# Numerical Geometric Integration 



## Ernst Hairer

## Contents

I Examples and Numerical Experiments ..... 1
I. 1 Two-Dimensional Problems ..... 1
I. 2 Kepler's Problem and the Outer Solar System ..... 4
I. 3 Molecular Dynamics ..... 8
I. 4 Highly Oscillatory Problems ..... 11
I. 5 Exercises ..... 12
II Numerical Integrators ..... 13
II. 1 Implicit Runge-Kutta and Collocation Methods ..... 13
II. 2 Order Conditions, Trees and B-Series ..... 17
II. 3 Adjoint and Symmetric Methods ..... 21
II. 4 Partitioned Runge-Kutta Methods ..... 23
II. 5 Nyström Methods ..... 27
II. 6 Methods Obtained by Composition ..... 27
II. 7 Linear Multistep Methods ..... 29
II. 8 Exercises ..... 29
III Exact Conservation of Invariants ..... 31
III. 1 Examples of First Integrals ..... 31
III. 2 Differential Equations on Lie Groups ..... 34
III. 3 Quadratic Invariants ..... 36
III. 4 Polynomial Invariants ..... 38
III. 5 Projection Methods ..... 40
III. 6 Magnus Series Expansion ..... 43
III. 7 Lie Group Methods ..... 46
III. 8 Exercises ..... 48
IV Symplectic Integration ..... 51
IV. 1 Hamiltonian Systems ..... 51
IV. 2 Symplectic Transformations ..... 52
IV. 3 Symplectic Runge-Kutta Methods ..... 56
IV. 4 Symplecticity for Linear Problems ..... 59
IV. 5 Campbell-Baker-Hausdorff Formula ..... 60
IV. 6 Splitting Methods ..... 62
IV. 7 Volume Preservation ..... 66
IV. 8 Generating Functions ..... 66
IV. 9 Variational Approach ..... 66
IV. 10 Symplectic Integrators on Manifolds ..... 66
IV. 11 Exercises ..... 66
V Backward Error Analysis ..... 68
V. 1 Modified Differential Equation - Examples ..... 68
V. 2 Structure Preservation ..... 72
V.2.1 Reversible Problems and Symmetric Methods ..... 72
V.2.2 Hamiltonian Systems and Symplectic Methods ..... 74
V.2.3 Lie Group Methods ..... 75
V. 3 Modified Equation Expressed with Trees ..... 75
V. 4 Modified Hamiltonian ..... 78
V. 5 Rigorous Estimates - Local Error ..... 78
V.5.1 The Analyticity Assumption ..... 79
V.5.2 Estimation of the Derivatives of the Numerical Solution ..... 79
V.5.3 Estimation of the Coefficients of the Modified Equation ..... 81
V.5.4 Choice of $N$ and the Estimation of the Local Error ..... 83
V.5.5 Estimates for Methods that are not B-series ..... 84
V. 6 Longtime Error Estimates ..... 84
V. 7 Variable Step Sizes ..... 87
V. 8 Exercises ..... 87
VI Kolmogorov-Arnold-Moser Theory ..... 88
VI. 1 Completely Integrable Systems ..... 88
VI. 2 KAM Theorem ..... 88
VI. 3 Linear Error Growth ..... 88
VI. 4 Exercises ..... 88
Bibliography ..... 89

This manuscript has been prepared for the lectures "Intégration géométrique des équations différentielles" given at the University of Geneva during the winter semester 1998/99 ( 2 hours lectures and 1 hour exercises each week). It is directed to students in the third or fourth year. Therefore, a basic knowledge in Analysis, Numerical Analysis and in the theory and solution techniques of differential equations is assumed.

The picture of the title shows parasitic solutions, when the 2-step explicit midpoint rule is applied to the pendulum problem with step size $h=\pi / 10.5$ ( 6000 steps), initial values $q(0)=0, \dot{q}(0)=1$, and exact starting values. It is taken from [Hai99].

At this place I want to thank all my friends and colleagues, with whom I had discussions on this topic and who helped me to a deeper understanding of the subject "Numerical Geometric Integration". In particular, I am grateful to Gerhard Wanner, Christian Lubich, Sebastian Reich, Stig Faltinsen, David Cohen, Robert Chan, ..., and to Pierre Leone who assisted me during the lectures.

## Preface

The numerical solution of ordinary differential equations has a long history. The first and most simple numerical approach was described by L. Euler (1768) in his "Institutiones Calculi Integralis". A.L. Cauchy (1824) established rigorous error bounds for Euler's method and, for this purpose, employed the implicit $\theta$-method (for more details see [HNW93, Sects.I.7, II.1, III.1], [BW96]).

The current most popular methods are Runge-Kutta methods and linear multistep methods. Runge-Kutta methods have been developed by G. Coriolis (1837), C. Runge (1895), K. Heun (1900), W. Kutta (1901), and their analysis has been brought to perfection by J.C. Butcher in the sixties and early seventies. Linear multistep methods have their origin in the work of J.C. Adams, which was carried out in 1855 and published in the book of F. Bashforth (1883). The important work of G. Dahlquist (1959) unified and completed the theory for multistep methods. For a comprehensive study of Runge-Kutta as well as multistep methods we refer the reader to the monographs [Bu87, HNW93, HW96].

The above-mentioned works mainly considered the approximation of one single solution trajectory of the differential equation. Later one became aware of the fact that much new insight can be gained by considering a numerical method as a discrete dynamical system which approximates the flow of the differential equation. This point of view has its origin in the study of symmetric methods (see G. Wanner [Wa73], who used the symbol $m_{h}\left(y_{0}\right)$ for the discrete flow) and is the central theme in the works of Feng Kang, collected in [FK95], U. Kirchgraber [Ki86], W.-J. Beyn [Be87], T. Eirola [Ei88] and in the mongraphs of P. Deuflhard \& F. Bornemann [DB94], J.M. Sanz-Serna \& M.P. Calvo [SC94] and A.M. Stuart \& A.R. Humphries [StH96]. One started to study whether geometrical properties of the flow (such as the preservation of first integrals, the symplecticity, reversibility, or volume-preservation of the flow) can carry over to the numerical discretization. Such properties are crucial for a qualitatively correct simulation of the dynamical system. They turn out to be also responsible for a more favourable error propagation in long-time integrations.

These lecture notes are organized as follows: we start with some examples (related to realistic problems) and illustrate the different qualitative behaviours of numerical methods in Chapter I. We then review numerical integrators and focus our attention on collocation, symmetric and partitioned methods (Chapter II). Geometric properties such as the conservation of first integrals (Chapter III) and the symplecticity for Hamiltonian systems (Chapter IV) are discussed in detail. A main technique for a deeper understanding of geometric properties of numerical methods is "backward error analysis" which will be explained in Chapter V. If time permits, we also consider KAM theory (Chapter VI) as far as it is needed for the explanation of numerical phenomena concerning the longtime integration of integrable systems. The title of these lecture notes is influenced by J.M. Sanz-Serna's survey article [SS97]. We have included the term "numerical" in order to distinguish it clearly from H. Whitney's "Geometric integration theory".

Geneva, date
Ernst Hairer

## Chapter I

## Examples and Numerical Experiments

This chapter introduces some interesting differential equations and illustrates the different qualitative behaviour of numerical methods. We deliberately consider only very simple numerical methods of orders 1 and 2 in order to emphasize the qualitative aspects of the experiments. The same effects (on a different scale) could be observed with more sophisticated higher order integration schemes. The presented experiments should serve as a motivation for the theoretical and practical investigations of later chapters. Every reader is encouraged to repeat the experiments or to invent similar ones.

## I. 1 Two-Dimensional Problems

Volterra-Lotka Problem Consider the problem

$$
\begin{equation*}
u^{\prime}=u(v-2), \quad v^{\prime}=v(1-u), \tag{1.1}
\end{equation*}
$$

which constitutes an autonomous system of two differential equations. It is a simple model for the development of two populations, where $u$ represents the predator and $v$ the prey. If we divide the two equations of (1.1) by each other and if we consider $u$ as a function of $v$ (or reversed), we get after separation of variables

$$
\frac{1-u}{u} d u-\frac{v-2}{v} d v=0 .
$$

Integration then leads to

$$
\begin{equation*}
I(u, v)=\ln u-u+2 \ln v-v=\text { Const. } \tag{1.2}
\end{equation*}
$$

This means that solutions of (1.1) lie on level curves of (1.2) or, equivalently, $I(u, v)$ is a first integral of (1.1). Some of the level curves are drawn in the pictures of Fig. 1.1. Since these level curves are closed, all solutions of (1.1) are periodic. Can we have the same property for the numerical solution?

Simple Numerical Methods The most simple numerical method for the solution of the initial value problem

$$
\begin{equation*}
y^{\prime}=f(y), \quad y\left(t_{0}\right)=y_{0} \tag{1.3}
\end{equation*}
$$



Fig. 1.1: Solutions of the Volterra-Lotka equations (1.1)
is Euler's method

$$
\begin{equation*}
y_{n+1}=y_{n}+h f\left(y_{n}\right) . \tag{1.4}
\end{equation*}
$$

It is also called explicit Euler method, because the approximation $y_{n+1}$ can be computed in an explicit straight-forward way from $y_{n}$ and from the step size $h$. Here, $y_{n}$ is an approximation to $y\left(t_{n}\right)$ where $y(t)$ is the exact solution of (1.3), and $t_{n}=t_{0}+n h$.

The implicit Euler method

$$
\begin{equation*}
y_{n+1}=y_{n}+h f\left(y_{n+1}\right) \tag{1.5}
\end{equation*}
$$

which has its name from its similarity to (1.4), is known from its excellent stability properties. In contrast to (1.4), the approximation $y_{n+1}$ is defined implicitly by (1.5), and the implementation needs the resolution of nonlinear systems.

Taking the mean of $y_{n}$ and $y_{n+1}$ in the argument of $f$, we get the implicit midpoint rule

$$
\begin{equation*}
y_{n+1}=y_{n}+h f\left(\frac{y_{n}+y_{n+1}}{2}\right) . \tag{1.6}
\end{equation*}
$$

It is a symmetric method, which means that the formula remains the same if we exchange $y_{n} \leftrightarrow y_{n+1}$ and $h \leftrightarrow-h$.

For partitioned systems

$$
\begin{equation*}
u^{\prime}=a(u, v), \quad v^{\prime}=b(u, v) \tag{1.7}
\end{equation*}
$$

such as the problem (1.1), we also consider the method

$$
\begin{equation*}
u_{n+1}=u_{n}+h a\left(u_{n+1}, v_{n}\right), \quad v_{n+1}=v_{n}+h b\left(u_{n+1}, v_{n}\right), \tag{1.8}
\end{equation*}
$$

which treats the $u$-variable by the implicit and the $v$-variable by the explicit Euler method. It is called symplectic Euler method (in Sect. IV it will be shown that it represents a symplectic transformation).

Numerical Experiment The result of our first numerical experiment is shown in Fig. 1.1. We have applied different numerical methods to (1.1), all with constant step size $h=0.12$. As initial values (the enlarged symbols in the pictures) we have chosen $\left(u_{0}, v_{0}\right)=(2,2)$ for the explicit Euler method, $\left(u_{0}, v_{0}\right)=(4,8)$ for the implicit Euler method, and $\left(u_{0}, v_{0}\right)=(4,2)$ respectively $(6,2)$ for the symplectic Euler method. The figure shows the numerical approximations of the first 125 steps. We observe that the
explicit and implicit Euler methods behave qualitatively wrong. The numerical solution either spirals outwards or it spirals inwards. The symplectic Euler method, however, gives a numerical solution that lies on a closed curve as does the exact solution. It is important to remark that the curves of the numerical and exact solution do not coincide, but they will be closer for smaller $h$. The implicit midpoint rule also shows the correct qualitative behaviour (we did not include it in the figure).

Pendulum Our next problem is the mathematical pendulum with a massless rod of length $\ell=1$ and mass $m=1$. Its movement is described by the equation $\alpha^{\prime \prime}+\sin \alpha=0$. With the coordinates $q=\alpha$ and $p=\alpha^{\prime}$ this becomes the two-dimensional system


$$
\begin{equation*}
q^{\prime}=p, \quad p^{\prime}=-\sin q \tag{1.9}
\end{equation*}
$$

As in the example above we can find a first integral, so that all solutions satisfy

$$
\begin{equation*}
H(p, q)=\frac{1}{2} p^{2}-\cos q=\text { Const } . \tag{1.10}
\end{equation*}
$$

Since the vector field (1.9) is $2 \pi$-perdiodic in $q$, it is natural to consider $q$ as a variable on the circle $S^{1}$. Hence, the phase space of elements $(p, q)$ becomes the cylinder $\mathbb{R} \times S^{1}$. In Fig. 1.2 level curves of $H(p, q)$ are drawn. They correspond to solution curves of the problem (1.9).


Fig. 1.2: Solutions of the pendulum problem (1.9)
Again we apply our numerical methods: the explicit Euler method with step size $h=0.2$ and initial value $\left(p_{0}, q_{0}\right)=(0,0.5)$; the symplectic Euler method and the implicit midpoint rule with $h=0.3$ and three different initial values $q_{0}=0$ and $p_{0} \in\{0.7,1.4,2.1\}$. Similar to the computations for the Volterra-Lotka equations we observe that only the symplectic Euler method and the implicit midpoint rule exhibit the correct qualitative behaviour. The numerical solution of the midpoint rule is closer to the exact solution, because it is a method of order 2 , whereas the other methods are only of order 1 .

Conclusion We have considered two-dimensional problems with the property that all solutions are periodic. In general, a discretization of the differential equation destroys this property. Surprisingly, there exist methods for which the numerical flow shows the same qualitative behaviour as the exact flow of the problem.

## I. 2 Kepler's Problem and the Outer Solar System

The evolution of the entire planetary system has been numerically integrated for a time span of nearly 100 million years. This calculation confirms that the evolution of the solar system as a whole is chaotic, ...
(G.J. Sussman and J. Wisdom 1992)

The Kepler problem (also called the two-body problem) describes the motion of two bodies which attract each other. If we choose one of the bodies as the center of our coordinate system, the motion will stay in a plane (Exercise 2). Denoting the position of the second body by $q=\left(q_{1}, q_{2}\right)^{T}$, Newton's law yields a second order differential equation which, with a suitable normalization, is given by

$$
\begin{equation*}
\ddot{q}_{1}=-\frac{q_{1}}{\left(q_{1}^{2}+q_{2}^{2}\right)^{3 / 2}}, \quad \ddot{q}_{2}=-\frac{q_{2}}{\left(q_{1}^{2}+q_{2}^{2}\right)^{3 / 2}} . \tag{2.1}
\end{equation*}
$$

One can check that this is equivalent to a Hamiltonian system

$$
\begin{equation*}
\dot{q}=H_{p}(p, q), \quad \dot{p}=-H_{q}(p, q) \tag{2.2}
\end{equation*}
$$

( $H_{p}$ and $H_{q}$ are the vectors of partial derivatives) with total energy

$$
\begin{equation*}
H\left(p_{1}, p_{2}, q_{1}, q_{2}\right)=\frac{1}{2}\left(p_{1}^{2}+p_{2}^{2}\right)-\frac{1}{\sqrt{q_{1}^{2}+q_{2}^{2}}} \tag{2.3}
\end{equation*}
$$

Exact Integration Kepler's problem can be solved analytically, i.e., it can be reduced to the computation of integrals. This is possible, because the system has not only the total energy $H(p, q)$ as invariant, but also the angular momentum

$$
\begin{equation*}
L\left(p_{1}, p_{2}, q_{1}, q_{2}\right)=q_{1} p_{2}-q_{2} p_{1} . \tag{2.4}
\end{equation*}
$$

This can be checked by differentiation. Hence, every solution of (2.1) satisfies the two relations

$$
\frac{1}{2}\left(\dot{q}_{1}^{2}+\dot{q}_{2}^{2}\right)-\frac{1}{\sqrt{q_{1}^{2}+q_{2}^{2}}}=H_{0}, \quad q_{1} \dot{q}_{2}-q_{2} \dot{q}_{1}=L_{0}
$$

where the constants $H_{0}$ and $L_{0}$ are determined by the initial values. Using polar coordinates $q_{1}=r \cos \varphi, q_{2}=r \sin \varphi$, this system becomes

$$
\begin{equation*}
\frac{1}{2}\left(\dot{r}^{2}+r^{2} \dot{\varphi}^{2}\right)-\frac{1}{r}=H_{0}, \quad r^{2} \dot{\varphi}=L_{0} . \tag{2.5}
\end{equation*}
$$

For its solution we consider $r$ as a function of $\varphi$ (assuming that $L_{0} \neq 0$ so that $\varphi$ is a monotonic function). Hence, we have $\dot{r}=\frac{d r}{d \varphi} \cdot \dot{\varphi}$ and the elimination of $\dot{\varphi}$ in (2.5) yields

$$
\frac{1}{2}\left(\left(\frac{d r}{d \varphi}\right)^{2}+r^{2}\right) \frac{L_{0}^{2}}{r^{4}}-\frac{1}{r}=H_{0}
$$

With the abbreviations

$$
\begin{equation*}
d=L_{0}^{2}, \quad e^{2}=1+2 H_{0} L_{0}^{2} \tag{2.6}
\end{equation*}
$$



Fig. 2.1: Exact and numerical solutions of Kepler's problem
and the substitution $u(\varphi)=1 / r(\varphi)$ we get

$$
\left(\frac{d u}{d \varphi}\right)^{2}=\frac{e^{2}}{d^{2}}-\left(u-\frac{1}{d}\right)^{2}
$$

This differential equation can be solved by separation of variables and yields

$$
\begin{equation*}
r(\varphi)=\frac{d}{1+e \cos \left(\varphi-\varphi^{*}\right)}, \tag{2.7}
\end{equation*}
$$

where $\varphi^{*}$ is determined by the initial values $r_{0}$ and $\varphi_{0}$. In the original coordinates this relation becomes

$$
\sqrt{q_{1}^{2}+q_{2}^{2}}+e\left(q_{1} \cos \varphi^{*}+q_{2} \sin \varphi^{*}\right)=d .
$$

Eliminating the square root, this gives a quadratic relation for $\left(q_{1}, q_{2}\right)$ which represents an ellipse with eccentricity $e$ for $H_{0}<0$ (see Fig. 2.1), a parabola for $H_{0}=0$, and a hyperbola for $H_{0}>0$. With the relation (2.7), the second equation of (2.5) gives

$$
\begin{equation*}
\frac{d^{2}}{\left(1+e \cos \left(\varphi-\varphi^{*}\right)\right)^{2}} d \varphi=L_{0} d t \tag{2.8}
\end{equation*}
$$

which, after integration, gives an implicit equation for $\varphi(t)$.
Numerical Integration We consider the problem (2.1) and we choose

$$
\begin{equation*}
q_{1}(0)=1-e, \quad q_{2}(0)=0, \quad \dot{q}_{1}(0)=0, \quad \dot{q}_{2}(0)=\sqrt{\frac{1+e}{1-e}} \tag{2.9}
\end{equation*}
$$



Fig. 2.2: Energy conservation and global error for Kepler's problem
with $0 \leq e<1$ as initial values. This implies that $H_{0}=-1 / 2, L_{0}=\sqrt{1-e^{2}}, d=1-e^{2}$ and $\varphi^{*}=0$. The period of the solution is $2 \pi$ (Exercise 4). Fig. 2.1 shows the exact solution with eccentricity $e=0.6$ and some numerical solutions. After our previous experience, it is no longer a surprise that the explicit Euler method spirals outwards and gives a completely wrong answer. For the symplectic Euler method and the implicit midpoint rule we take a step size 100 times larger in order to better observe their qualitative behaviour. We see that the numerical solution lies close to an ellipse which turns slowly around its focus, clockwise for the symplectic Euler method and anticlockwise for the implicit midpoint rule. The same behaviour can be observed for the exact solution of perturbed Kepler problems (Exercise 6).

Our next experiment (Fig. 2.2) studies the conservation of invariants and the global error. The main observation is that the error in the energy grows linearly for the explicit Euler method, and it remains bounded and small (no secular terms) for the symplectic Euler method. The global error, measured in the Euclidean norm, shows a quadratic growth (explicit Euler) compared to a linear growth (symplectic Euler and implicit midpoint rule). The findings of this experiment are collected in Table 2.1. We remark that the angular momentum $L(p, q)$ is exactly conserved for the symplectic Euler and the implicit midpoint rule.

Table 2.1: Qualitative long-time behaviour for Kepler's problem

| method | error in $H$ | error in $L$ | global error |
| :--- | :---: | :---: | :---: |
| explicit Euler | $\mathcal{O}(t h)$ | $\mathcal{O}(t h)$ | $\mathcal{O}\left(t^{2} h\right)$ |
| symplectic Euler | $\mathcal{O}(h)$ | 0 | $\mathcal{O}(t h)$ |
| implicit midpoint | $\mathcal{O}\left(h^{2}\right)$ | 0 | $\mathcal{O}\left(t h^{2}\right)$ |

Table 2.2: Data for the outer solar system

| planet | mass | initial position | initial velocity |
| :--- | :---: | ---: | ---: |
| Jupiter | $m_{1}=0.000954786104043$ | -3.5023653 | 0.00565429 |
|  |  | -3.8169847 | -0.00412490 |
|  |  | -1.5507963 | -0.00190589 |
| Saturn | $m_{2}=0.000285583733151$ | 9.0755314 | 0.00168318 |
|  |  | -3.0458353 | 0.00483525 |
|  |  | -1.6483708 | 0.00192462 |
| Uranus | $m_{3}=0.0000437273164546$ | -16.3101420 | 0.00354178 |
|  |  | -7.2521278 | 0.00137102 |
|  |  | 11.4707666 | 0.00055029 |
| Neptune | $=0.0000517759138449$ | -25.7294829 | 0.0011452727 |
|  |  | -10.8169456 | 0.00039677 |
|  |  | -15.5387357 | 0.00276725 |
|  |  | -25.2225594 | -0.00170702 |
|  |  | -3.1902382 | -0.00136504 |

Outer Solar System We next apply our methods to the system which describes the motion of the five outer planets relative to the sun. This system has been studied extensively by astronomers, who integrated it for a time span of nearly 100 million years and concluded the chaotic evolution of the solar system [SW92]. The problem is a Hamiltonian system (2.2) with

$$
\begin{equation*}
H(p, q)=\frac{1}{2} \sum_{i=0}^{5} m_{i}^{-1} p_{i}^{T} p_{i}-G \sum_{i=1}^{5} \sum_{j=0}^{i-1} \frac{m_{i} m_{j}}{\left\|q_{i}-q_{j}\right\|} \tag{2.10}
\end{equation*}
$$

Here $p$ and $q$ are the supervectors composed by the vectors $p_{i}, q_{i} \in \mathbb{R}^{3}$ (momenta and positions), respectively. The chosen units are: masses relative to the sun, so that the sun has mass 1. We have taken

$$
m_{0}=1.00000597682
$$

in order to take account of the inner planets. Distances are in astronomical units (1 [A.U.] = 149597870 [km]), times in earth days, and the gravitational constant is

$$
G=2.95912208286 \cdot 10^{-4}
$$

The initial values for the sun are taken as $q_{0}(0)=(0,0,0)^{T}$ and $\dot{q}_{0}(0)=(0,0,0)^{T}$. All other data (masses of the planets and the initial positions and initial velocities) are given in Table 2.2. The initial data are taken from "Ahnerts Kalender für Sternfreunde 1994", Johann Ambrosius Barth Verlag 1993, and they correspond to September 5, 1994 at 0h00. ${ }^{1}$

To this system we applied our four methods, all with step size $h=10$ (days) and over a time period of 200000 days. The numerical solution (see Fig. 2.3) behaves similarly to

[^0]

Fig. 2.3: Solutions of the outer solar system
that for the Kepler problem. With the explicit Euler method the planets increase their energy, they spiral outwards, Jupiter approaches Saturn which leaves the plane of the two-body motion. With the implicit Euler method the planets (first Jupiter and then Saturn) fall into the sun and are thrown far away. Both the symplectic Euler method and the implicit midpoint rule show the correct behaviour. An integration over a much longer time of say several million of years does not deteriorate this behaviour. Let us remark that Sussman \& Wisdom [SW92] have integrated the outer solar system with special methods which will be discussed in Chap.IV.

## I. 3 Molecular Dynamics

We do not need exact classical trajectories to do this, but must lay great emphasis on energy conservation as being of primary importance for this reason.
(M.P. Allen and D.J. Tildesley 1987)

Molecular dynamics requires the solution of Hamiltonian systems (2.2), where the total energy is given by

$$
\begin{equation*}
H(p, q)=\frac{1}{2} \sum_{i=1}^{N} m_{i}^{-1} p_{i}^{T} p_{i}+\sum_{i=2}^{N} \sum_{j=1}^{i-1} V_{i j}\left(\left\|q_{i}-q_{j}\right\|\right) \tag{3.1}
\end{equation*}
$$

and $V_{i j}(r)$ are given potential functions. Here, $q_{i}$ and $p_{i}$ denote the positions and momenta of atoms and $m_{i}$ is the atomic mass of the $i$ th atom. We remark that the outer solar system (2.10) is such an $N$-body system with $V_{i j}(r)=-G m_{i} m_{j} / r$. In molecular dynamics the Lennard-Jones potential

$$
\begin{equation*}
V_{i j}(r)=4 \varepsilon_{i j}\left(\left(\frac{\sigma_{i j}}{r}\right)^{12}-\left(\frac{\sigma_{i j}}{r}\right)^{6}\right) \tag{3.2}
\end{equation*}
$$

is very popular ( $\varepsilon_{i j}$ and $\sigma_{i j}$ are suitable constants depending on the atoms). This potential has an absolute minimum at distance $r=\sigma_{i j} \sqrt[6]{2}$. The force due to this potential strongly repulses the atoms when they are closer than this value, and they attract each other when they are farther.


Störmer-Verlet Scheme The Hamiltonian of (3.1) is of the form $H(p, q)=T(p)+V(q)$, where $T(p)$ is a quadratic function. Hence, the Hamiltonian system is of the form

$$
\dot{q}=M^{-1} p, \quad \dot{p}=-V^{\prime}(q)
$$

where $M=\operatorname{diag}\left(m_{1} I, \ldots, m_{N} I\right)$ and $I$ is the 3 -dimensional identity matrix. This system is equivalent to the special second order differential equation

$$
\begin{equation*}
\ddot{q}=f(q), \tag{3.3}
\end{equation*}
$$

where the right-hand side $f(q)=M^{-1} V^{\prime}(q)$ does not depend on $\dot{q}$. The most natural discretization of (3.3) is ${ }^{2}$

$$
\begin{equation*}
q_{n+1}-2 q_{n}+q_{n-1}=h^{2} f\left(q_{n}\right) \tag{3.4}
\end{equation*}
$$

This formula is either called Störmer's method (C. Störmer in 1907 used higher order variants for the numerical computation concerning the aurora borealis) or Verlet method. L. Verlet [Ver67] proposed this method for computations in molecular dynamics. An approximation to the derivative $v=\dot{q}$ is simply obtained by

$$
\begin{equation*}
v_{n}=\frac{q_{n+1}-q_{n-1}}{2 h} . \tag{3.5}
\end{equation*}
$$

For the second order problem (3.3) one usually has given initial values $q(0)=q_{0}$ and $\dot{q}(0)=v_{0}$. However, one also needs $q_{1}$ in order to be able to start the integration with the 3 -term recursion (3.4). Putting $n=0$ in (3.4) and (3.5), an elimination of $q_{-1}$ gives

$$
q_{1}=q_{0}+h v_{0}+\frac{h^{2}}{2} f\left(q_{0}\right)
$$

for the missing starting value.
The Störmer-Verlet method admits an interesting one-step formulation which is useful for numerical computations. Introducing the velocity approximation at the midpoint $v_{n+1 / 2}:=v_{n}+\frac{h}{2} f\left(q_{n}\right)$, an elimination of $q_{n-1}$ (as above) yields

$$
\begin{align*}
v_{n+1 / 2} & =v_{n}+\frac{h}{2} f\left(q_{n}\right) \\
q_{n+1} & =q_{n}+h v_{n+1 / 2}  \tag{3.6}\\
v_{n+1} & =v_{n+1 / 2}+\frac{h}{2} f\left(q_{n+1}\right)
\end{align*}
$$

[^1]which is an explicit one-step method $\Psi_{h}^{V}:\left(q_{n}, v_{n}\right) \mapsto\left(q_{n+1}, v_{n+1}\right)$ for the first order system $\dot{q}=v, \dot{v}=f(q)$. If one is not interested in the values $v_{n}$ of the derivative, the first and third equations in (3.6) can be replaced with
$$
v_{n+1 / 2}=v_{n-1 / 2}+h f\left(q_{n}\right) .
$$

Finally, let us mention an interesting connection between the Störmer-Verlet method and the symplectic Euler method (1.8). If the variable $q$ is discretized by the explicit Euler and $v$ by the implicit Euler method, we denote it by $\Psi_{h}^{e i}$; if $q$ is discretized by the implicit Euler and $v$ by the explicit Euler method, we denote it by $\Psi_{h}^{i e}$. Introducing $q_{n+1 / 2}:=q_{n}+\frac{h}{2} v_{n+1 / 2}$ as an approximation at the midpoint, one recognizes the mapping $\left(q_{n}, v_{n}\right) \mapsto\left(q_{n+1 / 2}, v_{n+1 / 2}\right)$ as an application of $\Psi_{h / 2}^{e i}$, and $\left(q_{n+1 / 2}, v_{n+1 / 2}\right) \mapsto\left(q_{n+1}, v_{n+1}\right)$ as an application of $\Psi_{h / 2}^{i e}$. Hence, the Störmer-Verlet method satisfies

$$
\begin{equation*}
\Psi_{h}^{V}=\Psi_{h / 2}^{i e} \circ \Psi_{h / 2}^{e i} . \tag{3.7}
\end{equation*}
$$

## Numerical Experiment with a Frozen Argon Crystal

 As in [BS93] we consider the interaction of seven argon atoms in a plane, where six of them are arranged symmetrically around a center atom. As mathematical model we take the Hamiltonian (3.1) with $N=7, m_{i}=m=66.34 \cdot 10^{-27}[\mathrm{~kg}]$,$$
\varepsilon_{i j}=\varepsilon=119.8 k_{B}[\mathrm{~J}], \quad \sigma_{i j}=\sigma=0.341[\mathrm{~nm}],
$$


where $k_{B}=1.380658 \cdot 10^{-23}[\mathrm{~J} / \mathrm{K}]$ is Boltzmann's constant (see [AT87, page 21]). As units for our calculations we take masses in $[\mathrm{kg}]$, distances in nanometers ( $\left.1[\mathrm{~nm}]=10^{-9}[\mathrm{~m}]\right)$, and times in nanoseconds ( $1[\mathrm{nsec}]=10^{-9}[\mathrm{sec}]$ ). Initial positions (in [nm]) and initial velocities (in $[\mathrm{nm} / \mathrm{nsec}]$ ) are given in Table 3.1. They are chosen such that neighbouring atoms have a distance that is close to the one with lowest potential energy, and such that the total momentum is zero and therefore the centre of gravity does not move. The energy at the initial position is $H\left(p_{0}, q_{0}\right) \approx-1260.2 k_{B}[\mathrm{~J}]$.

For computations in molecular dynamics one is usually not interested in the trajectories of the atoms, but one aims at macroscopic quantities such as temperature, pressure, internal energy, etc. We are interested in the total energy, given by the Hamiltonian, and in the temperature which can be calculated from the formula [AT87, page 46]

$$
\begin{equation*}
T=\frac{1}{2 N k_{B}} \sum_{i=1}^{N} m_{i}\left\|\dot{q}_{i}\right\|^{2} . \tag{3.8}
\end{equation*}
$$

We apply the explicit and symplectic Euler methods and also the Verlet method to this problem. Observe that for a Hamiltonian such as (3.1) all three methods are explicit, and

TABLE 3.1: Initial values for the simulation of a frozen Argon crystal

| atom | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| :--- | :---: | ---: | :---: | :---: | :---: | :---: | :---: |
| position | 0.00 | 0.02 | 0.34 | 0.36 | -0.02 | -0.35 | -0.31 |
|  | 0.00 | 0.39 | 0.17 | -0.21 | -0.40 | -0.16 | 0.21 |
| velocity | -30 | 50 | -70 | 90 | 80 | -40 | -80 |
|  | -20 | -90 | -60 | 40 | 90 | 100 | -60 |



Theorem 1.4 The collocation method of Definition 1.3 is equivalent to the s-stage RungeKutta method (1.2) with coefficients

$$
\begin{equation*}
a_{i j}=\int_{0}^{c_{i}} \ell_{j}(\tau) d \tau, \quad b_{i}=\int_{0}^{1} \ell_{i}(\tau) d \tau \tag{1.4}
\end{equation*}
$$

where $\ell_{i}(\tau)$ is the Lagrange polynomial $\ell_{i}(\tau)=\prod_{l \neq i}\left(\tau-c_{l}\right) /\left(c_{i}-c_{l}\right)$.
Proof. Let $u(t)$ be the collocation polynomial and define

$$
k_{i}:=u^{\prime}\left(t_{0}+c_{i} h\right) .
$$

By the Lagrange interpolation formula we have $u^{\prime}\left(t_{0}+\tau h\right)=\sum_{j=1}^{s} k_{j} \cdot \ell_{j}(\tau)$, and by integration we get

$$
u\left(t_{0}+c_{i} h\right)=y_{0}+h \sum_{j=1}^{s} k_{j} \int_{0}^{c_{i}} \ell_{j}(\tau) d \tau
$$

Inserted into (1.3) this gives the first formula of the Runge-Kutta equation (1.2). Integration from 0 to 1 yields the second one.

The above proof can also be read in reverse order. This shows that a Runge-Kutta method, whose coefficients are given by (1.4), can be interpreted as a collocation method. Since $\tau^{k-1}=\sum_{j=1}^{s} c_{j}^{k-1} \ell_{j}(\tau)$ for $k=1, \ldots, s$, the relations (1.4) are equivalent to the linear systems

$$
\begin{array}{ll}
C(q): & \sum_{j=1}^{s} a_{i j} c_{j}^{k-1}=\frac{c_{i}^{k}}{k}, \quad k=1, \ldots, q, \text { all } i \\
B(p): & \sum_{i=1}^{s} b_{i} c_{i}^{k-1}=\frac{1}{k}, \quad k=1, \ldots, p, \text { all } i
\end{array}
$$

with $q=s$ and $p=s$. What is the order of a Runge-Kutta method, whose coefficients $b_{i}, a_{i j}$ are given in this way?

Theorem 1.5 (Superconvergence) If the condition $B(p)$ holds for some $p \geq s$, then the collocation method (1.3) has order p. This means that the collocation method has the same order as the underlying quadrature formula.

Proof. The main idea is to consider the collocation polynomial $u(t)$ as the solution of a perturbed differential equation

$$
\begin{equation*}
u^{\prime}=f(t, u)+\delta(t) \tag{1.5}
\end{equation*}
$$

with defect $\delta(t):=u^{\prime}(t)-f(t, u(t))$. Subtracting (1.1) from (1.5) we get after linearization that

$$
\begin{equation*}
u^{\prime}(t)-y^{\prime}(t)=\frac{\partial f}{\partial y}(t, y(t))(u(t)-y(t))+\delta(t)+r(t) \tag{1.6}
\end{equation*}
$$

where, for $t_{0} \leq t \leq t_{0}+h$, the remainder $r(t)$ is of size $\mathcal{O}\left(\|u(t)-y(t)\|^{2}\right)=\mathcal{O}\left(h^{2 s+2}\right)$ by Lemma 1.6. The variation of constants formula then yields

$$
\begin{equation*}
y_{1}-y\left(t_{0}+h\right)=u\left(t_{0}+h\right)-y\left(t_{0}+h\right)=\int_{t_{0}}^{t_{0}+h} R\left(t_{0}+h, s\right)(\delta(s)+r(s)) d s \tag{1.7}
\end{equation*}
$$

where $R(t, s)$ is the resolvent of the linear part of the differential equation (1.6). The integral over $R\left(t_{0}+h, s\right) r(s)$ gives a $\mathcal{O}\left(h^{2 s+3}\right)$ contribution. The defect $\delta(s)$ vanishes at the collocation points $t_{0}+c_{i} h$ for $i=1, \ldots, s$. Hence, an application of the quadrature formula $\left(b_{i}, c_{i}\right)_{i=1}^{s}$ to the integral over $g(s)=R\left(t_{0}+h, s\right) \delta(s)$ gives zero as result and the quadrature error is bounded by $h^{p+1}$ times a bound of the $p$ th derivative of the function $g(s)$. This derivative is bounded, because by Lemma 1.6 all derivatives of the collocation polynomial are bounded uniformly for $h \rightarrow 0$. Since the order of the quadrature formula always satisfies $p \leq 2 s$, we get $y_{1}-y\left(t_{0}+h\right)=\mathcal{O}\left(h^{p+1}\right)$ from (1.7).

Lemma 1.6 The collocation polynomial $u(t)$ is an sth order approximation to the exact solution of (1.1) on the whole interval $\left[t_{0}, t_{0}+h\right]$, i.e.,

$$
\begin{equation*}
\|u(t)-y(t)\| \leq C \cdot h^{s+1} \tag{1.8}
\end{equation*}
$$

Moreover, the derivatives of $u(t)$ satisfy for $t \in\left[t_{0}, t_{0}+h\right]$

$$
\left\|u^{(k)}(t)-y^{(k)}(t)\right\| \leq C \cdot h^{s+1-k} \quad \text { for } \quad k=0, \ldots, s
$$

Proof. The collocation polynomial satisfies

$$
u^{\prime}\left(t_{0}+\tau h\right)=\sum_{i=1}^{s} f\left(t_{0}+c_{i} h, u\left(t_{0}+c_{i} h\right)\right) \ell_{i}(\tau)
$$

and for the exact solution of (1.1) it holds

$$
y^{\prime}\left(t_{0}+\tau h\right)=\sum_{i=1}^{s} f\left(t_{0}+c_{i} h, y\left(t_{0}+c_{i} h\right)\right) \ell_{i}(\tau)+h^{s} R(\tau, h)
$$

where the interpolation error $R(\tau, h)$ is bounded by $\max _{t \in\left[t_{0}, t_{0}+h\right]}\left\|y^{(s+1)}(t)\right\| / s!$ and its derivatives satisfy $\left\|R^{(k-1)}(\tau, h)\right\| \leq \max _{t \in\left[t_{0}, t_{0}+h\right]}\left\|y^{(s+1)}(t)\right\| /(s-k+1)$ !. This follows from the fact that by Rolle's theorem the polynomial $\sum_{i=1}^{s} f\left(t_{0}+c_{i} h, y\left(t_{0}+c_{i} h\right)\right) \ell_{i}^{(k-1)}(\tau)$ can be interpreted as the interpolation polynomial of $h^{k-1} y^{(k)}\left(t_{0}+\tau h\right)$ at $s-k+1$ points lying in $\left[t_{0}, t_{0}+h\right]$. Integrating the difference of the above two equations gives

$$
\begin{equation*}
y\left(t_{0}+\tau h\right)-u\left(t_{0}+\tau h\right)=h \sum_{i=1}^{s} \Delta f_{i} \int_{0}^{\tau} \ell_{i}(\sigma) d \sigma+h^{s+1} \int_{0}^{\tau} R(\sigma, h) d \sigma \tag{1.9}
\end{equation*}
$$

with $\Delta f_{i}=f\left(t_{0}+c_{i} h, y\left(t_{0}+c_{i} h\right)\right)-f\left(t_{0}+c_{i} h, u\left(t_{0}+c_{i} h\right)\right)$. Using a Lipschitz condition for $f(t, y)$, this relation yields

$$
\max _{t \in\left[t_{0}, t_{0}+h\right]}\|y(t)-u(t)\| \leq h C L \max _{t \in\left[t_{0}, t_{0}+h\right]}\|y(t)-u(t)\|+\text { Const } \cdot h^{s+1}
$$

implying the statement (1.8) for sufficiently small $h>0$.
The proof of the second statement follows from

$$
h^{k}\left(y^{(k)}\left(t_{0}+\tau h\right)-u^{(k)}\left(t_{0}+\tau h\right)\right)=h \sum_{i=1}^{s} \Delta f_{i} \ell_{i}^{(k-1)}(\tau)+h^{s+1} R^{(k-1)}(\tau, h)
$$

by using a Lipschitz condition for $f(t, y)$ and the estimate (1.8).

Example 1.7 (Gauss Methods) If we take for $c_{1}, \ldots, c_{s}$ the zeros of the shifted Legendre polynomial

$$
\frac{d^{s}}{d x^{s}}\left(x^{s}(x-1)^{s}\right)
$$

the interpolatory quadrature formula has order $p=2 s$, and so does the Runge-Kutta (or collocation) method based on these nodes. For $s=1$ we obtain the implicit midpoint rule. The Runge-Kutta coefficients for $s=2$ and $s=3$ are given in Table 1.1.

Table 1.1: Gauss methods of orders 4 and 6

|  |  | $\frac{1}{2}-\frac{\sqrt{15}}{10}$ | $\frac{5}{36}$ | $\frac{2}{9}-\frac{\sqrt{15}}{15}$ | $\frac{5}{36}-\frac{\sqrt{15}}{30}$ |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\frac{1}{2}-\frac{\sqrt{3}}{6}$ | $\frac{1}{4}$ | $\frac{1}{4}-\frac{\sqrt{3}}{6}$ | $\frac{1}{2}$ | $\frac{5}{36}+\frac{\sqrt{15}}{24}$ | $\frac{2}{9}$ | $\frac{5}{36}-\frac{\sqrt{15}}{24}$ |
| $\frac{1}{2}+\frac{\sqrt{3}}{6}$ | $\frac{1}{4}+\frac{\sqrt{3}}{6}$ | $\frac{1}{4}$ | $\frac{1}{2}+\frac{\sqrt{15}}{10}$ | $\frac{5}{36}+\frac{\sqrt{15}}{30}$ | $\frac{2}{9}+\frac{\sqrt{15}}{15}$ | $\frac{5}{36}$ |
|  | $\frac{1}{2}$ | $\frac{1}{2}$ |  | $\frac{5}{18}$ | $\frac{4}{9}$ | $\frac{5}{18}$ |

## II. 2 Order Conditions, Trees and B-Series

Even the standard notation has been found to be too heavy in dealing with fourth and higher order processes, ... (R.H. Merson 1957)

In this section we study the Taylor series of the exact solution of (1.1) and also that of the numerical solution. Their comparison will lead to the order conditions for RungeKutta methods. The computation is much simplified by the use of rooted trees (connected graphs without cycles and a distinguished vertex). This idea has its origin in the study of the composition of differential operators (A. Cayley 1857). Its importance for RungeKutta methods has been pointed out by R.H. Merson in 1957, and the theory has been fully developed by J.C. Butcher in the years 1963-72 (see [Bu87]) and by E. Hairer \& G. Wanner in 1973-74 (see [HNW93]).

In this section we consider autonomous problems

$$
\begin{equation*}
y^{\prime}=f(y), \quad y\left(x_{0}\right)=y_{0} \tag{2.1}
\end{equation*}
$$

where $f: E \rightarrow E\left(E=\mathbb{R}^{n}\right.$ or some Banach space) is sufficiently differentiable, and we denote the independent variable by $x$ (we reserve the letter $t$ for the notation of trees). A problem $y^{\prime}=f(x, y)$ can be brought into this form by adding the equation $x^{\prime}=1$. For the Taylor series expansion of the exact solution of (2.1) we need to compute its derivatives:

$$
\begin{align*}
y^{\prime}\left(x_{0}\right) & =f\left(y_{0}\right) \\
y^{\prime \prime}\left(x_{0}\right) & =\left(f^{\prime} f\right)\left(y_{0}\right)  \tag{2.2}\\
y^{\prime \prime \prime}\left(x_{0}\right) & =\left(f^{\prime \prime}(f, f)\right)\left(y_{0}\right)+\left(f^{\prime} f^{\prime} f\right)\left(y_{0}\right) \\
y^{(4)}\left(x_{0}\right) & =\left(f^{\prime \prime \prime}(f, f, f)\right)\left(y_{0}\right)+3\left(f^{\prime \prime}\left(f^{\prime} f, f\right)\right)\left(y_{0}\right)+\left(f^{\prime} f^{\prime \prime}(f, f)\right)\left(y_{0}\right)+\left(f^{\prime} f^{\prime} f^{\prime} f\right)\left(y_{0}\right)
\end{align*}
$$

The number of expressions that appear increases exponentially, and a systematic simple notation is indipensable. This can be done by associating a tree to every expression in the above formulas. We represent $f$ by a vertex, the first derivative $f^{\prime}$ by a vertex with one upwards leaving branch, and $f^{(k)}$ by a vertex with $k$ upwards leaving branches. The arguments of the $k$-linear mapping $f^{(k)}\left(y_{0}\right)$ correspond to trees that are attached on the ends of the upwards leaving branches. The tree to the right corresponds to $f^{\prime \prime}\left(f^{\prime} f, f\right)$. Other trees are plotted in Table 2.1. Let us now give the necessary formal definitions.


Table 2.1: Trees, elementary differentials, and coefficients

| $\rho(t)$ | $t$ | graph | $\alpha(t)$ | $\gamma(t)$ | $F(t)$ | $\Phi_{i}(t)$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | $\tau$ | - | 1 | 1 | $f$ | 1 |
| 2 | [ $\tau$ ] | $!$ | 1 | 2 | $f^{\prime} f$ | $\sum_{j} a_{i j}$ |
| 3 3 | $\begin{gathered} \hline[\tau, \tau] \\ {[[\tau]]} \\ \hline \end{gathered}$ | \% | 1 1 | 3 6 | $\begin{gathered} f^{\prime \prime}(f, f) \\ f^{\prime} f^{\prime} f \end{gathered}$ | $\begin{aligned} & \sum_{j k} a_{i j} a_{i k} \\ & \sum_{j k} a_{i j} a_{j k} \\ & \hline \end{aligned}$ |
| 4 | $[\tau, \tau, \tau]$ | V | 1 | 4 | $f^{\prime \prime \prime}(f, f, f)$ | $\sum_{j k l} a_{i j} a_{i k} a_{i l}$ |
| 4 | $[[\tau], \tau]$ | 6 | 3 | 8 | $f^{\prime \prime}\left(f^{\prime} f, f\right)$ |  |
| 4 | $[[\tau, \tau]]$ | $Y$ | 1 | 12 | $f^{\prime} f^{\prime \prime}(f, f)$ | $\sum_{j k l} a_{i j} a_{j k} a_{j l}$ |
| 4 | [[[ $[7]]]$ | ? | 1 | 24 | $f^{\prime} f^{\prime} f^{\prime} f$ | $\sum_{j k l} a_{i j} a_{j k} a_{k l}$ |

Definition 2.1 (Trees) The set of (rooted) trees $T$ is recursively defined as follows:
a) the graph $\tau=$ • with only one vertex (called the root) belongs to $T$;
b) if $t_{1}, \ldots, t_{m} \in T$, then the graph obtained by connecting the roots of $t_{1}, \ldots, t_{m}$ to a new vertex also belongs to $T$. It is denoted by

$$
t=\left[t_{1}, \ldots, t_{m}\right]
$$


and the new vertex is the root of $t$.
Examples of this construction are seen in Table 2.1. We remark that some of the trees among $t_{1}, \ldots, t_{m}$ may be equal and that $t$ does not depend on the order of $t_{1}, \ldots, t_{m}$. For example, we do not distinguish between $[[\tau], \tau]$ and $[\tau,[\tau]]$.

## Definition 2.2 (Elementary Differentials and Coefficients)

- For a tree $t \in T$ the elementary differential is a mapping $F(t): E \rightarrow E$, defined recursively by $F(\cdot)(y)=f(y)$ and

$$
F(t)(y)=f^{(m)}(y)\left(F\left(t_{1}\right)(y), \ldots, F\left(t_{m}\right)(y)\right) \quad \text { for } \quad t=\left[t_{1}, \ldots, t_{m}\right]
$$

- The order of a tree $t$, denoted by $\rho(t)$, is the number of its vertices.
- We define the coefficient $\alpha(t)$ by $\alpha(\cdot)=1$ and

$$
\alpha(t)=\binom{\rho(t)-1}{\rho\left(t_{1}\right), \ldots, \rho\left(t_{m}\right)} \alpha\left(t_{1}\right) \cdot \ldots \cdot \alpha\left(t_{m}\right) \frac{1}{\mu_{1}!\mu_{2}!\ldots} \quad \text { for } \quad t=\left[t_{1}, \ldots, t_{m}\right]
$$

where the integers $\mu_{1}, \mu_{2}, \ldots$ count equal trees among $t_{1}, \ldots, t_{m}$.
We shall see below (Theorem 2.5) that $\alpha(t)$ is the integer coefficient appearing in the differentiation process (2.2). It can be interpreted as the number of ways of labelling monotonically the vertices of $t$, starting at the root (Exercise 5).

The Taylor series expansion of the exact solution is $y\left(x_{0}+h\right)=\sum_{k \geq 0}\left(h^{k} / k!\right) y^{(k)}\left(x_{0}\right)$. Because of the relation (2.2) between derivatives $y^{(k)}\left(x_{0}\right)$ and elementary differentials, this motivates the following definition.

Definition 2.3 (B-Series) For a mapping $a: T \cup\{\emptyset\} \rightarrow \mathbb{R}$ a formal series of the form

$$
B(a, y)=a(\emptyset) y+\sum_{t \in T} \frac{h^{\rho(t)}}{\rho(t)!} \alpha(t) a(t) F(t)(y)
$$

is called a $B$-series. ${ }^{1}$
We shall prove that the exact solution $y\left(x_{0}+h\right)$ of (2.1) and the numerical solution $y_{1}$ of a Runge-Kutta method can both be written as B-series. The following lemma is fundamental to its proof.

Lemma 2.4 Let $a: T \cup\{\emptyset\} \rightarrow \mathbb{R}$ be a mapping satisfying $a(\emptyset)=1$. Then, we have

$$
h f(B(a, y))=B\left(a^{\prime}, y\right)
$$

where $a^{\prime}(\emptyset)=0, a^{\prime}(\cdot)=1$, and

$$
a^{\prime}(t)=\rho(t) a\left(t_{1}\right) \cdot \ldots \cdot a\left(t_{m}\right) \quad \text { for } \quad t=\left[t_{1}, \ldots, t_{m}\right]
$$

Proof. Since $a(\emptyset)=1$ we have $B(a, y)=y+\mathcal{O}(h)$, so that $h f(B(a, y))$ can be expanded into a Taylor series around $y$. Using the multilinearity of the derivative $f^{(m)}(y)$ we get

$$
\begin{aligned}
h f(B(a, y))= & h \sum_{m \geq 0} \frac{1}{m!} f^{(m)}(y)(B(a, y)-y)^{m} \\
= & h \sum_{m \geq 0} \frac{1}{m!} \sum_{t_{1} \in T} \cdots \sum_{t_{m} \in T} \frac{h^{\rho\left(t_{1}\right)+\ldots+\rho\left(t_{m}\right)}}{\rho\left(t_{1}\right)!\cdot \ldots \cdot \rho\left(t_{m}\right)!} \alpha\left(t_{1}\right) \cdot \ldots \cdot \alpha\left(t_{m}\right) \\
& \cdot a\left(t_{1}\right) \cdot \ldots \cdot a\left(t_{m}\right) f^{(m)}(y)\left(F\left(t_{1}\right)(y), \ldots, F\left(t_{m}\right)(y)\right) \\
= & \sum_{m \geq 0} \sum_{t_{1} \in T} \cdots \sum_{t_{m} \in T} \frac{h^{\rho(t)}}{(\rho(t)-1)!} \alpha(t) \frac{\mu_{1}!\mu_{2}!\cdot \ldots}{m!} a\left(t_{1}\right) \cdot \ldots \cdot a\left(t_{m}\right) F(t)(y) \\
= & \sum_{t \in T} \frac{h^{\rho(t)}}{(\rho(t)-1)!} \alpha(t) a\left(t_{1}\right) \cdot \ldots \cdot a\left(t_{m}\right) F(t)(y)=B\left(a^{\prime}, y\right) .
\end{aligned}
$$

[^2]The last equality follows from the fact that there are $\binom{m}{\mu_{1}, \mu_{2}, \ldots}$ possibilities for writing the tree $t$ in the form $t=\left[t_{1}, \ldots, t_{m}\right]$.

Theorem 2.5 (B-Series of Exact Solution) The exact solution of (2.1) is a B-series $y\left(x_{0}+h\right)=B\left(e, y_{0}\right)$, where $e(\emptyset)=1$ and

$$
\begin{equation*}
e(t)=1 \quad \text { for all } \quad t \in T \tag{2.3}
\end{equation*}
$$

Proof. From the computation (2.2) it is clear that $y\left(x_{0}+h\right)$ can be written as a B-series $y\left(x_{0}+h\right)=B\left(e, y_{0}\right)$ with $e(\emptyset)=1$ and $e(\cdot)=1$. Furthermore, by Lemma 2.4, this series is a solution of $h y^{\prime}\left(x_{0}+h\right)=h f\left(y\left(x_{0}+h\right)\right)$ if and only if

$$
\sum_{t \in T} \frac{h^{\rho(t)}}{(\rho(t)-1)!} \alpha(t) e(t) F(t)\left(y_{0}\right)=\sum_{t \in T} \frac{h^{\rho(t)}}{\rho(t)!} \alpha(t) e^{\prime}(t) F(t)\left(y_{0}\right) .
$$

A comparison of the coefficients gives

$$
\rho(t) e(t)=e^{\prime}(t)=\rho(t) e\left(t_{1}\right) \cdot \ldots \cdot e\left(t_{m}\right)
$$

which is satisfied by $e(t)$ of (2.3). Hence, by the uniqueness of the Taylor series expansion of the exact solution, $y\left(x_{0}+h\right)=B\left(e, y_{0}\right)$ holds.

Theorem 2.6 (B-Series of Numerical Solution) The numerical solution of a RungeKutta method (1.2) is a B-series $y_{1}=B\left(a, y_{0}\right)$, where $a(\emptyset)=1$ and

$$
\begin{equation*}
a(t)=\gamma(t) \sum_{i=1}^{s} b_{i} \Phi_{i}(t) \quad \text { for } \quad t \in T \tag{2.4}
\end{equation*}
$$

The integer coefficient $\gamma(t)$ is defined by $\gamma(\cdot)=1$ and

$$
\begin{equation*}
\gamma(t)=\rho(t) \gamma\left(t_{1}\right) \cdot \ldots \cdot \gamma\left(t_{m}\right) \quad \text { for } \quad t=\left[t_{1}, \ldots, t_{m}\right] \tag{2.5}
\end{equation*}
$$

and the expression $\Phi_{i}(t)$ by $\Phi_{i}(\cdot)=1$ and

$$
\begin{equation*}
\Phi_{i}(t)=\sum_{j_{1}, \ldots, j_{m}=1}^{s} a_{i j_{1}} \cdot \ldots \cdot a_{i j_{m}} \cdot \Phi_{j_{1}}\left(t_{1}\right) \cdot \ldots \cdot \Phi_{j_{m}}\left(t_{m}\right) \quad \text { for } \quad t=\left[t_{1}, \ldots, t_{m}\right] \tag{2.6}
\end{equation*}
$$

Proof. With the ansatz $h k_{i}=B\left(\Psi_{i}, y_{0}\right)$ the Runge-Kutta equations (1.2) are

$$
B\left(\Psi_{i}, y_{0}\right)=h k_{i}=h f\left(y_{0}+\sum_{j=1}^{s} a_{i j} h k_{j}\right)=h f\left(B\left(\sum_{j=1}^{s} a_{i j} \Psi_{j}, y_{0}\right)\right)
$$

where $\left(\sum_{j} a_{i j} \Psi_{j}\right)(\emptyset):=1$ and $\left(\sum_{j} a_{i j} \Psi_{j}\right)(t)=\sum_{j} a_{i j} \Psi_{j}(t)$ for $t \in T$. Lemma 2.4 gives

$$
\Psi_{i}(t)=\left(\sum_{j} a_{i j} \Psi_{j}\right)^{\prime}(t)=\rho(t)\left(\sum_{j} a_{i j} \Psi_{j}\right)\left(t_{1}\right) \cdot \ldots \cdot\left(\sum_{j} a_{i j} \Psi_{j}\right)\left(t_{m}\right)
$$

By definition of $\gamma(t)$ and $\Phi_{i}(t)$ we have $\Psi_{i}(t)=\gamma(t) \Phi_{i}(t)$ and the statement (2.4) follows from $y_{1}=y_{0}+\sum_{i=1}^{s} b_{i} h k_{i}=B\left(\sum_{i=1}^{s} b_{i} \Psi_{i}, y_{0}\right)$.

Order Conditions for Runge-Kutta Methods Comparing the B-series of the exact and numerical solutions we see that a Runge-Kutta method has order $p$, i.e., $y\left(x_{0}+h\right)-$ $y_{1}=\mathcal{O}\left(h^{p+1}\right)$ for all smooth problems (2.1), if and only if

$$
\begin{equation*}
\sum_{i=1}^{s} b_{i} \Phi_{i}(t)=\frac{1}{\gamma(t)} \quad \text { for } \quad \rho(t) \leq p \tag{2.7}
\end{equation*}
$$

The 'only if' part follows from the independency of the elementary differentials (Exercise 6). This order condition can be immediately read from a tree as follows: attach to every vertex a summation letter (' $i$ ' to the root), then the left-hand expression of (2.7) is a sum over all summation indices with the summand being a product of $b_{i}$, and $a_{j k}$ if the vertex ' $j$ ' is directly connected with ' $k$ ' by an upwards leaving branch. For the tree to the right we thus get

$$
\sum_{i, j, k, l, m, n, p, q, r} b_{i} a_{i j} a_{j m} a_{i n} a_{i k} a_{k l} a_{l q} a_{l r} a_{k p}=\frac{1}{9 \cdot 2 \cdot 5 \cdot 3}
$$

or, by using $\sum_{j} a_{i j}=c_{i}$,

$$
\sum_{i, j, k, l} b_{i} c_{i} a_{i j} c_{j} a_{i k} c_{k} a_{k l} c_{l}^{2}=\frac{1}{270}
$$



The order conditions up to order 4 can be read from Table 2.1, where $\Phi_{i}(t)$ and $\gamma(t)$ are tabulated.

## II. 3 Adjoint and Symmetric Methods

For an autonomous differential equation

$$
\begin{equation*}
y^{\prime}=f(y), \quad y\left(t_{0}\right)=y_{0} \tag{3.1}
\end{equation*}
$$

the solution $y\left(t, t_{0}, y_{0}\right)$ satisfies $y\left(t, t_{0}, y_{0}\right)=y\left(t-t_{0}, 0, y_{0}\right)$. Therefore, it holds for the flow of the differential equation, defined by $\varphi_{t}\left(y_{0}\right)=y\left(t, 0, y_{0}\right)$, that

$$
\varphi_{t} \circ \varphi_{s}=\varphi_{t+s}
$$

and in particular $\varphi_{0}=i d$ and $\varphi_{t} \circ \varphi_{-t}=i d$ (id is the identity map).
A numerical one-step method is a mapping $\Phi_{h}: y_{0} \mapsto y_{1}$, which approximates $\varphi_{h}$. It satisfies $\Phi_{0}=i d$, it is usually also defined for negative $h$, and by the inverse function theorem it is invertible for sufficiently small $h$. In the spirit of 'geometric integration' it is natural to study methods which share the property $\varphi_{t} \circ \varphi_{-t}=i d$ of the exact flow.

Definition 3.1 A numerical one-step method $\Phi_{h}$ is called symmetric, ${ }^{2}$ if it satisfies

$$
\Phi_{h} \circ \Phi_{-h}=i d \quad \text { or equivalently } \quad \Phi_{h}=\Phi_{-h}^{-1}
$$

The numerical method $\Phi_{h}^{*}:=\Phi_{-h}^{-1}$ is called the adjoint method.

[^3]The adjoint operator satisfies the usual properties such as $\left(\Phi_{h}^{*}\right)^{*}=\Phi_{h}$ and $\left(\Phi_{h} \circ \Psi_{h}\right)^{*}=$ $\Psi_{h}^{*} \circ \Phi_{h}^{*}$ for any two one-step methods $\Phi_{h}$ and $\Psi_{h}$.

For the computation of the adjoint method we observe that $y_{1}=\Phi_{h}^{*}\left(y_{0}\right)$ is implicitly defined by $\Phi_{-h}\left(y_{1}\right)=y_{0}$, i.e., $y_{1}$ is the value which yields $y_{0}$ when the method $\Phi_{h}$ is applied with negative step size $-h$. For example, the explicit Euler method in the role of $\Phi_{h}$ gives $y_{1}-h f\left(y_{1}\right)=y_{0}$, and we see that the adjoint of the explicit Euler method is the implicit Euler method. The implicit midpoint rule (I.1.6) is invariant with respect to the transformation $y_{0} \leftrightarrow y_{1}$ and $h \leftrightarrow-h$. Therefore, it is a symmetric method.

Theorem 3.2 Let $\varphi_{t}$ be the exact flow of (3.1) and let $\Phi_{h}$ be a one-step method of order p satisfying

$$
\begin{equation*}
\Phi_{h}\left(y_{0}\right)=\varphi_{h}\left(y_{0}\right)+C\left(y_{0}\right) h^{p+1}+\mathcal{O}\left(h^{p+2}\right) \tag{3.2}
\end{equation*}
$$

Then, the adjoint method $\Phi_{h}^{*}$ has the same order $p$ and it holds that

$$
\begin{equation*}
\Phi_{h}^{*}\left(y_{0}\right)=\varphi_{h}\left(y_{0}\right)+(-1)^{p} C\left(y_{0}\right) h^{p+1}+\mathcal{O}\left(h^{p+2}\right) . \tag{3.3}
\end{equation*}
$$

Moreover, if $\Phi_{h}$ is symmetric, its (maximal) order is even.
Proof. We replace $h$ and $y_{0}$ in Eq. (3.2) by $-h$ and $\varphi_{h}\left(y_{0}\right)$, respectively. This gives

$$
\begin{equation*}
\Phi_{-h}\left(\varphi_{h}\left(y_{0}\right)\right)=y_{0}+C\left(\varphi_{h}\left(y_{0}\right)\right)(-h)^{p+1}+\mathcal{O}\left(h^{p+2}\right) . \tag{3.4}
\end{equation*}
$$

Since $\varphi_{h}\left(y_{0}\right)=y_{0}+\mathcal{O}(h)$ and $\Phi_{-h}^{\prime}\left(y_{0}\right)=I+\mathcal{O}(h)$, it follows from the inverse function theorem that $\left(\Phi_{-h}^{-1}\right)^{\prime}\left(y_{0}\right)=I+\mathcal{O}(h)$. Applying the function $\Phi_{-h}^{-1}$ to (3.4) yields

$$
\begin{aligned}
\varphi_{h}\left(y_{0}\right) & =\Phi_{-h}^{-1}\left(y_{0}+C\left(\varphi_{h}\left(y_{0}\right)\right)(-h)^{p+1}+\mathcal{O}\left(h^{p+2}\right)\right) \\
& =\Phi_{h}^{*}\left(y_{0}\right)+C\left(y_{0}\right)(-h)^{p+1}+\mathcal{O}\left(h^{p+2}\right),
\end{aligned}
$$

implying (3.3). The statement for symmetric methods is an immediate consequence of this result, because $\Phi_{h}=\Phi_{h}^{*}$ implies $C\left(y_{0}\right)=(-1)^{p} C\left(y_{0}\right)$, and therefore $C\left(y_{0}\right)$ can be different from zero only for even $p$.
Theorem 3.3 ([Ste73, Wa73]) The adjoint method of an s-stage Runge-Kutta method (1.2) is again an s-stage Runge-Kutta method. Its coefficients are given by

If

$$
\begin{equation*}
a_{i j}^{*}=b_{s+1-j}-a_{s+1-i, s+1-j}, \quad b_{i}^{*}=b_{s+1-i} . \tag{3.5}
\end{equation*}
$$

$$
\begin{equation*}
a_{s+1-i, s+1-j}+a_{i j}=b_{j} \quad \text { for all } i, j, \tag{3.6}
\end{equation*}
$$

the Runge-Kutta method (1.2) is symmetric. ${ }^{3}$
Proof. Let $\Phi_{h}$ denote the Runge-Kutta method (1.2). The numerical solution of the adjoint method $y_{1}=\Phi_{h}^{*}\left(y_{0}\right)$ is given by $y_{0}=\Phi_{-h}\left(y_{1}\right)$. By exchanging $y_{0} \leftrightarrow y_{1}$ and $h \leftrightarrow-h$ we thus obtain

$$
\begin{equation*}
k_{i}=f\left(y_{0}+h \sum_{j=1}^{s}\left(b_{j}-a_{i j}\right) k_{j}\right), \quad y_{1}=y_{0}+h \sum_{i=1}^{s} b_{i} k_{i} . \tag{3.7}
\end{equation*}
$$

Since the values $\sum_{j=1}^{s}\left(b_{j}-a_{i j}\right)=1-c_{i}$ appear in reverse order, we replace $k_{i}$ by $k_{s+1-i}$ in (3.7), and then we substitute all indices $i$ and $j$ by $s+1-i$ and $s+1-j$, respectively. This proves (3.5).

The assumption (3.6) implies $a_{i j}^{*}=a_{i j}$ and $b_{i}^{*}=b_{i}$, so that $\Phi_{h}^{*}=\Phi_{h}$.

[^4]For the methods of Table 1.1 one can directly check that the condition (3.6) holds. Furthermore, all Gauss methods are symmetric (Exercise 10).

## II. 4 Partitioned Runge-Kutta Methods

Some interesting numerical methods introduced in Chapter I (symplectic Euler and the Störmer-Verlet method) do not belong to the class of Runge-Kutta methods. They are important examples of so-called partitioned Runge-Kutta methods. In this section we consider differential equations in the partitioned form

$$
\begin{equation*}
p^{\prime}=f(p, q), \quad q^{\prime}=g(p, q) \tag{4.1}
\end{equation*}
$$

We have chosen the letters $p$ and $q$ for the dependent variables, because Hamiltonian systems (I.2.2) are of this form and they are of particular interest in this lecture.

Definition 4.1 Let $b_{i}, a_{i j}$ and $\widehat{b}_{i}, \widehat{a}_{i j}$ be the coefficients of two Runge-Kutta methods. A partitioned Runge-Kutta method for the solution of (4.1) is given by

$$
\begin{align*}
k_{i} & =f\left(p_{0}+h \sum_{j=1}^{s} a_{i j} k_{j}, q_{0}+h \sum_{j=1}^{s} \widehat{a}_{i j} \ell_{j}\right) \\
\ell_{i} & =g\left(p_{0}+h \sum_{j=1}^{s} a_{i j} k_{j}, q_{0}+h \sum_{j=1}^{s} \widehat{a}_{i j} \ell_{j}\right)  \tag{4.2}\\
p_{1} & =p_{0}+h \sum_{i=1}^{s} b_{i} k_{i}, \quad q_{1}=q_{0}+h \sum_{i=1}^{s} \widehat{b}_{i} \ell_{i} .
\end{align*}
$$

Methods of this type have originally been proposed by Hofer (1976) and Griepentrog (1978) for problems with stiff and nonstiff parts (see [HNW93, Sect.II.15]). Their importance for Hamiltonian systems has been discovered only very recently.

An interesting example is the symplectic Euler method (I.1.8), where the implicit Euler method $b_{1}=1, a_{11}=1$ is combined with the explicit Euler method $\widehat{b}_{1}=1, \widehat{a}_{11}=0$. The Störmer-Verlet method (I.3.6) is of the form (4.2) with coefficients given in Table 4.1.

The theory of Runge-Kutta methods can be extended in a straight-forward way to partitioned methods. Since (4.2) is a one-step method $\left(p_{1}, q_{1}\right)=\Phi_{h}\left(p_{0}, q_{0}\right)$, the order, the adjoint method and symmetric methods are defined in the usual way.

Explicit Symmetric Methods An interesting feature of partitioned Runge-Kutta methods is the possibility of having explicit, symmetric methods for problems of the form

$$
\begin{equation*}
p^{\prime}=f(q), \quad q^{\prime}=g(p) \tag{4.3}
\end{equation*}
$$

e.g., if the problem is Hamiltonian with separable $H(p, q)=T(p)+V(q)$. This is not possible with classical Runge-Kutta methods (Exercise 9).

Table 4.1: Störmer-Verlet as partitioned Runge-Kutta method

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

| $1 / 2$ | $1 / 2$ | 0 |
| :---: | :---: | :---: |
| $1 / 2$ | $1 / 2$ | 0 |
|  | $1 / 2$ | $1 / 2$ |

Exactly as in the proof of Theorem 3.3 one can show that the partitioned method (4.2) is symmetric, if both Runge-Kutta methods are symmetric, i.e., if the coefficients of both methods satisfy (3.6). The method is explicit for problems (4.3), if $a_{i j}=\widehat{a}_{i j}=0$ for $i<j$, and if $a_{i i} \cdot \widehat{a}_{i i}=0$ for all $i$. An example of a symmetric method, which is also explicit for (4.3), is the Störmer-Verlet method (Table 4.1). We shall see later in Sect. II. 6 that explicit symmetric methods of order higher than 2 exist.

Bi-Colored Trees and P-Series We show here how the derivation of the order conditions of Sect.II. 2 (trees, elementary differentials, B-series) can be extended to partitioned Runge-Kutta methods. We start by computing the derivatives of the exact solution of (4.1) which are needed for the Taylor series expansion:

$$
\begin{align*}
p^{\prime}\left(x_{0}\right)= & f\left(p_{0}, q_{0}\right) \\
p^{\prime \prime}\left(x_{0}\right)= & \left(f_{p} f\right)\left(p_{0}, q_{0}\right)+\left(f_{q} g\right)\left(p_{0}, q_{0}\right)  \tag{4.4}\\
p^{\prime \prime \prime}\left(x_{0}\right)= & \left(f_{p p}(f, f)\right)\left(p_{0}, q_{0}\right)+2\left(f_{p q}(f, g)\right)\left(p_{0}, q_{0}\right)+\left(f_{q q}(g, g)\right)\left(p_{0}, q_{0}\right) \\
& +\left(f_{p} f_{p} f\right)\left(p_{0}, q_{0}\right)+\left(f_{p} f_{q} g\right)\left(p_{0}, q_{0}\right)+\left(f_{q} g_{p} f\right)\left(p_{0}, q_{0}\right)+\left(f_{q} g_{q} g\right)\left(p_{0}, q_{0}\right)
\end{align*}
$$

Here, $f_{p}, f_{q}, f_{p q}, \ldots$ denote partial derivatives. Similar expressions are obtained for the derivatives of $q(x)$.

Since two functions are involved in the expressions of (4.4), we need two different types of vertices for a graphical representation. We take a 'black' vertex for $f$ and a 'white' vertex for $g$. Upwards leaving branches again represent derivatives, with respect to $p$ if the upper end of the branch is a black vertex, and with respect to $q$ if it is a white vertex. With this convention, the graph to the
 right corresponds to the expression $f_{q p}\left(g_{p q}(f, g), f\right)$.

We denote by $T P$ the set of graphs obtained by the above procedure, and we call them (rooted) bi-colored trees. The graphs $\tau_{p}=\cdot$ and $\tau_{q}=\circ$ belong to $T P$ as well as

$$
\left[t_{1}, \ldots, t_{m}\right]_{p} \quad \text { and } \quad\left[t_{1}, \ldots, t_{m}\right]_{q}
$$

which, for $t_{1}, \ldots, t_{m} \in T P$, denote bi-colored trees as in Definition 2.1, and where the subscripts $p$ and $q$ indicate that the colors of the new roots are black and white, respectively (see Table 4.2 for some examples). Furthermore, we denote by $T P_{p}$ and $T P_{q}$ the subsets of $T P$ which are formed by trees with black and white roots, respectively. Hence, the trees of $T P_{p}$ correspond to derivatives of $p(x)$, whereas those of $T P_{q}$ correspond to derivatives of $q(x)$.

The elementary differentials $F(t)(p, q)$ (for $t \in T P$ ) are defined by the correspondence between bi-colored trees and the expressions in (4.4) explained above. Examples can be seen in Table 4.2. As in Definition 2.2 we denote by $\rho(t)$ the number of vertices of $t \in T P$, and we call it the order of $t$. The coefficient $\alpha(t)$ is defined by $\alpha\left(\tau_{p}\right)=\alpha\left(\tau_{q}\right)=1$, and for $t=\left[t_{1}, \ldots, t_{m}\right]_{p}$ or $t=\left[t_{1}, \ldots, t_{m}\right]_{q}$ by

$$
\begin{equation*}
\alpha(t)=\binom{\rho(t)-1}{\rho\left(t_{1}\right), \ldots, \rho\left(t_{m}\right)} \alpha\left(t_{1}\right) \cdot \ldots \cdot \alpha\left(t_{m}\right) \frac{1}{\mu_{1}!\mu_{2}!\ldots} \tag{4.5}
\end{equation*}
$$

where the integers $\mu_{1}, \mu_{2}, \ldots$ count equal trees among $t_{1}, \ldots, t_{m} \in T P$. This is formally the same definition as in Sect.II.2. Observe, however, that $\alpha(t)$ depends on the coloring of the vertices. For example, we have $\alpha\left(\left[\tau_{p}, \tau_{p}\right]_{p}\right)=1$, but $\alpha\left(\left[\tau_{p}, \tau_{q}\right]_{p}\right)=2$.

TABLE 4.2: Bi-colored trees, elementary differentials, and coefficients

| $\rho(t)$ | $t$ | graph | $\alpha(t)$ | $\gamma(t)$ | $F(t)$ | $\Phi_{i}(t)$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | $\tau_{p}$ | - | 1 | 1 | $f$ | 1 |
| 2 | $\left[\tau_{p}\right]_{p}$ | $\checkmark$ | 1 | 2 | $f_{p} f$ | $\sum_{j} a_{i j}$ |
| 2 | $\left[\tau_{q}\right]_{p}$ | \% | 1 | 2 | $f_{q} g$ | $\sum_{j} \widehat{a}_{i j}$ |
| 3 | $\left[\tau_{p}, \tau_{p}\right]_{p}$ | $V$ | 1 | 3 | $f_{p p}(f, f)$ | $\sum_{j k} a_{i j} a_{i k}$ |
| 3 | $\left[\tau_{p}, \tau_{q}\right]_{p}$ | $\gamma$ | 2 | 3 | $f_{p q}(f, g)$ | $\sum_{j k} a_{i j} \widehat{a}_{i k}$ |
| 3 | $\left[\tau_{q}, \tau_{q}\right]_{p}$ | $\vartheta$ | 1 | 3 | $f_{q q}(g, g)$ | $\sum_{j k} \widehat{a}_{i j} \widehat{a}_{i k}$ |
| 3 | $\left[\left[\tau_{p}\right]_{p}\right]_{p}$ | ? | 1 | 6 | $f_{p} f_{p} f$ | $\sum_{j k} a_{i j} a_{j k}$ |
| 3 | $\left[\left[\tau_{q}\right]_{p}\right]_{p}$ | . | 1 | 6 | $f_{p} f_{q} g$ | $\sum_{j k} a_{i j} \widehat{a}_{j k}$ |
| 3 | $\left[\left[\tau_{p}\right]_{q}\right]_{p}$ | ? | 1 | 6 | $f_{q} g_{p} f$ | $\sum_{j k} \widehat{a}_{i j} a_{j k}$ |
| 3 | $\left[\left[\tau_{q}\right]_{q}\right]_{p}$ | ? | 1 | 6 | $f_{q} g_{q} g$ | $\sum_{j k} \widehat{a}_{i j} \widehat{a}_{j k}$ |

Definition 4.2 (P-Series) For a mapping $a: T P \cup\left\{\emptyset_{p}, \emptyset_{q}\right\} \rightarrow \mathbb{R}$ a series of the form

$$
P(a,(p, q))=\binom{P_{p}(a,(p, q))}{P_{q}(a,(p, q))}=\binom{a\left(\emptyset_{p}\right) p+\sum_{t \in T P_{p}} \frac{h^{\rho(t)}}{\rho(t)!} \alpha(t) a(t) F(t)(p, q)}{a\left(\emptyset_{q}\right) q+\sum_{t \in T P_{q}} \frac{h^{\rho(t)}}{\rho(t)!} \alpha(t) a(t) F(t)(p, q)}
$$

is called a $P$-series.

The following results are obtained in exactly the same manner as the corresponding results for non-partitioned Runge-Kutta methods (Sect.II.2). We therefore omit their proofs.

Lemma 4.3 Let $a: T P \cup\left\{\emptyset_{p}, \emptyset_{q}\right\} \rightarrow \mathbb{R}$ satisfy $a\left(\emptyset_{p}\right)=a\left(\emptyset_{q}\right)=1$. Then, it holds

$$
h\binom{f(P(a,(p, q)))}{g(P(a,(p, q)))}=P\left(a^{\prime},(p, q)\right),
$$

where $a^{\prime}\left(\emptyset_{p}\right)=a^{\prime}\left(\emptyset_{q}\right)=0, a^{\prime}\left(\tau_{p}\right)=a^{\prime}\left(\tau_{q}\right)=1$, and

$$
\begin{equation*}
a^{\prime}(t)=\rho(t) a\left(t_{1}\right) \cdot \ldots \cdot a\left(t_{m}\right) \tag{4.6}
\end{equation*}
$$

if either $t=\left[t_{1}, \ldots, t_{m}\right]_{p}$ or $t=\left[t_{1}, \ldots, t_{m}\right]_{q}$.
Theorem 4.4 (P-Series of Exact Solution) The exact solution of (4.1) is a P-series $\left(p\left(x_{0}+h\right), q\left(x_{0}+h\right)\right)=P\left(e,\left(p_{0}, q_{0}\right)\right)$, where $e\left(\emptyset_{p}\right)=e\left(\emptyset_{q}\right)=1$ and

$$
\begin{equation*}
e(t)=1 \quad \text { for all } \quad t \in T P \tag{4.7}
\end{equation*}
$$

Theorem 4.5 (P-Series of Numerical Solution) The numerical solution of a partitioned Runge-Kutta method (4.2) is a P-series $\left(p_{1}, q_{1}\right)=P\left(a,\left(p_{0}, q_{0}\right)\right)$, where $a\left(\emptyset_{p}\right)=$ $a\left(\emptyset_{q}\right)=1$ and

$$
a(t)=\left\{\begin{array}{lll}
\gamma(t) \sum_{i=1}^{s} b_{i} \Phi_{i}(t) & \text { for } & t \in T P_{p}  \tag{4.8}\\
\gamma(t) \sum_{i=1}^{s} \widehat{b}_{i} \Phi_{i}(t) & \text { for } & t \in T P_{q}
\end{array}\right.
$$

The integer coefficient $\gamma(t)$ is the same as in (2.5). It does not depend on the color of the vertices. The expression $\Phi_{i}(t)$ is defined by $\Phi_{i}\left(\tau_{p}\right)=\Phi_{i}\left(\tau_{q}\right)=1$ and by

$$
\Phi_{i}(t)=\Psi_{i}\left(t_{1}\right) \cdot \ldots \cdot \Psi_{i}\left(t_{m}\right) \quad \text { with } \quad \Psi_{i}\left(t_{k}\right)= \begin{cases}\sum_{j_{k}=1}^{s} a_{i j_{k}} \Phi_{j_{k}}\left(t_{k}\right) & \text { if } t_{k} \in T P_{p}  \tag{4.9}\\ \sum_{j_{k}=1}^{s} \widehat{a}_{i j_{k}} \Phi_{j_{k}}\left(t_{k}\right) & \text { if } t_{k} \in T P_{q}\end{cases}
$$

for $t=\left[t_{1}, \ldots, t_{m}\right]_{p}$ or $t=\left[t_{1}, \ldots, t_{m}\right]_{q}$.

Order Conditions Comparing the P-series of the exact and numerical solutions we see that a partitioned Runge-Kutta method (4.2) has order $r$, i.e., $p\left(x_{0}+h\right)-p_{1}=\mathcal{O}\left(h^{r+1}\right)$, $q\left(x_{0}+h\right)-q_{1}=\mathcal{O}\left(h^{r+1}\right)$, if and only if

$$
\begin{array}{ll}
\sum_{i=1}^{s} b_{i} \Phi_{i}(t)=\frac{1}{\gamma(t)} & \text { for } t \in T P_{p} \text { with } \rho(t) \leq r  \tag{4.10}\\
\sum_{i=1}^{s} \widehat{b}_{i} \Phi_{i}(t)=\frac{1}{\gamma(t)} & \text { for } t \in T P_{q} \text { with } \rho(t) \leq r
\end{array}
$$

This means that not only every individual Runge-Kutta method has to be of order $r$, but also so-called coupling conditions between the coefficients of both methods have to hold.

As in Sect.II. 2 the order conditions can be directly read from the trees: we attach to every vertex a summation letter (' $i$ ' to the root). Then the left-hand expression of (4.10) is a sum over all summation indices with the summand being a product of $b_{i}$ or $\widehat{b}_{i}$ (depending on whether the root is black or white) and of $a_{j k}$ (if ' $k$ ' is black) or $\hat{a}_{j k}$ (if ' $k$ ' is white), if the vertex ' $k$ ' is directly above ' $j$ '. For the tree to the right we thus obtain

$$
\sum_{i, j, k, l, m, n, p, q, r} b_{i} \widehat{a}_{i j} \widehat{a}_{j m} \widehat{a}_{i n} a_{i k} \widehat{a}_{k l} a_{l q} a_{l r} a_{k p}=\frac{1}{9 \cdot 2 \cdot 5 \cdot 3}
$$

or, by using $\sum_{j} a_{i j}=c_{i}$ and $\sum_{j} \widehat{a}_{i j}=\widehat{c}_{i}$,

$$
\sum_{i, j, k, l} b_{i} \widehat{c}_{i} \widehat{a}_{i j} \widehat{c}_{j} a_{i k} c_{k} \widehat{a}_{k l} c_{l}^{2}=\frac{1}{270} .
$$



The order conditions up to order 3 can be read from Table 4.2, where $\Phi_{i}(t)$ and $\gamma(t)$ are tabulated.

Example 4.6 (Lobatto IIIA - IIIB Pair) We let $c_{1}, \ldots, c_{s}$ be the zeros of

$$
\frac{d^{s-2}}{d x^{s-2}}\left(x^{s-1}(x-1)^{s-1}\right)
$$

and we consider the interpolatory quadrature formula $\left(b_{i}, c_{i}\right)_{i=1}^{s}$ based on these nodes. The special cases $s=2$ and $s=3$ are the trapezoidal rule and Simpson's rule. We then define the Runge-Kutta coefficients $a_{i j}$ (Lobatto IIIA) and $\widehat{a}_{i j}$ (Lobatto IIIB) by the following conditions (see Sect.II. 2 for the definition of $B(p)$ and $C(q))$ :

| Lobatto IIIA | $B(2 s-2), C(s) \quad$ "collocation" |
| :--- | :--- |
| Lobatto IIIB | $B(2 s-2), C(s-2)$ and $\widehat{a}_{i 1}=b_{1}, \widehat{a}_{i s}=0$ for $i=1, \ldots, s$. |

For $s=2$ we get the Störmer-Verlet method of Table 4.1. The coefficients of the methods for $s=3$ are given in Table 4.3. One can prove that this partitioned Runge-Kutta method is of order $p=2 s-2$. Instead of giving the general proof (see e.g., [HW96, page 563]) we check the order for the case $s=3$. Due to the simplifying assumptions $C(s)$ for Lobatto IIIA and $C(s-2)$ for Lobatto IIIB the order conditions for all bi-colored trees up to order 3 are immediately verified (using the expressions of Table 4.2). Order 4 is then a consequence of Theorem 3.2 and the fact that both Runge-Kutta methods are symmetric.

## II. 5 Nyström Methods

as important special case of partitioned Runge-Kutta methods, M.P. Calvo

## II. 6 Methods Obtained by Composition

An interesting means for the construction of higher order integration methods is by composition of simple methods. We have already seen in Chapter I that the Störmer-Verlet method can be considered as the composition of two symplectic Euler methods. The results of this section are valid for general one-step methods (partitioned as well as nonpartitioned ones).

Theorem 6.1 ([Yo90]) Let $\Phi_{h}(y)$ be a symmetric one-step method of order $p=2 k$. If

$$
\begin{equation*}
2 b_{1}+b_{0}=1, \quad 2 b_{1}^{2 k+1}+b_{0}^{2 k+1}=0, \tag{6.1}
\end{equation*}
$$

then the composed method

$$
\Psi_{h}=\Phi_{b_{1} h} \circ \Phi_{b_{0} h} \circ \Phi_{b_{1} h}
$$

is symmetric and has order $p=2 k+2$.
Proof. The basic method satisfies $\Phi_{h}\left(y_{0}\right)=\varphi_{h}\left(y_{0}\right)+C\left(y_{0}\right) h^{2 k+1}+\mathcal{O}\left(h^{2 k+2}\right)$, where $\varphi_{t}\left(y_{0}\right)$ denotes the exact flow of the problem. Consequently, it holds

$$
\Phi_{b_{1} h} \circ \Phi_{b_{2} h} \circ \Phi_{b_{3} h}\left(y_{0}\right)=\varphi_{\left(b_{1}+b_{2}+b_{3}\right) h}\left(y_{0}\right)+\left(b_{1}^{2 k+1}+b_{2}^{2 k+1}+b_{3}^{2 k+1}\right) C\left(y_{0}\right) h^{2 k+1}+\mathcal{O}\left(h^{2 k+2}\right)
$$

The assumption (6.1) thus implies order at least $2 k+1$ for the composed method $\Psi_{h}$. Due to $b_{3}=b_{1}$ and the symmetry of $\Phi_{h}$, the method $\Psi_{h}$ is also symmetric, implying that the order of $\Psi_{h}$ is at least $2 k+2$ (Theorem 3.2).

Table 4.3: Coefficients of the 3-stage Lobatto IIIA - IIIB pair

| 0 | 0 | 0 | 0 | 0 | $1 / 6$ | $-1 / 6$ | 0 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $1 / 2$ | $5 / 24$ | $1 / 3$ | $-1 / 24$ | $1 / 2$ | $1 / 6$ | $1 / 3$ | 0 |
| 1 | $1 / 6$ | $2 / 3$ | $1 / 6$ | 1 | $1 / 6$ | $5 / 6$ | 0 |
|  | $1 / 6$ | $2 / 3$ | $1 / 6$ |  | $1 / 6$ | $2 / 3$ | $1 / 6$ |

The above theorem allows us to construct symmetric one-step methods of arbitrarily high order. We take a symmetric method $\Phi_{h}^{(2)}$ of order 2, e.g., the implicit midpoint rule (I.1.6) or the Störmer-Verlet method (I.3.6) or something else. With the choice

$$
b_{1}=\left(2-2^{1 / 3}\right)^{-1}, \quad b_{0}=1-2 b_{1},
$$

the method $\Phi_{h}^{(4)}:=\Phi_{b_{1} h}^{(2)} \circ \Phi_{b_{0} h}^{(2)} \circ \Phi_{b_{1} h}^{(2)}$ is symmetric and of order 4 (see Theorem 6.1). Observe that $b_{1} \approx 1.3512072$ and $b_{0} \approx-1.7024144$. This means that the method takes two positive step sizes $b_{1} h$ and one negative step size $b_{0} h$. We can now repeat this procedure. With $c_{1}=\left(2-2^{1 / 5}\right)^{-1}$ and $c_{0}=1-2 c_{1}$, the method $\Phi_{h}^{(6)}:=\Phi_{c_{1} h}^{(4)} \circ \Phi_{c_{0} h}^{(4)} \circ \Phi_{c_{1} h}^{(4)}$ is symmetric and of order 6 . For every step it requires 9 applications of the basic method $\Phi_{h}^{(2)}$. Continuing this procedure, we can construct symmetric methods of order $p=2 k$ which require $3^{k-1}$ applications of $\Phi_{h}^{(2)}$.

If we take as basic method the Störmer-Verlet method and if the differential equation is of the special type (4.3), then we obtain by this construction explicit symmetric methods of arbitrarily high order. The implementation of these methods is extremely simple. One only has to write a subroutine for the basic method of low order and one calls it several times with different step sizes.

Optimal Composition Methods The methods constructed above are of the form

$$
\begin{equation*}
\Psi_{h}=\Phi_{b_{m} h} \circ \ldots \circ \Phi_{b_{1} h} \circ \Phi_{b_{0} h} \circ \Phi_{b_{1} h} \circ \ldots \circ \Phi_{b_{m} h}, \tag{6.2}
\end{equation*}
$$

with $m=1$ for the 4 th order method, $m=4$ for the 6 th order method, and $m=13$ for the 8 th order method. It is natural to ask for optimal methods in the sense that a minimal number of $\Phi_{b_{i} h}$ (i.e., minimal $m$ ) is required for a given order.

There are several ways of finding the order conditions. One approach is that adopted in the proof of Theorem 6.1. In this way one sees that for a second order method $\Phi_{h}$ the composition $\Psi_{h}$ of (6.2) is of order at least 4 if

$$
\begin{align*}
& b_{0}+2\left(b_{1}+\ldots+b_{m}\right)=1 \\
& b_{0}^{3}+2\left(b_{1}^{3}+\ldots+b_{m}^{3}\right)=0 . \tag{6.3}
\end{align*}
$$

An extension of this approach to higher order is rather tedious. It is surprising that for order 6 only two additional order conditions

$$
\begin{align*}
& b_{0}^{5}+2\left(b_{1}^{5}+\ldots+b_{m}^{5}\right)=0 \\
& \sum_{j=1}^{m}\left(B_{j-1} b_{j}^{4}+B_{j-1}^{2} b_{j}^{3}-C_{j-1} b_{j}^{2}-B_{j-1} C_{j-1} b_{j}\right)=0, \tag{6.4}
\end{align*}
$$

where $B_{j}=b_{0}+2\left(b_{1}+\ldots+b_{j}\right)$ and $C_{j}=b_{0}^{3}+2\left(b_{1}^{3}+\ldots+b_{j}^{3}\right)$, have to be satisfied (compare with the large number of conditions for Runge-Kutta methods). We do not give details of the derivation of (6.4), because later in Chapter IV we shall see a very elegant approach to these order conditions with the help of backward error analysis and with the use of the Campbell-Baker-Hausdorff formula. It is interesting to note that the order conditions for (6.2) do not depend on the basic method $\Phi_{h}$.

Example 6.2 For a method (6.2) of order 6 the coefficients $b_{i}$ have to satisfy the four conditions (6.3) and (6.4). Yoshida [Yo90] solves these equations numerically with $m=3$. He finds three solutions, one of which is

$$
\begin{aligned}
& b_{1}=-1.17767998417887 \\
& b_{2}=0.235573213359357 \\
& b_{3}=0.784513610477560
\end{aligned}
$$

and $b_{0}=1-2\left(b_{1}+b_{2}+b_{3}\right)$.
A method of order 8 with $m=7$ is also constructed in [Yo90].

## II. 7 Linear Multistep Methods

In 1984 a graduate student, who took my ODE class at Stanford, wanted to work with symmetric multistep methods. I discouraged him, $\ldots{ }^{4}$ (G. Dahlquist in a letter to the author, 22. Feb. 1998)
symmetric, with a motivation, partitioned, underlying one-step method, Kirchgraber

## II. 8 Exercises

1. Compute all collocation methods with $s=2$ in dependence of $c_{1}$ and $c_{2}$. Which of them are of order 3 , which of order 4?
2. Prove that the collocation solution plotted in the right picture of Fig. 1.1 is composed of arcs of parabolas.
3. Find all trees of orders 5 and 6.
4. (A. Cayley [Ca1857]) Denote the number of trees of order $q$ by $a_{q}$. Prove that

| $a_{1}+a_{2} x+a_{3} x^{2}+a_{4} x^{3}+\ldots=(1-x)^{-a_{1}}\left(1-x^{2}\right)^{-a_{2}}\left(1-x^{3}\right)^{-a_{3}} \ldots$ |  |  |  |  |  |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $q$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| $a_{q}$ | 1 | 1 | 2 | 4 | 9 | 20 | 48 | 115 | 286 | 719 |

5. Prove that the coefficient $\alpha(t)$ of Definition 2.2 is equal to the number of possible monotonic labellings of the vertices of $t$, starting with the label 1 for the root. For example, the tree $[[\tau], \tau]$ has three different mono-
 tonic labellings.
6. Show that for every $t \in T$ there is a system (2.1) such that the first component of $F(t)(0)$ equals 1 , and the first component of $F(u)(0)$ is zero for all trees $u \neq t$.
Hint. Consider a monotonic labelling of $t$, and define $y_{i}^{\prime}$ as the product over all $y_{j}$, where $j$ runs through all labels of vertices that lie directly above the vertex ' $i$ '. For the first labelling of the tree of Exercise 5 this would be $y_{1}^{\prime}=y_{2} y_{3}, y_{2}^{\prime}=1, y_{3}^{\prime}=y_{4}$, and $y_{4}^{\prime}=1$.
7. Compute the adjoint of the symplectic Euler method (I.1.8).

[^5]8. Prove that the Störmer-Verlet method (I.3.6) is symmetric.

Hint. If $\Phi_{h}^{*}$ is the adjoint method of $\Phi_{h}$, then the compositions $\Phi_{h / 2} \circ \Phi_{h / 2}^{*}$ and $\Phi_{h / 2}^{*} \circ \Phi_{h / 2}$ are symmetric one-step methods.
9. Explicit Runge-Kutta methods cannot be symmetric.

Hint. If a one-step method applied to $y^{\prime}=\lambda y$ yields $y_{1}=R(h \lambda) y_{0}$ then, a necessary condition for the symmetry of the method is $R(z) R(-z)=1$ for all complex $z$.
10. A collocation method is symmetric if and only if (after a suitable permutation of the $c_{i}$ ) $c_{i}+c_{s+1-i}=1$ holds for all $i$, i.e., the collocation points are symmetrically distributed on the interval $[0,1]$.
11. Consider a one-step method $\Phi_{h}$ of order 2. Is it possible to construct a composition $\operatorname{method} \Psi_{h}=\Phi_{b_{1} h} \circ \ldots \circ \Phi_{b_{m} h}$ of order at least 3 under the restriction that all $b_{i}$ are positive?
12. The number of order conditions for partitioned Runge-Kutta methods (4.2) is $2 a_{r}$ for order $r$, where $a_{r}$ is given by (see [HNW93, page 311])

| $r$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $a_{r}$ | 1 | 2 | 7 | 26 | 107 | 458 | 2058 | 9498 | 44987 | 216598 |

Find a formula similar to that of Exercise 4.
13. Let $\left\{b_{i}, a_{i j}\right\}$ and $\left\{\widetilde{b}_{i}, \widetilde{a}_{i j}\right\}$ be the coefficients of two Runge-Kutta methods $\Phi_{h}$ and $\widetilde{\Phi}_{h}$, respectively. Prove that the composition $\Phi_{\alpha h} \circ \widetilde{\Phi}_{(1-\alpha) h}$ is again a Runge-Kutta method. What are its coefficients?
14. If $\Phi_{h}$ stands for the implicit midpoint rule, what are the Runge-Kutta coefficients of the composition method (6.2)? The general theory of Sect. II. 2 gives three order conditions for order 4 (those for the trees of order 2 and 4 are automatically satisfied by the symmetry of the method). Are they compatible with the two conditions (6.3)?

## Chapter III

## Exact Conservation of Invariants

This chapter is devoted to the conservation of invariants (first integrals) by numerical methods. Our investigation will follow two directions. We first study which of the methods introduced in Chapter II conserve first integrals automatically. We shall see that most of them conserve linear invariants, a few of them quadratic invariants, and none of them conserves cubic or general nonlinear invariants. We then construct new classes of methods, which are adapted to known invariants and which force the numerical solution to satisfy them. In particular, we study projection methods and so-called Lie group methods which are based on the Magnus expansion of the solution of non-autonomous linear systems.

## III. 1 Examples of First Integrals

Consider a differential equation

$$
\begin{equation*}
y^{\prime}=f(y), \quad y\left(t_{0}\right)=y_{0} \tag{1.1}
\end{equation*}
$$

where $y$ is a vector or eventually a matrix.
Definition 1.1 A non-constant function $I(y)$ is called a first integral of (1.1) if

$$
\begin{equation*}
I^{\prime}(y) f(y)=0 \quad \text { for all } y \tag{1.2}
\end{equation*}
$$

This implies that every solution $y(t)$ of (1.1) satisfies $I(y(t))=I\left(y_{0}\right)=$ Const.
In Chapter I we have seen many examples of differential equations with first integrals. E.g., The Volterra-Lotka problem (I.1.1) has $I(u, v)=\ln u-u+2 \ln v-v$ as first integral. The pendulum equation (I.1.9) has $H(p, q)=p^{2} / 2-\cos q$, and the Kepler problem (I.2.1) has even two first integrals, namely $H$ and $L$ of (I.2.3) and (I.2.4), respectively.

Example 1.2 (Conservation of the Total Energy) Every Hamiltonian system ${ }^{1}$

$$
p^{\prime}=-H_{q}^{T}(p, q), \quad q^{\prime}=H_{p}^{T}(p, q)
$$

has the Hamiltonian function $H(p, q)$ as first integral. This follows at once from $H^{\prime}(p, q)=$ $\left(H_{p}, H_{q}\right)$ and $H_{p}\left(-H_{q}\right)^{T}+H_{q} H_{p}^{T}=0$.

[^6]Example 1.3 (Conservation of Mass in Chemical Reactions) Suppose that three substances $A, B, C$ undergo a chemical reaction such as ${ }^{2}$

$$
\begin{array}{cccl}
A & \xrightarrow{0.04} & B & \text { (slow) } \\
B+B & \xrightarrow{3 \cdot 10^{7}} & C+B & \text { (very fast) } \\
B+C & \xrightarrow{10^{4}} & A+C & \text { (fast). }
\end{array}
$$

We denote the masses (or concentrations) of the substances $A, B, C$ by $y_{1}, y_{2}, y_{3}$, respectively. By the mass action law this leads to the equations

$$
\begin{array}{lll}
\mathrm{A}: & y_{1}^{\prime}=-0.04 y_{1}+10^{4} y_{2} y_{3} & y_{1}(0)=1 \\
\mathrm{~B}: & y_{2}^{\prime}=0.04 y_{1}-10^{4} y_{2} y_{3}-3 \cdot 10^{7} y_{2}^{2} & y_{2}(0)=0 \\
\mathrm{C}: & y_{3}^{\prime}=3 \cdot 10^{7} y_{2}^{2} & y_{3}(0)=0 .
\end{array}
$$

One can check that the total mass $I(y)=y_{1}+y_{2}+y_{3}$ is a first integral of the system.
Theorem 1.4 (Conservation of Linear First Integrals [Sh86]) All explicit and implicit Runge-Kutta methods as well as multistep methods conserve linear first integrals.

Partitioned Runge-Kutta methods (II.4.2) conserve linear first integrals only if $b_{i}=\widehat{b}_{i}$ for all $i$, or if the first integral only depends on $p$ alone or only on $q$ alone.

Proof. Let $I(y)=d^{T} y$ with a constant vector $d$. The assumption (1.2) implies $d^{T} f(y)=0$ for all $y$. In the case of Runge-Kutta methods we thus have $d^{T} k_{i}=0$, and consequently $d^{T} y_{1}=d^{T} y_{0}+h d^{T}\left(\sum_{i=1}^{s} b_{i} k_{i}\right)=d^{T} y_{0}$. The statement for multistep methods and partitioned methods is proved similarly.

Next we consider differential equations of the form

$$
\begin{equation*}
Y^{\prime}=B(Y) Y, \quad Y\left(t_{0}\right)=Y_{0} \tag{1.3}
\end{equation*}
$$

where $Y$ can be a vector, but usually it will be a square matrix. We then have the following result.

Theorem 1.5 If $B(Y)$ is skew-symmetric for all $Y$ (i.e., $B^{T}=-B$ ), then the quadratic function $I(Y)=Y^{T} Y$ is a first integral. In particular, if the initial value is an orthogonal matrix (i.e., $Y_{0}^{T} Y_{0}=I$ ), then the solution $Y(t)$ of (1.3) remains orthogonal for all $t$.

Proof. The derivative of $I(Y)$ is $I^{\prime}(Y) H=Y^{T} H+H^{T} Y$. Since $B(Y)$ is a skew-symmetric matrix, we have $I^{\prime}(Y) f(Y)=I^{\prime}(Y)(B(Y) Y)=Y^{T} B(Y) Y+Y^{T} B(Y)^{T} Y=0$ for all $Y$. This proves the statement.

Example 1.6 (Rigid Body Simulation) Consider a rigid body whose center of mass is fixed at the origin. Its movement is described by the Euler equations

$$
\begin{array}{ll}
y_{1}^{\prime}=a_{1} y_{2} y_{3}, & a_{1}=\left(I_{2}-I_{3}\right) /\left(I_{2} I_{3}\right) \\
y_{2}^{\prime}=a_{2} y_{3} y_{1}, & a_{2}=\left(I_{3}-I_{1}\right) /\left(I_{3} I_{1}\right)  \tag{1.4}\\
y_{3}^{\prime}=a_{3} y_{1} y_{2}, & a_{3}=\left(I_{1}-I_{2}\right) /\left(I_{1} I_{2}\right)
\end{array}
$$

where the vector $y=\left(y_{1}, y_{2}, y_{3}\right)^{T}$ represents the angular momentum in the body frame,

[^7]and $I_{1}, I_{2}, I_{3}$ are the principal moments of inertia (see [MaR94, Chap.15] for a detailed description). This problem can be written as
\[

\left($$
\begin{array}{l}
y_{1}^{\prime}  \tag{1.5}\\
y_{2}^{\prime} \\
y_{3}^{\prime}
\end{array}
$$\right)=\left($$
\begin{array}{ccc}
0 & I_{3}^{-1} y_{3} & -I_{2}^{-1} y_{2} \\
-I_{3}^{-1} y_{3} & 0 & I_{1}^{-1} y_{1} \\
I_{2}^{-1} y_{2} & -I_{1}^{-1} y_{1} & 0
\end{array}
$$\right)\left($$
\begin{array}{l}
y_{1} \\
y_{2} \\
y_{3}
\end{array}
$$\right)
\]

which is of the form (1.3) with a skew-symmetric matrix $B(Y)$. By Theorem $1.5, y_{1}^{2}+$ $y_{2}^{2}+y_{3}^{2}$ is a first integral. It is interesting to note that the system (1.4) is actually a Hamiltonian system on the sphere $\left\{y_{1}^{2}+y_{2}^{2}+y_{3}^{2}=1\right\}$ with Hamiltonian

$$
H\left(y_{1}, y_{2}, y_{3}\right)=\frac{1}{2}\left(\frac{y_{1}^{2}}{I_{1}}+\frac{y_{2}^{2}}{I_{2}}+\frac{y_{3}^{2}}{I_{3}}\right) .
$$

This is a second quadratic invariant of the system (1.4).


Fig. 1.1: Solutions of the Euler equations (1.4) for the rigid body

Inspired by the cover page of [MaR94], we present in Fig. 1.1 the sphere with some of the solutions of (1.4) corresponding to $I_{1}=2, I_{2}=1$ and $I_{3}=2 / 3$. In the left picture we have included the numerical solution ( 30 steps) obtained by the implicit midpoint rule with step size $h=0.3$ and initial value $y_{0}=(\cos (1.1), 0, \sin (1.1))^{T}$. It stays exactly on a solution curve. This follows from the fact that the implicit midpoint rule preserves quadratic invariants exactly (Sect. III.3).

For the explicit Euler method (right picture, 320 steps with $h=0.05$ and the same initial value as above) we see that the numerical solution shows the wrong qualitative behaviour (it should lie on a closed curve). The numerical solution even drifts away from the manifold.

## III. 2 Differential Equations on Lie Groups

Theorem 1.5 is a particular case of a more general result, which can be conveniently formulated with the concept of Lie groups and Lie algebras (see [Olv86] and [Var74] for an introduction to this subject).

Definition 2.1 A Lie group is a group $G$ which is a differentiable manifold, and for which the product is a differentiable mapping $G \times G \rightarrow G$.

If $I$ is the unit element of $G$, then the tangent space $T_{I} G$ is called the Lie algebra of $G$ and it is denoted by $\mathfrak{g}$.

We restrict our considerations to matrix Lie groups (Table 2.1). The matrix $J$, appearing in the definition of the symplectic group, is the matrix determining the symplectic structure on $R^{n}$ (see Chapter IV).

Table 2.1: Some matrix Lie groups and their corresponding Lie algebras

| Lie group | Lie algebra |
| :---: | :---: |
| $\begin{aligned} \mathrm{GL}(n)= & \{A \mid \operatorname{det} A \neq 0\} \\ & \text { general linear group } \end{aligned}$ | $\begin{aligned} \mathfrak{g l}(n)= & \{B \mid \text { arbitrary matrix }\} \\ & \text { Lie algebra of } n \times n \text { matrices } \end{aligned}$ |
| $\begin{aligned} \mathrm{SL}(n)= & \{A \mid \operatorname{det} A=1\} \\ & \text { special linear group } \end{aligned}$ | $\begin{aligned} \hline \mathfrak{s l}(n)= & \{B \mid \operatorname{trace}(B)=0\} \\ & \text { special linear Lie algebra } \end{aligned}$ |
| $\mathrm{O}(n)=\begin{aligned} & \left\{A \mid A^{T} A=I\right\} \\ & \\ & \text { orthogonal group } \end{aligned}$ | $\begin{aligned} \mathfrak{s o}(n)= & \begin{array}{l} \left\{B \mid B^{T}+B=0\right\} \\ \\ \text { skew-symmetric matrices } \end{array} \end{aligned}$ |
| $\mathrm{SO}(n)=\begin{gathered} \{A \in \mathrm{O}(n) \mid \operatorname{det} A=1\} \\ \text { special orthogonal group } \end{gathered}$ | $\begin{aligned} \hline \mathfrak{s o}(n)= & \begin{array}{l} \left\{B \mid B^{T}+B=0\right\} \\ \\ \text { skew-symmetric matrices } \end{array} \end{aligned}$ |
| $\begin{aligned} \mathrm{Sp}(n)= & \left\{A \mid A^{T} J A=J\right\} \\ & \text { symplectic group } \end{aligned}$ | $\mathfrak{s p}(n)=\left\{B \mid J B+B^{T} J=0\right\}$ |

Example 2.2 An interesting example of a Lie group is the set

$$
\mathrm{O}(n)=\left\{A \in \mathrm{GL}(n) \mid A^{T} A=I\right\}
$$

of all orthogonal matrices. It is the kernel of $g(A)=A^{T} A-I$, where we consider $g$ as a mapping from the set of all $n \times n$ matrices (i.e., $\mathbb{R}^{n \cdot n}$ ) to the set of all symmetric matrices (which can be identified with $\mathbb{R}^{n(n+1) / 2}$ ). The derivative $g^{\prime}(A)$ is surjective for $A \in \mathrm{O}(n)$, because for any symmetric matrix $K$ the choice $H=A K / 2$ solves the equation $g^{\prime}(A) H=K$. Therefore, $\mathrm{O}(n)$ defines a differentiable manifold of dimension $n^{2}-n(n+1) / 2=n(n-1) / 2$. The set $\mathrm{O}(n)$ is also a group with unit element $I$ (the identity). Since the matrix multiplication is a differentiable mapping, $\mathrm{O}(n)$ is a Lie group.

In order to compute its Lie algebra, we use the fact that for manifolds, which are defined as the kernel of a mapping $g$, the tangent space is given by $T_{I} \mathrm{O}(n)=\operatorname{Ker} g^{\prime}(I)$. Since $g^{\prime}(I) H=I^{T} H+H^{T} I=H+H^{T}$, it consists of all skew-symmetric matrices.

Lemma 2.3 (Exponential Map) Consider a Lie group $G$ and its Lie algebra $\mathfrak{g}$. The exponential function

$$
\exp (B)=\sum_{k \geq 0} \frac{1}{k!} B^{k}
$$

is a map $\exp : \mathfrak{g} \rightarrow G$, i.e., for $B \in \mathfrak{g}$ we have $\exp (B) \in G$.

Proof. For $B \in \mathfrak{g}$, it follows from the definition of the tangent space $\mathfrak{g}=T_{I} G$ that there exists a differentiable mapping $\gamma:(-\varepsilon, \varepsilon) \rightarrow G$ satisfying $\gamma(0)=I$ and $\gamma^{\prime}(0)=B$. For a fixed $Y \in G$, the function $\delta:(-\varepsilon, \varepsilon) \rightarrow G$, defined by $\delta(t):=\gamma(t) Y$, satisfies $\delta(0)=Y$ and $\delta^{\prime}(0)=B Y$. Consequently, $B Y \in T_{Y} G$ and $Y^{\prime}=B Y$ defines a differential equation on the manifold $G$. The solution $Y(t)=\exp (t B)$ is therefore in $G$.

The next lemma motivates the name "Lie algebra" for the tangent space $T_{I} G$.

Lemma 2.4 (Lie Bracket) Let $G$ be a Lie group and $\mathfrak{g}$ its Lie algebra. The Lie bracket (or commutator)

$$
\begin{equation*}
[A, B]=A B-B A \tag{2.1}
\end{equation*}
$$

defines an operation $\mathfrak{g} \times \mathfrak{g} \rightarrow \mathfrak{g}$ which is bilinear, skew-symmetric $([A, B]=-[B, A])$, and satisfies the Jacobi identity

$$
\begin{equation*}
[A,[B, C]]+[C,[A, B]]+[B,[C, A]]=0 . \tag{2.2}
\end{equation*}
$$

Proof. For $A, B \in \mathfrak{g}$, we consider the mapping $\gamma:(-\varepsilon, \varepsilon) \rightarrow G$, defined by

$$
\gamma(t)=\exp (\sqrt{t} A) \exp (\sqrt{t} B) \exp (-\sqrt{t} A) \exp (-\sqrt{t} B)
$$

Using $\exp (\sqrt{t} A)=I+\sqrt{t} A+\frac{1}{2} t A^{2}+\mathcal{O}\left(t^{3 / 2}\right)$, an elementary computation yields $\gamma(t)=I+$ $t[A, B]+\mathcal{O}\left(t^{3 / 2}\right)$. This is a differentiable mapping satisfying $\gamma(0)=I$ and $\gamma^{\prime}(0)=[A, B]$. Hence $[A, B] \in \mathfrak{g}$ by definition of the tangent space. The properties of the Lie bracket can be verified straight-forwardly.

Theorem 2.5 Let $G$ be a Lie group and $\mathfrak{g}$ its Lie algebra. If $B(Y) \in \mathfrak{g}$ for all $Y \in G$ and if $Y_{0} \in G$, then the solution of (1.3) satisfies $Y(t) \in G$ for all $t$.

If in addition $B(Y) \in \mathfrak{g}$ for all matrices $Y$, and if $G=\{Y \mid g(Y)=$ Const $\}$ is one of the Lie groups of Table 2.1, then $g(Y)$ is a first integral of the differential equation (1.3).

Proof. As in the proof of Lemma 2.3 we see that $B(Y) Y \in T_{Y} G$, so that (1.3) is a differential equation on the manifold $G$.

The second statement has already been proved in Theorem 1.5 for $G=\mathrm{O}(n)$. Let us prove it for $\operatorname{SL}(n)$. For $B \in \mathfrak{g}$ we let $\gamma(t)$ as in the proof of Lemma 2.3, and we put $\delta(t)=\gamma(t) Y$. In general, $\delta(t)$ will not be in $G$, but it holds $g(\delta(t))=\operatorname{det}(\delta(t))=$ $\operatorname{det}(\gamma(t)) \operatorname{det} Y=$ Const $\cdot \operatorname{det} Y$. Differentiation with respect to $t$ gives $g^{\prime}(Y)(B Y)=0$ for all $Y$, which means that $g(Y)$ is a first integral of (1.3).

## III. 3 Quadratic Invariants

Quadratic first integrals appear often in applications. Examples are: the conservation law of angular momentum ( $L(p, q)$ in Kepler's problem (I.2.1) and $L(p, q)=\sum_{i=1}^{N} q_{i} \times p_{i}$ in the Hamiltonian systems (I.2.10) and (I.3.1)), the two invariants in the rigid body simulation (Example 1.6), and the first integrals $Y^{T} Y-I$ and $Y^{T} J Y-J$ of Theorem 2.5. We therefore consider differential equations (1.1) and quadratic functions

$$
\begin{equation*}
Q(y)=y^{T} C y \tag{3.1}
\end{equation*}
$$

where $C$ is a symmetric square matrix.
Theorem 3.1 The Gauss methods of Example II.1.7 (collocation based on the shifted Legendre polynomials) conserve quadratic first integrals.

Proof. Let $u(t)$ be the collocation polynomial (Definition II.1.3). Since $\frac{d}{d t} Q(u(t))=$ $2 u(t)^{T} C u^{\prime}(t)$, it follows from $u\left(t_{0}\right)=y_{0}$ and $u\left(t_{0}+h\right)=y_{1}$ that

$$
\begin{equation*}
y_{1}^{T} C y_{1}-y_{0}^{T} C y_{0}=2 \int_{t_{0}}^{t_{0}+h} u(t)^{T} C u^{\prime}(t) d t \tag{3.2}
\end{equation*}
$$

The integrand $u(t)^{T} C u^{\prime}(t)$ is a polynomial of degree $2 s-1$, which is integrated without error by the $s$-stage Gaussian quadrature formula. It therefore follows from $C u^{\prime}\left(t_{0}+c_{i} h\right)=$ $C f\left(u\left(t_{0}+c_{i} h\right)\right)=0$ that the integral in (3.2) vanishes.

Since the implicit midpoint rule is the special case $s=1$ of the Gauss methods, the preceding theorem explains its good behaviour for the rigid body simulation in Fig 1.1.
Theorem 3.2 ([Co87]) If the coefficients of a Runge-Kutta method satisfy

$$
\begin{equation*}
b_{i} a_{i j}+b_{j} a_{j i}=b_{i} b_{j} \quad \text { for all } i, j=1, \ldots, s, \tag{3.3}
\end{equation*}
$$

then it conserves quadratic first integrals. ${ }^{3}$
Proof. The proof is the same as that for B-stability, given independently by Burrage \& Butcher and Crouzeix in 1979 (see [HW96, Sect.IV.12]).

The relation $y_{1}=y_{0}+h \sum_{i=1}^{s} b_{i} k_{i}$ of Definition II.1.1 yields

$$
\begin{equation*}
y_{1}^{T} C y_{1}=y_{0}^{T} C y_{0}+h \sum_{i=1}^{s} b_{i} k_{i}^{T} C y_{0}+h \sum_{j=1}^{s} b_{j} y_{0}^{T} C k_{j}+h^{2} \sum_{i, j=1}^{s} b_{i} b_{j} k_{i}^{T} C k_{j} . \tag{3.4}
\end{equation*}
$$

We then write $k_{i}=f\left(Y_{i}\right)$ with $Y_{i}=y_{0}+h \sum_{j=1}^{s} a_{i j} k_{j}$. The main idea is now to compute $y_{0}$ from this relation and to insert it into the central expressions of (3.4). This yields (using the symmetry of $C$ )

$$
y_{1}^{T} C y_{1}=y_{0}^{T} C y_{0}+2 h \sum_{i=1}^{s} b_{i} Y_{i}^{T} C f\left(Y_{i}\right)+h^{2} \sum_{i, j=1}^{s}\left(b_{i} b_{j}-b_{i} a_{i j}-b_{j} a_{j i}\right) k_{i}^{T} C k_{j} .
$$

The condition (3.3) together with the assumption $y^{T} C f(y)=0$, which states that $y^{T} C y$ is a first integral of (1.1), imply $y_{1}^{T} C y_{1}=y_{0}^{T} C y_{0}$.

[^8]We next consider partitioned Runge-Kutta methods for systems $p^{\prime}=f(p, q), q^{\prime}=$ $g(p, q)$. Usually such methods will not conserve general quadratic invariants (Exercise 6). We therefore concentrate on quadratic first integrals of the form

$$
\begin{equation*}
Q(p, q)=p^{T} D q \tag{3.5}
\end{equation*}
$$

where $D$ is an arbitrary matrix. Observe that the angular momentum in the Hamiltonian systems (I.2.10) and (I.3.1) is of this form.
Theorem 3.3 If the coefficients of a partitioned Runge-Kutta method (II.4.2) satisfy

$$
\begin{align*}
& b_{i} \widehat{a}_{i j}+\widehat{b}_{j} a_{j i}=b_{i} \widehat{b}_{j} \quad \text { for } \quad i, j=1, \ldots, s,  \tag{3.6}\\
& b_{i}=\widehat{b}_{i} \quad \text { for } i=1, \ldots, s \tag{3.7}
\end{align*}
$$

then it conserves quadratic first integrals of the form (3.5).
If the partitioned differential equation is of the special form $p^{\prime}=f(q), q^{\prime}=g(p)$, then the condition (3.6) alone implies that first integrals of the form (3.5) are conserved.

Proof. The proof is nearly identical to that of Theorem 3.2. Instead of (3.4) we get

$$
p_{1}^{T} D q_{1}=p_{0}^{T} D q_{0}+h \sum_{i=1}^{s} b_{i} k_{i}^{T} D q_{0}+h \sum_{j=1}^{s} \widehat{b}_{j} p_{0}^{T} D \ell_{j}+h^{2} \sum_{i, j=1}^{s} b_{i} \widehat{b}_{j} k_{i}^{T} D \ell_{j}
$$

Denoting by $\left(P_{i}, Q_{i}\right)$ the arguments of $k_{i}=f\left(P_{i}, Q_{i}\right)$ and $\ell_{i}=g\left(P_{i}, Q_{i}\right)$, the same trick as in the above proof gives

$$
\begin{align*}
p_{1}^{T} D q_{1}=p_{0}^{T} D q_{0} & +h \sum_{i=1}^{s} b_{i} f\left(P_{i}, Q_{i}\right)^{T} D Q_{i}+h \sum_{j=1}^{s} \widehat{b}_{j} P_{j}^{T} D g\left(P_{j}, Q_{j}\right)  \tag{3.8}\\
& +h^{2} \sum_{i, j=1}^{s}\left(b_{i} \widehat{b}_{j}-b_{i} \widehat{a}_{i j}-\widehat{b}_{j} a_{j i}\right) k_{i}^{T} D \ell_{j}
\end{align*}
$$

Since (3.5) is a first integral, we have $f(p, q)^{T} D q+p^{T} D g(p, q)=0$ for all $p$ and $q$. Consequently, the two conditions (3.6) and (3.7) imply $p_{1}^{T} D q_{1}=p_{0}^{T} D q_{0}$.

For the special case, where $f$ only depends on $q$ and $g$ only on $p$, the assumption $f(q)^{T} D q+p^{T} D g(p)=0$ (for all $p, q$ ) implies that $f(q)^{T} D q=-p^{T} D g(p)=$ Const. Therefore, the condition (3.7) is no longer necessary for the proof of the statement.

Example 3.4 ([Sun93]) The pair Lobatto IIIA - IIIB of Example II.4. 6 conserves quadratic first integrals of the form (3.5). For the proof of this statement we have to check the conditions (3.6) (observe that (3.7) is satisfied by the definition). We let $\left\{b_{i}, a_{i j}\right\}$ be the coefficients of Lobatto IIIA and $\left\{\widehat{b}_{i}, \widehat{a}_{i j}\right\}$ be those of Lobatto IIIB. We first prove that

$$
D(s): \quad \sum_{i=1}^{s} b_{i} c_{i}^{k-1} \widehat{a}_{i j}=\frac{b_{j}}{k}\left(1-c_{j}^{k}\right) \quad \text { for } k=1, \ldots, s \text { and all } j .
$$

For this we put $d_{j}:=\sum_{i=1}^{s} b_{i} c_{i}^{k-1} \widehat{a}_{i j}-b_{j}\left(1-c_{j}^{k}\right) / k$. From $\widehat{a}_{i s}=0$ and $c_{s}=1$ we have $d_{s}=0$, and from $\widehat{a}_{i 1}=b_{1}$ and $c_{1}=0$ we get $d_{1}=0$. Furthermore, it follows from $C(s-2)$ that for $l=1, \ldots, s-2$

$$
\sum_{j=2}^{s-1} d_{j} c_{j}^{l-1}=\sum_{i=1}^{s} b_{i} c_{i}^{k-1} \frac{c_{i}^{l}}{l}-\sum_{j=1}^{s} \frac{b_{j}}{k}\left(1-c_{j}^{k}\right) c_{j}^{l-1}=\frac{1}{l(k+l)}-\frac{1}{k l}+\frac{1}{k(k+l)}=0
$$

This Vandermonde type system implies $d_{j}=0$ for all $j$, and therefore also $D(s)$.

In order to prove (3.6) we put $e_{i}=b_{i} \widehat{a}_{i j}+\widehat{b}_{j} a_{j i}-b_{i} \widehat{b}_{j}$ for a fixed $j \in\{1, \ldots, s\}$. It follows from the conditions $C(s)$ for Lobatto IIIA and $D(s)$ for Lobatto IIIB that

$$
\sum_{i=1}^{s} e_{i} c_{i}^{k-1}=\frac{b_{j}}{k}\left(1-c_{j}^{k}\right)+b_{j} \frac{c_{j}^{k}}{k}-b_{j} \frac{1}{k}=0
$$

for $k=1, \ldots, s$. This implies $e_{i}=0$ for all $i$ and hence also (3.6).
Example 3.5 (Composition Methods) If a method $\Phi_{h}$ conserves quadratic first integrals, then so does the composition method

$$
\begin{equation*}
\Psi_{h}=\Phi_{b_{s} h} \circ \ldots \circ \Phi_{b_{1} h} \tag{3.9}
\end{equation*}
$$

This property is one of the most important motivations for considering composition methods. Since the $b_{i}$ need not be symmetric in (3.9), this example shows the existence of non-symmetric methods that conserve general quadratic integrals.

## III. 4 Polynomial Invariants

We next consider first integrals that are neither linear nor quadratic. An interesting example is $g(Y)=\operatorname{det} Y$ for problems $Y^{\prime}=B(Y) Y$ with trace $B(Y)=0$ (see Theorem 2.5 applied to $\operatorname{SL}(n)$ and $\mathfrak{s l}(n))$. Since $\operatorname{det} Y$ represents the volume of the parallelepiped generated by the columns of the matrix $Y$, the conservation of the first integral $g(Y)=\operatorname{det} Y$ is related to volume preservation. This topic will be further discussed in Chapter IV.

Lemma 4.1 ([FSh95]) Let $R(z)$ be a real analytic function defined in a neighbourhood of $z=0$, and assume that $R(0)=1$ and $R^{\prime}(0)=1$. Then, it holds

$$
R(\mathfrak{s l}(n)) \subset \mathrm{SL}(n) \quad \text { for some } n \geq 3
$$

if and only if $R(z)=\exp (z)$.
Proof. The "if" part follows from Theorem 2.5, because for constant $B$ the solution of $Y^{\prime}=B Y$ is given by $Y(t)=\exp (B t)$.

For the proof of the "only if" part, we consider diagonal matrices $B \in \mathfrak{s l}(n)$ of the form $B=\operatorname{diag}(\mu, \nu,-(\mu+\nu), 0, \ldots, 0)$ for which

$$
R(B)=\operatorname{diag}(R(\mu), R(\nu), R(-(\mu+\nu)), R(0), \ldots, R(0))
$$

The assumptions $R(0)=1$ and $R(\mathfrak{s l}(n)) \subset \mathrm{SL}(n)$ imply

$$
\begin{equation*}
R(\mu) R(\nu) R(-(\mu+\nu))=1 \tag{4.1}
\end{equation*}
$$

for all $\mu, \nu$ close to 0 . Putting $\nu=0$, this relation yields $R(\mu) R(-\mu)=1$ for all $\mu$, and therefore (4.1) can be written as

$$
\begin{equation*}
R(\mu) R(\nu)=R(\mu+\nu) \quad \text { for all } \mu, \nu \text { close to } 0 \tag{4.2}
\end{equation*}
$$

This functional equation can only be satisfied by the exponential function. This is seen as follows: from (4.2) we have

$$
\frac{R(\mu+\varepsilon)-R(\mu)}{\varepsilon}=R(\mu) \frac{R(\varepsilon)-R(0)}{\varepsilon} .
$$

Taking the limit $\varepsilon \rightarrow 0$ we obtain $R^{\prime}(\mu)=R(\mu)$, because $R^{\prime}(0)=1$. This implies $R(\mu)=\exp (\mu)$.

Theorem 4.2 For $n \geq 3$, no Runge-Kutta method can conserve all polynomial first integrals of degree $n$.

Proof. We consider linear problems $Y^{\prime}=B Y$ with constant $B \in \mathfrak{s l}(n)$, for which $g(Y)=\operatorname{det} Y$ is a polynomial invariant of degree $n$. Applying a Runge-Kutta method to such a differential equation yields $Y_{1}=R(h B) Y_{0}$, where

$$
R(z)=1+z b^{T}(I-z A)^{-1} \mathbb{1}
$$

$\left(b^{T}=\left(b_{1}, \ldots, b_{s}\right), \mathbb{1}=(1, \ldots, 1)^{T}\right.$ and $A=\left(a_{i j}\right)$ is the matrix of Runge-Kutta coefficients) is the so-called stability function. It is seen to be rational. By Lemma 4.1 it is therefore not possible that $\operatorname{det} R(h B)=1$ for all $B \in \mathfrak{s l}(n)$.

This negative result motivates the search for new methods which can conserve polynomial first integrals (see Sects. III.5, III. 7 and IV.7). We consider here another interesting class of problems with polynomial first integrals of order higher than two.

Isospectral Flows The general form of an isospectral flow is the differential equation

$$
\begin{equation*}
L^{\prime}=[B(L), L], \quad L(0)=L_{0} \tag{4.3}
\end{equation*}
$$

where $L_{0}$ is a given symmetric matrix, $B(L)$ is skew-symmetric and $[B, L]=B L-L B$ is the commutator of $B$ and $L$. An interesting example is the Toda flow, for which $B(L)=L_{+}-L_{+}^{T}$. Here, $L_{+}$denotes the strictly upper part of the matrix $L$. Further examples with a long list of references are given in [CIZ97].

Lemma 4.3 ([Fla74]) Let $L_{0}$ be symmetric and assume that $B(L)$ is skew-symmetric for all L. Then, the eigenvalues of the solution $L(t)$ of (4.3) are independent of $t$.

Proof. The idea is to consider the system

$$
\begin{equation*}
U^{\prime}=B\left(U L_{0} U^{T}\right) U, \quad U(0)=I \tag{4.4}
\end{equation*}
$$

It is straight-forward to check that for a solution $U(t)$ of (4.4) the matrix

$$
\begin{equation*}
L(t)=U(t) L_{0} U(t)^{T} \tag{4.5}
\end{equation*}
$$

is a solution of (4.3). Since $B\left(U L_{0} U^{T}\right)$ is skew-symmetric, $U^{T} U$ is a first integral of (4.4). Hence, $U(t)$ is orthogonal and (4.5) is a similarity transformation implying that $L(t)$ and $L_{0}$ have the same eigenvalues.

The above proof shows that the characteristic polynomial $\operatorname{det}(L-\lambda I)=\sum_{i=0}^{n} a_{i} \lambda^{i}$ and hence also the coefficients $a_{i}$ are independent of $t$. These coefficients are all polynomial invariants (e.g., $a_{0}=\operatorname{det} L, a_{n-1}=\operatorname{trace} L$ ). Because of Theorem 4.2 there is no hope that Runge-Kutta methods applied to (4.3) can conserve these invariants. The proof of Lemma 4.3, however, suggests an interesting approach for the numerical solution of (4.3).

For $n=0,1, \ldots$ we solve numerically

$$
\begin{equation*}
U^{\prime}=B\left(U L_{n} U^{T}\right) U, \quad U(0)=I \tag{4.6}
\end{equation*}
$$

and we put $L_{n+1}=U_{1} L_{n} U_{1}^{T}$, where $U_{1}$ is the numerical approximation $U_{1} \approx U(h)$. If $B(L)$ is skew-symmetric for all matrices $L$, then $U^{T} U$ is a quadratic first integral of (4.6) and the methods of Sect. III. 3 will produce an orthogonal $U_{1}$. Consequently, $L_{n+1}$ and $L_{n}$ have exactly the same eigenvalues.

## III. 5 Projection Methods

We consider general nonlinear differential equations

$$
\begin{equation*}
y^{\prime}=f(y), \quad y(0)=y_{0} \tag{5.1}
\end{equation*}
$$

and we assume throughout this section that one or several (say $m$ ) invariants are explicitly known. We write them as

$$
\begin{equation*}
g(y)=0, \tag{5.2}
\end{equation*}
$$

where $g: \mathbb{R}^{n} \rightarrow \mathbb{R}^{m}$. The individual invariants $g_{i}(y)$ need not be first integrals of (5.1). It is sufficient to know that the exact solution of (5.1) satisfies $g_{i}(y(t))=0$ for $t \geq 0$. This means that (5.1) defines a differential equation on the manifold

$$
\begin{equation*}
\mathcal{M}=\{y ; g(y)=0\} \tag{5.3}
\end{equation*}
$$

A very natural approach for the numerical solution of such problems is by projection (see e.g., [HW96, Sect. VII.2], [EF98, Sect.5.3.3]). The idea is to take an arbitrary one-step method $\Phi_{h}$ and to perform in every step the following two operations:

- compute $\widetilde{y}_{1}=\Phi_{h}\left(y_{0}\right)$,
- project the value $\widetilde{y}_{1}$ onto the manifold $\mathcal{M}$ to obtain $y_{1} \in \mathcal{M}$.

For $y_{0} \in \mathcal{M}$ the distance of $\widetilde{y}_{1}$ to the manifold $\mathcal{M}$ is of the size of the local error, i.e., $\mathcal{O}\left(h^{p+1}\right)$. We therefore also have $y_{1}-y(h)=\mathcal{O}\left(h^{p+1}\right)$, so that the projection does not deteriorate the convergence order of the method.

For the computation of $y_{1}$ we have to solve the constrained minimization problem

$$
\begin{equation*}
\left\|y_{1}-\widetilde{y}_{1}\right\| \rightarrow \min \quad \text { subject to } \quad g\left(y_{1}\right)=0 . \tag{5.4}
\end{equation*}
$$

In the case of the Euclidean norm, a standard approach is to introduce Lagrange multipliers $\lambda=\left(\lambda_{1}, \ldots, \lambda_{m}\right)^{T}$, and to consider the Lagrange function $\mathcal{L}\left(y_{1}, \lambda\right)=\left\|y_{1}-\widetilde{y}_{1}\right\|^{2} / 2-$ $g\left(y_{1}\right)^{T} \lambda$. The necessary condition $\partial \mathcal{L} / \partial y_{1}=0$ then leads to the system

$$
\begin{align*}
y_{1} & =\widetilde{y}_{1}+g^{\prime}\left(\widetilde{y}_{1}\right)^{T} \lambda \\
0 & =g\left(y_{1}\right) . \tag{5.5}
\end{align*}
$$

We have replaced $y_{1}$ by $\widetilde{y}_{1}$ in the argument of $g^{\prime}(y)$ in order to save some evaluations of $g^{\prime}(y)$. Inserting the first relation of (5.5) into the second, one gets a nonlinear equation for $\lambda$, which can be efficiently solved by simplified Newton iterations:

$$
\Delta \lambda_{i}=-\left(g^{\prime}\left(\widetilde{y}_{1}\right) g^{\prime}\left(\widetilde{y}_{1}\right)^{T}\right)^{-1} g\left(\widetilde{y}_{1}+g^{\prime}\left(\widetilde{y}_{1}\right)^{T} \lambda_{i}\right), \quad \lambda_{i+1}=\lambda_{i}+\Delta \lambda_{i} .
$$

For the choice $\lambda_{0}=0$ the first increment $\Delta \lambda_{0}$ is of size $\mathcal{O}\left(h^{p+1}\right)$, so that the convergence is usually extremely fast. Often, one simplified Newton iteration is sufficient.
explicit Euler, $h=0.03$

symplectic Euler, $h=0.03$




Fig. 5.1: Numerical solutions obtained without and with projections

Example 5.1 As a first example we consider the perturbed Kepler problem (see Exercise I.6) with Hamiltonian function

$$
H(p, q)=\frac{1}{2}\left(p_{1}^{2}+p_{2}^{2}\right)-\frac{1}{\sqrt{q_{1}^{2}+q_{2}^{2}}}-\frac{0.005}{2 \sqrt{\left(q_{1}^{2}+q_{2}^{2}\right)^{3}}}
$$

and initial values $q_{1}(0)=1-e, q_{2}(0)=0, p_{1}(0)=0$, $p_{2}(0)=\sqrt{(1+e) /(1-e)}($ eccentricity $e=0.6)$ on the interval $0 \leq t \leq 200$. The exact solution (plotted to the right) is approximately an ellipse that turns
 slowly around one of its foci. For this problem we know two first integrals: the Hamiltonian function $H(p, q)$ and the angular momentum $L(p, q)=q_{1} p_{2}-q_{2} p_{1}$.

We apply the explicit Euler method and the symplectic Euler method (I.1.8), both with constant step size $h=0.03$. The result is shown in Fig. 5.1. The numerical solution of the explicit Euler method (without projection) is completely wrong. The projection onto the manifold $\left\{H(p, q)=H\left(p_{0}, q_{0}\right)\right\}$ improves the numerical solution, but it has still a wrong qualitative behaviour. Only projection onto both invariants, $H(p, q)=$ Const and $L(p, q)=$ Const gives the correct behaviour. The symplectic Euler method shows already the correct behaviour without any projections (see Chapter V for an explanation).


Fig. 5.2: Explicit Euler method with projections applied to the outer solar system

Surprisingly, a projection onto $H(p, q)=$ Const destroys this behaviour, the numerical solution approaches the center and the simplified Newton iterations fail to converge beyond $t=25.23$. Projection onto both invariants re-establishes the correct behaviour.

Example 5.2 (Outer Solar System) Having encountered excellent experience with projections onto $H$ and $L$ for the perturbed Kepler problem (Example 5.1), let us apply the same idea to a more realistic problem in celestial mechanics. We consider the outer solar system as described in Sect.I.2. The numerical solution of the explicit Euler method, applied with constant step size $h=10$, once with projection onto $H=$ Const and once with projection onto $H=$ Const and $L=$ Const , is shown in Fig. 5.2 (observe that the conservation of the angular momentum $L(p, q)=\sum_{i=1}^{N} q_{i} \times p_{i}$ consists of three first integrals). We see a slight improvement in the orbits of Jupiter, Saturn and Uranus (compared to the explicit Euler method without projections, see Fig. I.2.3), but the orbit of Neptune becomes even worse. There is no doubt that this problem contains a structure which cannot be correctly simulated by methods that only preserve the total energy $H$ and the angular momentum $L$.

Example 5.3 (Special Linear Group) For the problem $Y^{\prime}=B(Y) Y$ with $B(Y) \in$ $\mathfrak{s l}(n)$ we know that $g(Y)=\operatorname{det} Y-\operatorname{det} Y_{0}$ is a first integral. Let $\tilde{Y}_{1}$ be the numerical approximation obtained with an arbitrary one-step method. If we consider the Frobenius norm $\|Y\|_{F}=\sqrt{\sum_{i, j}\left|y_{i j}\right|^{2}}$ for measuring the distance to the manifold $\{Y ; g(Y)=0\}$, then by using Cramer's rule the projection step (5.5) becomes (Exercise 11)

$$
\begin{equation*}
Y_{1}=\widetilde{Y}_{1}+\mu \widetilde{Y}_{1}^{-T} \tag{5.6}
\end{equation*}
$$

with the scalar $\mu=\lambda \operatorname{det} \tilde{Y}_{1}$. This leads to the nonlinear equation $\operatorname{det}\left(\widetilde{Y}_{1}+\mu \widetilde{Y}_{1}^{-T}\right)=\operatorname{det} Y_{0}$, for which simplified Newton iterations become

$$
\operatorname{det}\left(\tilde{Y}_{1}+\mu_{i} \tilde{Y}_{1}^{-T}\right)\left(1+\left(\mu_{i+1}-\mu_{i}\right) \operatorname{trace}\left(\left(\tilde{Y}_{1}^{T} \tilde{Y}_{1}\right)^{-1}\right)\right)=\operatorname{det} Y_{0}
$$

If the $Q R$-decomposition of $\tilde{Y}_{1}$ is available, the computation of trace $\left(\left(\tilde{Y}_{1}^{T} \tilde{Y}_{1}\right)^{-1}\right)$ can be done efficiently with $\mathcal{O}\left(n^{3} / 3\right)$ flops [GVL89, Sect. 5.3.9].

The above described projection is preferable over $Y_{1}=c \widetilde{Y}_{1}$, where $c \in \mathbb{R}$ is chosen such that $\operatorname{det} Y_{1}=\operatorname{det} Y_{0}$. This latter projection is ill-conditioned already for diagonal matrices with entries that differ in several magnitudes.

Example 5.4 (Orthogonal Matrices) As a final example let us consider $Y^{\prime}=F(Y)$, where the solution $Y(t)$ is known to be an orthogonal matrix. The projection step (5.4) requires the solution of the problem

$$
\begin{equation*}
\|Y-\tilde{Y}\|_{F} \rightarrow \min \quad \text { subject to } \quad Y^{T} Y=I \tag{5.7}
\end{equation*}
$$

where $\tilde{Y}$ is a given matrix, close to an orthogonal one. This projection can be computed as follows: compute the singular value decomposition $\tilde{Y}=U^{T} \Sigma V$, where $U$ and $V$ are orthogonal, $\Sigma=\operatorname{diag}\left(\sigma_{1}, \ldots, \sigma_{n}\right)$, and the singular values $\sigma_{1} \geq \ldots \geq \sigma_{n}$ are all close to 1 . Then the solution of (5.7) is given by $Y=U^{T} V$. See Exercise 12 for some hints.

A related procedure, where the $Q R$ decomposition of $\tilde{Y}$ is used instead of the singular value decomposition, is proposed in [DRV94].

As conclusion of these numerical experiments we see that a projection gives excellent results, if all invariants determining the longtime behaviour of the solution are known. If the original method already preserves some structure, then projection to a subset of invariants may destroy the good longtime behaviour.

## III. 6 Magnus Series Expansion



Wilhelm Magnus ${ }^{4}$

In this section we give an explicit formula for the solution of linear differential equations

$$
\begin{equation*}
Y^{\prime}=A(t) Y, \quad Y(0)=I \tag{6.1}
\end{equation*}
$$

This will be the basic ingredient for the Lie group methods discussed in Sect. III.7. For the scalar case, the solution of (6.1) is given by

$$
\begin{equation*}
Y(t)=\exp \left(\int_{0}^{t} A(\tau) d \tau\right) \tag{6.2}
\end{equation*}
$$

Also in the case, where the matrices $A(t)$ and $\int_{0}^{t} A(\tau) d \tau$ commute, (6.2) is the solution of (6.1). In the general non-commutative case we follow the approach of [Mag54] and we search for a matrix function $\Omega(t)$ such that

$$
Y(t)=\exp (\Omega(t))
$$

solves (6.1). This requires the use of matrix commutators (2.1) and of the adjoint operator (see [Var74, Sect. 2.13])

$$
\begin{equation*}
\operatorname{ad}_{\Omega}(A)=[\Omega, A] \tag{6.3}
\end{equation*}
$$

Let us start by computing the derivatives of $\Omega^{k}$. It is known that

$$
\begin{equation*}
\left(\frac{d}{d \Omega} \Omega^{k}\right) H=H \Omega^{k-1}+\Omega H \Omega^{k-2}+\ldots+\Omega^{k-1} H \tag{6.4}
\end{equation*}
$$

[^9]and this equals $k H \Omega^{k-1}$ only if $\Omega$ and $H$ commute. Therefore, it is natural to write (6.4) as $k H \Omega^{k-1}$ plus terms involving commutators and iterated commutators. In the cases $k=2$ and $k=3$ we have
\[

$$
\begin{aligned}
H \Omega+\Omega H & =2 H \Omega+\operatorname{ad}_{\Omega}(H) \\
H \Omega^{2}+\Omega H \Omega+\Omega^{2} H & =3 H \Omega^{2}+3\left(\operatorname{ad}_{\Omega}(H)\right) \Omega+\operatorname{ad}_{\Omega}^{2}(H)
\end{aligned}
$$
\]

where $\operatorname{ad}_{\Omega}^{i}$ denotes the iterated application of the linear operator $\mathrm{ad}_{\Omega}$. With the convention $\operatorname{ad}_{\Omega}^{0}(H)=H$ we obtain by induction on $k$ that

$$
\begin{equation*}
\left(\frac{d}{d \Omega} \Omega^{k}\right) H=\sum_{i=0}^{k-1}\binom{k}{i+1}\left(\operatorname{ad}_{\Omega}^{i}(H)\right) \Omega^{k-i-1} \tag{6.5}
\end{equation*}
$$

This is seen by applying Leibniz's rule to $\Omega^{k+1}=\Omega \cdot \Omega^{k}$ and by using $\Omega\left(\operatorname{ad}_{\Omega}^{i}(H)\right)=$ $\left(\operatorname{ad}_{\Omega}^{i}(H)\right) \Omega+\operatorname{ad}_{\Omega}^{i+1}(H)$.

Lemma 6.1 The derivative of $\exp \Omega=\sum_{k \geq 0} \frac{1}{k!} \Omega^{k}$ is given by

$$
\left(\frac{d}{d \Omega} \exp \Omega\right) H=\left(d \exp _{\Omega}(H)\right) \exp \Omega
$$

where

$$
\begin{equation*}
d \exp _{\Omega}(H)=\sum_{k \geq 0} \frac{1}{(k+1)!} \operatorname{ad}_{\Omega}^{k}(H) \tag{6.6}
\end{equation*}
$$

The series (6.6) converges for all matrices $\Omega$.
Proof. Multiplying (6.5) by ( $k$ ! $)^{-1}$ and summing up, then exchanging the sums and putting $j=k-i-1$ yields

$$
\begin{aligned}
\left(\frac{d}{d \Omega} \exp \Omega\right) H & =\sum_{k \geq 0} \frac{1}{k!} \sum_{i=0}^{k-1}\binom{k}{i+1}\left(\operatorname{ad}_{\Omega}^{i}(H)\right) \Omega^{k-i-1} \\
& =\sum_{i \geq 0} \sum_{j \geq 0} \frac{1}{(i+1)!j!}\left(\operatorname{ad}_{\Omega}^{i}(H)\right) \Omega^{j}
\end{aligned}
$$

The convergence of the series follows from the boundedness of the linear operator ad ${ }_{\Omega}$ (we have $\left\|\operatorname{ad}_{\Omega}\right\| \leq 2\|\Omega\|$ ).

Lemma 6.2 ([Ba05]) If the eigenvalues of the linear operator $\mathrm{ad}_{\Omega}$ are different from $2 \ell \pi i$ with $\ell \in\{ \pm 1, \pm 2, \ldots\}$, then $\exp _{\Omega}$ is invertible. Furthermore, we have for $\|\Omega\|<\pi$ that

$$
\begin{equation*}
d \exp _{\Omega}^{-1}(H)=\sum_{k \geq 0} \frac{B_{k}}{k!} \operatorname{ad}_{\Omega}^{k}(H) \tag{6.7}
\end{equation*}
$$

where $B_{k}$ are the Bernoulli numbers, defined by $\sum_{k \geq 0}\left(B_{k} / k!\right) x^{k}=x /\left(e^{x}-1\right)$.
Proof. The eigenvalues of $d \exp _{\Omega}$ are $\mu=\sum_{k>0} \lambda^{k} /(k+1)!=\left(e^{\lambda}-1\right) / \lambda$, where $\lambda$ is an eigenvalue of $\operatorname{ad}_{\Omega}$. By our assumption, the values $\mu$ are non-zero, so that $d \exp _{\Omega}$ is invertible. By definition of the Bernoulli numbers the composition of (6.7) with (6.6) gives the identity. Convergence for $\|\Omega\|<\pi$ follows from $\left\|\operatorname{ad}_{\Omega}\right\| \leq 2\|\Omega\|$ and from the fact that the convergence radius of the series for $x /\left(e^{x}-1\right)$ is $2 \pi$.

Theorem 6.3 ([Mag54]) The solution of (6.1) can be written as $Y(t)=\exp (\Omega(t))$ with $\Omega(t)$ defined by

$$
\begin{equation*}
\Omega^{\prime}=d \exp _{\Omega}^{-1}(A(t)), \quad \Omega(0)=0 \tag{6.8}
\end{equation*}
$$

As long as $\|\Omega(t)\|<\pi$, the convergence of the $d \exp _{\Omega}^{-1}$ expansion (6.7) is assured.
Proof. Comparing the equation

$$
Y^{\prime}(t)=\left(\frac{d}{d \Omega} \exp \Omega(t)\right) \Omega^{\prime}(t)=\left(d \exp _{\Omega(t)}\left(\Omega^{\prime}(t)\right)\right) \exp (\Omega(t))
$$

with (6.1) we obtain $A(t)=d \exp _{\Omega(t)}\left(\Omega^{\prime}(t)\right)$. Applying the inverse operator $d \exp _{\Omega}^{-1}$ to this relation yields the differential equation (6.8).

The first Bernoulli numbers are $B_{0}=1, B_{1}=-1 / 2, B_{2}=1 / 6, B_{3}=0$. The differential equation (6.8) therefore becomes

$$
\Omega^{\prime}=A(t)-\frac{1}{2}[\Omega, A(t)]+\frac{1}{12}[\Omega,[\Omega, A(t)]]+\ldots,
$$

which is nonlinear in $\Omega$. Applying Picard fixed point iteration after integration yields

$$
\begin{align*}
\Omega(t)=\int_{0}^{t} A(\tau) d \tau & -\frac{1}{2} \int_{0}^{t}\left[\int_{0}^{\tau} A(\sigma) d \sigma, A(\tau)\right] d \tau \\
& +\frac{1}{4} \int_{0}^{t}\left[\int_{0}^{\tau}\left[\int_{0}^{\sigma} A(\mu) d \mu, A(\sigma)\right] d \sigma, A(\tau)\right] d \tau  \tag{6.9}\\
& +\frac{1}{12} \int_{0}^{t}\left[\int_{0}^{\tau} A(\sigma) d \sigma,\left[\int_{0}^{\tau} A(\mu) d \mu, A(\tau)\right]\right] d \tau+\ldots,
\end{align*}
$$

which is the so-called Magnus expansion. For smooth matrices $A(t)$ the remainder in (6.9) is of size $\mathcal{O}\left(t^{5}\right)$ so that the truncated series inserted into $Y(t)=\exp (\Omega(t))$ gives for small $t$ an excellent approximation to the solution of (6.1).

Numerical Methods Based on the Magnus Expansion Such methods have been recently proposed by [IN97] and [Za97]. Consider the problem

$$
\begin{equation*}
y^{\prime}=A(t) y, \quad y\left(t_{0}\right)=y_{0} \tag{6.10}
\end{equation*}
$$

where $y$ may be a matrix or a vector. The idea is to replace $A(t)$ locally by an interpolation polynomial

$$
\widehat{A}(t)=\sum_{i=1}^{s} \ell_{i}(t) A\left(t_{n}+c_{i} h\right)
$$

and to solve $y^{\prime}=\widehat{A}(t) y$ on $\left[t_{n}, t_{n}+h\right]$ by the use of the truncated series (6.9).
Theorem 6.4 Consider a quadrature formula $\left(b_{i}, c_{i}\right)_{i=1}^{s}$ of order $p \geq s$, and let $y(t)$ and $z(t)$ be solutions of $y^{\prime}=A(t) y$ and $z^{\prime}=\widehat{A}(t) z$, respectively, satisfying $y\left(t_{n}\right)=z\left(t_{n}\right)$. Then, it holds $z\left(t_{n}+h\right)-y\left(t_{n}+h\right)=\mathcal{O}\left(h^{p+1}\right)$.
Proof. We write the differential equation for $z$ as $z^{\prime}=A(t) z+(\widehat{A}(t)-A(t)) z$ and use the variation of constants formula to get

$$
z\left(t_{n}+h\right)-y\left(t_{n}+h\right)=\int_{t_{n}}^{t_{n}+h} R\left(t_{n}+h, \tau\right)(\widehat{A}(\tau)-A(\tau)) z(\tau) d \tau
$$

Applying our quadrature formula to this integral gives zero as result, and the remainder is of size $\mathcal{O}\left(h^{p+1}\right)$. Details of the proof are as for Theorem 1.5.

Example 6.5 As a first example, we use the midpoint rule $\left(c_{1}=1 / 2, b_{1}=1\right)$. In this case the interpolation polynomial is constant, and the method becomes

$$
\begin{equation*}
y_{n+1}=\exp \left(h A\left(t_{n}+h / 2\right)\right) y_{n}, \tag{6.11}
\end{equation*}
$$

which is of order 2 .
Example 6.6 The two-stage Gauss quadrature is given by $c_{1,2}=1 / 2 \pm \sqrt{3} / 6, b_{1,2}=1 / 2$. The interpolation polynomial is of degree one and we have to apply (6.9) in order to get an approximation $y_{n+1}$. Since we are interested in a fourth order approximation, we can neglect the remainder term (indicated by ... in (6.9)). Computing carefully the iterated integrals over products of $\ell_{i}(t)$ we obtain

$$
\begin{equation*}
y_{n+1}=\exp \left(\frac{h}{2}\left(A_{1}+A_{2}\right)+\frac{\sqrt{3} h^{2}}{12}\left[A_{2}, A_{1}\right]\right) y_{n} \tag{6.12}
\end{equation*}
$$

where $A_{1}=A\left(t_{n}+c_{1} h\right)$ and $A_{2}=A\left(t_{n}+c_{2} h\right)$. This is a method of order four. The terms of (6.9) with triple integrals give $\mathcal{O}\left(h^{4}\right)$ expressions, whose leading term vanishes by the symmetry of the method (Exercise 15). Therefore, it need not be considered.

## III. 7 Lie Group Methods



Marius Sophus Lie ${ }^{5}$

The motivation of Lie group methods is similar to that of projection methods, discussed in Sect. III.5: if the solution of a differential equation is known to lie in a Lie group, one is interested in numerical approximations that stay in the same Lie group. Some early work on this subject is in the papers [Is84] and [CG93]. We present here the approach of [MK98] for the case of matrix Lie groups.

Consider the differential equation

$$
\begin{equation*}
Y^{\prime}=A(Y) Y, \quad Y(0)=Y_{0}, \tag{7.1}
\end{equation*}
$$

where $A(Y)$ is assumed to be in a Lie algebra, so that the solution $Y(t)$ stays in the corresponding Lie group. The crucial idea of Lie group methods is to write the solution as $Y(t)=\exp (\Omega(t)) Y_{0}$ and to solve numerically the differential equation for $\Omega(t)$. It sounds awkward to replace the differential equation (7.1) by a more complicated one. However, the nonlinear invariants $g(Y)=0$ of (7.1) defining the Lie group are replaced by linear invariants $g^{\prime}(I)(\Omega)=0$ defining the Lie algebra, and we know from Sect.III. 1 that essentially all numerical methods automatically conserve linear first integrals.

It follows from the proof of Theorem 6.3 that the solution of (7.1) can be written as $Y(t)=\exp (\Omega(t)) Y_{0}$, where $\Omega(t)$ is the solution of $\Omega^{\prime}=d \exp _{\Omega}^{-1}(A(Y(t))), \Omega(0)=0$. Since

[^10]it is not practical to work with the operator $d \exp _{\Omega}^{-1}$ we truncate suitably the series (6.7) and we consider the differential equation
\[

$$
\begin{equation*}
\Omega^{\prime}=A\left(\exp (\Omega) Y_{0}\right)+\sum_{k=1}^{q} \frac{B_{k}}{k!} \operatorname{ad}_{\Omega}^{k}\left(A\left(\exp (\Omega) Y_{0}\right)\right), \quad \Omega(0)=0 \tag{7.2}
\end{equation*}
$$

\]

The step $Y_{n} \mapsto Y_{n+1}$ of a Lie group method, as defined by [MK98], is given as follows:

- consider the differential equation (7.2) with $Y_{0}$ replaced by $Y_{n}$, and apply a RungeKutta method (explicit or implicit) in order to get an approximation $\Omega_{1} \approx \Omega(h)$,
- then compute $Y_{n+1}=\exp \left(\Omega_{1}\right) Y_{n}$.

Important properties of this method are given in the next two theorems.
Theorem 7.1 Let $G$ be a matrix Lie group and $\mathfrak{g}$ its Lie algebra. If $A(Y) \in \mathfrak{g}$ for $Y \in G$ and if $Y_{0} \in G$, then the numerical solution of the above Lie group method lies in $G$, i.e., $Y_{n} \in G$ for all $n=0,1,2, \ldots$.

Proof. It is sufficient to prove that for $Y_{0} \in G$ the numerical solution $\Omega_{1}$ of the RungeKutta method applied to (7.2) lies in $\mathfrak{g}$. Since the Lie bracket $[\Omega, A]$ is an operation $\mathfrak{g} \times \mathfrak{g} \rightarrow \mathfrak{g}$, and since $\exp (\Omega) Y_{0} \in G$ for $\Omega \in \mathfrak{g}$, the righthand expression of (7.2) is in $\mathfrak{g}$ for $\Omega \in \mathfrak{g}$. Hence, (7.2) is a differential equation on the linear manifold $\mathfrak{g}$ with solution $\Omega(t) \in \mathfrak{g}$. Since $\mathfrak{g}$ is a linear space, all operations in a Runge-Kutta method give results in $\mathfrak{g}$, so that also the numerical approximation $\Omega_{1}$ lies in $\mathfrak{g}$.

Theorem 7.2 If the Runge-Kutta method is of (classical) order $p$ and if the truncation index in (7.2) satisfies $q \geq p-2$, then the corresponding Lie group method is of order $p$.
Proof. For sufficiently smooth $A(Y)$ we have $\Omega(t)=t A\left(Y_{0}\right)+\mathcal{O}\left(t^{2}\right), Y(t)=Y_{0}+\mathcal{O}(t)$ and $[\Omega(t), A(Y(t))]=\mathcal{O}\left(t^{2}\right)$. This implies that $\operatorname{ad}_{\Omega(t)}^{k}(A(Y(t)))=\mathcal{O}\left(t^{k+1}\right)$, so that the truncation of the series in (7.2) induces an error of size $\mathcal{O}\left(h^{q+2}\right)$ for $|t| \leq h$. Hence, for $q+2 \geq p$, this truncation does not affect the order of convergence.

Example 7.3 We consider the Lie group method based on the explicit Euler scheme. Taking $q=0$ in (7.2) this leads to

$$
Y_{1}=\exp \left(h A\left(Y_{0}\right)\right) Y_{0}
$$

We apply this method with step size $h=0.1$ to the system (1.5) which is already of the form (7.1). Observe that $Y_{0}$ is a vector in $\mathbb{R}^{3}$ and not a matrix, but all results of this section remain valid for this case. For the computation of the matrix exponential we use the formula of Rodrigues (Exercise 14). The numerical result is shown in the picture to the right. We see that the numerical solution stays on the manifold (sphere), but on the sphere the qualitative behaviour is not correct. A similar behaviour can be observed for the projection method, also based


Euler Lie method

$$
00
$$

initial values satisfying $g\left(y_{0}\right)=0$, the solution of the differential-algebraic equation (DAE)

$$
\begin{aligned}
y^{\prime} & =f(y)+g^{\prime}(y)^{T} \mu \\
0 & =g(y)
\end{aligned}
$$

also solves the differential equation (5.1).
Remark Most methods for DAEs (e.g., stiffly accurate Runge-Kutta methods or BDF methods) lead to numerical integrators that preserve exactly the constraints $g(y)=0$. The difference from the projection method of Sect. III. 5 is that here the internal stages also satisfy the constraint.
14. (Rodrigues formula [MaR94, page 261]) Prove that

$$
\exp (\Omega)=I+\frac{\sin \alpha}{\alpha} \Omega+\frac{1}{2}\left(\frac{\sin (\alpha / 2)}{\alpha / 2}\right)^{2} \Omega^{2} \quad \text { for } \quad \Omega=\left(\begin{array}{ccc}
0 & -\omega_{3} & \omega_{2} \\
\omega_{3} & 0 & -\omega_{1} \\
-\omega_{2} & \omega_{1} & 0
\end{array}\right)
$$

where $\alpha=\sqrt{\omega_{1}^{2}+\omega_{2}^{2}+\omega_{3}^{2}}$. This formula allows for an efficient implementation of the Lie group methods in $\mathrm{O}(3)$.
15. For the numerical solution of (6.10) consider the method $y_{n} \mapsto y_{n+1}$ defined by $y_{n+1}=$ $z\left(t_{n}+h\right)$, where $z(t)$ is the solution of $z^{\prime}=\widehat{A}(t) z, z\left(t_{n}\right)=y_{n}$, and $\widehat{A}(t)$ is the interpolation polynomial based on symmetric nodes $c_{1}, \ldots, c_{s}$, i.e., $c_{s+1-i}+c_{i}=1$ for all $i$.
a) Prove that this method is symmetric.
b) It holds $y_{n+1}=\exp (\Omega(h)) y_{n}$, where $\Omega(h)$ has an expansion in odd powers of $h$.

This justifies the omission of the terms involving triple integrals in Example 6.6.
16. ([IN97]) Introducing $y_{1}=y$ and $y_{2}=y^{\prime}$, write the problem

$$
y^{\prime \prime}+t y=0, \quad y(0)=1, \quad y^{\prime}(0)=0
$$

in the form (6.10). Then apply the numerical method of Example 6.6 with different step sizes on the interval $0 \leq t \leq 100$. Compare the result with that obtained by fourth order classical (explicit or implicit) Runge-Kutta methods.
Remark If $A(t)$ in (6.10) (or $A(t, y)$ in (7.1)) are much smoother than the solution $y(t)$, then Lie group methods are usually superior to standard integrators, because Lie group methods approximate $A(t)$, whereas standard methods approximate the solution $y(t)$ by polynomials.
17. Prove that the Lie group method based on the implicit midpoint rule (Example 7.4) is symmetric.
18. Consider the 2 -stage Gauss method of order $p=4$. In the corresponding Lie group method, eliminate the presence of $\Omega$ in $[\Omega, A]$ by iteration, and neglect higher order commutators. Show that this leads to

$$
\begin{aligned}
& \Omega_{1}=h\left(\frac{1}{4} A_{1}+\left(\frac{1}{4}-\frac{\sqrt{3}}{6}\right) A_{2}\right)-\frac{h^{2}}{2}\left(-\frac{1}{12}+\frac{\sqrt{3}}{24}\right)\left[A_{1}, A_{2}\right] \\
& \Omega_{2}=h\left(\left(\frac{1}{4}+\frac{\sqrt{3}}{6}\right) A_{1}+\frac{1}{4} A_{2}\right)-\frac{h^{2}}{2}\left(\frac{1}{12}+\frac{\sqrt{3}}{24}\right)\left[A_{1}, A_{2}\right] \\
& y_{1}=\exp \left(h\left(\frac{1}{2} A_{1}+\frac{1}{2} A_{2}\right)-h^{2} \frac{\sqrt{3}}{12}\left[A_{1}, A_{2}\right]\right) y_{0},
\end{aligned}
$$

where $A_{i}=A\left(Y_{i}\right)$ and $Y_{i}=\exp \left(\Omega_{i}\right) y_{0}$. Prove that this is a Lie group method of order 4. Is it symmetric?
19. In [Za97] a Lie group method similar to that of Exercise 18 is presented. The only difference is that the coefficients $(-1 / 12+\sqrt{3} / 24)$ and $(1 / 12+\sqrt{3} / 24)$ in the formulas for $\Omega_{1}$ and $\Omega_{2}$ are replaced by $(-5 / 72+\sqrt{3} / 24)$ and $(5 / 72+\sqrt{3} / 24)$, respectively. Is there an error somewhere? Are both methods of order 4?

## Chapter IV

## Symplectic Integration

Hamiltonian systems form the most important class of ordinary differential equations in the context of 'Numerical Geometric Integration' (see the examples of Chapter I). In this chapter we start by discussing the origin of such systems and by studying their geometric properties such as symplecticity. We then turn our attention to numerical integrators which preserve the symplectic structure.

## IV. 1 Hamiltonian Systems



Sir William Rowan Hamilton ${ }^{1}$

Consider a mechanical system with $q=\left(q_{1}, \ldots, q_{d}\right)^{T}$ as generalized coordinates, and denote by $T=T(q, \dot{q})=$ $\frac{1}{2} \dot{q}^{T} M(q) \dot{q}$ its kinetic energy $(M(q)$ is assumed to be symmetric and positive definite) and by $U=U(q)$ its potential energy. The movement of such a system is described by the solution of the variational problem

$$
\begin{equation*}
\int L(q(t), \dot{q}(t)) d t \rightarrow \min \tag{1.1}
\end{equation*}
$$

where $L=T-U$ is the Lagrangian of the system. From the fundamental work of Euler (1744) and Lagrange (1755) at the age of 19 (see [HNW93, p. 8] for some historical remarks) we know that the solutions of (1.1) are determined by the second order differential equation

$$
\begin{equation*}
\frac{\partial L}{\partial q}-\frac{d}{d t}\left(\frac{\partial L}{\partial \dot{q}}\right)=0 \tag{1.2}
\end{equation*}
$$

which constitute the so-called Euler-Lagrange equations.
Example 1.1 (Pendulum) We consider the mathematical pendulum (see Sect.I.1) and we take the angle $\alpha$ as generalized coordinate. The kinetic and potential energies are given by $T=m\left(\dot{x}^{2}+\dot{y}^{2}\right) / 2=m \ell^{2} \dot{\alpha}^{2} / 2$ and $U=m g y=-m g \ell \cos \alpha$, respectively, so that the Euler-Lagrange equations become $-m g \ell \sin \alpha-m \ell^{2} \ddot{\alpha}=0$ or equivalently $\ddot{\alpha}+\frac{g}{\ell} \sin \alpha=0$.

[^11]With the aim of simplifying the structure of the Euler-Lagrange equations and of making them more symmetric, Hamilton [Ha1834] had the idea

- of introducing the new variables

$$
\begin{equation*}
p_{k}=\frac{\partial L}{\partial \dot{q}_{k}} \quad \text { for } \quad k=1, \ldots, d, \tag{1.3}
\end{equation*}
$$

the so-called conjugated generalized momenta. Observe that for a fixed $q$ we have $p=M(q) \dot{q}$, so that there is a bijection between $p=\left(p_{1}, \ldots, p_{d}\right)^{T}$ and $\dot{q}$, if $M(q)$ is invertible;

- of considering the Hamiltonian

$$
\begin{equation*}
H:=p^{T} \dot{q}-L(q, \dot{q}) \tag{1.4}
\end{equation*}
$$

as a function of $p$ and $q$, i.e., $H(p, q)$.
Theorem 1.2 Let $M(q)$ and $U(q)$ be continuously differentiable functions. Then, the Euler-Lagrange equations (1.2) are equivalent to the Hamiltonian system

$$
\begin{equation*}
\dot{p}_{k}=-\frac{\partial H}{\partial q_{k}}(p, q), \quad \dot{q}_{k}=\frac{\partial H}{\partial p_{k}}(p, q), \quad k=1, \ldots, d . \tag{1.5}
\end{equation*}
$$

Proof. The definitions (1.3) and (1.4) for the generalized momenta $p$ and for the Hamiltonian function $H$ imply that

$$
\begin{aligned}
& \frac{\partial H}{\partial p}=\dot{q}^{T}+p^{T} \frac{\partial \dot{q}}{\partial p}-\frac{\partial L}{\partial \dot{q}} \frac{\partial \dot{q}}{\partial p}=\dot{q}^{T}, \\
& \frac{\partial H}{\partial q}=p^{T} \frac{\partial \dot{q}}{\partial q}-\frac{\partial L}{\partial q}-\frac{\partial L}{\partial \dot{q}} \frac{\partial \dot{q}}{\partial q}=-\frac{\partial L}{\partial q} .
\end{aligned}
$$

The Euler-Lagrange equations (1.2) are therefore equivalent to (1.5).
If we replace the variable $\dot{q}$ by $M(q)^{-1} p$ in the definition (1.4) of $H(p, q)$, we obtain

$$
\begin{aligned}
H(p, q) & =p^{T} M(q)^{-1} p-L\left(q, M(q)^{-1} p\right)=p^{T} M(q)^{-1} p-\frac{1}{2} p^{T} M(q)^{-1} p+U(q) \\
& =\frac{1}{2} p^{T} M(q)^{-1} p+U(q)
\end{aligned}
$$

and the Hamiltonian is $H=T+U$, which is the total energy of the mechanical system.
In the following we consider Hamiltonian systems (1.5), where the Hamiltonian function $H(p, q)$ is arbitrary (not necessarily related to a mechanical problem).

## IV. 2 Symplectic Transformations

We have already seen in Example 1.2 of Sect. III. 1 that the Hamiltonian function $H(p, q)$ is a first integral of the system (1.5). In this section we shall study another important property of Hamiltonian systems - the symplecticity of its flow.

For two vectors

$$
\xi=\binom{\xi^{p}}{\xi^{q}}, \quad \eta=\binom{\eta^{p}}{\eta^{q}}
$$

in the $(p, q)$ space $\left(\xi^{p}, \xi^{q}, \eta^{p}, \eta^{q}\right.$ are in $\left.\mathbb{R}^{d}\right)$, we consider the parallelogram

$$
L=\{t \xi+s \eta \mid 0 \leq t \leq 1,0 \leq s \leq 1\} .
$$

We then consider its projection $L_{i}=\left\{\left(t \xi_{i}^{p}+s \eta_{i}^{p}, t \xi_{i}^{q}+s \eta_{i}^{q}\right)^{T} \mid 0 \leq t \leq 1,0 \leq s \leq 1\right\}$ onto the ( $p_{i}, q_{i}$ ) coordinate plane, and we let

$$
\left(d p_{i} \wedge d q_{i}\right)(\xi, \eta):=\operatorname{or.area}\left(L_{i}\right)=\operatorname{det}\left(\begin{array}{cc}
\xi_{i}^{p} & \eta_{i}^{p}  \tag{2.1}\\
\xi_{i}^{q} & \eta_{i}^{q}
\end{array}\right)=\xi_{i}^{p} \eta_{i}^{q}-\xi_{i}^{q} \eta_{i}^{p}
$$

be the oriented area of this projection. Here, $d p_{i}$ and $d q_{i}$ select the coordinates of the vectors $\xi$ and $\eta$. In an analogous way, we can also define $d p_{i} \wedge d q_{j}, d p_{i} \wedge d p_{j}$ or $d q_{i} \wedge d q_{j}$. This exterior product is a bilinear map acting on vectors of $\mathbb{R}^{2 d}$. It satisfies Grassmann's rules for exterior multiplication

$$
\begin{equation*}
d p_{i} \wedge d p_{j}=-d p_{j} \wedge d p_{i}, \quad d p_{i} \wedge d p_{i}=0 \tag{2.2}
\end{equation*}
$$

We further consider the differential 2-form

$$
\begin{equation*}
\omega^{2}:=\sum_{i=1}^{d} d p_{i} \wedge d q_{i} \tag{2.3}
\end{equation*}
$$

which will play a central role for Hamiltonian systems. This is again a bilinear mapping. In matrix notation it is given by

$$
\omega^{2}(\xi, \eta)=\xi^{T} J \eta \quad \text { with } \quad J=\left(\begin{array}{cc}
0 & I  \tag{2.4}\\
-I & 0
\end{array}\right)
$$

where $I$ is the identity matrix of dimension $d$.
Definition 2.1 A linear mapping $A: \mathbb{R}^{2 d} \rightarrow \mathbb{R}^{2 d}$ is called symplectic (a name suggested by H. Weyl, 1939), if

$$
\omega^{2}(A \xi, A \eta)=\omega^{2}(\xi, \eta) \quad \text { for all } \xi, \eta \in \mathbb{R}^{2 d}
$$

or, equivalently, if $A^{T} J A=J$.
In the case $d=1$, the expression $\omega^{2}(\xi, \eta)=\left(d p_{1} \wedge d q_{1}\right)(\xi, \eta)$ represents the area of the parallelogram spanned by the 2-dimensional vectors $\xi$ and $\eta$. Symplecticity of a linear mapping $A$ is therefore equivalent to area preservation. In the general case $(d>1)$, symplecticity means that the sum over the oriented areas of the projections $L_{i}$ is the same as that for the transformed parallelograms $A(L)_{i}$.

We now turn our attention to nonlinear mappings. Differentiable functions can locally be approximated by linear mappings. This justifies the following definition.

Definition 2.2 A differentiable function $g: \mathbb{R}^{2 d} \rightarrow \mathbb{R}^{2 d}$ is called symplectic at $(p, q) \in$ $\mathbb{R}^{2 d}$, if the Jacobian matrix $g^{\prime}(p, q)$ is symplectic, i.e., if

$$
\omega^{2}\left(g^{\prime}(p, q) \xi, g^{\prime}(p, q) \eta\right)=\omega^{2}(\xi, \eta) \quad \text { or } \quad g^{\prime}(p, q)^{T} J g^{\prime}(p, q)=J
$$

We next give a geometric interpretation of symplecticity for nonlinear mappings. Consider a 2-dimensional manifold $M$ in the $2 d$-dimensional phase space, and suppose that it is given as the image $M=\varphi(K)$ of a compact set $K \subset \mathbb{R}^{2}$, where $\varphi(s, t)$ is a continuously differentiable function. The manifold $M$ can then be considered as the limit of a union of small parallelograms spanned by the vectors

$$
\frac{\partial \varphi}{\partial s}(s, t) d s \quad \text { and } \quad \frac{\partial \varphi}{\partial t}(s, t) d t
$$

For one such parallelogram we consider (as above) the sum over the oriented areas of its projections onto the ( $p_{i}, q_{i}$ ) plane. We then sum over all parallelograms of the manifold. In the limit this gives the expression

$$
\begin{equation*}
\Omega(M)=\iint_{K} \omega^{2}\left(\frac{\partial \varphi}{\partial s}(s, t), \frac{\partial \varphi}{\partial t}(s, t)\right) d s d t . \tag{2.5}
\end{equation*}
$$

Lemma 2.3 If the mapping $g: \mathbb{R}^{2 d} \rightarrow \mathbb{R}^{2 d}$ is symplectic for all $(p, q) \in \mathbb{R}^{2 d}$, then it preserves the expression $\Omega(M)$, i.e.,

$$
\Omega(g(M))=\Omega(M)
$$

holds for all 2-dimensional manifolds $M$ that can be represented as the image of a continuously differentiable function $\varphi$.
Proof. The manifold $g(M)$ is parametrized by $g \circ \varphi$. The transformation formula for double integrals therefore implies

$$
\Omega(g(M))=\iint_{K} \omega^{2}\left(\frac{\partial(g \circ \varphi)}{\partial s}(s, t), \frac{\partial(g \circ \varphi)}{\partial t}(s, t)\right) d s d t=\Omega(M)
$$

because $(g \circ \varphi)^{\prime}(s, t)=g^{\prime}(\varphi(s, t)) \varphi^{\prime}(s, t)$ and $g$ is a symplectic transformation.
For $d=1, M$ is already a subset of $\mathbb{R}^{2}$ and we can take the identity map for $\varphi$, so that $M=K$. In this case, $\Omega(M)=\iint_{M} d s d t$ represents the area of $M$. Hence, Lemma 2.3 states that (for $d=1$ ) symplectic mappings are area preserving.

We are now able to prove the main result of this section. We use the notation $y=(p, q)$, and we write the Hamiltonian system (1.5) in the form

$$
\begin{equation*}
y^{\prime}=J^{-1} \nabla H(y) \tag{2.6}
\end{equation*}
$$

where $J$ is the matrix of $(2.4)$ and $\nabla H(y)=\operatorname{grad} H(y)^{T}$.
Recall that the flow $\varphi_{t}: \mathbb{R}^{2 d} \rightarrow \mathbb{R}^{2 d}$ of a Hamiltonian system is the mapping that advances the solution by time $t$, i.e., $\varphi_{t}\left(p_{0}, q_{0}\right)=\left(p\left(t, p_{0}, q_{0}\right), q\left(t, p_{0}, q_{0}\right)\right)$, where $p\left(t, p_{0}, q_{0}\right)$, $q\left(t, p_{0}, q_{0}\right)$ is the solution of the system corresponding to initial values $p(0)=p_{0}, q(0)=q_{0}$.
Theorem 2.4 (Poincaré [Po1899]) Let $H(p, q)$ be a twice continuously differentiable function. Then, the flow $\varphi_{t}$ is everywhere a symplectic transformation.
Proof. The derivative $\partial \varphi_{t} / \partial y_{0}$ (with $y_{0}=\left(p_{0}, q_{0}\right)$ ) is a solution of the variational equation which, for the Hamiltonian system (2.6), is given by $\Psi^{\prime}=J^{-1} H^{\prime \prime}\left(\varphi_{t}\left(y_{0}\right)\right) \Psi$, where $H^{\prime \prime}(p, q)$ is the Hessian matrix of $H(p, q)\left(H^{\prime \prime}(p, q)\right.$ is symmetric). We therefore obtain

$$
\begin{aligned}
\frac{d}{d t}\left(\left(\frac{\partial \varphi_{t}}{\partial y_{0}}\right)^{T} J\left(\frac{\partial \varphi_{t}}{\partial y_{0}}\right)\right)=\left(\frac{\partial \varphi_{t}}{\partial y_{0}}\right)^{\prime T} J\left(\frac{\partial \varphi_{t}}{\partial y_{0}}\right)+\left(\frac{\partial \varphi_{t}}{\partial y_{0}}\right)^{T} J\left(\frac{\partial \varphi_{t}}{\partial y_{0}}\right)^{\prime} \\
=\left(\frac{\partial \varphi_{t}}{\partial y_{0}}\right)^{T} H^{\prime \prime}\left(\varphi_{t}\left(y_{0}\right)\right) J^{-T} J\left(\frac{\partial \varphi_{t}}{\partial y_{0}}\right)+\left(\frac{\partial \varphi_{t}}{\partial y_{0}}\right)^{T} H^{\prime \prime}\left(\varphi_{t}\left(y_{0}\right)\right)\left(\frac{\partial \varphi_{t}}{\partial y_{0}}\right)=0
\end{aligned}
$$



Fig. 2.1: Area preservation of the flow of Hamiltonian systems
because $J^{T}=-J$ and $J^{-T} J=-I$. Since the relation

$$
\begin{equation*}
\left(\frac{\partial \varphi_{t}}{\partial y_{0}}\right)^{T} J\left(\frac{\partial \varphi_{t}}{\partial y_{0}}\right)=J \tag{2.7}
\end{equation*}
$$

is satisfied for $t=0$ ( $\varphi_{0}$ is the identity map), it is satisfied for all $t$ and all ( $p_{0}, q_{0}$ ), as long as the solution remains in the domain of definition of $H$.
Example 2.5 Consider the pendulum problem (Example 1.1) with the normalization $m=\ell=g=1$. We then have $q=\alpha, p=\dot{\alpha}$, and the Hamiltonian is given by

$$
H(p, q)=p^{2} / 2-\cos q
$$

Fig. 2.1 shows level curves of this function, and it also illustrates the area preservation of the flow $\varphi_{t}$. Indeed, by Theorem 2.4 and Lemma 2.3 the area of $A$ and $\varphi_{t}(A)$ as well as those of $B$ and $\varphi_{t}(B)$ are the same, although their appearance is completely different.

We next show that symplecticity of a flow is a characteristic property for Hamiltonian systems.
Theorem 2.6 Let $f: \mathbb{R}^{2 d} \rightarrow \mathbb{R}^{2 d}$ be continuously differentiable. Then, $y^{\prime}=f(y)$ is a Hamiltonian system, if and only if its flow $\varphi_{t}(y)$ is symplectic for all $y \in \mathbb{R}^{2 d}$ and for all sufficiently small t.

Proof. The necessity follows from Theorem 2.4. We therefore assume that the flow $\varphi_{t}$ is symplectic, and we have to prove the existence of a function $H(y)$ such that $f(y)=J^{-1} \nabla H(y)$. Differentiating (2.7) and using the fact that $\partial \varphi_{t} / \partial y_{0}$ is solution of the variational equation $\Psi^{\prime}=f^{\prime}\left(\varphi_{t}\left(y_{0}\right)\right) \Psi$, we obtain

$$
\frac{d}{d t}\left(\left(\frac{\partial \varphi_{t}}{\partial y_{0}}\right)^{T} J\left(\frac{\partial \varphi_{t}}{\partial y_{0}}\right)\right)=\left(\frac{\partial \varphi_{t}}{\partial y_{0}}\right)\left(f^{\prime}\left(\varphi_{t}\left(y_{0}\right)\right)^{T} J+J f^{\prime}\left(\varphi_{t}\left(y_{0}\right)\right)\right)\left(\frac{\partial \varphi_{t}}{\partial y_{0}}\right)=0 .
$$

Putting $t=0$, it follows from $J=-J^{T}$ that $J f^{\prime}\left(y_{0}\right)$ is a symmetric matrix for all $y_{0}$. Lemma 2.7 below shows that $J f(y)$ can be written as the gradient of a function $H(y)$.

Lemma 2.7 Let $f: \mathbb{R}^{2 d} \rightarrow \mathbb{R}^{2 d}$ be continuously differentiable, and assume that the Jacobian $f^{\prime}(y)$ is symmetric for all $y$. Then, there exists a function $H: \mathbb{R}^{2 d} \rightarrow \mathbb{R}$ such that $f(y)=\nabla H(y)$, i.e., the vector field $f(y)$ possesses a potential $H(y)$.

Proof. Since $f$ is defined on the whole space, we can define

$$
H(y)=\int_{0}^{1} y^{T} f(t y) d t+\text { Const }
$$

Differentiation with respect to $y_{k}$, and using the symmetry assumption $\partial f_{i} / \partial y_{k}=\partial f_{k} / \partial y_{i}$ yields

$$
\frac{\partial H}{\partial y_{k}}(y)=\int_{0}^{1}\left(f_{k}(t y)+y^{T} \frac{\partial f}{\partial y_{k}}(t y) t\right) d t=\int_{0}^{1} \frac{d}{d t}\left(t f_{k}(t y)\right) d t=f_{k}(y)
$$

which proves the statement.
Lemma 2.7 and Theorem 2.6 remain valid for functions $f: U \rightarrow \mathbb{R}^{2 d}$ with $U \subset \mathbb{R}^{2 d}$, if $U$ is star-shaped or, more generally, if $U$ is a simply connected domain. A counterexample, which shows that the statement of Theorem 2.6 is not true for general $U$, is given in Exercise 8.

## IV. 3 Symplectic Runge-Kutta Methods

Since the property of symplecticity is characteristic of Hamiltonian systems (Theorem 2.6), it is natural to search for numerical methods that share this property. After some pioneering work of de Vogelaere [Vo56], Ruth [Ru83] and Feng Kang [FeK85], the systematic study of symplectic methods started around 1988. A characterization of symplectic Runge-Kutta methods (Theorem 3.4 below) has been found independently by Lasagni [La88], Sanz-Serna [SS88] and Suris [Su89].

Definition 3.1 A numerical one-step method $y_{1}=\Phi_{h}\left(y_{0}\right)$ is called symplectic if, when applied to a smooth Hamiltonian system, the mapping $\Phi_{h}$ is everywhere a symplectic transformation.

Example 3.2 We consider the harmonic oscillator

$$
H(p, q)=\left(p^{2}+q^{2}\right) / 2
$$

so that the Hamiltonian system becomes $\dot{p}=-q, \dot{q}=p$. We apply six different numerical methods to this problem: the explicit Euler method (I.1.4), the symplectic Euler method (I.1.8), and the implicit Euler method (I.1.5), as well as the second order method of Runge

$$
\begin{equation*}
k_{1}=f\left(y_{0}\right), \quad k_{2}=f\left(y_{0}+h k_{1} / 2\right), \quad y_{1}=y_{0}+h k_{2} \tag{3.1}
\end{equation*}
$$

the Verlet scheme (I.3.6), and the implicit midpoint rule (I.1.6). For a set of initial values $\left(p_{0}, q_{0}\right)$ (the dark set in Fig. 3.1) we compute 16 steps with step size $h=\pi / 8$ for the first order methods, and 8 steps with $h=\pi / 4$ for the second order methods. Since the exact solution is periodic with period $2 \pi$, the numerical result of the last step approximates the set of initial values. One clearly observes that the explicit Euler, the implicit Euler and the second order explicit method of Runge are not symplectic (not area preserving).


Fig. 3.1: Area preservation of numerical methods for the harmonic oscillator

The other methods are symplectic (see Theorem 3.4 and Theorem 3.5), although the approximation at the end of the integration may be quite different from the initial set. Only the implicit midpoint rule preserves exactly the quadratic invariant $\left(p^{2}+q^{2}\right) / 2$.

For the study of symplecticity of numerical integrators we follow the approach of [BoS94], which is based on the following lemma.

Lemma 3.3 For Runge-Kutta methods and for partitioned Runge-Kutta methods the following diagram commutes:

$$
\begin{array}{ccc}
y^{\prime}=f(y), \quad y(0)=y_{0} & \longrightarrow & \begin{array}{c}
y^{\prime}=f(y), \quad y(0)=y_{0} \\
\Psi^{\prime}=f^{\prime}(y) \Psi, \quad \Psi(0)=I
\end{array} \\
\begin{array}{l}
\text { method } \\
\\
\left\{y_{n}\right\}
\end{array} & \longrightarrow & \downarrow \text { method }
\end{array}
$$

(horizontal fleches mean 'differentiation'). Therefore, the numerical result $y_{n}, \Psi_{n}$, obtained from applying the method to the problem augmented by its variational equation, is equal to the numerical solution for $y^{\prime}=f(y)$ augmented by its derivative $\Psi_{n}=\partial y_{n} / \partial y_{0}$.

Proof. This result is very important when the derivative of the numerical solution with respect to the initial value is needed. It is proved by implicit differentiation. Let us illustrate this for Euler's method

$$
y_{n+1}=y_{n}+h f\left(y_{n}\right) .
$$

We consider $y_{n}$ and $y_{n+1}$ as functions of $y_{0}$, and we differentiate the equation, defining the numerical method, with respect to $y_{0}$. For Euler's method this gives

$$
\frac{\partial y_{n+1}}{\partial y_{0}}=\frac{\partial y_{n}}{\partial y_{0}}+h f^{\prime}\left(y_{n}\right) \frac{\partial y_{n}}{\partial y_{0}}
$$

which is exactly the same relation that we get, when we apply the method to the variational equation. Since $\partial y_{0} / \partial y_{0}=I$, we have $\partial y_{n} / \partial y_{0}=\Psi_{n}$ for all $n$.

The main observation is now that the symplecticity condition (2.7) is a quadratic first integral of the variational equation. The following characterization of symplectic methods is therefore not surprising.

Theorem 3.4 If the coefficients of a Runge-Kutta method satisfy

$$
\begin{equation*}
b_{i} a_{i j}+b_{j} a_{j i}=b_{i} b_{j} \quad \text { for all } i, j=1, \ldots, s, \tag{3.2}
\end{equation*}
$$

then it is symplectic. ${ }^{2}$
Proof. We write the Hamiltonian system together with its variational equation as

$$
\begin{equation*}
y^{\prime}=J^{-1} \nabla H(y), \quad \Psi^{\prime}=J^{-1} H^{\prime \prime}(y) \Psi . \tag{3.3}
\end{equation*}
$$

It follows from

$$
\left(J^{-1} H^{\prime \prime}(y) \Psi\right)^{T} J \Psi+\Psi^{T} J\left(J^{-1} H^{\prime \prime}(y) \Psi\right)=0
$$

(see also the proof of Theorem 2.4) that $\Psi^{T} J \Psi$ is a first integral of the augmented system (3.3). Since this first integral is quadratic, it is exactly preserved by Runge-Kutta methods satisfying (3.2) (see Theorem III.3.2). Hence, $\Psi_{1}^{T} J \Psi_{1}=\Psi_{0}^{T} J \Psi_{0}$ holds. The symplecticity of the Runge-Kutta method $\Phi_{h}$ then follows from Lemma 3.3, because for $\Psi_{0}=I$ we have $\Psi_{1}=\Phi_{h}^{\prime}\left(y_{0}\right)$.

Theorem 3.5 If the coefficients of a partitioned Runge-Kutta method (II.4.2) satisfy

$$
\begin{align*}
& b_{i} \widehat{a}_{i j}+\widehat{b}_{j} a_{j i}=b_{i} \widehat{b}_{j} \quad \text { for } i, j=1, \ldots, s,  \tag{3.4}\\
& b_{i}=\widehat{b}_{i} \quad \text { for } i=1, \ldots, s, \tag{3.5}
\end{align*}
$$

then it is symplectic.
If the Hamiltonian is of the form $H(p, q)=T(p)+U(q)$, i.e., it is separable, then the condition (3.4) alone implies the symplecticity of the numerical flow.

Proof. We write the solution $\Psi$ of the variational equation as

$$
\Psi=\binom{\Psi^{p}}{\Psi^{q}} .
$$

Then, the Hamiltonian system together with its variational equation (3.3) is a partitioned system with variables $\left(p, \Psi^{p}\right)$ and $\left(q, \Psi^{q}\right)$. Every component of

$$
\Psi^{T} J \Psi=\left(\Psi^{p}\right)^{T} \Psi^{q}-\left(\Psi^{q}\right)^{T} \Psi^{p}
$$

is of the form (III.3.5), so that Theorem 3.3 can be applied.

[^12]
## IV. 4 Symplecticity for Linear Problems

For quadratic Hamiltonians $H(y)=\frac{1}{2} y^{T} C y$ (where $C$ is a symmetric real matrix) the corresponding system (2.6) is linear,

$$
\begin{equation*}
y^{\prime}=J^{-1} C y . \tag{4.1}
\end{equation*}
$$

Lemma 4.1 A Runge-Kutta method, applied with step size $h$ to a linear system $y^{\prime}=L y$, is equivalent to

$$
\begin{equation*}
y_{1}=R(h L) y_{0} \tag{4.2}
\end{equation*}
$$

where the rational function $R(z)$ is given by

$$
\begin{equation*}
R(z)=1+z b^{T}(I-z A)^{-1} \mathbb{1} \tag{4.3}
\end{equation*}
$$

$A=\left(a_{i j}\right), b^{T}=\left(b_{1}, \ldots, b_{s}\right)$, and $\mathbb{1}^{T}=(1, \ldots, 1)$. The function $R(z)$ is called the stability function of the method.
Proof. The Runge-Kutta method (Definition II.1.1), applied to $y^{\prime}=L y$, reads

$$
h k_{i}=h L\left(y_{0}+\sum_{j=1}^{s} a_{i j} h k_{j}\right)
$$

or, using the the supervector $K=\left(k_{1}^{T}, \ldots, k_{s}^{T}\right)^{T}$,

$$
(I-A \otimes h L) h K=\mathbb{1} \otimes h L y_{0}
$$

(here, $A \otimes B=\left(a_{i j} B\right)$ denotes the tensor product of two matrices or vectors). Computing $h K$ from this relation, and inserting it into $y_{1}=y_{0}+\sum_{i=1}^{s} b_{i} h k_{i}=y_{0}+\left(b^{T} \otimes I\right)(h K)$ proves the statement.

For the explicit Euler method, the implicit Euler method and the implicit midpoint rule, the stability function $R(z)$ is given by

$$
1+z, \quad \frac{1}{1-z}, \quad \frac{1+z / 2}{1-z / 2}
$$

Theorem 4.2 For Runge-Kutta methods the following statements are equivalent:

- the method is symmetric for linear problems $y^{\prime}=L y$;
- the method is symplectic for problems (4.1) with symmetric $C$;
- the stability function satisfies $R(-z) R(z)=1$ for all complex $z$.

Proof. The method $y_{1}=R(h L) y_{0}$ is symmetric, if and only if $y_{0}=R(-h L) y_{1}$ holds for all initial values $y_{0}$. But this is equivalent to $R(-h L) R(h L)=I$.

Since $\Phi_{h}^{\prime}\left(y_{0}\right)=R(h L)$, symplecticity of the method for the problem (4.1) is defined by $R\left(h J^{-1} C\right)^{T} J R\left(h J^{-1} C\right)=J$. For $R(z)=P(z) / Q(z)$ this is equivalent to

$$
\begin{equation*}
P\left(h J^{-1} C\right)^{T} J P\left(h J^{-1} C\right)=Q\left(h J^{-1} C\right)^{T} J Q\left(h J^{-1} C\right) \tag{4.4}
\end{equation*}
$$

By the symmetry of $C$, the matrix $L:=J^{-1} C$ satisfies $L^{T} J=-J L$ and hence also $\left(L^{k}\right)^{T} J=J(-L)^{k}$ for $k=0,1,2, \ldots$ Consequently, (4.4) is equivalent to

$$
P\left(-h J^{-1} C\right) P\left(h J^{-1} C\right)=Q\left(-h J^{-1} C\right) Q\left(h J^{-1} C\right)
$$

which is nothing else than $R\left(-h J^{-1} C\right) R\left(h J^{-1} C\right)=I$.

We remark that symmetry and symplecticity are equivalent properties of Runge-Kutta methods only for linear problems. For general nonlinear problems, there exist symmetric methods that are not symplectic, and there exist symplectic methods that are not symmetric. For example, the trapezoidal rule

$$
\begin{equation*}
y_{1}=y_{0}+\frac{h}{2}\left(f\left(y_{0}\right)+f\left(y_{1}\right)\right) \tag{4.5}
\end{equation*}
$$

is symmetric, but it does not satisfy the condition (3.2) for symplecticity. In fact, this is true for all Lobatto IIIA methods (see Example II.4.6). On the other hand, the method of Table 4.1 satisfies the symplecticity condition (3.2), but it is clearly not symmetric (the weights do not satisfy $b_{s+1-i}=b_{i}$ ).

Table 4.1: A symplectic Radau method of order 5 [Su93]

| $\frac{4-\sqrt{6}}{10}$ | $\frac{16-\sqrt{6}}{72}$ | $\frac{328-167 \sqrt{6}}{1800}$ | $\frac{-2+3 \sqrt{6}}{450}$ |
| :---: | :---: | :---: | :---: |
| $\frac{4+\sqrt{6}}{10}$ | $\frac{328+167 \sqrt{6}}{1800}$ | $\frac{16+\sqrt{6}}{72}$ | $\frac{-2-3 \sqrt{6}}{450}$ |
| 1 | $\frac{85-10 \sqrt{6}}{180}$ | $\frac{85+10 \sqrt{6}}{180}$ | $\frac{1}{18}$ |
|  | $\frac{16-\sqrt{6}}{36}$ | $\frac{16+\sqrt{6}}{36}$ | $\frac{1}{9}$ |

## IV. 5 Campbell-Baker-Hausdorff Formula

This section is devoted to the derivation of the Campbell-Baker-Hausdorff (short CBH or BCH) formula. It was claimed in 1898 by J.E. Campbell and proved independently by Baker [Ba05] and Hausdorff [Hau06]. This formula will be the essential ingredient for the discussion of splitting methods (Sect.IV.6).

Let $A$ and $B$ be two non-commuting matrices (or operators, for which the compositions $A^{k} B^{l}$ make sense). The problem is to find a matrix $C(t)$, such that

$$
\begin{equation*}
\exp (t A) \exp (t B)=\exp C(t) \tag{5.1}
\end{equation*}
$$

As long as we do not need the explicit form of $C(t)$, this is a simple task: the expression $\exp (t A) \exp (t B)$ is a series of the form $I+t(A+B)+\mathcal{O}\left(t^{2}\right)$ and, assuming that $C(0)=0$, $\exp C(t)$ is also close to the identity for small $t$. Therefore, we can apply the logarithm to (5.1) and we get $C(t)$. Using the series expansion $\log (1+x)=x-x^{2} / 2+\ldots$, this yields $C(t)$ as a series in powers of $t$. It starts with $C(t)=t(A+B)+\mathcal{O}\left(t^{2}\right)$ and it has a positive radius of convergence, because it is obtained by elementary operations of convergent series. Consequently, the obtained series for $C(t)$ will converge for bounded $A$ and $B$, if $t$ is sufficiently small.

The main problem of the derivation of the BCH formula is to get explicit formulas for the coefficients of the series of $C(t)$. With the help of the following lemma, recurrence relations for these coefficients will be obtained, which allow for an easy computation of the first terms.


John Edward Campbell ${ }^{3}$


Henry Frederick Baker ${ }^{4}$


Felix Hausdorff ${ }^{5}$

Lemma 5.1 Let $A$ and $B$ be (non-commuting) matrices. Then, (5.1) holds, where $C(t)$ is the solution of the differential equation

$$
\begin{equation*}
C^{\prime}=A+B+\frac{1}{2}[A-B, C]+\sum_{k \geq 2} \frac{B_{k}}{k!} \operatorname{ad}_{C}^{k}(A+B) \tag{5.2}
\end{equation*}
$$

with initial value $C(0)=0$. Recall that ad ${ }_{C} A=[C, A]=C A-A C$, and that $B_{k}$ denote the Bernoulli numbers as in Lemma III.6.2.

Proof. We follow [Var74, Sect. 2.15] and we consider a matrix function $Z(s, t)$ such that

$$
\begin{equation*}
\exp (s A) \exp (t B)=\exp Z(s, t) \tag{5.3}
\end{equation*}
$$

Using Lemma III.6.1, the derivative of (5.3) with respect to $s$ is

$$
A \exp (s A) \exp (t B)=d \exp _{Z(s, t)}\left(\frac{\partial Z}{\partial s}(s, t)\right) \exp Z(s, t)
$$

so that

$$
\begin{equation*}
\frac{\partial Z}{\partial s}=d \exp _{Z}^{-1}(A)=A-\frac{1}{2}[Z, A]+\sum_{k \geq 2} \frac{B_{k}}{k!} \operatorname{ad}_{Z}^{k}(A) \tag{5.4}
\end{equation*}
$$

We next take the inverse of (5.3)

$$
\exp (-t B) \exp (-s A)=\exp (-Z(s, t))
$$

and differentiate this relation with respect to $t$. As above we get

$$
\begin{equation*}
\frac{\partial Z}{\partial t}=d \exp _{-Z}^{-1}(B)=B+\frac{1}{2}[Z, B]+\sum_{k \geq 2} \frac{B_{k}}{k!} \operatorname{ad}_{Z}^{k}(B) \tag{5.5}
\end{equation*}
$$

[^13]because $\operatorname{ad}_{-Z}^{k}(B)=(-1)^{k} \operatorname{ad}_{Z}^{k}(B)$ and the Bernoulli numbers satisfy $B_{k}=0$ for odd $k \geq 2$. A comparison of (5.1) with (5.3) gives $C(t)=Z(t, t)$. The stated differential equation for $C(t)$ therefore follows from $C^{\prime}(t)=\frac{\partial Z}{\partial s}(t, t)+\frac{\partial Z}{\partial t}(t, t)$, and from adding the relations (5.4) and (5.5).

Using Lemma 5.1 we can compute the first coefficients of the series

$$
\begin{equation*}
C(t)=t C_{1}+t^{2} C_{2}+t^{3} C_{3}+t^{4} C_{4}+\ldots . \tag{5.6}
\end{equation*}
$$

Inserting this ansatz into (5.2) and comparing like powers of $t$ gives

$$
\begin{align*}
C_{1} & =A+B \\
2 C_{2} & =\frac{1}{2}[A-B, A+B]=[A, B] \\
3 C_{3} & =\frac{1}{2}\left[A-B, \frac{1}{2}[A, B]\right]=\frac{1}{4}[A,[A, B]]+\frac{1}{4}[B,[B, A]]  \tag{5.7}\\
4 C_{4} & =\cdots=\frac{1}{6}[A,[B,[B, A]]] .
\end{align*}
$$

For the simplification of the expression for $C_{4}$ we have made use of the Jacobi identity (III.2.2). The next coefficient $C_{5}$ contains already 6 independent terms, and for higher order the expressions become soon very complicated.

For later use (construction of splitting methods) we also need a formula for the symmetric composition

$$
\begin{equation*}
\exp (t A) \exp (t B) \exp (t A)=\exp D(t) \tag{5.8}
\end{equation*}
$$

Taking the inverse of (5.8), we see that $\exp (-D(t))=\exp (D(-t))$, so that $D(-t)=$ $-D(t)$, and the expansion of $D(t)$ is in odd powers of $t$ :

$$
\begin{equation*}
D(t)=t D_{1}+t^{3} D_{3}+t^{5} D_{5}+\ldots \tag{5.9}
\end{equation*}
$$

By repeated application of the BCH formula (5.1) with coefficients given by (5.7) we find that

$$
\begin{align*}
D_{1} & =2 A+B \\
3 D_{3} & =\frac{1}{2}[B,[B, A]]-\frac{1}{2}[A,[A, B]] \tag{5.10}
\end{align*}
$$

Remark 5.2 If $A$ and $B$ are bounded operators, the series (5.6) and (5.9) converge for sufficiently small $t$. We are, however, also interested in the situation, where $A$ and $B$ are unbounded differential operators. In this case, we still have a formal identity. This means that if we expand both sides of the identities (5.1) or (5.8) into powers of $t$, then the corresponding coefficients are equal. Truncation of these series therefore introduces a defect of size $\mathcal{O}\left(t^{N}\right)$, where $N$ can be made arbitrarily large.

## IV. 6 Splitting Methods

For a motivation of splitting methods, let us consider a Hamiltonian system with separable Hamiltonian $H(p, q)=T(p)+V(q)$. It is the sum of two Hamiltonians, which depend either only on $p$ or only on $q$. The corresponding Hamiltonian systems

$$
\begin{array}{rlrl}
\dot{p} & =0 & &  \tag{6.1}\\
\dot{q} & =T_{p}(p) & \text { and } & \\
p & =-V_{q}(q) \\
\dot{q} & =0
\end{array}
$$

can be solved exactly and yield

$$
\begin{align*}
& p(t)=p_{0} \quad \text { and } \quad p(t)=p_{0}-t V_{q}\left(q_{0}\right)  \tag{6.2}\\
& q(t)=q_{0}+t T_{p}\left(p_{0}\right) \quad \text { and } \quad q(t)=q_{0},
\end{align*}
$$

respectively. Denoting the flows of these two systems by $\varphi_{t}^{T}$ and $\varphi_{t}^{V}$, one can check that the symplectic Euler method (I.1.8) is nothing other than the composition $\varphi_{h}^{T} \circ \varphi_{h}^{V}$. Since $\varphi_{h}^{T}$ and $\varphi_{h}^{V}$ are both symplectic transformations, and since the composition of symplectic maps is again symplectic, this gives an elegant proof of the symplecticity of the symplectic Euler method. Furthermore, the adjoint of the symplectic Euler method can be written as $\varphi_{h}^{V} \circ \varphi_{h}^{T}$, and by (I.3.7) the Verlet scheme becomes $\varphi_{h / 2}^{V} \circ \varphi_{h}^{T} \circ \varphi_{h / 2}^{V}$.

The idea of splitting can be applied to very general situations. We consider an arbitrary (not necessarily Hamiltonian) system $y^{\prime}=f(y)$ in $\mathbb{R}^{n}$, which can be split as

$$
\begin{equation*}
y^{\prime}=f_{1}(y)+f_{2}(y) \tag{6.3}
\end{equation*}
$$

We further assume that the flows $\varphi_{t}^{[1]}$ and $\varphi_{t}^{[2]}$ of the systems $y^{\prime}=f_{1}(y)$ and $y^{\prime}=f_{2}(y)$ can be calculated explicitly (later in Chapter V we shall study splitting methods, where $\varphi_{t}^{[1]}$ and $\varphi_{t}^{[2]}$ are replaced with some numerical approximations). An extension of the symplectic Euler method to the new situation is

$$
\begin{equation*}
\varphi_{h}^{[1]} \circ \varphi_{h}^{[2]}, \tag{6.4}
\end{equation*}
$$

which is often called the Lie-Trotter formula [Tr59]. By Taylor expansion we find that $\left(\varphi_{h}^{[1]} \circ \varphi_{h}^{[2]}\right)\left(y_{0}\right)=\varphi_{h}\left(y_{0}\right)+\mathcal{O}\left(h^{2}\right)$, so that (6.4) gives an approximation of order 1 to the solution of (6.3). The analogue of the Verlet scheme is

$$
\begin{equation*}
\varphi_{h / 2}^{[1]} \circ \varphi_{h}^{[2]} \circ \varphi_{h / 2}^{[1]}, \tag{6.5}
\end{equation*}
$$

which is the so-called Strang splitting ${ }^{6}$ [Str68]. Due to its symmetry it is a method of order 2. The order can be still further increased by suitably composing the flows $\varphi_{t}^{[1]}$ and $\varphi_{t}^{[2]}$. According to [McL95] we distinguish the following cases:

- Non-symmetric. Such methods are of the form

$$
\begin{equation*}
\varphi_{b_{m} h}^{[2]} \circ \varphi_{a_{m} h}^{[1]} \circ \varphi_{b_{m-1} h}^{[2]} \circ \ldots \circ \varphi_{a_{2} h}^{[1]} \circ \varphi_{b_{1} h}^{[2]} \circ \varphi_{a_{1} h}^{[1]} \tag{6.6}
\end{equation*}
$$

( $a_{1}$ or $b_{m}$ or both of them are allowed to be zero).

- Symmetric. Symmetric methods are obtained by a composition of the form

$$
\begin{equation*}
\varphi_{a_{m} h}^{[1]} \circ \varphi_{b_{m} h}^{[2]} \circ \ldots \circ \varphi_{a_{1} h}^{[1]} \circ \varphi_{b_{1} h}^{[2]} \circ \varphi_{a_{1} h}^{[1]} \circ \ldots \circ \varphi_{b_{m} h}^{[2]} \circ \varphi_{a_{m} h}^{[1]} \tag{6.7}
\end{equation*}
$$

(here, $b_{1}$ or $a_{m}$ or both are allowed to be zero).

- Symmetric, composed of symmetric steps. We let $\Phi_{h}=\varphi_{h / 2}^{[1]} \circ \varphi_{h}^{[2]} \circ \varphi_{h / 2}^{[1]}$ or $\Phi_{h}=\varphi_{h / 2}^{[2]} \circ \varphi_{h}^{[1]} \circ \varphi_{h / 2}^{[2]}$, and we consider the composition

$$
\begin{equation*}
\Phi_{b_{m} h} \circ \Phi_{b_{m-1} h} \circ \ldots \circ \Phi_{b_{1} h} \circ \Phi_{b_{0} h} \circ \Phi_{b_{1} h} \circ \ldots \circ \Phi_{b_{m-1} h} \circ \Phi_{b_{m} h} \tag{6.8}
\end{equation*}
$$

[^14]An early contribution to this subject is the article of Ruth [Ru83], where, for the special case (6.1), a non-symmetric method (6.6) of order 3 with $m=3$ is constructed. A systematic study of such methods has started with the articles of Suzuki [Su90, Su92] and Yoshida [Yo90].

In all three situations the problem is the same: what are the conditions on the parameters $a_{i}$, $b_{i}$, such that the compositions (6.6), (6.7) or (6.8) approximate the flow $\varphi_{h}$ of (6.3) to a given order p?

In order to compare the expressions (6.6), (6.7) and (6.8) with the flow $\varphi_{h}$ of (6.3), it is convenient to introduce the differential operators $D_{i}$ (Lie derivative) which, for differentiable functions $F: \mathbb{R}^{n} \rightarrow \mathbb{R}^{n}$, are defined by

$$
\begin{equation*}
D_{i} F(y)=F^{\prime}(y) f_{i}(y) \tag{6.9}
\end{equation*}
$$

where $f_{i}(y)$ is the function of (6.3). This means that, if $y(t)$ is a solution of $y^{\prime}=f_{i}(y)$, then

$$
\begin{equation*}
\frac{d}{d t} F(y(t))=\left(D_{i} F\right)(y(t)) \tag{6.10}
\end{equation*}
$$

Applying iteratively this operator to the identity map $F(y)=y$, we obtain for the solution $y(t)$ of $y^{\prime}=f_{i}(y)$ that $y^{\prime}(t)=D_{i} y(t), y^{\prime \prime}(t)=D_{i}^{2} y(t)$, etc. Consequently, for analytic functions, the solution $\varphi_{t}^{[i]}\left(y_{0}\right)$ is given by

$$
\begin{equation*}
\varphi_{t}^{[i]}\left(y_{0}\right)=\sum_{k \geq 0} \frac{t^{k}}{k!} D_{i}^{k} y_{0}=\exp \left(t D_{i}\right) y_{0} \tag{6.11}
\end{equation*}
$$

Lemma 6.1 Let $\varphi_{t}^{[1]}$ and $\varphi_{t}^{[2]}$ be the flows of the differential equations $y^{\prime}=f_{1}(y)$ and $y^{\prime}=f_{2}(y)$, respectively. For their composition we then have

$$
\left(\varphi_{s}^{[1]} \circ \varphi_{t}^{[2]}\right)\left(y_{0}\right)=\exp \left(t D_{2}\right) \exp \left(s D_{1}\right) y_{0}
$$

(observe the reversed order of the operators).

Proof. For an arbitrary smooth function $F(y)$, it follows from an iterated application of (6.10) that

$$
\frac{d^{k}}{d t^{k}} F\left(\varphi_{t}^{[2]}\left(y_{0}\right)\right)=D_{2}^{k} F\left(\varphi_{t}^{[2]}\left(y_{0}\right)\right)
$$

so that by Taylor series expansion $F\left(\varphi_{t}^{[2]}\left(y_{0}\right)\right)=\sum_{k \geq 0} \frac{t^{k}}{k!} D_{2}^{k} F\left(y_{0}\right)$. Putting $F(y):=$ $\varphi_{s}^{[1]}(y)$ and using (6.11) gives

$$
\begin{equation*}
\left(\varphi_{s}^{[1]} \circ \varphi_{t}^{[2]}\right)\left(y_{0}\right)=\left(\sum_{k \geq 0} \frac{t^{k}}{k!} D_{2}^{k}\right)\left(\sum_{l \geq 0} \frac{s^{l}}{l!} D_{1}^{l}\right) y_{0} \tag{6.12}
\end{equation*}
$$

which proves the statement.

In general, the two operators $D_{1}$ and $D_{2}$ do not commute, so that the composition $\exp \left(t D_{2}\right) \exp \left(t D_{1}\right) y_{0}$ is different from $\exp \left(t\left(D_{1}+D_{2}\right)\right) y_{0}$, which represents the solution $\varphi_{t}\left(y_{0}\right)$ of $y^{\prime}=f(y)=f_{1}(y)+f_{2}(y)$.

Order Conditions The derivation of the order conditions for splitting methods can be done as follows: with the use of Lemma 6.1 we write the method as a product of exponentials, then we apply the Campbell-Baker-Hausdorff formula to get one exponential of a series in powers of $h$. Finally, we compare this series with $h\left(D_{1}+D_{2}\right)$, which corresponds to the exact solution of (6.3).

Let us illustrate this procedure with the methods of type (6.8) (see [Yo90]). Using Lemma 6.1 and formulas (5.8), (5.10), the second order integrator $\Phi_{h}=\varphi_{h / 2}^{[1]} \circ \varphi_{h}^{[2]} \circ \varphi_{h / 2}^{[1]}$ can be written as

$$
\begin{align*}
\Phi_{h} & =\exp \left(h D_{1} / 2\right) \exp \left(h D_{2}\right) \exp \left(h D_{1} / 2\right)  \tag{6.13}\\
& =\exp \left(h \alpha_{1}+h^{3} \alpha_{3}+h^{5} \alpha_{5}+\ldots\right),
\end{align*}
$$

where $\alpha_{1}=D_{1}+D_{2}, \alpha_{3}=\frac{1}{12}\left[D_{2},\left[D_{2}, D_{1}\right]\right]-\frac{1}{24}\left[D_{1},\left[D_{1}, D_{2}\right]\right]$. The Lie-bracket for differential operators is defined in the usual way, i.e., $\left[D_{1}, D_{2}\right]=D_{1} D_{2}-D_{2} D_{1}$. We next define $\Psi^{(j)}$ recursively by

$$
\begin{equation*}
\Psi^{(0)}=\Phi_{b_{0} h}, \quad \Psi^{(j)}=\Phi_{b_{j} h} \circ \Psi^{(j-1)} \circ \Phi_{b_{j} h} \tag{6.14}
\end{equation*}
$$

so that $\Psi^{(m)}$ is equal to the method (6.8).
Lemma 6.2 The operators $\Psi^{(j)}$, defined by (6.14) and (6.13), satisfy

$$
\begin{equation*}
\Psi^{(j)}\left(y_{0}\right)=\exp \left(A_{1, j} h \alpha_{1}+A_{3, j} h^{3} \alpha_{3}+A_{5, j} h^{5} \alpha_{5}+B_{5, j} h^{5}\left[\alpha_{1},\left[\alpha_{1}, \alpha_{3}\right]\right]+\mathcal{O}\left(h^{7}\right)\right) y_{0} \tag{6.15}
\end{equation*}
$$

where

$$
A_{1,0}=b_{0}, \quad A_{3,0}=b_{0}^{3}, \quad A_{5,0}=b_{0}^{5}, \quad B_{5,0}=0
$$

and

$$
\begin{aligned}
A_{1, j} & =A_{1, j-1}+2 b_{j} \\
A_{3, j} & =A_{3, j-1}+2 b_{j}^{3} \\
A_{5, j} & =A_{5, j-1}+2 b_{j}^{5} \\
B_{5, j} & =B_{5, j-1}+\frac{1}{6}\left(A_{1, j-1}^{2} b_{j}^{3}-A_{1, j-1} A_{3, j-1} b_{j}-A_{3, j-1} b_{j}^{2}+A_{1, j-1} b_{j}^{4}\right)
\end{aligned}
$$

Proof. We use the formulas (5.8), (5.9), (5.10) with $t A$ replaced with $b_{j} h \alpha_{1}+\left(b_{j} h\right)^{3} \alpha_{3}+\ldots$, and $t B$ replaced with $A_{1, j-1} h \alpha_{1}+A_{3, j-1} h^{3} \alpha_{3}+\ldots$. This gives $\Psi^{(j)}\left(y_{0}\right)=\exp (D(h)) y_{0}$ with

$$
\begin{aligned}
D(h) & =\left(2 b_{j}+A_{1, j-1}\right) h \alpha_{1}+\left(2 b_{j}^{3}+A_{3, j-1}\right) h^{3} \alpha_{3}+\left(2 b_{j}^{5}+A_{5, j-1}\right) h^{5} \alpha_{5}+B_{5, j-1} h^{5}\left[\alpha_{1},\left[\alpha_{1}, \alpha_{3}\right]\right] \\
& +\frac{1}{6}\left[A_{1, j-1} h \alpha_{1},\left[A_{1, j-1} h \alpha_{1}+A_{3, j-1} h^{3} \alpha_{3}, b_{j} h \alpha_{1}+b_{j}^{3} h^{3} \alpha_{3}\right]\right] \\
& -\frac{1}{6}\left[b_{j} h \alpha_{1},\left[b_{j} h \alpha_{1}+b_{j}^{3} h^{3} \alpha_{3}, A_{1, j-1} h \alpha_{1}+A_{3, j-1} h^{3} \alpha_{3}\right]\right]+\mathcal{O}\left(h^{7}\right) .
\end{aligned}
$$

A comparison of $D(h)$ with the argument in (6.15) proves the statement.

Theorem 6.3 The order conditions for the splitting method (6.8) are:

- order 2: $\quad A_{1, m}=1$,
- order 4: $\quad A_{1, m}=1, \quad A_{3, m}=0$,
- order 6: $\quad A_{1, m}=1, \quad A_{3, m}=0, \quad A_{5, m}=0, \quad B_{5, m}=0$.

The coefficients $A_{i, m}$ and $B_{5, m}$ are those defined in Lemma 6.2.

Proof. This is an immediate consequence of Lemma 6.2, because the conditions of order $p$ imply that the Taylor series expansion of $\Psi^{(m)}\left(y_{0}\right)$ coincides with that of the solution $\varphi_{h}\left(y_{0}\right)=\exp \left(h\left(D_{1}+D_{2}\right)\right) y_{0}$ up to terms of size $\mathcal{O}\left(h^{p}\right)$.

It is interesting to note that the order conditions of Theorem 6.3 do not depend on the special form of $\alpha_{3}$ and $\alpha_{5}$ in (6.13). We also remark that the composition methods of Sect. II. 6 are a special case of the splitting method (6.8). Theorem 6.3 therefore explains the order conditions (II.6.3) and (II.6.4), which were mysterious with the techniques of Chapter II.
[Yo90] solves the order conditions for order 6 with $m=3$ (four equations for the four parameters $\left.b_{0}, b_{1}, b_{2}, b_{3}\right)$. He finds three solutions, one of which is given in the end of Sect.II.6. [Yo90] also presents some methods of order 8. A careful investigation of symmetric splitting methods of orders 2 to 8 can be found in [McL95]. There, several new methods with small error constants are presented.

Remark 6.4 We emphasize that splitting methods are an important tool for the construction of symplectic integrators. If we split a Hamiltonian as $H(y)=H_{1}(y)+H_{2}(y)$, and if we consider the vector fields $f_{i}(y)=J^{-1} \nabla H_{i}(y)$, then the flows $\varphi_{t}^{[i]}$ are symplectic, and therefore all splitting methods are automatically symplectic.

## IV. 7 Volume Preservation

## IV. 8 Generating Functions

## IV. 9 Variational Approach

Marsden, etc

## IV. 10 Symplectic Integrators on Manifolds

## IV. 11 Exercises

1. Prove that a linear transformation $A: \mathbb{R}^{2} \rightarrow \mathbb{R}^{2}$ is symplectic, if and only if $\operatorname{det} A=1$.
2. Prove that the flow of a Hamiltonian system satisfies $\operatorname{det} \varphi_{t}^{\prime}(y)=1$ for all $y$ and all $t$. Deduce from this result that the flow is volume preserving, i.e., for $B \subset \mathbb{R}^{2 d}$ it holds that $\operatorname{vol}\left(\varphi_{t}(B)\right)=\operatorname{vol}(B)$ for all $t$.
3. Consider the Hamiltonian system $y^{\prime}=J^{-1} \nabla H(y)$ and a variable transformation $y=\varphi(z)$. Prove that, for a symplectic transformation $\varphi(z)$, the system in the $z$-coordinates is again Hamiltonian with $\widetilde{H}(z)=H(\varphi(z))$.
4. Consider a Hamiltonian system with $H(p, q)=\frac{1}{2} p^{T} p+V(q)$. Let $q=\chi(Q)$ be a change of position coordinates. How has one to define the variable $P$ (as a function of $p$ and $q$ ) so that the system in the new variables $(P, Q)$ is again Hamiltonian?
Result. $\quad P=\chi^{\prime}(Q)^{T} p$.
5. Let $\alpha$ and $\beta$ be the generalized coordinates of the double pendulum, whose kinetic and potential energies are

$$
\begin{aligned}
T & =\frac{m_{1}}{2}\left(\dot{x}_{1}^{2}+\dot{y}_{1}^{2}\right)+\frac{m_{2}}{2}\left(\dot{x}_{2}^{2}+\dot{y}_{2}^{2}\right) \\
U & =m_{1} g y_{1}+m_{2} g y_{2} .
\end{aligned}
$$

Determine the generalized momenta of the corresponding Hamiltonian system.

6. Consider the transformation $(r, \varphi) \mapsto(p, q)$, defined by

$$
p=\psi(r) \cos \varphi, \quad q=\psi(r) \sin \varphi .
$$

For which function $\psi(r)$ is it a symplectic transformation?
7. Write Kepler's problem with Hamiltonian

$$
H(p, q)=\frac{1}{2}\left(p_{1}^{2}+p_{2}^{2}\right)-\frac{1}{\sqrt{q_{1}^{2}+q_{2}^{2}}}
$$

in polar coordinates $q_{1}=r \cos \varphi, q_{2}=r \sin \varphi$. What are the conjugated generalized momenta $p_{r}, p_{\varphi}$ ? What is the Hamiltonian in the new coordinates.
8. On the set $U=\left\{(p, q) ; p^{2}+q^{2}>0\right\}$ consider the differential equation

$$
\begin{equation*}
\binom{\dot{p}}{\dot{q}}=\frac{1}{p^{2}+q^{2}}\binom{p}{q} . \tag{11.1}
\end{equation*}
$$

a) Prove that its flow is symplectic everywhere on $U$.
b) On every simply-connected subset of $U$ the vector field (11.1) is Hamiltonian (with $H(p, q)=\operatorname{Im} \log (p+i q)+$ Const $)$.
c) It is not possible to find a differentiable function $H: U \rightarrow \mathbb{R}$ such that (11.1) is equal to $J^{-1} \nabla H(p, q)$ for all $(p, q) \in U$.
Remark. The vector field (11.1) is called locally Hamiltonian.
9. Prove that the definition (2.5) of $\Omega(M)$ does not depend on the parametrization $\varphi$, i.e., the parametrization $\psi=\varphi \circ \alpha$, where $\alpha$ is a diffeomorphism between suitable domains of $\mathbb{R}^{2}$, leads to the same result.
10. Prove that the coefficient $C_{4}$ in the series (5.6) of the Campbell-Baker-Hausdorff formula is given by $[A,[B,[B, A]]] / 6$.
11. Deduce the BCH formula from the Magnus expansion (III.6.9).

Hint. For constant matrices $A$ and $B$ consider the matrix function $A(t)$, defined by $A(t)=B$ for $0 \leq t \leq 1$ and $A(t)=A$ for $1 \leq t \leq 2$.
12. Prove that the series (5.6) of the BCH formula converges for $|t|<\ln 2 /(\|A\|+\|B\|)$.
13. What are the conditions on the parameters $a_{i}$ and $b_{i}$, such that the splitting method (6.6) is of order 2 , of order 3 ?
14. How many order conditions have to be satisfied by a symmetric splitting method (6.7) to get order 4? The result is 4 .

## Chapter V

## Backward Error Analysis

One of the greatest virtues of backward analysis ... is that when it is the appropriate form of analysis it tends to be very markedly superior to forward analysis. Invariably in such cases it has remarkable formal simplicity and gives deep insight into the stability (or lack of it) of the algorithm.
(J.H. Wilkinson, IMA Bulletin 1986)

The origin of backward error analysis dates back to the work of Wilkinson [Wi60] in numerical linear algebra. For the study of integration methods for ordinary differential equations, its importance was seen much later. The present chapter is devoted to this theory. It is very useful, when the qualitative behaviour of numerical methods is of interest, and when statements over very long time intervals are needed. The formal analysis (construction of the modified equation, study of its properties) gives already a lot of insight into numerical methods. For a rigorous treatment, the modified equation, which is a formal series in powers of the step size, has to be truncated. The error, induced by such a truncation, can be made exponentially small, and the results remain valid on exponentially long time intervals.

## V. 1 Modified Differential Equation - Examples

Consider a differential equation

$$
y^{\prime}=f(y), \quad y(0)=y_{0}
$$

and a numerical method $\Phi_{h}(y)$ which produces the approximations

$$
y_{0}, y_{1}, y_{2}, \ldots .
$$

A forward error analysis consists of the study of the errors $y_{1}-\varphi_{h}\left(y_{0}\right)$ (local er-
 ror) and $y_{n}-\varphi_{n h}\left(y_{0}\right)$ (global error) in the solution space. The idea of backward error analysis is to search for a modified differential equation $\widetilde{y}^{\prime}=f_{h}(\widetilde{y})$ of the form

$$
\begin{equation*}
\widetilde{y}^{\prime}=f(\widetilde{y})+h f_{2}(\widetilde{y})+h^{2} f_{3}(\widetilde{y})+\ldots, \quad \widetilde{y}(0)=y_{0} \tag{1.1}
\end{equation*}
$$

such that $y_{n}=\widetilde{y}(n h)$, and in studying the difference of the vector fields $f(y)$ and $f_{h}(y)$. This then gives much insight into the qualitative behaviour of the numerical solution and into the global error $y_{n}-y(n h)=\widetilde{y}(n h)-y(n h)$. We remark that the series in (1.1) usually diverges and that one has to truncate it suitably. The effect of such a truncation will be studied in Sect. V.5. For the moment we content ourselves with a formal analysis without taking care of convergence issues. The idea of interpreting the numerical solution as the exact solution of a modified equation is common to many numerical analysts ("... This is possible since the map is the solution of some physical Hamiltonian problem which, in some sense, is close to the original problem", Ruth [Ru83], or ". . the symplectic integrator creates a numerical Hamiltonian system that is close to the original ...", Gladman, Duncan and Candy [GDC91]). A systematic study started with the work of Griffiths and Sanz-Serna [GSS86], Feng Kang [FeK91], Sanz-Serna [SS92], Yoshida [Yo93], Eirola [Ei93], Fiedler and Scheurle [FiS96], and many others.

For the computation of the modified equation (1.1) we put $y:=\widetilde{y}(t)$ for a fixed $t$, and we expand the solution of (1.1) into a Taylor series

$$
\begin{align*}
\widetilde{y}(t+h) & =y+h\left(f(y)+h f_{2}(y)+h^{2} f_{3}(y)+\ldots\right) \\
& +\frac{h^{2}}{2!}\left(f^{\prime}(y)+h f_{2}^{\prime}(y)+\ldots\right)\left(f(y)+h f_{2}(y)+\ldots\right)+\ldots \tag{1.2}
\end{align*}
$$

We assume that the numerical method $\Phi_{h}(y)$ can be expanded as

$$
\begin{equation*}
\Phi_{h}(y)=y+h f(y)+h^{2} d_{2}(y)+h^{3} d_{3}(y)+\ldots \tag{1.3}
\end{equation*}
$$

(the coefficient of $h$ is $f(y)$ for consistent methods). The functions $d_{j}(y)$ are known and are typically composed of $f(y)$ and its derivatives. For the explicit Euler method we simply have $d_{j}(y)=0$ for all $j \geq 2$. In order to get $\widetilde{y}(n h)=y_{n}$ for all $n$, we must have $\widetilde{y}(t+h)=\Phi_{h}(y)$. Comparing like powers of $h$ in the expressions (1.2) and (1.3) yields recurrence relations for the functions $f_{j}(y)$, namely,

$$
\begin{align*}
f_{2}(y) & =d_{2}(y)-\frac{1}{2!} f^{\prime} f(y) \\
f_{3}(y) & =d_{3}(y)-\frac{1}{3!}\left(f^{\prime \prime}(f, f)(y)+f^{\prime} f^{\prime} f(y)\right)-\frac{1}{2!}\left(f^{\prime} f_{2}(y)+f_{2}^{\prime} f(y)\right) \tag{1.4}
\end{align*}
$$

Example 1.1 Consider the scalar differential equation

$$
y^{\prime}=y^{2}, \quad y(0)=1
$$

with exact solution $y(t)=1 /(1-t)$. It has a singularity at $t=1$. We apply the explicit Euler method $y_{n+1}=y_{n}+h f\left(y_{n}\right)$ with step size $h=$ 0.02. The picture to the right presents the exact solution (dashed curve) together with the numerical solution (bullets). The above procedure for the computation of the modified equation, implemented as a Maple program (see [HaL99]) gives


```
>fcn := y -> y^2:
>nn := 6:
> fcoe[1] := fcn(y):
> for n from 2 by 1 to nn do
> modeq := sum(h^j*fcoe[j+1], j=0..n-2):
> diffy[0] := y:
> for i from 1 by 1 to n do
> diffy[i] := diff(diffy[i-1],y)*modeq:
> od:
> ytilde := sum(h^k*diffy[k]/k!, k=0..n):
> res := ytilde-y-h*fcn(y):
> tay := convert(series(res,h=0,n+1), polynom):
> fcoe[n] := -coeff(tay,h,n):
> od:
> simplify(sum(h^j*fcoe[j+1], j=0..nn-1));
```

Its output is

$$
\begin{equation*}
\widetilde{y}^{\prime}=\widetilde{y}^{2}-h \widetilde{y}^{3}+h^{2} \frac{3}{2} \widetilde{y}^{4}-h^{3} \frac{8}{3} \widetilde{y}^{5}+h^{4} \frac{31}{6} \widetilde{y}^{6}-h^{5} \frac{157}{15} \widetilde{y}^{7} \pm \ldots . \tag{1.5}
\end{equation*}
$$

The above picture also presents the solution of the modified equation, when truncated after $1,2,3$, and 4 terms. We observe an excellent agreement of the numerical solution with the exact solution of the modified equation.

Example 1.2 We next consider the Volterra-Lotka equations

$$
q^{\prime}=q(p-1), \quad p^{\prime}=p(2-q)
$$

and we apply (a) the explicit Euler method, and (b) the symplectic Euler method, both with constant step size $h=0.1$ The first terms of their modified equations are ([Hai99])
(a) $q^{\prime}=q(p-1)-\frac{h}{2} q\left(p^{2}-p q+1\right)+\mathcal{O}\left(h^{2}\right)$,

$$
p^{\prime}=-p(q-2)-\frac{h}{2} p\left(q^{2}-p q-3 q+4\right)+\mathcal{O}\left(h^{2}\right)
$$

(b) $\quad q^{\prime}=q(p-1)-\frac{h}{2} q\left(p^{2}+p q-4 p+1\right)+\mathcal{O}\left(h^{2}\right)$,

$$
p^{\prime}=-p(q-2)+\frac{h}{2} p\left(q^{2}+p q-5 q+4\right)+\mathcal{O}\left(h^{2}\right) .
$$




Fig. 1.1: Study of the truncation in the modified equation
(a) explicit Euler, $h=0.1$

(b) symplectic Euler, $h=0.1$




Fig. 1.2: Numerical solution compared to the exact and modified flows

Fig. 1.2 shows the numerical solutions for initial values indicated by a thick dot. In the pictures to the left they are embedded in the exact flow of the differential equation, whereas in those to the right they are embedded in the flow of the modified differential equation, truncated after the $h^{2}$ terms. As in the first example, we observe an excellent agreement of the numerical solution with the exact solution of the modified equation. For the symplectic Euler method, the solutions of the truncated modified equation are periodic, as it is the case for the unperturbed problem (Exercise 3).

In Fig. 1.1 we present the numerical solution and the exact solution of the modified equation, once truncated after the $h$ terms (dashed-dotted), and once truncated after the $h^{2}$ terms (dotted). The exact solution of the problem is included as a solid curve. This shows that taking more terms in the modified equation usually improves the agreement of its solution with the numerical approximation of the method.

Example 1.3 For a linear differential equation with constant coefficients

$$
y^{\prime}=A y, \quad y(0)=y_{0}
$$

we consider numerical methods which yield $y_{n+1}=R(h A) y_{n}$, where $R(z)$ is the stability function of the method (see Lemma IV.4.1). In this case we get $y_{n}=R(h A)^{n} y_{0}$, so that $y_{n}=\widetilde{y}(n h)$, where $\widetilde{y}(t)=R(h A)^{t / h} y_{0}=\exp \left(\frac{t}{h} \ln R(h A)\right) y_{0}$ is the solution of the modified differential equation

$$
\begin{equation*}
\widetilde{y}^{\prime}=\frac{1}{h} \ln R(h A) \widetilde{y}=\left(A+h b_{2} A^{2}+h^{2} b_{3} A^{3}+\ldots\right) \widetilde{y} \tag{1.6}
\end{equation*}
$$

with suitable constants $b_{2}, b_{3}, \ldots$. Since $R(z)=1+z+\mathcal{O}\left(z^{2}\right)$ and $\ln (1+x)=x-$ $x^{2} / 2+\mathcal{O}\left(x^{3}\right)$ both have a positive radius of convergence, the series (1.6) converges for $|h| \leq h_{0}$ with some small $h_{0}>0$. We shall see later that this is an exceptional situation. In general, the modified equation is a formal divergent series.

## V. 2 Structure Preservation

This section is devoted to the study of properties of the modified differential equation, and to the question of the extent to which structures (such as Hamiltonian) in the problem $y^{\prime}=f(y)$ can carry over to the modified equation.

Order Suppose that the method $y_{n+1}=\Phi_{h}\left(y_{n}\right)$ is of order $p$, so that

$$
\Phi_{h}(y)=\varphi_{h}(y)+h^{p+1} \delta_{p+1}(y)+\mathcal{O}\left(h^{p+2}\right),
$$

where $\varphi_{t}(y)$ denotes the exact flow of $y^{\prime}=f(y)$, and $h^{p+1} \delta_{p+1}(y)$ is the leading term of the local truncation error. From the considerations of the beginning of Sect. V. 1 it follows that $f_{j}(y)=0$ for $2 \leq j \leq p$, and that $f_{p+1}(y)=\delta_{p+1}(y)$. Consequently, the modified equation becomes

$$
\begin{equation*}
\widetilde{y}^{\prime}=f(\widetilde{y})+h^{p} f_{p+1}(\widetilde{y})+h^{p+1} f_{p+2}(\widetilde{y})+\ldots, \quad \widetilde{y}(0)=y_{0} . \tag{2.1}
\end{equation*}
$$

By the nonlinear variation of constants formula, the difference between its solution $\widetilde{y}(t)$ and the solution $y(t)$ of $y^{\prime}=f(y)$ satisfies

$$
\begin{equation*}
\widetilde{y}(t)-y(t)=h^{p} e_{p}(t)+h^{p+1} e_{p+1}(t)+\ldots . \tag{2.2}
\end{equation*}
$$

Since $y_{n}=\widetilde{y}(n h)+\mathcal{O}\left(h^{N}\right)$ for the solution of a truncated modified equation, this proves the existence of an asymptotic expansion in powers of $h$ for the global error $y_{n}-y(n h)$.

Modified Equation of Adjoint Method Consider a numerical method $\Phi_{h}$. Its adjoint $y_{n+1}=\Phi_{h}^{*}\left(y_{n}\right)$ is defined by the relation $y_{n}=\Phi_{-h}\left(y_{n+1}\right)$ (see Definition II.3.1). Hence, the solution $\widetilde{y}(t)$ of the modified equation for $\Phi_{h}^{*}$ has to satisfy $\widetilde{y}(t)=\Phi_{-h}(\widetilde{y}(t+h))$ or, equivalently, $\widetilde{y}(t-h)=\Phi_{-h}(y)$ with $y:=\widetilde{y}(t)$. We get this relation, if we replace $h$ with $-h$ in the formulas (1.1), (1.2) and (1.3). Hence, the coefficient functions $f_{j}^{*}(y)$ of the modified equation for the adjoint method $\Phi_{h}^{*}$ satisfy

$$
\begin{equation*}
f_{j}^{*}(y)=(-1)^{j+1} f_{j}(y) \tag{2.3}
\end{equation*}
$$

where $f_{j}(y)$ are the coefficient functions for the method $\Phi_{h}$.
Modified Equation for Symmetric Methods For symmetric methods we have $\Phi_{h}^{*}=$ $\Phi_{h}$, implying $f_{j}^{*}(y)=f_{j}(y)$. From (2.3) it therefore follows that $f_{j}(y)=0$ whenever $j$ is even, so that (1.1) has an expansion in even powers of $h$. As a consequence, the asymptotic expansion (2.2) of the global error is also in even powers of $h$. This property is responsible for the success of extrapolation methods.

## V.2.1 Reversible Problems and Symmetric Methods

Let $\rho$ be an invertible linear transformation in the phase space of $y^{\prime}=f(y)$. This differential equation is called reversible (more precisely, $\rho$-reversible) if

$$
\begin{equation*}
f(\rho y)=-\rho f(y) \quad \text { for all } y \tag{2.4}
\end{equation*}
$$

For reversible differential equations the exact flow $\varphi_{t}(y)$ satisfies

$$
\begin{equation*}
\rho \circ \varphi_{t}=\varphi_{-t} \circ \rho=\varphi_{t}^{-1} \circ \rho \tag{2.5}
\end{equation*}
$$

(see the picture to the right). This follows from

$$
\begin{aligned}
\frac{d}{d t}\left(\rho \circ \varphi_{t}\right)(y) & =\rho f\left(\varphi_{t}(y)\right)=-f\left(\left(\rho \circ \varphi_{t}\right)(y)\right) \\
\frac{d}{d t}\left(\varphi_{-t} \circ \rho\right)(y) & =-f\left(\left(\varphi_{-t} \circ \rho\right)(y)\right)
\end{aligned}
$$

because all expressions of (2.5) satisfy the same differential equation with the same initial value
 $\left(\rho \circ \varphi_{0}\right)(y)=\left(\varphi_{0} \circ \rho\right)(y)=\rho y$. A typical example is the partitioned system

$$
\begin{equation*}
\dot{p}=f(p, q), \quad \dot{q}=g(p, q) \tag{2.6}
\end{equation*}
$$

where $f(-p, q)=f(p, q)$ and $g(-p, q)=-g(p, q)$. Here, the transformation $\rho$ is given by $\rho(p, q)=(-p, q)$. Hamiltonian systems with a Hamiltonian satisfying $H(-p, q)=H(p, q)$ are reversible, as are all second order differential equations $\ddot{q}=f(q)$ written as $\dot{p}=f(q)$, $\dot{q}=p$. If $p$ and $q$ are scalar, and if (2.6) is $\rho$-reversible for $\rho(p, q)=(-p, q)$, then any solution that crosses the $q$-axis twice is periodic (Exercise 7, see also the solution of the pendulum problem in Fig. IV.2.1). For which methods does the numerical solution have the same geometric property?

Theorem 2.1 Consider a $\rho$-reversible differential equation $y^{\prime}=f(y)$, and a numerical method $\Phi_{h}(y)$ satisfying $\rho \circ \Phi_{h}=\Phi_{-h} \circ \rho$. Then, the modified differential equation is $\rho$-reversible, if and only if the method is symmetric.

Proof. We follow the ideas of [HSt97]. Let $f_{h}(y)=f(y)+h f_{2}(y)+h^{2} f_{3}(y)+\ldots$ be the formal vector field of the modified equation for $\Phi_{h}$. The value $z(h)=\left(\rho \circ \Phi_{h}\right)\left(y_{0}\right)$ is then obtained as the solution of

$$
z^{\prime}=\rho f_{h}\left(\rho^{-1} z\right), \quad z(0)=\rho y_{0}
$$

and the value $v(h)=\left(\Phi_{-h} \circ \rho\right)\left(y_{0}\right)$ is obtained from

$$
v^{\prime}=-f_{-h}(v), \quad v(0)=\rho y_{0}
$$

By our assumption $\rho \circ \Phi_{h}=\Phi_{-h} \circ \rho$, and by the uniqueness of the expansion of the modified equation, this implies that $\rho \circ f_{h} \circ \rho^{-1}=-f_{-h}$, so that

$$
\begin{equation*}
\rho \circ f_{j}=(-1)^{j} f_{j} \circ \rho \tag{2.7}
\end{equation*}
$$

for the coefficient functions $f_{j}(y)$ of the modified equation. Consequently, the modified equation is $\rho$-reversible, if and only if $f_{j}(y)=0$ whenever $j$ is even. But this is precisely the characterization of symmetric methods.

The assumption $\rho \circ \Phi_{h}=\Phi_{-h} \circ \rho$ is satisfied for all numerical methods for which $y_{1}$ depends only on the product $h f(y)$. This is the case for all Runge-Kutta methods. Hence, the relations (2.7) are satisfied for all such methods applied to a $\rho$-reversible system.

Partitioned Runge-Kutta methods (II.4.2) applied to partitioned systems (2.6) satisfy $\rho \circ \Phi_{h}=\Phi_{-h} \circ \rho$, if the transformation $\rho$ does not mix the $p$ and $q$ components, i.e., if $\rho(p, q)=\left(\rho_{1}(p), \rho_{2}(q)\right)$.

## V.2.2 Hamiltonian Systems and Symplectic Methods

We now present one of the most important results of this chapter. We consider a Hamiltonian system $y^{\prime}=J^{-1} \nabla H(y)$ with an infinitely differentiable Hamiltonian $H(y)$.

Theorem 2.2 If a symplectic method $\Phi_{h}(y)$ is applied to a Hamiltonian system with a smooth Hamiltonian $H: \mathbb{R}^{2 d} \rightarrow \mathbb{R}$, then the modified equation (1.1) is also Hamiltonian. More precisely, there exist smooth functions $H_{j}: \mathbb{R}^{2 d} \rightarrow \mathbb{R}$ for $j=2,3, \ldots$, such that $f_{j}(y)=J^{-1} \nabla H_{j}(y)$.

Proof. We prove the statement by induction (see [BG94] and also [Rei98] in the context of subspaces of Lie algebras). Assume that $f_{j}(y)=J^{-1} \nabla H_{j}(y)$ for $j=1,2, \ldots, r$ (this is satisfied for $r=1$, because $\left.f_{1}(y)=f(y)=J^{-1} \nabla H(y)\right)$. We have to prove the existence of a Hamiltonian $H_{r+1}(y)$. The idea is to consider the truncated modified equation

$$
\begin{equation*}
\widetilde{y}^{\prime}=f(\widetilde{y})+h f_{2}(\widetilde{y})+\ldots+h^{r-1} f_{r}(\widetilde{y}), \quad \widetilde{y}(0)=y_{0}, \tag{2.8}
\end{equation*}
$$

which is a Hamiltonian system with Hamiltonian $H(y)+h H_{2}(y)+\ldots+h^{r-1} H_{r}(y)$. Its flow $\varphi_{r, t}\left(y_{0}\right)$, compared to that of (1.1), satisfies $\Phi_{h}\left(y_{0}\right)=\varphi_{r, h}\left(y_{0}\right)+h^{r+1} f_{r+1}\left(y_{0}\right)+\mathcal{O}\left(h^{r+2}\right)$, so that also

$$
\Phi_{h}^{\prime}\left(y_{0}\right)=\varphi_{r, h}^{\prime}\left(y_{0}\right)+h^{r+1} f_{r+1}^{\prime}\left(y_{0}\right)+\mathcal{O}\left(h^{r+2}\right) .
$$

By our assumption on the method and by the induction hypothesis, $\Phi_{h}$ and $\varphi_{r, h}$ are symplectic transformations. This, together with $\varphi_{r, h}^{\prime}\left(y_{0}\right)=I+\mathcal{O}(h)$, therefore implies

$$
J=\Phi_{h}^{\prime}\left(y_{0}\right)^{T} J \Phi_{h}^{\prime}\left(y_{0}\right)=J+h^{r+1}\left(f_{r+1}^{\prime}\left(y_{0}\right)^{T} J+J f_{r+1}^{\prime}\left(y_{0}\right)\right)+\mathcal{O}\left(h^{r+2}\right) .
$$

Consequently, the matrix $J f_{r+1}^{\prime}(y)$ is symmetric and the existence of $H_{r+1}(y)$ satisfying $f_{r+1}(y)=J^{-1} \nabla H_{r+1}(y)$ follows from Lemma IV.2.7. This part of the proof is similar to that of Theorem IV.2.6.

For Hamiltonians $H: U \rightarrow \mathbb{R}$ the statement of the above theorem remains valid as long as Lemma IV.2.7 is applicable. This is the case for star-shaped $U \subset \mathbb{R}^{2 d}$ and for simply connected domains $U$, but not in general (see the discussion after the proof of Lemma IV.2.7). In Sect. V. 4 we shall give explicit formulas for the Hamiltonian of the modified vector field. This then shows the existence of a global Hamiltonian for all open $U$, as long as the numerical method can be represented as a B-series (which is the case for Runge-Kutta methods).

## V.2.3 Lie Group Methods

As in Sect. III. 7 we consider differential equations

$$
Y^{\prime}=A(Y) Y, \quad Y(0)=Y_{0}
$$

where $A(Y)$ is in some Lie algebra $\mathfrak{g}$, so that the solution $Y(t)$ lies in the corresponding Lie group $G$. For the Lie group methods of Sect. III. 7 the Taylor series is of the form

$$
\begin{equation*}
Y_{n+1}=\left(I+h A\left(Y_{n}\right)+\mathcal{O}\left(h^{2}\right)\right) Y_{n} \tag{2.9}
\end{equation*}
$$

so that the coefficient functions of the modified equation (1.1) satisfy $f_{j}(Y)=A_{j}(Y) Y$.
Theorem 2.3 Let $G$ be a matrix Lie group and let $\mathfrak{g}$ be its associated Lie algebra. If the method $Y_{n} \mapsto Y_{n+1}$ is of the form (2.9) and maps $G$ into $G$, then the modified equation is of the form

$$
\tilde{Y}^{\prime}=\left(A(\tilde{Y})+h A_{2}(\tilde{Y})+h^{2} A_{3}(\tilde{Y})+\ldots\right) \tilde{Y}, \quad \tilde{Y}(0)=Y_{0}
$$

where $A_{j}(Y) \in \mathfrak{g}$ for all $Y \in G$ and for all $j$.
Proof.

## V. 3 Modified Equation Expressed with Trees

By Theorem II.2.6 the numerical solution $y_{1}=\Phi_{h}\left(y_{0}\right)$ of a Runge-Kutta method can be written as a B-series

$$
\begin{align*}
\Phi_{h}(y)=y & +h f(y)+\frac{h^{2}}{2!} a(\ell)\left(f^{\prime} f\right)(y)  \tag{3.1}\\
& +\frac{h^{3}}{3!}\left(a(\boldsymbol{\bigvee}) f^{\prime \prime}(f, f)(y)+a(\vdots) f^{\prime} f^{\prime} f(y)\right)+\ldots
\end{align*}
$$

For consistent methods, i.e., methods of order at least 1 , we always have $a(\cdot)=1$, so that the coefficient of $h$ is equal to $f(y)$. In this section we exploit this special structure of $\Phi_{h}(y)$ in order to get practical formulas for the coefficient functions of the modified differential equation. Using (3.1) instead of (1.3), the equations (1.4) yield

$$
\begin{align*}
f_{2}(y) & =\frac{1}{2!}(a(\swarrow)-1)\left(f^{\prime} f\right)(y) \\
f_{3}(y) & =\frac{1}{3!}\left(a(\boldsymbol{\zeta})-\frac{3}{2} a(\boldsymbol{\ell})+\frac{1}{2}\right) f^{\prime \prime}(f, f)(y)  \tag{3.2}\\
& +\frac{1}{3!}(a(\vdots)-3 a(\boldsymbol{\ell})+2) f^{\prime} f^{\prime} f(y) .
\end{align*}
$$

Continuing this computation, one is quickly convinced of the general formula

$$
\begin{equation*}
f_{j}(y)=\frac{1}{j!} \sum_{\rho(t)=j} \alpha(t) b(t) F(t)(y) \tag{3.3}
\end{equation*}
$$

so that the modified equation (1.1) becomes

$$
\begin{equation*}
\widetilde{y}^{\prime}=\sum_{t \in T} \frac{h^{\rho(t)-1}}{\rho(t)!} \alpha(t) b(t) F(t)(\widetilde{y}) \tag{3.4}
\end{equation*}
$$

with $b(\cdot)=1, b(\boldsymbol{\ell})=a(\boldsymbol{\ell})-1, \ldots$ Since the coefficients $\rho(t)$ and $\alpha(t)$ are known from Definition II.2.2, all we have to do is to find suitable recursion formulas for the real coefficients $b(t)$. This will be done with the help of the following result.

Lemma 3.1 (Lie-derivative of B-series [Hai99]) Let $b(t)$ (with $b(\emptyset)=0$ ) and $c(t)$ be the coefficients of two B-series, and let $y(x)$ be a formal solution of the differential equation $h y^{\prime}(x)=B(b, y(x))$. The Lie derivative of the function $B(c, y)$ with respect to the vector field $B(b, y)$ is again a $B$-series

$$
h \frac{d}{d x} B(c, y(x))=B\left(\partial_{b} c, y(x)\right) .
$$

Its coefficients are given by $\partial_{b} c(\emptyset)=0$ and for $\rho(t) \geq 1$ by

$$
\begin{equation*}
\partial_{b} c(t)=\sum_{u \circ_{\gamma} v=t}\binom{\rho(t)}{\rho(u)} \frac{l a b\left(\left.t\right|_{\gamma} u\right)}{\alpha(t)} c(u) b(v) \tag{3.5}
\end{equation*}
$$

Here $u \circ_{\gamma} v$ denotes the tree $t$ which is obtained as follows: take the tree $u$, specify one of its vertices, say $\gamma$, and attach the root of the tree $v$ with a new branch to $\gamma$. We call this a splitting of $t$. The sum in (3.5) is over all such splittings ( $\emptyset \circ t:=t$ is also considered as a splitting of $T$ ). The integer lab $\left(\left.t\right|_{\gamma} u\right)$ denotes the number of monotonic labellings of $t=u \circ_{\gamma} v$, such that the first $\rho(u)$ labels are attached to the subtree $u$.


Proof. For the proof of this lemma it is convenient to work with monotonically labelled trees. This means that we attach to every vertex a number between 1 and $\rho(t)$, such that the root is labelled by 1 , and the labels are monotone in every branch. We denote the set of labelled trees by $L T$. By Exercise II.5, a sum $\sum_{t \in T} \alpha(t) \cdot / \cdot$ then becomes $\sum_{t \in L T} \cdot / \cdot$.

For the computation of the Lie derivative of $B(c, y)$ we have to differentiate the elementary differential $F(u)(y(x))$ with respect to $x$. Using Leibniz' rule, this yields $\rho(u)$ terms, one for every vertex of $u$. Then we insert the series $B(b, y(x))$ for $h y^{\prime}(x)$. This means that all the trees $v$ appearing in $B(b, y(x))$ are attached with a new branch to the distinguished vertex. Written out as formulas, this gives

$$
\begin{gathered}
h \frac{d}{d x} B(c, y(x))=\sum_{u \in L T \cup\{\emptyset\}} \frac{h^{\rho(u)}}{\rho(u)!} c(u) \sum_{\gamma} \sum_{v \in L T} \frac{h^{\rho(v)}}{\rho(v)!} b(v) F\left(u \circ_{\gamma} v\right)(y(x)) \\
=\sum_{t \in T} \frac{h^{\rho(t)}}{\rho(t)!} \sum_{u \circ_{\gamma} v=t}\binom{\rho(t)}{\rho(u)} \frac{\operatorname{lab(t|_{\gamma }u)}}{\alpha(t)} c(u) b(v) F(t)(y(x)),
\end{gathered}
$$

where $\sum_{\gamma}$ is a sum over all vertices of $u$. This proves formula (3.5).


Fig. 3.1: Illustration of the formula (3.5)
Let us illustrate this proof and the formula (3.5) with a tree of order 5. All possible splittings $t=u \circ_{\gamma} v$ together with their labellings are given in Fig. 3.1. Observe that $u$ may be the empty tree $\emptyset$ (the 6 trees in the upper row), that always $\rho(v) \geq 1$, and that $\operatorname{lab}\left(\left.t\right|_{\gamma} u\right)$ may be different from $\alpha(u) \alpha(v)$. We see that the tree $t$ is obtained in several ways: (i) differentiation of $F(\emptyset)(y)=y$ and adding $F(t)(y)$ as argument, (ii) differentiation of the factor corresponding to the root in $F(u)(y)=f^{\prime \prime}(f, f)(y)$ and adding $F(\ell)(y)=\left(f^{\prime} f\right)(y)$, (iii) differentiation of all $f^{\prime}$ 's in $F(u)(y)=f^{\prime \prime \prime}(f, f, f)(y)$ and adding $F(\cdot)(y)=f(y)$, and finally, (iv) differentiation of the factor for the root in $F(u)(y)=f^{\prime \prime}\left(f^{\prime} f, f\right)(y)$ and adding $F(\cdot)(y)=f(y)$. This proves that

$$
\partial_{b} c(\boldsymbol{\bigvee})=c(\emptyset) b(\boldsymbol{\bigvee})+\frac{5}{3} c(\boldsymbol{\bigvee}) b(\boldsymbol{\ell})+\frac{5}{2} c(\boldsymbol{V}) b(\cdot)+\frac{5}{2} c(\boldsymbol{\bigvee}) b(\cdot) .
$$

For the trees up to order 3 the formulas for $\partial_{b} c$ are:

$$
\begin{aligned}
\partial_{b} c(\cdot) & =c(\emptyset) b(\cdot) \\
\partial_{b} c(\boldsymbol{\ell}) & =c(\emptyset) b(\boldsymbol{\ell})+2 c(\cdot) b(\cdot) \\
\partial_{b} c(\boldsymbol{\bigvee}) & =c(\emptyset) b(\boldsymbol{\bigvee})+3 c(\boldsymbol{\ell}) b(\cdot) \\
\partial_{b} c(\boldsymbol{\ell}) & =c(\emptyset) b(\boldsymbol{\ell})+3 c(\cdot) b(\boldsymbol{\ell})+3 c(\boldsymbol{\ell}) b(\cdot) .
\end{aligned}
$$

The above lemma permits us to get recursion formulas for the coefficients $b(t)$ in (3.4).
Theorem 3.2 If the method $\Phi_{h}(y)$ is given by (3.1), the functions $f_{j}(y)$ of the modified differential equation (1.1) satisfy (3.3), where the real coefficients $b(t)$ are recursively defined by $b(\emptyset)=0, b(\cdot)=1$ and

$$
\begin{equation*}
b(t)=a(t)-\sum_{j=2}^{\rho(t)} \frac{1}{j!} \partial_{b}^{j-1} b(t) . \tag{3.6}
\end{equation*}
$$

Here, $\partial_{b}^{j-1}$ is the $(j-1)$-th iterate of the Lie derivative $\partial_{b}$ defined in Lemma 3.1.
Proof. The right-hand side of the modified equation (3.4) is the B-series $B(b, \widetilde{y}(x))$ divided by $h$. It therefore follows from an iterative application of Lemma 3.1 that

$$
h^{j} \widetilde{y}^{(j)}(x)=B\left(\partial_{b}^{j-1} b, \widetilde{y}(x)\right),
$$

so that by Taylor series expansion $\widetilde{y}(x+h)=y+B\left(\sum_{j \geq 1} \frac{1}{j!} \partial_{b}^{j-1} b, y\right)$, where $y:=\widetilde{y}(x)$. Since we have to determine the coefficients $b(t)$ such that $\widetilde{y}(x+h)=\Phi_{h}(y)=B(a, y)$, a comparison of the two B-series gives $\sum_{j \geq 1} \frac{1}{j!} \partial_{b}^{j-1} b(t)=a(t)$. This proves the statement, because $\partial_{b}^{0} b(t)=b(t)=0$ for $t \in T$, and $\partial_{b}^{j-1} b(t)=0$ for $j>\rho(t)$ (as a consequence of $b(\emptyset)=0)$.

Recurrence formulas for the coefficients $b(t)$ have first been given in [Hai94]. The original derivation does not use the Lie-derivative of B-series. A different recursion formula is given in [CMS94]. We present in Table 3.1 the formula (3.6) for trees up to order 3.

Table 3.1: Examples of formula (3.6)

| $t=$ • | $b(\cdot)=a(\cdot)$ | $b(\cdot)=a(\cdot)$ |
| :---: | :---: | :---: |
| $t=$ ! | $b(\boldsymbol{\ell})=a(\boldsymbol{\ell})-b(\cdot)^{2}$ | $b(\boldsymbol{\ell})=a(\boldsymbol{\ell})-a(\cdot)^{2}$ |
| $t=\gamma$ | $b(\boldsymbol{Y})=a(\boldsymbol{Y})-\frac{3}{2} b(\cdot) b(\cdot)-b(\cdot)^{3}$ | $b(\boldsymbol{Y})=a(\boldsymbol{Y})-\frac{3}{2} a(\boldsymbol{\ell}) a(\cdot)+\frac{1}{2} a(\cdot)^{3}$ |
| $t=$ ? |  |  |

## V. 4 Modified Hamiltonian

As an illustration use the CBH formula for the computation of the modified Hamiltonian for the symplectic Euler method and the Verlet scheme.

Introduce elementary Hamiltonians, etc.

## V. 5 Rigorous Estimates - Local Error

Up to now we have considered the modified equation (1.1) as a formal series without taking care about convergence issues. Here,

- we show that already in very simple situations the modified differential equation does not converge;
- we give bounds on the coefficient functions $f_{j}(y)$ of the modified equation (1.1), so that an optimal truncation index can be determined;
- we estimate the difference between the numerical solution $y_{1}=\Phi_{h}\left(y_{0}\right)$ and the exact solution $\widetilde{y}(h)$ of the truncated modified equation.

These estimates will be the basis for rigorous statements concerning the long-time behaviour of numerical solutions. The rigorous estimates of the present section have been obtained in the articles [BG94], [HaL97] and [Rei98]. We start our presentation with the approach of [HaL97], and we switch over to that of [BG94] in Subsection V.5.4.

Example 5.1 We consider the differential equation ${ }^{1} y^{\prime}=f(x), y(0)=0$, and we apply the trapezoidal rule $y_{1}=h(f(0)+f(h)) / 2$. In this case, the numerical solution has an expansion $\Phi_{h}(x, y)=y+h(f(x)+f(x+h)) / 2=y+h f(x)+h^{2} f^{\prime}(x) / 2+h^{3} f^{\prime \prime}(x) / 4+\ldots$, so that the modified equation is necessarily of the form

$$
\begin{equation*}
\widetilde{y}^{\prime}=f(x)+h b_{1} f^{\prime}(x)+h^{2} b_{2} f^{\prime \prime}(x)+h^{3} b_{3} f^{\prime \prime \prime}(x)+\ldots . \tag{5.1}
\end{equation*}
$$

[^15]The real coefficients $b_{k}$ can be computed by putting $f(x)=e^{x}$. The relation $\Phi_{h}(x, y)=$ $\widetilde{y}(x+h)$ (with initial value $\widetilde{y}(x)=y$ ) yields after division by $e^{x}$

$$
\frac{h}{2}\left(e^{h}+1\right)=\left(1+b_{1} h+b_{2} h^{2}+b_{3} h^{3} \ldots\right)\left(e^{h}-1\right) .
$$

This proves that $b_{1}=0$, and $b_{k}=B_{k} / k$ !, where $B_{k}$ are the Bernoulli numbers (see for example [HW97, Sect.II.10]). Since these numbers behave like $B_{k} / k!\approx$ Const $\cdot(2 \pi)^{-k}$ for $k \rightarrow \infty$, the series (5.1) diverges for all $h \neq 0$, as soon as the derivatives of $f(x)$ grow like $f^{(k)}(x) \approx k!M R^{-k}$. This is typically the case for analytic functions $f(x)$ with finite poles.

It is interesting to remark that the relation $\Phi_{h}(x, y)=\widetilde{y}(x+h)$ is nothing other than the Euler-MacLaurin summation formula.

## V.5.1 The Analyticity Assumption

The coefficient functions of the modified differential equation (1.1) depend on $f(y)$ and its derivatives. In order to be able to bound $f_{j}(y)$, we assume that in some fixed norm

$$
\begin{equation*}
\left\|f^{(k)}(y)\right\| \leq k!M R^{-k} \quad \text { for } \quad k=0,1,2, \ldots . \tag{5.2}
\end{equation*}
$$

This is the fundamental assumption throughout this chapter. It is motivated by Cauchy's integral formula for analytic functions. Indeed, if we assume that $f(y)$ is analytic in a neighbourhood of a polydisc $\left\{\left(z_{1}, \ldots, z_{d}\right) \in \mathbb{C}^{d} ;\left|z_{j}-y_{j}\right| \leq R\right.$ for $\left.j=1, \ldots, d\right\}$, then

$$
f\left(y_{1}, \ldots, y_{d}\right)=\frac{1}{(2 \pi i)^{d}} \int_{\gamma_{1}} \ldots \int_{\gamma_{d}} \frac{f\left(z_{1}, \ldots, z_{d}\right)}{\left(z_{1}-y_{1}\right) \cdot \ldots \cdot\left(z_{d}-y_{d}\right)} d z_{1} \ldots d z_{d}
$$

holds, where $\gamma_{j}$ denotes the circle of radius $R$ around $y_{j}$. Differentiation yields the estimate

$$
\left\|\frac{\partial^{k_{1}+\ldots+k_{d}} f(y)}{\partial y_{1}^{k_{1}} \ldots \partial y_{d}^{k_{d}}}\right\| \leq k_{1}!\cdot \ldots \cdot k_{d}!\cdot M \cdot R^{-\left(k_{1}+\ldots+k_{d}\right)}
$$

where $M$ is an upper bound of $f$ on the polydisc. The $k$ th derivative, considered as a $k$-linear mapping, is defined by $f^{(k)}(y)(u, v, \ldots)=\sum_{i, j, \ldots} \frac{\partial^{k} f(y)}{\partial y_{i} \partial y_{j} \ldots} \cdot u_{i} \cdot v_{j} \cdot \ldots$, and we obtain from $k_{1}!\cdot \ldots \cdot k_{d}!\leq\left(k_{1}+\ldots+k_{d}\right)$ ! that

$$
\left\|f^{(k)}(y)(u, v, \ldots)\right\| \leq k!M R^{-k} \sum_{i, j, \ldots}\left|u_{i}\right| \cdot\left|v_{j}\right| \cdot \ldots \leq k!M R^{-k} \cdot\|u\|_{1} \cdot\|v\|_{1} \cdot \ldots
$$

This proves (5.2) for the $\ell_{1}$ operator norm. Since all norms are equivalent in $\mathbb{R}^{d}$, we get (5.2) for a given norm by suitably changing the constants $M$ and $R$.

## V.5.2 Estimation of the Derivatives of the Numerical Solution

Let us first estimate the growth of the functions $d_{j}(y)$ in the Taylor expansion (1.3) of the numerical method $\Phi_{h}(y)$. We shall do this for Runge-Kutta methods, and for all methods that can be written as a B-series (see Definition II.2.3), i.e.,

$$
\begin{equation*}
\Phi_{h}(y)=y+h f(y)+h^{2} d_{2}(y)+h^{3} d_{3}(y)+\ldots \tag{5.3}
\end{equation*}
$$

where

$$
\begin{equation*}
d_{j}(y)=\frac{1}{j!} \sum_{t \in T, \rho(t)=j} \alpha(t) a(t) F(t)(y) \tag{5.4}
\end{equation*}
$$

For Runge-Kutta methods, the coefficients $a(t)$ are given in Theorem II.2.6, and it follows from (II.2.4) and (II.2.6) that

$$
\begin{equation*}
|a(t)| \leq \gamma(t) \mu \kappa^{\rho(t)-1} \tag{5.5}
\end{equation*}
$$

where $\gamma(t)$ is the integer defined in (II.2.5), and the constants $\mu, \kappa$ can be computed by

$$
\begin{equation*}
\mu=\sum_{i=1}^{s}\left|b_{i}\right|, \quad \kappa=\max _{i=1, \ldots, s} \sum_{j=1}^{s}\left|a_{i j}\right| \tag{5.6}
\end{equation*}
$$

from the coefficients of the Runge-Kutta method. Due to the consistency condition $\sum_{i=1}^{s} b_{i}=1$, methods with positive weights $b_{i}$ all satisfy $\mu=1$. The values $\mu, \kappa$ of some classes of Runge-Kutta methods are given in Table 5.1 (those for the Gauss methods and for the Lobatto IIIA methods have been checked for $s \leq 9$ and $s \leq 5$, respectively).

Table 5.1: The constants $\mu$ and $\kappa$ of formula (5.6)

| method | $\mu$ | $\kappa$ | method | $\mu$ | $\kappa$ |
| :--- | :---: | :---: | :--- | :---: | :---: |
| explicit Euler | 1 | 0 | implicit Euler | 1 | 1 |
| implicit midpoint | 1 | $1 / 2$ | trapezoidal rule | 1 | 1 |
| Gauss methods | 1 | $c_{s}$ | Lobatto IIIA | 1 | 1 |

Theorem 5.2 Under the conditions (5.2) and (5.5), the coefficients $d_{j}(y)$ of the numerical solution of a B-series method (5.3) are bounded by

$$
\left\|d_{j}(y)\right\| \leq \mu M\left(\frac{4 \kappa M}{R}\right)^{j-1}
$$

Proof. Due to the special structure of $d_{j}(y)$ we have

$$
\left\|d_{j}(y)\right\| \leq \frac{1}{j!} \sum_{\rho(t)=j} \alpha(t)|a(t)|\left\|F(t)\left(y_{0}\right)\right\| .
$$

Inserting the estimate (5.5), we see that it is sufficient to consider the case $\mu=\kappa=1$, i.e., the implicit Euler method, for which $a(t)=\gamma(t)$ for all trees $t$.

The main idea (method of majorants) is to consider the scalar differential equation

$$
\begin{equation*}
z^{\prime}=g(z) \quad \text { with } \quad g(z)=\frac{M}{1-z / R} \tag{5.7}
\end{equation*}
$$

with initial value $z(0)=0$. For the derivatives $g^{(k)}(0)$ of this function we have equality in (5.2) for all $k \geq 0$. Consequently,

$$
\sum_{\rho(t)=j} \alpha(t) \gamma(t)\|F(t)(y)\| \leq \sum_{\rho(t)=j} \alpha(t) \gamma(t) G(t)(0)=\frac{d^{j} z_{1}}{d h^{j}}(0)
$$

where $G(t)\left(z_{0}\right)$ are the elementary differentials corresponding to (5.7) and

$$
z_{1}=\frac{h M}{1-z_{1} / R}, \quad \text { or equivalently, } \quad z_{1}=\frac{R}{2}\left(1-\sqrt{1-\frac{4 M h}{R}}\right)
$$

is the numerical solution of the implicit Euler method applied to (5.7). From the series expansion $1-\sqrt{1-x}=\sum_{j \geq 1}(-1)^{j-1}\binom{1 / 2}{j} x^{j}$ and from $(-1)^{j-1}\binom{1 / 2}{j}=\left|\binom{1 / 2}{j}\right| \leq 1 / 2$, it follows that

$$
\frac{1}{j!} \frac{d^{j} z_{1}}{d h^{j}}(0)=(-1)^{j-1} \frac{R}{2}\binom{1 / 2}{j}\left(\frac{4 M}{R}\right)^{j} \leq M\left(\frac{4 M}{R}\right)^{j-1}
$$

which proves the statement. A slightly better estimate could be obtained with the help of Stirling's formula.

## V.5.3 Estimation of the Coefficients of the Modified Equation

For B-series methods (like Runge-Kutta methods), the coefficient functions $f_{j}(y)$ of the modified differential equation (1.1) are given by (see Sect. V.3)

$$
\begin{equation*}
f_{j}(y)=\frac{1}{j!} \sum_{\rho(t)=j} \alpha(t) b(t) F(t)(y) \tag{5.8}
\end{equation*}
$$

For a recursive definition of the coefficients $b(t)$ see Theorem 3.2. We first derive bounds for the coefficients $b(t)$, and then we estimate the functions $f_{j}(y)$ in a similar way as in the proof of Theorem 5.2.

Lemma 5.3 Suppose that

$$
\begin{equation*}
|a(t)| \leq \mu \cdot \kappa^{\rho(t)-1} \cdot \rho(t)!\quad \text { for all trees } t \tag{5.9}
\end{equation*}
$$

(due to $\gamma(t) \leq \rho(t)$ ! this is a consequence of (5.5)). Then, the coefficients $b(t)$ defined in Eq. (3.6) can be estimated by

$$
\begin{equation*}
|b(t)| \leq \ln 2 \cdot \nu^{\rho(t)} \cdot \rho(t)!\quad \text { with } \quad \nu=\kappa+\mu /(2 \ln 2-1) . \tag{5.10}
\end{equation*}
$$

Proof. We search for a generating function

$$
\begin{equation*}
b(\zeta)=b_{1} \zeta+b_{2} \zeta^{2}+b_{3} \zeta^{3}+\ldots \tag{5.11}
\end{equation*}
$$

whose coefficients $b_{\rho}$ majorize $b(t) / \rho(t)$ ! for $\rho(t)=\rho$.
Let the coefficients of another B-series satisfy $|c(t)| / \rho(t)!\leq c_{\rho}$ for $\rho(t)=\rho$, and put $c(\zeta)=c_{0}+c_{1} \zeta+c_{2} \zeta^{2}+\ldots$. Then the coefficients of the Lie derivative $\left|\partial_{b} c(t)\right|$ are majorized by $\rho(t)!d_{\rho(t)}$, where $d_{\rho}(\rho \geq 0)$ are the coefficients of the series $d(\zeta)=c(\zeta) \cdot b(\zeta)$. This follows from (3.5) and from the fact that for every pair ( $\rho_{1}, \rho_{2}$ ) satisfying $\rho_{1}+\rho_{2}=\rho(t)$, the sum over splittings $t=u \circ_{\gamma} v$ (with $\rho(u)=\rho_{1}$ and $\rho(v)=\rho_{2}$ ) of $l a b\left(\left.t\right|_{\gamma} u\right)$ is at most $\alpha(t)$. Consequently, the values $\left|\partial_{b}^{j-1} b(t)\right| / \rho(t)$ ! are majorized by the coefficients of $b(\zeta)^{j}$.

After this preparation, we let the series $b(\zeta)$ be implicitly defined by

$$
\begin{equation*}
b=\frac{\mu \zeta}{1-\kappa \zeta}+\sum_{j \geq 2} \frac{1}{j!} b^{j}=\frac{\mu \zeta}{1-\kappa \zeta}+e^{b}-1-b \tag{5.12}
\end{equation*}
$$



FIg. 5.1: Complex functions of the proof of Lemma $5.3(\mu=\kappa=1)$

Using the recursive definition (3.6) and the estimate (5.9), it follows by induction on $\rho$ that $|b(t)| / \rho(t)$ ! is bounded by the $\rho$ th coefficient of the series $b(\zeta)$ defined by (5.12).

Whenever $e^{b} \neq 2$ (i.e., for $\zeta \neq(2 b-1) /(\mu+\kappa(2 b-1))$ with $\left.b=\ln 2+2 k \pi i\right)$ the implicit function theorem can be applied to (5.12). This implies that $b(\zeta)$ is analytic in a disc with radius $1 / \nu=(2 \ln 2-1) /(\mu+\kappa(2 \ln 2-1))$ and centre at the origin. On the disc $|\zeta| \leq 1 / \nu$, the solution $b(\zeta)$ of $(5.12)$ with $b(0)=0$ is bounded by $\ln 2$. This is seen as follows (Fig. 5.1): with the function $w=-\mu \zeta /(1-\kappa \zeta)$ the disc $|\zeta| \leq 1 / \nu$ is mapped into a disc which, for all possible choices of $\mu \geq 0$ and $\kappa \geq 0$, lies in $|w| \leq 2 \ln 2-1$. The image of this disc under the mapping $b(w)$ defined by $e^{b}-1-2 b=w$ and $b(0)=0$ is completely contained in the disc $|b| \leq \ln 2$. Cauchy's inequalities therefore imply $\left|b_{j}\right| \leq \ln 2 \cdot \nu^{j}$, and the estimate (5.10) is a consequence of $|b(t)| / \rho(t)!\leq b_{\rho(t)}$.

Using the analyticity assumption on $f$ and the estimates of Lemma 5.3, we are in a position to bound the function $f_{j}$ of (5.8).
Theorem 5.4 If $f(y)$ satisfies (5.2) and if the coefficients $b(t)$ of (5.8) satisfy (5.10), then we have for $j \geq 2$ the estimate

$$
\begin{equation*}
\left\|f_{j}(y)\right\| \leq 0.5 \nu M\left(\frac{2 \nu M j}{e R}\right)^{j-1} \tag{5.13}
\end{equation*}
$$

Proof. The important observation is that

$$
\begin{equation*}
\sum_{\rho(t)=j} \alpha(t)\left\|F(t)\left(y_{0}\right)\right\| \leq \sum_{\rho(t)=j} \alpha(t) G(t)(0)=z^{(j)}(0) \tag{5.14}
\end{equation*}
$$

where, as in the poof of Theorem 5.2, $G(t)(z)$ are the elementary differentials corresponding to (5.7), and $z(x)$ is the solution of (5.7) with initial value $z(0)=0$. This solution is given by
$z(x)=R\left(1-\sqrt{1-\frac{2 M x}{R}}\right)=R \sum_{j \geq 1}(-1)^{j-1}\binom{1 / 2}{j}\left(\frac{2 M x}{R}\right)^{j}=\sum_{j \geq 1} \frac{R}{(2 j-1)}\binom{2 j}{j}\left(\frac{M x}{2 R}\right)^{j}$.
and we obtain for $j \geq 2$ that

$$
z^{(j)}(0)=\frac{R}{(2 j-1)} \cdot \frac{(2 j)!}{j!} \cdot\left(\frac{M}{2 R}\right)^{j} \leq 0.71 M\left(\frac{2 M j}{e R}\right)^{j-1}
$$

where Stirling's formula $\sqrt{2 \pi j} \cdot j^{j} \leq j!\cdot e^{j} \leq \sqrt{2 \pi j} \cdot j^{j} \cdot e^{1 / 12 j}$ (see [HW97, Theorem II.10.4]) has been used. The statement (5.13) is therefore a consequence of (5.8), (5.10) and (5.14).

## V.5.4 Choice of $N$ and the Estimation of the Local Error

In order to get rigorous estimates, we truncate the modified differential equation (1.1), and we consider

$$
\begin{equation*}
\widetilde{y}^{\prime}=F_{N}(\widetilde{y}), \quad F_{N}(\widetilde{y})=f(\widetilde{y})+h f_{2}(\widetilde{y})+\ldots+h^{N-1} f_{N}(\widetilde{y}) \tag{5.15}
\end{equation*}
$$


with initial value $\widetilde{y}(0)=y_{0}$. It is common in the theory of asymptotic expansions to truncate the series at the index where the corresponding term is minimal. Motivated by the bound (5.13) and by the fact that $(\varepsilon x)^{x}$ admits a minimum for $x=(\varepsilon e)^{-1}$ (see the picture to the left with $\varepsilon=0.15$ ), we suppose that the truncation index $N$ satisfies

$$
\begin{equation*}
h N \leq \frac{R}{2 \nu M} \tag{5.16}
\end{equation*}
$$

Under this assumption, (5.13) and (5.2) imply that

$$
\begin{align*}
\left\|F_{N}(y)\right\| & \leq M\left(1+0.5 \nu \sum_{j=2}^{N}\left(\frac{2 \nu M j h}{e R}\right)^{j-1}\right)  \tag{5.17}\\
& \leq M\left(1+0.5 \nu \sum_{j=2}^{N}\left(\frac{j}{e N}\right)^{j-1}\right) \leq M(1+0.5 \nu)
\end{align*}
$$

where we have used $j /(e N) \leq 1 / 2$ for the last inequality. We are now able to prove the main result of this section.

Theorem 5.5 Let $f(y)$ be analytic in $B_{\alpha R}\left(y_{0}\right)=\left\{y \in \mathbb{C}^{n} ;\left\|y-y_{0}\right\| \leq \alpha R\right\}$ for some $\alpha \leq 1 / 2$, and let (5.2) be satisfied for all $y \in B_{\alpha R}\left(y_{0}\right)$. For the numerical method we assume that it can be represented as a B-series with coefficients bounded by (5.5). If $h \leq \alpha h^{*}$ with $h^{*}=R /(4 e \nu M)$, then there exists $N=N(h)$ (namely $N$ equal to the largest integer satisfying $h N \leq h^{*}$ ) such that the difference between the numerical solution $y_{1}=\Phi_{h}\left(y_{0}\right)$ and the exact solution $\varphi_{N, t}\left(y_{0}\right)$ of the truncated modified equation (5.15) satisfies

$$
\left\|\Phi_{h}\left(y_{0}\right)-\varphi_{N, h}\left(y_{0}\right)\right\| \leq h \gamma M e^{-h^{*} / h}
$$

where $\gamma=e(2+0.5 \nu+\mu)$ depends only on the method (we have $\nu \leq 3.6$ and $\gamma \leq 13.1$ for the methods of Table 5.1).

Proof. We follow here the elegant proof of [BG94]. It is based on the fact that $\Phi_{h}\left(y_{0}\right)$ (as a convergent B-series) and $\varphi_{N, h}\left(y_{0}\right)$ (as the solution of an analytic differential equation) are both analytic functions of $h$. Hence,

$$
\begin{equation*}
g(h):=\Phi_{h}\left(y_{0}\right)-\varphi_{N, h}\left(y_{0}\right) \tag{5.18}
\end{equation*}
$$

is analytic in a complex neighbourhood of $h=0$. By definition of the functions $f_{j}(y)$ of the modified equation (1.1), the coefficients of the Taylor series for $\Phi_{h}\left(y_{0}\right)$ and $\varphi_{N, h}\left(y_{0}\right)$ are the same up to the $h^{N}$ term, but not further due to the truncation of the modified
equation. Consequently, the function $g(h)$ contains the factor $h^{N+1}$, and the maximum principle for analytic functions, applied to $g(h) / h^{N+1}$, implies that

$$
\begin{equation*}
\|g(h)\| \leq\left(\frac{h}{\varepsilon}\right)^{N+1} \max _{|z| \leq \varepsilon}\|g(z)\| \quad \text { for } \quad 0 \leq h \leq \varepsilon \tag{5.19}
\end{equation*}
$$

if $g(z)$ is analytic for $|z| \leq \varepsilon$. We shall show that we can take $\varepsilon=e h^{*} / N$, and we compute an upper bound for $\|g(z)\|$ by estimating separately $\left\|\Phi_{h}\left(y_{0}\right)-y_{0}\right\|$ and $\left\|\varphi_{N, h}\left(y_{0}\right)-y_{0}\right\|$.

The function $\Phi_{z}\left(y_{0}\right)$ is given by the series (5.3) which, due to the bounds of Theorem 5.2, converges at least for $|z|<R /(4 \kappa M)$, and therefore also for $|z| \leq \varepsilon$ (because $\kappa<\nu)$. Hence, it is analytic in $|z| \leq \varepsilon$. Furthermore, because of $4 \kappa M|z| / R \leq 1 / 2$ for $|z| \leq \varepsilon$ (due to $h \leq \alpha h^{*}$ and $\alpha \leq 1 / 2$ which implies $N \geq 2$ ), it follows from Theorem 5.2 that $\left\|\Phi_{z}\left(y_{0}\right)-y_{0}\right\| \leq|z| M(1+\mu) \leq \varepsilon M(1+\mu)$.

Because of the bound (5.17) on $F_{N}(y)$, which is valid for $y \in B_{\alpha R}\left(y_{0}\right)$, we have $\left\|\varphi_{N, z}\left(y_{0}\right)-y_{0}\right\| \leq|z| M(1+0.5 \nu)$ as long as the solution $\varphi_{N, z}\left(y_{0}\right)$ stays in the ball $B_{\alpha R}\left(y_{0}\right)$. Because of $\varepsilon M(1+0.5 \nu) \leq \alpha R$, which is a consequence of the definition of $\varepsilon$ and of $(1+0.5 \nu) \leq 2 \nu$ (because for consistent methods $\mu \geq 1$ holds and therefore also $\nu \geq 1$ ), this is the case for all $|z| \leq \varepsilon$. In particular, the solution $\varphi_{N, z}\left(y_{0}\right)$ is analytic in $|z| \leq \varepsilon$.

Inserting $\varepsilon=e h^{*} / N$ and the bound on $\|g(z)\| \leq\left\|\Phi_{z}\left(y_{0}\right)-y_{0}\right\|+\left\|\varphi_{N, z}\left(y_{0}\right)-y_{0}\right\|$ into (5.19) yields (with $C=2+0.5 \nu+\mu$ )

$$
\|g(h)\| \leq \varepsilon M C\left(\frac{h}{\varepsilon}\right)^{N+1} \leq h M C\left(\frac{h}{\varepsilon}\right)^{N}=h M C\left(\frac{h N}{e h^{*}}\right)^{N} \leq h M C e^{-N}
$$

because $h N \leq h^{*}$. The statement now follows from the fact that $N \leq h^{*} / h<N+1$, so that $e^{-N} \leq e \cdot e^{-h^{*} / h}$.

Remark 5.6 The quotient $L=M / R$ is an upper bound of the first derivative $f^{\prime}(y)$ and can be interpreted as a Lipschitz constant for $f(y)$. The condition $h \leq \alpha h^{*}$ is therefore equivalent to $h L \leq$ Const, where Const depends only on the method. Such a condition is natural and familiar from the study of convergence for nonstiff differential equations.

## V.5.5 Estimates for Methods that are not B-series

Proof of Benettin \& Giorgilli or of Reich.

## V. 6 Longtime Error Estimates

As a first application of Theorem 5.5 we study the long-time energy conservation of symplectic numerical schemes applied to Hamiltonian systems $y^{\prime}=J^{-1} \nabla H(y)$. It follows from Theorem 2.2 that the corresponding modified differential equation is also Hamiltonian. After truncation we thus get a modified Hamiltonian

$$
\begin{equation*}
\widetilde{H}(y)=H(y)+h^{p} H_{p+1}(y)+\ldots+h^{N-1} H_{N}(y) \tag{6.1}
\end{equation*}
$$

Theorem 6.1 Consider a Hamiltonian system with analytic $H: \mathbb{R}^{2 d} \rightarrow \mathbb{R}$, apply a symplectic numerical method $\Phi_{h}(y)$ with step size $h$, and assume that the numerical solution stays in the compact set $K \subset \mathbb{R}^{2 d}$. If the vector field $f(y)=J^{-1} \nabla H(y)$ satisfies (5.2) on the set

$$
K_{\alpha R}=\left\{y \in \mathbb{C}^{2 d} ;\left\|y-y_{0}\right\| \leq \alpha R \text { for some } y_{0} \in K\right\}
$$

with $\alpha=h / h^{*}$, where $h^{*}$ is as in Theorem 5.5, then there exists $N=N(h)$ (as in Theorem 5.5), such that

$$
\begin{aligned}
\widetilde{H}\left(y_{n}\right) & =\widetilde{H}\left(y_{0}\right)+\mathcal{O}\left(e^{-h^{*} / 2 h}\right) \quad \text { for } n h \leq T \\
H\left(y_{n}\right) & =H\left(y_{0}\right)+\mathcal{O}\left(h^{p}\right) \quad \text { for } \quad n h \leq T
\end{aligned}
$$

on exponentially long time intervals $T=e^{h^{*} / 2 h}$.
Proof. We let $\varphi_{N, t}\left(y_{0}\right)$ be the flow of the truncated modified equation. Since this differential equation is Hamiltonian with $\widetilde{H}$ of $(6.1), \widetilde{H}\left(\varphi_{N, t}\left(y_{0}\right)\right)=\widetilde{H}\left(y_{0}\right)$ holds for all times $t$. From Theorem 5.5 we know that $\left\|y_{n+1}-\varphi_{N, h}\left(y_{n}\right)\right\| \leq h \gamma M e^{-h^{*} / h}$ and, by using a global $h$-independent Lipschitz constant for $\widetilde{H}$ (which exists by Theorem 5.4), we also get $\widetilde{H}\left(y_{n+1}\right)-\widetilde{H}\left(\varphi_{N, h}\left(y_{n}\right)\right)=\mathcal{O}\left(h e^{-h^{*} / h}\right)$. From the identity

$$
\widetilde{H}\left(y_{n}\right)-\widetilde{H}\left(y_{0}\right)=\sum_{j=1}^{n}\left(\widetilde{H}\left(y_{j}\right)-\widetilde{H}\left(y_{j-1}\right)\right)=\sum_{j=1}^{n}\left(\widetilde{H}\left(y_{j}\right)-\widetilde{H}\left(\varphi_{N, h}\left(y_{j-1}\right)\right)\right)
$$

we thus get $\widetilde{H}\left(y_{n}\right)-\widetilde{H}\left(y_{0}\right)=\mathcal{O}\left(n h e^{-h^{*} / h}\right)$, and the statement on the long-time conservation of $\widetilde{H}$ is an immediate consequence. The statement for the Hamiltonian $H$ follows from (6.1), because $H_{p+1}(y)+h H_{p+2}(y)+\ldots+h^{N-p-1} H_{N}(y)$ is uniformly bounded on $K$ independently of $h$ and $N$. This follows from the proof of Lemma IV.2.7 and from the estimates of Theorem 5.4.

Example 6.2 Let us check explicitly the assumptions of Theorem 6.1 for the pendulum problem $\dot{q}=p, \dot{p}=-\sin q$. The vector field $f(p, q)=(p,-\sin q)^{T}$ is also well-defined for complex $p$ and $q$, and it is analytic everywhere on $\mathbb{C}^{2}$. We let $K$ be a compact subset of $\left\{(p, q) \in \mathbb{R}^{2} ;|p| \leq c\right\}$. As a consequence of $|\sin q| \leq e^{|\Im q|}$, we get the bounds

$$
\begin{aligned}
\|f(p, q)\| & \leq \max \left(c+\alpha R, e^{\alpha R}\right) \\
\left\|f^{(k)}(p, q)\right\| & \leq e^{\alpha R} \quad \text { for } k=1,2, \ldots
\end{aligned}
$$

for $(p, q) \in K_{\alpha R}$. Hence, we can choose $R=1$ and $M=\max \left(C+\alpha, e^{\alpha}\right)$ or $R=2$ and $M=\max \left(C+2 \alpha, 2 e^{2 \alpha}\right)$ in order to satisfy (5.2) for all $(p, q) \in K_{\alpha R}$. For $c \leq 2$ and for small $\alpha$ we get $M / R \approx 1$, so that $h^{*}$ of Theorem 5.5 is given by $h^{*} \approx 1 / 4 e \nu \approx 0.025$ for the methods of Table 5.1. For step sizes that are smaller than $h^{*} / 20$, Theorem 6.1 guarantees that the numerical Hamiltonian is well conserved on intervals $[0, T]$ with $T \approx e^{10} \approx 2 \cdot 10^{4}$.

The numerical experiment of Fig. 6.1 shows that the estimates for $h^{*}$, although qualitatively correct, are often too pessimistic. We have drawn 200000 steps of the numerical solution of the implicit midpoint rule for various step sizes $h$ and for initial values $\left(p_{0}, q_{0}\right)=$


Fig. 6.1: Numerical solutions of the implicit midpoint rule with large step sizes
$(0,-1.5),\left(p_{0}, q_{0}\right)=(0,-2.5),\left(p_{0}, q_{0}\right)=(1.5,-\pi)$, and $\left(p_{0}, q_{0}\right)=(2.5,-\pi)$. They are compared to the contour lines of the truncated modified Hamiltonian

$$
\widetilde{H}(p, q)=\frac{p^{2}}{2}-\cos q+\frac{h^{2}}{48}\left(\cos (2 q)-2 p^{2} \cos q\right) .
$$

This shows that for step sizes as large as $h \leq 0.7$ the Hamiltonian $\widetilde{H}$ is extremely well conserved. Beyond this value, the dynamics of the numerical method soon turns into chaotic behaviour (see also [Yo93] and [HNW93, page 336]).

Theorem 6.1 explains the nearly conservation of the Hamiltonian with symplectic integration methods as observed in Fig. I.1.2 for the pendulum problem, in Fig. I.2.2 for the Kepler problem, and in Fig. I.3.1 for the frozen argon crystal.

The linear drift of the numerical Hamiltonian for non-symplectic methods can be explained by a computation similar to that of the proof of Theorem 6.1. From a Lipschitz condition of the Hamiltonian and from the standard local error estimate, we obtain $H\left(y_{n+1}\right)-H\left(\varphi_{h}\left(y_{n}\right)\right)=\mathcal{O}\left(h^{p+1}\right)$. Since $H\left(\varphi_{h}\left(y_{n}\right)\right)=H\left(y_{n}\right)$, a summation of these terms leads to

$$
\begin{equation*}
H\left(y_{n}\right)-H\left(y_{0}\right)=\mathcal{O}\left(t h^{p}\right) \quad \text { for } \quad t=n h . \tag{6.2}
\end{equation*}
$$

This explains the linear growth in the error of the Hamiltonian observed in Fig. I.2.2 and in Fig. I.3.1 for the explicit Euler method.

More results from the life-span paper.

## V. 7 Variable Step Sizes

## V. 8 Exercises

1. Change the Maple program of Example 1.1 in such a way that the modified equations for the implicit Euler method, the implicit midpoint rule, or the trapezoidal rule are obtained. Observe that for symmetric methods one gets expansions in even powers of $h$.
2. Write a short Maple program which, for simple methods such as the symplectic Euler method, computes some terms of the modified equation for a two-dimensional system $p^{\prime}=f(p, q), q^{\prime}=g(p, q)$. Check the modified equations of Example 1.2.
3. Find a first integral of the truncated modified equation for the symplectic Euler method and the Volterra-Lotka problem (Example 1.2).
Hint. With the transformation $p=\exp P, q=\exp Q$ you will get a Hamiltonian system. Result. $\widetilde{I}(p, q)=I(p, q)-h\left((p+q)^{2}-8 p-10 q+2 \ln p+8 \ln q\right) / 4$.
4. Compute $\partial_{b} c(t)$ for the tree $t=[[\tau], \tau]$ of order 4 .
5. For the implicit mid-point rule compute the coefficients $a(t)$ of the expansion (3.1), and also a few coefficients $b(t)$ of the modified equation.
Result. $a(t)=\gamma(t) 2^{1-\rho(t)}, b(\cdot)=1, b(!)=0, b(t)=a(t)-1$ for $\rho(t)=3$.
6. Check the formulas of Table 3.1.
7. For the linear transformation $\rho(p, q)=(-p, q)$, consider a $\rho$-reversible problem (2.6) with scalar $p$ and $q$. Prove that every solution which crosses the $q$-axis twice is periodic.
8. Consider a differential equation $y^{\prime}=f(y)$ with a first integral $I(y)$, and assume that the numerical integrator $\Phi_{h}(y)$ preserves this invariant exactly. Prove that the corresponding modified equation has $I(y)$ as first integral.
9. Find at least two linear transformations $\rho$ for which the Kepler problem (I.2.1), written as a first order system, is $\rho$-reversible.
10. Consider explicit 2-stage Runge-Kutta methods of order 2, applied to the pendulum problem $\dot{q}=p, \dot{p}=-\sin q$. With the help of Exercise 2 compute $f_{3}(p, q)$ of the modified differential equation. Is there a choice of the free parameter $c_{2}$, such that $f_{3}(p, q)$ is a Hamiltonian vector field?
11. Consider Kepler's problem (I.2.1), written as a Hamiltonian system (I.2.2). Find constants $M$ and $R$ such that (5.2) holds for all $(p, q) \in \mathbb{R}^{4}$ satisfying

$$
\|p\| \leq 2 \quad \text { and } \quad 0.8 \leq\|q\| \leq 1.2
$$

12. [GeM88] Consider a Hamiltonian system with one degree of freedom, i.e., $d=1$. Prove that, if a numerical method $\Phi_{h}\left(y_{0}\right)$ is symplectic and if it preserves the Hamiltonian exactly, then it satisfies $\Phi_{h}(y)=\varphi_{\alpha(h, y)}(y)$, where $\alpha(h, y)=h+\mathcal{O}\left(h^{2}\right)$ is a formal series in $h$.

## Bibliography

[AT87] M.P. Allen and D.J. Tildesley, Computer Simulation of Liquids, Clarendon Press, Oxford, 1987.
[Ar78] V.I. Arnold, Mathematical Methods of Classical Mechanics, Springer-Verlag, New York, 1978.
[Ba05] H.F. Baker, Alternants and continuous groups, Proc. of London Math. Soc. 3 (1905) 24-47.
[BG94] G. Benettin and A. Giorgilli, On the Hamiltonian interpolation of near to the identity symplectic mappings with application to symplectic integration algorithms, J. Statist. Phys. 74 (1994) 1117-1143.
[Be87] W.-J. Beyn, On invariant closed curves for one-step methods, Numer. Math. 51 (1987) 103-122.
[BS93] J.J. Biesiadecki and R.D. Skeel, Dangers of multiple-time-step methods, J. Comput. Phys. 109 (1993) 318-328.
[BoS94] P.B. Bochev and C. Scovel, On quadratic invariants and symplectic structure, BIT 34 (1994) 337-345.
[BCP96] K.E. Brenan, S.L. Campbell and L.R. Petzold, Numerical Solution of InitialValue Problems in Differential-Algebraic Equations, Classics in Appl. Math., SIAM, Philadelphia, 1996.
[Bu87] J.C. Butcher, The Numerical Analysis of Ordinary Differential Equations. RungeKutta and General Linear Methods, John Wiley \& Sons, Chichester, 1987.
[BW96] J.C. Butcher and G. Wanner, Runge-Kutta methods: some historical notes, Appl. Numer. Math. 22 (1996) 113-151.
[CIZ97] M.P. Calvo, A. Iserles and A. Zanna, Numerical solution of isospectral flows, Math. Comput. 66 (1997) 1461-1486.
[CMS94] M.P. Calvo, A. Murua and J.M. Sanz-Serna, Modified equations for ODEs, Contemporary Mathematics 172 (1994) 63-74.
[Ca1857] A. Cayley, On the theory of the analytic forms called trees, Phil. Magazine XIII (1857) 172-176.
[Co87] G.J. Cooper, Stability of Runge-Kutta methods for trajectory problems, IMA J. Numer. Anal. 7 (1987) 1-13.
[CG93] P.E. Crouch and R. Grossman, Numerical integration of ordinary differential equations on manifolds, J. Nonlinear Sci. 3 (1993) 1-33.
[DB94] P. Deuflhard and F. Bornemann, Numerische Mathematik II, Integration gewöhnlicher Differentialgleichungen, Walter de Gruyter, Berlin, 1994.
[DRV94] L. Dieci, R.D. Russell and E.S. van Vleck, Unitary integrators and applications to continous orthonormalization techniques, SIAM J. Numer. Anal. 31 (1994) 261-281.
[EF98] E. Eich-Soellner and C. Führer, Numerical Methods in Multibody Dynamics, B. G. Teubner Stuttgart, 1998.
[Ei88] T. Eirola, Invariant curves of one-step methods, BIT 28 (1988), 113-122.
[Ei93] T. Eirola, Aspects of backward error analysis of numerical ODE's, J. Comp. Appl. Math. 45 (1993), 65-73.
[FeK85] Feng Kang, On difference schemes and symplectic geometry, Proceedings of the 5-th Intern. Symposium on differential geometry \& differential equations, August 1984, Beijing (1985) 42-58.
[FeK91] Feng Kang, Formal power series and numerical algorithms for dynamical systems. In Proceedings of international conference on scientific computation, Hangzhou, China, Eds. Tony Chan \& Zhong-Ci Shi, Series on Appl. Math. 1 (1991) 28-35.
[FK95] Feng Kang, Collected Works (II), National Defense Industry Press, Beijing, 1995.
[FSh95] Feng Kang and Shang Zai-jiu, Volume-preserving algorithms for source-free dynamical systems, Numer. Math. 71 (1995) 451-463.
[FiS96] B. Fiedler and J. Scheurle, Discretization of homoclinic orbits, rapid forcing and "invisible" chaos, Me. Amer. Math. Sov. 119, no. 570, 1996.
[Fla74] H. Flaschka, The Toda lattice, I. Phys. Rev. B9 (1974) 1924-1925.
[GeM88] Ge Zhong and J.E. Marsden, Lie-Poisson Hamilton-Jacobi theory and Lie-Poisson integrators, Phys. Lett. A 133 (1988) 134-139.
[GDC91] B. Gladman, M. Duncan and J. Candy, Symplectic integrators for long-term integrations in celestial mechanics, Celestial Mechanics and Dynamical Astronomy 52 (1991) 221-240.
[GVL89] G.H. Golub and C.F. Van Loan, Matrix Computations, 2nd edition, John Hopkins Univ. Press, Baltimore and London, 1989.
[GSS86] D.F. Griffiths and J.M. Sanz-Serna, On the scope of the method of modified equations, SIAM J. Sci. Stat. Comput. 7 (1986) 994-1008.
[GS69] A. Guillou and J.L. Soulé, La résolution numérique des problèmes différentiels aux conditions initiales par des méthodes de collocation. R.I.R.O. R-3 (1969) 17-44.
[Gra64] W.B. Gragg, Repeated extrapolation to the limit in the numerical solution of ordinary differential equations, Thesis, Univ. of California; see also SIAM J. Numer. Anal. 2 (1965) 384-403.
[Hai94] E. Hairer, Backward analysis of numerical integrators and symplectic methods, Annals of Numerical Mathematics 1 (1994) 107-132.
[Hai99] E. Hairer, Backward error analysis for multistep methods, Numer. Math. (1999) to appear.
[HaL97] E. Hairer and Ch. Lubich, The life-span of backward error analysis for numerical integrators, Numer. Math. 76 (1997), pp. 441-462.
Erratum: http://www.unige.ch/math/folks/hairer/
[HaL99] E. Hairer and Ch. Lubich, Asymptotic expansions and backward analysis for numerical integrators, IMA Proceedings (1999) to appear.
[HNW93] E. Hairer, S.P. Nørsett and G. Wanner, Solving Ordinary Differential Equations I. Nonstiff Problems, 2nd edition, Springer Series in Computational Mathematics 8, Springer Berlin, 1993.
[HSt97] E. Hairer and D. Stoffer, Reversible long-term integration with variable stepsizes, SIAM J. Sci. Comput. 18 (1997) 257-269.
[HW96] E. Hairer and G. Wanner, Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems, 2nd edition, Springer Series in Computational Mathematics 14, Springer-Verlag Berlin, 1996.
[HW97] E. Hairer and G. Wanner, Analysis by Its History, 2nd printing, Undergraduate Texts in Mathematics, Springer-Verlag New York, 1997.
[Ha1834] Sir W.R. Hamilton, On a general method in dynamics; by which the study of the motions of all free systems of attracting or repelling points is reduced to the search and differentiation of one central relation, or characteristic function, Phil. Trans. Roy. Soc. Part II for 1834, 247-308; Math. Papers, Vol. II, 103-161.
[Hau06] F. Hausdorff, Die symbolische Exponentialformel in der Gruppentheorie, Berichte der Sächsischen Akad. der Wissensch. 58 (1906) 19-48.
[Is84] A. Iserles, Solving linear ordinary differential equations by exponentials of iterated commutators, Numer. Math. 45 (1984) 183-199.
[IN97] A. Iserles and S.P. Nørsett, On the solution of linear differential equations in Lie groups, Technical Report 1997/NA3, DAMTP, University of Cambridge.
[Ki86] U. Kirchgraber, Multi-step methods are essentially one-step methods, Numer. Math. 48 (1986) 85-90.
[Ki88] U. Kirchgraber, An ODE-solver based on the method of averaging, Numer. Math. 53 (1988) 621-652.
[La88] F.M. Lasagni, Canonical Runge-Kutta methods, ZAMP 39 (1988) 952-953.
[Mag54] W. Magnus, On the exponential solution of differential equations for a linear operator, Comm. Pure Appl. Math. VII (1954) 649-673.
[MaR94] J.E. Marsden and T.S. Ratiu, Introduction to Mechanics and Symmetry, SpringerVerlag, New York, 1994.
[McL95] R.I. McLachlan, On the numerical integration of ordinary differential equations by symmetric composition methods, SIAM J. Sci. Comput. 16 (1995) 151-168.
[MK98] H. Munthe-Kaas, High order Runge-Kutta methods on manifolds, To appear in J. Appl. Num. Maths.
[Olv86] P.J. Olver, Applications of Lie Groups to Differential Equations, Graduate Texts in Mathematics 107, Springer-Verlag, New York, 1986.
[Po1899] H. Poincaré, Les méthodes nouvelles de la mécanique céleste. Tome III, GauthiersVillars, Paris, 1899.
[Rei98] S. Reich, Dynamical systems, numerical integration, and exponentially small estimates, Habilitationsschrift, Preprint SC 98-19 (July 1998), Konrad-Zuse-Zentrum Berlin.
[Ru83] R.D. Ruth, A canonical integration technique, IEEE Trans. Nuclear Science NS-30 (1983) 2669-2671.
[SS88] J.M. Sanz-Serna, Runge-Kutta schemes for Hamiltonian systems, BIT 28 (1988) 877883.
[SS92] J.M. Sanz-Serna, Symplectic integrators for Hamiltonian problems: an overview, Acta Numerica 1 (1992) 243-286.
[SS97] J.M. Sanz-Serna, Geometric integration, in: The state of the art in numerical analysis, Inst. Math. Appl. Conf. Ser. New Ser. 63, Oxford Univ. Press, New York, 1997, 121-143.
[SC94] J.M. Sanz-Serna and M.P. Calvo, Numerical Hamiltonian Problems, Chapman \& Hall, London, 1994.
[Sh86] L.F. Shampine, Conservation laws and the numerical solution of ODEs, Comp. Maths. Appls. 12B (1986) 1287-1296.
[Ste73] H.J. Stetter, Analysis of Discretization Methods for Ordinary Differential Equations, Springer-Verlag, Berlin, 1973.
[Str68] G. Strang, On the construction an comparison of difference schemes, SIAM J. Numer. Anal. 5 (1968) 506-517.
[StH96] A.M. Stuart and A.R. Humphries, Dynamical Systems and Numerical Analysis, Cambridge University Press, Cambridge, 1996.
[Su93] G. Sun, Construction of high order symplectic Runge-Kutta Methods, J. Comput. Math. 11 (1993) 250-260.
[Sun93] G. Sun, Symplectic partitioned Runge-Kutta methods, J. Comput. Math. 11 (1993) 365-372.
[Su89] Y.B. Suris, The canonicity of mappings generated by Runge-Kutta type methods when integrating the systems $\ddot{x}=-\partial U / \partial x$, Zh. Vychisl. Mat. i Mat. Fiz. 29, 202-211 (in Russian); same as U.S.S.R. Comput. Maths. Phys. 29 (1989) 138-144.
[SW92] G.J. Sussman and J. Wisdom, Chaotic evolution of the solar system, Science 257 (1992) 56-62.
[Su90] M. Suzuki, Fractal decomposition of exponential operators with applications to manybody theories and Monte Carlo simulations, Phys. Lett. A 146 (1990) 319-323.
[Su92] M. Suzuki, General theory of higher-order decomposition of exponential operators and symplectic integrators, Phys. Lett. A 165 (1992) 387-395.
[Tr59] H.F. Trotter, On the product of semi-groups of operators, Proc. Am. Math. Soc. 10 (1959) 545-551.
[Var74] V.S. Varadarajan, Lie Groups, Lie Algebras and Their Representations, PrenticeHall, Englewood Cliffs, New Jersey, 1974
[Ver67] L. Verlet, Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules, Physical Review 159 (1967) 98-103.
[Vo56] 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, Ind. (1956)
[Wa73] G. Wanner, Runge-Kutta-methods with expansion in even powers of $h$, Computing 11 (1973) 81-85.
[Wi60] J.H. Wilkinson, Error analysis of floating-point computation, Numer. Math. 2 (1960) 319-340.
[Wr70] K. Wright, Some relationships between implicit Runge-Kutta, collocation and Lanczos $\tau$ methods, and their stability properties, BIT 10 (1970) 217-227.
[Yo90] H. Yoshida, Construction of higher order symplectic integrators, Phys. Lett. A 150 (1990) 262-268.
[Yo93] H. Yoshida, Recent progress in the theory and application of symplectic integrators, Celestial Mechanics and Dynamical Astronomy 56 (1993) 27-43.
[Za97] A. Zanna, Collocation and relaxed collocation for the Fer and the Magnus expansions, Technical Report 1997/NA17, DAMTP, University of Cambridge.


[^0]:    ${ }^{1}$ We thank Alexander Ostermann, who provided us with all these data.

[^1]:    ${ }^{2}$ Attention. In (3.4) and in the subsequent formulas $q_{n}$ denotes an approximation to $q(n h)$, whereas $q_{i}$ in (3.1) denotes the $i$ th subvector of $q$.

[^2]:    ${ }^{1}$ In this section we are not concerned about the convergence of the series. We shall see later in Chapter V that the series converges for sufficiently small $h$, if $a(t)$ satisfies an estimate $|a(t)| \leq c d^{\rho(t)}$ and if $f(y)$ is an analytic function. If $f(y)$ is only $k$-times differentiable, then all formulas of this section remain valid for the truncated B-series $\sum_{t \in T, \rho(t) \leq k} \%$ with a suitable remainder term of size $\mathcal{O}\left(h^{k+1}\right)$ added.

[^3]:    ${ }^{2}$ The study of symmetric methods has its origin in the development of extrapolation methods [Gra64, Ste73], because the global error admits an asymptotic expansion in even powers of $h$.

[^4]:    ${ }^{3}$ For irreducible Runge-Kutta methods, the condition (3.6) is also necessary for symmetry (after a suitable permutation of the stages).

[^5]:    ${ }^{4} .$. because Dahlquist had proved the result that stable symmetric multistep methods with positive growth parameters have an order at most 2 . We shall see later that certain partitioned multistep methods can have an excellent long-time behaviour.

[^6]:    ${ }^{1}$ In contrast to the notation of the previous chapters we are here consistent with the usual notation and we write the vector of partial derivatives $H_{p}$ as a row vector.

[^7]:    ${ }^{2}$ This problem is very popular in testing codes for stiff differential equations.

[^8]:    ${ }^{3}$ For irreducible Runge-Kutta methods the condition (3.3) is also necessary for the conservation of quadratic first integrals.

[^9]:    ${ }^{4}$ Wilhelm Magnus, born: 5 february 1907 in Berlin (Germany), died: 15 october 1990. This picture is taken from http://www-groups.dcs.st-and.ac.uk/~history/Mathematicians/, where one can also find a short biographie.

[^10]:    ${ }^{5}$ Marius Sophus Lie, born: 17 december 1842 in Nordfjordeid (Norway), died: 18 february 1899. This picture is taken from http://www-groups.dcs.st-and.ac.uk/~history/Mathematicians/, where one can also find a short biographie.

[^11]:    ${ }^{1}$ William Rowan Hamilton, born: 4 August 1805 in Dublin (Ireland), died: 2 September 1865. Picture, copied from http://www-history.mcs.st-and.ac.uk/history/Mathematicians/Hamilton.html, where one can also find a short biography.

[^12]:    ${ }^{2}$ For irreducible Runge-Kutta methods the condition (3.2) is also necessary for symplecticity.

[^13]:    ${ }^{3}$ John Edward Campbell, born: 27 May 1862 in Lisburn, Co Antrim (Ireland), died: 1 October 1924.
    ${ }^{4}$ Henry Frederick Baker, born: 3 July 1866 in Cambridge (England), died: 17 March 1956.
    ${ }^{5}$ Felix Hausdorff, born: 8 November 1869 in Breslau (Germany), died: 26 January 1942. All three pictures are copied from http://www-history.mcs.st-and.ac.uk/~history/Mathematicians, where one can also find short biographies.

[^14]:    ${ }^{6}$ The article [Str68] deals with spatial discretizations of partial differential equations such as $u_{t}=$ $A u_{x}+B u_{y}$. There, the functions $f_{i}$ typically contain differences in only one spatial direction.

[^15]:    ${ }^{1}$ Observe that after adding the equation $x^{\prime}=1, x(0)=0$, we get for $Y=(x, y)^{T}$ the autonomous differential equation $Y^{\prime}=F(Y)$ with $F(Y)=(1, f(x))^{T}$. Hence, all results of this chapter are applicable.

