Symmetric multistep methods for constrained Hamiltonian systems

Console, Paola; Hairer, Ernst; Lubich, Christian Viktor

How to cite

CONSOLE, Paola, HAIRER, Ernst, LUBICH, Christian Viktor. Symmetric multistep methods for constrained Hamiltonian systems. In: Numerische Mathematik, 2013, vol. 124, n° 3, p. 517–539. doi: 10.1007/s00211-013-0522-z

This publication URL: https://archive-ouverte.unige.ch//unige:114859
Publication DOI: 10.1007/s00211-013-0522-z

Numer. Math. (2013) 124:517–539
DOI 10.1007/s00211-013-0522-z

Numerische Mathematik

Symmetric multistep methods for constrained Hamiltonian systems

Paola Console · Ernst Hairer · Christian Lubich

Received: 24 April 2012 / Revised: 14 November 2012 / Published online: 15 February 2013

Abstract A method of choice for the long-time integration of constrained Hamiltonian systems is the Rattle algorithm. It is symmetric, symplectic, and nearly preserves the Hamiltonian, but it is only of order two and thus not efficient for high accuracy requirements. In this article we prove that certain symmetric linear multistep methods have the same qualitative behavior and can achieve an arbitrarily high order with a computational cost comparable to that of the Rattle algorithm.

Mathematics Subject Classification (2010) 65L06 · 65L80 · 65P10

1 Introduction

The motion of mechanical systems is often constrained in the position coordinates (e.g., rigid body motion, frozen bonds in molecular dynamics). This typically leads to differential-algebraic equations of the form

$$\begin{aligned} M\ddot{q} &= -\nabla U(q) - G(q)^{\mathsf{T}}\lambda \\ 0 &= g(q), \end{aligned} \tag{1}$$

P. Console · E. Hairer (✉)
Dept. de Mathématiques, Univ. de Genève, 1211 Genève 4, Switzerland
e-mail: Ernst.Hairer@unige.ch

P. Console
e-mail: Paola.Console@unige.ch

C. Lubich
Mathematisches Institut, Universität Tübingen, Auf der Morgenstelle, 72076 Tübingen, Germany
e-mail: Lubich@na.uni-tuebingen.de

where $q \in \mathbf{R}^d$ is the vector of position coordinates, M is a positive definite mass matrix, $U(q)$ is a smooth real potential, $g(q) \in \mathbf{R}^m$ (with $m < d$) collects the constraints, and $G(q) = g'(q)$ is the matrix of partial derivatives. The term containing the Lagrange multiplier $\lambda \in \mathbf{R}^m$ forces the solution to satisfy the algebraic constraint. In addition to $g(q) = 0$ every solution of (1) also satisfies the differentiated relation $\frac{d}{dt} g(q) = G(q)\dot{q} = 0$. Initial values $q(0) = q_0, \dot{q}(0) = \dot{q}_0$ are said to be consistent if they satisfy both relations $g(q_0) = 0$ and $G(q_0)\dot{q}_0 = 0$. A second differentiation of the constraint leads to $\frac{\partial^2}{\partial q^2} g(q)(\dot{q}, \dot{q}) + G(q)\ddot{q} = 0$ which, after insertion of (1), permits to express the Lagrange multiplier λ in terms of q and $\dot{q}$, provided that the matrix

$$G(q)M^{-1}G(q)^{\mathsf{T}} \qquad \text{is invertible} \tag{2}$$

along the solution. This will be assumed throughout this article. It implies that the differential-algebraic equation is of index 3.

Introducing the momentum $p = M\dot{q}$, the problem is seen to be Hamiltonian with total energy

$$H(q, p) = \frac{1}{2}\, p^{\mathsf{T}} M^{-1} p + U(q). \tag{3}$$

Elimination of the Lagrange multiplier λ from the system yields a differential equation on the manifold

$$\mathcal{M} = \{(q, p)\,;\; g(q) = 0,\; G(q)M^{-1}p = 0\}. \tag{4}$$

The flow is symplectic on $\mathcal{M}$, and the energy $H(q, p)$ is preserved along solutions of the system. In the spirit of geometric numerical integration one is interested in numerical simulations that share these properties as far as possible.

The most natural discretization of (1) is obtained when the second derivative is replaced by a central difference. This leads to the so-called SHAKE algorithm [12]

$$\begin{aligned} q_{n+1} - 2q_n + q_{n-1} &= -h^2 M^{-1}\big(\nabla U(q_n) + G(q_n)^{\mathsf{T}}\lambda_n\big) \\ 0 &= g(q_{n+1}). \end{aligned} \tag{5}$$

The momentum approximation is given by $p_n = M(q_{n+1} - q_{n-1})/2h$ and does not enter the recursion (5). In general $G(q_n)M^{-1}p_n \neq 0$, so that the numerical solution (q_n, p_n) does not lie on the manifold $\mathcal{M}$.

An important modification, called RATTLE [1], consists in writing the algorithm as a one-step method and to add a projection step, so that $(q_n, p_n) \in \mathcal{M}$. The algorithm is given by

$$\begin{aligned} p_{n+1/2} &= p_n - \frac{h}{2}\Big(\nabla U(q_n) + G(q_n)^{\mathsf{T}}\theta_n\Big) \\ q_{n+1} &= q_n + hM^{-1}p_{n+1/2} \\ 0 &= g(q_{n+1}) \end{aligned}$$

$$p_{n+1} = p_{n+1/2} - \frac{h}{2}\Big(\nabla U(q_{n+1}) + G(q_{n+1})^{\mathsf{T}}\mu_{n+1}\Big)$$
$$0 = G(q_{n+1})M^{-1}p_{n+1}. \tag{6}$$

It is symmetric, symplectic on the manifold $\mathcal{M}$, and convergent of order 2 (see [7, Section VII.1] for details). Eliminating the momentum variables shows that the RATTLE approximation satisfies the two-term recursion (5) of SHAKE with $\lambda_n = (\theta_n + \mu_n)/2$.

The RATTLE algorithm is an excellent geometric integrator for low accuracy requirements (such as in molecular dynamics simulations). There are a few extensions of this algorithm to higher order. An easy way is by composition methods with the RATTLE scheme as basic integrator [11]. Another extension is the partitioned Runge–Kutta method based on the Lobatto IIIA–IIIB pair. It is of order $2s - 2$ and reduces to the RATTLE algorithm for $s = 2$ [8]. The present article proposes a new extension, based on symmetric multistep methods.

The long-time behavior of symmetric linear multistep methods for unconstrained Hamiltonian systems $\ddot{q} = -\nabla U(q)$ has been studied in [6], see also [3] for their applicability to more general Hamiltonian problems. Section 2 explains how these methods can be extended to constrained systems of the form (1). The main results on their long-time behaviour, in particular, the near-preservation of the total energy and the momentum over long time intervals, are reported in Sect. 3. The construction of stable symmetric methods is discussed in Sect. 4, and the coefficients of optimal-order methods are presented for orders 4, 6, and 8. The numerical experiments of Sect. 5 illustrate the excellent long-time behaviour of the methods in agreement with the theoretical results. Rigorous proofs are based on a backward error analysis. The long-time behaviour of "smooth" numerical solutions and their preservation of energy and momentum are discussed in Sect. 6. Bounds for parasitic solution components are the topic of Sect. 7. The results of Sects. 6 and 7 are then combined to yield the main results.

2 Symmetric linear multistep methods

With the notation $f(q) = -\nabla U(q)$ for the force, linear multistep methods for differential-algebraic equations (1) are given by

$$\sum_{j=0}^{k} \alpha_j q_{n+j} = h^2 \sum_{j=0}^{k} \beta_j M^{-1}\Big(f(q_{n+j}) - G(q_{n+j})^{\mathsf{T}}\lambda_{n+j}\Big)$$
$$0 = g(q_{n+k}). \tag{7}$$

For implicit methods ($\beta_k \neq 0$) this represents a nonlinear system for (q_{n+k}, λ_{n+k}). For explicit methods ($\beta_k = 0$) we insert q_{n+k} from the first relation into the second one to obtain a nonlinear equation for λ_{n+k-1}. As soon as λ_{n+k-1} is computed, the solution approximation q_{n+k} is given explicitly. The computational cost of an explicit multistep method is thus precisely the same as that for the SHAKE algorithm.

For the study of linear multistep methods it is convenient to introduce the generating polynomials

$$\rho(\zeta) = \sum_{j=0}^{k} \alpha_j \zeta^j, \qquad \sigma(\zeta) = \sum_{j=0}^{k} \beta_j \zeta^j.$$

Throughout this article we assume that $\rho(\zeta)$ and $\sigma(\zeta)$ do not have common zeros (irreducibility). The method (7) is *stable* if all zeros of $\rho(\zeta)$ satisfy $|\zeta| \leq 1$, and if those of modulus one have a multiplicity not exceeding two. It is *consistent of order* r, if

$$\frac{\rho(\zeta)}{(\log \zeta)^2} - \sigma(\zeta) = \mathcal{O}\big((\zeta - 1)^r\big) \qquad \text{for} \quad \zeta \to 1. \tag{8}$$

In the present article we focus our interest on *symmetric* methods, which means that the coefficients satisfy

$$\alpha_j = \alpha_{k-j}, \qquad \beta_j = \beta_{k-j} \qquad \text{for all } j.$$

If a multistep method (7) is stable and symmetric, all zeros of $\rho(\zeta)$ are on the unit circle, and the order r is even. Furthermore, it follows from the irreducibility assumption that k is even (because symmetry implies for odd k that $\rho(-1) = \sigma(-1) = 0$), and that -1 cannot be a simple zero of $\rho(\zeta)$. The construction of explicit symmetric methods of optimal order will be discussed in Sect. 4 below.

An approximation of the momentum $p = M\dot{q}$ can be computed *a posteriori* by symmetric finite differences supplemented with a projection onto the manifold $\mathcal{M}$:

$$p_n = M \, \frac{1}{h} \sum_{j=-l}^{l} \delta_j q_{n+j} + h \, G(q_n)^{\mathsf{T}} \mu_n. \tag{9}$$

together with $G(q_n) M^{-1} p_n = 0$, which gives a linear system for μ_n. One typically chooses $l = k/2$, so that the approximations p_n are of the same order as q_n. This is not essential, because errors in p_n do not propagate.

Comments on the implementation The formulation (7) is a straightforward extension of the SHAKE algorithm (5). To reduce the effect of round-off we consider momentum approximations $p_{n+1/2}$, as it was proposed in RATTLE. For explicit multistep methods this yields

$$\begin{aligned} \sum_{j=0}^{k-1} \hat{\alpha}_j \, p_{n+j+1/2} &= h \sum_{j=1}^{k-1} \beta_j \Big(f(q_{n+j}) - G(q_{n+j})^{\mathsf{T}} \lambda_{n+j} \Big) \\ q_{n+k} &= q_{n+k-1} + h \, M^{-1} p_{n+k-1/2} \\ 0 &= g(q_{n+k}), \end{aligned} \tag{10}$$

where $\hat{\alpha}_j$ are the coefficients of $\rho(\zeta)/(\zeta-1) = (\zeta-1)\tilde{\rho}(\zeta)$. The approximation of the momenta becomes

$$\begin{aligned} p_n &= \sum_{j=-l}^{l-1} \hat{\delta}_j \, p_{n+j+1/2} + h\, G(q_n)^{\mathsf{T}} \mu_n \\ 0 &= G(q_n) M^{-1} p_n, \end{aligned} \tag{11}$$

where the coefficients $\hat{\delta}_j$ are given by $(\zeta-1)\sum_{j=-l}^{l-1} \hat{\delta}_j \zeta^j = \sum_{j=-l}^{l} \delta_j \zeta^j$.

3 Main results

When linear multistep methods are applied to differential-algebraic equations of index 3, symmetric formulas are typically avoided because of their notorious weak instability and the standard choice is BDF schemes. There is some research on a partitioned treatment of the force term and the Lagrange multiplier (for example [2]) such that also non-stiff integrators can be applied. However, little attention has been paid to long-time energy and momentum preservation with these integrators. This requires the use of symmetric methods. The present work shows that the suspected weak instability is not harmful for problems of the form (1) and for a special class of integrators.

For a favourable long-time behaviour we need the following properties of the generating polynomials:

$$\rho(\zeta) = 0 \text{ has only simple roots with the exception of the double root 1; all roots are on the unit circle.} \tag{12}$$

$$\sigma(\zeta) = 0 \text{ has only simple non-zero roots; all non-zero roots are on the unit circle.} \tag{13}$$

Symmetry of the method together with condition (12) is essential for good long-time behaviour in unconstrained problems (see [6]), and condition (13) is familiar from the convergence analysis of multistep methods for index-3 differential-algebraic equations. For the starting values we assume

$$\begin{aligned} q_j - q(jh) &= \mathcal{O}(h^{r+2}) \quad \text{and} \quad g(q_j) = 0 \quad \text{for} \quad j = 0, \ldots, k-1 \\ \lambda_j - \lambda(jh) &= \mathcal{O}(h^r) \quad \text{for} \quad j = 1, \ldots, k-2 \end{aligned}$$

(the latter for the case of an explicit method with $\beta_{k-1} \neq 0$).

3.1 Energy conservation

It follows from differentiation of $H(q(t), p(t))$ that the total energy (3) is exactly preserved along solutions of the system (1). Recall that $p = M\dot{q}$.

Theorem 1 *Consider a symmetric linear multistep method (7) of order r with generating polynomials satisfying (12) and (13). Along the numerical solution of the constrained system (1) the total energy (3) is conserved up to $\mathcal{O}(h^r)$ over time $\mathcal{O}(h^{-r-1})$:*

$$H(q_n, p_n) = H(q_0, p_0) + \mathcal{O}(h^r) \quad \text{for } nh \le h^{-r-1}.$$

The constant symbolized by $\mathcal{O}$ is independent of n and h subject to $nh \le h^{-r-1}$.

3.2 Momentum conservation

Constrained N-body systems preserve the total angular momentum if both the potential $U(q)$ and the constraint function $g(q)$ are invariant under rotations. More generally, the invariance properties

$$U(\mathrm{e}^{\tau A} q) = U(q) \quad \text{and} \quad g(\mathrm{e}^{\tau A} q) = g(q) \qquad \text{for all } \tau, q \tag{14}$$

with a matrix A such that MA is skew-symmetric, implies that the Lagrange function

$$\mathcal{L}(q, \dot{q}, \lambda) = \frac{1}{2} \dot{q}^{\mathsf{T}} M \dot{q} - U(q) - g(q)^{\mathsf{T}} \lambda$$

is invariant under the symmetry $q \mapsto \mathrm{e}^{\tau A} q$. By Noether's theorem the momentum

$$L(q, p) = p^{\mathsf{T}} A q \tag{15}$$

is conserved along solutions of the constrained Hamiltonian system (1).

Theorem 2 *Consider a symmetric linear multistep method (7) of order r with generating polynomials satisfying (12) and (13). Along the numerical solution of the constrained system (1) satisfying (14) the momentum (15) is conserved up to $\mathcal{O}(h^r)$ over time $\mathcal{O}(h^{-r-1})$:*

$$L(q_n, p_n) = L(q_0, p_0) + \mathcal{O}(h^r) \quad \text{for } nh \le h^{-r-1}.$$

The constant symbolized by $\mathcal{O}$ is independent of n and h subject to $nh \le h^{-r-1}$.

Remark 1 Symplectic one-step methods (like the Rattle algorithm) conserve the momentum exactly. This is not the case for linear multistep methods, because their underlying one-step method cannot be symplectic (see [7, Section XV.4.1]).

4 Examples of higher order methods

Symmetric linear k-step multistep methods (7) with even k can be constructed as follows. We define the ρ-polynomial by

$$\rho(\zeta) = (\zeta - 1)^2 \prod_{j=1}^{k/2-1} (\zeta^2 + 2a_j\zeta + 1),$$

where a_j are distinct real numbers satisfying $-1 < a_j < 1$. This implies the assumption (12). The order condition (8) then uniquely determines the σ-polynomial of degree $k - 1$ such that the method is explicit and of order $r = k$. The resulting method is symmetric.

4.1 Coefficients of methods up to order 8

For methods up to order 8 we investigate for which values of a_j the corresponding σ-polynomial satisfies assumption (13).

Order $r = k = 4$: The σ-polynomial is given by

$$\sigma(\zeta) = (7 + a_1)(\zeta^3 + \zeta)/6 + (-1 + 5a_1)\zeta^2/3.$$

We see that condition (13) is satisfied for all choices of $-1 < a_1 < 1$.

Order $r = k = 6$: The σ-polynomial is given by

$$\sigma(\zeta) = \alpha(\zeta^5 + \zeta) + \beta(\zeta^4 + \zeta^2) + \gamma\zeta^3$$

with

$$\begin{aligned}
\alpha &= (79 + 9\,(a_1 + a_2) - a_1a_2)/60 \\
\beta &= (-14 + 26\,(a_1 + a_2) + 6\,a_1a_2)/15 \\
\gamma &= (97 + 7\,(a_1 + a_2) + 97\,a_1a_2)/30.
\end{aligned}$$

It has double zeros on the unit circle if $\beta^2 = 4\alpha(\gamma - 2\alpha)$. This curve separates the region where all non-zero roots of $\sigma(\zeta) = 0$ are of modulus 1, from that where at least one root is outside the unit disc, see Fig. 1.

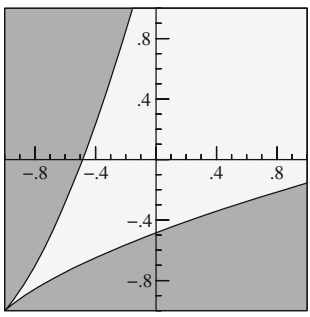

Fig. 1 The *dark grey region* shows the (a_1, a_2) values for which the corresponding σ-polynomial (case $k = 6$) has all non-zero roots on the unit circle

Order $r = k = 8$: The σ-polynomial is given by

$$\sigma(\zeta) = \alpha(\zeta^7 + \zeta) + \beta(\zeta^6 + \zeta^2) + \gamma(\zeta^5 + \zeta^3) + \delta\zeta^4$$

with

$$\begin{aligned}
\alpha &= (10993 + 1039\,s_1 - 95\,s_2 + 31\,s_3)/7560 \\
\beta &= (-2215 + 2279\,s_1 + 473\,s_2 - 73\,s_3)/1260 \\
\gamma &= (16661 + 491 s_1 + 8261\,s_2 + 2171\,s_3)/2520 \\
\delta &= (-8723 + 7027\,s_1 + 1357\,s_2 + 12067\,s_3)/1890,
\end{aligned}$$

where $s_1 = a_1 + a_2 + a_3$, $s_2 = a_1a_2 + a_1a_3 + a_2a_3$, and $s_3 = a_1a_2a_3$. We remark that none of the methods presented in Table 7.1 of [7, Sect. XV.7] (including a method proposed in [10]) satisfies the condition (13). However, if two among the parameters a_j are not too far from -1 and the third one is not far from 1, then condition (13) is satisfied. In particular, the choice

$$a_1 = -0.8, \qquad a_2 = -0.4, \qquad a_3 = 0.7$$

gives a method that satisfies both conditions (12) and (13).

Coefficients $\hat{\delta}_j$ *of* (11): Symmetric multistep methods of order $r = k$ are complemented by a difference formula (11) for the computation of the momenta. We use the coefficients $\hat{\delta}_j$, $j = -k/2, \ldots, k/2 - 1$ given by:

$$\begin{aligned}
k = 2: &\quad \frac{1}{2}(1, 1), \\
k = 4: &\quad \frac{1}{12}(-1, 7, 7, -1), \\
k = 6: &\quad \frac{1}{60}(1, -8, 37, 37, -8, 1), \\
k = 8: &\quad \frac{1}{840}(-3, 29, -139, 533, 533, -139, 29, -3).
\end{aligned}$$

4.2 Linear stability: interval of periodicity

When applied to the harmonic oscillator $\ddot{q} = -\omega^2 q$, the numerical solution of a symmetric linear multistep method is determined by the roots of the equation

$$\rho(\zeta) + (h\omega)^2\sigma(\zeta) = 0. \tag{16}$$

According to [9] we say that the method has interval of periodicity $(0, \Omega)$ if, for all $h\omega \in (0, \Omega)$, these roots are bounded by 1. For the method (5) of order 2 the interval of periodicity is $(0, 2)$, which implies that the method is stable only for $0 \le h\omega < 2$.

The assumption (12) and the symmetry of the method imply that the roots of (16) stay on the unit circle for small $h\omega > 0$. Consequently, the interval of periodicity is always non-empty.

Order $r = k = 4$: Studying the roots of (16) as a function of $h\omega$, one observes that a root can leave the unit circle only when two roots collapse at the point -1. This implies that

$$\Omega = \sqrt{-\frac{\rho(-1)}{\sigma(-1)}} = \sqrt{\frac{6(1-a_1)}{2-a_1}}.$$

For orders $r = k \geq 6$, the value Ω of the interval of periodicity can be computed numerically as function of the parameters a_j. For example, for values of (a_1, a_2) in the dark grey region of Fig. 1, we have $0 < \Omega < 1.05$, and the largest values of Ω are attained close to $a_1 = 0.66$ and $a_2 = -0.26$.

5 Numerical experiments

We have implemented symmetric linear multistep methods as proposed in Sect. 2. The following numerical experiments illustrate an excellent long-time behaviour for constrained Hamiltonian systems confirming our theoretical results.

Example 1 (Triple pendulum) We consider three connected mathematical pendulums moving in the plane and suspended at the origin. Denoting by $(q_1, q_2), (q_3, q_4), (q_5, q_6)$ their endpoints, the constraints $g_i(q) = 0$ are given by

$$q_1^2 + q_2^2 = 1, \quad (q_3 - q_1)^2 + (q_4 - q_2)^2 = 1, \quad (q_5 - q_3)^2 + (q_6 - q_4)^2 = 1.$$

The potential due to gravity is $U(q) = q_2 + q_4 + q_6$. We consider initial positions

$$q(0) = \Big(\frac{1}{2}, -\frac{\sqrt{3}}{2}, \frac{1}{2} + \frac{\sqrt{2}}{2}, -\frac{\sqrt{3}}{2} - \frac{\sqrt{2}}{2}, \frac{1}{2} + \frac{\sqrt{2}}{2} + 1, -\frac{\sqrt{3}}{2} - \frac{\sqrt{2}}{2}\Big),$$

which correspond to angles of 30°, 45°, and 90°, and momenta $p(0) = (0, \ldots, 0)$. This choice of initial values produces a chaotic behaviour of the solution.

To illustrate the necessity of the condition (13) we apply two symmetric multistep schemes of order $r = k = 6$, which are constructed as explained in Sect. 4:

(A) $a_1 = -0.7$, $a_2 = 0.4$, the σ-polynomial satisfies (13);
(B) $a_1 = -0.1$, $a_2 = 0.4$, the σ-polynomial does not satisfy (13).

The numerical Hamiltonian is shown in Fig. 2 for method (A). The error remains bounded without any drift, and an application with reduced step size shows that it is of size $\mathcal{O}(h^6)$. For the step size $h = 0.01$ this behaviour can be observed on much longer intervals than shown in Fig. 2 (numerically verified on [0, 200 000]). For method (B), the error explodes after about 130 steps (independent of the step size). This is due to the fact that the σ-polynomial has a zero of modulus larger than 1.

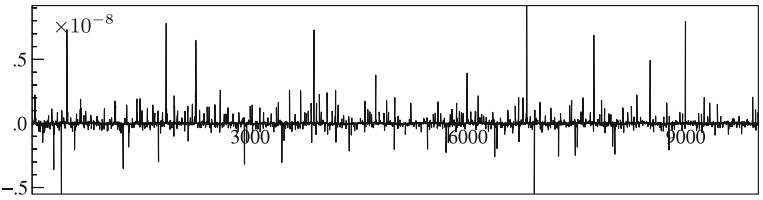

Fig. 2 Triple pendulum: error in the Hamiltonian for the symmetric multistep method (A) of order $r = k = 6$, applied with step size $h = 0.01$

Let us remark that the above description of the problem is extremely simple compared to the equations using minimal coordinates (angles). The long-time behaviour of method (A) in Fig. 2 should be compared with that of partitioned multistep methods applied to the equation in minimal coordinates (see [3, Section I.3]), where no energy preservation could be achieved in the chaotic regime.

Example 2 (Two-body problem on the sphere) We consider two particles moving on the unit sphere which are attracted by each other. As potential we take

$$U(q) = -\frac{\cos\vartheta}{\sin\vartheta}, \qquad \cos\vartheta = \langle Q_1, Q_2\rangle, \tag{17}$$

where $Q_1 = (q_1, q_2, q_3)^\mathsf{T}$, $Q_2 = (q_4, q_5, q_6)^\mathsf{T}$ are the positions of the two particles, and ϑ is their distance along a geodesics. The constraints are

$$g_1(q) = Q_1^\mathsf{T} Q_1 - 1, \qquad g_2(q) = Q_2^\mathsf{T} Q_2 - 1.$$

The equations of motion have the total angular momentum

$$L(p, q) = Q_1 \times P_1 + Q_2 \times P_2$$

as conserved quantity. Here, we use the notation $P_1 = (p_1, p_2, p_3)^\mathsf{T}$, $P_2 = (p_4, p_5, p_6)^\mathsf{T}$.

In view of a comparison with the experiments of [5] we consider initial values given in spherical coordinates by

$$Q_i = \big(\cos\phi_i \sin\theta_i, \sin\phi_i \sin\theta_i, \cos\theta_i\big)^\mathsf{T}$$

with $(\phi_1, \theta_1) = (0.8, 0.6)$ and $(\phi_2, \theta_2) = (0.5, 1.5)$ for the positions, and with $(\dot\phi_1, \dot\theta_1) = (1.1, -0.2)$ and $(\dot\phi_2, \dot\theta_2) = (-0.8, 0.0)$ for the velocities. In our numerical experiment we consider the multistep method of order $r = k = 8$ with parameters $a_1 = -0.8$, $a_2 = -0.4$, and $a_3 = 0.7$ (see Sect. 4). Figure 3 shows the error in the first component of the angular momentum. In perfect agreement with Theorem 2 we have an error of size $\mathcal{O}(h^8)$, and no drift can be observed over long time intervals (this is numerically checked on intervals as long as $T = 10^6$). A similar behavior is true for the other two components of the angular momentum and for the total energy.

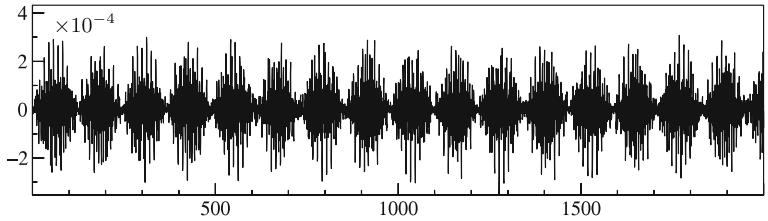

Fig. 3 Two-body problem on the sphere: error in the first component of the angular momentum for a symmetric multistep method of order $r = k = 8$ applied with step size $h = 0.02$

Since the same problem was treated numerically in [5, Section 5.3] with a composition method of order 8 and Rattle as basic integrator, this is the moment to say a few words on a comparison between symmetric linear multistep methods (as considered in the present work) and high order composition methods. Both are explicit and can have high order of accuracy. Which one is more efficient? From the experiment of [5] we see that an error in the energy of size 8×10^{-6} is obtained with step size $h = 0.15$. For the composition method of order 8 (with 17 Rattle applications per step) this corresponds to 226 666 force evaluations for an integration over an interval of length 2 000. With the multistep method we need a step size $h = 0.0125$ to achieve the same accuracy. This corresponds to 160 000 force evaluations, which is an improvement of about 30 %. Needless to say that such comparisons are problem dependent. We believe that it is important to consider both approaches.

Example 3 (Rigid body–heavy top) The configuration space of a rigid body with one point fixed is the rotation group $SO(3)$. The motion is described by an orthogonal matrix $Q(t)$ that satisfies

$$\begin{aligned} \ddot{Q}D &= -\nabla_Q U(Q) - Q\Lambda \\ 0 &= Q^{\mathsf{T}}Q - I, \end{aligned} \tag{18}$$

where the diagonal matrix $D = \mathrm{diag}(d_1, d_2, d_3)$ is related to the moments of inertia I_1, I_2, I_3 via

$$I_1 = d_2 + d_3, \quad I_2 = d_3 + d_1, \quad I_3 = d_1 + d_2,$$

and Λ is a symmetric matrix consisting of Lagrange multipliers. The potential, due to gravity, is given by $U(Q) = q_{33}$. For a more detailed description see [7, Section VII.5]. With $P = \dot{Q}D$, we are thus concerned with the Hamiltonian

$$H(P, Q) = \frac{1}{2}\,\mathrm{trace}(P D^{-1} P^{\mathsf{T}}) + U(Q).$$

The Eq. (18) is of the form (1) and satisfies the regularity condition (2).

With the abbreviation

$$\hat{\alpha}_{k-1}\widetilde{P}_{n+k-1/2} = -\sum_{j=0}^{k-2}\hat{\alpha}_j P_{n+j+1/2} - h\beta_{k-1}\nabla_Q U(Q_{n+k-1}) \\ -h\sum_{j=1}^{k-2}\beta_j\Big(\nabla_Q U(Q_{n+j}) + Q_{n+j}\Lambda_{n+j}\Big) \tag{19}$$

and $\gamma_{k-1} = \beta_{k-1}/\hat{\alpha}_{k-1}$ the multistep formula (10) becomes

$$P_{n+k-1/2} = \widetilde{P}_{n+k-1/2} - h\gamma_{k-1}Q_{n+k-1}\Lambda_{n+k-1} \\ Q_{n+k} = Q_{n+k-1} + hP_{n+k-1/2}D^{-1}.$$

These formulas are similar to those for the Rattle algorithm. We work with the auxiliary matrix

$$\Omega_{n+k-1} = Q_{n+k-1}^{\mathsf{T}}P_{n+k-1/2}D^{-1},$$

so that, for given $(Q_{n+j}, P_{n+j-1/2}, \Lambda_{n+j-1})$, $j \le k-1$, the approximations Q_{n+k}, $P_{n+k-1/2}$, Λ_{n+k-1} are obtained as follows:

- compute $\widetilde{P}_{n+k-1/2}$ from (19);
- find an orthogonal matrix $I + h\Omega_{n+k-1}$ such that

$$\Omega_{n+k-1}D = Q_{n+k-1}^{\mathsf{T}}\widetilde{P}_{n+k-1/2} - h\gamma_{k-1}\Lambda_{n+k-1}$$

  holds with a symmetric matrix Λ_{n+k-1};
- compute $Q_{n+k} = Q_{n+k-1}(I + h\Omega_{n+k-1})$;
- compute $P_{n+k-1/2} = Q_{n+k-1}\Omega_{n+k-1}D$.

Steps 1, 3, and 4 are straightforward computations. Step 2 requires the iterative solution of a nonlinear (quadratic) equation for Λ_{n+k-1}.

If an approximation P_n is required for output, it can be obtained from $P_n = Q_n\Omega_n$, where Ω_n and the symmetric matrix K_n are given by

$$\Omega_n = Q_n^{\mathsf{T}}\sum_{j=-l}^{l-1}\hat{\delta}_j P_{n+j+1/2} + hK_n \\ 0 = \Omega_n D^{-1} + D^{-1}\Omega_n^{\mathsf{T}}.$$

These two equations constitute a linear system for Ω_n and K_n. The computations can be done efficiently by representing orthogonal matrices in terms of quaternions (see [7, Section VII.5.3]).

6 Backward error analysis for smooth numerical solutions

For the proof of the main theoretical results we adapt the presentation of [6] to the case of constrained Hamiltonian systems. For the problem (1) we use the notation $f(q) = -\nabla U(q)$ and, without loss of generality, we assume the mass matrix M to be the identity, i.e., $M = I$.

6.1 Modified differential-algebraic system

Proposition 1 (Existence) *Let a consistent linear multistep method (7) be applied to the problem (1). Then, there exist unique h-independent functions $f_j(q, v)$ such that for every truncation index N, every solution $(y(t), \mu(t))$ of the modified differential-algebraic system*

$$\begin{aligned} \ddot{y} &= f(y) + hf_1(y, \dot{y}) + \cdots + h^{N-1} f_{N-1}(y, \dot{y}) - G(y)^{\mathsf{T}} \mu \\ 0 &= g(y) \end{aligned} \tag{20}$$

satisfies the multistep relation

$$\sum_{j=0}^{k} \alpha_j y(t + jh) = h^2 \sum_{j=0}^{k} \beta_j \Big(f\big(y(t + jh)\big) - G\big(y(t + jh)\big)^{\mathsf{T}} \mu(t + jh) \Big) + \mathcal{O}(h^{N+2}). \tag{21}$$

If the method is of order r, then $f_j(q, v) = 0$ for $j < r$. If it is symmetric, then $f_j(q, v) = 0$ for all odd j, and $f_j(q, -v) = f_j(q, v)$ for all even j.

Proof We write the Taylor series of a function as $z(t+h) = \mathrm{e}^{hD} z(t)$, where D denotes differentiation with respect to time. The identity (21) is then of the form

$$\rho(\mathrm{e}^{hD}) y = h^2 \sigma(\mathrm{e}^{hD}) \big(f(y) - G(y)^{\mathsf{T}} \mu \big) + \mathcal{O}(h^{N+2}). \tag{22}$$

With $x^2 \sigma(\mathrm{e}^x)/\rho(\mathrm{e}^x) = 1 + \vartheta_1 x + \vartheta_2 x^2 + \cdots$ this relation becomes

$$\ddot{y} = \big(1 + \vartheta_1 hD + \vartheta_2 h^2 D^2 + \cdots\big) \big(f(y) - G(y)^{\mathsf{T}} \mu \big) + \mathcal{O}(h^N). \tag{23}$$

With the exception of the h-independent term we replace $\mu(t)$ by $\mu(y(t), \dot{y}(t))$, where $\mu(q, v)$ is the expression obtained by differentiating twice the algebraic relation in (20). The coefficient functions $f_j(q, v)$ can then be obtained exactly as in the non-constrained case of [6]. □

In the modified differential-algebraic system (20) we have achieved uniqueness of the coefficient functions by imposing the term with the Lagrange multiplier to be independent of h.

6.2 Modified energy

We still assume that $M = I$ so that the momenta equal the velocities, $p = \dot{q}$. In this situation the total energy is given by

$$H(q, p) = \frac{1}{2} p^{\mathsf{T}} p + U(q).$$

It is preserved along the flow of the differential-algebraic system (1).

Proposition 2 (Energy preservation) *Consider a symmetric multistep method of order r applied to (1). Then, there exist unique h-independent functions $H_j(q, p)$ such that for every truncation index N the modified energy*

$$H_h(q, p) = H(q, p) + h^r H_r(q, p) + h^{r+2} H_{r+2}(q, p) + \cdots,$$

truncated at the $\mathcal{O}(h^N)$ term, satisfies

$$\frac{d}{dt} H_h\big(y(t), \dot{y}(t)\big) = \mathcal{O}(h^N)$$

along solutions of the modified differential-algebraic system (20).

Proof Instead of dividing (22) by $\rho(\mathrm{e}^{hD})$, we divide by $\sigma(\mathrm{e}^{hD})$. This yields

$$\big(1 + \gamma_1 hD + \gamma_2 h^2 D^2 + \cdots\big)\, \ddot{y} = -\nabla U(y) - G(y)^{\mathsf{T}} \mu + \mathcal{O}(h^N) \tag{24}$$

with coefficients γ_j given by $\rho(\mathrm{e}^x)/(x^2 \sigma(\mathrm{e}^x)) = 1 + \gamma_1 x + \gamma_2 x^2 + \cdots$. We take the scalar product with $\dot{y}$ and note that $G(y)\dot{y} = 0$, which follows from $g(y) = 0$ by differentiation with respect to time. The rest of the proof is the same as that of Proposition 1 in [6]. □

6.3 Modified momentum

We assume that $M = I$ and that A is a skew-symmetric matrix for which the invariance (14) holds.

Proposition 3 (Momentum preservation) *Consider a symmetric multistep method of order r applied to (1). Then, there exist unique h-independent functions $L_j(q, p)$ such that for every truncation index N the modified momentum*

$$L_h(q, p) = L(q, p) + h^r L_r(q, p) + h^{r+2} L_{r+2}(q, p) + \cdots,$$

truncated at the $\mathcal{O}(h^N)$ term, satisfies

$$\frac{d}{dt} L_h\big(y(t), \dot{y}(t)\big) = \mathcal{O}(h^N)$$

along solutions of the modified differential-algebraic system (20).

Proof We take the scalar product of (24) with Ay and note that the invariance (14) implies

$$f(y)^{\mathsf{T}} Ay = 0 \quad \text{and} \quad G(y)Ay = 0 \quad \text{for all } y.$$

The rest of the proof is the same as that of Proposition 2 in [6]. □

7 Long-term analysis of parasitic solution components

We consider irreducible, stable, symmetric linear multistep methods (7), we denote the double root of $\rho(\zeta) = 0$ by $\zeta_0 = 1$, and we assume that the remaining roots $\zeta_i, \zeta_{-i} = \overline{\zeta}_i$ for $1 \le i < k/2$ are simple. As a consequence of stability and symmetry we have $|\zeta_i| = 1$. Furthermore, we denote by $\zeta_i, \zeta_{-i} = \overline{\zeta}_i$ for $k/2 \le i < k$ complex pairs of roots of $\sigma(\zeta) = 0$ (not including 0 for explicit methods).

We consider the index set $\mathcal{I}_\rho = \{i \in \mathbf{Z};\ 1 \le |i| < k/2\}$ corresponding to the roots of $\rho(\zeta) = 0$ different from 1, and the index set $\mathcal{I}_\sigma = \{i \in \mathbf{Z};\ k/2 \le |i| < k - l\}$ (with $l = 0$ for implicit methods, and $l > 0$ else) corresponding to the non-zero roots of $\sigma(\zeta) = 0$. We denote $\mathcal{I} = \mathcal{I}_\rho \cup \mathcal{I}_\sigma$.

7.1 Linear problems with constant coefficients

To motivate the analysis of this section we consider the linear problem

$$\begin{aligned} \ddot{q} &= -A\,q - G^{\mathsf{T}}\lambda \\ 0 &= G\,q, \end{aligned} \tag{25}$$

where $q \in \mathbf{R}^d$, $\lambda \in \mathbf{R}^m$, the matrix A is symmetric, and G is of full rank. For this problem the multistep formula (7) reads

$$\sum_{j=0}^{k} \alpha_j q_{n+j} = -h^2 \sum_{j=0}^{k} \beta_j (A\,q_{n+j} + G^{\mathsf{T}}\lambda_{n+j}), \qquad G\,q_{n+k} = 0. \tag{26}$$

If the initial values are consistent, i.e., $G\,q_j = 0$ for $j = 0, \ldots, k-1$, then $G\,q_n = 0$ for all $n \ge 0$, and a multiplication by G of the multistep relation yields

$$\sum_{j=0}^{k} \beta_j\, G(A\,q_{n+j} + G^{\mathsf{T}}\lambda_{n+j}) = 0, \tag{27}$$

which permits to eliminate the Lagrange multipliers from the multistep formula. We thus obtain

$$\sum_{j=0}^{k} \alpha_j q_{n+j} = -h^2 \sum_{j=0}^{k} \beta_j \Big(I - G^{\mathsf{T}}(GG^{\mathsf{T}})^{-1}G\Big) A\,q_{n+j}.$$

This formula shows that the numerical solution $\{q_n\}$ depends only on the starting values $q_0, \ldots, q_{k-1}$, and is not affected by $\lambda_0, \ldots, \lambda_{k-1}$. Since we are concerned with a linear homogeneous difference equation with characteristic polynomial $\rho(\zeta)$ for $h = 0$, its solution is of the form

$$q_n = y(nh) + \sum_{i \in \mathcal{I}_\rho} \zeta_i^n \, z_i(nh), \tag{28}$$

where $y(t)$ and $z_i(t)$ are smooth functions in the sense that all their derivatives are bounded independently of h. The Lagrange multiplier is obtained from the difference relation (27) and satisfies

$$\lambda_n = -(GG^{\mathsf{T}})^{-1} G A \, q_n + \sum_{i \in \mathcal{I}_\sigma} \zeta_i^n \nu_i$$

with constant vectors ν_i that are determined by the initial approximations $\lambda_0, \ldots, \lambda_{k-1}$ for implicit methods, and by $\lambda_1, \ldots, \lambda_{k-2}$ for methods satisfying $\beta_k = 0$ and $\beta_{k-1} \neq 0$. Whereas only the zeros of the ρ-polynomial are important for the approximations $\{q_n\}$, also those of the σ-polynomial come into the game for the Lagrange multipliers $\{\lambda_n\}$.

7.2 Differential-algebraic system for parasitic solution components

Motivated by the analysis for the linear problem we aim at writing the numerical solution in the form (28) also for nonlinear problems. Due to the dependence of G on q we have to take the sum over $\mathcal{I}_\rho$ and $\mathcal{I}_\sigma$. It is easy to guess that $y(t)$ will be a solution of (20). It remains to study the smooth functions $z_i(t)$.

Proposition 4 (Differential-algebraic system) *Consider a symmetric linear multistep* (7) *of order r and assume that, with exception of the double root $\zeta_0 = 1$, all roots of $\rho(\zeta)$ are simple. For $i \in \mathcal{I}_\rho$ we let $\theta_i = \sigma(\zeta_i)/(\zeta_i \rho'(\zeta_i))$. We further assume that all non-zero roots of $\sigma(\zeta)$ are simple and of modulus* 1.

Then, there exist h-independent matrix-valued functions $A_{i,l}(y, v)$, $B_{i,l}(y, v)$, and $C_{i,l}(y, v)$, such that for every truncation index M and for every solution of the combined system (20) *and*

$$\begin{aligned} \dot{z}_i &= \big(h A_{i,1}(y, \dot{y}) + \cdots + h^{M-1} A_{i,M-1}(y, \dot{y})\big) z_i - \theta_i h \, G(y)^{\mathsf{T}} \nu_i \\ 0 &= G(y) \, z_i \end{aligned} \tag{29}$$

for $i \in \mathcal{I}_\rho$, and

$$\begin{aligned} \dot{\nu}_i &= \big(B_{i,0}(y, \dot{y}) + \cdots + h^{M-3} B_{i,M-3}(y, \dot{y})\big) \nu_i \\ z_i &= \big(h^3 C_{i,3}(y, \dot{y}) + \cdots + h^M C_{i,M}(y, \dot{y})\big) \nu_i \end{aligned} \tag{30}$$

for $i \in \mathcal{I}_\sigma$, with initial values satisfying $z_{-i}(0) = \overline{z}_i(0)$ and $v_{-i}(0) = \overline{v}_i(0)$ the following holds: as long as $\|z_i(t)\| \le \delta$ for all $i \in \mathcal{I}_\rho$ and $h^2 \|G(y(t))^\mathsf{T} v_i(t)\| \le \delta$ for $i \in \mathcal{I}_\sigma$ (with sufficiently small δ), the functions[1]

$$\widehat{y}(t) = y(t) + \sum_{i\in\mathcal{I}} \zeta_i^{t/h}\, z_i(t), \qquad \widehat{\mu}(t) = \mu(t) + \sum_{i\in\mathcal{I}} \zeta_i^{t/h}\, v_i(t) \tag{31}$$

satisfy $g(\widehat{y}(t)) = \mathcal{O}(\delta^2)$ and

$$\sum_{j=0}^{k} \alpha_j \widehat{y}(t+jh) = h^2 \sum_{j=0}^{k} \beta_j \Big(f\big(\widehat{y}(t+jh)\big) - G\big(\widehat{y}(t+jh)\big)^\mathsf{T} \widehat{\mu}(t+jh) \Big) + \mathcal{O}(h^{N+2} + h^{M+1}\delta + \delta^2). \tag{32}$$

Proof Taylor expansion yields

$$f\big(\widehat{y}(t)\big) = f\big(y(t)\big) + \sum_{i\in\mathcal{I}} \zeta_i^{t/h} f'\big(y(t)\big)\, z_i(t) + \mathcal{O}(\delta^2),$$

and similarly

$$G\big(\widehat{y}(t)\big)^\mathsf{T} \widehat{\mu}(t) = G\big(y(t)\big)^\mathsf{T} \mu(t) + \sum_{i\in\mathcal{I}} \zeta_i^{t/h} \Big(G\big(y(t)\big)^\mathsf{T} v_i(t) + \big(G'\big(y(t)\big)\, z_i(t)\big)^\mathsf{T} \mu(t) \Big) + \mathcal{O}(h^{-2}\delta^2),$$

because we have $h^2 v_i(t) = \mathcal{O}(\delta)$ on the considered interval. These relations show that (32) is satisfied if the functions $y(t)$ and $\mu(t)$ are solutions of (22) and the functions $z_i(t)$ and $v_i(t)$ satisfy the relation

$$\rho(\zeta_i \mathrm{e}^{hD}) z_i = h^2 \sigma(\zeta_i \mathrm{e}^{hD}) \Big(f'(y) z_i - G(y)^\mathsf{T} v_i - (G'(y) z_i)^\mathsf{T} \mu \Big) + \mathcal{O}(h^{M+1}\delta).$$

Similar to the proof of Proposition 1 we divide by $\rho(\zeta_i \mathrm{e}^{hD})$ and use the expansion

$$\frac{\sigma(\zeta_i \mathrm{e}^{x})}{\rho(\zeta_i \mathrm{e}^{x})} = \theta_{i,-1} x^{-1} + \theta_{i,0} + \theta_{i,1}\, x + \theta_{i,2}\, x^2 + \dots .$$

For $i \in \mathcal{I}_\rho$, where $\theta_{i,-1} \neq 0$, the above equation for z_i becomes

$$\dot{z}_i = h \Big(\theta_{i,-1} + \theta_{i,0} hD + \cdots \Big) \Big(f'(y) z_i - G(y)^\mathsf{T} v_i - (G'(y) z_i)^\mathsf{T} \mu \Big) + \mathcal{O}(h^M \delta). \tag{33}$$

[1] Note that the analogous expression in [4] and [6] has a sum over an index set that includes also finite products of ζ_i. This is not necessary for the investigations of the present work.

As in the proof of Proposition 1, the elimination of higher derivatives gives a differential equation of the form (29). The Lagrange multipliers ν_i are determined by the condition $G(y)z_i = 0$, which is needed for having $g(\widehat{y}) = \mathcal{O}(\delta^2)$.

For $i \in \mathcal{I}_\sigma$, where $\theta_{i,-1} = \theta_{i,0} = 0$ and $\theta_{i,1} \neq 0$, the equation for z_i becomes

$$z_i = h^2\Big(\theta_{i,1}hD + \theta_{i,2}(hD)^2 + \cdots\Big)\Big(f'(y)z_i - G(y)^{\mathsf{T}}\nu_i - (G'(y)z_i)^{\mathsf{T}}\mu\Big) + \mathcal{O}(h^{M+1}\delta). \tag{34}$$

We insert the equations (30) into (34) and express the higher derivatives of z_i and ν_i recursively in terms of ν_i. Equating powers of h yields for the h^3 term $C_{i,3} = -\theta_{i,1}((G'(y)\dot{y})^{\mathsf{T}} + G(y)^{\mathsf{T}}B_{i,0})$. The condition $G(y)z_i = 0$ yields $GC_{i,3} = 0$, so that multiplication of the above equation with $G(y)$ determines $B_{i,0}$, which in turn gives $C_{i,3}$. The same construction is used to determine the matrices for higher powers of h. This construction ensures that the relations (33) and (34) are satisfied, which completes the proof. □

Having found differential-algebraic equations for the smooth and parasitic solution components, we still need initial values for the combined system (20), (29), (30). We note that for given $y(0) = y_0$ and $\dot{y}(0) = \dot{y}_0$ satisfying $G(y_0)\dot{y}_0 = 0$, the function $\mu(t)$ is determined for all $t \geq 0$. For $i \in \mathcal{I}_\rho$, if in addition to y_0, $\dot{y}_0$ also $z_i(0) = z_{i,0}$ satisfying $G(y_0)z_{i,0} = 0$ is given, the functions $z_i(t)$ and $\nu_i(t)$ are determined for all $t \geq 0$ by (29). For $i \in \mathcal{I}_\sigma$ we need the initial value $\nu_i(0) = \nu_{i,0}$, which then determines $\nu_i(t)$ and $z_i(t)$ for all t by (30).

The next lemma shows how initial values $y_0, \dot{y}_0, z_{i,0}$ $(i \in \mathcal{I}_\rho)$, $\nu_{i,0}$ $(i \in \mathcal{I}_\sigma)$ can be obtained form starting approximations $q_0, q_1, \ldots, q_{k-1}$ and $\lambda_1, \ldots, \lambda_{k-2}$ for explicit methods satisfying $\beta_{k-1} \neq 0$. In general, there are $k - 2l$ starting values $\lambda_l, \ldots, \lambda_{k-l-1}$, where l is the multiplicity of the root 0 in $\sigma(\zeta)$. In the following, we only consider the most interesting case $l = 1$ for simplicity.

Proposition 5 (Initial values) *Under the assumptions of Proposition 4 consider the starting values $q_0, q_1, \ldots, q_{k-1}$ and $\lambda_1, \ldots, \lambda_{k-2}$. We assume that $g(q_j) = 0$, $q_j - q(jh) = \mathcal{O}(h^s)$ for $j = 0, 1, \ldots, k-1$, $\lambda_j - \lambda(jh) = \mathcal{O}(h^{s-2})$ for $j = 1, \ldots, k-2$, where $(q(t), \lambda(t))$ is a solution of (1) and $1 \leq s \leq r+2$. Then there exist (locally) unique consistent initial values $y_0, \dot{y}_0, z_{i,0}$ $(i \in \mathcal{I}_\rho)$, $\nu_{i,0}$ $(i \in \mathcal{I}_\sigma)$ of the combined system (20), (29), (30) such that its solution satisfies*

$$q_j = y(jh) + \sum_{i\in\mathcal{I}} \zeta_i^j\, z_i(jh) + G\big(y(jh)\big)^{\mathsf{T}}\kappa_j, \quad j = 0, \ldots, k-1 \tag{35}$$

$$\lambda_j = \mu(jh) + \sum_{i\in\mathcal{I}} \zeta_i^j\, \nu_i(jh), \qquad j = 1, \ldots, k-2, \tag{36}$$

where, with $\delta = h^s$, we have $\kappa_j = \mathcal{O}(\delta^2)$. The initial values satisfy $z_{-i,0} = \overline{z}_{i,0}$ for $i \in \mathcal{I}_\rho$ and $\nu_{-i,0} = \overline{\nu}_{i,0}$ for $i \in \mathcal{I}_\sigma$, and

$$\begin{aligned} &y_0 - q(0) = \mathcal{O}(\delta), \quad h\dot{y}_0 - h\dot{q}(0) = \mathcal{O}(\delta), \\ &z_{i,0} = \mathcal{O}(\delta),\ i \in \mathcal{I}_\rho, \quad h^2\nu_{i,0} = \mathcal{O}(\delta),\ i \in \mathcal{I}_\sigma. \end{aligned} \tag{37}$$

Proof The Eqs. (35)–(36) together with $g(y_0) = 0$, $G(y_0)\dot{y}_0 = 0$, and $G(y_0)z_{i,0} = 0$ constitute a nonlinear system $F(\mathbf{x}) = 0$ for the vector $\mathbf{x} = (y_0, h\dot{y}_0, (z_{i,0}; i \in \mathcal{I}_\rho), (h^2 v_{i,0}; i \in \mathcal{I}_\sigma), (\kappa_j; j = 0, \ldots, k-1))$. An approximation of its solution is $\mathbf{x}_0 = \big(q(0), h\dot{q}(0), 0, \ldots, 0\big)$. Using assumption (2), the inverse of the Jacobian matrix $F'(\mathbf{x}_0)$ can be shown to be bounded, and we have $F(\mathbf{x}_0) = \mathcal{O}(\delta)$. A convergence theorem for Newton's method thus proves the estimates (37). A sharper estimate for the variables κ_j follows from the fact that

$$\begin{aligned} 0 = g(q_j) - g(y(jh)) &= G(y(jh))(q_j - y(jh)) + \mathcal{O}\big(\|q_j - y(jh)\|^2\big) \\ &= G(y(jh))G(y(jh))^{\mathsf{T}}\kappa_j + \mathcal{O}(\delta^2), \end{aligned}$$

because $G(y)G(y)^{\mathsf{T}}$ has a bounded inverse. We have used that $q_j - y(jh) = q_j - q(jh) + q(jh) - y(jh)$ is bounded by $\mathcal{O}(\delta + h^{r+2})$. □

For given $q_0, \ldots, q_{k-1}$ and $\lambda_1, \ldots, \lambda_{k-2}$ (in the case of explicit methods) the numerical approximations q_k and λ_{k-1} are simultaneously obtained from (7).

Proposition 6 (Local error) *Under the assumptions of Propositions 4 and 5 consider the solution of the combined system* (20), (29), (30) *that corresponds to the starting approximations* $q_0, \ldots, q_{k-1}$ *and* $\lambda_1, \ldots, \lambda_{k-2}$. *Then the numerical approximation after one step satisfies*

$$\begin{aligned} q_k &= y(kh) + \sum_{i\in\mathcal{I}} \zeta_i^k\, z_i(kh) + \mathcal{O}(h^{N+2} + h^{M+1}\delta + \delta^2), \\ \lambda_{k-1} &= \mu((k-1)h) + \sum_{i\in\mathcal{I}} \zeta_i^k\, v_i((k-1)h) + \mathcal{O}(h^N + h^{M-1}\delta + h^{-2}\delta^2). \end{aligned}$$

Proof Using the notation (31) and subtracting (32) from the multistep formula (7), it follows from Proposition 5 that

$$\begin{aligned} \alpha_k(q_k - \widehat{y}(kh)) + \mathcal{O}(\delta^2) = h^2\beta_{k-1}G(q_{k-1})^{\mathsf{T}}(\lambda_{k-1} - \widehat{\mu}((k-1)h)) \\ + \mathcal{O}(h^{N+2} + h^{M+1}\delta + \delta^2). \end{aligned}$$

Inserting q_k from this formula into $g(q_k) = 0$ and using $g(\widehat{y}(kh)) = \mathcal{O}(\delta^2)$ yields the estimate for λ_{k-1}, and consequently also for q_k. □

7.3 Bounds on parasitic solution components

We next prove that the parasitic solution components $z_i(t)$ remain bounded and small on long time intervals.

Proposition 7 (Near-invariants) *Under the assumptions of Proposition 4 there exist* h*-independent matrix-valued functions* $E_{i,l}(y, v)$ *such that for every truncation index* M *and for every solution of the combined system* (20), (29), (30) *the functions*

$$K_i(y, v, z_i) = \|z_i\|^2 + \overline{z}_i^{\mathsf{T}}\Big(h^2 E_{i,2}(y, v) + \cdots + h^{M-1}E_{i,M-1}(y, v)\Big)z_i$$

for $i \in \mathcal{I}_\rho$ and

$$K_i(y, v, v_i) = \|h^2 G(y)^\mathsf{T} v_i\|^2 + h^4\, \overline{v}_i^\mathsf{T} \Big(h E_{i,1}(y, v) + \cdots + h^{M-1} E_{i,M-1}(y, v) \Big) v_i$$

for $i \in \mathcal{I}_\sigma$ are near-invariants of the system; more precisely, we have

$$K_i\big(y(t), \dot{y}(t), z_i(t)\big) = K_i\big(y(0), \dot{y}(0), z_i(0)\big) + \mathcal{O}(t h^M \delta^2), \quad i \in \mathcal{I}_\rho$$
$$K_i\big(y(t), \dot{y}(t), v_i(t)\big) = K_i\big(y(0), \dot{y}(0), v_i(0)\big) + \mathcal{O}(t h^M \delta^2), \quad i \in \mathcal{I}_\sigma$$

as long as $(y(t), \dot{y}(t))$ stays in a compact set and $\|z_i(t)\| \le \delta$ for $i \in \mathcal{I}_\rho$ and $h^2 \|G(y(t))^\mathsf{T} v_i(t)\| \le \delta$ for $i \in \mathcal{I}_\sigma$.

Proof We start as in the proof of Proposition 4. However, instead of dividing by $\rho(\zeta_i \mathrm{e}^{hD})$ we divide this time by $\sigma(\zeta_i \mathrm{e}^{hD})$. This yields

$$\left(\frac{\rho}{\sigma}\right)(\zeta_i \mathrm{e}^{hD})\, z_i = h^2 \Big(f'(y)\, z_i - G(y)^\mathsf{T} v_i - \big(G'(y) z_i\big)^\mathsf{T} \mu \Big) + \mathcal{O}(h^{M+1} \delta).$$

We multiply this relation with the transposed of $\overline{z}_i = z_{-i}$. The second term on the right-hand side vanishes, because of $G(y) z_{-i} = 0$. The first term on the right-hand side is real, because $f(y) = -\nabla U(y)$ so that $f'(y)$ is a symmetric matrix. This is also the case for the third term.

For the study of the left-hand side we consider the expansion (see [6, formula (4.16)])

$$\left(\frac{\rho}{\sigma}\right)(\zeta_i \mathrm{e}^{\mathrm{i}x}) = \sum_{l \ge -1} c_{i,l}\, x^l \qquad \text{with real coefficients} \quad c_{-i,l} = (-1)^l c_{i,l},$$

where $c_{i,-1} = c_{i,0} = 0$ and $c_{i,1} \neq 0$ for $i \in \mathcal{I}_\rho$, and $c_{i,-1} \neq 0$ for $i \in \mathcal{I}_\sigma$. We are thus concerned with the expression

$$\sum_{l \ge -1} c_{i,l} (-\mathrm{i}h)^l\, \overline{z}_i^\mathsf{T} z_i^{(l)}, \tag{38}$$

where for $l = -1$ we define in view of (34)

$$z_i^{(-1)} = h^3 \Big(\theta_{i,1} + \theta_{i,2}(hD) + \cdots \Big) \Big(f'(y) z_i - G(y)^\mathsf{T} v_i - (G'(y) z_i)^\mathsf{T} \mu \Big) + \mathcal{O}(h^{M+1} \delta) \tag{39}$$

such that $\dot{z}_i^{(-1)} = z_i$.

For $i \in \mathcal{I}_\rho$, we note that $2\Re(\overline{z}_i^{\mathsf{T}} \dot{z}_i) = z_{-i}^{\mathsf{T}} \dot{z}_i + \dot{z}_{-i}^{\mathsf{T}} z_i = \frac{d}{dt}\|z_i\|^2$. For the higher order expressions we have the telescoping sums

$$\Re\left(\overline{z}_i z_i^{(2m+1)}\right) = \frac{1}{2}\frac{d}{dt}\left(\sum_{j=0}^{2m}(-1)^j(\overline{z}_i^{(j)})^{\mathsf{T}} z_i^{(2m-j)}\right)$$
$$\Im\left(\overline{z}_i z_i^{(2m)}\right) = \frac{1}{2\mathrm{i}}\frac{d}{dt}\left(\sum_{j=0}^{2m-1}(-1)^j(\overline{z}_i^{(j)})^{\mathsf{T}} z_i^{(2m-j-1)}\right)$$

so that the imaginary part of (38) is a total derivative of a quadratic function in z_i and its derivatives. Using the system (29), first and higher order derivatives of z_i can be expressed as a linear function of z_i with coefficients depending on y and $\dot{y}$. Dividing the first formula of the present proof by $c_{i,1}(-\mathrm{i}h)/2$, and then taking the real part gives

$$\frac{d}{dt}K_i\big(y(t), \dot{y}(t), z_i(t)\big) = \mathcal{O}(h^M\delta^2)$$

with a quadratic function in z_i of the desired form.

For $i \in \mathcal{I}_\sigma$, we note that

$$2\,\Re\left(\overline{z}_i^{\mathsf{T}} z_i^{(-1)}\right) = 2\,\Re\left(\overline{\dot{z}_i^{(-1)}}^{\mathsf{T}} z_i^{(-1)}\right) = \frac{d}{dt}\left\|z_i^{(-1)}\right\|^2.$$

The same argument as above yields a near-invariant that is quadratic in $h^{-1}z_i^{(-1)}$. By formula (39) the leading term in $h^{-1}z_i^{(-1)}$ is given by $-h^2\theta_{i,1}G(y)^{\mathsf{T}}v_i$ and the higher-order terms can be expressed as linear functions in v_i. This proves the statement of the proposition. □

Let us collect the assumptions that are required for proving the boundedness of the parasitic solution components.

(A1) The multistep method (7) is symmetric and of order r. All roots of $\rho(\zeta)$, with the exception of the double root $\zeta_0 = 1$, are simple. All non-zero roots of $\sigma(\zeta)$ are simple and of modulus one.

(A2) The potential $U(q)$ and the constraint function $g(q)$ of (1) are defined and smooth in an open neighbourhood of a compact set K.

(A3) The starting approximations $q_0, \ldots, q_{k-1}$ and $\lambda_1, \ldots, \lambda_{k-2}$ are such that the initial values for the differential-algebraic system (20), (29), (30) obtained from Proposition 5 satisfy

$$y(0) \in K, \qquad \|\dot{y}(0)\| \le M,$$
$$\|z_i(0)\| \le \delta/2,\ i \in \mathcal{I}_\rho \quad \text{and} \quad \left\|h^2 G(y(0))^{\mathsf{T}} v_i(0)\right\| \le \delta/2,\ i \in \mathcal{I}_\sigma.$$

(A4) The numerical solution $\{q_n\}$, for $0 \le nh \le T$, stays in a compact set K_0 that has a positive distance to the boundary of K.

Theorem 3 (Long-time bounds for the parasitic components) *Assume* (A1)–(A4). *For sufficiently small h and δ and for fixed truncation indices N and M that are large enough such that $h^N = \mathcal{O}(\delta^2)$ and $h^M = \mathcal{O}(\delta)$, there exist functions $y(t)$, $\mu(t)$ and $z_i(t)$, $\nu_i(t)$ for $i \in \mathcal{I}$ on an interval of length*

$$T = \mathcal{O}(h\delta^{-1})$$

such that

- $q_n = y(nh) + \sum_{i\in\mathcal{I}} \zeta_i^n z_i(nh)$ *for* $0 \le nh \le T$;
- $\lambda_n = \mu(nh) + \sum_{i\in\mathcal{I}} \zeta_i^n \nu_i(nh)$ *for* $0 \le nh \le T$;
- *on every subinterval* $[nh, (n+1)h)$ *the functions* $y(t)$, $\mu(t)$ *and* $z_i(t)$, $\nu_i(t)$ *for* $i \in \mathcal{I}$ *are a solution of the system* (20), (29), (30);
- *the functions* $y(t)$, $h^2\mu(t)$ *and* $z_i(t)$, $h^2\nu_i(t)$ *for* $i \in \mathcal{I}$ *have jump discontinuities of size* $\mathcal{O}(\delta^2)$ *at the grid points* nh;
- *for* $0 \le t \le T$, *the parasitic components are bounded by*

$$\|z_i(t)\| \le \delta, \ i \in \mathcal{I}_\rho \quad \text{and} \quad \left\|h^2 G(y(t))^\mathsf{T} \nu_i(t)\right\| \le \delta, \ i \in \mathcal{I}_\sigma.$$

Proof To define the functions $y(t)$, $\mu(t)$, $z_i(t)$, $\nu_i(t)$ on the interval $[nh, (n+1)h)$ we consider the consecutive numerical solution values $q_n, q_{n+1}, \ldots, q_{n+k-1}$ and $\lambda_{n+1}, \ldots, \lambda_{n+k-2}$. We compute initial values for the system (20), (29), (30) according to Proposition 5, and we let $y(t)$, $\mu(t)$, $z_i(t)$, $\nu_i(t)$ be its solution on $[nh, (n+1)h)$. By Proposition 6 this construction yields jump discontinuities of size $\mathcal{O}(\delta^2)$ at the grid points.

It follows from Proposition 7 that $K_i(y(t), \dot{y}(t), z_i(t))$ for $i \in \mathcal{I}_\rho$ and $K_i(y(t), \dot{y}(t), \nu_i(t))$ for $i \in \mathcal{I}_\sigma$ remain constant up to an error of size $\mathcal{O}(h^{M+1}\delta^2)$ on the interval $[nh, (n+1)h)$. Taking into account the jump discontinuities of size $\mathcal{O}(\delta^2)$, we find that

$$\begin{aligned} K_i(y(t), \dot{y}(t), z_i(t)) &\le K_i(y(0), \dot{y}(0), z_i(0)) + C_1 t h^{-1}\delta^3 + C_2 t h^M \delta^2 \\ K_i(y(t), \dot{y}(t), \nu_i(t)) &\le K_i(y(0), \dot{y}(0), \nu_i(0)) + C_1 t h^{-1}\delta^3 + C_2 t h^M \delta^2 \end{aligned}$$

as long as $\|z_i(t)\| \le \delta$ for $i \in \mathcal{I}_\rho$ and $\|h^2 G(y(t))^\mathsf{T} \nu_i(t)\| \le \delta$ for $i \in \mathcal{I}_\sigma$. By Proposition 7 this then implies with $C_3 = C_1 + hC_2$, for $i \in \mathcal{I}_\rho$,

$$\|z_i(t)\|^2 \le \|z_i(0)\|^2 + C_3 t h^{-1}\delta^3 + C_4 h^2 \delta^2.$$

For $i \in \mathcal{I}_\sigma$ we obtain

$$\left\|h^2 G(y(t))^\mathsf{T} \nu_i(t)\right\|^2 \le \left\|h^2 G(y(0))^\mathsf{T} \nu_i(0)\right\|^2 + C_3 t h^{-1}\delta^3 + C_4 h\delta^2.$$

The assumptions $\|z_i(t)\| \le \delta$ and $\left\|h^2 G(y(t))^\mathsf{T} \nu_i(t)\right\| \le \delta$ are certainly satisfied as long as $C_3 t\delta \le h/4$ and $C_4 h \le 1/4$, so that the right-hand side of the above estimates is bounded by δ^2. This proves not only the estimate for $\|z_i(t)\|$ and $\left\|h^2 G(y(t))^\mathsf{T} \nu_i(t)\right\|$, but at the same time it guarantees recursively that the above construction of the functions $y(t)$, $\mu(t)$, $z_i(t)$, $\nu_i(t)$ is feasible. □

7.4 Proof of the main results

The proof of Theorem 1 combines Theorem 3 and Proposition 2. For the piecewise smooth function $y(t)$ of Theorem 3 we have

$$H_h(y(t), \dot{y}(t)) = H_h(y(0), \dot{y}(0)) + \mathcal{O}(th^N) + \mathcal{O}(th^{-1}\delta^2),$$

where the first error term comes from the truncation of the modified energy and the second error term comes from the discontinuity at the grid points. By the bounds for the parasitic components z_i we have

$$q_n = y(nh) + \mathcal{O}(\delta) \quad \text{and} \quad p_n = \dot{y}(nh) + \mathcal{O}(h^{-1}\delta + h^r)$$

because the differentiation formula is of order r. We therefore obtain

$$H_h(q_n, p_n) = H_h(q_0, p_0) + \mathcal{O}(th^N) + \mathcal{O}(th^{-1}\delta^2) + \mathcal{O}(h^{-1}\delta + h^r).$$

With $\delta = h^{r+2}$, Theorem 1 now follows by using the estimate between the modified energy H_h and the original energy H as given by Proposition 2.

Theorem 2 is obtained in the same way using Proposition 3.

Acknowledgments This work was partially supported by the Fonds National Suisse, Project No. 200020-126638 and 200021-129485.

References

1. Andersen, H.C.: Rattle: a "velocity" version of the shake algorithm for molecular dynamics calculations. J. Comput. Phys. **52**, 24–34 (1983)
2. Arévalo, C., Führer, C., Söderlind, G.: β-Blocked multistep methods for Euler–Lagrange DAEs: linear analysis. Z. Angew. Math. Mech. **77**(8), 609–617 (1997)
3. Console, P., Hairer, E.: Long-term stability of symmetric partitioned linear multistep methods. In: Current Challenges in Stability Issues for Numerical Differential Equations, C.I.M.E. Summer Sch., pp. 1–36. Springer, Berlin (2013)
4. Hairer, E.: Backward error analysis for multistep methods. Numer. Math. **84**, 199–232 (1999)
5. Hairer, E., Hairer, M.: GniCodes—Matlab programs for geometric numerical integration. In: Frontiers in Numerical Analysis (Durham 2002), pp. 199–240. Springer, Berlin (2003)
6. Hairer, E., Lubich, C.: Symmetric multistep methods over long times. Numer. Math. **97**, 699–723 (2004)
7. Hairer, E., Lubich, C., Wanner, G.: Geometric numerical integration. Structure-Preserving Algorithms for Ordinary Differential Equations. Springer Series in Computational Mathematics, vol. 31, 2nd edn. Springer, Berlin (2006)
8. Jay, L.: Symplectic partitioned Runge–Kutta methods for constrained Hamiltonian systems. SIAM J. Numer. Anal. **33**, 368–387 (1996)
9. Lambert, J.D., Watson, I.A.: Symmetric multistep methods for periodic initial value problems. J. Inst. Math. Appl. **18**, 189–202 (1976)
10. Quinlan, G.D., Tremaine, S.: Symmetric multistep methods for the numerical integration of planetary orbits. Astron. J. **100**, 1694–1700 (1990)
11. Reich, S.: Symplectic integration of constrained Hamiltonian systems by composition methods. SIAM J. Numer. Anal. **33**, 475–491 (1996)
12. Ryckaert, J.-P., Ciccotti, G., Berendsen, H.J.C.: Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J. Comput. Phys. **23**, 327–341 (1977)