Appl. Numer. Math. 203 (2024) 113–128

Contents lists available at ScienceDirect

Applied Numerical Mathematics

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

Research Paper

Weighted least squares collocation methods

Luigi Brugnano [a], Felice Iavernaro [b,*], Ewa B. Weinmüller [c]

[a] Dipartimento di Matematica e Informatica "U. Dini", Università di Firenze, Italy
[b] Dipartimento di Matematica, Università di Bari, Italy
[c] Institute for Analysis and Scientific Computing, Vienna University of Technology, A-1040 Wien, Austria

ARTICLE INFO

Dedicated to John Butcher, on the occasion of his 91st birthday

Keywords:
Least squares collocation methods
Gauss-Legendre collocation methods
Line integral methods
Hamiltonian boundary value methods

ABSTRACT

We consider overdetermined collocation methods and propose a weighted least squares approach to derive a numerical solution. The discrete problem requires the evaluation of the Jacobian of the vector field which, however, appears in a $O(h)$ term, h being the stepsize. We show that, by neglecting this infinitesimal term, the resulting scheme becomes a low-rank Runge–Kutta method. Among the possible choices of the weights distribution, we analyze the one based on the quadrature formula underlying the collocation conditions. A few numerical illustrations are included to better elucidate the potential of the method.

1. Introduction

In this paper, we apply an overdetermined collocation method to approximate the solution of an initial value problem in ordinary differential equations. This method is based on the following principle: the ansatz for the numerical solution is chosen in such a way that the number of equations to fix the ansatz function is larger than the number of unknown coefficients. The resulting overdetermined system of equations, in general nonlinear, is then solved in the least squares sense. This approach is a regularization technique and proves useful in cases when the analytical problem is not well-posed or whose data is noisy.

Historically, one of the most prominent fields for this approach is the numerical solution of boundary value problems (BVPs). In fact, an early attempt to provide professional software in the context of BVPs can be found in [2], where the ansatz functions were the piecewise polynomial solutions smoother at mesh points. The extra smoothness means fewer solution parameters, hence overdetermined systems of algebraic equations. It has been shown there, that nonetheless, some partial super-convergence still holds. Concerning the numerical solution of two-point boundary value problems, we also mention [1,13], and related software available in [21].

The approach has then been extended to the wider class of differential-algebraic equations (DAEs) and it can be found in the COLDAE software for solving boundary value problems in DAEs and extending the well-established code COLSYS [5]. Least squares collocation is also considered in the newest research for the numerical solution of higher index DAEs [15,16] (based on the development described in [3,4]), where a related technique has been discussed. The use of least squares collocation in the context of higher index DAEs was also proposed in [20], see [17–19] for the discussion of related issues.

With this premise, we emphasize that the present paper is not directly related to the above fields but, instead, it is aimed at finding an alternative approach that, still retaining the basic accuracy properties of least squares collocation methods, do have a

[*] Corresponding author.
E-mail addresses: luigi.brugnano@unifi.it (L. Brugnano), felice.iavernaro@uniba.it (F. Iavernaro), e.weinmueller@tuwien.ac.at (E.B. Weinmüller).

https://doi.org/10.1016/j.apnum.2024.05.017
Received 19 January 2024; Received in revised form 15 May 2024; Accepted 20 May 2024

Available online 24 May 2024

much more favorable computational complexity. As a matter of fact, one main drawback in using least squares collocation is that the discrete problem involves the Jacobian of the vector field evaluated along the *internal stages*, namely the values attained by the unknown polynomial approximation at the collocation abscissae. This significantly enhances the overall computational complexity associated with the integration procedure.

However, it turns out that the Jacobian matrix appears in a term that becomes negligible as the stepsize h approaches zero. Interestingly, for h small enough, this term acts as a pertubation to a special low-rank Runge–Kutta formula named Linear Integral Method (LIM), specifically designed for the numerical integration of conservative problems [7,8,10,12]. Given that LIMs have undergone extensive research, with the development of highly efficient techniques for their implementation [11,6], it makes sense to explore the connections between the solutions provided by the least squares collocation procedure and the associated LIM.

This investigation serves as the motivation for the present work and will be confined to ordinary differential equations coupled with initial values. In fact, at this very first stage, the primary objective is to examine the extent to which the more complex structure of the discrete problem arising from the least squares approach can be effectively replaced by the much simpler formulation represented by a Runge–Kutta method, as it is suggested by the numerical results reported in the sequel.

The paper is organized as follows. In Section 2 we introduce a weighted least squares approach to solve an overdetermined collocation problem. We also show that the resulting nonlinear system may be interpreted as a perturbation of a corresponding discrete problem defining the class of line integral methods named Hamiltonian Boundary Value Methods (HBVMs). Section 3 provides an in-depth discussion of the similarities between least squares and line integral methods. In Section 4, we present convergence results and explore the degree to which the two numerical solutions closely align. Section 5 offers a set of numerical illustrations that validate the theoretical findings. Finally, in Section 6, we draw some concluding remarks.

2. Overdetermined collocation methods

For a sufficiently regular function $f : [t_0, T] \times \mathbb{R}^m \to \mathbb{R}^m$, we consider the numerical approximation of the solution of the Initial Value Problem (IVP)

$$y' = f(t, y), \qquad y(t_0) = y_0 \in \mathbb{R}^m, \qquad t \in [t_0, T], \tag{1}$$

on the interval $[t_0, t_0 + h]$ (h stands for the stepsize) by means of a polynomial $u(t)$ of degree at most s obtained by imposing the following $k + 1$ collocation conditions:

$$\begin{cases} u(t_0) = y_0, \\ u'(t_0 + c_i h) = f(t_0 + c_i h, u(t_0 + c_i h)), \quad i = 1, \dots, k, \end{cases} \tag{2}$$

where $0 \le c_1 < c_2 < \cdots < c_k \le 1$ are the collocation abscissae. When $k > s$, system (2) is overdetermined and we handle it by solving a related weighted least squared problem.

Without loss of generality and to simplify the notation, we assume that the IVP (1) is scalar ($m = 1$) and autonomous. Denoting by Π_s the vector space of polynomials of degree at most s, for a given distribution of (positive) weights $\{w_i\}_{i=1,\dots,k}$ satisfying the normalization condition $\sum_{i=1}^k w_i = 1$, the approximating polynomial $u^* \in \Pi_s$ is defined by[1]

$$u^* = \underset{u \in \Pi_s, u(t_0)=y_0}{\arg\min} \sum_{i=1}^{k} w_i \left(u'(t_0 + c_i h) - f(u(t_0 + c_i h)) \right)^2. \tag{3}$$

We then advance the solution by setting $y_1 = u^*(t_0 + h)$. Note that we are imposing the quite natural constraint $u(t_0) = y_0$, thus the least squares problem only involves the collocation conditions on the derivative of the polynomial u.

In order to determine u^*, we represent $u'(t_0 + ch)$ using an arbitrary basis $\{P_j(c)\}_{j=0,\dots,s-1}$, of Π_{s-1}:

$$u'(t_0 + ch) = \sum_{j=0}^{s-1} P_j(c) \gamma_j, \tag{4}$$

from which, after integrating both sides of (4) in the interval $[0, c]$, we obtain

$$u(t_0 + ch) = y_0 + h \sum_{j=0}^{s-1} \int_0^c P_j(x) \mathrm{d}x \, \gamma_j. \tag{5}$$

Setting $\gamma = (\gamma_0, \dots, \gamma_{s-1})^\top$, and substituting (4) and (5) into (3), we need to find the stationary points of the scalar function

$$F(\gamma) = \sum_{i=1}^{k} w_i \left[\sum_{j=0}^{s-1} P_j(c_i) \gamma_j - f\left(y_0 + h \sum_{j=0}^{s-1} \int_0^{c_i} P_j(x) \mathrm{d}x \, \gamma_j \right) \right]^2.$$

[1] We recall that arg min denotes the solution of the minimization problem.

Differentiating with respect to γ_ℓ yields, for $\ell = 0, \dots, s-1$, the equations

$$0 = \frac{\partial F}{\partial \gamma_\ell} = 2\sum_{i=1}^k w_i \left[\sum_{j=0}^{s-1} P_j(c_i)\gamma_j - f\Big(y_0 + h\sum_{j=0}^{s-1}\int_0^{c_i} P_j(x)\mathrm{d}x\,\gamma_j\Big)\right] \cdot \left(P_\ell(c_i) - h\int_0^{c_i} P_\ell(x)\mathrm{d}x\, f'\Big(y_0 + h\sum_{j=0}^{s-1}\int_0^{c_i} P_j(x)\mathrm{d}x\,\gamma_j\Big)\right),$$

that may be recast as

$$\sum_{j=0}^{s-1}\left(\sum_{i=1}^k w_i P_j(c_i)P_\ell(c_i)\right)\gamma_j = \sum_{i=1}^k w_i P_\ell(c_i) f\Big(y_0 + h\sum_{j=0}^{s-1}\int_0^{c_i} P_j(x)\mathrm{d}x\,\gamma_j\Big) + h\sum_{i=1}^k w_i \int_0^{c_i} P_\ell(x)\mathrm{d}x\, g_i(\gamma), \tag{6}$$

where, for $i = 1, \dots, k$,

$$g_i(\gamma) = f'\Big(y_0 + h\sum_{j=0}^{s-1}\int_0^{c_i} P_j(x)\mathrm{d}x\,\gamma_j\Big)\left[\sum_{j=0}^{s-1} P_j(c_i)\gamma_j - f\Big(y_0 + h\sum_{j=0}^{s-1}\int_0^{c_i} P_j(x)\mathrm{d}x\,\gamma_j\Big)\right]. \tag{7}$$

We introduce the matrices

$$\mathcal{P}_s = \begin{pmatrix} P_0(c_1) & \dots & P_{s-1}(c_1) \\ \vdots & & \vdots \\ P_0(c_k) & \dots & P_{s-1}(c_k)\end{pmatrix}, \quad \mathcal{I}_s = \begin{pmatrix} \int_0^{c_1} P_0(x)\mathrm{d}x & \dots & \int_0^{c_1} P_{s-1}(x)\mathrm{d}x \\ \vdots & & \vdots \\ \int_0^{c_k} P_0(x)\mathrm{d}x & \dots & \int_0^{c_k} P_{s-1}(x)\mathrm{d}x \end{pmatrix} \in \mathbb{R}^{k\times s},$$

$\Omega = \mathrm{diag}(w_1, \dots, w_k)$ and the vectors $g(\gamma) = (g_1(\gamma), \dots, g_k(\gamma))^\top$ and $e = (1, \dots, 1)^\top \in \mathbb{R}^k$. Then, in vector form, system (6) reads

$$\mathcal{P}_s^\top \Omega \mathcal{P}_s \gamma = \mathcal{P}_s^\top \Omega f(e\otimes y_0 + h\mathcal{I}_s\gamma) + h\mathcal{I}_s^\top \Omega g(\gamma) \tag{8}$$

with the obvious meaning of $f(e\otimes y_0 + h\mathcal{I}_s\gamma)$ and

$$g(\gamma) = \mathrm{diag}(f'(e\otimes y_0 + h\mathcal{I}_s\gamma))\left(\mathcal{P}_s\gamma - f(e\otimes y_0 + h\mathcal{I}_s\gamma)\right). \tag{9}$$

After solving (8) in the unknown vector γ, we set

$$y_1 = u^*(t_0 + h) = y_0 + h\sum_{j=0}^{s-1}\int_0^1 P_j(x)\mathrm{d}x\gamma_j \tag{10}$$

and repeat the whole procedure to advance the solution in time.

We note that solving (8) requires the evaluation of the Jacobian matrix of the vector field appearing in $g(\gamma)$. The discrete problem may be considerably simplified by neglecting the last (infinitesimal) term in (8), thus obtaining the much simpler system

$$\mathcal{P}_s^\top \Omega \mathcal{P}_s \gamma = \mathcal{P}_s^\top \Omega f(e\otimes y_0 + h\mathcal{I}_s\gamma), \tag{11}$$

that, for an appropriate choice of weights, defines an integrator in the family of line integral methods (special low-rank Runge–Kutta methods), as is illustrated in the next section.

Remark 1. It is worth noting that, for pure quadrature problems $y' = f(t)$, the term $g(\gamma)$ vanishes, so that (8) and (11) reduce to the standard (linear) least squares problem

$$\mathcal{P}_s^\top \Omega \mathcal{P}_s \gamma = \mathcal{P}_s^\top \Omega f(t_0 + c_1 h, \dots, t_0 + c_k h)^\top.$$

This means that the residual $\mathcal{P}_s\gamma - f(t_0 + c_1 h, \dots, t_0 + c_k h)^\top$ is projected onto the orthogonal of the vector space spanned by the columns of $\mathcal{P}_s$ which, in turn, are the projection of the polynomial basis $\{P_j\}$ along the mesh $(c_1, \dots, c_k)$.

The resulting integrator (11) is indeed a Runge–Kutta method with rank-deficient coefficient matrix. In order to obtain the Runge–Kutta formulation, we introduce the stage vector $Y = (u(t_0 + c_1 h), \dots, u(t_0 + c_k h))^\top$ containing the evaluations of the polynomial at the internal points $t_0 + c_i h$, $i = 1, \dots, k$. Exploiting (5) and (11) we obtain

$$Y = e\otimes y_0 + h\mathcal{I}_s\gamma = e\otimes y_0 + h\mathcal{I}_s(\mathcal{P}_s^\top\Omega\mathcal{P}_s)^{-1}\mathcal{P}_s^\top\Omega f(Y). \tag{12}$$

Hence, the Butcher tableau associated with the method reads

$$\begin{array}{c|c} c & A \\ \hline & b^\top \end{array} \tag{13}$$

where (see also (10))

$$\begin{aligned} c &= (c_1, \dots, c_k)^\top, \\ A &= \mathcal{I}_s(\mathcal{P}_s^\top \Omega \mathcal{P}_s)^{-1}\mathcal{P}_s^\top \Omega \in \mathbb{R}^{k\times k}, \\ b^\top &= \left(\int_0^1 P_0(x)\mathrm{d}x, \dots, \int_0^1 P_{s-1}(x)\mathrm{d}x\right)^\top (\mathcal{P}_s^\top \Omega \mathcal{P}_s)^{-1}\mathcal{P}_s^\top \Omega \in \mathbb{R}^{1\times k}. \end{aligned} \tag{14}$$

Note that the k-dimensional coefficient matrix A has rank s independently of the number k of internal abscissae. For this reason, it is preferable to deal with system (11) in the unknown γ, whose dimension is indeed s, rather than solving the more classical system (12) in the unknown vector of stages Y.

3. Link with line integral methods

Line integral methods (LIMs) arise in the context of geometric integration and are particularly designed to simulate the dynamics of Hamiltonian systems for long times [7]. Indeed, their main feature is to preserve the energy function of canonical Hamiltonian systems defined by the vector field (hereafter m is an even integer)

$$f(y) = J\nabla H(y), \qquad J = \begin{pmatrix} & I_{m/2} \\ -I_{m/2} & \end{pmatrix},$$

where I_r denotes the identity matrix of dimension r. The coefficients of the polynomial $u(t)$ defined by (5), approximating the true solution $y(t)$ in the interval $[t_0, t_0 + h]$, are computed by requiring the related line integral to vanish:

$$\begin{aligned} H(y_1) - H(y_0) &= H(u(t_0+h)) - H(u(t_0)) = \int_{t_0}^{t_0+h} u'(t)^\top \nabla H(u(t))\mathrm{d}t \\ &= h\int_0^1 u'(t_0+ch)^\top \nabla H(u(t_0+ch))\mathrm{d}c \\ &= h\sum_{j=0}^{s-1}\gamma_j^\top \int_0^1 P_j(c)\nabla H(u(t_0+ch))\mathrm{d}c. \end{aligned}$$

However, in order to derive a numerical integrator, the involved integrals are approximated by a suitable quadrature formula based on the nodes and weights (c_i, w_i), $i = 1, \dots, k$, for a suitable $k \ge s$, having order q (i.e., exact for polynomial integrands of order $q-1$), which we shall assume at least (and usually much greater than) $2s$[2]:

$$\begin{aligned} &\int_0^1 P_j(c)\nabla H(u(t_0+ch))\mathrm{d}c \\ &= \sum_{i=1}^k w_i P_j(c_i)\nabla H(y_0 + h\sum_{\ell=0}^{s-1}\int_0^{c_i} P_\ell(x)\mathrm{d}x\,\gamma_\ell) + E_j(h), \qquad j = 0, \dots, s-1, \end{aligned}$$

where

$$E_j(h) = \begin{cases} 0, & \text{if } H \in \Pi_\nu, \text{ with } \nu \le q/s, \\ O(h^{q-j}), & \text{otherwise.} \end{cases}$$

Consequently, one obtains:

$$\begin{aligned} &H(y_1) - H(y_0) \\ &= h\sum_{j=0}^{s-1}\gamma_j^\top \left[\sum_{i=1}^k w_i P_j(c_i)\nabla H(y_0 + h\sum_{\ell=0}^{s-1}\int_0^{c_i} P_\ell(x)\mathrm{d}x\,\gamma_\ell) + E_j(h)\right] \end{aligned}$$

[2] As an example, by choosing a Gauss-Legendre formula, $q = 2k$.

$$= h\gamma^\top \left[(P_s^\top \Omega)\otimes I_m \cdot \nabla H(e\otimes y_0 + h\mathcal{I}_s \otimes I_m \cdot \gamma)\right] + E(h), \tag{15}$$

where we have denoted

$$\gamma = \left(\, \gamma_0^\top, \quad \dots, \quad \gamma_{s-1}^\top \,\right)^\top \qquad \text{and} \qquad E(h) = h\sum_{j=0}^{s-1} \gamma_j^\top E_j(h).$$

Setting

$$\gamma = (SP_s^\top \Omega)\otimes I_m \cdot f(e\otimes y_0 + h\mathcal{I}_s\otimes I_m\cdot\gamma) \tag{16}$$

where S is a positive definite symmetric matrix, one then obtains [12]

$$E(h) = \begin{cases} 0, & \text{if } H\in\Pi_\nu, \text{ with } \nu \le q/s, \\ O(h^{q+1}), & \text{otherwise.} \end{cases}$$

As is clear, by choosing k large enough, the order q of the quadrature can be increased so that $E(h)$ vanishes, in the polynomial case, or becomes smaller than the round-off error level (hence negligible) otherwise. As a result, (an at least practical) energy conservation is gained, since

$$\begin{aligned}\gamma &= (SP_s^\top\Omega)\otimes I_m\cdot(I_k\otimes J)\cdot\nabla H(e\otimes y_0 + h\mathcal{I}_s\otimes I_m\cdot\gamma)\\ &= (SP_s^\top\Omega)\otimes J\cdot\nabla H(e\otimes y_0 + h\mathcal{I}_s\otimes I_m\cdot\gamma)\\ &= (S\otimes J)\cdot((P_s^\top\Omega)\otimes I_m)\cdot\nabla H(e\otimes y_0 + h\mathcal{I}_s\otimes I_m\cdot\gamma)\end{aligned}$$

and, therefore (see (15)),

$$\begin{aligned}H(y_1) - H(y_0) \;=\; & h\left[(P_s^\top\Omega)\otimes I_m\cdot\nabla H(e\otimes y_0 + h\mathcal{I}_s\otimes I_m\cdot\gamma)\right]^\top\cdot(S\otimes J)^\top\\ & \cdot\left[(P_s^\top\Omega)\otimes I_m\cdot\nabla H(e\otimes y_0 + h\mathcal{I}_s\otimes I_m\cdot\gamma)\right] + E(h) \;=\; E(h),\end{aligned}$$

due to the skew-symmetry of matrix $S\otimes J$.

By setting $S = (\mathcal{P}_s^\top\Omega\mathcal{P}_s)^{-1}$ we obtain the method (11), which was derived as a simplification of the least squares problem (3). On the other hand, the most simple choice for S, which has been generally adopted in the context of line integral methods, is the identity matrix I_s. In this respect, we observe that the two choices become identical under the following assumptions:

(A_1) $(c_i, w_i)_{i=1,\dots,k}$ defines a quadrature formula of order $q \ge 2s$ (i.e. it is exact for polynomials of degree at least $2s-1$);

(A_2) for $j = 0,\dots,s-1$, the polynomials $P_j(c)$ are orthonormal on $[0,1]$, that is they are the Legendre polynomials scaled and normalized in the interval $[0,1]$:

$$\int_0^1 P_i(c)P_j(c)\mathrm{d}c = \delta_{ij}, \qquad i,j = 0,\dots,s-1,$$

where δ_{ij} is the Kronecker symbol.

In fact, in such a case, one has

$$(\mathcal{P}_s^\top\Omega\mathcal{P}_s)_{ij} = \sum_{\ell=1}^k w_\ell P_i(c_\ell)P_j(c_\ell) = \int_0^1 P_i(c)P_j(c)\mathrm{d}c = \delta_{ij}.$$

Once this relationship has been clarified, we wish to inspect how close are the problems (8) and (11) and whether we can transfer the convergence properties of LIMs (whose theory is well-established) to the related least squares collocation methods. This question is addressed in the next section.

4. Study of convergence

Hereafter, we denote by $\bar\gamma$ the solution of system (11), γ^* the solution of system (8) yielding the least squares approximation, and $\bar u(t_0 + ch)$ and $u^*(t_0 + ch)$ the associated polynomials, respectively (see (5) and (4)). We continue handling the scalar problem ($m = 1$), to simplify the notation. Their existence may be derived, as usual, by the fixed point theorem applied respectively to the schemes

$$\gamma^{(n+1)} = \Phi(\gamma^{(n)}) := (\mathcal{P}_s^\top\Omega\mathcal{P}_s)^{-1}\mathcal{P}_s^\top\Omega f(e\otimes y_0 + h\mathcal{I}_s\gamma^{(n)}), \tag{17}$$

$$\gamma^{(n+1)} = \Psi(\gamma^{(n)}) := \Phi(\gamma^{(n)}) + h(\mathcal{P}_s^\top\Omega\mathcal{P}_s)^{-1}\mathcal{I}_s^\top\Omega g(\gamma^{(n)}), \tag{18}$$

$\gamma^{(0)} \in \mathbb{R}^s$ being a given initial guess and assuming that f and f' are bounded and Lipschitz continuous. In fact, a direct computation shows that, choosing h small enough, both Φ and Ψ are contractions on $\mathbb{R}^s$. In the sequel, we denote by L the Lipschitz constant of the vector field $f(y)$.

Considering (8) as a perturbation of (11), we explore the closeness of the solutions of the two methods in order to assess the extent to which the latter may be considered a simplified version of the original one. To this end, we assume that the line integral method defined by (11) or, equivalently, by (13)-(14) has stage order p, which means that $h_0 > 0$ exists such that, for $h \le h_0$,

$$\max_{1\le i\le k} |\bar{u}(t_0 + c_i h) - y(t_0 + c_i h)| \le C_0 h^{p+1}, \tag{19}$$

and consequently

$$\max_{1\le i\le k} |\bar{u}'(t_0 + c_i h) - y'(t_0 + c_i h)| \le C_1 h^{p}, \tag{20}$$

with C_0, C_1 positive constants independent of h.[3]

Lemma 1. *The residual vector $R(\gamma) = \mathcal{P}_s\gamma - f(e \otimes y_0 + h\mathcal{I}_s\gamma)$ satisfies*

$$||R(\bar{\gamma})||_\infty = O(h^p), \tag{21}$$

$$||R(\gamma^*)||_\infty = O(h^p). \tag{22}$$

Proof. Considering (5) and (4), we obtain for $h \le h_0$,

$$\begin{aligned} ||R(\bar{\gamma})||_\infty &= \max_{1\le i\le k} |\bar{u}'(t_0 + c_i h) - f(\bar{u}(t_0 + c_i h))| \\ &\le \max_{1\le i\le k} |\bar{u}'(t_0 + c_i h) - y'(t_0 + c_i h)| + \max_{1\le i\le k} |f(y(t_0 + c_i h)) - f(\bar{u}(t_0 + c_i h))| \\ &\le C_1 h^p + L \max_{1\le i\le k} |y(t_0 + c_i h) - \bar{u}(t_0 + c_i h)| \le C_1 h^p + C_0 L h^{p+1}. \end{aligned}$$

Hence (21) holds. Concerning (22), we first observe that the weighted norm in (3) may be recast, in terms of the residual vector $R(\gamma)$, as $||\Omega^{1/2} R(\gamma)||_2$. Therefore, from the equivalence of vector norms, for suitable constants $\kappa_1, \kappa_2 > 0$, and from (20), we conclude

$$||R(\gamma^*)||_\infty \le \kappa_1 ||\Omega^{1/2} R(\gamma^*)||_2 \le \kappa_1 ||\Omega^{1/2} R(\bar{\gamma})||_2 \le \kappa_2 ||R(\bar{\gamma})||_\infty = O(h^p). \quad \square$$

Theorem 1. *The approximating polynomials $\bar{u}(t_0 + ch)$ and $u^*(t_0 + ch)$ satisfy the relation*

$$\max_{0\le c\le 1} |u^*(t_0 + ch) - \bar{u}(t_0 + ch)| = O(h^{p+2}). \tag{23}$$

Consequently, the least squares collocation method (8) inherits the same stage order as (11), namely

$$\max_{1\le i\le k} |u^*(t_0 + c_i h) - y(t_0 + c_i h)| = O(h^{p+1}).$$

Proof. From (17) and (18), subtracting $\bar{\gamma} = \Phi(\bar{\gamma})$ from $\gamma^* = \Psi(\gamma^*)$ yields

$$\gamma^* - \bar{\gamma} = \Phi(\gamma^*) - \Phi(\bar{\gamma}) + h(\mathcal{P}_s^\top \Omega \mathcal{P}_s)^{-1} \mathcal{I}_s^\top \Omega g(\gamma^*).$$

From (9) and Lemma 1, we see that $g(\gamma^*) = O(h^p)$. Consequently, setting (see (17))

$$\kappa_3 = ||(\mathcal{P}_s^\top \Omega \mathcal{P}_s)^{-1} \mathcal{P}_s^\top \Omega||_\infty, \qquad \kappa_4 = ||\mathcal{I}_s||_\infty,$$

we obtain

$$\begin{aligned} ||\gamma^* - \bar{\gamma}||_\infty &\le \kappa_3 ||f(e \otimes y_0 + h\mathcal{I}_s\gamma^*) - f(e \otimes y_0 + h\mathcal{I}_s\bar{\gamma})||_\infty + O(h^{p+1}) \\ &\le h\kappa_3\kappa_4 L ||\gamma^* - \bar{\gamma}||_\infty + O(h^{p+1}) \end{aligned}$$

and consequently, for a suitable $h_0 > 0$ and for $h \le h_0$ we obtain

$$||\gamma^* - \bar{\gamma}||_\infty \le \frac{1}{1 - h\kappa_3\kappa_4 L} O(h^{p+1}). \tag{24}$$

Relation (23) is a direct consequence of (24):

[3] The bounds (19) and (20) may be extended to the whole interval $[t_0, t_0 + h]$.

$$|u^*(t_0+ch)-\bar{u}(t_0+ch)| \le h\sum_{j=0}^{s-1}\left|\int_0^c P_j(x)\mathrm{d}x\right| \cdot |\gamma_j^*-\bar{\gamma}_j| = O(h^{p+2}).$$

Finally, taking into account (5) and (19),

$$\begin{aligned}\max_{1\le i\le k}|u^*(t_0+c_ih)-y(t_0+c_ih)| &\le \max_{1\le i\le k}|u^*(t_0+c_ih)-\bar{u}(t_0+c_ih)| \\ &\quad + \max_{1\le i\le k}|\bar{u}(t_0+c_ih)-y(t_0+c_ih)| \\ &\le O(h^{p+2})+O(h^{p+1})=O(h^{p+1})\end{aligned}$$

follows. □

Remark 2. Relation (23) suggests that, for $h\to 0$, the approximations $u^*(t_0+ch)$ and $\bar{u}(t_0+ch)$ get close to each other at a faster rate than their speed of convergence to the true solution $y(t_0+ch)$, so that, for h small enough, they are expected to essentially yield the same numerical approximation.

By virtue of (21), we may conclude that $g(\bar{\gamma})=O(h^p)$ (see (9)). This suggests using $\bar{\gamma}$ as a starting guess for the iteration (18), since each element in the sequence (18) would then be a $O(h^{p+1})$ correction of the previous one. Since very efficient techniques for the solution of (11) are available (see, for example, [11,6]), this procedure would result in an easy and effective strategy for implementing least squares collocation formulae. Both aspects elucidated above are confirmed by the numerical experiments discussed in Section 5.

The previous results may be significantly improved under the assumptions (A$_1$) and (A$_2$). When working with the scaled Legendre polynomials, which form an orthonormal basis of $L_2([0,1])$, line integral methods may be equivalently derived by first considering the Fourier expansion of the vector field $f(y)$

$$f(y(t_0+ch)) = \sum_{j\ge 0} P_j(c)\gamma_j(y), \tag{25}$$

with Fourier coefficients

$$\gamma_j(y) = \int_0^1 P_j(c)f(y(t_0+ch))\mathrm{d}c, \quad j\ge 0,$$

and then defining the polynomial approximation $u(t_0+ch)$ for the true solution $y(t_0+ch)$ as the solution of the following IVP obtained by truncating the expansion (25) up to the first s terms:

$$u'(t_0+ch) = \sum_{j=0}^{s-1} P_j(c)\gamma_j(u), \qquad c\in[0,1], \qquad u(t_0)=y_0,$$

with

$$\begin{aligned}\gamma_j(u) &= \int_0^1 P_j(c)f(u(t_0+ch))\mathrm{d}c \\ &= \int_0^1 P_j(c)f\Big(y_0+h\sum_{i=0}^{s-1}\int_0^c P_i(x)\mathrm{d}x\,\gamma_i(u)\Big)\mathrm{d}c, \qquad j=0,\dots,s-1.\end{aligned} \tag{26}$$

Approximating the integrals by means of the quadrature rule (c_i,w_i) and recasting the equations (26) in vector form, we finally get

$$\gamma = \mathcal{P}_s^\top \Omega f(e\otimes y_0 + h\mathcal{I}_s\gamma), \tag{27}$$

namely system (11) after observing that $\mathcal{P}_s^\top\Omega\mathcal{P}_s = I_s$, due to assumption (A$_1$). In system (27),

$$\gamma_j = \sum_{i=1}^k w_iP_j(c_i)f(u(t_0+c_ih)) = \gamma_j(u)+\Delta_j(h),$$

with the quadrature error $\Delta_j(h)=O(h^{q-j})$, q being the order of the quadrature rule. In fact, $\Delta_j(h)$ depends on the qth derivative of the integrand function with respect to c and we can apply the following property which we state in a more general form for later use.

Lemma 2. *Let $g:\mathbb{R}\to\mathbb{R}$ be sufficiently smooth in a neighborhood of t_0, $v_j(c)$ a polynomial of degree $j\ge 0$, q a nonnegative integer and $h>0$. Then*

$$\frac{d^q}{dc^q}(v_j(c)g(t_0+ch)) = O(h^{\max\{q-j,0\}}). \tag{28}$$

Proof. Formula (28) directly follows from the general Leibniz rule for the qth derivative of a product of functions:

$$\begin{aligned}\frac{d}{dc^q}(v_j(c)g(t_0+ch)) &= \sum_{i=0}^{q}\binom{q}{i}v_j^{(i)}(c)\frac{d}{dc^{q-i}}g(t_0+ch)\\ &= \sum_{i=0}^{\min\{j,q\}}\binom{q}{i}h^{q-i}v_j^{(i)}(c)g^{(q-i)}(x)\Big|_{x=t_0+ch}. \quad \square\end{aligned}$$

Exploiting the orthogonality properties of Legendre polynomials, we can simplify the approximation y_1 to $y(t_0+h)$ as follows ($\bar{\gamma}$ denotes the solution of (27)):

$$y_1 = \bar{u}(t_0+h) = y_0 + h\sum_{j=0}^{s-1}\int_0^1 P_j(x)\mathrm{d}x\,\bar{\gamma}_j = y_0 + h\bar{\gamma}_0. \tag{29}$$

Usually, in order to maximize the order of the quadrature formula, the nodes c_i, $i=1,\dots,k$, are chosen as the roots of $P_k(c)$, the scaled Legendre polynomial of degree k, which yields the Gauss-Legendre quadrature rule of order $2k$. We note that, under this choice, assumption A_1 is satisfied if $k \geq s$.

The resulting Runge–Kutta method, depending on the two parameters k and s is denoted by HBVM(k,s) since its first instance arose in the context of Hamiltonian Boundary Value Methods [7]. The following list summarizes the main properties of a HBVM(k,s), with $k \geq s$ (see [7,12] for more details),

- it is symmetric with order $2s$, that is $y_1 - y(t_0+h) = O(h^{2s+1})$, while the stage order is s;
- for $k=s$ it becomes the s-stage Gauss collocation method;
- it is energy conserving when applied to canonical Hamiltonian problems with polynomial Hamiltonian of degree not larger than $2k/s$;
- for generic Hamiltonian functions, one has $H(y_1)-H(y_0) = O(h^{2k+1})$. As was observed in the previous section, a *practical* energy conservation property may be easily obtained by choosing k sufficiently large for the previous error to be within the round-off error level of the used finite precision arithmetic. Since the dimension of system (27) is independent of k (indeed γ has block dimension s), the computational effort of the whole procedure is not strongly affected by the order of the applied quadrature formula.

It turns out that, due to the orthogonality of the polynomials $P_j(c)$, for a HBVM(k,s) or, more in general, for a LIM satisfying the assumptions (A$_1$) and (A$_2$), the relationship with the weighted least squares collocation problem becomes even more meaningful, in that (see (29) and (10)) the numerical approximations $\bar{y}_1 = \bar{u}(t_0+h)$ and $y_1^* = u^*(t_0+h)$ get much closer to each other. To see this, we need some preliminary results.

Lemma 3. *Let $g : \mathbb{R} \to \mathbb{R}$ be sufficiently smooth in a neighborhood of t_0 and $v_\ell(c)$ a polynomial of degree ℓ. Then, under the assumptions (A$_1$) and (A$_2$), for $c \in [0,1]$ and $h > 0$, one has*

$$\sum_{i=1}^{k} w_i P_j(c_i)v_\ell(c_i)g(t_0+c_ih) = \begin{cases} O(h^{\max\{j-\ell,0\}}), & \text{if } j \leq s,\\ O(h^{\max\{2s-j-\ell,0\}}), & \text{if } j \geq s.\end{cases}$$

Proof. By virtue of Lemma 2 and considering $q \geq 2s$, we have

$$\begin{aligned}\sum_{i=1}^{k} w_i P_j(c_i)v_\ell(c_i)g(t_0+c_ih) &= \int_0^1 P_j(c)v_\ell(c)g(t_0+ch)\mathrm{d}c + O(h^{\max\{2s-j-\ell,0\}})\\ &= \int_0^1 P_j(c)v_\ell(c)\Big(\sum_{i=0}^{\infty}\frac{g^{(i)}(t_0)}{i!}(ch)^i\Big)\mathrm{d}c + O(h^{\max\{2s-j-\ell,0\}})\\ &= \sum_{i=0}^{\infty}\frac{g^{(i)}(t_0)}{i!}h^i\int_0^1 P_j(c)c^i v_\ell(c)\mathrm{d}c + O(h^{\max\{2s-j-\ell,0\}})\\ &= O(h^{\max\{j-\ell,0\}}) + O(h^{\max\{2s-j-\ell,0\}}),\end{aligned}$$

where the latter integral vanishes when $j > i+\ell$, due to the orthogonality conditions from assumption (A$_2$). □

It is worth recalling that, due to (A$_1$) and (A$_2$), systems (8) and (11) become

$$\gamma = \mathcal{P}_s^\top\Omega f(e\otimes y_0 + h\mathcal{I}_s\gamma) + h\mathcal{I}_s^\top\Omega g(\gamma) \tag{30}$$

and

$$\gamma \;=\; \mathcal{P}_s^\top \Omega f(e \otimes y_0 + h\mathcal{I}_s \gamma), \tag{31}$$

respectively. We also consider the Fourier expansion

$$f(\bar u(t_0+ch)) = \sum_{j\ge 0} P_j(c)\bar\gamma_j, \qquad \bar\gamma_j = \int_0^1 P_j(c) f(\bar u(t_0+c_ih))\mathrm{d}c.$$

Theorem 2. *Under the assumptions (*A_1*) and (*A_2*), the polynomials* $\bar u(t_0+ch)$ *and* $u^*(t_0+ch)$ *generated by the line integral and weighted least squares collocation methods satisfy*

$$\max_{0\le c\le 1} |u^*(t_0+ch) - \bar u(t_0+ch)| = O(h^{s+2}). \tag{32}$$

$$|u^*(t_0+h) - \bar u(t_0+h)| = O(h^{2s+1}). \tag{33}$$

Proof. Relation (32) is nothing but (23) and has been proved in Theorem 1. Concerning the superconvergence property (33) we begin with evaluating the term $h\mathcal{I}_s^\top \Omega g(\gamma)$ in (30) at $\bar\gamma$. For its $(\ell+1)$st component, $\ell = 0,\dots,s-1$, we have

$$\begin{aligned}
(h\mathcal{I}_s^\top \Omega g(\bar\gamma))_{\ell+1} \;&=\; h\sum_{i=1}^k w_i \int_0^{c_i} P_\ell(x)\mathrm{d}x g_i(\bar\gamma)\\
&=\; h\sum_{i=1}^k w_i \int_0^{c_i} P_\ell(x)\mathrm{d}x\big[f'(\bar u(t_0+c_ih))(\bar u'(t_0+c_ih) - f(\bar u(t_0+c_ih)))\big]\\
&=\; h\sum_{i=1}^k w_i \int_0^{c_i} P_\ell(x)\mathrm{d}x\big[f'(\bar u(t_0+c_ih))(\sum_{j=0}^{s-1} P_j(c_i)\Delta_j(h) + \sum_{j=s}^{\infty} P_j(c_i)\bar\gamma_j)\big]\\
&=\; h\sum_{j=0}^{s-1} \underbrace{\Delta_j(h)}_{O(h^{2s-j})} \underbrace{\sum_{i=1}^k w_i P_j(c_i) \int_0^{c_i} P_\ell(x)\mathrm{d}x f'(\bar u(t_0+c_ih))}_{O(h^{j-\ell-1}) \text{ from Lemma 3}}\\
&\quad +h\sum_{j=s}^{\infty} \underbrace{\bar\gamma_j}_{O(h^j)} \underbrace{\sum_{i=1}^k w_i P_j(c_i) \int_0^{c_i} P_\ell(x)\mathrm{d}x f'(\bar u(t_0+c_ih))}_{O(h^{2s-j-\ell-1}) \text{ from Lemma 3}} = O(h^{2s-\ell}).
\end{aligned}$$

Consequently, in vector notation,

$$|h\mathcal{I}_s^\top \Omega g(\bar\gamma)| = (O(h)^{2s}, O(h^{2s-1}),\dots,O(h^{s+1}))^\top. \tag{34}$$

We now consider the sequence (see (17) and (18))

$$\begin{cases} \gamma^{(0)} = \bar\gamma,\\ \gamma^{(n+1)} = \Psi(\gamma^{(n)}) = \Phi(\gamma^{(n)}) + h\mathcal{I}_s^\top\Omega g(\gamma^{(n)}), \quad n = 1,2,\dots, \end{cases}$$

and show that, positive constants h_0 and $M(h_0)$ exist such that for any $h \le h_0$, the sequence $\{\gamma^{(n)}\}$ is entirely contained in the neighborhood $B_{\bar\gamma}$ defined as

$$z \in B_{\bar\gamma} \;\Leftrightarrow\; |z-\bar\gamma| \le r(h_0) := M(h_0)(h_0^{2s}, h_0^{2s-1},\dots,h_0^{s+1})^\top, \tag{35}$$

and is a contraction there. Then, the contraction mapping theorem assures that the sequence converges to the unique fixed point γ^* of Ψ in $B_{\bar\gamma}$.

Let $\nu_1,\nu_2,\dots,\nu_s$ be the Lipschitz constants of the s components of the iteration function Ψ, and set $\nu = \max_i\{\nu_i\}$. Since $\nu = O(h)$, we can choose h_0 and $M(h_0)$ such that $\nu < 1$ in $B_{\bar\gamma}$ and, by (34),

$$|h\mathcal{I}_s^\top\Omega g(\bar\gamma)| \le (1-\nu)r(h_0) \qquad \text{for } h \le h_0.$$

Now, we use an induction argument. For $n = 1$ we have

$$|\gamma^{(1)} - \bar\gamma| = |\Psi(\bar\gamma) - \Phi(\bar\gamma)| = |h\mathcal{I}_s^\top\Omega g(\bar\gamma)| \le (1-\nu)r(h_0).$$

Assuming now that $\gamma^{(n)} \in B_{\bar\gamma}$, we obtain

$$\begin{aligned}
|\gamma^{(n+1)}-\bar\gamma| &\le |\gamma^{(n+1)}-\gamma^{(1)}|+|\gamma^{(1)}-\bar\gamma|\\
&\le |\Psi(\gamma^{(n)})-\Psi(\bar\gamma)|+(1-\nu)r(h_0)\\
&\le \nu|\gamma^{(n)}-\bar\gamma|+(1-\nu)r(h_0)\le r(h_0).
\end{aligned}$$

Having shown that $\gamma^*\in B_{\bar\gamma}$, from (35) we deduce that, in particular, $|\bar\gamma_0-\gamma_0^*|=O(h^{2s})$ and hence, by (29) and (10) with $\gamma_j=\gamma_j^*$,

$$|u^*(t_0+h)-\bar u(t_0+h)|=|y_0+h\gamma_0^*-y_0-h\bar\gamma_0|=h|\gamma_0^*-\bar\gamma_0|=O(h^{2s+1}),$$

which completes the proof. □

As an immediate consequence of Theorem 2 we have the following convergence result.

Corollary 1. *A least squares collocation method satisfying conditions* (A_1) *and* (A_2) *inherits the same stage order* $s+1$ *as the underlying LIM integrator, as well as the same superconvergence property, namely*

$$|u^*(t_0+h)-y(t_0+h)|=O(h^{2s+1}). \tag{36}$$

Remark 3. Strictly speaking, since formula (36) means that

$$|u^*(t_0+h)-y(t_0+h)|\le Ch^{2s+1},$$

for a given constant C independent of h, this result does not prevent the least squares method to exhibit an order higher than $2s$. Indeed, as we will show in the next section, we have experienced an order precisely equal to $2s$ for all least squares integrators, with the only exception of the method corresponding to the choice $s=1$ which, for many tested IVPs (both autonomous and non autonomous) yields an “ultraconvergence” property with order $p=3$, whenever $k\ge 2$.

This remark does not apply to the stage order property, since the $O(h^{s+2})$ closeness between the polynomials $\bar u(t_0+ch)$ and $u^*(t_0+ch)$, combined with the fact that the LIM has stage order $s+1$, unambiguously determines the stage order of the corresponding least squares formula.

Due to the super-convergence behavior, we can extend property (32) to the whole integration interval, provided that, in view of Remark 3, the least squares formula has order precisely equal to $2s$. For a given $N>0$, we set

$$h=\frac{T-t_0}{N},\qquad t_n=t_0+nh,\ n=0,\dots,N,$$

and denote by $\bar u(t_n+ch)$ and $u^*(t_n+ch)$ the numerical approximations obtained by applying the line integral and least squares methods to problem (1) sequentially on the intervals $[t_n,t_n+h]$, assuming $\bar y_n=\bar u(t_{n-1}+h)$ and $y_n^*=u^*(t_{n-1}+h)$ as initial conditions at the current step.

Corollary 2. *Under the assumptions* (A_1) *and* (A_2), *if the least squares method has order* $2s$, *the following estimation holds true for* $n=0,\dots,N-1$:

$$\max_{0\le c\le 1}|u^*(t_n+ch)-\bar u(t_n+ch)|=O(h^{s+2}). \tag{37}$$

5. Numerical illustrations

We present a few numerical illustrations to confirm the theoretical findings and, in particular, the behavior of least squares methods w.r.t. the underlying LIMs, as explained in Theorem 2 and Corollaries 1 and 2, with due attention to Remark 3. For this purpose, we introduce the following notations

$$\begin{aligned}
\bar e_n(c) &= |\bar u(t_{n-1}+ch)-y(t_{n-1}+ch)|,\\
e_n^*(c) &= |u^*(t_{n-1}+ch)-y(t_{n-1}+ch)|,\\
e_n(c) &= |\bar u(t_{n-1}+ch)-u^*(t_{n-1}+ch)|,
\end{aligned} \tag{38}$$

for $n=1,\dots,N$. The first two examples are simple test models whose exact solution $y(t)$ is known. In addition, in the last example, we compare the two methods in terms of their geometric properties, when applied to a simple canonical Hamiltonian system.

All the numerical tests have been implemented in Matlab (R2023a) on a 3.6 GHz Intel I9 core computer with 32 GB of memory.

5.1. Example 1

We solve the initial value problem

$$y'=t^3\exp(y)-t,\qquad y(0)=1,\qquad t\in[0,1], \tag{39}$$

Table 1
Convergence behavior of the numerical solutions $\bar{y}_N$ and y^*_N generated by the line integral and least squares methods of order four ($s = 2$) applied to Problem (39).

N	$\lvert\bar{y}_N - y(T)\rvert$	order	$\lvert y^*_N - y(T)\rvert$	order	$\lvert\bar{y}_N - y^*_N\rvert$	order
2	5.81e-03		3.90e-02		4.48e-02	
2^2	4.56e-04	3.67	2.63e-03	3.89	3.09e-03	3.86
2^3	3.08e-05	3.89	1.36e-04	4.27	1.67e-04	4.21
2^4	1.97e-06	3.97	7.05e-06	4.27	9.02e-06	4.21
2^5	1.24e-07	3.99	3.91e-07	4.17	5.14e-07	4.13
2^6	7.75e-09	4.00	2.28e-08	4.10	3.06e-08	4.07
2^7	4.84e-10	4.00	1.38e-09	4.05	1.86e-09	4.04
2^8	3.03e-11	4.00	8.46e-11	4.03	1.15e-10	4.02
2^9	1.89e-12	4.00	5.24e-12	4.01	7.13e-12	4.01
2^{10}	1.18e-13	4.01	3.24e-13	4.02	4.41e-13	4.01

Table 2
Convergence behavior of the polynomial approximations $\bar{u}(t_{N-1}+ch)$ and $u^*(t_{N-1}+ch)$ to the true solution $y(t_{N-1}+ch)$ generated by the line integral and least squares methods of order four ($s = 2$) applied to Problem (39).

N	$\max_i \bar{e}_N(c_i)$	order	$\max_i e^*_N(c_i)$	order	$\max_i e_N(c_i)$	order
2	2.62e-02		3.43e-02		4.38e-02	
2^2	5.66e-03	2.21	5.72e-03	2.58	3.01e-03	3.86
2^3	1.07e-03	2.40	1.05e-03	2.45	1.61e-04	4.23
2^4	1.76e-04	2.61	1.72e-04	2.61	8.53e-06	4.23
2^5	2.56e-05	2.78	2.55e-05	2.75	4.78e-07	4.16
2^6	3.48e-06	2.88	3.47e-06	2.87	2.81e-08	4.09
2^7	4.54e-07	2.94	4.54e-07	2.94	1.70e-09	4.05
2^8	5.80e-08	2.97	5.80e-08	2.97	1.05e-10	4.02
2^9	7.33e-09	2.98	7.33e-09	2.98	6.48e-12	4.01
2^{10}	9.21e-10	2.99	9.21e-10	2.99	4.00e-13	4.02

whose true solution is

$$y(t) = \log\left(\frac{1}{\exp(t^2/2)(\exp(-1)-2)+t^2+2}\right),$$

using the least squares and line integral methods of order 4 ($s = 2$), 6 ($s = 3$) and 8 ($s = 4$), with $k = 10$ abscissae. To assess the convergence properties of these formulae, we progressively halve the stepsize h, and consequently double the number N of the subdivisions of the integration interval, until an error close to the machine epsilon is attained. These formulae satisfy the assumptions of Corollary 2, allowing us to focus our attention on the errors (38) evaluated in the rightmost subinterval $[t_{N-1}, t_N] = [1-h, 1]$.

Tables 1 and 2 summarize the results for the methods of order four. Regarding Table 1, the following details are provided:

- The first column reports the number of subdivisions N: for the problem at hand, selecting $N = 2^{10}$ yields an approximation accuracy of about 10^{-13} at the final time $t = T = 1$ for both methods.
- The second and third columns display the errors $|\bar{y}_N - y(T)|$ produced by the line integral method at time $t = 1$ and the associated order of convergence, which is four, as expected.
- In the fourth and fifth columns are the errors $|y^*_N - y(T)|$ produced by the least squares method at time $t = 1$ and the corresponding order of convergence which is also four, consistently with what stated in Corollary 1.
- The last two columns present the distances $|\bar{y}_N - y^*_N|$ between the two computed approximations at $t = 1$, confirming the result (33) in Theorem 2.

With reference to Table 2, the following details are provided:

- The second and third columns show the maximum errors $\max_i \bar{e}_N(c_i)$ produced by the line integral method at the internal abscissae $t_{N-1} + c_i h$ located within the last subinterval $[t_{N-1}, t_N]$. It is well-known from the literature [7] that the corresponding order of convergence (stage order) is $s + 1 = 3$.
- The fourth and fifth columns contain the errors $\max_i e^*_N(c_i)$ of the polynomial approximation $u^*(t_{N-1} + ch)$ generated by the least squares method along with the associated order of convergence which is as well $s + 1 = 3$, consistently with what stated in Corollary 1.
- The last two columns report the distance $\max_i e_N(c_i)$ between the two polynomial approximations $\bar{u}(t_{N-1}+ch)$ and $u^*(t_{N-1}+ch)$ evaluated at the internal abscissae, confirming the result (32) in Theorem 2.

Tables 3-4 and 5-6 summarize the results for the methods of order six and eight respectively, confirming the same conclusions discussed above for these higher order methods.

Table 3
Convergence behavior of the numerical solutions $\bar{y}_N$ and y_N^* generated by the line integral and least squares methods of order six ($s = 3$) applied to Problem (39).

N	$\lvert\bar{y}_N - y(T)\rvert$	order	$\lvert y_N^* - y(T)\rvert$	order	$\lvert\bar{y}_N - y_N^*\rvert$	order
2	7.99e-05		3.88e-03		3.96e-03	
2^2	1.19e-06	6.07	1.03e-04	5.24	1.04e-04	5.25
2^3	1.63e-08	6.19	1.75e-06	5.87	1.77e-06	5.88
2^4	2.36e-10	6.11	2.53e-08	6.11	2.56e-08	6.11
2^5	3.59e-12	6.04	3.63e-10	6.13	3.67e-10	6.12
2^6	5.71e-14	5.97	5.36e-12	6.08	5.41e-12	6.08
2^7	1.33e-15	5.42	8.08e-14	6.05	8.22e-14	6.04

Table 4
Convergence behavior of the polynomial approximations $\bar{u}(t_{N-1}+ch)$ and $u^*(t_{N-1}+ch)$ to the true solution $y(t_{N-1}+ch)$ generated by the line integral and least squares methods of order six ($s = 3$) applied to Problem (39).

N	$\max_i \bar{e}_N(c_i)$	order	$\max_i e_N^*(c_i)$	order	$\max_i e_N(c_i)$	order
2	4.38e-03		5.70e-03		3.82e-03	
2^2	6.91e-04	2.66	7.04e-04	3.02	9.60e-05	5.31
2^3	7.89e-05	3.13	7.77e-05	3.18	2.39e-06	5.32
2^4	7.09e-06	3.48	7.01e-06	3.47	1.45e-07	4.04
2^5	5.42e-07	3.71	5.39e-07	3.70	6.05e-09	4.59
2^6	3.77e-08	3.85	3.76e-08	3.84	2.20e-10	4.78
2^7	2.49e-09	3.92	2.49e-09	3.92	7.43e-12	4.89
2^8	1.60e-10	3.96	1.60e-10	3.96	2.41e-13	4.94
2^9	1.02e-11	3.98	1.02e-11	3.98	7.77e-15	4.96
2^{10}	6.39e-13	3.99	6.39e-13	3.99	2.22e-16	5.13

Table 5
Convergence behavior of the numerical solutions $\bar{y}_N$ and y_N^* generated by the line integral and least squares methods of order eight ($s = 4$) applied to Problem (39).

N	$\lvert\bar{y}_N - y(T)\rvert$	order	$\lvert y_N^* - y(T)\rvert$	order	$\lvert\bar{y}_N - y_N^*\rvert$	order
2	1.84e-06		3.63e-04		3.62e-04	
2^2	1.65e-08	6.80	3.85e-06	6.56	3.83e-06	6.56
2^3	9.17e-11	7.49	2.18e-08	7.46	2.17e-08	7.46
2^4	4.07e-13	7.81	8.95e-11	7.93	8.91e-11	7.93
2^5	6.66e-16	9.26	3.32e-13	8.07	3.32e-13	8.07

Table 6
Convergence behavior of the polynomial approximations $\bar{u}(t_{N-1} + ch)$ and $u^*(t_{N-1}+ch)$ to the true solution $y(t_{N-1}+ch)$ generated by the line integral and least squares methods of order eight ($s = 4$) applied to Problem (39).

N	$\max_i \bar{e}_N(c_i)$	order	$\max_i e_N^*(c_i)$	order	$\max_i e_N(c_i)$	order
2	9.40e-04		1.10e-03		3.37e-04	
2^2	9.32e-05	3.33	9.37e-05	3.55	4.57e-06	6.20
2^3	6.25e-06	3.90	6.23e-06	3.91	1.57e-07	4.86
2^4	3.10e-07	4.34	3.09e-07	4.33	4.40e-09	5.16
2^5	1.25e-08	4.63	1.25e-08	4.63	9.57e-11	5.52
2^6	4.49e-10	4.80	4.49e-10	4.80	1.78e-12	5.75
2^7	1.50e-11	4.90	1.50e-11	4.90	3.04e-14	5.87

5.2. Example 2

As was anticipated in Remark 3, the least squares method corresponding to the choice $s = 1$ and $k > 1$, exhibits an order equal to three, so that Corollary 2 does not apply for this method. With $k = 10$, we use this method to solve the IVP

$$y' = -\sin(y), \qquad y(0) = 1, \qquad t \in [0, T], \tag{40}$$

whose true solution is

$$y(t) = 2\arctan(\exp(\log(\tan(1/2)) - t)).$$

Table 7
Convergence behavior of the numerical solutions $\bar{y}_1$ and y_1^* generated by the line integral and least squares methods corresponding to the choice $s = 1$, applied to Problem (40).

N	$\lvert\bar{y}_1 - y(T)\rvert$	order	$\lvert y_1^* - y(T)\rvert$	order	$\lvert\bar{y}_1 - y_1^*\rvert$	order
2	2.98e-03		1.41e-05		2.96e-03	
2^2	3.63e-04	3.04	1.28e-05	0.13	3.76e-04	2.98
2^3	4.33e-05	3.07	1.35e-06	3.25	4.46e-05	3.07
2^4	5.22e-06	3.05	1.02e-07	3.72	5.32e-06	3.07
2^5	6.39e-07	3.03	6.96e-09	3.88	6.46e-07	3.04
2^6	7.90e-08	3.02	4.53e-10	3.94	7.95e-08	3.02
2^7	9.82e-09	3.01	2.88e-11	3.97	9.85e-09	3.01
2^8	1.22e-09	3.00	1.82e-12	3.99	1.23e-09	3.01
2^9	1.53e-10	3.00	1.14e-13	3.99	1.53e-10	3.00
2^{10}	1.91e-11	3.00	7.22e-15	3.98	1.91e-11	3.00

Table 8
Convergence behavior of the polynomial approximations $\bar{u}(t_0 + ch)$ and $u^*(t_0 + ch)$ to the true solution $y(t_0 + ch)$ generated by the line integral and least squares methods corresponding to the choice $s = 1$, applied to Problem (40).

N	$\max_i \bar{e}_1(c_i)$	order	$\max_i e_1^*(c_i)$	order	$\max_i e_1(c_i)$	order
2	1.38e-02		1.51e-02		2.93e-03	
2^2	3.55e-03	1.96	3.72e-03	2.02	3.71e-04	2.98
2^3	8.84e-04	2.00	9.06e-04	2.04	4.40e-05	3.07
2^4	2.19e-04	2.01	2.22e-04	2.03	5.25e-06	3.07
2^5	5.46e-05	2.01	5.49e-05	2.02	6.38e-07	3.04
2^6	1.36e-05	2.00	1.37e-05	2.01	7.84e-08	3.02
2^7	3.40e-06	2.00	3.40e-06	2.00	9.72e-09	3.01
2^8	8.49e-07	2.00	8.49e-07	2.00	1.21e-09	3.01
2^9	2.12e-07	2.00	2.12e-07	2.00	1.51e-10	3.00
2^{10}	5.30e-08	2.00	5.30e-08	2.00	1.88e-11	3.00

We now focus on the asymptotic behavior of the solution in the first subinterval $[0, h]$, as h approaches zero. To this end, we choose $T = 1/N$, with $N = 2^k$, $k = 1, \dots,$ until the accuracy gets close enough to the machine epsilon. The results are summarized in Tables 7 and 8, with the same meaning as in the previous example, except that now the errors have a local nature. The following conclusions may be drawn:

- The third column of Table 7 displays the order of convergence related to the local truncation error associated with the line integral method. We can observe a third-order convergence (the stage order remains $s + 1$ as in the previous example due to the superconvergence property).
- The local truncation error of the least squares method and the associated order of convergence appear in the fourth and fifth columns of Table 7: it turns out that the method exhibits local order four and hence global order three.
- The last two columns of Table 7 show the distances $|\bar{y}_1 - y_1^*|$ between the two computed approximations at the very first step, confirming the result (33) in Theorem 2.
- Finally, the results in Table 8 are consistent with the stage order result in Corollary 1 and the property (32) of Theorem 2.

5.3. Example 3

To provide a conclusive comparison, we examine the geometric properties of two methods applied to the Kepler problem [7,14]

$$\dot{q} = p, \qquad \dot{p} = -\|q\|_2^{-3} q, \tag{41}$$

$q = (q_1, q_2)^\top, p = (p_1, p_2)^\top \in \mathbb{R}^2$ being the generalized coordinates and momenta, respectively. Problem (41) is a canonical Hamiltonian system with two degrees of freedom that describes the (planar) motion of two massive bodies, under their mutual gravitational attraction, around their center of mass. The initial condition

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

yields a periodic orbit of period $T = 2\pi$ that, in the q-plane, is given by an ellipse of eccentricity ε. Two well-known constants of motion of system (41) are:

- the total energy: $H(q, p) = \frac{1}{2}\|p\|_2^2 - \|q\|_2^{-1}$;

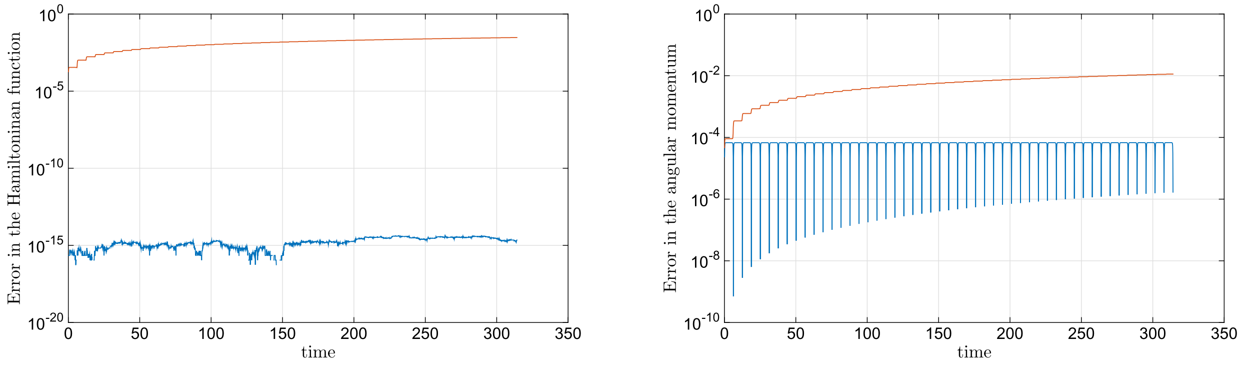

Fig. 1. Errors $|H(q_n, p_n) - H(q_0, p_0)|$ in the Hamiltonian function (left picture) and $|M(q_n, p_n) - M(q_0, p_0)|$ in the angular momentum (right picture) produced by the line integral method (bottom line, blue) and least squares method (upper line, red). (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

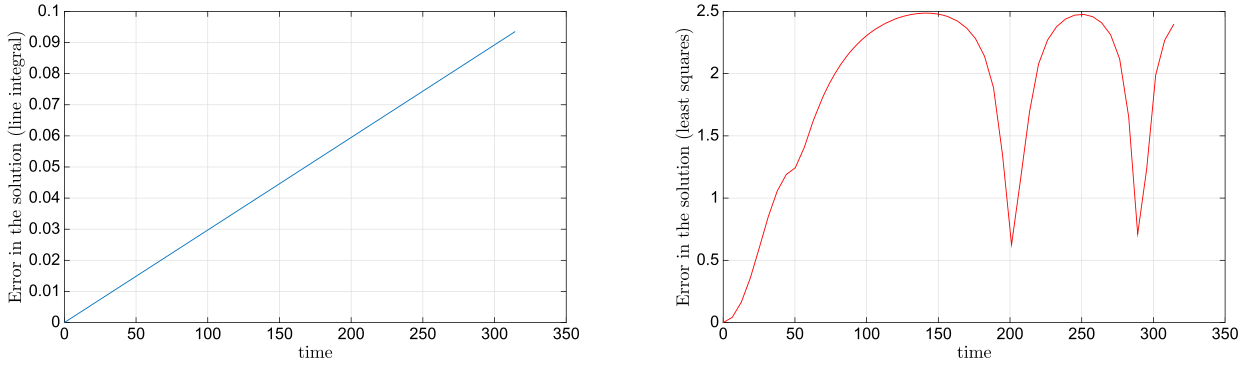

Fig. 2. Errors in the solution at multiples of the period T produced by the line integral method (left picture) and least squares method (right picture).

- the angular momentum: $M(q, p) = q_1 p_2 - p_1 q_2$.

For our experiment, we set $\epsilon = 0.6$ and integrate the problem over the time interval $[0, 50T]$ using fourth-order line integral and least squares methods ($s = 2$) with $k = 10$ abscissae and a constant stepsize $h = T/50$. Our interest lies in comparing the long-term behavior of solutions produced by the two methods in terms of conserving the aforementioned first integrals. The results are displayed in Fig. 1.

As was illustrated in Section 3, the line integral method, turns out to be energy-conserving for the problem at hand, up to an error that is below the machine epsilon. This is confirmed by the left picture of Fig. 1 which displays the error $|H(q_n, p_n) - H(q_0, p_0)|$ of the Hamiltonian function evaluated along the numerical solution: the bottom line, colored blue, is of the order of 10^{-15} throughout the considered time interval. This is not the case for the least squares method, for which a drift in the error may be observed (upper red line). The right picture of Fig. 1 shows the errors $|M(q_n, p_n) - M(q_0, p_0)|$ in the angular momentum. Despite the line integral method does not preserve this constant of motion precisely, the error remains uniformly bounded over time. Conversely, the least squares method produces a drift. We conclude that the least squares method (2)-(3) under the conditions (A$_1$)-(A$_2$) is not recommended for geometric integration purposes.

The absence of conservation, or near-conservation of first integrals also impacts the accuracy of the numerical solution. In Fig. 2, we present the errors $||(q_n, p_n) - (q_0, p_0)||_\infty$ evaluated at times $t_n = kT$, $k = 0, \dots, 50$, multiples of the period T, where the solution is exactly (q_0, p_0) due to its periodic behavior. In the left picture, displaying the results of the line integral method, we observe a linear growth of the error, which remains at the order of 10^{-1}. Conversely, the error in the solution generated by the least squares method, displayed in the right panel of Fig. 2, exhibits a more erratic behavior and is one order greater.

We conclude that the least squares method (2)-(3) under the conditions (A$_1$)-(A$_2$) is not recommended for geometric integration purposes.

6. Conclusions

We have explored a weighted least squares approach to tackle an overdetermined collocation problem arising from the requirement that a polynomial of degree s satisfies an ODE-IVP over $k \geq s$ collocating points $t_0 + c_i$, $i = 1, \dots, k$. Our analysis revealed that the resulting nonlinear system can be interpreted as a perturbation of a corresponding system defining line integral methods (LIMs), which are special low-rank Runge–Kutta methods introduced in the context of geometric integration. This interpretation has led to several interesting insights:

- The LIM solution can serve as a convenient initial guess for solving the least squares problem.
- The convergence properties of LIMs can be extended to least squares collocation methods, including superconvergence behavior when using the Legendre basis to represent the approximating polynomials.
- These polynomial approximations converge to each other at a rate exceeding the underlying order of convergence. Consequently, for small stepsizes h, the two methods essentially yield the same approximation.

Given that the perturbation term involves the Jacobian matrix of the vector field, LIMs may be viewed as a computationally simplified variant of least squares methods. Considering the recent successful application of the latter class of integrators to differential algebraic equations (DAEs) of higher order, a potential future research could be investigating the performance of LIMs applied to DAEs, starting from the preliminary results in [9].

CRediT authorship contribution statement

Luigi Brugnano: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Felice Iavernaro:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Ewa B. Weinmüller:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Acknowledgements

L. Brugnano and F. Iavernaro are members of GNCS-INdAM. F. Iavernaro wishes to thank "*Fondo acquisto e manutenzione attrezzature per la ricerca*" D.R. 3191/2022, Università degli Studi di Bari Aldo Moro.

References

[1] E.L. Albasiny, A subroutine for solving a system of differential equations in Chebyshev series, in: B. Childs, M. Scott, J.W. Daniel, E. Denman, P. Nelson (Eds.), Codes for Boundary-Value Problems in Ordinary Differential Equations, in: Lecture Notes in Computer Science, vol. 76, Springer-Verlag, Berlin Heidelberg New York, 1979, pp. 280–286.

[2] U. Ascher, Discrete least squares approximations for ordinary differential equations, SIAM J. Numer. Anal. 15 (3) (1978) 478–496.

[3] U. Ascher, G. Bader, A new basis implementation for a mixed order boundary-value ode solver, SIAM J. Sci. Stat. Comput. 8 (4) (1987) 483–500, https://doi.org/10.1137/0908047.

[4] U. Ascher, S. Pruess, R.D. Russell, On spline basis selection for solving differential equations, SIAM J. Numer. Anal. 20 (1) (1983) 121–142, https://doi.org/10.1137/0720009.

[5] U. Ascher, R. Spiteri, Collocation software for boundary-value differential-agebraic equations, SIAM J. Sci. Comput. 15 (1994) 938–952, https://doi.org/10.1137/0915056.

[6] L. Brugnano, G. Frasca-Caccia, F. Iavernaro, Efficient implementation of Gauss collocation and Hamiltonian boundary value methods, Numer. Algorithms 65 (2014) 633–650, https://doi.org/10.1007/s11075-014-9825-0.

[7] L. Brugnano, F. Iavernaro, Line Integral Methods for Conservative Problems, Chapman et Hall/CRC, Boca Raton, FL, USA, 2016, http://web.math.unifi.it/users/brugnano/LIMbook/.

[8] L. Brugnano, F. Iavernaro, Line integral solution of differential problems, Axioms 7 (2) (2018) 36, https://doi.org/10.3390/axioms7020036.

[9] L. Brugnano, F. Iavernaro, A general framework for solving differential equations, Ann. Univ. Ferrara 68 (2) (2022) 243–258, https://doi.org/10.1007/s11565-022-00409-6.

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

[11] L. Brugnano, F. Iavernaro, D. Trigiante, A note on the efficient implementation of Hamiltonian BVMs, J. Comput. Appl. Math. 236 (2011) 375–383, https://doi.org/10.1016/j.cam.2011.07.022.

[12] L. Brugnano, F. Iavernaro, D. Trigiante, A simple framework for the derivation and analysis of effective one-step methods for ODEs, Appl. Math. Comput. 218 (2012) 8475–8485, https://doi.org/10.1016/j.amc.2012.01.074.

[13] I. Gladwell, The development of the boundary-value codes in the ordinary differential equations chapter of the NAG library, in: B. Childs, M. Scott, J.W. Daniel, E. Denman, P. Nelson (Eds.), Codes for Boundary-Value Problems in Ordinary Differential Equations, in: Lecture Notes in Computer Science, vol. 76, Springer-Verlag, Berlin Heidelberg New York, 1979, pp. 122–143.

[14] E. Hairer, Ch. Lubich, G. Wanner, Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations, 2nd ed., Springer, Berlin, Germany, 2006.

[15] M. Hanke, R. März, Towards a reliable implementation of least squares collocation for higher-index differential-algebraic equations. Part 1: basics and ansatz function choices, Numer. Algorithms 89 (2022) 931–963, https://dx.doi.org/10.1007/s11075-021-01140-7.

[16] M. Hanke, R. März, Towards a reliable implementation of least squares collocation for higher index linear differential-algebraic equations. Part 2: the discrete least squares problem, Numer. Algorithms 89 (2022) 965–986, https://doi.org/10.1007/s11075-021-01141-6.

[17] M. Hanke, R. März, Convergence analysis of least squares collocation methods for nonlinear higher index differential-algebraic equations, J. Comput. Appl. Math. 387 (2021) 112514, https://doi.org/10.1016/j.cam.2019.112514.

[18] M. Hanke, R. März, A reliable direct numerical treatment of differential-algebraic equations by overdetermined collocation: an operator approach, J. Comput. Appl. Math. 387 (2021) 112520, https://doi.org/10.1016/j.cam.2019.112520.

[19] M. Hanke, R. März, C. Tischendorf, Least squares collocation for higher-index linear differential algebraic equations: estimating the stability threshold, Math. Comput. 88 (318) (2019) 1647–1683, https://doi.org/10.1090/mcom/3393.

[20] M. Hanke, R. März, C. Tischendorf, E.B. Weinmüller, S. Wurm, Least squares collocation for linear higher-index differential-algebraic equations, J. Comput. Appl. Math. 317 (2017) 403–431, https://doi.org/10.1016/j.cam.2016.12.017.

[21] The Numerical Algorithms Group (NAG), the nag library for Fortran, 2019.