APPUNTI DI CALCOLO NUMERICO

Equazioni differenziali ordinarie

Introduzione

Moltissimi fenomeni reali possono essere modellizzati tramite equazioni differenziali ordinarie, ossia equazioni in cui sono correlate una funzione incognita e le sue derivate fino all'ordine m. Espressa in forma normale, una equazione differenziale ordinaria di ordine m si presenta in questo modo

$$y^{(m)}(t) = f\left(t, y(t), y'(t), \ldots, y^{(m-1)}(t)\right).$$

Si può dimostrare che ogni equazione di ordine m si può sempre scrivere come un sistema di m equazioni differenziali ordinarie di ordine 1. Un sistema di ordine uno generico è il seguente

$$\begin{cases} y_1'(t) = f_1(t, y_1(t), y_2(t), \ldots, y_m(t)) \\ y_2'(t) = f_2(t, y_1(t), y_2(t), \ldots, y_m(t)) \\ \ldots \\ y_m'(t) = f_m(t, y_1(t), y_2(t), \ldots, y_m(t)) \end{cases}$$

in cui si vede che tutte le equazioni dipendono da tutte le incognite.

Come passare da una equazione differenziale di ordine m ad un sistema di m equazioni di ordine uno? Semplicemente applicando una sostituzione standard

$$z_1(t) = y(t),\ z_2 = y'(t),\ \ldots,\ z_m(t) = y^{(m-1)}(t)$$

da cui si ottiene che l'equazione $y^{(m)}(t) = f\left(t, y(t), y'(t), \ldots, y^{(m-1)}(t)\right)$ è equivalente a

$$\begin{cases} z_1'(t) = z_2(t) \\ z_2'(t) = z_3(t) \\ \ldots \\ z_{m-1}'(t) = z_m(t) \\ z_m'(t) = f(t, z_1(t), z_2(t), \ldots, z_m(t)) \end{cases}.$$

Ai sistemi del primo ordine di equazioni differenziali ordinarie si applicano i metodi numerici per determinarne la soluzione.

Esempio:

Ricondurre a un sistema del primo ordine la seguente equazione differenziale ordinaria

$$4y^{(4)}(t)-3y''(t)+6t^2y'(t)-t^3=0\,.$$

Per prima cosa conviene scrivere l'equazione assegnata in forma normale

$$y^{(4)}(t)=\frac{3}{4}y''(t)-\frac{6}{4}t^2y'(t)+\frac{1}{4}t^3\,.$$

È ora possibile operare la sostituzione

$$z_1=y(t)$$
$$z_2=y'(t)$$
$$z_3=y''(t)$$
$$z_4=y'''(t)$$

e ottenere il sistema richiesto

$$\begin{cases} z_1'(t)=z_2(t) \\ z_2'(t)=z_3(t) \\ z_3'(t)=z_4(t) \\ z_4'(t)=\frac{3}{4}z_3(t)-\frac{6}{4}t^2z_2(t)+\frac{1}{4}t^3 \end{cases}$$

e si vede subito che è un sistema accoppiato.

La derivata prima di una funzione rappresenta geometricamente la tangente alla funzione data. In particolare rappresenta il modo in cui la funzione varia nel tempo: più la derivata prima è elevata, maggiore sarà la velocità di variazione della funzione. Sia che la funzione stia crescendo, sia che stia diminuendo. Quindi la derivata prima dà un'indicazione sul tasso di variazione della funzione.

Esempio: modello di Malthus per la dinamica di una popolazione isolata.

Una popolazione è isolata se non interagisce con altre popolazioni.

Sia $p(t)$ il numero di individui che compongono la popolazione isolata al tempo t. Se N è il tasso di natalità (dato dal numero di nati moltiplicato per l'unità di tempo moltiplicato ancora per l'unità di popolazione) e M è il tasso di natalità (dato dal numero di morti moltiplicato per l'unità di tempo moltiplicato ancora per l'unità di popolazione), allora

$$\frac{p(t+\Delta t)-p(t)}{\Delta t}=(N-M)p(t)=ap(t).$$

Per conoscere l'andamento istantaneo della popolazione si può far tendere a zero l'intervallo di tempo considerato:

$$\Delta t\to 0 \Rightarrow \lim_{\Delta t\to 0}\frac{p(t+\Delta t)-p(t)}{\Delta t}=p'(t)$$

quindi, detto I il contributo dovuto alle immigrazioni si ha

$$p'(t)=ap(t)+I.$$

Per $t\to\infty$ accade che

- se $M<N\Rightarrow a>0$, la popolazione tende a ∞;
- se $N<M\Rightarrow a<0$, la popolazione tende a 0.

Esempio: modelli di Lotka–Volterra per la dinamica di due popolazioni che si influenzano vicendevolmente.

$$\begin{cases} p_1'(t)=a_{11}p_1(t)+a_{12}p_1(t)p_2(t) \\ p_2'(t)=a_{21}p_1(t)p_2(t)+a_{22}p_2(t) \end{cases}$$

significato dei coefficienti:

- $a_{11}=N_1-M_1$ e $a_{22}=N_2-M_2$: differenza tra i tassi di natalità e di mortalità delle due popolazioni prese singolarmente (termini malthusiani);
- a_{12} e a_{21}: termini di interazione.

In base ai segni dei coefficienti è possibile avere tre differenti modelli di interazione:

1. <u>cooperazione</u>: $a_{11}<0$, $a_{22}<0$, $a_{12}>0$, $a_{21}>0$. Entrambe le popolazioni si estinguerebbero se fossero isolate l'una dall'altra. Se interagiscono, si influenzano positivamente ed entrambe possono accrescersi.
2. <u>competizione</u>: $a_{11}>0$, $a_{22}>0$, $a_{12}<0$, $a_{21}<0$. Entrambe le popolazioni possono espandersi se non avvengono interazioni fra di loro. Qualunque interazione fra le due popolazioni è negativa.
3. <u>preda–predatore</u>: sia p_1 la popolazione preda e sia p_2 la popolazione predatrice, allora $a_{11}>0$, $a_{22}<0$, $a_{12}<0$, $a_{21}>0$. La popolazione preda potrebbe esistere ed accrescersi senza alcun problema, se fosse isolata. Al contrario, la popolazione predatrice si estinguerebbe se non interagisse in alcun modo con la popolazione preda. Quindi l'interazione fra le due popolazioni è negativa per la popolazione preda, ma è positiva per quella predatrice.

Problema di Cauchy

Un problema è ben posto se viene assegnata una condizione iniziale. Infatti, per ogni equazione del tipo

$$y'(t) = y(t)$$

ha infinite soluzioni del tipo $y(t) = ce^t \quad \forall c \in \Re$.

Un Problema di Cauchy è composto da un'equazione differenziale ordinaria e da un dato iniziale che permette di localizzare una di queste infinite soluzioni:

$$\begin{cases} y'(t) = f(t, y(t)) \\ y(a) = y_0 \end{cases}$$

con $t \in [a,b]$, $f : \Re \times \Re^m \mapsto \Re^m$, $y_0 \in \Re^m$. Si possono trattare con il Problema di Cauchy sia le funzioni scalari, sia le funzioni vettoriali. Non vi è alcuna differenza nel procedimento.

Si può dimostrare l'esistenza e unicità della soluzione del Problema di Cauchy, sotto opportune ipotesi di regolarità. Il seguente teorema fornisce queste ipotesi di regolarità.

Teorema: sia $f : S = [\alpha, \beta] \times \Re^m \mapsto \Re^m$ continua nella striscia S. Se esiste una costante $L > 0$ tale per cui

$$\| f(t, y_1) - f(t, y_2) \| \le L \| y_1 - y_2 \| \quad \forall t \in [\alpha, \beta] \text{ e } \forall y_1, y_2 \in \Re^m$$

allora $\forall a \in [\alpha, \beta]$ e $\forall y_0$ esiste una e una sola soluzione del Problema di Cauchy su tutto $[\alpha, \beta]$.

La condizione si chiama condizione di Lipschitz rispetto alla seconda variabile.

La condizione è certamente soddisfatta se le derivate parziali della funzione $\frac{\partial f_i}{\partial y_j}$ esistono e sono continue e limitate in S.

Controesempio:

$$\begin{cases} y'(t) = \sqrt{|y(t)|} \\ y(0) = 0 \end{cases} \quad t \in [0, +\infty)$$

$$f(t, y(t)) = \sqrt{|y(t)|}$$

Calcolo delle derivate parziali

$$\frac{\partial f}{\partial y} = \begin{cases} \dfrac{1}{2\sqrt{y}} & y > 0 \\ -\dfrac{1}{2\sqrt{-y}} & y < 0 \end{cases}$$

$\dfrac{\partial f}{\partial y}$ è illimitata per $y = 0$.

Quindi la funzione non è lipschitziana in $y = 0$. Non è perciò garantita l'unicità della soluzione del Problema di Cauchy.

Infatti, in questo caso, esistono infinite soluzioni:

$$y(t) = \begin{cases} 0 & t \le c \\ \dfrac{1}{4}(t-c)^2 & t > c \end{cases}$$

al variare di c.

Metodi numerici

La soluzione di un'equazione differenziale è una funzione. Un metodo numerico fornisce una formula per calcolare un'approssimazione di questa funzione, ossia dà come risultati dei numeri.

Per prima cosa si introduce una griglia di punti sull'intervallo di integrazione $[\alpha, \beta]$. Questi punti sono detti nodi e sono così disposti:

$$a \equiv t_0 < t_1 < t_2 < \ldots < t_n \equiv b \,.$$

Per semplicità di trattazione, si supponga di considerare nodi equidistanti. In questo caso

$$h = \frac{b-a}{n} = t_{k+1} - t_k \quad \forall k \,.$$

Un metodo numerico permette di calcolare l'approssimazione $y_k \cong y(t_k)$ per $k = 0,1,\ldots,n$. In sostanza, un metodo numerico fornisce un numero, approssimazione della soluzione esatta (funzione) calcolata nel punto t_k.
Appare chiaro che non si potrà mai conoscere esattamente la funzione soluzione. Infatti, un calcolatore (a meno che non si utilizzi un software in grado di eseguire calcoli simbolici) non è in grado di eseguire un'operazione di integrazione in modo

esatto, ma solamente applicando un metodo numerico e fornendo un'approssimazione della soluzione.

Un metodo numerico si dice implicito se y_{k+1} è definita in modo implicito attraverso se stessa, cioè è funzione di se stessa. Tipicamente si ottengono metodi non lineari.
Un metodo numerico si dice esplicito se y_{k+1} è definita solamente in funzione di y_k.

Esempio: metodo di Eulero esplicito.

Sia $y'(t_k) = f(t_k, y(t_k))$. È possibile approssimare la derivata prima attraverso le differenze in avanti. Per $h > 0$

$$y'(t_k) \cong \frac{y(t_k + h) - y(t_k)}{h}.$$

Quindi, sostituendo

$$\frac{y(t_k + h) - y(t_k)}{h} \cong f(t_k, y(t_k)).$$

Se h rappresenta la distanza tra due nodi consecutivi, ossia vale $y(t_k + h) = y(t_{k+1})$, si ottiene

$$y(t_{k+1}) \cong y(t_k) + hf(t_k, y(t_k)).$$

La formula del metodo numerico si ottiene dalla precedente equazione, semplicemente sostituendone il circa–uguale con una uguaglianza e i termini $y(\cdot)$ con le loro approssimazioni

$$y_{k+1} = y_k + hf(t_k, y_k).$$

Quanto vale l'errore commesso con questo metodo?

Il metodo di Eulero esplicito è sintetizzato nella formula $y_{k+1} = y_k + hf(t_k, y_k)$ con $k = 0,1,...,n-1$. Lo sviluppo di Taylor arrestato al primo ordine della funzione $y(t)$, centrato in t_k e calcolato in t_{k+1}, con resto di Lagrange, è il seguente

$$y(t_{k+1}) = y(t_k) + y'(t_k)(t_{k+1} - t_k) + \frac{1}{2}(t_{k+1} - t_k)^2 y''(\xi_k)$$

e ricordando che $t_{k+1} - t_k = h$

$$y(t_{k+1}) = y(t_k) + y'(t_k)h + \frac{1}{2}h^2 y''(\xi_k).$$

Immaginando che non si commetta alcun errore al passo k, cioè che si abbia $y_k = y(t_k)$, e ricordando che $y'(t_k) = f(t_k, y(t_k))$, si sottrae l'espressione del metodo di Eulero esplicito dall'ultima espressione trovata. L'errore risulta

$$y(t_{k+1}) - y_{k+1} = y(t_k) - y_k + h\left(y'(t_k) - f(t_k, y_k)\right) + \frac{1}{2}h^2 y''(\xi_k) = \frac{1}{2}h^2 y''(\xi_k)$$

ossia si comporta come h^2 e si dice errore locale di troncamento.
Se si utilizza l'approssimazione della derivata prima mediante differenze all'indietro, si ottiene un altro metodo numerico che si chiama metodo di Eulero implicito, la cui formula è la seguente

$$y_{k+1} = y_k + hf(t_{k+1}, y_{k+1}).$$

Partendo dalla forma generale di una equazione differenziale ordinaria $y'(t) = f(t, y(t))$, è possibile trovare altri metodi numerici.

In particolare, se si approssima la derivata prima con le differenze centrate si trova

$$y'(t) \cong \frac{y(t+\Delta t) - y(t-\Delta t)}{2\Delta t}.$$

Se si chiamano $t_{k+1} = t + \Delta t$ e $t_k = t - \Delta t$, e di conseguenza $t_{k+1} - t_k = h = 2\Delta t$, si nota che il punto t non appartiene alla griglia dei nodi. Per convenzione tale punto viene chiamato $t_{k+\frac{1}{2}}$ e si ha che

$$y\left(t_{k+\frac{1}{2}}\right) \cong \frac{y(t_{k+1}) - y(t_k)}{h}.$$

Allora

$$\begin{aligned} \frac{y(t_{k+1}) - y(t_k)}{h} &\cong f\left(t_{k+\frac{1}{2}}, y\left(t_{k+\frac{1}{2}}\right)\right) \\ y(t_{k+1}) &\cong y(t_k) + hf\left(t_{k+\frac{1}{2}}, y\left(t_{k+\frac{1}{2}}\right)\right) \end{aligned}.$$

Il punto $t_{k+\frac{1}{2}}$, non essendo un nodo, non è "interessante". A partire dall'ultima relazione scritta, è possibile ottenere tre differenti metodi numerici, a seconda di come si sostituisca il termine $y\left(t_{k+\frac{1}{2}}\right)$.

Primo metodo: il termine $y\left(t_{k+\frac{1}{2}}\right)$ viene sostituito dal valore medio fra $y(t_k)$ e $y(t_{k+1})$

$$f\left(t_{k+\frac{1}{2}}, y\left(t_{k+\frac{1}{2}}\right)\right) \cong \frac{1}{2}\left(f(t_k, y(t_k)) + f(t_{k+1}, y(t_{k+1}))\right)$$

da cui si ottiene il metodo dei trapezi

$$y_{k+1} = y_k + \frac{h}{2}\left(f(t_k, y_k) + f(t_{k+1}, y_{k+1})\right)$$

che è strettamente legato alla formula dei trapezi per la risoluzione numerica di integrali definiti.

Dimostrazione: una equazione differenziale ordinaria generica è la seguente

$$y'(t) = f(t, y(t)).$$

Tale equazione è espressa nella forma differenziale. È possibile esprimerla anche in forma integrale, semplicemente integrandola fra t_0 e t

$$\int_{t_0}^{t} y'(t)dt = \int_{t_0}^{t} f(t, y(t))dt$$

$$y(t) - y(t_0) = \int_{t_0}^{t} f(t, y(t))dt$$

$$y(t) = y(t_0) + \int_{t_0}^{t} f(t, y(t))dt$$

e, analogamente, considerando due istanti generici t_k e t_{k+1}

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(t, y(t))dt.$$

Utilizzando la regola dei trapezi per approssimare l'integrale si ottiene

$$y(t_{k+1}) \cong y(t_k) + \frac{t_{k+1} - t_k}{2}\left(f(t_k, y(t_k)) + f(t_{k+1}, y(t_{k+1}))\right)$$

da cui si ricava immediatamente la formula del metodo dei trapezi

$$y_{k+1} = y_k + \frac{h}{2}\left(f(t_k, y_k) + f(t_{k+1}, y_{k+1})\right).$$

c.v.d.

Il metodo dei trapezi è un metodo implicito.

La generalizzazione di questo metodo si chiama $\vartheta-$metodo. In questo caso il termine in $t_{k+\frac{1}{2}}$ viene sostituito da una combinazione lineare convessa di un termine in t_k e un termine in t_{k+1}. Combinazione lineare convessa di due termini significa che la loro somma è sempre pari a 1. La formula generale per il $\vartheta-$metodo è la seguente

$$y_{k+1} = y_k + h\left(\vartheta f(t_k, y_k) + (1-\vartheta) f(t_{k+1}, y_{k+1})\right)$$

con $\vartheta \in [0,1]$.

Si può dimostrare che se $\vartheta = \frac{1}{2}$ si ottiene la formula del metodo dei trapezi; che se $\vartheta = 0$ si ottiene la formula del metodo di Eulero implicito e che se $\vartheta = 1$ si ottiene la formula del metodo di Eulero esplicito.

Secondo metodo: è possibile rendere esplicito il metodo dei trapezi approssimando il termine y_{k+1} indicato con la freccia nella seguente formula

$$y_{k+1} = y_k + \frac{h}{2}\left(f(t_k, y_k) + f\left(t_{k+1}, \underset{\Uparrow}{y_{k+1}}\right)\right).$$

Se tale termine viene approssimato mediante l'uso del metodo di Eulero esplicito

$$y_{k+1} = y_k + hf(t_k, y_k)$$

si ottiene il metodo di Heun

$$y_{k+1} = y_k + \frac{h}{2}\left(f(t_k, y_k) + f(t_{k+1}, y_k + hf(t_k, y_k))\right)$$

che è chiaramente esplicito.

Terzo metodo: analogamente a quanto fatto per determinare il secondo metodo, si può approssimare $y_{k+\frac{1}{2}}$ con un passo del metodo di Eulero esplicito lungo $\frac{h}{2}$

$$y_{k+1} = y_k + hf\left(t_{k+\frac{1}{2}}, y_{k+\frac{1}{2}}\right)$$
$$y_{k+\frac{1}{2}} = y_k + \frac{h}{2} f(t_k, y_k)$$

e, sostituendo, si ottiene il seguente metodo numerico, detto metodo di Eulero modificato

$$y_{k+1} = y_k + hf\left(t_{k+\frac{1}{2}}, y_k + \frac{h}{2} f(t_k, y_k)\right).$$

Tutti i metodi finora descritti appartengono ad un'unica famiglia con caratteristiche comuni definita come famiglia dei metodi one–step. I metodi appartenenti a questa famiglia possono essere espliciti, allora hanno la seguente forma generale

$$y_{k+1} = y_k + h\Phi(t_k, y_k; h);$$

oppure possono essere impliciti e allora hanno la seguente forma generale

$$y_{k+1} = y_k + h\Phi(t_k, y_k, y_{k+1}; h).$$

La funzione Φ si dice funzione generatrice del metodo.

Contrapposti ai metodi one–step esistono i metodi multi–step che necessitano delle informazioni su p nodi, precedenti al $k-$esimo che si sta considerando, per calcolare y_{k+1}.

Riassumendo, i metodi finora trattati sono i seguenti

- metodo di Eulero esplicito;
- metodo di Eulero implicito;
- metodo dei trapezi;
- metodo di Heun;
- metodo di Eulero modificato.

Metodi di Runge–Kutta espliciti

I metodi di Eulero esplicito, di Heun e di Eulero modificato rientrano nella categoria dei metodi di Runge–Kutta espliciti, in particolare essi sono dei casi particolari dei metodi espliciti di Runge–Kutta.

I metodi di Runge–Kutta espliciti hanno la seguente forma

$$y_{k+1} = y_k + h\sum_{i=1}^{s} a_i \kappa_i$$

dove $a_i \in \Re^s$ è un parametro che dipende dal metodo. Scrivendo in maniera esplicita il parametro κ_i

$$\kappa_i = f\left(t_k + b_i h, y_k + \sum_{j=1}^{i-1} c_{ij} \kappa_j \right) \quad i = 2,3,\ldots,s$$

si evidenziano altri due parametri $b \in \Re^s$ e $c \in \Re^{s\times s}$, anch'essi dipendenti dal metodo. Si nota anche che l'indice j è sempre strettamente minore dell'indice i.

Per convenzione si pone $\kappa_1 = f(t_k, y_k)$.

Il parametro s si dice numero di stadi ed è esattamente uguale al numero di valutazioni di funzioni richieste dal metodo.

Riunendo i passaggi precedenti in un'unica equazione si ha

$$y_{k+1} = y_k + h\sum_{i=1}^{s} a_i f\left(t_k + b_i h, y_k + h\sum_{j=1}^{i-1} c_{ij} \kappa_j \right).$$

Casi particolari:

$s=1$: $y_{k+1} = y_k + h a_1 \underbrace{f(t_k + b_1 h, y_k)}_{\kappa_1}$.

$s=2$: $y_{k+1} = y_k + h\left(a_1 \underbrace{f(t_k + b_1 h, y_k)}_{\kappa_1} + a_2 \underbrace{f\left(t_k + b_2 h, y_k + c_{21} \underbrace{f(t_k + b_1 h, y_k)}_{\kappa_1} \right)}_{\kappa_2} \right)$.

Esempio: Eulero esplicito.

$$y_{k+1} = y_k + hf(t_k, y_k)$$

non è altro che un metodo di Runge–Kutta ad uno stadio in cui si pone

$$\begin{cases} s = 1 \\ a_1 = 1 \\ b_1 = 0 \end{cases}.$$

Esempio: Heun.

$$y_{k+1} = y_k + \frac{h}{2}\left(f(t_k, y_k) + f(t_{k+1}, y_k + hf(t_k, y_k))\right)$$

non è altro che un metodo di Runge–Kutta a due stadi in cui si pone

$$\begin{cases} s = 2 \\ a_1 = \frac{1}{2} \\ a_2 = \frac{1}{2} \\ b_1 = 0 \\ b_2 = 1 \\ c_{21} = 1 \end{cases}.$$

Esempio: Eulero modificato.

$$y_{k+1} = y_k + hf\left(t_{k+\frac{1}{2}}, y_k + \frac{h}{2} f(t_k, y_k)\right)$$

non è altro che un metodo di Runge–Kutta a due stadi in cui si pone

$$\begin{cases} s = 2 \\ a_1 = 0 \\ a_2 = 1 \\ b_1 = 0 \\ b_2 = \frac{1}{2} \\ c_{21} = \frac{1}{2} \end{cases}.$$

Tutti i parametri dei metodi espliciti di Runge–Kutta, ossia i parametri a, b e c, sono raggruppati nella seguente tabella, detta Tableau di Butcher, per una più rapida memorizzazione

$$\begin{array}{c|c} b \in \Re^s & c \in \Re^{s\times s} \\ \hline & a^T \in \Re^s \end{array} \Rightarrow \begin{array}{c|ccc} b_1 & & & \\ b_2 & c_{21} & & \\ \vdots & \vdots & & \\ b_s & c_{s1} & \cdots & c_{s,s-1} \\ \hline & a_1 & \cdots & a_s \end{array} .$$

I parametri soddisfano le seguenti relazioni: $\sum_{i=1}^{s} a_1 = 1$ e $b_i = \sum_{j=1}^{s} c_{ij} \quad \forall i = 1,2,\ldots,s$. Un metodo, i cui parametri soddisfino tali condizioni, è detto un <u>buon metodo</u> di Runge–Kutta.

Esempio: Tableau di Butcher.

- Eulero esplicito (Runge–Kutta ad uno stadio)

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

- Eulero modificato (Runge–Kutta a due stadi)

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

- Heun (Runge–Kutta a due stadi)

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

- Runge–Kutta a più stadi: $s = 3$

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ \hline & 1/4 & 0 & 3/4 \end{array}$$

- Runge–Kutta a più stadi: $s = 4$

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

Comportamento dei metodi numerici

In pratica, si parla di convergenza e stabilità dei metodi numerici analizzati.

Definizione di convergenza per un metodo uniforme: un metodo numerico è detto convergente in un punto $x \in [a,b]$ se, data una suddivisione di $[a,x]$ in N intervalli di ampiezza $h = \frac{x-a}{N}$, si ha

$$\lim_{N\to\infty} y_N = y(x).$$

Il metodo si dirà convergente in $[a,b]$ se è convergente $\forall x \in [a,b]$.

Tutti i metodi analizzati precedentemente risultano convergenti. Questo accade perché sono tutti metodi one–step consistenti.

Per essere convergente, un metodo deve essere consistente e deve presentare $0-$stabilità. Unicamente per i metodi one–step la consistenza e la convergenza sono la stessa cosa, in quanto la $0-$stabilità è sempre garantita.

Occorre ora definire gli errori numerici di approssimazione.

Si consideri il generico metodo one–step $y_{k+1} = y_k + h\Phi(t_k, y_k, y_{k+1}; h)$. Sia $\bar{t} \in [a,b]$ un punto fissato e considerando il seguente Problema di Cauchy

$$\begin{cases} y'(t) = f(t, y(t)) \\ y(\bar{t}) = \bar{y} \end{cases} \quad t \in [\bar{t}, \bar{t}+h]$$

è possibile dare la seguente

Definizione di errore locale di troncamento (o di discretizzazione):

$$\tau(\bar{t}, \bar{y}; h) = \frac{y(\bar{t}+h) - y(\bar{t})}{h} - \Phi(\bar{t}, \bar{y}, y(\bar{t}+h); h).$$

L'espressione "errore locale" indica l'errore che si commette in un solo passo del metodo numerico, supponendo che tale metodo commetta errore nullo nel primo estremo dell'intervallo di variazione di t.

Dimostrazione:

Si riscrive l'equazione del generico metodo numerico in questo modo

$$\frac{y_{k+1}-y_k}{h}=\Phi(t_k,y_k,y_{k+1};h).$$

A questo punto si sostituisce nella formula la soluzione esatta del Problema di Cauchy e l'errore di troncamento è il residuo prodotto

$$\begin{aligned} y_k &\to y(\bar{t}) \\ y_{k+1} &\to y(\bar{t}+h). \\ t_k &\to \bar{t} \end{aligned}$$

In pratica, l'uguaglianza non sarà verificata e la differenza fra i due membri è esattamente l'errore di troncamento τ che rappresenta la bontà dello schema numerico nell'approssimare (in un solo passo) l'equazione differenziale ordinaria.

c.v.d.

Definizione di <u>consistenza</u>: un metodo numerico si dice <u>consistente</u> se $\forall \bar{t} \in [a,b]$ e $\forall \bar{y} \in \Re^m$ si verifica che

$$\lim_{h\to 0}\tau(\bar{t},\bar{y};h)=0.$$

La consistenza è la proprietà di un metodo numerico di convergere alla soluzione cercata, in caso di convergenza del metodo stesso.

Definizione di <u>ordine</u> di un metodo: si dice che un metodo ha <u>ordine di consistenza</u> p se

$$\tau(\bar{t},\bar{y};h)=O(h^p).$$

Teorema: per i metodi numerici <u>one–step</u> la consistenza equivale alla convergenza. Per i metodi multi–step la convergenza è assicurata se è garantita la consistenza e la $0-$stabilità.

Per i metodi espliciti di Runge–Kutta i parametri devono soddisfare la seguente condizione

$$\sum_{i=1}^{s} a_i = 1.$$

Tale condizione è la proprietà che garantisce la consistenza dei metodi espliciti di Runge–Kutta.

Esempio: ordine di alcuni metodi.

- Eulero implicito: $p=1$
- Eulero esplicito: $p=1$
- Trapezi: $p=2$
- Heun: $p=2$
- Runge–Kutta: variabile, ma comunque $p \le s$
  - $p^{\max} = s$ se $s = 1,2,3,4$
  - $p^{\max} = s-1$ se $s = 5,6,7$
  - $p^{\max} = s-2$ se $s = 8,9$
  - $p^{\max} \le s-3$ se $s \ge 10$

Esempio:

$$\begin{cases} y' = y \\ y(0) = 1 \end{cases} \quad t \in (0,T]$$

Soluzione esatta: $y(t) = e^t$.

Confronto fra i metodi di Eulero esplicito (EE), di Heun (H) e di Runge–Kutta di ordine 4 ($RK4$): errori calcolati nell'ultimo nodo della griglia (nodo N) i quali, per la definizione di convergenza, dovrebbero tendere a zero.

h	$\lvert y_N{}^{EE} - y(x_N)\rvert$	$\lvert y_N{}^{H} - y(x_N)\rvert$	$\lvert y_N{}^{RK4} - y(x_N)\rvert$
$1/2$	$4.6828e-01$	$7.7657e-02$	$9.3564e-04$
$1/4$	$2.7688e-01$	$2.3426e-02$	$7.1889e-05$
$1/8$	$1.5250e-01$	$6.4406e-03$	$4.9840e-06$
$1/16$	$8.0353e-02$	$1.6883e-03$	$3.2812e-07$
$1/32$	$4.1292e-02$	$4.3215e-04$	$2.1048e-08$
$1/64$	$2.0937e-02$	$1.0932e-04$	$1.3327e-09$

Si deduce che, se un metodo ha ordine p, $\|y_N - y(x_N)\| = O(h^p)$ e allora

$$h \to \text{errore} \sim O(h^p)$$

$$\frac{h}{2} \to \text{errore} \sim O\left(\left(\frac{h}{2}\right)^p\right) = \frac{1}{2^p} O(h^p)$$

cioè dimezzando il passo, l'errore diminuisce come h^p diviso per 2^p. Quindi, più p è elevato, maggiore sarà la velocità di diminuzione dell'errore.

Stabilità

La convergenza garantisce che l'errore è piccolo se $h \to 0$. Tuttavia, nella pratica, h è di solito ben diverso da zero. Tutti i valori di h sono accettabili?

Esempio:

$\begin{cases} y' = -10y \\ y(0) = 1 \end{cases}$ $t \in (0, T]$ (problema autonomo, ossia non ha dipendenza esplicita da t).

Soluzione esatta: $y(t) = e^{-10t}$.

Eulero esplicito: $y_{k+1} = y_k + h(-10y_k) = (1-10h)y_k$.

- Con $h = 0.15$ la soluzione numerica converge alla soluzione esatta.
- Con $h = 0.3$ la soluzione numerica diverge completamente dalla soluzione esatta.
- Con $h = 0.5$ la soluzione numerica diverge ancora più vistosamente dalla soluzione esatta, con errori superiori di almeno due ordini di grandezza agli errori del caso precedente.

Generalizzando: dato il problema test

$$\begin{cases} y' = \lambda y \\ y(t_0) = y_0 \end{cases} \quad \Re e\, \lambda < 0$$

con soluzione esatta $y(t) = y_0 e^{\lambda(t-t_0)}$ si dice che il problema è asintoticamente stabile se

$$\lim_{t \to \infty} y(t) = 0$$

e il metodo numerico si dice assolutamente stabile se, per un h fissato,

$$\lim_{k \to \infty} y_k = 0 .$$

Come stabilire per quali valori di h un metodo numerico risulta stabile per un dato problema?

Applicando un metodo one–step al problema test con $f(t,h) = \lambda y$ ci si riconduce sempre ad un'espressione del tipo

$$y_{k+1} = F(h\lambda)y_k .$$

Esempi:

- Eulero esplicito

$y_{k+1} = y_k + h\lambda y_k = (1+h\lambda)y_k$
$F(h\lambda) = 1+h\lambda$

- Eulero implicito

$$y_{k+1} = y_k + h\lambda y_{k+1} = \frac{1}{1-h\lambda} y_k$$

$$F(h\lambda) = \frac{1}{1-h\lambda}$$

- Heun

$$y_{k+1} = y_k + \frac{h}{2}(\lambda y_k + \lambda(y_k + h\lambda y_k)) = y_k + h\lambda y_k + \frac{1}{2}(h\lambda)^2 y_k$$

$$F(h\lambda) = 1 + h\lambda + \frac{1}{2}(h\lambda)^2$$

- Runge–Kutta ad s stadi di ordine s

$$F(h\lambda) = 1 + h\lambda + \frac{1}{2}(h\lambda)^2 + \ldots + \frac{1}{s!}(h\lambda)^s$$

Allora, si deduce che

$$y_{k+1} = F(h\lambda)y_k = F(h\lambda)^2 y_{k-1} = \ldots = F(h\lambda)^{k+1} y_0 .$$

E perché il metodo sia assolutamente stabile deve valere

$$\lim_{k\to\infty} y_k = 0 \Leftrightarrow |F(h\lambda)| < 1 .$$

Si definisce <u>regione di assoluta stabilità</u> del metodo numerico la regione così definita

$$R_a = \{h\lambda \in C : |F(h\lambda)| < 1\}$$

dove con C si indica l'insieme dei numeri complessi.

Pertanto, le condizioni da imporre sono

- Eulero esplicito: $|1+h\lambda<1|$;
- Eulero implicito: $\left|\frac{1}{1-h\lambda}\right|<1$;
- Heun: $\left|1+h\lambda+\frac{1}{2}(h\lambda)^2\right|<1$.

I metodi impliciti hanno una regione di assoluta stabilità più grande di quella dei metodi espliciti.

Esempio: conseguenze pratiche.

Per Eulero esplicito la regione di assoluta stabilità è

$$R_a=\{h\lambda\in C:|1+h\lambda|<1\}$$

$|h\lambda-(-1)|<1$ sono tutti i punti del piano complesso che distano meno di 1 dal punto -1. In sostanza, sono i punti interni alla circonferenza di raggio 1 centrata in -1.

Se $\lambda=-10\in\Re^-$ si ha

$$|1-h\lambda|<1\Rightarrow -1<h\lambda<1\Rightarrow\begin{cases}1+h\lambda<1\\1+h\lambda>-1\end{cases}\Rightarrow\begin{cases}h\lambda<0\\h\lambda>-2\end{cases}\Rightarrow\begin{cases}h>0\\h<-\frac{2}{\lambda}\end{cases}$$

poiché, essendo $\lambda<0$, i segni di disequazione si invertono nell'ultimo passaggio.

Nel caso reale, per avere assoluta stabilità, è necessario imporre $-2<h\lambda<0$.

Per Eulero implicito si ha

$$R_a=\left\{h\lambda\in C:\left|\frac{1}{1-h\lambda}\right|<1\right\}$$

$$\left|\frac{1}{1-h\lambda}\right|<1\Leftrightarrow\frac{1}{|1-h\lambda|}<1\Leftrightarrow|1-h\lambda|>1\Leftrightarrow|h\lambda-1|>1$$

che è l'insieme di tutti i punti esterni alla circonferenza di centro $+1$ e raggio 1.

Quindi si verifica che $\Re e(h\lambda)\setminus[0,2]\subset R_a\cap\Re e(h\lambda)$. Dato che $h>0$, che $\Re e\,\lambda<0$ e che quindi $\lambda\in\Re^-$, allora anche $h\lambda\in\Re^-$.

Per questo motivo $h\lambda\in R_a\ \ \forall h>0$.

Si deduce che il metodo di Eulero implicito è incondizionatamente stabile, cioè garantisce la stabilità senza porre alcuna condizione sul passo. I metodi impliciti sono metodi incondizionatamente stabili, oppure pongono restrizioni molto meno stringenti rispetto a quanto accade con i metodi espliciti sul passo.

Il metodo di Heun genera una regione di assoluta stabilità di forma ellittica che, pur essendo più grande di quella del metodo di Eulero esplicito, non migliora le condizioni poste sul passo

$$R_a = \left\{ h\lambda \in C : \left| 1 + h\lambda + \frac{1}{2}(h\lambda)^2 \right| < 1 \right\}.$$

I metodi di Runge–Kutta ad s stadi, di ordine s si comportano allo stesso modo: aumentando l'ordine del metodo non migliora sensibilmente la stabilità

$$R_a = \left\{ h\lambda \in C : \left| 1 + h\lambda + \ldots + \frac{1}{s!}(h\lambda)^s \right| < 1 \right\}.$$

Il metodo dei trapezi, invece, ha come regione di assoluta stabilità tutto il semipiano delle ascisse negative del piano complesso:

$$R_a = C^-.$$

Sistemi di equazioni differenziali ordinarie

Un sistema di equazioni differenziali ordinarie può essere espresso in forma matriciale nel seguente modo

$$y'(t) = A(t)y(t) + g(t).$$

La matrice $A(t)$ si chiama matrice dei coefficienti, il termine $g(t)$ rappresenta i termini sorgente non autonomi, ossia quelli in funzione della variabile t, ma che non sono parte della funzione incognita.

Nel caso particolare di un sistema lineare di equazioni differenziali ordinarie a coefficienti costanti, senza termini autonomi, si pone

$$y'(t) = Ay(t).$$

Si può dimostrare che gli autovalori della matrice A sono i responsabili della stabilità. Se la parte reale di tutti gli autovalori è negativa, allora la stabilità è assicurata.

Se A è diagonalizzabile e λ_i sono i suoi autovalori, con autovettori v_i, per $i=1,2,...,m$ si ha

$$y(t)=c_1e^{\lambda_1(t-t_0)}v_1+...+c_me^{\lambda_m(t-t_0)}v_m.$$

Se $\Re e(\lambda_i)<0$ per $i=1,2,...,m$ il problema è asintoticamente stabile. Affinché un metodo numerico sia stabile occorre che per ogni autovalore λ_i si abbia $h\lambda_i \in R_a$.

Esempio: in caso di più autovalori, quello più negativo fornisce la restrizione sul passo.

$$y'=Ay;\ A=\begin{pmatrix}0 & 1\\ -6 & -5\end{pmatrix} \Rightarrow \begin{cases}y_1'=y_2\\ y_2'=-6y_1-5y_2\end{cases}$$

$$\det(A-\lambda I)=\begin{vmatrix}-\lambda & 1\\ -6 & -5-\lambda\end{vmatrix}=\lambda(5+\lambda)+6=\lambda^2+5\lambda+6=(\lambda+2)(\lambda+3)$$

$\lambda_1=-2$
$\lambda_2=-3$ $\Rightarrow$ il problema è asintoticamente stabile.

Utilizzando Eulero esplicito:

$$|1+h\lambda|<1 \Rightarrow h<-\frac{2}{\lambda} \Rightarrow \begin{matrix}h<-\frac{2}{-2}=1\\ h<-\frac{2}{-3}=\frac{2}{3}\end{matrix}$$ la seconda condizione è più restrittiva ed è quindi quella che influisce sulla scelta del passo h. Corrisponde all'autovalore più negativo.

Eventuali termini sorgente $g(t)$ non autonomi non influiscono in alcun modo sulla stabilità. Semplicemente si avrà

$$y'=Ay(t)+g(t).$$

Problemi stiff

Un problema stiff è ostico da integrare, ma rappresenta molti dei problemi reali (ad esempio, in chimica).

Definizione: un sistema di equazioni differenziali ordinarie lineari a coefficienti costanti del tipo

$$y'(t)=Ay(t)+g(t)$$

si dice stiff in $I=[t_0,t_0+L]$ se

- ci sono autovalori λ_i con $\Re e(\lambda_i)>0$ che soddisfino $\Re e(\lambda_i)\cdot L$ non grande;
- esiste almeno un autovalore λ_i con $\Re e(\lambda_i)<0$ e si verifica che $\Re e(\lambda_i)\cdot L << -1$.

Si dice grado di stiffness il $\max_i |\Re e(\lambda_i)\cdot L|$.

A causa dell'autovalore che soddisfa il secondo punto della definizione, se R_a è piccola, allora h risulta essere molto piccolo, spesso troppo piccolo. Ecco perché i metodi impliciti risultano utili, pur essendo più dispendiosi dal punto di vista computazionale e meno pratici da utilizzare. Infatti, hanno una regione di assoluta stabilità che è maggiore di quanto accade con i metodi espliciti.

Esempio:

$$\begin{cases} y_1'(t)=y_2(t) \\ y_2'(t)=-100y_1(t)-101y_2(t) \\ y_1(0)=\sqrt{7} \\ y_2(0)=\pi \end{cases} \text{ in } t\in[0,100].$$

La matrice del sistema è $A=\begin{pmatrix} 0 & 1 \\ -100 & -101 \end{pmatrix}$.

Gli autovalori sono $\lambda_1(A)=-10$ e $\lambda_2(A)=-1$.

L'autovalore λ_1 porta ad avere $\Re e(\lambda_1)\cdot L << -1$. Se si risolvesse il problema applicando il metodo di Eulero esplicito sarebbe necessario usare un passo $h\le 0.02$ e quindi sarebbe necessario compiere almeno 5000 passi prima di coprire l'intero intervallo $[0,100]$.

Esempio:

$$\begin{cases} y_1'(t)=y_2(t) \\ y_2'(t)=-10y_1(t)-11y_2(t) \\ y_1(0)=\sqrt{7} \\ y_2(0)=\pi \end{cases} \text{ in } t\in[0,10000].$$

La matrice del sistema è $A=\begin{pmatrix} 0 & 1 \\ -10 & -11 \end{pmatrix}$.

Gli autovalori sono $\lambda_1(A)=-10$ e $\lambda_2(A)=-1$.

L'autovalore λ_1 porta ad avere $\Re e(\lambda_1)\cdot L << -1$. Se si risolvesse il problema applicando il metodo di Eulero esplicito sarebbe necessario usare un passo $h \le 0.2$ e quindi sarebbe necessario compiere <u>almeno</u> 50000 passi prima di coprire l'intero intervallo $[0,10000]$.

Quanto contribuisce l'autovalore più piccolo sulla soluzione?

Esempio:

$$\begin{cases} y_1'(t) = y_2(t) \\ y_2'(t) = -1000\,y_1(t) - 1001\,y_2(t) \\ y_1(0) = 2 \\ y_2(0) = -1 \end{cases}.$$

La matrice del sistema è $A = \begin{pmatrix} 0 & 1 \\ -1000 & -1001 \end{pmatrix}$.

Gli autovalori sono $\lambda_1(A) = -1$ e $\lambda_2(A) = -1000$.

Se si risolvesse il problema applicando il metodo di Eulero esplicito sarebbe necessario usare un passo $h \le \frac{1}{500}$.

La soluzione complessiva del problema è

$$y(t) = c_1 e^{-(t-t_0)} v_1 + \underbrace{c_2 e^{-1000(t-t_0)} v_2}_{transiente}.$$

Il <u>transiente</u> "dura" pochissimo. Il termine corrispondente all'autovalore -1000 va a zero in un brevissimo tempo. Eppure condiziona tantissimo la scelta del passo: non dà praticamente alcun contributo percettibile alla soluzione (solo ingrandendo moltissimo i primi istanti di tempo si nota una differenza tra l'uscita completa e quella legata al solo autovalore -1), ma determina il passo massimo utilizzabile e la stabilità.

Quindi il termine legato all'autovalore più piccolo <u>NON è semplificabile</u>.

Per risolvere efficacemente ed efficientemente questi problemi si utilizzano spesso i metodi impliciti, che pongono minori restrizioni sulla scelta del passo.

Generalizzando: se il problema non è lineare, oppure non è assolutamente stabile, è necessario introdurre concetti di stabilità differenti, volti comunque a garantire che la soluzione numerica abbia il medesimo comportamento qualitativo della soluzione esatta.