UNIVERSITY OF CAMBRIDGE

Numerical Analysis Reports

# Numerical solution of isospectral flows

**Mari Paz Calvo, Arieh Iserles and Antonella Zanna**

DAMTP 1995/NA03
August 1995

Department of Applied Mathematics and Theoretical Physics
Silver Street
Cambridge
England CB3 9EW

# Numerical solution of isospectral flows

Mari Paz Calvo,[1] Arieh Iserles[2] and Antonella Zanna[3]

**Abstract** In this paper we are concerned with the problem of solving numerically isospectral flows. These flows are characterized by the differential equation

$$L' = [B(L), L], \quad L(0) = L_0,$$

where $L_0$ is a $d \times d$ symmetric matrix, $B(L)$ is a skew-symmetric matrix function of $L$ and $[B, L]$ is the Lie bracket operator.
We show that standard Runge–Kutta schemes fail in recovering the main qualitative feature of these flows, that is isospectrality, since they cannot recover arbitrary cubic conservation laws. This failure motivates us to introduce an alternative approach and establish a framework for generation of isospectral methods of arbitrarily high order.

[1]Departamento de Matemática Aplicada y Computación, Universidad de Valladolid, Spain
[2]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, England
[3]Newnham College, University of Cambridge, England

# 1 Background and notation

## 1.1 Introduction

The interest in solving isospectral flows is motivated by their relevance in a wide range of applications, from molecular dynamics to micromagnetics to linear algebra. The general form of an isospectral flow is the differential equation

$$L' = [B(L), L], \quad L(0) = L_0, \tag{1}$$

where $L_0$ is a given $d \times d$ symmetric matrix, $B(L)$ is a skew-symmetric matrix function of $L$ and $[B(L), L] = B(L)L - LB(L)$ is the commutator of $B(L)$ and $L$.

The choice of the matrix function $B(L)$ characterizes the dynamics of the underlying flow $L(t)$. Important special cases are the *Toda lattice equations*, *double-bracket flows* and *KvM flows*.

Toda lattice equations in the Lax formulation (1) were considered by Toda [T], Flaschka [F] and Moser [Mo] and their relation with the QR algorithm for finding eigenvalues by Symes [Sy] and then extensively by Deift, Nanda, Tomei et al., Lagarias, in [Na1], [Na2], [DNT], [L], [DRTW]. It has been finally generalized to the nonsymmetric case by Chu, Watkins and Elsner in [Ch], [W], [WE]. The double bracket flow was introduced by Brockett in [B1] and then investigated by Brockett et al. in [BBR]. Its relation with the singular value decomposition (SVD) was considered by Chu, Driessel, Moore, Mahony, Helmke, Watkins, and others (cf. [ChD1], [D], [HM], [MMH], [WE]). Driessel and Chu in [ChD2] have also investigated another isospectral flow of the form (1) in relation with the inverse eigenvalue problem for Toeplitz symmetric matrices. Finally we mention the KvM flows studied by Kac and von Moerbeke (cf. [KvM], [T]). We will return time and again to these flows in the sequel.

It is important to point out that the aforementioned flows are obtained for very special choices of the matrix $B(L)$. In the most general case the dynamics of (1) is still unknown or not yet fully understood.

The most important qualitative feature of (1) is the isospectrality of the solution $L(t)$. In other words the eigenvalues of the matrix $L(t)$ are independent of time. This has been shown by Flaschka for the Toda lattice equations (see [F], [T]) but with greater generality his proof applies to all the flows that can be written in the form (1). Therefore it is often of essence to require that a numerical method for the initial value problem (1) retains isospectrality. So far, Moore–Mahony–Helmke have proposed in [MMH] an algorithm which produces an isospectral solution. Their algorithm is aimed to evaluate the eigenvalues of $L_0$ rather than to approximate the solution of (1) with any degree of precision, and it is applicable just to the double-bracket flows. Instead we propose a considerably more general approach which allows us to produce an isospectral solution for the initial value problem (1) with an arbitrarily high order of accuracy. This new class of methods, which we call *modified Gauss–Legendre Runge–Kutta* (MGLRK) schemes, is based on a rendition of Flaschka's theoretical approach in a computational form.

This different technique is strongly motivated by the failure of standard ODE methods for the problem in hand. Isospectrality of (1) can be interpreted in the following way. The solution $L(t)$ lies on an intersection of several manifolds, each one corresponding to an integral for (1) that can be expressed in terms of a conservation law. We show that the most likely candidates, Runge–Kutta (RK) methods which retain quadratic conservation laws, fail since they cannot recover cubic integrals.

This paper is organized as follows. Section §1 introduces some basic concepts for the problem in hand and describes the most ubiquitous isospectral flows. Section §2 is concerned with standard methods for ODEs. We derive the conditions that the coefficients of the numerical method have to obey in order to recover conservation laws. In particular, we prove that for Runge–Kutta schemes, conservation of quadratic and cubic integrals are conflicting requirements, thereby concluding that for $d \geq 3$ no RK method can be isospectral. In Section §3 we introduce the modified Gauss–Legendre RK methods and, finally, Section §4 is concerned with numerical examples.

### 1.2 The QR flow and the Toda flow

Given a function $f$ which is analytic on the spectrum $\sigma(L_0) = \{\lambda_1, \lambda_2, \ldots, \lambda_d\}$ of the matrix $L_0$, we refer to (1) as a *QR flow* when

$$B(L) = f(L)_+ - f(L)_-. \tag{2}$$

The subscripts '$\pm$' denote the (strictly) upper and lower triangular part of the matrix $f(L)$ respectively. The name QR flow (cf. [Sy], [Na1], [Na2], [DMT]) is due to the fact that for symmetric and positive definite $L_0$ and $f(x) = \log x$ at integer time-steps the flow produces exactly the iterations of the familiar QR method for finding eigenvalues. Reversing the order in (2) is equivalent to reversing integration in time. Moser in [Mo] has shown that for $t \to \pm\infty$ the flow $L(t)$ tends to a diagonal matrix whose elements are the eigenvalues of $L(t)$, while Deift, Nanda and Tomei in [DNT] have shown that the convergence to the asymptotically stable equilibrium point is exponential. For $t \to +\infty$, the eigenvalues of the flow (2) are arranged from the largest to the smallest, the other way around for reverse time integration. The flow retains the bandwidth of the initial matrix $L_0$. This is clear for the Toda flow (see [T]) but, by virtue of the analyticity of $f$ on $\sigma(L_0)$ it applies with greater generality also to (2) (cf. [DNT]). For $f(x) = x$, the identity function, and for tridiagonal $L_0$, we obtain the Toda flow in a notation originally due to Lax (cf. [T]), in which case we denote the matrix $L(t)$ in the form

$$L(t) = \begin{bmatrix} \beta_1 & \alpha_1 & 0 & \cdots & 0 \\ \alpha_1 & \beta_2 & \alpha_2 & \ddots & \vdots \\ 0 & \alpha_2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \alpha_{d-1} \\ 0 & \cdots & 0 & \alpha_{d-1} & \beta_d \end{bmatrix}_{d\times d}. \tag{3}$$

Occasionally we refer directly to the differential equations for the Toda flow (cf. [T]), namely

$$\begin{aligned} \beta_k' &= 2(\alpha_k^2 - \alpha_{k-1}^2), \\ \alpha_k' &= \alpha_k(\beta_{k+1} - \beta_k), \end{aligned} \qquad k = 1, \ldots, d, \tag{4}$$

where here, as well as in the remaining part of this paper, we use the convention that $\alpha_k, \beta_k = 0$ whenever $k \notin \{1, 2, \ldots, d-1\}$ and $k \notin \{1, \ldots, d\}$ respectively.

## 1.3 The double-bracket flow

We refer to a *double-bracket flow* when

$$L' = [L, [L, N]], \tag{5}$$

where $N$ is a given $d \times d$ symmetric matrix. Without loss of generality and observing that $[L, [L, N]] = [[N, L], L]$, the flow can be written in the form (1), where

$$B(L) = [N, L] = NL - LN.$$

The flow has been firstly introduced by Brockett in [B1], where the author shows that it can be formulated as a gradient flow evolving in a Riemannian manifold. In the same paper he has also proved that, for diagonal $N$ and an initial matrix $L_0$, both of them with distinct eigenvalues, the matrix function $L(t)$ tends exponentially to a diagonal matrix as $t \to +\infty$ and the eigenvalues are then sorted accordingly to the diagonal entries of $N$. If $L_0$ or $N$ have multiple eigenvalues, exponential (but not asymptotical) convergence is lost. He also showed how this flow can be used to diagonalize matrices, sort lists and solve linear programming problems. This flow in general does not retain the bandwidth of the initial matrix $L_0$ except in the case

$$N = \kappa I \pm \mathrm{diag}\{1, 2, \ldots, d\}.$$

In particular, when $N = \mathrm{diag}\{1, 2, \ldots, d\}$, the double-bracket flow (5) is a reformulation of the Toda lattice.

When $N$ is nondiagonal, the analysis of convergence is essentially the same (cf. [B1]). To verify this, observe that $N$ is symmetric, therefore it can be diagonalized by means of an orthogonal transformation. In other words, there exist an orthogonal matrix $Q$ and a diagonal matrix $\Lambda$ such that

$$N = Q\Lambda Q^T, \qquad QQ^T = Q^TQ = I.$$

Next observe that, if $Q$ is orthogonal and $A, B$ are arbitrary $d \times d$ matrices, it is true that

$$[QAQ^T, QBQ^T] = Q[A, B]Q^T,$$

from which, letting $\tilde{L} = Q^TLQ$, we deduce that the problem (5) is equivalent to

$$\tilde{L}' = [\tilde{L}, [\tilde{L}, \Lambda]],$$

with diagonal $\Lambda$. Thus, the previous analysis holds. Finally note that, when $N$ is nondiagonal, the equilibria of the flow need not to be diagonal matrices. This can be observed by transforming, by means of the orthogonal matrix $Q$, the equilibria of the flow corresponding to the diagonal matrix $\Lambda$.

## 1.4 An isospectral flow for inverse eigenvalue problems

Chu and Driessel have shown in [ChD2] that isospectral flows can also be used for the inverse eigenvalue problem for symmetric Toeplitz matrices. Given a set of $d$ arbitrary real numbers, say $\{\lambda_1, \ldots, \lambda_d\}$, the problem consist of finding a symmetric $d \times d$ Toeplitz matrix whose eigenvalues are exactly the given numbers. To this aim, they formulate an isospectral flow whose equilibria are only Toeplitz symmetric matrices. In more detail, assume that $L$ is a given symmetric matrix. Then (cf. [ChD2]) it can be uniquely decomposed as

$$L = T(L) + P(L),$$

where $T(L)$ is a symmetric Toeplitz matrix and $P(L)$ is the projection of $L$ on the set $\mathcal{P}$, namely

$$\mathcal{P} := \{X : X \text{ is } d \times d, \text{ symmetric, and } \forall j = 1, \ldots, d-1, X_{1,j} + X_{2,j+1} = 0; X_{1,d} = 0\}.$$

Letting $Z$ stand for the $d \times d$ *shift matrix*,

$$Z = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & 0 & 1 & 0 \\ \vdots & & \ddots & 0 & 1 \\ 0 & \ldots & \ldots & 0 & 0 \end{bmatrix},$$

the isospectral flow for the symmetric eigenvalue problem for Toeplitz matrices is

$$L' = [L, [P(L), V]], \qquad L_0 = \operatorname{diag}\{\lambda_1, \ldots, \lambda_d\}, \quad V = Z + Z^T. \tag{6}$$

Unfortunately, Chu and Driessel have not yet proved the convergence of the flow. This flow is closely related to the double-bracket flow. If we write

$$P(L) = L - T(L),$$

we find

$$L' = [L, [L, V]] - [L, [T(L), V]],$$

whereby the first term is just a double-bracket flow with symmetric (nondiagonal) matrix $N = V$.

## 1.5 The Kac–von Moerbeke flow.

Given

$$L(t) = \begin{bmatrix} 0 & \alpha_1 & 0 & \ldots & 0 \\ \alpha_1 & 0 & \alpha_2 & \ddots & \vdots \\ 0 & \alpha_2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \alpha_{d-1} \\ 0 & \ldots & 0 & \alpha_{d-1} & 0 \end{bmatrix},$$

the KvM flow corresponds to the choice

$$B(L) = \begin{bmatrix} 0 & 0 & \alpha_1\alpha_2 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \alpha_2\alpha_3 & \ddots & \vdots \\ -\alpha_1\alpha_2 & 0 & 0 & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \alpha_{d-2}\alpha_{d-1} \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -\alpha_{d-2}\alpha_{d-1} & 0 & 0 \end{bmatrix}. \tag{7}$$

Kac and von Moerbeke have shown that this flow remains tridiagonal with zero diagonal entries. Moser (cf. [T], p. 72) has generalized this result to problems whose matrix $B(L)$ has up to the $p$-th off-diagonal element. For $t \to \infty$, $L(t)$ tends to a block diagonal matrix. Each block is $2 \times 2$ and the spectrum of $L(t)$ is obtained by evaluating the eigenvalues of each block. Kac and von Moerbeke have shown (cf. [T], [KvM]) that this flow is related to the Toda flow since it can be interpreted as two different motions of the same lattice. The equations for the $\alpha_k$'s are given explicitely by

$$\alpha_k' = \alpha_k(\alpha_{k+1}^2 - \alpha_{k-1}^2), \qquad k = 1, \dots, d-1. \tag{8}$$

# 2 Failure of conventional ODE methods

## 2.1 Conserved integrals

Following Toda [T], we associate with the matrix $L(t)$ the $d$-degree polynomial

$$p(\lambda) = \det(\lambda I - L) = \lambda^d - p_{d-1}\lambda^{d-1} + \cdots + (-1)^d p_0, \tag{9}$$

whereby the zeros of $p(\lambda)$ are the eigenvalues of the matrix $L$. The coefficients appearing in (9) can be expressed in terms of the matrix $L$ and its principal minors, or as elementary symmetric polynomials in its eigenvalues. Explicitely we have

$$\begin{aligned} p_{d-1} &= \lambda_1 + \cdots + \lambda_d, & (10)\\ p_{d-2} &= \sum_{i=1}^{d} \sum_{j=i+1}^{d} \lambda_i \lambda_j, & (11)\\ &\vdots & \\ p_0 &= \lambda_1 \lambda_2 \cdots \lambda_d. & (12) \end{aligned}$$

Since the flow is isospectral, the coefficients given by (10)–(12) are independent of time. Therefore they constitute $d$ integrals associated with the flow $L(t)$. It is shown in [T] how these constants are related to Henon's conserved quantities. However, we observe that the conservation of (10)–(12) is equivalent to the conservation of

$$\kappa_j = \sum_{i=1}^{d} \lambda_i^j, \qquad j = 1, \dots, d. \tag{13}$$

Since it is known from classical matrix theory (cf. [G]) that

$$\sum_{i=1}^{d} \lambda_i^j = \mathrm{tr}\,(L^j), \qquad j = 1, \ldots, d, \tag{14}$$

we use the latter integrals rather than (10)–(12) to carry out our theoretical analysis for numerical methods and their retention of isospectrality.

### 2.2 General approximation of the eigenvalues and conservation of the trace

Given an arbitrary numerical method for ODEs of order $p$, we expect the eigenvalues of $L$ to be approximated with the same precision as the entries of the matrix. When an eigenvalue is multiple, in all likehood its approximation will be less precise. However this cannot happen for irreducible Toda flows, since in this case the eigenvalues are all distinct. This is true since, expanding $\det(\lambda I - L)$ in the bottom row, we obtain a three-term recurrence relation. Provided that $L_0$ is irreducible, the Favard theorem implies that the recurrence relation generates orthogonal polynomials (in $\lambda$). The eigenvalues of $L_0$ are the zeros of the orthogonal polynomial of degree $d$ and they are all distinct by virtue of the separation theorem for zeros of orthogonal polynomials (cf. [C]).

Insofar as the approximation of the trace is concerned, we first introduce the following result.

**Lemma 1** *Given a skew-symmetric matrix $B$ and a symmetric matrix $C$, it is true that*

$$\mathrm{tr}\,[B, C] = 0.$$

*Proof.* Observe first that

$$BC - CB = BC + (BC)^T.$$

Therefore it is immediate to verify that

$$\mathrm{tr}\,[B, C] = \sum_{i=1}^{d} (BC - CB)_{i,i} = 2 \sum_{i=1}^{d} (BC)_{i,i}.$$

But

$$\sum_{i=1}^{d} (BC)_{i,i} = \sum_{i=1}^{d} \sum_{j=1}^{d} B_{i,j} C_{j,i} = -\sum_{i=1}^{d} \sum_{j=1}^{d} B_{j,i} C_{i,j},$$

where the latter equality holds by virtue of our hypothesis on $B$ and $C$. Changing the order of summation, it is clear that

$$\sum_{i=1}^{d} (BC)_{i,i} = 0.$$

$\square$

It is worthwhile to introduce the notation that we use in the remaining part of this paper. Given the autonomous differential system

$$\boldsymbol{y}' = \boldsymbol{f}(\boldsymbol{y}), \qquad \boldsymbol{y}(0) = \boldsymbol{y}_0, \qquad t \geq 0, \tag{15}$$

with $\boldsymbol{y}_0 \in \mathbb{R}^q$, $\boldsymbol{f} : \Omega \subseteq \mathbb{R}^q \to \mathbb{R}^q$, an $s$-stage RK scheme, defined by the following Butcher tableau

$$\begin{array}{c|c} \boldsymbol{c} & A \\ \hline & \boldsymbol{b}^T \end{array}, \tag{16}$$

produces the following time-stepping formula,

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h \sum_{i=1}^{s} b_i \boldsymbol{k}_i, \tag{17}$$

where

$$\boldsymbol{k}_i = \boldsymbol{f}(\boldsymbol{\phi}_i), \qquad i = 1, \ldots, s, \tag{18}$$

and

$$\boldsymbol{\phi}_i = \boldsymbol{y}_n + h \sum_{j=1}^{s} A_{i,j} \boldsymbol{k}_j, \qquad i = 1, \ldots, s, \tag{19}$$

are the internal stages vectors. The $\boldsymbol{k}_i$'s and $\boldsymbol{\phi}_i$'s depend on $n$ even if it is not explicitely stated in the latter formulae.

Similarly, a linear $s$-step method produces

$$\sum_{i=0}^{s} a_i \boldsymbol{y}_{n+i} = h \sum_{i=0}^{s} b_i \boldsymbol{f}(\boldsymbol{y}_{n+i}), \qquad a_s = 1. \tag{20}$$

This method can be characterized by the polynomials $(\rho, \sigma)$, where

$$\rho(z) = \sum_{i=0}^{s} a_i z^i, \qquad \sigma(z) = \sum_{i=0}^{s} b_i z^i, \tag{21}$$

and $z \in \mathbb{C}$. Both the RK scheme (16) and the linear multistep method (20) can be applied to differential equations in matrix form.

For further details we suggest the reader to consult texts in computational differential equations, for example [HNW], [Lb].

We have at this point all the necessary technical tools to introduce the following results.

**Theorem 2** *Every $s$-stage RK method with Butcher tableau* (16)*, when applied to the isospectral problem* (1)*, retains the trace of $L_0$.*

*Proof.* Consider an $s$-stage RK method defined by the Butcher tableau (16). Then, at each step,

$$L_{n+1} = L_n + h \sum_{i=1}^{s} b_i K_i.$$

Taking into account that the trace is a linear operation on matrices and that the $K_i$ are of the form $[B, C]$ with $B$ skew-symmetric and $C$ symmetric, it follows from Lemma 1 that

$$\mathrm{tr}\,(L_{n+1}) = \mathrm{tr}\,(L_n),$$

and the trace is retained. □

**Theorem 3** *Every consistent linear multistep method* (20) *when applied to the problem* (1)*, retains the trace of* $L_0$*.*

*Proof.* Assume that $\mathrm{tr}\,(L_n) = \mathrm{tr}\,(L_{n+1}) = \cdots = \mathrm{tr}\,(L_{n+s-1}) = \mathrm{tr}\,(L_0)$. Then, from Lemma 1 we have

$$\mathrm{tr}\,(L_{n+s}) = -\left(\sum_{i=0}^{s-1} a_i\right) \mathrm{tr}\,(L_0).$$

The theorem follows since $a_s = 1$ and, by virtue of consistency, $\rho(1) = 0$. □

It is evident that the conservation of the trace is shared by all standard numerical ODE methods. This is because the conservation of the trace is a linear conservation law, which is obeyed whenever we approximate the solution with linear combinations of its derivative values.

### 2.3 Quadratic conservation laws

Given the autonomous system of differential equations

$$\boldsymbol{y}' = \boldsymbol{f}(\boldsymbol{y}), \qquad \boldsymbol{y}(0) = \boldsymbol{y}_0 \in \mathbb{R}^q, \tag{22}$$

we say that the underlying flow $\boldsymbol{y}(t)$ obeys a quadratic conservation law if there exists a symmetric matrix $S \neq O$ such that

$$\boldsymbol{y}(t)^T S \boldsymbol{y}(t) = \text{const} \qquad \text{for all } t \geq 0. \tag{23}$$

If $\Omega \subseteq \mathbb{R}^q$ is the domain of definition of $\boldsymbol{f}$, differentiation affirms that (23) is equivalent to

$$\boldsymbol{\omega}^T S \boldsymbol{f}(\boldsymbol{\omega}) = 0 \qquad \forall \boldsymbol{\omega} \in \Omega. \tag{24}$$

Assume next that we are given the isospectral flow (1). Since the matrix $L$ is symmetric, it is true that

$$\mathrm{tr}\,(L^2) = \|L\|_F^2, \tag{25}$$

where the latter is the Frobenius norm on matrices, $\|L\|_F^2 = \sum_{i,j=1}^{d} L_{i,j}^2$. Therefore, this norm of $L(t)$ is retained. Furthermore observe that the Frobenius norm is equivalent to the Euclidean norm on vectors if we order the matrix $L$ column-wise, say. Let $\boldsymbol{y}$ be the vector obtained in this manner. The isospectral problem (1) can be written in the autonomous form (22) where $\boldsymbol{y}(t)$ obeys the quadratic conservation law (23) with $S \equiv I$, since

$$\boldsymbol{y}^T \boldsymbol{y} = \|\boldsymbol{y}\|_2^2 = \text{const}.$$

Hence, in order to have an isospectral method for (1), we require that it recovers quadratic integrals when applied to the autonomous vector system (22).

In the sequel we restrict our attention to RK methods, since Eirola and Sanz-Serna have shown in [ES] (cf. also [DRV]) that linear multistep methods that retain quadratic integrals have poor stability features.

Suppose that we have an $s$-stage RK method defined by the tableau (16), with which we associate the symmetric matrix $M$ whose entries are

$$M_{i,j} = b_i A_{i,j} + b_j A_{j,i} - b_i b_j, \qquad i,j = 1,\ldots,s. \tag{26}$$

Then the following result holds.

**Theorem 4** *The RK method* (16) *recovers all the quadratic conservation laws of the form* (23) *if and only if*

$$M = O. \tag{27}$$

*Proof.* The sufficiency has been essentially shown by Cooper in [Co] and follows from algebraic stability type considerations. However, in reporting his result, we prefer to follow Sanz-Serna (see [Sz], [SzC]). Recalling the time-stepping formula (16) for a RK method, by virtue of (24), it is possible to show that

$$\begin{aligned}
\boldsymbol{y}_{n+1}^T S \boldsymbol{y}_{n+1} &= \boldsymbol{y}_n^T S \boldsymbol{y}_n + h^2 \sum_{i,j=1}^{s} (b_i b_j - b_i A_{i,j} - b_j A_{j,i}) \boldsymbol{k}_i^T S \boldsymbol{k}_j \\
&= \boldsymbol{y}_n^T S \boldsymbol{y}_n - h^2 \sum_{i,j=1}^{s} M_{i,j} \boldsymbol{k}_i^T S \boldsymbol{k}_j.
\end{aligned} \tag{28}$$

If $M = O$ it is clear that the method recovers the quadratic integral (23) since

$$\boldsymbol{y}_{n+1}^T S \boldsymbol{y}_{n+1} = \boldsymbol{y}_n^T S \boldsymbol{y}_n,$$

and sufficiency follows. Insofar as necessity is concerned, Sanz-Serna and Verwer in [SzV] (cf. the Appendix) have mentioned the condition $M = O$ as sufficient and necessary for conservation of all quadratic laws of the form (23). As far as we are aware [Sanz-Serna, 1995, personal communication], the proof of necessity has not been published yet. Our presentation of necessity proof is further motivated since it applies to isospectral flows, as well as to higher order conservation laws. Thus, we consider the differential equation

$$\begin{bmatrix} y^{(1)} \\ y^{(2)} \end{bmatrix}' = \begin{bmatrix} f^{(1)}(\boldsymbol{y}) \\ f^{(2)}(\boldsymbol{y}) \end{bmatrix}, \qquad \boldsymbol{y}(0) = \begin{bmatrix} \beta_0 \\ \alpha_0 \end{bmatrix},$$

where $\boldsymbol{y} = [y^{(1)}, y^{(2)}]^T \in \mathbb{R}^2$. To avoid confusion with indices that stand for time-stepping, we denote the vector components using superscripts. Assume that $\boldsymbol{y}$ obeys the conservation law (23) with $S \equiv I$. Hence, by (24),

$$\boldsymbol{y}^T \boldsymbol{y}' = 0.$$

Therefore,

$$y^{(2)}f^{(2)}(\boldsymbol{y}) = -y^{(1)}f^{(1)}(\boldsymbol{y}),$$

or, in other words,

$$\boldsymbol{y}' = g(\boldsymbol{y})\begin{bmatrix} y^{(2)} \\ -y^{(1)} \end{bmatrix},$$

where $g : \mathbb{R}^2 \to \mathbb{R}$ is an arbitrary function. For our purposes, it is sufficient to consider one step of the method, from $\boldsymbol{y}_0$ to $\boldsymbol{y}_1$. We fix an index $1 \le \ell \le s$ and choose a smooth $g$ (for example by Lagrangian interpolation) such that

$$\begin{aligned} g(\boldsymbol{\phi}_\ell) &= 1, \\ g(\boldsymbol{\phi}_i) &= 0, \quad i = 1, \ldots, s, \quad i \neq \ell. \end{aligned}$$

This means, for (18), that

$$\boldsymbol{k}_i = \boldsymbol{f}(\boldsymbol{\phi}_i) = g(\boldsymbol{\phi}_i)\begin{bmatrix} \phi_i^{(2)} \\ -\phi_i^{(1)} \end{bmatrix} = \begin{cases} [\phi_i^{(2)}, -\phi_i^{(1)}]^T, & : i = \ell, \\ 0, & : i \neq \ell, \end{cases}$$

and consequently,

$$\boldsymbol{k}_i^T \boldsymbol{k}_j = \delta_{i,\ell}\delta_{j,\ell}\|\boldsymbol{k}_\ell\|^2, \qquad i,j = 1, \ldots, s.$$

Furthermore, we recall (19), namely

$$\boldsymbol{\phi}_i = \boldsymbol{y}_0 + hA_{i,\ell}\boldsymbol{k}_\ell, \qquad i = 1, \ldots, s.$$

Since $\boldsymbol{y}_0 = [\beta_0, \alpha_0]^T$, we use the $\ell$-th equation, $\boldsymbol{\phi}_\ell = \boldsymbol{y}_0 + hA_{\ell,\ell}\boldsymbol{k}_\ell$, to express the components of $\boldsymbol{\phi}_\ell$ in terms of the initial condition,

$$\begin{aligned} \phi_\ell^{(1)} &= \frac{\beta_0 + h\alpha_0 A_{\ell,\ell}}{1 + h^2 A_{\ell,\ell}^2}, \\ \phi_\ell^{(2)} &= \frac{\alpha_0 - h\beta_0 A_{\ell,\ell}}{1 + h^2 A_{\ell,\ell}^2}. \end{aligned}$$

These values can be used to find the remaining vectors $\boldsymbol{\phi}_i$. Specifically, for all $i \neq \ell$, we find

$$\begin{aligned} \phi_i^{(1)} &= \frac{\beta_0 + h\alpha_0 A_{i,\ell} + h^2\beta_0 A_{\ell,\ell}(A_{\ell,\ell} - A_{i,\ell})}{1 + h^2 A_{\ell,\ell}^2}, \\ \phi_i^{(2)} &= \frac{\alpha_0 - h\beta_0 A_{i,\ell} + h^2\alpha_0 A_{\ell,\ell}(A_{\ell,\ell} - A_{i,\ell})}{1 + h^2 A_{\ell,\ell}^2}. \end{aligned}$$

Hence, in correspondence to this initial condition and our function $g$, we deduce that

$$\sum_{i,j=1}^{s} M_{i,j}\boldsymbol{k}_i^T\boldsymbol{k}_j = M_{\ell,\ell}\|\boldsymbol{k}_\ell\|^2.$$

Since $\alpha_0, \beta_0$ can be chosen so that $\boldsymbol{k}_\ell \neq \mathbf{0}$, it is clear from (28) that $M_{\ell,\ell}$ must vanish. Repeating the same construction for all $\ell = 1, 2, \ldots, s$, we obtain

$$M_{\ell,\ell} = 0, \qquad \ell = 1, 2, \ldots, s,$$

and this proves the necessity of the condition for the diagonal entries of $M$. Next we fix again $1 \leq \ell \leq s-1$, and choose a function $g$ such that

$$g(\boldsymbol{\phi}_\ell) = g(\boldsymbol{\phi}_{\ell+1}) = 1, \qquad \text{and} \qquad g(\boldsymbol{\phi}_i) = 0, \quad i \neq \ell, \ell+1.$$

After the first step the method produces

$$\boldsymbol{\phi}_i = \boldsymbol{y}_0 + h(A_{i,\ell}\boldsymbol{k}_\ell + A_{i,\ell+1}\boldsymbol{k}_{\ell+1}), \qquad i = 1, \ldots, s.$$

From the $\ell$-th and $(\ell+1)$-st equations we can find $\boldsymbol{\phi}_\ell$ and $\boldsymbol{\phi}_{\ell+1}$ in terms of $\alpha_0, \beta_0$,

$$\begin{bmatrix} 1 & 0 & -hA_{\ell,\ell} & -hA_{\ell,\ell+1} \\ hA_{\ell,\ell} & hA_{\ell,\ell+1} & 1 & 0 \\ 0 & 1 & -hA_{\ell+1,\ell} & -hA_{\ell+1,\ell+1} \\ hA_{\ell+1,\ell} & hA_{\ell+1,\ell+1} & 0 & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\phi}_\ell \\ \boldsymbol{\phi}_{\ell+1} \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \alpha_0 \\ \beta_0 \\ \alpha_0 \end{bmatrix}.$$

Exchanging the second and third row, we obtain

$$\left( I + h \begin{bmatrix} O & -D \\ D & O \end{bmatrix} \right) \begin{bmatrix} \phi_\ell^{(1)} \\ \phi_{\ell+1}^{(1)} \\ \phi_\ell^{(2)} \\ \phi_{\ell+1}^{(2)} \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_0 \\ \alpha_0 \\ \alpha_0 \end{bmatrix},$$

where

$$D = \begin{bmatrix} A_{\ell,\ell} & A_{\ell,\ell+1} \\ A_{\ell+1,\ell} & A_{\ell+1,\ell+1} \end{bmatrix},$$

and $O$ is a zero $2 \times 2$ block. Hence, for sufficiently small $h > 0$, the system has a solution, and it can be used to find the other $\boldsymbol{\phi}_i$'s. Thus, taking into account the symmetry of $M$,

$$\begin{aligned} \sum_{i,j=1}^{s} M_{i,j}\boldsymbol{k}_i^T\boldsymbol{k}_j &= M_{\ell,\ell}\|\boldsymbol{k}_\ell\|^2 + M_{\ell+1,\ell+1}\|\boldsymbol{k}_{\ell+1}\|^2 + 2M_{\ell,\ell+1}\boldsymbol{k}_\ell^T\boldsymbol{k}_{\ell+1} \\ &= 2M_{\ell,\ell+1}\boldsymbol{k}_\ell^T\boldsymbol{k}_{\ell+1}, \end{aligned} \tag{29}$$

since we have already proved that the diagonal elements of $M$ do vanish. In order for the contribution of the $h^2$ terms in (28) to vanish, necessarily

$$M_{\ell,\ell+1} = 0,$$

since, for $h$ sufficiently small, $\boldsymbol{k}_\ell^T\boldsymbol{k}_{\ell+1} \approx \|\boldsymbol{k}_\ell\|^2 > 0$. This procedure generalizes for all the diagonals of $M$. For example, to prove that $M_{\ell,\ell+2} = 0$, we can take $g(\boldsymbol{\phi}_\ell) = g(\boldsymbol{\phi}_{\ell+1}) = g(\boldsymbol{\phi}_{\ell+2}) = 1$ and $g(\boldsymbol{\phi}_i) = 0$ otherwise and so on. This completes the proof of our assertion that $M = O$ is necessary for the recovery of all quadratic conservation laws. □

Theorem 4 ensures that the RK method recovers all quadratic integrals *a priori*, that is to say regardless of the matrix $S$ and of the function $\boldsymbol{f}$. However, for isospectral flows the condition $M = O$ is necessary as well. The necessity is shown in the following result.

**Theorem 5** *If a RK method is isospectral for all flows* (1)*, then necessarily* $M = O$*.*

*Proof.* Consider the flow
$$L' = g(L)[\tilde{B}(L), L],$$
where $g : \mathbb{R}^{d\times d} \to \mathbb{R}, g \neq 0$, is an arbitrary function. The flow is isospectral, since, by letting
$$B(L) = g(L)\tilde{B}(L),$$
it can be written in the form (1), the matrix $B(L)$ being skew-symmetric. The proof follows similarly to Theorem 4, by virtue of the arbitrariness of the function $g$. $\square$

Moreover, since for $2 \times 2$ systems the conservation of quadratic laws is necessary and sufficient condition for isospectrality (recall that, by Theorem 1, $\mathrm{tr}\,(L_0)$ is conserved by all RK schemes), we have the following result.

**Corollary 5.1** *All RK methods for which* $M = O$ *are isospectral for* $2\times 2$ *systems.* $\square$

## 2.4 RK methods and higher order conservation laws

Quadratic conservation laws for RK schemes have been widely investigated in the last few years, mostly because the condition $M = O$ is the same as the *symplecticity* condition (cf. [SzC]). However no attention has been devoted to cubic conservation laws.

Consider the problem
$$\boldsymbol{y}' = \boldsymbol{f}(\boldsymbol{y}), \quad \boldsymbol{y}(0) = \boldsymbol{y}_0 \in \mathbb{R}^q.$$
We say that $\boldsymbol{y}$ obeys a cubic conservation law if there exists a trilinear function
$$S(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{\omega}) : \mathbb{R}^q \times \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}, \qquad S \not\equiv 0,$$
which is symmetric in $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{\omega}$, i.e. assumes the same value on all the permutations of $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{\omega})$, such that
$$S(\boldsymbol{y}(t), \boldsymbol{y}(t), \boldsymbol{y}(t)) = \text{const}, \tag{30}$$
for all $t \geq 0$. Equivalently
$$S(\boldsymbol{\omega}, \boldsymbol{\omega}, \boldsymbol{f}(\boldsymbol{\omega})) = 0 \qquad \forall \boldsymbol{\omega} \in \Omega. \tag{31}$$

Isospectral flows in a vector formulation obey cubic conservation laws. We assume in the remainder of this section that
$$M = O,$$
a condition that guarantees that the RK method recovers quadratic integrals. Since $S$ is symmetric and linear in each of its variables, it is easy to verify that
$$\begin{aligned} S(\boldsymbol{y}_{n+1}, \boldsymbol{y}_{n+1}, \boldsymbol{y}_{n+1}) - S(\boldsymbol{y}_n, \boldsymbol{y}_n, \boldsymbol{y}_n) \quad = \quad & S(\boldsymbol{y}_{n+1} - \boldsymbol{y}_n, \boldsymbol{y}_{n+1}, \boldsymbol{y}_{n+1}) \\ & + S(\boldsymbol{y}_{n+1} - \boldsymbol{y}_n, \boldsymbol{y}_n, \boldsymbol{y}_{n+1}) \\ & + S(\boldsymbol{y}_{n+1} - \boldsymbol{y}_n, \boldsymbol{y}_n, \boldsymbol{y}_n), \end{aligned}$$

therefore it follows that the numerical method obeys (30) if and only if the right hand side vanishes. In other words, we require

$$\mathcal{I} = S(\boldsymbol{y}_{n+1} - \boldsymbol{y}_n, \boldsymbol{y}_{n+1}, \boldsymbol{y}_{n+1}) + S(\boldsymbol{y}_{n+1} - \boldsymbol{y}_n, \boldsymbol{y}_n, \boldsymbol{y}_{n+1}) + S(\boldsymbol{y}_{n+1} - \boldsymbol{y}_n, \boldsymbol{y}_n, \boldsymbol{y}_n) \tag{32}$$

to vanish. Because of the time-stepping formula (17), we can write

$$\boldsymbol{y}_{n+1} - \boldsymbol{y}_n = h \sum_{i=1}^{s} b_i \boldsymbol{k}_i,$$

hence substitution in (32) yields

$$\mathcal{I} = h \sum_{i=1}^{s} b_i \left\{ S(\boldsymbol{k}_i, \boldsymbol{y}_{n+1}, \boldsymbol{y}_{n+1}) + S(\boldsymbol{k}_i, \boldsymbol{y}_{n+1}, \boldsymbol{y}_n) + S(\boldsymbol{k}_i, \boldsymbol{y}_n, \boldsymbol{y}_n) \right\}.$$

The next step consists of expanding $\boldsymbol{y}_{n+1}$ in terms of $\boldsymbol{y}_n$ using again the time-stepping formula (17). We obtain

$$\begin{aligned} S(\boldsymbol{k}_i, \boldsymbol{y}_{n+1}, \boldsymbol{y}_{n+1}) &= S(\boldsymbol{k}_i, \boldsymbol{y}_n, \boldsymbol{y}_n) + 2h \sum_{j=1}^{s} b_j S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{y}_n) \\ &\quad + h^2 \sum_{j,m=1}^{s} b_j b_m S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{k}_m), \qquad (33) \\ S(\boldsymbol{k}_i, \boldsymbol{y}_{n+1}, \boldsymbol{y}_n) &= S(\boldsymbol{k}_i, \boldsymbol{y}_n, \boldsymbol{y}_n) + h \sum_{j=1}^{s} b_j S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{y}_n), \qquad (34) \end{aligned}$$

and substitution results in

$$\mathcal{I} = h \sum_{i=1}^{s} b_i \{ 3S(\boldsymbol{k}_i, \boldsymbol{y}_n, \boldsymbol{y}_n) + 3h \sum_{j=1}^{s} b_j S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{y}_n) + h^2 \sum_{j,m=1}^{s} b_j b_m S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{k}_m) \}.$$

Using (19) we can write $\boldsymbol{y}_n$ in terms of the $\boldsymbol{\phi}_i$'s, since

$$\boldsymbol{y}_n = \boldsymbol{\phi}_i - h \sum_{m=1}^{s} A_{i,m} \boldsymbol{k}_m.$$

Therefore,

$$\begin{aligned} S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{y}_n) &= S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{\phi}_i) - h \sum_{m=1}^{s} A_{i,m} S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{k}_m), \qquad (35) \\ S(\boldsymbol{k}_i, \boldsymbol{y}_n, \boldsymbol{y}_n) &= S(\boldsymbol{k}_i, \boldsymbol{\phi}_i, \boldsymbol{\phi}_i) - 2h \sum_{j=1}^{s} A_{i,j} S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{\phi}_i) \qquad (36) \\ &\quad + h^2 \sum_{j,m=1}^{s} A_{i,j} A_{i,m} S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{k}_m) \\ &= -2h \sum_{j=1}^{s} A_{i,j} S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{\phi}_i) + h^2 \sum_{j,m=1}^{s} A_{i,j} A_{i,m} S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{k}_m) \qquad (37) \end{aligned}$$

since $\boldsymbol{k}_i = \boldsymbol{f}(\boldsymbol{\phi}_i)$ and (31) implies that the first term in (36) vanishes. Substitution in $\mathcal{I}$ produces

$$\begin{aligned}\mathcal{I} &= 3h^2(-2\sum_{i,j=1}^{s} b_i A_{i,j} S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{\phi}_i) + \sum_{i,j=1}^{s} b_i b_j S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{\phi}_i)) \\ &\quad + h^3 \sum_{i,j,m=1}^{s} \{3b_i A_{i,j} A_{i,m} - 3b_i b_j A_{i,m} + b_i b_j b_m\} S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{k}_m).\end{aligned} \tag{38}$$

Let us focus on $S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{\phi}_i)$, observing that

$$S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{\phi}_i) = S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{\phi}_j) + h\sum_{m=1}^{s} (A_{i,m} - A_{j,m}) S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{k}_m).$$

This allows us to write

$$\begin{aligned}2\sum_{i,j=1}^{s} b_i A_{i,j} S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{\phi}_i) &= \sum_{i,j=1}^{s} b_i A_{i,j} S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{\phi}_i) + \sum_{i,j=1}^{s} b_i A_{i,j} S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{\phi}_j) \\ &\quad + h\sum_{i,j,m=1}^{s} b_i A_{i,j} (A_{i,m} - A_{j,m}) S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{k}_m).\end{aligned}$$

We find that the $\mathcal{O}(h^2)$ term in $\mathcal{I}$ vanishes, since

$$3\sum_{i,j=1}^{s} (b_i b_j - b_i A_{i,j} - b_j A_{j,i}) S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{\phi}_i) = 0,$$

because of the assumption $M = O$. Next we consider the $\mathcal{O}(h^3)$ terms. We are left with

$$\begin{aligned}\mathcal{I} &= -3h^3 \sum_{i,j,m=1}^{s} b_i A_{i,j} (A_{i,m} - A_{j,m}) S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{k}_m) + 3h^3 \sum_{i,j,m=1}^{s} b_i A_{i,j} A_{j,m} S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{k}_m) \\ &\quad - 3h^3 \sum_{i,j,m=1}^{s} b_i b_j A_{i,m} S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{k}_m) + h^3 \sum_{i,j,m=1}^{s} b_i b_j b_m S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{k}_m) \\ &= h^3 \sum_{i,j,m=1}^{s} \{3b_i A_{i,j} A_{j,m} - 3b_i b_j A_{i,m} + b_i b_j b_m\} S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{k}_m).\end{aligned} \tag{39}$$

We use once more $M = O$ to deduce that

$$b_i b_j = b_i A_{i,j} + b_j A_{j,i},$$

so that

$$\mathcal{I} = h^3 \sum_{i,j,m=1}^{s} \{3b_i A_{i,j} A_{j,m} - 3b_i A_{i,j} A_{i,m} - 3b_j A_{j,i} A_{i,m} + b_i b_j b_m\} S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{k}_m).$$

We observe that the first and the third term in the latter sum cancel since they coincide, $S$ being symmetric in its arguments. This, finally, leads to

$$\mathcal{I} = -h^3 \sum_{i,j,m=1}^{s} \Upsilon_{i,j,m} S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{k}_m), \tag{40}$$

where, by virtue of the symmetry of $S$,

$$\Upsilon_{i,j,m} = b_i A_{i,j} A_{i,m} + b_j A_{j,i} A_{j,m} + b_m A_{m,j} A_{m,i} - b_i b_j b_m, \qquad i,j,m = 1,\ldots,s. \tag{41}$$

Thus we deduce that

$$\Upsilon_{i,j,m} = 0, \qquad i,j,m = 1,2,\ldots,s, \tag{42}$$

is a sufficient condition for the conservation of all cubic integrals associated to the given differential equation.

**Necessity.** First let us consider the case of the implicit midpoint rule. In the sequel we generalize our analysis but the case study of the implicit midpoint rule is a good preparation for this task. The implicit midpoint (IMR) is a one-stage method ($s = 1$) and its Butcher tableau is

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}.$$

In other words

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h\boldsymbol{k}_1,$$

where

$$\begin{aligned} \boldsymbol{k}_1 &= \boldsymbol{f}(\boldsymbol{\phi}_1), \\ \boldsymbol{\phi}_1 &= \boldsymbol{y}_n + \tfrac{1}{2}h\boldsymbol{k}_1. \end{aligned}$$

A more popular form is

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h\boldsymbol{f}\left(\frac{\boldsymbol{y}_n + \boldsymbol{y}_{n+1}}{2}\right),$$

which can be easily obtained from the RK formalism.

So far, this method is a good candidate for isospectrality, since it is the only one-stage method that satisfies the condition $M = O$, which is equivalent to isospectrality in the $2 \times 2$ case. We find that $\mathcal{I}/h^3$ in (40) reduces to

$$(b_1^3 - b_1 A_{1,1}^3) S(\boldsymbol{k}_1, \boldsymbol{k}_1, \boldsymbol{k}_1).$$

However first we have to find the tensor $S$ in an explicit form. For a $3 \times 3$ Toda flow the matrix $L$ is

$$L = \begin{bmatrix} \beta_1 & \alpha_1 & 0 \\ \alpha_1 & \beta_2 & \alpha_2 \\ 0 & \alpha_2 & \beta_3 \end{bmatrix}.$$

We order $L$ as a vector, $\boldsymbol{y}$, so that

$$\boldsymbol{y} = [\beta_1, \beta_2, \beta_3, \alpha_1, \alpha_2]^T. \tag{43}$$

The differential system obtained in this way is

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \alpha_1 \\ \alpha_2 \end{bmatrix}' = \begin{bmatrix} 2\alpha_1^2 \\ 2(\alpha_2^2 - \alpha_1^2) \\ -2\alpha_2^2 \\ \alpha_1(\beta_2 - \beta_1) \\ \alpha_2(\beta_3 - \beta_2) \end{bmatrix}, \tag{44}$$

which is equivalent to the formulation (4). The corresponding tensor $S$ has 125 components $S_{i_1,i_2,i_3}$, with $i_1, i_2, i_3 \in \{1, \ldots, 5\}$. To find them, recall that $S$ is symmetric and satisfies the condition

$$S(\boldsymbol{y}, \boldsymbol{y}, \boldsymbol{y}) = \operatorname{tr}(L^3) = \sum_{\ell=1}^{3} \beta_\ell^3 + 3\sum_{\ell=1}^{2} \alpha_\ell^2(\beta_\ell + \beta_{\ell+1}).$$

Since

$$S(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \sum_{i_1,i_2,i_3=1}^{5} S_{i_1,i_2,i_3} x^{(i_1)} y^{(i_2)} z^{(i_3)},$$

it is clear that

$$S_{i,i,i} = 1, \qquad i = 1, 2, 3,$$

and

$$S_{1,4,4} = S_{2,4,4} = S_{2,5,5} = S_{3,5,5} = 1,$$

where the latter formula extends to all the possible permutations of the same indices. All the remaining components of $S$ are zero. Then the vector $\boldsymbol{k}_1$ is given by

$$\boldsymbol{k}_1 = [2\tilde{\alpha}_1^2, 2(\tilde{\alpha}_2^2 - \tilde{\alpha}_1^2), -2\tilde{\alpha}_2^2, \tilde{\alpha}_1(\tilde{\beta}_2 - \tilde{\beta}_1), \tilde{\alpha}_2(\tilde{\beta}_3 - \tilde{\beta}_2)]^T,$$

where the tilde means that the functions $\beta_i, \alpha_i$ are evaluated at the point $h/2$. By (40),

$$\frac{\mathcal{I}}{h^3} = (b_1^3 - 3b_1A_{1,1}^2)S(\boldsymbol{k}_1, \boldsymbol{k}_1, \boldsymbol{k}_1),$$

whereby explicit calculation yields

$$S(\boldsymbol{k}_1, \boldsymbol{k}_1, \boldsymbol{k}_1) = -6\tilde{\alpha}_1^2\tilde{\alpha}_2^2\{4(\tilde{\alpha}_2^2 - \tilde{\alpha}_1^2) - \tilde{\beta}_1^2 + \tilde{\beta}_3^2 + 2\tilde{\beta}_1\tilde{\beta}_2 - 2\tilde{\beta}_2\tilde{\beta}_3\}. \tag{45}$$

It is evident that we can choose an initial condition such that this quantity is nonzero. For example, choosing $\alpha_1(0), \beta_1(0)$ sufficiently large and the other quantities, including the stepsize $h$, sufficiently small, $S(\boldsymbol{k}_1, \boldsymbol{k}_1, \boldsymbol{k}_1)$ in (45) remains strictly positive. Therefore, in order for $\mathcal{I}$ to vanish, we must require that

$$\Upsilon_{1,1,1} = 3b_1A_{1,1}^2 - b_1^3 = 0,$$

but this is impossible, since $b_1 = 1$ and $A_{1,1} = 1/2$. Let us consider next a generic $s$-stage method. To prove necessity we start from the Toda lattice equations. Consider the problem (44) that we generalize in the following way

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \alpha_1 \\ \alpha_2 \end{bmatrix}' = g(\boldsymbol{y}) \begin{bmatrix} 2\alpha_1^2 \\ 2(\alpha_2^2 - \alpha_1^2) \\ -2\alpha_2^2 \\ \alpha_1(\beta_2 - \beta_1) \\ \alpha_2(\beta_3 - \beta_2) \end{bmatrix}, \tag{46}$$

where $\boldsymbol{y}$ is the vector in (43) and $g : \mathbb{R}^5 \to \mathbb{R}$ is an arbitrary Lipschitz function. In the first time step the method produces

$$\begin{aligned}
\boldsymbol{y}_1 &= \boldsymbol{y}_0 + h\sum_{i=1}^{s} b_i \boldsymbol{k}_i \\
\boldsymbol{\phi}_i &= \boldsymbol{y}_0 + h\sum_{j=1}^{s} A_{i,j} \boldsymbol{k}_j \qquad i = 1,\ldots,s, && (47)\\
\boldsymbol{k}_j &= g(\boldsymbol{\phi}_j)\boldsymbol{\psi}_j, \qquad j = 1,\ldots,s && (48)
\end{aligned}$$

where

$$\boldsymbol{\psi}_j = \begin{bmatrix} 2\phi_j^{(4)^2} \\ 2(\phi_j^{(5)^2} - \phi_j^{(4)^2}) \\ -2\phi_j^{(5)^2} \\ \phi_j^{(4)}(\phi_j^{(2)} - \phi_j^{(1)}) \\ \phi_j^{(5)}(\phi_j^{(3)} - \phi_j^{(2)}) \end{bmatrix} \qquad j = 1,\ldots,s.$$

We fix an index $\ell \in \{1, 2, \ldots, s\}$ and choose a smooth function $g$ such that

$$\begin{aligned}
g(\boldsymbol{\phi}_\ell) &= 1, \\
g(\boldsymbol{\phi}_i) &= 0, \qquad i = 1,\ldots,s, \quad i \neq \ell.
\end{aligned}$$

Then, from (48) we find

$$\boldsymbol{k}_i = 0, \qquad i \neq \ell,$$

and (47) reduces to

$$\boldsymbol{\phi}_i = \boldsymbol{y}_0 + hA_{i,\ell}\boldsymbol{\psi}_\ell, \qquad i = 1,\ldots,s. \tag{49}$$

Let $\boldsymbol{y}_0 = [\beta_1(0), \beta_2(0), \beta_3(0), \alpha_1(0), \alpha_2(0)]^T$. Then, for $i = \ell$, the equality (49) produces

$$\begin{aligned}
\phi_\ell^{(1)} - \beta_1(0) - 2hA_{\ell,\ell}\phi_\ell^{(4)^2} &= 0, \\
\phi_\ell^{(2)} - \beta_2(0) - 2hA_{\ell,\ell}(\phi_\ell^{(5)^2} - \phi_\ell^{(4)^2}) &= 0, \\
\phi_\ell^{(3)} - \beta_3(0) + 2hA_{\ell,\ell}\phi_\ell^{(5)^2} &= 0, \\
\phi_\ell^{(4)} - \alpha_1(0) - 2hA_{\ell,\ell}\phi_\ell^{(4)}(\phi_\ell^{(2)} - \phi_\ell^{(1)}) &= 0, \\
\phi_\ell^{(5)} - \alpha_2(0) - 2hA_{\ell,\ell}\phi_\ell^{(5)}(\phi_\ell^{(3)} - \phi_\ell^{(2)}) &= 0,
\end{aligned}$$

that can be written in the compact form as

$$\boldsymbol{F}(\boldsymbol{\phi}_\ell) = \boldsymbol{0}.$$

We can easily evaluate the Jacobian matrix $\partial\boldsymbol{F}/\partial\boldsymbol{\phi}_\ell$ of $\boldsymbol{F}$ to find

$$\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{\phi}_\ell} = I + hA_{\ell,\ell}\begin{bmatrix} O & -4D \\ D^T & G \end{bmatrix}, \tag{50}$$

where

$$G = -\begin{bmatrix} (\phi_\ell^{(2)} - \phi_\ell^{(1)}) & 0 \\ 0 & (\phi_\ell^{(3)} - \phi_\ell^{(2)}) \end{bmatrix}, \qquad D = \begin{bmatrix} \phi_\ell^{(4)} & 0 \\ \phi_\ell^{(4)} & \phi_\ell^{(5)} \\ 0 & \phi_\ell^{(5)} \end{bmatrix},$$

and $O$ is a zero $3 \times 3$ block. For sufficiently small $h$, the matrix (50) is invertible, hence, the implicit function theorem affirms that it is possible to find $\boldsymbol{\phi}_\ell$ in terms of $\boldsymbol{y}_0$. Then, by substitution in (48), we derive all the remaining $\boldsymbol{\phi}_i$'s. Note moreover that

$$\boldsymbol{k}_i = \mathbf{0} \quad \text{for } i \neq \ell \quad \Longrightarrow \quad S(\boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{k}_m) = 0 \quad \text{for } (i,j,m) \neq (\ell,\ell,\ell).$$

Hence $\mathcal{I}/h^3 = 0$ in (40) reduces to

$$-\Upsilon_{\ell,\ell,\ell} S(\boldsymbol{k}_\ell, \boldsymbol{k}_\ell, \boldsymbol{k}_\ell) = 0, \tag{51}$$

whereby $S(\boldsymbol{k}_\ell, \boldsymbol{k}_\ell, \boldsymbol{k}_\ell)$ is given, as for the implicit midpoint rule, by (45), the tilde meaning that the underlying functions are evaluated at $t = c_\ell h$. As we have seen for the IMR, the initial condition $\boldsymbol{y}_0$ can be chosen in order to have $S(\boldsymbol{k}_\ell, \boldsymbol{k}_\ell, \boldsymbol{k}_\ell) \neq 0$. Thus, we conclude that

$$\Upsilon_{\ell,\ell,\ell} = 0. \tag{52}$$

Here the index $\ell$ is arbitrary in $\{1, \ldots, s\}$, therefore (52) must be true for all $\ell$.

We do not intend to prove the necessity for the other components of $\Upsilon$. In fact, even though $\Upsilon = O$ is consistent with the order conditions, since, summing up the indices $i, j, m$ we find

$$\sum_{i,j,m=1}^{s} (3b_i A_{i,j} A_{i,m} - b_i b_j b_m) = 3 \sum_{i=1}^{s} b_i c_i^2 - 1,$$

hence consistency with order 2, unfortunately $\Upsilon_{\ell,\ell,\ell} = 0, \ell = 1, \ldots, s$, is contradictory with the assumption $M = O$. To verify this assertion, we consider an index $m \in \{1, 2, \ldots, s\}$ such that $b_m \neq 0$ (it exists because of the consistency condition $\sum_{i=1}^{s} b_i = 1$). Letting $i = j = m$, $\Upsilon_{m,m,m} = 0$ implies

$$A_{m,m} = \frac{b_m}{\sqrt{3}},$$

while $M_{m,m} = 0$ yields

$$A_{m,m} = \frac{b_m}{2}.$$

Since $b_m \neq 0$ this leads to a contradiction. This concludes the proof of the main results of the current section.

**Theorem 6** *No RK method can recover all cubic conservation laws.* □

**Corollary 6.1** *No RK method can be isospectral for all flows* (1) *when* $d \geq 3$.

*Proof.* The proof follows by considering that, in order to have isospectrality for $d \geq 3$, the method must recover cubic integrals. □

Apart from its importance to future analysis of isospectral integration, the latter results provide us also with a better insight into general conservative systems. Not only we have proved that quadratic and cubic conservation laws are conflicting requirements for RK schemes, but also we expect, in numerical integration, that even if the method obeys $M = O$, the error corresponding to the cubic integral will be of order $\mathcal{O}(h^{p+1})$. Note finally that we have used the condition $M = O$ to conclude that the $\mathcal{O}(h^2)$ term in $\mathcal{I}$ vanishes, hence, in a matter of fact, a RK method cannot conserve cubic laws even if we do not insist that the quadratic conservation laws are retained as well.

# 3 A family of isospectral methods: modified Gauss–Legendre Runge–Kutta schemes

## 3.1 Isospectral methods via Flaschka's formalism

Flaschka has shown (cf. [F], [T]) that the problem (1) is equivalent to solving

$$U' = B(L(t))U, \qquad U(0) = I, \qquad t \geq 0, \tag{53}$$

in tandem with

$$L(t) = U(t)L_0U(t)^T, \qquad t \geq 0. \tag{54}$$

As a consequence of skew-symmetry of $B(L)$, the matrix $U(t)$ in (53) remains unitary for all $t \geq 0$. Hence the main idea of this section is to use (53)–(54) and render them in a computational form. In particular we are interested in solving

$$U' = B(L)U, \qquad U(t_n) = I, \qquad t_n \leq t \leq t_{n+1}, \tag{55}$$

and then obtaining

$$L(t_{n+1}) = U(t_{n+1})L(t_n)U(t_{n+1})^T. \tag{56}$$

In the sequel we denote the numerical approximants using subscripts, that is $U_{n+1} \approx U(t_{n+1}), L_n \approx L(t_n), L_{n+1} \approx L(t_{n+1})$. As long as (55) is solved with a *unitary* method, the substitution in (56) produces an approximation $L_{n+1}$ which is orthogonally similar to $L_n$, hence isospectrality is retained.

Essentially, there are two ways of solving (55) preserving unitarity of $U$ (see [DRV]). The first one is to use *structural unitary integrators*, the second is to use *projected unitary methods*. Briefly (we recommend that the reader who is interested on this point consults [DRV]), structural unitary integrators are schemes which preserve unitarity whenever the underlying flow is unitary, while projected unitary methods consist of projecting (by means of a QR factorization) the numerical solution produced by an arbitrary method for ODEs, into the manifold of orthogonal matrices. The latter approach produces orthogonal matrices even when applied to non-unitary flows and might destroy time-reversibility, therefore we regard it as unsatisfactory in the present context. For this reason we prefer to consider structural unitary integrators. Insofar as standard ODE methods are concerned, the following result holds.

**Theorem 7** *Given a general $d \times d$ linear system of the form*

$$U' = S(t,U)U, \qquad U(0) = U_0, \qquad U_0^T U_0 = I, \tag{57}$$

*where $S(t,U)$ is skew-symmetric, let $\{U_n\}_{n=0}^{\infty}$ be the numerical approximation produced by a RK scheme. If $M = O$ then $U_n$ is orthogonal for all $n \geq 0$.*

*Proof.* Recall that

$$U_{n+1} = U_n + h \sum_{i=1}^{s} b_i K_i,$$

where, having denoted $S(t_n + c_\ell h, \Phi_\ell)$ by $S_\ell$,

$$\begin{aligned} K_\ell &= S_\ell \Phi_\ell, \qquad\qquad \ell = 1, \ldots, s, \\ \Phi_\ell &= U_n + h \sum_{j=1}^{s} A_{\ell,j} K_j. \end{aligned}$$

Hence

$$\begin{aligned} U_{n+1}^T U_{n+1} &= U_n^T U_n + h \sum_{\ell=1}^{s} b_\ell \{U_n^T K_\ell + K_\ell^T U_n\} \\ &\quad + h^2 \sum_{\ell=1}^{s} \sum_{j=1}^{s} b_\ell b_j K_\ell^T K_j. \end{aligned} \tag{58}$$

But

$$\begin{aligned} U_n^T K_\ell &= \Phi_\ell^T K_\ell - h \sum_{j=1}^{s} A_{\ell,j} K_j^T K_\ell, \\ K_\ell^T U_n &= K_\ell^T \Phi_\ell - h \sum_{j=1}^{s} A_{\ell,j} K_\ell^T K_j. \end{aligned}$$

However,

$$\begin{aligned} \Phi_\ell^T K_\ell + K_\ell^T \Phi_\ell &= \Phi_\ell^T S_\ell \Phi_\ell + \Phi_\ell^T S_\ell^T \Phi_\ell = \Phi_\ell^T (S_\ell + S_\ell^T) \Phi_\ell \\ &= O, \end{aligned}$$

since $S_\ell$ is a skew-symmetric matrix. Hence, (58) reduces to

$$U_{n+1}^T U_{n+1} = U_n^T U_n + h^2 \sum_{\ell=1}^{s} \sum_{j=1}^{s} \{b_\ell b_j - b_\ell A_{\ell,j} - b_j A_{j,\ell}\} K_j^T K_\ell.$$

Assuming that $U_n^T U_n = I$, it is clear that if $M = O$ then

$$U_{n+1}^T U_{n+1} = I.$$

Moreover, since the same holds when considering $U_{n+1} U_{n+1}^T$, the scheme retains unitarity while stepping from $t_n$ to $t_{n+1}$. The proof then follows by induction on $n$. □

Dieci, Russell and van Vleck have proved in [DRV] that $M = O$ and $\boldsymbol{b} \geq 0$ are sufficient conditions to retain unitarity for (57). Their result follows from stability type considerations, requiring that the RK scheme is B-stable for $t \to +\infty$ as well as for $t \to -\infty$. Our result shows that the condition on the vector of weights is not necessary and that unitarity is determined solely in terms of the matrix $M$.

Gauss–Legendre RK (GLRK) are examples of structural unitary integrators. However, there is a major problem in solving (55) even with GLRK schemes: the function $B(L)$ depends on $L(t)$. We need to know not only $L_n$ and eventually $L_{n+1}$ (and this can be done implicitly or using numerical approximation), but also the values of $L$ at the Gauss–Legendre nodes $t_n + c_\ell h, \ell = 1, \ldots, s$, the $c_\ell$'s being the zeros of the Legendre polynomials linearly transformed to $[0, 1]$. This information is not available in the standard formulation of a Runge–Kutta method. To overcome this difficulty, the theoretical problem (55) is hence replaced by an approximate one by introducing a polynomial $\tilde{L}(t)$ which interpolates the exact flow $L(t)$ at the points $t_n$ and $t_{n+1}$. Assume that we use an $s$-stage GLRK scheme (order $2s$) and that the function $B(L)$ is sufficiently smooth with respect to the variable $L$. Let

$$\tilde{L}^{(j)}(t_{n+i}) = L^{(j)}(t_{n+i}), \qquad i = 0, 1, \quad j = 0, 1, \ldots, s-1,$$

be the Hermite interpolant of $L$ of degree $2s - 1$ at $\{t_n, t_{n+1}\}$. Then the interpolation error is given by

$$\tilde{L}(t) - L(t) = \frac{1}{(2s)!}(t - t_n)^s (t - t_{n+1})^s L^{(2s)}(\theta_t)$$

for some $\theta_t \in (t_n, t_{n+1})$. If we replace $B(L)$ with $B(\tilde{L})$, the leading error term is

$$E(t) = \frac{1}{(2s)!}(t - t_n)^s (t - t_{n+1})^s L^{(2s)}(\theta_t) B'(\tilde{L}). \tag{59}$$

Therefore it is clear that $E(t) = \mathcal{O}(h^{2s})$. Next we show that this kind of interpolation does not affect the order of the GLRK method used.

## 3.2 Effects of the interpolation on the theoretical problem

Let $\tilde{L}(t)$ be the polynomial of degree $2s - 1$ in $t$ which interpolates $L^{(j)}(t)$, for $j = 0, 1, \ldots, s-1$, at $t_n$ and $t_{n+1}$ (order $2s$-interpolant). Moreover, we consider the following two differential equations, namely

$$\begin{aligned} U' &= B(L(t))U, \qquad U(t_n) = I, & (60)\\ V' &= B(\tilde{L}(t))V, \qquad V(t_n) = I. & (61) \end{aligned}$$

Obviously the latter can be regarded as a perturbation of (60). Denoting by $\Psi_B(t)$ the fundamental solution of (60), as a consequence of the Alekseev–Gröbner lemma we deduce

$$V(t) - U(t) = \int_{t_n}^{t} \Psi_B(t - \tau)\{V'(\tau) - BV(\tau)\}\, \mathrm{d}\tau,$$

therefore, $\Psi_B(t)$ being unitary for all $t$, we have in the 2-norm

$$\begin{aligned} \|V(t) - U(t)\| &\leq (t - t_n) \max_{t \in [t_n, t_{n+1}]} \|V' - B(t)V\| \leq h\, \|B(\tilde{L}(t)) - B(L(t))\|_\infty \times \|V\| \\ &= h\, \|B(\tilde{L}(t)) - B(L(t))\|_\infty = h\, \|E(t)\|_\infty. \qquad (62) \end{aligned}$$

Bearing in mind that, as we have formerly seen in (59), $\|E(t)\|_\infty$ is of order $\mathcal{O}(h^{2s})$, we deduce that the error in $V(t)$ is of order $\mathcal{O}(h^{2s+1})$, provided that we can bound $B'$ and the higher derivatives of $L$ appearing in the leading error term. Hence letting

$$L_{n+1} = V(t_{n+1})L(t_n)V(t_{n+1})^T,$$

we are committing at most an error of $\mathcal{O}(h^{2s+1})$, which is subsumed in the error of a numerical method of order $2s$.

### 3.3 Modified Gauss–Legendre RK methods

In the sequel we refer to a *modified* Gauss–Legendre RK method of order $2s$ whenever we apply the classical Gauss–Legendre RK method of order $2s$ to (61) in tandem with $L_{n+1} = V(t_{n+1})L(t_n)V(t_{n+1})^T$. The underlying equations are implicit, since we need $L_{n+1}$ and in general up to the $(s-1)$-st derivative at the point $t_{n+1}$ to derive the interpolating polynomial $\tilde{L}(t)$. Thus, we need to iterate and our choice is the simplest, namely the Picard iteration. We set (as an initial guess)

$$L_{n+1}^{(j)\,[0]} = L_n^{(j)}, \qquad j = 0, 1, \ldots, s-1, \tag{63}$$

then we can use $L_n^{(j)}, L_{n+1}^{(j)\,[k]}$, for $k = 0, 1, \ldots$, to solve

$$V^{[k]\prime} = B(\tilde{L}_{n+1}^{[k]})V^{[k]}, \qquad V^{[k]}(t_n) = I, \tag{64}$$

where $B(\tilde{L}_{n+1}^{[k]})$ denotes the function $B$ evaluated with the Hermite interpolant, in tandem with

$$L_{n+1}^{[k+1]} = V_{n+1}^{[k]} L_n V_{n+1}^{[k]\,T}. \tag{65}$$

A forthcoming paper will debate the convergence of the modified Gauss–Legendre RK methods for various isospectral flows.

### 3.4 The modified implicit midpoint rule

Consider the case $s = 1$, the implicit midpoint rule. Denoting

$$B_{n+\frac{1}{2}} = B\left(\tfrac{1}{2}(L_n + L_{n+1})\right), \tag{66}$$

the scheme yields

$$V_{n+1} = (I - \tfrac{1}{2}hB_{n+\frac{1}{2}})^{-1}(I + \tfrac{1}{2}hB_{n+\frac{1}{2}}). \tag{67}$$

Hence we let

$$L_{n+1} = V_{n+1} L_n V_{n+1}^T. \tag{68}$$

**Theorem 8** *The algorithm* (66)–(68) *is a second order isospectral modification of the implicit midpoint rule when applied to* (1).

*Proof.* Since $V$ is unitary, it follows that the algorithm (66)–(68) is isospectral. Moreover we can use (67) to eliminate $V_{n+1}$ in (68). This leads to

$$L_{n+1} = \left(I - \tfrac{1}{2}hB_{n+\frac{1}{2}}\right)^{-1} \left(I + \tfrac{1}{2}hB_{n+\frac{1}{2}}\right) L_n \left(I - \tfrac{1}{2}hB_{n+\frac{1}{2}}\right) \left(I + \tfrac{1}{2}hB_{n+\frac{1}{2}}\right)^{-1},$$

and, multiplying both sides of the latter equation by $\left(I - \tfrac{1}{2}hB_{n+\frac{1}{2}}\right)$ on the left and by $\left(I + \tfrac{1}{2}hB_{n+\frac{1}{2}}\right)$ on the right, we obtain

$$\left(I - \tfrac{1}{2}hB_{n+\frac{1}{2}}\right) L_{n+1} \left(I + \tfrac{1}{2}hB_{n+\frac{1}{2}}\right) = \left(I + \tfrac{1}{2}hB_{n+\frac{1}{2}}\right) L_n \left(I - \tfrac{1}{2}hB_{n+\frac{1}{2}}\right)$$

Thus

$$\begin{aligned} L_{n+1} - \tfrac{1}{2}hB_{n+\frac{1}{2}}L_{n+1} + \tfrac{1}{2}hL_{n+1}B_{n+\frac{1}{2}} - \tfrac{1}{4}h^2 B_{n+\frac{1}{2}}L_{n+1}B_{n+\frac{1}{2}} \\ = L_n + \tfrac{1}{2}hB_{n+\frac{1}{2}}L_n - \tfrac{1}{2}hL_nB_{n+\frac{1}{2}} - \tfrac{1}{4}h^2 B_{n+\frac{1}{2}}L_nB_{n+\frac{1}{2}} \end{aligned}$$

and, bearing in mind the definition of the commutator operator,

$$L_{n+1} - \tfrac{1}{2}h[B_{n+\frac{1}{2}}, L_{n+1}] - \tfrac{1}{4}h^2 B_{n+\frac{1}{2}}L_{n+1}B_{n+\frac{1}{2}} = L_n + \tfrac{1}{2}h[B_{n+\frac{1}{2}}, L_n] - \tfrac{1}{4}h^2 B_{n+\frac{1}{2}}L_nB_{n+\frac{1}{2}}.$$

Furthermore,

$$L_{n+1} = L_n + h\left[B_{n+\frac{1}{2}}, \frac{L_n + L_{n+1}}{2}\right] + \tfrac{1}{4}h^2 B_{n+\frac{1}{2}}(L_{n+1} - L_n)B_{n+\frac{1}{2}}. \tag{69}$$

It is clear from (69) that the method (66)–(68) adds to the implicit midpoint rule an extra term, namely

$$\tfrac{1}{4}h^2 B_{n+\frac{1}{2}}(L_{n+1} - L_n)B_{n+\frac{1}{2}}.$$

This extra term retains the second order of the method since it is $\mathcal{O}(h^3)$, while rendering the scheme isospectral. This is very important because, as we have proved in the previous section, the implicit midpoint rule is not isospectral, except in the $2 \times 2$ case. □

# 4 Numerical results

In order to illustrate some of the results in this paper, let us consider as a test problem the Toda lattice equations [F], [T]. This Hamiltonian system models the motion of a finite number of particles on the line with exponential interactions of nearest neighbours. We are interested in the nonperiodic case with $d$ particles. The Hamiltonian function is

$$H(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\sum_{k=1}^{d} p_k^2 + \sum_{k=1}^{d-1}\{\exp\left(2(q_k - q_{k+1})\right) - 1\},$$

where $q_k$ is the position of the $k$-th particle and $p_k$ is its momentum, $1 \le k \le d$. The equations of motion are

$$\begin{aligned} q_k' &= p_k, \\ p_k' &= 2\left(\exp\left(2(q_{k-1} - q_k)\right) - \exp\left(2(q_k - q_{k+1})\right)\right), \end{aligned} \qquad k = 1, \ldots, d, \tag{70}$$

where it should be understood that $q_0 = q_{d+1} = 0$.

After introducing the new variables [F]

$$\begin{array}{rcll} \beta_k & = & -p_k, & 1 \leq k \leq d, \\ \alpha_k & = & \exp(q_k - q_{k+1}), & 1 \leq k \leq d-1, \end{array} \tag{71}$$

the differential system may be rewritten as (4), or as the matrix equation (1) with $B(L) = L_+ - L_-$ and $L_0$ a symmetric tridiagonal matrix. As it was mentioned in Section §2, there is a set of $d$ integrals of motion [T] which are related to the eigenvalues of the symmetric matrix $L_0$. These $d$ integrals are in involution and then the differential system (70) is a completely integrable Hamiltonian system.

In the experiments reported here we integrate a lattice with three particles in the interval $0 \leq t \leq 640$ with initial conditions

$$\mathbf{q}(0) = [0, 0, 0]^T, \qquad \mathbf{p}(0) = [1, -0.5, -0.5]^T.$$

Note that three is the minimum number of particles we must consider in order to see the different behaviour of isospectral methods and conventional ODE integrators which preserve quadratic invariants.

In the numerical experiments the integration has been performed using two different methods of order two. The first method is just the implicit midpoint rule (IMR) applied to the differential equation (1), implemented with fixed point iteration for solving the implicit equations. The second method is the modified implicit midpoint rule (MIMR) introduced in §3.4 implemented as explained in §3.3. Both methods have been implemented using the same stepsizes and the same tolerances in order to solve the implicit equations.

In order to make a comparison between both methods, we have measured the errors in the variables $\alpha_k$'s and $\beta_k$'s at $t = 5, 10, 20, 40, \ldots, 640$, using the Euclidean norm of $\mathbb{R}^5$. This is equivalent to measuring the Frobenius norm of the error in the numerical approximation to the matrix $L(t)$ at the same time levels.

Figure 1 gives error against time for the methods being considered, with stepsizes 1/8 and 1/32. The stars joined by a solid line correspond to the implicit midpoint rule while the circles joined by the dotted line represent the errors for the isospectral method. We see that the errors for the implicit midpoint rule are almost constant but the errors for the modified method decay with time. In fact, the main contribution to the errors corresponding to the IMR comes from the error in the diagonal elements of the solution, i.e. in the $\beta_k$'s. The error in the off-diagonal elements becomes almost negligible with time.

In Figure 2 we have plotted the Euclidean norm of the off-diagonal elements of the numerical solutions with stepsize 1/8 against time. In order to represent the data generated with both methods we use the same symbols as in Figure 1. In fact, IMR and MIMR display very similar behaviour. For both methods we observe convergence of the numerical solution to a diagonal matrix as $t$ increases. This fact is also true for the exact solution as it was proved by Moser in [Mo]. However, the almost constant errors for the implicit midpoint rule indicates that the limit matrix for IMR is not the right diagonal matrix of the eigenvalues of $L_0$ as it should be. Recall that although both methods preserve the

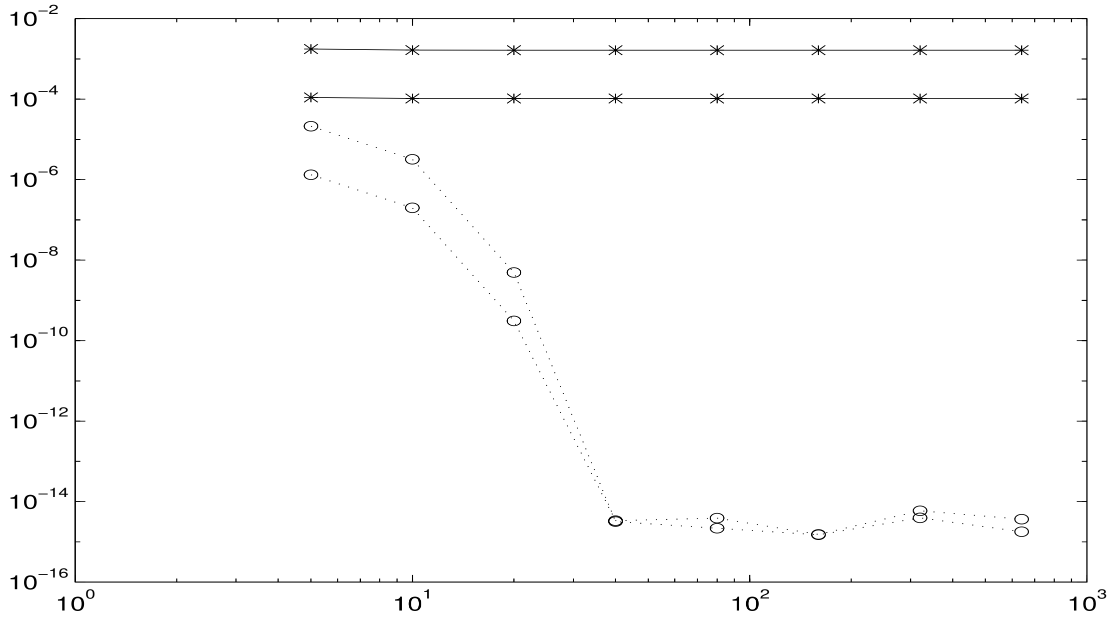


Figure 1: Error against time

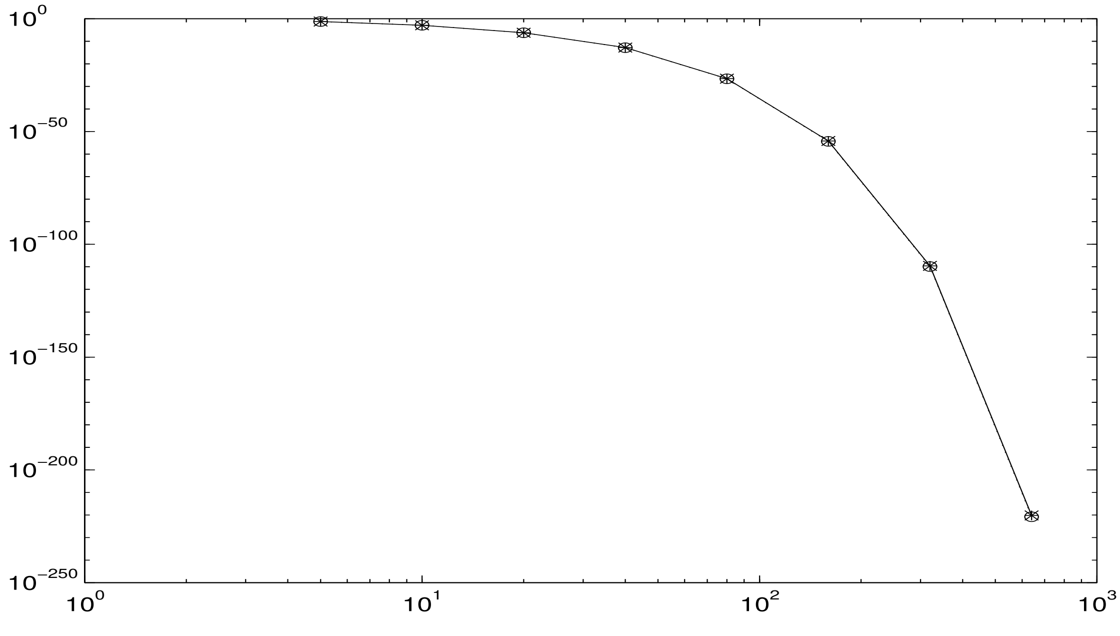


Figure 2: Norm of the off-diagonal elements against time

trace and the quadratic invariants of the solution, only the modified implicit midpoint rule recovers the third integral of motion.

**Acknowledgements.** The authors wish to express their gratitude to a number of colleagues with whom they have discussed the subject matter of this paper and whose comments have been most helpful and welcome: Brad Baxter, Moody Chu, Jim Demmel, Jeffrey Lagarias, Dirk Laurie, Ben Leimkuhler, Luciano Lopez, Beresford Parlett, Andrew Stuart, David Watkins and Andre Weidemann. The work of the first author has been supported by DGICYT under project PB92-254.

## References


[B1] R. W. Brockett, "Dynamical systems that sort lists, diagonalize matrices, and solve linear programming problems", *Lin. Alg. Appl.*, **146** (1991), 79–91.

[BBR] A. M. Bloch, R. W. Brockett and T. Ratiu, "A new formulation of the generalized Toda lattice equations and their fixed point analysis via the momentum map", *Bull. Amer. Math. Soc.*, **23** (1990), 477–485.

[C] T. S. Chihara, *An Introduction to Orthogonal Polynomials*, Gordon and Breach Science Publishers, New York, London, Paris, 1978.

[Ch] M. T. Chu, "The generalized Toda flow, the QR algorithm and the center manifold theory", *SIAM J. Discr. Math.*, **5** (1984), 187–803.

[ChD1] M. T. Chu and K. R. Driessel, "The projected gradient method for least squares matrix approximations with spectral constraints", *SIAM J. Numer. Anal.*, **27** (1990), 1050–1060.

[ChD2] M. T. Chu and K. R. Driessel, "Can real symmetric Toeplitz matrices have arbitrary real spectra?", Technical report, Idaho State University.

[Co] G. J. Cooper, "Stability of Runge–Kutta methods for trajectory problems", *IMA J. Numer. Anal.*, **7** (1987), 1–13.

[D] K. R. Driessel, "On isospectral gradient flows solving matrix eigenproblems using differential equations", *Inverse Problems*, J. R. Cannon and U. Hournung eds., Birkhauser-Verlag (1986), 69–90.

[DNT] P. Deift, T. Nanda and C. Tomei, "Ordinary differential equations and the symmetric eigenvalue problem", *SIAM J. Numer. Anal.*, **20** (1983), 1–22.

[DRTW] P. Deift, S. Rivera, C. Tomei and D. Watkins, "A monotonicity property for Toda-Type flows", *SIAM J. Matrix Anal. Appl.*, **12** (1991), 463–468.

[DRV] L. Dieci, R. D. Russell and E. S. van Vleck, "Unitary integrators and applications to continuous orthonormalization techniques", *SIAM J. Numer. Anal.*, **31** (1994), 261–281.

[ES] T. Eirola and J. M. Sanz-Serna, "Conservation of integrals and symplectic structure in the integration of differential equations by multistep methods", *Numer. Math.*, **61** (1992), 281–290.

[F] H. Flaschka, "The Toda lattice", I. Phys. Rev. **B9** (1974), 1924–25.

[G] F. R. Gantmacher, *The Theory of Matrices*, Chelsea, New York, 1977.

[GRPD] G. Gaeta, C. Reiss, M. Peyrard and T. Dauxois, "Simple models for nonlinear DNA dynamics", *Rivista Nuovo Cimento*, **17** (1994), 1–47.

[HM] U. Helmke and J. B. Moore, "Singular value decomposition via gradient flows", *Systems Control Lett.*, **14** (1990), 369–377.

[HNW] E. Hairer, S. P. Norsett and G. Wanner, *Solving Ordinary Differential Equations I. Nonstiff Problems. Second Revised Edition*, Springer, Berlin, 1993.

[I] A. Iserles, "Solving linear ordinary differential equations by exponentials of iterated commutators", *Numer. Math.*, **45** (1984), 183–199.

[KvM] M. Kac and P. von Moerbeke, "On explicitly solvable systems of differential equations related to certain Toda lattices", *Adv. in Math.*, **16** (1975), 160–169.

[L] J. Lagarias, "Monotonicity properties of the Toda flow, the QR-flow and subspace iteration", *SIAM J. Matrix Anal. Appl.*, **12**, **3** (1991), 449–462.

[Lb] J. D. Lambert *Numerical Methods for Ordinary Differential Systems*, John Wiley & Sons, Chichester-New York-Brisbane-Toronto-Singapore (1991).

[MMH] J. B. Moore, R. E. Mahony and U. Helmke, "Numerical gradient algorithms for eigenvalue and singular value calculations", *SIAM J. Matrix Anal. Appl.*, **15** (1994), 881–902.

[Mo] J. Moser, "Finitely many mass points on the line under the influence of an exponential potential – An integrable system", *Dynamic Systems Theory and Applications*, (J. Moser, ed.) Springer-Verlag, New York, Berlin, Heildelberg, 1975, 467–497.

[Na1] T. Nanda, Ph.D. Thesis, New York Univ., New York, 1982.

[Na2] T. Nanda, "Differential equations and the QR algorithm", *SIAM J. Numer. Anal.*, **22** (1985), 310–321.

[T] M. Toda, *Theory of Nonlinear Lattices*, Springer-Verlag, Berlin, Heildelberg, New York (1981).

[Sy] W. W. Symes, "The QR algorithm and scattering for the finite nonperiodic Toda lattice", *Phys. D.*, **4** (1982), 275–180.

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

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

[SzV] J. M. Sanz-Serna and J. G. Verwer, "Conservative and nonconservative schemes for the solution of the nonlineal Schrödinger equation", *IMA J. Numer. Anal.*, **6** (1986), 25–42.

[W] D. S. Watckins, "Isospectral flows", *SIAM Review,* **26** (1984), 379–391.

[WE] D. Watkins and L. Elsner, "Self-equivalent flows associated with the singular value decomposition", *SIAM J. Matrix Anal. Appl.*, **10** (1989), 244–258.