Capitolo 9

Metodi Runge-Kutta

I metodi Runge-Kutta (RK) costituiscono una alternativa ai metodi LMF per superare le barriere di ordine di Dahlquist. Li tratteremo in modo semplificato, applicandoli al problema autonomo

$$y'(t) = f(y(t)), \qquad y(0) = y_0 \in \mathbb{R}^m, \tag{9.1}$$

sebbene con questo non si perda in generalità, in quanto t può essere incluso nel vettore di stato.

9.1 Derivazione di un metodo Runge-Kutta

Tal teorema fondamentale del calcolo, si ottiene, in virtù della (9.1):

$$y(h) = y(0) + \int_0^h f(y(s))\mathrm{d}s.$$

Approssimando l'integrale con una formula di quadratura definita su s nodi $\{c_i\}$ e con pesi $\{b_i\}$, si ottiene

$$y_1 = y_0 + h\sum_{i=1}^{s} b_i f(Y_i), \tag{9.2}$$

dove i vettori

$$Y_i \approx y(c_i h), \qquad i = 1, \ldots, s,$$

sono detti *stadi* del metodo. Approssimando anche questi con corrispondenti formule di quadratura, si ottiene:

$$Y_i = y_0 + h\sum_{j=1}^{s} a_{ij} f(Y_j), \qquad i = 1, \ldots, s. \tag{9.3}$$

Le formule (9.2)–(9.3) defniscono un *metodo Runge-Kutta (RK, nel seguito) a s-stadi.* Questo è altresì caratterizzato dal cosiddetto *tableau di Butcher*:[1]

$$\begin{array}{c|c} \mathbf{c} & A \\ \hline & \mathbf{b}^T \end{array} \tag{9.4}$$

[1] In onore di John C. Butcher, che è stato il pioniere nello studio dei metodi RK.

in cui

$$\mathbf{b} = (b_1, \ldots, b_s)^T$$

è il *vettore dei pesi* della quadratura,

$$\mathbf{c} = (c_1, \ldots, c_s)^T$$

è il *vettore dei nodi*, e la matrice

$$A = (a_{ij}) \in \mathbb{R}^{s \times s}$$

è detta *matrice di Butcher* (o *Butcher array*). Riguardo alla complessità computazionale del metodo (9.2)-(9.3), si distinguono i seguenti tre casi significativi:

metodo RK esplicito: in questo caso la matrice A è strettamente triangolare inferiore (ovvero, $a_{ij} = 0$, se $j \geq i$). In tal caso, la (9.3) diviene

$$Y_i = y_0 + h \sum_{j=1}^{i-1} a_{ij} f(Y_j), \qquad i = 1, \ldots, s, \tag{9.5}$$

ovvero gli stadi $\{Y_i\}$ si ottengono in sequenza mediante sostituzioni successive;

metodo RK semi-implicito: in questo caso la matrice A è triangolare inferiore (ovvero, $a_{ij} = 0$, se $j > i$). In tal caso, la (9.3) diviene

$$Y_i - h a_{ii} f(Y_i) = y_0 + h \sum_{j=1}^{i-1} a_{ij} f(Y_j), \qquad i = 1, \ldots, s, \tag{9.6}$$

ovvero gli stadi $\{Y_i\}$ si ottengono risolvendo s equazioni nonlineari;

metodo RK implicito: in questo caso la matrice A non ha struttura di sparsità. In tal caso, la (9.3) diviene

$$Y - hA \otimes I_m f(Y) = \mathbf{e} \otimes y_0, \tag{9.7}$$

dove abbiamo indicato con

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_s \end{pmatrix}, \qquad f(Y) = \begin{pmatrix} f(Y_1) \\ \vdots \\ f(Y_s) \end{pmatrix}, \qquad \mathbf{e} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^s.$$

In tal caso, gli stadi $\{Y_i\}$ si ottengono risolvendo un sistema di s equazioni nonlineari.

Ci si convince facilmente che, sotto le usuali ipotesi che f sia continua e Lipschitziana in y, le equazioni (9.6) e (9.7) ammettono sempre soluzione, e questa è unica, per h sufficientemente piccolo.

9.1.1 Ordine di un metodo Runge-Kutta

Riguardo all'ordine di accuratezza, il metodo si dirà avere ordine p se, $y(h) - y_1 = O(h^{p+1})$ (assumendo che f sia di classe $C^{(p+1)}$). Si dimostra che l'ordine massimo di un metodo RK a s stadi non può eccedere $2s$. Tuttavia, le condizioni sui coefficienti del tableau (9.4) per ottenere un metodo di ordine p divengono assai complicate e numerose, al crescere dell'ordine

p. Infatti, se indichiamo con $n(p)$ il numero delle condizioni per esso richieste, si può costruire la seguente tabella [11, pag. 154]:

p	1	2	3	4	5	6	7	8	9	10
$n(p)$	1	2	4	8	17	37	85	200	486	1205

Un classico esempio di metodo Runge-Kutta esplicito è dato seguente metodo di ordine 4,[2] descritto dal *tableau*:

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

Si ottiene, pertanto,

$$\begin{aligned} Y_1 &= y_0, \\ Y_2 &= y_0 + \frac{h}{2} f(Y_1), \\ Y_3 &= y_0 + \frac{h}{2} f(Y_2), \\ Y_4 &= y_0 + h f(Y_3), \\ y_1 &= y_0 + \frac{h}{6}(f(Y_1) + 2f(Y_2) + 2f(Y_3) + f(Y_4)). \end{aligned}$$

9.1.2 Analisi di stabilità lineare per un metodo Runge-Kutta

Un metodo Runge-Kutta è 0-stabile per costruzione. Infatti, si tratta di un metodo *one-step* per cui (vedi (9.2)) $\rho(z) = z - 1$. Se si considera, invece, l'equazione test

$$y' = \lambda y, \qquad \Re(\lambda) < 0,$$

e ponendo, al solito, $q = h\lambda$, dalle (9.2) e (9.3) si ottiene che:

$$\begin{pmatrix} I - qA & 0 \\ -q\mathbf{b}^T & 1 \end{pmatrix} \begin{pmatrix} Y \\ y_1 \end{pmatrix} = \begin{pmatrix} \mathbf{e} \\ 1 \end{pmatrix} y_0.$$

Applicando la regola di Kramer si ottiene, pertanto:

$$y_1 = \frac{\det \begin{pmatrix} I - qA & \mathbf{e} \\ -q\mathbf{b}^T & 1 \end{pmatrix}}{\det(I - qA)} y_0 = \frac{\det\left(I - qA + q\mathbf{e}\mathbf{b}^T\right)}{\det(I - qA)} y_0 \equiv R(q) y_0.$$

La funzione razionale $R(q)$ è detta *funzione di stabilità* del metodo. La sua regione di assoluta stabilità sarà pertanto definita da

$$\mathcal{D} = \{q \in \mathbb{C} : |R(q)| < 1\}.$$

Il metodo sarà, al solito, A-stabile se $\mathbb{C}^- \subseteq \mathcal{D}$.

[2] Detto, appunto, *Runge-Kutta classico*.

Osservazione 9.1 *Osserviamo che, per un metodo esplicito, $R(q)$ è un polinomio di grado s e, pertanto, $R(q) \to \infty$ per $q \to \infty$. Pertanto, la regione di assoluta stabilità di un metodo RK esplicito è limitata.*[3] *Questo sarà altresì vero quando il grado del denominatore di $R(q)$ è inferiore a quello del numeratore.*

Per ottenere metodi RK A-stabili, che saranno necessariamente impliciti o semi-impliciti, come appena argomentato, sono richieste ulteriori condizioni sui coefficienti del tableau (9.4): tuttavia, le cosiddette *formule di Gauss-Legendre* a s stadi raggiungono l'ordine massimo $2s$, e risultano essere perfettamente A-stabili.[4]

In generale, è però difficile ottenere in modo immediato dei metodi RK A-stabili. Cercheremo di ovviare a questo inconveniente, utilizzando un diverso approccio per definire dei metodi di approssimazione per il problema (9.1).

9.2 Sviluppo di Fourier locale

Ai fini della nostra trattazione, introduciamo la seguente famiglia di polinomi:[5]

$$\begin{aligned} &P_0(t) \equiv 1, \qquad P_1(t) = \sqrt{3}(t-1), \\ &P_{i+1}(t) = (2t-1)\frac{2i+1}{i+1}\sqrt{\frac{2i+3}{2i+1}}P_i(t) - \frac{i}{i+1}\sqrt{\frac{2i+3}{2i-1}}P_{i-1}(t), \qquad i \geq 1. \end{aligned} \tag{9.8}$$

Per essi vale il seguente risultato, che si dimostra per induzione.

Teorema 9.1 *I polinomi definiti in (9.8) soddisfano:*

$$\deg P_i = i, \qquad \int_0^1 P_i(t)P_j(t)\mathrm{d}t = \delta_{ij}, \qquad \forall i,j \geq 0. \tag{9.9}$$

In virtù della (9.9), i polinomi $\{P_i\}$ costituiscono una *famiglia di polinomi ortonormali sull'intervallo [0,1]*, rispetto al prodotto scalare definito dalla seconda equazione in (9.9). Essi costituiscono una base per le funzioni continue definte nell'intervallo $[0,1]$. Vale, inoltre, il seguente risultato preliminare.

Teorema 9.2 *Sia V uno spazio vettoriale e $g : [0,h] \to V$ sufficientemente regolare. Allora*

$$\int_0^1 P_j(t)g(th)\mathrm{d}t = O(h^j), \qquad j \geq 0. \tag{9.10}$$

Dimostrazione. Assumeremo, per semplicità, che g sia sviluppabile in serie di Taylor. Segue quindi che, $\forall j \geq 0$:

$$\begin{aligned} \int_0^1 P_j(t)g(th)\mathrm{d}t &= \int_0^1 P_j(t)\sum_{k\geq 0}\frac{g^{(k)}(0)}{k!}(th)^k\mathrm{d}t = \sum_{k\geq 0}\frac{g^{(k)}(0)}{k!}h^k\int_0^1 P_j(t)t^k\mathrm{d}t \\ &= \sum_{k\geq j}\frac{g^{(k)}(0)}{k!}h^k\int_0^1 P_j(t)t^k\mathrm{d}t = O(h^j), \end{aligned}$$

[3] Conseguentemente, non esistono metodi RK espliciti A-stabili.
[4] Ovvero, la loro regione di assoluta stabilità coincide con $\mathbb{C}^-$.
[5] Si tratta dei *polinomi di Legendre* scalati e *shiftati* sull'intervallo $[0,1]$.

dove la penultima uguaglianza segue dalla (9.9). □

Avremo anche bisogno del seguente risultato relativo alle equazioni differenziali.

Lemma 9.1 *Sia $y(t;t_0,y_0)$ la soluzione del problema ai valori iniziali*

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

Allora:

$$\frac{\partial}{\partial y_0} y(t;t_0,y_0) = \Phi(t,t_0), \qquad \frac{\partial}{\partial t_0} y(t;t_0,y_0) = -\Phi(t,t_0) f(y_0),$$

dove $\Phi(t,t_0)$ è la matrice fondamentale del problema variazionale associato:

$$\Phi'(t,t_0) = J_f(y(t;t_0,y_0))\Phi(t,t_0), \qquad \Phi(t_0,t_0) = I.$$

Dimostrazione. Per dimostrare la prima parte della tesi, si consideri una perturbazione infinitesima δy_0 della condizione iniziale. Pertanto, la soluzione del problema perturbato soddisferà il problema variazionale, la cui soluzione è

$$\delta y(t) \equiv y(t;t_0,y_0+\delta y_0) - y(t;t_0,y_0) = \Phi(t,t_0)\delta y_0.$$

Pertanto,

$$\frac{\partial}{\partial y_0} y(t;t_0,y_0) = \Phi(t,t_0).$$

Per la dimostrare la seconda parte della tesi, fissato $\varepsilon \approx 0$, osserviamo che, denotando con $y(t) \equiv y(t;t_0,y_0)$, allora evidentemente risulterà

$$y(t;t_0+\varepsilon,y_0) \equiv y(t-\varepsilon;t_0,y_0). \tag{9.11}$$

Ricerchiamo ora la condizione iniziale, sia essa $y_0(\varepsilon)$, tale che si abbia:

$$y(t;t_0+\varepsilon,y_0) \equiv y(t;t_0,y_0(\varepsilon)).$$

Al primo ordine, tenendo conto della (9.11), si avrà: $y_0(\varepsilon) = y_0 - \varepsilon f(y_0)$. Segue quindi che:

$$\begin{aligned}\frac{\partial}{\partial t_0} y(t;t_0,y_0) &= \lim_{\varepsilon\to 0} \frac{y(t;t_0+\varepsilon,y_0) - y(t;t_0,y_0)}{\varepsilon} \\ &= \frac{\partial}{\partial y_0} y(t;t_0,y_0(\varepsilon))\Big|_{\varepsilon=0} \frac{\mathrm{d}}{\mathrm{d}\varepsilon} y_0(\varepsilon) = \Phi(t,t_0)\left[-f(y_0)\right]. \square\end{aligned}$$

Avendo premesso questi risultati, consideriamo il seguente sviluppo di Fourier locale del problema (9.1) sull'intervallo $[0,h]$, lungo la base definita dai polinomi (9.8), che è sicuramente definito per h sufficientemente piccolo e f sufficientemente regolare:[6]

$$\begin{aligned} y'(ch) &= \sum_{j\geq 0} P_j(c)\gamma_j(y), \qquad c \in [0,1], \\ \gamma_j(y) &= \int_0^1 P_j(\tau) f(y(\tau h))\mathrm{d}\tau, \qquad j \geq 0. \end{aligned} \tag{9.12}$$

[6] Si può dimostrare che è sufficiente che si abbia $f \in C^{(1)}$.

Otterrmo una approssimazione polinomiale, sia essa $\sigma \in \Pi_s$, troncando questo sviluppo in serie:

$$\begin{aligned} \sigma'(ch) &= \sum_{j=0}^{s-1} P_j(c)\gamma_j(\sigma), \qquad c \in [0,1], \\ \gamma_j(\sigma) &= \int_0^1 P_j(\tau) f(\sigma(\tau h))\mathrm{d}\tau, \qquad j = 0, \ldots, s-1. \end{aligned} \tag{9.13}$$

Imponendo che sia soddisfatta la condizione iniziale $\sigma(0) = y_0$, si ottiene, quindi,

$$\sigma(ch) = y_0 + h\sum_{j=0}^{s-1} \int_0^c P_j(x)\mathrm{d}x\, \gamma_j(\sigma), \qquad c \in [0,1]. \tag{9.14}$$

La approssimazione a $y(h)$ è quindi definita, in virtù delle (9.9)-(9.14), come

$$\sigma(h) \equiv y_0 + h\gamma_0(\sigma) = y_0 + h\int_0^1 f(\sigma(\tau h))\mathrm{d}\tau \equiv y_0 + \int_0^h f(\sigma(x))\mathrm{d}x. \tag{9.15}$$

Supponedo per semplicità, nel seguito, che f sia sviluppabile in serie di Taylor, valgono i seguenti risultati.

Lemma 9.2 $\gamma_j(\sigma) = O(h^j), \qquad j \geq 0.$

Dimostrazione. La dimostrazione discende immediatamente dal Teorema 9.2. □

Teorema 9.3 $\sigma(h) - y(h) = O(h^{2s+1})$.

Dimostrazione. Denoteremo, al solito, con $y(t; t_0, y_0)$ la soluzione di (9.1). In virtù dei Lemmi 9.1 e 9.2, e dalle (9.12)-(9.13), segue quindi che:

$$\begin{aligned} \sigma(h) - y(h) &= y(h; h, \sigma(h)) - y(h; 0, \sigma(0)) = \int_0^h \frac{\mathrm{d}}{\mathrm{d}t} y(h; t, \sigma(t))\mathrm{d}t \\ &= \int_0^h \frac{\partial}{\partial t} y(h; t, \sigma(t)) + \frac{\partial}{\partial \sigma} y(h; t, \sigma(t))\sigma'(t)\mathrm{d}t \\ &= h\int_0^1 \Phi(h, \tau h)[-f(\sigma(\tau h)) + \sigma'(\tau h)]\mathrm{d}\tau \\ &= -h\int_0^1 \Phi(h, \tau h)\sum_{j\geq s} P_j(\tau)\gamma_j(\sigma)\mathrm{d}\tau = -h\sum_{j\geq s} \underbrace{\left[\int_0^1 \overbrace{\Phi(h, \tau h)}^{\equiv g(\tau h)} P_j(\tau)\mathrm{d}\tau\right]}_{=O(h^j)} \overbrace{\gamma_j(\sigma)}^{=O(h^j)} \\ &= h\sum_{j\geq s} O(h^{2j}) = O(h^{2s+1}). \square \end{aligned}$$

Osserviamo che il metodo definito dalle (9.13)-(9.15) non è ancora un metodo operativo, perchè gli integrali per ottenere i $\gamma_j(\sigma)$ devono essere approssimati mediante una idonea formula di quadratura. A questo fine, consideriamo le k ascisse

$$0 \leq c_1 < c_2 < \ldots < c_k \leq 1, \tag{9.16}$$

ed i corrispondenti pesi della quadratura definita su queste ascisse,

$$b_i = \int_0^1 \prod_{j=1,\ j\neq i}^{k} \frac{t-c_j}{c_i-c_j}\mathrm{d}t, \qquad i=1,\ldots,k, \tag{9.17}$$

ottenendo, pertanto,

$$\int_0^1 P_j(\tau) f(\sigma(\tau h))\mathrm{d}\tau \approx \sum_{i=1}^{k} b_i P_j(c_i) f(\sigma(c_i h)), \qquad j=0,\ldots,s-1. \tag{9.18}$$

Chiaramente, così facendo si ottiene una nuova approssimazione polinomiale, sia essa u, definita dalla seguente equazione:

$$\begin{aligned} u(ch) &= y_0 + h\sum_{j=0}^{s-1} \int_0^c P_j(x)\mathrm{d}x\,\hat{\gamma}_j, \qquad c\in[0,1], \\ \hat{\gamma}_j &= \sum_{i=1}^{k} b_i P_j(c_i) f(u(c_i h)), \qquad j=0,\ldots,s-1. \end{aligned} \tag{9.19}$$

La approssimazione a $y(h)$ è quindi definita, in virtù delle (9.9)-(9.19), come

$$y_1 = u(h) \equiv y_0 + h\hat{\gamma}_0 = y_0 + h\sum_{i=1}^{k} b_i f(u(c_i h)). \tag{9.20}$$

Definendo

$$Y_i = u(c_i h), \qquad i=1,\ldots,k,$$

dalla (9.19) si ottiene:

$$\begin{aligned} Y_i &= y_0 + h\sum_{j=0}^{s-1} \int_0^{c_i} P_j(x)\mathrm{d}x \sum_{\ell=1}^{k} b_\ell P_j(c_\ell) f(Y_\ell) \\ &\equiv y_0 + h\sum_{j=1}^{k} \left[b_j \sum_{\ell=0}^{s-1} P_\ell(c_j) \int_0^{c_i} P_\ell(x)\mathrm{d}x \right] f(Y_j), \qquad i=1,\ldots,k, \end{aligned} \tag{9.21}$$

mentre la (9.20) diviene:

$$y_1 = y_0 + h\sum_{i=1}^{k} b_i f(Y_i). \tag{9.22}$$

È evidente che le (9.21)-(9.22) definiscono il seguente metodo RK a k-stadi:

$$\begin{array}{c|c} c_1 & \\ \vdots & A=(a_{ij}) \equiv \left(b_j \sum_{\ell=0}^{s-1} P_\ell(c_j) \int_0^{c_i} P_\ell(x)\mathrm{d}x\right) \\ c_k & \\ \hline & b_1 \quad \ldots \quad b_k \end{array} \tag{9.23}$$

Osservazione 9.2 *Come vedremo nel seguito, con una opportuna scelta delle ascisse (9.16), (9.23) definisce la forma Runge-Kutta di un metodo HBVM(k,s).*[7]

[7] Si tratta dell'acronimo di *Hamiltonian Boundary Value Method* a k stadi e grado s.

9.2.1 Ordine del metodo

Al fine di studiare l'ordine del metodo RK definito dal *tableau* (9.23) si premette il seguente risultato, riguardante l'errore della formula di quadratura (9.16)-(9.17), che assumeremo avere *ordine* q, ovvero esatta per polinomi di grado non maggiore di $q-1$.

Lemma 9.3 *Sia $f \in C^{(q)}$. Allora*

$$\int_0^1 P_j(\tau) f(\tau h) \mathrm{d}\tau - \sum_{i=1}^{k} b_i P_j(c_i) f(c_i h) = O(h^{q-j}), \qquad j = 0, \ldots, q.$$

Dimostrazione. Poiché la formula di quadratura è esatta per polinomi di grado $q-1$, l'errore dipenderà dalla derivata q-esima dell'integrando. La tesi segue considerando che

$$\begin{aligned} \frac{\mathrm{d}^q}{\mathrm{d}\tau^q} P_j(\tau) f(\tau h) &= [P_j(\tau) f(\tau h)]^{(q)} = \sum_{i=0}^{q} \binom{q}{i} P_j^{(i)}(\tau) f^{(q-i)}(\tau h) h^{q-i} \\ &= \sum_{i=0}^{j} \binom{q}{i} P_j^{(i)}(\tau) f^{(q-i)}(\tau h) h^{q-i} = O(h^{q-j}), \end{aligned}$$

perché $P_j^{(i)}(\tau) \equiv 0$, per $i > j$. □

Come conseguenza, si ottiene immediatamente il seguente risultato.

Corollario 9.1 *Se l'ordine della quadratura (9.16)-(9.17) è q, allora (vedi (9.13)-(9.19)),*

$$\hat{\gamma}_j = \gamma_j(u) - \Delta_j(h), \qquad \Delta_j(h) = O(h^{q-j}). \tag{9.24}$$

Possiamo ora discutere l'ordine del metodo (9.23).

Teorema 9.4 *Sia q l'ordine della quadratura (9.16)-(9.17). Allora $y_1 - y(h) = O(h^{p+1})$, con $p = \min(q, 2s)$.*

Dimostrazione. Procedendo in modo analogo a quanto visto nella dimostrazione del Teorema 9.3, si ha, in virtù delle (9.19)-(9.24):

$$\begin{aligned} y_1 - y(h) &\equiv u(h) - y(h) = y(h; h, u(h)) - y(h; 0, y_0) \\ &\equiv y(h; h, u(h)) - y(h; 0, u(0)) = \int_0^h \frac{\mathrm{d}}{\mathrm{d}t} y(h; t, u(t)) \mathrm{d}t \\ &= \int_0^h \frac{\partial}{\partial t} y(h; t, u(t)) + \frac{\partial}{\partial u} y(h; t, u(t)) u'(t) \mathrm{d}t \\ &= h \int_0^1 \Phi(h, \tau h)[-f(u(\tau h)) + u'(\tau h)] \mathrm{d}\tau \\ &= h \int_0^1 \Phi(h, \tau h) \left[-\sum_{j \geq 0} P_j(\tau) \gamma_j(u) + \sum_{j=0}^{s-1} P_j(\tau)(\gamma_j(u) - \Delta_j) \right] \mathrm{d}\tau \\ &= h \int_0^1 \Phi(h, \tau h) \left[\sum_{j=0}^{s-1} P_j(\tau) \Delta_j - \sum_{j \geq s} P_j(\tau) \gamma_j(u) \right] \mathrm{d}\tau \end{aligned}$$

$$= h\sum_{j=0}^{s-1}\underbrace{\left[\int_0^1 \overbrace{\Phi(h,\tau h)}^{\equiv g(\tau h)} P_j(\tau)\mathrm{d}\tau\right]}_{=O(h^j)} \overbrace{\Delta_j}^{=O(h^{q-j})} - h\sum_{j\geq s}\underbrace{\left[\int_0^1 \overbrace{\Phi(h,\tau h)}^{\equiv g(\tau h)} P_j(\tau)\mathrm{d}\tau\right]}_{=O(h^j)} \overbrace{\gamma_j(u)}^{=O(h^j)}$$

$$= hO(h^q) - h\sum_{j\geq s} O(h^{2j}) = O(h^{q+1}) + O(h^{2s+1}) \equiv O(h^{p+1}),$$

avendo posto $p = \min(q, 2s)$. □

Corollario 9.2 *Se l'ordine della quadratura (9.16)-(9.17) è $q \geq 2s$, Allora il metodo (9.23) ha ordine $2s$.*

Nel seguito, assumeremo che sia $q \geq 2s$.

9.2.2 Costo computazionale

Dall'analisi fatta, si intuisce che è auspicabile che la quadratura (9.16)-(9.17) abbia ordine $q \geq 2s$. Questo potrebbe essere ottenuto ponendo $k = s$ e scegliendo le ascisse (9.16) come le radici del polinomio $P_s(x)$. Infatti, questa scelta origina la formula di quadratura di Gauss-Legendre di ordine $2s$.

Con questa scelta si ottengono i metodi *RK di Gauss-Legendre a s stadi.*
Essi sono impliciti e raggiungono l'ordine massimo $2s$.

Riportiamo i *tableau di Butcher* dei metodi RK di Gauss per $s = 1, 2$:

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array} \qquad \begin{array}{c|cc} \frac{1}{2}-\frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4}-\frac{\sqrt{3}}{6} \\ \frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{1}{4}+\frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Il primo di questi metodi, che può essere riscritto come

$$y_1 = y_0 + hf\left(\frac{y_0+y_1}{2}\right), \tag{9.25}$$

è detto *mid-point implicito* e ha ordine 2. Infatti, esso è il seguente metodo RK a 1 stadio,

$$Y = y_0 + \frac{h}{2}f(Y), \qquad y_1 = y_0 + hf(Y).$$

Moltiplicando la prima equazione per 2, e sottraendovi la seconda, si ottiene, infatti, $Y = (y_0 + y_1)/2$ che, sostituita nella seconda, fornisce la (9.25). Osserviamo che, nel caso la f sia lineare, allora la (9.25) si riduce al metodo dei trapezi. Come vedremo successivamente, può essere talora opportuno che la fomula di quadratura abbia ordine maggiore di $2s$: ponendo le $k \geq s$ ascisse (9.16) nelle radici di $P_k(x)$, si ottiene una formula di quadratura di Gauss-Legendre di ordine $2k$ che, pertanto, può essere resa di ordine arbitrariamente elevato, scegliendo k abbastanza grande.

Definizione 9.1 *Denomineremo questi metodi HBVM(k, s). Essi sono definiti per $k \geq s$.*

Osservazione 9.3 *HBVM è l'acronimo di* "Hamiltonian Boundary Value Methods" *[3]. Questa denominazione sarà giustificata nel seguito.*

In virtù del Corollario 9.2, questi metodi possono essere riguardati come una generalizzazione delle formule di Gauss-Legendre ad s stadi, che raggiungono l'ordine massimo $2s$. Tuttavia, trattandosi di metodi RK a k stadi (si veda il *tableau* (9.23)), il costo computazionale per la loro implementazione, che consiste nel risolvere ad ogni passo un sistema di k equazioni nonlineari nella forma (9.7), aumenta al crescere di k.

In realtà, vedremo come sia *sempre* possibile riformulare il problema discreto (9.7) in modo tale che esso abbia *dimensione s, indipendentemente dal valore di k considerato.* Questo implica che la complessità computazionale di un metodo HBVM(k,s) dipende essenzialmente da s, per ogni $k \geq s$. Esso sarà, pertanto, assimilabile a quello della corridpondente formula RK di Gauss-Legendre di ordine $2s$. Si premette il seguente risultato.

Teorema 9.5 *La matrice del* tableau *di Butcher (9.23) può essere scritta nella forma*

$$A = \mathcal{I}\mathcal{P}^T\Omega, \tag{9.26}$$

dove:

$$\mathcal{I} = \left(\int_0^{c_i} P_{j-1}(\tau)\mathrm{d}\tau\right)_{\substack{i=1,\ldots,k\\ j=1,\ldots,s}}, \quad \mathcal{P} = \left(P_{j-1}(c_i)\right)_{\substack{i=1,\ldots,k\\ j=1,\ldots,s}}, \quad \Omega = \mathrm{diag}(b_1,\ldots,b_k).$$

Dimostrazione. La dimostrazione si ottiene uguagliando il generico elemento a_{ij} delle due matrici (9.26) e (9.23). □

In virtù della (9.26), il sistema nonlineare (9.7) può essere scritto nella forma:

$$Y = \mathbf{e} \otimes y_0 - h\mathcal{I}\mathcal{P}^T\Omega \otimes I_m f(Y). \tag{9.27}$$

Definendo il vettore a blocchi di dimensione s contenente i coefficienti $\hat{\gamma}_j$, incogniti, nella (9.19),

$$\boldsymbol{\gamma} = \begin{pmatrix} \hat{\gamma}_0^T & \cdots & \hat{\gamma}_{s-1}^T \end{pmatrix}^T,$$

si verifica facilmente, da quest'ultima equazione, che

$$\boldsymbol{\gamma} = \mathcal{P}^T\Omega \otimes I_m f(Y). \tag{9.28}$$

Dalle precedenti equazioni (9.27)-(9.28) si ottiene, pertanto,

$$\boldsymbol{\gamma} = \mathcal{P}^T\Omega \otimes I_m f\left(\mathbf{e} \otimes y_0 - h\mathcal{I} \otimes I_m\boldsymbol{\gamma}\right), \tag{9.29}$$

che è un *sistema nonlineare di s equazioni, qualunque sia il valore scelto per k.* Una volta risolta (9.29), la nuova approssimazione sarà data da

$$y_1 = y_0 + h\hat{\gamma}_0. \tag{9.30}$$

Dalle (9.29)-(9.30) si può pertanto concludere che la complessità computazionale per l'implementazione di un metodo HBVM(k,s) è essenzialmente dipendente da s, qualunque sia il valore scelto di k. Questo aspetto sarà molto importante, quando tratteremo di problemi Hamiltoniani.

9.2.3 Analisi di stabilità lineare

Per l'analisi di stabilità lineare, consideriamo, al solito, l'equazione test

$$y' = \lambda y, \qquad \Re(\lambda) < 0.$$

Ponendo

$$\lambda = \alpha + i\beta, \qquad y = x_1 + ix_2,$$

l'equazione test diviene:

$$\mathbf{x}' \equiv \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}' = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \equiv A\mathbf{x}, \qquad \mathbf{x}(0) = \mathbf{x}_0 \neq \mathbf{0}. \tag{9.31}$$

Definendo la funzione a valori scalari

$$V(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{x} \equiv \frac{1}{2}\|\mathbf{x}\|_2^2, \tag{9.32}$$

si ottiene che l'applicazione di un metodo HBVM(k,s) per risolvere (9.31) produce il polinomio σ tale che $\sigma(0) = \mathbf{x}(0)$ e, inoltre,

$$\begin{aligned} \sigma'(ch) &= \sum_{j=0}^{s-1} P_j(c) \sum_{i=1}^{k} b_i P_j(c_i) A\sigma(c_i h) = A \sum_{j=0}^{s-1} P_j(c) \sum_{i=1}^{k} b_i \overbrace{P_j(c_i)\sigma(c_i h)}^{\text{grado} \le 2s-1} \\ &\equiv A \sum_{j=0}^{s-1} P_j(c) \int_0^1 P_j(\tau)\sigma(\tau h)\mathrm{d}\tau \equiv A \sum_{j=0}^{s-1} P_j(c) \int_0^1 P_j(\tau)\nabla V(\sigma(\tau h))\mathrm{d}\tau. \end{aligned} \tag{9.33}$$

Questo implica che (9.32) sia una funzione di Lyapunov per il sistema dinamico discreto indotto dal metodo. Infatti, ponendo $\sigma(0) = \mathbf{x}_0$, e la nuova approssimazione data da $\mathbf{x}_1 \equiv \sigma(h)$, si ottiene, tenendo conto del fatto che

$$A = \alpha I_2 + \beta J_2, \qquad J_2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = -J_2^T,$$

$$\begin{aligned} \Delta V(\mathbf{x}_0) &= V(\mathbf{x}_1) - V(\mathbf{x}_0) = V(\sigma(h)) - V(\sigma(0)) \\ &= \int_0^h \frac{\mathrm{d}}{\mathrm{d}t} V(\sigma(t))\mathrm{d}t = \int_0^h \nabla V(\sigma(t))^T \sigma'(t)\mathrm{d}t \\ &= h \int_0^1 \nabla V(\sigma(\tau h))^T A \sum_{j=0}^{s-1} P_j(\tau) \left[\int_0^1 P_j(c)\nabla V(\sigma(ch))\mathrm{d}c \right] \mathrm{d}\tau \\ &= h \sum_{j=0}^{s-1} \left[\int_0^1 P_j(c)\nabla V(\sigma(ch))\mathrm{d}c \right]^T A \left[\int_0^1 P_j(c)\nabla V(\sigma(ch))\mathrm{d}c \right] \\ &= \alpha h \sum_{j=0}^{s-1} \left\| \int_0^1 P_j(\tau)\nabla V(\sigma(\tau h))\mathrm{d}\tau \right\|_2^2 = \alpha h \sum_{j=0}^{s-1} \left\| \int_0^1 P_j(\tau)\sigma(\tau h)\mathrm{d}\tau \right\|_2^2 \\ &\equiv \alpha h \Gamma. \end{aligned}$$

Osserviamo che, poichè $\sigma(t) \not\equiv \mathbf{0}$, deve aversi $\Gamma > 0$. Diversamente, avendosi $\sigma \in \Pi_s$, si avrebbe $\sigma(ch) = \rho P_s(c)$ con $\rho \neq 0$ che, tuttavia, in virtù della (9.33), darebbe $P_s'(c) \equiv 0$, $c \in [0,1]$ che è, evidentemente, falso. Pertanto $\Gamma > 0$ e, quindi,

$$\|\mathbf{x}_1\|_2^2 = \|\mathbf{x}_0\|_2^2 + \alpha h \Gamma < \|\mathbf{x}_0\|_2^2 \qquad \Leftrightarrow \qquad \alpha = \Re(\lambda) < 0.$$

Pertanto, un HBVM(k,s) risulta essere perfettanemte A-stabile, in quanto la sua regione di assoluta stabilità coincide con $\mathbb{C}^-$, per ogni $k \geq s \geq 1$.

Osservazione 9.4 *In realtà, qualunque scelta delle ascisse che dia una formula di quadratura sufficientemente accurata può essere utilizzata per definire un metodo HBVM. Ad esempio, nel caso $s = 1$, e scegliendo delle ascisse equidistanti, si ottengono i seguenti metodi, denominati* k-stage trapezoidal rules, *in quanto generano una generalizzazione della formula di base dei trapezi. In questo caso, si vede facilmente che*

$$u(ch) = (1-c)y_0 + cy_1, \qquad c \in [0,1].$$

Scegliendo delle k ascisse equidistanti,

$$c_i = (i-1)/(k-1), \qquad i = 1, \ldots, k,$$

ed i pesi dati, al solito, da (9.17), si ottiene:

$k = 2$: *è la formula di base dei trapezi,*

$$y_1 = y_0 + \frac{h}{2}\left(f(y_0) + f(y_1)\right).$$

$k = 3$:

$$y_1 = y_0 + \frac{h}{6}\left(f(y_0) + 4f\left(\frac{y_0+y_1}{2}\right) + f(y_1)\right).$$

$k = 5$:

$$y_1 = y_0 + \frac{h}{90}\left(7f(y_0) + 32f\left(\frac{3y_0+y_1}{4}\right) + 12f\left(\frac{y_0+y_1}{2}\right) + 32f\left(\frac{y_0+3y_1}{4}\right) + 7f(y_1)\right).$$

In generale, il tableau di Butcher *corrispondente a queste formule si vede essere dato da:*

$$\begin{array}{c|c} \mathbf{c} & \mathbf{c}\mathbf{b}^T \\ \hline & \mathbf{b}^T \end{array}$$

Il corrispondente problema discreto può essere quindi riscritto come

$$\begin{aligned} \hat{\gamma}_0 &= \mathbf{b}^T \otimes I_m\, f\left(\mathbf{e} \otimes y_0 + h\mathbf{c} \otimes \hat{\gamma}_0\right), \\ y_1 &= y_0 + h\hat{\gamma}_0. \end{aligned}$$

9.3 Problemi Hamiltoniani

Molti problemi derivanti dalle applicazioni, sono problemi del secondo ordine del tipo:

$$q'' = \nabla U(q) \in \mathbb{R}^m, \tag{9.34}$$

in cui $U : \mathbb{R}^m \to \mathbb{R}$. Ponendo

$$p = q', \qquad H(q,p) = \frac{1}{2}p^T p - U(q) \equiv H(y), \qquad y = \begin{pmatrix} q \\ p \end{pmatrix}, \tag{9.35}$$

l'equazione (9.34) può essere scritta in forma del primo ordine come

$$q' = p, \quad p' = \nabla U(q), \qquad \Leftrightarrow \qquad y' = J\nabla H(y), \qquad J = \begin{pmatrix} O & I_m \\ -I_m & O \end{pmatrix}. \tag{9.36}$$

Quest'ultima è la *forma Hamiltoniana* del problema originale. La funzione $H(y)$ nella (9.35) è la *funzione Hamiltoniana* che definisce il problema. Quando il problema Hamiltoniano deriva dalla modellizzazione di un sistema meccanico isolato, la funzione Hamiltoniana rappresenta generalmente la sua energia totale: per questo motivo, spesso si fa riferimento alla funzione Hamiltoniana come all'"energia" del sistema. Osserviamo che la matrice J nella (9.36) è ortogonale ed antisimmetrica:

$$J^T = -J, \qquad J^T J = I.$$

Per un problema Hamiltoniano (9.36), l'Hamiltoniana si mantiene costante lungo una traiettoria, ovvero:

$$H(y(t)) \equiv H(y(0)), \qquad \forall t \geq 0.$$

Infatti:

$$\frac{\mathrm{d}}{\mathrm{d}t}H(y(t)) = \nabla H(y(t))^T y'(t) = \nabla H(y(t))^T J \nabla H(y(t)) = 0,$$

essendo J antisimmetrica. Visto il significato fisico, è auspicabile che questa caratteristica sia preservata da un metodo di approssimazione. Quando questo non avviene, la dinamica riprodotta dal metodo numerico potrebbe essere non corretta, come illustrato nel seguente esempio.

Il problema di Keplero

Si tratta del moto relativo (planare) di due corpi di massa unitaria che si muovono in interazione gravitazionale tra loro. Ponendo uno dei corpi nell'origine del sistema di riferimento, e dopo normalizzazione delle costanti fisiche, si ottiene la funzione Hamiltoniana

$$H(q,p) = \frac{1}{2}\|p\|_2^2 - \frac{1}{\|q\|_2}, \qquad q, p \in \mathbb{R}^2,$$

in cui il vettore q contiene le coordinate del secondo corpo. In particolare, partendo dal punto iniziale

$$y_0 = (q_0^T, p_0^T) = \begin{pmatrix} 1-\varepsilon & 0 & 0 & \sqrt{\frac{1+\varepsilon}{1-\varepsilon}} \end{pmatrix}, \qquad \varepsilon \in [0,1),$$

la soluzione è periodica di periodo $T = 2\pi$ e, nel piano (q_1, q_2) risulta essere un'ellisse di eccentricità ε. L'Hamiltoniana, lungo questa traiettoria, ha un valore costante pari a $-\frac{1}{2}$. Utilizzando il metodo di Eulero esplicito per approssimare questa traiettoria, con un passo $h = 10^{-3}T$, si ottiene la soluzione numerica riprodotta in Figura 9.1. Quest'ultima è chiaramente errata, e questo è confermato dal fatto che l'errore nell'Hamiltoniana cresce, come si può osservare in Figura 9.2.[8]

[8] Questo fenomeno è denominato *drift numerico dell'Hamiltoniana*, o *"energy drift"*.

Le cose migliorano notevolmente utilizzando il *mid-point* implicito (9.25):[9] in questo caso, utilizzando lo stesso passo h, la traiettoria è qualitativamente corretta, come si evince da Figura 9.3, e questo ha un preciso riscontro nel fatto che l'errore nell'Hamiltoniana numerica è limitato, come si può vedere in Figura 9.4. Tuttavia, sebbene non vi sia un *drift*, l'ampiezza dell'errore dell'Hamiltoniana sarà in questo caso proporzionale all'ordine del metodo.

Conservazione dell'Hamiltoniana per un metodo HBVM(k,s)

Consideriamo il valore dell'Hamiltoniana numerica in $y_1 = u(h)$, vedi (9.19), per un metodo HBVM(k,s), in cui le k ascisse $\{c_i\}$ sono poste nelle radici di $P_k(x)$. In questo modo, come abbiamo visto innanzi,

$$\begin{aligned} \hat{\gamma}_j &= \sum_{i=1}^{k} P_j(c_i) J\nabla H(u(c_i h)) = \gamma_j(u) - \Delta_j(h), \\ \Delta_j(h) &= O(h^{2k-j}), \qquad j = 0,\dots,s-1, \end{aligned} \tag{9.37}$$

avendo, al solito, denotato con ($f = J\nabla H$, in questo caso)

$$\gamma_j(u) = \int_0^1 P_j(\tau) J\nabla H(u(\tau h))\mathrm{d}\tau = O(h^j), \qquad j = 0,\dots,s-1. \tag{9.38}$$

Si ottiene, pertanto, che:

$$\begin{aligned} H(y_1) - H(y_0) \;\equiv\; H(u(h)) - H(u(0)) \;&=\; \int_0^h \nabla H(u(t))^T u'(t)\mathrm{d}t \\ &= h\int_0^1 \nabla H(u(\tau h))^T \sum_{j=0}^{s-1} P_j(\tau)\hat{\gamma}_j \mathrm{d}\tau \;=\; h\sum_{j=0}^{s-1} \gamma_j(u)^T J\hat{\gamma}_j \;\equiv\; E_H(h). \end{aligned}$$

Si distinuono i seguenti casi significativi:

- $H \in \Pi_\nu$ e $\deg\left[P_j(\tau)\nabla H(u(\tau h))\right] \le 2k-1$, $j = 0,\dots,s-1$. Poiché $u \in \Pi_s$, questo avviene se
$$\nu \le \frac{2k}{s}.$$
In questo caso, $\hat{\gamma}_j = \gamma_j(u)$ e, pertanto, $E_H(h) = 0$, ovvero *l'Hamiltoniana è conservata esattamente lungo la traiettoria numerica*;

- in ogni altro caso, dalle (9.37)-(9.38) segue che:
$$E_H(h) = -h\sum_{j=0}^{s-1} \gamma_j(u)^T J\hat{\Delta}_j(h) = O(h^{2k+1}).$$
Pertanto, *anche se il metodo ha ordine* $2s$, *l'errore sull'Hamiltoniana è* $2k$. Conseguentemente, esso può essere reso arbitrariamente piccolo, scegliendo k sufficientemente elevato. Questo, infatti, non comporta un sostanziale aumento del costo computazionale del metodo, come esposto in Sezione 9.2.2.

[9] Osserviamo che questo metodo coincide con HBVM(1,1).

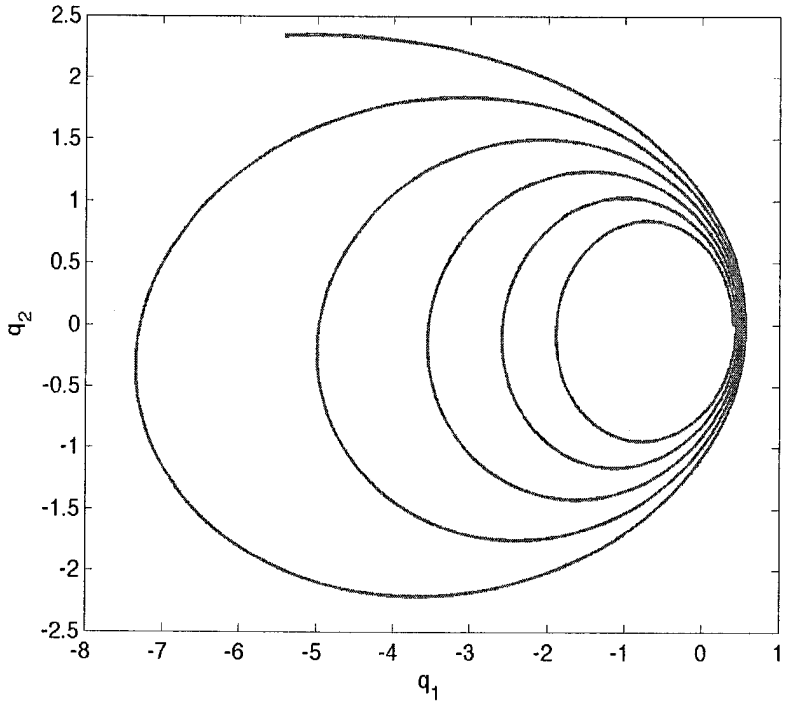

Figura 9.1: Problema di Keplero, $\varepsilon = 0.6$, metodo di Eulero esplicito, $h = \pi/500$.

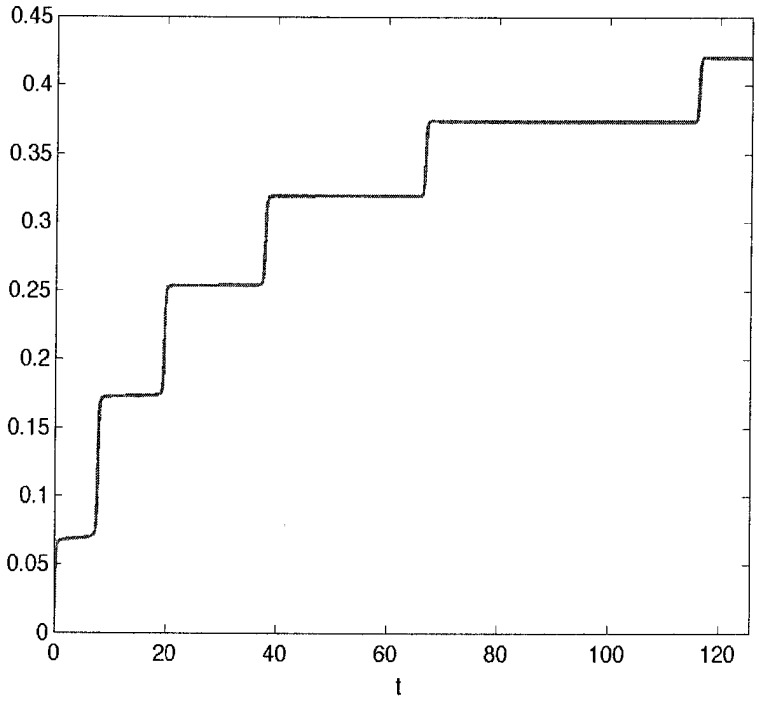

Figura 9.2: Problema di Keplero, $\varepsilon = 0.6$, metodo di Eulero esplicito, $h = \pi/500$, errore nell'Hamiltoniana.

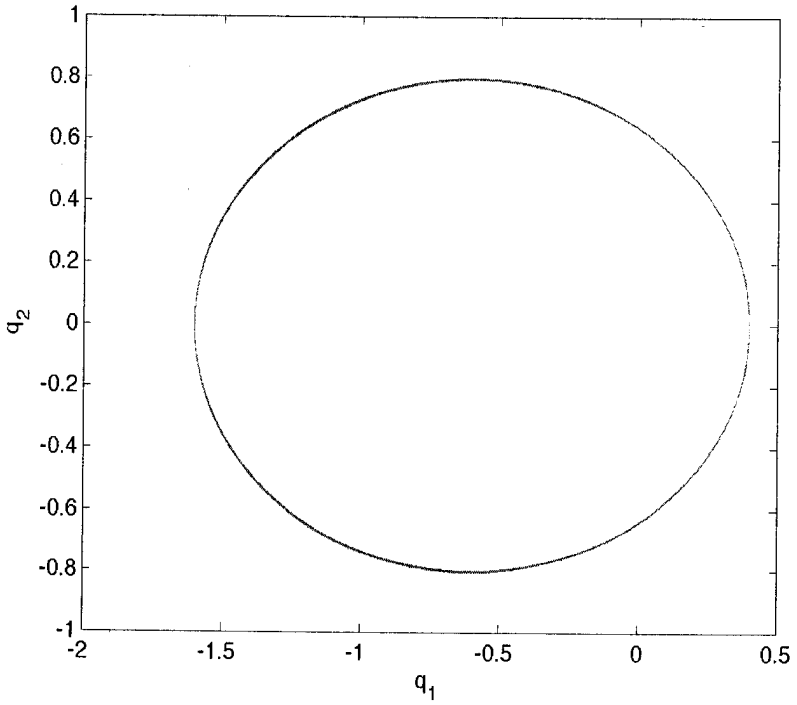

Figura 9.3: Problema di Keplero, $\varepsilon = 0.6$, *mid-point* implicito, $h = \pi/500$.

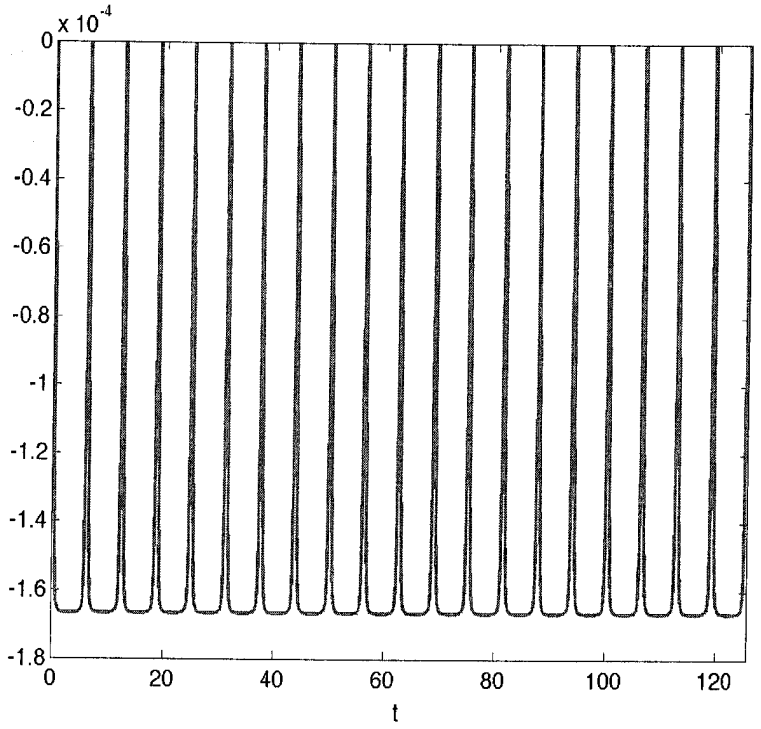

Figura 9.4: Problema di Keplero, $\varepsilon = 0.6$, *mid-point* implicito, $h = \pi/500$, errore nell'Hamiltoniana.

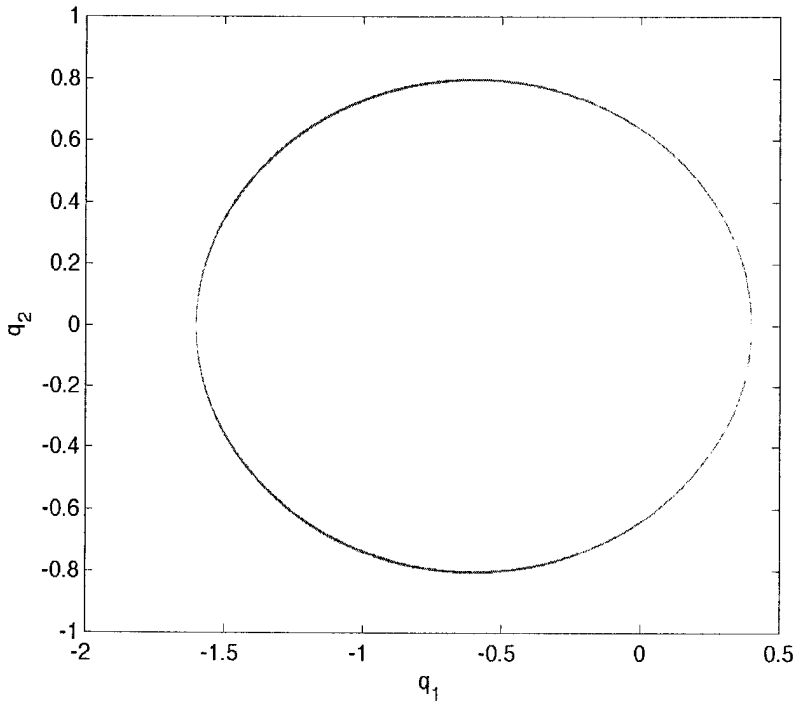

Figura 9.5: Problema di Keplero, $\varepsilon = 0.6$, HBVM(4,1), $h = \pi/500$.

Osservazione 9.5 *È opportuno sottolineare, riguardo all'ultimo punto, che in realtà è sufficiente approssimare l'Hamiltoniana fino alla precisione di macchina, dopoché, non vi è differenza tra un polinomio o una funzione sufficientemente regolare, utilizzando l'aritmetica finita. Questo è confermato dalle Figure 9.5 e 9.6, ottenute con il metodo HBVM(4,1), in cui la approssimazione dell'Hamiltoniana è dell'ordine della precisione di macchina, per il passo h utilizzato. Osserviamo, infine, che la conservazione dell'Hamiltoniana si concretizza in una crescita più lenta dell'errore (misurato ad ogni periodo), rispetto al* mid-point *implicito. Questo aspetto è documentato dal grafico in Figura 9.7. Da queste considerazioni, si comprende il perché della denominazione di questi metodi (HBVM), che sono metodi cosiddetti* "energy-preserving".[10]

[10]Ovvero, che conservano l'energia.

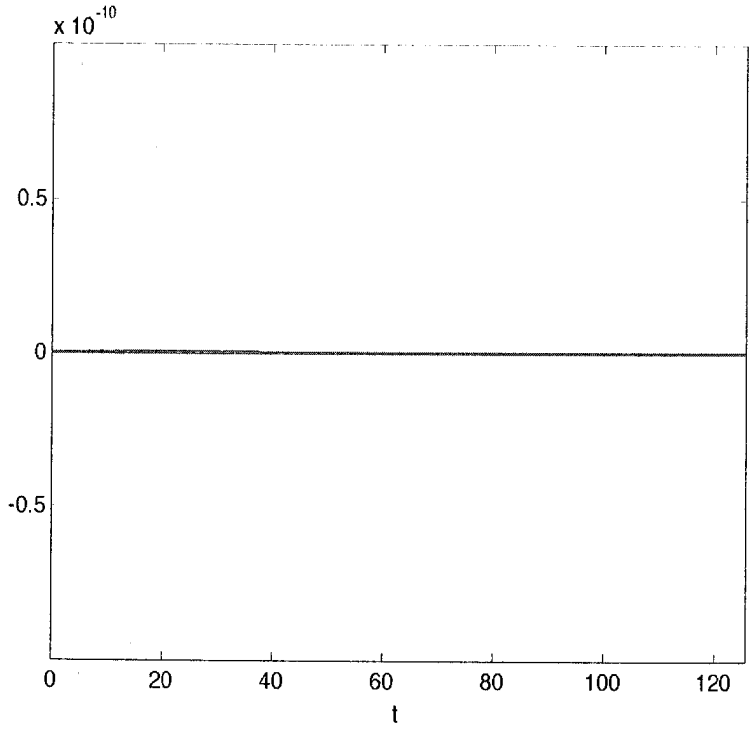

Figura 9.6: Problema di Keplero, $\varepsilon = 0.6$, HBVM(4,1), $h = \pi/500$, errore nell'Hamiltoniana.

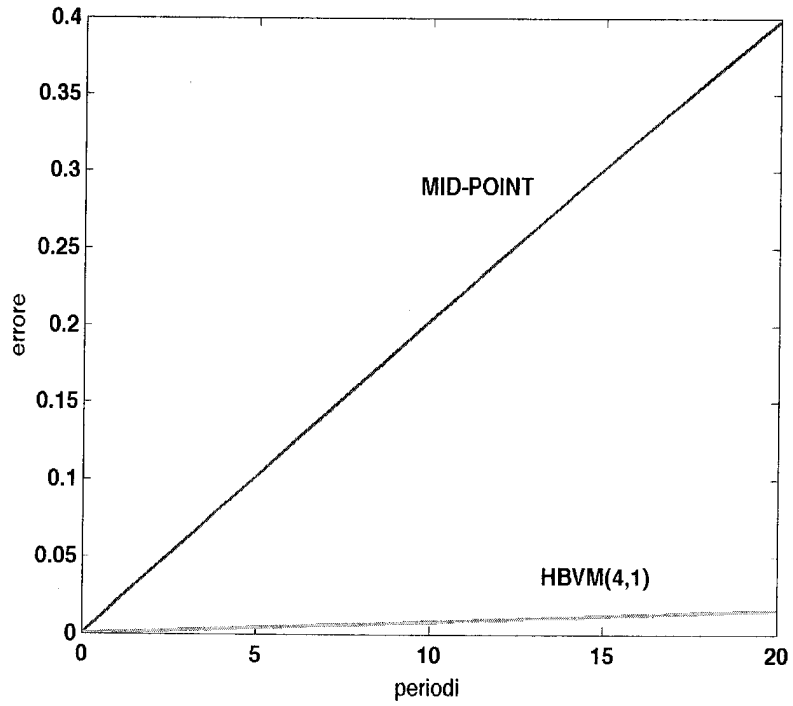

Figura 9.7: Problema di Keplero, $\varepsilon = 0.6$, errore per il *mid-point* implicito e HBVM(4,1), $h = \pi/500$.