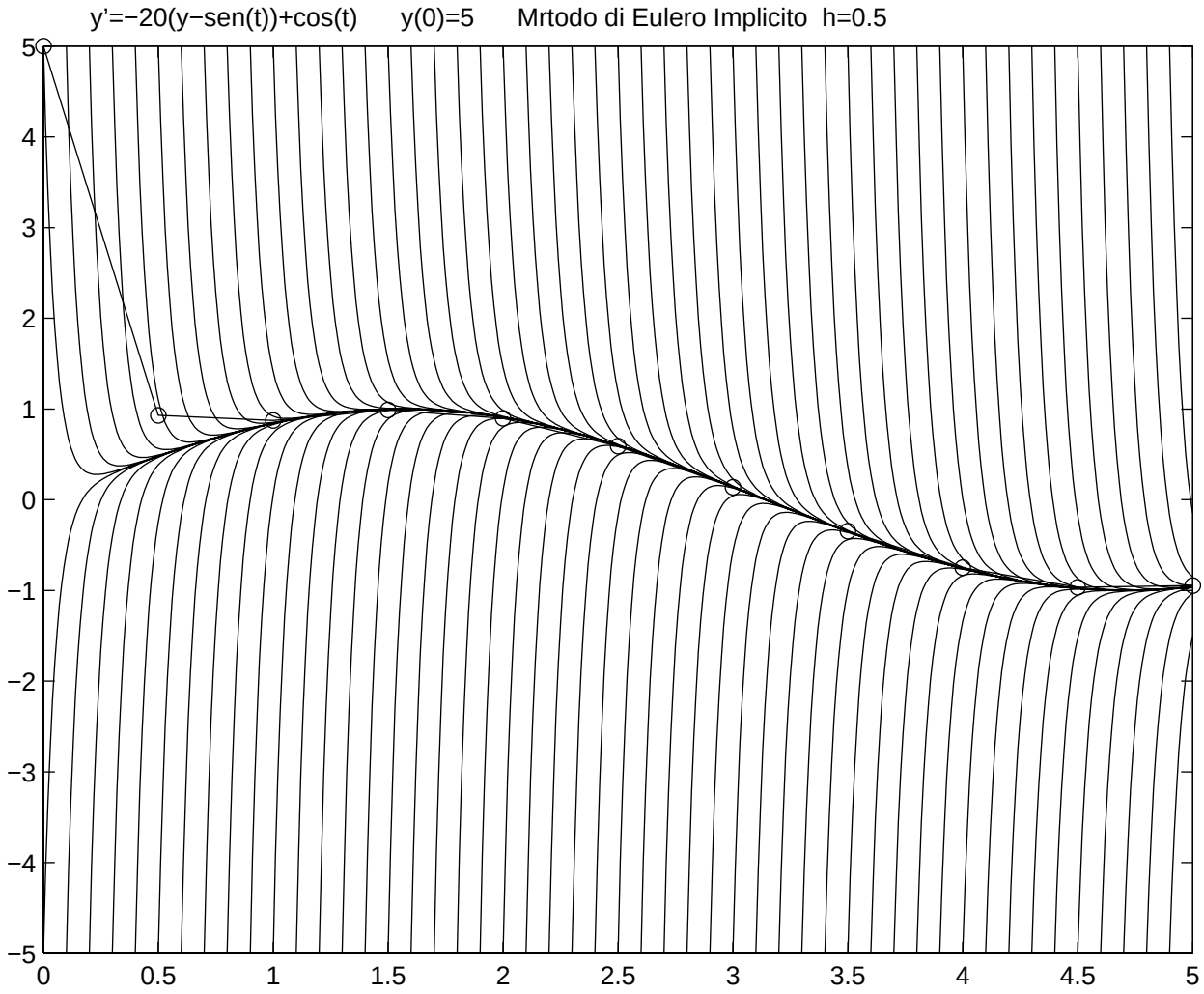
Poiche' il metodo ha una regione finita di assoluta stabilita',bisogna procedere con un passo tale che $|\varphi(h\lambda)|\leq 1$. Nel nostro caso $|1+20h|\leq 1$ e quindi $h\leq 0.1$.

Nella figura si osserva che con un passo di poco inferiore a 0.1 l'errore e' dovuto essenzialmente alla componente di propagazione anche nella fase stazionaria, dove invece l'errore di troncamento e' molto piccolo. Per un valore appena superiore ad 1 si avrebbe una crescita esponenziale dell'errore.



Viceversa, per il metodo di Eulero Implicito, che e' assolutamente stabile, la condizione $|\varphi(h\lambda)|\leq 1$ è verificata per ogni valore del passo.

Nella figura successiva e' illustrato il caso di un passo h=0.5 per il quale, nella fase stazionaria, l'errore propagato e' trascurabile e l'errore globale e' dovuto essenzialmente all'errore di troncamento .



Equazioni differenziali le cui soluzioni hanno un comportamento simile a questo sono dette **equazioni stiff** (rigide) . Nella fase transitoria esse potranno essere integrate indifferentemente con metodi espliciti o impliciti in quanto l'errore locale sara' dominato dall'errore di troncamento ed il passo sara' modulato su di esso. Invece nella fase stazionaria, se il metodo non e' stabile, l'errore locale sara' dominato dall'errore di propagazione che imporra' un passo forzatamente piccolo, molto piu' piccolo di quanto sarebbe richiesto dall'errore di troncamento. Un metodo stabile consentira', viceversa, di procedere con passi di integrazione piu' ampi vincolati essenzialmente dall'errore di troncamento.

Sistemi altamente oscillatori:

Abbiamo visto nel paragrafo precedente le difficolta' che insorgono ed i rimedi da adottare per l'integrazione di equazioni stiff caratterizzati da una fase transitoria, nella quale la soluzione ha una forte variazione ed una fase stazionaria nella quale la soluzione e' sostanzialmente liscia.

Consideriamo ora una classe di problemi dove, contrariamente alle equazioni stiff, la fase di alta oscillazione della soluzione e' permanente lungo l'intervallo di integrazione.

Si consideri per esempio l'equazione:

$$y''(t)+\lambda^2 y(t)=(\lambda^2-1)\,\mathrm{sen}(t) \qquad t\epsilon[0,T]$$

che possiede la seguente famiglia di soluzioni

$$y(t)=c\ \mathrm{sen}(\lambda t)+\mathrm{sen}(t).$$

Per c piccolo e λ sufficientemente grande, la soluzione e' costituita da un'onda ad alta frequenza, c sen(λt), modulata da un'onda lenta, sen(t).

La componente a forte variazione della soluzione permane lungo tutto l'intervallo di integrazione e quindi neanche l'impiego di un metodo "stiff" consente di svincolarsi dalla limitazione sul passo, imposto dall'errore di troncamento.

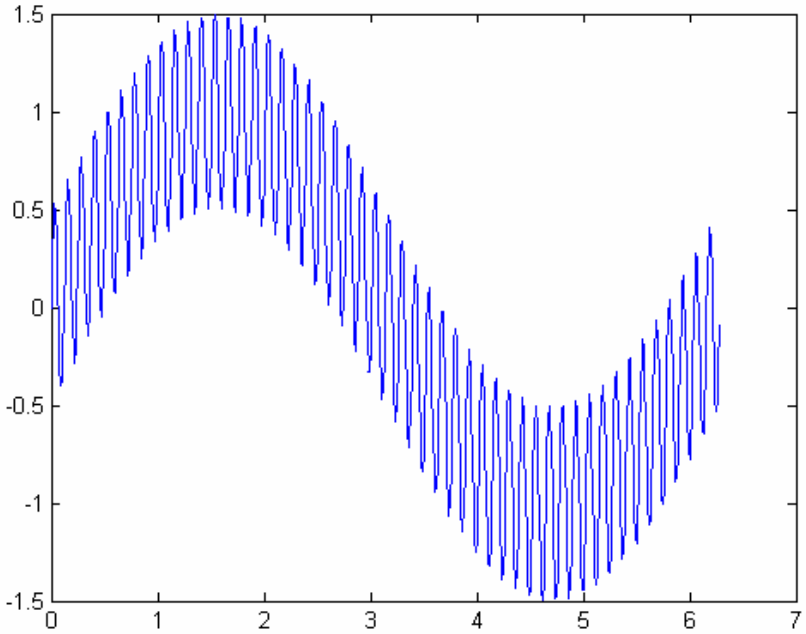


Grafico della funzione 0.5sen(50t)+sen(t)

Stabilità dei sistemi di equazioni differenziali:

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

$$y'(t) = Ay(t)$$
$$y(0)= u \qquad (1.13)$$

dove $A \in R^{m\times m}$ e $u=(1,1,...,1)^T \in R^m$. In questo caso ogni componente della soluzione si esprime come $p_i(t)e^{\lambda_i t}$ dove λ_i sono gli autovalori della matrice A e $p_i(t)$ e' un polinomio di grado m-1, se m e' la molteplicita' algebrica dell'autovalore λ_i. Il sistema risulta quindi stabile se e solo se $\mathrm{Re}(\lambda_i)\leq 0$ per ogni autovalore λ_i di A

Il metodo di Eulero esplicito applicato al sistema (1.13) 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 $||\varphi(hA)||\leq 1$. Poiche' il raggio spettrale di una matrice e' l'estremo inferiore delle norme della matrice stessa, ma potrebbe non coincidere con nessuna di esse, la condizione $||\varphi(hA)||\leq 1$ e' garantita dalla disuguaglianza stretta $\rho(\varphi(hA))<1$.
Ricordando ora che se λ è autovalore di A allora $\varphi(h\lambda)$ è autovalore di $\varphi(hA)$, è necessario che sia $|\varphi(h\lambda)|<1$ 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 (1.12) è sufficiente anche per il caso vettoriale. Infatti un metodo risulta stabile se $h\lambda$ è *interno* 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

$$h_{new} = h_{n+1}\cdot\max\{\tfrac{1}{2}, R\}$$

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

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

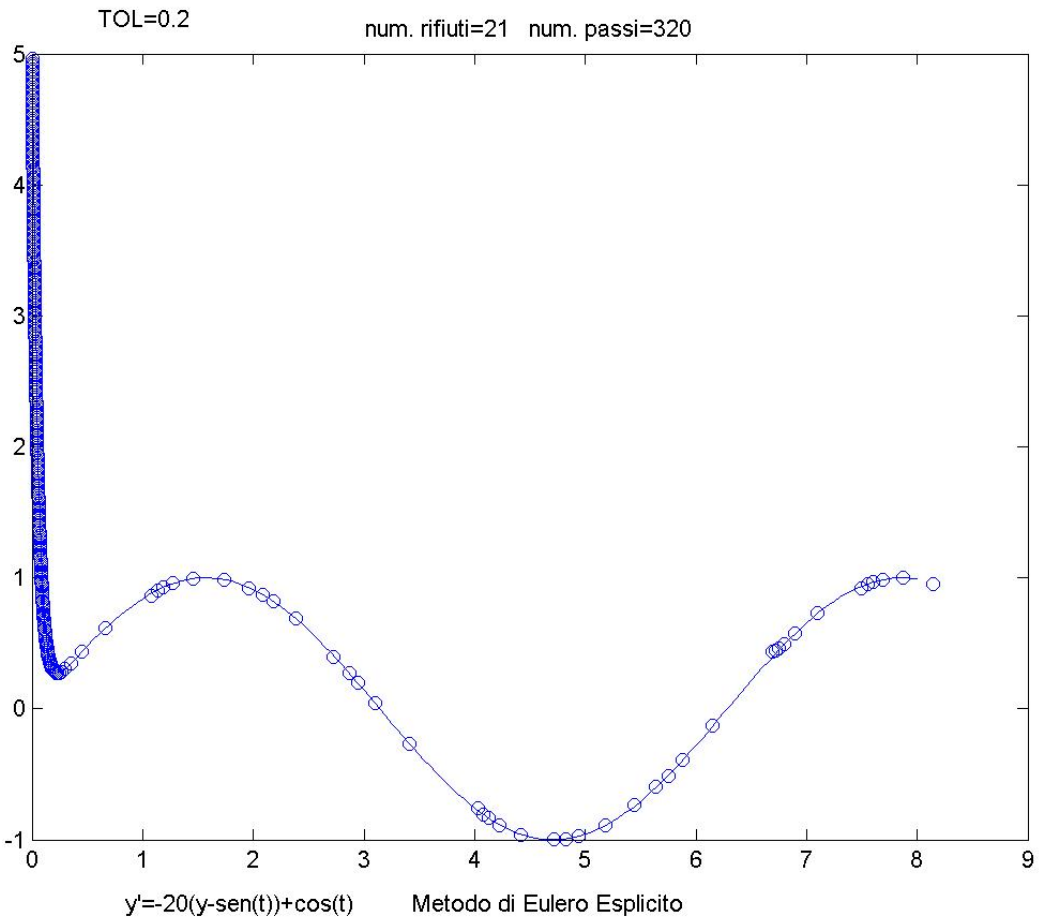
che ne limita l'allungamento.

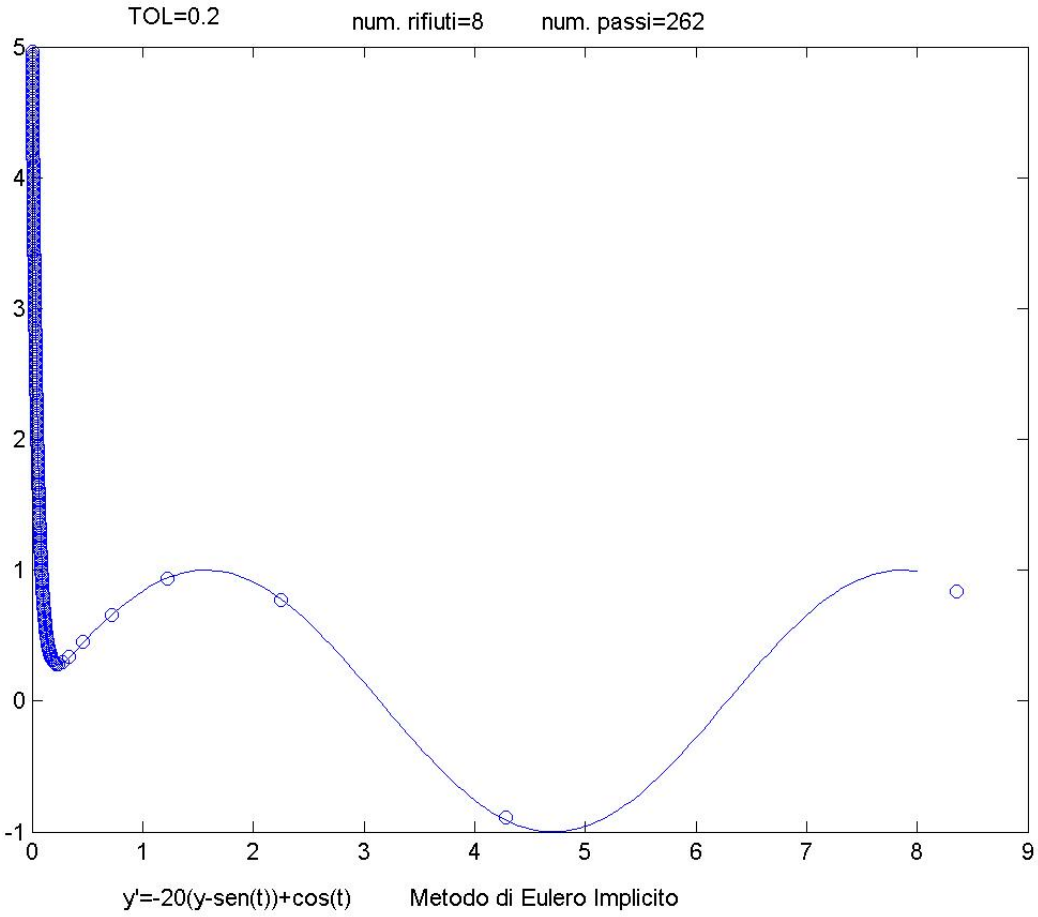
Con questo passo di tentativo, si applica la procedura descritta e si procede fino all'esaurimento dell'intervallo di integrazione.

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.





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

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ & & & \end{array}$$

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

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

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

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 (1.12), 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.

In particolare, consideriamo l'equazione test con $\lambda=i$ (unita' immaginaria)

$$y'=i\,y$$
$$y(0)=1$$

la cui soluzione

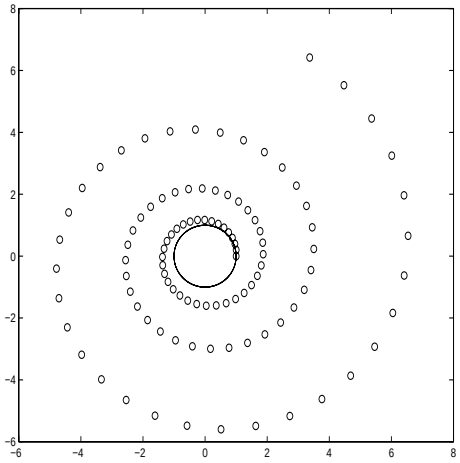
$$y(t)= \cos(t) + i\,\mathrm{sen}(t)$$

descrive un cerchio unitario $|y(t)|=1$ nel piano complesso C.

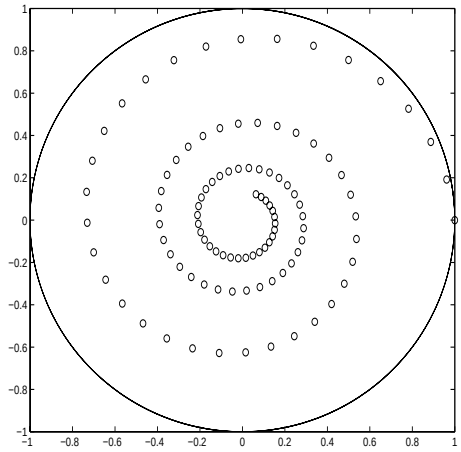
Per ogni metodo numerico, la soluzione e' data dalla sequenza di punti $y_{n+1}=\varphi(i\,h)y_n$. In figura sono riportate le traiettorie per i tre metodi di Eulero esplicito, implicito e dei trapezi.

Si osservi che, dei tre, solo il metodo dei trapezi e' capace di conservare il modulo unitario della soluzione numerica, e fornire quindi una traiettoria chiusa ***per ogni scelta del passo***

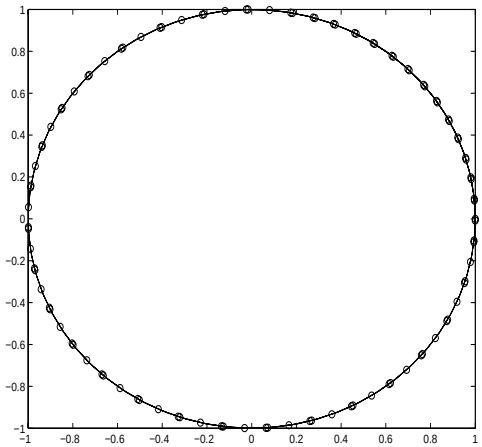
h. Al contrario, gli altri due metodi forniscono, per ogni passo, una spirale che implode o esplode. Perche?



Eulero Esplicito



Eulero Implicito



Trapezi

<u>Analisi asintotica dei sistemi:</u>

Analogamente al caso scalare, l'analisi asintotica è fatta sull'equazione test (1.13):

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

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 equazione test (1.13) 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 (metodo di semi-discretizzazione)

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 $x_i=ih$, $i=0,\ldots,m$ con passo $h=1/m$.

Per ogni t e per ogni x_i, approssimiamo la derivata seconda rispetto ad x con la differenza centrale seconda:

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

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

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

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:

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

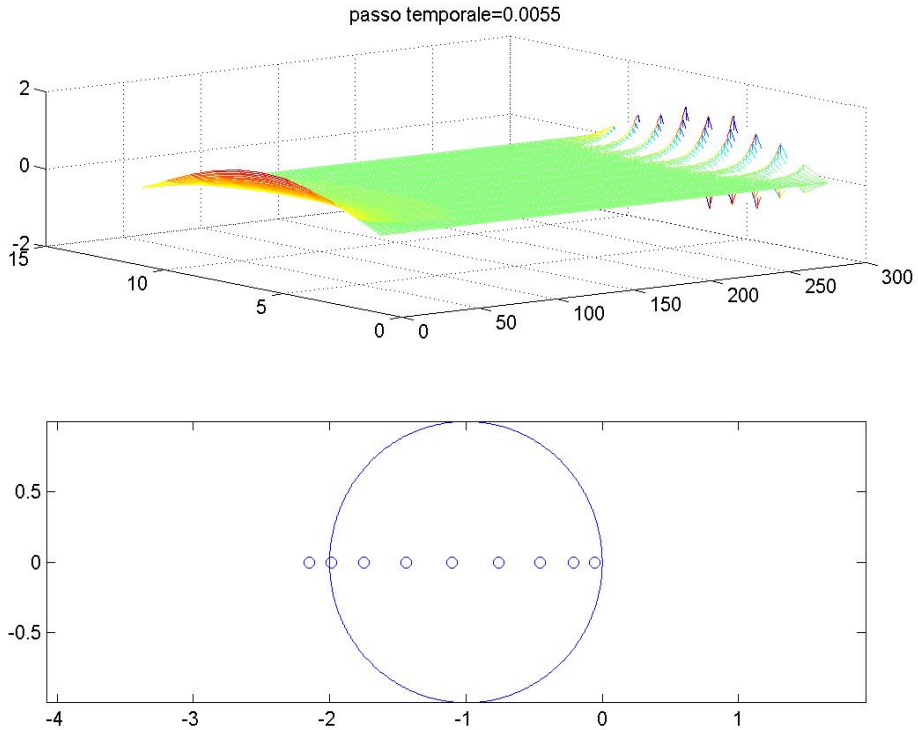
con le condizioni iniziali: $y_i(0) = y(0,x_i) = g(x_i) \qquad i = 1,\dots,m-1$.

Gli autovalori della matrice A del sistema sono:

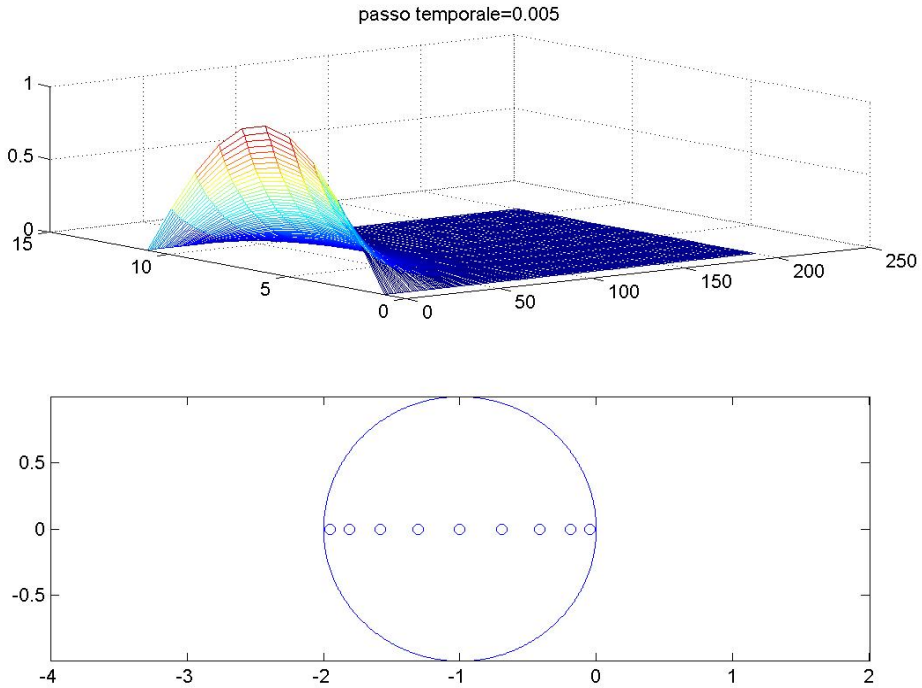
$$\lambda_i = m^2\left(-2 + 2\cos\left(\frac{i}{m}\pi\right)\right) \qquad i = 1,\dots,m-1$$

Essi sono compresi tra $\lambda_{m-1} \cong -4m^2$ e $\lambda_1 \le -\pi^2$ ed, essendo tutti negativi, il sistema è stabile e, per m grande, richiede l'impiego di un metodo assolutamente stabile anche in fase stazionaria. In particolare occorre che la condizione $|\varphi(h\lambda)|\le 1$ sia verificata per ogni λ autovalore di A.

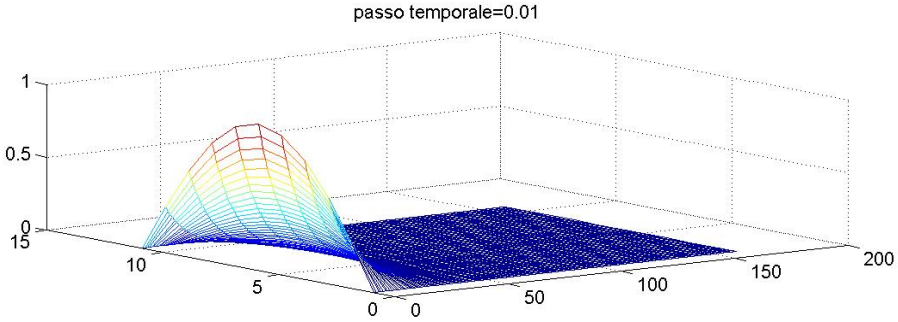
Volendo usare il metodo di Eulero esplicito ($\varphi(h\lambda)=1+h\lambda$) e supponendo di discretizzare l'intervallo spaziale in 10 parti (m=10) , il metodo risultera' stabile per quei valori del passo h tali che $-4\ h\ m^2 = -400\ h \ge -2$, quindi, $h \le 0.005$. Raffinando ulteriormente la discretizzazione spaziale, il passo temporale decresce drammaticamente poiche' l'autovalore di modulo massimo cresce col quadrato di m. Bisogna quindi usare un metodo assolutamente stabile.

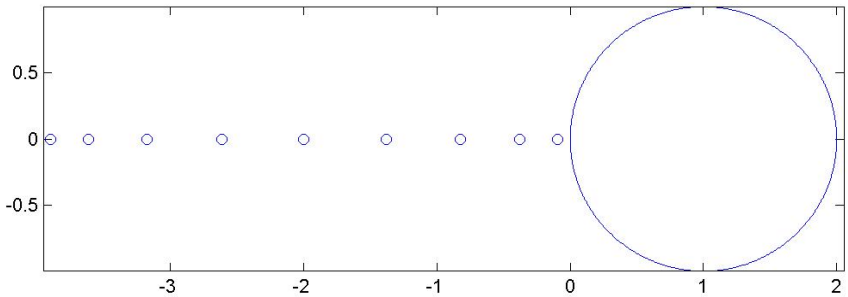


Metodo di Eulero esplicito con passo temporale tale che un autovalore esce dalla regione di stabilita'



Metodo di Eulero esplicito con passo temporale tale che tutti gli autovalori stanno nella regione di stabilita'





Metodo di Eulero implicito. Poiche' gli autovalori sono tutti negativi, il metodo e' stabile per ogni valore del passo h.