Journal of Computational and Applied Mathematics 236 (2012) 3890–3904

Contents lists available at SciVerse ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Energy-preserving methods for Poisson systems[☆]

L. Brugnano[a], M. Calvo[b], J.I. Montijano[b,*], L. Rández[b]

[a] Dipartimento di Matematica "U. Dini", Università di Firenze, Italy
[b] IUMA – Departamento de Matemática Aplicada, Universidad de Zaragoza, Spain

ARTICLE INFO

Article history:
Received 15 September 2011
Received in revised form 21 January 2012

In memory of Professor Donato Trigiante

MSC:
65P10
65L05

Keywords:
Ordinary differential equations
One-step methods
Poisson problems
Hamiltonian boundary value methods
Energy-preserving methods
Line integral methods

ABSTRACT

We present and analyze energy-conserving methods for the numerical integration of IVPs of Poisson type that are able to preserve some Casimirs. Their derivation and analysis is done following the ideas of Hamiltonian BVMs (HBVMs) (see Brugnano et al. [10] and references therein). It is seen that the proposed approach allows us to obtain the methods recently derived in Cohen and Hairer (2011) [17], giving an alternative derivation of such methods and a new proof of their order. Sufficient conditions that ensure the existence of a unique solution of the implicit equations defining the formulae are given. A study of the implementation of the methods is provided. In particular, order and preservation properties when the involved integrals are approximated by means of a quadrature formula, are derived.

1. Introduction

In this paper we deal with the numerical solution of Initial Value Problems (IVPs) in ordinary differential systems where the vector field can be written in the gradient form, i.e.,

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{y}(t) = \mathbf{f}(\boldsymbol{y}(t)) \equiv B(\boldsymbol{y}(t))\ \nabla H(\boldsymbol{y}(t)), \quad t \in [0,h], \\ \boldsymbol{y}(0) = \boldsymbol{y}_0 \in \mathbb{R}^m, \end{cases} \tag{1}$$

where $B(\boldsymbol{y})$ is a skew-symmetric matrix and $H = H(\boldsymbol{y})$ a scalar function that will be called the Hamiltonian function. We shall assume that (1) has a unique solution $\boldsymbol{y} = \boldsymbol{y}(t)$, $t \in [0,h]$ and that both $B(\boldsymbol{y})$ and $H(\boldsymbol{y})$ are sufficiently smooth in a suitable neighborhood around such a solution.

The differential system of (1) has the first integral $H(\boldsymbol{y}) = \text{const}$, and numerical methods that preserve this integral are usually called energy-preserving methods. Further, any scalar function $C = C(\boldsymbol{y})$ such that $\nabla C(\boldsymbol{y})^T\ B(\boldsymbol{y}) \equiv 0$, called a Casimir function, is as well a first integral for (1).

In this paper we study numerical methods that provide a vector-valued polynomial approximation $\boldsymbol{u} = \boldsymbol{u}(t) \simeq \boldsymbol{y}(t)$, $t \in [0,h]$ to the solution of (1) preserving the energy in the sense that $H(\boldsymbol{u}(0)) = H(\boldsymbol{u}(h))$ and eventually possible Casimirs of (1).

[☆] Work developed within the project "Numerical methods and software for differential equations".

[*] Corresponding author.
E-mail address: monti@unizar.es (J.I. Montijano).

0377-0427/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2012.02.033

A natural way to preserve a first integral $G(\boldsymbol{y})$ of a general differential system $\dot{\boldsymbol{y}} = \mathbf{f}(\boldsymbol{y})$ in the numerical integration with a one-step method φ_h is, after each step $(t_0, \boldsymbol{y}_0) \rightarrow (t_1 = t_0 + h, \boldsymbol{y}_1)$, to project the numerical solution $\boldsymbol{y}_1 = \varphi_h(\boldsymbol{y}_0)$ onto the manifold $\{\boldsymbol{y}; G(\boldsymbol{y}) = G(\boldsymbol{y}_0)\}$. In the case of orthogonal projection the new approximation $\tilde{\boldsymbol{y}}_1$ is defined by $\tilde{\boldsymbol{y}}_1 = \boldsymbol{y}_1 + \lambda\ \nabla G(\boldsymbol{y}_1)$, where λ is a Lagrange multiplier to be chosen so that $G(\tilde{\boldsymbol{y}}_1) = G(\boldsymbol{y}_0)$. This approach can be used for any numerical method but it requires the solution of a non linear equation at each step. Although this projection preserves the order of the original method its main drawback is that it may destroy other good properties of the original method, in particular the affine invariance. Because of this the authors [1] have proposed for Runge–Kutta methods a different projection that is affine invariant and, as shown in some numerical experiments, provides better results than the standard orthogonal projection.

On the other hand, in the special case $m = 2d$ and $B(\boldsymbol{y}) = J = \begin{pmatrix} 0_d & I_d \\ -I_d & 0_d \end{pmatrix}$, the standard $(2d)$-dim skew-symmetric matrix of Hamiltonian dynamics, (1) defines a canonical Hamiltonian system and the derivation of methods that preserve the energy in Hamiltonian systems has been the subject of an extensive research in the last years. Among them, we may mention the so called "averaged vector field" of Quispel and McLaren [2] in which the solution of (1) at $t = \tau h$ is approximated by the first degree polynomial $\boldsymbol{u}(\tau h) = (1-\tau)\boldsymbol{y}_0 + \tau \boldsymbol{y}_1,\ \tau \in [0, 1]$, where $\boldsymbol{y}_1$ is the solution of the implicit equation

$$\boldsymbol{y}_1 = \boldsymbol{y}_0 + h \int_0^1 J\ \nabla H((1-\tau)\boldsymbol{y}_0 + \tau \boldsymbol{y}_1) \mathrm{d}\tau. \tag{2}$$

It can be seen that this method is energy-preserving because it satisfies $H(\boldsymbol{y}_1) = H(\boldsymbol{y}_0)$. Observe that (2) amounts to substitute in the integral equation equivalent to (1) with $B = J$, the average along the exact solution $\boldsymbol{y}(t)$ by the average along the first order polynomial $\boldsymbol{u}(\tau h)$. As a further remark, note that for some Hamiltonian functions the right hand side of (2) can be integrated exactly and the computational cost of (2) is similar to an implicit one-stage Runge–Kutta (RK) method but in general this is not the case and to solve the integro-algebraic equation (2) we must combine a quadrature formula together with a non linear solver.

The above approach, based on a first degree polynomial approximation, has been generalized by Hairer in [3] to s-degree polynomials $\boldsymbol{u}(t)$ that satisfy $\boldsymbol{u}(0) = \boldsymbol{y}_0$ together with s collocation conditions at $t_j = c_j h, j = 1, \ldots, s$ where $\{c_j\}_{j=1}^s$ is a given set of non confluent nodes ($c_j \neq c_k$ for all $j \neq k$). This author has studied the order and the conjugate simplecticity of their proposed methods by using appropriate modification techniques employed in the theory of RK methods. In more detail, [3] describes the limit formulae of a particular instance of *Hamiltonian BVMs (HBVMs, see below)*, as is shown in [4].

An alternative approach to the derivation of Hamiltonian conservative methods, to the best of our knowledge the first instance of energy-preserving Runge–Kutta methods for polynomial Hamiltonian dynamical systems, was given in an early paper by Iavernaro and Pace [5] and, as later observed in [6], they can also be derived by the discretization of the *averaged vector field method* in [2]. HBVMs were formerly presented in [7] (see also [8–10] and references therein), as a generalization of the *s-Stage Trapezoidal Methods* [5] and of the energy-preserving methods for polynomial Hamiltonians derived in [11, 12]. In this class of methods, the polynomial approximate solution $\boldsymbol{u}(t)$ of (1) with degree $\leq r$ is given as the solution of the implicit equation

$$\boldsymbol{u}(0) = \boldsymbol{y}_0, \qquad \dot{\boldsymbol{u}}(\tau h) = h \sum_{j=0}^{r-1} \gamma_j P_j(\tau), \qquad \gamma_j = \int_0^1 P_j(c) f(\boldsymbol{u}(ch))\, \mathrm{d}c, \quad j = 0, \ldots, r-1,$$

where $\dot{\boldsymbol{u}}(\tau h)$ is the derivative of $\boldsymbol{u}(t)$ with respect to its argument at $t = \tau h$ and the right hand side is a truncated series expansion of the vector field $\mathbf{f}(\boldsymbol{u}(\tau h))$ along the Legendre polynomials basis $P_j(\tau)$, shifted to the interval $\tau \in [0, 1]$ and normalized to be orthonormal on that interval (see [13,14]). The authors show that such methods preserve the energy and attain order $2r$. Here again the explicit computation of $\boldsymbol{u}(t)$ leads to a set of implicit integro-algebraic equations that, as mentioned above, requires combining a quadrature formula together with a non linear solver. We also mention that HBVMs admit a straightforward Runge–Kutta formulation which, in turn, is closely related to that of Gauss collocation formulae [15].

On the other hand, *Line Integral Methods*, recently introduced in [16], represent a straight generalization of the approach which has led to define HBVMs, where the key idea of imposing energy conservation through a line integral, has been extended to any invariant and to any conservative problem.

For Poisson type systems in which $B(\boldsymbol{y})$ satisfies also the Jacobi identity or, more generally, for gradient type systems, there are several proposals to preserve energy. Thus in a recent paper of Cohen and Hairer [17], a generalization of [3] is proposed in order to preserve the energy of systems (1) as well as quadratic Casimirs. A different approach by Dahlby et al., based on Discrete Gradients [18], allows to construct methods that preserve simultaneously several invariants in systems of type (1).

In this paper we derive and analyze a class of methods which are energy-preserving and are also able to preserve quadratic Casimirs. It is found that these methods are equivalent to those recently proposed in [17], giving an alternative derivation of such methods and further a new proof of their order. This paper is also motivated by the recent research in [19], which is concerned with the error growth in the numerical solution of periodic orbits. Indeed, for problems in the form (1), methods able to preserve more invariants of the dynamical system have a more favorable error growth.

With these premises, in Section 2 we introduce the new class of energy-preserving formulae, associated to a Gaussian set of nodes and they are analyzed in Section 3. In Section 4 the above analysis is extended to an arbitrary set of nodes.

In Section 5 we then state and analyze the numerical methods obtained by the discretization of such formulae. Finally, a few numerical tests, as well as some concluding remarks, are contained in Section 6.

2. The energy-preserving formulae

We denote by $\{P_j(\tau)\}_{j\geq 0}$ the Legendre orthonormal polynomials shifted to [0, 1] that satisfy

$$\deg(P_i) = i, \quad \text{and} \quad \int_0^1 P_i(\tau)P_j(\tau)\mathrm{d}\tau = \delta_{ij}, \tag{3}$$

for all $i, j \geq 0$.

For all $h > 0$ and real continuous function $g(t)$, $t \in [0, h]$ the series expansion of $g(\tau h)$, $\tau \in [0, 1]$, in terms of Legendre polynomials is given by

$$\sum_{j\geq 0} \gamma_j(g)P_j(\tau), \quad \tau \in [0, 1], \tag{4}$$

where the coefficients $\gamma_j(g)$ are

$$\gamma_j(g) = \int_0^1 P_j(\tau)g(\tau h)\mathrm{d}\tau. \tag{5}$$

The coefficient $\gamma_j(g)$ can be interpreted as the density of the jth Legendre component in $g(\tau h)$, $\tau \in [0, 1]$ and also as the projection of $g(\tau h)$ on the jth Legendre polynomial.

It is well known that, for a fixed $h > 0$, $\gamma_j(g) \to 0$ as $j \to \infty$ and (4) converges to $g(\tau h)$, $(t = \tau h)$ in the mean norm in the sense that

$$\lim_{n\to\infty} \int_0^1 \left[g(\tau h) - \sum_{j=0}^{n} \gamma_j(g)P_j(\tau) \right]^2 \mathrm{d}\tau = 0.$$

Further, if $g(\tau h)$ has a continuous second derivative on [0, 1], then the series (4) converges uniformly to $g(\tau h)$ (see, e.g., Isaacson–Keller, pp. 206 [20]) so that we have $g(\tau h) = \sum_{j\geq 0} \gamma_j(g)P_j(\tau)$, $\tau \in [0, 1]$. Recall that in the more general case of an analytic function $g(t)$ of t (see Whittaker and Watson, pp. 322 [21]) the series (4) is called Neumann's expansion of $g(t)$ in a series of Legendre polynomials, and it is proved that if $g(\tau h)$ is analytic inside and on an ellipse C of the τ complex plane whose foci are the points $\tau = 0, 1$, the expansion (4) converges uniformly to g in any domain inside C.

In the following we will consider flows $\mathbf{f} : \mathbb{R}^m \to \mathbb{R}^m$, and vector valued functions $\boldsymbol{u} : [0, h] \to \mathbb{R}^m$ so that the series expansion of each component of $\mathbf{f}(\boldsymbol{u}(t))$, $t \in [0, h]$ converges uniformly, so that

$$\mathbf{f}(\boldsymbol{u}(\tau h)) = \sum_{j\geq 0} \gamma_j(\mathbf{f}(\boldsymbol{u}))P_j(\tau),$$

holds for all $\tau \in [0, 1]$.

In the use of vector valued polynomials we will denote by $\Pi_r^m(t)$ the set of m-vector polynomials in t with degree $\leq r$. With the above notations our proposed methods are defined as follows.

Definition 1. Let r be a positive integer, the polynomial approximation $\boldsymbol{u} = \boldsymbol{u}_h(\tau h) \in \Pi_r^m(t)$, to the solution $\boldsymbol{y}(\tau h)$, $\tau \in [0, 1]$, of (1) is defined by the $r + 1$ (vector) conditions

$$\begin{cases} \boldsymbol{u}_h(0) = \boldsymbol{y}_0, \\ \dot{\boldsymbol{u}}_h(c_i h) = B(\boldsymbol{u}_h(c_i h)) \displaystyle\sum_{j=0}^{r-1} P_j(c_i)\gamma_j(\nabla H(\boldsymbol{u}_h)), \quad i = 1, \ldots, r, \end{cases} \tag{6}$$

where c_i, $i = 1, \ldots, r$, are the Gauss nodes of $P_r(\tau)$, i.e. the roots of the Legendre polynomial $P_r(\tau)$. The solution at $t = h$ is $\boldsymbol{y}_1 = \boldsymbol{u}_h(h)$.

Note that we have used the subscript h in the polynomial $\boldsymbol{u}_h(t) = \sum_{j=0}^r \mathbf{a}_j\, t^j$ to make clear that the coefficients $\mathbf{a}_j$ may depend on h, but to simplify the notation we will omit this subscript in the sequel.

In the next section, we prove that these formulae are well defined, that is, the implicit equations that define them have a unique solution for h small enough, and they are energy-conserving, namely $H(\boldsymbol{y}_1) = H(\boldsymbol{y}_0)$. Moreover, quadratic Casimir $C(\boldsymbol{y})$ are also conserved $C(\boldsymbol{u}(h)) = C(\boldsymbol{y}_0)$.

We also prove that they have a $2r$ order of convergence, for all $r \geq 1$. Later on, in Section 5, we discuss the order of approximation of the numerical methods obtained by approximating the involved integrals by suitable quadratures.

3. Analysis of the energy-preserving formulae: Gaussian nodes

Since $\boldsymbol{u}(\tau)$ is defined by the implicit equation (6), our first concern is to give sufficient conditions to ensure the existence of a unique solution. In order to do that, let us give first some preliminary results.

By construction, the polynomial $\boldsymbol{u}$ satisfying (6) has degree $\le r$. Consequently, $\dot{\boldsymbol{u}}(\tau h)$ can be expanded in terms of the polynomial basis $\{P_j\}_{j=0}^{r-1}$ as

$$\dot{\boldsymbol{u}}(\tau h) = \sum_{j=0}^{r-1} \Gamma_j P_j(\tau), \quad \tau \in [0,1], \tag{7}$$

where the (vector) coefficients are (see (5)) $\Gamma_j = \gamma_j(\dot{\boldsymbol{u}})$. By imposing the initial condition in (6) it is clear that

$$\boldsymbol{u}(\tau h) = \boldsymbol{y}_0 + h \sum_{j=0}^{r-1} \Gamma_j \int_0^\tau P_j(s)\,\mathrm{d}s, \quad \tau \in [0,1]. \tag{8}$$

The following result then holds true.

Theorem 1. *Let b_i and c_i, $i = 1, \dots, r$, be the coefficients and the nodes of the Gaussian quadrature formula in* $[0,1]$. *The vectors*

$$\boldsymbol{\Gamma} = (\Gamma_0^T, \dots, \Gamma_{r-1}^T)^T \in \mathbb{R}^{rm}, \qquad \boldsymbol{\gamma} = (\gamma_0(\nabla H(\boldsymbol{u}))^T, \dots, \gamma_{r-1}(\nabla H(\boldsymbol{u}))^T)^T \in \mathbb{R}^{rm}$$

whose (block) entries are those appearing in (7) *and* (6)*, respectively, are related by*

$$\boldsymbol{\Gamma} = S\,\boldsymbol{\gamma}, \tag{9}$$

where

$$-S^T = S = (S_{ij}), \qquad S_{ij} = S_{ij}(\boldsymbol{u}) = \sum_{l=1}^{r} b_l P_{i-1}(c_l) P_{j-1}(c_l) B(\boldsymbol{u}(c_l h)), \quad i,j = 1, \dots, r. \tag{10}$$

Proof. Let us define the following matrices:

$$\mathcal{P} = (P_{j-1}(c_i)) \in \mathbb{R}^{r\times r}, \qquad \Omega = \mathrm{diag}(b_1, \dots, b_r). \tag{11}$$

From (6), and evaluating (7) at the Gauss nodes $\{c_i\}$, one obtains:

$$(\mathcal{P} \otimes I_m)\boldsymbol{\Gamma} = \begin{pmatrix} B(\boldsymbol{u}(c_1 h)) & & \\ & \ddots & \\ & & B(\boldsymbol{u}(c_r h)) \end{pmatrix} (\mathcal{P} \otimes I_m)\boldsymbol{\gamma}.$$

By considering that the Gaussian quadrature formula has order $2r$ (and thus it is exact for polynomials of degree $2r-1$), $\mathcal{P}^T \Omega \mathcal{P} = I_r$, and solving for $\boldsymbol{\Gamma}$ the result holds. Note that the matrix S is skew-symmetric because the matrices $B(\boldsymbol{u}(c_j h))$ are skew-symmetric. □

Now, we can state the following result.

Theorem 2. *Assume that, for a given $\rho > 0$, the functions $B = B(\boldsymbol{y})$ and $H = H(\boldsymbol{y})$ are $\mathcal{C}^1(\mathcal{V}_0)$ and $\mathcal{C}^2(\mathcal{V}_0)$, respectively, where*

$$\mathcal{V}_0 = \{\boldsymbol{y} \in \mathbb{R}^m;\ \|\boldsymbol{y} - \boldsymbol{y}_0\| \le \rho\},$$

for a given norm $\|\cdot\|$ in $\mathbb{R}^m$. Then, there exist $h_0 > 0$ such that, for all $0 < h \le h_0$, there is a unique vector valued polynomial satisfying (6).

Proof. Let

$$\beta_0 = \sup\{\|B(\boldsymbol{y})\|; \boldsymbol{y} \in \mathcal{V}_0\}, \qquad \widehat{\beta}_0 = \sup\{\|B'(\boldsymbol{y})\|; \boldsymbol{y} \in \mathcal{V}_0\},$$
$$\beta_1 = \sup\{\|\nabla H(\boldsymbol{y})\|; \boldsymbol{y} \in \mathcal{V}_0\}, \qquad \widehat{\beta}_1 = \sup\{\|\nabla H(\boldsymbol{y})'\|; \boldsymbol{y} \in \mathcal{V}_0\},$$

where in each case we have used the corresponding norm associated to the given vector norm.

Since $\mathcal{V}_0$ is a convex and compact set, β_j, $\widehat{\beta}_j$, $j = 1, 2$ are bounded and by the mean value theorem $B(\boldsymbol{y})$ and $\nabla H(\boldsymbol{y})$ satisfy the Lipschitz condition in $\mathcal{V}_0$ with constants $\widehat{\beta}_0$ and $\widehat{\beta}_1$ respectively.

After Theorem 1 we can see that the existence of a unique solution of (6) is equivalent to the existence and uniqueness of the solution of (8) with Γ_j given by (9)–(10).

Let h be a positive constant and $\Phi : \Pi_r^m(\tau) \to \Pi_r^m(\tau)$ be an operator such that for all $\boldsymbol{v} = \boldsymbol{v}(\tau h) \in \Pi_r^m(\tau)$, $\tau \in [0, 1]$, $\Phi(\boldsymbol{v})$ is defined by

$$\Phi(\boldsymbol{v})(\tau h) \equiv \boldsymbol{y}_0 + h \sum_{i=0}^{r-1} \Gamma_i \left(\int_0^\tau P_i(s)\,\mathrm{d}s \right), \quad \tau \in [0, 1], \tag{12}$$

with $\Gamma_i = \sum_{j=0}^{r-1} S_{i+1,j+1}\gamma_j$, $i = 0, \ldots, r-1$, and (see (10))

$$S_{ij} = \sum_{l=1}^{r} b_l P_{i-1}(c_l) P_{j-1}(c_l) B(\boldsymbol{v}(c_l h)), \qquad \gamma_j = \gamma_j(\nabla H(\boldsymbol{v})),$$

where b_i and c_i are the coefficients and nodes of the Gaussian quadrature formula with r nodes in [0, 1]. We now consider the m-dim vector valued space of polynomials $\Pi_r^m(\tau)$ of degree $\le r$ in the interval $\tau \in [0, 1]$ with the uniform norm

$$\|\boldsymbol{v}\|_\infty = \sup\{\|\boldsymbol{v}(\tau h)\|;\ \tau \in [0, 1]\}.$$

Since $\max_{t\in[0,1]} |P_j(t)| = \sqrt{2j+1}$, then, for all $j = 0, \ldots, r-1$:

$$\begin{aligned}\|\gamma_j(\nabla H(\boldsymbol{v}))\| &= \left\| \int_0^1 P_j(\tau)[\nabla H(\boldsymbol{v}(\tau h))]\,\mathrm{d}\tau \right\| \le \int_0^1 |P_j(\tau)|\ \|[\nabla H(\boldsymbol{v}(\tau h))]\|\,\mathrm{d}\tau \\ &\le \beta_1 \int_0^1 |P_j(\tau)|\,\mathrm{d}\tau \le \beta_1\sqrt{2j+1} \le \beta_1\sqrt{2r-1}.\end{aligned}$$

Further, for all $i, j = 0, \ldots, r-1$,

$$\sum_{l=1}^{r} b_l |P_j(c_l)P_i(c_l)| \le 2r-1$$

and, consequently, $\|S_{ij}\| \le (2r-1)\beta_0$. Therefore, for $\boldsymbol{v}(\tau h) \in \mathcal{V}_0$, $\tau \in [0, 1]$ it follows from (12) that

$$\|\Phi(\boldsymbol{v}) - \boldsymbol{y}_0\| = \left\| h \sum_{i=0}^{r-1}\sum_{j=0}^{r-1} S_{i+1,j+1}\gamma_j(\nabla H(\boldsymbol{v})) \int_0^\tau P_i(s)\,\mathrm{d}s \right\| \le h\ \beta_0\ \beta_1\ r^2(2r-1)^2.$$

Consequently, if

$$h \le \frac{\rho}{r^2(2r-1)^2\ \beta_0\beta_1}, \tag{13}$$

then $\|\Phi(\boldsymbol{v}) - \boldsymbol{y}_0\| \le \rho$ and the polynomial $\Phi(\boldsymbol{v})$ is also contained in $\mathcal{V}_0$. Moreover, the set $\Pi_r^m(\tau)$ is a Banach space with the uniform norm. Denoting by $\mathcal{M}$ the closed set $\mathcal{M} = \{\boldsymbol{v} \in \Pi_r^m(\tau)$ such that $\boldsymbol{v}(0) = \boldsymbol{y}_0$ and $\boldsymbol{v}(\tau h) \in \mathcal{V}_0\ \forall\tau \in [0, 1]\}$, under condition (13), $\Phi(\mathcal{M}) \subset \mathcal{M}$. On the other hand, for $\boldsymbol{v},\ \tilde{\boldsymbol{v}} \in \mathcal{M}$, we have

$$\|\Phi(\boldsymbol{v}) - \Phi(\tilde{\boldsymbol{v}})\| = \left\| h \sum_{i=0}^{r-1}\sum_{j=0}^{r-1} (S_{i+1,j+1}\ \gamma_j - \tilde{S}_{i+1,j+1}\ \tilde{\gamma}_j) \int_0^\tau P_j(s)\,\mathrm{d}s \right\|,$$

where

$$\tilde{S}_{i,j} = \sum_{l=1}^{r} b_l P_{i-1}(c_l) P_{j-1}(c_l) B(\tilde{\boldsymbol{v}}(c_l h)), \qquad \tilde{\gamma}_j = \gamma_j(\nabla H(\tilde{\boldsymbol{v}}(\tau h))).$$

Observe that

$$\begin{aligned}\|S_{i+1,j+1}\ \gamma_j - \tilde{S}_{i+1,j+1}\tilde{\gamma}_j\| &= \|S_{i+1,j+1}\ \gamma_j - \tilde{S}_{i+1,j+1}\ \gamma_j + \tilde{S}_{i+1,j+1}\ \gamma_j - \tilde{S}_{i+1,j+1}\tilde{\gamma}_j\| \\ &\le \|S_{i+1,j+1} - \tilde{S}_{i+1,j+1}\|\ \|\gamma_j\| + \|\tilde{S}_{i+1,j+1}\|\ \|\gamma_j - \tilde{\gamma}_j\|,\end{aligned}$$

and

$$\|\gamma_j - \tilde{\gamma}_j\| = \left\| \int_0^1 [\nabla H(\boldsymbol{v}(\tau h)) - \nabla H(\tilde{\boldsymbol{v}}(\tau h))]P_j(\tau)\mathrm{d}\tau \right\| \le \sqrt{2j+1}\ \widehat{\beta}_1\ \|\boldsymbol{v} - \tilde{\boldsymbol{v}}\|,$$

$$\|S_{i+1,j+1} - \tilde{S}_{i+1,j+1}\| \le (2r-1)\widehat{\beta}_0\ \|\boldsymbol{v} - \tilde{\boldsymbol{v}}\|.$$

Hence,

$$\begin{aligned}\|S_{i+1,j+1}\ \gamma_j - \tilde{S}_{i+1,j+1}\tilde{\gamma}_j\| &\le (2r-1)^{3/2}\ \beta_1\widehat{\beta}_0\|\boldsymbol{v} - \tilde{\boldsymbol{v}}\| + \beta_0(2r-1)^{3/2}\ \widehat{\beta}_1\ \|\boldsymbol{v} - \tilde{\boldsymbol{v}}\| \\ &= (2r-1)^{3/2}\ (\beta_1\widehat{\beta}_0 + \beta_0\widehat{\beta}_1)\|\boldsymbol{v} - \tilde{\boldsymbol{v}}\|\end{aligned}$$

and, then,

$$\|\Phi(\boldsymbol{v})-\Phi(\tilde{\boldsymbol{v}})\| \le hr^2(2r-1)^{3/2}(\beta_1\widehat{\beta}_0+\beta_0\widehat{\beta}_1)\|\boldsymbol{v}-\tilde{\boldsymbol{v}}\|,$$

which implies that Φ is a contractive mapping on $\mathcal{M}$ if

$$h<\frac{1}{r^2(2r-1)^{3/2}(\beta_1\widehat{\beta}_0+\beta_0\widehat{\beta}_1)}. \tag{14}$$

Consequently, the solution of (6) exists and is unique, for all $h>0$ such that (13)–(14) hold. □

Next, we will analyze the energy conservation.

Theorem 3. *The method proposed in Definition 1 preserves the energy of (1).*

Proof. We will show that

$$\Delta H \equiv H(\boldsymbol{u}(h))-H(\boldsymbol{u}(0))=\int_0^1 \frac{\mathrm{d}}{\mathrm{d}\tau}H(\boldsymbol{u}(\tau h))\mathrm{d}\tau = h\int_0^1 \nabla H(\boldsymbol{u}(\tau h))^T\,\dot{\boldsymbol{u}}(\tau h)\mathrm{d}\tau, \tag{15}$$

vanishes. Substituting (7) into (15), and using Theorem 1, we have

$$\Delta H = h\sum_{j=1}^r\int_0^1 \nabla H(\boldsymbol{u}(\tau h))^T P_j(\tau)\;\Gamma_j\mathrm{d}\tau = h\boldsymbol{\gamma}^T\,\boldsymbol{\Gamma} = h\boldsymbol{\gamma}^T\,S\,\boldsymbol{\gamma}=0, \tag{16}$$

where the last equality follows from the skew-symmetry of S. □

Remark 1. Observe that the essential point in the vanishing of (16) is the skew-symmetry of matrix S. By taking into account (10), it then follows that if in (6) we replace $B(\boldsymbol{u}(c_ih))$ by any skew-symmetric matrix B_i, then the method would be energy preserving as well (though, its order could be affected).

Let us consider now the preservation of Casimir functions.

Theorem 4. *The method proposed in Definition 1 preserves all Casimir functions $C(\boldsymbol{y})$ that are polynomial functions of degree ≤ 2.*

Proof. It follows the same line as the above theorem. Starting from the identity

$$C(\boldsymbol{y}_1)-C(\boldsymbol{y}_0)=C(\boldsymbol{u}(h))-C(\boldsymbol{u}(0))=\int_0^1\frac{\mathrm{d}}{\mathrm{d}\tau}C(\boldsymbol{u}(h\tau))\mathrm{d}\tau = h\int_0^1\nabla C(\boldsymbol{u}(\tau h))^T\,\dot{\boldsymbol{u}}(\tau h)\mathrm{d}\tau, \tag{17}$$

if $C(\boldsymbol{y})$ is a polynomial of degree ≤ 2, all components of $\nabla C(\boldsymbol{y})$ are polynomials of degree ≤ 1 and the integrand function in the right hand side of (17) is a polynomial of degree $\le 2r-1$ and will be integrated exactly by the Gauss formula with r nodes, i.e.,

$$C(\boldsymbol{y}_1)-C(\boldsymbol{y}_0)=h\sum_{i=1}^r b_i\;\nabla C(\boldsymbol{u}(hc_i))^T\;\dot{\boldsymbol{u}}(hc_i). \tag{18}$$

Then, by substituting $\dot{\boldsymbol{u}}(hc_i)$ from the Definition 1 each term of the right hand side of (18) vanishes. □

The next part of this section will be devoted to discuss the order of accuracy of the formulae (6). Before that, we need the following preliminary results.

Lemma 1. *Let $g:\mathbb{R}\to V$ be a suitably smooth function. Then*

$$\gamma_j(g)=\int_0^1 g(\tau h)P_j(\tau)\,\mathrm{d}\tau = O(h^j).$$

Proof. See [13, Lemma 1] or [14, Lemma 2.1]. □

We are now able to state the result concerning the accuracy. The proof strictly follows that of [14, Th. 1] which, in turn, is based on that of [22, Th. 6.5.1, pp. 165–166].

Theorem 5. $\boldsymbol{y}(h)-\boldsymbol{u}(h)=O(h^{2r+1})$.

Proof. Let us denote by $\boldsymbol{y}(t,s,\boldsymbol{v})$ the solution of problem (1) satisfying the initial condition $\boldsymbol{y}(s) = \boldsymbol{v}$. We also denote by $\Phi(t,s,\boldsymbol{v})$ the fundamental matrix function of the corresponding variational problem, satisfying the initial condition $\Phi(s,s,\boldsymbol{v}) = I$. Then (see (5)),

$$\begin{aligned}
\boldsymbol{y}(h) - \boldsymbol{u}(h) &= \boldsymbol{y}(h,0,\boldsymbol{y}_0) - \boldsymbol{y}(h,h,\boldsymbol{u}(h)) = -\int_0^h \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{y}(h,t,\boldsymbol{u}(t))\,\mathrm{d}t \\
&= -\int_0^h \left[\frac{\partial}{\partial t}\boldsymbol{y}(h,t,\boldsymbol{u}(t)) + \frac{\partial}{\partial \boldsymbol{u}}\boldsymbol{y}(h,t,\boldsymbol{u}(t))\,\dot{\boldsymbol{u}}(t)\right]\mathrm{d}t \\
&= h\int_0^1 \Phi(h,\tau h,\boldsymbol{u}(\tau h))[B(\boldsymbol{u}(\tau h))\nabla H(\boldsymbol{u}(\tau h)) - \dot{\boldsymbol{u}}(\tau h)]\,\mathrm{d}\tau \\
&= h\int_0^1 \Phi(h,\tau h,\boldsymbol{u}(\tau h))\left[B(\boldsymbol{u}(\tau h))\nabla H(\boldsymbol{u}(\tau h)) - B(\boldsymbol{u}(\tau h))\sum_{j=0}^{r-1}\gamma_j(\nabla H(\boldsymbol{u}))\,P_j(\tau)\right]\mathrm{d}\tau \\
&\quad + h\int_0^1 \Phi(h,\tau h,\boldsymbol{u}(\tau h))\left[B(\boldsymbol{u}(\tau h))\sum_{j=0}^{r-1}\gamma_j(\nabla H(\boldsymbol{u}))P_j(\tau) - \dot{\boldsymbol{u}}(\tau h)\right]\mathrm{d}\tau \\
&\equiv h(E_1(h) + E_2(h)).
\end{aligned}$$

Concerning $E_1(h)$, by setting $G(\tau h) = \Phi(h,\tau h,\boldsymbol{u}(\tau h))$, and considering that

$$\nabla H(\boldsymbol{u}(\tau h)) = \sum_{j\ge 0}\gamma_j(\nabla H(\boldsymbol{u}))P_j(\tau),$$

one obtains, by virtue of Lemma 1:

$$E_1(h) = \sum_{j\ge r}\left[\int_0^1 G(\tau h)B(\boldsymbol{u}(\tau h))P_j(\tau)\,\mathrm{d}\tau\right]\gamma_j(\nabla H(\boldsymbol{u})) = \sum_{j\ge r} O(h^j)\,O(h^j) = O(h^{2r}).$$

Concerning $E_2(h)$, one obtains:

$$\begin{aligned}
E_2(h) &= \sum_{i=1}^r b_i G(c_i h)\left[B(\boldsymbol{u}(c_i h))\sum_{j=0}^{r-1}\gamma_j(\nabla H(\boldsymbol{u}))\,P_j(c_i) - \dot{\boldsymbol{u}}(c_i h)\right] + O(h^{2r}) \\
&= O(h^{2r}),
\end{aligned}$$

since the Gaussian quadrature formula has order $2r$. Consequently, the thesis follows. □

4. Analysis of the energy-preserving formulae: Non Gaussian nodes

In the above analysis the Definition 1 of $\boldsymbol{u}(\tau h)$ is given with respect to the Gaussian nodes in $[0, 1]$. In this section, the analysis will be extended to any other quadrature formula. Let (b_i, c_i), $i = 1,\ldots,r$ be a quadrature formula of order $q \ge r$ and we will denote $\mathcal{P}$ and Ω the corresponding matrices defined by (11). Observe that for the Gaussian formula $\mathcal{P}^T\Omega\mathcal{P} = I_r$, but now, this is no longer true. In this case, we have.

Lemma 2. *If the quadrature formula has order $q \ge r$ (degree of precision $q-1$) then*

$$\mathcal{P}^T\Omega\mathcal{P} = I_r + M, \qquad M = (m_{ij}) = M^T, \quad m_{ij} = 0 \quad \textit{for } i+j \le q+1. \tag{19}$$

Proof. Matrix M is clearly symmetric. The result then follows taking into account that the (i,j)-th component of the matrix $\mathcal{P}^T\Omega\mathcal{P}$ is

$$\sum_{l=1}^r b_l P_{i-1}(c_l)P_{j-1}(c_l) = \int_0^1 P_{i-1}(\tau)P_{j-1}(\tau)\,\mathrm{d}\tau = \delta_{ij},$$

whenever the degree of the integrand, $i+j-2$, is not larger than the degree of precision, $q-1$, of the quadrature. That is, when $i+j \le q+1$. □

If we define the methods in the same way as in the case of the Gauss quadrature formulae, the Hamiltonian is not preserved in general. Then, to ensure the preservation of the energy, instead of (6), we will define the method by.

Definition 2. Let r be a positive integer and (b_i, c_i), $i = 1, \dots, r$ the coefficients and nodes of a quadrature formula of order $q \geq r$. The polynomial approximation $\boldsymbol{u} = \boldsymbol{u}_h(\tau h) \in \Pi_r^m(t)$, to the solution $\boldsymbol{y}(\tau h)$, $\tau \in [0, 1]$, of (1) is defined by the $r + 1$ (vector) conditions

$$\begin{cases} \boldsymbol{u}_h(0) = \boldsymbol{y}_0, \\ \dot{\boldsymbol{u}}(c_i h) = B(\boldsymbol{u}(c_i h)) \displaystyle\sum_{j=0}^{r-1} q_{i,j+1}\ \gamma_j(\nabla H(\boldsymbol{u})), \quad i = 1, \dots, r, \end{cases} \tag{20}$$

where $q_{i,j+1}$ are the coefficients of the matrix

$$\mathcal{Q} = (q_{ij}) = \mathcal{P}(I_r + M)^{-1} \in \mathbb{R}^{r\times r}. \tag{21}$$

The solution at $t = h$ is $\boldsymbol{y}_1 = \boldsymbol{u}_h(h)$.

From (19), we have

$$(I_r + M)^{-1} = \begin{pmatrix} I_s & 0 \\ 0 & \hat{M} \end{pmatrix} = I_r + \begin{pmatrix} 0 & 0 \\ 0 & \tilde{M} \end{pmatrix}, \quad \tilde{M} \in \mathbb{R}^{r-s\times r-s}, \tag{22}$$

with

$$s = q - r + 1, \tag{23}$$

and, therefore, the first s columns of $\mathcal{Q}$ (see (21)) are the same as those of $\mathcal{P}$. Hence (20) can be written as

$$\dot{\boldsymbol{u}}(c_i h) = B(\boldsymbol{u}(c_i h)) \left[\sum_{j=0}^{s-1} P_j(c_i)\gamma_j(\nabla H(\boldsymbol{u})) + \sum_{j=s}^{r-1} q_{i,j+1}\ \gamma_j(\nabla H(\boldsymbol{u})) \right], \quad i = 1, \dots, r.$$

A comparison with (6) shows that to preserve the energy for non Gaussian nodes, we have perturbed the last $r - s$ terms of the expansion of $\nabla H(\boldsymbol{u})$ in terms of the orthogonal polynomials $\{P_j\}$. This may imply, as it is seen below, a reduction of the order with respect to the order of the quadrature formula.

In a similar way to the Gaussian nodes, it can be proved that the Eqs. (20) have a unique solution.

Expanding again $\dot{\boldsymbol{u}}(\tau h) = \sum_{j=0}^{r-1} \Gamma_j P_j(\tau)$ in terms of the polynomial basis $\{P_j\}$, we have, in place of (9)–(10), the following result.

Theorem 6. *The coefficients Γ_j satisfy*

$$\boldsymbol{\Gamma} = S\ \boldsymbol{\gamma},$$

with

$$-S^T = S = (S_{ij}), \qquad S_{ij} = \sum_{l=1}^{r} b_l\ q_{li}\ q_{lj}\ B(\boldsymbol{u}(c_l h)), \quad i, j = 1, \dots, r. \tag{24}$$

Proof. Using the matrices and vectors defined in (11) and substituting (20) into (7), one obtains:

$$(\mathcal{P} \otimes I_m)\boldsymbol{\Gamma} = \begin{pmatrix} B(\boldsymbol{u}(c_1 h)) & & \\ & \ddots & \\ & & B(\boldsymbol{u}(c_r h)) \end{pmatrix} (\mathcal{Q} \otimes I_m)\boldsymbol{\gamma}$$

and taking into account that (see (19)–(21)) $\mathcal{P}^{-1} = \mathcal{Q}^T\Omega$,

$$\boldsymbol{\Gamma} = (\mathcal{Q}^T \otimes I_m)(\Omega \otimes I_m) \begin{pmatrix} B(\boldsymbol{u}(c_1 h)) & & \\ & \ddots & \\ & & B(\boldsymbol{u}(c_r h)) \end{pmatrix} (\mathcal{Q} \otimes I_m)\boldsymbol{\gamma},$$

so that (24) follows. At last, the matrix S is skew-symmetric because the matrices $B(\boldsymbol{u}(c_i h))$ are skew-symmetric. □

With this result, Theorem 3 can be applied and we conclude that the formulae (20) preserve the energy. Conversely, quadratic Casimirs turn out to be preserved only when $q = 2r$, whereas, in the general case, only linear Casimirs are preserved.

To discuss the order of accuracy of the formula (20), we need a couple of preliminary results.

Lemma 3. *If the quadrature formula (b_i, c_i) has order $q \geq r$ then*

$$\sum_{j=0}^{r-1}[P_j(c_i) - q_{i,j+1}]\gamma_j(\nabla H(\boldsymbol{u})) = O(h^s), \quad i = 1, \ldots, r$$

with s given by (23).

Proof. The above expression can be written in compact form as

$$\sum_{j=0}^{r-1}[P_j(c_i) - q_{i,j+1}]\gamma_j(\nabla H(\boldsymbol{u})) = ([\mathcal{P} - \mathcal{Q}] \otimes I_m)\boldsymbol{\gamma}. \tag{25}$$

From (21)–(23) we have

$$\mathcal{P} - \mathcal{Q} = \mathcal{P}\begin{pmatrix} 0 & 0 \\ 0 & \tilde{M} \end{pmatrix}.$$

This means that the expression (25) only depends on $\boldsymbol{\gamma}_j(\nabla H(\boldsymbol{u}))$, $j = s, \ldots, r-1$ and, since $\boldsymbol{\gamma}_j(\nabla H(\boldsymbol{u})) = O(h^j)$, the thesis follows. □

Note that if $q = 2r$ or $q = 2r - 1$, then $\tilde{M} = 0$ and, therefore, (25) vanishes. Consequently, no perturbation must be introduced in the method.

Lemma 4. *If the quadrature formula (b_i, c_i) has order $q \geq r$ (that is, degree of precision $q - 1$), then*

$$\sum_{i=1}^{r} b_i c_i^l [P_j(c_i) - q_{i,j+1}] = 0, \quad j = 0, \ldots, r-1, \ l = 0, \ldots, s-1,$$

with s given by (23).

Proof. Denoting $b = (b_1, \ldots, b_r)^T$ and $c = \mathrm{diag}(c_1, \ldots, c_r)$, the above expression can be written in compact form as

$$\sum_{i=1}^{r} b_i c_i^l [P_j(c_i) - q_{i,j+1}] = (b^T c^l [\mathcal{P} - \mathcal{Q}])_{j+1} = \left(b^T c^l \mathcal{P}\begin{pmatrix} 0 & 0 \\ 0 & \tilde{M} \end{pmatrix}\right)_{j+1}, \quad j = 0, \ldots, r-1, \tag{26}$$

so that the first s entries ($j = 0, \ldots, s-1$) vanish for all $l \geq 0$. On the other hand, for $j \geq s$, by the order q of the quadrature formula and the orthogonality of the polynomials, it is clear that

$$\sum_{i=1}^{r} b_i c_i^l P_j(c_i) = 0, \quad \text{whenever} \quad l < j \quad \text{and} \quad l + j \leq q - 1,$$

so that (see (23)) $l \leq s - 1$. This means that the row vector $b^T c^l \mathcal{P}$ has its last $r - s$ components zero for all $j = 0, \ldots, r-1$ and $l = 0, \ldots, s-1$. Consequently, the thesis follows. □

Theorem 7. *If the quadrature formula (b_i, c_i) has order $q \geq r$ then $\boldsymbol{y}(h) - \boldsymbol{u}(h) = O(h^{p+1})$ with $p = \min\{q, 2(q-r)+2\}$.*

Proof. Proceeding as in the proof of Theorem 5, we arrive at

$$\boldsymbol{y}(h) - \boldsymbol{u}(h) = h(E_1(h) + E_2(h))$$

with $E_1(h) = O(h^{2r})$, and

$$\begin{aligned} E_2(h) &= \sum_{i=1}^{r} b_i G(c_i h)\left[B(\boldsymbol{u}(c_i h))\sum_{j=0}^{r-1}\gamma_j(\nabla H(\boldsymbol{u}))\, P_j(c_i) - \dot{\boldsymbol{u}}(c_i h)\right] + O(h^q) \\ &= \sum_{i=1}^{r} b_i G(c_i h) B(\boldsymbol{u}(c_i h))\sum_{j=0}^{r-1}[P_j(c_i) - q_{i,j+1}]\ \gamma_j(\nabla H(\boldsymbol{u})) + O(h^q) \end{aligned}$$

since the quadrature formula with weights $\{b_i\}$ has order q. Let us assume that the (matrix) function $G(c_i h)B(\boldsymbol{u}(c_i h))$ is smooth enough so that it can be expanded as

$$G(c_i h)B(\boldsymbol{u}(c_i h)) = T_0 + (c_i h)T_1 + \cdots + (c_i h)^s T_s + \cdots.$$

Then, by virtue of Lemmas 4 and 3, one obtains, respectively,

$$\begin{aligned} E_2(h) &= \sum_{l\ge s} T_l h^l \sum_{j=s}^{r-1} \left(\sum_{i=1}^{r} b_i c_i^l [P_j(c_i) - q_{i,j+1}] \right) \gamma_j(\nabla H(\boldsymbol{u})) + O(h^q) \\ &= O(h^{2s}) + O(h^q). \end{aligned}$$

Consequently, from (23) the thesis easily follows. □

In the remaining part of this section, we show that the formulae (6) are actually those defined in [17], so that we have here provided an alternative proof of their properties, which relies on the use of the orthonormal basis (3), in place of the Lagrange basis $\{\lambda_i(\tau)\}$ defined at the abscissae $\{c_i\}$. Before that, we need the following preliminary result.

Lemma 5. *Let $\lambda_i(\tau)$, $i = 1, \dots, r$, be the Lagrange polynomials defined at the Gauss abscissae $c_1, \dots, c_r$. Then*

$$\sum_{j=0}^{r-1} P_j(c_i) P_j(\tau) = \frac{\lambda_i(\tau)}{b_i}, \quad i = 1, \dots, r. \tag{27}$$

Proof. By expanding the Lagrange polynomial $\lambda_i(\tau)$ along the orthonormal basis (3), one obtains:

$$\lambda_i(\tau) = \sum_{j=0}^{r-1} P_j(\tau) \int_0^1 P_j(c)\lambda_i(c)\,\mathrm{d}c = \sum_{j=0}^{r-1} P_j(\tau) \sum_{\ell=1}^{r} b_\ell P_j(c_\ell)\lambda_i(c_\ell) = \sum_{j=0}^{r-1} P_j(\tau) b_i P_j(c_i),$$

since $\lambda_i(c_\ell) = \delta_{i\ell}$, the Kronecker delta, so that (27) follows. □

Consequently, one obtains that formula (6) can be rewritten as

$$\dot{\boldsymbol{u}}(c_i h) = B(\boldsymbol{u}(c_i h)) \int_0^1 \frac{\lambda_i(\tau)}{b_i} \nabla H(\boldsymbol{u}(\tau h))\,\mathrm{d}\tau, \quad i = 1, \dots, r,$$

i.e., the formulae defined in [17]. However, the choice of the orthonormal basis (3), in addition to providing an alternative way to study the properties of the methods, will in turn make more intuitive (in our opinion) the analysis and the implementation of the numerical methods obtained by discretizing (6). □

Remark 2. It has been proved by Quispel and Capel [23] that an autonomous differential system $\dot{y} = f(y)$, $y \in \mathbf{R}^m$ has a first integral $I(y)$ iff it can be formally written as a skew gradient system

$$\frac{\mathrm{d}}{\mathrm{d}t} y = B(y)\nabla I(y), \quad \text{where } B(y)^T = -B(y). \tag{28}$$

Moreover for an arbitrary vector field $g(y)$ (not orthogonal to $\nabla I(y)$) the components of $B(y) = (B_{ij}(y))$ are given by

$$B_{ij}(y) = \frac{f_i(y)g_j(y) - f_j(y)g_i(y)}{g(y) \cdot \nabla I(y)}. \tag{29}$$

In fact, by taking into account that $I(y)$ is a first integral iff

$$\frac{\mathrm{d}}{\mathrm{d}t} I(y) = \sum_{j=1}^{m} \frac{\partial I}{\partial y_j}(y) f_j(y) = 0,$$

we have

$$\sum_{j=1}^{m} B_{ij}(y) \frac{\partial I}{\partial y_j}(y) = f_i(y),$$

and, then, (28)–(29) is equivalent to the original differential system. This implies that an arbitrary differential system with a first integral can be written in a Poisson type form and, then, the above methods can be used preserving the first integral.

5. Discretization

Clearly, formulae (6) do not yet provide numerical methods, since the integrals appearing in them cannot, in general, be directly computed. Numerical methods will be obtained once such integrals are conveniently approximated by means of a quadrature formula. This aspect will be studied in the sequel.

Let us suppose to approximate the integrals defining $\gamma_j(\nabla H(\boldsymbol{u}))$ by means of a quadrature with $k \geq r$ nodes $\{\hat{c}_\ell\}$ and weights $\{\hat{b}_\ell\}$. In such a case, the solution will not be $\boldsymbol{u}(t)$ anymore but another polynomial of the same degree, let us denote it by $\boldsymbol{\omega}(t)$, satisfying

$$\dot{\boldsymbol{\omega}}(c_i h) = B(\boldsymbol{\omega}(c_i h)) \sum_{j=0}^{r-1} P_j(c_i)\ \hat{\gamma}_j(\nabla H(\boldsymbol{\omega})), \quad i = 1, \ldots, r, \tag{30}$$

where the coefficients $\hat{\gamma}_j(\nabla H(\boldsymbol{\omega}))$ are computed by means of the above mentioned quadrature formula. Clearly, the best choice for the nodes is that of placing them at the Gauss points, thus providing a quadrature of order $2k$ (i.e., exact for polynomials of degree $2k-1$), which we assume hereafter:

$$\hat{\gamma}_j(\nabla H(\boldsymbol{\omega})) \equiv \sum_{\ell=1}^{k} \hat{b}_\ell P_j(\hat{c}_\ell) \nabla H(\boldsymbol{\omega}(\hat{c}_\ell h)) = \gamma_j(\nabla H(\boldsymbol{\omega})) - \Delta_j(h), \quad j = 0, \ldots, r-1 \tag{31}$$

Since the quadrature has order $2k$, then

$$\Delta_j(h) = O(h^{2k-j}), \quad j = 0, \ldots, r-1. \tag{32}$$

Remark 3. In the particular case where H is a polynomial of degree ν, by considering that $\boldsymbol{\omega}$ is itself a polynomial of degree r, one obtains that $\gamma_j = \hat{\gamma}_j, j = 0, \ldots, r-1$, provided that

$$\nu \leq \frac{2k}{r}. \tag{33}$$

As already observed in [10,7,4], this provides a *practical conservation* for all suitably regular Hamiltonians, by choosing k large enough, since it suffices that the integrals are approximated within machine precision.

By following similar steps as those in Section 3, let us set

$$\begin{aligned}
\dot{\boldsymbol{\omega}}(ch) &= \sum_{i=0}^{r-1} P_i(c)\ \hat{\Gamma}_i, \\
\hat{\Gamma}_i &= \sum_{j=0}^{r-1} S_{i+1,j+1}\ \hat{\gamma}_j(\nabla H(\boldsymbol{\omega})), \quad i = 0, \ldots, r-1, \\
S_{ij} &= \sum_{l=1}^{r} b_l\ P_{i-1}(c_l) P_{j-1}(c_l) B(\boldsymbol{\omega}(c_l h)), \quad i, j = 1, \ldots, r.
\end{aligned} \tag{34}$$

Moreover, we define the following matrices and vectors, besides those defined in (11):

$$\begin{aligned}
&\hat{\mathcal{P}} = (P_{j-1}(\hat{c}_i)), \qquad \hat{\mathcal{I}} = \left(\int_0^{\hat{c}_i} P_{j-1}(x)\,\mathrm{d}x\right) \in \mathbb{R}^{k\times r}, \qquad \hat{\Omega} = \mathrm{diag}(\hat{b}_1, \ldots, \hat{b}_k), \\
&\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_0(\nabla H(\boldsymbol{\omega}))^T, \ldots, \hat{\gamma}_{r-1}(\nabla H(\boldsymbol{\omega}))^T)^T, \qquad \boldsymbol{W} = (\boldsymbol{\omega}(c_1 h)^T, \ldots, \boldsymbol{\omega}(c_r h)^T)^T, \\
&B(\boldsymbol{W}) = \mathrm{diag}(B(\boldsymbol{\omega}(c_1 h)), \ldots, B(\boldsymbol{\omega}(c_r h))), \\
&e = (1, \ldots, 1)^T \in \mathbb{R}^r, \qquad \hat{e} = (1, \ldots, 1)^T \in \mathbb{R}^k, \qquad \mathcal{I} = \left(\int_0^{c_i} P_{j-1}(x)\,\mathrm{d}x\right) \in \mathbb{R}^{r\times r}.
\end{aligned}$$

Then, the discrete problem (30)–(31) can be written as

$$\begin{aligned}
\hat{\boldsymbol{\gamma}} &= (\hat{\mathcal{P}}^T \hat{\Omega} \otimes I)\ \nabla H(\hat{e} \otimes y_0 + h(\hat{\mathcal{I}} \mathcal{P}^T \Omega \otimes I)\, B(\boldsymbol{W})\, (\mathcal{P} \otimes I)\, \hat{\boldsymbol{\gamma}}), \\
\boldsymbol{W} &= \hat{e} \otimes y_0 + h(\hat{\mathcal{I}} \mathcal{P}^T \Omega \otimes I)\, B(\boldsymbol{W})\, (\mathcal{P} \otimes I)\, \hat{\boldsymbol{\gamma}}.
\end{aligned} \tag{35}$$

We observe that the discrete problem (35) can be more efficiently cast, by introducing the block vector

$$\hat{\boldsymbol{\Gamma}} = (\hat{\Gamma}_0^T, \ldots, \hat{\Gamma}_{r-1}^T)^T \equiv ((\mathcal{P}^T \Omega \otimes I)\, B(\boldsymbol{W})\, (\mathcal{P} \otimes I))\, \hat{\boldsymbol{\gamma}},$$

as

$$\hat{\boldsymbol{\Gamma}} = (\mathcal{P}^T \Omega \otimes I)\, B(e \otimes y_0 + h(\mathcal{I} \otimes I)\, \hat{\boldsymbol{\Gamma}})\, (\mathcal{P} \hat{\mathcal{P}}^T \hat{\Omega} \otimes I)\, \nabla H(\hat{e} \otimes y_0 + h(\hat{\mathcal{I}} \otimes I)\, \hat{\boldsymbol{\Gamma}}), \tag{36}$$

which has block size r, *independently of* k. Once (36) has been solved, the new approximation (see (8) and (34)) is given by

$$\boldsymbol{y}_1 = \boldsymbol{y}_0 + \sum_{j=0}^{r-1} \hat{\Gamma}_j \int_0^1 P_j(\sigma)\,\mathrm{d}\sigma = \boldsymbol{y}_0 + \hat{\Gamma}_0. \tag{37}$$

Remark 4. In the case where in (1) $B(\boldsymbol{y}) \equiv B$, a constant skew-symmetric matrix (e.g., in the case of Hamiltonian problems), by considering that $\mathcal{P}^T\Omega\mathcal{P} = I_r$, then (36) reduces to the discrete problem generated by a HBVM(k, r) method [24]:

$$\hat{\boldsymbol{\Gamma}} = (\hat{\mathcal{P}}^T\hat{\Omega}\otimes B)\,\nabla H(\hat{e}\otimes y_0 + h(\hat{\mathcal{I}}\otimes I)\,\hat{\boldsymbol{\Gamma}}).$$

Remark 5. As already observed above, it is worth noting that the (block) dimension of the discrete problem (36) is r, independently of k. This means that we can always choose k large enough, for getting an accurate enough quadrature, without increasing the computational cost of the method too much.

Clearly, the order of the resulting method will depend on the order $q \equiv 2k$ of the quadrature formula. The following theorem, which follows the proof of [14, Th. 3.1], addresses this issue.

Theorem 8. *For all $k \ge r$, one has $\boldsymbol{y}(h) - \boldsymbol{\omega}(h) = O(h^{2r+1})$.*

Proof. The steps are quite similar to those followed in the proof of Theorem 5. By using a notation similar to that used in that theorem, one has:

$$\begin{aligned}
\boldsymbol{y}(h) - \boldsymbol{\omega}(h) &= \boldsymbol{y}(h, 0, \boldsymbol{y}_0) - \boldsymbol{y}(h, h, \boldsymbol{\omega}(h)) = -\int_0^h \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{y}(h, t, \boldsymbol{\omega}(t))\,\mathrm{d}t \\
&= -\int_0^h \left[\frac{\partial}{\partial t}\boldsymbol{y}(h, t, \boldsymbol{\omega}(t)) + \frac{\partial}{\partial \boldsymbol{\omega}}\boldsymbol{y}(h, t, \boldsymbol{\omega}(t))\,\dot{\boldsymbol{\omega}}(t)\right]\mathrm{d}t \\
&= h\int_0^1 \Phi(h, \tau h, \boldsymbol{\omega}(\tau h))[B(\boldsymbol{\omega}(\tau h))\nabla H(\boldsymbol{\omega}(\tau h)) - \dot{\boldsymbol{\omega}}(\tau h)]\,\mathrm{d}\tau \\
&= h\int_0^1 \Phi(h, \tau h, \boldsymbol{\omega}(\tau h))\left[B(\boldsymbol{\omega}(\tau h))\nabla H(\boldsymbol{\omega}(\tau h)) - B(\boldsymbol{\omega}(\tau h))\sum_{j=0}^{r-1}\gamma_j(\nabla H(\boldsymbol{\omega}))P_j(\tau)\right]\mathrm{d}\tau \\
&\quad + h\int_0^1 \Phi(h, \tau h, \boldsymbol{\omega}(\tau h))\left[B(\boldsymbol{\omega}(\tau h))\sum_{j=0}^{r-1}\gamma_j(\nabla H(\boldsymbol{\omega}))\,P_j(\tau) - \dot{\boldsymbol{\omega}}(\tau h)\right]\mathrm{d}\tau \\
&\equiv h(E_1(h) + E_2(h)).
\end{aligned}$$

Concerning $E_1(h)$, by using the same arguments as in the proof of Theorem 5, we arrive to state that $E_1(h) = O(h^{2r})$. Concerning $E_2(h)$, by setting $G(\tau h) = \Phi(h, \tau h, \boldsymbol{\omega}(\tau h))$ one obtains, by virtue of (31)–(32),

$$\begin{aligned}
E_2(h) &= \sum_{i=1}^r b_i G(c_i h)B(\boldsymbol{\omega}(c_i h))\left[\sum_{j=0}^{r-1}\gamma_j(\nabla H(\boldsymbol{\omega}))\,P_j(c_i) - \sum_{j=0}^{r-1}\hat{\gamma}_j(\nabla H(\boldsymbol{\omega}))\,P_j(c_i)\right] + O(h^{2r}) \\
&= \sum_{i=1}^r b_i G(c_i h)B(\boldsymbol{\omega}(c_i h))\sum_{j=0}^{r-1}\Delta_j(h)\,P_j(c_i) + O(h^{2r})
\end{aligned}$$

since the Gaussian quadrature formula has order $2r$. Let us assume that the (matrix) function $G(c_i h)B(\boldsymbol{\omega}(c_i h))$ is smooth enough so that it can be expanded as

$$G(c_i h)B(\boldsymbol{\omega}(c_i h)) = T_0 + (c_i h)T_1 + \cdots + (c_i h)^s T_s + \cdots.$$

Then, considering that

$$\sum_{i=1}^r b_i c_i^l P_j(c_i) = 0, \quad \text{for } l = 0, \ldots, j-1,$$

one obtains:

$$\begin{aligned}
E_2(h) &= \sum_{l\ge 0} T_l h^l \sum_{i=1}^r b_i c_i^l \sum_{j=0}^{r-1} P_j(c_i)\,\Delta_j(h) + O(h^{2r}) \\
&= \sum_{j=0}^{r-1}\sum_{l\ge j} T_l h^l \left(\sum_{i=1}^r b_i c_i^l P_j(c_i)\right)\Delta_j(h) + O(h^{2r}) \\
&= \sum_{j=0}^{r-1} O(h^j)\,\Delta_j(h) + O(h^{2r}) \ = O(h^{2k}) + O(h^{2r}).
\end{aligned}$$

Consequently, the thesis follows. □

Table 1
Problem (38), error after one period for the 2-stages Gauss method (e_G) and for the (12, 2) conservative variant (e_{CV}), when using a stepsize $h = T/n$, along with the estimated order of convergence.

n	e_{CV}	p	e_G	p
20	1.287e−02	–	6.556e−01	–
40	2.124e−03	2.60	4.509e−02	3.86
60	4.589e−04	3.78	1.331e−02	3.01
80	1.510e−04	3.86	4.298e−03	3.93
100	6.300e−05	3.92	1.796e−03	3.91
120	3.068e−05	3.95	8.751e−04	3.94

The next result concerns the order of approximation of the Hamiltonian, in case it is not a polynomial and/or (33) is not satisfied, thus extending to the discrete setting the result of Theorem 3. The proof strictly relies on the concept of *discrete line integral*, as defined in [9,8,10,7,4,5,11,12].

Theorem 9. *For all $k \ge r$, $H(\boldsymbol{\omega}(h)) - H(\boldsymbol{y}_0) = O(h^{2k+1})$. In addition, if $H(\boldsymbol{y})$ is a polynomial of degree ν satisfying (33), then $H(\boldsymbol{\omega}(h)) = H(\boldsymbol{y}_0)$.*

Proof. From (31) and (34), one has:

$$\begin{aligned} H(\boldsymbol{\omega}(h)) - H(\boldsymbol{y}_0) &= \int_0^h \nabla H(\boldsymbol{\omega}(t))^T\, \dot{\boldsymbol{\omega}}(t)\,\mathrm{d}t \;=\; h\int_0^1 \nabla H(\boldsymbol{\omega}(\tau h))^T\, \dot{\boldsymbol{\omega}}(\tau h)\,\mathrm{d}\tau \\ &= h\int_0^1 \nabla H(\boldsymbol{\omega}(\tau h))^T \sum_{j=0}^{r-1} \hat{\Gamma}_j P_j(\tau)\,\mathrm{d}\tau = h\sum_{j=0}^{r-1} \hat{\Gamma}_j^T\, \gamma_j(\nabla H(\boldsymbol{\omega})) \\ &= -h\sum_{j=0}^{r-1} \hat{\Gamma}_j^T \Delta_j(h). \end{aligned}$$

If H is a polynomial of degree ν satisfying (33), then $\Delta_j(h) = 0, j = 0, \dots, r-1$. Conversely, by using (34), and expanding $B(\boldsymbol{\omega}(c_l h))$ as a series power of $(c_l h)$, it is easy to see that $\hat{\Gamma}_j = \mathcal{O}(h^j)$ and, therefore, $H(\boldsymbol{\omega}(h)) - H(\boldsymbol{y}_0) = O(h^{2k+1})$. □

On the other hand, it can be readily proved that the result of Theorem 4 continues to be valid after the discretization. Moreover, for a general Casimir $C(\boldsymbol{y})$, one obtains $C(\boldsymbol{\omega}(h)) = C(\boldsymbol{y}_0) + O(h^{2r+1})$, for all $k \ge r$.

6. Numerical tests

We consider the Poisson problem defined as follows:

$$B(\boldsymbol{y}) = \begin{pmatrix} 0 & c_3 y_3 & -c_2 y_2 \\ -c_3 y_3 & 0 & c_1 y_1 \\ c_2 y_2 & -c_1 y_1 & 0 \end{pmatrix}, \qquad H(\boldsymbol{y}) = y_1^{12} + \frac{1}{2}[(y_2 - y_3)^2 + (y_1 - y_3)^2], \tag{38}$$

with $c_1 = 1$, $c_2 = 5$, $c_3 = -4$. The solution started at $(1, 1, 1)^T$ turns out to be periodic with period $T \approx 0.53102669598427$. The problem admits also the quadratic Casimir $C(\boldsymbol{y}) = (c_1 y_1^2 + c_2 y_2^2 + c_3 y_3^2)/2$.

We use the methods derived by formula (6) with $r = 2$, by considering $k = 12$ Gauss points for the numerical quadrature approximating the integrals. Consequently, according to (33), the quadrature turns out to be exact for this polynomial Hamiltonian of degree $\nu = 12$, so that the method is energy conserving. Moreover, also the quadratic Casimir $C(\boldsymbol{y})$ turns out to be conserved. In the practice, we obtain a method closely related to the HBVM (12, 2) method (for Hamiltonian problems, they are, indeed the *same* method, as observed in Remark 4). In Table 1, we list the errors after one period by using the 2-stages Gauss method, and the (12, 2) conservative variant (i.e., the method defined by (30)–(31) with $k = 12$ and $r = 2$). As one can see, the order four of the methods is numerically confirmed. Moreover, the error for the conserving method is smaller than that for the Gauss method, which turns out to preserve only the Casimir. This fact is further confirmed by considering a longer time interval, where the conserving method exhibits a linear error growth, whereas the Gauss method has a quadratic error growth, along with a drift in the Hamiltonian, as is shown in Figs. 1 and 2, respectively. We observe that the linear growth of the error with the (12, 2) conserving method (which, actually, preserves two invariants of the problem), confirms the analysis in [19]. In addition to this, in Fig. 3 we plot the error growth by using the (12, 2) method and the (4, 2) method. According to Theorem 9, the latter method preserves the Hamiltonian only up to order 8, which is doubled with respect to that of the method (the Casimir is, instead, exactly preserved). In such a case, even though a drift in the Hamiltonian still occurs (though, much smaller that that observed for the fourth-order Gauss method), according to the analysis in [19] it turns out that a linear growth of the error is to be expected, provided that Nh^4 (N being the number

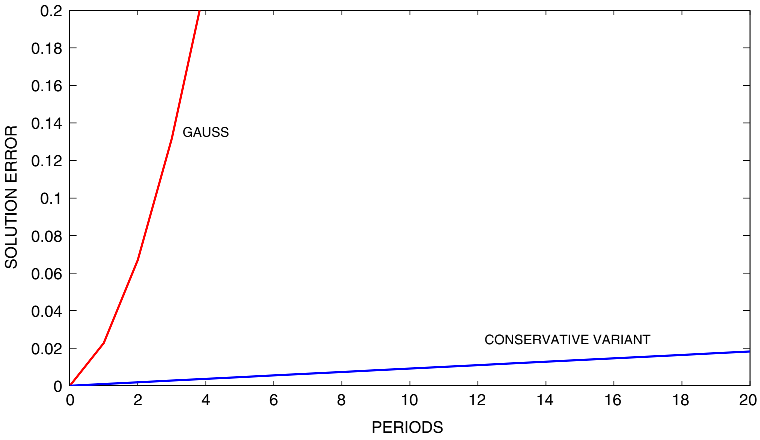

Fig. 1. Problem (38), error in the computed solution by using the 2-stage Gauss method and the $(12, 2)$ conservative variant with a constant stepsize $h = T/50$.

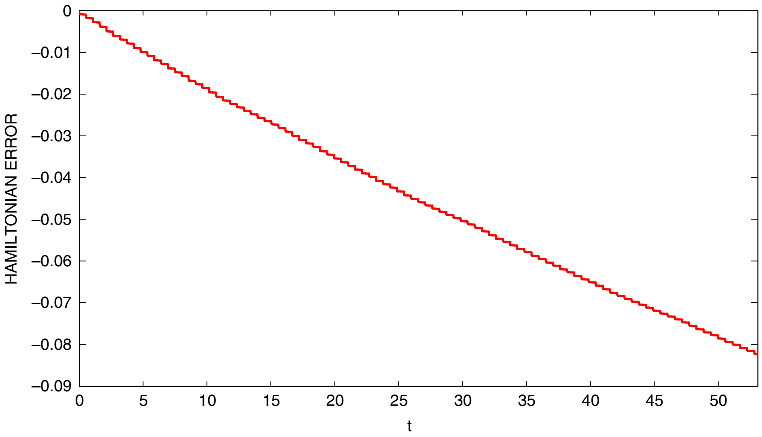

Fig. 2. Problem (38), drift in the Hamiltonian by using the Gauss method with stepsize $h = T/50$.

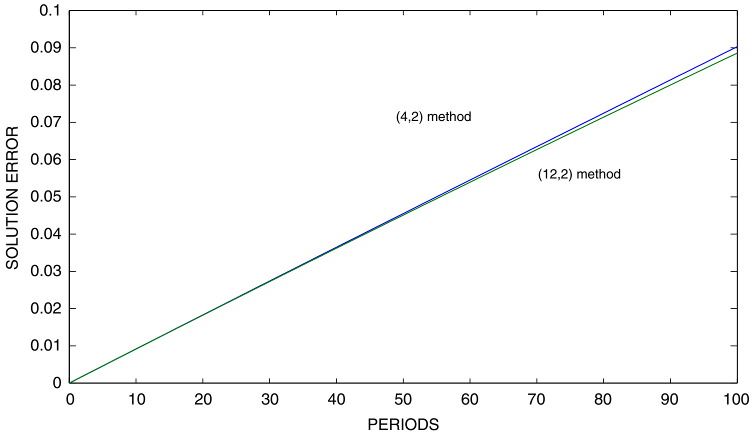

Fig. 3. Problem (38), linear growth of the error by using the energy conserving method $(12, 2)$ and the almost energy conserving method $(4, 2)$, with stepsize $h = T/50$.

of periods) is "small" (say, less than 1). This fact is duly confirmed by the plot in Fig. 3, where the error turns out to grow linearly for both the methods.

Conclusions. In this paper, we studied energy preserving methods for Poisson type systems. The proposed methods approximate the local solution in a step by a truncated Fourier series with coefficients defined by some integrals. We show that the proposed approach allows us to obtain the methods recently derived in [17], giving an alternative derivation of such methods and a new proof of their order. In addition, sufficient conditions that guarantee the existence and uniqueness of the solution of the implicit equations defining the method are given. To provide effective numerical methods quadrature formulae with k nodes are considered to approximate the involved integrals. We show that the order is maintained and the methods are energy-conserving for polynomial Hamiltonians of suitable degree and "practically" energy-conserving, for k large enough, for all sufficiently regular Hamiltonians. Quadratic Casimirs are also preserved. Some numerical experiments are presented to confirm the above theory.

Acknowledgments

This paper emerged from a visit of the first author to I.U.M.A., in February 2011. This visit has been possible because of the financial support received by that Institute.

References

[1] M. Calvo, D. Hernández-Abreu, J.I. Montijano, L. Rández, On the preservation of invariants by explicit Runge–Kutta methods, SIAM J. Sci. Comput. 28 (3) (2006) 868–885.

[2] G.R.W. Quispel, D.I. McLaren, A new class of energy-preserving numerical integration methods, J. Phys. A: Math. Theor. 41 (045206) (2008) 7pp.

[3] E. Hairer, Energy-preserving variant of collocation methods, J. Numer. Anal. Ind. Appl. Math. 5 (1–2) (2010) 73–84.

[4] L. Brugnano, F. Iavernaro, D. Trigiante, Hamiltonian boundary value methods (energy preserving discrete line integral methods), J. Numer. Anal. Ind. Appl. Math. 5 (1–2) (2010) 17–37. arXiv:0910.3621.

[5] F. Iavernaro, B. Pace, S-Stage trapezoidal methods for the conservation of Hamiltonian functions of polynomial type, AIP Conf. Proc. 936 (2007) 603–606.

[6] E. Celledoni, R.I. McLachlan, D. McLaren, B. Owren, G.R.W. Quispel, W.M. Wright, Energy preserving Runge–Kutta methods, M2AN 43 (2009) 645–649.

[7] L. Brugnano, F. Iavernaro, D. Trigiante, Analysis of Hamiltonian boundary value methods (HBVMs): a class of energy-preserving Runge–Kutta methods for the numerical solution of polynomial Hamiltonian dynamical systems, 2009, Preprint. arXiv:0909.5659.

[8] L. Brugnano, F. Iavernaro, D. Trigiante, Hamiltonian BVMs (HBVMs): a family of "drift-free" methods for integrating polynomial Hamiltonian systems, AIP Conf. Proc. 1168 (2009) 715–718.

[9] L. Brugnano, F. Iavernaro, T. Susca, Numerical comparisons between Gauss–Legendre methods and Hamiltonian BVMs defined over Gauss points, Monogr. Real Aced. Ci. Zaragoza 33 (2010) 95–112.

[10] L. Brugnano, F. Iavernaro, D. Trigiante, The Hamiltonian BVMs (HBVMs) homepage. arXiv:1002.2757. URL: http://www.math.unifi.it/~brugnano/HBVM/.

[11] F. Iavernaro, B. Pace, Conservative block-boundary value methods for the solution of polynomial Hamiltonian systems, AIP Conf. Proc. 1048 (2008) 888–891.

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

[13] L. Brugnano, F. Iavernaro, D. Trigiante, Numerical solution of ODEs and the Columbus' egg: three simple ideas for three difficult problems, Math. Eng. Sci. Aerosp. 1 (4) (2010) 105–124. arXiv:1008.4789.

[14] L. Brugnano, F. Iavernaro, D. Trigiante, A simple framework for the derivation and analysis of effective classes of one-step methods for ODEs, Appl. Math. Comput. (in press) 10.1016/j.amc.2012.01.074.

[15] L. Brugnano, F. Iavernaro, D. Trigiante, The lack of continuity and the role of infinite and infinitesimal in numerical methods for ODEs: the case of symplecticity, Appl. Math. Comput. (in press) 10.1016/j.amc.2011.03.022.

[16] L. Brugnano, F. Iavernaro, Line integral methods which preserve all invariants of conservative problems. (This volume).

[17] D. Cohen, E. Hairer, Linear energy-preserving integrators for Poisson systems, BIT Numer. Math. 51 (1) (2011) 91–101.

[18] M. Dahlby, B. Owren, T. Yaguchi, Preserving multiple first integrals by discrete gradients, Technical report, Norwegian University of Science and Technology, Trondheim, Numerics no. 11/2010. arXiv:1011.0478v4.

[19] M. Calvo, M.P. Laburta, J.I. Montijano, L. Rández, Error growth in the numerical integration of periodic orbits, Math. Comput. Simulation 81 (2011) 2646–2661.

[20] E. Isaacson, H.B. Keller, Analysis of Numerical Methods, Wiley & Sons, 1966.

[21] E.T. Whittaker, G.N. Watson, A Course in Modern Analysis, Fourth edition, Cambridge University Press, 1950.

[22] V. Lakshmikantham, D. Trigiante, Theory of Difference Equations. Numerical Methods and Applications, Academic Press, 1988.

[23] G.R.W. Quispel, H.W. Capel, Solving ODE's numerically while preserving a first integral, Phys. Lett. 218A (1996) 223–228.

[24] L. Brugnano, F. Iavernaro, D. Trigiante, A note on the efficient implementation of Hamiltonian BVMs, J. Comput. Appl. Math. 236 (2011) 375–383.