

Capitolo 4

Una introduzione alla risoluzione numerica di ODE: problemi a valori iniziali

4.1 Alcuni esempi di problemi a valori iniziali

Le *equazioni differenziali* rappresentano il modello matematico di numerosi fenomeni, in vari campi della scienza e della tecnica, quali, ad esempio, la meccanica, la chimica, la biologia, l'astronomia, le scienze ambientali, l'economia etc etc. Risulta quindi di importanza fondamentale avere a disposizione strumenti per la risoluzione di tali equazioni, che consentano di calcolare in maniera efficiente i valori numerici delle soluzioni.

In generale, un'equazione differenziale è un'equazione che coinvolge una funzione incognita e le sue derivate. In particolare, si parla di *equazioni differenziali ordinarie* (ODE)¹ se l'incognita è funzione di una sola variabile indipendente.

♣ **Esempio 4.1. (Moto rettilineo uniforme)** Si consideri un punto che si muove di moto rettilineo uniforme con velocità v . La posizione $x(t)$ del punto ad un generico istante t , lungo la retta che costituisce la traiettoria del moto, soddisfa la seguente ODE:

$$x'(t) = v \tag{4.1}$$



¹Dall'inglese *Ordinary Differential Equations*.

♣ **Esempio 4.2. (Decadimento radioattivo)** In natura esistono alcuni elementi chimici che decadono spontaneamente in loro isotopi o in altri elementi, mediante emissione di particelle α , particelle β o fotoni. Tali elementi sono detti radioattivi. Ad esempio il carbonio ^{14}C decade nel suo isotopo ^{12}C . Se si considera un numero elevato di atomi radioattivi, cioè dell'ordine del numero di Avogadro ($\sim 10^{23}$), la velocità con cui tale numero varia è direttamente proporzionale al numero stesso di atomi, ovvero il numero $N(t)$ di atomi radioattivi all'istante t soddisfa la ODE:

$$N'(t) = -kN(t) \quad (4.2)$$

dove k è una costante positiva, generalmente determinata in maniera sperimentale.²

♣

Le equazioni (4.1) e (4.2) costituiscono un esempio di ODE del *primo ordine*. In generale, una ODE del primo ordine è un'equazione del tipo:

$$y'(t) = f(t, y(t)), \quad t \in I \quad (4.3)$$

dove I è un intervallo dell'asse reale, y , l'incognita, è una funzione reale di variabile reale, derivabile in I , ed f è una funzione reale nota, definita in $I \times \mathbb{R}$. L'equazione si dice del primo ordine in quanto la derivata di ordine massimo della funzione incognita che compare in essa è la derivata prima. L'equazione (4.3) è inoltre in *forma esplicita*, cioè è esplicitata rispetto alla derivata di ordine massimo di y .

Spesso un'equazione di questo tipo rappresenta l'evoluzione temporale di qualche fenomeno e la variabile indipendente t rappresenta il tempo. Nel seguito si considera $I = [t_0, T]$, con $t_0 < T < +\infty$ oppure $I = [t_0, +\infty)$.

Una soluzione di (4.3) è denominata anche *integrale* di (4.3); da (4.3) discende infatti che

$$y(t) = \int_{t_0}^t f(\tau, y(\tau)) d\tau + c, \quad t \in I,$$

con c costante arbitraria. Le eventuali soluzioni di (4.3) costituiscono quindi una famiglia di funzioni, dipendenti da un parametro.

♣ **Esempio 4.3. (Moto rettilineo uniforme - continuazione)** La generica soluzione dell'equazione (4.1) ha l'espressione

$$x(t) = vt + c,$$

dove c è una costante arbitraria. Tale soluzione è rappresentata, al variare di c in Figura 4.1 (sinistra).

E' chiaro che la posizione del punto all'istante t dipende dalla posizione del punto ad un istante precedente t_0 , che si può assumere come istante iniziale, cioè dipende dalla condizione iniziale

$$x(t_0) = x_0.$$

Imponendo tale condizione si determina un'unica soluzione

$$x(t) = v(t - t_0) + x_0,$$

che fornisce la posizione del punto nel caso particolare considerato.

♣

²Si noti che il numero di atomi è trattato come una grandezza continua e non discreta, quale è in realtà. Tale approssimazione si può ritenere valida perché si considera un numero elevato di atomi.

♣ **Esempio 4.4. (Decadimento radioattivo - continuazione)** La generica soluzione dell'equazione (4.2) è del tipo

$$N(t) = c e^{-kt},$$

dove c è una costante arbitraria. Tale soluzione è rappresentata, al variare di c e per $k = 1$, in Figura 4.1 (destra).

Analogamente all'esempio precedente, il numero di atomi ad un certo istante t dipende dal numero iniziale di atomi, cioè dipende dalla condizione iniziale

$$N(t_0) = N_0,$$

dove t_0 indica l'istante iniziale. Imponendo tale condizione si determina la soluzione particolare

$$N(t) = N_0 e^{-k(t-t_0)}.$$

♣

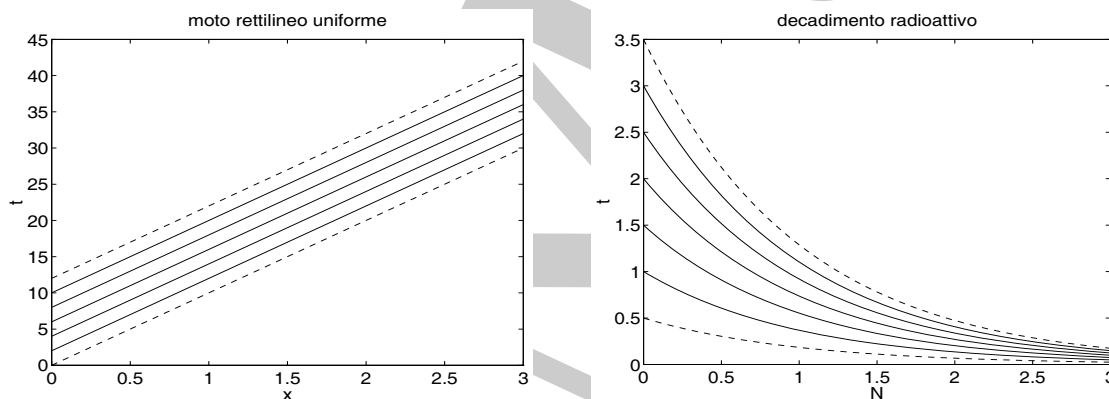


Figura 4.1: Famiglia di soluzioni dell'equazione del moto rettilineo uniforme e dell'equazione del decadimento radioattivo.

In generale, sotto ipotesi opportune su f , l'assegnazione di una condizione aggiuntiva sull'incognita y , quale ad esempio la condizione

$$y(t_0) = y_0, \quad (4.4)$$

consente di determinare una ed una sola soluzione dell'equazione. La condizione (4.4) si dice *condizione iniziale* ed il problema

$$M(P) \equiv \begin{cases} y'(t) &= f(t, y(t)), & t \in I \\ y(t_0) &= y_0 \end{cases} \quad (4.5)$$

si dice *problema di Cauchy* o *problema a valori iniziali (IVP)*³. Nel seguito si suppone che f verifichi le ipotesi che garantiscono l'esistenza e l'unicità della soluzione, ovvero le ipotesi del teorema:

Teorema 4.1. [Esistenza e unicità]

Sia $f(t, y)$ continua nella striscia $S = I \times \mathbb{R} = \{(t, y) : t \in I, y \in \mathbb{R}\}$, con $I = [a, b]$. Si supponga che f soddisfi in S una condizione di Lipschitz rispetto ad y , uniformemente rispetto a t , cioè che esista $L > 0$ tale che:

$$|f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2| \quad (4.6)$$

per ogni $t \in I$ e $y_1, y_2 \in \mathbb{R}$.

Allora, per ogni $t_0 \in I$ e $y_0 \in \mathbb{R}$, esiste una ed una sola soluzione del problema (4.5), cioè esiste una ed una sola funzione $y(t)$, derivabile in I , tale che

$$\begin{cases} y'(t) &= f(t, y(t)) & t \in I \\ y(t_0) &= y_0 \end{cases}$$

Più in generale, nella costruzione di modelli matematici di problemi reali, si possono ottenere problemi a valori iniziali relativi a *sistemi di ODE* del primo ordine.

♣ **Esempio 4.5. (Problema predatore-preda)** Si considerino due popolazioni di individui appartenenti a specie che interagiscono secondo una relazione predatore-preda, quali ad esempio squali e pesci più piccoli nell'oceano, oppure lupi e lepri. Si supponga che la popolazione di prede sia in grado di procurarsi cibo a sufficienza e che quindi il numero di prede $x(t)$, in assenza di predatori, cresca con una velocità $x'(t)$ direttamente proporzionale al numero stesso. Analogamente, si supponga che il numero di predatori $y(t)$ diminuisca, in assenza di prede, ovvero per mancanza di cibo, con una velocità $y'(t)$ direttamente proporzionale ad $y(t)$ stesso. Si supponga inoltre che il numero di prede uccise dai predatori dipenda dalla possibilità che predatore e preda si incontrino, più precisamente che $x(t)$ diminuisca con una velocità direttamente proporzionale al prodotto $x(t)y(t)$ del numero di prede per il numero di predatori, e che $y(t)$ cresca in maniera analoga. In tali ipotesi, la dinamica delle due popolazioni è descritta dal seguente sistema di due ODE, note come equazioni di *Lotka-Volterra*:⁴

$$\begin{cases} x'(t) &= ax(t) - bx(t)y(t) \\ y'(t) &= -cy(t) + dx(t)y(t) \end{cases}$$

dove a, b, c, d sono costanti positive. A tali equazioni vanno aggiunte le condizioni iniziali:

$$x(t_0) = x_0, \quad y(t_0) = y_0,$$

che specificano il numero di individui ad un certo istante t_0 , assunto come istante iniziale.

In Figura 4.2 sono rappresentate le traiettorie $(x, y) = (x(t), y(t))$, per valori fissati dei parametri a, b, c, d . L'andamento delle traiettorie mostra che un aumento di prede comporta un aumento di predatori e quindi una successiva diminuzione di prede e poi di predatori. Il punto $(\bar{x}, \bar{y}) = (c/d, a/b)$ costituisce un punto stazionario, cioè un punto in cui $(x'(t), y'(t)) = (0, 0)$, e rappresenta quindi lo stato stazionario del sistema preda-predatore. ♣

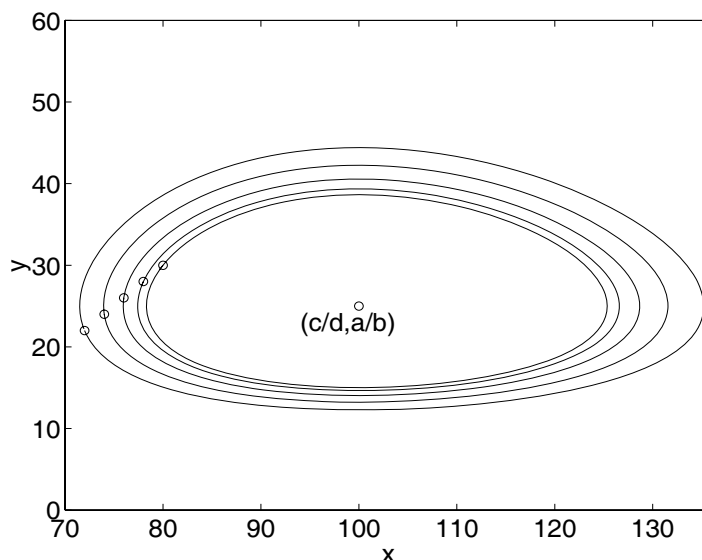


Figura 4.2: Evoluzione delle popolazioni nel problema predatore-preda (il simbolo $\circ$ indica il punto iniziale (x_0, y_0)).

In generale un problema a valori iniziali per un sistema di ODE del primo ordine, in forma esplicita, è del tipo:

$$\begin{cases} y'_i(t) = f_i(t, y_1(t), y_2(t), \dots, y_n(t)), & t \in I \\ y_i(t_0) = y_{i,0} & i = 1, \dots, n \end{cases} \quad (4.7)$$

Posto

$$Y(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_n(t) \end{bmatrix}, \quad F(t, Y) = \begin{bmatrix} f_1(t, y_1, y_2, \dots, y_n) \\ f_2(t, y_1, y_2, \dots, y_n) \\ \vdots \\ f_n(t, y_1, y_2, \dots, y_n) \end{bmatrix},$$

il problema (4.7) si può riscrivere nella forma:

$$\begin{cases} Y'(t) = F(t, Y(t)) & t \in I \\ Y(t_0) = Y_0 \end{cases},$$

con $Y_0 = [y_{1,0}, y_{2,0}, \dots, y_{n,0}]^T$, cioè si può rappresentare come problema a valori inizia-

³Dall'inglese *Initial Value Problem*.

⁴Come nell'Esempio 4.2, il modello in esame presuppone che $x(t)$ e $y(t)$ siano grandezze continue, mentre esse sono discrete.

li per una singola ODE vettoriale. Se l'ordine massimo delle derivate della funzione incognita è m , si parla di ODE di ordine m .

♣ **Esempio 4.6. (Moto del pendolo semplice)** Si consideri un pendolo semplice, cioè una particella di massa m sospesa ad un punto O mediante una fune inestensibile di lunghezza L e di massa trascurabile. La variabile s misura la lunghezza dell'arco descritto dal pendolo a partire dalla posizione di equilibrio, ovvero dal punto più basso della circonferenza di centro O e raggio L (Fig. 4.3).

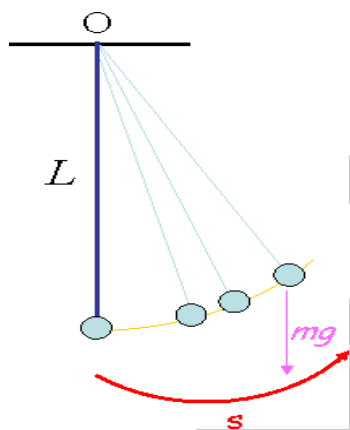


Figura 4.3: Pendolo semplice.

Applicando la legge di Newton, il moto della particella è regolato dalla ODE del secondo ordine:

$$s''(t) = -\frac{\lambda}{mL}s'(t) - \frac{g}{L}\sin s(t), \quad (4.8)$$

dove $s(t)$ è l'arco che la fune descrive con la verticale per O all'istante t , g è l'accelerazione di gravità e λ è un coefficiente dipendente dalla viscosità dell'aria. All'equazione (4.8) sono associate le condizioni iniziali

$$s(t_0) = \hat{s}, \quad s'(t_0) = 0,$$

che indicano posizione e velocità della particella all'istante iniziale t_0 . La soluzione

$$s(t) = \epsilon a_0 e^{-\frac{\lambda}{2mL}t} \cos \left[\omega_0 t - \frac{\epsilon^2 a_0^2}{\omega_0 \frac{\lambda}{2mL} 32} e^{-\frac{\lambda}{mL}t} + b_0 \right] + \mathcal{O}(\epsilon^3)$$

con ϵ , a_0 e b_0 costanti e con $\omega_0 = \sqrt{\frac{g}{L}}$, descrive il moto del pendolo nell'intervallo di tempo $[t_0, T]$ in cui esso percorre, in senso orario, l'arco di lunghezza $s_0 = \epsilon a_0$, partendo dalla posizione iniziale, occupata all'istante t_0 (Fig. 4.4). Successivamente, a partire dalla posizione $s_0 = \epsilon a_0$, per $t \geq T$ il moto cambia verso e prosegue regolato dalla legge

$$s(t) = s_0 e^{-\frac{\lambda}{2mL}t} \cos \left[\omega_0 t - \frac{s_0^2}{\omega_0 \frac{\lambda}{2mL} 32} \left(e^{-\frac{\lambda}{mL}t} - 1 \right) \right] + \mathcal{O}(\epsilon^2).$$

Per effetto della resistenza dell'aria il pendolo oscilla descrivendo archi la cui lunghezza diminuisce gradualmente (Fig. 4.5), fino a quando il pendolo si ferma.

Aggiungendo l'ipotesi "s piccolo" rispetto a L , ovvero

$$\sin \frac{s}{L} \approx \frac{s}{L},$$

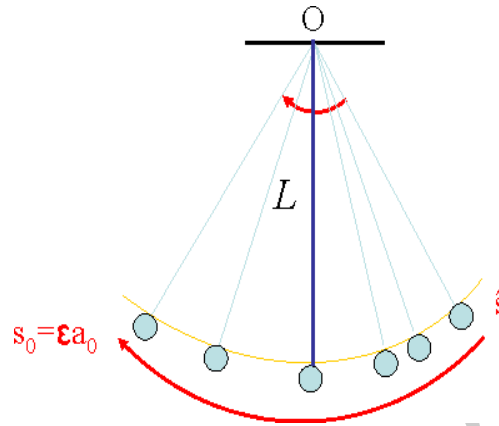


Figura 4.4: Pendolo semplice.

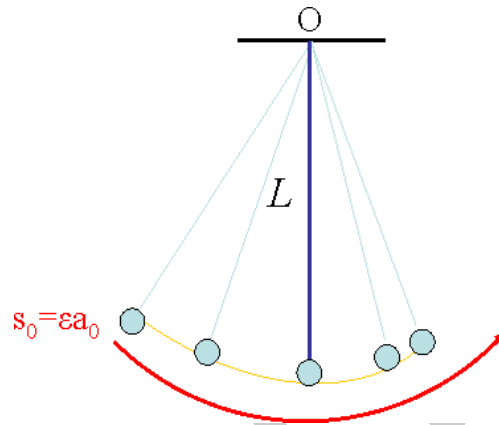


Figura 4.5: Pendolo semplice.

il modello nella (4.8) risulta semplificato, ovvero descritto dalla ODE:

$$s''(t) = -\omega^2 s(t),$$

con $\omega^2 = \frac{g}{L}$, la cui soluzione è

$$s = s_0 + \cos(\omega t + \delta),$$

con δ spostamento angolare all'istante $t = 0$ e s_0 arco di lunghezza massima, descritto dal pendolo. ♣

In generale, una ODE di ordine m , in forma esplicita, ha un'espressione del tipo:

$$y^{(m)}(t) = f(t, y(t), y'(t), \dots, y^{(m-1)}(t)) \quad (4.9)$$

e per ottenere un'unica soluzione di tale equazione è necessario assegnare m condizioni in un punto t_0 , su y e tutte le sue derivate fino a quella di ordine $m - 1$.

Si osservi che una ODE di ordine superiore al primo si può ricondurre ad un sistema

di ODE del primo ordine. Infatti, posto

$$\begin{cases} w_1(t) &= y(t) \\ w_2(t) &= y'(t) \\ &\vdots \\ w_m(t) &= y^{(m-1)}(t) \end{cases}$$

l'equazione (4.9) si può riscrivere nel modo seguente:

$$\begin{cases} w'_1(t) &= w_2(t) \\ w'_2(t) &= w_3(t) \\ &\vdots \\ w'_m(t) &= f(t, w_1(t), w_2(t), \dots, w_m(t)) \end{cases}$$

Per la ODE dell'Esempio 4.6 si ha:

$$\begin{cases} s'_1(t) &= s_2(t) \\ s'_2(t) &= -\frac{\lambda}{mL}s_2(t) - \frac{g}{L}\sin s_1(t) \end{cases}$$

Nel seguito si rivolge quindi l'attenzione alle ODE del primo ordine.

Osservazione 4.1. È interessante notare che la necessità di risolvere una ODE può risultare anche dall'applicazione di metodi numerici ad altri modelli matematici. Ad esempio, se si applica il metodo delle linee [8] ad una equazione differenziale a derivate parziali, quale:

$$\frac{\partial}{\partial t}y(t, x) = \frac{\partial^2}{\partial x^2}(y(t, x)),$$

si ottiene un sistema di ODE del tipo:

$$Y'(t) = AY(t) + b(t),$$

dove $Y(t) = (y_1(t), y_2(t), \dots, y_n(t))^T$, con $\{y_i(t), i = 1, 2, \dots, n\}$ approssimazioni di $y(t, x)$ in un insieme finito di valori $\{x_i, i = 1, 2, \dots, n\}$.

4.2 Un metodo numerico di risoluzione: il metodo di Eulero

La necessità di avere a disposizione metodi numerici e software per la risoluzione di problemi a valori iniziali per equazioni differenziali ordinarie (ODE) nasce non solo dal fatto che in generale non è possibile ottenere le soluzioni in forma analitica ma soprattutto dal fatto che, in numerose applicazioni, è sentita l'esigenza di avere strumenti in grado di calcolare la soluzione numerica in maniera automatica e ripetitiva oltre che, chiaramente, affidabile ed efficiente ed il software numerico gioca quindi un ruolo fondamentale.

Si consideri il problema a valori iniziali (4.5), dove $I = [t_0, T]$. Il più semplice dei metodi numerici per la risoluzione di tale problema si può ottenere applicando la formula di Taylor alla funzione incognita y . Si consideri lo sviluppo in serie di Taylor della funzione y di punto iniziale x_0 arrestato al termine del primo ordine:

$$y(x) = y(x_0) + (x - x_0)y'(x_0) + \tau(x_0, x - x_0), \quad (4.10)$$

dove $\tau(x_0, x - x_0)$ è il resto della serie. Ponendo nella (4.10)

$$x_0 = t, \quad x - x_0 = h$$

e quindi $x = t + h$, si ottiene:

$$y(t + h) = y(t) + hy'(t) + \tau(t, h) = y(t) + hf(t, y(t)) + \tau(t, h). \quad (4.11)$$

Si consideri l'espressione di $y(t + h)$ ottenuta trascurando il resto $\tau(t, h)$:

$$y(t + h) \simeq y(t) + hf(t, y(t)). \quad (4.12)$$

Posto

$$h = \frac{T - t_0}{n}, \quad (4.13)$$

con n intero positivo, si considerino i punti dell'asse t $t_0, t_1, \dots, t_n = T$, con

$$t_i = t_0 + ih,$$

cioè si *discretizzi* l'intervallo $[t_0, T]$ utilizzando un *passo di discretizzazione* h . Sostituendo t_i a t e t_{i+1} a $t + h$ nella (4.12) si ha la relazione:

$$y(t_{i+1}) \simeq y(t_i) + hf(t_i, y(t_i)),$$

che consente di calcolare un'approssimazione y_i di y nel punto t_i , noto il valore iniziale $y(t_0) = y_0$:

$$y_{i+1} = y_i + hf(t_i, y_i), \quad i = 0, 1, 2, \dots, n - 1. \quad (4.14)$$

La (4.14), unitamente al valore noto $y(t_0) = y_0$, descrive un metodo numerico per la risoluzione del problema (4.5), noto come *metodo di Eulero*.

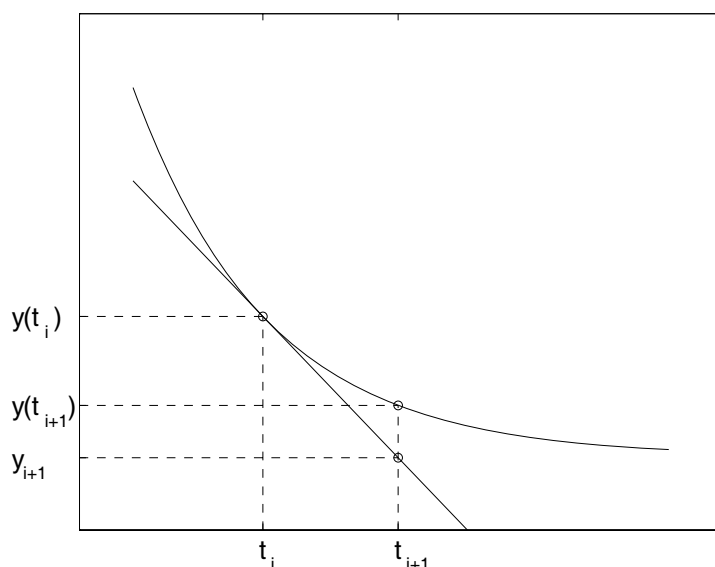


Figura 4.6: Interpretazione geometrica del metodo di Eulero.

Da un punto di vista geometrico si ha la seguente interpretazione del metodo di Eulero: nel punto $(t_i, y(t_i))$ si sostituisce la curva che rappresenta la soluzione $y(t)$ con la retta ad essa tangente e si calcola y_{i+1} come ordinata del punto di ascissa t_{i+1} . Tale situazione è illustrata graficamente in Figura 4.6.

La relazione (4.14) mostra che y_{i+1} dipende da y_i , ma non da y_j con $j < i$; per tale motivo il metodo di Eulero è detto *ad un passo*. Inoltre, il secondo membro della (4.14) non dipende da y_{i+1} e quindi il metodo è detto *esplicito*⁵. A partire dalla relazione (4.14) si ottiene immediatamente un algoritmo per la risoluzione numerica del problema (4.5) (Procedura 4.1).

⁵In generale, sussistono le definizioni seguenti.

Definizione 4.1. (Metodo ad un passo)

Un metodo numerico per la risoluzione del problema (4.5) si dice *ad un passo*, o *one-step*, se ha un'espressione del tipo

$$y_{i+1} = y_i + h\varphi(t_i, y_i, y_{i+1}, h, f), \quad (4.15)$$

cioè se y_{i+1} dipende solo da y_i .

Definizione 4.2. (Metodo ad un passo esplicito o implicito)

Un metodo numerico ad un passo per la risoluzione del problema (4.5) si dice *esplicito* se ha un'espressione del tipo

$$y_{i+1} = y_i + h\varphi(t_i, y_i, h, f),$$

cioè se φ non dipende da y_{i+1} . Altrimenti il metodo si dice *implicito*.

Si osservi che per il metodo di Eulero si ha $\varphi(t_i, y_i, h, f) = f(t_i, y_i)$.

```

procedure eulero(input:  $t_0, T, y_0, f, n$ , out:  $y$ )
  /# SCOPO: Risoluzione numerica di equazioni differenziali
  /# ordinarie del primo ordine in forma esplicita
  /# SPECIFICHE PARAMETRI:
  var:  $t_0$  : reale           {estremo sinistro dell'intervallo}
  var:  $T$  : reale           {estremo destro dell'intervallo}
  var:  $y_0$  : reale         {condizione iniziale}
  var:  $f$  : funzione esterna {funzione  $f(t, y)$ }
  var:  $n$  : intero         {numero di sottointervalli}
  var:  $y$  : array di reali   {restituisce la soluzione calcolata.}
  /# INIZIO ISTRUZIONI:
     $h := (T - t_0)/n$ ;
     $y(0) := y_0$ ;
    for  $i = 1, n$  do
       $t := t_0 + (i - 1) * h$ ;
       $y(i) := y(i - 1) + h * f(t, y(i - 1))$ ;
    endfor
end procedure eulero

```

Procedura 4.1: Algoritmo per la risoluzione numerica del problema (4.5) con il metodo di Eulero

La *complessità di tempo* della Procedura 4.1 dipende essenzialmente dal numero di valutazioni della funzione $f(t, y)$, in quanto il numero di operazioni ad essa relative è in generale predominante rispetto alle altre operazioni eseguite nell'algoritmo. In questo caso il numero di valutazioni di f è n , cioè è uguale al numero di intervalli in cui è stato suddiviso l'intervallo di definizione del problema continuo. L'algoritmo ha dunque una complessità di tempo

$$T_{\text{Eulero}}(n) = \mathcal{O}(n) \text{ valutazioni di funzione.}$$

La *complessità di spazio* è

$$S_{\text{Eulero}}(n) = \mathcal{O}(n) \text{ variabili reali,}$$

i	t_i	$\tilde{u}(t_i)$	$y(t_i)$
0	0.0	1.000	1.000
1	0.1	0.900	0.909
2	0.2	0.820	0.835
3	0.3	0.756	0.774
4	0.4	0.705	0.725
5	0.5	0.664	0.684
6	0.6	0.631	0.651
7	0.7	0.605	0.623
8	0.8	0.584	0.601
9	0.9	0.567	0.583
10	1.0	0.554	0.568

Tabella 4.1: Applicazione del metodo di Eulero al problema (4.16). Valori della soluzione calcolata $\tilde{u}(t_i)$ e della soluzione effettiva $y(t_i)$.

in quanto l'algoritmo utilizza un array di reali, y , di dimensione $n + 1$.

♣ **Esempio 4.7.** Si consideri il problema a valori iniziali:

$$\begin{cases} y'(t) &= -2y(t) + 1, & t \in [0, 1] \\ y(0) &= 1 \end{cases} \quad (4.16)$$

la cui soluzione è:

$$y(t) = \frac{1}{2} (e^{-2t} + 1) .$$

Applicando ad esso la Procedura 4.1 con $n = 10$, ovvero $h = 0.1$, si ottengono i valori $\tilde{u}(t_i)$ riportati in Tabella 4.1.

♣

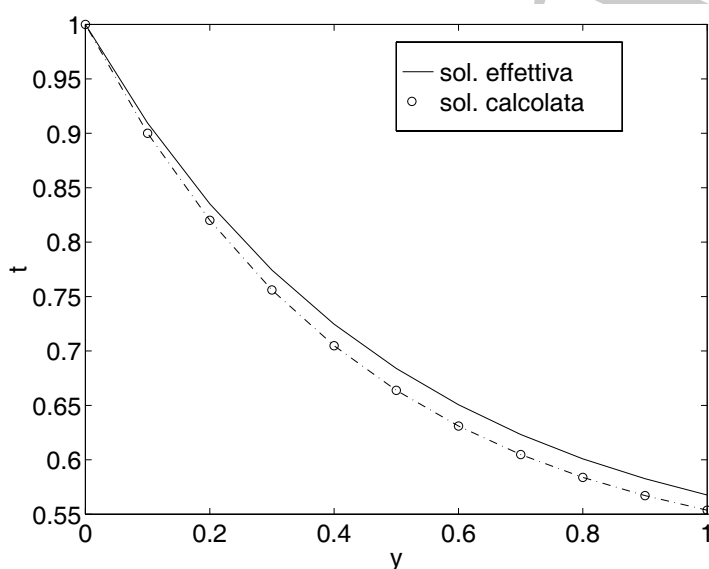


Figura 4.7: Applicazione del metodo di Eulero al problema (4.16): soluzione calcolata e soluzione effettiva. I grafici sono relativi ai dati in Tabella 4.1.

4.3 Analisi degli errori introdotti dal metodo di Eulero

Nell'Esempio 4.7 i valori della soluzione effettiva e di quella calcolata con il metodo di Eulero sono diversi. Ciò è in accordo col fatto che la curva soluzione è approssimata ad ogni passo con la retta tangente ad essa in un determinato punto. In altri termini, l'espressione (4.12), dalla quale si ottiene il metodo di Eulero, corrisponde alla sostituzione della derivata prima della funzione incognita con il rapporto incrementale:

$$y'(t) \simeq \frac{y(t+h) - y(t)}{h}.$$

Si osservi che, detta $u(t)$ una funzione di I in $\mathfrak{R}$ e posto $u(t_0) = y_0$, il metodo di Eulero fornisce la soluzione del problema discreto

$$M_h(P) \equiv \begin{cases} \frac{u(t_{i+1}) - u(t_i)}{h} = f(t_i, u(t_i)), & i = 0, 1, \dots, n-1 \\ u(t_0) = y(t_0) \end{cases} \quad (4.17)$$

In altri termini si sostituisce il problema continuo $M(P)$ (4.5) con il problema discreto $M_h(P)$ (4.17)⁶. Tale approssimazione non è però l'unica fonte di errore nel calcolo della soluzione.

♣ **Esempio 4.8.** Applicando la Procedura 4.1 al problema:

$$\begin{cases} y'(t) = 0.01, & t \in [0, 1] \\ y(0) = 1 \end{cases}, \quad (4.18)$$

con $n = 10^5$, ovvero $h = 10^{-5}$, si ottengono i risultati riportati in Tabella 4.2 (i valori della soluzione effettiva $y(t_i)$ e quelli della soluzione calcolata $\tilde{u}(t_i)$ sono rappresentati con sette cifre significative). Si noti che, anche se la soluzione del problema (4.18) è $y(t) = 0.01t + 1$ e quindi la retta tangente alla curva soluzione in un qualsiasi punto coincide con la curva stessa, cioè la soluzione del problema discreto $u(t_i)$ coincide con la soluzione del problema continuo $y(t_i)$, la soluzione calcolata $\tilde{u}(t_i)$ è diversa. Tale situazione è illustrata anche graficamente, in Figura 4.8. Ciò è dovuto al fatto che l'algoritmo è stato eseguito in un sistema aritmetico a precisione finita e quindi i valori calcolati $\tilde{u}(t_i)$ sono affetti dall'errore di roundoff. ♣

Gli errori che si verificano nel calcolo della soluzione numerica di un problema a valori iniziali mediante il metodo di Eulero sono da addebitarsi a due cause fondamentali:

⁶L'equazione in (4.17) è detta anche *equazione alle differenze*, in quanto la derivata $y'(t)$ è sostituita con la *differenza finita* $(u(t+h) - u(t))/h$. A tal proposito, si noti che esiste una classe di metodi detti *metodi alle differenze finite* per la risoluzione di equazioni differenziali, i quali si basano proprio sull'idea di approssimare il problema continuo con un problema discreto sostituendo l'equazione differenziale con una opportuna equazione alle differenze.

i	t_i	$\tilde{u}(t_i)$	$y(t_i)$
91000	0.91	1.010848	1.009100
92000	0.92	1.010967	1.009200
93000	0.93	1.011086	1.009300
94000	0.94	1.011206	1.009400
95000	0.95	1.011325	1.009500
96000	0.96	1.011444	1.009600
97000	0.97	1.011563	1.009700
98000	0.98	1.011683	1.009800
99000	0.99	1.011802	1.009900
100000	1.00	1.011921	1.010000

Tabella 4.2: Applicazione del metodo di Eulero al problema (4.18). Valori della soluzione calcolata $\tilde{u}(t_i)$ e della soluzione effettiva $y(t_i)$.

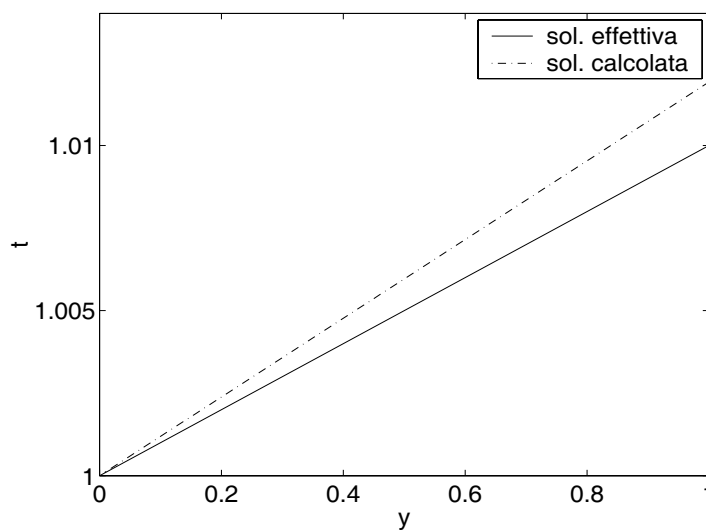


Figura 4.8: Applicazione del metodo di Eulero al problema (4.18): soluzione calcolata e soluzione effettiva. I grafici sono relativi ai dati in Tabella 4.2.

- (1) la sostituzione del problema continuo $M(P)$ (4.5) con il problema discreto $M_h(P)$ (4.17) (errore di discretizzazione);
- (2) la risoluzione del problema discreto $M_h(P)$ (4.17) in un sistema aritmetico floating-point a precisione finita (errore di round-off).

4.3.1 Il problema discreto come approssimazione del problema continuo

Si supponga di eseguire un solo passo del metodo di Eulero. Da (4.11) e (4.17) segue che l'errore commesso sostituendo ad $y(t_{i+1})$ il valore $u(t_{i+1})$, è:

$$y(t_{i+1}) - u(t_{i+1}) = y(t_i) + hf(t_i, y(t_i)) + \tau(t_i, h) - [u(t_i) + hf(t_i, u(t_i))] .$$

dove $\tau(t_i, h)$ esprime il resto di Taylor della funzione $y(t)$ nel punto t_i . Quindi supponendo nota la $y(t)$ in t_i , e posto $u(t_i) = y(t_i)$, si ottiene

$$y(t_{i+1}) - u(t_{i+1}) = \tau(t_i, h) .$$

L'errore $\tau(t_i, h)$, detto *errore locale di troncamento del metodo di Eulero* nel punto t_i , si commette ad ogni passo del metodo di Eulero e influenza il valore ottenuto al passo successivo. Sussiste la seguente:

Definizione 4.3. Si dice *errore di discretizzazione*⁷ del metodo di Eulero la quantità:

$$T(t, h) = \frac{\tau(t, h)}{h} , \quad (4.19)$$

cioè la differenza tra l'operatore differenziale del problema continuo (4.5) (derivata prima di $y(t)$) e l'operatore alle differenze del problema discreto (4.17) (rapporto incrementale).

Consistenza

Si vuole dare una stima dell'errore di discretizzazione del metodo di Eulero. Si ha:

Lemma 4.1. Se $y(t) \in C^{(2)}[t_0, T]$ è la soluzione del problema continuo $M(P)$ (4.5), per l'errore di discretizzazione del metodo di Eulero si ha:

$$T(t, h) = \mathcal{O}(h) .$$

Dimostrazione Dalla continuità della derivata seconda si ha per l'errore locale di troncamento:

$$\tau(t, h) = h^2 \frac{y''(\xi)}{2} \quad (t < \xi < t + h) .$$

⁷In generale sussiste la seguente definizione:

Definizione 4.4. (Errore di discretizzazione - metodo ad un passo)

Si dice *errore di discretizzazione di un metodo ad un passo del tipo (4.15)* la quantità:

$$T(t, h, \varphi) = \frac{y(t+h) - y(t)}{h} - \varphi(t, y(t), y(t+h), h, f) .$$

Si utilizzerà la notazione $T(t, h)$ in luogo di $T(t, h, \varphi)$, cioè non si esplicherà la dipendenza da φ , sia per il metodo di Eulero sia per un qualsiasi metodo, ogni qual volta dal contesto si evincerà chiaramente di quale metodo si sta parlando.

Posto

$$C = \max_{[t_0, T]} |y''| ,$$

risulta

$$|\tau(t, h)| \leq \frac{C}{2} h^2 \quad (4.20)$$

Quindi da (4.19) e (4.20) si ha per l'errore di discretizzazione:

$$|T(t, h)| = \frac{|\tau(t, h)|}{h} \leq \frac{C}{2} h$$

da cui la tesi. ■

Dal Lemma 4.1 si evince che per ogni punto t fissato, al tendere del passo h a 0 l'errore di discretizzazione tende a 0 con ordine di infinitesimo uguale a 1. A tal proposito si dice che *il metodo di Eulero è consistente di ordine 1*⁸. La consistenza esprime il fatto che il problema discreto (4.17) è un'approssimazione del problema continuo (4.5) e l'ordine esprime quanto tale approssimazione sia accurata. La consistenza costituisce una condizione necessaria per ottenere un'approssimazione della soluzione.

Convergenza

Anche se il metodo di Eulero è consistente, ovvero il problema discreto (4.17) è un'approssimazione del problema continuo (4.5), non è detto che la soluzione $u(t_i) = y_i$ del problema discreto (4.17), in un punto t_i , sia una buona approssimazione della soluzione $y(t_i)$ del problema continuo (4.5). È naturale allora chiedersi quale sia la differenza tra tali soluzioni. In particolare, per il metodo di Eulero, si noti che calcolare y_1 a partire da y_0 equivale ad applicare un solo passo del metodo di Eulero al problema continuo

$$\begin{cases} y'(t) &= f(t, y(t)) \\ y(t_0) &= u(t_0) = y_0 \end{cases} .$$

Analogamente, all' i -mo passo, calcolare y_i a partire dal valore y_{i-1} equivale ad applicare solo un passo del metodo di Eulero al problema continuo

$$\begin{cases} y'(t) &= f(t, y(t)) \\ y(t_{i-1}) &= u(t_{i-1}) = y_{i-1} \end{cases} .$$

⁸In generale si ha la seguente definizione:

Definizione 4.5. (Consistenza-metodo ad un passo)

Un metodo del tipo (4.15) si dice consistente se, per t fissato, risulta

$$\lim_{h \rightarrow 0} T(t, h) = 0.$$

Se, inoltre, si ha

$$|T(t, h)| \leq Dh^p ,$$

con $D > 0$ costante, il metodo si dice consistente di ordine p . In generale, p è detto ordine del metodo.

Il metodo di Eulero comporta quindi, ad ogni passo, il passaggio da una soluzione ad un'altra della famiglia di soluzioni dell'equazione differenziale (4.3), come illustrato in Figura 4.9.

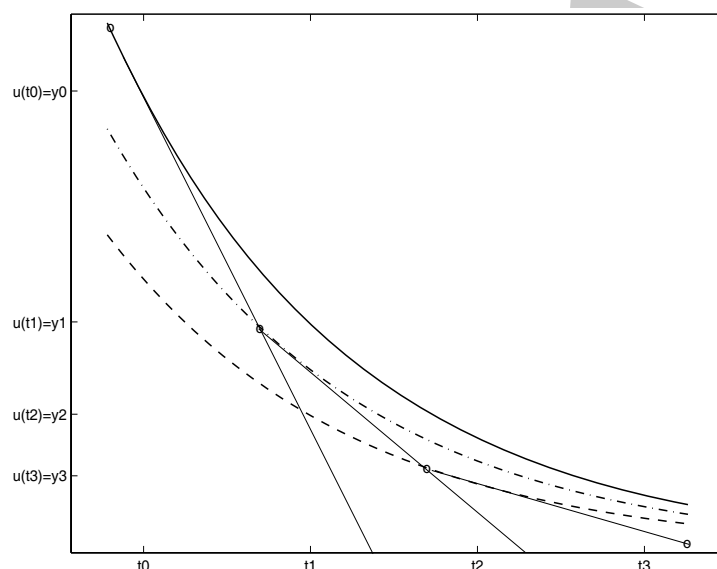


Figura 4.9: Esecuzione del metodo di Eulero: passaggio da una soluzione all'altra della famiglia di soluzioni dell'equazione (4.3).

La soluzione ottenuta dal metodo di Eulero, in un generico punto t_i , risente così di tutti gli errori commessi nei punti precedenti $(t_{i-1}, t_{i-2}, \dots, t_2, t_1)$. Si introduce pertanto la seguente:

Definizione 4.6. Si dice *errore globale di troncamento*⁹ del metodo di Eulero, nel punto t_i , la quantità

$$E_i(h) = y(t_i) - u(t_i), \quad (4.21)$$

cioè la differenza tra la soluzione del problema continuo $M(P)$ (4.5) e la soluzione del problema discreto $M_h(P)$ (4.17).

♣ **Esempio 4.9.** Si consideri nuovamente il problema (4.16) (Esempio 4.7) e si applichi ad esso la Procedura 4.1 con differenti valori del passo h , precisamente con $h = 0.2, 0.1, 0.05$. Si considerino ad esempio i valori calcolati della soluzione per $t = 0.4, 1$. I risultati riportati in Tabella 4.3 mostrano che

⁹In generale sussiste la definizione seguente:

Definizione 4.7. (Errore globale di troncamento - metodo ad un passo)

Si dice *errore globale di troncamento* di un metodo del tipo (4.15), nel punto t_i , la quantità

$$E_i(h) = y(t_i) - y_i.$$

t	$y(t)$	h	$\tilde{u}(t)$	errore
0.4	0.725	0.20	0.680	0.447×10^{-1}
		0.10	0.705	0.199×10^{-1}
		0.05	0.715	0.943×10^{-2}
1.0	0.568	0.20	0.539	0.288×10^{-1}
		0.10	0.554	0.140×10^{-1}
		0.05	0.561	0.688×10^{-2}

Tabella 4.3: Applicazione del metodo di Eulero al problema (4.16). Soluzione calcolata ed errore globale di troncamento in $t = 0.4, 1$, per $h = 0.2, 0.1, 0.05$.

l'errore globale di troncamento decresce al decrescere di h e, in particolare, al dimezzarsi di h si riduce di circa la metà.

Tale situazione si verifica in generale in ogni punto t_i della discretizzazione, come mostrato dai grafici in Figura 4.10. ♣

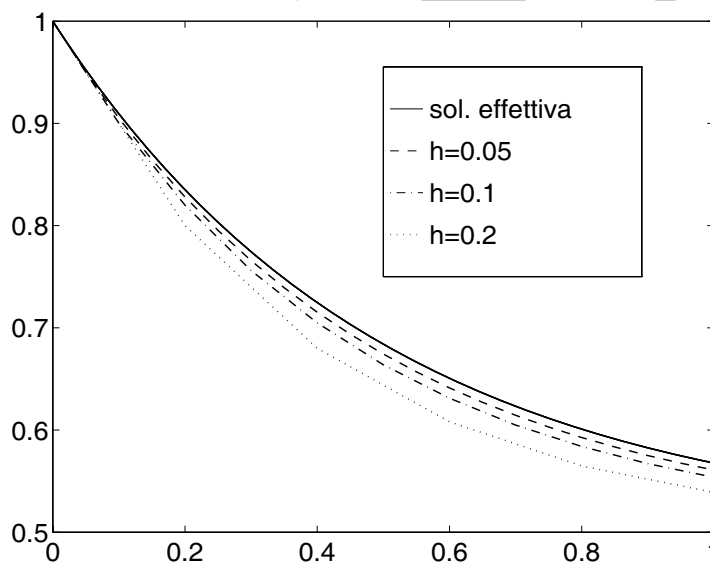


Figura 4.10: Applicazione del metodo di Eulero al problema (4.16). Soluzioni calcolate con $h = 0.2, 0.1, 0.05$ e soluzione effettiva.

Al fine di studiare l'errore globale di troncamento si premette il lemma:

Lemma 4.2. *Siano dati un intero positivo n e le costanti $K > 0$ e $H \geq 0$. Se, per le costanti M_i con $i = 0, 1, 2, \dots, n$, si ha:*

$$|M_i| \leq (1 + K)|M_{i-1}| + H, \quad \text{per } i = 1, 2, \dots, n$$

allora:

$$|M_i| \leq e^{nK} |M_0| + \frac{e^{nK} - 1}{K} H \quad \text{per } i = 1, 2, \dots, n. \quad (4.22)$$

Dimostrazione Mostriamo preliminarmente che

$$|M_i| \leq (1 + K)^i |M_0| + \frac{(1 + K)^i - 1}{K} H \quad \text{per } i = 1, 2, \dots, n. \quad (4.23)$$

Procediamo per induzione. Per $i = 1$, si ha la base d'induzione. Infatti:

$$|M_1| \leq (1 + K) |M_0| + \frac{(1 + K) - 1}{K} H = (1 + K) |M_0| + H.$$

Supponiamo la (4.23) valida per un dato $i < n$. Si ha:

$$\begin{aligned} |M_{i+1}| &\leq (1 + K) |M_i| + H \leq (1 + K) \left((1 + K)^i |M_0| + \frac{(1 + K)^i - 1}{K} H \right) + H = \\ &= (1 + K)^{i+1} |M_0| + \left\{ (1 + K) \frac{(1 + K)^i - 1}{K} + 1 \right\} H = \\ &= (1 + K)^{i+1} |M_0| + \frac{(1 + K)^{i+1} - 1}{K} H. \end{aligned}$$

Quindi la (4.23) vale anche per $i + 1$ e si ha la tesi. Dallo sviluppo in serie di MacLaurin di e^K , si ha:

$$e^K = 1 + K + \dots + \frac{K^r}{r!} + \dots > 1 + K,$$

da cui

$$(1 + K)^i < e^{iK} \leq e^{nK}. \quad (4.24)$$

Sostituendo la (4.24) nella (4.23) si ottiene la (4.22). ■

Per l'errore globale di troncamento del metodo di Eulero si ha:

Teorema 4.2. [Errore globale di troncamento del metodo di Eulero]

Si supponga che $f(t, y)$ soddisfi le ipotesi del Teorema 4.1 e sia $y \in C^2[t_0, T]$ la soluzione del problema continuo $M(P)$ (4.5). Per l'errore globale di troncamento $E_i(h)$ del metodo di Eulero, definito in (4.21), si ha:

$$|E_i(h)| \leq e^{L(T-t_0)} |E_0| + \frac{(e^{L(T-t_0)} - 1)}{L} \frac{C}{2} h, \quad (4.25)$$

dove E_0 indica l'errore nel dato iniziale e $C = \max_{[t_0, T]} |y''|$.

Dimostrazione Dalla (4.11) e dalla (4.17), si ha:

$$\begin{aligned} E_i(h) &= y(t_i) - u(t_i) = \\ &= y(t_{i-1}) - u(t_{i-1}) + h [f(t_{i-1}, y(t_{i-1})) - f(t_{i-1}, u(t_{i-1}))] + \tau(t_{i-1}, h). \end{aligned}$$

Passando ai valori assoluti e usando la (4.6) si ottiene:

$$|E_i(h)| \leq (1 + hL) |y(t_{i-1}) - u(t_{i-1})| + |\tau(t_{i-1}, h)| = (1 + hL) |E_{i-1}(h)| + |\tau(t_{i-1}, h)|,$$

ricordando la (4.20), si ha:

$$|E_i(h)| \leq (1 + hL)|E_{i-1}(h)| + \frac{C}{2}h^2.$$

Applicando il Lemma 4.2, con $K = hL$ e $H = \frac{C}{2}h^2$, si ricava:

$$|E_i(h)| \leq e^{nhL}|E_0| + \frac{(e^{nhL} - 1)}{Lh} \frac{C}{2}h^2.$$

Dalla (4.13) segue la tesi. ■

Dal Teorema 4.2 si ricava inoltre:

Corollario 4.1. *Nelle ipotesi del Teorema 4.2, supposto $E_0 = 0$, per l'errore globale di troncamento del metodo di Eulero si ha:*

$$E_i(h) = \mathcal{O}(h).$$

Dimostrazione Ponendo $E_0 = 0$ nella (4.25), si ha la tesi. ■

Dal Corollario 4.1 si evince che se si fa tendere h a 0, $E_i(h)$ tende a zero con ordine di infinitesimo 1. A tal proposito si dice che *il metodo di Eulero è convergente di ordine 1*¹⁰.

La convergenza di un metodo esprime il fatto che la soluzione del problema discreto descritto dal metodo stesso è un'approssimazione della soluzione del problema continuo. L'ordine di convergenza esprime l'accuratezza di tale approssimazione. Si osservi che tale ordine coincide con l'ordine di consistenza.

4.3.2 La risoluzione del problema discreto

L'analisi effettuata nel §4.3.1 mostra che il metodo di Eulero può fornire un'approssimazione della soluzione del problema (4.5) con un grado di accuratezza arbitrario, a

¹⁰In generale sussiste la seguente definizione:

Definizione 4.8. (Convergenza - metodo ad un passo)

Un metodo del tipo (4.15) si dice convergente se, posto $E_0 = 0$, per $t_i = t_0 + ih$ fissato risulta

$$\lim_{h \rightarrow 0} E_i(h) = 0.$$

Se, inoltre, si ha

$$|E_i(h)| \leq Kh^p,$$

con $K > 0$ costante, il metodo si dice convergente di ordine p .

Come nel caso dell'errore globale di troncamento, questa definizione non si applica solo ad i metodi ad un passo, ma ha una validità più generale.

patto che il parametro h possa essere scelto "sufficientemente piccolo". Tale analisi non tiene però conto degli errori introdotti nell'esecuzione delle operazioni aritmetiche. In effetti, si calcola, con un numero finito di passi del metodo, una soluzione numerica del problema discreto (4.17), usando un opportuno valore di h in un sistema aritmetico a precisione finita. In questo paragrafo l'attenzione è quindi posta sulla risoluzione del problema discreto definito dal metodo e sulla scelta dell'opportuno valore di h .

Stabilità

La convergenza di un metodo assicura che, per $h \rightarrow 0$, la soluzione del problema discreto converge alla soluzione del problema continuo, ma non assicura che, per un valore di h fissato, pur se scelto sufficientemente piccolo, il metodo sia in grado di calcolare una soluzione "accettabile".

♣ **Esempio 4.10.** Si consideri il problema a valori iniziali:

$$\begin{cases} y'(t) = -15y(t) & t \in [0, 0.8] \\ y(0) = 0.6 \end{cases} \quad (4.26)$$

Utilizzando la Procedura 4.1 con $h = 0.16, 0.08, 0.04$ si ottengono i risultati riportati in Figura 4.11. Si osserva che per $h = 0.16$ l'errore si amplifica in modo tale da condurre ad una soluzione che è completamente inaccettabile. ♣

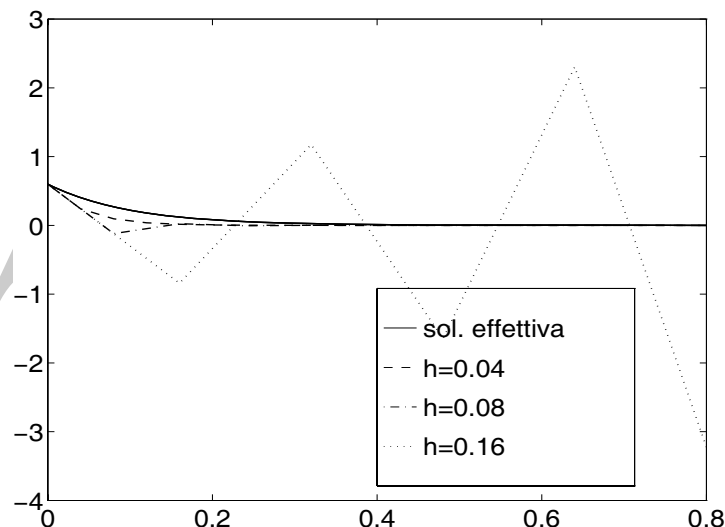


Figura 4.11: Applicazione del metodo di Eulero al problema (4.26). Soluzione calcolata per $h = 0.16, 0.08, 0.04$ e soluzione effettiva.

La relazione (4.25) mostra che l'errore globale di troncamento non è la somma degli errori locali di troncamento. Ciò è illustrato graficamente in Figura 4.12. La stima

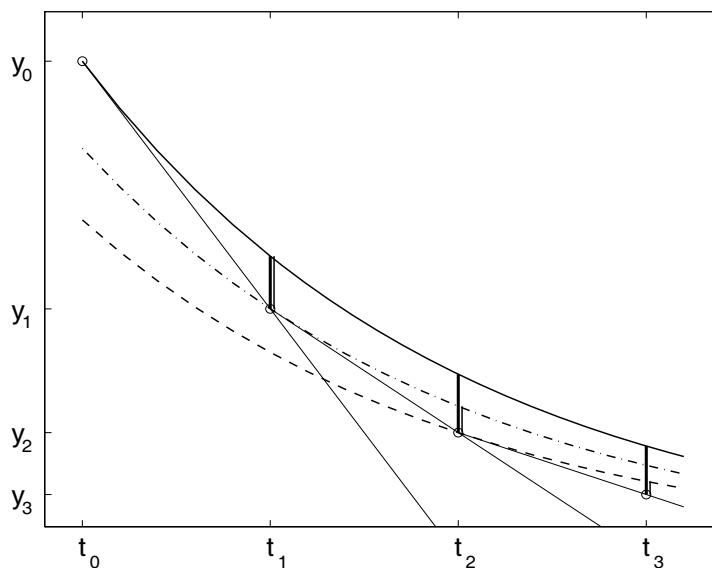


Figura 4.12: Errore globale ed errore locale di troncamento nel metodo di Eulero.

dell'errore globale di troncamento mostra che l'errore è somma di due componenti: una dovuta "all'accumulo" degli errori locali di troncamento $\tau(t_i, h)$, l'altra dovuta alla propagazione dell'errore presente nel dato iniziale. Siamo ora interessati ad indagare sulla propagazione degli errori durante l'esecuzione del metodo di Eulero. In pratica si è interessati all'analisi della stabilità del metodo.

Siano $u(t_i) = y_i$ ($i = 0, 1, 2, \dots$) la soluzione del problema discreto (4.17), e $v(t_i) = v_i$ ($i = 0, 1, 2, \dots$) la soluzione del problema discreto perturbato

$$\begin{cases} \frac{v(t_{i+1}) - v(t_i)}{h} = f(t_i, v(t_i)), & i = 0, 1, \dots \\ v(t_0) = y(t_0) + \epsilon_0 \end{cases} \quad (4.27)$$

dove ϵ_0 esprime la perturbazione introdotta sul dato iniziale $y(t_0)$. La quantità

$$\epsilon_i(h) = v(t_i) - u(t_i) = v_i - y_i$$

rappresenta quindi l'errore dovuto alla propagazione di ϵ_0 ed è detta *errore di propagazione*. Lo studio della propagazione di tale errore equivale ad indagare sulle caratteristiche di stabilità del metodo di Eulero e, a tal fine, si pongono due problemi:

- Sia $t \in [t_0, \infty[$, posto $t = t_i = t_0 + ih$, come si comporta l'errore di propagazione $\epsilon_i(h)$ per h fissato e per $i \rightarrow \infty$ (cioè per $t \rightarrow \infty$) ?
(Studio della *stabilità assoluta*)
- Sia $t \in [t_0, T]$, posto $t = t_i = t_0 + ih$, come si comporta l'errore di propagazione $\epsilon_i(h)$ per t fissato (e quindi ih fissato) e per $h \rightarrow 0$ (cioè $i \rightarrow \infty$) ?
(Studio della *zero-stabilità*)

Si osservi che, sebbene i due problemi abbiano una formulazione simile, il primo, dove il valore di h è fissato, è legato alla sola soluzione del problema discreto, mentre il secondo, dove $h \rightarrow 0$, è ovviamente rilevante ai fini della convergenza.

(a) Stabilità assoluta

Analizziamo in particolare come si propaga $\epsilon_i(h)$ nell'intervallo $[t_0, \infty[$, nel caso in cui il metodo di Eulero sia applicato al seguente problema a valori iniziali, detto *problema test*.¹¹

$$\begin{cases} y'(t) = \lambda y(t), & t \in [t_0, +\infty[\quad \lambda \in \mathbb{R} \\ y(t_0) = y_0 \end{cases} \quad (4.28)$$

Si ha:

Proposizione 4.1. *Il metodo di Eulero è assolutamente stabile¹² se e solo se*

$$(\lambda < 0, \quad h \leq -2/\lambda) \quad \text{oppure} \quad \lambda = 0.$$

Dimostrazione Essendo $f(t, y) = \lambda y$, i problemi discreti (4.17) e (4.27) assumono la forma equivalente

$$\begin{cases} y_i = (1 + h\lambda)y_{i-1}, & i = 1, 2, 3, \dots \\ y_0 = y(t_0) \end{cases} \quad (4.29)$$

e

$$\begin{cases} v_i = (1 + h\lambda)v_{i-1}, & i = 1, 2, 3, \dots \\ v_0 = y(t_0) + \epsilon_0 \end{cases}$$

¹¹La scelta del problema test come problema di riferimento è giustificata dal fatto che ci si può ricondurre localmente a tale problema, mediante una linearizzazione opportuna della funzione f in (4.5). Infatti in un intorno di un generico punto y si ha:

$$f(t, v) \simeq f(t, y) + \frac{\partial f}{\partial y}(t, y)(v - y)$$

e quindi, posto $\epsilon(t) = v(t) - y(t)$, con $v(t)$ soluzione della ODE in (4.5) corrispondente ad un dato v_0 “sufficientemente vicino” a y_0 , risulta

$$\epsilon'(t) = v'(t) - y'(t) = f(t, v(t)) - f(t, y(t)) \simeq \frac{\partial f}{\partial y}(t, y(t))\epsilon(t).$$

Supposto, per semplicità,

$$\frac{\partial f}{\partial y}(t, y) = \lambda = \text{costante},$$

si ha, a meno delle approssimazioni effettuate,

$$\epsilon'(t) = \lambda \epsilon(t),$$

ovvero si ottiene l'equazione test, della quale, come osservato precedentemente, $\epsilon_i(h)$ è la soluzione ottenuta con il metodo di Eulero.

¹²In generale sussiste la seguente definizione:

Definizione 4.9. (Stabilità assoluta - metodo ad un passo)

Siano y_i e v_i le soluzioni, determinate con un metodo del tipo (4.15), dei problemi test corrispondenti, rispettivamente, ai dati iniziali y_0 e $v_0 \neq y_0$. Il metodo considerato si dice assolutamente stabile, per un dato valore di h e per un dato valore di λ , se

$$|\epsilon_0| = |y_0 - v_0| < \delta \implies |\epsilon_i(h)| = |y_i - v_i| < \delta \quad i = 1, 2, 3, \dots$$

Quindi per l'errore di propagazione $\epsilon_i(h)$ si ha:

$$\epsilon_i(h) = v_i - y_i = (1 + h\lambda)v_{i-1} - (1 + h\lambda)y_{i-1} = (1 + h\lambda)\epsilon_{i-1}(h). \quad (4.30)$$

Applicando ripetutamente la (4.30), si ricava:

$$\epsilon_i(h) = (1 + h\lambda)^i \epsilon_0. \quad (4.31)$$

Dalla (4.31) si deduce:

$$\lim_{i \rightarrow \infty} |\epsilon_i(h)| = \begin{cases} 0 & \text{se } |1 + h\lambda| < 1 \\ |\epsilon_0| & \text{se } |1 + h\lambda| = 1 \\ +\infty & \text{se } |1 + h\lambda| > 1 \end{cases}.$$

Quindi l'errore di propagazione non si amplifica, e il metodo di Eulero è assolutamente stabile, se e solo se

$$|1 + h\lambda| \leq 1 \iff -1 \leq 1 + h\lambda \leq 1 \iff -2 \leq h\lambda \leq 0. \quad (4.32)$$

Dalla (4.32) segue la tesi. ■

La stabilità assoluta di un metodo dipende dunque dal problema attraverso λ , e dal metodo attraverso il passo di discretizzazione h , ed esprime il fatto che, fissati tali parametri, l'errore nel dato iniziale si propaga senza amplificarsi in ogni valore della soluzione calcolata. Nell'Esempio 4.10 si ha $\lambda = -15$ e $-2/\lambda \simeq 0.13$; il metodo di Eulero non è quindi assolutamente stabile per $h = 0.16$ e ciò spiega perché la soluzione calcolata con tale valore di h sia inaccettabile. Si noti infine che il metodo di Eulero non è mai assolutamente stabile se $\lambda > 0$, ovvero l'errore di propagazione cresce illimitatamente¹³.

(b) Zero-stabilità e teorema di Dahlquist

Analizziamo come si comporta l'errore di propagazione nel caso in cui si risolvono i problemi discreti (4.17) e (4.27) nell'intervallo finito $[t_0, T]$. Si ha:

Proposizione 4.2. *Il metodo di Eulero è zero-stabile qualunque sia h .*

Dimostrazione Ricordando che $u(t_i) = y_i$ e $v(t_i) = v_i$, si ha:

$$\begin{aligned} \epsilon_i(h) &= v_{i-1} + hf(t_{i-1}, v_{i-1}) - y_{i-1} - hf(t_{i-1}, y_{i-1}) = \\ &= v_{i-1} - y_{i-1} + h(f(t_{i-1}, v_{i-1}) - f(t_{i-1}, y_{i-1})). \end{aligned}$$

¹³Tuttavia se si considera l'errore relativo di propagazione, $\bar{\epsilon}_i(h) = \epsilon_i(h)/y_i(h)$, per il metodo di Eulero applicato al problema test (4.28), si ha:

$$|\bar{\epsilon}_i(h)| = \frac{|(1 + h\lambda)^i \epsilon_0|}{|(1 + h\lambda)^i y_0|} = \frac{|e_0|}{|y_0|},$$

Cioè l'errore relativo resta costante qualunque sia il passo h . E' bene però osservare che esiste una classe di problemi, di particolare interesse per le applicazioni, noti come *problemi stiff*, per cui il concetto di errore relativo non fornisce informazioni sulla bontà delle soluzioni e l'errore assoluto gioca un ruolo fondamentale

Passando ai valori assoluti e usando la (4.6) si ottiene:

$$|\epsilon_i(h)| \leq |v_{i-1} - y_{i-1}| + hL|v_{i-1} - y_{i-1}| = (1 + hL)|\epsilon_{i-1}(h)|$$

Applicando il Lemma 4.2, con $K = hL$ e $H = 0$, si ottiene:

$$|\epsilon_i(h)| \leq e^{nhL}|\epsilon_0|,$$

e ricordando la (4.13):

$$|\epsilon_i(h)| \leq e^{L(T-t_0)}|\epsilon_0|. \quad (4.33)$$

Dalla (4.33) si deduce che esiste $M = e^{(T-t_0)L} > 0$ tale che

$$|\epsilon_i(h)| \leq M|\epsilon_0|, \text{ per } h \text{ qualunque e } i = 1, 2, \dots, n, \text{ con } n = \frac{T-t_0}{h}. \quad (4.34)$$

Dalla (4.34) si ha che se ϵ_0 tende a zero anche $\epsilon_i(h)$ tende a zero, da cui se $|\epsilon_0|$ è “sufficientemente piccolo”, $|\epsilon_i(h)|$ si può mantenere al di sotto di un valore fissato ovvero si ha la zero-stabilità¹⁴. ■

La zero-stabilità è detta anche *stabilità* oppure *D-stabilità*, in onore di G.Dahlquist, che formalizzò tale concetto e lo analizzò per una classe di metodi alle differenze finite, a cui il metodo di Eulero appartiene¹⁵. Tra l'altro a Dahlquist va il merito di aver provato che la zero-stabilità è fortemente connessa alle proprietà di consistenza e di convergenza, enunciando il teorema:

Teorema 4.3. [di Dahlquist o di equivalenza di Lax-Richtmeyer]

Un metodo numerico del tipo¹⁶ (4.15) consistente è convergente se e solo se è zero-stabile.¹⁷

Il Teorema 4.3 ha una validità generale nell'ambito dei metodi numerici per la risoluzione di problemi ai valori iniziali¹⁸ e costituisce uno strumento di importanza fondamentale nello studio di tali metodi. Infatti, è spesso “più facile” verificare la consistenza e la zero-stabilità di un metodo anziché direttamente la sua convergenza.

¹⁴In generale sussiste la seguente definizione:

Definizione 4.10. (Zero-stabilità - metodo ad un passo)

Un metodo del tipo (4.15) si dice zero-stabile se esistono $h_0 > 0$ e $M > 0$ tali che

$$|\epsilon_i(h)| \leq M|\epsilon_0|, \text{ per } h < h_0 \text{ e } i = 1, 2, \dots, n \text{ con } n = \frac{T-t_0}{h}.$$

La zero-stabilità esprime il fatto che, se h è sufficientemente piccolo ($h < h_0$), l'errore di propagazione $|\epsilon_i(h)|$ si mantiene limitato, in tutto l'intervallo $[t_0, T]$.

¹⁵Per un approfondimento si vedano [4] ed i riferimenti bibliografici ivi citati.

¹⁶Vedere note 5, 8 e 10

¹⁷Il teorema è noto come *teorema di Dahlquist*, in quanto dimostrato da Dahlquist nel 1956 [3] per una particolare classe di metodi alle differenze finite, dei quali il metodo di Eulero fa parte, e come *teorema di Lax-Richtmeyer*, in quanto pubblicato da Lax e Richtmeyer nello stesso anno [7], con equazioni alle differenze finite lineari derivanti dalla discretizzazione di problemi ai valori iniziali con equazioni differenziali alle derivate parziali.

¹⁸Per una sua formulazione e dimostrazione in casi più generali si vedano, ad esempio, [2, 3, 6].

Condizionamento del problema discreto

E' noto che nella risoluzione numerica di un problema è fondamentale avere una misura del *condizionamento* del problema discreto. Analizziamo ora come si comporta l'indice di condizionamento nel metodo di Eulero. Si ha:

Proposizione 4.3. *Il metodo di Eulero è ben condizionato¹⁹ se e solo se*

$$(\lambda < 0, \quad h \leq -2/\lambda) \quad \text{oppure} \quad \lambda = 0.$$

Dimostrazione Indicata con F_i la funzione che al dato iniziale y_0 associa il valore della soluzione y_i del problema discreto (4.29), ottenuto applicando il metodo di Eulero al problema test (4.28), si ha:

$$F_i(y_0) = y_i = (1 + h\lambda)^i y_0$$

quindi l'indice di condizionamento del problema (4.29), nel punto t_i , è dato da

$$\mu_i(h, i, \lambda) = |F'_i(y_0)| = |1 + h\lambda|^i. \quad (4.35)$$

Dalla (4.35) si deduce che il metodo di Eulero è ben condizionato se e solo se

$$|\mu_i(h, i, \lambda)| \leq 1 \iff |1 + h\lambda|^i \leq 1 \iff |1 + h\lambda| \leq 1 \iff -2 \leq h\lambda \leq 0. \quad (4.36)$$

Dalla (4.36) segue la tesi. ■

Si osservi che il risultato della Proposizione 4.3 coincide con quello della Proposizione 4.1. D'altra parte, in questo caso, lo studio della stabilità assoluta del metodo di Eulero coincide con lo studio del condizionamento del problema discreto individuato dal metodo stesso. Per quanto riguarda la scelta del problema test, o meglio del problema discreto ad esso corrispondente, nell'analisi del condizionamento, valgono le stesse motivazioni addotte nel §4.3.2.

Si osservi infine che, se si considera il problema discreto (4.17) ottenuto applicando il metodo di Eulero al problema differenziale (4.5), i dati del problema sono il valore iniziale y_0 e la funzione f e quindi un'analisi completa del condizionamento deve tener conto degli effetti generati da perturbazioni su entrambi i dati.

Effetti dell'errore di roundoff

Si consideri il problema test (4.28). L'analisi effettuata nel paragrafo precedente mostra che, se $\lambda < 0$, per $h \leq -2/\lambda$ il metodo di Eulero non amplifica l'errore introdotto sul dato iniziale. Inoltre la soluzione calcolata dal metodo, in generale, costituisce un'approssimazione della soluzione del problema continuo con un errore $\mathcal{O}(h)$, che decresce linearmente al decrescere di h . Tali considerazioni possono indurre a scegliere il valore del parametro h piccolo in modo arbitrario. Tuttavia, se si eseguono le operazioni in un

¹⁹In tale analisi si fa riferimento al condizionamento assoluto del problema.

t	$y(t)$	h	$\tilde{u}(t)$	errore
0.2	0.6703200	10^{-4}	0.6702936	0.265×10^{-4}
		10^{-5}	0.6703188	0.125×10^{-5}
		10^{-6}	0.6702805	0.395×10^{-4}
0.4	0.4493290	10^{-4}	0.4492932	0.357×10^{-4}
		10^{-5}	0.4493268	0.224×10^{-5}
		10^{-6}	0.4494673	0.138×10^{-3}
0.6	0.3011942	10^{-4}	0.3011584	0.358×10^{-4}
		10^{-5}	0.3011920	0.221×10^{-5}
		10^{-6}	0.3013079	0.114×10^{-3}
0.8	0.2018965	10^{-4}	0.2018644	0.322×10^{-4}
		10^{-5}	0.2018946	0.189×10^{-5}
		10^{-6}	0.2020078	0.111×10^{-3}
1.0	0.1353353	10^{-4}	0.1353082	0.271×10^{-4}
		10^{-5}	0.1353334	0.186×10^{-5}
		10^{-6}	0.1354239	0.886×10^{-4}

Tabella 4.4: Applicazione del metodo di Eulero al problema (4.37). Soluzione calcolata $\tilde{u}(t)$ ed errore in $t = 0.2, 0.4, 0.6, 0.8, 1$ per $h = 10^{-4}, h = 10^{-5}, h = 10^{-6}$.

sistema aritmetico a precisione finita si osserva che riducendo il valore di h non sempre si ha una riduzione analoga dell'errore sulla soluzione calcolata.

♣ **Esempio 4.11. (Errore di roundoff nel metodo di Eulero)** Si consideri il problema differenziale

$$\begin{cases} y'(t) = -2y(t), & t \in [0, 1] \\ y(0) = 1 \end{cases} \quad (4.37)$$

e si applichi ad esso la Procedura 4.1 con passo $h = 10^{-4}, h = 10^{-5}, h = 10^{-6}$ (cioè con $n = 10^4, 10^5, 10^6$). In Tabella 4.4 sono riportati i valori in $t = 0.2, 0.4, 0.6, 0.8, 1$ delle soluzioni calcolate con i diversi passi di discretizzazione. Si noti che per $h = 10^{-4}$ e $h = 10^{-5}$ l'errore nella soluzione calcolata è dello stesso ordine di grandezza di h , mentre per $h = 10^{-6}$ tale errore è dell'ordine di 10^{-3} o 10^{-4} ed è maggiore degli errori corrispondenti agli altri valori di h . In altri termini, al decrescere di h l'accuratezza del metodo di Eulero diminuisce.

♣

Indichiamo con $\tilde{u}(t_i)$ il valore calcolato, in un sistema aritmetico floating-point a precisione finita, della soluzione $u(t_i)$ del problema discreto (4.17) nel punto t_i . Sussiste la seguente:

Definizione 4.11. Si dice *errore globale di round-off del metodo di Eulero*, nel punto t_i , la quantità

$$R_i(h) = u(t_i) - \tilde{u}(t_i) \quad (4.38)$$

Applicando un solo passo del metodo di Eulero e indicando con ρ_i l'errore locale di roundoff del metodo di Eulero nel punto t_i , ovvero l'errore di round-off commesso nella valutazione di f e nell'esecuzione delle operazioni aritmetiche, si ha:

$$\tilde{u}(t_i) = \tilde{u}(t_{i-1}) + hf(t_{i-1}, \tilde{u}(t_{i-1})) + \rho_i. \quad (4.39)$$

L'errore globale di round-off $R_i(h)$ risente pertanto di tutti gli errori locali di round-off $(\rho_{i-1}, \rho_{i-2}, \dots, \rho_2, \rho_1)$, commessi nei punti precedenti $(t_{i-1}, t_{i-2}, \dots, t_2, t_1)$. Per l'errore globale di round-off del metodo di Eulero si ha:

Teorema 4.4. [Errore globale di round-off del metodo di Eulero]

Si supponga che $f(t, y)$ soddisfi le ipotesi del Teorema 4.1 e sia $u(t_i)$ ($i = 1, 2, \dots, n$) la soluzione del problema discreto $M_h(P)$ (4.17). Per l'errore globale di round-off $R_i(h)$ del metodo di Eulero, definito in (4.38), si ha:

$$|R_i(h)| \leq e^{L(T-t_0)} |R_0| + \frac{(e^{L(T-t_0)} - 1)}{L} \frac{\rho}{h}, \quad (4.40)$$

dove R_0 indica l'errore di round-off nel dato iniziale, $\rho = \max_{i=1,2,\dots,n} |\rho_i|$ e ρ_i è l'errore locale di round-off nel punto t_i .

Dimostrazione Dalla (4.17) e dalla (4.39) segue

$$\begin{aligned} R_i(h) &= u(t_i) - \tilde{u}(t_i) = u(t_{i-1}) + hf(t_{i-1}, u(t_{i-1})) - [\tilde{u}(t_{i-1}) + hf(t_{i-1}, \tilde{u}(t_{i-1})) + \rho_i] = \\ &= u(t_{i-1}) - \tilde{u}(t_{i-1}) + h[f(t_{i-1}, u(t_{i-1})) - f(t_{i-1}, \tilde{u}(t_{i-1}))] - \rho_i. \end{aligned}$$

Passando ai valori assoluti e usando la (4.6), si ha:

$$|R_i(h)| \leq |u(t_{i-1}) - \tilde{u}(t_{i-1})| + hL|u(t_{i-1}) - \tilde{u}(t_{i-1})| + |\rho_i| = (1 + hL)|R_{i-1}(h)| + |\rho_i|,$$

da cui, posto $\rho = \max_{i=1,2,\dots,n} |\rho_i|$:

$$|R_i(h)| \leq (1 + hL)|R_{i-1}(h)| + \rho.$$

Applicando il Lemma 4.2, con $K = hL$ e $H = \rho$, si ricava:

$$|R_i(h)| \leq e^{nhL} |R_0| + \frac{(e^{nhL} - 1)}{Lh} \rho.$$

Dalla (4.13) segue la tesi. ■

La stima (4.40) mostra che l'errore di roundoff cresce, al decrescere del passo h . Si introduce a questo punto la seguente definizione:

Definizione 4.12. Si dice errore totale del metodo di Eulero, nel punto t_i , la quantità

$$E_i^{tot}(h) = |y(t_i) - \tilde{u}(t_i)|. \quad (4.41)$$

$E_i^{tot}(h)$ esprime l'errore totale commesso nel calcolare la soluzione del problema (4.5) con il metodo di Eulero. Si osserva che

$$E_i^{tot}(h) = |y(t_i) - u(t_i) + u(t_i) - \tilde{u}(t_i)| = |E_i(h) + R_i(h)| \leq |E_i(h)| + |R_i(h)|. \quad (4.42)$$

Per l'errore totale del metodo di Eulero si ha:

Corollario 4.2. Nelle ipotesi del Teorema 4.2 e del Teorema 4.4, per l'errore totale del metodo di Eulero, definito in (4.41), si ha:

$$E_i^{tot}(h) \leq e^{L(T-t_0)}(|E_0| + |R_0|) + \frac{(e^{L(T-t_0)} - 1)}{L} \left(\frac{C}{2}h + \frac{\rho}{h} \right). \quad (4.43)$$

Dimostrazione Sostituendo le stime (4.25) e (4.40) in (4.42), si ha la tesi. ■

In Figura 4.13 è mostrato l'andamento delle stime degli errori globali di round-off, di troncamento, e dell'errore totale in funzione del passo di discretizzazione h . Dalla figura si evince che esiste un valore ottimale del passo, h_{opt} , che rende minima la stima (4.43). Per determinare tale valore è sufficiente minimizzare, in $]0, +\infty[$, la funzione di h :

$$\psi(h) = e^{L(T-t_0)}(|E_0| + |R_0|) + \frac{(e^{L(T-t_0)} - 1)}{L} \left(\frac{C}{2}h + \frac{\rho}{h} \right).$$

Dall'espressione

$$\psi'(h) = \frac{(e^{L(T-t_0)} - 1)}{L} \left(\frac{C}{2} - \frac{\rho}{h^2} \right) = 0$$

si ricava che l'unico zero di ψ' , in $]0, +\infty[$, è $h_* = \sqrt{\frac{2\rho}{C}}$. Inoltre si ha:

$$\psi''(h) = \frac{(e^{L(T-t_0)} - 1)}{L} \frac{2\rho}{h^3} > 0 \quad (0 < h < +\infty),$$

per cui ψ è convessa in $]0, +\infty[$ ed h_* è minimo assoluto. Pertanto il valore di h_{opt} cercato è h_* . Si osserva che se $h < h_{opt}$, l'errore di roundoff diventa dominante rispetto all'errore globale di troncamento. Tuttavia il calcolo effettivo di h_{opt} richiede la conoscenza di C e di ρ , i quali non sono facilmente determinabili. In generale, la determinazione di h_{opt} non è affatto un problema di semplice risoluzione. Nella pratica sono realizzati algoritmi che scelgono dinamicamente h , passo dopo passo, tenendo conto in qualche modo dell'errore totale.

Si osservi infine che la stima suddetta è pessimistica in quanto in generale gli errori ρ_i non raggiungono il valore massimo ρ e possono compensarsi assumendo a volta valore positivo a volte valore negativo. Una stima più stringente si può ottenere con un'analisi statistica dell'errore (si vedano, ad esempio, [5, 6]).

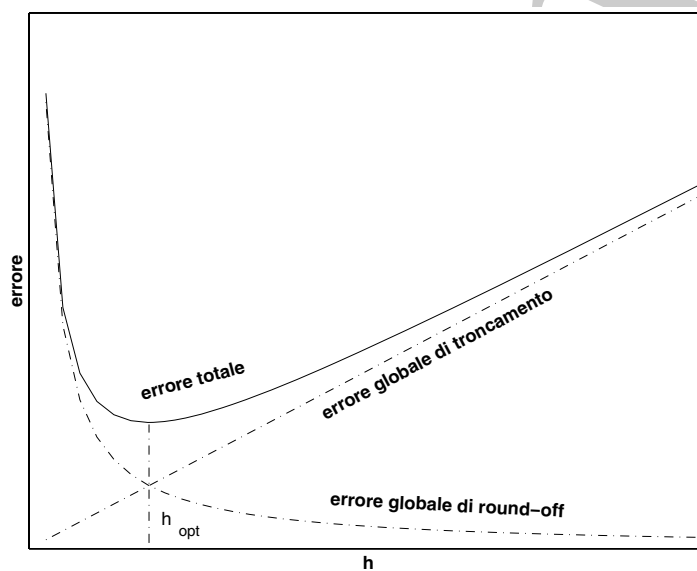


Figura 4.13: Stima dell'andamento dell'errore globale di troncamento, dell'errore globale di roundoff e dell'errore totale in funzione del passo di discretizzazione h .

4.4 Un metodo implicito: il metodo di Eulero all'indietro

Si faccia riferimento ancora al problema continuo $M(P)$ (4.5) e sia y la funzione incognita soluzione di tale problema. Considerato lo sviluppo in serie di Taylor ottenuto in (4.10), vi si ponga

$$x = t \quad x_0 - x = h$$

e quindi $x_0 = t + h$. Si ottiene:

$$y(t) = y(t + h) - hy'(t + h) + \tau(t, h) = y(t + h) - hf(t + h, y(t + h)) + \tau(t, h). \quad (4.44)$$

Da cui, trascurando il resto ed esplicitando la (4.44) rispetto ad $y(t + h)$, si ha:

$$y(t + h) \simeq y(t) + hf(t + h, y(t + h)). \quad (4.45)$$

Discretizzando l'intervallo $I = [t_0, T]$ come per il metodo di Eulero esplicito, utilizzando un passo di discretizzazione h , sostituendo t_i a t , t_{i+1} a $t + h$, indicando con y_i un'approssimazione di y nel punto t_i , si ottiene la relazione:

$$y_{i+1} = y_i + hf(t_{i+1}, y_{i+1}), \quad i = 0, 1, \dots, n-1. \quad (4.46)$$

La (4.46), unitamente al valore noto $y(t_0) = y_0$, costituisce il *metodo di Eulero all'indietro* o, più brevemente, *metodo EI*. Il metodo EI è un metodo implicito ad un passo, pertanto, ad ogni passo, la sua applicazione a partire dal punto t_i , richiede la valutazione di f in y_{i+1} che non è noto. La (4.46) equivale al calcolo di un valore di y_{i+1} che sia soluzione dell'equazione:

$$\Phi(x) = hf(t_{i+1}, x) - x + y_i = 0 \quad (4.47)$$

L'equazione (4.47) è, in generale, non lineare²⁰ e le tecniche per "risolverla" sono molteplici, la preferenza di una rispetto ad un'altra dipende dal problema.

Scelta dunque una tecnica di risoluzione per (4.47), a partire dalla relazione (4.46) si ottiene immediatamente un algoritmo per la risoluzione numerica del problema (4.5) (Procedura 4.2).

²⁰Tale equazione è lineare se e solo se lo è la funzione f rispetto alla variabile x .


```

procedure eulerol(input:  $t_0, T, y_0, f, n$ , out:  $y$ )
  /# SCOPO: Risoluzione numerica di equazioni differenziali
  /#          ordinarie del primo ordine in forma esplicita
  /# SPECIFICHE PARAMETRI:
  var:  $t_0$   : reale           {estremo sinistro dell'intervallo}
  var:  $T$     : reale           {estremo destro dell'intervallo}
  var:  $y_0$   : reale           {condizione iniziale}
  var:  $f$     : funzione esterna {funzione  $f(t, y)$ }
  var:  $n$     : intero          {numero di punti di sottointervalli}
  var:  $y$     : array di reali   {restituisce la soluzione calcolata.}
  /# INIZIO ISTRUZIONI:
     $h := (T - t_0)/n$ ;
     $y(0) := y_0$ ;
    for  $i = 0, n - 1$  do
       $t := t_0 + (i + 1) * h$ 
      call NonLinSolve( $f, t, y(i), h, y(i + 1)$ ) {Routine per la risoluzione
                                                    {dell'equazione non lineare.}
                                                    {Richiede in input  $f, t, y(i), h$ }
                                                    {Restituisce in output  $y(i + 1)$ }
    endfor
end procedure eulerol

```

Procedura 4.2: Algoritmo per la risoluzione numerica del problema (4.5) con il metodo EI

Come per la Procedura 4.1, la *complessità di tempo* della Procedura 4.2 dipende essenzialmente dal numero di valutazioni della funzione $f(t, y)$. In questo caso però, il numero di valutazioni dipende oltre che dal numero di passi n , anche da quante valutazioni sono richieste, ad ogni passo, per la risoluzione numerica dell'equazione (4.47). Se indichiamo con m tale quantità, allora la procedura ha una complessità di tempo

$$T_{EI}(n) = \mathcal{O}(n \cdot m) \text{ valutazioni di funzione.}$$

La *complessità di spazio* è ancora

$$S_{EI}(n) = \mathcal{O}(n) \text{ variabili reali,}$$

in quanto l'algoritmo utilizza un array di reali, y , di dimensione $n + 1$.

4.5 Analisi degli errori introdotti dal metodo di Eulero all'indietro

L'espressione (4.45), dalla quale si deriva il metodo EI, corrisponde alla sostituzione della derivata con il rapporto incrementale:

$$y'(t+h) \simeq \frac{y(t+h) - y(t)}{h}.$$

Sostanzialmente si sostituisce il problema continuo (4.5) con il problema discreto

$$M'_h(P) \equiv \begin{cases} \frac{u(t_{i+1}) - u(t_i)}{h} = f(t_{i+1}, u(t_{i+1})), & i = 0, 1, \dots, n-1 \\ u(t_0) = y(t_0) \end{cases} \quad (4.48)$$

Inoltre la soluzione del problema discreto (4.48) è calcolata con un sistema aritmetico floating-point a precisione finita. Quindi la risoluzione numerica del problema continuo (4.5) con il metodo EI è, in analogia al metodo di Eulero esplicito, affetta da errori dovuti alle stesse cause fondamentali:

- (1) la sostituzione del problema continuo $M(P)$ (4.5) con il problema discreto $M'_h(P)$ (4.48) (errore di discretizzazione);
- (2) la risoluzione del problema discreto $M'_h(P)$ (4.48) in un sistema aritmetico floating-point a precisione finita (errore di round-off) ²¹.

4.5.1 Il problema discreto come approssimazione del problema continuo

Come per il metodo di Eulero esplicito, l'errore $\tau(t_i, h)$, che si commette in un solo passo del metodo EI, è detto *errore locale di troncamento del metodo EI*. Analogamente sussistono le seguenti definizioni:

Definizione 4.13. Si dice *errore di discretizzazione del metodo EI* la quantità:

$$T(t, h) = \frac{\tau(t, h)}{h}, \quad (4.49)$$

cioè la differenza tra operatore differenziale del problema continuo (4.5) (derivata prima di $y(t)$) e operatore alle differenze del problema discreto (4.48) (rapporto incrementale).

²¹Nel computo totale degli errori si dovrebbe tener conto, oltre che di (1) e (2), anche dell'errore dovuto alla risoluzione dell'equazione (4.47). Tuttavia tale errore è assimilabile, ad ogni passo, all'*errore locale di troncamento* e pertanto non viene considerato.

Definizione 4.14. Si dice *errore globale di troncamento del metodo EI*, nel punto t_i , la quantità

$$E_i(h) = y(t_i) - u(t_i) , \quad (4.50)$$

cioè la differenza tra la soluzione del problema continuo $M(P)$ (4.5) e la soluzione del problema discreto $M_h(P)$ (4.48).

Consistenza

In perfetta analogia al §4.3.1, si ha:

Lemma 4.3. Se $y(t) \in C^{(2)}[t_0, T]$ è la soluzione del problema continuo $M(P)$ (4.5), per l'errore di discretizzazione del metodo EI si ha:

$$T(t, h) = \mathcal{O}(h) .$$

Dal Lemma 4.3 si evince che per ogni punto t fissato, al tendere del passo h a 0 l'errore di discretizzazione tende a 0 con ordine di infinitesimo uguale a 1. Pertanto si dice che il metodo EI è consistente di ordine 1.

Convergenza

Al fine di studiare l'errore globale di troncamento si premette il lemma:

Lemma 4.4. Se h e L sono costanti positive tali che

$$h \leq \frac{1}{2L} , \quad (4.51)$$

allora:

$$1 + 2hL \leq 2 \quad (4.52)$$

$$1 - hL > 0 \quad (4.53)$$

$$\frac{1}{1-hL} \leq 1 + 2hL \quad (4.54)$$

Dimostrazione Da (4.51) si ottiene

$$hL \leq \frac{1}{2} , \quad (4.55)$$

da cui, si ricava

$$1 + 2hL \leq 1 + 2 \cdot \frac{1}{2} , \quad (4.56)$$

ovvero la (4.52). Da (4.55) si ha

$$1 - hL \geq 1 - \frac{1}{2} = \frac{1}{2} , \quad (4.57)$$

da cui segue la (4.53). Invertendo la (4.57) si ricava

$$\frac{1}{1-hL} \leq 2 . \quad (4.58)$$

Quindi da (4.55) e (4.58):

$$\frac{hL}{1-hL} \leq \frac{1}{2} = 1. \quad (4.59)$$

Applicando la (4.59) all'identità

$$\frac{1}{1-hL} = 1 + hL \left(1 + \frac{hL}{1-hL} \right)$$

si ha la (4.54). ■

Per l'errore globale di troncamento del metodo EI si ha:

Teorema 4.5. [Errore globale di troncamento del metodo EI]

Si supponga che $f(t, y)$ soddisfi le ipotesi del Teorema 4.1 e sia $y \in C^2[t_0, T]$ la soluzione del problema continuo $M(P)$ (4.5). Se per il passo di discretizzazione vale²²

$$h \leq \frac{1}{2L},$$

con L definita da (4.6), per l'errore globale di troncamento $E_i(h)$ del metodo EI, definito in (4.50), si ha:

$$|E_i(h)| \leq e^{2L(T-t_0)} |E_0| + \frac{(e^{2L(T-t_0)} - 1)C}{L} \frac{C}{2} h, \quad (4.60)$$

dove E_0 indica l'errore nel dato iniziale e $C = \max_{[t_0, T]} |y''|$.

Dimostrazione Dalla (4.44) e dalla (4.46), si ha:

$$E_i(h) = y(t_i) - u(t_i) = y(t_{i-1}) - u(t_{i-1}) + h[f(t_i, y(t_i)) - f(t_i, u(t_i))] + \tau(t_{i-1}, h).$$

Passando ai valori assoluti e usando la (4.6) si ottiene:

$$\begin{aligned} |E_i(h)| &\leq |y(t_{i-1}) - u(t_{i-1})| + hL|y(t_i) - u(t_i)| + |\tau(t_{i-1}, h)| = \\ &= |E_{i-1}(h)| + hL|E_i(h)| + |\tau(t_{i-1}, h)|, \end{aligned}$$

Dalla continuità della derivata seconda si ha per l'errore locale di troncamento:

$$|\tau(t_{i-1}, h)| \leq \frac{C}{2} h^2$$

da cui si ha:

$$|E_i(h)|(1-hL) \leq |E_{i-1}(h)| + \frac{C}{2} h^2.$$

Per la (4.53) si ha:

$$|E_i(h)| \leq \frac{1}{1-hL} |E_{i-1}(h)| + \frac{1}{1-hL} \frac{C}{2} h^2.$$

²²Analoghi risultati si possono ottenere sotto la condizione più generale:

$$h < \frac{1}{L}.$$

Inoltre dalla (4.54) segue:

$$|E_i(h)| \leq (1 + 2hL)|E_{i-1}(h)| + (1 + 2hL)\frac{C}{2}h^2.$$

Applicando il Lemma 4.2, con $K = 2hL$ e $H = (1 + 2hL)\frac{C}{2}h^2$, si ricava:

$$|E_i(h)| \leq e^{2nhL}|E_0| + \frac{(e^{2nhL} - 1)}{2hL}(1 + 2hL)\frac{C}{2}h^2.$$

Ricordando la (4.13):

$$|E_i(h)| \leq e^{2L(T-t_0)}|E_0| + \frac{(e^{2L(T-t_0)} - 1)}{L}(1 + 2hL)\frac{C}{4}h.$$

Dalla (4.52) segue infine la tesi. ■

Dal Teorema 4.5 si ricava inoltre:

Corollario 4.3. *Nelle ipotesi del Teorema 4.5, supposto $E_0 = 0$, per l'errore globale di troncamento del metodo EI si ha:*

$$E_i(h) = \mathcal{O}(h). \quad (4.61)$$

Dimostrazione Ponendo $E_0 = 0$ nella (4.60), si ha la tesi. ■

Dal Corollario 4.3 si evince che se si fa tendere h a 0, $E_i(h)$ tende a zero con ordine di infinitesimo 1. Pertanto *il metodo EI è convergente di ordine 1*

4.5.2 La risoluzione del problema discreto

Come per il metodo di Eulero esplicito si analizzano stabilità, condizionamento ed effetti dell'errore di round-off.

Stabilità

Al fine di indagare sulla stabilità del metodo EI, in analogia al §4.3.2, indichiamo con $u(t_i) = y_i$ ($i = 1, 2, \dots$) la soluzione del problema discreto (4.48), e con $v(t_i) = v_i$ ($i = 1, 2, \dots$) la soluzione del problema discreto perturbato

$$\begin{cases} \frac{v(t_{i+1}) - v(t_i)}{h} = f(t_{i+1}, v(t_{i+1})), & i = 0, 1, \dots \\ v(t_0) = y(t_0) + \epsilon_0 \end{cases} \quad (4.62)$$

dove ϵ_0 è la perturbazione sul dato iniziale $y(t_0)$. Siamo interessati allo studio dell'errore di propagazione

$$\epsilon_i(h) = v(t_i) - u(t_i) = v_i - y_i.$$

Analogamente al metodo di Eulero esplicito, si pongono due problemi:

- (a') Sia $t \in [t_0, \infty[$, posto $t = t_i = t_0 + ih$, come si comporta l'errore di propagazione $\epsilon_i(h)$ per h fissato e per $i \rightarrow \infty$ (cioè per $t \rightarrow \infty$) ?
(Studio della *stabilità assoluta*)
- (b') Sia $t \in [t_0, T]$, posto $t = t_i = t_0 + ih$, come si comporta l'errore di propagazione $\epsilon_i(h)$ per t fissato (e quindi ih fissato) e per $h \rightarrow 0$ (cioè $i \rightarrow \infty$) ?
(Studio della *zero-stabilità*)

(a') Stabilità assoluta

Analizziamo come si propaga $\epsilon_i(h)$ nell'intervallo $[t_0, \infty[$, nel caso in cui il metodo EI è applicato al problema test (4.28). Si ha:

Proposizione 4.4. *Il metodo di EI è assolutamente stabile se e soltanto se*

$$\lambda \leq 0 \quad \text{oppure} \quad (\lambda > 0, \quad h \geq 2/\lambda) .$$

Dimostrazione Essendo $f(t, y) = \lambda y$, i problemi discreti (4.48) e (4.62) assumono la forma equivalente

$$\begin{cases} y_i = y_{i-1} + h\lambda y_i, & i = 1, 2, 3, \dots \\ y_0 = y(t_0) \end{cases} \quad (4.63)$$

e

$$\begin{cases} v_i = v_{i-1} + h\lambda v_i, & i = 1, 2, 3, \dots \\ v_0 = y(t_0) + \epsilon_0 \end{cases} .$$

Quindi per l'errore di propagazione $\epsilon_i(h)$ si ha:

$$\epsilon_i(h) = v_i - y_i = (v_{i-1} - y_{i-1}) + h\lambda(v_i - y_i) = \epsilon_{i-1}(h) + h\lambda\epsilon_i(h),$$

da cui, supposto $1 - h\lambda \neq 0$,

$$\epsilon_i(h) = \epsilon_{i-1}(h) \frac{1}{1 - h\lambda} . \quad (4.64)$$

Applicando ripetutamente la (4.64), si ricava:

$$\epsilon_i(h) = \epsilon_0 \left(\frac{1}{1 - h\lambda} \right)^i . \quad (4.65)$$

Dalla (4.65) si deduce:

$$\lim_{i \rightarrow \infty} |\epsilon_i(h)| = \begin{cases} 0 & \text{se } |1 - h\lambda| > 1 \\ |\epsilon_0| & \text{se } |1 - h\lambda| = 1 \\ +\infty & \text{se } 0 < |1 - h\lambda| < 1 \end{cases} .$$

Quindi l'errore di propagazione non si amplifica e il metodo EI è assolutamente stabile, se e solo se

$$|1 - h\lambda| \geq 1 \implies h\lambda \leq 0 \quad \text{oppure} \quad h\lambda \geq 2 \quad (4.66)$$

Dalla (4.66) segue la tesi. ■

Si osserva che se $\lambda < 0$ il metodo EI è assolutamente stabile comunque si scelga h . Se invece $\lambda > 0$, affinché ci sia assoluta stabilità, è necessario che $h > 2/\lambda$, cioè non si può scegliere h piccolo quanto si vuole ²³.

²³Si ricorda che per $\lambda > 0$ il metodo di Eulero esplicito non è assolutamente stabile per alcun valore di h .

(b') Zero-stabilità

Analizziamo come si comporta l'errore di propagazione nel caso in cui si risolvono i problemi discreti (4.48) e (4.62) nell'intervallo finito $[t_0, T]$. Si ha:

Proposizione 4.5. *Il metodo di Eulero all'indietro è zero-stabile se*

$$h \leq h_0 = 1/(2L) .$$

Dimostrazione Ricordando che $u(t_i) = y_i$ e $v(t_i) = v_i$, si ha:

$$\begin{aligned} \epsilon_i(h) &= v_{i-1} + hf(t_i, v_i) - y_{i-1} - hf(t_i, y_i) = \\ &= v_{i-1} - y_{i-1} + h(f(t_i, v_i) - f(t_i, y_i)) . \end{aligned}$$

Passando ai valori assoluti e usando la (4.6) si ottiene:

$$\begin{aligned} |\epsilon_i(h)| &\leq |y_{i-1} - v_{i-1}| + hL|v_i - y_i| = \\ &= |\epsilon_{i-1}(h)| + hL|\epsilon_i(h)| , \end{aligned}$$

da cui si ha

$$|\epsilon_i(h)|(1 - hL) \leq |\epsilon_{i-1}(h)|$$

Essendo $h \leq \frac{1}{2L}$, da (4.53) e (4.54) segue:

$$|\epsilon_i(h)| \leq |\epsilon_{i-1}(h)|(1 + 2hL)$$

Applicando il Lemma 4.2, con $K = 2hL$ e $H = 0$, si ottiene:

$$|\epsilon_i(h)| \leq e^{2nhL} |\epsilon_0| ,$$

e ricordando la (4.13):

$$|\epsilon_i(h)| \leq e^{2(T-t_0)L} |\epsilon_0| . \quad (4.67)$$

Dalla (4.67) si deduce che esistono $M = e^{2(T-t_0)L} > 0$ e $h_0 = 1/(2L)$ tali che

$$|\epsilon_i(h)| \leq M |\epsilon_0| , \text{ per } h \leq h_0 \text{ e } i = 1, 2, \dots, n, \text{ con } n = \frac{T-t_0}{h} . \quad (4.68)$$

Dalla (4.68) si ha che se ϵ_0 tende a zero anche $\epsilon_i(h)$ tende a zero, da cui se $|\epsilon_0|$ è “sufficientemente piccolo”, $|\epsilon_i(h)|$ si può mantenere al di sotto di un valore fissato ovvero si ha la zero-stabilità. ■

Si ricorda che la stima ottenuta per l'errore globale di troncamento richiede la stessa condizione sul passo h . Questa analogia è in accordo con il Teorema 4.3. Infatti essendo il metodo EI consistente, esso è zero-stabile quando e solo quando risulta convergente (cioè per $h \leq 1/(2L)$).

Condizionamento del problema discreto

Analizziamo ora come si comporta l'indice di condizionamento nel metodo EI. In analogia al §4.3.2, si ha:

Proposizione 4.6. *Il metodo EI è ben condizionato se e solo se*

$$\lambda \leq 0 \quad \text{oppure} \quad (\lambda > 0, \quad h \geq 2/\lambda) .$$

Dimostrazione Indicata con F_i la funzione che al dato iniziale y_0 associa il valore della soluzione del problema discreto (4.63), ottenuto applicando il metodo EI al problema test (4.28), si ha:

$$F_i(y_0) = y_i = \left(\frac{1}{1 - h\lambda} \right)^i y_0 ,$$

quindi l'indice di condizionamento del problema (4.63), nel punto t_i , è dato da

$$\mu_i(h, i, \lambda) = |F'_i(y_0)| = \left| \frac{1}{1 - h\lambda} \right|^i . \quad (4.69)$$

Dalla (4.69) si deduce che il metodo EI è ben condizionato se e solo se

$$|\mu_i(h, i, \lambda)| \leq 1 \iff |1 - h\lambda|^i \geq 1 \iff |1 - h\lambda| \geq 1 \iff h\lambda \leq 0 \text{ oppure } h\lambda \geq 2. \quad (4.70)$$

Dalla (4.70) segue la tesi. ■

Dalla Proposizione 4.6 si evince che il problema discreto (4.63) è ben condizionato se $h\lambda \leq 0$ oppure $h\lambda \geq 2$ e mal condizionato altrimenti (in particolare per $h\lambda \simeq 1$). Si osservi infine che $\mu_i(h, i, \lambda)$ coincide con il fattore di amplificazione dell'errore di propagazione ottenuto in (4.65). Quindi, analogamente al metodo di Eulero esplicito, lo studio del condizionamento del problema discreto (4.48) può essere ricondotto a quello dell'assoluta stabilità.

Effetti dell'errore di round-off

Si vogliono analizzare gli effetti dell'errore di round-off nell'applicazione del metodo di Eulero all'indietro in un sistema aritmetico floating-point a precisione finita.

In analogia al §4.3.2, indichiamo con $\tilde{u}(t_i)$ il valore calcolato²⁴, in un sistema aritmetico floating-point a precisione finita, della soluzione $u(t_i)$ del problema discreto (4.48) nel punto t_i . Sussiste la seguente:

Definizione 4.15. *Si dice errore globale di round-off del metodo EI, nel punto t_i , la quantità*

$$R_i(h) = u(t_i) - \tilde{u}(t_i) \quad (4.71)$$

²⁴In tale analisi si assume per semplicità di non commettere errori (di troncamento analitico) nella risoluzione dell'equazione non lineare (4.47).

Indicando con ρ_i l'errore locale di roundoff del metodo EI nel punto t_i , ovvero l'errore di round-off commesso, applicando un solo passo del metodo EI, nella valutazione di f e nell'esecuzione delle operazioni aritmetiche, si ha:

$$\tilde{u}(t_i) = \tilde{u}(t_{i-1}) + hf(t_i, \tilde{u}(t_i)) + \rho_i. \quad (4.72)$$

Per l'errore globale di round-off del metodo EI si ha:

Teorema 4.6. [Errore globale di round-off del metodo EI]

Si supponga che $f(t, y)$ soddisfi le ipotesi del Teorema 4.1 e sia $u(t_i)$ ($i = 1, 2, \dots, n$) la soluzione del problema discreto $M'_h(P)$ (4.48). Se per il passo di discretizzazione vale²²

$$h \leq \frac{1}{2L},$$

con L definita da (4.6), per l'errore globale di round-off $R_i(h)$ del metodo EI, definito in (4.71), si ha:

$$|R_i(h)| \leq e^{2L(T-t_0)} |R_0| + \frac{(e^{2L(T-t_0)} - 1)}{L} \frac{\rho}{h}, \quad (4.73)$$

dove R_0 indica l'errore di round-off nel dato iniziale, $\rho = \max_{i=1,2,\dots,n} |\rho_i|$ e ρ_i è l'errore locale di round-off nel punto t_i .

Dimostrazione Dalla (4.48) e dalla (4.72) segue

$$\begin{aligned} R_i(h) &= u(t_i) - \tilde{u}(t_i) = u(t_{i-1}) + hf(t_i, u(t_i)) - [\tilde{u}(t_{i-1}) + hf(t_i, \tilde{u}(t_i)) + \rho_i] = \\ &= u(t_{i-1}) - \tilde{u}(t_{i-1}) + h[f(t_i, u(t_i)) - f(t_i, \tilde{u}(t_i))] - \rho_i. \end{aligned}$$

Passando ai valori assoluti e usando la (4.6), si ha:

$$|R_i(h)| \leq |u(t_{i-1}) - \tilde{u}(t_{i-1})| + hL|u(t_i) - \tilde{u}(t_i)| + |\rho_i| = |R_{i-1}(h)| + hL|R_i(h)| + |\rho_i|,$$

da cui, posto $\rho = \max_{i=1,2,\dots,n} |\rho_i|$:

$$|R_i(h)|(1 - hL) \leq |R_{i-1}(h)| + \rho.$$

Per la (4.53) si ha:

$$|R_i(h)| \leq \frac{1}{1 - hL} |R_{i-1}(h)| + \frac{1}{1 - hL} \rho.$$

Inoltre dalla (4.54) segue:

$$|R_i(h)| \leq (1 + 2hL)|R_{i-1}(h)| + (1 + 2hL)\rho.$$

Applicando il Lemma 4.2, con $K = 2hL$ e $H = (1 + 2hL)\rho$, si ricava:

$$|R_i(h)| \leq e^{2nhL} |R_0| + \frac{(e^{2nhL} - 1)}{2hL} (1 + 2hL)\rho.$$

Ricordando la (4.13):

$$|R_i(h)| \leq e^{2L(T-t_0)} |R_0| + \frac{(e^{2L(T-t_0)} - 1)}{L} (1 + 2hL) \frac{\rho}{2h}.$$

Dalla (4.52) segue infine la tesi. ■

La stima (4.73) mostra che l'errore di roundoff cresce, al decrescere del passo h .

Si introduce a questo punto la seguente definizione:

Definizione 4.16. Si dice errore totale del metodo di Eulero, nel punto t_i , la quantità

$$E_i^{tot}(h) = |y(t_i) - \tilde{u}(t_i)|. \quad (4.74)$$

$E_i^{tot}(h)$ esprime l'errore totale commesso nel calcolare la soluzione del problema (4.5) con il metodo EI. Si osserva che

$$E_i^{tot}(h) = |y(t_i) - u(t_i) + u(t_i) - \tilde{u}(t_i)| = |E_i(h) + R_i(h)| \leq |E_i(h)| + |R_i(h)|. \quad (4.75)$$

Per l'errore totale del metodo EI si ha:

Corollario 4.4. Nelle ipotesi del Teorema 4.5 e del Teorema 4.6, per l'errore totale del metodo EI, definito in (4.74), si ha:

$$E_i^{tot}(h) \leq e^{2L(T-t_0)}(|E_0| + |R_0|) + \frac{(e^{2L(T-t_0)} - 1)}{L} \left(\frac{C}{2}h + \frac{\rho}{h} \right). \quad (4.76)$$

Dimostrazione Sostituendo le stime (4.60) e (4.73) in (4.75), si ha la tesi. ■

La stima (4.76), valida per $h \leq 1/(2L)$, è analoga alla (4.43). Pertanto, come per il metodo di Eulero esplicito, è possibile determinare un valore ottimale del passo di discretizzazione, h_{opt} , che minimizza nell'intervallo $]0, 1/(2L)]$, la seguente funzione

$$\psi(h) = e^{2L(T-t_0)}(|E_0| + |R_0|) + \frac{(e^{2L(T-t_0)} - 1)}{L} \left(\frac{C}{2}h + \frac{\rho}{h} \right).$$

In perfetta analogia al §4.3.2 si ricava che l'unico zero di ψ' , in $]0, +\infty[$, è $h_* = \sqrt{\frac{2\rho}{C}}$. Quindi, se $h_* \leq 1/(2L)$, allora $h_{opt} = h_*$. Altrimenti, se $h_* > 1/(2L)$, essendo ψ decrescente in $]0, 1/(2L)]$, il minimo assoluto è $h_{opt} = 1/(2L)$.

In ogni caso, indipendentemente dal valore di h_{opt} , la stima (4.76) conduce alla stessa conclusione raggiunta per il metodo di Eulero esplicito. Cioè non è possibile scegliere h piccolo in modo arbitrario, in quanto per $h < h_{opt}$ si ha una corrispondente crescita della stima dell'errore totale.

4.6 Alcuni problemi da risolvere con il calcolatore

Esercizio 1 Si consideri il problema di Cauchy:

$$\begin{cases} y' = -te^{-y} & t \in (0, 1] \\ y(0) = 0 \end{cases}$$

Lo si risolva con il metodo di Eulero esplicito, con $h = 1/100$, e si stimi il numero di cifre significative corrette della soluzione approssimata, per $t = 1$, sapendo che la soluzione esatta si mantiene limitata fra -1 e 0.

Esercizio 2 Si risolva il problema modello

$$\begin{cases} y' = \lambda y & t \in (0, +\infty) \\ y(0) = 1 \end{cases}$$

con $\lambda = -1 + i$, con il metodo di Eulero in avanti. Per quali valori di h il metodo è assolutamente stabile?

Esercizio 3 Si consideri il problema di Cauchy:

$$\begin{cases} y' = -5y \\ y(0) = 1 \end{cases}$$

Si risolva con il metodo di Eulero in avanti, con $h = 0.5$.

- (a) Il metodo è stabile?
- (b) Si calcoli la soluzione approssimata per $t = 0.5$.
- (c) Il metodo di Eulero all'indietro è stabile con $h = 0.5$?
- (d) Si calcoli la soluzione approssimata, con il metodo di Eulero all'indietro, per $t = 0.5$.

Esercizio 4 Si calcoli la soluzione per $t = 1$, del problema di Cauchy:

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

con il metodo di Eulero in avanti e con il metodo di Eulero all'indietro.

Esercizio 5 Si calcoli la soluzione dei problemi di Cauchy seguenti:

$$\begin{cases} y' = te^{3t} - 2y & t \in [0, 1] \\ y(0) = 0 \end{cases} \quad \text{con} \quad h = 0.5$$

$$\begin{cases} y' = 1 + (t - y)^2 & t \in [2, 3] \\ y(2) = 1 \end{cases} \quad \text{con} \quad h = 0.5$$

$$\begin{cases} y' = 1 + \frac{y}{t} & t \in [1, 2] \\ y(1) = 2 \end{cases} \quad \text{con} \quad h = 0.25$$

$$\begin{cases} y' = \cos(2t) + \sin(3t) & t \in [0, 1] \\ y(0) = 1 \end{cases} \quad \text{con} \quad h = 0.25$$

Esercizio 6 Si consideri il problema di Cauchy:

$$\begin{cases} y' = -y + t + 1 & t \in [0, 5] \\ y(0) = 1 \end{cases}$$

la cui soluzione esatta è $y(t) = e^{-t} + t$.

- (a) Calcolare $y(5)$, utilizzando il metodo di Eulero in avanti, con $h = 0.2$, $h = 0.1$, $h = 0.05$.
- (b) Determinare il valore ottimale di h nel calcolo di $y(5)$, che minimizzi l'errore globale.

Bibliografia

- [1] Ascher U.M., Petzold L.R. - *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations* - SIAM, 1998.
- [2] Butcher J.C. - *The Numerical Analysis of Ordinary Differential Equations, Runge-Kutta and General Linear Methods* - John Wiley & Sons, 1987.
- [3] Dahlquist G. - *Convergence and Stability in the Numerical Integration of Ordinary Differential Equations*.
- [4] Dahlquist G. - *33 Years of Numerical Instability, Part I* - BIT 25 (1985), pp. 188-204.
- [5] Gear C.W. - *Numerical initial Value problems in Ordinary Differential Equations* - Prentice-Hall, 1971.
- [6] Henrici P. - *Discrete Variable Methods in ODE* - John Wiley & Sons, 1962.
- [7] Lax P.D., Richtmeyer R.D. - *Survey of the Stability of Linear Finite Difference Equations* - Communications on Pure and Applied Mathematics, vol. IX (1956), pp. 267-293.
- [8] Schiesser W.E. - *The Numerical Method of Lines* - Academic Press, 1991.