Chapter 7
B-series and geometric integration

7.1 Introduction

In geometric integration, the success of a numerical method is judged, not only by its order of accuracy, but also by its ability to preserve structural properties of specific problems and problem classes. For example, numerical methods for the Kepler problem are regarded as preferable if they respect the conservation of energy and of angular momentum. More complicated gravitational models, as used in the simulations of the solar system over extended time periods, will inevitably drift from the true solution but the principles of geometric integration require us to understand and limit the loss of integrity of important physical characteristics, both qualitative and quantitative. The practical aim will be to control, and possibly even eliminate completely, the drift away from exact preservation of invariants.

Geometric integration has grown into a significant subject in its own right. This short review will comment on just a handful of questions, each of which has a close connection with B-series, and possibly some connection with integration methods in the sense of Chapter 4. Reference is made to the treatise [49] (Hairer, Lubich, Wanner, 2006), for a broad coverage of geometric integration.

Chapter outline

In Section 7.2 we will discuss some of the problem classes for which conservation properties are fundamentally important.

We then devote several sections to numerical methods which are known to be successful in solving many of these problems in a geometric manner. The first of these, introduced in Section 7.3, focusses on canonical and symplectic Runge–Kutta methods. The theoretical questions are expressed in B-series terms and the development leads to important methods and families of methods. Also, Runge–Kutta methods which use "processing" to obtain improved performance are discussed.

J. C. Butcher, *B-Series*, Springer Series in Computational Mathematics 55,
https://doi.org/10.1007/978-3-030-70956-3_7

The so-called "G-symplectic methods", introduced in Section 7.4, represent an attempt to generalize symplectic properties to a multivalue setting. B-series play a central role in the development of these relatively new methods and this application illustrates many of the techniques and theoretical B-series results. However, a detailed study of G-symplectic methods aims for more than this.

The methods actually constructed are more efficient than Runge–Kutta methods, of comparable order and accuracy, and their fundamental limitation, the destruction of useful accuracy because of incipient parasitism, does not seem to manifest itself within a large time-period, covering millions of time steps. In particular, Section 7.6 outlines the derivation of a sixth order method which satisfies many of the requirements of slow loss of integrity, because of parasitism.

The simulations presented in Section 7.8 provide some reasons for confidence in the quality of the methods, in spite of the evident dangers.

Section 7.9 contains an introduction to energy preserving methods. It is known that Runge–Kutta methods cannot preserve energy but, as exemplified by the Average Vector Field method, integration methods, in the sense of Chapter 4, also known as Continuous Stage Runge–Kutta methods, overcome this limitation. For these methods, necessary and sufficient conditions based on B-series are given for energy preservation. For continuous methods in general, a sufficient condition for energy preservation is given.

7.2 Hamiltonian and related problems

Quadratic invariants

Inequality

In the study of non-linear stability, differential equations on an inner-product space were considered in [15] (Butcher, 75) with special application to Runge–Kutta methods (see Section 7.3 (p. 252)). For the system

$$y'(x) = f(y), \qquad y(x_0) = y_0,$$

where

$$\langle f(Y), Y \rangle \leq 0, \tag{7.2 a}$$

the value of $\|y(x)\| := \sqrt{\langle y(x), y(x) \rangle}$ is bounded for all time, because

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}x} \langle y(x), y(x) \rangle &= 2 \langle y'(x), y(x) \rangle \\ &= 2 \langle f\big(y(x)\big), y(x) \rangle \\ &\leq 0. \end{aligned} \tag{7.2 b}$$

If the inner product in (7.2 a) is replaced by a positive smi-definiite iinner product, so that the norm becomes a semi-norm, the non-increasing property of $\|y(x)\|^2$ still holds.

The way this generalization can give a useful stability condition is based on the $2N$-dimensional system

$$\begin{aligned} y'(x) &= f(y), \qquad y(x_0) = y_0, \\ z'(x) &= f(z), \qquad z(x_0) = z_0. \end{aligned}$$

This models the behaviour of $\|y(x) - z(x)\|^2$ for increasing x, using the indefinite inner product

$$\left\langle \begin{bmatrix} Y \\ Z \end{bmatrix}, \begin{bmatrix} \widetilde{Y} \\ \widetilde{Z} \end{bmatrix} \right\rangle := \langle Y, \widetilde{Y} \rangle - \langle Y, \widetilde{Z} \rangle - \langle Z, \widetilde{Y} \rangle + \langle Z, \widetilde{Z} \rangle.$$

The indefinite inner product is based on the partitioned symmetric positive indefinite matrix

$$\begin{bmatrix} I & -I \\ -I & I \end{bmatrix},$$

and the stability rule becomes

$$\|y(x) - z(x)\| \leq \|y_0 - z_0\|.$$

Equality

If (7.2 a) is replaced by

$$\big\langle f(Y), Y \big\rangle = 0,$$

then (7.2 b) becomes

$$\frac{\mathrm{d}}{\mathrm{d}x} \big\langle y(x), y(x) \big\rangle = 0$$

and $\|y(x)\|$ is constant over all time.

In the case of exact conservation of a quadratic form, there is no special reason why this should be a positive definite, or even a positive semi-definite, form and we will therefore look at problems for which

$$[Y, f(y)]_Q = 0, \tag{7.2 c}$$

with

$$[Y, Y]_Q := \langle Y, QY \rangle,$$

for Q symmetric.

A particular family of problems which satisfy (7.2 c) are of the form

$$y' = S(y)y, \tag{7.2 d}$$

where $S(y)$ is skew-symmetric. In this case, Q is the identity matrix.

Euler rigid body rotations

Two famous examples of (7.2 d) are derived from the Euler equations of rigid body rotation in the absence of external moments:

$$\begin{aligned} Au' &= (B-C)vw, \\ Bv' &= (C-A)wu, \\ Cw' &= (A-B)uv, \end{aligned} \tag{7.2 e}$$

with A, B, C the (positive) principal moments of inertia and u, v, w the components of angular velocity about the principal inertial axes in the body. The conservation of rotational energy is found by writing $y^1 = \sqrt{A}u$, $y^2 = \sqrt{B}v$, $y^3 = \sqrt{C}w$. We have

$$y' = \begin{bmatrix} 0 & -\sqrt{\frac{C}{AB}}y^3 & \sqrt{\frac{B}{AC}}y^2 \\ \sqrt{\frac{C}{AB}}y^3 & 0 & -\sqrt{\frac{A}{BC}}y^1 \\ -\sqrt{\frac{B}{AC}}y^2 & \sqrt{\frac{A}{BC}}y^1 & 0 \end{bmatrix} y = Sy,$$

with S skew-symmetric, so that $\frac{1}{2}[y,y] = \frac{1}{2}(Au^2 + Bv^2 + Cw^2)$ is conserved. The conservation of the norm of the angular momentum vector form of the equations is found from $y^1 = Au$, $y^2 = Bv$, $y^3 = Cw$, with

$$y' = \begin{bmatrix} 0 & \frac{1}{C}y^3 & -\frac{1}{B}y^2 \\ -\frac{1}{C}y^3 & 0 & \frac{1}{A}y^1 \\ \frac{1}{B}y^2 & -\frac{1}{A}y^1 & 0 \end{bmatrix} y,$$

Poisson and Hamiltonian problems

Poisson systems

A Poisson system has the form

$$y' = S\nabla H := S(H')^{\intercal},$$

where H, the Hamiltonian, is a function of y and S is skew-symmetric and may also depend on y. For such a system, H is conserved because

$$\frac{\mathrm{d}}{\mathrm{d}x}H = H'y' = H'S(H')^{\intercal} = 0,$$

because of the skew-symmetry of S. In the case of (7.2 e), define $y = \begin{bmatrix} u & v & w \end{bmatrix}^\intercal$, and this problem can be formulated in Poisson form in two different ways:

$$y' = S\nabla H, \quad H = \tfrac{1}{2} y^\intercal \operatorname{diag}(A,B,C) y, \qquad S = \begin{bmatrix} 0 & -\frac{Cy^3}{AB} & \frac{By^2}{AC} \\ \frac{Cy^3}{AB} & 0 & -\frac{Ay^1}{BC} \\ -\frac{By^2}{AC} & \frac{Ay^1}{BC} & 0 \end{bmatrix},$$

$$y' = S\nabla H, \quad H = \tfrac{1}{2} y^\intercal \operatorname{diag}(A^2,B^2,C^2) y, \quad S = \begin{bmatrix} 0 & \frac{y^3}{AB} & -\frac{y^2}{AC} \\ -\frac{y^3}{AB} & 0 & \frac{y^1}{BC} \\ \frac{y^2}{AC} & -\frac{y^1}{BC} & 0 \end{bmatrix}.$$

The Poisson formulation is a direct path to the conservation laws of rigid body rotation.

Hamiltonian problems

The Hamiltonian formulation of classical mechanics is a special case of a Poisson system based on a "Hamiltonian function", $H(y)$, and the governing equations take the form

$$y' = S\nabla H,$$

where S is skew-symmetric, constant and non-singular. The non-singularity of S implies that the dimension N is even. Write $N = 2n$.

To adhere to the classical formulation of Hamiltonian mechanics, attention is restricted to the special case

$$S = J^\intercal = \begin{bmatrix} \mathbf{0} & I \\ -I & \mathbf{0} \end{bmatrix}, \qquad J = \begin{bmatrix} \mathbf{0} & -I \\ I & \mathbf{0} \end{bmatrix},$$

so that (7.2 a) can be written

$$\begin{aligned} q_i' &= \frac{\partial H}{\partial p_i}, \\ p_i' &= -\frac{\partial H}{\partial q_i}, \end{aligned} \qquad i = 1,2,\ldots,n,$$

where

$$y = \begin{bmatrix} q \\ p \end{bmatrix}.$$

In many applications in mechanics, H is separable:

$$H(p.q) = T(p) + V(q).$$

Variational problems

For an initial value problem

$$y' = f(y), \qquad y(x_0) = y_0, \tag{7.2 f}$$

with solution $x \mapsto y(x)$, the variational problem

$$Y'(x) = f'(y)Y(x), \qquad Y(x_0) = I, \tag{7.2 g}$$

is used to model the behaviour of infinitesimal deviations from the trajectory in (7.2 f). That is, if the initial value is modified by $y_0 \to y_0 + \delta$, then the solution at x is modified by $y(x) \to y(x) + Y(x)\delta + \mathscr{O}(\delta^2)$, where Y is the solution to (7.2 g).

Now consider the case that (7.2 f) is a Hamiltonian problem $f(x) = J^{\mathsf{T}}\nabla H(y)$. In this case $f'(x) = J^{\mathsf{T}}W(y)$, where $W(y)$ is the Hessian matrix with (i, j) element equal to $\partial^2 H/\partial y_i \partial y_j$. Note that $W(x)$ is symmetric.

Symplectic flows

Theorem 7.2A Let $Y(x)$ denote the solution to the initial value problem (7.2 g), where the underlying problem (7.2 f) is Hamiltonian, then

$$Y(x)^{\mathsf{T}} J Y(x) = J. \tag{7.2 h}$$

Proof. Because $Y(x_0) = I$ we see that (7.2 h) holds when $x = x_0$. It remains to show that

$$\frac{\mathrm{d}}{\mathrm{d}x} Y(x)^{\mathsf{T}} J Y(x) = 0.$$

We have

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x} Y(x)^{\mathsf{T}} J Y(x) &= Y'(x)^{\mathsf{T}} J Y(x) + Y(x)^{\mathsf{T}} J Y'(x) \\ &= Y(x)^{\mathsf{T}} W(y) J J Y(x) + Y(x)^{\mathsf{T}} J J^{\mathsf{T}} W(y) Y(x) \\ &= -Y(x)^{\mathsf{T}} W(y) Y(x) + Y(x)^{\mathsf{T}} W(y) Y(x) = 0. \end{aligned}$$

□

Numerical methods which conserve quantities related to invariants of a given problem are the main subject of this chapter.

7.3 Canonical and symplectic Runge–Kutta methods

Canonical Runge–Kutta methods, also known as symplectic Runge–Kutta methods, have the ability to preserve quadratic invariants and symplectic structures. They will be defined in terms of a matrix, usually denoted by M.

The matrix M and an identity

Given a Runge–Kutta tableau, $(A, b^\intercal, c)$, we will be interested in the symmetric matrix $M = [m_{ij}]_{i,j=1}^{s}$ defined as

$$M = \operatorname{diag}(b)A + A^\intercal \operatorname{diag}(b) - bb^\intercal$$
$$= \begin{bmatrix} 2b_1a_{11} - b_1^2 & b_1a_{12} + b_2a_{21} - b_1b_2 & \cdots & b_1a_{1s} + b_sa_{s1} - b_1b_s \\ b_2a_{21} + b_1a_{12} - b_2b_1 & 2b_2a_{22} - b_2^2 & \cdots & b_2a_{2s} + b_sa_{s2} - b_2b_s \\ \vdots & \vdots & & \vdots \\ b_sa_{s1} + b_1a_{1s} - b_sb_1 & b_sa_{s2} + b_2a_{2s} - b_sb_2 & \cdots & 2b_sa_{ss} - b_s^2 \end{bmatrix}. \tag{7.3 a}$$

The quadratic identity

Theorem 7.3A Let the stages and the stages derivatives be given by the usual Runge–Kutta equations

$$\begin{aligned} Y_i &= y_0 + h\sum_{j=1}^{s} a_{ij}F_j, \quad i = 1, 2, \ldots, s, \\ F_i &= f(Y_i), \\ y_1 &= y_0 + h\sum_{i=1}^{s} b_iF_i, \end{aligned}$$

then, for any symmetric inner product, Q,

$$[y_1, y_1]_Q = [y_0, y_0]_Q + h\sum_{i=1}^{s} b_i[Y_i, F_i]_Q + h\sum_{j=1}^{s} b_j[F_j, Y_j]_Q - h^2\sum_{i=1}^{s}\sum_{j=1}^{s} m_{ij}[F_i, F_j]_Q.$$

Proof. Evaluate three expressions as follows:

$$\begin{aligned} E_1 &:= [y_1, y_1]_Q - [y_0, y_0]_Q \\ &= \left[\left(y_0 + h\sum_{i=1}^{s} b_iF_i\right), \left(y_0 + h\sum_{j=1}^{s} b_jF_j\right)\right]_Q - [y_0, y_0]_Q \\ &= h^2\sum_{i,j=1}^{s} b_ib_j[F_i, F_j]_Q + h\sum_{i=1}^{s} b_i[F_i, y_0]_Q + h\sum_{j=1}^{s} b_j[y_0, F_j]_Q, \end{aligned}$$

$$\begin{aligned} E_2 &:= -h\sum_{i=1}^{s} b_i[Y_i, F_i]_Q \\ &= -h^2\sum_{i,j=1}^{s} b_ia_{ij}[F_j, F_i]_Q - \sum_{i=1}^{s} b_i[y_0, F_i]_Q, \end{aligned}$$

$$\begin{aligned} E_3 &:= -h\sum_{j=1}^{s} b_j[F_j, Y_j]_Q \\ &= -h^2\sum_{i,j=1}^{s} b_ja_{ji}[F_j, F_i]_Q - \sum_{j=1}^{s} b_j[F_j, y_0]_Q, \end{aligned}$$

with $E_1+E_2+E_3=-h^2\sum_{i=1}^{s}\sum_{j=1}^{s}m_{ij}[F_i,F_j]_Q$. □

Stability applications

A problem satisfying $[Y,F]_Q\le 0$, where Q is positive indefinite is "dissipative". The following result is from [15] (Butcher, 1975).

Theorem 7.3B Let $(A,b^{\intercal},c)$ be a Runge–Kutta method satisfying (i) M is positive indefinite and (ii) $b_i>0$, $i=1,2,\ldots,s$. Then for any dissipative problem, $[y_n,y_n]_Q$ is non-increasing.

Proof. Because M is positive indefinite, it is the sum of squares of linear forms and hence

$$\sum_{i=1}^{s}\sum_{j=1}^{s}m_{ij}[F_i,F_j]_Q\ge 0.$$

From Theorem 7.3A, $[y_n,y_n]_Q-[y_{n-1},y_{n-1}]_Q$ is the sum of three non-positive terms. □

The conservation case

For problems satisfying $[Y,F]_Q=0$, where it is not necessary to assume that Q has any special properties other than symmetry, the stability condition becomes a conservation property of the differential equation.

Theorem 7.3C Let $(A,b^{\intercal},c)$ be a Runge–Kutta method with $M=0$. Then for a problem satisfying $[Y,F]_Q=0$, $[y_n,y_n]_Q$ is constant.

Proof. From Theorem 7.3A, $[y_n,y_n]_Q-[y_{n-1},y_{n-1}]_Q$ is the sum of three zero terms. □

For applications of this result to problems possessing quadratic invariants, see [32] (Cooper, 1987) and [68] (Lasagni. 1988). For applications to Hamiltonian problems see [83] (Sanz-Serna,1988).

Order conditions

The order conditions for Runge–Kutta methods have a remarkable property in the case of symplectic methods (see [84] (Sanz-Serna, Abia, 1991)). Rather than impose sufficiently many additional restrictions as to make canonical methods elusive, and difficult to construct, the conditions $M=0$ actually lead to simplifications.

To illustrate this effect, look at the usual conditions for order 4, where the underlying trees are also shown

$$b^{\intercal}\mathbf{1}=1, \tag{7.3 b}$$

$$b^{\intercal}c=\tfrac{1}{2}, \tag{7.3 c}$$

$$b^{\intercal}c^2=\tfrac{1}{3}, \tag{7.3 d}$$

$$b^{\intercal}Ac=\tfrac{1}{6}, \tag{7.3 e}$$

$$b^{\intercal}c^3=\tfrac{1}{4}, \tag{7.3 f}$$

$$b^{\intercal}cAc=\tfrac{1}{8}, \tag{7.3 g}$$

$$b^{\mathsf{T}}Ac^2 = \tfrac{1}{12}, \tag{7.3 h}$$

$$b^{\mathsf{T}}A^2c = \tfrac{1}{24}. \tag{7.3 i}$$

Write $M = 0$ in the form

$$\operatorname{diag}(b)A + A^{\mathsf{T}}\operatorname{diag} b = bb^{\mathsf{T}} \tag{7.3 j}$$

and form the inner product

$$u^{\mathsf{T}}\operatorname{diag}(b)Av + u^{\mathsf{T}}A^{\mathsf{T}}\operatorname{diag}(b)v = u^{\mathsf{T}}bb^{\mathsf{T}}v,$$

for various choices of u and v, to obtain the results

$$u = \mathbf{1},\quad v = \mathbf{1},\quad \text{yields}\quad 2b^{\mathsf{T}}c = (b^{\mathsf{T}}\mathbf{1})^2, \tag{7.3 k}$$

$$u = \mathbf{1},\quad v = c,\quad \text{yields}\quad b^{\mathsf{T}}c^2 + b^{\mathsf{T}}Ac = (b^{\mathsf{T}}\mathbf{1})(b^{\mathsf{T}}c), \tag{7.3 l}$$

$$u = c,\quad v = c,\quad \text{yields}\quad 2b^{\mathsf{T}}cAc = (b^{\mathsf{T}}c)^2, \tag{7.3 m}$$

$$u = \mathbf{1},\quad v = c^2,\quad \text{yields}\quad b^{\mathsf{T}}c^3 + b^{\mathsf{T}}Ac^2 = (b^{\mathsf{T}}\mathbf{1})(b^{\mathsf{T}}c^2), \tag{7.3 n}$$

$$u = \mathbf{1},\quad v = Ac,\quad \text{yields}\quad b^{\mathsf{T}}cAc + b^{\mathsf{T}}A^2 = (b^{\mathsf{T}}\mathbf{1})(b^{\mathsf{T}}Ac). \tag{7.3 o}$$

Starting from (7.3 b), we find in turn

$$\begin{aligned}
(7.3\,\mathrm{k}) &\implies (7.3\,\mathrm{c}),\\
(7.3\,\mathrm{l}) &\implies (7.3\,\mathrm{d}) + (7.3\,\mathrm{e}),\\
(7.3\,\mathrm{m}) &\implies (7.3\,\mathrm{g}),\\
(7.3\,\mathrm{n}) &\implies (7.3\,\mathrm{f}) + (7.3\,\mathrm{h}),\\
(7.3\,\mathrm{o}) &\implies (7.3\,\mathrm{g}) + (7.3\,\mathrm{i}).
\end{aligned}$$

In summary, instead of the 8 independent order conditions (7.3 b)–(7.3 i), it is only necessary to impose the three conditions (7.3 b), (7.3 d), (7.3 f) to obtain order 4, *given that the method is canonical*.

To extend this approach to any order, consider a sequence of steps that could be taken to verify that all order conditions are satisfied..

1. Show that the order 1 condition is satisfied.
2. For $p = 2, \ldots,$ up to the required order, show that the order condition for one tree within each non-superfluous class of order p is satisfied.
3. For $p = 2, \ldots,$ up to the required order, show that the order condition for one tree within each superfluous class of order p is satisfied.
4. Show that if the order condition for one tree within each order p class is satisfied then the same is true for all trees in the class.

In the case of canonical methods, for which Theorem 7.3A holds, Steps 3 and 4 in this sequence are automatically satisfied and only Steps 1 and 2 needs to be verified.

Theorem 7.3D For a canonical Runge–Kutta method, of order $p-1$, let $\mathrm{t}_1 = \mathrm{t} * \mathrm{t}'$, $\mathrm{t}_2 = \mathrm{t}' * \mathrm{t}$ where $|\mathrm{t}| + |\mathrm{t}'| = p$. Then

$$\big(\Phi(\mathrm{t}_1) - \tfrac{1}{\mathrm{t}_1!}\big) + \big(\Phi(\mathrm{t}_2) - \tfrac{1}{\mathrm{t}_2!}\big) = \Phi(\mathrm{t})\Phi(\mathrm{t}') - \tfrac{1}{\mathrm{t}!}\tfrac{1}{\mathrm{t}'!}. \qquad (7.3\,\mathrm{p})$$

Proof. To show that $\Phi(\mathrm{t}_1) + \Phi(\mathrm{t}_2) = \Phi(\mathrm{t})\Phi(\mathrm{t}')$, write $\Phi(\mathrm{t}) = b^{\intercal}\phi$, $\Phi(\mathrm{t}') = b^{\intercal}\phi'$, so that

$$\Phi(\mathrm{t}_1) = b^{\intercal}\phi A\phi' = \phi^{\intercal}\,\mathrm{diag}(b)A\phi',$$
$$\Phi(\mathrm{t}_2) = b^{\intercal}\phi' A\phi = \phi^{\intercal}A^{\intercal}\,\mathrm{diag}(b)\phi'.$$

From (7.3 j), it follows that

$$\phi^{\intercal}\big(\,\mathrm{diag}(b)A + A^{\intercal}\,\mathrm{diag}(b)\big) = \phi^{\intercal}bb^{\intercal}\phi' = \Phi(\mathrm{t})\Phi(\mathrm{t}').$$

To show that $(\mathrm{t}_1!)^{-1} + (\mathrm{t}_2!)^{-1} = (\mathrm{t}!)^{-1}(\mathrm{t}'!)^{-1}$, use the recursions

$$\mathrm{t}_1! = \frac{\mathrm{t}!\mathrm{t}'!|\mathrm{t}_1|}{|\mathrm{t}'|}, \quad \mathrm{t}_2! = \frac{\mathrm{t}'!\mathrm{t}!|\mathrm{t}_2|}{|\mathrm{t}|},$$

so that

$$\frac{1}{\mathrm{t}_1!} + \frac{1}{\mathrm{t}_2!} = \frac{1}{\mathrm{t}!}\frac{1}{\mathrm{t}'!}\Big(\frac{\mathrm{t}'!}{\mathrm{t}_1!} + \frac{\mathrm{t}!}{\mathrm{t}_2!}\Big) = \frac{1}{\mathrm{t}!}\frac{1}{\mathrm{t}'!}. \qquad \square$$

Theorem 7.3E For a canonical Runge–Kutta method, the number of independent conditions for order p is equal to the number of non-superfluous free trees of order up to p.

Proof. From (7.3 p), we deduce that the order conditions for t_1 and t_2 are equivalent and hence only one condition is required for each non-superfluous free tree. In the case of a superfluous tree $\mathrm{t}_1 = \mathrm{t}_2 = \mathrm{t} * \mathrm{t}$, (7.3 j) implies $2\Phi(\mathrm{t}_1) = 2(\mathrm{t}_1!)^{-1}$. $\square$

Particular methods

Gauss methods

For the classical Gauss methods it was shown in [15] (Butcher, 1975) that M is positive indefinite and therefore that the method is algebraically stable. But in the present context, these methods are symplectic because $M = 0$.

Theorem 7.3F Let $(A, b^{\intercal}, c)$ be the Gauss method with s stages,.then

$$\mathrm{diag}(b)A + A^{\intercal}\,\mathrm{diag}(b) = bb^{\intercal}.$$

Proof. Let V denote the Vandermonde matrix with (i,j) element equal to c_i^{j-1}. From the order conditions for the trees $[\tau^{i-1}[\tau^{j-1}]]$, the product $V^{\intercal}\,\mathrm{diag}(b)AV$ has (i,j)

element equal to $1/j(i+j)$. Hence, the (i,j) element of $V^{\intercal}(\mathrm{diag}(b)A + A^{\intercal}\mathrm{diag}(b) - bb^{\intercal})V$ is equal to $1/j(i+j) + 1/i(i+j) - 1/ij = 0$. Because V is non-singular, the result follows. □

Diagonally implicit methods

Methods in which A is lower triangular are canonical only if they have the form

$$\begin{array}{c|cccc} \frac{1}{2}b_1 & \frac{1}{2}b_1 & & & \\ b_1+\frac{1}{2}b_2 & b_1 & \frac{1}{2}b_2 & & \\ \vdots & \vdots & \vdots & \ddots & \\ b_1+b_2+\cdots+\frac{1}{2}b_s & b_1 & b_2 & \cdots & \frac{1}{2}b_s \\ \hline & b_1 & b_2 & \cdots & b_s \end{array}$$

This can also be looked at as the product of a sequence of s scaled copies of the implicit mid-point rule method. That is, the product method

$$\begin{array}{c|c} \frac{1}{2}b_1 & \frac{1}{2}b_1 \\ \hline & b_1 \end{array} \quad \begin{array}{c|c} \frac{1}{2}b_2 & \frac{1}{2}b_2 \\ \hline & b_2 \end{array} \quad \cdots \quad \begin{array}{c|c} \frac{1}{2}b_s & \frac{1}{2}b_s \\ \hline & b_s \end{array}.$$

For consistency, which will guarantee order 2, we must have $\sum_{i=1}^{s} b_i = 1$. To obtain order 3, we must have $\sum_{i=1}^{s} b_i^3 = 0$ and, assuming this holds, order 4 is also possible if $b^{\intercal}$ is symmetric, in the sense that $b_i = b_{s+1-i}$.

The simplest case of order 4 can then be found with $b_3 = b_1$ and satisfying

$$2b_1 + b_2 = 1, \tag{7.3 q}$$

$$2b_1^3 + b_2^3 = 0. \tag{7.3 r}$$

From (7.3 r), $b_2 = -\sqrt[3]{2}\,b_1$ and from (7.3 q) we then find

$$b^{\intercal} = \left[\begin{array}{ccc} \frac{1}{2-\sqrt[3]{2}} & \frac{-\sqrt[3]{2}}{2-\sqrt[3]{2}} & \frac{1}{2-\sqrt[3]{2}} \end{array} \right].$$

This gives the method [34] (Creutz, Gocksch, 1989), [88] (Suzuki, 1990), [92] (Yoshida, 1990)

$$\begin{array}{c|ccc}
\frac{1}{4-2\sqrt[3]{2}} & \frac{1}{4-2\sqrt[3]{2}} & & \\
\frac{1}{2} & \frac{1}{2-\sqrt[3]{2}} & \frac{-\sqrt[3]{2}}{4-2\sqrt[3]{2}} & \\
\frac{3-2\sqrt[3]{2}}{4-2\sqrt[3]{2}} & \frac{1}{2-\sqrt[3]{2}} & \frac{-\sqrt[3]{2}}{2-\sqrt[3]{2}} & \frac{1}{4-2\sqrt[3]{2}} \\
\hline
 & \frac{1}{2-\sqrt[3]{2}} & \frac{-\sqrt[3]{2}}{2-\sqrt[3]{2}} & \frac{1}{2-\sqrt[3]{2}}
\end{array} . \tag{7.3 s}$$

Many similar schemes exist of which the following is particularly convenient and efficient [88] (Suzuki, 1990)

$$b^{\intercal} = \left[\begin{array}{ccccc} \frac{1}{4-\sqrt[3]{4}} & \frac{1}{4-\sqrt[3]{4}} & \frac{-\sqrt[3]{4}}{4-\sqrt[3]{4}} & \frac{1}{4-\sqrt[3]{4}} & \frac{1}{4-\sqrt[3]{4}} \end{array} \right].$$

Block diagonally implicit

Nesting of known methods to obtain higher orders is possible using block diagonal structures. For example, if $(A, b^{\intercal}, c)$ is a symmetric canonical method with order 4, then the composition of three methods forming the product

$$\begin{array}{c|c} \theta c & \theta A \\ \hline & \theta b^{\intercal} \end{array} \cdot \begin{array}{c|c} (1-2\theta)c & (1-2\theta)A \\ \hline & (1-2\theta)b^{\intercal} \end{array} \cdot \begin{array}{c|c} \theta c & \theta A \\ \hline & \theta b^{\intercal} \end{array},$$

where $\theta = (2-\sqrt[5]{2})^{-1}$, will be canonical and have order 6.

For example, the method $(A, b^{\intercal}, c)$ could be the 2-stage Gauss method or the method (7.3 s).

Order with processing

In [69] (López-Marcos, Skeel, Sanz-Serna, 1996), it was proposed to precede a sequence of symplectic Runge–Kutta steps with a "processing step", which can have its effects reversed at the conclusion of the integration steps. This makes it possible to obtain adequate accuracy with an inexpensive integrator. This can be seen as an application of effective order, or conjugate order [13] (Butcher, 1969).

Let ξ denote the B-series for the input to each step so that the order conditions become

$$\begin{aligned} \eta &= A(\eta \mathrm{D}) + \mathbf{1}\xi, \\ \mathrm{E}\xi &= b^{\intercal}(\eta \mathrm{D}) + \xi + \mathsf{O}_{p+1}, \end{aligned} \tag{7.3 t}$$

where η is the stage B-series vector. For classical order, $\xi = 1$.

Conformability and weak conformability

The conformability and weak conformability conditions refer to the starting method (that is, the processor). To obtain the highest possible order, the values of ξ for any pair of equivalent trees need to be related in a special way.

Definition 7.3G The starting method ξ is conformable of order p if, for t, t$'$, such that $|\mathrm{t}|+|\mathrm{t}'] \le p-1$,

$$\xi(\mathrm{t}*\mathrm{t}')+\xi(\mathrm{t}'*\mathrm{t})=\xi(\mathrm{t})\xi(\mathrm{t}').$$

Definition 7.3H The starting method ξ is weakly conformable of order p if, for t, t$'$, such that $|\mathrm{t}|+|\mathrm{t}'] \le p$,

$$(\mathrm{E}\xi)(\mathrm{t}*\mathrm{t}')+(\mathrm{E}\xi)(\mathrm{t}'*\mathrm{t})-(\mathrm{E}\xi)(\mathrm{t})(\mathrm{E}\xi)(\mathrm{t}') = \xi(\mathrm{t}*\mathrm{t}')+\xi(\mathrm{t}'*\mathrm{t})-\xi(\mathrm{t})\xi(\mathrm{t}'). \quad (7.3\,\mathrm{u})$$

We now present a series of results interconnecting the two levels of conformability and order. Write O to mean that a method has order p relative to a specific choice of ξ, WC to mean that ξ is weakly conformable, C to mean that ξ is conformable and P to mean that if the order condition holds for a tree in each non-superfluous class, then the order is p.

The results can be summarized in the diagram.

$$\mathsf{O} \Longrightarrow \mathsf{WC} \Longleftrightarrow \mathsf{C} \Longrightarrow \mathsf{P}. \quad (7.3\,\mathrm{v})$$

Theorem 7.3I Let $(A,b^{\intercal},c)$ be a canonical Runge–Kutta method with order p relative to the starting method ξ. Then ξ is weakly conformable of order p.

Proof. Write (7.3 t) in the form

$$(\mathrm{E}\xi)(\mathrm{t})-\xi(\mathrm{t})=b^{\intercal}(\eta\mathrm{D})(\mathrm{t}), \qquad |\mathrm{t}|\le p,$$

and substitute $\mathrm{t}\to\mathrm{t}*\mathrm{t}'$, noting that

$$b^{\intercal}(\eta\mathrm{D})(\mathrm{t}*\mathrm{t}')=(\eta\mathrm{D})(\mathrm{t})^{\intercal}\,\mathrm{diag}(b)\eta(\mathrm{t}').$$

This gives

$$\begin{aligned}(\mathrm{E}\xi)(\mathrm{t}*\mathrm{t}')-\xi(\mathrm{t}*\mathrm{t}') &= (\eta\mathrm{D})(\mathrm{t})^{\intercal}\,\mathrm{diag}(b)\eta(\mathrm{t}')\\ &= (\eta\mathrm{D})(\mathrm{t})^{\intercal}\,\mathrm{diag}(b)\big(A(\eta\mathrm{D})(\mathrm{t}')+\mathbf{1}\xi(\mathrm{t}')\big)\\ &= (\eta\mathrm{D})(\mathrm{t})^{\intercal}\,\mathrm{diag}(b)A(\eta\mathrm{D})(\mathrm{t}')+b^{\intercal}(\eta\mathrm{D})(\mathrm{t})\xi(\mathrm{t}').\end{aligned}$$

Add a copy of this equation, with t and t$'$ interchanged, and the result is

$$
\begin{aligned}
&(\mathrm{E}\xi)(\mathrm{t}*\mathrm{t}')-\xi(\mathrm{t}*\mathrm{t}')+(\mathrm{E}\xi)(\mathrm{t}'*\mathrm{t})-\xi(\mathrm{t}'*\mathrm{t})\\
&\quad=(\eta\mathrm{D})(\mathrm{t})^{\intercal}\operatorname{diag}(b)A(\eta\mathrm{D})(\mathrm{t}')+(\eta\mathrm{D})(\mathrm{t}')^{\intercal}\operatorname{diag}(b)A(\eta\mathrm{D})(\mathrm{t})\\
&\qquad\quad+b^{\intercal}(\eta\mathrm{D})(\mathrm{t})\xi(\mathrm{t}')+b^{\intercal}(\eta\mathrm{D})(\mathrm{t}')\xi(\mathrm{t})\\
&\quad=(\eta\mathrm{D})(\mathrm{t})^{\intercal}\big(\operatorname{diag}(b)A+A^{\intercal}\operatorname{diag}(b)\big)(\eta\mathrm{D})(\mathrm{t}')\\
&\qquad\quad+b^{\intercal}(\eta\mathrm{D})(\mathrm{t})\xi(\mathrm{t}')+b^{\intercal}(\eta\mathrm{D})(\mathrm{t}')\xi(\mathrm{t})\\
&\quad=(\eta\mathrm{D})(\mathrm{t})^{\intercal}\big(bb^{\intercal}\big)(\eta\mathrm{D})(\mathrm{t}')+b^{\intercal}(\eta\mathrm{D})(\mathrm{t})\xi(\mathrm{t}')+b^{\intercal}(\eta\mathrm{D})(\mathrm{t}')\xi(\mathrm{t})\\
&\quad=\big(b^{\intercal}(\eta\mathrm{D})(\mathrm{t})+\xi(\mathrm{t})\big)\big(b^{\intercal}(\eta\mathrm{D})(\mathrm{t}')+\xi(\mathrm{t}')\big)-\xi(\mathrm{t})\xi(\mathrm{t}')\\
&\quad=(\mathrm{E}\xi)(\mathrm{t})(\mathrm{E}\xi)(\mathrm{t}')-\xi(\mathrm{t})\xi(\mathrm{t}'),
\end{aligned}
$$

which is equivalent to (7.3 u). □

Before showing the equivalence of conformability and weak conformability, we establish a utility definition and a utility lemma.

Definition 7.3J Let $\mathrm{t}=[\mathrm{t}_1\mathrm{t}_2\cdots\mathrm{t}_m\tau^n]$, where $\mathrm{t}_i\neq\tau$, $i=1,2,\ldots,m$. Then the bushiness of t is defined by $\operatorname{bush}(\mathrm{t})=n$.

Lemma 7.3K For $\xi\in\mathsf{B}$ and $\mathrm{t},\mathrm{t}'\in\mathrm{T}$,

$$
\begin{aligned}
&(\mathrm{E}\xi)(\mathrm{t}*\mathrm{t}')+(\mathrm{E}\xi)(\mathrm{t}'*\mathrm{t})-(\mathrm{E}\xi)(\mathrm{t})(\mathrm{E}\xi)(\mathrm{t})\\
&\quad=\sum_{\mathrm{x}\le\mathrm{t},\mathrm{x}'\le\mathrm{t}'}\mathrm{E}(\mathrm{t}\smallsetminus\mathrm{x})\mathrm{E}(\mathrm{t}'\smallsetminus\mathrm{x}')\big(\xi(\mathrm{x}*\mathrm{x}')+\xi(\mathrm{x}'*\mathrm{x})-\xi(\mathrm{x})\xi(\mathrm{x}')\big).
\end{aligned}
\tag{7.3 w}
$$

Proof. The subtrees of $\mathrm{t}*\mathrm{t}'$ are of the form $\mathrm{x}*\mathrm{x}'$ and x and hence

$$
\begin{aligned}
(\mathrm{E}\xi)(\mathrm{t}*\mathrm{t}')&=\sum_{\mathrm{x}\le\mathrm{t},\mathrm{x}'\le\mathrm{t}'}\mathrm{E}(\mathrm{t}\smallsetminus\mathrm{x})\mathrm{E}(\mathrm{t}'\smallsetminus\mathrm{x}')\xi(\mathrm{x}*\mathrm{x}')\\
&\qquad+\mathrm{E}(\mathrm{t}')\sum_{\mathrm{x}\le\mathrm{t}}\mathrm{E}(\mathrm{t}\smallsetminus\mathrm{x})\xi(\mathrm{x})+\mathrm{E}(\mathrm{t}*\mathrm{t}').
\end{aligned}
$$

Using this and the same formula, with t and t' interchanged, we find

$$
\begin{aligned}
&(\mathrm{E}\xi)(\mathrm{t}*\mathrm{t}')+(\mathrm{E}\xi)(\mathrm{t}'*\mathrm{t})-(\mathrm{E}\xi)(\mathrm{t})(\mathrm{E}\xi)(\mathrm{t})\\
&=\sum_{\mathrm{x}\le\mathrm{t},\mathrm{x}'\le\mathrm{t}'}\mathrm{E}(\mathrm{t}\smallsetminus\mathrm{x})\mathrm{E}(\mathrm{t}'\smallsetminus\mathrm{x}')\big(\xi(\mathrm{x}*\mathrm{x}')\\
&\quad+\xi(\mathrm{x}'*\mathrm{x})\big)+\big(\mathrm{E}(\mathrm{t}')\sum_{\mathrm{x}\le\mathrm{t}}\mathrm{E}(\mathrm{t}\smallsetminus\mathrm{x})\xi(\mathrm{x})\\
&\qquad+\mathrm{E}(\mathrm{t})\sum_{\mathrm{x}'\le\mathrm{t}'}\mathrm{E}(\mathrm{t}'\smallsetminus\mathrm{x}')\xi(\mathrm{x}')\big)+\big(\mathrm{E}(\mathrm{t}*\mathrm{t}')+\mathrm{E}(\mathrm{t}'*\mathrm{t})\big)\\
&\qquad-\big(\sum_{\mathrm{x}\le\mathrm{t}}\mathrm{E}(\mathrm{t}\smallsetminus\mathrm{x})\xi(\mathrm{x})+\mathrm{E}(\mathrm{t})\big)\big(\sum_{\mathrm{x}'\le\mathrm{t}'}\mathrm{E}(\mathrm{t}'\smallsetminus\mathrm{x}')\xi(\mathrm{x}')+\mathrm{E}(\mathrm{t}')\big).
\end{aligned}
$$

Noting that $E(t' * t) + E(t' * t) = E(t)E(t')$, we see that this reduces to the result of the lemma. □

We now have:

Theorem 7.3L The starting method ξ is weakly conformable of order p, if and only if it is conformable of order p.

Proof. The 'if' part of the proof follows from Lemma 7.3K because, if ξ is conformable of order p, all terms on the right of (7.3 w) are zero. To prove the only if result by induction, assume that ξ is conformable of order $p-1$, so that it is only necessary to show that for $|x|+|x'| = p-1$, $\xi(x * x') + \xi(x' * x) - \xi(x)\xi(x') = 0$. Without loss of generality, assume $\text{bush}(x) \geq \text{bush}(x')$. Note that $\text{bush}(x) \leq p-3$, corresponding to $x = [\tau^{p-3}]$, $x' = \tau$. We will carry out induction on $K = p-3, p-4, \ldots, 0$. For each K consider all x, x' pairs such that $\text{bush}(x) = K$. Define $t = x * \tau$, $t' = x'$, and substitute into (7.3 w). All terms on the right-hand side vanish because they correspond to a higher value of K and the single term corresponding to the current value of K. Hence we have

$$(K+1)\big(\xi(x * x') + \xi(x' * x) - \xi(x)\xi(x')\big) = 0. \qquad \square$$

Theorem 7.3M Let $(A, b^{\intercal}, c)$ be a canonical Runge–Kutta method such that, for each non-superfluous free tree, at least one of the trees has order p relative to a conformable starting method ξ, then all trees have order p relative to ξ.

Proof. Use an induction argument, so that the result can be assumed for all trees up to order $p-1$. It remains to show that if

$$\begin{aligned}&(E\xi)(t * t') - \xi(t * t') - b^{\intercal}(\eta D)(t * t') = 0, \text{ then}\\ &(E\xi)(t' * t) - \xi(t' * t) - b^{\intercal}(\eta D)(t' * t) = 0.\end{aligned}$$

Add these expressions and use the fact that

$$b^{\intercal}(\eta D)(t * t') + b^{\intercal}(\eta D)(t' * t) = (E\xi)(t)(E\xi)(t')\xi(t)\xi(t').$$

It is found that

$$\begin{aligned}&(E\xi)(t * t') - \xi(t * t') - b^{\intercal}(\eta D)(t * t') + (E\xi)(t' * t') - \xi(t' * t) - b^{\intercal}(\eta D)(t' * t)\\ &= (E\xi)(t * t') - \xi(t * t') + (E\xi)(t' * t') - \xi(t' * t)(E\xi)(t)(E\xi)(t') + \xi(t)\xi(t')\\ &= 0. \qquad \square\end{aligned}$$

7.4 G-symplectic methods

The multivalue form of the matrix M, and an identity

For a general linear method,

$$\left[\begin{array}{cc} A & U \\ B & V \end{array}\right],$$

the partitioned matrix

$$M = \left[\begin{array}{cc} DA + A^{\intercal}D - B^{\intercal}GB & DU - B^{\intercal}GV \\ U^{\intercal}D - V^{\intercal}GB & G - V^{\intercal}GV \end{array}\right] \tag{7.4 a}$$

was introduced in [6] (Burrage, Butcher, 1980) to characterize quadratic stability for multivalue methods and it has a similar role in the general linear case as the matrix (7.3 a) with the same name.

The matrix G appearing in M has a similar role as in the definition of G-stability [38]. In the general linear case, G is used to construct the quadratic form

$$[y^{[n]}, y^{[n]}]_{G\otimes Q} := \sum_{i,j=1}^{r} g_{ij}\,[y_i^{[n]}, y_j^{[n]}]_Q,$$

whose behaviour, as n increases, can be used to study non-linear stability ([6]) or conservation.

The quadratic identity for multivalue methods

The result given in Theorem 7.3A has a natural extension to the G-symplectic case

Theorem 7.4A

$$[y^{[n]}, y^{[n]}]_{G\otimes Q} = [y^{[n-1]}, y^{[n-1]}]_{G\otimes Q} + h[Y, F]_{D\otimes Q} + h[F, Y]_{D\otimes Q} - [v, v]_{M\otimes Q},$$

where

$$v = \left[\begin{array}{c} hF \\ y^{[n-1]} \end{array}\right].$$

Proof. Rewrite (7.4 a) in the form

$$\left[\begin{array}{c} B^{\intercal} \\ V^{\intercal} \end{array}\right] G \left[\begin{array}{cc} B & V \end{array}\right] = \left[\begin{array}{cc} 0 & 0 \\ 0 & G \end{array}\right] + \left[\begin{array}{c} A^{\intercal} \\ U^{\intercal} \end{array}\right] \left[\begin{array}{cc} D & 0 \end{array}\right] + \left[\begin{array}{c} D \\ 0 \end{array}\right] \left[\begin{array}{cc} A & U \end{array}\right] - M.$$

Apply the linear operation

$$X \mapsto [v, Xv]_Q$$

to each term in (7.4 a) and the result follows. □

Non-linear stability

As for Runge–Kutta methods, we will consider problems for which $[Y, f(Y)] \leq 0$ with the aim of achieving stable behaviour.

Definition 7.4B A general linear method (A,U,B,V) for which there exist D, a non-negative diagonal matrix, and G a positive semi-definite symmetric matrix is algebraically stable if M, given by (7.4 a), is positive semi-definite.

Theorem 7.4C For a problem for which $[Y, f(Y)]_Q$ is non-positive for positive indefinite Q, a numerical solution $y^{[n]}$, found from an algebraically stable Runge–Kutta method, has the property that $[y^{[n]}, y^{[n]}]_{G\otimes Q}$ is non increasing for $n = 0, 1, 2, \ldots$.

Proof. This follows from Theorem 7.4A. □

Conservation properties

We now turn our attention to problems for which $[Y, f(Y)]_Q = 0$ and methods for which $M = 0$, where M is given by (7.4 a). That is, we are considering methods covered by the following definition:

Definition 7.4D A general linear method (A,U,B,V) for which there exist D, a non-negative diagonal matrix and G such that

$$DA + A^{\intercal}D = B^*GB, \tag{7.4 b}$$

$$DU = B^*GV, \tag{7.4 c}$$

$$G = V^*GV.$$

is G-symplectic.

In this definition we have allowed for complex coefficients in U, B and V, by writing Hermitian transposes.

Theorem 7.4E Let (A,U,B,V) denote a G-symplectic method. Then for a problem for which $[Y, f(Y)]_Q = 0$, $[y^{[n]}, y^{[n]}]_Q$ is constant for $n = 0, 1, 2, \ldots$.

Proof. The result follows from the identity in Theorem 7.4A, with $[Y, F]_{D\otimes Q}$ and M deleted. □

Two methods based on Gaussian quadrature

The two methods **P** and **N** were introduced in [21] and differ only in the sign of $\sqrt{3}$ which appears in the coefficients. For **P** we have the defining matrices

$$\left[\begin{array}{cc} A & U \\ B & V \end{array}\right] = \left[\begin{array}{cc|cc} \frac{3+\sqrt{3}}{6} & 0 & 1 & -\frac{3+2\sqrt{3}}{6} \\ -\frac{\sqrt{3}}{3} & \frac{3+\sqrt{3}}{6} & 1 & \frac{3+2\sqrt{3}}{6} \\ \hline \frac{1}{2} & \frac{1}{2} & 1 & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 & -1 \end{array}\right].$$

This method can be verified to be G-symplectic with

$$G = \mathrm{diag}(1, \tfrac{3+2\sqrt{3}}{6}). \quad D = \mathrm{diag}(\tfrac{1}{2}, \tfrac{1}{2}).$$

Dealing with parasitism

Parasitism, and methods for overcoming its deleterious effects. were discussed in [21] (Butcher, Habib, Hill, Norton, 2014).

The stability function for **P** is

$$V + zB(I - zA)^{-1}U = V + zBU + \mathscr{O}(z^2) = \left[\begin{array}{cc} 1+z & 0 \\ 0 & -1-\frac{3+2\sqrt{3}}{6}z \end{array}\right] + \mathscr{O}(z^2).$$

For a high-dimensional problem, z represents the value of an eigenvalue of the Jacobian of f at points in the step being taken. In general, it is not possible to guarantee that the real parts of these eigenvalues will not be positive and hence the method cannot be guaranteed to be stable.

In numerical experiments with the simple pendulum, unstable behaviour does occur both for **P** and **N**. This is manifested by the loss of apparently bounded behaviour of the deviation of the Hamiltonian from its initial value. The onset depends on the initial amplitude of the pendulum swings and also appears later for **N**, compared with **P**. This behaviour is illustrated, in the case of **P**, in Figure 13. This shows the deviation of H from its initial value for the simple pendulum problem with $p = 0$ and two different values of q_0.

Conformability properties for general linear methods

We will extend Definitions 7.3G (p. 259) and 7.3H to the multivalue case. For Runge–Kutta methods, the need for these concepts only arose for methods with processing but, in the more general case, they are always needed because, even if the principal input might have a trivial starting method, the supplementary components will not. Recall (7.3 v) (p. 259) which applies, suitably interpreted, also to G-symplectic methods.

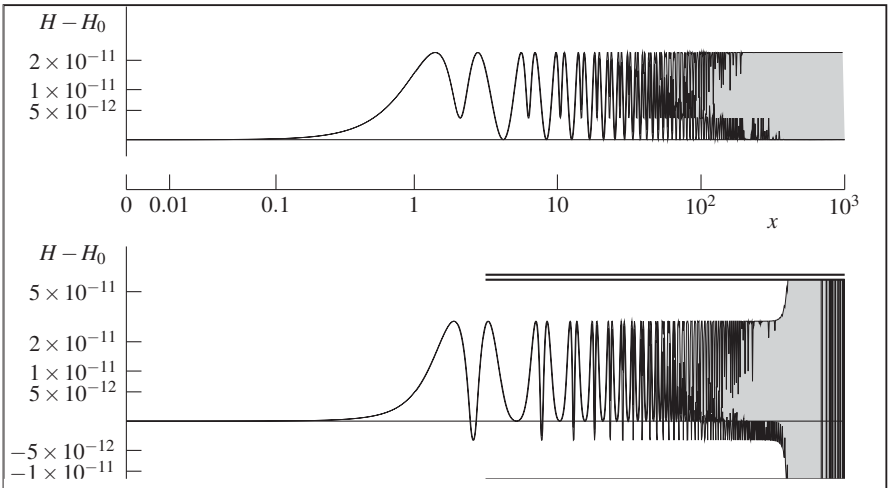


Figure 13 Variation in the Hamiltonian in attempts to solve the simple pendulum problem using method **P** with $h = 0.01$ and 10^5 time steps, with initial value $y_0 = [1.5, 0]^\intercal$ (upper figure) and $y_0 = [2, 0]^\intercal$ (lower figure)

Definition 7.4F The starting method ξ is conformable of order p if, for t, t$'$, such that $|\mathrm{t}| + |\mathrm{t}'| \leq p - 1$,

$$\xi_1(\mathrm{t} * \mathrm{t}') + \xi_1(\mathrm{t}' * \mathrm{t}) = \xi(\mathrm{t})^\intercal G \xi(\mathrm{t}').$$

Definition 7.4G The starting method ξ is weakly conformable of order p if, for t, t$'$, such that $|\mathrm{t}| + |\mathrm{t}'] \leq p$,

$$\begin{aligned}(\mathrm{E}\xi)_1(\mathrm{t} * \mathrm{t}') + (\mathrm{E}\xi)_1(\mathrm{t}' * \mathrm{t}) - (\mathrm{E}\xi)(\mathrm{t})^\intercal G (\mathrm{E}\xi)(\mathrm{t}') \\ = \xi_1(\mathrm{t} * \mathrm{t}') + \xi_1(\mathrm{t}' * \mathrm{t}) - \xi(\mathrm{t})^\intercal G \xi(\mathrm{t}').\end{aligned}$$

Theorem 7.4H The starting method ξ is conformable of order p if and only if ξ is weakly conformable of order p.

Theorem 7.4I Let (A, U, B, V) be a G-symplectic method with order p relative to the starting method ξ. Then ξ is weakly conformable of order p.

Theorem 7.4J Let (A,U,B,V) be a G-symplectic method with order at least $p-1$ relative to a starting method ξ, which is conformable of order p. Then the method satisfies the order condition for $\mathrm{t}*\mathrm{t}'$, where $|\mathrm{t}*\mathrm{t}'|=p$, if and only if it satisfies the order condition for $\mathrm{t}'*\mathrm{t}$.

Theorems 7.4H, 7.4I, 7.4J are proved in [24] (Butcher, Imran, 2015).

7.5 Derivation of a fourth order method

The method G4123

The method G4123, with $pqrs=4123$, was derived in [24] (Butcher, Imran, 2015). We will consider methods with a partitioned coefficient matrix

$$\left[\begin{array}{cc} A & U \\ B & V \end{array}\right] = \left[\begin{array}{c|cc} A & 1 & \widehat{U} \\ \hline b^{\mathsf{T}} & 1 & 0 \\ \widehat{B} & 0 & \widehat{V} \end{array}\right],$$

with the eigenvalues of $\widehat{V}$ distinct from 1 but lying on the unit circle.

It will be found that, with $s=3$ and $r=2$, fourth order G-symplectic methods exist such that A is lower-triangular with only a single non-zero diagonal element, and such that the parasitic growth factors are zero. A suitable ansatz is

$$\left[\begin{array}{cc} A & U \\ B & V \end{array}\right] = \left[\begin{array}{ccc|cc} \frac{1}{2}b_1(1+gx_1^2) & 0 & 0 & 1 & -gx_1 \\ b_1(1+gx_1x_2) & \frac{1}{2}b_2(1+gx_2^2) & 0 & 1 & -gx_2 \\ b_1(1+gx_1x_3) & b_2(1+gx_2x_3) & \frac{1}{2}b_3(1+gx_3^2) & 1 & -gx_3 \\ \hline b_1 & b_2 & b_3 & 1 & 0 \\ b_1x_1 & b_2x_2 & b_3x_3 & 0 & -1 \end{array}\right], \quad (7.5\,\mathrm{a})$$

based on

$$D=\mathrm{diag}(b_1,b_2,b_3), \qquad G=\mathrm{diag}(1,g).$$

For efficiency, we will attempt to obtain order 4 with $a_{11}=a_{22}=0$. We achieve this by choosing $g=-1$ (for simplicity, noting that g cannot be positive), together with $x_1=1$, $x_2=-1$. Substitute into (7.5 a) to obtain the simplified coefficient matrices

$$\left[\begin{array}{cc} A & U \\ B & V \end{array}\right] = \left[\begin{array}{ccc|cc} 0 & 0 & 0 & 1 & 1 \\ 2b_1 & 0 & 0 & 1 & -1 \\ b_1(1-x_3) & b_2(1+x_3) & \frac{1}{2}b_3(1-x_3^2) & 1 & x_3 \\ \hline b_1 & b_2 & b_3 & 1 & 0 \\ b_1 & -b_2 & b_3x_3 & 0 & -1 \end{array}\right].$$

Table 19 Solution and verification of (7.5 b)

	∅	•	⁝	∨	⁝	∨	∨	Y	⁝
ξ	1	0	$-\frac{1}{32}$	$-\frac{7}{4320}$	$\frac{149}{8640}$	0	0	0	0
$\widehat{\zeta}$	0	$\frac{1}{4}$	$-\frac{1}{16}$	$-\frac{49}{960}$	$-\frac{13}{384}$	$\frac{2543}{57600}$	$\frac{193}{7680}$	$\frac{619}{34560}$	$\frac{163}{69120}$
η_1	1	$\frac{1}{4}$	$-\frac{3}{32}$	$-\frac{91}{1728}$	$-\frac{287}{17280}$	$\frac{2543}{57600}$	$\frac{193}{7680}$	$\frac{619}{34560}$	$\frac{163}{69120}$
η_2	1	$\frac{5}{12}$	$\frac{19}{96}$	$\frac{787}{8640}$	$-\frac{197}{17280}$	$-\frac{1943}{57600}$	$-\frac{313}{7680}$	$-\frac{5497}{103680}$	$-\frac{557}{41472}$
η_3	1	$\frac{11}{20}$	$\frac{37}{160}$	$\frac{1147}{8640}$	$\frac{739}{17280}$	$\frac{377}{6400}$	$\frac{313}{12800}$	$\frac{2489}{172800}$	$-\frac{15487}{345600}$
$\eta_1 \mathrm{D}$	0	1	$\frac{1}{4}$	$\frac{1}{16}$	$-\frac{3}{32}$	$\frac{1}{64}$	$-\frac{3}{128}$	$-\frac{91}{1728}$	$-\frac{287}{17280}$
$\eta_2 \mathrm{D}$	0	1	$\frac{5}{12}$	$\frac{25}{144}$	$\frac{19}{96}$	$\frac{125}{1728}$	$\frac{95}{1152}$	$\frac{787}{8640}$	$-\frac{197}{17280}$
$\eta_3 \mathrm{D}$	0	1	$\frac{11}{20}$	$\frac{121}{400}$	$\frac{37}{160}$	$\frac{1331}{8000}$	$\frac{407}{3200}$	$\frac{1147}{8640}$	$\frac{739}{17280}$
$\mathrm{E}\xi$	1	1	$\frac{15}{32}$	$\frac{1163}{4320}$	$\frac{1319}{8640}$	$\frac{109}{720}$	$\frac{3}{32}$	$\frac{187}{2160}$	$\frac{187}{4320}$
$\mathrm{E}\widehat{\zeta}$	0	$\frac{1}{4}$	$\frac{3}{16}$	$\frac{71}{960}$	$\frac{11}{384}$	$-\frac{2677}{57600}$	$-\frac{73}{2560}$	$-\frac{1001}{34560}$	$-\frac{1457}{69120}$

It was shown in [24] (Butcher, Imran, 2015) how the free parameters and the starting vectors can be chosen to achieve order 4 accuracy and also to ensure that parasitic growth factors are zero. The method parameters, are

$$\left[\begin{array}{cc} A & U \\ B & V \end{array}\right] = \left[\begin{array}{ccc|cc} 0 & 0 & 0 & 1 & 1 \\ \frac{2}{3} & 0 & 0 & 1 & -1 \\ \frac{2}{5} & -\frac{3}{10} & \frac{1}{2} & 1 & -\frac{1}{5} \\ \hline \frac{1}{3} & -\frac{3}{8} & \frac{25}{24} & 1 & 0 \\ \frac{1}{3} & \frac{3}{8} & -\frac{5}{24} & 0 & -1 \end{array}\right], \qquad c = \left[\begin{array}{c} \frac{1}{4} \\ \frac{5}{12} \\ \frac{11}{20} \end{array}\right],$$

and they were chosen to satisfy the order conditions

$$\begin{aligned} \eta &= A\eta\mathrm{D} + 1\xi + \widehat{U}\widehat{\zeta}, \\ \mathrm{E}\xi &= b^{\mathsf{T}}\eta\mathrm{D} + \xi + \mathsf{O}_5, \\ \mathrm{E}\widehat{\zeta} &= \widehat{B}\eta\mathrm{D} + \widehat{V}\widehat{\zeta} + \mathsf{O}_5. \end{aligned} \tag{7.5 b}$$

The values of the starting methods, ξ and $\widehat{\zeta}$, and the stage values and derivatives, η and $\eta\mathrm{D}$, are given in Table 19, together with a verification that the conditions are satisfied. It was assumed from the start, without loss of generality, that ξ_1 and ξ_5, ξ_6, ξ_7, ξ_8 are zero. Note that the entries for $\mathrm{E}\xi$ and $\mathrm{E}\widehat{\zeta}$ in (7.5 b) are identical, to within O_5 and these lines are the final steps of the order verification.

Implementation questions

Starting and finishing methods for ξ

We will write $\mathcal{S}_h$ and $\mathcal{F}_h$ for the mappings corresponding to the starting and finishing methods, respectively, for $y_1^{[0]}$. Because each of the mappings is only required to be correct to within O_4, with the proviso that $\mathcal{F}_h \circ \mathcal{S}_h = \mathrm{id} + O_5$, we will first construct a Runge–Kutta tableau with only three stages which gives the B-series $\xi^{-1} + O_4$ from which a corresponding approximation to ξ can be found cheaply to within O_5.

Calculate the coefficients of ξ^{-1} for the first 4 trees and write down the order conditions for the required tableau

$$\begin{aligned} b_1 + b_2 + b_3 &= \xi^{-1}(\bullet) &&= 0, \\ b_2c_2 + b_3c_3 &= \xi^{-1}(t_2) &&= \tfrac{1}{32}, \\ b_2c_2^2 + b_3c_3^2 &= \xi^{-1}(t_3) &&= \tfrac{7}{4320}, \\ b_3a_{32}c_2 &= \xi^{-1}(t_4) &&= -\tfrac{149}{8640}. \end{aligned}$$

A possible solution to this system is

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{2} & \frac{1}{2} & & \\ 1 & -\frac{28}{121} & \frac{149}{121} & \\ \hline & -\frac{391}{4320} & \frac{16}{135} & -\frac{121}{4320} \end{array} .$$

If the finishing method for the first component is given by $\mathcal{F}_h$, then $\mathcal{S}_h$ can be approximated by

$$\mathcal{S}_h = 3\mathrm{id} - \mathcal{F}_h - \mathcal{F}_h \circ (2\mathrm{id} - \mathcal{F}_h).$$

Let

$$a = \begin{bmatrix} 1 & 0 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 \end{bmatrix}$$

be the B-series coefficients for $\mathcal{F}_h$, for $\emptyset \ldots \mathbf{t}_8$, so that the corresponding coefficient vector for $(2\mathrm{id} - \mathcal{F}_h)$ is

$$b = \begin{bmatrix} 1 & 0 & -a_2 & -a_3 & -a_4 & -a_5 & -a_6 & -a_7 & -a_8 \end{bmatrix}.$$

The series for $\mathcal{F}_h \circ (2\mathrm{id} - \mathcal{F}_h)$ is found to be

$$ba = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & -a_2^2 & 0 & -a_2^2 \end{bmatrix},$$

with the final result

$$\begin{bmatrix} 1 & 0 & -a_2 & -a_3 & -a_4 & -a_5 & a_2^2 - a_6 & -a_7 & a_2^2 - a_8 \end{bmatrix},$$

which is identical to the series for a^{-1}, corresponding to $\mathcal{S}_h$.

Now consider the starting method for $y_2^{[0]}$. This can be found using a generalized Runge–Kutta method with order conditions $\Phi(\mathrm{t}) = \widehat{\zeta}(\mathrm{t})$ for $|\mathrm{t}| \leq 4$ and coefficient of y_0 equal to zero. A suitable tableau for this starter is

$$\begin{array}{r|llll}
0 & & & & \\
-\frac{1}{4} & -\frac{1}{4} & & & \\
-\frac{1}{4} & \frac{9319973}{11609760} & \frac{6417533}{11609760} & & \\
\frac{1}{4} & -\frac{7417}{6432} & 0 & \frac{9025}{6432} & \\
-\frac{3}{4} & -\frac{887}{5536} & 0 & 0 & -\frac{3265}{5536} \\
\hline
0 & \frac{28}{675} & -\frac{1583059879}{5775779700} & \frac{43875218}{57757797} & -\frac{67}{450} \quad -\frac{173}{1350}
\end{array}.$$

Exercise 58 Find an alternative starting method for $y_2^{[0]}$, using a generalized five stage Runge–Kutta method with $a_{42} = a_{52} = a_{53} = 0$, and with $c_2 = c_3 = -\frac{1}{3}$, $c_4 = -\frac{2}{3}$, $c_5 = -1$.

Exercise 59 If a four stage generalized Runge–Kutta method is used for the starting method for $y_2^{[0]}$ and $c_2 = \frac{15924}{14305}$, what is c_4?

Initial approximation for Y_3

Because the first two stages of the method are explicit, the most important implementation question is the evaluation of the third stage. We will consider only the Newton method for this evaluation and we will need to find the most accurate method for obtaining an initial estimate to commence the iterations.

Information available when the first two stage derivatives have been computed in, for example, the first step of the solution, includes $y_1^{[0]}$, $y_2^{[0]}$, hF_1 and hF_2 and we will need to obtain a useful approximation to F_3. In terms of B-series coefficients we have

$$\begin{array}{rrrrrr}
\xi(\varnothing) = 1, & \xi(\bullet) = 0, & \xi(\!\!\downarrow) = -\frac{1}{32}, & \xi(\curlyvee) = -\frac{7}{4320}, \\
\widehat{\zeta}(\varnothing) = 0, & \widehat{\zeta}(\bullet) = \frac{1}{4}, & \widehat{\zeta}(\!\!\downarrow) = -\frac{1}{16}, & \widehat{\zeta}(\curlyvee) = -\frac{49}{960}, \\
\eta_1\mathrm{D}(\varnothing) = 0, & \eta_1\mathrm{D}(\bullet) = 1, & \eta_1\mathrm{D}(\!\!\downarrow) = \frac{1}{4}, & \eta_1\mathrm{D}(\curlyvee) = \frac{1}{16}, \\
\eta_2\mathrm{D}(\varnothing) = 0, & \eta_2\mathrm{D}(\bullet) = 1, & \eta_2\mathrm{D}(\!\!\downarrow) = \frac{5}{12}, & \eta_2\mathrm{D}(\curlyvee) = \frac{25}{144}, \\
\eta_3(\varnothing) = 1, & \eta_3(\bullet) = \frac{11}{20}, & \eta_3(\!\!\downarrow) = \frac{37}{160}, & \eta_3(\curlyvee) = \frac{1147}{8640}.
\end{array}$$

A short calculation suggests that the approximation

$$\eta \approx \xi + \widehat{\eta} - \tfrac{6}{5}\eta_1\mathrm{D} + \tfrac{3}{2}\eta_2\mathrm{D}$$

Table 20 Trees to order 6, grouped together as free trees with superfluency, symmetry and possible deletion if the $C(2)$ condition holds

order	serial number	free tree	tree count	superfluous	symmetric	C(2)
1	1		1			
2	2		1	X	X	
3	3		2			X
4	4		2		X	
	5		2	X	X	X
5	6		2			
	7		4			X
	8		3			X
6	9		2		X	
	10		4		X	X
	11		4		X	X
	12		2	X	X	X
	13		5		X	X
	14		3	X	X	X

is exact, based on just $\emptyset$, $\bullet$, I and V. Accordingly, the approximation

$$Y_3 \approx y_1^{[0]} + y_2^{[0]} - \tfrac{6}{5}hF_1 + \tfrac{3}{2}hF_2$$

is suggested to initialize the iterative computation of Y_3.

7.6 Construction of a sixth order method

This discussion is based on [25] (Butcher, Imran, Podhaisky, 2017). Of the 37 trees up to order 6, which contribute to the order requirements, these can be immediately reduced to 14 because of the role played by the equivalences which define free trees. Some of these can immediately be discarded because of superfluency. If the method is symmetric then further trees become candidates for deletion from the set of required order conditions [23] (Butcher, Hill, Norton, 2016). If it is possible to impose the $C(2)$ condition, further deletions are possible. These simplifications are summarized in Table 20.

Design requirements

Time-reversal symmetry

Methods with time-reversal symmetry were considered in [23] (Butcher, Hill, Norton, 2016). This property is an important attribute of numerical schemes for the long-term integration of mechanical problems. Furthermore, the symmetric general linear methods perform well over long time intervals. We can define a general linear method to be symmetric in a similar fashion to a Runge–Kutta method. A general linear method is symmetric if it is equal to its adjoint general linear method, where the adjoint general linear method takes the stepsize with opposite sign. However, symmetry in general linear methods is not as simple as for Runge–Kutta methods, because the output approximations contain the matrix V, which is multiplied by the input approximations, and it is possible that the inverse matrix V^{-1} is not equal to V. For this reason, an involution matrix L is introduced, such that $L^2 = I$ and $LV^{-1}L = V$. We also introduce the stage reversing permutation P defined as $P_{ij} = \delta_{i,s+1-j}$ for $i, j = 1, \ldots, s$.

In particular, because of time-reversal symmetry, trees with even order can be ignored because the corresponding conditions will be automatically satisfied.

Definition 7.6A A method (A, U, B, V) is time-reversal symmetric with respect to the involution L if

$$A + PAP = UV^{-1}B, \tag{7.6a}$$

$$VLBP = B, \tag{7.6b}$$

$$PULV = U, \tag{7.6c}$$

$$(LV)^2 = I. \tag{7.6d}$$

From results in [23], it follows that, for a method with this property, with starting method S_h, it can be assumed that $S_h = LS_{-h}$. Methods which are both G-symplectic and symmetric have many advantages, and some of these were derived in [23]. For methods with lower-triangular A, the two properties are closely related.

Theorem 7.6B Let (A, U, B, V) be a method with the properties

1. A is lower triangular,
2. The method is G-symplectic,
3. (7.6c) is satisfied,

then (7.6a), (7.6b) and (7.6d) are satisfied.

This result is proved in [20] (Butcher, 2016).

Structure of the method G6245

The method, which will be referred to as G6245, because $pqrs = 6245$, was originally derived in [25]. It achieves order 6 by combining symmetry, C(2) with G-symplecticity.

Symmetry requirements

An arbitrary choice is made to define

$$\begin{aligned} V &= \mathrm{diag}(1, \mathrm{i}, -\mathrm{i}, -1), \\ G &= \mathrm{diag}(1, -\tfrac{1}{2}, -\tfrac{1}{2}, 1), \\ U &= \left[\begin{array}{ccc} \mathbf{1} & \frac{1}{2}(-\beta - \mathrm{i}\alpha) & \frac{1}{2}(-\beta + \mathrm{i}\alpha) \quad -\gamma \end{array} \right], \end{aligned} \tag{7.6e}$$

where

$$\alpha := \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ 0 \\ -\alpha_2 \\ -\alpha_1 \end{bmatrix}, \qquad \beta := \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_2 \\ \beta_1 \end{bmatrix}, \qquad \gamma := \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_2 \\ \gamma_1 \end{bmatrix}.$$

Also write $b^{\mathsf{T}} = \left[\begin{array}{ccccc} b_1 & b_2 & b_3 & b_2 & b_1 \end{array} \right]$, $D = \mathrm{diag}(b)$. From (7.4c), we deduce

$$B = \begin{bmatrix} b^{\mathsf{T}} \\ (\alpha^{\mathsf{T}} + \mathrm{i}\beta^{\mathsf{T}})D \\ (\alpha^{\mathsf{T}} - \mathrm{i}\beta^{\mathsf{T}})D \\ \gamma^{\mathsf{T}} D \end{bmatrix}.$$

Define the 5×5 symmetric matrix W with elements $w_{ij} = \alpha_i \alpha_j + \beta_i \beta_j - \gamma_i \gamma_j$, $i, j = 1, 2, \ldots, 5$, which can be written

$$W = \alpha\alpha^{\mathsf{T}} + \beta\beta^{\mathsf{T}} - \gamma\gamma^{\mathsf{T}}. \tag{7.6f}$$

Because of the symmetries and anti-symmetries in α, β, γ, it follows that W has the form

$$W = \begin{bmatrix} w_{11} & w_{21} & w_{31} & w_{41} & w_{51} \\ w_{21} & w_{22} & w_{32} & w_{42} & w_{41} \\ w_{31} & w_{32} & w_{33} & w_{32} & w_{31} \\ w_{41} & w_{42} & w_{32} & w_{22} & w_{21} \\ w_{51} & w_{41} & w_{31} & w_{21} & w_{11} \end{bmatrix}. \tag{7.6g}$$

From (7.4 b) (p. 263), assuming A is lower triangular, the elements of this matrix are found to be

$$a_{ij} = \begin{cases} b_j(1-\alpha_i\alpha_j-\beta_i\beta_j+\gamma_i\gamma_j) = \frac{1}{2}b_j(1-w_{ij}), & j<i, \\ \frac{1}{2}b_j(1-\alpha_i\alpha_j-\beta_i\beta_j+\gamma_i\gamma_j) = b_j(1-w_{ij}), & j=i, \\ 0, & j>i. \end{cases} \tag{7.6 h}$$

Symmetry also requires $Pc+c=\mathbf{1}$ and $b^{\intercal}P=b^{\intercal}$ and we choose the abscissae vector as

$$c = \begin{bmatrix} 0 & \frac{1}{2}(1-t) & \frac{1}{2} & \frac{1}{2}(1+t) & 1 \end{bmatrix}$$

and the vector $b^{\intercal} = \begin{bmatrix} b_1 & b_2 & b_3 & b_2 & b_1 \end{bmatrix}$ such that

$$\begin{aligned} b^{\intercal}\mathbf{1} &= 1, \\ b^{\intercal}c^2 &= \tfrac{1}{3}, \\ b^{\intercal}c^4 &= \tfrac{1}{5}. \end{aligned}$$

The choice of $t = 1-2c_2$

The choice of t must yield a negative coefficient amongst b_1, b_2, b_3 to ensure that the parasitism growth factors can be eliminated. It is found that this is possible in three cases

Case 1:	$0<t^2<\frac{1}{5}$:	$b_3<0$,	
Case 2:	$\frac{3}{5}<t^2<1$:	$b_1<0$,	(7.6 i)
Case 3:	$1<t^2$:	$b_2<0$.	

From the C(2) conditions, the first three rows of A are given by

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2}c_2 & \frac{1}{2}c_2 & 0 & 0 & 0 \\ a_{31} & a_{32} & a_{33} & 0 & 0 \end{bmatrix}, \tag{7.6 j}$$

where a_{31} and a_{32} are written in terms of the parameter a_{33}, as solutions of the system

$$\begin{aligned} a_{31}+a_{32} &= c_3-a_{33}, \\ a_{32}c_2 &= \tfrac{1}{2}c_3^2-a_{33}c_3. \end{aligned}$$

To find a_{33}, use (7.6 h) and the symmetry of W to see that

$$\sum_{i=1}^{5} b_i w_{ii} = 2(b_1 - 2a_{11}) + 2(b_2 - 2a_{22}) + (b_3 - 2a_{33})$$
$$= 1 - 2c_2 - 2a_{33}.$$

However, to guarantee that the parasitism growth factors are zero, we must have

$$\sum_{i=1}^{5} b_i w_{ii} = \sum_{i=1}^{5} b_i \alpha_i^2 + \sum_{i=1}^{5} b_i \beta_i^2 - \sum_{i=1}^{5} b_i \gamma_i^2 = 0.$$

Hence,

$$a_{33} = \tfrac{1}{2} - c_2 = \tfrac{1}{2} t,$$

and (7.6 j) can be rewritten as

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2}c_2 & \frac{1}{2}c_2 & 0 & 0 & 0 \\ c_2 - \frac{1}{2} + \frac{1}{8c_2} & \frac{1}{2} - \frac{1}{8c_2} & \frac{1}{2} - c_2 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4}(1-t) & \frac{1}{4}(1-t) & 0 & 0 & 0 \\ \frac{1-2t+2t^2}{4-4t} & \frac{1-2t}{4-4t} & \frac{1}{2}t & 0 & 0 \end{bmatrix}.$$

Derivation of the method

Transform to the W formulation, impose the symmetry pattern given by (7.6 g) and transform back to find all elements of A except a_{41}, a_{42} a_{51}, a_{52}. These are found in turn from the C(2) conditions for the final two rows.

We will now show that the rank of W cannot exceed 3. This follows because $b^{\intercal} W = b^{\intercal} \operatorname{diag}(c) W = 0$, which can be verified by detailed calculations. A consequence of this is

$$\sum_{i=1}^{5} b_i \alpha_i = \sum_{i=1}^{5} b_i \beta_i = \sum_{i=1}^{5} b_i \gamma_i = \sum_{i=1}^{5} b_i c_i \alpha_i = \sum_{i=1}^{5} b_i c_i \beta_i = \sum_{i=1}^{5} b_i c_i \gamma_i = 0,$$

implying

$$b^{\intercal} W = 0, \tag{7.6 k}$$
$$b^{\intercal} \operatorname{diag}(c) W = 0. \tag{7.6 l}$$

A special case

As a special case, choose $t = \frac{1}{3}$ in Case 1 of (7.6 i). This gives

$$c = \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{2} & \frac{2}{3} & 1 \end{bmatrix}^{\intercal},$$

$$b^{\mathsf{T}} = \begin{bmatrix} \frac{11}{120} & \frac{27}{40} & -\frac{8}{15} & \frac{27}{40} & \frac{11}{120} \end{bmatrix},$$

$$W = \begin{bmatrix} 1 & -\frac{9}{11} & -\frac{14}{11} & -\frac{83}{297} & -\frac{39}{121} \\ -\frac{9}{11} & \frac{41}{81} & \frac{22}{27} & \frac{209}{729} & -\frac{83}{297} \\ -\frac{14}{11} & \frac{22}{27} & \frac{13}{8} & \frac{22}{27} & -\frac{14}{11} \\ -\frac{83}{297} & \frac{209}{729} & \frac{22}{27} & \frac{41}{81} & -\frac{9}{11} \\ -\frac{39}{121} & -\frac{83}{297} & -\frac{14}{11} & -\frac{9}{11} & 1 \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{6} & \frac{1}{6} & 0 & 0 & 0 \\ \frac{5}{24} & \frac{1}{8} & \frac{1}{6} & 0 & 0 \\ \frac{19}{162} & \frac{13}{27} & -\frac{8}{81} & \frac{1}{6} & 0 \\ \frac{4}{33} & \frac{19}{22} & -\frac{40}{33} & \frac{27}{22} & 0 \end{bmatrix}.$$

To recover the vectors α, β, γ from (7.6 f), form the two symmetric matrices $\widehat{W} = \widehat{T}^{\mathsf{T}} W \widehat{T}$, $\widetilde{W} = \widetilde{T}^{\mathsf{T}} W \widetilde{T}$, where

$$\widehat{T} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \\ 0 & 0 \\ 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 \end{bmatrix} \qquad \widetilde{T} = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 \\ 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 \end{bmatrix}.$$

Note that each of $\widehat{W}$ and $\widetilde{W}$ is singular because, making use of (7.6 k) and (7.6 l),

$$\begin{aligned} \begin{bmatrix} b_1(1-2c_1) & b_2(1-2c_2) \end{bmatrix} \widehat{W} &= \begin{bmatrix} b_1(1-2c_1) & b_2(1-2c_2) \end{bmatrix} \widehat{T}^{\mathsf{T}} W \widehat{T} \\ &= \left(\tfrac{1}{2} b^{\mathsf{T}} - b^{\mathsf{T}} \operatorname{diag}(c)\right) W \widehat{T} \\ &= 0, \\ \begin{bmatrix} 2b_1 & 2b_2 & b_3 \end{bmatrix} \widetilde{W} &= \begin{bmatrix} 2b_1 & 2b_2 & b_3 \end{bmatrix} \widetilde{T}^{\mathsf{T}} W \widetilde{T} \\ &= b^{\mathsf{T}} W \widetilde{T} \\ &= 0. \end{aligned}$$

It is found that

$$\widehat{W} = \widehat{\alpha}\widehat{\alpha}^{\mathsf{T}}, \quad \widetilde{W} = \widetilde{\beta}\widetilde{\beta}^{\mathsf{T}} - \widetilde{\gamma}\widetilde{\gamma}^{\mathsf{T}}, \quad \widehat{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}, \quad \widetilde{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \quad \widetilde{\gamma} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix},$$

leading to

$$\widehat{W} = \begin{bmatrix} \frac{80}{121} & -\frac{80}{297} \\ -\frac{80}{297} & \frac{80}{729} \end{bmatrix}, \quad \alpha = \begin{bmatrix} \frac{4\sqrt{5}}{11} \\ -\frac{4\sqrt{5}}{27} \\ 0 \\ \frac{4\sqrt{5}}{27} \\ -\frac{4\sqrt{5}}{11} \end{bmatrix}, \quad \widetilde{W} = \begin{bmatrix} \frac{41}{121} & -\frac{163}{297} & -\frac{14}{11} \\ -\frac{163}{297} & \frac{289}{729} & \frac{22}{27} \\ -\frac{14}{11} & \frac{22}{27} & \frac{13}{8} \end{bmatrix}.$$

Choose $\widetilde{\gamma}$ by the conditions that $\widetilde{W}+\widetilde{\gamma}\widetilde{\gamma}^{\mathsf{T}}$ has rank 1 and that $b_1\widetilde{\gamma}_1^2+b_2\widetilde{\gamma}_2^2+\frac{1}{2}b_3\widetilde{\gamma}_3^2=0$. This gives

$$\beta=\left(\tfrac{65274\sqrt{330}-347009}{1265902}\right)^{-\frac{1}{2}}\begin{bmatrix}\frac{65274\sqrt{330}-347009}{1265902}\\ -\frac{70518\sqrt{330}+318613}{3107214}\\ -\frac{18285\sqrt{330}+162856}{460328}\\ -\frac{70518\sqrt{330}+318613}{3107214}\\ \frac{65274\sqrt{330}-347009}{1265902}\end{bmatrix},$$

$$\gamma=\left(\tfrac{5934\sqrt{330}-70541}{115082}\right)^{-\frac{1}{2}}\begin{bmatrix}\frac{5934\sqrt{330}-70541}{115082}\\ -\frac{23506\sqrt{330}-462231}{1035738}\\ -\frac{18285\sqrt{330}-423016}{460328}\\ -\frac{23506\sqrt{330}-462231}{1035738}\\ \frac{5934\sqrt{330}-70541}{115082}\end{bmatrix}.$$

This completes the construction of the method G6245.

7.7 Implementation

For practical use the method is first transformed to real coefficients so that (A,U,B,V) is replaced by $(A,UT,T^{-1}B,T^{-1}VT)$, where

$$T=\begin{bmatrix}1&0&0&0\\0&1&\mathrm{i}&0\\0&1&-\mathrm{i}&0\\0&0&0&1\end{bmatrix},\quad T^{-1}=\begin{bmatrix}1&0&0&0\\0&\frac{1}{2}&\frac{1}{2}&0\\0&-\frac{\mathrm{i}}{2}&\frac{\mathrm{i}}{2}&0\\0&0&0&1\end{bmatrix}.$$

For the remainder of this section, the notations (A,U,B,V) will refer to the transformed matrices with real coefficients. That is,

$$\left[\begin{array}{c}A\ U\\ B\ V\end{array}\right]=\left[\begin{array}{ccccc|cccc}0&0&0&0&0&1&\beta_1&\alpha_1&-\gamma_1\\ a_{21}&a_{22}&0&0&0&1&\beta_2&\alpha_2&-\gamma_2\\ a_{31}&a_{32}&a_{33}&0&0&1&\beta_3&0&-\gamma_3\\ a_{41}&a_{42}&a_{43}&a_{44}&0&1&\beta_2&-\alpha_2&-\gamma_2\\ a_{51}&a_{52}&a_{53}&a_{54}&0&1&\beta_1&-\alpha_1&-\gamma_1\\ \hline b_1&b_2&b_3&b_2&b_1&1&0&0&0\\ b_1\alpha_1&b_2\alpha_2&0&-b_2\alpha_2&-b_1\alpha_1&0&0&-1&0\\ b_1\beta_1&b_2\beta_2&b_3\beta_3&b_2\beta_2&b_1\beta_1&0&1&0&0\\ b_1\gamma_1&b_2\gamma_2&b_3\gamma_3&b_2\gamma_2&b_1\gamma_1&0&0&0&-1\end{array}\right],$$

with G transformed to

$$G = T^* \operatorname{diag}(1, -\tfrac{1}{2}, -\tfrac{1}{2}, 1) T = \operatorname{diag}(1, -1, -1, 1).$$

To satisfy the technical requirements of order six, a starting method needs to be supplied such that

$$\mathcal{M}_h \circ \mathcal{S}_h = \mathcal{S}_h \circ \mathcal{E}_h + O(h^7),$$

where $\mathcal{S}_h : \mathbb{R}^N \to (\mathbb{R}^N)^4$ is the mapping corresponding to the starting method, $\mathcal{M}_h : (\mathbb{R}^N)^4 \to (\mathbb{R}^N)^4$ is the mapping corresponding to a single step of the main method and $\mathcal{E}_h : \mathbb{R}^N \to \mathbb{R}^N$ is the mapping corresponding to the flow through a stepsize h. In the case of this method, the first component of $\mathcal{S}_h$ should be chosen as the identity mapping.

Trivial and enhanced starting methods

In our first experiments, the remaining components were, for simplicity, set to the zero mappings. Although this worked well it is possible to gain improvements with little additional effort.

Let R_h be a given starting method for the non-principal values. Calculate $y^{[0]}$. $\widehat{y}^{[0]} = R_h y_0$. Use the method to find $y^{[1]}$. Evaluate $\widehat{y}^{[1]}$. Evaluate $R_h y_1^{[1]}$. Then evaluate $(I - V)^{-1}\big(\widehat{y}^{[1]} - R_h y_1^{[1]}\big)$. Add this to $\widehat{y}^{[0]}$ to get R_h^+.

7.8 Numerical simulations

These experiments are intended to test the ability of the new G-symplectic method to approximately conserve the Hamiltonian in both short and long time integrations. In each case the constant stepsize is chosen to be 0.1. For the short runs, 100 steps are taken and, for the long runs, the number of steps is 10^6, with the deviation of the Hamiltonian from its initial value, $\Delta H = H(y(x)) - H(y_0)$, sampled every 1000 steps.

Test problems

The two test problems are:

$$\begin{aligned}
&\text{Simple pendulum,} && H(y) = \tfrac{1}{2} y_2^2 - \cos(y_1), \\
& && y(0) = \begin{bmatrix} \frac{5}{2} & 0 \end{bmatrix}^{\intercal}, \\
&\text{Hénon–Heiles,} && H(y) = \tfrac{1}{2}(y_3^2 + y_4^2) + \tfrac{1}{2}(y_1^2 + y_2^2) + y_1^2 y_2 - \tfrac{1}{3} y_2^3, \\
& && y(0) = \begin{bmatrix} 0 & \frac{3}{10} & \frac{9}{25} & \frac{11}{50} \end{bmatrix}^{\intercal}.
\end{aligned}$$

Tests will be given for the fourth order G4123 method with the standard Gauss method, with $s = 2$, $p = 4$, used for calibration. Tests will also be given for the sixth order G6245, calibrated against the Gauss method with $s = 3$, $p = 6$.

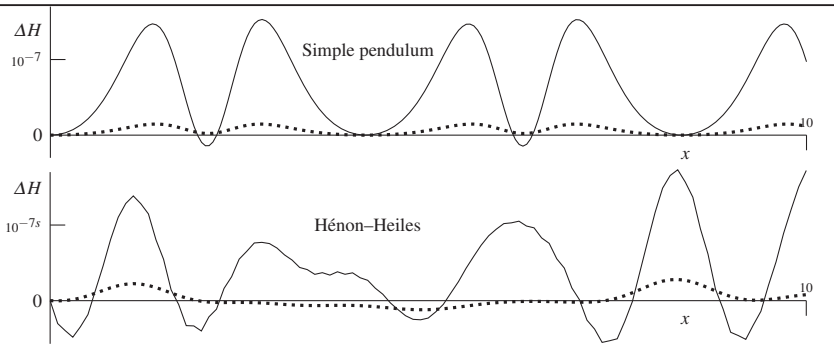


Figure 14 A comparison between the G4123 method (full line) and the 2-stage Gauss method (dotted line) for the Simple Pendulum and the Hénon–Heiles problems. For each method, $h = 0.1$, $n = 100$ (lower figure)

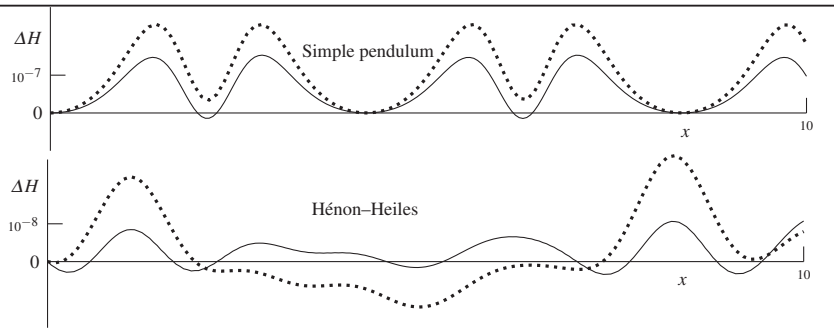


Figure 15 Deviation of H from its initial value for the simple pendulum and the Hénon–Heiles problems, using G4123, with $h = 0.05$ and $n = 200$, compared with Gauss order 4 (dotted line), with $h = 0.1$ and $n = 100$

Fourth order methods

The first experiment presented here, with results shown in Figure 14, compares G4123 with the 2 stage Gauss method. These results are misleading because the G4123 method is less costly by a factor of at least two, assessed in terms of the effort expended on the iterations. To obtain a more realistic comparison, the simulations are repeated with $h = 0.05$ and $n = 200$ steps for G4123 compared with $h = 0.1$ and $n = 100$ for the Gauss method. These adjusted results are shown in Figure 15. Further comparisons between G4123 with $h = 0.05$ and the Gauss order 4 method with $h = 0.1$ are presented in Figure 16. In this case, the results are for the time interval $[0, 10^5]$. Based on this experiment, there seems to be no reason to discount the advantages of G4123.

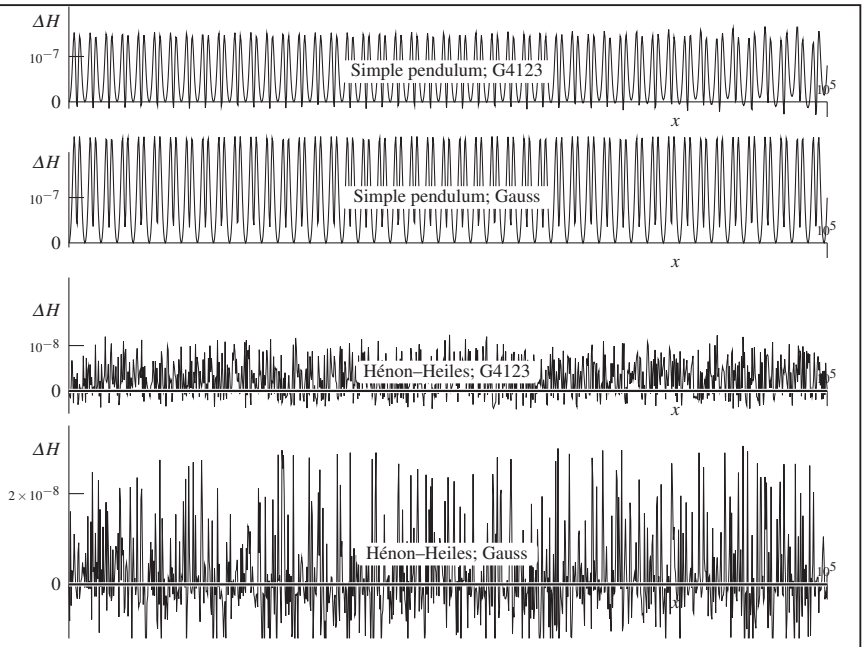


Figure 16 Deviation of H from its initial value for the simple pendulum and Hénon–Heiles problems using G4123 ($n = 2 \times 10^6$ time steps and $h = 0.05$), with Gauss order 4 ($n = 10^6$, $h = 0.1$, dotted line) shown for comparison

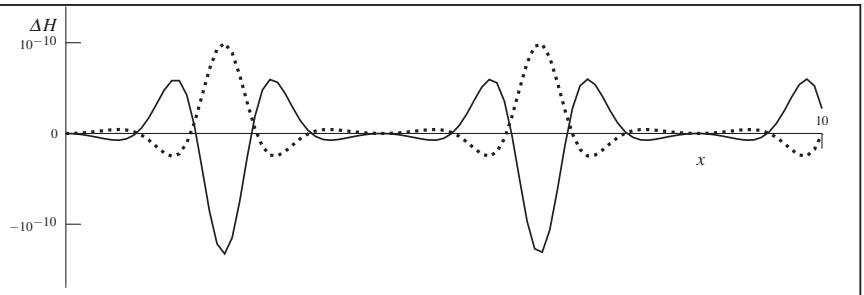


Figure 17 Deviation of H from initial value for the simple pendulum using G6245, with Gauss order 6 (dotted line) shown for comparison

Experiments with G6245

The results of the short term simulation for the simple pendulum are given in Figure 17 and for the Hénon–Heiles problem in Figure 18 (p. 280).

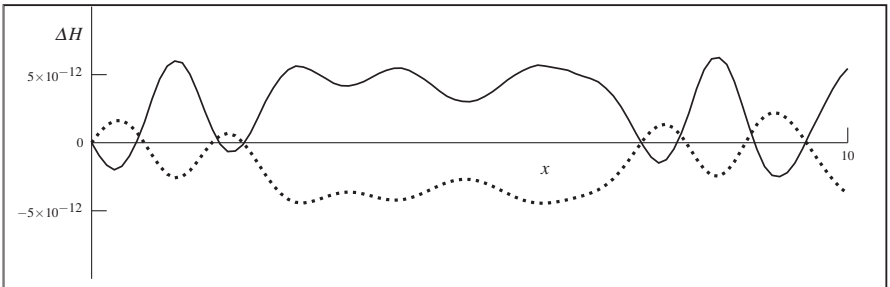


Figure 18 Deviation of H from initial value for the Hénon–Heiles problem using G6245, with Gauss order 6 method (dotted line) shown for comparison

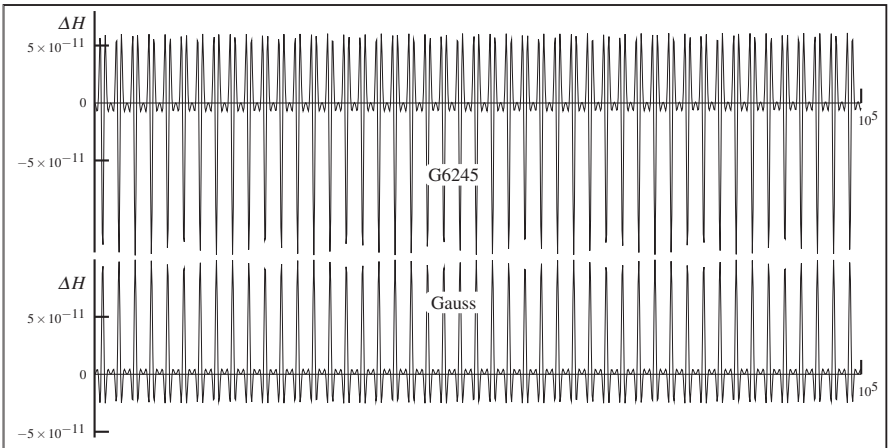


Figure 19 Deviation of H from initial value for the simple pendulum using G6245 (upper figure), with the Gauss order 6 method (lower figure) shown for comparison

Integrations through only 100 time steps are not demanding but we can at least conclude that the new sixth order method has comparable conservation behaviour as for the much more expensive fully implicit Gauss method of the same order.

For more stringent tests, the same problems can be attempted with the same methods but for 10^6 time steps. These are given in Figure 19 for the simple pendulum and Figure 20 for the Hénon–Heiles problem.

Promising results to be treated with caution

The result of this and other numerical tests have given very encouraging results for millions of time steps and it is tempting to assume that there is no real limit as to how far stable behaviour would continue.

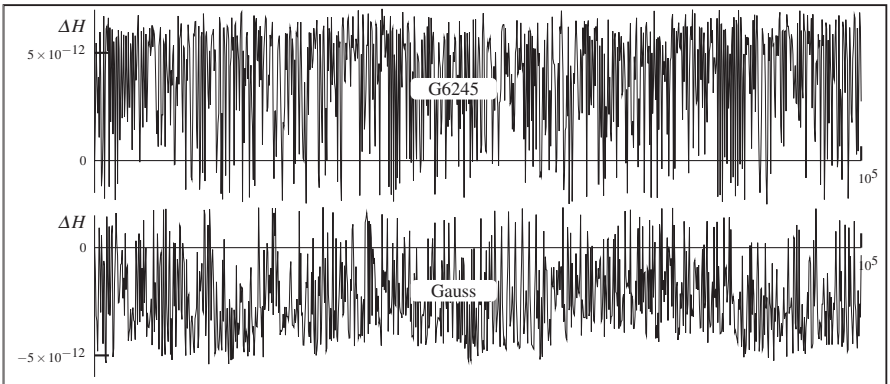


Figure 20 Deviation of H from initial value for the Hénon–Heiles problem using G6245 (upper figure), with the Gauss order 6 method (lower figure) shown for comparison

However, this is an unrealistic expectation because, from the analysis in [40] (D'Ambrosio, Hairer, 2014), parasitism will eventually take over and destroy the integrity of the numerical results.

7.9 Energy preserving methods

We will consider Poisson problems $y' = S\nabla H$ in the case that the skew matrix S is constant. This includes Hamiltonian problems for which the dimension is even and $S = J^{\intercal}$.

For a given $\mathrm{t} = [\mathrm{t}_1 \mathrm{t}_2 \cdots \mathrm{t}_m]$, define

$$\mathsf{h}(\mathrm{t}) = H^{(m)}\mathbf{F}(\mathrm{t}_1)\mathbf{F}(\mathrm{t}_2)\cdots\mathbf{F}(\mathrm{t}_m),$$

with $\mathsf{h}(\tau) = H$. No meaning is given to $\mathsf{h}(\varnothing)$.

Examples to order 4 are

$$\begin{aligned}
\mathsf{h}(\bullet) &= \mathsf{h}(\mathbf{t}_1) &&= H,\\
\mathsf{h}(\bullet\!\!-\!\!\bullet) &= \mathsf{h}(\mathbf{t}_2) &&= H'\mathbf{f},\\
\mathsf{h}(\vee) &= \mathsf{h}(\mathbf{t}_3) &&= H''\mathbf{f}\mathbf{f},\\
\mathsf{h}(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet) &= \mathsf{h}(\mathbf{t}_4) &&= H'\mathbf{f}'\mathbf{f},\\
\mathsf{h}(\curlyvee) &= \mathsf{h}(\mathbf{t}_5) &&= H'''\mathbf{f}\mathbf{f}\mathbf{f},\\
\mathsf{h}(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet) &= \mathsf{h}(\mathbf{t}_6) &&= H''\mathbf{f}\mathbf{f}'\mathbf{f},\\
\mathsf{h}(\curlyvee) &= \mathsf{h}(\mathbf{t}_7) &&= H'\mathbf{f}''\mathbf{f}\mathbf{f},\\
\mathsf{h}(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet) &= \mathsf{h}(\mathbf{t}_8) &&= H'\mathbf{f}'\mathbf{f}'\mathbf{f}.
\end{aligned}$$

Exercise 60 Find $\mathsf{h}(\mathbf{t}_{13})$.

Analogously to the infinite row vector $\mathbf{B}_h$, indexed on $\mathrm{T}^{\#}$, define

$$\mathsf{H}_h = \left[\begin{array}{cccccc} \mathsf{h}(\mathbf{t}_1) & h\mathsf{h}(\mathbf{t}_2) & h^2\tfrac{1}{2}\mathsf{h}(\mathbf{t}_3) & h^2\mathsf{h}(\mathbf{t}_4) & \cdots \end{array} \right],$$

with typical term

$$\mathsf{H}_h(\mathrm{t}) = h^{|\mathrm{t}|-1}\sigma(\mathrm{t})^{-1}\mathsf{h}(\mathrm{t}).$$

Because there is no term corresponding to y_0, we use a truncated version of Λ, denoted by $\widehat{\Lambda}$, with the first row and column deleted and the remaining rows and columns indexed by $\mathrm{T}\times\mathrm{T}$.

Theorem 7.9A Let $y_1 = (\mathbf{B}_h y_0)a$, where $a \in \mathsf{B}$. Then

$$\mathsf{H}_h y_1 = (\mathsf{H}_h y_0)\widehat{\Lambda}(a).$$

Proof. Use Theorem 3.3A (p. 110), applied to

$$H^{(n)}\big(y_0 + \sum_{\mathrm{t}\in\mathrm{T}} \sigma(\mathrm{t})^{-1}h^{|\mathrm{t}|}\mathbf{F}(\mathrm{t})\big).$$

In particular

Corollary 7.9B For a Runge–Kutta method with elementary weights $\mathrm{t}\mapsto\Phi(\mathrm{t})$,

$$\mathsf{H}_h(y_1) = (\mathsf{H}_h y_0)\Phi D.$$

For conservation of energy, we would need $H(y_1) = H(y_0)$ and our aim now is to find conditions for this.

Operating with S on $H^{(n+1)}$

Expressing the $n+1$-fold derivative of $H^{(n+1)}$ and then operating on this with the matrix S requires some care because we really need to combine the contravariant derivative ∇H with an n-fold covariant derivative. The operation that needs to be performed gives a result

$$\frac{\partial^n}{\partial y^n}S\nabla H = S\frac{\partial^n}{\partial y^n}\nabla H, \tag{7.9 a}$$

with component $i \in \{1,2,\ldots,N\}$ of the n-linear operator in (7.9 a) acting on $v_1, v_2, \ldots v_n$ given by

$$\sum_{j=1}^{N} S^{ij} H_{j,k_1 k_2 \cdots k_n} v_1^{k_1} v_2^{k_2} \cdots v_2^{k_2}.$$

In the remainder of this chapter, we will denote by $\dot{S}$, the operator which acts on $H^{(n)}$ to produce the quantity given by (7.9 a). That is

$$\dot{S} H^{(n)} := S \frac{\partial^n}{\partial y^n} \nabla H.$$

Taylor expansion for $H(\mathbf{B}_h y_0 a)$

Because of the special form of Poisson problems, we have

Theorem 7.9C Given $\mathsf{t}, \mathsf{t}' \in \mathsf{T}$,

$$\mathsf{h}(\mathsf{t} * \mathsf{t}') = -\mathsf{h}(\mathsf{t}' * \mathsf{t})$$

Proof. Let $\mathsf{t} = [\mathsf{t}_1 \mathsf{t}_2 \cdots \mathsf{t}_m]$, $\mathsf{t}' = [\mathsf{t}'_1 \mathsf{t}'_2 \cdots \mathsf{t}'_m]$ and write $\mathbf{F}_i := \mathbf{F}(\mathsf{t}_i)$, $\overline{\mathbf{F}}_i := \mathbf{F}(\mathsf{t}'_i)$. Then, assuming the summation convention,

$$\mathsf{h}(\mathsf{t} * \mathsf{t}') = H_{i j_1 j_2 \cdots j_m} \mathbf{F}_1^{j_1} \mathbf{F}_2^{j_2} \cdots \mathbf{F}_m^{j_m} S^{ik} H_{k \ell_1 \ell_2 \cdots \ell_n} \overline{\mathbf{F}}_1^{\ell_1} \overline{\mathbf{F}}_2^{\ell_1} \cdots \overline{\mathbf{F}}_n^{\ell_n},$$
$$\mathsf{h}(\mathsf{t}' * \mathsf{t}) = H_{k \ell_1 \ell_2 \cdots \ell_n} \overline{\mathbf{F}}_1^{\ell_1} \overline{\mathbf{F}}_2^{\ell_1} \cdots \overline{\mathbf{F}}_n^{\ell_n} S^{ki} H_{i j_1 j_2 \cdots j_m} \mathbf{F}_1^{j_1} \mathbf{F}_2^{j_2} \cdots \mathbf{F}_m^{j_m}$$

and the result follows because $S^{ik} = -S^{ki}$. □

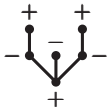
By applying this result, step-by-step, through all trees in a free tree class, we find that

$$\mathsf{h}(\mathsf{t}) = \pm \mathsf{h}(\mathsf{t}') \quad \text{whenever} \quad \mathsf{t} \sim \mathsf{t}',$$

with the actual sign determined by the parity of the number of steps between the two roots.

A non-superfluous example

We will give two examples. The first is given by a diagram in which a representative tree has a + sign attached to the root and an appropriate sign attached to the other vertices which would apply if this vertex became the root.



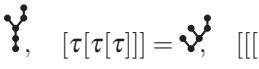
(7.9 b)

This diagram indicates that

$$\mathsf{h}([[[\tau[\tau]]]]) = -\mathsf{h}([\tau[\tau[\tau]]]) = -\mathsf{h}([[[\tau]^2]]) = \mathsf{h}([\tau[\tau]^2]), \tag{7.9 c}$$

where we recall the notation

$$[[[\tau[\tau]]]] = \quad, \quad [\tau[\tau[\tau]]] = \quad, \quad [[[\tau]^2]] = \quad, \quad [\tau[\tau]^2] = \quad$$



In the Taylor expansion of $H(y_1)$, given in Theorem 7.9A, four terms involving the trees arising in the present discussion will be

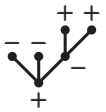
$$h^5\Big(\tfrac{1}{2}(\Phi\mathrm{D})\big([\tau[\tau]^2]\big)\mathsf{h}\big([\tau[\tau]^2]\big)+(\Phi\mathrm{D})\big([[[\tau[\tau]]]]\big)\mathsf{h}\big([[[\tau[\tau]]]]\big)$$
$$+(\Phi\mathrm{D})\big([\tau[\tau[\tau]]]\big)\mathsf{h}\big([\tau[\tau[\tau]]]\big)+\tfrac{1}{2}(\Phi\mathrm{D})\big([[[\tau]^2]]\big)\Big)\big(\mathsf{h}([[[\tau]^2]])\big). \quad (7.9\,\mathrm{d})$$

Because of (7.9 c), (7.9 d) collapses to a single term

$$\tfrac{1}{2}h^5\Big((\Phi\mathrm{D})\big([\tau[\tau]^2]\big)+2(\Phi\mathrm{D})\big([[[\tau[\tau]]]]\big)$$
$$-2(\Phi\mathrm{D})\big([\tau[\tau[\tau]]]\big)-(\Phi\mathrm{D})\big([[[\tau]^2]]\big)\Big)\mathsf{h}\big([\tau[\tau]^2]\big)$$
$$=\tfrac{1}{2}h^5\Big(\Phi(\mathbf{t}_1)\Phi(\mathbf{t}_2)^2+2\Phi(\mathbf{t}_{15})-2\Phi(\mathbf{t}_1)\Phi(\mathbf{t}_6)-\Phi(\mathbf{t}_{13})\Big)\mathsf{h}(\mathbf{t}_{22}).$$

An example based on a superfluous tree

Now consider the example of a superfluous tree, again with signs attached to the vertices, in accordance with Theorem 7.9C and its consequences.



If we now list the corresponding terms in the Taylor expansion of $H(y_1)$, it can be seen that each of these terms disappears because, for any tree t in the equivalence class, any term $(\Phi\mathrm{D})(\mathrm{t})\mathsf{h}(\mathrm{t})$ is matched by a corresponding $-(\Phi\mathrm{D})(\mathrm{t})\mathsf{h}(\mathrm{t})$.

A sufficient condition for energy preservation

Theorem 7.9D An integration method is energy preserving if for every similarity class the following is true based on a representative t' in this class. For every tree $\mathrm{t}\sim\mathrm{t}'$ let $n(\mathrm{t})$ be the number of steps from the root of t' to the root of t. Then

$$\sum_{\mathrm{t}\sim\mathrm{t}'}(-1)^{n(\mathrm{t})}\Phi\mathrm{D}(\mathrm{t})=0. \quad (7.9\,\mathrm{e})$$

The Average Vector Field method

The Average Vector Field method [80] (Quispel, McLachlan, 2008) is an integration method $(A, b^{\intercal}, c)$ on the index set $[0,1]$ with

$$\begin{aligned}(A\phi)_\xi &= \xi \int_0^1 \phi_\eta \,\mathrm{d}\,\eta,\\ b^{\intercal}\phi &= \int_0^1 \phi_\eta \,\mathrm{d}\,\eta,\\ c_\xi &= \xi.\end{aligned}$$

The stages and output in a step are given by

$$Y_\xi = y_0 + \xi h \int_0^1 f(Y_\eta)\,\mathrm{d}\,\eta, \tag{7.9 f}$$

$$y_1 = y_0 + h \int_0^1 f(Y_\eta)\,\mathrm{d}\,\eta, \tag{7.9 g}$$

from which it follows that $Y_\xi = (1-\xi)y_0 + \xi y_1$ and the method can be written in the compact form

$$y_1 = y_0 + h \int_0^1 f\big((1-\xi)y_0 + \xi y_1\big)\,\mathrm{d}\,\xi.$$

Energy preservation of the Average Vector Field method

Let $\Phi(\mathrm{t})$, $\mathrm{t} \in \mathrm{T}$, be the elementary weights, that is, the B-series coefficients for y_1.

The values of $\Phi(\mathrm{t})$

The B-series coefficients for stage ξ are $\xi\Phi(\mathrm{t})$ and hence,

$$\begin{aligned}\Phi(\mathrm{t}) &= \int_0^1 \xi^m \prod_{i=1}^{m} \Phi(\mathrm{t}_i)\,\mathrm{d}\,\xi\\ &= \frac{1}{m+1}\prod_{i=1}^{m}\Phi(\mathrm{t}_i).\end{aligned}$$

It now follows that

> Theorem 7.9E For the Average Vector Field method, $\Phi(\mathrm{t})^{-1}$ is the product over each vertex of t of (1+ the number of children of this vertex). Furthermore, $(\Phi\mathrm{D})(\mathrm{t})^{-1}$ is the product over each vertex of t, except for the root, of (1+ the number of children of this vertex).

Approaching this result in a different way by writing t in the form (V,E,r), where V and E are fixed and r runs through all members of V, makes it possible to use the criterion for energy preservation given by Theorem 7.9D.

Theorem 7.9F The AVF method is energy preserving.

Proof. Given a free tree (V,E) and $r \in V$, define valency(r) as the number of members $\{x,y\} \in E$ such that $x = r$ or $y = r$. Then from Theorem 7.9E,

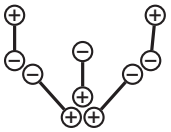
$$(aD)(\mathrm{t})^{-1} = \frac{\text{valency}(r)}{\prod_{j\in V} \text{valency}\, j}.$$

It now follows that the coefficient in (7.9 e) is

$$\frac{\sum_{i\in V} \pm \text{valency}(i)}{\prod_{\in V} \text{valency}(j)}, \tag{7.9 h}$$

where the sign $\pm$ alternates along each edge. There is now a cancellation of the terms in the numerator of (7.9 h) □

To illustrate the cancellation that takes place in this proof, consider the following diagram, based on (7.9 b).



The number of edges ending at any vertex is equal to the valency and these edges also meet a vertex with the opposite sign. Hence the cancellation.

An alternative proof

The second proof is a special case of the proof of Theorem 7.9E, below, and should be looked on as an introduction to that result.

Proof. We have

$$\begin{aligned} H(y_1) - H(y_0) &= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}\eta} H(Y_\eta)\,\mathrm{d}\eta \\ &= \int_0^1 H'(Y_\eta) \frac{\mathrm{d}}{\mathrm{d}\eta} Y_\eta\,\mathrm{d}\eta \\ &= \int_0^1 \big(\nabla H_\eta\big)^\intercal h\Big(\int_0^1 S\nabla H_\xi\,\mathrm{d}_\xi\Big)\,\mathrm{d}\eta \\ &= h\Big(\int_0^1 \nabla H_\eta\,\mathrm{d}\eta\Big)^\intercal S\Big(\int_0^1 \nabla H_\eta\,\mathrm{d}\eta\Big) \\ &= 0, \end{aligned}$$

because of the skew-symmetry of S. □

A generalization of the AVF method

We consider the method defined by replacing (7.9 f) and (7.9 g) by

$$Y_\xi = y_0 + h\int_0^1 \Psi(\xi,\eta) f(Y_\eta)\,\mathrm{d}\eta, \tag{7.9 i}$$

$$y_1 = Y_1, \tag{7.9 j}$$

where the partial derivative $\widetilde{\Psi}(\xi,\eta) := \partial\Psi(\xi,\eta)/\partial\xi$ exists and is continuous for $\xi,\eta \in [0,1]$.

Theorem 7.9G If $\widetilde{\Psi}(\xi,\eta) = \widetilde{\Psi}(\eta,\xi)$, then the method defined by (7.9 i) and (7.9 j) preserves energy.

Proof. Given $\varepsilon > 0$, by the Weierstrass approximation theorem there exists a polynomial in two variables Π such that $\widetilde{\Psi}(\xi,\eta) = \Pi(\xi,\eta) + E(\xi,\eta)$, with $|E(\xi,\eta)| \le \varepsilon$ for $\xi,\eta \in [0,1]$. Without loss of generality (because $\Pi(\xi,\eta)$ can be replaced by $\big(\Pi(\xi,\eta)+\Pi(\eta,\xi)\big)/2$), assume that $\Pi(\xi,\eta) = \Pi(\eta,\xi)$.

Let

$$\Pi(\xi,\eta) = \begin{bmatrix} 1 & \xi & \xi^2 & \cdots & \xi^{n-1} \end{bmatrix}^{\mathsf{T}} M \begin{bmatrix} 1 & \eta & \eta^2 & \cdots & \eta^{n-1} \end{bmatrix},$$

where M is an $n \times n$ symmetric matrix. From standard decomposition results for symmetric matrices, there exists an $m \times n$ matrix N and a diagonal $m \times m$ matrix D, such that $M = N^{\mathsf{T}} D N$. It then follows that $\Pi(\xi,\eta) = \sum_{i=1}^m d_i\,\varpi_i(\xi)\varpi_i(\eta)$, where the polynomial ϖ_i has coefficients given by row number i in N. We can now write

$$\widetilde{\Psi}(\xi,\eta) = \sum_{i=1}^m d_i \varpi_i(\xi)\varpi_i(\eta) + E(\xi,\eta)$$

and we obtain

$$\begin{aligned}
&H(y_1) - H(y_0)\\
&= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}\xi} H(Y_\xi)\,\mathrm{d}\xi\\
&= \int_0^1 H'(Y_\xi)\frac{\mathrm{d}}{\mathrm{d}\xi} Y_\xi\,\mathrm{d}\xi\\
&= \int_0^1 \big(\nabla H_\xi\big)^{\mathsf{T}} h\Big(\int_0^1 \widetilde{\Psi}(\xi,\eta) S \nabla H_\eta\,\mathrm{d}\eta\Big)\,\mathrm{d}\xi\\
&= \int_0^1 \big(\nabla H_\xi\big)^{\mathsf{T}} h\Big(\int_0^1 \Big(\sum_{i=1}^m d_i \varpi_i(\xi)\varpi_i(\eta) + E(\xi,\eta)\Big) S \nabla H_\eta\,\mathrm{d}\eta\Big)\,\mathrm{d}\xi.
\end{aligned} \tag{7.9 k}$$

The coefficient of hd_i in (7.9 k) is

$$\int_0^1 (\nabla H_\xi)^\intercal \Big(\int_0^1 \varpi_i(\xi)\varpi_i(\eta) S \nabla H_\eta \,\mathrm{d}\eta \Big) \,\mathrm{d}\xi$$
$$= \Big(\int_0^1 \varpi_i(\xi) \nabla H_\xi \,\mathrm{d}\xi \Big) S \Big(\int_0^1 \varpi_i(\eta) \nabla H_\eta \,\mathrm{d}\eta \Big),$$

which vanishes because of skew-symmetry of S. Hence,

$$\begin{aligned} &\big| H(y_1) - H(y_0) \big| \\ &\le h \left| \int_0^1 \int_0^1 E(\xi,\eta) (\nabla H_\xi)^\intercal S \nabla H_\eta \,\mathrm{d}\eta \,\mathrm{d}\xi \right| \\ &\le \varepsilon h \int_0^1 \int_0^1 \left| (\nabla H_\xi)^\intercal S \nabla H_\eta \right| \mathrm{d}\eta \,\mathrm{d}\xi. \end{aligned}$$

Because this can be made arbitrarily small, $H(y_1) = H(y_0)$. □

A fourth order method

We will construct an energy preserving method based on a polynomial

$$\Psi(\xi,\eta) = 2a\xi^2\eta + 2b\xi\eta + b\xi^2 + (1-a-2b)\xi,$$

where the coefficients are chosen subject to the symmetry of $\partial\Psi/\partial\xi$ and the consistency condition $\Phi(\mathbf{t}_1) = \int_0^1 \Psi(1,\eta)\,\mathrm{d}\eta = 1$. Evaluation of the remaining elementary differentials up to order 4 give

$$\begin{aligned} \Phi(\mathbf{t}_2) &= \tfrac{1}{2}, \\ \Phi(\mathbf{t}_3) &= \tfrac{1}{3}, \\ \Phi(\mathbf{t}_4) &= \tfrac{1}{4} - \tfrac{1}{36}a + \tfrac{1}{36}(a+b)^2, \\ \Phi(\mathbf{t}_5) &= \tfrac{1}{4}, \\ \Phi(\mathbf{t}_6) &= \tfrac{1}{6} - \tfrac{1}{72}a + \tfrac{1}{360}(a+b)(6a+5b) - \tfrac{1}{360}(a+b)^3, \\ \Phi(\mathbf{t}_7) &= \tfrac{1}{6} - \tfrac{1}{36}a + \tfrac{1}{180}(a+b)(6a+5b) - \tfrac{1}{180}(a+b)^3, \\ \Phi(\mathbf{t}_8) &= \tfrac{1}{8} - \tfrac{1}{36}a + \tfrac{1}{36}(a+b)^2, \end{aligned}$$

and to obtain order 4, by requiring that $\Phi(\mathrm{t}) = 1/\mathrm{t}!$, up to this order we need to satisfy

$$\begin{aligned} (a+b)^2 - a &= 3, \\ -(a+b)^3 + (a+b)(6a+5b) - 5a &= 15, \end{aligned}$$

with solution $-a = b = 3$.

Summary of Chapter 7

Although it has not been possible to survey all aspects of the burgeoning subject of Geometric Integration, symplectic Runge–Kutta methods and their generalization to general linear methods are introduced to the extent that their main properties are studied and explained. It is perhaps surprising that G-symplectic methods perform well over millions of time steps, even though, according to [40], they will eventually fail.

In Section 7.9, energy preserving methods were introduced, based on integration methods, in the sense of Chapter 4, also known as Continuous Stage Runge–Kutta methods.

Teaching and study notes

The following books and articles are essential reading, and provide a starting point for further studies on Geometric Integration.

Cohen, D. and Hairer, E. *Linear energy-preserving integrators for Poisson systems*, (2011) [31]

Hairer, E. *Energy-preserving variant of collocation methods* (2010) [48]

Hairer, E., Lubich, C. and Wanner, G. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, (2006) [49]

Iserles, A., Munthe-Kaas, H.Z., Nørsett, S.P. and Zanna, A. *Lie-group methods*, (2000) [62]

Miyatake, Y. *An energy-preserving exponentially-fitted continuous stage Runge–Kutta method* (2014) [73]

McLachlan, R. and Quispel, G. *Six lectures on the geometric integration of ODEs* (2001) [71]

Sanz-Serna, J.M. and Calvo, M.P. *Numerical Hamiltonian Problems*, (1994) [85]

Projects

Project 26 Derive a method similar to G6245 but with $c_2 = \frac{1}{10}$.

Project 27 Consider the consequence of replacing (7.6 e) (p. 272) by $V = \mathrm{diag}(1, \exp(\mathrm{i}\theta), \exp(-\mathrm{i}\theta), -1)$, for $0 < \theta < \pi$, in the G6245 method.