Capitolo 8

Metodi numerici per equazioni differenziali ordinarie

8.1 Introduzione

Il problema trattato nel presente capitolo riguarda l'approssimazione numerica di una funzione $y(t) : [a, b] \subseteq \mathbb{R} \rightarrow \mathbb{R}^m$, soluzione del seguente *problema di valori iniziali*, o *di Cauchy*, del primo ordine

$$\begin{aligned} y'(t) &= f(t, y(t)), \quad t \in [a, b], \\ y(t_0) &= y_0 \, , \end{aligned} \tag{8.1}$$

dove $f(t, y) : [a, b] \times \mathbb{R}^m \rightarrow \mathbb{R}^m$, $t_0 \in [a, b]$ e $y_0 \in \mathbb{R}^m$.

Prima di presentare alcune delle principali tecniche di approssimazione nel discreto, giova premettere qualche considerazione teorica sul problema continuo (8.1).

Teorema 8.1.1 *Sia $f(t, y)$ definita e continua nell'insieme*

$$D = \{(t, y) \mid -\infty < a \leq t \leq b < +\infty, \, \|y\| < +\infty\}$$

ed esista una costante L tale che

$$\|f(t, y) - f(t, y^*)\| \leq L\|y - y^*\| \tag{8.2}$$

per ogni $(t, y), (t, y^*) \in D$. Allora esiste un'unica soluzione $y(t) \in C^1([a, b])$ del problema (8.1), per ogni y_0 assegnato.

La costante L e la relazione (8.2) diconsi rispettivamente *costante* e *condizione di Lipschitz*.

Il precedente teorema vale sotto l'ipotesi, più restrittiva, che esista e sia continua la matrice jacobiana $J(t,y) : D \to \mathbb{R}^{m\times m}$ di f, definita da

$$J(t,y) = \frac{\partial f}{\partial y} \ :$$

in tal caso nella (8.2) può assumersi $L = \sup_{(t,y)\in D} \|\partial f/\partial y\|$.

Talvolta il problema di Cauchy è dato nella *forma autonoma*

$$\begin{aligned} y'(t) &= f(y), \\ y(t_0) &= y_0 \ . \end{aligned} \tag{8.3}$$

Ciò non lede la generalità in quanto ogni problema (8.1) può ricondursi alla forma (8.3) con la sostituzione $z_1(t) = y(t)$, $z_2(t) = t$, aggiungendo l'equazione $z_2'(t) = 1$ e completando le condizioni iniziali con $z_2(t_0) = t_0$; ponendo $z^T = (z_1^T, z_2)$, $g^T = (f(z)^T, 1)$, $z_0^T = (y_0^T, t_0)$ risulta infatti

$$\begin{aligned} z'(t) &= g(z) \ , \\ z(t_0) &= z_0 \ , \end{aligned}$$

che è, appunto, della forma (8.3).

Problemi differenziali di ordine superiore al primo possono essere trasformati in problemi equivalenti del primo ordine con una opportuna sostituzione. Si consideri, infatti, il problema di ordine $r > 1$

$$\begin{aligned} &y^{(r)} = f(t, y, y', \ldots, y^{(r-1)}) \ , \\ &y(t_0) = \eta_1 \ , \\ &y'(t_0) = \eta_2 \ , \\ &\qquad \vdots \\ &y^{(r-1)}(t_0) = \eta_r \ ; \end{aligned}$$

se si introduce il vettore ausiliario $z \in \mathbb{R}^{mr}$, $z^T = (z_1^T, z_2^T, \ldots, z_r^T)$, definito da

$$\begin{aligned} z_1 &= y \,, \\ z_2 &= z_1' = y' \,, \\ z_3 &= z_2' = y'' \,, \\ &\vdots \\ z_r &= z_{r-1}' = y^{(r-1)} \,, \end{aligned}$$

e si pone $z_r' = y^{(r)} = f(t, z_1, z_2, \ldots, z_r)$ e $z(t_0)^T = (\eta_1^T, \eta_2^T, \ldots, \eta_r^T) = z_0^T$, il precedente problema può scriversi come

$$\begin{aligned} z'(t) &= g(t, z) \,, \\ z(t_0) &= z_0 \,, \end{aligned}$$

con $g^T = (z_2^T, z_3^T, \ldots, z_r^T, f(t, z)^T)$, che è della forma (8.1).

Il problema (8.1) si dice *lineare* se è $f(t, y(t)) = K(t)y(t) + \alpha(t)$ con $K(t) : [a, b] \to \mathbb{R}^{m \times m}$ e $\alpha(t) : [a, b] \to \mathbb{R}^m$; si dice *lineare a coefficienti costanti* se K non dipende da t: con ciò il problema assume la forma

$$\begin{aligned} y'(t) &= Ky(t) + \alpha(t) \,, \\ y(t_0) &= y_0 \,. \end{aligned} \tag{8.4}$$

Nel caso importante che K sia diagonalizzabile, la soluzione generale di (8.4) è

$$y(t) = \sum_{i=1}^{m} d_i x^{(i)} e^{\lambda_i t} + \beta(t) \tag{8.5}$$

dove λ_i ed $x^{(i)}$, $i = 1, 2, \ldots, m$, sono rispettivamente gli autovalori e i corrispondenti autovettori di K, $\beta(t)$ è una soluzione particolare di (8.4) e le d_i, $i = 1, 2, \ldots, m$, sono costanti arbitrarie.

Il problema (8.4) si dice *omogeneo* se $\alpha(t) = 0$: nell'ipotesi fatta su K, la sua soluzione generale è la (8.5) con $\beta(t)$ identicamente nulla.

Nei metodi numerici considerati nel seguito, si farà riferimento ad un sottoinsieme discreto dell'intervallo $[a, b]$, $t_0 = a < t_1 < \cdots < t_N = b$, ottenuto con una progressione aritmetica di ragione $h > 0$, definita da $t_n = t_0 + nh$, $n = 0, 1, 2, \ldots, N$. La ragione h si dice *passo della discretizzazione.*

In corrispondenza ad una data discretizzazione, si indica con y_n una approssimazione di $y(t_n)$ ottenuta con un metodo numerico specifico in assenza di errori di arrotondamento, mentre, in presenza di tali errori (dovuti, in genere, ad una macchina da calcolo), l'approssimazione di $y(t_n)$ è indicata con $\tilde{y}_n$.

8.2 Metodi a un passo

8.2.1 Generalità

I *metodi a un passo* sono della forma generale

$$y_{n+1} = y_n + h\phi(h, t_n, y_n), \quad n = 0, 1, \ldots, \tag{8.6}$$

in cui la funzione ϕ dipende dalla funzione f del problema (8.1).

Posto $y_0 = y(t_0)$, la (8.6) serve a calcolare y_{n+1} conoscendo y_n.

Se ϕ dipende anche da y_{n+1} il metodo si dice *implicito* e, se ϕ non è lineare, y_{n+1} si calcola con una tecnica iterativa (cfr. (8.36)). Se ϕ non dipende da y_{n+1} il metodo è detto *esplicito* e la sua applicazione è immediata.

Due semplici metodi sono i seguenti:

la *formula trapezoidale*

$$y_{n+1} = y_n + \frac{h}{2}\left[f(t_n, y_n) + f(t_{n+1}, y_{n+1})\right] \quad \text{(implicito)};$$

il *metodo di Eulero*

$$y_{n+1} = y_n + hf(t_n, y_n) \quad \text{(esplicito)}.$$

Si introducono ora alcune definizioni.

Definizione 8.2.1 *Si dice* errore globale di discretizzazione *nel punto t_{n+1}, la differenza $e_{n+1} = y(t_{n+1}) - y_{n+1}$.*

Definizione 8.2.2 *Dicesi* errore locale di troncamento *la quantità τ_{n+1} definita da*

$$\tau_{n+1} = y(t_{n+1}) - u_{n+1} \tag{8.7}$$

dove

$$u_{n+1} = y(t_n) + h\phi(h, t_n, y(t_n))\,.$$

L'errore τ_{n+1} è quindi l'errore introdotto dal metodo al passo da t_n a t_{n+1} ed è uguale alla differenza fra il valore esatto $y(t_{n+1})$ e quello teorico u_{n+1} che si ottiene usando il metodo (8.6) col valore esatto $y(t_n)$ al posto di y_n.

Si consideri un passaggio al limite facendo tendere h a zero e n all'infinito in modo che $t_0 + nh$ resti fisso e si indichi tale operazione con $\lim_{\substack{h\to 0 \\ t=t_n}}$.

Definizione 8.2.3 *Un metodo (8.6) si dice* coerente *se vale la condizione*

$$\lim_{\substack{h\to 0 \\ t=t_{n+1}}} \frac{\tau_{n+1}}{h} = 0 \,.$$

Dalla (8.7) segue che la coerenza implica

$$\phi(0, t, y(t)) = f(t, y(t)) \,. \tag{8.8}$$

Si definisce *ordine del metodo* il più grande intero positivo p per cui risulta

$$\tau_{n+1} = O\left(h^{p+1}\right) \,.$$

Si noti che un metodo coerente ha ordine almeno $p = 1$ e che, in linea di principio, l'accuratezza del metodo cresce al crescere di p.

Definizione 8.2.4 *Un metodo si dice* convergente *se, applicato a un qualunque problema di Cauchy soddisfacente le ipotesi del Teorema 8.1.1, risulta, per ogni* $t \in [a, b]$,

$$\lim_{\substack{h\to 0 \\ t=t_{n+1}}} y_{n+1} = y(t_{n+1}) \,,$$

e quindi

$$\lim_{\substack{h\to 0 \\ t=t_{n+1}}} e_{n+1} = 0 \,.$$

Occorre considerare, infine, gli errori che nascono dagli arrotondamenti.

La differenza $\tilde{e}_{n+1} = y_{n+1} - \tilde{y}_{n+1}$ dicesi *errore globale di arrotondamento*. Tale errore è originato dagli errori di arrotondamento introdotti dalla macchina ad ogni passo. Questi si suppongono di solito indipendenti da h e quindi, per un fissato intervallo, il loro contributo cresce con il numero dei passi.

Si definisce *errore totale* $\hat{e}_{n+1} = y(t_{n+1}) - \tilde{y}_{n+1}$ l'errore che si accumula al passo t_{n+1} per l'effetto degli errori che si sono prodotti in ognuno dei passi precedenti.

Il seguente teorema stabilisce condizioni necessarie e sufficienti perché un metodo esplicito (8.6) sia convergente.

Teorema 8.2.1 *La funzione $\phi(h,t,y)$ sia continua nella regione*

$$\mathcal{D} = \{(h,t,y) \mid 0 < h \leq h_0,\ -\infty < a \leq t \leq b < +\infty\ ,\ \|y\| < +\infty\}$$

e inoltre soddisfi la seguente condizione di Lipschitz

$$\|\phi(h,t,y^*) - \phi(h,t,y)\| \leq M\|y^* - y\|$$

per ogni $(h,t,y^*), (h,t,y) \in \mathcal{D}$, allora il metodo (8.6) è convergente se e solo se è coerente.

Per i metodi considerati in 8.2.2 e 8.2.3 la funzione ϕ verifica le condizioni del Teorema 8.2.1 se la $f(t,y)$ soddisfa quelle del Teorema 8.1.1.

8.2.2 Metodi di Runge-Kutta

I *metodi di Runge-Kutta* costituiscono una importante classe di metodi della forma (8.6). La struttura generale di tali metodi è

$$y_{n+1} = y_n + h\sum_{i=1}^{s} b_i k_i \tag{8.9}$$

dove

$$k_i = f(t_n + c_i h, y_n + h\sum_{j=1}^{s} a_{ij}k_j), \quad i = 1,2,\ldots,s. \tag{8.10}$$

I parametri reali b_i, c_i, a_{ij} definiscono il metodo ed s è detto *numero di stadi*.

Pertanto un metodo di Runge-Kutta è un metodo a un passo in cui risulta

$$\phi(h,t_n,y_n) = \sum_{i=1}^{s} b_i k_i\ . \tag{8.11}$$

Si osservi che, dalla convergenza del metodo, segue

$$\lim_{\substack{h\to 0 \\ t=t_{n+1}}} k_i = f(t,y(t)), \quad i = 1,2,\ldots,s,$$

per cui, tenuto conto della (8.11), la condizione di coerenza (8.8) equivale a

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

Per convenzione, comunemente adottata, si pone

$$c_i = \sum_{j=1}^{s} a_{ij}, \quad i = 1, 2, \ldots, s\,. \tag{8.12}$$

Si ottiene una utile rappresentazione compatta di un metodo di Runge-Kutta per mezzo della seguente *tavola di Butcher*

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array} = \begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \cdots & \cdots & \cdots & \cdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array}\,.$$

Si distinguono due classi di metodi, riconoscibili dalla forma della matrice A:

metodi espliciti se $a_{ij} = 0$, per ogni coppia i, j con $1 \le i \le j \le s$;

metodi impliciti se $a_{ij} \neq 0$ per qualche coppia i, j con $i \le j$.

Nel primo caso le (8.10) hanno la forma

$$k_i = f(t_n + c_i h, y_n + h \sum_{j=1}^{i-1} a_{ij} k_j), \quad i = 1, 2, \ldots, s,$$

e quindi ciascun vettore k_i si può calcolare esplicitamente in funzione dei precedenti k_j, $j = 1, 2, \ldots, i-1$.

Nel secondo caso, introducendo il vettore $k^T = (k_1^T, \ldots, k_s^T) \in \mathbb{R}^{ms}$, le (8.10) si possono scrivere in forma implicita

$$k = \varphi(k) \tag{8.13}$$

con $\varphi_i(k) = f(t_n + c_i h, y_n + h \sum_{j=1}^{s} a_{ij} k_j)$, $i = 1, 2, \ldots, s$.

Il sistema (8.13) è lineare o non lineare a seconda che lo sia $f(t, y)$ rispetto a y e può essere risolto con un algoritmo specifico.

Una sottoclasse dei metodi impliciti è costituita dai *metodi semi-impliciti* nei quali è $a_{ij} = 0$, $1 \le i < j \le s$, e $a_{ii} \neq 0$ per almeno un i. In tal caso si ha

$$k_i = f(t_n + c_i h, y_n + h \sum_{j=1}^{i-1} a_{ij} k_j + h a_{ii} k_i), \quad i = 1, 2, \ldots s,$$

e quindi le (8.10) possono essere risolte singolarmente con un costo computazionale più contenuto di quello richiesto per il sistema (8.13).

L'errore locale di troncamento di un metodo di Runge-Kutta di ordine p è della forma

$$\tau_{n+1} = h^{p+1}\psi$$

dove ψ è una funzione dipendente in modo non semplice da y_n, c, b e dagli elementi di A.

Di seguito si riportano alcuni metodi espliciti, con $s \leq 4$, scelti fra i più noti.

Metodo di Eulero, $p = 1$:

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}. \tag{8.14}$$

Metodo di Eulero modificato, $p = 2$:

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

Metodo della poligonale, $p = 2$:

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

Formula di Heun, $p = 3$:

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

Formula di Kutta, $p = 3$:

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 1 & -1 & 2 & 0 \\ \hline & 1/6 & 2/3 & 1/6 \end{array}.$$

Metodo di Runge-Kutta classico, $p = 4$:

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array} . \tag{8.15}$$

In generale i metodi impliciti sono classificati in base al tipo di formula di quadratura a cui danno luogo allorché vengono applicati al problema $y' = f(t)$. Si ha in questo caso

$$y_{n+1} - y_n = h \sum_{i=1}^{s} b_i f(t_n + c_i h) .$$

Il secondo membro può intendersi come una formula di quadratura, con pesi b_i e nodi c_i, che approssima l'integrale

$$\int_{t_n}^{t_{n+1}} f(t)dt ;$$

infatti si ha

$$y_{n+1} - y_n \simeq y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} f(t)dt = h \int_0^1 f(t_n + ch)dc .$$

Di seguito si riportano alcuni esempi con $s \leq 3$.

Metodi di Gauss-Legendre, $p = 2s$:

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

$$\begin{array}{c|cc} (3-\sqrt{3})/6 & 1/4 & (3-2\sqrt{3})/12 \\ (3+\sqrt{3})/6 & (3+2\sqrt{3})/12 & 1/4 \\ \hline & 1/2 & 1/2 \end{array} .$$

Metodi di Radau IA, $p = 2s - 1$:

$$\begin{array}{c|c} 0 & 1 \\ \hline & 1 \end{array} , \tag{8.16}$$

$$\begin{array}{c|cc} 0 & 1/4 & -1/4 \\ 2/3 & 1/4 & 5/12 \\ \hline & 1/4 & 3/4 \end{array}.$$

Metodi di Radau IIA, $p = 2s - 1$:

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

$$\begin{array}{c|cc} 1/3 & 5/12 & -1/12 \\ 1 & 3/4 & 1/4 \\ \hline & 3/4 & 1/4 \end{array}.$$

Metodi di Lobatto IIIA, $p = 2s - 2$:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}, \tag{8.17}$$

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/2 & 5/24 & 1/3 & -1/24 \\ 1 & 1/6 & 2/3 & 1/6 \\ \hline & 1/6 & 2/3 & 1/6 \end{array}.$$

Metodi di Lobatto IIIB, $p = 2s - 2$:

$$\begin{array}{c|cc} 0 & 1/2 & 0 \\ 1 & 1/2 & 0 \\ \hline & 1/2 & 1/2 \end{array}, \tag{8.18}$$

$$\begin{array}{c|ccc} 0 & 1/6 & -1/6 & 0 \\ 1/2 & 1/6 & 1/3 & 0 \\ 1 & 1/6 & 5/6 & 0 \\ \hline & 1/6 & 2/3 & 1/6 \end{array}.$$

Metodi di Lobatto IIIC, $p = 2s - 2$:

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

$$\begin{array}{c|ccc} 0 & 1/6 & -1/3 & 1/6 \\ 1/2 & 1/6 & 5/12 & -1/12 \\ 1 & 1/6 & 2/3 & 1/6 \\ \hline & 1/6 & 2/3 & 1/6 \end{array} .$$

Si osservi che i metodi (8.16) e (8.18) sono un esempio di eccezione alla condizione (8.12).

Esempi di metodi semi-impliciti sono il metodo (8.17), già riportato in 8.2.1 come "formula trapezoidale", e il metodo (8.18).

È da rilevare che i metodi impliciti hanno, in genere, a parità di numero di stadi, un ordine superiore a quello dei metodi espliciti. Inoltre per i metodi di Gauss-Legendre, Radau e Lobatto l'ordine p dipende in modo semplice e diretto da s. Un risultato di questo tipo non è disponibile per i metodi espliciti per i quali, invece, si hanno teoremi che stabiliscono per ogni ordine fissato p il minimo numero s degli stadi (o, per contro, per ogni s fissato, il massimo ordine ottenibile). Al riguardo la situazione desumibile dalla letteratura attuale è riassunta nella tavola seguente.

p	1	2	3	4	5	6	7	8	9	10
s minimo	1	2	3	4	6	7	9	11	$12 \leq s \leq 17$	$13 \leq s \leq 17$

Tavola 8.1: Corrispondenza tra ordine e numero minimo di stadi.

8.2.3 Stabilità dei metodi di Runge-Kutta

Si applichi un metodo di Runge-Kutta, nel caso scalare $m = 1$, al seguente *problema test*

$$y'(t) = \lambda y(t), \quad \lambda \in \mathbb{C}\,,\ Re(\lambda) < 0, \tag{8.19}$$

la cui soluzione è $y(t) = de^{\lambda t}$, con d costante arbitraria.

Poiché $\lim_{t\to\infty} y(t) = 0$, è naturale richiedere che la soluzione numerica abbia un comportamento analogo a quello della soluzione continua, cioè che sia, per ogni n,

$$| \, y_{n+1} \, | \leq c \, | \, y_n \, | \tag{8.20}$$

per qualche costante positiva $c < 1$.
Tale relazione implica che, per il passo h prescelto, gli errori di discretizzazione si mantengono limitati col procedere dei calcoli.

Questa proprietà riguarda la sensibilità agli errori di uno schema discreto, ovvero la sua *stabilità numerica*.

Per giungere ad una condizione che garantisca la (8.20) si faccia riferimento, ad esempio, al più elementare dei metodi di Runge-Kutta, il metodo di Eulero (8.14), il quale può scriversi

$$y_{n+1} = y_n + hf(t_n, y_n) \, . \tag{8.21}$$

Questo metodo, applicato al problema (8.19), fornisce

$$y_{n+1} = (q+1)y_n$$

dove si è posto

$$q = h\lambda \, .$$

Pertanto la (8.20) sarà verificata se e solo se

$$| \, q+1 \, | < 1 \, ,$$

ovvero se i valori del parametro q, nel piano complesso, sono interni al cerchio di centro $[-1, 0]$ e raggio unitario.

Generalizzando quanto sopra esposto per il metodo esplicito (8.21) al caso di un metodo di Runge-Kutta qualunque, cioè applicando un metodo (8.9)-(8.10) al problema (8.19), si ottiene un'equazione della forma

$$y_{n+1} = R(q)y_n \, . \tag{8.22}$$

$R(q)$ è detta *funzione di stabilità* ed è verificata la (8.20) se e solo se

$$| \, R(q) \, | < 1 \, . \tag{8.23}$$

Definizione 8.2.5 *Un metodo di Runge-Kutta si dice* assolutamente stabile*, per un dato q, se la sua funzione di stabilità soddisfa la condizione (8.23).*

Definizione 8.2.6 *L'insieme del piano complesso*

$$S_A = \{q \in \mathbb{C} \; | \; | \, R(q) \, | < 1\}$$

si chiama regione di assoluta stabilità *del metodo.*

Si può dimostrare che, introdotto il vettore $u = (1, 1, \ldots, 1)^T \in \mathbb{R}^s$, la funzione di stabilità di un metodo di Runge-Kutta è data da

$$R(q) = \frac{\det(I - qA + qub^T)}{\det(I - qA)} \,. \tag{8.24}$$

A titolo di verifica qui si ricavano la (8.22) e (8.24) per la formula trapezoidale

$$y_{n+1} = y_n + \frac{h}{2}\left[f(y_n) + f(y_{n+1})\right] \,:$$

applicandola al problema test (8.19) si ha

$$y_{n+1} = \frac{1 + q/2}{1 - q/2}\, y_n \,.$$

Facendo riferimento alla formula espressa nella forma (8.17), la stessa funzione di stabilità $R(q) = \frac{1+q/2}{1-q/2}$ si può ottenere direttamente dalla (8.24) ponendo

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ A = \begin{pmatrix} 0 & 0 \\ 1/2 & 1/2 \end{pmatrix},\ u = \begin{pmatrix} 1 \\ 1 \end{pmatrix},\ b = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix} \,.$$

Nella (8.24) $\det(I - qA + qub^T)$ e $\det(I - qA)$ sono polinomi in q, a coefficienti reali, di grado $\leq s$, quindi $R(q)$ è una funzione razionale (cfr. 6.5). Poiché allora $R(\bar{q}) = \overline{R(q)}$ ne viene che S_A è simmetrica rispetto all'asse reale.

Si vede subito che per la formula trapezoidale la regione S_A è l'intero semipiano $Re(q) < 0$.

Si noti che per la soluzione esatta del problema test si ha

$$y(t_{n+1}) = de^{\lambda t_{n+1}} = y(t_n)e^q \,,$$

mentre dalla (8.22), per definizione di errore locale di troncamento, si ottiene

$$y(t_{n+1}) = R(q)y(t_n) + \tau_{n+1} \,,$$

da cui, confrontando con la precedente e supponendo $\tau_{n+1} = O(h^{p+1})$, si ha

$$e^q - R(q) = O(h^{p+1}) \,. \tag{8.25}$$

Dalla (8.25) segue che, per $Re(q) > 0$ ed h sufficientemente piccolo, si ha $|\, R(q)\, | > 1$, cioè $q \notin S_A$ e quindi l'intersezione di S_A con l'asse reale è del tipo $]\alpha, 0[$ con $\alpha < 0$.

Definizione 8.2.7 *Un metodo si dice* A_0-stabile *se*

$$S_A \supseteq \{q \mid \ Re(q) < 0,\ Im(q) = 0\} \ ,$$

cioè se S_A *contiene l'intero semiasse reale negativo.*

Definizione 8.2.8 *Un metodo si dice* A-stabile *se*

$$S_A \supseteq \{q \mid \ Re(q) < 0\} \ .$$

In un metodo A-stabile, quindi, la condizione di stabilità $\mid R(q) \mid < 1$ è garantita indipendentemente dal passo h purché sia $Re(q) < 0$. Inoltre la A-stabilità implica la A_0-stabilità.

Definizione 8.2.9 *Un metodo si dice* L-stabile *se è A-stabile e se*

$$\lim_{Re(q)\to-\infty} \mid R(q) \mid = 0 \ .$$

Si noti, per esempio, che la formula trapezoidale è A-stabile, ma, essendo $\lim_{Re(q)\to-\infty} R(q) = -1$, non è L-stabile.

Poiché nei metodi espliciti la matrice A è triangolare inferiore con elementi diagonali nulli, risulta $\det(I - qA) = 1$ e pertanto $R(q)$ è un polinomio di grado compreso fra 1 e s; per questi metodi risulta $\lim_{Re(q)\to-\infty} \mid R(q) \mid = +\infty$. Resta quindi provato il seguente teorema.

Teorema 8.2.2 *Non esistono metodi di Runge-Kutta espliciti A-stabili.*

È da notare che le varie definizioni di stabilità date per il caso scalare $m = 1$, si estendono anche al caso $m > 1$ facendo riferimento al problema lineare a coefficienti costanti

$$y' = Ky \ , \tag{8.26}$$

dove si suppone che K abbia autovalori $\lambda_1, \lambda_2, \ldots, \lambda_m$ distinti e che sia

$$Re(\lambda_i) < 0 \ , \quad i = 1, 2, \ldots, m. \tag{8.27}$$

Posto

$$q_i = h\lambda_i \ , \quad i = 1, 2, \ldots, m,$$

l'equivalenza del sistema (8.26) ad un sistema di m equazioni indipendenti $z_i' = \lambda_i z_i$, $i = 1, 2, \ldots, m$, che si ottiene diagonalizzando la matrice K, giustifica la seguente definizione.

Definizione 8.2.10 *Un metodo di Runge-Kutta, applicato al problema (8.26), si dice assolutamente stabile per un insieme di valori* $q_i \in \mathbb{C}$, $i = 1, 2, \ldots, m$, *se*

$$q_i \in S_A \,, \quad i = 1, 2, \ldots, m, \tag{8.28}$$

dove S_A *è la regione di assoluta stabilità definita nel caso scalare (cfr. Definizione 8.2.6).*

Si può dimostare che nei metodi espliciti con $p = s$, cioè (cfr. Tavola 8.1) per $s = 1, 2, 3, 4$, si ha

$$R(q) = 1 + q + \frac{1}{2!}q^2 + \cdots + \frac{1}{s!}q^s \,. \tag{8.29}$$

Nella Tavola 8.2 si riportano il valore minimo di $Re(q)$ e il valore massimo di $\mid Im(q) \mid$ per le regioni di assoluta stabilità associate alla (8.29).

s	$Re(q)$	$\mid Im(q) \mid$
1	-2	1
2	-2	1.75
3	-2.5	2.4
4	-2.785	2.95

Tavola 8.2: Valori estremi di S_A per i metodi espliciti con $p = s$.

Le regioni S_A di assoluta stabilità per i metodi (8.29) con $s = 1, 2, 3, 4$ sono riportate in Fig. 8.1, e sono costituite dai punti interni alle porzioni di piano delimitate dalle curve date per ogni valore di s.

Si osservi ora che dalla (8.25) si ha $e^q \simeq R(q)$, e quindi $R(q)$ è una funzione razionale che approssima l'esponenziale e^q. Si comprende quindi come possa essere interessante confrontare le funzioni di stabilità $R(q)$ dei vari metodi di Runge-Kutta con le cosiddette *approssimazioni razionali di Padé* della funzione esponenziale. Esse sono date da

$$R_j^k(q) = \frac{P_k(q)}{Q_j(q)} \,,$$

dove

$$P_k(q) = \sum_{i=0}^{k} \frac{k!(j+k-i)!}{(k-i)!(j+k)!i!} q^i \,,$$

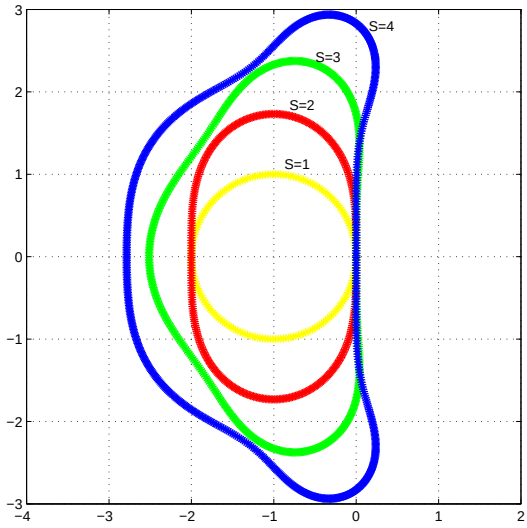

Figura 8.1: Regioni S_A dei metodi di Runge Kutta espliciti con $s = 1, 2, 3, 4$.

$$Q_j(q) = \sum_{i=0}^{j} \frac{j!(j+k-i)!}{(j-i)!(j+k)!i!}(-q)^i \, .$$

$R_j^k(q)$ è l'unica funzione razionale con numeratore di grado k e denominatore di grado j tale che

$$e^q = R_j^k(q) + O(q^{k+j+1}) \tag{8.30}$$

quando $q \to 0$. Si riportano nella Tavola 8.3 le espressioni di $R_j^k(q)$ con $0 \le k, j \le 3$.

	$k=0$	$k=1$	$k=2$	$k=3$
$j=0$	1	$1+q$	$1+q+\frac{1}{2}q^2$	$1+q+\frac{1}{2}q^2+\frac{1}{6}q^3$
$j=1$	$\frac{1}{1-q}$	$\frac{1+\frac{1}{2}q}{1-\frac{1}{2}q}$	$\frac{1+\frac{2}{3}q+\frac{1}{6}q^2}{1-\frac{1}{3}q}$	$\frac{1+\frac{3}{4}q+\frac{1}{4}q^2+\frac{1}{24}q^3}{1-\frac{1}{4}q}$
$j=2$	$\frac{1}{1-q+\frac{1}{2}q^2}$	$\frac{1+\frac{1}{3}q}{1-\frac{2}{3}q+\frac{1}{6}q^2}$	$\frac{1+\frac{1}{2}q+\frac{1}{12}q^2}{1-\frac{1}{2}q+\frac{1}{12}q^2}$	$\frac{1+\frac{3}{5}q+\frac{3}{20}q^2+\frac{1}{60}q^3}{1-\frac{2}{5}q+\frac{1}{20}q^2}$
$j=3$	$\frac{1}{1-q+\frac{1}{2}q^2-\frac{1}{6}q^3}$	$\frac{1+\frac{1}{4}q}{1-\frac{3}{4}q+\frac{1}{4}q^2-\frac{1}{24}q^3}$	$\frac{1+\frac{2}{5}q+\frac{1}{20}q^2}{1-\frac{3}{5}q+\frac{3}{20}q^2-\frac{1}{60}q^3}$	$\frac{1+\frac{1}{2}q+\frac{1}{10}q^2+\frac{1}{120}q^3}{1-\frac{1}{2}q+\frac{1}{10}q^2-\frac{1}{120}q^3}$

Tavola 8.3: Approssimazioni di Padé dell'esponenziale.

Si constata che molte approssimazioni di Padé coincidono con altrettante funzioni di stabiltà. Per esempio, dalla Tavola 8.3 si vede che le approssimazioni $R_0^k(q)$, $k = 1, 2, 3$, coincidono con le funzioni di stabilità dei metodi di Runge-Kutta espliciti con $p = s$.

Nella definizione seguente si caratterizzano le varie approssimazioni di Padé .

Definizione 8.2.11 *Una approssimazione $R_j^k(q)$ di Padé si dice:*
A_0-accettabile *se* $\mid R_j^k(q) \mid < 1$ *quando* $Re(q) < 0$ *e* $Im(q) = 0$*;*
A-accettabile *se* $\mid R_j^k(q) \mid < 1$ *quando* $Re(q) < 0$*;*
L-accettabile *se è A-accettabile e* $\lim_{Re(q)\to-\infty} \mid R_j^k(q) \mid = 0$.

Evidentemente $R_j^k(q)$ non può essere A-accettabile se $k > j$. Al riguardo si ha il seguente risultato che dimostra la cosiddetta *congettura di Ehle*.

Teorema 8.2.3 *Le approssimazioni di Padé $R_j^k(q)$ sono A-accettabili se e solo se $j - 2 \leq k \leq j$.*

Ne segue che se $j - 2 \leq k < j$, $R_j^k(q)$ sono anche L-accettabili.
Vale poi il seguente teorema.

Teorema 8.2.4 *$R_j^k(q)$ è A_0-accettabile se e solo se $k \leq j$.*

Se per un metodo risulta $R(q) = R_j^k(q)$, con $k \leq j$, allora esso sarà A_0-stabile, A-stabile oppure L-stabile a seconda che $R_j^k(q)$ sia A_0-accettabile, A-accettabile o L-accettabile.

Infine si possono dimostrare le proposizioni che seguono, le quali, per semplicità, si sono riunite in un unico teorema:

Teorema 8.2.5 *Risulta:*

$R(q) = R_s^s(q)$ per i metodi di Gauss-Legendre a s stadi;

$R(q) = R_s^{s-1}(q)$ per i metodi di Radau IA e Radau IIA a s stadi;

$R(q) = R_{s-1}^{s-1}(q)$ per i metodi di Lobatto IIIA e Lobatto IIIB a s stadi;

$R(q) = R_s^{s-2}(q)$ per i metodi di Lobatto IIIC a s stadi.

Quindi i metodi riportati da questo teorema sono A-stabili per ogni valore di s e, in particolare, i metodi di Radau IA, Radau IIA e di Lobatto IIIC sono anche L-stabili.

Una classe di problemi particolarmente importanti, per i quali sono utili i metodi A-stabili, è quella dei cosiddetti *problemi stiff*. Limitandosi, per semplicità, al caso di un problema lineare, siano λ^* e λ^{**} autovalori della matrice K di (8.4) tali che

$$| Re(\lambda^*) | = \min \{| Re(\lambda_1) |, | Re(\lambda_2) |, \ldots, | Re(\lambda_m) |\} ,$$

$$| Re(\lambda^{**}) | = \max \{| Re(\lambda_1) |, | Re(\lambda_2) |, \ldots, | Re(\lambda_m) |\} .$$

Il problema (8.4) si dice stiff se, per gli autovalori di K, vale l'ipotesi (8.27) e risulta anche:

$$| Re(\lambda^{**}) | \gg | Re(\lambda^*) | ,$$

$$| Re(\lambda^{**}) | \gg 1 .$$

In tal caso, nella soluzione (8.5), la parte $\sum_{i=1}^{m} d_i x^{(i)} e^{\lambda_i t}$, che prende il nome di *soluzione transitoria*, contiene la componente $e^{\lambda^{**} t}$ che tende a zero per $t \to \infty$, variando rapidamente in $T^{**} \equiv [t_0, t_0 + \mid \lambda^{**} \mid^{-1}]$.

È chiaro che per approssimare la componente $e^{\lambda^{**} t}$ sull'intervallo T^{**} è necessario un passo h^{**} dell'ordine di $\mid \lambda^{**} \mid^{-1}$. Un tale passo risulta troppo piccolo nei riguardi della componente $e^{\lambda^* t}$: infatti questa decresce lentamente in $T^* \equiv [t_0, t_0 + \mid \lambda^* \mid^{-1}]$ e risulta che l'ampiezza di T^* è molto maggiore di quella di T^{**}. Fuori dell'intervallo T^{**} sarebbe quindi opportuno aumentare la lunghezza del passo, compatibilmente con l'accuratezza che si vuole ottenere. Tuttavia, per i metodi con S_A limitata, il rispetto della condizione di stabilità (8.28) costringe ad usare il passo h^{**} anche su tutto l'intervallo T^*, facendo crescere in modo inaccettabile il numero dei passi necessari. Analoghe considerazioni possono farsi nei riguardi della componente $e^{\lambda^{**} t}$ confrontata con $\beta(t)$, cioè con quella parte della (8.5) che, nell'ipotesi (8.27), è detta *soluzione stazionaria*.

Al riguardo si consideri il seguente esempio.

Si vuole approssimare la soluzione del problema omogeneo (8.26) con un errore locale di troncamento dell'ordine di 10^{-4}; siano gli autovalori di K reali e $\lambda^{**} = -1000$, $\lambda^* = -0.1$; si applichi il metodo di Runge-Kutta classico (8.15), di ordine $p = 4$, il cui intervallo reale di stabilità è $]-2.785, 0[$. Il passo $h = 0.1$ è sufficiente per la precisione che si richiede, tuttavia le condizioni (8.28) sono soddisfatte solo se $q^{**} = h\lambda^{**} \in]-2.785, 0[$, ovvero se $h < 0.002785$. Assumendo $t_0 = 0$, il termine della soluzione relativo a

λ^{**} tende rapidamente a zero, mentre quello relativo a λ^* sarà prossimo a zero per un valore t_ν nettamente maggiore; sia esso $t_\nu \simeq \mid \lambda^* \mid^{-1} = 10$. Occorrono perciò $\nu = 10/h \simeq 3600$ passi per approssimare la soluzione $y(t)$ sull'intrevallo $[t_0, t_\nu]$. Poiché ad ogni passo si effettuano $s = 4$ valutazioni della funzione Ky, il costo computazionale risulta elevato e può divenire apprezzabile il fenomeno di accumulo degli errori di arrotondamento. Questi inconvenienti si possono evitare se si usa una tecnica di variazione del passo e se si utilizza un metodo A-stabile. In questo caso, poiché le (8.28) sono soddisfatte qualunque sia h, l'unico vincolo al passo è dato dal valore richiesto per l'errore locale di troncamento e a parità di ordine, con 100 passi si ottiene il risultato voluto.

8.3 Metodi a più passi

8.3.1 Equazioni alle differenze

Si consideri l'equazione

$$\sum_{j=0}^{k} \gamma_j y_{n+j} = b_n \,, \quad n = 0, 1, \ldots, \tag{8.31}$$

dove γ_j sono costanti scalari e b_n un vettore di $\mathbb{R}^m$ assegnato. La (8.31) prende il nome di *equazione lineare alle differenze a coefficienti costanti di ordine* k e la sua soluzione è una successione di vettori y_n di $\mathbb{R}^m$. Sia y_n^* una sua soluzione particolare e z_n la soluzione generale dell'equazione omogenea associata

$$\sum_{j=0}^{k} \gamma_j y_{n+j} = 0 \,, \quad n = 0, 1, \ldots, \tag{8.32}$$

allora la soluzione generale della (8.31) è data da

$$y_n = z_n + y_n^* \,.$$

Per sostituzione si trova che $z_n = d\mu^n$, $d \in \mathbb{R}^m$, è una soluzione della (8.32) se μ è radice del *polinomio caratteristico*

$$\pi(\mu) = \sum_{j=0}^{k} \gamma_j \mu^j \,.$$

Se $\pi(\mu)$ ha k radici distinte $\mu_1, \mu_2, \ldots, \mu_k$, l'insieme $\{\mu_1^n, \mu_2^n, \ldots, \mu_k^n\}$ forma un *sistema fondamentale di soluzioni* e la soluzione generale della (8.32) è

$$z_n = \sum_{i=1}^{k} d_i \mu_i^n \tag{8.33}$$

essendo d_i vettori arbitrari di $\mathbb{R}^m$. Se una radice, ad esempio μ_j, ha molteplicità ν e le rimanenti sono tutte distinte, l'insieme

$$\{\mu_1^n, \ldots, \mu_{j-1}^n, \mu_j^n, n\mu_j^n, n^2\mu_j^n, \ldots, n^{\nu-1}\mu_j^n, \mu_{j+1}^n, \ldots, \mu_{k-\nu+1}^n\}$$

forma ancora un sistema fondamentale di soluzioni e si ha

$$z_n = \sum_{i=1}^{j-1} d_i \mu_i^n + \sum_{i=j}^{j+\nu-1} d_i n^{i-j} \mu_j^n + \sum_{i=j+\nu}^{k} d_i \mu_{i-\nu+1}^n \, . \tag{8.34}$$

È immediata l'estensione al caso generale che $\pi(\mu)$ abbia r radici distinte μ_i, $i = 1, 2, \ldots, r$, ciascuna con molteplicità ν_i con $\sum_{i=1}^{r} \nu_i = k$.

Si indichi con E l'*operatore di avanzamento* definito da

$$E y_n = y_{n+1} \, ;$$

ammettendo che l'operatore E sia lineare, cioè risulti $E(\alpha y_n + \beta y_{n+1}) = \alpha E y_n + \beta E y_{n+1}$, e convenendo di porre $E^2 y_n = E(E y_n)$, la (8.31) può formalmente scriversi

$$\pi(E) y_n = b_n \, .$$

8.3.2 Metodi lineari

Un *metodo lineare a più passi*, o *a k passi*, con $k \geq 1$, per approssimare la soluzione del problema (8.1) ha la struttura seguente

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f(t_{n+j}, y_{n+j}) \, , \quad n = 0, 1, \ldots, \tag{8.35}$$

cioè ha la forma di una equazione alle differenze, lineare rispetto a y_{n+j} e $f(t_{n+j}, y_{n+j})$, $j = 0, 1, \ldots, k$.

I coefficienti α_j e β_j sono costanti reali e si ammette che sia

$$| \, \alpha_0 \, | + | \, \beta_0 \, | \neq 0 \, , \quad \alpha_k = 1 \, .$$

Per ogni n, la (8.35) fornisce il vettore y_{n+k} in funzione dei k vettori precedenti $y_{n+k-1}, y_{n+k-2}, \ldots, y_n$.
Si suppongono noti (dati o calcolati) i k *vettori iniziali* $y_0, y_1, \ldots, y_{k-1}$.

Se $\beta_k \neq 0$ il metodo si dice *implicito*: posto

$$w = -\sum_{j=0}^{k-1} \alpha_j y_{n+j} + h \sum_{j=0}^{k-1} \beta_j f(t_{n+j}, y_{n+j}) \,,$$

si calcola, per $n = 0, 1, \ldots$, una approssimazione z^* della soluzione dell'equazione

$$z = h\beta_k f(t_{n+k}, z) + w \tag{8.36}$$

e si assume quindi $y_{n+k} = z^*$. Se $f(t, y)$ è lineare rispetto a y la (8.36) si riduce ad un sistema lineare; in caso contrario si può utilizzare, ad esempio, il seguente procedimento iterativo

$$z^{(r+1)} = h\beta_k f(t_{n+k}, z^{(r)}) + w \,, \quad r = 0, 1, \ldots,$$

la cui convergenza è garantita se (cfr. Teorema 4.6.1)

$$h \mid \beta_k \mid L < 1 \,,$$

dove L è la costante di Lipschitz della funzione $f(t, z)$. Se esiste la matrice $\partial f/\partial z$ si può porre $L = \sup_{(t,z)\in D} \|\partial f/\partial z\|$.

Se $\beta_k = 0$ il metodo si dice *esplicito* e il calcolo di y_{n+k} è diretto.

Al metodo (8.35) sono associati i seguenti due polinomi, detti *primo* e *secondo polinomio caratteristico*,

$$\rho(\mu) = \sum_{j=0}^{k} \alpha_j \mu^j \,, \quad \sigma(\mu) = \sum_{j=0}^{k} \beta_j \mu^j \,.$$

Convenendo di porre

$$f_{n+j} = f(t_{n+j}, y_{n+j}) \,,$$

la (8.35) può formalmente scriversi

$$\rho(E) y_n = h\sigma(E) f_n \,.$$

Definizione 8.3.1 *Dicesi* errore locale di troncamento *di un metodo a più passi la quantità* τ_{n+k} *definita da*

$$\tau_{n+k} = \sum_{j=0}^{k} \alpha_j y(t_{n+j}) - h \sum_{j=0}^{k} \beta_j f(t_{n+j}, y(t_{n+j})) \,. \tag{8.37}$$

Questa definizione, nell'ambito dei metodi lineari, coincide con la Definizione 8.2.2 se il metodo è esplicito, mentre, se il metodo è implicito, le due definizioni forniscono errori che differiscono per termini dell'ordine di h. Analogamente si hanno poi le definizioni di coerenza, ordine e convergenza. Per un metodo a più passi la condizione di coerenza è data da

$$\lim_{\substack{h\to 0 \\ t=t_{n+k}}} \frac{\tau_{n+k}}{h} = 0 \,, \tag{8.38}$$

e l'ordine è il più grande intero p per cui risulta

$$\tau_{n+k} = O(h^{p+1}) \,.$$

Restano invariate, rispetto ai metodi a un passo, le definizioni di errore globale di discretizzazione, di errore globale di arrotondamento e di errore totale (riferite al calcolo di y_{n+k} nel punto t_{n+k}).

In generale i vettori $y_1, \ldots, y_{k-1}$ dipendono da h e si dice che formano un *insieme compatibile di vettori iniziali* se vale la proprietà

$$\lim_{h\to 0} y_i = y_0 \,, \quad i = 1, 2, \ldots, k-1.$$

Definizione 8.3.2 *Il metodo (8.35) si dice* convergente *se, applicato a un qualunque problema soddisfacente le ipotesi del Teorema 8.1.1, è tale che, per ogni* $t \in [a, b]$*, si abbia*

$$\lim_{\substack{h\to 0 \\ t=t_{n+k}}} y_{n+k} = y(t_{n+k}) \,,$$

per ogni insieme compatibile di vettori iniziali.

Come già osservato per i metodi ad un passo, anche per questi metodi, la riduzione di h su un dato intervallo può produrre un aumento dei contributi all'errore totale: questo fenomeno è dovuto all'accumulo degli errori locali di arrotondamento i quali possono ritenersi indipendenti da h.

Nell'ipotesi che $y(t) \in C^\infty(]a,b[)$, sviluppando il secondo membro della (8.37) si può scrivere formalmente

$$\tau_{n+k} = c_0 y(t_n) + c_1 y'(t_n)h + c_2 y''(t_n)h^2 + \cdots + c_r y^{(r)}(t_n)h^r + \cdots ,$$

dove le c_i, $i = 0, 1, \ldots$, sono costanti date da

$$\begin{aligned} c_0 &= \sum_{j=0}^{k} \alpha_j = \rho(1) , \\ c_1 &= \sum_{j=0}^{k} (j\alpha_j - \beta_j) = \rho'(1) - \sigma(1) , \\ c_r &= \sum_{j=0}^{k} \left(\frac{1}{r!} j^r \alpha_j - \frac{1}{(r-1)!} j^{r-1} \beta_j \right) , \ r = 2, 3, \ldots . \end{aligned} \tag{8.39}$$

Ne segue che per un metodo di ordine p deve essere $c_0 = c_1 = \cdots = c_p = 0$, $c_{p+1} \neq 0$; quindi risulta

$$\tau_{n+k} = c_{p+1} y^{(p+1)}(t_n) h^{p+1} + O(h^{p+2}) , \tag{8.40}$$

dove $c_{p+1} y^{(p+1)}(t_n) h^{p+1}$ si dice *parte principale* di τ_{n+k} mentre c_{p+1} prende il nome di *costante di errore del metodo*. Dalle (8.39) e (8.40) discende il seguente teorema.

Teorema 8.3.1 *Un metodo a più passi è coerente se e solo se*

$$\rho(1) = 0, \quad \rho'(1) = \sigma(1) . \tag{8.41}$$

L'applicazione della (8.35) al problema test

$$y' = 0 , \quad y(t_0) = y_0 , \tag{8.42}$$

la cui soluzione è $y(t) = y_0$, dà luogo all'equazione lineare omogenea alle differenze

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = 0 \tag{8.43}$$

il cui polinomio caratteristico coincide con il primo polinomio caratteristico $\rho(\mu)$ del metodo stesso. La soluzione generale y_n della (8.43) è della forma (8.33) o (8.34) a seconda che le radici $\mu_1, \mu_2, \ldots, \mu_k$ di $\rho(\mu) = 0$ siano

distinte o meno. Si può verificare che la soluzione numerica della (8.43), per un insieme compatibile di valori iniziali, converge alla soluzione del problema continuo (8.42) solo se vale la seguente *condizione delle radici*

$$\begin{gathered} |\,\mu_i\,| \leq 1\,, \quad i = 1, 2, \ldots, k, \\ \text{dove, se } |\,\mu_i\,| = 1, \text{ allora } \mu_i \text{ è semplice.} \end{gathered} \tag{8.44}$$

Si conclude che un metodo della forma (8.35) non è convergente se gli zeri del suo polinomio $\rho(\mu)$ non soddisfano la condizione (8.44).

Definizione 8.3.3 *Un metodo lineare (8.35) si dice* zero-stabile *se gli zeri del suo polinomio $\rho(\mu)$ soddisfano la condizione (8.44).*

Si può dimostrare il seguente teorema.

Teorema 8.3.2 *Un metodo lineare a più passi è convergente se e solo se è coerente e zero-stabile.*

In linea di principio è possibile costruire metodi lineari a k passi fino ad un ordine massimo $p = 2k$ determinando le $2k+1$ costanti $\alpha_0, \ldots, \alpha_{k-1}, \beta_0, \ldots, \beta_k$ in modo che, in base alle (8.39), risulti $c_0 = c_1 = \cdots = c_{2k} = 0$: tuttavia tali metodi possono risultare non zero-stabili. Sussiste infatti il seguente teorema detto *prima barriera di Dahlquist.*

Teorema 8.3.3 *Non esistono metodi lineari a k passi zero-stabili di ordine superiore a $k+1$ se k è dispari e a $k+2$ se k è pari.*

Si riportano qui alcuni dei più noti metodi lineari con relative costanti d'errore.

I *metodi di Adams* sono caratterizzati da $\rho(\mu) = \mu^k - \mu^{k-1}$. Se $\beta_k = 0$ si ha la sottoclasse dei *metodi* espliciti *di Adams-Bashforth*; per $k = 1$ si ha il metodo di Eulero (costante di errore $c_2 = 1/2$), mentre per $k = 2$ risulta

$$y_{n+2} - y_{n+1} = \frac{h}{2}(3f_{n+1} - f_n)\,, \quad c_3 = \frac{5}{12}\,.$$

Se $\beta_k \neq 0$ si ha la sottoclasse dei *metodi* impliciti *di Adams-Moulton*; per $k = 1$ si ha la formula trapezoidale (costante di errore $c_3 = -1/12$), per $k = 2$ si ottiene

$$y_{n+2} - y_{n+1} = \frac{h}{12}(5f_{n+2} + 8f_{n+1} - f_n)\,, \quad c_4 = -\frac{1}{24}\,.$$

I *metodi BDF* (Backward Differentiation Formulae) hanno $\sigma(\mu) = \beta_k \mu^k$, sono di ordine $p = k$ e zero-stabili solo per $k \leq 6$. Per esempio, per $k = 1, 2, 3$ si ha rispettivamente

$$y_{n+1} - y_n = h f_{n+1} \,, \quad c_2 = -\frac{1}{2} \,, \tag{8.45}$$

$$y_{n+2} - \frac{4}{3} y_{n+1} + \frac{1}{3} y_n = h \frac{2}{3} f_{n+2} \,, \quad c_3 = -\frac{2}{3} \,,$$

$$y_{n+3} - \frac{18}{11} y_{n+2} + \frac{9}{11} y_{n+1} - \frac{2}{11} y_n = h \frac{6}{11} f_{n+3}, c_4 = -\frac{3}{22}.$$

La (8.45) è nota come *formula di Eulero implicita*.

Nella classe dei metodi con $\rho(\mu) = \mu^k - \mu^{k-2}$ quelli aventi $\beta_k = 0$ si dicono *metodi di Nyström*; per $k = 2$ si ottiene la *formula del punto centrale*

$$y_{n+2} - y_n = 2h f_{n+1} \,, \quad c_3 = \frac{1}{3} \,. \tag{8.46}$$

Se invece $\beta_k \neq 0$ si hanno i *metodi generalizzati di Milne-Simpson*; in particolare per $k = 2$ si ha la *regola di Simpson*

$$y_{n+2} - y_n = \frac{h}{3}(f_{n+2} + 4 f_{n+1} + f_n) \,, \quad c_5 = -\frac{1}{90} \,. \tag{8.47}$$

I *metodi di Newton-Cotes* hanno $\rho(\mu) = \mu^k - 1$; con $k = 4$ e $\beta_k = 0$ si ha, per esempio,

$$y_{n+4} - y_n = \frac{4}{3} h (2 f_{n+3} - f_{n+2} + 2 f_{n+1}) \,, \quad c_5 = \frac{28}{90} \,. \tag{8.48}$$

8.3.3 Metodi a più passi: assoluta stabilità

Si è visto che la coerenza e la zero-stabilità garantiscono la convergenza di un metodo lineare (cfr. Teorema 8.3.2), ma, essendo la convergenza una proprietà limite per $h \to 0$, può accadere che un metodo convergente, usato con un fissato passo $h > 0$, anche se piccolo, produca errori globali relativamente grandi.

Ciò accade per esempio se si applica al problema test (8.19) la formula del punto centrale (8.46), che pure è coerente e zero-stabile.

Occorre pertanto definire un concetto di stabilità che garantisca non solo la convergenza del metodo per $h \to 0$ ma anche il contenimento degli errori per un dato h.

Si consideri il caso scalare $m = 1$.

Il metodo (8.35) applicato al problema test (8.19) fornisce

$$\sum_{j=0}^{k}(\alpha_j - q\beta_j)y_{n+j} = 0\,. \tag{8.49}$$

Alla (8.49) è associato il polinomio caratteristico, detto *polinomio di stabilità* ,

$$\pi(q,\mu) = \rho(\mu) - q\sigma(\mu)\,. \tag{8.50}$$

Se $\mu_1(q), \mu_2(q), \ldots, \mu_k(q)$ sono le radici di $\pi(q,\mu) = 0$, che si suppone si mantengano semplici, la soluzione generale della (8.49) è data da

$$y_n = \sum_{i=1}^{k} d_i \mu_i^n(q)\,, \quad n = 0, 1, \ldots, \tag{8.51}$$

dove d_i, $i = 1, 2, \ldots, k$, sono costanti arbitrarie.

D'altra parte, ponendo nella (8.49) il valore esatto $y(t_{n+j})$ al posto di y_{n+j}, per definizione di errore di troncamento locale (vedi la (8.37)) si ha

$$\sum_{j=0}^{k}(\alpha_j - q\beta_j)y(t_{n+j}) = \tau_{n+k}\,;$$

sottraendo membro a membro da questa equazione la (8.49) si ottiene

$$\sum_{j=0}^{k}(\alpha_j - q\beta_j)e_{n+j} = \tau_{n+k}$$

che può ritenersi un caso perturbato della (8.49). Ne segue che l'errore globale di discretizzazione $e_n = y(t_n) - y_n$ ha un andamento analogo a quello di y_n e si può quindi affermare che e_n non cresce, al crescere di n, per i valori di q per cui risulta

$$|\,\mu_i(q)\,| < 1\,, \quad i = 1, 2, \ldots, k. \tag{8.52}$$

Definizione 8.3.4 *Un metodo della forma (8.35) si dice* assolutamente stabile *per un dato $q \in \mathbb{C}$, se gli zeri del suo polinomio di stabilità $\pi(q,\mu)$ verificano la condizione (8.52).*

Definizione 8.3.5 *L'insieme del piano complesso*

$$S_A = \{q \in \mathbb{C} \;|\; |\,\mu_i(q)\,| < 1,\ i = 1, 2, \ldots, k\}$$

si chiama regione di assoluta stabilità *del metodo lineare a più passi.*

S_A è simmetrico rispetto all'asse reale: infatti se μ^* è tale che $\pi(q, \mu^*) = 0$, per la (8.50), risulta anche $\pi(\bar{q}, \bar{\mu}^*) = 0$.

Si osservi che si ha $\lim_{h\to 0} \pi(q, \mu) = \pi(0, \mu) = \rho(\mu)$; pertanto le radici caratteristiche $\mu_i(q)$, $i = 1, 2, \ldots, k$, per $h \to 0$ tendono alle radici di $\rho(\mu) = 0$. D'altra parte, una di queste radici, per le condizioni di coerenza (8.41), è uguale a 1. Quindi una delle radici $\mu_i(q)$, che si indicherà con $\mu_1(q)$, tende a 1 per $h \to 0$ e, a causa della condizione (8.44), essa è unica. A $\mu_1(q)$ si dà il nome di *radice principale* di $\pi(q, \mu) = 0$, perché è quella che nella (8.51) approssima la soluzione $y(t) = de^{\lambda t}$ del problema test.

Le altre radici $\mu_2(q), \ldots, \mu_k(q)$ si chiamano *radici spurie* o *parassite* perché nascono dalla sostituzione dell'equazione differenziale, del primo ordine, con l'equazione alle differenze (8.49), di ordine k, e il loro effetto è solo quello di accrescere l'errore.

Per la (8.37) ove si ponga $y(t) = e^{\lambda t}$ ed $f(t, y) = \lambda e^{\lambda t}$, si ottiene

$$\tau_{n+k} = e^{\lambda t_n} \pi(q, e^q) = O(h^{p+1}) .$$

D'altra parte vale l'identità

$$\pi(q, \mu) = (1 - q\beta_k)(\mu - \mu_1(q))(\mu - \mu_2(q)) \cdots (\mu - \mu_k(q));$$

ne segue quindi

$$\pi(q, e^q) = (1 - q\beta_k)(e^q - \mu_1(q))(e^q - \mu_2(q)) \cdots (e^q - \mu_k(q)) = O(h^{p+1}) .$$

Quando $q \to 0$ l'unico fattore di $\pi(q, e^q)$ che tende a zero, per quanto detto sopra, è $e^q - \mu_1(q)$: si ha pertanto

$$\mu_1(q) = e^q + O(h^{p+1}) . \tag{8.53}$$

Dalla (8.53) segue che se $Re(q) > 0$, ed h è sufficientemente piccolo, risulta $| \mu_1(q) | > 1$ e quindi $q \notin S_A$, cioè l'intersezione di S_A con l'asse reale è del tipo $]\alpha, 0[$ con $\alpha < 0$.

In Fig. 8.2 sono riportate le regioni S_A delle formule BDF con $k = 1, 2, 3$, e sono costituite dai punti esterni alle porzioni di piano delimitate dalle curve date per ogni valore di k.

Nel teorema seguente sono riuniti i risultati più significativi della teoria della stabilità per i metodi a più passi, avvertendo che anche per questi metodi le definizioni di A_0-stabilità e di A-stabilità sono le medesime date per i metodi di Runge-Kutta (cfr. Definizioni 8.2.7 e 8.2.8).

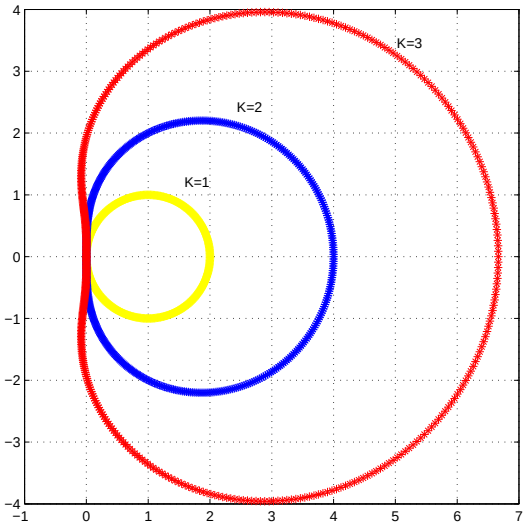

Figura 8.2: Regioni S_A delle prime tre fomule BDF.

Teorema 8.3.4 (di Dahlquist)

Non esistono metodi lineari a più passi espliciti A-stabili;

*l'ordine massimo di un metodo lineare a più passi A-stabile non può superare due (*seconda barriera di Dahlquist*);*

la formula trapezoidale è il metodo lineare A-stabile del secondo ordine con la costante di errore più piccola in modulo.

Successivamente la prima proposizione di questo teorema è stata ulteriormente precisata come segue.

Teorema 8.3.5 *Non esistono metodi lineari a più passi espliciti A_0-stabili.*

La definizione di assoluta stabilità data nel caso $m = 1$, si estende in modo naturale al caso $m > 1$ considerando il problema test (8.26): si dice ora che un metodo a più passi applicato al problema (8.26) è assolutamente stabile per un dato insieme di valori q_i, $i = 1, 2, \ldots, m$, se e solo se

$$q_i \in S_A\,, \quad i = 1, 2, \ldots, m,$$

dove S_A è la regione di assoluta stabilità introdotta con la Definizione 8.3.5.

8.3.4 Metodi di predizione e correzione

In 8.3.2 si è visto che l'uso di un metodo (8.35) di tipo implicito richiede, ad ogni passo, la risoluzione dell'equazione (8.36), che può essere non lineare. In tal caso l'incognita y_{n+k} viene approssimata con il processo iterativo

$$z^{(r+1)} = h\beta_k f(t_{n+k}, z^{(r)}) + w\ , \quad r = 0, 1, \dots\ .$$

È chiaro che, ammesso che vi sia convergenza, quanto più vicina alla soluzione si sceglie l'approssimazione iniziale $z^{(0)}$, tanto più piccolo è il numero delle iterazioni necessarie per ottenere una certa accuratezza.

Su quest'idea si basano i cosiddetti *metodi di predizione e correzione*, nei quali si sceglie $z^{(0)} = y^*_{n+k}$ dove y^*_{n+k} si calcola in precedenza con un metodo esplicito.

In questo contesto il metodo esplicito viene detto *predittore* mentre il metodo implicito prende il nome di *correttore*. In quello che segue si fa riferimento, per semplicità, al caso in cui predittore e correttore hanno lo stesso ordine; inoltre si suppone che il correttore venga usato una sola volta ad ogni passo.

In queste ipotesi, la struttura generale di un metodo di predizione e correzione è quindi data da

$$\begin{aligned} y^*_{n+k} &= -\sum_{j=0}^{k-1} \alpha^*_j y_{n+j} + h \sum_{j=0}^{k-1} \beta^*_j f_{n+j}\ , \\ y_{n+k} &= -\sum_{j=0}^{k-1} \alpha_j y_{n+j} + h \sum_{j=0}^{k-1} \beta_j f_{n+j} + h\beta_k f^*_{n+k}\ , \end{aligned} \tag{8.54}$$

dove si è posto $f^*_{n+k} = f(t_{n+k}, y^*_{n+k})$.

Si osservi che un metodo di predizione e correzione nella forma (8.54) ha un costo computazionale non superiore a quello del correttore usato da solo iterativamente.

L'algoritmo (8.54) viene designato con la sigla PEC per indicare le varie fasi che costituiscono un passo, cioè

P (Prediction): calcolo di y^*_{n+k} mediante il predittore,

E (Evaluation): valutazione del termine $f(t_{n+k}, y^*_{n+k})$,

C (Correction): calcolo di y_{n+k} mediante il correttore.

Al passo successivo si utilizza $f(t_{n+k}, y^*_{n+k})$ nel predittore per dare corso alla nuova fase P.

Una variante è costituita dall'algoritmo $PECE$ in cui compare l'ulteriore fase

E: valutazione del termine $f(t_{n+k}, y_{n+k})$;

in tal caso, nella fase P del passo successivo, il predittore utilizza la valutazione $f(t_{n+k}, y_{n+k})$ che, di solito, è più corretta della $f(t_{n+k}, y^*_{n+k})$.

Anche la stabilità di un metodo di predizione e correzione si studia mediante la sua applicazione al problema test (8.19) e, in generale, è diversa da quella del predittore e del correttore supposti usati singolarmente. Si consideri, ad esempio, il seguente metodo di predizione e correzione in cui il predittore è la formula del punto centrale (8.46) e il correttore è la formula trapezoidale

$$\begin{aligned} y^*_{n+2} &= y_n + 2hf_{n+1}\,, \\ y_{n+2} &= y_{n+1} + \frac{h}{2}(f_{n+1} + f^*_{n+2})\,. \end{aligned} \tag{8.55}$$

La formula del punto centrale non è assolutamente stabile mentre la formula trapezoidale è A-stabile.
Applicato al problema test, il metodo (8.55) diviene

$$\begin{aligned} y^*_{n+2} &= y_n + 2qy_{n+1}\,, \\ y_{n+2} &= y_{n+1} + \frac{1}{2}qy_{n+1} + \frac{1}{2}qy^*_{n+2}\,, \end{aligned}$$

da cui, eliminando y^*_{n+2}, si ottiene

$$y_{n+2} - \left(1 + \frac{1}{2}q + q^2\right) y_{n+1} - \frac{1}{2}qy_n = 0\,.$$

Il metodo (8.55) è dotato di una regione non vuota di assoluta stabilità come si verifica dalla sua regione S_A in Fig 8.3 costituita dai punti interni alla porzione di piano delimitata dalla curva data.

Particolare importanza hanno i metodi di predizione e correzione costruiti con due formule dello stesso ordine: in tal caso si può avere ad ogni passo una buona stima della parte principale dell'errore locale di troncamento sia del

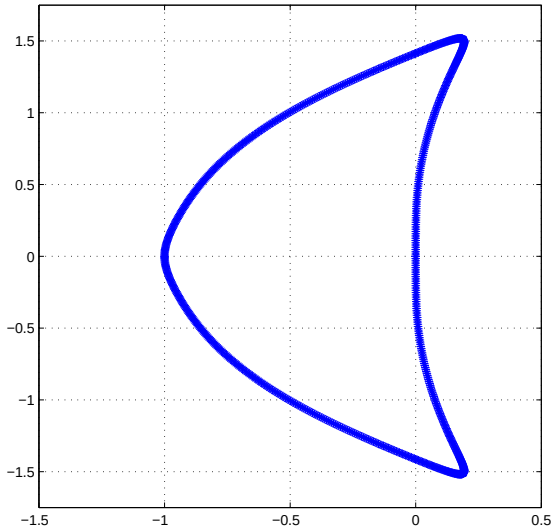

Figura 8.3: Regione S_A del metodo (8.55).

predittore che del correttore. Basandosi sull'assunto che nei membri destri delle (8.54) sia

$$y_{n+j} = y(t_{n+j}) \,, \quad j = 0, 1, \ldots, k-1,$$

tenuto conto della definizione (8.37) di errore locale di troncamento e della (8.40), può scriversi, per un predittore e un correttore entrambi di ordine p,

$$\begin{aligned} y(t_{n+k}) &= y^*_{n+k} + c^*_{p+1} y^{(p+1)}(t_n) h^{p+1} + O(h^{p+2}) \,, \\ y(t_{n+k}) &= y_{n+k} + c_{p+1} y^{(p+1)}(t_n) h^{p+1} + O(h^{p+2}) \,. \end{aligned} \tag{8.56}$$

Dalle precedenti uguaglianze si ricava la relazione

$$c_{p+1} y^{(p+1)}(t_n) h^{p+1} = \frac{c_{p+1}}{c^*_{p+1} - c_{p+1}} (y_{n+k} - y^*_{n+k}) + O(h^{p+2}) \tag{8.57}$$

che fornisce la detta stima per il correttore e richiede soltanto la conoscenza del valore predetto e del valore corretto.

Eliminando $y^{(p+1)}(t_n)$ tra la (8.57) e la (8.56) si ottiene una ulteriore approssimazione di $y(t_{n+k})$ data da

$$\hat{y}_{n+k} = y_{n+k} + \frac{c_{p+1}}{c^*_{p+1} - c_{p+1}} (y_{n+k} - y^*_{n+k}) \,, \tag{8.58}$$

e si ha $y(t_{n+k}) - \hat{y}_{n+k} = O(h^{p+2})$.

La (8.58), detta *correzione di Milne* o *estrapolazione locale*, può essere una ulteriore relazione in un metodo di predizione e correzione PEC o $PECE$: ad esempio, usando come predittore la (8.48) e come correttore la (8.47), si ottiene il seguente metodo

$$\begin{aligned} y^*_{n+4} &= y_n + \frac{4}{3}h(2f_{n+1} - f_{n+2} + 2f_{n+3}) \,, \\ y_{n+4} &= y_{n+2} + \frac{h}{3}(f_{n+2} + 4f_{n+3} + f^*_{n+4}) \,, \\ \hat{y}_{n+4} &= y_{n+4} - \frac{1}{29}(y_{n+4} - y^*_{n+4}) \,. \end{aligned} \tag{8.59}$$

La stima che si ottiene dalla (8.57) può servire anche per controllare ad ogni passo l'entità dell'errore locale, ai fini di una strategia di variazione del passo. In tal caso si riduce il passo se l'errore è troppo grande e si aumenta in caso contrario.

8.3.5 Metodi a più passi: stabilità relativa

Per tutti i metodi esposti in questo capitolo, l'assoluta stabilità viene definita per problemi con soluzioni $y(t)$ tali che $\lim_{t\to+\infty} y(t) = 0$, essendo questo tipo di problema molto frequente nelle applicazioni. Da qui l'uso di problemi test della forma

$$y' = \lambda y \,, \quad \text{o} \quad y' = Ky$$

con le rispettive condizioni

$$Re(\lambda) < 0 \,, \quad \text{e} \quad Re(\lambda_i) < 0 \,,\ i = 1, 2, \ldots, m.$$

Per i problemi con soluzione $y(t)$ tale che $\lim_{t\to+\infty} \|y(t)\| = +\infty$, si fa riferimento agli stessi problemi test, dove ora si assume rispettivamente

$$Re(\lambda) > 0 \,, \quad \text{e} \quad Re(\lambda_j) > 0 \,, \text{ per qualche } \ j \in \{1, 2, \ldots, m\}.$$

In tal caso si richiede, per la soluzione numerica, che sia $\lim_{n\to\infty} \|y_n\| = +\infty$ e a tale scopo si introduce il concetto di *stabilità relativa*, per significare che l'errore relativo $\frac{\|y_n - y(t_n)\|}{\|y(t_n)\|}$ si mantiene "accettabile" (cioè piccolo rispetto a 1) al crescere di n. Per esempio, nel caso dei metodi lineari e per $m = 1$, si propone, di solito, la seguente definizione.

Definizione 8.3.6 *Il metodo lineare (8.35) si dice* relativamente stabile *per un dato $q \in \mathbb{C}$, se, applicato al problema test $y' = \lambda y$, $Re(\lambda) > 0$, gli zeri del suo polinomio di stabilità $\pi(q, \mu)$ verificano le condizioni*

$$| \mu_i(q) | < | \mu_1(q) | , \quad i = 2, 3, \ldots, k,$$

dove $\mu_1(q)$ è la radice principale di $\pi(q, \mu) = 0$.

Definizione 8.3.7 *L'insieme del piano complesso*

$$S_R = \{q \in \mathbb{C} \ \ | \ | \mu_i(q) | < | \mu_1(q) |, \ i = 2, 3, \ldots, k\}$$

si dice regione di relativa stabilità *del metodo (8.35).*

Si consideri, ad esempio, la formula del punto centrale (8.46): il suo polinomio di stabilità è

$$\pi(q, \mu) = \mu^2 - 2q\mu - 1 ,$$

i cui zeri sono $\mu_1(q) = q + \sqrt{q^2 + 1}$ e $\mu_2 = q - \sqrt{q^2 + 1}$. Poiché $| \mu_1 \mu_2 | = 1$, la formula non ha una regione di assoluta stabilità. Tuttavia, limitandosi per semplicità al caso $q \in \mathbb{R}$, risulta $\mu_2(q) < \mu_1(q)$ se $q > 0$: essa, quindi, è dotata di un intervallo reale di relativa stabilità coincidente con il semiasse positivo.

Si conclude questa rassegna di metodi per problemi di valori iniziali osservando che le proprietà di stabilità qui riportate fanno parte della cosiddetta teoria della stabilità lineare, in cui si fa riferimento a problemi test lineari a coefficienti costanti, ma se ne estendono i risultati a problemi più generali.

Anche se la stabilità lineare è un requisito necessario, tuttavia essa può rivelarsi inadeguata quando un metodo, stabile in senso lineare, venga applicato ad un problema (8.1) con $f(t, y)$ non lineare, o anche lineare con la matrice K dipendente da t oppure non diagonalizzabile. In questi casi sarebbe più appropriata una forma di *stabilità non lineare*, la cui trattazione va oltre gli scopi del presente testo ed è reperibile in opere specializzate.

8.4 Problemi ai limiti e BV-metodi

8.4.1 Introduzione

Nell'Esempio 6.8.8 si è considerato un problema continuo per una equazione differenziale del secondo ordine, che differisce da un problema di Cauchy

per il fatto che la soluzione è assoggettata, invece che ad una condizione in un punto iniziale, a due condizioni poste agli estremi di un intervallo. Indipendentemente dall'ordine dell'equazione differenziale, che può sempre ricondursi ad un sistema del primo ordine (cfr. 8.1), i problemi continui con condizioni che coinvolgono i valori della soluzione in più punti si dicono *problemi ai limiti*.

La discretizzazione di questi problemi si può fare mediante un *problema ai limiti discreto*, cioè costituito da una equazione alle differenze di ordine k a cui sono associate condizioni su valori iniziali e valori finali della soluzione.

Più in generale, un problema ai limiti discreto si può sempre ottenere da un problema continuo, sia ai limiti che di valori iniziali, applicando una formula lineare a k passi del tipo (8.35)

$$\sum_{j=0}^{k} \alpha_j y_{n+j} - h \sum_{j=0}^{k} \beta_j f_{n+j} = 0 \,. \tag{8.60}$$

In tal caso la formula prende il nome di BV-*metodo* o, più brevemente, BVM (Boundary Value Method). Per contro quando la (8.60) è usata come nei paragrafi precedenti (cioè quando trasforma un problema continuo di valori iniziali in un problema discreto di valori iniziali), allora viene indicata con la sigla IVM (Initial Value Method). Nel caso dei problemi di valori iniziali, l'uso dei BVM presenta alcuni notevoli vantaggi rispetto ai tradizionali IVM. Per esempio il superamento della seconda barriera di Dahlquist (cfr. Teorema 8.3.4) e la maggiore facilità di stima dell'errore globale. I due usi della (8.60) come IVM e come BVM sono schematizzati nella Tavola 8.4. Si noti che esistono anche metodi che riconducono un problema continuo ai limiti ad un equivalente problema di valori iniziali che poi viene discretizzato con un IVM. Tali metodi, che qui non sono considerati, sono noti come *metodi shooting*.

Problema continuo	Metodo	Problema discreto
Valori iniziali	$\xrightarrow{\text{IVM}}$	Valori iniziali
	$\searrow^{\text{BVM}}$	
Valori ai limiti	$\xrightarrow{\text{BVM}}$	Valori ai limiti

Tavola 8.4: Caratterizzazione degli IVM e dei BVM.

8.4.2 Modo di impiego dei BVM

Di solito una formula del tipo (8.60) usata come BVM si dice *metodo base* e viene associata ad altre formule dello stesso tipo, generalmente implicite e con un numero di passi minore, che si dicono *metodi ausiliari*. Per delineare come si impiegano tali formule, si considererà il problema continuo di valori iniziali: il caso di un problema di valori ai limiti necessita solo di poche e semplici modifiche.
Si considera un passo di discretizzazione costante $h > 0$ sull'intervallo $[a, b]$. Se $t_0 = a$, $t_N = b$ e $t_n = t_0 + nh$, $n = 0, 1, \ldots, N$, si assume $h = (t_N - t_0)/N$. Applicando al problema il metodo base (8.60), per $n = 0, 1, \ldots, N - k$, si scrivono $N - k + 1$ relazioni indipendenti.

Queste $N - k + 1$ relazioni si possono considerare come un sistema nelle N incognite $y_1, \ldots, y_N$. Vi sono quindi ancora $k - 1$ incognite del problema discreto che devono essere determinate e ciò si può fare aggiungendo al sistema altrettanti metodi ausiliari. Se $f(t, y)$ è lineare rispetto a y tale è anche il sistema che così si ottiene: la soluzione esiste ed è unica per l'indipendenza lineare delle equazioni che lo costituiscono. Tale soluzione si assume come approssimazione discreta della soluzione del problema continuo (cfr. 8.5.5)

Se il problema continuo non è lineare, il procedimento ora descritto rimane valido, ma conduce ad un sistema discreto non lineare. In tal caso si può dimostrare, sotto opportune ipotesi, l'esistenza di una soluzione unica del problema discreto e la convergenza delle iterazioni di Newton (cfr. 4.6).

In particolare il sistema viene strutturato come segue.
Alle $N - k + 1$ equazioni ottenute dal metodo base vengono aggiunte $k - 1$ equazioni date da altrettanti metodi ausiliari. Tali metodi, per motivi che si chiariranno meglio nel paragrafo successivo, sono generalmente organizzati in due gruppi: $k_1 - 1$ metodi ausiliari "di testa" da porre come equazioni iniziali del sistema e k_2 metodi ausiliari "di coda" da porre come equazioni finali, con $k = k_1 + k_2$. Pertanto il sistema di N equazioni nelle N incognite $y_1, \ldots, y_N$ assume la forma seguente:

$$\begin{cases} \sum_{j=0}^{r} \alpha_{j\nu} y_j - h \sum_{j=0}^{r} \beta_{j\nu} f_j = 0, & \nu = 1, \ldots, k_1 - 1, \\ \sum_{j=0}^{k} \alpha_j y_{\nu+j-k_1} - h \sum_{j=0}^{k} \beta_j f_{\nu+j-k_1} = 0, & \nu = k_1, \ldots, N - k_2, \\ \sum_{j=0}^{s} \alpha_{j\nu} y_{N+j-s} - h \sum_{j=0}^{s} \beta_{j\nu} f_{N+j-s} = 0, & \nu = N - k_2 + 1, \ldots, N. \end{cases} \tag{8.61}$$

Si noti che nelle prime $k_1 - 1$ equazioni ausiliarie del sistema sono implicate le incognite $y_1, \ldots, y_r$, nelle equazioni del metodo di base le incognite $y_1, \ldots, y_N$, mentre nelle ultime k_2 equazioni ausiliarie le incognite $y_{N-s}, \ldots, y_N$. Tale sistema costituisce un *metodo BVM con* (k_1, k_2)*-condizioni al contorno* o, più semplicemente $\mathrm{BVM}_{k_1 k_2}$.
Poiché, in generale, è $N \gg k$, si può provare che le caratteristiche di convergenza e di stabilità del metodo (8.61) sono essenzialmente regolate dal metodo base il cui ordine p, per ovvi motivi di omogeneità, è il medesimo delle formule ausiliarie. Pertanto, in quello che segue, $\rho(\mu)$, $\sigma(\mu)$, $\pi(q, \mu)$ sono i polinomi del metodo base.

8.4.3 Stabilità e convergenza dei BVM

Le condizioni che si richiedono per la stabilità di un $\mathrm{BVM}_{k_1 k_2}$ sono alquanto diverse da quelle già viste per un IVM.

Si premettono alcune definizioni.

Definizione 8.4.1 *Sia* $k = k_1 + k_2 + k_3$ *con* k_1, k_2, k_3 *interi non negativi. Il polinomio a coefficienti reali*

$$p(\mu) = \sum_{j=0}^{k} a_j \mu^j$$

si dice del tipo $(\mathsf{k}_1,\mathsf{k}_2,\mathsf{k}_3)$ *se ha* k_1 *zeri in modulo minori di* 1, k_2 *zeri in modulo uguali a* 1 *e* k_3 *zeri in modulo maggiori di* 1.
Sono polinomi di Schur *quelli del tipo* $(k, 0, 0)$, *sono* polinomi di Von Neumann *quelli del tipo* $(k_1, k-k_1, 0)$ *con i* $k-k_1$ *zeri di modulo* 1 *tutti semplici, mentre si dicono* polinomi conservativi *quelli del tipo* $(0, k, 0)$.

Definizione 8.4.2 *Sia* $k = k_1 + k_2$ *con* k_1, k_2 *interi non negativi. Il polinomio a coefficienti reali*

$$p(\mu) = \sum_{j=0}^{k} a_j \mu^j$$

si dice un $\mathsf{S}_{k_1 k_2}$-polinomio *se per i suoi zeri risulta*

$$|\mu_1| \leq |\mu_2| \leq \ldots |\mu_{k_1}| < 1 < |\mu_{k_1+1}| \leq \ldots \leq |\mu_k|,$$

mentre si dice un $\mathsf{N}_{k_1 k_2}$-polinomio *se*

$$|\mu_1| \leq |\mu_2| \leq \ldots |\mu_{k_1}| \leq 1 < |\mu_{k_1+1}| \leq \ldots \leq |\mu_k|$$

e gli zeri di modulo unitario sono semplici.

Si osservi che un $S_{k_1k_2}$-polinomio è del tipo $(k_1, 0, k - k_1)$ e quindi un S_{k0}-polinomio è un polinomio di Schur, mentre un N_{k0}-polinomio è un polinomio di Von Neumann.
Di conseguenza un metodo lineare a k passi usato come IVM è zero-stabile se il suo polinomio caratteristico $\rho(\mu)$ è un polinomio di Von Neumann (ovvero un N_{k0}-polinomio) mentre è assolutamente stabile se $\pi(q, \mu)$ è un polinomio del tipo $(k, 0, 0)$ (ovvero un polinomio di Schur, ovvero ancora un S_{k0}-polinomio).

Per la zero-stabilità di un IVM si è fatto riferimento al problema test $y' = 0, \quad y(t_0) = y_0$. Ma lo stesso risultato si ottiene assumendo in un metodo a k passi $h = 0$. In questo senso la zero-stabilità è una stabilità asintotica per h che tende a 0. Un analogo concetto può essere dato nel caso di un $BVM_{k_1k_2}$. Precisamente vale il seguente risultato.

Teorema 8.4.1 *Un metodo $BVM_{k_1k_2}$ è convergente nel senso che*

$$\| y(t_n) - y_n \| = O(h^p)$$

se $\rho(1) = 0$, $\rho'(1) = \sigma(1)$ (coerenza) e se $\rho(\mu)$ è un $N_{k_1k_2}$-polinomio.

Ciò conduce alla seguente definizione.

Definizione 8.4.3 *Un $BVM_{k_1k_2}$ si dice* $0_{k_1k_2}$-stabile *se il corrispondente polinomio $\rho(z)$ è un $N_{k_1k_2}$-polinomio.*

Quindi coerenza e $0_{k_1k_2}$-stabilità sono condizioni sufficienti, ma non anche necessarie come nel caso di un IVM (cfr. Teorema 8.3.2), per la convergenza di un $BVM_{k_1k_2}$.

Si esamina ora la assoluta stabilità di un $BVM_{k_1k_2}$, cioè nel caso di h fisso.

Si consideri il problema test

$$y'(t) = \lambda y(t),\ y(t_0) = y_0,\ \lambda \in \mathbb{C},\ Re(\lambda) < 0$$

già utilizzato per studiare l'assoluta stabilità di un metodo di Runge-Kutta e di un IVM.
Il problema, quindi, è quello di approssimare con un $BVM_{k_1k_2}$ la soluzione $y(t_n) = y_0 e^{\lambda(nh)} = y_0(e^q)^n$.
Vale il seguente teorema.

Teorema 8.4.2 *Si supponga che per gli zeri del polinomio $\pi(q,\mu)$ si abbia*

$$\begin{aligned} |\mu_1(q)| \leq \ldots \leq |\mu_{k_1-1}(q)| &< |\mu_{k_1}(q)| < |\mu_{k_1+1}(q)| \leq \ldots \leq |\mu_k(q)| \\ |\mu_{k_1-1}(q)| &< 1 < |\mu_{k_1+1}(q)| \end{aligned}$$

con $\mu_{k_1}(0) = \mu_{k_1} = 1$, ovvero si ammetta che $\mu_{k_1}(q)$ sia radice principale, tale quindi da avere $\mu_{k_1}(q) = e^q + O(h^{p+1})$, allora si ha

$$y_n = \mu_{k_1}^n(q)(y_0 + O(h^p)) + O(h^p)(O(|\mu_{k_1-1}(q)|^n) + O(|\mu_{k_1+1}(q)|^{-(N-n)})).$$

Questo risultato conduce alla seguente definizione di *Assoluta stabilità* per un $\mathrm{BVM}_{k_1k_2}$

Definizione 8.4.4 *Un $BVM_{k_1k_2}$ si dice (k_1,k_2)-*Assolutamente stabile *per un dato q se il polinomio $\pi(q,\mu)$ è del tipo $(k_1,0,k_2)$, cioè un $S_{k_1k_2}$-polinomio.*

Definizione 8.4.5 *La regione del piano complesso*

$$S_{A_{k_1k_2}} = \{q \in \mathbb{C} \,|\pi(q,\mu) \text{è del tipo } (k_1,0,k_2)\}$$

*si dice regione di (k_1,k_2)-*Assoluta stabilità.

Si noti che se $k_2 = 0$ queste definizioni coincidono con quelle ordinarie per IVM.

Definizione 8.4.6 *Un $BVM_{k_1k_2}$ si dice* $\mathsf{A}_{k_1k_2}$-stabile *se*

$$S_{A_{k_1k_2}} \supseteq \{q \in \mathbb{C} \,|Re(q) < 0\}.$$

Per illustrare come una medesima formula possa avere comportamenti diversi se usata come IVM o $\mathrm{BVM}_{k_1k_2}$ si consideri la formula del punto centrale (8.46) per la quale risulta $\rho(\mu) = \mu^2 - 1$ e $\pi(q,\mu) = \mu^2(q) - 2q\mu(q) - 1$. Come già notato è zero-stabile essendo $\rho(\mu)$ un polinomio di Von Neumann, ma S_A è vuoto avendosi in ogni caso $|\mu_1(q)| > 1$ oppure $|\mu_2(q)| > 1$: tale formula non può essere usata da sola come IVM. Inoltre essa non è 0_{11}-stabile perché $\rho(\mu)$ non è un N_{11}-polinomio (dove è richiesto $|\mu_1| < |\mu_2|$). Tuttavia per $Re(q) < 0$ risulta $|\mu_1(q)| < 1 < |\mu_2(q)|$ e quindi $\pi(q,\mu)$ è un S_{11}-polinomio: per $Re(q) < 0$, quindi, usata con una equazione ausiliaria come BVM_{11} fornisce un metodo A_{11}-stabile (vedi paragrafo 8.5.5).

Come già accennato, uno degli aspetti più importanti dei BVM consiste nell'esistenza di metodi $\mathrm{A}_{k_1k_2}$-stabili di ordine $2k$, cioè del massimo ordine

possibile per formule lineari a k passi: ciò esclude di fatto ogni barriera d'ordine per metodi $\mathrm{BVM}_{k_1k_2}$ stabili.

Di seguito si forniscono tre esempi di metodi $\mathrm{A}_{k_1k_2}$-stabili: altri possono essere trovati nel testo *Brugnano-Trigiante* [4].

Metodo ETR (Extended Trapezoidal Rule),
p=4, $k_1 = 2, k_2 = 1$:

$$\begin{cases} y_1 - y_0 = \frac{h}{24}(f_3 - 5f_2 + 19f_1 + 9f_0), \\ \\ y_n - y_{n-1} = \frac{h}{24}(-f_{n+1} + 13f_n + 13f_{n-1} - f_{n-2}), \quad n = 2, \ldots, N-1, \\ \\ y_N - y_{N-1} = \frac{h}{24}(f_{N-3} - 5f_{N-2} + 19f_{N-1} + 9f_N). \end{cases}$$

Metodo GAM (Generalized Adams Method),
p=5, $k_1 = 2, k_2 = 2$:

$$\begin{cases} y_1 - y_0 = \frac{h}{720}(-19f_4 + 106f_3 - 264f_2 + 646f_1 + 251f_0), \\ \\ y_n - y_{n-1} = \frac{h}{720}(-19f_{n-2} + 346f_{n-1} + 456f_n - 74f_{n+1} + 11f_{n+2}), \\ \qquad\qquad n = 2, \ldots, N-1, \\ \\ y_N - y_{N-1} = \frac{h}{720}(-19f_{N+1} + 346f_N + 456f_{N-1} - 74f_{N-2} + 11f_{N-3}), \\ \\ y_{N+1} - y_N = \frac{h}{720}(-19f_{N-3} + 106f_{N-2} - 264f_{N-1} + 646f_N + 251f_{N+1}). \end{cases}$$

Metodo TOM (Top Order Method),
p=6, $k_1 = 2, k_2 = 1$:

$$\begin{cases} y_1 - y_0 = \frac{h}{1440}(27f_5 - 173f_4 + 482f_3 - 798f_2 + 1427f_1 + 475f_0), \\ \\ \frac{1}{60}(11y_{n+1} + 27y_n - 27y_{n-1} - 11y_{n-2}) = \\ \qquad\qquad \frac{h}{20}(f_{n+1} + 9f_n + 9f_{n-1} + f_{n-2}), \quad n = 2, \ldots, N-1, \\ \\ y_N - y_{N-1} = \\ \quad \frac{h}{1440}(27f_{N-5} - 173f_{N-4} + 482f_{N-3} - 798f_{N-2} + 1427f_{N-1} + 475f_N). \end{cases}$$

Si osservi che la formula trapezoidale

$$y_{n+1} - y_n = \frac{h}{2}(f_{n+1} + f_n),$$

che è A-stabile se usata come IVM, risulta essere A_{10}-stabile come BVM, cioè è un metodo BVM_{10}.
Infatti il suo polinomio di stabilità ha un unico zero $|\mu_1(q)| < 1$ per $Re(q) < 0$. Essa, pertanto, può essere usata da sola, cioè senza formule ausiliarie, con $n = 0, 1, \ldots, N-1$, avendosi $k_1 - 1 = 0$ e $k_2 = 0$.

8.5 Complementi ed esempi

8.5.1 I metodi di Runge-Kutta impliciti come metodi di collocazione

Il *metodo di collocazione* per un problema di valori iniziali della forma (8.1) consiste nel determinare un polinomio di grado s, con coefficienti in $\mathbb{R}^m$, che approssimi la soluzione sull'intervallo $[t_n, t_n + h]$.

Dati i numeri reali due a due distinti $c_1, c_2, \ldots, c_s \in [0, 1]$, il corrispondente *polinomio di collocazione* $u(t)$, di grado s, è definito univocamente dalle condizioni

$$u(t_n) = y_n \,, \tag{8.62}$$

$$u'(t_n + c_i h) = f(t_n + c_i h, u(t_n + c_i h)), i = 1, 2, \ldots, s \,. \tag{8.63}$$

Si assume come soluzione numerica della equazione differenziale nel punto t_{n+1} il valore

$$y_{n+1} = u(t_n + h) \,. \tag{8.64}$$

Il metodo di collocazione (8.62)-(8.63) è equivalente ad un metodo di Runge-Kutta implicito ad s stadi ove si ponga

$$a_{ij} = \int_0^{c_i} l_j(\tau) d\tau \,, \quad b_j = \int_0^1 l_j(\tau) d\tau \,, \quad i, j = 1, 2, \ldots, s, \tag{8.65}$$

essendo

$$l_j(\tau) = \frac{(\tau - c_1) \cdots (\tau - c_{j-1})(\tau - c_{j+1}) \cdots (\tau - c_s)}{(c_j - c_1) \cdots (c_j - c_{j-1})(c_j - c_{j+1}) \cdots (c_j - c_s)} \,, \; j = 1, 2, \ldots, s,$$

i polinomi fondamentali della interpolazione di Lagrange relativi ai punti $c_1, c_2, \ldots, c_s$. Infatti, poiché il polinomio $u'(t_n + \tau h)$ si può scrivere come polinomio di interpolazione relativo ai detti punti, posto

$$k_i = u'(t_n + c_i h) \,, \tag{8.66}$$

si ha (cfr. 6.2)

$$u'(t_n + \tau h) = \sum_{j=1}^{s} l_j(\tau) k_j \, ;$$

integrando entrambi i membri, rispetto a t, sugli intervalli $[t_n, t_n + c_i h]$, $i = 1, 2, \ldots, s$ e $[t_n, t_{n+1}]$, si ottiene rispettivamente

$$u(t_n + c_i h) = u(t_n) + h \sum_{j=1}^{s} \left(\int_0^{c_i} l_j(\tau) d\tau \right) k_j \, , \quad i = 1, 2, \ldots, s, \tag{8.67}$$

e

$$u(t_n + h) = u(t_n) + h \sum_{j=1}^{s} \left(\int_0^{1} l_j(\tau) d\tau \right) k_j \, . \tag{8.68}$$

La (8.68) si può anche scrivere, tenendo conto delle (8.64), (8.62) e (8.65),

$$y_{n+1} = y_n + h \sum_{j=1}^{s} b_j k_j \, ,$$

mentre l'equazione (8.63), utilizzando nel primo membro la (8.66) e nel secondo la (8.67), diventa

$$k_i = f(t_n + c_i h, y_n + h \sum_{j=1}^{s} a_{ij} k_j) \, , \quad i = 1, 2, \ldots, s.$$

8.5.2 Sulla stabilità e l'ordine dei metodi di Runge-Kutta

Esempio 8.5.1 Si vuole studiare la stabilità e calcolare l'ordine del seguente metodo di Runge-Kutta semi-implicito

$$A = \begin{pmatrix} 1/s & 0 & \cdots & 0 \\ 1/s & 1/s & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 1/s & 1/s & \cdots & 1/s \end{pmatrix}, \; b = \begin{pmatrix} 1/s \\ 1/s \\ \vdots \\ 1/s \end{pmatrix}, \; c \in I\!R^s \, .$$

Si verifica che

$$I - qA = \begin{pmatrix} 1 - q/s & 0 & \cdots & 0 \\ -q/s & 1 - q/s & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ -q/s & -q/s & \cdots & 1 - q/s \end{pmatrix}$$

e

$$I - qA + qub^T = \begin{pmatrix} 1 & q/s & \cdots & q/s \\ 0 & 1 & \cdots & q/s \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} .$$

Segue, dalla (8.24),

$$R(q) = \frac{1}{(1 - q/s)^s} . \tag{8.69}$$

Il metodo, quindi, è A-stabile.

Il calcolo dell'ordine può essere fatto in base alla (8.25). Sviluppando in serie di Taylor $R(q)$ si ottiene

$$R(q) = 1 + q + \frac{s+1}{2s} q^2 + O(q^3) .$$

Confrontando con la serie esponenziale si trova

$$e^q - R(q) = -\frac{1}{2s} q^2 + O(q^3) ,$$

da cui, essendo $q = h\lambda$, si ha $p = 1$.

Si osservi che, dalla (8.69), risulta

$$\lim_{s \to \infty} R(q) = e^q .$$

□

Esempio 8.5.2 Si vuole determinare l'ordine del metodo semi-implicito

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/2 & 1/4 & 1/4 & 0 \\ 1 & 0 & 1 & 0 \\ \hline & 1/6 & 4/6 & 1/6 \end{array}$$

e calcolare un passo h idoneo ad approssimare, con tale metodo, la soluzione del problema

$$y' = Ky$$

dove

$$K = \begin{pmatrix} -4 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -4 \end{pmatrix} \in \mathbb{R}^{m\times m} .$$

Dalla (8.24) si ottiene

$$R(q) = \frac{1 + \frac{3}{4}q + \frac{1}{4}q^2 + \frac{1}{24}q^3}{1 - \frac{1}{4}q} .$$

Pertanto risulta (cfr. Tavola 8.3) $R(q) = R_1^3(q)$: da ciò e dalle (8.25) e (8.30) si ricava che l'ordine del metodo è $p = 4$.

Gli autovalori $\lambda_1, \lambda_2, \ldots, \lambda_m$ di K sono reali e, per il primo teorema di Gershgorin, si ha

$$\lambda_i \in]-6, -2[, \quad i = 1, 2, \ldots, m.$$

D'altro canto si verifica che $| \ R(q) \ | < 1$ se $q \in]-5.42, 0[$. Segue che la condizione (8.28) è soddisfatta se $h < \frac{-5.42}{-6} \simeq 0.9$.

Si osservi infine che, nell'applicazione del metodo al problema dato, k_1 si calcola direttamente e così k_3, una volta noto k_2. Il calcolo di k_2 richiede la risoluzione del sistema lineare

$$(I - \frac{1}{4}hK)k_2 = Ky_n + \frac{1}{4}hKk_1 .$$

Poiché gli autovalori di $I - \frac{1}{4}hK$ sono dati da $1 - \frac{1}{4}h\lambda_i$, $i = 1, 2, \ldots, m$, (cfr. Teorema 2.7.6) ed essendo $\lambda_i < 0$, $i = 1, 2, \ldots, m$, risulta sicuramente $\det(I - \frac{1}{4}hK) \neq 0$ per ogni $h > 0$ (cfr. Osservazione 2.7.1). □

8.5.3 Costruzione dei metodi a più passi ed esame della stabilità

Un metodo a più passi può essere costruito, a partire dalla sua forma generale (8.35), calcolandone i coefficienti mediante le condizioni di ordine (8.39) e imponendo che l'ordine p risulti massimo, compatibilmente con la condizione di zero-stabilità.

Esempio 8.5.3 Si vogliono determinare i coefficienti della formula BDF a due passi

$$\alpha_0 y_n + \alpha_1 y_{n+1} + \alpha_2 y_{n+2} = h\beta_2 f_{n+2} \, .$$

Per definizione $\alpha_2 = 1$. Per la coerenza deve aversi $\rho(1) = 0$ e $\rho'(1) - \sigma(1) = 0$ ovvero $\alpha_0 + \alpha_1 + 1 = 0$ e $2 + \alpha_1 - \beta_2 = 0$, da cui, esprimendo α_1 e β_2 in funzione di α_0, $\alpha_1 = -1 - \alpha_0$ e $\beta_2 = 1 - \alpha_0$.
Risulta quindi $\rho(\mu) = \mu^2 - (1+\alpha_0)\mu + \alpha_0 = (\mu - 1)(\mu - \alpha_0)$: la zero-stabilità è garantita se $\mid \alpha_0 \mid < 1$ oppure $\alpha_0 = -1$. L'ordine massimo si ottiene scegliendo α_0 in modo che risulti $c_2 = 0$. Dalle (8.39) con $r = 2$ si ha $c_2 = -\frac{1}{2} + \frac{3}{2}\alpha_0$ da cui $\alpha_0 = \frac{1}{3}$.
Allo stesso risultato si giunge, in virtù della (8.40), osservando che per un metodo di ordine p l'errore locale di troncamento è nullo se $y(t) = t^r$, $r = 0, 1, \ldots, p$, e quindi, utilizzando la (8.37), scrivendo $p + 1$ relazioni lineari nelle incognite α_j, β_j.

Il polinomio di stabilità del metodo è $\pi(q, \mu) = \left(1 - \frac{2}{3}q\right)\mu^2 - \frac{4}{3}\mu + \frac{1}{3}$ i cui zeri sono

$$\mu_1 = \frac{2 + \sqrt{1 + 2q}}{3 - 2q} \, , \quad \mu_2 = \frac{2 - \sqrt{1 + 2q}}{3 - 2q} \, .$$

Limitandosi al caso $q \in \mathbb{R}$, con semplici calcoli, si trova che il metodo è assolutamente stabile su tutto l'asse reale ad eccezione dell'intervallo $[0, 4]$ (il metodo, quindi, è A_0-stabile e si può, poi, dimostrare che è anche A-stabile), mentre è relativamente stabile per $q > -\frac{1}{2}$. □

8.5.4 Sulla stabilità dei metodi di predizione e correzione

Come si è visto in 8.3.4, le caratteristiche di stabilità di un metodo di predizione e correzione sono, in genere, diverse da quelle dei metodi che lo compongono. Per esempio, il metodo (8.55), pur essendo dotato di una regione di assoluta stabilità, perde la A-stabilità posseduta dal solo correttore. In altri casi, tuttavia, la stabilità è migliore di quella del predittore e di quella del correttore.

Esempio 8.5.4 Si consideri il metodo (8.59) senza la correzione di Milne.
Il polinomio di stabilità del predittore è

$$\pi^*(q, \mu) = \mu^4 - \frac{8}{3}q\mu^3 + \frac{4}{3}q\mu^2 - \frac{8}{3}q\mu - 1 \, .$$

Poiché gli zeri di $\pi^*(q,\mu)$ verificano la relazione $\mu_1\mu_2\mu_3\mu_4 = -1$ (cfr. la (4.50) con $i = m$), ne segue che almeno uno di essi è, in modulo, ≥ 1.

Il polinomio di stabilità del correttore è

$$\pi(q,\mu) = \left(1 - \frac{1}{3}q\right)\mu^2 - \frac{4}{3}q\mu - \left(1 + \frac{1}{3}q\right)$$

e si constata facilmente che per esso risulta $\mid \mu_1(q) \mid \geq 1$ se $Re(q) \geq 0$ e $\mid \mu_2(q) \mid > 1$ se $Re(q) < 0$. Se ne conclude che né il predittore né il correttore sono assolutamente stabili.

Il polinomio di stabilità del metodo di predizione e correzione definito dai due metodi è (cfr. la procedura già adottata per il metodo (8.55))

$$\hat{\pi}(q,\mu) = \mu^4 - \left(\frac{8}{9}q^2 + \frac{4}{3}q\right)\mu^3 + \left(\frac{4}{9}q^2 - \frac{1}{3}q - 1\right)\mu^2 - \frac{8}{9}q^2\mu - \frac{1}{3}q\,.$$

Limitandosi per semplicità al caso $q \in \mathbb{R}$, si verifica che il metodo è assolutamente stabile per $q = -\frac{1}{2}$. Posto $P_0(\mu) = \hat{\pi}\left(-\frac{1}{2},\mu\right)$, se ne consideri la successione di Sturm (cfr. Definizione e Teorema 4.7.1)

$$\begin{aligned}
P_0(\mu) &= \mu^4 + \tfrac{4}{9}\mu^3 - \tfrac{13}{18}\mu^2 - \tfrac{2}{9}\mu + \tfrac{1}{6},\\
P_1(\mu) &= \mu^3 + \tfrac{1}{3}\mu^2 - \tfrac{13}{36}\mu - \tfrac{1}{18},\\
P_2(\mu) &= \mu^2 + \tfrac{41}{129}\mu - \tfrac{56}{129},\\
P_3(\mu) &= \mu + \tfrac{1625}{164},\\
P_4(\mu) &= costante < 0.
\end{aligned}$$

Poiché risulta $V(-\infty) - V(0) = 0$ e $V(0) - V(+\infty) = 2$, il polinomio $\hat{\pi}\left(-\frac{1}{2},\mu\right)$ ha due zeri reali e positivi e due complessi coniugati. Per gli zeri reali si ha $0.6 < \mu_1 < 0.7$ e $0.4 < \mu_2 < 0.5$.
Indicato con ρ il modulo comune dei due zeri complessi coniugati, vale la relazione $\mu_1\mu_2\rho^2 = \frac{1}{6}$ (cfr. ancora la (4.50) con $i = m$) e quindi, poiché $0.24 < \mu_1\mu_2 < 0.35$, risulta $\rho^2 < 1$.

Uno studio più completo di $\hat{\pi}(q,\mu)$ mostra che il metodo in esame è dotato dell'intervallo reale di assoluta stabilità $]-0.8, -0.3[$. □

Un altro caso di miglioramento della stabilità è fornito dal *metodo di Hermite*

$$y_{n+2} = -4y_{n+1} + 5y_n + h(4f_{n+1} + 2f_n)\,,$$

addirittura non zero-stabile, e la regola di Simpson (8.47) che è zero-stabile, ma non assolutamente stabile. Usati insieme come predittore e correttore, danno luogo ad un metodo dotato di una regione di assoluta stabilità i cui estremi sono:

$$\text{minimo di } Re(q) = -1 \text{ e massimo di } \mid Im(q) \mid = 0.5.$$

8.5.5 Esempi di applicazione dei BVM

Esempio 8.5.5 Si abbia il problema di valori iniziali, scalare e lineare,

$$y'(t) = p(t)y(t) + q(t),\ p(t) < 0,\ a \le t \le b\,, \tag{8.70}$$

$$y(a) = y_0\,. \tag{8.71}$$

Volendo usare un $\mathrm{BVM}_{k_1k_2}$, seguendo quanto detto in 8.4.2, si ponga $t_0 = a$, $t_N = b$ ed $h = \frac{b-a}{N}$. L'intero N può essere fissato in base all'accuratezza che si vuole ottenere, tenendo conto che, se il metodo base prescelto è di ordine p, il $\mathrm{BVM}_{k_1k_2}$ produce errori dell'ordine di grandezza di $h^p = \left(\frac{b-a}{N}\right)^p$.

Essendo $p(t) < 0$ conviene usare un metodo base $\mathrm{A}_{k_1k_2}$-stabile. Scegliendo la formula del punto centrale (8.46), che è A_{11}-stabile, si ha

$$-y_n + y_{n+2} = 2h\left[p(t_{n+1})y_{n+1} + q(t_{n+1})\right]\ ,\ n = 0, 1, \ldots N-2. \tag{8.72}$$

Avendosi $k_1 = k_2 = 1$ occorre un solo metodo ausiliario finale, quale, ad esempio, la formula di Eulero implicita (8.45) con $n = N-1$:

$$-y_{N-1} + y_N = h\left[p(t_N)y_N + q(t_N)\right]\ . \tag{8.73}$$

Si ottiene perciò il sistema $Gy = c$ dove

$$G = \begin{pmatrix} -2hp(t_1) & 1 & & & & \\ -1 & -2hp(t_2) & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -1 & -2hp(t_{N-1}) & 1 & \\ & & & -1 & 1 - hp(t_N) & \end{pmatrix},$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N-1} \\ y_N \end{pmatrix}, \quad c = \begin{pmatrix} 2hq(t_1) + y_0 \\ 2hq(t_2) \\ \vdots \\ 2hq(t_{N-1}) \\ hq(t_N) \end{pmatrix}.$$

Invece di un problema di valori iniziali, si consideri ora il problema ai limiti continuo formato dalla (8.70) con la condizione ai limiti

$$\alpha y(a) + \beta y(b) = 0\,. \tag{8.74}$$

Il problema ai limiti discreto cui si giunge, usando le stesse formule, è costituito dalle (8.72) e (8.73). La corrispondente della (8.74)

$$\alpha y_0 + \beta y_N = 0$$

viene utilizzata ponendo $y_0 = -(\beta/\alpha)y_N$. Il sistema lineare che si ottiene in questo caso è $\tilde{G}y = \tilde{c}$ dove

$$\tilde{G} = \begin{pmatrix} -2hp(t_1) & 1 & & & \beta/\alpha \\ -1 & -2hp(t_2) & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & -2hp(t_{N-1}) & 1 \\ & & & -1 & 1 - hp(t_N) \end{pmatrix},$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N-1} \\ y_N \end{pmatrix}, \quad \tilde{c} = \begin{pmatrix} 2hq(t_1) \\ 2hq(t_2) \\ \vdots \\ 2hq(t_{N-1}) \\ hq(t_N) \end{pmatrix}.$$

□

Esempio 8.5.6 Si consideri la formula

$$y_{n+1} - y_n = h(\frac{5}{12}f_n + \frac{8}{12}f_{n+1} - \frac{1}{12}f_{n+2}).$$

Se ne studia la stabilità come BVM.
Si considera il problema test $y' = \lambda y$, per semplicità, solo nel caso reale, cioè $0 > \lambda \in$R. Il polinomio di stabilità della formula è

$$\pi(q, \mu) = \frac{1}{12}q\mu^2 + (1 - \frac{2}{3}q)\mu - (1 + \frac{5}{12}q).$$

Gli zeri sono quindi

$$\mu_1(q) = (-1 + \frac{2}{3}q + \sqrt{\Delta})/(\frac{1}{6}q), \quad \mu_2(q) = (-1 + \frac{2}{3}q - \sqrt{\Delta})/(\frac{1}{6}q),$$

avendo posto $\Delta = \frac{7}{12}q^2 - q + 1$. Si verifica facilmente che

$$|\mu_1(q)| < 1 < |\mu_2(q)|, \qquad \lim_{q\to 0^-} \mu_1(q) = 1,$$

per cui $\mu_1(q)$ è la radice principale, $k_1 = k_2 = 1$ e la formula usata come metodo base in un BVM è A_{11}-stabile.
Si trova poi $c_4 = \frac{1}{24}$ e quindi $p = 3$. Avendosi $k_1 - 1 = 0$ e $k_2 = 1$ occorre un metodo ausiliario "di coda". Allo scopo si può usare la formula del terzo ordine

$$y_{n+2} - y_{n+1} = h(-\frac{1}{12}f_n + \frac{8}{12}f_{n+1} + \frac{5}{12}f_{n+2}),$$

per la quale si ha $c_4 = -\frac{1}{24}$.
Il sistema, riferito al problema $y' = f(t, y), y(t_0) = y_0$, risulta quindi essere

$$\begin{cases} y_1 - \frac{8}{12}hf_1 + \frac{1}{12}hf_2 - y_0 - \frac{5}{12}hf_0 & = 0 \\ -y_1 - \frac{5}{12}hf_1 + y_2 - \frac{8}{12}hf_2 + \frac{1}{12}hf_3 & = 0 \\ -y_2 - \frac{5}{12}hf_2 + y_3 - \frac{8}{12}hf_3 + \frac{1}{12}hf_4 & = 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \ldots \\ -y_{N-2} - \frac{5}{12}hf_{N-2} + y_{N-1} - \frac{8}{12}hf_{N-1} + \frac{1}{12}hf_N & = 0 \\ \frac{1}{12}hf_{N-2} - y_{N-1} - \frac{8}{12}hf_{N-1} + y_N - \frac{5}{12}hf_N & = 0 \end{cases}.$$

La sua risoluzione numerica può essere effettuata con il metodo di Newton. Alternativamente si osserva che tale sistema è, in realtà, semilineare in quanto le incognite $y_1, y_2, \ldots, y_N$ si presentano sia in forma lineare sia coinvolte non linearmente come argomento della funzione f. Si verifica subito che il sistema si lascia scrivere nella forma

$$By = hF(y) + c,$$

dove:

$$B = \begin{pmatrix} 1 & & \\ -1 & 1 & \\ & \ddots & \ddots \\ & -1 & 1 \end{pmatrix},$$

il vettore delle incognite è

$$y = (y_1^T, y_2^T, \ldots, y_N^T)^T,$$

le componenti del vettore $F(y)$ sono

$$\begin{array}{l} F_1(y) = \frac{8}{12}f_1 - \frac{1}{12}f_2, \\ F_{i+1}(y) = \frac{5}{12}f_i + \frac{8}{12}f_{i+1} - \frac{1}{12}f_{i+2}, \quad i = 1, 2, \ldots, N-2, \\ F_N(y) = -\frac{1}{12}f_{N-2} + \frac{8}{12}f_{N-1} + \frac{5}{12}f_N, \end{array}$$

e il vettore dei dei termini noti è dato da

$$c = ((y_0 + h\frac{5}{12}f_0)^T, 0^T, \ldots, 0^T)^T.$$

Resta quindi definito il procedimento iterativo

$$y^{(s+1)} = hB^{-1}F(y^{(s)}) + B^{-1}c, \quad s = 0, 1, \ldots,$$

essendo, come si riscontra immediatamente, la matrice inversa di B triangolare inferiore data da

$$B^{-1} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 1 & 1 & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ 1 & 1 & \ldots & 1 \end{pmatrix}.$$

Per h sufficientemente piccolo, in virtù del Teorema 4.6.1, tale procedimento risulta convergente e pertanto può adottarsi il criterio di arrestare le iterazioni allorché $\|y^{(s+1)} - y^{(s)}\| \leq \epsilon$ con ϵ opportunamente prefissato. □

Esempio 8.5.7 La formula trapezoidale, $y_{n+1} - y_n = \frac{h}{2}(f_{n+1} + f_n)$, come si è osservato alla fine del paragrafo 8.4.3, può essere usata come BVM da sola. Facendo riferimento al problema di valori iniziali lineare

$$y' = K(t)y, \quad y(t_0) = y_0,$$

dove $K(t) \in \mathrm{R}^{m\times m}$, ponendo $K(t_n) = K_n$, il metodo si può scrivere

$$y_n + \frac{1}{2}hK_n y_n - y_{n+1} + \frac{1}{2}hK_{n+1}y_{n+1} = 0, \quad n = 0, 1, \ldots, N-1.$$

Introducendo il vettore delle incognite $y = (y_1^T, y_2^T, \ldots, y_N^T)^T$, il vettore dei termini noti $c = ((-y_0 - \frac{1}{2}hK_0y_0)^T, 0^T, \ldots, 0^T)^T$ e la matrice

$$G = \begin{pmatrix} -I + \frac{1}{2}hK_1 & & & \\ I + \frac{1}{2}hK_1 & -I + \frac{1}{2}hK_2 & & \\ & \ddots & \ddots & \\ & & I + \frac{1}{2}hK_{N-1} & -I + \frac{1}{2}hK_N \end{pmatrix},$$

si è condotti alla risoluzione del sistema lineare $Gy = c$.
Si osserva poi che

$$det(G) = det(-I + \frac{1}{2}hK_1) \cdots det(-I + \frac{1}{2}hK_N),$$

per cui $\lim_{h\to 0} |det(G)| = 1$. Ne segue che per h sufficientemente piccolo (e $K(t)$ sufficientemente regolare) il sistema ha un'unica soluzione.
Si supponga ora $K(t) = K =$ costante. Risulta

$$det(G) = [det(-I + \frac{1}{2}hK)]^N = [(-1 + \frac{1}{2}h\lambda_1) \cdots (-1 + \frac{1}{2}h\lambda_m)]^N,$$

essendo $\lambda_1, \ldots, \lambda_m$ gli autovalori di K. Pertanto se gli autovalori sono reali non positivi o complessi si ha $det(G) \neq 0$ per qualunque valore di h. Se vi sono autovalori reali e positivi si potrà assumere $h < 2/\rho(K)$. □

Bibliografia: [4], [3], [6], [8], [16], [17], [18], [20], [21], [22], [24].