European Society of Computational Methods
in Sciences and Engineering (ESCMSE)

Journal of Numerical Analysis,
Industrial and Applied Mathematics
(JNAIAM)
vol. ?, no. ?, 2010, pp. 1-12
ISSN 1790–8140

Energy-preserving variant of collocation methods[1]

E. Hairer[2]

Université de Genève,
Section de Mathématiques, 2-4 rue du Lièvre,
CH-1211 Genève 4, Switzerland

Received 15 October, 2009

Abstract: We propose a modification of collocation methods extending the 'averaged vector field method' to high order. These new integrators exactly preserve energy for Hamiltonian systems, are of arbitrarily high order, and fall into the class of B-series integrators. We discuss their symmetry and conjugate-symplecticity, and we compare them to energy-preserving composition methods.

Keywords: Hamiltonian systems, energy-preserving integrators, B-series, Runge–Kutta methods, collocation, conjugate-symplecticity, symmetry, Gaussian quadrature.

Mathematics Subject Classification: 65P10, 65L06

PACS: 45.20.Jj, 02.30.Hq, 02.60.Lj

1 Introduction

We consider Hamiltonian differential equations, written in the form

$$\dot y = J^{-1}\nabla H(y), \tag{1}$$

where J is a skew-symmetric constant matrix, and the Hamiltonian $H(y)$ is assumed to be sufficiently differentiable. For its numerical integration one would ideally like to have energy-preservation and symplecticity of the discrete flow. Both properties cannot be satisfied at the same time unless the integrator produces the exact solution (see [6, page 379]). There is a huge literature on symplectic integrators, and backward error analysis permits to prove that they nearly preserve the energy (they exactly preserve a modified Hamiltonian).

Methods that exactly preserve energy (energy-momentum methods, discrete gradient methods) have been considered since several decades. More recently, the existence of energy-preserving B-series methods has been shown in [5], and a practical integrator has been proposed in [10]. It is called 'averaged vector field method' and is defined by

$$y_{n+1} = y_n + h\int_0^1 J^{-1}\nabla H\Big((1-\tau)\,y_n + \tau\,y_{n+1}\Big)\mathrm{d}\tau. \tag{2}$$

[1] This work is supported by the Fonds National Suisse, project No. 200020-126638.

[2] Corresponding author. E-mail: Ernst.Hairer@unige.ch

This method exactly preserves the energy for an arbitrary Hamiltonian, and in contrast to projection-type integrators, only requires evaluations of the vector field. It is symmetric and its Taylor series has the structure of a B-series. For polynomial Hamiltonians, the integral can be evaluated exactly, and the implementation is comparable to that of the implicit mid-point rule. There is much activity applying this method and various modifications to polynomial Hamiltonian systems (see [2, 3]) and also to Hamiltonian partial differential equations. Another interesting class of integrators are 'extended collocation methods' and 'Hamiltonian Boundary Value Methods', which exactly preserve energy of polynomial Hamiltonian systems (see [9, 1]).

In this article we combine the ideas of both approaches. At the one hand we extend the method (2) to higher order, similar to the way how the implicit mid-point rule can be extended to collocation Runge–Kutta methods. On the other hand, we consider a limit of the method proposed in [9] to cover also non-polynomial Hamiltonians. In Section 2 we give the precise definition of the class of considered methods, we interpret them as Runge-Kutta methods with a continuum of stages, we study their order, we prove their exact energy conservation, and we investigate their symmetry. Section 3 is devoted to the study how close the methods are to symplectic integrators. We interpret the numerical solution as a B-series. We show that many of the algebraic conditions on the B-series coefficients, which characterize symplecticity, are satisfied. We then prove that the methods of maximal order $2s$ are conjugate-symplectic up to order at least $2s + 2$, and we prove that composition methods of order 4 based on (2) cannot be conjugate-symplectic up to an order higher than 4. We terminate this article with a discussion of the new class of integrators (Section 4).

2 The new class of integrators

Although our interest is mainly in Hamiltonian systems (1), we define the methods for general initial value problems $\dot y = f(y)$, $y(t_0) = y_0$.

2.1 Definition

In the following we use the notation

$$\ell_i(\tau) = \prod_{j=1,j\neq i}^{s} \frac{\tau - c_j}{c_i - c_j}, \qquad b_i = \int_0^1 \ell_i(\tau)\,\mathrm{d}\tau$$

for the Lagrange basis polynomials in interpolation.

Energy-preserving collocation methods. *Let $c_1, \ldots, c_s$ be distinct real numbers (usually $0 \le c_i \le 1$) for which $b_i \neq 0$ for all i. We consider a polynomial $u(t)$ of degree s satisfying*

$$u(t_0) \;=\; y_0 \tag{3}$$

$$\dot u(t_0 + c_i h) \;=\; \frac{1}{b_i}\int_0^1 \ell_i(\tau)\, f\bigl(u(t_0 + \tau h)\bigr)\,\mathrm{d}\tau. \tag{4}$$

The numerical solution after one step is then defined by $y_1 = u(t_0 + h)$.

Approximating the integral with the interpolatory quadrature formula corresponding to the nodes $c_1, \ldots, c_s$, we obtain $\dot u(t_0 + c_i h) = f\bigl(u(t_0 + c_i h)\bigr)$ in place of (4), and the method reduces to a classical collocation method. Because of this connection we call the new methods (with abuse of notation) energy-preserving collocation methods.

Denoting $k_i := \dot u(t_0 + c_i h)$, the derivative of $u(t)$ and the polynomial $u(t)$ itself become

$$\begin{aligned} \dot u(t_0 + \tau h) &= \sum_{i=1}^{s} \ell_i(\tau)\, k_i && (5)\\ u(t_0 + \tau h) &= y_0 + h \sum_{i=1}^{s} \int_0^{\tau} \ell_i(\sigma)\,\mathrm{d}\sigma\, k_i. && (6) \end{aligned}$$

Inserting (6) into (4) yields a nonlinear equation $\mathbf{k} = G(\mathbf{k})$ for the finite dimensional vector $\mathbf{k} = (k_1, \ldots, k_s)$. This equation can be solved by fixed-point iteration or by Newton techniques. Every evaluation of $G(\mathbf{k})$ requires the computation of integrals which can be computed exactly for polynomial vector fields or approximated numerically by an accurate quadrature formula. Instead of writing the polynomial $u(t)$ in terms of y_0 and $\dot u(t_0 + c_i h)$, we can also express it in terms of y_0 and $\mathbf{u} = \big(u(t_0 + \widehat c_i h), i = 1, \ldots, s\big)$, where the $\widehat c_i$ are distinct numbers that need not coincide with the collocation points c_i. With k_i from (4) inserted into (6) we obtain a nonlinear equation $\mathbf{u} = \widehat G(\mathbf{u})$ which can again be solved iteratively. If the appearing integrals are replaced by a fixed high-order quadrature, the resulting method is closely related to the approach presented in [1].

2.2 Interpretation as Runge–Kutta method and order

For theoretical investigations of the method, it is more convenient to interpret it as a Runge–Kutta method with a continuum of stages $\tau \in [0, 1]$:

$$Y_\tau = y_0 + h \int_0^1 A_{\tau,\sigma}\, f(Y_\sigma)\,\mathrm{d}\sigma, \qquad y_1 = y_0 + h \int_0^1 B_\sigma\, f(Y_\sigma)\,\mathrm{d}\sigma. \qquad (7)$$

Here, $Y_\tau \approx y(t_0 + C_\tau h)$, where $C_\tau = \int_0^1 A_{\tau,\sigma}\,\mathrm{d}\sigma$. In our situation, the internal stages Y_τ are the values of the polynomial $u(t_0 + \tau h)$, and the coefficients are given by

$$C_\tau = \tau, \qquad A_{\tau,\sigma} = \sum_{i=1}^{s} \frac{1}{b_i} \int_0^{\tau} \ell_i(\alpha)\,\mathrm{d}\alpha\ \ell_i(\sigma), \qquad B_\sigma = 1. \qquad (8)$$

Theorem 1. *Let r be the order of the interpolatory quadrature formula based on the nodes $c_1, \ldots, c_s$. Then, the energy-preserving collocation method has*

$$\textit{order} \quad = \quad \begin{cases} 2s & \textit{for}\ \ r \ge 2s - 1 \\ 2r - 2s + 2 & \textit{for}\ \ r \le 2s - 2. \end{cases}$$

Proof. Using the facts that $\int_0^1 g(\sigma)\,\mathrm{d}\sigma = \sum_{i=1}^s b_i g(c_i)$ for polynomials of degree $r - 1$ and that $\sum_{i=1}^s \ell_i(\tau) g(c_i) = g(\tau)$ for polynomials of degree $s - 1$, one sees that our method satisfies the simplifying assumptions (partial integration is helpful for the verification of $D(\zeta)$)

$$\begin{aligned} B(\rho):&\quad \int_0^1 B_\tau\, C_\tau^{k-1}\mathrm{d}\tau = \frac{1}{k}, && k = 1, \ldots, \rho \\ C(\eta):&\quad \int_0^1 A_{\tau,\sigma} C_\sigma^{k-1}\mathrm{d}\sigma = \frac{1}{k}\, C_\tau^{k}, && k = 1, \ldots, \eta \\ D(\zeta):&\quad \int_0^1 B_\tau\, C_\tau^{k-1} A_{\tau,\sigma}\,\mathrm{d}\tau = \frac{B_\sigma}{k}\,(1 - C_\sigma^{k}), && k = 1, \ldots, \zeta \end{aligned}$$

with $\rho = \infty$, $\eta = \min(s, r - s + 1)$ and $\zeta = \min(s - 1, r - s)$. The statement now follows from a classical result by Butcher (see for example [7, p. 208]), which states that the order of a Runge–Kutta method satisfying the simplifying assumptions $B(\rho)$, $C(\eta)$ and $D(\zeta)$ is at least $\min(\rho, 2\eta + 2, \eta + \zeta + 1)$. This proves the statement. □

It is surprising that the order of our new method is not the same as for the corresponding classical collocation method (which is equal to the order r of the quadrature formula). For $r = 2s-1$ it is higher, whereas for $r \le 2s-3$ it is lower. We notice that the order based on Newton–Côtes quadrature drops to 2 when s is even. We also see that the order of the new methods is always even.

2.3 Examples

Case $\mathbf{s = 1}$. Writing the polynomial $u(t)$ as $u(t_0 + \tau h) = y_0 + \tau(y_1 - y_0)$, we recover the averaged vector field method (2), independently of the choice of c_1.

Case $\mathbf{s = 2}$. We consider nodes c_1, c_2 satisfying

$$\frac{1}{3} - \frac{1}{2}(c_1 + c_2) + c_1 c_2 = 0,$$

so that the corresponding quadrature formula has at least order $r = 3$, e.g., Gaussian nodes $c_{1,2} = \frac{1}{2} \mp \frac{\sqrt{3}}{6}$ or Radau nodes $c_1 = \frac{1}{3}, c_2 = 1$. The Runge–Kutta coefficients become

$$C_\tau = \tau, \qquad A_{\tau,\sigma} = \tau\big((4-3\tau) - 6\,(1-\tau)\sigma\big), \qquad B_\sigma = 1,$$

and we notice with surprise that the method is independent of the choice of c_1. If we write the polynomial $u(t)$ (which is of degree 2) as a linear combination of $y_0 = u(t_0)$, $y_1 = u(t_0 + h)$ and $Y_{1/2} := u(t_0 + h/2)$, then this method becomes

$$\begin{aligned} Y_{1/2} &= y_0 + h\int_0^1 \Big(\frac{5}{4} - \frac{3}{2}\,\sigma\Big)\, f\big(u(t_0+\sigma h)\big)\,\mathrm{d}\sigma \\ y_1 &= y_0 + h\int_0^1 f\big(u(t_0+\sigma h)\big)\,\mathrm{d}\sigma. \end{aligned} \tag{9}$$

This method is of order 4, it is symmetric and exactly preserves the energy for Hamiltonian systems as will be shown in Sections 2.4 and 2.5 below. There is an interesting connection to the methods studied in [9]. If we approximate the integrals in (9) by the Lobatto quadrature of order eight, we obtain precisely the "extended Lobatto IIIA method of order four" [9].

Case $\mathbf{s = 3}$. We can either take any quadrature formula of order $r \ge 2s - 1 = 5$ and proceed as for the case $s = 2$. Another possibility is to exploit the fact that $A_{\tau,\sigma}$ is known by (8) to be a polynomial of degree 2 in σ. The simplifying assumption $C(s)$ thus uniquely determines its coefficients and yields

$$A_{\tau,\sigma} = \tau\big((9 - 18\tau + 10\tau^2) - 12\,(3 - 8\tau + 5\tau^2)\,\sigma + 30\,(1 - 3\tau + 2\tau^2)\,\sigma^2\big)$$

together with $C_\tau = \tau$ and $B_\sigma = 1$. Representing the polynomial $u(t)$ of degree 3 as linear combination of $y_0, Y_{1/3}, Y_{2/3}$, and y_1, the method reads

$$\begin{aligned} Y_{1/3} &= y_0 + h\int_0^1 \Big(\frac{37}{27} - \frac{32}{9}\,\sigma + \frac{20}{9}\,\sigma^2\Big)\, f\big(u(t_0+\sigma h)\big)\,\mathrm{d}\sigma \\ Y_{2/3} &= y_0 + h\int_0^1 \Big(\frac{26}{27} + \frac{8}{9}\,\sigma - \frac{20}{9}\,\sigma^2\Big)\, f\big(u(t_0+\sigma h)\big)\,\mathrm{d}\sigma \\ y_1 &= y_0 + h\int_0^1 f\big(u(t_0+\sigma h)\big)\,\mathrm{d}\sigma \end{aligned} \tag{10}$$

and is of order 6. If we represent the polynomial $u(t)$ by different values, say $y_0, Y_\alpha, Y_\beta, y_1$, we get another representation of the same method.

2.4 Exact energy-preservation

The proof of energy conservation for the new class of methods is as simple as for method (2). From the fundamental theorem of calculus we have

$$H\big(u(t_0+h)\big)-H\big(u(t_0)\big)=h\int_0^1 \dot u(t_0+\tau h)^{\mathsf T}\nabla H\big(u(t_0+\tau h)\big)\,\mathrm d\tau.$$

Substituting (5) for $\dot u(t_0+\tau h)$ and replacing $k_i=\dot u(t_0+c_ih)$ with (4), where $f(y)=J^{-1}\nabla H(y)$, yields for this expression

$$h\sum_{i=1}^s\frac{1}{b_i}\int_0^1 \ell_i(\tau)\nabla H\big(u(t_0+\tau h)\big)^{\mathsf T}\mathrm d\tau\,J^{-\mathsf T}\int_0^1 \ell_i(\tau)\nabla H\big(u(t_0+\tau h)\big)\,\mathrm d\tau$$

which vanishes by the skew-symmetry of the matrix J. As a consequence our method satisfies $H(y_1)=H(y_0)$, so that $H(y_n)=H(y_0)$ is exactly preserved along the numerical solution.

2.5 Symmetry

A numerical one-step method $y_{n+1}=\Phi_h(y_n)$ is symmetric, if $\Phi_h^{-1}=\Phi_{-h}$. For a standard s-stage Runge–Kutta method this is the case, if their coefficients satisfy $a_{s+1-i,s+1-j}+a_{i,j}=b_j$. For a continuous Runge–Kutta method (7) this condition can be written as

$$A_{1-\tau,1-\sigma}+A_{\tau,\sigma}=B_\sigma. \tag{11}$$

For the energy-preserving collocation method with coefficients given by (8), this condition is equivalent to (after differentiation with respect to τ)

$$\sum_{i=1}^s\frac{1}{b_i}\Big(\ell_i(\tau)\ell_i(\sigma)-\ell_i(1-\tau)\ell_i(1-\sigma)\Big)=0. \tag{12}$$

Lemma 2. *If the nodes $c_1,\dots,c_s$ are symmetric, i.e., $c_{s+1-i}=1-c_i$ for all i, then the energy-preserving collocation method is symmetric.*

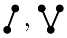
Proof. This follows from (12), because for symmetric nodes $c_1,\dots,c_s$ we have $b_i=b_{s+1-i}$ and $\ell_i(1-\tau)=\ell_{s+1-i}(\tau)$. □

As a consequence, all methods presented in Section 2.3 are symmetric. Notice that, in contrast to standard collocation methods, the symmetry of the nodes is a sufficient but not necessary condition for the symmetry of the method (in fact, a method based on the non-symmetric Radau nodes is identical to the method based on the symmetric Gaussian nodes).

3 On their conjugate-symplecticity

Numerical methods in the class of B-series integrators, which exactly conserve the energy for Hamiltonian systems, cannot be symplectic. It is of interest to investigate whether such methods are conjugate to a symplectic integrator (up to a certain order). The proof of our results is based on a representation of the numerical solution in terms of trees and elementary differentials. Our notation closely follows the monograph [6, Chap. III] (see also [4] for a recent survey on this approach).

3.1 Trees and symplecticity

Let $T = \{\,\bullet\,,$ $,$ $,\ldots\}$ be the set of rooted trees. We denote $u=[u_1,\ldots,u_m]$ the tree that is obtained by grafting the roots of the trees $u_1,\ \ldots,\ u_m$ to a new vertex which becomes the root of u. The order $|u|$ is the number of vertices of the tree u, the symmetry coefficient is defined recursively by

$$\sigma(\,\bullet\,)=1, \qquad \sigma(u)=\sigma(u_1)\cdots\sigma(u_m)\mu_1!\mu_2!\cdots\ , \tag{13}$$

where the integers $\mu_1,\mu_2,\ldots$ count equal trees among u_1, ..., u_m, and elementary differentials $F(u)$ are given by

$$F(\,\bullet\,)(y)=f(y), \qquad F(u)(y)=f^{(m)}(y)\big(F(u_1)(y),\ldots,F(u_m)(y)\big).$$

From Runge–Kutta theory it is known that the Taylor series of the exact solution, of the numerical solution, and of the internal stages of the method (7) can be written as B-series, $y(t_0{+}h)=B(e,y_0)$, $y_1=B(\phi,y_0)$ and $Y_\tau=B(\phi_\tau,y_0)$, where

$$B(a,y)=y+\sum_{u\,\in\,T}\frac{h^{|u|}}{\sigma(u)}\,a(u)\,F(u)(y). \tag{14}$$

The coefficients for the exact solution are recursively defined by

$$e(\,\bullet\,)=1, \qquad e(u)=\frac{1}{|u|}\,e(u_1)\cdot\ldots\cdot e(u_m) \qquad \text{for} \quad u=[u_1,\ldots,u_m]. \tag{15}$$

Those for the numerical solution and the internal stages are defined by $\phi_\tau(\,\bullet\,)=C_\tau,\ \phi(\,\bullet\,)=1$ and

$$\phi_\tau(u)=\int_0^1 A_{\tau,\sigma}\phi_\sigma(u_1)\cdot\ldots\cdot\phi_\sigma(u_m)\,\mathrm{d}\sigma, \quad \phi(u)=\int_0^1 B_\tau\phi_\tau(u_1)\cdot\ldots\cdot\phi_\tau(u_m)\,\mathrm{d}\tau.$$

For the study of symplecticity of a numerical B-series method, the expressions

$$a(u,v):=a(u\circ v)+a(v\circ u)-a(u)a(v) \qquad \text{for} \quad u,v\in T \tag{16}$$

play an important role. Here, the so-called Butcher product $u\circ v$ of two trees $u=[u_1,\ldots,u_m]$ and v is defined by $u\circ v=[u_1,\ldots,u_m,v]$. The mapping $\Phi_h(y)=B(a,y)$ is a symplectic transformation for Hamiltonian vector fields (i.e., its Jacobian matrix satisfies $\Phi_h'(y)^{\mathsf T}J\Phi_h'(y)=J$), if and only if

$$a(u,v)=0 \qquad \text{for all} \quad u,v\in T.$$

This characterization, due to Calvo and Sanz-Serna, is explained and discussed in [6, Chap. VI.7].

Theorem 3. *Assume that a Runge–Kutta method (with finitely many or a continuum of stages) is of order r and satisfies the simplifying conditions $C(\eta)$ and $D(\zeta)$ with $r\ge\eta\ge\zeta$. Then, we have*

$$\phi(u,v)=0 \qquad \textit{for} \quad \min(|u|,|v|)\le\zeta. \tag{17}$$

Proof. In a first step we prove the statement for $u=\tau_k$ with $k\le\zeta$, where $\tau_k=[\,\bullet,\ldots,\bullet\,]$ is the bushy tree of order k. For a tree $v=[v_1,\ldots,v_m]$ we multiply the simplifying assumption $D(\zeta)$ with $\phi_\sigma(v_1)\cdot\ldots\cdot\phi_\sigma(v_m)$, and integrate over σ. This yields for $k\le\zeta$ (for a continuum of stages)

$$\int_0^1\!\int_0^1 B_\tau\,C_\tau^{k-1}A_{\tau,\sigma}\,\mathrm{d}\tau\,\phi_\sigma(v_1)\cdot\ldots\cdot\phi_\sigma(v_m)\,\mathrm{d}\sigma=\frac{1}{k}\int_0^1 B_\sigma\,(1-C_\sigma^k)\,\phi_\sigma(v_1)\cdot\ldots\cdot\phi_\sigma(v_m)\,\mathrm{d}\sigma.$$

On the other hand the condition $C(\eta)$ implies for $k \le \eta$ that

$$\int_0^1 B_\tau \int_0^1 A_{\tau,\sigma} C_\sigma^{k-1} \mathrm{d}\sigma\, \phi_\tau(v_1) \cdot \ldots \cdot \phi_\tau(v_m)\, \mathrm{d}\tau = \frac{1}{k} \int_0^1 B_\tau C_\tau^k \phi_\tau(v_1) \cdot \ldots \cdot \phi_\tau(v_m)\, \mathrm{d}\tau.$$

Adding both equations and using the fact that $\phi(\tau_k) = 1/k$ for $k \le r$ (the method is of order r), we obtain

$$\phi(\tau_k \circ v) + \phi(v \circ \tau_k) = \phi(\tau_k)\, \phi(v)$$

which proves $\phi(\tau_k, v) = 0$ for $k \le \zeta$ and for $v \in T$.

We next notice that a recursive application of the simplifying assumption $C(\eta)$ implies that $\phi_\tau(w) = e(w) C_\tau^{|w|}$ for all trees w with $|w| \le \eta$, where $e(w)$ is the coefficient defined in (15). We now fix a tree $u = [u_1, \ldots, u_m]$ satisfying $|u| = k \le \eta$, and we apply this property to u_j and u. This implies $\phi(u) = k\, e(u) \phi(\tau_k)$, $\phi(u \circ v) = e(u_1) \cdot \ldots \cdot e(u_m) \phi(\tau_k \circ v)$, and $\phi(v \circ u) = k\, e(u) \phi(v \circ \tau_k)$ for all trees v. Consequently, we have $\phi(u, v) = k e(u) \phi(\tau_k, v)$. Since $\phi(\tau_k, v) = 0$ for $k \le \zeta$ and $\zeta \le \eta$, this completes the proof of the theorem. □

For the energy-preserving collocation method of maximal order $2s$, the simplifying conditions $C(s)$ and $D(s-1)$ hold. Therefore the symplecticity condition $\phi(u, v) = 0$ is satisfied when at least one tree among u and v has $\le s-1$ vertices. For the leading error term, where trees of order $2s+1$ are involved, only pairs (u, v) with $|u| = s$ and $|v| = s+1$ give thus rise to non-zero $\phi(u, v)$, and due to the simplifying assumption $C(s)$ they are related by

$$\phi(u, v) = \frac{e(u)\, e(v)}{e(\tau_s)\, e(\tau_{s+1})}\, \phi(\tau_s, \tau_{s+1}) \qquad \text{for} \quad |u| = s, \ |v| = s+1. \tag{18}$$

3.2 Conditional conjugate-symplecticity

We consider the energy-preserving collocation method of order $2s$. It cannot be symplectic (see also Section 3.1). It is therefore interesting to study whether it can be conjugate to a symplectic method: does there exist a change of coordinates $\chi_h(y) = y + \mathcal{O}(h^{2s})$, such that $\chi_h^{-1} \circ \Phi_h \circ \chi_h$ is a symplectic integrator? If this happens, its long-time behavior is the same as that of a symplectic method. Unfortunately, it turns out that this aim is too ambitious.

We therefore study whether our methods satisfy some weaker condition. We call the method $y_{n+1} = \Phi_h(y_n)$ of order r *conjugate-symplectic up to order q* ($q > r$), if there exists a change of coordinates $\chi_h(y) = y + \mathcal{O}(h^r)$, such that the Jacobian matrix of $\Psi_h = \chi_h^{-1} \circ \Phi_h \circ \chi_h$ satisfies $\Psi_h'(y)^\mathsf{T} J \Psi_h'(y) = J + \mathcal{O}(h^{q+1})$. In such a situation the modified differential equation (in the sense of backward error analysis) is Hamiltonian up to terms of size $\mathcal{O}(h^q)$. For nearly integrable systems the global error is thus bounded by $\mathcal{O}(th^r + t^2 h^q)$, so that the method behaves like a symplectic integrator on intervals of length $\mathcal{O}(h^{r-q})$, see [6, page 436].

Lemma 4. *If a symmetric integrator Ψ_h satisfies $\Psi_h'(y)^\mathsf{T} J \Psi_h'(y) = J + \mathcal{O}(h^{2m})$, then it automatically also satisfies $\Psi_h'(y)^\mathsf{T} J \Psi_h'(y) = J + \mathcal{O}(h^{2m+1})$.*

Proof. An integrator is symmetric if $\Psi_h\big(\Psi_{-h}(y)\big) = y$. Differentiating this relation with respect to y yields $\Psi_h'\big(\Psi_{-h}(y)\big) \Psi'_{-h}(y) = I = \textit{identity}$. The assumption implies that

$$\Psi_h'(\hat{y})^\mathsf{T} J \Psi_h'(\hat{y}) = J + C(\hat{y}) h^{2m} + \mathcal{O}(h^{2m+1}) \quad \text{with} \quad \hat{y} = \Psi_{-h}(y).$$

We multiply this relation from the right with $\Psi'_{-h}(y)$ and from the left with $\Psi'_{-h}(y)^\mathsf{T}$. Since $\hat{y} = y + \mathcal{O}(h)$ and $\Psi_h'(y) = I + \mathcal{O}(h)$, this yields $J = \Psi'_{-h}(y)^\mathsf{T} J \Psi'_{-h}(y) + C(y) h^{2m} + \mathcal{O}(h^{2m+1})$. However, the assumption written with $-h$ in place of h gives $\Psi'_{-h}(y)^\mathsf{T} J \Psi'_{-h}(y) = J + C(y) h^{2m} + \mathcal{O}(h^{2m+1})$. Both relations together can be true only when $C(y) = 0$. This proves the statement. □

The main result of this section is the following property concerning conjugate-symplecticity. It shows that besides exact energy-preservation, the new class of methods has also excellent symplecticity features.

Theorem 5. *The energy-preserving collocation method of order* $2s$ *is conjugate-symplectic up to order* $2s+2$.

Proof. We have seen in Section 3.1 that the numerical method is a B-series $\Phi_h(y) = B(\phi, y)$. We therefore search for a change of coordinates that can also be written as a B-series $\chi_h(y) = B(c, y)$. To satisfy $\chi_h(y) = y + \mathcal{O}(h^{2s})$ and because we are only interested in the leading error term, we require $c(u) = 0$ for trees with $|u| \neq 2s$. The study of conjugate-symplecticity heavily relies on the fact that the composition of two B-series is again a B-series, $B(a, B(b, y)) = B(b \cdot a, y)$, where the product on the coefficients is given by

$$(b \cdot a)(u) = b(u) + \sum_{\sigma \in S(u)} b(u \setminus \sigma)\, a(\sigma_u). \qquad (19)$$

Here, $S(u)$ denotes the set of ordered subtrees of u (an ordered subtree is a connected subset of the vertices containing the root), $u \setminus \sigma$ is the collection of trees that remains when σ is removed, and σ_u is the rooted tree given by the vertices of σ with edges induced from the tree u (we use the notation of [4]). The proof of the theorem is in several steps.

a) As composition of B-series, the conjugated method $\Psi_h = \chi_h^{-1} \circ \Phi_h \circ \chi_h$ is also a B-series $\Psi_h(y) = B(\psi, y)$, and a transcription of the relation $\chi_h \circ \Psi_h = \Phi_h \circ \chi_h$ into coefficients of B-series yields

$$(\psi \cdot c)(w) = (c \cdot \phi)(w) \qquad \text{for trees } \ w \in T. \qquad (20)$$

The statement of the theorem, written in terms of the coefficients of B-series, is as follows: find coefficients $c(w)$ for trees w satisfying $|w| = 2s$ (we assume $c(w) = 0$ for trees with $|w| \neq 2s$), such that the coefficients $\psi(w)$, defined by (20), satisfy

$$\psi(u, v) = 0 \qquad \text{for pairs } u, v \text{ with } \ |u| + |v| = 2s + 1. \qquad (21)$$

For trees with $|u| + |v| \leq 2s$, the relation (21) is satisfied, because the method $\Psi_h(y)$ is of order $2s$, and trees with $|u| + |v| = 2s + 2$ need not be considered by Lemma 4, because $\chi_{-h}(y) = \chi_h(y)$ guarantees the symmetry of $\Psi_h(y)$.

b) Exploiting the fact that $c(w) \neq 0$ only for trees with $|w| = 2s$ and using the composition law (19), the relation (20) becomes for the trees $u \circ v$ and $v \circ u$ (with $|u| + |v| = 2s + 1$)

$$\begin{aligned} \psi(u \circ v) + \sum_{\sigma \in S_{2s}(u \circ v)} c(\sigma_{u \circ v}) &= \phi(u \circ v) + \begin{cases} c(v) & \text{if } \ u = \bullet \\ 0 & \text{else,} \end{cases} \\ \psi(v \circ u) + \sum_{\sigma \in S_{2s}(v \circ u)} c(\sigma_{v \circ u}) &= \phi(v \circ u) + \begin{cases} c(u) & \text{if } \ v = \bullet \\ 0 & \text{else,} \end{cases} \end{aligned}$$

where $S_{2s}(w)$ denotes the set of subtrees of w with exactly $2s$ vertices. We add both relations and then we subtract the product $\psi(u)\psi(v) = \phi(u)\phi(v)$ from both sides. The term $c(v)$ (resp. $c(u)$) on the right-hand side, which is present only for $u = \bullet$ (resp. $v = \bullet$), cancels with the subtree $\sigma \in S_{2s}(v \circ u)$, where $u = \bullet$ is removed (resp. $\sigma \in S_{2s}(u \circ v)$, where $v = \bullet$ is removed). Since all other $\sigma \in S_{2s}(v \circ u)$ can be written either as $\sigma = \hat{u} \circ v$ with $\hat{u} \in S_{|u|-1}(u)$ or as $\sigma = u \circ \hat{v}$ with $\hat{v} \in S_{|v|-1}(v)$, we see that the symplecticity condition (21) becomes equivalent to

$$\phi(u, v) \ = \sum_{\hat{u} \in S_{|u|-1}(u)} c(\hat{u}, v) \ + \sum_{\hat{v} \in S_{|v|-1}(v)} c(u, \hat{v}) \qquad \text{for pairs } u, v \text{ with } \ |u| + |v| = 2s + 1 \qquad (22)$$

(notice that the set $S_{|u|-1}(u)$ is empty when $u = \bullet$). The proof of the theorem is complete if we can find coefficients $c(w)$ that satisfy this relation.

c) In this part of the proof we define the coefficients $c(w)$ of the conjugacy mapping $\chi_h(y)$. We shall show in part (d) that they are solution of the linear system (22). We start with assuming

$$\begin{aligned} c(\tau_s,\tau_s) &= \frac{1}{s}\,\phi(\tau_s,\tau_{s+1}) && (23)\\ c(u,v) &= \frac{e(u)\,e(v)}{e(\tau_s)\,e(\tau_s)}\,c(\tau_s,\tau_s) \qquad \text{if} \quad |u|=|v|=s && (24)\\ c(u,v) &= 0 \qquad \text{else,} && (25) \end{aligned}$$

where $e(u)$ are the coefficients of the exact solution, defined in (15). Let us first explain the meaning of defining values for the expressions $c(u,v) = c(u \circ v) + c(v \circ u)$ (notice that $c(u) = 0$ for $|u| < 2s$). We let $\sim$ be the equivalence relation defined by $u \circ v \sim v \circ u$, so that equivalence classes are free (or un-rooted) trees. This means that the root can be moved arbitrarily in the tree. If the value $c(w)$ is known for one rooted tree in an equivalence class, then the knowledge of $c(u,v)$ uniquely defines the coefficients $c(w)$ for all other trees in this equivalence class.

For a tree of the form $u \circ u$, the value of $c(u \circ u)$ is determined by (23)-(25), because $c(u,u) = 2c(u \circ u)$. Therefore, whenever an equivalence class contains a trees of the form $u \circ u$, the values of $c(w)$ are uniquely defined for all trees w in this equivalence class. For equivalence classes that do not contain a tree of the form $u \circ u$, we can arbitrarily fix the value $c(w)$ for one tree in the equivalence class, which then uniquely defines the values for the remaining trees.

This freedom in the choice of the coefficients $c(w)$ reflects the fact that whenever a change of coordinates χ_h transforms the method $\Phi_h(y)$ to a symplectic mapping up to a certain order, the same will be true for $\chi_h \circ \varphi_h$, if φ_h is a symplectic transformation.

d) We prove that the coefficients defined in part (c) are a solution of the linear system (22). By the symmetry of $\phi(u,v)$ we can assume $|u| < |v|$, so that the first sum in (22) vanishes by (25), because $|\hat u| \le s-1$. If $|u| < s$, and hence $|v| > s+1$, also the second sum in (22) vanishes, and the equation is satisfied by Theorem 3, because $\phi(u,v) = 0$ in this case. It remains to consider the situation, where $|u| = s$ and $|v| = s+1$. Using (24), then Lemma 6 below, then $e(\tau_s) = 1/s$ and $e(\tau_{s+1}) = 1/(s+1)$, and finally (23), yields for the right-hand side of (22)

$$\sum_{\hat v \in S_s(v)} c(u,\hat v) \;=\; \sum_{\hat v \in S_s(v)} \frac{e(u)\,e(\hat v)}{e(\tau_s)\,e(\tau_s)}\,c(\tau_s,\tau_s) \;=\; \frac{(s+1)\,e(u)\,e(v)}{e(\tau_s)\,e(\tau_s)}\,c(\tau_s,\tau_s) \;=\; \frac{e(u)\,e(v)}{e(\tau_s)\,e(\tau_{s+1})}\,\phi(\tau_s,\tau_{s+1}).$$

Comparing this relation with (18) proves that every set of coefficients $c(w)$ satisfying (23)-(25) solves the linear system (22). □

Lemma 6. *The coefficients $e(v)$ of the B-series for the exact solution (cf. definition (15)) satisfy*

$$(s+1)\,e(v) = \sum_{\hat v \in S_s(v)} e(\hat v) \qquad \textit{for trees } v \textit{ with} \quad |v| = s+1. \tag{26}$$

Proof. For the bushy tree $v = \tau_{s+1} = [\bullet,\ldots,\bullet]$ of order $s+1$, the sum in (26) is over s copies of $\hat v = \tau_s$. Since $e(\tau_s) = 1/s$, this proves (26) for $v = \tau_{s+1}$.

For an arbitrary tree v, the proof is by induction on its order. Let $v = [v_1,\ldots,v_m]$ and assume the formula (26) be true for $v_1,\ldots,v_m$. We split the sum over $\hat v \in S_s(v)$ as

$$\sum_{\hat v \in S_s(v)} e(\hat v) \;=\; \frac{1}{s}\Big(\sum_{\hat v_1 \in S_{s_1}(v_1)} e(\hat v_1)\,e(v_2)\ldots e(v_m) \;+\ldots+\; \sum_{\hat v_m \in S_{s_m}(v_m)} e(v_1)\,e(v_2)\ldots e(\hat v_m)\Big),$$

where $s_j + 1 = |v_j|$. By induction hypothesis the right-hand expression becomes

$$\frac{1}{s}\,\big(|v_1| + \ldots + |v_m|\big)\, e(v_1)\, e(v_2) \ldots e(v_m) = e(v_1)\, e(v_2) \ldots e(v_m) = (s+1)\, e(v)$$

which completes the proof of the lemma. □

The obvious question is now whether the energy-preserving collocation methods can be conjugate-symplectic up to an order higher than $2s+2$. The answer is negative for $s=1$: the recent publication [2] shows that the method (2) is not conjugate-symplectic up to an order higher than 4. The question is still open for $s \geq 2$. Since many symplecticity conditions are satisfied by the new class of methods (see Theorem 3), we expect the energy-preserving collocation method of order $2s$ to be conjugate-symplectic up to an order higher than $2s+2$ (it could even by $4s$). We do not pursue this question in the present article.

Taking up the discussion of the beginning of Section 3.2, we see that the methods of order $2s$ proposed in this article not only preserve exactly (up to round-off errors) the total energy, but also behave like symplectic methods on long time intervals. For nearly integrable Hamiltonian systems, Theorem 5 implies that the global error is bounded by $\mathcal{O}(th^{2s} + t^2h^{2s+2})$ and, if our conjecture is true, by $\mathcal{O}(th^{2s})$ as long a $th^{2s} \leq$ const.

3.3 Comparison with composition methods

An alternative way of designing high order energy-preserving integrators is by composition. Let $\Phi_h(y)$ denote the averaged vector field method (2), which is symmetric and of order two. The symmetric composition method

$$\Phi_h^{[4]} = \Phi_{\gamma_m h} \circ \ldots \circ \Phi_{\gamma_2 h} \circ \Phi_{\gamma_1 h} \circ \Phi_{\gamma_2 h} \circ \ldots \circ \Phi_{\gamma_m h} \tag{27}$$

is of order four if (see [6, Section V.3.2])

$$\gamma_1 + 2\big(\gamma_2 + \ldots + \gamma_m\big) = 1, \qquad \gamma_1^3 + 2\big(\gamma_2^3 + \ldots + \gamma_m^3\big) = 0. \tag{28}$$

It is not of order six for linear differential equations (e.g., the harmonic oscillator) if

$$\gamma_1^5 + 2\big(\gamma_2^5 + \ldots + \gamma_m^5\big) \neq 0. \tag{29}$$

Theorem 7. *A symmetric composition (27) based on the averaged vector field method (2) and satisfying (28)-(29) cannot be conjugate-symplectic up to an order higher than* 4.

Proof. The proof is much simplified if we work with modified differential equations in the sense of backward error analysis. It has been shown in [3] that the modified differential equation of the averaged vector field method (2) is

$$\dot y = f(y) + h^2 f_3(y) + h^4 f_5(y) + \ldots \tag{30}$$

where

$$f_3(y) = \frac{1}{12}(f'f'f)(y), \qquad f_5(y) = \sum_{|u|=5} \frac{a(u)}{\sigma(u)}\, F(u)(y)$$

with coefficients $a(u)$ given in Table 1. It follows from Lemma 5.9 of [6, page 94] that the modified differential equation of the composition method $\Phi_h^{[4]}$ is

$$\dot y = f(y) + h^4\big(\alpha\, f_5(y) + \beta\,\big[f,[f,f_3]\big](y)\big) + \mathcal{O}(h^6), \tag{31}$$

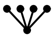
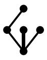
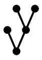
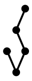
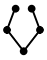
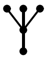
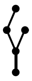
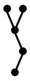
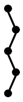
u									
720 a(u)	0	-2	-8	-4	4	-6	2	8	9
12 b(u)	0	2	4	1	2	6	1	-2	0

Table 1: B-series coefficients for the modified differential equation (31)

where $\alpha = \gamma_1^5 + 2\big(\gamma_2^5 + \ldots + \gamma_m^5\big) \neq 0$ and β is some constant depending on the coefficients γ_i of the method $\Phi_h^{[4]}$. Recall that $[f, f_3] = f_3' f - f' f_3$. Applying this operation a second time shows that the double commutator has the structure of a B-series

$$\big[f, [f, f_3]\big](y) = \sum_{|u|=5} \frac{b(u)}{\sigma(u)}\, F(u)(y)$$

with coefficients $b(u)$ of Table 1. This implies that the leading term of the local truncation error is $h^5\big(\alpha\, f_5(y) + \beta\, \big[f, [f, f_3]\big](y)\big)$, so that the B-series coefficients of the method $\Phi_h^{[4]}$ are

$$\phi(u) = e(u) + \alpha\, a(u) + \beta\, b(u) \qquad \text{for trees } u \text{ with } \ |u| = 5.$$

Theorem 8.3 of [6, page 224] gives necessary (and sufficient) conditions for a fourth order method to be conjugate-symplectic up to order 5. These conditions are fulfilled if and only if $\alpha = 0$. This proves the statement of the theorem. □

4 Conclusion and discussion

For Hamiltonian differential equations there is a long-standing dispute on the question whether in a numerical simulation it is more important to preserve energy or symplecticity. Many people give more weight on symplecticity, because it is known (by backward error analysis arguments) that symplectic integrators conserve a modified Hamiltonian. The present article proposes a class of integrators that guarantees energy conservation up to machine precision. The methods are not symplectic, but those of maximal order $2s$ are conjugate-symplectic up to at least order $2s + 2$ (possibly to a much higher order). Therefore, we expect these methods to behave like symplectic methods when they are applied to an integration over very long time intervals.

It is not common to use numerical methods, where integrals over the vector field have to be evaluated. However, for polynomial vector fields this work is reduced to the computation of integrals over polynomials, and this can be done once for all in the beginning of the integration. Therefore, the complexity of the methods is comparable to that of implicit Runge–Kutta methods with the same number of stages. In a parallel computing environment, vector field evaluations can be done in parallel, when the integrals are approximated by a high order quadrature formula.

The proposed methods of maximal order $2s$ are symmetric. This implies that for nearly integrable reversible differential equations their long-time behavior is comparable to that of symplectic integrators (see [6, Chapter XI]). The use of variable step sizes retains energy conservation, but destroys symplecticity (and conjugate-symplecticity). If the step sizes are selected in a reversible way (for example, using the techniques of [8]), the excellent long-time behavior of symmetric integrators can be conserved.

Acknowledgment

This work emerged during the international conference “ICNAAM 2009”, Rethymno, Crete, Greece, 18-22 September 2009. We are in particular grateful to Luigi Brugnano whose presentation was the stimulus of the present work, and to Brynjulf Owren for pointing out the pseudo-conjugate-symplecticity of the second order averaged vector field method.

References

[1] L. Brugnano, F. Iavernaro, and D. Trigiante. Analysis of Hamiltonian boundary value methods (HBVMs) for the numerical solution of polynomial Hamiltonian dynamical systems. *Submitted for publication*, 2009.

[2] E. Celledoni, R. I. McLachlan, D. I. McLaren, B. Owren, G. R. W. Quispel, and W. M. Wright. Energy-preserving Runge–Kutta methods. *M2AN Math. Model. Numer. Anal.*, 43(4):645–649, 2009.

[3] E. Celledoni, R. I. McLachlan, B. Owren, and G. R. W. Quispel. Energy-preserving integrators and the structure of B-series. *NTNU Report No. 5*, 2009.

[4] P. Chartier, E. Hairer, and G. Vilmart. Algebraic structures of B-series. *To appear in Foundations of Comput. Math.*, 2010.

[5] E. Faou, E. Hairer, and T.-L. Pham. Energy conservation with non-symplectic methods: examples and counter-examples. *BIT*, 44:699–709, 2004.

[6] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations.* Springer Series in Computational Mathematics 31. Springer-Verlag, Berlin, 2nd edition, 2006.

[7] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I. Nonstiff Problems.* Springer Series in Computational Mathematics 8. Springer, Berlin, 2nd edition, 1993.

[8] E. Hairer and G. Söderlind. Explicit, time reversible, adaptive step size control. *SIAM J. Sci. Comput.*, 26:1838–1851, 2005.

[9] F. Iavernaro and D. Trigiante. High-order symmetric schemes for the energy conservation of polynomial Hamiltonian problems. *JNAIAM J. Numer. Anal. Ind. Appl. Math.*, 4(1-2):87–101, 2009.

[10] G. R. W. Quispel and D. I. McLaren. A new class of energy-preserving numerical integration methods. *J. Phys. A*, 41(4):045206, 7, 2008.