Numer. Math. (1997) 76: 441–462

Numerische
Mathematik
© Springer-Verlag 1997
Electronic Edition

The life-span of backward error analysis for numerical integrators

E. Hairer[1], **Ch. Lubich**[2]

[1] Département de Mathématiques, Université de Genève, CH-1211 Genève 24, Switzerland; e-mail: Ernst.Hairer@math.unige.ch

[2] Mathematisches Institut, Universität Tübingen, Auf der Morgenstelle 10, D-72076 Tübingen, Germany; e-mail: lubich@na.uni-tuebingen.de

Received October 18, 1995 / Revised version received February 28, 1996

Summary. Backward error analysis is a useful tool for the study of numerical approximations to ordinary differential equations. The numerical solution is formally interpreted as the exact solution of a perturbed differential equation, given as a formal and usually divergent series in powers of the step size. For a rigorous analysis, this series has to be truncated.

In this article we study the influence of this truncation to the difference between the numerical solution and the exact solution of the perturbed differential equation. Results on the long-time behaviour of numerical solutions are obtained in this way. We present applications to the numerical phase portrait near hyperbolic equilibrium points, to asymptotically stable periodic orbits and Hopf bifurcation, and to energy conservation and approximation of invariant tori in Hamiltonian systems.

Mathematics Subject Classification (1991): 65L05, 65L70

1. Introduction

Since the work of Wilkinson (1960), backward error analysis has become a well-established tool in numerical linear algebra. For the numerical treatment of ordinary differential equations, its use is only very recent. The papers by Eirola (1993), Feng Kang (1991), Fiedler and Scheurle (1996), Sanz-Serna (1992), and Yoshida (1993) appear to be among the first studies on this topic. The idea is to interpret the numerical solution as the exact solution of a perturbed differential equation. Properties of the perturbed equation can then be converted into properties of the numerical solution yielding new insight into numerical integrators.

We consider the system of ordinary differential equations

Correspondence to: E. Hairer

$$y' = f(y)\ , \qquad y(0) = y_0 \tag{1.1}$$

and assume that f is sufficiently differentiable. A numerical one-step method is a recursion $y_{n+1} = \Phi_h(y_n)$ that gives an approximation y_n to the solution of (1.1) at $t_n = nh$. We are looking for a perturbed differential equation

$$\widetilde{y}\,' = f(\widetilde{y}) + hf_2(\widetilde{y}) + h^2 f_3(\widetilde{y}) + \ldots\ , \qquad \widetilde{y}(0) = y_0, \tag{1.2}$$

such that the numerical solution y_1 is formally equal to $\widetilde{y}(h)$. Expanding y_1 and $\widetilde{y}(h)$ into a Taylor series and comparing equal powers of h yields recursion formulas for the coefficient functions $f_j(y)$ (explicit formulas for $f_j(y)$ are given in Hairer (1994)). For a method of order p, the functions $f_2(y), \ldots, f_p(y)$ vanish identically.

Unfortunately, the series in (1.2) does not converge in general. This is already the case for the simple differential equation $y' = f(t)$ and the trapezoidal rule, for which (1.2) is equivalent to the Euler-MacLaurin summation formula. For a truncated series (1.2), the values y_1 and $\widetilde{y}(h)$ are no longer equal.

In Sect. 2 we study the difference $y_1 - \widetilde{y}(h)$ when the series in Eq. (1.2) is suitably truncated. A rigorous estimate is given for the case of analytic f (Theorem 1). Its proof is postponed to Sect. 4. The global error $y_n - \widetilde{y}(nh)$ is analyzed in Sect. 3, and estimates are given for the time span on which this error remains small. We give applications to the approximation of phase portraits near hyperbolic equilibrium points, to asymptotically stable periodic orbits and Hopf bifurcation, and to long-time energy conservation and approximation of invariant tori in Hamiltonian systems.

A result similar to Theorem 1, together with applications to Hamiltonian systems (in particular long-time energy conservation like in Corollary 6 below), has recently been published by Benettin and Giorgilli (1994). Unlike their result, our bounds are directly in terms of data of the original differential equation (1.1). Our approach is however more restrictive in the sense that we consider numerical approximations which can be represented as a B-series. This also allows us to obtain slightly sharper estimates. Our proof is completely different from that of Benettin and Giorgilli (1994).

2. Local error of backward analysis

We consider for concreteness a Runge-Kutta method of order p applied to (1.1),

$$y_1 = y_0 + h\sum_{i=1}^{s} b_i k_i\ , \qquad k_i = f\Big(y_0 + h\sum_{j=1}^{s} a_{ij} k_j\Big), \qquad i = 1, \ldots, s\ . \tag{2.1}$$

Here, $h > 0$ is the step size, and b_i, a_{ij} are the Runge-Kutta coefficients. We truncate the series of the corresponding perturbed differential equation (1.2) after the Nth term. This yields

$$\widetilde{y}\,' = \widetilde{f}(\widetilde{y})\ , \qquad \widetilde{y}(0) = y_0 \tag{2.2a}$$

with $\widetilde{f}(y)$ (depending on h and on N) of the form

$$\text{(2.2b)} \qquad \widetilde{f}(y) = f(y) + h^p f_{p+1}(y) + h^{p+1} f_{p+2}(y) + \ldots + h^{N-1} f_N(y) \ .$$

By construction of the $f_j(y)$ we have that

$$\text{(2.3)} \qquad y_1 - \widetilde{y}(h) = O(h^{N+1}) \ .$$

The truncation index N is arbitrary when f is infinitely differentiable. In general, the series (2.2b) diverges as $N \to \infty$, and the constants hidden in the $O(h^{N+1})$ bounds of (2.3) tend to infinity with N, even if f is analytic. It is thus of interest to know how close the numerical solution y_1 can come to the solution $\widetilde{y}(h)$ of a perturbed differential equation (2.2).

In the following theorem this question is studied for complex functions $f : \mathbb{C}^d \to \mathbb{C}^d$ which are analytic in a neighbourhood of $y_0 \in \mathbb{C}^d$. We call "polydisc of radius R around y_0" the set $\{(z_1, \ldots, z_d) : |z_j - y_{0j}| \leq R \text{ for } j = 1, \ldots, d\}$, and we shall work with the ℓ_1-norm on $\mathbb{C}^d$ (because of the bound (4.3) below). Moreover, we consider the method dependent constants κ, μ, and ν defined by

$$\text{(2.4)} \qquad \|(a_{ij})\| \leq \kappa \ , \quad \|(b_i)\| \leq \mu \ , \quad \nu = \kappa + \mu/(2\log 2 - 1) \ .$$

The norm used in $\|(a_{ij})\|$ is the matrix norm corresponding to $\|(b_i)\|$.

Theorem 1. *Let f be analytic and bounded by M in a polydisc of radius R around y_0. If $h \leq h_0/2$ with $h_0 = R/(2\mathrm{e}^2\nu M)$, then there exists $N = N(h)$ (namely N equal to the largest integer satisfying $hN \leq h_0$) such that the difference between the solution of* (2.2) *and the numerical result after one step is bounded by*

$$\text{(2.5)} \qquad \|y_1 - \widetilde{y}(h)\| \leq 0.032\, R \cdot \mathrm{e}^{-h_0/h}.$$

Theorem 1 is not restricted to Runge-Kutta methods. In fact, the proof shows that the result is valid for any method whose result after one step is given as a B-series with suitably (and not restrictively) bounded coefficients. See, e.g., Hairer, Nørsett and Wanner (1993) for the definition of B-series. As in Hairer (1994), the result extends also to partitioned methods.

The proof of Theorem 1 also yields that the derivatives of y_1 and $\widetilde{y}(h)$ with respect to the initial value differ by no more than the right-hand side of (2.5) (with different h_0). This result could alternatively be obtained by a direct application of Theorem 1, by noting first that $(y_1, \partial y_1/\partial y_0)$ is the numerical solution of the Runge-Kutta method applied to the system consisting of the original differential equation (1.1) and its variational equation, and by noting further that the associated perturbed system consists of the perturbed differential equation (2.2) together with *its* variational equation. Although we will not use it in the present paper, such a C^1 approximation result is of interest in a variety of situations in the numerical analysis of dynamical systems. In particular, it would allow us to apply all the results in Stuart's (1994) review article to the comparison of the dynamics of the numerical method and the perturbed differential equation (2.2).

We defer the proof of Theorem 1 to Sect. 4, and study first some implications of Theorem 1 to the long-time behaviour of numerical solutions of differential equations.

3. Global error of backward analysis

3.1. General bounds

We now ask over which time interval the solution of the perturbed equation (2.2) remains close to the numerical solution, or in other words: What is the ‘life-span’ of backward error analysis? We denote by $y(t,s,z)$ the solution at time t of the differential equation $y' = f(y)$ that satisfies $y(s) = z$.

Corollary 2. *Let the numerical solution y_n and the exact solution $y(t)$ of* (1.1) *for sufficiently long time be contained in a compact subset K of the region of analyticity of f. Let U be a neighbourhood of K, and assume that for $z \in U$ and $t > s$ with $y(t,s,z) \in U$*

$$\left\| \frac{\partial y}{\partial z}(t,s,z) \right\| \le C\,\mathrm{e}^{\mu(t-s)} , \qquad \left\| \frac{\partial^2 y}{\partial z^2}(t,s,z) \right\| \le C\,\mathrm{e}^{\mu(t-s)} . \tag{3.1}$$

Over a time interval of length

$$T = \begin{cases} \infty & \\ O(h^{-(p+1)}) & \textit{for} \\ O(h^{-1}) & \end{cases} \begin{matrix} \mu < 0 \\ \mu = 0 \\ \mu > 0 \, , \end{matrix}$$

the numerical solution and the solution of the perturbed equation of Theorem 1 then remain exponentially close:

$$y_n - \widetilde{y}(nh) = O(\mathrm{e}^{-h^*/2h}) \qquad \textit{for} \quad nh < T \ ,$$

where h^* is the minimum, taken over the compact set K, of the numbers $h_0 = h_0(y_0)$ of Theorem 1.

Proof. (a) We use the Gröbner-Alekseev formula (see Hairer, Nørsett and Wanner (1993), p. 96) to show that the perturbed equation (2.2) propagates errors similarly to the original equation (1.1). Let $y(t), z(t)$ be solutions of (1.1) to initial values y_0, z_0, and let $\widetilde{y}(t), \widetilde{z}(t)$ be the solutions of the perturbed equation (2.2) to these initial values. Writing $\widetilde{f}(y) = f(y) + h^p g(y)$, the Gröbner-Alekseev formula reads

$$\widetilde{y}(t) = y(t) + h^p \int_0^t \frac{\partial y}{\partial z}\big(t,s,\widetilde{y}(s)\big)\, g\big(\widetilde{y}(s)\big)\, ds \ ,$$

and similarly for $\widetilde{z}(t)$ and $z(t)$. Forming the difference of these two formulas and taking norms, we obtain by the assumption (3.1)

$$\|\widetilde{y}(t) - \widetilde{z}(t)\| \le C\,\mathrm{e}^{\mu t}\|y_0 - z_0\| + h^p \int_0^t L\mathrm{e}^{\mu(t-s)}\|\widetilde{y}(s) - \widetilde{z}(s)\|\, ds \ ,$$

with $L = CB_g + CL_g$, where B_g and L_g are respectively a bound and a Lipschitz bound of g. (These can be obtained, uniformly in h, from Lemma 11 in Sect. 4.) The Gronwall lemma then gives us

$$\|\widetilde{y}(t) - \widetilde{z}(t)\| \le C\,\mathrm{e}^{(\mu+Lh^p)t}\,\|y_0 - z_0\| \ . \tag{3.2}$$

(b) We use Lady Windermere's fan (see Hairer, Nørsett and Wanner (1993), p. 160) to bound the difference between y_n and $\widetilde{y}(t_n)$. Let us denote by $\widetilde{y}(t,s,z)$ the solution of (2.2) with $y(s) = z$. By (3.2), we have

$$\begin{aligned} \|\widetilde{y}(t_n) - y_n\| &\leq \sum_{k=1}^{n} \|\widetilde{y}(t_n, t_{k-1}, y_{k-1}) - \widetilde{y}(t_n, t_k, y_k)\| \\ &\leq \sum_{k=1}^{n} C\,\mathrm{e}^{(\mu+Lh^p)(t_n-t_k)} \|\widetilde{y}(t_k, t_{k-1}, y_{k-1}) - \widetilde{y}(t_k, t_k, y_k)\| \ , \end{aligned}$$

where we note that $\widetilde{y}(t_k, t_k, y_k) = y_k$. By Theorem 1, we have

$$\|\widetilde{y}(t_k, t_{k-1}, y_{k-1}) - y_k\| \leq C\,\mathrm{e}^{-h^*/h} \ .$$

Inserting this bound and summing up the geometric series gives the result. □

Remark. The conclusion of Corollary 2 remains valid when condition (3.1) is replaced by the one-sided Lipschitz condition

$$\mathrm{Re}\,\langle f(y) - f(z), y - z\rangle \leq \mu\|y - z\|^2 \qquad \text{for} \ \ y, z \in U \ ,$$

where $\langle\cdot,\cdot\rangle$ is some inner product with associated norm $\|\cdot\|$.

3.2. Near hyperbolic equilibrium points

In some situations, the numerical solution and a solution of the associated perturbed differential equation remain close for arbitrarily long times even when solutions diverge exponentially. This is achieved by shadowing, where one gives up the strict requirement that $\widetilde{y}(0) = y_0$. For example, this works near hyperbolic equilibrium points of (1.1). We recall that such points are characterized by the property that f vanishes there and the Jacobian has no eigenvalues on the imaginary axis.

Corollary 3. *Let y^* be a hyperbolic equilibrium point of* (1.1). *Then, there is a neighbourhood U of y^* such that for sufficiently small step sizes there exists a perturbed differential equation* (2.2) *with the following property: To every numerical solution (y_n) in U, there exists a solution $\widetilde{y}(t)$ of the perturbed differential equation satisfying*

$$\|y_n - \widetilde{y}(nh)\| \leq \mathrm{e}^{-c/h} \quad \text{\textit{as long as} } y_n \in U$$

with $c > 0$. Conversely, to every solution $\widetilde{y}(t)$ in U, there is a numerical solution (y_n) satisfying the above inequality.

With Theorem 1, the result follows from the shadowing result of one-step methods, see Beyn (1987a), Sanz-Serna and Larsson (1993).

Up to exponentially small terms, the phase portrait of the numerical method is therefore identical to that of a perturbed differential equation near the hyperbolic equilibrium. In particular, the stable and unstable manifold of the discrete scheme are exponentially close to those of a perturbed differential equation. See also Eirola (1993) and Garay (1993) for different backward analysis results near hyperbolic fixed points.

3.3. Asymptotically stable periodic orbits

Corollary 4. *Assume that the differential equation* (1.1) *with analytic right-hand side has an asymptotically stable periodic orbit. For sufficiently small step sizes, the numerical method then has an asymptotically stable invariant closed curve which is exponentially close to the periodic orbit of a perturbed differential equation* (2.2)*.*

Proof. The existence of a uniformly (in h) asymptotically stable invariant closed curve for the numerical method is shown by Braun and Hershenov (1977) and Beyn (1987b). For the perturbed differential equation (2.2), the existence of an asymptotically stable periodic orbit follows by considering the Poincaré map of (1.1) at a cross-section, which is a contraction in a suitable norm and hence preserves fixed points under small perturbations. It remains to be shown that the two curves are exponentially close. We choose a cross-section S at a point x^* on the invariant curve of the numerical method. As a substitute for the Poincaré map, we consider the family of maps $P_\theta : S \to S$ defined as follows for $0 \le \theta \le 1$: Starting from a point x on S, we first go a time interval of length θh backwards via (2.2) and take $y_0 = \widetilde{y}(-\theta h, 0, x)$ as starting value of the numerical method, with which we compute $y_1, \ldots, y_M$ until there exists $\tau \in [0, 1)$ such that $z = \widetilde{y}(\tau h, 0, y_M) \in S$. We then set $P_\theta(x) = z$. From the asymptotic stability of the periodic orbit of (2.2) and Theorem 1 (or rather Corollary 2 on a bounded time interval) we conclude that there exists $\kappa < 1$ such that in a suitable norm

$$\|P_\theta(x) - P_\sigma(y)\| \le \kappa \|x - y\| + \mathrm{e}^{-c/h} \tag{3.3}$$

with $c > 0$ for any $x, y \in S$ in a neighbourhood of the periodic orbit, and for any $\theta, \sigma \in [0, 1]$. Again by Theorem 1 (or Corollary 2), we obtain

$$\|P_\theta(\widetilde{x}) - \widetilde{x}\| \le \mathrm{e}^{-c/h} \tag{3.4}$$

for $\widetilde{x} \in S$ on the periodic orbit of (2.2) and arbitrary $\theta \in [0, 1]$. By (3.3), we have for any sequence $x_{k+1} = P_{\theta_k}(x_k)$

$$\|x_k - x_0\| \le C \|x_1 - x_0\| + C\,\mathrm{e}^{-c/h}$$

with $C = 1/(1-\kappa)$. Choosing $x_0 = \widetilde{x}$ and letting k become large, we obtain with (3.4) for a suitable choice of the shifts θ_j

$$\|x^* - \widetilde{x}\| \le 3C\,\mathrm{e}^{-c/h} ,$$

and hence the result follows. □

3.4. Hopf bifurcation

Via the arguments of the preceding proof, Theorem 1 provides a relatively simple means to study Hopf bifurcation under discretization. Consider a parametrized differential equation

$$y' = f(y, \lambda) \tag{3.5}$$

that has y^* as an equilibrium point for all real λ in a neighbourhood of 0. Let f be analytic with respect to y and at least five times continuously differentiable with respect to (y, λ). We assume that the Jacobian $\partial f / \partial y$ has a complex conjugate pair of eigenvalues $\lambda \pm \mathrm{i}\omega(\lambda)$ with $\omega(0) \neq 0$, and that all other eigenvalues are in the left half-plane and bounded away from the imaginary axis. If a certain coefficient a is negative, which depends in a complicated way on derivatives of f at $(y^*, 0)$ up to order 3, then the system undergoes a *supercritical Hopf bifurcation* at $\lambda^* = 0$ (see, e.g., Wiggins (1990), Ch.3.1B): For sufficiently small $\lambda > 0$, the system has an asymptotically stable periodic orbit which, in suitable coordinates, is $O(\lambda)$ close to a circle of radius $\sqrt{\lambda/(-a)}$.

Corollary 5. *In the above situation, let Eq.* (3.5) *be discretized by a* symmetric *Runge-Kutta method. Then, there exist* $\overline{\lambda} > 0$ *and* $\overline{h} > 0$ *such that the following is valid for* $h \leq \overline{h}$*: There exists a perturbed differential equation (parametrized by* λ*) that undergoes a supercritical Hopf bifurcation at the* same *parameter value* $\lambda^* = 0$ *and satisfies the local error bound of Theorem 1. The numerical scheme has for* $0 < \lambda \leq \overline{\lambda}$ *an asymptotically stable invariant closed curve which is* $\mathrm{e}^{-c/h}$ *close to the periodic orbit of the perturbed differential equation. Here* $c > 0$ *is independent of* λ *and* h.

Proof. (a) We show first that there exists a perturbed parametrized differential equation which has a supercritical Hopf bifurcation at $\lambda^* = 0$ and satisfies the local error bound of Theorem 1. We may assume that the equilibrium point is $y^* = 0$, and we rewrite (3.5) as

$$y' = A(\lambda)y + g(y, \lambda) \tag{3.6}$$

with $A(\lambda) = \partial f / \partial y(0, \lambda)$, so that $g(0, \lambda) = 0$ and $\partial g/\partial y(0, \lambda) = 0$. Then, the numerical method takes the form

$$y_{n+1} = R(hA(\lambda))y_n + h\varphi(y_n, \lambda) \ ,$$

where $R(z)$ is the stability function of the method, and φ is a h-dependent function with $\varphi(0, \lambda) = 0$ and $\partial\varphi/\partial y(0, \lambda) = 0$. Instead of (2.2), we choose a modified perturbed equation

$$\widetilde{y}\,' = \widetilde{A}(\lambda)\widetilde{y} + \widetilde{g}(\widetilde{y}, \lambda) \tag{3.7a}$$

where $\widetilde{A}(\lambda) = h^{-1} \log R(hA(\lambda))$ and $\widetilde{g}$ is of the form

$$\widetilde{g}(y, \lambda) = g(y, \lambda) + h^p g_{p+1}(y, \lambda) + \ldots + h^{N-1} g_N(y, \lambda) \tag{3.7b}$$

with $\widetilde{g}(0,\lambda) = 0$ and $\partial\widetilde{g}/\partial y(0,\lambda) = 0$. The above choice of $\widetilde{A}(\lambda)$ gives $\widetilde{y}(t_n) \equiv y_n$ when $g \equiv 0$. We expand $\widetilde{A}(\lambda)$ as a (convergent) series in h,

$$\widetilde{A} = A + h^p A_{p+1} + h^{p+1} A_{p+2} + \ldots\ ,$$

and we note that the expansion terms of (2.2), as constructed in Hairer (1994), are of the form

$$f_k(y,\lambda) = A_k(\lambda)y + g_k(y,\lambda)$$

with $g_k(0,\lambda) = 0$ and $\partial g_k/\partial y(0,\lambda) = 0$. Hence, (3.7) differs from (2.2) only in the remainder term of the convergent series for $\widetilde{A}$, which is bounded by $(Ch)^N$ (another difference is in the obvious dependence on λ). Therefore, the estimate of Theorem 1 remains valid for (3.7) with a suitable N, and moreover it is uniform for λ varying in bounded intervals.

Since the method is assumed to be symmetric, the stability function has unit modulus along the imaginary axis. It follows that for small λ also $\widetilde{A}(\lambda)$ has a pair of complex conjugate eigenvalues with nonvanishing imaginary part which cross the imaginary axis at $\lambda^* = 0$ with positive speed as λ increases, while the other eigenvalues have strictly negative real part for small λ and h. Therefore, (3.7) undergoes a Hopf bifurcation at $\lambda^* = 0$. Because of the genericity assumption $a < 0$ about f, the bifurcation is again supercritical.

(b) For a given small positive λ, let $\widetilde{P}$ be the Poincaré map of (3.7) associated with a cross-section S, and let P_θ be the Poincaré-like map on S for the numerical method, as defined in the proof of Corollary 4. By asymptotic stability and Theorem 1 we have in a suitable norm

$$\|\widetilde{P}(x) - P_\theta(y)\| \le \kappa\|x - y\| + \mathrm{e}^{-2c/h} \tag{3.8}$$

for $x, y \in S$ near the periodic orbit of (3.7) and for arbitrary $\theta \in [0,1]$, with

$$\kappa \le 1 - K\lambda$$

for some positive constant K. The norm in (3.8) can be chosen independently of λ for small λ, as can be seen from the normal form (Wiggins (1990), p. 271). Consequently, also $c > 0$ in (3.8) is independent of λ. Defining the sequences $x_{k+1} = P_{\theta_k}(x_k)$ and $\widetilde{x}_{k+1} = \widetilde{P}(\widetilde{x}_k)$ with starting value $x_0 = \widetilde{x}_0$ in the basin of attraction of the periodic orbit of (3.7), we obtain

$$\|x_{k+1} - \widetilde{x}_{k+1}\| \le \frac{1}{1-\kappa}\,\mathrm{e}^{-2c/h} = O(\mathrm{e}^{-c/h}) \qquad \text{for} \quad \lambda \ge \mathrm{e}^{-c/h}\ .$$

Letting k become large, we conclude that the numerical solution is for sufficiently large times exponentially close to the periodic orbit of (3.7), uniformly for $\mathrm{e}^{-c/h} \le \lambda \le \overline{\lambda}$.

(c) We can show that there is actually an invariant closed curve for the numerical method for small positive λ by invoking the Naimark-Sacker theorem (Wiggins (1990), Ch. 3.2C) (“Hopf bifurcation for maps”). This yields immediately that for sufficiently small h, there is a $\overline{\lambda}_h > 0$ such that there exists an asymptotically stable invariant closed curve for $0 < \lambda \le \overline{\lambda}_h$. A more detailed investigation, which we omit here, shows that $\overline{\lambda}_h$ can in fact be chosen independent of h. □

Remark. When the method is not symmetric, then the perturbed differential equation (3.7) has a Hopf bifurcation at $\widetilde{\lambda} = \lambda^* + O(h^p)$. Otherwise, Corollary 5 remains valid for non-symmetric methods.

3.5. Hamiltonian systems: energy conservation

We consider long-time energy approximation of symplectic numerical schemes for the integration of Hamiltonian systems

$$p' = -\frac{\partial H}{\partial q}(p,q)\ , \qquad q' = \frac{\partial H}{\partial p}(p,q)\ . \tag{3.9}$$

It was shown independently by Benettin and Giorgilli (1994) and Hairer (1994) that for a *symplectic* numerical method the perturbed differential equation (2.2) is again a Hamiltonian system, with perturbed Hamiltonian

$$\widetilde{H}(p,q) = H(p,q) + h^p H_{p+1}(p,q) + \ldots + h^{N-1} H_N(p,q)\ . \tag{3.10}$$

Hairer's proof shows in addition that the perturbation functions H_k are composed of derivatives of H. This has the important global consequence that the perturbed Hamiltonian $\widetilde{H}$ is analytic (and single-valued) on the same domain as the original Hamiltonian H.

The following is a variant of a result that was previously obtained by the same argument as below by Benettin and Giorgilli (1994) as a corollary to their version of Theorem 1.

Corollary 6. *Suppose that the numerical solution* (p_n, q_n) *given by a symplectic method applied to* (3.9) *stays in a compact subset K of the region of analyticity of H. Over a time interval of length*

$$T = h\, \mathrm{e}^{h^*/2h}\ ,$$

the perturbed Hamiltonian $\widetilde{H}$, with $N = N(h)$ of Theorem 1, remains nearly constant along the numerical solution:

$$\widetilde{H}(p_n, q_n) = \widetilde{H}(p_0, q_0) + O(\mathrm{e}^{-h^*/2h}) \qquad \textit{for} \ \ nh \le T\ .$$

Here h^* is the minimum, taken over the compact set K, of the numbers h_0 of Theorem 1.

In particular, the original Hamiltonian is well conserved over exponentially long time:

$$H(p_n, q_n) = H(p_0, q_0) + O(h^p) \qquad \text{for} \ \ nh \le T\ .$$

Proof. We denote by $\widetilde{y}(t,s,z)$ the solution at time t of the system with Hamiltonian $\widetilde{H}$ that starts at z at time s. Since the flow of this system conserves $\widetilde{H}$, the values $\widetilde{H}(\widetilde{y}(t,s,z)) = \widetilde{H}(z)$ are independent of t and s, and hence we have with $y_k = (p_k, q_k)$

$$\widetilde{H}(p_n,q_n) - \widetilde{H}(p_0,q_0) = \sum_{k=1}^{n} \Big(\widetilde{H}(y_k) - \widetilde{H}(\widetilde{y}(t_k, t_{k-1}, y_{k-1})) \Big) \ .$$

The result now follows from Theorem 1 upon using an h-independent Lipschitz bound for $\widetilde{H}$, which is obtained from Lemma 11 below. □

3.6. Hamiltonian systems: KAM tori

We consider a Hamiltonian system whose Hamiltonian can be written in suitable symplectic coordinates $(a,\varphi) \in \mathbb{R}^d \times \mathbb{T}^d$ (with $\mathbb{T}^d = \mathbb{R}^d/2\pi\mathbb{Z}^d$ the standard d-dimensional torus) as

$$(3.11) \qquad H(a,\varphi) = \omega^{\mathrm{T}} a + a^{\mathrm{T}} M(\varphi) a + O(\|a\|^3) \qquad \text{for } a \text{ near } 0,$$

with $\omega \in \mathbb{R}^d$ and $M(\varphi) \in \mathbb{R}^{d\times d}$. The system thus has the invariant torus $a = 0$. The celebrated Kolmogorov-Arnold-Moser (KAM) theory concerns the persistence of invariant tori under perturbations of the Hamiltonian. We will use the following version of the KAM theorem, cf. Thirring (1977), Ch. 3.6. Let the following conditions be satisfied:

(3.12a) The Hamiltonian (3.11) is real-analytic
in a complex neighbourhood U of $\{0\} \times \mathbb{T}^d$.

The vector of frequencies ω satisfies the non-resonance condition

$$(3.12b) \qquad |\omega^{\mathrm{T}} k| > \gamma \, \|k\|^{-\nu} \qquad \text{for all } \ k \in \mathbb{Z}^d, \quad k \neq 0$$

for some $\gamma > 0$ and $\nu > d - 1$. Further, there is a non-degeneracy condition:

$$(3.12c) \qquad \int_{\mathbb{T}^d} M(\varphi)\, d\varphi \quad \text{is an invertible matrix.}$$

Consider now a perturbed Hamiltonian

$$\widetilde{H}(a,\varphi) = H(a,\varphi) + \varepsilon G(a,\varphi) \ ,$$

where G is real-analytic and bounded by unity on U, and ε is a small scaling factor. Then, there exists $\varepsilon_0 > 0$, such that for every ε with $|\varepsilon| \le \varepsilon_0$ there is a real-analytic symplectic transformation of coordinates between complex neighbourhoods of $\{0\} \times \mathbb{T}^d$, such that in the new coordinates the perturbed Hamiltonian has again the form (3.11), with the same ω. The coordinate transform tends to the identity mapping as $\varepsilon \to 0$. (The proof in Thirring (1977) provides a bound of the difference with a power of ε that is inverse proportional to the dimension

d. Since the coordinate transform depends analytically on ε, one concludes that the difference to the identity mapping is in fact $O(\varepsilon)$, where the constant depends on ε_0.) The perturbed system therefore has an invariant torus near that of the original system. The threshold ε_0 and the domain of the coordinate transform depend on the quantities in (3.12) and on the dimension d, but can be chosen independently of G.

In our context, this applies to the Hamiltonian (3.10) of the perturbed system associated with a symplectic numerical integrator. The following result shows in particular that the invariant torus of the perturbed system is over exponentially long times nearly invariant under the numerical method.

Corollary 7. *Consider a Hamiltonian system satisfying the KAM conditions* (3.12)*, and the discretization of its equations of motion* (3.9) *by a* symplectic *numerical method. For sufficiently small step size h, there exists a perturbed Hamiltonian system* (3.10) *possessing an invariant torus on which the flow has frequencies ω, such that the following is valid: Over a time interval of length*

$$T = h\,\mathrm{e}^{h^*/4h} ,$$

there is the error bound

$$y_n - \widetilde{y}(nh) = O(\mathrm{e}^{-h^*/2h}) \qquad \text{for } \; nh \le T$$

between an arbitrary numerical solution y_n starting on the invariant torus of the perturbed system, and the solution $\widetilde{y}(t)$ of the perturbed system with the same starting value.

The constant h^* is given as the minimum over a neighbourhood of the torus of the numbers h_0 of Theorem 1.

Proof. (a) Lemma 11 below assures that the perturbed system with $N = N(h)$ of Theorem 1 fits into the KAM framework described above, with $\varepsilon = O(h^p)$. The KAM theorem then yields that the perturbed Hamiltonian (3.10) can be written in the transformed coordinates (b, ψ) as

$$\widetilde{H} = \omega^{\mathrm{T}} b + S(b, \psi)$$

with $S(b, \psi) = O(\|b\|^2)$. The perturbed system thus has the invariant torus $b = 0$. The transformation from the original coordinates (p, q) of (3.9) to (b, ψ) is bounded uniformly in h. The equations of motion read in these coordinates

$$b' = u(b, \psi) \; , \qquad \psi' = \omega + v(b, \psi) \tag{3.13}$$

where $u(b, \psi) = O(\|b\|^2)$ and $v(b, \psi) = O(\|b\|)$, and similarly for the derivatives $\partial u/\partial b = O(\|b\|)$, $\partial u/\partial \psi = O(\|b\|^2)$, and $\partial v/\partial b = O(1)$, $\partial v/\partial \psi = O(\|b\|)$. The constants in these O-terms are independent of the step size h. Let $(b(t), \psi(t))$ and $(\widehat{b}(t), \widehat{\psi}(t))$ be two solutions of (3.13) such that $\|b(t)\| \le \beta$, $\|\widehat{b}(t)\| \le \beta$ (β sufficiently small) for all t under consideration. Then, a simple argument based on Gronwall's lemma shows that their difference is bounded over a time interval $0 \le t \le 1/\beta$ by

$$\begin{aligned} \|b(t) - \widehat{b}(t)\| &\le C \left(\|b(0) - \widehat{b}(0)\| + \beta \, \|\psi(0) - \widehat{\psi}(0)\|\right) \\ \|\psi(t) - \widehat{\psi}(t)\| &\le C \left(t \, \|b(0) - \widehat{b}(0)\| + \|\psi(0) - \widehat{\psi}(0)\|\right) \ , \end{aligned} \tag{3.14}$$

for some constant C that does not depend on β or h.

(b) With the previous estimate, the result is obtained with Lady Windermere's fan similarly to the proof of Corollary 2. With the notation used there, we have

$$\|y_k - \widetilde{y}(t_k, t_{k-1}, y_{k-1})\| \le \delta := \text{Const.}\, \mathrm{e}^{-h^*/h}$$

by Theorem 1. We further denote the b-coordinates of y_n, $\widetilde{y}(t)$ and $\widetilde{y}(t, t_k, y_k)$ by b_n, $\widetilde{b}(t)$ and $\widetilde{b}(t, t_k, y_k)$, respectively. In order to be able to apply the error propagation estimate (3.14), we assume that

$$\|\widetilde{b}(t, t_k, y_k)\| \le \beta \qquad \text{for} \ \ t_k \le t \le 1/\beta \ , \tag{3.15}$$

and for all k satisfying $t_k = kh \le 1/\beta$. This assumption will be justified later and the value of β will be specified in Eq. (3.17) below. By (3.14) we thus obtain the bound

$$\|\widetilde{y}(t, t_k, y_k) - \widetilde{y}(t, t_{k-1}, y_{k-1})\| \le C\, (1 + (t - t_k))\, \delta \qquad \text{for} \ \ t_k \le t \le 1/\beta \ .$$

Summing up from $k = 1$ to $k = n$ gives for $t_n \le t \le 1/\beta$

$$\begin{aligned} \|\widetilde{y}(t, t_n, y_n) - \widetilde{y}(t)\| \le \sum_{k=1}^{n} C \left(1 + (t - t_k)\right) \delta &\le Ch^{-1} \delta \, (t_n + tt_n - t_n^2/2) \\ &< Ch^{-1} \delta \, t^2 \le Ch^{-1} \delta/\beta^2 \end{aligned} \tag{3.16}$$

(these estimates hold at least for $t > 2$). We now set

$$\beta = (2Ch^{-1}\delta)^{1/3} \ , \tag{3.17}$$

so that $Ch^{-1}\delta/\beta^2 = \beta/2$, and we obtain the desired estimate from (3.16) by putting $t = t_n$.

(c) We still have to justify the assumption (3.15). This will be done by induction. For $k = 0$ nothing has to be shown, because $\widetilde{b}(t, t_0, y_0) = \widetilde{b}(t) \equiv 0$ as a consequence of the fact that $\widetilde{y}(t)$ stays on the torus $b = 0$. Suppose now that (3.15) holds for $k \le n$. It then follows from (3.16) that

$$\|\widetilde{b}(t, t_n, y_n)\| < Ch^{-1}\delta/\beta^2 = \beta/2 \qquad \text{for} \ \ t_n \le t \le 1/\beta$$

(again because of $\widetilde{b}(t) \equiv 0$). Consequently we also have

$$\|b_{n+1}\| \le \|b_{n+1} - \widetilde{b}(t_{n+1}, t_n, y_n)\| + \|\widetilde{b}(t_{n+1}, t_n, y_n)\| < \delta + \beta/2 \le \beta,$$

provided that h is sufficiently small so that $\delta \le \beta/2$. By continuity, $\widetilde{b}(t, t_{n+1}, y_{n+1})$ is bounded by β on a non-empty interval $[t_{n+1}, T_{n+1}]$. The computation of part (b) shows that $\|\widetilde{b}(t, t_{n+1}, y_{n+1})\| \le \beta/2$ on this interval. Hence, T_{n+1} can be increased until $T_{n+1} \ge 1/\beta$. This proves the estimate (3.15) for $k = n + 1$. □

Using the preceding arguments together with a normal form of the Hamiltonian derived by Perry and Wiggins (1994), we can study also the behaviour of numerical solutions which start a distance ρ away from an invariant torus of the perturbed Hamiltonian.

Corollary 8. *In the situation of Corollary 7, the following is additionally valid: Over a time interval of length*

$$T = \min\left(C\rho^{-1}\mathrm{e}^{(\rho^*/\rho)^\alpha}, h\,\mathrm{e}^{h^*/4h} \right) \ ,$$

there is the error bound

$$y_n - \widetilde{y}(nh) = O(\mathrm{e}^{-h^*/2h}) \qquad \textit{for} \;\; nh \le T$$

between an arbitrary numerical solution y_n starting at a sufficiently small distance ρ from the invariant torus $b = 0$ of the perturbed system, and the solution $\widetilde{y}(t)$ of the perturbed system with the same starting value.

This holds with a constant ρ^* that is independent of h and ρ, and with $\alpha = 1/(d+2)$, where d is the dimension of the system. The constant h^* is given as the minimum over a neighbourhood of the torus of the numbers h_0 of Theorem 1.

Remark. It is known from Perry and Wiggins (1994) that $\widetilde{y}(t)$ remains $O(\rho)$-close to the invariant torus for $t \le T$.

Proof. (a) According to Theorem 5 of Perry and Wiggins (1994) there is, for every $r > 0$ (sufficiently small), a real-analytic symplectic transformation of coordinates $(b, \psi) \mapsto (c, \theta)$ between complex neighbourhoods of $\{0\} \times \mathbb{T}^d$, such that in the new coordinates the Hamiltonian $\widetilde{H}$ takes the form

$$\widetilde{H} = \omega^{\mathrm{T}} c + Z(c) + R(c, \theta) \ , \qquad \|c\| \le r \ , \tag{3.18}$$

where

$$\|Z(c)\| \le K\, r^2 \ , \qquad \|R(c,\theta)\| \le K\, r^2\, \mathrm{e}^{-(k/r)^\alpha} \tag{3.19}$$

for (c, θ) in the complex neighbourhood, with some constants K and k. The explicit bounds in Perry and Wiggins (1994) show that the neighbourhood can be chosen independent of h, and the coordinate transform as well as K and k are bounded independently of h and r. Let now $(c(t), \theta(t))$ and $(\widehat{c}(t), \widehat{\theta}(t))$ be two solutions of the equations of motion for the Hamiltonian $\widetilde{H}$ such that $\|c(t)\| \le r$, $\|\widehat{c}(t)\| \le r$ for all t under consideration. With (3.19), an argument based on Gronwall's lemma then shows that the difference of the two solutions is bounded for $t \le \mathrm{Const.}\, r^{-1}\mathrm{e}^{(k/r)^\alpha}$ (hence, much longer than (3.14)!) by

$$\begin{aligned} \|c(t) - \widehat{c}(t)\| &\le C\left(\|c(0) - \widehat{c}(0)\| + r\,\mathrm{e}^{-(k/r)^\alpha}\, \|\theta(0) - \widehat{\theta}(0)\| \right) \\ \|\theta(t) - \widehat{\theta}(t)\| &\le C\left(t\, \|c(0) - \widehat{c}(0)\| + \|\theta(0) - \widehat{\theta}(0)\| \right) \ , \end{aligned} \tag{3.20}$$

for some constant C that does not depend on r or h.

(b) The result now follows by a slight variation of parts (b) and (c) of the preceding proof, using (3.20) instead of (3.14). □

4. Proof of Theorem 1

Theorem 1 is a direct consequence of the following bound: If $hN \le R/(2\mathrm{e}^2\nu M)$, then the difference between the numerical result y_1 and the solution $\widetilde{y}(t,h)$ of the perturbed differential equation (2.2) is bounded by

$$\|y_1 - \widetilde{y}(h,h)\| \le 0.012\,R\,\mathrm{e}^{-(N+1)} + 0.02\,R\,\mathrm{e}^{-(N+1)}\ . \tag{4.1}$$

Choosing N as large as permissible, i.e., as the largest integer such that $hN \le R/(2\mathrm{e}^2\nu M)$, then gives the bound of Theorem 1. The error bound (4.1) will be shown as follows: By analyticity, we have the convergent Taylor series expansions

$$\begin{aligned} y_1 = y_1(h) &= \sum_{i=0}^{\infty} \frac{h^i}{i!}\,\frac{d^i y_1}{dh^i}(0) \qquad \text{and} \\ \tilde{y}(h,h) &= \sum_{i=0}^{\infty} \frac{h^i}{i!} \sum_{k=0}^{i} \binom{i}{k} \frac{\partial^i \widetilde{y}(0,0)}{\partial t^k \partial h^{i-k}}\ . \end{aligned} \tag{4.2}$$

The construction of (2.2) yields that the two series coincide up to the Nth term. The remainder terms will be bounded separately. Lemma 12 yields an upper bound for the remainder of the lower series, and Lemma 9 shows that the remainder of the upper series is bounded by (for $N \ge 2$)

$$2\mu M h \sum_{\rho \ge N+1} \left(\frac{4\kappa M h}{R}\right)^{\rho-1} \le \frac{(2\log 2 - 1)R}{N(\mathrm{e}^2-1)}\,\mathrm{e}^{-2N} \le 0.012\,R\,\mathrm{e}^{-(N+1)}.$$

Here we have used the estimates $hN \le R/(2\mathrm{e}^2\nu M)$, $\kappa \le \nu$, and $\mu \le (2\log 2 - 1)\nu$ (which follow from (2.4)). The bounds of Lemmas 9 and 12 also show that the series in (4.2) are absolutely convergent for $hN \le R/(2\mathrm{e}^2\nu M)$.

4.1. The analyticity assumption

We assume that f is analytic in the polydisc of radius R around y_0, so that Cauchy's integral formula can be applied:

$$f(x_1,\ldots,x_d) = \frac{1}{(2\pi i)^d} \int_{\gamma_1} \cdots \int_{\gamma_d} \frac{f(z_1,\ldots,z_d)}{(z_1-x_1)\cdot\ldots\cdot(z_d-x_d)}\,dz_1\ldots dz_d\ .$$

Here we write $x = (x_1,\ldots,x_d)$ instead of y_0, and γ_j denotes the circle of radius R around x_j. Differentiation yields the estimate

$$\left\| \frac{\partial^{k_1+\ldots+k_d} f(x)}{\partial x_1^{k_1}\ldots\partial x_d^{k_d}} \right\| \le k_1!\cdot\ldots\cdot k_d!\cdot M\cdot R^{-(k_1+\ldots+k_d)},$$

where M is an upper bound of f on the polydisc. The kth derivative, considered as a k-linear mapping, is defined by

$$f^{(k)}(x)(u,v,\ldots) = \sum_{i,j,\ldots} \frac{\partial^k f(x)}{\partial x_i \partial x_j \ldots} \cdot u_i \cdot v_j \cdot \ldots ,$$

and we obtain from $k_1! \cdot \ldots \cdot k_d! \le (k_1 + \ldots + k_d)!$ that

$$\|f^{(k)}(x)(u,v,\ldots)\| \le k!\,M\,R^{-k} \sum_{i,j,\ldots} |u_i| \cdot |v_j| \cdot \ldots \le k!\,M\,R^{-k} \cdot \|u\|_1 \cdot \|v\|_1 \cdot \ldots .$$

For the ℓ_1 operator norm we thus obtain (we write again y_0 in place of x)

$$\|f^{(k)}(y_0)\| \le k!\,M\,R^{-k} , \qquad k \ge 0 , \tag{4.3}$$

which will be the fundamental assumption for the following estimates.

4.2. Estimation of the derivatives of the numerical solution

We need to recall some notation (for more details see Butcher (1987) or Hairer, Nørsett and Wanner (1993)). The numerical solution y_1 is given as a B-series

$$y_1 = y_0 + \sum_{\tau \in T} \frac{h^{\rho(\tau)}}{\rho(\tau)!}\, \alpha(\tau)\, a(\tau)\, F(\tau)(y_0) . \tag{4.4}$$

Here, T is the set of rooted trees, $\rho(\tau)$ is the number of vertices of a tree $\tau \in T$, and the integer $\alpha(\tau)$ counts the number of monotonic labellings of the tree τ. The set of all trees with exactly ρ vertices will be denoted by T_ρ. In (4.4), $F(\tau)(y)$ is the elementary differential of $f(y)$ associated with the tree τ. This is recursively defined by

$$F(\tau)(y) = f(y) \qquad \text{for } \tau = \bullet$$

$$F(\tau)(y) = f^{(m)}(y) \cdot \big(F(\tau_1)(y), \ldots, F(\tau_m)(y)\big) \qquad \text{for } \tau = [\tau_1, \ldots, \tau_m]$$

where $\bullet$ is the tree with only one vertex, and $[\tau_1, \ldots, \tau_m]$ is the tree which decomposes into the trees $\tau_1, \ldots, \tau_m$ when the root is chopped off.

The real number $a(\tau)$ depends on the coefficients of the integration method. With κ and μ of (2.4), it satisfies

$$|a(\tau)| \le \gamma(\tau)\, \mu\, \kappa^{\rho(\tau)-1} , \tag{4.5}$$

where the integer $\gamma(\tau)$ is bounded by $\rho(\tau)!$, and is such that $a(\tau) = \gamma(\tau)$ for the implicit Euler method. There exist recursive definitions of both $a(\tau)$ and $\gamma(\tau)$, which will however not be used below.

Lemma 9. *Under conditions (4.3) and (4.5), the ρth derivative of the numerical solution is bounded by*

$$\frac{1}{\rho!} \left\| \frac{d^\rho y_1}{dh^\rho}(0) \right\| \le 2\mu M \left(\frac{4\kappa M}{R} \right)^{\rho-1} . \tag{4.6}$$

Proof. By (4.4) we have

$$\left\| \frac{d^\rho y_1}{dh^\rho}(0) \right\| \leq \sum_{\tau \in T_\rho} \alpha(\tau)\, |a(\tau)|\, \|F(\tau)(y_0)\| \ .$$

For $\kappa = 0$ (i.e., for the explicit Euler method) the estimates are obvious, if one interprets $\kappa^0 = 1$. For $\kappa > 0$ we may assume without loss of generality that $\kappa = \mu = 1$. This means that it is sufficient to prove the result for the implicit Euler method, for which $a(\tau) = \gamma(\tau)$ for all trees τ.

The main idea is to consider the scalar differential equation

$$z' = g(z) \qquad \text{with} \qquad g(z) = \frac{M}{1 - z/R}, \tag{4.7}$$

and initial value $z(0) = 0$. For the derivatives $g^{(k)}(0)$ of this function we have equality in (4.3) for all $k \geq 0$. Consequently,

$$\sum_{\tau \in T_\rho} \alpha(\tau)\,\gamma(\tau)\,\|F(\tau)(y_0)\| \leq \sum_{\tau \in T_\rho} \alpha(\tau)\,\gamma(\tau)\,G(\tau)(0) = \frac{d^\rho z_1}{dh^\rho}(0)\ ,$$

where $G(\tau)(z_0)$ are the elementary differentials corresponding to (4.7) and

$$z_1 = \frac{hM}{1 - z_1/R}\ , \qquad \text{or equivalently,} \qquad z_1 = \frac{R}{2}\left(1 - \sqrt{1 - \frac{4Mh}{R}}\ \right) \tag{4.8}$$

is the numerical solution of the implicit Euler method applied to (4.7). From the Taylor series expansion

$$\sqrt{1-x} = \sum_{\rho \geq 0} \binom{1/2}{\rho} (-x)^\rho = 1 - \sum_{\rho \geq 1} \frac{1}{(2\rho - 1)} \binom{2\rho}{\rho} \left(\frac{x}{4}\right)^\rho, \tag{4.9}$$

applied to (4.8), we see that

$$\frac{1}{\rho!}\frac{d^\rho z_1}{dh^\rho}(0) = \frac{R}{2(2\rho-1)} \binom{2\rho}{\rho} \left(\frac{M}{R}\right)^\rho \leq 2M \left(\frac{4M}{R}\right)^{\rho - 1}. \tag{4.10}$$

The second inequality of (4.10) follows from the fact that $|\binom{1/2}{\rho}| \leq 1$. A slightly better estimate could be obtained with Stirling's formula. □

Remark. If, in the proof of Lemma 9, we apply the implicit Euler method to the scalar problem (4.7) with arbitrary initial value z_0, we get estimates for derivatives of y_1 with respect to both h and y_0, in particular,

$$\frac{1}{\rho!}\left\| \frac{\partial^\rho}{\partial h^\rho}\left(\frac{\partial y_1}{\partial y_0}\right)(0, y_0) \right\| \leq \frac{\mu}{2}\binom{2\rho}{\rho}\left(\frac{\kappa M}{R}\right)^\rho \leq \frac{\mu}{2}\left(\frac{4\kappa M}{R}\right)^\rho.$$

Together with the estimates of the proof of Lemma 11 below, this allows us to prove that besides the difference $\|y_1 - \widetilde{y}(h)\|$ also the difference of the derivatives with respect to the initial value, $\|\partial y_1/\partial y_0(h) - \partial \widetilde{y}/\partial y_0(h)\|$, is exponentially small.

4.3. Estimation of coefficients of the perturbed differential equation

In Hairer (1994), the expansion terms f_ρ of (2.2) were constructed as

$$f_\rho(y) = \frac{1}{\rho!} \sum_{\tau \in T_\rho} \alpha(\tau)\, b(\tau)\, F(\tau)(y)\ . \tag{4.11}$$

Here, the coefficients $b(\tau)$ are recursively defined by

$$a(\tau) = \sum_{k=1}^{\rho(\tau)} \frac{1}{k!} \sum_{(\tau,S)} \binom{\rho(\tau)}{\rho(\sigma_1), \ldots, \rho(\sigma_k)} \frac{\alpha(\tau,S)}{\alpha(\tau)}\, b(\sigma_1) \cdot \ldots \cdot b(\sigma_k). \tag{4.12}$$

The second sum in (4.12) is over all partitions S of τ into k subtrees $\sigma_1, \ldots, \sigma_k$. The integer $\alpha(\tau,S)$, associated with a partition S of τ, counts the number of possible monotonic labellings of τ such that the vertices of the subtrees in the partition are labelled consecutively.

Lemma 10. *Suppose that*

$$|a(\tau)| \leq \mu \cdot \kappa^{\rho(\tau)-1} \cdot \rho(\tau)! \qquad \textit{for all trees } \tau. \tag{4.13}$$

Then the coeffcients $b(\tau)$, defined in Eq. (4.12), can be estimated by

$$|b(\tau)| \leq \log 2 \cdot \nu^{\rho(\tau)} \cdot \rho(\tau)! \qquad \textit{with} \quad \nu = \kappa + \mu/(2\log 2 - 1). \tag{4.14}$$

Proof. In the sum of (4.12) the term for $k = 1$ is equal to $b(\tau)$. We extract this term from the formula, apply the triangle inequality, divide by $\rho!$, and thus obtain

$$\frac{|b(\tau)|}{\rho!} \leq \frac{|a(\tau)|}{\rho!} + \sum_{k=2}^{\rho} \frac{1}{k!} \sum_{(\tau,S)} \frac{\alpha(\tau,S)}{\alpha(\tau)} \cdot \frac{|b(\sigma_1)|}{\rho_1!} \cdot \ldots \cdot \frac{|b(\sigma_k)|}{\rho_k!},$$

where $\rho = \rho(\tau)$ and $\rho_i = \rho(\sigma_i)$. We now fix $k \geq 2$, a k-tupel $(\rho_1, \ldots, \rho_k)$ satisfying $\rho_1 + \ldots + \rho_k = \rho$ and a labelling of the tree τ. If the sets of vertices $\{1, \ldots, \rho_1\}$, $\{\rho_1 + 1, \ldots, \rho_1 + \rho_2\}$, $\{\rho_1 + \rho_2 + 1, \ldots, \rho_1 + \rho_2 + \rho_3\}$, $\ldots$ are all connected in the tree τ, then we are led to a "labelled" partition of τ and every labelled partition can be obtained in this way. Therefore it holds

$$|b(\tau)|/\rho! \leq d_\rho, \tag{4.15}$$

where the numbers d_ρ are recursively defined by

$$d_\rho = \mu\kappa^{\rho-1} + \sum_{k=2}^{\rho} \frac{1}{k!} \sum_{\rho_1+\ldots+\rho_k=\rho} d_{\rho_1} \cdot \ldots \cdot d_{\rho_k} \tag{4.16}$$

(the second sum is over k-tupels $(\rho_1, \ldots, \rho_k)$ with $\rho_i \geq 1$). Multiplying (4.16) by ζ^ρ and summing up for $\rho \geq 1$, we get for the generating function

$$d(\zeta) = d_1\zeta + d_2\zeta^2 + d_3\zeta^3 + \ldots$$

the relation

$$d = \frac{\mu\zeta}{1-\kappa\zeta} + \mathrm{e}^d - 1 - d\,. \tag{4.17}$$

Whenever $\mathrm{e}^d \neq 2$ (i.e., for $\zeta \neq (2d-1)/\bigl(\mu + \kappa(2d-1)\bigr)$ with $d = \log 2 + 2k\pi i$) the implicit function theorem can be applied. This implies that $d(\zeta)$ is analytic in a disc with radius $1/\nu = (2\log 2 - 1)/\bigl(\mu + \kappa(2\log 2 - 1)\bigr)$ and center at the origin. On the disc $|\zeta| \le 1/\nu$, the solution $d(\zeta)$ of (4.17) with $d(0) = 0$ is bounded by $\log 2$. This is seen as follows: The complex function $\mathrm{e}^d - 2d - 1$ maps the disc $|d| < \log 2$ conformally onto a domain that contains the disc $|w| < 2\log 2 - 1$, which is the image of the disc $|\zeta| < 1/\nu$ under the mapping $\mu\zeta/(1-\kappa\zeta)$ appearing in (4.17). The preimage of this latter disc is thus contained in the disc $|d| < \log 2$, and by continuity $|\zeta| \le 1/\nu$ therefore implies $|d(\zeta)| \le \log 2$. The estimate (4.14) is thus a consequence of Cauchy's inequalities for the coefficients d_ρ, and of (4.15). □

4.4. Estimation of the expansion functions of the perturbed differential equation

Using the analyticity assumption on f and the estimates of Lemma 10, we are in the position to bound the function f_ρ of (4.11) and its derivatives.

Lemma 11. *If f satisfies (4.3), then we have for $\rho \ge 2$ and $k \ge 0$ the estimate*

$$\frac{1}{k!}\,\|f_\rho^{(k)}(y_0)\| \le \frac{0.2\,R}{\rho}\Bigl(\frac{2\mathrm{e}\nu M\rho}{R}\Bigr)^{\rho}\Bigl(\frac{\mathrm{e}}{R}\Bigr)^{k}. \tag{4.18}$$

Proof. The important observation is that

$$\sum_{\tau\in T_\rho}\alpha(\tau)\,\Bigl\|\frac{d^k}{dy^k}\,F(\tau)(y_0)\Bigr\| \le \sum_{\tau\in T_\rho}\alpha(\tau)\,\frac{d^k}{dz^k}\,G(\tau)(0) = \frac{\partial^{\rho+k}z}{\partial t^\rho \partial z_0^k}\,(0,0) \tag{4.19}$$

where, as in the poof of Lemma 9, $G(\tau)(z)$ are the elementary differentials corresponding to (4.7), and $z(t,z_0)$ is the solution of (4.7) with initial value z_0 for $t = 0$. This solution is given by

$$z(t,z_0) = R\left(1 - \sqrt{\Bigl(1-\frac{z_0}{R}\Bigr)^2 - \frac{2Mt}{R}}\;\right)$$

and has the Taylor series expansion

$$z(t,z_0) = z_0 + \sum_{\rho\ge 1}\frac{R}{(2\rho-1)}\binom{2\rho}{\rho}\Bigl(\frac{Mt}{2R}\Bigr)^{\rho}\sum_{k\ge 0}\binom{2\rho+k-2}{k}\Bigl(\frac{z_0}{R}\Bigr)^{k}$$

so that the desired derivative is

$$\frac{\partial^{\rho+k} z}{\partial t^{\rho}\partial z_0^k}(0,0) = \frac{\rho!\,k!\,R}{(2\rho-1)}\binom{2\rho}{\rho}\binom{2\rho+k-2}{k}\Big(\frac{M}{2R}\Big)^{\rho}\Big(\frac{1}{R}\Big)^{k}.$$

Using Stirling's formula and the well-known inequality $(1+x/n)^n \le \mathrm{e}^x$ (for $x \ge 0$), which implies $(2\rho+k)^{2\rho} \le (2\rho)^{2\rho}\mathrm{e}^k$ and $(2\rho+k)^k \le k^k\mathrm{e}^{2\rho}$, this unhandy expression can be estimated to give

$$\frac{1}{k!}\,\frac{\partial^{\rho+k} z}{\partial t^{\rho}\partial z_0^k}(0,0) \le \frac{0.287\,R}{\rho}\Big(\frac{2\mathrm{e}M\rho}{R}\Big)^{\rho}\Big(\frac{\mathrm{e}}{R}\Big)^{k}.$$

Together with (4.11), (4.14) and (4.19) this yields the result. □

4.5. Choice of N and estimation of solution derivatives of the perturbed equation

In order to estimate the derivatives of the solution of the perturbed differential equation (2.2) we have to fix N. The estimate (4.18) with $k=0$ shows that the ρth term of the truncated series (2.2b) is bounded by $Const\cdot(\varepsilon\rho)^{\rho}$ with $\varepsilon = 2\mathrm{e}\nu Mh/R$. This bound is a convex function of ρ, having its minimum at $\rho = (\varepsilon\mathrm{e})^{-1}$. For a fixed h, it is therefore natural to choose N close to the value $(\varepsilon\mathrm{e})^{-1}$. Hence, for the following we assume

$$hN \le \frac{R}{2\mathrm{e}^2\nu M}\,. \tag{4.20}$$

We are interested in estimating the expression (see Eq. (4.2))

$$E_N := \sum_{i\ge N+1}\frac{h^i}{i!}\sum_{k=0}^{i}\binom{i}{k}\frac{\partial^i\widetilde{y}(0,0)}{\partial t^k\partial h^{i-k}}\,. \tag{4.21}$$

Lemma 12. *Under condition (4.20) the remainder term (4.21) is bounded by*

$$\|E_N\| \le 0.02\,R\,\mathrm{e}^{-(N+1)}. \tag{4.22}$$

Proof. (a) We use

$$\frac{1}{k!}\,\|f_{\rho}^{(k)}(y_0)\| \le 0.1\,R\Big(\frac{2\mathrm{e}\nu M\rho}{R}\Big)^{\rho}\Big(\frac{\mathrm{e}}{R}\Big)^{k}, \tag{4.23}$$

which follows for $\rho \ge 2$ from (4.18) and for $\rho = 1$ (we use the notation $f_1(y) := f(y)$) from (4.3) provided that $0.2\mathrm{e}\nu \ge 1$. This condition is satisfied for methods of order ≥ 1 (for which $a(\bullet) = 1$), because by (4.5) $\mu \ge 1$ and by (4.14) $\nu \ge 1/(2\log 2 - 1) \ge 2.5$.

We now consider the scalar differential equation (with parameter h)

$$z' = g_1(z) + hg_2(z) + \ldots + h^{N-1}g_N(z),$$

where

$$g_\rho(z) = \frac{0.1 \cdot R \cdot (B\rho)^\rho}{1 - \mathrm{e}z/R} \qquad \text{with} \qquad B = 2\mathrm{e}\nu M/R.$$

For the functions $g_\rho(z)$ we have equality in (4.23) for all ρ and k. Expressing the derivatives of $\widetilde{y}(t,h)$ and of $z(t,h)$ in terms of the derivatives of $f_\rho(y)$ and $g_\rho(z)$, respectively, we see that

$$\left\| \frac{\partial^i \widetilde{y}(0,0)}{\partial t^k \partial h^{i-k}} \right\| \le \frac{\partial^i z(0,0)}{\partial t^k \partial h^{i-k}} \qquad \text{for all} \quad i,k \ge 0\ , \tag{4.24}$$

and hence

$$\|E_N\| \le \sum_{i \ge N+1} \frac{h^i}{i!} \sum_{k=0}^{i} \binom{i}{k} \frac{\partial^i z(0,0)}{\partial t^k \partial h^{i-k}}\ . \tag{4.25}$$

The differential equation for z can be solved analytically and we have for the initial value $z_0 = 0$ that

$$z(t,h) = \frac{R}{\mathrm{e}} \Bigl(1 - \sqrt{1 - 0.2\,\mathrm{e}\,t\,h^{-1} \textstyle\sum_{\rho=1}^{N} (B\rho h)^\rho}\Bigr).$$

From the Taylor series expansion (4.9) we thus get

$$z(h,h) = \frac{R}{\mathrm{e}} \sum_{j\ge1} \frac{1}{(2j-1)} \binom{2j}{j} \Bigl(\frac{0.2\,\mathrm{e}}{4}\Bigr)^j \Bigl(\sum_{\rho=1}^{N} (B\rho h)^\rho\Bigr)^j.$$

Removing all terms which have $h^0, h^1, \dots, h^N$ as factor, we obtain from (4.25) that

$$\|E_N\| \le \frac{R}{\mathrm{e}} \sum_{j\ge2} \frac{1}{(2j-1)} \binom{2j}{j} \Bigl(\frac{0.2\,\mathrm{e}}{4}\Bigr)^j \Bigl(\sum_{\rho \ge (N+1)/j}^{N} (B\rho h)^\rho\Bigr)^j. \tag{4.26}$$

We shall prove below (in part (b)) that for $k \ge 1$ and for $N \ge k$

$$\Bigl(\sum_{\rho=k}^{N} \Bigl(\frac{\rho}{N\mathrm{e}}\Bigr)^\rho\Bigr)^{\frac{N+1}{k}} \le \mathrm{e}^{-(N+1)}. \tag{4.27}$$

If for a given $j \ge 2$ we denote by k the smallest integer satisfying $k \ge (N+1)/j$, then we have from (4.27), from the definition of B, and from (4.20) that

$$\Bigl(\sum_{\rho \ge (N+1)/j}^{N} (B\rho h)^\rho\Bigr)^j \le \Bigl(\sum_{\rho=k}^{N} \Bigl(\frac{\rho}{N\mathrm{e}}\Bigr)^\rho\Bigr)^{\frac{N+1}{k}} \le \mathrm{e}^{-(N+1)}.$$

Inserted into (4.26) we get from (4.9) the estimate

$$\begin{aligned} \|E_N\| &\le \frac{R}{\mathrm{e}} \Bigl(\sum_{j\ge2} \frac{1}{(2j-1)} \binom{2j}{j} \Bigl(\frac{0.2\mathrm{e}}{4}\Bigr)^j\Bigr) \mathrm{e}^{-(N+1)} \\ &\le \frac{R}{\mathrm{e}} \Bigl(1 - \sqrt{1 - 0.2\,\mathrm{e}} - 0.1\,\mathrm{e}\Bigr) \mathrm{e}^{-(N+1)}, \end{aligned}$$

which completes the proof.

(b) In order to prove the inequality (4.27), we first remark that x^x is decreasing on the interval $[0, \mathrm{e}^{-1}]$. Consequently, the term for $\rho = k$ is the largest one in the sum of (4.27) and we have

$$\left(\sum_{\rho=k}^{N}\left(\frac{\rho}{N\mathrm{e}}\right)^{\rho}\right)^{\frac{N+1}{k}} \le \left((N-k+1)\left(\frac{k}{N\mathrm{e}}\right)^{k}\right)^{\frac{N+1}{k}}.$$

The estimate (4.27) thus follows from the fact that

$$(N-k+1)k^k \le N^k \qquad \text{for } \; k \ge 1 \; \text{ and for } \; N \ge k,$$

which, for fixed $k \ge 1$, can be proved by induction on N. □

Acknowledgement. We are grateful to D. Stoffer for making us aware of the article by Benettin and Giorgilli (1994) during the preparation of this paper, and to St. Tompaidis for helpful discussions about KAM theory. Our special thanks go to Claudia Wulff for pointing out an error in a previous version of this article.

References

Benettin, G., Giorgilli, A. (1994): On the Hamiltonian interpolation of near to the identity symplectic mappings with application to symplectic integration algorithms. J. Statist. Phys. **74**, 1117–1143

Beyn, W.-J. (1987a): On the numerical approximation of phase portraits near stationary points. SIAM J. Numer. Anal. **24**, 1095–1113

Beyn, W.-J. (1987b): On invariant closed curves for one-step methods. Numer. Math. **51**, 103–122

Braun, M., Hershenov, J. (1977): Periodic solutions of finite difference equations. Quart. Appl. Maths. **35**, 139–147

Butcher, J.C. (1987): The Numerical Analysis of Ordinary Differential Equations. Wiley, New York

Eirola, T. (1993): Aspects of backward error analysis of numerical ODE's. J. Comp. Appl. Math. **45**, 65–73

Feng Kang (1991): Formal power series and numerical algorithms for dynamical systems. Proceedings of international conference on scientific computation, Hangzhou, China, Eds. Tony Chan and Zhong-Ci Shi, Series on Appl. Math. **1**, 28–35

Fiedler, B., Scheurle, J. (1996): Discretization of homoclinic orbits, rapid forcing and "invisible" chaos. Mem. Amer. Math. Soc. **119**, no. 570

Garay, B.M. (1993): Discretization and some qualitative properties of ordinary differential equations about equilibria. Acta Math. Univ. Comenian.. (N.S.) **62**, 249–275

Hairer, E. (1994): Backward analysis of numerical integrators and symplectic methods. Annals of Numerical Mathematics **1**, 107–132

Hairer, E., Nørsett, S.P., Wanner, G. (1993): Solving Ordinary Differential Equations I. Nonstiff Problems. Second Revised Edition. Springer Series in Computational Mathematics **8**, Springer-Verlag

Perry, A.D., Wiggins, S. (1994): KAM tori are very sticky: rigorous lower bounds on the time to move away from an invariant Lagrangian torus with linear flow. Physica D **71**, 102–121

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

Sanz-Serna, J.M., Larsson, S. (1993): Shadows, chaos and saddles. Appl. Numer. Math. **13**, 181–190

Stuart, A.M. (1994): Numerical analysis of dynamical systems. In: Iserles, A. (ed.) Acta Numerica 1994, 467–572. Cambridge University Press, Cambridge

Thirring, W. (1977): Lehrbuch der Mathematischen Physik 1. Klassische Dynamische Systeme. Springer-Verlag, Wien

Wiggins, S. (1990): Introduction to Applied Nonlinear Dynamical Systems and Chaos. Springer-Verlag, New York

Wilkinson, J.H. (1960): Error analysis of floating-point computation. Numer. Math. **2**, 319–340

Yoshida, H. (1993): Recent progress in the theory and application of symplectic integrators. Celestial Mechanics and Dynamical Astronomy **56**, 27–43

This article was processed by the author using the LaTeX style file *pljour1m* from Springer-Verlag.