UNIVERSITY OF CAMBRIDGE

Numerical Analysis Reports

Semi-explicit methods for isospectral flows

Mari Paz Calvo, Arieh Iserles and Antonella Zanna

DAMTP NA1997/16
August 1997

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

Semi-explicit methods for isospectral flows

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

Abstract

In this paper we propose semi-explicit schemes based on Taylor methods for the solution of the isospectral equation $L' = [B, L]$ for $d \times d$ real matrices L, while reproducing the isospectrality of the exact equation. Although the theoretical solution may be symmetric, the proposed schemes usually do not retain symmetry of the underlying flow. We present techniques that allow us to decrease the breakdown in symmetry.

1 Introduction

Isospectral flows are characterized in terms of the matrix differential equation

$$L' = [B, L], \qquad L(0) = L_0, \tag{1.1}$$

where $L, B \in \mathbb{R}^{d\times d}$ and L_0 is a given $d \times d$ initial matrix. $B \equiv B(t, L)$ is a matrix function, which is allowed to depend on L and on the time t. The square brackets denote the commutator or *Lie bracket* on matrices, namely $[B, L] = BL - LB$. It is well known (cf. Toda, 1981; Perelomov, 1990) that the solution of (1.1) can be written in the form

$$L(t) = U(t)L_0[U(t)]^{-1},$$

whereby the matrix U is the theoretical solution of the linear differential equation

$$U' = B(t, L)U, \qquad U_0 = I.$$

It is easy to verify that, if we introduce a partition $0 \equiv t_0 < t_1 < t_2 < \cdots$, the theoretical solution L can also be written in the form

$$L(t_{n+1}) = U(t_{n+1})L(t_n)[U(t_{n+1})]^{-1}, \tag{1.2}$$

the function U being the solution of

$$U' = B(t, L)U, \quad t_n \le t \le t_{n+1}, \qquad U(t_n) = I. \tag{1.3}$$

In the particular case when B is a skew-symmetric matrix function, the solution of (1.3) is orthogonal, therefore we have

$$L(t_{n+1}) = U(t_{n+1})L(t_n)[U(t_{n+1})]^{\mathrm{T}}, \tag{1.4}$$

[1]Departamento de Matemática Aplicada y Computación, Universidad de Valladolid, Valladolid, Spain.
The research of M.P. Calvo has been supported by *Junta de Castilla y León* and by project DGICYT PB95-0705.
[2]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, England.
[3]Newnham College, University of Cambridge, Cambridge, England.

and, provided that L_0 is symmetric, so is $L(t)$ for all $t \geq 0$.

We are particularly interested in the case when B is skew-symmetric, and will assume, unless explicitely mentioned, that $B^{\mathrm{T}} = -B$ throughout the rest of the present paper. This is typical of isospectral flows (e.g. Toda and Kac–van Moerbeke flows) arising in physics (Toda, 1981) and many isospectral flows with relevance to numerical linear algebra (Chu, 1994).

Although classical schemes for ordinary differential equations (ODEs) do not retain isospectrality, the authors have proposed an alternative approach to retain isospectrality under discretization (Calvo et al., 1995). The basic idea consists of solving the associated orthogonal flow

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

with an orthogonal Runge–Kutta (RK) method, e.g. an order-2ν Gauss–Legendre Runge–Kutta scheme, that, in tandem with (1.4), produces an isospectral approximation to $L(t)$. We have called these methods *modified Gauss–Legendre Runge–Kutta schemes* (MGLRK) because it is possible to show (Calvo et al., 1995) that they can be formulated as an *isospectral* modification of their classical RK counterpart when applied to (1.1). The MGLRK schemes are in general quite expensive. On the other hand, unless (1.3) is solved with an orthogonal method, the substitution in (1.4) does not produce an isospectral approximation!

It is easy to observe that isospectrality is motivated by the formula (1.2). Therefore, the correct approach is that, even if (1.3) is not solved with an orthogonal method, one should update the matrix L by means of (1.2) instead of (1.4). It follows that we might solve (1.3) with an arbitrary scheme for ODEs, whereby substitution in (1.2) produces an isospectral approximation.

Although this approach is very appealing, since it allows us to obtain an isospectral approximation in a cheaper way, e.g. with explicit RK methods, there are two main problems. Firstly, the numerical approximation in general fails to retain the symmetry of the exact solution. In many cases, symmetry is important because it is related to physical features that we want to reproduce under discretization (Perelomov, 1990; Toda, 1981) and since many isospectral problems originate in numerical algebra calculations for symmetric matrices (Chu, 1994). Secondly, loss of symmetry implies that, although the flow remains isospectral, the numerical approximation is no longer a normal matrix and its Euclidean norm is not equal to its spectral radius. Instead, we have the inequality

$$\|L\|_2 \leq \kappa \|L_0\|_2, \qquad \kappa = \|U\|_2 \cdot \|U^{-1}\|_2, \tag{1.5}$$

where κ is the spectral condition number of the matrix U. If the approximation for U is orthogonal then the above relation becomes an equality and no problems occur. Otherwise, ill-conditioning and rapid error accumulation might occur if κ becomes large.

The paper is organized as follows. In §2 we briefly review the MGLRK schemes that have been proposed in (Calvo et al., 1995). In §3 we present the semi-explicit isospectral schemes based on Taylor methods for ODEs and we also introduce time-stepping techniques that allow us to obtain higher order of symmetry for the numerical approximations and better conditioning, therefore a smaller condition number κ. Finally in §4 we present some examples and provide numerical results.

2 Modified Gauss–Legendre RK schemes

Gauss–Legendre RK are a useful example of structural orthogonal integrators (Dieci, Russell and van Vleck, 1994; Iserles and Zanna, 1995) and we may wish to implement such methods to perform the numerical integration of

$$U' = B(t, L(t))U, \qquad U(t_n) = I, \tag{2.1}$$

in tandem with

$$L_{n+1} = U_{n+1} L_n U_{n+1}^{\mathrm{T}}. \tag{2.2}$$

Since $L(t)$ is unknown for $t > t_n$ and we need to evaluate $B(t, L(t))$ at the Gauss–Legendre nodes $t_n + c_\ell h$, $\ell = 1, \ldots, \nu$, we replace the function $L(t)$ by a polynomial $\tilde{L}(t)$ which interpolates the exact flow $L(t)$ at the points t_n and t_{n+1}. Thus, for a ν-stage GLRK scheme (order 2ν), we let

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

be the Hermite polynomial interpolant of L of degree $2\nu - 1$ at $\{t_n, t_{n+1}\}$. Replacing $L(t)$ by $\tilde{L}(t)$ we are committing at most an error of $\mathcal{O}(h^{2\nu+1})$, which is subsumed in the error of a numerical method of order 2ν (Calvo et al., 1995).

In the sequel we refer to a *modified* Gauss–Legendre RK method (MGLRK) of order 2ν whenever we apply the classical Gauss–Legendre RK method of order 2ν to

$$U' = B(t, \tilde{L}(t))U, \qquad U(t_n) = I, \tag{2.3}$$

in tandem with (2.2). The schemes are implicit, and we solve them by fixed point iteration, since a Newton iteration (unless performed to machine accuracy) would not produce an orthogonal approximation. Our choice is the simplest, namely the Picard iteration. We set, as an initial guess,

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

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

$$[U^{[k]}]' = B(t, \tilde{L}^{[k]})U^{[k]}, \qquad U^{[k]}(t_n) = I.$$

Finally we update the numerical approximation of the unknown function L by means of

$$L_{n+1}^{[k+1]} = U_{n+1}^{[k]} L_n [U_{n+1}^{[k]}]^{\mathrm{T}}. \tag{2.4}$$

The simplest method in the family of modified Gauss–Legendre RK is obtained by letting $\nu = 1$. In this case,

$$\tilde{L}(t) = \frac{1}{t_{n+1} - t_n}\Big[(t_{n+1} - t)L_n + (t - t_n)L_{n+1}\Big],$$

therefore, denoting

$$B_{n+\frac{1}{2}} = B\left(t_n + \tfrac{1}{2}h, \tfrac{1}{2}(L_n + L_{n+1})\right), \tag{2.5}$$

the scheme yields

$$U_{n+1} = \left(I - \tfrac{1}{2}hB_{n+\frac{1}{2}}\right)^{-1} \left(I + \tfrac{1}{2}hB_{n+\frac{1}{2}}\right). \tag{2.6}$$

Hence we let

$$L_{n+1} = U_{n+1} L_n U_{n+1}^{\mathrm{T}}. \tag{2.7}$$

We refer to (2.5)–(2.7) as the *modified implicit midpoint rule* (MIMR), since it yields

$$L_{n+1} = L_n + h\Big[B_{n+\frac{1}{2}}, \tfrac{1}{2}(L_n + L_{n+1})\Big] + \tfrac{1}{4}h^2 B_{n+\frac{1}{2}}(L_{n+1} - L_n)B_{n+\frac{1}{2}}, \tag{2.8}$$

an isospectral $\mathcal{O}(h^3)$ modification of the implicit midpoint rule (IMR).

The formula (2.8) also suggests an alternative way to implement the MIMR. Specifically,

$$\begin{aligned}
L_{n+1}^{[0]} &= L_n, \\
B_{n+\frac{1}{2}}^{[k]} &= B\left(t_n + \tfrac{1}{2}h, \tfrac{1}{2}(L_n + L_{n+1}^{[k]})\right), \qquad\qquad k = 0, 1, \ldots, \\
L_{n+1}^{[k+1]} &= L_n + h\Big[B_{n+\frac{1}{2}}^{[k]}, \tfrac{1}{2}(L_n + L_{n+1}^{[k]})\Big] + \tfrac{1}{4}h^2 B_{n+\frac{1}{2}}^{[k]}(L_{n+1}^{[k]} - L_n)B_{n+\frac{1}{2}}^{[k]},
\end{aligned} \tag{2.9}$$

a procedure that avoids the direct solution of the associated orthogonal flow. However, an isospectral solution is obtained only if the latter scheme is iterated to convergence, because otherwise the numerical approximation is not necessarily isospectral.[4] The standard implicit midpoint rule, however, is not isospectral, even if iterated to convergence.

The advantage of the procedure (2.5)–(2.7) is that, by solving directly the associated orthogonal flow, it yields isospectrality at each inner iteration and need not be iterated to convergence.

3 Semi-explicit schemes for isospectral flows

3.1 Taylor methods and orthogonal flows

A Taylor method of order p for the orthogonal flow

$$U' = BU, \qquad U(0) = U_0, \tag{3.1}$$

with a skew-symmetric matrix function B, where U_0 is a $d \times d$ orthogonal matrix, produces the approximation

$$U_{n+1} = U_n + hU_n' + \frac{h^2}{2}U_n'' + \cdots + \frac{h^p}{p!}U_n^{(p)}, \tag{3.2}$$

where

$$U_n^{(\ell)} = U^{(\ell)}(t)|_{t=t_n}, \qquad \ell = 1, 2, \ldots, p,$$

[4] To verify this statement, it is sufficient to consider the first inner iteration in (2.9), for which we have exactly the same outcome as produced by the IMR implemented with fixed point iterations.

and

$$
\begin{aligned}
U' &= BU, \\
U'' &= B' + BU' = (B' + B^2)U, \\
U''' &= (B'' + 2B'B + BB' + B^3)U
\end{aligned}
$$

etc. In general, due to linearity of (3.1), it is easy to verify that the derivatives of U satisfy the equation

$$U^{(\ell)} = F_\ell U, \qquad \ell = 1, 2, \ldots,$$

where the sequence of functions $\{F_\ell\}$ obeys the recurrence relation

$$F_{\ell+1} = F'_\ell + F_\ell B, \qquad F_0 = I. \tag{3.3}$$

It follows from (3.2) that, provided that $U_n = U(t_n)$,

$$U_{n+1} - U(t_{n+1}) = -\frac{h^{p+1}}{(p+1)!} U^{(p+1)}(t_n) + \mathcal{O}\left(h^{p+2}\right).$$

Lemma 1 *The only Taylor method that can be orthogonal for all orthogonal flows that satisfy* (3.1) *is the exact solution ($p = \infty$).*

Proof. Assume without loss of generality that $U_n = U(t_n) = I$. Then, taking into account that the theoretical solution $U(t)$ is orthogonal, we have

$$
\begin{aligned}
U_{n+1}^{\mathrm{T}} U_{n+1} &= \left[U^{\mathrm{T}}(t_{n+1}) - \frac{h^{p+1}}{(p+1)!} F_{p+1,n}^{\mathrm{T}}\right] \left[U(t_{n+1}) - \frac{h^{p+1}}{(p+1)!} F_{p+1,n}\right] + \mathcal{O}\left(h^{p+2}\right) \\
&= I - \frac{h^{p+1}}{(p+1)!} \left[U^{\mathrm{T}}(t_{n+1}) F_{p+1,n} + F_{p+1,n}^{\mathrm{T}} U(t_{n+1})\right] + \mathcal{O}\left(h^{p+2}\right),
\end{aligned}
$$

where $F_{p+1,n} \equiv F_{p+1}(t_n)$. Bearing in mind that $U(t_{n+1}) = I + \mathcal{O}(h)$ ant that $F_{p+1,n}$ is in general not skew-symmetric, it follows that the h^{p+1} term is non-zero and the Taylor method with finite p is not orthogonal. □

3.2 The semi-explicit isospectral Taylor methods

The result in Lemma 1 implies that Taylor methods applied to (1.1) cannot be isospectral for all isospectral flows. In fact, since they do not retain orthogonality, they do not retain quadratic conservation laws that are necessary for isospectrality, as proved in (Calvo et al., 1995).

In order to obtain an isospectral numerical approximation for (1.1), we can solve (1.3) with a Taylor method, but since the numerical approximation U_{n+1} is not orthogonal, we must use the formula

$$L_{n+1} = U_{n+1} L_n U_{n+1}^{-1}.$$

The simplest example is the *forward Euler* method ($p = 1$). For (1.3) this yields

$$U_{n+1} = I + hB_n,$$

where $B_n = B(t_n, L_n)$. Therefore,

$$L_{n+1} = (I + hB_n)\, L_n\, (I + hB_n)^{-1}.$$

Thus, while the standard forward Euler for (1.1) yields the scheme

$$L_{n+1} = L_n + h(B_n L_n - L_n B_n),$$

its modified isospectral counterpart is

$$L_{n+1} = L_n + h(B_n L_n - L_{n+1} B_n). \tag{3.4}$$

The scheme is *semi-explicit* because although the matrix function B_n uses only past values of the function L, in order to find L_{n+1} we need to invert a matrix that depends only on past values of L and not on L_{n+1}. If we want to avoid matrix inversions, we can apply the functional iteration to the semi-explicit scheme (3.4). As for the MIMR, we get

$$\begin{aligned} L_{n+1}^{[0]} &= L_n, \\ L_{n+1}^{[k+1]} &= L_n + h\left(B_n L_n - L_{n+1}^{[k]} B_n\right), \qquad k = 0, 1, 2, \ldots, \end{aligned} \tag{3.5}$$

whereby the first step coincides with the classical forward Euler scheme. We remark that, as for the MIMR, the above scheme must be iterated to convergence, otherwise the numerical approximation need not be isospectral. Furthermore, some restrictions on the choice of the stepsize h might be necessary to force the convergence of (3.5).

All this can be generalized to higher order Taylor methods. Since

$$L_{n+1} U_{n+1} = U_{n+1} L_n$$

and U_{n+1} is given by the formula

$$U_{n+1} = I + hB_n + \frac{h^2}{2}(B_n' + B_n^2) + \cdots + \frac{h^p}{p!} F_{p,n},$$

where F_p has been defined in (3.3) and $F_{p,n} = F_p(t_n)$, we have

$$\begin{aligned} L_{n+1} &= L_n + h(B_n L_n - L_{n+1} B_n) + \frac{h^2}{2}[(B_n' + B_n^2)L_n - L_{n+1}(B_n' + B_n^2)] \\ &\quad + \cdots + \frac{h^p}{p!}(F_{p,n} L_n - L_{n+1} F_{p,n}), \end{aligned} \tag{3.6}$$

which is an isospectral semi-explicit modification of the underlying Taylor method when applied to (1.1). Needless to say, also for this scheme it is possible to use fixed point iteration to avoid matrix inversions, iterations being carried to convergence.

As we have already mentioned beforehand, we see that the schemes are generally non-symmetric. This can also be deduced directly from (3.4) and (3.6). We introduce the *symmetry error matrix*

$$E_n = \tfrac{1}{2}(L_n - L_n^{\mathrm{T}}), \qquad n \in \mathbb{Z}^+.$$

The above matrix will be used in order to study time-stepping strategies for decreasing the loss of symmetry while retaining isospectrality.

3.3 A strategy for decreasing E_n

Instead of writing the theoretical solution $L(t)$ of the differential equation (1.1) as in (1.2), we may wish to use

$$L(t_{n+1}) = [V(t_{n+1})]^{-1} L(t_n) V(t_{n+1}). \tag{3.7}$$

This follows at once from (1.2) by letting $V(t) = [U(t)]^{-1}$. Recalling that

$$[U^{-1}]' = -U^{-1}U'U^{-1},$$

we conclude that the matrix V satisfies the ordinary differential equation

$$V' = -VBUU^{-1} = -VB, \tag{3.8}$$

with the initial condition

$$V(t_n) = [U(t_n)]^{-1} = I.$$

It is convenient to commence with the following preliminary result.

Lemma 2 *Let $m \in \mathbb{Z}^+$ and*

$$P_{n+1} = U_{m(n+1)} \cdots U_{mn+2} U_{mn+1}, \tag{3.9}$$

where each U_s is an approximation to the solution of either (1.3) *or* (3.8) *at t_s. If*

$$P_{n+1} P_{n+1}^{\mathrm{T}} = I + \mathcal{O}\left(h^{q+1}\right), \qquad q \in \mathbb{Z}^+, \tag{3.10}$$

then it is true, for all $n \in \mathbb{Z}^+$, that

$$E_{mn} = \mathcal{O}(h^q).$$

Proof. Since

$$L_{m(n+1)} = P_{n+1} L_{mn} P_{n+1}^{-1}, \qquad n = 0, 1, \ldots,$$

we have

$$E_{m(n+1)} = \tfrac{1}{2}(P_{n+1} L_{mn} P_{n+1}^{-1} - P_{n+1}^{-T} L_{mn}^{\mathrm{T}} P_{n+1}^{\mathrm{T}}).$$

Thus, the symmetry error matrices obey the recurrence relation

$$\begin{aligned} P_{n+1}^{\mathrm{T}} E_{m(n+1)} P_{n+1} &= \tfrac{1}{2}\left[(P_{n+1}^{\mathrm{T}} P_{n+1}) L_{mn} - L_{mn}^{\mathrm{T}} (P_{n+1}^{\mathrm{T}} P_{n+1})\right] \\ &= \tfrac{1}{2}(L_{mn} - L_{mn}^{\mathrm{T}}) + \mathcal{O}\left(h^{q+1}\right) = E_{mn} + \mathcal{O}\left(h^{q+1}\right). \end{aligned}$$

Since $E_0 = O$, and $P_{n+1} = I + \mathcal{O}(h)$, it follows by induction that (3.10) implies $E_{mn} = \mathcal{O}(h^q)$ for all $n \in \mathbb{Z}^+$. □

Strategy 1: A first idea is to approximate L by alternating between (1.2) and (3.7). In other words,

$$\begin{aligned} L_{2n+1} &= U_{2n+1} L_{2n} U_{2n+1}^{-1}, \\ L_{2n+2} &= V_{2n+2}^{-1} L_{2n+1} V_{2n+2}, \end{aligned} \tag{3.11}$$

$m = 2$, and

$$P_n = V_{2n+2}^{-1} U_{2n+1}. \tag{3.12}$$

Theorem 3 *Strategy 1 yields*

$$q = p + 1,$$

in other words, we gain an extra order of symmetry for the same computational cost.

Proof. We are interested in investigating the order condition (3.10). Recalling the definition of the matrix P_n, (3.10) is equivalent to

$$U_{2n+1}^{\mathrm{T}}(V_{2n+2}V_{2n+2}^{\mathrm{T}})^{-1}U_{2n+1} = I + \mathcal{O}\left(h^{q+1}\right),$$

or, in other words, to

$$V_{2n+2}V_{2n+2}^{\mathrm{T}} - U_{2n+1}U_{2n+1}^{\mathrm{T}} = \mathcal{O}\left(h^{q+1}\right). \tag{3.13}$$

This, in fact, is the form suitable for our analysis. Since the local errors are

$$U_{2n+1} - U(t_{2n+1}) = -\frac{1}{(p+1)!}U^{(p+1)}(t_{2n})h^{p+1} + \mathcal{O}\left(h^{p+2}\right)$$

and

$$V_{2n+2} - V(t_{2n+2}) = -\frac{1}{(p+1)!}V^{(p+1)}(t_{2n+1})h^{p+1} + \mathcal{O}\left(h^{p+2}\right)$$

respectively, and the exact solutions U and V are unitary, we have

$$U_{2n+1}U_{2n+1}^{\mathrm{T}} = I - \frac{1}{(p+1)!}\left[U^{(p+1)}(t_{2n}) + [U^{(p+1)}(t_{2n})]^{\mathrm{T}}\right]h^{p+1} + \mathcal{O}\left(h^{p+2}\right)$$

and

$$V_{2n+2}V_{2n+2}^{\mathrm{T}} = I - \frac{1}{(p+1)!}\left[V^{(p+1)}(t_{2n+1}) + [V^{(p+1)}(t_{2n+1})]^{\mathrm{T}}\right]h^{p+1} + \mathcal{O}\left(h^{p+2}\right),$$

whereby we have taken into account that $U(t_{2n+1}), V(t_{2n+2}) = I + \mathcal{O}(h)$. Thus, substitution in (3.13) yields

$$\begin{aligned} V_{2n+2}V_{2n+2}^{\mathrm{T}} - U_{2n+1}U_{2n+1}^{\mathrm{T}} &= \frac{1}{(p+1)!}\left\{\left[U^{(p+1)}(t_{2n}) + [U^{(p+1)}(t_{2n})]^{\mathrm{T}}\right]\right. \\ &\quad \left. - \left[V^{(p+1)}(t_{2n+1}) + [V^{(p+1)}(t_{2n+1})]^{\mathrm{T}}\right]\right\}h^{p+1} + \mathcal{O}\left(h^{p+2}\right). \end{aligned}$$

Since both equations (1.3) and (3.7) are linear, there exist functions F_ℓ and G_ℓ, $0 \le \ell \le p+1$, such that

$$U^{(\ell)}(t) = F_\ell(t)U(t), \qquad V^{(\ell)}(t) = V(t)G_\ell(t).$$

Differentiation affirms at once that

$$\begin{aligned} F_{\ell+1} &= F_\ell' + F_\ell B, \\ G_{\ell+1} &= G_\ell' - BG_\ell. \end{aligned}$$

Therefore, due to the skew-symmetry of B,

$$G_{\ell+1}^{\mathrm{T}} = [G_\ell']^{\mathrm{T}} - G_\ell^{\mathrm{T}}B^{\mathrm{T}} = [G_\ell']^{\mathrm{T}} + G_\ell^{\mathrm{T}}B.$$

Furthermore $F_0 = G_0^{\mathrm{T}} = I$ and we thus deduce that the sequences $\{F_\ell\}$ and $\{G_\ell^{\mathrm{T}}\}$ coincide, since they satisfy identical recurrences and starting conditions. In other words

$$\begin{aligned}
&V_{2n+2}V_{2n+2}^{\mathrm{T}} - U_{2n+1}U_{2n+1}^{\mathrm{T}} \\
&= \frac{1}{(p+1)!}\left[(F_{p+1} + F_{p+1}^{\mathrm{T}}) - (F_{p+1} + F_{p+1}^{\mathrm{T}})\right] h^{p+1} + \mathcal{O}\Big(h^{p+2}\Big).
\end{aligned}$$

In fact, since $t_{2n+2} - t_{2n+1} = h$, we are allowed to evaluate all derivatives at the same point, e.g. at t_{2n+1}. From the above relation we see that the coefficient of h^{p+1} vanishes and the result follows. □

In other words, at the same computational cost, we can increase the order of symmetry from p to $p+1$, and $E_{2n} = \mathcal{O}(h^{p+1})$ instead of $E_{2n} = \mathcal{O}(h^p)$.

3.4 The three-step time advancing technique

Motivated by the result in Theorem 3, we consider the following three-step time advancing technique. Assuming without loss of generality $n = 0$, we consider the initial value problems

$$\begin{aligned}
U' &= BU, & U(0) &= I, \\
V' &= -VB, & V(\alpha h) &= I, \\
W' &= BW, & W(\beta h) &= I
\end{aligned} \tag{3.14}$$

and write the exact solution of (1.1) in the form

$$L(h) = W(h)[V(\beta h)]^{-1}U(\alpha h)L_0[U(\alpha h)]^{-1}V(\beta h)[W(h)]^{-1}. \tag{3.15}$$

Strategy 2: We seek values of α and β such that the approximation

$$L_1 = \tilde{W}\tilde{V}^{-1}\tilde{U}L_0\tilde{U}^{-1}\tilde{V}\tilde{W}^{-1}, \tag{3.16}$$

of (3.15) is 'good', namely $q \geq p+1$. The matrices W, V and U denote the theoretical values $W(h), V(\beta h)$ and $U(\alpha h)$ in (3.14) respectively. The tilde means that we are considering their approximations obtained with the Taylor method of order p.

Theorem 4 *Strategy 2 yields $q = 3$ for forward Euler with*

$$\alpha = 0.31313067266396\cdots, \qquad \beta = 0.72794049770602\cdots,$$

or

$$\alpha = 0.68686932733604\cdots, \qquad \beta = 1.59677745953869\cdots.$$

For any other Taylor method of order $p \geq 2$, $q = p+2$ is not attainable.

Proof. We write (3.16) as

$$L_1 = PL_0P^{-1},$$

the matrix P being

$$P = \tilde{W}\tilde{V}^{-1}\tilde{U}.$$

To simplify our notation, we let

$$C_U := \frac{1}{(p+1)!}U^{(p+1)}(0), \qquad D_U := \frac{1}{(p+2)!}U^{(p+2)}(0),$$
$$C_V := \frac{1}{(p+1)!}V^{(p+1)}(\alpha h), \qquad D_V := \frac{1}{(p+2)!}V^{(p+2)}(\alpha h),$$
$$C_W := \frac{1}{(p+1)!}W^{(p+1)}(\beta h), \qquad D_W := \frac{1}{(p+2)!}W^{(p+2)}(\beta h)$$

Therefore, since our method consists of a Taylor expansion truncated to order p, we have

$$\tilde{U}(\alpha h) = U(\alpha h) - \alpha^{p+1}C_U h^{p+1} - \alpha^{p+2}D_U h^{p+2} - \cdots, \tag{3.17}$$
$$\tilde{V}(\beta h) = V(\beta h) - (\beta-\alpha)^{p+1}C_V h^{p+1} - (\beta-\alpha)^{p+2}D_V h^{p+2} - \cdots, \tag{3.18}$$
$$\tilde{W}(h) = W(h) - (1-\beta)^{p+1}C_W h^{p+1} - (1-\beta)^{p+2}D_W h^{p+2} - \cdots. \tag{3.19}$$

Moreover we let Y be V^{-1}. Then Y obeys the ordinary differential equation

$$Y' = BY, \qquad Y(\alpha h) = I.$$

It follows that

$$\begin{aligned} Y(t) &= U(t)[U(\alpha h)]^{-1} \\ W(t) &= U(t)[U(\beta h)]^{-1}. \end{aligned} \tag{3.20}$$

We let $C = C_U$, $D = D_U$, $B_0 = B(0, L_0)$. Our aim is to render, by using (3.20), the other derivatives C_W, D_W, C_V and D_V in terms of C, D and B_0. We have

$$\begin{aligned} C_W &= \frac{1}{(p+1)!}U^{(p+1)}(\beta h)[U(\beta h)]^{-1} \\ &= \frac{1}{(p+1)!}\Big[U^{(p+1)}(0) + \beta h U^{(p+2)}(0)\Big](I - \beta h B_0) + \mathcal{O}\Big(h^2\Big) \\ &= C + \beta h\Big[(p+2)D - CB_0\Big] + \mathcal{O}\Big(h^2\Big), \\ D_W &= D + \mathcal{O}(h), \end{aligned}$$

and, by virtue of the orthogonality of U and V,

$$\begin{aligned} C_V &= \frac{1}{(p+1)!}U(\alpha h)[U^{(p+1)}(\alpha h)]^{\mathrm{T}} \\ &= \frac{1}{(p+1)!}(I + \alpha h B_0)\Big\{[U^{(p+1)}(0)]^{\mathrm{T}} + \alpha h [U^{(p+2)}(0)]^{\mathrm{T}}\Big\} + \mathcal{O}\Big(h^2\Big) \\ &= C^{\mathrm{T}} + \alpha h\Big[(p+2)D^{\mathrm{T}} + B_0 C^{\mathrm{T}}\Big] + \mathcal{O}\Big(h^2\Big), \\ D_V &= D^{\mathrm{T}} + \mathcal{O}(h). \end{aligned}$$

Furthermore, for the numerical approximation $\tilde{Y} = \tilde{Y}(\beta h)$, the relation $Y(\beta h) = I + (\beta - \alpha)hB_0 + \mathcal{O}(h^2)$ implies

$$\begin{aligned} \tilde{Y}(\beta h) &= [V(\beta h) - (\beta-\alpha)^{p+1}C_V h^{p+1} - (\beta-\alpha)^{p+2}D_V h^{p+2} - \cdots]^{-1} \\ &= [I + (\beta-\alpha)^{p+1}Y(\beta h)C_V h^{p+1} + (\beta-\alpha)^{p+2}Y(\beta h)D_V h^{p+2} + \cdots]Y(\beta h) \\ &= Y(\beta h) + (\beta-\alpha)^{p+1}Y(\beta h)C_V Y(\beta h)h^{p+1} + (\beta-\alpha)^{p+2}D_V h^{p+2} + \cdots, \end{aligned}$$

allowing us to write

$$\begin{aligned}
\tilde{U}(\alpha h) &= U(\alpha h) - \alpha^{p+1} C h^{p+1} - \alpha^{p+2} D h^{p+2} + \mathcal{O}\Big(h^{p+3}\Big)\,, \\
\tilde{Y}(\beta h) &= Y(\beta h) + (\beta-\alpha)^{p+1} C^{\mathrm{T}} h^{p+1} + \Big[(\beta-\alpha)^{p+1}\beta B_0 C^{\mathrm{T}} \\
&\quad + (\beta-\alpha)^{p+2} C^{\mathrm{T}} B_0 + (\beta-\alpha)^{p+1}(\alpha(p+1)+\beta) D^{\mathrm{T}}\Big] h^{p+2} + \mathcal{O}\Big(h^{p+3}\Big)\,, \\
\tilde{W}(h) &= W(h) - (1-\beta)^{p+1} C h^{p+1} + \Big[(1-\beta)^{p+1}\beta C B_0 \\
&\quad - (1-\beta)^{p+1}(\beta(p+1)+1) D\Big] h^{p+2} + \mathcal{O}\Big(h^{p+3}\Big)\,.
\end{aligned}$$

Finally, it is convenient to introduce the matrices

$$\begin{aligned}
\Phi &:= C + C^{\mathrm{T}}, \\
\Psi &:= D + D^{\mathrm{T}}, \\
\Theta &:= B_0 C^{\mathrm{T}} - C B_0, \\
\Lambda &:= C^{\mathrm{T}} B_0 - B_0 C.
\end{aligned}$$

Note that the above matrices are all symmetric and that Θ and Λ satisfy the identity

$$B_0 \Phi - \Phi B_0 = \Theta - \Lambda. \tag{3.21}$$

In particular it follows that if Φ and B_0 commute, then $\Theta = \Lambda$.

Using these matrices, we wish to estimate the departure from the identity matrix of

$$P P^{\mathrm{T}} = \tilde{W}\tilde{Y}\tilde{U}\tilde{U}^{\mathrm{T}}\tilde{Y}^{\mathrm{T}}\tilde{W}^{\mathrm{T}}.$$

To this end, our first step consists of evaluating

$$\begin{aligned}
\tilde{U}\tilde{U}^{\mathrm{T}} &= \Big[U(\alpha h) - \alpha^{p+1} C h^{p+1} - \alpha^{p+2} D h^{p+2} + \cdots\Big] \cdot \Big[[U(\alpha h)]^{-1} - \alpha^{p+1} C^{\mathrm{T}} h^{p+1} \\
&\quad - \alpha^{p+2} D^{\mathrm{T}} h^{p+2} + \cdots\Big] \\
&= I - \alpha^{p+1}\Phi h^{p+1} - \alpha^{p+2}(\Psi + \Theta) h^{p+2} + \mathcal{O}\Big(h^{p+3}\Big)\,.
\end{aligned}$$

Next, using the identity (3.21), skew-symmetry of the matrix function B, and the first order expansion $Y(\beta h) = I + (\beta - \alpha) h B_0 + \mathcal{O}(h^2)$, we obtain

$$\begin{aligned}
\tilde{Y}\tilde{U}\tilde{U}^{\mathrm{T}}\tilde{Y}^{\mathrm{T}} &= I + [(\beta-\alpha)^{p+1} - \alpha^{p+1}]\Phi h^{p+1} + \{\beta[(\beta-\alpha)^{p+1} - \alpha^{p+1}]\Theta \\
&\quad + \alpha^{p+1}(\beta-\alpha)\Lambda + [(\beta-\alpha)^{p+1}(\alpha(p+1)+\beta) - \alpha^{p+2}]\Psi\} h^{p+2} + \mathcal{O}\Big(h^{p+3}\Big)\,.
\end{aligned}$$

Finally, similar calculation results in an expansion of PP^{T}, namely

$$\begin{aligned}
PP^{\mathrm{T}} &= I + [(\beta-\alpha)^{p+1} - \alpha^{p+1} - (1-\beta)^{p+1}]\Phi h^{p+1} \\
&\quad + \Big\{[(\beta-\alpha)^{p+1} - \alpha^{p+1} - (1-\beta)^{p+1}]\Theta \\
&\quad + [(1-\alpha)\alpha^{p+1} - (1-\beta)(\beta-\alpha)^{p+1}]\Lambda \\
&\quad + [(\alpha(p+1)+\beta)(\beta-\alpha)^{p+1} - \alpha^{p+2} - (\beta(p+1)+1)(1-\beta)^{p+1}]\Psi\Big\} h^{p+2} \\
&\quad + \mathcal{O}\Big(h^{p+3}\Big)\,.
\end{aligned} \tag{3.22}$$

Before proceeding further, we observe that the coefficients of Φh^{p+1} and Θh^{p+2} coincide, therefore, if we let the first equal zero, automatically so does the second.

Let us first consider the simplest case, e.g. when $p=1$, corresponding to the forward Euler scheme. We have

$$\begin{aligned}\Phi &= B_0^2, \\ \Psi &= \frac{1}{2}(B_0 B_0' + B_0' B_0).\end{aligned}$$

Furthermore, Φ is a polynomial of B_0 and the two matrices commute. It follows from (3.21) that $\Theta = \Lambda$ and

$$\Theta = \Lambda = -\Psi.$$

Thus, in order to annihilate the coefficients of h^{p+1} and h^{p+2} in (3.22), it is necessary and sufficient that α and β satisfy the equations

$$(\beta-\alpha)^2 = \alpha^2 + (1-\beta)^2, \tag{3.23}$$

$$\alpha(\beta-\alpha)^2 = \beta(1-\beta)^2. \tag{3.24}$$

We deduce from (3.23) that

$$\beta = \frac{1}{2(1-\alpha)}$$

and substitution of the latter in (3.24) results in the equation

$$8\alpha^6 - 24\alpha^5 + 32\alpha^4 - 24\alpha^3 + 14\alpha^2 - 6\alpha + 1 = 0. \tag{3.25}$$

We let $x = \left(\alpha - \frac{1}{2}\right)^2$. Substitution in (3.25) yields the cubic equation

$$8x^3 + 2x^2 + \tfrac{7}{2}x - \tfrac{1}{8} = 0.$$

Using the Cardano formulae (cf. Abramowitz and Stegun, 1965) we can find its roots and verify that only one of them is real. Backward substitution gives the following values for α,

$$\alpha = \frac{1}{2} \pm \left\{ \frac{1}{6}\left[\left(\frac{3\sqrt{69}+11}{2}\right)^{1/3} - \left(\frac{3\sqrt{69}-11}{2}\right)^{1/3} - \frac{1}{2}\right]\right\}^{1/2}.$$

Consequently, there are two choices of real (α, β), namely

$$\alpha = 0.31313067266396\cdots, \qquad \beta = 0.72794049770602\cdots,$$

and

$$\alpha = 0.68686932733604\cdots, \qquad \beta = 1.59677745953869\cdots.$$

Even if both choices are in principle equally good, the second means that we are getting from $t_0 = 0$ to $t_1 = h$ stepping twice forward and then once backward. Both choices ensure, by virtue of (3.10), that the symmetry error matrix E_n is of order $\mathcal{O}(h^3)$ instead of $\mathcal{O}(h)$ as we might expect from a numerical method of order one.

In the case $p \geq 2$, the analysis becomes technically more difficult. We require the following equations to be satisfied by α and β:

$$\begin{aligned}
\alpha^{p+1} + (1-\beta)^{p+1} &= (\beta-\alpha)^{p+1}, & (3.26)\\
(1-\alpha)\alpha^{p+1} &= (1-\beta)(\beta-\alpha)^{p+1}, & (3.27)\\
[\alpha(p+1)+\beta](\beta-\alpha)^{p+1} &= \alpha^{p+2} + [\beta(p+1)+1](1-\beta)^{p+1}. & (3.28)
\end{aligned}$$

The relations (3.26)–(3.28) are sufficient conditions to get a symmetry order $p+2$ instead of $p+1$ as in the previous section. In order to prove that they are also necessary, we wish to show that the matrices Ψ, Λ are in general linearly independent. To prove this, we show that there exists a skew-symmetric function B for which it is true that

$$\gamma\Psi + \delta\Lambda = O \qquad \Rightarrow \qquad \gamma = \delta = 0.$$

To this end, let $B(t) = B_0 + B_1 t^{p-1} + B_2 t^p$ (needless to say, B_0, B_1 and B_2 are skew-symmetric). We consider the Taylor expansion of U, namely

$$U(t) = \sum_{\ell=0}^{\infty} \frac{F_{\ell,0}}{\ell!} t^\ell.$$

and compare

$$U'(t) = \sum_{\ell=0}^{\infty} \frac{F_{\ell+1,0}}{\ell!} t^\ell \tag{3.29}$$

with

$$\begin{aligned}
U'(t) &= (B_0 + B_1 t^{p-1} + B_2 t^p)U & (3.30)\\
&= B_0 \sum_{l=0}^{\infty} \frac{F_{\ell,0}}{\ell!} t^\ell + B_1 \sum_{\ell=p-1}^{\infty} \frac{F_{\ell-p+1,0}}{(\ell-p+1)!} t^\ell + B_2 \sum_{\ell=p}^{\infty} \frac{F_{\ell-p,0}}{(\ell-p)!} t^\ell.
\end{aligned}$$

For simplicity sake, since all functions are evaluated at $t=0$, we drop the second subscript. Equating identical powers of t, we deduce the recurrence relation

$$\frac{F_{\ell+1}}{(\ell+1)!} = \frac{1}{(\ell+1)} \left[B_0 \frac{F_\ell}{\ell!} + B_1 \frac{F_{\ell-p+1}}{(\ell-p+1)!} + B_2 \frac{F_{\ell-p}}{(\ell-p)!} \right], \quad \ell = p, p+1, \ldots, \tag{3.31}$$

with the starting conditions

$$\begin{aligned}
F_k &= B_0^k, \quad k = 0, \ldots, p-1,\\
F_p &= B_0^p + (p-1)! B_1.
\end{aligned}$$

In order to evaluate the matrices $\Psi \equiv \Psi_p = D + D^{\mathrm{T}}$ and $\Lambda \equiv \Lambda_p = C^{\mathrm{T}} B_0 - B_0 C$, we need to calculate

$$D = \frac{F_{p+2}}{(p+2)!}$$

and

$$C = \frac{F_{p+1}}{(p+1)!}.$$

Using (3.31) to find F_{p+1} and F_{p+2}, we obtain

$$\begin{aligned} F_{p+1} &= B_0^{p+1} + (p-1)!B_0B_1 + p!B_1B_0 + p!B_2, \\ F_{p+2} &= B_0^{p+2} + (p-1)!B_0^2B_1 + p!B_0B_1B_0 + \frac{1}{2}(p+1)!B_1B_0^2 + p!B_0B_2 + (p+1)!B_2B_0. \end{aligned}$$

Therefore

$$\Psi_p = \frac{1}{(p+2)!}[1+(-1)^p]B_0^{p+2} + \frac{p-1}{2p(p+1)}[B_1, B_0^2] + \frac{1}{p+1}(B_0B_2 + B_2B_0),$$

while

$$\Lambda_p = -\frac{1}{(p+1)!}[1+(-1)^p]B_0^{p+2} + \frac{p-1}{2p(p+1)}[B_1, B_0^2] - \frac{1}{p+1}(B_0B_2 + B_2B_0).$$

When p is even, we let $B_2 = O$, hence

$$\begin{aligned} \Psi_p &= \frac{1}{(p+2)!}[1+(-1)^p]B_0^{p+2} + \frac{p-1}{2p(p+1)}[B_1, B_0^2], \\ \Lambda_p &= -\frac{1}{(p+1)!}[1+(-1)^p]B_0^{p+2} + \frac{p-1}{2p(p+1)}[B_1, B_0^2]. \end{aligned}$$

We choose

$$B_0 = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad B_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}. \tag{3.32}$$

Note that $B_0^{2m+2} = (-1)^m B_0^2, m = 0, 1, \ldots$. It follows that B_0^{p+2} and $[B_1, B_0^2]$ have the sparsity structure

$$B_0^{p+2} = \begin{bmatrix} \times & 0 & 0 \\ 0 & \times & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad [B_1, B_0^2] = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \times \\ 0 & \times & 0 \end{bmatrix}, \qquad p \text{ even},$$

respectively. Here, the symbol '$\times$' denotes a nonzero entry. It follows that B_0^{p+2} is not a scalar multiple of $[B_1, B_0^2]$, and, since

$$\det \begin{bmatrix} \frac{2}{(p+2)!} & \frac{p-1}{2p(p+1)} \\ \frac{-2}{(p+1)!} & \frac{p-1}{2p(p+1)} \end{bmatrix} = \frac{(p-1)}{p(p+1)(p+2)!} \neq 0,$$

we deduce linear independence of Ψ_p and Λ_p.

For odd values of $p \geq 3$,

$$\begin{aligned} \Psi_p &= \frac{p-1}{2p(p+1)}[B_1, B_0^2] + \frac{1}{p+1}(B_0B_2 + B_2B_0), \\ \Lambda_p &= \frac{p-1}{2p(p+1)}[B_1, B_0^2] - \frac{1}{p+1}(B_0B_2 + B_2B_0), \end{aligned}$$

and we choose $B_2 = B_0$, whereby B_0 and B_1 are the same as in (3.32). Once again, since

$$\det \begin{bmatrix} \frac{p-1}{2p(p+1)} & \frac{1}{(p+1)} \\ \frac{p-1}{2p(p+1)} & -\frac{1}{(p+1)} \end{bmatrix} = -\frac{(p-1)}{p(p+1)^2} \neq 0,$$

we deduce linear independence.

We conclude that the conditions (3.26)–(3.28) are also necessary for (α, β) if we require a symmetry error matrix

$$E_n = \mathcal{O}\left(h^{p+2}\right).$$

A number of remarks with regard to (3.26)–(3.28):

1. Consider the choices
$$\alpha, \beta \in \{0, 1\}.$$
These correspond to using exclusively $U' = BU$ or $V' = -VB$. Since these parameters do no satisfy (3.26), it follows that it is not possible to increase the order in symmetry without alternating between the solution of the two associated orthogonal flows.

2. Consider the choices
$$\alpha = \frac{1}{2}, \quad \beta = 1, \qquad \text{and} \qquad \alpha = 0, \quad \beta = \frac{1}{2}.$$
Then (3.26) is satisfied but not (3.27) or (3.28). This means that the h^{p+1} term in PP^{T} vanishes but not the h^{p+2} term. These choices of (α, β) are in effect the ones presented in §3.3.

We therefore exclude the aforementioned values of α and β. From (3.26) we substitute $(\beta - \alpha)^{p+1}$ in (3.27) to find that

$$(\beta - \alpha)^{p+1} = \frac{1-\alpha}{1-\beta}\alpha^{p+1}.$$

Substituting the latter in (3.26) we have

$$(1-\beta)^{p+1} = \frac{\beta - \alpha}{1-\beta}\alpha^{p+1}.$$

Thus, we can substitute both $(\beta - \alpha)^{p+1}$ and $(1-\beta)^{p+1}$ in (3.28) to find

$$[\alpha(p+1) + \beta]\frac{1-\alpha}{1-\beta}\alpha^{p+1} = \alpha^{p+2} + [\beta(p+1) + 1]\frac{\beta-\alpha}{1-\beta}\alpha^{p+1}.$$

We remove the multiple root $\alpha = 0$ and simple algebraic manipulation of the above equation produces

$$(\beta - \alpha)^2 = \alpha(1-\beta). \tag{3.33}$$

Let

$$\begin{aligned} a &= \beta - \alpha, \\ b &= \alpha, \\ c &= 1 - \beta. \end{aligned}$$

Then, (3.33) is equivalent to

$$a^2 = bc,$$

from which we deduce that $bc \geq 0$. Furthermore, with these new variables, (3.26) is given by

$$a^{p+1} = b^{p+1} + c^{p+1},$$

therefore, substituting the value of a, we find

$$(bc)^{\frac{p+1}{2}} = b^{p+1} + c^{p+1}.$$

Squaring both sides of the above relation, we deduce that

$$b^{2(p+1)} + (bc)^{p+1} + c^{2(p+1)} = 0,$$

an equation with the unique real solution

$$b = c = 0,$$

or, in the language of α and β,

$$\alpha = 0, \qquad \beta = 1,$$

which contradicts our assumption that $\alpha \neq 0$ and $\beta \neq 1$. We therefore conclude that (3.26)–(3.28) are incompatible.

□

3.5 On the condition number κ

As already observed in §1, when we solve (1.3) with a non-orthogonal method, at each step we have the 2-norm inequality

$$\|L_n\|_2 \leq \|L_{n+1}\|_2 \leq \kappa_{U_{n+1}} \|L_n\|_2, \qquad \kappa_{U_{n+1}} = \|U_{n+1}\|_2 \|U_{n+1}^{-1}\|_2. \tag{3.34}$$

The number $\kappa_{U_{n+1}}$ is the *spectral condition number* of the matrix U_{n+1}. In other words, (3.34) means that, although the numerical approximation L_{n+1} obtained with (1.2) is isospectrally similar to L_n, its 2-norm is allowed to increase with n and this may cause numerical instability for large values of n.

We recall that, for a normal nonsingular $d \times d$ matrix A,

$$\kappa_A = \frac{|\lambda_d|}{|\lambda_1|}, \tag{3.35}$$

where the eigenvalues $\lambda_1, \ldots, \lambda_d$ of A are ordered in such a way that $|\lambda_1| \le |\lambda_2| \le \cdots \le |\lambda_d|$. When A is not normal, we need to replace (3.35) with

$$\kappa_A = \frac{|\sigma_d|}{|\sigma_1|}, \tag{3.36}$$

where $\sigma_1, \ldots, \sigma_d$ are the singular values of A, sorted accordingly to increasing size.

Supposing that (1.1) is integrated with the time-stepping technique (3.11), an obvious bound is given by

$$\|L_{2n+2}\|_2 \le \kappa_{U_{2n+1}} \kappa_{V_{2n+2}^{-1}} \|L_{2n}\|_2,$$

where

$$\begin{aligned} \kappa_{U_{2n+1}} &= 1 + \mathcal{O}\left(h^{p+1}\right), \\ \kappa_{V_{2n+2}^{-1}} &= 1 + \mathcal{O}\left(h^{p+1}\right), \end{aligned}$$

therefore

$$\kappa_{U_{2n+1}} \kappa_{V_{2n+2}^{-1}} = 1 + \mathcal{O}\left(h^{p+1}\right).$$

However, the following result holds.

Theorem 5 *Using the time-stepping technique* (3.11), *it is true that*

$$\|L_{2n+2}\|_2 \le \kappa \|L_{2n}\|_2,$$

where

$$\kappa = 1 + \mathcal{O}\left(h^{p+2}\right),$$

while the exclusive use of either (1.2) *or* (3.7) *yields*

$$\kappa = 1 + \mathcal{O}\left(h^{p+1}\right).$$

Proof. We first consider the case $p = 1$, corresponding to the forward Euler method. We integrate the $(2n+1)$-st step with

$$U_{2n+1} = I + hB_{2n},$$

obtaining

$$L_{2n+1} = U_{2n+1} L_{2n} U_{2n+1}^{-1}.$$

Then we evaluate

$$\begin{aligned} U_{2n+2} &= I + hB_{2n+1}, \\ V_{2n+2} &= I - hB_{2n+1}, \end{aligned}$$

to obtain, at the $(2n+2)$-nd step,

$$\begin{aligned} L_{2n+2} &= U_{2n+2} U_{2n+1} L_{2n} U_{2n+1}^{-1} U_{2n+2}^{-1}, \\ \tilde{L}_{2n+2} &= V_{2n+2}^{-1} U_{2n+1} L_{2n} U_{2n+1}^{-1} V_{2n+2}. \end{aligned}$$

We claim that

$$\kappa_{V_{2n+2}^{-1}U_{2n+1}} = 1 + \mathcal{O}\Big(h^3\Big)\,,$$

while

$$\kappa_{U_{2n+2}U_{2n+1}} = 1 + \mathcal{O}\Big(h^2\Big)\,.$$

To this end, we observe that

$$\begin{aligned} U_{2n+2}U_{2n+1} &= I + h(B_{2n+1} + B_{2n}) + h^2 B_{2n+1}B_{2n}, \\ V_{2n+2}^{-1}U_{2n+1} &= (I - hB_{2n+1})^{-1}(I + hB_{2n}). \end{aligned}$$

Although the individual matrices $U_{2n+1}, U_{2n+2}, V_{2n+2}^{-1}$ are normal, their product is generally not normal, therefore, in order to evaluate the condition numbers, we need to use (3.36). Recalling that $B_{2n+1} = B_{2n} + \mathcal{O}(h)$, easy expansions yield

$$\begin{aligned} (U_{2n+2}U_{2n+1})^{\mathrm{T}}(U_{2n+2}U_{2n+1}) &= I - h^2(B_{2n+1}^2 + B_{2n}^2) + \mathcal{O}\Big(h^3\Big) \\ (V_{2n+2}^{-1}U_{2n+1})^{\mathrm{T}}(V_{2n+2}^{-1}U_{2n+1}) &= I + h^2(B_{2n+1}^2 - B_{2n}^2) + \mathcal{O}\Big(h^3\Big) = I + \mathcal{O}\Big(h^3\Big)\,. \end{aligned}$$

It follows that the singular values are given by

$$\begin{aligned} \sigma_i(U_{2n+2}U_{2n+1}) &= 1 + \mathcal{O}\Big(h^2\Big)\,, \\ \sigma_i(V_{2n+2}^{-1}U_{2n+1}) &= 1 + \mathcal{O}\Big(h^3\Big)\,, \qquad i = 1, \ldots, d \end{aligned}$$

and the proof follows for $p = 1$. Generalization to $p \geq 2$ is just as easy. Therefore, the time-stepping technique (3.11) reduces the departure from orthogonality of the matrix $V_{2n+2}^{-1}U_{2n+1}$ approximating the theoretical orthogonal matrix $U(t_{2n+2})U(t_{2n+1})$ from $\mathcal{O}(h^{p+1})$ to order $\mathcal{O}(h^{p+2})$. $\square$

4 Examples and numerical results

As an example, we present the results obtained with the forward Euler method. We have integrated a symmetric[5] QR flow with the semi-explicit isospectral forward Euler (continuous line), then using the time-stepping strategy (3.11) (dashed line) and finally with the three step strategy as explained in §3.4 (dash-dotted line). In Figure 1 we compare the error in symmetry obtained for the three strategies, plotting

$$\|E_n\|_2 = \tfrac{1}{2}\|L_n - L_n^{\mathrm{T}}\|_2.$$

We have used the same 10×10 randomly generated symmetric matrix L_0 and stepsize $h = \frac{1}{100}$. The circles denote the results obtained for the two and three time-stepping strategies after two and three steps respectively. These have to be compared with the solution obtained using the isospectral forward Euler scheme.

In order to evaluate the order of the method, we use the following technique. Assuming that the method is of order p, namely that the local error behaves as h^{p+1}, we integrate

[5] The theoretical solution is symmetric but the numerical solution is likely to lose symmetry

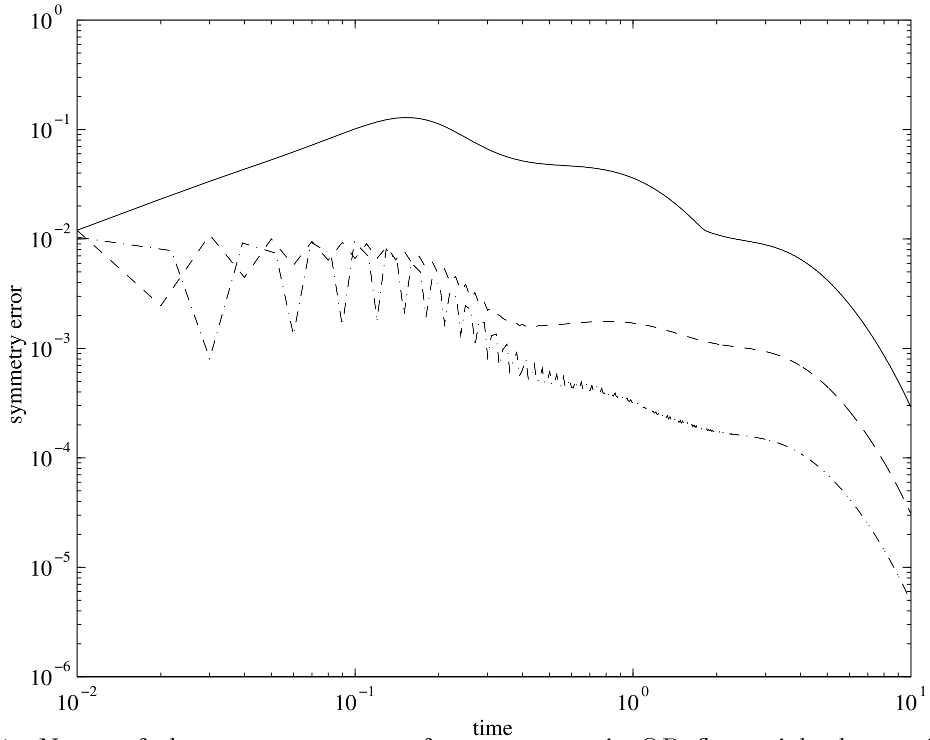

Figure 1: Norm of the symmetry error for a symmetric QR flow with the semi-explicit forward Euler (continuous line), with the time-stepping strategy (3.11) (dashed line) and the three step strategy as in §3.4 (dash-dotted line). Stepsize of integration is $h = \frac{1}{100}$. The matrix function B is as in (4.2).

$h - h/2$	FE	FE1	FE2
$1/256 - 1/512$	1.0161	2.0173	2.1658
$1/512 - 1/1024$	1.0080	2.0089	2.1605
$1/1024 - 1/2048$	1.0040	2.0045	2.0141

Table 1: Sequences converging to the order for forward Euler (FE), forward Euler implemented with strategy 1 (FE1) and forward Euler implemented with strategy 2 (FE2) for the matrix function B in (4.2).

$h - h/2$	FE	FE1	FE2
$1/256 - 1/512$	1.0009	2.0068	3.0038
$1/512 - 1/1024$	1.0005	2.0034	3.0026
$1/1024 - 1/2048$	1.0002	2.0017	2.9918

Table 2: Sequences converging to the order for forward Euler (FE), forward Euler implemented with strategy 1 (FE1) and forward Euler implemented with strategy 2 (FE2) for the matrix function B in (4.3).

in the interval $[0, T]$, hence the global error behaves as h^p. We integrate again halving the stepsize. Therefore, letting s_h be the norm of the symmetry error matrix at T for stepsize h, we have

$$p \approx -\frac{\log(s_{h/2}) - \log(s_h)}{\log 2}. \tag{4.1}$$

Letting $T = 2$ and choosing $h = \frac{1}{256}, \frac{1}{512}, \frac{1}{1024}$ and $\frac{1}{2048}$, we obtain Table 1. It is straightforward to observe that the sequences converge to one for forward Euler (FE), to two for forward Euler strategy 1 (FE1), and to two for forward Euler strategy 2 (FE2). The last result is in fact in contrast with what we have derived in §3.4. However, recall that throughout the paper we have assumed that the function B is skew-symmetric. In these numerical experiments, the matrix B has the form

$$B(L) = L_- - L_+, \tag{4.2}$$

therefore is no longer skew-symmetric when L is not symmetric. In fact, the result for strategy 2 still holds, but the error in symmetry affects the next term in the expansion and the theoretical result obtained for strategy 3 is no longer valid. Note, however, that strategy 2 still produces a better approximation at the same numerical cost as the standard isospectral forward Euler.

Next we consider the skew-symmetric matrix function

$$B(L) = \tfrac{1}{2}\Big[(L + L^{\mathrm{T}})_- - (L + L^{\mathrm{T}})_+\Big]. \tag{4.3}$$

When L is symmetric, this function coincides with $B(L) = L_- - L_+$. We evaluate again the order to find Table 2.

Table 2 confirms the theoretical results presented in §3.3 and §3.4, namely that we gain one and two orders of symmetry respectively. However, the error in symmetry behaves as in Figure 2, which is theoretically correct, but not necessarily better than the result we have obtained in Figure 1. The explanation is to be found in the fact that, as t tends to infinity, the matrix L tends to a diagonal matrix and, the flow being isospectral, it 'auto-corrects' itself. When we introduce the perturbation in (4.3) this is no longer valid for the numerical solution. We have also tried with an 'average' expression for function B. In this example,

$$B(L) = \begin{cases} \frac{1}{2}\Big((L + L^{\mathrm{T}})_- - (L + L^{\mathrm{T}})_+\Big), & \frac{1}{2}\|L - L^{\mathrm{T}}\| > 10^{-2}, \\ L_- - L_+, & \text{otherwise.} \end{cases} \tag{4.4}$$

$h - h/2$	FE	FE1	FE2
1/256 – 1/512	0.9908	2.0173	2.1658
1/512 – 1/1024	0.9603	2.0089	2.1605
1/1024 – 1/2048	5.1645	2.0045	2.0141

Table 3: Sequences converging to the order for forward Euler (FE), forward Euler implemented with strategy 1 (FE1) and forward Euler implemented with strategy 2 (FE2) for the matrix function B in (4.4).

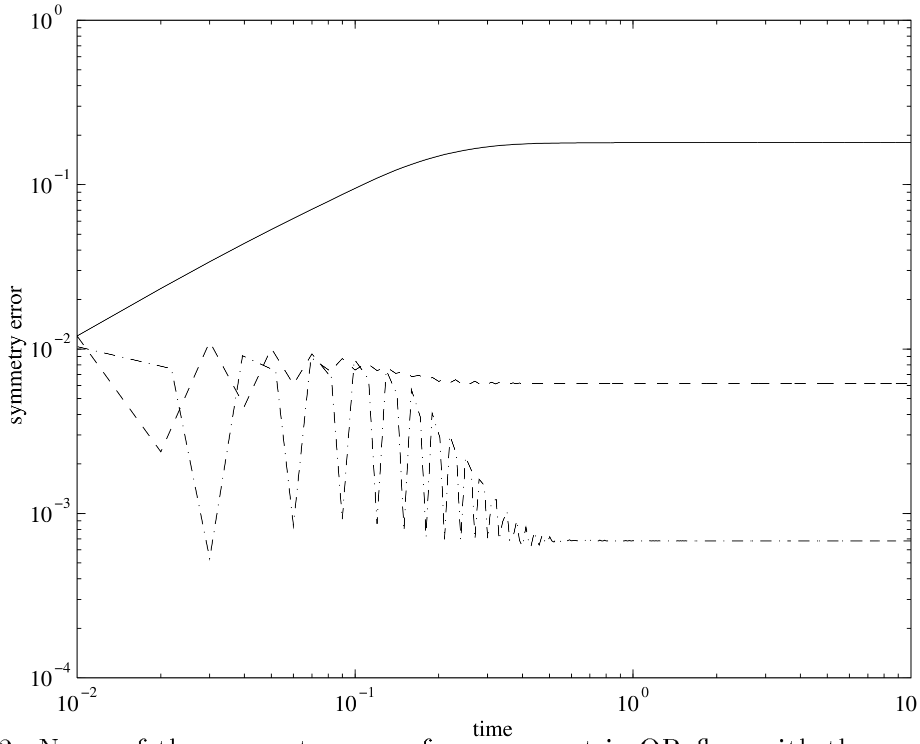

Figure 2: Norm of the symmetry error for a symmetric QR flow with the semi-explicit forward Euler (continuous line), with the time-stepping strategy (3.11) (dashed line) and the three step strategy as in §3.4 (dash-dotted line). Stepsize of integration is $h = \frac{1}{100}$. The matrix function B is as in (4.3).

The results are presented in Table 3 and the error in symmetry for the given example is in Figure 3.

Note that, although its formal order reduces to two, the forward Euler method implemented with strategy 2 produces a considerably smaller error in symmetry by using this choice of $B(L)$.

The last result for forward Euler in Table 3 is not surprising. Assume that we consider the qualitative behaviour of the symmetry breaking in function of the parameter $\lambda \in \mathbb{R}^+$ in

$$B(L) = \begin{cases} \frac{1}{2}\Big[(L+L^{\mathrm{T}})_- - (L+L^{\mathrm{T}})_+\Big], & \frac{1}{2}\|L-L^{\mathrm{T}}\| > 10^{-\lambda}, \\ L_- - L_+, & \text{otherwise.} \end{cases} \tag{4.5}$$

For the value of the parameter $\lambda = 2$ as in (4.4), there is a change of behaviour for a stepsize between 1/1024 and 1/2048. In fact, when $t \approx .01$, only with the second stepsize the numerical solution recovers the 'auto-correcting' mode of the true solution. For the same value of λ, the other two strategies recover the correct behaviour for smaller values of the stepsize.

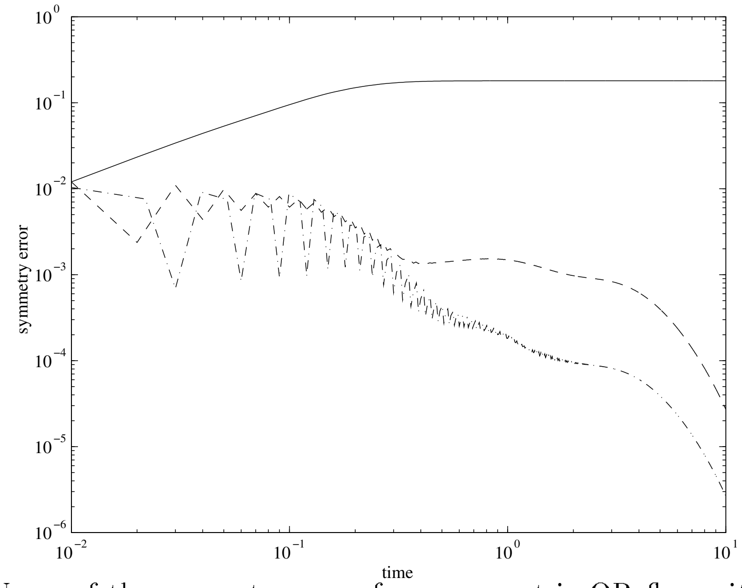

Figure 3: Norm of the symmetry error for a symmetric QR flow with the semi-explicit forward Euler (continuous line), with the time-stepping strategy (3.11) (dashed line) and the three step strategy as in §3.4 (dash-dotted line). Stepsize of integration is $h = \frac{1}{100}$. The matrix function B is as in (4.4).

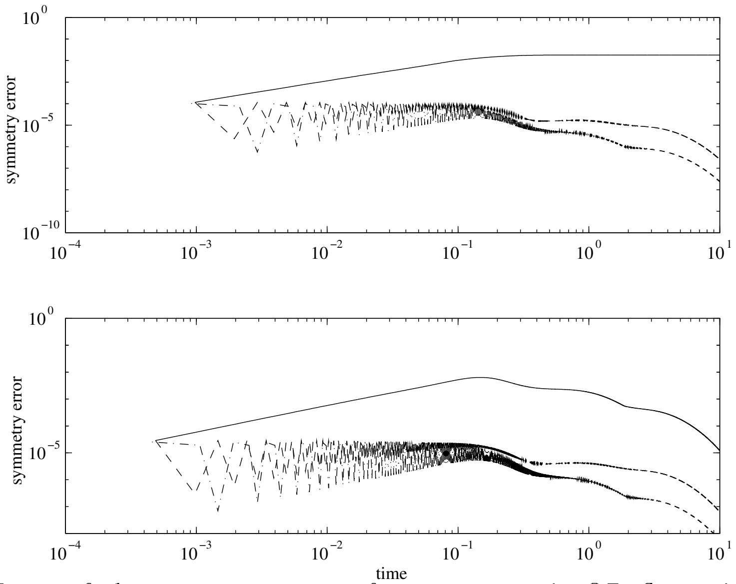

Figure 4: Norm of the symmetry error for a symmetric QR flow with the semi-explicit forward Euler (continuous line), with the time-stepping strategy (3.11) (dashed line) and the three step strategy as in §3.4 (dash-dotted line). Stepsizes of integration are $h = \frac{1}{2048}$ and $h = \frac{1}{1024}$ respectively. The matrix function B is as in (4.4).

References

Abramowitz, M., and Stegun, I.A. (1965), *Handbook of Mathematical Functions,* Dover, New York.

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

Calvo, M.P., Iserles, A. and Zanna, A. (1995), "Numerical solution of isospectral flows", DAMTP Tech. Rep. 1995/NA03, to appear in *Math. Comp.*

Chu, M.T. (1994), "A list of matrix flows with applications", in *Hamiltonian and Gradient Flows, Algorithms and Control* (A. Bloch, ed.), Fields Institute Communications, American Mathematical Society, Providence RI, 87–97.

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

Iserles A. and Zanna A. (1995), " Qualitative numerical analysis of ordinary differential equations", in *Lectures in Applied Mathematics* (J. Renegar, M. Shub & S. Smale, eds), *Lectures in Applied Maths* **32**, American Mathematical Society, Providence RI, 421–442.

Perelomov, A.M. (1990), *Integrable Systems of Classical Mechanics and Lie Algebras,* vol. 1, Birkhäuser Verlag, Basel.

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