

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

Arno Berger

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

Keywords: Explicit RK-methods; Truncation and Round-off Error; Conley Index Theory

Abstract

Complicated dynamical systems can be rigorously analysed by means of Conley index theory. Necessary for such an analysis are 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.

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 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

*E-mail: aberger@mch2ws1.tuwien.ac.at

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 to 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 different 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

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

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 ,

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

If this method is of order p , the inequality

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

constitutes a well-known rigorous bound for the local error ([2]). For technical reasons $\|\cdot\|$ always denotes 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 i -th component of V is denoted by V^i while its partial derivative with respect to the i -th coordinate is symbolized by $V_{,i}$. Let $K \subseteq U$ be a compact set containing x_0 and

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

The $(p+1)$ -th 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

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

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

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

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

$$\|V_{,i_1 \dots i_{j_0}} V_{,i_{(j_0+1)} \dots i_{(j_0+j_1)}}^{i_1} \dots V_{,i_{(j_0+\dots+j_{p-1}+1)} \dots i_{(j_0+\dots+j_p)}}^{i_p}\| \leq D_{j_0} D_{j_1} \dots D_{j_p} ;$$

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

$$\|x^{(p+1)}\| \leq \sum_{\mathbf{t} \in T_{p+1}} \alpha(\mathbf{t}) D_{j_0(\mathbf{t})} D_{j_1(\mathbf{t})} \dots D_{j_p(\mathbf{t})} . \quad (6)$$

With increasing order p the determination of $\alpha(\mathbf{t})$ 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 $\alpha(\mathbf{t})$ up to $p=4$.)

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

$$k_i^{(r)} = \sum_{m=1}^r \sum_{\substack{i_1 \geq \dots \geq i_m \geq 1 \\ i_1 + \dots + i_m = r}} \beta_{m;i_1, \dots, i_m}^{(r)} V_{,j_1 \dots j_m} (hl_{i_1})^{(i_1); j_1} \dots (hl_{i_m})^{(i_m); j_m} . \quad (7)$$

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

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

As a consequence of the representation (7) we see that

$$\|k_i^{(r)}\| \leq \sum_{m=1}^r \sum_{\substack{i_1 \geq \dots \geq i_m \geq 1 \\ i_1 + \dots + i_m = r}} \beta_{m;i_1, \dots, i_m}^{(r)} D_m \| (hl_{i_1})^{(i_1)} \|_1 \dots \| (hl_{i_m})^{(i_m)} \|_1 . \quad (8)$$

Taking into account the obvious relations

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

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

$$\begin{aligned} \Phi(x_1, \dots, x_p) &:= \left(\beta_{1;1}^{(1)} D_1 x_1 , \beta_{1;2}^{(2)} D_1 x_2 + \beta_{2;1,1}^{(2)} D_2 x_1^2 , \dots \right. \\ &\quad \left. \dots , \sum_{m=1}^p \sum_{\substack{i_1 \geq \dots \geq i_m \geq 1 \\ i_1 + \dots + i_m = p}} \beta_{m;i_1, \dots, i_m}^{(p)} D_m x_{i_1} \dots x_{i_m} \right) \quad \text{and} \\ \Psi(x_1, \dots, x_p) &:= (D_0 + h x_1 , 2x_1 + h x_2 , \dots , p x_{p-1} + h x_p) \end{aligned}$$

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

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

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

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

we can replace (3) by the concise expression

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

Clearly $C_K(h)$ is a polynomial in h with coefficients calculated from the quantities D_r . An induction argument shows that

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

The subscript K in $C_K(h)$ displays the fact that the constructed error bound also depends on the compact set K (i.e. via the quantities D_r). In applications one typically tries to define this set as small as possible. One therefore has to check whether the solution $x(t)$ of (1) remains in K for all $0 \leq t \leq h$. 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 ($p \geq 5$) 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 $p \leq 4$.

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 $M_2 : \mathbb{R} \rightarrow \{2^k | k \in \mathbb{Z}\}$ by

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

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 $w(X_1, \dots, X_n)$ consisting of a finite number of symbols and implemented non-polynomial functions f_i from $\mathbb{M} \cup \{+, -, \cdot, (\cdot, \cdot)\} \cup \{(f_i)_{i=1}^m\}$ one defines recursively the evaluation $w(x_1, \dots, x_n)$ and the machine evaluation $\langle w \rangle(x_1, \dots, x_n)$ of w at (x_1, \dots, 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

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

and

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

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), (14) and (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

$$|w(x_1, \dots, x_n) - \langle w \rangle(x_1, \dots, x_n)| \leq \varepsilon_w(m_0, \dots, m_0) H_w(M_2(x_1), \dots, M_2(x_n)) \quad . \quad (16)$$

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 $H_w := (H_{w^1}, \dots, H_{w^n})$ (and analogously ε_w) for an arithmetic expression w having n components w^i . The n -tuple $(1, \dots, 1)$ is denoted by $\mathbf{1}$. Expressions like $M_2(u)$ and $\max\{u, v\}$ with real n -tuples u, v should be read coordinatewise.

Setting $\xi^i := \max_{x \in K} |x^i|$ ($1 \leq i \leq n$) with the compact set K discussed in the preceding section and using consequently the notation just introduced, we can find

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

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

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

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

$$E_K(h) := \max_{i=1}^n \left(H_{x_1}^i(M_2(\xi)) \varepsilon_{x_1}^i(m_0 \mathbf{1}) \right) \quad . \quad (19)$$

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

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

as the desired bound for the round-off error. Observe that due to (17) $E_K(h)$ will effectively not 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

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

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) obeys

$$\|x(Nh) - \langle x_1 \rangle_N\| \leq e^{hL} \|x((N-1)h) - \langle x_1 \rangle_{(N-1)}\| + h^{p+1} C_K(h) + E_K(h) .$$

It is well known from numerical mathematics that this recursion leads to

$$\|x(Nh) - \langle x_1 \rangle_N\| \leq \frac{e^{hLN} - 1}{e^{hL} - 1} (h^{p+1} C_K(h) + E_K(h)) =: G_K(h) . \quad (20)$$

With the functions $C_K(h)$ and $E_K(h)$ constructed earlier, (20) constitutes $G_K(h)$ as a ready-made rigorous error bound for the numerical integration of (1) by means of the RK-method (2). It comes as no surprise that due to our simple construction this error bound is very crude. In particular, the influence of rounding is usually dramatically overestimated in (20). Applications indicate however that $G_K(h)$ does quite satisfyingly reflect some important aspects of error analysis.

Example [Classical RK4-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

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

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

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

the coefficients of which are

$$\begin{aligned} c_0 &= \frac{49}{2880} D_0^4 D_4 + \frac{169}{1440} D_0^3 D_1 D_3 + \frac{31}{480} D_0^3 D_2^2 + \frac{47}{240} D_0^2 D_1^2 D_2 + \frac{1}{120} D_0 D_1^4 , \\ c_1 &= \frac{19}{1151} D_0^4 D_1 D_4 + \frac{7}{384} D_0^4 D_2 D_3 + \frac{31}{288} D_0^3 D_1^2 D_3 + \frac{5}{48} D_0^3 D_1 D_2^2 + \frac{5}{64} D_0^2 D_1^3 D_2 , \\ c_2 &= \frac{91}{4608} D_0^4 D_1^2 D_4 + \frac{31}{768} D_0^4 D_1 D_2 D_3 + \frac{11}{1536} D_0^4 D_2^2 + \frac{1}{12} D_0^3 D_1^3 D_3 + \frac{11}{128} D_0^3 D_1^2 D_2^2 + \frac{1}{32} D_0^2 D_1^4 D_2 , \\ c_3 &= \frac{23}{1536} D_0^4 D_1^3 D_4 + \frac{163}{4608} D_0^4 D_1^2 D_2 D_3 + \frac{29}{3072} D_0^4 D_1 D_2^2 + \frac{25}{576} D_0^3 D_1^4 D_3 + \frac{1}{24} D_0^3 D_1^3 D_2^2 , \\ c_4 &= \frac{157}{18432} D_0^4 D_1^4 D_4 + \frac{191}{9216} D_0^4 D_1^3 D_2 D_3 + \frac{19}{3072} D_0^4 D_1^2 D_2^2 + \frac{5}{384} D_0^3 D_1^5 D_3 + \frac{5}{768} D_0^3 D_1^4 D_2^2 , \end{aligned}$$

$$\begin{aligned}
c_5 &= \frac{43}{12288}D_0^4D_1^5D_4 + \frac{17}{2304}D_0^4D_1^4D_2D_3 + \frac{5}{3072}D_0^4D_1^3D_2^3 + \frac{1}{384}D_0^3D_1^6D_3 , \\
c_6 &= \frac{5}{4608}D_0^4D_1^6D_4 + \frac{1}{576}D_0^4D_1^5D_2D_3 + \frac{1}{12288}D_0^4D_1^4D_2^3 , \\
c_7 &= \frac{1}{4608}D_0^4D_1^7D_4 + \frac{1}{6144}D_0^4D_1^6D_2D_3 , \\
c_8 &= \frac{1}{36864}D_0^4D_1^8D_4 .
\end{aligned}$$

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

$$\begin{aligned}
H_{x_1}(M_2(\xi)) &= \max\left\{2M_2(\xi), M_2(h) \max\{8H_{k_1}(M_2(\xi)), 8H_{k_2}(M_2(\xi)), 4H_{k_3}(M_2(\xi)), H_{k_4}(M_2(\xi))\}\right\} , \\
\eta_{i+1} &= H_V(2 \max\{M_2(\xi), 2a_{i+1,i}M_2(h)\eta_i\}) .
\end{aligned}$$

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

$$V(x_1, x_2) := (x_1 - x_1^2, -\frac{1}{2}x_2 + \frac{1}{2}x_1x_2) . \quad (21)$$

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

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

and consequently

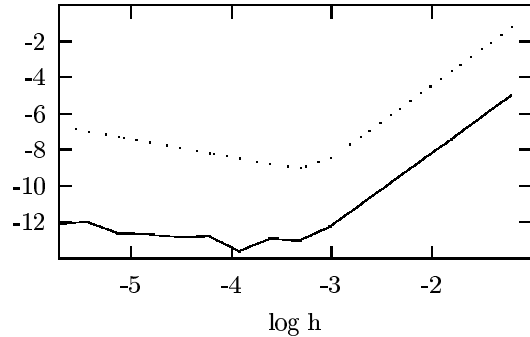
$$C_K(h) = \frac{16816}{15} + 3440h + \frac{35264}{3}h^2 + 20288h^3 + 24576h^4 + 11520h^5 + 3456h^6 .$$

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

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

A Lipschitz constant of the (local) flow generated by (21) 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 $h_i := 2^{-i}$ ($4 \leq i \leq 20$) 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 G_K provides a reasonable (upper) estimate for the optimal stepsize (figure 1).

Figure 1: The real error $\log \|x(1) - \langle x_1 \rangle_{[\frac{1}{h}]}\|$ (full line) and the error bound $\log G_K(h)$ for the vectorfield (21)



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] and [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 $\alpha \cos \omega t$ (figure 2). Its equation of motion reads

$$\ddot{\varphi} + \frac{\beta}{ml^2} \dot{\varphi} + \left(\frac{g}{l} + \frac{\alpha \omega^2}{l} \cos \omega t \right) \sin \varphi = 0, \quad (22)$$

where only *linear* frictional effects have been considered via the parameter β . By introducing nondimensional time and frequency, $\tau := t\sqrt{g/l}$ and $\nu := \omega\sqrt{l/g}$, (22) can be equivalently written as

$$\frac{d}{d\tau} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_2 \\ -(1 + A \cos x_3) \sin x_1 - Bx_2 \\ \nu \end{pmatrix} \quad (23)$$

with the abbreviations $A := \frac{\alpha \omega^2}{g}$ and $B := \frac{\beta}{ml\sqrt{gl}}$. Integration of (23) over $[0; \frac{2\pi}{\nu}]$ 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]).

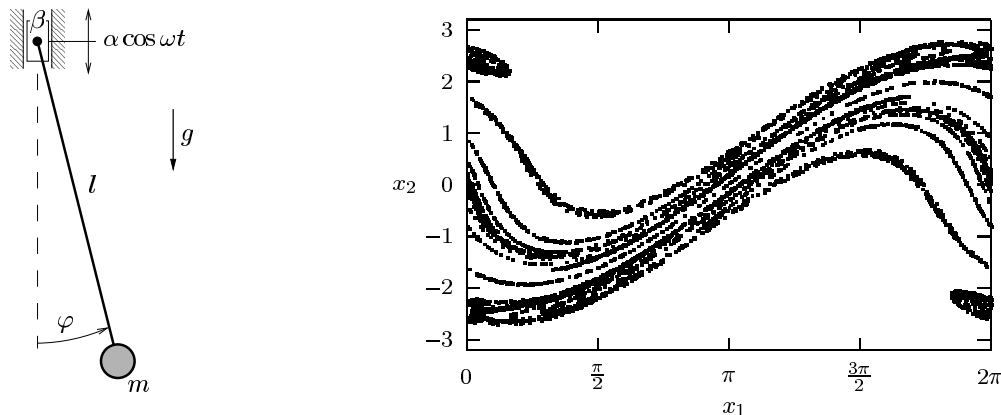


Figure 2: The pendulum with oscillating support (left) and the attractor for the Poincaré map $\Psi_{0.94;0.15}$ displayed via 7500 iterates of the (arbitrarily chosen) point (1,1)

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] and [4] for details and other parameter values); in addition we make the usual choice $\nu = \frac{\pi}{2}$. In this special setting most trajectories of (23) 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 (figure 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 (23) 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 (23), 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't 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 (23). With the relevant logarithmic norm ([2])

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

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 $4.1 \cdot 10^5$ is obtained. With only the availability of standard workstations and PCs, this growth rate is definitely too large in order to find any nontrivial 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 603/120Mhz took about 57 minutes. (The interesting region in phase space was covered by more than 25,000 cells.)

The main result is depicted in figure 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 $CH_{\Psi_{0.94;0.15}}^k$ at $k = 1$,

$$CH_{\Psi_{0.94;0.15}}^1((\text{inv}_{\Psi_{0.94;0.15}} \overline{A_1 \setminus A_2} \cap B_i)_{i \in \{1, \dots, 7\}}; \mathbb{Q}) = [(\mathbb{Q}^7, (p_I \circ \psi)_{I \subseteq \{1, \dots, 7\}})] ; \quad (24)$$

the indices at $k \neq 1$ are trivial. Here p_I denotes the projection of \mathbb{Q}^7 onto $\bigoplus_{i \in I} [e_i]$. The mapping ψ corresponding to the matrix representation (26) describes (on a cohomological level) the action of $\Psi_{0.94;0.15}$ on its invariant part in $\overline{A_1 \setminus A_2}$. From (24) 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 $\overline{A_1 \setminus A_2}$ 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 $K \subseteq \text{inv}_{\Psi_{0.94;0.15}} \overline{A_1 \setminus A_2}$ and a continuous surjection $\rho : K \rightarrow \Sigma_{7, A_\psi}$ such that the diagram

$$\begin{array}{ccc} K & \xrightarrow{\Psi_{0.94;0.15}|_K} & K \\ \rho \downarrow & & \downarrow \rho \\ \Sigma_{7, A_\psi} & \xrightarrow{\sigma} & \Sigma_{7, A_\psi} \end{array} \quad (25)$$

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 (25) we obtain a *positive* lower bound for the entropy of the dynamical system $(K, \Psi_{0.94;0.15}|_K)$.

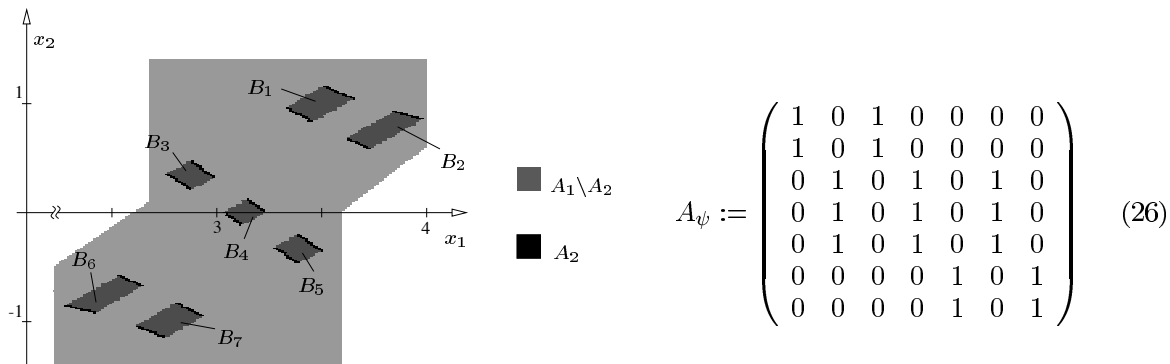


Figure 3: The (weak) index pair (A_1, A_2) for $\Psi_{0.94;0.15}$ consists of seven disjoint parts B_i . (Algorithmic investigations were performed within the lightly shaded union of cells.)

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*, Diss. TU Wien, 1997 ;
- [2] E. Hairer, S.P. Nørsett, G. Wanner, *Solving Ordinary Differential Equations I, Nonstiff Problems (2nd ed.)*, 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. AMS* **32**(1995), 66-72 ;
- [6] M. Mrozek, Rigorous error analysis of numerical algorithms via symbolic computations, *J. Symb. Comp.* **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 .