UNIVERSITY OF CAMBRIDGE

Numerical Analysis Reports

Qualitative numerical analysis of ordinary differential equations

Arieh Iserles and Antonella Zanna

DAMTP 1995/NA5
September 1995

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

Qualitative numerical analysis of ordinary differential equations

Arieh Iserles[1] and Antonella Zanna[2]

Abstract The traditional goal of numerical analysis is to devise and understand algorithms that approximate fast, robustly and within given error tolerance the solution of an underlying mathematical problem. In this paper we advance a philosophy that, while not at variance with this goal, calls for an important change of emphasis. In many problems of interest, mathematical analysis provides us with a partial knowledge of the solution, its properties and possibly its invariants. We suggest that such problems be computed by numerical methods that share their qualitative attributes. Thus, the recovery of the correct qualitative behaviour of a mathematical problem should be seen as a major criterion in devising and analysing numerical algorithms.

We specialise our discussion to three examples, all within the broader framework of ordinary differential equations and Runge–Kutta methods: unitary flows, symplectic flows and isospectral flows.

[1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, England.

[2]Newnham College, University of Cambridge, Cambridge, England.

1 A new paradigm for computation

What is the goal of numerical analysis? A typical text on numerical solution of ordinary differential equations, say, is likely to commence by stating the problem. Thus,

$$\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y}), \quad t \geq t_0, \qquad \boldsymbol{y}(t_0) = \boldsymbol{y}_0 \in \mathbb{R}^d,$$

say, where $\boldsymbol{f} : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ is globally Lipschitz. Next step consists of a thorough and substantive discussion of the different kinds of numerical methods – multistep, Runge–Kutta, extrapolation etc. – and their attributes: order, stability and so on. All this is underlied by the implicit assumption that the subject of the enquiry is the above 'general equation'. We need the Lipschitz condition to argue existence and uniqueness of the exact solution and, subsequently, this condition is hardly ever mentioned. Occasionally extra requirements are demanded from $\boldsymbol{f}$, so as to assist in the choice of the correct numerical algorithm. For example, if $\boldsymbol{f}$ is sufficiently differentiable then the problem can be solved by high-order methods, if the eigenvalues of its Jacobian differ widely in size then the equation is probably stiff and requires a 'stiff solver' and so on. Note a glaring absence in the aforementioned list: we never ask what is known about the *particular* equation that we attempt to solve. In other words, we never request assistance from the theory of differential equations.

The textbook of J.D. Lambert [18], perhaps the most popular modern exposition of numerical ordinary differential equations, concludes with 125 items of bibliography, *none on the theory of differential equations.* This is typical: numerical analysis starts at the exact point when mathematical analysis leaves and the twain shall never meet. By this unstated consensus, the goal of numerical analysis is to devise, analyse and finally code a 'master method', that can cater for the maximal range of possible inputs.

It is our contention that this state of affairs is unsatisfactory. It is often beneficial and occasionally essential to consider theory and computation of ordinary differential equations in tandem, as part of a broader canvass. For decades numerical analysts have been repeating Hamming's mantra, 'the purpose of computing is insight, not numbers' [13], yet observing it mainly in the breach and maintaining strict separation between the qualitative and the quantitative. But analysis and computation need each other! Inasmuch as mathematical analysis can seldom describe a solution completely, it usually comes up with important partial information. Particular mention deserve two types of qualitative results for which the theory is particularly well-suited, namely the determination of asymptotic behaviour and the investigation of invariants and conservation laws which are obeyed by the underlying flow. It is wasteful to abandon this knowledge as soon as computation is being attempted and it seems to us (and to an increasingly larger proportion of our colleagues) that a preferable course of action is to incorporate analytic results into the numerical method.

We mention in passing that this linkage of analysis and computation bears benefits to theory as well. Numerical results do not obviate the need for rigourous theorems and complete proofs, but they surely point us in the right direction. A whole range of interesting phenomena – chaotic trajectories, metastability, inertial manifolds...– have been revealed first in numerical experimentation, only subsequently succumbing to mathematical analysis. We will not pursue this issue further in the present paper.

There is an important price tag to the incorporation of qualitative information into numerical algorithms: generality is abandoned. We are no longer seeking methods and software that can cater for just about every ordinary differential equation of interest. The focus shifts to families of differential equations that share an important structural feature and a basic rule of the thumb is that different equations deserve different numerical treatment.

Needless to say, general methods and software remain a highly worthwhile subject of research and they are useful in a multitide of applications. In particular, they are important when the function $\boldsymbol{f}$ is imprecisely known or excessively complicated or when the user (a scientist or an engineer) does not wish to be bothered with mathematical intricacies, just to produce a reasonably good approximation to the solution.

We have already mentioned the two most important kinds of qualitative information, namely asymptotics and conservation laws. There is nothing new, in principle, insofar as incorporation of asymptotic considerations into computational ordinary differential equations is concerned. It is well known since the Fifties that the presence of processes that progress and decay at vastly different time scales – a phenomenon known as *stiffness* – has important ramifications on the choice of a method and of a step-size. Specifically, at least for some popular methods, retention of correct asymptotic behaviour necessitates radical reduction in the step-size, which often renders the computation unviable. Understanding of stiffness has been traditionally based on a linear model, although this has changed with the introduction of the monotone model by Germund Dahlquist [7]. Given an inner product $\langle\,\cdot\,,\,\cdot\,\rangle$, we say that the ODE

$$\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y}), \qquad t \geq t_0, \tag{1.1}$$

is *monotone* if

$$\langle \boldsymbol{f}(t,\boldsymbol{x}) - \boldsymbol{f}(t,\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle \leq 0, \qquad t \geq t_0, \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d.$$

It is easy to prove that monotonicity implies that any two solutions $\boldsymbol{x}$ and $\boldsymbol{y}$ of (1.1) obey

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{x}(t) - \boldsymbol{y}(t)\|^2 \leq 0$$

(where $\|\cdot\|$ is the induced norm), hence (1.1) is contractive. Substantial effort has been spent in the last two decades in singling out numerical methods that recover contractivity of monotone systems without any need to depress the step-size.

The last few years have witnessed growing sophistication of nonlinear stability models. Perhaps the most interesting novel model is when $\alpha, \beta > 0$ exist so that

$$\langle \boldsymbol{f}(t,\boldsymbol{y}), \boldsymbol{y}\rangle \leq \alpha - \beta\|\boldsymbol{y}\|^2, \qquad t \geq t_0, \quad \boldsymbol{y} \in \mathbb{R}^d. \tag{1.2}$$

In that case it is not difficult to prove that there exists a bounded set $\mathcal{A} \subset \mathbb{R}^d$ which attracts all solution trajectories – a *global attractor.*

It is not our intention, however, to address ourselves in this article to questions of asymptotic behaviour, that have been already described at great length and erudition in a survey by Andrew Stuart [26]. Our focus here is on the retention of invariants by the numerical solution. This is the place to clarify that this subject area is still in its infancy. Having said this, we believe that there is enough material for a preliminary exposition – and that the results so far whet the appetite for future research.

Suppose that it is known on theoretical grounds that the solution of (1.1) obeys an algebraic equation of the form

$$\varphi(\boldsymbol{y}(t)) = 0, \qquad t \geq t_0. \tag{1.3}$$

The function $\varphi \in C[\mathbb{R}^d]$ is allowed to depend on the initial values $(t_0, \boldsymbol{y}_0)$. Often it is important to require that the numerical solution $\{\boldsymbol{y}_n\}_{n=0}^{\infty}$ also satisfies the conservation law, that is $\varphi(\boldsymbol{y}_n) = 0$, $n \in \mathbb{Z}^+$. The obvious reason is that the functional φ represents qualitative information about the solution, which we might be keen to retain under discretization. However, there often exist more subtle reasons for numerical conservation. Thus, sometimes a conservation of the 'right' quantity implies that the numerical error accumulates slower (cf. Section 4). In other cases, the whole purpose of the exercise is to reach a certain point on the manifold

$$\mathcal{M} = \{\boldsymbol{x} \in \mathbb{R}^d : \varphi(\boldsymbol{x}) = 0\}$$

and it is essential to prevent a numerical method from departing off $\mathcal{M}$ (cf. Section 5).

On the face of it, we can formulate the ODE, together with the conservation law (1.3), as a DAE (differential-algebraic equation)

$$\begin{aligned} \boldsymbol{y}' &= \boldsymbol{f}(t, \boldsymbol{y}), \qquad t \geq 0, \\ 0 &= \varphi(\boldsymbol{y}), \end{aligned}$$

and apply standard DAE methods. This is quite appealing since many reliable DAE methods are available [20] but, unfortunately, completely wrong. The above DAE system is overdetermined: the conservation law does not represent a genuine addition of intrinsic information to the ordinary differential equation. After all, it has been deduced from the differential equation without adding any new information! Hence, it lies outside the scope of usual DAE theory and methods.

Two basic approaches to numerical recovery of conservation laws are available. Firstly, we may use an arbitrary numerical method to derive a candidate solution $\tilde{\boldsymbol{y}}_n \approx \boldsymbol{y}(t_n)$, say, and subsequently project it on the manifold $\mathcal{M}$ by methods of linear algebra and optimization. Secondly, we can build the conservation law into the numerical method or, alternatively, choose a numerical method that provably reproduces (1.3). Both approaches, which we term *projective* and *structural* respectively, are viable, although we must confess a certain affinity to the second one, mainly on æsthetical grounds. Moreover, projection is sometimes liable to corrupt advantageous features of the numerical solution, e.g. time-reversibility of Hamiltonian equations (cf. Section 4).

The remainder of this paper is devoted to three sets of ODEs, each with its own conservation laws, and to their numerical treatment. Although these are early days in 'qualitative numerical analysis', we wish to make the case that there is already enough novelty and usefulness in the reported results to justify further research and, perhaps more importantly, further identification and numerical treatment of other sets of ODEs that share important structural attributes. Above everything else, we wish to promote the idea that exciting new numerical analysis is bound to follow from better knowledge of ODE theory.

Section 2 is devoted to a brief overview of Runge–Kutta methods. Although many such reviews are available [11, 12], we include it for completeness and also because it provides us with an excellent opportunity to introduce concepts, terminology and notation that will be useful in the sequel.

In Section 3 we describe the numerical theory of *unitary flows*

$$Y' = F(Y)Y, \qquad t \geq 0, \quad Y(0) = Y_0,$$

where Y_0 is a $d \times d$ orthogonal matrix and F is a skew-symmetric matrix function. In that case it is easy to prove that for every $t \geq 0$ the matrix $Y(t)$ is orthogonal. We review two alternative approaches, one structural and the other projective, to recover orthogonality in the course of computation.

The theme of Section 4 is one of the best known developments in contemporary numerical analysis of ODEs, namely the computation of *Hamiltonian systems*

$$\begin{aligned} \boldsymbol{p}' &= -\frac{\partial H(\boldsymbol{p},\boldsymbol{q})}{\partial \boldsymbol{q}}, \\ \boldsymbol{q}' &= \frac{\partial H(\boldsymbol{p},\boldsymbol{q})}{\partial \boldsymbol{p}}. \end{aligned} \tag{1.4}$$

Although Hamiltonian systems possess many invariants, conservation of the symplectic form $\mathrm{d}\boldsymbol{p} \wedge \mathrm{d}\boldsymbol{q}$ is arguably the most important and accounts for many qualitative features of (1.4). We describe recent advances in symplectic solvers, that have already spawned a large volume of research and literally hundreds of papers.

Finally, Section 5 focuses on *isospectral* flows

$$L' = [B(L), L], \qquad t \geq 0, \quad L(0) = L_0,$$

where L_0 is a $d \times d$ symmetric matrix, the matrix function B is skew-symmetric and $[\,\cdot\,,\,\cdot\,]$ is the standard *Lie bracket.* It has been proved by Flaschka that the eigenvalues are an invariant of the flow [27]. In other words, for all $t \geq 0$ the spectrum of $L(t)$ is the same as the spectrum of L_0. Since many such systems originate in problems of linear algebra (specifically, eigenvalue and inverse eigenvalue problems) or in spectral problems in quantum chemistry, it is essential to recover the correct spectrum in the course of discretization. The authors, working in collaboration with Mari Paz Calvo, have recently introduced numerical methods that retain isospectrality and we describe briefly some of our results.

2 Overview of Runge–Kutta methods

Let $t_0 < t_1 < t_2 < \cdots$ be a given (finite or infinite) mesh and set $h_n = t_{n+1} - t_n$, $n \in \mathbb{Z}^+$. We denote by $\boldsymbol{y}_n$ a numerical approximant to the exact solution $\boldsymbol{y}(t_n)$ of the ordinary differential equation (1.1) and assume that is is produced by the one-step method

$$\boldsymbol{y}_{n+1} = \Phi(\boldsymbol{y}_n, h_n), \qquad n \in \mathbb{Z}^+. \tag{2.1}$$

A general ν-stage Runge–Kutta method advances from t_n to t_{n+1} by solving the set of $\nu \times d$ algebraic equations,

$$\begin{aligned} \boldsymbol{\psi}_\ell &= \boldsymbol{y}_n + h_n \sum_{j=1}^{\nu} a_{\ell,j} \boldsymbol{k}_j, \\ \boldsymbol{k}_\ell &= \boldsymbol{f}(t_n + c_\ell h_n, \boldsymbol{\psi}_\ell), \qquad \ell = 1, 2, \ldots, \nu, \end{aligned} \tag{2.2}$$

subsequently letting

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h_n \sum_{\ell=1}^{\nu} b_\ell \boldsymbol{k}_\ell. \tag{2.3}$$

Such a method is uniquely characterized by the *RK matrix*

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,\nu} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,\nu} \\ \vdots & \vdots & & \vdots \\ a_{\nu,1} & a_{\nu,2} & \cdots & a_{\nu,\nu} \end{bmatrix}$$

and by the two vectors

$$\boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_\nu \end{bmatrix}, \qquad \boldsymbol{c} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_\nu \end{bmatrix},$$

the *RK weights* and the *RK nodes* respectively.

Provided that no higher-order derivatives of $\boldsymbol{y}$ are employed in the numerical method, Runge–Kutta schemes (2.2–3) present a useful form for a general one-step algorithm Φ: general enough to include all methods used in practice, yet with sufficient structure to enable meaningful analysis.

If the matrix A is strictly lower triangular then the system (2.2) can be solved explicitly, otherwise a numerical procedure, e.g. Newton's method, is required in its solution. We do not delve further into this issue, except for remarking that the results reported in this paper often require (2.2) to be solved exactly or, at any rate, to machine accuracy.

The most basic attribute of a numerical method is the (local) *order* of accuracy. Suppose that $\boldsymbol{f} \in C^{p+1}[\mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d]$, where $p \in \mathbb{N}$. We say that (2.1) is of order p if $\boldsymbol{y}_n = \boldsymbol{y}(t_n)$ implies that $\boldsymbol{y}_{n+1} = \boldsymbol{y}(t_{n+1}) + \mathcal{O}(h_n^{p+1})$ for $h_n \to 0$.

Order analysis of Runge–Kutta methods is always the first step and the foremost criterion in choosing a good triple $\{A, \boldsymbol{b}, \boldsymbol{c}\}$. Yet, it is surprisingly intricate. The main difficulty rests in the expansion of composite functions. This becomes apparent as soon as we attempt to derive higher-order derivatives of (1.1) by repeated differentiation: thus

$$\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y}), \qquad \boldsymbol{y}'' = \frac{\partial \boldsymbol{f}(t, \boldsymbol{y})}{\partial t} + \frac{\partial \boldsymbol{f}(t, \boldsymbol{y})}{\partial \boldsymbol{y}} \boldsymbol{f}(t, \boldsymbol{y}),$$

and complexity grows fast. Although a general expansion of $\boldsymbol{y}(t_{n+1})$ about the point $(t_n, \boldsymbol{y}(t_n))$ *via* (1.1) is available in terms of the Faà di Bruno formula [11], expanding

$\boldsymbol{y}_{n+1}$ about $(t_n, \boldsymbol{y}(t_n))$ is more of a challenge. The first unified treatment of order conditions for Runge–Kutta methods has been presented by John Butcher, who used symbolic and combinatorial calculus. This has been subsequently simplified by John Butcher and Gerhard Wanner, who have used graph theory and identified order conditions with enumerations on rooted trees [11]. Recently, Hans Munthe-Kaas elucidated an alternative framework for order analysis by using the formalism of Lie series [19]. We mention in passing that the latter approach can be formulated in a natural way for ODEs defined in a Lie group, rather than the real line, and this possesses clear advantages insofar as differential equations 'living' on manifolds are concerned.

An alternative approach to order analysis exploits the connection between Runge–Kutta schemes and numerical quadrature and formulates order conditions in terms of orthogonality [22]. It applies only to a subset of all possible methods (2.2–3) but possesses a great virtue of neatness and simplicity. In its most straightforward, yet useful, manifestation, we choose arbitrary distinct RK nodes $c_1, c_2, \ldots, c_\nu \in [0,1]$ and let

$$\begin{aligned} a_{\ell,j} &= \int_0^{c_\ell} \frac{\omega_j(\tau)}{\omega_j(c_j)}\,\mathrm{d}\tau, \qquad \ell, j = 1, 2, \ldots, \nu, \\ b_\ell &= \int_0^1 \frac{\omega_\ell(\tau)}{\omega_\ell(c_\ell)}\,\mathrm{d}\tau, \qquad \ell = 1, 2, \ldots, \nu, \end{aligned}$$

where

$$\omega(t) = \prod_{\ell=1}^{\nu}(t - c_\ell) \qquad \text{and} \qquad \omega_\ell(t) = \frac{\omega(t)}{t - c_\ell}, \quad \ell = 1, 2, \ldots, \nu.$$

It is possible to prove that the order of the method characterized by those coefficients $\{A, \boldsymbol{b}, \boldsymbol{c}\}$ is the largest integer p such that

$$\int_0^1 \tau^{j-1}\omega(\tau)\,\mathrm{d}\tau = 0, \qquad j = 1, 2, \ldots, p - \nu. \tag{2.4}$$

In particular, choosing the RK nodes as zeros of the Legendre polynomial P_ν, linearly translated to $[0,1]$, results in a method of order $p = 2\nu$, the Gauß–Legendre Runge–Kutta method (GLRK). Thus,

$$\nu = 1: \qquad A = \left[\ \tfrac{1}{2}\ \right], \qquad \boldsymbol{b} = \left[\ 1\ \right], \qquad \boldsymbol{c} = \left[\ \tfrac{1}{2}\ \right],$$

$$\nu = 2: \qquad A = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \end{bmatrix}, \qquad \boldsymbol{b} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}, \qquad \boldsymbol{c} = \begin{bmatrix} \frac{1}{2} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} \end{bmatrix}$$

and so on. The GLRK schemes possess the highest attainable order per number of stages and they display many other desirable properties. We will encounter them time and again in this paper.

Asymptotic behaviour of Runge–Kutta methods has attracted much attention in the last two decades. The simplest asymptotic model is linear, namely $\boldsymbol{y}' = B\boldsymbol{y}$, where all the eigenvalues of the $d \times d$ matrix B reside in the left complex half-plane. In that case every exact solution tends to $\mathbf{0}$ as $t \to +\infty$ and it is easy to prove that the method (2.2–3) inherits this property if and only if $|r(z)| < 1$ for all $z \in \mathbb{C}$, $\mathrm{Re}\, z < 0$, where

$$r(z) = 1 + z\boldsymbol{b}^T(I - zA)^{-1}\mathbf{1}, \qquad z \in \mathbb{C}.$$

Such a method is said to be *A-stable.*

There exists an extensive and mathematically appealing theory of one-step A-stable methods. In particular, it is possible to prove that all GLRK methods are A-stable [15].

We have already mentioned the simplest nonlinear stability model,

$$\langle \boldsymbol{f}(t,\boldsymbol{x}) - \boldsymbol{f}(t,\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle \le 0, \qquad t \ge t_0, \quad \boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^d.$$

Such equations are contractive and we say that a Runge–Kutta method (2.2–3) is *algebraically stable* if $\|\boldsymbol{x}_n - \boldsymbol{y}_n\|$ is a monotonically decreasing function of $n \in \mathbb{Z}^+$ for every two solution sequences $\{\boldsymbol{x}_n\}$ and $\{\boldsymbol{y}_n\}$. John Butcher proved in [3] that a Runge–Kutta method is algebraically stable if $\boldsymbol{b} \ge \boldsymbol{0}$ and the matrix $M = (m_{\ell,j})$, where

$$m_{\ell,j} = b_\ell a_{\ell,j} + b_j a_{j,\ell} - b_\ell b_j, \qquad \ell, j = 1,2,\ldots,\nu, \tag{2.5}$$

is positive semi-definite. This feature is shared by all GLRK methods, since then $\boldsymbol{b} > \boldsymbol{0}$ and $M = O$.

There is more to algebraic stability than meets the eye. In particular, it is relevant for the considerably more general nonlinear model (1.2), when the underlying ODE possesses a bounded global attractor. If a method is algebraically stable then it shares this feature [26].

An asymptotic attribute of a different kind is the retention of fixed points of an autonomous ODE

$$\boldsymbol{y}' = \boldsymbol{f}(\boldsymbol{y}), \quad t \ge t_0, \qquad \boldsymbol{y}(t_0) = \boldsymbol{y}_0,$$

where $\boldsymbol{f} \in C[\mathbb{R}^d \to \mathbb{R}^d]$, under discretization. Naturally, if a numerical trajectory tends to a bounded limit $\hat{\boldsymbol{y}}$, say, as $n \to \infty$, we want the assurance that $\hat{\boldsymbol{y}}$ is a genuine fixed point of the underlying flow, i.e. that $\boldsymbol{f}(\hat{\boldsymbol{y}}) = \boldsymbol{0}$. In general, Runge–Kutta methods perform exceedingly badly in that respect and they may possess spurious fixed points. There are few exceptions, that have been identified in [10] but, in general, if avoidance of spurious fixed points is at issue then multistep methods are likely to perform significantly better [16].

Multistep methods might be superior in the case of hyperbolic fixed points but Runge–Kutta have a definite edge insofar as elliptic fixed points are concerned. In particular, it is possible to prove that a Runge–Kutta method undergoes a Hopf bifurcation if and only if the original ODE displays this feature, subject to the condition $M = O$ [16]. Note that the latter is true with regard to GLRK.

We conclude this brief guide to Runge–Kutta methods with two remarks. Firstly, we have completely disregarded the 'usual' explicit Runge–Kutta methods, familiar from all undergraduate textbooks, since, by and large, they are of limited utility within the context of this paper. Secondly, we happily acknowledge the existence of many methods in the numerical ODE menagerie which do not conform with (2.2–3), e.g. multistep methods, multiderivative methods and extrapolation. It is our contention, however, that Runge–Kutta methods are the most promising when it comes to the recovery of qualitative features. This can be illustrated with greatest ease in the case of multistep methods. For simplicity we consider the fixed-step case $t_n = t_0 + nh$,

$h_n = h > 0$, $n \in \mathbb{Z}^+$. A multistep method is of the form

$$\sum_{\ell=0}^{m} \rho_\ell \boldsymbol{y}_{n+\ell} = h \sum_{\ell=0}^{m} \sigma_\ell \boldsymbol{f}(t_{n+\ell}, \boldsymbol{y}_{n+\ell}), \qquad n \in \mathbb{Z}^+. \tag{2.6}$$

Let

$$\rho(w) = \sum_{\ell=0}^{m} \rho_\ell w^\ell, \qquad \sigma(w) = \sum_{\ell=0}^{m} \sigma_\ell w^\ell, \qquad w \in \mathbb{C}.$$

By a classical result of Dahlquist, the scheme (2.6) is convergent (i.e., the numerical solution tends uniformly to the exact solution of (1.1) within any compact time interval) if and only if $\rho(1) = 0$, $\rho'(1) = \sigma(1)$, all the zeros of the polynomial ρ reside in the closed complex unit disc and the zeros on the unit circle are simple. Note that convergence implies that ρ always has at least one zero on $|w| = 1$. It is an accepted wisdom that a good multistep scheme has exactly one zero there, since further such zeros result in marginal convergence and very slow decay of perturbations and errors [18]. On the other hand, the equivalent condition to $M = O$ in a multistep setting is that *all* the zeros of ρ reside on the unit circle [25]. This is hardly surprising, since 'zeros in $|w| < 1$' correspond to contractivity, while 'zeros on $|w| = 1$' yield conservation (consider a linear equation!), but the conclusion is that multistep methods with $m \geq 2$ are of little use in the present setting.

3 Unitary flows

The ODE

$$Y' = G(Y), \quad t \geq 0, \qquad Y(0) = I, \tag{3.1}$$

where Y is a $d \times d$ matrix and G is a Lipschitz matrix function, is said to be a *unitary flow* if $Y(t)$ is an orthogonal matrix for every $t \geq 0$.

Theorem 1 *The ODE* (3.1) *is a unitary flow if and only if* $G(Y) = F(Y)Y$, *where the* $d \times d$ *matrix function* F *maps orthogonal matrices to skew-symmetric matrices.*

Proof Let us assume first that $G(Y) = F(Y)Y$ with a skew-symmetric F. Then

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} Y^T(t)Y(t) &= \left(\frac{\mathrm{d}}{\mathrm{d}t} Y^T\right) Y + Y^T \left(\frac{\mathrm{d}}{\mathrm{d}t} Y\right) = Y^T F(Y) Y + (F(Y)Y)^T Y \\ &= Y^T (F(Y) + F^T(Y)) Y = O, \qquad t \geq 0. \end{aligned}$$

Therefore

$$Y^T(t)Y(t) = Y^T(0)Y(0) = I, \qquad t \geq 0,$$

and the flow is unitary.

The proof in the other direction is just as easy. Thus, let us assume that $Y(t)$ is orthogonal for all $t \geq 0$, therefore $Y^T Y \equiv I$. We deduce that

$$Y^T G(Y) + G^T(Y) Y = \frac{\mathrm{d}}{\mathrm{d}t} Y^T Y \equiv O.$$

Let $F(Y)=G(Y)Y^T$. Therefore, orthogonality of Y implies that $G(Y)=F(Y)Y$ and we deduce that

$$O \equiv Y^T F(Y) Y + Y^T F^T(Y) Y = Y^T (F(Y)+F^T(Y))Y.$$

Multiplying this identity by Y and Y^T from the left and the right, respectively, concludes the proof that $F(Y)+F^T(Y)=O$, hence that F maps orthogonal to skew-symmetric matrices. □

Numerical solution of unitary flows has been described in some detail by Luca Dieci, Bob Russell and Eric van Vleck in [8]. In particular, they have proved that $M=O$ and $\boldsymbol{b} \geq \boldsymbol{0}$ imply that the Runge–Kutta method (2.2–3) is unitary, i.e. that $Y_n^T Y_n = I$ for all $n \in \mathbb{Z}^+$. Herewith we prove a theorem that ensures unitarity of such a method subject to somewhat weaker conditions. The proof of the theorem has the added virtue of illustrating how the somewhat mysterious matrix M is so ubiquitous in the analysis of Runge–Kutta methods.

Theorem 2 *If* $M=O$ *then the Runge–Kutta method* (2.2–3) *is unitary.*

Proof Application of the method to the flow

$$Y' = F(Y)Y, \quad t \geq 0,$$

results in

$$\begin{aligned}
\Psi_\ell &= Y_n + h_n \sum_{j=1}^{\nu} a_{\ell,j} K_j, \\
K_\ell &= F(\Psi_\ell)\Psi_\ell, \qquad\qquad \ell = 1,2,\ldots,\nu, \\
Y_{n+1} &= Y_n + h_n \sum_{\ell=1}^{\nu} b_\ell K_\ell.
\end{aligned}$$

Therefore

$$\begin{aligned}
Y_{n+1}^T Y_{n+1} &= \left(Y_n + h_n \sum_{\ell=1}^{\nu} b_\ell K_\ell\right)^T \left(Y_n + h_n \sum_{\ell=1}^{\nu} b_\ell K_\ell\right) \\
&= Y_n^T Y_n + h_n \left(Y_n^T \sum_{\ell=1}^{\nu} b_\ell K_\ell + \sum_{\ell=1}^{\nu} b_\ell K_\ell^T Y_n\right) + h_n^2 \sum_{\ell=1}^{\nu}\sum_{j=1}^{\nu} b_\ell b_j K_\ell^T K_j.
\end{aligned}$$

Assuming by induction that Y_n is orthogonal, hence $Y_n^T Y_n = I$, it is thus enough to prove that

$$\sum_{\ell=1}^{\nu} b_\ell (Y_n^T K_\ell + K_\ell^T Y_n) + h_n \sum_{\ell=1}^{\nu}\sum_{j=1}^{\nu} b_\ell b_j K_\ell^T K_j = O. \tag{3.2}$$

But

$$Y_n = \Psi_\ell - h_n \sum_{j=1}^{\nu} a_{\ell,j} K_j, \qquad \ell = 1,2,\ldots,\nu,$$

and substitution yields

$$\begin{aligned}\sum_{\ell=1}^{\nu} b_\ell Y_n^T K_\ell &= \sum_{\ell=1}^{\nu} b_\ell \left(\Psi_\ell - h_n \sum_{j=1}^{\nu} a_{\ell,j} K_j\right)^T K_\ell \\ &= \sum_{\ell=1}^{\nu} b_\ell \Psi_\ell^T K_\ell - h_n \sum_{\ell=1}^{\nu}\sum_{j=1}^{\nu} b_\ell a_{\ell,j} K_j^T K_\ell.\end{aligned}$$

Likewise, exchanging the summation index,

$$\sum_{\ell=1}^{\nu} b_\ell K_\ell^T Y_n = \sum_{\ell=1}^{\nu} b_\ell K_\ell^T \Psi_\ell - h_n \sum_{\ell=1}^{\nu}\sum_{j=1}^{\nu} b_j a_{j,\ell} K_j^T K_\ell.$$

Since $F(Y)$ is skew-symmetric and $K_\ell = F(\Psi_\ell)$, $\ell = 1, 2, \ldots, \nu$, we have

$$\Psi_\ell^T K_\ell + K_\ell^T \Psi_\ell = \Psi_\ell^T (F(\Psi_\ell) + F^T(\Psi_\ell))\Psi_\ell = O, \qquad \ell = 1, 2, \ldots, \nu,$$

therefore

$$\sum_{\ell=1}^{\nu} b_\ell (Y_n^T K_\ell + K_\ell^T Y_n) = -\sum_{\ell=1}^{\nu}\sum_{j=1}^{\nu} (b_\ell a_{\ell,j} + b_j a_{j,\ell}) K_\ell^T K_j.$$

Substitution in (3.2) and $M = O$ demonstrate $Y_{n+1}^T Y_{n+1} = I$, thereby concluding an inductive proof of unitarity. □

Corollary *[8] All GLRK methods are unitary.* □

GLRK methods are an example of a structural approach: unitarity is built into the method. An alternative is a projective technique. Dieci, Russell and van Vleck recommend using an arbitrary p-order numerical method to produce a candidate solution $\tilde{Y}_{n+1}$, say, QR-factorize it $\tilde{Y}_{n+1} = QR$ and choose $Y_{n+1} = Q$ [8]. In this case $R = I + \mathcal{O}(h_n^{p+1})$, hence this approach does not disturb local accuracy.

However, the operation of projection is norm-dependent. In the case of the manifold $O(d)$ of $d \times d$ orthogonal matrices it is natural to consider the Frobenius norm (i.e., the Euclidean norm on $\mathbb{R}^{d^2}$). With respect to the Frobenius norm the QR factorization is not a projection onto $O(d)$. Such a projection is attainable if, instead of QR factorization, we employ the polar decomposition $\tilde{Y}_{n+1} = QP$, where Q is orthogonal and P symmetric and positive semi-definite [14]. Having said this, we wish not to hide our bias in favour of structural methods.

4 Symplectic flows

An ODE is said to be Hamiltonian if it is of the form

$$\begin{aligned}\boldsymbol{p}' &= -\frac{\partial H(\boldsymbol{p},\boldsymbol{q})}{\partial \boldsymbol{q}}, \\ \boldsymbol{q}' &= \frac{\partial H(\boldsymbol{p},\boldsymbol{q})}{\partial \boldsymbol{p}},\end{aligned} \tag{4.1}$$

where $H(\,\cdot\,,\,\cdot\,):\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ is a sufficiently smooth *Hamiltonian function.* Hamiltonian functions are ubiquitous in mathematical physics. As Sir Roger Penrose writes, "In fact, for suitable choices of H, Hamilton's equations still hold true for *any* system of classical equations whatever... ...Hamilton's equations also hold true for special relativity. Even general relativity can, if due care is exercised, be subsumed into the Hamiltonian framework." [23].

Hamiltonian ODEs possess a large number of invariants and it is this feature that makes them such an important tool in so many applications. For example, it is easy to prove that

$$\frac{\mathrm{d}}{\mathrm{d}t}H(\boldsymbol{p}(t),\boldsymbol{q}(t))\equiv 0,\qquad t\ge 0$$

and the total mechanical energy H is conserved. The proof is straightforward and it follows by rewriting (4.1) in the form

$$\boldsymbol{y}'=J\nabla H,$$

where

$$\boldsymbol{y}=\begin{bmatrix}\boldsymbol{p}\\ \boldsymbol{q}\end{bmatrix},\qquad J=\begin{bmatrix}O & -I\\ I & O\end{bmatrix}.$$

Since J is skew-symmetric,

$$\frac{\mathrm{d}}{\mathrm{d}t}H(\boldsymbol{p}(t),\boldsymbol{q}(t))=(\nabla H)^T\frac{\mathrm{d}\boldsymbol{y}(t)}{\mathrm{d}t}=(\nabla H)^T J\nabla H\equiv 0,\qquad t\ge 0,$$

hence H is conserved.

More important conserved quantity of Hamiltonian ODEs – arguably, the most important – is *symplecticity.* Although it is possible to define it in terms of the external product $\mathrm{d}\boldsymbol{p}\wedge\mathrm{d}\boldsymbol{q}$, we prefer a more transparent and intuitive geometric route. Given $s>0$, the *Hamiltonian flow* $\boldsymbol{\varphi}_{H,s}$ maps the initial value to the value at $t=s$,

$$\boldsymbol{\varphi}_{H,s}\left(\begin{bmatrix}\boldsymbol{p}(0)\\ \boldsymbol{q}(0)\end{bmatrix}\right)=\begin{bmatrix}\boldsymbol{p}(s)\\ \boldsymbol{q}(s)\end{bmatrix}.$$

Let Ω be a measurable subset of $\mathbb{R}^{2d}$. The definition of the Hamiltonian flow can be readily generalized to Ω by letting

$$\boldsymbol{\varphi}_{H,s}(\Omega)=\{\boldsymbol{\varphi}_{H,s}(\boldsymbol{x})\,:\,\boldsymbol{x}\in\Omega\}.$$

Denote by $\gamma_k(\Omega)$ the projection of Ω on the two-dimensional space $[\,p_k\quad q_k\,]$, $k=1,2,\ldots,d$. The measurability of Ω ensures that the (oriented) area of each projection is well defined and this allows us to introduce the *Poincaré integral invariant* of Ω,

$$m(\Omega):=\sum_{k=1}^{d}\operatorname{area}\gamma_k(\Omega).$$

The conservation of the symplectic form is equivalent to

$$m(\Omega)=m(\boldsymbol{\varphi}_{H,s}(\Omega))$$

for every measurable $\Omega\subseteq\mathbb{R}^{2d}$ and $s>0$ [19, 25].

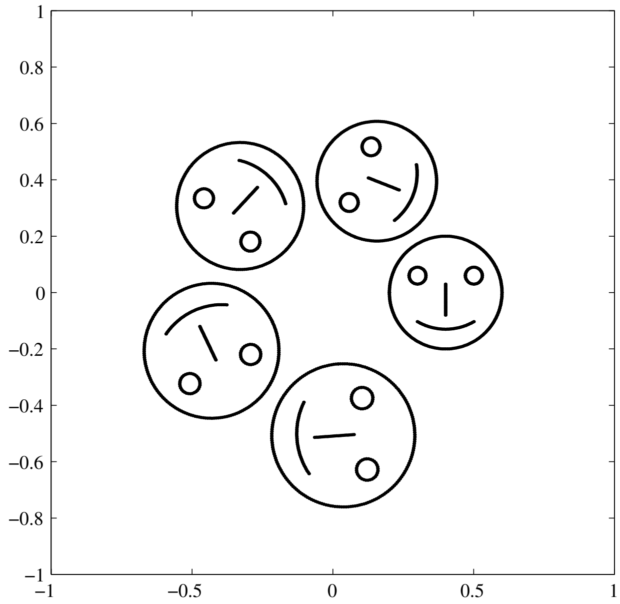

Figure 1 Rigid Smiley dynamics: The harmonic oscillator $y'' + y = 0$ (that is, $H(p,q) = \frac{1}{2}(p^2 + q^2)$), as approximated by the Euler method. The original 'Smiley' shape is mapped four times by $\varphi_{H,\frac{5}{4}}$.

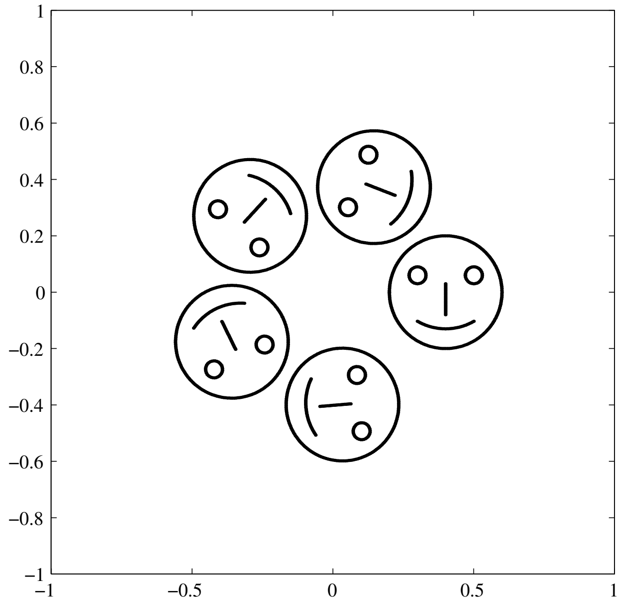

Figure 2 Discretization of the harmonic oscillator by implicit midpoint rule (cf. Section 2).

An immediate consequence of generalized area conservation is that fixed points of (4.1) can be neither sinks nor sources: they are centres. In other words, each fixed point is surrounded by periodic orbits, a situation familiar in the case of the linear harmonic oscillator $y'' + y = 0$ (possibly the simplest Hamiltonian equation of interest). Centres, however, are perhaps the most volatile dynamical objects insofar as discretization is concerned. Every small perturbation is liable to destroy this feature.

An adverse effect of discretization is vividly illustrated in Figure 1, where we display the effect of discretization on the linear harmonic oscillator $y'' + y = 0$, which corresponds to the Hamiltonian $H(p, q) = \frac{1}{2}(p^2 + q^2)$. The equation is solved by the familiar Euler method

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h_n \boldsymbol{f}(t_n, \boldsymbol{y}_n), \qquad n \in \mathbb{Z}^+,$$

with $h_n \equiv 1/10$ and we display the evolution of a Smiley-shaped (measurable) set at times $t = 0, \frac{5}{4}, \frac{5}{2}, \frac{15}{4}, 5$. Clearly, the area (in two dimensions the symplectic form coincides with area in the phase plane) is not conserved, Smiley grows fatter and the centre at the origin is replaced by a source.

Our faith in computation is restored upon an examination of Figure 2, which displays the evolution of the same system when we use the implicit midpoint rule. This method has already been described in Section 2 as GLRK with $\nu = 1$ and we can write it in a more compact form as

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h \boldsymbol{f}(t_n + \tfrac{1}{2}h_n, \tfrac{1}{2}(\boldsymbol{y}_n + \boldsymbol{y}_{n+1})), \qquad n \in \mathbb{Z}^+. \tag{4.2}$$

With this scheme, Smiley retains its area (in fact, its shape) and the centre at origin seems to have survived the discretization. As we will see soon, this is not an accident.

Further evidence that the Euler method and the implicit midpoint rule differ in their treatment of symplecticity is furnished by our second example, the equations

$$p' = -\tfrac{3}{2}q^2, \qquad q' = p,$$

which originate in the Hamiltonian $H(p, q) = \frac{1}{2}(p^2 + q^3)$. As we show in Figure 3, this time Smiley is stretched but, insofar as the implicit midpoint rule is concerned, the area is not amended and the discretization does not corrupt the centre. This is not the case with Euler's method at the top figure: as before, the area grows and the origin is a source.

We say that a Runge–Kutta method is *symplectic* (or *canonical*) if, when applied with a constant step-size sequence, it preserves the symplectic structure of each and every Hamiltonian equation (4.1). Remarkably, such methods can be characterized by techniques similar to the proof of Theorem 2.

Theorem 3 (Lasagni, Suris, Sanz-Serna) *A Runge–Kutta method* (2.2–3) *is symplectic if $M = O$. [25]* □

Corollary *All GLRK methods are symplectic.* □

Incidentally, unless the Runge–Kutta method has redundant stages, the condition $M = O$ is also necessary for symplecticity.

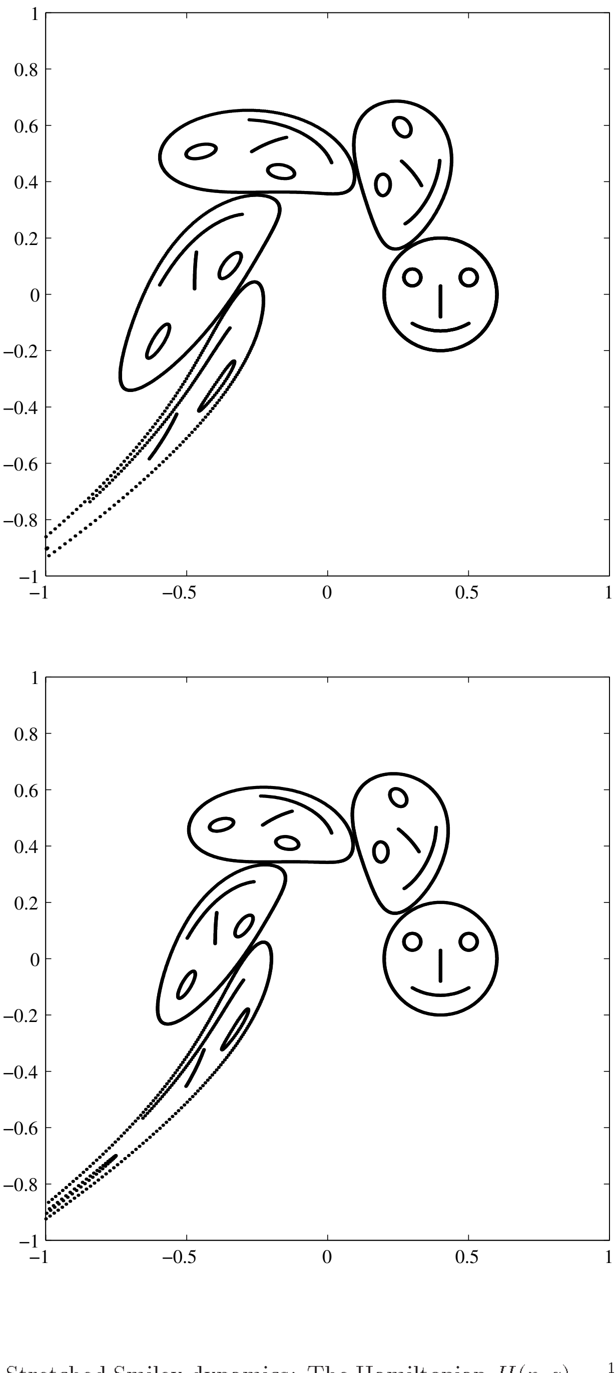

Figure 3 Stretched Smiley dynamics: The Hamiltonian $H(p,q) = \frac{1}{2}(p^2+q^3)$ under discretization by Euler's method (top) and the implicit midpoint rule (bottom), each with step $h_n \equiv \frac{1}{10}$.

The condition $M = O$ requires the method to be implicit, and this is a severe restriction in many applications. Equations of molecular dynamics, say, may involve an exceedingly large number of degrees of freedom and the solution need be computed for a long time. Calculating a large system of nonlinear algebraic equations in each step results in a prohibitive cost. The situation is somewhat easier for *separable* Hamiltonians,

$$H(\boldsymbol{p}, \boldsymbol{q}) = T(\boldsymbol{p}) + Q(\boldsymbol{q}), \qquad \boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^d.$$

In that case it is possible to use *separable* Runge–Kutta and (for $T(\boldsymbol{p}) = \frac{1}{2}\boldsymbol{p}^T\boldsymbol{p}$) Runge–Kutta–Nyström methods which may be both symplectic and explicit. We wish not to discuss further this issue, which has been admirably described in [24, 25].

The restriction of constant step-size is fundamental to our argument. As soon as variable step-size is allowed, benefits of symplecticity disappear. The reason is that, as long as $h_n \equiv h$, the discretization recovers exactly the solution of a nearby Hamiltonian system but this property is generically lost as soon as steps are allowed to vary. There is a double benefit of staying on a nearby Hamiltonian orbit. Firstly, the solution is likely to recover the dynamics of the underlying system with high degree of faithfulness. Since Hamiltonian dynamics is often exceedingly complicated [19], this is an important feature. Secondly, it is possible to prove that the error accumulates at a linear speed [5]. This is a most welcome attribute of the method, at variance with the behaviour of variable-step ODE solvers, where even the most rigourous state-of-the-art (local) error control cannot prevent quadratic error accumulation [25].

An alternative to symplecticity is conservation of energy. We should perhaps emphasize that the options are exclusive since, according to a theorem of Zhong and Marsden [29], the only symplectic 'method' that conserves energy is, essentially, the exact solution. Whereas a symplectic method is, by its very nature, structural, an energy-conserving method can be projective. Thus, we can solve (4.1) by an arbitrary algorithm and project the solution on the manifold

$$\{[\ \boldsymbol{p} \quad \boldsymbol{q}\ ] : H(\boldsymbol{p}, \boldsymbol{q}) = H(\boldsymbol{p}(0), \boldsymbol{q}(0))\}$$

by methods of optimization theory.

Conservation of energy, like symplecticity, implies linear error accumulation [9]. The choice between these alternative approaches is a matter of current discussion. Clearly, energy conservation can be cheaper to attain in practice. Equally clearly, there is much more to symplecticity than just slow error accumulation.

5 Isospectral flows

Like a unitary flow, an *isospectral flow* is defined over $d \times d$ matrices. Specifically, we say that

$$L' = H(L), \quad t \geq 0, \qquad L(0) = L_0, \tag{5.1}$$

where L_0 is an arbitrary symmetric matrix and H maps symmetric $d \times d$ matrices to themselves, is isospectral if

$$\sigma(L(t)) \equiv \sigma(L_0), \qquad t \geq 0.$$

Originally, such flows occurred in the analysis of certain Toda lattice equations and their isospectrality was proved by Hermann Flaschka [27].

Theorem 4 *The flow* (5.1) *is isospectral if and only if there exists a Lipschitz function* B *that maps symmetric matrices into skew-symmetric matrices, such that*

$$L' = H(L) = [B(L), L] = B(L)L - LB(L). \tag{5.2}$$

Proof Let us first assume that H is in the form (5.2) and consider the equation

$$U' = B(L(t))U, \quad t \geq 0, \qquad U(0) = I. \tag{5.3}$$

Since $B(L(t))$ is skew-symmetric for all $t \geq 0$, according to Theorem 1 the solution of (5.3) is a unitary flow.

Since

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} U^T L U &= \left(\frac{\mathrm{d}}{\mathrm{d}t} U\right)^T LU + U^T \left(\frac{\mathrm{d}}{\mathrm{d}t} L\right) U + U^T L \left(\frac{\mathrm{d}}{\mathrm{d}t} U\right) \\ &= -U^T B(L) LU + U^T \Big(B(L)L - LB(L)\Big) U + U^T LB(L)U = O, \end{aligned}$$

we deduce that

$$U^T(t)L(t)U(t) = U^T(0)L(0)U(0) = L_0.$$

Consequently, $L(t) = U(t)L_0U^T(t)$, hence isospectrality.

To prove in the opposite direction we note that the identity $\sigma(L(t)) \equiv \sigma(L_0)$ is the same as the assertion that there exists a differentiable matrix function $V(t)$, $t \geq 0$, such that $\det V(t) \neq 0$ and

$$L(t) = V(t)L_0V^{-1}(t), \qquad t \geq 0.$$

Differentiation yields

$$L' = V'L_0V^{-1} + VL_0(-V^{-1}V'V^{-1}) = V'V^{-1}L - LV'V^{-1} = B(L)L - LB(L),$$

where $B(L) = V'V^{-1}$, thereby deriving the asserted form of H. □

As we have already mentioned, an important example of isospectral flows is *Toda lattice equations.* In general, the equations of motion of a one-dimensional lattice are

$$y_n'' = \phi'(y_{n+1} - y_n) - \phi'(y_n - y_{n-1}), \qquad n = 1, 2, \ldots, d,$$

where y_n is the displacement of the nth particle (of unit mass) and ϕ is the interaction potential. The Toda lattice equations are obtained when the interaction potential is related to the exponential function: letting $\phi(r) = \mathrm{e}^{-r} + r$, $r \in \mathbb{R}$, yields the Hamiltonian system

$$\begin{aligned} p_n' &= \mathrm{e}^{-(q_n - q_{n-1})} - \mathrm{e}^{-(q_{n+1} - q_n)}, \\ q_n &= p_n. \end{aligned}$$

Moreover, the change of variables

$$\alpha_n = \tfrac{1}{2}\mathrm{e}^{-\frac{1}{2}(q_{n+1}-q_n)}, \qquad \beta_n = \tfrac{1}{2}p_n,$$

results in the equations

$$\begin{aligned} \alpha_n' &= \alpha_n(\beta_n - \beta_{n+1}), \qquad n = 1, 2, \ldots, d-1, \\ \beta_n' &= 2(\alpha_{n-1}^2 - \alpha_n^2), \qquad n = 1, 2, \ldots, d, \end{aligned}$$

where $\alpha_0, \alpha_d \equiv 0$. Finally, letting

$$L = \begin{bmatrix} \beta_1 & \alpha_1 & 0 & \cdots & 0 \\ \alpha_1 & \beta_2 & \alpha_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \alpha_{d-2} & \beta_{d-1} & \alpha_{d-1} \\ 0 & \cdots & 0 & \alpha_{d-1} & \beta_d \end{bmatrix},$$

it is trivial to verify that (5.2) holds with the skew-symmetric matrix

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

Given a $d \times d$ symmetric matrix K, we denote by K_- its strictly lower-triangular part. In this notation $B(L) = L_- - L_-^T$. With greater generality, let $f : \mathbb{C} \to \mathbb{C}$ be an arbitrary function, analytic on the spectrum of L_0, and set $B(L) = f(L)_- - f(L)_-^T$. This is termed a *QR flow.* Obviously, $f(z) = z$ corresponds simply to Toda lattice equations. A more interesting observation is that the function $f(z) = \ln z$ recovers, at integer time steps, the familiar QR method of numerical linear algebra when applied to compute the eigenvalues of the (positive definite) matrix L_0. It is possible to prove that for every such f the function $L(t)$ converges to a diagonal matrix [28].

Another isospectral flow originating in modelling particles in a lattice is the *Kac–von Moerbeke flow*, where

$$\begin{aligned} L_{k,j} &= \begin{cases} a_k & : |k-j| = 1, \\ 0 & : \text{otherwise}, \end{cases} \\ B_{k,j}(L) &= \begin{cases} a_k a_{k+1} & : k = j-2, \\ -a_{k-2}a_{k-1} & : k = j+2, \\ 0 & : \text{otherwise}, \end{cases} \qquad k, j = 1, 2, \ldots, d. \end{aligned}$$

Our final example, the *double bracket flow,* has been introduced by Roger Brockett [1] and has a significant number of interesting applications. Let N be a $d \times d$ symmetric matrix. The double bracket flow defines $B(L)$ in terms of the Lie bracket, $B(L) = [N, L]$. It is possible to generalize the double bracket flow in several ways and this leads to further intriguing applications, e.g. inverse eigenvalue problems [6] and the determination of the active set in linear programming [17].

A standard order-p ODE method is likely to produce an error of similar order in approximating the eigenvalues. Sometimes this is perfectly satisfactory. However, if the purpose of the computation is to solve an eigenvalue problem or an inverse eigenvalue problem, say, then it is essential that all d eigenvalues of L are recovered exactly.

Let us denote $\sigma(L(t)) = \{\lambda_1(t), \lambda_2(t), \ldots, \lambda_d(t)\}$, $t \geq 0$, and ponder for a moment how to formulate relations for the invariants $\lambda_k(t) \equiv \lambda_k(0)$, $t \geq 0$, $k = 1, 2, \ldots, d$, in a form suitable for analysis. In general, the eigenvalues are unknown and we do not wish to evaluate them first. Instead, it is a good idea to express conservation of eigenvalues in terms of the elements of the matrix L. We commence by noting that conserving $\lambda_k(t)$, $k = 1, 2, \ldots, d$, is identical to conserving the quantities

$$\sum_{\ell=1}^{d} \lambda_\ell^k(t), \qquad k = 1, 2, \ldots, d.$$

Next, we note that

$$\sum_{\ell=1}^{d} \lambda_\ell^k(t) = \operatorname{tr} L^k(t), \qquad k = 1, 2, \ldots, d.$$

Finally, we observe that, for every $t \geq 0$, $\operatorname{tr} L^k(t) = \text{const}$ is a k-degree manifold $\mathcal{M}_k$ in $\mathbb{R}^{d^2}$, which can be described in terms of a k-degree covariant tensor S_k, say. Therefore, isospectrality is equivalent to

$$L(t) \in \bigcap_{k=1}^{d} \mathcal{M}_k, \qquad t \geq 0.$$

In other words, isospectrality is equivalent to $L(t)$ lying at the intersection of d polynomial manifolds for all $t \geq 0$. This justifies posing a more general question, namely how well can Runge–Kutta methods (our best hope for an isospectral algorithm) stay on a polynomial manifold?

Suppose that it is known that the exact solution of the ODE $\boldsymbol{y}' = \boldsymbol{f}(\boldsymbol{y})$, where $\boldsymbol{f} : \mathbb{R}^s \to \mathbb{R}^s$ and $s \in \mathbb{N}$ is the dimension of the system, lies for all $t \geq 0$ on the k-degree manifold

$$\mathcal{M}_k = \{\boldsymbol{x} \in \mathbb{R}^s \,:\, \sum_{|\boldsymbol{i}|=k} S_{\boldsymbol{i}} \boldsymbol{x}^{\boldsymbol{i}} = \text{const}\},$$

where $S_{\boldsymbol{\pi}(\boldsymbol{i})} = S_{\boldsymbol{i}}$ for every permutation $\boldsymbol{\pi}$. Given a Runge–Kutta method $\{A, \boldsymbol{b}, \boldsymbol{c}\}$, we query whether it is true that

$$\boldsymbol{y}_n \in \mathcal{M}_k, \qquad n \in \mathbb{Z}^+. \tag{5.4}$$

It is trivial to verify that every Runge–Kutta method of order $p \geq 1$ stays, when required, on the linear manifold $\mathcal{M}_1$.

Theorem 5 (Cooper) *If $M = O$ then a Runge–Kutta method obeys* (5.4) *with $k = 2$ [25].* □

It is possible further to prove that for every Runge–Kutta method (2.2–3) such that $M \neq O$ there exists an isospectral problem for which this method departs off the quadratic manifold $\mathcal{M}_2$ (defined by the tensor S_2) [4]. We deduce thus that $M = O$ is necessary for isospectrality. Unfortunately, there is much more to the conservation of eigenvalues than simply $M = O$. In [4] we have proved, together with Mari Paz Calvo, that for every Runge–Kutta method the departure from the cubic manifold $\mathcal{M}_3$ can be expanded in the form

$$\gamma h^2 + \delta h^3 + \mathcal{O}(h^4),$$

where h is the step-size. Moreover, $\gamma = 0$ for all isospectral problems if and only if

$$m_{i,j} = b_i a_{i,j} + b_j a_{j,i} - b_i b_j = 0, \qquad i, j = 1, 2, \ldots, \nu, \tag{5.5}$$

whereby $\delta = 0$ for all isospectral problems is equivalent to requiring

$$n_{i,j,\ell} = b_i a_{i,j} a_{i,\ell} + b_j a_{j,i} a_{j,\ell} + b_\ell a_{\ell,i} a_{\ell,j} - b_i b_j b_\ell = 0, \qquad i, j, \ell = 1, 2, \ldots, \nu. \tag{5.6}$$

Theorem 6 (Calvo, Iserles & Zanna) *No Runge–Kutta method of order $p \geq 1$ may stay on the cubic manifold $\mathcal{M}_3$ (defined by the tensor S_3) for all isospectral problems.*

Proof. We show that the conditions (5.5) and (5.6) are contradictory. Since the underlying method is of order $p \geq 1$, not all the weights $\boldsymbol{b}$ may vanish. Let us choose $i \in \{1, 2, \ldots, \nu\}$ such that $b_i \neq 0$. Then

$$\begin{aligned} m_{i,i} = 0 \quad &\Rightarrow \quad b_i a_{i,i} = \tfrac{1}{2} b_i^2 \quad \Rightarrow \quad a_{i,i} = \frac{b_i}{2}, \\ n_{i,i,i} = 0 \quad &\Rightarrow \quad b_i a_{i,i}^2 = \tfrac{1}{3} b_i^3 \quad \Rightarrow \quad a_{i,i} = \frac{b_i}{\sqrt{3}}. \end{aligned}$$

Hence contradiction and we deduce that $\gamma = \delta = 0$ is impossible. □

Corollary (Calvo, Iserles & Zanna) *No Runge–Kutta method can conserve isospectrality for all possible ODEs of the form* (5.2). □

In order to present an isospectral algorithm, we hark back to the proof of Theorem 4. Specifically, to step from t_n to t_{n+1}, we consider the two equations

$$U' = B(L(t))U, \quad t \geq t_n, \qquad U(t_n) = I, \tag{5.7}$$

$$L_{n+1} = U(t_{n+1}) L_n U^T(t_{n+1}). \tag{5.8}$$

Our first idea for an algorithm sets an inner iteration between (5.7) and (5.8). We set $L_{n+1}^{[0]} = L_n$ and, for $m = 0, 1, \ldots$, solve

$$U^{[m]\prime} = B(L^{[m]}) U^{[m]}, \quad t \geq t_n, \qquad U^{[m]}(t_n) = I,$$

with a *unitary* Runge-Kutta method (cf. Section 3) and let

$$L_{n+1}^{[m+1]} = U^{[m]}(t_{n+1}) L_n {U^{[m]}}^T (t_{n+1}).$$

The iteration is terminated as soon as convergence is attained to given tolerance or, alternatively, when convergence fails (in which case we repeat the calculation with a smaller h_n).

Unfortunately, this algorithm cannot be executed in this form. The reason is that, while a Runge–Kutta method, as applied to (5.7), requires $B(L^{[m]}(t))$ to be known at $t = t_n + c_\ell h_n$, $\ell = 1, 2, \ldots, \nu$, the matrix function $L^{[m]}$ is provided by (5.8) only at t_{n+1} (of course, it is known also at t_n). This, however, can be rectified by a modification of the usual Runge–Kutta approach.

Let us suppose that the order of the GLRK method is an even integer p and that $B(L(\cdot)) \in C^r[\mathbb{R} \to \mathbb{R}^{d^2}]$, where $r \geq p/2 - 1$. Given $L^{[m]}$ at $t = t_n, t_{n+1}$, we let P be the Hermite (matrix) polynomial interpolant to the data

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k} B(L^{[m]}(t_i)), \qquad k = 0, 1, \ldots, p/2 - 1, \quad i = n, n+1.$$

Therefore, P approximates $B(L^{[m]}(\cdot))$, $t_n \leq t \leq t_{n+1}$, to order $\mathcal{O}(h_n^p)$. Moreover, P is also a skew-symmetric matrix. In other words, applying the Runge–Kutta method to the unitary flow

$$U^{[m]\prime} = P(t) U^{[m]}, \quad t \geq t_n, \qquad U^{[m]}(t_n) = I,$$

we have enough data to derive an approximation to $U^{[m]}(t_{n+1})$, subsequently advancing the inner iteration to $L_{n+1}^{[m+1]}$, and the order of the modified method remains p.

The simplest example is provided by the *modified implicit midpoint rule,*

$$\begin{array}{rcll} L_{n+1}^{[0]} & = & L_n, & \\ L_{n+1/2}^{[m]} & = & \frac{1}{2}(L_n + L_{n+1}^{[m]}), & m = 0, 1, \ldots \\ U^{[m]}(t_{n+1}) & = & (I - \frac{1}{2} h_n L_{n+1/2}^{[m]})^{-1} (I + \frac{1}{2} h_n L_{n+1/2}^{[m]}), & \\ L_{n+1}^{[m+1]} & = & U^{[m]}(t_{n+1}) L_n U^{[m]^T}(t_{n+1}). & \end{array}$$

This iteration converges in a norm $\|\cdot\|$ and $L_{n+1} = \lim_{m\to\infty} L_{n+1}^{[m]}$ exists, provided that $\mu_B h_n < 1$, where μ_B is a Lipschitz constant of the function B, namely μ_B satisfies

$$\|B(L) - B(K)\| \leq \mu_B \|L - K\|, \qquad L, K \in \mathbb{R}^{d^2}.$$

In that case

$$\begin{array}{rcl} L_{n+1} & = & L_n + h_n [B(L_{n+1/2}), L_{n+1/2}] \\ & & \quad + \frac{1}{4} h_n^2 B(L_{n+1/2})(L_{n+1} - L_n) B(L_{n+1/2}), \end{array}$$

where $L_{n+1/2} = \frac{1}{2}(L_n + L_{n+1})$. Comparison with (4.2) affirms that the modification to the conventional implicit midpoint rule is to be found in the second line, in a term which, being $\mathcal{O}(h_n^3)$, does not interfere with the order of accuracy. It is this perturbation that renders the modified implicit midpoint rule isospectral.

References

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

[2] J.C. Butcher, "Coefficients for the study of Runge–Kutta integration processes", *J. Aust. Math. Soc.* **3** (1963), 185–201.

[3] J.C. Butcher, "A stability property of implicit Runge–Kutta methods", *BIT* **15** (1975), 358–361.

[4] M.P. Calvo, A. Iserles and A. Zanna, "Numerical solution of isospectral flows", DAMTP Tech. Rep. 1995/NA03 (1995).

[5] B. Cano and J.M. Sanz-Serna, "Error growth in the numerical integration of periodic orbits, with application to Hamiltonian and reversible systems", Tech. Rep., Universidad de Valladolid (1995).

[6] M.T. Chu and K.R. Driessel, "Can real symmetric Toeplitz matrices have arbitrary real spectra?", Tech. Rep., Idaho State Univ.

[7] G. Dahlquist, "G-stability is equivalent to A-stability", *BIT* **18** (1978), 384–401.

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

[9] D. Estep and A.M. Stuart, "The rate of error growth in Hamiltonian-conserving integrators", Tech. Rep. SCCM-94-09, Stanford University (1994).

[10] E. Hairer, A. Iserles and J.M. Sanz-Serna, "Equilibria of Runge–Kutta methods", *Numer. Math.* **58** (1990), 243–254.

[11] E. Hairer, S.P. Nørsett and G. Wanner, *Solving Ordinary Differential Equations I: Nonstiff Problems,* Springer-Verlag, Berlin (1987).

[12] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II: Stiff Problems and Differential-Algebraic Equations*, Springer-Verlag, Berlin, 1991.

[13] R.W. Hamming, *Introduction to Applied Numerical Analysis,* McGraw–Hill, New York (1971).

[14] D. Higham, "Time-stepping and preserving orthonormality", Tech. Rep. NA/161, University of Dundee (1995).

[15] A. Iserles and S.P. Nørsett, *Order Stars,* Chapman and Hall, London (1991).

[16] A. Iserles, A. Peplow and A.M. Stuart, "A unified approach to spurious solutions introduced by time discretization. Part I: basic theory", *SIAM J. Num. Anal.* **28**, 1723–1751.

[17] J. Lagarias, Personal communication (1995).

[18] J.D. Lambert, *Numerical Methods for Ordinary Differential Systems,* John Wiley & sons, Chichester (1991).

[19] R.S. MacKay, "Some aspects of the dynamics and numerics of Hamiltonian systems", in *The Dynamics of Numerics and the Numerics of Dynamics* (D.S. Broomhead & A. Iserles, eds), Oxford University Press (1992), 137–193.

[20] R. März, "Numerical methods for differential algebraic equations", *Acta Numerica* **1** (1992), 141–198.

[21] H. Munthe-Kaas, "Lie–Butcher theory for Runge–Kutta methods", Tech. Rep. 105, University of Bergen, Norway (1995).

[22] S.P. Nørsett and G. Wanner, "The real-pole sandwich for rational approximations and oscillation equations", *BIT* **19** (1979) 89–94.

[23] R. Penrose, *The Emperor's New Mind,* Oxford University Press, Oxford (1989).

[24] J.M. Sanz-Serna, "Symplectic integrators for Hamiltonian problems: an overview", *Acta Numerica* **1** (1992), 243–286.

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

[26] A.M. Stuart, "Numerical analysis of dynamical systems", *Acta Numerica* **3** (1994), 467–572.

[27] M. Toda, *Theory of Nonlinear Lattices,* Springer-Verlag, Berlin (1981).

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

[29] G. Zhong and J.E. Marsden, "Lie–Poisson Hamilton–Jacobi theory and Lie–Poisson integrators", *Phys. Lett. A* **133**, 134–139.