The numerical integration of relative equilibrium solutions. Geometric theory

The numerical integration of relative equilibrium solutions. Geometric theory

A Durán[†] and J M Sanz-Serna[‡]

Departamento de Matemática Aplicada y Computación, Universidad de Valladolid, Valladolid, Spain

Received 17 February 1998
Recommended by R Krasny

Abstract. We study the propagation of errors in the numerical integration of relative equilibria solutions of differential equations with symmetries. In the Hamiltonian case and for stable equilibria, the error growth is typically quadratic for 'general' schemes and linear for schemes that preserve the invariant quantities of the problem. Numerical results are presented.

AMS classification scheme numbers: 65L05, 70H05

1. Introduction

The purpose of this paper is to study the propagation of errors in the numerical integration of relative equilibria. In recent years there has been a growing interest in the construction and analysis of so-called geometric integrators [15], i.e. of numerical methods for the integration of initial-value problems that take into account geometric properties of the system of differential equations under consideration. The most extensively researched case is that of symplectic integrators for Hamiltonian problems [11, 16]. Geometric integrators may be superior not only from a qualitative point of view but also quantitatively; some cases studied in the literature include periodic orbits [2–5, 8] and integrable systems [2, 10]. A reference of particular relevance to us is [9], that deals with the time-integration of travelling wave solutions of partial differential equations. The study in [9] uses the Korteweg–de Vries equation as a model problem: it is shown that while for 'general' schemes the error growth is quadratic, conservative schemes exhibit linear error growth. The results of [9] are based on an investigation of the direction in phase space of the local error: conservative methods align their local errors along directions that are not very harmful. The analysis of [9] does not identify the geometric mechanisms that lead to favourable error propagation in geometric integrators; the reader may be left with the impression that the cancellations obtained are a matter of luck and would only hold for the specific case of the Korteweg–de Vries equation and travelling wave solutions. In this paper and in the subsequent article [7] (both based on the thesis [6]) we show that the situation investigated in [9] is not exceptional and typically holds for stable relative equilibria of Hamiltonian problems, including travelling wave solutions of Hamiltonian partial differential equations. Earlier studies of error growth in the simulation of integrable systems and periodic orbits are also subsumed in the geometric

[†] E-mail: angel@mac.cie.uva.es
[‡] E-mail: sanzserna@cpd.uva.es

0951-7715/98/061547+21$19.50

framework considered in this paper. Our framework clearly identifies the geometric reasons that underlie the ‘success’ of geometric integrators in many applications. This paper deals with the more geometric aspects of the matter and, to avoid technicalities, treats the case of ordinary differential equations. In [7] we shall consider travelling wave solutions. The main conclusion is that, for a wide class of equations, the error in the integration of a relative equilibrium solution grows quadratically for ‘general’ methods and only linearly for ‘conservative methods’. While the mathematical proofs are restricted to relative equilibria, the conclusions of the study on the benefits of conservative schemes appear experimentally to be valid for other classes of solutions (see section 4 below).

This paper is structured as follows. Section 2 summarizes definitions and basic results about differential equations with symmetries and their relative equilibria [1, 13, 14]. In order to be accessible to as wide a readership as possible, we have tried to present a self-contained approach that can be read without a deep geometric background. The main results are presented in section 3 and section 4 contains some numerical illustrations.

2. Relative equilibria

2.1. Vector fields and flows

We consider vector fields g, defined in a given domain Ω of $\mathcal{R}^D$, and the corresponding differential systems

$$\dot{u} = g(u). \tag{1}$$

For each t, the phase flow $\varphi_{t,g}$ is a mapping in $\mathcal{R}^D$ such that $\varphi_{t,g}(u)$ is the value at time t of the solution of (1) that at time 0 takes the initial value u. For simplicity, we assume that for each $t \in \mathcal{R}$ the domain of $\varphi_{t,g}$ is the whole Ω, i.e. that all solutions of (1) exist for all $t \in (-\infty, \infty)$. As t varies, the mappings $\varphi_{t,g}$ form a one-parameter group of diffeomorphisms $\varphi_{t,g} : \Omega \to \Omega$,

$$\varphi_{t+s,g} = \varphi_{t,g} \circ \varphi_{s,g}, \qquad \varphi_{0,g} = \mathrm{Id},$$

whose phase velocity field is g,

$$\frac{\mathrm{d}}{\mathrm{d}t}\varphi_{t,g}(u) = g(\varphi_{t,g}(u)), \qquad u \in \Omega. \tag{2}$$

In particular, g can be recovered by evaluating (2) at $t = 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\varphi_{t,g}(u)|_{t=0} = g(u), \qquad u \in \Omega. \tag{3}$$

The vector field g is said to be the infinitesimal generator of the group $\{\varphi_{t,g}\}$.

If f and g are vector fields on Ω, the corresponding flows generally do not commute. The commutativity condition for the flows can be written in terms of the Lie bracket of the vector fields [14]:

$$\forall s, t, \qquad \varphi_{t,f} \circ \varphi_{s,g} = \varphi_{s,g} \circ \varphi_{t,f} \Leftrightarrow [f, g] \equiv 0, \tag{4}$$

where $[f, g]$ is the vector field

$$[f, g](u) = g'(u)f(u) - f'(u)g(u), \qquad u \in \Omega.$$

Here and later, the prime denotes the corresponding Jacobian matrix.

2.2. Symmetry groups of differential equations

Let f be a vector field on Ω. A group $\mathcal{G}$ of diffeomorphisms in Ω is a *symmetry group* [14] of the system $\dot{u} = f(u)$ or f admits $\mathcal{G}$ as a symmetry group if

$$\forall t, \forall G \in \mathcal{G}, \forall u \in \Omega, \qquad \varphi_{t,f}(G(u)) = G(\varphi_{t,f}(u)). \tag{5}$$

The condition (5) means that the solution $t \to \varphi_{t,f}(G(u))$ corresponding to the transformed initial datum $G(u)$ is obtained by transforming by G the solution $\varphi_{t,f}(u)$.

If we differentiate (5) with respect to t, evaluate at $t = 0$ and use (3), we obtain the symmetry condition in terms of the field f of the system; namely

$$\forall G \in \mathcal{G},\ \forall u \in \Omega \qquad f(G(u)) = G'(u)f(u). \tag{6}$$

We can also use Lie brackets to express the symmetry condition. If g is an infinitesimal generator of the group $\mathcal{G}$ (i.e. $\{\varphi_{t,g} : t \in \mathcal{R}\}$ is a one-parameter subgroup of $\mathcal{G}$) and f admits $\mathcal{G}$ as a symmetry group then, by (4) and (6) $[f, g] \equiv 0$. Conversely, if $[f, g] \equiv 0$ for all the generators of $\mathcal{G}$ then f admits the group $\mathcal{G}$.

2.3. The Abelian case

From now on, we consider vector fields $f, g_1, \ldots, g_\nu$ defined in a domain $\Omega \subset \mathcal{R}^D$ and satisfying the following conditions.

(H1) For each $u \in \Omega$, the vectors $g_1(u), \ldots, g_\nu(u)$ are linearly independent.

(H2) $[f, g_j] \equiv 0,\ \ j = 1, \ldots \nu$.

(H3) $[g_i, g_j] \equiv 0,\ \ i, j = 1, \ldots, \nu$.

We still suppose that the phase flows $\varphi_{t,f}, \varphi_{t,g_i}, i = 1, \ldots, \nu$ are defined in the whole of Ω. For each $(\tau_1, \ldots, \tau_\nu) \in \mathcal{R}^\nu$ we have a diffeomorphism of Ω

$$G_{(\tau_1,\ldots,\tau_\nu)} = \varphi_{\tau_1,g_1} \circ \varphi_{\tau_2,g_2} \circ \cdots \circ \varphi_{\tau_\nu,g_\nu}. \tag{7}$$

The hypothesis (H3) implies that the transformations (7) form an Abelian ν-parameter group $\mathcal{G}$: $G_0 = \mathrm{Id}$, $G_{(\tau_1,\ldots,\tau_\nu)} \circ G_{(\sigma_1,\ldots,\sigma_\nu)} = G_{(\tau_1+\sigma_1,\ldots,\tau_\nu+\sigma_\nu)}$. The condition (H2) shows that $\mathcal{G}$ is a symmetry group for the system

$$\dot{u} = f(u), \qquad u \in \Omega. \tag{8}$$

Due to the presence of the symmetry group $\mathcal{G}$, the dynamics of (8) can be studied by means of a simpler system called the *reduced system*. Since (H1) holds, all the orbits $\{G_{\tau_1,\ldots,\tau_\nu}(u) : (\tau_1, \ldots, \tau_\nu) \in \mathcal{R}^\nu\}$ of the group are ν-dimensional submanifolds (this is a simple consequence of the implicit function theorem). We can construct the *reduced phase space* with dimension $D - \nu$ by identifying points in Ω that belong to the same orbit of the group $\{G(u) : G \in \mathcal{G}\}$. Thus a point in the reduced space is an orbit of the group. The system (8) in Ω gives rise in a natural way to a new system in the reduced phase space, the so-called reduced system. We describe how to write the reduced system in local coordinates. Given $u_0 \in \Omega$, due to (H1), (H3) [14] there is a neighbourhood U of u_0 and a local change of variables in U that carries $g_i, i = 1, \ldots, \nu$ into the constant vector field equal to the ith coordinate vector $(0, 0, \ldots, 1, 0, \ldots, 0)^T$ (the 1 is in the ith place). Thus, there are coordinates $(x, y) = (x_1, \ldots, x_\nu, y_1, \ldots, y_{D-\nu})$ in U for which the elements of the group $\mathcal{G}$ can be expressed in the form

$$G_{(\tau_1,\ldots,\tau_\nu)}(x, y) = (x_1 + \tau_1, \ldots, x_\nu + \tau_\nu, y_1, \ldots, y_{D-\nu}), \qquad (x, y) \in U. \tag{9}$$

The system (8) can then be written as

$$\dot{x} = F_1(y), \tag{10}$$

$$\dot{y} = F_2(y), \tag{11}$$

(x is not an argument of F_1, F_2 as these admit the group of translations (9)) and the reduced system is locally (11). Therefore, the integration of the original D-dimensional system is reduced to the integration of (11), which is $(D-\nu)$-dimensional, and to ν quadratures to recover $x(t) = \int F_1(y(t))\,\mathrm{d}t$.

An example will now be presented. The planar system

$$\dot{u_1} = u_1^2/u_2, \qquad \dot{u_2} = u_2^2/u_1, \tag{12}$$

is invariant with respect to the one-parameter group of dilations $G_\tau(u_1, u_2) = (\mathrm{e}^\tau u_1, \mathrm{e}^\tau u_2)$. (To see this, note that the infinitesimal generator of the group is $g(u_1, u_2) = (u_1, u_2)$ and check that $[f, g] = 0$; alternatively one may see geometrically that (6) holds.) The group orbits are rays through the origin and the reduced phase space may be identified with the unit circumference. In this example, the x variable is given by $x = \log\sqrt{u_1^2 + u_2^2}$ because the transformation $(u_1, u_2) \to (\mathrm{e}^\tau u_1, \mathrm{e}^\tau u_2)$ increases x by τ. Furthermore we take $y = u_1/u_2$ (any function of the quotient $y = u_1/u_2$ may do as well). In terms of the new variables (12) becomes

$$\dot{x} = \frac{1+y^4}{y(1+y^2)}, \qquad \dot{y} = y^2 - 1. \tag{13}$$

Essentially, the second equation in (13) (i.e. the reduced system) describes the evolution of the argument of the point (u_1, u_2) in the plane. The first equation governs the evolution of the modulus $\sqrt{u_1^2 + u_2^2}$, i.e. the drift along group orbits. After integration of the y-equation by separation of variables one may find $x(t)$ by quadrature. We have more or less reproduced the elementary technique of integration of the 'homogeneous' equation $\mathrm{d}u_1/\mathrm{d}u_2 = u_1^3/u_2^3$ by taking u_1/u_2 as a new variable.

2.4. Relative equilibria and their variational equations

The equilibria of the reduced system play an important role in many applications. In the situation described in the preceding section, a point $u_0 \in \Omega$ is a *relative equilibrium* [1, 13] if there exist real numbers $\lambda_0^1, \ldots, \lambda_0^\nu$ such that

$$f(u_0) - \sum_{i=1}^{\nu} \lambda_0^i g_i(u_0) = 0. \tag{14}$$

The following lemma shows the implications of the condition (14).

Lemma 2.1. *Assume that, for the system (8), the conditions (H1)–(H3) hold.*

(i) Let u_0 be a relative equilibrium as in (14). Then

$$\varphi_{t,f}(u_0) = G_{(t\lambda_0^1,\ldots,t\lambda_0^\nu)}(u_0), \tag{15}$$

which in particular shows that

$$\{\varphi_{t,f}(u_0) : t \in \mathcal{R}\} \subset \{G_{(\tau_1,\ldots,\tau_\nu)}(u_0) : (\tau_1, \ldots, \tau_\nu) \in \mathcal{R}^\nu\}, \tag{16}$$

i.e. that, for each t, $\varphi_{t,f}(u_0)$ is contained in the group orbit through u_0.

(ii) Conversely, if (16) holds then there are real numbers λ_0^i for which (14) is satisfied.

Thus, relative equilibria are precisely the points that project to an equilibrium of the reduced system. For instance, if $\nu = 1$, $D = 2$ and $\mathcal{G}$ consists of all planar rotations around the origin, a relative equilibrium would be a point u_0 such that for all t, $\varphi_{t,f}(u_0)$ can be obtained by rotating u_0 by a suitable angle $\alpha(t)$. Note that by (15) $\alpha(t)$ is necessarily of the form $\alpha(t) = \lambda_0^1 t$, i.e. $\varphi_{t,f}(u_0)$ rotates at a uniform rate.

In the particular case where in (14) all λ_0^i vanish we have $f(u_0) = 0$ and the relative equilibrium is in fact an equilibrium.

In the example (12) the condition (14) is given by

$$\left(\frac{u_1^2}{u_2}, \frac{u_2^2}{u_1}\right) - \lambda(u_1, u_2) = 0. \tag{17}$$

Elimination of λ leads to $u_1^2 = u_2^2$, so that the relative equilibria are given by the points on the diagonals $u_1 = \pm u_2$. These points have $y = u_1/u_2 = \pm 1$ so that they project to equilibria of the reduced system (see (13)) $\dot{y} = y^2 - 1$. Furthermore, if $u_1 = u_2$, from (17), we have $\lambda = 1$ leading to the solutions of the form $u_1(t) = \mathrm{e}^t\alpha$, $u_2(t) = \mathrm{e}^t\alpha$. For $u_1 = -u_2$, (17) yields $\lambda = -1$, leading to $u_1(t) = \mathrm{e}^{-t}\alpha$, $u_2(t) = -\mathrm{e}^{-t}\alpha$. These are the only solutions of (12) whose trajectories on the (u_1, u_2) plane are rays through the origin, i.e. orbits of the group.

Proof of lemma 2.1. Due to the invariance of $G_{(t\lambda_0^1,\ldots,t\lambda_0^\nu)}$ with respect to the phase flows of the vector fields g_i, we can write (see (6))

$$g_i(G_{(t\lambda_0^1,\ldots,t\lambda_0^\nu)}(u_0)) = G'_{(t\lambda_0^1,\ldots,t\lambda_0^\nu)}(u_0)g_i(u_0), \qquad i = 1, \ldots, \nu. \tag{18}$$

Therefore, using (H2), (14) and (18), we have

$$\begin{aligned}\dot{u}(t) &= \sum_{i=1}^{\nu}\lambda_0^i g_i(G_{(t\lambda_0^1,\ldots,t\lambda_0^\nu)}(u_0)) = G'_{(t\lambda_0^1,\ldots,t\lambda_0^\nu)}(u_0)\left(\sum_{i=1}^{\nu}\lambda_0^i g_i(u_0)\right)\\ &= G'_{(t\lambda_0^1,\ldots,t\lambda_0^\nu)}(u_0)f(u_0) = f(G_{(t\lambda_0^1,\ldots,t\lambda_0^\nu)}(u_0))\\ &= f(u(t)).\end{aligned}$$

That proves (i). If a solution $u(t)$ of (8) is contained in the orbit $\{G(u_0) : G \in \mathcal{G}\}$ then the phase velocity vector $f(u)$ must be tangent at u_0 to the orbit. This implies (14) because the $g_i(u_0)$ span the tangent space to the orbit. □

The variational equation governs the way in which the local errors in a numerical integration build up to give the global error. The next result describes the properties of the variational equation near a relative equilibrium.

Lemma 2.2. *Assume that, for the system (8), the conditions (H1)–(H3) hold. If u_0 is a relative equilibrium as in (14) and $\varphi_{t,f}(u_0) = u(t)$ then the following results hold.*

(i) The solutions of the homogeneous variational equation

$$\dot{\delta}(t) = f'(u(t))\delta(t)$$

resulting from linearizing (8) around $u(t)$ are of the form

$$\delta(t) = G'_{(t\lambda_0^1,\ldots,t\lambda_0^\nu)}(u_0)\Delta(t), \tag{19}$$

where $\Delta(t)$ is a solution of the linear system with constant coefficients

$$\dot{\Delta}(t) = L\Delta(t), \qquad L = f'(u_0) - \sum_{i=1}^{\nu}\lambda_0^i g_i'(u_0) \tag{20}$$

resulting from linearizing the differential equation

$$\dot{u} = f(u) - \sum_{i=1}^{\nu}\lambda_0^i g_i(u)$$

around its equilibrium u_0.

(ii) 0 *is an eigenvalue of* L *and* $g_i(u_0) \in \mathrm{Ker} L$, $i = 1, \dots, \nu$.

(iii) Let s *be a vector field that admits the one-parameter group* $\{G_{(t\lambda_0^1,\dots,t\lambda_0^\nu)} : t \in \mathcal{R}\}$ *as a symmetry group. Then, the solution of the nonhomogeneous variational problem*

$$\dot{\delta}(t) = f'(u(t))\delta(t) + s(u(t)), \qquad \delta(0) = 0 \tag{21}$$

can be written in the form (19) where $\Delta(t)$ *satisfies*

$$\dot{\Delta}(t) = L\Delta(t) + s(u_0), \qquad \Delta(0) = 0, \tag{22}$$

that is

$$\Delta(t) = \int_0^t \mathrm{e}^{(t-\tau)L}\, \mathrm{d}\tau\, s(u_0). \tag{23}$$

Proof. Note first that since for each value of t, $G_{(t\lambda_0^1,\dots,t\lambda_0^\nu)}$ is a diffeomorphism, the linear operator $G'_{(t\lambda_0^1,\dots,t\lambda_0^\nu)}$ is an isomorphism and we may interpret the relation (19) as a (time-dependent) change of variables $\delta(t) \longmapsto \Delta(t)$. Now, we can write, by differentiating (19),

$$\dot{\delta}(t) = \sum_{i=1}^{\nu} \lambda_0^i g_i'(G_{(t\lambda_0^1,\dots,t\lambda_0^\nu)}(u_0))G'_{(t\lambda_0^1,\dots,t\lambda_0^\nu)}(u_0)\Delta(t) + G'_{(t\lambda_0^1,\dots,t\lambda_0^\nu)}(u_0)\dot{\Delta}(t). \tag{24}$$

On the other hand, since δ satisfies the homogeneous variational equation around $u(t)$, we have

$$\dot{\delta}(t) = f'(G_{(t\lambda_0^1,\dots,t\lambda_0^\nu)}(u_0))\delta(t) = f'(G_{(t\lambda_0^1,\dots,t\lambda_0^\nu)}(u_0))G'_{(t\lambda_0^1,\dots,t\lambda_0^\nu)}(u_0)\Delta(t). \tag{25}$$

Comparing (24) and (25), we obtain

$$\begin{aligned}\dot{\Delta}(t) = (G'_{(t\lambda_0^1,\dots,t\lambda_0^\nu)}(u_0))^{-1}(f'(G_{(t\lambda_0^1,\dots,t\lambda_0^\nu)}(u_0))\\ - \sum_{i=1}^{\nu} \lambda_0^i g_i'(G_{(t\lambda_0^1,\dots,t\lambda_0^\nu)}(u_0)))G'_{(t\lambda_0^1,\dots,t\lambda_0^\nu)}(u_0)\Delta(t).\end{aligned} \tag{26}$$

If $\bar{f}$ is the vector field given by $\bar{f} = f - \sum_{i=1}^{\nu} \lambda_0^i g_i$, then $\bar{f}$ admits $G_{(t\lambda_0^1,\dots,t\lambda_0^\nu)}$ as a symmetry group and therefore

$$\bar{f}'(G_{(t\lambda_0^1,\dots,t\lambda_0^\nu)}(u))G'_{(t\lambda_0^1,\dots,t\lambda_0^\nu)}(u) = G''_{(t\lambda_0^1,\dots,t\lambda_0^\nu)}(u)\bar{f}(u) + G'_{(t\lambda_0^1,\dots,t\lambda_0^\nu)}(u)\bar{f}'(u), \qquad u \in \Omega.$$

Evaluating this expression at $u = u_0$, using (14) and substituting in (25), we obtain

$$\begin{aligned}\dot{\Delta}(t) &= (G'_{(t\lambda_0^1,\dots,t\lambda_0^\nu)}(u_0))^{-1}\bar{f}'(G_{(t\lambda_0^1,\dots,t\lambda_0^\nu)}(u_0)G'_{(t\lambda_0^1,\dots,t\lambda_0^\nu)}(u_0)\Delta(t)\\ &= \bar{f}'(u_0)\Delta(t) = L\Delta(t)\end{aligned}$$

with L given by (20).

Now, using (H2), (H3) and (14), observe that

$$\begin{aligned}Lg_j(u_0) &= f'(u_0)g_j(u_0) - \sum_{i=1}^{\nu} \lambda_0^i g_i'(u_0)g_j(u_0)\\ &= g_j'(u_0)f(u_0) - \sum_{i=1}^{\nu} \lambda_0^i g_j'(u_0)g_i(u_0)\\ &= g_j'(u_0)(f(u_0) - \sum_{i=1}^{\nu} \lambda_0^i g_i(u_0)) = 0.\end{aligned}$$

for $j = 1, \dots, \nu$. That proves (ii). Finally, the proof of (iii) is analogous to that of (i). □

According to (19), there are two sources of growth with time in the solution of (21). First, the growth of $||G'_{(t\lambda_0^1,\ldots,t\lambda_0^\nu)}(u_0)||$ and, on the other hand, the growth of $\Delta(t)$. In many applications the symmetry group consists of isometries such as rotations and translations and then $||G'_{(t\lambda_0^1,\ldots,t\lambda_0^\nu)}(u_0)|| = 1$. In what follows we concentrate on the behaviour of $\Delta(t)$. The constant coefficient system (22) can be solved easily. Let M be the unique L-invariant supplementary subspace in $\mathcal{R}^D$ of the generalized kernel of L (i.e M is the sum of the invariant subspaces corresponding to the nonzero eigenvalues of L). We can decompose the vector $s(u_0)$ in the form

$$s(u_0) = s_M + \sum_{j=1}^{\sigma} s^{(j)}, \tag{27}$$

where $s_M \in M$, σ is the maximum of the sizes of the Jordan blocks of L associated with the eigenvalue 0, $s^{(1)} \in \mathrm{Ker}L$ and $s^{(j)} \in \mathrm{Ker}L^j \backslash \mathrm{Ker}L^{j-1}$, $j = 2, 3, \ldots, \sigma$. There exists a unique $\widehat{L}^{-1}s_M \in M$ where $\widehat{L}$ denotes the restriction $L\big|_M : M \to M$ and, after simple manipulations, (23) can be written as

$$\Delta(t) = \Delta_M(t) + \Delta_{\mathrm{Ker}}(t), \tag{28}$$

with

$$\Delta_{\mathrm{Ker}}(t) = \sum_{j=1}^{\sigma} \sum_{k=0}^{j-1} \frac{t^{k+1}}{(k+1)!} L^k s^{(j)},$$

$$\Delta_M(t) = (\mathrm{e}^{t\widehat{L}} - I)(\widehat{L}^{-1} s_M).$$

The following result shows the behaviour of $\Delta(t)$ in the (typical) case of nondegenerate relative equilibrium.

Lemma 2.3. *Assume that, in addition to the conditions of lemma 2.2, the relative equilibrium u_0 is a nondegenerate equilibrium of the reduced system, i.e.* 0 *is not an eigenvalue of the operator L_R of the linearization of the reduced system around (the orbit of) u_0. Then we have the following.*

(i) The geometric and algebraic multiplicities of 0 *as an eigenvalue of L equal ν and the vectors $g_i(u_0)$, $i = 1, \ldots, \nu$ form a basis of* Ker L. *The eigenvalues and Jordan structure of $\widehat{L} = L\big|_M$ coincide with those of L_R.*

(ii) Near u_0, every relative equilibrium is of the form $G_{(\tau_1,\ldots,\tau_\nu)}(u_0)$ with $\tau_1, \ldots, \tau_\nu \in \mathcal{R}$ and has the same multipliers λ_0^i as u_0.

(iii) $\Delta_{\mathrm{Ker}}(t) = ts^{(1)}$.

Proof. We first note that, since the system (8) can be written, in a neighbourhood of u_0, in the form (10), (11), the matrix L in (20) is similar to the $D \times D$ matrix

$$\begin{pmatrix} 0 & F_1'(y_0) \\ 0 & F_2'(y_0) \end{pmatrix},$$

where $u_0 = (x_0, y_0)$ and, since u_0 is nondegenerate, $F_2'(y_0)$ is a nonsingular $D - \nu$ block. This proves that the algebraic and geometric multiplicities of the eigenvalue 0 are ν. By (ii) in lemma 2.2, the $g_i(u_0)$ are a basis of Ker L. Furthermore the operator $\widehat{L}$ can be expressed by the matrix $F_2'(y_0)$, which in turn is the Jacobian of the reduced system $\mathrm{d}y/\mathrm{d}t = F_2(y)$ at the reduced equilibrium $u_0 = (x_0, y_0)$. This proves (i).

The implicit function theorem implies that in a sufficiently small neighbourhood of u_0 there is no equilibrium of the reduced system different from y_0. Hence, near u_0, every relative equilibrium is of the form

$$G_{(\tau_1,\ldots,\tau_\nu)}(u_0), \qquad \tau_1,\ldots,\tau_\nu \in \mathcal{R}.$$

Furthermore, by (6)

$$f(G_{(\tau_1,\ldots,\tau_\nu)}(u_0)) - \sum_{i=1}^{\nu} \lambda_0^i g_i(G_{(\tau_1,\ldots,\tau_\nu)}(u_0)) = G'_{(\tau_1,\ldots,\tau_\nu)}(u_0)\left(f(u_0) - \sum_{i=1}^{\nu}\lambda_0^i g_i(u_0)\right) = 0,$$

and that proves (ii). The part (iii) is trivial. □

For a nondegenerate relative equilibrium, $\Delta(t)$ consists of a part $\Delta_{\mathrm{Ker}}(t) = ts^{(1)}$ that grows linearly in time and is tangent to the group orbit and of a complementary term $\Delta_M(t)$. The complementary term grows like $\exp(tL_R)$; in particular, if u_0 is a linearly stable equilibrium of the reduced system, $\Delta_M(t)$ is a bounded function for $0 \leqslant t < \infty$.

2.5. The Hamiltonian case

Let us now assume that the system (8) being integrated is Hamiltonian. This means [1, 14] that the dimension D is even $D = 2d$ and the vector field f is of the form $f(u) = \Xi\nabla H(u)$, where Ξ is the constant, skew-symmetric, invertible matrix

$$\Xi = \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix}$$

and $\nabla H(u)$ is the gradient of a function $H : \Omega \to \mathcal{R}$ called the Hamiltonian. We additionally assume that the group generators g_i are also Hamiltonian vector fields, $g_i = \Xi\nabla I_i$ associated with Hamiltonian functions I_i satisfying the conditions:

(H2′) $\{I_i, H\} = 0, \qquad i = 1,\ldots,\nu,$

(H3′) $\{I_i, I_j\} = 0, \qquad i, j = 1,\ldots,\nu,$

where $\{\cdot,\cdot\}$ denotes the Poisson bracket determined by Ξ:

$$\{F, G\}(u) = \nabla F(u)^T\,\Xi\nabla G(u), \qquad u \in \Omega.$$

Since $[f, g_i] = \Xi\nabla\{H, I_i\}$, the relation (H2′) implies (H2). Furthermore (H2′) implies that each I_i is a first integral of (8). In a similar way, (H3′) implies (H3) and that each I_i is a first integral of the Hamiltonian system with Hamiltonian function I_j.

The existence of the conserved quantities I_i plays an important role in the construction of the reduced phase space for this case. The phase space Ω is foliated by level sets of $I_1,\ldots,I_\nu$, that are manifolds invariant by the symmetry group. In turn each of these level manifolds is foliated by group orbits. This is described next in terms of local coordinates. (The reader may now wish to see the example in section 4 below.)

Let us introduce locally new coordinates (p_x, p_y, x, y), $p_x, x \in \mathcal{R}^\nu$, $p_y, y \in \mathcal{R}^{d-\nu}$ in such a way that $p_{x_i} = I_i$, $i = 1,\ldots\nu$ and that the change $u \longmapsto u^* = (p_x(u), p_y(u), x(u), y(u))$ is canonical (i.e. any two components of u^* possess zero Poisson bracket except for $\{p_{x_i}, x_i\} = 1, i = 1,\ldots,\nu$, $\{p_{y_j}, y_j\} = 1$, $j = 1,\ldots,d-\nu$). Since the Hamiltonian system with Hamiltonian function $p_{x_i} = I_i$ is $\dot{p}_{x_j} = 0$, $\dot{p}_{y_j} = 0$, $\dot{x_j} = \delta_{ij}$, $\dot{y}_j = 0$, the corresponding flow is translation in x_i; this implies that, in the new variables a symmetric Hamiltonian depends only on p_x, p_y, y. Therefore the equations of motion are

$$\dot{p}_{x_i} = 0, \qquad i = 1,\ldots,\nu, \tag{29}$$

$$\dot{p}_{y_i} = -\frac{\partial}{\partial y_i} H(p_x, p_y, y), \qquad i = 1,\ldots,d-\nu, \tag{30}$$

$$\dot{x}_i = \frac{\partial}{\partial p_{x_i}} H(p_x, p_y, y), \qquad i = 1, \ldots, \nu, \tag{31}$$

$$\dot{y}_i = \frac{\partial}{\partial p_{y_i}} H(p_x, p_y, y), \qquad i = 1, \ldots, d - \nu, \tag{32}$$

and the reduced system considered in section 2.3 would be given by (29), (30), (32). However, in the Hamiltonian case one proceeds in a slightly different way [1, 14]. One fixes $c = (c^1, \ldots, c^\nu) \in \mathcal{R}^\nu$ and considers the given system (29)–(32) restricted to the invariant level set $p_{x_i} = c_i$, $i = 1, \ldots, \nu$, i.e.

$$\dot{p}_{y_i} = -\frac{\partial}{\partial y_i} H(c, p_y, y), \qquad i = 1, \ldots, d - \nu, \tag{33}$$

$$\dot{x}_i = \frac{\partial}{\partial p_{x_i}} H(c, p_y, y), \qquad i = 1, \ldots, \nu, \tag{34}$$

$$\dot{y}_i = \frac{\partial}{\partial p_{y_i}} H(c, p_y, y), \qquad i = 1, \ldots, d - \nu. \tag{35}$$

The system (33)–(35) is $(2d - \nu)$-dimensional and is still symmetric with respect to x-translations. Then we reduce (33)–(35) by ignoring the x variables; this yields the reduced $2(d - \nu)$-dimensional system given by (33), (35). With this reduction the reduced system is Hamiltonian with Hamiltonian function $\tilde{H}(p_y, y) = H(c, p_y, y)$.

Having discussed the Hamiltonian reduction, we note that, if we restrict the attention to a level set $I_i = c_0^i$, $i = 1, \ldots, \nu$, then a relative equilibrium u_0 must satisfy (recall that now in (14) $f = \Xi \nabla H$, $g_i = \Xi \nabla I_i$)

$$\nabla \left(H(u_0) - \sum_{i=1}^{\nu} \lambda_0^i I_i(u_0) \right) = 0, \tag{36}$$

$$I_i(u_0) = c_0^i, \qquad i = 1, \ldots, \nu. \tag{37}$$

Thus u_0 is a stationary point of H restricted to the level set.

We can now study the solutions of the variational equation of a symmetric Hamiltonian system around a relative equilibrium. As in lemma 2.2 we denote by L the linearized operator

$$L = f'(u_0) - \sum_{i=1}^{\nu} \lambda_0^i g_i'(u_0),$$

by $\widehat{L}$ the restriction of L to the invariant supplement M of the generalized kernel of L and by L_R the linearization of the reduced system.

The following lemma summarizes the structure of L in the Hamiltonian case.

Lemma 2.4. *Assume that (H1), (H2′) and (H3′) hold and let u_0 be a relative equilibrium as in (36), (37). Then we have the following.*

(i) $\nabla I_j(u_0)$, $j = 1, \ldots, \nu$ are left eigenvectors of L with 0 eigenvalue.

Furthermore, if u_0 is a nondegenerate equilibrium of the reduced Hamiltonian system (33), (35), then:

(ii) There are (nonunique) smooth mappings $u = u(c)$, $\lambda = \lambda(c)$ such that $u(c_0) = u_0$, $\lambda(c_0) = \lambda_0$ and that for c close to c_0

$$\nabla \left(H(u(c)) - \sum_{i=1}^{\nu} \lambda^i(c) I_i(u(c)) \right) = 0, \tag{38}$$

$$I_i(u(c)) = c_i, \qquad i = 1, \ldots, \nu, \tag{39}$$

i.e. for each fixed c, $u(c)$ is a relative equilibrium with multipliers $\lambda^i(c)$.

(iii) The algebraic multiplicity of zero as an eigenvalue of L is 2ν. Moreover, $\frac{\partial u(c)}{\partial c_i}|_{c=c_0} \in \mathrm{Ker} L^2$, $i = 1, \ldots, \nu$ *with*

$$\nabla I_i(u_0)^T \frac{\partial u(c)}{\partial c_j}\bigg|_{c=c_0} = \delta_{ij}, \qquad i, j = 1, \ldots, \nu. \tag{40}$$

and the 2ν vectors $\frac{\partial u(c)}{\partial c_j}|_{c=c_0}$, $\Xi \nabla I_j(u_0)$, $j = 1, \ldots, \nu$ *form a basis of the generalized kernel of L,* $\mathrm{Ker}\, L^2$. *The eigenvalues and Jordan structure of $\widehat{L}$ coincide with those of L_R.*

(iv) Near u_0 every relative equilibrium is of the form $G_{(\tau_1,\ldots,\tau_\nu)}(u(c))$ with multipliers $\lambda^i = \lambda^i(c)$, $i = 1, \ldots, \nu$.

(v) If the matrix $(\frac{\partial \lambda^i(c)}{\partial c_j}|_{c=c_0})_{i,j=1,\ldots,\nu}$ *is nonsingular, the geometric multiplicity of zero as an eigenvalue of L is ν; the vectors $g_j(u_0) = \Xi \nabla I_j(u_0)$, $j = 1, \ldots, \nu$ form a basis of* $\mathrm{Ker}\, L$ *and* $\frac{\partial u(c)}{\partial c_j}|_{c=c_0}$, $j = 1, \ldots, \nu$ *form a basis of a supplement of* $\mathrm{Ker}\, L$ *in* $\mathrm{Ker}\, L^2$. *Moreover*

$$\Delta_{\mathrm{Ker}}(t) = ts^{(1)} + (tI + \frac{t^2}{2}L)s^{(2)},$$

so that $\Delta_{\mathrm{Ker}}(t)$ grows quadratically with t unless $s^{(2)} = 0$, which happens if and only if

$$\nabla I_i(u_0)^T s(u_0) = 0, \qquad i = 1, \ldots, \nu, \tag{41}$$

Proof. Note that, since (36) holds, for $j = 1, \ldots, \nu$ we have

$$\nabla I_j(u_0)^T \left(f(u_0) - \sum_{i=1}^{\nu} \lambda_0^i g_i(u_0) \right) = 0,$$

which, after differentiating and applying (36), leads to

$$\nabla I_j(u_0)^T L = 0, \qquad j = 1, \ldots, \nu.$$

On the other hand, if u_0 is a nondegenerate equilibrium of (33), (35), we can obtain (ii) by using the implicit function theorem. In local coordinates, we have to write, for a relative equilibrium, the local variables p_x, p_y, x, y and the multipliers λ as functions of c. By setting $\dot{p}_{y_i} = 0$, $\dot{y}_i = 0$, $i = 1, \ldots, d - \nu$ in the reduced system (36), (37) we obtain uniquely $y(c)$, $p_y(c)$; furthermore in local coordinates $p_x = c$. The x variables can be chosen freely and the multipliers are given by $\partial H / \partial p_{x_i}$.

Now, differentiating (38) with respect to c_j and evaluating at $c = c_0$, we have

$$L \frac{\partial u}{\partial c_j}\Big|_{c=c_0} - \sum_{i=1}^{\nu} \frac{\partial \lambda^i}{\partial c_j}\Big|_{c=c_0} g_i(u_0) = 0. \tag{42}$$

Hence, $\frac{\partial u}{\partial c_j}\big|_{c=c_0} \in \mathrm{Ker} L^2$, $j = 1, \ldots, \nu$ and if we differentiate (39) with respect to c_j and evaluate the resulting expression at $c = c_0$, we obtain (40). Note that, since (H3$'$) holds, each $g_i(u_0)$ is orthogonal to each $\nabla I_j(u_0)$. This result and (40) imply that the 2ν vectors $\frac{\partial u}{\partial c_j}\big|_{c=c_0}$, $g_j(u_0)$, $j = 1, \ldots, \nu$ are linearly independent. Therefore, the algebraic multiplicity of zero as an eigenvalue of L is at least 2ν. But, in local coordinates, the matrix L is similar to a block matrix with a nonsingular submatrix of order $D - 2\nu$ (because u_0 is nondegenerate) so that the algebraic multiplicity is exactly 2ν. The proof of (iv) is based on the implicit function theorem and (ii).

On the other hand, observe that if $(\frac{\partial \lambda^i(c)}{\partial c_j}|_{c=c_0})_{i,j=1,\ldots,\nu}$ is invertible, from (42) we have $\frac{\partial u}{\partial c_j}|_{c=c_0} \in \mathrm{Ker} L^2 \backslash \mathrm{Ker} L$, $j = 1, \ldots, \nu$. To conclude the proof note that, by (40), the component $s^{(2)}$ vanishes if and only if (41) holds. □

Thus for a nondegenerate relative equilibrium, the behaviour of the complementary term $\Delta_M(t)$ is governed by $\exp(tL_R)$; the difference with the situation in the preceding section is that now L_R has dimension $D-2\nu$ rather than $D-\nu$. On the other hand the generalized kernel has dimension 2ν rather than ν and we expect quadratic growth in $\Delta_{\mathrm{Ker}}(t)$. Note that if the source vector $s(u_0)$ is orthogonal to the surfaces $I_i(u) = I_i(u_0)$, the growth of $\Delta_{\mathrm{Ker}}(t)$ is only linear. Note also that in (vi) the dominant term $\frac{t^2}{2}Ls^{(2)}$ belongs to $\mathrm{Ker}\, L$ and therefore (see (v)) lies in a direction spanned by the $g_j(u_0)$s. Thus the leading part of the error Δ can be interpreted as an error tangent to the group orbit $\{G_{(\tau_1,\ldots,\tau_\nu)}(u_0) : (\tau_1, \ldots, \tau_\nu) \in \mathcal{R}^\nu\}$. This will be used below to show that the leading error in computing the solution $G_{(t\lambda_0^1,\ldots,t\lambda_0^\nu)}(u_0)$ can be interpreted as an error in the values of the group parameters $t\lambda_0^1, \ldots, t\lambda_0^\nu$. If $s^{(2)} = 0$, then the leading term is $ts^{(1)}$, which also belongs to $\mathrm{Ker}\, L$.

3. Numerical approximation

Our purpose now is to apply the preceding results to analyse the time propagation of the errors in the numerical integration of differential equations. We are interested in one-step methods for the system (8), given by mappings $\psi_{h,f} : \Omega \to \Omega$ that advance the solution h units of time, where h is the step size. The map $\psi_{h,f}$ approximates the flow $\varphi_{h,f}$ of the system and the numerical solution is obtained by iteration of $\psi_{h,f}$,

$$U_{n+1} = \psi_{h,f}(U_n), \qquad n = 0, 1, \ldots, \tag{43}$$

so that, if $U_0 = u_0$, U_n is an approximation to the solution value $u(t_n) = \varphi_{t_n,f}(u_0)$, $t_n = nh$.

The local (truncation) error at a point $u_0 \in \Omega$ is, by definition, the difference $\varphi_{h,f}(u_0) - \psi_{h,f}(u_0)$. A numerical method of the form (43) has order of consistency r if the local error is, for each $u_0 \in \Omega$, $\mathrm{O}(h^{r+1})$ as $h \to 0$. The mapping $\psi_{h,f}$ is therefore consistent with $\varphi_{h,f}$ of order r and the global errors $U_n - u(t_n)$ are $\mathrm{O}(h^r)$ as $h \to 0$. The smoothness of (8) and the consistency of order r of (43) guarantee that the local error has an asymptotic expansion of the form

$$\varphi_{h,f}(u) - \psi_{h,f}(u) = h^{r+1} l_{r+1}(u) + h^{r+1} R(h, u),$$

where l_{r+1}, R are smooth functions such that l_{r+1} is independent of h and $R(h, u) \to 0$ as $h \to 0$ with u fixed.

3.1. Error propagation

We assume that a scheme of the form (43) with order $r \geqslant 1$ is used to approximate the solution of (8) $u(t) = G_{(t\lambda_0^1,\ldots,t\lambda_0^\nu)}(u_0)$, where u_0 is a relative equilibrium and $U_0 = u_0$. We make some additional hypotheses about (43).

(A1) The global error admits an expansion

$$U_n - u(t_n) = h^r e(t_n) + h^r Q(t_n, h), \tag{44}$$

where e is a smooth function that satisfies the variational problem [4]:

$$\begin{aligned} \dot{e} &= f'(u(t)) \cdot e - l_{r+1}(u(t)), \\ e(0) &= 0, \end{aligned} \tag{45}$$

and Q is a remainder that, for fixed t, tends to zero as $h \to 0$.

(A2) The mapping $\psi_{h,f}$ is invariant by $\{G_{(t\lambda_0^1,\ldots,t\lambda_0^\nu)} : t \in \mathcal{R}\}$. This implies that the source term of (45) admits this group as a symmetry group.

The assumption (A1) is satisfied by virtually all methods used in practice. As far as (A2) is concerned, this is valid for most methods if the elements of the group are linear transformations, see e.g. [17]. We can now state the following result, where the notation is as in Lemma 2.3 and l_M is the projection of l_{r+1} onto M parallel to the generalized kernel of L.

Theorem 3.1. *Under the assumptions of lemma 2.3, suppose that (A1), (A2) hold. If u_0 is a nondegenerate equilibrium of the reduced system, then*

$$U_n = G_{(t_n(\lambda_0^1+h^r\alpha_1),\ldots,t_n(\lambda_0^\nu+h^r\alpha_\nu))}(u_0) + h^r G'_{(t_n\lambda_0^1,\ldots,t_n\lambda_0^\nu)}(u_0)(\mathrm{e}^{t_n\widehat{L}} - I)\widehat{L}^{-1} l_M + h^r q(h, t_n), \tag{46}$$

for suitable coefficients α_i, $i = 1, \ldots, \nu$. The function q is a remainder that, for fixed t, tends to zero as $h \to 0$.

If u_0 is linearly stable as an equilibrium of the reduced vector field and the elements of the group are isometries, the second term of the right-hand side of (46) is bounded for $t \geqslant 0$.

Proof. From (19), the function e that satisfies (45) can be expressed in the form

$$e(t) = G'_{(t\lambda_0^1,\ldots,t\lambda_0^\nu)}(u_0)[(\mathrm{e}^{t\widehat{L}} - I)\widehat{L}^{-1} l_M + t l^1],$$

where $-l_{r+1}(u_0) = l_M + l^1, l_M, \widehat{L}^{-1} l_M \in M, l^1 \in \mathrm{Ker} L$. By using (i) of lemma 2.3, we can write

$$l^1 = \sum_{i=1}^{\nu} \alpha_i g_i(u_0),$$

for some $\alpha_1, \ldots, \alpha_\nu$ and substituting into (44) we have

$$U_n = G_{(t_n\lambda_0^1,\ldots,t_n\lambda_0^\nu)}(u_0) + G'_{(t_n\lambda_0^1,\ldots,t_n\lambda_0^\nu)}(u_0)\left(h^r t_n \sum_{i=1}^{\nu} \alpha_i g_i(u_0)\right)$$
$$+ h^r G'_{(t_n\lambda_0^1,\ldots,t_n\lambda_0^\nu)}(u_0)(\mathrm{e}^{t_n\widehat{L}} - I)\widehat{L}^{-1} l_M + h^r Q(t_n, h). \tag{47}$$

Now the first two terms on the right-hand side of (47) can be written in the form

$$\begin{aligned} G_{(t_n\lambda_0^1,\ldots,t_n\lambda_0^\nu)}(u_0) &+ G'_{(t_n\lambda_0^1,\ldots,t_n\lambda_0^\nu)}(u_0)\left(h^r t_n \sum_{i=1}^{\nu} \alpha_i g_i(u_0)\right) \\ &= G_{(t_n\lambda_0^1,\ldots,t_n\lambda_0^\nu)}(u_0) + \sum_{i=1}^{\nu} h^r t_n \alpha_i g_i(G_{(t_n\lambda_0^1,\ldots,t_n\lambda_0^\nu)}(u_0)) \\ &= G_{(t_n\lambda_0^1,\ldots,t_n\lambda_0^\nu)}(u_0) + \sum_{i=1}^{\nu} h^r t_n \alpha_i \frac{\mathrm{d}}{\mathrm{d}\tau_i} G_{(\tau_1,\ldots,\tau_\nu)}(u_0)|_{\tau_i = t_n\lambda_0^i}, \end{aligned}$$

that differs from $G_{(t_n(\lambda_0^1+h^r\alpha_1),\ldots,t_n(\lambda_0^\nu+h^r\alpha_\nu))}(u_0)$ in $\mathrm{O}(h^{2r})$ terms, which can be hidden in the remainder, along with the function Q. The last part of the Theorem is proved by using (i) of lemma 2.3. □

Note that the numerical solution consists of three components: (i) a term $G_{(t_n(\lambda_0^1+h^r\alpha_1),\ldots,t_n(\lambda_0^\nu+h^r\alpha_\nu))}(u_0)$ that describes a transformation of u_0 by elements of the group. The velocities $\lambda_0^i + h^r\alpha_i$ differ in $\mathrm{O}(h^r)$ from the true velocities λ_0^i in the relative equilibrium. The differences $t_n h^r \alpha_i$ in parameters grow linearly with time. (ii) A *complementary term* $h^r G'_{(t_n\lambda_0^1,\ldots,t_n\lambda_0^\nu)}(u_0)(\mathrm{e}^{t_n\widehat{L}} - I)\widehat{L}^{-1} l_M$ which comprises contributions that while being of leading order cannot be interpreted as changes in the group parameters. This term, under suitable hypothesis, remains bounded. (iii) A third term $h^r q(h, t_n)$, that is an $\mathrm{o}(h^r)$ remainder.

The expansion (46) is not uniform, in the sense that the remainder $h^r q(h, t_n)$ will in general grow with time; a discussion of the growth of this remainder has been presented in the final section of [4].

3.2. The Hamiltonian case

In the Hamiltonian case, we have the following result, where $u(c), \lambda(c)$ are the mappings in lemma 2.4 (ii). The proof is similar to that of theorem 3.1 and will not be given.

Theorem 3.2. *Assume that (H1), (H2′), (H3′), (A1), (A2) hold and let u_0 be a relative equilibrium as in (36), (37) such that is a nondegenerate equilibrium of the reduced system. Then*

$$U_n = G_{(t_n\bar{\lambda}^1,\ldots,t_n\bar{\lambda}^\nu)}(u(c_0 + t_n h^r\theta)) + h^r G'_{(t_n\bar{\lambda}_0^1,\ldots,t_n\bar{\lambda}_0^\nu)}(u_0)(\mathrm{e}^{t_n\widehat{L}} - I)\widehat{L}^{-1} l_M + h^r q(h, t_n) \qquad (48)$$

for suitable $\theta = (\theta_1, \ldots, \theta_\nu), \alpha = (\alpha_1, \ldots, \alpha_\nu)$ and

$$\bar{\lambda}^i = \lambda^i\left(c_0 + \frac{t_n}{2}h^r\theta\right) + \alpha_i h^r, \qquad c_0^i = I_i(u_0), \qquad i = 1, \ldots, \nu.$$

The function q is a remainder that, for fixed t, tends to zero as $h \to 0$.

If u_0 is linearly stable as an equilibrium of the reduced system and the elements of the group are isometries, the second term of the right hand side of (48) is bounded for $t \geqslant 0$.

If $(\frac{\partial\lambda^i(c)}{\partial c_j}|_{c=c_0})_{i,j=1,\ldots,\nu}$ is invertible then $\theta = 0$ if and only if the method (43) satisfies the conditions

$$\nabla I_j(u_0)^T l_{r+1}(u_0) = 0,\ j = 1, \ldots, \nu. \qquad (49)$$

In particular, (49) holds if the method preserves exactly the invariant quantities $I_i(\psi_{h,f}(u_0)) = I_i(u_0)$.

Comparing this result with theorem 3.1 we see that now the error in velocities $\bar{\lambda}^i - \lambda^i$ that were bounded in time in theorem 3.1, now grow linearly. This leads to a quadratic growth in the parameters $t\lambda$.

Before closing this section we point out that it is straightforward to extend our analysis in several directions. As in [4], first we may have considered variable step sizes. Secondly, we may have considered not only the leading $\mathrm{O}(h^r)$ term of the expansion of the local error but all terms $\mathrm{O}(h^\sigma)$, $\sigma = r, \ldots, 2r - 1$, as these satisfy variational problems similar to that satisfied by the leading term.

4. Numerical experiments

We now show some numerical examples to illustrate the preceding results. We focus on the Hamiltonian case.

4.1. Test problem

We integrate numerically the Hamiltonian problem with four degrees of freedom and Hamiltonian function $H = T + V$.

$$\begin{aligned} T &= \tfrac{1}{2}(p_1^2 + p_2^2 + p_3^2 + p_4^2), \\ V &= -\frac{1}{\sqrt{q_1^2 + q_2^2}} - \frac{1}{\sqrt{q_3^2 + q_4^2}} - \frac{\epsilon}{\sqrt{(q_1 - q_3)^2 + (q_2 - q_4)^2}}, \end{aligned} \qquad (50)$$

where ϵ is a positive parameter. We are thus studying the planar motion of two bodies attracted to the origin and to each other with forces inversely proportional to the distance squared. The Hamiltonian H and the total angular momentum

$$M = p_2q_1 - p_1q_2 + p_4q_3 - p_3q_4$$

are conserved quantities of the problem. The invariant quantity M is the Hamiltonian function that generates the one-parameter group of rotations that maps $u = (p_1, p_2, p_3, p_4, q_1, q_2, q_3, q_4)$ into $G_\tau(u) = \mathcal{R}_\tau u$ where $\mathcal{R}_\tau$ is the block matrix

$$\mathcal{R}_\tau = \begin{pmatrix} R_\tau & 0 & 0 & 0 \\ 0 & R_\tau & 0 & 0 \\ 0 & 0 & R_\tau & 0 \\ 0 & 0 & 0 & R_\tau \end{pmatrix}, \qquad R_\tau = \begin{pmatrix} \cos\tau & -\sin\tau \\ \sin\tau & \cos\tau \end{pmatrix}.$$

Therefore, this is a group of symmetries of the equations of motion for (50).

In order to obtain the reduced system that will be required for analytical purposes, it is advisable to introduce polar coordinates (r_1, θ_1) and (r_2, θ_2) in the planes (q_1, q_2), (q_3, q_4) respectively. Then the Hamiltonian (50) becomes $H = T + V$ with

$$T = \frac{1}{2}\left(p_{r_1}^2 + p_{r_2}^2 + \frac{p_{\theta_1}^2}{r_1^2} + \frac{p_{\theta_2}^2}{r_2^2}\right),$$
$$V = -\frac{1}{r_1} - \frac{1}{r_2} - \frac{\epsilon}{\sqrt{r_1^2 + r_2^2 - 2r_1r_2\cos(\theta_1 - \theta_2)}},$$

where the momenta are

$$p_{r_i} = \dot{r}_i \qquad p_{\theta_i} = r_i^2\dot{\theta}_i, \qquad i = 1, 2.$$

Observe that in polar coordinates $M = p_{\theta_1} + p_{\theta_2}$. We now use the new set of canonical variables

$$\begin{aligned} p_x &= \tfrac{1}{2}(p_{\theta_1} + p_{\theta_2}), \qquad & x &= -(\theta_1 + \theta_2), \\ p_z &= \tfrac{1}{2}(p_{\theta_2} - p_{\theta_1}), \qquad & z &= (\theta_2 - \theta_1), \\ p_{r_1}, & & r_1&, \\ p_{r_2}, & & r_2&. \end{aligned}$$

(Note that the new momentum $p_x = M/2$ is a conserved quantity and that in this example $y = (z, r_1, r_2)$.) Now $H = T + V$ with

$$\begin{aligned} T &= \frac{1}{2}\left(p_{r_1}^2 + p_{r_2}^2 + \frac{(p_x - p_z)^2}{r_1^2} + \frac{(p_x + p_z)^2}{r_2^2}\right), \\ V &= -\frac{1}{r_1} - \frac{1}{r_2} - \frac{\epsilon}{\sqrt{r_1^2 + r_2^2 - 2r_1r_2\cos z}}, \end{aligned} \tag{51}$$

and the reduced Hamiltonian is obtained by setting $p_x = c$ in the last expression for H. It is easy to see that, at relative equilibria, z is a multiple of π, i.e. the two bodies are aligned with the origin.

4.2. Numerical methods

The numerical schemes being considered are as follows.

(RK) The classical third-order Runge–Kutta method with Butcher tableau [11]

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{2} & \frac{1}{2} & & \\ 1 & -1 & 2 & \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array}$$

which is chosen as an example of a nonconservative method, because it does not preserve either of the two invariants of the equations of motion for (50).

(V) Verlet's algorithm

$$\begin{aligned} p^{n+1/2} &= p^n + \frac{h}{2} f^n, \\ q^{n+1} &= q^n + h p^{n+1/2}, \\ p^{n+1} &= p^{n+1/2} + \frac{h}{2} f^{n+1}, \end{aligned}$$

where $p = (p_1, p_2, p_3, p_4)$, $q = (q_1, q_2, q_3, q_4)$, $f^n = (f_1^n, f_2^n, f_3^n, f_4^n)$ and the ith component of the force is given by $f_i^n = -V_{q_i}(q^n)$, $i = 1(1)4$. This method has order two, is time symmetric and conserves the momentum M but not the Hamiltonian [16].

(EC) The second-order method

$$p^{n+1/2} = p^n + \frac{h}{2} f^{n(+)}, \tag{52}$$

$$q^{n+1/2} = q^n + \frac{h}{4}(p^{n+1/2} + p^n), \tag{53}$$

$$p^{n+1} = p^{n+1/2} + \frac{h}{2} f^{n+1(-)}, \tag{54}$$

$$q^{n+1} = q^{n+1/2} + \frac{h}{4}(p^{n+1} + p^{n+1/2}), \tag{55}$$

where the components of the forces $f^{n(+)}$ and $f^{n+1(-)}$ are given by

$$\begin{aligned} f_1^{n(+)} &= -\frac{V(q_1^{n+1/2}, q_2^n, q_3^n, q_4^n) - V(q_1^n, q_2^n, q_3^n, q_4^n)}{q_1^{n+1/2} - q_1^n}, \\ f_2^{n(+)} &= -\frac{V(q_1^{n+1/2}, q_2^{n+1/2}, q_3^n, q_4^n) - V(q_1^{n+1/2}, q_2^n, q_3^n, q_4^n)}{q_2^{n+1/2} - q_2^n}, \\ f_3^{n(+)} &= -\frac{V(q_1^{n+1/2}, q_2^{n+1/2}, q_3^{n+1/2}, q_4^n) - V(q_1^{n+1/2}, q_2^{n+1/2}, q_3^n, q_4^n)}{q_3^{n+1/2} - q_3^n}, \\ f_4^{n(+)} &= -\frac{1}{q_4^{n+1/2} - q_4^n}(V(q_1^{n+1/2}, q_2^{n+1/2}, q_3^{n+1/2}, q_4^{n+1/2}) \\ &\qquad -V(q_1^{n+1/2}, q_2^{n+1/2}, q_3^{n+1/2}, q_4^n)), \\ f_1^{n+1(-)} &= -\frac{V(q_1^{n+1}, q_2^{n+1}, q_3^{n+1}, q_4^{n+1}) - V(q_1^{n+1/2}, q_2^{n+1}, q_3^{n+1}, q_4^{n+1})}{q_1^{n+1} - q_1^{n+1/2}}, \\ f_2^{n+1(-)} &= -\frac{V(q_1^{n+1/2}, q_2^{n+1}, q_3^{n+1}, q_4^{n+1}) - V(q_1^{n+1/2}, q_2^{n+1/2}, q_3^{n+1}, q_4^{n+1})}{q_2^{n+1} - q_2^{n+1/2}}, \end{aligned}$$

$$
\begin{aligned}
f_3^{n+1(-)} &= -\frac{1}{q_3^{n+1} - q_3^{n+1/2}}(V(q_1^{n+1/2}, q_2^{n+1/2}, q_3^{n+1}, q_4^{n+1}) \\
&\qquad -V(q_1^{n+1/2}, q_2^{n+1/2}, q_3^{n+1/2}, q_4^{n+1})), \\
f_4^{n+1(-)} &= -\frac{1}{q_4^{n+1} - q_4^{n+1/2}}(V(q_1^{n+1/2}, q_2^{n+1/2}, q_3^{n+1/2}, q_4^{n+1}) \\
&\qquad -V(q_1^{n+1/2}, q_2^{n+1/2}, q_3^{n+1/2}, q_4^{n+1/2}).
\end{aligned}
$$

Note that the first component of the vector equation (52) and the first component of (53) give a system of two scalar equations for $p_1^{n+1/2}, q_1^{n+1/2}$. Once $p_1^{n+1/2}, q_1^{n+1/2}$ are known, the second components of (52), (53) give two scalar equations for $p_2^{n+1/2}, q_2^{n+1/2}$ etc. To solve (54), (55) one considers the components in reverse order and successively finds (p_4^{n+1}, q_4^{n+1}), (p_3^{n+1}, q_3^{n+1}) etc. The idea behind EC, i.e. to obtain force values by numerically differentiating the potential can be found in [12, 18] and is used to ensure conservation of the Hamiltonian. The scheme EC is time symmetric but does not exactly conserve momentum.

4.3. Numerical results

We use the methods above to approximate some solutions of the problem (50) described in section 4.1. The validity of the conditions (A1), (A2) for the schemes can be easily proved. Our aim is to see the difference in error propagation. We do not try to compare the efficiency of the methods nor imply that any of these three methods is a practical method for the problem at hand.

4.3.1. Relative equilibrium case. We start our study of the error propagation by taking the relative equilibrium (in Cartesian coordinates)

$$
u_0 = (0, \lambda, 0, -\lambda, 1, 0, -1, 0), \lambda^2 = 1 + \frac{\epsilon}{4}. \qquad (56)
$$

In the corresponding solution the two bodies move with angular velocity λ around a circumference of unit radius. An easy analysis based on (51) shows that this relative equilibrium is linearly stable for $0 \leqslant \epsilon < \frac{1}{2}$. Taking $\epsilon = 0.1$ we integrate with the three methods up to 100 periods of time $t = 100T$, $T = 2\pi/\lambda$ and step sizes $h = T/1280, T/2560, T/5120$. Figure 1 gives, in a log–log scale, the Euclidean norm of the global error as a function of time, with the full lines corresponding to EC, the broken lines to V and the dotted lines to RK; plotted is the error at the end of every period. The distance between parallel lines corresponding to a given method shows the $\mathrm{O}(h^2)$ behaviour of the errors of the Verlet's method and EC method, while for RK errors behave as $\mathrm{O}(h^3)$. For fixed t, the third-order scheme gives smaller errors than the other methods, but if we focus on the error propagation, we see that for the V and EC methods, errors grow like t and, for RK, they grow like t^2. This confirms the results stated in theorem 3.2. For the leading term l of the local error of the third-order method, $\nabla M(u_0)^T \cdot l(u_0) \neq 0$, which leads to the quadratic growth. In the case of Verlet's method, the conservation of M assures the orthogonality condition (49) and therefore the linear growth. Since EC conserves the Hamiltonian (52), the leading term l of its local error at the relative equilibrium is orthogonal to ∇H at this point and hence to ∇M as, at relative equilibria, ∇H and ∇M are parallel (see (14)).

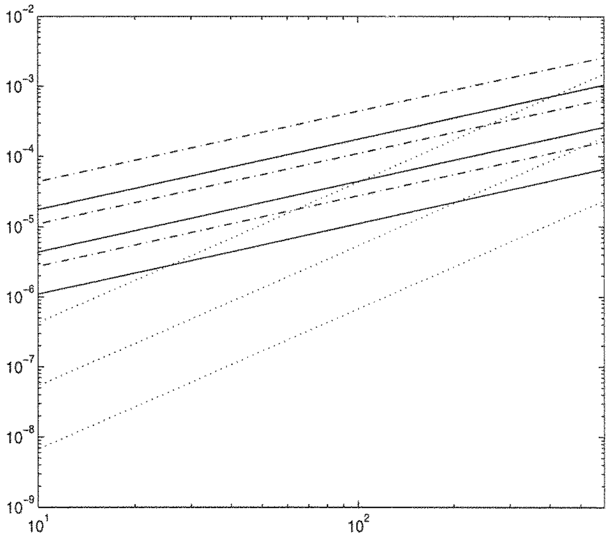

Figure 1. Error as a function of time in the integration of a stable relative equilibrium. The full lines correspond to EC, the chain lines to V and the dotted lines to RK.

4.3.2. Perturbation of the relative equilibrium. We next take the initial condition

$$u_0 = (0, \lambda, 0, -\lambda, 1 + \epsilon', 0, -1, 0), \qquad \lambda^2 = 1 + \frac{\epsilon}{4}, \tag{57}$$

with $\epsilon = 0.1$, $\epsilon' = 1E - 04$, i.e. we take a small perturbation of the relative equilibrium (56). We integrated the problem with a standard variable step code up to $t = 200$ and sufficiently small tolerance to obtain an 'exact' solution. The error propagation for this case can be seen in figure 2. No significant change is observed with respect to figure 1, so that the better error propagation that we have proved for relative equilibria also holds for neighbouring solutions as one may have expected. In figure 2 and in later figures the error has been plotted at intervals of unit length.

4.3.3. Instability. We can also study unstable relative equilibria. We take an initial condition of the form (57) but with $\epsilon = 1$. Figure 3 gives the behaviour in time of the error for V (chain curves) and for RK (dotted curves). The results for the EC method are similar to those of V and we do not include them. Note that, for all methods, errors grow, eventually, in an exponential fashion.

4.3.4. Asymptotically uncoupled motion. The initial condition

$$u_0 = (6, 4, 1, 0, -1, 0, 1, 1),$$

with $\epsilon = 1$ gives rise to a solution in which, asymptotically as $t \uparrow \infty$, one body describes a Keplerian elliptic orbit around the origin, while the other approaches infinity at a constant velocity. The system behaves for large t as two uncoupled Kepler problems, one with positive energy and the other with negative energy. Figure 4 displays the global errors for the three methods. Observe that, after a transient with exponential growth, errors for the RK scheme grow quadratically and, in the case of the second-order methods, they grow linearly, as if we were really integrating a Kepler problem [3].

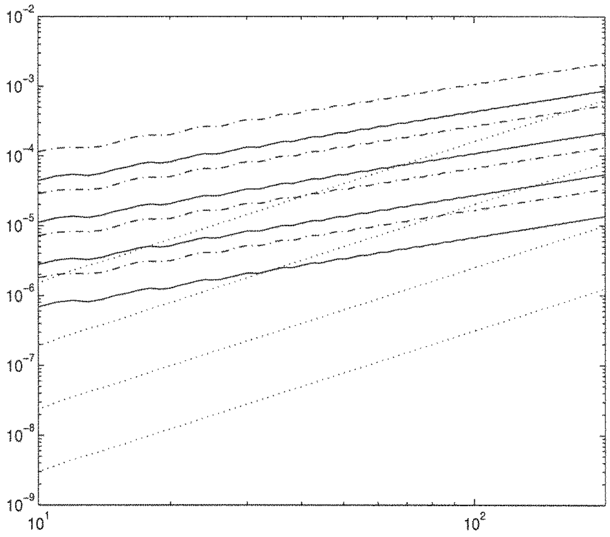

Figure 2. Error as a function of time near a stable relative equilibrium. The full lines correspond to EC, the chain lines to V and the dotted lines to RK.

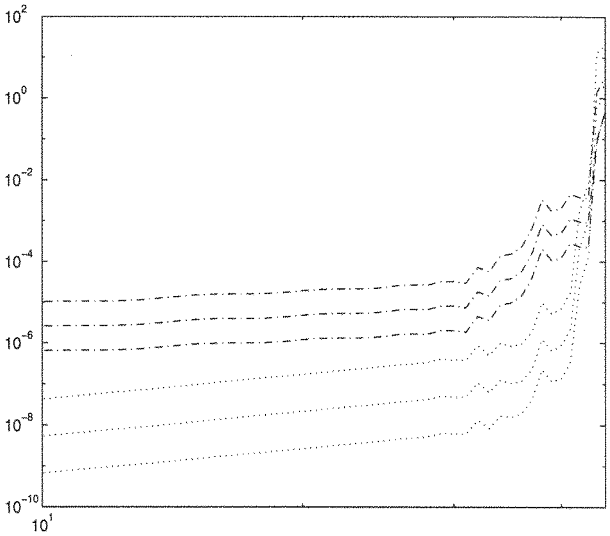

Figure 3. Error as a function of time near an unstable relative equilibrium. The chain curves correspond to V and the dotted curves to RK.

4.3.5. A symmetric solution Finally, we show an example for which the two second-order methods behave in very different ways. For the initial condition

$$u_0 = (1, 0, -1, 0, 1, 1, -1, -1),$$

($\epsilon = 0.1$) the solution represents the two bodies moving symmetrically with respect to the origin

$$q_1(t) = -q_3(t), \qquad q_2(t) = -q_4(t), \qquad p_1(t) = -p_3(t), \qquad p_2(t) = -p_4(t). \quad (58)$$

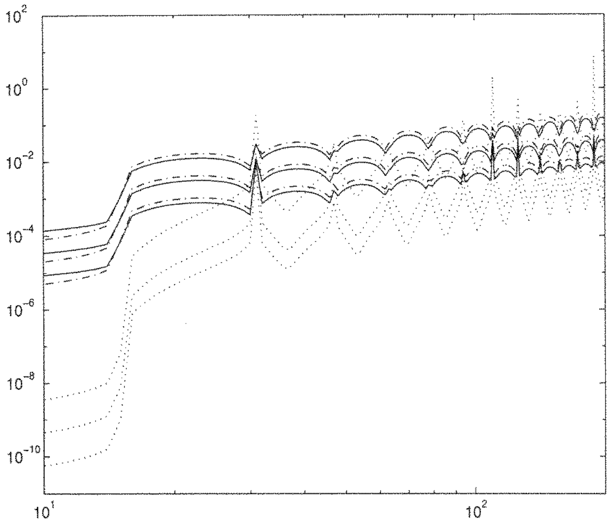

Figure 4. Error as a function of t for a case where for large t one body describes a Keplerian ellipse and the other approaches infinity. The full curves correspond to EC, the chain curves to V and the dotted curves to RK.

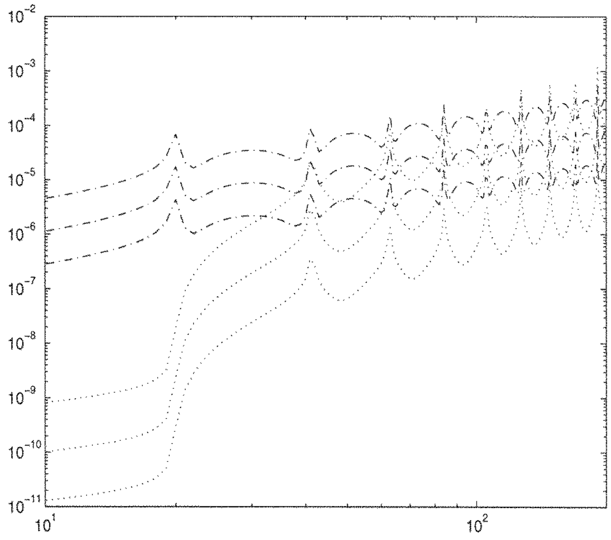

Figure 5. Error as a function of t for a solution where both bodies describe symmetric solutions. The chain curves corresponds to V and the dotted curves to RK.

Each body describes a Keplerian ellipse with focus at the origin. Due to the symmetry (58), the eight equations of motion could be reduced to four equations for q_1, q_2, p_1, p_2. The error propagation for this case is shown in figures 5 and 6. Figure 5 corresponds to RK and V, while errors for EC are displayed in figure 6. The RK and V schemes (and any other Runge–Kutta or partitioned Runge–Kutta method) have the property that if at one step $q_1^n = -q_3^n, q_2^n = -q_4^n, p_1^n = -p_3^n, p_2^n = -p_4^n$ then the same is true at the next step. They thus behave as if they were integrating a Kepler problem for one of the bodies. Therefore,

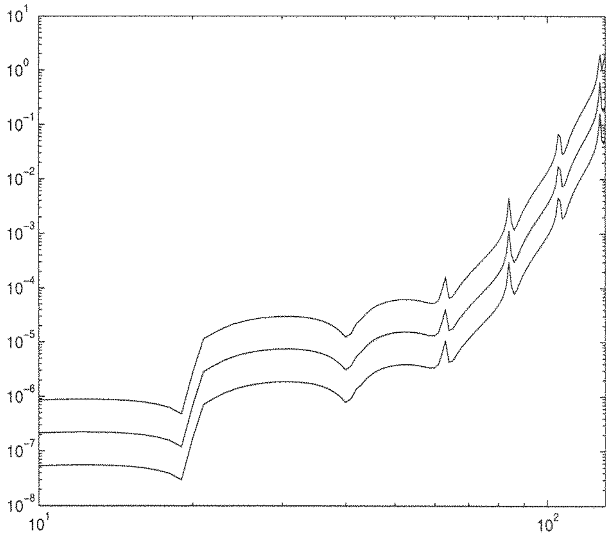

Figure 6. Error as a function of t for a solution where both bodies describe symmetric solutions. The integrator is EC.

we have quadratic growth for the nonconservative scheme and linear growth for Verlet's method [4]. But figure 6 shows a very different behaviour in the case of EC. This method does not preserve the symmetry of the solution: it is clear from (52)–(55) that q_1 and q_3, q_2 and q_4 do not play a symmetric role in the algorithm. The lack of symmetry in the local error triggers an exponential growth.

Acknowledgment

This research has been supported by project no PB95-705.

References

[1] Arnold V I 1989 *Mathematical Methods of Classical Mechanics* (Berlin: Springer)

[2] Calvo M P and Hairer E 1995 Accurate long-term integration of dynamical systems *Appl. Numer. Math.* **18** 95–105

[3] Calvo M P and Sanz-Serna J M 1993 The development of variable-step symplectic integrators, with application to the two-body problem *SIAM J. Sci. Comput.* **14** 936–52

[4] Cano B and Sanz-Serna J M 1997 Error growth in the numerical integration of periodic orbits, with application to Hamiltonian and reversible systems *SIAM J. Numer. Anal.* **34** 1391–417

[5] Cano B and Sanz-Serna J M 1998 Error growth in the numerical integration of periodic orbits by multistep methods, with application to reversible systems *IMA J. Numer. Anal.* **18** 57–75

[6] Durán A 1997 Propagación del error en la integración numérica de la ecuación no lineal de Schroedinger *PhD Thesis* Universidad de Valladolid, Spain

[7] Durán A and Sanz-Serna J M 1998 The numerical integration of relative equilibrium solutions. The nonlinear Schrödinger equation, in preparation

[8] Estep D J and Stuart A M 1995 The rate of error growth in Hamiltonian-conserving integrators *Z. Angew. Math. Phys.* **46** 407–18

[9] Frutos J de and Sanz-Serna J M 1997 Accuracy and conservation properties in numerical integration: the case of the Korteweg–de Vries equation *Numer. Math.* **75** 421–45

[10] Hairer E and Lubich Ch 1997 The life-span of backward error analysis for numerical integrators *Numer. Math.* **76** 441–62

[11] Hairer E, Nørsett S P and Wanner W G 1993 *Solving Ordinary differential equations I, Nonstiff problems* 2nd edn (Berlin: Springer)

[12] Itoh T and Abe K 1988 Hamiltonian-conserving discrete canonical equations based on variational difference quotients *J. Comput. Phys.* **77** 85–102

[13] Marsden J E and Ratiu T S 1994 *Introduction to Mechanics and Symmetry* (New York: Springer)

[14] Olver P J 1986 *Applications of Lie Groups to Differential Equations* (New York: Springer)

[15] Sanz-Serna J M 1997 Geometric integration *The State of the Art in Numerical Analysis* ed I S Duff and G A Watson (Oxford: Clarendon) pp 121–43

[16] Sanz-Serna J M and Calvo M P 1994 *Numerical Hamiltonian Problems* (London: Chapman and Hall)

[17] Stoffer D 1995 Variable steps for reversible integration methods *Computing* **55** 1–22

[18] Strauss W A and Vázquez L 1978 Numerical solution of a nonlinear Klein-Gordon equation *J. Comput. Phys.* **28** 271–8