Chapter 1

An Introduction to Numerical Integrators Preserving Physical Properties

Martin J. Gander* and Rita Meyer-Spasche†

Most real life applications lead to equations which cannot be solved analytically. Numerical methods are used to find solutions and it is not evident how to choose a method from the wide variety of methods available. Classical criteria include order of accuracy, stability and ease of implementation. We emphasize the important criterion of how much of the physics of the underlying problem the numerical method can preserve.

We first analyze classical numerical schemes to integrate ordinary differential equations and show that they are capable of reproducing the exact solution for corresponding classes of problems. Thus for those problems, all the underlying physical properties are preserved by the numerical scheme. We show that such schemes are of interest in applications if the main part of the problem is in the class integrated exactly.

Then we analyze how much of the underlying dynamics a scheme can preserve if it is not exact. This leads us to schemes which preserve fixed points and their stability, closed orbits by preserving the energy and symplectic schemes which preserve the area under the evolution map given by the physical problem.

*Department of Mathematics and Statistics, McGill University, 805 Sherbrooke Street West, Montreal, QC, H3A 2K6, Canada, (mgander@math.mcgill.ca).

†Max-Planck-Institut für Plasmaphysik, EURATOM-Association, D-85748 Garching, Germany (meyer-spasche@ipp-garching.mpg.de).

1.1 Introduction

There is a wide variety of methods available to integrate systems of ordinary differential equations and the choice of a suitable method is not evident. Methods are usually chosen for their

(a) order of accuracy,
(b) stability properties,
(c) code availability or ease of implementation.

Stability is the crucial issue, since systems of ordinary differential equations are often **stiff**, in particular when they come from discretizations of partial differential equations. Thus when choosing an explicit scheme, the step size of the time integration has to be much smaller than required by the structure of the approximated solution, because otherwise the numerical scheme becomes unstable. In addition it is common practice to solve initial boundary value problems for $t \to \infty$ in search of steady states. This saves computer storage and thus allows finer spatial discretizations, but uses up much computing time for the time stepping. It is thus desirable to compute with a large time step k to advance quickly to the steady state, the transient solution not being of interest. Implicit methods have the advantage that they allow larger time steps without going unstable. But they require much more computational effort per time step than explicit methods and there is the problem of superstability, where the numerical method delivers a decaying solution even though the underlying physical problem has an exponentially growing one. If the computational cost of a fully implicit method becomes too high, a mixture of explicit and implicit methods is common practice: implicit where necessary, explicit where possible.

Part of the problem is that the traditional theory of difference schemes does neither treat differential equations nor difference equations as dynamical systems. If the dynamical properties of the continuous equations are very different from the dynamical properties of their discretizations, we can still obtain convergence for bounded, closed time intervals and step-size $k \to 0$. But on open time intervals and for long time integration it is often difficult to guarantee small global errors. Over the last few years, numerical methods for evolution problems have been investigated using dynamical systems theory [3; 28; 32; 35, and the references therein]. Cases are known, for which a common method (forward Euler) produces discrete chaotic trajectories even though the underlying differential equation does

not have chaotic solutions (see Section 1.3.3). Other cases are known in which the differential equation does have chaotic trajectories which cannot be detected by a common method (backward Euler) [4].

In our investigations presented here we focus on the question of how much of the physics of the underlying problem the numerical method can preserve. We show that every choice of a numerical integration scheme involves a decision if the scheme to be used

(d) is exact for a subclass of problems,
(e) preserves the dynamics of the underlying problem,
(f) is conserving energy or flow like the underlying problem.

At present, this decision is made unconsciously in most cases, because the knowledge necessary for this type of decision is not available yet. It was our aim to increase this knowledge so that in the future it will be possible to include criteria (d), (e) and (f) when choosing the numerical method, in addition to the classical criteria (a), (b) and (c) mentioned above. We treat these three additional criteria separately in the different sections of this chapter. We chose typical model problems to do so and we use schemes of constant order with constant step sizes to simplify the exposition. The problems and phenomena we encounter and discuss are currently avoided in practice by very small step sizes, step size control and order control.

We start by analyzing in Section 1.2 classical numerical schemes to integrate ordinary differential equations and show that they are capable of reproducing the exact solution for corresponding classes of problems. Thus for those problems, all the underlying physical properties are preserved by the numerical scheme. We show that such schemes are of interest in applications if the main part of the problem is in the class of problems which are integrated exactly. Then we analyze in Sections 1.3 and 1.4 how much of the underlying dynamics a scheme which is not exact can preserve. This leads us to schemes which preserve fixed points and their stability, closed orbits by conserving the energy and symplectic schemes which preserve the area under the map given by the physical problem.

1.2 Exact Difference Schemes

In this section we consider standard schemes for scalar initial value problems

$$\dot{u} = f(u), \qquad u(0) = u_0 \in I\!R. \tag{1.1}$$

We investigate if standard schemes can reproduce the exact solution for certain classes of problems, i.e. for given right hand sides $f(u)$. Following [26] we discuss two different ways of finding exact schemes. They are related to the two questions:

(1) Given a standard numerical scheme, on which differential equations is it exact?
(2) Given a differential equation, which schemes are exact for it?

We start by answering the first question for several classical schemes. Traditionally, error expansions are used to obtain convergence results for all differential equations. We use these error expansions to find the most general differential equation for which a given numerical scheme is exact.

The second question is the traditional one asked when dealing with exact schemes. Answering this question is essentially the same as finding explicit solutions to the differential equation. Unfortunately this is not possible in general. If, however, a low-dimensional function space is known to contain the solution to be approximated, then a variable-coefficient Runge-Kutta scheme can be constructed which is exact on this function space. We present this approach following Ozawa [31].

We conclude this section by showing how exact schemes can be used to approximate blow-up solutions. We report on work by Le Roux [19] who used exact schemes for ordinary differential equations with blow-up solutions to obtain nonstandard schemes for parabolic differential equations with blow-up solutions.

1.2.1 *Standard Numerical Schemes as Exact Schemes*

We start with the first question raised above: given a numerical scheme, for which differential equations is it exact ? To this end, we define the truncation error of a numerical scheme. We restrict ourselves to discussing one step schemes, although the approach we take is more general. Suppose

we are given a numerical scheme to solve the differential equation (1.1),

$$u_{n+1} = \mathcal{A}(f)(u_n, k), \tag{1.2}$$

where k is the time step, u_n is an approximation to the exact solution $u(t_n)$ at time $t_n = nk,\ n = 0, 1, 2, \ldots$, and $\mathcal{A}(f)$ denotes the evolution map given by the numerical scheme. For the **forward Euler scheme**, for example, the evolution map $\mathcal{A}(f)$ is given by

$$u_{n+1} = \mathcal{A}(f)(u_n, k) = u_n + kf(u_n). \tag{1.3}$$

Note that for implicit methods, the evolution map $\mathcal{A}(f)$ requires a non-linear solve. In this exposition we assume that the explicit form (1.2) can always be obtained in exact arithmetic and we neglect the presence of rounding errors. We will also consider numerical methods for which the evolution map $\mathcal{A}(f)$ involves derivatives of f.

To estimate the numerical error of a scheme, we define the **truncation error** $T(u, k)$ of scheme (1.2) using the first step of the numerical method,

$$T(u, k) := \frac{1}{k}(u(k) - u_1) = \frac{1}{k}(u(k) - \mathcal{A}(f)(u_0, k)). \tag{1.4}$$

For the analysis we expand $T(u, k)$ in a Taylor series in k

$$T(u, k) = \sum_{j=0}^{\infty} \mathcal{B}_j(f(u))(0)k^j. \tag{1.5}$$

For forward Euler, for example, we find

$$\mathcal{B}_0 = 0, \qquad \mathcal{B}_j = \frac{1}{(j+1)!} \left.\frac{d^j f(u(t))}{dt^j}\right|_{t=0} \text{ for } j \geq 1. \tag{1.6}$$

Forward Euler is a **first order method**, because $\mathcal{B}_0 = 0$, $\mathcal{B}_1 \neq 0$.

Definition 1.1 A numerical scheme is an **exact scheme** for a given differential equation (1.1), if $\mathcal{B}_j(f(u))(0) = 0$ for all $j \geq 0$.

In the following we investigate several standard numerical schemes and derive the differential equations on which they are exact.

1.2.1.1 *First Order Schemes*

The two first order schemes mostly used are the forward Euler scheme and the backward Euler scheme. For the forward Euler scheme to be exact, we

have to obtain $T(u,k) = 0$. Using (1.6), we see that this is achieved if all total derivatives of $f(u(t))$ with respect to t vanish at $t = 0$. It thus suffices that the first total derivative vanishes since then all the higher derivatives will vanish as well. The first total derivative is given by

$$\left.\frac{df(u(t))}{dt}\right|_{t=0} = f'(u(t))\dot{u}(t)|_{t=0} = f'(u_0)f(u_0)$$

where we used the differential equation (1.1) in the last step. We thus find that the r.h.s. function f has to satisfy a nonlinear differential equation for the numerical scheme to be exact for any initial condition u_0,

$$f'f = 0. \tag{1.7}$$

This differential equation has the solutions $f(u) = 0$ and $f(u) = C$ where C is an arbitrary constant. Thus the forward Euler method is exact for any r.h.s. function which is a constant, which is not surprising, since such an equation has a linear solution and Euler is exact for linear solutions. Nevertheless this approach generalizes to arbitrary schemes and we will derive differential equations, nonlinear in general, to be satisfied by the r.h.s. function f such that the numerical method becomes exact.

For **backward Euler**,

$$u_{n+1} = u_n + kf(u_{n+1}) \tag{1.8}$$

we obtain for the truncation error (1.5) after a short calculation

$$\mathcal{B}_j = \begin{cases} 0 & j = 0 \\ \left(-\dfrac{j}{(j+1)!}\right) \left.\dfrac{d^j f(u(t))}{dt^j}\right|_{t=0} & j > 0 \end{cases} \tag{1.9}$$

and thus backward Euler is also a first order method, $\mathcal{B}_0 = 0$, $\mathcal{B}_1 \neq 0$. It is exact, if all the total derivatives with respect to t of $f(u(t))$ at $t = 0$ vanish, i.e. if we obtain $T(u,k) = 0$. We thus obtain the same differential equation to be satisfied by f as for forward Euler,

$$f'f = 0$$

Backward Euler is an exact scheme for $f(u) = C$, C an arbitrary constant and thus for all linear solutions $u(t)$.

1.2.1.2 *Second Order Schemes*

We start with the **trapezoidal rule**

$$\frac{u_{n+1}-u_n}{k} = \frac{f(u_{n+1})+f(u_n)}{2}, \qquad u_0 = u(0) \tag{1.10}$$

which is a method used very often in practice. As follows from the formulas given in [26, Lemma 3.3], its truncation error is given by (1.5) and

$$\mathcal{B}_j = \begin{cases} 0 & j \le 1, \\ \left(\dfrac{1}{(j+1)!} - \dfrac{1}{2j!}\right) \left.\dfrac{d^j f(u(t))}{dt^j}\right|_{t=0} & j > 1 \end{cases} \tag{1.11}$$

and the method is second order in general, $\mathcal{B}_0 = \mathcal{B}_1 = 0$, $\mathcal{B}_2 \neq 0$.

Lemma 1.1 *The trapezoidal rule is exact on equation (1.1) for any time step k if the right hand side function $f(u)$ satisfies the differential equation*

$$f''f + (f')^2 = 0. \tag{1.12}$$

Proof. We have to show that (1.12) implies that $\mathcal{B}_j = 0$ for all $j \ge 0$. As for the Euler methods, it suffices for the first non-zero term $\mathcal{B}_2$ in the expansion to vanish for all the others to vanish as well, since the higher order terms consist of higher total derivatives of the first non-zero term. So to make the first non-zero term $\mathcal{B}_2$ vanish, we need

$$\begin{aligned} \left.\frac{d^2 f(u(t))}{dt^2}\right|_{t=0} &= \frac{d}{dt}(f'(u(t))f(u(t)))|_{t=0} \\ &= f''(u_0)f^2(u_0) + (f'(u_0))^2 f(u_0) \end{aligned} \tag{1.13}$$

to vanish and thus f has either to vanish itself, $f(u) = 0$, or to satisfy the differential equation (1.12) which concludes the proof. □

Corollary 1.1 *The trapezoidal rule is exact for all r.h.s. functions f of the form*

$$f(u) = \sqrt{Cu + D}, \quad C > 0, D \in \mathbb{R} \tag{1.14}$$

and equivalently for solutions $u(t)$ satisfying

$$u(t) \in span\{1, t, t^2\}. \tag{1.15}$$

Proof. It suffices to solve (1.12) for f. Using

$$(\log f)' = \frac{f'}{f}, \quad (\log f')' = \frac{f''}{f'}$$

we divide (1.12) by ff', integrate and find

$$\log f = C_1 - \log f', \quad C_1 \in I\!R.$$

Taking the exponential on both sides, we are lead to the first order differential equation

$$f' = \frac{C_2}{f}, \quad C_2 = \exp(C_1) > 0,$$

which gives using separation of variables

$$\frac{1}{2} f^2 = C_2 u + C_3$$

and solving for f result (1.14) follows. To get result (1.15) we have from the differential equation (1.1)

$$\frac{d^2}{dt^2} f(u(t)) = \frac{d^3}{dt^3} u(t)$$

and thus the vanishing of the second total derivative of f corresponds to the vanishing of the third total derivative of u with respect to t which holds for polynomials of degree at most two. □

The next second order method we investigate is the linearly implicit **lintrap scheme**

$$\frac{u_{n+1} - u_n}{k} = f(u_n) + \frac{1}{2} f'(u_n)(u_{n+1} - u_n), \qquad u_0 = u(0). \tag{1.16}$$

Its truncation error is found after a short calculation in expanded form (1.5) with the coefficients

$$\mathcal{B}_j = \begin{cases} 0 & j \le 1, \\ \dfrac{1}{(j+1)!} \dfrac{d^j f(u(t))}{dt^j}\Big|_{t=0} - \dfrac{1}{2j!} f'(u_0) \dfrac{d^{j-1} f(u(t))}{dt^{j-1}}\Big|_{t=0} & j > 1 \end{cases} \tag{1.17}$$

and hence the method is second order in general, $\mathcal{B}_0 = \mathcal{B}_1 = 0$, $\mathcal{B}_2 \neq 0$.

Lemma 1.2 *The lintrap scheme is exact for (1.1) if f satisfies the differential equation*

$$2f''f - (f')^2 = 0. \tag{1.18}$$

Proof. We investigate first the differential equation imposed on f to have the first non-zero term $\mathcal{B}_2$ vanish and then show by induction, that this implies that all terms vanish in the error expansion. The first term is given by

$$\frac{1}{6}\left.\frac{d^2 f(u(t))}{dt^2}\right|_{t=0} - \frac{1}{4}f'(u_0)\left.\frac{df(u(t))}{dt}\right|_{t=0} = \frac{1}{6}f''(u_0)f^2 - \frac{1}{12}(f'(u_0))^2 f(u_0)$$

and thus it vanishes if either f vanishes or f satisfies the differential equation (1.18). By induction we show now that under this condition all the error terms vanish. Using the differential equation, the terms $\mathcal{B}_j$ can be written as functions of u instead of f,

$$\mathcal{B}_j = \frac{1}{(j+1)!}\left.\frac{d^{j+1}u(t)}{dt^{j+1}}\right|_{t=0} - \frac{1}{2j!}\frac{\ddot{u}(t)}{\dot{u}(t)}\left.\frac{d^j u(t)}{dt^j}\right|_{t=0}$$

and we have just shown that under condition (1.18) $\mathcal{B}_2 = 0$, which means in terms of $u(t)$

$$\dddot{u}(t) = \frac{3(\ddot{u}(t))^2}{2\dot{u}(t)}. \tag{1.19}$$

Hence for induction we assume for a given $j > 2$ that $\mathcal{B}_j = 0$, which is in terms of $u(t)$

$$u(t)^{(j+1)} = \frac{j+1}{2}\frac{\ddot{u}(t)}{\dot{u}(t)}u^{(j)}(t), \tag{1.20}$$

and we have to show that $\mathcal{B}_{j+1} = 0$, which means in terms of $u(t)$ the equality

$$u(t)^{(j+2)} = \frac{j+2}{2}\frac{\ddot{u}(t)}{\dot{u}(t)}u^{(j+1)}(t). \tag{1.21}$$

But using (1.20) we have

$$u(t)^{(j+2)} = \frac{d}{dt}(u^{(j+1)}) = \frac{j+1}{2}\left(u^{(j+1)}\frac{\ddot{u}}{\dot{u}} + \frac{\dddot{u}\dot{u} - (\ddot{u})^2}{(\dot{u})^2}u^{(j)}\right)$$

and using (1.19) to simplify the fraction we find

$$u(t)^{(j+2)} = \frac{1}{2}(j+1)\left(\frac{\ddot{u}}{\dot{u}}u^{(j+1)} + \frac{1}{2}\frac{\ddot{u}^2}{\dot{u}^2}u^{(j)}\right).$$

Now using the induction assumption (1.20) to replace $u^{(j)}$ we obtain

$$u(t)^{(j+2)} = \frac{1}{2}(j+1)\left(\frac{\ddot{u}}{\dot{u}}u^{(j+1)} + \frac{1}{2}\frac{\ddot{u}^2}{u^2}\frac{2}{j+1}\frac{\dot{u}}{\ddot{u}}u^{(j+1)}\right) = \frac{j+2}{2}\frac{\ddot{u}}{\dot{u}}u^{(j+1)}$$

which concludes the induction argument. □

Thus in the case of the lintrap scheme as well, it suffices to require that the r.h.s. function f is such that the first non-zero term in the error expansion vanishes to have all terms vanish. We get immediately

Corollary 1.2 *The lintrap scheme is exact for all right hand side functions f of the form*

$$f(u) = (Cu + D)^2, \quad C > 0,\ D \in I\!R \tag{1.22}$$

or equivalently for solutions $u(t)$ of the form

$$u(t) = -\frac{D}{C} - \frac{1}{C^2(t + C_1)}, \qquad C > 0, C_1, D \in I\!R, \tag{1.23}$$

Proof. To show (1.22) we simply integrate (1.18),

$$(\log f)' = 2(\log f')' \quad \Longrightarrow \quad f = C_1(f')^2, \quad C_1 > 0$$

which gives the first order differential equation

$$f' = \sqrt{\frac{f}{C_1}}$$

Integrating this differential equation using separation of variables leads to result (1.22). To show (1.23) we integrate the differential equation

$$\dot{u} = f(u) = (Cu + D)^2.$$

□

The **implicit midpoint rule** is the third example of second order we investigate. It is given by

$$\frac{u_{n+1} - u_n}{k} = f(\frac{u_{n+1} + u_n}{2}), \qquad u_0 = u(0), \tag{1.24}$$

and has the error expansion (1.5) with

$$\mathcal{B}_j = \begin{cases} 0 & j \le 1, \\ \dfrac{1}{(j+1)!}\left.\dfrac{d^j f(u(t))}{dt^j}\right|_{t=0} - \dfrac{1}{j!}\left.\dfrac{d^j f((u(t) - u_0)/2)}{dt^j}\right|_{t=0} & j > 1. \end{cases} \tag{1.25}$$

The method is thus second order accurate in general, $\mathcal{B}_0 = \mathcal{B}_1 = 0$.

Lemma 1.3 *The implicit midpoint rule is exact for (1.1) if f satisfies the differential equation*

$$f''f - 2(f')^2 = 0. \tag{1.26}$$

Proof. For the first non-zero term to vanish, we have to impose $\mathcal{B}_2 = 0$, which means

$$\begin{aligned}
\mathcal{B}_2 &= \frac{1}{6}\frac{d^2 f(u(t))}{dt^2}\bigg|_{t=0} - \frac{1}{2}\,\frac{d^2 f((u(t)-u_0)/2)}{dt^2}\bigg|_{t=0} \\
&= \frac{1}{6}\,\frac{df(u(t))f'(u(t))}{dt}\bigg|_{t=0} - \frac{1}{2}\,\frac{df'((u(t)-u_0)/2)\dot{u}(t)/2}{dt}\bigg|_{t=0} \\
&= \frac{f''(u_0)(f(u_0))^2 + (f'(u_0))^2 f}{3} - \frac{f''(u_0)(f(u_0))^2}{4} - \frac{(f'(u_0))^2 f(u_0)}{2} \\
&= \frac{1}{12}(f''(u_0)(f(u_0))^2 - 2(f'(u_0))^2 f(u_0)) \\
&= 0.
\end{aligned}$$

Again f either has to vanish, $f(u) = 0$, or it has to satisfy the differential equation (1.26). Integrating this differential equation,

$$2(\log f)' = (\log f')' \quad \Longrightarrow \quad f^2 = C_1 f', \quad C_1 > 0$$

we find the first order differential equation

$$f' = \frac{f^2}{C_1}$$

which integrated using separation of variables leads to the r.h.s function

$$f(u) = \frac{1}{D - Cu}, \quad C > 0, D \in I\!R \tag{1.27}$$

for which the first non-zero term vanishes. Integrating the differential equation with this right hand side,

$$\dot{u} = f(u) = \frac{1}{D - Cu}, \quad u(0) = u_0 \tag{1.28}$$

we find the solutions $u(t)$

$$u(t) = \frac{1}{C}\left(D \pm \sqrt{D^2 + C^2 u_0^2 - 2C(Cu_0 + t)}\right), \quad C > 0,\ D \in I\!R. \tag{1.29}$$

To show that the midpoint rule is exact, we apply the midpoint rule (1.24) to (1.28) and simplify,

$$\frac{u(k)-u_0}{k} - \frac{1}{D - C\frac{u(k)+u_0}{2}} = \frac{u^2(k) - \frac{2D}{C}u(k) + \frac{2}{C}k + \frac{2D}{C}u_0 - u_0^2}{k(u(k)+u(0)-2D/C)} \quad (1.30)$$

and show that the discrete scheme has the same solutions in this case as the underlying differential equation (1.28). Solving (1.30) for $u(k)$ we find indeed

$$u(k) = \frac{1}{C}\left(D \pm \sqrt{D^2 + C^2 u_0^2 - 2C(Du_0 + k)}\right)$$

which means that $u(k)$ lies in the solution space (1.29). Hence the implicit midpoint rule is exact for all r.h.s. functions of the form (1.27) which is equivalent to (1.26). □

Corollary 1.3 *The implicit midpoint rule is exact for all right hand sides f of the form*

$$f(u) = \frac{1}{D - Cu}, \quad C > 0, D \in I\!R.$$

and equivalently for all solutions $u(t)$ of the form

$$u(t) = \frac{1}{C}\left(D \pm \sqrt{D^2 - 2C(C_1 + t)}\right), \quad C > 0,\ C_1,\ D \in I\!R.$$

Proof. The proof is already contained in equations (1.27) and (1.29) of Lemma 1.3. □

1.2.1.3 *Higher Order Schemes*

We consider the family of **Taylor methods**, see for instance [1, p. 215ff]. The higher-order Taylor methods are obtained by adding more and more terms of the Taylor expansion to the numerical method, so the **Taylor method of order** m is given by

$$u_1 = \sum_{j=0}^{m} \frac{1}{j!} \frac{d^j f(u(t))}{dt^j}\bigg|_{t=0} k^j$$

where the total derivatives with respect to time are evaluated to lead to a numerical scheme. For example the *first order Taylor method* is identical to

the forward Euler method, while the *second order Taylor method* is given by

$$u_{n+1} = u_n + kf(u_n) + \frac{k^2}{2} f'(u_n) f(u_n).$$

The error expansion (1.5) for the Taylor method of order m contains the coefficients

$$\mathcal{B}_j = \begin{cases} 0 & j < m, \\ \dfrac{1}{(j+1)!} \dfrac{d^j f(u(t))}{dt^j}\Big|_{t=0} & j \geq m \end{cases} \tag{1.31}$$

and hence the method is in general m-th order accurate, $\mathcal{B}_0 = \mathcal{B}_1 = \ldots = \mathcal{B}_{m-1} = 0$, $\mathcal{B}_m \neq 0$.

Lemma 1.4 *The Taylor method of order 2 is exact for all f satisfying the differential equation*

$$f'' f + (f')^2 = 0. \tag{1.32}$$

Proof. It suffices for the Taylor methods that the first non-zero term of the error expansion vanishes for the method to become exact, because all the higher order terms are derivatives of the previous ones. For the second order Taylor method, the first non-zero error term is given by

$$\frac{1}{6} \frac{d^2 f(u(t))}{dt^2}\Big|_{t=0} = \frac{1}{6}(f''(u_0)(f(u_0))^2 + (f'(u_0))^2 f(u_0))$$

and thus the Taylor method of second order is exact if either f vanishes or if f satisfies the differential equation (1.32). □

Note that this is the same differential equation that we found for the trapezoidal rule and thus the second order Taylor method is exact on the same problems the trapezoidal rule is.

Corollary 1.4 *The second order Taylor method is exact for right hand side functions*

$$f(u) = \sqrt{Cu + D}, \quad C > 0, D \in I\!R.$$

and equivalently for solutions $u(t)$ satisfying

$$u(t) \in span\{1, t, t^2\}.$$

Proof. The proof is identical to the proof of Corollary 1.1. □

Next we consider the third order Taylor method.

Lemma 1.5 *The third order Taylor method is exact for f which satisfy the differential equation*

$$f'''f^2 + 4f''f'f + f^3 = 0. \tag{1.33}$$

Proof. The first non-zero term in the error expansion is $\mathcal{B}_3$ which is given by

$$\begin{aligned}\mathcal{B}_3 &= \frac{1}{24}\left.\frac{d^3 f(u(t))}{dt^3}\right|_{t=0} \\ &= \frac{1}{24}(f'''(u_0)(f(u_0))^3 + 4f''(u_0)f'(u_0)(f(u_0))^2 + (f'(u_0))^3 f(u_0).\end{aligned}$$

Since all higher order terms vanish if $\mathcal{B}_3 = 0$ the result (1.33) follows. □

Thus for each Taylor method a higher and higher order non-linear differential equation can be defined and if f satisfies this differential equation, the Taylor method will be exact for this type of problems. However it becomes difficult to solve the nonlinear differential equations for f for higher order Taylor methods. Looking at the solution space however, we find the following, intuitive

Lemma 1.6 *The Taylor methods of order m are exact for solutions satisfying*

$$u(t) \in span\{1, t, t^2, \ldots, t^m\}.$$

Proof. It is convenient to transform the error terms into derivatives of the solution u using the differential equation,

$$\frac{d^{j+1}}{dt^{j+1}}u(t) = \frac{d^j}{dt^j}f(u(t)).$$

We thus find for the error terms of the m-th order Taylor method

$$\mathcal{B}_j = \begin{cases} 0 & j < m, \\ \dfrac{1}{(j+1)!}\left.\dfrac{d^{j+1}u(t)}{dt^{j+1}}\right|_{t=0} & j \geq m. \end{cases}$$

Therefore the m-th order Taylor method is exact whenever the solutions are polynomials of degree m, as one expects from the construction of the Taylor methods. □

1.2.1.4 *Runge-Kutta Schemes*

We have observed so far that whenever the first non-zero term in the error expansion of a numerical scheme vanished because of a particular choice of the r.h.s. function f then the scheme became exact and all the error terms vanished. So one might wonder if this is the case for *all* numerical schemes. We show in this section that the answer is 'no'. To do so, we use the framework of Runge-Kutta methods. The trapezoidal rule can be written as a Runge-Kutta method,

$$\begin{aligned} U_1 &= u_n, \\ U_2 &= u_n + k(f(U_1) + f(U_2))/2, \\ u_{n+1} &= u_n + k(f(U_1) + f(U_2))/2, \end{aligned} \tag{1.34}$$

and the implicit midpoint rule as well,

$$\begin{aligned} U_1 &= u_n + kf(U_1)/2, \\ u_{n+1} &= u_n + kf(U_1). \end{aligned} \tag{1.35}$$

The lintrap scheme belongs to the closely related family of Rosenbrock schemes [13; 33]. We review briefly the results we need in the sequel about Runge-Kutta or RK schemes in one dimension. For the general case and for more details see for example [35, p. 214ff]. The general s-stage constant-coefficient RK scheme is given by

$$\begin{aligned} U_i &= u_n + k\sum_{j=1}^{s} a_{ij} f(U_i), \qquad i = 1, \ldots s, \\ u_{n+1} &= u_n + k\sum_{i=1}^{s} b_i f(U_i), \end{aligned} \tag{1.36}$$

for some given initial value u_0. A RK scheme is called *explicit* if

$$a_{ij} = 0 \text{ for } i \leq j, \quad i, j = 1, \ldots s. \tag{1.37}$$

For explicit schemes, every stage-value U_i is defined by one explicit equation for it. A scheme that is not explicit is called *implicit*. We continue to assume that implicit equations are solved exactly.

An s-stage scheme has $s^2 + s$ independent coefficients, a_{ij} and b_i, i.e. 6 independent coefficients if $s = 2$. It often happens that schemes with formally different coefficients produce the same numerical approximation. To reduce this redundancy, conditions have been formulated when a scheme

is 'S-reducible' or 'DJ-reducible'. A sufficient condition for a scheme to be S-irreducible is to be nonconfluent. An RK-scheme is called **nonconfluent** if $c_i \neq c_j$ for $i \neq j$, with

$$c_i := \sum_{j=1}^{s} a_{ij}, \qquad i = 1, \ldots s. \tag{1.38}$$

An **RK scheme** is said to be **of order** r if r is the largest integer such that for all functions $f \in C^\infty(I\!R)$ and all $u \in I\!R$

$$\lim_{k \to 0} \frac{||T(u,k)||}{k^r} < \infty \tag{1.39}$$

where the truncation error was defined in (1.4) to be

$$T(u,k) := \frac{1}{k}(u(k) - u_1),$$

with $u(k)$ denoting the exact solution of the differential equation (1.1) at time k and u_1 being the solution of the RK scheme with time step k and initial value $u_0 = u(0)$. For a scheme of order r, there is some f and some initial value u_0 such that inequality (1.39) does not hold any more if r is replaced by $r+1$. A scheme of order one is called **consistent** . An explicit s-stage scheme has order $r \leq s$, an implicit s-stage scheme has order $r \leq 2s$. Computing the error expansion (1.5) for the general RK method (1.36) we find

$$\mathcal{B}_0 = \left(\sum_{i=1}^{s} b_i - 1 \right) f \tag{1.40}$$

$$\mathcal{B}_1 = \left(\sum_{i}^{s} b_i c_i - \frac{1}{2} \right) f' f \tag{1.41}$$

$$\mathcal{B}_2 = \left(\sum_{i,j=1}^{s} b_i a_{ij} c_j - \frac{1}{6} \right) f f'^2 + \left(\frac{1}{2} \sum_{i=1}^{s} b_i c_i^2 - \frac{1}{6} \right) f^2 f'' \tag{1.42}$$

$$\begin{aligned} \mathcal{B}_3 = & \left(\sum_{i,j,k=1}^{s} b_i a_{ij} a_{jk} c_k - \frac{1}{24} \right) f f'^3 + \left(\frac{1}{6} \sum_{i=1}^{s} b_i c_i^3 - \frac{1}{24} \right) f''' f^3 \\ & + \left(\frac{1}{2} \sum_{i,j=1}^{s} b_i a_{ij} c_j^2 + \sum_{i,j=1}^{s} b_i c_i a_{ij} c_j - \frac{1}{6} \right) f^2 f' f'' \end{aligned} \tag{1.43}$$

$\vdots \qquad \vdots$

Hence the necessary and sufficient condition for 1st order accuracy (or consistency) is

$$\sum_{i=1}^{s} b_i = 1. \tag{1.44}$$

Necessary and sufficient for 2nd order accuracy is in addition

$$\sum_{i=1}^{s} b_i c_i = \frac{1}{2}, \tag{1.45}$$

and for 3rd order in addition

$$\sum_{i=1}^{s} b_i c_i^2 = \frac{1}{3}, \qquad \sum_{i=1}^{s}\sum_{j=1}^{s} b_i a_{ij} c_j = \frac{1}{6}, \tag{1.46}$$

and for 4th order:

$$\begin{aligned} \sum_{i=1}^{s} b_i c_i^3 &= \frac{1}{4}, & \sum_{i=1}^{s}\sum_{j=1}^{s} b_i c_i a_{ij} c_j &= \frac{1}{8}, \\ \sum_{i=1}^{s}\sum_{j=1}^{s} b_i a_{ij} c_j^2 &= \frac{1}{12}, & \sum_{i=1}^{s}\sum_{j=1}^{s}\sum_{k=1}^{s} b_i a_{ij} a_{jk} c_k &= \frac{1}{24}. \end{aligned} \tag{1.47}$$

Conditions (1.47) are given here in the version which is correct for systems of order N. In our case $N = 1$ of one scalar equation it suffices to require the three conditions following from (1.43).

We return now to the initial question: does the vanishing of the first non-zero term in the error expansion of any RK scheme imply that it becomes exact in this case ? The answer is *no*, as the following example shows: consider the family of implicit 2-stage second order schemes depending on a parameter μ,

$$\begin{aligned} U_1 &= u_n \\ U_2 &= u_n + k(\mu f(U_1) + (1-\mu) f(U_2)) \\ u_{n+1} &= u_n + k(f(U_1) + f(U_2))/2 \end{aligned} \tag{1.48}$$

For $\mu = 1/2$ this is the trapezoidal rule. We now choose $\mu = 2/3$ for which we obtain $\mathcal{B}_0 = \mathcal{B}_1 = 0$ and

$$\mathcal{B}_2 = (\frac{1}{2} \cdot 0 \cdot 1 + \frac{1}{2} \cdot \frac{1}{3} \cdot 1 - \frac{1}{6}) f f'^2 + (\frac{1}{2} \cdot \frac{1}{2} \cdot 1 - \frac{1}{6}) f^2 f'' = \frac{1}{12} f^2 f''. \quad (1.49)$$

Hence the scheme is in general second order. The first non-zero term in the error expansion $\mathcal{B}_2$ vanishes if either $f = 0$ or $f'' = 0$ which means $f = Cu + D$ for arbitrary constants C and D. Thus for the differential equation

$$\dot{u} = Cu + D \quad (1.50)$$

the scheme is at least third order. However the next higher order term $\mathcal{B}_3$ in the error expansion does not vanish for this differential equation. We find

$$\mathcal{B}_3 = \frac{1}{72} f f'^3 + \frac{1}{24} f''' f^3 + \frac{1}{12} f^2 f' f'' = \frac{1}{72} f f'^3$$

where we used for the last step that $f'' = 0$ because we forced $\mathcal{B}_2$ to vanish. Hence $\mathcal{B}_3$ does not vanish when $\mathcal{B}_2$ does and thus the 2nd order scheme (1.48) with $\mu = 2/3$ is only third order for eq. (1.50) and not an exact scheme. In the next subsection we will see that there is no constant-coefficient RK scheme which is exact on equation (1.50). There are however variable coefficient schemes for equation (1.50). An example is given in (1.57).

Thus some RK schemes are exact for larger classes of differential equations, others only for the trivial case where f is constant. The vanishing of the first non-zero term in the error expansion by particular choice of the r.h.s. function f does not guarantee exactness as one might have hoped for from the analysis of the classical schemes at the beginning of this section. The approach discussed in the next subsection shows that the question *Given s linearly independent functions, which s-stage RK scheme allows these functions as solutions to be represented exactly?* has the potential to shed new light on the understanding of RK schemes. This approach originates in the technique of **functional fitting** which is of practical importance in numerical analysis and scientific computing.

1.2.2 *Functional Fitting RK-Methods*

Functional fitting RK-methods approach exactness from the solution side. They do not look at the r.h.s. function f to find conditions under which a given scheme becomes exact, but they construct schemes which allow given functions $u(t)$ to be represented exactly. To present this approach, we must consider non-autonomous differential equations and variable-coefficient RK schemes, i.e. schemes whose coefficients a_{ij}, b_i depend on the independent variable t and the step size k. Very recently, Ozawa [31] proved the following results:

Theorem 1.1 *Let $\{c_i\}_{i=1}^s \in I\!R$ be given, $c_i \neq c_j$ for $i \neq j$. Let $\{u_m(t)\}_{m=1}^s \in C^s[t_0, T]$ be linearly independent functions, sufficiently smooth such that each of them satisfies*

$$\dot{u}_m(t + c_i k) = \sum_{j=1}^{s} \frac{(c_i k)^{j-1}}{(j-1)!} u_m^{(j)}(t) + \mathcal{O}(k^s) \tag{1.51}$$

and suppose that they solve in $[t_0, T]$ a homogeneous linear differential equation

$$\sum_{m=0}^{s} p_m(t)\, u^{(m)}(t) = 0 \qquad \text{with} \quad p_s(t) \equiv 1, \quad p_0(t) \neq 0, \tag{1.52}$$

with continuous coefficients $p_m \in C[t_o, T]$. Then the linear system

$$\begin{array}{rcl} u_m(t + c_i k) & = & u_m(t) + k \sum_{j=1}^{s} a_{ij}(t,k)\, \dot{u}_m(t + c_j k) \\ u_m(t + k) & = & u_m(t) + k \sum_{i=1}^{s} b_i(t,k)\, \dot{u}_m(t + c_i k) \end{array} \tag{1.53}$$

is uniquely solvable for $a_{ij}(t,k)$ and $b_i(t,k)$, with $t \in [t_0, T]$ and $0 < k < k_0$ for k_0 small enough.

Proof. The idea of the proof given by Ozawa is the following: For fixed t and k, system (1.53) is a collection of $s+1$ linear systems of order s with matrix

$$(\dot{u}_m(t + c_j k))_{m,j=1,\ldots s} =: \bar{U}(t,k),$$

inhomogeneities

$$\left((\frac{u_m(t + c_i k) - u_m(t)}{k})_m \right)_i \quad \text{and} \quad (\frac{u_m(t+k) - u_m(t)}{k})_m,$$

and s^2+s unknowns $a_{ij}(t,k)$, $b_i(t,k)$. These systems are uniquely solvable if the matrix $\bar{U}(t,k)$ is nonsingular. To prove that $\bar{U}(t,k)$ is nonsingular for all $t \in [t_o, T]$ and small enough k, conditions (1.51) and (1.52) are used. Condition (1.52) ensures that the Wronskian matrix of the linearly independent functions $\{u_1, \ldots, u_s\}$ is nonsingular [12, p. 64ff]. □

Now assume that a function $u \in U := \text{span}\{u_1, ...u_s\}$ satisfies the non-autonomous differential equation

$$\dot{u} = f(t,u), \qquad u(t_0) = u_0, \quad t \in [t_0, T]. \tag{1.54}$$

Then we expect that the RK scheme with coefficients attained according to Theorem 1.1 will be exact on $u(t)$. The surprising result is: from this exactness on the s-dimensional linear space U it follows that the scheme has order s:

Theorem 1.2 *Let the coefficients of the variable-coefficient s-stage RK scheme*

$$\begin{array}{rcl} Y_i & = & y_n + k\sum_{j=1}^{s} a_{ij}(t_n,k) f(t_n + c_j k, Y_j) \\ y_{n+1} & = & y_n + k\sum_{i=1}^{s} b_i(t_n,k) f(t_n + c_i k, Y_i) \\ & & i = 1, \ldots, s \qquad t_n = t_0 + nk, \qquad y_0 = u_0, \end{array} \tag{1.55}$$

be obtained according to Theorem 1.1. Then the order of the scheme is at least s. If the abscissae c_i, $i = 1, \ldots s$ are taken to satisfy

$$\int_0^1 t^{q-1} \prod_{i=1}^{s} (t - c_i) dt = 0, \qquad q = 1, \ldots, \nu, \qquad 1 \le \nu \le s, \tag{1.56}$$

then the order of accuracy is $s + \nu$. The maximum attainable order is $2s$.

Proof. The proof of these statements uses results on RK collocation methods with constant coefficients [12, p. 212] and can be found in Ozawa [31]. If the abscissae c_i satisfy the additional condition (1.56), both the RK collocation scheme and the scheme obtained according to (1.53) have order $s + \nu$. □

We emphasize that the s-stage RK scheme obtained with those linearly independent functions $\{u_1, ...u_s\}$ is exact whenever the solution $u(t) \in U = \text{span}\{u_1, ...u_s\}$. If all solutions of (1.54) happen to belong to U, the scheme is exact on (1.54), no matter how nonlinear f is, because we can first construct the linear combination of the basis functions and afterwards

we replace $\dot{u}$ by $f(t,u)$. It is thus of interest to use functional fitting RK-schemes whenever there is some knowledge about the solution in advance. If one knows that certain low frequencies will be part of the solution, it pays to use a functional fitting RK-scheme which is exact on those frequencies. The remaining part of the solution is still captured by the order of the RK-scheme.

Given an s-stage RK scheme which is exact on $U = \text{span}\{u_1, \ldots u_s\}$, there is a whole family of nonconfluent schemes which depend on s parameters $c_1, \ldots c_s$. All these schemes are exact on the same function space U. Though all these schemes are equivalent when the scheme is used as an exact scheme, they differ in their numerical performance when the scheme is used on a problem where it is not exact. This follows from the second statement of Theorem 1.2.

For constant-coefficient schemes for non-autonomous differential equations, it is a *convention* to satisfy eq. (1.38) when designing new schemes. Because condition (1.38) implies that $t_n + c_i k = y_n + c_i k$ for $u(t) = t$ [6, p. 56]. In the case of Theorem 1.1, condition (1.38) is ensured if $u(t) = t$ is one of the chosen basis functions.

We expect schemes associated to non-autonomous differential equations to have variable coefficients. If we apply the theorem to autonomous differential equations with known solutions, the resulting scheme might have constant or variable coefficients, depending on f. This is illustrated by the following examples.

Example 1.1 We chose $s = 2$, $u_1(t) = t$, $u_2(t) = t^2$ and use the c_i as parameters. Solving the system (1.53) we find the coefficients in the RK-scheme to be

$$\begin{aligned} a_{i1} &= \frac{c_i^2 - 2c_i c_2}{2(c_1 - c_2)}, \qquad a_{i2} = c_i - a_{i1}, \qquad i = 1, 2 \\ b_1 &= \frac{1 - 2c_2}{2(c_1 - c_2)}, \qquad b_2 = 1 - b_1. \end{aligned}$$

For $c_1 = 0$, $c_2 = 1$ we obtain the coefficients of the trapezoidal rule, as expected. We obtain the trapezoidal rule also for $c_1 = 1, c_2 = 0$. For varying c_1, c_2 with $c_1 \neq c_2$ we get a 2–parameter family of nonconfluent RK schemes which are exact on the same family of differential equations on which the trapezoidal rule is exact.

Example 1.2 We now show that there is no constant-coefficient 2-stage RK scheme which is exact on equation (1.50), but there are variable-coefficient schemes for it. The general solution of $\dot{u} = Cu + D,\ u(0) = u_0$ is given by

$$u(t) = (u_0 + \tilde{D}) \exp Ct - \tilde{D}, \qquad \tilde{D} := D/C.$$

A basis function for the 1-dimensional solution space is $u(t) = \exp Ct$. Note that the constant $\tilde{D}$ does not influence the coefficients: it cannot be a basis function for computing RK coefficients since it satisfies the homogeneous differential equation $\dot{u} = 0$ with $p_0(t) = 0$. If we put $u(t) = \tilde{D} + \exp Ct$ we get

$$\tilde{D} + \exp(Ct + c_i k) = \tilde{D} + \exp Ct + k \sum \ldots,$$

and the sum contains derivatives of u and thus no $\tilde{D}$.

To get a 2-stage scheme satisfying eq. (1.38), we chose $u(t) = t$ as second function. With $s = 2,\ c_1 = 0,\ c_2 = 1$ and $u_1(t) = t,\ u_2(t) = \exp Ct$ with $1/C \notin [0, T]$ we obtain the coefficients

$$\begin{aligned} a_{11} &= 0 & a_{12} &= 0 \\ a_{21}(k) &= \frac{1 - (1 - Ck)\exp Ck}{kC(\exp Ck - 1)}, & a_{22}(k) &= \frac{-1 - Ck + \exp Ck}{kC(\exp Ck - 1)} \\ b_1(k) &= \frac{1 - (1 - Ck)\exp Ck}{kC(\exp Ck - 1)}, & b_2(k) &= \frac{-1 - Ck + \exp Ck}{kC(\exp Ck - 1)} \end{aligned} \tag{1.57}$$

These coefficients have an apparent singularity in the limit $k \to 0$. This will be discussed elsewhere.

Example 1.3 With $s = 2,\ u_1(t) = t,\ u_2(t) = 1/t,\ 0 \notin [t_0, T]$ and the c_i as parameters, we obtain the coefficients

$$\begin{aligned} a_{i1}(t,k) &= \frac{\dfrac{1}{t + c_i k} - \dfrac{1}{t} + \dfrac{k c_i}{(t + c_2 k)^2}}{\left(\dfrac{1}{(t + c_2 k)^2} - \dfrac{1}{(t + c_1 k)^2}\right) k}, \quad a_{i2} = c_i - a_{i1}, \quad i = 1, 2 \\ b_1(t,k) &= \frac{\dfrac{1}{t + k} - \dfrac{1}{t} + \dfrac{k}{(t + c_2 k)^2}}{\left(\dfrac{1}{(t + c_2 k)^2} - \dfrac{1}{(t + c_1 k)^2}\right) k}, \quad b_2 = 1 - b_1. \end{aligned} \tag{1.58}$$

Again, we see that the coefficients are undefined in the case $c_1 = c_2$. This example also shows that quite simple functions u_m can lead to complicated coefficients which depend on t and k. Remembering that the lintrap scheme is exact on $u(t) = -1/t$, we think that these complicated formulae might indicate that it is more appropriate to use a Rosenbrock scheme in this case.

Note that we arrived for initial value problems at the point which is standard for the numerical treatment of boundary value problems: We chose some complete system of functions, take the first s of these functions and approximate the Banach space containing the solutions of the given differential equation with this s-dimensional finite space.

1.2.3 *Schemes for Given Differential Equations*

Now we address the second question raised initially: given a differential equation, which schemes are exact for it ? There are many different methods that may lead to exact schemes. They are essentially the same methods as those for finding closed form solutions to differential equations. For a comprehensive classic collection see [15]. Here we discuss only one method which is often applicable and which leads us to exact schemes for polynomial ordinary differential equations. Then we report on how Le Roux used such schemes to obtain nonstandard schemes for parabolic equations with blow-up solutions.

1.2.3.1 *Exact Schemes for Given Differential Equations*

This method starts with a known exact scheme and generates others by transformations. It was applied in [26] to prove the following result for $C = 1$ and $D = 0$.

Lemma 1.7 *Let m be an integer, $m \neq 0, -1$. Assume that the equation*

$$\dot{u} = \frac{1}{mC}(Cu + D)^{m+1}, \qquad u(0) = u_0, \quad Cu_0 + D \neq 0, \tag{1.59}$$

has a solution $u(t)$ such that $Cu(t)+D \neq 0$ in $[0, T)$. Then $u_{n+1} = u(t_{n+1})$, $t_{n+1} = (n + 1)k < T$ is given by

$$\frac{u_{n+1} - u_n}{k} = \frac{(Cu_n + D)^m (Cu_{n+1} + D)^m}{\sum_{j=0}^{m-1} (Cu_{n+1} + D)^j (Cu_n + D)^{m-1-j}}, \quad m > 0, \tag{1.60}$$

$$\frac{u_{n+1}-u_n}{k} = \frac{-1}{\sum_{j=0}^{|m|-1}(Cu_{n+1}+D)^j(Cu_n+D)^{|m|-1-j}}, \quad m<-1. \tag{1.61}$$

Proof. We consider two cases. **Case 1:** Let $m>0$, $Cu_0+D\neq 0$. Then $Cu(t)+D$ is non-zero as long as it exists. With $v:=(Cu+D)^m$ equation (1.59) is equivalent to

$$\frac{dv}{dt}=v^2, \qquad v(0)=(Cu_0+D)^m. \tag{1.62}$$

As discovered independently by many authors, this equation has the exact scheme

$$\frac{v_{n+1}-v_n}{k}=v_nv_{n+1}, \qquad v_0=v(0) \text{ given}. \tag{1.63}$$

This is equivalent to

$$\frac{(Cu_{n+1}+D)^m-(Cu_n+D)^m}{k}=(Cu_n+D)^m(Cu_{n+1}+D)^m, \qquad u_0 \text{ given}, \tag{1.64}$$

where that m-th root has to be taken which produces the correct initial condition. Now we notice that

$$a^q-b^q=\sum_{j=0}^{q-1}(a^{j+1}b^{q-(j+1)}-a^jb^{q-j})=\left(\sum_{j=0}^{q-1}a^jb^{q-1-j}\right)(a-b) \tag{1.65}$$

and that the sum on the r.h.s. is non-zero whenever $a^q-b^q\neq 0$. We get the desired result (1.60) with $q=m$, $a=Cu_{n+1}+D$ and $b=Cu_n+D$ and by division of both sides of eq. (1.64) by the r.h.s. sum of (1.65).
Case 2: Let $m<-1$, $p:=-m>0$, $Cu_0+D\neq 0$. By assumption $Cu(t)+D$ is non-zero in $[0,T)$. With $v:=(Cu+D)^{-p}$ eq. (1.59) is equivalent to

$$\frac{dv}{dt}=v^2, \qquad v(0)=(Cu_0+D)^{-p}. \tag{1.66}$$

As in the previous case, this equation has the exact scheme

$$\frac{v_{n+1}-v_n}{k}=v_nv_{n+1}, \qquad v_0=v(0) \text{ given}. \tag{1.67}$$

This is equivalent to

$$\frac{(Cu_{n+1}+D)^{-p}-(Cu_n+D)^{-p}}{k} = (Cu_n+D)^{-p}(Cu_{n+1}+D)^{-p}, \quad u_0 \text{ given}, \tag{1.68}$$

where that p-th root has to be taken which produces the correct initial condition. Now we multiply both sides by $(Cu_n+D)^p(Cu_{n+1}+D)^p$ and get

$$\frac{(Cu_{n+1}+D)^{p}-(Cu_n+D)^{p}}{k} = -1, \qquad u_0 \text{ given}. \tag{1.69}$$

We apply (1.65) with $q = p = |m|$, and we get the desired result (1.61) after division of both sides of eq. (1.69) by the sum. □

Remark 1.1 The procedure used in the proof can also be used on other equations than (1.59): let $m \in \mathbb{Q}$, for example, as in eq. (1.14); use the same or other substitutions on any differential equation for which an exact scheme is known; etc.

Remark 1.2 For all integers $m > 0$ scheme (1.60) with $C = \alpha$ and $D = 0$ is equivalent to the scheme

$$\frac{1}{m}\left(w_n^{-m}w_{n+1}^{m+1} - w_{n+1}\right) = \alpha k w_{n+1}^{m+1} \tag{1.70}$$

given by Le Roux [20] for the differential equation

$$\dot{w} = \alpha w^{m+1}, \qquad w(0) = w_0 > 0, \tag{1.71}$$

with arbitrary $\alpha \in \mathbb{R}$: multiplication of eq. (1.70) by $m w_n^m w_{n+1}^{-1} k^{-1}$ leads to eq. (1.64) for the $\{w_n\}$. The procedure used by Le Roux for obtaining scheme (1.70) is essentially the same as ours: she put $W := w^{-m}$ and thus transformed eq. (1.71) into

$$\dot{W} = -\alpha m, \qquad W(0) = W_0 = w_0^{-m}, \tag{1.72}$$

applied the exact scheme trivially known for this equation and then did the inverse transformation to obtain the $\{w_n\}$. Le Roux used the exact scheme (1.70) to derive the approximate semi-discrete scheme (1.75). This approach is discussed next.

1.2.3.2 *Nonstandard Schemes for Parabolic Equations with Blow-Up Solutions: Le-Roux Schemes*

Consider the parabolic problem

$$\begin{aligned} v_t - \Delta v^m &= \alpha v^m \qquad \text{for } x \in \Omega \subset I\!R^d,\ t > 0, \\ v(x,t) &= 0 \qquad\quad \text{for } x \in \partial\Omega,\ t > 0, \\ v(x,0) &= v_0(x) > 0 \qquad \text{for } x \in \Omega, \end{aligned} \tag{1.73}$$

where Ω is a smooth bounded domain, $\alpha \geq 0$ real and $m > 1$ an integer. Let (λ_1, u_1) be the principal eigenpair of

$$-\Delta u = \lambda u, \qquad u|_{\partial\Omega} = 0, \tag{1.74}$$

satisfying $u_1(x) > 0$ in Ω and $||u_1||_{L^1(\Omega)} = 1$. Let v_1 be a real smooth function satisfying $v_1^m(x) = u_1(x)$ in Ω. Then $\theta v_1,\ \theta > 0$, is a steady state solution of (1.73) for $\alpha = \lambda_1$. A steady state solution for all α is $v(x,t) \equiv 0$. The time-dependent solutions of (1.73) for given initial function $v_0(x) > 0$ were investigated by Sacks and others, and the results are [20]:

- If $0 \leq \alpha < \lambda_1$ and $v_0 \in L^q(\Omega),\ q > 1$, problem (1.73) has a solution existing for all times and decaying to zero for $t \to \infty$.
- If $\alpha = \lambda_1$ and $v_0 \in L^q(\Omega),\ q > 1$, problem (1.73) has a solution existing for all times and tending to $\theta u_1^{1/m}$ for $t \to \infty$. The factor θ depends on the initial function v_0.
- If $\alpha > \lambda_1$ there exists $T > 0$ such that problem (1.73) has for given v_0 a unique weak solution v on $[0,T]$ with $\lim_{t \to T_-} ||v(\cdot,t)||_{L^\infty(\Omega)} = +\infty$. Such solutions are called **blow-up solutions**. The only nonnegative solution of problem (1.73) which exists for all times is $v(x,t) \equiv 0$.

To construct a numerical scheme whose solution has similar properties as the solution of the continuous problem, Le Roux [20] used the exact scheme (1.70) for $\dot{w} = \alpha w^m$ to derive the approximate semi-discrete scheme

$$\frac{1}{m-1}\left(V_n^{1-m} V_{n+1}^m - V_{n+1}\right) - k\Delta V_{n+1}^m = \alpha k V_{n+1}^m \tag{1.75}$$

for eq. (1.73). Here k is the time step and $V_n = V(x, nk)$ approximates $v(x,t)$ at $t = nk$. Note that solving eq. (1.75) for V_{n+1} with given V_n means solving a nonlinear elliptic boundary value problem with $V_{n+1|\partial\Omega} = 0$, and

this has to be done at each time step. With

$$p := \frac{1}{m}, \quad q := 1 - p, \qquad U_n := V_n^m \tag{1.76}$$

and $U_n \in I\!B := H_0^1(\Omega) \cap C^2(\bar{\Omega})$ where all elements satisfy the given boundary condition, (1.75) becomes

$$\begin{aligned} k\Delta U_{n+1} &= \frac{p}{q}(U_{n+1}U_n^{-q} - U_{n+1}^p) - \alpha k U_{n+1} \\ &=: f(U_{n+1}; U_n), \end{aligned} \tag{1.77}$$

which is a standard quasilinear elliptic problem for U_{n+1}. Le Roux [20] proved existence and uniqueness of the solution of scheme (1.77) for

$$\begin{aligned} t_n = nk < T_1 &:= \frac{p}{\alpha q}||U_0||_\infty^{-q} \\ &= \frac{1}{\alpha(m-1)}||v_0^m||_\infty^{(1-m)/m}, \end{aligned} \tag{1.78}$$

and formulated conditions on Y_0 under which the iterative scheme

$$\Delta Y_{j+1} = \frac{1}{k} f(Y_j; U_n), \qquad Y_0 \text{ given }, \tag{1.79}$$

converges for $j \to \infty$ (monotonic) to U_{n+1}. Le Roux proved stability and convergence of the time discretization for a wide class of initial conditions, gave for fixed $k = \Delta t$ the estimates

$$||U_n||_\infty \leq \begin{cases} c t_n^{-1/q} & \text{if } \alpha < \lambda_1 \\ c(t_n^{-1/q} + ||U_0||_{p+1}) & \text{if } \alpha = \lambda_1 \end{cases}, \tag{1.80}$$

where c is a constant, $c = c(\Omega, p, \alpha, U_0)$, and showed:

- If $\alpha \leq \lambda_1$, then there exists a constant $\Delta t_0 > 0$ depending only on Ω, p, α, U_0 such that the numerical solution U_n exists for all $n \to \infty$ for every time step $\Delta t < \Delta t_0$. From the estimate (1.80) it then follows that $||U_n|| \to 0$ if $\alpha < \lambda_1$, as desired. We also see that the norm of the initial function U_0 specifies the numerical value of θ in the case $\alpha = \lambda_1$.
- If $\alpha > \lambda_1$ then there exists T^* depending on the time step Δt and on U_0 such that the numerical solution U_n exists for $n\Delta t < T^*$ and

becomes infinite at T^*. The following estimate is valid:

$$||U_n||_{p+1} \leq \left(\frac{T^*}{T^* - t_n}\right)^{1/q} ||U_0||_{p+1}, \tag{1.81}$$

and this estimate has also been obtained for the exact solution.

Thus we see that, for sufficiently small $\Delta t < \Delta t_0$, scheme (1.77) produces qualitatively correct numerical solutions which satisfy the estimates known for the exact solutions. In further work this scheme and its mathematical analysis were extended to the more general case

$$\begin{aligned} v_t - \Delta v^{1+\delta} &= \alpha v^p \qquad \text{for } x \in \Omega \subset I\!R^d,\ t > 0, \\ v(x,t) &= 0 \qquad\quad \text{for } x \in \partial\Omega,\ t > 0, \\ v(x,0) &= v_0(x) > 0 \qquad \text{for } x \in \Omega, \end{aligned} \tag{1.82}$$

where $\Omega \subset I\!R^d$ is a smooth bounded domain, δ is a parameter describing diffusion, $\delta \in (-1, 0)$ for fast diffusion and $\delta > 0$ for slow diffusion, $\alpha \geq 0$ real and $p \geq 1 + \delta$. This mathematical work is reviewed in [19].

The usefulness of scheme (1.77) for Computational Plasma Physics is demonstrated in investigations of fusion plasmas, where diffusion equations with slow diffusion (e.g. $\delta = 2$) are used for the density of particles and with fast diffusion (e.g. $\delta = -1/2$) for their temperature. In reference [21] a coupled system for density and temperature of ions is solved for various parameter values, while in reference [22] the two equations are solved separately for various cases (decay of the solution, evolution to a constant profile, explosive case – blow-up).

1.3 Dynamics of Difference Schemes

In this section we first recall basic definitions and facts from the theory of dynamical systems. Then we investigate the performance of various difference methods on the logistic differential equation, which has one stable and one unstable steady state. Some readers might wonder why we also discuss what happens for 'huge' step sizes: knowing the solution of a problem, it is easy to decide which step size is adequate and which one is too large. When solving real-life problems, it is not always clear which step size is adequate. Depending on the structure of the differential system, it is possible that $k = 10^{-5}$ is too large or that $k = 1$ is unnecessarily small. We have to

know what happens when step sizes are too large in order to detect them when dealing with real-life problems. This kind of investigations might help to find additional criteria for deciding which step size is adequate for a given scheme and to single out schemes which allow larger step sizes than others because the discrete dynamics is similar to the continuous dynamics. At the end of the section we discuss the dynamical properties of the lintrap scheme in general. In the next section it will be applied to Hamiltonian systems.

1.3.1 *Continuous Dynamical Systems*

Consider

$$\dot{u} = f(u), \quad u(0) = u_0 \in I\!R^N, \tag{1.83}$$

f continuously differentiable. For such fs eq. (1.83) has a unique solution $u(t; u_0)$ which exists in some maximum interval $[0, T(u_0))$. Recall the following definitions and facts:
$\bar{u}$ is a **stationary state** of (1.83) iff $f(\bar{u}) = 0$ for all $t > 0$.
$\bar{u}$ is a **stable** stationary state of (1.83) iff for any given $\epsilon > 0$ there exists a $\delta > 0$ such that $u(t; u_0) \in U_\epsilon(\bar{u})$ for all $u_0 \in U_\delta(\bar{u})$ and all $t \geq 0$, where $U_\mu(\bar{u}) := \{u \in I\!R^N : |u - \bar{u}| < \mu\}$.
$\bar{u}$ is an **asymptotically stable** stationary state of (1.83) iff $\bar{u}$ is stable and

$$\lim_{t\to\infty} |u(t; u_0) - \bar{u}| = 0 \quad \text{for all} \quad u_0 \in U_\delta(\bar{u}) \qquad \text{for some} \qquad \delta > 0.$$

$\bar{u}$ is a **marginally stable** stationary state of (1.83) iff it is stable but not asymptotically stable.
$\bar{u}$ is a **hyperbolic** stationary state of (1.83) iff $Re\ \mu \neq 0$ for all eigenvalues μ of the Jacobian $f'(\bar{u})$. For hyperbolic stationary states a **Principle of Linearized Stability** is valid, called *Theorem of Hartman-Grobman* by Guckenheimer/Holmes [11, p. 13]:

Theorem 1.3 *Let $\bar{u}$ be a hyperbolic stationary state of (1.83) $\dot{u} = f(u)$. Then there are neighborhoods $U(\bar{u})$ and $V(0)$ such that the dynamics of $\dot{u} = f(u)$ in $U(\bar{u})$ and of $\dot{v} = f'(\bar{u})v$ in $V(0)$ are equivalent, i.e. there is a homeomorphism between $U(\bar{u})$ and $V(0)$ which preserves the sense of orbits and can also be chosen to preserve parameterization by time.*

Remark 1.3 If $\bar{u}$ is a hyperbolic stationary state, it is thus asymptotically stable if $Re\ \mu_i < 0$ for all eigenvalues μ_i of $f'(\bar{u})$, $i = 1, \ldots, N$, and it is unstable if one μ_{i_0} satisfies $Re\ \mu_{i_0} > 0$. If $\bar{u}$ is a non-hyperbolic stationary state, it might be a **bifurcation point** (stationary-stationary or stationary-periodic (**Hopf bifurcation**)). In this case a nonlinear analysis is necessary to decide on the stability of $\bar{u}$ and on the dynamics of (1.83) in a neighborhood of $\bar{u}$.

Example 1.4 The **logistic differential equation**

$$\dot{u} = \lambda u(1-u), \qquad u(0) = u_0 \tag{1.84}$$

has the solution

$$u(t) = \frac{u_0 e^{\lambda t}}{1 + u_0(e^{\lambda t} - 1)} = \frac{u_0}{(1-u_0)e^{-\lambda t} + u_0}. \tag{1.85}$$

It has two stationary states for all λ: $\bar{u} = 0$ and $\hat{u} = 1$. The principle of linearized stability reveals that
$\bar{u} = 0$ is asymptotically stable for $\lambda < 0$ and unstable for $\lambda > 0$,
$\hat{u} = 1$ is unstable for $\lambda < 0$ and asymptotically stable for $\lambda > 0$.
For $\lambda = 0$, every constant is a stationary state. They all are non-hyperbolic.
For $\lambda < 0$, all solutions of (1.84) with $u_0 < 1$ are attracted by $\bar{u} = 0$, and convergence is monotonic; and all $u_0 > 1$ lead to trajectories that grow unbounded in finite time, i.e. to **blow-up solutions**. The *blow-up time* is

$$T(u_0; \lambda) = \frac{1}{-\lambda} \ln \frac{u_0}{u_0 - 1} = \ln \left(\frac{u_0 - 1}{u_0} \right)^{1/\lambda} > 0. \tag{1.86}$$

For $\lambda > 0$, all solutions with $u_0 > 0$ are attracted by $\hat{u} = 1$, and convergence is monotonic; and all $u_0 < 0$ lead to trajectories that tend to $-\infty$ in finite time T. The blow-up time is

$$T(u_0; \lambda) = \frac{1}{\lambda} \ln \frac{u_0 - 1}{u_0} = \ln \left(\frac{u_0 - 1}{u_0} \right)^{1/\lambda} > 0. \tag{1.87}$$

The logistic differential equation (and its name) were introduced by Verhulst in 1838 to model the growth of populations in environments with limited resources. Under certain conditions (no major wars, epidemics (like the plague) or other catastrophes inside the country), it is indeed a very good model. See for instance [14, p. 103], where the values computed for the US population by Pearl and Read in 1920 are compared with census

data for the years 1790 to 1950. No chance to model the development of European populations the same way.

1.3.2 *Discrete Dynamical Systems*

In this section we consider discrete dynamical systems and recall basic definitions and facts used later on. Consider

$$y_{n+1} = g(y_n), \qquad y_0 \in I\!R^N, \tag{1.88}$$

with continuously differentiable g. Such difference equations are explicit and thus uniquely solvable.
$\bar{y}$ is a **fixed point** of (1.88) iff $g(\bar{y}) = \bar{y}$.
$\bar{y}$ is a **periodic point with period** m of (1.88) iff $\bar{y} = g^m(\bar{y})$.
$\bar{y}$ is a **stable** fixed point of (1.88) iff for any given $\epsilon > 0$ there exists a $\delta > 0$ such that $g^n(y_0) \in U_\epsilon(\bar{y})$ for all $y_0 \in U_\delta(\bar{y})$ and all $n \geq 0$.
$\bar{y}$ is an **asymptotically stable fixed point** of (1.88) iff $\bar{y}$ is a stable fixed point of (1.88), and

$$\lim_{n\to\infty} |g^n(y_0) - \bar{y}| = 0 \quad \text{for all} \quad y_0 \in U_\delta(\bar{y}) \qquad \text{for some} \qquad \delta > 0.$$

$\bar{y}$ is an **(asymptotically) stable periodic point** of (1.88) with period m iff $\bar{y},\ g(\bar{y}) \ldots,\ g^{m-1}(\bar{y})$ are (asymptotically) stable fixed points of (1.88).
$\bar{y}$ is a **marginally stable** (periodic) point iff it is stable but not asymptotically stable.
$\bar{y}$ is a **hyperbolic** fixed point of (1.88) iff $|\mu| \neq 1$ for all eigenvalues μ of the Jacobian $g'(\bar{y})$.

By the implicit function theorem, hyperbolic fixed points $\bar{y}$ have a neighborhood $U(\bar{y})$ in which $g - id$ is invertible. If the local inverse is differentiable, it is a diffeomorphism. For hyperbolic fixed points and sufficiently smooth g a **Principle of Linearized Stability** is valid, called *Theorem of Hartman-Grobman* by Guckenheimer/Holmes [11, p. 18]:

Theorem 1.4 *Let $\bar{y}$ be a hyperbolic fixed point of (1.88) $y_{n+1} = g(y_n)$ and let g be a diffeomorphism. Then there are neighborhoods $U(\bar{y})$ and $V(0)$ such that the dynamics of $y_{n+1} = g(y_n)$ in $U(\bar{y})$ and of $v_{n+1} = g'(\bar{y})v_n$ in $V(0)$ are equivalent.*

Remark 1.4 If $\bar{y}$ is hyperbolic, it is thus asymptotically stable if the spectral radius ρ of the Jacobian $g'(\bar{y})$ satisfies $\rho(g'(\bar{u})) < 1$; it is unstable if

$\rho(g'(\bar{y})) > 1$. If $\bar{y}$ is non-hyperbolic, it might be a **bifurcation point** (fixed point – fixed point or fixed point – periodic point (**flip bifurcation**)). In this case, a nonlinear analysis is necessary to decide on the dynamics in a neighborhood of $\bar{y}$.

Example 1.5 The **logistic difference equation**

$$y_{n+1} = \mu y_n(1 - y_n), \qquad y_0 \in [0,1], \qquad 0 < \mu < 4. \tag{1.89}$$

For $\mu \in [0,4]$, all iterates lie in the interval $[0,1]$ if y_0 does.
$v_1 = 0$ is a fixed point for all $\mu > 0$. For $0 < \mu < 1$ it is the only fixed point in $[0,1]$ and asymptotically stable. For $\mu > 1 =: a_1$ it is unstable.
For $\mu = 1$ there is a bifurcation with exchange of stability. A second branch of fixed points, $v_2(\mu)$, appears in the interval $[0,1]$:
$v_2(\mu) = \mu - 1/\mu \in [0,1]$ for $\mu \geq 1$. $v_2(\mu)$ is unstable for $\mu < 1$ and asymptotically stable for $1 < \mu < 3$. For $1 < \mu < 2$ convergence to v_2 is monotonic, for $2 < \mu < 3 =: a_2$ it is a damped oscillation.
In $\mu = 3$ this branch of fixed points loses stability in a flip bifurcation:
for $3 < \mu < 1 + \sqrt{6} =: a_3$ there is an asymptotically stable 2-cycle $v_3 = g_\mu(v_4)$, $v_4 = g_\mu(v_3)$. For $\mu = a_3$ there is another flip bifurcation to a 4-cycle. This 4-cycle is asymptotically stable for $a_3 < \mu < a_4$, etc.
The sequence of period-doubling bifurcations accumulates in $a_\infty \approx 3.5699\ldots$ with an aperiodic solution.

$$\lim_{n\to\infty} \frac{a_n - a_{n-1}}{a_{n+1} - a_n} =: \delta \approx 4.669\ldots \tag{1.90}$$

is the **Feigenbaum** constant. For $\mu > a_\infty$ periods other than powers of 2 are possible; first even periods, then also odd periods. For $\mu = 1 + \sqrt{8}$ period 3 occurs. For $\mu \geq 1 + \sqrt{8}$ all periods m are possible, and the iterates are **chaotic** in the sense of Li and Yorke [25; 11]. For $\mu > 4$, part of the iterates leave the interval $[0,1]$ and converge to $-\infty$.

1.3.3 *Forward Euler Scheme*

We discretize (1.84) $\dot{u} = \lambda u(1-u)$, $u(0) = u_0$ by Euler's method with fixed time step k and get

$$\begin{aligned} y_{n+1} &= y_n + \lambda\, k\, y_n\, (1 - y_n), \qquad y_0 = u(0), \\ &= F_k(y_n). \end{aligned} \tag{1.91}$$

Fig. 1.1 *Mapping properties of forward Euler in the* $(\lambda k, y_n)$ *plane* [10, Fig. 1].
Figure **a** on the left shows to where y_n is mapped after one iteration with scheme (1.91). The limiting curves are $y_n = 0;\ y_n = 1;\ y_n = 1/(\lambda k),\ y_n = 1 + 1/(\lambda k)$.
Figure **b** on the right shows to where it is mapped after two iterations. The additional borders are given by $y_n = \left(1 + \lambda k \pm \sqrt{(-1+\lambda k)(3+\lambda k)}\right)/(2\lambda k)$.
Qualitatively correct trajectories are obtained for $k\lambda y_n < 1$ and $k\lambda < 1$. Oscillations occur for $k\lambda > 1$ and $0 < y_n < 1 + 1/(k\lambda)$. For larger y_n the iterates tend to $-\infty$.

The fixed points of (1.91) satisfy $\lambda k y(1-y) = 0$ and are thus the same as for the continuous case, $\bar{y} = \bar{u} = 0$ and $\hat{y} = \hat{u} = 1$ for all λk. The Jacobian is

$$F_k'(y_n) = 1 + \lambda k - 2\lambda k y_n. \tag{1.92}$$

Let $\lambda = 1$ for the following analysis. The analysis for arbitrary $\lambda > 0$ only requires a rescaling of k. The analysis for $\lambda < 0$ is also similar, but $\bar{y}$ and $\hat{y}$ then exchange their roles.

For $\bar{y} = 0$ we get $F_k'(0) = 1 + k > 1$. Thus $\bar{y} = 0$ is unstable for all k, as is $\bar{u} = 0$. For $\hat{y} = 1$ we get $F_k'(1) = 1 - k$ and we have to consider several different cases:

Case 1: For $0 < k < 1$ we get $0 < F_k'(1) < 1$ and $\hat{y} = 1$ is stable. Now we consider the trajectories:

If $y_0 < 0$, the iterates tend monotonically to $-\infty$ for $n \to \infty$. Though the continuous solution exists only for $t < T$ as given by (1.87), the discrete iterates exist for all $t_n = nk,\ n \to \infty$. This was already noticed by Dahlquist in 1959 [32].
If $0 < y_0 < 1$, all trajectories $\{y_n\}_{n \in I\!N}$ grow monotonically to $\hat{y} = 1$ and thus behave qualitatively correctly.
If $y_0 > 1$, the qualitative behavior of the iterates depends both on y_0 and on k. For all (y_0, k)-values above the curve $y_0 = 1 + \frac{1}{k}$, the iterates "overshoot" and the trajectories tend to $-\infty$ for $n \to \infty$. For all (y_0, k)-values satisfying $\frac{1}{k} < y_0 < 1 + \frac{1}{k}$, the iterates enter the region $0 < y < 1$ and continue monotonically towards 1. But $y_0 = 1$ does have a neighborhood in which the discrete trajectories tend monotonically to $\hat{y} = 1$ and thus behave qualitatively like the continuous trajectories (see Fig. 1.1).
Case 2: For $1 < k < 2$ we get $-1 < F_k'(1) < 0$. Thus $\hat{y} = 1$ is stable, but the iterates oscillate in all neighborhoods of $\hat{y} = 1$. Hence $\hat{y} = 1$ does not have a neighborhood where trajectories behave qualitatively correctly. But they still converge to the correct limit for certain initial values. The dependence of the limit on the initial value y_0 and on the step size k is illustrated in Fig. 1.1. The curves were computed using MATHEMATICA.
Case 3: For $k > 2$ we find $|F_k'(1)| > 1$, and $\hat{y} = 1$ is unstable. For $k = 2$ there is a flip bifurcation to the 2-cycle

$$\bar{y}_{3,4} = \frac{k + 2 \pm \sqrt{k^2 - 4}}{2k} \in I\!R, \tag{1.93}$$

which is stable for $2 < k < \sqrt{6}$. What happens for larger k can best be seen from the map [36]

$$v_n = \frac{k}{1+k}\ y_n, \tag{1.94}$$

which is a homeomorphism for $k > 0$ and maps the discrete dynamical system defined by the iteration

$$y_{n+1} = y_n + k y_n (1 - y_n) \tag{1.95}$$

to the discrete dynamical system with iteration

$$v_{n+1} = (1 + k)\, v_n (1 - v_n). \tag{1.96}$$

This is the logistic map (1.89) with $\mu = 1 + k$. The discretization parameter k can thus produce all the peculiar behavior which is known for the logistic

map, and which was briefly described in section 1.3.2. For $k > \sqrt{8}$ we get chaotic trajectories. A Feigenbaum diagram of (1.95) was shown several times, see for instance [17, Fig. 3] and [37, Fig. 3.5].

Note that the homeomorphism (1.94) *must* break down for $k = 0$: the fixed points 0 and 1 of (1.95) are different from each other for all k, but the fixed points 0 and $\frac{k}{1+k}$ of (1.96) meet in a bifurcation point for $k = 0$.

Summary: With the forward Euler scheme, the discrete analog of the unstable stationary state is an unstable fixed point for all λk. The discrete analog of the stable stationary state is a stable fixed point for $-2 < \lambda k < 2$. For $\lambda > 0$, it turns unstable in a flip bifurcation at $\lambda k = 2$. This flip bifurcation is the beginning of a Feigenbaum cascade of period-doubling bifurcations. Already for $\lambda k > 1$ the discrete scheme is a very poor model: there is no neighborhood of the stable fixed point with correct dynamic behavior. For $0 < \lambda k < 1$ such a neighborhood exists. It depends on the initial value y_0 and on the step size k whether the dynamic behavior of the discrete solution is qualitatively correct.

It should be noted that the curves separating the different regimes for the initial values y_0 and shown in Fig. 1.1 are either branches of fixed points or closely related to the branches of spurious fixed points for the midpoint Euler scheme to be discussed next.

1.3.4 *Midpoint Euler Scheme*

For smooth one-step methods Beyn [3] proved the following

Theorem 1.5 *Let $\Omega \subset \mathbb{R}^N$ be compact and assume that*

$$\dot{u} = f(u), \qquad u(0) = u_0 \in \mathbb{R}^N \tag{1.97}$$

has finitely many stationary solutions v_i, $i = 1, \ldots, M$ in the interior of Ω, and that all v_i are regular, i.e. $f'(v_i)$ is invertible for $i = 1, \ldots, M$. Let ϕ be a smooth one-step method of order $p \geq 1$. Then there exists a $k_0 > 0$ such that the discrete system

$$y_{n+1} = \phi(k, y_n), \tag{1.98}$$

$k \leq k_0$, has exactly M fixed points $v_i(k)$, $i = 1, \ldots, M$ in Ω, and these satisfy

$$v_i(k) = v_i + \mathcal{O}(k^p), \qquad i = 1, \ldots, M. \tag{1.99}$$

Moreover, if $Re\ \mu > 0$ *for some eigenvalue* μ *of* $f'(v_i)$*, then* $v_i(k)$ *is an unstable fixed point of (1.98); and if* $Re\ \mu < 0$ *for all eigenvalues* μ *of* $f'(v_i)$ *then it is an asymptotically stable fixed point.*

For Runge-Kutta schemes, (1.99) is too pessimistic: RK schemes exactly reproduce all stationary states of the differential equation, i.e. $v_i(k) = v_i$ [3; 16], but they often add some spurious fixed points. Bifurcation points between branches of proper fixed points and branches of spurious fixed points were characterized by Iserles et al. [16]. We shall apply these results to the scheme

$$\begin{aligned} Y_1 &= y_n, \\ Y_2 &= y_n + \frac{k}{2} f(Y_1), \\ y_{n+1} &= y_n + k f(Y_2) \end{aligned} \tag{1.100}$$

for the logistic differential equation (1.84). It is a Runge-Kutta scheme sometimes called 'midpoint Euler scheme' since it is derived by using the midpoint rule (or first Gauss formula) for integration [12, Chap. II, (1.4)]. Another formulation of (1.100) is

$$\begin{aligned} y_{n+1} &= y_n + k f(y_n + \frac{k}{2} f(y_n)) \\ &=: F_k(y_n). \end{aligned} \tag{1.101}$$

Since $f(u) = \lambda u(1-u)$ is a polynomial of 2nd order, $F_k(y_n)$ is a polynomial of 4th order, namely

$$F_k(y_n) = y_n + k\lambda y_n(1-y_n)(2+k\lambda(1-y_n))(2-k\lambda y_n)/4. \tag{1.102}$$

The equation $F_k(y) - y = 0$ thus always has four complex solutions. These turn out to be real for all k. They are [10]

$$0, \quad \frac{2}{\lambda k}, \quad 1, \quad 1 + \frac{2}{\lambda k}. \tag{1.103}$$

The **spurious fixed points** $\frac{2}{\lambda k}$, $1 + \frac{2}{\lambda k}$ are unbounded for $k \to 0$. Both of them converge to the proper fixed points 0, 1 for $k \to \infty$. This might indicate that the implicitness of scheme (1.101) with (1.102) is not optimal in the sense of [26]. Note also the connection between these spurious fixed points and the spurious curves governing the convergence for forward Euler (Fig. 1.1a): the factor 2 is due to the factor $\frac{k}{2}$ in the middle line of (1.100).

Fig. 1.2 *Feigenbaum diagram for midpoint Euler* in the $(y, \lambda k)$-plane, $\lambda = 1$ [10, Fig. 2]. The 200th to 700th iterates are shown as obtained for two initial values y_0 per k. The unstable fixed points are not seen.

Applying the principle of linearized stability in the case $\lambda = 1$ gives:
$\bar{y} = 0$ is unstable for all k.
$\hat{y} = 1$ is stable for $0 < k < 2$, and convergence is monotonic for $0 < y_0 < \frac{2}{k}$.
$\hat{y}$ loses its stability to $\bar{y}_3 = \frac{2}{k}$ in a bifurcation point [17]: the two stationary states $\hat{y}(k) \equiv 1$ and $\bar{y}_3(k) = \frac{2}{k}$ meet for $k = 2$ and exchange stability there.
$\bar{y}_3 = \frac{2}{k}$ is stable for $2 < k < 1 + \sqrt{5} \approx 3.24$ and loses stability to a Feigenbaum cascade of period-doubling bifurcations.
$\bar{y}_4 = 1 + \frac{2}{k}$ is stable for $0 < k < -1 + \sqrt{5} \approx 1.24$ and loses stability to a Feigenbaum cascade of period-doubling bifurcations.

This example demonstrates a close relationship between the size of the compact domain Ω and the step size k_0 in Beyn's theorem: if we choose $\Omega = [-\omega, 1 + \epsilon]$, then $k_0 < 2/(1 + \epsilon)$, in order to exclude the spurious unstable fixed point $\bar{y}_3(k) = \frac{2}{k}$. Thus for small $\epsilon > 0$, k_0 is nearly given by the stability limit of the method. If we choose $\Omega = [-\omega, 3]$, then $k_0 < \frac{2}{3}$. Figure 1.2 shows the stable fixed points of (1.101) with $f(y) = \lambda y(1 - y)$, and their transition to chaos. It can also be found in [37, Fig. 3.10], its lower part is given in [17, Fig. 4].

Fig. 1.3 *Trajectories of midpoint Euler* in the $(y, \lambda k)$-plane, $\lambda = 1$, [9, Figs. 3.6ff]. Solutions of (1.102) are shown for $k\lambda = 0.8,\ 1.5,\ 1.9,\ 2.5$ and two initial values for each $k\lambda$, satisfying $y_{o,1} < \bar{y}_3 = 2/(k\lambda) < y_{o,2}$. The unstable fixed point $\bar{y}_3$ separates the domains of attraction of the different trajectories. The stability behavior of the scheme depends both on the value of $k\lambda$ and of y_0.

Figure 1.3 comments on Figure 1.2. Part *a* ($k\lambda = 0.8$): For $y_0 = 2.49 < 2/k = 2.5$ the trajectory converges to the proper fixed point $\hat{y} = 1$. For $y_0 = 2.51 > 2/k$ it converges to the spurious stable fixed point $\bar{y}_4 = 1 + 2/k = 3.5$.
Part *b* ($k\lambda = 1.5 > \sqrt{5} - 1 \approx 1.24$): For $y_0 = 1.32 < 2/k = 4/3$, the trajectory converges to the proper fixed point $\hat{y} = 1$. For $y_0 > 2/k$, it converges to the stable spurious solution of period 4.
Part *c* ($k\lambda = 1.9$): For $y_0 = 1.4 > 2/k$, the iterates first wander in the chaotic regime of the spurious stable branch, then they enter the domain of attraction of $\hat{y} = 1$. For $y_0 = 0.9$, they converge monotonically to $\hat{y} = 1$.
Part *d* ($k\lambda = 2.5$): For both initial values $y_0 = 0.5$ and $y_0 = 1.4$ the trajectory converges to the stable spurious solution $\bar{y} = 0.8$.

Spurious fixed points are very unwelcome. In computations with fixed k at least two runs with different k are required in order to detect their k-dependence and thus the fact that they are spurious. Also, they distort the dynamics considerably: unstable spurious fixed points diminish domains of attraction by being an additional 'continental water divide'. Stable spurious fixed points attract trajectories that should go somewhere else.

In today's standard, Runge-Kutta methods are used as variable order, variable step size schemes because these allow to keep the local error uniformly small and thus optimize the time stepping. They are thus more efficient than fixed step-size fixed order schemes. In addition, they will destroy spurious fixed points present in fixed step-size fixed order Runge-Kutta schemes.

Summary: With the midpoint Euler scheme, the discrete analog of the unstable stationary state is an unstable fixed point for all λk. The discrete analog of the stable stationary state has a neighborhood with correct dynamic behavior for $-2 < \lambda k < 2$. With $\lambda > 0$, it loses its stability for $\lambda k = 2$ through an exchange of stability with a branch of unstable spurious fixed points. There is another spurious branch of stable fixed points. Both branches of spurious fixed points lose stability to a Feigenbaum cascade of period-doubling bifurcations independently of each other (Fig. 1.2).
This time, the difference equation is a good model up to $\lambda k = 2$, but only in a small domain Ω owing to the spurious fixed points. As a consequence of Beyn's theorem [3], both branches of spurious fixed points become unbounded for $\lambda k \to 0$.

1.3.5 *Linearly Implicit Euler Schemes*

For our model problem (1.84), Twizell et. al. [36] introduced the following schemes:

$$y_{n+1} = y_n + \lambda k y_{n+1}(1 - y_n) \tag{1.104}$$

and

$$y_{n+1} = y_n + \lambda k y_n(1 - y_{n+1}). \tag{1.105}$$

Written explicitly, they read

$$y_{n+1} = \frac{y_n}{1 - \lambda k(1 - y_n)} =: g_0(y_n; \lambda k) \tag{1.106}$$

and

$$y_{n+1} = \frac{(1+\lambda k)y_n}{1+\lambda k y_n} =: g_1(y_n;\lambda k). \tag{1.107}$$

They are related to each other in an obvious way: each of them treats one of the two stationary states of eq. (1.84) implicitly and the other one explicitly. They are **adjoint** to each other according to the definition discussed in section 1.3.6.2: the map (1.130) replaces scheme (1.104) by scheme (1.105), and scheme (1.105) by scheme (1.104). Schemes adjoint to each other always have the same order of accuracy, and the same global error expansion, with k replaced by $-k$.

As was shown in [27] by induction, the difference equation (1.106) has the solution

$$y_n = \frac{y_0}{(1-\lambda k)^n(1-y_0)+y_0}, \tag{1.108}$$

and the difference equation (1.107) has the solution

$$y_n = \frac{(1+\lambda k)^n y_0}{1+y_0((1+\lambda k)^n-1)}. \tag{1.109}$$

The iterates of scheme (1.106) are defined as long as the denominator of eq. (1.106) is non-zero, i.e. as long as condition

$$(1-k\lambda)^n \neq \frac{y_0}{y_0-1} \tag{1.110}$$

is satisfied. Similarly, the iterates of scheme (1.107) exist as long as

$$(1+k\lambda)^n \neq \frac{y_0-1}{y_0}.$$

Both (1.108) and (1.109) are approximations to (1.85), with $e^{\pm\lambda k}$ replaced by the first two terms of their Taylor expansion. $1\pm\lambda k$ is a qualitatively correct approximation to $e^{\pm\lambda k}$ for those λk for which $1\pm\lambda k > 0$, i.e. for $\pm\lambda k > -1$.

Schemes (1.106) and (1.107) were investigated in detail in [10]. Here we review the results for scheme (1.106). Details can be found in the next subsection.

For $\lambda < 0$, $\bar{y} = 0$ is stable for all k, and $\hat{y} = 1$ is unstable for all k. Trajectories with initial value $y_0 \leq 1$ behave qualitatively correctly for all

k. Trajectories with initial value $y_0 > 1$ behave qualitatively correctly as long as the blow-up time T has not passed, i.e. as long as

$$t_{_N} := \sum_{n=1}^{N} nk < T(y_0) = \ln \frac{y_0 - 1}{y_0}. \tag{1.111}$$

For $\lambda > 0$ and $\lambda k < 2$, $\bar{y} = 0$ is unstable and $\hat{y} = 1$ is stable. Trajectories with arbitrary initial value y_0 behave qualitatively correctly for $\lambda k < 1$ (as long as the blow-up time has not passed in the blow-up case). For $1 < \lambda k < 2$, convergence to the correct limit is oscillatory. For $\lambda k > 2$, in contrast to the continuous case, $\bar{y} = 0$ is *stable* and $\hat{y} = 1$ is *unstable*, and the spuriously stable fixed point $\bar{y} = 0$ is globally attracting. Blow-up cannot be seen any more. Hence *the whole dynamics is incorrect* for $\lambda k > 2$, *but 'looks perfectly alright' if there is no pre-knowledge of the behavior of the trajectories* and if blow-up solutions are not expected.

If the discrete image of an unstable steady state of the differential equation is a stable fixed point, we call the difference scheme **superstable**. Superstability will be further discussed below.

That the dynamics is incorrect is much harder to detect for this scheme on a 'real life problem' than for the other schemes investigated here: in the previous two cases, the stable spurious solutions are k-dependent (see (1.93), (1.103)) and thus disappear in computations with step-size control or when computations are repeated with different step size k. In the present case, substantially smaller step sizes or a different scheme must be used for detection of the failure of the method.

As far as scheme (1.107) is concerned, everything is very similar. Scheme (1.107) converges monotonically to the correct fixed points for $\lambda k > -1$ and all initial values y_0, as long as the blow-up time has not passed. For $-2 < \lambda k < -1$ the stable fixed point $\bar{y} = 0$ has a neighborhood of oscillating convergence. For $\lambda k < -2$ the scheme is unstable in a neighborhood of the fixed point $\bar{y} = 0$ and it is superstable in a neighborhood of $\hat{y} = 1$ which is globally attracting now. The stability of both fixed points thus disagrees with the stability of the stationary states of the differential equation, blow-up is disguised, and the whole dynamics is incorrect.

1.3.5.1 *Details of the Dynamics*

We now consider scheme (1.106) in detail, first for $\lambda < 0$ and then for $\lambda > 0$.

Lemma 1.8 *Let $\lambda < 0$. Then $\bar{y} = 0$ is a stable fixed point of scheme (1.106) for all k and $\hat{y} = 1$ is an unstable fixed point of (1.106) for all k.*

Proof. From (1.106) we get

$$g_0'(y_n; \lambda k) = \frac{1 - \lambda k}{(1 - \lambda k + \lambda k y_n)^2} \tag{1.112}$$

and thus

$$0 < g_0'(0; \lambda k) = \frac{1}{1 - \lambda k} < 1 \quad \text{for all} \quad \lambda k < 0 \tag{1.113}$$

and

$$g_0'(1; \lambda k) = 1 - \lambda k > 1 \qquad \text{for all } \lambda k < 0. \tag{1.114}$$

□

We now discuss the behavior of the trajectories for given initial value y_0.
Case 1: Trajectories with initial value $y_0 < 1$ converge monotonically to $\bar{y} = 0$:

$$y_n < 1 \Rightarrow 1 - y_n > 0 \Rightarrow -\lambda k(1 - y_n) > 0 \Rightarrow 1 - \lambda k(1 - y_n) > 1 \Rightarrow$$

$$|y_{n+1}| = \frac{|y_n|}{1 - \lambda k(1 - y_n)} < |y_n|. \tag{1.115}$$

Case 2: For $y_0 > 1$, the qualitatively correct behavior of the iterates depends on the size of $|\lambda k|$ and of the iteration index n: If $-\lambda k > 0$ is small enough, it follows that $0 < 1 - \lambda k(1 - y_0) < 1$, and thus $y_1 > y_0 > 1$. For all λk and n with $0 < 1 - \lambda k(1 - y_n) < 1$ we thus get $y_{n+1} > y_n$ and $1 - \lambda k(1 - y_n) > 1 - \lambda k(1 - y_{n+1})$. For computations with fixed step size k, either there is an N with

$$1 - \lambda k(1 - y_N) > 0 \quad \text{and} \quad 1 - \lambda k(1 - y_{N+1}) < 0, \tag{1.116}$$

or it happens that

$$1 - \lambda k(1 - y_N) = 0. \tag{1.117}$$

In the case of eq. (1.117), the iteration comes to a stop, blow-up has happened. In the case of (1.116), the denominator changes sign without vanishing. The following iterates are negative and approach $\bar{y} = 0$ from below. This is a discrete analog of "a rational function passes a pole and returns from $-\infty$". In the case considered here, however, iteration for

$n > N$ does not make sense. The iterates do not approximate the solution $u(t; y_0)$ of the differential equation anymore. They do approximate the solution $u(t; y_{N+2})$ with initial value $y_{N+2} < 0$ for $n \geq N + 2$.

Lemma 1.9 *Let $\lambda > 0$. Then $\bar{y} = 0$ is unstable and $\hat{y} = 1$ is stable for $\lambda k < 2$. For $\lambda k > 2$, both fixed points of (1.106) show incorrect stability properties: $\bar{y} = 0$ is stable and $\hat{y} = 1$ is unstable.*

Proof. From (1.112) we get

$$g_0'(0; \lambda k) = \frac{1}{1 - \lambda k},$$

and this satisfies
$g_0'(0; \lambda k) > 1$ for $0 < \lambda k < 1$,
$g_0'(0; \lambda k) < -1$ for $1 < \lambda k < 2$, and
$g_0'(0; \lambda k) > -1$ for $\lambda k > 2$.
Note that $g_0'(0; \lambda k)$ is singular for $\lambda k = 1$.
For $\hat{y} = 1$ we get

$$g_0'(1; \lambda k) = 1 - \lambda k \tag{1.118}$$

and this satisfies
$0 < g_0'(1; \lambda k) < 1$ for $0 < \lambda k < 1$,
$-1 < g_0'(1; \lambda k) < 0$ for $1 < \lambda k < 2$,
$g_0'(1; \lambda k) < -1$ for $2 < \lambda k$. □

We now discuss the trajectories for given initial value y_0. We show that convergence is monotonic for $0 < \lambda k < 1$ and *all initial values* y_0, as long as the blow-up time has not passed. This readily follows by using formula (1.106)

$$y_{n+1} = \frac{y_n}{1 - \lambda k(1 - y_n)} \tag{1.119}$$

and

$$1 - y_{n+1} = \frac{(1 - \lambda k)(1 - y_n)}{1 - \lambda k(1 - y_n)}. \tag{1.120}$$

Case 1: Let $0 < y_n < 1$. Then $y_n < y_{n+1} < 1$:
$0 < 1 - y_n < 1 \Rightarrow 0 < 1 - \lambda k(1 - y_n) < 1$ and $y_{n+1} > y_n$ from (1.119).
From (1.120) it follows that $y_{n+1} < 1$.
Case 2: Let $1 < y_n$. Then $1 < y_{n+1} < y_n$:
$y_n > 1 \Rightarrow 1 - \lambda k(1 - y_n) > 1 \Rightarrow y_{n+1} < y_n$ by using eq. (1.119).

From (1.120) we now get $1 - y_{n+1} < 0$. Thus the stable fixed point $\hat{y} = 1$ attracts all trajectories with initial value $y_0 > 0$.
Case 3: If $y_n < 0$, it follows from (1.119) and (1.120) that
$y_{n+1} < y_n$ if $1 - \lambda k(1 - y_n) > 0$ and $y_{n+1} > 0$ if $1 - \lambda k(1 - y_n) < 0$.
What has been said earlier about approximation of blow-up solutions applies here analogously.

1.3.5.2 *Superstability*

Superstability was discussed by Lindberg in 1974 ('a dangerous property of methods for stiff differential equations'), by Dahlquist et al (1982) and by Dieci and Estep [5, and references therein]. Dieci et al (1991) used the following working definition: 'Superstability is that situation in which a numerical integrator does not detect that the underlying solution is physically unstable'. They report that

- superstability was observed in the adaptive integration of stiff initial value problems and in the integration of parabolic PDEs by the method of lines;
- it is known how to avoid it for Riccati equations (use information on the eigenvalues of the differential equation in the step size control), but
- in general, it is not known how to avoid this phenomenon.

In our case here, it is easy to avoid superstability and instability simultaneously by using the **self-adjoint** scheme related to (1.104) and (1.105): we obtain it by adding the two adjoint schemes (1.104) and (1.105). It has the additional advantage of being second order accurate, as shown in section 1.3.6.2. We get

$$\frac{y_{n+1} - y_n}{k} = \frac{\lambda}{2}(y_n(1 - y_{n+1}) + y_{n+1}(1 - y_n)). \tag{1.121}$$

As was shown in [27], an alternate way of obtaining scheme (1.121) is to apply the lintrap method (1.16) to the logistic differential equation (1.84).

The difference scheme (1.121) gives qualitatively correct approximations for $|\lambda k| < 2$. It produces oscillatory trajectories and thus incorrect dynamic behavior for $|\lambda k| > 2$, but the stability of both fixed points is correct for all λk. This follows from the dynamical properties of scheme (1.16) for general functions f. These are discussed next.

1.3.6 *The Linearly Implicit Lintrap Scheme*

In several case studies, so-called 'unconventional' or 'non-standard' difference schemes were 'custom-tailored' for individual differential equations and were shown to preserve qualitative properties of the differential equation: some schemes are exact, symplectic or feature discrete blow-up [29; 34; 27]. It turned out [27] that several of these ad-hoc schemes can be obtained by applying the scheme

$$\frac{\mathbf{y}_{n+1}-\mathbf{y}_n}{k} = \mathbf{f}(\mathbf{y}_n) + \mathbf{f}'(\mathbf{y}_n)\frac{\mathbf{y}_{n+1}-\mathbf{y}_n}{2}, \qquad \mathbf{y}_0 = \mathbf{u}_0 \in I\!R^N, \tag{1.122}$$

to the differential system

$$\dot{\mathbf{u}} = \mathbf{f}(\mathbf{u}), \qquad \mathbf{u}(0) = \mathbf{u}_0. \tag{1.123}$$

We call (1.122) the *linearized trapezoidal rule* or the *lintrap scheme*. It can be obtained by applying one Newton step to the term $(\mathbf{f}(\mathbf{y}_n)+\mathbf{f}(\mathbf{y}_{n+1}))/2$ in the trapezoidal rule or to $\mathbf{f}((\mathbf{y}_n+\mathbf{y}_{n+1})/2)$ in the implicit midpoint rule. It has been used on systems of partial differential equations in Computational Fluid Dynamics [2] and Plasma Physics [18; 7]. It can also be written as

$$\begin{aligned} \mathbf{y}_{n+1} &= \mathbf{y}_n + k\mathbf{K}_1, \qquad \mathbf{y}_0 = \mathbf{u}_0, \\ (I - \frac{k}{2}\mathbf{f}'(\mathbf{y}_n))\mathbf{K}_1 &= \mathbf{f}(\mathbf{y}_n), \end{aligned} \tag{1.124}$$

and is a member of the family of Rosenbrock-Wanner (ROW) schemes which are standard in the numerical treatment of systems of stiff ordinary differential equations and of differential-algebraic equations [13; 24; 33]. Here I is the identity matrix, k the time step and $\mathbf{f}'(\mathbf{y}_n)$ the Jacobian of system (1.123), evaluated at the n-th iterate.

The scheme is linearly implicit. If

$$A(k,\mathbf{y}) := (I - \frac{k}{2}\mathbf{f}'(\mathbf{y})) \tag{1.125}$$

is non-singular for all $k \in I\!R^+$ and all $\mathbf{y} \in I\!R^N$, all iterates exist. A point $\bar{\mathbf{u}} \in I\!R^N$ with non-singular $A(k,\bar{\mathbf{u}})$ is fixed point of (1.124) iff $\mathbf{f}(\bar{\mathbf{u}}) = 0$, i.e. iff it is a steady state of (1.123). There are no spurious fixed points.

If the function $\mathbf{f}$ is such that the Jacobian $\mathbf{f}'(\mathbf{y})$ is negative definite for all $\mathbf{y} \in I\!R^N$, then $A(k,\mathbf{y})$ is non-singular for all k and all $\mathbf{y}$. If the function $\mathbf{f}$ does not have this property, $A(k,\mathbf{y}_n)$ turns singular for values of k that match eigenvalues of $\mathbf{f}'(\mathbf{y}_n)$. The iteration given by (1.124) is thus

undefined for certain combinations of step-size and initial value. This is a discrete version of blow-up, as already discussed following eq. (1.117).

In special cases, the a priori bound for the blow-up time obtained from the singularity of $A(k, \mathbf{y}_0)$ is extremely good [28]. This is due to the fact that the scheme is exact for functions f satisfying (1.18), and then, of course, the error is small for functions which are close to satisfying eq. (1.18), in the norm of some appropriate Banach space. In the general case, the accuracy of the scheme guarantees a good approximation 'for sufficiently small $k \leq k_0$'. In addition to that, there is no relation between the continuous and the discrete blow-up time in general. This is shown by the following theorem and examples.

1.3.6.1 *Existence Intervals*

We now compare the size of the bounding time step in the n-th iterate with the time interval for which existence of the solution to the differential equation can be ensured:

Theorem 1.6 *Let $B(\mathbf{y}_n) := \{\mathbf{u} \in I\!R^N : |\mathbf{u} - \mathbf{y}_n| \leq b\}$, and assume that $\mathbf{f} : I\!R^N \to I\!R^N$ is continuously differentiable with $|\mathbf{f}(\mathbf{u})| \leq C$ and $|\mathbf{f}'(\mathbf{u})| \leq L$ in the closed domain B. Then the differential system*

$$\dot{\mathbf{u}} = \mathbf{f}(\mathbf{u}), \qquad \mathbf{u}(0) = \mathbf{y}_n, \tag{1.126}$$

has a unique solution $\mathbf{u}(t) \in B$ which is defined for $t \leq t_b := b/C$.
Under the same assumptions, the linear system

$$\begin{aligned} (I - k\theta \mathbf{f}'(\mathbf{y}_n))\mathbf{K}_{n+1} &= \mathbf{f}(\mathbf{y}_n), \\ \mathbf{y}_{n+1} &= \mathbf{y}_n + k\mathbf{K}_{n+1}, \end{aligned} \tag{1.127}$$

has a unique solution $\mathbf{y}_{n+1}(\theta, k)$ for all (θ, k) with $\theta k < k_b := 1/L$ and $0 \leq \theta \leq 1$. If $\mathbf{f}'(\mathbf{y}_n)$ is negative (semi)-definite, k_b can be replaced by $k_b^- := \infty$.

Proof. The result on the solution of the differential system follows from the familiar Theorem of Picard-Lindelöf [35, p. 104], the result on the solution of the linear system (1.127) follows from basic theorems in Linear Algebra, including the Perturbation Lemma [30, p. 45]. □

Both bounds are crude and could be improved. However, already one-dimensional examples show that sharp bounds necessarily differ as well.

Example 1.6 We start with the linear equation

$$\dot{u} = Lu, \qquad u(0) = y_n. \tag{1.128}$$

With $L > 0$, $f(u) = Lu$ and $|u - y_n| \leq b$ we get
$|f(u)| = |Lu| \leq L|u - y_n| + L|y_n| \leq Lb + L|y_n| =: C$.
The bound C is attained by $f(u)$ if $y_n > 0$.
Thus $k_b = 1/L$ and $t_b = b\ L^{-1}\ (b + |y_n|)^{-1} = k_b\ b\ (b + |y_n|)^{-1} \leq k_b$.
Since we can write down the solutions of (1.128) explicitly, we know that they exist for all $t \geq 0$. Thus it might be surprising at first that scheme (1.127) does not allow arbitrarily large steps if $\theta \neq 0$ and $L > 0$. We see, however, that the interval of existence ensured by the theorem of Picard-Lindelöf is even smaller than the bound k_b. This might be related to the fact that equations $\dot{u} = Lu^\alpha$, $\alpha > 1$ can have blow-up solutions and can be viewed as neighboring to eq. (1.128).

Example 1.7 This example shows that the bound k_b can be much larger as well as much smaller than the actual blow-up time of the solution. Consider

$$\dot{u} = u^2 + c, \qquad u(0) = u_0. \tag{1.129}$$

For $c = -\gamma^2 < 0$, eq. (1.129) has two steady states: $\bar{u}_- = -\gamma$ (asymptotically stable) and $\bar{u}_+ = +\gamma$ (unstable). Trajectories with initial value $u_0 < -\gamma$ converge monotonically increasing to $\bar{u}_-$, trajectories with initial value $-\gamma < u_0 < \gamma$ converge monotonically decreasing to $\bar{u}_-$, and trajectories with initial value $u_0 > \gamma$ blow up.
For $c = 0$, $-\gamma = +\gamma$ and the two steady states $\bar{u}_-$, $\bar{u}_+$ merged to the steady state $\bar{u} = 0$: blow-up for $u_0 > 0$, monotonic convergence to $\bar{u} = 0$ for $u_0 < 0$.
For $c > 0$, there is no steady state. All solutions blow up.

Now we consider the discretization of eq. (1.129) by the lintrap method (1.127): we notice that the constant c does not contribute to the derivative and thus to the bound k_b: for $u_0 < 0$ we get $k_b = k_b^- = \infty$ for all $c \in I\!R$, and for $u_0 > 0$ we get $k_b = (2u_0)^{-1}$ for all $c \in I\!R$. This produces the correct blow-up time for $c = 0$, $\theta = 1/2$ and is incorrect for all other values of c.

1.3.6.2 *Adjoint and Self-Adjoint Schemes*

As we have seen earlier, the unconventional schemes (1.104) and (1.105) for the logistic differential equation (1.84) are intimately related to each other: each of them treats one of the two stationary states of eq. (1.84) implicitly, the other one explicitly. They are **adjoint** to each other in the sense of Definition 8.2 of [12, Chap. II]: the map

$$k \mapsto -k; \qquad \begin{array}{c} y_{n+1} \mapsto y_n; \\ y_n \mapsto y_{n+1} \end{array} \tag{1.130}$$

replaces scheme (1.104) by scheme (1.105), and scheme (1.105) by scheme (1.104). Other schemes adjoint to each other are forward Euler (1.3) and backward Euler (1.8). Schemes adjoint to each other always have the same order of accuracy, and the same asymptotic expansion of the global error, with k replaced by $-k$ [12]. Schemes which are invariant under the map (1.130) are called **self-adjoint**. Self-adjoint schemes have always an even order of accuracy and an asymptotic expansion in even powers of k [12]. The trapezoidal rule and the implicit midpoint rule always produce self-adjoint schemes because their general formula is invariant under map (1.130).

The general formula for lintrap is not invariant under (1.130), but several schemes obtained by applying lintrap to a given differential system are self-adjoint: scheme (1.121) obtained by applying lintrap to the logistic differential equation, for instance, and also scheme (1.144) obtained for the Lotka Volterra system (1.140). If we apply lintrap to $\dot{u} = u^3$, however, we obtain a scheme which is not self-adjoint: it is

$$\frac{y_{n+1} - y_n}{k} = -\frac{1}{2} y_n^3 + \frac{3}{2} y_n^2 y_{n+1}.$$

Applying lintrap to $\dot{u} = u^p$ with general p and requiring self-adjointness leads, after a short calculation, to $p = 1$ or $p = 2$.

1.3.6.3 *Convergence of the Scheme*

Applied to ordinary differential equations with arbitrary θ, $0 \leq \theta \leq 1$, scheme (1.127) is first order; it is second order for $\theta = 1/2$. Its order can be raised in the framework of Runge-Kutta-ROW schemes [33; 13] or according to the methods developed by Kahan and Li for palindromic schemes [23]. The scheme was also applied to systems of nonlinear parabolic equations, and its performance on such systems was investigated theoretically. Error

bounds, convergence proofs and discussions of the convergence properties of scheme (1.122) can be found in [7; 24].

1.3.6.4 *Stability*

Assuming $k < 2k_b$ in each step such that the discrete trajectories considered exist, the stability function of scheme (1.124) is given by

$$R(z) = (1 + \frac{z}{2})(1 - \frac{z}{2})^{-1}. \tag{1.131}$$

We find that $|R(z)| < 1$ for $Re\ z < 0$ and $|R(z)| > 1$ for $Re\ z > 0$: the scheme is A-stable [13]. Moreover it gives correct linear stability to all fixed points corresponding to hyperbolic steady states. There is thus no stability limit for k introduced by (1.131). As we have seen near eq. (1.121), however, the converging trajectories might perform damped oscillations for large k.

If the differential system approximated by scheme (1.124) has periodic orbits, the stability bound on k ensuring existence of discrete periodic orbits can depend on the orbit to be approximated and can be more stringent than the bound k_b ensuring existence of the iterates. This is shown below for the Lotka-Volterra system (1.144). In that case we [28] found that the admissible size of the step-size strongly depends on the initial value.

If applied to the linear heat equation, scheme (1.122) results in the Crank-Nicholson scheme. This was derived by J. Crank and P. *Nicolson*[‡] in 1947 by applying the trapezoidal rule to the linear heat equation. The Crank-Nicholson scheme is always stable; if, however, the ratio of the time step Δt to the spatial step Δx becomes too large, the iterates perform damped oscillations.

The scheme was used extensively in laser fusion investigations, on four coupled quasilinear parabolic equations from gas dynamics (with different temperatures for electrons and ions, viscosity, the dynamics of shocks, and other effects). The heat equation used for the electron gas essentially was

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(u^{\delta}\frac{\partial u}{\partial x}\right), \tag{1.132}$$

with $\delta = 5/2$. In the testing phase, numerical solutions were compared to known similarity solutions, and the scheme was found to be robust and

[‡]Unfortunately, the misspelling of her name in the book by Richtmyer/Morton (1967) propagated and became common practice.

very useful. No numerical instabilities were observed [18, 1972 - 1978]. In later investigations the stability properties of the scheme were investigated systematically in numerical experiments. It was found that the scheme does turn unstable on quasilinear parabolic systems if the time step is too large [7, Section 6], [26].

1.4 Symplectic and Energy-Conserving Schemes

In this section we investigate a particular class of continuous dynamical systems which posses periodic solutions. When integrating such systems numerically, it is important to obtain periodic solutions as well, to preserve the underlying structure. We restrict our discussion to two dimensional problems. The general case can be found in [35].

1.4.1 *Canonical Hamiltonian Systems*

To introduce ideas, we study the equation of a circle with center $(0,0)$ and going through the point $(u_0, v_0) \in I\!R^2$,

$$\begin{array}{rcll} \dot{u} & = & -v, & u(0) = u_0, \\ \dot{v} & = & u, & v(0) = v_0. \end{array} \tag{1.133}$$

This dynamical system has one steady state, namely $u = v = 0$. The steady state is non-hyperbolic with eigenvalues $\lambda_{1,2} = \pm i$ and is marginally stable. The system is a two-dimensional **Hamiltonian system** since it can be written in the form

$$\begin{array}{rcll} \dot{u} & = & -\dfrac{\partial H}{\partial v}, & u(0) = u_0 \\ \dot{v} & = & \dfrac{\partial H}{\partial u}, & v(0) = v_0. \end{array} \tag{1.134}$$

with the Hamiltonian function [35, chap. 8]

$$H(u,v) = \frac{1}{2}(u^2 + v^2)$$

Lemma 1.10 *The solutions of Hamiltonian System (1.134) conserve energy,*

$$H(u(t), v(t)) = H(u_0, v_0). \tag{1.135}$$

Fig. 1.4 *Evolution of an area under a symplectic map:* the area of Ω_0 and Ω_1 are the same.

Proof. Equations (1.134) show, that the solution trajectories are always orthogonal to the gradient of $H(u, v)$. Thus the solution must be a level set of H and therefore the value of the Hamiltonian is preserved for all time□

For our simple example (1.133) this means that the radius is independent of t and thus the solutions are circles with center at the origin, which is the steady state.

A Hamiltonian system has a second interesting property: it is area preserving. Area preserving maps are called **symplectic**.

Lemma 1.11 *The map described by (1.134) is area preserving.*

Proof. Let Ω_0 be a subset of $\mathbb{R}^2$ at time t_0 and Ω_1 the set into which Ω_0 is mapped by (1.134) at time t_1. Preservation of area is equivalent to

$$\int_{\Omega_0} dudv = \int_{\Omega_1} dudv.$$

We now look at the domain D in u, v, t space with the boundary ∂D given by Ω_0 at t_0, Ω_1 at t_1 and the set of trajectories emerging from the boundary of Ω_0 and ending on the boundary of Ω_1 as given in Fig. 1.4. Consider the

vector field

$$\boldsymbol{w} := \begin{pmatrix} \dot{u} \\ \dot{v} \\ 1 \end{pmatrix}$$

in u, v, t space. Integrating this vector field over the boundary ∂D of D, we obtain

$$\begin{aligned} \int_{\partial D} \boldsymbol{w} \cdot \boldsymbol{n} &= \int_{\Omega_0} \boldsymbol{w} \cdot \boldsymbol{n_0} + \int_{\Omega_1} \boldsymbol{w} \cdot \boldsymbol{n_1} \\ &= \int_{\Omega_0} du dv - \int_{\Omega_1} du dv, \end{aligned}$$

where $\boldsymbol{n_0} = (0, 0, -1)^T$ and $\boldsymbol{n_1} = (0, 0, 1)^T$ denote the unit outward normal of Ω_0 and Ω_1. There is no other contribution to the surface integral, because by construction the vector field $\boldsymbol{w}$ is parallel to the trajectories, which form the rest of the boundary ∂D. Applying the divergence theorem to the left hand side of the same equation, we get

$$\begin{aligned} \int_{\partial D} \boldsymbol{w} \cdot \boldsymbol{n} &= \int_D \nabla \cdot \boldsymbol{w} \\ &= \int_D -\frac{\partial H^2}{\partial u \partial v} + \frac{\partial H^2}{\partial u \partial v} \\ &= 0\ , \end{aligned}$$

which shows that the area is preserved. □

When we apply *forward Euler* with step size k to the model problem of the circle (1.133), we get the discrete dynamical system

$$\begin{aligned} \frac{u_{n+1} - u_n}{k} &= -v_n\,, \\ \frac{v_{n+1} - v_n}{k} &= u_n\,. \end{aligned}$$

It has the same steady state $(\bar{u}, \bar{v}) = (0, 0)$ as the underlying continuous system. The eigenvalues in $(\bar{u}, \bar{v})$ are $\lambda_{1,2} = 1 \pm ik$. Thus $|\lambda_{1,2}| > 1$ for $k > 0$, and $(\bar{u}, \bar{v})$ is unstable here in contrast to the steady state of the continuous dynamical system. The scheme cannot produce closed solution trajectories. This can be seen by solving for u_{n+1} and v_{n+1} and adding their squares,

$$u_{n+1}^2 + v_{n+1}^2 = (1 + k^2)(u_n^2 + v_n^2).$$

Thus the numerical solution will spiral outward and the scheme does not conserve the Hamiltonian. Neither does the scheme preserve area: when computing the infinitesimal area element given by the wedge product $du_{n+1} \times dv_{n+1}$ [35], we find using $dx \times dx = 0$ and $dx \times dy = -dy \times dx$

$$\begin{aligned} du_{n+1} \times dv_{n+1} &= (du_n - k dv_n) \times (dv_n + k du_n) \\ &= du_n \times dv_n + k^2 du_n \times dv_n \\ &\neq du_n \times dv_n \end{aligned}$$

and thus area is not preserved by the scheme.

1.4.1.1 *Symplectic Euler*

A small change to the forward Euler scheme, however, namely using the newest solution obtained from the first equation in the second one in a Gauss Seidel fashion

$$\begin{aligned} \frac{u_{n+1} - u_n}{k} &= -v_n, \\ \frac{v_{n+1} - v_n}{k} &= u_{n+1}, \end{aligned} \tag{1.136}$$

makes the method symplectic. The following Lemma shows that this method is also symplectic for more general problems.

Lemma 1.12 *For a given 2-dimensional Hamiltonian system*

$$\begin{aligned} \dot{u} &= -H_v(u,v), & u(0) &= u_0, \\ \dot{v} &= H_u(u,v), & v(0) &= v_0, \end{aligned}$$

with a **separable Hamiltonian**, $H(u,v) = p(u) + q(v)$, *the scheme*

$$\begin{aligned} \frac{u_{n+1} - u_n}{k} &= -H_v(u_n, v_n), \\ \frac{v_{n+1} - v_n}{k} &= H_u(u_{n+1}, v_n), \end{aligned} \tag{1.137}$$

is symplectic.

We call scheme (1.137) **symplectic Euler**.

Proof. We compute the infinitesimal changes using the chain rule

$$\begin{aligned} du_{n+1} &= du_n - kH_{uv}(u_n, v_n)du_n - kH_{vv}(u_n, v_n)dv_n \\ dv_{n+1} &= dv_n + kH_{uu}(u_{n+1}, v_n)du_{n+1} + kH_{uv}(u_{n+1}, v_n)dv_n. \end{aligned}$$

Noting that H_{uv} vanishes for separable Hamiltonians, we get for the infinitesimal area element

$$\begin{aligned} du_{n+1} \times dv_{n+1} &= (du_n - kH_{vv}(u_n, v_n)dv_n) \\ &\quad \times (dv_n + kH_{uu}(u_{n+1}, v_n)du_{n+1}) \\ &= (du_n - kH_{vv}(u_n, v_n)dv_n) \\ &\quad \times (dv_n + kH_{uu}(u_{n+1}, v_n)(du_n - kH_{vv}(u_n, v_n)dv_n)) \\ &= du_n \times dv_n. \end{aligned}$$

Thus the method is symplectic for any separable Hamiltonian system. □

The scheme (1.136) is therefore symplectic for the equations of the circle. Symplectic schemes produce closed curve solutions as well, since they preserve a nearby Hamiltonian. In our example, multiplying the first equation in (1.136) by $u_{n+1} + u_n$, the second equation in (1.136) by $v_{n+1} + v_n$, and adding the resulting equations we find

$$u_{n+1}^2 + v_{n+1}^2 - ku_{n+1}v_{n+1} = u_n^2 + v_n^2 - ku_nv_n$$

and thus the nearby Hamiltonian which is preserved by the scheme is

$$\widetilde{H}(u_n, v_n) = u_n^2 + v_n^2 - ku_nv_n \tag{1.138}$$

an ellipse which approaches the circle as the time step k is refined. This ellipse is visible in Fig. 1.5, where the numerical solution is given by plus signs and the exact solution is drawn as a solid line.

1.4.1.2 *The Lintrap Scheme*

This is different for **lintrap** applied to the model problem of the circle (1.133). After a short calculation, we find the scheme

$$\begin{aligned} \frac{u_{n+1} - u_n}{k} &= -\frac{v_{n+1} + v_n}{2}, \\ \frac{v_{n+1} - v_n}{k} &= \frac{u_{n+1} + u_n}{2}. \end{aligned} \tag{1.139}$$

Surprisingly this scheme is symplectic as well. To compute the infinitesimal change in area, we first solve (1.139) for the new values as functions of the old ones,

$$u_{n+1} = \frac{4 - k^2}{4 + k^2}u_n - \frac{4k}{4 + k^2}v_n,$$

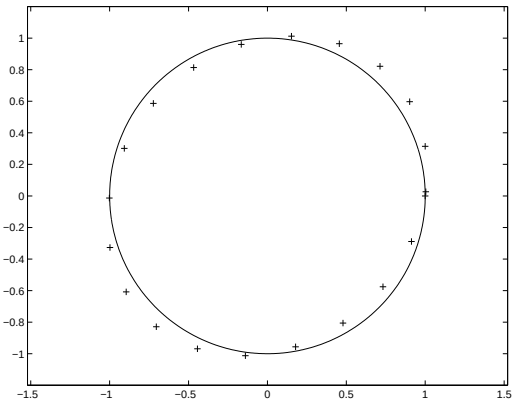

Fig. 1.5 Symplectic Euler applied to the equations of the circle. The numerical solution given by plus signs is an ellipse, preserving the slightly changed Hamiltonian

$$v_{n+1} = \frac{4k}{4+k^2}u_n + \frac{4-k^2}{4+k^2}v_n,$$

differentiate,

$$\begin{aligned} du_{n+1} &= \frac{4-k^2}{4+k^2}du_n - \frac{4k}{4+k^2}dv_n, \\ dv_{n+1} &= \frac{4k}{4+k^2}du_n + \frac{4-k^2}{4+k^2}dv_n, \end{aligned}$$

and compute the wedge product

$$\begin{aligned} du_{n+1} \times dv_{n+1} &= \left(\frac{4-k^2}{4+k^2}du_n - \frac{4k}{4+k^2}dv_n\right) \times \left(\frac{4k}{4+k^2}du_n + \frac{4-k^2}{4+k^2}dv_n\right) \\ &= \left(\left(\frac{4-k^2}{4+k^2}\right)^2 + \left(\frac{4k}{4+k^2}\right)^2\right) du_n \times dv_n \\ &= du_n \times dv_n. \end{aligned}$$

The scheme thus preserves infinitesimal area and is therefore symplectic. But it also preserves the Hamiltonian for the equation of the circle, as one can see from multiplying the first equation in (1.139) by $u_{n+1}+u_n$ and the second equation in (1.139) by $v_{n+1}+v_n$ and adding the resulting equations. One finds

$$v_{n+1}^2 + u_{n+1}^2 = v_n^2 + u_n^2$$

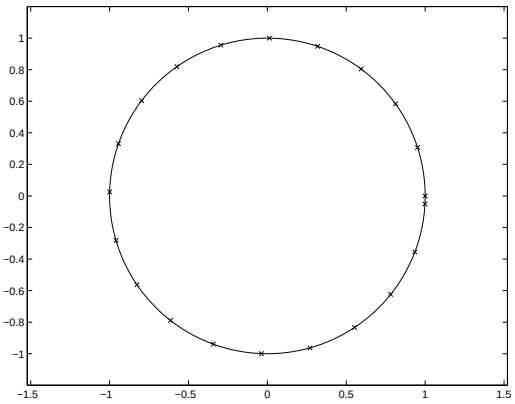

Fig. 1.6 The lintrap scheme applied to the model problem of the circle, where it is 'symplectic and Hamiltonian preserving'

and thus $H(u_{n+1}, v_{n+1}) = H(u_n, v_n)$, and the Hamiltonian is preserved. In two dimensions a scheme is exact up to a time reparametrization, if it preserves the Hamiltonian. A more general result in N dimensions says that if a scheme is symplectic and at the same time preserves the Hamiltonian, then it produces the exact solution up to a time reparametrization [38].

The error in time is visible for our example in the Figure 1.6: the numerical solution shown by crosses is lying exactly on the circle, but on the right the two crosses do not coincide after one integration around the circle, showing that the solution is only exact up to a time reparametrization.

1.4.2 *Non-Canonical Hamiltonian Systems*

Non-Canonical Hamiltonian systems include a weight in front of the derivatives of the Hamiltonian when forming the system of ordinary differential equations. They arise in many applications. In the following we consider a Lotka-Volterra system as our model problem,

$$\begin{aligned} \dot{u} &= -\alpha u + \beta uv, \quad u(0) = u_0, \\ \dot{v} &= \gamma v - \delta uv, \quad v(0) = v_0, \end{aligned} \tag{1.140}$$

with given constants $\alpha, \beta, \gamma, \delta > 0$. Equation (1.140) models the evolution of a predator population u and its prey population v. There are two steady

states:

$$\bar{\mathbf{u}}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{with} \quad \mathbf{f}'(\bar{\mathbf{u}}_1) = \begin{pmatrix} -\alpha & 0 \\ 0 & \gamma \end{pmatrix} \tag{1.141}$$

and

$$\bar{\mathbf{u}}_2 = \begin{pmatrix} \gamma/\delta \\ \alpha/\beta \end{pmatrix} \quad \text{with} \quad \mathbf{f}'(\bar{\mathbf{u}}_2) = \begin{pmatrix} 0 & \beta\gamma/\delta \\ -\delta\alpha/\beta & 0 \end{pmatrix}. \tag{1.142}$$

$\bar{\mathbf{u}}_1$ is a saddle point and unstable. $\bar{\mathbf{u}}_2$ is non-hyperbolic with eigenvalues $\lambda_{1,2} = \pm i\sqrt{\alpha\gamma}$ and known to be marginally stable: it is a center, surrounded by nested closed curves (periodic orbits) as we will see below. Both coordinate axes are invariant under (1.140): an initial value $\mathbf{u}_0 = (u, 0)^t$ leads to $\dot{v} = 0$ and to $\dot{u} = -\alpha u$: the iterates approach the origin $\bar{\mathbf{u}}_1$; an initial value $\mathbf{u}_0 = (0, v)^t$ leads to $\dot{u} = 0$ and to $\dot{v} = \gamma v$: the iterates move away from the origin $\bar{\mathbf{u}}_1$.

The Lotka-Volterra system (1.140) can be written as

$$\begin{aligned} \dot{u} &= uv\frac{\partial H}{\partial v}, \quad u(0) = u_0 \\ \dot{v} &= -uv\frac{\partial H}{\partial u}, \quad v(0) = v_0 \end{aligned} \tag{1.143}$$

with the Hamiltonian

$$H(u, v) = -\alpha \ln v + \beta v - \gamma \ln u + \delta u$$

and therefore it is a **non-canonical Hamiltonian system** . It would be Hamiltonian, if the weighting factor uv was not present in (1.143). The trajectories are still level sets of H, since they evolve along a vector orthogonal to the gradient of H and thus Lemma 1.10 applies. But the system does not preserve area. A related quantity, however, is preserved, as the following Lemma shows.

Lemma 1.13 *The map described by (1.143) preserves area weighted by the factor* $\frac{1}{uv}$.

Proof. Let Ω_0 be a subset of $\mathbb{R}^2$ at time t_0 and Ω_1 the set into which Ω_0 is mapped by (1.143) at time t_1, as in Fig. 1.4. Preservation of weighted area means

$$\int_{\Omega_0} \frac{1}{uv}\, dudv = \int_{\Omega_1} \frac{1}{uv}\, dudv.$$

Proceeding as in the classical Hamiltonian case, we look at the domain D in x, y, t space with the boundary ∂D given by Ω_0 at t_0, Ω_1 at t_1 and the set of trajectories emerging from the boundary of Ω_0 and ending on the boundary of Ω_1. Consider the vector field

$$\boldsymbol{w} := \frac{1}{uv} \begin{pmatrix} \dot{u} \\ \dot{v} \\ 1 \end{pmatrix}$$

in u, v, t space. Integrating this vector field over the boundary ∂D of D, we obtain

$$\begin{aligned} \int_{\partial D} \boldsymbol{w} \cdot \boldsymbol{n} &= \int_{\Omega_0} \boldsymbol{w} \cdot \boldsymbol{n}_0 + \int_{\Omega_1} \boldsymbol{w} \cdot \boldsymbol{n}_1 \\ &= \int_{\Omega_0} \frac{1}{uv}\, dudv - \int_{\Omega_1} \frac{1}{uv}\, dudv \ , \end{aligned}$$

where $\boldsymbol{n}_0 = (0,0,-1)^T$ and $\boldsymbol{n}_1 = (0,0,1)^T$ denote the outward unit normal of Ω_0 and Ω_1. There is no other contribution to the surface integral, because by construction the vector field $\boldsymbol{w}$ is parallel to the trajectories, which form the rest of the boundary ∂D. Applying the divergence theorem to the left hand side of the same equation, we get

$$\begin{aligned} \int_{\partial D} \boldsymbol{w} \cdot \boldsymbol{n} &= \int_D \nabla \cdot \boldsymbol{w} \\ &= \int_D -\frac{\partial H^2}{\partial u \partial v} + \frac{\partial H^2}{\partial u \partial v} \\ &= 0 \ , \end{aligned}$$

which shows that the weighted area is preserved. □

1.4.2.1 *Lintrap for Lotka-Volterra*

W. Kahan considered in his (unpublished) lecture notes on 'unconventional numerical methods' the Lotka-Volterra system (1.140) and gave the following difference scheme for it:

$$\begin{aligned} \frac{u_{n+1} - u_n}{k} &= -\frac{\alpha}{2}(u_{n+1} + u_n) + \frac{\beta}{2}(u_{n+1} v_n + u_n v_{n+1}) \\ \frac{v_{n+1} - v_n}{k} &= \frac{\gamma}{2}(v_{n+1} + v_n) - \frac{\delta}{2}(u_{n+1} v_n + u_n v_{n+1}) \ . \end{aligned} \tag{1.144}$$

He noticed that, for positive u and v, the discrete trajectories are closed curves. Sanz-Serna [34] then showed that the scheme (1.144) is symplectic with respect to the non-canonical Hamiltonian given above.

Scheme (1.144) can be obtained by applying the **lintrap** method to eq. (1.140) [27]. The scheme is symplectic and thus produces closed curve solutions as long as the symplectic structure is maintained, i.e. as long as $1/(u_n v_n) > 0$. In the following we describe what happens if the step size is allowed to be large [28]. We show that closed curves can only be produced by (1.144) in a neighborhood of the fixed point $\mathbf{u}_2$, and the size of the neighborhood depends on the step size k. As soon as iterates violate the condition $u_{n+1} > 0$, $v_{n+1} > 0$, the step size k for obtaining them was too large, the discrete model is not an acceptable model of the continuous system any more.

To simplify notation, we assume

$$\alpha = \beta = \gamma = \delta = 1. \tag{1.145}$$

The time step k is considered a varying parameter now, we consider a family of dynamical systems. In explicit form, iteration (1.144) with (1.145) reads

$$\begin{aligned} u_{n+1} &= u_n + \frac{2k u_n (2(v_n - 1) + k(1 - u_n))}{4 - k^2 + 2k u_n + k^2 u_n - 2k v_n + k^2 v_n} \\ v_{n+1} &= v_n + \frac{2k v_n (2(1 - u_n) + k(1 - v_n))}{4 - k^2 + 2k u_n + k^2 u_n - 2k v_n + k^2 v_n} \end{aligned} \tag{1.146}$$

which we also write as

$$\mathbf{y}_{n+1} = \mathbf{F}(\mathbf{y}_n), \qquad \mathbf{y}_0 = \mathbf{u}_0. \tag{1.147}$$

The map $\mathbf{F}$ has two fixed points:

$$\bar{\mathbf{y}}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{with} \quad \mathbf{F}'(\bar{\mathbf{y}}_1) = \begin{pmatrix} \frac{2-k}{2+k} & 0 \\ 0 & \frac{2+k}{2-k} \end{pmatrix} \tag{1.148}$$

and

$$\bar{\mathbf{y}}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{with} \quad \mathbf{F}'(\bar{\mathbf{y}}_2) = \begin{pmatrix} \frac{4-k^2}{4+k^2} & \frac{4k}{4+k^2} \\ -\frac{4k}{4+k^2} & \frac{4-k^2}{4+k^2} \end{pmatrix}. \tag{1.149}$$

The eigenvalues of $\mathbf{F}'(\bar{\mathbf{y}}_2)$ are $\lambda_{1,2} = (2 \pm ik)^2/(4 + k^2)$ and they satisfy $|\lambda_{1,2}| = 1$. Points on the u- and v- axes are iterated according to

$$\mathbf{y}_{n+1} = \begin{pmatrix} u_n \frac{2-k}{2+k} \\ 0 \end{pmatrix} \qquad \text{and} \qquad \mathbf{y}_{n+1} = \begin{pmatrix} 0 \\ v_n \frac{2+k}{2-k} \end{pmatrix}. \tag{1.150}$$

Fig. 1.7 *Dependence of the dynamics of (1.144) on the step size* [28, Fig. 1]
White: the 20-th iterate with initial value in this area is still in Q_1; Shaded: the m-th iterate, $m \leq 20$, has left Q_1 for Q_4 (dark grey) or for Q_2 (light grey). In addition, the periodic orbits with initial values (1.0,1.2) and (1.0,2.1) are shown.

Fig. 1.8 Similar to Fig. 1.7. In addition, the first 10 iterates of the trajectory with initial value $\mathbf{y}_0 = (0.44, 2.0)^t$ are shown. [28, Fig. 2]

As we see from this, the scheme has a singularity for $k = 2$.

We first consider the case $0 < k < 2$. As already known, the scheme preserves periodic orbits [34]. The size of the neighborhood of $\bar{\mathbf{y}}_2 = (1, 1)^t$ where this is true, however, shrinks with increasing k. This is shown in Fig. 1.7. The area where periodic orbits exist is bounded by a straight line

which is also shown in both parts of Fig. 1.7: it is given by

$$g_2(k): \qquad v_n = \frac{2+k}{k} - u_n \frac{2-k}{2+k} \tag{1.151}$$

and has the property that all $(u_n, v_n)^t \in g_2(k)$ satisfy

$$v_{n+1} = 0, \quad u_{n+1} = \frac{k+2}{k} \tag{1.152}$$

i.e. all points on $g_2(k)$ are mapped into the same point $(u_{n+1}, 0)^t$ on the u-axis. For $k \to 0$, $g_2(k)$ moves to infinity: its inclination approaches -1 and its intersection points with the u-axis and with the v-axis

$$u_k = \frac{(2+k)^2}{k(2-k)} \qquad \text{and} \qquad v_k = \frac{2+k}{k} \tag{1.153}$$

both converge to infinity. For $k \to 2$, $g_2(k)$ converges to the line $v \equiv 2$.

There are two more straight lines with special properties:

$$g_1(k): \qquad v_n = -\frac{(k-2)^2}{(k+2)k} - u_n \frac{2-k}{2+k} \tag{1.154}$$

has the property that all $(u_n, v_n)^t \in g_1(k)$ satisfy

$$u_{n+1} = 0, \quad v_{n+1} = \frac{k-2}{k}, \tag{1.155}$$

i.e. all points on $g_1(k)$ are mapped into the same point $(0, v_{n+1})^t$ on the v-axis. $g_1(k)$ is parallel to $g_2(k)$ and cannot be seen in Fig. 1.7 because we show only the positive sector Q_1 of the plane. For $k \to 0$, $g_1(k)$ moves to infinity in the negative sector Q_3; in the limiting case $k \to 2$ it coincides with the u-axis. All points on

$$g_3(k): \qquad v_n = \frac{2+k}{k} + u_n \frac{2+k}{2-k} \tag{1.156}$$

satisfy $4 - k^2 + 2ku_n + k^2 u_n - 2kv_n + k^2 v_n = 0$, i.e. they have the property that the denominator in (1.146) vanishes: they cause blow-up in one step. $g_3(k)$ intersects with $g_2(k)$ on the v-axis and is perpendicular to $g_2(k)$. For $k \to 0$ it also disappears to infinity, for $k \to 2$ it coincides with the v-axis. It is shown in Fig. 1.7 for $k = 1$.

Together with the axes, these three straight lines partition the plane. All points above $g_2(k)$ and to the right of $g_3(k)$ are mapped into the sector $Q_4 = \{(u, v) \in I\!R^2 : u > 0,\ v < 0\}$. All points above $g_2(k)$ and between $g_3(k)$ and

Fig. 1.9 *Degenerate dynamics of (1.144) for step size* $k = 2$ [28, Fig. 3].

the v-axis are mapped into the sector $Q_2 = \{(u, v) \in I\!R^2 : u < 0,\ v > 0\}$. For the points in the triangle between the u-axis, the v-axis and $g_2(k)$ things are more complicated (see Figs. 1.7 and 1.8): The orbits of certain points stay in Q_1, others are mapped into the other sectors Q_i, $i \neq 1$. White: the 20-th iterate with initial value $\mathbf{y}_0$ is still in Q_1; shaded: the m-th iterate, $m \leq 20$, has left Q_1 for Q_4 (dark grey) or for Q_2 (light grey). Figure 1.7 shows in addition the two orbits with initial value $(1.0, 1.2)^t$ and with $(1.0, 2.1)^t$ for $k = 0.5$ (left) and for $k = 1.$(right). Figure 1.8 shows in addition the first 10 iterates with $k = 1$ and initial value $\mathbf{y}_0 = (0.44, 2.0)^t$. The trajectory 'visits' subsets associated to $m = 10, 9, 8, \ldots$ until it leaves Q_1. Note that each subset *'the trajectory with initial value in this set leaves* Q_1 *in the m-th step'* is not connected but fractal. Figures 1.7 and 1.8 were obtained numerically by iterating the map up to 20 times for each initial value on a fine grid, and then painting the pixel at the initial value according to the result.

In the case $k = 2$, the iteration (1.144) degenerates to

$$\begin{aligned} u_{n+1} &= v_n, \\ v_{n+1} &= \frac{2v_n - v_n^2}{u_n}, \end{aligned} \tag{1.157}$$

and the periodic orbits degenerate to 4-cycles:

$$\begin{pmatrix} a \\ a \end{pmatrix} \mapsto \begin{pmatrix} a \\ 2-a \end{pmatrix} \mapsto \begin{pmatrix} 2-a \\ 2-a \end{pmatrix} \mapsto \begin{pmatrix} 2-a \\ a \end{pmatrix} \mapsto \begin{pmatrix} a \\ a \end{pmatrix}. \tag{1.158}$$

Orbits with initial values not lying on

$$g_4(2): \quad v_n = u_n \quad \text{or on} \quad g_5(2): \quad v_n = 2 - u_n \tag{1.159}$$

but 'close' to them (white area in Fig. 1.9) stay longer in Q_1 than the orbits with other initial values, but eventually leave the slab on a 4-spiral (see Fig. 1.9b).

1.4.2.2 *Symplectic Euler for Lotka-Volterra*

Now we consider symplectic Euler for the Lotka-Volterra predator-pray system (1.140). The special case with $\alpha = \beta = \gamma = \delta = 1$ was already treated in [8]. We have

$$\begin{aligned} \frac{u_{n+1} - u_n}{k} &= \alpha u_n - \beta u_n v_n \\ \frac{v_{n+1} - v_n}{k} &= -\gamma v_n + \delta u_{n+1} v_n \end{aligned} \tag{1.160}$$

Lemma 1.14 *Symplectic Euler (1.160) is also symplectic for the non canonical Hamiltonian system (1.140), the weighted area* $\frac{du \times dv}{uv}$ *is preserved.*

Proof. To compute the infinitesimal change in u_{n+1} and v_{n+1} we take derivatives in (1.160),

$$\begin{aligned} du_{n+1} &= du_n + k\alpha du_n - k\beta du_n v_n - k\beta u_n dv_n \\ dv_{n+1} &= dv_n - k\gamma dv_n + k\delta du_{n+1} dv_n + k\delta u_{n+1} dv_n \end{aligned}$$

and then compute the wedge product $du_{n+1} \times dv_{n+1}$. Observing again that $dx \times dx = 0$ and $dx \times dy = -dy \times dx$ we get after a short calculation

$$\frac{du_{n+1} \times dv_{n+1}}{u_{n+1} v_{n+1}} = \frac{du_n \times dv_n}{u_n v_n}.$$

Thus the weighted area is preserved. □

1.5 Acknowledgement

The authors thank Ron Mickens for inviting them to contribute to this volume, and for his constant support.

Bibliography

R.L. Burden, J.D. Faires (1985): *Numerical Analysis*, 3rd ed., Prindle, Weber & Schmidt, Boston

R.M. Beam, R.F. Warming (1978): An implicit factored scheme for the compressible Navier-Stokes equations. AIAA J. **16**, 393 – 402

W.-J. Beyn (1990): Numerical methods for dynamical systems. In: *Proceedings of the SERC Summer School at Lancaster (UK),* Oxford University Press

R.M. Corless, C. Essex, M.A.H. Nerenberg (1991): Numerical methods can suppress chaos. Physics Letters A, **157**, 27–36

L. Dieci, D. Estep (1991): Some stability aspects of schemes for the adaptive integration of stiff initial value problems. SIAM J. Sci. Stat. Comput. **12**, 1284 – 1303

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

K. v. Finckenstein (1987): Difference methods for quasilinear parabolic systems from plasma physics. Numer. Meth. PDEs **3**, 289 – 311

M.J. Gander (1994): A non spiraling integrator for the Lotka Volterra equations. Il Volterriano **4**, 21 – 28

K. Grote (1996): *Dynamische Eigenschaften von Einschrittverfahren.* Diplomarbeit, Fakultät für Mathematik, TU München, 81 Seiten

K. Grote, R. Meyer-Spasche (1998): *On Euler-like discrete models of the logistic differential equation.* In *Advances in Difference Equations II*, R.P. Agarwal, (ed.), Computers Math. Applic. **36**, 211-225,

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

E. Hairer, S.P. Nørsett, G. Wanner (1992): Solving Ordinary Differential Equations, vol. I, 2nd ed., Springer Verlag

E. Hairer, G. Wanner (1991): *Solving Ordinary Differential Equations*, vol. II, Springer Verlag

J.H. Hubbard, B.H. West (1991): *Differential Equations. A Dynamical Systems*

Approach; Part I: Ordinary Differential Equations. Springer Verlag, New York

E. Kamke (1967): *Differentialgleichungen: Lösungsmethoden und Lösungen. Teil I, Gewöhnliche Differentialgleichungen.* Akademische Verlagsgesellschaft Geest& Portig, Leipzig, 8^{th} edition

A. Iserles, A. Peplow, A.M. Stuart (1991): A unified approach to spurious solutions introduced by time discretization. Part I: Basic theory. SIAM J. Num. Anal. **28**, 1723–1751

A. Iserles (1994): Dynamics of Numerics. Bull. IMA **30**, 106–115

K. Lackner, MPI für Plasmaphysik, private communication

M.N. Le Roux (1998): Numerical solution of fast diffusion or slow diffusion equations. J. Comp. Appl. Math. **97**, 121 – 136 (1998)

M.N. Le Roux (1994): Semidiscretization in time of nonlinear parabolic equations with blowup of the solution, SIAM J. Numer. Anal. **31**, 170-195

M.N. Le Roux, J. Weiland, H. Wilhelmsson (1992): Simulation of a coupled dynamic system of temperature and density in a fusion plasma, Phys. Scripta **46**, 457 – 462

M.N. Le Roux, H. Wilhelmsson (1989): External boundary effects on simultaneous diffusion and reaction processes, Phys. Scripta **40**, 674 – 681

R.-C. Li (1995): *Raising the Orders of Unconventional Schemes for Ordinary Differential Equations.* Ph.D. dissertation in Applied Mathematics, UC Berkeley, 251 pages

Ch. Lubich, A. Ostermann (1995): Linearly implicit time discretization of nonlinear parabolic equations. IMA J. Numer. Anal. **15**, 555-583

R.M. May (1976): Simple mathematical models with very complicated dynamics. Review article, Nature **261**, 459 – 467

R. Meyer-Spasche (1998): Difference schemes of optimum degree of implicitness for a family of simple ODEs with blow-up solutions. J. Comp. Appl. Math. **97**, 137 – 152 (1998)

R. Meyer-Spasche, D. Düchs (1997): A general method for obtaining unconventional and nonstandard difference schemes. Dyn. Cont., Discr. Impulsive Systems **3**, 453 – 467

R. Meyer-Spasche, K. Grote (1997): Dynamical Properties of a Linearly Implicit Scheme, pp. 581 – 586 in: *Proc. 15th IMACS World Congress on Scientific Computation, Modelling and Applied Mathematics, vol 2*, A. Sydow (ed.); Wissenschaft & Technik Verlag, Berlin

R.E. Mickens (1994): *Nonstandard Finite Difference Models of Differential Equations.* World Scientific, Singapore

J.M. Ortega, W.C. Rheinboldt (1970): *Iterative Solution of Nonlinear Equations in Several Variables* Academic Press, New York

K. Ozawa (1999): Functional fitting Runge-Kutta method with variable coefficients, Japan J. of Industr. and Appl. Math., to appear

Proc. *The Dynamics of Numerics and the Numerics of Dynamics*, D.S. Broomhead and A. Iserles (eds.), Oxford University Press, 1992

P. Rentrop, K. Strehmel, R. Weiner (1996): Ein Überblick über Einschrittverfahren zur numerischen Integration in der technischen Simulation. GAMM-Mitteilungen **19**, 9 – 43

J.M. Sanz-Serna (1994): An unconventional symplectic integrator of W. Kahan. Appl. Numer. Math. **16**, 245 - 250

A.M. Stuart and A.R. Humphries (1996): *Dynamical Systems and Numerical Analysis,* Cambridge University Press

Y. Wang, E.H. Twizell, W.G. Price (1992): Second order numerical methods for the solution of equations in population modelling. Comm. Appl. Numer. Meth. **8**, 511 – 518

H.C. Yee, P.K. Sweby, D.F. Griffiths (1991): Dynamical approach study of spurious steady-state numerical solutions of nonlinear differential equations I. J. Comp. Phys. **97**, 249 – 310

G. Zhong and J.E. Marsden (1998): Lie-Poisson Hamilton-Jacobi theory and Lie-Poisson integrators. Phys. Lett. A, 133:134–139