Commun Nonlinear Sci Numer Simulat 20 (2015) 650–667

Contents lists available at ScienceDirect

Commun Nonlinear Sci Numer Simulat

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

Analysis of Hamiltonian Boundary Value Methods (HBVMs): A class of energy-preserving Runge–Kutta methods for the numerical solution of polynomial Hamiltonian systems

Luigi Brugnano [a,*], Felice Iavernaro [b], Donato Trigiante [1]

[a] *Dipartimento di Matematica e Informatica "U. Dini", Università di Firenze, Italy*
[b] *Dipartimento di Matematica, Università di Bari, Italy*

ARTICLE INFO

Article history:
Available online 20 June 2014

Keywords:
Collocation methods
Energy-preserving methods
Runge–Kutta methods

ABSTRACT

One main issue, when numerically integrating autonomous Hamiltonian systems, is the long-term conservation of some of its invariants; among them the Hamiltonian function itself. For example, it is well known that classical symplectic methods can only exactly preserve, at most, quadratic Hamiltonians. In this paper, we report the theoretical foundations which have led to the definition of the new family of methods, called *Hamiltonian Boundary Value Methods* (*HBVMs*). HBVMs are able to *exactly* preserve, in the discrete solution, Hamiltonian functions of polynomial type of arbitrarily high degree. These methods turn out to be symmetric and can have arbitrarily high order. A few numerical tests confirm the theoretical results.

1. Foreword

The numerical solution of Hamiltonian problems is a relevant issue of investigation since many years: we refer to the recent monographs [1,2] for a comprehensive description of this topic, and to the references therein.

In a certain sense, the use of a numerical method acts as introducing a small perturbation in the original system which, in general, destroys all of its first integrals. The study of the preservation of invariant tori in the phase space of nearly integrable Hamiltonian systems has been a central theme in the research since the pioneering work of Poincaré, the final goal being to asses the stability of the solar system. From a numerical point of view, results in this respect are still poor, and this is justified by considering the delicacy of the problem: as testified by KAM theory, even small Hamiltonian perturbations of completely integrable systems do not prevent the disappearance of most of the tori, unless a Diophantine condition on the frequencies of the unperturbed system is satisfied.

At the times when research on this topic was started, there were no available numerical methods possessing such conservation features. A main approach to the problem was the devising of symplectic methods. However, though the numerical solution generated by symplectic (and/or symmetric) methods shows some interesting long-time behavior (see, for example, [1, Theorems X.3.1 and XI.3.1]), it was observed that symplecticity alone can only assure, at most, the conservation of quadratic Hamiltonian functions, unless they are coupled with some projection procedure (see, e.g., [1,20, Section IV.4]). In the general case, conservation cannot be assured, even though a quasi-preservation can be expected for reversible problems,

[*] Corresponding author.
E-mail addresses: luigi.brugnano@unifi.it (L. Brugnano), felice.iavernaro@uniba.it (F. Iavernaro).
[1] Deceased author.

http://dx.doi.org/10.1016/j.cnsns.2014.05.030
1007-5704/© 2014 Elsevier B.V. All rights reserved.

when symmetric methods are used. On the other hand, a numerical "drift" can be sometimes observed in the discrete solution [3,4]. One of the first successful attempts to solve the problem of loss of conservation of the Hamiltonian function by the numerical solution, is represented by *discrete gradient methods* (see [5] and references therein). Purely algebraic approaches have been also introduced (see, e.g., [6]), without presenting any energy-preserving method.

A further approach was considered in [7], where the *averaged vector field method* was proposed and shown to conserve the energy function of canonical Hamiltonian systems. As was recently outlined (see [8]), approximating the integral appearing in such method by means of a quadrature formula (based upon polynomial interpolation) yields a family of second order Runge–Kutta methods. These latter methods represent an instance of energy-preserving Runge–Kutta methods for polynomial Hamiltonian problems: their first appearance may be found in [9], under the name of *s-stage trapezoidal methods*. Additional examples of fourth and sixth-order Runge–Kutta methods were presented in [10,11].

In [9–11], the derivation of such energy-preserving Runge–Kutta formulae relies on the definition of the so called "discrete line integral", first introduced in [12]. However, a comprehensive analysis of such methods has not been carried out so far, so that their properties were not known and, moreover, their practical construction was difficult.

This was the situation when the results in the unpublished work [17] were obtained. Later on, there has been a flourishing of new results on energy-preserving methods, which we do not mention here. One of the aims of the present paper is to give an account about the theoretical foundations of the class of energy-preserving Runge–Kutta methods, named *Hamiltonian Boundary Value Methods (HBVMs)*. Even though they were derived in 2009, the results reported here have remained unpublished, so far. The first two authors agreed to publish them, in memory of the third author (passed away on September 18, 2011) on the occasion of the second death anniversary. The remaining part of the paper is essentially unchanged, with respect to the original version, including the references (apart [17,18], the latter being the proceedings of the conference where the methods were presented, and [19], introduced to answer to one of the referees), though the arguments have been slightly rearranged, to improve clarity.

2. Introduction

In this paper we derive and analyse symmetric methods, of arbitrarily high order, able to preserve Hamiltonian functions of polynomial type, of any specified degree. Such methods are named *Hamiltonian Boundary Value Methods (HBVMs)*, since the above approach has been at first studied in the framework of *block Boundary Value Methods* (see, e.g., [10,11]). The latter are block one-step methods [13]. However, the equivalent Runge–Kutta formulation of HBVMs will be here also considered. Before that, we need to introduce the background information concerning the approach. Let then

$$y' = J\nabla H(y), \quad y(0) = y_0 \in \mathbb{R}^{2m}, \tag{1}$$

be a Hamiltonian problem in canonical form, where, by setting as usual I_m the identity matrix of dimension m,

$$J = \begin{pmatrix} & I_m \\ -I_m & \end{pmatrix} \tag{2}$$

and where the Hamiltonian function, $H(y)$, is hereafter assumed to be a polynomial of degree ν. It is well known that, for any $y^* \in \mathbb{R}^{2m}$,

$$H(y^*) - H(y_0) = \int_{y_0 \to y^*} \nabla H(y)^T dy = \int_0^1 \sigma'(t)^T \nabla H(\sigma(t))dt, \tag{3}$$

where $\sigma : [0,1] \to R^{2m}$ is any smooth function such that

$$\sigma(0) = y_0, \quad \sigma(1) = y^*.$$

In particular, over a trajectory, $y(t)$, of (1), one has

$$H(y(t)) - H(y(0)) = \int_0^t \nabla H(y(\tau))^T y'(\tau)d\tau = \int_0^t \nabla H(y(\tau))^T J \nabla H(y(\tau))d\tau = 0,$$

due to the fact that matrix J in (2) is skew-symmetric.

Here we consider the case where $\sigma(t)$ is a polynomial of degree s yielding an approximation to the true solution $y(t)$ in the time interval $[0,h]$ which, without loss of generality, is hereafter normalized to $[0,1]$. More specifically, given the $s+1$ abscissae

$$0 = c_0 < c_1 < \cdots < c_s = 1 \tag{4}$$

and the approximations $y_i \approx y(c_i), \sigma(t)$ is meant to be defined by the interpolation conditions

$$\sigma(c_i) = y_i, \quad i = 0, \ldots, s. \tag{5}$$

Actually, the approximations $\{y_i\}$ will be unknown, until the new methods will be fully derived.

A different, though related concept, is that of collocating polynomial for the problem, at the abscissae (4). It is the unique polynomial $u(t)$, of degree $s+1$, satisfying

$$u(0) = y_0, \quad \text{and} \quad u'(c_i) = J\nabla H(u(c_i)), \quad i = 0, \ldots, s. \tag{6}$$

It is well known that (6) define a Runge–Kutta *collocation method*. Moreover, the set of abscissae (4) defines a corresponding quadrature formula with weights

$$b_i = \int_0^1 \prod_{j=0, j\neq i}^{s} \frac{t-c_j}{c_i-c_j} dt, \quad i = 0, 1, \ldots, s, \tag{7}$$

which has degree of precision ranging from s to $2s-1$, depending on the choice of the abscissae (4). In particular, the highest degree of precision is obtained by using the Lobatto abscissae, which we shall consider in the sequel.[2] The underlying collocation method has, then, order $2s$.

Remark 1. Choosing a Gauss distribution of the abscissae $\{c_i\}$ raises the degree of precision of the related quadrature formula to $2s+1$. In such a case, it is interesting to observe that applying (3) along the trajectory $u(t)$ and exploiting the collocation conditions (6), one gets

$$H(u(1)) - H(u(0)) = \int_0^1 u'(t)^T \nabla H(u(t)) dt = \sum_{i=0}^{s} b_i u'(c_i)^T \nabla H(u(c_i)) + R_s = R_s, \tag{8}$$

where R_s is the error in the approximation of the line integral. Therefore, $H(u(1)) = H(u(0))$ if and only if $R_s = 0$, which is implied by assuming that the quadrature formula with abscissae $\{c_i\}$ and weights $\{b_i\}$ is exact when applied to the integrand $u'(t)^T \nabla H(u(t))$. However, since the integrand has degree

$$s + (\nu - 1)(s+1) = \nu(s+1) - 1,$$

it follows that the maximum allowed value for ν is 2. Indeed, it is well known that quadratic invariants are preserved by symmetric collocation methods. On the other hand, when $\nu > 2$, in general R_s does not vanish, so that $H(u(1)) \neq H(u(0))$.

The above remark gives us a hint about how to approach the problem. Note that in (8) demanding that each term of the sum representing the quadrature formula is null (i.e., the conditions (6)), is an excessive requirement to obtain the conservation property, which causes the observed low degree of precision. A weaker assumption, that would leave the result unchanged, is to relax conditions (6) so as to devise a method whose induced quadrature formula, evaluated on a suitable line integral that links two successive points of the numerical solution, is exact and, at the same time, makes the corresponding sum vanish, without requiring that each term is zero. More precisely, in the new methods, conditions (6) will turn out to be replaced by relations of the form

$$\sigma'(c_i) = \sum_j \beta_{ij} J \nabla H(\sigma(c_j)),$$

which resemble a sort of *extended collocation condition* (see also [11, Section 2]) since $\sigma'(c_i)$ brings information from the global behavior of the problem in the time interval $[0, h]$ (see (17)–(33) in Section 3 and the analogues in Section 4). In this sense, the methods that we shall devise can be regarded as a kind of *extended collocation methods*.

If we use $\sigma(t)$ instead of $u(t)$, the integrand function in (3) has degree $\nu s - 1$ so that, in order for the quadrature formula to be exact, one would need say, $k+1$ points, where

$$k = \left\lceil \frac{\nu s}{2} \right\rceil, \tag{9}$$

if the corresponding Lobatto abscissae are used. Of course, in such a case, the vanishing of the quadrature formula is no longer guaranteed by conditions (6) and must be obtained by a different approach. For this purpose, let

$$r = k - s, \tag{10}$$

be the number of the required additional points, and let

$$0 < \tau_1 < \cdots < \tau_r < 1, \tag{11}$$

be r additional abscissae distinct from (4). Moreover, let us define the following *silent stages* [11],

$$w_i \equiv \sigma(\tau_i), \quad i = 1, \ldots, r. \tag{12}$$

Consequently, the polynomial $\sigma(t)$, which interpolates the couples (c_i, y_i), $i = 0, 1, \ldots, s$, also interpolates the couples (τ_i, w_i), $i = 1, \ldots, r$. That is, $\sigma(t)$ interpolates at $k+1$ points, even though it has only degree s. If we define the abscissae

$$\{t_0 < t_1 < \cdots < t_k\} = \{c_i\} \cup \{\tau_i\} \tag{13}$$

and dispose them according to a Lobatto distribution in $[0, 1]$ in order to get a formula of degree $2k$, we have that

[2] Different choices of the abscissae will be the subject of future investigations.

$$\int_0^1 \sigma'(t)^T \nabla H(\sigma(t))dt = \sum_{i=0}^{k} b_i \sigma'(t_i)^T \nabla H(\sigma(t_i)) \tag{14}$$

and, consequently, the conservation condition becomes

$$\sum_{i=0}^{k} b_i \sigma'(t_i)^T \nabla H(\sigma(t_i)) = 0, \tag{15}$$

where, now,

$$b_i = \int_0^1 \prod_{j=0, j\neq i}^{k} \frac{t-t_j}{t_i - t_j} dt, \quad i = 0, 1, \ldots, k. \tag{16}$$

The left-hand side of (15) is called "discrete line integral" because, as will be clear in the sequel, the choice of the path $\sigma(t)$ is dictated by the numerical method by which we will solve problem (1) (see [11] for details).

With these premises, in Section 4, we devise such methods, able to fulfill (15), after having set some preliminary results in Section 3. Section 5 contains the analysis of the energy-preserving methods. A few numerical tests are then reported in Section 6 and, finally, some conclusions are given in Section 7. For sake of completeness, some properties of shifted Legendre polynomials are listed in Appendix A.

3. Matrix form of collocation methods

In this section we deliberately do not care of the exactness of the discrete line integral, as stated by (14), and in fact we choose $k = s$ (and hence $t_i = c_i, \ i = 0, \ldots, s$). We show that imposing the vanishing of the discrete line integral (condition (15)) leads to the definition of the classical Lobatto IIIA methods. The reason why we consider this special situation is that the technique that we are going to exploit is easier to be explained, but at the same time is straightforwardly generalizable to the case $k > s$. As a by-product, we will gain more insight about the link between the new methods and the Lobatto IIIA class. For example, we will deduce that Lobatto IIIA methods may be defined by means of a polynomial $\sigma(t)$ of degree lower than that of the collocation polynomial $u(t)$ (indeed, $\deg \sigma(t) = \deg u(t) - 1$). For completeness, the link between u and σ will be elucidated in Section 3.3. To begin with, let us consider the following expansion of $\sigma'(c)$:

$$\sigma'(c) = \sum_{j=0}^{s-1} \gamma_j P_j(c), \tag{17}$$

where the (vector) coefficients γ_j are to be determined. Then, (15) becomes

$$\sum_{j=0}^{s-1} \gamma_j^T \sum_{i=0}^{s} b_i P_j(c_i) \nabla H(\sigma(c_i)) = 0, \tag{18}$$

which will clearly hold true, provided that the following set of orthogonality conditions are satisfied:

$$\gamma_j = \eta_j \sum_{i=0}^{s} b_i P_j(c_i) J \nabla H(\sigma(c_i)), \quad j = 0, \ldots, s-1, \tag{19}$$

where $\{\eta_j\}$ are suitable scaling factors. We now impose that the polynomial

$$\sigma(c) = y_0 + \sum_{j=0}^{s-1} \gamma_j \int_0^c P_j(x)\, dx \tag{20}$$

satisfies (5). We shall do this in Section 3.2, by using a matrix formulation of the methods, after setting some notation in Section 3.1.

3.1. Notations and preliminary results

The *shifted Legendre polynomials*, in the interval $[0, 1]$, constitute a family of polynomials, $\{P_n\}_{n\geq 0}$, for which a number of known properties, named **P1–P7**, are reported in Appendix A. Let us then set:

$$\boldsymbol{\gamma} = \begin{pmatrix} \gamma_0 \\ \vdots \\ \gamma_{s-1} \end{pmatrix}, \quad e = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^s, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_s \end{pmatrix}, \quad \hat{\mathbf{y}} = \begin{pmatrix} y_0 \\ \mathbf{y} \end{pmatrix}. \tag{21}$$

Moreover, with reference to the abscissae (4), let:

$$\mathbf{p}_j = \begin{pmatrix} P_j(c_1) \\ \vdots \\ P_j(c_s) \end{pmatrix}, \quad \hat{\mathbf{p}}_j = \begin{pmatrix} P_j(c_0) \\ \mathbf{p}_j \end{pmatrix}, \quad j = 0, \ldots, s, \tag{22}$$

$$\mathcal{P}_j = \begin{pmatrix} \mathbf{p}_0 & \ldots & \mathbf{p}_j \end{pmatrix} \in \mathbb{R}^{s \times j+1}, \quad \widehat{\mathcal{P}}_j = \begin{pmatrix} \hat{\mathbf{p}}_0 & \ldots & \hat{\mathbf{p}}_j \end{pmatrix} \in \mathbb{R}^{s+1 \times j+1}, \tag{23}$$

$$\mathbf{I}_j = \begin{pmatrix} \int_0^{c_1} P_j(x)dx \\ \vdots \\ \int_0^{c_s} P_j(x)dx \end{pmatrix}, \quad \hat{\mathbf{I}}_j = \begin{pmatrix} \int_0^{c_0} P_j(x)dx \\ \mathbf{I}_j \end{pmatrix} \equiv \begin{pmatrix} 0 \\ \mathbf{I}_j \end{pmatrix}, \quad j = 0, \ldots, s. \tag{24}$$

Furthermore, we set:

$$\mathcal{I}_j = \begin{pmatrix} \mathbf{I}_0 & \ldots \mathbf{I}_j \end{pmatrix} \in \mathbb{R}^{s \times j+1}, \quad \widehat{\mathcal{I}}_j = \begin{pmatrix} \hat{\mathbf{I}}_0 & \ldots \hat{\mathbf{I}}_j \end{pmatrix} \in \mathbb{R}^{s+1 \times j+1}, \tag{25}$$

$$D_j = \begin{pmatrix} 1 & & & \\ & 3 & & \\ & & \ddots & \\ & & & 2j-1 \end{pmatrix} \in \mathbb{R}^{j \times j}, \quad \Omega = \begin{pmatrix} b_0 & & \\ & \ddots & \\ & & b_s \end{pmatrix} \tag{26}$$

and

$$G_j = \begin{pmatrix} 1 & -1 & & \\ 1 & 0 & \ddots & \\ & 1 & \ddots & -1 \\ & & \ddots & 0 \\ & & & 1 \end{pmatrix} \in \mathbb{R}^{j+1 \times j}. \tag{27}$$

By virtue of **P2** and **P5**, we deduce that

$$\widehat{\mathcal{P}}_{j-1}^T \Omega \widehat{\mathcal{P}}_j = \left[D_j^{-1} \; \mathbf{0} \right], \quad j = 1, \ldots, s \tag{28}$$

and

$$\widehat{\mathcal{I}}_{j-1} = \frac{1}{2} \widehat{\mathcal{P}}_j G_j D_j^{-1}, \quad \mathcal{I}_{j-1} = \frac{1}{2} \mathcal{P}_j G_j D_j^{-1}, \quad j = 1, 2, \ldots. \tag{29}$$

The following result holds true.

Lemma 1. *Matrices* $\widehat{\mathcal{P}}_s = \begin{pmatrix} \hat{\mathbf{p}}_0 & \ldots & \hat{\mathbf{p}}_s \end{pmatrix} \in \mathbb{R}^{s+1 \times s+1}$ *and* $\mathcal{I}_{s-1} \in \mathbb{R}^{s \times s}$ *are nonsingular.*

Proof. $\widehat{\mathcal{P}}_s$ is the transpose of the Gramian matrix defined by the linearly independent polynomials $P_0(c), \ldots, P_s(c)$ at the distinct abscissae $c_0, \ldots, c_s$ and is, therefore, nonsingular. The matrix $\mathcal{I}_{s-1}$ is nonsingular since, from (27)–(29),

$$\widehat{\mathcal{I}}_{s-1} = \begin{pmatrix} \mathbf{0}^T \\ \mathcal{I}_{s-1} \end{pmatrix} = \frac{1}{2} \widehat{\mathcal{P}}_s G_s D_s^{-1} \in \mathbb{R}^{s+1 \times s} \tag{30}$$

with $\widehat{\mathcal{P}}_s$ and D_s nonsingular, and $\text{rank}(G_s) = s$. □

3.2. Matrix formulation

By imposing that the polynomial (20) satisfies (5), one obtains (see (24)–(29))

$$\mathcal{I}_{s-1} \otimes I_{2m} \gamma = \left(\frac{1}{2} \mathcal{P}_s G_s D_s^{-1} \right) \otimes I_{2m} \gamma = \mathbf{y} - e \otimes y_0. \tag{31}$$

Consequently,

$$\gamma = \left[2 D_s (\mathcal{P}_s G_s)^{-1} \begin{pmatrix} -e & I_s \end{pmatrix} \right] \otimes I_{2m} \hat{\mathbf{y}}. \tag{32}$$

On the other hand, the vector form of relations (19) reads

$$\gamma = \left(\Gamma \widehat{\mathcal{P}}_{s-1}^T \Omega\right) \otimes I_{2m} \hat{\mathbf{f}}, \tag{33}$$

where $\Gamma = \mathrm{diag}(\eta_0, \ldots, \eta_{s-1}) \in \mathbb{R}^{s\times s}$ and

$$\hat{\mathbf{f}} = (f_0 \quad \ldots \quad f_s)^T, \quad f_i = J\nabla H(\sigma(c_i)), \quad i = 0,\ldots,s. \tag{34}$$

Since Γ contains free parameters, we set

$$\Gamma = D_s. \tag{35}$$

Comparing (32) and (33), we arrive at the following block method, where now h denotes, in general, the used stepsize,

$$A \otimes I_{2m} \hat{\mathbf{y}} = hB \otimes I_{2m} \hat{\mathbf{f}} \tag{36}$$

with (see (29))

$$A = (-e \quad I_s), \quad B = \left(\frac{1}{2}\mathcal{P}_s G_s \widehat{\mathcal{P}}_{s-1}^T \Omega\right) \equiv \left(\mathcal{I}_{s-1} D_s \widehat{\mathcal{P}}_{s-1}^T \Omega\right). \tag{37}$$

The following noticeable result holds true.

Theorem 1. *Each row of the block method (36)–(37) defines a linear multistep formula of order* $s+1$. *The last row corresponds to the Lobatto quadrature formula of order* $2s$.

Proof. For the first part of the proof, it suffices to show that the method is exact for polynomials of degree $s+1$. Clearly, it is exact for polynomials of degree 0, due to the form of the matrix A. We shall then prove that $A\widehat{\mathcal{I}}_s = B\widehat{\mathcal{P}}_s$, that is (see (24), (25), and (37)), $\mathcal{I}_s = B\widehat{\mathcal{P}}_s$. By virtue of (37), (28), and considering that from property **P7** one obtains $\mathbf{I}_s = \mathbf{0}$, we have

$$B\widehat{\mathcal{P}}_s = \mathcal{I}_{s-1} D_s \widehat{\mathcal{P}}_{s-1}^T \Omega \widehat{\mathcal{P}}_s = \mathcal{I}_{s-1} D_s \left[D_s^{-1} \ \mathbf{0}\right] = [\mathcal{I}_{s-1} \ \mathbf{I}_s] = \mathcal{I}_s,$$

which completes the first part of the proof. For the second part, one has to show, by setting as usual e_i the ith unit vector, that

$$e_s^T B = (b_0 \quad \ldots \quad b_s),$$

the vector containing the coefficients of the quadrature formula. From (37), exploiting property **P4** (see also (27)), we obtain

$$e_s^T B = \frac{1}{2} e_s^T \mathcal{P}_s G_s \widehat{\mathcal{P}}_{s-1}^T \Omega = \frac{1}{2}(1 \quad \ldots \quad 1) G_s \widehat{\mathcal{P}}_{s-1}^T \Omega = e_1^T \widehat{\mathcal{P}}_{s-1}^T \Omega = (1 \quad \ldots \quad 1)\Omega = (b_0 \quad \ldots \quad b_s). \qquad \square$$

As an immediate consequence, the following result follows.

Corollary 1. *The block method (36)–(37) collocates at the Lobatto abscissae (4) and has global order* $2s$.

Proof. The proof follows from known results about collocation methods (see, e.g., [1, Theorem II.1.5]). □

Remark 2. In conclusion, the method corresponding to the pencil (A, B), as defined by (37), is nothing but the Lobatto IIIA method of order $2s$.

3.3. Link between $\sigma(c)$ and the collocation polynomial $u(c)$

An important consequence of Theorem 1 and Corollary 1 is that the Lobatto IIIA method of order $2s$ may be also defined by means of an underlying polynomial, namely $\sigma(c)$, of degree s instead of $s+1$, as is the collocation polynomial associated with the method (36).

The main aim of the present subsection is to elucidate the relation between these two polynomials. In what follows, we deliberately ignore the result obtained in Theorem 1 and Corollary 1, so as to provide, among other things, an alternative proof of part of the statements they report.

Let $u(c)$ be the polynomial (6) (of degree $s+1$) that collocates problem (1) at the abscissae (4). The expansion of $u'(c)$ along the shifted Legendre polynomials basis reads

$$u'(c) = \sum_{j=0}^{s} \zeta_j P_j(c). \tag{38}$$

Consequently, by setting

$$\hat{\mathbf{g}} = \begin{pmatrix} g_0 \\ \vdots \\ g_s \end{pmatrix}, \quad g_i = J\nabla H(u(c_i)), \quad \text{and} \quad \hat{\zeta} \equiv \begin{pmatrix} \zeta \\ \zeta_s \end{pmatrix} \equiv \begin{pmatrix} \begin{pmatrix} \zeta_0 \\ \vdots \\ \zeta_{s-1} \end{pmatrix} \\ \zeta_s \end{pmatrix},$$

one obtains that (6) may be recast in matrix notation as $\widehat{\mathcal{P}}_s \otimes I_{2m}\hat{\zeta} = \hat{\mathbf{g}}$, or

$$\hat{\zeta} = \widehat{\mathcal{P}}_s^{-1} \otimes I_{2m}\hat{\mathbf{g}}. \tag{39}$$

We get the expression of $u(c)$ by integrating both sides of (38) on the interval $[0, c]$:

$$u(c) = y_0 + \sum_{j=0}^{s-1}\zeta_j \int_0^c P_j(x)dx + \zeta_s \int_0^c P_s(x)dx. \tag{40}$$

By virtue of property **P7**, we get

$$u(c_i) = y_0 + \sum_{j=0}^{s-1}\zeta_j \int_0^{c_i} P_j(x)dx, \quad i = 0,\ldots,s. \tag{41}$$

Setting $z_i = u(c_i),\ i = 1,\ldots,s,\ \mathbf{z} = (z_1,\ldots,z_s)^T$, and $\hat{\mathbf{z}} = (y_0, \mathbf{z}^T)^T$, allows us to recast (41) in matrix form. This is done by exploiting a similar argument used to get (31) starting from (20). By taking into account (37), one then obtains:

$$A \otimes I_{2m}\hat{\mathbf{z}} = \mathbf{z} - e \otimes y_0 = \mathcal{I}_{s-1} \otimes I_{2m}\zeta = \left(\frac{1}{2}\mathcal{P}_s G_s D_s^{-1}\right) \otimes I_{2m}\zeta = \left(\frac{1}{2}\mathcal{P}_s G_s \left[D_s^{-1}\ \mathbf{0}\right]\right) \otimes I_{2m}\hat{\zeta}. \tag{42}$$

Inserting (39) into (42), and exploiting (28), yields

$$A \otimes I_{2m}\hat{\mathbf{z}} = \frac{1}{2}\mathcal{P}_s G_s \left[D_s^{-1}\ \mathbf{0}\right]\widehat{\mathcal{P}}_s^{-1} \otimes I_{2m}\hat{\mathbf{g}} = \frac{1}{2}\mathcal{P}_s G_s \widehat{\mathcal{P}}_{s-1}^T \Omega \otimes I_{2m}\hat{\mathbf{g}} = B \otimes I_{2m}\hat{\mathbf{g}}.$$

Thus, the collocation problem (6) defines the very same method arising from the polynomial $\sigma(c)$ (see (36) and (37)) with $h = 1$. This implies that system (36) is a collocation method defined on the Lobatto abscissae $c_i,\ i = 0,\ldots,s$ (therefore, a Lobatto IIIA method), and provides an alternative proof of Corollary 1. In particular, we deduce that

$$u(c_i) = y_i = \sigma(c_i), \quad i = 0,\ldots,s.$$

It follows that (40) becomes

$$u(c) = \sigma(c) + \zeta_s \int_0^c P_s(x)dx \tag{43}$$

and, after differentiating,

$$u'(c) = \sigma'(c) + \zeta_s P_s(c). \tag{44}$$

We can obtain the expression of the unknown ζ_s by imposing a collocation condition at any of the abscissae c_i. For example, choosing $c = c_s = 1$, yields

$$\zeta_s = u'(1) - \sigma'(1) = f(y_s) - \sum_{j=0}^{s-1}\gamma_j = f(y_s) - e^T \otimes I_{2m}\gamma. \tag{45}$$

This latter expression can be slightly simplified by considering that:

(i) $f(y_s) = (e^T\ 1)\widehat{\mathcal{P}}_s^{-1} \otimes I_{2m}\hat{\mathbf{f}}$, which comes from the fact that the system $\widehat{\mathcal{P}}_s^T x = \begin{pmatrix} e \\ 1 \end{pmatrix}$ has solution $x = e_{s+1}$ (the nonsingularity of $\widehat{\mathcal{P}}_s$ being assured by Lemma 1);

(ii) from (32) and (36), (37), one has

$$\gamma = (D_s\widehat{\mathcal{P}}_{s-1}^T\Omega) \otimes I_{2m}\hat{\mathbf{f}} = (D_s\widehat{\mathcal{P}}_{s-1}^T\Omega\widehat{\mathcal{P}}_s\widehat{\mathcal{P}}_s^{-1}) \otimes I_{2m}\hat{\mathbf{f}} = (D_s(D_s^{-1}\ \mathbf{0})\widehat{\mathcal{P}}_s^{-1}) \otimes I_{2m}\hat{\mathbf{f}} = (I_s\ \mathbf{0})\widehat{\mathcal{P}}_s^{-1} \otimes I_{2m}\hat{\mathbf{f}}.$$

Thus, from (45) we get

$$\zeta_s = \left[(e^T\ 1) - (e^T\ 0)\right]\widehat{\mathcal{P}}_s^{-1} \otimes I_{2m}\hat{\mathbf{f}} = e_{s+1}^T\widehat{\mathcal{P}}_s^{-1} \otimes I_{2m}\hat{\mathbf{f}}. \tag{46}$$

The remaining collocation conditions, $u'(c_i) = J\nabla H(u(c_i)),\ i = 0,\ldots,s-1$, are clearly satisfied since the collocation polynomial $u(c)$ is uniquely identified by the $s+2$ linearly independent conditions in (6). Nonetheless, they can be easily checked after observing that, from (43), (ii), and (46),

$$\hat{\zeta} = \begin{pmatrix} \gamma \\ \zeta_s \end{pmatrix} = \widehat{\mathcal{P}}_s^{-1} \otimes I_{2m} \hat{\mathbf{f}}.$$

Therefore, from (38), (39) and (43), one obtains,

$$\hat{\mathbf{u}}' \equiv \begin{pmatrix} u'(c_0) \\ \vdots \\ u'(c_s) \end{pmatrix} = \widehat{\mathcal{P}}_s \otimes I_{2m} \zeta = \widehat{\mathcal{P}}_s \widehat{\mathcal{P}}_s^{-1} \otimes I_{2m} \hat{\mathbf{f}} = \hat{\mathbf{f}}.$$

That is (see (34)), $u'(c_i) = J\nabla H(u(c_i))$, $i = 0, \ldots, s$.

4. Derivation of the energy-preserving methods

In Section 3, we have considered the particular case $k = s$. In the general case, i.e., when $k \geqslant s$, condition (15) can be recast as

$$\sum_{j=0}^{s-1} \gamma_j^T \sum_{i=0}^{k} b_i P_j(t_i) \nabla H(\sigma(t_i)) = 0, \tag{47}$$

which will clearly hold true, provided that the following set of orthogonality conditions are satisfied:

$$\gamma_j = \eta_j \sum_{i=0}^{k} b_i P_j(t_i) J \nabla H(\sigma(t_i)), \quad j = 0, \ldots, s-1, \tag{48}$$

where $\{\eta_j\}$ are suitable scaling factors. According to (35), we choose them as $\eta_j = 2j+1,\ j = 0, \ldots, s-1$.[3] The vector γ (see (21)) is then obtained by imposing that the polynomial $\sigma(c)$ in (20) satisfies the interpolation constrains (5) and (12). In so doing, one obtains a block method characterized by the pencil (A, B), where the two $k \times k+1$ matrices A and B are defined as follows. In order to simplify the notation, we shall use a "Matlab-like" notation: let $ind_s \in \mathbb{R}^{s+1}$ and $ind_r \in \mathbb{R}^r$ be the vectors whose entries are the indexes, belonging to $\{1, \ldots, k+1\}$, of the main abscissae $c_0 < \cdots < c_s$ in (4) and of the silent ones $\tau_1 < \cdots < \tau_r$ in (11), respectively, within the Lobatto abscissae $t_0 < \cdots < t_k$, as defined in (13). Then, the orthogonality conditions (48) will define the first s rows of A and B[4] (compare with (37)):

$$A(1:s, ind_s) = (-e \quad I_s), \quad B(1:s, :) = \left(\mathcal{I}_{s-1} D_s \bar{\mathcal{P}}^T \bar{\Omega}\right), \tag{49}$$

where (see (25),(26) and (13))

$$\bar{\mathcal{P}} = \begin{pmatrix} P_0(t_0) & \ldots & P_{s-1}(t_0) \\ \vdots & & \vdots \\ P_0(t_k) & \ldots & P_{s-1}(t_k) \end{pmatrix} \in \mathbb{R}^{k+1 \times s} \tag{50}$$

and (see (7))

$$\bar{\Omega} = \begin{pmatrix} b_0 & & \\ & \ddots & \\ & & b_k \end{pmatrix} \in \mathbb{R}^{k+1 \times k+1}. \tag{51}$$

On the other hand, the interpolation conditions for the silent stages (12) define the last r rows of the matrix A (the corresponding rows of B are obviously zero):

$$\begin{aligned} &A(s+1:k, ind_r) = I_r, \\ &A(s+1:k, ind_s) = -\bar{\mathcal{I}}_r \left[\mathcal{I}_{s-1}^{-1} (-e \; I_s)\right] - \bar{e} \cdot e_1^T, \end{aligned} \tag{52}$$

where I_r is the identity matrix of dimension $r, \bar{e} = (1, \ldots, 1)^T \in \mathbb{R}^r,\ e_1$ is the first unit vector (of dimension $s+1$), and

$$\bar{\mathcal{I}}_r = \begin{pmatrix} \int_0^{\tau_1} P_0(x)dx & \ldots & \int_0^{\tau_1} P_{s-1}(x)dx \\ \vdots & & \vdots \\ \int_0^{\tau_r} P_0(x)dx & \ldots & \int_0^{\tau_r} P_{s-1}(x)dx \end{pmatrix} \in \mathbb{R}^{r \times s}.$$

The following result generalizes Theorem 1 to the present setting (the proof being similar).

[3] It is worth mentioning that, even though any choice for the $\{\eta_j\}$ is in principle allowed, choosing $\eta_j = 2j+1$ maximizes the order of the resulting method, according to what proved in Corollary 2.

[4] As a further convention, the entries not explicitly set are assumed to be 0.

Theorem 2. *Each row of the block method (49)–(52) defines a linear multistep formula of order at least s. The s-th row corresponds to the Lobatto quadrature formula of order* $2k$*.*

Definition 1. We call the method defined by the pencil (A,B) in (49)–(52) a "*Hamiltonian BVM with k steps and degree s*", hereafter *HBVM* (k,s).[5]

Remark 3. The structure of the nonlinear system associated with the HBVM(k,s) is better visualized by performing a permutation of the stages that splits, into two block sub-vectors, the fundamental stages and the silent ones. More precisely, the permuted vector of stages, say $\mathbf{z}$, is required to be:

$$\mathbf{z} = \Bigg[\underbrace{y_0^T, y_1^T, \ldots, y_s^T}_{\text{fundamental stages}}, \underbrace{w_1^T, w_2^T, \ldots, w_r^T}_{\text{silent stages}}\Bigg]^T \equiv [y_0^T, \mathbf{y}^T, \mathbf{w}^T]^T.$$

This is accomplished by introducing the permutation matrices $W \in \mathbb{R}^{k\times k}$ and $W_1 \in \mathbb{R}^{k+1\times k+1}$, such that

$$W\begin{pmatrix} 2 \\ \vdots \\ k+1 \end{pmatrix} = \begin{pmatrix} ind_s(2:s+1) \\ ind_r \end{pmatrix}, \quad W_1\begin{pmatrix} 1 \\ \vdots \\ k+1 \end{pmatrix} = \begin{pmatrix} ind_s \\ ind_r \end{pmatrix}.$$

It is easy to realize that

$$WAW_1^T = \begin{pmatrix} -e & I_s & 0_{s\times r} \\ -a_0 & -A_1 & I_r \end{pmatrix}, \quad WBW_1^T = \begin{pmatrix} b_0 & B_1 & B_2 \\ \mathbf{0} & 0_{r\times s} & 0_{r\times r} \end{pmatrix},$$

where $[-a_0, -A_1]$ coincides with $A(s+1:k, ind_s)$ in (52), while $[b_0, B_1, B_2]$ matches the matrix $B(1:s,:)$ in (49). The HBVM(k,s) then takes the form:

$$\begin{pmatrix} -e & I_s & 0_{s\times r} \\ -a_0 & -A_1 & I_r \end{pmatrix} \otimes I_{2m}\mathbf{z} = h\begin{pmatrix} b_0 & B_1 & B_2 \\ \mathbf{0} & 0_{r\times s} & 0_{r\times r} \end{pmatrix} \otimes J\nabla H(\mathbf{z}). \tag{53}$$

The presence of the null blocks in the lower part of WBW_1^T clearly suggests that the (generally nonlinear) system (53) of (block) size k is actually equivalent to a system having (block) size s. Indeed, we can easily remove the silent stages,

$$\mathbf{w} = a_0 \otimes y_0 + A_1 \otimes I_{2m}\mathbf{y}$$

and obtain

$$\mathbf{y} = e \otimes y_0 + hb_0 \otimes (J\nabla H(y_0)) + hB_1 \otimes J\nabla H(\mathbf{y}) + hB_2 \otimes J\nabla H(a_0 \otimes y_0 + A_1 \otimes I_{2m}\mathbf{y}). \tag{54}$$

(We refer to [19] for an alternative technique to reduce the dimension of system (53). The main idea, in this case, is to reformulate the discrete problem in terms of the coefficients $\{\gamma_j\}$ (see (48)) of the polynomial σ, which are s, independently of k.)

Remark 4. As was shown in the previous section, when $k = s$, the HBVM (s,s) coincides with the Lobatto IIIA method of order $2s$. More in general, for $k \geqslant s$, by summing up (49)–(52), we can cast HBVM(k,s) as a $(k+1)$-stage Runge–Kutta method with the following tableau:

$$\begin{array}{c|ccc} t_0 & & & \\ \vdots & & \bar{\mathcal{I}} D_s \bar{\mathcal{P}}^T \bar{\Omega} & \\ t_k & & & \\ \hline & b_0 & \ldots & b_k \end{array} \tag{55}$$

where

$$\bar{\mathcal{I}} = \begin{pmatrix} \int_0^{t_0} P_0(x)dx & \ldots & \int_0^{t_0} P_{s-1}(x)dx \\ \vdots & & \vdots \\ \int_0^{t_k} P_0(x)dx & \ldots & \int_0^{t_k} P_{s-1}(x)dx \end{pmatrix} \in \mathbb{R}^{k+1\times s}.$$

We observe that the $k+1 \times k+1$ matrix

$$C = \bar{\mathcal{I}} D_s \bar{\mathcal{P}}^T \bar{\Omega} \tag{56}$$

[5] Indeed, the pencil (A,B) perfectly fits the framework of block BVMs (see, e.g., [13]).

appearing in (55) has rank s, thus confirming that the computational cost per iteration depends on s, rather than on k (see [14] for more details and a practical example of Butcher tableau concerning the method HBVM (6,2)).

By the way, we observe that, when $s = 1$, HBVM$(k,1)$ are nothing but the "s-stage trapezoidal methods", defined in [9], based on the Lobatto abscissae. In such a case, the matrix C becomes

$$C = \begin{pmatrix} t_0 \\ \vdots \\ t_k \end{pmatrix} (\, b_0 \quad \ldots \quad b_k \,).$$

Similarly, for $s = 2$ and $k = 4$, HBVM (4,2) coincides with the fourth-order method presented in [11, Section 4.2], able to preserve polynomial Hamiltonians of degree four.

5. Analysis of the methods

Concerning the order of convergence of HBVM(k,s) methods, the following result generalizes that of Corollary 1.

Corollary 2. *For all $k \geqslant s$, the HBVM (k,s) (49)–(52) has order of convergence $p = 2s$.*

Proof. By virtue of Theorem 2, the corresponding Runge–Kutta method (55) satisfies the usual simplifying assumptions $B(2k)$ and $C(s)$. If we are able to prove $D(s-1)$, from the classical result of Butcher (see, e.g., [15, Theorem 5.1]), it will follow that the method has order $p = 2s$. With reference to (55), the condition $D(s-1)$ can be cast in matrix form, by introducing the vectors $e = (1,\ldots,1)^T \in \mathbb{R}^{s-1}$, $\bar{e} = (1,\ldots,1)^T \in \mathbb{R}^{k+1}$, and the matrices

$$Q = \mathrm{diag}(1,\ldots,s-1), \quad T = \mathrm{diag}(t_0,\ldots,t_k), \quad V = (t_{i-1}^{j-1}) \in \mathbb{R}^{k+1\times s-1},$$

as

$$QV^T\bar{\Omega}(\bar{\mathcal{I}}D_s\bar{\mathcal{P}}^T\bar{\Omega}) = \left(e\bar{e}^T - V^TT\right)\bar{\Omega},$$

i.e.,

$$\bar{\mathcal{P}}D_s\bar{\mathcal{I}}^T\bar{\Omega}VQ = (\bar{e}e^T - TV). \tag{57}$$

Since the quadrature is exact for polynomials of degree $2s-1 \leqslant 2k-1$, one has

$$\left(\bar{\mathcal{I}}^T\bar{\Omega}VQ\right)_{ij} = \left(\sum_{\ell=0}^{k} b_\ell \int_0^{t_\ell} P_{i-1}(x)\mathrm{d}x\,(jt_\ell^{j-1})\right) = \left(\int_0^1\int_0^t P_{i-1}(x)\mathrm{d}x(jt^{j-1})\mathrm{d}t\right) = \left(\delta_{i1} - \int_0^1 P_{i-1}(x)x^j\mathrm{d}x\right),$$
$$i = 1,\ldots,s, \quad j = 1,\ldots,s-1,$$

where the last equality is obtained by integrating by parts, with δ_{i1} the Kronecker symbol. Consequently,

$$\left(\bar{\mathcal{P}}D_s\bar{\mathcal{I}}^T\bar{\Omega}VQ\right)_{ij} = \left(1 - \sum_{\ell=0}^{s-1}\eta_\ell P_\ell(t_i)\int_0^1 P_\ell(x)x^j\mathrm{d}x\right) = (1 - t_{i-1}^j), \quad i = 1,\ldots,k+1, \quad j = 1,\ldots,s-1,$$

that is, (57), where the last equality follows from the fact that

$$\sum_{\ell=0}^{s-1}\eta_\ell P_\ell(t)\int_0^1 P_\ell(x)x^j\mathrm{d}x = t^j, \quad j = 1,\ldots,s-1. \qquad \square$$

An additional, remarkable property of such methods is gained, provided that the abscissae $\{t_0,\ldots,t_k\}$ (13) are symmetrically distributed (as is the case of the Lobatto abscissae here considered). For this purpose, we need to introduce some notations and preliminary results. Let us define the matrix

$$E_n = \begin{pmatrix} & & & 1 \\ & & \cdot & \\ & \cdot & & \\ 1 & & & \end{pmatrix} \in \mathbb{R}^{n\times n}, \quad n = 1,2,\ldots,$$

which, when applied to a vector of length n, reverses the order of its entries. We also set

$$\begin{aligned} L &= \begin{pmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix} \in \mathbb{R}^{k\times k+1}, \\ F &= \begin{pmatrix} (-1)^0 & & & \\ & (-1)^1 & & \\ & & \ddots & \\ & & & (-1)^{s-1} \end{pmatrix} \in \mathbb{R}^{s\times s}. \end{aligned} \tag{58}$$

The following preliminary result holds true.

Lemma 2. *If the abscissae (13) are symmetric, then matrix (56) satisfies:*

$$E_k L C E_{k+1} = LC.$$

Proof. From the symmetry of the abscissae it easily follows that (see (16) and (51))

$$E_{k+1}\bar{\Omega}E_{k+1} = \bar{\Omega}.$$

From property **P3**, we have that (see (50))

$$\bar{\mathcal{P}}^T E_{k+1} = F\bar{\mathcal{P}}^T.$$

Moreover, by considering that (see (4))

$$L\mathcal{I} = \begin{pmatrix} \int_{t_0}^{t_1} P_0(x)dx & \dots & \int_{t_0}^{t_1} P_{s-1}(x)dx \\ \vdots & & \vdots \\ \int_{t_{k-1}}^{t_k} P_0(x)dx & \dots & \int_{t_{k-1}}^{t_k} P_{s-1}(x)dx \end{pmatrix},$$

again from **P3** we see that

$$E_s L\mathcal{I} = L\mathcal{I}F.$$

Finally, from (56) we obtain

$$E_k L C E_{k+1} = (E_k L\mathcal{I})D_s(\bar{\mathcal{P}}^T E_{k+1})(E_{k+1}\bar{\Omega}E_{k+1}) = L\mathcal{I}FD_sF\bar{\mathcal{P}}^T\bar{\Omega} = L\mathcal{I}D_s\bar{\mathcal{P}}^T\bar{\Omega} = LC. \qquad \Box$$

As a consequence, we have the following result.

Theorem 3. *If the abscissae (13) are symmetric, then the method (49)–(52) (i.e., (55)) is symmetric, that is, it is self-adjoint.*

Proof. Indeed, the discrete solution, $\hat{\mathbf{y}}$, satisfies the equation (see (55),(56) and (58))

$$L\otimes I_{2m}\hat{\mathbf{y}} = hLC\otimes I_{2m}f(\hat{\mathbf{y}}).$$

Considering that $E_k L E_{k+1} = -L$ and, from Lemma 2, $E_k LCE_{k+1} = LC$, one then obtains

$$L\otimes I_{2m}(E_{k+1}\otimes I_{2m}\hat{\mathbf{y}}) = -hLC\otimes I_{2m}(E_{k+1}\otimes I_{2m}f(\hat{\mathbf{y}})) = -hLC\otimes I_{2m}f(E_{k+1}\otimes I_{2m}\hat{\mathbf{y}}).$$

The thesis then follows by observing that the vector $E_{k+1}\otimes I_{2m}\hat{\mathbf{y}}$ contains the time-reversed discrete solution. □

The next theorem summarizes the results about HBVM(k,s).

Theorem 4 (*Main Result*). *For all $s = 1,2,\ldots$, and $k \geqslant s$, the HBVM (k,s) method:*

1. is symmetric;
2. has order of accuracy $2s$;
3. is energy-preserving for polynomial Hamiltonians of degree not larger than $2k/s$;
4. for general $C^{(2k+1)}$ Hamiltonians, the energy error at each integration step is $O(h^{2k+1})$, if h is the used stepsize.[6]

[6] Consequently, on any finite interval the global energy error is not larger than $O(h^{2k})$.

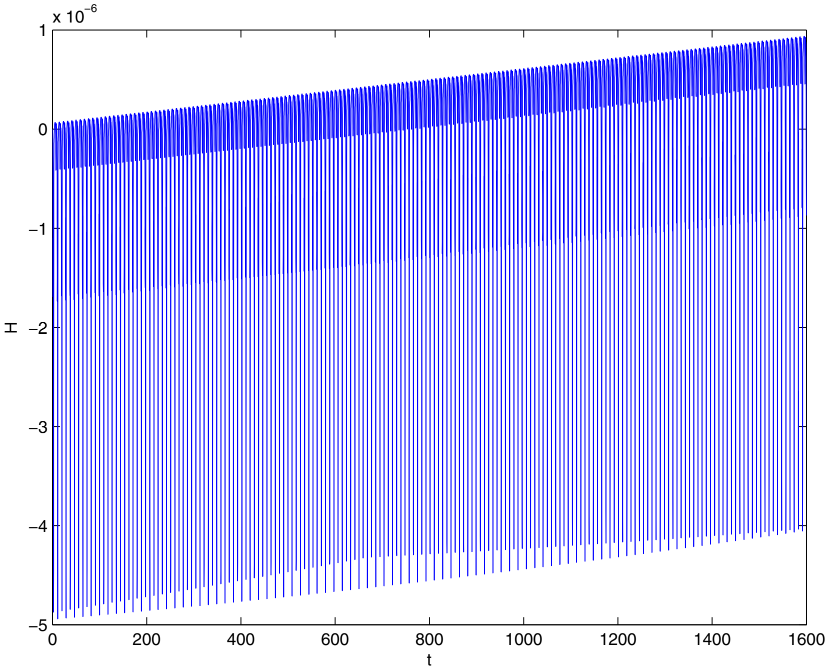

Fig. 1. Fourth-order Lobatto IIIA method, $h = 0.16$, problem (59).

Proof. Item 1 follows from Theorem 3, since the Lobatto abscissae $\{t_i\}$ are symmetrically distributed. Item 2 follows from Corollary 2. Item 3 follows from the fact that, for such polynomial Hamiltonians, the vanishing discrete line integral equals the continuous line integral (see (14) and (15)). Finally, Item 4 follows from the fact that, by using arguments similar to those used in Remark 1 (see (8)), the energy error per integration step equals the quadrature error of the Gauss–Lobatto formula of order $2k$. Indeed, for a general stepsize h, one would obtain, by taking into account (20)–(48):

$$\begin{aligned} H(y_1) - H(y_0) &= H(\sigma(h)) - H(\sigma(0)) = h\int_0^1 \sigma'(\tau h)^T \nabla H(\sigma(\tau h))d\tau = h\left(\sum_{i=0}^k b_i \nabla H(\sigma(t_i h))^T \sigma'(t_i h) + R_k(h)\right) \\ &= h\sum_{i=0}^k b_i \nabla H(\sigma(t_i h))^T \sum_{j=0}^{s-1} P_j(t_i)\gamma_j + hR_k(h) = h\sum_{j=0}^{s-1} \underbrace{\left[\sum_{i=0}^k b_i P_j(t_i)\nabla H(\sigma(t_i h))\right]^T}_{=[\eta_j^{-1} J^T \gamma_j]^T} \gamma_j + hR_k(h) \\ &= h\sum_{j=0}^{s-1} \eta_j^{-1}\gamma_j^T J\gamma_j + hR_k(h) = hR_k(h). \end{aligned}$$

The thesis completes by recalling that, when choosing the $k+1$ Lobatto abscissae, then $R_k(h) = O(h^{2k})$. □

Remark 5. Since HBVM(k,s) is a one-step method (indeed, a Runge–Kutta method), the result of Theorem 4 continues to hold in the case where the stepsize h is dynamically changed at each integrations step.

6. Numerical tests

We here report a few numerical tests, in order to show the potentialities of HBVM(k,s).

Let then consider, at first, the Hamiltonian problem characterized by the polynomial Hamiltonian (4.1) in [3],

$$H(p,q) = \frac{p^3}{3} - \frac{p}{2} + \frac{q^6}{30} + \frac{q^4}{4} - \frac{q^3}{3} + \frac{1}{6} \qquad (59)$$

having degree $\nu = 6$, starting at the initial point $y_0 \equiv (q(0), p(0))^T = (0,1)^T$. For such a problem, in [3] it has been experienced a numerical drift in the discrete Hamiltonian, when using the fourth-order Lobatto IIIA method[7] with stepsize $h = 0.16$. This is confirmed by the plot in Fig. 1, where a linear drift in the numerical Hamiltonian is clearly observable. On the other hand, by

[7] Such method coincides with the HBVM (2,2) above described.

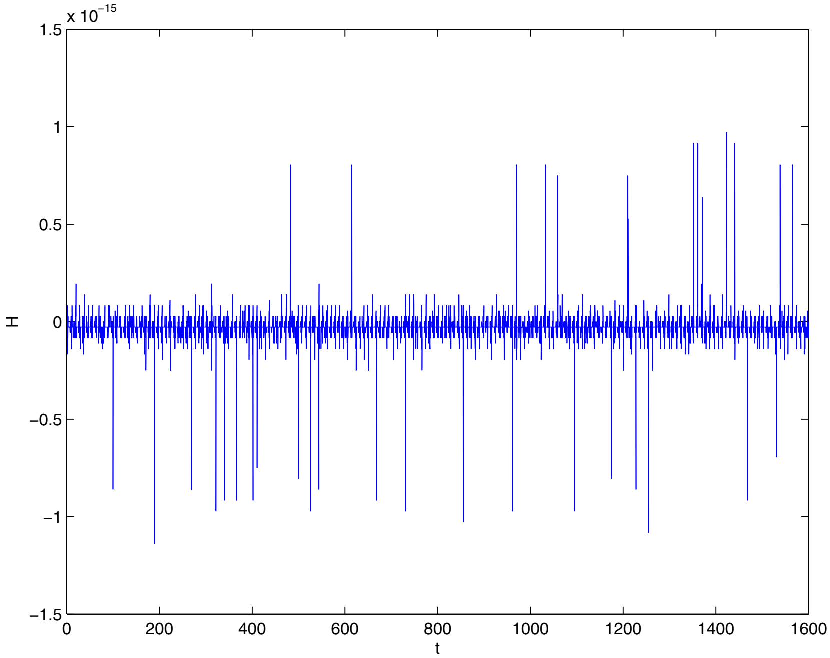

Fig. 2. Fourth-order HBVM (6,2) method, $h = 0.16$, problem (59).

Table 1
Numerical order of convergence for the HBVM (6,2) method, problem (59).

h	0.32	0.16	0.08	0.04	0.02
Error	$2.288 \cdot 10^{-2}$	$1.487 \cdot 10^{-3}$	$9.398 \cdot 10^{-5}$	$5.890 \cdot 10^{-6}$	$3.684 \cdot 10^{-7}$
Order	–	3.94	3.98	4.00	4.00

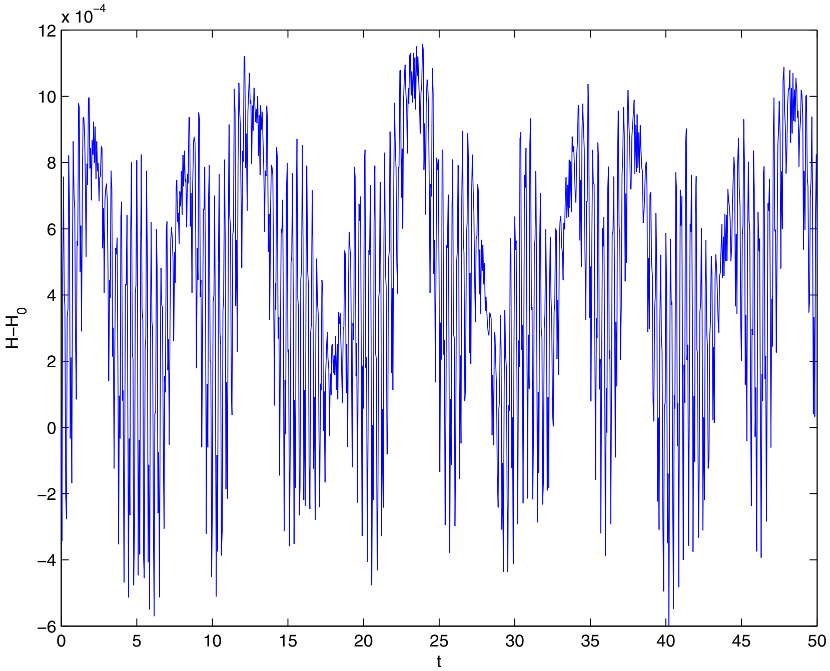

Fig. 3. Fourth-order Lobatto IIIA method, $h = 0.05$, problem (60).

Table 2
Numerical order of convergence for the HBVM (4,2) method, problem (60).

h	$1.6 \cdot 10^{-2}$	$8 \cdot 10^{-3}$	$4 \cdot 10^{-3}$	$2 \cdot 10^{-3}$	10^{-3}
Error	3.030	$1.967 \cdot 10^{-1}$	$1.240 \cdot 10^{-2}$	$7.761 \cdot 10^{-4}$	$4.853 \cdot 10^{-5}$
Order	–	3.97	3.99	4.00	4.00

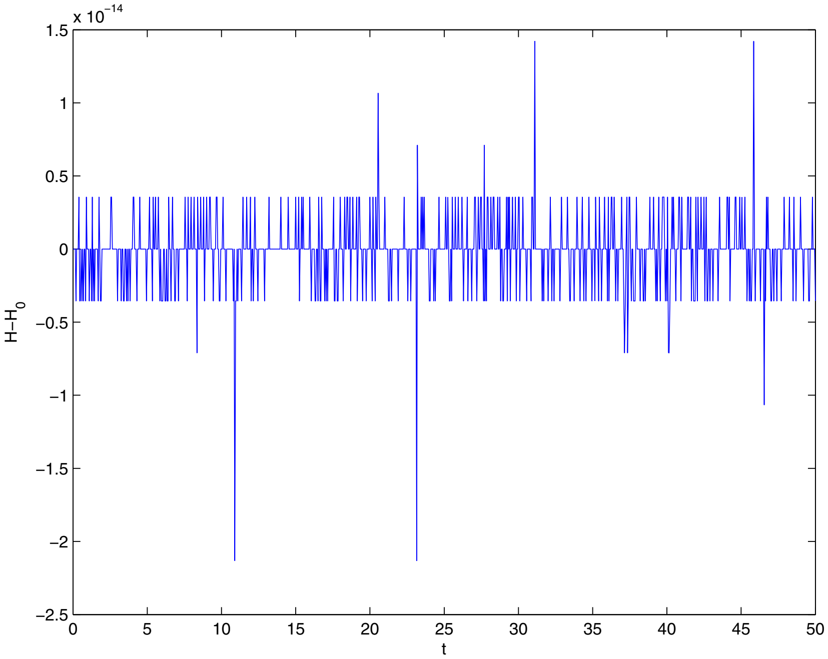

Fig. 4. Fourth-order HBVM (4,2) method, $h = 0.05$, problem (60).

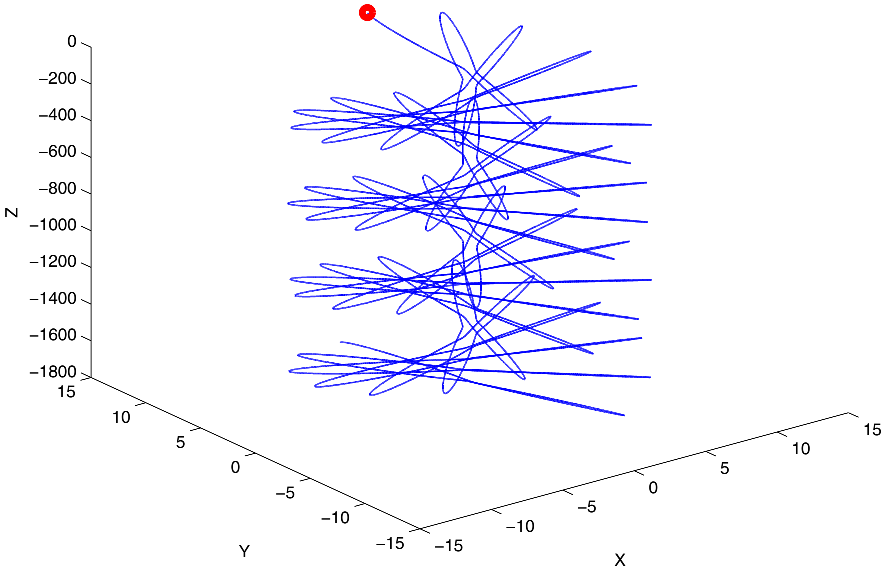

Fig. 5. Phase-space plot of the solution of problem (61) for $0 \leqslant t \leqslant 10^3$ (the circle denotes the starting point of the trajectory).

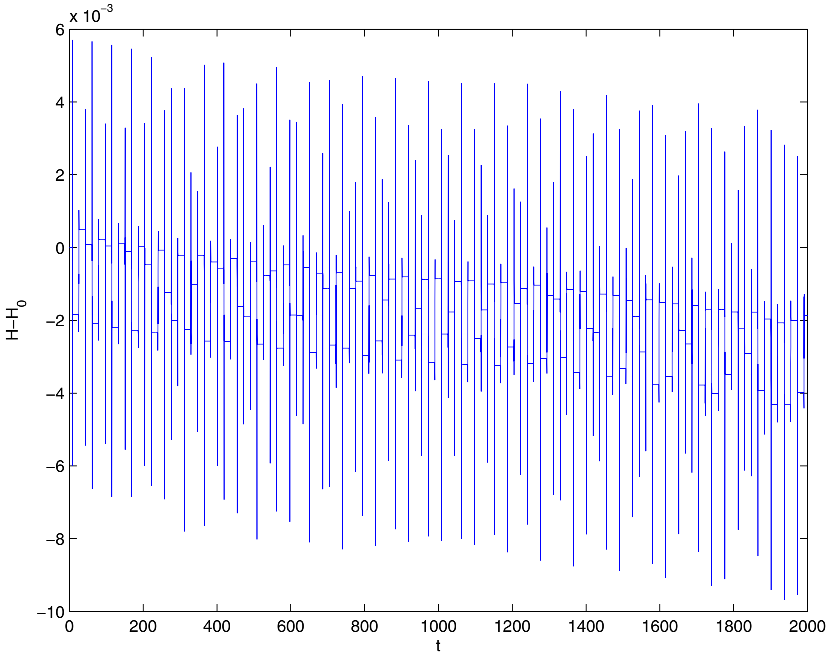

Fig. 6. Fourth-order Lobatto IIIA method, $h = 0.1$, problem (61).

using the fourth-order HBVM (6,2) with the same stepsize, the drift disappears, as shown in Fig. 2, since such method exactly preserves polynomial Hamiltonians of degree up to 6. Moreover, the order of convergence $p = 4$ is (numerically) confirmed by the results listed in Table 1, where the used stepsizes h, the maximum estimated error (obtained as the difference of two consecutive solutions), and the estimated order of convergence are listed.

The second test problem is the Fermi–Pasta–Ulam problem (see [1, Section I.5.1]), defined by the Hamiltonian

$$H(p,q) = \frac{1}{2}\sum_{i=1}^{m}(p_{2i-1}^2 + p_{2i}^2) + \frac{\omega^2}{4}\sum_{i=1}^{m}(q_{2i} - q_{2i-1})^2 + \sum_{i=0}^{m}(q_{2i+1} - q_{2i})^4 \tag{60}$$

with $q_0 = q_{2m+1} = 0,\ m = 3,\ \omega = 50$, and starting vector

$$p_i = 0, \quad q_i = (i-1)/10, \quad i = 1, \ldots, 6.$$

In such a case, the Hamiltonian function is a polynomial of degree 4, so that the fourth-order HBVM (4,2) method, which is used with stepsize $h = 0.05$, is able to exactly preserve the Hamiltonian, as confirmed by the plot in Fig. 4, whereas the fourth-order Lobatto IIIA method provides the result plotted in Fig. 3. Moreover, in Table 2 we list corresponding results as in Table 1, again confirming the fourth-order convergence.

In the previous examples, the Hamiltonian function was a polynomial. Nevertheless, as is easily argued from Theorem 4, HBVM(k,s) are expected to produce a *practical* conservation of the energy when applied to systems defined by a non-polynomial Hamiltonian function which are sufficiently differentiable. As an example, we consider the motion of a charged particle in a magnetic field with Biot–Savart potential.[8] It is defined by the Hamiltonian

$$H(x,y,z,\dot{x},\dot{y},\dot{z}) = \frac{1}{2m}\left[\left(\dot{x} - \alpha\frac{x}{\rho^2}\right)^2 + \left(\dot{y} - \alpha\frac{y}{\rho^2}\right)^2 + (\dot{z} + \alpha\log(\rho))^2\right] \tag{61}$$

with $\rho = \sqrt{x^2 + y^2},\ \alpha = eB_0,\ m$ is the particle mass, e is its charge, and B_0 is the magnetic field intensity. We have used the values

$$m = 1, \quad e = -1, \quad B_0 = 1$$

with starting point

$$x = 0.5, \quad y = 10, \quad z = 0, \quad \dot{x} = -0.1, \quad \dot{y} = -0.3, \quad \dot{z} = 0.$$

In Fig. 5, the trajectory of the particle in the interval $[0, 10^3]$ is plotted in the phase space. As one can see, it is a helix that wings downward. By using the fourth-order Lobatto IIIA method with stepsize $h = 0.1$, a drift in the numerical Hamiltonian can be again observed (see Fig. 6), so that the method does introduce a friction. When using the HBVM (4,2) method with the

[8] As an example, this kind of motion causes the well known phenomenon of *aurora borealis*.

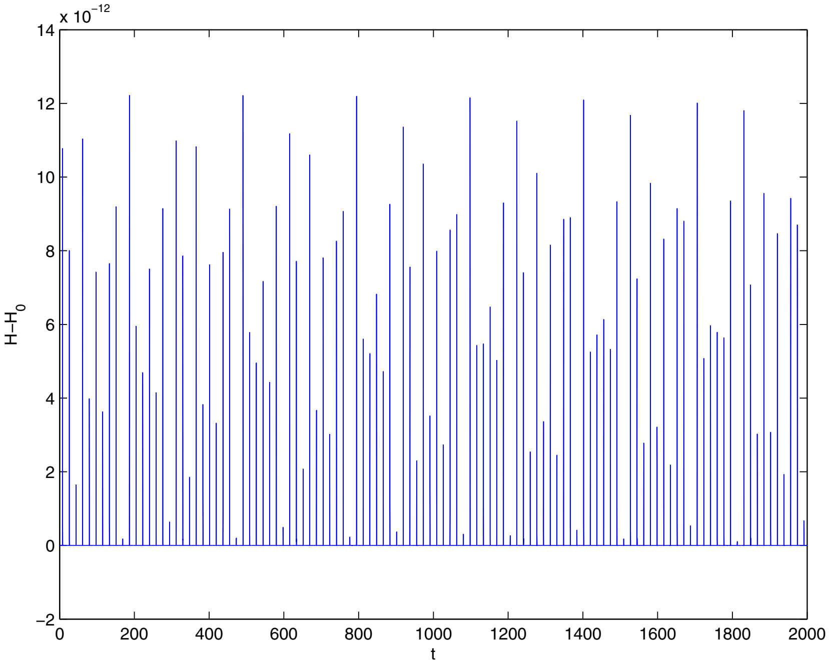

Fig. 7. Fourth-order HBVM (4,2) method, $h = 0.1$, problem (61).

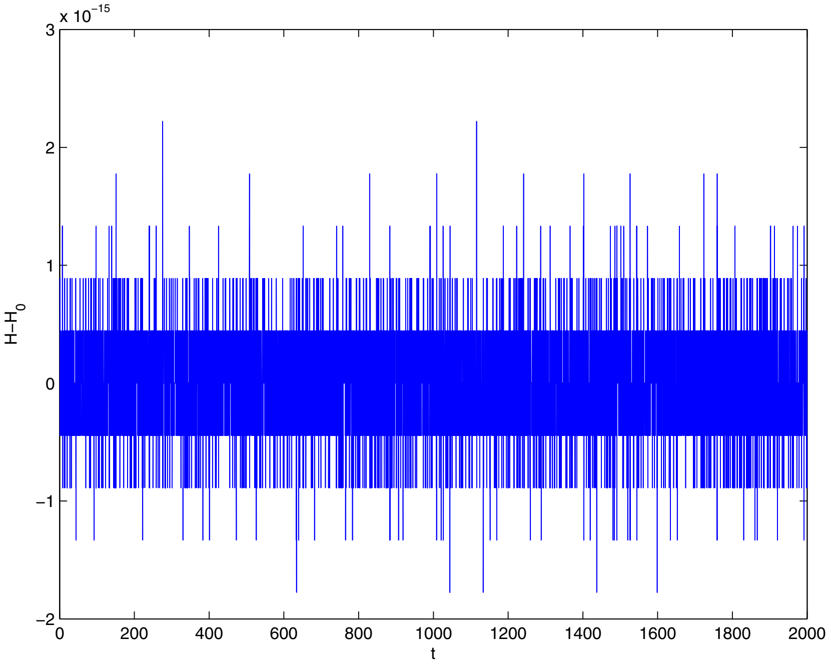

Fig. 8. Fourth-order HBVM (6,2) method, $h = 0.1$, problem (61).

Table 3
Numerical order of convergence for the HBVM (6,2) method, problem (61).

h	$3.2 \cdot 10^{-2}$	$1.6 \cdot 10^{-2}$	$8 \cdot 10^{-3}$	$4 \cdot 10^{-3}$	$2 \cdot 10^{-3}$
Error	$3.944 \cdot 10^{-6}$	$2.635 \cdot 10^{-7}$	$1.729 \cdot 10^{-8}$	$1.094 \cdot 10^{-9}$	$6.838 \cdot 10^{-11}$
Order	–	3.90	3.93	3.98	4.00

same stepsize, the drift disappears and the Hamiltonian turns out to be almost preserved (see Fig. 7). As expected, the result improves if we increase k: the plot in Fig. 8 has been obtained by using the HBVM (6,2), from which one realizes that a practical preservation of the Hamiltonian is reached. Finally, the data listed in Table 3 confirm the fourth-order convergence of the latter method.

7. Conclusions

In this paper a new class of numerical methods, able to preserve polynomial Hamiltonians, has been studied in detail. From the analysis, it turns out that such methods can be regarded as a generalization of collocation Runge–Kutta Lobatto IIIA methods. Nevertheless, the fact of being characterized by a matrix pencil, perfectly fits the framework of block BVMs, so that we have named them Hamiltonian BVMs (HBVMs). A number of numerical tests prove their effectiveness in preserving the Hamiltonian function when evaluated along the numerical solution, as well as confirm the predicted order of convergence. Possible different choices of the abscissae, as well as the actual efficient implementation of the methods, will be the subject of future investigations.

Acknowledgments

We wish to thank Prof. Mario Trigiante for providing us with problem (61). We are grateful to Ernst Hairer, for his comments on the proof of Corollary 2, and to the referees, for the useful comments.

Appendix A. Some properties of shifted Legendre polynomials

The shifted Legendre polynomials $\{P_n\}_{n\geqslant 0}$, can be obtained recursively as follows:

$$\begin{aligned}
&P_0(x) \equiv 1,\\
&P_1(x) = 2x-1,\\
&(n+1)P_{n+1}(x) = (2n+1)(2x-1)P_n(x) - nP_{n-1}(x), \quad n = 1,2,\ldots .
\end{aligned}$$

A number of useful properties of such polynomials are here recalled: for their proof see any book on special functions (e.g., [16]).

P1. Lobatto quadrature: the Lobatto abscissae $\{c_i\}$ (4), of the formula of degree $2s$, are the zeros of the polynomial

$$(x^2-x)P_s'(x),$$

where $P_s'(x)$ denotes the derivative of $P_s(x)$. The corresponding weights (7) are given by:

$$b_i = \frac{1}{s(s+1)(P_s(c_i))^2}, \quad i = 0,1,\ldots,s,$$

which are, therefore, all positive.

P2. Orthogonality:

$$\int_0^1 P_n(x)P_m(x)\,dx = \frac{1}{2n+1}\delta_{nm}, \quad n = 0,1,\ldots,$$

where, as usual, δ_{nm} denotes the Kronecker delta.

P3. Symmetry:

$$P_n(1-x) = (-1)^n P_n(x), \quad n = 0,1,\ldots$$

P4. Symmetry at the end-points:

$$P_n(0) = (-1)^n, \quad P_n(1) = 1, \quad n = 0,1,\ldots$$

P5. Integrals:

$$2\int_0^x P_0(t)\,dt = 2x = P_1(x) + P_0(x),$$

$$2(2n+1)\int_0^x P_n(t)\,dt = P_{n+1}(x) - P_{n-1}(x), \quad n = 1,2,\ldots .$$

P6. Shifted Legendre differential equations. The shifted Legendre polynomials satisfy the second order differential equation:

$$\frac{d}{dx}\left[(x^2-x)P_n'(x)\right] + n(n+1)P_n(x) = 0, \quad n = 0,1,\ldots .$$

P7. From **P1** and **P6**, it follows that, if (4) are the Lobatto abscissae of the formula of order $2s$ (i.e., exact for polynomials of degree $2s-1$), then

$$\int_0^{c_i} P_s(x)\,dx = 0, \quad i = 0, 1, \ldots, s.$$

References

[1] Hairer E, Lubich C, Wanner G. Geometric numerical integration. 2nd ed. Berlin: Springer; 2006.

[2] Leimkuhler B, Reich S. Simulating Hamiltonian dynamics. Cambridge Univ. Press; 2004.

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

[4] Brugnano L, Trigiante D. Energy drift in the numerical integration of Hamiltonian problems. J Numer Anal Ind Appl Math 2009;4(3–4):153–70.

[5] McLachlan RI, Quispel GRW, Robidoux N. Geometric integration using discrete gradient. Philos Trans R Soc London A 1999;357:1021–45.

[6] Chartier P, Faou E, Murua A. An algebraic approach to invariant preserving integrators: the case of quadratic and Hamiltonian invariants. Numer Math 2006;103:575–90.

[7] Quispel GRW, McLaren DI. A new class of energy-preserving numerical integration methods. J Phys A Math Theor 2008;41:045206 (7pp).

[8] Celledoni E, McLachlan RI, McLaren D, Owren B, Quispel GRW, Wright WM. Energy preserving Runge–Kutta methods. M2AN 2009;43:645–9.

[9] Iavernaro F, Pace B. s-Stage trapezoidal methods for the conservation of Hamiltonian functions of polynomial type. AIP Conf Proc 2007;936:603–6.

[10] Iavernaro F, Pace B. Conservative block-boundary value methods for the solution of polynomial Hamiltonian systems. AIP Conf Proc 2008;1048:888–91.

[11] Iavernaro F, Trigiante D. High-order symmetric schemes for the energy conservation of polynomial Hamiltonian problems. J Numer Anal Ind Appl Math 2009;4(1–2):87–111.

[12] Iavernaro F, Trigiante D. Discrete conservative vector fields induced by the trapezoidal method. J Numer Anal Ind Appl Math 2006;1:113–30.

[13] Brugnano L, Trigiante D. Solving differential problems by multistep initial and boundary value methods. Amsterdam: Gordon and Breach Science Publ.; 1998.

[14] Brugnano L, Iavernaro F, Susca T. Hamiltonian BVMs (HBVMs): implementation details and applications. AIP Conf Proc 2009;1168:723–6.

[15] Hairer E, Wanner G. Solving ordinary differential equations II. 2nd ed. Berlin: Springer; 1996.

[16] Abramovitz M, Stegun IA. Handbook of mathematical functions. Dover; 1965.

[17] L. Brugnano, F. Iavernaro, D. Trigiante. Analisys of Hamiltonian boundary value methods (HBVMs) for the numerical solution of polynomial Hamiltonian dynamical systems; 2009. <arXiv:0909.5659[math.NA]>.

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

[19] Brugnano L, Iavernaro F, Trigiante D. A note on the efficient implementation of Hamiltonian BVMs. J Comput Appl Math 2011;236(3):375–83.

[20] Calvo M, Hernández-Abreu D, Montijano JI, Rández L. On the preservation of invariants by explicit Runge–Kutta methods. SIAM J Sci Comput 2006;28(3):868–85.