FLORE
Repository istituzionale dell'Università degli Studi di Firenze

Analysis of Energy and QUadratic Invariant Preserving (EQUIP) methods

(Article begins on next page)

18 April 2024

Corresponding Author: Professor Felice IAVERNARO,

Corresponding Author's Institution: University of Bari

First Author: Luigi Brugnano

Order of Authors: Luigi Brugnano; Gianmarco Gurioli; Felice Iavernaro

Abstract: In this paper we are concerned with the analysis of a class of geometric integrators, at first devised in [14, 18], which can be regarded as an energy-conserving variant of Gauss collocation methods. With these latter they share the property of conserving quadratic first integrals but, in addition, they also conserve the Hamiltonian function itself. We here reformulate the methods in a more convenient way, and propose a more refined analysis than that given in [18] also providing, as a by-product, a practical procedure for their implementation. A thorough comparison with the original Gauss methods is carried out by means of a few numerical tests solving Hamiltonian and Poisson problems.

Analysis of Energy and QUadratic Invariant Preserving (EQUIP) methods

Luigi Brugnano[a], Gianmarco Gurioli[a], Felice Iavernaro[b*]

[a] Dipartimento di Matematica e Informatica "U. Dini", Viale Morgagni 67/a, 50134 Firenze, Italy
[b] Dipartimento di Matematica, Via Orabona 4, 70125 Bari, Italy

November 25, 2017

Abstract

In this paper we are concerned with the analysis of a class of geometric integrators, at first devised in [14, 18], which can be regarded as an energy-conserving variant of Gauss collocation methods. With these latter they share the property of conserving quadratic first integrals but, in addition, they also conserve the Hamiltonian function itself. We here reformulate the methods in a more convenient way, and propose a more refined analysis than that given in [18] also providing, as a by-product, a practical procedure for their implementation. A thorough comparison with the original Gauss methods is carried out by means of a few numerical tests solving Hamiltonian and Poisson problems.

Keywords: Gauss collocation methods, symplectic methods, energy-conserving methods, line integral methods, Hamiltonian problems, Poisson problems.

MSC: 65P10, 65L05.

1 Introduction

The study of numerical methods for differential equations able to retain relevant geometric properties of the solutions, is the core of what nowadays is named *geometric integration* (see, e.g., the classical monographs [40, 37, 29] and, more recently, [10, 4]). Accordingly, numerical methods able to reproduce specific geometric properties of the continuous dynamical system in the discrete one generated by their application, are often referred to as *geometric integrators*. In the framework of ODEs, this approach dates back to the work of G. Dahlquist, who devised the linear stability analysis of numerical methods for ODEs in order to assess their behavior when solving dissipative problems. In this respect, *A*-stable methods can be considered as geometric integrators, when dealing with such problems.

For a large class of problems which are of interest in the applications, the geometric properties of the solutions can be taken back to the presence of *constants of motion*, which are conserved along

---

*Corresponding author, felice.iavernaro@uniba.it.

the evolution. Such problems are, therefore, named *conservative problems*. In more detail, let us consider an ODE-IVP in the form

$$\dot{y} = f(y), \qquad t > 0, \qquad y(0) = y_0 \in \Omega \subseteq \mathbb{R}^m. \tag{1}$$

The problem admits $\nu < m$ constants of motion when there exists

$$C : \Omega \to \mathbb{R}^\nu \tag{2}$$

such that

$$y(t) \text{ solves } (1) \quad \Rightarrow \quad C(y(t)) \equiv C(y_0), \quad \forall t \geq 0. \tag{3}$$

Hereafter, we shall assume that both f and C are suitably regular.[1] From a local point of view, the conservation property (3) is equivalent to require that

$$\frac{\mathrm{d}}{\mathrm{d}t} C(y(t)) = \nabla C(y(t))^\top \dot{y}(t) = \nabla C(y(t))^\top f(y(t)) = 0, \tag{4}$$

i.e., since the initial point of the trajectory is arbitrary,

$$\nabla C(y)^\top f(y) = 0, \qquad \forall y \in \Omega. \tag{5}$$

An alternative way for requiring the conservation of $C(y(t))$ is through the use of a *line integral*,

$$0 = C(y(t)) - C(y(0)) = \int_0^t \nabla C(y(s))^\top \dot{y}(s) \mathrm{d}s, \qquad \forall t > 0, \tag{6}$$

which is clearly equivalent to (4). The reformulation (6) of the invariance property is at the basis of the so-called *(discrete) Line Integral Methods* (see, e.g., the recent monograph [10]), and here we will exploit this new framework to carry out a theoretical analysis of the methods and to devise a practical implementation strategy.

The methods we are going to consider can be regarded as a suitable modification of the basic Gauss-Legendre collocation methods [24], which we rewrite here by using the well-known W-transformation [30]. In more details, introducing the orthonormal Legendre polynomial basis $\{P_i(x)\}$ on $[0,1]$ (Π_i denotes the subspace of polynomials of degree at most i)

$$P_i(x) \in \Pi_i, \qquad \int_0^1 P_i(x) P_j(x) \mathrm{d}x = \delta_{ij}, \qquad \forall i, j = 0, 1, \dots, \tag{7}$$

the matrices

$$\mathcal{P}_s = \begin{pmatrix} P_0(c_1) & \dots & P_{s-1}(c_1) \\ \vdots & & \vdots \\ P_0(c_s) & \dots & P_{s-1}(c_s) \end{pmatrix}, \qquad \Omega = \begin{pmatrix} b_1 & & \\ & \ddots & \\ & & b_s \end{pmatrix}, \tag{8}$$

with

$$P_s(c_i) = 0, \qquad b_i = \int_0^1 \ell_i(x) \mathrm{d}x, \qquad \ell_i(x) = \prod_{j \neq i} \frac{x - c_j}{c_i - c_j}, \qquad i = 1, \dots, s, \tag{9}$$

---

[1]E.g., analytical.

2

the Legendre nodes and weights, and matrix

$$X_s = \begin{pmatrix} \xi_0 & -\xi_1 & & \\ \xi_1 & 0 & \ddots & \\ & \ddots & \ddots & -\xi_{s-1} \\ & & \xi_{s-1} & 0 \end{pmatrix}, \qquad \xi_i = \frac{1}{2\sqrt{|4i^2-1|}}, \quad i = 0, \ldots, s-1, \qquad (10)$$

the s-stage Gauss method is defined by the Butcher tableau

$$\begin{array}{c|c} \boldsymbol{c} & A := \mathcal{P}_s X_s \mathcal{P}_s^\top \Omega \\ \hline & \boldsymbol{b}^\top \end{array}, \qquad \boldsymbol{b} = \left(\begin{array}{ccc} b_1, & \ldots, & b_s \end{array} \right)^\top, \qquad \boldsymbol{c} = \left(\begin{array}{ccc} c_1, & \ldots, & c_s \end{array} \right)^\top. \qquad (11)$$

As is well known (see, e.g., [23]), such methods are algebraically stable, in that

$$\Omega A + A^\top \Omega - \boldsymbol{b}\boldsymbol{b}^\top = O, \qquad (12)$$

and, in particular, they preserve any quadratic invariant of (1). In the framework of geometric integration, condition (12) is equivalent to say that such methods are *symplectic* [36, 38]. The importance of symplectic methods relies to their application to canonical Hamiltonian systems, i.e. problems of the form

$$\dot{y} = f(y) \equiv J\nabla H(y), \qquad J = \begin{pmatrix} O_d & I_d \\ -I_d & O_d \end{pmatrix} = -J^\top = -J^{-1}, \qquad (13)$$

with $m = 2d$ and O_d and I_d the zero and identity matrices of dimension d, respectively (hereafter I_r will denote the identity matrix of dimension r). Moreover,

$$H : \Omega \subseteq \mathbb{R}^m \to \mathbb{R} \qquad (14)$$

is the Hamiltonian function defining the problem, which we assume to be suitably regular. As is well-known, in such a case, the Hamiltonian function is a constant of motion for the dynamical system defined by (13)–(14). By virtue of (12) a symplectic method does preserve quadratic first integrals including the Hamiltonian function itself in case (13) is linear. The capability of conserving quadratic invariants has been recently recognized as one of the main advantages in using symplectic integrators (see [39]).

With this premise, assuming hereafter that only one constant of motion exists for (1) (i.e., $\nu = 1$ in (2)), besides possible quadratic invariants,[2] the methods we shall study have a Runge-Kutta representation in the form (cf. (11))

$$\begin{array}{c|c} \boldsymbol{c} & A(\alpha) := \mathcal{P}_s X_s(\alpha) \mathcal{P}_s^\top \Omega \\ \hline & \boldsymbol{b}^\top \end{array}, \qquad (15)$$

---

[2]We observe that the methods could in principle be generalized to cope with the conservation of more invariants (i.e., $\nu > 1$), as sketched in [18]. Nevertheless, the arguments become more involved, so that we shall not consider this case, here.

where $\mathcal{P}_s$ and Ω are the same matrices defined in (8), and (see (10)), by setting $\boldsymbol{e}_i \in \mathbb{R}^s$ the ith unit vector,

$$X_s(\alpha) := X_s - \alpha W_s \equiv \begin{pmatrix} \xi_0 & \alpha - \xi_1 & & \\ \xi_1 - \alpha & 0 & \ddots & \\ & \ddots & \ddots & -\xi_{s-1} \\ & & \xi_{s-1} & 0 \end{pmatrix}, \quad W_s = \boldsymbol{e}_2\boldsymbol{e}_1^\top - \boldsymbol{e}_1\boldsymbol{e}_2^\top = -W_s^\top. \quad (16)$$

We observe that, for $\alpha = 0$, (15)–(16) coincides with the Butcher tableau (11) of the s-stage Gauss collocation method. In the present case, the scalar parameter α is chosen, at each step of application of the method for numerically solving (1), in order to enforce the conservation of the invariant (2) (or (14)) in the numerical solution (see also [18], for the Hamitonian case). To this purpose, exploiting Theorem 5.1 in [31, Chapter IV.5] and the symmetry [3] of the method (15), we see that it has order 2 for any fixed $\alpha \neq 0$ and, clearly, order $2s$ when $\alpha = 0$. More precisely, denoting by $y_1(\alpha)$ the solution yielded by the application of a single step of method (15) to problem (1), we have

$$||y_1(\alpha) - y(h)|| = \alpha O(h^3) + O(h^{2s+1}). \quad (17)$$

In Section 3 we will show that an $\alpha^* = O(h^{2s-2})$ actually exists such that $C(y_1(\alpha^*)) = C(y_0)$, which will confer an order $p = 2s$ on the resulting method, as is the case for the corresponding Gauss formula. Since $s \geq 2$, we may confine the study of existence and uniqueness of such an α^* inside a ball of radius ch^2. Therefore in the sequel we make the following assumption

$$|\alpha| \leq ch^2, \qquad \text{for a given positive constant } c. \quad (18)$$

The formulae resulting from the choice $\alpha = \alpha^*$ have been named *Energy and QUadratic Invariants Preserving* (EQUIP) methods in [18] since, beside preserving the Hamiltonian function, they continue to preserve any quadratic invariant just like the Gauss collocation methods they come from. It must be stressed that this double feature seems to confer great robustness to the methods, as we will see later in the numerical tests. For sake of completeness, we mention that the idea of modifying Runge-Kutta methods, by introducing a suitable parameter to enforce energy conservation, has been also considered, e.g., in [25, 35].

An aspect which has not yet been studied neither in [18] nor in the subsequent research concerns the implementation of the methods. Motivated by this issue, in the present work we exploit the characterization (6) of a constant of motion to show that the methods (15) may be indeed interpreted as discrete line integral methods. This will allow us to carry out a more refined theoretical analysis than that given in [18] and to devise a practical implementation strategy.

The structure of the paper is as follows: in Section 2 we provide full details about the derivation of the parameter α; in Section 3 we provide a novel analysis for the corresponding method, besides that given in [18]; in Section 4 we consider the fully discrete method; Section 5 is devoted to the solution of the generated discrete problem; Section 6 provides numerical evidence for the theoretical findings; at last, Section 7 contains a few conclusions and further directions of investigation.

---

[3]See Theorem 3 below.

2 Energy and QUadratic Invariants Preserving (EQUIP) methods

In this section, we derive an expression for the parameter α in (15)–(16), chosen in order to enforce the conservation of $C(y)$. Since we are speaking about a one-step method, it will suffice to consider the very first application of the method, in order to obtain the approximation

$$y_1 := y_1(\alpha) \approx y(h), \tag{19}$$

where h is the time step and $y(t)$ is the solution of (1). The derivation of the parameter α will be then derived in the framework of *line integral methods*, namely methods which enforce the conservation of the invariants through an approximation of the line integral (6). Such methods, at first derived for the numerical solution of Hamiltonian problems [32, 33, 34], eventually resulted in the class of *Hamiltonian Boundary Value Methods (HBVMs)* [6, 11, 12, 13, 15, 16, 17, 20], and have been then generalized along several directions [1, 2, 5, 7, 9, 19, 21, 41], including the methods here studied.[4]

To begin with, we start recalling that, when $\alpha = 0$, the method defined by (15)–(16) reduces to the s-stage Gauss method, which is defined by the corresponding *collocation polynomial* $\sigma \in \Pi_s$ such that

$$\sigma(0) = y_0, \qquad \dot\sigma(c_i h) = f(\sigma(c_i h)), \quad i = 1, \ldots, s, \qquad y_1 \equiv y_1(0) := \sigma(h).$$

Denoting by Y_i, $i = 1, \ldots, s$ the stages associated with the method, an equivalent way of defining the collocation polynomial $\sigma(ch)$ is via the interpolation conditions

$$\sigma(0) = y_0, \qquad Y_i = \sigma(c_i h), \quad i = 1, \ldots, s. \tag{20}$$

When $\alpha \neq 0$, the collocation polynomial is replaced with a *piecewise continuous polynomial path*, which we continue to denote by σ,[5]

$$\sigma : [0, h] \cup [h, h+1] \to \mathbb{R}^m,$$

such that the interpolation conditions (20) are still fulfilled for the stages Y_i associated with (15) and, moreover,

$$y_1 \equiv y_1(\alpha) := \sigma(h+1), \qquad s.t. \qquad C(\sigma(1+h)) = C(\sigma(0)) \iff C(y_1) = C(y_0). \tag{21}$$

For sake of simplicity, let us denote

$$\sigma_1(t) \equiv \sigma(t), \quad t \in [0, h], \qquad \sigma_2(t) \equiv \sigma(t+h), \quad t \in [0, 1], \tag{22}$$

and notice that the interpolation conditions (20) only involve the polynomial σ_1. The following results hold true.

---

[4]Despite the common *line integral* framework to gain the conservation property, the methods that we shall describe here greatly differ from HBVMs. As matter of fact, the latter methods are not symplectic and are conservative only for Hamiltonian problems, whereas the former ones are defined, at each integration step, by a symplectic map and can be conservative for more general problems.

[5]For sake of notation simplicity, in the sequel we will omit to explicitly highlight the dependence of the continuous polynomial path σ and some related quantities on the parameter α, if this is clear from the context.

Theorem 1 *For fixed α and h, and with reference to the polynomial basis (7), let us denote*

$$\gamma_j := \sum_{i=1}^{s} b_i P_j(c_i) f(Y_i), \qquad j = 0, \ldots, s - 1, \tag{23}$$

where Y_i are the stages associated with the Runge-Kutta method (15). Moreover, see (10), we define the vectors

$$\boldsymbol{\phi}_i \equiv \begin{pmatrix} \phi_{i0} \\ \vdots \\ \phi_{i,s-1} \end{pmatrix} := X_s^{-1} \boldsymbol{e}_i, \qquad i = 1, 2. \tag{24}$$

Then, the polynomial $\sigma_1 \in \Pi_s$ defined as

$$\sigma_1(ch) = y_0 + h \sum_{j=0}^{s-1} \int_0^c P_j(x) \mathrm{d}x \, [\gamma_j - \alpha \left(\phi_{2j}\gamma_0 - \phi_{1j}\gamma_1 \right)], \qquad c \in [0,1], \tag{25}$$

satisfies the interpolation conditions (20).

<u>Proof</u>   The polynomial σ_1 defined at (25) clearly satisfies the first condition in (20). Setting

$$Y := \begin{pmatrix} Y_1 \\ \vdots \\ Y_s \end{pmatrix},$$

the stage vector of (15)–(16), and

$$\sigma_1(\boldsymbol{c}h) := \begin{pmatrix} \sigma_1(c_1 h) \\ \vdots \\ \sigma_1(c_s h) \end{pmatrix}, \tag{26}$$

the stage interpolation conditions in (20) read, in compact notation, $Y = \sigma_1(\boldsymbol{c}h)$. Denoting

$$\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^s, \tag{27}$$

one obtains:

$$\begin{aligned} Y &= \mathbf{1} \otimes y_0 + h\mathcal{P}_s X_s(\alpha)\mathcal{P}_s^\top \Omega \otimes I_m f(Y) = \mathbf{1} \otimes y_0 + h\mathcal{P}_s X_s \left[I_s - \alpha X_s^{-1} W_s \right] \mathcal{P}_s^\top \Omega \otimes I_m f(Y) \\ &= \mathbf{1} \otimes y_0 + h\mathcal{P}_s X_s \left[I_s - \alpha(\boldsymbol{\phi}_2 \boldsymbol{e}_1^\top - \boldsymbol{\phi}_1 \boldsymbol{e}_2^\top) \right] \mathcal{P}_s^\top \Omega \otimes I_m f(Y). \end{aligned} \tag{28}$$

Now, considering the vector form of (23),

$$\boldsymbol{\gamma} := \begin{pmatrix} \gamma_0 \\ \vdots \\ \gamma_s \end{pmatrix} = \mathcal{P}_s^\top \Omega \otimes I_m f(Y) \tag{29}$$

and the property

$$\mathcal{P}_s X_s = \mathcal{I}_s := \begin{pmatrix} \int_0^{c_1} P_0(x)\mathrm{d}x & \dots & \int_0^{c_1} P_{s-1}(x)\mathrm{d}x \\ \vdots & & \vdots \\ \int_0^{c_s} P_0(x)\mathrm{d}x & \dots & \int_0^{c_s} P_{s-1}(x)\mathrm{d}x \end{pmatrix}, \tag{30}$$

from (28) and (25) we deduce that

$$Y = \mathbf{1} \otimes y_0 + h\mathcal{P}_s X_s \left[I_s - \alpha(\boldsymbol{\phi}_2 \boldsymbol{e}_1^\top - \boldsymbol{\phi}_1 \boldsymbol{e}_2^\top) \right] \otimes I_m \boldsymbol{\gamma} = \sigma_1(\boldsymbol{ch}). \tag{31}$$

This concludes the proof. $\square$

Concerning the polynomial $\sigma_2(t)$ in (22), it will be used to gain the conservation property. In more details, one has:

$$\sigma_2(0) \equiv \sigma(h) = \sigma_1(h) = y_0 + h\left[\gamma_0 - \alpha\left(\phi_{20}\gamma_0 - \phi_{10}\gamma_1\right)\right], \tag{32}$$

$$\sigma_2(1) \equiv \sigma(h+1) =: y_1 = y_0 + h\sum_{i=1}^s b_i f(Y_i) \equiv y_0 + h\gamma_0. \tag{33}$$

Consequently, we can choose a linear polynomial:

$$\sigma_2(t) = y_1 + (t-1)\,\alpha h\left(\phi_{20}\gamma_0 - \phi_{10}\gamma_1\right), \qquad t \in [0,1]. \tag{34}$$

We observe that, from (25) and (34), we obtain: [6]

$$\dot{\sigma}_1(ch) = \sum_{j=0}^{s-1} P_j(c)\left[\gamma_j - \alpha\left(\phi_{2j}\gamma_0 - \phi_{1j}\gamma_1\right)\right], \qquad \dot{\sigma}_2(c) \equiv \alpha h\left(\phi_{20}\gamma_0 - \phi_{10}\gamma_1\right), \qquad c \in [0,1]. \tag{35}$$

Concerning the choice of the parameter α such that the conservation condition $C(y_1) = C(y_0)$ in (21) is fulfilled, the following result holds true.

Theorem 2 *Let us consider the polynomials σ_1 and σ_2 defined at (25) and (34), respectively, and the vectors γ_j defined at (23). Morever, we define the vectors*

$$\rho_j(\sigma_1) = \int_0^1 P_j(c)\nabla C(\sigma_1(ch))\mathrm{d}c, \quad j = 0, \dots, s-1, \qquad \bar{\rho}(\sigma_2) = \int_0^1 \nabla C(\sigma_2(c))\mathrm{d}c. \tag{36}$$

Then $C(y_1) = C(y_0)$, provided that:

$$\alpha = \frac{\sum_{j=0}^{s-1} \rho_j(\sigma_1)^\top \gamma_j}{\left(\rho_0(\sigma_1) - \bar{\rho}(\sigma_2)\right)^\top \left(\phi_{20}\gamma_0 - \phi_{10}\gamma_1\right) + \sum_{j=1}^{s-1} \rho_j(\sigma_1)^\top \left(\phi_{2j}\gamma_0 - \phi_{1j}\gamma_1\right)}. \tag{37}$$

---

[6]Clearly, from (34)–(35) one obtains that $\sigma_2(t) \equiv y_1$, when $\alpha = 0$.

<u>Proof</u> By using the framework provided by line integral methods, from (22), (25), (34), and (35), one has:

$$
\begin{aligned}
0 &= C(y_1) - C(y_0) \\
&= C(\sigma(h+1)) - C(\sigma(0)) \;=\; C(\sigma_2(1)) - C(\sigma_2(0)) + C(\sigma_1(h)) - C(\sigma_1(0)) \\
&= \int_0^1 \nabla C(\sigma_2(c))^\top \dot\sigma_2(c)\mathrm{d}c + \int_0^h \nabla C(\sigma_1(t))^\top \dot\sigma_1(t)\mathrm{d}t \\
&= \int_0^1 \nabla C(\sigma_2(c))^\top \dot\sigma_2(c)\mathrm{d}c + h\int_0^1 \nabla C(\sigma_1(ch))^\top \dot\sigma_1(ch)\mathrm{d}c \\
&= \alpha h\left[\int_0^1 \nabla C(\sigma_2(c))\mathrm{d}c\right]^\top (\phi_{20}\gamma_0 - \phi_{10}\gamma_1) \\
&\quad + h\int_0^1 \nabla C(\sigma_1(ch))^\top \sum_{j=0}^{s-1} P_j(c)\mathrm{d}c\,[\gamma_j - \alpha(\phi_{2j}\gamma_0 - \phi_{1j}\gamma_1)]\,\mathrm{d}c \\
&= \alpha h\bar\rho(\sigma_2)^\top(\phi_{20}\gamma_0 - \phi_{10}\gamma_1) + h\sum_{j=0}^{s-1}\rho_j(\sigma_1)^\top[\gamma_j - \alpha(\phi_{2j}\gamma_0 - \phi_{1j}\gamma_1)],
\end{aligned}
$$

from which (37) immediately follows. $\square$

In case of the Hamiltonian problem (13)–(14), the previous result is modified as follows.

Corollary 1 *Let us consider the polynomials σ_1 and σ_2 defined at (25) and (34), respectively, and the vectors γ_j defined at (23). Moreover, with reference to (13), let us define the vectors*

$$
\gamma_j(\sigma_1) = \int_0^1 P_j(c)f(\sigma_1(ch))\mathrm{d}c, \quad j = 0,\dots,s-1, \qquad \bar\gamma(\sigma_2) = \int_0^1 f(\sigma_2(c))\mathrm{d}c. \tag{38}
$$

Then $H(y_1) = H(y_0)$, provided that:

$$
\alpha = \frac{\sum_{j=0}^{s-1}\gamma_j(\sigma_1)^\top J\gamma_j}{(\gamma_0(\sigma_1) - \bar\gamma(\sigma_2))^\top J(\phi_{20}\gamma_0 - \phi_{10}\gamma_1) + \sum_{j=1}^{s-1}\gamma_j(\sigma_1)^\top J(\phi_{2j}\gamma_0 - \phi_{1j}\gamma_1)}. \tag{39}
$$

<u>Proof</u> The expression (39) follows immediately from (37) by taking into account (13) and setting $C(y) = H(y)$. In fact, from (36) one obtains: $\gamma_j(\sigma_1) = J\rho_j(\sigma_1)$, $j = 0,\dots,s-1$, and $\bar\gamma(\sigma_2) = J\bar\rho(\sigma_2)$. $\square$

The method defined by (15)–(16), with α given by (37) (i.e., (39), in the Hamiltonian case (13)), are named [18] s-stage EQUIP methods: the name being accounted for by their analysis in the next section. We conclude this section observing that the discrete line integral tool provides us with a new nonlinear system associated with an EQUIP method, where the stage vector Y is replaced by the vector $\boldsymbol\gamma$. In fact, inserting (31) in (29) we obtain that the unknown vectors γ_j are determined by solving the system

$$
\boldsymbol\gamma = \mathcal{P}_s^\top \Omega \otimes I_m\, f(\mathbf{1}\otimes y_0 + h(\mathcal{I}_s - \alpha\mathcal{P}_s W_s)\otimes I_m\,\boldsymbol\gamma), \tag{40}
$$

with $f(\mathbf{1}\otimes y_0 + h(\mathcal{I}_s - \alpha\mathcal{P}_s W_s)\otimes I_m\,\boldsymbol\gamma)$ denoting the block vector containing the application of the function f to each block entry of the argument. This new representation of an EQUIP method

made up of (40), and either (37) or (39), will be exploited in the sequel to derive a number of theoretical results. Compared with [18], where only the existence of an α yielding energy conservation was proved without giving details about how α should be searched, here equation (37) yields an implicit equation the parameter α must satisfy, and will provide us with a practical strategy for the implementation of the methods.

3 Analysis of EQUIP methods

In what follows, we shall obviously assume $s \geq 2$ in (15)–(16).[7] As already observed, when $\alpha = 0$, the s-stage EQUIP method (15)–(16) reduces to the s-stage Gauss-collocation method. Another relevant property is symmetry.

Theorem 3 *For any fixed $\alpha \in \mathbb{R}$ the s-stage Runge-Kutta method (15)–(16) is symmetric.*

<u>Proof</u>   Since there are no redundant stages, the symmetry of the method is equivalent to require (see (27))

$$PA(\alpha) + A(\alpha)P = \mathbf{1}\boldsymbol{b}^\top, \qquad \text{and} \qquad P\boldsymbol{b} = \boldsymbol{b},$$

with

$$P := \begin{pmatrix} & & 1 \\ & \cdot^{\cdot^{\cdot}} & \\ 1 & & \end{pmatrix} = P^\top = P^{-1} \in \mathbb{R}^{s \times s}.$$

The latter condition is trivial. Concerning the former, one has:

$$
\begin{aligned}
PA(\alpha) + A(\alpha)P &= PA(0) + A(0)P - \alpha \left(P\mathcal{P}_s W_s \mathcal{P}_s^\top \Omega + \mathcal{P}_s W_s \mathcal{P}_s^\top \Omega P \right) \\
&= \mathbf{1}\boldsymbol{b}^\top - \alpha \left(P\mathcal{P}_s W_s \mathcal{P}_s^\top \Omega + \mathcal{P}_s W_s \mathcal{P}_s^\top \Omega P \right).
\end{aligned}
$$

Since (see (8)) $P\Omega P = \Omega$ and $W_s^\top = -W_s$, we obtain

$$
\begin{aligned}
\left(P\mathcal{P}_s W_s \mathcal{P}_s^\top \Omega + \mathcal{P}_s W_s \mathcal{P}_s^\top \Omega P \right) &= \left(P\mathcal{P}_s W_s \mathcal{P}_s^\top + \mathcal{P}_s W_s \mathcal{P}_s^\top P \right) \Omega \\
&= \left((P\mathcal{P}_s W_s)\mathcal{P}_s^\top - \mathcal{P}_s (P\mathcal{P}_s W_s)^\top \right) \Omega =: \Psi.
\end{aligned}
$$

By considering that $P_j(c_i) = (-1)^j P_j(c_{s-i+1})$ and $W_s = \boldsymbol{e}_2\boldsymbol{e}_1^\top - \boldsymbol{e}_1\boldsymbol{e}_2^\top$, one obtains:

$$P\mathcal{P}_s W_s = P\mathcal{P}_s \left(\boldsymbol{e}_2\boldsymbol{e}_1^\top - \boldsymbol{e}_1\boldsymbol{e}_2^\top \right) = -\mathcal{P}_s \left(\boldsymbol{e}_2\boldsymbol{e}_1^\top + \boldsymbol{e}_1\boldsymbol{e}_2^\top \right) =: -\mathcal{P}_s G,$$

with $G = G^\top$. Consequently, matrix $\mathcal{P}_s G \mathcal{P}_s^\top$ is symmetric and, therefore:

$$\Psi = -\mathcal{P}_s G \mathcal{P}_s^\top + \mathcal{P}_s G \mathcal{P}_s^\top = O. \;\square$$

A further relevant property of the s-stage method (15)–(16) is algebraic stability.

Theorem 4 *For any fixed $\alpha \in \mathbb{R}$ the s-stage Runge-Kutta method (15)–(16) satisfies (12).*

---

[7]In general, $s \geq \nu + 1$, in case of ν constants of motion.

<u>Proof</u> In fact, for any fixed $\alpha \in \mathbb{R}$, from (8)–(10) and (15)–(16) one has:

$$
\begin{aligned}
\Omega A(\alpha) + A(\alpha)^\top \Omega &= \Omega \mathcal{P}_s (X_s - \alpha W_s) \mathcal{P}_s^\top \Omega + \Omega \mathcal{P}_s (X_s - \alpha W_s)^\top \mathcal{P}_s^\top \Omega \\
&= \Omega \mathcal{P}_s \overbrace{(X_s + X_s^\top)}^{=e_1 e_1^\top} \mathcal{P}_s^\top \Omega - \alpha \Omega \mathcal{P}_s \overbrace{(W_s + W_s^\top)}^{=O} \mathcal{P}_s^\top \Omega = \boldsymbol{b} \boldsymbol{b}^\top + O. \ \square
\end{aligned}
$$

As a straightforward consequence, one obtains the following result.

Corollary 2 *For any fixed $\alpha \in \mathbb{R}$ the s-stage Runge-Kutta method (15)–(16) is symplectic and, therefore, conserves all quadratic invariants of (1).*

<u>Proof</u> See [23, 36, 38]. $\square$

Remark 1 *The name* Energy and QUadratic Invariants Preserving (EQUIP) method *derives from the conservation properties associated with Theorem 2 (i.e., Corollary 1, in the Hamiltonian case), and Corollary 2.*

We notice that all the properties discussed in Theorems 3 and 4, and in Corollary 2 above, hold true independently of the choice of the parameter α. If we let equation (37) (or its variant (39)) come into play, a primary issue is to ascertain the existence and uniqueness of the solution of the nonlinear system associated with the method, namely (37)–(40) (or (39)–(40)). For this purpose, we need some additional preliminary results.

Lemma 1 *Let $G : \mathbb{R} \to V$, with V a vector space, be a function admitting a Taylor expansion at 0. Then,*

$$
\int_0^1 P_j(c) G(ch) \mathrm{d}c = O(h^j), \qquad j = 0, 1, \ldots.
$$

<u>Proof</u> See [17, Lemma 1]. $\square$

Lemma 2 *With reference to (23), (36), and (38), one has, under regularity assumptions on f and C:*

$$
\gamma_j(\sigma_1) = O(h^j), \qquad \rho_j(\sigma_1) = O(h^j), \qquad \gamma_j = \gamma_j(\sigma_1) + O(h^{2s-j}), \qquad j \geq 0, \tag{41}
$$

and

$$
\bar{\rho}(\sigma_2) = O(1), \qquad \bar{\gamma}(\sigma_2) = O(1). \tag{42}
$$

<u>Proof</u> The thesis easily follows from Lemma 1. $\square$

We need now to discuss the right-hand sides of the expressions at (37) and (39): we shall then study the corresponding numerators and denominators.

Analysis of the numerators in (37) and (39)

Hereafter, let us denote $\mathcal{N}(h)$ the numerator at the right-hand side of (37) or (39).

Lemma 3 *Concerning the numerator in (37), under regularity assumptions on f and C, one has*

$$
\mathcal{N}(h) := \sum_{j=0}^{s-1} \rho_j(\sigma_1)^\top \gamma_j = O(h^{2s}). \tag{43}
$$

10

<u>Proof</u> Following the arguments in [9], let us consider the following expansions along the orthonormal basis (7):

$$\nabla C(\sigma_1(ch)) = \sum_{j\geq0} P_j(c)\rho_j(\sigma_1), \qquad f(\sigma_1(ch)) = \sum_{j\geq0} P_j(c)\gamma_j(\sigma_1), \qquad c \in [0,1],$$

with $\rho_j(\sigma_1)$ and $\gamma_j(\sigma_1)$ given by (36) and (38), respectively. Because of (5) and (7), one has, then:

$$0 = \int_0^1 \nabla C(\sigma_1(ch))^\top f(\sigma_1(ch))\mathrm{d}c = \sum_{i,j\geq0} \rho_i(\sigma_1)^\top \gamma_j(\sigma_1) \underbrace{\int_0^1 P_i(c)P_j(c)\mathrm{d}c}_{=\,\delta_{ij}} = \sum_{j\geq0} \rho_j(\sigma_1)^\top \gamma_j(\sigma_1).$$

Consequently, by virtue of (41), one obtains

$$\sum_{j=0}^{s-1} \rho_j(\sigma_1)^\top \gamma_j(\sigma_1) = -\sum_{j\geq s} \rho_j(\sigma_1)^\top \gamma_j(\sigma_1) = O(h^{2s}).$$

Then,

$$\sum_{j=0}^{s-1} \rho_j(\sigma_1)^\top \gamma_j = \sum_{j=0}^{s-1} \rho_j(\sigma_1)^\top \left[\gamma_j(\sigma_1) + O(h^{2s-j})\right] = O(h^{2s}) + \sum_{j=0}^{s-1} O(h^j)O(h^{2s-j}) = O(h^{2s}). \;\square$$

For the particular case of Hamiltonian problems, one has the following simpler result.

Lemma 4 *Concerning the numerator in (39), under regularity assumptions on H, one has*

$$\mathcal{N}(h) := \sum_{j=0}^{s-1} \gamma_j(\sigma_1)^\top J\gamma_j = O(h^{2s}). \tag{44}$$

<u>Proof</u> In fact, from (41), one has, by taking into account that matrix J is skew-symmetric:

$$\sum_{j=0}^{s-1} \gamma_j(\sigma_1)^\top J\gamma_j = \sum_{j=0}^{s-1} \gamma_j(\sigma_1)^\top J\left[\gamma_j(\sigma_1) + O(h^{2s-j})\right] = 0 + \sum_{j=0}^{s-1} O(h^j)O(h^{2s-j}) = O(h^{2s}). \;\square$$

Analysis of the denominators in (37) and (39)

Before studying the denominators at the right-hand sides of (37) and (39), which we shall hereafter denote by $\mathcal{D}(h)$, we need to state the following result concerning the leading entries of the vectors defined at (24).

Lemma 5 *With reference to (10) and (24), one has (discarding ϕ_{ij}, when $j \geq s$):*

$$\left(\phi_{10}, \quad \phi_{11}, \quad \phi_{12}, \quad \phi_{13} \right) = \begin{cases} (\xi_0^{-1}, 0, \xi_1(\xi_0\xi_2)^{-1}, 0), & s \quad odd, \\ (0, -\xi_1^{-1}, 0, \neq 0), & s \quad even, \end{cases} \tag{45}$$

$$\left(\phi_{20}, \quad \phi_{21}, \quad \phi_{22}, \quad \phi_{23} \right) = \begin{cases} (0, 0, -\xi_2^{-1}, 0), & s \quad odd, \\ (\xi_1^{-1}, \xi_0\xi_1^{-2}, 0, \neq 0), & s \quad even. \end{cases} \tag{46}$$

11

Moreover, it is quite straightforward, though lengthy, to prove the following expansions, which hold under the assumption (18).

Lemma 6 *Let us consider the vectors defined at (23), (36), and (38), and the parameter α defined at either (37) or (39). Moreover, let us set*

$$f_0 := f(y_0), \qquad f_0' := f'(y_0), \qquad f_0'' := f''(y_0),$$

$$g(y) := \nabla C(y) \qquad g_0 := g(y_0), \qquad g_0' := g'(y_0), \qquad g_0'' := g''(y_0). \tag{47}$$

Then, under regularity assumptions on the involved functions, and assuming that (18) is satisfied, the following expansions hold true for the following coefficients defined at (23), (36), and (38):[8]

$$
\begin{aligned}
\gamma_0(\sigma_1), \gamma_0 &= f_0 + \xi_0 h f_0' f_0 + h^2[(\xi_0^2 - \xi_1^2)(f_0')^2 f_0 + \xi_0^{-1}\xi_1^2 f_0''(f_0, f_0)] + O(h^3), \\
\gamma_1(\sigma_1), \gamma_1 &= \xi_1 h f_0' f_0 + \xi_0 \xi_1 h^2[(f_0')^2 f_0 + f_0''(f_0, f_0)] + O(h^3), \\
\gamma_2(\sigma_1), \gamma_2 &= \xi_1 \xi_2 h^2[(f_0')^2 f_0 + f_0''(f_0, f_0)] + O(h^3), \\
\bar{\gamma}(\sigma_2) &= f_0 + h f_0' f_0 + \xi_0 h^2[(f_0')^2 f_0 + f_0''(f_0, f_0)] + O(h^3),
\end{aligned}
\tag{48}
$$

and

$$
\begin{aligned}
\rho_0(\sigma_1) &= g_0 + \xi_0 h g_0' f_0 + h^2[(\xi_0^2 - \xi_1^2) g_0' f_0' f_0 + \xi_0^{-1}\xi_1^2 g_0''(f_0, f_0)] + O(h^3), \\
\rho_1(\sigma_1) &= \xi_1 h g_0' f_0 + \xi_0 \xi_1 h^2[g_0' f_0' f_0 + g_0''(f_0, f_0)] + O(h^3), \\
\rho_2(\sigma_1) &= \xi_1 \xi_2 h^2[g_0' f_0' f_0 + g_0''(f_0, f_0)] + O(h^3), \\
\bar{\rho}(\sigma_2) &= g_0 + h g_0' f_0 + \xi_0 h^2[g_0' f_0' f_0 + g_0''(f_0, f_0)] + O(h^3).
\end{aligned}
\tag{49}
$$

The next results then follow.

Lemma 7 *Let us consider the denominator at the right-hand side in (37). Then, under regularity assumptions on f and assuming (18) be satisfied, one has:*

$$\mathcal{D}(h) = \xi_1 h^2 \left[f_0^\top ((g_0')^\top f_0' - (g_0' f_0')^\top) f_0 - (g_0''(f_0, f_0))^\top f_0 \right] + O(h^3). \tag{50}$$

<u>Proof</u> We set $\mathcal{D}(h) = D_1(h) + D_2(h),$ with

$$
\begin{aligned}
D_1(h) &:= (\rho_0(\sigma_1) - \bar{\rho}(\sigma_2))^\top (\phi_{20}\gamma_0 - \phi_{10}\gamma_1) + \rho_1(\sigma_1)^\top (\phi_{21}\gamma_0 - \phi_{11}\gamma_1), \\
D_2(h) &:= \sum_{j=2}^{s-1} \rho_j(\sigma_1)^\top (\phi_{2j}\gamma_0 - \phi_{1j}\gamma_1).
\end{aligned}
$$

Then, by taking into account (41) and (45)–(49), we discuss separately the case s odd and s even.

<u>s odd:</u> in such a case, one has, by taking into account that $\phi_{22}\xi_1\xi_2 = -\xi_1,$

$$
\begin{aligned}
D_1(h) &= -\xi_0^{-1}(\rho_0(\sigma_1) - \bar{\rho}(\sigma_2))^\top \gamma_1 = -\xi_0^{-1}[-\xi_0 h g_0' f_0 + O(h^2)]^\top (\xi_1 h f_0' f_0 + O(h^2)) \\
&= \xi_1 h^2 (g_0' f_0)^\top (f_0' f_0) + O(h^3), \\
D_2(h) &= \phi_{22}\rho_2(\sigma_1)^\top \gamma_0 + O(h^3) = \phi_{22}\xi_1\xi_2 h^2[g_0' f_0' f_0 + g_0''(f_0, f_0)]^\top f_0 + O(h^3) \\
&= -\xi_1 h^2[(g_0' f_0' f_0)^\top f_0 + (g_0''(f_0, f_0))^\top f_0] + O(h^3).
\end{aligned}
$$

Consequently, (50) follows.

___
[8]They are obtained by expanding $f(y)$ and $\nabla C(y)$ at $y = y_0$.

<u>s even</u>: in such a case, one has, by taking into account that $(\xi_0^{-1}\xi_1 - \xi_0\xi_1^{-1}) = -\xi_0^{-2}\xi_1$ and $(\xi_0^2\xi_1^{-1} - \xi_0^{-2}\xi_1) = -\xi_1$,

$$
\begin{aligned}
D_1(h) &= \xi_1^{-1}(\rho_0(\sigma_1) - \bar{\rho}(\sigma_2))^\top \gamma_0 + \rho_1(\sigma_1)^\top(\xi_0\xi_1^{-2}\gamma_0 + \xi_1^{-1}\gamma_1) \\
&= \xi_1^{-1}[-\xi_0 h g_0' f_0 - \xi_0^{-2}\xi_1^2 h^2 g_0' f_0' f_0 + (\xi_0^{-1}\xi_1^2 - \xi_0)h^2 g_0''(f_0, f_0)]^\top(f_0 + \xi_0 h f_0' f_0) + \\
&\quad \xi_0\xi_1^{-2}[\xi_1 h g_0' f_0 + \xi_0\xi_1 h^2[g_0' f_0' f_0 + g_0''(f_0, f_0)]]^\top(f_0 + \xi_0 h f_0' f_0) + \\
&\quad \xi_1^{-1}(\xi_1 h g_0' f_0)^\top(\xi_1 h f_0' f_0) + O(h^3) \\
&= -\xi_1 h^2 (g_0' f_0' f_0)^\top f_0 - \xi_1 h^2 (g_0''(f_0, f_0))^\top f_0 + \xi_1 h^2(g_0' f_0)^\top(f_0' f_0) + O(h^3), \\
D_2(h) &= \phi_{23}\rho_3(\sigma_1)^\top \gamma_0 + O(h^4) = O(h^3),
\end{aligned}
$$

from which (50) follows. $\square$

In the Hamiltonian case, by considering (13)–(14) (i.e., $Jg(y) \equiv f(y)$), and the expansions (48), one proves the following result in a similar way.

Lemma 8 *Let us consider the denominator at the right-hand side in (39). Then, under regularity assumptions on f and assuming (18) be satisfied, one has:*

$$
\mathcal{D}(h) = -\xi_1 h^2 \left[f_0''(f_0, f_0)\right]^\top J f_0 + O(h^3). \tag{51}
$$

Existence and uniqueness

As a simple consequence of the previous Lemmas 3–8, one obtains the following result, which provides a different, and more general, proof than what proved in [18, Theorem 3.5] in the Hamiltonian case.

Theorem 5 *Assume that α satisfies equation (37) and*

$$
\left[f_0^\top((g_0')^\top f_0' - (g_0' f_0')^\top)f_0 - (g_0''(f_0, f_0))^\top f_0\right] \neq 0. \tag{52}
$$

Then,

$$
\alpha = O(h^{2(s-1)}). \tag{53}
$$

The same result holds true in the Hamiltonian case, when α is given by the right-hand side in (39), and

$$
\left[f_0''(f_0, f_0)\right]^\top J f_0 \neq 0. \tag{54}
$$

<u>Proof</u> Depending on the two cases, one obtains $\alpha = \mathcal{N}(h)/\mathcal{D}(h)$, with $\mathcal{N}(h)$ and $\mathcal{D}(h)$ defined according to (43) and (50), when α is given by (37), or (44) and (51), when α is given by (39). In both cases, (53) easily follows, due to the fact that $\mathcal{N}(h) = O(h^{2s})$ and $\mathcal{D}(h) = O(h^2)$. $\square$

Remark 2 *We observe that when (52) is not true, then the leading term in (50) vanishes. Practically, this event can be recognized by the slow down or failure of the convergence of the nonlinear iteration later described in Section 5 which, in turn, relies on the result of Theorem 6 below. In such a case, one may set $\alpha = 0$, thus performing a single step of the basic s-stage Gauss method. The same argument applies, in the Hamiltonian case, when (54) is not satisfied, so that the leading term in (51) vanishes. This strategy will allow to retain the order $2s$ of the underlying s-stage Gauss method, as we shall see later in the analysis of the method.*

13

We are now in the right position to prove the existence and uniqueness result. The nonlinear system to be solved at each step of the integration procedure is (see (37)–(39) and (40)) [9]

$$\begin{cases} \alpha = \dfrac{N(\alpha,\boldsymbol{\gamma})}{D(\alpha,\boldsymbol{\gamma})} := \dfrac{\mathcal{N}(h)}{\mathcal{D}(h)}, \\[3mm] \boldsymbol{\gamma} = \Psi(\alpha,\boldsymbol{\gamma}) := \mathcal{P}_s^\top \Omega \otimes I_m \, f(\mathbf{1} \otimes y_0 + h(\mathcal{I}_s - \alpha \mathcal{P}_s W_s) \otimes I_m \, \boldsymbol{\gamma}). \end{cases} \tag{55}$$

Setting $\alpha = 0$ in the first equation of (55) leads back to the nonlinear system associated with the Gauss method of order $2s$ that we write as

$$\boldsymbol{\gamma} = \Psi_0(\boldsymbol{\gamma}) := \Psi(0,\boldsymbol{\gamma}). \tag{56}$$

Assuming $f : \mathbb{R}^m \to \mathbb{R}^m$ Lipschitz continuous with constant $L > 0$, we see that $\Psi_0(\boldsymbol{\gamma})$ is Lipschitz continuous in $\mathbb{R}^{sm}$ with constant [10]

$$K_0 := hL\|\mathcal{P}_s^\top \Omega\| \, \|\mathcal{I}_s\| = O(h), \tag{57}$$

therefore $\Psi_0(\boldsymbol{\gamma})$ is a contraction on $\mathbb{R}^{sm}$ for h sufficiently small, and the Banach fixed-point theorem assures the existence of a unique solution $\bar{\boldsymbol{\gamma}}$ of (56) which can be determined as the limit of the sequence $\boldsymbol{\gamma}^{\ell+1} = \Psi_0(\boldsymbol{\gamma}^\ell)$, $\ell = 0, 1, \ldots$, starting from any $\boldsymbol{\gamma}_0 \in \mathbb{R}^{sm}$.

Theorem 6 *Under regularity assumptions on the functions f, and assuming either (52) or (54) holds true, so that $D(\alpha,\boldsymbol{\gamma})$ is bounded away from zero in a neighbourhood of $(0,\bar{\boldsymbol{\gamma}})$, system (55) admits a unique solution $(\alpha^*,\boldsymbol{\gamma}^*)$ which can be found as the limit of the sequence*

$$\begin{cases} \alpha^{\ell+1} = \dfrac{N(\alpha^\ell,\boldsymbol{\gamma}^\ell)}{D(\alpha^\ell,\boldsymbol{\gamma}^\ell)}, \\[3mm] \boldsymbol{\gamma}^{\ell+1} = \Psi(\alpha^\ell,\boldsymbol{\gamma}^\ell), \qquad \ell = 0, 1, \ldots, \end{cases} \tag{58}$$

starting at

$$(\alpha^0,\boldsymbol{\gamma}^0) = (0,\bar{\boldsymbol{\gamma}}).$$

Moreover the solution $(\alpha^*,\boldsymbol{\gamma}^*)$ satisfies

$$|\alpha^*| = O(h^{2s-2}), \qquad \|\boldsymbol{\gamma}^* - \bar{\boldsymbol{\gamma}}\| = O(h^{2s-1}). \tag{59}$$

<u>Proof</u> In accord with the contraction mapping theorem, we show that constants $h_0 > 0$ and $r > 0$ exist such that, for $h \in [0, h_0)$, the iteration function,

$$F(\alpha,\boldsymbol{\gamma}) := \begin{pmatrix} N(\alpha,\boldsymbol{\gamma})/D(\alpha,\boldsymbol{\gamma}) \\ \Psi(\alpha,\boldsymbol{\gamma}) \end{pmatrix} : \mathbb{R}^{sm+1} \to \mathbb{R}^{sm+1}, \tag{60}$$

defined at (58) satisfies the two properties:

  (a) $F(\alpha,\boldsymbol{\gamma})$ is a contraction in the closed ball $B((0,\bar{\boldsymbol{\gamma}}),r)$ of center $(0,\bar{\boldsymbol{\gamma}})$ and radius $r > 0$;

---

[9]We now emphasize, in the first equation in (55), the dependence of the numerator and denominator defining α from the given arguments, rather than from the stepsize h.

[10]Hereafter, $\|\cdot\|$ denotes the infinity norm.

14

(b) the sequence $(\alpha^\ell, \boldsymbol{\gamma}^\ell) \in B((0, \bar{\boldsymbol{\gamma}}), r)$, for all $\ell = 0, 1, \dots$.

Expanding $\Psi(\alpha, \boldsymbol{\gamma})$ with respect to the variable α in a neighbourhood of zero yields

$$\Psi(\alpha, \boldsymbol{\gamma}) = \Psi_0(\gamma) + R(\alpha, \gamma),$$

with

$$R(\alpha, \gamma) := -h\alpha\, \mathcal{P}_s^\top \Omega \otimes I_m\, f'(\mathbf{1} \otimes y_0 + h\mathcal{I}_s \otimes I_m \boldsymbol{\gamma})\, \mathcal{P}_s W_s \otimes I_m\, \boldsymbol{\gamma} \ + \ O((\alpha h)^2),$$

where f' stands for the Jacobian matrix of f. The function $R(\alpha, \boldsymbol{\gamma})$ satisfies $R(0, \boldsymbol{\gamma}) = 0$ and, since by virtue of Theorem 5 we may assume $|\alpha| \le c_1 h^{2s-2}$, we also get

$$\|R(\alpha, \gamma)\| \le c_2 h^{2s-1}, \tag{61}$$

with c_1 and c_2 positive constants independent of h, α, and $\boldsymbol{\gamma}$, in a closed ball $B((0, \bar{\boldsymbol{\gamma}}), r)$ of center $(0, \bar{\boldsymbol{\gamma}})$ and given radius $r > 0$. Consequently, in a neighbourhood of $(0, \bar{\boldsymbol{\gamma}})$, system (55) may be regarded as a perturbation of system (56), the perturbing term being the function $R(\alpha, \boldsymbol{\gamma})$.

From (61), it follows that $R(\alpha, \boldsymbol{\gamma})$ is Lipschitz continuous with constant $K_1 = O(h^{2s-1})$ and hence $\Psi(\alpha, \boldsymbol{\gamma})$ is Lipschitz continuous with constant $\mu \le K_0 + K_1 = O(h)$, where K_0 is defined in (57). This property, together with the Lipschitz continuity of $N(\alpha, \boldsymbol{\gamma})/D(\alpha, \boldsymbol{\gamma})$ with constant $O(h^{2s-2})$, assures that $h_0 > 0$ exists such that, for $h \in [0, h_0)$, the iteration function $F(\alpha, \boldsymbol{\gamma})$ defined at (60) is indeed a contraction in $B((0, \bar{\boldsymbol{\gamma}}), r)$, with constant μ.

To show that system (55) actually admits a solution $\boldsymbol{\gamma}^*$ which is $O(h^{2s-1})$-close to $\bar{\boldsymbol{\gamma}}$, we set

$$r := |N(0, \bar{\boldsymbol{\gamma}})/D(0, \bar{\boldsymbol{\gamma}})|/(1 - \mu) = O(h^{2s-2}),$$

and prove that the sequence $(\alpha^\ell, \boldsymbol{\gamma}^\ell)$ defined at (58) is entirely contained in the closed ball $B((0, \bar{\boldsymbol{\gamma}}), r)$. We proceed by induction. For $\ell = 1$ we get

$$\|\boldsymbol{\gamma}^1 - \bar{\boldsymbol{\gamma}}\| = \|\Psi(0, \boldsymbol{\gamma}^0) - \bar{\boldsymbol{\gamma}}\| = \|\Psi(0, \bar{\boldsymbol{\gamma}}) - \bar{\boldsymbol{\gamma}}\| = \|\Psi_0(\bar{\boldsymbol{\gamma}}) - \bar{\boldsymbol{\gamma}}\| = 0,$$

and

$$|\alpha^1 - \alpha^0| = |\alpha^1| = \left| \frac{N(0, \boldsymbol{\gamma}^0)}{D(0, \boldsymbol{\gamma}^0)} \right| = \left| \frac{N(0, \bar{\boldsymbol{\gamma}})}{D(0, \bar{\boldsymbol{\gamma}})} \right| = (1 - \mu)r.$$

Assuming now that $(\alpha^\ell, \boldsymbol{\gamma}^\ell) \in B((0, \bar{\boldsymbol{\gamma}}), r)$, we obtain

$$\|\boldsymbol{\gamma}^{\ell+1} - \bar{\boldsymbol{\gamma}}\| = \|\Psi(\alpha^\ell, \boldsymbol{\gamma}^\ell) - \Psi(0, \bar{\boldsymbol{\gamma}})\| \le \mu \max\{\|\boldsymbol{\gamma}^\ell - \bar{\boldsymbol{\gamma}}\|, |\alpha^\ell|\} \le \mu r \ < r,$$

and

$$
\begin{aligned}
|\alpha^{\ell+1} - \alpha^0| &\le |\alpha^{\ell+1} - \alpha^1| + |\alpha^1 - \alpha^0| = \left| \frac{N(\alpha^\ell, \boldsymbol{\gamma}^\ell)}{D(\alpha^\ell, \boldsymbol{\gamma}^\ell)} - \frac{N(\alpha^0 \boldsymbol{\gamma}^0)}{D(\alpha^0, \boldsymbol{\gamma}^0)} \right| + |\alpha^1| \\
&\le \mu \max\{\|\boldsymbol{\gamma}^\ell - \bar{\boldsymbol{\gamma}}\|, |\alpha^\ell|\} + |\alpha^1| \le \mu r + (1 - \mu)r = r.
\end{aligned}
$$

Since $(\alpha^*, \boldsymbol{\gamma}^*) \in B((0, \bar{\boldsymbol{\gamma}}), r)$, then clearly $|\alpha^*| \le r = O(h^{2s-2})$. Moreover, subtracting $\bar{\boldsymbol{\gamma}} = \Psi(0, \bar{\boldsymbol{\gamma}})$ from $\boldsymbol{\gamma}^* = \Psi(\alpha^*, \boldsymbol{\gamma}^*)$ we finally obtain

$$\|\boldsymbol{\gamma}^* - \bar{\boldsymbol{\gamma}}\| = \|\Psi(\alpha^*, \boldsymbol{\gamma}^*) - \Psi(0, \bar{\boldsymbol{\gamma}})\| \le \mu \max\{\|\boldsymbol{\gamma}^* - \bar{\boldsymbol{\gamma}}\|, |\alpha^*|\} = O(h^{2s-1}). \ \square$$

15

Order of convergence

Next, we show that method (15)–(16), with α given by (37) (or (39), in case the problem is Hamiltonian), retains the order $2s$ of the underlying s-stage Gauss method (i.e., the method corresponding to $\alpha = 0$).

Theorem 7 *Let $y_1 \approx y(h)$ be the approximation provided by the s-stage EQUIP method (15)–(16), with α given by (37) (respectively, (39) when the problem is Hamiltonian). Then, assuming that the hypotheses of Theorem 6 hold true,*

$$y_1 - y(h) = O(h^{2s+1}), \tag{62}$$

i.e., the method has order $2s$.

<u>Proof</u>   The proof will follow the novel approach defined in [17]. For this purpose, let us denote by $y(t, \theta, \xi)$ the solution of the problem

$$\dot{y} = f(y), \qquad t > \theta, \qquad y(\theta) = \xi.$$

Moreover, let $\Phi(t, \theta)$ be the fundamental matrix solution of the corresponding variational problem

$$\dot{\Phi}(t, \theta) = f'(y(t, \theta, \xi))\Phi(t, \theta), \qquad t > \theta, \qquad \Phi(\theta, \theta) = I_m.$$

We also recall that, according to Lemma 1,

$$\Gamma_j(h) := \int_0^1 P_j(c)\Phi(h, ch)\mathrm{d}c = O(h^j), \qquad j \geq 0. \tag{63}$$

By taking into account (20)–(34), let us then define the polynomial

$$u(ch) = y_0 + h \sum_{j=0}^{s-1} \int_0^c P_j(x)\mathrm{d}x\, \gamma_j \qquad \left[\quad \Rightarrow \quad \dot{u}(ch) = \sum_{j=0}^{s-1} P_j(c)\gamma_j \right], \qquad c \in [0, 1], \tag{64}$$

such that $u(0) = y_0$ and $u(h) = y_1$. Consqeuently,

$$
\begin{aligned}
y_1 - y(h) &= y(h, h, u(h)) - y(h, 0, u(0)) = \int_0^h \frac{\mathrm{d}}{\mathrm{d}t} y(h, t, u(t))\mathrm{d}t \\
&= \int_0^h \left[\frac{\partial}{\partial\theta} y(h, \theta, u(t)) \Big|_{\theta=t} + \frac{\partial}{\partial\xi} y(h, t, \xi) \Big|_{\xi=u(t)} \dot{u}(t) \right] \mathrm{d}t \\
&= \int_0^h \Phi(h, t) \left[-f(u(t)) + \dot{u}(t) \right] \mathrm{d}t = h \int_0^1 \Phi(h, ch) \left[-f(u(ch)) + \dot{u}(ch) \right] \mathrm{d}c =: (65)
\end{aligned}
$$

By considering that

$$f(u(ch)) = \sum_{j\geq 0} P_j(c)\gamma_j(u), \qquad \gamma_j(u) = \int_0^1 P_j(\tau)f(u(\tau h))\mathrm{d}\tau = O(h^j),$$

16

and taking into account (41), (63), and (64), one obtains:

$$
\begin{aligned}
(65) \quad &= \quad h \int_0^1 \Phi(h, ch) \left[\sum_{j=0}^{s-1} P_j(c)(\gamma_j(\sigma_1) - \gamma_j(u)) - \sum_{j \geq s} P_j(c)\gamma_j(u) + \sum_{j=0}^{s-1} P_j(c)(\gamma_j - \gamma_j(\sigma_1)) \right] dc \\
&= \quad h \left[\sum_{j=0}^{s-1} \Gamma_j(h)(\gamma_j(\sigma_1) - \gamma_j(u)) - \sum_{j \geq s} \Gamma_j(h)\gamma_j(u) + \sum_{j=0}^{s-1} \Gamma_j(h)(\gamma_j - \gamma_j(\sigma_1)) \right].
\end{aligned}
$$

Due to the fact that $\Gamma_j(h) = O(h^j)$, $\gamma_j(u) = O(h^j)$, $\gamma_j - \gamma_j(\sigma_1) = O(h^{2s-j})$, one then concludes that

$$
y_1 - y(h) \;=\; h \sum_{j=0}^{s-1} \Gamma_j(h)(\gamma_j(\sigma_1) - \gamma_j(u)) + O(h^{2s+1}). \tag{66}
$$

Consequently, (62) follows provided that

$$
\sum_{j=0}^{s-1} \Gamma_j(h)(\gamma_j(\sigma_1) - \gamma_j(u)) \;=\; O(h^{2s}). \tag{67}
$$

By taking into account (25), (64), and (53), one has

$$
\sigma_1(ch) - u(ch) = -h\alpha \sum_{i=0}^{s-1} \int_0^c P_i(x)dx \, (\phi_{2i}\gamma_0 - \phi_{1i}\gamma_1) = O(h^{2s-1}), \tag{68}
$$

so that:

$$
\begin{aligned}
\gamma_j(\sigma_1) - \gamma_j(u) \quad &= \quad \int_0^1 P_j(c) \left[f(\sigma_1(ch)) - f(u(ch)) \right] dc \\
&= \quad \int_0^1 P_j(c) f'(u(ch)) \left(\sigma_1(ch) - u(ch) \right) dc + O(h^{4s-2}) \\
&= \quad -h\alpha f'(y_0) \int_0^1 P_j(c) \sum_{i=0}^{s-1} \int_0^c P_i(x)dx \, (\phi_{2i}\gamma_0 - \phi_{1i}\gamma_1) dc \;+\; O(h^{2s}) \\
&= \quad -h\alpha f'(y_0) \int_0^1 g_j(c)dc \;+\; O(h^{2s}),
\end{aligned}
$$

having set

$$
g_j(c) := P_j(c) \sum_{i=0}^{s-1} \int_0^c P_i(x)dx \, (\phi_{2i}\gamma_0 - \phi_{1i}\gamma_1).
$$

Since g_j is a polynomial of degree $s + j$, for $j = 0, \ldots, s-1$, one has that

$$
\int_0^1 g_j(c)dc = \sum_{\ell=1}^{s} b_\ell g_j(c_\ell).
$$

Consequently, by setting

$$
\Gamma(h) := \big(\; \Gamma_0(h), \quad \ldots, \quad \Gamma_{s-1}(h) \; \big),
$$

17

and taking into account (8), (30), (24), (29), one has:

$$\sum_{j=0}^{s-1} \Gamma_j(h)(\gamma_j(\sigma_1) - \gamma_j(u))$$

$$= -h\alpha \sum_{j=0}^{s-1} \Gamma_j(h) f'(y_0) \sum_{\ell=1}^{s} b_\ell P_j(c_\ell) \sum_{i=0}^{s-1} \int_0^{c_\ell} P_i(x) \mathrm{d}x \, (\phi_{2i}\gamma_0 - \phi_{1i}\gamma_1) \, + \, O(h^{2s})$$

$$= -h\alpha \, \Gamma(h) \, (I_s \otimes f'(y_0)) \, (\mathcal{P}_s^\top \Omega \mathcal{I}_s \otimes I_m) \left[X_s^{-1} \boldsymbol{e}_2 \otimes \gamma_0 - X_s^{-1} \boldsymbol{e}_1 \otimes \gamma_1 \right] \, + \, O(h^{2s})$$

$$= -h\alpha \, \Gamma(h) \, (I_s \otimes f'(y_0)) \, (\mathcal{P}_s^\top \Omega \mathcal{P}_s \otimes I_m) \left[(\boldsymbol{e}_2 \boldsymbol{e}_1^\top - \boldsymbol{e}_1 \boldsymbol{e}_2^\top) \otimes I_m \right] \boldsymbol{\gamma} \, + \, O(h^{2s}) \quad =: \quad (69)$$

By considering that $\mathcal{P}_s^\top \Omega \mathcal{P}_s = I_s$, one then obtains:

$$(69) = h\alpha \left[\Gamma_0(h) f'(y_0)\gamma_1 - \Gamma_1(h) f'(y_0)\gamma_0 \right] + O(h^{2s}). \tag{70}$$

(67) then follows by taking into account that $\alpha = O(h^{2s-2})$, because of Theorem 6, and, moreover,

$$\gamma_1 = O(h) = \Gamma_1(h), \qquad \gamma_0 = O(1) = \Gamma_0(h). \; \square$$

4 Discretization of the involved integrals

We observe that the actual implementation of the s-stage EQUIP method (15)–(16) requires the evaluation of the integrals (36) and (38). Consequently, it is not yet an operative method since, quoting e.g., Dahlquist and Björk [27, p. 521], *as is well known, even many relatively simple integrals cannot be expressed in finite terms of elementary functions, and thus must be evaluated by numerical methods.* For this purpose, one may use a k-point Gauss-Legendre quadrature, and we denote EQUIP(k,s) the corresponding method, whose analysis is now completed, w.r.t. that done in the previous section. Hereafter, we shall assume $k \geq s$. Consequently, by setting

$$P_k(\hat{c}_i) = 0, \qquad \hat{b}_i = \int_0^1 \hat{\ell}_i(x)\mathrm{d}x, \qquad \hat{\ell}_i(x) = \prod_{j \neq i} \frac{x - \hat{c}_j}{\hat{c}_i - \hat{c}_j}, \qquad i = 1, \ldots, k, \tag{71}$$

the nodes and weights of the Gauss-Legendre formula of order $2k$, one approximates (36) and (38) by means of

$$\hat{\rho}_j(\sigma_1) = \sum_{\ell=1}^k \hat{b}_\ell P_j(\hat{c}_\ell) \nabla C(\sigma_1(\hat{c}_\ell h)), \quad j = 0, \ldots, s-1, \qquad \hat{\rho}(\sigma_2) = \sum_{\ell=1}^k \hat{b}_\ell \nabla C(\sigma_2(\hat{c}_\ell)), \tag{72}$$

and

$$\hat{\gamma}_j(\sigma_1) = \sum_{\ell=1}^k \hat{b}_\ell P_j(\hat{c}_\ell) f(\sigma_1(\hat{c}_\ell h)), \quad j = 0, \ldots, s-1, \qquad \hat{\gamma}(\sigma_2) = \sum_{\ell=1}^k \hat{b}_\ell f(\sigma_2(\hat{c}_\ell)), \tag{73}$$

respectively. It is quite straightforward to prove that, under regularity assumptions on f and C,

$$\begin{aligned}
\hat{\rho}_j(\sigma_1) &= \rho_j(\sigma_1) + O(h^{2k-j}), \\
\hat{\gamma}_j(\sigma_1) &= \gamma_j(\sigma_1) + O(h^{2k-j}), \qquad j = 0, \ldots, s-1, \\
\hat{\rho}(\sigma_2) &= \bar{\rho}(\sigma_2) + O(h^{2k}), \\
\hat{\gamma}(\sigma_2) &= \bar{\gamma}(\sigma_2) + O(h^{2k}).
\end{aligned} \tag{74}$$

18

As a result, the new approximation is formally still defined by the two polynomials σ_1 and σ_2 defined in (25) and (32)–(34). However, in such formulae, the parameter α, defined in (37) or (39), has to be replaced by an approximated one, say $\hat{\alpha}$, computed by using the quadratures (71)–(73). That is,

$$\hat{\alpha} = \frac{\sum_{j=0}^{s-1} \hat{\rho}_j(\sigma_1)^\top \gamma_j}{(\hat{\rho}_0(\sigma_1) - \hat{\rho}(\sigma_2))^\top (\phi_{20}\gamma_0 - \phi_{10}\gamma_1) + \sum_{j=1}^{s-1} \hat{\rho}_j(\sigma_1)^\top (\phi_{2j}\gamma_0 - \phi_{1j}\gamma_1)} \tag{75}$$

or

$$\hat{\alpha} = \frac{\sum_{j=0}^{s-1} \hat{\gamma}_j(\sigma_1)^\top J\gamma_j}{(\hat{\gamma}_0(\sigma_1) - \hat{\gamma}(\sigma_2))^\top J (\phi_{20}\gamma_0 - \phi_{10}\gamma_1) + \sum_{j=1}^{s-1} \hat{\gamma}_j(\sigma_1)^\top J (\phi_{2j}\gamma_0 - \phi_{1j}\gamma_1)}, \tag{76}$$

respectively. The following result can then be proved.

Lemma 9 *Let the parameter α be defined by either (75), when solving problem (1)–(2), or (76), when solving problem (13)–(14). Then, for all $k \geq s$ one has*

$$\hat{\alpha} = \alpha + O(h^{2(k-1)}) \qquad \Rightarrow \qquad \hat{\alpha} = O(h^{2(s-1)}), \tag{77}$$

with α given by (37) and (39), respectively. In particular, when the problem is Hamiltonian and $k = s$, one has $\hat{\alpha} = 0$, so that EQUIP(s, s) reduces to the s-stage Gauss-method.

Proof   The first part of the proof follows straightforwardly, by taking into account (74), from the line integral formulation used to derive (37) and (39), respectively. In particular, one finds that

$$\hat{\alpha} = \frac{\mathcal{N}(h) + O(h^{2k})}{\mathcal{D}(h) + O(h^{2k-s+1})} \equiv \alpha \cdot \left(1 + O(h^{2(k-s)})\right), \tag{78}$$

with $\mathcal{N}(h) = O(h^{2s})$ and $\mathcal{D}(h) = O(h^2)$ formally given by (43) and (50), in case of problem (1)–(2), or (44) and (51), in case of problem (13)–(14), respectively.

The fact that $\hat{\alpha} = 0$, when the problem is Hamiltonian and $k = s$, follows from (76), by considering that matrix J is skew-symmetric and (see (23) and (73)) $\hat{\gamma}_j(\sigma_1) \equiv \gamma_j$. $\square$

Moreover, the following results hold true, whose straightforward proofs, which strictly follow that of Theorem 2, are omitted for sake of brevity.

Theorem 8 *With reference to problem (1)–(2) and under regularity assumptions on both f and C, and for $k \geq s$, one obtains that the EQUIP(k, s) method defined by the parameter $\hat{\alpha}$ in (75) satisfies:*

$$C(y_1) - C(y_0) = \begin{cases} 0, & \text{if } C \text{ is a polynomial of degree not larger than } 2k/s, \\ O(h^{2k+1}), & \text{otherwise.} \end{cases}$$

Theorem 9 *With reference to problem (13)–(14) and under regularity assumptions on H, and for $k \geq s$, one obtains that the EQUIP(k, s) method defined by the parameter $\hat{\alpha}$ in (76) satisfies:*

$$H(y_1) - H(y_0) = \begin{cases} 0, & \text{if } H \text{ is a polynomial of degree not larger than } 2k/s, \\ O(h^{2k+1}), & \text{otherwise.} \end{cases}$$

Remark 3 *As a consequence of Theorems 8 and 9, one has that an* exact *conservation can be always gained, in case of a polynomial invariant, by taking k large enough. On the other hand, even in the general case a* practical *conservation can still be obtained, by taking k large enough so that the quadrature error is within round-off error level. It is worth mentioning that by considering higher values of k does not increase the complexity of the nonlinear problem to be solved, which amounts to compute the s coefficients of the polynomial σ_1, and the parameter α.*

The previous results allows us to prove that, for all $k \geq s$, the EQUIP(k, s) method still has order $2s$.

Theorem 10 *Let $y_1 \approx y(h)$ be the approximation provided by the EQUIP(k, s) method (15)–(16), with $\alpha = \hat{\alpha}$ given by (75) (respectively, (76) when the problem is Hamiltonian). Then, for all $k \geq s$, and assuming the hypotheses of Theorem 7 hold true,*

$$y_1 - y(h) = O(h^{2s+1}), \tag{79}$$

i.e., the method retains the order $2s$.

<u>Proof</u>   The proof strictly follows that done for Theorem 7, the only difference being that, starting from (68), α has to be formally replaced by $\hat{\alpha}$. Nevertheless, because of (77), the same arguments of that proof apply also in this case, so that (79) follows. □

Remark 4 *It is worth noticing that, by taking into account that $\hat{\alpha} = O(h^{2s-2})$, from (31) it follows that the stage order of EQUIP(k, s) methods is s, for all $k \geq s$. I.e., the methods also retain the same stage order as that of the underlying s-stage Gauss method.*

5   The nonlinear iteration

Next, let us duscuss the use of a fixed-point iteration for solving the discrete problem generated by the EQUIP(k, s) method, with $k \geq s$. In principle, this is given by (58), with the only difference that we have to consider $\hat{\alpha}$ in place of α. Therefore, by taking into account (77)–(78), convergence can be proved by using the same arguments in the proof of Theorem 6, provided that $k > s$, which we assume hereafter. There are, however, two improvements over the basic iteration (6), as explained below.

First of all, in order to improve the convergence properties of the iteration, we consider the following *Gauss-Seidel type* variant of (58),

$$\begin{cases} \hat{\alpha}^{\ell+1} = \dfrac{\hat{N}(\hat{\alpha}^\ell, \boldsymbol{\gamma}^\ell)}{\hat{D}(\hat{\alpha}^\ell, \boldsymbol{\gamma}^\ell)}, \\ \boldsymbol{\gamma}^{\ell+1} = \Psi(\hat{\alpha}^{\ell+1}, \boldsymbol{\gamma}^\ell), \qquad \ell = 0, 1, \ldots, \end{cases} \tag{80}$$

where we have formally used $\hat{N}$ and $\hat{D}$ in place of N and D, respectively, in order to take into account of the quadratures (see (71)–(76)).

Secondly, the quadratures involved in the computation of $\hat{\alpha}$ in place of α, potentially introduce a $O(h^{2k+1})$ error in the invariant at each step, according to Theorems 8 and 9. In order to avoid

20

their accumulation, it suffices to compute $\hat{\alpha}$ so that the variation of the invariant, at the current step, matches the opposite of the error until the previous step. This latter error, which we denote ΔC, can be easily computed by evaluating the invariant. As a result, following steps similar to those used in the proof of Theorem 2, one obtains that the iteration (80) becomes:

$$
\begin{cases}
\hat{\alpha}^{\ell+1} = \dfrac{\hat{N}(\hat{\alpha}^\ell, \boldsymbol{\gamma}^\ell) + \frac{\Delta C}{h}}{\hat{D}(\hat{\alpha}^\ell, \boldsymbol{\gamma}^\ell)}, \\
\boldsymbol{\gamma}^{\ell+1} = \Psi(\hat{\alpha}^{\ell+1}, \boldsymbol{\gamma}^\ell), \qquad \ell = 0, 1, \ldots .
\end{cases}
\tag{81}
$$

In so doing, the errors in the invariant remain uniformly $O(h^{2k+1})$, since their accumulation is prevented, and moreover, the magnitude of $\hat{\alpha}$ remains $O(h^{2(s-1)})$.

6 Numerical results

In this section we consider a few numerical tests concerning the numerical approximation of some conservative problems. In particular, two Hamiltonian problems and two Poisson problems. They are listed below.

1. The well-known Kepler problem [29, 10], which is Hamiltonian, i.e. in the form (13), with

$$
H(q, p) = \frac{1}{2}\|p\|_2^2 - \frac{1}{\|q\|_2}, \qquad q, p \in \mathbb{R}^2.
\tag{82}
$$

It admits also the additional quadratic invariant (i.e., the *angular momentum*),

$$
M(q, p) = q^\top J p, \qquad J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.
\tag{83}
$$

Its solution is periodic of period $T = 2\pi$, and resulting into an ellipse of eccentricity ε in the q-plane, when starting at

$$
q(0) = (1 - \varepsilon, 0)^\top, \qquad p(0) = \left(0, \sqrt{\frac{1 + \varepsilon}{1 - \varepsilon}} \right)^\top.
\tag{84}
$$

In particular, we shall consider the value $\varepsilon = 0.5$.

2. The pendulum problem [10], which is also Hamiltonian with

$$
H(q, p) = \frac{1}{2}p^2 - \cos(q).
\tag{85}
$$

Its solution is periodic of period $T \approx 28.57109480185544$, when starting at

$$
q(0) = 0, \qquad p(0) = 1.99999.
\tag{86}
$$

3. The Poisson problem [5]

$$
\dot{y} = B(y)\nabla H(y), \qquad B(y)^\top = -B(y),
\tag{87}
$$

with $y \in \mathbb{R}^3$ and

$$B(y) = \begin{pmatrix} 0 & c_3 y_3 & -c_2 y_2 \\ -c_3 y_3 & 0 & c_1 y_1 \\ c_2 y_2 & -c_1 y_1 & 0 \end{pmatrix}, \qquad H(y) = y_1^{12} + \frac{1}{2}\left[(y_2 - y_3)^2 + (y_1 - y_3)^2\right], \quad (88)$$

also admitting, as a further invariant besides H, the quadratic *Casimir*

$$C(y) = y^\top \begin{pmatrix} c_1 & & \\ & c_2 & \\ & & c_3 \end{pmatrix} y. \qquad (89)$$

The solution turns out to be periodic, of period $T \approx 0.53102669598427$, when choosing

$$c_1 = 1, \quad c_2 = 5, \quad c_3 = -4, \qquad \text{and} \qquad y(0) = (1,\, 1,\, 1)^\top. \qquad (90)$$

4. The *Lotka-Volterra* problem, which is in the form (87) with $y \in \mathbb{R}^2$ and

$$B(y) = \begin{pmatrix} 0 & y_1 y_2 \\ -y_1 y_2 & 0 \end{pmatrix}, \qquad H(y) = a \log y_1 - y_1 + b \log y_2 - y_2. \qquad (91)$$

The solution turns out to be periodic, of period $T \approx 7.720315563434113$, when choosing

$$a = 1, \qquad b = 2, \qquad \text{and} \qquad y(0) = (0.1,\, 0.1)^\top. \qquad (92)$$

For such problems, we compare the EQUIP(k, s) methods with the corresponding s-stage Gauss collocation methods, $s = 2, 3$, using a given constant stepsize of the form $h = T/n$, with T the period of the solution and increasing values of n. This allows us to easily asses the accuracy of the solution, by checking it at the end of each period. As a measure of the error on the Hamiltonian, by setting H_i its numerical value at the ith integration step, we consider the mean square value of the error over N steps,

$$e_H := \sqrt{\frac{1}{N} \sum_{i=1}^{N} (H_i - H_0)^2}.$$

A similar measure will be used for the invariants (83), e_M, and (89), e_C. When needed, by setting α_i the computed parameter (75)–(76) for the EQUIP method, we set

$$\bar{\alpha} := \sqrt{\frac{1}{N} \sum_{i=1}^{N} \alpha_i^2},$$

as a measure of its magnitude. The EQUIP methods are implemented via the iteration (81), whereas only the second equation, having fixed $\hat{\alpha} = 0$, is used for the Gauss methods. For the EQUIP(k, s) methods, the value of k is chosen in order to obtain a practical energy-conservation, as soon as the stepsize is decreased, for the considered values of s: in the present case, $k = 6$ is sufficient for our purposes.

We start considering the Kepler problem (82)–(84). In Table 1 we list the obtained results by solving the problem with the EQUIP(6,2) and the 2-stage Gauss (GAUSS2, hereafter) methods. From such results, one deduces that:

- both methods are 4-th order, with EQUIP(6,2) one order of magnitude more accurate than GAUSS2, and preserve the quadratic invariant M;

- for EQUP(6,2), $\bar{\alpha}$ turns out to be $O(h^2)$ as expected and, moreover, the error on H soon falls within round-off error level, whereas it decreases with order 4 for GAUSS2;

- the number of iterations per step (iter/step) is essentially the same for both methods, and it decreases (as is expected) with the stepsize h.

Similarly, in Table 2 we list the obtained results by solving the problem with the EQUIP(6,3) and the 3-stage Gauss (GAUSS3, hereafter) methods, showing that:

- both methods are 6-th order, with EQUIP(6,3) one order of magnitude more accurate than GAUSS3, and preserve the quadratic invariant M;

- for EQUP(6,3), $\bar{\alpha}$ turns out to be $O(h^4)$ as expected and, moreover, the error on H soon falls within round-off error level, whereas it decreases with order 6 for GAUSS3;

- the number of iterations per step (iter/step) is essentially the same for both methods, and it decreases (as is expected) with the stepsize h.

Table 1: Numerical solution of the Kepler problem (82)–(84) by using the EQUIP(6,2) and the 2-stage Gauss (GAUSS2) methods with stepsize $h = 2\pi/n$ over 10 periods.

	EQUIP(6,2)							GAUSS2					
n	error	rate	H-error	M-error	$\bar{\alpha}$	rate	iter/ step	error	rate	H-error	rate	M-error	iter/ step
20	1.34e-1	−	1.64e-09	3.66e-15	1.51e-3	−	19.6	1.55e0	−	1.95e-3	−	5.75e-15	17.4
30	2.61e-2	4.0	6.10-12	7.88e-15	6.81e-4	2.0	15.6	2.37e-1	4.6	2.27e-4	5.3	1.15e-14	14.2
40	8.36e-3	4.0	1.86e-13	4.22e-16	3.84e-4	2.0	13.6	8.00e-2	3.8	7.65e-5	3.8	1.24e-15	12.8
50	3.45e-3	4.0	1.84e-14	1.58e-15	2.45e-4	2.0	12.5	3.41e-2	3.8	3.28e-5	3.8	2.39e-15	11.7
60	1.67e-3	4.0	2.05e-15	4.70e-16	1.70e-4	2.0	11.8	1.68e-2	3.9	1.61e-5	3.9	3.16e-15	11.3
70	9.01e-4	4.0	1.44e-15	4.28e-16	1.25e-4	2.0	11.4	9.17e-3	3.9	8.83e-6	3.9	4.10e-15	10.6
80	5.29e-4	4.0	1.11e-15	2.46e-15	9.58e-5	2.0	10.8	5.41e-3	3.9	5.22e-6	3.9	1.36e-15	10.3
90	3.31e-4	4.0	2.44e-15	8.11e-16	7.57e-5	2.0	10.5	3.40e-3	4.0	3.27e-6	4.0	3.33e-15	10.0
100	2.18e-4	4.0	9.77e-16	2.98e-15	6.13e-5	2.0	10.2	2.24e-3	4.0	2.16e-6	4.0	6.65e-15	9.7

Table 2: Numerical solution of the Kepler problem (82)–(84) by using the EQUIP(6,3) and the 3-stage Gauss (GAUSS3) methods with stepsize $h = 2\pi/n$ over 10 periods.

	EQUIP(6,3)							GAUSS3					
n	error	rate	H-error	M-error	$\bar{\alpha}$	rate	iter/ step	error	rate	H-error	rate	M-error	iter/ step
20	2.67e-3	−	1.15e-09	5.44e-15	4.62e-5	−	15.3	5.16e-2	−	6.72e-5	−	1.55e-15	15.4
30	3.11e-4	5.3	1.68e-11	4.43e-15	1.17e-5	3.4	13.1	7.41e-3	4.8	8.44-6	5.1	2.17e-15	13.1
40	5.63e-5	6.0	4.61e-13	2.00e-15	3.81e-6	3.9	11.9	1.22e-3	6.3	1.38e-6	6.3	4.08e-15	11.9
50	1.47e-5	6.0	2.38e-14	9.81e-16	1.55e-6	4.0	11.3	3.09e-4	6.2	3.48e-7	6.2	8.59e-16	11.3
60	4.94e-6	6.0	2.01e-15	7.91e-16	7.47e-7	4.0	10.5	1.02e-4	6.1	1.15e-7	6.1	9.51e-16	10.5
70	1.96e-6	6.0	7.58e-16	5.67e-16	4.02e-7	4.0	10.1	4.01e-5	6.0	4.51e-8	6.1	6.47e-16	10.1
80	8.78e-7	6.0	3.53e-16	8.32e-16	2.35e-7	4.0	9.7	1.79e-5	6.0	2.01e-8	6.0	9.04e-16	9.7
90	4.33e-7	6.0	3.92e-16	1.13e-15	1.47e-7	4.0	9.3	8.82e-6	6.0	9.90e-9	6.0	2.22e-15	9.3
100	2.30e-7	6.0	2.46e-16	4.30e-16	9.62e-8	4.0	9.1	4.68e-6	6.0	5.25e-9	6.0	6.18e-16	9.1

Next, we briefly study the numerical solution of the pendulum problem (85)–(86) by means of EQUIP(k,s) and the corresponding Gauss method. In particular:

23

Table 3: Numerical solution of the pendulum problem (85)–(86) by using the EQUIP(6,2) and the 2-stage Gauss (GAUSS2) methods with stepsize $h = T/n$ over 10 periods.

| | EQUIP(6,2) | | | | | | GAUSS2 | | | | |
n	error	rate	H-error	$\bar{\alpha}$	rate	iter/ step	error	rate	H-error	rate	iter/ step
50	4.54e-1	–	4.71e-06	4.55e-3	–	40.6	1.48e2	–	2.61e-4	–	17.5
60	2.23e-1	3.9	1.38e-11	3.22e-3	1.9	34.4	1.35e2	0.5	1.05e-4	5.0	16.1
70	1.22e-1	3.9	7.47e-12	2.40e-3	1.9	32.5	1.23e2	0.6	5.16e-5	4.6	15.0
80	7.22e-2	3.9	6.78e-09	1.84e-3	2.0	30.4	2.86e0	28	2.72e-5	4.8	14.0
90	4.54e-2	3.9	9.33e-13	1.50e-3	1.8	27.8	3.63e0	-2.0	1.72e-5	3.9	13.4
100	3.01e-2	3.9	4.73e-13	1.22e-3	2.0	25.2	3.72e0	-0.2	1.12e-5	4.1	12.9
110	2.09e-2	3.8	2.46e-13	1.01e-3	1.9	23.5	3.70e0	0.0	7.53e-6	4.2	12.4
120	1.52e-2	3.7	1.03e-13	8.71e-4	1.8	23.3	3.58e0	0.4	5.25e-6	4.1	12.0
130	1.11e-2	3.9	5.51e-14	7.48e-4	1.9	22.0	3.29e0	1.1	3.78e-6	4.1	11.6
140	8.36e-3	3.9	3.34e-14	6.50e-4	1.9	20.5	2.86e0	1.9	2.79e-6	4.1	11.4
150	6.31e-3	4.1	2.49e-14	5.65e-4	2.0	20.5	2.37e0	2.7	2.10e-6	4.1	11.2

Table 4: Numerical solution of the pendulum problem (85)–(86) by using the EQUIP(6,3) and the 3-stage Gauss (GAUSS3) methods with stepsize $h = T/n$ over 10 periods.

| | EQUIP(6,3) | | | | | | GAUSS3 | | | | |
n	error	rate	H-error	$\bar{\alpha}$	rate	iter/ step	error	rate	H-error	rate	iter/ step
50	7.67e-3	–	9.20e-12	2.0972e-4	–	16.4	3.72e0	–	1.01e-5	–	15.8
60	1.80e-3	7.9	6.04e-12	6.6569e-5	6.3	14.7	3.36e0	0.6	2.86e-6	6.9	14.6
70	5.69e-4	7.5	3.75e-12	2.7582e-5	5.7	13.9	1.97e0	3.5	1.07e-6	6.4	13.8
80	2.16e-4	7.2	2.30e-12	1.3510e-5	5.3	13.1	9.29e-1	5.6	4.65e-7	6.2	13.2
90	1.28e-4	4.5	1.13e-12	9.2616e-6	3.2	12.6	4.56e-1	6.0	2.24e-7	6.2	12.6
100	6.19e-5	6.9	8.93e-13	5.6190e-6	4.7	12.1	2.40e-1	6.1	1.18e-7	6.1	12.3
110	3.15e-5	7.1	5.72e-13	3.5084e-6	4.9	11.8	1.34e-1	6.1	6.57e-8	6.1	11.8
120	1.72e-5	7.0	3.74e-13	2.3050e-6	4.8	11.4	7.92e-2	6.1	3.87e-8	6.1	11.5
130	9.83e-6	7.0	2.50e-13	1.5774e-6	4.7	11.1	4.88e-2	6.1	2.38e-8	6.1	11.2
140	5.88e-6	6.9	1.70e-13	1.1162e-6	4.7	10.9	3.11e-2	6.1	1.52e-8	6.1	10.9
150	3.65e-6	6.9	1.22e-13	8.1196e-7	4.6	10.6	2.05e-2	6.0	1.00e-8	6.1	10.6

- in Table 3 we list the obtained results for the EQUIP(6,2) and GAUSS2 methods, showing that, even though GAUSS2 is able to conserve the Hamiltonian with the prescribed order, nonetheless, the EQUIP(6,2) method is much more accurate in conserving the Hamiltonian, and this reflects into a much more accurate numerical solution;

- in Table 4 we list the obtained results for the EQUIP(6,3) and GAUSS3 methods, showing that also in this case, the energy-conserving EQUIP(6,3) is much more reliable than the GAUSS3 method.

We now consider the numerical solution of the Poisson problem (87)–(90) by means of the EQUIP(6,2) and GAUSS2 methods, and the EQUIP(6,3) and GAUSS3 methods, over 50 periods, by using a stepsize $h = T/100$. In Figure 1 are the computed results for the 4th-order (left-plot) and 6th-order (right-plot) methods, all of them exactly conserving the quadratic Casimir (89). As one may see, the EQUIP methods exhibit a linear error growth, whereas the Gauss methods show a quadratic error growth.

A similar result is obtained when numerically solving the Lotka-Volterra problem (91)–(92) by means of the EQUIP(6,2) and GAUSS2 methods, and the EQUIP(6,3) and GAUSS3 methods, over 50 periods, by using a stepsize $h = T/100$. In Figure 2 are the computed results for the 4th-order

(left-plot) and 6th-order (right-plot) methods. As one may see, also in this case the EQUIP methods exhibit a linear error growth, whereas the Gauss methods show a quadratic error growth.

7   Conclusions

In this paper we reported a novel analysis for EQUIP(k, s) methods, a symplectic variant of the s-stage Gauss collocation methods with energy-preserving properties. Such methods, which retain the order $2s$ of the underlying Gauss method, were originally devised for Hamiltonian problems and are here extended to cope with general conservative problems. The practical implementation of the methods has been another major focus of the paper, ranging from the discretization of the involved integrals, by means of suitable high-order Gaussian rules, to the solution of the generated discrete problem. In particular, the fixed-point iteration used for the analysis of the methods has been improved, in order to maintain a uniform error level in the invariant. A few numerical tests, on a couple of Hamiltonian and Poisson problems, confirm the theoretical findings and show that such methods can be far more reliable than the original Gauss formulae, due to their better conservation properties.

As a future direction of investigation we mention that, upon devising a Newton-type iteration more efficient than (80), EQUIP(k, s) methods could be conveniently applied also for numerically solving Hamiltonian PDEs, after a proper space semi-discretization. In such a case, in fact, they will conserve possible quadratic invariants of the solution, besides the (semi)-discrete Hamiltonian (see, e.g., [2]).

References

[1] P. Amodio, L. Brugnano, F. Iavernaro. Energy-conserving methods for Hamiltonian Boundary Value Problems and applications in astrodynamics. *Adv. Comput. Math.* **41** (2015) 881–905.

[2] L. Barletti, L. Brugnano, G. Frasca Caccia, F. Iavernaro. Energy-conserving methods for the nonlinear Schrödinger equation. *Applied Mathematics and Computation* **318** (2018) 3–18.

[3] G. Benettin, A. Giorgilli. On the Hamiltonian interpolation of near-to-the-identity symplectic mappings with application to symplectic integration algorithms. *J. Statist. Phys.* **74** (1994) 1117–1143.

[4] S. Blanes, F. Casas. *A Concise Introduction to Geometric Numerical Integration.* CRC Press, Boca Raton, FL, 2016.

[5] L. Brugnano, M. Calvo, J.I. Montijano, L. Rández. Energy-preserving methods for Poisson systems. *J. Comput. Appl. Math.* **236** (2012) 16, 3890–3904.

[6] L. Brugnano, G. Frasca Caccia, F. Iavernaro. Efficient implementation of Gauss collocation and Hamiltonian Boundary Value Methods. *Numer. Algorithms* **65** (2014) 633–650.

[7] L. Brugnano, G. Frasca Caccia, F. Iavernaro. Energy conservation issues in the numerical solution of the semilinear wave equation. *Appl. Math. Comput.* **270** (2015) 842–870.

[8] L. Brugnano, G. Frasca Caccia, F. Iavernaro. Energy and QUadratic Invariants Preserving (EQUIP) methods for Hamiltonian Systems. *AIP Conference Proc.* **1738** (2016) 100002.

[9] L. Brugnano, F. Iavernaro. Line integral methods which preserve all invariants of conservative problems. *J. Comput. Appl. Math.* **236** (2012) 3905–3919.
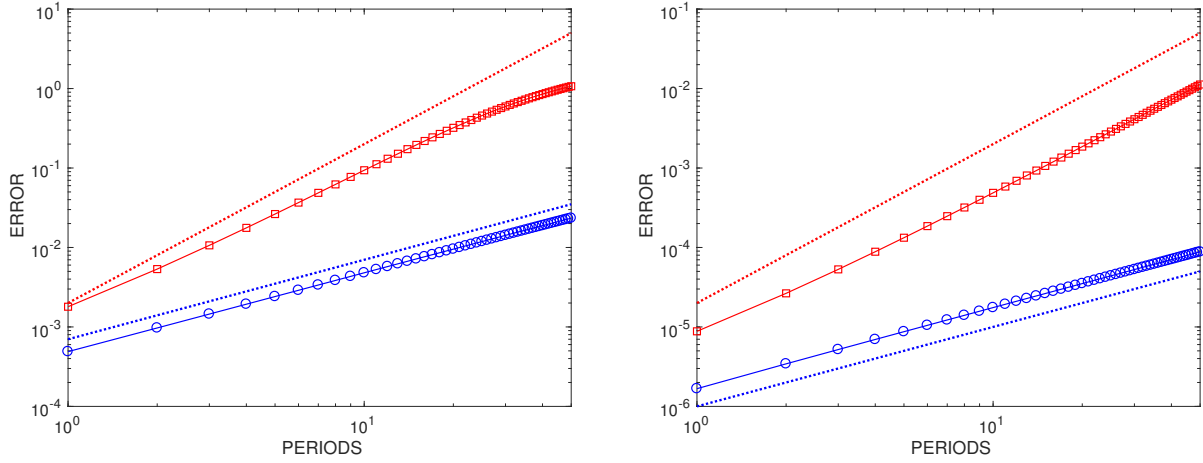
Figure 1: Poisson problem (87)–(90). Left-plot: solution error along 50 periods for EQUIP(6,2) (circles and solid line) and GAUSS2 (squares and solid line). Right-plot: solution error along 50 periods for EQUIP(6,3) (circles and solid line) and GAUSS2 (squares and solid line). In both plots, the linear and quadratic grows are reported in dotted-line.
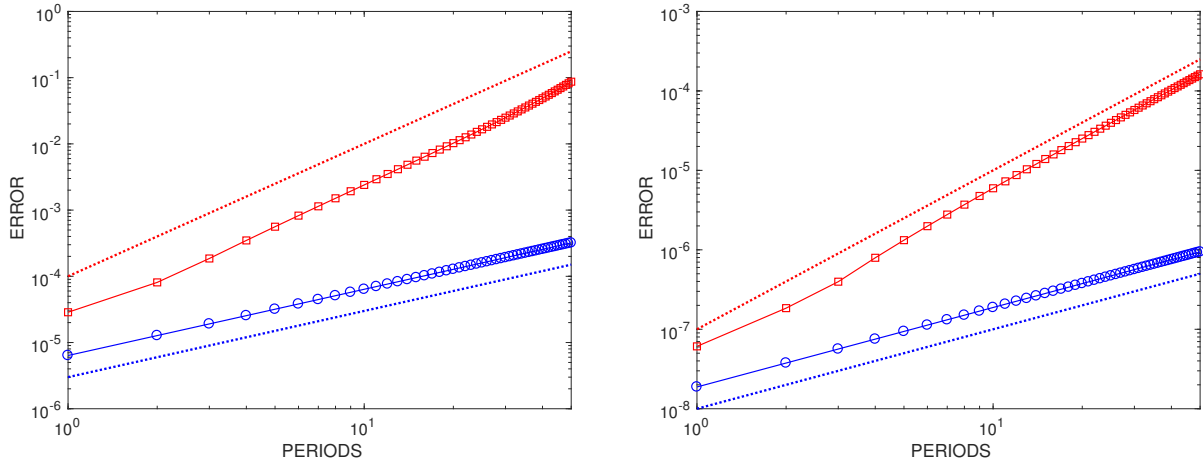

Figure 2: Lotka-Volterra problem (91)–(92). Left-plot: solution error along 50 periods for EQUIP(6,2) (circles and solid line) and GAUSS2 (squares and solid line). Right-plot: solution error along 50 periods for EQUIP(6,3) (circles and solid line) and GAUSS2 (squares and solid line). In both plots, the linear and quadratic grows are reported in dotted-line.

[10] L. Brugnano, F. Iavernaro. *Line Integral Methods for Conservative Problems*. CRC Press, Boca Raton, FL, 2016.

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

[12] L. Brugnano, F. Iavernaro, T. Susca. Numerical comparisons between Gauss-Legendre methods and Hamiltonian BVMs defined over Gauss points. *Proceedings of the MCalv6(5) Conference, 95–112*, Monogr. Real Acad. Ci. Exact. Fs.-Qum. Nat. Zaragoza, 33, *Acad. Cienc. Exact. Fs. Qum. Nat. Zaragoza, Zaragoza*, 2010.

[13] L. Brugnano, F. Iavernaro, D. Trigiante. Hamiltonian Boundary Value Methods (energy preserving discrete line methods). *J. Numer. Anal. Ind. Appl. Math.* **5**(1–2) (2010) 17–37.

[14] L. Brugnano, F. Iavernaro, D. Trigiante. Energy and quadratic invariants preserving integrators of Gaussian type. *AIP Conf. Proc.* **1281** (2010) 227–230.

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

[16] 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.* **218** (2012) 8053–8063.

[17] L. Brugnano, F. Iavernaro, D. Trigiante. A simple framework for the derivation and analysis of effective one-step methods for ODEs. *Appl. Math. Comput.* **218** (2012) 8475–8485.

[18] L. Brugnano, F. Iavernaro, D. Trigiante. Energy and quadratic invariants preserving integrators based upon Gauss collocation formulae. *SIAM J. Numer. Anal.* **50**(6) (2012) 2897–2916.

[19] L. Brugnano, F. Iavernaro, D. Trigiante. A two-step, fourth-order method with energy preserving properties. *Comput. Phys. Commun.* **183** (2012) 1860–1868.

[20] L. Brugnano, F. Iavernaro, D. Trigiante. Analysis of Hamiltonian Boundary Value Methods (EQUIPs): a class of energy-preserving Runge–Kutta methods for the numerical solution of polynomial Hamiltonian systems. *Commun. Nonlin. Sci. Numer. Simul.* **20** (2015) 650–667.

[21] L. Brugnano, Y. Sun. Multiple invariants conserving Runge-Kutta type methods for Hamiltonian problems. *Numer. Algorithms* **65** (2014) 611–632.

[22] K. Burrage, K. Burrage. Low rank Runge–Kutta methods, symplecticity and stochastic Hamiltonian problems with additive noise. *J. Comput. Appl. Math.* **236** (2012) 3920–3930.

[23] K. Burrage, J.C. Butcher. Stability criteria for implicit Runge-Kutta methods. *SIAM J. Numer. Anal.* **16** (1979) 46–57.

[24] J.C. Butcher. Implicit Runge-Kutta processes. *Math. Comp.* **18** (1964) 50–64.

[25] 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** (2006) 868–885.

[26] P. Chartier, E. Faou, A. Murua. An algebraic approach to invariant preserving integrators: the case of quadratic and Hamiltonian invariants. *Numer. Math.* **103** (2006) 575–590.

[27] G. Dahlquist, A. Björk. *Numerical Methods in Scientific Computing, Volume I*. SIAM, Philadelphia, 2008.

[28] Z. Ge, J.E. Marsden. Lie-Poisson Hamilton-Jacobi theory and Lie-Poisson integrators. *Phys. Lett. A* **133** (1988) 134–139.

[29] E. Hairer, C. Lubich, G. Wanner. *Geometric Numerical Integration, 2nd ed.*, Springer, Berlin, 2006.

[30] E. Hairer, G. Wanner. Algebraically stable and implementable Runge-Kutta methods of high horder. *SIAM J. Numer. Anal.* **18** (1981) 1098–1108.

[31] E. Hairer, G. Wanner. *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems, 2nd ed.*, Springer Ser. Comput. Math. **14**, Springer, Berlin, 1996.

[32] F. Iavernaro, B. Pace. *s*-Stage Trapezoidal Methods for the Conservation of Hamiltonian Functions of Polynomial Type. *AIP Conf. Proc.* **936** (2007) 603–606.

[33] F. Iavernaro, B. Pace. Conservative Block-Boundary Value Methods for the Solution of Polynomial Hamiltonian Systems. *AIP Conf. Proc.* **1048** (2008) 888–891.

[34] 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.

[35] H. Kojima. Invariants preserving schemes based on explicit Runge-Kutta methods. *BIT* **56**, No. 4 (2016) 1317–1337.

[36] F.M. Lasagni. Canonical Runge-Kutta methods. *Z. Angew. Math. Phys.* **39** (1988) 952–953.

[37] B. Leimkulher, S. Reich. *Simulating Hamiltonian Dynamics.* Cambridge University Press, Cambridge, 2004.

[38] J.M. Sanz-Serna. Runge-Kutta schemes for Hamiltonian systems. *BIT* **28** (1988) 877–883.

[39] J.M. Sanz-Serna. Symplectic Runge-Kutta schemes for adjoint equations, automatic differentiation, optimal control, and more. *SIAM Rev.* **58** (2016) 3–33.

[40] J.M. Sanz-Serna, M.P. Calvo. *Numerical Hamiltonian Problems.* Chapman & Hall, London, 1994.

[41] J.L. Yan, Z.Y. Zhang. New energy-preserving schemes using Hamiltonian Boundary Value and Fourier pseudospectral methods for the numerical solu- tion of the "good" Boussinesq equation. *Computer Physics Communications* **201** (2016) 33–42.