ELSEVIER

Journal of Computational and Applied Mathematics 111 (1999) 13–24

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Rigorous error bounds for RK methods in the proof of chaotic behaviour

Arno Berger

Institute of Mechanics, Vienna University of Technology, Vienna, Austria

Received 27 April 1998; received in revised form 24 February 1999

Abstract

Complicated dynamical systems can be rigorously analysed by means of Conley index theory. Due to its partly numerical nature such an analysis necessitates bounds on the truncation and the round-off error. These are provided for explicit RK methods in the form of iteration schemes ready-made for applications. The presentation is aimed to simplify error bounds already available so that different error sources can be clearly overlooked. As an immediate application, a computer-assisted analysis elucidates the intricate dynamics of a simple mechanical system.

Keywords: Explicit RK method; Truncation and round-off error; Conley index theory

1. Introduction

Many interesting features of complicated dynamical systems are nowadays studied by means of computers. Despite all numerical evidence it is usually highly demanding to describe the observed phenomena in a mathematically rigorous manner. Finding theoretical results that are accessible to the computational power of modern computers is a major challenge of applied mathematics. Recently, a couple of attempts in this direction have been carried out [6].

Since the famous paper of Mischaikow and Mrozek [5] on chaos in the Lorenz equations *Conley index theory* has been taken notice of beyond a small number of specialists. Due to its topological nature the Conley index can be used to develop elegant tools for analysing dynamical systems by means of computers. (Although highly plausible, this statement is not at all trivial, cf. [5,6,8].) Since the theory is intended to give completely rigorous statements about dynamics, all numerical calculations have to be accompanied by rigorous error considerations. Without additional information

E-mail address: aberger@mch2ws1.tuwien.ac.at (A. Berger)

0377-0427/99/$ - see front matter
PII: S0377-0427(99)00128-4

on the problem under consideration such rigorous error treatment is known to be quite costly [6]. Many interesting systems from applications indicate however that errors occur at different orders of magnitude [1,6]. In double precision arithmetic round-off errors are negligible in many cases when compared with the error effects of discretization (i.e., obtaining finite models of spaces and maps). Even the truncation errors due to the usage of finite algorithms can be seen to be of minor importance if only reasonable stepsizes are used. Developing bounds for all error sources that are neither too crude nor too complicated on one hand and completely rigorous on the other hand therefore is an important task in applied Conley index theory.

Sections 2, 3 and 4, respectively, present rigorous bounds on the truncation, the round-off and the global error for the numerical integration of ordinary differential equations by means of explicit RK methods. The results may even be suited for other purposes and partly simplify the calculations in [6]. In order to demonstrate how the findings fit into the framework of Conley index theory, the dynamics of a pendulum with oscillating support is investigated in the final section. With little effort, the existence of periodic orbits and the factorization onto a chaotic system can be proved.

2. The truncation error

Let V denote a C^p-vectorfield on $U \subseteq \mathbb{R}^d$ and consider the autonomous initial-value problem

$$\dot{x} = V(x), \quad x(0) = x_0. \tag{1}$$

In order to get a numerical approximation for the solution of (1), we apply an s-stage explicit RK method with stepsize h and real coefficients a_{ij} and b_i,

$$\begin{aligned} k_1 &:= V(x_0), \\ k_i &:= V\left(x_0 + h\sum_{j=1}^{i-1} a_{ij}k_j\right), \quad i = 2,\dots,s, \\ x_1 &:= x_0 + h\sum_{i=1}^{s} b_i k_i. \end{aligned} \tag{2}$$

If this method is of order p, the inequality

$$\|x(h) - x_1\| \leqslant h^{p+1}\left(\frac{1}{(p+1)!}\max_{t\in[0;1]}\|x^{(p+1)}(th)\| + \frac{1}{p!}\sum_{i=1}^{s}|b_i|\max_{t\in[0;1]}\|k_i^{(p)}(th)\|\right) \tag{3}$$

constitutes a well-known rigorous bound for the local error [2]. For technical reasons $\|\cdot\|$ always denote the max-norm on $\mathbb{R}^d$; other norms are made recognizable by subscript.

The right-hand side of (3), although easily written down, is commonly regarded as being of no practical importance because its evaluation turns out to be arduous even for simple non-linear vectorfields [2,6]. Other methods permit a more tractable and realistic view of the local error. However, the usage of (3) becomes inevitable if definitely rigorous error considerations are essential.

Rather than following directly the approach in [6] which relies heavily on the availability of symbolic computation software, we shall refine an idea from [3]. To this end the ith component of V is denoted by V^i while its partial derivative with respect to the ith coordinate is symbolized

by $V_{,i}$. Let $K \subseteq U$ be a compact set containing x_0 and

$$D_r := \max_{\substack{\xi \in K \\ 1 \leqslant i_1,\ldots,i_r \leqslant d}} \|V_{,i_1\ldots i_r}(\xi)\|_1 \quad \text{with } 0 \leqslant r \leqslant p. \tag{4}$$

The derivative $x^{(p+1)}$ of the solution of (1) can be easily bounded by means of the quantities D_r, once a representation of $x^{(p+1)}$ in terms of V and its derivatives is given. This is most conveniently achieved by means of trees which give rise to the concise expression

$$x^{(p+1)} = \sum_{\boldsymbol{t} \in T_{p+1}} \alpha(\boldsymbol{t}) \partial_{\boldsymbol{t}} V. \tag{5}$$

Here, T_{p+1} denotes the set of trees of order $p+1$, and $\alpha(\boldsymbol{t})$ is the number of different monotonic labellings of $\boldsymbol{t} \in T_{p+1}$ (see [2] for details). In our notation one term of (5) typically takes the form

$$V_{,i_1\ldots i_{j_0}} V^{i_1}_{,i_{(j_0+1)}\ldots i_{(j_0+j_1)}} \ldots V^{i_p}_{,i_{(j_0+\cdots+j_{p-1}+1)}\ldots i_{(j_0+\cdots+j_p)}} \quad \text{with } j_0+\cdots+j_p = p.$$

There is a one-to-one correspondence between the $(p+1)$-tuples $(j_0,\ldots,j_p)$ and the monotonic labellings of $\boldsymbol{t}$ [2]. Using (4) we get

$$\|V_{,i_1\ldots i_{j_0}} V^{i_1}_{,i_{(j_0+1)}\ldots i_{(j_0+j_1)}} \ldots V^{i_p}_{,i_{(j_0+\cdots+j_{p-1}+1)}\ldots i_{(j_0+\cdots+j_p)}}\| \leqslant D_{j_0} D_{j_1} \ldots D_{j_p};$$

and denoting by $(j_0(\boldsymbol{t}),\ldots,j_p(\boldsymbol{t}))$ the $(p+1)$-tuple corresponding to a representative of $\boldsymbol{t}$,

$$\|x^{(p+1)}\| \leqslant \sum_{\boldsymbol{t} \in T_{p+1}} \alpha(\boldsymbol{t}) D_{j_0(\boldsymbol{t})} D_{j_1(\boldsymbol{t})} \ldots D_{j_p(\boldsymbol{t})}. \tag{6}$$

With increasing order p the determination of $\alpha(\boldsymbol{t})$ becomes lengthy. Nevertheless, it is a purely combinatorial task. We therefore regard (6) as a satisfactory bound for the first summand on the right-hand side of (3). (In [2] one can find values of $\alpha(\boldsymbol{t})$ up to $p = 4$.)

We now turn our attention towards the evaluation of $\|k_i^{(p)}\|$. Obviously $k_1^{(r)} = 0$ for $1 \leqslant r \leqslant p$. If $i \geqslant 2$ let for sake of brevity $l_i := \sum_{j=1}^{i-1} a_{ij} k_j$. Differentiation leads to the general expression

$$k_i^{(r)} = \sum_{m=1}^{r} \sum_{\substack{i_1 \geqslant \cdots \geqslant i_m \geqslant 1 \\ i_1+\cdots+i_m = r}} \beta^{(r)}_{m;i_1,\ldots,i_m} V_{,j_1\ldots j_m} (hl_i)^{(i_1),j_1} \ldots (hl_i)^{(i_m),j_m}. \tag{7}$$

Counting identical derivatives, the positive integers $\beta^{(r)}_{m;i_1,\ldots,i_m}$ can be determined inductively. Via differentiation $\beta^{(r)}_{m;i_1,\ldots,i_m}$ contributes to $\beta^{(r+1)}_{m+1;i_1,\ldots,i_m,1}$ as well as to $\beta^{(r+1)}_{m;i_1+1,\ldots,i_m}, \ldots, \beta^{(r+1)}_{m;i_1,\ldots,i_m+1}$ (where the latter will have to be rearranged if the indices i_j are not in the decreasing order). For example up to $r = 3$ we find

$$\beta^{(1)}_{1;1} = 1, \quad \beta^{(2)}_{1;2} = 1, \quad \beta^{(2)}_{2;1,1} = 1, \quad \beta^{(3)}_{1;3} = 1, \quad \beta^{(3)}_{2;2,1} = 3, \quad \beta^{(3)}_{3;1,1,1} = 1.$$

As a consequence of representation (7) we see that

$$\|k_i^{(r)}\| \leqslant \sum_{m=1}^{r} \sum_{\substack{i_1 \geqslant \cdots \geqslant i_m \geqslant 1 \\ i_1+\cdots+i_m = r}} \beta^{(r)}_{m;i_1,\ldots,i_m} D_m \|(hl_i)^{(i_1)}\|_1 \ldots \|(hl_i)^{(i_m)}\|_1. \tag{8}$$

Taking into account the obvious relations

$$\|(hl_i)^{(r)}\| \leqslant r\|l_i^{(r-1)}\| + h\|l_i^{(r)}\| \quad \text{and} \quad \|l_i^{(r)}\| \leqslant \sum_{j=1}^{i-1} |a_{ij}| \|k_j^{(r)}\|, \tag{9}$$

we can construct an iteration scheme to obtain the required bounds for $\|k_i^{(p)}\|$. To this end define two mappings $\Phi, \Psi : \mathbb{R}^p \to \mathbb{R}^p$ by

$$\Phi(x_1,\dots,x_p) := \left(\beta_{1;1}^{(1)} D_1 x_1, \beta_{1;2}^{(2)} D_1 x_2 + \beta_{2;1,1}^{(2)} D_2 x_1^2, \dots, \sum_{m=1}^{p} \sum_{\substack{i_1 \geqslant \cdots \geqslant i_m \geqslant 1 \\ i_1+\cdots+i_m=p}} \beta_{m;i_1,\dots,i_m}^{(p)} D_m x_{i_1} \dots x_{i_m}\right)$$

and

$$\Psi(x_1,\dots,x_p) := (D_0 + hx_1, 2x_1 + hx_2, \dots, px_{p-1} + hx_p),$$

corresponding to (8) and the first relation in (9), respectively. Assume that we have already constructed upper bounds K_i^r for $\|k_i^{(r)}\| (1 \leqslant r \leqslant p)$ and let $K_i := (K_i^1, \dots, K_i^p)$. It is easy now to show that the difference scheme

$$K_{i+1} := \Phi\left(\sum_{j=1}^{i} |a_{i+1,j}| \Psi(K_j)\right), \tag{10}$$

in fact, generates K_{i+1} correctly. As initial value for (10) we have $K_1 := (0,\dots,0)$. Collecting our results and defining

$$C_K(h) := \frac{1}{(p+1)!} \sum_{t \in T_{p+1}} \alpha(t) D_{j_0(t)} \dots D_{j_p(t)} + \frac{1}{p!} \sum_{i=1}^{s} |b_i| K_i^p, \tag{11}$$

we can replace (3) by the concise expression

$$\|x(h) - x_1\| \leqslant h^{p+1} C_K(h).$$

Clearly, $C_K(h)$ is a polynomial in h with nonnegative coefficients multinomial in the quantities D_r. One should note that due to this fact $C_K(h)$ may be evaluated numerically by *directed rounding* [6]. If the latter is not available, rigorous upper bounds for $C_K(h)$ will be provided by (a simplified version of) the round-off treatment sketched in the next section.

An induction argument shows that

$$\deg C_K(h) = (s-2)p \quad \text{for } s \geqslant 2. \tag{12}$$

The subscript K displays the fact that the constructed error bound also depends on the compact set K (via the quantities D_r). In applications one typically tries to define this set as small as possible and therefore has to check whether the solution $x(t)$ of (1) remains in K for all $0 \leqslant t \leqslant h$. If not so, the set K will have to be enlarged.

The calculations leading to (11) clearly become more and more arduous for higher orders p. This is mainly due to the nonlinearity sewed in the definition of Φ (note that Ψ is just an affine map). But also the cardinality of T_p increases rapidly, as can be seen from [2]. Nevertheless, it is not at all difficult to calculate $C_K(h)$ with the help of symbolic computation software. In case of higher orders ($p \geqslant 5$) this approach turns out to be inevitable. However, it may be interesting to point out that (contrary to [6]) all the necessary calculations in fact can be performed by hand if $p < 5$.

3. The round-off error

As was mentioned earlier we do not expect the round-off error to play a dominant role in our analysis of dynamical systems. We will therefore content ourselves with quite crude rigorous error bounds, as long as they can be calculated efficiently. An elegant method in this direction was developed in [6]. We refer the reader to this article and to [1] for any details concerning the following definitions and results.

Let $\mathbb{M}(2, p, e_{\min}, e_{\max})$ denote a fixed system of machine numbers and define the function $M_2 : \mathbb{R} \to \{2^k | k \in \mathbb{Z}\}$ by

$$M_2(x) := \begin{cases} 2^{e_{\min}-1}/m_0 & \text{if } |x| \leqslant 2^{e_{\min}-1}/m_0, \\ \min\{2^k \,|\, |x| \leqslant 2^k\} & \text{otherwise.} \end{cases} \tag{13}$$

Here m_0 measures the accuracy of representing real numbers by elements in $\mathbb{M}$. In case of optimal rounding we have $m_0 = 2^{-(p+1)}$, while in any other case $m_0 = 2^{-p}$.

For any arithmetic expression $w(X_1, \ldots, X_n)$ consisting of a finite number of symbols and implemented non-polynomial functions f_i from $\mathbb{M} \cup \{+, -, \cdot, (,)\} \cup \{(f_i)_{i=1}^m\}$ one defines recursively the evaluation $w(x_1, \ldots, x_n)$ and the machine evaluation $\langle w \rangle(x_1, \ldots, x_n)$ of w at $(x_1, \ldots, x_n)$. It is well known from numerical analysis that different arithmetic expressions having the same evaluation (i.e., being mathematically equivalent) can give rise to different machine evaluations. In order to obtain a rigorous bound for the absolute round-off error $|w - \langle w \rangle|$, two functions H_w and ε_w are defined recursively as

$$H_w(x_1, \ldots, x_n) := \begin{cases} M_2(m) & \text{if } w = m \in \mathbb{M}, \\ M_2(x_i) & \text{if } w = X_i, \\ 2\max\{H_{w_1}, H_{w_2}\} & \text{if } w = w_1 \pm w_2, \\ H_{w_1} H_{w_2} & \text{if } w = w_1 \cdot w_2, \\ M_2(\|f_i\|) & \text{if } w = f_i \end{cases} \tag{14}$$

and

$$\varepsilon_w(x_1, \ldots, x_n) := \begin{cases} 0 & \text{if } w = m \in \mathbb{M}, \\ |x_i| & \text{if } w = X_i, \\ m_0 + \max\{\varepsilon_{w_1}, \varepsilon_{w_2}\} & \text{if } w = w_1 \pm w_2 \ , \\ m_0 + \varepsilon_{w_1} + \varepsilon_{w_2} & \text{if } w = w_1 \cdot w_2, \\ \varepsilon_{f_i} & \text{if } w = f_i, \end{cases} \tag{15}$$

where ε_{f_i} denotes an upper bound for the relative evaluation error of $\langle f_i \rangle$. In some sense ε_w measures the relative evaluation error of $\langle w \rangle$ while H_w represents a bound on the absolute value of w. For sake of simplicity, definitions (13)–(15) have not been formulated in full generality here (cf. [6]). However, if there is no exponential overflow during our calculation, we have the fundamental inequality

$$|w(x_1, \ldots, x_n) - \langle w \rangle(x_1, \ldots, x_n)| \leqslant \varepsilon_w(m_0, \ldots, m_0) H_w(M_2(x_1), \ldots, M_2(x_n)). \tag{16}$$

Let $\langle x_1 \rangle$ denote the numerical result of one RK step (2) calculated in the finite arithmetic of $\mathbb{M}$. We shall now turn to the estimation of the round-off error $\|x_1 - \langle x_1 \rangle\|$. In order to maintain a certain amount of lucidity we shall write $H_w := (H_{w^1}, \ldots, H_{w^d})$ (and analogously ε_w) for an arithmetic

expression w having d components w^i. The d-tuple $(1,\dots,1)$ is denoted by $\mathbf{1}$. Expressions like $M_2(u)$ and $\max\{u,v\}$ with real d-tuples u,v should be read coordinatewise.

Setting $\xi^i := \max_{x\in K}|x^i|$ $(1\leqslant i\leqslant d)$ with the compact set K discussed in Section 3 and using systematically the notation just introduced, we can find

$$\begin{aligned} &H_{x_1}(M_2(\xi)) = 2\max\{M_2(\xi), 2^s M_2(h)M_2(b_1)H_{k_1}(M_2(\xi)),\dots,2M_2(h)M_2(b_s)H_{k_s}(M_2(\xi))\},\\ &\varepsilon_{x_1}(m_0\mathbf{1}) = 6m_0\mathbf{1} + \max\{(s-1)m_0\mathbf{1}+\varepsilon_{k_1},\dots,m_0\mathbf{1}+\varepsilon_{k_{s-1}},\varepsilon_{k_s}\}. \end{aligned} \tag{17}$$

The quantities $\eta_i := H_{k_i}(M_2(\xi))$ and $\zeta_i := \varepsilon_{k_i}(m_0\mathbf{1})$ can be determined by means of the iteration scheme

$$\begin{aligned} &\eta_{i+1} := H_V(2\max\{M_2(\xi), 2^i M_2(h)M_2(a_{i+1,1})\eta_1,\dots,2M_2(h)M_2(a_{i+1,i})\eta_i\}),\\ &\zeta_{i+1} := \varepsilon_V(6m_0\mathbf{1} + \max\{(i-1)m_0+\zeta_1,\dots,\zeta_i\}) \end{aligned} \tag{18}$$

with initial values $\eta_1 = H_V(M_2(\xi)), \zeta_1 = \varepsilon_V(m_0\mathbf{1})$. Having calculated η_i and ζ_i from this scheme we may define

$$E_K(h) := \max_{i=1}^{d}(H^i_{x_1}(M_2(\xi))\varepsilon^i_{x_1}(m_0\mathbf{1})). \tag{19}$$

Again the dependence on K (via ξ) has been emphasized by subscript. Combining (16) and (19) we finally get

$$\|x_1 - \langle x_1\rangle\| \leqslant E_K(h)$$

as the desired bound for the round-off error. Observe that due to (17) $E_K(h)$ will not effectively depend on the stepsize h, if the latter is sufficiently small.

4. The global error

It is now an easy task to combine the error bounds constructed in the previous sections. Consequently, the one-step error of the RK method (2) admits the rigorous bound

$$\|x(h) - \langle x_1\rangle\| \leqslant h^{p+1}C_K(h) + E_K(h).$$

A numerical integration of (1) usually requires more than one RK step. Using a (local) Lipschitz constant e^{hL} for the (local) flow φ_h generated by (1), the global error after N iterations of the RK scheme (2) with stepsizes $h_1,\dots,h_N$, respectively, obeys

$$\left\|x\left(\sum_{n=1}^{N}h_n\right) - \langle x_1\rangle_N\right\| \leqslant \mathrm{e}^{h_N L}\left\|x\left(\sum_{n=1}^{N-1}h_n\right) - \langle x_1\rangle_{(N-1)}\right\| + h_N^{p+1}C_K(h_N) + E_K(h_N).$$

Setting $F_K(h) := h^{p+1}C_K(h) + E_K(h)$ for sake of brevity this recursion immediately leads to

$$\left\|x\left(\sum_{n=1}^{N}h_n\right) - \langle x_1\rangle_N\right\| \leqslant \sum_{n=1}^{N}F_K(h_n)\mathrm{e}^{L\sum_{m=n+1}^{N}h_m} =: G_K(h_1,\dots,h_N). \tag{20}$$

In case of *variable* stepsize, clearly (20) cannot be evaluated until the integration process has come to an end. Additionally, a rigorous stepsize documentation is indispensable. Although these aspects will possibly cause no serious difficulties in many applications, one should notice that in case of *constant* stepsize h (20) simply reads

$$\|x(Nh) - \langle x_1\rangle_N\| \leqslant \frac{\mathrm{e}^{hLN}-1}{\mathrm{e}^{hL}-1} F_K(h) = G_K(h,\ldots,h) =: \bar{G}_K(h). \tag{21}$$

Obviously (21) constitutes $\bar{G}_K(h)$ as a ready-made error bound, the evaluation of which can be performed *before* the process of integration. In any case (20) and (21) provide rigorous error bounds for the numerical integration of (1) by means of the RK method (2). It comes as no surprise that due to our simple construction these error bounds are very crude. (In particular, the influence of rounding is usually considerably overestimated.) A refined analysis may thus focus for example on specialised growth restrictions to the flow and on round-off effects. In the general situation of (1) such considerations tend to be intractable. The applications we bear in mind, however, demand flexible and efficiently computable error bounds rather than very tight ones. We therefore consider (20) and (21) as a compromise for practical reasons. After all, applications indicate that G_K and $\bar{G}_K$ do quite satisfyingly reflect some important aspects of error analysis.

Example (*Classical RK*4 *method*)**.** Let us sketch a few results of the outlined procedure in case of the classical RK method of fourth order most conveniently represented by the tableau

0				
$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{1}{2}$	0	$\frac{1}{2}$		
1	0	0	1	
	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

where $p=4$ and $s=4$. In accordance to (12) $C_K(h)$ is a polynomial of degree 8,

$$C_K(h) = c_0 + c_1 h + \cdots + c_8 h^8,$$

whose coefficients are

$$c_0 = \tfrac{49}{2880}D_0^4D_4 + \tfrac{169}{1440}D_0^3D_1D_3 + \tfrac{31}{480}D_0^3D_2^2 + \tfrac{47}{240}D_0^2D_1^2D_2 + \tfrac{1}{120}D_0D_1^4,$$
$$c_1 = \tfrac{19}{1151}D_0^4D_1D_4 + \tfrac{7}{384}D_0^4D_2D_3 + \tfrac{31}{288}D_0^3D_1^2D_3 + \tfrac{5}{48}D_0^3D_1D_2^2 + \tfrac{5}{64}D_0^2D_1^3D_2,$$
$$c_2 = \tfrac{91}{4608}D_0^4D_1^2D_4 + \tfrac{31}{768}D_0^4D_1D_2D_3 + \tfrac{11}{1536}D_0^4D_2^3 + \tfrac{1}{12}D_0^3D_1^3D_3 + \tfrac{11}{128}D_0^3D_1^2D_2^2 + \tfrac{1}{32}D_0^2D_1^4D_2,$$
$$c_3 = \tfrac{23}{1536}D_0^4D_1^3D_4 + \tfrac{163}{4608}D_0^4D_1^2D_2D_3 + \tfrac{29}{3072}D_0^4D_1D_2^3 + \tfrac{25}{576}D_0^3D_1^4D_3 + \tfrac{1}{24}D_0^3D_1^3D_2^2,$$
$$c_4 = \tfrac{157}{18\,432}D_0^4D_1^4D_4 + \tfrac{191}{9216}D_0^4D_1^3D_2D_3 + \tfrac{19}{3072}D_0^4D_1^2D_2^3 + \tfrac{5}{384}D_0^3D_1^5D_3 + \tfrac{5}{768}D_0^3D_1^4D_2^2,$$
$$c_5 = \tfrac{43}{12\,288}D_0^4D_1^5D_4 + \tfrac{17}{2304}D_0^4D_1^4D_2D_3 + \tfrac{5}{3072}D_0^4D_1^3D_2^3 + \tfrac{1}{384}D_0^3D_1^6D_3,$$
$$c_6 = \tfrac{5}{4608}D_0^4D_1^6D_4 + \tfrac{1}{576}D_0^4D_1^5D_2D_3 + \tfrac{1}{12\,288}D_0^4D_1^4D_2^3,$$
$$c_7 = \tfrac{1}{4608}D_0^4D_1^7D_4 + \tfrac{1}{6144}D_0^4D_1^6D_2D_3,$$
$$c_8 = \tfrac{1}{36\,864}D_0^4D_1^8D_4.$$

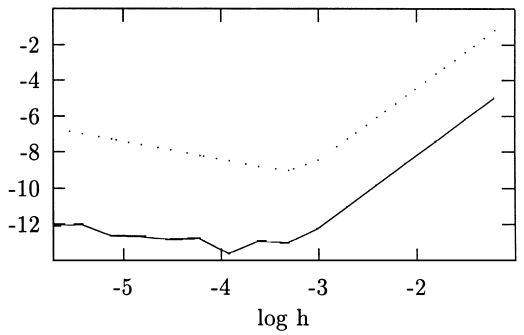


Fig. 1. The real error $\log\|x(1)-\langle x_1\rangle_{[1/h]}\|$ (full line) and the error bound $\log\bar{G}_K(h)$ for the vectorfield (22).

Relations (17) and (18) can partly be simplified to give

$$H_{x_1}(M_2(\xi))=\max\{2M_2(\xi),M_2(h)\max\{8H_{k_1}(M_2(\xi)),8H_{k_2}(M_2(\xi)),4H_{k_3}(M_2(\xi)),H_{k_4}(M_2(\xi))\}\},$$

$$\eta_{i+1}=H_V(2\max\{M_2(\xi),2a_{i+1,i}M_2(h)\eta_i\}).$$

As a specific example we shall determine the quantities $C_K(h)$ and $E_K(h)$ for the simple vectorfield

$$V(x_1,x_2):=(x_1-x_1^2,-\tfrac{1}{2}x_2+\tfrac{1}{2}x_1x_2). \tag{22}$$

Setting $x(0):=(2,1)$ and expecting the solution of $\dot{x}=V(x)$ not to diverge from this point too rapidly we choose $K:=[1;3]\times[0;2]$. We then find

$$D_0=8,\quad D_1=6,\quad D_2=2,\quad D_r=0\quad(r\geqslant 3)$$

and consequently

$$C_K(h)=\tfrac{16\,816}{15}+3440\,h+\tfrac{35\,264}{3}h^2+20\,288\,h^3+24\,576\,h^4+11\,520\,h^5+3456\,h^6.$$

If we restrict ourselves to the case $M_2(h)\leqslant 2^{-4}$ we get (after some tedious but straightforward calculations)

$$E_K(h)=\begin{cases}17\times 2^{14}m_0 & \text{if } M_2(h)=2^{-4},\\ 17\times 2^{13}m_0\max\{2^{-7},M_2(h)\} & \text{if } M_2(h)\leqslant 2^{-5}.\end{cases}$$

A Lipschitz constant of the (local) flow generated by (22) is most conveniently determined by means of logarithmic norms [2]. For the system under consideration we have $L=2$. Using an IEEE arithmetic $\mathbb{M}(2,53,-1021,1024)$ we apply the classical RK4 method with stepsizes $h_i:=2^{-i}$ $(4\leqslant i\leqslant 20)$ in order to get a numerical approximation of $x(1)=(2/\gamma,\sqrt{\gamma})$ with $\gamma=2-e^{-1}$. Despite its crude character the rigorous error bound $\bar{G}_K$ provides a reasonable (upper) estimate for the optimal stepsize (Fig. 1).

5. A view towards applications

The goal of this final section is to briefly discuss the usage of the presented techniques in calculating an important algebraic-topological invariant from the theory of dynamical systems. This invariant,

the so-called *Conley index*, may be thought of as an algebraic-topological quantity (roughly) describing some structural features of invariant sets, which for technical reasons are always assumed to be compact and isolated, i.e., maximal invariant within an open neighbourhood. Having actually calculated the Conley index, one can often gain some insight into the dynamics taking place on the isolated invariant set under consideration. The analysis of such sets may be vital for globally understanding complicated dynamical systems. Due to our focus on rigorous error considerations and computational aspects we shall not give mathematical precision to these statements but encourage the reader to look up the presentations [5,7]. For our purpose, it is sufficient to know that in order to calculate the Conley index one has to find a so-called *index pair*, i.e., a pair of compact sets satisfying some topological assumptions and thereby carrying a certain amount of information about the underlying dynamics. Starting from this index pair an algebraic-topological procedure gives the index.

In the sequel, we shall concentrate on the dynamical behaviour of a mathematical pendulum with a support oscillating according to $\alpha\cos\omega t$ (Fig. 2). Its equation of motion reads

$$\ddot{\varphi} + \frac{\beta}{ml^2}\dot{\varphi} + \left(\frac{g}{l} + \frac{\alpha\omega^2}{l}\cos\omega t\right)\sin\varphi = 0, \tag{23}$$

where only *linear* frictional effects have been considered via the parameter β. By introducing nondimensional time and frequency, $\tau := t\sqrt{g/l}$ and $\nu := \omega\sqrt{l/g}$, (23) can be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_2 \\ -(1 + A\cos x_3)\sin x_1 - Bx_2 \\ \nu \end{pmatrix} \tag{24}$$

with the abbreviations $A := \alpha\omega^2/g$ and $B := \beta/ml\sqrt{gl}$. Integration of (24) over $[0; 2\pi/\nu]$ yields a Poincaré map $\Psi_{A;B}$ naturally acting on $S^1 \times \mathbb{R}$. The resulting dynamical system exhibits a great variety of different phenomena and therefore has been studied repeatedly by theoretical as well as numerical means [4].

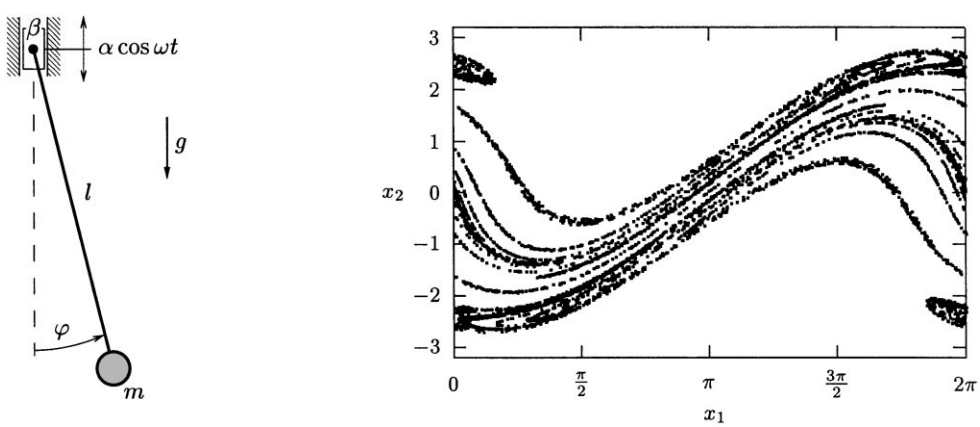


Fig. 2. The pendulum with oscillating support (left) and the attractor for the Poincaré map $\Psi_{0.94;0.15}$ displayed via 7500 iterates of the (arbitrarily chosen) point (1,1).

To be more concrete we shall investigate the pendulum with oscillating support for parameter values $A = 0.94$ and $B = 0.15$ only (see [1,4] for details and other parameter values); in addition we make the usual choice $v = \pi/2$. In this special setting most trajectories of (24) show complicated aperiodic behaviour. Simulations and experiments suggest the existence of a certain randomness according to the observation of a strange looking attractor of $\Psi_{0.94;0.15}$ in the Poincaré section (Fig. 2, cf. [4]).

The analysis to be performed naturally consists of three steps:

- discretization of the (interesting region in) phase space yielding a finite number of *cells*;
- numerical integration of (24) starting at one point (usually near the center) in each cell;
- construction of an index pair and interpretation of the resulting Conley index.

By means of the rigorous error bounds developed in the previous sections as well as growth considerations for the flow generated by (24), one obtains from the first two steps a *finite* model of the Poincaré section and the map $\Psi_{A;B}$. This rigorous but finite model turns the last step into a purely combinatorial task [8]. Although one is inevitably faced by multivalued maps, the index theory for multivalued dynamical systems successfully applied in [5] is not used throughout our analysis.

Despite the apparent simplicity of the outlined procedure some difficulties have to be overcome. Care must be taken due to the fact that the phase space $S^1 \times \mathbb{R}$ is *cylindrical* and has circumference 2π which can not be represented as a machine number. By means of the covering space $\mathbb{R}^2$ and some rescaling these difficulties can be avoided [1]. A more serious problem arises from the exponential divergence of trajectories of (24). With the relevant logarithmic norm [2]

$$\mu_\infty(DV(x_1, x_2, x_3)) \leqslant \max\{1, 1 + |A| - B\}$$

and the parameter values under consideration a *growth rate*, i.e., the number of cells rigorously containing the image of one cell under $\Psi_{A;B}$, of about 4.1×10^5 is obtained. With only the availability of standard workstations and PCs, this growth rate is definitely too large in order to find any non-trivial index information. Following [5] we use *intermediate sections*, the composition of which gives a much better approximation of the Poincaré map. However, one should note that the concept of intermediate sections will for example not be appropriate in the case of uniformly exponential growth. (In the latter, of course, one does not expect any interesting recurrent dynamics.)

In accordance to the procedure sketched above the computations were actually performed using the standard RK4 method (with stepsize $h = 2^{-10}$) and eight intermediate sections. The calculations in double precision arithmetic on an IBM power PC 603/120 MHz took about 57 min. (The interesting region in phase space was covered by more than 25,000 cells.) Due to the considerable growth rate numerical error bounds turn out to be negligible when compared with discretization effects. Such an observation seems to be not accidental: many chaotic systems (including the well-known Lorenz equations, cf. [1,6]) clearly exhibit errors on different orders of magnitude. Carefully choosing and discretizing an interesting region in phase space therefore is most important!

The main result is depicted in Fig. 3. The (weak) index pair (A_1, A_2) consists of seven disjoint parts giving rise (in the terminology of [7]) to a seven-dimensional (cohomological) Conley index $\mathrm{CH}^k_{\Psi_{0.94;\,0.15}}$ at $k = 1$,

$$\mathrm{CH}^1_{\Psi_{0.94;\,0.15}}((\mathrm{inv}_{\Psi_{0.94;\,0.15}}\overline{A_1 \setminus A_2} \cap B_i)_{i \in \{1,\ldots,7\}}; \mathbb{Q}) = [(\mathbb{Q}^7, (p_I \circ \psi)_{I \subseteq \{1,\ldots,7\}})]; \tag{25}$$

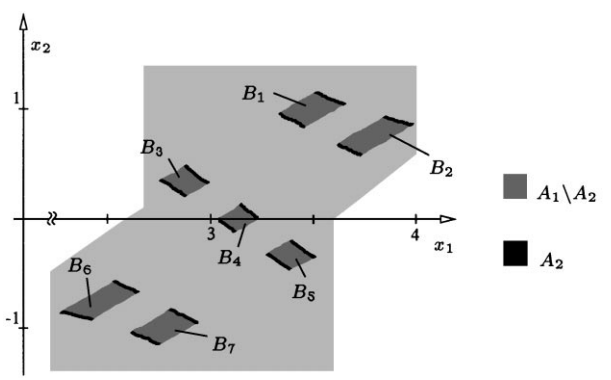
the indices at $k \neq 1$ are trivial. Here p_I denotes the projection of $\mathbb{Q}^7$ onto $\oplus_{i\in I}[e_i]$. The mapping ψ corresponding to the matrix representation (27) describes (on a cohomological level) the action of $\Psi_{0.94;0.15}$ on its invariant part in $\overline{A_1\backslash A_2}$. (Notice how the symmetry in (24) is reflected by A_ψ.) From (25) we may finally deduce some nontrivial insights concerning the Poincaré map:

(i) For each $p \in \mathbb{N}$ there is a periodic point for $\Psi_{0.94;0.15}$ in $\overline{A_1\backslash A_2}$ with primitive period p. (This observation is mainly based on the Lefschetz fixed point theorem; see [1,7,8].)

(ii) One can find a compact set $K \subseteq \mathrm{inv}_{\Psi_{0.94;0.15}}\overline{A_1\backslash A_2}$ and a continuous surjection $\rho : K \to \Sigma_{7,A_\psi}$ such that the diagram

$$\begin{array}{ccc} K & \xrightarrow{\Psi_{0.94;0.15}|_K} & K \\ \rho\downarrow & & \downarrow\rho \\ \Sigma_{7,A_\psi} & \xrightarrow{\sigma} & \Sigma_{7,A_\psi} \end{array} \tag{26}$$

commutes. Here σ denotes the shift operator on Σ_{7,A_ψ}, i.e., the space of all seven-symbol sequences that satisfy an admissibility condition represented by the matrix A_ψ [1]. As a consequence of (26) we obtain a *positive* lower bound for the topological entropy of the dynamical system $(K, \Psi_{0.94;0.15}|_K)$.



$$A_\psi := \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix} \tag{27}$$

Fig. 3. The (weak) index pair (A_1, A_2) for $\Psi_{0.94;0.15}$ consists of seven disjoint parts B_i. (Algorithmic investigations were performed within the lightly shaded union of cells.).

Using the efficient algebraic-topological tools of Conley index theory and the power of modern computers (thereby heavily relying on the rigorous error bounds from above) we have established several important dynamical features of a simple mechanical system. Due to (ii) we expect the latter to behave in a more or less unpredictable manner. This is exactly what can be observed by real-world experiments or numerical simulations [4].

References

[1] A. Berger, Anwendungen der Conley-Index-Theorie zum Nachweis chaotischen Systemverhaltens, Dissertation, TU Wien, 1997.
[2] E. Hairer, S.P. Nørsett, G. Wanner, Solving Ordinary Differential Equations I, Nonstiff Problems, 2nd Edition, Springer, Berlin, New York, Heidelberg, 1993.
[3] P. Henrici, Discrete Variable Methods in Ordinary Differential Equations, Wiley, New York, London, Sidney, 1962.
[4] R. Leven, B. Koch, B. Pompe, Chaos in Dissipativen Systemen, Akademie, Berlin, 1989.
[5] M. Mischaikow, M. Mrozek, Chaos in the Lorenz equations: a computer assisted proof, Bull. American Mathematical Society 32 (1995) 66–72.
[6] M. Mrozek, Rigorous error analysis of numerical algorithms via symbolic computations, J. Symbolic Comput. 22 (1996) 435–458.
[7] A. Szymczak, The Conley index for decomposition of isolated invariant sets, Fund. Math. 148 (1995) 71–90.
[8] A. Szymczak, A combinatorial procedure for finding isolating neighbourhoods and index pairs, preprint, 1996.