References

[1] Arrowsmith D.K. and Place C. M., *An Introduction to Dynamical Systems*, Cambridge University Press (1990).

[2] Chillingworth D.R.J., *Differential Topology with a View to Applications*, Pitman: London (1976).

[3] Chow, S.-N. and K.J. Palmer, On the numerical computation of orbits of dynamical systems: the one-dimensional case, Preprint (1989).

[4] Constantin P., Foias C., Nicolaenko B., and Temam R., *Integral Manifolds and Inertial Manifolds for Dissipative Partial Differential Equations*, Springer-Verlag: Berlin (1989).

[5] Devaney R.L., *An Introduction to Chaotic Dynamical Systems*, Benjamin-Cummings Publ. Co. Inc. (1986).

[6] Eckmann J.-P. and Ruelle D., *Rev. Mod. Phys.* **57** (1985), 617.

[7] Hammel S.M., Yorke J.A., Grebogi C., *Bull. Am. Math. Soc. (NS)* **19** (1988), 465.

[8] Glendinning P.A. and Sparrow C.T., *J. Stat. Phys.* **35** (1984) 645.

[9] Grebogi C., Ott E. and Yorke J.A., *Physica D* **7** (1983), 181.

[10] Guckenheimer J. and Holmes P., *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, Springer-Verlag: Berlin (1983).

[11] Keller H.B., *Applications of Bifurcation Theory* (P. Rabinowitz, ed.), Academic Press: New York (1977), 359.

[12] Newhouse S.E., *Prog. in Math.* **8** (1980), 1.

[13] Ruelle D., *Annals of Mathematics* **115** (1982), 243.

[14] Sanz-Serna J.M., *Acta Numerica* **1**, (1992), to appear.

[15] Silnikov L.P., *Sov. Math. Dokl.* **6** (1965), 163.

[16] Temam R., *Infinite Dimensional Dynamical Systems*, Springer-Verlag: Berlin (1988).

[17] Wiggins S., *Global Bifurcations and Chaos: Analytic Methods*, Springer-Verlag: Berlin (1988).

Numerical Ordinary Differential Equations vs. Dynamical Systems

J.M. Sanz-Serna

Universidad de Valladolid, Spain

Abstract. In this expository paper we are concerned with the following question: A given system of ordinary differential equations S is integrated by means of a given numerical method M. To what extent is the dynamics of the approximate solutions generated by M a faithful description of the dynamics of S?

1 Introduction

The present paper is devoted to the study of the relations between two mathematical fields: time-continuous dynamical systems and numerical methods for ordinary differential equations. In more concrete terms, we are concerned with the following question.

(Q): A given system of ordinary differential equations S is integrated by means of a given numerical method M. To what extent is the dynamics of the approximate solutions generated by M a faithful description of the dynamics of S?

The relevance of this question is clear: very often numerical simulations are used to discover the dynamics of systems of differential equations S and we would like to be sure that the behaviour of the numerical results corresponds to the behaviour of S, rather than being an artifact introduced by the discretization.

The paper is mostly expository. Little background on numerical methods or dynamical systems is assumed on the reader. It is therefore hoped that the article may be easily read both by numerical analysts and by experts on dynamical systems.

We begin with the presentation, in Section 2, of the numerical methods referred to later in the paper. In Sections 3 and 4 we provide a necessarily very sketchy review of the main developments on the analysis of numerical

ODE methods in the last thirty-five years. It turns out that most of the available results are not directly relevant in connection with question *(Q)*. In Section 5 we reconsider *(Q)* in the light of a concrete example. Sections 6–8 are devoted to answering *(Q)*. The final Section 9 briefly refers to aspects of the dynamical systems/numerical methods interface not covered elsewhere in the paper.

2 Numerical methods

2.1 Preliminaries

We consider initial-value problems of the form

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}, t), \quad t \in I, \quad \mathbf{y}(0) = \boldsymbol{\alpha} \in \mathcal{R}^d, \tag{2.1}$$

where I denotes either a compact interval $[0, T]$, or the half-line $[0, \infty)$. Often, (2.1) may be the result of the space-discretization of an initial boundary-value problem in partial differential equations. In such a case the dimension d is typically very high. The restriction to real unknowns is not essential: the methods described below are easily adapted to the complex case or, alternatively, complex d dimensional systems can be numerically integrated as real $2d$-dimensional systems.

All numerical methods for (2.1) generate approximations

$$\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_n, \ldots$$

to the 'true' values

$$\mathbf{y}(t_0), \mathbf{y}(t_1), \ldots, \mathbf{y}(t_n), \ldots,$$

where $0 = t_0 < t_1 < \cdots < t_n < \cdots$ is a *grid* in I. If I is bounded, the grid is assumed to have a finite number of *grid-points* t_n. If $I = [0, \infty)$ we suppose that n takes all positive integer values and $t_n \uparrow \infty$: while it is clearly not feasible to actually compute infinitely many $\mathbf{y}_n$, it is convenient to conceive of numerical integrations that cover arbitrarily long time-intervals. The increments $h_n := t_{n+1} - t_n$ are called the *step-sizes*. We will almost exclusively refer to *constant step-size* situations (i.e. $t_{n+1} - t_n = h$ for all n), in spite of the fact that, in practice, *variable step-sizes* should always be used (more on this later).

The (explicit) *Euler rule* recursively defines the numerical solution by

$$\mathbf{y}_{n+1} := \mathbf{y}_n + h\mathbf{f}(\mathbf{y}_n, t_n) \tag{2.2}$$

($\mathbf{y}_0 = \boldsymbol{\alpha}$) and provides the canonical example of an integration method. If the solution $\mathbf{y}(\cdot)$ of (2.1) exists, then we can define the associated *truncation error*

$$\mathbf{TE}_{n+1} := \mathbf{y}(t_{n+1}) - \mathbf{y}(t_n) - h\mathbf{f}(\mathbf{y}(t_n), t_n) \tag{2.3}$$

which clearly has a Taylor expansion

$$\mathbf{TE}_{n+1} = \frac{1}{2}h^2\mathbf{y}''(t_n) + \cdots \tag{2.4}$$

provided that $\mathbf{y}(\cdot)$ is smooth. Since (2.4) starts with h^2, we say that (2.2) is *consistent of the first order* with (2.1). The truncation error possesses a nice interpretation: $\mathbf{TE}_{n+1}$ is the difference between the true solution $\mathbf{y}(t_{n+1})$ and the result $\mathbf{y}(t_n) + h\mathbf{f}(\mathbf{y}(t_n), t_n)$ of an Euler step taken from $\mathbf{y}(t_n)$.

Another useful example is given by the *implicit Euler rule*

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(\mathbf{y}_{n+1}, t_{n+1}). \tag{2.5}$$

For each n (2.5) provides d real equations to be solved for the d real components of $\mathbf{y}_{n+1}$. Usually some iterative procedure must be employed to find $\mathbf{y}_{n+1}$ numerically and as a result the cost of a step $t_n \mapsto t_{n+1}$ with (2.5) is considerably higher than that of a step with (2.2). Since the truncation error of (2.5),

$$\mathbf{TE}_{n+1} = \mathbf{y}(t_{n+1}) - \mathbf{y}(t_n) - h\mathbf{f}(\mathbf{y}(t_{n+1}), t_{n+1}), \tag{2.6}$$

satisfies

$$\mathbf{TE}_{n+1} = -\frac{1}{2}h^2\mathbf{y}''(t_n) + \cdots, \tag{2.7}$$

comparison with (2.4) makes it clear that, in general, there is no reason why (2.2) should not be preferred to (2.5) (see, however, Section 4 below).

It is perhaps useful to point out that for the implicit method $\mathbf{TE}_{n+1}$ cannot be interpreted as the difference between 'true' $\mathbf{y}(t_{n+1})$ and the numerical solution, integrating from $\mathbf{y}(t_n)$.

In spite of their simplicity, both (2.2) and (2.5) are still 'state of the art' methods e.g. in cases where the dimension d is so high as to preclude the use of more sophisticated schemes. However for most problems (2.2) and (2.5) are too naive and some of the methods in the next subsections should definitely be preferred.

2.2 Linear multistep methods

A linear multistep method (LM) is specified by a positive integer k (the number of steps) and constants $\alpha_i, \beta i$, $i = 0, 1, \ldots, k$ with $\alpha_k = 1$. Once $\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_{n+k-1}$, $n \geq 0$, have been found, $\mathbf{y}_{n+k}$ is defined through

$$\sum_{i=0}^{k} \alpha_i \mathbf{y}_{n+i} = h \sum_{i=0}^{k} \beta_i \mathbf{f}(\mathbf{y}_{n+i}, t_{n+i}). \tag{2.8}$$

(See e.g. Lambert (1973), Sections 2.1–4)

If $\beta_k = 0$ the method is *explicit*. Otherwise the method is *implicit* and at each step a d-dimensional system must be solved. In either case, $\mathbf{y}_1, \ldots, \mathbf{y}_{k-1}$ should be suitably chosen before the application of (2.8) can begin. It is customary to associate with (2.8) the polynomials

$$\rho(z) = \alpha_k z^k + \ldots + \alpha_0, \quad \sigma(z) = \beta_k z^k + \ldots + \beta_0. \qquad (2.9)$$

These specify the method and are referred to as *characteristic polynomials*.

The *truncation error* is defined by

$$\mathbf{TE}_{n+k} := \sum_{i=0}^{k} \alpha_i y(t_{n+i}) - h \sum_{i=0}^{k} \beta_i \mathbf{f}(\mathbf{y}(t_{n+i}), t_{n+i}). \qquad (2.10)$$

For explicit methods $\mathbf{TE}_{n+1}$ is the difference between the true $\mathbf{y}(t_{n+1})$ and the approximation that (2.8) would yield if $\mathbf{y}_{n+k-1}, \ldots, \mathbf{y}_n$ were 'exact', i.e. $\mathbf{y}_{n+k-1} = \mathbf{y}(t_{n+k-1}), \ldots, \mathbf{y}_n = \mathbf{y}(t_n)$.

The method is said to be *consistent of order* p if

$$\mathbf{TE}_{n+k} = C_{p+1} h^{p+1} \mathbf{y}(t_n) + \ldots, \quad C_{p+1} \neq 0 \qquad (2.11)$$

whenever the solution $\mathbf{y}$ of (2.1) is sufficienty smooth. *Consistent* means consistent of order p for some $p \geq 0$. In (2.11) C_{p+1} represents a constant depending only on $\{\alpha_i, \beta_i\}$. (See e.g. Lambert (1973), Section 2.6). It is a simple matter to see that the requirement 'order $\geq p$' imposes $p+1$ independent linear constraints on the $2k+1$ parameters $\alpha_{k-1}, \ldots, \alpha_0, \beta_k, \ldots, \beta_0$. Hence order of consistency $2k$ is possible with k-step formulae. The constraints for order at least 1 (i.e. consistency) are

$$\alpha_0 + \alpha_1 + \cdots + \alpha_k = 0, \quad \alpha_1 + \cdots + k\alpha_k - \beta_0 - \beta_1 - \cdots - \beta_k = 0 \quad (2.12)$$

The best known LM formulae are the so-called Adams methods. These have $\alpha_k = 1, \alpha_{k-1} = -1, \alpha_{k-2} = \ldots = \alpha_0 = 0$ and the β_i chosen so as to maximize the order of consistency. With k steps, the explicit Adams (or Adams–Bashforth) method is of order k and its implicit counterpart (Adams–Moulton method) is of order $k+1$. In practice implicit Adams formulae are preferred, but the equations for $\mathbf{y}_{n+k}$ are only solved approximately by computing one or two iterants of the obvious fixed point iteration in (2.8). An accurate initial guess to start such an iteration is found by applying an explicit Adams formula. The overall algorithm is called a predictor-corrector pair (Lambert (1973), Section 3.9).

The sophisticated 'state of the art' software is written around Adams predictor-corrector pairs and, as the integration proceeds, changes both the step-lengths and the number of steps of the formula so as to increase efficiency (see e.g. Hairer *et al.* (1987), Section III.7 and Shamphine and Gordon (1975)). The number of steps of the formulae built-in a given code ranges from 1 tc 12, say.

2.3 Runge–Kutta methods

A Runge–Kutta (RK) method is specified by an integer s (the number of *stages*) and constants a_{ij}, $1 \leq i, j \leq s$, b_i, $1 \leq i \leq s$. When $\mathbf{y}_n$ has been found, auxiliary vectors $\mathbf{Y}_{n,i}$, $1 \leq i \leq s$, are defined through

$$\mathbf{Y}_{n,i} = \mathbf{y}_n + h \sum_{j=1}^{s} \alpha_{ij} \mathbf{f}(\mathbf{Y}_{n,j}, t_n + c_j h), \qquad (2.13)$$

with

$$c_i = \sum_j \alpha_{ij}, 1 \leq i \leq s. \qquad (2.14)$$

Then one sets

$$\mathbf{y}_{n+1} := \mathbf{y}_n + h \sum_{i=1}^{s} \beta_i \mathbf{f}(\mathbf{Y}_{n,i}, t_n + c_i h). \qquad (2.15)$$

If $a_{ij} = 0$ for $i \leq j$, the vectors

$$\mathbf{Y}_{n,1} = \mathbf{y}_n, \quad \mathbf{Y}_{n,2} = \mathbf{y}_n + h a_{21} \mathbf{f}(\mathbf{Y}_{n,1}, t_n + c_2 h) \quad \text{etc.}$$

can be easily computed. Then the method is called *explicit*. Note however that to find $\mathbf{y}_{n+1}$, with an explicit RK method demands s evaluations of the function $\mathbf{f}$, whereas the cost of an explicit linear k-step formula (2.8) is of *one* evaluation per grid point, regardless of the value of k.

For implicit RK methods (2.13) provides a system of ds real equations for the ds real components of the vectors $\mathbf{Y}_{n,j}$. This should be compared with the situation for (2.8) where the system is only d-dimensional, regardless of the value of k.

To analyze the consistency of (2.13–14) it is customary to observe that, for smooth $\mathbf{f}$, the system

$$\mathbf{Y}_i = \mathbf{y} + h \sum_{j=1}^{s} \alpha_{ij} \mathbf{f}(\mathbf{Y}_j, t + c_j h), \quad 1 \leq i \leq s \qquad (2.16)$$

implicitly defines, for small h, functions $\mathbf{Y}_i = \mathbf{Y}_i(\mathbf{y}, t, h)$ with $\mathbf{Y}_i(\mathbf{y}, t, 0) = \mathbf{y}$. On setting

$$\Phi(\mathbf{y}, t, h) := \sum_{i=1}^{s} \beta_i \mathbf{f}(\mathbf{Y}_i, t + c_i h), \qquad (2.17)$$

the step $t_n \mapsto t_{n+1}$ in (2.14) can be written as

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h \Phi(\mathbf{y}_n, t_n, h), \qquad (2.18)$$

an extension of (2.2)

Both for explicit and implicit RK methods the truncation error is defined by

$$\mathbf{TE}_{N+1} = \mathbf{y}(t_{n+1}) - \mathbf{y}(t_n) - h\Phi(\mathbf{y}(t_n), t_n, h). \tag{2.19}$$

and possesses the interpretation 'true $\mathbf{y}(t_{n+1})$ minus numerical if $\mathbf{y}_n = \mathbf{y}(t_n)$'.

The method is said to have order of consistency p if the Taylor expansion of $\mathbf{TE}$ as a function of h ($h \to 0$) begins with h^{p+1} terms whenever $\mathbf{f}$ and $\mathbf{y}$ are smooth. Even though this is analogous to (2.11), there are two important differences. (i) While in (2.11) the Taylor expansion involves only derivatives of the solution $\mathbf{y}$, in the RK case one also finds partial derivatives of $\mathbf{f}$. (ii) While the conditions for (2.8) to have order p are very easy to derive, it is a major task to systematically obtain conditions on a_{ij}, b_i that guarantee that (2.13–14) possess order p. This was first achieved by J. Butcher (1963). See Hairer *et al.* (1987), Section II.2 for a simple presentation and Butcher (1987) for a more comprehensive treatment. With s stages, order $2s$ can be achieved (Gauss–Legendre methods).

Standard codes employ variable step-sizes. The actual value of h_n to be used at t_n is determined by comparing the approximations at t_n obtained by the method being employed and an auxiliar RK method (embedded pairs, see Hairer *et al.* (1987), Section 11.4). A pair due to Prince and Dormand (1981) incorporating a 13 stages, order 8 formula is very popular.

Before we close this section, we observe that it is by no means obvious that in LM or RK implicit methods the equations (2.8) or (2.13) defining the numerical approximation possess a unique solution. Even for explicit methods the existence of the numerical solution is not to be taken for granted: If s is not defined in the whole of $\mathcal{R}^d \times I$, (2.8) or (2.13) may require that $\mathbf{f}$ be evaluated outside its domain of definition. Due to lack of space, these issues cannot be discussed here (see Sanz-Serna (1985a), Lopez-Marcos and Sanz-Serna (1988)) and we will always assume that (2.8) or (2.13–14) uniquely define the numerical approximation.

3 Classical error bounds

The classical theory of error bounds for numerical ODE methods was derived in the fifties, mainly by Dahlquist (1959). It is assumed that, in (2.1), I is compact ($I = [0,T]$) and $\mathbf{f}$ is globally Lipschitz continuous with respect to $\mathbf{y}$ in $\mathcal{R}^d \times [0,T]$. (Extensions to cases where $\mathbf{f}$ is Lipschitz only in a tube around the solution $\mathbf{y}(t)$ are feasible, see page 25 of Shampine and Gordon (1975) and López-Marcos and Sanz-Serna (1988)). The boundedness of I however cannot be dispensed with.)

Consider first Euler's formula (2.2). Along with $\{\mathbf{y}_n\}$ we take a pertubed Euler solution $\{\mathbf{v}_n\}$ satisfying

$$\mathbf{v}_{n+1} = \mathbf{v}_n + h\mathbf{f}(\mathbf{v}_n, t_n) + \delta_{n+1}, \tag{3.1}$$

where δ_{n+1} represents any pertubation. Subtraction of (3.1) from (2.2) yields ($L > 0$ denotes the Lipschitz constant)

$$\|\mathbf{v}_{n+1} - \mathbf{y}_{n+1}\| \le (1 + Lh)\|\mathbf{v}_n - \mathbf{y}_n\| + \|\delta_{n+1}\|, \tag{3.2}$$

and recursion leads to

$$\|\mathbf{v}_n - \mathbf{y}_n\| \le (1 + Lh)^n\|\mathbf{v}_0 - \mathbf{y}_0\| + (1 + Lh)^{n-1}\|\delta_1\| + \ldots + \|\delta_n\|. \tag{3.3}$$

Now, since $nh = t_n \le T$ and $(1 + Lh)^m \le \exp(Lmh)$ for $m \ge 0$, we obtain

$$\|\mathbf{v}_n - \mathbf{y}_n\| \le \exp(LT)\{\|\mathbf{v}_0 - \mathbf{y}_0\| + \sum_{m=1}^{n} \|\delta_m\|\}, \tag{3.4}$$

a *stability estimate* that bounds the change in Euler solution in terms of the perturbations. The key fact is that the factor $\exp(LT)$ does not depend on h. If we let the theoretical vectors $\{\mathbf{y}(t_n)\}$ play the role of $\{\mathbf{v}_n\}$, then (2.3) shows that the corresponding pertubations are $\delta_n = \mathbf{TE}_n$ and (3.3) reads

$$\|\mathbf{y}(t_n) - \mathbf{y}_n\| \le \exp(LT) \sum_{m=1}^{n} \|\mathbf{TE}_m\|, \tag{3.5}$$

which, according to (2.5) implies, for any grid point $t_n \le T$ and if $\mathbf{y}(t)$ is smooth,

$$\|\mathbf{y}(t_n) - \mathbf{y}_n\| = \mathcal{O}(nh^2) = \mathcal{O}(h), \quad h \to 0. \tag{3.6}$$

This shows *first order of convergence*, i.e. an $\mathcal{O}(h)$ behaviour for the errors in the approximate solution $\{\mathbf{y}_n\}$. To sum up, (3.5) (convergence) results from (2.5) (consistency) and (3.3) (stability).

The stability estimate (3.3) is a discrete counterpart of the Gronwall bound

$$\|\mathbf{v}(t) - \mathbf{y}(t)\| \le \exp(LT)\{\|\mathbf{v}(0)\| + \int_0^T \|\delta(\tau)\|d\tau\}, \tag{3.7}$$

$0 \le t \le T$, for the difference between the solution $\mathbf{y}$ of (2.1) and the solution of a perturbed problem

$$\mathbf{v}'(t) = \mathbf{f}(\mathbf{v}(t), t) + \delta(t), \quad \mathbf{v}(0) \text{ given.} \tag{3.8}$$

In ODE circles (3.6) would be referred to as a 'well-posedness' estimate rather than as a stability estimate, the word stability being used in connection with $t \to \infty$ situations. *Thus, when a numerical analyst says 'Euler's rule is stable', a person with an ODE background should interpret it as meaning 'the Euler recursion is well posed (uniformly in h) in any compact interval'.*

Turning now to the implicit Euler rule, convergence of the first order as in (3.5) is easily proved from (2.7) and a stability estimate. The latter is derived by considering along with (2.5) a perturbed solution.

$$\mathbf{v}_{n+1} = \mathbf{v}_n + h\mathbf{f}(\mathbf{v}_{n+1}, t_{n+1} + \boldsymbol{\delta}_{n+1}, \tag{3.9}$$

Now, instead of (3.2), we have, from (2.5) and (3.7)

$$\|\mathbf{v}_{n+1} - \mathbf{y}_{n+1}\| \leq (1 - Lh)^{-1}\{\|\mathbf{v}_n - \mathbf{y}_n\| + \|\boldsymbol{\delta}_{n+1}\|\}, \tag{3.10}$$

provided that $Lh < \frac{1}{2}$, say. Recursion in (3.8) easily leads to an estimate like (3.3), with $\exp(LT)$ replaced by $\exp(2LT)$.

All RK methods possess a stability estimate similar to (3.3). This is derived very much as in (3.1–2), starting from the format (2.15). (Φ inherits its Lipschitz character from $\mathbf{f}$). Consequently all RK methods consistent of the p-th order satisfy $\|\mathbf{y}(t_n) - \mathbf{y}_n\| = O(h^p)$, i.e. they are convergent of order p.

For LM methods the situation is more complex. The stability analysis is best performed by rewriting (2.8) as a one-step recursion. To do so, it is of course enough to consider the kd-dimensional vectors $\mathbf{Z}_{n+1} = [\mathbf{y}_{n+k}^T, \ldots, \mathbf{y}_{n+1}^T]^T$ that satisfy

$$\mathbf{Z}_{n+1} = M\mathbf{Z}_n + h\mathbf{F}(\mathbf{Z}_{n+1}, \mathbf{Z}_n, t_n, h) \tag{3.11}$$

where M is the companion matrix of the characteristic polynomial $\rho(z)$ in (2.9) and $\mathbf{F}$ is a Lipschitz function of its first and second arguments. The stability of the recursion (3.9) depends on the spectrum of M (Hairer *et al.* (1987), Section III.4), i.e. on the roots of P. It turns out that a stability bound for (3.9) (or equivalently (2.8)) exists if and only if p satisfies the so-called *root condition*, namely if all its roots are in the closed unit circle and roots with unit modulus are simple. (Note that (2.12) shows that, for the consistent methods, 1 is always a root.) Dahlquist (1959) proved that high order of consistency is not compatible with the root condition. Although, as we saw, there are k-step methods or orders of consistency up to $2k$, the order of a stable method is at most $k+1$ if k is odd and $k+2$ if k is even. For stable methods of order of consistency p, convergence or order p holds provided that the missing starting values $\mathbf{y}_1, \ldots, \mathbf{y}_{k-1}$ are sufficiently accurate. This is easily established by an argument like that leading to (3.4–5).

4 Absolute stability

It is a remarkable fact that most information on numerical methods has traditionally been obtained by looking at their performance on the simple scalar equation

$$y' = \lambda y, \quad \lambda \text{ a complex constant.} \tag{4.1}$$

This performance is in principle easily analyzed because for (4.1) the numerical solution y_n can be found in closed form in terms of h and λ. Rather than studying arbitrary LM or RK methods (see e.g. Lambert (1973)), we will present some illuminating particular cases. Also, for simplicity we restrict our attention here to the case where λ in (4.1) is *real and negative*.

4.1 Always wrong

As a first example consider the explicit mid-point rule, i.e. the 2-step method with $\rho(z) = z^2 - 1$, $\sigma(z) = 2z$. The order of consistency is 2 and the root condition is 'just' satisfied: the roots of ρ, namely ± 1, are on the boundary to the unit disk. The application to (4.1) reads, when written as a one-step recursion,

$$\mathbf{Z}_{n+1} = \begin{bmatrix} -2h\lambda & 1 \\ 1 & 0 \end{bmatrix} \mathbf{Z}_n \tag{4.2}$$

Since λ has been assumed to be negative, solutions of (4.1) approach each other exponentially and the Gronwall estimate (3.6), that predicts exponential *growth* of pertubations, is far too pessimistic. By analogy, one would expect that the corresponding classical error bound for the mid-point rule would also be a gross overestimation. However this is not the case, as the matrix in (4.2) possesses an eigenvalue with modulus > 1, so that the numerical solution, and hence the error, *do* grow exponentially with n. Obviously, this shows that the mid-point rule in general cannot be recommended as a good numerical method. From our point of view is important to emphasize that *convergent methods may well generate, for any chosen value of the step-length, sequences* $\{\mathbf{y}_n\}$ *with the wrong qualitative behaviour*. There is no contradiction: convergence refers to compact time intervals and $h \to 0$, qualitative behaviour refers to fixed h, t_n growing unboundedly. This phenomenon has been known, at least, since Dahlquist (1959).

Note that as h tends to 0, the eigenvalues of the matrix in (4.2) tend to ± 1, the roots of ρ. It is easy to see that 'always wrong' behaviour like this studied here cannot take place either for RK methods or for LM methods that satisfy the so-called *strong root condition*: the roots of ρ are 1 (simple) and $k-1$ complex numbers with moduli < 1 (Stetter (1973), Theorem 4.6.4, Lambert (1973) p.67). For such methods the magnitude of the numerical approximation to (4.1), $\mathrm{Re}\,\lambda < 0$, decreases exponentially for *h sufficiently small*.

4.2 Always right

We now take the implicit Euler method (2.5). Its application to (4.1), $\lambda < 0$, results in the recurrence $y_{n+1} = (1 - h\lambda)^{-1} y_n$. Since $0 < (1 - h\lambda)^{-1} < 1$, the numerical solution tends monotically to 0, thus exhibiting the right qualitative behaviour for all values of h.

The good behaviour of (4.5) shown here for the model problem (4.1) holds for any dissipative problem (2.1) where $\mathbf{f}$ satisfies (angular brackets denote an inner product)

$$\langle \mathbf{f}(\mathbf{v},t)-\mathbf{f}(\mathbf{w},t),\mathbf{v}-\mathbf{w}\rangle \leq 0, \tag{4.3}$$

in its domain of definition (Dekker and Verwer (1984), Sections 2.4, 2.5). Clearly, (4.3) implies that solutions of the system in (2.1) become closer to each other as t increases. Numerical solutions computed by the backward Euler method also become closer to each other and it is possible to improve substantially the classical error bounds (which grow exponentially with T).

4.3 Sometimes right. Stiffness

Our final example is given by Euler's rule (2.2). The application to (4.1), $\lambda < 0$, leads to the recusion $y_{n+1} = (1+h\lambda)y_n$ and the numerical solution grows exponentially unless h is taken $< 2/|\lambda|$. If $\lambda \ll -1$, this is a severe restriction on h. However, it is less severe than the restriction on h deriving from the requirement that the local truncation error should be reasonably small, i.e. the requirement that h should be in line with the time scale in which the solution itself varies. For instance, we would roughly require, according to (2.4), $h < 0.3/|\lambda|$ to have local truncation errors of about 5%. Therefore the existence of a so-called *absolute stability restriction* $h < 2/|\lambda|$ is not a serious drawback of the method, *as applied to (4.1)*. However let us know consider the slightly more complicated nonhomogeneous *stiff* problem

$$y' = \lambda y + g(t), \quad \lambda = -10^6, \quad g(t) = \cos t - \lambda \sin t, \quad y(0) = 1, \tag{4.4}$$

with the solution

$$y(t) = \sin t + \exp(-10^6 t). \tag{4.5}$$

After a short transient, (4.5) is virtually identical to the $\sin t$ function, and steps of length $h = 0.1$, say, would be reasonable to keep the truncation error small. Nevertheless, subtraction of (2.2) and (2.3) yields

$$y(t_{n+1}) - y_{n+1} = (1-h10^6)[y(t_n)-y_n] + TE_{n+1}, \tag{4.6}$$

and accordingly the errors $y(t_n) - y_n$ will grow exponentially with n unless h satisfies the absolute stability restriction $h < 2/|\lambda| = 2\times 10^{-6}$. This renders Euler's rule unsuitable for (4.4). Many other numerical methods, including all explicit RK and LM formulae, cannot accurately integrate stiff problems unless the time step is chosen unreasonably small. Unfortunately, so-called stiff problems like (4.4) occur frequently in many areas of application, including time-integration of evolutionary partial differential equations, see e.g. Sanz-Serna and Verwer (1989). As a consequence the numerical treatment of stiff problems requires special RK or LM formulae and has attracted enormously the interest of numerical analysts, starting with the important 1963 paper by Dahlquist. A complete bibliography on numerical stiff problems would certainly include many hundreds of items. Initially, only the scalar model problem (4.1) was considered along with constant coefficient linear systems that reduce to (4.1) after diagonalization. Nevertheless, extensions to dissipative nonlinear problems like (4.3) have recently received much attention starting from the work of Dahlquist in the mid-seventies, Dahlquist (1978). See Dekker and Verwer (1984) for a summary.

It is important to bear in mind that in the developments just surveyed the term *stability* has a different meaning than in Section 3. There it referred to $h \to 0$ in connection with the idea of convergence. Here absolute stability refers to behaviour for fixed h, as t_n increases unboundedly. I would also like to emphasize that the analysis of the performance of numerical methods on the simple model (4.1) is mathematically deeper than the material in this section may suggest. See e.g. the theory of order stars (Wanner *et al.*, 1978).

5 The main question

The theory described so far has focused on the *quantitative* approximation of the solution of an initial value problem (2.1), mostly in a compact time interval. While it is true that numerical analysts have studied the *qualitative* behaviour (of families) of numerical solutions, such studies have traditionally been centred around dissipative problems (4.3) and have been considered relevant in as much as they helped to identify the behaviour of methods when applied to stiff problems. It is only in the last decade that the question *(Q)* posed in the introduction has attracted some attention in the numerical analysis community. Early works in that direction include Brezzie *et al.* (1984), Sanz-Serna (1985b), Mitchell and Griffiths (1986), Sanz-Serna and Vadillo (1986).

It is expedient to review here the example studied by Brezzi *et al.* (1984). They consider the complex equation

$$\frac{\mathrm{d}z}{\mathrm{d}t} = \left(i + s - |z|^2\right) z, \tag{5.1}$$

where s is a real parameter. Due to the rotational symmetry of (5.1), it is possible to derive a scalar real equation for the evolution of the variable $q = |z|^2$, namely

$$\frac{\mathrm{d}q}{\mathrm{d}t} = 2(s-q)q. \tag{5.2}$$

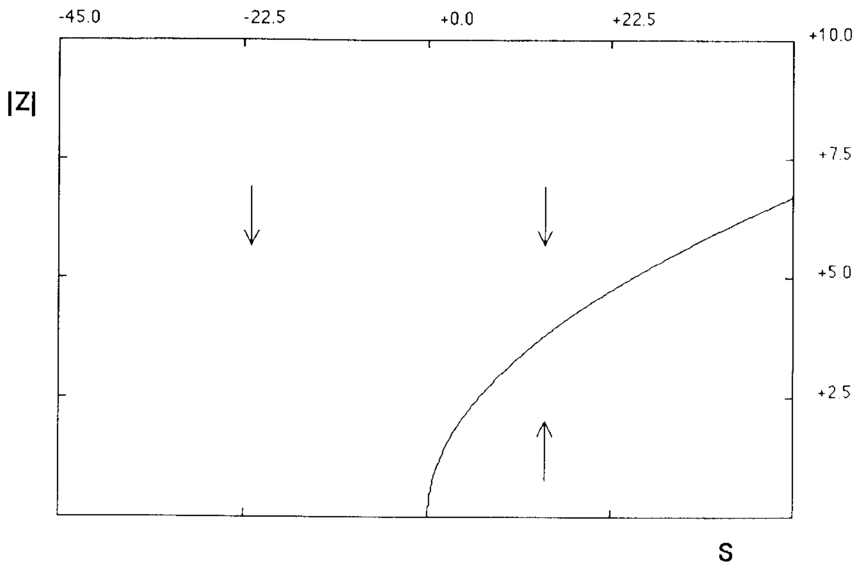

Figure 1.

The dynamics of (5.2) is easily described (T below stands for true):

(T1) For $s < 0$, $q = 0$ is the only equilibrium; all other solutions decrease towards 0. In terms of the original equation (5.1), all the solutions spiral towards the origin in the complex plane.

(T2) For $s > 0$ there are two equilibria given by $q_0 = 0$ and $q_H = s$. All trajectories, other than the equilibria, tend toward q_H. In terms of (5.1), we have an equilibrium at the origin and an invariant curve with equation $|z| = s^{\frac{1}{2}}$. The latter attracts all trajectories other than $z = 0$. Clearly the orbitally stable invariant curves originate from a Hopf bifurcation at $s = 0$ (Chow and Hale (1982), p.8, Guckenheimer and Holmes (1986), p.150).

Figure 1 depicts the dynamics of (5.1) in the plane $(s, |z|)$.

Let us study the application of Euler's method (2.2) to (5.1). The rotational symmetry of (5.1) is retained by the discretization and easy algebra shows that the Euler formula implies the following recursion for the approximations q_n to $q(t_n)$:

$$q_{n+1} = g(q_n) := [(1 + h(s - q_n))^2 + h^2]q_n \tag{5.3}$$

The dynamics of (5.3) with h *fixed*, $h < \frac{1}{2}$ and s varying is as follows (N stands for numerical):

(N1) For $s < s_+ = -\left(1 + \sqrt{1 - h^2}\right)/h$ one finds $g' > 0$ and $g'' > 0$. Therefore the origin is a repellor and the nontrivial numerical solutions increase monotonically towards ∞. This is just the opposite to the true *(T1)*.

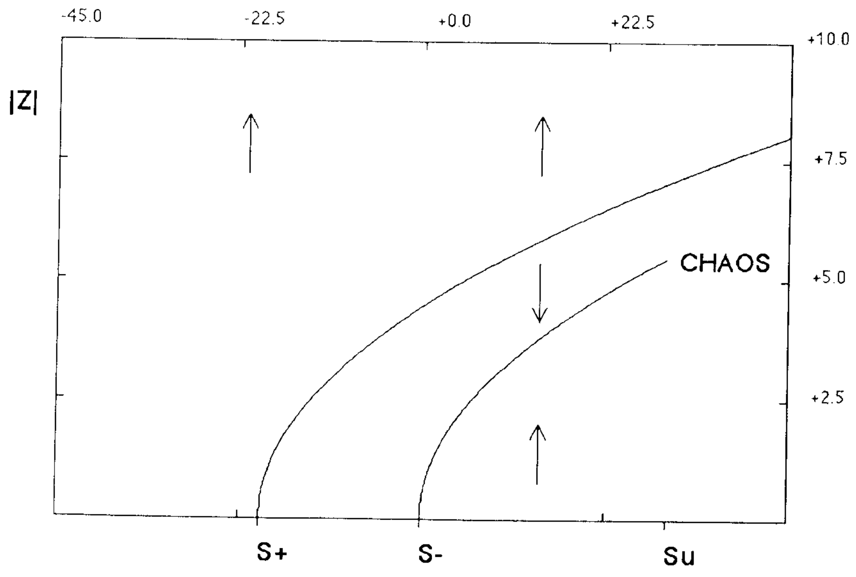

Figure 2.

(N2) For $s_+ < s < s_- := -\left(1 - \sqrt{1 - h^2}\right)/h$ (5.3) possesses fixed points at the origin and at $q_+ = s + \left(1 + \sqrt{1 - h^2}\right)/h$. The latter has no counterpart in (5.2). An initial point larger than q_+ generates a solution that increases to ∞. Initial points below q_+ give rise to the right qualitative behaviour: attraction towards the origin.

(N3) For $s_- < s$ there are three equilibria of (5.3). These are the origin, q_+ (the spurious equilibrium found above) and $q_- = s + \left(1 - \sqrt{1 - h^2}\right)/h$. The latter is an $\mathcal{O}(h)$ approximation to the true equilibrium q_H of (5.2). There are two subcases to be considered.

(N3a) $s_- < s < s_u := (-2 + h^2)/\left(h\sqrt{1 - h^2}\right) - 1/h$. Here q_- is stable (just as its true counterpart q_H). Its basin of attraction is the interval $(0, q_+)$.

(N3b) $s_u < s$. Here q_- is unstable. The dynamics can be very complicated including chaos and the reader is referred to Brezzi *et al.* (1984) for more information.

Figure 2 corresponds to the Euler dynamics with $h = 0.1$ and should be compared with Figure 1. The branch corresponding to the Hopf equilibrium q_- is only presented for $s < s_u$.

The following conclusions may be drawn, depending on whether you are a pessimist *(P)* or an optimist *(O)*.

(P) Whatever the value of s and whatever the value of h, no matter how small, the z-plane dynamics of the Euler discretization of (5.1) is widely different form the true dynamics.

(O) If the attention is restricted to a bounded region $|z| < R$, $|s| < S$, then for h small enough (how small would of course depend on R and S), the Euler discretization of the parameterized equation (5.1) approximates the true dynamics. In fact, as h decreases, $s_- \rightarrow -\infty$ and $s_u \rightarrow \infty$. As a result, for h small, only the regimes *(N2)* and *(N3a)* are found in $|s| < S$. Furthermore $q_+ \rightarrow \infty$ as $h \rightarrow 0$, so that the spurious q_+ eventually leaves $|z| < R$. For example, the value $h = 0.1$ in Figure 2 is small enough for the true and numerical dynamics to coincide for, say $|z| < 2.5$, $|s| < 10$. Note in particular that the Hopf bifurcation at $s = 0$ is faithfully inherited by the discretization.

The next two sections are devoted to exploring the optimistic and pessimistic points of view respectively.

6 Optimism

For simplicity, in the remainder of the paper we only consider *autonomous systems*

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}). \tag{6.1}$$

We organize the presentation around the invariant objects of (6.1).

6.1 Equilibria

Equilibria are duly inherited by both RK and LM methods. If $\mathbf{y}^*$ is an equilibrium of (6.1) $\mathbf{f}(\mathbf{y}^*) = \mathbf{0}$, then for any method and any step-length h, $\mathbf{y}^*$ is also a zero of the function Φ in (2.15) (which for (6.1) does not depend on t) and hence an equilibrium of the RK dynamics. This is trivial to check. For LM methods, the kd-dimensional vector $\mathbf{Z}* = [\mathbf{y}^{*T}, \dots, \mathbf{y}^{*T}]^T$ is also easily seen to be an equilibrium of the associated recursion (3.9).

Assume furthermore that $\mathbf{y}^*$ is a hyperbolic sink of (6.1). The stability of $\mathbf{y}^*$ or $\mathbf{Z}^*$ as equilibria of the numerical recursions (2.6) or (3.9) is of course studied by linearizing (2.6) or (3.9) around the equilibrium. Such linearizations turn out to coincide with the result of the application of the numerical method to the linearization around $\mathbf{y}^*$ of the system (6.1), i.e. the processes of linearization and discretization commute. In this way we are led to the analysis of the qualitative behaviour of numerical solutions of asymptotically stable, linear, constant coefficient problems, a task which, as we saw, is familiar to numerical analysts. According to the discussion in Section 4, $\mathbf{y}^*$ and $\mathbf{Z}^*$ are hyperbolic sinks of (2.6), (3.9) respectively, provided that h is *small enough* and, for LM methods, that the strong root condition is satisfied. It should also be pointed out that in this case it is possible to derive error estimates for the difference $\mathbf{y}(t_n) - \mathbf{y}_n$ that hold uniformly for $0 \leq t < \infty$, cf. Stetter (1973), Section 3.5. To sum up, near sinks, (sensible) numerical methods have the right dynamics, provided that the step-length is rightly chosen.

If $\mathbf{y}^*$ is not a sink of (6.1), the situation is more subtle and has been studied by Beyn (1987a). For simplicity we only review here the results for RK methods. If $\mathbf{y}^*$ is hyperbolic and the RK method (2.13–14) is consistent of order p, the local stable and unstable manifolds of $\mathbf{y}^*$ as an equilibrium of the numerical recursion (2.15) approximate their counterparts in the ODE system (6.1) with errors $\mathcal{O}(h^p)$. Furthermore, in a neighbourhood Ω of $\mathbf{y}^*$ the numerical dynamics reproduces the true dynamics in the following sense. There exist constants C and h_0 such that if $\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_N$ are in Ω and satisfy (2.15) with $h < h_0$, then there is a suitable initial condition $\mathbf{x}_0$ such that if $\mathbf{x}_n$ denotes the value at $t = nh$ of the solution of (6.1) starting at $\mathbf{x}_0$, then, for $0 \leq n \leq N$, $\mathbf{x}_n$ is well defined and $\|\mathbf{x}_n - \mathbf{y}_n\| \leq Ch^p$. Thus, each numerical orbit O_h is uniformly close to an orbit O (of the h-flow) of the system of ODEs. However, the 'true' orbit O being approximated will possess an initial value $\mathbf{x}_0$ different from the starting vector $\mathbf{y}_0$. This is analogous to what in dynamical systems is known as *shadowing* (Bowen (1975), Chapter 3B).

If the equilibrium $\mathbf{y}^*$ of (6.1) is not hyperbolic there is a centre manifold (Guckenheimer and Holmes (1983), Section 3.2), a situation whose discretization has been studied by Beyn and Lorenz (1987).

6.2 Hyperbolic periodic orbits

The behaviour of numerical methods near a periodic orbit P of (6.1) was first investigated by Braun and Hershenov (1977), who only considered one-step methods like (2.15) and *stable* orbits. They showed that, for h small, there is a closed curve P_h in the $\mathbf{y}$-space which is close to P and invariant for the numerical iteration. A similar result was given by Doan (1985) for multistep methods satisfying the strong root condition and general *hyperbolic* periodic orbits P. For k-step methods like (2.8) the invariance of P_h must be understood in the following sense: for any point $\mathbf{y}_{k-1}$ on P_h there exist $\mathbf{y}_0, \dots \mathbf{y}_{k-2}$ on P_h such that the numerical solution of (2.8) with starting vectors $\mathbf{y}_0, \dots \mathbf{y}_{k-1}$ stays on P_h. Doan's results have in turn been improved by Beyn (1987b), Eirola (1988), (1989) and Eirola and Nevanlinna (1989). The first papers deal with one-step methods and the last with the multistep situation. All of them provide $\mathcal{O}(h^p)$ bounds for the distance between P and P_h in suitable norms (here, as before, p denotes the order of the method).

It is clear that in the result just quoted the hypothesis that P is hyperbolic is essential. When P is not hyperbolic, systems in the neighbourhood of (6.1) may or may not have a periodic orbit near P and accordingly P is likely to disappear in the process of discretization. An example is provided by the (non-hyperbolic) closed orbits of the linear centre $y' = iy$, (y complex). Working as in Section 4, it is easily seen that most methods generate orbits that spiral either towards the origin or towards infinity.

When (6.1) depends on a parameter s, periodic oribts are often born from a branch $\mathbf{y}^*(s)$ of equilibria *via* Hopf bifurcation at a critical value s_c of the parameter. In the case of Euler's method Brezzi *et al.* (1984) showed that for h sufficiently small there is a critical value $s_c(h)$ of the parameter so that the numerical recursion undergoes a Hopf bifurcation (in the sense of mappings) at $s_c(h)$. Furthermore $s_c(h) - s_c = \mathcal{O}(h)$. This situation has been illustrated in Figure 2, where the Euler dynamics has a Hopf bifurcation at $s_c(h) = s_- = s_-(h)$. Hofbauer and Iooss (1984) study general RK methods applied to systems with a Hopf bifurcation, but their investigation is limited to the behaviour of the numerical method at $s = s_c$. Eirola and Nevanlinna in a paper presented at the 1989 Numerical ODE meeting in London analyzed the behaviour of general numerical methods near s_c.

Other useful references in this connection are Mahar (1982a), (1982b).

6.3 Other invariant objects

Kloeden and Lorenz (1986), (1990) have shown that, if Λ is a compact attracting set of arbitrary shape for (6.1), then numerical discretization possesses a nearby attracting set Λ_h. Beyn (1987c) studies the effect of discretization on homoclinic orbits.

7 Pessimism

As mentioned in Section 4, it is now a long time since the literature first presented cases where a convergent numerical method generates the wrong qualitative behaviour, either for a given time-step or for all choices of time-step. However such misbehaviour had traditionally been studied in linear problems, such as (4.1), where the trouble used to be that the method would approximate an exponentially decreasing true solution by an exponentially increasing numerical solution. As expected, the class of possible pathologies grows dramatically when moving to nonlinear problems. Yamaguti and Ushiki (1980), (1981) and Ushiki (1982) proved that even for simple equations such as $y' = -y(1-y)$ and simple methods such as Euler's rule (2.2) or the m -point rule (cf. Section 4.1), the numerical solution could exhibit chaos. in the case of Euler's rule, chaotic orbits appear for values of h too large to be considered meaningful from a numerical analysis point of view (see below). However for the mid-point rule, chaos may appear for any choice of h. The general study of the dynamics of the mid-point rule was taken up by Sanz-Serna (1985b), Sanz-Serna and Vadillo (1986), (1987). Prüfer (1985) considers the logistic equation $y' = y(1-y)$ and shows the chaoticity of some orbits generated by Adams–Bashforth LM methods. See also Sleeman *et al.* (1988).

A more systematic analysis is performed by Iserles (1987), (1990) who undertakes a study of the equilibria of numerical methods. Among other things, Iserles notes that for RK methods the function Φ in (2.15) usually has spurious zeros $\mathbf{y}^*$, i.e. vectors $\mathbf{y}^*$ such that, for a given h, $\Phi(\mathbf{y}^*, h) = \mathbf{0}$, while $\mathbf{f}(\mathbf{y}^*) \neq \mathbf{0}$. Such a $\mathbf{y}^*$ would represent an equilibrium (possibly asymptotically stable) of the RK dynamics which does not approximate an equilibrium of the true dynamics. Iserles emphasizes that the pathology consisting of a numerical solution being attracted by a spurious equilibrium may not be easily discovered by some numerical analysts. In fact many numerical analysts have been brought up with the linear theory, where the only pathology to be feared is spurious growth. Such individuals are not likely to suspect that a numerical orbit nicely setting into a equilibrium may be completely spurious. It is of some interest to note that there are certain *implicit* RK methods for which spurious equilibria cannot occur. These methods are called regular and have been characterized by Hairer *et al.* (1989).

Spurious dynamics arise not only from spurious equilibria, but also from spurious periodic orbits, spurious invariant curves *etc.* Further references are Stuart (1989), Stuart and Peplow (1989), Iserles *et al.* (1990), Iserles and Stuart (1990).

It may be useful to present a simple example to illustrate the foregoing ideas. Consider the equation $y' = -y(1-y)$ integrated by Euler's rule. The equilibria $y = 0, 1$ are duly inherited by the Euler map

$$y_{n+1} = [1 - h(1 - y_n)]y_n. \tag{7.1}$$

Linearization of (7.1) around 0 yields $y_{n+1} = (1-h)y_n$. Clearly, 0 is a stable equilibrium of (7.1) if $0 < h < 2$ (see Section 4.3). At the critical value $h = 2$, the relevant eigenvalue of (7.1) leaves the unit disk through -1 and thus $y = 0$ suffers a flip bifurcation (Guckenheimer and Holmes (1983), Section 3.5), whereby the stability of the origin is transferred to a spurious period 2 solution of (7.1). In turn this period 2 solution becomes unstable at a higher value of h to give rise to a period 4 solution, etc. In fact, on setting $y_n = -(1+h)z_n/h+1$, (7.1) becomes the familiar $z_{n+1} = (1+h)z_n(1-z_n)$, one of the most often quoted examples in nonlinear dynamics.

If the 2-stage RK discretization $Y_2 = y_n - (h/2)y_n(1-y_n)$, $y_{n+1} = y_n - hY_2(1-Y_2)$, is used, the origin at 0 again becomes unstable at $h = 2$. The relevant eigenvalue now leaves the unit circle through 1 and a stable spurious equilibrium is present for h near 2, $h > 2$ (transcritical bifurcation, Guckenheimer and Holmes (1983), Section 5.3).

In more general terms the situation is as follows. Given a sensible numerical method (e.g. an RK method or a LM method satisfying the strong root condition), if $\mathbf{y}^*$ is a sink of (6.1) then, for h small enough, say $h < h_c$, $\mathbf{y}^*$ is also an asymptotically stable equilibrium of the numerical recursion. At the critical value $h = h_c$ the point $\mathbf{y}^*$ loses stability because one or more of the relevant eigenvalues μ cross the unit disk. The value h_c is easily determined from the linear theory of numerical methods as in Section 4. Generically, if μ crossed through 1, a spurious equilibrium is born that inherits the stability previously enjoyed by $\mathbf{y}^*$. A crossing through -1 results in a bifurcation to a stable spurious period 2 solution. Two complex conjugate eigenvalues leaving together the unit disk lead to a spurious Hopf bifurcation. In practice, and as mentioned before, the critical value h_c obtained *via* linear absolute stability theory is in general larger than the value of h one would use with a view to having an accurate integration. It is nevertheless possible that the branches of stable spurious objects born at $h = h_c$ turn back and exist for values of $h \ll h_c$.

8 Discussion

The material above shows that it is only recently that numerical ODE researchers have turned their attention to the question *(Q)* posed in the introduction. One of the reasons why most classical analysis of ODE numerical methods is not useful in connection with *(Q)* is the following. Traditionally, numerical ODE researchers perform *forward error analysis*, i.e. they see the result $\mathbf{y}_n$ of a numerical computation as an approximation to the true solution $\mathbf{y}(t_n)$. In other branches of numerical mathematics, notably in numerical linear algebra, error analysis is done in a *backward* manner, whereby the numerical solution of a problem P is seen as the *exact* solution of an approximate problem P_h. In an ODE context a backward error analysis would interpret numerical orbits $\{\mathbf{y}_n\}$ as true orbits of a system of ODEs close to that being solved. If a complete backward error analysis of numerical ODE methods were available, then *(Q)* would be equivalent to the standard question of whether the dynamics of given system of differential equations is the same as the dynamics of its neighbouring systems. The advantages of a backward error analysis over the forward error analysis sketched in Section 3 would not be confined to the question being addressed so far. In fact, in most instances the system S being integrated is only a model of complex real world situation, so that the true solution $\mathbf{y}(t_n)$ itself may be suspected to be only an approximation. In such a situation seeing $\mathbf{y}_n$ as a true solution of a nearby model S_h is clearly advantageous.

The results reported in Section 6 may endorse the optimistic view that the answer to *(Q)* is affirmative, at least locally, under the assumption that the step-length h is chosen to be sufficiently small. Pessimists argue that, in practice, it may be difficult to decide when the value of h being used is 'sufficiently small'. The obvious idea would be to use smaller and smaller values of h until things appear to converge. However, pessimists point out that, in many practical situations, one works at the limit of the capacity of the machine, so that a reduction in h may not be feasible (see e.g. Stuart and Peplow (1989)).

In my opinion, the pessimistic school may well have overstated their case. Many of the pathologies studied by them occur in situations where the environment of the experiment is not really a *bona fide* numerical simulation set-up. I shall try to illustrate this with an experiment to be presented later. There is another reason why I would rather answer *(Q)* affirmatively: Whilst it is a fact that numerical methods may exhibit wild spurious dynamics if h is *not chosen judiciously*, there is also factual evidence that a significant part of our knowledge of nonlinear dynamics has been (rightly) derived by observing the dynamics of numerical methods.

On the other hand, it is fair to say that articles of the pessimistic school are likely to have a positive impact on numerical analysts. Traditionally numerical analysts have been brought up in a linear world and efforts tending to make people aware of truly nonlinear phenomena should be welcome. Numerical analysts should also be made aware of the fact that the dynamics of a mapping is, in general, very different from ODE dynamics (see e.g. Guckenheimer and Holmes (1983), Sections 1.4, 3.5).

To end the section, let us take up again the Euler discretization of (5.1). For any fixed value of s, we saw that, regardless of the choice of h, Euler's rule leads to the wrong dynamics if $|z|$ is large. This is hardly surprising. In the Euler formula

$$z_{n+1} = z_n + \left[h\left(i + s - |z_n|^2\right) z_n\right] \tag{8.1}$$

the term is square brackets should represent a small correction being added to z_n to obtain z_{n+1}. If h and s are fixed, then for $|z_n|$ large, the term in brackets is actually much larger than z_n and accordingly the numerical method is not used in the set-up it was meant to work. Equivalently, h should be chosen in line with the rate of change in z, and this rate is strongly dependent on s and $|z|$. Even if a numerical analyst decided, for some strange reason, to use Euler's rule to discretize (5.1), he would realize that a fixed value of h will simply not do for all values of s and $|z|$ and that some form of adaptive step-length should be used.

For the sake of the argment, assume that our numerical analyst implements the following variable step-size strategy. He choses a small number $\mu > 0$ and sets

$$z_{n+1} = z_n + \left[h_n\left(i + s - |z_n|^2\right) z_n\right], \tag{8.2}$$

with

$$h_n = \frac{\mu}{|i + s - |z_n|^2|}. \tag{8.3}$$

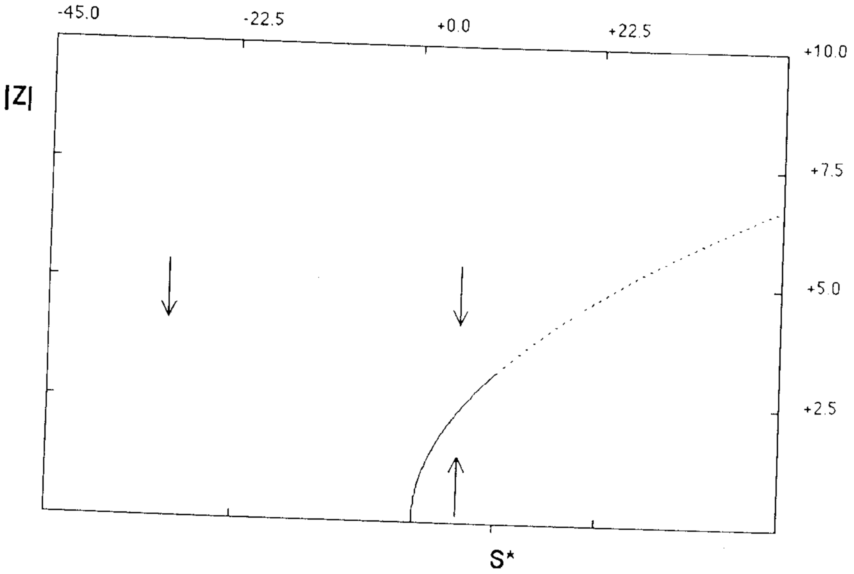

Figure 3.

Note that now the magnitude term in brackets in (8.1) is a fraction μ of the magnitude of z_n. Formulae (8.1–2) may be interpreted as an imbedded RK pair based on Euler's rule and on the 0-th order method $z_{n+1} = z_n$; then the term in brackets in (8.1) is the estimation of the truncation error in the 0-th order method and μ plays the role of a prescribed tolerance on the truncation error.

The dynamics of (8.1–2) is easily found to be as follows (see Figure 3, where $\mu = 0.1$).

*(N*1)* For $s < s_c(h) := -\mu/\sqrt{4-\mu^2} = \mathcal{O}(\mu)$ all solutions spiral towards the origin.

*(N*2)* For $s > s_c(h)$, the origin is unstable and so is the point at infinity. There is a Hopf branch of invariant curves with $|z|^2 = s + \mu/\sqrt{4-\mu^2}$, i.e. an $\mathcal{O}(\mu)$ approximation to the Hopf orbits of (5.1). There are two subcases:

*(N*a)* $0 < s < s^*$, where s^* satisfies $s^* = \mu^{-1} + \mathcal{O}(\mu)$. Here the Hopf branch of (8.1–2) is asymptotically stable, just as its continuous counterpart.

*(N*b)* $s^* < s$. Here the Hopf branch of (8.1–2) is unstable.

Comparison of Figures 2 and 3 shows that the variable time-step strategy is a big improvement. Now the numerical dynamics is 'almost' right: only in the case *(N*2)* there is a discrepancy in the stability of the Hopf orbits. However this is not too serious. First of all this only happens in a regime $s > s^* = \mu^{-1} + \mathcal{O}(\mu)$ that will not be seen if μ is small and large values of s are not of interest. Secondly *even for large s* the numerical dynamics is to some extent right. In fact consider, for $s > 0$, the scaled variable $r := |z|^2/s$. For all $s > 0$, the value $r = 1$ corresponds to the invariant circle of (5.1). Now (8.1–2) imply for the approximations r_n to $r(t_n)$ a recursion $r_{n+1} = \Psi_s(r_n)$ with

$$\Psi_s(r) = \left[1 + 2\mu\frac{1-r}{\sqrt{(1-r)^2 + s^{-2}}} + \mu^2\right], \quad r \geq 0. \tag{8.4}$$

In the limit $s \to \infty$, Ψ_s tends to the piecewise linear function Ψ_∞ given by

$$\Psi_\infty(r) = \begin{cases} (1-2\mu+\mu^2)r & : r > 1, \\ (1+2\mu+\mu^2)r & : r < 1 \end{cases}. \tag{8.5}$$

The dynamics of the iteration $r_{n+1} = \Psi_\infty(r_n)$ is as follows. Initial points near the origin increase monotonically until they enter the interval $[1 - 2\mu + \mu^2, 1 + 2\mu + \mu^2]$; initial points near infinity decrease until they enter this interval. Hence, in the r variable, instead of the attractor $r = 1$ of (5.1), we find an interval around $r = 1$, with width $\mathcal{O}(\mu)$, that is eventually entered by all solutions (other than the trivial equilibrium). Returning to the z-plane, we find that the dynamics of (8.1–2) is correct, with the only proviso that, in the regime *(N*b)*, the stable invariant curve of (5.1) at $|z| = s^{\frac{1}{2}}$ is approximated by an invariant annulus of width $s^{\frac{1}{2}}\mathcal{O}(\mu)$.

In conclusion, the pathologies of Euler's rule studied in Section 5 disappear (almost) completely as soon as the step-lengths are chosen judicisously. Variable time-step strategies do provide a sensible way of chosing h_n. Therefore the advantages of *using library software packages with variable step-lengths* rather than writing our own's fixed h software cannot be overemphasized. Many people working with PDEs must deal with very large problems which unfortunately cannot be directly plugged to a package. Even in such cases, I believe, it is always feasible to implement some sort of simple variable step strategy, a course of action which enhances the efficiency of the computation and the reliability of the results.

9 Other questions

(i) Following Beyn (1987c), it may be said that the long time behaviour of systems of ODEs can be numerically investigated *via* two different approaches. In the *indirect approach* the system of ODEs is integrated with several initial conditions and the long time behaviour of the numerical trajectories is observed. It is in this set-up that question *(Q)* is relevant. In the *direct approach* one numerically solves the defining equations for limit sets of the system (such as steady states or periodic orbits), determines their stability properties, etc. This approach is out of our scope here.

(ii) In the important particular of Hamiltonian systems, many features of the dynamics are determined by the symplectic or canonical character of the flow (see e.g. Arnold (1989)). Recently, there has been much interest in producing numerical methods that induce, for all values of the step-length, a canonical transformation in phase space. Such so-called symplectic or canonical methods automatically inherit some qualitative properties of the ODE flow. Some references are Ruth (1983), Feng (1986), Sanz-Serna (1988), Lasagni (1988), Suris (1989), Sanz-Serna (1990), Sanz-Serna and Abia (1990). When a completely integrable Hamiltonian system (Arnold, 1989) is integrated by means of a canonical method, KAM theory can be used to show that the numerical dynamics preserve most invariant tori (Sanz-Serna and Vadillo, 1986, 1987).

(iii) The procedure of time-step selection in numerical methods has been analyzed from a dynamical system point of view by Hall (1985), (1986), Griffiths (1988), Higham and Hall (1990).

(iv) Other useful papers are Kirchgraber (1986), Kirchgraber and Pospiech (1986), Kirchgraber (1988).

Acknowledgement. The author has been partly supported by project CICYT PB-86-0313.

References

1. Arnold, V.I. (1989). *Mathematical methods of classical Mechanics (2nd ed.)*. Springer-Verlag, Berlin.

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

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

4. Beyn, W.J. (1987c). The effect of discretization on homoclinic orbits. In: *Bifurcation, Analysis, Algorithms, Applications* (Küpper, T., Seydel, R. and Troger, H., eds.), 1–8. Birkhauser-Verlag, Basel.

5. Beyn, W.J. and Lorenz, J. (1987). Center manifolds of dynamical systems under discretizations. *Numer. Funct. Anal. and Optimiz.* **9**, 381–414.

6. Bowen, R. (1975). *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*. Lecture Notes in Mathematics 470, Springer-Verlag, Berlin.

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

8. Brezzi, F., Ushiki, S. and Fujii, H. (1984). "Real" and "ghost" bifurcation dynamics in difference schemes of ODEs. In: *Numerical Methods for Bifurcation Problems* (Küpper, T., Mittleman, H.D. and Weber, H., eds.), 79–104. Birkhauser-Verlag, Basel.

9. Butcher, J.C. (1963). Coefficients for the study of Runge–Kutta integration processes. *J. Austral. Math. Soc.* **3**, 185–201.

10. Butcher, J.C. (1987). *The Numerical Analysis of Ordinary Differential Equations*. John Wiley, Chichester.

11. Chow, Sh-N. and Hale, J.K. (1982). *Methods of Bifurcation Theory*. Springer-Verlag, New York.

12. Dahlquist, G. (1959). *Stability and Error Bounds in the Numerical Integration of Ordinary Differential Equations*. Trans. of the Royal Inst. of Techn. (No. 130), Stockholm.

13. Dahlquist, G. (1963). A special stability problem for linear multistep methods. *BIT* **3**, 27–43.

14. Dahlquist, G. (1978). G-stability is equivalent to A-stability. *BIT* **18**, 384–401.

15. Dekker, K. and Verwer, J.G. (1984). *Stability of Runge–Kutta Methods for Stiff Nonlinear Differential Equations*. North-Holland, Amsterdam.

16. Doan, H.T. (1985). Invariant curves for numerical methods. *Quart. Appl. Math.* **43**, 385–393.

17. Eirola, T. (1988). Invariant curves of one-step methods. *BIT* **28**, 113–122.

18. Eirola, T. (1989). Two concepts for numerical periodic solutions of ODEs. *Appl. Math. Comput.* **31**, 121–131.

19. Eirola, T. and Nevanlinna, O. (1988). What do multistep methods approximate? *Numer. Math.* **53**, 559–569.

20. Feng, K. (1986). Difference schemes for Hamiltonian formalism and symplectic geometry. *J. Comput. Mat.* **4**, 279–289.

21. Griffiths, D.F. (1988). The dynamics of some linear multistep methods with step-size control. In *Numerical Analysis 1987* (Griffiths, D.F. and Watson, G.A., eds), Pitman, Harlow.

22. Guckenheimer, J. and Holmes, P. (1983). *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Springer-Verlag, New York.

23. Hairer, E., Nørsett, S.P. and Wanner, G. (1987). *Solving Ordinary Differential Equations I. Nonstiff Problems*. Springer-Verlag, Berlin.

24. Hairer, E., Iserles, A. and Sanz-Serna, J. M. (1989). Equilibria of Runge–Kutta methods. *Numerische Math.* (to appear).

25. Hall, G. (1985). Equilibrium states of Runge–Kutta schemes. *ACM Trans. Math. Software* **11**, 289–301.

26. Hall, G. (1986). Equilibrium states of Runge–Kutta schemes: Part II. *ACM Trans. Math. Software* **212**, 183–102.

27. Higham, D.J. and Hall, G. (1990). Embedded Runge–Kutta formulae with stable equilibrium states. *J. Comput. Appl. Math.* **29**, 25–33.

28. Hofbauer, J. and Iooss, G. (1984). A Hopf bifurcation theorem for difference equations approximating a differential equation. *Mh. Math.* **98**, 99–113.

29. Iserles, A. (1987). Dynamical systems and nonlinear stability theory for numerical ODEs. In *Numerical Treatement of Differential Equations* (Strehmel, K., ed.), 84–94. Teubner, Leipzig.

30. Iserles, A. (1990). Stability and dynamics of numerical methods of nonlinear ordinary differential equations. *IMA J. Numer. Anal.* **10**, 1–30.

31. Iserles, A., Peplow, A.T. and Stuart, A.M. (1990). A unified approach to spurious solutions introduced by time discretisation. Part I: Basic theory. *University of Cambridge, Numerical Analysis Report DAMTP/1990 NA4*.

32. Iserles, A., and Stuart, A.M. (1990). A unified approach to spurious solutions introduced by time discretisation. Part II: BDF-like methods, *University of Cambridge, Numerical Analysis Report DAMTP/1990 NA6*.

33. Kirchgraber, U. (1986). Multi-step methods are essentially one-step methods. *Numer. Math.* **48**, 85–90.

34. Kirchgraber, U. (1988). An ODE-solver based on the method of averaging. *Numer. Math.* **53**, 621–625.

35. Kirchgraber, U. and Pospiech, G. (1988). An extrapolation method for the efficient composition of maps with applications to non-linear oscillations. *Computing* **36**, 343–354.

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

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

38. Lambert, J.D. (1973). *Computational Methods in Ordinary Differential Equations*. John Wiley, London.

39. Lasagni, F. (1988). Canonical Runge–Kutta methods. *ZAMP* **39**, 952–953.

40. López-Marcos, J.C. and Sanz-Serna, J.M. (1988). A definition of stability for nonlinear problems. In *Numerical Treatment of Differential Equations* (Strehmel, K., ed.), 216–226. Teubner, Leipzig.

41. Mahar, T.J. (1982a). Discrete conservative oscillators: periodic and asymptotically periodic solutions. *SIAM J. Numer. Anal.* **19**, 231–236.

42. Mahar, T.J. (1982b). Discrete almost-linear oscillators. *SIAM J. Numer. Anal.* **19**, 237–244.

43. Mitchell, A.R. and Griffiths, D.F. (1986), Beyond the linearised stability limit in nonlinear problems. In *Numerical Analysis* (Griffiths, D.F. and Watson, G.A., eds), 187–199, Longman, London.

44. Prince, P.J. and Dormand, J.R. (1981). High order embedded Runge–Kutta formulae. *J. Comp. Appl. Math.* **7**, 67–75.

45. Prüfer, M. (1985). Turbulence in multistep methods for initial value problems. *SIAM J. Appl. Math.* **45**, 32–69.

46. Ruth, R. (1983). A canonical integration technique. *IEEE Trans. Nucl. Sci.* **30**, 2669–2671.

47. Sanz-Serna, J.M. (1985a). Stability and convergence in numerical analysis I: Linear problems, a simple comprehensive account. In *Nonlinear Differential Equations* (Hale, J.K. and Martinez Amores, P., eds), 64–113. Pitman, Boston.

48. Sanz-Serna, J.M. (1985b). Studies in numerical nonlinear instability I. Why do leapfrog schemes go unstable? *SIAM J. Sci. Stat. Comput.* **6**, 923–938.

49. Sanz-Serna, J.M. (1988). Runge–Kutta schemes for Hamiltonian systems. *BIT* **28**, 877–883.

50. Sanz-Serna, J.M. (1990). The numerical integration of Hamiltonian systems. To appear.

51. Sanz-Serna, J.M. & Abia, L. (1990). Order conditions for canonical Runge–Kutta schemes. *SIAM J. Numer. Anal.* (to appear).

52. Sanz-Serna, J.M. and Vadillo, F. (1986). Nonlinear instability, the dynamic approach. In *Numerical Analysis* (Griffiths, D.F. and Watson, G.A., eds.), 187–199. Longman, London.

53. Sanz-Serna, J.M. and Vadillo, F. (1987). Studies in nonlinear instability III: augmented Hamiltonian systems. *SIAM J. Appl. Math.* **47**, 92–108.

54. Sanz-Serna, J.M. and Verwer, J.G. (1989). Stability and convergence at the PDE/stiff ODE interface. *Appl. Numer. Meth.* **5**, 117–132.

55. Shampine, L.F. and Gordon, M.K. (1975). *Computer Solution of Ordinary Differential Equations*. W.H. Freeman and Co., San Francisco.

56. Sleeman, B.D., Griffiths, D.F., Mitchell, A.R. and Smith, P.D. (1988). Period doubling bifurcation in nonlinear difference equations. *SIAM J. Sci. Stat. Comput.* 9, 543–557.

57. Stetter, H.J. (1973). *Analysis of Discretization Methods for Ordinary Differential Equations.* Springer-Verlag, Berlin–Heidelberg–New York.

58. Stuart, A.M. (1989). The global attractor under discretisation. Preprint.

59. Stuart, A.M. and Peplow, A. (1989). The dynamics of the theta method. Preprint.

60. Suris, Y.B. (1989). Canonical transformations generated by methods of Runge–Kutta type for the numerical integration of the system $x'' = -\partial U/\partial x$. *Zh. Vychisl. Mat. i Fiz.* **29**, 202–211 (in Russian).

61. Ushiki, S. (1982). Central difference scheme and chaos. *Physica D* **4**, 407–424.

62. Wanner, G., Hairer, E. and Nørsett, S.P. (1978), Order stars and stability theorems. *BIT* **18**, 475–489.

63. Yamaguti, M. and Ushiki, S. (1980). Discrétisation et chaos. *C.R. Acad. Sc. Paris.* **290**, 637–640.

64. Yamaguti, M. and Ushiki, S. (1981). Chaos in numerical analysis of ordinary differential equations. *Physica D* **3**, 618–626

Sensitivity of Bifurcations to Discretization

D.R. Moore and N.O. Weiss

Imperial College London and *University of Cambridge, England*

Abstract Bifurcations and transitions to chaos found in numerical studies of nonlinear PDEs may be artifacts introduced by discretization. A systematic procedure is developed which makes it possible to determine whether the bifurcation structure persists as truncation errors are consistently reduced. In particular, the presence of chaos can be established by precise tracking of narrow periodic windows within the chaotic regime. This procedure is applied to numerical experiments on two-dimensional thermosolutal convection.

1 Introduction

This review is slanted towards applied mathematicians. We shall only consider dissipative systems, with the aim of providing a bridge between the discussions of dynamical systems (Broomhead, 1991; Stewart, 1991) and of numerical analysis (Sanz-Serna, 1991) elsewhere in these Proceedings. The general problem arises if we wish to investigate a continuous macroscopic system governed by nonlinear partial differential equations (PDEs). Then we usually have to rely on numerical experiments, so we construct some discrete approximation to the PDEs, which is a related (but different) nonlinear system. If chaos appears, is it a property of the original PDEs or just a consequence of discretization?

The classic example is two-dimensional Rayleigh-Bénard convection, where a minimal Galerkin expansion reduces the PDEs to the well-known Lorenz (1963) system of ordinary differential equations (ODEs). This third-order system correctly describes the pitchfork bifurcation at the onset of convection but the nontrivial steady solutions undergo a subcritical Hopf bifurcation which is followed by a wealth of chaotic behaviour (Sparrow, 1982). Accurate numerical solutions of the PDEs show that the Hopf bifurcation is indeed there but it is supercritical and there is no chaos (Moore & Weiss, 1973; Curry *et al.*, 1984). In this instance, the approximation