RUNGE–KUTTA METHODS ON MANIFOLDS

MARIPAZ CALVO
Departamento de Matemática Aplicada y Computación
Universidad de Valladolid
Valladolid, Spain
E-mail: Maripaz@mac.cie.uva.es

ARIEH ISERLES
Department of Applied Mathematics and Theoretical Physics
University of Cambridge
Cambridge, England
E-mail: A.Iserles@amtp.cam.ac.uk

and

ANTONELLA ZANNA
Newnham College
University of Cambridge
Cambridge, England
E-mail: A.Zanna@amtp.cam.ac.uk

ABSTRACT

The subject matter of this paper is the recovery of invariants and conservation laws of ordinary differential systems by numerical methods. We prove that the most likely candidates for this task, Runge–Kutta schemes, fail to stay on manifolds defined by r-tensors with $r \geq 3$. As an alternative, we suggest diffeomorphically mapping complicated manifolds to simpler ones. This procedure allows for recovery of invariants that are intractable in a classical setting and it emphasizes the crucial role of the topology of underlying manifolds.

1. Invariants and numerical methods

Let us suppose that an autonomuos system of ordinary differential equations (ODEs)

$$\boldsymbol{y}' = \boldsymbol{f}(\boldsymbol{y}), \qquad t \geq 0, \tag{1.1}$$

is given in the Euclidean space $\mathbb{R}^d$ and that it is known that for every initial condition $\boldsymbol{y}(0) = \boldsymbol{y}_0$ its exact solution stays for all $t \geq 0$ on a manifold $\mathcal{M} = \mathcal{M}(\boldsymbol{y}_0) \subset \mathbb{R}^d$. In other words,

$$\boldsymbol{y}_0 \in \mathcal{M} \qquad \Rightarrow \qquad \boldsymbol{y}(t) \in \mathcal{M} \quad \forall\, t \geq 0. \tag{1.2}$$

We say that the equation (1.1) is $\mathcal{M}$-*invariant.*

$\mathcal{M}$-invariance often conveys an important information about the underlying differential system, for example the satisfaction of a conservation law, an integral, etc. It is usually desirable and often essential to retain this feature under discretization.

Classical approach to numerical methods for ODEs pays little attention to the issue of invariance, the emphasis being on deriving accuracy within a specified error tolerance and on the recovery of correct asymptotic behaviour of the solution. This is perfectly satisfactory in many situations and we have no intention to denigrate the impressive body of contemporary work in numerical solution of ODEs. However, there exists a wide range of problems that require the exact retention of invariants, e.g. Hamiltonian systems and isospectral flows, whereby the classical approach falls short of our requirements. We refer the reader to (Iserles and Zanna, 1995), where this issue is debated at a greater length.

Recent years have witnessed a growing amount of attention devoted to numerical retention of invariance. Much of the work has been specialized to specific manifolds: the symplectic manifold for Hamiltonian equations (Sanz-Serna and Calvo, 1994), the unitary group $O(d)$ for unitary flows (Dieci, Russell and Van Vleck, 1994) and the isospectral manifold for Toda lattice equations, double-bracket equations and other isospectral flows (Calvo, Iserles and Zanna, 1995). However, outlines of a more general theory have started to emerge lately, mainly in the work of Peter Crouch and R. Grossman (1993) and Hans Munthe-Kaas (1995). The aim of the present paper is to pursue the general theme, to understand how well can existing methods retain $\mathcal{M}$-invariance for different manifolds $\mathcal{M}$ and what can be done to improve their efficacy.

Section 2 expands upon a theme to which we have already addressed ourselves in a somewhat lesser generality in (Calvo et al., 1995), namely manifolds that are defined as a level set of a r-tensor, $r \in \mathbb{N}$. The case $r = 1$ is trivial, while $r = 2$ has been already determined by Cooper (1987). We prove that no nonconfluent Runge–Kutta method can retain $\mathcal{M}$-invariance for all manifolds described by a symmetric r-tensor whenever $r \geq 3$. In particular, no Runge–Kutta method can stay on the manifold

$$\mathcal{M} = \left\{ \boldsymbol{y} \in \mathbb{R}^d \,:\, \sum_{k=1}^{d} y_k^r = \sum_{k=1}^{d} y_k^r(0) \right\}$$

for $r \geq 3$.

In Section 3 we propose a remedy which allows us to retain $\mathcal{M}$-invariance for a large set of interesting manifolds and opens up an approach which might well have a substantially wider scope. Mapping the ODE (1.1) by a diffeomorphism, we obtain a differential equation which is $\mathcal{N}$-invariant and a manifold $\mathcal{N}$ which, while being topologically equivalent to $\mathcal{M}$, is tractable by Runge–Kutta methods. This technique retains $\mathcal{M}$-invariance under discretization when $\mathcal{M}$ is diffeomorphic to $\mathbb{R}^q$, $q \le d-1$, or to the sphere $\boldsymbol{S}_q$, $q \le d-1$. We hope to identify in the future numerical methods which retain invariance for manifolds that are diffeomorphic to torii or to manifolds of higher genera.

We refer the reader to (Guillemin and Pollack, 1974) for basic concepts and terminology of differential topology, which might not be familiar to all numerical analysts.

2. Runge–Kutta methods and manifolds

We discretize the ODE system (1.1) by means of the ν-stage Runge–Kutta method

$$\begin{aligned} \boldsymbol{\varphi}_\ell &= \boldsymbol{y}_n + h\sum_{j=1}^{\nu} a_{\ell,j}\boldsymbol{k}_j, \qquad & \ell = 1,2,\ldots,\nu, \\ \boldsymbol{k}_\ell &= \boldsymbol{f}(\boldsymbol{\varphi}_\ell), & \\ \boldsymbol{y}_{n+1} &= \boldsymbol{y}_n + h\sum_{\ell=1}^{\nu} b_\ell \boldsymbol{k}_\ell. & \end{aligned} \tag{2.1}$$

In conformity with standard terminology, we designate $A = (a_{k,\ell})$ as the *RK matrix* and call $\boldsymbol{b} = (b_\ell)$ the *RK weights* (Iserles, 1996). We say that the method (2.1) is *nonconfluent* if the *RK nodes* $c_\ell = \sum_{j=1}^{\nu} a_{\ell,j}$, $\ell = 1,2,\ldots,\nu$, are distinct.

Assuming that the solution of (1.1) is $\mathcal{M}$-invariant, it is of interest to pose the question whether the solution sequence, as discretized by the Runge–Kutta method (2.1), shares this feature, i.e.

$$\boldsymbol{y}_0 \in \mathcal{M} \qquad \Rightarrow \qquad \boldsymbol{y}_n \in \mathcal{M} \quad \forall\, n \in \mathbb{Z}^+. \tag{2.2}$$

Without much abuse of terminology, we refer to the Runge–Kutta method (2.1) as $\mathcal{M}$-*invariant* if it obeys (2.2) whenever the underlying ODE system is $\mathcal{M}$-invariant. Moreover, given a family $\mathbb{M} = \{\mathcal{M}_\alpha\}$ of manifolds in $\mathbb{R}^d$, we say that a Runge–Kutta method is $\mathbb{M}$-*invariant* if it is $\mathcal{M}_\alpha$-invariant for all α.

In this section we do not attempt to address the question of $\mathcal{M}$-invariance in its full generality. Instead, we just consider a subset of all possible manifolds which, while broad enough to be of interest in numerous issues of genuine practical importance, possesses enough structure to lend itself to our analysis. Our assumption is that

$$\mathcal{M} = \{\boldsymbol{x} \in \mathbb{R}^d \,:\, S(\boldsymbol{x}, \boldsymbol{x}, \ldots, \boldsymbol{x}) \equiv c\} \tag{2.3}$$

where $c \in \mathbb{R}$ is a constant and S is a *symmetric r-tensor.* In other words,

$$S : \overbrace{\mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d}^{r \text{ times}} \to \mathbb{R},$$

S is linear in all its components and

$$S(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_r) = S(\boldsymbol{x}_{\pi(1)}, \boldsymbol{x}_{\pi(2)}, \ldots, \boldsymbol{x}_{\pi(r)}), \qquad \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_r \in \mathbb{R}^d,$$

for all permutations $[\pi(1)\ \pi(2)\ \cdots\ \pi(r)]$ of the indices $[\,1\ 2\ \cdots\ r\,]$. We denote the set of all manifolds in $\mathbb{R}^d$ generated by symmetric r-tensors by $\mathcal{S}_r$.

The insistence on symmetry does not lead to loss of generality. Thus, let T be an r-tensor, which is not necessarily symmetric, and set

$$S(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_r) := \frac{1}{r!} \sum T(\boldsymbol{y}_{\pi(1)}, \boldsymbol{y}_{\pi(2)}, \ldots, \boldsymbol{y}_{\pi(r)}),$$

where the sum is carried out across all $r!$ permutations of the indices $[\,1\ 2\ \cdots\ r\,]$. It is easy to verify that S is a symmetric r-tensor and that

$$\{\boldsymbol{x} \in \mathbb{R}^d \,:\, S(\boldsymbol{x}, \boldsymbol{x}, \ldots, \boldsymbol{x}) \equiv c\} = \{\boldsymbol{x} \in \mathbb{R}^d \,:\, T(\boldsymbol{x}, \boldsymbol{x}, \ldots, \boldsymbol{x}) \equiv c\}.$$

Therefore we may assume without further ado that $\mathcal{M}$ is a level set of a symmetric r-tensor. This construction, although perfectly correct from the formal point of view, makes no sense when the symmetrization results in a null tensor, e.g. when T is an alternating tensor.

The case $r = 1$ corresponds to $\boldsymbol{y}(t)$ lying in an affine space: there exists $\boldsymbol{s} \in \mathbb{R}^d$ such that $\boldsymbol{s}^T \boldsymbol{y}(t) \equiv c$ for all $t \geq 0$. It is trivial to verify that every consistent numerical method is $\mathcal{S}_1$-invariant (Calvo et al, 1995).

For the quadratic case, $r = 2$, there exists a symmetric matrix $\tilde{S}$ such that $S(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{x}_1^T \tilde{S} \boldsymbol{x}_2$. It has been proved by Graeme Cooper that the Runge–Kutta method (2.1) is, in our terminology, $\mathcal{S}_2$-invariant if $M = O$ (Cooper, 1987), where the matrix M, ubiqitous in stability analysis of Runge–Kutta methods (Dekker and Verwer, 1984; Sanz-Serna and Calvo, 1994) is defined by

$$m_{k,\ell} = b_k a_{k,\ell} + b_\ell a_{\ell,k} - b_k b_\ell, \qquad k, \ell = 1, 2, \ldots, \nu.$$

Moreover, unless $M = O$, there always exists a matrix $\tilde{S}$ so that (2.2) is violated by the underlying nonconfluent Runge–Kutta method (Calvo et al., 1995).

Cubic manifolds have been already considered in (Calvo et al., 1995), where we have proved that for every nonconfluent Runge–Kutta method there exists a symmetric 3-tensor S such that the method is not $\mathcal{M}$-invariant with respect to the manifold $\mathcal{M}$, as defined by (2.3). In other words, although we cannot rule out $\mathcal{M}$-invariance for some Runge–Kutta methods and cubic manifolds, no nonconfluent Runge–Kutta method can be $\mathcal{S}_3$-invariant. In the present paper we adopt a more general approach and consider the case $r \geq 3$. Our objective, the main result of this section, is that no nonconfluent Runge–Kutta method can be $\mathcal{S}_r$-invariant for any $r \geq 3$.

We denote by $S_{k_1,k_2,\ldots,k_q}$, where $1 \leq k_1 < k_2 < \cdots < k_q \leq r$, the value of the r-tensor S when each k_jth coordinate has been replaced by $\boldsymbol{k}_{\ell_j}$, $j = 1,2,\ldots,q$, while all the remaining coordinates are $\boldsymbol{y}_n$. By virtue of linearity of S, (2.1) implies that

$$\begin{aligned}
& S(\boldsymbol{y}_{n+1}, \boldsymbol{y}_{n+1}, \ldots, \boldsymbol{y}_{n+1}) \\
= \; & S\left(\boldsymbol{y}_n + h\sum_{\ell=1}^{\nu} b_\ell \boldsymbol{k}_\ell, \boldsymbol{y}_n + h\sum_{\ell=1}^{\nu} b_\ell \boldsymbol{k}_\ell, \ldots, \boldsymbol{y}_n + h\sum_{\ell=1}^{\nu} b_\ell \boldsymbol{k}_\ell\right) \\
= \; & S(\boldsymbol{y}_n, \boldsymbol{y}_n, \ldots, \boldsymbol{y}_n) + h\sum_{k_1=1}^{r}\sum_{\ell_1=1}^{\nu} b_{\ell_1} S_{k_1} + h^2 \sum_{k_1=1}^{r-1}\sum_{k_2=k_1+1}^{r}\sum_{\ell_1,\ell_2=1}^{\nu} b_{\ell_1} b_{\ell_2} S_{k_1,k_2} \\
& + h^3 \sum_{k_1=1}^{r-2}\sum_{k_2=k_1+1}^{r-1}\sum_{k_3=k_2+1}^{r}\sum_{\ell_1,\ell_2,\ell_3=1}^{\nu} b_{\ell_1} b_{\ell_2} b_{\ell_3} S_{k_1,k_2,k_3} + \mathcal{O}\left(h^4\right).
\end{aligned}$$

Assuming that $\boldsymbol{y}_n \in \mathcal{M}$, we wish to check whether this is also true for $\boldsymbol{y}_{n+1}$. In other words, we examine whether

$$I := S(\boldsymbol{y}_{n+1}, \boldsymbol{y}_{n+1}, \ldots, \boldsymbol{y}_{n+1}) - S(\boldsymbol{y}_n, \boldsymbol{y}_n, \ldots, \boldsymbol{y}_n)$$

vanishes. By virtue of symmetry,

$$S_{k_1,k_2,\ldots,k_q} = S(\overbrace{\boldsymbol{y}_n, \boldsymbol{y}_n \ldots, \boldsymbol{y}_n}^{r-q \text{ times}}, \boldsymbol{k}_{\ell_1}, \boldsymbol{k}_{\ell_2}, \ldots, \boldsymbol{k}_{\ell_q}),$$

therefore

$$I = \sum_{q=1}^{r} \binom{r}{q} h^q \sum_{\ell_1,\ell_2,\ldots,\ell_q=1}^{\nu} b_{\ell_1} b_{\ell_2} \cdots b_{\ell_q} S(\boldsymbol{y}_n, \ldots, \boldsymbol{y}_n, \boldsymbol{k}_{\ell_1}, \ldots, \boldsymbol{k}_{\ell_q}). \quad (2.4)$$

We commence with the $q = 1$ term, noting that, according to (2.1), we may substitute

$$\boldsymbol{y}_n = \boldsymbol{\varphi}_\ell - h\sum_{j=1}^{\nu} a_{\ell,j} \boldsymbol{k}_j$$

for any $\ell \in \{1, 2, \ldots, \nu\}$. Therefore

$$\begin{aligned}
&\sum_{\ell=1}^{\nu} b_\ell S(\boldsymbol{y}_n, \ldots, \boldsymbol{y}_n, \boldsymbol{k}_\ell) \\
&= \sum_{\ell=1}^{\nu} b_\ell S\left(\boldsymbol{\varphi}_\ell - h\sum_{j_1=1}^{\nu} a_{\ell,j_1}\boldsymbol{k}_{j_1}, \ldots, \boldsymbol{\varphi}_\ell - h\sum_{j_{r-1}=1}^{\nu} a_{\ell,j_{r-1}}\boldsymbol{k}_{j_{r-1}}, \boldsymbol{k}_\ell\right) \\
&= \sum_{p=0}^{r-1}(-1)^p \binom{r-1}{p} h^p \sum_{\ell=1}^{\nu}\sum_{j_1,\ldots,j_p=1}^{\nu} b_\ell a_{\ell,j_1} a_{\ell,j_2}\cdots a_{\ell,j_p} \\
&\qquad \times S(\overbrace{\boldsymbol{\varphi}_\ell, \ldots, \boldsymbol{\varphi}_\ell}^{r-p-1 \text{ times}}, \boldsymbol{k}_{j_1}, \boldsymbol{k}_{j_2}, \ldots, \boldsymbol{k}_{j_p}, \boldsymbol{k}_\ell).
\end{aligned}$$

Note that, differentiating $S(\boldsymbol{y}, \boldsymbol{y}, \ldots, \boldsymbol{y}) \equiv c$ and exploiting symmetry, we deduce $S(\boldsymbol{y}, \ldots, \boldsymbol{y}, \boldsymbol{f}(\boldsymbol{y})) \equiv 0$, therefore, according to (2.1),

$$S(\boldsymbol{\varphi}_\ell, \ldots, \boldsymbol{\varphi}_\ell, \boldsymbol{k}_\ell) = 0.$$

Consequently,

$$\begin{aligned}
&\sum_{\ell=1}^{\nu} b_\ell S(\boldsymbol{y}_n, \ldots, \boldsymbol{y}_n, \boldsymbol{k}_\ell) = -(r-1)h\sum_{\ell,j=1}^{\nu} b_\ell a_{\ell,j} S(\boldsymbol{\varphi}_\ell, \ldots, \boldsymbol{\varphi}_\ell, \boldsymbol{k}_j, \boldsymbol{k}_\ell) \\
&\quad + \frac{(r-2)(r-1)}{2}h^2 \sum_{\ell,j_1,j_2=1}^{\nu} b_\ell a_{\ell,j_1} a_{\ell,j_2} S(\boldsymbol{\varphi}_\ell, \ldots, \boldsymbol{\varphi}_\ell, \boldsymbol{k}_{j_1}, \boldsymbol{k}_{j_2}, \boldsymbol{k}_\ell) \quad (2.5) \\
&\quad + \mathcal{O}\left(h^3\right).
\end{aligned}$$

Next, we proceed to the $q = 2$ term. By similar substitution we readily obtain

$$\begin{aligned}
&\sum_{\ell_1,\ell_2=1}^{\nu} b_{\ell_1} b_{\ell_2} S(\boldsymbol{y}_n, \ldots, \boldsymbol{y}_n, \boldsymbol{k}_{\ell_1}, \boldsymbol{k}_{\ell_2}) \\
&= \sum_{\ell_1,\ell_2=1}^{\nu} b_{\ell_1} b_{\ell_2} S\left(\boldsymbol{\varphi}_{\ell_1} - h\sum_{j=1}^{\nu} a_{\ell_1,j}\boldsymbol{k}_j, \ldots, \boldsymbol{\varphi}_{\ell_1} - h\sum_{j=1}^{\nu} a_{\ell_1,j}\boldsymbol{k}_j, \boldsymbol{k}_{\ell_1}, \boldsymbol{k}_{\ell_2}\right) \\
&= \sum_{\ell_1,\ell_2=1}^{\nu} b_{\ell_1} b_{\ell_2} S(\boldsymbol{\varphi}_{\ell_1}, \ldots, \boldsymbol{\varphi}_{\ell_1}, \boldsymbol{k}_{\ell_1}, \boldsymbol{k}_{\ell_2}) \\
&\quad - (r-2)h\sum_{\ell_1,\ell_2,j=1}^{\nu} b_{\ell_1} b_{\ell_2} a_{\ell_1,j} S(\boldsymbol{\varphi}_{\ell_1}, \ldots, \boldsymbol{\varphi}_{\ell_1}, \boldsymbol{k}_j, \boldsymbol{k}_{\ell_1}, \boldsymbol{k}_{\ell_2}) \\
&\quad + \mathcal{O}\left(h^2\right). \qquad (2.6)
\end{aligned}$$

Finally, the contribution of the $q = 3$ term is

$$\begin{aligned}
&\frac{(r-2)(r-1)r}{6}\sum_{\ell_1,\ell_2,\ell_3=1}^{\nu} b_{\ell_1} b_{\ell_2} b_{\ell_3} S(\boldsymbol{\varphi}_{\ell_1}, \boldsymbol{\varphi}_{\ell_1}, \ldots, \boldsymbol{\varphi}_{\ell_1}, \boldsymbol{k}_{\ell_1}, \boldsymbol{k}_{\ell_2}, \boldsymbol{k}_{\ell_3}) \\
&\quad + \mathcal{O}(h). \qquad (2.7)
\end{aligned}$$

Collecting together the contributions of (2.5)–(2.7) and making a frequent, albeit trivial, use of symmetry, we expand I in powers of h,

$$
\begin{aligned}
I = h^2 & \left[-(r-1)r \sum_{\ell,j=1}^{\nu} b_\ell a_{\ell,j} S(\boldsymbol{\varphi}_\ell, \ldots, \boldsymbol{\varphi}_\ell, \boldsymbol{k}_j, \boldsymbol{k}_\ell) \right. \\
& \left. + \frac{(r-1)r}{2} \sum_{\ell,j=1}^{\nu} b_\ell b_j S(\boldsymbol{\varphi}_\ell, \ldots, \boldsymbol{\varphi}_\ell, \boldsymbol{k}_j, \boldsymbol{k}_\ell) \right] \\
+ h^3 & \left[\frac{(r-2)(r-1)r}{2} \sum_{\ell,j,i=1}^{\nu} b_\ell a_{\ell,j} a_{\ell,i} S(\boldsymbol{\varphi}_\ell, \ldots, \boldsymbol{\varphi}_\ell, \boldsymbol{k}_j, \boldsymbol{k}_i, \boldsymbol{k}_\ell) \right. \\
& - \frac{(r-2)(r-1)r}{2} \sum_{\ell,j,i=1}^{\nu} b_\ell b_j a_{\ell,i} S(\boldsymbol{\varphi}_\ell, \ldots, \boldsymbol{\varphi}_\ell, \boldsymbol{k}_j, \boldsymbol{k}_i, \boldsymbol{k}_\ell) \\
& \left. + \frac{(r-2)(r-1)r}{6} \sum_{\ell,j,i=1}^{\nu} b_\ell b_j b_i S(\boldsymbol{\varphi}_\ell, \ldots, \boldsymbol{\varphi}_\ell, \boldsymbol{k}_j, \boldsymbol{k}_i, \boldsymbol{k}_\ell) \right] \\
+ \mathcal{O}&\left(h^4\right).
\end{aligned}
$$

In order to obtain $I = 0$ we need to set to zero the coefficients of all powers of h. We start with the $\mathcal{O}(h^2)$ term. Before proceding further, observe that

$$
\begin{aligned}
& S(\boldsymbol{\varphi}_\ell, \ldots, \boldsymbol{\varphi}_\ell, \boldsymbol{k}_j, \boldsymbol{k}_\ell) \qquad (2.8) \\
= \;& S\left(\boldsymbol{\varphi}_j + h \sum_{i=1}^{\nu} (a_{\ell,i} - a_{j,i}) \boldsymbol{k}_i, \ldots, \boldsymbol{\varphi}_j + h \sum_{i=1}^{\nu} (a_{\ell,i} - a_{j,i}) \boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{k}_\ell \right) \\
= \;& S(\boldsymbol{\varphi}_j, \ldots, \boldsymbol{\varphi}_j, \boldsymbol{k}_j, \boldsymbol{k}_\ell) + h \sum_{i=1}^{\nu} (a_{\ell,i} - a_{j,i}) S(\boldsymbol{\varphi}_j, \ldots, \boldsymbol{\varphi}_j, \boldsymbol{k}_i, \boldsymbol{k}_j, \boldsymbol{k}_\ell) \\
& + \mathcal{O}\left(h^2\right).
\end{aligned}
$$

Hence, the $\mathcal{O}(h^2)$ term reduces to

$$
\binom{r}{2} \sum_{\ell,j=1}^{\nu} (b_\ell b_j - b_\ell a_{\ell,j} - b_j a_{j,\ell}) S(\boldsymbol{\varphi}_\ell, \ldots, \boldsymbol{\varphi}_\ell, \boldsymbol{k}_j, \boldsymbol{k}_\ell),
$$

while adding a further contribution to the $\mathcal{O}(h^3)$ term. Consequently, $M = O$ is sufficient for setting the coefficient of h^2 to zero and later we verify that it is also necessary.

Next we turn our attention to the cubic term. We recall that, as for (2.8), it is true that

$$
S(\boldsymbol{\varphi}_\ell, \ldots, \boldsymbol{\varphi}_\ell, \boldsymbol{k}_j, \boldsymbol{k}_i, \boldsymbol{k}_\ell) = S(\boldsymbol{\varphi}_j, \ldots, \boldsymbol{\varphi}_j, \boldsymbol{k}_j, \boldsymbol{k}_i, \boldsymbol{k}_\ell) + \mathcal{O}(h),
$$

thus protracted, yet easy, algebra yields

$$\binom{r}{3}\sum_{\ell,j,i=1}^{\nu}\{b_\ell b_j b_i-(b_\ell b_j a_{\ell,i}+b_j b_i a_{j,\ell}+b_i b_\ell a_{i,j})$$
$$-[b_j a_{j,\ell}(a_{j,i}-a_{\ell,i})+b_\ell a_{\ell,i}(a_{\ell,j}-a_{i,j})+b_i a_{i,j}(a_{i,\ell}-a_{j,\ell})]$$
$$+(b_\ell a_{\ell,j}a_{\ell,i}+b_j a_{j,\ell}a_{j,i}+b_i a_{i,\ell}a_{i,j})\}\,S(\boldsymbol{\varphi}_\ell,\ldots,\boldsymbol{\varphi}_\ell,\boldsymbol{k}_j,\boldsymbol{k}_i,\boldsymbol{k}_\ell),$$

which, after further simplification, reduces to

$$\binom{r}{3}\sum_{\ell,j,i=1}^{\nu}[b_\ell b_j b_i-(b_\ell b_j a_{\ell,i}+b_j b_i a_{j,\ell}+b_i b_\ell a_{i,j})$$
$$+b_j a_{j,\ell}a_{\ell,i}+b_\ell a_{\ell,i}a_{i,j}+b_i a_{i,j}a_{j,\ell})]\,S(\boldsymbol{\varphi}_\ell,\ldots,\boldsymbol{\varphi}_\ell,\boldsymbol{k}_j,\boldsymbol{k}_i,\boldsymbol{k}_\ell).$$

We thus deduce that the conditions for eliminating the h^2 and h^3 terms are

$$b_\ell a_{\ell,j}+b_j a_{j,\ell}-b_\ell b_j=0,\qquad \ell,j=1,2,\ldots,\nu, \tag{2.9}$$

and

$$\begin{aligned}&(b_j a_{j,\ell}a_{\ell,i}+b_\ell a_{\ell,i}a_{\ell,j}+b_i a_{i,j}a_{i,\ell})\\&\quad-(b_\ell b_j a_{\ell,i}+b_j b_i a_{j,\ell}+b_i b_\ell a_{i,j})+b_\ell b_j b_i=0,\end{aligned}\qquad \ell,j,i=1,2,\ldots,\nu, \tag{2.10}$$

respectively. Incidentally, if we impose that the numerical scheme satisfies (2.9), we can use the latter to express the terms which are quadratic in the components of the vector $\boldsymbol{b}$ and obtain a relation equivalent to (2.10), namely

$$b_i b_j b_\ell-(b_i a_{i,j}a_{i,\ell}+b_j a_{j,\ell}a_{j,i}+b_\ell a_{\ell,i}a_{\ell,j})=0,\qquad i,j,\ell=1,\ldots,\nu,$$

as already derived in (Calvo et al., 1995).

Our contention is that the conditions (2.9) and (2.10) conflict for every Runge–Kutta method of nontrivial order. To this end we note that order $p\geq 1$ implies that $\mathbf{1}^T\boldsymbol{b}=1$, therefore there exists $s\in\{1,2,\ldots,\nu\}$ such that $b_s\neq 0$. We let $\ell=j=i=s$ in (2.9), (2.10), and this yields

$$2b_s a_{s,s}-b_s^2=0, \tag{2.11}$$
$$3b_s a_{s,s}^2-3b_s^2 a_{s,s}+b_s^3=0. \tag{2.12}$$

Since $b_s\neq 0$, we deduce from (2.11) that $a_{s,s}=\frac{1}{2}b_s$ and substitution in (2.12) results in $\frac{1}{4}b_s^3=0$, hence a contradiction.

The statement that (2.9) and (2.10) conflict does not mean, on its own, that no Runge–Kutta method can solve the underlying ODE in an $\mathcal{S}_r$-invariant manner for $r \geq 3$. To be able to argue the latter statement, we need to demonstrate that there exists $\mathcal{M} \in \mathcal{S}_r$ and an $\mathcal{M}$-invariant ODE (1.1) for which neither $S(\boldsymbol{\varphi}_s, \ldots, \boldsymbol{\varphi}_s, \boldsymbol{f}(\boldsymbol{\varphi}_s), \boldsymbol{f}(\boldsymbol{\varphi}_s))$ nor $S(\boldsymbol{\varphi}_s, \ldots, \boldsymbol{\varphi}_s, \boldsymbol{f}(\boldsymbol{\varphi}_s), \boldsymbol{f}(\boldsymbol{\varphi}_s), \boldsymbol{f}(\boldsymbol{\varphi}_s))$ vanish (recall that, by (2.1), $\boldsymbol{k}_s = \boldsymbol{f}(\boldsymbol{\varphi}_s)$). Fortunately, a relatively simple member of $\mathcal{S}_r$ furnishes us with the required example. For the sake of simplicity we thereafter assume that the underlying method (2.1) is nonconfluent.

Letting $d = 3$, we choose

$$\mathcal{M} = \{\boldsymbol{y} \in \mathbb{R}^3 \,:\, y_1^r + y_2^r = y_1^r(0) + y_2^r(0)\}.$$

In other words, S is a scalar multiple of the unit r-tensor (which is of course symmetric). We consider the ODE

$$\begin{aligned} y_1' &= y_2^{r-1} g(\boldsymbol{y}), \\ y_2' &= -y_1^{r-1} g(\boldsymbol{y}), \\ y_3' &\equiv 1, \end{aligned} \tag{2.13}$$

where $g : \mathbb{R}^3 \to \mathbb{R}$ is for the time being an arbitrary Lipschitz function. It is trivial to verify that

$$y_1^{r-1}(t) y_1'(t) + y_2^{r-1}(t) y_2'(t) \equiv 0, \qquad t \geq 0,$$

hence, as long as $\boldsymbol{y}(0) \in \mathcal{M}$, also $\boldsymbol{y}(t) \in \mathcal{M}$ for all $t \geq 0$ and the ODE (2.13) is $\mathcal{M}$-invariant. Moreover, simple calculation confirms that

$$\begin{aligned} S(\boldsymbol{\varphi}, \ldots, \boldsymbol{\varphi}, \boldsymbol{f}(\tilde{\boldsymbol{\varphi}}), \boldsymbol{f}(\boldsymbol{\varphi})) &= g(\boldsymbol{\varphi}) g(\tilde{\boldsymbol{\varphi}}) (\varphi_1 \varphi_2)^{r-2} \\ &\quad \times (\varphi_2 \tilde{\varphi}_2^{r-1} + \varphi_1 \tilde{\varphi}_1^{r-1}), \\ S(\boldsymbol{\varphi}, \ldots, \boldsymbol{\varphi}, \boldsymbol{f}(\tilde{\boldsymbol{\varphi}}), \boldsymbol{f}(\hat{\boldsymbol{\varphi}}), \boldsymbol{f}(\boldsymbol{\varphi})) &= g(\boldsymbol{\varphi}) g(\tilde{\boldsymbol{\varphi}}) g(\hat{\boldsymbol{\varphi}}) \\ &\quad \times (\varphi_1 \varphi_2)^{r-3} (\varphi_2^2 \tilde{\varphi}_2^{r-1} \hat{\varphi}_2^{r-1} - \varphi_1^2 \tilde{\varphi}_1^{r-1} \hat{\varphi}_1^{r-1}) \end{aligned}$$

where $\boldsymbol{\varphi}^T = [\, \varphi_1 \quad \varphi_2 \quad \varphi_3 \,]$.

We choose $\boldsymbol{y}_0 \in \mathbb{R}^3$ so that $y_{0,1}, y_{0,2}, y_{0,2}^r - y_{0,1}^r \neq 0$ and $y_{0,3} = 0$. Let $\boldsymbol{\varphi}_s$ be a real solution of the polynomial system

$$\begin{aligned} \varphi_{s,1} &= y_{0,1} + h a_{s,s} \varphi_{s,2}^{r-1}, \\ \varphi_{s,2} &= y_{0,2} - h a_{s,s} \varphi_{s,1}^{r-1}. \end{aligned}$$

Note that there exists such solution as long as $h > 0$ is sufficiently small. Next, set

$$\begin{aligned} \varphi_{\ell,1} &:= y_{0,1} + h a_{\ell,s} \varphi_{s,2}^{r-1}, \\ \varphi_{\ell,2} &:= y_{0,2} - h a_{\ell,s} \varphi_{s,1}^{r-1} \end{aligned}$$

for all $\ell = 1, 2, \ldots, \nu$, $\ell \neq s$ and choose a Lipschitz function g such that

$$g(\boldsymbol{\varphi}_\ell) = \begin{cases} 1, & \ell = s, \\ 0, & \ell \in \{1, 2, \ldots, \nu\},\ \ell \neq s. \end{cases}$$

Note that

$$\varphi_{\ell,3} = y_{0,3} + h \sum_{j=1}^{\nu} a_{\ell,j} = hc_\ell, \qquad \ell = 1, 2, \ldots, \nu,$$

therefore confluence implies that $\boldsymbol{\varphi}_\ell \neq \boldsymbol{\varphi}_m$, $\ell, m = 1, 2, \ldots, \nu$, $\ell \neq m$, and such function g exists.

It follows from (2.1) and our construction that, as long as $h > 0$ is sufficiently small, we have

$$S(\boldsymbol{\varphi}_s, \ldots, \boldsymbol{\varphi}_s, \boldsymbol{f}(\boldsymbol{\varphi}_s), \boldsymbol{f}(\boldsymbol{\varphi}_s)), S(\boldsymbol{\varphi}_s, \ldots, \boldsymbol{\varphi}_s, \boldsymbol{f}(\boldsymbol{\varphi}_s), \boldsymbol{f}(\boldsymbol{\varphi}_s), \boldsymbol{f}(\boldsymbol{\varphi}_s)) \neq 0$$

and

$$S(\boldsymbol{\varphi}_\ell, \ldots, \boldsymbol{\varphi}_\ell, \boldsymbol{f}(\boldsymbol{\varphi}_j), \boldsymbol{f}(\boldsymbol{\varphi}_\ell)), S(\boldsymbol{\varphi}_\ell, \ldots, \boldsymbol{\varphi}_\ell, \boldsymbol{f}(\boldsymbol{\varphi}_j), \boldsymbol{f}(\boldsymbol{\varphi}_i), \boldsymbol{f}(\boldsymbol{\varphi}_\ell)) = 0$$

for all $\ell \neq s$, $i, j = 1, 2, \ldots, \nu$. Consequently, both (2.11) and (2.12) must be true in order for the $\mathcal{O}(h^2)$ and $\mathcal{O}(h^3)$ terms to be eliminated. This, however, is impossible according to our analysis.

We deduce the central result of this section.

Theorem 1 *Let $r \geq 3$ be a given integer. No nonconfluent Runge–Kutta method* (2.1) *can be $\mathcal{S}_r$-invariant when $\mathcal{S}_r$ is the set of all manifolds described by symmetric r-tensors.* $\square$

3. Invariant modification of Runge–Kutta methods

Let us suppose that it is known (typically, as a consequence of mathematical analysis) that the ODE (1.1) is $\mathcal{M}$-invariant for some q-dimensional manifold $\mathcal{M} \subset \mathbb{R}^d$, $q \leq d-1$. Further, we assume that $\mathcal{M}$ is diffeomeorphic to another manifold, $\mathcal{N}$, say. In other words, there exists a smooth bijection

$$\boldsymbol{g} : \mathcal{M} \to \mathcal{N}$$

whose inverse is also smooth. Letting $\boldsymbol{x}(t) := \boldsymbol{g}(\boldsymbol{y}(t))$, $t \geq 0$, we readily derive from (1.1) a differential equation for the function $\boldsymbol{x}$,

$$\boldsymbol{x}' = \frac{\partial \boldsymbol{g}(\boldsymbol{g}^{-1}(\boldsymbol{x}))}{\partial \boldsymbol{y}} \boldsymbol{f}(\boldsymbol{g}^{-1}(\boldsymbol{x})), \quad t \geq 0, \qquad \boldsymbol{x}(0) = \boldsymbol{g}(\boldsymbol{y}_0) \in \mathcal{N}. \quad (3.1)$$

Since $\boldsymbol{y}(t) \in \mathcal{M}$, $t \geq 0$, it follows from our construction that $\boldsymbol{x}(t) \in \mathcal{N}$, $t \geq 0$, therefore (3.1) is $\mathcal{N}$-invariant. Thus, provided that the manifold $\mathcal{N}$ is sufficiently simple and we can integrate the ODE (3.1) with an $\mathcal{N}$-invariant numerical method, we can construct an $\mathcal{M}$-invariant solution of (1.1) by the inverse mapping $\boldsymbol{y}_n = \boldsymbol{g}^{-1}(\boldsymbol{x}_n)$, $n \in \mathbb{N}$.

One way of expressing our argument is that, whenever we know that an ODE is $\mathcal{M}$-invariant, it is possible to map $\mathcal{M}$ diffeomorphically into a simpler manifold, $\mathcal{N}$, for which we have our numerical analysis sorted out. An alternative point of view is that, whenever we can construct $\mathcal{N}$-invariant methods for some manifold $\mathcal{N}$, we can extend our result to all manifolds $\mathcal{M}$ that are diffeomorphic to $\mathcal{N}$.

On the face of it, we have just (in the best mathematical tradition) replaced one problem, how to construct $\mathcal{M}$-invariant methods, by two: firstly, how to find a diffeomorphism from $\mathcal{M}$ to some 'convenient' manifold $\mathcal{N}$ and, secondly, how to integrate along the second manifold in an $\mathcal{N}$-invariant manner. However, this difficulty is often illusory since many interesting manifolds lend themselves easily to a 'natural' diffeomorphism to a simple manifold. Moreover, this approach helps us to discard redundant information and focus on the real problem in numerical $\mathcal{M}$-invariance: *not the precise shape of $\mathcal{M}$ but its topology!*

We mention in passing that the idea of changing variables, to map an invariant manifold $\mathcal{M}$ into a manifold $\mathcal{N}$ with simpler topology, has been already considered in a different context and with a readically different approach in (Leimkuhler and Partick, 1995).

The remainder of this section is devoted to two simple examples which illustrate its main idea. Firstly, let $\mathcal{M}$ be the unit sphere with respect to the ℓ_p norm, $1 < p < \infty$,

$$\mathcal{M} = \left\{ \boldsymbol{y} \in \mathbb{R}^d \,:\, \sum_{k=1}^{d} |y_k|^p = 1 \right\}.$$

The $(d-1)$-dimensional manifold $\mathcal{M}$ is trivially diffeomorphic to the sphere $\boldsymbol{S}_{d-1}$ (which, of course, is nothing else but the unit sphere with respect to the Euclidean norm ℓ_2) and it is easy to find a diffeomorphism. Let

$$\boldsymbol{g}(\boldsymbol{y}) = \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|_2}, \qquad \boldsymbol{y} \in \mathcal{M},$$

whereby the inverse map is

$$\boldsymbol{g}^{-1}(\boldsymbol{x}) = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_p}, \qquad \boldsymbol{x} \in \boldsymbol{S}_{d-1}.$$

We thus obtain the ODE system

$$\boldsymbol{x}' = \|\boldsymbol{x}\|_p (I - \boldsymbol{x}\boldsymbol{x}^T) \boldsymbol{f}(\boldsymbol{x}/\|\boldsymbol{x}\|_p). \tag{3.2}$$

Note that, although this might not be apparent at a glance, the equation (3.2) reduces in the special case $p = 2$ to the original ODE (1.1) (as it should) because $\boldsymbol{x}^T \boldsymbol{f}(\boldsymbol{x}) \equiv 0$ for equations contrained to $\boldsymbol{S}_{d-1}$.

The solution of (3.2) is $\boldsymbol{S}_{d-1}$-invariant. As we have already stated in Section 2, employing a different terminology, any Runge–Kutta method with $M = O$ is $\boldsymbol{S}_{d-1}$-invariant. Therefore we may solve (3.2) with such a method and an $\mathcal{M}$-invariant solution of the original ODE follows by letting $\boldsymbol{y}_n = \boldsymbol{g}^{-1}(\boldsymbol{x}_n)$ for all relevant values of n.

An interesting example of such a manifold $\mathcal{M}$, which harks back to the work of Section 2, is when $p \geq 4$ is an even integer. In that case $\mathcal{M}$ follows by letting S in (2.3) be a unit p-tensor. This is an example of a manifold which, according to Theorem 1, is intractable directly by Runge–Kutta methods in an invariant manner. Yet, subjecting it to a transformation of variables, we can solve it whilst maintaining $\mathcal{M}$-invariance.

Odd values of r in (2.3) can be addressed in a similar manner, except that the topology of $\mathcal{M}$ is no longer the same as of the sphere $\boldsymbol{S}_{d-1}$. Consider the manifold

$$\mathcal{M} = \left\{ \boldsymbol{y} \in \mathbb{R}^d \,:\, \sum_{k=1}^{d} y_k^r = 1 \right\},$$

where r is odd. Fixing $y_1, y_2, \ldots, y_{d-1}$, the equation

$$y_d^r = 1 - \sum_{k=1}^{d-1} y_k^r$$

has exactly one real solution. Therefore, $\mathcal{M}$ is diffeomorphic to $\mathbb{R}^{d-1}$,

$$\boldsymbol{g}(\boldsymbol{y}) = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d-1} \end{bmatrix}$$

and

$$\boldsymbol{g}^{-1}(\boldsymbol{x}) = \begin{bmatrix} x_1 \\ \vdots \\ x_{d-1} \\ \operatorname{sgn}\left(1 - \sum_{k=1}^{d-1} x_k^r\right) \times \left|1 - \sum_{k=1}^{d-1} x_k^r\right|^{1/r} \end{bmatrix}.$$

The problem of $\mathcal{M}$-invariance is now, if at all, even easier. We map $\boldsymbol{y}$ into $\mathbb{R}^{d-1}$ by the natural projection $\boldsymbol{g}$, solve the ODE (3.1) by an arbitrary consistent numerical method and map back from $\{\boldsymbol{x}_n\}$ to $\{\boldsymbol{y}_n\}$.

Needless to say, an Euclidean subspace and a sphere represent the simplest possible topologies. As things stand, it is an open question how to retain invariance on torii, for example, saying nothing of more complicated topologies. Much remains to be done in elucidating numerical methods that capture qualitative attributes of the underlying ODE system. Having said this, we emphasize again our conclusion, namely that the critical factor in the retention of invariance under discretization is not the shape but the topology of the underlying manifold.

The technique of diffeomorphisms is implicit in an earlier work of the present authors. Given an arbitrary $d \times d$ matrix L_0, we consider the manifold of matrices similar to L_0, that we denote by

$$\mathcal{M}_{L_0} = \{V L_0 V^{-1} : V \in \mathrm{GL}(d, \mathbb{R})\},$$

where $\mathrm{GL}(d,\mathbb{R})$ is the group of $d \times d$ invertible matrices with real entries. Letting

$$L(t) = V(t) L_0 V(t)^{-1}, \qquad V(t) \in \mathrm{GL}(d,\mathbb{R}),$$

differentiation yields

$$L' = V'V^{-1}L - LV'V^{-1}, \tag{3.3}$$

thus (3.3) characterizes flows on $\mathcal{M}_{L_0}$. However the function V is not known a priori. An important flow on $\mathcal{M}_{L_0}$, closely allied with (3.3), has the form

$$L' = B(L)L - LB(L) = [B(L), L], \quad t \geq 0, \qquad L(0) = L_0. \tag{3.4}$$

It is known as an *isospectral flow* and its numerical solution has been debated at length in (Calvo et al., 1995).

Comparing (3.3) and (3.4) we deduce that

$$(B(L) - V'V^{-1})L - L(B(L) - V'V^{-1}) = O,$$

thus $B(L) - V'V^{-1}$ and L commute. This means that

$$B(L) - V'V^{-1} = p(L),$$

where $p(L)$ is a polynomial in L. Hence

$$V' = (B(L) - p(L))V. \tag{3.5}$$

Once the initial condition for (3.5) is fixed, the function V is uniquely determined. The initial condition $V(0) = V_0$ must commute with L_0, therefore V_0 is a polynomial in L_0. Choosing $V_0 = I$ and letting $p(L) = O$, we obtain that V is the unique solution of the differential equation

$$V' = B(VL_0V^{-1})V, \qquad V(0) = I. \tag{3.6}$$

Thus,

$$\mathcal{N} = \mathrm{GL}(d, \mathbb{R})$$

and the natural diffeomorfism

$$g : \mathcal{M}_{L_0} \to \mathcal{N}$$

is $g(L) = V$, where V is the solution of (3.6). The inverse map is

$$g^{-1}(V) = VL_0V^{-1}. \tag{3.7}$$

It follows that, solving (3.6) with any RK method and then mapping back to $\mathcal{M}_{L_0}$ by means of (3.7), we obtain a solution that is similar (an isospectral solution) to the initial condition L_0. The mapping g from $\mathcal{M}_{L_0}$ to $\mathcal{N} = \mathrm{GL}(d, \mathbb{R})$ has the advantage that classical RK methods, subject to adequate choices of parameters (for example, the stepsize h) can solve (3.6) whilst staying on $\mathrm{GL}(d, \mathbb{R})$, while, as proved in (Calvo et al. 1995), they cannot be isospectral, because they fail in recovering cubic conservation laws that are essential for isospectrality as soon as $d \geq 3$.

Another special case is when the matrix function $B(L)$ is skew-symmetric. In this situation the solution of (3.6) is orthogonal and $L(t) = V(t)L_0V(t)^T$, therefore the most obvious diffeomorfism is

$$g(L) = V \in \mathcal{N}, \qquad \mathcal{N} = O(d),$$

where $O(d)$ is the orthogonal group of $d \times d$ matrices. Since orthogonality can be viewed as a quadratic conservation law obeyed by $V(t)$, it follows that RK schemes with $M = O$ can be used to solve (3.6) on $O(d)$ and, mapping back by means of $L = g^{-1}(V) = VL_0V^T$, we obtain a solution L which is orthogonally similar to the initial condition. This is, as a matter of fact, the main idea of (Calvo et al., 1995).

The latter two examples differ from our earlier discussion, because the diffeomorfism depends in an implicit manner on the solution. Yet, as evident from (Calvo et al., 1995), this is not an unsurmountable impediment to the implementation of the main idea of this section.

Acknowledgements This paper has been written during the second author's visit to Departamento de Matemática Aplicada y Computación, Universidad de Valladolid as an IBERDROLA Visiting Professor.

References

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

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

P.E. Crouch and R. Grossman (1993), "Numerical integration of ordinary differential equations on manifolds", *J. Nonlinear Sci.* **3**, 1–33.

K. Dekker and J.G. Verwer (1984), *Stability of Runge–Kutta Methods for Stiff Nonlinear Differential Equations,* North Holland, Amsterdam.

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

V. Guillemin and A. Pollack (1974), *Differential Topology,* Prentice–Hall, Englewood Cliffs NJ.

A. Iserles (1996), *A First Course in the Numerical Analysis of Differential Equations,* Cambridge University Press, Cambridge.

A. Iserles and A. Zanna (1995), "Qualitative numerical analysis of ordinary differential equations", DAMTP Tech. Rep. 1995/NA05, to appear in *Lectures in Applied Mathematics,* American Mathematical Society, Providence RI.

B. Leimkuhler and G.W. Patrick (1995), "A symplectic integrator for Riemannian manifolds", Kansas University Technical Report.

H. Munthe-Kaas (1995), "Lie–Butcher theory for Runge–Kutta methods", *BIT* **35**, 572–587.

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