MATHEMATICS OF COMPUTATION
Volume 76, Number 260, October 2007, Pages 1941–1953
S 0025-5718(07)01967-9
Article electronically published on May 9, 2007

NUMERICAL INTEGRATORS BASED ON MODIFIED DIFFERENTIAL EQUATIONS

PHILIPPE CHARTIER, ERNST HAIRER, AND GILLES VILMART

Abstract. Inspired by the theory of modified equations (backward error analysis), a new approach to high-order, structure-preserving numerical integrators for ordinary differential equations is developed. This approach is illustrated with the implicit midpoint rule applied to the full dynamics of the free rigid body. Special attention is paid to methods represented as B-series, for which explicit formulae for the modified differential equation are given. A new composition law on B-series, called substitution law, is presented.

1. Introduction

For an accurate numerical integration of a system of differential equations

$$\dot{y} = f(y), \qquad y(0) = y_0 \tag{1}$$

it is important to use methods of high order (say, at least order 4). Classical approaches for getting high order are multistep, Runge–Kutta, Taylor series, extrapolation, composition, and splitting methods. In this article we present a new approach for constructing high order methods by using modified differential equations.

The idea is the following: for a given one-step method $y_{n+1} = \Phi_{f,h}(y_n)$ (typically very simple to implement, and of order 1 or 2), find a modified differential equation, written as a formal series in powers of the step size h,

$$\dot{y} = \widetilde{f}(y) = f(y) + hf_2(y) + h^2 f_3(y) + \cdots, \qquad y(0) = y_0, \tag{2}$$

such that the numerical solution of the method Φ_h applied to the modified differential equation (2) yields the exact solution of (1) in the sense of formal power series, i.e.,

$$\Phi_{\widetilde{f},h}(y) = \varphi_{f,h}(y). \tag{3}$$

Here, $\varphi_{f,t}(y)$ denotes the exact time-t flow of the problem $\dot{y} = f(y)$.

Once a few coefficient functions $f_j(y)$ are known, this permits us to construct high order integration methods for (1). We suggest the name *modifying integrators* for this approach, because the vector field (1) is modified into (2) before the basic method is applied.

Received by the editor December 5, 2005 and, in revised form, August 1, 2006.
2000 *Mathematics Subject Classification.* Primary 65L06, 65P10, 70E15.
Key words and phrases. Geometric numerical integration, modified differential equation, backward error analysis, modifying integrator, rigid body integrator, B-series, substitution law.

Modifying integrator. *For $r > 1$, consider the truncation*

$$\dot{y} = f^{[r]}(y) = f(y) + hf_2(y) + \cdots + h^{r-1}f_r(y) \tag{4}$$

of the modified equation (2) *for which* (3) *holds. Then,*

$$y_{n+1} = \Phi_{f^{[r]},h}(y_n) \tag{5}$$

defines a numerical method of order r for (1).

An intrinsic feature of this approach is that geometric properties of the flow of (1) which are conserved by the basic method, are in general retained by the high order modifying integrator (see Section 2 below).

There are a few methods that can be cast into the framework of modifying integrators. This is the case for the generating function methods of Feng Kang [4], Feng, Wu, Qin and Wang [5], and Channel and Scovel [1]. There, Hamiltonian systems $f(y) = J^{-1}\nabla H(y)$ in canonical form are considered together with simple symplectic integrators (e.g., symplectic Euler method, or the implicit midpoint rule). It turns out that the modified differential equation is Hamiltonian and can be obtained as a formal solution of the Hamilton–Jacobi partial differential equation (see [6, Sect. VI.5.4]). A recent modification by McLachlan and Zanna [11] of the discrete Moser–Veselov algorithm for solving the Euler equations for the free rigid body can also be interpreted as a modifying method (although it is not constructed in this way).

Modifying integrators will be efficient when the evaluation of the truncated vector field in (4) is not much more expensive than that of $f(y)$. This is definitely the case for the equations of motion for the full dynamics of a rigid body (see Section 3). We shall see later in Section 4 that the coefficient functions $f_j(y)$ depend on derivatives of $f(y)$. McLachlan [10] discusses situations (N-body problems, lattice systems) where the computation of derivatives is cheap when it is performed together with the evaluation of $f(y)$. In these situations the modifying integrators have a large potential.

This paper is organized as follows: The construction of the modified differential equation (2) is discussed in Section 2, where also some important geometric properties are presented. As an example of modifying integrators, a new efficient high-order method (based on the implicit midpoint rule) is developed in Section 3 for the motion of a free rigid body. Many numerical one-step methods (e.g., all Runge–Kutta and Taylor series methods) can be represented as a B-series. For this case, a substitution law for B-series is introduced, which yields general formulae for the modified equation (Section 4), with technical details postponed to Section 5.

2. The modified differential equation

We explain the construction of the modified equation (2), and we discuss how the modified equation inherits the geometric properties of the numerical integrator.

2.1. Construction of the modified equation.

In the following, we assume that the vector field of (1) is infinitely differentiable, and that the numerical integrator $\Phi_{f,h}$ is smooth in h and in f, and of order at least one.

If the basic integrator $\Phi_{f,h}(y)$ is well-defined for all smooth vector fields $f(y)$, then one can simply develop both sides of (3) into a Taylor series around $h = 0$. A comparison of equal powers of h then yields recursively the functions $f_j(y)$ of the

modified differential equation (2). This can conveniently be done with a formula manipulation program such as MAPLE.

It may happen that the basic integrator is only defined for a subclass of differential equations (e.g., the Discrete Moser–Veselov algorithm for the motion of a free rigid body; cf. [7]). In this case, the following recursive construction is in general possible. Suppose that the functions $f_j(y)$ are known for $j = 1, \dots, r$ (we use $f_1(y) = f(y)$). If the basic method is well-defined for the vector field $f^{[r]}(y)$ of (4) (this is certainly the case for $r = 1$) and if it satisfies $\Phi_{f+\varepsilon g,h}(y) = \Phi_{f,h}(y) + h\varepsilon g(y) + \mathcal{O}(h^2\varepsilon)$, the function $f_{r+1}(y)$ is obtained from the relation

$$\Phi_{f^{[r]},h}(y) = \varphi_{f,h}(y) - h^{r+1} f_{r+1}(y) + \mathcal{O}(h^{r+2}). \tag{6}$$

Remark 2.1. The above construction is similar to that for modified differential equations considered in the theory of backward error analysis. There, one interprets the numerical solution $\Phi_{f,h}(y)$ as the exact solution of a modified differential equation of the form (2), i.e.,

$$\varphi_{\widetilde{f},h}(y) = \Phi_{f,h}(y). \tag{7}$$

The only difference between (3) and (7) is that the roles of the integrator Φ and of the exact flow φ are interchanged. Backward error analysis is fundamental for the study of geometric integrators and it is treated in much detail in the monographs of Sanz-Serna and Calvo [13], Hairer, Lubich, and Wanner [6], and Leimkuhler and Reich [9].

2.2. **Geometric properties.** The importance of backward error analysis in the context of geometric numerical integration lies in the fact that properties of numerical integrators are transfered to corresponding properties of modified equations (see [6, Chap. IX]). Due to the close relationship between backward error analysis and our approach of modifying integrators, it is not a surprise that most results of backward error analysis can be extended to our situation. Let us collect the most important properties of the modified equation (2):

- if the numerical integrator $\Phi_{f,h}(y)$ has order p, i.e., the local error satisfies $\Phi_{f,h}(y) - \varphi_{f,h}(y) = \mathcal{O}(h^{p+1})$, then we have $f_j(y) = 0$ for $j = 2, \dots, p$;
- if the integrator $\Phi_{f,h}(y)$ is symmetric, i.e., $\Phi_{f,-h}(y) = \Phi^{-1}_{f,h}(y)$, then the modified differential equation has an expansion in even powers of h, i.e., $f_{2j}(y) = 0$ for all j, and the modifying integrator is symmetric;
- if the basic method $\Phi_{f,h}(y)$ exactly conserves a first integral $I(y)$ of (1), then the modified differential equation has $I(y)$ as first integral, and the modifying integrator exactly conserves $I(y)$;
- if the basic method is symplectic for Hamiltonian systems of the form $\dot{y} = J^{-1}\nabla H(y)$, then the modified differential equation is also Hamiltonian, i.e., $\widetilde{f}(y) = J^{-1}\nabla \widetilde{H}(y)$; the modifying integrator is also symplectic;
- if the basic method is a Poisson integrator for Poisson systems of the form $\dot{y} = B(y)\nabla H(y)$, then the modified differential equation is also a Poisson system with the same structure matrix $B(y)$, and the modifying integrator is a Poisson integrator;
- if the basic method is reversible for reversible differential equations, then the modified differential equation and the modifying integrator are reversible;

- if the basic method is volume preserving for divergence-free differential equations, then the modified differential equation is also divergence-free, and the modifying integrator is volume preserving.

Rigorous proofs of these statements are obtained by adapting those of Theorems IX.1.2, IX.2.2, IX.2.3, IX.3.1, IX.3.5, and Corollary IX.5.4 in [6]. One only has to interchange the roles of the numerical and the exact flows.

3. Modifying midpoint rule for the rigid body

As an example of a modifying integrator, we introduce a new efficient high-order method for the dynamics of a free rigid body. As basic numerical integrator $y_{n+1} = \Phi_{f,h}(y_n)$, we choose the implicit midpoint rule,

$$y_{n+1} = y_n + hf\Big(\frac{y_n + y_{n+1}}{2}\Big). \tag{8}$$

It is a simple symmetric method that exactly preserves quadratic first integrals. For simplicity, we present the modifying implicit midpoint rule of order 6, but the procedure can be extended straightforwardly to higher orders.

3.1. Solving the Euler equations of the rigid body.

The Euler equations of motion for the free rigid body are

$$\begin{aligned} \dot y_1 &= \alpha\, y_2 y_3, & \alpha &= I_3^{-1} - I_2^{-1}, \\ \dot y_2 &= \beta\, y_3 y_1, & \beta &= I_1^{-1} - I_3^{-1}, \\ \dot y_3 &= \gamma\, y_1 y_2, & \gamma &= I_2^{-1} - I_1^{-1}, \end{aligned} \tag{9}$$

where $y_1(t), y_2(t), y_3(t)$ are the angular momenta of the rigid body, and the constants I_1, I_2, I_3 are the three moments of inertia. This system has two quadratic first integrals (Casimir and Hamiltonian)

$$C(y) = \frac{1}{2}\Big(y_1^2 + y_2^2 + y_3^2\Big) \qquad \text{and} \qquad H(y) = \frac{1}{2}\Big(\frac{y_1^2}{I_1} + \frac{y_2^2}{I_2} + \frac{y_3^2}{I_3}\Big). \tag{10}$$

Since the midpoint rule exactly conserves $C(y)$ and $H(y)$, the modified differential equation (2) has these two functions as first integrals (see Section 2.2). Therefore, it is a time transformation of (9). Since the method is also symmetric, it is in even powers of h, and the truncated modified equation (order 6) reduces to

$$\dot y = f^{[5]}(y) = \big(1 + h^2 s_3(y) + h^4 s_5(y)\big) f(y), \tag{11}$$

where $f(y)$ is the right-hand side of (9). The scalar functions $s_3(y), s_5(y)$ can be computed using MAPLE and are given by

$$\begin{aligned} s_3(y) &= -\frac{1}{12}\Big(\beta\gamma y_1^2 + \alpha\gamma y_2^2 + \alpha\beta y_3^2\Big), \\ s_5(y) &= \frac{6}{5} s_3^2(y) + \frac{1}{60}\,\alpha\,\beta\,\gamma\,\left(\beta\, {y_1}^2 {y_3}^2 + \gamma\, {y_2}^2 {y_1}^2 + \alpha\, {y_3}^2 {y_2}^2\right). \end{aligned} \tag{12}$$

Notice that the scalar functions $s_3(y)$ and $s_5(y)$ are not constant along a particular solution (except in the case of a symmetric body). A modified equation of the same structure has been studied in [14] in the context of backward error analysis.

3.2. **The full dynamics: the configuration update.** To obtain the full dynamics of the free rigid body, one has to solve the augmented differential equation

$$\begin{pmatrix}\dot y\\ \dot Q\end{pmatrix} = \begin{pmatrix} f(y)\\ QW(y)\end{pmatrix} \qquad \text{with} \qquad W(y) = \begin{pmatrix} 0 & -\frac{y_3}{I_3} & \frac{y_2}{I_2}\\ \frac{y_3}{I_3} & 0 & -\frac{y_1}{I_1}\\ -\frac{y_2}{I_2} & \frac{y_1}{I_1} & 0\end{pmatrix}, \tag{13}$$

where $Q(t)$ is an orthogonal matrix that gives the position of the body in the fixed coordinate system at time t. The modified vector field for the implicit midpoint rule is given by

$$\begin{pmatrix}\dot y\\ \dot Q\end{pmatrix} = \begin{pmatrix} f^{[5]}(y)\\ QW^{[5]}(y)\end{pmatrix}, \tag{14}$$

where $f^{[5]}(y)$ is the vector field of (11) and the skew-symmetric matrix $W^{[5]}(y)$ is given by

$$W^{[5]}(y) = W(y^{[5]}). \tag{15}$$

Here, $y^{[5]}$ is the vector with components

$$y_j^{[5]} \;=\; y_j\Big(1 + h^2\big(s_3(y) + I_j\, d_3(y)\big) + h^4\big(s_5(y) + I_j\, d_5(y)\big)\Big), \quad j = 1,2,3,$$

where $s_3(y)$ and $s_5(y)$ are the functions of (12), and (using MAPLE)

$$\begin{aligned} d_3(y) \;&=\; \frac{1}{3\Delta}\Big(-C(y) + \delta_0 H(y)\Big),\\ d_5(y) \;&=\; \frac{1}{30\Delta}\Big(\delta_1 C(y)^2 + \delta_2 C(y)H(y) + \delta_3 H(y)^2 + y_1^2\big(\delta_4 C(y) + \delta_5 H(y)\big)\Big).\end{aligned}$$

The constants Δ, $\delta_0, \ldots, \delta_5$ only depend on the three moments of inertia I_1, I_2, I_3, and are given by

$$\begin{aligned} &\Delta = I_1 I_2 I_3, && \delta_2 = \frac{1}{\Delta}\big(2I_2^2 + 2I_3^2 - 3I_1^2\big) + \frac{8}{I_1} - \frac{7}{I_2} - \frac{7}{I_3},\\ &\delta_0 = \frac{1}{2}\big(I_1 + I_2 + I_3\big), && \delta_3 = 3 + 2\frac{I_1 + I_3}{I_2} + 2\frac{I_1 + I_2}{I_3} - 3\frac{I_2 + I_3}{I_1},\\ &\delta_1 = \frac{1}{\Delta}\big(10I_1 - 6\delta_0\big), && \delta_4 = 5\Big(\frac{1}{I_1} - \frac{1}{I_3}\Big)\Big(\frac{1}{I_2} - \frac{1}{I_1}\Big), \qquad \delta_5 = -\delta_0\delta_4.\end{aligned}$$

Since the vectors y and $y^{[5]}$ are not collinear, the modified equation (14) is not a time transformation of the original system (except in the case of a symmetric body).

Applying the implicit midpoint rule to the system (14) thus yields a numerical integrator of order 6 for the full dynamics of the free rigid body.

3.3. **Efficient implementation.** Since $C(y)$ and $H(y)$ are two invariants, for the modifying integrator of order 4 (and similarly for higher orders) it is possible to avoid some costly multiplications in (12) for the computation of $s_3(y)$ by writing it in the form

$$s_3(y) = c_1 C(y) + c_2 H(y) + c_3\, y_1^2, \tag{16}$$

where the constants c_j only depend on I_1, I_2, I_3, and can be calculated once for all. Then, when using a fixed point iteration to compute the internal stage,

$$Y = \frac{y_n + y_{n+1}}{2}, \tag{17}$$

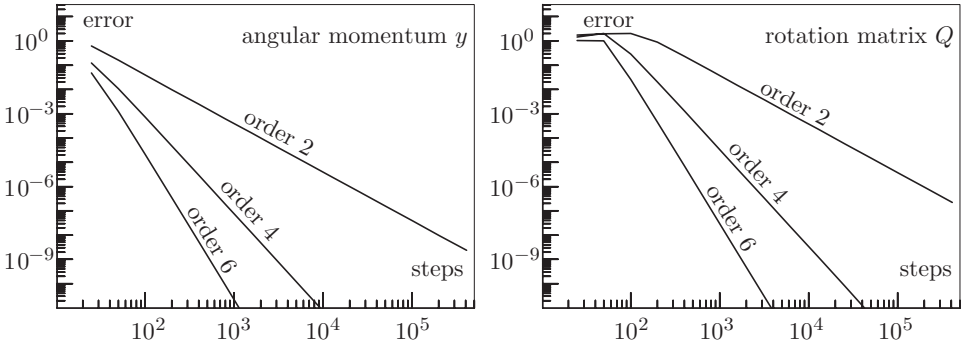

FIGURE 1. Work-precision diagram for the implicit midpoint rule (order 2) and its modifying versions of orders 4 and 6.

it is not necessary to evaluate $C(Y)$ and $H(Y)$ with the formulae (10). Indeed, one can use the estimates $C(y_n)$ and $H(y_n)$ instead of $C(Y)$ and $H(Y)$,

$$s_3(Y) \approx c_1 C(y_n) + c_2 H(y_n) + c_3\, Y_1^2,$$

where Y_1 is the first component of Y. The method is still symmetric because $C(y_n) = C(y_{n+1})$, and the order remains 4 since $C(Y) = C(y_n) + \mathcal{O}(h^2)$ (and similarly for $H(Y)$).

We now turn our attention to the computation of the configuration update. For an efficient implementation, it is a standard approach to use quaternions to represent orthogonal matrices (see [6] in the context of rigid body integrators implementations). This reduces the midpoint rule,

$$Q_{n+1} = Q_n + h\Big(\frac{Q_n + Q_{n+1}}{2}\Big)\, W^{[5]}(Y),$$

where $W^{[5]}$ in given in (15) and Y is defined in (17), to a simple multiplication of quaternions through the equivalent formulation

$$Q_{n+1} = Q_n\, \Omega.$$

Here, Ω is the orthogonal matrix defined by the Cayley transform

$$\Omega = \Big(I + \frac{h}{2} W^{[5]}(Y)\Big)\Big(I - \frac{h}{2} W^{[5]}(Y)\Big)^{-1}$$

which can be represented by the quaternion $\frac{\omega}{\|\omega\|}$ of norm 1 given by

$$\omega = 1 + \frac{h}{2}\Big(i\frac{Y_1^{[5]}}{I_1} + j\frac{Y_2^{[5]}}{I_2} + k\frac{Y_3^{[5]}}{I_3}\Big).$$

Numerical experiment. We consider the system (13) for the free rigid body on the interval $[0, 100]$, and we use $I_1 = 0.9144$, $I_2 = 1.0980$, $I_3 = 1.6600$, and initial values $y(0) = (0.4165, 0.9072, 0.0577)^T$ as in [11]. As numerical integrators we apply the standard implicit midpoint rule and also the modifying versions of orders 4 and 6. The errors as a function of the computational work (number of steps) are drawn as solid lines in Figure 1.

We are also curious to see how much work the modifying versions require with respect to the standard application of the midpoint rule. For this, we have carefully implemented the implicit midpoint rule IMR2 and the modifying versions IMR4 and

IMR6 of orders 4 and 6 (using quaternions for the rotation matrices). Table 1 shows the cpu time (normalized with respect to that of IMR2) of the different implementations, and also the error in the angular momentum for three different choices of the step size. Although the numbers should not be overestimated, one clearly sees that IMR4 needs not more than twice and IMR6 not more than 2.5 times the work of IMR2. This is cheaper than what can be expected for either s-stage Runge–Kutta methods of the same order or for composition methods. FORTRAN codes for the modifying implicit midpoint rule introduced in this article can be obtained from the authors on request.

TABLE 1. Normalized computational work and accuracy

	IMR2		IMR4		IMR6	
nstep	work	error	work	error	work	error
100	1.0	$4.0 \cdot 10^{-2}$	1.5	$7.4 \cdot 10^{-4}$	1.8	$2.1 \cdot 10^{-5}$
400	1.0	$2.5 \cdot 10^{-3}$	1.9	$3.0 \cdot 10^{-6}$	2.5	$5.4 \cdot 10^{-9}$
1600	1.0	$1.5 \cdot 10^{-4}$	1.8	$1.2 \cdot 10^{-8}$	2.2	$1.3 \cdot 10^{-12}$

4. Analysis for B-series methods

The discrete flow of many numerical integrators (including Runge–Kutta methods) can be expanded into a B-series as introduced and studied in [8]. We follow the notation of [6, Chap. III], where a more comprehensive presentation of this theory is given.

4.1. **Substitution law for B-series vector fields.** Let $T = \{\bullet, \tau_2, \tau_3, \ldots\}$ be the set of rooted trees, and let $\emptyset$ be the empty tree. For $\tau_1, \ldots, \tau_m \in T$, we denote by $\tau = [\tau_1, \ldots, \tau_m]$ the tree obtained by grafting the roots of $\tau_1, \ldots, \tau_m$ to a new vertex which becomes the root of τ. The order $|\tau|$ of a tree τ is its number of vertices and its symmetry coefficient is defined recursively by

$$\sigma(\bullet) = 1, \qquad \sigma(\tau) = \sigma(\tau_1) \cdots \sigma(\tau_m) \mu_1! \mu_2! \cdots , \tag{18}$$

where the integers $\mu_1, \mu_2, \ldots$ count equal trees among $\tau_1, \ldots, \tau_m$. Eventually, elementary differentials $F_f(\tau)$ are given by

$$F_f(\bullet)(y) = f(y), \qquad F_f(\tau)(y) = f^{(m)}(y)\big(F_f(\tau_1)(y), \ldots, F_f(\tau_m)(y)\big). \tag{19}$$

For real coefficients $a(\emptyset)$ and $a(\tau), \tau \in T$, a B-series is a series of the form

$$\mathcal{B}(f, a) = a(\emptyset)\, Id + \sum_{\tau \in T} \frac{h^{|\tau|}}{\sigma(\tau)} a(\tau) F_f(\tau), \tag{20}$$

where Id stands for the identity $Id\,(y) = y$. The Taylor series of the exact solution of (1) can be written as a B-series $y(h) = \mathcal{B}(f, e)(y_0)$ with coefficients $e(\tau) = \gamma(\tau)^{-1}$, where

$$\gamma(\bullet) = 1, \qquad \gamma(\tau) = |\tau|\, \gamma(\tau_1) \cdots \gamma(\tau_m) \ . \tag{21}$$

The flow $y_{n+1} = \Phi_{f,h}(y_n)$ of a Runge–Kutta method is of the form $\Phi_{f,h} = \mathcal{B}(f, a)$ with $a(\tau)$ depending only on the coefficients of the method (see [6, Chap. III] for more details).

With the aim of unifying the theory of this article with backward error analysis, we let (2) be the modified equation defined by

$$\Phi_{\widetilde{f},h}(y) = \Psi_{f,h}(y) \tag{22}$$

where Φ and Ψ are two numerical integrators that can be expressed as B-series $\Phi_{f,h} = \mathcal{B}(f,a)$ and $\Psi_{f,h} = \mathcal{B}(f,c)$. For $\Psi_{f,h}(y) = \varphi_{f,h}(y)$ we recover formula (3), and for $\Phi_{\widetilde{f},h}(y) = \varphi_{\widetilde{f},h}(y)$ we get (7).

In terms of B-series, formula (22) becomes $\mathcal{B}(\widetilde{f},a) = \mathcal{B}(f,c)$. When computing recursively some of the coefficient functions of (2), one is quickly convinced that they are linear combinations of elementary differentials and that $\widetilde{f}(y) = h^{-1}\mathcal{B}(f,b)(y)$ with coefficients $b(\tau)$ that have to be determined (notice that we necessarily have $b(\emptyset) = 0$). This motivates the following theorem, introduced in [2].

Theorem 4.1. *For $b(\emptyset) = 0$, the vector field $h^{-1}\mathcal{B}(f,b)$ inserted into $\mathcal{B}(\cdot,a)$ gives a B-series*

$$\mathcal{B}\big(h^{-1}\mathcal{B}(f,b),a\big) = \mathcal{B}(f,b\star a).$$

We have $(b\star a)(\emptyset) = a(\emptyset)$, some further coefficients are given in Table 2, and a general formula for $(b\star a)(\tau)$ is given in (26) of Section 5 below.

TABLE 2. Coefficients of the substitution law for B-series vector fields.

$(b\star a)(\emptyset)$	$=$	$a(\emptyset)$
$(b\star a)(\bullet)$	$=$	$a(\bullet)b(\bullet)$
$(b\star a)(\bullet\!\!-\!\!\bullet)$	$=$	$a(\bullet)b(\bullet\!\!-\!\!\bullet) + a(\bullet\!\!-\!\!\bullet)b(\bullet)^2$
$(b\star a)(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet)$	$=$	$a(\bullet)b(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet) + 2a(\bullet\!\!-\!\!\bullet)b(\bullet)b(\bullet\!\!-\!\!\bullet) + a(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet)b(\bullet)^3$
$(b\star a)(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet)$	$=$	$a(\bullet)b(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet) + 2a(\bullet\!\!-\!\!\bullet)b(\bullet)b(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet) + a(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet)b(\bullet)^3$

We postpone the proof of this theorem to Section 5, and briefly discuss some of the most important properties and applications. Further properties may be found in [2].

The question of finding the modified equation defined by (22), i.e., of finding the coefficients $b(\tau)$ for given $a(\tau)$ and $c(\tau)$ in the relation

$$\mathcal{B}\big(h^{-1}\mathcal{B}(f,b),a\big) = \mathcal{B}(f,c),$$

results in solving for $b(\tau)$ the algebraic system

$$(b\star a)(\tau) = c(\tau) \qquad \text{for} \quad \tau \in T. \tag{23}$$

We notice that

$$(b\star a)(\tau) = a(\bullet)b(\tau) + \cdots + a(\tau)b(\bullet)^{|\tau|},$$

where the three dots involve only trees of order strictly less than $|\tau|$. Consequently, for consistent integrators $\Phi_{f,h} = \mathcal{B}(f,a)$ and $\Psi_{f,h} = \mathcal{B}(f,c)$, for which $a(\emptyset) = a(\bullet) = 1$ and $c(\emptyset) = c(\bullet) = 1$, the coefficients $b(\tau)$ can be computed recursively from (23). In this way, the computation of the vector fields $f_j(y)$ in the modified differential equation (2) is reduced to that of real coefficients.

Modifying integrators. In this case $\Psi_{f,h}$ is the exact h-flow of (1) which is a B-series with coefficients $e(\tau) = \gamma(\tau)^{-1}$. Consequently, the coefficients $b(\tau)$ of the modified differential equation for $\Phi_{f,h} = \mathcal{B}(f, a)$ are obtained from

$$(b \star a)(\tau) = e(\tau) \qquad \text{for} \quad \tau \in T. \tag{24}$$

Backward error analysis. The modified differential equation of a method $\Psi_{f,h} = \mathcal{B}(f, c)$ is obtained by putting $\Phi_{f,h}$ equal to the exact flow. Its coefficients $b(\tau)$ are therefore obtained from

$$(b \star e)(\tau) = c(\tau) \qquad \text{for} \quad \tau \in T.$$

Remark 4.2. The B-series $h^{-1}\mathcal{B}(f, b)$ of mappings $b : T \cup \{\emptyset\} \to \mathbb{R}$ with $b(\emptyset) = 0$ represent vector fields. The product $b \star a$ defines a group structure on the set $\big\{c : T \cup \{\emptyset\} \to \mathbb{R}\,;\, c(\emptyset) = 0, c(\bullet) = 1\big\}$ which represents such vector fields. Its unit element is given by $c(\bullet) = 1$ and $c(\tau) = 0$ for $|\tau| > 1$, and it corresponds to the original vector field $f(y)$.

We mention that the presented theory can be extended straightforwardly to partitioned integration methods (P-series). This is particularly important for the consideration of symplectic integrators.

4.2. **Modifying implicit midpoint rule.** As an example, consider the implicit midpoint rule (8) which admits a B-series expansion $\mathcal{B}(f, a)$ with $a(\tau) = \big(\frac{1}{2}\big)^{|\tau|-1}$. Determining the functions $f_3(y)$ and $f_5(y)$ in the modified differential equation (2) amounts to computing (up to order 5) the coefficients $b(\tau)$ of the B-series $\mathcal{B}(f, b)$ from the relation (24). The formulae of Table 2 yield

$$b(\bullet) = 1, \qquad b(\rotatebox{90}{τ}) = 0, \qquad b(\rotatebox{90}{τ^2}) = \frac{1}{12}, \qquad b(\rotatebox{90}{γ}) = -\frac{1}{12}.$$

The coefficients for trees of order 4 vanish due to the symmetry of the method, and those for order 5 can be calculated from (26). We thus arrive at the following modified vector field:

$$\begin{aligned} f^{[5]} = f \;+\; & \frac{h^2}{12}\Big(- f'f'f + \frac{1}{2}f''(f,f)\Big) \\ & + \frac{h^4}{120}\Big(f'f'f'f'f - f''(f,f'f'f) + \frac{1}{2}f''(f'f,f'f)\Big) \\ & + \frac{h^4}{240}\Big(- \frac{1}{2}f'f'f''(f,f) + f'f''(f,f'f) + \frac{1}{2}f''(f,f''(f,f))\Big) \\ & + \frac{h^4}{240}\Big(- \frac{1}{2}f^{(3)}(f,f,f'f) - \frac{1}{2}f'f^{(3)}(f,f,f) + \frac{1}{8}f^{(4)}(f,f,f,f)\Big). \end{aligned} \tag{25}$$

This formula reduces to (11) for the Euler equations and to (14) for the full dynamics of the rigid body.

4.3. **Elementary differential Runge–Kutta methods.** The idea of modifying integrators applied to Runge–Kutta methods provides an easy way to construct high-order methods for the numerical solution of (1). Methods obtained in this manner are a particular case of the so-called *elementary differential Runge-Kutta methods* (EDRK), introduced by Murua [12].

Consider an s-stage Runge–Kutta method $y_{n+1} = \Phi_{f,h}(y_n)$ of order p. It admits a B-series expansion $\Phi_{f,h} = \mathcal{B}(f, a)$. Applying this method to the modified vector field $f^{[r]}(y)$, truncated at some order r greater than p, leads to an r-derivative EDRK method of order (at least) r given by

$$\begin{aligned} Y_i &= y_n + h\sum_{j=1}^{s} a_{ij} f^{[r]}(Y_j), \quad i = 1, \dots, s, \\ y_{n+1} &= y_n + h\sum_{j=1}^{s} b_j f^{[r]}(Y_j). \end{aligned}$$

By Theorem 4.1 the modified vector field is a B-series

$$\begin{aligned} f^{[r]}(y) &= f(y) + h^p f_{p+1}(y) + \cdots + h^{r-1} f_r(y), \\ f_j(y) &= \sum_{|\tau|=j} \frac{b(\tau)}{\sigma(\tau)} F_f(\tau)(y), \quad j = 1, \dots, r. \end{aligned}$$

Its coefficients $b(\tau)$ are obtained from the relation (24).

Example 4.3. Consider the s-stage Runge–Kutta method of order $p = 2s$ (Gauss method). Since it is symplectic and symmetric, the modified vector field $f^{[r]}(y)$ is Hamiltonian for all Hamitonian systems $\dot{y} = J^{-1}\nabla H(y)$, and we have $f_{2j}(y) = 0$ for all j (see Section 2.2). Then, if we take an odd integer r, we obtain an implicit symplectic and symmetric r-derivative EDRK method of order (at least) $r+1$. The special case $s = 1$ yields the symplectic generating function methods based on the implicit midpoint rule [5].

For instance, for $s = 2,\ r = 5$, we obtain a 5-derivative EDRK method of order 6, and coefficients $b(\tau)$ for trees of order $|\tau| = 5$ are given by

$b($  $) = \frac{1}{180}, \quad b($  $) = \frac{1}{360}, \quad b($  $) = \frac{1}{720},$

together with the algebraic conditions on the coefficients $b(\tau)$ for $f^{[r]}(y)$ to be a Hamiltonian vector field (see [6, Section IX.9.2]).

Remark 4.4. It would be interesting to know whether there exist symplectic (and symmetric) EDRK methods that are not modifying classical Runge–Kutta methods and have an order higher than $\max(2s, r + 1)$.

5. An explicit formula for the substitution law

In this section, we give a computation formula for the substitution law of B-series introduced in Section 4.1. We begin with some definitions.

5.1. **Partitions and skeletons.** A *partition* p^τ of a tree τ is obtained by cutting some of its edges [2]. The resulting list of trees is denoted $P(p^\tau)$. Eventually, the set of all partitions p^τ of τ is denoted $\mathcal{P}^\tau$. Now, given a partition p^τ, the corresponding *skeleton* $\chi(p^\tau)$, as introduced in [3], is the tree obtained by contracting each tree of $P(p^\tau)$ to a single vertex $\bullet$ and by re-establishing the cut edges (see Table 3). We observe that a tree $\tau \in T$ has exactly $2^{|\tau|-1}$ partitions $p^\tau \in \mathcal{P}^\tau$, and that different partitions may lead to the same list $P(p^\tau)$.

TABLE 3. The 8 partitions of a tree of order 4 with associated functions

p^τ								
$\chi(p^\tau)$								
$P(p^\tau)$								

5.2. **The substitution law formula.** We are now in a position to state the main result of this section. The coefficients $(b \star a)(\tau)$ of the substitution law can be expressed in terms of the coefficients $a(\theta)$ and $b(\theta)$ with $|\theta| \leq |\tau|$ in the following polynomial expression:

$$(b \star a)(\tau) = \sum_{p^\tau \in \mathcal{P}(\tau)} a\big(\chi(p^\tau)\big) \prod_{\delta \in P(p^\tau)} b(\delta) \tag{26}$$

for $\tau \in T$. For the example of Table 3, this formula yields

$$\begin{aligned} (b \star a)(\ ) &= a(\ )b(\ ) + a(\ )b(\ )b(\ ) + 2a(\ )b(\ )b(\ ) \\ &+ a(\ )b(\ )^2 b(\ ) + 2a(\ )b(\ )^2 b(\ ) + a(\ )b(\ )^4. \end{aligned}$$

5.3. **Proof of the substitution law formula.** Multiplying (27) with $a(\theta)$ and summing up yields $\mathcal{B}(g, a) = \mathcal{B}(f, b \star a)$. Formula (26) is thus obtained by multiplying (28) with $a(\theta)$ and summing up. It therefore remains to prove the following lemma.

Lemma 5.1. *Let $g(y) = h^{-1}\mathcal{B}(f, b)(y)$ be a (h-dependent) vector field defined by a B-series with $b(\emptyset) = 0$. Then, for $\theta \in T$, we have*

$$\frac{h^{|\theta|}}{\sigma(\theta)} F_g(\theta) = \mathcal{B}(f, b_\theta), \tag{27}$$

where the coefficients $b_\theta(\tau)$ are given by $b_\theta(\emptyset) = 0$, and for $\tau \in T$,

$$b_\theta(\tau) = \sum_{p^\tau \in \mathcal{P}^\tau,\ \chi(p^\tau) = \theta} \ \prod_{\delta \in P(p^\tau)} b(\delta). \tag{28}$$

Before proving this lemma, we need Lemma 5.2 below, which requires a few more definitions illustrated in Table 4. Given a partition p^τ of a tree τ, the tree of $P(p^\tau)$ which contains the root of τ is denoted $r(p^\tau)$. For brevity of formulae, we further use $P^*(p^\tau) = P(p^\tau) \backslash \{r(p^\tau)\}$. A partition p^τ is said to be *admissible* if the path from the root to any vertex has at most one cut. The set of admissible partitions of τ is denoted $\mathcal{AP}^\tau$.

Lemma 5.2. *Let $g(y)$ be defined by $g(y) = h^{-1}\mathcal{B}(f, b)(y)$ with $b(\emptyset) = 0$. Then, for $\delta = [\delta_1, \ldots, \delta_m] \in T$, we have*

$$\frac{h^{|\delta|}}{\sigma(\delta)}\ g^{(m)}(y)\ \Big(F_f(\delta_1)(y), \ldots, F_f(\delta_m)(y)\Big) = \mathcal{B}(f, d_\delta b)(y), \tag{29}$$

TABLE 4. The 8 partitions of a tree of order 4 with other associated functions

p^τ								
$r(p^\tau)$								
$P^*(p^\tau)$	$\emptyset$							
$p^\tau \in \mathcal{AP}^\tau$?	yes	yes	yes	yes	yes	no	no	no

where $d_\delta b(\tau)$ is defined by $d_\delta b(\emptyset) = 0$, and for $\tau \in T$,

$$d_\delta b(\tau) = \sum_{p^\tau \in \mathcal{AP}^\tau,\, P^*(p^\tau)=\{\delta_1,\ldots,\delta_m\}} b\big(r(p^\tau)\big). \tag{30}$$

Proof. The proof follows closely that of Lemma IX.9.1 in [6] and it is thus omitted. Notice that admissible partitions correspond to ordered subtrees in [6]. □

Proof of Lemma 5.1. We proceed by induction on $|\theta|$. From $F_g(\bullet) = g = h^{-1}\mathcal{B}(f,b)$ we have $b_\bullet(\tau) = b(\tau)$ for all $\tau \in T$. Consider now a tree $\theta = [\theta_1, \ldots, \theta_m]$ with $|\theta| \geq 2$, and assume (27) and (28) are satisfied for trees of order strictly less than $|\theta|$. By definition of $F_g(\theta)$ and multi-linearity of $g^{(m)}(y)(\cdot,\ldots,\cdot)$, we have

$$\begin{aligned}\frac{h^{|\theta|}}{\sigma(\theta)} F_g(\theta)(y) &= \frac{\sigma(\theta_1)\cdots\sigma(\theta_m)}{\sigma(\theta)} \sum_{\tau_1,\ldots,\tau_m \in T} \frac{1}{\sigma(\tau_1)\cdots\sigma(\tau_m)} \left(\prod_{i=1}^m b_{\theta_i}(\tau_i)\right) \\ &\quad \cdot h^{|\tau|} g^{(m)}(y)(F_f(\tau_1)(y), \ldots, F_f(\tau_m)(y))\end{aligned}$$

with $\tau = [\tau_1, \ldots, \tau_m]$. Formula (29) of Lemma 5.2 then gives, for $v \in T$,

$$b_\theta(v) = \frac{\sigma(\theta_1)\cdots\sigma(\theta_m)}{\sigma(\theta)} \sum_{\tau_1,\ldots,\tau_m \in T} \frac{\sigma(\tau)}{\sigma(\tau_1)\cdots\sigma(\tau_m)} \left(\prod_{i=1}^m b_{\theta_i}(\tau_i)\right) d_\tau b(v).$$

Now, taking into account the fact that permutations among $\tau_1, \ldots, \tau_m$ do not change the tree $\tau = [\tau_1, \ldots, \tau_m]$ (and similarly for θ), it follows that

$$b_\theta(v) = \sum_{\tau=[\tau_1,\ldots,\tau_m]\in T} \sum_{\substack{\theta_1,\ \ldots,\ \theta_m \in T,\\ [\theta_1,\ \ldots,\ \theta_m] = \theta}} \left(\prod_{i=1}^m b_{\theta_i}(\tau_i)\right) d_\tau b(v)$$

and formula (30) allows one to write

$$b_\theta(v) = \sum_{[\tau_1,\ldots,\tau_m]\in T} \sum_{\substack{p^v \in \mathcal{AP}^v,\\ P^*(p^v) = \{\tau_1,\ \ldots,\ \tau_m\}}} \sum_{\substack{\theta_1,\ \ldots,\ \theta_m \in T,\\ [\theta_1,\ \ldots,\ \theta_m] = \theta}} b\big(r(p^v)\big) \prod_{i=1}^m b_{\theta_i}(\tau_i).$$

Using the induction hypothesis we eventually obtain

$$\begin{aligned} b_\theta(v) &= \sum_{\substack{[\tau_1,\ldots,\tau_m]\in T,\\ p^v\in\mathcal{AP}^v,\\ P^*(p^v)=\{\tau_1,\ \ldots,\ \tau_m\}}} \ \sum_{\substack{p^{\tau_1}\in\mathcal{P}^{\tau_1},\ \ldots,\ p^{\tau_m}\in\mathcal{P}^{\tau_m},\\ [\chi(p_1^\tau),\ \ldots,\ \chi(p_m^\tau)]=\theta}} b\big(r(p^v)\big) \prod_{\delta\in\bigcup_{i=1}^m P(p^{\tau_i})} b(\delta) \\ &= \sum_{p^v\in\mathcal{P}^v,\,\chi(p^v)=\theta} \ \prod_{\delta\in P(p^v)} b(\delta), \end{aligned}$$

which proves the statement of Lemma 5.1. □

Acknowledgment

We are grateful to the participants of the numerical analysis seminar in Geneva for helpful discussions. This work was partially supported by the Fonds National Suisse, project No. 200020-109158.

References

[1] P. J. Channell and J. C. Scovel. Symplectic integration of Hamiltonian systems. *Nonlinearity*, 3:231–259, 1990. MR1054575 (91g:58073)

[2] P. Chartier, E. Hairer, and G. Vilmart. A substitution law for B-series vector fields. *INRIA Report, No. 5498*, 2005.

[3] P. Chartier and E. Lapôtre. Reversible B-series. *INRIA Report, No. 1221*, 1998.

[4] K. Feng. Difference schemes for Hamiltonian formalism and symplectic geometry. *J. Comp. Math.*, 4:279–289, 1986. MR860157 (88a:65094)

[5] K. Feng, H. M. Wu, M.-Z. Qin, and D. L. Wang. Construction of canonical difference schemes for Hamiltonian formalism via generating functions. *J. Comp. Math.*, 7:71–96, 1989. MR1017182 (90j:58043)

[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, second edition, 2006. MR2221614 (2006m:65006)

[7] E. Hairer and G. Vilmart. Preprocessed discrete Moser–Veselov algorithm for the full dynamics of the free rigid body. J. Phys. A 39 (2006), no. 42, 13225–13235. MR2266054

[8] E. Hairer and G. Wanner. On the Butcher group and general multi-value methods. *Computing*, 13:1–15, 1974. MR0403225 (53:7037)

[9] B. Leimkuhler and S. Reich. *Simulating Hamiltonian Dynamics*. Cambridge Monographs on Applied and Computational Mathematics 14. Cambridge University Press, Cambridge, 2004. MR2132573 (2006a:37078)

[10] R. I. McLachlan. A new implementation of symplectic Runge–Kutta methods. To appear in *SIAM J. Sci. Comput.*

[11] R. I. McLachlan and A. Zanna. The discrete Moser–Veselov algorithm for the free rigid body, revisited. *Found. Comput. Math.*, 5:87–123, 2005. MR2125692 (2005k:37189)

[12] A. Murua. *Métodos simplécticos desarrollables en P-series*. Ph.D. thesis, Univ. Valladolid, 1994.

[13] J. M. Sanz-Serna and M. P. Calvo. *Numerical Hamiltonian Problems*. Chapman & Hall, London, 1994. MR1270017 (95f:65006)

[14] A. Zanna. A note on the implicit midpoint rule and the Euler equations for the rigid body. *Private communication*, 2005.

INRIA Rennes, Campus Beaulieu, F-35042 Rennes, Cedex, France
E-mail address: Philippe.Chartier@irisa.fr

Section de Mathématiques, Université de Genève, CH-1211 Genève 4, Switzerland
E-mail address: Ernst.Hairer@math.unige.ch

ENS Cachan Bretagne, Campus Ker-Lann, av. Robert Schumann, F-35170 Bruz, France
E-mail address: Gilles.Vilmart@irisa.fr