High-order symmetric schemes for the energy conservation of polynomial Hamiltonian problems.*

Felice Iavernaro† and Donato Trigiante‡

Abstract

We define a class of arbitrary high order symmetric one-step methods that, when applied to Hamiltonian systems, are capable to precisely conserve the Hamiltonian function when this is a polynomial, whatever the initial condition and the stepsize h used.

The key idea to devise such methods is the use of the so called *discrete line integral*, the discrete counterpart of the the line integral in conservative vector fields. This approach naturally suggests a formulation of such methods in terms of block Boundary Value Methods, although they can be as well recast as Runge-Kutta methods, if preferred.

Key words: Hamiltonian and conservative systems, block-Boundary Value Methods, quadrature formulae, discrete line integral.

Subject classification: 65P10, 65L05.

1 Introduction and motivations

We are concerned with the numerical integration of Hamiltonian systems with m degrees of freedom

$$\begin{cases} \dot{y} = J\nabla H(y), \\ y(t_0) = y_0, \end{cases} \qquad J = \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix}, \tag{1}$$

*Work developed within the project "Numerical methods and software for differential equations".

†Dipartimento di Matematica, Università di Bari, Via Orabona 4, I-70125 Bari (Italy), felix@dm.uniba.it.

‡Dipartimento di Energetica, Università di Firenze, via C. Lombroso 6/17, I-50134 Firenze (Italy), trigiant@unifi.it.

where I is the identity matrix of dimension m. The state vector y is partitioned into two m-length vectors p and q, the conjugate momenta and the generalized coordinates respectively. Our aim is to devise new families of one-step high order methods $y_n = \Phi_h(y_{n-1})$ (h is the stepsize of integration) capable of providing numerical approximations y_n to the true solution $y(t_n)$ such that

$$H(y_{n+1}) = H(y_n), \qquad \text{for all } n \text{ and } h > 0, \tag{2}$$

in the case where $H(p, q)$ is a polynomial in the variables p and q. From a topological point of view this means that the discrete orbit generated by the numerical method is demanded to lie on the same manifold of the continuous one, namely $H(p, q) = H(p_0, q_0)$. Thus we get the relevant property that the global portrait in the $2m$ dimensional phase space is preserved after the discretization of (1), whatever the initial point y_0 and the stepsize h used. Many interesting evolutionary systems deriving from different application fields are defined by polynomial Hamiltonian functions.

Example 1.1 *The Fermi-Pasta-Ulam Problem.* It is defined by the Hamiltonian function

$$H(p, q) = \frac{1}{2}\sum_{i=1}^{m}(p_{2i-1}^2 + p_{2i}^2) + \frac{\omega^2}{4}\sum_{i=1}^{m}(q_{2i} - q_{2i-1})^2 + \sum_{i=0}^{m}(q_{2i+1} - q_{2i})^4. \tag{3}$$

This problem arises from molecular dynamics and describes the interaction of $2m$ mass points linked with alternating soft nonlinear and stiff linear springs, in a one-dimensional lattice with fixed end points ($q_0 = q_{2m+1} = 0$) [6].

In the last section we will consider further examples for numerical tests[1]. Alternately, some nonlinear Hamiltonian systems may be well approximated by polynomials. Taylor expansion is a common tool to get polynomial approximations to dynamical systems in a neighborhood of equilibrium points, especially when simple linearization does not help in studying their stability character, as is the case of marginally stable equilibria.

Example 1.2 *Polynomial pendulum oscillator.* Starting from the Hamiltonian function of the nonlinear pendulum equation

$$H(p, q) = 1/2p^2 + 1 - \cos q, \tag{4}$$

we retain a finite number of terms in the Taylor expansion of the cosine, thus obtaining:

$$\begin{aligned}
&H(p, q) = \tfrac{1}{2}p^2 + \tfrac{1}{2}q^2 - \tfrac{1}{24}q^4 && \text{(quartic pendulum oscillator)},\\
&H(p, q) = \tfrac{1}{2}p^2 + \tfrac{1}{2}q^2 - \tfrac{1}{24}q^4 + \tfrac{1}{720}q^6 && \text{(pend. oscillator of degree 6)},\\
&H(p, q) = \tfrac{1}{2}p^2 + \tfrac{1}{2}q^2 - \tfrac{1}{24}q^4 + \tfrac{1}{720}q^6 - \tfrac{1}{40320}q^8 && \text{(pend. oscillator of degree 8)},
\end{aligned} \tag{5}$$

[1] A more classical but puzzling example is the so called *Infinitesimal Hilbert 16th problem.*

and so on. The approximation may be indefinitely improved by adding more and more terms in the expansion. On a computer a polynomial pendulum oscillator of high enough degree would be undistinguishable from the original problem. For example, the last polynomial in (5) produces an approximation of order 10^{-10} for $|q| < 1/2$ (this issue will be discussed again in the last section).

As well known, symplectic or symmetric RK-methods only conserve quadratic Hamiltonian functions $H(y) = \frac{1}{2}y^T Cy$, but, in general, they fail to yield conservation for higher degree. In the general case, taking aside the roundoff errors deriving from the use of floating point arithmetic, one can expertise two different behaviors in a neighborhood of a critical point:

(a) *nearly conservation*, which means that, although the energy of the system is not strictly conserved by the numerical method, the sequence $H(y_n)$ displays an oscillating behavior around the theoretical value $H(y_0)$, where the amplitude of oscillation is bounded with respect to the time n and of size $O(h^p)$, p being the order of convergence.

(b) *energy drift*, that is the numerical method alters the marginal stability character of the equilibrium by dissipating or absorbing energy thus leading to asymptotic stability (that characterize dissipative systems) or instability.

The two situations listed above are visible in Figure 1 where we report the energy function $H(y_n)$ obtained by applying the LobattoIIIB method of order four to the quartic pendulum oscillator (5) (left picture) and to the system defined by the following Hamiltonian function:

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

This problem was introduced in [5] as a counterexample showing how symmetric methods may display the energy drift phenomenon even when the problem is reversible: $H(-p,q) = H(p,q)$. [2]

Of course the point (b) refers to a much more dangerous situation with respect to the one described in point (a). Nevertheless, even the assumption of *nearly conservation* would not prevent the occurrence of completed misleading results about the asymptotic behavior of the solution in the phase space, as the next example shows.

[2] In fact, in [5] the author show that the system deriving from (6) is equivalent to a reversible system. See also [2] for a discussion about the definition and the role of symmetry of Hamiltonian systems in the numerical integration.

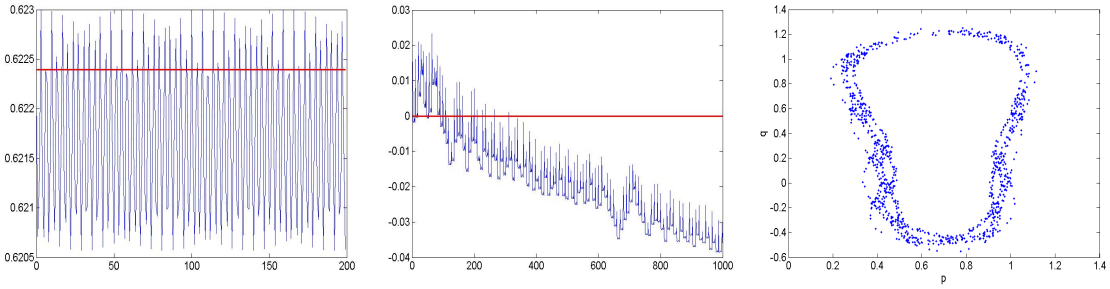

Figure 1: Energy function $H(y_n)$ evaluated over the numerical solution obtained by the Lobatto IIIB method of order four applied to two Hamiltonian systems. Left picture: the quartic pendulum oscillator has been solved with stepsize $h = 1$, number of points $n = 200$ and initial condition $[p_0, q_0] = [1, 0.5]$. Central picture: problem (6) has been solved with stepsize $h = 1$, number of points $n = 1000$ and initial condition $[p_0, q_0] = [1, 0]$. The right picture confirms that the solution in the phase plane is far from describing a closed orbit. In the first two plots the horizontal line denotes the theoretical value of the Hamiltonian function.

Example 1.3 The system with one degree of freedom defined by the cubic Hamiltonian function

$$H(p, q) = p^2 + q^2 + \frac{1}{10}(p + q)^3 \tag{7}$$

admits the origin and the point $P^* = (p^*, q^*) = (-\frac{5}{3}, -\frac{5}{3})$ as equilibrium points, the former being a center and the second a saddle point.

The left upper plot of Figure 2 shows the phase space portrait associated to system (7): there are closed orbits surrounding the origin which are enveloped by open orbits which embrace P^* and eventually depart to infinity.

This picture also reports the orbit (big dots) computed by the LobattoIIIA method of order 4 starting from the initial point $y_0 = [p_0, q_0] = [-1, -1]$ and stepsize $h = 1$. The numerical solution undergoes small oscillations around the level curve $H(p, q) = H(p_0, q_0)$ (see the upper right picture), but since y_0 is close enough to the origin, any point of the numerical solution remains strictly inside the stability region associated to the origin and the stability character of the true solution is preserved.

On the other hand, as is easily argued, even a small deviation from a closed level curve may drastically change the fate of an orbit if the dynamics takes place near the boundary of a stability region of a given equilibrium point. This is emphasized in the bottom pictures which show the numerical solution obtained by choosing $y_0 = [-1.5643, -1.6430]$ (the orbit generated by such initial point is indeed closed). After a few cycles around the origin, the numerical solution comes out of the stability region and is pulled away towards infinity.

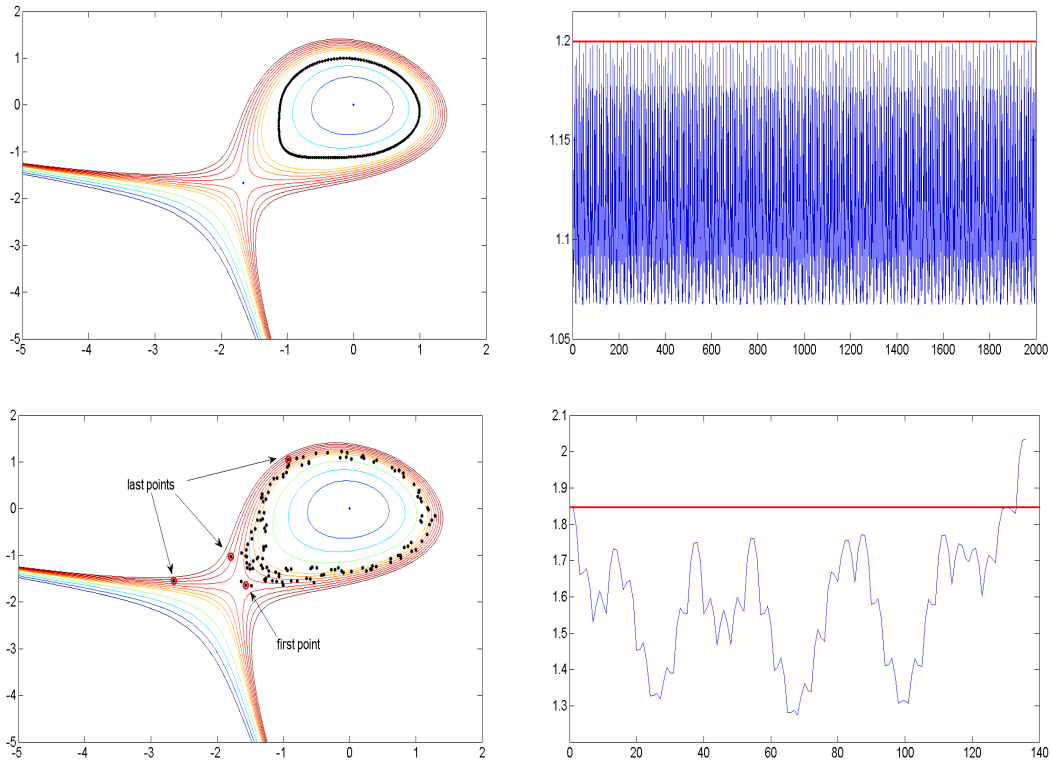

Figure 2: Two orbits generated by the Lobatto IIIA method, with stepsize $h = 1$, applied to system (7). The left pictures display these orbits as dots in the phase plane (p, q), while the right pictures report the values $H(y_n)$ (the constant lines correspond to the values of the energy of the theoretical solution). The method does introduce oscillations around the theoretical closed orbit that may destroy the correct stability behavior of the solution, when the amplitude of oscillations is large or when the dynamics takes place near the boundary of the stability region of a given marginally stable equilibrium point.

2 Extended collocation methods

The key idea to devise methods satisfying (2) is based upon the combination of the following two ingredients: the definition of *discrete line integral* (introduced in [11] and [10]) and the *extended collocation conditions*.

Given a vector field $f : \mathbb{R}^n \to \mathbb{R}^n$, and a path $\sigma : t \in [0, 1] \to \mathbb{R}^{2m}$ joining two pints in $\mathbb{R}^n$, say y_0 and y_1, the *discrete line integral* associated to a given method for the numerical solution of Initial Value Problems, is defined by the solution of the pure quadrature problem $\dot{z} = (\dot{\sigma}(t))^T f(\sigma(t))$ in the time interval $[0, 1]$, using stepsize

$h = 1$. This means that the line integral

$$\int_{y_0 \to y_1} f(y) \cdot \mathrm{d}y \equiv \int_0^1 (\dot{\sigma}(t))^T \, f(\gamma(t)) \, \mathrm{d}t. \tag{8}$$

is approximated by the simple quadrature formula induced by the numerical method. By analogy, the notation

$$\sum_{\gamma} f \cdot \Delta y \equiv \sum_{i=0}^{k} b_i (\dot{\gamma}(t_i))^T \, f(\gamma(t_i)),$$

will denote a discrete line integral: the coefficients b_i and $t_i \in [0, 1]$ are the weights and abscissae of the induced quadrature formula.

The interplay between the continuous and the discrete line integrals in association with the differential problem and its discrete counterpart, is summarized by the following diagram:

$$\begin{array}{ccc} \dot{y} = f(y) & \longrightarrow & \displaystyle\int_{y_0 \to y_1} f(y) \cdot \mathrm{d}y \\ \Big\downarrow & & \Big\downarrow \\ y_1 = \Phi_h(y_0) & \longrightarrow & \displaystyle\sum_{\gamma} f \cdot \Delta y \end{array}$$

Before introducing more formally such tools and their use, we make a preliminary remark in order to better elucidate the guide line of our investigation.

Remark 2.1 A collocation Runge-Kutta method applied to (1) and defined on the collocation abscissae $t_i = t_0 + c_i h$, with $0 \leq c_1 < \cdots < c_s \leq 1$ is defined by means of the following polynomial interpolation problem:

$$\begin{cases} \gamma(t_0) = y_0, \\ \dot{\gamma}(t_i) = J \nabla H(\gamma(t_i)), \quad i = 1, \ldots, s. \end{cases} \tag{9}$$

Conditions (9) uniquely define a polynomial $\gamma(t)$ of degree s which is used to advance the solution by posing $y_1 = \gamma(t_0 + h)$. The weights b_i and the coefficients a_{ij} of the Butcher array are defined as

$$b_i = \int_0^1 \ell_i(c) \mathrm{d}c, \qquad a_{ij} = \int_0^{c_i} \ell_j(c) \mathrm{d}c, \qquad \text{with} \;\; \ell_i(c) = \prod_{j \neq i} \frac{c - c_j}{c_i - c_j}.$$

As well known, the order p of the resulting RK is at least $s + 1$ (see for example [6]).

The s-degree polynomial $\gamma(t)$ may be thought of as a path in the phase space joining the state vectors y_0 and $y_1 = \gamma(t_0 + h)$.[3] Due to conservativeness of the vector field, we have that $H(y_1) - H(y_0) = \int_\sigma \nabla H(y) \cdot \mathrm{d}y$, where $\sigma : [0, 1] \to \mathbb{R}^{2m}$ is any path such that $\sigma(0) = y_0$ and $\sigma(1) = y_1$. Choosing $\sigma(c) = \gamma(t_0 + ch)$ yields

$$H(y_1) - H(y_0) = \int_{t_0}^{t_0+h} (\dot\gamma(t))^T \nabla H(y(t)) \mathrm{d}t = \sum_{i=1}^{s} b_i (\dot\gamma(t_i))^T \nabla H(y(t_i)) + E_s, \quad (10)$$

where in the last equality we have approximated the integral by using the quadrature formula $\sum_{i=1}^{s} b_i f(t_i)$ induced by the RK method. The error E_s is proportional to $h^{p+1} f^{(p)}(\xi)$ where ξ is a suitable point in the interval $[0, 1]$ and $f(t) = (\dot\gamma(t))^T \nabla H(y(t))$. From (9) we have

$$(\dot\gamma(t_i))^T \nabla H(y(t_i)) = -\nabla^T H(y(t_i)) J \nabla H(y(t_i)) = 0, \qquad \text{for all } i = 1, \ldots, s,$$

which imply that the quadrature approximation in (10) vanishes. Therefore $H(y_1) = H(y_0)$ if and only if $E_s = 0$, which means that the quadrature formula must be exact when applied to the integrand polynomial $(\dot\gamma(t))^T \nabla H(y(t))$. Since the order p cannot exceed $2s$, such condition is equivalent to requiring that the degree of the integrand is less than or equal to $2s - 1$:

$$\deg(\gamma) - 1 + (\deg(H) - 1) \deg(\gamma) \leq p - 1 \leq 2s - 1. \quad (11)$$

However, since $\deg(\gamma) = s$, this inequality imposes that $\deg(H) \leq 2$, which states the well known result that this class of methods conserve quadratic Hamiltonian functions while fail to conserve polynomial Hamiltonian functions of higher degree.

The computation carried out in Remark 2.1 gives us very useful suggestions on how to modify things in order to retrive the conservativeness property (2) for high degree polynomials. The trick is to observe that a method may be defined in such a way as to have order p when applied to the differential equation $\dot y = f(t, y)$ and order $d > p$ when applied to the pure quadrature formula $\dot y = f(t)$. For such a method, inequality (11) would read

$$\deg(\gamma) - 1 + (\deg(H) - 1) \deg(\gamma) \leq d - 1, \quad (12)$$

which would be satisfied if the path γ has moderate degree and/or d is large enough.

The idea is then to allow the stages to be partly used to confer the method a given order and partly to make the quadrature formula in (10) exact and null. More precisely, as for standard collocation methods, the approximating polynomial will interpolate all the stages even though its degree will be less than the number of stages minus one. This means that some of the stages are deliberately positioned on the polynomial individuated by the remaining ones. The term *extended collocation method* has been coined to indicate such kind of relaxation.

[3] Since $K_i = \gamma(t_0 + c_i h)$, $i = 1, \ldots, s$, the path γ also links the intermediate stages which are as well approximations to the true solution: $K_i \simeq y(t_0 + c_i h)$.

3 Polynomial Conservativeness

In order to construct the class of conservative methods we are looking for, it turns out to be advantageous to reverse the flow of reasoning presented in the previous section, that is we start by a generic polynomial path $\gamma(t)$ in the phase space $\mathbb{R}^{2m}$, impose conservativeness and then use the remaining free parameters to impose the order conditions. As we will see, this way to proceed naturally leads to a formulation of the methods in terms of *block-Boundary Value Methods* (block-BVMs), therefore hereafter we slightly change the notation. We will recall later how to transform a block-BVM into a Runge-Kutta method, if desired.

Consider $s+1$ points $z_i \in \mathbb{R}^{2m}$, $i = 0, \dots, s$, not better specified apart from z_0 which is assumed to be equal to the initial condition y_0. Let $\sigma : [0,1] \to \mathbb{R}^{2m}$ be the vector polynomial that interpolates the data (c'_i, z_i), $i = 0, \dots, s$, where $0 = c'_0 < c'_1 < \cdots < c'_s = 1$. Using Newton's basis and divided differences, the polynomial σ reads:

$$\begin{aligned} \sigma(c) = \quad & z_0 + (z_s - z_0)c + z[c'_0, c'_1, c'_s]c(c-1) + \dots \\ & + z[c'_0, c'_1, \dots, c'_{s-1}, c'_s]c(c-1)\cdots(c-c'_{s-2}), \end{aligned} \tag{13}$$

where $z[c'_0, \cdots, c'_\ell, c'_s]$ are the divided differences of the z_i, defined on the nodes c'_i, for example

$$z[c'_0, c'_1, c'_s] = \frac{1}{c'_1(1-c'_1)} \left(c'_1 z_s - z_1 + (1-c'_1) z_0 \right). \tag{14}$$

Now we add a number of auxiliary points w_j, $j = 1, \dots, r$, on the curve σ, on locations different from those corresponding to the points z_i. To do that, we consider additional abscissae $0 < \bar{c}_1 < \cdots < \bar{c}_r < 1$, with $\bar{c}_j \neq c'_i$ for all indices i and j, and set

$$w_j = \sigma(\bar{c}_j). \tag{15}$$

It is worth noticing that (15) implies that each point w_j is actually a linear combination of the points z_i, $i = 0, \dots, s$ (see also the examples in the next section).

As will be clear in a while, both the points z_i and w_j will act as stages in the numerical method that is going to be defined. However, after specifying the nodes $\bar{c}_j$, the additional points w_j only depend on the choice of the points y_i and do not alter the degree of the polynomial σ. For this reason we call these points *silent stages*: they do not contribute in increasing the order of the method but serve to get conservativeness.

Let $k = s + r$ and $0 = c_0 < c_1 < \cdots < c_k = 1$ be the abscissae c'_i and $\bar{c}_j$ gathered and sorted in ascending order. We consider the interpolation quadrature formula with weights b_i, $i = 0, \dots, k$, defined on the nodes c_i. Its degree of precision, that we denote by $d-1$ in order to be consistent with the notation used in (12), is at least k and at most $2k-1$: this latter case corresponds to the choice of Gauss-Lobatto

abscissae.[4] Finally, by $y_1, y_2, \ldots, y_k$, we denote all the points z_i, $i = 1, \ldots, s$, and w_j, $j = 1, \ldots, r$, sorted accordingly to the order introduced above for the abscissae c_i. Hence the polynomial $\sigma : [0,1] \to \mathbb{R}^{2m}$ has degree $s < k$ but interpolates all the data (c_i, y_i), $i = 0, \ldots, k$. We assume to use a sufficient number of silent stages to ensure that (12) be satisfied, which implies that our quadrature formula will be exact if applied to the integral $\int_0^1 (\dot{\sigma}(c))^T \nabla H(\sigma(c)) \mathrm{d}c$, for all Hamiltonian functions of degree up to d/s.

Now we are ready to retrieve the conditions for conservativeness: we observe that now relation (2) reads $H(y_k) = H(y_0)$. We have

$$\begin{aligned} H(y_k) - H(y_0) = & \int_{y_0 \to y_1} \nabla H(y) \cdot \mathrm{d}y = \int_0^1 (\dot{\sigma}(c))^T \nabla H(\sigma(c)) \mathrm{d}c = \sum_{i=0}^{k} b_i (\dot{\sigma}(c_i))^T \nabla H(\sigma(c_i)) \\ = & (y_k - y_0)^T \sum_{i=0}^{k} b_i \nabla H(y_i) + z[c'_0, c'_1, c'_s]^T \sum_{i=0}^{k} b_i (2c_i - 1) \nabla H(y_i) \\ & + z[c'_0, c'_1, \ldots, c'_{s-1}, c'_s]^T \sum_{i=0}^{k} b_i \frac{\mathrm{d}}{\mathrm{d}c} c(c-1) \cdots (c - c'_{s-2}) \Big|_{c=c_i} \nabla H(y_i). \end{aligned}$$

Hence the conservativeness conditions are

$$\begin{cases} (y_k - y_0)^T \displaystyle\sum_{i=0}^{k} b_i \nabla H(y_i) = 0, \\ z[c'_0, c'_1, c'_s]^T \displaystyle\sum_{i=0}^{k} b_i (2c_i - 1) \nabla H(y_i) = 0, \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ z[c'_0, c'_1, \ldots, c'_{s-1}, c'_s]^T \displaystyle\sum_{i=0}^{k} b_i \frac{\mathrm{d}}{\mathrm{d}c} c(c-1) \cdots (c - c'_{s-2}) \Big|_{c=c_i} \nabla H(y_i) = 0. \end{cases} \tag{16}$$

The first condition in (16) is satisfied if we assume that the state vector y_k is computed by the formula

$$y_k - y_0 = h \sum_{i=0}^{k} b_i J \nabla H(y_i). \tag{17}$$

In fact, multiplying both terms on the left by $(y_k - y_0)^T J^T$ just yields this first condition.

[4] In this paper the endpoints of the interval $[0,1]$ have been included in the set of abscissae, since we have assumed $c_0 = 0$ and $c_k = 1$. A further generalization of this approach would be to eliminate such constraints.

Analogously, the second condition suggests the following relation on the involved stages:

$$z[c_0', c_1', c_s'] = h \sum_{i=0}^{k} \eta b_i (2c_i - 1) \nabla H(y_i). \tag{18}$$

As is clear by looking at relation (14), the left hand side of (18) is nothing but a linear combination of the stages y_0, z_1 and y_k. The free parameter η will be suitably chosen in order to maximize the order of (18). Again, one easily verifies that this equation implies the second condition.

Continuing this reasoning, we see that all the remaining conditions in (16) will dictate analogous equations as the ones above. We have obtained s linear multistep equations involving y_0 and the k unknown stages $y_1 \ldots, y_k$. We need $k - s = r$ extra conditions in order to close the system, but these were already settled up in (15) and, as observed, consist of linear homogeneous difference equations with constant coefficients in the unknown y_i.

Collecting all these k equations gives rise to a system of k (vector) equations in the k unknowns y_i, $i = 1, \ldots, k$, of the form

$$([a_0|A] \otimes I)\widehat{Y} - h\big([b_0|B] \otimes J\big)\nabla H(\widehat{Y}) = 0, \tag{19}$$

where

$$\widehat{Y} \equiv \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_k \end{pmatrix}, \qquad \nabla H(\widehat{Y}) \equiv \begin{pmatrix} \nabla H(y_0) \\ \nabla H(y_1) \\ \vdots \\ \nabla H(y_k) \end{pmatrix},$$

I is the identity matrix of dimension $2m$, while the two k vectors a_0 and b_0 and the $k \times k$ matrices A and B contain the coefficients that form the linear combination of the vectors y_i and $J\nabla H(y_i)$ respectively. System (19) represents the standard form of a block-BVM (refer to [3] for the general theory on Boundary Value Methods). The columns a_0 and b_0 have been explicitly indicated to emphasize that they will multiply the known vector y_0; pulling them out of the coefficient matrices yields

$$(A \otimes I)Y - h\big(B \otimes J\big)\nabla H(Y) = -a_0 \otimes y_0 + h b_0 \otimes J\nabla H(y_0), \tag{20}$$

where all the known terms have been moved to the right hand side and the block-vector Y now only contains unknown data: $Y^T = [y_1^T, \ldots, y_k^T]$.

Though a systematic study of the convergence properties of these methods and their appropriate implementation will be carried out in a different paper, in the next sections we use the technique described above to derive the explicit formulation of a number of methods, compute their order, and finally apply them to solve some test problems.

We conclude this section by recalling how the block-BVM (19) may be recast in Runge-Kutta form. First of all we see that the sums over each row of the matrix $[a_0|A]$ are all null. This follows from the obvious fact that each divided difference with at least two arguments computed along the constant values $z_i = 1$ is null. Hence we have $-A^{-1}a_0 = [1, \ldots, 1]^T \equiv e$ (the pre-consistency condition), which allows us to recast (20) in the form

$$\widehat{Y} = e \otimes y_0 + h \begin{pmatrix} 0 \ldots\ldots 0 \\ A^{-1}[b_0|B] \end{pmatrix} \otimes J\nabla H(\widehat{Y}). \tag{21}$$

Equation (21) represents a $(k+1)$-stages Runge-Kutta method with the coefficients b_i, $i = 0, \ldots, k$, taken from the last row of the matrix $A^{-1}[b_0|B]$.

4 Extended LobattoIIIA methods and numerical tests

In this section we derive some methods of order 2 and 4 obtained by choosing $\sigma(c)$ of degree 1 and 2. Since the points $\sigma(0) = y_0$ and $\sigma(1) = y_k$ have been included in the vector of stages, we can look at the following methods as extensions of LobattoIIIA formulae, because they become the standard LobattoIIIA methods when the set of silent stages is empty. In a sense, each one of the new methods listed below is generated by an underlying Lobatto formula.[5] Of course, one can derive arbitrary high order conservative methods by acting on both the degree of σ and the number of silent stages. Clearly, it turns out to be advantageous to choose a Lobatto distribution for all the abscissae c_i in the interval $[0, 1]$, since such choice maximizes the degree of precision of the underlying quadrature formula.

In this paper we will not discuss about how the implementation of the derived methods should be carried out, but it is worth observing that, by the very definition, all the silent stages may be expressed in terms of the fundamental stages so that the resulting nonlinear system associated to the method has dimension independent on the number of silent stages introduced.

4.1 s-stages Trapezoidal Methods

These methods correspond to the choice $\deg(\sigma) = 1$ and an arbitrary number of silent stages. They have been introduced and described in [8] where they have been applied to the polynomial pendulum equation for several degrees. The associated Butcher

[5] The algebraic and topological link that there exists between the extended collocation methods and their generating formulae will be object of a different study.

array is

$$\begin{array}{c|ccccc} 0 & 0 & 0 & \dots & \dots & 0 \\ c_1 & c_1b_0 & c_1b_1 & \dots & \dots & c_1b_k \\ c_2 & c_2b_0 & c_2b_1 & \dots & \dots & c_2b_k \\ \vdots & \vdots & \vdots & & & \vdots \\ c_{k-1} & c_{k-1}b_0 & c_{k-1}b_1 & \dots & \dots & c_{k-1}b_k \\ 1 & b_0 & b_1 & \dots & \dots & b_k \\ \hline & b_0 & b_1 & \dots & \dots & b_k \end{array} = \begin{array}{c|c} c & c\,b^T \\ \hline & b^T \end{array}.$$

It is easily seen that each method under consideration is symmetric. When $k = 2$ we obtain the trapezoidal method. For $k = 3$ and $k = 5$, we obtain the methods

$$y_1 = y_0 + \frac{h}{6}\left(f(y_0) + 4f(\frac{y_0 + y_1}{2}) + f(y_1)\right) \tag{22}$$

and, after setting $c_0 = 0$, $c_1 = (\frac{1}{2} - \frac{1}{14}\sqrt{21})$, $c_2 = \frac{1}{2}$, $c_3 = (\frac{1}{2} + \frac{1}{14}\sqrt{21})$, $c_4 = 1$,

$$y_1 = y_0 + \frac{h}{180}\left(9f(y_0) + 49f(c_2y_0 + c_1y_1) + 64f(\frac{y_0 + y_1}{2}) + 49f(c_1y_0 + c_2y_1) + 9f(y_1)\right), \tag{23}$$

where we have removed the explicit presence of the silent stages by replacing their expression in terms of the first stage y_0, and the last stage approximating the solution in $t_0 + h$ that now we denote by y_1. When applied to $y' = f(t)$, the above schemes become the Lobatto quadrature formulae of order 4 and 8 (and therefore with degree of precision 3 and 7) respectively. On the other hand, when applied to general ODE problems, their order falls down to two. These formulae look similar to Mono-Impicit Runge-Kutta schemes in that they are implicit only in y_1 (see for example [4]).

The pictures in Figure 3 refer to the application of the 3-stages Trapezoidal method (22) to the Fermi-Pasta-Ulam problem (3). The underlying quadrature rule is the Simpson formula that has degree of precision 3. Consequently, on the basis of formula (12), such method is appropriate for the conservation of Hamiltonian polynomial functions of degree up to four, as confirmed by the right picture that reports the quantity $H(y_n)$.

The aim of the second experiment is to show how these methods, under suitable conditions, can provide a *practical conservation* of the Hamiltonian function even when $H(y)$ is not a polynomial. This is true in all the cases when the Hamiltonian function $H(y)$ may be approximated by a polynomial $\widetilde{H}(y)$ within a given tolerance ε in a set $\mathcal{D}$ of the phase space that contains the level curve $H(y) = H(y_0)$:

$$|H(y) - \widetilde{H}(y)| < \varepsilon, \qquad \text{for all } y \in \mathcal{D}. \tag{24}$$

It is clear that adding a suitable number of silent stages assures the conservation of the energy function $\widetilde{H}(y)$. Hence, from (24) it follows that $|H(y_n) - H(y_0)| < \varepsilon$ independently of the stepsize h used. If ε matches the machine precision, the computer

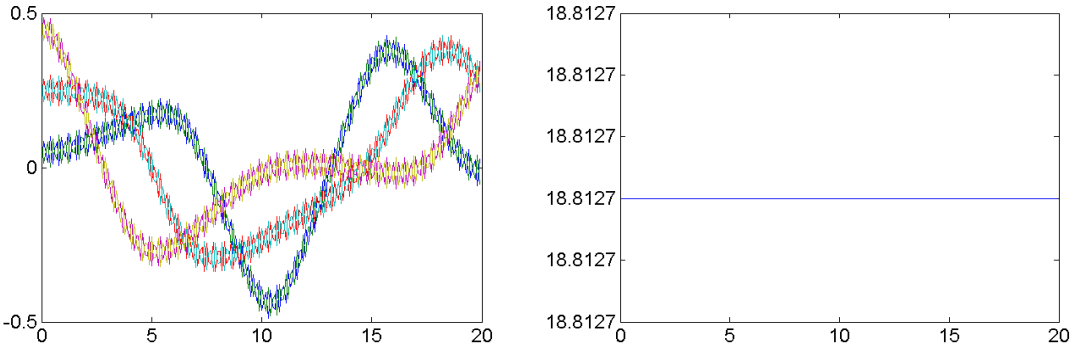

Figure 3: Application of the 3-stages Trapezoidal Method to the solution of problem (3) with stepsize $h = 1/10$, number of points $n = 200$ and initial condition $p_i = 0$ and $q_i = (i-1)/10$, for $i = 1, \ldots, 6$. In particular the left picture reports the components $q_i,\ i = 1, \ldots, 6$ of the numerical solution, while the central picture confirms the conservation of the Hamiltonian function along the numerical solution: $H(y_n) = H(y_0)$.

makes no practical difference between the Hamiltonian function and its polynomial approximation. This is illustrated in Figure 4 which displays the relative error in the numerical Hamiltonian function $H(y_n)$ computed after applying the s-stages Trapezoidal Methods corresponding to the values $s = 2,\ 3,\ 5,\ 7$, to the nonlinear pendulum equation (4).

4.2 Extended Lobatto IIIA methods of order four

These methods correspond to the choice $\deg(\sigma) = 2$ and an arbitrary number of silent stages defined on a set of abscissae c_i chosen according to a Lobatto distribution in the interval $[0,\ 1]$.[6] The quadratic curve $\sigma(c)$ is determined by the tree fundamental stages corresponding to the abscissae $c_0 = 0$, $c_{(k+1)/2} = \frac{1}{2}$ and $c_k = 1$.

To better elucidate the argument presented in the previous section, we repeat the steps to devise the method in this class defined on 5 stages: 3 fundamental stages y_0, y_2 and y_4 corresponding to the abscissae $c_0 = 0$, $c_2 = \frac{1}{2}$ and $c_4 = 1$, and 2 silent stages y_1 and y_3 defined on the nodes $c_1 = (\frac{1}{2} - \frac{1}{14}\sqrt{21})$ and $c_3 = (\frac{1}{2} + \frac{1}{14}\sqrt{21})$ (see the left picture of Figure 5). On the basis of (12) this method is appropriate for the conservation of the energy for all Hamiltonian polynomial functions with degree up to four. Therefore, in this example we assume $\deg(H(p, q)) \leq 4$.

The interpolation conditions $\sigma(0) = y_0$, $\sigma(\frac{1}{2}) = y_2$, and $\sigma(1) = y_4$ yield, after

[6] An extended collocation method of order four with 5 stages corresponding to a uniform distribution of the abscissae c_i, $i = 1, \ldots, 5$, has been presented in [9].

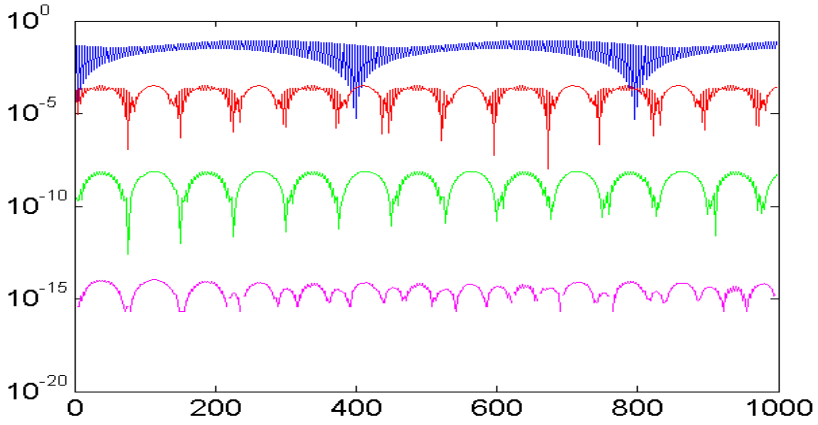

Figure 4: Relative error $|H(y_n) - H(y_0)|/|H(y_0)|$ of the numerical Hamiltonian function related to the application of four s-stages Trapezoidal Methods to the nonlinear pendulum (4), with stepsize $h = 1$, number of points $n = 1000$ and initial condition $[p_0, q_0] = [1/2, \pi/2]$. The methods differ for the number of silent stages introduced that is, starting from the top plot, 0 ($s = 2$, the trapezoidal method), 1 ($s = 3$, method (22)), 3 ($s = 5$, method (23)) and 5 ($s = 7$). Increasing the number of silent stages results in a significant reduction of the error, independently of the choice of the stepsize h.

ordering the nodes as 0, 1 and 1/2,

$$\sigma(c) = y_0 + (y_4 - y_0)c + 2(y_4 - 2y_2 + y_0)c(c-1).$$

Consequently, we define the two additional stages y_1 and y_3 as

$$y_1 = \sigma(c_1) = \frac{1}{14}(3+\sqrt{21})y_0 + \frac{4}{7}y_2 + \frac{1}{14}(3-\sqrt{21})y_4 \tag{25}$$

and

$$y_3 = \sigma(c_3) = \frac{1}{14}(3-\sqrt{21})y_0 + \frac{4}{7}y_2 + \frac{1}{14}(3+\sqrt{21})y_4. \tag{26}$$

The line integral along the curve σ is:

$$\begin{aligned} H(y_4) - H(y_0) &= \int_{y_0 \to y_4} \nabla H(y) \cdot dy = \int_0^1 (\dot{\sigma}(c))^T \nabla H(\sigma(c))\, \mathrm{d}c \\ &= (y_4 - y_0)^T \int_0^1 \nabla H(\sigma(c))\, \mathrm{d}c + 2(y_4 - 2y_2 + y_0)^T \int_0^1 (2c-1)\nabla H(\sigma(c))\, \mathrm{d}c. \end{aligned}$$

The integrand has degree less than or equal to 7 and therefore it is exactly computed by the Lobatto quadrature formula with 5 nodes which, referred to the quadrature

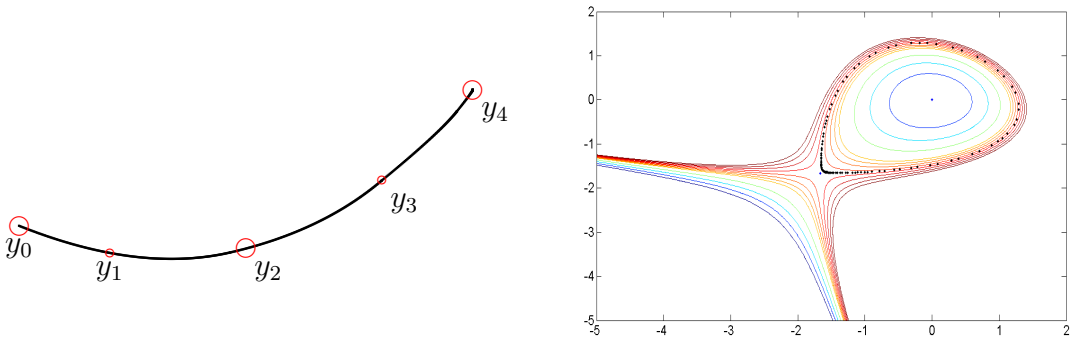

Figure 5: Left picture: the path $\sigma(c)$ is the quadratic curve that interpolates the points y_0 y_2 and y_4 (big circles). The remaining stages, in this example y_1 and y_3 (small circles), are to be selected on σ in order to improve the degree of the underlying quadrature formula. Right picture: the block-BVM (29)-(30) applied to the test problem (7) provides a numerical solution (p_n, q_n) lying on the theoretical orbit in the phase plane. Compare with Figure 2.

problem $y' = f(t)$, $t \in [0,1]$, $y(0) = y_0$, reads:

$$y_4 = y_0 + \sum_{i=0}^{4} b_i f(c_i), \qquad \text{with} \qquad [b_0, b_1, \ldots, b_4] = \left[\frac{1}{20}, \frac{49}{180}, \frac{16}{45}, \frac{49}{180}, \frac{1}{20}\right].$$

As a consequence we have:

$$H(y_4) - H(y_0) = (y_4 - y_0)^T \sum_{i=0}^{4} b_i \nabla H(\gamma(c_i)) + 2(y_4 - 2y_2 + y_0)^T \sum_{i=0}^{4} b_i (2c_i - 1) \nabla H(\gamma(c_i)).$$

Requiring that $H(y_4) = H(y_0)$ results in the following two conditions:

$$\begin{cases} (y_4 - y_0)^T \displaystyle\sum_{i=0}^{4} b_i \nabla H(\gamma(c_i)) = 0, \\ (y_4 - 2y_2 + y_0)^T \displaystyle\sum_{i=0}^{4} b_i (2c_i - 1) \nabla H(\gamma(c_i)) = 0. \end{cases}$$

As seen in the previous section, they come out from assuming that the stages satisfy the following two linear multistep formulae

$$y_4 - y_0 = h \sum_{i=0}^{4} b_i J \nabla H(\gamma(c_i)) \tag{27}$$

and

$$y_4 - 2y_2 + y_0 = h \sum_{i=0}^{4} \eta b_i (2c_i - 1) J \nabla H(\gamma(c_i)). \tag{28}$$

Choosing $\eta = \frac{3}{2}$ maximizes the order of formula (28). The resulting conservative block-BVM is the collection of linear multistep formulae (27), (28), (25) and (26). Referring to the notation in (19), it is defined by the following coefficient matrices:

$$[a_0|A] = \left(\begin{array}{c|cccc} -\frac{1}{14}(\sqrt{21}+3) & 1 & -\frac{4}{7} & 0 & \frac{1}{14}(\sqrt{21}-3) \\ 1 & 0 & -2 & 0 & 1 \\ \frac{1}{14}(\sqrt{21}-3) & 0 & -\frac{4}{7} & 1 & -\frac{1}{14}(\sqrt{21}+3) \\ -1 & 0 & 0 & 0 & 1 \end{array}\right) \tag{29}$$

and

$$[b_0|B] = \left(\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ -\frac{3}{40} & -\frac{7}{120}\sqrt{21} & 0 & \frac{7}{120}\sqrt{21} & \frac{3}{40} \\ 0 & 0 & 0 & 0 & 0 \\ \frac{1}{20} & \frac{49}{180} & \frac{16}{45} & \frac{49}{180} & \frac{1}{20}. \end{array}\right) \tag{30}$$

Written as a Runge-Kutta formula, this method is defined by means of the following Butcher tableau:

0	0	0	0	0	0
$\frac{1}{2}-\frac{1}{14}\sqrt{21}$	$\frac{13}{280}-\frac{1}{280}\sqrt{21}$	$\frac{49}{360}-\frac{1}{360}\sqrt{21}$	$\frac{8}{45}-\frac{8}{315}\sqrt{21}$	$\frac{49}{360}-\frac{13}{360}\sqrt{21}$	$\frac{1}{280}-\frac{1}{280}\sqrt{21}$
$\frac{1}{2}$	$\frac{1}{16}$	$\frac{49}{360}+\frac{7}{240}\sqrt{21}$	$\frac{8}{45}$	$\frac{49}{360}-\frac{7}{240}\sqrt{21}$	$-\frac{1}{80}$
$\frac{1}{2}+\frac{1}{14}\sqrt{21}$	$\frac{13}{280}+\frac{1}{280}\sqrt{21}$	$\frac{49}{360}+\frac{13}{360}\sqrt{21}$	$\frac{8}{45}+\frac{8}{315}\sqrt{21}$	$\frac{49}{360}+\frac{1}{360}\sqrt{21}$	$\frac{1}{280}+\frac{1}{280}\sqrt{21}$
1	$\frac{1}{20}$	$\frac{49}{180}$	$\frac{16}{45}$	$\frac{49}{180}$	$\frac{1}{20}$
	$\frac{1}{20}$	$\frac{49}{180}$	$\frac{16}{45}$	$\frac{49}{180}$	$\frac{1}{20}$

The order four can be stated by checking, for example, the simplifying conditions listed in [7, Theorem 5.1, page 71]. This method has been applied to the test problem (7): the absence of oscillation of $H(y_n) - H(y_0)$ prevent the numerical orbit to escape from the stability region (see the right picture of Figure 5).

By an analogous computation we have derived the order 4 method with 7 stages, suitable for the conservation of Hamiltonian polynomials of degree up to 6 and applied it to the test problem (6) (see Figure 6).

Conlusions

We have derived new one step symmetric methods that, applied to Hamiltonian problems, provide a precise conservation of the Hamiltonian function in the case where this is a polynomial in the momenta p and in the generalized coordinates q. The definition of such methods exploits the properties of the so called *discrete line integral*,

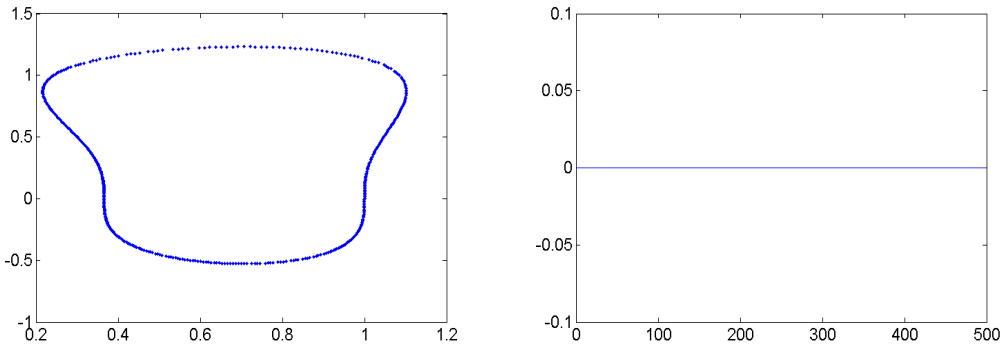

Figure 6: The extended LobattoIIIA formula of order 4 with 7 stages has been applied to the problem defined by the 6-degree polynomial (6). We have used stepsize $h = 1$, number of points $n = 500$ and initial condition $[1, 0]$. The two pictures display the numerical orbit in the phase plane and the numerical Hamiltonian function $H(y_n)$.

which is the discrete counterpart of line integrals that represent a common tool to define and study conservative fields.

These methods are naturally set up in block-BVM form and may be interpreted as extensions of well-known formulae obtained by introducing a certain number of additional stages to get conservativeness. However, as outlined in the last section, the presence of these extra-stages does not alter the dimension of the associated nonlinear system that must be solved at each step.

A general convergence and stability theory on such methods, their appropriate implementation, and their geometrical features will be the subject of a future research [1].

References

[1] L. Brugnano, F. Iavernaro and D. Trigiante, *Analisys of Hamiltonian Boundary Value Methods for the numerical solution of polynomial Hamiltonian dynamical systems*, (in preparation).

[2] L. Brugnano and D. Trigiante, *Energy drift in the numerical integration of Hamiltonian problems*, J. Numer. Anal. Ind. Appl. Math., (in press).

[3] L. Brugnano and D. Trigiante, *Solving ODEs by Linear Multistep Initial and Boundary Value Methods*, Gordon & Breach: Amsterdam, 1998.

[4] J. R. Cash, A. Singhal, *Mono-implicit Runge-Kutta formulae for the numerical integration of stiff differential systems*, IMA J. Numer. Anal. **2** no. 2 (1982), 211–227.

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

[6] E. Hairer, C. Lubich, and G. Wanner, *Geometric numerical integration. Structure-preserving algorithms for ordinary differential equations.*, Springer Series in Computational Mathematics, **31**, Springer-Verlag, Berlin, 2002.

[7] E. Hairer, G. Wanner., *Solving Ordinary Differential Equations II. Stiff and Differential–Algebraic Equations*, Springer Series in Computational Mathematics **14**, Springer-Verlag, Berlin, 1996.

[8] F. Iavernaro and B. Pace, *s-stage trapezoidal methods for the conservation of Hamiltonian functions of polynomial type*, AIP Conf. Proc. **936** (2007), 603–606.

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

[10] F. Iavernaro and D. Trigiante, *Discrete conservative vector fields induced by the trapezoidal method*, J. Numer. Anal. Ind. Appl. Math. **1** (no. 1) (2006), 113–130.

[11] F. Iavernaro and D. Trigiante, *On some conservation properties of the trapezoidal method applied to Hamiltonian systems*, in: International Conference on Numerical Analysis and Applied Mathematics 2005, T.E. Simos et al. eds., Wiley-Vch Verlag GmbH & Co., Weinheim 2005, 254–257.