ELSEVIER

Journal of Computational and Applied Mathematics 115 (2000) 503–517

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Conserving first integrals under discretization with variable step size integration procedures

Johannes Schropp

Department of Mathematics and Computer Science, University of Konstanz, P.O. Box 5560, D-78434 Konstanz, Germany

Received 10 August 1998; received in revised form 10 May 1999

Abstract

It is well known that the application of one-step or linear multistep methods to an ordinary differential equation with first integrals will destroy the conserving quantities. With the use of stabilization techniques similar to Ascher, Chin, Reich (Numer. Math. 67 (1997) 131–149) we derive stabilized variants of our original problem. We show that variable step size one-step and linear multistep methods applied to the stabilized equation will reproduce that phase portrait correctly. In particular, this technique will conserve first integrals over an infinite time interval within the local error of the used method. © 2000 Elsevier Science B.V. All rights reserved.

1. Introduction

We consider an autonomous smooth ordinary differential equation

$$\dot{x} = f(x), \tag{1.1}$$

on $\mathbb{R}^N$ which possesses $0 < l < N$ first integrals $g = (g_1, \ldots, g_l)$, that is,

$$g(\phi(t, x_0)) = g(x_0), \quad t \in J(x_0),\ x_0 \in \mathbb{R}^N. \tag{1.2}$$

Here $\phi(t, x_0)$ denotes the solution of (1.1) with the initial value $\phi(0, x_0) = x_0$ on its open maximal interval of existence $J(x_0)$.

One of the most interesting questions one could ask about a dynamical system is: What are the properties of its solution flow in the longtime run?

From that point of view, for problem (1.1), (1.2) it is natural to provide discretization methods which conserve the quantities g_i, $i = 1, \ldots, l$. For constant step sizes the reader may find one-step discretizations in [1,11]. In the book of Stuart and Humphries [11] the reader will also find results for linear multistep methods with constant step size. But if one is interested in longtime properties,

PII: S0377-0427(99)00178-8

efficient variable step size discretizations of (1.1), (1.2) which conserve the first integrals $g=(g_1,\dots,g_l)$ are much more appropriate. Such one-step and linear multistep methods can be established and its our purpose here to show how.

The reader may keep in mind that as an interesting special case Hamiltonian systems fit into framework (1.1),(1.2). For such systems an enormous research characterizing numerical methods that preserve the symplectic Hamiltonian structure has been undertaken. An excellent overview is presented in Sanz-Serna [9]. For constant step size these symplectic methods work excellent but, unfortunately, not for variable step size (see, e.g. [2]). Recently, however, Hairer [4] presented a partial solution of the step size problem using Poincare transformations.

Since it is well known that first integrals of ordinary differential equations do not persist under discretization, due to Ascher et al. [1] we change the vectorfield of (1.1) to stabilize the invariant set $M_{x_0}=\{x\in\mathbb{R}^N \mid g(x)=g(x_0)\}$. To be more precise, we replace (1.1) by a differential equation which possesses M_{x_0} as globally exponentially attractive invariant set. Then we can apply the techniques of Kloeden and Lorenz [7,8] to obtain results for constant step size one-step, respectively, linear multistep methods. We will develop our conserving results by extending the Kloeden and Lorenz [7,8] tools to variable step size numerical methods. Finally, we apply our methods to the restricted three-body problem in physics and compare it with the classical approach.

2. The main results

We consider a smooth initial value problem

$$\dot{x}=f(x),\quad x(0)=x_0, \tag{2.1}$$

on $\mathbb{R}^N$ with solution flow $\phi(t,x_0)$. To discretize (2.1), we choose an arbitrary grid $[t_n]_{n\in\mathbb{N}}$, $t_0=0$ and denote the step sizes by $h_n=t_{n+1}-t_n$. We assume that the infimal, respectively, supremal step size satisfy

$$h_{inf}:=\inf\{h_j \mid j\in\mathbb{N}\}>0,$$
$$h_{sup}:=\sup\{h_j \mid j\in\mathbb{N}\}<\infty \text{ and sufficiently small.}$$

Furthermore, we denote by $y_n=y(t_n)$ the approximations to the exact solution $\phi(t_n,x_0)$. Then our one-step method reads

$$y_{n+1}=y_n+h_nV(h_n,y_n),\quad n=0,1,2,\dots,\ y_0=x_0 \tag{2.2}$$

for an appropriate V. Following Hairer et al. [5, p. 401] we write the variable step size linear k-step method in the form

$$y_{n+k}+\sum_{j=0}^{k-1}\alpha_{jn}y_{n+j}=h_{n+k-1}\sum_{j=0}^{k}\beta_{jn}f(y_{n+j}),\quad n=0,1,2,\dots\ . \tag{2.3}$$

The values α_{jn}, $j=0,\dots,k-1$, β_{jn}, $j=0,\dots,k$ here depend on the ratios $\omega_i=h_i/(h_{i-1})$, $i=n+1,\dots,n+k-1$. We assume henceforth that the α_{jn}, β_{jn} are bounded. The linear multistep method (2.3) is then completed by a starting procedure

$$y_{n+1}=y_n+h_nV(h_n,y_n),\quad n=0,\dots,k-2,\ y_0=x_0. \tag{2.4}$$

It is well known, that if f is globally lipschitzian and $h_{\sup}$ small enough, the solution of Eq. (2.3) has the form

$$y_{n+k} = g_{n+k-1}(y_n, \dots, y_{n+k-1}) := -\sum_{i=0}^{k-1} \alpha_{in} y_{n+i} + h_{n+k-1}\psi_{n+k-1},$$

where ψ_{n+k-1} solves

$$\psi_{n+k-1} = \sum_{i=0}^{k-1} \beta_{in} f(y_{n+i}) + \beta_{kn} f\left(h_{n+k-1}\psi_{n+k-1} - \sum_{i=0}^{k-1} \alpha_{in} y_{n+i}\right).$$

To guarantee stability of a variable step size linear multistep method, according to [5, p. 403], we need the following additional stability conditions:

(i) $1 + \sum_{j=0}^{k-1} \alpha_{jn} = 0$.
(ii) The values $\alpha_{jn} = \alpha_j(\omega_{n+1}, \dots, \omega_{n+k-1})$ are continuous in a neighbourhood of $(1, \dots, 1)$.
(iii) The underlying constant step size formula $q(\xi) = \xi^k + \sum_{j=0}^{k-1} \alpha_j(1, \dots, 1)\xi^j$ satisfies the strong root condition, i.e., $q(\xi)$ possesses 1 as simple zero and all other zeroes $\bar{\bar{\xi}}$ of q satisfy $|\bar{\bar{\xi}}| < 1$.

Then, Theorem 5.5, Ch. III.5 in [5] assures the existence of real numbers $\omega_d < 1 < \omega_u$ such that the method is stable if $\omega_d \leqslant h_n/h_{n-1} \leqslant \omega_u$ for $n \in \mathbb{N}$.

In addition, we shall assume that the function f in Eq. (2.1) is $p \geqslant 1$ times continuously differentiable and that methods (2.2) and (2.3),(2.4) are of order p. In the one-step case we can establish the inequality

$$\| \phi(h,x) - (x + hV(h,x)) \| \leqslant C_p h^{p+1} \quad \text{for } x \in \Omega,\ 0 < h \leqslant h_{\sup}, \tag{2.5}$$

$\Omega \subset \mathbb{R}^N$ bounded and in the multistep case we obtain

$$\left\| \phi(h_{n+k-1}, x) - g_{n+k-1}\left(\phi\left(-\sum_{j=1}^{k-1} h_{n+j-1}, x \right), \dots, \phi(-h_{n+k-2}, x), x \right) \right\|_2 \leqslant C_p h_{n+k-1}^{p+1} \tag{2.6}$$

for step sizes $0 < h_{n+j} < h_{\sup}$, $j = 0, \dots, k-1$, $n \in \mathbb{N}$ and $x \in \Omega \subset \mathbb{R}^N$, Ω a bounded set.

Note that by definition g_{n+k-1} depends on $h_n, h_{n+1}, \dots, h_{n+k-1}$. But with $0 < h_{inf} \leqslant h_n \leqslant h_{sup}$ for $n \in \mathbb{N}$ and $j = 0, \dots, k-1$ the quotients h_{n+j}/h_n remain bounded for $n \in \mathbb{N}$ and $j = 0, \dots, k-1$ and, hence, an upper bound of the local error only depending on h_{n+k-1} can be established (see, e.g., Hairer, Nørsett, Wanner [5], p. 401).

We now focus our interest to the smooth autonomous initial value problem (2.1) with solution flow $\phi(t, x_0)$. We assume that (2.1) possesses $0 < l < N$ first integrals $g = (g_1, \dots, g_l) \in C^{p+1}(\mathbb{R}^N, \mathbb{R}^l)$, that is,

$$g(\phi(t, x_0)) = g(x_0), \quad t \in J(x_0),\ x_0 \in \mathbb{R}^N. \tag{2.7}$$

To stabilize the invariant set

$$M_{x_0} = \{x \in \mathbb{R}^N \mid g(x) = g(x_0)\}, \tag{2.8}$$

we proceed as follows. Let $\tau > 0$, let

$$D_\tau := \{x \in \mathbb{R}^N \mid \ \| g(x) - g(x_0) \|_2 < \tau\} \tag{2.9}$$

and assume that $Dg(x)$ has full rank on $\bar{D}_\tau$. For a smooth family $A(x)$, $x \in D_\tau$ of $(l \times l)$-matrices we consider the initial value problem

$$\dot{x} = T(x) := f(x) - Dg(x)^{\mathrm{T}} A(x)(g(x) - g(x_0)), \quad x(0) = \hat{x} \in D_\tau. \tag{2.10}$$

Then, the stabilized variant of Eq. (2.1) reads as follows.

Lemma 2.1. *We consider the initial value problem* (2.10) *and assume that D_τ is bounded. Let $B(x) := Dg(x)Dg(x)^{\mathrm{T}} A(x)$ and let $\mu_2(-B(x)) \leqslant -\eta$, $x \in D_\tau$, $\eta > 0$ hold. Then, every solution $\phi(t, \hat{x})$ of* (2.10) *exists for all $t \geqslant 0$. Moreover, M_{x_0} from* (2.8) *is a globally, exponentially attractive invariant set with attraction exponent η for the equation $\dot{x} = T(x)$.*

Here $\mu_2(C)$ denotes the logarithmic norm $\mu_2(C) := \lim_{\delta \to 0} 1/\delta(\| I + \delta C \|_2 - 1)$ of a matrix C, and can be expressed in terms of the spectrum of C by $\mu_2(C) = \max\{\lambda \in \mathbb{R} \mid \lambda \in \sigma(\frac{1}{2}(C + C^{\mathrm{T}}))\}$ (see, e.g., [3, p. 41]).

Remark. Natural possible choices are $A(x) = (Dg(x)Dg(x)^{\mathrm{T}})^{-1}$ or $A(x) = I$.

Now we are in the position to formulate our main result.

Theorem 2.2. *Let the assumptions of Lemma* 2.1 *hold and let $\hat{\tau} < \tau$. We assume $\Gamma(x) := \| g(x) - g(x_0) \|_2$ to be lipschitzian with L. Then for $h_{\sup} > 0$ sufficiently small, the iterates y_n exist for $n \in \mathbb{N}$ and belong to $D_{\hat{\tau}}$* when $y_0 = \hat{x} \in D_{\hat{\tau}}$ *for a* pth *order variable step size one-step or linear multistep method applied to* (2.10), *provided assumptions* (i)–(iii) *hold in the multistep case. The set*

$$M(h_{\sup}) = \left\{ x \in D_{\hat{\tau}} \mid \| g(x) - g(x_0) \|_2 \leqslant \frac{2LC_p h_{\sup}^{p+1}}{1 - \exp(-\eta h_{\sup})} \right\}$$

is positive invariant and globally attractive for the discrete dynamical system. Furthermore, with $\hat{x} = x_0$ the estimate

$$\| g(y_n) - g(y_0) \|_2 \leqslant C h_{\sup}^p, \quad n \in \mathbb{N}$$

holds for the conserving quantities g.

Theorem 2.2 states that the quantities g are preserved forever within the magnitude of the local error.

Remark. An inequality of the type $\Gamma(y_{n+1}) \leqslant C h_n^{p+1} + \exp(-\eta h_n)\Gamma(y_n), n \in \mathbb{N}$ (see formula (3.4)) will be established in the process of proving Theorem 2.2. Thus, it is obvious how to generalize the results of Kloeden and Lorenz [7,8] on the behaviour of stable sets under discretization to variable step size one-step or linear multistep methods (compare [8, formula (1.5)]).

3. Proof of the main results

Proof of Lemma 2.1. For $\hat{x} \in D_\tau$ arbitrary we define $r : J(\hat{x}) \to \mathbb{R}^l$ via $r(t) := g(\phi(t,\hat{x}) - g(x_0)$. Then, $r(t)$ solves the initial value problem

$$\dot{u} = -Dg(\phi(t,\hat{x}))Dg(\phi(t,\hat{x}))^{\mathrm{T}}A(\phi(t,\hat{x}))u = -B(\phi(t,\hat{x}))u,$$
$$u(0) = g(\hat{x}) - g(x_0).$$

Now, using Theorem 5.1.3 in [10] with $v = 0$, we obtain

$$\| r(t) \|_2 \leqslant \| r(0) \|_2 \exp(-\eta t), \quad t \in [0, t^+(\hat{x})[, \ \hat{x} \in D_\tau, \tag{3.1}$$

since $\mu_2(-B(\phi(t,\hat{x}))) \leqslant -\eta$ holds.

A direct consequence of (3.1) is that the compact set $C(\hat{x}) := \{x \in D_\tau \mid \| g(x) - g(x_0) \|_2 \leqslant \| g(\hat{x}) - g(x_0) \|_2\}$ is positive invariant for every initial value $\hat{x} \in D_\tau$. Hence, $t^+(\hat{x}) = \infty$ follows for $\hat{x} \in D_\tau$. Moreover, the set $M_{x_0} = \{x \in D_\tau \mid g(x) = g(x_0)\}$ is globally, exponentially attractive for the dynamics of $\dot{x} = T(x)$.

In our next step we show the existence of the discrete dynamics.

Lemma 3.1. *Let the assumptions of Theorem* 2.2 *be fulfilled, and let $\hat{\tau} < \tau$. Then, for $h_{\sup} > 0$ small enough, the variable step size one-step method* (2.2) *as well as the variable step size linear multistep method* (2.3), (2.4) *is defined for all $n \in \mathbb{N}$.*

Proof. Let $\delta > 0$ be defined by

$$\delta := \inf\{\| u - v \|_2 \mid u \in \overline{D_{\hat{\tau}}}, v \in \mathbb{R}^N \setminus D_\tau\}$$

and let $\tilde{T}$ denote the trivial extension of T onto $\mathbb{R}^N$ via 0 on $\mathbb{R}^N \setminus D_\tau$. With the smoothing function $\beta \in C^\infty(\mathbb{R}^N, [0,1])$, $\beta(x) = 1$ if $\mathrm{dist}(x, \overline{D_{\hat{\tau}}}) \leqslant \delta/3$, $\beta(x) = 0$ if $\mathrm{dist}(x, \mathbb{R}^N \setminus D_\tau) \leqslant \delta/3$ (see [6, Section 2.2, Exercise 1]) we define $\hat{T} \in C^p(\mathbb{R}^N, \mathbb{R}^N)$ via $\hat{T}(x) := \beta(x)\tilde{T}(x)$. Then $y_n \in D_{\hat{\tau}} \to y_{n+1} \in D_\tau$ holds. This follows from the inequality

$$\| y_{n+1} - y_n \|_2 \leqslant \| y_{n+1} - \phi(h_n, y_n) \|_2 + h_n \| T(y_n) \|_2 + \| \phi(h_n, y_n) - y_n - h_n T(y_n) \|_2 = \mathrm{O}(h_n). \tag{3.2}$$

The last term on the right-hand side in (3.2) is $\mathrm{O}(h_n^2)$ and the first term is $\mathrm{O}(h_n^{p+1})$. This follows from consistency condition (2.5) in the one-step case and it will be shown under a lipschitz condition for the vectorfield in the multistep case in Lemma A.1 in the appendix. Moreover, using bump function techniques the Lipschitz condition in Lemma A.1 is obviously verified in applications.

For $y_n \in D_{\hat{\tau}}$ and $\Gamma(x) = \| g(x) - g(x_0) \|_2$ we can calculate

$$\begin{aligned}
\Gamma(y_{n+1}) &\leqslant |\Gamma(y_{n+1}) - \Gamma(\phi(h_n, y_n))| + \Gamma(\phi(h_n, y_n)) \\
&\leqslant L \| y_{n+1} - \phi(h_n, y_n) \|_2 + \exp(-\eta h_n)\Gamma(y_n) \\
&\leqslant LCh_n^{p+1} + \exp(-\eta h_n)\hat{\tau} \\
&\leqslant (1 - \exp(-\eta h_n))\hat{\tau} + \exp(-\eta h_n)\hat{\tau} = \hat{\tau},
\end{aligned}$$

possibly after diminishing $h_{\sup} > 0$.

Proof of Theorem 2.2. Let the sequence $(y_n)_{n\in\mathbb{N}}$ be generated via applying a variable step size one-step method (2.2) of order p or a variable step size pth-order linear multistep method onto Eq. (2.10). Moreover, we assume

$$\| y_{n+1} - \phi(h_n, y_n) \| \leqslant Ch_n^{p+1}, \quad n \in \mathbb{N}. \tag{3.3}$$

Eq. (3.3) follows directly from the consistency condition (2.5) in the one-step case and will be proven in Lemma $A1$ for the linear multistep case.

Now we show that the quantities g are preserved with in the magnitude of the local error of the used method. With the Lipschitz constant L of Γ and C from (3.3), following [7], we define

$$\sigma(h) := \frac{2LCh^{p+1}}{1-\exp(-\eta h)}$$

and the discrete analog of M_{x_0} to be

$$M(h) = \{x \in D_{\hat{\tau}} \mid \Gamma(x) \leqslant \sigma(h)\}.$$

The reader may keep in mind that $M(0) = M_{x_0}$ holds and that $M(h)$ is a small neighbourhood of M_{x_0} for $h > 0$. Then we can compute

$$\begin{aligned} \Gamma(y_{n+1}) &\leqslant |\Gamma(y_{n+1}) - \Gamma(\phi(h_n, y_n))| + \Gamma(\phi(h_n, y_n)) \\ &\leqslant LCh_n^{p+1} + \exp(-\eta h_n)\Gamma(y_n) \\ &= \tfrac{1}{2}\sigma(h_n)(1-\exp(-\eta h_n)) + \exp(-\eta h_n)\Gamma(y_n) \\ &\leqslant \tfrac{1}{2}\sigma(h_{\text{sup}})(1-\exp(-\eta h_n)) + \exp(-\eta h_n)\Gamma(y_n), \quad n \in \mathbb{N}. \end{aligned} \tag{3.4}$$

Now, let $y_n \in M(h_{\text{sup}})$, that is, $\Gamma(y_n) \leqslant \sigma(h_{\text{sup}})$. We obtain

$$\Gamma(y_{n+1}) \leqslant \sigma(h_{\text{sup}})\tfrac{1}{2}(1-\exp(-\eta h_n)) \leqslant \sigma(h_{\text{sup}}).$$

Thus, $y_{n+1} \in M(h_{\text{sup}})$ follows.

On the other hand, we consider the case $y_n \notin M(h_{\text{sup}})$, that is, $\Gamma(y_n) > \sigma(h_{\text{sup}})$. A simple calculation shows

$$\begin{aligned} \Gamma(y_{n+1}) &\leqslant \tfrac{1}{2}\Gamma(y_n)(1+\exp(-\eta h_n)) \\ &\leqslant \Gamma(y_n)\exp(-\eta h_n/4) \quad \text{for } h_n \leqslant h_{\text{sup}} \text{ sufficiently small.} \end{aligned}$$

Assuming $y_n, y_{n-1}, \ldots, y_1, y_0 = \hat{x} \notin M(h_{\text{sup}})$ we obtain inductively

$$\Gamma(y_{n+1}) \leqslant \Gamma(y_0)\exp\left(-\eta\left(\sum_{i=0}^{n} h_i\right)\bigg/ 4\right).$$

Since $\exp(-\eta(\sum_{i=0}^{n} h_i)/4)$ tends to zero as $n \to \infty$, there exists an $\bar{n} \in \mathbb{N}$ such that $y_n \in M(h_{\text{sup}})$ for $n \geqslant \bar{n}$.

For the important special case $y_0 = x_0$ we finally obtain

$$\| g(y_n) - g(y_0) \|_2 = \Gamma(y_n) \leqslant \sigma(h_{\text{sup}}) \leqslant Ch_{\text{sup}}^{p} \quad \forall n \in \mathbb{N}.$$

It remains to prove formula (3.3) in the multistep case. This will be done in the appendix.

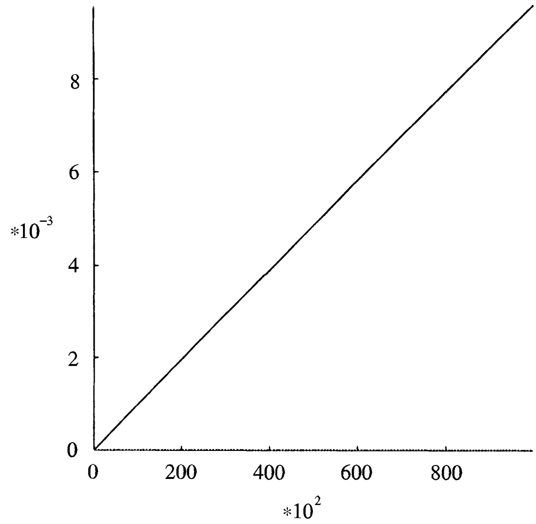

Fig. 1. $|g(\phi(t,x(0)^1)) - g(x(0)^1)|$ vs. t (for $x(0)^2$ same picture). Without stabilization.

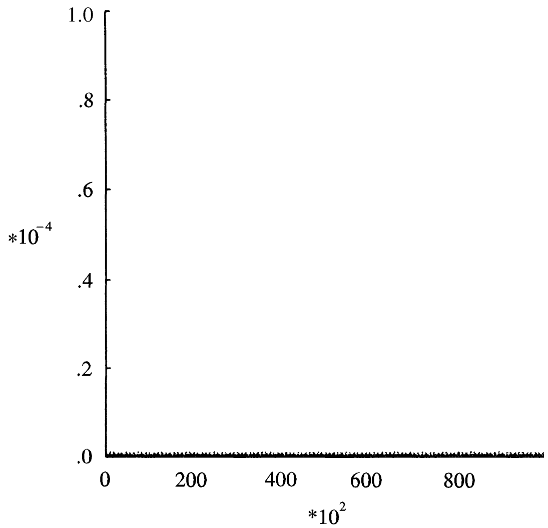

Fig. 2. $|g(\phi(t,x(0)^1)) - g(x(0)^1)|$ vs. t (for $x(0)^2$ same picture). With stabilization.

4. Numerical applications

In this section we apply our stabilization techniques to the restricted three-body problem from physics. There one considers the motion of three bodies in their gravitational field under the following simplyfying assumptions:

- The motion of the three bodies takes place in a plane.
- Two of the three bodies rotate on a circle with the same period.
- The mass of the third body is zero.

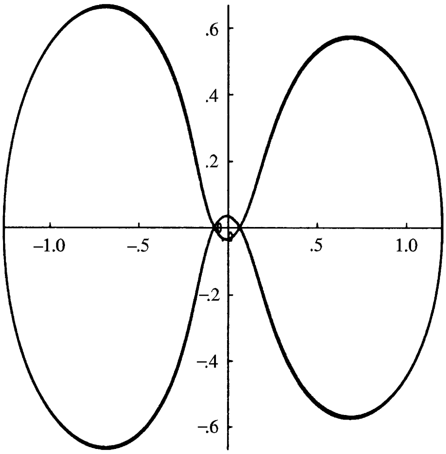

Fig. 3. Phase portrait of the trajectory starting at $x(0)^1$. Without stabilization.

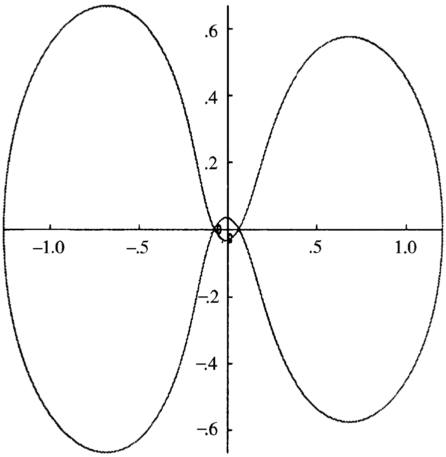

Fig. 4. Phase portrait of the trajectory starting at $x(0)^1$. With stabilization.

Then the equations of motion for the third body read

$$\begin{aligned}
\dot{x}_1 &= x_3, \qquad \dot{x}_2 = x_4,\\
\dot{x}_3 &= x_1 + 2x_4 - (1-\mu)\frac{x_1+\mu}{r_1^3} - \mu\frac{x_1-1+\mu}{r_2^3},\\
\dot{x}_4 &= x_2 - 2x_3 - (1-\mu)\frac{x_2}{r_1^3} - \mu\frac{x_2}{r_2^3}
\end{aligned} \tag{4.1}$$

with

$$r_1 = \sqrt{(x_1+\mu)^2 + x_2^2}, \qquad r_2 = \sqrt{(x_1-1+\mu)^2 + x_2^2} \tag{4.2}$$

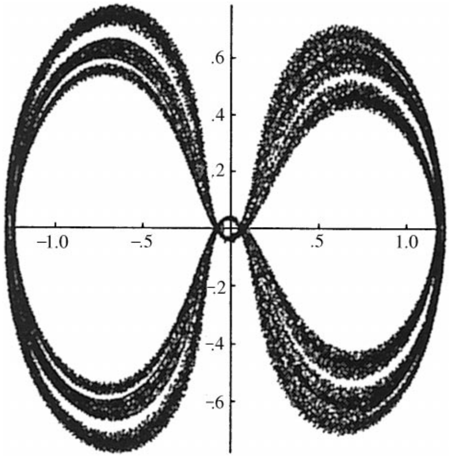

Fig. 5. Phase portrait of the trajectory starting at $x(0)^2$. Without stabilization.

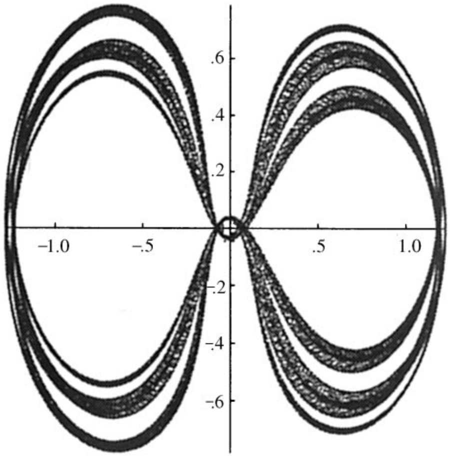

Fig. 6. Phase portrait of the trajectory starting at $x(0)^2$. With stabilization.

and $\mu = m_1/m_2$. The appropriate conserving quantity is the physical energy of the system. It reads

$$g(x) = \frac{1}{2}(x_3^2 + x_4^2) - \frac{1-\mu}{r_1} - \frac{\mu}{r_2} - \frac{1}{2}(x_1^2 + x_2^2).$$

We choose $\mu = 1/82.45$ (mass ratio earth–moon) as well as the initial values

$$\begin{aligned} x(0)^1 &= (1.2, 0, 0, -1.04935751), \\ x(0)^2 &= (1.22, 0.02, 0.02, -1.05481939), \end{aligned} \tag{4.3}$$

which both lie on the same energy level $g(x(0)^i) = -1.041588$, $i = 1,2$. We then solve problem (4.1)–(4.3) for $t \in [0, 10^5]$ with and without stabilization using NAG-routine D02BBF with local error parameter $\mathrm{TOL} = 10^{-7}$. We show in Figs. 1 and 2 the conservation of the energy along the solution and in Figs. 3–6 projections of the phase portraits to the (x_1, x_2)-plane of the solution for each initial value in (4.3). Figs. 1,3,5 show the results without stabilization and Figs. 2,4,6 are generated with the stabilization matrix $A(x)=(Dg(x)Dg(x)^{\mathrm{T}})^{-1}$ in (2.10).

It is now obvious that, using the stabilization, the energy is conserved within the local error (TOL$=10^{-7}$), whereas in the other case the energy rises linearly in time. Moreover, the plots computed with stabilization show much more structure than the pictures generated without the stabilization technique.

Appendix

Here we establish formula (3.3) in the multistep case.

Lemma A.1. *Let the sequence* $(y_n)_{n\in\mathbb{N}}$ *be generated via applying a variable step size* pth *order linear multistep method* (2.3), (2.4) *to an arbitrary smooth initial value problem* $\dot{x}=f(x)$, $x(0)=x_0$ *with solution flow* $\phi(t,x_0)$. *Moreover, let the stability assumptions* (i)–(iii) *be fulfilled and let* f *be globally lipschitzian. Then, the inequality* $\|y_{n+k}-\phi(h_{n+k-1},y_{n+k-1})\|\leqslant Ch_{n+k-1}^{p+1}$ *holds for* $n\in\mathbb{N}$.

Proof. We know

$$y_{n+k}=g_{n+k-1}(y_n,\dots,y_{n+k-1})=-\sum_{i=0}^{k-1}\alpha_{in}y_{n+i}+h_{n+k-1}\psi_{n+k-1}, \tag{A.1}$$

where ψ_{n+k-1} is the solution of

$$\psi_{n+k-1}=\sum_{i=0}^{k-1}\beta_{in}f(y_{n+i})+\beta_{kn}f\left(h_{n+k-1}\psi_{n+k-1}-\sum_{i=0}^{k-1}\alpha_{in}y_{n+i}\right). \tag{A.2}$$

For the local error (see (2.6)) in the $(n+k-1)$th step we find the representation

$$\begin{aligned} &g_{n+k-1}\left(\phi\left(-\sum_{j=1}^{k-1}h_{n+j-1},y_{n+k-1}\right),\dots,\phi(-h_{n+k-2},y_{n+k-1}),y_{n+k-1}\right)\\ &\quad=-\sum_{i=0}^{k-1}\alpha_{in}\phi\left(-\sum_{j=i+1}^{k-1}h_{n+j-1},y_{n+k-1}\right)+h_{n+k-1}\hat{\psi}_{n+k-1} \end{aligned} \tag{A.3}$$

and $\hat{\psi}_{n+k-1}$ fulfills

$$\begin{aligned} \hat{\psi}_{n+k-1}=&\sum_{i=0}^{k-1}\beta_{in}f\left(\phi\left(-\sum_{j=i+1}^{k-1}h_{n+j-1},y_{n+k-1}\right)\right)\\ &+\beta_{kn}f\left(h_{n+k-1}\hat{\psi}_{n+k-1}-\sum_{i=0}^{k-1}\alpha_{in}\phi\left(-\sum_{j=i+1}^{k-1}h_{n+j-1},y_{n+k-1}\right)\right). \end{aligned}$$

Thus, we get

$$\begin{aligned}\psi_{n+k-1}-\hat{\psi}_{n+k-1}=&\sum_{i=0}^{k-1}\beta_{in}\varDelta_i\left(y_{n+i}-\phi\left(-\sum_{j=i+1}^{k-1}h_{n+j-1},y_{n+k-1}\right)\right)\\&+\beta_{kn}\varDelta_k\Bigg[h_{n+k-1}(\psi_{n+k-1}-\hat{\psi}_{n+k-1})\\&-\sum_{i=0}^{k-1}\alpha_{in}\left(y_{n+i}-\phi\left(-\sum_{j=i+1}^{k-1}h_{n+j-1},y_{y+k-1}\right)\right)\Bigg],\end{aligned}\tag{A.4}$$

where the matrices $\varDelta_i$, $i=0,\dots,k$ arise from an application of the mean value theorem. Next, we remind the reader of the flow property

$$\phi(t,y)-\phi(t,z)=(I+\mathrm{O}(t))(y-z)\quad\text{for } t\in J(y)\cap J(z).$$

With $\hat{t}_i=\sum_{j=i+1}^{k-1}h_{n+j-1}$ and $h_{inf}\leqslant h_l\leqslant h_{\sup}$, $l\in\mathbb{N}$ we can deduce $\mathrm{O}(\hat{t}_i)=\mathrm{O}(h_{n+k-1})$. Using this and $(I+\mathrm{O}(t))^{-1}=I+\mathrm{O}(t)$, we can calculate for $i=0,\dots,k-2$

$$y_{n+i}-\phi\left(-\sum_{j=i+1}^{k-1}h_{n+j-1},y_{n+k-1}\right)=(-I+\mathrm{O}(h_{n+k-1}))\left(y_{n+k-1}-\phi\left(\sum_{j=i+1}^{k-1}h_{n+j-1},y_{n+i}\right)\right).\tag{A.5}$$

Inserting this into Eq. (A.4) and noticing that the summand for $i=k-1$ in (A.4) is zero, we obtain

$$\begin{aligned}\psi_{n+k-1}-\hat{\psi}_{n+k-1}=&(I+\mathrm{O}(h_{n+k-1}))\sum_{i=0}^{k-2}[\alpha_{in}\beta_{kn}\varDelta_k-\beta_{in}\varDelta_i+\mathrm{O}(h_{n+k-1})]\\&\times\left(y_{n+k-1}-\phi\left(\sum_{j=i+1}^{k-1}h_{n+j-1},y_{n+i}\right)\right).\end{aligned}\tag{A.6}$$

Next, we compare a discrete forward step (A.1) with one forward step (A.3). Using the representations (A.1), (A.3) and (A.6) and applying (A.5) we get

$$\begin{aligned}&g_{n+k-1}(y_n,\dots,y_{n+k-1})-g_{n+k-1}\left(\phi\left(-\sum_{j=1}^{k-1}h_{n+j-1},y_{n+k-1}\right),\dots,\phi(-h_{n+k-2},y_{n+k-1}),y_{n+k-1}\right)\\&\quad=-\sum_{i=0}^{k-2}\alpha_{in}\left(y_{n+i}-\phi\left(-\sum_{j=i+1}^{k-1}h_{n+k-1},y_{n+k-1}\right)\right)+h_{n+k-1}(I+\mathrm{O}(h_{n+k-1}))\\&\qquad\times\sum_{i=0}^{k-2}[\alpha_{in}\beta_{kn}\varDelta_k-\beta_{in}\varDelta_i+\mathrm{O}(h_{n+k-1})]\left(y_{n+k-1}-\phi\left(\sum_{j=i+1}^{k-1}h_{n+j-1},y_{n+i}\right)\right)\\&\quad=\sum_{i=0}^{k-2}(\alpha_{in}I+\mathrm{O}(h_{n+k-1}))\left(y_{n+k-1}-\phi\left(\sum_{j=i+1}^{k-1}h_{n+j-1},y_{n+i}\right)\right).\end{aligned}\tag{A.7}$$

Now, with definition (2.6) of the local error we obtain from (A.7)

$$\begin{aligned} & y_{n+k} - \phi(h_{n+k-1}, y_{n+k-1}) \\ &\quad = \sum_{i=0}^{k-2} (\alpha_{in} I + \mathrm{O}(h_{n+k-1})) \left(y_{n+k-1} - \phi\left(\sum_{j=i+1}^{k-1} h_{n+j-1}, y_{n+i} \right) \right) + \mathrm{O}(h_{n+k-1}^{p+1}). \end{aligned} \tag{A.8}$$

Moreover, with formula (A.8) we can calculate for $i = 1,\dots,k-2$

$$\begin{aligned} & y_{n+k} - \phi\left(\sum_{j=i+1}^{k} h_{n+j-1}, y_{n+i} \right) - y_{n+k-1} + \phi\left(\sum_{j=i+1}^{k-1} h_{n+j-1}, y_{n+i} \right) \\ &\quad = (I + \mathrm{O}(h_{n+k-1})) \left(y_{n+k-1} - \phi\left(\sum_{j=i+1}^{k-1} h_{n+j-1}, y_{n+i} \right) \right) \\ &\qquad + \sum_{i=0}^{k-2} (\alpha_{in} I + \mathrm{O}(h_{n+k-1})) \left(y_{n+k-1} - \phi\left(\sum_{j=i+1}^{k-1} h_{n+j-1}, y_{n+i} \right) \right) - y_{n+k-1} \\ &\qquad + \phi\left(\sum_{j=i+1}^{k-1} h_{n+j-1}, y_{n+i} \right) + \mathrm{O}(h_{n+k-1}^{p+1}) \\ &\quad = \sum_{i=0}^{k-2} (\alpha_{in} I + \mathrm{O}(h_{n+k-1})) \left(y_{n+k-1} - \phi\left(\sum_{j=i+1}^{k-1} h_{n+j-1}, y_{n+i} \right) \right) + \mathrm{O}(h_{n+k-1}^{p+1}). \end{aligned} \tag{A.9}$$

In our next step to the sequence $(y_n)_{n\in\mathbb{N}}$ generated by the linear k-step method we assign the sequence $(v_{n+k-1})_{n\in\mathbb{N}}$,

$$v_{n+k-1} = \begin{pmatrix} y_{n+k-1} - \phi(\sum_{j=1}^{k-1} h_{n+j-1}, y_n) \\ y_{n+k-1} - \phi(\sum_{j=2}^{k-1} h_{n+j-1}, y_{n+1}) \\ \vdots \\ y_{n+k-1} - \phi(h_{n+k-2}, y_{n+k-2}) \end{pmatrix} = \begin{pmatrix} v_{n+k-1,1} \\ v_{n+k-1,2} \\ \vdots \\ v_{n+k-1,k-1} \end{pmatrix} \in \mathbb{R}^{N(k-1)}. \tag{A.10}$$

Using Eqs. (A.8) and (A.9), we can express v_{n+k} in terms of v_{n+k-1} as follows:

$$v_{n+k,k-1} = \sum_{i=0}^{k-2} (\alpha_{in} I + \mathrm{O}(h_{n+k-1})) v_{n+k-1,i+1} + \mathrm{O}(h_{n+k-1}^{p+1}).$$

In the case $i = 1,\dots,k-2$ we can calculate with (A.9)

$$\begin{aligned} v_{n+k,i} &= y_{n+k} - \phi\left(\sum_{j=i+1}^{k} h_{n+j-1}, y_{n+i} \right) \\ &= v_{n+k-1,i+1} + \sum_{j=0}^{k-2} (\alpha_{jn} I + \mathrm{O}(h_{n+k-1})) v_{n+k-1,j+1} + \mathrm{O}(h_{n+k-1}^{p+1}). \end{aligned}$$

Putting these two formulae together in a matrix notation yields

$$v_{n+k} = ((\hat{A}_n \otimes I) + \mathrm{O}(h_{n+k-1})) v_{n+k-1} + \mathrm{O}(h_{n+k-1}^{p+1}) \tag{A.11}$$

with the matrix

$$\hat{A}_n = \begin{pmatrix} \theta & I_{k-2} \\ 0 & \theta^{\mathrm{T}} \end{pmatrix} + \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} (\alpha_{0n}, \dots, \alpha_{(k-2)n})^{\mathrm{T}} \in \mathbb{R}^{k-1,k-1}.$$

The matrix $\hat{A}_n$ is related to the multistep method in the following crucial way. Let

$$A_n = \begin{pmatrix} 0 & 1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -\alpha_{0n} & -\alpha_{1n} & \cdots & -\alpha_{(k-2)n} & -\alpha_{(k-1)n} \end{pmatrix} \in \mathbb{R}^{k,k}$$

denote the matrix of the linear multistep method in a one-step formulation (see, e.g. [5, p. 403]). By assumption we have $A_n(1,\dots,1)^{\mathrm{T}} = (1,\dots,1)^{\mathrm{T}}$, $n \in \mathbb{N}$. Let now $\gamma^n = (\gamma_0^n, \dots, \gamma_{k-1}^n)$ be chosen via $A_n^{\mathrm{T}}\gamma^n = \gamma^n$ and $\sum_{i=0}^{k-1} \gamma_i^n = 1$. With

$$T_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} (1, \gamma_0^n, \dots, \gamma_{k-2}^n)^{\mathrm{T}} - \begin{pmatrix} \theta & I_{k-1} \\ 0 & \theta^{\mathrm{T}} \end{pmatrix} \in \mathbb{R}^{k,k},$$

we find that

$$T_n^{-1} = \begin{pmatrix} \gamma_0^n & \cdots & \gamma_{k-2}^n & \gamma_{k-1}^n \\ & & & 1 \\ & -I_{k-1} & & \vdots \\ & & & 1 \end{pmatrix}$$

and obtain the relationship

$$T_n^{-1} A_n T_n = \begin{pmatrix} 1 & \theta^{\mathrm{T}} \\ \theta & \hat{A}_n \end{pmatrix} \in \mathbb{R}^{k,k}.$$

We now want to ensure the existence of a norm $\|\cdot\|_*$ on $\mathbb{R}^{N(k-1)}$ such that

$$\|(\hat{A}_n \otimes I)\|_* \leqslant \rho < 1, \quad \forall n \in \mathbb{N}.$$

Let

$$\hat{A} = \begin{pmatrix} \theta & I_{k-2} \\ 0 & \theta^{\mathrm{T}} \end{pmatrix} + \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} ((\alpha_0, \dots, \alpha_{k-2})(1, \dots, 1))^{\mathrm{T}}.$$

Condition (iii) guarantees that there is a norm $\|\cdot\|_*$ on $\mathbb{R}^{N(k-1)}$ such that

$$\|\hat{A} \otimes I\|_* \leqslant \hat{\rho} < 1.$$

Let now ρ satisfy $\hat{\rho} < \rho < 1$, and let $U := \{C \in \mathbb{R}^{N(k-1),N(k-1)} \mid \| C \|_* < \rho\}$ be a neighbourhood of $(\hat{A} \otimes I)$ in $\mathbb{R}^{N(k-1),N(k-1)}$. Now, for $v = (v_1, \dots, v_{k-1}) \in \mathbb{R}^{k-1}$ we consider the continuous matrix family

$$B(v) = \begin{pmatrix} \theta & I_{k-2} \\ 0 & \theta^{\mathrm{T}} \end{pmatrix} + \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} (\alpha_0(v), \dots, \alpha_{k-2}(v))^{\mathrm{T}}.$$

By continuity, there exists a neighbourhood V of $(1, \dots, 1)$ in $\mathbb{R}^{k-1}$ such that $B(v) \in U$ for $v \in V$. If ω_d, ω_u are sufficiently close to 1 we have $v \in V$ for $\omega_d \leqslant v_i \leqslant \omega_u$, $i = 1, \dots, k-1$. This assures $\hat{A}_n = B(\omega_{n+1}, \dots, \omega_{n+k-1}) \in U$, $n \in \mathbb{N}$, if $\omega_d \leqslant \omega_n \leqslant \omega_u$ for $n \in \mathbb{N}$. Hence,

$$\| \hat{A}_n \otimes I \|_* < \rho < 1 \quad \forall n \in \mathbb{N}$$

follows. Using (A.11), we can calculate

$$\| v_{n+k} \|_* \leqslant \rho \| v_{n+k-1} \|_* + Ch_{n+k-1}^{p+1} \leqslant \rho \| v_{n+k-1} \|_* + Ch_{\sup}^{p+1}$$

for $h_{\sup} > 0$ sufficiently small. Recursively, this gives us

$$\begin{aligned} \| v_{n+k} \|_* &\leqslant (1 + \rho + \rho^2 + \cdots + \rho^n) Ch_{\sup}^{p+1} + \rho^{n+1} \| v_{k-1} \|_* \\ &\leqslant \frac{1}{1-\rho} Ch_{\sup}^{p+1} + \rho^{n+1} \hat{C} h_{\sup}^{p+1} \leqslant \tilde{C} h_{\sup}^{p+1}. \end{aligned}$$

Since we have $0 < h_{\inf} \leqslant h_j \leqslant h_{\sup}$, we obtain

$$\| v_{n+k} \|_* \leqslant C_1 h_{n+k-1}^{p+1} \quad \forall n \in \mathbb{N} \quad \text{and} \quad y_0 \in \mathbb{R}^N. \tag{A.12}$$

Now we are in the position to derive (3.3). Since f is lipschitzian, the discrete step forward map g_{n+k-1} is lipschitzian with L_g, and we obtain

$$\begin{aligned} &\| y_{n+k} - \phi(h_{n+k-1}, y_{n+k-1}) \| \\ &\quad \leqslant \Bigg\| g_{n+k-1}(y_n, \dots, y_{n+k-1}) \\ &\qquad - g_{n+k-1}\left(\phi\left(-\sum_{j=1}^{k-1} h_{n+j-1}, y_{n+k-1}\right), \dots, \phi(-h_{n+k-2}, y_{n+k-1}), y_{n+k-1}\right)\Bigg\| \\ &\qquad + \Bigg\| g_{n+k-1}\left(\phi\left(-\sum_{j=1}^{k-1} h_{n+j-1}, y_{n+k-1}\right), \dots, \phi(-h_{n+k-2}, y_{n+k-1}), y_{n+k-1}\right) \\ &\qquad - \phi(h_{n+k-1}, y_{n+k-1})\Bigg\| \leqslant L_g \| \Gamma_{n+k-1} \| + C_p h_{n+k-1}^{p+1} \end{aligned} \tag{A.13}$$

with

$$\Gamma_{n+k-1} = \begin{pmatrix} y_n \\ y_{n+1} \\ \vdots \\ y_{n+k-2} \\ y_{n+k-1} \end{pmatrix} - \begin{pmatrix} \phi(-\sum_{j=1}^{k-1} h_{n+j-1}, y_{n+k-1}) \\ \phi(-\sum_{j=2}^{k-1} h_{n+j-1}, y_{n+k-1}) \\ \vdots \\ \phi(-h_{n+k-2}, y_{n+k-1}) \\ y_{n+k-1} \end{pmatrix} = \begin{pmatrix} \Gamma_{n+k-1,1} \\ \Gamma_{n+k-1,2} \\ \vdots \\ \Gamma_{n+k-1,k-1} \\ \Gamma_{n+k-1,k} \end{pmatrix}.$$

Then, using $\phi(t,y) - \phi(t,z) = (I + \mathrm{O}(t))(y - z)$ we can calculate

$$\begin{aligned} -v_{n+k-1,i} &= \phi\left(\sum_{j=i}^{k-1} h_{n+j-1}, y_{n+i-1}\right) - \phi\left(\sum_{j=i}^{k-1} h_{n+j-1}, \phi\left(-\sum_{j=i}^{k-1} h_{n+j-1}, y_{n+k-1}\right)\right) \\ &= (I + \mathrm{O}(h_{n+k-1}))\left(y_{n+i-1} - \phi\left(-\sum_{j=i}^{k-1} h_{n+j-1}, y_{n+k-1}\right)\right), \quad i = 1,\dots,k-1. \end{aligned}$$

Hence, $v_{n+k-1,i} = (-I + \mathrm{O}(h_{n+k-1}))\Gamma_{n+k-1,i}$, $i = 1,\dots,k-1$ follows. Since the last component of Γ_{n+k-1} vanishes, formula (A.12) assures

$$\| \Gamma_{n+k-1} \| = \| (-I + \mathrm{O}(h_{n+k-1}))(v_{n+k-1}, 0) \| \leqslant C_2 h_{n+k-1}^{p+1}.$$

Finally, we insert this into Eq. (A.13) and obtain

$$\| y_{n+k} - \phi(h_{n+k-1}, y_{n+k-1}) \| \leqslant C_3 h_{n+k-1}^{p+1} \quad \forall n \in \mathbb{N} \text{ and } y_0 \in \mathbb{R}^N,$$

which is the desired result.

References

[1] U.M. Ascher, H. Chin, S. Reich, Stabilization of DAE's and invariant manifolds, Numer. Math. 67 (1994) 131–149.

[2] M.P. Calvo, J.M. Sanz-Serna, Reasons for a failure. The integration of the two-body problem with a symplectic Runge-Kutta method with step changing facilities, in: C. Simo, J. de Sola-Marales (Eds.), Equadiff-91, Singapore, 1993, pp. 93–102.

[3] W.A. Coppel, Stability and Asymptotic Behaviour of Differential Equations, Boston Englewood Chicago, 1965.

[4] E. Hairer, Variable time step integration with symplectic methods, Appl. Numer. Math. 25 (1997) 219–227.

[5] E. Hairer, S.P. Nørsett, G. Wanner, Solving Ordinary Differential Equations I, 2nd Edition, Wiley-Interscience, New York, 1993.

[6] M.W. Hirsch, Differential Topology, Springer, New York, 1976.

[7] P.E. Kloeden, J. Lorenz, Stable attracting sets in dynamical systems and in their one-step discretizations, SIAM J. Numer. Anal. 23 (1986) 986–995.

[8] P.E. Kloeden, J. Lorenz, A note on multistep methods and attracting sets of dynamical systems, Numer. Math. 56 (1990) 667–673.

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

[10] K. Strehmel, R. Weiner, Numerik gewöhnlicher Differentialgleichungen, Teubner Verlag, Stuttgart, 1995.

[11] A.M. Stuart, A.R. Humphries, Dynamical Systems and Numerical Analysis, Monographs on Applied and Computational Mathematics, 1996, Cambridge University Press, Cambridge.