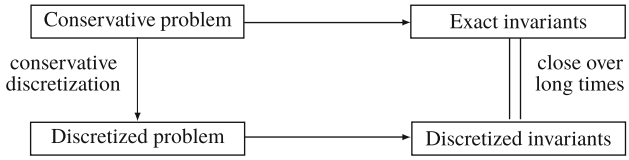
Chapter 8
Geometric Numerical Integration

It turned out that the preservation of geometric properties of the flow not only produces an improved qualitative behaviour, but also allows for a more accurate long-time integration than with general-purpose methods.

(Ernst Hairer, Christian Lubich, Gerhard Wanner, Preface of [192])

Modern Numerical Analysis is not only devoted to approximating the solutions of various problems through accurate and efficient numerical schemes, but also to retaining qualitative properties of the continuous problem over long times. Sometimes such conservation properties naturally characterize the numerical schemes, while in more complex situations preservation issues have to be conveyed into the numerical approximations. The numerical preservation of invariants is at the basis of the so-called *geometric numerical integration*. A classical reference to this topic is the monograph [192] by E. Hairer, C. Lubich and G. Wanner, which provides a comprehensive treatise on several aspects of geometric numerical integration.

The basic principle of geometric numerical integration can be briefly explained through the following diagram:



Indeed, suppose that a numerical method is applied to solve a conservative problem, i.e., a problem showing some invariants along the dynamics generated by its exact solution. A geometric numerical method provides a discretized problem that, along its solution, possesses invariants that are close to the exact ones over long time windows. Such a long-term preservation is not always automatically provided by any numerical method, hence it is relevant to analyze the conditions to impose

R. D'Ambrosio, *Numerical Approximation of Ordinary Differential Problems*,
La Matematica per il 3+2 148, https://doi.org/10.1007/978-3-031-31343-1_8

on a numerical scheme in order to make it a geometric numerical method. Before entering into the details of the topic, let us give an example.

Example 8.1 Let us consider the system of ODEs for the harmonic oscillator (1.20). As we have proved (see Example 1.7), the total energy (1.21) is a first integral of the system. We now aim to check if such a first integral remains invariant also along the numerical solutions computed by the following three methods:

- the explicit Euler method (2.19);
- the implicit Euler method (2.32);
- the two-stage Gaussian RK method (4.25).

Figures 8.1, 8.2 and 8.3 show the phase portrait of the approximate solutions to (1.20) with $\omega = 10$, computed over the time window $[0, 1000]$ by applying the aforementioned methods with constant stepsize 10^{-2}. As visible from these figures, both explicit and implicit Euler methods are not able to retain the symplecticity of the phase space, since they cannot reconstruct the periodic orbit characterizing the dynamics of (1.20). More specifically, the dynamics described by Fig. 8.1 is an outward spiral, due to the unstable behavior of the employed explicit method. On the contrary, the employ of an implicit method as in Fig. 8.2 yields an inward spiral dynamics. This is not the case of the two-stage Gaussian RK method (4.25) since, as visible from Fig. 8.3, it nicely maintains the symplecticity of the phase space.

A similar behavior can also be visible from the pattern of the deviation between the energy in the final integration point and that referred to the initial point. Indeed, Fig. 8.4 shows that the only method able to preserve the energy along time is the two-stage Gaussian RK method. The reason why this situation occurs will be clarified in the remainder of this chapter.

8.1 Historical Overview

The denomination *geometric numerical integration* strongly recalls the approach to geometry formulated by Felix Klein in his Erlangen program [238]. Klein describes geometry as the study of invariants under certain transformations. Similarly, geometric numerical methods were launched as structure-preserving schemes, able to retain peculiar features of a dynamical system along its discretizations. As addressed by Robert Mc Lachlan in his review [260] of the book by Hairer, Lubich and Wanner [192], the connection with the so-called *geometric integration theory* by Hassler Whitney [343] is even more subtle than that suggested by the name itself. Indeed, as

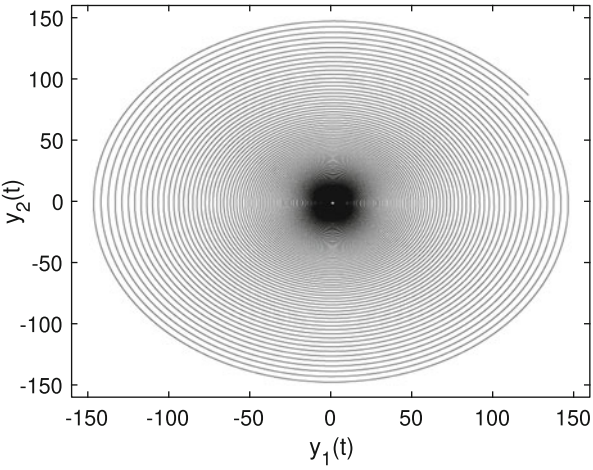


Fig. 8.1 Phase portrait of the approximate solution to the harmonic oscillator (1.20) with $\omega = 10$, initial values $y_1(0) = 0$ and $y_2(0) = 1$, computed by the Euler method (2.19) with stepsize 10^{-2}

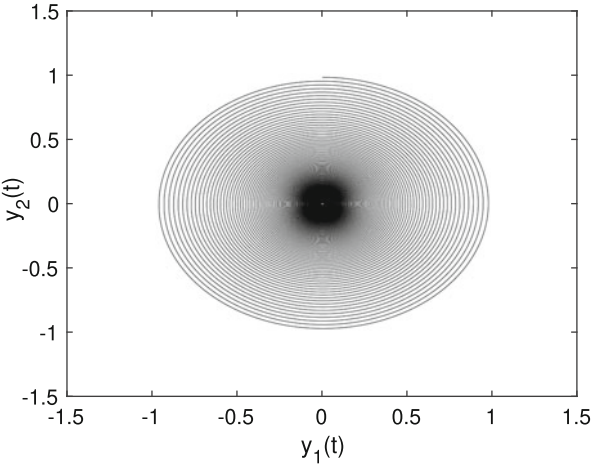


Fig. 8.2 Phase portrait of the approximate solution to the harmonic oscillator (1.20) with $\omega = 10$, initial values $y_1(0) = 0$ and $y_2(0) = 1$, computed by the implicit Euler method (2.32) with stepsize 10^{-2}

stated by Arnold [16] in his speech addressed to the participants of the International Congress of Mathematicians in Beijing, "*The design of stable discretizations of systems of PDEs often hinges on capturing subtle aspects of the structure of the system in the discretization. This new geometric viewpoint has provided a unifying understanding of a variety of innovative numerical methods developed over recent decades*". In his talk, Arnold shows that the function spaces introduced by Whitney in [343] (the so-called Whitney elements) represent what is required for a geometric discretization of many PDEs.

A famous method, well-known in the context of geometric numerical integration, is the so-called leapfrog method, also known as Störmer-Verlet method [192, 196].

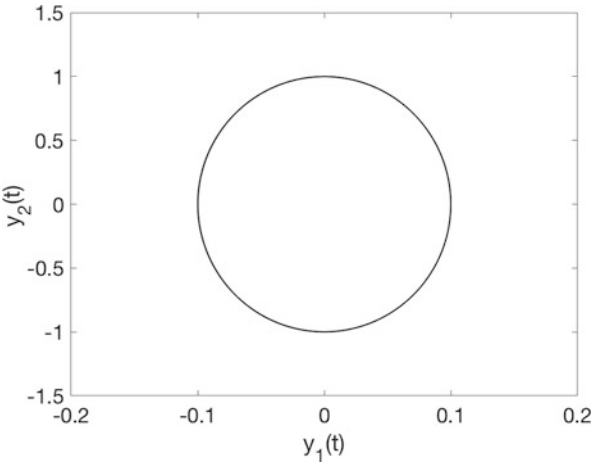


Fig. 8.3 Phase portrait of the approximate solution to the harmonic oscillator (1.20) with $\omega = 10$, initial values $y_1(0) = 0$ and $y_2(0) = 1$, computed by the two-stage Gaussian RK method (4.25) with stepsize 10^{-2}

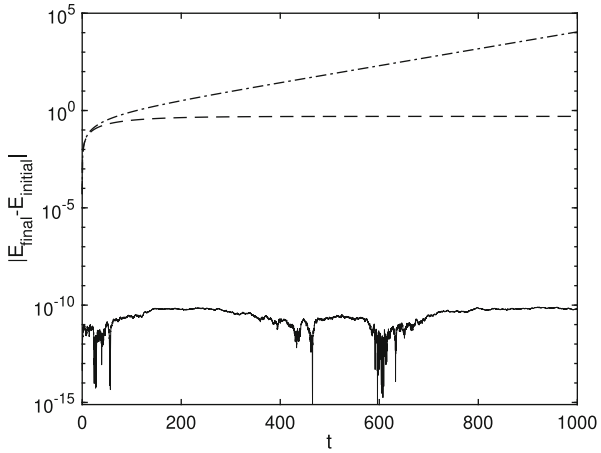


Fig. 8.4 Energy deviations in time along the approximate solutions to the harmonic oscillator (1.20) with $\omega = 10$, initial values $y_1(0) = 0$ and $y_2(0) = 1$, computed by the explicit Euler method (2.19, dashed-dotted line), the implicit Euler method (2.32, dashed line), the two-stage Gaussian RK method (4.25, solid line) with stepsize 10^{-2}. The deviation is computed as the absolute value of the difference between the energy in the final integration point $t = 1000$, minus that in the initial point $t = 0$

This method, for the discretization of the second order problem

$$\ddot{q} = f(q),$$

is given by

$$q_{n+1} - 2q_n + q_{n-1} = h^2 f(q_n).$$

This method is extensively used in many fields, such as celestial mechanics and molecular dynamics, and it is first due to Störmer that, in 1907, used a variant of this scheme for the computation of the motion of ionized particles in the Earth's magnetic field (aurora borealis). Above formulation is that developed by Verlet in 1967 [339] in his pioneering papers on the computer simulation of molecular dynamics models. Verlet was also particularly interested in the history of science, through which he was able to discover that his scheme was previously used by several authors (see [196] and references therein): for instance, by Delambre in 1792 for the computation of logarithms and astronomical tables (see [263]) and by Newton, who used it in his Principia (1687) to prove Kepler's second law (see [340]).

As highlighted in [196], a seminal contribution regarding geometric numerical integration was given by De Vogelaere in 1956 [144], "*a marvellous paper, short, clear, elegant, written in one week, submitted for publication and never published*". In particular, this paper provides examples of numerical methods (such as the symplectic Euler method) retaining the symplecticity of Hamiltonian problems. Still regarding Hamiltonian problems, successive contributions on their structure-preserving integrations are due to Ruth [305] in 1983 and Kang [232] in 1985.

A criterion for the numerical conservation of the symplecticity via Runge-Kutta methods (leading to the family of so-called *symplectic Runge-Kutta methods*) has independently been proved in 1988 by Lasagni [244], Sanz-Serna [307] and Suris [331], depending on a similar condition discovered by Cooper [98] for the numerical conservation of quadratic first integrals. To some extent, 1988 is the starting date for the spread out and the establishment of a theory of conservative numerical methods for Hamiltonian problems (on this topic, the interested reader can refer, for instance, to the monographs [26, 32, 192, 223, 233, 248, 249, 308], the survey papers [40, 41, 189, 261, 262, 264] and references therein).

Symplecticity is a prerogative of RK methods: in fact, Tang proved in 1993 [335] that linear multistep methods cannot be symplectic, as well as Hairer and Leone in 1997 [190, 250] and Butcher and Hewitt in 2009 [71] proved that genuine multivalue numerical methods cannot be symplectic. However, nearly-conserving linear multistep methods exhibiting excellent long-time behaviors have been developed by Hairer and Lubich [191, 192], Eirola and Sanz-Serna [158], while a theory of nearly-preserving multivalue methods has been explored in [67, 69, 70, 73, 122, 133, 134].

Other relevant classes of geometric numerical integrators fall in the field of the so-called *energy preserving* numerical integrators that are not considered here for the sake of brevity, but the interested reader can refer, for instance, to [31, 32, 34–36, 81–84, 92, 274–276, 294] and references therein.

This short historical overview of geometric numerical integration is clearly very far from being exhaustive and also the mentioned references are a small portion of the very wide scientific literature on the topic. However, it is in the author's opinion that even a brief glance at the historical frame is important to contextualize the results, better understand their genesis and the developments of new ideas.

8.2 Principles of Nonlinear Stability for Runge-Kutta Methods

We have introduced in Sect. 1.3 the relevant property of dissipativity of a differential problem, arising from a one-sided Lipschitz property of its vector field. In particular, we have proved that negative one-sided Lipschitz functions guarantee, according to Theorem 1.5, that contractive solutions with respect to a given norm are generated.

We now aim to understand under which conditions this feature is preserved along the solutions computed by a Runge-Kutta method, according to the following definition, given by Butcher in [61].

Definition 8.1 Let us consider a Runge-Kutta method applied to a differential problem (1.1) satisfying the contractivity condition

$$\langle f(t, y(t)) - f(t, \widetilde{y}(t)), y(t) - \widetilde{y}(t)\rangle \leq 0, \tag{8.1}$$

where $y(t)$ and $\widetilde{y}(t)$ are two solutions of (1.1), obtained with respect to the distinct initial values y_0 and $\widetilde{y}_0$, respectively. The method is *B-stable* if, for any stepsize h,

$$\|y_{n+1} - \widetilde{y}_{n+1}\| \leq \|y_0 - \widetilde{y}_0\|, \quad n \geq 0.$$

B-stable methods are certainly A-stable; this evidence can be proved by a simple check, obtained with respect to the Dahlquist test problem (6.1). The vice versa is not true. All Gaussian Runge-Kutta methods (see Sect. 4.4.1) are B-stable; the interested reader can find a detailed proof in [195].

Clearly, Definition 8.6 needs a practical way to check whether a Runge-Kutta method is B-stable or not. As usual, we present an algebraic condition on the coefficients of the method, ensuring its B-stability. Such a conditions has been independently proved by Burrage, Butcher [49] and Crouzeix [103].

Theorem 8.1 *For a given Runge-Kutta method (4.8), let us consider the matrix*

$$M = BA + A^{\intercal}B - bb^{\intercal}, \tag{8.2}$$

where $B = diag(b)$. If $b_i \geq 0$, $i = 1, 2, \ldots, s$ and M is non-negative definite, then the Runge-Kutta method is B-stable.

Proof According to Definition 8.6 of B-stability, let us consider a differential problem (1.1) generating contractive solutions and denote two of its solutions by $y(t)$ and $\widetilde{y}(t)$. Side-by-side subtraction between two applications of the Runge-Kutta method (4.8) for the approximation of $y(t)$ and $\widetilde{y}(t)$ yields

$$y_{n+1} - \widetilde{y}_{n+1} = y_n - \widetilde{y}_n + h \sum_{i=1}^{s} b_i \left(f(t_n + c_i h, Y_i) - f(t_n + c_i h, \widetilde{Y}_i) \right), \tag{8.3}$$

and

$$Y_i - \widetilde{Y}_i = y_n - \widetilde{y}_n + h \sum_{j=1}^{s} a_{ij} \left(f(t_n + c_j h, Y_j) - f(t_n + c_j h, \widetilde{Y}_j) \right). \tag{8.4}$$

Squaring side-by-side in (8.3) leads to

$$\begin{aligned} \|y_{n+1} - \widetilde{y}_{n+1}\|^2 = \|y_n - \widetilde{y}_n\|^2 & \\ +2h \sum_{i=1}^{s} b_i \langle f(t_n + c_i h, Y_i) - f(t_n + c_i h, \widetilde{Y}_i), y_n - \widetilde{y}_n \rangle & \\ +h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i b_j \langle f(t_n + c_i h, Y_i) - f(t_n + c_i h, \widetilde{Y}_i), & \\ f(t_n + c_j h, Y_j) - f(t_n + c_j h, \widetilde{Y}_j) \rangle. & \end{aligned}$$

Let us replace the value of $y_n - \widetilde{y}_n$ computed from (8.4) in the first scalar product appearing in the right-hand side of last equation, obtaining

$$\begin{aligned} \|y_{n+1} - \widetilde{y}_{n+1}\|^2 = \|y_n - \widetilde{y}_n\|^2 & \\ +2h \sum_{i=1}^{s} b_i \langle f(t_n + c_i h, Y_i) - f(t_n + c_i h, \widetilde{Y}_i), Y_i - \widetilde{Y}_i \rangle & \\ -h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} m_{ij} \langle f(t_n + c_i h, Y_i) - f(t_n + c_i h, \widetilde{Y}_i), & \\ f(t_n + c_j h, Y_j) - f(t_n + c_j h, \widetilde{Y}_j) \rangle. & \end{aligned}$$

Taking into account the contractivity condition (8.1), the hypothesis $b_i \geq 0$, $i = 1, 2, \ldots, s$, and the characteristic property of non-negative matrices

$$\sum_{i=1}^{s} \sum_{j=1}^{s} m_{ij} \langle u_i, v_j \rangle \geq 0, \quad u_i, v_j \in \mathbb{R}^d, \quad i = 1, 2, \ldots, s,$$

the thesis holds true. □

Definition 8.2 A Runge-Kutta method (4.8) such that $b_i \geq 0$, $i = 1, 2, \ldots, s$, and whose matrix M defined by (8.2) is non-negative definite, is said to be *algebraically stable*.

According to Theorem 8.1 an algebraically stable RK method is B-stable. The vice versa is not true in general, unless the method is non-confluent, i.e., $c_i \neq c_j$, for any $i \neq j$. In this case, the following result holds true.

Theorem 8.2 *A non-confluent Runge-Kutta method is B-stable if and only if it is algebraically stable.*

The interested reader can find a complete proof of this result in [195]. An equivalence theorem for confluent methods has been proved by Hundsdorfer and Spijker in [220].

The concepts and the results contained in this section are a very brief introduction of the building blocks of the so-called *nonlinear stability* theory of numerical methods, i.e., the analysis of the properties of numerical methods applied to nonlinear problems and the ability of numerical discretizations to retain the qualitative properties of nonlinear test problems. Pioneering papers on nonlinear stability analysis for numerical methods approximating the solutions of ODEs have been provided by G. Dahlquist [110, 111], starting from the notion of G-stability (also see [65, 195]).

Let us now specialize our presentation to conservation issues for numerical methods approximating nonlinear problems with selected specific features.

8.3 Preservation of Linear and Quadratic Invariants

We have introduced the notion of first integral for a d-dimensional autonomous ODE (1.17) in Sect. 1.4. We now aim to analyze the conservative behavior of Runge-Kutta methods (4.8) if such a first integral is linear, i.e., it is of the form

$$I(y(t)) = v^{\intercal} y(t), \tag{8.5}$$

with $v \in \mathbb{R}^d$. The following result holds true.

Theorem 8.3 *Any Runge-Kutta method* (4.8) *preserves linear invariants* (8.5)*, i.e.,*

$$v^{\intercal} y_{n+1} = v^{\intercal} y_n, \quad n \geq 0.$$

Proof According to Definition 1.4, a first integral satisfies

$$\nabla I(y(t)) f(y(t)) = 0,$$

that means, for the linear case (8.5)

$$v^{\intercal} f(y(t)) = 0.$$

Let us compute $v^{\intercal} y_{n+1}$, where y_{n+1} is provided by a RK method (4.8), obtaining

$$v^{\intercal} y_{n+1} = v^{\intercal} y_n + h \sum_{i=1}^{s} b_i v^{\intercal} f(Y_i).$$

Since $v^{\intercal} f(Y_i) = 0$, $i = 1, 2, \ldots, s$, the thesis holds true. □

Let us now analyze the conservation of quadratic functions

$$Q(y(t)) = y(t)^{\intercal} C y(t), \tag{8.6}$$

where $C \in \mathbb{R}^{d \times d}$ is a symmetric matrix. Such a quadratic form is a first integral of (1.17), according to Definition 1.4, if

$$y(t)^{\intercal} C f(y(t)) = 0. \tag{8.7}$$

This condition is useful to prove the following result, proved by Cooper in [98].

Theorem 8.4 *If the coefficients of a Runge-Kutta method* (4.8) *fulfill the condition*

$$b_i a_{ij} + b_j a_{ji} = b_i b_j, \quad i, j = 1, 2, \ldots, s, \tag{8.8}$$

then it preserves quadratic invariants (8.6)*, i.e.,*

$$y_{n+1}^{\intercal} C y_{n+1} = y_n^{\intercal} C y_n, \quad n \geq 0.$$

Proof Let us compute the quadratic form $y_{n+1}^{\intercal} C y_{n+1}$, obtaining

$$y_{n+1}^{\intercal} C y_{n+1} = y_n^{\intercal} C y_n + h \sum_{i=1}^{s} b_i f(Y_i)^{\intercal} C y_n + h \sum_{i=1}^{s} b_i y_n^{\intercal} C f(Y_i)$$
$$+ h^2 \sum_{i,j=1}^{s} b_i b_j f(Y_i)^{\intercal} C f(Y_j).$$

Let us analyze the $O(h)$ terms in the right-hand side of last equation, by recasting y_n using the formula of the internal stages in (4.8), i.e.,

$$y_n = Y_i - h \sum_{j=1}^{s} a_{ij} f(Y_j).$$

We correspondingly obtain

$$h \sum_{i=1}^{s} b_i f(Y_i)^{\intercal} C y_n = h \sum_{i=1}^{s} b_i f(Y_i)^{\intercal} C Y_i - h^2 \sum_{i,j=1}^{s} b_i a_{ij} f(Y_i)^{\intercal} C f(Y_j),$$
$$h \sum_{i=1}^{s} b_i y_n^{\intercal} C f(Y_i) = h \sum_{i=1}^{s} b_i Y_i^{\intercal} C f(Y_i) - h^2 \sum_{i,j=1}^{s} b_j a_{ji} f(Y_i)^{\intercal} C f(Y_j),$$

i.e., by means of (8.7),

$$h \sum_{i=1}^{s} b_i f(Y_i)^{\intercal} C y_n = -h^2 \sum_{i,j=1}^{s} b_i a_{ij} f(Y_i)^{\intercal} C f(Y_j),$$
$$h \sum_{i=1}^{s} b_i y_n^{\intercal} C f(Y_i) = -h^2 \sum_{i,j=1}^{s} b_j a_{ji} f(Y_i)^{\intercal} C f(Y_j).$$

We finally get

$$y_{n+1}^{\intercal} C y_{n+1} = y_n^{\intercal} C y_n - h^2 \sum_{i,j=1}^{s} (b_i a_{ij} + b_j a_{ji} - b_i b_j) f(Y_i)^{\intercal} C f(Y_j),$$

leading to the thesis. □

It is worth observing that Eq. (8.8) provides an algebraic condition on the coefficients of RK methods that can more compactly be written as $M = 0$, where the matrix M is defined by (8.2). In other terms, the matrix M plays a role both in retaining the contractive character of solutions to dissipative problems and in conserving quadratic first integrals. However, the story does not end here, as we recognize in next section: indeed, RK methods satisfying (8.8) are particularly relevant in the numerical approximation of Hamiltonian problems.

We have realized that any Runge-Kutta method is able to exactly preserve linear invariants, while quadratic invariants are preserved only by a family of Runge-Kutta methods. A natural question to ask is what happens to polynomial invariants of degree greater than or equal to 3. This (negative) result gives the answer related to RK methods, whose complete proof can be found in [192]. Clearly, as aforementioned, since Runge-Kutta methods are not able to cover themselves all possible conservation issues, other relevant classes of geometric numerical integrators have been introduced, most of them falling in the general field of *energy-preserving* numerical methods (the reader can refer, for instance, to [31, 32, 34–36, 81–84, 92, 274–276, 294] and references therein).

8.4 Symplectic Methods

We have introduced a relevant class of conservative problems in Sect. 1.4, i.e., Hamiltonian problems (1.28). A characteristic property of these problems, as proved in Theorem 1.6 is the symplecticity of the corresponding flow map. In the spirit of geometric numerical integration we are interested in understanding under which conditions a numerical method is able to retain the same property along discretized dynamics. Let us particularly focus on one-step methods; we represent them as a map φ_h that associates y_{n+1} to y_n and give the following definition.

Definition 8.3 A one-step method is *symplectic* if the one-step map φ_h is a symplectic transformation when applied to a smooth Hamiltonian problem (1.28), i.e., if

$$\varphi_h'(y_n)^\intercal J \varphi_h'(y_n) = J.$$

We now provide important examples of symplectic methods, starting from the famous *symplectic Euler method*, introduced by de Vogelaere in [144].

Theorem 8.5 (de Vogelaere) *The* symplectic Euler method

$$\begin{aligned} p_{n+1} &= p_n - h\mathcal{H}_q(p_{n+1}, q_n), \\ q_{n+1} &= q_n + h\mathcal{H}_p(p_{n+1}, q_n), \end{aligned} \tag{8.9}$$

for the numerical solution of Hamiltonian problems (1.22) *is a symplectic method of order* 1.

Proof We first differentiate (8.9) side-by-side with respect to (p_n, q_n), obtaining

$$\begin{aligned} \frac{\partial p_{n+1}}{\partial p_n} &= \frac{\partial p_n}{\partial p_n} - h\mathcal{H}_{qp}\frac{\partial p_{n+1}}{\partial p_n}, \\ \frac{\partial p_{n+1}}{\partial q_n} &= -h\mathcal{H}_{qp}\frac{\partial p_{n+1}}{\partial q_n} - h\mathcal{H}_{qq}\frac{\partial q_n}{\partial q_n}, \\ \frac{\partial q_{n+1}}{\partial p_n} &= h\mathcal{H}_{pp}\frac{\partial p_{n+1}}{\partial p_n} \\ \frac{\partial q_{n+1}}{\partial q_n} &= \frac{\partial q_n}{\partial q_n} + h\mathcal{H}_{pp}\frac{\partial p_{n+1}}{\partial q_n} + h\mathcal{H}_{pq}\frac{\partial q_n}{\partial q_n} \end{aligned}$$

being $I \in \mathbb{R}^{d\times d}$ the identity matrix and avoiding to explicitly write the dependence of the Hamiltonian function on (p_{n+1}, q_n) for the sake of brevity. As a consequence,

$$\begin{aligned} \left(I + h\mathcal{H}_{qp}\right)\frac{\partial p_{n+1}}{\partial p_n} &= I, \\ \left(I + h\mathcal{H}_{qp}\right)\frac{\partial p_{n+1}}{\partial q_n} &= -h\mathcal{H}_{qq}, \\ -h\mathcal{H}_{pp}\frac{\partial p_{n+1}}{\partial p_n} + \frac{\partial q_{n+1}}{\partial p_n} &= 0, \\ -h\mathcal{H}_{pp}\frac{\partial p_{n+1}}{\partial q_n} + \frac{\partial q_{n+1}}{\partial q_n} &= I + h\mathcal{H}_{pq}. \end{aligned}$$

Recasting above relations in a compact matrix form yields

$$\begin{bmatrix} I + h\mathcal{H}_{qp} & 0 \\ -h\mathcal{H}_{pp} & I \end{bmatrix} \begin{bmatrix} \frac{\partial p_{n+1}}{\partial p_n} & \frac{\partial p_{n+1}}{\partial q_n} \\ \frac{\partial q_{n+1}}{\partial p_n} & \frac{\partial q_{n+1}}{\partial q_n} \end{bmatrix} = \begin{bmatrix} I & -h\mathcal{H}_{qq} \\ 0 & I + h\mathcal{H}_{pq} \end{bmatrix},$$

from which we compute

$$\begin{bmatrix} \frac{\partial p_{n+1}}{\partial p_n} & \frac{\partial p_{n+1}}{\partial q_n} \\ \frac{\partial q_{n+1}}{\partial p_n} & \frac{\partial q_{n+1}}{\partial q_n} \end{bmatrix} = \begin{bmatrix} I + h\mathcal{H}_{qp} & 0 \\ -h\mathcal{H}_{pp} & I \end{bmatrix}^{-1} \begin{bmatrix} I & -h\mathcal{H}_{qq} \\ 0 & I + h\mathcal{H}_{pq} \end{bmatrix}$$

$$= \begin{bmatrix} D & -hD\mathcal{H}_{qq} \\ h\mathcal{H}_{pp}D & -h^2\mathcal{H}_{pp}D\mathcal{H}_{qq} + D^{-1} \end{bmatrix},$$

where $D = (I + h\mathcal{H}_{qp})^{-1}$. The reader can easily check that the symplecticity condition

$$\begin{bmatrix} \frac{\partial p_{n+1}}{\partial p_n} & \frac{\partial p_{n+1}}{\partial q_n} \\ \frac{\partial q_{n+1}}{\partial p_n} & \frac{\partial q_{n+1}}{\partial q_n} \end{bmatrix}^{\intercal} J \begin{bmatrix} \frac{\partial p_{n+1}}{\partial p_n} & \frac{\partial p_{n+1}}{\partial q_n} \\ \frac{\partial q_{n+1}}{\partial p_n} & \frac{\partial q_{n+1}}{\partial q_n} \end{bmatrix} = J$$

holds true. □

We observe that the symplectic Euler method (8.9) is implicit with respect to p. An alternative version implicit in q also exists, given by

$$\begin{aligned} p_{n+1} &= p_n - h\mathcal{H}_q(p_n, q_{n+1}), \\ q_{n+1} &= q_n + h\mathcal{H}_p(p_n, q_{n+1}) \end{aligned} \tag{8.10}$$

and the reader can check its symplecticity, applying similar arguments as those used in the proof of Theorem 8.5, see Exercise 1 at the end of this chapter.

Let us now provide a Matlab implementation of the symplectic Euler method (8.9) applied to (1.22), given in Program 8.1. The code requires defining the right-hand side of (1.22) through the functions `fp.m` and `fq.m`. Moreover, the built-in function `fsolve` is used to handle the implicitness of (8.9).

Program 8.1 (Symplectic Euler Method)

```
% Function implementing the symplectic Euler method (8.9)
% for the numerical solution of a Hamiltonian problem
% on a uniform grid.

% Inputs
% - problem: label of the problem to be solved;
% - tspan: vector of two components, storing the extrema
%          of the integration interval;
```

(continued)

Program 8.1 (continued)

```
% - p0: initial momentum;
% - q0: initial position;
% - h: constant stepsize.

% Outputs
% - t: set of N equidistant grid points (tspan(1) excluded);
% - p: d×N matrix whose i-th column p(:,i) stores the
%      approximate momentum in the i-th grid point;
% - q: d×N matrix whose i-th column p(:,i) stores the
%      approximate position in the i-th grid point.

function [t,p,q]=symplecticEuler(problem,tspan,p0,q0,h)
N=(tspan(2)-tspan(1))/h;
t=linspace(h,tspan(2),N);
d=length(p0);
p=zeros(d,N);
q=zeros(d,N);
options=optimset('Display','off','TolFun',eps,'TolX',eps);
p(:,1)=fsolve(@(x) x-p0-h*fp(problem,x,q0),p0,options);
q(:,1)=q0+h*fq(problem,p(:,1),q0);
for i=2:N
    p(:,i)=fsolve(@(x) x-p(:,i-1)+...
           h*fp(problem,x,q(:,i-1)),p(:,i-1),options);
    q(:,i)=q(:,i-1)+h*fq(problem,p(:,i),q(i-1));
end
```

Example 8.2 Let us solve the system of ODEs for the mathematical pendulum (1.23) by the symplectic Euler method (8.9), in order to check if the symplecticity of the continuous flow is also retained along the numerical dynamics. The numerical evidence is provided by using Program 8.1 and displayed in Fig. 8.5, showing that the symplecticity of the phase space is nicely preserved by (8.9) that provides the periodic orbit characterizing the dynamics of (1.23). This property is not visible if a non-symplectic method is used: for instance, computing the numerical dynamics by means of the explicit Euler method (2.19) provides the phase portrait depicted in Fig. 8.6, where the symplecticity of the original problem is totally lost.

Let us now analyze the property of symplecticity for Runge-Kutta methods, applied to Hamiltonian problems (1.22). This topic has been object of seminal papers, all dated 1988, independently authored by Lasagni [244], Sanz-Serna [307], Suris [331]. The proof of symplecticity for Runge-Kutta methods relies on the following lemma [27, 192].

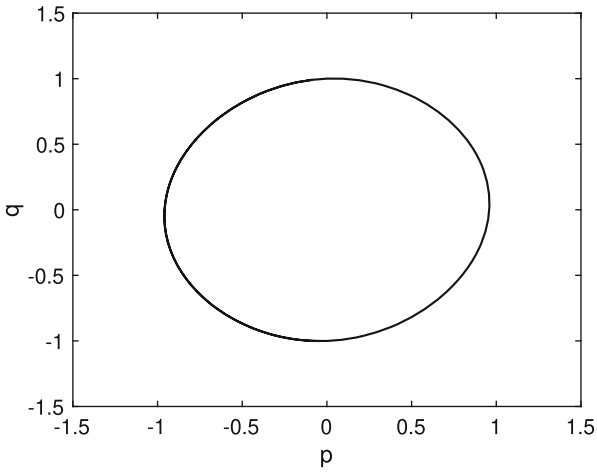


Fig. 8.5 Phase portrait associated to the approximate solution to the mathematical pendulum (1.23) with initial values $p(0) = 0$ and $q(0) = 1$, computed by the symplectic Euler method (8.9) with stepsize 10^{-1}

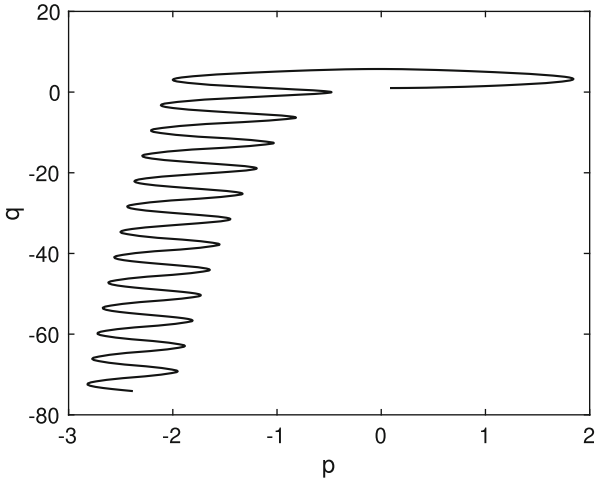
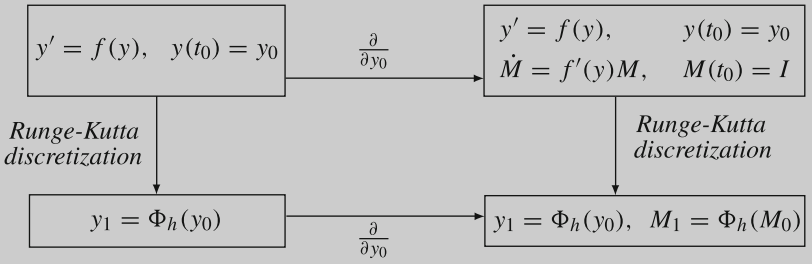


Fig. 8.6 Phase portrait associated to the approximate solution to the mathematical pendulum (1.23) with initial values $p(0) = 0$ and $q(0) = 1$, computed by the explicit Euler method (2.19) with stepsize 10^{-1}

Lemma 8.1 *Consider an autonomous problem* (1.17) *and its variational equation* (1.29). *Correspondingly, let us denote by* $y_{n+1} = \Phi_h(y_n)$ *the map associating a single step of a given Runge-Kutta method from the point* t_n *to* t_{n+1} *of the grid. Then, the following diagram commutes:*

(continued)

Lemma 8.1 (continued)



where the horizontal arrows denote differentiation with respect to y_0 and the vertical arrows the application of Φ_h. In other terms, the numerical result $\{y_1, M_1\}$ obtained by applying a single step of the method to the problem augmented by its variational equation is equal to the numerical solution of $\dot{y} = f(y)$ augmented by its derivative $M_1 = \partial y_1/\partial y_0$.

Proof We first compute a single step of a RK method (4.8) applied to (1.17) and side-by-side differentiate with respect to y_0, obtaining

$$\begin{aligned} \frac{\partial y_1}{\partial y_0} &= I + h \sum_{i=1}^{s} b_i f'(Y_i) \frac{\partial Y_i}{\partial y_0}, \\ \frac{\partial Y_i}{\partial y_0} &= I + h \sum_{j=1}^{s} a_{ij} f'(Y_j) \frac{\partial Y_j}{\partial y_0}, \quad i = 1, 2, \ldots, s. \end{aligned} \tag{8.11}$$

We observe that the last equation is a linear system in the unknowns $\frac{\partial Y_i}{\partial y_0}$, $i = 1, 2, \ldots, s$.

We now aim to prove that side-by-side differentiating (1.17) and then applying (4.8) lead to the same result. So, we apply (4.8) directly to the variational equation (1.29), getting

$$\begin{aligned} \frac{\partial y_1}{\partial y_0} &= I + h \sum_{i=1}^{s} b_i f'(Y_i) \widetilde{M}_i, \\ \widetilde{M}_i &= I + h \sum_{j=1}^{s} a_{ij} f'(Y_j) \widetilde{M}_j, \quad i = 1, 2, \ldots, s. \end{aligned} \tag{8.12}$$

We observe that last equation is also a linear system in the unknowns $\widetilde{M}_i$, $i = 1, 2, \ldots, s$. Moreover, the two linear systems displayed as second equations of (8.11) and (8.12) act exactly in the same way. For sufficiently small values of h, both systems have unique solution and, since they are the same system, we have $\widetilde{M}_i = \partial Y_i / \partial y_0$ and, consequently, $M_1 = \partial y_1 / \partial y_0$. So the diagram in the statement of the lemma commutes. □

Theorem 8.6 *Any RK method* (4.8) *preserving quadratic first integrals* (8.6) *is a symplectic method.*

Proof Let us consider the augmented system

$$\begin{aligned} \dot{y} &= J^{-1}\nabla\mathcal{H}(y), \\ \dot{M} &= J^{-1}\nabla^2\mathcal{H}(y)M, \end{aligned} \tag{8.13}$$

containing the Hamiltonian problem (1.28) and its variational equation. Let us prove that $M^\intercal J M$ is a first integral for (8.13). Indeed,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}(M^\intercal J M) &= \dot{M}^\intercal J M + M^\intercal J \dot{M} \\ &= \left(J^{-1}\nabla^2\mathcal{H}(y)M\right)^\intercal J M + M^\intercal J J^{-1}\nabla^2\mathcal{H}(y)M \\ &= M^\intercal \left(\nabla^2\mathcal{H}(y)\right)^\intercal (J^{-1})^\intercal J\psi + M^\intercal\nabla^2\mathcal{H}(y)M \\ &= -M^\intercal\nabla^2\mathcal{H}(y)M + M^\intercal\nabla^2\mathcal{H}(y)M = 0. \end{aligned}$$

In other terms, $M^\intercal J M$ is a quadratic first integral of (8.13) and is preserved by any RK method fulfilling the condition (8.8) of conservation of quadratic invariants described in Theorem 8.4. The conserved value of $M^\intercal J M$ is then equal to its initial value, i.e., $M^\intercal J M = J$, that is the symplecticity condition. So, all RK conserving quadratic invariants are symplectic. □

It is worth highlighting that condition (8.8) is then also a symplecticity condition. For this reason, the literature directly denotes RK methods satisfying (8.8) as *symplectic RK methods*. A consequence of this result is that all Gaussian RK methods (see Sect. 4.4.1) are symplectic methods; Program 8.2 implements one of them, namely that depending on two internal stages (4.25), to solve a given Hamiltonian problem.

Program 8.2 (Symplectic RK Method (2-Stage Gaussian Method))

```
% Function implementing the 2-stage Gaussian method (4.25)
% for the numerical solution of a Hamiltonian problem
% on a uniform grid.

% Inputs
% - problem: label of the problem to be solved;
% - tspan: vector of two components, storing the extrema
%          of the integration interval;
% - y0: vector of initial momenta (stored in y0(1:d)) and
%       initial positions (stored in y0(d+1:2d))
% - h: constant stepsize.

% Outputs
% - t: set of N equidistant grid points (tspan(1) excluded);
% - y: 2d×N matrix whose i-th column stores approximate
%      momenta (in y0(1:d)) and coordinates (in y0(d+1:2d)),
%      referring to the i-th grid point;
% - hamDev: N-dimensional vector storing the deviation
%          of the Hamiltonian function in each grid point
%          from the initial Hamiltonian;

function [t,y,hamDev]=GaussRK2s(problem,tspan,y0,h)
N=(tspan(2)-tspan(1))/h;
t=linspace(h,tspan(2),N);
d=length(y0)/2;
Id=eye(2*d);
y=zeros(2*d,N);
hamDev=zeros(N,1);
c=[(3-sqrt(3))/6; (3+sqrt(3))/6]; e=ones(length(c),1);
A=[1/4 1/4-sqrt(3)/6; 1/4+sqrt(3)/6 1/4];
b=[1; 1]/2;
options=optimset('Display','off','TolFun',eps,'TolX',eps);
Y=fsolve(@(Z) Z-kron(e,Id)*y0-h*kron(A,Id)*...
  [f(problem,[],Z(1:2*d)); f(problem,[],Z(2*d+1:4*d))],...
  [y0; y0],options);
y(:,1)=y0+h*kron(b',Id)*...
       [f(problem,[],Y(1:2*d)); f(problem,[],Y(2*d+1:4*d))];
ham0=hamiltonian(problem,y0);
hamDev(1)=abs(hamiltonian(problem,y(:,1))-ham0);
for i=2:N
    Y=fsolve(@(Z) Z-kron(e,Id)*y(:,i-1)-h*kron(A,Id)*...
      [f(problem,[],Z(1:2*d)); f(problem,[],Z(2*d+1:4*d))],...
      [y(:,i-1); y(:,i-1)],options);
    y(:,i)=y(:,i-1)+h*kron(b',Id)*[f(problem,[],Y(1:2*d));
          f(problem,[],Y(2*d+1:4*d))];
    hamDev(i)=abs(hamiltonian(problem,y(:,i))-ham0);
end
```

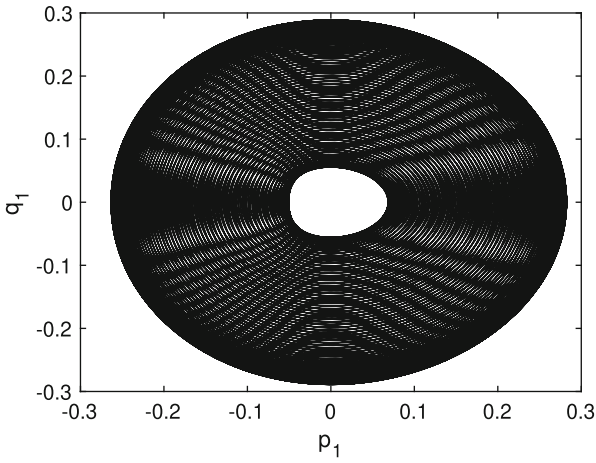


Fig. 8.7 Phase portrait of (1.27) in the (p_1, q_1)-plane, with initial values $p_1(0) = 0.2$, $p_2(0) = 0$, $q_1(0) = -0.2$, $q_2(0) = 0$. The displayed dynamics originates from the application of the symplectic two-stage Gaussian RK method (4.25) with stepsize $h = 0.1$

A numerical evidence of the symplecticity of Gaussian RK method is certainly given by Example 8.1. An additional one is reported in the following example, whose results have been obtained via Program 8.2.

Example 8.3 Let us consider Hénon-Heiles problem (1.26), already analyzed in Example 1.9 in order to provide a numerical evidence of the symplecticity of the two-stage Gaussian RK method (4.25). Figures 8.7, 8.8, 8.9, and 8.10 display the phase portrait in several planes and provide a confirmation of the symplecticity of the numerical scheme, able to recover the symplecticity of the original problem along the numerical dynamics. We observe that the chosen time window is [0, 4000] and the employed stepsize is $h = 0.1$.

8.5 Symmetric Methods

A relevant property of mechanical systems is their time reversibility; in terms of flow map, this property is equivalent to say that $\Phi_t \circ \Phi_{-t}$ is the identity map. In other terms, for a reversible system with initial value y_0, the dynamics starting from $y(t)$ with reverse time goes back to y_0. In this section, we aim to understand under which conditions this property is recovered by a one-step method. Then, the following definitions are given.

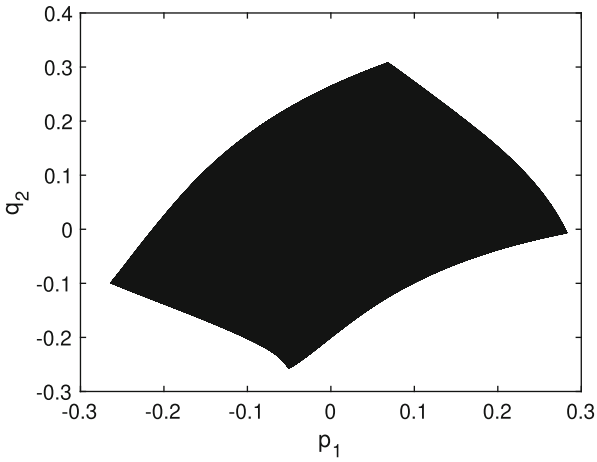


Fig. 8.8 Phase portrait of (1.27) in the (p_1, q_2)-plane, with initial values $p_1(0) = 0.2$, $p_2(0) = 0$, $q_1(0) = -0.2$, $q_2(0) = 0$. The displayed dynamics originates from the application of the symplectic two-stage Gaussian RK method (4.25) with stepsize $h = 0.1$

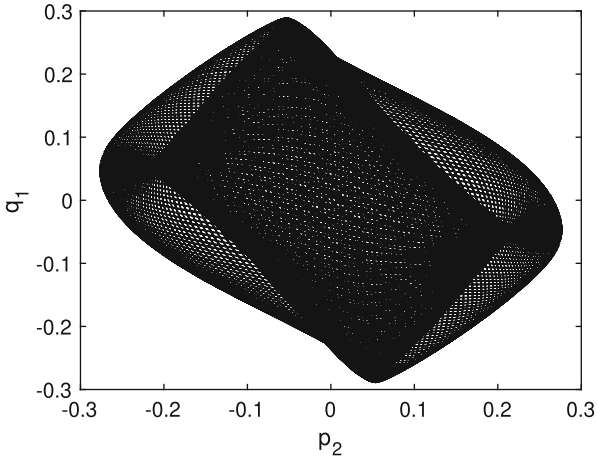


Fig. 8.9 Phase portrait of (1.27) in the (p_2, q_1)-plane, with initial values $p_1(0) = 0.2$, $p_2(0) = 0$, $q_1(0) = -0.2$, $q_2(0) = 0$. The displayed dynamics originates from the application of the symplectic two-stage Gaussian RK method (4.25) with stepsize $h = 0.1$

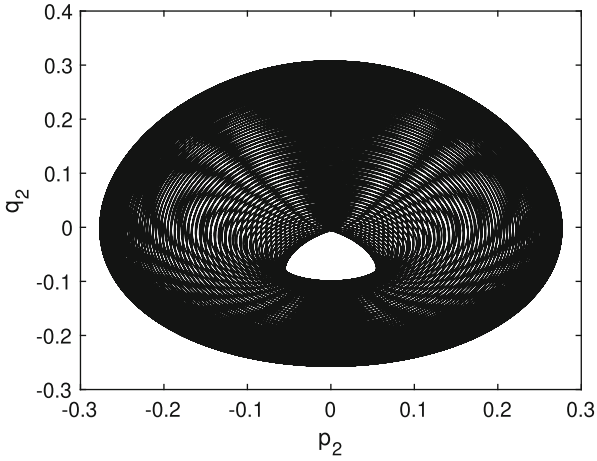


Fig. 8.10 Phase portrait of (1.27) in the (p_2, q_2)-plane, with initial values $p_1(0) = 0.2$, $p_2(0) = 0$, $q_1(0) = -0.2$, $q_2(0) = 0$. The displayed dynamics originates from the application of the symplectic two-stage Gaussian RK method (4.25) with stepsize $h = 0.1$

Definition 8.4 Given a one-step method φ_h, its *adjoint method* is the one-step map

$$\varphi_h^\star = \varphi_{-h}^{-1}.$$

Definition 8.5 A one-step method φ_h is *symmetric* if it is equal to its adjoint.

Example 8.4 Let us compute the adjoint of the explicit Euler method (2.19), i.e.,

$$y_n = y_{n+1} - hf(y_{n+1}).$$

Rearranging the terms in the last equation leads to the implicit Euler method (2.32). Hence, the explicit Euler method is not self-adjoint, so it is not symmetric.

The implicit midpoint method (4.24) is symmetric since its adjoint method is given by

$$y_n = y_{n+1} - hf\left(\frac{1}{2}(y_{n+1} + y_n)\right),$$

i.e., it is the implicit midpoint method as well.

The following theorem provides a relevant accuracy property of symmetric methods, useful for their construction and analysis. Indeed, we now prove that the order of convergence of a symmetric method is always even, then their construction requires to fulfill a restricted number of order conditions.

Theorem 8.7 *The order of a symmetric one-step method is even.*

Proof Let us denote by p the order of convergence of the method. Then (also see Theorem 3.2, Section II.3 in [192]), a single step of length h satisfies

$$\varphi_h(y_0) = \Phi_h(y_0) + Ch^{p+1} + O(h^{p+2}),$$

where C is the error constant of the method. Performing a step in reverse time leading to y_0 yields

$$y_0 = \varphi_{-h}(\Phi_h(y_0)) + (-1)^p Ch^{p+1} + O(h^{p+2}).$$

Inverting the operator

$$\varphi_h^\star(y_0) = \Phi_h(y_0) + (-1)^p Ch^{p+1} + O(h^{p+2}).$$

Therefore, the adjoint of a method of order p has order p as well. Moreover, since the method is symmetric, then $C = (-1)^p C$ and, as a consequence, the error constant C is different from 0 only for even values of p. □

We now aim to give a characterization of symmetric Runge-Kutta methods, provided in terms of algebraic conditions on their coefficients, as usual.

Theorem 8.8 *If the coefficients of a given Runge-Kutta method* (4.8) *satisfy the conditions*

$$a_{s+1-i,s+1-j} + a_{ij} = b_j, \quad i, j = 1, 2, \ldots, s, \tag{8.14}$$

then, the method is symmetric.

Proof The first step of the proof consists in computing the coefficients of the adjoint of a Runge-Kutta method (4.8). Referring to a single step with stepsize $-h$, leading to y_n if we start from y_{n+1}, the internal stages $Y_i^\star$ of the adjoint method are given by

$$\begin{aligned} Y_i^\star &= y_{n+1} - h\sum_{j=1}^{s} a_{ij} f(Y_j) = y_n + h\sum_{j=1}^{s} b_j f(Y_j) - h\sum_{j=1}^{s} a_{ij} f(Y_j) \\ &= y_n + h\sum_{j=1}^{s} (b_j - a_{ij}) f(Y_j). \end{aligned}$$

Observing that the internal stages of the adjoint method appear in reverse order with respect to those of the original method, i.e.,

$$Y_i^\star = Y_{s+1-i}, \quad i = 1, 2, \ldots, s,$$

the coefficients of the adjoint are then given by

$$a_{ij}^\star = b_{s+1-j} - a_{s+1-i,s+1-j}, \quad i, j = 1, 2, \ldots, s.$$

Proceeding similarly with the advancing law, we obtain

$$b_i^\star = b_{s+1-i}, \quad i = 1, 2, \ldots, s.$$

The second step of the proof is a trivial check of the conditions guaranteeing that the method is equal to its adjoint, i.e., $a_{ij}^\star = a_{ij}$ and $b_i^\star = b_i$, leading to the thesis. □

Example 8.5 Let us specialize the symmetry conditions (8.14) to specific values of s, in order to check the symmetry of some methods presented in the previous chapters.

For $s = 1$, (8.14) yields

$$b_1 = 2a_{11}.$$

This condition is certainly satisfied by the one-stage Gaussian Runge-Kutta method (4.23), i.e., the implicit midpoint method, that is a symmetric method of order 2. This result is not surprising, since we have already given a direct proof of symmetry for the implicit midpoint method in Example 8.4.

For $s = 2$, (8.14) yields

$$a_{11} + a_{22} = a_{12} + a_{21}, \quad b_1 = b_2.$$

These conditions are satisfied by the two-stage Gaussian Runge-Kutta method (4.25), as well as by the two-stage Lobatto IIIA and Lobatto IIIB methods, presented in Sect. 4.4.3. Hence, these methods are symmetric.

Actually, the property is more general: all Gaussian Runge-Kutta methods (see Sect. 4.4.1) are symmetric. Similarly, all Lobatto IIIA and Lobatto IIIB (presented in Sect. 4.4.3) are symmetric as well. The interested reader can find a detailed proof in [192].

We finally aim to understand which is the connection between symplecticity and symmetry for RK methods. In some cases (as it happens for Gaussian RK methods),

the two notions coexist, while in other cases (think of Lobatto IIIA methods) they do not. The following result holds true.

Theorem 8.9 *For a given Runge-Kutta method (4.8) the following statements are equivalent:*

- *the method is symmetric for linear problems $y' = Ly$, with $L \in \mathbb{R}^{d\times d}$;*
- *the method is symplectic for problems of the type $y' = JCy$, where C is a symmetric matrix;*
- *the stability function $R(z)$ of the method, defined in (6.9), satisfies $R(-z)R(z) = 1$, for any $z \in \mathbb{C}$.*

Proof Applying a RK method to a linear problem $y' = Ly$ leads to the recurrence $y_{n+1} = R(hL)y_n$, where $R(hL)$ is the matrix version of the stability function (6.9) of the employed RK method, defined for linear scalar test problems. Symmetry holds true if and only if $y_n = R(-hL)y_{n+1}$, leading to $R(-hL)R(hL) = I$, being $I \in \mathbb{R}^{d\times d}$ is the identity matrix.

Applying a RK method to the problem $y' = JCy$ leads to $y_{n+1} = R(hJC)y_n$. As a consequence, since $\varphi_h'(y_n) = R(hJC)$, the symplecticity condition reads

$$R(hJC)^\intercal J R(hJC) = J \tag{8.15}$$

and, since for implicit Runge-Kutta methods $R(z)$ is a rational function, its matrix counterpart can be factored out as

$$R(hJC) = P(hJC)Q(hJC)^{-1}.$$

Consequently, condition (8.15) is equivalent to

$$Q(hJC)^{-\intercal} P(hJC)^\intercal J P(hJC) Q(hJC)^{-1} = J,$$

i.e.,

$$P(hJC)^\intercal J P(hJC) = Q(hJC)^\intercal J Q(hJC).$$

Algebraic manipulations of the last expression (left to the reader, see Exercise 3 at the end of this chapter) lead to $R(-hJC)R(hJC) = I$. □

Let us observe that symmetry and symplecticity are equivalent concepts if the problem is of type $y' = JCy$. This is certainly true for Hamiltonian problems with quadratic Hamiltonian function $\mathcal{H}(y) = \frac{1}{2}y^\intercal Cy$, where C is a symmetric matrix, since $\nabla\mathcal{H}(y) = Cy$.

8.6 Backward Error Analysis

As highlighted at the beginning of this chapter, a geometric numerical method is able to retain characteristic features of a dynamical system over long times. Studying the long-term character of numerical methods for ODEs has already regarded, for instance, the analysis of their linear and nonlinear stability properties, presented in the previous sections. A very effective tool in order to investigate the long-term conservative property of candidate geometric numerical methods is the *backward error analysis*, extensively presented in [192] and references therein, whose origin comes from numerical linear algebra (in particular the work of Wilkinson [345]).

The main ingredient of backward error analysis consists in inspecting the properties of differential equations associated to a numerical method, well known as *modified differential equations*, whose role is clarified in the following section.

8.6.1 Modified Differential Equations

Let us focus on the solution of an autonomous problem (1.17) by a one-step method that, over a single step, is briefly denoted as the map

$$y_n = \varphi_h(y_{n-1}).$$

Forward error analysis is performed after computing the numerical solution, by estimating the local error (i.e., the local on a single step, such as $y_1 - \Phi_h(y_0)$, being Φ the flow map of the continuous problem) or the global error (i.e., the error overall the integration interval so far, without localizing assumptions, given by $y_n - \Phi_{t_0+nh}(y_0)$).

Backward error analysis is the analysis of a continuous problem relying on the so-called modified differential equations, whose exact solution is the numerical solution of the original ODEs. More specifically, we search for an ordinary differential equation $\tilde{y}' = f_h(\tilde{y})$, written in terms of a formal power series of h, i.e.,

$$\tilde{y}' = f(\tilde{y}) + hf_2(\tilde{y}) + h^2 f_3(\tilde{y}) + \dots, \tag{8.16}$$

such that $y_n = \tilde{y}(t_0 + nh)$. The error is then measured as difference between the vector field $f(y)$ of the original problem (1.17) and that of the modified differential equation (8.16), namely $f_h(y)$. In other terms, the idea is to interpret the numerical solution computed by a given numerical method as the exact solution of a continuous problem. The right-hand side in (8.16) may generally give rise to a divergent series, so we will later employ just a truncation of it.

Under suitable regularity assumptions, the computation of modified differential equations can be provided, for instance, by means of Taylor series arguments and

using the expressions of the elementary differentials introduced in Sect. 4.2.2, as follows. Let us first expand $\tilde{y}(t+h)$ around t, leading to

$$\begin{aligned}
\tilde{y}(t+h) &= \tilde{y}(t) + h\tilde{y}'(t) + \frac{h^2}{2}\tilde{y}''(t) + \frac{h^3}{6}\tilde{y}'''(t) + \dots \\
&= \tilde{y}(t) + h\left(f + hf_2 + h^2 f_3 + \dots\right) + \frac{h^2}{2}\left(f'\tilde{y}'(t) + hf_2'\tilde{y}'(t) + \dots\right) \\
&\qquad + \frac{h^3}{6}\left(f''(f,f) + f'f'f + \dots\right) + \dots \\
&= \tilde{y}(t) + h\left(f + hf_2 + h^2 f_3 + \dots\right) \\
&\qquad + \frac{h^2}{2}\left(f' + hf_2' + \dots\right)\left(f + hf_2 + \dots\right) \\
&\qquad + \frac{h^3}{6}\left(f''(f,f) + f'f'f + \dots\right) + \dots
\end{aligned} \tag{8.17}$$

or, equivalently,

$$\begin{aligned}
\tilde{y}(t+h) = \tilde{y}(t) + hf + h^2\left(f_2 + \frac{1}{2}f'f\right) \\
+ h^3\left(f_3 + \frac{1}{2}\left(f'f_2 + f_2'f\right) + \frac{1}{6}\left(f''(f,f) + f'f'f\right)\right) + \dots
\end{aligned} \tag{8.18}$$

In the expressions above we have omitted the dependence of f, f_2, f_3 and their derivatives on $\tilde{y}(t)$, in order to simplify the notation.

Supposing that the one-step map $\phi_h(y)$ can be expanded itself in power series of h, with coefficient $f(y)$ for the power 1 due to the consistency of the method, i.e.,

$$\varphi_h(y) = y + hf(y) + h^2 d_2(y) + h^3 d_3(y) + \dots \tag{8.19}$$

yields

$$\begin{aligned}
f_2 &= d_2(y) - \frac{1}{2}f'f, \\
f_3 &= d_3(y) - \frac{1}{6}\left(f''(f,f) + f'f'f\right) - \frac{1}{2}\left(f'f_2 + f_2'f\right),
\end{aligned} \tag{8.20}$$

and so on, by comparison of (8.18) and (8.19).

Let us provide an example of computation of modified differential equations for selected numerical methods aimed to solve a scalar problem.

Example 8.6 Let us consider the following differential equation

$$y'(t) = y(t)^4, \tag{8.21}$$

assuming $y(0) = 1$ as initial value, the exact solution is

$$y(t) = \sqrt[3]{\frac{1}{1-3t}}.$$

We aim to compute the modified differential equation associated to the explicit Euler method (2.19). Clearly, in this case we have $d_j(y) = 0$ for all $j \geq 2$ in (8.19). The coefficients given in (8.20) assume the form

$$f_2(y) = -\frac{3}{2}y^5, \quad f_3(y) = \frac{19}{3}y^{10}.$$

As a consequence, the modified differential equation for the explicit Euler method applied to the logistic equation (8.21) reads

$$\tilde{y}' = \tilde{y}^4 - \frac{3}{2}h\tilde{y}^5 + \frac{19}{3}h^2\tilde{y}^{10} + \dots \tag{8.22}$$

Figure 8.11 compares the solution of the original problem based on the ODE (8.21) with the solution of the modified differential equations truncated after the h and h^2 terms. We observe that taking more terms in the modified differential equation improves the agreement between numerical and exact solutions.

The following theorem highlights an important, though expectable, property: the perturbation term in the modified differential equation of an order p method has magnitude $O(h^p)$.

Theorem 8.10 *The modified differential equation (8.16) of a one-step method $y_{n+1} = \varphi_h(y_n)$ of order p has the form*

$$\tilde{y}' = f(\tilde{y}) + h^p f_{p+1}(\tilde{y}) + h^{p+1} f_{p+2}(\tilde{y}) + \dots,$$

with $f_{p+1}(y)$ equal to the principal error term of the method.

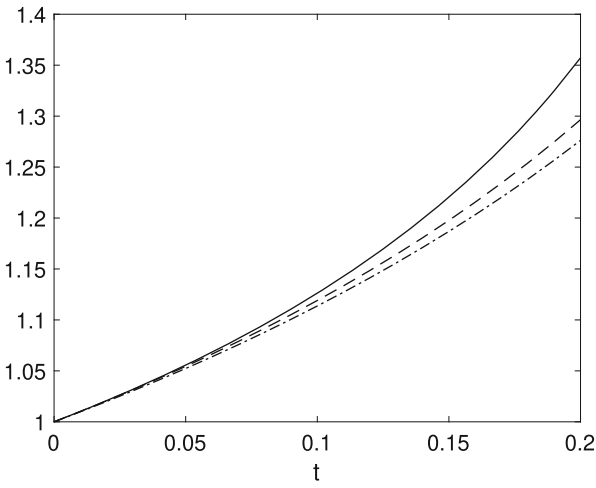


Fig. 8.11 Exact solution of Eq. (8.21) (solid line) vs solutions of the modified differential equation (8.22) of the explicit Euler method, truncated at the $O(h)$ (dashed-dotted line) and $O(h^2)$ (dashed line) terms

Proof The proof follows straightforwardly from the fact that $f_j(y) = 0$, for $2 \leq j \leq p$, if and only if $\varphi_h(y) - \Phi_h(y) = O(h^{p+1})$. □

A special case worth being considered regards the analysis of modified differential equations of symplectic methods [23, 192, 277, 336], hence with a focus on Hamiltonian problems (1.28). To this purpose, it is useful introducing the following lemma [192].

Lemma 8.2 *Let Ω be an open set of $\mathbb{R}^d$ and $f : \Omega \to \mathbb{R}^d$ be a continuously differentiable function, whose Jacobian is symmetric. Then, for any $y_0 \in \Omega$ there exists a neighborhood of y_0 and a function $\mathcal{H}(y)$ such that $f(y) = \nabla\mathcal{H}(y)$ on this neighborhood.*

Theorem 8.11 *Consider a symplectic method $\varphi_h(y)$ applied to a Hamiltonian system (1.28) with smooth Hamiltonian. Then, the corresponding modified differential equation*

$$\dot{\tilde{y}} = f(\tilde{y}) + hf_2(\tilde{y}) + h^2 f_3(\tilde{y}) + \dots$$

is also Hamiltonian. In particular, there exist smooth functions $\mathcal{H}_j : \mathbb{R}^{2d} \to \mathbb{R}$ for $j = 2, 3, \dots$, such that $f_j(y) = J\nabla\mathcal{H}_j(y)$.

Proof The proof is given by induction. In particular, since $f_1(y) = f(y) = J\nabla\mathcal{H}(y)$, we assume that $f_j(y) = J\nabla\mathcal{H}_j(y)$ is satisfied for $j = 1, 2, \ldots, r$ and aim to prove the existence of a Hamiltonian $\mathcal{H}_{r+1}(y)$. According to the inductive hypothesis, the truncated modified differential equation

$$\dot{\tilde{y}} = f(\tilde{y}) + hf_2(\tilde{y}) + \cdots + h^{r-1} f_r(\tilde{y})$$

is Hamiltonian, with Hamiltonian function given by $\mathcal{H}(y) + h\mathcal{H}_2(y) + \cdots + h^{r-1}\mathcal{H}_r(y)$. Defining its flow by $\Phi_{r,t}(y_0)$, we have

$$\begin{aligned}
\varphi_h(y_0) &= \Phi_{r,t}(y_0) + h^{r+1} f_{r+1}(y_0) + O(h^{r+2}),\\
\varphi_h'(y_0) &= \Phi_{r,t}'(y_0) + h^{r+1} f_{r+1}'(y_0) + O(h^{r+2}) \ .
\end{aligned}$$

Since the method is symplectic and the inductive hypothesis holds true, both φ_h and $\Phi_{r,h}$ are symplectic maps. Taking into account that $\Phi_{r,h}'(y_0) = I + O(h)$, we have that

$$\begin{aligned}
J &= \varphi_h'(y_0)^\intercal J \varphi_h'(y_0)\\
&= \left(\Phi_{r,t}'(y_0) + h^{r+1} f_{r+1}'(y_0)\right)^\intercal J \left(\Phi_{r,t}'(y_0) + h^{r+1} f_{r+1}'(y_0)\right) + O(h^{r+2})\\
&= \left(I + h^{r+1} f_{r+1}'(y_0)\right)^\intercal J \left(I + h^{r+1} f_{r+1}'(y_0)\right) + O(h^{r+2})\\
&= J + h^{r+1}\left(f_{r+1}'(y_0)^\intercal J + J f_{r+1}'(y_0)\right) + O(h^{r+2}).
\end{aligned}$$

This means that the matrix $J^\intercal f_{r+1}'(y_0)$ is symmetric and, by means of Lemma 8.2, there exists $\mathcal{H}_{r+1}(y)$ such that

$$J^\intercal f_{r+1}(y_0) = \nabla\mathcal{H}_{r+1}(y)$$

or, equivalently,

$$f_{r+1}(y_0) = J\nabla\mathcal{H}_{r+1}(y),$$

that completes the proof. □

We complete this section presenting a couple of results regarding the construction of the modified differential equation for the adjoint of a numerical method and, as a consequence, we provide an important result concerning the modified differential equations of symmetric methods.

Theorem 8.12 *Considering a one-step method $\varphi_h(y)$, whose modified differential equation (8.19) has coefficients $f_j(y)$, the coefficients of the modified equations of its adjoint $\varphi_h^\star(y)$ satisfy*

$$f_j^\star(y) = (-1)^{j+1} f_j(y).$$

Proof The thesis holds true in straightforward way, by considering that $\tilde{y}(t-h) = \varphi_{-h}(\tilde{y}(t))$. Consequently, it is enough to replace h by $-h$ in formulae (8.16), (8.17) and (8.19) to obtain the thesis. □

Corollary 8.1 *The right-hand side of the modified differential equation of a symmetric method only consists in even powers of h.*

Proof The thesis is direct consequence of Theorem 8.12, since a symmetric method coincides with its adjoint and, therefore, the same happens to their modified differential equations. Thus, any $f_j(y)$ is null, whenever j is even; coefficients of (8.16) with even subindices are those related to odd powers of h that, consequently, disappear from (8.16) if the method is symmetric. □

8.6.2 Truncated Modified Differential Equations

As aforementioned, the presentation of modified differential equations so far has been based on considering their right-hand side as a formal series of powers of h, without taking into account its convergence. Unfortunately, as clearly highlighted in [192], such a power series is almost never convergent, actually even in very simple situations. As a consequence, we should consider a proper truncation of the modified differential equations, up to an optimal index to be properly chosen. Such a choice is based on rigorous error estimates, described in details in [192] and references therein. Here we report them without their proofs, that can be found in the mentioned monograph by Hairer, Lubich and Wanner.

We aim to find an optimal truncation index N for the modified differential equation (8.16) leading to

$$\tilde{y}' = F_N(\tilde{y}) = f(\tilde{y}) + hf_2(\tilde{y}) + \ldots + h^{N-1} f_N(\tilde{y}),$$

with $\tilde{y}(0) = y_0$. To this purpose, the following bound on the coefficients of (8.16), whose proof can be found in [192], is particularly useful.

Theorem 8.13 *Suppose that $f(y)$ is analytic in $\mathcal{B}_{2R}(y_0)$ and the coefficients of (8.19) are also analytic in $\mathcal{B}_R(y_0)$. Assume that there exists a positive M such that $\|f(y)\| \leq M$, for any $\|y - y_0\| \leq 2R$. Moreover, assume that each $d_j(y)$ in (8.19) satisfies*

$$\|d_j(y)\| \leq \mu M \left(\frac{2\kappa M}{R}\right)^{j-1},$$

for any $\|y - y_0\| \leq R$, where

$$\mu = \sum_{i=1}^{s} |b_i|, \quad \kappa = \max_{i=1,2,\ldots,s} \sum_{j=1}^{s} |a_{ij}|.$$

Then, the following bound holds true

$$\|f_j(y)\| \leq \ln 2 \, \eta M \left(\frac{\eta M j}{R}\right)^{j-1}, \tag{8.23}$$

assuming that $\|y - y_0\| \leq R/2$ and being $\eta = 2\max(\kappa, \mu/(2\ln 2 - 1))$.

Taking into account the bound (8.23) and since the function $(\varepsilon x)^x$ has a minimum at $x = (\varepsilon \mathrm{e})^{-1}$, it makes sense assuming as truncation index the integer N such that

$$\frac{\eta M N}{R} \leq \frac{1}{h\mathrm{e}}$$

or, in less restrictive way,

$$hN \leq \mathrm{e}h_0,$$

being $h_0 = \frac{R}{\mathrm{e}\eta M}$. In this way, since $\|f(y)\| \leq M$ and using (8.23), we have

$$\|F_N(y)\| \leq M \left(1 + \eta \ln 2 \sum_{j=2}^{N} \left(\frac{\eta M j}{R}\right)^{j-1}\right) \leq M \left(1 + \eta \ln 2 \sum_{j=2}^{N} \left(\frac{j}{hN}\right)^{j-1}\right),$$

leading to

$$\|F_N(y)\| \leq M(1 + 1.65\eta).$$

The following result holds true (see [192]).

Theorem 8.14 *Let $f(y)$ be analytic in $\mathcal{B}_{2R}(y_0)$ and the coefficients $d_j(y)$ of (8.19) analytic in $\mathcal{B}_R(y_0)$. If $h \leq h_0/4$, then there exists $N = N(h)$ (the largest integer satisfying $hN \leq h_0$), such that*

$$\|\varphi_h(y_0) - \Phi_{N,h}(y_0)\| \leq h\gamma M e^{-h_0/h},$$

with $\gamma = \mathrm{e}(2 + 1.65 + \mu)$ only depending on the method.

In other terms, for problems with analytic vector fields, the numerical solution computed by a one-step method and the solution of the corresponding modified differential equation, truncated after $N \sim \frac{1}{h}$ terms, differ by a term that is exponentially small.

8.6.3 Long-Term Analysis of Symplectic Methods

The core of backward error analysis in the context of geometric numerical integration certainly involves the study of the long-time conservative character of symplectic numerical methods applied to Hamiltonian problems (1.28). We know from Theorem 8.11 that the corresponding modified differential equation is also Hamiltonian and, after truncation, the modified Hamiltonian is given by

$$\widetilde{\mathcal{H}}(y) = \mathcal{H}(y) + h^p \mathcal{H}_{p+1}(y) + \cdots + h^{N-1}\mathcal{H}_N(y). \tag{8.24}$$

The following fundamental result, proved by Benettin and Giorgilli in [23], provides information on the long-term conservative character of symplectic methods.

Theorem 8.15 (Benettin-Giorgilli Theorem) *Consider a Hamiltonian system (1.28) with analytic Hamiltonian function $\mathcal{H} : D \to \mathbb{R}$, with $D \subset \mathbb{R}^{2d}$. Suppose that a symplectic numerical method $\varphi_h(y)$ of order p is used to solve this problem and assume that the corresponding numerical solution lies in a compact set $K \subset D$. Then, there exists h_0 and $N = N(h)$ (as in Theorem 8.13)*

(continued)

Theorem 8.15 (continued)
such that

$$\begin{aligned}\widetilde{\mathcal{H}}(y_n) &= \widetilde{\mathcal{H}}(y_0) + \mathcal{O}(e^{-h_0/2h}),\\ \mathcal{H}(y_n) &= \mathcal{H}(y_0) + \mathcal{O}(h^p),\end{aligned} \tag{8.25}$$

for exponentially long time intervals of length $nh - t_0 \leq e^{h_0/2h}$.

Proof Let $\Phi_{N,t}(y_0)$ be the flow of the truncated modified equation (8.24), that is also Hamiltonian with Hamiltonian function $\widetilde{\mathcal{H}}$ satisfying $\widetilde{\mathcal{H}}\left(\Phi_{N,t}(y_0)\right) = \widetilde{\mathcal{H}}(y_0)$, for any t. As a consequence of Theorem 8.14, we have that

$$\|y_{n+1} - \Phi_{N,h}(y_n)\| \leq h\gamma M e^{-h_0/h}$$

and again, from Theorem 8.13, we deduce that there exists a global Lipschitz constant (independent from h) for $\widetilde{\mathcal{H}}$, such that

$$\widetilde{\mathcal{H}}(y_n) - \widetilde{\mathcal{H}}(\Phi_{N,h}(y_n)) = \mathcal{O}(he^{-h_0/h}) \ .$$

Since

$$\widetilde{\mathcal{H}}(y_n) - \widetilde{\mathcal{H}}(y_0) = \sum_{j=1}^{n}\left(\widetilde{\mathcal{H}}(y_j) - \widetilde{\mathcal{H}}(y_{j-1})\right) = \sum_{j=1}^{n}\left(\widetilde{\mathcal{H}}(y_j) - \widetilde{\mathcal{H}}(\Phi_{N,h}(y_{j-1}))\right) \ ,$$

we obtain $\widetilde{\mathcal{H}}(y_n) - \widetilde{\mathcal{H}}(y_0) = \mathcal{O}(nhe^{-h_0/h})$, that proves the statement for $\widetilde{\mathcal{H}}$, recalling that $nh \leq e^{h_0/2h}$.

The result for $\mathcal{H}$ follows from (8.24), since

$$\begin{aligned}\widetilde{\mathcal{H}}(y) &= \mathcal{H}(y) + h^p\mathcal{H}_{p+1}(y) + \cdots + h^{N-1}\mathcal{H}_N(y)\\ &= \mathcal{H}(y) + h^p\left(\mathcal{H}_{p+1}(y) + h\mathcal{H}_{p+2}(y) + \cdots + h^{N-p-1}\mathcal{H}_N(y)\right)\end{aligned}$$

and considering the fact that

$$\mathcal{H}_{p+1}(y) + h\mathcal{H}_{p+2}(y) + \cdots + h^{N-p-1}\mathcal{H}_N(y)$$

is uniformly bounded on K, independently of h and N. This is a consequence of the fact that

$$\mathcal{H}_j(y) = \int_0^1 y^{\intercal} f_j(ty)\mathrm{d}t + \text{constant}$$

on a ball centered in y_0 contained in D and, moreover, of the estimate on f_j given by (8.23). □

Benettin-Giorgilli theorem 8.15 is a gifted result in understanding the long-term conservative character of a symplectic method: as long as the numerical solution lies in a compact set, the Hamiltonian function of the optimally truncated modified differential equation is almost conserved up to errors of exponentially small size. Moreover, for a symplectic method of order p, the modified Hamiltonian function is close to the original Hamiltonian function over exponentially long time windows, with a deviation comparable to the accuracy in the computation of the solution, i.e., $O(h^p)$. Let us test the usefulness of this result through the following highly didactic example.

Example 8.7 Let us apply Benettin-Giorgilli theorem to the mathematical pendulum (1.23), with $p_0 = 0$ and $q_0 = 1$. The reader can find a detailed verification of the hypothesis of Theorem 8.15 for this problem in [192] (Example VI.8.2). Actually, the stepsize restriction dictated by Theorem 8.14 is too severe and definitely not sharp. Indeed, symplectic methods may have excellent conservation properties even if used with large values of the stepsize.

We use the symplectic Euler method (8.9) and the two-stage Gaussian method (4.25) with several values of the stepsize. As visible in Fig. 8.12, the conservation of the symplectic structure is achieved also for large values of h.

Let us now check the accuracy in conserving the Hamiltonian function. Figures 8.13 and 8.14 reveal an excellent long-term conservation of the Hamiltonian, measured for several values of the stepsize, in the intervals [0,1000] and [0,10000]. The accuracy of the second equation in (8.25) is also confirmed, as visible in Tables 8.1 and 8.2, where the orders of both methods are very well recovered. They have been computed through the following formula, analogous to (3.23),

$$p \approx \log_2 \left| \frac{\mathcal{H}(y_N) - \mathcal{H}(y_0)}{\mathcal{H}(y_{2N}) - \mathcal{H}(y_0)} \right|, \tag{8.26}$$

i.e., as the logarithm in basis 2 of the ratio of the deviations between the Hamiltonian in the numerical solution computed with stepsize h from the initial Hamiltonian, divided by the analogous deviation with stepsize $h/2$. Both values are listed in the table with reference to the final integration point.

Let us finally make an observation on non-symplectic methods, motivated by Fig. 8.4, where a linear energy drift is visible for the explicit Euler method. This fact can be motivated through arguments very similar to those provided in the proof of Benettin and Giorgilli theorem (8.15). Indeed, one can prove (also see Exercise 6

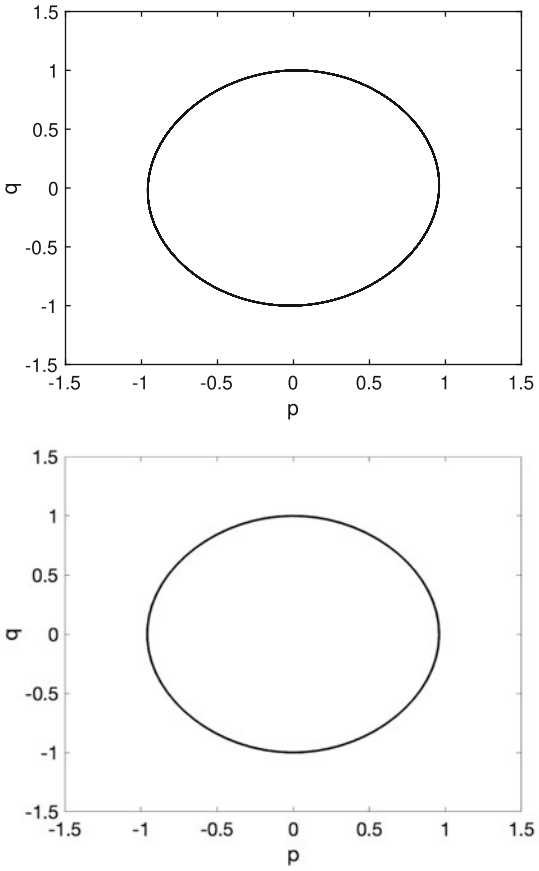


Fig. 8.12 Example 8.7: phase portrait associated to the numerical dynamics generated by applying the symplectic Euler method (8.9) (top) and the two-stage Gaussian method (4.25) (bottom) to the mathematical pendulum (1.23). The graphs are obtained in correspondence of $h = 0.05$ (top) and $h = 0.1$ (bottom)

at the end of the chapter) that

$$\mathcal{H}(y_n) = \mathcal{H}(y_0) + O(th^p).$$

We finally observe that alternatives to symplecticity or relaxed notions of symplecticity have been treated in the literature, e.g., through the notion of conjugate symplectic method [133, 192, 197].

8.7 Long-Term Analysis of Multivalue Methods

This section is devoted to providing a comprehensive analysis of the long-term stability properties of multivalue numerical methods, described in Chap. 5. The presented analysis is based on the results contained in [122].

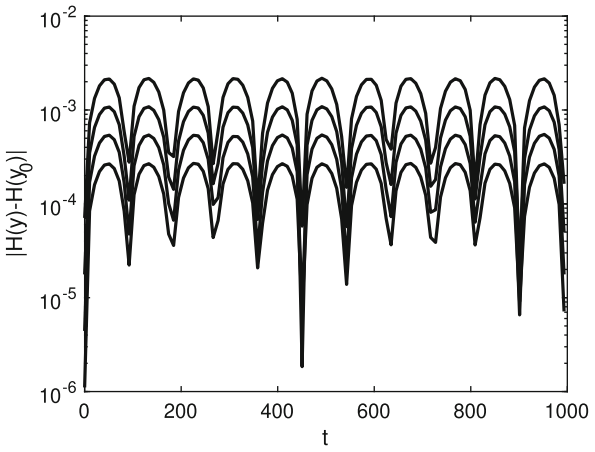


Fig. 8.13 Example 8.7: Hamiltonian deviations along the numerical dynamics generated by applying the symplectic Euler method (8.9) to the mathematical pendulum (1.23). The graphs are obtained in correspondence of four values of the stepsize: $h = 0.01$ (top), $h = 0.005, 0.0025$ (middle) and $h = 0.00125$ (bottom). The plot displays the graph obtained considering a grid point every hundred

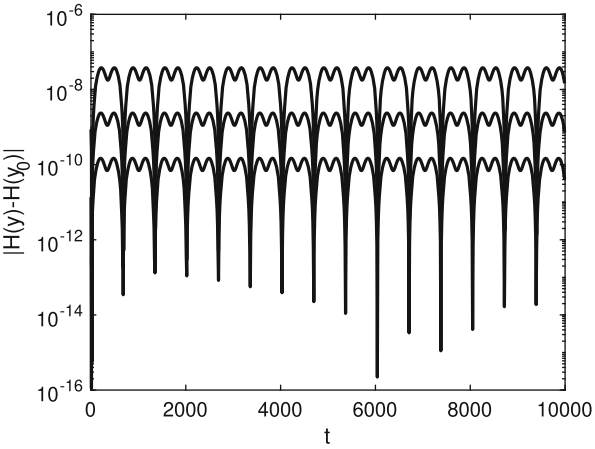


Fig. 8.14 Example 8.7: Hamiltonian deviations along the numerical dynamics generated by applying the two-stage Gaussian method (4.25) to the mathematical pendulum (1.23). The three graphs are obtained in correspondence of three values of the stepsize: $h = 0.1$ (top), $h = 0.05$ (middle) and $h = 0.025$ (bottom). The plot displays the graph obtained considering a grid point every hundred

Table 8.1 Example 8.7: Hamiltonian deviations along the numerical dynamics generated by applying the symplectic Euler method (8.9) to the mathematical pendulum (1.23), computed in the final integration point $t = 1000$. The displayed Hamiltonian deviations measure the gap at the final step point from the initial Hamiltonian. Order estimation is also reported, computed as suggested by Eq. (8.26)

h	Hamiltonian deviation (final point)	p
0.01	$1.64 \cdot 10^{-4}$	
0.005	$7.27 \cdot 10^{-5}$	1.17
0.0025	$3.49 \cdot 10^{-5}$	1.06
0.00125	$1.69 \cdot 10^{-5}$	1.05

Table 8.2 Example 8.7: Hamiltonian deviations along the numerical dynamics generated by applying the two-stage Gaussian method (4.25) to the mathematical pendulum (1.23), computed in the final integration point $t = 10000$. The displayed Hamiltonian deviations measure the gap at the final step point from the initial Hamiltonian. Order estimation is also reported, computed as suggested by Eq. (8.26)

h	Hamiltonian deviation (final point)	p
0.05	$6.74 \cdot 10^{-9}$	
0.025	$4.23 \cdot 10^{-10}$	3.99
0.0125	$2.64 \cdot 10^{-11}$	4.00

To perform the long-term analysis of multivalue methods, it is worth using the following representation for the forward step procedure

$$Y_{n+1} = V\, Y_n + h\, \Phi(h, Y_n). \tag{8.27}$$

We also remind that the method requires a starting procedure

$$Y_0 = \mathcal{S}_h(y_0),$$

and a finishing procedure

$$y_n = \mathcal{F}_h(Y_n),$$

which permits to extract the numerical approximation from Y_n. If d is the dimension of the differential equation (1.17) and V is a matrix of dimension $r \times r$ (by abuse of notation we write V in (8.27) instead of the correct $V \otimes I$, where I is the d-dimensional identity matrix), then the vector Y_n is of dimension rd.

If $r > 1$, the recursion of the forward step procedure has parasitic solutions. Our aim is to study the long-time behavior of these parasitic solutions. We are mainly interested in stable methods having good conservation properties. We therefore assume that all eigenvalues of V are simple and lie on the unit circle. We denote them by $\zeta_1 = 1, \zeta_2, \ldots, \zeta_r$. We let v_j and v_j^* be right and left eigenvectors ($V v_j = \zeta_j v_j$ and $v_j^* V = \zeta_j v_j^*$) satisfying $v_j^* v_j = 1$.

To relate the forward step procedure (8.27) to the differential equation (1.17) we assume the pre-consistency condition

$$\Phi(0, Y) = Bf(UY), \qquad Uv_1 = e, \tag{8.28}$$

where B is an $r \times s$ matrix, U an $s \times r$ matrix, and e is the unit vector in $\mathbb{R}^s$. Again, by abuse of notation, we avoid the heavy tensor notation and use matrices B and U instead of $B \otimes I$ and $U \otimes I$. For $UY = W = (W_i)_{i=1}^s \in \mathbb{R}^{sd}$ the vector $f(W) \in \mathbb{R}^{sd}$ is defined by $f(W) = (f(W_i))_{i=1}^s$. We assume throughout this article that the forward step method is consistent, i.e.,

$$v_1^* \Phi(0, yv_1) = f(y), \tag{8.29}$$

and, for pre-consistent methods (8.28), it is equivalent to $v_1^* Be = 1$.

8.7.1 Modified Differential Equations

As discussed for one-step methods, a crucial tool for the study of the long-time behavior of numerical integrators is the backward error analysis, extended to the case of multivalue methods in [122]. This analysis relies on describing the dynamics of the smooth and parasitic components characterizing the numerical solution computed by genuine multivalue methods (i.e., those with $r > 1$).

With the aim of separating the smooth and parasitic components in the numerical solution $y_n = \mathcal{F}_h(Y_n)$, we consider approximations to Y_n of the form

$$\widehat{Y}_n = Y(t_n) + \sum_{j=2}^{r} \zeta_j^n \, Z_j(t_n), \tag{8.30}$$

where $t_n = nh$, and the coefficient functions $Y(t)$, $Z_j(t)$ are independent of n, but depend smoothly on h. Such expansions have first been considered for the study of the long-time behavior of linear multistep methods [187] (also refer to [191, 193] for highly oscillatory problems).

We introduce a system of modified differential equations for the smooth functions $Y(t)$ and $Z_j(t)$. These modified equations only depend on the forward step procedure and are independent of the starting and finishing procedures.

Theorem 8.16 *Consider a forward step procedure (8.27) with matrix V having simple eigenvalues of modulus* 1. *Then, there exist h-independent real functions $f_l(y_1)$ and complex functions $g_{kl}(y_1), a_{jl}(y_1)$ and $b_{jkl}(y_1)$ such*

(continued)

Theorem 8.16 (continued)
that, for an arbitrarily chosen truncation index N and for any solution $y_k(t)$, $z_{kj}(t)$, $j,k = 1,2,\ldots,r$, of the system

$$\begin{aligned}
\dot{y}_1 &= f(y_1) + h\, f_1(y_1) + \ldots + h^{N-1} f_{N-1}(y_1),\\
y_k &= h\, g_{k1}(y_1) + \ldots + h^N g_{k,N}(y_1), \qquad k > 1,\\
\dot{z}_{jj} &= \big(a_{j0}(y_1) + h\, a_{j1}(y_1) + \ldots + h^{N-1} a_{j,N-1}(y_1)\big) z_{jj},\\
z_{jk} &= \big(h\, b_{jk1}(y_1) + \ldots + h^N b_{j,k,N}(y_1)\big) z_{jj}, \qquad k \neq j,
\end{aligned} \tag{8.31}$$

the approximations (8.30), with

$$Y(t) = \sum_{k=1}^{r} y_k(t)\, v_k, \qquad Z_j(t) = \sum_{k=1}^{r} z_{kj}(t)\, v_k, \tag{8.32}$$

satisfy (8.27) with a small defect, i.e.,

$$\widehat{Y}_{n+1} = V\, \widehat{Y}_n + h\, \Phi(h, \widehat{Y}_n) + O(h^{N+1}), + O(h\|\mathbf{Z}\|^2),$$

as long as $y_1(t_n)$ remains in a compact set. The constant symbolized by $O(\cdot)$ is independent of h, but depends on the truncation index N. We use the notation $\|\mathbf{Z}\| = \max\{|z_{jk}(t_n)|\,;\ j,k = 1,\ldots,r\}$.

Proof Inserting (8.30) into the forward step procedure and expanding the nonlinearity around $Y(t_n)$ yields

$$\begin{aligned}
Y(t+h) &= V\, Y(t) + h\, \Phi\big(h, Y(t)\big) + O(h\|\mathbf{Z}\|^2)\\
\zeta_j\, Z_j(t+h) &= V\, Z_j(t) + h\, \Phi'\big(h, Y(t)\big)\, Z_j(t) + O(h\|\mathbf{Z}\|^2).
\end{aligned} \tag{8.33}$$

Neglecting terms of size $O(h\|\mathbf{Z}\|^2)$ and using (8.32), from the previous relation we get

$$y_k(t+h) = \zeta_k\, y_k(t) + h\, v_k^* \Phi\big(h, Y(t)\big).$$

We expand the left-hand side into a Taylor series around $h = 0$ and thus obtain (omitting the argument t)

$$\begin{aligned} \dot{y}_1 + \frac{h}{2}\,\ddot{y}_1 + \cdots &= \Psi_1(h, y_1, \ldots, y_r) \\ (1-\zeta_k)\,y_k + h\,\dot{y}_k + \frac{h^2}{2}\,\ddot{y}_k + \cdots &= h\Psi_k(h, y_1, \ldots, y_r), \quad k = 2, \ldots, r. \end{aligned} \tag{8.34}$$

Differentiation of the relations for y_k $(k = 2, \ldots, r)$ and recursive elimination of the first and higher derivatives, and also of $y_2, \ldots, y_r$ on the right-hand side, yield the second relation of (8.31) with a defect of size $O(h^{N+1})$. In the same way one can eliminate the second and higher derivatives in the first equation of (8.34) and thus obtains a differential equation for y_1. By the consistency assumption (8.29), the h-independent term of this differential equation becomes $f(y_1)$.

Neglecting terms of size $O(h\|\mathbf{Z}\|^2)$ in the second relation of (8.33) yields

$$\zeta_j z_{kj}(t+h) = \zeta_k z_{kj}(t) + h\, v_k^* \Phi'\big(h, Y(t)\big)\, Z_j(t). \tag{8.35}$$

We expand the left-hand side into a Taylor series, and apply the same elimination procedure as for the smooth component $Y(t)$. This then gives a first order differential equation for z_{jj} and algebraic relations for z_{kj} $(k \neq j)$, and terminates the proof of (8.31). □

It is now worth equipping modified differential equations by suitable initial conditions. For $n = 0$ and $\widehat{Y}_0 = Y_0 = \mathcal{S}_h(y_0)$ the relation (8.30) gives

$$\mathcal{S}_h(y_0) = Y(0) + \sum_{j=2}^{r} Z_j(0).$$

Because of the algebraic relations in (8.31), this represents a nonlinear algebraic equation for the h-dependent vectors $y_1(0), z_{22}(0), \ldots, z_{rr}(0)$. For $h = 0$, we get

$$y_1(0)\big|_{h=0} = v_1^* \mathcal{S}_0(y_0), \qquad z_{jj}(0)\big|_{h=0} = v_j^* \mathcal{S}_0(y_0),$$

and the implicit function theorem guarantees the existence of a local unique solution for sufficiently small h.

The initial values $z_{jj}(0)$, for $j = 2, \ldots, r$, determine, on intervals of length $O(1)$, the size of the parasitic solution components. We shall investigate how they depend on the choice of the starting procedure. Let us denote the forward step procedure (8.27) by $Y_{n+1} = \mathcal{G}_h(Y_n)$. We know from Sect. 5.1 (also see Theorem XV.8.2 of [192]) that, for a given $\mathcal{G}_h(Y)$ and a given finishing procedure $\mathcal{F}_h(Y)$, there exist a unique (as formal power series in h) starting procedure $\mathcal{S}_h^*(y)$ and a unique one-step

method $y_{n+1} = \Phi_h^*(y_n)$, such that

$$\mathcal{G}_h \circ \mathcal{S}_h^* = \mathcal{S}_h^* \circ \Phi_h^* \qquad \text{and} \qquad \mathcal{F}_h \circ \mathcal{S}_h^* = \text{identity}. \tag{8.36}$$

This means that for the choice $Y_0 = \mathcal{S}_h^*(y_0)$ the numerical solution obtained by the multivalue method is (formally) equal to that of the one-step method Φ_h^*, the so-called underlying one-step method.

For all common multivalue methods, the underlying one-step method and the components of the starting procedure are B-series. Their coefficients can be computed recursively from the relations (8.36) by using the composition formula for B-series.

Theorem 8.17 *Let the starting procedure $\mathcal{S}_h(y_0)$ satisfy*

$$\mathcal{S}_h(y_0) = \mathcal{S}_h^*(y_0) + O(h^q), \tag{8.37}$$

and assume that the finishing procedure is given by $F_h(Y) = v_1^* Y = y_1$. Then, the initial values for the system of modified equations (8.31) satisfy

$$y_1(0) = y_0 + O(h^q), \qquad z_{jj}(0) = O(h^q).$$

Proof For the exact starting procedure $\mathcal{S}_h^*(y_0)$, the numerical solution $\{y_n\}_{n\geq 0}$ is that of the underlying one-step method and does not have parasitic components. Consequently, we have $y_1(0) = y_0$ and $z_{kj}(0) = 0$ for all k and j. A perturbation of this starting procedure implies, by the implicit function theorem, a perturbation of the same size in the initial values $y_1(0), z_{22}(0), \ldots, z_{rr}(0)$. □

We conclude this section by providing a result regarding the modified differential equations of symmetric multivalue methods, according to the following definition of symmetry.

Definition 8.6 A given multivalue method (8.27) is *symmetric* if its underlying one-step method is a symmetric method.

Theorem 8.18 *Consider a forward step procedure (8.27), where V is of dimension 2 with eigenvalues 1 and -1, and assume that the method is*

(continued)

Theorem 8.18 (continued)
symmetric, therefore mathematically equivalent to

$$Y_n = V\, Y_{n+1} - h\, \Phi(-h, Y_{n+1}).$$

Then, Eq. (8.31) only contain expressions with even powers of h.

Proof Neglecting terms of size $O(h^{N+1})$ and $O(h\|\mathbf{Z}\|^2)$, the functions $Y(t)$ and $Z_j(t)$ of Theorem 8.16 satisfy

$$\begin{aligned} Y(t+h) &= V\, Y(t) + h\, \Phi\big(h, Y(t)\big), \\ \zeta_j\, Z_j(t+h) &= V\, Z_j(t) + h\, \Phi'\big(h, Y(t)\big)\, Z_j(t), \end{aligned} \tag{8.38}$$

where the prime in $\Phi'(h, Y)$ stands for a derivative with respect to Y. Our assumption on the forward step procedure implies that

$$\begin{aligned} Y(t) &= V\, Y(t+h) - h\, \Phi\big(-h, Y(t+h)\big), \\ Z_j(t) &= V\, \zeta_j\, Z_j(t+h) - h\, \Phi'\big(-h, Y(t+h)\big)\, \zeta_j\, Z_j(t+h), \end{aligned}$$

and, replacing $t - h$ for t, leading to

$$\begin{aligned} Y(t-h) &= V\, Y(t) - h\, \Phi\big(-h, Y(t)\big), \\ \zeta_j^{-1} Z_j(t-h) &= V\, Z_j(t) - h\, \Phi'\big(-h, Y(t)\big)\, Z_j(t). \end{aligned} \tag{8.39}$$

Let us first consider the components of the vector $Y(t)$. Comparing the upper relations of (8.38) and (8.39) we notice that the components $y_k(t)$ of $Y(t)$ have to satisfy the same equations for h and for $-h$.

Since, by assumption, $\zeta_2 = -1$ is the only eigenvalue of V different from 1, we have $\zeta_2^{-1} = \zeta_2$. The lower relation of (8.38) is therefore equal to the lower relation of (8.39), where h is replaced by $-h$. Consequently, also the components of $Z_2(t)$ have to satisfy the same equations for h and for $-h$. This implies that all equations of (8.31) are in even powers of h. □

8.7.2 Bounds on the Parasitic Components

The parasitic solution components are determined by the functions $z_{jj}(t)$. To study their long-time behavior we first examine the leading term in the differential

equation (8.31) for z_{jj}. For $k = j$, Eq. (8.35) yields

$$\zeta_j \, \dot{z}_{jj} = v_j^* \Phi'(0, y_1 v_1) \, v_j z_{jj} + O(h|z_{jj}|).$$

Subject to the pre-consistency assumption (8.28), we obtain

$$\dot{z}_{jj} = \mu_j \, f'(y_1) \, z_{jj} + O(h|z_{jj}|), \qquad \mu_j = \zeta_j^{-1} \, v_j^* B U v_j. \tag{8.40}$$

The coefficients μ_j are called *growth parameters* of the multivalue method. They determine to a large extent the long-term behavior of the parasitic components $Z_j(t)$.

It follows from Theorem 8.16 that the coefficient functions of the parasitic solution components (8.32) satisfy

$$\begin{aligned} \dot{z}_{jj} &= h^M A\big(h, y_1(t)\big) \, z_{jj}, \\ z_{jk} &= h \, B\big(h, y_1(t)\big) \, z_{jj}, \qquad k \neq j. \end{aligned} \tag{8.41}$$

In general we have $M = 0$, but if the growth parameters (8.40) of the method are zero we have $M = 1$, and if in addition to zero growth parameters the assumptions of Theorem 8.18 are satisfied we have $M = 2$. If the vector field $f(y)$ of (1.17) is smooth and has bounded derivatives (which excludes stiff and highly oscillatory problems), the functions $A(h, y_1)$ and $B(h, y_1)$ are bounded as long as $y_1(t)$ stays in a compact set. Grönwall lemma then implies

$$\|z_{jj}(t)\| \leq \|z_{jj}(0)\| \, \exp(h^M L t), \tag{8.42}$$

where L is a bound on the norm or, better, the logarithmic norm of $A(h, y_1)$. For $k \neq j$ the functions $z_{jk}(t)$ are bounded by the same expression with an additional factor Ch.

8.7.3 Long-Time Conservation for Hamiltonian Systems

We have built the necessary tools to prove a conservation result for multivalue methods applied to Hamiltonian problems (1.22), as follows.

Theorem 8.19 *Consider a multivalue method of order p, a starting procedure satisfying (8.37) with q, and let $0 \leq M \leq q$ be the integer such that the modified equations for z_{jj}, $j = 2, \ldots, r$, satisfy (8.41). Furthermore, assume the existence of a modified Hamiltonian $\widetilde{\mathcal{H}}(y)$ satisfying $\widetilde{\mathcal{H}}(y) - \mathcal{H}(y) =$*

(continued)

Theorem 8.19 (continued)
$O(h^p)$ which is well preserved by the flow $\widetilde{\varphi}_t(y)$ of the underlying one-step method, more precisely,

$$\widetilde{\mathcal{H}}\big(\widetilde{\varphi}_h(y)\big) = \widetilde{\mathcal{H}}(y) + O(h^{\gamma+1}), \tag{8.43}$$

with $p \le \gamma \le 2q$. We then have, for $t = nh$,

$$\mathcal{H}(y_n) - \mathcal{H}(y_0) = O(h^p) + O(th^\gamma) + O\big(h^{q+1}\exp(h^M Lt)\big),$$

as long as $t = O(h^{-M})$.

Proof Recall that for a given initial value y_0 the numerical solution is obtained from $Y_0 = \mathcal{S}_h(y_0)$, the forward step procedure $Y_{n+1} = VY_n + h\Phi(h, Y_n)$, and the finishing procedure $y_n = \mathcal{F}_h(Y_n)$. The proof consists in several steps.

(a) We use the expansion (8.30) only locally, on one step. This means that, for any n, we compute functions $Y^{[n]}(t)$ and $Z_j^{[n]}(t)$ satisfying the modified equations (8.31), such that

$$Y_n = Y^{[n]}(0) + \sum_{j=2}^{r} Z_j^{[n]}(0).$$

It follows from Theorem 8.16 that (with the choice $N = 2q$)

$$Y_{n+1} = Y^{[n]}(h) + \sum_{j=2}^{r} \zeta_j Z_j^{[n]}(h) + O(h^{2q+1}),$$

as long as the parasitic components are bounded as $\|Z(t)\| = O(h^q)$. By the uniqueness of the initial values, we have that

$$Y^{[n+1]}(0) = Y^{[n]}(h) + O(h^{2q+1}), \qquad Z_j^{[n+1]}(0) = \zeta_j Z_j^{[n]}(h) + O(h^{2q+1}). \tag{8.44}$$

(b) The estimates (8.42) and (8.44) yield

$$\|z_{jj}^{[n+1]}(0)\| \le \|z_{jj}^{[n]}(h)\| + Ch^{2q+1} \le \|z_{jj}^{[n]}(0)\| \exp(h^{M+1}L) + Ch^{2q+1}.$$

Applying a discrete Gronwall Lemma we obtain for $t = nh$

$$\|z_{jj}^{[n]}(0)\| \le \|z_{jj}^{[0]}(0)\| \exp(h^M Lt) + Ch^{2q}\, t\, \exp(h^M Lt). \tag{8.45}$$

(c) We assume that the finishing procedure is given by $\mathcal{F}_h(Y) = v_1^* Y$, so that the flow of the modified equation for y_1 in (8.31) represents the underlying one-step method. We consider the telescoping sum

$$\widetilde{\mathcal{H}}\big(y_1^{[n]}(0)\big) - \widetilde{\mathcal{H}}\big(y_1^{[0]}(0)\big) = \sum_{l=0}^{n-1} \Big(\widetilde{\mathcal{H}}\big(y_1^{[l+1]}(0)\big) - \widetilde{\mathcal{H}}\big(y_1^{[l]}(0)\big) \Big).$$

From the estimate (8.44) and the assumption (8.43) we obtain that every summand is bounded by $O(h^{2q+1}) + O(h^{\gamma+1})$ (the first term can be removed, because $\gamma \le 2q$), which yields an error term of size $O(th^{\gamma})$. In the left-hand side we substitute $y_1^{[n]}(0)$ from the relation

$$y_n = y_1^{[n]}(0) + \sum_{j=2}^{r} z_{1j}^{[n]}(0).$$

The statement now follows from $\|z_{1j}(0)\| \le ch\|z_{jj}(0)\|$, from the bounds (8.45) for $z_{jj}^{[n]}(0)$, and from the assumption $\widetilde{\mathcal{H}}(y) - \mathcal{H}(y) = O(h^p)$.

□

The crucial ingredient of the previous theorem is the existence of a modified Hamiltonian function. Let us discuss some relevant situations where such a modified Hamiltonian is known to exist.

- If the underlying one-step method is a symplectic transformation, there exists a modified Hamiltonian satisfying (8.43) with arbitrarily large γ (see Sect. IX.3 in [192]; also see Theorem 8.15). Unfortunately, the underlying one-step method of multivalue methods cannot be symplectic [190];
- if (1.22) is an integrable reversible system, and if the underlying one-step method is symmetric (reversible), under mild non-resonance conditions there exists a modified Hamiltonian satisfying (8.43) with arbitrarily large γ (see Chapter 9 in [192]);
- if the underlying one-step method is a B-series (this is the case for all general linear methods), necessary and sufficient conditions for the existence of a modified Hamiltonian satisfying (8.43) with a given γ are presented in [192] (Chapter IX.9.4). For example, only one condition is necessary for symmetric methods of order 4 to satisfy condition (8.43) with $\gamma = 6$.

Example 8.8 Let us consider a multivalue method in the following form

$$Y_{n+1} = VY_n + hBf(W), \qquad W = UY_n + hAf(W).$$

(continued)

Example 8.8 (continued)
with

$$\left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right] = \left[\begin{array}{cccc|cc} \frac{1}{12} & 0 & 0 & 0 & 1 & \frac{1}{2} \\ -\frac{1}{3} & \frac{1}{6} & 0 & 0 & 1 & 1 \\ \frac{5}{3} & -\frac{2}{3} & \frac{1}{6} & 0 & 1 & -1 \\ \frac{7}{6} & -\frac{5}{12} & \frac{1}{12} & \frac{1}{12} & 1 & -\frac{1}{2} \\ \hline \frac{2}{3} & -\frac{1}{6} & -\frac{1}{6} & \frac{2}{3} & 1 & 0 \\ 1 & -\frac{1}{2} & \frac{1}{2} & -1 & 0 & -1 \end{array}\right],$$

corresponding to a multivalue method proposed in [73] and analyzed in [122]. The vector

$$Y_n = \begin{bmatrix} y_n \\ a_n \end{bmatrix}$$

provides an approximation y_n to the solution and an approximation a_n to a scaled second derivative. If we denote by $R_h(y_0)$ the result of one step of the Runge-Kutta method

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{2} & \frac{1}{2} & & \\ 1 & \frac{373}{550} & \frac{177}{550} & \\ 0 & \frac{8233}{50976} & -\frac{30749}{152928} & \frac{3025}{76464} \\ \hline & 0 & -\frac{383}{648} & \frac{275}{1296} \quad 1 \end{array}$$

then the starting procedure is given by

$$\mathcal{S}_h(y_0) = \begin{bmatrix} y_0 \\ \frac{1}{2}\big(R_h(y_0) + R_{-h}(y_0)\big) - y_0 \end{bmatrix}.$$

Let us collect some essential properties of this method:

- the method has order $p = 4$, implying that the underlying one-step method has also order 4;
- the method is symmetric in the sense of Theorem 8.18. As a consequence all equations in (8.31) are in even powers of h;

(continued)

Example 8.8 (continued)
- the eigenvalues of V are $\zeta_1 = 1$ and $\zeta_2 = -1$. By construction, the growth parameter corresponding to the parasitic root $\zeta_2 = -1$ is zero. Together with the symmetry of the method this implies that $M = 2$ in (8.41);
- the analysis of $\mathcal{S}_h(y)$ leads to $q = 6$ in the formula (8.17) for the starting procedure (the detailed proof is given in [122]);
- Equation (8.43) is satisfied with $\gamma = 8$ (detailed computations are again given in [122]).

Proposition 8.1 *If the method regarding this example is applied to a Hamiltonian system (1.22), then the Hamiltonian function is nearly preserved according to*

$$\mathcal{H}(y_n) - \mathcal{H}(y_0) = O(h^4) + O(th^8) + O\big(h^8 \exp(h^2 Lt)\big),$$

as long as $t = nh = O(h^{-2})$.

Proof The first two error terms follow directly from Theorem 8.19. From Theorem 8.17 we have that the parasitic solution components satisfy $z_{jj}(0) = O(h^6)$, so that $z_{jj}(t) = O\big(h^6 \exp(h^2 Lt)\big)$. To justify the factor h^8 in front of the exponential term we note that only the functions z_{1j} enter the formula for y_n. By symmetry of the method, we have a factor h^2 in the modified equation (8.31) for z_{1j}. This proves that $z_{1j}(t) = O\big(h^8 \exp(h^2 Lt)\big)$. □

Let us illustrate with numerical experiments that the bounds of Theorem 8.19 and, in particular, those for the parasitic solution components are sharp. In particular we aim to observe that, for multivalue methods for which the order q of the starting procedure is larger or equal than the order p of the method, the parasitic solution components can be neglected on time intervals of length $O(h^{-M})$. On such intervals the underlying one-step method completely describes the qualitative behavior of the method. In particular, if the problem is an integrable reversible system and if the underlying one-step method is symmetric (and reversible), then all action variables are preserved up to an error of size $O(h^p)$. Moreover, the global error increases at most linearly with time.

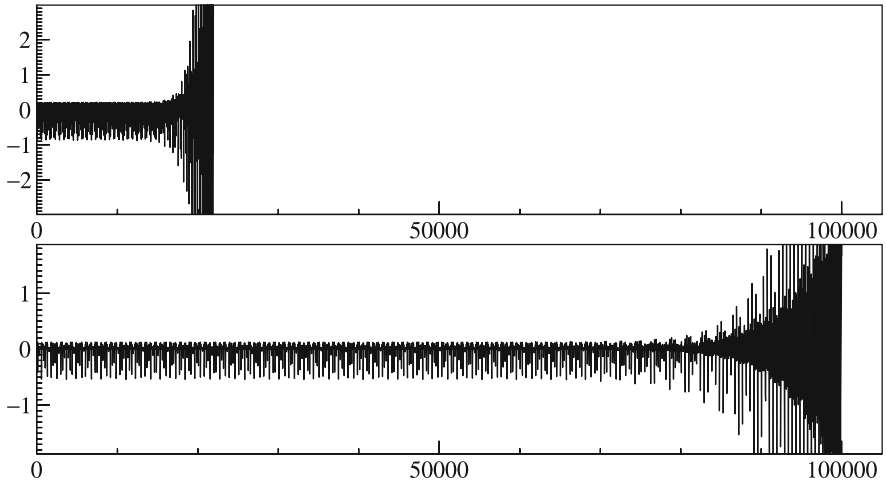


Fig. 8.15 Error in the Hamiltonian for the method in Example 8.8 applied to the mathematical pendulum (1.23), with initial values $q(0) = 3$, $p(0) = 0$. The employed values of h are $h = 0.25$ (top) and $h = 0.125$ (bottom)

Example 8.9 To prove that the estimate of Theorem 8.1 is sharp, we apply the method described in Example 8.8 to the mathematical pendulum (1.23), with initial values $q(0) = 3$, $p(0) = 0$. Figure 8.15 (see [122]) shows the error in the Hamiltonian as a function of time for the step sizes $h = 0.25$ and $h = 0.125$. The scales on the vertical axis differ by a factor 16, so that the $O(h^4)$ behavior of the error can be observed. As predicted by the estimate of Theorem 8.1 the error behaves like $O(h^4)$ on intervals of length $O(h^{-2})$, and then follows an exponential growth. We notice that halving the step size increases the interval of good energy preservation by a factor of 4. This confirms the factor h^2 in the exponential term. The constant L in the estimate, which depends on the problem and on the coefficients of the method, seems to be rather small.

8.8 Exercises

1. Prove that the symplectic Euler method (8.10) is symplectic. The proof requires similar arguments as those used to prove Theorem 8.5.

2. Prove that the implicit midpoint method applied to (1.22), i.e.,

$$y_{n+1} = y_n + hJ^{-1}\nabla\mathcal{H}\left(\frac{y_n + y_{n+1}}{2}\right).$$

   is a symplectic method.
3. Complete the proof of Theorem 8.9, by providing the requested algebraic manipulations.
4. With reference to Example 8.6, compute the modified differential equation associated to the implicit midpoint method (4.24).
5. As highlighted in [192], prove that symplectic Runge-Kutta methods preserve all invariants of the form

$$I(y) = y^{\intercal}Cy + d^{\intercal}y + c.$$

6. As remarked in the explanation of Fig. 8.4, a linear energy drift is visible for the explicit Euler method, that is a non-symplectic method. Give a proof of this fact, i.e.,

$$\mathcal{H}(y_n) = \mathcal{H}(y_0) + O(th^p),$$

   through similar arguments as those provided in the proof of Benettin-Giorgilli theorem (8.15).
7. By using Program 8.2, solve the non-separable Hamiltonian problem whose Hamiltonian is given by

$$\mathcal{H}(p,q) = \frac{p^2}{2(1+U'(q))} + U(q),$$

   being $U(q) = 0.1(q(q-2))^2 + 0.008q^3$, with initial values $p(0) = 0.49$ and $q(0) = 0$, describing the path of a particle of unit mass moving on a wire of shape $U(q)$ [15]. In the numerical solution, focus on the conservation of the Hamiltonian and comment the results.
8. By using Program 8.2, solve the separable Hamiltonian problem whose Hamiltonian is the following polynomial of degree 6 [164]:

$$\mathcal{H}(p,q) = \frac{p^3}{3} - \frac{p}{2} + \frac{q^6}{30} + \frac{q^4}{4} - \frac{q^3}{3} + \frac{1}{6},$$

   by choosing several initial values. In the numerical solution, focus on the conservation of the Hamiltonian and comment the results.

9. Can explicit Runge-Kutta methods be symmetric? Give a proof motivating your answer.
10. Prove that the underlying one-step method of a multivalue method cannot be symplectic. As aforementioned, proofs on non-symplecticity for multivalue method have been given in [71, 190, 250].