Contents

Part I
Ordinary Differential Equations

In this part, we discuss the standard numerical techniques used to integrate systems of ordinary differential equations (ODEs).

Chapter 1
Preliminary Concepts

By definition, an ordinary differential equation (ODE), is a differential equation in which all dependent variables are functions of a *single* independent variable.

Chapter 2
Initial Value Problem

An initial value problem (IVP) is a system of differential equation

$$\frac{d\mathbf{x}}{dt} = f(t, \mathbf{x}(t)), \tag{2.1}$$

$\mathbf{x}(t) = (x_1(t), x_2(t), \ldots x_n(t))^T$, $f \in [a,b] \times \mathbb{R}^n \to \mathbb{R}^n$, together with specified value

$$\mathbf{x}(a) = \mathbf{x}_0, \tag{2.2}$$

called the initial condition. We are interested in a numerical approximation of the continuously differentiable solution $\mathbf{x}(t)$ of the IVP (2.1)–(2.2) over the time interval $t \in [a,b]$. To this aim we subdivide the interval $[a,b]$ into M equal subintervals and select *the mesh points* t_j [45, 30]

$$t_j = a + jh, \qquad j = 0, 1, \ldots, M, \quad h = \frac{b-a}{M}. \tag{2.3}$$

The value h is called *a step size*.

2.1 Explicit Euler's Method

The simplest method to approximate IVP (2.1)–(2.2) was devised by Leonhard Euler in 1768. The idea of the method is straightforward: From the initial condition (2.2) we know that at $t = t_0 = a$ the slope of the solution curve $d\mathbf{x}/dt$ is $f(t_0, \mathbf{x}_0)$. Therefore we can try to obtain the next approximation $\mathbf{x}_1 := \mathbf{x}(t_1)$ at a small time h later by adding $h\, f(t_0, \mathbf{x}_0)$ to $\mathbf{x}_0$, namely

$$\mathbf{x}_1 = \mathbf{x}_0 + h\, f(t_0, \mathbf{x}_0).$$

Now we can take another step forward in the same way, using the slope $f(t_1, \mathbf{x}_1)$, corresponding to the new time $t_1 = t_0 + h$, i.e.,

$$\mathbf{x}_2 = \mathbf{x}_1 + h\, f(t_1, \mathbf{x}_1)\,.$$

The process is repeated and generates a sequaence of points $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_M$ that approximates the solution $\mathbf{x}(t)$. The general step of the Euler's method is [**?**, 45]

$$\begin{aligned} \mathbf{x}_{j+1} &= \mathbf{x}_j + h\, f(t_j, \mathbf{x}_j)\,, \\ t_{j+1} &= t_j + h, \quad j = 0, 1, \ldots, M-1. \end{aligned} \tag{2.4}$$

Notice that the Euler method (2.4) is *an explicit method*, i.e., $\mathbf{x}_{j+1}$ is given *explicitly* in terms of known quantities such as $\mathbf{x}_j$ and $f(t_j, \mathbf{x}_j)$.

From geometrical point of view, one starts at the point $(t_0, \mathbf{x}_0)$ of the $(t, \mathbf{x})$-plane and is moving along the tangent line to the solution $\mathbf{x}(t)$ and will end up at the point $(t_1, \mathbf{x}_1)$. Now this point is used to compute the next slope $f(t_1, \mathbf{x}_1)$ and to locate the next approximation point $(t_2, \mathbf{x}_2)$ etc.

Example 1

Let us use Euler's method (2.4) to solve approximately a simple IVP [**?**]

$$\dot{x} = x, \quad \text{over} \quad t \in [0, 1], \quad x(0) = 1. \tag{2.5}$$

The exact solution is $x(t) = \exp(t)$, so we can calculate the correct value at the end of the time interval, i.e.,

$$x(1) = e = 2.71828...$$

Let us find the numerical approximation of (2.5) for different step sizes h=$\{0.1, 1\text{e-}2, 1\text{e-}3, 1\text{e-}4, 1\text{e-}5\}$ and calculate the difference between obtained numerical value at the end of the time interval x_{end} and the exact value e. The results are shown in Table (2.1) The presented results demonstrate that the error at the end of interval is

Table 2.1 Numerical results, obtained by Euler's methhod for Eq. (2.5)

h	x_{end}	$\lvert x_{end} - 1\rvert$
0.1	2.5937	0.12
1e-2	2.7048	1.35e-2
1e-3	2.7169	1.4e-3
1e-4	2.7181	1.35e-4
1e-5	2.7183	1.35e-5

propotional to the step size h.

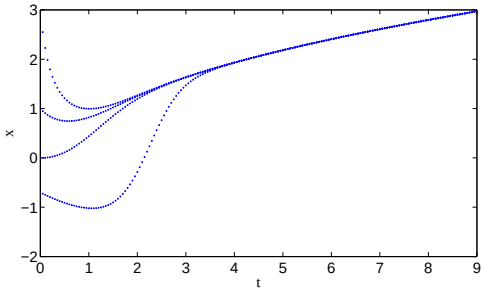

Fig. 2.1 Numerical solution of Eq. (2.6) over the interval $[0, 9]$ by the method (2.4) with the step size $h = 0.05$ for four different initial values $x_0 = \{-0.7, 0, 1, 3\}$.

Example 2

Now let us solve a nonlinaer IVP [**?**]

$$\dot{x} = t - x^2, \quad \text{over} \quad t \in [0, T], \quad x(0) = x_0 \tag{2.6}$$

for different values of x_0 and T using the method (2.4). Figure 2.1 shows numerical solutions of Eq. (2.6) over the time interval $[0, 9]$ for four different initial values $x_0 = \{-0.7, 0, 1, 3\}$. One can see that the solutions corresponding to different x_0 converge to the same curve. But if we compute the solution again for a longer time interval, say $t \in [0, 900]$ for, e.g., $x_0 = 0$, the numerical solution starts to oscillate from some time moment on (see Fig. 2.2 (a)) and the oscillations character becomes chaotic. This effect indicate the instability of the Euler's method at least at the choosen value of the time step. However, the effect disappears if we repeat the calculation with a smaller h (see Fig. 2.2 (b) for details).

The presented examples raises a number of questions. One of these is the question of *convergence*. That is, as the step size h tends to zero, do the values of the numerical solution approach the corresponding values of the actual solution? Assuming that the answer is affirmative, there remains the important practical question of how rapidly the numerical approximation converges to the solution. In other words, how small a step size is needed to guarantee a given level of accuracy? We discuss these questions below.

2.2 Numerical errors

Generally there are two major sources of error associated with a numerical integration scheme for ODE's, namely, *truncation error* and *rounding error*, which is due to finite precision of floating-point arithmetic [**?**, 30].

For the sake of simplicity, let us suppose that the IVP (2.1)–(2.2) is posed for the first order ODE, i.e., $\mathbf{x}(t)$ is a scalar value.The *local discretization (truncation) error* of IVP (2.1)–(2.2) is the error committed in the single step from t_j to t_{j+1} and is defined by [**?**]

$$\varepsilon_{j+1} = \frac{1}{h}\big(\mathbf{x}(t_{j+1}) - \mathbf{x}(t_j)\big) - f(t_j, \mathbf{x}_j)\,, \qquad j = 0, \ldots, M\,. \tag{2.7}$$

In addition, the integration scheme is *consistent* [**?**], if

$$\max_{t_j} \|\varepsilon_j\| \to 0\,, \quad \text{for} \quad h_{\max} \to 0\,, \tag{2.8}$$

where $h_{\max} = \max_j h_j$ and $h_j = t_{j+1} - t_j$.

The explicit Euler method (2.4) is based on a truncated Taylor series expansion, i.e., if we expand $\mathbf{x}(t)$ in the neighborhood of $t = t_j$, we get

$$\mathbf{x}(t_j + h) = \mathbf{x}(t_j) + h\mathbf{x}'(t_j) + \frac{h^2}{2!}\mathbf{x}''(t_j) + \ldots\,. \tag{2.9}$$

Thus, every time we take a step using Euler's method (2.4), we incur a truncation error of $\mathscr{O}(h^2)$, i.e., the local truncation error for the Euler method is proportional to the square of the step size h and the proportionality factor depends on the second derivative of the solution .

The local truncation error (2.7) is different from the *global discretization (truncation) error* e_j, which is defined as the difference between the true solution and the computed solution, i.e.,

$$e_j = \mathbf{x}(t_j) - \mathbf{x}_j\,, \qquad j = 0, \ldots, M, \tag{2.10}$$

where $\mathbf{x}(t_j)$ denotes the exact solution on the step j and $\mathbf{x}_j$ stands for its numerical approximation. The concept of the global discretization error is connected with the notion of *convergency* of the method, namely, the numerical scheme is convergent,

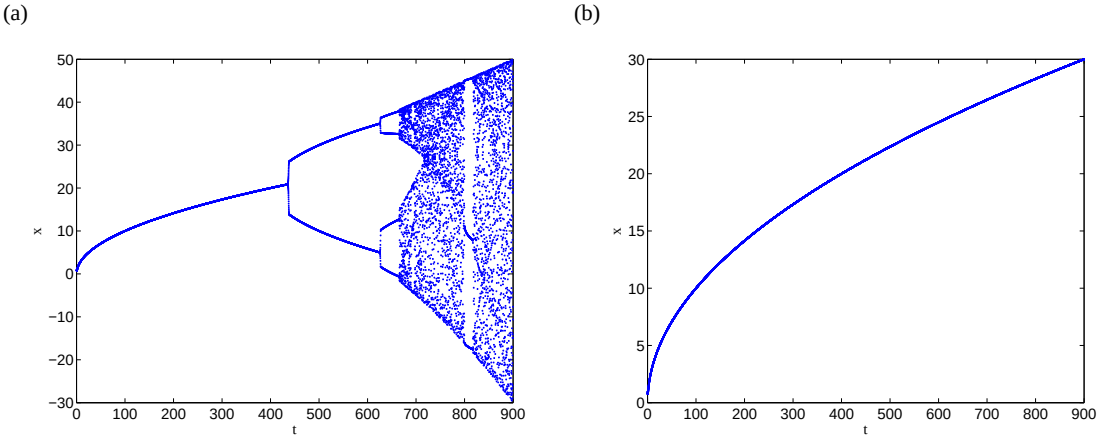

Fig. 2.2 Numerical implementation of the Euler's method for Eq. (2.6) over a long time interval, $T = 900$ and initial value $x_0 = 0$. (a) with $h = 0.05$, illustrating numerical instability of the scheme (2.4). (b) The instability desappeares for a smaller step size $h = 0.025$.

if

$$\max_{t_j} \|e_j\| \to 0\,, \quad \text{for} \quad h_{\max} \to 0\,. \tag{2.11}$$

Moreover, one can also say that the scheme possesses *the order of convergency p*, if

$$\max_{t_j} \|e_j\| \leq K h_{\max}^p\,, \qquad h_{\max} \to 0\,, \tag{2.12}$$

where K is some constant value.

If we are interested in study of the behavior of the error for various step sizes we can also consider the *final global error*, which is the global truncation error at the end of the integration intervall $[a, b]$ [30], i.e.,

$$E(\mathbf{x}(b), h) = \mathbf{x}(b) - \mathbf{x}_M. \tag{2.13}$$

In most cases, we do not know the exact solution and hence the final global error (2.13) is not possible to be evaluated. However, if we neglect round-off errors, it is reasonable to assume that the global error after M time step is M times the local truncation error (2.7), since M is proportional to $1/h$, $E(\mathbf{x}(b), h)$ should be proportional to ε_{j+1}/h. For example, for the Euler's method (2.4) the accumulated error would be

$$E(\mathbf{x}(b), h) = \sum_{j=1}^{M} \frac{h^2}{2}\mathbf{x}'' \approx M\mathbf{x}''\frac{h^2}{2} = \frac{hM}{2}\mathbf{x}''h = \frac{b-a}{2}\mathbf{x}''h = \mathcal{O}(h)$$

Thus, the explicit Euler method is of the *first order of convergency*.

Let us consider two approximations, made by the method (2.4), using the steps sizes h and $h/2$. Then we obtain

$$E(\mathbf{x}(b), h) \approx Kh, \qquad K = \text{const}\,,$$

and

$$E(\mathbf{x}(b), \frac{h}{2}) \approx K\frac{h}{2} = \frac{1}{2}Kh \approx \frac{1}{2}E(\mathbf{x}(b), h)\,.$$

Hence if the step size in Euler's method (2.4) is reduced by a factor of $1/2$ we can expect that the final global truncation error (2.13) will be reduced by the same factor (see also Example 2 of Section (2.1)).

2.3 Heun's Method

We have seen that the Euler method (2.4) is not sufficiently accurate to be an efficient problem-solving procedure, e.g., the rate of convergence scaling linearly with the step size h, so it is desirable to develop more accurate methods. To this end, let us consider IVP (2.1)–(2.2). Integrating both sides of Eq. (2.1) over one time step from t_j to t_{j+1} we obtain the *exact* relation

$$\mathbf{x}(t_{j+1}) - \mathbf{x}(t_j) = \int_{t_j}^{t_{j+1}} f(t, \mathbf{x}(t))\, dt\,. \tag{2.14}$$

Now a numerical integration method can be used to approximate the definite integral in Eq. (2.14). From the geometrical point of view, the right-hand side of (2.14) corresponds to the area S under the curve $f(t, \mathbf{x}(t))$, between t_j and t_{j+1}. For example, Euler's method (2.4) consists of approximation of right-hand side of (2.14) by the area of the rectangle S_r with the height $f(t_j, \mathbf{x}(t_j))$ and width h, i.e., one obtains Eq. (2.4), namely

$$\mathbf{x}_{j+1} = \mathbf{x}_j + S_r = \mathbf{x}_j + h\, f(t_j, \mathbf{x}(t_j))\,.$$

Clearly, a better approximation to the area S can be obtained if we use the trapezium with area

$$S_t = \frac{h}{2}\Big(f(t_j, \mathbf{x}(t_j)) + f(t_{j+1}, \mathbf{x}(t_{j+1}))\Big)\,,$$

yielding

$$\mathbf{x}_{j+1} = \mathbf{x}_j + \frac{h}{2}\Big(f(t_j, \mathbf{x}(t_j)) + f(t_{j+1}, \mathbf{x}(t_{j+1}))\Big)\,. \tag{2.15}$$

Notice that the r.h.s. of Eq. (2.15) contains the yet unknown value $\mathbf{x}_{j+1}$. In order to overcome this difficulty we use the Euler's approximation (2.4) to replace $f(t_{j+1}, \mathbf{x}(t_{j+1}))$ with $f(t_{j+1}, \mathbf{x}_j + h\, f(t_j, \mathbf{x}(t_j)))$. After it is substituted into Eq. (2.15), the resulting expression is called *Heun's*, *trapezoid* or *improved Euler's method*:

$$\mathbf{x}_{j+1} = \mathbf{x}_j + \frac{h}{2}\Big(f\big(t_j, \mathbf{x}(t_j)\big) + f\big(t_j + h, \mathbf{x}_j + h\, f(t_j, \mathbf{x}(t_j))\big)\Big)\,. \tag{2.16}$$

The improved Euler formula [30, **?**] is an example of a two-stage method: First, Euler's method (2.4) is used as a *prediction*, and then the trapezoidal rule is used as a *correction*,i.e.,

$$\begin{aligned} \mathbf{y}_{j+1} &= \mathbf{x}_j + h\, f(t_j, \mathbf{x}_j)\,, \\ t_{j+1} &= t_j + h\,, \\ \mathbf{x}_{j+1} &= \mathbf{x}_j + h\, f(t_{j+1}, \mathbf{y}_{j+1})\,. \end{aligned} \tag{2.17}$$

The local truncation error ε_{j+1} for the trapezoidal formula (2.18) is $\mathscr{O}(h^3)$ as opposed to $\mathscr{O}(h^2)$ for the Euler's method (2.4) [45, 30]. It can also be shown that for a finite interval, the global truncation error for (2.18) is bounded by $\mathscr{O}(h^2)$, so this method is *a second order method*. Indeed, if we take into acount only the local error [37, 30]

$$\varepsilon_{j+1} = -\frac{h^3}{12}\mathbf{x}''(t_j)\,,$$

after M steps the accumulated error $E(\mathbf{x}(b), h)$ for the method (2.18) is

$$E(\mathbf{x}(b),h) = -\sum_{j=1}^{M} \frac{h^3}{12}\mathbf{x}'' \approx -\frac{b-a}{12}\mathbf{x}''h^2 = \mathscr{O}(h^2)\,.$$

Again, if we perform two computations using the step sizes h and $h/2$ we obtain

$$E(\mathbf{x}(b),h) = Kh^2\,, \qquad K = \text{const}\,.$$

and

$$E(\mathbf{x}(b),\frac{h}{2}) \approx K\frac{h^2}{4} = \frac{1}{4}E(\mathbf{x}(b),h)\,.$$

Thus, if the step size in Heun's method is reduced by a factor of $1/2$, we can expect that the final global truncation error would by reduced by a factor of $1/4$.

Example 1

Use Euler's method (2.4) and Heun's method (2.18) to solve the IVP for the ODE, describing the behavior of the simple harmonic oscillator

$$\ddot{x} + \omega^2 x = 0, \qquad x(0) = 0, \quad \dot{x}(0) = v_0\,, \tag{2.18}$$

over the time interval $t \in [0,T]$, and where the frequency ω and initial velocity v_0 are given constants. The exact solution

$$x(t) = \frac{v_0}{\omega}\sin(\omega t)\,. \tag{2.19}$$

represents simple harmonic motion: sinusoidal oscillations about the equilibrium point, with a constant amplitude and a constant frequency.
First of all we rewrite Eq. (2.18) as a system of first-order ODE's:

$$\begin{aligned} &\dot{x} = y\,, \\ &\dot{y} = -\omega^2 x\,, \qquad x(0) = 0\,, \quad y(0) = v_0\,. \end{aligned} \tag{2.20}$$

We begin with analysis of system (2.20) using the ideas of *phase space* [46]. If we multiply both sides of the first equation of (2.20) by $\omega^2 x$ and both sides of the second equation of the system by y and add the two together we get the following relation

$$y\dot{y} + \omega^2 x\dot{x} = 0\,.$$

Notice that the l.h.s. of the relation above is the time derivative, so one can rewrite the last relation as

$$\frac{d}{dt}\left(\frac{1}{2}y^2 + \frac{1}{2}\omega^2 x^2\right) = 0 \Leftrightarrow \frac{1}{2}y^2 + \frac{1}{2}\omega^2 x^2 := I_1\,, \tag{2.21}$$

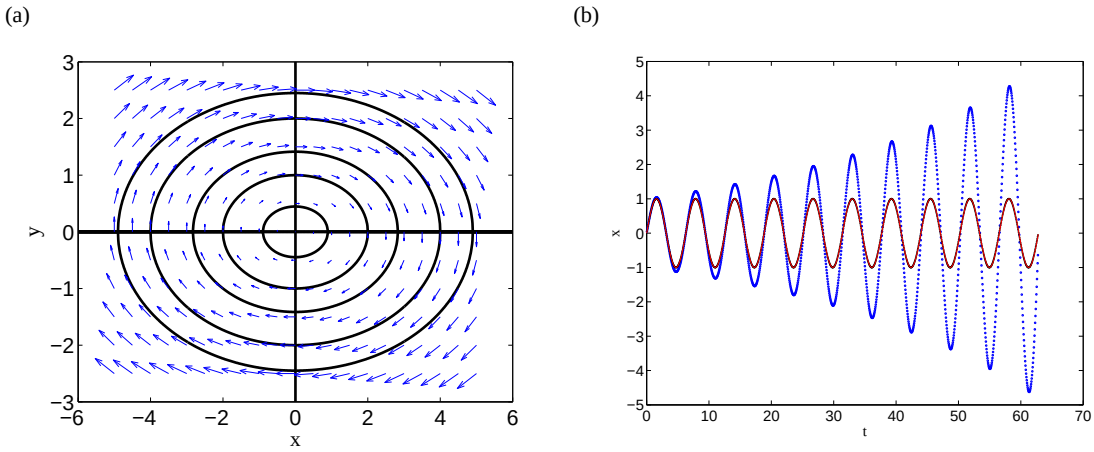

Fig. 2.3 (a) Phase diagramm for the linear oscillator (2.18), corresponding to the different values of the energy $I_1 = \{0.1, 0.5, 1, 3\}$ and $\omega = 0.5$. (b) Numerical solution of (2.18) over the interval $[0, 20\pi]$ by methods (2.4) (blue points) and (2.18) (red line) with the step size $h = 0.05$. The black curve corresponds to the exact solution of Eq. (2.18).

where $I_1 = \text{const}$ is usually called a *constant of motion* or a *first integral*, which can be interpreted as the *mechanical energy* of the system. From the geometrical point of view, one can speak about *phase curves* that form a set of ellipses in the *phase space* with coordinates (x, y). These ellipses cut the x-axis at $x = \pm\sqrt{2I_1}/\omega$ and the y-axis at $y = \pm\sqrt{2I_1}$. The origin of the phase plane corresponds to an equilibrium point of the motion (see Fig. 2.3 (a)).

Now we can solve system (2.20) numerically, using approximation methods (2.4) and (2.18) to the system (2.20). For the sake of simplisity we choose $v_0 = 1$, $\omega = 1$, so that the exact solution of (2.18) is just $x(t) = \sin(t)$, and integrate the system (2.20) over the time interval $t \in [0, 20\pi]$ with the time step $h = 0.05$. Figure 2.3 (b) shows a comparison between two methods, indicating the much better accuracy of the Heun's method (2.18).

2.4 Runge-Kutta Methods

The Runge-Kutta methods are an important family of iterative methods for the approximation of solutions of ODE's, that were develoved around 1900 by the german mathematicians C. Runge (1856–1927) and M.W. Kutta (1867–1944). We start with the considereation of the explicit methods.

Let us consider the IVP (2.1)–(2.2). The family of explicit Runge–Kutta (RK) methods of the m'th stage is given by [45, 37]

$$\mathbf{x}(t_{n+1}) := \mathbf{x}_{n+1} = \mathbf{x}_n + h\sum_{i=1}^{m} c_i k_i, \tag{2.22}$$

where

$$\begin{aligned}
k_1 &= f(t_n, \mathbf{x}_n),\\
k_2 &= f(t_n + \alpha_2 h, \mathbf{x}_n + h\beta_{21} k_1(t_n, \mathbf{x}_n)),\\
k_3 &= f(t_n + \alpha_3 h, \mathbf{x}_n + h(\beta_{31} k_1(t_n, \mathbf{x}_n) + \beta_{32} k_2(t_n, \mathbf{x}_n))),\\
&\vdots\\
k_m &= f(t_n + \alpha_m h, \mathbf{x}_n + h\sum_{j=1}^{m-1} \beta_{mj} k_j).
\end{aligned}$$

To specify a particular method, we need to provide the integer m (the number of stages), and the coefficients α_i (for $i = 2,3,...,m$), β_{ij} (for $1 \le j < i \le m$), and c_i (for $i = 1,2,...,m$). These data are usually arranged in a co-called *Butcher tableau* (after John C. Butcher) [45, 37]:

Table 2.2 The Butcher tableau.

0					
α_2	β_{21}				
α_3	β_{31}	β_{32}			
$\vdots$	$\vdots$	$\vdots$	$\ddots$		
$\vdots$	$\vdots$	$\vdots$			
α_m	β_{m1}	β_{m2}	.....	β_{mm-1}	
	c_1	c_2	.....	c_{m-1}	c_m

Examples

1. Let $m = 1$. Then

$$\begin{aligned}
k_1 &= f(t_n, \mathbf{x}_n),\\
\mathbf{x}_{n+1} &= \mathbf{x}_n + h c_1 f(t_n, \mathbf{x}_n).
\end{aligned}$$

   On the other hand, the Taylor expansion yields

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h\dot{\mathbf{x}}\big|_{t_n} + \cdots = \mathbf{x}_n + h f(t_n, \mathbf{x}_n) + \mathscr{O}(h^2) \Rightarrow c_1 = 1.$$

   Thus, the first-stage RK-method is equivalent to the explicit Euler's method (2.4). Note that the method (2.4) is of the first order of accuracy. Thus we can speak about the RK method of the first order.

2. Now consider the case $m = 2$. In this case Eq. (2.22) is equivalent to the system

$$\begin{aligned} k_1 &= f(t_n, \mathbf{x}_n)\,, \\ k_2 &= f(t_n + \alpha_2 h, \mathbf{x}_n + h\beta_{21} k_1)\,, \\ \mathbf{x}_{n+1} &= \mathbf{x}_n + h(c_1 k_1 + c_2 k_2)\,. \end{aligned} \tag{2.23}$$

Now let us write down the Taylor series expansion of $\mathbf{x}$ in the neighborhood of t_n up to the h^2 term, i.e.,

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h\left.\frac{d\mathbf{x}}{dt}\right|_{t_n} + \frac{h^2}{2}\left.\frac{d^2\mathbf{x}}{dt^2}\right|_{t_n} + \mathscr{O}(h^3)\,.$$

However, we know that $\dot{\mathbf{x}} = f(t, \mathbf{x})$, so that

$$\frac{d^2\mathbf{x}}{dt^2} := \frac{df(t,\mathbf{x})}{dt} = \frac{\partial f(t,\mathbf{x})}{\partial t} + f(t,\mathbf{x})\frac{\partial f(t,\mathbf{x})}{\partial \mathbf{x}}\,.$$

Hence the Taylor series expansion can be rewritten as

$$\mathbf{x}_{n+1} - \mathbf{x}_n = h\,f(t_n, \mathbf{x}_n) + \frac{h^2}{2}\left(\frac{\partial f}{\partial t} + f\frac{\partial f}{\partial \mathbf{x}}\right)\Bigg|_{(t_n,\mathbf{x}_n)} + \mathscr{O}(h^3)\,. \tag{2.24}$$

On the other hand, the term k_2 in the proposed RK method can also expanded to $\mathscr{O}(h^3)$ as

$$k_2 = f(t_n + \alpha_2 h, \mathbf{x}_n + h\beta_{21}k_1) = h\,f(t_n,\mathbf{x}_n) + h\,\alpha_2\left.\frac{\partial f}{\partial t}\right|_{(t_n,\mathbf{x}_n)} + h\beta_{21}\,f\left.\frac{\partial f}{\partial \mathbf{x}}\right|_{(t_n,\mathbf{x}_n)} + \mathscr{O}(h^3)\,.$$

Now, substituting this relation for k_2 into the last equation of (2.23), we achieve the following expression:

$$\mathbf{x}_{n+1} - \mathbf{x}_n = h\,(c_1 + c_2)\,f(t_n,\mathbf{x}_n) + h^2\,c_2\,\alpha_2\left.\frac{\partial f}{\partial t}\right|_{(t_n,\mathbf{x}_n)} + h^2\,c_2\beta_{21}\,f\left.\frac{\partial f}{\partial \mathbf{x}}\right|_{(t_n,\mathbf{x}_n)} + \mathscr{O}(h^3)\,.$$

Making comparision the last equation and Eq. (2.24) we can write down the system of algebraic equations for unknown coefficients

$$\begin{aligned} c_1 + c_2 &= 1\,, \\ c_2\,\alpha_2 &= \frac{1}{2}\,, \\ c_2\,\beta_{21} &= \frac{1}{2}\,. \end{aligned}$$

The system involves four unknowns in three equations. That is, one additional condition must be supplied to solve the system. We discuss two useful choices, namely

a) Let $\alpha_2 = 1$. Then $c_2 = 1/2$, $c_1 = 1/2$, $\beta_{21} = 1$. The corresponding Butcher tableau reads:

$$\begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & 1/2 & 1/2 \end{array}$$

Thus, in this case the two-stages RK method takes the form

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \frac{h}{2}\Big(f(t_n, \mathbf{x}_n) + f(t_n + h, \mathbf{x}_n + h f(t_n, \mathbf{x}_n)) \Big),$$

and is equivalent to the Heun's method (2.18), so we refer the last method to as RK-method of the second order.

b) Now let $\alpha_2 = 1/2$. In this case $c_2 = 1$, $c_1 = 0$, $\beta_{21} = 1/2$. The corresponding Butcher tableau reads:

$$\begin{array}{c|cc} 0 & & \\ 1/2 & 1/2 & \\ \hline & 0 & 1 \end{array}$$

In this case the second-order RK method (2.22) can be written as

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h\, f\big(t_n + \frac{h}{2}, \mathbf{x}_n + \frac{h}{2} f(t_n, \mathbf{x}_n)\big)$$

and is called the *RK2 method*.

RK4 Methods

One member of the family of Runge–Kutta methods (2.22) is often referred to as *RK4 method* or *classical RK method* and represents one of the solutions corresponding to the case $m = 4$. In this case, by matching coefficients with those of the Taylor series one obtains the following system of equations [30]

$$
\begin{aligned}
c_1 + c_2 + c_3 + c_4 &= 1\,, \\
\beta_{21} &= \alpha_2\,, \\
\beta_{31} + \beta_{32} &= \alpha_3\,, \\
c_2\alpha_2 + c_3\alpha_3 + c_4\alpha_4 &= \frac{1}{2}\,, \\
c_2\alpha_2^2 + c_3\alpha_3^2 + c_4\alpha_4^2 &= \frac{1}{3}\,, \\
c_2\alpha_2^3 + c_3\alpha_3^3 + c_4\alpha_4^3 &= \frac{1}{4}\,, \\
c_3\alpha_2\beta_{32} + c_4(\alpha_2\beta_{42} + \alpha_3\beta_{43}) &= \frac{1}{6}\,, \\
c_3\alpha_2\alpha_3\beta_{32} + c_4\alpha_4(\alpha_2\beta_{42} + \alpha_3\beta_{43}) &= \frac{1}{8}\,, \\
c_3\alpha_2^2\beta_{32} + c_4(\alpha_2^2\beta_{42} + \alpha_3^2\beta_{43}) &= \frac{1}{12}\,, \\
c_4\alpha_2\beta_{32}\beta_{43} &= \frac{1}{24}\,.
\end{aligned}
$$

The system involves thirteen unknowns in eleven equations. That is, two additional condition must be supplied to solve the system. The most useful choices is [37]

$$
\alpha_2 = \frac{1}{2}\,, \qquad \beta_{31} = 0\,.
$$

The corresponding Butcher tableau is presented in Table 2.3. The tableau 2.3 yields

Table 2.3 The Butcher tableau corresponding to the RK4 method.

0				
1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	1/3	1/3	1/6

the equivalent corresponding equations defining the classical RK4 method:

$$
\mathbf{x}_{n+1} = \mathbf{x}_n + \frac{h}{6}\big(k_1 + 2k_2 + 2k_3 + k_4\big), \tag{2.25}
$$

where

$$
\begin{aligned}
k_1 &= f(t_n, \mathbf{x}_n),\\
k_2 &= f(t_n + \frac{h}{2}, \mathbf{x}_n + \frac{h}{2}k_1),\\
k_3 &= f(t_n + \frac{h}{2}, \mathbf{x}_n + \frac{h}{2}k_2),\\
k_4 &= f(t_n + h, \mathbf{x}_n + hk_3).
\end{aligned}
$$

This method is reasonably simple and robust and is a good general candidate for numerical solution of ODE's when combined with an intelligent adaptive step-size routine or an embedded methods (,e.g., so-called Runge-Kutta-Fehlberg methods (RKF45)).

Remark:

Notice that except for the classical method (2.25), one can also construct other RK4 methods. We mention only so-called $3/8$-*Runge-Kutta method.* The Brutcher tableau, corresponding to this method is presented in Table 2.4.

Table 2.4 The Butcher tableau corresponding to the $3/8$- Runge-Kutta method.

0				
1/3	1/3			
2/3	-1/3	1		
1	1	-1	1	
	1/8	3/8	3/8	1/8

Geometrical interpretation of the RK4 method

Let us consider a curve $\mathbf{x}(t)$, obtained by (2.25) over a single time step from t_n to t_{n+1}. The next value of approximation $\mathbf{x}_{n+1}$ is obtained ty integrating the slope function, i.e.,

$$
\mathbf{x}_{n+1} - \mathbf{x}_n = \int_{t_n}^{t_{n+1}} f(t, \mathbf{x})\, dt\,. \tag{2.26}
$$

Now, if the Simpson's rule is applied, the approximation to the integral of the last equation reads [**?**]

$$
\int_{t_n}^{t_{n+1}} f(t, \mathbf{x})\, dt \approx \frac{h}{6}\left(f(t_n, \mathbf{x}(t_n)) + 4\, f(t_n + \frac{h}{2}, \mathbf{x}(t_n + \frac{h}{2})) + f(t_{n+1}, \mathbf{x}(t_{n+1}))\right). \tag{2.27}
$$

On the other hand, the values k_1, k_2, k_3 and k_4 are approximations for slopes of the curve $\mathbf{x}$, i.e., k_1 is the slope of the left end of the interval, k_2 and k_3 describe two estimations of the slope in the middle of the time interval, whereas k_4 corresponds to the slope at the right. Hence, we can choose $f(t_n, \mathbf{x}(t_n)) = k_1$ and $f(t_{n+1}, \mathbf{x}(t_{n+1})) = k_4$, whereas for the value in the middle we choose the average of k_2 and k_3, i.e.,

$$f(t_n + \frac{h}{2}, \mathbf{x}(t_n + \frac{h}{2})) = \frac{k_2 + k_3}{2}.$$

Then Eq. (2.26) becomes

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \frac{h}{6}\left(k_1 + \frac{4(k_2 + k_3)}{2} + k_4\right),$$

which is equivalent to the RK4 schema (2.25).

Stage versus Order

The local truncation error (2.7) for the method (2.25) can be estimated from the error term for the Simpson's rule (2.27) and equals [**?**, 30]

$$\varepsilon_{n+1} = -h^5 \frac{\mathbf{x}^{(4)}}{2880}.$$

Now we can estimate the final global error (2.13), if we suppose that only the error above is presented. After M steps the accumulated error for the RK4 method reads

$$E(\mathbf{x}(b), h) = -\sum_{k=1}^{M} h^5 \frac{\mathbf{x}^{(4)}}{2880} \approx \frac{b-a}{2880}\mathbf{x}^{(4)} h = \mathscr{O}(h^4).$$

That is, the RK4 method (2.25) is of the fourth order. Now, let us compare two appximations, obtained using the time steps h and $h/2$. For the step size h we have

$$E(\mathbf{x}(b), h) \approx K h^4,$$

with $K = \text{const}$. Hence, for the step $h/2$ we get

$$E(\mathbf{x}(b), \frac{h}{2}) = K \frac{h^4}{16} \approx \frac{1}{16} E(\mathbf{x}(b), h).$$

That is, if the step size in (2.25) is reduced by the factor of two, the global error of the method will be reduced by the factor of $1/16$.

Remark:

In general there are two ways to improve the accuracy:

1. One can reduce the time step h, i.e., the amount of steps increases;
2. The method of the higher convergency order can be used.

However, increasing of the convergency order p is reasonable only up to some limit, given by so-called *Butcher barrier* [45], which says, that the amount of stages m grows faster, as the order p. In other words, *for $m \geq 5$ there are no explicit RK methods with the convergency order $p = m$ (the corresponding system is unsolvable)*. Hence, in order to reach convergency order five one needs six stages. Notice that further increasing of the stage $m = 7$ leads to the convergency order $p = 5$ as well.

2.4.1 Adaptive stepsize control and embedded methods

As mentioned above, one way to guarantee accuracy in the solution of (2.1)–(2.1) is to solve the problem twice using step sizes h and $h/2$. To illustrate this approach, let us consider the RK method of the order p and denote an exact solution at the point $t_{n+1} = t_n + h$ by $\widetilde{\mathbf{x}}_{n+1}$, whereas $\mathbf{x}_1$ and $\mathbf{x}_2$ represent the approximate solutions, corresponding to the step sizes h and $h/2$. Now let us perform one step with the step size h and after that two steps each of size $h/2$. In this case the true solution and two numerical approximations are related by

$$\begin{aligned}
\widetilde{\mathbf{x}}_{n+1} &= \mathbf{x}_1 + C h^{p+1} + \mathscr{O}(h^{p+2}), \\
\widetilde{\mathbf{x}}_{n+1} &= \mathbf{x}_2 + 2C \left(\frac{h}{2}\right)^{p+1} + \mathscr{O}(h^{p+2}).
\end{aligned}$$

That is,

$$|\mathbf{x}_1 - \mathbf{x}_2| = C h^{p+1} \left(1 - \frac{1}{2^p}\right) \Leftrightarrow C = \frac{|\mathbf{x}_1 - \mathbf{x}_2|}{(1 - 2^{-p}) h^{p+1}}.$$

Substituing the relation for C in the second estimate for the true solution we get

$$\widetilde{\mathbf{x}}_{n+1} = \mathbf{x}_2 + \varepsilon + \mathscr{O}(h^{p+2}),$$

where

$$\varepsilon = \frac{|\mathbf{x}_1 - \mathbf{x}_2|}{2^p - 1}$$

can be considered as a convenient *indicator* of the truncation error. That is, we have improved our estimate to the order $p + 1$. For example, for $p = 4$ we get

$$\widetilde{\mathbf{x}}_{n+1} = \mathbf{x}_2 + \frac{|\mathbf{x}_1 - \mathbf{x}_2|}{15} + \mathscr{O}(h^6).$$

This estimate is accurate to fifth order, one order higter than with the original step h. However, this method is not efficient. First of all, it requires a significant amount

of computation (we should solve the equation three times at each time step). The second point is, that we have no possibility to control the truncation error of the method (higher order means not always higher accuracy).
However we can use an estimate ε for the *step size control*, namely we can compare ε with some *desired accuracy* ε_0 (see Fig 2.4).

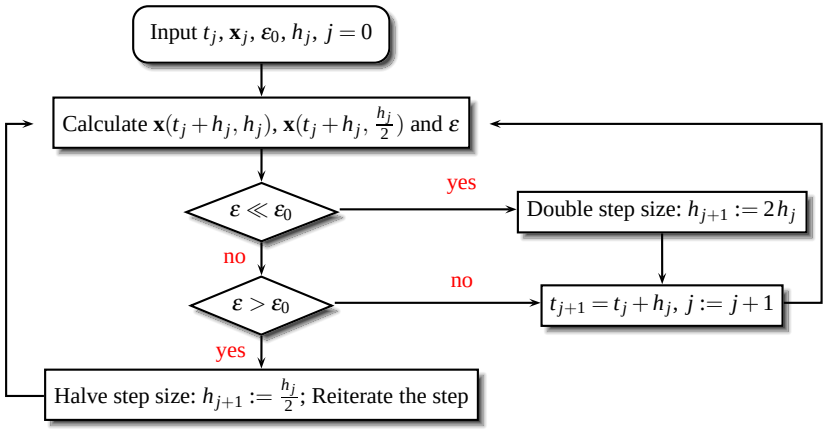

Fig. 2.4 Flow diagramm of the step size control by use of the step doubling method.

Alternatively, using the estimate ε, we can try to formulate the following problem of the *adaptive step size control*, namely: Using the given values $\mathbf{x}_j$ and t_j, find the largest possible step size h_{new}, so that the truncation error after the step with this step size remains below some given desired accuracy ε_0, i.e,

$$Ch_{new}^{p+1} \le \varepsilon_0 \Leftrightarrow \left(\frac{h_{new}}{h}\right)^{p+1} \frac{|\mathbf{x}_1 - \mathbf{x}_2|}{1-2^{-p}} \le \varepsilon_0 .$$

That is,

$$h_{new} = h\left(\frac{\varepsilon_0}{\varepsilon}\right)^{1/p+1} .$$

Then if the two answers are in close agreement, the approximation is accepted. If $\varepsilon > \varepsilon_0$ the step size has to be decreased, whereas the relation $\varepsilon < \varepsilon_0$ means, that the step size has to be increased in the next step.

Notice that because our estimate of error is not exact, we should put some "safety" factor $\beta \simeq 1$ [45, 37]. Usually, $\beta = 0.8, 0.9$. The flow diagramm, corresponding to the the adaptive step size control is shown on Fig. 2.5

Notice one additional technical point. The choise of the desired error ε_0 depends on the IVP we are interested in. In some applications it is convinient to set ε_0 propotional to h [37]. In this case the exponent $1/p+1$ in the estimate of the new time step is no longer correct (if h is reduced from a too-large value, the new predicted value h_{new} will fail to meet the desired accuracy, so instead of $1/p+1$ we should scale with $1/p$ (see [37] for details)). That is, the optimal new step size can be written as

$$h_{new} = \begin{cases} \beta h\left(\frac{\varepsilon_0}{\varepsilon}\right)^{1/p+1}, & \varepsilon \ge \varepsilon_0, \\ \beta h\left(\frac{\varepsilon_0}{\varepsilon}\right)^{1/p}, & \varepsilon < \varepsilon_0, \end{cases} \tag{2.28}$$

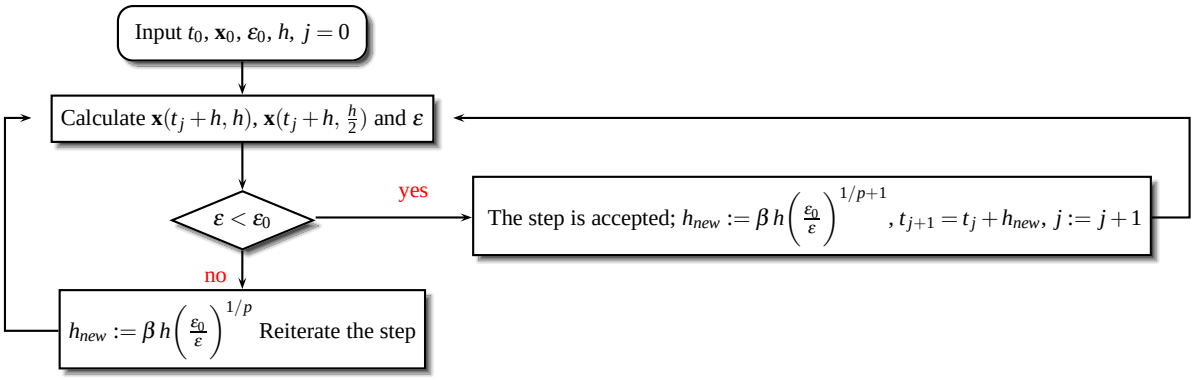

Fig. 2.5 Flow diagramm of the adaptive step size control by use of the step doubling method.

where β is a "safety" factor.

Runge-Kutta-Fehlberg method

The alternative stepsize adjustment algorithm is based on the *embedded Runge-Kutta formulas*, originally invented by Fehlberg and is called *the Runge-Kutta-Fehlberg methods (RKF45)* [45, **?**]. At each step, two different approximations for the solution are made and compared. Usually an fourth-order method with five stages is used together with an fifth-order method with six stages, that uses all of the points of the first one. The general form of a fifth-order Runge-Kutta with six stages is

$$\begin{aligned} k_1 &= f(t,\mathbf{x}), \\ k_2 &= f(t+\alpha_2 h, \mathbf{x}+h\beta_{21}k_1), \\ &\vdots \\ k_6 &= f(t+\alpha_6 h, \mathbf{x}+h\sum_{j=1}^{5}\beta_{6j}k_j)\,. \end{aligned}$$

The embedded fourth-order formula is

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h\sum_{i=1}^{6} c_i k_i + \mathscr{O}(h^5)\,.$$

And a better value for the solution is determined using a Runge-Kutta method of fifth-order:

$$\mathbf{x}^*_{n+1} = \mathbf{x}_n + h\sum_{i=1}^{6} c^*_i k_i + \mathscr{O}(h^6)$$

The two particlular choises of unknown parametrs of the method are given in Tables 2.5–2.6.

The error estimate is

$$\varepsilon = |\mathbf{x}_{n+1} - \mathbf{x}^*_{n+1}| = \sum_{i=1}^{6}(c_i - c^*_i)k_i\,.$$

Table 2.5 Fehlberg parameters of the Runge-Kutta-Fehlberg 4(5) method.

1/4	1/4					
3/8	3/32	9/32				
12/13	1932/2197	-7200/2197	7296/2197			
1	439/216	-8	3680/513	-845/4104		
1/2	-8/27	2	-3544/2565	1859/4104	-11/40	
	25/216	0	1408/2565	2197/4104	-1/5	
	16/135	0	6656/12825	28561/56430	-9/50	2/55

Table 2.6 Cash-Karp parameters of the Runge-Kutta-Fehlberg 4(5) method.

1/5	1/5					
3/10	3/40	9/40				
3/5	3/10	-9/10	6/5			
1	-11/54	5/2	-70/27	35/27		
7/8	1631/55296	175/512	575/13828	44275/110592	253/4096	
	37/378	0	250/621	125/594	512/1771	
	2825/27648	0	18575/48384	13525/55296	277/14336	1/4

As was mentioned above, if we take the current step h and produce an error ε, the corresponding "optimal" step h_{opt} is estimated as

$$h_{opt} = \beta h \left(\frac{\varepsilon_{tol}}{\varepsilon} \right)^{0.2},$$

where ε_{tol} is a desired accuracy and β is a "safety" factor, $\beta \simeq 1$. Then if the two answers are in close agreement, the approximation is accepted. If $\varepsilon > \varepsilon_{tol}$ the step size has to be decreased, whereas the relation $\varepsilon < \varepsilon_{tol}$ means, that the step size are to be increased in the next step.
Using Eq. (2.28), the optimal step can be often written as

$$h_{opt} = \begin{cases} \beta h \left(\frac{\varepsilon_{tol}}{\varepsilon} \right)^{0.2}, & \varepsilon \geq \varepsilon_{tol}, \\ \beta h \left(\frac{\varepsilon_{tol}}{\varepsilon} \right)^{0.25}, & \varepsilon < \varepsilon_{tol}, \end{cases}$$

2.4.2 Examples

2.4.2.1 Lotka-Volterra competition model

The Lotka–Volterra competition equations are a simple model of the population dynamics of species competing for some common resource. For given two populations with sizes x and y the model equations are [**?**]

$$\begin{aligned} \dot{x} &= ax(b - x - cy), \\ \dot{y} &= dy(e - y - fx), \end{aligned} \tag{2.29}$$

Here, positive constant c represents the effect species two has on the population of species one and positive constant f describes the effect species one has on the population of species two. Let us analyse the system (2.29) using parameters

$$a = 0.004,\, b = 50,\, c = 0.75,\, d = 0.001,\, e = 100,\; f = 3\,.$$

Fixed points

Equations (2.29) have four fixed points (x^*, y^*):

$$(0,0)\,, \quad (0,e) = (0,100)\,, \quad (b,0) = (50,0)\,, \quad \left(\frac{b-ce}{1-fc}, \frac{bf-1}{cf-1}\right) = (20,40).$$

Linear stability

In order to analyse the linear stability of (2.29) one derives the corresponding Jacobian

$$J = \begin{pmatrix} ab - 2ax^* - acy^* & -acx^* \\ -dfy^* & de - 2y^* - dfx^* \end{pmatrix}$$

Now one can calculate J for all fixed point values (x^*, y^*) and derive the eigenvalues (λ_1, λ_2) (see Table 2.7).

Table 2.7 Eigenvalues and linear stability analysis for four fixed points of the system (2.29)

(x^*, y^*)	(λ_1, λ_2)	stability
$(0, 0)$	$(0.2, 0.1)$	-
$(0, 100)$	$(-0.1, -0.1)$	+
$(50, 0)$	$(-0.2, -0.05)$	+
$(20, 40)$	$(0.027, -0.14)$	-

Numerical results

Table (2.29) indicates that the trivial fixed point, coresponding to the case, that both populations die out, is unstable. Furthermore, the fixed point $(20, 40)$, corresponding to the case, that both populations survive, is unstable too. That is, both populations will neither die out or survive. Which population will survive (or die out) depends on initial conditions (see Fig. 2.6).

2.4.2.2 Predator-Prey Model

Now let us consider another model of Lotka-Volterra type, which describes prey's (x) and predator's (y) population dynamics in the presence of one another. Equations are:

$$\dot{x} = ax - bxy, \qquad (2.30)$$
$$\dot{y} = cxy - dy.$$

Here $a > 0$ and $c > 0$ are prey's and predator's growth rates whereas $d > 0$ and $b > 0$ describe prey's and predator's death rates respectively.
A typical numerical coefficients are $a = 2$, $b = 0.02$, $c = 0.0002$, $d = 0.8$.
The model (2.30) predicts a cyclical relationship between predator and prey numbers. To see this effect, first we find two fixed points (x^*, y^*) of the system. The fixed points are

$$(0,0), \qquad \left(\frac{d}{c}, \frac{a}{b}\right) = (4 \cdot 10^3, 10^2).$$

The Jacobian of (2.30) is

$$J = \begin{pmatrix} a - by^* & -bx^* \\ cy^* & cx^* - d \end{pmatrix}$$

Now one can calculate eigenvalues for both fixed points (see Table 2.8). Furthermore, one can also

Table 2.8 Eigenvalues and linear stability analysis for four fixed points of the system (2.29)

(x^*, y^*)	(λ_1, λ_2)	stability
$(0,0)$	$(a, -d)$	no, saddle point
$\left(\frac{d}{c}, \frac{a}{b}\right)$	$(i\sqrt{ad}, -i\sqrt{ad})$	neutral stable

calculate the first integral V of the system (2.30):

$$c\dot{x} + b\dot{y} - \frac{d\dot{x}}{x} - \frac{a\dot{y}}{y} = 0 \Rightarrow V := cx + by - d\ln(x) - a\ln(y) = \text{const}.$$

The total derivative of V with respect to time reads

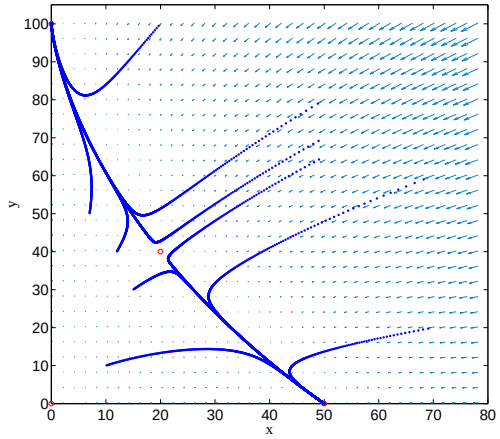

Fig. 2.6 Numerical solution of (2.29) over the time interval $t \in [0, 150]$ on the phase plane (x, y) by the classical RK4 method with the step size $h = 0.025$ for different initial values. Open red circles denote unstable fixed points, whereas filled red circles represent stable fixed points.

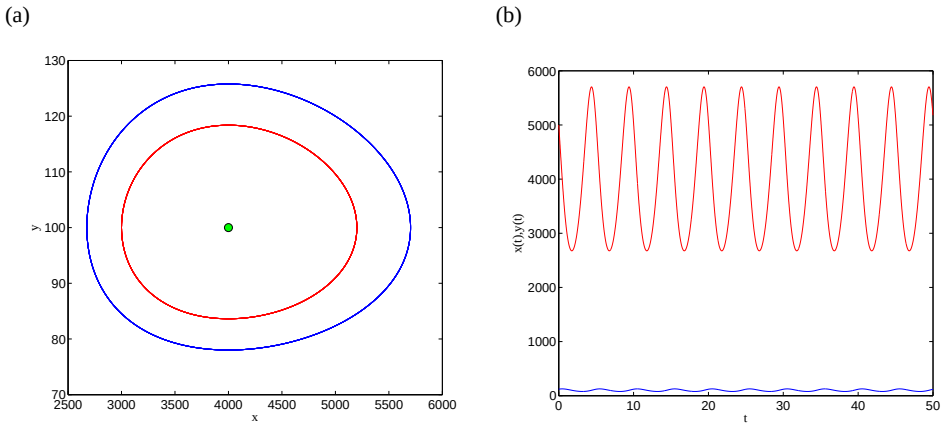

Fig. 2.7 Numerical solution of the system (2.30), calculated with RK4 method. (a) Solutions on the phase plane, corresponding to two different initial conditions: $(5 \cdot 10^3, 120)$ and $(3 \cdot 10^3, 10^2)$. (b) A cyclical relationship between predator and prey numbers, calculated for the initial condition $(5 \cdot 10^3, 120)$.

$$\frac{dV}{dt} = \left(c - \frac{d}{x}\right) x \left(a - by\right) + \left(b - \frac{a}{y}\right) y \left(cx - d\right) = 0$$

That is, solutions of (2.30) can not leave levels of V. This is illustrated on Fig. 2.7 (a), where two numerical solutions, corresponding to two different initial conditions $(5 \cdot 10^3, 120)$ and $(3 \cdot 10^3, 10^2)$ are presented. The green point denotes the neutral stable fixed point. Oscillations of both populations, corresponding to the initial value $(5 \cdot 10^3, 120)$ is presented on Fig. 2.7 (b).
Now suppose that prey's growth rate is periodic in time, e.g,

$$a := a\left(1 + \varepsilon \sin(\omega t)\right),$$

where $\varepsilon \in [0, 1)$ and let be $\omega = \pi$. In this case, depending on control parameter ε, quasiperiodic or even chaotic behaviour can be expected. Figure 2.8 illustates an example of quasiperiodic behaviour.

2.4.2.3 Forced oscillations: Pohl's pendulum

The Pohl's wheel is a rotating pendulum with electromagnetic brake, spiral spring and variable stimulation, which can demonstrate harmonic oscillations as well as chaotic motion. The equation of motion of the wheel reads

$$J\ddot{\varphi} + K\dot{\varphi} + D\varphi - N\sin(\varphi) = \hat{F}\sin(\omega t + \Omega). \tag{2.31}$$

Here φ denotes the rotation angle, J is a inertia moment of the pendulum about the axis of rotation, K is a damping constant, D stays for the torque per unit angel, and, finally, $N = mgr\sin(\varphi)$ is a projection of the variable stimulation's moment (m is a external mass, r is a radius of the wheel). In addition, $\hat{F}\sin(\omega t + \Omega)$ is an external forsing of the amplitude $\hat{F}$, frequency ω and the free phase Ω.

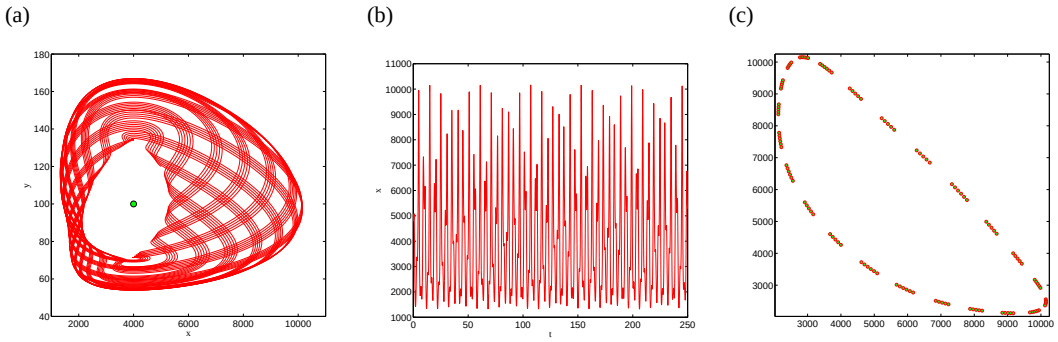

Fig. 2.8 Numerical solution for (2.30) calculated with RK4 method for the case $\varepsilon = 0.4$. Initial condition is $(5 \cdot 10^3, 120)$. (a) Solutions on phase plane; (b) Quasiperiodic oscillations of the preys population; (c) The first return map.

In order to solve Eq. (2.31) numerically we rewrite it to a system of first order ODE's. Substitution $x := \varphi$, $y := \dot{\varphi}$, $z := t$ leads to the system

$$\begin{aligned} \dot{x} &= y\,, \\ \dot{y} &= -ay - bx + c\sin(x) + d\sin(\omega z)\,, \\ \dot{z} &= 1\,, \end{aligned} \tag{2.32}$$

with

$$a := \frac{K}{J}\,, \quad b := \frac{D}{J}\,, \quad c := \frac{N}{J}\,, \quad d := \frac{\hat{F}}{J}\,.$$

We solve the system (2.32) with the classical RK4 method (2.25) with the time step $h = 0.025$ over the time interval $t \in [0, 150]$. Other parameters are

$$a = 0.799\,, \quad b = 9.44\,, \quad c = 14.68, \quad d = 2.1\,.$$

Furthermore, we use the frequency of the external forcing ω as a control parameter. We start at $\omega = 2.5$. The result is shown on Fig. 2.9 (a)-(c). Figure 2.9 (a) shows the solution of (2.32) on the phase space (φ, φ'). One can see, that the solution corresponds to forced oscillations with the period one (see Fig. 2.9 (b) as well). Period one oscillations can also be recognised from the first return map (Fig. 2.9 (c)). Now we increase the control parameter ω to $\omega = 2.32$. The results can be seen on Fig. 2.10 (a)–(c). In this case the system oscillates between two values, so one can speak about period two oscillations (or about periode-doubling bifurcation). Further increasing of ω leads to second periode-doubling bifurcation and period four oscillation sets in (see Fig. 2.11 (a)–(c)). Finally, we increase ω further and chaotic oscillations can be observed (see Fig. 2.12 (a)–(c)). The first return map, shown on Fig. (2.12) (c) indicates the structure of the chaotic motion: the n'th maximal value of φ is predicted by the $n-1$'th one.

2.4.2.4 Lorenz system

Let us consider the so-called *Lorenz equations*

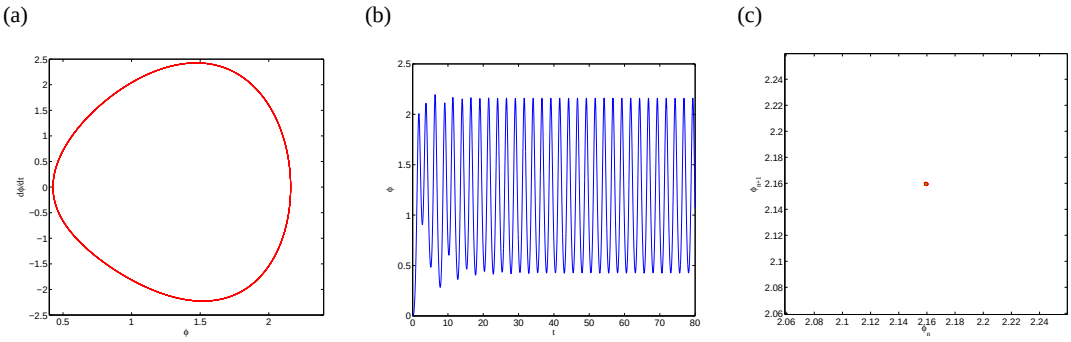

Fig. 2.9 Solution of Eq. (2.31) corresponding to $\omega = 2.5$. (a) Solution on the phase plane. (b) One period oscillations on the φ, t plot. (c) The first return map.

$$\begin{aligned} \dot{x} &= \sigma(y-x), \\ \dot{y} &= rx - x - xz, \\ \dot{z} &= xy - bz. \end{aligned} \tag{2.33}$$

Here $\sigma > 0$ is Prandtl number, $r > 0$ stays for normalized Rayleigh number, whereas $b > 0$ is a geometric factor. The function $x(t)$ is propotional to the intensity of convection motion, $y(t)$ is propotional to the temperature difference between ascending and descending currents and $z(t)$ is propotional to the distortion of vertical tmperature profile from the linear one.
This system was investigated by Ed Lorenz in 1963. Its purpose was to provide a simplifed model of atmospheric convection [**?**, 46].

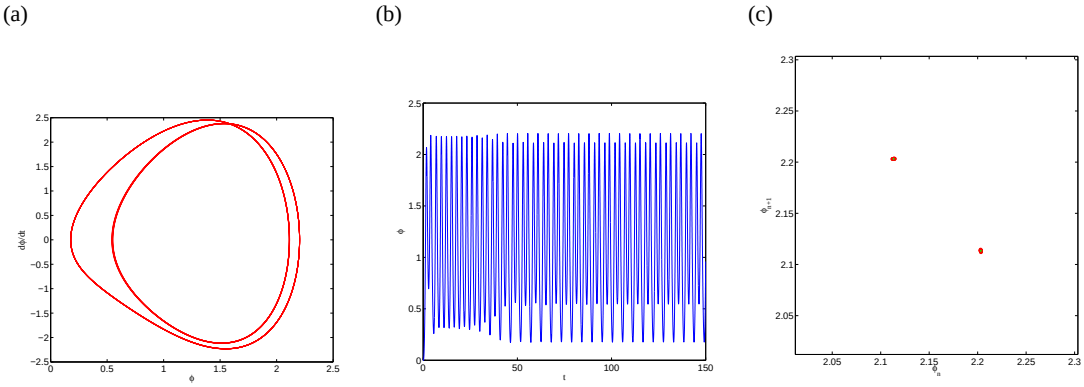

Fig. 2.10 Solution of Eq. (2.31), corresponding to the periode-doubling bifurcation for $\omega = 2.32$. (a) Solution on the phase plane. (b) Two period oscillations on the φ, t plane . (c) The first return map.

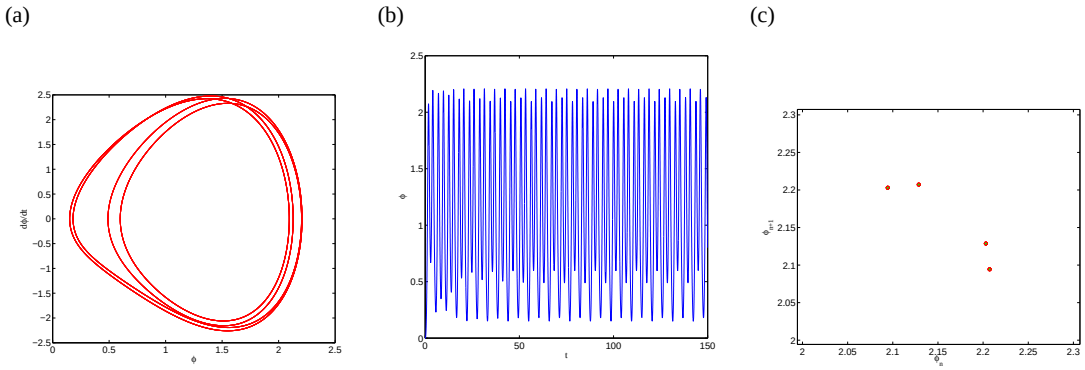

Fig. 2.11 Solution of Eq. (2.31), corresponding to the second periode-doubling bifurcation for $\omega = 2.3$. (a) Solution on the phase plane. (b) Period four oscillations on the φ, t plane. (c) The first return map.

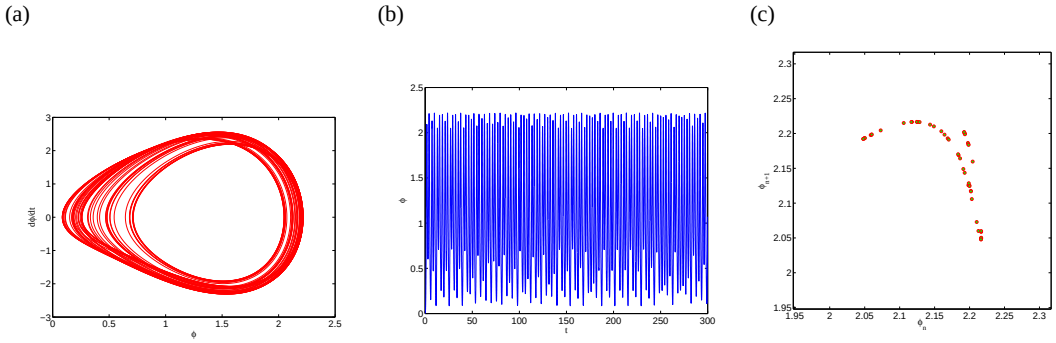

Fig. 2.12 Solution of Eq. (2.31), corresponding to the chaotic oscillation for $\omega = 2.25$.(a) Solution on the phase plane. (b) Chaotic oscillations on the φ, t plane. (c) The first return map, indicationg the chaotic regime.

Symmetry

The system (2.33) admits a symmetry

$$(x, y, z) \rightarrow (-x, -y, z)\,.$$

Fixed Points

The fixed points (x^*, y^*, z^*) are

(a) $x^* = y^* = z^* = 0$ – corresponds to the state of no convection;
(b) $C^+ = (\sqrt{b(r-1)}, \sqrt{b(r-1)}, r-1)$ and $C^- = (-\sqrt{b(r-1)}, -\sqrt{b(r-1)}, r-1)$ – correspond to the state of steady convection. Note that both solutions exist only for $r > 1$.

Linear Stability

The Jacobian of the system (2.33) reads

$$J(x^*, y^*, z^*) = \begin{pmatrix} -\sigma & \sigma & 0 \\ \sigma - z^* & -1 & -x^* \\ y^* & x^* & -b \end{pmatrix}$$

(a) The trivial solution $(x^*, y^*, z^*) = (0, 0, 0)$: In this case the matrix J can be written as 2×2 matrix,

$$J_0 = \begin{pmatrix} -\sigma & \sigma \\ r & -1 \end{pmatrix}$$

as an linearized equation for $z(t)$ is

$$\dot{z} = -bz$$

decoupled. The stability of (2.33) can be determined using the trace and the deteminant of J_0:

$$\mathrm{Sp}(J_0) = -\sigma - 1 < 0\,, \qquad \det(J_0) = \sigma(1-r) > 0 \Rightarrow r < 1\,.$$

That is, the trivial solution is stable if $r < 1$.

(b) Stability of C^+ and C^-: Consider the case $r > 1$, so both nontrivial solutions exist. The characteristic polynomial reads

$$\lambda^3 + (\sigma + b + 1)\lambda^2 + (r + \sigma) b \lambda + 2b\sigma(r-1) = 0\,.$$

The eigenvalues consist of one real negative root and a pair of complex conjugate roots [**?**]. The complex roots can be found using the ansatz $\lambda = i\omega$. Substitution into characteristic polynomial leads to the expression for the crtical Rayleigh number r_H

$$r_H = \frac{\sigma + b + 3}{\sigma - b - 1}\,, \qquad \sigma > b + 1.$$

The third eigenvalue λ_3 can be found as

$$\lambda_3 = -(\sigma + b + 1) < 0$$

That is, the nontrivial solutions C^+ and C^- are stable for

$$1 < r < r_H\,, \qquad \sigma > b + 1.$$

The nontrivial solutions loose stability at r_H via the Hopf bifurcation. One can show that this bifurcation is subcritical [**?**]. That is, the limit circles are unstable and exist only for $r < r_H$.

Time behaviour for different r's

In his study, Lorenz chose the parameter values

$$b = \frac{8}{3}\,, \qquad \sigma = 10\,.$$

With this choise the steady state becomes unstable at

$$r = r_H = \frac{470}{19} \approx 27.74\ldots.$$

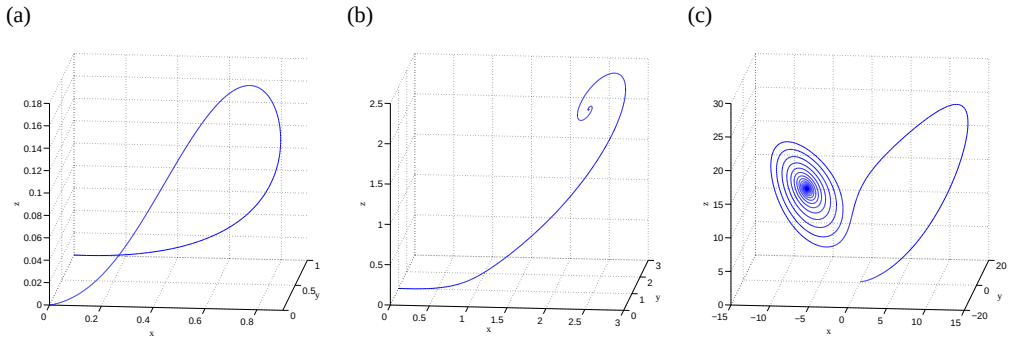

Fig. 2.13 Solutions of the Lorenz equations (2.33), corresponding to different values of r. (a) $r = 0.5$ - the origin is stable; (b) $r = 3$ - the origin is unstable. All trajectories converge to one of stable nontrivial fixed points C^+ or C^-; (c) $r = 16$ -the basin of attraction around C^+ and C^- are no longer distinct.

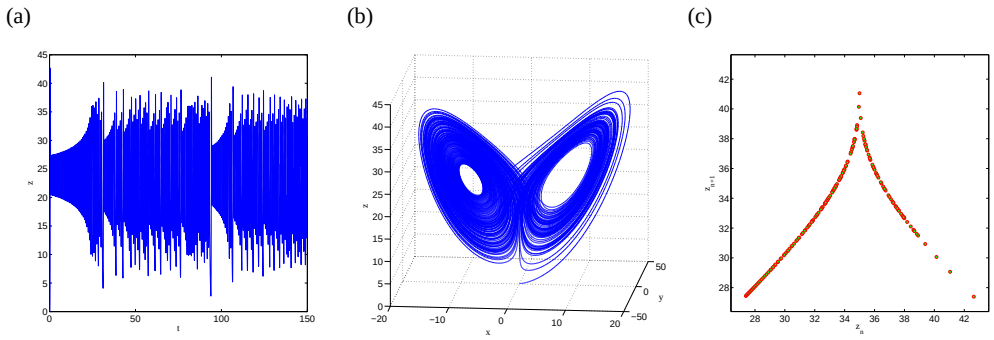

Fig. 2.14 (a) Solution of the Lorenz equations (2.33) on (t, z) plane, computed at $r = 26$. (b) Solution of (2.33) at $r = 26$ on the tree-dimensional phase space. (c) The Lorenz map.

The initial value was $(0, 1, 0)$.
Now let us summarize what happens to the solutions of (2.33) as r is increased [**?**, 46]:

- $0 < r < 1$: The origin is stable node (see Fig. 2.13 (a)).
- $1 < r < 24.74$: The origin becomes unstable and bifurcates into a pair of solutions C^+ and C^-. All trajectories converge to either one or another point (see Fig. 2.13 (b)). At $r \approx 13.926$ the origin becomes a homoclinic point, i.e., beyond this point trajectories can cross forward and backward between C^+ and C^- befor settle down to them (see Fig. 2.13 (c)).
- $r = 24.74$: Both C^+ and C^- become unstable via subcritical Hopf bifurcation.
- $r > 24.74$: After initial transient the solution settes into irregular oscillation and is aperiodic (see Fig. 2.14 (a)). On the phase space, the time spent wandering near sets around C^+ and C^- becomes infinite and the set becomes *a strange attractor* (see Fig. 2.14 (b)).

Lorenz map

Lorenz found a way to analyse the dynamics on the strange attractor. He has considered a projection of the three-dimensional phase space on the (t, z) plane. The idea was that if we consider the n'th local maximum of the function $z(t)$, z_n, it should predict z_{n+1}. To check this, one can estimate the local maxima of the function $z(t)$ and plot z_{n+1} versus z_n. The resulting function, presented on Fig. 2.14 (c) is now called *the Lorenz map*.

2.5 Predictor-Corrector Methods

The Runge-Kutta methods, introduced in Section 2.4 are referred to as *single-step methods*, because they use only the information from one previous point to evolve from x_n to x_{n+1}. In addition to single-step methods, there is also a broad class of so-called *multi-step* integration methods, which use information at more than one previous point to estimate solution at next point.

The main advantages of Runge-Kutta methods (2.22) are that they are easy to implement, rather stable, and "self-starting", (i.e., we do not have to treat the first few steps taken by a single-step integration method as special cases). On the other hand, the primary disadvantage of Runge-Kutta methods (2.22) compared to multi-step methods is that they require significantly more computer time than multi-step methods of comparable accuracy. In addition, the local truncation error of a milti-step method can be determinated and a correction term can be included, which improves the accuracy of the numarical approximation *at each step* [30]. One of the example of multi-step methods are the various *predictor-corrector methods*, which proceed by extrapolating a polynomial fit to the derivative from the previous points to the new point (the predictor step), then using this to interpolate the derivative (the corrector step) [37].

2.5.1 The Adams-Bashforth-Moulton method

Again, let us consider IVP (2.1)–(2.2). Integrating both sides of Eq. (2.1) over one time step from t_n to t_{n+1} we obtain the *exact* relation (2.14):

$$\mathbf{x}(t_{n+1}) - \mathbf{x}(t_n) = \int_{t_n}^{t_{n+1}} f(t, \mathbf{x}(t))\, dt\,.$$

Now a numerical integration method can be used to approximate the definite integral in the last equation. The Adams-Bashforth-Moulton method is a multi-step method that proceeds in two steps [30, 37]. The first step is called *the Adams-Bashforth predictor*. The predictor uses the Lagrange polynomial approximation for the function $f(t, \mathbf{x}(t))$ based on the nodes $(t_{n-3},\, f_{n-3})$, $(t_{n-2},\, f_{n-2})$, $(t_{n-1},\, f_{n-1})$ and $(t_n,\, f_n)$. After integrating over the interval $[t_n, t_{n+1}]$ the predictor reads

$$p_{n+1} = \mathbf{x}_n + \frac{h}{24}\left(-9\, f_{n-3} + 37\, f_{n-2} - 59\, f_{n-1} + 55\, f_n\right). \tag{2.34}$$

The second step is *the Adams-Moulton corrector* and is developed similary. A seecond Lagrange polynomial for the function $f(t, \mathbf{x}(t))$ is constracted. In this case, it based on the points $(t_{n-2},\, f_{n-2})$, $(t_{n-1},\, f_{n-1})$, $(t_n,\, f_n)$ and the new point $(t_{n+1},\, f_{n+1}) = (t_{n+1},\, f(t_{n+1},\, p_{n+1}))$. After integrating over the interval $[t_n, t_{n+1}]$ the following relation for the corrector is obtained

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \frac{h}{24}\left(f_{n-2} - 5\, f_{n-1} + 19\, f_n + 9\, f_{n+1} \right). \tag{2.35}$$

Notice that the method (2.34)–(2.35) is not "self-starting", i.e., four initial points $(t_n, \mathbf{x}_n)$, $n = 0, 1, 2, , 3$ must be given in order to estimate the points $(t_n, \mathbf{x}_n)$ for $n \geq 4$.

Error Estimation

The local truncation error for both predictor (2.34) and corrector (2.35) terms are of the order $\mathscr{O}(h^5)$, namely [30]

$$\begin{aligned} \mathbf{x}(t_{n+1}) - p_{n+1} &= \frac{251}{720}\mathbf{x}^{(5)} h^5, \\ \mathbf{x}(t_{n+1}) - \mathbf{x}_{n+1} &= -\frac{19}{720}\mathbf{x}^{(5)} h^5. \end{aligned}$$

That is, for small values of h one can eliminate terms with fifth derivative and the error estimate reads

$$\mathbf{x}(t_{n+1}) - \mathbf{x}_{n+1} \approx \frac{-19}{270}\left(\mathbf{x}_{n+1} - p_{n+1} \right). \tag{2.36}$$

Equation (2.36) gives an estimation of the local truncation error based on the *two computed values* p_{n+1} and $\mathbf{x}_{n+1}$, but $\mathbf{x}^{(5)}$.

Example 1

Solve an IVP

$$\dot{x} = t^2 - x, \qquad x_0 := x(0) = 1 \tag{2.37}$$

over the time interval $t \in [0, 5]$ with the Adams-Bashforth-Moulton method (2.34)–(2.35) using the time step $h = 0.05$. The three starting x_1, x_2 and x_3 values can be calculated via the classical RK4 method. The exact solution of the problem is [30]

$$x(t) = t^2 - 2t + 2 - e^{-t}.$$

The result of the calculation is presented on Fig. 2.15.

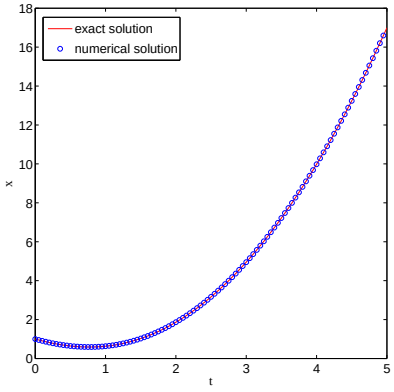

Fig. 2.15 Numerical solution of Eq. (2.37) over the interval $[0, 5]$ by the Adams-Bashforth-Moulton method (2.34)–(2.35) with the step size $h = 0.05$ (blue open circles). The exact solution of the problem is depicted by the red line.

Chapter 3
Boundary Value Problem

A boundary value problem (BVP) is a problem, typically an ODE or a PDE, which has values assigned on the physical boundary of the domain in which the problem is specified. Let us consider a genearal ODE of the form

$$\mathbf{x}^{(n)} = f(t, \mathbf{x}, \mathbf{x}', \mathbf{x}'', \cdots, \mathbf{x}^{(n-1)}), \qquad t \in [a, b] \tag{3.1}$$

At $t = a$ and $t = b$ the solution is supposed to satisfy

$$\begin{aligned} r_1(\mathbf{x}(a)\mathbf{x}'(a), \cdots, \mathbf{x}^{(n-1)}(a), \mathbf{x}(b)\mathbf{x}'(b), \cdots, \mathbf{x}^{(n-1)}(b)) = 0\,, \\ \vdots \\ r_n(\mathbf{x}(a)\mathbf{x}'(a), \cdots, \mathbf{x}^{(n-1)}(a), \mathbf{x}(b)\mathbf{x}'(b), \cdots, \mathbf{x}^{(n-1)}(b)) = 0\,. \end{aligned} \tag{3.2}$$

The resulting problem (3.1)–(3.2) is called a *two point boundary value problem* [**?**].
In order to be useful in applications, a BVP (3.1)–(3.2) should be *well posed*. This means that given the input to the problem there exists a unique solution, which depends continuously on the input. However, questions of existence and uniqueness for BVPs are much more difficult than for IVPs and there is no general theory.

3.1 Single shooting methods

3.1.1 Linear shooting method

Consider a linear two-point second-order BVP of the form

$$x''(t) = p(t)x'(t) + q(t)x(t) + r(t)\,, \qquad t \in [a, b] \tag{3.3}$$

with

$$x(a) = \alpha\,, \qquad x(b) = \beta\,.$$

The main idea of the method is to reduce the solution of the BVP (3.3) to the solution of an initial value problem [45, 30]. Namely, let us consider two special IVPs for two functions $u(t)$ and $v(t)$. Suppose that $u(t)$ is a solution of the IVP

$$u''(t) = p(t)\,u'(t) + q(t)\,u(t) + r(t)\,, \qquad u(a) = \alpha\,, \qquad u'(a) = 0$$

and $v(t)$ is the unique solution to the IVP

$$v''(t) = p(t)\,v'(t) + q(t)\,v(t)\,, \qquad v(a) = 0\,, \qquad v'(a) = 1\,.$$

Then the linear combination

$$x(t) = u(t) + c\,v(t)\,, \qquad c = \text{const}\,. \tag{3.4}$$

is a solution to BVP (3.3). The unknown constant c can be found from the boundary condition on the right end of the time interval, i.e.,

$$x(b) = u(b) + c\,v(b) = \beta \;\Rightarrow c = \frac{\beta - u(b)}{v(b)}\,.$$

That is, if $v(b) \neq 0$ the unique solution of (3.3) reads

$$x(t) = u(t) + \frac{\beta - u(b)}{v(b)}\,v(t)\,.$$

Example 1

Let us solve a BVP [30]

$$\begin{aligned} x''(t) &= \frac{2t}{1+t^2}\,x'(t) - \frac{2}{1+t^2}\,x(t) + 1\,, \\ x(0) &= 1.25, \qquad x(1) = -0.95\,. \end{aligned} \tag{3.5}$$

over the time interval $t \in [0, 4]$ using the linear shooting method (3.4).
According to Eq. (3.4) the solution of this equation has the form

$$x(t) = u(t) - \frac{0.95 + u(4)}{v(4)}\,v(t)\,,$$

where $u(t)$ and $v(t)$ are solutions of two IVPs

$$u''(t) = \frac{2t}{1+t^2}\,u'(t) + \frac{2}{1+t^2}\,u(t) + 1\,, \qquad u(0) = 1.25\,, \qquad u'(0) = 0$$

and

$$v''(t) = \frac{2t}{1+t^2}\,v'(t) + \frac{2}{1+t^2}\,v(t)\,, \qquad v(0) = 0\,, \qquad v'(0) = 1\,.$$

Numerical solution of the problem 3.5 as well as both dunctions $u(t)$ and $v(t)$ are presented on Fig. 3.1

3.1.2 Single shooting for general BVP

For a general BVP for a second-order ODE, the simple shooting method is stated as follows: Let

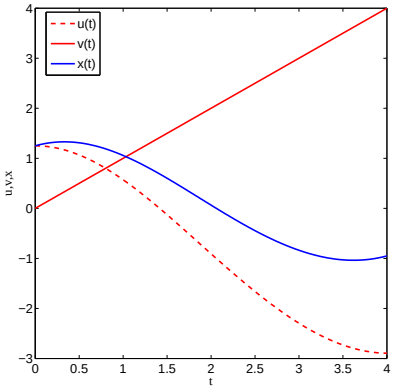

Fig. 3.1 Numerical solution of Eq. (3.5) over the interval $[0,4]$ by the linear shooting method (3.4).

$$\begin{aligned} x''(t) &= f(t,x(t),x'(t)), \qquad t \in [a,b] \\ x(a) &= \alpha, \qquad x(b) = \beta\,. \end{aligned} \tag{3.6}$$

be the BVP in question and let $x(t,s)$ denote the solution of the IVP

$$\begin{aligned} x''(t) &= f(t,x(t),x'(t)), \qquad t \in [a,b] \\ x(a) &= \alpha, \qquad x'(a) = s, \end{aligned} \tag{3.7}$$

where s is a parameter that can be varied. The IVP (3.7) is solved with different values of s with, e.g., RK4 method till the boundary condition on the right side $x(b) = \beta$ becomes fulfilled. As mentioned above, the solution $x(t,s)$ of (3.7) depends on the parameter s. Let us define a function

$$F(s) := x(b,s) - \beta\,.$$

If the BVP (3.6) has a solution, then the function $F(s)$ has a root, which is just the value of the slope $x'(a)$ giving the solution $x(t)$ of the BVP in question. The zeros of $F(s)$ can be found with, e.g., *Newton's method* [37].
The Newton's method is probably the best known method for finding numerical approximations to the zeroes of a real-valued function. The idea of the method is to use the first few terms of the Taylor series of a function $F(s)$ in the vicinity of a suspected root, i.e.,

$$F(s_n + h) = F(s_n) + F'(s_n)\,h + \mathcal{O}(h^2)\,.$$

where s_n is a n'th approximation of the root. Now if one inserts $h = s - s_n$, one obtains

$$F(s) = F(s_n) + F'(s_n)\,(s - s_n)\,.$$

As the next approximation s_{n+1} to the root we choose the zero of this function, i.e.,

$$F(s_{n+1}) = F(s_n) + F'(s_n)\,(s_{n+1} - s_n) = 0 \Rightarrow s_{n+1} = s_n - \frac{F(s_n)}{F'(s_n)}\,. \tag{3.8}$$

The derivative $F'(s_n)$ can be calculated using the forward difference formula

$$F'(s_n) = \frac{F(s_n + \delta s) - F(s_n)}{\delta s}$$

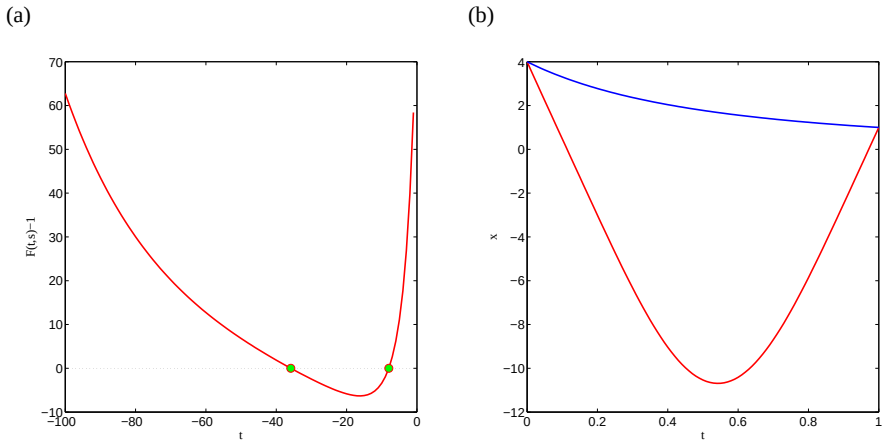

Fig. 3.2 Numerical solution of BVP (3.9) with single shooting method. (a) The Function $F(s) = x(t,s) - 1$ is presented. Green points depict two zeros of this function, which can be found with Newton's method. (b) Two solutions of (3.9) corresponding to two different values of parameter s (the red line corresponds to $s = -35.8$, whereas the blue one – to $s = -8.0$).

where δs is small. Notice that this procedure can be unstable near a horizontal asymptote or a local extremum.

Example 1

Consider a simple nonlinear BVP [45]

$$\begin{aligned} x''(t) &= \frac{3}{2}x(t)^2, \\ x(0) &= 4, \qquad x(1) = 1 \end{aligned} \tag{3.9}$$

over the interval $t \in [0, 1]$ and let us solve it numerically with the single shooting method discussed above. First of all we define a corresponding IVP

$$x''(t) = \frac{3}{2}x(t)^2 \qquad x(0) = 4\,, \quad x'(0) = s$$

over $t \in [0, 1]$ and solve it for different values of s, e.g., $s \in [-100, 0]$ with the classical RK4 method. The result of calculation is presented on Fig. 3.2 (a). One can see, that the function $F(s) = x(t,s) - 1$ admits two zeros, depicted on Fig. 3.2 (a) as green points. In order to find them we use the Newton's method, discussed above. The method gives an approximation to both zeros of the function $F(s)$: $s = \{-35.8, -8.0\}$, which give the right slope $x'(0)$. Both solutions, corresponding to two different values of s are presented on Fig. 3.2 (b).

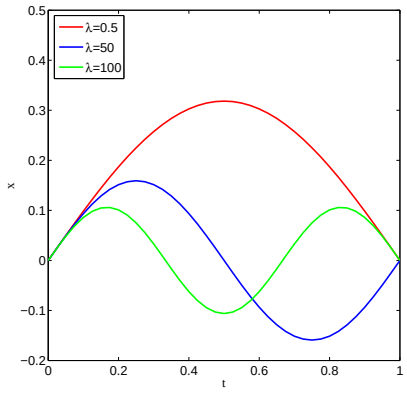

Fig. 3.3 Numerical solutions of Eq. (3.10) over the interval $[0,1]$ by single shooting method. First three eigenfunctions, corresponding to eigenvalues $\lambda = \{\pi^2, (2\pi)^2, (3\pi)^2\}$ are presented.

Example 2

Let us consider a linear eigenvalue problem of the form

$$x'' + \lambda x = 0, \qquad x(0) = x(1) = 0, \quad x'(0) = 1 \tag{3.10}$$

over $t \in [0,1]$ with the simple shooting method. The exact solution is

$$\lambda = n^2\pi^2, \quad n \in \mathbb{N}.$$

In order to apply the simple shooting method we consider a corresponding IVP of the first order with additional equation for the unknown function $\lambda(t)$:

$$x' = y, \qquad y' = -\lambda x, \quad \lambda' = 0$$

with

$$x(0) = 0, \quad x'(0) = 1, \quad \lambda(0) = s.$$

where s is a free shooting parameter. Here we choose $s = \{0.5, 50, 100\}$. Results of the shooting with these initial parameters are shown on Fig. 3.3. One can see, that numerical solutions correspond to first three eigenvalues $\lambda = \{\pi^2, (2\pi)^2, (3\pi)^2\}$.

Example 3

Consider a nonlinear BVP of the fourth order [**?**]

$$x^{(4)}(t) - (1+t^2)x''(t)^2 + 5x(t)^2 = 0, \qquad t \in [0,1] \tag{3.11}$$

with

$$x(0) = 1, \quad x'(0) = 0, \quad x''(1) = -2, \quad x'''(1) = -3.$$

Our goal is to solve this equation with the simple shooting method. To this end, first we rewrite the equation as a system of four ODE's of the first order:

$$
\begin{aligned}
x_1' &= x_2\,, \\
x_2' &= x_3\,, \qquad & x_1(0) = 1\,, \qquad & x_3(1) = -2\,, \\
x_3' &= x_4\,, \qquad & x_2(0) = 0\,, \qquad & x_4(1) = -3\,, \\
x_4' &= (1+t^2)x_3^2 - 5x_1^2\,.
\end{aligned}
$$

As the second step we consider correspondig IVP

$$
\begin{aligned}
x_1' &= x_2\,, \\
x_2' &= x_3\,, \qquad & x_1(0) = 1\,, \qquad & x_3(1) = p\,, \\
x_3' &= x_4\,, \qquad & x_2(0) = 0\,, \qquad & x_4(1) = q\,, \\
x_4' &= (1+t^2)x_3^2 - 5x_1^2
\end{aligned}
$$

with two free shooting parameters p and q. The solution of this IVP fulfilles following two requirements:

$$
\begin{aligned}
F_1(p,q) &:= x_3(1,p,q) + 2 = 0\,, \\
F_2(p,q) &:= x_4(1,p,q) + 3 = 0\,.
\end{aligned}
$$

That is, a system of nonlinear algebraic equations should be solved to find (p,q). The zeros of the system can be found with the Newton's method (3.8). In this case the iteration step reads

$$
s_{i+1} = s_i - \frac{F(s_i)}{DF(s_i)}
$$

where $s = (p,q)^T$, $F = (F_1, F_2)^T$ and

$$
DF(s_i) = \begin{pmatrix} \frac{\partial F_1}{\partial p} & \frac{\partial F_1}{\partial q} \\ \frac{\partial F_2}{\partial p} & \frac{\partial F_2}{\partial q} \end{pmatrix}
$$

is a Jacobian of the system and

$$
\begin{aligned}
\frac{\partial F_i}{\partial p} &= \frac{F_i(p+\Delta p, q) - F_i(p,q)}{\Delta p}\,, \\
\frac{\partial F_i}{\partial q} &= \frac{F_i(p, q+\Delta q) - F_i(p,q)}{\Delta q}\,,
\end{aligned}
$$

where $i = 1, 2$ and Δp, Δq are given values. Numerical solution of the problem in question is presented on Fig. 3.4.

3.2 Finite difference Method

One way to solve a given BVP over the time interval $t \in [a,b]$ numerically is to approximate the problem in question by *finite differences* [**?**, 45, 30]. We form a partition of the domain $[a,b]$ using *mesh points* $a = t_0, t_1, \ldots, t_N = b$, where

$$
t_i = a + i\,h\,, \qquad h = \frac{b-a}{N}\,, \qquad i = 0, 1, \ldots N\,.
$$

Difference quotient approximations for derivatives can be used to solve BVP in question [45, 30]. In particular, using a Taylor expansion in the vicinity of the point t_j, for the first derivative one

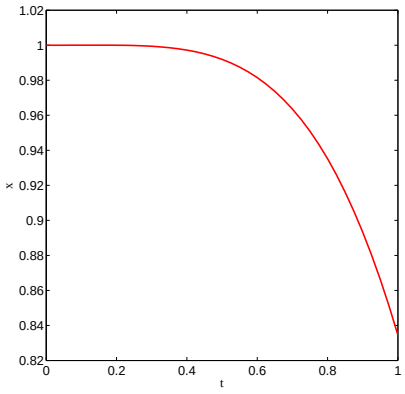

Fig. 3.4 Numerical solutions of (3.11) over the interval $[0,1]$ by single shooting method. Parameters are: $\Delta p = \Delta q = 0.05$, the time step $h = 0.025$, initial shooting parameters $(p_0, q_0) = (0,0)$.

obtains *a forward difference*

$$x'(t_i) = \frac{x(t_{i+1}) - x(t_i)}{h} + \mathscr{O}(h)\,. \tag{3.12}$$

In a similar way one gets *a backward difference*

$$x'(t_i) = \frac{x(t_i) - x(t_{i-1})}{h} + \mathscr{O}(h)\,. \tag{3.13}$$

We can combine these two approaches and derive *a central difference*, which yields a more accurate approximation:

$$x'(t_i) = \frac{x(t_{i+1}) - x(t_{i-1})}{2h} + \mathscr{O}(h^2)\,. \tag{3.14}$$

The second derivative $x''(t_i)$ can be found in the same way using the linear combination of different Taylor expansions. For example, a central difference reads

$$x''(t_i) = \frac{x(t_{i+1}) - 2x(t_i) + x(t_{i-1})}{h^2} + \mathscr{O}(h^2)\,. \tag{3.15}$$

3.2.1 Finite Difference for linear BVP

Let us consider a linear BVP of the second order (3.3)

$$x'' = p(t)x'(t) + q(t)x(t) + r(t)\,, \quad t \in [a,b]\,, \quad x(a) = \alpha\,, \quad x(b) = \beta\,.$$

and introduce the notation $x(t_i) = x_i$, $p(t_i) = p_i$, $q(t_i) = q_i$ and $r(t_i) = r_i$. Then, using Eq. (3.14) and Eq. (3.15) one can rewrite Eq. (3.3) as *a difference equation*

$$\begin{aligned} x_0 &= \alpha\,, \\ \frac{x_{i+1} - 2x_i + x_{i-1}}{h^2} &= p_i \frac{x_{i+1} - x_{i-1}}{2h} + q_i x_i + r_i\,, \qquad i = 1, \ldots, N-1\,, \\ x_N &= \beta\,. \end{aligned}$$

Now we can multiply both sides of the second equation with h^2 and collect terms, involving x_{i-1}, x_i and x_{i+1}. As result we get a system of linear equations

$$\left(1+\frac{h}{2}p_i\right)x_{i-1}-(2+h^2q_i)x_i+\left(1-\frac{h}{2}p_i\right)x_{i+1}=h^2r_i\,,\qquad i=1,2,\dots N-1\,.$$

or, in matrix notation

$$A\mathbf{x}=b\,, \tag{3.16}$$

or, more precisely

$$\begin{pmatrix} -(2+h^2q_1) & 1-\frac{h}{2}p_1 & 0 & \dots & \dots & 0 \\ 1+\frac{h}{2}p_2 & -(2+h^2q_2) & 1-\frac{h}{2}p_2 & 0 & \dots & 0 \\ 0 & 1+\frac{h}{2}p_3 & -(2+h^2q_3) & 1-\frac{h}{2}p_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & 1-\frac{h}{2}p_{N-2} \\ 0 & \dots & \dots & 0 & 1+\frac{h}{2}p_{N-1} & -(2+h^2q_{N-1}) \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ \cdot \\ x_{N-1} \end{pmatrix}=\begin{pmatrix} h^2r_1-\gamma_1 \\ h^2r_2 \\ h^2r_3 \\ \cdot \\ \cdot \\ h^2r_{N-1}-\gamma_N \end{pmatrix},$$

where

$$\gamma_1=\alpha\left(\frac{h}{2}p_1+1\right),\qquad \gamma_N=\beta\left(1-\frac{h}{2}p_{N-1}\right).$$

Our goal is to find unknown vector **x**. To this end we should invert the matrix A. This matrix has a band structure and is *tridiagonal*. For matrices of this kind a tridiagonal matrix algorithm (TDMA), also known als *Thomas algorithm* can be used (see Appendix A for details).

Example

Solve a linear BVP [**?**]

$$\begin{aligned} -x''(t)-(1+t^2)x(t)&=1\,,\\ x(-1)=x(1)&=0 \end{aligned} \tag{3.17}$$

over $t\in[-1,1]$ with finite difference method. First we introduce discrete set of nodes $t_i=-1+ih$ with given time step h. According to notations used in previous section, $p(t)=0$, $q(t)=-(1+t^2)$, $r(t)=-1$, $\alpha=\beta=0$. Hence, the linear system (3.16) we are interested in reads

$$\begin{pmatrix} -(2+h^2q_1) & 1 & 0 & \dots & \dots & 0 \\ 1 & -(2+h^2q_2) & 1 & 0 & \dots & 0 \\ 0 & 1 & -(2+h^2q_3) & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & 1 \\ 0 & \dots & \dots & 0 & 1 & -(2+h^2q_{N-1}) \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ \cdot \\ x_{N-1} \end{pmatrix}=-\begin{pmatrix} h^2 \\ h^2 \\ h^2 \\ \cdot \\ \cdot \\ h^2 \end{pmatrix}.$$

The numerical solution of the problem in question is presented on Fig. 3.5.

3.2.2 Finite difference for linear eigenvalue problems

Consider a Sturm-Liouville problem of the form

$$-x''(t)+q(t)x(t)=\lambda\,v(t)x(t)\,, \tag{3.18}$$

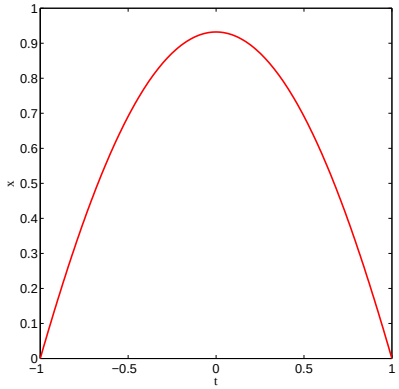

Fig. 3.5 Numerical solutions of (3.17) over the interval $[-1, 1]$ by finite difference method.

over $t \in [a, b]$ with

$$x(a) = 0\,, \qquad x(b) = 0\,.$$

Introducing notation $x_i := x(t_i)$, $q_i := q(t_i)$, $v_i := v(t_i)$, we can write down a difference equation for Eq. (3.18)

$$\begin{aligned} x_0 &= 0 \\ -\frac{x_{i+1} - 2x_i + x_{i-1}}{h^2} + q_i x_i - \lambda\, v_i x_i &= 0\,, \qquad i = 1, \ldots N-1\,, \\ x_N &= 0\,. \end{aligned}$$

If $v_i \neq 0$ for all i we can rewrite the difference equation above as an eigenvalue problem

$$(A - \lambda\, I)\, x = 0 \tag{3.19}$$

for a tridiagonal matrix A

$$A = \begin{pmatrix} \frac{2}{h^2 v_1} + \frac{q_1}{v_1} & \frac{-1}{h^2 v_1} & 0 & \ldots & \ldots & 0 \\ \frac{-1}{h^2 v_2} & \frac{2}{h^2 v_2} + \frac{q_2}{v_2} & \frac{-1}{h^2 v_2} & \ldots & \ldots & 0 \\ 0 & \frac{-1}{h^2 v_3} & \frac{2}{h^2 v_3} + \frac{q_3}{v_3} & \frac{-1}{h^2 v_3} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \frac{-1}{h^2 v_{N-2}} \\ 0 & \ldots & \ldots & 0 & \frac{-1}{h^2 v_{N-1}} & \frac{2}{h^2 v_{N-1}} + \frac{q_3}{v_{N-1}} \end{pmatrix}$$

and a vector $x = (x_1, x_2, \ldots, x_{N-1})^T$.

Part II
Partial Differential Equations

In this part, we discuss the standard numerical techniques used to integrate partial differential equations (PDEs). Here we focus on finite difference method.

Chapter 4
Introduction

4.1 Definition, Notation and Classification

A differential equation involving more than one independent variable and its partial derivatives with respect to those variables is called a *partial differential equation* (PDE).
Consider a simple PDE of the form:

$$\frac{\partial}{\partial x}u(x,y) = 0.$$

This equation implies that the function $u(x,y)$ is independent of x. Hence the general solution of this equation is $u(x,y) = f(y)$, where f is an arbitrary function of y. The analogous ordinary differential equation is

$$\frac{du}{dx} = 0,$$

its general solution is $u(x) = c$, where c is a constant. This example illustrates that general solutions of ODEs involve arbitrary constants, whereas solutions of PDEs involve *arbitrary functions*.

In general, one can classify PDEs with respect to different criterion, e.g.:

- Order;
- Dimension;
- Linearity;
- Initial/Boundary value problem, etc.

By *order* of PDE we will understand the order of the highest derivative that occurs. A PDE is said to be *linear* if it is linear in unknown functions and their derivatives, with coefficients depending on the independent variables. The independent variables typically include one or more *space dimensions* and sometimes time dimension as well.
For example, the wave equation

$$\frac{\partial^2 u(x,t)}{\partial t^2} = a^2 \frac{\partial^2 u(x,t)}{\partial x^2}$$

is a one-dimensional, second-order linear PDE. In contrast, the Fisher Equation of the form

$$\frac{\partial u(\mathbf{r},t)}{\partial t} = \triangle u(\mathbf{r},t) + u(\mathbf{r},t) - u(\mathbf{r},t)^2,$$

where $\mathbf{r} = (x, y)$ is a two-dimensional, second-order nonlinear PDE.

Linear Second-Order PDEs

For linear PDEs in two dimensions there is a simple classification in terms of the general equation

$$au_{xx} + bu_{xy} + cu_{yy} + du_x + eu_y + fu + g = 0,$$

where the coefficients a, b, c, d, e, f and g are real and in general can also be functions of x and y. The PDE's of this type are classified by the value of discriminant $D_\lambda = b^2 - 4ac$ of the eigenvalue problem for the matrix

$$A = \begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix},$$

build from the coefficients of the highest derivatives. A simple classification is shown on the following table [45, 30]:

D_λ	Typ	Eigenvalues
$D_\lambda < 0$	*elliptic*	the same sign
$D_\lambda > 0$	*hyperbolic*	different signs
$D_\lambda = 0$	*parabolic*	zero is an eigenvalue

For instance, *the Laplace equation* for the electrostatic potential φ in the space without a charge

$$\frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} = 0$$

is elliptic, as $a = c = 1$, $b = 0$, $D_\lambda = -4 < 0$. In general, elliptic PDEs describe processes that have already reached steady state, and hence are time-independent.
One-dimensional *wave equation* for some amplitude $A(x,t)$

$$\frac{\partial^2 A}{\partial t^2} - v^2 \frac{\partial^2 A}{\partial x^2} = 0$$

with the positive dispersion velocity v is hyperbolic ($a = 1$, $b = 0$, $c = -v^2$, $D_\lambda = 4v^2 > 0$). Hyperbolic PDEs describe time-dependent, conservative processes, such as convection, that are not evolving toward steady state.
The next example is a *diffusion equation* for the patricle's density $\rho(x,t)$

$$\frac{\partial \rho}{\partial t} = D \frac{\partial^2 \rho}{\partial x^2},$$

where $D > 0$ is a diffusion coefficient. This equation is called to be parabolic ($a = -D$, $b = c = 0$,$D_\lambda = 0$). Parabolic PDEs describe time-dependent, dissipative processes, such as diffusion, that are evolving toward steady state.

Each of these classes should be investigated separately as different methods are required for each class. The next point to emphasize is that as all the coefficients of the PDE can depend on x and y, this classification concept is *local*.

Initial and Boundary-Value Problems

As it was mentioned above the solution of PDEs involve arbitrary functions. That is, in order to solve the system in question completely, additional conditions are needed. These conditions can be given in the form of *initial* and *boundary* conditions. Initial conditions define the values of the dependent variables at the initial stage (e.g. at $t = 0$), whereas the boundary conditions give the information about the value of the dependent valiable or its derivative on the boundary of the area of interest. Typically, one distinguishes

- *Dirichlet conditions* specify the values of the dependent variables of the boundary points.
- *Neumann conditions* specify the values of the normal gradients of the boundary.
- *Robin conditions* defines a linear combination of the Drichlet and Neumann conditions.

Moreover, it is useful to classify the PDE in question in terms of *initial value problem (IVP)* and *boundary value problem (BVP)*.

- *Initial value problem:* PDE in question describes *time evolution*, i.e., the initial space-distribution is given; the goal is to find how the dependent variable propagates in time (e.g., the diffusion equation).
- *Boundary value problem:* A static solution of the problem should be found in some region-and the dependent variable is specified on its boundary (e.g., the Laplace equation).

4.2 Finite difference method

Let us consider a one-dimensional PDE for the unknown function $u(x,t)$. One way to numerically solve the PDE is to approximate all the derivatives by *finite differences*. We partition the domain in space using *a mesh* $x_0, x_1, \ldots, x_N$ and in time using a mesh $t_0, t_1, \ldots, t_T$. Fisrt we assume *a uniform partition* both in space and in time, so that the difference between two consecutive space points will be $\triangle x$ and between two consecutive time points will be $\triangle t$, i.e.,

$$\begin{aligned} x_i &= x_0 + i\triangle x, \qquad i = 0,1,\ldots,M; \\ t_j &= t_0 + j\triangle t, \qquad j = 0,1,\ldots,T; \end{aligned}$$

The Taylor series method

Consider a Taylor expansion of an analytical function $u(x)$.

$$u(x+\triangle x) = u(x) + \sum_{n=1}^{\infty} \frac{\triangle x^n}{n!} \frac{\partial^n u}{\partial x^n} = u(x) + \triangle x \frac{\partial u}{\partial x} + \frac{\triangle x^2}{2!} \frac{\partial^2 u}{\partial x^2} + \frac{\triangle x^3}{3!} \frac{\partial^3 u}{\partial x^3} + \ldots \tag{4.1}$$

Then for the first derivative one obtains:

$$\frac{\partial u}{\partial x} = \frac{u(x+\triangle x) - u(x)}{\triangle x} - \frac{\triangle x}{2!} \frac{\partial^2 u}{\partial x^2} - \frac{\triangle x^2}{3!} \frac{\partial^3 u}{\partial x^3} - \ldots \tag{4.2}$$

If we break the right hand side of the last equation after the first term, for $\triangle x \ll 1$ the last equation becomes

$$\boxed{\frac{\partial u}{\partial x} = \frac{u(x+\triangle x) - u(x)}{\triangle x} + \mathcal{O}(\triangle x) = \frac{\triangle_i u}{\triangle x} + \mathcal{O}(\triangle x)}, \tag{4.3}$$

where

$$\triangle_i u = u(x+\triangle x) - u(x) := u_{i+1} - u_i$$

is called *a forward difference*. The backward expansion of the function u can be written as $\triangle x \ll 1$ the last equation reads

$$u(x+(-\triangle x)) = u(x) - \triangle x \frac{\partial u}{\partial x} + \frac{\triangle x^2}{2!} \frac{\partial^2 u}{\partial x^2} - \frac{\triangle x^3}{3!} \frac{\partial^3 u}{\partial x^3} + \ldots, \tag{4.4}$$

so for the first derivative one obtains

$$\boxed{\frac{\partial u}{\partial x} = \frac{u(x) - u(x - \triangle x)}{\triangle x} + \mathscr{O}(\triangle x) = \frac{\nabla_i u}{\triangle x} + \mathscr{O}(\triangle x)}, \tag{4.5}$$

where

$$\nabla_i u = u(x) - u(x - \triangle x) := u_i - u_{i-1}$$

is called *a backward difference*. One can see that both forward and backward differences are of the order $\mathscr{O}(\triangle x)$. We can combine these two approaches and derive *a central difference*, which yields a more accurate approximation. If we substract Eq. (4.5) from Eq. (4.3) one obtains

$$u(x + \triangle x) - u(x - \triangle x) = 2\triangle x \frac{\partial u}{\partial x} + 2\frac{\triangle x^3}{3!}\frac{\partial^3 u}{\partial x^3} + \ldots, \tag{4.6}$$

what is equivalent to

$$\boxed{\frac{\partial u}{\partial x} = \frac{u(x + \triangle x) - u(x - \triangle x)}{2\triangle x} + \mathscr{O}(\triangle x^2)} \tag{4.7}$$

Note that the central difference (4.7) is of the order of $\mathscr{O}(\triangle^2 x)$.
The second derivative can be found in the same way using the linear combination of different Taylor expansions. For instance, consider

$$u(x + 2\triangle x) = u(x) + 2\triangle x \frac{\partial u}{\partial x} + \frac{(2\triangle x)^2}{2!}\frac{\partial^2 u}{\partial x^2} + \frac{(2\triangle x)^3}{3!}\frac{\partial^3 u}{\partial x^3} + \ldots \tag{4.8}$$

Substracting from the last equation Eq. (4.1), multiplied by two, one gets the following equation

$$u(x + 2\triangle x) - 2u(x + \triangle x) = -u(x) + \triangle x^2 \frac{\partial^2 u}{\partial x^2} + \triangle x^3 \frac{\partial^3 u}{\partial x^3} + \ldots \tag{4.9}$$

Hence one can approximate the second derivative as

$$\boxed{\frac{\partial^2 u}{\partial x^2} = \frac{u(x + 2\triangle x) - 2u(x + \triangle x) + u(x)}{\triangle x^2} + \mathscr{O}(\triangle x).} \tag{4.10}$$

Similarly one can obtain the expression for the second derivative in terms of backward expansion, i.e.,

$$\boxed{\frac{\partial^2 u}{\partial x^2} = \frac{u(x - 2\triangle x) - 2u(x - \triangle x) + u(x)}{\triangle x^2} + \mathscr{O}(\triangle x).} \tag{4.11}$$

Finally, if we add Eqn. (4.3) and (4.5) an expression for the cental second derivative reads

$$\boxed{\frac{\partial^2 u}{\partial x^2} = \frac{u(x + \triangle x) - 2u(x) + u(x - \triangle x)}{\triangle x^2} + \mathscr{O}(\triangle x^2).} \tag{4.12}$$

One can see that approximation (4.12) provides more accurate approximation as (4.10) and (4.11). In an analogous way one can obtain finite difference approximations to higher order derivatives and differential operators. The coefficients for first three derivatives for the case of forward, backward and central differences are given in Tables 4.1, 4.2, 4.3.

Mixed derivatives

A finite difference approximations for the mixed partial derivatives can be calculated in the same way. For example, let us find the central approximation for the derivative

	u_i	u_{i+1}	u_{i+2}	u_{i+3}	u_{i+4}
$\triangle x\frac{\partial u}{\partial x}$	-1	1			
$\triangle x^2\frac{\partial^2 u}{\partial x^2}$	1	-2	1		
$\triangle x^3\frac{\partial^3 u}{\partial x^3}$	-1	3	-3	1	
$\triangle x^4\frac{\partial^4 u}{\partial x^4}$	1	-4	6	-4	1

Table 4.1 Forward difference quotient, $\mathcal{O}(\triangle x)$

	u_{i-4}	u_{i-3}	u_{i-2}	u_{i-1}	u_i
$\triangle x\frac{\partial u}{\partial x}$				-1	1
$\triangle x^2\frac{\partial^2 u}{\partial x^2}$			1	-2	1
$\triangle x^3\frac{\partial^3 u}{\partial x^3}$		-1	3	-3	1
$\triangle x^4\frac{\partial^4 u}{\partial x^4}$	1	-4	6	-4	1

Table 4.2 Backward difference quotient, $\mathcal{O}(\triangle x)$

	u_{i-2}	u_{i-1}	u_i	u_{i+1}	u_{i+2}
$2\triangle x\frac{\partial u}{\partial x}$		-1	0	1	
$\triangle x^2\frac{\partial^2 u}{\partial x^2}$		1	-2	1	
$2\triangle x^3\frac{\partial^3 u}{\partial x^3}$	-1	2	0	-2	1
$\triangle x^4\frac{\partial^4 u}{\partial x^4}$	1	-4	6	-4	1

Table 4.3 Central difference quotient, $\mathcal{O}(\triangle x^2)$

$$\frac{\partial^2 u}{\partial x\partial y}=\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial y}\right)=\frac{\partial}{\partial x}\left(\frac{u(x,y+\triangle y)-u(x,y-\triangle y)}{2\triangle y}+\mathcal{O}(\triangle y^2)\right)=$$
$$=\frac{u(x+\triangle x,y+\triangle y)-u(x-\triangle x,y+\triangle y)-u(x+\triangle x,y-\triangle y)+u(x-\triangle x,y-\triangle y)}{4\triangle x\triangle y}+\mathcal{O}(\triangle x^2\triangle y^2).$$

A nonequidistant mesh

In the section above we have considered different numerical approximations for the derivatives using the equidistant mesh. However, in many applications it is convinient to use a nonequidistant mesh, where the spatial steps fulfill the following rule:

$$\triangle x_i=\alpha\triangle x_{i-1}.$$

If $\alpha=1$ the mesh is said to be equidistant. Let us now calculate the first derivative of the function $u(x)$ of the second-order accurance:

$$u(x+\alpha\triangle x)=u(x)+\alpha\triangle x\frac{\partial u}{\partial x}+\frac{(\alpha\triangle x)^2}{2!}\frac{\partial^2 u}{\partial x^2}+\frac{(\alpha\triangle x)^3}{3!}\frac{\partial^3 u}{\partial x^3}+\ldots \tag{4.13}$$

Adding the last equation with Eq. (4.4) multiplied by α one obtains the expression for the second derivative

$$\frac{\partial^2 u}{\partial x^2}=\frac{u(x+\alpha\triangle x)-(1+\alpha)u(x)+\alpha u(x-\triangle x)}{\frac{1}{2}\alpha(\alpha+1)\triangle x^2}+\mathcal{O}(\triangle x) \tag{4.14}$$

Substitution of the last equation into Eq. (4.4) yields

$$\boxed{\frac{\partial u}{\partial x} = \frac{u(x+\alpha\triangle x) - (1-\alpha^2)u(x) - \alpha^2 u(x-\triangle x)}{\alpha(\alpha+1)\triangle x} + \mathscr{O}(\triangle x^2)}. \tag{4.15}$$

4.3 von Neumann stability analysis

One of the central questions arising by numerical treatment of the problem in question is stability of the numerical scheme we are interested in [37]. An algorithm for solving an evolutionary partial differential equation is said to be *stable*, if the numerical solution at a fixed time remains bounded as the step size goes to zero, so the perturbations in form of, e.g., rounding error does not increase in time. Unfortunately, there are no general methods to verify the numerical stability for the partial differential equations in general form, so one restrict oneself to the case of linear PDE's. The standard method for linear PDE's with periodic boundary conditions was proposed by John von Neumann [9, 5] and is based on the representation of the rounding error in form of the Fouirer series.

In order to illustrate the procedure, let us introduce the following notation:

$$u^{j+1} = \mathscr{T}[u^j]. \tag{4.16}$$

Here $\mathscr{T}$ is a nonlinear operator, depending on the numerical scheme in question. The successive application of $\mathscr{T}$ results in a sequence of values

$$u^0, u^1, u^2, \ldots,$$

that approximate the exact solution of the problem. However, at each time step we add a small error ε^j, i.e., the sequence above reads

$$u^0 + \varepsilon^0, u^1 + \varepsilon^1, u^2 + \varepsilon^2, \ldots,$$

where ε^j is a cumulative rounding error at time t_j. Thus we obtain

$$u^{j+1} + \varepsilon^{j+1} = \mathscr{T}(u^j + \varepsilon^j). \tag{4.17}$$

After linearization of the last equation (we suppose that Taylor expansion of $\mathscr{T}$ is possible) the linear equation for the perturbation takes the form:

$$\boxed{\varepsilon^{j+1} = \frac{\partial \mathscr{T}(u^j)}{\partial u^j}\varepsilon^j := G\varepsilon^j}. \tag{4.18}$$

This equation is called *the error propagation law*, whereas the linearization matrix G is said to be *an amplification matrix* [21]. Now, the stability of the numerical scheme in question depends on the eigenvalues g_μ of the matrix G. In other words, the scheme is stable if and only if

$$|g_\mu| \leq 1 \quad \forall \mu$$

Now the question is how this information can be used in practice. The first point to emphasize is that in general one deals with the $u(x_i, t_j) := u_i^j$, so one can write

$$\varepsilon_i^{j+1} = \sum_{i'} G_{ii'}\varepsilon_{i'}^j, \tag{4.19}$$

where

$$G_{ii'} = \frac{\partial \mathscr{T}(u^j)_i}{\partial u^j_{i'}}.$$

Futhermore, the spatial variation of ε_i^j (rounding error at the time step t_j at the point x_i) can be expanded in a finite Fourier series in the intreval $[0, L]$:

$$\varepsilon_i^j = \sum_k e^{ikx_i} \tilde{\varepsilon}^j(k), \tag{4.20}$$

where k is the wavenumber and $\tilde{\varepsilon}^j(k)$ are the Fourier coefficients. Since the rounding error tends to grow or decay exponentially with time, it is reasonable to assume that $\tilde{\varepsilon}^j(k)$ varies exponentially with time, i.e.,

$$\varepsilon_i^j = \sum_k e^{\omega t_j} e^{ikx_i},$$

where ω is a constant. The next point to emphasize is that the functions e^{ikx_i} are eigenfunctions of the matrix G, so the last expansion can be interpreted as the expansion in eigenfunctions of G. In addition, the equation for the error is linear, so it is enough to examine the grows of the error of a typical term of the sum. Thus, from the practical point of view one take the error ε_i^j just as

$$\varepsilon_i^j = e^{\omega t_j} e^{ikx_i}.$$

The substitution of this expression into Eq. (4.20) results in the following relation

$$\varepsilon_i^{j+1} = g(k)\varepsilon_i^j. \tag{4.21}$$

That is, one can interpert e^{ikx_i} as an eigenvector corresponding to the eigenvalue $g(k)$. The value $g(k)$ is often called *an amplification factor*. Finally, the stability criterium is then given by

$$\boxed{|g(k)| \leq 1 \quad \forall k}. \tag{4.22}$$

This criterium is called *von Neumann stablity criterium*.
Notice that presented stability analysis can be applied only in certain cases. Namely, the linear PDE in question schould be with constant coefficients and satisfies periodic boundary conditions. In addition, the corresponding difference scheme should possesses no more than two time levels [44]. However, it is often used in more complicated situations as a good estimation for the step sizes used in the approximation.

Chapter 5
Advection Equation

Let us consider a continuity equation for the one-dimensional drift of incompressible fluid. In the case that a particle density $u(x,t)$ changes only due to convection processes one can write

$$u(x, t+\triangle t) = u(x - c\triangle t, t).$$

If $\triangle t$ is sufficient small, the Taylor-expansion of both sides gives

$$u(x,t) + \triangle t \frac{\partial u(x,t)}{\partial t} \simeq u(x,t) - c\triangle t \frac{\partial u(x,t)}{\partial x}$$

or, equivalently

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0\,. \tag{5.1}$$

Here $u = u(x,t)$, $x \in \mathbb{R}$, and c is a nonzero constant velocity. Equation (5.1) is called to be *an advection equation* and describes the motion of a scalar u as it is advected by a known velocity field. According to the classification given in Sec. 4.1, Eq. (5.1) is a hyperbolic PDE. The unique solution of (5.1) is determined by an initial condition $u_0 := u(x,0)$

$$u(x,t) = u_0(x - ct)\,, \tag{5.2}$$

where $u_0 = u_0(x)$ is an arbitrary function defined on $\mathbb{R}$.
One way to find this exact solution is the method of characteristics (see App. B). In the case of Eq. (5.1) the coefficients $A = c$, $B = 1$, $C = 0$ and Eqn. (B.2) read

$$\begin{aligned} \frac{dt}{ds} &= 1 \Leftrightarrow |t(0) = 0| \Leftrightarrow t = s, \\ \frac{dx}{ds} &= c \Leftrightarrow |x(0) = x_0| \Leftrightarrow x = x_0 + ct. \end{aligned}$$

That is, for the advection equation (5.1) characteristic curves are represented by straight lines (see Fig. 5.1). Hence, Eq. (B.3) becomes

$$\frac{du}{ds} = 0 \quad \text{with} \quad u(0) = u_0(x_0)\,.$$

Alltogether the solution of (5.1) takes the form (5.2). The solution (5.2) is just an initial function u_0 shifted by ct to the right (for $c > 0$) or to the left ($c < 0$), which remains constant along the characteristic curves ($du/ds = 0$).

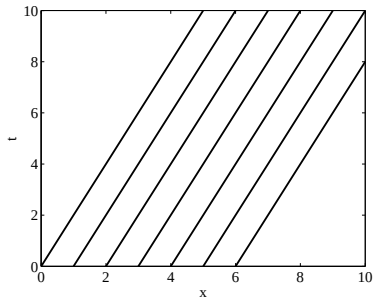

Fig. 5.1 Chareacteristic curves $x = x_0 + cs,\ s = t$ for advection equation (5.1) are shown for different values of c.

Now we focus on different explicit methods to solve advection equation (5.1) numerically on the periodic domain $[0, L]$ with a given initial condition $u_0 = u(x, 0)$.

5.1 FTCS Method

We start the discussion of Eq. (5.1) with a so-called FTCS (forward in time, centered in space) method. As discussed in Sec. 4.2 we introduce the discretization in time on the uniform grid

$$t_j = t_0 + j\triangle t\,, \qquad j = 0 \ldots T\,.$$

Furthermore, in the x-direction, we discretize on the uniform grid

$$x_i = a + i\triangle x\,, \qquad i = 0 \ldots M\,, \quad \triangle x = \frac{L}{M}\,.$$

Adopting a forward temporal difference scheme (4.3), and a centered spatial difference scheme (4.7), Eq. (5.1) yields

$$\begin{aligned} \frac{u_i^{j+1} - u_i^j}{\triangle t} &= -c\frac{u_{i+1}^j - u_{i-1}^j}{2\triangle x} \Leftrightarrow \\ u_i^{j+1} &= u_i^j - \frac{c\triangle t}{2\triangle x}\left(u_{i+1}^j - u_{i-1}^j\right). \end{aligned} \tag{5.3}$$

Here we use a notation $u_i^j := u(x_i, t_j)$. Shematic representation of the FTCS approximation (5.3) is shown on Fig. 5.2.

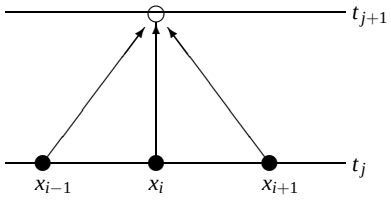

Fig. 5.2 Schematic visualization of the FTCS-method (5.3).

von Neumann Stability Analysis

To investigate stability of the scheme (5.3) we follow the concept of von Neumann, introduced in Sec. 4.3. The usual ansatz

$$\varepsilon_i^j \sim e^{ikx_i}$$

leads to the following relation

$$\varepsilon_i^{j+1} = e^{ikx_i} - \frac{c\triangle t}{2\triangle x}\left(e^{ik(x_i+\triangle x)} - e^{ik(x_i-\triangle x)}\right) = \underbrace{\left(1 - \frac{c\triangle t}{2\triangle x}\left(e^{ik\triangle x} - e^{-ik\triangle x}\right)\right)}_{g(k)}\varepsilon_i^j,$$

where ε_i^{j+1} stands for the cumulative rounding error at time t_j. The von Neumann's stability condition (4.22) for the amplification factor $g(k)$ reads:

$$|g(k)| \leq 1 \qquad \forall k.$$

In our case one obtains:

$$|g(k)|^2 = 1 + \frac{c^2\triangle t^2}{\triangle x^2}\sin^2(k\triangle x),$$

One can see that the magnitude of the amplification factor $g(k)$ is greater than unity for all k. This implies that the instability occurs for all given c, $\triangle t$ and $\triangle x$, i.e., the FTCS scheme (5.3) is *unconditionally unstable*.

5.2 Upwind Methods

The next simple scheme we are intersted in belongs to the class of so-called *upwind methods* – numerical discretization schemes for solving hyperbolic PDEs. According to such a scheme, the spatial differences are skewed in the "upwind" direction, i.e., the direction from which the advecting flow originates. The origin of upwind methods can be traced back to the work of R. Courant et al. [8].
The simplest upwind schemes possible are given by

$$\begin{aligned}\frac{u_i^{j+1} - u_i^j}{\triangle t} &= c\frac{u_i^j - u_{i-1}^j}{\triangle x} \Leftrightarrow \\ u_i^{j+1} &= u_i^j - \frac{c\triangle t}{\triangle x}\left(u_i^j - u_{i-1}^j\right), \qquad (c>0).\end{aligned} \tag{5.4}$$

and

$$\begin{aligned}\frac{u_i^{j+1} - u_i^j}{\triangle t} &= c\frac{u_{i+1}^j - u_i^j}{\triangle x} \Leftrightarrow \\ u_i^{j+1} &= u_i^j - \frac{c\triangle t}{\triangle x}\left(u_{i+1}^j - u_i^j\right) \qquad (c<0).\end{aligned} \tag{5.5}$$

Note that the upwind scheme (5.4) corresponds to the case of positive velocities c, whereas Eq. (5.5) stands for the case $c<0$. The next point to emphasize is that both schemes (5.4)–(5.5) are only first-order in space and time. Shematic representations of both upwind methods is presented on Fig. 5.3

In the matrix form the upwind scheme (5.4) takes the form

$$\mathbf{u}^{j+1} = A\mathbf{u}^{j}, \tag{5.6}$$

where $\mathbf{u}^j$ is a vector on the time step j and A is a $n \times n$ matrix ($h := \triangle t/\triangle x$),

$$A = \begin{pmatrix} 1-ch & 0 & 0 & \dots & \boxed{ch} \\ ch & 1-ch & 0 & \dots & 0 \\ 0 & ch & 1-ch & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & ch & 1-ch & \end{pmatrix}$$

The boxed element A_{1n} indicates the influence of the periodic boundary conditions. Similary, one can also represent the scheme (5.5) in the form (5.6) with matrix

$$A = \begin{pmatrix} 1+ch & -ch & 0 & \dots & 0 \\ 0 & 1+ch & -ch & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1+ch & -ch & \\ \boxed{-ch} & \dots & 0 & 1+ch & \end{pmatrix}$$

Again, the boxed element A_{n1} displays the influence of periodic boundary conditions.

von Neumann Stability Analysis

In order to investigate the stability of the upwind scheme (5.4) (or (5.5)) we start with the usual ansatz

$$\varepsilon_i^j \sim e^{ikx_i},$$

leading to the equation for the cumulative rounding error at time t_{j+1}

$$\varepsilon_i^{j+1} = g(k)\varepsilon_i^j,$$

where the amplification factor $g(k)$ for, e.g., the upwind scheme (5.4) is given by

$$g(k) = 1 - \frac{c\triangle t}{\triangle x}\left(1 - e^{-ik\triangle x}\right) = \left|\alpha = \frac{c\triangle t}{\triangle x},\ \varphi = -k\triangle x\right| = 1 - \alpha + \alpha e^{i\varphi}.$$

The stability condition (4.22) is fulfilled for all k as long as

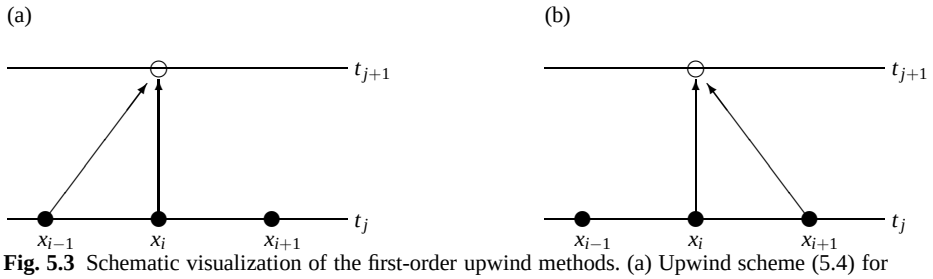

Fig. 5.3 Schematic visualization of the first-order upwind methods. (a) Upwind scheme (5.4) for $c > 0$. (b) Upwind scheme (5.5) for $c < 0$.

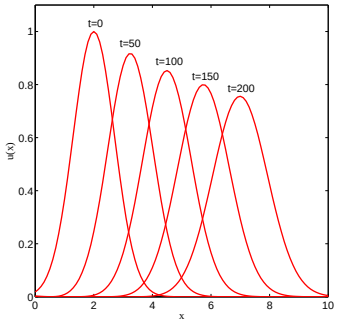

Fig. 5.4 Advection of a one-dimensional Gauß–pulse $u_0 = \exp(-(x-0.2)^2)$ with the scheme (5.4). Numerical calculation performed on the interval $x \in [0, 10]$ using $c = 0.5$, $\triangle t = 0.05$, $\triangle x = 0.1$. Numerical solutions at different times $t = 0$, $t = 50$, $t = 100$, $t = 150$, $t = 200$ are shown.

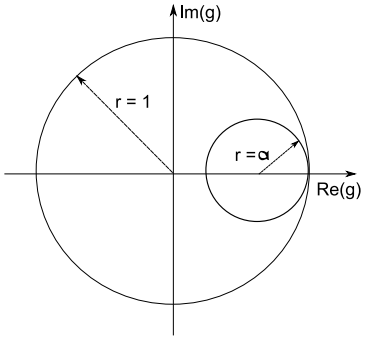

Fig. 5.5 Scematic illustration of the stability condition (5.7) for the upwind-method (5.4).

$$|g(k)| \leq 1 \Leftrightarrow 1-\alpha \leq 0 \Leftrightarrow \boxed{\frac{c\triangle t}{\triangle x} \leq 1 \Leftrightarrow c \leq \frac{\triangle x}{\triangle t}.} \tag{5.7}$$

That is, the method (5.4) is *conditionally stable*, i.e., is stable if and only if the "physical" velocity c is not bigger than the spreading velocity $\triangle x/\triangle t$ of the numerical method. This is equivalent to the condition that the time step, $\triangle t$, must be smaller than the time taken for the wave to travel the distance of the spatial step, $\triangle x$. Schematic illustration of stability condition (5.7) is shown on Fig. . Condition (5.7) is called *a Courant-Friedrichs-Lewy (CFL)* stability criterion whereas α is. The condition (5.7) is named after R. Courant, K. Friedrichs, and H. Lewy, who described it in their paper in 1928 [38].

Numerical results

Figure 5.5 shows an example of the calculation in which the upwind scheme (5.4) is used to advect a Gauß–pulse. Parameters of the calculation are choosen as

`Space interval`	L=10
`Initial condition`	$u_0(x) = \exp(-(x-2)^2)$
`Space discretization step`	$\triangle x = 0.1$
`Time discretization step`	$\triangle t = 0.05$
`Velocity`	$c = 0.5$
`Amount of time steps`	$T = 200$

For parameter values given above the stability condition (5.7) is fulfilled, so the scheme (5.4) is stable. On the other hand, one can see, that the wave-form shows evidence of dispersion. We discuss this problem in details in the next section.

5.3 The Lax Method

Let us consider a minor modification of the FTCS-method (5.3), in which the term u_i^j has been replaced by an average over its two neighbours (see Fig. 5.6):

$$u_i^{j+1} = \frac{1}{2}\left(u_{i+1}^j + u_{i-1}^j\right) - \frac{c\triangle t}{2\triangle x}\left(u_{i+1}^j - u_{i-1}^j\right). \tag{5.8}$$

In this case the matrix A of the linear system (5.6) is given by a sparse matrix with zero main diagonal

$$A = \begin{pmatrix} 0 & a & 0 & 0 & \dots & 0 & 0 & \boxed{\text{b}} \\ b & 0 & a & 0 & \dots & 0 & 0 & 0 \\ 0 & b & 0 & a & \dots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \dots & b & 0 & a \\ \boxed{\text{a}} & 0 & 0 & 0 & \dots & 0 & b & 0 \end{pmatrix},$$

where

$$a = \frac{1}{2} - \frac{c\triangle t}{2\triangle x},$$
$$b = \frac{1}{2} + \frac{c\triangle t}{2\triangle x}.$$

and the boxed elements represent the influence of periodic boundary conditions.

von Neumann stability analysis

In the case of the scheme (5.8) the amplification factor $g(k)$ becomes

$$g(k) = \cos k\triangle x - i\frac{c\triangle t}{\triangle x}\sin k\triangle x.$$

With $\alpha = \frac{c\triangle t}{\triangle x}$ and $\varphi(k) = k\triangle x$ one obtains

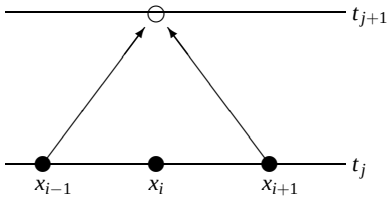

Fig. 5.6 Schematic visualization of the Lax method (5.8).

$$|g(k)|^2 = \cos^2\varphi(k) + \alpha^2 \sin^2\varphi(k) = 1 - (1-\alpha^2)\sin^2\varphi(k)\,.$$

The stability condition (4.22) is fulfilled for all k as long as

$$1 - \alpha^2 \geq 0 \Leftrightarrow \frac{c\triangle t}{\triangle x} \leq 1,$$

which is again the Courant-Friedrichs-Lewy condition (5.7). In fact, all stable *explicit* differencing schemes for solving the advection equation (5.1) are subject to the CFL constraint, which determines the maximum allowable time-step $\triangle t$.

Numerical results

Consider a realization of the Lax method (5.8) on the concrete numerical example:

`Space interval`	L=10
`Initial condition`	$u_0(x) = \exp(-10(x-2)^2)$
`Space discretization step`	$\triangle x = 0.05$
`Time discretization step`	$\triangle t = 0.05$
`Velocity`	$c = 0.5$
`Amount of time steps`	$T = 200$

As can be seen from Fig. 5.7 (a) like the upwind method (5.4), the Lax scheme introduces a spurious *dispersion* effect into the advection problem (5.1). Although the pulse is advected at the correct speed (i.e., it appears approximately stationary in the co-moving frame $x - ct$ (see Fig. 5.7 (b))), it does not remain the same shape as it should.

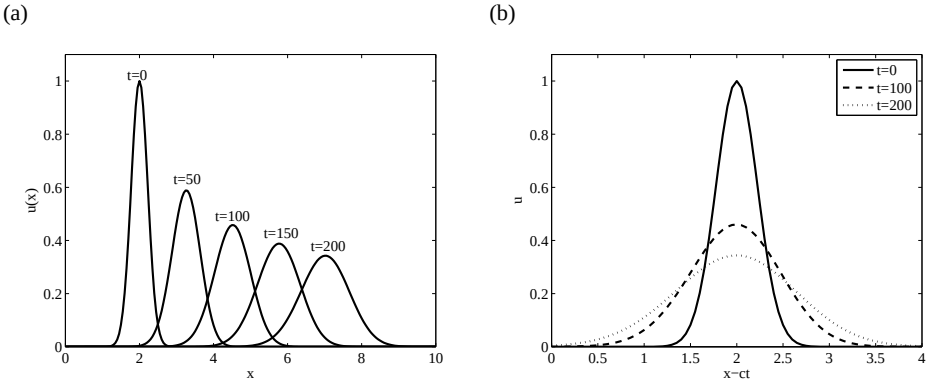

Fig. 5.7 Numerical implementation of the Lax method (5.8). Parameters are: Advection velocity is $c = 0.5$, length of the space interval is $L = 10$, space and time discretization steps are $\triangle x = 0.05$ and $\triangle t = 0.05$, amount of time steps is $T = 200$, and initial condition is $u_0(x) = \exp(-10(x-2)^2)$. (a) Time evolution of $u(x,t)$ for different time moments. Solutions at $t = 0$, 100, 150, 200 are shown. (b) Time evolution in the co-moving frame $x - ct$ at $t = 0$, 100, 200.

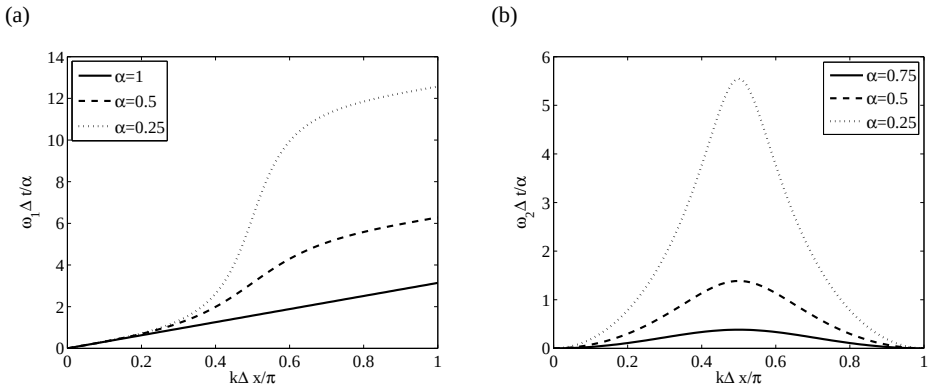

Fig. 5.8 Illustration of the dispersion relation for the Lax method calculated for different values of the Courant number α. (a) Real part of ω. (b) Imaginary part of ω.

Fourier Analysis

One can try to understand the origin of the dispersion effect with the help of the dispersion relation. The ansatz of the Fourier mode of the form

$$u_i^j \sim e^{ikx_i - i\omega t_j}$$

for Eq. (5.8) results in the following relation

$$e^{-i\omega\triangle t} = \cos k\triangle x - i\alpha \sin k\triangle x\,,$$

where again $\alpha = c\triangle t/\triangle x$. For $\alpha = 1$ the right hand side of this relation is equal to $\exp(-ik\triangle x)$ and one otbaines

$$\omega = k\frac{\triangle x}{\triangle t} = k\cdot c\,.$$

That is, in this case the Lax method (5.8) is exact (the phase velocity ω/k is equal c). However, in general case one should suppose $\omega = \omega_1 + i\omega_2$, i.e., the Fourier modes are of the form

$$u(x,t) \sim e^{ikx - i(\omega_1 + i\omega_2)t} = e^{i(kx-\omega_1 t)}e^{-\omega_2 t}$$

and the corresponding dispersion relation reads

$$\omega\triangle t = (\omega_1 - i\omega_2)\triangle t = i\ln\bigl(\cos k\triangle x - i\alpha\sin k\triangle x\bigr)\,. \tag{5.9}$$

Hence, if $\omega_2 \geq 0$ one has deal with damped waves, that decay exponentially with the time contstant $1/\omega_2$. Furthermore, from Eq. (5.9) can be seen, that for $\alpha < 1$ Fouirer modes with wavelength about some grid constants ($\lambda = 2\pi/k \approx 4\triangle x$) are not only damped (see Fig 5.8 (b)) but, on the other hand, propagate with the essential greater phase velocity ω_1/k as long-wave components (see Fig. 5.8 (a)). Now the question we are interested in is what is the reason for this unphysical behavior? To answer this question let us rewrite the differencial scheme (5.8):

$$\underbrace{\frac{1}{2}u_i^{j+1}+\frac{1}{2}u_i^{j+1}}_{u_i^{j+1}}\underbrace{-\frac{1}{2}u_i^{j-1}+\frac{1}{2}u_i^{j-1}}_{0}=\frac{1}{2}\left(u_{i+1}^j+u_{i-1}^j\right)+\underbrace{u_i^j-u_i^j}_{0}-\frac{c\triangle t}{2\triangle x}\left(u_{i+1}^j-u_{i-1}^j\right)\Leftrightarrow$$

$$\frac{1}{2}\left(u_i^{j+1}-u_i^{j-1}\right)=\frac{1}{2}\left(u_{i+1}^j-2u_i^j+u_{i-1}^j\right)-\frac{c\triangle t}{2\triangle x}\left(u_{i+1}^j-u_{i-1}^j\right)-\frac{1}{2}\left(u_i^{j+1}-2u_i^j+u_i^{j-1}\right),;$$

or, in the continuous limit,

$$\frac{\partial u}{\partial t}=\frac{\triangle x^2}{2\triangle t}\frac{\partial^2 u}{\partial x^2}-c\frac{\partial u}{\partial x}-\frac{\triangle t^2}{2}\frac{\partial^2 u}{\partial t^2} \tag{5.10}$$

Althougth the last term in (5.10) tends to zero as $\triangle t\to 0$, the behavior of the first term depends on the behavior of $\triangle t$ and $\triangle x$. That is, the Lax method is not a consistent way to solve Eq. (5.1). This message becomes clear if one calculates the partial derivative

$$\frac{\partial^2 u}{\partial t^2}\overset{(5.1)}{=}c^2\frac{\partial^2 u}{\partial x^2}.$$

Substitution of the last expresssion into Eq. (5.10) relults in the equation, which in addition to the advection term includes diffusion term as well,

$$\frac{\partial u}{\partial t}=-c\frac{\partial u}{\partial x}+D\frac{\partial^2 u}{\partial x^2},$$

where

$$D=\frac{\triangle x^2}{2\triangle t}-c^2\frac{\triangle t}{2}$$

is a positive diffusion constant. Now the unphysical behavior of the Fourier modes becomes clear– we have integrated *the wrong equation!* That is, other numerical approximations should be used to solve Eq. (5.1) in a more correct way.

5.4 The Lax-Wendroff method

The Lax-Wendroff method, named after P. Lax and B. Wendroff [29], can be derived in a variety of ways. Let us consider two of them. The first way is based on the idea of so-called *multistep* methods. First of all let us calculate u_i^{j+1} using the information on the half time step:

$$u_i^{j+\frac{1}{2}}=u_i^j+\frac{\triangle t}{2}\left(-c\left.\frac{\partial u}{\partial x}\right|_{(i,j)}\right),$$

$$u_i^{j+1}=u_i^j+\triangle t\left(-c\left.\frac{\partial u}{\partial x}\right|_{(i,j+\frac{1}{2})}\right).$$

Now we use the central difference to approximate the derivative $u_x|_{i,j+\frac{1}{2}}$, i.e.,

$$u_i^{j+1}=u_i^j-\frac{c\triangle t}{\triangle x}\left(u_{i+\frac{1}{2}}^{j+\frac{1}{2}}-u_{i-\frac{1}{2}}^{j+\frac{1}{2}}\right).$$

On the second step, both quantities $u_{i\pm\frac{1}{2}}^{j+\frac{1}{2}}$ can be calculated using the Lax method (5.8). As the result, following two-steps scheme is obtained (see Fig. 5.9):

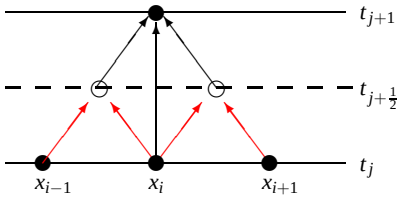

Fig. 5.9 Schematical visualization of the Lax-Wendroff method (5.11).

$$
\begin{aligned}
u_{i-\frac{1}{2}}^{j+\frac{1}{2}} &= \frac{1}{2}\left(u_i^j + u_{i-1}^j\right) - \frac{c\triangle t}{2\triangle x}\left(u_i^j - u_{i-1}^j\right),\\
u_{i+\frac{1}{2}}^{j+\frac{1}{2}} &= \frac{1}{2}\left(u_i^j + u_{i+1}^j\right) - \frac{c\triangle t}{2\triangle x}\left(u_{i+1}^j - u_i^j\right),\\
u_i^{j+1} &= u_i^j - \frac{c\triangle t}{\triangle x}\left(u_{i+\frac{1}{2}}^{j+\frac{1}{2}} - u_{i-\frac{1}{2}}^{j+\frac{1}{2}}\right).
\end{aligned}
\tag{5.11}
$$

The approximation scheme (5.11) can also be rewritten as

$$
u_i^{j+1} = b_{-1}u_{i-1}^j + b_0 u_i^j + b_1 u_{i+1}^j, \tag{5.12}
$$

where constants b_{-1}, b_0 and b_1 are given by

$$
\begin{aligned}
b_{-1} &= \frac{\alpha}{2}(\alpha+1),\\
b_0 &= 1-\alpha^2,\\
b_1 &= \frac{\alpha}{2}(\alpha-1)
\end{aligned}
$$

and α is the Courant number. The matrix A of the linear system (5.6) is a sparse matrix of the form

$$
A = \begin{pmatrix}
b_0 & b_1 & 0 & 0 & \dots & 0 & 0 & \boxed{b_{-1}}\\
b_{-1} & b_0 & b_1 & 0 & \dots & 0 & 0 & 0\\
0 & b_{-1} & b_0 & b_1 & \dots & 0 & 0 & 0\\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots\\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots\\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots\\
0 & 0 & 0 & 0 & \dots & b_{-1} & b_0 & b_1\\
\boxed{b_1} & 0 & 0 & 0 & \dots & 0 & b_{-1} & b_0
\end{pmatrix},
$$

where boxed elements stays for influence of the periodic boundary conditions.

Notice that the three-point scheme (5.12) is second-order accurate in space and time. The distinguishing feature of the Lax–Wendroff method is, that for the linear advection equation (5.1) it is the only *explicit* scheme of second-order accuracy in space and time.

The second way to derive the Lax-Wendroff diffrential scheme is based on the idea that we would like to get a scheme with second-order accurate in space and time. First of all, we use Taylor series expansion in time, namely

$$
u(x_i,t_{j+1}) = u(x_i,t_j) + \triangle t\,\partial_t u(x_i,t_j) + \frac{\triangle t^2}{2}\partial_t^2 u(x_i,t_j) + \mathscr{O}(\triangle t^2).
$$

In the next place one replaces time derivatives in the last expression by space derivatives according to the relation

$$
\partial_t^{(n)} u = (-c)^n \partial_x^{(n)} u.
$$

Hence

$$u(x_i,t_{j+1}) = u(x_i,t_j) - c\triangle t\partial_x u(x_i,t_j) + \frac{c^2\triangle t^2}{2}\partial_x^2 u(x_i,t_j) + \mathscr{O}(\triangle t^2)\,.$$

Finally, the space derivatives are approximated by central differences (4.7), (4.12), resulting in the Lax-Wendroff scheme (5.12).

von Neumann stability analysis

In the case of the method (5.12) the amplification factor $g(k)$ becomes

$$g(k) = \big(1+\alpha^2(\cos(k\triangle x)-1)\big) - i\alpha\sin(k\triangle x)$$

and

$$|g(k)|^2 = 1-\alpha^2(1-\alpha^2)\big(1-\cos(k\triangle x)\big)^2\,.$$

Hence, the stability condition (4.22) reads

$$1-\alpha^2 \geq 0 \;\Leftrightarrow\; \alpha = \frac{c\triangle x}{\triangle t} \leq 1\,,$$

and one becomes (as expected) the CFL-condition (5.7) again.

Fourier analysis

In order to check availability of dispersion, let us calculate the dispersion relation for the scheme (5.12). The ansatz of the form $\exp(i(kx_i-\omega t_j))$ results in

$$e^{-i\omega\triangle t} = \big(1+\alpha^2(\cos(k\triangle x)-1)\big) - i\alpha\sin(k\triangle x)\,,$$

and with $\omega = \omega_1 + i\omega_2$ one obtaines

$$\omega\triangle t = \omega_1\triangle t - i\omega_2\triangle t = i\ln\Big(\big(1+\alpha^2(\cos(k\triangle x)-1)\big) - i\alpha\sin(k\triangle x)\Big)\,.$$

One can easily see, that in the case of (5.12) dispersion (see Fig. 5.10 (a)) as well as damping (diffusion) (see Fig. 5.10 (b)) of Fourier modes take place. However, as can be seen on Fig. 5.10 and Fig. 5.11, dispersion and diffusion are weaker as for the Lax method (5.8) and appear by much smaller wave lengths. Because of these properties and taking into account the fact that the method (5.12) is of the second order, it becomes a standard scheme to approximate Eq. (5.1). Moreover, the scheme (5.12) can be generalized to the case of conservation equation in general form.

Lax-Wendroff method for 1D conservation equations

A typical one-dimensional evolution equation takes the form

$$\frac{\partial u}{\partial t} + \frac{\partial F(u)}{\partial x} = 0\,, \tag{5.13}$$

where $u = u(x,t)$ and the form of a function $F(u)$ depends on the problem we are interested in. One can try to apply the Lax-Wendroff method (5.12) to Eq. (5.13). With $F_i^j := F(u_i^j)$ one obtains the following differential scheme

$$
\begin{aligned}
u_{i-\frac{1}{2}}^{j+\frac{1}{2}} &= \frac{1}{2}\left(u_i^j + u_{i-1}^j\right) - \frac{\triangle t}{2\triangle x}\left(F_i^j - F_{i-1}^j\right), \\
u_{i+\frac{1}{2}}^{j+\frac{1}{2}} &= \frac{1}{2}\left(u_i^j + u_{i+1}^j\right) - \frac{\triangle t}{2\triangle x}\left(F_{i+1}^j - F_i^j\right), \\
u_i^{j+1} &= u_i^j - \frac{\triangle t}{\triangle x}\left(F_{i+\frac{1}{2}}^{j+\frac{1}{2}} - F_{i-\frac{1}{2}}^{j+\frac{1}{2}}\right).
\end{aligned}
\tag{5.14}
$$

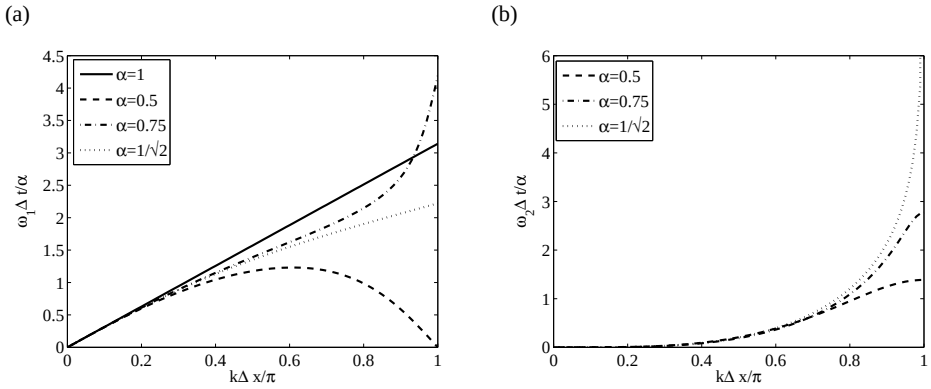

Fig. 5.10 Illustration of the dispersion relation for the Lax-Wendroff method calculated for different values of α. (a) Real part of ω (dispersion). (b) Imaginary part of ω (diffusion).

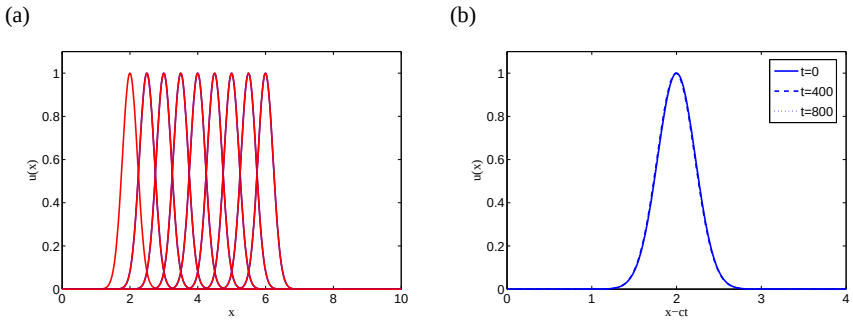

Fig. 5.11 Numerical implementation of the Lax-Wendroff method (5.12). Parameters are: Advection velocity is $c = 0.5$, length of the space interval is $L = 10$, space and time discretization steps are $\triangle x = 0.05$ and $\triangle t = 0.05$, amount of time steps is $T = 800$, and initial condition is $u_0(x) = \exp(-(x-2)^2)$. (a) Time evolution of $u(x,t)$ for different time moments. (b) Time evolution in the co-moving frame $x - ct$ at $t = 0, 400, 800$.

Chapter 6
Burgers Equation

One of the major challenges in the field of complex systems is a thorough understanding of the phenomenon of turbulence. Direct numerical simulations (DNS) have substantially contributed to our understanding of the disordered flow phenomena inevitably arising at high Reynolds numbers. However, a successful theory of turbulence is still lacking which whould allow to predict features of technologically important phenomena like turbulent mixing, turbulent convection, and turbulent combustion on the basis of the fundamental fluid dynamical equations. This is due to the fact that already the evolution equation for the simplest fluids, which are the so-called Newtonian incompressible fluids, have to take into account nonlinear as well as nonlocal properties:

$$\begin{aligned} \frac{\partial}{\partial t}\mathbf{u}(\mathbf{x},t)+\mathbf{u}(\mathbf{x},t)\cdot\nabla\mathbf{u}(\mathbf{x},t) &= -\nabla p(\mathbf{x},t)+\nu\Delta\mathbf{u}(\mathbf{x},t)\,, \\ \nabla\cdot\mathbf{u}(\mathbf{x},t) &= 0\,. \end{aligned} \tag{6.1}$$

Nonlinearity stems from the convective term and the pressure term, whereas nonlocality enters due to the pressure term. Due to incompressibility, the pressure is defined by a Poisson equation

$$\Delta p(\mathbf{x},t) = -\nabla\cdot\mathbf{u}(\mathbf{x},t)\cdot\nabla\mathbf{u}(\mathbf{x},t)\,. \tag{6.2}$$

In 1939 the dutch scientist J.M. Burgers [4] simplified the Navier-Stokes equation (6.1) by just dropping the pressure term. In contrast to Eq. (6.1), this equation can be investigated in one spatial dimension (Physicists like to denote this as 1+1 dimensional problem in order to stress that there is one spatial and one temporal coordinate):

$$\frac{\partial}{\partial t}u(x,t)+u(x,t)\frac{\partial}{\partial x}u(x,t) = \nu\frac{\partial^2}{\partial x^2}u(x,t)+F(x,t) \tag{6.3}$$

Note that usually the Burgers equation is considered without external force $F(x,t)$. However, we shall include this external force field.

The Burgers equation 6.3 is nonlinear and one expects to find phenomena similar to turbulence. However, as it has been shown by Hopf [20] and Cole [6], the homogeneous Burgers equation lacks the most important property attributed to turbulence: The solutions do not exhibit chaotic features like sensitivity with respect to initial conditions. This can explicitly shown using the *Hopf-Cole transformation* which transforms Burgers equation into a linear parabolic equation. From the numerical point of view, however, this is of importance since it allows one to compare numerically obtained solutions of the nonlinear equation with the exact one. This comparison is important to investigate the quality of the applied numerical schemes. Furthermore, the equation has still interesting applications in physics and astrophysics. We will briefly mention some of them.

Growth of interfaces: Deposition models

The Burgers equation (6.3) is equivalent to the so-called *Kardar-Parisi-Zhang (KPZ-) equation* which is a model for a solid surface growing by vapor deposition, or, the opposite case, erosion of material from a solid surface. The location of the surface is described in terms of a height function $h(\mathbf{x},t)$. This height evolves in time according to the KPZ-equation

$$\frac{\partial}{\partial t}h(\mathbf{x},t) - \frac{1}{2}\left(\nabla h(\mathbf{x},t)\right)^2 = \nu\frac{\partial^2}{\partial x^2}h(x,t) + F(x,t)\,. \tag{6.4}$$

This equation is obtained from the simple advection equation for a surface at $z = h(\mathbf{x},t)$ moving with velocity $\mathbf{U}(\mathbf{x},t)$

$$\frac{\partial}{\partial t}h(\mathbf{x},t) + \mathbf{U}\cdot\nabla h(\mathbf{x},t) = 0\,. \tag{6.5}$$

The velocity is assumed to be proportional to the gradient of $h(\mathbf{x},t)$, i.e. the surface evolves in the direction of its gradient. Surface diffusion is described by the diffusion term.

Burgers equation (6.3) is obtained from the KPZ-equation just by forming the gradient of $h(\mathbf{x},t)$:

$$\mathbf{u}(\mathbf{x},t) = -\nabla h(\mathbf{x},t)\,. \tag{6.6}$$

6.1 Hopf-Cole Transformation

The Hopf-Cole transformation is a transformation, which maps the solution of the Burgers equation (6.3) to the heat equation

$$\frac{\partial}{\partial t}\psi(\mathbf{x},t) = \nu\Delta\psi(\mathbf{x},t)\,. \tag{6.7}$$

We perform the ansatz

$$\psi(\mathbf{x},t) = e^{h(\mathbf{x},t)/2\nu} \tag{6.8}$$

and determine

$$\Delta\psi = \frac{1}{2\nu}\left[\Delta h + \frac{1}{2\nu}(\nabla h)^2\right]e^{h/2\nu} \tag{6.9}$$

leading to

$$\frac{\partial}{\partial t}h - \frac{1}{2}(\nabla h)^2 = \nu\Delta h\,. \tag{6.10}$$

However, this is exactly the Kardar-Parisi-Zhang equation (6.4). The complete transformation is then obtained by combining

$$\mathbf{u}(\mathbf{x},t) = -\frac{1}{2\nu}\nabla\ln\psi(\mathbf{x},t)\,. \tag{6.11}$$

We explicitly see that the Hopf-Cole transformation turns the nonlinear Burgers equation into the linear heat conduction equation. Since the heat conduction equation is explicitly solvable in terms of the so-called heat kernel we obtain a general solution of the Burgers equation. Before we construct this general solution, we want to emphasize that the Hopf-Cole transformation applied to the multi-dimensional Burgers equation only leads to the general solution provided the initial condition $\mathbf{u}(\mathbf{x},0)$ is a gradient field. For general initial conditions, especially for initial fields with $\nabla\times\mathbf{u}(\mathbf{x},t)$, the solution can not be constructed using the Hopf-Cole transformation and, consequently, is not known in analytical terms. In one dimension spatial dimension it is not necessary to distinguish between these two cases.

6.2 General Solution of the 1D Burgers Equation

We are now in the position to formulate the general solution of the Burgers equation (6.3) in one spatial dimension with initial condition

$$u(x,0)\,, \qquad \psi(x,0) = e^{-\frac{1}{2\nu}\int^x dx' u(x',0)}\,. \tag{6.12}$$

The solution of the 1D heat equation can be expressed by the heat-kernel

$$\psi(x,t) = \int dx' G(x-x',t)\psi(x',0) \tag{6.13}$$

with the kernel

$$G(x-x',t) = \frac{1}{\sqrt{4\pi t}} e^{-\frac{(x-x')^2}{4\nu t}} \tag{6.14}$$

In terms of the initial condition (6.12) the solution explicitly reads

$$\psi(x,t) = \frac{1}{\sqrt{4\pi t}} \int dx' e^{-\frac{(x-x')^2}{4\nu t} - \frac{1}{2\nu}\int^{x'} dx'' u(x'',0)}\,. \tag{6.15}$$

The n-dimensional solution of the Burgers equation (6.3) for initial fields, which are gradient fields, are obtained analogously:

$$\psi(x,t) = \frac{1}{(4\pi t)^{d/2}} \int d\mathbf{x}' e^{-\frac{(\mathbf{x}-\mathbf{x}')^2}{4\nu t} - \frac{1}{2\nu}\int^{\mathbf{x}'} d\mathbf{x}''\cdot\mathbf{u}(\mathbf{x}'',0)}\,. \tag{6.16}$$

Agian, we see that the solution exist provided the integral is independent of the integration contour:

$$\int^{\mathbf{x}'} d\mathbf{x}''\cdot\mathbf{u}(\mathbf{x}'',0) = h(\mathbf{x}',t)\,. \tag{6.17}$$

We can investigate the limiting case of vanishing viscosity, $\nu \to 0$. In the expression for $\psi(x,t)$, eq. (6.16), the integral is dominated by the minimum of the exponential function,

$$\min_{x'}\left[-\frac{(x-x')^2}{4\nu t} - \frac{1}{2\nu}\int^{x'} dx'' u(x'',0)\right]. \tag{6.18}$$

This leads to the so-called characteristics (see App. (B))

$$x = x' - tu(x',0)\,, \tag{6.19}$$

which we have already met in the discussion of the advection equation (5.1) (see Chapter 5). A special solution for the viscid Burgers equation is

$$u(x,t) = 1 - \tanh\left(\frac{x-x_c-t}{2\nu}\right). \tag{6.20}$$

6.3 Forced Burgers Equation

The Hopf-Cole transformation can be applied to the forced Burgers equation. It is straightforward to show that this leads to the parabolic differential equation

$$\frac{\partial}{\partial t}\psi(x,t) = \nu\Delta\psi(\mathbf{x},t) - U(\mathbf{x},t)\psi(\mathbf{x},t)\,, \tag{6.21}$$

where the potential is related to the force

$$\mathbf{F}(\mathbf{x},t) = -\frac{1}{2\nu}\nabla U(\mathbf{x},t)\,. \tag{6.22}$$

The relationship with the Schrödinger equation for a particle moving in the potential $U(\mathbf{x},t)$ is obvious. Recently, the Burgers equation with a fluctuating force has been investigated [36]. Interestingly, Burgers equation with a linear force, i.e. a quadratic potential

$$U(x,t) = a(t)x^2 \tag{6.23}$$

for an arbitrary time dependent coefficient $a(t)$ could be solved analytically [15].

6.4 Numerical Treatment

Let us consider a one-dimensional Burgers equation (6.3) without forcing.

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \nu\frac{\partial^2 u}{\partial x^2}.$$

When $\nu = 0$, Burgers equation becomes *the inviscid Burgers equation:*

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0, \tag{6.24}$$

which is a prototype for equations for which the solution can develop discontinuities *(shock waves)*. As was mentioned above, as the solution of the advection equation (5.1), the solution of Eq. (6.24) can be constructed by the method of characteristics (see App. B). Suppose we have an initial value problem, i.e., a smooth function $u(x,0) = u_0(x)$, $x \in \mathbb{R}$ is given. In this case the coefficients A, B and C are

$$A = u, \quad B = 1, \quad C = 0.$$

Equations (B.2-B.3) read

$$\begin{aligned}
\frac{dt}{ds} &= 1 \Leftrightarrow |t(0) = 0| \Leftrightarrow t = s,\\
\frac{du}{ds} &= 0 \Leftrightarrow |u(0) = u_0(x_0)| \Leftrightarrow u(s,x_0) = u_0(x_0),\\
\frac{dx}{ds} &= u \Leftrightarrow |x(0) = x_0| \Leftrightarrow x = u_0(x_0)t + x_0.
\end{aligned}$$

Hence the general solution of (6.24) takes the form

$$u(x,t) = u_0(x - u_0(x_0)t, t). \tag{6.25}$$

Eq. (6.25) is an implicit relation that determines the solution of the inviscid Burgers' equation. Note that the characteristics are straight lines, but not all the lineas have the same slope. It will be possible for the characteristics to intersect. If we write the characteristics as

$$t = \frac{u}{u_0(x_0)} - \frac{x_0}{u_0(x_0)},$$

one can see, that the slope $1/u_0(x_0)$ of the characteristics depends on the point x_0 and on the initial function u_0. For inviscid Burgers' equation (6.24), the time T_c at which the characteristics cross and a shock forms, the "breaking" time, can be determined exactly as

$$T_c = \frac{-1}{\min\{u_x(x,0)\}}$$

This relation can be used if Eq. (6.24) has smooth initial data (so that it is differentiable). From the formula for T_c, we can see that the solution will break and a shock will form if $u_x(x,0)$ is negative at some point. From numerical point of view it is convenient to rewrite the Burgers' equation as

$$\frac{\partial u}{\partial t} + \frac{1}{2}\frac{\partial}{\partial x}(u^2) = 0 \tag{6.26}$$

Equation (6.26) describes a one-dimensional conservation law (5.13) with $F = \frac{1}{2}u^2$ and can be solve, e.g., with the upwind method (5.4) or with the Lax-Wendroff method (5.14).

`Space interval`	L=10
`Initial condition`	$u_0(x) = \exp(-(x-3)^2)$
`Space discretization step`	$\triangle x = 0.05$
`Time discretization step`	$\triangle t = 0.05$
`Amount of time steps`	$T = 36$

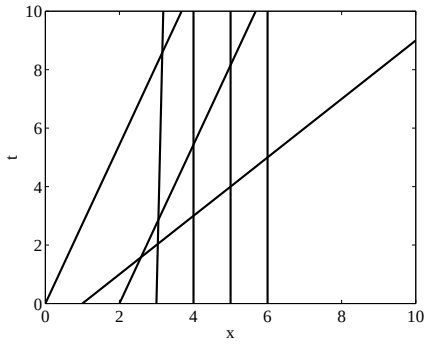

Fig. 6.1 Characteristics curves for the inviscid Burgers' equation (6.24)

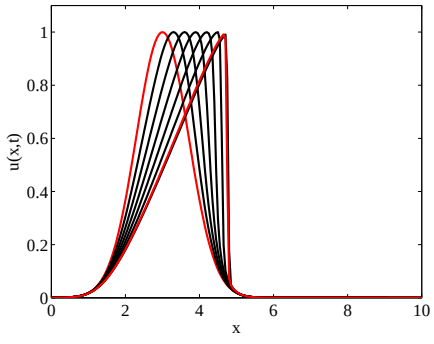

Fig. 6.2 Numerical solution of the inviscid Burgers' equation (6.24)

6.4.0.1 The Riemann Problem

A Riemann problem, named after *Bernhard Riemann*, consists of a conservation law, e.g., Eq. (6.24) together with a piecewise constant data having a single discontinuity, i.e.,

$$u(x,0) = u_0(x) = \begin{cases} u_l, & x < a; \\ u_r, & x \geq a. \end{cases} \tag{6.27}$$

The form of the solution depends on the relation between u_l and u_r.

- $\mathbf{u_l} > \mathbf{u_r}$: The unique weak solution (see Fig. 6.2 (a)) is

$$u(x,0) = u_0(x) = \begin{cases} u_l, & x < a + ct; \\ u_r, & x \geq a + ct \end{cases} \tag{6.28}$$

with the *shock velocity*

$$c = \frac{1}{2}(u_l + u_r).$$

Note, that in this case the characteristics in each of the region where u is constatnt *go into the shock* as time advances (see Fig. 6.3 (b)).

`Space interval`	L=10
`Initial condition`	$u_l = 0.8, u_r = 0.2$
`Space discretization step`	$\triangle x = 0.05$
`Time discretization step`	$\triangle t = 0.05$
`Amount of time steps`	$T = 100$

The initial condition is:

$$u(x,0) = u_0(x) = \begin{cases} 0.8, & x < 5; \\ 0.2, & x \geq 5. \end{cases} \tag{6.29}$$

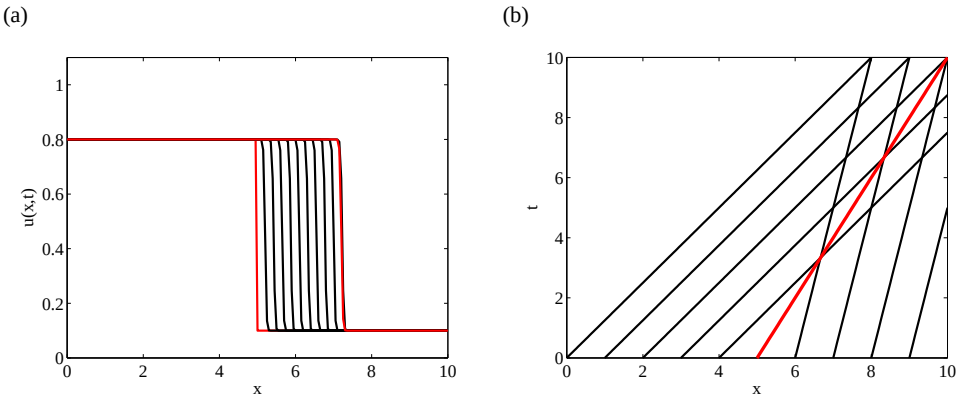

Fig. 6.3 a) Numerical solution of the inviscid Burgers' equation (6.24) for the Riemann problem for $u_l < u_r$. b) Characterics of Eq. (6.24) with initial conditions (6.29). The red line indicates the curve $x = a + ct$.

- $\mathbf{u_l} < \mathbf{u_r}$: In this case there are infinitely many weak solutions. One of them is again (6.28) with the same velocity (see Fig. 6.4 (a)). Note that in this case the characteristics *go out of the shock* (Fig. 6.4 (b)) and the solution is not stable to perturbations.

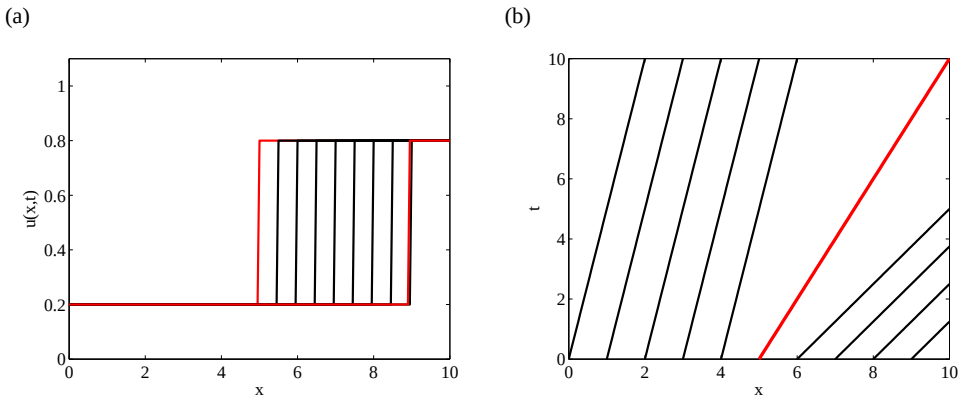

Fig. 6.4 a) Numerical solution of the inviscid Burgers' equation (**??**) for the Riemann problem for $u_l < u_r$. b) Characterics of the inviscid Burgers' equation with initial conditions (**??**). The red line indicates the curve $x = a + ct$.

Chapter 7
The Wave Equation

Another classical example of a hyperbolic PDE is a wave equation. The wave equation is a second-order linear hyperbolic PDE that describes the propagation of a variety of waves, such as sound or water waves. It arises in different fields such as acoustics, electromagnetics, or fluid dynamics. In its simplest form, the wave equation refers to a scalar function $u = u(\mathbf{r},t)$, $\mathbf{r} \in \mathbb{R}^n$ that satisfies:

$$\frac{\partial^2 u}{\partial t^2} = c^2 \nabla^2 u\,. \tag{7.1}$$

Here ∇^2 denotes the Laplacian in $\mathbb{R}^n$ and c is a constant speed of the wave propagation. An even more compact form of Eq. (7.1) is given by

$$\Box^2 u = 0\,,$$

where $\Box^2 = \nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}$ is the d'Alembertian.

7.1 The Wave Equation in 1D

The wave equation for the scalar u in the one dimensional case reads

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}. \tag{7.2}$$

The one-dimensional wave equation (7.2) can be solved exactly by d'Alembert's method, using a Fourier transform method, or via separation of variables. To illustrate the idea of the d'Alembert method, let us introduce new coordinates (ξ,η) by use of the transformation

$$\xi = x - ct\,, \qquad \eta = x + ct\,. \tag{7.3}$$

In the new coordinate system one can write

$$u_{xx} = u_{\xi\xi} + 2u_{\xi\eta} + u_{\eta\eta}\,, \qquad \frac{1}{c^2}u_{tt} = u_{\xi\xi} - 2u_{\xi\eta} + u_{\eta\eta}\,,$$

and Eq. (7.2) becomes

$$\frac{\partial^2 u}{\partial \xi \partial \eta} = 0\,. \tag{7.4}$$

That is, the function u remains constant along the curves (7.3), i.e., Eq. (7.3) describes characteristic curves of the wave equation (7.2) (see App. B). Moreover, one can see that the derivative $\partial u/\partial \xi$ does not depends on η, i.e.,

$$\frac{\partial}{\partial \eta}\left(\frac{\partial u}{\partial \xi}\right) = 0 \Leftrightarrow \frac{\partial u}{\partial \xi} = f(\xi)\,.$$

After integration with respect to ξ one obtains

$$u(\xi,\eta) = F(\xi) + G(\eta)\,,$$

where F is the primitive function of f and G is the "constant" of integration, in general the function of η. Turning back to the coordinates (x,t) one obtains the general solution of Eq. (7.2)

$$\boxed{u(x,t) = F(x-ct) + G(x+ct)\,.} \tag{7.5}$$

7.1.1 Solution of the IVP

Now let us consider an initial value problem for Eq. (7.2):

$$\begin{aligned} u_{tt} &= c^2 u_{xx}\,, \quad t \geq 0\,, \\ u(x,0) &= f(x)\,, \\ u_t(x,0) &= g(x)\,. \end{aligned} \tag{7.6}$$

To write down the general solution of the IVP for Eq. (7.2), one needs to exspress the arbitrary function F and G in terms of initial data f and g. Using the relation

$$\frac{\partial}{\partial t}F(x-ct) = -cF'(x-ct)\,, \quad \text{where} \quad F'(x-ct) := \frac{\partial}{\partial \xi}F(\xi)$$

one becomes:

$$\begin{aligned} u(x,0) &= F(x) + G(x) = f(x); \\ u_t(x,0) &= c(-F'(x) + G'(x)) = g(x)\,. \end{aligned}$$

After differentiation of the first equation with respect to x one can solve the system in terms of $F'(x)$ and $G'(x)$, i.e.,

$$F'(x) = \frac{1}{2}\left(f'(x) - \frac{1}{c}g(x)\right), \qquad G'(x) = \frac{1}{2}\left(f'(x) + \frac{1}{c}g(x)\right).$$

Hence

$$F(x) = \frac{1}{2}f(x) - \frac{1}{2c}\int_0^x g(y)\,dy + C, \quad G(x) = \frac{1}{2}f(x) + \frac{1}{2c}\int_0^x g(y)\,dy - C\,,$$

where the integration constant C is chosen in such a way that the initial condition $F(x) + G(x) = f(x)$ is fullfield. Alltogether one obtains:

$$\boxed{u(x,t) = \frac{1}{2}\left(f(x-ct) + f(x+ct)\right) + \frac{1}{2c}\int_{x-ct}^{x+ct} g(y)\,dy}\,. \tag{7.7}$$

7.1.2 Numerical Treatment

7.1.2.1 A Simple Explicit Method

The first idea is just to use central differences for both time and space derivatives, i.e.,

$$\frac{u_i^{j+1}-2u_i^j+u_i^{j-1}}{\triangle t^2}=c^2\,\frac{u_{i+1}^j-2u_i^j+u_{i-1}^j}{\triangle x^2}\,, \tag{7.8}$$

or, with $\alpha=c\triangle t/\triangle x$

$$\boxed{u_i^{j+1}=-u_i^{j-1}+2(1-\alpha^2)u_i^j+\alpha^2(u_{i+1}^j+u_{i-1}^j)}\,. \tag{7.9}$$

Schematical representation of the scheme (7.9) is shown on Fig. 7.1.

Note that one should also implement initial conditions (7.6). In order to implement the second initial condition one needs the virtual point u_i^{-1},

$$u_t(x_i,0)=g(x_i)=\frac{u_i^1-u_i^{-1}}{2\triangle t}+\mathscr{O}(\triangle t^2)\,.$$

With $g_i:=g(x_i)$ one can rewrite the last expression as

$$u_i^{-1}=u_i^1-2\triangle t g_i+\mathscr{O}(\triangle t^2)\,,$$

and the second time row can be calculated as

$$\boxed{u_i^1=\triangle t g_i+(1-\alpha^2)f_i+\frac{1}{2}\alpha^2(f_{i-1}+f_{i+1})\,,} \tag{7.10}$$

where $u(x_i,0)=u_i^0=f(x_i)=f_i$.

von Neumann Stability Analysis

In order to investigate the stability of the explicit scheme (7.9) we start with the usual ansatz (4.21)

$$\varepsilon_i^{j+1}=g^j e^{ikx_i}\,,$$

which leads to the following expression for the amplification factor $g(k)$

$$g^2=2(1-\alpha^2)g-1+2\alpha^2 g\cos(k\triangle x)\,.$$

After several transformations the last expression becomes just a quadratic equation for g, namely

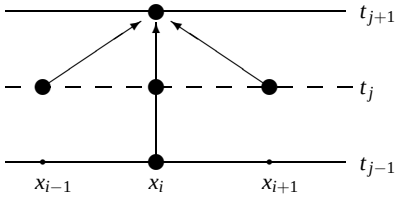

Fig. 7.1 Schematical visualization of the numerical scheme (7.9) for (7.2).

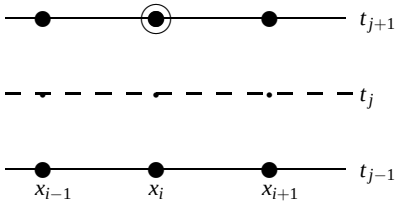

Fig. 7.2 Schematical visualization of the implicit numerical scheme (7.12) for (7.2).

$$g^2 - 2\beta g + 1 = 0\,, \tag{7.11}$$

where

$$\beta = 1 - 2\alpha^2 \sin^2\big(\frac{k\triangle x}{2}\big)\,.$$

Solutions of the equation for $g(k)$ read

$$g_{1,2} = \beta \pm \sqrt{\beta^2 - 1}\,.$$

Notice that if $\beta > 1$ then at least one of absolute value of $g_{1,2}$ is bigger that one. Therefor one should desire for $\beta < 1$, i.e.,

$$g_{1,2} = \beta \pm i\sqrt{\beta^2 - 1}$$

and

$$|g|^2 = \beta^2 + 1 - \beta^2 = 1\,.$$

That is, the scheme (7.9) is conditional stable. The stability condition reads

$$-1 \le 1 - 2\alpha^2 \sin^2\left(\frac{k\triangle x}{2}\right) \le 1,$$

what is equivalent to the standart CFL condition (5.7)

$$\alpha = \frac{c\triangle t}{\triangle x} \le 1\,.$$

7.1.2.2 An Implicit Method

One can try to overcome the problems with conditional stability by introducing *an implicit scheme*. The simplest way to do it is just to replace all terms on the right hand side of (7.8) by an average from the values to the time steps $j+1$ and $j-1$, i.e,

$$\frac{u_i^{j+1} - 2u_i^j + u_i^{j-1}}{\triangle t^2} = \frac{c^2}{2\triangle x^2}\left(u_{i+1}^{j-1} - 2u_i^{j-1} + u_{i-1}^{j-1} + u_{i+1}^{j+1} - 2u_i^{j+1} + u_{i-1}^{j+1}\right). \tag{7.12}$$

Schematical diagramm of the numerical scheme (7.12) is shown on Fig. (7.2).

Let us check the stability of the implicit scheme (7.12). To this aim we use the standart ansatz

$$\varepsilon_i^{j+1} = g^j e^{ikx_i}$$

leading to the equation for $g(k)$

$$\beta g^2 - 2g + \beta = 0$$

with

$$\beta = 1 + \alpha^2 \sin^2\left(\frac{k\triangle x}{2}\right).$$

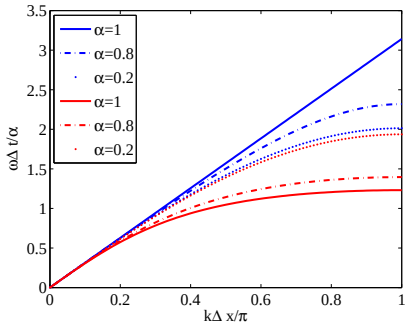

Fig. 7.3 Dispersion relation for the one-dimensional wave equation (7.2), calculated using the explicit (blue curves) and implicit (red curves) methods (7.9) and (7.12).

One can see that $\beta \geq 1$ for all k. Hence the solutions $g_{1,2}$ take the form

$$g_{1,2} = \frac{1 \pm i\sqrt{1-\beta^2}}{\beta}$$

and

$$|g|^2 = \frac{1-(1-\beta^2)}{\beta^2} = 1\,.$$

That is, the implicit scheme (7.12) is *absolute stable*.
Now, the question is, whether the implicit scheme (7.12) is better than the explicit scheme (7.9) form numerical point of view. To answer this question, let us analyse dispersion relation for the wave equation (7.2) as well as for both schemes (7.9) and (7.12). The exact dispersion relation is

$$\omega = \pm ck\,,$$

i.e, all Fourier modes propagate without dispersion with the same phase velocity $\omega/k = \pm c$. Using the ansatz $u_i^j \sim e^{ikx_i - i\omega t_j}$ for the explicit method (7.9) one obtains:

$$\cos(\omega\triangle t) = 1 - \alpha^2(1-\cos(k\triangle x))\,, \tag{7.13}$$

while for the implicit method (7.12)

$$\cos(\omega\triangle t) = \frac{1}{1+\alpha^2(1-\cos(k\triangle x))}\,. \tag{7.14}$$

One can see that for $\alpha \to 0$ both methods provide the same result, otherwise the explicit scheme (7.9) always exceeds the implicit one (see Fig. (7.3)). For $\alpha = 1$ the scheme (7.9) becomes exact, while (7.12) deviates more and more from the exact value of ω for increasing α. Hence, for Eq. (7.2) there are no motivation to use implicit scheme instead of the explicit one.

7.1.3 Examples

Example 1.

Use the explicit method (7.9) to solve the one-dimansional wave equation (7.2):

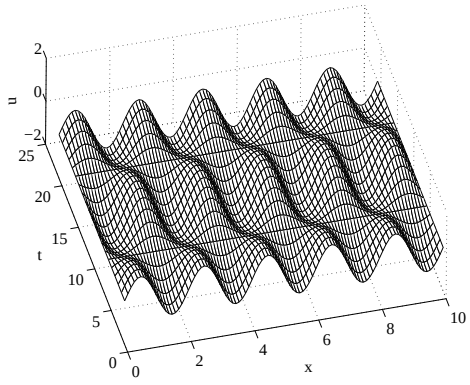

Fig. 7.4 Space-time evolution of Eq. (7.15) with the initial distribution $u(x,0) = \sin(\pi x)$, $u_t(x,0) = 0$.

$$u_{tt} = 4u_{xx} \quad \text{for} \quad x \in [0, L] \quad \text{and} \quad t \in [0, T] \tag{7.15}$$

with boundary conditions

$$u(0,t) = 0 \qquad u(L,t) = 0\,.$$

Assume that the initial position and velocity are

$$u(x,0) = f(x) = \sin(\pi x)\,, \quad \text{and} \quad u_t(x,0) = g(x) = 0\,.$$

Other parameters are:

`Space interval`	L=10
`Space discretization step`	$\triangle x = 0.1$
`Time discretization step`	$\triangle t = 0.05$
`Amount of time steps`	$T = 20$

First one can find the d'Alambert solution. In the case of zero initial velocity Eq. (7.7) becomes

$$u(x,t) = \frac{f(x-2t) + f(x+2t)}{2} = \frac{\sin\pi(x-2t) + \sin\pi(x+2t)}{2} = \sin(\pi x)\cos(2\pi t)\,,$$

i.e., the solution is just a sum of a travelling waves with initial form, given by $\frac{f(x)}{2}$. Numerical solution of (7.15) is shown on Fig. (7.4).

Example 2.

Solve Eq. (7.15) with the same boundary conditions. Assume now, that initial distributions of position and velocity are

$$u(x,0) = f(x) = 0 \quad \text{and} \quad u_t(x,0) = g(x) = \begin{cases} 0, & x \in [0, x_1]; \\ g_0, & x \in [x_1, x_2]; \\ 0, & x \in [x_2, L]\,. \end{cases}$$

Other parameters are:

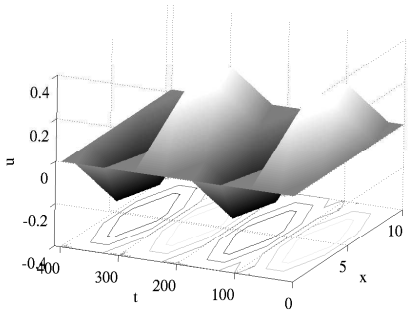

Fig. 7.5 Space-time evolution of Eq. (7.15) with the initial distribution $u(x,0)=0$, $u_t(x,0)=g(x)$.

`Initial nonzero velocity`	g_0=0.5
`Initial space intervals`	$x_1 = L/4,\ x_2 = 3L/4$
`Space interval`	L=10
`Space discretization step`	$\triangle x = 0.1$
`Time discretization step`	$\triangle t = 0.05$
`Amount of time steps`	$T = 400$

Numerical solution of the problem is shown on Fig. (7.5).

Example 3. Vibrating String

Use the explicit method (7.9) to solve the wave equation for a vibrating string:

$$u_{tt} = c^2 u_{xx} \quad \text{for} \quad x \in [0, L] \quad \text{and} \quad t \in [0, T]\,, \tag{7.16}$$

where $c = 1$ with the boundary conditions

$$u(0,t) = 0 \qquad u(L,t) = 0\,.$$

Assume that the initial position and velocity are

$$u(x,0) = f(x) = \sin(n\pi x/L)\,, \quad \text{and} \quad u_t(x,0) = g(x) = 0\,, \quad n = 1,2,3,\ldots\,.$$

Other parameters are:

`Space interval`	L=1
`Space discretization step`	$\triangle x = 0.01$
`Time discretization step`	$\triangle t = 0.0025$
`Amount of time steps`	$T = 2000$

Usually a vibrating string produces a sound whose frequency is constant. Therefore, since frequency characterizes the pitch, the sound produced is a constant note. Vibrating strings are the basis of any string instrument like guitar or cello. If the speed of propagation c is known, one can calculate the frequency of the sound produced by the string. The speed of propagation of a wave c is equal to the wavelength multiplied by the frequency f:

$$c = \lambda f$$

If the length of the string is L, the fundamental harmonic is the one produced by the vibration whose nodes are the two ends of the string, so L is half of the wavelength of the fundamental

harmonic, so

$$f = \frac{c}{2L}$$

Solutions of the equation in question are given in form of standing waves. The standing wave is a wave that remains in a constant position. This phenomenon can occur because the medium is moving in the opposite direction to the wave, or it can arise in a stationary medium as a result of interference between two waves traveling in opposite directions (see Fig. (7.6))

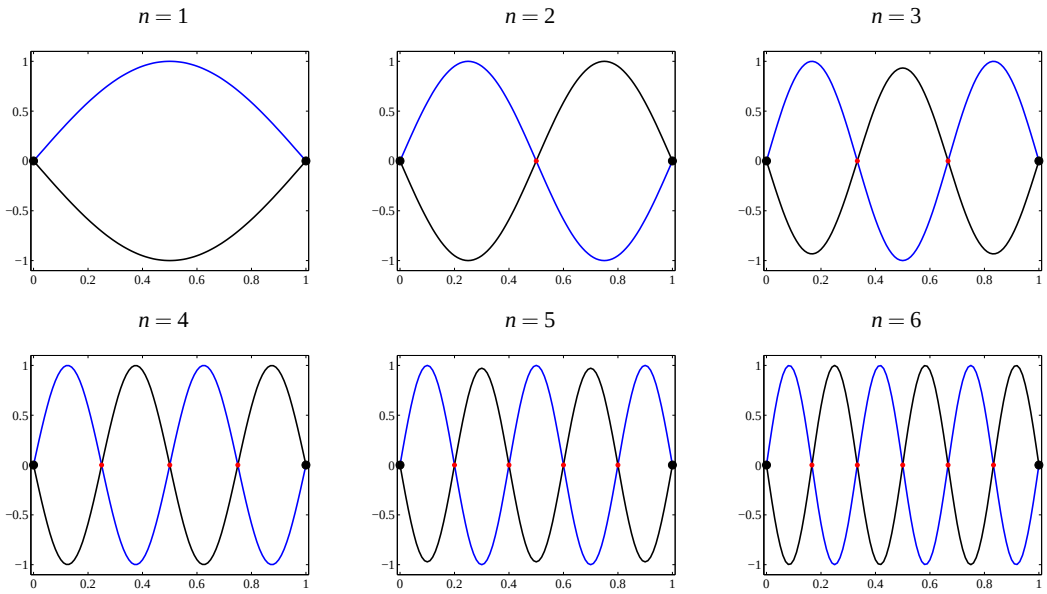

Fig. 7.6 Standing waves in a string. The fundamental mode and the first five overtones are shown. The red dots represent the wave nodes.

7.2 The Wave Equation in 2D

7.2.1 Examples

7.2.1.1 Example 1.

Use the standart five-point explicit method (7.9) to solve a two-dimansional wave equation

$$u_{tt} = c^2(u_{xx} + u_{yy}), \quad u = u(x,y,t)$$

on the rectangular domain $[0,L] \times [0,L]$ with Dirichlet boundary conditions. Other parameters are:

`Space interval`	L=1
`Space discretization step`	$\triangle x = \triangle y = 0.01$
`Time discretization step`	$\triangle t = 0.0025$
`Amount of time steps`	$T = 2000$
`Initial condition`	$u(x,y,0) = 4x^2 y(1-x)(1-y)$

Numerical solution of the problem for two different time moments $t = 0$ and $t = 500$ can be seen on Fig. (7.7)

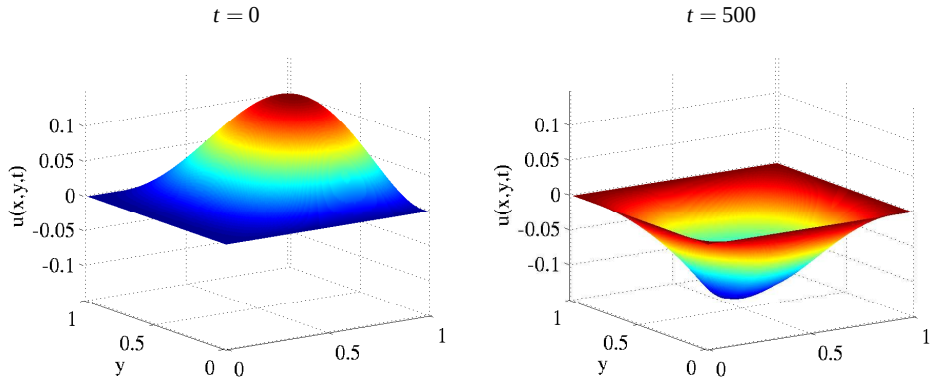

Fig. 7.7 Numerical solution of the two-dimensional wave equation, shown for $t = 0$ and $t = 500$.

Chapter 8
Sine-Gordon Equation

The sine-Gordon equation is a nonlinear hyperbolic partial differential equation involving the d'Alembert operator and the sine of the unknown function. The equation, as well as several solution techniques, were known in the nineteenth century in the course of study of various problems of differential geometry. The equation grew greatly in importance in the 1970s, when it was realized that it led to *solitons* (so-called "kink" and "antikink"). The sine-Gordon equation appears in a number of physical applications [27, 11, 46], including applications in relativistic field theory, Josephson junctions [39] or mechanical trasmission lines [42, 39].
The equation reads

$$u_{tt} - u_{xx} + \sin u = 0\,, \tag{8.1}$$

where $u = u(x,t)$. In the case of mechanical trasmission line, $u(x,t)$ describes an angle of rotation of the pendulums. Note that in the low-amplitude case ($\sin u \approx u$) Eq. (8.1) reduces to the Klein-Gordon equation

$$u_{tt} - u_{xx} + u = 0\,,$$

admiting solutions in the form

$$u(x,t) = u_0 \cos(kx - \omega t)\,, \qquad \omega = \sqrt{1+k^2}\,.$$

Here we are interested in large amplitude solutions of Eq. (8.1).

8.1 Kink and antikink solitons

Let us look for travelling wave solutions of the sine-Gordon equation (8.1) of the form

$$u(\xi) := u(x - ct)\,,$$

where c is an arbitrary velocity of propagation and $u \to 0$, $u_\xi \to 0$, when $\xi \to \pm\infty$ [39, 46]. In the co-moving frame Eq. (8.1) reads

$$(1-c^2)\,u_{\xi\xi} = \sin u\,.$$

Multiplying both sides of the last equation by u_ξ and integrating yields

$$\frac{1}{2}\,u_\xi^2\,(1-c^2) = -\cos u + c_1\,, \tag{8.2}$$

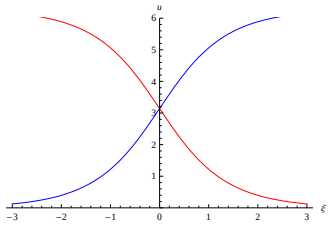

Fig. 8.1 Representation of the kink (blue) and antikink (red) solutions (8.4)

where c_1 is an arbitrary constant of integration. Notice that we look for solutions for which $u \to 0$ and $u_\xi \to 0$ when $\xi \to \pm\infty$, so $c_1 = 1$. Now we can rewrite the last equation as

$$\frac{d\,u}{\sin\frac{u}{2}} = \pm\frac{2}{\sqrt{1-c^2}}\,d\,\xi\,. \tag{8.3}$$

Integrating Eq. (8.3) yields

$$\pm\frac{2}{\sqrt{1-c^2}}\,(\xi-\xi_0) = 2\ln\left(\tan\frac{u}{4}\right),$$

or

$$u(\xi) = 4\arctan\left(\exp\left(\pm\frac{\xi-\xi_0}{\sqrt{1-c^2}}\right)\right).$$

That is, the solution of Eq. (8.1) becomes

$$u(x,t) = 4\arctan\left(\exp\left(\pm\frac{x-x_0-ct}{\sqrt{1-c^2}}\right)\right). \tag{8.4}$$

Equation (8.4) represents a localized solitary wave, travelling at any velocity $|c| < 1$. The $\pm$ signs correspond to localized solutions which are called *kink* and *antikink*, respectively. For the mechanical transmission line, when c increases from $-\infty$ to $+\infty$ the pendlums rotate from 0 to 2π for the kink and from 0 to -2π for the antikink. (see Fig. 8.1)

One can show [27, 39], that Eq. (8.1) admits more solutions of the form

$$u(x,t) = 4\arctan\left(\frac{F(x)}{G(t)}\right).$$

where F and G are arbitrary functions. Namely, one distinguishes the kink-kink and the kink-antikink collisions as well as the breather solution. The *kink-kink collision* solution reads

$$u(x,t) = 4\arctan\left(\frac{c\sinh\left(\frac{x}{\sqrt{1-c^2}}\right)}{\cosh\left(\frac{ct}{\sqrt{1-c^2}}\right)}\right) \tag{8.5}$$

and describes the collision between two kinks with respective velocities c and $-c$ and approaching the origin from $t \to -\infty$ and moving away from it with velocities $\pm c$ for $t \to \infty$ (see Fig. 8.2). In a similar way, one can construct solution, corresponding to the *kink-antikink collision*. The solution has the form:

$$u(x,t) = 4\arctan\left(\frac{\sinh\left(\frac{ct}{\sqrt{1-c^2}}\right)}{c\cdot\cosh\left(\frac{x}{\sqrt{1-c^2}}\right)}\right) \tag{8.6}$$

The *breather* soliton solution, which is also called a *breather mode* or *breather soliton* [39], is given by

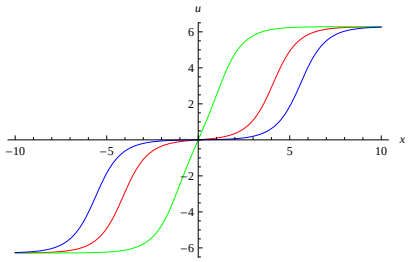

Fig. 8.2 The kink-kink collision, calculated at three different times: At $t=-7$ (red curve) both kinks propagate with opposite velocities $c=\pm 0.5$; At $t=0$ they collide at the origin (green curve); At $t=10$ (blue curve) they move away from the origin with velocities $c=\mp 0.5$.

$$u_B(x,t) = 4\arctan\left(\frac{\sqrt{1-\omega^2}\,\sin(\omega t)}{\omega\cosh(\sqrt{1-\omega^2}\,x)}\right) \tag{8.7}$$

which is periodic for frequencies $\omega < 1$ and decays exponentially when moving away from $x=0$. Now we are in the good position to look for numerical solutions of Eq. (8.1).

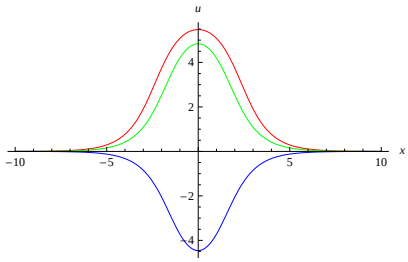

Fig. 8.3 The breather solution, oscillating with the frequency $\omega = 0.2$, calculated for three different times $t=0$ (red curve), $t=5$ (green curve) and $t=10$ (blue curve).

8.2 Numerical treatment

A numerical scheme

Consider an IVP for the sine-Gordon equation (8.1):

$$u_{tt} - u_{xx} + \sin(u) = 0$$

on the interval $x \in [a,\, b]$ with initial conditions

$$u(x,0) = f(x), \qquad u_t(x,0) = g(x)\,, \tag{8.8}$$

and with, e.g., no-flux boundary conditions

$$\left.\frac{\partial u}{\partial x}\right|_{x=a,\,b} = 0\,.$$

Let us try to apply a simple explicit scheme (7.9) to Eq. (8.1). The discretization scheme reads

$$u_i^{j+1} = -u_i^{j-1} + 2\,(1-\alpha^2)\,u_i^j + \alpha^2\,(u_{i+1}^j + u_{i-1}^j) - \triangle t^2\,\sin(u_i^j) \tag{8.9}$$

with $\alpha = \triangle t/\triangle x$, $i = 0,\ldots,M$ and $t = 0,\ldots,T$. To the implementation of the second initial condition one needs again the virtual point u_i^{-1},

$$u_t(x_i,0) = g(x_i) = \frac{u_i^1 - u_i^{-1}}{2\triangle t} + \mathscr{O}(\triangle t^2)\,.$$

Hence, one can rewrite the last expression as

$$u_i^{-1} = u_i^1 - 2\triangle t g(x_i) + \mathscr{O}(\triangle t^2)\,,$$

and the second time row u_i^1 can be calculated as

$$u_i^1 = \triangle t\, g(x_i) + (1-\alpha^2)\, f(x_i) + \frac{1}{2}\,\alpha^2 (f(x_{i-1}) + f(x_{i+1})) - \frac{\triangle t^2}{2}\sin(f(x_i))\,. \tag{8.10}$$

In addition, no-flux boundary conditions lead to the following expressions for two virtual space points u_{-1}^j and u_{M+1}^j:

$$\begin{aligned}
\left.\frac{\partial u}{\partial x}\right|_{x=a} &= 0 \Leftrightarrow \frac{u_1^j - u_{-1}^j}{2\triangle x} = 0 \Leftrightarrow u_{-1}^j = u_1^j\,,\\
\left.\frac{\partial u}{\partial x}\right|_{x=b} &= 0 \Leftrightarrow \frac{u_M^j - u_{M+1}^j}{2\triangle x} = 0 \Leftrightarrow u_{M+1}^j = u_M^j\,.
\end{aligned}$$

One can try to rewrite the differential scheme to more general matrix form. In matrix notation the second time-row is given by

$$\boxed{\mathbf{u}^1 = \triangle t\,\gamma_1 + A\mathbf{u}^0 - \frac{\triangle t^2}{2}\beta_1,} \tag{8.11}$$

where

$$\begin{aligned}
\gamma_1 &= \big(g(a), g(x_1), g(x_2), \ldots, g(x_{M-1}), g(b)\big)^T \quad \text{and}\\
\beta_1 &= \big(\sin(u_0^0), \sin(u_1^0), \ldots, \sin(u_{M-1}^0), \sin(u_M^0)\big)^T
\end{aligned}$$

are $M+1$-dimensional vectors and A is a tridiagonal square $M+1\times M+1$ matrix of the form

$$A = \begin{pmatrix}
1-\alpha^2 & \boxed{\alpha^2} & 0 & \ldots 0\\
\alpha^2/2 & 1-\alpha^2 & \alpha^2/2 & \ldots 0\\
0 & \alpha^2/2 & 1-\alpha^2 & \ldots 0\\
\ldots & \ldots & \ldots & \ldots\\
0 & \ldots & \boxed{\alpha^2} & 1-\alpha^2
\end{pmatrix}$$

The boxed elements indicate the influence of boundary conditions. Other time rows can also be written in the matrix form as

$$\boxed{\mathbf{u}^{j+1} = -\mathbf{u}^{j-1} + B\mathbf{u}^j - \triangle t^2\beta, \quad j = 1,\ldots,T-1} \tag{8.12}$$

Here

$$\beta = \big(\sin(u_0^j), \sin(u_1^j), \ldots, \sin(u_{M-1}^j), \sin(u_M^j)\big)^T$$

is a $M+1$-dimensional vector and B is a square matrix, defined by an equation

$$B = 2A.$$

Now we can apply the explicit scheme (8.9) described above to Eq. (8.1). Let us solve it on the interval $[-L, L]$ with no-flux boundary conditions using the following parameters set:

`Space interval`	L=20
`Space discretization step`	$\triangle x = 0.1$
`Time discretization step`	$\triangle t = 0.05$
`Amount of time steps`	$T = 1800$
`Velocity of the kink`	$c = 0.2$

We start with the numerical representation of kink and antikink solutions. The initial condition for the kink is

$$f(x) = 4\arctan\left(\exp\left(\frac{x}{\sqrt{1-c^2}}\right)\right),$$
$$g(x) = -2\frac{c}{\sqrt{1-c^2}}\operatorname{sech}\left(\frac{x}{\sqrt{1-c^2}}\right).$$

Figure 8.4 (a) shows the space-time plot of the numerical kink solution. For the antikink the initial condition reads

$$f(x) = 4\arctan\left(\exp\left(-\frac{x}{\sqrt{1-c^2}}\right)\right),$$
$$g(x) = -2\frac{c}{\sqrt{1-c^2}}\operatorname{sech}\left(\frac{x}{\sqrt{1-c^2}}\right).$$

Numerical solutions is shown on Fig. 8.4 (b).

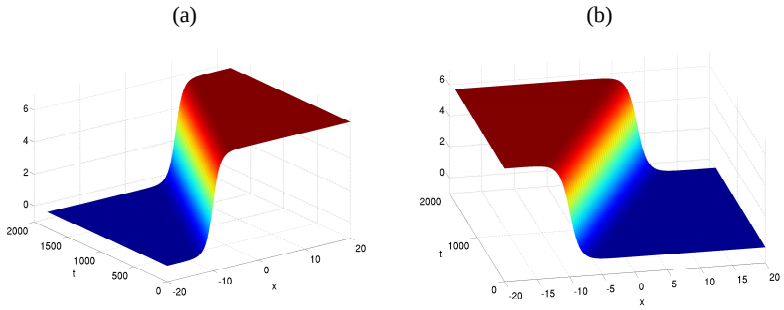

Fig. 8.4 Numerical solution of Eq. (8.1), calculated with the scheme (8.9) for the case of (a) the kink and (b) antikink solitons, moving with the velocity $c = 0.2$. Space-time information is shown.

Now we are in position to find numerical solutions, corresponding to kink-kink and kink-antikink collisions. For the kink-kink collision we choose

$$f(x) = 4\arctan\left(\exp\left(\frac{x+L/2}{\sqrt{1-c^2}}\right)\right) + 4\arctan\left(\exp\left(\frac{x-L/2}{\sqrt{1-c^2}}\right)\right),$$
$$g(x) = -2\frac{c}{\sqrt{1-c^2}}\operatorname{sech}\left(\frac{x+L/2}{\sqrt{1-c^2}}\right) + 2\frac{c}{\sqrt{1-c^2}}\operatorname{sech}\left(\frac{x-L/2}{\sqrt{1-c^2}}\right),$$

whereas for the kink-antikink collision the initial conditions are

$$f(x) = 4\arctan\left(\exp\left(\frac{x+L/2}{\sqrt{1-c^2}}\right)\right) + 4\arctan\left(\exp\left(-\frac{x-L/2}{\sqrt{1-c^2}}\right)\right),$$
$$g(x) = -2\frac{c}{\sqrt{1-c^2}}\,\mathrm{sech}\left(\frac{x+L/2}{\sqrt{1-c^2}}\right) - 2\frac{c}{\sqrt{1-c^2}}\,\mathrm{sech}\left(\frac{x-L/2}{\sqrt{1-c^2}}\right).$$

Numerical solutions, corresponding to both cases is presented on Fig. 8.5 (a)-(b), respectively. Finally, for the case of breather we choose

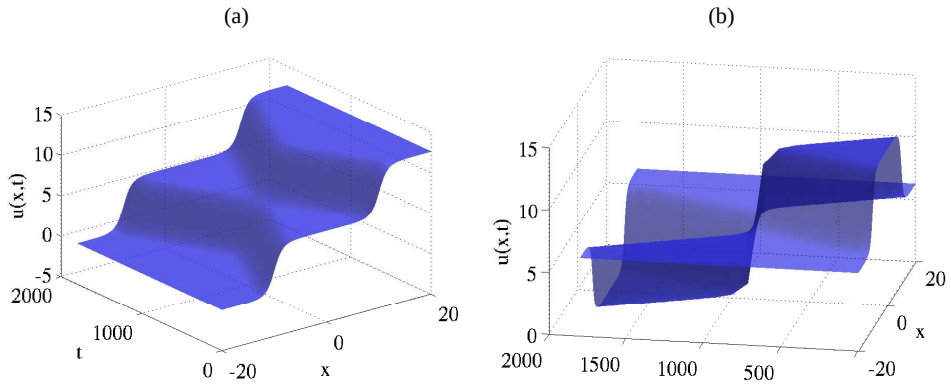

Fig. 8.5 Space-time representation of the numerical solution of Eq. (8.1) for (a) kink-kink collision and (b) kink-antikink collision.

$$f(x) = 0,$$
$$g(x) = 4\sqrt{1-c^2}\,\mathrm{sech}\left(x\sqrt{1-c^2}\right).$$

Corresponding numerical solution is presented on Fig. 8.6.

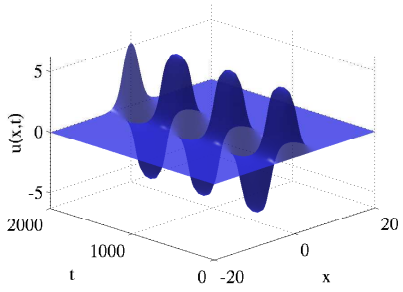

Fig. 8.6 Space-time plot of the numerical breather solution, oscillating with the frequency $\omega = 0.2$.

Chapter 9
Korteweg-de Vries Equation

The Korteweg-de Vries (KdV) equation is the partial differential equation, derived by Korteweg and de Vries [26] to describe weakly nonlinear shallow water waves. The nondimensionalized version of the equation reads

$$\frac{\partial u}{\partial t} = 6u\frac{\partial u}{\partial x} - \frac{\partial^3 u}{\partial x^3}\,, \tag{9.1}$$

where $u = u(x,t)$. The factor of 6 is convenient for reasons of complete integrability, but can easily be scaled out if desired. Equation (9.1) was found to have *solitary wave solutions*, vindicating the observations of a solitary channel wave made by Russell [41].

9.1 Traveling wave solution

In the same way as in Sec. 8.1 we look for a right traveling wave solution of the form [46]

$$u(\xi) := u(x - ct)\,,$$

such as $u \to 0$, $u_\xi \to 0$ and $u_{\xi\xi} \to 0$ as $\xi \to \pm\infty$. Substitution into Eq. (9.1) leads to the ODE

$$u_{\xi\xi\xi} - 6uu_\xi - cu_\xi = 0\,.$$

An integration with respect to ξ yields

$$u_{\xi\xi} = 3u^2 + cu + c_1\,,$$

where c_1 is a constant of integration. Since $u \to 0$, $u_\xi \to 0$ and $u_{\xi\xi} \to 0$ as $\xi \to \pm\infty$, $c_1 = 0$. A second integration yields

$$\frac{1}{2}u_\xi^2 = u^3 + \frac{1}{2}cu^2 + c_2\,,$$

where $c_2 = \text{const} = 0$. That is, the last equation can be written as

$$d\xi = \frac{du}{u\sqrt{2u+c}}\,,$$

which can be integrated, yielding

$$u(\xi) = -\frac{c}{2}\,\mathrm{sech}^2\left(\frac{1}{2}\sqrt{c}\,(\xi-\xi_0)\right),$$

where ξ_0 is an arbitrary constant. In (x,t) coordinates the traveling wave solution reads

$$u(x,t) = -\frac{c}{2}\,\mathrm{sech}^2\left(\frac{1}{2}\sqrt{c}\,(x-x_0-ct)\right). \tag{9.2}$$

Equation (9.2) describes the localized traveling wave solution with a negative amplitude (see Fig. 9.1 (a)), which is called *a soliton*. The term soliton was first introduced by Zabusky and Kruskal [53], who studied Eq. (9.1) with periodic boundary conditions numerically. They found [53, 46, 27] that initial condition of the form $u(x,0) = \cos(2\pi x/L)$, $x \in [0,L]$ broke up into a train of solitary waves with successively large amplitude. Moreover the solitons seems to be almost unaffected in shape by passing through each other (though this could cause a change in their position). An example of two-soliton solution is shown on Fig. 9.1 (b).

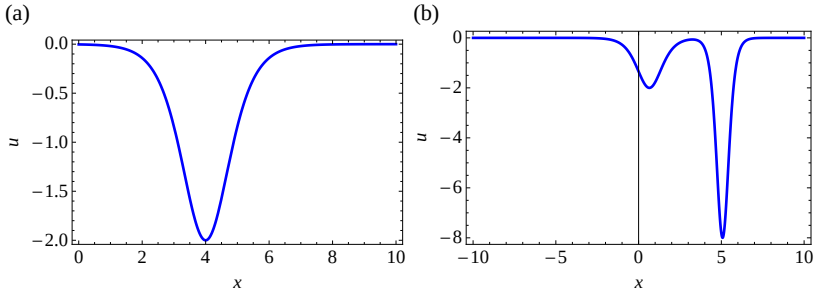

Fig. 9.1 Solitary solutions of KdV equation (9.1). (a)A single-soliton solution (9.2) for $c = 2$, calculated for $t = 1$. (b) A two-soliton solution, calculated at $t = 0.3$.

9.2 Numerical treatment

Consider the KdV Eq. (9.1) on the interval $x \in [-L, L]$ with initial condition

$$u(x,0) = f(x) := -N(N+1)\,\mathrm{sech}^2(x),$$

where N is an amount of solitons and periodic boundary conditions [24]. The first idea is to apply central difference to spatial derivatives on the right hand side and forward difference to the time derivative on the left, as in contrast to the wave equation (7.1) or the sine-Gordon equation (8.1) only the information about initial position of u is known and the artificial point u_i^{-1} can not be calculated. That is, the simple explicit schema reads:

$$\frac{u_i^{j+1}-u_i^j}{\triangle t} = 3u_i^j\,\frac{u_{i+1}^j-u_{i-1}^j}{\triangle x} - \frac{u_{i+2}^j-2u_{i+1}^j+2u_{i-1}^j-u_{i-2}^j}{2\triangle x^3},$$

or, with $h = \triangle t/\triangle x$

$$u_i^{j+1} = u_i^j + 3hu_i^j\left(u_{i+1}^j-u_{i-1}^j\right) - \frac{h}{2\triangle x^2}\left(u_{i+2}^j-2u_{i+1}^j+2u_{i-1}^j-u_{i-2}^j\right). \tag{9.3}$$

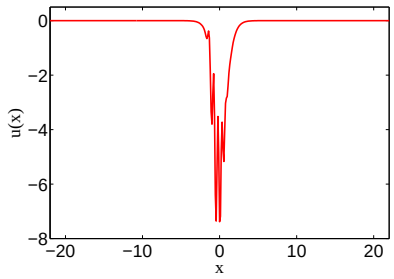

Fig. 9.2 Numerical solution of the KdV Eq. (9.1) on the interval $t \in [-22, 22]$, using the explicit schema (9.3). Space and time discretization steps are $\triangle x = 0.11$, $\triangle t = 5e-4$, respectively. After several intergation steps numerical instability can be observed.

Since Eq. (9.1) is nonlinear, the direct verification of the stability of the scheme (9.3) with the help of von Neumann analysis (see Sec. 4.3). However, one can examine the stability of the *liner equation*

$$u_t = -u_{xxx}\,. \tag{9.4}$$

Using the usual ansatz (4.21) the following criterium for Eq. (9.4) can be obtained [24]

$$\triangle t \le \frac{1}{m}\triangle x^3\,, \tag{9.5}$$

where

$$m = \max|\sin(2k\triangle x) - 2\sin(k\triangle x)| = \frac{3\sqrt{3}}{2} \simeq 2.6\,.$$

That is, the linear equation (9.4) is conditionally stable, what is not surprising for explicit schemata. However, if we apply the schema (9.3), one can see that after several intergation steps a numerical instability occurs (see Fig. 9.2). That is, the schema (9.3) is unstable and has to be modified.

The first idea is to modify the relation for the time derivative on the right hand side. As was mentioned above, the direct usage of the central difference formula is impossible due to initial condition. On the other hand, the artificial point u_i^{-1} is essential only on the first time step. Hence, on the first time step ($j = 0$) the schema (9.3) can be used, whereas for $j = 1, \ldots, T$ the central difference formula is applied:

$$\frac{\partial u}{\partial t} \to \frac{u_i^{j+1} - u_i^{j-1}}{2\triangle t}\,.$$

In addition, we replace u_i^j on the right hand side by the average, namely

$$u_i^j \to \frac{1}{3}\left(u_{i-1}^j + u_i^j + u_{i+1}^j\right).$$

That is, the final modified schema reads

$$u_i^{j+1} = u_i^{j-1} + 2h\left(u_{i-1}^j + u_i^j + u_{i+1}^j\right)\left(u_{i+1}^j - u_{i-1}^j\right) - \frac{h}{\triangle x^2}\left(u_{i+2}^j - 2u_{i+1}^j + 2u_{i-1}^j - u_{i-2}^j\right). \tag{9.6}$$

Let us apply the modified schema (9.6) to Eq. (9.1) for the case of two-soliton solution. That is, we solve Eq. (9.1 on the interval $x \in [-L, L]$ according to

`Space interval`	$L = 10$
`Space discretization step`	$\triangle x = 0.18$
`Time discretization step`	$\triangle t = 2e-3$
`Amount of time steps`	$T = 1e+5$

We start with initial condition

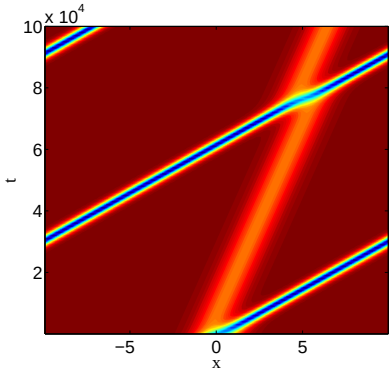

Fig. 9.3 Space-time representation of the numerical two-soliton solution of the KdV Eq. (9.1) on the interval $t \in [-10, 10]$, using the modified schema (9.6). Space and time discretization steps are $\triangle x = 0.18$, $\triangle t = 2e-5$, respectively.

$$u(x,0) = f(x) := -6\,\mathrm{sech}^2(x)\,,$$

and apply periodic boundary condition. Notice that in the presented case the linear stability condition (9.5) is fulfilled. The result of calculation is presented on Fig. 9.3. The localized initial condition decomposes into two solitons with different depths and velocities, moving in the same direction. In addition, the solitons collide at some time moment, and the deeper soliton overtank the smaller one.

Chapter 10
The Diffusion Equation

The diffusion equation is a partial differential equation which describes density fluctuations in a material undergoing diffusion. The equation can be written as:

$$\frac{\partial u(\mathbf{r},t)}{\partial t} = \nabla \cdot \big(D(u(\mathbf{r},t),r)\nabla u(\mathbf{r},t)\big)\,, \tag{10.1}$$

where $u(\mathbf{r},t)$ is the density of the diffusing material at location $\mathbf{r} = (x,y,z)$ and time t. $D(u(\mathbf{r},t),r)$ denotes the collective *diffusion coefficient* for density u at location $\mathbf{r}$. If the diffusion coefficient doesn't depend on the density, i.e., D is constant, then Eq. (10.1) reduces to the following linear equation:

$$\frac{\partial u(\mathbf{r},t)}{\partial t} = D\nabla^2 u(\mathbf{r},t)\,. \tag{10.2}$$

Equation (10.2) is also called *the heat equation* and also describes the distribution of a heat in a given region over time.

Equation (10.2) can be derived in a straightforward way from the *continuity equation*, which states that a change in density in any part of the system is due to inflow and outflow of material into and out of that part of the system. Effectively, no material is created or destroyed:

$$\frac{\partial u}{\partial t} + \nabla \cdot \Gamma = 0\,,$$

where Γ is the flux of the diffusing material. Equation (10.2) can be obtained easily from the last equation when combined with the phenomenological Fick's first law, which assumes that the flux of the diffusing material in any part of the system is proportional to the local density gradient:

$$\Gamma = -D\nabla u(\mathbf{r},t)\,.$$

10.1 The Diffusion Equation in 1D

Consider an IVP for the diffusion equation in one dimension:

$$\frac{\partial u(x,t)}{\partial t} = D\,\frac{\partial^2 u(x,t)}{\partial x^2} \tag{10.3}$$

on the interval $x \in [0,\, L]$ with initial condition

$$u(x,0) = f(x), \qquad \forall x \in [0,L] \tag{10.4}$$

and Dirichlet boundary conditions

$$u(0,t) = u(L,t) = 0 \qquad \forall t > 0\,. \tag{10.5}$$

10.1.1 Analytical Solution

Let us attempt to find a nontrivial solution of (10.3) satisfying the boundary conditions (10.5) using separation of variables [7], i.e.,

$$u(x,t) = X(x)T(t)\,.$$

Substituting u back into Eq. (10.3) one obtains:

$$\frac{1}{D}\frac{T'(t)}{T(t)} = \frac{X''(x)}{X(x)}\,.$$

Since the right hand side depends only on x and the left hand side only on t, both sides are equal to some constant value $-\lambda$ ($-$ sign is taken for convenience reasons). Hence, one can rewrite the last equation as a system of two ODE's:

$$X''(x) + \lambda X(x) = 0, \tag{10.6}$$

$$T'(t) + D\lambda T(t) = 0. \tag{10.7}$$

Let us consider the first equation for $X(x)$. Taking into account the boundary conditions (10.5) one obtains ($T(t) \neq 0$ as we are loocking for nontrivial solutions)

$$u(0,t) = X(0)T(t) = 0 \Rightarrow X(0) = 0\,,$$
$$u(L,t) = X(L)T(t) = 0 \Rightarrow X(L) = 0\,.$$

That is, the problem of finding of the solution of (10.3) reduces to the solving of linear ODE and consideration of three different cases with respect to the sign of λ:

1. $\lambda < 0$:
$$X(x) = C_1 e^{\sqrt{-\lambda}x} + C_2 e^{-\sqrt{-\lambda}x}\,.$$
Taking into account the boundary conditions one gets $C_1 = C_2 = 0$, so for $\lambda < 0$ only the trivial solution exists.
2. $\lambda = 0$:
$$X(x) = C_1 x + C_2$$
Again, due to the boundary conditions, one gets only trivial solution of the problem ($C_1 = C_2 = 0$).
3. $\lambda > 0$:
$$X(x) = C_1 \cos(\sqrt{\lambda}x) + C_2 \sin(\sqrt{\lambda}x)\,.$$
Substituting of the boundary conditions leads to the following equations for the constants C_1 and C_2:
$$X(0) = C_1 = 0\,,$$
$$X(L) = C_2 \sin(\sqrt{\lambda}L) = 0 \Rightarrow \sin(\sqrt{\lambda}L) = 0 \Rightarrow \lambda_n = \left(\frac{\pi n}{L}\right)^2, \quad n = 1,2,\ldots.$$
Hence,

$$X(t) = C_n \sin\left(\frac{\pi n}{L} x\right).$$

That is, the second equation for the function $T(t)$ takes the form:

$$T'(t) + D\left(\frac{\pi n}{L}\right) T(t) = 0 \Rightarrow T(t) = B_n \exp\left(-D\left(\frac{\pi n}{L}\right)^2 t\right),$$

where B_n is constant.

Altogether, the general solution of the problem (10.3) can be written as

$$u(x,t) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{\pi n}{L} x\right) \exp\left(-D\left(\frac{\pi n}{L}\right)^2 t\right), \quad A_n = \text{const}.$$

In order to find A_n one can use the initial condition (10.4). Indeed, if we write the function $f(x)$ as a Fourier series, we obtain:

$$f(x) = \sum_{n=1}^{\infty} F_n \sin(\frac{\pi n}{L} x) = \sum_{n=1}^{\infty} A_n \sin(\frac{\pi n}{L} x),$$
$$A_n = F_n = \frac{2}{L} \int_0^L f(\xi) \sin(\frac{\pi n}{L} \xi)\, d\xi.$$

Hence, the genetal solution of Eq. (10.3) reads:

$$\boxed{u(x,t) = \sum_{n=1}^{\infty} \left(\frac{2}{L} \int_0^L f(\xi) \sin\left(\frac{\pi n}{L} \xi\right) d\xi\right) \sin\left(\frac{\pi n}{L} x\right) \exp\left(-D(\frac{\pi n}{L})^2 t\right).} \tag{10.8}$$

10.1.2 Numerical Treatment

The FTCS Explicit Method

Consider Eq. (10.3) with the initial condition (10.4). The first simple idea is an explicit forward in time, central in space (FTCS) method [47, 40] (see Fig. (10.1)):

$$\frac{u_i^{j+1} - u_i^j}{\triangle t} = D \frac{u_{i+1}^j - 2u_i^j + u_{i-1}^j}{\triangle x^2},$$

or, with $\alpha = D\frac{\triangle t}{\triangle x^2}$

$$\boxed{u_i^{j+1} = (1 - 2\alpha) u_i^j + \alpha (u_{i+1}^j + u_{i-1}^j).} \tag{10.9}$$

In order to check the stability of the schema (10.9) we apply again the ansatz (4.21) (see Sec. 4.3), considering a single Fourier mode in x space and obtain the following equation for the amplification factor $g(k)$:

$$g^2 = (1 - 2\alpha) g + 2g\alpha \cos(k\triangle x),$$

from which

$$g(k) = 1 - 4\alpha \sin^2 \frac{k\triangle x}{2}.$$

The stability condition for the method (10.9) reads

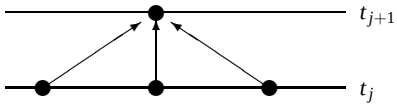

Fig. 10.1 Schematical representation of the FTCS finite difference scheme (10.9) for solving the 1-d diffusion equation (10.3).

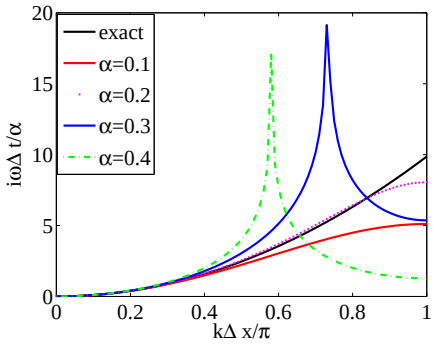

Fig. 10.2 Dispersion relation by means of the schema (10.9) for different values of α, compared with the exact dispersion relation for Eq. (10.3).

$$|g(k)| \leq 1 \quad \forall k \Leftrightarrow \alpha \leq \frac{1}{2} \Leftrightarrow \boxed{\triangle t \leq \frac{1}{2}\frac{\triangle x^2}{D}\,.} \tag{10.10}$$

Although the method (10.9) is conditionally stable, the derived stability condition (10.10), however, hides an uncomfortable property: A doubling of the spatial resolution $\triangle x$ requires a simultaneous reduction in the time-step $\triangle t$ by a factor of four in order to maintain numerical stability. Certainly, the above constraint limits us to absurdly small time-steps in high resolution calculations.

The next point to emphasize is the numerical dispersion. Indeed, let us compare the exact dispersion relation for Eq. (10.3) and relation, obtained by means of the schema (10.9). If we consider the perturbations in form $\exp(ikx - i\omega t)$ the dispersion relation for Eq. (10.3) reads

$$i\omega = Dk^2\,.$$

On the other hand, the FTCS schema (10.3) leads to the following relation

$$e^{i\omega\triangle t} = 1 - 4\alpha\sin^2\left(\frac{k\triangle x}{2}\right),$$

or, in other words

$$\boxed{i\omega\triangle t = -\ln\left(1 - 4\alpha\sin^2\left(\frac{k\triangle x}{2}\right)\right).}$$

The comparison between exact and numerical disperion relations is shown on Fig. (10.2). One can see, that both relations are in good agreement only for $k\triangle x \ll 1$. For $\alpha > 0.25$ the method is stable, but the values of ω can be complex, i.e., the Fourier modes drops off, performing damped oscillations (see Fig. (10.2) for $\alpha = 0.3$ and $\alpha = 0.4$). Now, if we try to make the time step smaler, in the limit $\triangle t \to 0$ (or $\alpha \to 0$) we obtain

$$i\omega\triangle t \approx 4\alpha\sin^2\left(\frac{k\triangle x}{2}\right) = k^2 D\triangle t \frac{\sin^2\left(\frac{k\triangle x}{2}\right)}{\left(\frac{k\triangle x}{2}\right)^2},$$

i.e., we get the correct dispersion relation only if the space step $\triangle x$ is small enougth too.

The Richardson Method

The first idea to improve the approximation order of the schema is to use the central diferences for the time derivative of Eq. (10.3), namely [47]

$$\frac{u_i^{j+1} - u_i^{j-1}}{2\triangle t} = D\frac{u_{i+1}^j - 2u_i^j + u_{i-1}^j}{\triangle x^2},$$

or, with $\alpha = D\triangle t/\triangle x^2$

$$\boxed{u_i^{j+1} = u_i^{j-1} + 2\alpha\left(u_{i+1}^j - 2u_i^j + u_{i-1}^j\right).} \tag{10.11}$$

Unfortunately, one can show that the schema (10.11) is unconditional unstable. Indeed, amplification factor $g(k)$ in this case fulfilles the following equation:

$$g^2 + 2\beta g - 1 = 0, \qquad \beta = 4\alpha\sin^2\frac{k\triangle x}{2},$$

giving

$$g_{1,2} = -\beta \pm \sqrt{\beta^2 + 1}.$$

Since $|g_2(k)| > 1$ for all values of k, the schema (10.11) is absolut unstable.

The DuFort-Frankel Method

Let us consider one of many alternative algorithms which have been designed to overcome the stability problems of the simple FTCS and Richardson methods. We modify Eq. (10.9) as (see Fig. (10.3)) [47]

$$\frac{u_i^{j+1} - u_i^{j-1}}{2\triangle t} = D\frac{u_{i+1}^j - 2\frac{u_i^{j+1}+u_i^{j-1}}{2} + u_{i-1}^j}{\triangle x^2},$$

which can be solved explicitly for u_i^{j+1}:

$$\boxed{u_i^{j+1} = \frac{1-\alpha}{1+\alpha}u_i^{j-1} + \frac{\alpha}{1+\alpha}\left(u_{i+1}^j + u_{i-1}^j\right),} \tag{10.12}$$

where $\alpha = 2D\triangle t/\triangle x$. When the usual von Neumann stability analysis is applied to the method (10.12), the amplification factor $g(k)$ can be found from

$$(1+\alpha)g^2 - 2g\alpha\cos(k\triangle x) + (\alpha - 1) = 0.$$

It can be easily shown, that stability condition is fulfilled for all values of α, so the method (10.12) is unconditionally stable. However, this does not imply that $\triangle x$ and $\triangle t$ can be made indefinitely large; we must still worry about the accuracy of the method. Indeed, consider the Taylor expansion

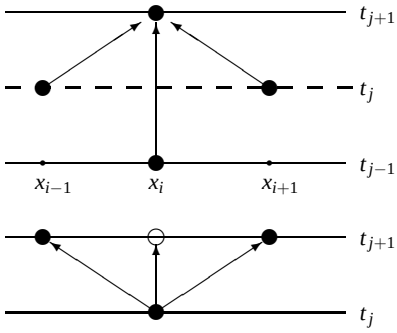

Fig. 10.3 Schematical representation of the DuFort-Frankel method (10.12).

Fig. 10.4 Schematical representation of the implicit BTCS method (10.13).

for Eq. (10.3) by means of (10.12):

$$\begin{aligned}
\frac{u_i^{j+1}-u_i^{j-1}}{2\triangle t} &= D\frac{u_{i+1}^{j}-u_i^{j+1}-u_i^{j-1}+u_{i-1}^{j}}{\triangle x^2} \Leftrightarrow \\
u_t - \frac{\triangle x^2}{3!}u_{ttt}+\ldots &= \frac{D}{\triangle x^2}\left(\triangle x^2 u_{xx}+\frac{2\triangle x^4}{4!}u_{xxxx}-\triangle t^2 u_{tt}-\frac{2\triangle t^4}{4!}u_{tttt}+\ldots\right) \Leftrightarrow \\
u_t+\mathscr{O}(\triangle t^2) &= Du_{xx}+\mathscr{O}(\triangle x^2)-D\left(\frac{\triangle t^2}{\triangle x^2}\right)u_{tt}+\mathscr{O}\left(\frac{\triangle t^4}{\triangle x^2}\right).
\end{aligned}$$

In other words, the method (10.12) has order of accuracy

$$\mathscr{O}\left(\triangle t^2, \triangle x^2, \frac{\triangle t^2}{\triangle x^2}\right).$$

For cosistency, $\triangle t/\triangle x \to 0$ as $\triangle t \to 0$ and $\triangle x \to 0$, so (10.12) is inconsistent. This constitutes an effective restriction on $\triangle t$. For large $\triangle t$, however, the scheme (10.12) is consistent with *another* equation of the form

$$Du_{tt}+u_t = Du_{xx}.$$

10.1.2.1 The BTCS Implicit Method

One can try to overcome problems, described above by introducing an implicit method. The simplest example is a BTCS (backward in time, central in space) method (see Fig. 10.4) [48]. The differential schema reads:

$$\frac{u_i^{j+1}-u_i^{j}}{\triangle t} = D\,\frac{u_{i+1}^{j+1}-2u_i^{j+1}+u_{i-1}^{j+1}}{\triangle x^2}+\mathscr{O}(\triangle t, \triangle x^2),$$

or, with $\alpha = D\triangle t/\triangle x^2$

$$\boxed{-u_i^j = \alpha u_{i+1}^{j+1}-(1+2\alpha)u_i^{j+1}+\alpha u_{i-1}^{j+1}.} \tag{10.13}$$

In this case the amplification factor $g(k)$ is given by

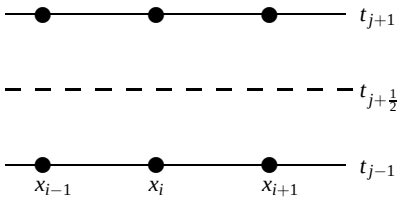

Fig. 10.5 Schematical representation of the Crank-Nicolson method (10.14).

$$g(k) = \left(1 + 4\alpha \sin^2 \frac{k\triangle x^2}{2}\right)^{-1}.$$

That is, the schema (10.13) is unconditionally stable. However, the method has order of accuracy $\mathcal{O}(\triangle t, \triangle x^2)$, i.e., first order in time, and second in space. Is it possible to improve it? The answer to is given below.

The Crank-Nicolson Method

An implicit scheme, introduced by J. Crank and P. Nicolson in 1947 [9] is based on the central approximation of Eq. (10.3) at the point $(x_i, t_j + \frac{1}{2}\triangle t)$ (see Fig. (10.5)):

$$\frac{u_i^{j+1} - u_i^j}{2\frac{\triangle t}{2}} = D\,\frac{u_{i+1}^{j+\frac{1}{2}} - 2\,u_i^{j+\frac{1}{2}} + u_{i-1}^{j+\frac{1}{2}}}{\triangle x^2}.$$

The approximation used for the space derivative is just an average of approximations in points (x_i, t_j) and (x_i, t_{j+1}):

$$\frac{u_i^{n+1} - u_i^n}{\triangle t} = D\,\frac{(u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}) + (u_{i+1}^n - 2u_i^n + u_{i-1}^n)}{2\triangle x^2}.$$

Introducing $\alpha = D\triangle t/\triangle x^2$ one can rewrite the last equation as

$$\boxed{-\alpha u_{i+1}^{j+1} + 2(1+\alpha)u_i^{j+1} - \alpha u_{i-1}^{j+1} = \alpha u_{j+1}^n + 2(1-\alpha)u_i^j + \alpha u_{j-1}^n.} \qquad (10.14)$$

All terms on the right-hand side of Eq. (10.14) are known. Hence, the equations in (10.14) form a tridiagonal linear system

$$A\mathbf{u} = \mathbf{b}.$$

The amplification factor for Eq. (10.14) reads

$$g(k) = \frac{1 - \alpha(1 - \cos k\triangle x)}{1 + \alpha(1 - \cos k\triangle x)}.$$

Since α and $1 - \cos k\triangle x$ are positive, the denominator of the last expression is always greater than the numerator. That is, the absolute value of g is less than one, i.e., the method (10.14) is unconditionaly stable.

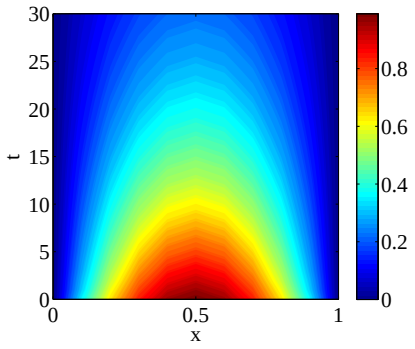

Fig. 10.6 Contour plot of the heat distribution after the time $T = 30$, calculated with the FTCS method (10.9).

10.1.3 Examples

Example 1

Use the FTCS explicit method (10.9) to solve the one-dimensional heat equation

$$u_t = u_{xx},$$

on the interval $x \in [0, L]$, if the initial heat distribution is given by $u(x,0) = f(x)$, and the temperature on both ends of the interval is $u(0,t) = T_l$, $u(L,t) = T_r$. Other parameters are choosen according to the table below:

`Space interval`	$L = 1$
`Amount of space points`	$M = 10$
`Amount of time steps`	$T = 30$
`Boundary conditions`	$T_l = T_r = 0$
`Initial heat distribution`	$f(x) = 4x(1-x)$

The result of the calculation is shown on Fig 10.6.

Example 2

Use the implicit BTCS method (10.13) to solve the one-dimensional diffusion equation

$$u_t = u_{xx},$$

on the interval $x \in [-L, L]$, if the initial distribution is a Gauss pulse of the form $u(x,0) = \exp(-x^2)$ and the density on both ends of the interval is given as $u_x(-L,t) = u_x(L,t) = 0$. For the other parameters see the table below:

`Space interval`	$L = 5$
`Space discretization step`	$\triangle x = 0.1$
`Time discretization step`	$\triangle t = 0.05$
`Amount of time steps`	$T = 200$

Solution of the problem is shown on Fig. (10.7).

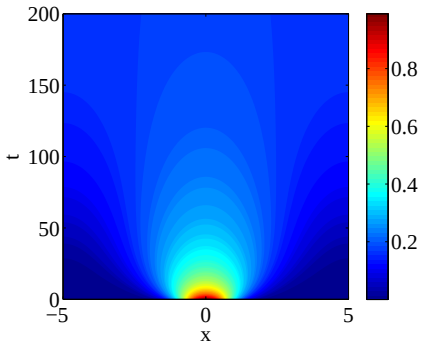

Fig. 10.7 Contour plot of the diffusion of the initial Gauss pulse, calculated with the BTCS implicit method (10.13).

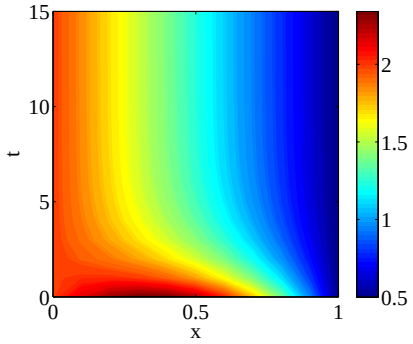

Fig. 10.8 Contour plot of the heat distribution, calculated with the Crank-Nicolson method (10.14).

Example 3

Use the Crank-Nicolson method (10.14) to solve the one-dimensional heat equation

$$u_t = 1.44\,u_{xx},$$

on the interval $x \in [0, L]$, if the initial heat distribution is $u(x,0) = f(x)$ and again, the temperature on both ends of the interval is given as $u(0,t) = T_l$, $u(L,t) = T_r$. Other parameters are choosen as:

`Space interval`	L=1
`Space discretization step`	$\triangle x = 0.1$
`Time discretization step`	$\triangle t = 0.05$
`Amount of time steps`	$T = 15$
`Boundary conditions`	$T_l = 2$, $T_r = 0.5$
`Initial heat distribution`	$f(x) = 2 - 1.5x + \sin(\pi x)$

Numerical solution of the problem in question is shown on Fig. (10.8).

10.2 The Diffusion Equation in 2D

Let us consider the solution of the diffusion equation (10.2) in two dimensions

$$\frac{\partial u}{\partial t} = D\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right), \tag{10.15}$$

where $u = u(x, y, t)$, $x \in [a_x, b_x]$, $y \in [a_y, b_y]$. Suppose, that the inittial condition is given and function u satisfies boundary conditions in both x- and in y-directions.
As before, we discretize in time on the uniform grid $t_n = t_0 + n\triangle t$, $n = 0, 1, 2, \ldots$. Futhermore, in the both x- and y-directions, we use the uniform grid

$$x_i = x_0 + i\triangle x, \quad i = 0, \cdots, M, \quad \triangle x = \frac{b_x - a_x}{M+1},$$
$$y_j = y_0 + j\triangle y, \quad j = 0, \cdots, N, \quad \triangle y = \frac{b_y - a_y}{N+1}.$$

10.2.1 Numerical Treatment

The FTCS Method in 2D

In the case of two dimensions the explicit FTCS scheme reads

$$\frac{u_{ij}^{n+1} - u_{ij}^n}{\triangle t} = D\left(\frac{u_{i+1j}^n - 2u_{ij}^n + u_{i-1j}^n}{\triangle x^2} + \frac{u_{ij+1}^n - 2u_{ij}^n + u_{ij-1}^n}{\triangle y^2}\right),$$

or, with $\alpha = D\triangle t/\triangle x^2$ and $\beta = D\triangle t/\triangle y^2$

$$\boxed{u_{ij}^{n+1} = \alpha(u_{i+1j}^n + u_{i-1j}^n) + \beta(u_{ij+1}^n + u_{ij-1}^n) + (1 - 2\alpha - 2\beta)u_{ij}^n.} \tag{10.16}$$

The ansatz

$$\varepsilon_{ij}^n = g^n e^{i(k_x x_i + k_y x_j)}$$

leads to the following relation for the amplification factor $g(k)$

$$g(k) = 1 - 4\alpha \sin^2\left(\frac{k_x\triangle x}{2}\right) - 4\beta \sin^2\left(\frac{k_y\triangle y}{2}\right).$$

In this case the stability condition reads

$$\alpha + \beta \le \frac{1}{2}. \tag{10.17}$$

This stability condition imposes a limit on the time step:

$$\boxed{\triangle t \le \frac{\triangle x^2 \triangle y^2}{2D(\triangle x^2 + \triangle y^2)}.}$$

In particular, for the case $\triangle x = \triangle y$ we have

$$\triangle t \le \frac{\triangle x^2}{4D},$$

which is even more restrictive, than in the one-dimensional case.

The BTCS Method in 2D

To overcome the stability restriction of the FTCS method (10.16), we can use an implicit BTCS schema in the two-dimensional case. The schema reads:

$$\frac{u_{ij}^{n+1}-u_{ij}^{n}}{\triangle t}=D\left(\frac{u_{i+1j}^{n+1}-2u_{ij}^{n+1}+u_{i-1j}^{n+1}}{\triangle x^2}+\frac{u_{ij+1}^{n+1}-2u_{ij}^{n+1}+u_{ij-1}^{n+1}}{\triangle y^2}\right),$$

or

$$\boxed{-\alpha(u_{i+1j}^{n+1}+u_{i-1j}^{n+1})+(1+2\alpha+2\beta)u_{ij}^{n+1}-\beta(u_{ij+1}^{n+1}+u_{ij-1}^{n+1})=u_{ij}^{n}.} \tag{10.18}$$

Let us consider the approximation (10.18) on the 5×5 grid, i.e., $i=j=0,\ldots,4$. Moreover, suppose that Dirichlet boundary conditions are given, that is, all values u_{0j}, u_{4j}, u_{i0}, u_{i4} are known. Suppose also that $n=1$ and define $\gamma=1+2\alpha+2\beta$. Then the approximation above leads to the neun algebraic equations:

$$\begin{aligned}
-\alpha u_{21}^{2}+\gamma u_{11}^{2}-\beta u_{12}^{2} &= u_{11}^{1}+\alpha u_{01}^{2}+\beta u_{10}^{2},\\
-\alpha u_{22}^{2}+\gamma u_{12}^{2}-\beta(u_{13}^{2}+u_{11}^{2}) &= u_{12}^{1}+\alpha u_{02}^{2},\\
-\alpha u_{23}^{2}+\gamma u_{13}^{2}-\beta u_{12}^{2} &= u_{13}^{1}+\alpha u_{03}^{2}+\beta u_{14}^{2},\\
-\alpha(u_{31}^{2}+u_{11}^{2})+\gamma u_{21}^{2}-\beta u_{22}^{2} &= u_{21}^{1}+\beta u_{20}^{2},\\
-\alpha(u_{32}^{2}+u_{12}^{2})+\gamma u_{22}^{2}-\beta(u_{23}^{2}+u_{21}^{2}) &= u_{22}^{1},\\
-\alpha u_{21}^{2}+\gamma u_{31}^{2}-\beta u_{32}^{2} &= u_{31}^{1}+\alpha u_{41}^{2}+\beta u_{30}^{2},\\
-\alpha u_{22}^{2}+\gamma u_{32}^{2}-\beta(u_{33}^{2}+u_{31}^{2}) &= u_{32}^{1}+\alpha u_{42}^{2},\\
-\alpha u_{23}^{2}+\gamma u_{33}^{2}-\beta u_{32}^{2} &= u_{33}^{1}+\alpha u_{44}^{2}+\beta u_{34}^{2}.
\end{aligned}$$

Formally, one can rewrite the system above to the matrix form $A\mathbf{u}=\mathbf{b}$, i.e.,

$$\left(\begin{array}{ccc|ccc|ccc}
\gamma & -\beta & 0 & -\alpha & 0 & 0 & 0 & 0 & 0\\
-\beta & \gamma & -\beta & 0 & -\alpha & 0 & 0 & 0 & 0\\
0 & -\beta & \gamma & 0 & 0 & -\alpha & 0 & 0 & 0\\
\hline
-\alpha & 0 & 0 & \gamma & -\beta & 0 & -\alpha & 0 & 0\\
0 & -\alpha & 0 & -\beta & \gamma & -\beta & 0 & -\alpha & 0\\
0 & 0 & -\alpha & 0 & -\beta & \gamma & 0 & 0 & -\alpha\\
\hline
0 & 0 & 0 & -\alpha & 0 & 0 & \gamma & -\beta & 0\\
0 & 0 & 0 & 0 & -\alpha & 0 & -\beta & \gamma & -\beta\\
0 & 0 & 0 & 0 & 0 & -\alpha & 0 & -\beta & \gamma
\end{array}\right)
\begin{pmatrix} u_{11}^{2}\\ u_{12}^{2}\\ u_{13}^{2}\\ \hline u_{21}^{2}\\ u_{22}^{2}\\ u_{23}^{2}\\ \hline u_{31}^{2}\\ u_{32}^{2}\\ u_{33}^{2}\end{pmatrix}
=
\begin{pmatrix} u_{11}^{1}+\alpha u_{01}^{2}+\beta u_{10}^{2}\\ u_{12}^{1}+\alpha u_{02}^{2}\\ \hline u_{13}^{1}+\alpha u_{03}^{2}+\beta u_{14}^{2}\\ \hline u_{21}^{1}+\beta u_{20}^{2}\\ u_{22}^{1}\\ u_{23}^{1}+\beta u_{24}^{2}\\ \hline u_{31}^{1}+\alpha u_{41}^{2}+\beta u_{30}^{2}\\ u_{32}^{1}+\alpha u_{42}^{2}\\ u_{33}^{1}+\alpha u_{44}^{1}+\beta u_{34}^{2}\end{pmatrix}$$

The matrix A is a five-band matrix. Nevertheless, despite of the fact that the schema is absolute stable, two of five bands are desposed so far apart from the main diagonal, that simple $\mathscr{O}(n)$ algorithms like TDMA are difficult or even impossible to apply.

The ADI Method

The idea of the ADI-method (*alternating direction implicit*) is to alternate direction and thus solve two one-dimensional problem at each time step [35]. The first step keeps y-direction fixed:

$$\frac{u_{ij}^{n+1/2}-u_{ij}^{n}}{\triangle t/2}=D\left(\frac{u_{i+1j}^{n+1/2}-2u_{ij}^{n+1/2}+u_{i-1j}^{n+1/2}}{\triangle x^2}+\frac{u_{ij+1}^{n}-2u_{ij}^{n}+u_{ij-1}^{n}}{\triangle y^2}\right).$$

In the second step we keep x-direction fixed:

$$\frac{u_{ij}^{n+1}-u_{ij}^{n+1/2}}{\triangle t/2}=D\left(\frac{u_{i+1j}^{n+1/2}-2u_{ij}^{n+1/2}+u_{i-1j}^{n+1/2}}{\triangle x^2}+\frac{u_{ij+1}^{n+1}-2u_{ij}^{n+1}+u_{ij-1}^{n+1}}{\triangle y^2}\right).$$

Both equations can be written in a triadiagonal form. Define

$$\alpha=\frac{D\triangle t}{2\triangle x^2},\qquad \beta=\frac{D\triangle t}{2\triangle y^2}.$$

Than we get:

$$\begin{aligned}-\alpha u_{i+1j}^{n+1/2}+(1+2\alpha)u_{ij}^{n+1/2}-\alpha u_{i-1j}^{n+1/2}&=\beta u_{ij+1}^{n}+(1-2\beta)u_{ij}^{n}+\beta u_{ij-1}^{n}\\ -\beta u_{ij+1}^{n+1}+(1+2\beta)u_{ij}^{n+1}-\beta u_{ij-1}^{n+1}&=\alpha u_{i+1j}^{n+1/2}+(1-2\alpha)u_{ij}^{n+1/2}+\alpha u_{i+1j}^{n+1/2}.\end{aligned}\tag{10.19}$$

Instead of five-band matrix in BTCS method (10.18), here each time step can be obtained in two sweeps. Each sweep can be done by solving a tridiagonal system of equations. The ADI-method is second order in time and space and is absolute stable [22] (however, the ADI method in 3D is conditional stable only).

10.2.2 Examples

Use the ADI method (10.19) to solve the two-dimensional diffusion equation

$$\partial_t u(\mathbf{r},t)=\triangle u(\mathbf{r},t)\,,$$

where $u=u(\mathbf{r},t)$, $\mathbf{r}\subseteq\mathbb{R}^2$ on the interval $r\in[0,L]\times[0,L]$, if the initial distribution is a Gauss pulse of the form $u(x,0)=\exp(-20\,(x-L/2)^2-20\,(y-L/2)^2)$ and the density on both ends of the interval is given as $u_r(0,t)=u_r(L,t)=0$. Other parameters are choosen according to the table below.

`Space interval`	$L=1$
`Amount of points`	$M=100$, $(\triangle x=\triangle y)$
`Time discretization step`	$\triangle t=0.001$
`Amount of time steps`	$T=40$

Solution of the problem is shown on Fig. (10.9).

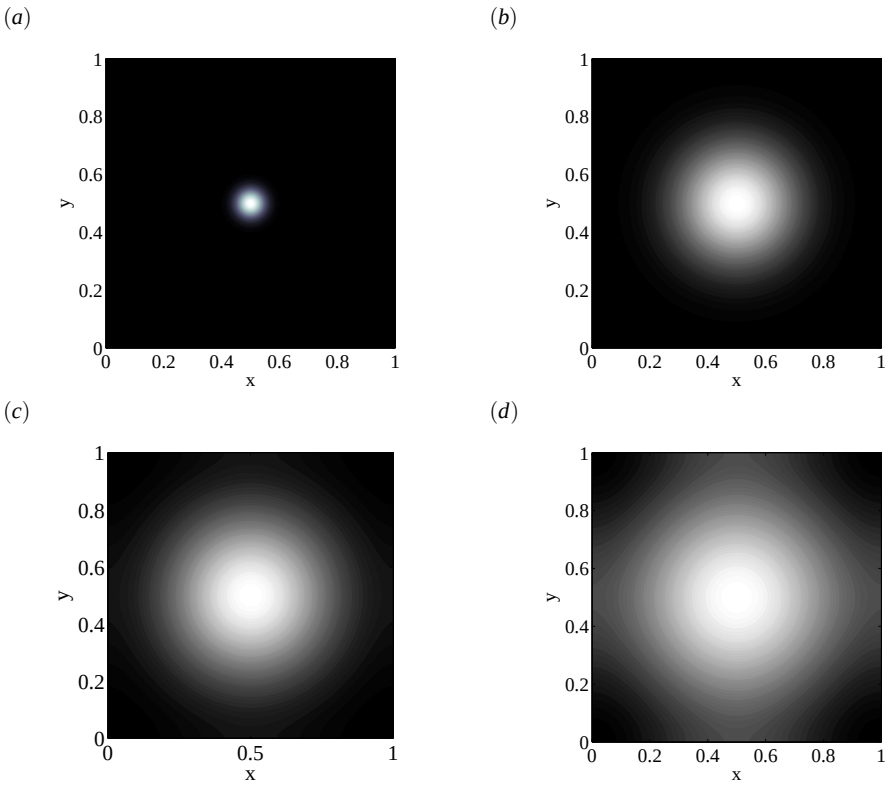

Fig. 10.9 Numerical solution of the two-dimensional diffusion equation 10.2 by means of the ADI method (10.19), calculated at four different time moments: (a) t=0; (b) t=10; (c) t=20; (d) t=40.

Chapter 11
The Reaction-Diffusion Equations

Reaction-diffusion (RD) equations arise naturally in systems consisting of many interacting components, (e.g., chemical reactions) and are widely used to describe pattern-formation phenomena in variety of biological, chemical and physical systems. The principal ingredients of all these models are equation of the form

$$\partial_t \mathbf{u} = \mathbf{D}\nabla^2 \mathbf{u} + \mathbf{R}(\mathbf{u}), \tag{11.1}$$

where $\mathbf{u} = \mathbf{u}(\mathbf{r}, \mathbf{t})$ is a vector of concentration variables, $\mathbf{R}(\mathbf{u})$ describes a local reaction kinetics and the Laplace operator ∇^2 acts on the vector $\mathbf{u}$ componentwise. $\mathbf{D}$ denotes a diagonal diffusion coefficient matrix. Note that we suppose the system (11.1) to be isotropic and uniform, so $\mathbf{D}$ is represented by a scalar matrix, independent on coordinates.

11.1 Reaction-diffusion equations in 1D

In the following sections we discuss different nontrivial solutions of this system (11.1) for different number of components, starting with the case of one component RD system in one spatial dimension, namely

$$u_t = D u_{xx} + R(u), \tag{11.2}$$

where $D = \text{const}$. Suppose, that initial distribution $u(x,0)$ is given on the whole space interval $x \in (-\infty, +\infty)$.

11.1.1 The FKPP-Equation

Investigation in this field starts form the classical papers of Fisher [17] and Kolmogorov, Petrovsky and Piskunoff [25] motivated by population dynamics issues, where authors arrived at a modified diffusion equation:

$$\partial_t u(x,t) = D\,\partial_x^2 u(x,t) + R(u), \tag{11.3}$$

with a nonlinear source term $R(u) = u - u^2$. A typical solution of the Eq. (11.3) is a propagating front, separating two non-equilibrium homogeneous states, one of which ($u = 1$) is stable and another one ($u = 0$) is unstable [10, 13, 51]. Such fronts behavior is often said to be *front propagation into unstable state* and fronts as such are referred to as *waves (or fronts) of transition from an unstable state.*

Initially the subject was discussed and investigated mostly in mathematical society (see, e.g., [16] where nonlinear diffusion equation was discussed in details). The interest in physics in these type of fronts was stimulated in the early 1980s by the work of G. Dee and coworkers on the theory of dendritic solidification [12]. Examples of such fronts can be found in various physical [28, 52], chemical [43, 14] as well as biological [3] systems.

Notice that for Eq. (11.3) the propagating front always relaxes to a unique shape and velocity

$$c^* = 2\sqrt{D}, \tag{11.4}$$

if the initial profile is well-localized [1, 2, 50].

Numerical treatment

Let us consider Eq. (11.3) and suppose that initial distribution $u(x,0) = f(x)$ as well as no-flux boundary conditions are given. We can try to apply an implicit BTCS-method (10.13) (see Chapter 10) for the linear part of the equation, taking the nonlinearilty explicitly, i.e.,

$$\frac{u_i^{j+1} - u_i^j}{\triangle t} = D\,\frac{u_{i+1}^{j+1} - 2u_i^{j+1} + u_{i-1}^{j+1}}{\triangle x^2} + R(u_i^j)\,,$$

where $R(u_i^j) = u_i^j - (u_i^j)^2$. We can rewrite the last equation to the matrix form

$$A\,\mathbf{u^{n+1}} = \mathbf{u^n} + \triangle t \cdot R(\mathbf{u^n})\,, \tag{11.5}$$

where matrix A is a tridiagonal $M+1 \times M+1$ matrix of the form

$$A = \begin{pmatrix} 1+2\alpha & \boxed{-2\alpha} & 0 & \ldots 0 \\ -\alpha & 1+2\alpha & -\alpha & \ldots 0 \\ 0 & -\alpha & 1+2\alpha & \ldots 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & \boxed{-2\alpha} & 1+2\alpha \end{pmatrix},$$

$\alpha = D\triangle t/\triangle x^2$. The boxed elements indicate the influence of no-flux boundary conditions.

As an example, let us solve Eq. (11.3) on the interval $x \in [-L, L]$ with the scheme (11.5). Parameters are:

`Space interval`	$L = 50$
`Space discretization step`	$\triangle x = 0.2$
`Time discretization step`	$\triangle t = 0.05$
`Amount of time steps`	$T = 800$
`Diffusion coefficient`	$D = 1$
`Initial distribution`	$f(x) = 0.05\exp(-5x^2)$

Numerical solution for six different time moments is shown on Fig. (11.1). One can see, that a small local initial fluctuation around $u = 0$ leads to an instability, that develops in a nonlinear way: a front propagates away from the initial perturbation. Finally the uniform stable state with $u = 1$ is established on the whole space interval.

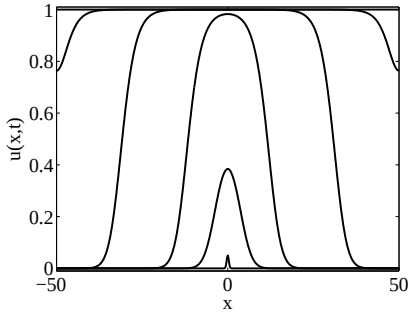

Fig. 11.1 Numerical solution of (11.3) calculated with the method (11.5) for six different time moments $t = 0, 100, 200, 400, 600, 800$.

11.1.2 Switching waves

Another important class of one-component RD systems is so-called *bistable systems*. They possess two stable states, say $u = u_-$ and $u = u_+$, separated by an unstable state $u = u_0$.

An example of bistable system is the Zeldovich–Frank–Kamenetsky–Equation, namely Eq. (11.2) with the reaction term

$$R(u) = u\,(1-u)\,(u-\beta)\,, \quad \beta \in (0,\,1)\,,$$

describing the flame propagation [54]

$$u_t = D u_{xx} + u(1-u)(u-\beta)\,, \quad \beta \in (0,\,1)\,. \tag{11.6}$$

The fundamental form of a pattern in bistable infinite one-component media is *a trigger wave*, which represents a propagating front of transition from one stationary state into the other. In the literature other nomenclature, e.g., *switching waves* is also used. The propagation velocity of a flat front is uniquely determined by the properties of the bistable medium. Indeed, moving to a frame, moving with a constant velocity $\xi := x - ct$, and considering partial solution of the form $u = u(\xi)$ one obtains an equation

$$D u_{\xi\xi} + c\,u_\xi + R(u) = 0$$

with boundary conditions

$$u(\xi \to -\infty) = u_-\,, \qquad u(\xi \to +\infty) = u_+\,.$$

Introducing the potential $R(u) = \frac{\partial V(u)}{\partial u}$ one can show that in this situation the velocity of the front can be determined as [16]

$$c = \frac{V(u_+) - V(u_-)}{\int\limits_{-\infty}^{+\infty} (u_\xi)^2 d\,\xi}\,.$$

The numerator of the last equation uniquely defines the velocity direction. In particular, if $V(u_+) = V(u_-)$ the front velocity equals zero, so *stationary front* is also a solution in bistable one-component media. However, the localized states in form of a domain, which can be produced by a connection of two fronts propagating in opposite directions, are normally unstable. Indeed, for the arbitrary choice of parameters one state ($V(u_+)$ or $V(u_-)$) will be dominated. This causes either collapse or expansion of the two-front solution.

Example 1: Moving fronts

Let us solve Eq. (11.6) on the interval $x \in [-L, L]$ with no-flux boundary conditions by means of numerical scheme (11.5). Other parameters are:

`Space interval`	$L = 10$
`Space discretization step`	$\triangle x = 0.04$
`Time discretization step`	$\triangle t = 0.05$
`Amount of time steps`	$T = 150$
`Diffusion coefficient`	$D = 1$

Consider four different cases, corresponding to different behaviors of the front:

a) A front moving to the right: $\beta = 0.8$;
b) A front moving to the left: $\beta = 0.1$.
   Initial distribution are:

$$u(x,0) = \begin{cases} u_-\,, & \text{for} \quad x \in [-L, 0] \\ u_+\,, & \text{for} \quad x \in (0, L]\,. \end{cases}$$

c) Front collision: $\beta = 0.8$;
d) Front scattering: $\beta = 0.1$.
   Initial distribution are:

$$u(x,0) = \begin{cases} u_-\,, & \text{for} \quad x \in [-L, -L/3] \\ u_+\,, & \text{for} \quad x \in (-L/3, L/3) \\ u_-\,, & \text{for} \quad x \in [L/3, L]\,. \end{cases}$$

Results of the numerical calculation is shown on Fig. 11.2.

Example 2: Stationary fronts

Now let us consider a one-dimensional RD equation (11.6), describing a bistable media for the case $\beta = -1$, i.e,

$$u_t = D u_{xx} + u(1 - u^2)\,, \qquad x \in [-L, L]\,. \tag{11.7}$$

Equation (11.7) has three steady state solutions: two stable $u_\pm = \pm 1$, separated with an unstable state $u_0 = 0$. One can calculate the potential values at $u = u_\pm$,

$$V(u_-) = V(u_+) \Rightarrow c = 0\,.$$

That is, a stationary front, connecting stable steady state is expected to be a solution of the problem. Moreover, one can constuct a localized pulse by a connection of two stable fronts. The form of the stationary front can be found analytically [16, 10], namely

$$u(x) = \tanh\left(\frac{x - x_0}{\sqrt{2D}}\right).$$

From numerical point of view one can use again the scheme (11.5) for the reaction term $R(u) = u - u^3$. That is, let us solve Eq. (11.7) on the interval $x \in [-L, L]$ with no-flux boundary conditions. Parameters are:

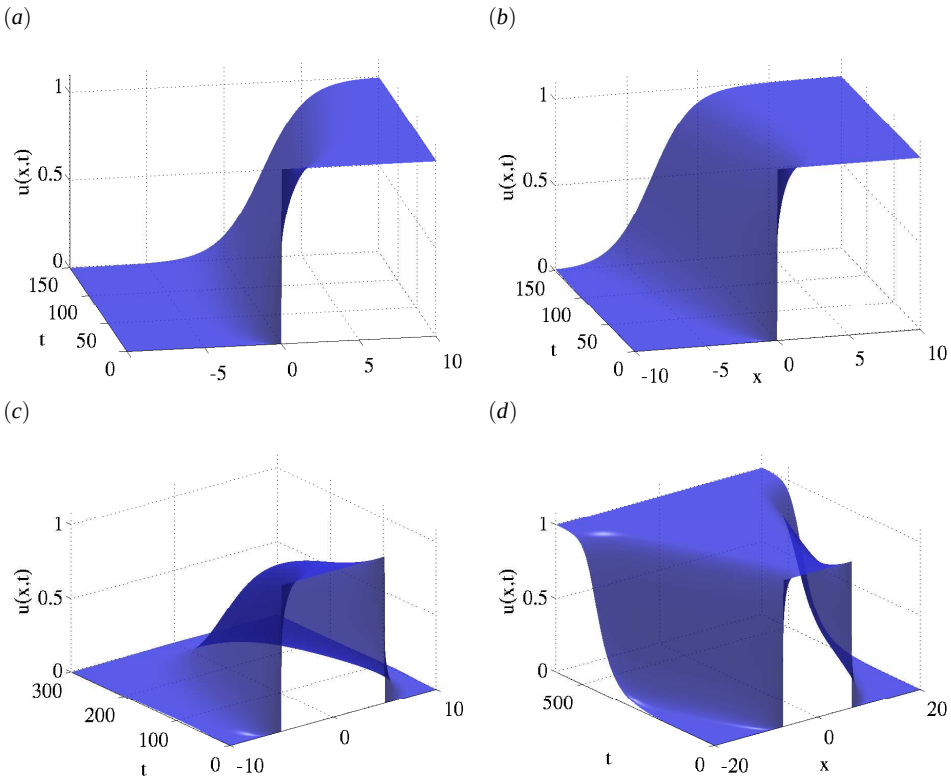

Fig. 11.2 Numerical solution of Eq. (11.6), calculated with the scheme (11.5) for four different cases: a) a front, propagating to the right for $\beta = 0.8$; b) a front, propagating to the left for $\beta = 0.1$; c) collision of two fronts, $\beta = 0.8$; d) scattering of two fronts, $\beta = 0.1$.

`Space interval`	$L = 10$
`Space discretization step`	$\triangle x = 0.04$
`Time discretization step`	$\triangle t = 0.05$
`Amount of time steps`	$T = 100$
`Diffusion coefficient`	$D = 1$

Initial distribution is:

a) A stationary front:

$$u(x,0) = \begin{cases} u_-\,, & \text{for} \quad x \le 0\,, \\ u_+\,, & \text{for} \quad x > 0\,. \end{cases}$$

b) A stationary pulse:

$$u(x,0) = \begin{cases} u_-\,, & \text{for} \quad x \in [-L, -L/4]\,, \\ u_+\,, & \text{for} \quad x \in (-L/4, L/4)\,, \\ u_-\,, & \text{for} \quad x \in [L/4, L]\,. \end{cases}$$

Solutions of the problem, corresponding to both cases are shown on Fig. 11.3.

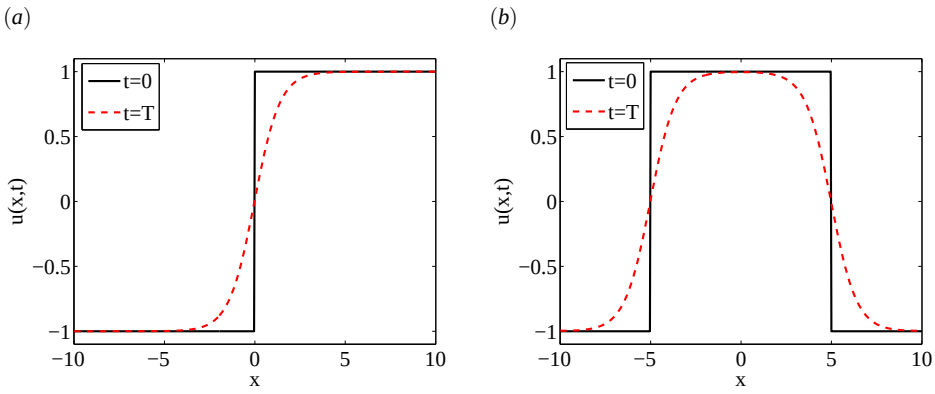

Fig. 11.3 Numerical solution of Eq. (11.7) by means of scheme (11.5): a) A stable stationary front. b) A stable stationary pulse.

11.2 Reaction-diffusion equations in 2D

11.2.1 Two-component RD systems: a Turing bifurcation

A Turing instability (or bifurcation) involves the destabilization of a homogeneus solution to form a static periodic spatial pattern (Turing pattern), whose wavelength depends on the local reaction kinetic parameters, diffusion coefficients of the system and is its intrinsic property. The hypothesis that just a difference in diffusion constants of components could be enough to destabilize the homogeneous solution was put forward by *A. M. Turing* in 1952 [49]. By studying the problem of biological morphogenesis he showed that a reaction-diffusion system with a different diffusion constants can autonomously produce stationary spatial patterns.

We start our analysis of Turing instability from by considering a reaction-diffusion system in general form, restricting ourself first to the case of two components, i.e.,

$$\partial_t \mathbf{u} = \mathbf{D}\nabla^2\mathbf{u} + \mathbf{R}(\mathbf{u}) \tag{11.8}$$

where $\mathbf{u} = \mathbf{u}(\mathbf{r},\mathbf{t}) = (u,v)^T$ is a vector of concentration variables, $\mathbf{R}(\mathbf{u}) = (f(u,v), g(u,v))^T$ describes as before a local reaction kinetics and the Laplace operator ∇^2 acts on the vector $\mathbf{u}$ componentwise. $\mathbf{D}$ denotes a diagonal diffusion coefficient matrix,

$$\mathbf{D} = \begin{pmatrix} D_u & 0 \\ 0 & D_v \end{pmatrix}.$$

Let $\mathbf{u_0} = (u_0, v_0)^T$ be a homogeneous solution (or steady-state solution) of the system (11.8), i.e. $f(u_0, v_0) = g(u_0, v_0) = 0$. Suppose that this solution is stable in absence of diffusion, namely the real parts of all eigenvalues of the Jacobi matrix

$$\mathbf{A} = (\partial\mathbf{R}/\partial\mathbf{u})_{\mathbf{u}=\mathbf{u_0}} = \begin{pmatrix} f_u & f_v \\ g_u & g_v \end{pmatrix},$$

describing the local dynamics of the system (11.8) are less that zero. For the case of a 2×2 matrix this is equivalent to the simple well-known condition for the trace and the determinant of the matrix $\mathbf{A}$ (Vieta's formula), namely

$$\begin{aligned} \mathrm{Sp}(\mathbf{A}) &= \lambda_1 + \lambda_2 = f_u + g_v < 0\,, \\ \det(\mathbf{A}) &= \lambda_1\lambda_2 = = f_u g_v - f_v g_u > 0\,. \end{aligned} \tag{11.9}$$

Keeping Eq. (11.9) in mind, let us see if the presence of diffusion term can change the stability of $\mathbf{u_0}$. To this end, consider a small perturbation $\tilde{\mathbf{u}}$, i.e. $\mathbf{u} = \mathbf{u_0} + \tilde{\mathbf{u}}$ and the corresponding linear equation for it:

$$\partial_t \tilde{\mathbf{u}} = \mathbf{D}\nabla^2 \tilde{\mathbf{u}} + \mathbf{A}\tilde{\mathbf{u}}\,. \tag{11.10}$$

After decomposition $\tilde{\mathbf{u}}$ into modes $\tilde{\mathbf{u}} \sim \mathbf{a_k} e^{ikr}$ we get the equation

$$\dot{\mathbf{a}}_\mathbf{k} = \mathbf{B}\mathbf{a}_\mathbf{k}\,, \tag{11.11}$$

where $\mathbf{B} = \mathbf{A} - k^2\mathbf{D}$.

As mentioned above, the stability conditions for the system (11.11) with a 2×2 matrix $\mathbf{B}$ can be written as:

$$\begin{aligned} \mathrm{Sp}(\mathbf{B}) &< 0 \quad \forall k\,, \\ \det(\mathbf{B}) &> 0 \quad \forall k\,, \end{aligned} \tag{11.12}$$

where

$$\mathrm{Sp}(\mathbf{B}) = -(D_u + D_v)k^2 + \mathrm{Sp}(\mathbf{A})\,, \tag{11.13}$$

$$\det(\mathbf{B}) = D_u D_v k^4 - (D_u g_v + D_v f_u)k^2 + \det(\mathbf{A})\,. \tag{11.14}$$

Notice, that for $k = 0$ the conditions (11.12) are equivalent to the stability criterion (11.9) for the local dynamics. In particular this implies that $\mathrm{Sp}(\mathbf{B}) < 0$ for all k (see gray curve in Fig. 11.4 for illustration), so the instability of the homogeneous solution can occur only due to violation of the second condition (11.12), that is, $\det(\mathbf{B})$ should be equal to zero for some k. It means that the instability occur at the point where the equation $\det(\mathbf{B}) = 0$ has a multiple root. To find it we can simply calculate a minimum of the function $T(k) = \det(\mathbf{B})$:

$$T'(k) = 4D_u D_v k^3 - 2(D_u g_v + D_v f_u)k = 0 \quad \Rightarrow \quad k^2 = \frac{1}{2}\left(\frac{f_u}{D_u} + \frac{g_v}{D_v}\right).$$

From the last equation can be seen that the situation described above is possible if

$$D_u g_v + D_v f_u > 0\,. \tag{11.15}$$

In this case the critical wavenumber is

$$k_c = \sqrt{\frac{1}{2}\left(\frac{f_u}{D_u} + \frac{g_v}{D_v}\right)} \tag{11.16}$$

and instability occurs on condition that

$$T(k_c) \le 0 \quad \Leftrightarrow \quad k_c^4 = \left(\frac{1}{2}\left(\frac{f_u}{D_u} + \frac{g_v}{D_v}\right)\right)^2 > \frac{\det\mathbf{A}}{D_u D_v}\,. \tag{11.17}$$

The instability scenario, described above is illustrated in Fig. 11.4, where three different cases of dependence of the function $T(k) = \det(\mathbf{B})$ on the wave vector k are presented. In Fig. 11.4 (a) the function $T(k)$ has no roots, so the stability of $\mathbf{u_0}$ is not affected as well as in the case (b). Here $T(k) > 0$ for all k, but minimum of this function exists. Finally, in Fig. 11.4 (c) $T(k) = 0$ for $k = k_c$, indicating the onset of instability.

Hence, the full system of the conditions for instability of the homogeneous solution $\mathbf{u_0}$ is

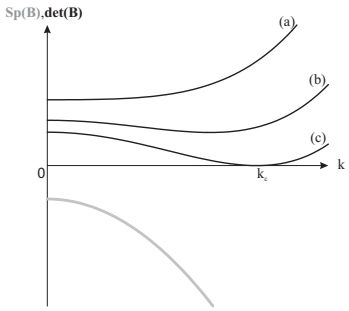

Fig. 11.4 Three different cases of dependence of the function $T(k) = \det(\mathbf{B})$ on the wave vector k are presented. (a) the function $T(k)$ has no roots, so the stability of $\mathbf{u_0}$ is not affected as well as in the case (b). $T(k) > 0$ for all k, but minimum of this function exists. (c) $T(k) = 0$ for $k = k_c$, indicating the onset of instability.

$$\boxed{\begin{aligned} f_u + g_v &< 0\,, \\ f_u g_v - f_v g_u &> 0\,, \\ D_u g_v + D_v f_u &> 0\,, \\ \left(\frac{f_u}{D_u} + \frac{g_v}{D_v}\right)^2 &> \frac{4\det\mathbf{A}}{D_u D_v}\,. \end{aligned}} \tag{11.18}$$

A detailed description of the mechanism of Turing instability can also be found in [32, 31, 23].

While the conditions for the onset of a Turing bifurcation are rather simple, the determination of the nature of the pattern that is selected is a more difficult problem since beyond the bifurcation point a finite band of wavenumbers is unstable. Pattern selection is usually approached by studying *amplitude equations* that are valid near the onset of the instability. To determine which modes are selected, modes and their complex conjugates are usually treated in pairs so that the concentration field, expanded about the homogeneous solution, reads

$$\mathbf{u}(\mathbf{r},t) = \mathbf{u_0} + \sum_{j=1}^{n} \left(A_j(t) e^{i\mathbf{k_j}\mathbf{r}} + c.c.\right),$$

where $\mathbf{k_j}$ are different wavevectors such that $|\mathbf{k_j}| = k_c$. In one dimensional space the situation is rather simple, as result of the instability is represented by a periodic in space structure. In two space dimension this form leads to *stripes* for $n = 1$, *rhombi* (or *squares*) for $n = 2$ and *hexagons* for $n = 3$. The pattern and wavelength that is selected depends on coefficients in the nonlinear amplitude equation for the complex amplitude A_j, but some conclusions about selected pattern can be made using, e.g., symmetry arguments. In particular, in the case of hexagonal pattern, in which three wave vectors are mutually situated at an angle of $2\pi/3$, i.e., $\mathbf{k_1} + \mathbf{k_2} + \mathbf{k_3} = 0$, the absence of inversion symmetry ($\mathbf{u} \mapsto -\mathbf{u}$) leads to additional quadratic nonlinearity in the amplitude equation. The latter, in its turn, ends in a fact, that hexagonal pattern has the maximum growth rate near the threshold and is therefor preferred (for details see [10]).

The general procedure in details for the derivation of such amplitude equations based on mode projection techniques can be found in [19]. Another approach, using multi scale expansion was evolved in [33].

11.2.1.1 The Brusselator Model

The Brusselator model is a classical reaction-diffusion system, proposed by I. Prigogine and co-workers in Brussels in 1971 [18, 34]. The model describes some chemical reaction with two components

$$u_t = D_u \triangle u + a - (b+1)\,u + u^2 v, \tag{11.19}$$
$$v_t = D_v \triangle v + b\,u - u^2 v. \tag{11.20}$$

Here $u = u(x,y,t)$, $v = v(x,y,t)$, a, b are posotive constants. The steady state solution is

$$u_0 = a, \quad v_0 = \frac{b}{a}.$$

For the system (11.19) the matrices **D**, **A** and **B** are given by

$$\mathbf{D} = \begin{pmatrix} D_u & 0 \\ 0 & D_v \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} b-1 & a^2 \\ -b & -a^2 \end{pmatrix},$$

and

$$\mathbf{B} = \begin{pmatrix} b-1-D_u k^2 & a^2 \\ -b & -D_v k^2 - a^2 \end{pmatrix}.$$

Suppose that the system (11.19) is local stable, i.e.,

$$\mathrm{Sp}(\mathbf{A}) = b-1-a^2 < 0,$$
$$\mathrm{Det}(\mathbf{A}) = -(b-1)\,a^2 + a^2 b = a^2 > 0.$$

Note that the violation of the first condition above leads to tthe Hopf bifurcation, i.e., the onset of Hopf instability is

$$\boxed{\mathrm{Sp}(\mathbf{A}) \geq 0 \Leftrightarrow b \geq b_H = 1 + a^2.}$$

The critical wavenumber is

$$k_c = \sqrt{\frac{1}{2}\left(\frac{b-1}{D_u} - \frac{a^2}{D_v}\right)}.$$

The existence of k_c is equivalent to the following condition

$$b > 1 + \frac{D_u}{D_v} a^2 + 1 \Rightarrow \frac{D_u}{D_v} < 1.$$

The instability occurs, if

$$\mathrm{Det}(\mathbf{B}(k_c)) \leq 0 \Leftrightarrow b > b_T = \left(1 + a\sqrt{\frac{D_u}{D_v}}\right)^2.$$

Hence, the conditions (11.18) for the system (11.19) takes the form

$$\boxed{\begin{array}{l} b < b_H = 1 + a^2, \\ b > b_T = \left(1 + a\sqrt{\dfrac{D_u}{D_v}}\right)^2, \\ \dfrac{D_u}{D_v} < 1. \end{array}} \tag{11.21}$$

On Fig. 11.5 both b_H (blue line), b_T (red line) as functions of a are shown. The thresholds of these two instabilities coincide at codimensional-two Turing-Hopf point $b_H = b_T$

$$a_c = \frac{2\sqrt{\sigma}}{1-\sigma},$$

where $\sigma = D_u/D_v$.

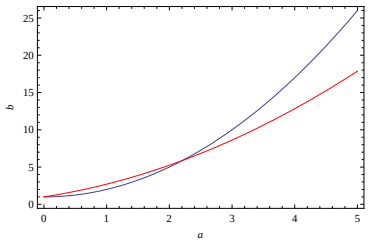

Fig. 11.5 Bifurcation diagram in (a, b) parameter space, indicating the onset of Hopf (blue line) and Turing (red line) instabilities. Here $D_u = 5$, $D_v = 12$.

From a numerical point of view, one can apply the scheme (10.19), taking the nonlinear terms explicitly. Parameters are

`Space interval`	$L = 50$
`Space discretization step`	$\triangle x = 0.5$
`Time discretization step`	$\triangle t = 0.05$
`Amount of time steps`	$T = 4000$
`Diffusion coefficients`	$D_u = 5$, $D_v = 12$
`Reaction kinetics`	$a = 3$, $b = 9$

The result of calculation is shown on Fig. 11.6. The uniform state becomes unstable in favor of finite wave number perturbation. That is, starting with random perturb homogeneous solution (see Fig. 11.6 (a)) one obtains a high-amplitude stripe pattern, shown in Fig. 11.6 (c).

(a) (b) (c)

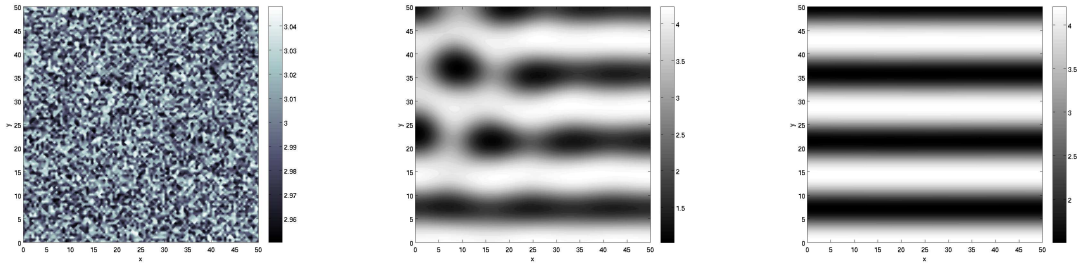

Fig. 11.6 Stripe pattern, obtained as a numerical solution of Eq. (11.19) by means of the modified ADI scheme (10.19) for three different time moments: a) $t = 0$; b) $t = 2000$; c) $t = 4000$.

Appendix A
Tridiagonal matrix algorithm

The tridiagonal matrix algorithm (TDMA), also known als *Thomas algorithm*, is a simplified form of Gaussian elimination that can be used to solve tridiagonal system of equations

$$a_i x_{i-1} + b_i x_i + c_i x_{i+1} = y_i, \qquad i = 1, \ldots n, \tag{A.1}$$

or, in matrix form ($a_1 = 0$, $c_n = 0$)

$$\begin{pmatrix} b_1 & c_1 & 0 & \ldots & \ldots & 0 \\ a_2 & b_2 & c_2 & \ldots & \ldots & 0 \\ 0 & a_3 & b_3 & c_3 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & c_{n-1} \\ 0 & \ldots & \ldots & 0 & a_n & b_n \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{pmatrix}$$

The TDMA is based on the Gaussian elimination procedure and consist of two parts: a forward elimination phase and a backward substitution phase [37]. Let us consider the system (A.1) for $i = 1 \ldots n$ and consider following modification of first two equations:

$$\mathrm{Eq}_{i=2} \cdot b_1 - \mathrm{Eq}_{i=1} \cdot a_2$$

which relults in

$$(b_1 b_2 - c_1 a_2) x_2 + c_2 b_1 x_3 = b_1 y_2 - a_2 y_1 .$$

The effect is that x_1 has been eliminated from the second equation. In the same manner one can eliminate x_2, using the modified second equation and the third one (for $i = 3$):

$$(b_1 b_2 - c_1 a_2)\mathrm{Eq}_{i=3} - a_3(\text{mod. } \mathrm{Eq}_{i=2}),$$

which would give

$$(b_3(b_1 b_2 - c_1 a_2) - c_2 b_1 a_3) x_3 + c_3(b_1 b_2 - c_1 a_2) x_4 = y_3(b_1 b_2 - c_1 a_2) - (y_2 b_1 - y_1 a_2) a_3$$

If the procedure is repeated until the n'th equation, the last equation will involve the unknown function x_n only. This function can be then used to solve the modified equation for $i = n - 1$ and so on, until all unknown x_i are found (backward substitution phase). That is, we are looking for a backward ansatz of the form:

$$x_{i-1} = \gamma_i x_i + \beta_i . \tag{A.2}$$

If we put the last ansatz in Eq. (A.1) and solve the resulting equation with respect to x_i, the following relation can be obtained:

$$x_i = \frac{-c_i}{a_i\gamma_i + b_i} x_{i+1} + \frac{y_i - a_i\beta_i}{a_i\gamma_i + b_i} \tag{A.3}$$

This relation possesses the same form as Eq. (A.2) if we identify

$$\boxed{\gamma_{i+1} = \frac{-c_i}{a_i\gamma_i + b_i}, \qquad \beta_{i+1} = \frac{y_i - a_i\beta_i}{a_i\gamma_i + b_i}}. \tag{A.4}$$

Equation (A.4) involves the recursion formula for the coefficients γ_i and β_i for $i = 2, \ldots, n-1$. The missing values γ_1 and β_1 can be derived from the first ($i = 1$) equation (A.1):

$$x_1 = \frac{y_1}{b_1} - \frac{c_1}{b_1} x_2 \Rightarrow \gamma_2 = -\frac{c_1}{b_1}, \beta_2 = \frac{1}{b_1} \Rightarrow \boxed{\gamma_1 = \beta_1 = 0.}$$

The last what we need is the value of the function x_n for the first backward substitution. We can obtain if we put the ansatz

$$x_{n-1} = \gamma x_n + \beta_n$$

into the last ($i = n$) equation (A.1):

$$a_n(\gamma x_n + \beta_n) + b_n x_n = y_n,$$

yielding

$$x_n = \frac{y_n - a_n\beta_n}{a_n\gamma_n + b_n}.$$

One can get this value directly from Eq. (A.2), if one formal puts

$$x_{n+1} = 0.$$

Altogether, the TDMA can be written as:

1. `Set` $\gamma_1 = \beta_1 = 0$;
2. `Evaluate for` $i = 1, \ldots, n-1$

$$\gamma_{i+1} = \frac{-c_i}{a_i\gamma_i + b_i}, \qquad \beta_{i+1} = \frac{y_i - a_i\beta_i}{a_i\gamma_i + b_i};$$

3. `Set` $x_{n+1} = 0$;
4. `Find for` $i = n+1, \ldots, 2$

$$x_{i-1} = \gamma_i x_i + \beta_i.$$

The algorithm admits $\mathcal{O}(n)$ operations instead of $\mathcal{O}(n^3)$ required by Gaussian elimination.

Limitation

The TDMA is only applicable to matrices that are diagonally dominant, i.e.,

$$|b_i| > |a_i| + |c_i|, \qquad i = 1, \ldots, n.$$

Appendix B
The Method of Characteristics

The method of characteristics is a method which can be used to solve *an initial value problem* for general first order PDEs [7]. Let us consider a quasilinear equation of the form

$$A\frac{\partial u}{\partial x}+B\frac{\partial u}{\partial t}+Cu=0, \qquad u(x,0)=u_0, \tag{B.1}$$

where $u=u(x,t)$, and A, B and C can be functions of independent variables and u. The idea of the method is to change coordinates from (x,t) to a new coordinate system (x_0,s), in which Eq. (B.1) becomes *an ordinary differential equation* along certain curves in the (x,t) plane. Such curves, $(x(s),t(s))$ along which the solution of (B.1) reduces to an ODE, are called the *characteristic curves*. The variable s can be varied, whereas x_0 changes along the line $t=0$ on the plane (x,t) and remains constant along the characteristics. Now if we choose

$$\frac{dx}{ds}=A, \quad \text{and} \quad \frac{dt}{ds}=B, \tag{B.2}$$

then we have

$$\frac{du}{ds}=u_x\frac{dx}{ds}+u_t\frac{dt}{ds}=Au_x+Bu_t,$$

and Eq. (B.1) becomes the ordinary differential equation

$$\frac{du}{ds}+Cu=0 \tag{B.3}$$

Equations (B.2) and (B.3) give the characteristics of (B.1).
That is, a general strategy to find out the characteristics of the system like (B.1) is as follows:

- Solve Eq. (B.2) with initial conditions $x(0)=x_0$, $t(0)=0$. Solutions of (B.2) give the transformation $(x,t)\rightarrow(x_0,s)$;
- Solve Eq. (B.3) with initilal condition $u(0)=u_0(x_0)$ (where x_0 are the initial points on the characteristic curves along the $t=0$ axis). So, we have a solution $u(x_0,s)$;
- Using the results of the first step find s and x_0 in terms of x and t and substitute these values in $u(x_0,s)$ to get the solution $u(x,t)$ of the original equation (B.1).

References

1. David Acheson. *From Calculus to Chaos*. Oxford University Press, New York, 1997.
2. D. G. Aronson and H. F. Weinberger. Multidimensional nonlinear diffusion arising in population genetics. *Advances in Mathematics*, 30:33–76, 1978.
3. E. Ben-Jacob, H. Brand, G. Dee, L. Kramer, and J. S. Lange. Pattern propagation in nonlinear dissipative systems. *Physica D*, 14:348–364, 1985.
4. N.F. Britton. *Reaction–Diffusion Equations and their Applications to Biology*. Academic Press, New York, 1986.
5. J. M. Burgers. *The Nonlinear Diffusion Equation*. D. Reidel, Dordrecht, 1974.
6. J. G. Charney, R. Fjoertoft, and J. von Neumann. Numerical integration of the barotropic vorticity equation. *Tellus*, 2:237–254, 1950.
7. J. D. Cole. On a quasi-linear parabolic equation occurring in aerodynamics. *Quarterly of Applied Mathematics*, 9:225–236, 1951.
8. R. Courant and D. Hilbert. *Methods of Mathematical Physics II*. Wiley, 1962.
9. R. Courant, E. Isaacson, and M. Rees. On the solution of non-linear hyperbolic differential equations. *Communications on Pure and Applied Mathematics*, 5:243–255, 1952.
10. J. Crank and P. Nicolson. A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Proceedings of the Cambridge Philosophical Society*, 43:50–67, 1947.
11. M. C. Cross and P. C. Hohenberg. Pattern formation outside of equilibrium. *Reviews of Modern Physics*, 65(3):851–1112, 1993.
12. A. S. Davydov. *Solitons in Molecular Systems (Mathematics and its Applications)*. Dordrecht, Netherlands: Reidel, 1985, 1985.
13. G. Dee and J. S. Langer. Propagating pattern selection. *Phys. Rev. Lett.*, 50:383–386, 1983.
14. U. Ebert and W. van Saarloos. Front propagation into unstable states: universal algebraic convergence towards uniformly translating pulled fronts. *Physica D*, 146:1–99, 2000.
15. I. R. Epstein and K. Showalter. Nonlinear chemical dynamics: Oscillations, patterns, and chaos. *J. of Phys. Chem.*, 100:13132–13147, 1996.
16. S. Eule and R. Friedrich. A note on the forced burgers equation. *Physics Letters A*, 351:238, 2006.
17. P. C. Fife. *Mathematical Aspects of Reacting and Diffusing Systems. Lecture Notes in Biomathematics*, volume 28. Springer, Berlin, 1979.
18. R. A. Fisher. The wave of advance of advantageous genes. *Ann. Eugenics*, 7:355, 1937.
19. P. Glandsdorff and I. Prigogine. *Thermodynamic theory of structure, stability and fluctuations*. Wiley, New York, 1971.
20. H. Haken. *Synergetics, An Introduction*. 3rd ed. Springer Ser. Synergetics, Berlin, Heidelberg, New York, 1983.
21. E. Hopf. The partial differential equation $u_t + uu_x = u_{xx}$. *Communications on Pure and Applied Mathematics*, 3:201–230, 1950.
22. E. Isaacson and H. B. Keller. *Analysis of numerical Method*. Wiley, 1965.
23. J. Douglas Jr. and J. E. Gunn. A general formulation of alternating direction methods: Part i. parabolic and hyperbolic problems. *Numerische Mathematik*, 6(1):428–453, 1964.
24. R. Kapral. Pattern formation in chemical systems. *Physica D*, 86(1-2):149–157, 1995.
25. W. Kinzel and G. Reents. *Physik per Computer. Programmierung physikalisher Probleme mit Mathematica und C*. Spektrum Akademischer Verlag, 1996.
26. A. Kolmogorov, I. Petrovsky, and N. Piscounov. A study of the equation of dissusion with increase in the quantity of matter, and its application to a bilological problem. *Moscow Univ. Bull. Math. A*, 1:1, 1937.
27. D. J. Korteweg and F. de Vries. On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves. *Philosophical Magazine*, 39:422–443, 1895.
28. G. L. Jr. Lamb. *Elements of Soliton Theory*. New York: Wiley, 1980.

29. J. S. Langer and H. Mueller-Krumbhaar. Mode selection in a dendritelike nonlinear system. *Phys. Rev. A*, 27:499, 1983.
30. P. D. Lax and B. Wendroff. Systems of conservation laws. *Communications on Pure and Applied Mathematics*, 13:217–237, 1960.
31. John H. Mathews and Kurtis D. Fink. *Numerical Methods Using Matlab*. Prentice Hall, New York, 1999.
32. A. S. Mikhailov. *Foundations of Synergetics I. Distributed Active Systems*, volume 51 of *Springer Series in Synergetics*. Springer, Berlin, 1990.
33. J. D. Murray. *Mathematical Biology*. Springer, Berlin, 1993.
34. A. C. Newell and J. A. Whitehead. *Review of the finite bandwidth concept*, pages 284–289. Springer-Verlag, Berlin, 1971.
35. G. Nicolis. Stability and dissipative structures in open systems far from equilibrium. *Advan. Chem. Phys*, 19:209, 1971.
36. D.W. Peaceman and H. H. Rachford Jr. The numerical solution of parabolic and elliptic differential equations. *Journal of the Society for Industrial and Applied Mathematics*, 3:28–41, 1955.
37. A. M. Polyakov. Turbulence without pressure. *Physical Review E*, 52:6183–6188, 1995.
38. William H. Press, Saul A. Teukolsky, and William T. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, www.nr.com, 1993.
39. K. Friedrichs R. Courant and H. Lewy. Ueber die partiellen differenzengleichungen der mathematischen physik. *Mathematische Annalen*, 100(1):32–74, 1928.
40. M. Remoissenet. *Waves Called Solitons. Concepts and Experiments*. Springer, Berlin, 2003.
41. P. J. Roache. *Fundamentals of Computational Fluid Dynamics*. Hermosa Publishers, 1998.
42. J. S. Russell. *Report on Waves. Report of the 14th Meeting of the British Association for the Advancement of Science*. London: John Murray, 1844.
43. Hans Rudolf Schwarz and Norbert Koeckler. *Numerische Mathematik*. Teubner, Wiesbaden, 2006.
44. A. C. Scott. A nonlinear klein-gordon equation. *American Journal of Physics*, 37:52–61, 1969.
45. K. Showalter and J.J. Tyson. Luther's 1906 discovery and analysis of chemical waves. *J. Chem. Educ.*, 64:742–744, 1987.
46. G. D. Smith. *Numerical Solution of Partial Differential Equations: Finite Difference Methods*. Oxford University Press, 3rd edition, 1985.
47. Josef Stoer and Roland Bulirsch. *Numerische Mathematik 2*. Springer, Berlin, 2000.
48. Steven H. Strogaz. *Nonlinear Dynamics and Chaos*. Perseus Books Publishing, NewYork, 1994.
49. Michael Tabor. *Chaos and Integrability in Nonlinear Dynamics: An Introduction*. A Wiley-Interscience Publication, 1989.
50. J. C. Tannehill, D. A. Anderson, and R. H. Pletcher. *Computational Fluid Mechanics and Heat Transfer*. Taylor Francis, 1997.
51. J. W. Thomas. *Numerical Partial Differential Equations: Finite Difference Methods*. Springer, 1995.
52. A. M. Turing. The chemical basis of morphogenesis. *Phil. Trans. Roy. Soc. B*, 237:37–72, 1952.
53. W. van Saarloos. Front propagation into unstable states. ii. linear versus nonlinear marginal stability and rate of convergence. *Phys. Rev. A*, 39:6367, 1989.
54. W. van Saarloos. Front propagation into unstable states. *Physics Reports*, 386:29–222, 2003.
55. W. van Saarloos, M. van Hecke, and R. Hołyst. Front propagation into unstable and metastable states in smectic-c* liquid crystals: Linear and nonlinear marginal-stability analysis. *Phys. Rev. E.*, 52:1773, 1995.
56. N. J. Zabusky and M. D. Kruskal. Interaction of solitons in a collisionless plasma and the recurrence of initial states. *Physical Review Letters*, 15:240–243, 1965.
57. Y. B. Zeldovich and D. A. Frank-Kamenetsky. A theory of thermal propagation of flame. *Acta Physicochim. URSS*, 9:341–350, 1938.